Universitatea de Ştiinţe Agricole şi Medicină Veterinară Cluj-Napoca
Şcoala Doctorală
Facultatea de Horticultură

Lorentz JÄNTSCHI

Teză de doctorat

Algoritmi genetici şi aplicații ale acestora

PhD Thesis

Genetic algorithms and their applications

Conducător Ştiinţific:
Prof. Univ. Dr. Radu E. SESTRAŞ

Cluj-Napoca
2010

Cuprins

Introducere **1**

STADIUL ACTUAL AL CUNOAŞTERII

1. Probleme de optimizare a relațiilor structură-activitate **4**

§1.1 ..... Scurt istoric

§1.2 ..... Ce se pretează la relații structură-activitate

§1.3 ..... Ce sunt relațiile structură-activitate

§1.4 ..... Elaborarea și valorificarea de relații structură-activitate

2. Simularea evoluției cu algoritmi genetici **11**

§2.1 ..... Scurt istoric

§2.2 ..... Ce se pretează la algoritmi genetici

§2.3 ..... Ce sunt algoritmii genetici

§2.4 ..... Complexitatea algoritmică

§2.5 ..... Construcția algoritmilor genetici

§2.6 ..... Elaborarea și valorificarea algoritmilor genetici

§2.7 ..... Optimizare vs. simulare

§2.8 ..... Cadrul cercetării

§2.9 ..... Domeniul de aplicabilitate

SCOPUL & OBIECTIVELE CERCETĂRII

3. Simularea evoluției cu algoritmi genetici în probleme de optimizare a relațiilor structură-activitate **50**

MATERIAL & METODĂ

4. Definirea problemei de optimizare a relației structură-activitate **51**

§4.1 ..... Seria de compuși incluși în studiu

§4.2 ..... Modelare moleculară

§4.3 ..... Familia de descriptori

§4.4 ..... Relația structură-activitate

5. Definirea problemei genetice de evoluție **55**

§5.1 ..... Spațiul de căutare

§5.2 ..... Algoritmul genetic

§5.3 ..... Obiectivul evoluției

6. Definirea experimentului de simulare a evoluției **68**

§6.1 ..... Observabile

§6.2 ..... Informații cu privire la evoluție în fișierul PCB_XXXX_evo.txt

§6.3 ..... Informații cu privire la evoluție în fișierul PCB_XXXX_cfg.txt

§6.4 ..... Configurație program

§6.5. ..... Design experimental

REZULTATE & DISCUȚII

7. Implementare algoritm genetic; rezultate & verificarea datelor **80**

§7.1 ..... Strategii pentru selecție și supraviețuire

§7.2 ..... Rezultate

§7.3 ..... Verificarea datelor

8. Analiza variabilității **87**

§8.1 ..... Testarea independenței între strategii

§8.2 ..... Compararea evoluțiilor

§8.3 ..... Diversitatea genotipurilor

9. Analiza calitativă (observații pe observabile calitative) **102**

§9.1 ..... Genotipuri

§9.2 ..... Fenotipuri

§9.3 ..... Asocieri de fenotipuri
10. Analiza obiectivului evoluției folosind un eşantion întâmplător de generații...121
§10.1 .....Obținerea eşantionului din distribuția uniformă discretă
§10.2 .....Căutarea legii de distribuție a coeficientului de determinare al unei generații
§10.3 .....Media nu este o statistică suficientă
§10.4 ..... Distribuția Gumbel nu este suficient de generală
11. Analiza legii de distribuție a obiectivului evoluției aplicând un studiu sistematic în cadru generalizat 129
§11.1 ..... Legea de distribuție Fisher-Tippett a obiectivului evoluției
§11.2 .....Evoluția nu este polinomială
§11.3 Evoluția locației, scalei şi formei distribuției Fisher-Tippett a obiectivului evoluției
§11.4 ..... Caracterizarea statistică a distribuției evoluției
12. Ce lege urmează momentele de apariție a evoluției?.....157
§12.1 ..... Familia de curbe log-Pearson de tipul III
§12.2 ..... Asocieri în familia de curbe de evoluție
§12.3 .....Diferențe în familia de curbe de evoluție
§12.4 .....Ce lege urmează numărul de evoluții?

Concluzii & Recomandări 174

Referințe 188

Publicații.....208
§P.1 ..... "On about what Can Be Done and what Cannot Be Done with Genetic ..."
§P.2 ..... "Hard Problems in Gene Sequence Analysis ..."
§P.3 ....."Classical Approaches of Genetic Algorithms ..."
§P.4 ..... "Distribution Fitting 1. Parameters Estimation ..."
§P.5 ....."Distribution Fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, ..."
§P.6 ..... "Distribution Fitting 3. Analysis under Normality Assumption"
§P.7 ....."Distribution Fitting 4. Benford test on a sample of observed genotypes ..."
§P.8 ....."(Distribution Fitting 5.)..."
§P.9 ....."(Distribution Fitting 6.)..."
§P.10 ....."(Distribution Fitting 7.)..."
§P.11 ....."(Distribution Fitting 8.)..."
§P.12 ....."(Distribution Fitting 9.)..."
§P.13 ....."(Distribution Fitting 10.)..."
§P.14 ....."Meta-heuristics on quantitative structure-activity relationships..."
§P.15 ....."A Study of Genetic Algorithm Evolution..."
§P.16 ..... "A genetic algorithm for structure-..."
§P.17 ..... "Recording Evolution Supervised by a Genetic Algorithm..."
§P.18 ..... Rezultate în premieră - posibil de publicat

Anexe 226

Contents

Introduction **1**

LITERATURE REVIEW

1. Problems of structure-activity relationships optimization **4**

§1.1 ..... Short history

§1.2 ..... The opportunity and suitability of structure-activity relationships

§1.3 ..... What the structure-activity relationships are

§1.4 ..... Design and usage of structure-activity relationships

2. Simulating evolution with genetic algorithms **11**

§2.1 ..... Short history

§2.2 ..... The opportunity and suitability of genetic algorithms

§2.3 ..... What the genetic algorithms are

§2.4 ..... Algorithmic complexity

§2.5 ..... Design of genetic algorithms

§2.6 ..... Design and usage of genetic algorithms

§2.7 ..... Optimization vs. simulation

§2.8 ..... Research frame

§2.9 ..... Applicability domain

AIM & OBJECTIVES OF THE RESEARCH

3. Simulating evolution with genetic algorithms in problems of structure-activity relationships optimization **50**

MATERIAL & METHOD

4. Definition of the structure-activity relationship optimization problem **51**

§4.1 ..... The series of compounds included in the study

§4.2 ..... Molecular modelling

§4.3 ..... Family of descriptors

§4.4 ..... Structure-activity relationship

5. Definition of the genetic problem of evolution **55**

§5.1 ..... Search space

§5.2 ..... Genetic algorithm

§5.3 ..... Evolution objective

6. Definition of the experiment simulating evolution **68**

§6.1 ..... Observables

§6.2 ..... Information regarding evolution in PCB_XXXX_evo.txt file

§6.3 ..... Information regarding evolution in PCB_XXXX_cfg.txt file

§6.4 ..... Program configuration

§6.5. ..... Experimental design

RESULTS & DISCUSSIONS

7. Genetic algorithm implementation; results & checking of data **80**

§7.1 ..... Strategies for selection and survival

§7.2 ..... Results

§7.3 ..... Data checking

8. Variability analysis **87**

§8.1 ..... Testing the independence between strategies

§8.2 ..... Comparing evolutions

§8.3 ..... Diversity of the genotypes

9. Qualitative analysis (observations of qualitative observables) **102**

§9.1 ..... Genotypes

§9.2 ..... Phenotypes

§9.3 ..... Phenotype associations
10. Evolution objective's analysis using a random sample of generations 121
§10.1 ..... Obtaining of the sample from discrete uniform distribution
§10.2 ..... Searching for the distribution law of determination coefficient of a generation
§10.3 ..... The mean is not a sufficiency statistic
§10.4 ..... Gumbel distribution is not general enough
11. Evolution objective's distribution law analysis in a systematic study in a generalized frame.....129
§11.1 ..... Fisher-Tippett distribution of evolution's objective
§11.2 ..... Evolution is not polynomial
§11.3 .... Location, scale, and shape of Fisher-Tippett distribution of evolution's objective
§11.4 ..... Statistical description of the evolution distribution
12. Which law follows the moments of evolution?.....157
§12.1 ..... log-Pearson of type III family of curves
§12.2 ..... Associations in the family of evolution curves
§12.3 ..... Differences in the family of evolution curves
§12.4 ..... The distribution law for the number of evolutions?

Conclusions & Recommendations.....174

References 188

Publications 208
§P.1 ..... "On about what Can Be Done and what Cannot Be Done with Genetic ..."
§P.2 ..... "Hard Problems in Gene Sequence Analysis ..."
§P.3 ..... "Classical Approaches of Genetic Algorithms ..."
§P.4 ..... "Distribution Fitting 1. Parameters Estimation ..."
§P.5 ..... "Distribution Fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, ..."
§P.6 ..... "Distribution Fitting 3. Analysis under Normality Assumption"
§P.7 ..... "Distribution Fitting 4. Benford test on a sample of observed genotypes ..."
§P.8 ..... "(Distribution Fitting 5.)..."
§P.9 ..... "(Distribution Fitting 6.)..."
§P.10 ..... "(Distribution Fitting 7.)..."
§P.11 ..... "(Distribution Fitting 8.)..."
§P.12 ..... "(Distribution Fitting 9.)..."
§P.13 ..... "(Distribution Fitting 10.)..."
§P.14 ..... "Meta-heuristics on quantitative structure-activity relationships..."
§P.15 ..... "A Study of Genetic Algorithm Evolution..."
§P.16 ..... "A genetic algorithm for structure-..."
§P.17 ..... "Recording Evolution Supervised by a Genetic Algorithm..."
§P.18 ..... Original results - to be published

Annexes.....226

Introducere

Teza “Algoritmi genetici şi aplicaţii ale acestora” reprezintă un demers de cercetare fundamentală modernă, situat la interdisciplinaritatea genetică – informatică – chimie, având printre principalele obiective declarate “Simularea evoluției cu algoritmi genetici în probleme de optimizare a relațiilor structură-activitate”. Demersul ştiinţific urmăreşte proiectarea, implementarea şi analiza statistică a influenței diferitelor metode de selecție şi de supraviețuire a indivizilor în cadrul unei populații asupra evoluției unui algoritm genetic utilizat pentru obținerea de relații structură-activitate în serii de compuşi chimici biologic activi.

Modelele informatice elaborate prin intermediul algoritmilor genetici emulează modelele biologice evoluționiste, asigurând rezolvarea unor probleme concrete de optimizare sau căutare în experiențele de genetică şi ameliorare a plantelor. Prin intermediul elementelor individuale, reprezentate sub forma şirurilor binare, precum şi a operatorilor de natură biologică definiți asupra populației şi a modelului molecular, algoritmii genetici manipulează cele mai promițătoare şiruri, evaluate conform unei funcții obiectiv, căutând soluții mai bune, tinzând în esență spre cea “optimă”, dar acceptând în final una apropiată de optim.

Prezenta lucrare demonstrează că algoritmii genetici, ca tehnici adaptive de căutare euristică, bazate pe principiile geneticii şi selecției naturale, pot fi eficient utilizați în simularea procesului biologic al evoluției şi în cel de ameliorare a plantelor. În acest sens, se propun soluții proprii biochimiei prin care să se abordeze problematicile dificile, cu o complexitate exponențială, de optimizare a relațiilor dintre structura compuşilor chimici şi activitatea lor biologică.

Metoda experimentală concepută şi aplicată a fost proprie bioinformaticii, urmărind elaborarea premizelor necesare pentru construcția şi aplicarea unui algoritm genetic care să rezolve problemele de optimizare. Ulterior realizării acestui deziderat, algoritmul genetic a fost aplicat pe un set de date experimentale, folosind un eşantion de genotipuri asupra căruia s-au efectuat observații în vederea analizării modelului evolutiv.

Designul experimental a fost construit cu scopul de a asigura transferul de la probleme de optimizare, la probleme de simulare, în mod concret la simularea evoluției conform unor strategii diferite, de selecție şi de supraviețuire. În acest sens, a fost creată o contingență de 3×3 strategii distincte pentru selecție şi supraviețuire, urmărindu-se evoluția în mod repetat, de 46 de ori în fiecare strategie în parte. Din acest moment, conținutul demersului a fost orientat către domeniul biostatisticii, cuprinzând inferențe statistice în observabile calitative şi cantitative ale procesului de evoluție controlată de diferitele strategii de evoluție.

Rezultatele, respectiv informațiile şi cunoştințele obținute au în mare măsură un caracter

fundamental, oferind atât răspunsuri, cât şi suportul teoretic şi practic pentru noi abordări ştiinţifice a unor problematici precum: *Care este legea de distribuţie a obiectivului evoluţiei? Care este legea de distribuţie a momentelor evoluţiei? Care este legea de distribuţie a numărului de evoluţii? Cum este influenţată variabilitatea genotipică de alegerea unei strategii de evoluţie?* etc.

De asemenea, cercetările au permis obţinerea unor rezultate cu un caracter pronunţat aplicativ, cum sunt: implementarea algoritmului genetic într-un program evolutiv capabil să rezolve problemele dificile de optimizare; implementarea rutinelor automate pentru optimizarea geometriei moleculare; implementarea programelor pentru calculul statisticii Anderson-Darling de agrement între model şi observaţie şi a procedurii Grubbs de identificare şi eliminare a observaţiilor în eroare faţă de model.

Teza oferă şi soluţii de transfer tehnologic, cuprinzând răspunsuri la o serie de probleme de optimizare în domeniul agricol, respectiv horticol, în care evoluţia către un obiectiv de ameliorare sau de planificare este influenţată de o serie de parametrii specifici materialului genetic şi/sau ai arealului, şi în care obiectivul fixat este influenţat de strategia aleasă; în acest cadru experimentul de simulare a evoluţiei, teza oferă soluţii evidenţiate statistic cu privire la influenţa strategiilor de selecţie şi supravieţuire.

Structura lucrării este în concordanţă cu cerinţele actuale, specifice tezelor de doctorat.

Stadiul actual al cunoaşterii cuprinde analiza problemelor de optimizare a relaţiilor structură-activitate şi analiza problemelor de simulare a evoluţiei cu algoritmi genetici, inventariindu-se o literatură de specialitate adecvată.

Cadrul cercetării, scopul şi obiectivele au fost încadrate într-un capitol separat, intitulat “Simularea evoluţiei cu algoritmi genetici în probleme de optimizare a relaţiilor structură-activitate”.

În secţiunea de material şi metodă au fost definite problema de optimizare şi problema genetică de evoluţie, precum şi spaţiul integrat în care acestea interacţionează în cadrul experimentului de simulare a evoluţiei.

În secţiunea de rezultate şi discuţii, un prim capitol a fost dedicat detaliilor referitoare la strategiile de selecţie şi de supravieţuire care constituie factorul ce influenţează observabila - procesul de evoluţie, precum şi verificărilor necesare pentru asigurarea certitudinii şi validităţii datelor colectate. Următorul capitol din rezultate şi discuţii cuprinde analiza variabilităţii şi diversităţii cu privire la genotipurile reprezentate în populaţie şi eşantion. Analiza calitativă axată pe observabile cantitative pe parcursul evoluţiei a fost prezentată pe parcursul unui capitol separat, iar analiza obiectivului evoluţiei a fost prezentată pe parcursul a două capitole: folosind un eşantion întâmplător de generaţii şi folosind un studiu sistematic pentru fiecare generaţie. Observaţiile asupra altor variabile ce caracterizează evoluţia (momentele relative ale evoluţiei şi numărul de evoluţii) încheie secţiunea de rezultate şi discuţii într-un capitol distinct.

Investigațiile ştiințifice şi cercetările efectuate pe parcursul doctoranturii, precum şi discutarea, argumentarea şi interpretarea rezultatelor obținute, sunt susținute în teză de un număr de 310 referințe din literatura de specialitate, redate ca atare într-un capitol separat.

Teza cuprinde un număr de 69 de figuri şi 70 de tabele distribuite pe capitole astfel: (14 figuri, 6 tabele) - capitolul 2; (0, 0) - capitolul 3; (0, 1) - capitolul 4; (4, 4) - capitolul 5; (0, 5) - capitolul 6; (1, 5) - capitolul 7; (6, 14) - capitolul 8; (13, 7) - capitolul 9; (0, 5) - capitolul 10; (26, 8) - capitolul 11; (5, 15) - capitolul 12.

Ultimul capitol al tezei este reprezentat de concluziile şi recomandările desprinse din cercetările efectuate.

Lucrările reprezentative publicate în cadrul doctoratului, care reprezintă o valorificare ilustrativă a studiilor şi cercetărilor proprii realizate, precum şi confirmarea şi validarea acestora printr-un proces riguros din punct de vedere ştiințific de "peer review" (în reviste cotate ISI, cu IF, respectiv BDI) sunt prezentate sintetic în secțiunea de anexe. Dintre lucrările publicate, unele sunt de tip "review" (P.1 & P.2), altele originale (restul), incluzând rezultatele obținute pe baza cercetărilor efectuate în cadrul tezei (P.8, P.9, P.11, P.12, P.13, P.17) sau derivate din cercetările cuprinse în teză. Este de menționat faptul că o parte din rezultatele cercetării urmează a fi valorificate în continuare (P.18).

O serie de anexe (totalizând 179 de pagini) sunt disponibile în format electronic online şi au rolul de a furnizare repere, cunoştințe şi chiar alte rezultate proprii ale cercetărilor, originale, suplimentare informațiilor cuprinse în teză. Acestea sunt reprezentate de argumentații, demonstrații, analize de detaliu etc., uşor de accesat prin link-uri, care susțin ipotezele formulate pe parcursul tezei şi permit completarea unei imagini de ansamblu, clară şi coerentă, a problematicilor abordate.

1. Probleme de optimizare a relațiilor structură-activitate

Scurt istoric: Relații între structura şi activitatea seriilor de compuşi chimici

Abordarea matematică a SAR pentru BAC (începută în secolul nouăsprezece) s-a concretizat prin apariția conceptului de relații cantitative structură-activitate (QSAR = Quantitative Structure-Activity Relationships), metodă matematică care permite identificarea legăturii cantitative dintre structura chimică şi activitatea biologică a compuşilor (BAC) investigați - (Hammett, 1935). Observații SAR au fost publicate în literatura de specialitate încă din 1868, când Crum-Brown & Fraser au stipulat ideea că activitatea compuşilor este o funcție a structurii şi compoziției chimice (Crum-Brown & Fraser, 1868). Zece ani după introducerea termenului QSAR (Hansch & Leo, 1979), această paradigmă şi-a găsit aplicabilitatea practică în agrochimie, chimia farmaceutică, toxicologie şi ale domenii conexe chimiei.

Multe metode: CoMFA - Comparative Molecular Field Analysis (şi varianta sa CoMSIA, MSIA - Molecular Similarity Indices Analysis) - (Cunningham & Day, 2005), WHIM - Weighted Holistic Invariant Molecular (şi varianta sa MS-WHIM - Molecular Surface WHIM) - (Todeschini & alții, 1994), MTD (Simon & Bohl, 1992) - Minimal Topological Distance (şi varianta sa MSD, S - Steric) - (Plauchithiu & alții, 1978), MDF - Molecular Descriptor Family (Jäntschi, 2004), FPIF - Fragmental Property Index Family (Jäntschi, 1999) şi mulți descriptori moleculari au fost introduşi şi şi-au dovedit calitățile de predicție şi estimare (Jäntschi & Bolboacă, 2007-Results) şi (Bolboacă & Jäntschi, 2007-HowGood). Alte apariții de actualitate, cum ar fi S(Q)SAR (Putz & Lacrămă, 2007) provoacă din nou abordările clasice.

Rezultatele investigațiilor din domeniul geneticii celulare, a mecanismelor dinamicii celulare, a mutațiilor genetice, sub aspectul secvențializării şi codificării informatizate a topologiei macromoleculare a determinat înregistrarea şi depozitarea în baze de date a geometriei moleculare a unui număr foarte mare de compuşi chimici biologic activi [1]. Integrarea acestor informații cu metode combinatoriale adecvate poate determina identificarea unor compuşi activi cu potență ridicată.

Metodele statistice de evaluare şi validare internă şi externă (Eriksson & alții, 2003), analiza corelațiilor corelate (Steiger, 1980) şi analiza componentelor principale (Hemmateenejad, 2006)

---

[1] Baze de date online conținând structuri 3D de compuşi chimici:
- ÷ Cambridge Structural Database (CSD). http://www.ccdc.cam.ac.uk/products/csd/
- ÷ Protein Data Bank (PDB). http://www.pdb.org/
- ÷ Visual Molecular Dynamics (VMD). http:// www.ks.uiuc.edu/Research/vmd/
- ÷ Molecular Modeling DataBase (MMDB). http://130.14.29.110/Structure/MMDB/mmdb.shtml
- ÷ PubChem Compounds (PCC). www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pccompound
- ÷ PubChem Substance (PCS). www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pcsubstance

sunt metode care au rol în identificarea modelelor de relație cantitativă structură-proprietate care să permită selecția compușilor cu cea mai bună activitate biologică de interes.

Ce se pretează la relații structură-activitate: rolul relației structură-activitate

Procesul de dezvoltare a noi medicamente implică identificarea țintei, screening-ul, generare, optimizare, studii preclinice şi clinice (faza 1, 2, 3) şi în final înregistrarea (Bleicher & alții, 2003). Dezvoltarea tehnologiei calculatoarelor şi a informației a determinat modificarea conceptului de cercetare ducând la apariția a tot mai multor metode de analiză şi experimentare in silico (Plant, 2004; Baranczewski & alții, 2006) utilizate în determinarea şi predicția metabolismului medicamentelor (Schuster & alții, 2006). Aceste metode au o serie de avantaje, dintre care imediate sunt: permit determinarea profilului metabolic în etapele inițiale ale descoperirii medicamentului; experimentele se realizează într-un timp mai scurt şi sunt mai puțin costisitoare (Baranczewski & alții, 2006).

Putz & Lacrămă (Putz & Lacrămă, 2007) clasifică metodele (Q)SAR în patru clase: clasice (Hansch-Fujita, Free-Wilson, topologice): (Jäntschi & Bolboacă, 2007-Results); tri-dimensionale (CoMFA: comparative molecular field analysis, MSD-MTD: mininal steric difference-minimal topological difference, topografic) - (Cunningham & Day, 2005), (Duda-Seiman & alții, 2006), decizional şi rețele neuronale (Micheli & alții, 2007), Fuzzy Adaptive Least Squares (Kumar şi alții, 2007), Support Vector Machine (Li & alții, 2007), ortogonale - analiza componentelor principale sau PCA (Ursu & alții, 2006), descriptori ortogonali - (Randic & Pompe, 2005), şi în final spectral SAR (Putz & Lacrămă, 2007). Este necesar însă să subliniem aici că o categorie aparte este creată de Molecular Descriptor Family (MDF) - (Jäntschi, 2005; Jäntschi & Bolboacă, 2007-Results; Jäntschi & Bolboacă, 2007-Modeling) şi Fragmental Property Index Family (FPIF) - (Jäntschi & alții, 2000; Opriş & Diudea, 2001), care integrează deopotrivă informația topologică cât şi cea topografică. Metodele de analiză a relației structură-activitate şi-au găsit utilitatea în dezvoltarea de noi medicamente - (Narasimhan & alții, 2006), şi în analiza şi caracterizarea activităților compușilor (Saliner & alții, 2006).

Există o serie de reglementări cu privire la caracterizarea modelelor QSAR, cum este regulamentul Parlamentului şi Consiliului European nr. 1272/2008 din 16 dec. 2008. Cinci tipuri de informații trebuie să însoțească modelele QSAR: definirea scopului; algoritm clar; domeniul de aplicabilitate; metode adecvate folosite în analiza validității şi a capacităților predictive; interpretarea modelului. Metodele statistice recomandate a fi utilizate în analiza modelelor QSAR sunt: MLR (multiple linear regression - regresie liniară multiplă), PLS (partial least square regression - regresie parțială a celor mai mici pătrate), MS (classification models - modele de clasificare).

Ce sunt relațiile structură-activitate (SAR): MLR pe BAC

Activitatea biologică sau bioactivitatea este termenul comun pentru efectul benefic sau advers al unui compus (sau amestec de compuşi) chimici asupra materiei vii.

Manifestarea şi cuantificarea calitativă şi/sau cantitativă a activității biologice a unui anumit compus chimic este un proces extrem de complex prin natura foarte variată a efectelor pe care un compus chimic le poate avea asupra diferitelor organisme vii. Procedura de determinare a activității acestuia asupra organismului este standardizată (ex. Schema ADMET, Tabelul 1 - Anexa 2-1). Activitatea biologică este evaluată prin intermediul unor proceduri specifice, supuse standardizării (Tabelul 2 - Anexa 2-1 redă două astfel de activități biologice). O serie de procese biologice sunt referite distinct atunci când se exprimă activitatea biologică a unui compus chimic (Tabelul 3 - Anexa 2-1 redă definiții specifice mediului acvatic). Standardizarea care se referă la determinarea toxicității acvatice acute (UNE-CE-4, 2009) stabileşte că aceasta trebuie determinată folosind una din manierele prezentate în Tabelul 4 - Anexa 2-1.

Importanța compuşilor biologic activi (BAC) se justifică prin utilizarea acestora în domeniul agricol şi horticol. Tabelul 5 - Anexa 2-1 redă o clasificare a regulatorilor de creştere, în acord cu Societatea Americană a Ştiințelor Horticole (*American Society for Horticultural Science; website: http://www.ashs.org/*).

Serii de compuşi înrudiți atât în ceea ce priveşte structura cât şi proprietățile/activitățile acestora sunt prezenți în mediul care ne înconjoară atât în materia vie cât şi în materia nevie. Lantanidele (şi actinidele) sistemului periodic sunt un astfel de exemplu binecunoscut, ce are ca consecință utilizarea unor metode extrem de elaborate pentru separarea şi izolarea acestora din amestecuri. În aceeaşi situație se află şi seriile de aminoacizi esențiali (în număr de 20) al căror rol în procesele vii este binecunoscut. Atât structura cât şi proprietățile celor 20 de aminoacizi diferă (într-o măsură mai mare sau mai mică), unii fiind hidrofobi, alții fiind hidrofili, unii au masă relativ mare, alții relativ mică, având însă în comun prezența ambelor grupări amino ($-NH_2$) şi acid (-COOH) şi având de asemenea în comun prezența lor în proteine, în succesiuni în care locul unuia poate fi foarte uşor luat de altul.

Atunci când se supune observației o serie de compuşi, se porneşte de la ipoteza că aceştia au în comun atât elemente de structură cât şi elemente de proprietate/activitate ce fac ca valorile acestora să fie relativ apropiate.

Pentru ca rezultatele observației să capete consistență în ceea ce priveşte interpretarea statistică, mai este necesar ca să se asume ipoteza de convergență la normalitate asupra valorilor observate în eşantionul seriei de compuşi, spațiul complet al acestora fiind în acest caz un exemplu tipic de populație finită distribuită normal.

În aceste ipoteze, de înrudire a compuşilor atât sub aspect structural cât şi sub aspect al

proprietății/activității măsurate, şi de distribuire normală a valorilor observate, se pot formula şi verifica (cu ajutorul testelor statistice) ipoteze de inferență (dependență) între structură şi activitatea/proprietatea măsurată. Relațiile structură-activitate (SAR) şi respectiv relațiile structură-proprietate (SPR) stabilesc legături funcționale între structura compuşilor chimici şi proprietățile măsurate de natură biologică (SAR) şi fizico-chimică (SPR) ale acestora.

Relațiile cantitative (q) care se stabilesc între structură şi activitate (qSAR) sau respectiv proprietate (qSPR) se exprimă prin intermediul unor ecuații care au un domeniu de aplicabilitate definit cel mai frecvent de seria de compuşi pe care au fost obținute şi de proprietatea sau activitatea supusă observației.

Unul din avantajele majore pe care le conferă obținerea unei relații cantitative structură-activitate/proprietate (qSPAR) este posibilitatea efectuării de predicții cu privire la proprietatea/activitatea unor compuşi congeneri cu cei din seria pe care s-a obținut relația qSPAR.

De cele mai multe ori, o relație qSPAR are asociată, pe lângă domeniul de aplicabilitate şi o semnificație statistică. Cu cât semnificația statistică a unei relații qSPAR este mai mare, cu atât predicția activității/proprietății pentru compuşi congeneri cu cei ai seriei pe care s-a obținut relația qSPAR are şanse de reuşită mai mari.

Aşa cum un singur atom sau o singură regiune a moleculei nu este singura responsabilă pentru manifestarea unei proprietăți moleculare, nici relațiile qSPAR nu se exprimă cu ajutorul unui singur descriptor de structură. În mod frecvent în expresia unei ecuații qSPAR intră 2 sau mai mulți descriptori de structură. Este important însă de ştiut că odată cu creşterea numărului de descriptori într-o ecuație qSPAR scade şi semnificația statistică a parametrilor acesteia. Astfel, pe lângă căutarea celei mai potrivite ecuații sub aspectul maximizării determinării statistice a proprietății/activității măsurate, mai apare şi problema minimizării numărului de variabile ce intră în această ecuație.

Ecuația qSPAR poate fi aproximată de (şi astfel obținută prin) o ecuație de regresie multiplă (MLR) între descriptorii de structură asupra cărora se face ipoteza că nu sunt afectați de erori întâmplătoare (atâta vreme cât structura compusului respectiv este de asemenea o certitudine) şi proprietatea/activitatea măsurată (care este variabila dependentă şi în acelaşi timp supusă erorii experimentale întâmplătoare).

Ecuațiile de regresie liniară multiplă (Anexa 3-4) au avantajul descompunerii liniare a varianței experimentale observate în variațe aditive pentru fiecare variabilă independentă a ecuației (şi asociate acestora). Existența a două sau mai multe variabile independente nu presupune absența corelației liniare între acestea, corelația putând fi pusă pe seama întâmplării, sau, mai exact, pe seama măsurării (sau exprimării unei măsuri) a două sau mai multe mărimi parțial corelate.

Dacă variabilele ce descriu structura nu sunt corelate liniar cu proprietatea măsurată se

preferă linearizarea acestora şi exprimarea unei ecuații de regresie liniare cu aceste variabile ce descriu structura transformate (linearizate) corespunzător.

O ecuație de regresie multiplă descriind o relație qSPAR este semnificativă statistic dacă toți parametrii (coeficienții) acesteia sunt semnificativ statistic diferiți de zero (în caz contrar expresia ecuației se reduce corespunzător prin eliminarea variabilelor ale căror coeficienți nu sunt semnificativ diferiți de zero) şi coeficientul de corelație este de asemenea semnificativ diferit de zero (în caz contrar respingându-se întreaga ecuație de regresie). Aprecierea semnificației parametrilor ecuației de regresie se face prin raportarea valorii parametrului la varianța acestuia, folosind pentru interpretare testul Student t.

Elaborarea şi valorificarea de relații structură-activitate: Studii de elaborare şi valorificare de SAR raportate în literatură

Fluxul de informații de specialitate face numeroase referiri la metodologia de obținere a noilor compuşi biologic activi. În acest sens monografia (Diudea & alții, 2001) este cuprinzătoare.

În ultima perioada de timp, indicii structurali folosiți pentru modelarea de compuşi biologic activi prin studii integrate structura-activitate (SAR - structure activity relationship) sunt tot mai frecvent calculați din parametrii sterici (geometrici) şi/sau electrostatici (sarcini parțiale) - (Rienzo & alții, 2004) în detrimentul parametrilor pur topologici (Wikler & alții, 1998). Sunt preferate calcule semiempirice si cuantice cu programe ca Hondo95, Gaussian94, Gamess, Icon08, Tx90, Polyrate, Unichem/Dgauss, Allinger`s MM3, Mopac93, Mozyme si HyperChem (Chen & MacKerell, 2000).

Analiza de regresie structură/activitate foloseşte metode clasice ca regresia liniară, liniară multiplă, neliniară, sisteme expert sau rețele neuronale pentru baze mari de date (Chen & alții, 2003).

Elaborarea, implementarea şi folosirea modelului molecular pentru obținerea de noi compuşi biologic activi este referită frecvent în fluxul de informații. Iată un rezumat al acestor preocupări.

Ca metodă preliminară de analiză, unii autori aliniază setul de molecule. Mai mult, metoda CoMFA introduce un algoritm in 6 paşi pentru analiza structură-activitate (Cunningham & Day, 2005):

1. construieşte setul şcoala de molecule cu activitate biologică cunoscută şi generează structura 3D a moleculelor, folosind programe ca: Mopac (Tzimopoulos & alții, 2010), Sybyl (Mouchlis & alții, 2010), HyperChem (Koba & alții, 2010), Alchemy 2000 (Sârbu & alții, 2004), MolConn (Chu & alții, 2009);
2. alege o metodă de suprapunere, care poate fi suprapunerea de fragmente alese din molecule (Horwitz & alții, 1993; Waller & alții, 1994) sau suprapunerea de grupuri farmacofore

(McGaughey & Mewshaw, 1998) şi suprapune virtual coordonatele spaţiale;

3. construieşte un grid care înconjoară moleculele suprapuse la pasul (2) într-o formă standard sau modificata (Chuman & alţii, 1998) şi alege un atom de probă pentru interacţiunea cu punctele gridului (Walter & Kellogg, 1996);
4. utilizează o metoda empirică (Kellogg & alţii, 1991), un model specific de suprapunere farmacoforă (Myers & alţii, 1994), energia potenţială clasică (Lennard-Jones, Coulomb), energia legăturilor de hidrogen (Srivastava & alţii, 2008; Kapou & alţii, 2008), câmpul generat de orbitalii moleculari (Waller & Marshall, 1993; Durst, 1998) sau orice alt model definit de utilizator (Rault & alţii, 1992) şi calculează valorile de interacţiune in gridul de la pasul (3) al câmpului de interacţiune ales cu atomul proba de la pasul (3);
5. foloseşte valorile calculate ale interacţiunii de la pasul (4) între punctele gridului şi atomul de probă pentru a face estimarea activităţii cunoscute construind o relaţie structură-activitate (Filimonov & alţii, 2009);
6. foloseşte relaţia structură-activitate obţinută la pasul (5) pentru a face predicţia activităţii pentru molecule cu aceeaşi suprapunere cu setul şcoală de la pasul (1) - (Zhang & alţii, 2009).

Metoda CoMFA este un bun instrument pentru predicţia unui variat tip de activităţi biologice cum sunt citotoxicitatea (Wiese & Pajeva, 1998), inhibiţia (Czaplinski & alţii, 1995), sau proprietăţile de formare (Czaplinski & Grunewald, 1994). Mai mult, metoda serveşte în modelarea compuşilor cu efecte farmaceutice (Doweyko, 1994) şi inhibitorilor HIV (Debnath, 1994).

Una din modalităţile de realizare a căutării substructurilor active biologic este identificarea invarianţilor moleculari. In acest sens metoda WHIM (Weighted Holistic Invariant Molecular) calculează un set de indicatori statistici derivat din proprietăţile sterice şi electrostatice ale moleculelor (Todeschini & alţii, 1994; Todeschini & alţii, 1995; Todeschini & alţii, 1996). O variantă a acestei metode, MS-WHIM (de la Molecular Surface) serveşte în analiza suprafeţei moleculare (Zaliani & Gancia, 1999). MS-WHIM este o colecţie de 36 indicatori statistici derivaţi din proprietăţile sterice şi electrostatice şi este orientată spre parametrizarea suprafeţei moleculare (Bravi & alţii, 1997).

Relaţiile cantitative structură-activitate QSAR (Q - quantitative) referă în general o activitate biologica măsurabilă iar tehnica de investigare ce foloseşte QSAR este o tehnică modernă folosită astăzi în multe domenii prioritare, incluzând farmaceutica, mediul, biotehnologia şi microbiologia. Literatura de specialitate conţine astăzi un număr mare de relaţii structură-activitate care evaluează impactul produselor chimice asupra mediului (Sacan & alţii, 2009; Luis & alţii, 2010).

Modelul matematic folosit pentru obţinerea relaţiilor structură-activitate este cel mai frecvent bazat pe regresia liniară (Sharma & alţii, 2010) şi pe reţelele neuronale artificiale (Rao & alţii, 2009).

Seturile de compuşi supuse unei analize structură-activitate în mod uzual sunt evaluate în prealabil în funcție de scopul propus; astfel, pentru calificarea unui compus cu potență biologică într-un studiu preclinic având ca scop elaborarea unui medicament, Lipinski (Lipinski & alții, 1997&2001) a propus aşa-numita "regulă a lui 5" (nu mai mult de 5 legături de hidrogen donoare; nu mai mult de 10 legături de hidrogen acceptoare; masă moleculară mai mică de 500 g/mol; coeficient de partiție octanol/apă mai mic de 5).

2. Simularea evoluției cu algoritmi genetici

Scurt istoric: Genetica - origine a Algoritmilor Genetici

Genetica, aşa cum o cunoaştem astăzi, este o ştiință complexă derivată din biologie mereu în căutarea de răspunsuri cu privire la origine şi evoluție şi la legătura cauză-efect dintre caracterele observabile şi amprenta nucleotidică, servind la caracterizarea proceselor metabolice care au loc şi la identificarea actorilor cheie ai acestora, producând mereu cunoaştere capabilă să amelioreze starea de sănătate a organismelor vii.

Primele studii sistematice cu privire la trăsăturile şi caracterele plantelor care au dus la formularea ipotezelor de moştenire a eredității se regăsesc în tratatul lui Jean-Baptiste Pierre Antoine de Monet, Chevalier de LAMARCK (1 August 1744 - 18 Decembrie 1829) intitulat "Philosophie zoologique, ou Exposition ..." (Lamarck, 1809) lucrare apărută în 1809 şi reeditată în 1830. În această vastă lucrare (în două volume, cumulând 928 de pagini, Lamarck face prima dată referiri exacte despre moştenirea caracterelor dobândite, fiind în acelaşi timp prima cu adevărat coerentă teorie a evoluției (p. 58-59 vol. II, ed. 1830 în Anexa 1-1). Aşa cum se observă şi din textul original Lamarck face pentru prima dată asocierea între trăsăturile plantelor şi acele părți din plante care sunt responsabile cu reproducerea, pe care astăzi le cunoaştem drept purtătoare a informației genetice (RNA). În acest tratat Lamarck aduce argumente observaționale ale eredității, în el regăsindu-se şi clasificări ale speciilor pe care astăzi le cunoaştem sub numele de arbori filogenetici, Lamarck rămânând astfel în istorie drept proprietarul teoriei eredității caracterelor dobândite, numită ***soft inheritance*** (în traducere: 'moştenire uşoară').

Cincizeci de ani mai târziu Charles Robert DARWIN (12 Februarie 1809 - 19 Aprilie 1882) a propus şi demonstrat evidența ştiințifică că toate speciile vii au evoluat de-a cursul timpului din aceeaşi organisme vii printr-un proces de selecție naturală. Lucrarea intitulată "The origin of species by means of natural selection or the preservation of favoured races in the struggle for life" (Darwin, 1859) publicată în primă ediție în 1859 (p. 460 & 475, ed. 1859 în Anexa 1-2) şi reeditată apoi în 1860, 1861, 1866, 1869 şi 1872 (Darwin, 1872) aduce pentru prima dată argumente pentru teoria evoluției speciilor, Darwin rămânând astfel în istorie drept proprietarul acestei teorii, oferind explicația unificatoare pentru diversitatea vieții, şi oferind drept principiu al evoluției ***survival of the fittest*** (în traducere: 'supraviețuirea celui mai tare').

Câțiva ani mai târziu, Gregor Johann MENDEL (20 Iulie 1822 - 6 Ianuarie 1884) arată că moştenirea urmează legi specifice, legi care acum îi poartă numele (Mendel, 1866). Lucrarea (p. 1-32 ed. 1866 în Anexa 1-3) a cărei valoare a fost redescoperită abia 35 de ani mai târziu, cuprinde rezultatele obținute în urma experiențelor de hibridare desfăşurate în perioada dintre 1856 şi 1863 pe în jur de 28000 de plante de mazăre.

Friedrich Leopold August WEISMANN (17 Ianuarie 1834 - 5 Noiembrie 1914) printr-o serie de prelegeri susținute începând cu 1883 şi materializate prin publicarea cărții (Weismann, 1893) formulează teoria ***hard inheritance*** (în traducere: 'moştenire dură'). În formulare modernă, teoria lui Weismann postulează că tranformarea/moştenirea la un organism (multicelular) se realizează prin intermediul celulelor germinale (p. 453, 469, 473, ed. 1893 în Anexa 1-4).

Aproape cincizeci de ani mai târziu (în 1915) de la apariția lucrării lui Mendel, Thomas Hunt MORGAN (25 Septembrie 1866 - 4 Decembrie 1945) integrează moştenirea Mendeliană în ***teoria cromozomială a moştenirii*** (Morgan & alții, 1915), acesta astfel nucleul geneticii clasice (p. i, ed. 1915 în Anexa 1-5).

La confirmarea rezultatelor obținute de Mendel şi Morgan s-a remarcat Sir Ronald Aylmer FISHER (17 Februarie 1890 - 29 Julie 1962), care pe lângă extraordinarele sale contribuții la fundarea statisticii moderne, a avut cele mai notabile contribuții în demonstrarea legilor lui Mendel ale eredității şi teoriei cromozomiale a moştenirii - (Fisher, 1918), (Fisher, 1922) şi (Fisher, 1954), fiind reprezentative pentru subiectul abordat, şi fragmente din textul în limba engleză este redat în Anexa 1-6.

Ce se pretează la algoritmi genetici: primele studii de simulare a evoluției

Baza de dezvoltare a algoritmilor genetici o constituie genetica. Sintetizând, Lamarck (Lamarck, 1809) după eforturi extraordinare de clasificare a organismelor vii, remarcă că atât în ceea ce privește animalele (studiind mai cu seamă animalele nevertebrate) cât şi plantele clasificarea pe specii şi varietăți este mai mult sau mai puțin arbitrară, aducând numeroase argumente în acest sens. Făcând apel la taxonomia modernă, astăzi sunt recunoscute 5 domenii, care mai apoi se subîmpart pe divizii, clase, ordine, familii, genuri şi specii. Însă studiile lui Lamarck despre specii rămân de o deosebită importanță, în ciuda faptului că explicația pe care a găsit-o Lamarck asupra observațiilor sale bazată pe superstițiile vremii şi anume că golurile constatate în schema de clasificare a speciilor ar fi în fapt datorată dispariției acestora a fost contrazisă ulterior de studiile lui Darwin (Darwin, 1859). Chiar şi Darwin a fost influențat de concluziile la care a ajuns Lamarck, aşa cum nota Fisher în 1954 (Fisher, 1954). Realizările majore ale epocii de început a geneticii sunt completate de studiile de încrucişare de varietăți ale lui Mendel (Mendel, 1866) stabileşte legile care astăzi îi poartă numele, formularea teoriei 'moştenirii dure' care vine să completeze elementele lipsă lăsate de predecesori (Weismann, 1893), şi în final studiile pe musculița de oțet ale lui Morgan (Morgan & alții, 1915) în urma cărora acesta elaborează teoria cromozomială a eredității, care este şi astăzi la baza geneticii moderne. Nu în ultimul rând se menționează studiile de o deosebită valoare ştiințifică ale lui Fisher (Fisher, 1918; Fisher, 1922), care a adus argumente ştiințifice de necontestat

(chiar dacă la vremea când acestea au fost scrise de către Fisher au fost din plin contestate, aşa cum se poate desprinde şi din fragmentele alese pentru exemplificare) cu privire la originea speciilor şi teoria evoluționistă. Rezumând, moştenirea 'dură' (Weismann, 1893) şi 'uşoară' (Lamarck, 1809), selecția şi supraviețuirea (Darwin, 1859), genele şi recombinarea genelor (Morgan & alții, 1915) şi caracterelor (Mendel, 1866) îndelung dezbătute şi disputate de-a lungul secolului 19 (Fisher, 1954) constituie toate piese dintr-un puzzle care construieşte astăzi genetica modernă (Ayala & alții, 1994), şi reprezintă sursele de inspirație ale algoritmilor genetici.

Primele simulări ale evoluției se regăsesc în studiile lui Nils Aall BARRICELLI (Barricelli, 1954). Puțin timp mai târziu, Alex FRASER (1923-2002) a publicat o serie de lucrări despre simularea selecției artificiale a organismelor cu locuşi multipli (*locus: poziția pe care o anumită genă o ocupă pe un cromozom*) ce controlează o trăsătură măsurabilă. Simulările lui Fraser (Fraser, 1957-1970) includ toate elementele esențiale ale algoritmilor genetici moderni.

Instrumentul de dezvoltare a algoritmilor genetici îl constituie informatica. Astfel, uzual, în viața noastă de zi cu zi la fel ca şi în cercetarea ştiințifică noi operăm cu ***probleme***. În informatică şi ramurile derivate ale acesteia (cum e cazul bio-informaticii şi chemo-informaticii) o problemă are o semnificație precisă, foarte apropiată cu cea de ***algoritm***. Un algoritm este în esență o rețetă specificând ce să facem în anumite condiții pentru a obține un anumit obiectiv. Un algoritm necesită două resurse pentru a ***rezolva*** o problemă, şi anume timp (cu sensul de timp de execuție, mărime corelată cu numărul de instrucțiuni elementare) şi spațiu (pentru stocarea datelor de intrare şi a variabilelor).

Nu toate problemele sunt de aceeaşi ***complexitate*** şi acelaşi lucru este valabil şi pentru algoritmii de rezolvare. Astfel, unele probleme au complexitate exponențială, ceea ce înseamnă că cel mai bun algoritm rezolvă problema într-un timp de execuție ce creşte exponențial în funcție de dimensiunea (volumul, mărimea) datelor de intrare. Acest tip de probleme sunt numite ***dificile***, deoarece chiar şi cel mai bun algoritm (care există sau ar putea exista) va fi probabil nepractic cu date de intrare din practică (Falkenauer, 1998). De exemplu, o problemă dificilă este următoarea (în care timpul de explorare al spațiului de căutare a posibilităților este exponențial):

Fiind date un număr de maşini de recoltat, un număr de câmpuri de recoltat, un număr de şoferi (şi dacă dorim un număr de spații de depozitare) să se găsească cea mai bună cale să se organizeze recoltarea la o fermă, pentru fiecare maşină de recoltat implicând un câmp şi un şofer

Dacă o problemă este dificilă, atunci căutarea ***optimului*** frecvent iese în afara timpului disponibil pentru aplicațiile reale. Chiar dacă există această problemă, există totuşi o serie de probleme întâlnite în practică când obținerea optimului nu este necesară (obligatorie). De cele mai multe ori o ***soluție bună*** este suficientă. Într-adevăr, presupunând că problema dificilă este organizarea recoltatului la o fermă, un algoritm permițând ca costul de recoltare să fie redus de la 40000 lei/săptămână la 10000 lei/săptămână este de un real folos pentru fermă, chiar dacă un

algoritm optimal (care găseşte minimul global) ar mai putea încă să îmbunătățească organizarea reducând costul la 8000 lei/săptămână. Mai mult, desigur că algoritmul care permite reducerea costului la 10000 lei/săptămână este preferat celui care reduce costul la 8000 lei/săptămână dacă timpul de execuție al acestuia din urmă este excesiv de mare, de exemplu mai mare decât săptămâna care se organizează.

În concluzie, căutarea pentru algoritmi de aproximare buni este perfect legitimată, chiar dacă aceşti algoritmi nu garantează atingerea optimului global pentru orice instanță cu care sunt hrăniți, dar care produc soluții aproape de soluția optimă.

Ce sunt algoritmii genetici: meta-euristici

Deoarece cele mai multe probleme dificile au fost împrejurul nostru de foarte mulți ani, pentru o varietate de probleme dificile unul sau mai mulți ***euristici*** au fost deja concepuți. Aceştia sunt seturi de reguli gândite pentru a rezolva o problemă anume, uzual bazați pe bunul simț (în ceea ce priveşte soluția aşteptată) prin evitarea erorilor grosolane, dar care nu sunt gândiți pentru a produce totdeauna soluția cu exactitate şi respectiv să fie capabili să producă o soluție pentru orice valori de intrare.

Chiar dacă cei mai mulți euristici sunt foarte mult ad-hoc şi dependenți de problema dată, odată cu dezvoltarea informaticii cercetătorii au reuşit să formuleze trei euristici care sunt foarte generali, şi anume aplicabili la o mare varietate de probleme dificile. Din cauza acestei generalități pe care o posedă, aceştia au căpătat numele de ***meta-euristici***. Toți trei sunt stocastici în natura lor (*a fi stocastic: Implicând sau conținând una sau mai multe variabile aleatoare, implicând şansa sau probabilitatea*), doi dintre aceştia (SA şi GA) fiind bazați pe procese naturale care au loc în jurul nostru din totdeauna. Împreună cu '***călirea simulată***' (în engleză SA - Simulated Annealing; van Laarhoven and Aarts, 1987; Davis, 1987) şi '***căutarea tabu***' (în engleză TS - Tabu Search; Glower, 1977; Glover, 1986; Glover & alții, 1993) sunt şi ***algoritmii genetici*** (în engleză GA - Genetic Algorithm).

Chiar dacă primele studii în care au apărut algoritmii genetici se găsesc în anul 1954, studii de amploare ale acestora au apărut după 1970 (Bosworth & alții, 1972; Holland, 1975) şi algoritmii genetici au fost re-inventați din nou ceva mai târziu (Davis, 1991; Holland, 1992).

Complexitatea algoritmică: Domeniu de aplicabilitate (AD)

Înainte de a defini ceea ce este un algoritm genetic, să definim mai întâi prin ce ***evaluează calitatea*** unul algoritm euristic. Sunt trei criterii care trebuie considerate:

÷ viteza: cât de repede obține soluția;

÷ precizia: cât de departe de află acea soluție de optimul global;

÷ scopul: cât de mare este subsetul datelor de intrare în raport cu setul tuturor valorilor posibile pentru care algoritmul performează în raport cu anterioarele două criterii;

O problemă importantă legată de ***complexitatea algoritmică*** este reprezentată de teorema '***inexistenței mesei pe gratis***' (în engleză NFLT - No Free Lunch Theorem; Wolpert & Macready, 1995&1997; English, 1996), teoremă care utilizând aceste trei criterii de mai sus arată că toți algoritmii sunt strict echivalenți, ceea ce înseamnă că pentru doi algoritmi A şi B, pentru fiecare set de date pentru care A performează mai bine decât B există un set de date pentru care B performează mai bine decât A.

Interpretarea simplă care se dă acestei teoreme în termeni comuni şi anume că oricât ai încerca să-ți faci algoritmii tăi mai isteți, este un efort în van deoarece ei vor performa la fel ca orice alt algoritm, nu este una corectă. Ceea ce teorema într-adevăr spune este că dacă se mediază performanțele tuturor algoritmilor pe toate datele posibile, atunci ei vor performa la fel. Revenind la termeni comuni, şmecheria este desigur să nu încerci să hrăneşti toți algoritmii pe care îi realizezi pe toate datele cu putință, ci să încerci să îți dedici algoritmul la un ***domeniu de aplicabilitate***, şi aici să iei în considerare şi să valorifici prin implementare în algoritm orice structură specială este posibil să existe în datele cu care intenționezi să hrăneşti algoritmul.

De aici rezultă că scopul algoritmului care performează bine trebuie să fie restrâns la setul de date care prezintă structurile speciale identificate.

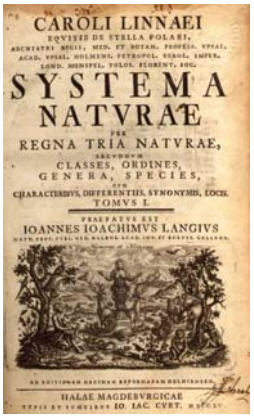
Următoarele categorii de probleme pot fi subiect de rezolvare pentru algoritmii genetici:

÷ ***Probleme de decizie***

- O problemă de decizie este definită pentr-o întrebare cu răspuns de tipul da/nu pe un set (infinit) de date de intrare; din acest motiv problemele de decizie sunt echivalente cu obținerea setului de date de intrare pentru care răspunsul problemei este da. Problemele de decizie sunt legată de problemele de optimizare atâta timp cât problema este obținerea celui mai bun răspuns la problemă.

÷ ***Probleme de clasificare***

- O problemă de clasificare pentru obiecte dintr-un domeniu dat este în separarea acestor obiecte în clase mai mici, şi producerea de criterii de determinare dacă un obiect anume dintr-un domeniu este într-o anume clasă sau nu. Una dintre cele mai faimoase probleme de clasificare este problema formulată de Carl LINNAEUS (23 Mai 23 1707 - 10 Januarie 10 1778) a clasificării viețuitoarelor după clase, ordine, genuri şi specii (Linnaei, 1735).


CAROLI LINNAEI
SYSTEMA
NATVRAE
REGNA TRIA NATVRAE,
CLASSES, ORDINES,
GENERA, SPECIES,
TOMVS I.
IOANNES IOACHIMVS LANGIVS
HALAE MAGDEBVRGICAE

÷ ***Probleme de optimizare***

- O problemă de optimizare este o problemă de găsire a celei mai bune soluții dintre toate soluțiile posibile. În mod formal, o problemă de optimizare este un cvadruplu (I,f,m,g) unde:
  - I - set de instanțe;
  - f(·) - setul soluțiilor posibile definite pe I;
  - m(·,·) - măsura definită pe produsul soluțiilor posibile şi instanțelor
  - g - min. sau max. - funcția obiectiv
  - scopul este găsirea optimului lui x: $m(x,f(x))=g\{m(y,f(y), y \in I\}$
- Pentru fiecare problemă de optimizare există o problemă de decizie care este asociată şi a cărei întrebare este dacă există o soluție posibilă pentru o anumită măsură m_0.

÷ ***Probleme de simulare***

- Simularea este imitarea unui fapt real, unei stări de fapt, sau a unui proces. Actul simulării a ceva în general implică reprezentarea unor anumite caracteristici sau comportamente cheie ale unui sistemul fizic sau abstract. Simularea este folosită în multe contexte incluzând modelarea sistemelor naturale, pentru a pătrunde funcționalitatea acestora. Elementele cheie în simulare includ achiziția unei surse valide de informație despre subiectul de studiu, selecția caracteristicilor şi comportamentelor cheie, utilizarea de aproximații şi presupuneri simplificatoare în cadrul simulării, şi evaluarea fidelității şi validității rezultatelor simulării.

Construcția algoritmilor genetici: Ce sunt algoritmii genetici

Algoritmii genetici sunt algoritmi de căutare euristici adaptivi bazați pe ideile teoriei evoluției şi anume aduce conceptele de selecție naturală şi genetică în arena simulării matematice cu ajutorul calculatorului. Mimica proceselor observate în evoluția naturală a materiei organice în general serveşte drept instrument algoritmilor genetici în scopul de a rezolva probleme de decizie, clasificare, optimizare şi simulare. Elementele cheie la care se face apel în algoritmii genetici sunt:

÷ Modelul genetic (dualismul genotip - fenotip) aşa cum a fost el formulat şi argumentat încă de la primii paşi ai geneticii (Morgan & alții, 1915; Fisher, 1918);

÷ Încrucişarea (dualismul caractere - gene) aşa cum a fost ea observată încă de la precursorii geneticii moderne (Lamarck, 1809; Mendel, 1866; Weismann, 1893);

÷ Mutația, aşa cum a fost ea observată încă de la precursorii geneticii moderne şi până în zilele noastre:

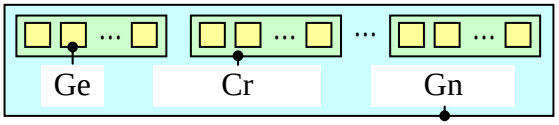
- întâmplătoare (Veies, 1902);
- deliberată prin expunerea la anumite condiții (Patterson, 1928; Muller, 1928; Auerbach & alții, 1947);
- sub presiunea factorilor de mediu: (Cains & alții, 1988);

÷ Selecția naturală sau supraviețuirea celui mai tare (Darwin, 1859).

Algoritmii genetici se materializează sub forma de programe evolutive şi sunt simulări pe calculator în care:

(spațiul de căutare)

÷ Se operează asupra unei populații de reprezentări abstracte (Figura 2-1) numite (după elementele genetice pe baza cărora au fost imaginate) cromozomi sau genotipuri ale unui ***genom***, la rândul său fiecare reprezentare abstractă a unui ***cromozom*** fiind compusă din ***gene***.



Legendă (*Legend*): Ge - Genă (*Gene*); Cr - Cromozom (*Chromosome*); Gn - Genom (*Genome*)

Fig. 2-1. Spațiul de căutare al unui algoritm genetic
The search space of a genetic algorithm

÷ Fiecare ***generație*** este compusă dintr-o populație de şiruri de caractere (sau alte forme de reprezentare abstractă) analog cu cromozomii ADN-ului. Fiecare element al populației reprezintă un punct în spațiul de căutare şi în acelaşi timp o soluție posibilă.

÷ Ceea ce Figura 2-1 reprezintă formal, şi anume spațiul de căutare al unui algoritm genetic, poate avea multe variante de implementare, trei dintre ele fiind următoarele:

- Dacă algoritmul genetic are ca subiect rezolvarea unei probleme dificile formulate în sistemul S (în engleză: S-system formalism, Savageau, 1976) care este un tip de formalism derivat din modelul de proces al reacțiilor stoechiometrice cu pre-echilibru ($\Sigma_i R_i \leftrightarrow \Sigma_j I_j \rightarrow \Sigma_k P_k$, unde R_i reactanți, I_j intermediari, P_k produşi ai unei reacții în care constantele de proces - constantă de viteză şi ordine parțiale de reacție - sunt necunoscute şi se doresc a fi determinate), atunci următoarea este o posibilă implementare:
  - O genă: o constantă (un ordin parțial sau o constantă de viteză de reacție) subiect al găsirii (optimizării);
  - Un cromozom: o posibilă cale de desfăşurare a reacției, având specificate toate ordinele parțiale şi constantele de viteză specificate;
  - Genomul: toate căile de desfăşurarea a reacției prezente într-o iterație a algoritmului genetic;
- Dacă algoritmul genetic are ca subiect rezolvarea unei probleme dificile de aliniament de secvențe genetice (Notredame & Higgins, 1996) de ADN, ARN sau proteine în scopul identificării regiunilor de similaritate care pot fi sursă de relații structurale, funcționale sau

evolutive între secvențe, atunci următoarea este o implementare posibilă:

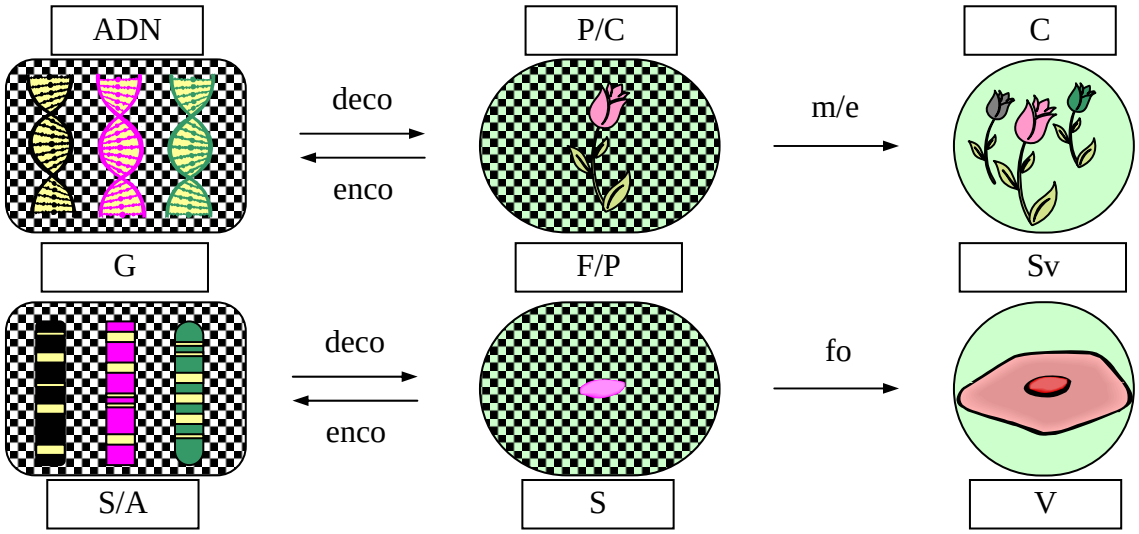
- O genă: două (sau mai multe) poziții corespunzătoare la două (sau mai multe) sub-secvențe aliniate (sau mai exact pseudo-aliniate) şi lungimea aliniamentului acestora;
- Un cromozom: o posibilitate de aliniament pentru cele două (sau mai multe) secvențe;
- Genomul: toate posibilitățile de aliniament de secvențe stocate într-o iterație a algoritmului genetic;

o Dacă algoritmul genetic are ca scop o problemă de setare în managementul efectuat în scopul maximizării randamentului de producție în câmp (Liu & alții, 2001), o problemă dificilă de setare a parametrilor controlabili (sau alteori predictibili) pentru obținerea unei productivități maxime, atunci următoarea este o implementare posibilă:

- O genă: una dintre următoarele: pH-ul solului, fertilizatori în termeni de cantitate de N, P şi K, cantitatea de materie organică din sol, gradul de creştere termică zilnică (o mărime medie între temperatura minimă şi maximă a zilei), potențialul genetic (ce poate fi exprimat în termeni de randament care s-ar obține dacă vremea, solul şi fertilitatea sunt toate optime), cantitățile de precipitații pe perioada de maximă vegetație pe lunile Mai, Iunie, Iulie şi August, densitatea de plantare şi factorul de rotație;
- Un cromozom: o stare de fapt care poate apare în practică în câmp;
- Genomul: toate stările de fapt stocate într-o iterație a algoritmului genetic;

(selecția şi supraviețuirea)

÷ Un scor sau ***şansă de supraviețuire*** a fiecărei soluții este calculată (Figura 2-2) pentru fiecare genotip cu ajutorul unei funcții, numită şi ***funcție obiectiv***.



Legendă (*Legend*): AND - Acid dezoxiribonucleic *(deoxyribonucleic acid)*; P/C - Plantă (*Crop*); C - Cultivar (*Cultivar*); deco - decodare (*decoding*); enco - codare (*encoding*); m/e - mediu (*environment*); G - genotip (*genotype*); F/P - fenotip (*phenotype*); Sv - supraviețuire (*Survival*); S/A - Şir (*Array*); S - Soluție (*Solution*); fo - funcție obiectiv (*objective function*); V - valoare (*value)*

Fig. 2-2. Selecția: genotip, fenotip şi supraviețuire
Selection: genotype, phenotype and survival

- Valoarea funcției obiectiv este asociată cu abilitatea individului să supraviețuiască şi defineşte astfel ***fenotipul*** asociat genotipului. Dacă fiecare genotip reprezintă un punct în spațiul de căutare şi în acelaşi timp o soluție posibilă, prin intermediul selecției genotipul este concretizat în fenotip (operație care iterează reprezentarea soluțiilor posibile în spațiul soluțiilor şi evaluează valoarea acestora). Principiul selecției naturale se exprimă astfel:
  - Indivizii (fenotipurile) din populație concurează pentru supraviețuire (selecție).
  - Genele indivizilor selectați se propagă de la o generație la alta (datorită selecției);
  - Fiecare generație devine mai potrivită mediului în care se află (prin penalizarea indivizilor care eşuează a supraviețui).
- Scorul este asociat fiecărui fenotip (soluție) reprezentând abilitatea acestuia să concureze pentru resurse în mediu, pentru ***supraviețuire***. Scopul algoritmului genetic este ca să aplice încrucişarea şi mutația selectivă a fenotipurilor (prin intermediul decodării lor în genotipurile din care provin) pentru a produce descendenți mai buni decât părinții lor.
- Evoluția prin intermediul algoritmului genetic se realizează prin menținerea unui eşantion din populație de un număr dat (sau uneori variabil) de genotipuri candidate la selecție, care se poate face aplicând acelaşi operator. Astfel, selecția şi supraviețuirea sunt două concepte asociate. Selecție se face pentru operațiile de încrucişare şi mutație asupra genotipurilor din eşantionul de material genetic, şi selecție se face şi pentru supraviețuirea fenotipurilor în cultivarul de dimensiune limitată.
- Pe parcursul evoluției, o parte din indivizii populației sunt înlocuiți de alții. În acest mod se speră că de-a lungul generațiilor soluții mai bune vor ***răsări*** în timp ce cele mai slabe soluții sunt înlăturate. Odată cu trecerea de la o generație la alta eşantionul va conține din ce în ce mai bune soluții decât generația anterioară.
- În Tabelul 2-1 este redată legătura care se stabileşte între scorul (exprimat prin funcția Fitness(·) în tabel) şi regula de selecție în funcție de strategia (aşa cum este ea cunoscută în literatura de specialitate) folosită.

Tabelul 2-1

Selecție şi scor de selecție în algoritmii genetici
Score and selection score in genetic algorithms

Strategie *Strategy*	Expresia funcției de scor *Score function expression*	Selecție *Selection*	Comentarii *Comments*
Proporțional *Proportional*	f_i=Fitness(Cromozom_i)	$p_i=f_i/\Sigma_i f_i$	Șansa de selecție este proporțională cu scorul (utilizând probabilitatea p_i în selecție)
Deterministic *Deterministic*		i | f_i = max. sau min.	Selecția indivizilor este făcută pe baza celui mai tare (sau celui mai slab) individ (elitism)
Turnir *Tournament*		(f_i,f_j) max. sau min.	Perechi de indivizi concurează între ei pentru selecție (din nou cel mai tare sau cel mai slab)
Normalizare *Normalization*	$g_i=(f_i-N_0)(f_{max.}-f_{min.})/(N_1-N_0)$	$p_i=g_i/\Sigma_i g_i$	O scală fixă $[N_0,N_1]$ normalizează scorul fenotipurilor între generații diferite
Ranguri *Ranks*	h_i=Rank(f_i)($f_{max.}-f_{min.}$)/Size	$p_i=h_i/\Sigma_i h_i$	Șansa este proporțională cu rangul scorului unde: Rank(·): rangul; Size: volum genom

÷ Ceea ce Figura 2-2 reprezintă formal, şi anume selecţia şi supravieţuirea fenotipurilor poate avea multe variante de implementare, trei dintre ele fiind următoarele:

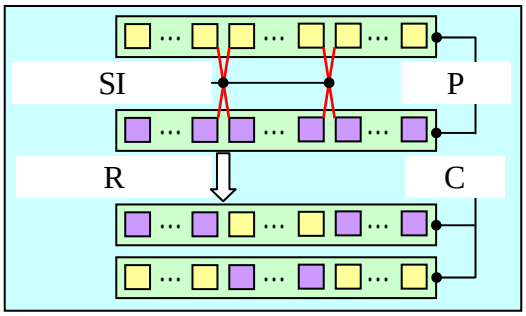
- Dacă algoritmul genetic are ca subiect rezolvarea unei probleme dificile formulate în sistemul S (Savageau, 1976), atunci următoarea este o posibilă implementare:
  - Şirul, corespunzător unui genotip: o listă de valori constante subiect al optimizării şi asociat cu un experiment virtual;
  - Soluţia, corespunzătoare genotipului (şi cromozomului din Figura 2-1): seria de timp a elementelor experimentului virtual (pentru o reacţie chimică prin soluţie se înţeleg seriile de timp ale concentraţiilor reactanţilor, intermediarilor şi produşilor de reacţie pe parcursul desfăşurării reacţiei);
  - Valoarea, corespunzătoare scorului: suma pătrată a diferenţelor dintre valorile observate (ca serie sau serii de timp) şi valorile estimate (de fenotip) ale uneia (sau mai multor) observabile (cum ar fi concentraţie sau concentraţii de intermediari);
- Dacă algoritmul genetic are ca subiect rezolvarea unei probleme dificile de aliniament de secvenţe de aminoacizi (Notredame & Higgins, 1996), atunci următoarea este o posibilă implementare:
  - Şirul, corespunzător unui genotip: o listă de perechi (sau de mai multe) poziţii de sub-secvenţe aliniate urmată de lungimea fiecărei sub-secvenţe;
  - Soluţia, corespunzătoare fenotipului (şi genotipului din Figura 2-1): o serie de valori conţinând poziţii de rupere şi lungimi de translatare necesare pentru a alinia secvenţele;
  - Valoarea, corespunzătoare scorului: o funcţie de scor dând (uzual sub forma unei sume) costul total pentru toate ruperile şi deplasările necesare pentru a alinia secvenţele, utilizând un cost predefinit pentru o rupere şi pentru deplasarea unei unităţi în secvenţă;
- Dacă algoritmul genetic setarea parametrilor necesari pentru obţinerea unei bune producţii în câmp (Liu & alţii, 2001), atunci următoarea este o posibilă implementare:
  - Şirul, corespunzător unui genotip: o listă de valori ce corespund unui experiment virtual şi constituie obiect al optimizării; Valorile din şir pot fi: pH-ul solului, fertilizatori în termeni de cantitate de N, P şi K, cantitatea de materie organică din sol, gradul de creştere termică zilnică (o mărime medie între temperatura minimă şi maximă a zilei), potenţialul genetic (ce poate fi exprimat în termeni de randament care s-ar obţine dacă vremea, solul şi fertilitatea sunt toate optime), cantităţile de precipitaţii pe perioada de maximă vegetaţie pe lunile Mai, Iunie, Iulie şi August, densitatea de plantare şi factorul de rotaţie;
  - Soluţia, corespunzătoare fenotipului (şi genotipului din Figura 2-1): un şir de valori caracterizând soluţia, cuprinzând valori obţinute prin aplicarea de funcţii care să

exprime: calitatea solului, calitatea vremii, managementul de cultivare, potențialul genetic şi efectul unor evenimente întâmplătoare;

- Valoarea, corespunzătoare scorului: suma pătratelor diferențelor între randamente observate (în serii de experimente anterioare) şi estimate (de fenotip) ale randamentelor;

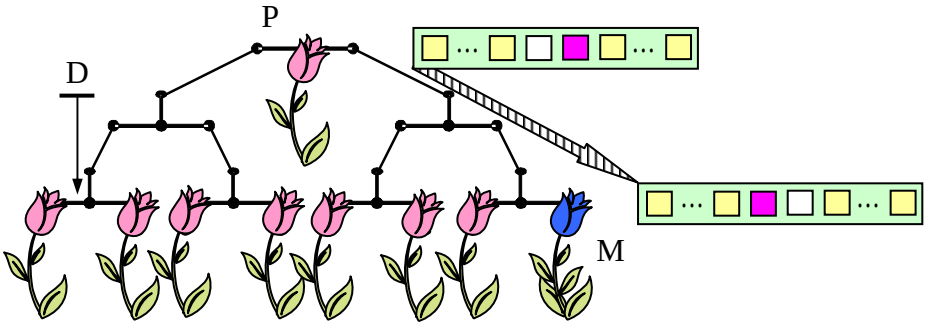
(încrucişarea şi mutația)

÷ Operatorul de încrucişare realizează împerecherea între fenotipuri; fenotipurile (uzual două) sunt ***selectionate*** din populație folosind operatorul de selecție; o porțiune de recombinat de-a lungul şirului de gene ale genotipurilor asociate fenotipurilor este aleasă (întâmplător sau deterministic) şi valorile celor două porțiuni de şiruri sunt schimbate între ele (Figura 2-3), rezultând astfel din această împerechere doi descendenți care sunt direct selectați pentru a face parte din noua generație de populație; încrucişarea este făcută în speranța că dacă se recombină porțiuni de genotipuri de succes, atunci acest proces este probabil să producă descendenți chiar mai buni decât părinții din care provin;



Legendă (*Legend*): SI: Site încrucişare (*crossover site*); P: Părinți (*Parents*); R: Recombinare (*Recombination*); C: Copii (*Childs*)

Fig. 2-3. O încrucişare dublă implicând ruperea şi reunirea cromozomilor părinților
A double crossover involving crossing and joining of parent's chromosomes



Legendă (*Legend*): P: Părinte (*Parent*); D: Descendent (*Descendant*); M: Mutant (*Mutated*)

Fig. 2-4. Mutația
Mutation

÷ Mutația este operatorul care introduce modificări noi (inexistente în populația unei generații); ceea ce este caracteristic în general mutației şi implicit şi operatorului acesteia corespondent în algoritmii genetici este că ea se petrece cu o probabilitate scăzută, fiind deci aplicată cu o probabilitate scăzută (cu probabilitatea de 1/8 în Figura 2-4); operatorul de mutație poate implementa o mutație:

- Întâmplătoare: când o porțiune a unui individ selectat va suferi schimbarea valorilor stocate în genele sale cu alte valori existente în materialul genetic al populației şi are rolul menținerii diversității în populație pentru a preveni populația să prezinte o convergență prematură;
- Deliberată: când expunerea la anumite condiții se transpune în folosirea unei reguli predeterminate de modificare a valorilor genelor;
- Sub presiunea factorilor de mediu, când valorile genelor se schimbă în raport cu scorul fenotipului supus modificării genetice;

(evoluția)

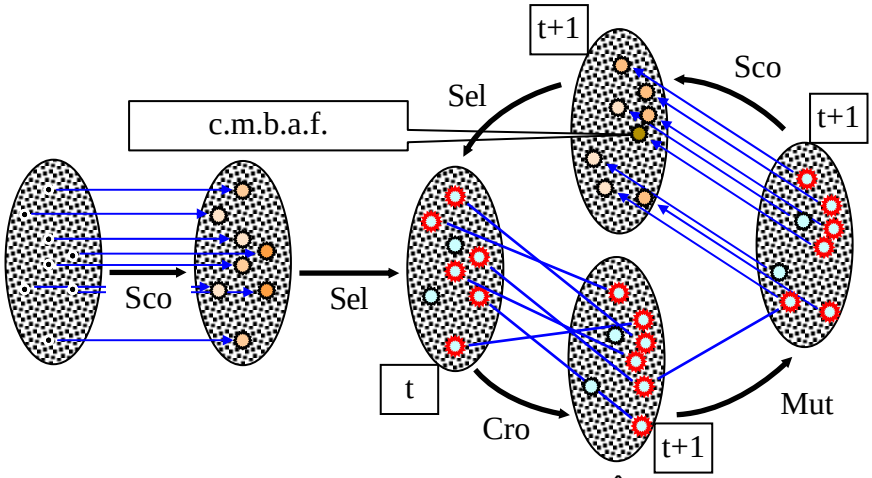
÷ Utilizând doar selecția singură un algoritm nu va reuşi decât să copieze (cloneze) cel mai bun individ (fenotip) al său în întreaga populație;

÷ Utilizând mutația singură un algoritm va reuşi doar să inducă parcurgerea întâmplătoare a spațiului de căutare;

÷ Utilizând încrucişarea şi selecția un algoritm va reuşi să conveargă către o soluție bună dar nu sub-optimală (în apropierea celei optime);

÷ Mutația şi selecția (fără încrucişare) într-un algoritm creează algoritmi paraleli, toleranți la perturbații în căutarea de puncte de maxim local (în terminologia în engleză: hill-climbing);

÷ Utilizarea tuturor operatorilor (mutație, încrucişare si selecție) asigură unui algoritm toate caracteristicile de definire ale unui algoritm genetic (Figura 2-5);



Legendă (*Legend*): Sco - Scor (Score); Sel - Selecție (Selection); Cro - Încrucişare (Crossover); Mut - Mutație (Mutation); c.m.b.f. - Cel mai bine adaptat fenotip (*Best adapted phenotype*)

Fig. 2-5. Schema ilustrativă a modului de lucru al unui algoritm genetic clasic
Illustrative scheme for workflow of a classical genetic algorithm

÷ Într-un algoritm genetic clasic (de genul celui ilustrat în Figura 2-5), pentru a rezolva o problemă, se generează întâmplător sau se inițiază cu valori predefinite o populație de un volum dat de genotipuri (Figura 2-1); cerințele preliminare algoritmului genetic este existența funcției obiectiv cu ajutorul căreia se evaluează scorul unui fenotip în populație; algoritmul genetic

iterează astfel:

- Repetă
  - Pasul_1: Utilizând operatorul de selecție (Figura 2-2) selectează doi cromozomi;
  - Pasul_2: Utilizând o funcție discretă de probabilitate pentru alegerea porțiunii de încrucișat încrucișează cei doi părinți şi creează descendenții acestora (Figura 2-3);
  - Pasul_3: Cu o mică probabilitate şi utilizând o funcție discretă de probabilitate pentru alegerea porțiunii de mutat efectuează mutația unui genotip (Figura 2-4), eventual un descendent al încrucișării din pasul anterior;
  - Pasul_4: Inițializează o nouă populație cu noile fenotipuri (de la paşii 2 şi 3 anteriori);
  - Pasul_5: Completează utilizând operatorul de selecție aplicat populației de părinți noua populație cu fenotipuri (până cel puțin la refacerea numărului inițial de membrii);
  - Pasul_6: Refă valorile funcției de scor ale noii populații în conformitate cu noua compoziție a acesteia;
- Până când cel mai bun fenotip al populației satisface o condiție impusă (condiție care reprezintă condiția de sfârşit a algoritmului).

O problemă dificilă de proces chimic şi formalismul algoritmilor genetici

Procesele naturale la fel ca şi procesele controlate evoluează după un mecanism. Dacă se referă la chimie, atunci mecanismul referă explicarea căii de reacție, în timp ce în biologie referă cum o caracteristică este formată (obținută). Pornind de la un formalism numeric al cineticii reacțiilor biochimice (Jäntschi, 2003), reacția rapidă dintre cupru (Cu^{2+}) şi tiosulfat ($S_2O_3^{2-}$) a fost investigată (Stoenoiu & alții, 2008). Formularea problemei dificile este:

÷ Doi reactanți (A şi B), prin intermediul unui intermediar (X) formează produsul de reacție (P);

÷ Concentrația intermediarului (X) poate fi corelată cu absorbția intensității luminoase la o anumită lungime de undă (430 mn) utilizând relația Lambert-Beer (Denisov & alții, 2003);

÷ Ca în general în orice studiu de cinetică, de interes este obținerea constantelor de viteză şi ordinelor de reacție parțiale.

Formularea modelului se începe prin asocierea de constante de viteză şi ordine de reacție necunoscute pentru fiecare reacție elementară a procesului şi scrierea cineticilor de reacție. Astfel, pentru o reacție cu preechilibru între reactanții A şi B, implicând intermediarul X şi formând produsul P, următoarele (Tabelul 2-2) reacții chimice şi ecuații matematice se obțin, unde [·] denotă concentrație în soluție apoasă.

Dacă procesul evoluează într-un mediu controlat şi porneşte de la două concentrații inițiale ale reactanților A şi B, în care nu se adaugă reactanți pe parcursul desfăşurării reacției, atunci principiul conservării masei se aplică şi are loc relația: $[A]_1-[A]_2=[B]_1-[B]_2=[P]_2-[P]_1+[X]_2-[X]_1$

între două concentrații instantanee $[\cdot]_1$ şi $[\cdot]_2$ ale compuşilor chimici implicați.

Tabelul 2-2

Reacții elementare şi ecuații asociate pentru A+B↔X→P
Elementary reactions and their equations for A+B↔X→P

Reacție chimică elementară *Ellementary chemical reaction*	Ecuație matematică asociată *Associated mathematical equation*
$A + B \rightarrow X$	$d[X]/dt = k_0 \cdot [A]^{y0}[B]^{y1}$
$X \rightarrow A + B$	$d[X]/dt = -k_1 \cdot [X]^{y2}$
$X \rightarrow P$	$d[X]/dt = -k_2 \cdot [X]^{y3}$

Modelul cinetic poate fi deci scris astfel:

÷ Reacția chimică: $A+B \overset{k1}{\leftrightarrow}_{k0} X \rightarrow_{k2} P$;

÷ Presupunerea fundamentală (că concentrațiile compuşilor sunt serii de timp): A=A(t), B=B(t), X=X(t), P=P(t)

÷ Ecuațiile matematice (în care $[\cdot]$ a fost înlocuit cu $\cdot$ pentru simplitate):

- $dA/dt = - k_0A^{y0}B^{y1} + k_1X^{y2}$;
- $dB/dt = - k_0A^{y0}B^{y1} + k_1X^{y2}$;
- $dX/dt = k_0A^{y0}B^{y1} - k_1X^{y2} - k_2X^{y3}$;
- $E \sim X \Leftrightarrow \hat{E} = aX+b$ (din legea Lambert-Beer).

÷ Inecuațiile (inegalitățile) matematice:

- $A, B, X, P \geq 0$ (fiind concentrații);
- $y_0, y_1, y_2, y_3 \geq 0$ (fiind ordine parțiale de reacție);
- $k_0, k_1, k_2 \geq 0$ (fiind constante de viteză);

÷ Serii de date experimentale E=E(t), fiind extincții experimentale pentru diferite valori ale concentrațiilor inițiale ale reactanților (A şi B);

÷ A doua presupunere fundamentală: pe durata desfăşurării reacției chimice, valorile k_0, k_1, k_2, y_0, y_1, y_2, y_3 rămân constante;

÷ Funcția obiectiv: $(aX+b-E)^2$ = min. - minimizarea abaterii valorilor experimentale față de valorile calculate de model; aceeaşi presupunere fundamentală că pe durata desfăşurării reacției chimice valoarea rămâne constantă se aplică şi coeficientului de extincție din legea Lambert-Beer (ε); presupunere care se reflectă (transferă) asupra pantei (a) din funcția obiectiv; de asemenea valoarea intercepției (b) este o valoare constantă şi are semnificația decalajului în timp între începutul efectiv al reacției chimice (t=0; X=0) şi începutul monitorizării extincției E.

Problema de cinetică a reacției chimice A+B↔X→P este o problemă grea de optimizare, având o serie de ecuații matematice şi o serie de inecuații matematice care trebuie satisfăcute, o funcție obiectiv de minimizat şi nu mai puțin de nouă parametrii necunoscuți de determinat. Astfel, la cele şapte valori ai parametrilor cinetici (k_0, k_1, k_2, y_0, y_1, y_2, y_3) se adaugă cei doi parametrii de

regresie (a şi b).

Evident că rezolvarea unei probleme de acest tip implică şi elementele de bun simț, care evită erorile grosolane în estimarea parametrilor, şi care în acelaşi timp, constituie şi elementele cheie care pot asigura succesul aplicării unui algoritm genetic. Iată câteva dintre acestea:

÷ Având la dispoziție mai multe serii de măsurători experimentale (cum e cazul şi în problema expusă aici), ceea ce cadrează în categoria elementelor de bun simț şi evită erorilor grosolane este că valorile parametrilor cinetici (k_0, k_1, k_2, y_0, y_1, y_2, y_3) şi valoarea coeficientului de extincție (ε translatat în a) obținute prin optimizarea independentă a parametrilor modelului pentru cele trei în cazul de față serii de măsurători experimentale ***trebuie să nu difere statistic semnificativ*** una de cealaltă.

÷ Al doilea element intuitiv este mai mult un element observațional (rezultat din observații), şi anume valoarea extincției (ε translatat în a) pentru intermediarul X (care este un intermediar conținând ionul de cupru) ***trebuie să se coreleze*** cu valori similare ale extincției ionului de cupru observate în alte studii din literatura de specialitate şi ***trebuie să se coreleze*** cu tranziția între nivelele energetice ale atomului de cupru (şi aici abaterea poate fi de maxim un ordin de mărime, datorită stării excitate, instabile, a intermediarului, şi datorită stării ionice în care se află acesta în intermediar). Cele două elemente de corelație expuse mai sus au servit pentru stabilirea unui domeniu pentru valorile coeficientului de extincție (ε) şi implicit ale pantei (a), domeniul fiind [100-400] $mol^{-1}m^{-1}$, în timp ce (Ungureşan & Niac, 2007) obținând o valoare de 175 $mol^{-1}m^{-1}$ la o lungime de undă de 450 nm. Acest din urmă element intuitiv a folosit la inițializarea valorilor inițiale ale pantei de regresie (fiind inițializată în toate cazurile cu 200 $mol^{-1}m^{-1}$).

÷ În Figura 2-6 este redată schema de rezolvare a problemei dificile în abordarea clasică aşa cum este ea redată în (Jäntschi & Bolboacă, 2008-rET108).

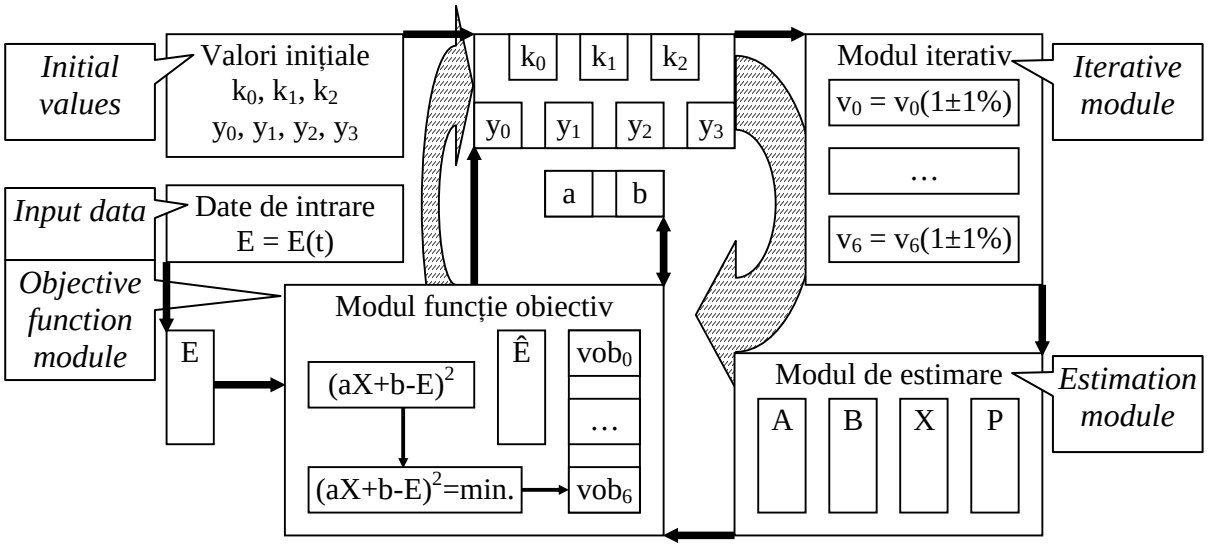


Fig. 2-6. Schema clasică de rezolvare a problemei dificile de cinetică A+B↔X→P
Classical design for solving of A+B↔X→P kinetics hard problem

Explicațiile care susțin schema de rezolvare din Figura 2-6 sunt următoarele:

÷ Valorile inițiale ale constantelor (obiect al optimizării) sunt stocate într-un şir de date, care e optimizat iterativ;

÷ Modulul iterativ foloseşte valorile $k_0...k_2$, $y_0...y_3$ să genereze o tabelă cu 3^7 ($3 = |\{-1;0;1\}|$) linii şi 7 ($7 = |\{ k_0, k_1, k_2, y_0, y_1, y_2, y_3\}|$) coloane;

÷ Fiecare linie din tabel generată în modulul iterativ este folosită în modulul estimativ pentru a face o estimare a concentrațiilor compuşilor (A, B, X, P) în funcție de timp;

÷ Modulul funcției obiectiv foloseşte estimările (A, B, X, P) pentru a obține coeficienții regresiei (a şi b);

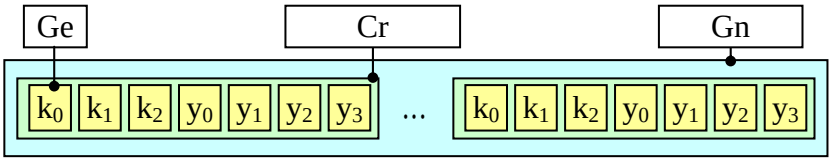
÷ Şirul $v_0..v_6$ (valori temporare pentru $k_0...k_2$, $y_0...y_3$) care produc o cea mai mică valoare a funcției obiectiv $\Sigma(aX+b-E)^2$ sunt salvate în şirul $vob_0..vob_6$;

÷ La sfârşitul unei iterații complete a modulului iterativ (3^7 iterații) valorile $vob_0..vob_6$ înlocuiesc vechile valori $k_0...k_2$, $y_0...y_3$;

÷ Ciclul Modul iterativ → Modul estimativ → Modul al funcției obiectiv este repetat până când estimarea parametrilor necunoscuți $k_0...k_2$, $y_0...y_3$ produce o aproximație stabilă, adică până când între două iterații consecutive în ciclu valorile parametrilor necunoscuți au o variație nesemnificativă statistic.

Problema de mai sus se poate reformula în formalismul algoritmilor genetici astfel:

÷ Codul genetic al unui cromozom este compus din 7 gene (Figura 2-7);



Legendă (*Legend*): Ge - Genă (*Gene*); Cr - Cromozom (*Chromosome*); Gn - Genom (*Genome*)

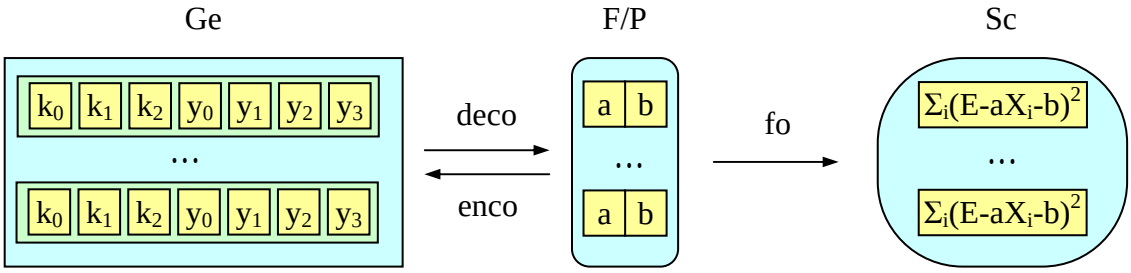
Fig. 2-7. Spațiul de căutare al algoritmului genetic în problema de cinetică A+B↔X→P

The search space of the genetic algorithm for A+B↔X→P kinetics problem

÷ Populația inițială se poate obține generând 100 de valori pentru $k_0...k_2$, $y_0...y_3$;

÷ Selecția se poate face bazat pe funcția obiectiv $\Sigma(aX+b-E)^2$ folosind una din metodele expuse în Tabelul 2-1, unde funcția Fitness(·) este chiar $\Sigma(aX+b-E)^2$;

÷ Încrucişarea a doi indivizi din populație se face în urma selecției acestora bazată pe o funcție de scor (Figura 2-8) când o încrucişare dublă se aplică alegând capetele secvenței de încrucişat generând 2 secționări ale cromozomului aleatorii (2 numere întâmplătoare între 0 şi 6); valorile genelor cuprinse între aceste două numere sunt schimbate între cei doi indivizi şi astfel se obțin descendenții încrucişării (idem Figura 2-3);



Legendă (Legend): Ge - Genotip (*Genotype*); F/P - Fenotip (*Phenotype*); Sc - Scor (*Score*); deco - decodare (*decoding*); enco - codare (*encoding*); fo - funcție obiectiv (*objective function*)

Fig. 2-8. Procesul de selecție la algoritmul genetic din problema de cinetică A+B↔X→P
The selection process of the genetic algorithm for A+B↔X→P hard problems of kinetics

÷ Mutația poate fi aleasă să introducă doar mutații întâmplătoare; astfel cu o anumită probabilitate (ce poate fi aleasă 5%) se poate decide că unul dintre descendenții încrucișării va fi supus mutației; se alege atunci o genă a acestuia folosind din nou un număr aleator (între 0 și 6) care să dea gena care va fi supusă mutației; se modifică aleator valoarea genei selectate folosind un interval de variație între cea mai mică și cea mai mare valoare a genei existentă în populație în acest moment (Figura 2-9).

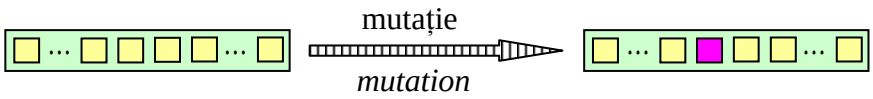


Fig. 2-9. Procesul de mutație la algoritmul genetic din problema de cinetică A+B↔X→P
Mutation process of the genetic algorithm for A+B↔X→P hard problems of kinetics

Probleme dificile de proces în biotehnologie (Jäntschi & alţii, 2009-DPA)

(Lee & alţii, 1999) au realizat estimarea parametrilor folosind o aplicație hibridă având înglobată metoda simplex deal-coborâtoare (*metoda simplex deal-coborâtoare: aici se face referire la metoda elaborată de Nelder & Mead cunoscută şi sub numele de metoda Nelder-Mead şi este o metodă de optimizare numerică pentru optimizarea problemelor fără constrângeri multidimensionale, metodă care aparține unei clase mai generale de algoritmi de căutare*) - vezi şi (Nelder & Mead, 1964) ca operator adițional într-un algoritm genetic. În timpul evoluției, la fiecare pas al iterației algoritmul hibrid operează astfel încât metoda simplex este folosită pentru selecția unui procent din porțiunea superioară a populației (în acord cu funcția de scor) pentru a produce noi soluții candidate pentru generația următoare. Restul populației este generat folosind schema de reproducere a unui algoritm genetic clasic (cuprinzând selecție, încrucişare şi mutație). Algoritmul a fost aplicat pentru optimizarea a trei cinetici de reacție, şi autorii au remarcat îmbunătățiri semnificative comparat cu cazul clasic. Reacțiile investigate au fost după cum urmează:

÷ Carboxilarea fosfo-enol-piruvatului (PEP) la oxalo-acetat (OAA) catalizat de P-enol-piruvat (PPC), când dioxidul de carbon este transformat la fosfat (Pi): $CO_2 + PEP \rightarrow OAA + Pi$

÷ Transformarea adenozin-tri-fosfatului (ATP) la adenozin-di-fosfat (ADP) în prezența OOA transformat în PEP catalizat de carboxi-kinaza fosfo-enol-piruvatului (PCK): $OAA + ATP \rightarrow PEP + ADP + CO_2$

÷ Transformarea PEP la piruvat (Pyr) în prezența ADP (transformat la ATP) catalizată de kinaza piruvatului (PyKi): $PEP + ADP \rightarrow Pyr + ATP$

Pizarro şi alții au raportat o transformare a unui algoritm genetic clasic adaptată la caracteristicile unui model capabil să explice rata de creştere în fermentarea industrială la fermentarea acidului acetic (Pizarro & alții, 2001). În abordarea făcută de autori, fiecare cromozom reprezintă o posibilă combinare a valorilor pentru fiecare din cei cinci parametrii de optimizat, reprezentați în cod binar. S-a definit aici un domeniu permis pentru valorile fiecărui parametru prin implementarea acestuia în codificarea binară a valorilor. Populația inițială a fost constituită din valori selectate la întâmplare. Programul realizat decodează aceste valori ale parametrilor pentru fiecare cromozom şi apoi le foloseşte pentru a simula un proces de fermentare cu fiecare secvență de parametrii. Algoritmul de simulare rezolvă un sistem de ecuații diferențiale având date constantele de viteză şi concentrații viabile ale biomasei, pe baza relațiilor între formarea produsului, consumul de substrat şi creşterea celulară utilizând algoritmul Runge-Kutta (*Metodele Runge-Kutta: sunt o importantă familie de metode implicite şi explicite de analiză numerică care au ca scop aproximarea soluțiilor ecuațiilor diferențiale ordinare. Aceste tehnici au fost propuse de C. Runge*

în 1895 şi completate de M. W. Kutta în 1902). Concentrațiile inițiale sunt considerate ca fiind acelea ale secvenței reprezentative a procesului, în timp ce rația inițială biomasă viabilă/total este codificată în parametrii cromozomului. Algoritmul genetic implementat are două condiții de stop importante: când nu se mai înregistrează valori reale pozitive pentru una din concentrații, şi când timpul de proces în simulare a atins timpul total de proces al secvenței reprezentative. O nouă generație cu acelaşi număr de cromozomi este formată prin aplicarea operatorilor de reproducere (aici înțeles cu sensul de copiere), încrucişare şi mutație. Cromozomii cu cea mai bună abilitate de supraviețuire obțin cel mai mare scor şi cea mai mare probabilitate de succes în adaptare (identic est: mai aproape de 1), şi au mai mare şansă să fie selectați şi copiați în noua generație. Încrucişarea uniformă este folosită şi cei mai buni cinci cromozomi din fiecare generație trec în generația următoare neschimbați. Aceşti cromozomi sunt numiți elitişti. Cromozomii frați şi cei care ies din domeniu sunt blocați folosind o buclă de repetiție cu filtru. Când aceşti cromozomi sunt descoperiți după încrucişare, alți cromozomi obținuți de asemenea din încrucişare îi substituie, iar dacă sunt descoperiți după mutație, ei sunt înlocuiți de cromozomii originali în aceleaşi poziții dar sunt mutați din nou cu aceeaşi şansă de mutație. În acest proces, mutabilitatea nu este mărită, şi numărul de cromozomi rămâne constant. Procesul se opreşte după cinci generații fără schimbări mai mari decât un procent fixat al răspunsului mediu al cromozomilor elitişti. Algoritmul este rulat de cinci ori la fiecare rulare a programului. O execuție finală are loc în care populația este compusă din cei mai buni cromozomi găsiți în fiecare din execuțiile anterioare. Concentrațiile acetice în fermentatoarele fabricii Vinagrerıas Riojanas SA (Logrono, Spain), obținute prin NIR (infraroşu apropiat), au fost studiate aplicând această metodologie. Datele au fost culese pe o perioadă de 4 luni fără a interveni în parametrii proceselor industriale, cum e cazul condițiilor de oxigenare şi temperaturii. Temperatura medie a fost de 29.5°C şi condițiile de oxigenare au fost suficiente pentru a asigura necesarul de oxigen, astfel încât oxigenul a devenit un substrat ne-limitator. Astăzi fermentatoarele industriale lucrează discontinuu cu schimbări (în parametrii de mediu). Bazinele bioreactoarelor studiate au fost tot timpul hrănite cu vin alb de aceeaşi origine. Timpul de proces a fost de aproximativ 30-31 ore şi pe această durată de timp 218 secvențe complete au fost obținute. O concentrație medie a secvențelor de fermentare a fost calculată din datele experimentale şi a servit în modelarea procesului. Variabilitatea concentrației în cadrul secvențelor este datorată erorilor analitice şi factorilor care nu pot fi controlați în procesul industrial, cum ar fi concentrația în etanol a vinului între procesele de fermentare. Astfel, modelul obținut folosind valoarea medie nu modelează această varianță.

Estimarea parametrilor cinetici ai poli-esterificării între acidul gras dimeric şi etilen-glicol a constituit subiectul investigației folosind un algoritm genetic clasic uşor modificat (Guangzhu & alții, 2006). Lucrarea arată că modelul dezvoltat de autori este util pentru poli-esterificarea acidului

dimeric cu etilen-glicol catalizat de acidul para-toluen-sulfonic. Astfel, autorii au folosit 28.1g (0.05moli) de acid gras dimeric, 3.11g (0.05moli) de etilen-glicol şi 0.5% acid para-toluen-sulfonic (ca şi catalizator) amestec care a fost pus într-un balon cu fund rotund (cu trei capete), care a fost echipat cu un agitator şi un tub pentru azot. Azotul a fost introdus în balon pentru a înlătura oxigenul şi a preveni oxidarea materialelor. Balonul a fost plasat într-o baie de ulei cu o temperatură de 170°C. După 30 min. de reacţie, azotul a fost oprit şi o pompă de vid a fost folosită pentru a scoate apa din reactant. Reacţia a continuat 8-10 ore în vid. Cantitatea de acid în reactant a fost măsurată la anumite momente de timp pe durata desfăşurării reacţiei. Estimarea parametrilor a fost realizată în trei paşi. În primul rând, ordinul de reacţie a fost confirmat utilizând presupunerea unei activităţi chimice egale. În al doilea rând, experimentele au fost gândite să permită estimarea parametrilor de constantă de viteză a reacţiei între carboxil şi monomer şi respectiv între hidroxil şi polimeri. Excesul de monomer a fost adăugat reactantului după ce acesta a reacţionat pentru câteva ore cu o proporţie de materiale de 1:1, şi reacţiile au putut fi ignorate exceptând cea pentru monomerul adăugat şi polimeri. În final, valorile obţinute au fost introduse pentru a obţine ecuaţiile de viteză şi a obţine valorile vitezelor de reacţie între carboxil pe monomer şi hidroxil pe monomer, şi între carboxil pe polimeri şi hidroxil pe polimeri.

Modificând un algoritm genetic prin utilizarea de operaţii genetice ARN asupra unui model ADN şi utilizarea de operatori de mutaţie şi încrucişare îmbunătăţită, (Tao & Wang, 2007) au reuşit să realizeze estimarea parametrilor pentru două cazuri: cracarea termică a unui ulei greu folosind un model cu trei mase şi unitate cu fluid catalitic de cracare cu fracţionator (care converteşte uleiuri cu masă moleculară mare în produse hidrocarbonate mai uşoare). În ambele cazuri s-a arătat că metodologia dezvoltată este efectivă în estimarea parametrilor procesului chimic.

Un algoritm genetic pentru căutare cu ajutorul calculatorului a fost raportată recent în (Wollman & alţii, 2008). Algoritmul la care se face referire utilizează măsurătorile experimentale pentru a descoperi maşinăria de mecanică moleculară care se află în spatele procesului. Au fost efectuate măsurători în serii de timp mari efectuate in vivo şi pe celule ou perturbate experimental cu scopul de a identifica modelele mecaniciste ce stau la baza coordonării generatoarelor de forţe mitotice în celulele ou de *Drosofila*. Algoritmul a fost capabil să caute şi să elimine mii de modele posibile şi să identifice şase strategii distincte pentru integrarea motorului microtubulelor care corelează cu datele avute la dispoziţie. Multe caracteristici ale acestor şase strategii prezise au fost conservate, incluzând un mecanism persistent condus de kinesin-5 combinat cu inhibiţia anafazei B-specifică a caracteristicilor kinesinice şi profile de activare-deactivare pentru motoarele mitotice cheie. Abordarea de inginerie inversă a utilizat în mod indirect date cantitative pentru a realiza o căutare exhaustivă cu calculatorul şi a identifica astfel construcţia mecanică a celulei ou care poate să explice datele observate. Strategia a permis examinarea unui număr mare de parametrii posibili şi

mecanisme alternative utilizând modele grosiere dintre care ulterior au fost rafinate modelele promițătoare incluzând componente adiționale şi rezultând astfel modele mult mai detaliate. Aşa cum autorii subliniază, schema de lucru sugerată poate fi uşor adaptată şi la celule ou mitotice ale altor organisme şi in vitro (unde poate fi gândită diferit), şi, în fapt, la multe alte sisteme biomecanice pentru care există suficiente date cantitative.

Probleme dificile de aliniere a amprentelor genetice (Jäntschi & alții, 2009-GSA)

Încă de la început trebuie precizat că nu există o funcție obiectiv (o măsură a calității globale a alinierii) care să fie de utilizat pentru a spune că o aliniere (Tabelul 2-3) este preferabilă alteia, sau pentru a spune că cea mai bună posibilă aliniere a fost obținută.

Tabelul 2-3

Aliniere de secvențe genetice
Alignment of genetic sequences

Două secvențe nealiniate *Two unaligned sequences*	Secvența_1 *Sequence_1*	U	A	A	G	C	C	U	C	A	G	U	A	A
	Secvența_2 *Sequence_2*	A	A	C	C	C	U	C	A	U	A			
O aliniere posibilă a două secvențe *A possible alignment of the two sequences*	Secvența_1 *Sequence_1*	U	**A**	**A**	G	**C**	**C**	**U**	**C**	**A**	G	**U**	**A**	A
	Secvența_2 *Sequence_2*		**A**	**A**	C	**C**	**C**	**U**	**C**	**A**		**U**	**A**	

O abordare este a realiza un aliniament progresiv, care are două alternative principale:

÷ Utilizarea de modele Markov ascunse (Krogh & alții, 1994) care simultan găsesc un model şi un model probabilistic de substituții, inserări şi ştergeri care sunt cele mai auto-consistente;

÷ Utilizarea funcțiilor obiectiv care măsoară calitatea aliniamentului multiplu şi găsesc folosind scorul cel mai bun aliniament.

Această din urmă abordare are avantajul că poate fi optimizată orice funcție obiectiv care poate fi concepută, atâta timp cât căutarea de funcții obiectiv pentru aliniamentul de secvențe, şi probabil pentru alte scopuri decât cel de față, este cu siguranță o arie cheie de cercetare. Evaluarea aliniamentului poate fi făcută din nou utilizând o funcție obiectiv care este în mod simplu o măsură a calității alinierii.

Principalele elemente ale problemei alinierii secvențelor de gene cu care funcția obiectiv trebuie să opereze sunt:

÷ Potrivirile şi golurile: sunt aici două obiective principale - maximizarea potrivirilor şi minimizarea golurilor; o potrivire poate avea o semnificație biologică diferită de cea a unui gol;

÷ Lungimea secvenței: o simplă numărare a potrivirilor (şi golurilor) va produce un număr care creşte odată cu creşterea lungimii secvenței;

÷ Deplasarea secvențelor: deplasând o secvență se va produce un gol la începutul şi unul la sfârşitul secvențelor aliniate, goluri care trebuie tratate diferit de celelalte goluri din interior;

Există multe abordări în ceea ce priveşte construcția unei funcții obiectiv. Astfel, (Karlin & Altschul, 1990) numără patru tipuri de scoruri:

÷ Bazate pe sarcini electrice: nu toți aminoacizii prezintă aceeaşi sarcină electrică parțială (sau aparentă) într-un mediu dat (cum ar fi serul fiziologic sau celulele musculare) şi valorile sarcinilor pot fi obținute mediind valorile ce rezultă dintr-un experiment; alternativ, valoarea pK (constanta de disociere a acidului) sau mai exact pK-7 poate fi folosită;

÷ Bazat pe potrivirile unui aminoacid (cum ar fi A în Tabelul 2-3);

÷ Derivat din frecvențele țintă dând diferite ponderi potrivirilor de diferiți aminoacizi;

÷ Bazat pe alfabete de structură când aminoacizii sunt partiționați în clase (cum ar fi intern, extern, ambivalent);

Pentru a rezolva o problemă specifică, se aduc frecvent modificări algoritmilor genetici clasici. Astfel, în (Notredame & Higgins, 1996) se comunică realizarea unui produs soft numit SAGA (acronim al denumii din limba engleză: Sequence Alignment by Genetic Algorithm) care utilizează o schemă de planificare pentru a controla utilizarea a 22 de operatori diferiți pentru combinarea aliniamentelor sau comutarea lor între generații. Aceştia au implementat ca funcție obiectiv o funcție de cost pentru aliniamentul multiplu (A) ca suprapunere liniară de costuri între perechi de sub-secvențe aliniate:

$$OF(A) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} W_i \cdot W_j \cdot Cost(A_i, A_j)$$

unde W_i şi W_j sunt ponderi ale sub-secvențelor A_i şi A_j, în timp ce funcția Cost(·,·) cumulează penalități pentru deschiderea unui gol şi penalități pentru extinderea acestuia. Pentru un rezumat extins al diferitelor moduri de a penaliza golurile în aliniamente multiple se poate consulta lucrarea (Altschul, 1989). Programul este îmbunătățit ulterior pentru a performa în alinierea secvențelor de aminoacizi ai ARN (Notredame & alții, 1997).

După aliniamentul secvențelor, două întrebări se pot formula:

÷ Este alinierea semnificativă făcând apel la un anume model statistic?

÷ Cât de stabilă este alinierea? (cu referire la întrebarea: Care sunt aliniamentele alternative care produc scoruri identice sau similare?)

Prima întrebare este despre cunoaşterea probabilității observării unui anume alinieri din pură întâmplare. Aceasta este o problemă foarte dificilă şi are soluții numai în anumite condiții (Waterman & Vingron, 1994). A doua întrebare se referă la interpretarea acestor alinieri şi au fost deja obținute unele rezultate (Vingron & Argos, 1990).

Când se dispune de informație adițională (cum ar fi structura secundară a unei proteine dintre cele două care se doreşte a fi aliniate) este posibil să se reducă complexitatea problemei. Astfel de aliniamente includ interacțiuni ne-locale şi problema s-a dovedit a fi o problemă dificilă (Lathrop, 1994). În acest caz, funcțiile obiectiv trebuie să ia în considerare această nouă provocare. Pentru a reflecta acest fapt în funcția obiectiv, în lucrarea (Corpet & Michot, 1994) se propune o funcție obiectiv având două penalități specifice de poziție: GOS (o penalitate pentru deschiderea unui gol între două perechi alăturate) şi GO (o penalitate pentru deschiderea unui gol în regiuni nestructurate), şi o a treia penalitate, GEP (care penalizează lungimea golului). Este uzual să se definească şi folosească ponderi predefinite. Astfel, în (Corpet & Michot, 1994) se propun următoarele valori: GO=5; GOS=8; GEP=0.3. Oricum, penalitatea totală a golurilor se calculează cu expresii de forma: $GAP(\cdot)=a(\cdot)\cdot GOS+b(\cdot)\cdot GO+c(\cdot)\cdot GEP$, unde a este numărul de goluri între perechi în lanțurile duble, b este numărul de goluri neterminale în restul lanțurilor, şi c este lungimea totală a golurilor neterminale (incluzându-le şi pe cele din lanțurile duble). Într-o astfel de abordare, scorul de aliniere (funcția obiectiv) este calculat ca: $OF(\cdot)=Pr(\cdot)+\lambda\cdot Se(\cdot)-GapP(\cdot)$, unde $Pr(\cdot)$ este o funcție de perechi de reziduuri aliniate în aliniament, $Se(\cdot)$ este bazată pe structura secundară, şi evaluează stabilitatea împachetării induse de către secvența de referință asupra secvenței subordonate. Parametrul λ are efectul balansării contribuției care vine de la informația structurii primare şi de la cea a structurii secundare, fiind o constantă pozitivă.

Pornind de la modelul propus în (Corpet & Michot, 1994), şi îmbunătățind o versiune anterioară de algoritm genetic (Notredame & Higgins, 1996), s-a propus (Notredame & alții, 1997) un model pentru alinierea secvențelor de ARN, când s-a observat că optimizarea este foarte dificilă pentru $\lambda>0$ (când structura secundară este luată în considerare). Totuşi, s-a reuşit (Notredame & alții, 1997) obținerea de rezultate foarte bune când s-a folosit drept proteină cu structură cunoscută o mitocondrie de *Homo sapiens* (X03205 şi V00702), iar ca proteină cu structură secundară necunoscută o mitocondrie provenită de la câteva specii diferite (*Drosophila virilis* X05914, *Apis mellifera* S51650, *Penicillium chrysogenum* L01493, şi aşa mai departe). S-a arătat astfel că cea mai bună potrivire de perechi obținută judecând după alinierea de referință a variat de la 66.6% la 84.9% pentru nouă experimente (cu o statistică medie pe care am calculat-o şi este de 79.3±4.6% la încredere de 95%) şi când cea mai bună valoare a parametrului λ a variat de la 1 (pentru *Oxytrichia nova* X03948, *Latimeria chalumnae* Z21921, *Xenopus laevis* M27605) la 6 (la *Saccharomyces cerevisiae* V00702).

În (Pan & alții, 2002) se comunică realizarea aplicației GABRIEL (de la denumirea în limba engleză: Genetic Analysis By Rules Incorporating Expert Logic), un sistem bazat pe reguli (incluzând similaritate, tipar şi reguli bazate pe probant), aplicație dedicată a aplica cunoaştere domeniu-specifică şi procedurală în mod sistematic şi uniform pentru analiza şi interpretarea datelor

din micro-şiruri de ADN. Utilizând informația cunoscută despre efectele adăugării de ser în biologia fibroblastului uman, şi anume (Greenberg & Ziff, 1984; Lau & Nathans, 1987; Ryder & alții, 1988; Jähner & Hunter, 1991), autorii au analizat un set conținând 517 gene care au prezentat o transcripție alterată în fibroblastul uman din prepuț, după adăugarea culturilor inhibate la creştere aşa cum a fost raportat în literatură în (Iyer & alții, 1999). Utilizând reguli bazate pe tipar a fost posibil să se obțină (Pan & alții, 2002) valorile la o serie de parametrii care nu au fost explicit definiți în analiza din (Iyer & alții, 1999), aceştia fiind: elevația (în engleză: elevated), linie de bază (în engleză: baseline), imediat (în engleză: immediately), rămas (în engleză: remained), şi perioadă scurtă (în engleză: short period) - aici făcându-se referire la folosirea acestor termeni în (Iyer & alții, 1999) pentru exprimarea plastică a elementelor ce caracterizează observațiile efectuate asupra desfăşurării proceselor experimentale). Astfel, s-a observat în (Pan & alții, 2002) că:

÷ elevația necesită cel puțin o schimbare de 2 împachetări;

÷ linia de bază la -1.2 produce o configurație mai bună pe scara logaritmică produce o configurație mai bună decât -1.0 găsită în (Iyer & alții, 1999);

÷ timpul de răspuns imediat (de la: I/E - immediate/early) al genei a fost obținut utilizând un arbore de decizie şi a variat pentru diferite configurații experimentale menționate în (Iyer & alții, 1999) între 15 min. şi 1 oră;

Pornind de la date disponibile în literatura de specialitate cu privire la afinitatea de legătură proteică a moleculelor din clasa I MHC (de la termenii în engleză: major histocompatibility complex), şi utilizând un algoritm genetic, s-a obținut în (Carpio-Muñoz & alții, 2002-Bioinformatic) o nouă metodă de calcul a gradului antigenic al peptidelor, ceea ce facilitează designul de medicamente peptidice utile clinic şi fără efect imunologic, precum şi a medicamentelor de terapie imunologică şi a vaccinelor pentru maladii autoimunitare şi a cancerului. O metodologie pe mai multe nivele pentru evaluarea afinității peptidei pentru clasa I de molecule MHC a fost propusă:

÷ Primul nivel: leagă structura primară a peptidelor de activitatea măsurată experimental a acestora. Acest pas implică un algoritm genetic combinat cu analiza de profil pentru detecția proteinelor înrudite aşa cum este descris în (Gribskov & alții, 1987) şi duce la un set de coeficienți care exprimă cantitativ contribuția fiecărui aminoacid din secvența peptidică la activitatea de legătură. În (Carpio-Muñoz & alții, 2002-Bioinformatic) se raportează rezultate obținute pentru peptidele 9-merice şi construcția profilului de activitate de legătură pentru moleculele clasei I de MHC, şi se sugerează că profilul rezultat poate fi folosit pentru a prezice activitatea de legătură a oricărei secvențe 9-merice.

÷ Al doilea nivel: analizează complexul format de peptidă şi regiunea receptoare a moleculei MHC, prin implicarea altui meta-euristic, călirea simulată, pentru evaluarea interacțiunii bio-

macromoleculare ce a fost comunicată anterior (Carpio-Muñoz & alții, 2002-Miax). Analiza este efectuată pe un set de mai multe tipuri de peptide cu afinitate de legătură la clasa I a MHC, şi tendințele în corelație hidrofobă şi complementaritățile electrostatice sunt evidențiate că conduc legarea moleculelor şi formarea complexului. Tendințele derivate din această aproximație sunt apoi aplicate pentru a evalua probabilitatea de legătură ligand-receptor la nivel atomic.

Un alt rezultat important obținut în (Pan & alții, 2002) a fost utilizarea algoritmilor genetici implementați în GABRIEL pentru analiza de căutare a tiparelor (de la: PSA - Pattern Search Analysis), pentru a identifica tiparele (*tipar (în context statistic): referă sistemele în care datele sunt clasificate în funcție de inter-relațiile care se stabilesc între părțile componente*) printre profilurile expresiilor genetice (*profil al expresiei genetice: set de date care portretizează caracteristicile expresiilor de gene în anumite condiții*). Cu ajutorul studiului PSA s-a reuşit reconstrucția rezultatelor din (Iyer & alții, 1999). Atunci când s-a aplicat regula de continuitate a probantului GABRIEL a detectat şi alte continuități care nu au fost găsite în (Iyer & alții, 1999), care au folosit o analiză de clasificare ierarhică cu dendrograme. S-a remarcat de asemenea că algoritmul genetic a fost capabil să distingă între profilurile expresiilor având diferențe subtile care nu sunt evidente la o scanare vizuală a datelor. Mai mult, în anumite instanțe când rezultatele au diferit de cele din (Iyer & alții, 1999), funcția de reguli de explicare a GABRIEL a indicat parametrii statistici sau limită responsabili pentru diferență. Se sugerează astfel (Pan & alții, 2002) că caracteristicile cheie ale lui GABRIEL pot fi utile pentru analiza seturilor de date mari generate de alte abordări genomice şi proteomice.

O abordare interesantă este utilizarea formalismului sistemului S (Savageau, 1976) pentru modelarea dinamicii genelor regulatoare şi efectoare în circuitele induse (Hlavacek & Savageau, 1996). Deoarece metoda clasică (fără formalismul sistemului S) a fost găsită prezicând numai un număr mic de parametrii (Tominaga & alții, 1999; Tominaga & alții, 2000), şi rata de convergență a fost mică, au fost astfel aplicate o serie de îmbunătățiri în (Kikuchi & alții, 2003) asupra funcției obiectiv, metodei de încrucişare, şi strategiei de optimizare, elemente care sunt redate în continuare:

÷ Funcția obiectiv (scor), ca sumă de:
  - Scorul metodei clasice (Tominaga & Okamoto, 1998) reproducând anumite evoluții în timp (care converge la mai multe minime locale şi rar reține scheletul structurii;
  - Un termen de scurtare, ce permite detectarea scheletelor de structură cu găsirea aşteptării unei căi necunoscute, fiind suma valorilor absolute ale parametrilor de model (Savageau, 1976).

÷ Încrucişare simplă, cum se propune în (Tsutsui & alții, 1999), deoarece în (Higuchi & alții, 2000) s-a arătat că produce creşterea vitezei de optimizare, având următoarele caracteristici:

- Moşteneşte independența de sistemul de coordonate;
- Vectorul de valori al descendentului moşteneşte caracteristicile părinților şi eşantionarea reflectă o anumită legătură între parametrii;
- Balansează între explorare şi exploatare în generarea descendenților, şi lucrează bine pentru funcții multimodale (*Funcție multimodală: o funcție care are mai mult de un optim (minim sau maxim) şi în acelaşi timp pot avea unul sau mai multe optime globale*) şi/sau în prezența epistasisului (*Epistasis: mascarea expresiei unei gene într-o poziție (locus) într-un cromozom de una sau mai multe gene aflate în alte poziții; termen intodus de William Bateson in 1907 (Bateson, 1907)*) printre parametrii;
- Este un operator simplu şi nu este consumator de timp.

÷ Strategie de optimizare gradată, fiind ştiut că este dificil să se optimizeze toți parametrii odată, parametrii cu o importanță relativă mică, care devin aproape 0, sunt astfel detectabili. Aceste valori sunt fixate la 0 şi optimizarea este repetată ulterior de la început, când mulți parametrii de importanță mai mică sunt detectați.

O adaptare a unui algoritm genetic numită Lamarckian GA este comunicată şi folosită în (Kadirvelraj & alții, 2006) pentru a ancora fragmentul trizaharidic GBSIII la locația de legătură antigenică a Fv1B1, când s-a reuşit să se furnizeze o interpretare vastă pentru un volum mare de date biochimice şi imunologice obținute cu privire la recunoaşterea Ab (*legarea Ab a peptidelor: proprietate măsurată folosind metoda ELISA competitivă*)- cu detalii în (Uray & alții, 2003) a bacteriilor polizaharidice şi autorii notează că rezultatele ar trebui să fie aplicabile şi la alte interacțiuni Ab-carbohidrat.

Alinierea de secvențe are aplicații importante, una dintre acestea fiind clasificarea. Astfel, în (Moore & alții, 2007) s-a realizat o căutare de tip Bayes şi a maximei probabilități (ML - de la terminologia în limba engleză: maximum-likelihood) pe 61 de gene plastide codificatoare de proteine pe 5 linii majore de *mesangiosperme* pentru 45 de taxe. Pentru căutările de tip ML, autorii au folosit aplicația GARLI (de la denumirea în limba engleză: Genetic Algorithm for Rapid Likelihood Inference) care implementează (Zwickl & Hills, 2006) un algoritm genetic pentru realizarea de căutări euristice ML rapide. Pentru căutările de tip Bayes, s-a optat pentru MrBayes3 (Ronquist & Huelsenbeck, 2003) care implementează o variantă a metodei MCMC (de la denumirea în limba engleză: Markov Chain Monte Carlo) numită MCMC cuplată Metropolis (Gilks & Roberts, 1996) aşa cum în (Huelsenbeck & alții, 2001) se sugerează a fi cea mai utilă metodă numerică ce permite aproximația probabilității posterioare a unui arbore. Evaluarea parametrilor estimați de ML în (Moore & alții, 2007) a fost realizată utilizând abordarea BS neparametrică (Baldwin & Sanderson, 1998). Analiza filogenetică din (Moore & alții, 2007) a arătat că parametrii estimați de GARLI sunt totdeauna foarte aproape de valorile unei optimizări complete.

Studii care vin să confirme identitatea folosind algoritmi genetici sunt numeroase, (Tiptara & alții, 2008) fiind doar unul dintre ele. În acest caz modalitatea uzuală de investigare este folosind aplicații dedicate, cum sunt MS BLAST (Shevchenko & alții, 2001) şi Mascot (David & alții, 1999).

Serii de timp din măsurători repetate ale micro-şirurilor ADN responsabile de migrarea factor-indusă a creşterii hepatocitelor din cheratinocitele umane primare au servit la alinierea repetată cu ajutorul unul algoritm genetic (Busch & alții, 2008). Autorii au reuşit să dovedească că modelarea inversă a rețelei genetice stabileşte un model abstract al unui sistem biologic complex capabil să captureze bine proprietățile dinamice şi prezentând capacitate predictivă. Analiza cineticii expresiilor genetice descifrează dinamica unui mic, dar extrem de complex model regulator genetic de migrare prin modelarea inversă. Aşa cum autorii subliniază, designul experimental (Fisher, 1926; Taguchi, 1995; Bolboacă & Jäntschi, 2007-DoE; Nguyen & Liu, 2008) este esențial şi a fost elementul cheie de succes în descoperirea orchestrației complexe a căilor multiple ce controlează migrația celulară:

÷ Alegerea timpului de observare axat pe momentul migrației şi măsurând la o suficient de mare rată pentru a captura dinamica rețelei genetice;
÷ Alegerea genelor candidate din ranguri obținute prin ordonarea acestora;
÷ Alegerea funcției biologice şi validarea încrucişată împotriva stimulilor proliferativi;
÷ Limitarea numărului de gene dominante în dinamica rețelei;
÷ Combinarea criteriilor de tărie şi robustețe în căutarea modelului de interacțiune genetică plauzibil;

Probleme dificile de evoluție şi arbori filogenetici (Jäntschi & alții, 2008-PTA)

Răspunsul la probleme dificile de ***evoluție*** se caută adesea folosind algoritmi genetici. Astfel, genomul cloroplastului din *Manihot esculenta* şi evoluția atpF în familia *Malpighiales* sunt subiectul cercetărilor în (Daniell & alții, 2008), coniferele genului *Taxus* şi evoluția genelor paclitaxe biosintetice TS şi DBAT sunt subiectul lucrării (Hao & alții, 2009), evoluția parfumului trandafirilor chinezești sunt subiect al (Scalliet & alții, 2008), iar al plantelor cățărătoare *Hemiptera* şi *Psylloidea* în asociere cu *Anacardiaceae* sunt subiect al unui studiu sistematic în (Burckhardt & Basset, 2000).

Studiul arborilor filogenetici utilizând corespondența cu setul potrivirilor perfecte în grafuri complete (Jäntschi & Diudea, 2009) a constituit subiectul lucrării (Diaconis & Holmes, 1998). Autorii au arătat că corespondența menționată produce o metrică de distanță între arborii filogenetici, şi devine astfel o cale pentru enumerarea tuturor arborilor într-un număr minim de paşi. Identificarea arborelui filogenetic este o problemă dificilă, şi în cadrul acesteia autorii au arătat că efectuând produsul a două potriviri care este cunoscut sub denumirea de algebra Brauer (Brauer,

1937), se permite o implementare simplă a unui algoritm genetic.

Problemele legate de eşantioane mari de taxoni în estimarea filogenetică sunt discutate în (Lemmon & Milinkovitch, 2002), unde un algoritm genetic meta-populațional (metaGA) implicând mai multe populații de arbori care sunt forțate să coopereze în căutarea arborelui optim a fost găsit potrivit. Un rezultat important se desprinde din (Lemmon & Milinkovitch, 2002), şi anume că frecvențele cu care arborii şi clicile prelevate utilizând algoritmul metaGA pot corespunde la estimatorii nedeplasați ai probabilităților ulterioare (Huelsenbeck & alții, 2001).

O altă analiză de arbore filogenetic în liniile majore ale *Brachycera* a fost realizat în (Wiegmann & alții, 2003) şi indică că *Brachycera* este originată în Triasicul târziu sau în Mezozoicul timpuriu şi toate liniile majore inferioare ale zburătoarelor *Brachycera* au avut origini contemporane în Jurasicul mijlociu înainte de originile plantelor de flori (angiospermelor). Autorii au obținut o rezoluție mărită a filogeniei pentru *Brachycera*, şi estimările revizuite ale epocii zborului îmbunătățește contextul temporal al interferențelor evolutive şi comparațiilor genomice între organismele zburătoare model. Secvențele de nucleotide au fost aliniate manual cu un editor de aliniere interactiv numit Genetic Data Environment 2.2 (Smith & alții, 1994). Datele filogenetice au inclus 2220 de caractere din 28S rDNA (cuprinzând 608 variabile şi 294 parsimonic (*parsimonie: adoptarea celor mai simple presupuneri în formularea teoriei sau interpretarea datelor, în special în acord cu regula lamei de ras a lui Ockham (principiu atribuit logicianului William of OCKHAM, care subliniază că trebuie eliminate toate acele presupuneri care nu fac nici o diferență în predicțiile observate ale ipotezelor explicatoare sau teoriei); în latină: lex parsimoniae - entia non sunt multiplicanda praeter necessitatem*) informative corespunzător la toate datele; 493 variabile şi 296 informative în *Brachycera* şi 101 caractere morfologice (Yeates, 2002). Analiza filogenetică a setului de date combinat a fost efectuată cu opțiunea parsimonie din programul PAUP (Fink, 1986).

Un studiu extins cu privire la evoluția timpurie şi diversificarea furnicilor a fost raportat în (Brady & alții, 2006). O importantă parte a acestui studiu este reprezentată de elaborarea metodelor de studiu care au fost folosite, şi care se regăsesc descrise în informația suplimentară lucrării menționate. Astfel, autorii au folosit o serie de programe, toate acestea operând cu algoritmi genetici:

÷ Pentru alinierea secvențelor: Clustal X (Larkin & alții, 2007);

÷ Pentru datarea divergenței (estimarea lungimii ramificației) şi inferență filogenetică (analiză parsimonică; inferența arborilor optimi de probabilitate maximă; comparația unui set de amplasări ale grupurilor de ieşire în arborele de grupări interne folosind testul Shimodaira-Hasegawa): PAUP* v4.0b10 (Fink, 1986);

÷ Pentru obținerea modelelor de substituție nucleotidică: ModelTest v3.06 (Posada & Crandall, 1998);

÷ Pentru analiza neparametrică de încărcare a probabilității maxime: GARLI v0.94 (Schultz & alții, 2006), derivat din GAML (Lewis, 1998);

÷ Pentru analiza Bayes: MrBayes v3.1.2 (Ronquist & Huelsenbeck, 2003);

÷ Pentru datarea divergenței (estimarea lungimii ramificației) utilizând abordarea probabilității penalizate: r8s v1.7 (Sanderson, 2002; Sanderson, 2003).

Într-un studiu ulterior (Schultz & Brady, 2008), cercetările asupra furnicilor au avut ca rezultat comunicarea identificării de relicve încă în viață de specii de furnici *attine* care ocupă poziții filogenetice care sunt de tranziție între sistemele agricole. Metodologia folosită include ca mai sus analiza filogenetică (parsimonie, probabilitate maximă şi datarea divergenței), un model nucleotidic de tip Bayes şi un model MCMC al codonului, şi în plus o nouă abordare, topografia filogenetică a sistemelor agricole:

÷ Taxelor terminale le-au fost asociate stări într-un caracter cu şase stări reprezentând patru sisteme agricole de *attine* şi agricultura tăietorilor de frunze (nu, inferior, mediu, superior, tăietor de frunze, *coral-fungus*);

÷ Cinci specii (*Myrmicocrypta n. sp. Brazil*, *Mycetagroicus triangularis*, *Cyphomyrmex n. sp.*, *Cyphomyrmex morschi*, *Trachymyrmex irmgardae*, şi *Pseudoatta n. sp.*) ale căror stări au fost asociate la 'necunoscut' şi *Trachymyrmex papulatus* a primit starea 'agricultură inferioară', asocieri de stări bazate pe o colecție de grădină din Argentina (o a doua colecție din aceeaşi localitate a cultivat o grădină tipică de *attine* înalte);

÷ Evoluția caracterelor a fost optimizată într-un arbore de consens Bates codon-model (cu lungimile ramurilor) sub ambele parsimonie folosind MacClade [2] şi probabilitate maximă folosind modulul StochChar al programului Mesquite [3];

÷ În parsimonie, optimizările stărilor ancestrale au fost neambigue. În ipoteza modelului Markov cu k stări şi 1 parametru (Lewis, 2001), probabilitatea ca fiecare sistem agricol să se ridice din cel mai recent strămoş al clicii de furnici corespunzătoare a fost, ca proporție din probabilitatea totală distribuită între cele şase stări ale caracterului, de 0.9831 pentru inferior, 0.9995 pentru mediu, 0.9905 pentru superior, 0.9924 pentru tăietorii de frunze şi 0.9998 pentru *coral-fungus*.

Altă analiză filogenetică au fost condusă utilizând algoritmi genetici pentru producție setului de reguli necesare pentru a modela distribuțiile populaționale geografice ale maimuțelor păianjen şi bocitoare prin caracterizarea nişelor sale ecologice (Ortiz-Martinez & alții, 2008). Datorită proceselor întâmplătoare implicate în model, fiecare model obținut cu un singur set de date este diferit; pentru a captura variabilitatea autorii au elaborat 100 de modele pentru fiecare specie şi apoi au selectat 10 modele care dau cea mai mică eroare de suprapunere şi omisiune, urmând procedura

[2] Maddison DR, Maddison WP. 2000. MacClade v4.0. http://macclade.org/
[3]Maddison WP, Maddison DR. 2006. Mesquite v1.1. http://mesquiteproject.org/

descrisă în (Anderson & alţii, 2003). Autorii au putut să obţină că maimuţele păianjen ocupă un areal mai mare şi o diferenţă de altitudine mai mare decât maimuţele plângătoare. Validarea modelului a fost făcute pentru maimuţele păianjen, fiind suficiente date disponibile pentru această specie; validarea modelului a indicat că distribuţia prezisă a speciei este statistic mai mare decât cea aşteptată de întâmplare.

Probleme dificile în sisteme agricole şi horticole

Aplicaţiile algoritmilor genetici în probleme specifice sistemelor agronomice au constituit subiectul unor analize critice ale literaturii de specialitate (Hashimoto, 1997; Mayer & alţii, 1999). O serie de aplicaţii importante au fost raportate în literatura de specialitate de atunci încoace, şi noi perspective de cercetare au fost anunţate (Anisimova & Liberles, 2007).

Sisteme de decizie bazate pe algoritmi genetici pot elabora modele capabile să stabilească priorităţi (Smith, 2001), să configureze sisteme de producţie, şi să elaboreze managementul resurselor (Kuo & Liu, 2003; Wardlaw & Bhaktikul, 2004).

Orientat pe aspectele fundamentale, în (Annevelink, 1992) se comunică realizarea unui sistem menit să asiste decizia şi managementul în sisteme horticole şi implementat sub formă de program utilizabil pe un calculator personal (PC). O remarcă este necesară aici: în general programele bazate pe algoritmi genetici sunt mari consumatoare de resurse de memorie şi timp; adaptarea acestor programe pentru a fi folosite pe calculatoarele PC obişnuite este astfel notabilă). Sistemul, denumit IMAG IPP posedă un nivel de planificare tactică, şi un mediu interactiv pentru planificarea spaţiului în nivelul de planificare operaţională.

Crearea unui sistem de decizie care să fie utilizat în cadrul unei metodologii de control optimal a constituit subiectul lucrării (Seginer & alţii, 2007). Sistemul de decizie a fost elaborat pentru operarea unui sistem de control al umidităţii într-un solar cu ventilaţie, în care umiditatea a fost folosită drept caracter dominant de control.

Formularea considerentelor teoretice care trebuie să stea la baza elaborării a unui model dinamic pentru controlul producţiei (Buwalda & alţii, 2006) şi utilizarea acestuia în scopul optimizării randamentului şi consumului energetic (Henten & alţii, 2006) pentru cultivarea ardeilor dulci (*Capsicum annuum*) sunt subiecte ale preocupărilor actuale.

Optimizarea irigării (Montazar & alţii, 2008) şi identificarea regulilor optime de cultivare (Bozorg-Haddad & alţii, 2009) pentru exploatarea zonelor aride prin cultivarea de grâu, orz, porumb, sfeclă de zahăr, floarea soarelui, castraveţi, ceapă, cartofi, roşii, fasole, linte, lucernă şi peri sunt dintre cele mai recente comunicări în ceea ce priveşte utilizarea algoritmilor genetici.

Optimizarea sistemelor de producţie folosind un model de vegetaţie cu variabile independente pentru sistemele de producţie a salatei verzi în două medii de dezvoltare: în solar şi pe parcele, au

constituit subiectul cercetărilor în (Seginer & Ioslovich, 1999), când o serie de concluzii de importanță practică au fost obținute:

÷ Plantele de toate vârstele (situate în diferite stadii de dezvoltare) pot creşte împreună într-un singur compartiment climatizat;

÷ Spațierea trebuie planificată pentru a menține o densitate de plantare constantă;

÷ Densitatea optimă de plantare este o funcție crescătoare de cantitatea de lumină şi o funcție descrescătoare de temperatura disponibile;

÷ Dacă prețul de producție este mare în raport cu prețul de întreținere a suprafeței cultivate (în textul lucrării făcându-se aici referire la chirie) şi costul energetic, atunci intensitatea optimă de cultivare se înregistrează pentru o operare în solar în defavoarea operării pe parcele; opusul este adevărat când chiria este mare;

÷ Diferența de preț care se cere a fi plătită pentru suplimentarea iluminării este mică atunci când lumina naturală este mai intensă şi de durată.

Modelele de creştere pot fi folosite ca instrumente ale ***simulării*** pentru estimare cantitativă. Astfel, recent (Rodkaew & alții, 2004) s-a raportat un algoritm genetic care înglobează teoria matematică a lui Lindenmayer (Lindenmayer, 1968) pentru creşterea de soia pentru boabe.

Extinzând rezultate anterioare şi bazat pe măsurători experimentale întinse pe durata a doi ani de zile consecutivi, s-a elaborat (Salomez & Hofman, 2007) un model de creştere a salatei desfăcute (în engleză: Butterhead lettuce) care exprimă greutatea în funcție de schimbările de temperatură în sol şi radiațiile cu lungime de undă mică.

Simulatoare bazate pe algoritmi genetici au fost aplicate cu succes în predicția producției de alune supusă la contaminarea cu alfa-toxine (Henderson & alții, 2000), monitorizarea creşterii utilizând date obținute de la sateliți (Boken & alții, 2008), evaluarea efectului metalelor grele şi PCBs (bifenili policloruraţi) asupra pico-planctonului (fracțiunea din plancton compusă din celule cu diametrul între 0.2 şi 2 μm care pot fi deopotrivă fotosintetice şi heterotrofice) marin (Caroppo & alții, 2006), a deşeurilor militare asupra organismelor marine în (Jäntschi & Bolboacă, 2008-Marine), a toxicității de fenolii para-substituiți asupra *Tetrahymena pyriformis* (Jäntschi & alții, 2008-Tetrahymena), precum şi analiza asocierilor complexe între proprietățile solului şi abundența de ovăz sălbatic (Diaz & alții, 2005). Evenimentele rare cum este cazul temperaturilor extreme pot fi înglobate în modelele bazate pe algoritmi genetici care simulează creşterea plantelor, aşa cum se arată în lucrarea (Kysely & Dubrovsky, 2005).

Studii sistematice ale relațiilor care se stabilesc între fenotipuri şi proprietățile acestora au fost recent realizate la vinurile de masă pentru câteva componente ale acestora (Larsen & alții, 2006), la epistasisul plantelor cu autopolenizare (Cui & Wu, 2005), la activitatea hemoglutinativă a extracțiilor de *Curcuma aromatica* în raport cu identitatea secvenței putative (Tiptara & alții, 2008),

precum şi pentru genotipizarea *Ficus carica L* (Masi & alţii, 2005).

O abordare recentă (Letort & alţii, 2008) reţine atenţia realizând predicţia trăsăturilor fenotipice sub diferite condiţii de mediu în vederea elaborării strategiilor de înmulţire şi îmbunătăţirii trăsăturilor dorite.

Maşini capabile de învăţare bazate pe algoritmi genetici pot servi în ***clasificare***. Astfel, se raportează obţinerea de astfel de sisteme capabile de discriminarea automată a semințelor (Chtioui & alţii, 1996; Chtioui & alţii, 1997; Chtioui & alţii, 1998), ciupercilor (Hruschka & alţii, 2003), şi a imaginilor de plante stocate în baze de date (Zhu & alţii, 2008), precum şi pentru diferenţierea secvenţelor la genomii unor specii şi varietăţi de iarbă (Saski & alţii, 2007).

Nu în cele din urmă, algoritmii genetici îşi găsesc utilizarea în probleme de decizie, clasificare, optimizare şi simulare pentru resursele naturale aşa cum rezultă din cercetările care au fost realizate care sunt menţionate în continuare.

Astfel, decizia este subiectul abordat pentru construcţia politicilor strategice energetice în (Dagdeviren & Eraslan, 2008), clasificarea la forme de relief în (Moore & alţii, 2003), pentru datare geologică bazată pe raţia uraniu/plumb în (Lundmark & alţii, 2007), în timp ce optimizarea sistemelor de asigurare a resursei energetice în horticultură este subiectul cercetărilor raportate în (Husmann & Tantau, 2001), optimizarea tratamentului termic la fructe în (Morimoto & alţii, 1997), şi managementul resurselor de apă în (Chen, 1997). Simularea serveşte pentru predicţia potenţialului solar (Bălan & alţii, 2008; Sirdas & Sahin, 2008), precum şi al potenţialului resurselor de apă (Anandhi & alţii, 2008; Chen & alţii, 2008).

Identificarea seturilor de resurse naturale care maximizează reprezentarea diversităţii regionale şi menţinerea pe termen lung a biodiversităţii (Cabeza & Moilanen, 2001), precum şi rolul schimbărilor climatice în modelarea studiilor de impact (Fowler & alţii, 2007) sunt alte preocupări de actualitate care au fost abordate cu ajutorul algoritmilor genetici.

Variante, adaptări şi alternative ale formalismului algoritmilor genetici

Există multe variante şi adaptări ale algoritmilor genetici menite să îmbunătăţească performanţele acestora pentru un anume tip de probleme. Menţionarea tehnicilor derivate şi/sau bazate pe tehnica algoritmilor genetici este suficientă pentru problematica abordată:

÷ Optimizarea bazată pe strategia coloniilor de furnici (în engleză: Ant colony optimization) - (Bouktir & Slimani, 2005);

÷ Algoritmi bacteriologici (în engleză: Bacteriologic algorithms) - (Benoit & alţii, 2005);

÷ Metoda entropiei încrucişării (în engleză: cross-entropy method) - (Boer & alţii, 2005);

÷ Algoritmi culturali (în engleză: Cultural algorithms) - (Kobti & alţii, 2004);

÷ Strategii evolutive (în engleză: Evolution strategies) - (Schwefel, 1995);

÷ Programare evolutivă (în engleză: Evolutionary programming) - (Fogel & alții, 1966);

÷ Optimizare extremistă (în engleză: Extremal optimization) - (Bak & Sneppen, 1993);

÷ Adaptare Gausiană (în engleză: Gaussian adaptation) - (Kjellström, 1991);

÷ Programare genetică (în engleză: Genetic programming) - (Banzhaf & alții, 1997);

÷ Algoritmi memetici (în engleză: Memetic algorithm) - (Smith, 2007);

÷ Alte variate, colectate în (Davis, 1991).

Alte abordări conjugă algoritmii genetici cu alte concepte. Următoarele se pot menționa:

÷ Utilizarea maşinilor cu suport vectorial (în engleză: Support Vector Machines) - (Brown & alții, 2000);

÷ Analiza de localizare a asemănărilor structurale prin histograme secvențiale (acronim în engleză: SPLASH) - (Califano, 2000);

÷ Setul neregulat (în engleză: Rough set) - (Hvidsten & alții, 2001).

Optimizare vs. simulare: studii argumentative

Problema de optimizare aleasă în studiu, şi anume a relațiilor structură-activitate este în specialitatea pregătirii, situându-se la interdisciplinaritatea chimie - informatică - biologie. Dezvoltarea continuă a depozitelor de cunoştințe de genul celor administrate de NIH (National Institute of Health, USA), cum sunt PubMed, PubChem, Genome, etc accentuează necesitatea de a poseda instrumente eficace de a relaționa aceste cunoştințe, iar relațiile structură-activitate reprezintă unul dintre aceste instrumente.

Problema de simulare aleasă în studiu, şi anume a evoluției (prin intermediul diferiților parametrii ce caracterizează eşantionul supus evoluției) este o problemă insuficient explorată în literatura de specialitate (şi anume literatura al cărui subiect sunt algoritmii genetici).

Studii asupra altor operatorilor esențiali pentru evoluție se regăsesc în literatura de specialitate şi sunt axate pe eficiența algoritmică (văzută din punctul de vedere al vitezei cu care se atinge obiectivul şi al apropierii de maximul global al optimizării). O colecție de lucrări de acest tip este reprezentativă în acest sens (Martin & Spears, 2001). Astfel subiectul îl constituie diferiții operatori de încrucişare în (Prügel-Bennett, 2001), mutația şi încrucişarea în (Spears, 2001), alți parametrii dinamici în (Droste & alții, 2001).

Prea frecvent studiile îşi concentrează atenția spre rezolvarea problemei dificile cu ajutorul algoritmului genetic, câteodată discutând şi eficiența acestuia (timp de execuție, resurse de memorie necesare) şi foarte puțin spre influența diferitei strategii de evoluție asupra obiectivului urmărit (şi aici în special din nou asupra eficienței algoritmului) şi aproape niciodată asupra altor parametrii ce caracterizează eşantionul supus evoluției.

Mai rămâne de motivat legătura optimizare → simulare. Şi aici o căutare sistematică în literatura de specialitate nu este capabilă să producă prea multe rezultate evidente şi notabile, o singură referință către o monografie putând fi remarcată (Stender & alții, 1994). Literatura de specialitate este din nou mult mai bogată însă pe calea inversă de la simulare către optimizare.

Cadrul cercetării: Teze de doctorat susținute pe subiectul algoritmi genetici

O serie de teze de doctorat au fost susținute având ca subiect algoritmii genetici, în toate domeniile de cercetare şi preocupările au vizat deopotrivă aspectul fundamental şi aplicativ.

O serie de cercetări doctorale sub aspect fundamental îşi au ca punct de pornire teza (Jong & Holland, 1975) susținută sub îndrumarea unuia dintre părinții algoritmilor genetici moderni [4]. Axată pe probleme de optimizare, lucrarea (Jong & Holland, 1975) analizează eficiența unui algoritm genetic pentru câteva probleme clasice de optimizare (Figura 2-10), în care este cunoscut din literatura de specialitate că algoritmii clasici de optimizare eşuează frecvent. Seria de cercetări doctorale mai cuprinde studiul algoritmilor genetici clasici - şi aici rolul mutației şi recombinării e subiectul cercetării în (Spears & Jong, 1998) - pe fond unul dintre foarte puținele studii similare cu tema propriei cercetări (în care se abordează rolul selecției şi supraviețuirii), al unei forme modificate - co-evoluție cooperativă (Potter & Jong, 1997; Wiegand & Jong, 2003), şi respectiv al cazului în care populația inițială este împărțită în sub-populații numite insule şi evoluția se produce pe fiecare insulă, permițând însă migrația indivizilor de pe o insulă pe alta (Skolicki & Jong, 2007).

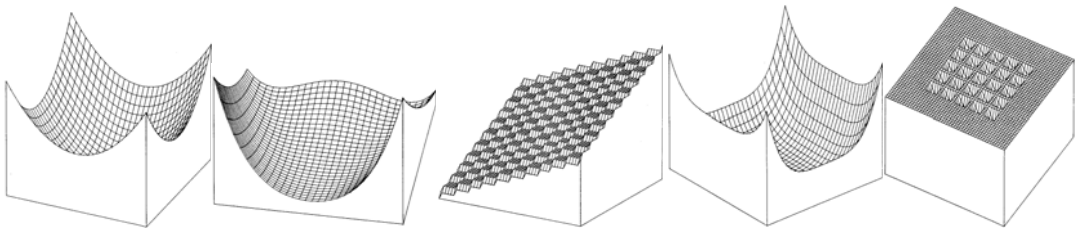


Fig. 2-10. Reprezentarea 3D a funcțiilor de test F1-F5 pentru comportamentul algoritmului genetic în lucrarea (Jong & Holland, 1975)
Representation of 3D test functions from F1 to F5 used for assessing of the genetic algorithm in (Jong & Holland, 1975)'s work.

Un rezultat valoros al (Spears & Jong, 1998), valorificat ceva mai târziu în (Spears, 2001) este ilustrat în Figura 2-11 şi referă efectul ratei de mutație asupra evoluției către echilibru.

---

[4] John Henry HOLLAND (n. 2 Februarie 1929) este un om de ştiință american, profesor de Psihologie, profesor de Inginerie Electrică şi Ştiința Calculatoarelor la University of Michigan, Ann Arbor; el este pionier în ştiința sistemelor complexe şi neliniare; este cunoscut ca unul din părinții algoritmilor genetici. http://en.wikipedia.org/wiki/John_Henry_Holland

Theorem 1 *Let S be any string of L alleles: $(a_1, ..., a_L)$. If a population is mutated repeatedly (without selection or recombination) then:*

$$\lim_{t\to\infty} p_S(t) = \prod_{i=1}^{L} \frac{1}{C}$$

where $p_S(t)$ is the expected proportion of string S in the population at time t and C is the cardinality of the alphabet.

Theorem 2 *Let S be any string of L alleles: $(a_1, ..., a_L)$. If a population is recombined repeatedly (without selection or mutation) then:*

$$\lim_{t\to\infty} p_S(t) = \prod_{i=1}^{L} p_{a_i}(0)$$

where $p_S(t)$ is the expected proportion of string S in the population at time t and $p_{a_i}(0)$ is the proportion of allele a at locus (position) i in the initial population.

Theorem 3 *Let S be any string of L alleles: $(a_1, ..., a_L)$. If a population is mutated and recombined repeatedly (without selection) then:*

$$\lim_{t\to\infty} p_S(t) = \prod_{i=1}^{L} \frac{1}{C}$$

where $p_S(t)$ is the expected proportion of string S in the population at time t and C is the cardinality of the alphabet.

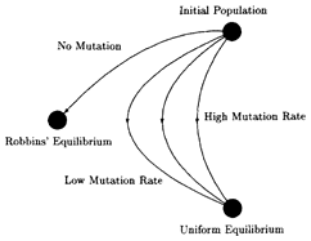


Fig. 2-11. Efectul mutației asupra evoluției către echilibru în lucrarea (Spears, 2001)
The effect of the mutation on evolution to the equlibrium in (Spears, 2001)'s work.

În Figura 2-11 echilibrul Robbins referă o populație care este recombinată fără selecție şi mutație (Robbins, 1918), iar echilibrul uniform referă o populație care este mutată repetat fără selecție şi încrucişare care duce într-un punct în care valoarea funcției obiectiv se anulează (Solan, 2009).

Alte cercetări doctorale axate pe componenta de cercetare fundamentală cuprind diferite adaptări ale algoritmilor genetici făcute pentru a rezolva categorii speciale de probleme dificile, reținând atenția: optimizarea multi-obiectiv (Zitzler & Thiele, 1999), optimizarea Pareto (Knowles & Corne, 2002), şi sistemele multi-agent (Panait & Luke, 2006).

Algoritmii genetici au depăşit granițele domeniului informatică datorită potențialui de valorificare al rezultatelor.

Teze de doctorat având ca obiectiv proiectarea de algoritmi genetici, implementarea de programe evolutive şi realizarea de studii având la bază pe aceştia se regăsesc practic în toate domeniile de cercetare. În continuare se redau câteva exemple reprezentative în acest sens.

În domeniul agricultură şi-au găsit utilizarea la planificarea culturilor (Matthews & Kraw, 2001), în construcții la evaluarea riscului de eroziune a solului (Osman & McManus, 2007), în bioinginerie la controlul eficient al poluării la nivelul unui bazin hidrografic (Veith & Wolfe, 2002), în chimie la designul proceselor controlate senzorial (Dai & Lodder, 2007), în economie la probleme de optimizare cu opțiuni multiple (Aickelin & Dowsland, 1999), în management la modelarea proceselor multi-scală (Sastry & alții, 2007), în mecanică la optimizarea structurilor compozite (Gantovnik & Gürdal, 2005) şi în mediu la alegerea strategiei pentru controlul calității apei (Tufail & Ormsbee, 2006).

În cele din urmă, dar nu în ultimul rând, în domeniul biologie, două direcții se desprind în elaborarea şi utilizarea algoritmilor genetici: la probleme de evoluție (Suzuki & Iwasa, 1998) şi la studiul filogenetic (Zwickl & Hills, 2006).

Pe partea pur aplicativă, de utilizare a algoritmilor genetici în domeniul agricol şi horticol, aplicațiile algoritmilor genetici vizează de la studii de creştere (Venard &

Vaillancourt, 2006), la clasificări taxonomice (Sarmiento-Monroy & Sharkey, 2006) şi la analiza diversității genetice (Zhang & Ghabrial, 2006).

Domeniul de aplicabilitate: Serii de compuşi biologic activi

Revenind asupra investigațiilor structurale pe compuşi biologic activi, o activitate foarte importantă este căutarea substructurilor active biologic din compuşii biologic activi care produc cea mai mare parte a răspunsului biologic măsurat (Kellogg & alții, 1991).

Bifenilii policlorurați (PCBs - PolyChlorinated Biphenils) sunt o familie de compuşi cu potență biologică distructivă asupra mediului vegetal şi animal intens studiată şi cu o bogată documentație în literatura de specialitate, şi din acest punct de vedere încrederea în rezultatele experimentale folosite în modelare este maximă; mai mult, este unul dintre puținele seturi de date numeroase (conține 209 compuşi) de activitate biologică a compuşilor chimici care reprezintă într-adevăr o familie de compuşi, atât sub aspectul structurii (toți compuşii au o structură de schelet organic identică) cât şi sub aspectul activității biologice alese (aici impunerea este şi mai drastică, şi anume să se verifice ipoteza de distribuție normală a valorilor experimentale observate - analiză redată mai jos).

O scurtă incursiune în istoria acestor compuşi ne arată că aceştia şi-au găsit utilizări sub formă de lichide dielectrice în transformatoare şi condensatoare, ca agenți de răcire şi ungere, aditivi stabilizatori la lipiturile de țevi PVC pentru rețelele electrice, în componente electronice, ca agenți de diluare ale pesticidelor, uleiuri de răcire pentru scule de tăiere, fluide hidraulice, agenți de protecție şi lustruire, adezivi, lacuri (de parchet), vopsele, agenți de curățare şi chiar ca alternativă la hârtia de celuloză [5].

Chiar dacă producția de PCB a fost stopată în 1970 datorită toxicității ridicate a celor mai mulți congeneri ai familiei PCB, efectele acestora sunt încă prezente în mediu, datorită faptului că PCBs sunt poluanți organici persistenți - clasificați ca atare [6] care se acumulează (bioacumulare) în animale.

Coeficientul de partiție octanol/apă (K_{ow}) este raportul concentrațiilor unui compus chimic între octanol şi apă aflate în contact la o anumită temperatură. Este un parametru adimensional (fiind un raport de concentrații) care frecvent se exprimă pe scară logaritmică ($\log K_{ow}$ sau mai simplu lkow). Această proprietate fizico-chimică este utilizată în multe studii de mediu în determinarea efectului compuşilor chimici în mediu, un exemplu fiind utilizarea acestuia pentru prezicerea

[5] UNEP Chemicals. 1999. Guidelines for the Identification of PCBs and Materials Containing PCBs. United Nations Environment Programme. http://www.chem.unep.ch/pops/pdf/PCBident/pcbid1.pdf
[6] Office of International Affairs. 2002. Persistent Organic Pollutants: A Global Issue, A Global Response. Environmental Protection Agency. http://www.epa.gov/oia/toxics/pop.pdf

magnitudinii de bioacumulare în peşti [7]. Valorile activității măsurate se bazează pe un studiu datat în 1996 (Eisler & Belisle, 1996) care sintetizează rezultatele obținute şi raportate în literatura de specialitate în studiul PCBs de mai mulți autori.

Structura chimică a bifenililor policlorurați se poate schematiza ca în Figura 2-12.

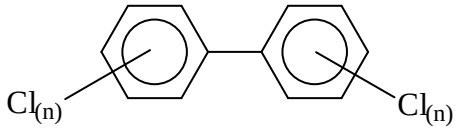


Fig. 2-12. PCBs: Seria bifenililor policlorurați (structuri & activități în Anexa 2-2)
PCBs: Series of Polichlorinated Biphenils (structures and activities in Annex 2-2)

Măsurătorile experimentale ale coeficientului de partiție octanol/apă ale unei serii de 206 PCBs (din totalul de 209 reprezentate generic în Figura 2-13), exprimate în scară logaritmică - lkow = ln(K_{OW}) - aşa cum au fost ele raportate în (Eisler & Belisle, 1996) sunt redate în Anexa 2-2.

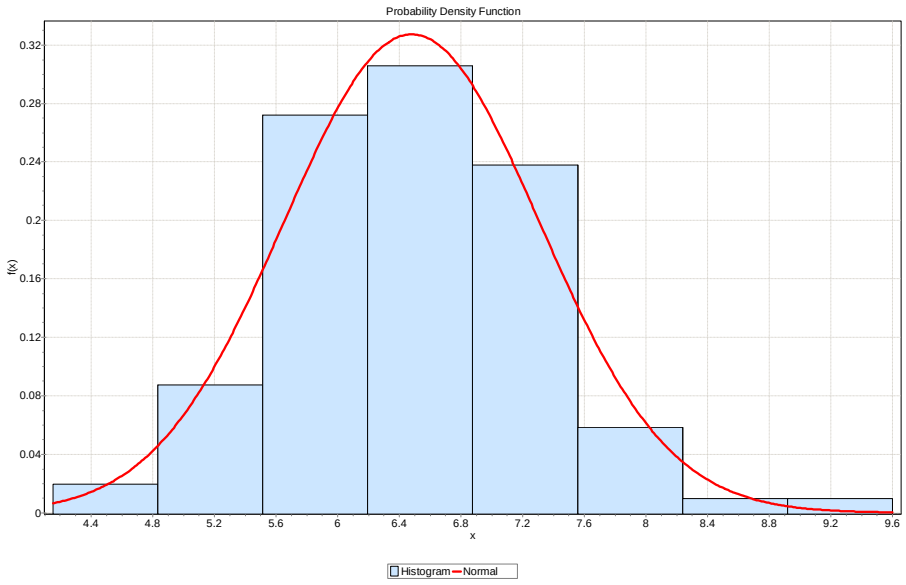


Fig. 2-13. Histograma valorilor observate ale activității biologice măsurate pentru PCBs
The hisogram of observed values for measured biological activities of PCBs

Ceea ce este esențial pentru experimentul derulat de analiză structură-activitate prin intermediul algoritmilor genetici este distribuția observată a valorilor experimentale. Din acest punct de vedere verificarea statistică a distribuției normale acestora oferă garanția că setul de date ales este potrivit pentru analiza de regresie. În acest sens, Figura 2-13 prezintă histograma valorilor observate ale activității măsurate pentru cei 206 compuşi. Analiza statistică a distribuției observate presupune folosirea a cel puțin unui test statistic pentru verificarea ipotezei de normalitate (de preferat este folosirea şi validarea prin intermediul mai multora). În selecția testelor statistice care oferă cea mai mare încredere în rezultat, este esențială compararea acestora pe o multitudine de seturi de date. Astfel, o lucrare recentă (Steele & alții, 2005) a arătat că cele mai potrivite teste statistice sunt

[7] U.S. Geological Survey. 2008. Octanol-Water Partition Coefficient (K_{OW}). U.S. Department of the Interior. http://toxics.usgs.gov/definitions/kow.html

Pearson-Fisher Chi Square χ^2 (CS) - (Pearson, 1900; Fisher, 1922-X2; Fisher, 1924; Fisher, 1935) şi Anderson-Darling (AD) - (Anderson & Darling, 1952; Scholz & Stephens, 1986 & 1987), fiecare dintre acestea obținând performanțe superioare celeilalte pentru diferite cazuri de legi de distribuție, dar amândouă fiind superioare celorlalte teste statistice, şi foarte aproape de acestea se situează Kolmogorov-Smirnov (KS) - (Kolmogorov, 1941; Smirnov, 1948). Testul Jarque-Bera este frecvent folosit pentru verificarea depărtării de la normalitate (Jarque & Bera, 1980 & 1981). Descrierea acestor teste statistice este redată în Anexa 3-3.

În Tabelul 2-4 sunt redate o serie de statistici aplicate pentru verificarea ipotezei de normalitate pe seria de 206 activități biologice observate.

Tabelul 2-4

Ipoteza de normalitate a activității biologice măsurate pentru PCBs
Hypothesis of normality of measured biological activity for PCBs

Statistică *Statistic*	Valoare *Value*	Probabilitate de observație *Probability of observation*	Concluzie *Conclusion*
KS	0.03348	96.91%	Nu se respinge ipoteza de normalitate
AD	0.44432	27.2%; 25.2%; 19.2%	Nu se respinge ipoteza de normalitate
CS	11(df=7)	13.8%	Nu se respinge ipoteza de normalitate
WS	0.98709	5.8%	Nu se respinge ipoteza de normalitate
$Z_{Skewness}$	1.48	14%	Nu se respinge ipoteza de normalitate
$Z_{Kurtosis}$	2.51	1.2%	Se respinge ipoteza de normalitate
JB	7.577	2.3%	Se respinge ipoteza de normalitate
Nota: testul JB nu este afectat de valori pereche (tied); KS şi AD sunt afectate			

Pentru comparație s-a ales setul de compuşi farmaceutici din studiul (Duchowicz & alții, 2008) a căror histogramă a valorilor măsurate ale solubilității sunt redate în Figura 2-14, iar analiza de normalitate similară cu cea din Tabelul 2-4 este redată în Tabelul 2-5.

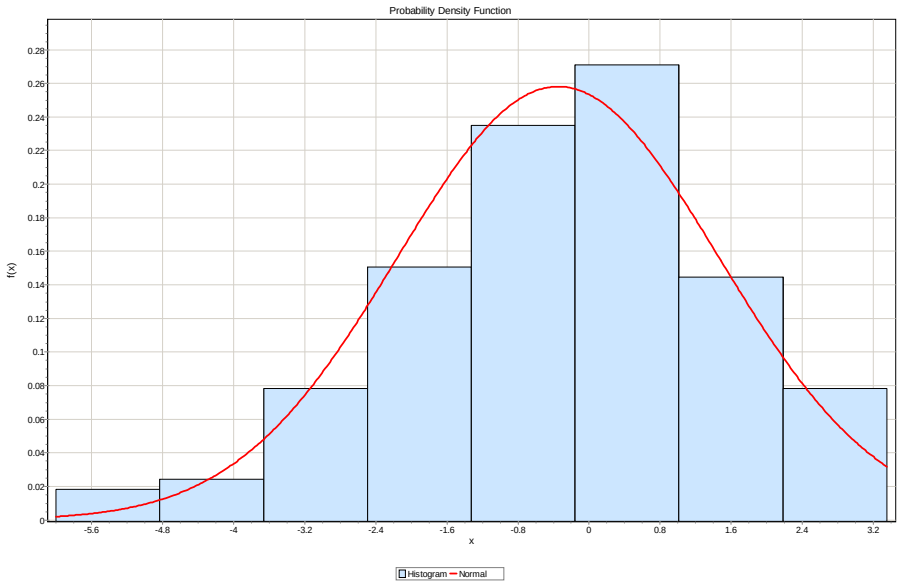


Fig. 2-14. Histograma valorilor observate ale activității în (Duchowicz & alții, 2008)
The histogram of observed activities in (Duchowicz & others, 2008)

Tabelul 2-5

Ipoteza de normalitate a activității biologice măsurate în (Duchowicz & alții, 2008)
Hypothesis of normality of measured biological activity from (Duchowicz & others, 2008)

Statistică *Statistic*	Valoare *Value*	Probabilitate de observație *Probability of observation*	Concluzie *Conclusion*
KS	0.05508	67.43%	Nu se respinge ipoteza de normalitate
AD	0.56539	14.1%; 12.5%; 14.3%	Nu se respinge ipoteza de normalitate
CS	3(df=7)	88.6%	Nu se respinge ipoteza de normalitate
WS	0.98173	2.8%	Se respinge ipoteza de normalitate
$Z_{Skewness}$	-2.58	1‰	Se respinge ipoteza de normalitate
$Z_{Kurtosis}$	0.53	59.5%	Nu se respinge ipoteza de normalitate
JB	6.61	3.7%	Se respinge ipoteza de normalitate
Nota: testul JB nu este afectat de valori pereche (tied); KS și AD sunt afectate			

Analiza din Tabelele 2-4 şi 2-5 arată că în ambele cazuri eşantioanele conțin outlieri. Testul Grubbs poate fi folosit pentru identificarea acestor outlieri (Grubbs, 1969), această analiză regăsindu-se în (Jäntschi, 2009-DistFit1). Pentru datele din Tabelul 2-4 outlierii identificați sunt:

÷ 9.603 (5%);

÷ Y\{9.603}: No outliers (5%);

iar pentru datele din Tabelul 2-5 outlierii identificați sunt:

÷ No outliers (5%).

Eliminând outlierul (9.603) rezultatele din Tabelul 2-4 se schimbă dramatic (Tabelul 2-6).

Tabelul 2-6. Ipoteza de normalitate a activității biologice măsurate ($\mu = 6.4653$; $\sigma = 0.80344$)

Statistică	Valoare	Probabilitate de observație	Concluzie
KS	0.03579	94.68%	Nu se respinge ipoteza de normalitate
AD	0.37878	40.3%; 39.5%; 21.0%	Nu se respinge ipoteza de normalitate
CS	8.64(df=7)	27.9%	Nu se respinge ipoteza de normalitate
WS	0.98709	47.8%	Nu se respinge ipoteza de normalitate
$Z_{Skewness}$	1.48	79.2%	Nu se respinge ipoteza de normalitate
$Z_{Kurtosis}$	2.51	41.5%	Nu se respinge ipoteza de normalitate
JB	0.56146	75.5%	Nu se respinge ipoteza de normalitate

Se desprind o serie de remarci în urma analizei de distribuție normală a datelor (Tabelele 2-4, 2-5 & 2-6; astfel cel mai puțin afectat de outlieri este KS (variație procentuală pozitivă, sub 2%); outlierii aduc în KS erori de tipul II (se acceptă ipoteza nulă chiar dacă ea nu este adevărată); îl urmează testul AD (variație procentuală de 10%); outlierii aduc în AD erori de tipul I (se respinge ipoteza nulă chiar dacă ea este adevărată); Celelalte teste sunt din ce în ce mai puternic afectate de outlieri în ordinea Chi-Square (100%), Z-Skewness (470%), Jarque-Bera (3200%), Z-Kurtosis (3400%).

SCOPUL & OBIECTIVELE CERCETĂRII

3. Simularea evoluției cu algoritmi genetici în probleme de optimizare a relațiilor structură-activitate

structură de lucru

÷ ***Scop***

- Proiectarea, implementarea şi analiza statistică a influenței diferitelor metode de selecție şi de supraviețuire asupra evoluției unui algoritm genetic utilizat pentru obținerea de relații structură-activitate în serii de compuşi chimici biologic activi.

÷ ***Obiective***

a. Elaborarea algoritmului genetic:

i. Definirea problemei dificile (cu complexitate exponențială);

ii. Formularea problemei în termeni genetici;

iii. Proiectarea algoritmului genetic;

iv. Implementarea şi documentarea programului evolutiv bazat pe algoritmul genetic;

b. Simularea evoluției:

i. Evoluție: definirea observabilelor;

ii. Definirea contingenței selecție vs. supraviețuire;

iii. Proiectarea experimentului statistic;

iv. Realizarea experimentului statistic;

c. Analiza şi interpretarea rezultatelor:

i. Observabile calitative;

ii. Obiectivul evoluției - r2 - observabilă cantitativă;

4. Definirea problemei de optimizare a relației structură-activitate

Seria de compuşi incluşi în studiu: Bifenili policloruraţi (PCBs)

Măsurătorile experimentale ale coeficientului de partiție octanol/apă ale seriei de 206 PCBs (din totalul de 209), exprimate în scară logaritmică - lkow = ln(K_{OW}) - aşa cum au fost ele raportate în (Eisler & Belisle, 1996) au fost folosite pentru optimizarea relației structură-activitate.

Un fişier (text) conținând valorile experimentale observate ale klow pentru 206 PCBs reprezintă date de intrare în analiza structură-activitate realizată (vezi şi Anexa 2-2).

Modelare moleculară: Optimizare geometrie

Modelarea structurii moleculare este condiția obligatorie pentru o analiză structură-activitate. Realizarea unui model tridimensional (3D) se impune pentru cazurile când descriptorii calculați uzează de geometria moleculară, ceea ce este cazul aici.

Obținerea modelului 3D se poate realiza folosind un program de modelare moleculară cum ar fi: HyperChem, Spartan, Gaussian, Molecular Modelling Pro, Mopac.

În modelarea structurii PCBs s-a folosit programul de modelare moleculară HyperChem (licență v. 8.0/2007).

Obținerea informației de structură 3D presupune parcurgerea unei serii de paşi, care cuprinde definirea unui model de mecanică moleculară (a fost definit AMBER [1]), optimizarea geometriei moleculare până la convergență folosind un algoritm de optimizare (a fost folosit POLAK-RIBIERE [2]), definirea unei metode semiempirice de calcul energetic (a fost definită AM1 [3]) şi obținerea unei serii de parametrii energetici ce caracterizează optimul împreună cu sarcinile electrice parțiale.

Rezultatul optimizării şi calculului energetic este salvat în fişierul ce defineşte structura chimică a moleculei al cărui model a fost construit. În cazul HyperChem, fişierele sunt cu extensia *.hin, având o structură cu mai multe linii, câte o linie pentru fiecare atom al moleculei cuprinzând numărul, numele, sarcina parțială, coordonatele spațiale, numărul de legături pe care le formează şi lista legăturilor împreună cu tipul acestora.

---

[1] AMBER (acronim pentru *Assisted Model Building with Energy Refinement* - este o familie de câmpuri de forță pentru dinamica moleculară a biomoleculelor dezvoltat de succesorii grupului lui Peter Kollman la University of California, San Francisco. http://ambermd.org/

[2] Polak-Ribiere este o metodă de optimizare ce foloseşte gradienți conjugați (Polak B, Ribiere G, 1969. Note surla convergence des méthodes de directions conjuguées. *Rev. Fr. Imform. Rech. Oper* 16:35-43).

[3] AM1: Austin Model 1 (dezvoltată de Dewar & alții) în mod similar cu metoda tradițională MNDO de aproximare a integralelor de doi electroni dar foloseşte o expresie modificată pentru repulsia nucleu-nucleu (Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP, 1985. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J Am Chem Soc* 107:3902-3909).

Aceste fişiere (pentru fiecare moleculă în parte) reprezintă datele de intrare în analiza de structură realizată.

Uzual metodologia de analiză structură-activitate foloseşte structurile moleculare în care atomii de hidrogen sunt neglijați (şterşi). O sumă de motive stau la baza acestei alegeri:

÷ activitățile biologice în mod uzual determinate in vivo au ca mediu de analiză mediul apos, în care au loc procese de disociere (parțială) în care atomii de hidrogen trec sub formă de protoni în soluție, părăsind locul pe care îl ocupă în structura moleculară, deci o formă depletizată de hidrogen îşi găseşte corespondentul în starea de fapt a experimentului;

÷ atomii de hidrogen pot forma o singură legătură; aceştia sunt singurii şterşi, deci exceptând poziția geometrică a acestora informația poate fi oricând reconstruită; mai mult, fiindcă formează o singură legătură, ei nu contribuie la complexitatea moleculară (nu creează lanțuri şi ramificații, sunt doar terminatori de structură) şi din acest punct de vedere nu aduc informație care să caracterizeze suplimentar o structură în raport cu informația care a fost deja produsă de ceilalți atomi;

÷ nu în ultimul rând se reduce considerabil volumul de calcule pe o structură fără atomi de hidrogen; considerând doar un alcan cu formula generală C_nH_{2n+2} se observă că volumul de calcule se reduce la aproape o treime (în ipoteza ideală că volumul de calcule e proporțional cu dimensiunea moleculei, ceea ce de cele mai multe ori este nerealist, reducerea fiind şi mai drastică, existând dependențe pătratice sau chiar exponențiale în timp de execuție cu volumul molecular) dacă atomii de hidrogen sunt eliminați.

Seria de fişiere (cu extensia `hin`) conținând modelele structurilor 3D ale moleculelor din setul de 206 PCBs reprezintă date de intrare în analiza structură-activitate realizată (vezi şi Anexa 2-2).

Familia de descriptori: MDF

Pentru ca problema de optimizare aleasă să poată fi asociată unei probleme de evoluție cu ajutorul unui algoritm genetic, este obligatoriu ca descriptorii moleculari să provină dintro populație care posedă un cod genetic (de aici denumirea de familie). Cu alte cuvinte, posibilitatea de a asocia un cod genetic populației de descriptori este cea care creează oportunitatea execuției unui algoritm genetic.

Elaborarea algoritmului genetic s-a făcut având ca model patru familii de descriptori de concepție proprie, redați în Tabelul 4-1.

Tabelul 4-1

Familii de descriptori moleculari (vezi şi Anexa 4)
Families of molecular descriptors (see Annex 4 too)

Familie (*Family*)	©	Autori (*Authors*)
FPIF (Fragmental Property Index Family)	1999-2000	Jäntschi L & Diudea MV
MDF (Molecular Descriptors Family)	2004-2005	Jäntschi L
MDFV (Molecular Descriptors Family Vertex)	2008-2009	Bolboacă SD & Jäntschi L
SAPF (Structural Atomic Property Family)	2010-2011	Bolboacă SD & Jäntschi L

FPIF (Jäntschi & Diudea, 2000) este o metodă matriceală bazată pe matrice pătratice de proprietăți derivate din structură. MDF (Jäntschi, 2004) este o metodă bazată pe fragmente moleculare obținute din perechi de vârfuri. MDFV (Bolboacă & Jäntschi, 2009-Comparison) este o metodă bazată pe vârfuri în locul perechilor de vârfuri, în timp ce SAPF (metodă încă nedocumentată) este o metodă cumulativă de proprietăți atomice la nivel molecular. Descrierea detaliată a familiilor de descriptori moleculari este redată în Anexa 4.

Obținerea unei familii de descriptori de structură presupune execuția unor programe. Familiile de descriptori din Tabelul 3-4 se generează prin execuția unor programe de concepție proprie implementate în limbajul PHP. Atunci când se stochează familia de descriptori într-o bază de date (cazul pentru MDF şi MDFV) întâi se construieşte structura unei constelații de tabele într-o bază de date, tabele care ulterior sunt depozitarul pentru activitate/proprietate, familia de descriptori, şi forma linearizată a acesteia.

Pentru analiza structură-activitate implicând PCBs şi activitatea măsurată a acestora lkow s-a ales familia MDF (vezi şi Anexa 4-2) să definească informația structurală în asocierea cu activitatea biologică, această familie fiind cea mai bine documentată şi ale cărei rezultate sunt cel mai bine susținute (peste 50 de seturi de compuşi au fost analizați cu ajutorul acesteia).

Relația structură-activitate: QSAR 4v cu MDF

Studiul anterior al lkow cu familia de descriptori moleculari (Jäntschi & Bolboacă, 2006-MDF-SAR-6) a arătat că se poate obține o ecuație de regresie liniară multiplă în 4 variabile care să explice activitatea măsurată în procent de 91%, ecuație care însă nu respectă toate condițiile impuse de viabilitate fenotipică (variabilitate, abatere de la normalitate, şi determinare rezonabile - impuneri definite în secțiunea de material şi metodă pentru algoritmul genetic implementat), în fapt 2 dintre genotipurile acestei ecuații eşuând la acest test de viabilitate: *HMmjQt* şi *SMMjQg*.

Oricum, ecuația obținută în (Jäntschi & Bolboacă, 2006-MDF-SAR-6) se constituie într-un element de referință în ceea ce priveşte performanța algoritmului genetic, şi din acest motiv este redată în continuare:

$$\hat{Y} = 3.04 - 0.42 \cdot IIDDKGg + 0.04 \cdot IHDRKEg + 0.07 \cdot aHMmjQt - 37.5 \cdot aSMMjQg$$
$$r^2(Y,\hat{Y}) \approx 0.91, F=554$$

Căutarea relației de mai sus a fost realizată folosind un euristic care a căutat asocieri de perechi de câte doi descriptori. Motivul este bineînțeles, complexitatea calculului reducând-se în mod dramatic în acest caz (complexitatea calculului fiind în acest caz proporțională cu dublul puterii a 2-a a volumului populației) față de cazul în care s-ar fi căutat asocieri de fiecare cu fiecare (complexitatea calculului fiind în acest caz proporțională cu puterea a 4-a a volumului populației). Chiar şi în cazul de mai sus, durata de execuție a unei căutări este de ordinul a câteva zile, valoarea exactă depinzând de performanțele maşinii de calcul.

Problema de optimizare propusă spre rezolvare a fost găsirea unei cât mai bune relații structură-activitate care să caracterizeze activitatea biologică folosind o regresie liniară multiplă (MLR) - Anexa 3-4 - cu 4 descriptori membrii ai MDF - Anexa 4-2 - având la dispoziție un model (cel anterior) obținut drept referință.

Căutarea unei soluții folosind un euristic mai eficient este perfect legitimată pentru cazul de mai sus, când căutarea necesită relații cu un număr mare de descriptori (prin mare înțelegând aici un număr mai mare ca 2; în doi descriptori căutarea sistematică fiecare cu fiecare necesită câteva zile; în 3 descriptori deja câteva zeci de ani pentru un volum al populației de descriptori de ordinul zecilor de mii).

5. Definirea problemei genetice de evoluție

Spațiul de căutare: Relațiile structură-activitate MLR în 4 variabile cu MDF

Lucrarea (Jäntschi & alții, 2007-Chromatogr) face referire pentru prima dată la algoritmul genetic implementat şi astfel serveşte drept sursă a exemplificării problemei dificile de asociere între structura biochimică şi activitatea măsurată. Astfel, în lucrare autorii pornesc de la datele experimentale ale timpilor de retenție observați pentru separarea cromatografică a bifenililor policloruraţi (PCBs) raportați în (Mullins & alții, 1984). Folosind informația structurală, o bază de date (Figura 5-1) conținând descriptori de structură a fost construită.

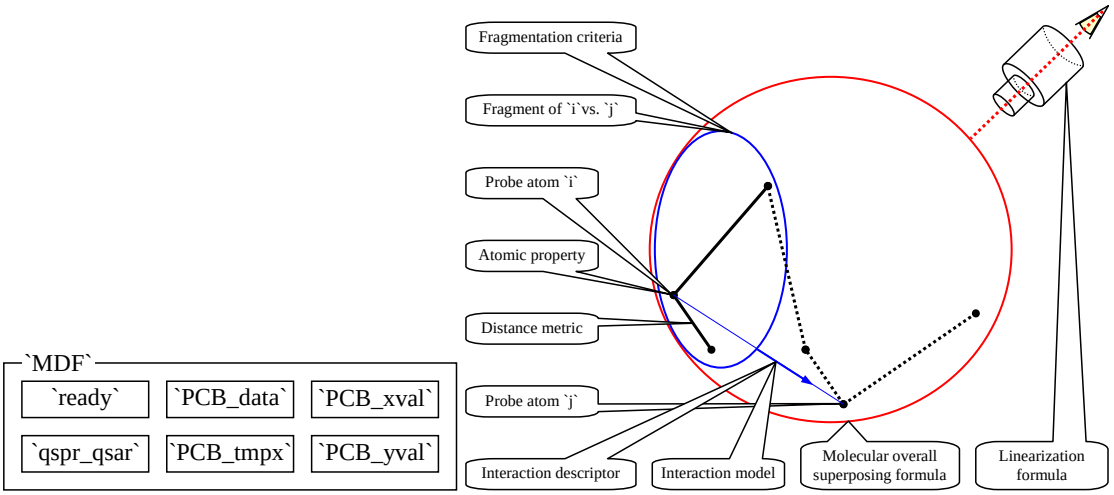


Fig. 5-1. Baza de date şi metodologia relațiilor structură-activitate (SAR) în abordarea Familiei de Descriptori Moleculari (MDF) - din (Jäntschi & Bolboacă, 2008-Posters)
Database and the methodology of structure-activity relationships (SAR) in the approach of Molecular Descriptors Family - from (Jäntschi & Bolboacă, 2008-Posters)

Problema dificilă de relație structură chimică - activitate/proprietate măsurată se formulează astfel: având la dispoziție o informație structurală (obținută din realizarea topologiei moleculare a fiecărui compus pe baza legăturilor chimice care se stabilesc între atomi şi realizarea geometriei moleculare pe baza aplicării modelelor aproximative ale fizicii cuantice şi moleculare) şi o informație rezultată în urma observației experimentale a unei proprietăți/activități pentru fiecare din cei 209 (de data aceasta) compuşi în aceleaşi condiții de mediu, care este cea mai bună relație structură-activitate care să descrie proprietatea/activitatea compuşilor în funcție de structură.

În acord cu metodologia descrisă în (Jäntschi & alții, 2007-Chromatogr) informația structurală se poate valorifica construind o familie de descriptori. Abordarea din perspectiva unui algoritm genetic aplicat problemei este următoarea:

÷ Populația este familia de descriptori moleculari având fiecare genotip (cromozom) descris de următoarele gene (vezi şi Tabelul 5-2):
  o gena *d* conținând operatorul de distanță folosit la construirea descriptorului molecular

(fenotipului), având 2 valori posibile: `g` - distanță geometrică; `t` - distanță topologică;

- gena *p* conținând proprietatea atomică folosită la construirea fenotipului, cu 6 valori posibile: `M` - masa atomică relativă; `Q` - sarcina electrică parțială a atomului (obținută în aproximația Hückel extinsă); `C` - cardinalitatea (proprietate atomică trivială, totdeauna valoarea sa pentru orice atom este 1); `E` - electronegativitatea atomică (valoare relativă la scara de electronegativitate Sanderson); `G` - electronegativitatea de grup (valoare obținută prin calcularea unei medii geometrice a electronegativității grupului de atomi situat în imediata vecinătate a atomului cercetat); `H` - numărul de atomi de hidrogen ce înconjoară atomul cercetat;
- gena *I* conținând descriptorul de interacțiune, implicând 2 atomi participanți şi având una din următoarele 24 de valori (în care s-au folosit notațiile: d - operatorul de distanță; p - proprietatea atomică): $`D(d)`$, $`d(1/d)`$, $`O(p_1)`$, $`o(1/p_1)`$, $`P(p_1p_2)`$, $`p(1/p_1p_2)`$, $`Q(\sqrt{p_1p_2})`$, $`q(1/\sqrt{p_1p_2})`$, $`J(p_1d)`$, $`j(1/p_1d)`$, $`K(p_1p_2d)`$, $`k(1/p_1p_2d)`$, $`L(d\sqrt{p_1p_2})`$, $`l(1/d\sqrt{p_1p_2})`$, $`V(p_1/d)`$, $`E(p_1/d_2)`$, $`W(p_1^2/d)`$, $`w(p_1p_2/d)`$, $`F(p_1^2/d^2)`$, $`f(p_1p_2/d^2)`$, $`S(p_1^2/d^3)`$, $`s(p_1p_2/d^3)`$, $`T(p_1^2/d^4)`$, $`t(p_1p_2/d^4)`$, în care prima literă dă valoarea genei şi între paranteze este dată formula de calcul asociată valorii acesteia;
- gena *O* codificând suprapunerea interacțiunilor, fiind implementate 6 valori ale acesteia, două pentru modele în care interacțiunile sunt rare şi la distanță (`R` şi `r`), două pentru modelele în care interacțiunile sunt dense dar la distanță (`M` şi `m`), şi două pentru modelele în care interacțiunile sunt dense şi apropiate (`D` şi `d`);
- gena *f* care codifică algoritmul de fragmentare moleculară pe perechi de atomi, care poate avea una din următoarele 4 valori: `P` - pentru fragmentare bazată pe căi; `D` - pentru fragmentare bazată pe distanțe; `M` - pentru fragmentare în fragmente maximale; `m` - pentru fragmentare în fragmente minimale (fragmente triviale conținând un singur atom);
- gena *M* care codifică modalitatea de suprapunere globală a interacțiunilor fragmentelor, care poate lua una din următoarele 19 valori împărțite în 4 grupuri: grupul de mărimi (`*m*` - selectează cea mai mică valoare; `*M*` - cea mai mare valoare; `*n*` - cea mai mică valoare absolută; `N` - cea mai mare valoare absolută); grupul de medii (`*S*` - suma; `*A*` - media aritmetică după numărul de proprietăți de fragmente; `*a*` - media aritmetică după numărul de fragmente; `*B*` - media aritmetică după numărul de atomi; `*b*` - media aritmetică după numărul de legături); grupul geometric (`*P*` - multiplicare; `*G*` - medie geometrică după numărul de proprietăți de fragmente; `*g*` - medie geometrică după numărul de fragmente; `*F*` - medie geometrică după numărul de atomi; `*f*`` - media geometrică după numărul de legături); grupul armonic (`*s*` - suma armonică; `*H*` - medie armonică după numărul de proprietăți de fragmente; `*h*` - medie armonică după numărul de fragmente; `*I*` - medie

armonică după numărul de atomi; `*i*`` - media armonică după numărul de legături);

- gena L care rezultă în algoritm în urma unei mutații deterministe asupra fenotipului şi reprezintă unul din cei 6 operatori de linearizare ce urmează: `I` - identitate, `i` - inversa, `A` - valoarea absolută, `a` - inversa valorii absolute, `L` - logaritmul valorii absolute, `l` - logaritmul;

÷ Scorul (funcția obiectiv): având proprietatea măsurată pentru cei 209 compuşi exprimată în şirul Y (cu 209 valori), urmează:

- Fenotipul, caracterizat de genele care compun cromozomul este transformat în fenotip când se calculează valorile corespunzătoare genotipului pentru fiecare din cele 209 molecule ale setului de PCBs.
- Pentru o funcție obiectiv de estimare a proprietății măsurate folosind 2 genotipuri (în cazul general putând fi un număr oarecare fixat de genotipuri, r), ecuația de regresie este de forma: $\hat{Y} = a_0 + a_1 \cdot Fenotip_1 + a_2 \cdot Fenotip_2$, (în general: $\hat{Y} = a_0 + \Sigma_{1 \le i \le r} a_i \cdot Fenotip_i$), unde:
  - $\hat{Y}$ este estimatorul proprietății măsurate;
  - $a_0..a_2$ sunt coeficienții care rezultă din minimizarea pătratelor diferențelor între valorile măsurate Y şi cele estimate prin intermediul lui $\hat{Y}$ pentru setul de 209 compuşi;
  - $0Fenotip_1=Fenotip(Genotip_1)$ şi $Fenotip_2=Fenotip(Genotip_2)$ sunt fenotipurile (şirurile de valori ale descriptorilor moleculari obținuți prin operațiile prescrise de genotipurile $Genotip_1=d_1p_1I_1O_1f_1M_1L_1$ şi $Genotip_2=d_2p_2I_2O_2f_2M_2L_2$ pentru cei 209 compuşi;
- Tăria (scorul) unei perechi de genotipuri (sau mai multe, pentru regresia multivariată) poate fi exprimată de una din funcțiile de scor date în Tabelul 5-1.

Tabelul 5-1

Funcții de scor pentru penalizarea regresiilor cu perechi de fenotipuri
Score functions for penalizing the regressions with pairs of descriptors

Nr.	Scor *Score*	Semnificație *Meaning*	Obiectiv *Objective*	Remarci *Remarks*
1	$\Sigma(Y-\hat{Y})^p$	Suma erorilor de estimare	minim	Uzual p=2; pentru p = 1 şi mai mult pentru p =1/2 există avantajul că se favorizează tendința generală în regresie în defavoarea abaterilor grosolane de la ecuația de regresie
2	$r^2(Y,\hat{Y})$	Coeficient de determinare	maxim	Are avantajul legăturii cu scopul general al optimizării (maximizarea coeficientului de determinare)
3	$Mt(t_i)$	Media Minkowski a semnificațiilor	maxim	Are avantajul maximizării importanței fiecărui parametru al ecuației de regresie
4	$Hr(r^2)$	Entropia Renyi a determinării	minim	Are avantajul utilizării scării logaritmice
$\hat{Y}=a_0+\Sigma_i a_i \cdot Fenotip_i$ sau $\hat{Y}=\Sigma_i a_i \cdot Fenotip_i$ (când a_0 nu este semnificativ statistic diferit de zero)				

÷ Încrucişarea (vezi şi Figura 5-2 - Procedure GC_Cr în Anexa 5):

- Se selectează (folosind una din metodele de selecție descrise în Tabelul 5-4) două genotipuri, $Genotip_1=d_1p_1I_1O_1f_1M_1L_1$ şi $Genotip_2=d_2p_2I_2O_2f_2M_2L_2$;

- Se generează două numere aleatoare (din distribuția uniformă) între 0 şi 6 (fie ele 2 şi 4);
- Se încrucişează $Genotip_1$ cu $Genotip_2$ când se obțin descendenții $Descendent_1 = d_1p_1I_2O_2f_2M_1L_1$ şi $Descendent_2 = d_2p_2I_1O_1f_1M_2L_2$; se obțin fenotipurile acestor descendenți şi se adaugă la cultivar;

÷ Mutația (vezi şi Figura 5-2 - Procedure GC_Mu în Anexa 5):

- Se foloseşte o probabilitate de mutație α=1/7 (numărul de gene) pentru a decide dacă se aplică sau nu mutația;
- Se alege unul din descendenți (cel mai tare sau cel mai slab în raport cu scorul, sau aleatoriu);
- Se generează un număr aleatoriu între 0 şi 6 pentru a decide care genă va suferi mutația (fie acest număr 3);
- Corespunzător lui 3 este gena I; se generează atunci un număr aleatoriu între 1 şi 24 (numărul de valori posibile pe care le permite gena I); folosind valoarea generată aleatoriu între 1 şi 24 (fie acesta 12) se înlocuieşte valoarea genei I cu noua valoare (pentru 12 valoarea genei I este `$k(1/p_1p_2d)$`);

÷ Evoluția (algoritmul genetic, vezi şi Figura 5-4 & Algoritm 4 - Anexa 5):

- Aşa cum rezultă în acest punct, din întregul populației (care numără, după un calcul simplu) 787968 membrii, implementarea în abordarea algoritmilor genetici necesită generarea şi păstrarea unui eşantion (cultivar) restrâns din întreaga populație (maxim 100 de genotipuri); fie volumul cultivarului N;
- Se generează aleatoriu N genotipuri;
- Se obțin N fenotipuri asociate celor N genotipuri;
- Se repetă:
  - Pentru fiecare pereche (în cazul regresiei bivariate) de genotipuri
    - Se obține scorul asociat perechii de fenotipuri: $r^2(Y,\hat{Y})$
  - Sfârşit pentru.
  - Pentru fiecare genotip
    - Se obține scorul genotipului ca valoarea minimă a scorului pe care el îl obține prin împerechere cu celelalte genotipuri;
  - Sfârşit pentru.
  - Se aleg două genotipuri din cultivar folosind operatorul de selecție (Tabelul 5-4) şi se încrucişează; li se generează fenotipurile şi se adaugă cultivarului (volumul cultivarului devine N+2);
  - Se decide dacă unul dintre genotipuri va fi mutat folosind probabilitatea α=1/7; dacă da,

atunci se alege unul dintre genotipurile încrucişate şi se aplică operatorul de mutație; mutantului i se generează fenotipul şi se adaugă cultivarului (în caz favorabil mutației volumul cultivarului devine N+3);

- Folosind scorul genotipurilor se înlătură din cultivar exact atâtea genotipuri până când volumul cultivarului devine din nou N (minim 0, maxim 3);

- Până când scorul asociat celei mai bune perechi (în cazul regresiei bivariate) de genotipuri satisface condiția impusă (de exemplu $r^2 > 0.99$).

÷ Remarcă cu privire la cultivar, mutație şi încrucişare:

- Nu orice combinație de valori ale genelor produce un fenotip sănătos; la fel mutația poate produce un fenotip mort; acesta este motivul pentru care există posibilitatea ca după rularea unui ciclu complet de evoluție, cultivarul să rămână neschimbat (atunci când încrucişarea urmată de o eventuală mutație produc 3 fenotipuri care fie nu se nasc, fie nu se adaptează mediului);
- Semnificația este următoarea: aplicarea secvențelor de construcție a fenotipului descrise de cromozom pentru o anumită moleculă (sau pentru mai multe dintre acestea) poate, în traseul de construcție să întâlnească nedeterminări (împărțiri la 0, logaritm din număr negativ);
- Chiar dacă este născut fenotipul, există posibilitatea să nu se adapteze mediului (având de exemplu valori identice pentru toate moleculele setului, ceea ce-l face nefolositor în exprimarea unei ecuații de regresie cu proprietatea măsurată).

Analiza de mai sus este argumentativă pentru elaborarea algoritmului genetic care să rezolve o problemă dificilă de asociere a structurii date de o familie de descriptori de structură cu o activitate biologică măsurată a acestora.

O serie de instrumente statistice ce servesc în caracterizarea observabilelor cantitative (cum este cazul pentru valorile activității biologice observate lkow şi coeficientul de determinare r^2) sunt redate în Anexa 3-1 şi referă eşantionul statistic şi populația din care provine, fiind folosite în implementarea algoritmului genetic.

Algoritmul genetic: Definirea problemei genetice şi proiectarea algoritmului

Fiecare ***genă*** [4] codifică câte un operator folosit în construcția ***cromozomului*** [5] unui descriptor molecular (Tabelul 5-2). Fiecare descriptor al unei familii de descriptori este un ***genotip*** [6] şi toți împreună constituie ***materialul genetic*** [7] al familiei respective.

[4] genă = una din valorile de pe coloana Gene a Tabelului 36; ex. I_M pentru FPIF
[5] cromozom = secvența de gene a unei familii în Tabelul 36; ex. $D_M A_P I_D I_M F_C S_M L_O$ pentru MDF
[6] genotip = o concretizare posibilă a valorilor fiecărei gene a unui cromozom; ex. TCJtAAfDI pentru MDFV

Tabelul 5-2

Spațiul de căutare al familiilor de descriptori moleculari

Search space of molecular descriptors families

Familie *Family*	Gene *Gene*	Genom *Genome*																												
FPIF (©2000)	I_M	R	D																											
	D_M	T	G																											
	A_P	M	E	C	Q																									
	P_D	__p__			__d__			_1/p_		_1/d_			_p*d_			_p/d_			_p/d2			p2/d2							
	F_C	si		se		ji		je		fi		fe																		
	S_M	S	P	A	G	H																								
	M_I	P_		P2		E_		E2																						
	L_O	I	R	L																										
MDF (©2005)	D_M	t	g																											
	A_P	C	H	M	E	G	Q																							
	I_D	D	d	O	o	P	p	Q	q	J	j	K	k	L	l	V	E	W	w	F	f	S	s	T	t					
	I_M	r	R	m	M	d	D																							
	F_C	m	M	D	P																									
	S_M	m	M	n	N	S	A	a	B	b	P	G	g	F	f	s	H	h	I	i										
	L_O	I	i	A	a	L	l																							
MDFV (©2008)	D_O	T	G																											
	A_P	C	H	M	E	Q	L	A																						
	I_D	J	j	O	o	P	p	Q	q	R	r	K	k	L	l	M	m	N	n	W	w	X	x	Y	y	Z	z	S	s	T
		t	U	u	V	v	F	f	G	g	H	h	I	i	A	a	B	b	C	c	D	d	0	1	2	3	4	5	6	7
	S_F	A	a	I	i	F	P	C																						
	S_M	A	a	I	i	F	P	C																						
	I_T	f	F	c	C	p	P	a	A	i	I																			
	E_U	D	d																											
	L_O	I	R	L																										
SAPF (©2009)	C_F	D	P	C																										
	D_O	T	G																											
	A_P	C	H	M	E	A																								
	D_P	I	E	H	G	A	Q	S																						
	P_P	I	E	H	G	A	Q	S																						
	O_M	S	M																											
	M_P	I	E	H	G	A	Q	S																						
	L_O	I	A	S	T	Q	R	L																						

Numărul de valori pe care le codifică fiecare genă variază de la 2 valori (în cazul genei ce codifică tipul de metrică de distanță - topologică şi geometrică - D_M pentru FPIF şi MDF şi D_O pentru MDFV şi SAPF) până la 58 de valori în cazul descriptorului de interacțiune I_D al familiei MDFV. Volumul materialului genetic variază şi Tabelul 5-3 sumarizează aceste variații:

[7] material genetic = mulțimea tuturor combinațiilor posibile de valori de pe coloana Genom în tabelul 36; ex. {D, P, C} × {T,G} × {C,H,M,E,A} × {I,E,H,G,A,Q,S} × {S,M} × {I,E,H,G,A,Q,S} × {I,E,H,G,A,Q,S} × {I,A,S,T,Q,R,L} pentru SAPF

Tabelul 5-3

Volumele familiilor de descriptori moleculari
Volumes of the molecular descriptors families

Familie (*Family*)	Gene (*Gene*)								Volume (N)
FPIF (©2000)	I_M:2	D_M:2	A_P:4	P_D:8	F_C:6	S_M:5	M_I:4	L_O:3	46080
MDF (©2005)	D_M:2	A_P:6	I_D:6	I_M:24	F_C:4	S_M:19	L_O:6		787968
MDFV (©2008)	D_O:2	A_P:7	I_D:58	S_F:7	S_M:7	I_T:10	E_U:2	L_O:3	2387280
SAPF (©2009)	C_F:3	D_O:2	A_P:5	D_P:6	P_P:6	O_M:2	M_P:6	L_O:6	77760

Mărimea spațiului de căutare (V_S) se calculează în funcție de volumul familiei de descriptori moleculari (N) şi ordinul de multiplicitate al regresiei multiple (`n`) cu formula:

$$V_S = \prod_{j=1}^{n} \frac{N-j+1}{j} = \binom{N}{n}$$

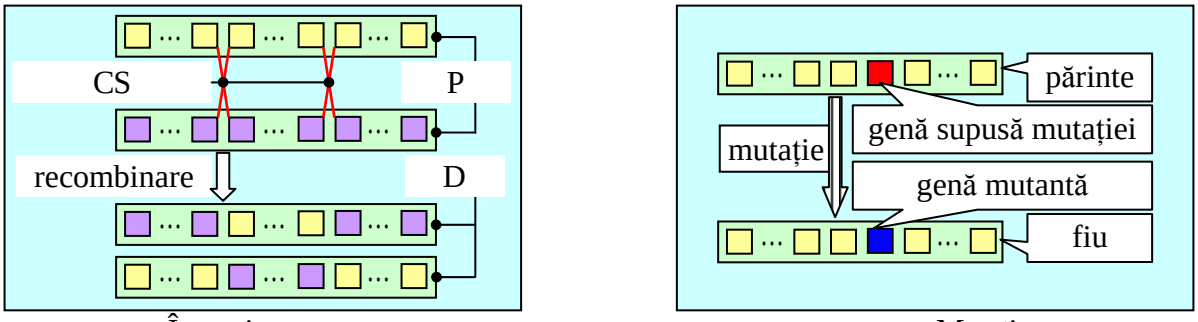
Volumul spațiului de căutare dat de ecuația de mai sus permite exprimarea complexității calculului necesar pentru a parcurge întreg spațiul de căutare, această valoare fiind dublată deoarece căutarea se poate face pe baza ecuației (1) sau ecuației (2), ecuații care folosesc expresii diferite de calcul al parametrilor ecuației de regresie.

Metodologia algoritmilor genetici presupune prelevarea inițială (întâmplător sau deterministic) a unui ***eşantion*** [8] de cromozomi din materialul genetic format dintr-un şir de descriptori X_1, ..., X_p care este supus procesului de ***evoluție*** [9] în ***cultivar*** [10]. ***Algoritmul genetic*** [11] operează astfel asupra eşantionului care suferă modificări în fiecare ***generație*** [12]. Fiecare mulțime de `n` descriptori distincți reprezintă un punct în ***spațiul de căutare*** [13] şi în acelaşi timp o ***soluție posibilă*** [14].

Operatorii de bază ai unui algoritm genetic sunt încrucişarea şi mutația cromozomilor. ***Încrucişarea*** [15] (reprezentată în Figura 5-2) a două genotipuri presupune alegerea unei porțiuni de recombinat de-a lungul şirului de gene (întâmplător sau deterministic) şi valorile celor două porțiuni de şiruri sunt schimbate între ele, când rezultă doi descendenți. ***Mutația*** [16] (reprezentată în Figura 5-2) unui genotip presupune modificarea unei valori a unei gene a cromozomului cu o altă

[8] eşantion = submulțime a materialului genetic al familiei de descriptori moleculari; ex. {DTCIISII, DTCIESII, DTCGISII} reprezintă un eşantion de volum 3 al materialului genetic al SAPF
[9] evoluție = proces genetic complex care implică procese de selecție, încrucişare şi mutație
[10] cultivar = spațiu (de memorie, virtual) în care genotipurile sunt transformate în fenotipuri prin aplicarea operatorilor definiți de valorile genelor pentru întreg setul de `m` molecule supus studiului; fenotipul asociat genotipului este astfel un şir de `m` valori numerice (câte una pentru fiecare moleculă a setului)
[11] algoritm genetic = algoritm care descrie prin instrucțiuni procesul de evoluție asupra eşantionului
[12] generație = una din iterațiile algoritmului genetic
[13] spațiul de căutare = mulțimea posibilităților de selecție a `n` descriptori din `$V_S$` posibilități (relația 20)
[14] soluție posibilă = o ecuație de regresie cu `n` descriptori distincți definită de relația (1) sau (2)
[15] încrucişarea = procesul prin care o porțiune a materialului genetic al unui cromozom este înlocuită de porțiunea corespunzătoare a altui cromozom şi viceversa; încrucişarea este făcută în speranța că dacă se recombină porțiuni de genotipuri de succes, atunci acest proces este probabil să producă descendenți chiar mai buni decât părinții din care provin
[16] Mutație = operatorul care introduce modificări noi (inexistente în eşantionul unei generații); ceea ce este caracteristic în general mutației şi implicit şi operatorului acesteia corespondent în algoritmii genetici este că ea se petrece cu o probabilitate scăzută, fiind deci aplicată cu o probabilitate scăzută

valoare din lista valorilor posibile pentru gena respectivă. Rezultatul încrucişării şi al mutaţiei este obţinerea de ***descendenţi*** [17] sau fii.



Încrucişare Mutaţie

Legendă (Legend): SC - site încrucişare (*crossover site*); P - Părinţi (*Parents*); R - Recombinare (*Recombination*); D - Descendenţi (*Descendent*)

Fig. 5-2. Operatorii de încrucişare şi mutaţie

Operators for crossover and mutation

Selecţia [18] genotipurilor este operaţia prealabilă necesară încrucişării şi mutaţiei şi se face pe baza unui ***scor de selecţie*** [19]. Cel puţin o parte a descendenţilor reprezintă descriptori ***viabili*** [20] putând face atunci parte din soluţii candidate ale generaţiilor următoare. Descendenţii viabili înlocuiesc o parte corespunzătoare a indivizilor din eşantion în urma unui proces de ***supravieţuire*** [21] aplicat indivizilor din eşantion pe baza unui ***scor de supravieţuire*** [22]. Un alt parametru al algoritmului genetic îl reprezintă ***obiectivul evoluţiei*** [23] care este urmărit pe baza unei ***funcţii obiectiv*** [24].

Urmărirea obiectivului evoluţiei se face odată la fiecare generaţie selectând din eşantion acei indivizi care maximizează sau, după caz minimizează valoarea funcţiei obiectiv (acei indivizi care

[17] Descendenţi = genotipurile obţinute din încrucişarea şi eventual mutaţia indivizilor din eşantion
[18] Selecţie = operatorul cu ajutorul căruia se extrage din eşantion mai mulţi indivizi care participă la înmulţire
[19] Scor de selecţie = valoare numerică asociată individului din eşantion calculată pe baza (sau exprimată din) tăriei fenotipului în cultivar
[20] Viabilitatea (unui descriptor molecular) = referă potenţialul acestuia de a fi folosit în regresii; un descriptor este viabil dacă (ceea ce urmează referă fenotipul acestuia, viabilitatea referind astfel manifestarea fenotipică) cel puţin are valori reale şi finite pentru toate moleculele din set şi nu are toate valorile identice; suplimentar i se pot impune şi alte condiţii, cum ar fi o variabilitate rezonabilă (prin intermediul unui coeficientul de variaţie), o abatere de la normalitate rezonabilă (prin intermediul unui test de normalitate cum este Jarque-Bera) şi o capacitate de explicare a proprietăţii măsurate rezonabilă (prin intermediul coeficientului de determinare din regresia liniară simplă cu proprietatea măsurată).
[21] Supravieţuire = operatorul cu ajutorul căruia se extrage din eşantion mai mulţi indivizi care vor fi înlocuiţi în eşantion de către descendenţi
[22] Scor de supravieţuire = valoare numerică asociată individului din eşantion care poate fi o valoare obţinută atât pe baza caracteristicilor genotipice ale individului (asociind o măsură a similarităţii acestuia cu alte genotipuri din cultivar în scopul menţinerii diversităţii materialului genetic) cât şi pe baza trăsăturilor fenotipice ale individului (asociind o măsură a similarităţii acestuia cu alte fenotipuri din cultivar în scopul menţinerii diversităţii fenotipice)
[23] Obiectivul evoluţiei = parametrul sau caracteristica (unei ecuaţii de regresie) care constituie subiectul optimizării (minimizare - ex. suma pătratelor diferenţelor între (erorilor) valoarea măsurată şi cea explicată de model; maximizare - ex. coeficientul de determinare)
[24] Funcţia obiectiv = algoritmul (procedura) de calcul al parametrului sau caracteristicii ce constituie obiectivul evoluţiei eşantionului.

fac parte din cea mai bună ecuație de regresie obținută pe baza indivizilor din cultivar). Se poate opta ca indivizii care ating obiectivul evoluției într-o generație să fie păstrați în eșantion, caz în care acestora nu li se mai aplică procesul de supraviețuire, ei fiind automat incluși în eșantionul generației următoare.

Așa cum rezultă din aplicarea procesului de supraviețuire, nu toți indivizii unei generații supraviețuiesc și sunt incluși în generația următoare. Motivul acestui fapt este păstrarea unui număr constant de genotipuri în eșantionul dezvoltat în cultivar, astfel încât numărul de indivizi înlocuiți din eșantion este egal cu numărul de descendenți viabili obținuți în urma selecției, încrucișării și mutației.

Selecția și supraviețuirea având la bază scorurile de selecție și de supraviețuire se realizează folosind o ***strategie de selecție și supraviețuire*** [25]. Tabelul 5-4 redă modalitățile de selecție și supraviețuire folosite.

Tabelul 5-4

Selecție (și supraviețuire) în algoritmii genetici
Selection (and survival) in genetic algorithms

Strategie *Strategy*	Expresia funcției de scor *Expression of score function*	Selecție *Selection*	Comentarii *Comments*
Proporțional ÷ SL_Pro ÷ SV_Pro	f_i=Fitness(Cromozom_i)	$p_i=f_i/\Sigma_i f_i$	Șansa de selecție este proporțională cu scorul (utilizând probabilitatea p_i în selecție)
Deterministic ÷ SL_Det ÷ SV_Det		i | f_i = max. sau min.	Selecția indivizilor este făcută pe baza celui mai tare (sau celui mai slab) individ (elitism)
Turnir ÷ SL_Tur ÷ SV_Tur		(f_i,f_j) max. sau min.	Perechi de indivizi concurează între ei pentru selecție (din nou este selectat cel mai tare sau cel mai slab)
Normalizare ÷ SL_Fit ÷ SV_Fit	$g_i=(f_i-N_0)(f_{max.}-f_{min.})/(N_1-N_0)$	$p_i=g_i/\Sigma_i g_i$	O scală fixă [N_0,N_1] normalizează scorul fenotipurilor între generații diferite
Ranguri ÷ SL_Rnk ÷ SV_Rnk	h_i=Rank(f_i)($f_{max.}$-$f_{min.}$)/Size	$p_i=h_i/\Sigma_i h_i$	Șansa este proporțională cu rangul scorului unde: Rank(·): rangul; Size: volum genom

Următoarele remarci cu privire la utilizarea unui algoritm genetic definesc cadrul de operare al acestuia:

÷ Utilizând doar selecția nu va reuși decât să copieze (cloneze) cel mai bun individ al său în

[25] modalitate de selecție și modalitate de supraviețuire = metodă de extragere a unui individ din eșantion ce folosește drept parametru valorile scorurilor (de selecție și respectiv de supraviețuire) ale indivizilor ce compun eșantionul; ceea ce Tabelul 38 prezintă în mod formal exprimă faptul că se folosesc trei alternative de selecție (proporțional, deterministic și turnir) care se pot aplica valorilor scorurilor sau rangurilor scorurilor (când selecția e bazată pe valorile rangurilor în locul valorilor scorurilor); de asemenea, valoarea scorului poate fi supusă unui proces de normalizare care corectează (ajustează relativ) scorurile indivizilor din cultivar în raport cu două valori (una de minim și alta de maxim) care se actualizează global în fiecare generație pe parcursul întregii evoluții a eșantionului în cultivar

întregul eşantion;

÷ Utilizând doar mutația nu va reuşi decât să parcurgă întâmplător spațiul de căutare;

÷ Utilizând doar selecția urmată de încrucişare va reuşi să conveargă către o soluție bună (un optim local) dar nu sub-optimală (în apropierea celei optime);

÷ Utilizând doar selecția urmată de mutație se creează algoritmi paraleli, toleranți la perturbații în căutarea de punctelor de maxim local (în terminologia în engleză: hill-climbing);

÷ Utilizarea selecției urmate de mutație şi încrucişare asigură toate caracteristicile de definire ale unui algoritm genetic;

Se poate verifica că ecuația (20) defineşte o ***problemă dificilă*** [26]. Astfel, reprezentând grafic ecuația (20) pentru diferite valori ale volumului familiei de descriptori moleculari (`N`) date în Tabelul 5-4 şi pentru diferite ordine de multiplicitate (`n`) se obțin reprezentările de mai jos (Figura 5-3) care susțin faptul că ecuația (20) defineşte o problemă dificilă.

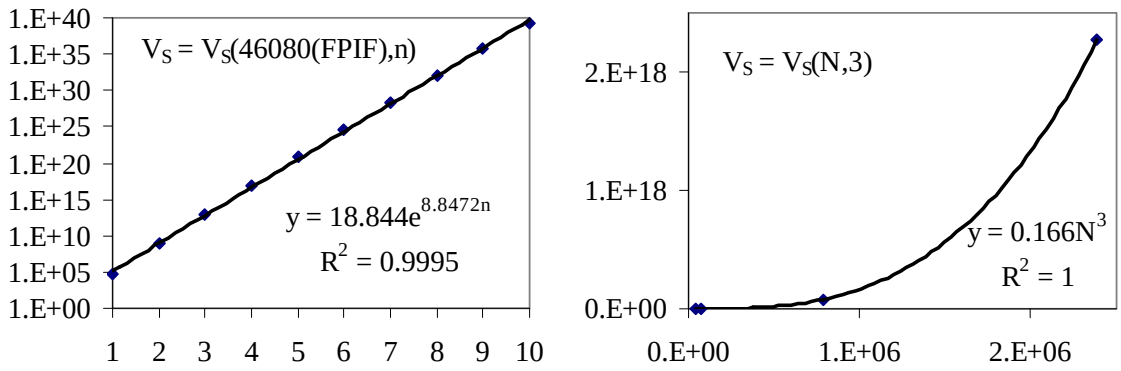


Fig. 5-3. Complexitatea exponențială a regresiilor multiple cu familii de descriptori moleculari
The exponential complexity of multiple regressions with molecular descriptors families

Pentru a rezolva o ***problemă dificilă*** un algoritm genetic generează întâmplător (sau inițiază cu valori predefinite) un eşantion de un volum dat de genotipuri (volum ce va fi păstrat constant pe toată durata evoluției).

Algoritmul genetic (reprezentat în Figura 5-4 & detaliat în Algoritm 4 - Anexa 5) evoluează astfel:

÷ Repetă

- Pasul_1: Utilizând operatorul de selecție selectează perechi de cromozomi;
- Pasul_2: Calculează scorul de selecție al genotipurilor; calculează scorul de supraviețuire al genotipurilor; calculează funcția obiectiv a fenotipurilor şi obține grupul descriptorilor din eşantion ce întrunesc obiectivul în generație şi eventual include în generația următoare automat genotipurile din acest grup;

[26] problemă dificilă = problemă a cărei rezolvare de către cel mai bun algoritm (care ar putea fi imaginat) necesită un timp de execuție care creşte exponențial în funcție de volumul datelor problemei (şi în acelaşi timp de intrare ale algoritmului); denumirea de probleme dificile vine de la faptul că deoarece chiar şi cel mai bun algoritm (care ar putea fi imaginat) va fi probabil ineficient pe date de intrare reale (din practică)

- Pasul_3: Cu o mică probabilitate şi utilizând o funcție discretă de probabilitate uniformă alege porțiunea de mutat şi mută genotipurile selectate (părinți);
- Pasul_4: Utilizând o funcție discretă de probabilitate uniformă alege porțiunea de încrucişat şi produce descendenți ai acestora (descendenți);
- Pasul_5: Cu o mică probabilitate şi utilizând o funcție discretă de probabilitate uniformă alege porțiunea de mutat şi mută genotipurile descendenților (fii);
- Pasul_6: Utilizând operatorul de supraviețuire înlocuieşte o parte din părinți cu fii;

÷ Până când (condiție care reprezintă condiția de sfârşit a algoritmului)

- Se satisface o condiție impusă cu privire la valoarea funcției obiectiv (atinge o valoare impusă) sau se realizează un număr dat de iterații (evoluții).

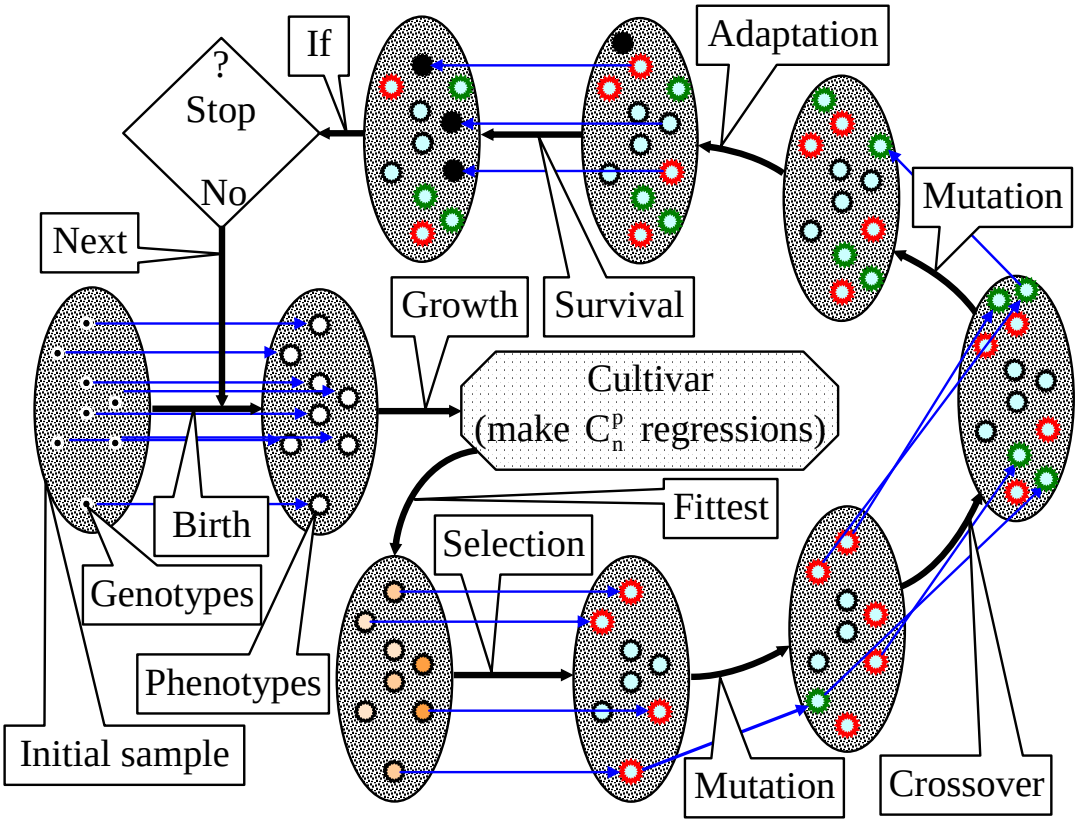


Fig. 5-4. Ilustrarea algoritmului genetic implementat
imagine din (Jäntschi & Bolboacă, 2009-Manuscript)
Illustration of the implemented genetic algorithm
- image from (Jäntschi & Bolboacă, 2009-Manuscript)

În implementarea algoritmului s-a ținut seama de următoarele considerente:

÷ Calculele necesare pentru obținerea valorilor unui descriptor molecular sunt complexe, implicând fragmentarea grafului molecular, calcularea distanțelor topologice şi geometrice în moleculă, aplicare de modele de suprapunere a interacțiunilor, fiecare necesitând un timp de execuție şi o memorie de lucru considerabile, astfel încât populația de descriptori trebuie generată şi stocată în prealabil execuției algoritmului genetic; este stocată într-o bază de date

pentru MDF şi MDFV; este stocată în fişiere text cu format impus pentru FPIF şi SAPF.

÷ Volumul de informație al populației de descriptori este considerabil (ex. tabela `PCB_lkow_tmpx` care stochează a şasea parte din descriptorii MDF (131328, Tabelul 5-3) calculați pentru 206 PCBs - din totalul de 209 - pentru care există măsurat în aceleaşi condiții experimentale coeficientul de partiție octanol-apă ocupă peste 200Mb - atât în stocare fizică pe disc cât şi în memoria internă) astfel încât manipularea acestui volum de informație ridică serioase provocări de optimizare a memoriei de lucru şi vitezei de calcul; alegerea mediului de programare a ținut seama de aceste considerente; s-a ales drept mediu de implementare FreePascal [27] un mediu de programare freeware multi-platformă bazat pe limbajul Pascal, recunoscut pentru viteza de calcul şi economia de memorie.

÷ Întrucât evoluția algoritmului genetic se doreşte a fi rapidă, accesul şi interogarea repetată a bazei de date pentru obținerea valorilor unui descriptor molecular la moleculele din setul de lucru încetineşte considerabil viteza de lucru, astfel încât soluția aleasă a fost stocarea în memoria internă a imaginii bazei de date, cu cost suplimentar în ceea ce priveşte memoria de lucru, dar cu câştig deosebit în ceea ce priveşte viteza de lucru.

÷ Nu toți descriptorii din familie (una dintre FPIF, MDF, MDFV sau SAPF - vezi Tabelul 4-4 şi Tabelul 5-2) întrunesc condițiile reale de existență (să aibă valori reale şi finite pentru toate moleculele din setul considerat; să nu aibă valori banale - identice pentru toate moleculele din set); mai mult, impunerea de condiții suplimentare cu privire la variabilitate, capacitate de explicare şi normalitate reduce şi mai mult setul de descriptori candidați în regresii multiple pentru setul de molecule considerat; stocarea valorilor acestor descriptori în memoria internă ar fi lipsită de sens, şi astfel s-a impus realizarea acestei selecții în preambulul execuției efective a algoritmului genetic, introducând o variabilă care conține două valori de adevăr (T/F) pentru validitatea descriptorului molecular la setul de molecule considerat; mai mult, practic stocarea unei fracții din totalul descriptorilor moleculari (fapt valabil pentru toate familiile de descriptori) s-a făcut pe seama renunțării la o genă din secvența cromozomului familiei - operatorul de linearizare - fapt care însă trebuie compensat în execuția algoritmului genetic; astfel s-au mai introdus un număr de variabile cu două valori de adevăr (T/F) pentru fiecare operator de linearizare (câte 6 pentru MDF şi SAPF şi câte 3 pentru FPIF şi MDFV), fiecare conținând valoarea de adevăr a validității descriptorului molecular respectiv în urma linearizării - numit în continuare fenotip; astfel, variabila globală definită pentru fracția de descriptori stocată intern a căpătat o nouă semnificație, şi anume adevărat (T) dacă cel puțin unul dintre fenotipurile sale este valid.

÷ O serie de parametrii de statistică descriptivă sunt esențiali (şi implicați în mod repetat) pentru

[27] FreePascal, http://freepascal.org

analiza de regresie (valori medii, momente) astfel încât s-a impus calcularea prealabilă a acestora pentru fenotipurile valide incluse în eşantion.

Programul evolutiv [28] rezultat trebuie să fie gândit astfel încât să poată lucra cu oricare familie de descriptori moleculari (incluzând desigur cele 4 familii FPIF, MDF, MDFV, SAPF - Tabelul 4-4, Tabelul 5-2 & Anexa 4) şi să poată fi parametrizat; soluția de implementare aleasă este crearea şi utilizarea de fişiere de configurare.

Prezentarea şi documentarea programului evolutiv ce implementează algoritmul genetic realizat este redată în Anexa 5.

Obiectivul evoluției: r^2 = max.

Procesul de optimizare a relației SAR sub formă de MLR cu 4 variabile MDF pentru predicția lkow a PCBs are ca scop obținerea unei cât mai bune relații, care să posede bune capacități de estimare (reziduuri teoretic vs. experimental cât mai mici) şi de predicție (pentru molecule care aparțin din aceeaşi clasă de compuşi dar care nu au fost incluşi în analiză.

Care este cel mai bun instrument cu ajutorul căruia se poate obține acest lucru este o discuție deschisă în literatura de specialitate.

Se poate alege să se minimizeze reziduurile, să se maximizeze determinarea şi lista de posibilități poate continua (de exemplu suma pătratelor reziduurilor poate fi înlocuită cu suma modulelor reziduurilor, determinarea în estimare poate fi înlocuită cu determinarea în validarea încrucişată, corelația Pearson cu corelația Spearman, ş.a.m.d.) - (Bolboacă & Jäntschi, 2006-PSK).

A fost aleasă soluția de a maximiza coeficientul de determinare în expresia sa clasică (pătratul coeficientului de corelație multiplă Pearson).

Caracterizarea obiectivului evoluției referă două aspecte:

÷ identificarea legii de distribuție a acestuia într-o generație - şi pentru aceasta o sumă de ipoteze de distribuție formează materialul cercetării (legi de distribuție) - acestea fiind redate în Anexa 3-2 - şi respectiv măsurarea asocierii între observație şi model - şi pentru aceasta o sumă de statistici de asociere formează metoda cercetării (teste de semnificație statistică) - acestea fiind redate în Anexa 3-3.

[28] program evolutiv = în accepțiunea generală este un program ce implementează un algoritm genetic

6. Definirea experimentului de simulare a evoluției

Observabile: Parametrii definiți în implementarea programului evolutiv

Pe lângă alegerea familiei de descriptori (MDF), a setului de molecule (206 PCBs), şi a proprietății măsurate (coeficientul de partiție octanol/apă în unități logaritmice lkow) execuția programului evolutiv a impus alegerea unei serii de parametrii intrinseci de funcționare a algoritmului genetic.

Fişierul de configurare a execuției algoritmului genetic excluzând opțiunile de afişare numără nu mai puțin de 30 de parametrii (redați în Tabelul 6-1) dintre care parametrii ordinali un număr de 19 şi dintre care parametrii cu valori dintr-o listă de valori impuse (finită şi definită) un număr de 12.

Tabelul 6-1

Opțiuni de configurare pentru algoritmul genetic realizat (vezi şi Tabelul 5 - Anexa 5)
Configuration options for the realized genetic algorithm (see Table 5 - Annex 5 too)

Parametru *Parameter*	Tip (şi valori pentru listă) *Type (and its values for lists)*
a_v_ADAPT_Variance	Real
ajb_ADAPT_JarqueBera	Real
a_c_ADAPT_Correlation	Real
sn0_SAMPLE_Size	Întreg (natural)
rn0_REGRESSION_Multiple	Întreg (natural)
e1n_GENERATIONS_max	Întreg (natural)
g_r_GENERATIONS_first_rich	Listă: {Yes, No}
cn0_CROSSOVER_Pairs	Întreg (natural)
m_m_MUTATION_Genes	Întreg (natural)
mpp_MUTATION_Parent_probability	Real
mcp_MUTATION_Child_probability	Real
b_p_SELECTION_parameter	Listă: {r2, se, Mt, Hr}
b_o_SELECTION_objective	Listă: {min, max}
sfs_FITNESS_strategy	Listă: {proportional, deterministic, tournament}
sfn_FITNESS_normalized	Listă: {Yes, No}
sfr_FITNESS_ranks	Listă: {Yes, No}
sfa_FITNESS_accuracy	Întreg (natural)
sff_FITTEST_function	Listă: {nalive, r2_min, se_min, Mt_min, Hr_min, r2_max, se_max, Mt_max, Hr_max, r2_avg, se_avg, Mt_avg, Hr_avg}
sfo_FITTEST_objective	{min, max}
fr2_FITTEST_r2_p	Real
fse_FITTEST_se_p	Real
fMt_FITTEST_Mt_p	Real
fHr_FITTEST_Hr_p	Real
v_p_SURVIVAL_phenotyping_p	Real
v_g_SURVIVAL__genotyping_p	Real
vfs_SURVIVAL_strategy	Listă: {proportional, deterministic, tournament}
vfr_SURVIVAL_ranks	Listă: {Yes, No}
e0n_RUNS_number	Întreg (natural)
b_k_RUNS_kepp_best_in_sample	Listă: {Yes, No}
b_f_RUNS_get_best_from_file	Listă: {Yes, No} + listă genotipuri în fişier c_galg.txt

Un simplu calcul matematic efectuat asupra parametrilor din urmă, cu valori dintr-o listă finită şi definită, dă numărul total posibil de combinații de investigat de 134784 (vezi Tabelul 4-4), fără a socoti că pentru fiecare dintre aceste combinații, există practic o infinitate de posibilități de configurare pentru parametrii rămaşi, în număr de 18.

Argumentele de mai sus pot părea descurajatoare în ceea ce privește posibilitățile de investigare sistematică a rezultatelor obținute prin execuția algoritmului genetic. Trebuie însă văzută partea favorabilă a situației, și anume că se oferă practic o diversitate foarte mare de stări algoritmului genetic prin care se asigură variabilitatea acestuia, care este benefică atâta timp cât algoritmul genetic este menit să investigheze activități biologice și proprietăți fizico-chimice și ele cu o variabilitate foarte mare, atât sub aspectul provenienței măsurătorilor (din observații umane, instrumentale), scări absolute și relative de măsură, variabile discrete și continue observate.

O altă remarcă se poate face cu privire la valorile parametrilor de configurare (Tabelul 5-1). Astfel, nu toți parametrii pot lua orice valori (cum e cazul pentru *sn0_SAMPLE_Size* care în mod obligatoriu trebuie să fie cel puțin cât *rn0_REGRESSION_Multiple* în timp ce *cn0_CROSSOVER_Pairs* nu poate depăși jumătate din *sn0_SAMPLE_Size*) în timp ce definirea valorilor altor parametrii este restricționată la un domeniu de valori pentru simplul fapt că efectul definirii unei valori din afara domeniului este fie lipsită de sens (de exemplu definirea unei valori întregi negative pentru *e0n_RUNS_number* sau pentru *e1n_GENERATIONS_max* face ca algoritmul genetic să nu mai fie executat), fie nu produce nici un fel de efecte diferite de o anumită valoare din domeniu (atribuirea unei valori negative pentru *a_c_ADAPT_Correlation* este similară cu atribuirea valorii 0 și semnificând că nu există valoare minimă impusă pentru coeficientul de determinare între fenotipurile ce compun o ecuație de regresie și proprietatea măsurată), fie nu poate duce la atingerea obiectivului (atribuirea unei valori mai mari ca unitatea pentru *a_c_ADAPT_Correlation* este similară cu atribuirea valorii 1 și are ca efect impunerea determinării de 100%, ceea ce pe cale de consecință face lipsită de sens căutarea de regresii multiple atâta timp cât ar exista un descriptor molecular capabil să explice singur în proporție de 100% proprietatea măsurată), în timp valorile de configurare a unor parametrii impune valorile altora (de exemplu *b_p_SELECTION_parameter=r2* impune *b_o_SELECTION_objective=max*, pentru simplul fapt că obținerea unei ecuații de regresie cu cel mai mic coeficient de determinare nu poate constitui un obiectiv de interes, tot așa cum *b_p_SELECTION_parameter=se* impune *b_o_SELECTION_objective=min*), în timp ce limitele valorilor unor parametrii sunt impuse doar de soluția de memorare aleasă la implementare (de exemplu *e1n_GENERATIONS_max* admite ca valoare maximă 2147483647 impusă de tipul de dată *LongInt* în care este stocată valoarea, dar care poate fi însă modificat fără dificultate la tip de dată *Int64* cu limita superioară de 9223372036854775807.

Performanța algoritmului genetic a fost evaluată prin prisma rezultatelor colectate în fișierele de ieșire. Configurația parametrilor de ieșire este redată în Tabelul 6-2.

În configurația precizată mai sus programul evolutiv produce 2 fișiere la o execuție a programului: PCB_XXXX_evo.txt și PCB_XXXX_cfg.txt unde XXXX este un număr generat la

întâmplare pentru a permite execuția simultană a programului în diferite configurații pe aceeaşi maşină de calcul.

Tabelul 6-2

Configurații de afişare în experiment (vezi şi Tabelul 5 - Anexa 5)
Experiment's displaying configurations (see Table 5 - Annex 5 too)

Parametru *Parameter*	Valoare *Value*
d_d_SHOW_descriptive_XX (XX = m0, m1, m2, m3, m4, mx, my, v0, g1, g2, jb, r1, r2)	Yes
d_f_SHOW_fitness_YY (YY = nalive, r2_min, se_min, Mt_min, Hr_min, r2_max, se_max, Mt_max, Hr_max, r2_avg, se_avg, Mt_avg, Hr_avg)	Yes
d_c_SHOW_configuration	Yes
d_m_SHOW_mols	No
d_r_SHOW_regressions	No
d_p_SHOW_phenotypes	No
d_s_SHOW_genotypes	No
d_t_SHOW_fittests	No
d_g_SHOW_generations	No
d_e_SHOW_evolutions	Yes

În fişierul PCB_XXXX_evo.txt se scrie câte o linie la fiecare evoluție (obținerea unei relații mai bune decât cea anterior existentă) în timp ce în fişierul PCB_XXXX_cfg.txt conținând la partea sa de început opțiunile de configurare folosite în execuție se scrie o singură linie la încheierea procesului de evoluție.

Pentru ca analiza statistică a datelor obținută din simularea evoluției în cele 9 perechi diferite de strategii de selecție şi supraviețuire să fie hrănită cu suficientă informație pentru atingerea pragului de semnificație statistică au fost făcute un număr de 46 de execuții independente ale algoritmului genetic în aceeaşi configurație de execuție a programului evolutiv. De fiecare dată, o nouă execuție porneşte de la o nouă generație inițială, total independentă de evoluția anterioară.

Datele scrise în fişierele PCB_XXXX_evo.txt (pentru o evoluție) şi PCB_XXXX_cfg.txt (pentru o execuție) sunt structurate într-o formă tabelară, şi semnificațiile acestor valori sunt date de parametrii ce apar în capul de tabel din fişierul text.

Informații cu privire la evoluție în fişierul PCB_XXXX_evo.txt

÷ e1i: generația (număr întreg) în care s-a produs o nouă evoluție (evoluția referă valoarea funcției obiectiv r2 → max.; prin evoluție înțelegând obținerea unei relații structură activitate cu un coeficient de determinare mai mare decât cel obținut anterior);
÷ sn2: număr de genotipuri viabile (nu totdeauna toate genotipurile obținute în urma încrucişărilor şi mutațiilor sunt viabile adică prin aplicarea operatorului de linearizare se obține măcar unul din cei 6 posibili indivizi care posedă ca şi caractere valori reale, finite, nenule şi cu o variabilitate rezonabilă pentru toate moleculele setului de compuşi; numărul maxim de genotipuri viabile e dat de volumul eşantionului, şi anume variabila de configurare în program sn0);
÷ pn2: număr de fenotipuri (indivizi, descriptori) în viață din cultivar; numărul maxim de

fenotipuri viabile la un anumit moment de timp este de 6 ori numărul de genotipuri viabile sn2; numărul maxim de fenotipuri viabile ce pot exista în cultivar este de 6 ori volumul eşantionului sn0);

÷ nrali: număr de asocieri fenotipice (regresii) obținute cu indivizii din cultivar; asocierile fenotipice se obțin din combinația a 4 fenotipuri distincte din cultivar şi pentru ca asocierea să fie viabilă i se impun aceleaşi condiții ca cele impuse unui individ; prima limită maximă este dată de numărul de fenotipuri viabile existente: combinări de pn2 luate câte 4; a doua limită maximă este combinări de 6*sn2 luate câte 4; a treia limită maximă este 6*sn2 luate câte 4);

÷ Gen11: genotipul aflat pe poziția 12 în eşantion într-un anumit moment al evoluției;

÷ Gen10: genotipul aflat pe poziția 11 în eşantion într-un anumit moment al evoluției;

÷ Gen9: genotipul aflat pe poziția 10 în eşantion într-un anumit moment al evoluției;

÷ Gen8: genotipul aflat pe poziția 9 în eşantion într-un anumit moment al evoluției;

÷ Gen7: genotipul aflat pe poziția 8 în eşantion într-un anumit moment al evoluției;

÷ Gen6: genotipul aflat pe poziția 7 în eşantion într-un anumit moment al evoluției;

÷ Gen5: genotipul aflat pe poziția 6 în eşantion într-un anumit moment al evoluției;

÷ Gen4: genotipul aflat pe poziția 5 în eşantion într-un anumit moment al evoluției;

÷ Gen3: genotipul aflat pe poziția 4 în eşantion într-un anumit moment al evoluției;

÷ Gen2: genotipul aflat pe poziția 3 în eşantion într-un anumit moment al evoluției;

÷ Gen1: genotipul aflat pe poziția 2 în eşantion într-un anumit moment al evoluției;

÷ Gen0: genotipul aflat pe poziția 1 în eşantion într-un anumit moment al evoluției;

÷ rr2: valoarea funcției obiectiv "coeficient de determinare" al evoluției; coeficientul de determinare al celei mai bune SAR sub formă de MLR cu 4 descriptori MDF prezenți în cultivar;

÷ rse: valoarea funcției obiectiv (alternative) a evoluției "suma erorilor"; suma erorilor de determinare al celei mai bune SAR sub formă de MLR cu 4 descriptori MDF prezenți în cultivar;

÷ rMt: valoarea funcției obiectiv (alternative) a evoluției "medie semnificații Student t parametrii de regresie"; idem cu rr2 şi rse;

÷ rHr: valoarea funcției obiectiv (alternative) a evoluției "entropie a determinării"; idem cu rr2, rse şi rMt;

÷ rdr: număr de coeficienți în ecuația de regresie care dă valoarea funcției obiectiv (cea mai bună regresie cu indivizi din cultivar); rdr=rn0+1 dacă această ecuație are termen liber (i.e. acesta are semnificație statistică) şi rdr=rn0 dacă această ecuație nu are termen liber (i.e. în ecuația alternativă termenul liber nu are semnificație statistică);

÷ rme: media erorii din ecuația de regresie ce dă funcția obiectiv (MSE);

÷ rdt: număr de grade de libertate în ecuația de regresie ce dă funcția obiectiv (număr de molecule în set (nmo) minus număr de coeficienți în ecuația de regresie (rdr);

÷ rm0: media modulului estimatei M(|$\hat{Y}$|);

÷ rm1: media estimatei M($\hat{Y}$);

÷ rm2: media M((Y-$\hat{Y}$)^2);

÷ rm3: media M((Y-$\hat{Y}$)^3);

÷ rm4: media M((Y-$\hat{Y}$)^4);

÷ rmx: media M($\hat{Y}^2$);

÷ rmy: media M(Y$\hat{Y}$);

÷ rv0: rm2/rm0^2;

÷ rg1: oblicitatea (g_1) estimatei ($\hat{Y}$);

÷ rg2: excesul de boltire (g_2) al estimatei ($\hat{Y}$);

÷ rjb: valoarea statisticii Jarque-Bera (JB) pentru estimată ($\hat{Y}$);

÷ rr1: r(Y,$\hat{Y}$);

÷ rr2: r^2(Y,$\hat{Y}$);

÷ rnalive: 0 (Ŷ nu este un fenotip viabil) sau 1 (Ŷ este un fenotip viabil);
÷ rr2_min: valoarea "r2_min" a scorului de selecție FS dat de Ŷ descriptorilor ce îl compun (min, max şi avg identice, referind minimul, media şi maximul aici cu o singură valoare);
÷ rse_min: valoarea alternativei "se_min" la scorul de selecție FS dat de Ŷ descriptorilor ce îl compun (min, max şi avg identice, referind minimul, media şi maximul aici cu o singură valoare);
÷ rMt_min: valoarea alternativei "Mt_min" la scorul de selecție FS dat de Ŷ descriptorilor ce îl compun (min, max şi avg identice, referind minimul, media şi maximul aici cu o singură valoare);
÷ rHr_min: valoarea alternativei "Hr_min" la scorul de selecție FS dat de Ŷ descriptorilor ce îl compun (min, max şi avg identice, referind minimul, media şi maximul aici cu o singură valoare);
÷ rr2_max: idem rr2_min;
÷ rse_max: idem rse_min;
÷ rMt_max: idem rMt_min;
÷ rHr_max: idem rHr_min;
÷ rr2_avg: idem rr2_min;
÷ rse_avg: idem rse_min;
÷ rMt_avg: idem rMt_min;
÷ rHr_avg: idem rHr_min;
÷ gnalive: în generațiile care produc evoluție (şi doar acestea sunt listate în configurația definită) este identic cu nrali; nrali păstrează informația cu privire la numărul de regresii obținute cu indivizii din cultivar în ultima (incluzând-o pe cea pe care o analizează la un moment dat) generație care a produs evoluție) în timp ce gnalive referă generația analizată la un anumit moment de timp;
÷ gr2_min: valoarea "r2_min" a scorului de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gse_min: valoarea alternativei "se_min" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gMt_min: valoarea alternativei "Mt_min" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gHr_min: valoarea alternativei "Hr_min" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gr2_max: valoarea alternativei "r2_max" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gse_max: valoarea alternativei "se_max" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gMt_max: valoarea alternativei "Mt_max" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gHr_max: valoarea alternativei "Hr_max" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gr2_avg: valoarea alternativei "r2_avg" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gse_avg: valoarea alternativei "se_avg" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gMt_avg: valoarea alternativei "Mt_avg" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ gHr_avg: valoarea alternativei "Hr_avg" la scorul de selecție FS dat de descriptorii prezenți în cultivar în generație;
÷ regression_equation: cea mai bună ecuație de regresie obținută până la un anumit moment de timp;

Informații cu privire la evoluție în fișierul PCB_XXXX_cfg.txt

÷ Partea de configurare (vezi şi Tabelul 1 - Anexa 5, Tabelul 2 - Anexa 5 & Tabelul 5 - Anexa 5):
- My_Data=172.27.211.5/MDFSARs[PCB_lkow_data;PCB_lkow_tmpx] - IP server baza de date, baza de date, tabelele de date;
- Genomes=Genes:[mp;fc;oi;id;ap;dm]/Addre:[fc;ap;id;oi;dm;mp] - definiția genomului şi relația de adresabilitate în baza de date (permutare a definiției genomului pentru calcul de adresare rapidă a datelor stocate în baza de date);
- G_Codes=mMnNSPsAaBbGgFfHhIi/mMDP/RrMmDd/DdOoPpQqJjKkLlVEWwFfSsTt/CHMEGQ/gt - lista codurilor de gene pentru fiecare genă a genomului;
- N_Sizes=Sample:12/CrossO:2/RegreM:4/MutGen:2 - volumul eşantionului; număr de încrucişări (implicând 2 genotipuri) per generație; număr de variabile în ecuația de regresie optimizată; număr de gene mutate atunci când mutația apare (probabilistic);
- Adapted=AbsVariance: 1.6353599751539131E-0003(0.100)/Jarque-Bera: 7.5766145463847375E+0000(1.000)/Determinate:0.100(%)/StudentTVal:[1.9721423227715891E+0000(201); 1.9720824746250930E+0000(202)] - v0 (corespondentul lui rv0 din PCB_XXXX_evo.txt) pentru observată (Y); jb (corespondentul lui rjb din PCB_XXXX_evo.txt) pentru observată (Y); coeficientul minim de determinare pentru calificarea de viabilitate; valoarea Student t din distribuția Student corespunzătoare unui model de regresie cu termen liber (df=nmol-RegreM-1) şi valoarea Student t din distribuția Student corespunzătoare unui model de regresie fără termen liber (df=nmol-RegreM) ce servesc drept valori de prag pentru semnificația statistică a unui parametru al ecuației de regresie pentru cele două cazuri (cu şi fără termen liber);
- Mutated=ParentsProb:0.050/ChildreProb:0.050 - probabilitatea (în procente) ca mutația să apară înainte de încrucişare (este aplicată părinților); probabilitatea (în procente) ca mutația să apară după încrucişare (este aplicată descendenților);
- Objectiv=Parameter:r2/To:max - care dintre alternativele de funcție obiectiv a fost aleasă în procesul de evoluție şi optimizare; care este obiectivul optimizării pentru funcția obiectiv aleasă;
- Selected=Parameter:r2_min/To:max/Using:tournament/Normalized?:No/SortedRank?:No/First_Rich?:Yes - care este observabila în selecție; care este obiectivul selecției; care este strategia de selecție; este strategia aplicată la valori normalizate între generații?; se folosesc ranguri în locul valorilor în strategie pentru observabilă?; prima generație se construieşte bogată în genotipuri (însemnând că generația inițială are toate genotipurile viabile)?
- Survival=Phenotyping:1.000/Genotyping:1.000/Using:tournament/SortedRank?:No - valoarea ponderii similarității fenotipice în scorul de supraviețuire; valoarea ponderii similarității genotipice în scorul de supraviețuire; care este strategia de selecție; este strategia aplicată la valori normalizate între generații?; se folosesc ranguri în locul valorilor în strategie pentru observabilă?;
- Evolution=nalive;r2_min;se_min;Mt_min;Hr_min;r2_max;se_max;Mt_max;Hr_max;r2_avg;se_avg;Mt_avg;Hr_avg; - parametrii urmăriți pe parcursul evoluției
- Alive=Genotypes:60337/Phenotypes(I):26355;/Phenotypes(A):26324;/Phenotypes(l):24085;/Phenotypes(L):29973;/Phenotypes(i):30178;/Phenotypes(a):30174; - număr total de viabili în întreaga populație în configurația de viabilitate definită: genotipuri şi reprezentări fenotipice ale acestora în cultivar (prin aplicarea celor 6 operatori de linearizare);
- Repeated=Sampling:45 - numărul de repetiții minus unu (valoarea următoare pragului limită de repetiții pentru construcția unei semnificații statistice);

÷ Repetition: numărul repetiției în execuția independentă de mai multe ori a evoluției de la o generație inițială oarecare prin aplicarea algoritmului genetic de un anumit număr de ori

(numărul de ori este parametrul e1n_GENERATIONS_max din fişierul de configurare);

÷ Found: numărul generației în repetiția curentă din în execuția independentă de mai multe ori a evoluției de la o generație inițială oarecare prin aplicarea algoritmului genetic de un anumit număr de ori (Found ≤ e1n_GENERATIONS_max din fişierul de configurare);

÷ restul parametrilor au corespondenți în fişierul PCB_XXXX_evo.txt: r2=rr2, se=rse, Mt=rMt, Hr=rHr, dr=rdr, me=rme, dt=rdt, m0=rm0, m1=rm1, m2=rm2, m3=rm3, m4=rm4, mx=rmx, my=rmy, v0=rv0, g1=rg1, g2=rg2, jb=rjb, r1=rr1, r2=rr2, nalive=rnalive, r2_min=rr2_min, se_min=rse_min, Mt_min=rMt_min, Hr_min=rHr_min, r2_max=rr2_max, se_max=rse_max, Mt_max=rMt_max, Hr_max=rHr_max, r2_avg=rr2_avg, se_avg=rse_avg, Mt_avg=rMt_avg, Hr_avg=rHr_avg, regression_equation= regression_equation).

Configurație program: Valorile parametrilor de execuție păstrate constante

Ceilalți parametrii ce configurează execuția algoritmului genetic au primit valori care au fost păstrate pe toată durata experimentului aceleaşi. Valorile acestora (care au fost date ținând seama de considerentele menționate mai sus) sunt redate în Tabelul 6-3.

Tabelul 6-3

Valori definite şi păstrate pe toată durata experimentului pentru parametrii rămaşi
Defined and keept constant values of the reminder parameters in the experiment

Parametru	Valoare
a_v_ADAPT_Variance	0.1
ajb_ADAPT_JarqueBera	1.0
a_c_ADAPT_Correlation	0.1
sn0_SAMPLE_Size	12
rn0_REGRESSION_Multiple	4
e1n_GENERATIONS_max	20000
g_r_GENERATIONS_first_rich	Yes
cn0_CROSSOVER_Pairs	2
m_m_MUTATION_Genes	2
mpp_MUTATION_Parent_probability	5%
mcp_MUTATION_Child_probability	5%
b_p_SELECTION_parameter	r2
b_o_SELECTION_objective	max
sfn_FITNESS_normalized	No
sfr_FITNESS_ranks	No
sfa_FITNESS_accuracy	10000
sff_FITTEST_function	r2_min
sfo_FITTEST_objective	max
fr2_FITTEST_r2_p	1.0
fse_FITTEST_se_p	1.0
fMt_FITTEST_Mt_p	1.0
fHr_FITTEST_Hr_p	1.0
v_p_SURVIVAL_phenotyping_p	1.0
v_g_SURVIVAL__genotyping_p	1.0
vfr_SURVIVAL_ranks	No
e0n_RUNS_number	46
b_k_RUNS_kepp_best_in_sample	Yes
b_f_RUNS_get_best_from_file	No

Design experimental:
Contingența selecției cu supraviețuirea în 46 de execuții independente ale algoritmului genetic

În implementarea algoritmului genetic s-au parametrizat o serie de variabile esențiale pentru evoluția acestuia, cât şi pentru portabilitatea sa:

÷ variabile esențiale pentru portabilitate:
  - definirea topologiei familiei de descriptori - prin parametrii din fişierul de configurare *c_galg.cgt* - care reprezintă în esență genotipizarea familiei de descriptori moleculari (*Genes*, *Addre* şi lista valorilor genelor ce variază în funcție de valoarea dată parametrilor *Genes* şi *Addre*);
  - definirea topologiei infrastructurii informaționale - prin parametrii definiți în fişierul de configurare *c_galg.cfg* - conectivitatea la baza de date ce conține populația studiată (*Host*, *User*, *Pass*, *Mydb*, *TabE*, *TabM*);

÷ variabile esențiale pentru evoluție - prin parametrii definiți în fişierul de configurare `c_galg.cga`, incluzând:
  - parametrizarea selecției, prin intermediul următoarelor valori de configurare:
    - definirea metodei de selecție (*sfs_FITNESS_strategy*) fiind implementate trei metode de selecție - proporțional (*proportional*), în turnir (*tournament*) şi deterministic (*deterministic*) - fiecare cu câte patru sub-variante date de contingența dintre *sfn_FITNESS_normalized* şi *sfr_FITNESS_ranks* - bazat pe valori (*normalized=No*, *ranks=No*), pe valori normalizate (*normalized=Yes*, *ranks=No*), ranguri (*normalized=No*, *ranks=Yes*), şi ranguri de valori normalizate (*normalized=Yes*, *ranks=Yes*);
    - definirea scorului de selecție (*sff_FITTEST_function*) fiind implementate treisprezece funcții de scor de selecție: *nalive* - număr de fenotipuri viabile ale genotipului, şi alte douăsprezece (în forma *XX_YYY*) din combinarea a trei reguli de selecție (*YYY=min* - minim, *YYY=max* - maxim şi *YYY=avg* - medie) aplicate la patru scoruri fenotipice (*XX=r2* - coeficient de determinare, *XX=se* - suma erorilor de estimare, *XX=Mt* - media semnificației Student t a parametrilor ecuației de regresie şi *XX=Hr* - entropia determinării); fiecare din cele patru scoruri fenotipice (*r2*, *se*, *Mt* şi *Hr*) este parametrizat de o valoare reală după cum urmează: pentru coeficientul de determinare puterea la care se ridică (*fr2_FITTEST_r2_p*), pentru suma erorilor de estimare puterea la care se ridică diferențele între măsurat şi estimat (*fse_FITTEST_se_p*), pentru media semnificațiilor parametrul familiei de medii Minkowski care defineşte tipul mediei (*fMt_FITTEST_Mt_p*), şi pentru entropie parametrul familiei de entropii Renyi care defineşte tipul entropiei (*fHr_FITTEST_Hr_p*);

- definirea obiectivului de selecție (*sfo_FITTEST_objective*) fiind implementate două obiective de selecție (*min* - minim şi *max* - maxim) aplicabile la scorurile de selecție (expresia definită de *sff_FITTEST_function*);
- definirea acurateței observației (*sfa_FITNESS_accuracy*), fiind implementată acuratețea de observație asupra scorului de selecție identică cu acuratețea de observare asupra scorului de supraviețuire;

o parametrizarea supraviețuirii, prin intermediul următoarelor valori de configurare:
- definirea metodei de supraviețuire (*vfs_SURVIVAL_strategy*) fiind implementate trei metode de supraviețuire - proporțional (*proportional*), în turnir (*tournament*) şi deterministic (*deterministic*) - fiecare cu câte două sub-variante - bazat pe valori (*vfr_SURVIVAL_ranks=No*) sau bazat pe ranguri (*vfr_SURVIVAL_ranks=Yes*);
- parametrizarea funcției de scor de supraviețuire fiind implementate două funcții de scor de supraviețuire bazate pe similaritate - genotipică - ce răspunde la întrebarea cât de diferite sunt două genotipuri (*v_g_SURVIVAL__genotyping_p*) şi fenotipică - ce răspunde la întrebarea cât de diferite sunt manifestările fenotipice a două genotipuri (*v_p_SURVIVAL_phenotyping_p*), fiecare dintre ele fiind parametrizată de o valoare reală care defineşte ponderea acestor similarități în scorul global de similaritate definit ca măsură aditivă a similarității genotipice şi fenotipice;
- definirea acurateței observației, fiind implementată acuratețea de observație asupra scorului de supraviețuire identică cu acuratețea de observare asupra scorului de selecție (*sfa_FITNESS_accuracy*);

o parametrizarea înmulțirii, prin intermediul următoarelor valori de configurare:
- definirea numărului de încrucişări între genotipuri permise într-o generație (*cn0_CROSSOVER_Pairs*);
- definirea numărului de gene care suferă mutație atunci când mutația are loc (*m_m_MUTATION_Genes*);
- definirea probabilității de apariție a mutației înainte de încrucişare (*mpp_MUTATION_Parent_probability*);
- definirea probabilității de apariție a mutației după încrucişare (*mcp_MUTATION_Child_probability*);

o parametrizarea materialului genetic şi a cultivarului, prin intermediul:
- definirii numărului de genotipuri înmulțite în cultivar (*sn0_SAMPLE_Size*);
- parametrizării viabilității (sau adaptării) fenotipurilor şi genotipurilor prin intermediul capacității minime de explicare (*a_c_ADAPT_Correlation*), variabilității minime (*a_v_ADAPT_Variance*) şi abaterii maxime de la normalitate

(*ajb_ADAPT_JarqueBera*);

- restricționării (*Yes*) sau nerestricționării (*No*) conținutului inițial al cultivarului la fenotipuri viabile (*g_r_GENERATIONS_first_rich*);
- condiționării (*Yes*) sau necondiționării (*No*) celui mai bun material genetic în cultivar (*b_k_RUNS_kepp_best_in_sample*);
- importării (*Yes*) sau ne-importării (*No*) de material genetic în prima generație (*b_f_RUNS_get_best_from_file*) din fișier de intrare (*c_galg.txt*);

o parametrizarea obiectivului evoluției, prin:

- definirea multiplicității regresiei multiple (*rn0_REGRESSION_Multiple*);
- definirea obiectivului evoluției (*b_p_SELECTION_parameter*) fiind implementate coeficientul de determinare între măsurat și estimat (*r2*), suma pătratelor erorilor de estimare (*se*), media aritmetică a semnificațiilor Student t ai parametrilor ecuației de regresie (*Mt*) și entropia Shannon a determinării de model (*Hr*);
- definirea scopului evoluției (*b_o_SELECTION_objective*) fiind implementate două scopuri minim (*min*) și maxim (*max*) aplicate obiectivului evoluției;

o parametrizarea evoluției, prin:

- definirea numărului maxim de generații (*e1n_GENERATIONS_max*);
- definirea numărului de execuții independente ale algoritmului genetic (*e0n_RUNS_number*);

o parametrizarea observației asupra materialului genetic și cultivarului prin definirea variabilelor monitorizate printr-o listă cu treizeci și patru de opțiuni de afișare cu privire la rezultatele de statistică descriptivă (*d_d_SHOW_descriptive_XX, XX = m0, m1, m2, m3, m4, mx, my, v0, g1, g2, jb, r1, r2*), valorile scorurilor de selecție *d_f_SHOW_fitness_YYYYYY, YYYYYY = nalive, r2_min, se_min, Mt_min, Hr_min, r2_max, se_max, Mt_max, Hr_max, r2_avg, se_avg, Mt_avg, Hr_avg*), configurație și rezultate finale de evoluție (*d_c_SHOW_configuration*), valori fenotipice pentru fiecare moleculă (*d_m_SHOW_mols*), ecuații de regresie obținute (*d_r_SHOW_regressions*), fenotipuri prezente în cultivar (*d_p_SHOW_phenotypes*), genotipuri ale fenotipurilor prezente în cultivar (*d_s_SHOW_genotypes*), scoruri de selecție (*d_t_SHOW_fittests*), generații (*d_g_SHOW_generations*) și evoluții (*d_e_SHOW_evolutions*) într-o listă de șapte fișiere posibile de ieșire (*cfg.txt, reg.txt, phe.txt, sam.txt, gen.txt, fit.txt, evo.txt*);

Algoritmul genetic a fost implementat (în limbajul *FreePascal*) și compilat formă de program executabil (devenind program evolutiv); face parte din categoria de aplicații client-server (execuție simultană pe aceeaşi sau pe mai multe maşini de calcul).

În evaluarea algoritmului genetic s-a considerat de importanță teoretică și practică

compararea performanțelor obținute pentru două elemente esențiale în procesul de evoluție şi anume strategia de selecție a indivizilor generatori de descendenți prin încrucişare şi mutație şi strategia de selecție a indivizilor pentru înlocuire de către descendenți (supraviețuire).

Principiul parsimoniei [29] este esența care stă la baza asocierii *Optimizare (SAR) → Evoluție (Observabile)*. Principiul a fost aplicat în simularea evoluției controlate către obiectivul evoluției folosind diferitele strategii de selecție şi supraviețuire. Principiul a avut ca scop evidențierea diferențelor în valorile observabilelor pe parcursul evoluției.

Conform cu acest principiu, am definit şi păstrat aceeaşi pe toată durata experimentului o serie de parametrii intrinseci execuției programului evolutiv şi evoluției algoritmului genetic (Tabelul 6-4).

Tabelul 6-4

Parametri de configurare în experimentul de execuție a programului evolutiv

The configuration parameters in the experiment of execution of the evolutionary program

Clasă *Class*	Parametru *Parameter*	Valoare *Value*
Topologia familiei de descriptori moleculari *Topology of the molecular descriptors family*	Genes	mp/fc/oi/id/ap/dm
	Addre	fc/ap/id/oi/dm/mp
	mp	mMnNSPsAaBbGgFfHhIi
	fc	mMDP
	oi	RrMmDd
	id	DdOoPpQqJjKkLlVEWwFfSsTt
	ap	CHMEGQ
	dm	gt
Topologia infrastructurii informaționale *Topology of the informational infrastructure*	Mydb	MDFSARs
	TabM	PCB_lkow_data
	TabE	PCB_lkow_tmpx
Material genetic şi cultivar *Genetic material and the cultivar*	sn0_SAMPLE_Size	12
	a_v_ADAPT_Variance	0.1
	ajb_ADAPT_JarqueBera	0.1
	a_c_ADAPT_Correlation	0.1
	g_r_GENERATIONS_first_rich	Yes
	b_k_RUNS_kepp_best_in_sample	Yes
	b_f_RUNS_get_best_from_file	No
Înmulțire *Breading*	cn0_CROSSOVER_Pairs	2
	m_m_MUTATION_Genes	2
	mpp_MUTATION_Parent_probability	5%
	mcp_MUTATION_Child_probability	5%
Obiectivul evoluției *Evolution objective*	rn0_REGRESSION_Multiple	4
	b_p_SELECTION_parameter	r2
	b_o_SELECTION_objective	max
Evoluție *Evolution*	e1n_GENERATIONS_max	20000
	e0n_RUNS_number	46
Selecție *Selection*	sfn_FITNESS_normalized	No
	sfr_FITNESS_ranks	No
	sfa_FITNESS_accuracy	10000
	sff_FITTEST_function	r2_min
	sfo_FITTEST_objective	max
	fr2_FITTEST_r2_p	1.0
	fse_FITTEST_se_p	1.0
	fMt_FITTEST_Mt_p	1.0
	fHr_FITTEST_Hr_p	1.0
Supraviețuire *Survival*	v_p_SURVIVAL_phenotyping_p	1.0
	v_g_SURVIVAL__genotyping_p	1.0
	vfr_SURVIVAL_ranks	No

[29] parsimonie - adoptarea celor mai simple presupuneri în formularea teoriei sau interpretarea datelor, în special în acord cu regula lamei de ras a lui Ockham (principiu atribuit logicianului William of OCKHAM, care subliniază că trebuie eliminate toate acele presupuneri care nu fac nici o diferență în predicțiile observate ale ipotezelor explicatoare sau teoriei); în latină: *lex parsimoniae - entia non sunt multiplicanda praeter necessitatem*.

Doi parametrii: strategia de selecție (*sfs_FITNESS_strategy* în fișierul de configurare - vezi Tabelul 6-1) și strategia de supraviețuire (*vfs_SURVIVAL_strategy* în fișierul de configurare - vezi Tabelul 6-1) au luat pe rând valorile: *proportional* (pentru strategie proporțională - vezi Tabelul 5-4), *deterministic* (pentru strategie deterministă - vezi Tabelul 5-4) și *tournament* (pentru strategie în turnir - vezi Tabelul 5-4). S-au proiectat astfel nouă execuții ale programului evolutiv, în fiecare din cele nouă execuții dând valori diferite celor doi parametrii ce definesc cele două strategii (Tabelul 6-5), ceilalți parametrii rămânând cu aceleași valori (Tabelul 6-4) pe toată durata experimentului.

Tabelul 6-5

Strategii de selecție și supraviețuire: design experimental de execuție program
Strategies of selection and survival: experimental design of program execution

Supraviețuire *Survival* / Selecție *Selection*	Proporțional (P) SV_Pro *Proportional*	Deterministic (D) SV_Det *Deterministic*	Turnir (T) SV_Tur *Tournament*
Proporțional (P) - SL_Pro - *Proportional*	P:P (1)	P:D (2)	P:T (3)
Deterministic (D) - SL_Det - *Deterministic*	D:P (4)	D:D (5)	D:T (6)
Turnir (T) - SL_Tur - *Tournament*	T:P (7)	T:D (8)	T:T (9)

7. Implementare algoritm genetic; rezultate & verificarea datelor

Strategii pentru selecție și supraviețuire: Proporțional, Turnir, Deterministic

Pentru a realiza o selecție bazată pe un scor de selecție se impune parcurgerea unei serii de pași. Algoritmul ***FS*** (Algoritm 1 - pentru scorurile de selecție) parcurge o astfel de serie de pași și realizează astfel premizele pentru aplicarea unei strategii de selecție.

În ceea ce privește numărul de selecții (notat N_Sel în algoritmii de mai jos), pentru fiecare încrucișare sunt necesare două selecții, deci N_Sel este totdeauna număr par. Numărul de încrucișări nu este un parametru în programul evolutiv ce implementează algoritmul genetic (deși ar fi putut fi definit ca atare în fișierul g_alg.cga). Există mai multe condiționări asupra acestui număr cum ar fi că trebuie să fie mai mic sau egal cu numărul de genotipuri viabile reprezentate în cultivar n_of_alive_g - număr care este totdeauna mai mic sau egal cu numărul de genotipuri din eșantion n_of_sample). Aceeași serie de condiționări se aplică și numărului de descriptori din ecuația de regresie n_of_reg, astfel încât soluția de implementare aleasă este că numărul de selecții N_Sel este cel mai mare număr par mai mic decât n_of_reg, număr care se obține după formula: $2*[^{n_of_reg-1}/_{2}]$.

Algoritmul ***PS*** (Algoritm 2 - pentru strategia proporțională) realizează o selecție proporțională folosind un șir de scoruri de selecție și dă o șansă de selecție proporțională cu scorul de selecție.

Algoritmul ***DS*** (Algoritm 3 - pentru strategia deterministă) realizează o selecție deterministă folosind un șir de scoruri de selecție extrăgând cele mai mari N_Sel scoruri de selecție.

Algoritmul ***TS*** (Algoritm 4 - pentru strategia în turnir) realizează o selecție în turnir folosind un șir de scoruri de selecție extrăgând acele valori care se califică în urma unui turnir între 2 valori candidate repetat de un număr de ori dat de N_Sel.

Scorul de supraviețuire (***VS***) este o valoare compozită, menită să asigure deopotrivă diversitatea genotipică și cea fenotipică în cultivar. Pentru aceasta două măsuri de similaritate între doi descriptori intră în expresia scorului de supraviețuire: o măsură de similaritate fenotipică (***VSP*** - dată de diferența între valorile scorurilor de selecție) și o măsură de similaritate genotipică (***VSG*** - dată de diferența între codurile genetice ale celor doi descriptori). Tabelul 7-1 redă expresiile de calcul folosite.

Odată obținute valorile ***VS*** ale scorurilor de supraviețuire în modalitatea expusă în tabelul de mai sus (în care scara de similaritate individuală are acum același obiectiv - cele mai mari valori dau cei mai potenți candidați pentru înlocuire în materialul genetic) pentru șirul de indivizi reprezentați în cultivar, algoritmul proiectat pentru pregătirea scorurilor de selecție ***FS*** este legitimat a fi folosit și pentru scorul de supraviețuire ***VS***, ceea ce a și fost făcut (motiv pentru care nu există descris un

algoritm separat pentru ***VS***). Idem algoritmii ce implementează strategiile proporțională, deterministă şi în turnir sunt perfect legitimați pentru a fi folosiți şi pentru şirul scorurilor de supraviețuire (***PV*** = ***PS***, ***DV*** = ***DS***, ***TV*** = ***TS***) - şi nici pentru aceştia nu există descriere separată.

Tabelul 7-1

Scorul de supraviețuire - instrument al păstrării variabilității
Survival score - the tool for keeping the variability

Scor *Score*	Semnificație *Meaning*	Obiectiv *Objective*	Remarci *Remarks*
$VSP_q(X_i, X_j) = \lvert f(X_i) - f(X_j) \rvert^q$	disimilaritatea fenotipică	minimum	f(·) este scorul de selecție
$VSG_r(X_i, X_j) = \left(\frac{NCD(X_i, X_j)}{NC} \right)^r$	disimilaritatea genotipică	minimum	NCD(·,·) - număr de valori de gene diferite NC - număr de gene în cromozom
$VS(X_i, X_j) = \frac{2}{VSP_q(X_i, X_j) + VSG_r(X_i, X_j)}$	similaritatea perechii	maximum	VS(·,·) este o măsură a şansei
$VS(X_i) = \min_{1 \le j \le n}^{j \ne i} VS(X_i, X_j)$	similaritatea unui individ	maximum	Cele mai mici diferențe (cu alții) definesc scorul individual

Algoritm 1. Algoritmul ***FS*** pentru obținerea scorului de selecție

Scorul de selecție se calculează astfel:
- Date de intrare: O matrice conținând valorile calculate ale fiecărui descriptor viabil (din eşantionul de dimensiune n_of_alive_p) şi moleculă (din setul de dimensiune n_of_mol)
- Calculează (folosind un modul de regresie multiplă) toate regresiile posibile pentru activitatea măsurată Y cu n_of_reg descriptori (în număr de combinări de n_of_alive_p luate câte n_of_reg) obținând activitatea estimată Ŷ;
- Aplică adaptarea (folosind valorile definite pentru adaptare în variabilele ADAPT_Variance, ADAPT_JarqueBera şi ADAPT_Correlation din fişierul de configurare g_alg.cga) şi construieşte o listă cu regresiile (asocierile fenotipice) adaptate;
- Calculează scorul de selecție pentru fiecare fenotip (din cele n_of_alive_p) pe baza valorilor obținute în regresiile (asocierile) ce îl conțin (folosind valorile definite pentru selecție în variabilele SELECTION_parameter şi SELECTION_score din fişierul de configurare g_alg.cga)
- Calculează scorul de selecție pentru fiecare genotip (din cele n_of_sample) pe baza valorilor obținute pentru fenotipurile generate de acesta (prin unul din cei 6 operatori de linearizare) folosind aceeaşi valoare definită pentru selecție în variabila SELECTION_score din fişierul de configurare g_alg.cga)
- Dacă SELECTION_normalized=Yes în fişierul de configurare g_alg.cga atunci normalizează cele n_of_sample valori obținute;
- Rotunjeşte la numărul de cifre semnificative dat de SELECTION_accuracy în fişierul de configurare g_alg.cga cele n_of_sample valori obținute;
- Calculează rangurile pentru cele n_of_sample valori obținute;
- Dacă SELECTION_ranks=Yes în fişierul de configurare g_alg.cga atunci înlocuieşte cele n_of_sample valori obținute cu rangurile acestora obținute mai sus;
- Ordonează (păstrând asocierea cu genotipurile din eşantion) cele n_of_sample valori obținute;
- Date de ieşire:
  - FS_Array - şirul scorurilor de selecție pentru fiecare genotip (n_of_sample valori)
  - FSD_Array - şirul valorilor distincte de scor de selecție (cel mult n_of_sample valori)
  - FSC_Array - şirul numărului de apariții pentru fiecare scor distinct de selecție

Algoritm 2. Algoritmul ***SP*** ce realizează o strategie de selecție proporțională

Selecția proporțională se realizează astfel:
- Date de intrare:
  - FS_Array - şirul scorurilor de selecție pentru fiecare genotip (n_of_sample valori)
  - FSD_Array - şirul valorilor distincte de scor de selecție (cel mult n_of_sample valori)
  - FSC_Array - şirul numărului de apariții pentru fiecare scor distinct de selecție
  - N_Sel - număr de selecții pentru încrucişare şi mutație;

```
├ Inițializează Selected_Genotypes_Array la ∅ (mulțimea vidă);
├ Pentru fiecare selecție (de la 1 la N_Sel)
        ├ Calculează suma scorurilor pentru genotipurile încă neselectate în FS_Sum;
        ├ Generează întâmplător (distribuție uniformă) un număr FS_Freq între 0 și FS_Sum (inclusiv);
        ├ Găsește primul (indice) Group din FSD_Array pentru care FS_Freq ≤ Σ_{i≤Group} FSD_Array[i]*FSC_Array[i];
        ├ Generează întâmplător (distribuție uniformă) un număr FSD_Next între 1 și FSC_Array[Group] (inclusiv)
        ├ Adaugă a FSD_Next-a valoare egală cu FSD_Array[Group] neselectată încă din FS_Array în
              Selected_Genotypes_Array și scade o unitate din FSC_Array[Group];
├ Sfârșit 'Pentru'
┌ Date de ieșire: Selected_Genotypes_Array - șirul genotipurilor selectate (în număr de N_Sel)
```

Algoritm 3. Algoritmul ***SD*** ce realizează o strategie de selecție deterministă

```
Strategia deterministă se realizează astfel:
└ Date de intrare:
÷ FS_Array - șirul scorurilor de selecție pentru fiecare genotip (n_of_sample valori)
÷ FSD_Array - șirul valorilor distincte de scor de selecție (cel mult n_of_sample valori)
÷ FSC_Array - șirul numărului de apariții pentru fiecare scor distinct de selecție
÷ N_Sel - număr de selecții pentru încrucișare și mutație;
├ Inițializează Selected_Genotypes_Array la ∅ (mulțimea vidă);
├ Inițializează Already_Selected la 0;
├ Inițializează Group la n_of_sample;
├ Cât timp Already_Selected + FSC_Array[Group] ≤ N_Sel
        ├ Pune primul indice din FS_Array egal cu FSD_Array[Group] în Selected_Genotypes_Array;
        ├ Dacă FSC_Array[Group] > 0 atunci dec(FSC_Array[Group]) altfel dec(Group);
├ Sfârșit 'Cât timp'
├ Cât timp Already_Selected ≤ N_Sel (au mai rămas câteva scoruri identice într-un ultim grup din care trebuie făcută o
      selecție):
        ├ Generează întâmplător (distribuție uniformă) un număr FSD_Next între 1 și FSC_Array[Group] (inclusiv)
        ├ Adaugă a FSD_Next-a valoare egală cu FSD_Array[Group] neselectată încă din FS_Array în
              Selected_Genotypes_Array și scade o unitate din FSC_Array[Group];
├ Sfârșit 'Cât timp'
┌ Date de ieșire: Selected_Genotypes_Array - șirul genotipurilor selectate (în număr de N_Sel)
```

Algoritm 4. Algoritmul ***ST*** ce realizează o strategie de selecție în turnir

```
Strategia în turnir se realizează astfel:
└ Date de intrare:
÷ FS_Array - șirul scorurilor de selecție pentru fiecare genotip (n_of_sample valori)
÷ FSD_Array - șirul valorilor distincte de scor de selecție (cel mult n_of_sample valori)
÷ FSC_Array - șirul numărului de apariții pentru fiecare scor distinct de selecție
÷ N_Sel - număr de selecții pentru încrucișare și mutație;
├ Inițializează Selected_Genotypes_Array la o permutare întâmplătoare (distribuție uniformă) a mulțimii
      {1..n_of_sample}
├ Pentru fiecare i_Sel de la 2 la N_Sel (primele N_Sel genotipuri din permutare concurează în turnir):
        ├ Dacă FS_Array[i_Sel] ≤ FS_Array[i_Sel-1] atunci
            ├ Dacă FS_Array[i_Sel] == FS_Array[i_Sel-1] atunci dacă Random({0,1}) == 0 atunci continuă de la
                  începutul iterației 'Pentru';
            ├ Permută în FS_Array valorile de pe pozițiile i_Sel & i_Sel-1;
        ├ Sfârșit 'Dacă'
├ Sfârșit 'Pentru'
├ Dacă N_Sel < n_of_sample atunci (ultimul nu a participat în turnir și încă mai sunt genotipuri cu care să concureze în
      eșantion)
        ├ Generează întâmplător (distribuție uniformă) un număr i_Sel între N_Sel + 1 și n_of_sample;
        ├ Dacă FS_Array[N_Sel] ≤ FS_Array[i_Sel] atunci
            ├ Dacă FS_Array[N_Sel] == FS_Array[i_Sel] atunci
                  ├ Dacă Random({0,1}) == 0 atunci  Stop (turnir complet);
            ├ Permută în FS_Array valorile de pe pozițiile N_Sel & i_Sel;
        ├ Sfârșit 'Dacă'
├ Sfârșit 'Dacă'
┌ Date de ieșire: Selected_Genotypes_Array - șirul genotipurilor selectate (în număr de N_Sel)
```

Implementarea strategiilor de selecție și supraviețuire au pus în evidență o serie de situații care nu au fost anticipate până în acest punct al proiectării algoritmului și/sau realizării programului.

Un exemplu de acest tip care a trebuit să fie tratat în implementare referă selecția proporțională, în care șansa de selecție este proporțională cu valoarea scorului, și există posibilitatea teoretică de a avea scoruri de selecție nule ca date de intrare în algoritm (poate chiar toate să fie nule la un anumit moment al execuției) și în aceste condiții algoritmul să producă totuși o selecție (și aceasta să fie proporțională).

Soluția de proiectare aleasă, în care construcția șirurilor scorurilor de selecție (FS) și de supraviețuire (VS) a permis aplicarea aceloraşi algoritmi pentru ambele operații (de selecție și de supraviețuire) creează o contingență a celor trei strategii (proporțional, determinist, în turnir) cu cele două operații (selecție și supraviețuire). Ea este argumentul principal al tranziției *Optimizare* → *Simulare* care stă la baza analizei observabilelor la contingența dintre selecție și supraviețuire din secțiunile de analiză statistică.

Rezultate: Fişiere rezultat

Execuția programului evolutiv s-a făcut pe calculatoare din generația P6 (Dual P5) în perioada Ianuarie - Februarie 2009 și rezultatele au fost salvate de program în fișierele date în Tabelul 7-2.

Tabelul 7-2

Fişiere rezultat (configurare şi evoluție) după designul experimental din Tabelul 6-5
Result files (configuration and evolution) by the experimental design from Table 6-5

Selecție *Selection*	Supraviețuire *Survival*	Configurare *Configuration*	Evoluție *Evolution*
Proporțional *Proportional*	Proporțional *Proportional*	PCB_4044_cfg.txt	PCB_4044_evo.txt
Proporțional *Proportional*	Deterministic *Deterministic*	PCB_2441_cfg.txt	PCB_2441_evo.txt
Proporțional *Proportional*	Turnir *Tournament*	PCB_9878_cfg.txt	PCB_9878_cfg.txt
Deterministic	Proporțional *Proportional*	PCB_5108_cfg.txt	PCB_5108_evo.txt
Deterministic	Deterministic *Deterministic*	PCB_6369_cfg.txt	PCB_6369_evo.txt
Deterministic	Turnir *Tournament*	PCB_6690_cfg.txt	PCB_6690_evo.txt
Turnir *Tournament*	Proporțional *Proportional*	PCB_5828_cfg.txt	PCB_5828_evo.txt
Turnir *Tournament*	Deterministic *Deterministic*	PCB_4872_cfg.txt	PCB_4872_evo.txt
Turnir *Tournament*	Turnir *Tournament*	PCB_1758_cfg.txt	PCB_1758_evo.txt

Disponibile pentru descărcare de la adresa:
Available for download from the address:
http://l.academicdirect.org/Horticulture/GAs/MLR_MDF_selection_vs_survival/

Verificarea datelor: Testul Benford asupra numărului de genotipuri reprezentate în cultivar

Tabelul 7-3 cumulează rezultatele obținute pentru numărul de genotipuri viabile observate (*sum_obs*) în numărul de evoluții observate (*num_obs*) grupate după mii de generații (de la 0..1000

până la 19001-20000). Figurile 7-1(a-f) ilustrează comparațiile între probabilitățile de distribuție teoretice (Benford) și proporțiile observate ale cifrelor numerelor din Tabelul 7-3. Tabelul 7-4 conține experimentul de verificare a ipotezei de distribuție a primelor trei cifre după legea Benford folosind statistica χ^2 iar Tabelul 7-5 statistica Kolmogorov-Smirnov.

Tabelul 7-3

Frecvențe cumulate din 46 de execuții independente ale numărului (num_obs) și sumei (sum_obs) aparițiilor genotipurilor viabile reprezentate în cultivar în generațiile ce au produs evoluție pentru fiecare asociere de selecție (Sel) și supraviețuire (Srv): Srv, Sel ∈ {P, T, D}

Cumulated frequencies from 46 independent executions for the number (num_obs) and the sum (sum_obs) of the viable genotypes occurrences being represented in cultivar in the generations in which the evolution were recorded for every association of selection (Sel) and survival (Srv): Srv, Sel ∈{P, T, D}

1000G	SelSrv	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
num_obs	PP	732	141	97	62	36	44	25	30	35	32	29	16	20	15	8	7	12	21	17	20
sum_obs	PP	8366	1580	1073	700	417	500	285	339	390	358	330	186	221	166	92	84	134	245	201	235
num_obs	PT	713	171	133	89	54	44	40	33	32	32	11	28	32	26	10	21	6	8	11	15
sum_obs	PT	8207	1960	1505	1013	609	505	455	383	363	369	127	313	362	279	111	239	69	90	128	161
num_obs	PD	748	159	101	85	60	71	24	25	28	21	22	33	24	25	10	15	15	22	14	18
sum_obs	PD	8825	1858	1166	990	702	829	273	286	326	243	257	384	278	292	116	171	176	256	162	207
num_obs	TP	655	159	110	88	57	31	37	33	34	27	25	23	17	9	7	8	9	12	10	13
sum_obs	TP	7470	1741	1202	992	639	331	410	363	372	300	278	262	191	100	79	92	99	134	115	147
num_obs	TT	740	178	92	62	56	37	38	31	21	23	19	25	14	17	18	25	9	1	7	12
sum_obs	TT	8475	1988	1025	700	625	424	427	352	237	248	218	269	159	195	199	278	102	12	82	136
num_obs	TD	757	110	98	83	61	66	49	39	21	31	29	25	31	16	5	4	14	9	11	16
sum_obs	TD	8969	1282	1145	963	704	769	569	453	244	364	333	291	364	187	59	48	165	105	129	188
num_obs	DP	422	94	44	33	43	28	15	43	31	15	15	17	12	17	16	9	15	17	6	5
sum_obs	DP	4739	987	450	349	467	289	150	451	344	167	166	183	116	163	157	99	150	190	68	57
num_obs	DT	431	110	65	51	48	39	38	45	36	43	14	25	29	13	8	16	8	10	11	2
sum_obs	DT	4883	1223	719	558	533	440	411	477	385	468	153	273	312	139	92	169	87	111	119	20
num_obs	DD	466	120	81	66	53	41	24	39	51	24	11	19	18	23	23	14	19	12	24	2
sum_obs	DD	5511	1402	949	772	627	486	283	459	596	283	130	225	212	271	270	166	222	143	285	24

num_obs: număr de observații (*observation's number*); sum_obs: suma observațiilor (*sum of observations*); 1000G = 1000 de generații (*1000 generations*)

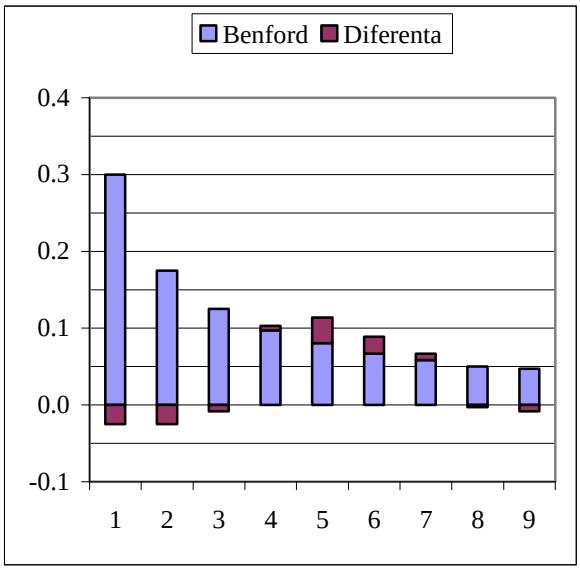
Tabelul 7-4

Testul χ^2 aplicat frecvențelor observate din Tabelul 7-3 pentru respingerea ipotezei că prima, a doua și a treia cifră semnificativă a numerelor urmează distribuția Benford

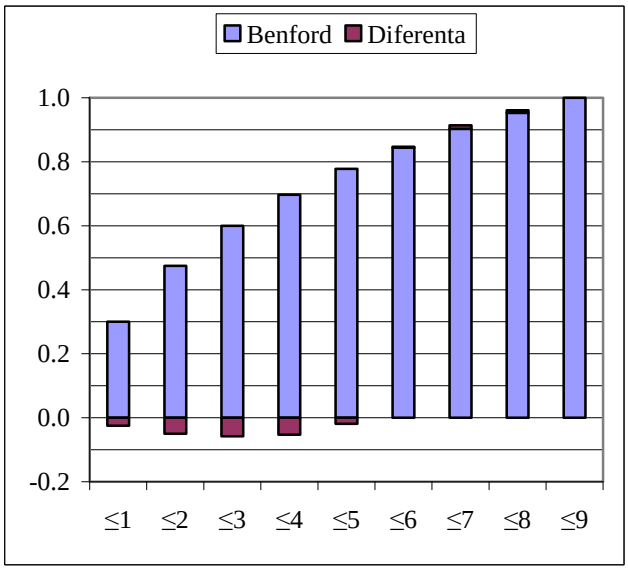
χ^2 test applied on the observed frequencies given in Table 7-3 to reject the hypothesis that first, second and third significant digit of the numbers follows the Benford distribution

Cifra			Frecvență așteptată			Frecvență observată			$(O_i-E_i)^2$			$(O_i-E_i)^2/E_i$		
d_0	d_1	d_2	d_0	d_1	d_2	d_0	d_1	d_2	d_0	d_1	d_2	d_0	d_1	d_2
0	0	0	-	40	19	-	28	25	-	144	36	-	3.60	1.89
1	1	1	108	38	19	117	41	18	81	9	1	0.75	0.24	0.05
2	2	2	63	37	18	72	37	18	81	0	0	1.29	0.00	0.00
3	3	3	45	35	18	48	33	20	9	4	4	0.20	0.11	0.22
4	4	4	35	34	18	33	34	11	4	0	49	0.11	0.00	2.72
5	5	5	29	33	18	17	42	15	144	81	9	4.97	2.45	0.50
6	6	6	24	32	18	16	34	18	64	4	0	2.67	0.13	0.00
7	7	7	21	31	18	18	30	18	9	1	0	0.43	0.03	0.00
8	8	8	18	30	18	19	31	16	1	1	4	0.06	0.03	0.22
9	9	9	17	29	18	20	29	23	9	0	25	0.53	0.00	1.39
Σ	Σ	Σ	360	339	182	360	339	182	402	244	128	11.0	6.60	7.00

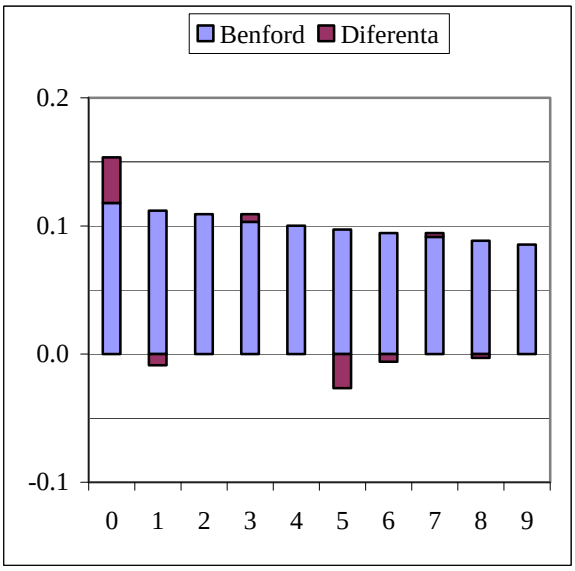
Numărul gradelor de libertate ale legii de distribuție Benford este 1 (baza de numerație, 10)
d_0: $X^2=11<14.7=\chi^2(9\text{-}2, 5\%)$; d_1: $X^2=6.6<15.5=\chi^2(10\text{-}2, 5\%)$; d_2: $X^2=7<15.5=\chi^2(10\text{-}2, 5\%)$;



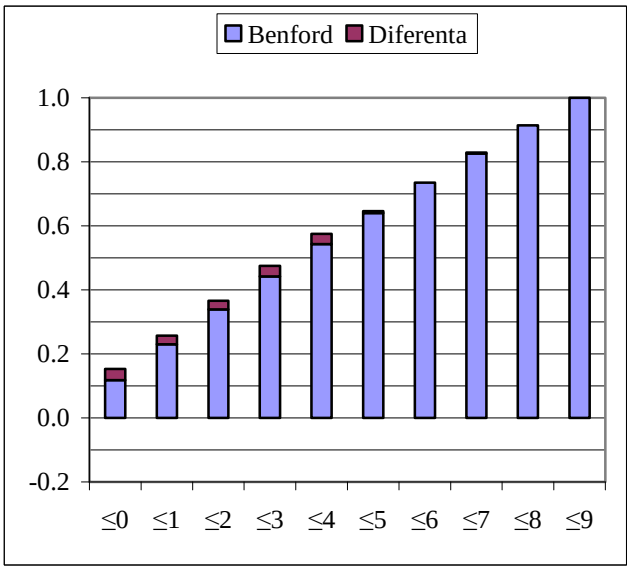
(a) d_0: Funcția de distribuție de probabilitate
Probability density function (PDF)



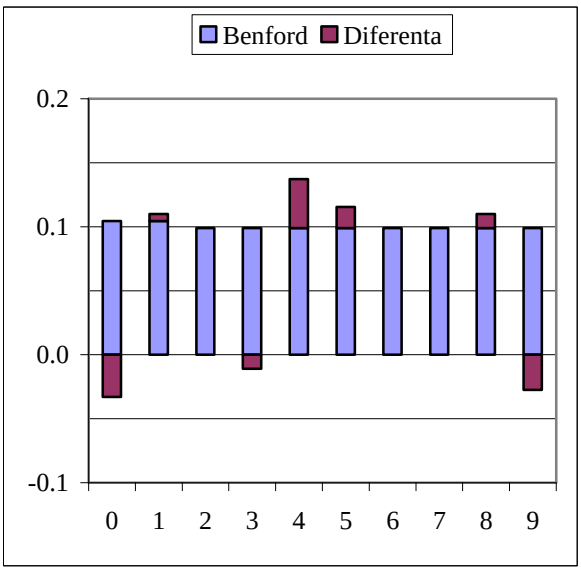
(b) d_0: Funcția cumulativă de probabilitate
Cummulative distribution function (CDF)



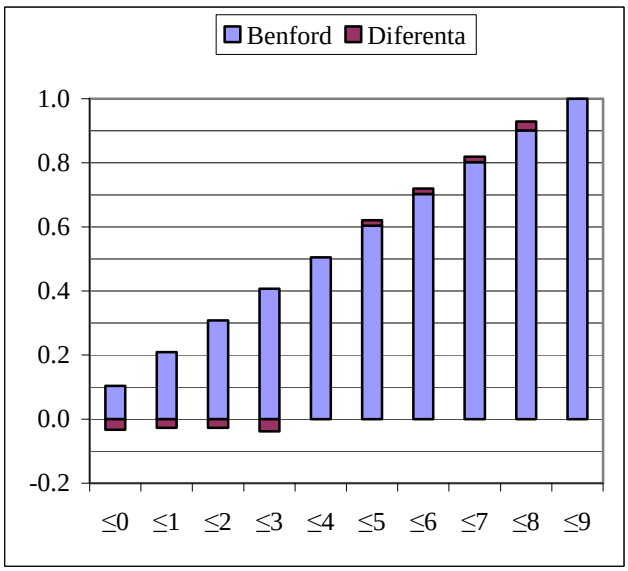
(c) d_1: Funcția de distribuție de probabilitate (*PDF*)



(d) d_1: Funcția cumulativă de probabilitate (*CDF*)



(e) d_2: Funcția de distribuție (*PDF*)



(f) d_2: Funcția cumulativă (*CDF*)

Figura 7-1. Distribuția teoretică (Benford) şi distribuția observată (datele din Tabelul 7-3)
Theoretical distribution (Benford) and observed distribution (the data from Table 7-3)

Tabelul 7-5

Testul K-S aplicat frecvențelor observate din Tabelul 7-3 pentru respingerea ipotezei că prima, a doua și a treia cifră semnificativă a numerelor urmează distribuția Benford

K-S test applied on the observed frequencies in Table 7-3 to reject the hypothesis that first, second and third digits follows Benford distribution

Cifra *Digit*			Frecvență cumulată așteptată și observată *Cummulated and observed frequency*						Diferență *Difference*			| Diferență | *|Diference|*		
d_0	d_1	d_2	d_0a	d_1a	d_2a	d_0o	d_1o	d_2o	d_0	d_1	d_2	d_0	d_1	d_2
0	0	0	0	40	19	0	28	25	0	12	-6	0	12	6
1	1	1	108	78	38	117	69	43	-9	9	-5	9	9	5
2	2	2	171	115	56	189	106	61	-18	9	-5	18	9	5
3	3	3	216	150	74	237	139	81	-21	11	-7	21	11	7
4	4	4	251	184	92	270	173	92	-19	11	0	19	11	0
5	5	5	280	217	110	287	215	107	-7	2	3	7	2	3
6	6	6	304	249	128	303	249	125	1	0	3	1	0	3
7	7	7	325	280	146	321	279	143	4	1	3	4	1	3
8	8	8	343	310	164	340	310	159	3	0	5	3	0	5
9	9	9	360	339	182	360	339	182	0	0	0	0	0	0
Σ	Σ	Σ	-	-	-	-	-	-	-66	55	-9	82	55	37

d_0: $D\sqrt{n} = \frac{21}{360}\sqrt{9} = \frac{14}{80} < \frac{31}{80} = K(9, 5\%); \frac{14}{80} = K(9, 90.82\%)$

d_1: $D\sqrt{n} = \frac{12}{339}\sqrt{10} \approx \frac{17}{152} < \frac{56}{152} \approx K(10, 5\%); \frac{17}{152} = K(10, 95.16\%)$

d_2: $D\sqrt{n} = \frac{7}{182}\sqrt{10} \approx \frac{62}{510} < \frac{188}{510} \approx K(10, 5\%); \frac{62}{510} = K(10, 94.60\%)$

Valorile χ^2 obținute în Tabelul 7-4 (și ilustrate în Figura 7-1) arată că pentru fiecare dintre primele trei cifre semnificative nu se poate respinge semnificativ statistic ipoteza de distribuție după legea Benford la un nivel de semnificație de 5%. Pentru ca să se poată asigura semnificativ statistic distribuția cifrelor după legea Benford, ar fi fost necesar ca valoarea X^2 să fi fost de cel mult 2.17 = $\chi^2(7, 95\%)$ pentru prima cifră, și de cel mult 2.73 = $\chi^2(8, 95\%)$ pentru următoarele două cifre semnificative ale numerelor.

Se impune o măsură prealabilă în aplicarea testului χ^2 între observat și așteptat la datele din Tabelul 7-4, și anume asigurarea apropierii de normalitate a distribuției pătratelor diferențelor (Fisher, 1920-Accuracy). Astfel, dacă valoarea $b_2 = m_4/m_2^2$ este apropiată de 3, atunci distribuția pătratelor erorilor se poate aproxima de legea de distribuție normală și pătratele diferențelor constituie o statistică suficientă (Fisher, 1922-Estimation); dacă b_2 este apropiată de 6, atunci distribuția pătratelor erorilor se poate aproxima de legea de distribuție dublu exponențială și suma diferențelor modulelor constituie o statistică suficientă. Pentru diferențele din Tabelul 7-4 valorile β_2 sunt: $\beta_2(d_0) = 2.13$; $\beta_2(d_1) = 4.60$; $\beta_2(d_2) = 2.71$ cu o valoare medie de 3.15, deci aplicarea testului χ^2 pentru pătratele diferențelor este consistentă cu legea de distribuție a acestora.

Concluzia analizei din Tabelul 7-4 este că nu se poate respinge ipoteza distribuției Benford cu o probabilitate de 95% (la nivelul de semnificație de 5%) pentru datele prezentate în Tabelul 7-3. Mai mult, valorile probabilităților din distribuția Kolmogorov arată că un număr relativ scăzut de date experimentale concordă mai bine decât cele din Tabelul 7-3 cu legea de distribuție Benford (9.18% pentru d_0; 4.84% pentru d_1; 5.4% pentru d_2).

8. Analiza variabilității:
Cu privire la genotipurile reprezentate în cultivar

Frecvența de apariție a genotipurilor în eșantion de-a lungul evoluțiilor permite aprecieri cu privire la capacitatea de adaptare a acestora, și în același timp o măsură a variabilității materialului genetic al eșantionului pe care o induc metoda de selecție și metoda de supraviețuire.

Numărul de genotipuri viabile (reprezentate fenotipic în cultivar) s-a urmărit folosind schema de contingență {*Top23*, *Total*}×{*Num*, *Apar*, *Part*}, unde *Top23* - referind cele mai frecvente 23 iar *Total* - toate genotipurile reprezentate în cultivar în 46 de execuții independente, *Num* - numărul de genotipuri distincte, *Apar* - suma numărului de genotipuri și *Part* - numărul de genotipuri ale fenotipurilor participante în regresii (valori date în Tabelul 8-1).

Tabelul 8-1

Cele mai frecvente genotipuri prezente în generații care au produs evoluții în 46 de execuții independente ale programului evolutiv

Most frequent genotypes in the generations producing evolution in the 46 runs of the genetic algorithm

Sel	Srv	Geno	Nr	Apar	Particip
P	P	**Top 23 apariții**	**13**	**406**	**389**
		mMdlHg	1	46	43
		MDMKHt	1	40	39
		nDRLHt	1	40	39
		iPDKCg	1	39	39
		ADDJCg	1	35	35
		mDdjGg	1	31	30
		bDDDGg	1	28	19
		bDDJCg	1	27	27
		sDdLHg	1	25	25
		BDDDGg	1	24	22
		bDMLEg	1	24	24
		bDMLGg	1	24	24
		MMDPMt	1	23	23
		Total	**6760**	**16788**	**15902**
	D	**Top 23 apariții**	**13**	**378**	**371**
		iPMDHg	1	39	37
		bPRjCg	1	38	38
		IPMDEg	1	37	36
		mMdoHt	1	30	29
		IPRKCg	1	29	29
		MDRLHt	1	29	29
		MMdlHg	1	29	29
		MDmWHg	1	26	26
		BPRjCg	1	26	25
		NDRlHt	1	25	25
		iPMDCg	1	24	23
		bmrVCt	1	23	23
		IPMDCg	1	23	22
		Total	**8070**	**18240**	**17797**
	T	**Top 23 apariții**	**6**	**214**	**207**
		MMdlHg	1	47	47
		mMdlHg	1	46	43
		sPDLEg	1	38	38
		AMdwGg	1	29	29
		IPMDHg	1	29	27
		mMdqGt	1	25	23
		Total	**7466**	**16599**	**15739**
D	P	**Top 23 apariții**	**3**	**89**	**72**
		MDRLHt	1	31	31
		ImrWCg	1	30	19
		ImrWHg	1	28	22
		Total	**3922**	**10764**	**9742**
D	D	**Top 23 apariții**	**32**	**893**	**893**
		gmdKHg	1	48	48
		iPDDGg	1	43	43
		bmRkHg	1	37	37
		gMdEQg	1	34	34
		sDRDGg	1	34	34
		HDmLQt	1	33	33
		MDMKHt	1	33	33
		mMdLMt	1	30	30
		MMmwCg	1	29	29
		bmdFEt	1	29	29
		hDDJCg	1	27	27
		hDDpCg	1	27	27
		hPmEMg	1	27	27
		sPmJMt	1	27	27
		NmdlQg	1	26	26
		SMMFEg	1	26	26
		bMddEg	1	26	26
		sPRDHt	1	26	26
		BDrsGt	1	25	25
		hDMKEg	1	25	25
		smdoQg	1	25	25
		AMMpHt	1	24	24
		GPmVCg	1	24	24
		SMMjEt	1	24	24
		BPMkHg	1	23	23
		GmmlQt	1	23	23
		bPmjMg	1	23	23
		hDDDHg	1	23	23
		hMdWGt	1	23	23
		hPmSEg	1	23	23
		hmddCt	1	23	23
		imMtGg	1	23	23
		Total	**4385**	**13560**	**13316**
D	T	**Top 23 apariții**	**5**	**152**	**152**
		NDRkHt	1	37	37
		sDDEMg	1	30	30
		hMrkGg	1	29	29
		MDDKHt	1	28	28
		sMrLCg	1	28	28

		Total	**4965**	**12504**	**11572**
T	P	Top 23 apariţii	13	419	405
		sPDJEg	1	64	64
		mMdlHg	1	44	42
		MMdlHg	1	40	40
		MDdjEg	1	32	30
		sDMDMg	1	29	28
		mMdqGt	1	29	23
		sDDKCg	1	28	28
		sPDLEg	1	28	28
		aDDKEg	1	27	27
		sDRKCg	1	26	26
		sPRKGg	1	25	22
		sDMLGg	1	24	24
		MDRLHt	1	23	23
		Total	**6537**	**16368**	**15317**
	D	**Top 23 apariţii**	**21**	**714**	**687**
		MDRLHt	1	88	87
		IPMJCg	1	46	45
		IPMDEg	1	42	38
		sDRJEg	1	41	39
		iPMKCg	1	36	36
		iPDJCg	1	35	33
		sPDLEg	1	34	34
		mDRlHt	1	33	33
		nDRLHt	1	32	31
		sDMLCg	1	31	29
		iPDDGg	1	31	28
		iPDDEg	1	29	27
		mDRkHt	1	28	28
		IPRKCg	1	27	26
		IPDJCg	1	27	25
		iPDKCg	1	27	25
		bPmkEt	1	26	26
		sDDJEg	1	26	26
		MDDKHt	1	26	22
		IPDKCg	1	25	25
		sDDLHg	1	24	24
		Total	**7964**	**17700**	**17331**
	T	**Top 23 apariţii**	**8**	**217**	**213**
		IDRwHt	1	34	34
		mMdlHg	1	28	28
		nMRSEt	1	28	27
		mPRDHt	1	27	26
		MDRLHt	1	26	26
		smmLCt	1	26	24
		AMDEQt	1	24	24
		IDRwGt	1	24	24
		Total	**7529**	**17100**	**16151**

Sel: Selecţie (*Selection*); Srv: Supravieţuire (*Survival*);
P: Proporţional (*Proportional*); D: Deterministic (*Deterministic*);
T: în Turnir (*Tournament*); Nr: Distincte (*Distrinct*);
Apar: Apariţii (*Occurences*);
Particip: Participări în regresii (*Participations in regressions*)

Testarea independenţei între metodele de selecţie şi supravieţuire în ceea ce priveşte numărul de genotipuri

Pentru a verifica această ipoteză se foloseşte testul χ^2 aplicat la o tabelă de contingenţă de 3X3 pentru fiecare serie de valori numerice din Tabelul 8-1 şi rezultatele sunt în Tabelele 8-2..8-7.

Tabelul 8-2

Sunt independente metoda de selecţie faţă de metoda de supravieţuire în ceea ce priveşte numărul de genotipuri distincte din cultivar în generaţiile ce produc evoluţie? - NU

Selection and survival strategy are independent events when the number of distinct viable genotypes are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 70$);

χ^2	P	T	D	Σ
P	6760 (6665)	7466 (7726)	8070 (7904)	22296
T	6537 (6586)	7529 (7634)	7964 (7810)	22030
D	3922 (3968)	4965 (4599)	4385 (4705)	13272
Σ	17219	19960	20419	57598

$$X^2(4) = \frac{95^2}{6665} + \frac{260^2}{7726} + \frac{166^2}{7904} + \frac{49^2}{6586} + \frac{105^2}{7634} + \frac{154^2}{7810} + \frac{46^2}{3968} + \frac{366^2}{4599} + \frac{320^2}{4705} \cong 70\,;\ p \cong 2 \cdot 10^{-14}$$

Tabelul 8-3

Sunt independente metoda de selecţie faţă de metoda de supravieţuire în ceea ce priveşte numărul total de genotipuri din cultivar în generaţiile ce produc evoluţie? - NU

Selection and survival strategy are independent events when the total number of viable genotypes are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 135$);

χ^2	P	T	D	Σ
P	16788 (16240)	16599 (17084)	18240 (18303)	51627
T	16368 (16095)	17100 (16932)	17700 (18140)	51168
D	10764 (11585)	12504 (12187)	13560 (13056)	36828
Σ	43920	46203	49500	139623

$$X^2(4) \cong 135\,;\ p \cong 3 \cdot 10^{-28}$$

Tabelul 8-4

Sunt independente metoda de selecție față de metoda de supraviețuire în ceea ce privește genotipurile participante la regresiile ce produc evoluție? - NU

Selection and survival strategy are independent events when the number of viable genotypes having phenotypes in regression equations are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 187$);

χ^2	P	T	D	Σ
P	15902 (15241)	15739 (16172)	17797 (18025)	49438
T	15317 (15044)	16151 (15963)	17331 (17792)	48799
D	9742 (10676)	11572 (11328)	13316 (12626)	34630
Σ	40961	43462	48444	132867

$X^2(4) \cong 187\ ;\ p \cong 2 \cdot 10^{-39}$

Tabelul 8-5

Sunt independente metoda de selecție față de metoda de supraviețuire în ceea ce privește numărul de genotipuri distincte din Top 23 în generațiile ce produc evoluție? - NU

Selection and survival strategy are independent events when the number of distinct viable genotypes with over 23 occurrences in 46 runs (Top23) are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 14.6$);

χ^2	P	T	D	Σ
P	13 (8)	6 (5)	13 (19)	32
T	13 (11)	8 (7)	21 (24)	42
D	3 (10)	5 (7)	32 (23)	40
Σ	29	19	66	114

$X^2(4) \cong 14.6\ ;\ p \cong 6 \cdot 10^{-3}$

Tabelul 8-6

Sunt independente metoda de selecție față de metoda de supraviețuire în ceea ce privește numărul total de genotipuri din Top 23 în generațiile ce produc evoluție? - NU

Selection and survival strategy are independent events when the total number of viable genotypes with over 23 occurrences in 46 runs (Top23) are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 420$);

χ^2	P	T	D	Σ
P	406 (262)	214 (167)	378 (569)	998
T	419 (354)	217 (226)	714 (770)	1350
D	89 (298)	152 (190)	893 (646)	1134
Σ	914	583	1985	3482

$X^2(4) \cong 420\ ;\ p \cong 1 \cdot 10^{-89}$

Tabelul 8-7

Sunt independente metoda de selecție față de metoda de supraviețuire în ceea ce privește genotipurile din Top 23 participante la regresiile ce produc evoluție? - NU

Selection and survival strategy are independent events when the number of viable genotypes with over 23 occurrences in 46 runs (Top23) having phenotypes in regression equations are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 440$);

χ^2	P	T	D	Σ
P	389 (247)	207 (163)	371 (557)	967
T	405 (333)	213 (220)	687 (751)	1305
D	72 (285)	152 (189)	893 (643)	1117
Σ	866	572	1951	3389

$X^2(4) \cong 440\ ;\ p \cong 6 \cdot 10^{-94}$

Se remarcă că confidența în dependența dintre metoda de selecție și metoda de supraviețuire

creşte în ordinea Număr de genotipuri distincte; Număr total de genotipuri; Număr de genotipuri participante la regresii în acelaşi timp cu faptul că numărul de observaţii nu creşte în aceeaşi ordine.

În baza dependenţei remarcate în Tabelele 8-2..8-7 între metoda de selecţie şi cea de supravieţuire pentru toţi parametrii ce caracterizează numărul de genotipuri pe parcursul evoluţiei, se impune o caracterizare a acestei dependenţe.

La întrebarea "Există legătură între cele 3 serii de numere de genotipuri?" se răspunde calculând coeficientul de corelaţie (calcul dat în Tabelul 8-8).

Tabelul 8-8

Există legătură între numărul de genotipuri distincte (NGD), numărul total de genotipuri (NTG) şi numărul de genotipuri participante la regresii (NGR)? - DA

It exists a link between the number of distinct genotypes (NGD), the number of genotypes (NTG) and the number of viable genotypes which participates to regressions (NGR)? - YES

Serii *Series*	Coeficient de determinare *Determination coefficient*	Valoare F; probabilitate de a greşi *F value; probability being in error*
NGD vs. NTG	0.982 (y=ax)	924; 10^{-15}
NGD vs. NGR	0.982 (y=ax)	951; 10^{-15}
NTG vs. NGR	0.999 (y=ax)	16110; 10^{-25}

Urmează obţinerea de evidenţe că anumite asocieri între metoda de selecţie şi metoda de supravieţuire sunt semnificativ diferite decât celelalte aşa cum rezultă şi de mai jos:

	P	T	D
P	**13**	**6**	**13**
T	**13**	**8**	**21**
D	**3**	**5**	**32**

Figura 8-1. Număr de genotipuri distincte cu cel puţin 23 de apariţii în 46 de execuţii independente

Number of distinct genotypes with at least 23 occurrences from 46 independent executions

Ceea ce se observă în Figura 8-1 de mai sus, poate fi pus în evidenţă statistic. Astfel, Tabelele 8-9 şi 8-10 redau această analiză.

Analiza din Tabelele 8-9 şi 8-10 se bazează pe calculul intervalului de încredere pentru medie şi deviaţie standard în ipoteza că distribuţia de eşantionare induce o distribuţie normală a acestor statistici (ale eşantionului) în jurul parametrului statistic (al populaţiei) asociat. Aşa cum s-a (mai) arătat Teorema Limită Centrală [1] ne asigură că se pot folosi în această analiză statisticile

[1] Teorema Limită Centrală

÷ Cronologia contribuţiilor majore:

- Abraham DE MOIVRE. 1733. Approximatio ad Summam Terminorum Binomii $(a+b)^n$ in Seriem expansi. In: Abraham DE MOIVRE. The Doctrine of Chance: or The Method of Calculating the Probability of Events in Play. W. Pearforn 1738: 235-243.
- Joseph L. LAGRANGE. 1776. Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observations; dans lequel on examine les avantages de cette méthode par le calcul des probabilités; et où l'on résoud différents problèmes relat ifs à cette matière. Miscellanea Taurinensia 5:167-232.
- Pierre S. LAPLACE. 1812. Théorie Analytique des Probabilités. Courcier, 465 p.
- Aleksandr M. LIAPUNOV. 1901. Nouvelle forme du théoreme sur la limite des probabilités. Mémoires de l'Académie Impériale des Sciences de St. Pétersbourg 12(5):1-24.

valoare medie (m) şi abatere standard (s), ale căror medii şi varianțe au fost exprimate în Tabelul 3 & Tabelul 4 - Anexa 3-1, şi pe baza cărora se poate exprima intervalul de încredere al acestora din distribuția Student t folosind relațiile [2]:

$$CI(1-\alpha, m) = m \pm t(\alpha/2, N-1)s/\sqrt{N} \quad (24)$$

$$CI(1-\alpha, s) = \left(s\sqrt{\frac{n-1}{\chi^2(\alpha/2, n-1)}}, s\sqrt{\frac{n-1}{\chi^2((1-\alpha)/2, n-1)}} \right) \quad (25)$$

Tabelul 8-9

Abateri semnificative statistic (α=5%) în numărul de genotipuri [3]

Statistically significant deviations (at 5% being in error) in the number of genotypes

Genotipuri *Genotypes*	Observabilă *Observable*	Medie *Mean*	CI(95%,Medie) *CI(95%,Mean)*		În afara intervalului* *Outside the interval*
Top 23	Num	12.7	6	20	(T,D), (D,D) > CIU; (D,P), (D,T) < CIL
Top 23	Apar	387	182	592	
Top 23	Part	377	173	580	
Total	Num	6400	5183	7617	(D,·) < CIL; (P,D), (T,D) > CIU
Total	Apar	15514	13522	17505	(D,P), (D,T) < CIL; (P,D), (T,D) > CIU
Total	Part	14763	12697	16829	(D,P), (D,T) < CIL; (P,D), (T,D) > CIU

* (Sel,Srv) - Observația făcută folosind metoda de selecție *Sel* şi de supraviețuire *Srv*
(Sel,·) - Observațiile făcute folosind metoda de selecție *Sel* şi orice metodă de supraviețuire
(·,Srv) - Observațiile făcute folosind metoda de supraviețuire *Srv* şi orice metodă de selecție
CIL - Limita inferioară a intervalului de încredere de 95% (*Lower 95% CI limit*)
CIU - Limita superioară a intervalului de încredere de 95% (*Upper 95% CI limit*)

Rezultatele din Tabelul 8-9 arată că:

÷ Selecția deterministă (D) face ca:
- Indiferent de metoda de supraviețuire numărul total de genotipuri distincte să scadă semnificativ statistic;
- Folosind supraviețuirea turnir (T) sau proporțională (P) se remarcă scădere semnificativă statistic la toți parametrii observați (Top 23 şi Total; Distincți, Apariții şi Participări în regresii) în timp ce folosind supraviețuirea deterministă (D) se remarcă creştere semnificativ statistică numai în ceea ce priveşte cele mai frecvente genotipuri pentru toți parametrii (Distincți, Apariții şi Participări în regresii);

÷ Supraviețuirea deterministă (D) face ca:
- Folosind supraviețuirea turnir (T) sau proporțională (P) să mărească semnificativ numărul total de genotipuri pentru toți parametrii (Distincți, Apariții şi Participări în regresii).

÷ Enunțul teoremei (fie $(X_n)_{n\geq 1}$ variabile independente şi $\exists\delta>0$ a.î. $\mu_{2+\delta}(X_n)<\infty$):
- dacă $\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \mu_{2+\delta}(X_k)}{\left(\sum_{k=1}^{n} \sigma_k^{\,2}\right)^{(2+\delta)/2}} = 0$ atunci $\frac{\sum_{i=1}^{n} (X_n - \mu_1(X_n))}{\sqrt{\sum_{k=1}^{n} \sigma_k^{\,2}}} \xrightarrow[n\to\infty]{} N(0,1)$

[2] Ronald A. FISHER. 1956. Statistical Methods and Scientific Inference. Edinburgh: Oliver & Boyd. p. 32.
[3] Valori calculate cu STATISTICA v.8.0 (©1984-2008), StatSoft Inc. în baza formulelor (24) şi (25).

Tabelul 8-10

Diferențe semnificative statistic (α=5%) în numărul de genotipuri (medii, abateri)

Statistically significant differences (α=5%) in number of genotypes (means, deviations)

Geno	Obs	Grup: Medie; Deviație	CI(95%,Medie)		Dev	CI(95%,Dev)		În afara intervalului
		Group: mean; deviation	CI(95%,Mean)					*Outside of the interval*
Top23	Num	(P,·): 10.7; 4.0	6	20	9.10	6	17	Dev(P,·)
Top23	Apar	(P,·): 333; 104	182	592	266	180	510	Dev(P,·)
Top23	Part	(P,·): 322; 100	173	580	265	179	508	Dev(P,·)
Total	Num	(P,·): 7432; 656	5183	7617	1583	1069	3033	Dev(P,·)
Total	Apar	(P,·): 17209; 898	13522	17505	2591	1750	4963	Dev(P,·)
Total	Part	(P,·): 16479; 1144	12697	16829	2687	1815	5148	Dev(P,·)
Top23	Num	(T,·): 14.0; 6.6	6	20	9.10	6	17	-
Top23	Apar	(T,·): 450; 250	182	592	266	180	510	-
Top23	Part	(T,·): 435; 238	173	580	265	179	508	-
Total	Num	(T,·): 7343; 731	5183	7617	1583	1069	3033	Dev(T,·)
Total	Apar	(T,·): 17056; 667	13522	17505	2591	1750	4963	Dev(T,·)
Total	Part	(T,·): 16266; 1012	12697	16829	2687	1815	5148	Dev(T,·)
Top23	Num	(D,·): 13.3; 16.2	6	20	9.10	6	17	-
Top23	Apar	(D,·): 378; 447	182	592	266	180	510	-
Top23	Part	(D,·): 372; 453	173	580	265	179	508	-
Total	Num	(D,·): 4424; 523	5183	7617	1583	1069	3033	Med(D,·); Dev(D,·)
Total	Apar	(D,·): 12276; 1412	13522	17505	2591	1750	4963	Med(D,·); Dev(D,·)
Total	Part	(D,·): 11543; 1787	12697	16829	2687	1815	5148	Med(D,·); Dev(D,·)
Top23	Num	(·,P): 9.7; 5.8	6	20	9.10	6	17	Dev(·,P)
Top23	Apar	(·,P): 305; 187	182	592	266	180	510	-
Top23	Part	(·,P): 289; 188	173	580	265	179	508	-
Total	Num	(·,P): 5740; 1578	5183	7617	1583	1069	3033	-
Total	Apar	(·,P): 14640; 3363	13522	17505	2591	1750	4963	-
Total	Part	(·,P): 13654; 3400	12697	16829	2687	1815	5148	-
Top23	Num	(·,T): 6.3; 1.5	6	20	9.10	6	17	Dev(·,T)
Top23	Apar	(·,T): 194; 37	182	592	266	180	510	Dev(·,T)
Top23	Part	(·,T): 191; 34	173	580	265	179	508	Dev(·,T)
Total	Num	(·,T): 6653; 1462	5183	7617	1583	1069	3033	-
Total	Apar	(·,T): 15401; 2521	13522	17505	2591	1750	4963	-
Total	Part	(·,T): 14487; 2533	12697	16829	2687	1815	5148	-
Top23	Num	(·,D): 22.0; 9.5	6	20	9.10	6	17	Med(·,D)
Top23	Apar	(·,D): 662; 261	182	592	266	180	510	Med(·,D)
Top23	Part	(·,D): 650; 263	173	580	265	179	508	Med(·,D)
Total	Num	(·,D): 6806; 2098	5183	7617	1583	1069	3033	-
Total	Apar	(·,D): 16500; 2560	13522	17505	2591	1750	4963	-
Total	Part	(·,D): 16148; 2464	12697	16829	2687	1815	5148	-

Rezultatele din Tabelul 8-10 arată că:

÷ Supraviețuirea deterministă (D) îmbogățește semnificativ statistic grupul celor mai frecvente genotipuri (Top23) din generațiile ce produc evoluție în timp ce selecția deterministă (D) sărăcește semnificativ statistic numărul total al genotipurilor din generațiile care produc evoluție;

÷ Practic fiecare metodă de selecție definește câte o populație genotipică în generațiile care produc

evoluție; argumentul este că oricare ar fi parametrul urmărit pentru numărul total de genotipuri (Distincți, Apariții şi Participări în regresii), şi luând pentru exemplificare numărul de genotipuri distincte (Num), avem:

- Varianța totală: 1583^2 cu intervalul de încredere de 95%: [1069^2, 3033^2];
- Varianța populației produse de selecția proporțională (P): $656^2 < 1069^2$;
- Varianța populației produse de selecția turnir (T): $731^2 < 1069^2$;
- Varianța populației produse de selecția proporțională (P): $523^2 < 1069^2$;

÷ Nu aceeaşi concluzie se poate trage cu privire la metoda de supraviețuire, pentru care se produce segregare populațională doar pentru supraviețuirea deterministă (D), care creează o populație cu un număr mediu de genotipuri semnificativ statistic mai mare decât supraviețuirea proporțională (P) şi respectiv turnir (T).

Concluzii cu privire la variabilitate

S-au interpretat abaterile semnificative statistic (α=5%) şi s-a observat că:

÷ selecția deterministă face ca indiferent de metoda de supraviețuire numărul total de genotipuri distincte reprezentate în cultivar să scadă semnificativ statistic;

÷ supraviețuirea deterministă asociată cu selecția turnir sau proporțională măreşte semnificativ statistic numărul total de genotipuri reprezentate în cultivar pentru toate cele trei statistici urmărite;

÷ supraviețuirea deterministă îmbogățeşte semnificativ statistic grupul celor mai frecvente genotipuri din generațiile ce produc evoluție în timp ce selecția deterministă sărăceşte semnificativ statistic numărul total al genotipurilor din generațiile care produc evoluție;

÷ practic fiecare metodă de selecție defineşte câte o populație genotipică în generațiile care produc evoluție; argumentul este că oricare ar fi statistica urmărită al numărului total (*Total*) de genotipuri (*Num*, *Apar* sau *Part*), fiecare dintre variațele produse de variația metodei de supraviețuire în grupul în care metoda de selecție este aceeaşi sunt semnificativ statistic mai mici decât varianța totală; nu aceeaşi concluzie se poate trage cu privire la metoda de supraviețuire, pentru care se produce segregare populațională doar pentru supraviețuirea deterministă, care creează o populație cu un număr mediu de genotipuri semnificativ statistic mai mare decât supraviețuirea proporțională şi respectiv în turnir;

Compararea evoluțiilor bazate pe diferite metode de selecție şi supraviețuire

Pe baza rezultatelor experimentale un alt parametru important al evoluției poate fi interpretat: numărul de generații care produc evoluție în cursul execuției cu număr impus de generații (20000) ca măsură a capacității de adaptare definită de combinația celor două metode (de

selecție şi supraviețuire) precum şi valoarea medie a numerelor generațiilor care produc evoluție ca măsură a vitezei de adaptare.

Tabelul 8-11

Generații ce produc evoluții (număr şi medie) pentru 46 de execuții independente

Generations producing evolution (number and average) for 46 independent runs

run	(T,T)		(T,D)		(T,P)		(D,T)		(D,D)		(D,P)		(P,T)		(P,D)		(P,P)	
1	17	3744	26	2884	12	412	12	2662	17	1424	7	2632	52	2840	22	1734	37	3423
2	27	2096	27	4704	40	5115	13	3854	19	2295	25	2780	20	1565	36	4446	19	4916
3	26	1134	32	3531	28	1545	30	2714	24	2099	13	1653	29	1808	30	4095	38	2499
4	27	2746	18	241	16	413	17	2981	36	1084	26	3317	16	3952	27	5396	36	2856
5	28	3299	42	1655	13	2314	32	3079	27	3133	22	4913	20	977	36	5462	23	817
6	20	1717	49	2398	17	1704	25	1500	8	4697	25	3235	27	488	24	653	28	4255
7	19	1059	38	2103	36	4749	34	6709	25	3382	20	4528	28	1764	39	2646	28	1710
8	23	3173	38	1781	20	852	23	5272	23	8355	17	3865	35	2491	23	2060	20	2335
9	37	1034	34	3449	44	3596	12	822.1	33	5814	21	3523	30	1485	28	3264	40	7250
10	18	5046	39	5952	22	1762	14	3647	36	4260	19	5019	17	3345	38	7288	61	5790
11	36	2516	19	3191	15	599	27	3174	11	2849	24	2135	31	564	33	2193	33	4844
12	25	3190	35	4744	24	4942	27	2511	26	6400	24	4675	33	1795	38	4015	33	8315
13	35	907	26	1479	38	2803	21	9086	6	32	14	6338	39	4880	49	5570	33	3103
14	48	4672	20	3958	25	4738	15	8139	32	8734	5	718	22	5116	18	1607	43	3602
15	36	1657	44	4893	39	2839	19	3227	32	5209	4	232	40	6560	20	2896	27	736
16	21	991	24	246	36	1749	12	2503	20	1317	10	2984	34	11086	50	3292	28	3195
17	56	6242	41	1806	27	1707	33	4121	29	6367	31	5316	26	1555	28	964	46	4860
18	24	1773	33	4545	23	5670	18	4119	27	4520	28	3645	45	2394	49	7685	19	1589
19	29	1059	19	6112	23	1141	23	5796	11	992	16	1458	36	4459	14	1378	21	5115
20	23	3851	30	3557	28	710	23	1558	22	3669	6	3335	42	4816	45	3543	35	3731
21	31	1928	25	1302	38	2767	17	5730	20	3094	18	8798	36	2526	18	2795	45	1216
22	37	699	25	2320	31	4964	18	4596	29	5753	15	576	41	1993	17	790	33	3505
23	37	2148	20	4900	14	2292	32	3934	42	5110	14	5037	47	8925	47	2020	34	3427
24	31	5706	20	8773	26	555	10	2962	46	4439	19	3209	13	1954	24	1985	28	1930
25	21	2297	35	2497	47	1827	17	2966	25	2590	14	3056	20	1558	40	2493	26	1915
26	17	4759	29	2069	28	6925	21	5171	20	7232	24	6650	25	1430	46	3584	17	985
27	37	5362	28	2554	29	6977	19	5248	32	4013	13	2644	31	4939	32	3002	36	5643
28	18	2531	38	1395	34	1976	4	877	18	3256	24	10910	33	1617	32	1810	18	798
29	50	3231	38	975	18	3132	22	2462	25	7589	34	6637	26	937	56	5874	32	3690
30	37	3887	51	6828	43	6310	20	5512	18	4664	10	3450	44	1598	28	4147	26	3217
31	23	3474	29	1921	20	1908	21	1825	26	4596	11	1884	34	2606	22	2664	14	1854
32	18	1130	34	3467	20	353	35	7361	32	2518	6	15	32	2099	40	5477	12	931
33	36	3597	20	1300	23	2535	15	1754	33	4615	25	5651	26	4976	19	12799	21	2169
34	32	516	22	5830	30	3108	43	8655	6	4893	30	4578	27	4334	42	2904	13	4336
35	27	3186	42	5272	64	3288	15	3252	31	5858	13	1222	32	3064	29	1348	30	2081
36	36	4600	39	4125	22	4620	14	3278	14	3301	14	5456	57	4468	28	3003	32	3050
37	30	5061	16	3154	31	1345	23	4881	28	5105	27	887	33	8163	35	8225	23	4708
38	25	2055	40	4141	32	7258	24	4122	13	2382	22	2577	44	4311	20	819	38	4541
39	27	1018	23	3269	17	2075	43	4545	18	5421	29	2232	35	6333	33	4784	40	853
40	26	1039	46	1892	50	4306	24	5563	22	631	10	177	23	887	26	1155	8	792
41	31	8549	20	1763	18	1954	15	3522	34	3061	21	6100	26	2593	29	360	25	4508
42	19	2272	22	772	56	7329	21	2441	22	3782	14	5815	26	1904	35	3286	24	1084
43	47	5746	23	3095	21	1484	13	2313	19	2516	23	6353	25	3230	44	1418	40	3994
44	38	2142	39	6810	26	1454	36	4964	8	3765	10	2492	34	495	29	2954	34	3491
45	34	1432	32	4117	30	2757	27	2673	24	6675	34	2471	38	6027	26	2764	31	4802
46	39	4481	39	4332	24	1277	17	1635	15	6036	20	5650	33	944	30	758	25	2566

Rezultatele au fost prelucrate în Tabelul 8-11 astfel: pentru fiecare execuție independentă a

algoritmului genetic s-au consemnat numerele consecutive ale generațiilor care au produs îmbunătățirea valorii funcției obiectiv, şi anume s-a obținut o ecuație de regresie validă (toți coeficienții sunt semnificativi statistic) cu un coeficient de determinare mai mare decât cel obținut în generațiile anterioare.

S-a calculat apoi valoarea medie a acestui număr (care totdeauna este mai mic decât numărul maxim de generații al unei execuții) şi numărul de evoluții distincte (numărul de valori), informații care sunt prezentate în tabel. Întrucât ambele valori (media şi numărul de valori) au fost obținute printr-o repetare (de 46 de ori) a experimentului pentru fiecare pereche de metode (selecție, supraviețuire) valorile obținute aproximează distribuția de eşantionare, astfel încât s-a putut presupune aproximația la normalitate a acestora (atât ele ca valori ale eşantionului de 46 de observații) cât şi populația din care provin, care aşa cum s-a dovedit mai sus este caracteristică (distinctă) cel puțin după metoda de selecție.

Pentru a se realiza compararea perechilor de metode (selecție, supraviețuire) informațiile din Tabelul 8-11 au fost supuse unei analize statistice descriptive, care a inclus calcularea valorilor medii şi a deviațiilor standard, împreună cu intervalele de încredere la un nivel de semnificație de 95%, rezultate care sunt redate în Tabelul 8-12.

Tabelul 8-12

Statistici ale evoluției supervizate de algoritmul genetic în funcție de selecție şi supraviețuire

Statistics of genetic algorithm supervised evolution depending on selection and survival strategies

Parametru	Medie	CI(95%,Medie)		Deviație	CI(95%,Deviație)	
m(D,D)	4120	3518	4722	2027	1681	2553
m(D,T)	3907	3326	4488	1957	1623	2465
m(D,P)	3714	3032	4396	2296	1904	2892
m(P,D)	3335	2631	4039	2369	1965	2984
m(T,D)	3307	2748	3866	1882	1561	2371
m(P,T)	3214	2520	3908	2338	1939	2945
m(P,P)	3196	2671	3722	1770	1468	2229
m(T,T)	2929	2400	3458	1781	1478	2244
m(T,P)	2916	2322	3510	2001	1660	2520
n(P,D)	32.0	29.0	35.1	10.1	8.4	12.8
n(P,T)	31.8	29.0	34.6	9.3	7.7	11.7
n(T,D)	31.1	28.3	33.8	9.2	7.7	11.6
n(T,T)	30.0	27.2	32.7	9.3	7.7	11.7
n(P,P)	29.4	26.4	32.4	10.2	8.4	12.8
n(T,P)	28.7	25.2	32.1	11.5	9.5	14.5
n(D,D)	23.6	20.8	26.3	9.2	7.6	11.6
n(D,T)	21.7	19.1	24.2	8.5	7.1	10.8
n(D,P)	18.5	16.2	20.8	7.9	6.5	9.9

Datele din Tabelul 8-12 au servit pentru caracterizarea genotipurilor reprezentate în cultivar (numărul şi media acestora). Caracterizarea s-a făcut pentru fiecare parametru statistic calculat în Tabelul 8-12, atribuind semnificațiile fizice măsurilor folosite şi interpretând valorile obținute. Rezultatele acestei caracterizări sunt redate în continuare.

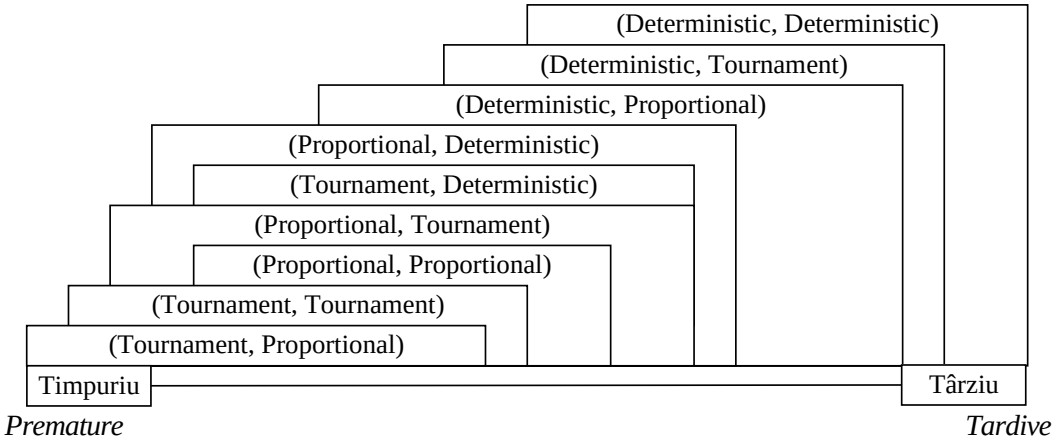
Media generațiilor care produc evoluții (Medie şi CI(95%,Medie pentru m(·,·) în Tabelul 8-12)

Media generațiilor care produc evoluții este o măsură a ***cât de timpuriu / târziu se produc evoluțiile***. Figura următoare (Figura 8-2) reprezintă valorile mediei şi intervalului său de încredere.

Analizând valorile prezentate în Tabelul 8-12 în ceea ce priveşte valorile medii şi intervalele de încredere ale acestora pentru media generațiilor care produc evoluții, se remarcă că (Figura 8-2):

÷ selecția deterministă (şi cu atât mai mult însoțită de supraviețuirea deterministă) produce cele mai târzii evoluții;

÷ ce selecția turnir (şi cu atât mai mult însoțită de supraviețuirea proporțională sau turnir) produce cele mai timpurii evoluții;



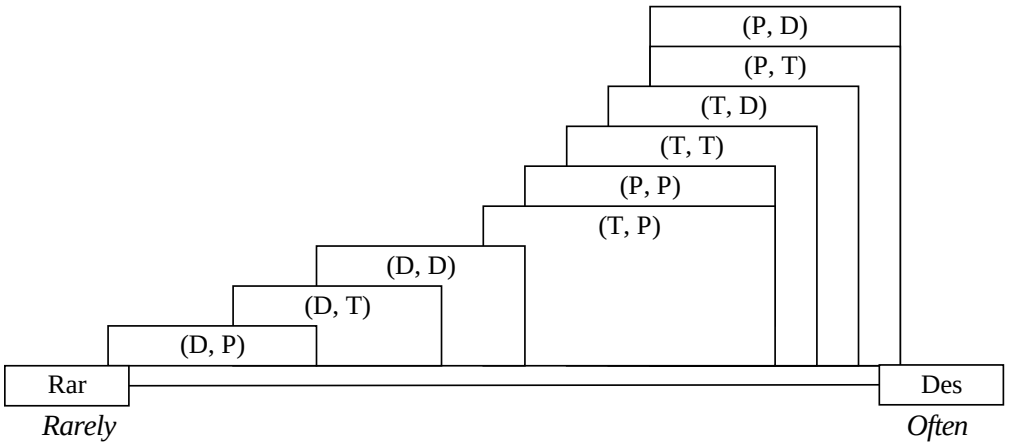
(Selecție, Supraviețuire): [CI(95%,Medie$_{n=46}$(generație medie))]

Figura 8-2. Cât de timpuriu se produc evoluțiile?

(Selection, Survival): [CI(95%,Mean$_{n=46}$(evolution moments mean))] - How fast the evolutions occurs?

Media numărului de evoluții (Medie şi CI(95%,Medie pentru n(·,·) în Tabelul 8-12)

Numărul de evoluții dintr-un număr dat de generații este o măsură a ***cât de frecvent se produc evoluțiile***. Figura următoare (Figura 8-3) reprezintă valorile mediei şi intervalului său de încredere.



(Selecție, Supraviețuire) [CI(95%,Medie$_{n=46}$(număr evoluții))]

Figura 8-3. Cât de frecvent se produc evoluțiile?

(Selection, Survival): [CI(95%,Mean$_{n=46}$(evolution number))] - How often the evolution occurs?

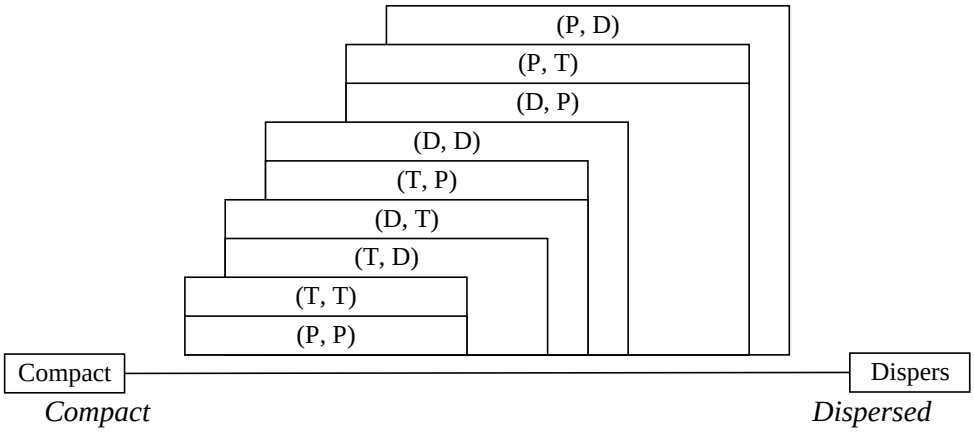
Analizând valorile prezentate în Tabelul 8-12 în ceea ce priveşte valorile medii şi intervalele de încredere ale acestora pentru numărul de evoluții, se remarcă că (Figura 8-3):

÷ selecția proporțională (şi cu atât mai mult însoțită de supraviețuirea deterministă sau turnir) produce cele mai multe evoluții, în timp ce selecția deterministă (şi cu atât mai mult însoțită de supraviețuirea proporțională sau turnir) produce cele mai rare evoluții;

÷ în ceea ce priveşte frecvența evoluțiilor (Figura 8-3) o selecție proporțională aproape că nu face diferența între supraviețuirea deterministă şi supraviețuirea în turnir (medii 32 şi 31.8, cu intervale de încredere rotunjite la întreg egale) în timp ce viteza evoluțiilor (Figura 8-2) aceeaşi selecție proporțională nu face diferența între valoarea medie pentru supraviețuirea proporțională şi turnir (diferența de 150 între valorile medii reprezentând cel mult 30% din lărgimea intervalului de încredere la oricare dintre ele.

Dispersia momentului de evoluție (Deviație şi CI(95%,Deviație pentru m(·,·) în Tabelul 8-12)

Variabilitatea momentului evoluției pe parcursul generațiilor este o măsură a ***cât de compact / dispers se produc evoluțiile***. Figura următoare (Figura 8-4) reprezintă valorile deviației standard şi intervalului său de încredere.



(Selecție, Supraviețuire) [CI(95%,Deviație$_{n=46}$(generație medie))]

Figura 8-4. Cât de dispers se produc evoluțiile?

(Selection, Survival): [CI(95%,StD$_{n=46}$(evolution moments mean))] - How much dispersed the evolutions are?

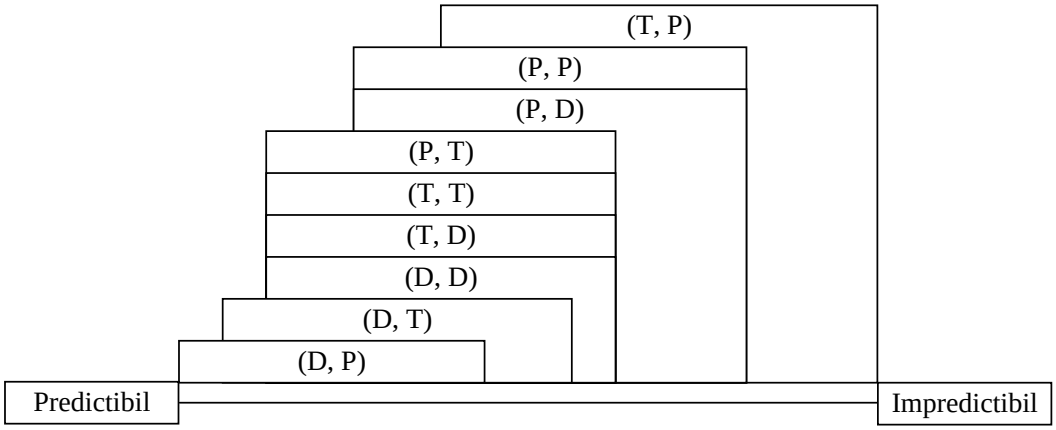
Analizând valorile prezentate în Tabelul 8-12 în ceea ce priveşte deviațiile standard şi intervalele de încredere ale acestora pentru numărul de evoluții, se remarcă că (Figura 8-4):

÷ cea mai bună compactitate de evoluție o au selecția proporțională asociată cu supraviețuirea proporțională şi selecția în turnir asociată cu supraviețuirea în turnir;

÷ ce cea mai mare împrăştiere în timp a evoluției se observă la selecția proporțională asociată cu supraviețuirea deterministă urmată îndeaproape de selecția deterministă asociată cu supraviețuirea proporțională;

Dispersia numărului de evoluții (Deviație şi CI(95%,Deviație pentru n(·,·) în Tabelul 8-12)

Variabilitatea numărului evoluțiilor este o măsură a ***cât de predictibil / impredictibil se produc evoluțiile***. Figura următoare (Figura 8-5) reprezintă valorile deviației standard şi intervalului său de încredere.



(Selecție, Supraviețuire) [CI(95%,Deviație$_{n=46}$(număr evoluții))]

Figura 8-5. Cât de predictibil se produc evoluțiile?

(Selection, Survival): [CI(95%,StD$_{n=46}$(evolution number))] - How much predictable the evolutions are?

Analizând valorile prezentate în Tabelul 8-12 în ceea ce priveşte deviațiile standard şi intervalele de încredere ale acestora pentru numărul de evoluții, se remarcă că (Figura 8-5):

÷ patru asocieri de selecție şi supraviețuire au rezultate similare sub aspectul predictibilității evoluției: (selecție deterministic, supraviețuire deterministic), (selecție în turnir, supraviețuire deterministic), (selecție în turnir, supraviețuire în turnir) şi (selecție proporțională, supraviețuire în turnir);

÷ predictibilități extreme ale evoluției obțin selecția deterministă asociată cu supraviețuirea proporțională având cea mai mare predictibilitate (cea mai mică variabilitate) şi selecția în turnir asociată cu supraviețuirea proporțională având cea mai mică predictibilitate (cea mai mare variabilitate) de evoluție;

Diversitatea genotipurilor reprezentate în cultivar în momentele de evoluție

Diversitatea genotipurilor reprezentate în cultivar în momentele evoluției se poate cuantifica folosind entropia informațională. Cum în cazul calculului unei valori medii, unde avem la dispoziție o familie de măsuri medii, date de expresia mediei generalizate (sau Hölder sau Kolmogorov) - pentru o discuție a mediilor vezi (Marichal, 2000):

$$H(X_1,...,X_n;p)=\left(\frac{1}{n}\sum_{i=1}^{n}X_i^{\,p}\right)^{1/p},$$

având loc relația:

$H(\{X\};-\infty)=\min(\{X\})\le H(\{X\};-1)=HM(\{X\})\le H(\{X\};0)=GM(\{X\})\le H(\{X\};1)=AM(\{X\})\le H(\{X\};-\infty)=\max(\{X\})$

min: valoarea minimă; HM: media armonică; GM: media geometrică; AM: media aritmetică; max: valoarea maximă

la fel şi în cazul entropiei informaționale avem la dispoziție o familie de măsuri entropice, date de expresia entropiei generalizate (sau Rényi) - (Rényi, 1961):

$$H(p_1,...,p_n;\alpha)=\frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n}p_i^{\alpha}\right), \text{ unde } \sum_{i=1}^{n}p_i\le 1,$$

cazul $\Sigma p_i<1$ numindu-se al unei distribuții incomplete, şi unde are loc relația:

$H(\{p\};\infty)=H_\infty(\{p\}) \le H(\{p\};2)=H_2(\{p\}) \le H(\{p\};1)=H_1(\{p\}) \le H(\{p\};0)=H_0(\{p\}) \le 2H(\{p\};\infty)=2H_\infty(\{p\})$
H_∞: entropia min; H_2: - logaritmul diversității Simpson; H_1: entropia Shannon; H_0: entropia Hartley (entropia max);

Măsurile H_0, H_1 şi H_2 sunt frecvent folosite în literatura de specialitate pentru a descrie gradul de dezordine sau diversitatea. Astfel, entropia Hartley (H_0) - (Hartley, 1928) - este aplicată în informatică la reconcilierea informației, entropia Shannon (H_1) - (Shannon, 1948) - este aplicată în fizică şi chimie la caracterizarea stării materiei, iar entropia derivată din indicele de diversitate Simpson (H_2) - (Simpson, 1949) - în biologie şi științele conexe la caracterizarea diversității populațiilor de organisme.

Dacă logaritmul din expresia entropiei se calculează în baza 2 atunci unitatea de măsură a entropiei este biți, dacă se calculează în baza e (logaritm natural) atunci măsura este nați, iar dacă se calculează în baza 10 atunci măsura este diți.

Fişierele de rezultate au fost prelucrate pentru calculul măsurilor entropice H_0, H_1 şi H_2 după cum urmează:

÷ Au fost extrase pentru fiecare execuție, fiecare asociere (selecție, supraviețuire) şi fiecare evoluție lista genotipurilor reprezentate în cultivar în momentul evoluției (coloanele Gen0..Gen11 în fişierele de rezultate);

÷ S-au concatenat listele de genotipuri pentru toate evoluțiile unei execuții şi unei asocieri;

÷ S-au sortat şi apoi numărat aparițiile acestor genotipuri pentru fiecare execuție şi asociere;

÷ Rezultatul obținut reprezintă frecvențe genotipuri la o execuție şi o asociere;

÷ S-au calculat probabilitățile {p} din frecvențe; s-au aplicat formulele de mai sus pentru calculul lui H_0, H_1 şi H_2; valorile calculate sunt redate în Anexa 6;

÷ Pentru valorile calculate ale entropiilor H_0, H_1 şi H_2 şi pentru fiecare asociere de strategie de selecție şi supraviețuire s-au calculat valorile medii şi lărgimea intervalelor de încredere pentru medii; rezultatele sunt redate în Tabelul 8-13;

÷ Cele trei valori medii redate în Tabelul 8-13 s-au reprezentat grafic în ordinea crescătoare a valorii medii împreună cu intervalul de încredere în Figura 8-6.

Tabelul 8-13

Entropii medii şi intervale de încredere 95% *Means and CI's at 95% for entropies*

SS	H0LCM	H0Mean	H0SCM	H1LCM	H1Mean	H1SCM	H2LCM	H2Mean	H2SCM
DD	6.42	6.59	6.76	5.82	5.97	6.11	5.25	5.37	5.50
DP	6.37	6.54	6.71	5.98	6.14	6.29	5.59	5.73	5.87
DT	6.90	7.05	7.20	6.53	6.67	6.81	6.06	6.18	6.31
PD	7.43	7.56	7.69	6.94	7.07	7.19	6.38	6.50	6.61
PP	7.18	7.32	7.46	6.74	6.87	7.01	6.25	6.38	6.51
PT	7.47	7.59	7.70	7.06	7.17	7.28	6.57	6.67	6.78
TD	7.47	7.60	7.72	7.04	7.15	7.27	6.49	6.60	6.71
TP	7.10	7.25	7.40	6.67	6.81	6.94	6.18	6.31	6.43
TT	7.35	7.46	7.58	6.93	7.04	7.15	6.42	6.53	6.65

H0Mean, H1Mean, H2Mean: valorile medii (din 46 de execuţii) pentru H_0, H_1 şi H_2
mean values (from 46 executions) for
H0LCM, H1LCM, H2LCM: valoarea inferioară a limitei de probabilitate 95% pentru valoarea medie a H_0, H_1 şi respectiv H_2 (*lower 95% CI for the mean of*)
H0SCM, H1SCM, H2SCM: valoarea superioară a limitei de probabilitate 95% pentru valoarea medie a H_0, H_1 şi respectiv H_2 (*upper 95% CI for the mean of*)

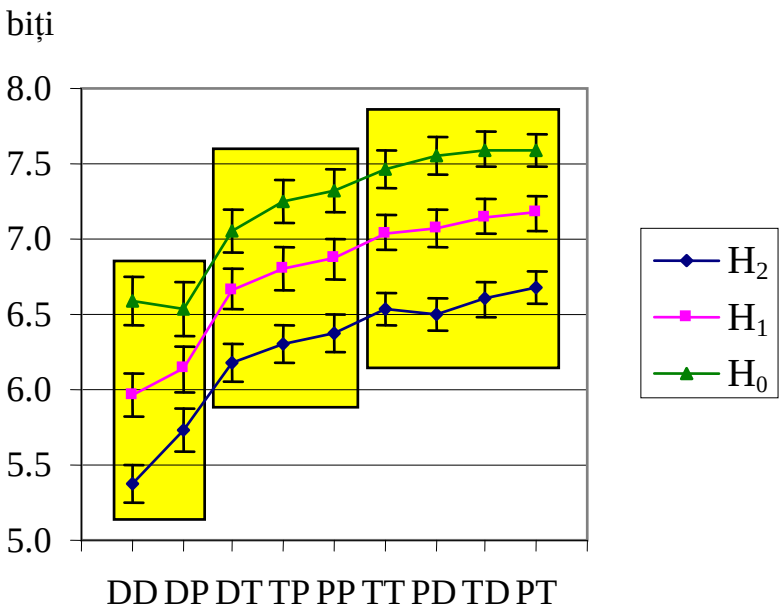


Figura 8-6. Diversitatea genotipică pe parcursul evoluţiei
Genotypic diversity during evolution

Graficul din Figura 8-6 arată că cele nouă strategii de evoluţie se grupează în 3 grupe după diversitatea genotipică care o produc în cultivar:

÷ Grupul (DD, DP) - selecţie deterministă, supravieţuire deterministă sau proporţională este semnificativ statistic diferit în diversitate de toate celelalte strategii în toate trei măsurile entropice; produc cea mai scăzută diversitate genotipică;

÷ Pentru grupul (DT,TP,PP) - (selecţie deterministă, supravieţuire în turnir) împreună cu (supravieţuire proporţională, selecţie în turnir sau proporţională) se poate evidenţia statistic că este în medie semnificativ statistic diferit de grupul de ce produce mai mare diversitate (TT, PD, TD, PT) chiar dacă nu fiecare metodă a grupului este semnificativ statistică diferită de metodele

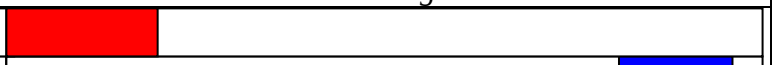
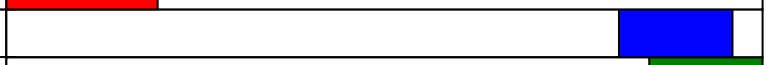
celuilalt grup - de exemplu pentru H_2 calculul mediei arată că Medie(DT,TP,PP) = 6.29±0.07(df=138) în timp ce Medie(TT,PD,TD,PT) = 6.58±0.05(df=184);

÷ Grupul (TT, PD, TD, PT) este compus din cele mai favorabile strategii pentru păstrarea diversității: (selecție și supraviețuire în turnir), (selecție proporțională și supraviețuire deterministă), (selecție în turnir și supraviețuire deterministă), și respectiv selecție proporțională și supraviețuire în turnir;

O remarcă generală cu privire la strategia de selecție se poate face concatenând observațiile de la cele 3 strategii de supraviețuire și calculând din nou valoarea medie împreună cu intervalul său de încredere și în mod similar pentru strategia de supraviețuire. Tabelul următor cumulează aceste rezultate pentru H_1 (Tabelul 8-14).

Tabelul 8-14

Entropia Shannon (H_1) pentru selecție și pentru supraviețuire
Shannon (H_1) entropy for selection and survival

Strategie *Strategy*	Medii și intervale de încredere la riscul de a fi în eroare de 5% *Means and CI's at 5% risk being in error*	
Selecție *Selection*	D(6.26±0.10)	
	T(7.00±0.07)	
	P(7.04±0.07)	
Supraviețuire *Survival*	P(6.61±0.10)	
	D(6.73±0.12)	
	T(6.96±0.08)	
D: strategie deterministă; T: strategie în turnir; P: strategie proporțională *D: deterministic strategy; T: tournament strategy; P: proportional strategy*		

Analiza statistică din Tabelul 8-14 arată extrem de clar că influență decisivă asupra reducerii dramatice a diversității în cultivar o are strategia de selecție deterministă; diferența între diversitatea produsă de strategia de selecție proporțională și cea în turnir nu este semnificativă statistic la riscul de a fi în eroare de 5%.

Și strategia de supraviețuire are o influență decisivă asupra diversității genotipurilor. Strategia de supraviețuire în turnir produce o diversitate mai mare (fapt evidențiat statistic cu un risc de a fi în eroare de 5%) decât strategia deterministă și respectiv proporțională; acestea din urmă nu se disting statistic la riscul de a fi în eroare de 5%, cu toate că observația arată că strategia deterministă produce o mai mare diversitate.

9. Analiza calitativă:
Interpretarea rezultatelor din observațiile pe observabile calitative

Analiza numărului de genotipuri reprezentate în cultivar în generațiile ce au produs evoluție

O ipoteză importantă poate fi formulată pe baza datelor prezentate în Tabelul 8-1, şi anume dacă numărul mediu de genotipuri viabile (ce rezultă din împărțirea la numărul de generații ce au produs evoluție observate *num_obs* a numărului de genotipuri viabile observate în aceste generații *sum_obs*) este sau nu independent de strategia de selecție şi strategia de supraviețuire.

Pentru această analiză (care să răspundă la întrebarea *"În ce măsură numărul mediu de genotipuri viabile este sau nu independent de strategia de selecție şi strategia de supraviețuire?"*) s-a implementat şi folosit un program pentru analiza de distribuție a perechilor de eşantioane folosind testul Anderson-Darling descris în Anexa 3-3.

Un calcul al numărului total de inferențe posibile arată că pentru media genotipurilor viabile (dată în Tabelul 9-1) trebuie investigate un număr de 2^9-C_9^0-C_9^1 = 502 inferențe. Rezultatul analizei este redat în Tabelul din Anexa 7-1.

Tabelul 9-1

Distribuția mediei numărului de genotipuri viabile în cultivar după strategia de selecție şi strategia de supraviețuire (*Distribution of the mean of the number of viable genotypes in cultiver by the selection and survival strategy*)

Obs\Grup	PP	PT	PD	TP	TT	TD	DP	DT	DD
1	11.429	11.511	11.798	11.405	11.453	11.848	11.23	11.329	11.826
2	11.206	11.462	11.686	10.95	11.169	11.655	10.5	11.118	11.683
3	11.062	11.316	11.545	10.927	11.141	11.684	10.227	11.062	11.716
4	11.29	11.382	11.647	11.273	11.29	11.602	10.576	10.941	11.697
5	11.583	11.278	11.7	11.211	11.161	11.541	10.86	11.104	11.83
6	11.364	11.477	11.676	10.677	11.459	11.652	10.321	11.282	11.854
7	11.4	11.375	11.375	11.081	11.237	11.612	10	10.816	11.792
8	11.3	11.606	11.44	11	11.355	11.615	10.488	10.6	11.769
9	11.143	11.344	11.643	10.941	11.286	11.619	11.097	10.694	11.686
10	11.188	11.531	11.571	11.111	10.783	11.742	11.133	10.884	11.792
11	11.379	11.545	11.682	11.12	11.474	11.483	11.067	10.929	11.818
12	11.625	11.179	11.636	11.391	10.76	11.64	10.765	10.92	11.842
13	11.05	11.313	11.583	11.235	11.357	11.742	9.667	10.759	11.778
14	11.067	10.731	11.68	11.111	11.471	11.688	9.588	10.692	11.783
15	11.5	11.1	11.6	11.286	11.056	11.8	9.813	11.5	11.739
16	12	11.381	11.4	11.5	11.12	12	11	10.563	11.857
17	11.167	11.5	11.733	11	11.333	11.786	10	10.875	11.684
18	11.667	11.25	11.636	11.167	12	11.667	11.176	11.1	11.917
19	11.824	11.636	11.571	11.5	11.714	11.727	11.333	10.818	11.875
20	11.75	10.733	11.5	11.308	11.333	11.75	11.4	10	12
Obs: medie din parcursul a 1000 de generații; Grup: Strategie Selecție & Supraviețuire									

Pe baza rezultatelor din Tabelul din Anexa 7-1) se desprinde că:

÷ Analiza grupurilor de câte două perechi de metode evidențiază statistic că:
  - Nu poate fi respinsă ipoteza apartenenței la populații identice pentru: DT şi DP (cu un raport între valoarea critică şi statistică c/k = 1.2); PP şi PT (c/k = 3.0); PP şi TT (3.2); TT şi PT (2.1); TT şi TP (1.3);

÷ Concluzia analizei grupurilor de câte două perechi de metode este:
  - Cu un risc de a fi în eroare de 5% următoarele provin din populații diferite: DD, PD, TD;
  - Nu poate fi respinsă ipoteza că DT şi DP provin din populații identice;
  - Analiza populațională pentru PP, PT, TT şi TP necesită investigația grupurilor de ordin superior;

÷ Analiza grupurilor de câte trei perechi de metode evidențiază statistic că:
  - Nu poate fi respinsă ipoteza că PP, PT şi TT (c/k = 2.2) provin din populații identice;
  - Cu un risc de a fi în eroare de 5% toate grupurile de trei perechi de metode ce conțin metoda TP provin din populații diferite;

÷ Concluzia analizei grupurilor de câte cel mult trei perechi de metode este:
  - Cu un risc de a fi în eroare de 5% următoarele provin din populații diferite: DD, PD, TD;
  - Nu poate fi respinsă ipoteza că DT şi DP provin din populații identice;
  - Nu poate fi respinsă ipoteza că TT şi TP provin din populații identice;
  - Nu poate fi respinsă ipoteza că PP, PT şi TT provin din populații identice;

÷ Analiza grupurilor de câte patru perechi de metode evidențiază statistic că:
  - Cu un risc de a fi în eroare de 5% în grupul de perechi de metode PP, PT, TP, TT cel puțin una provine din populație diferită (c/k = 0.9);

÷ Concluzia analizei grupurilor de perechi de metode este:
  - Cu un risc de a fi în eroare de 5% DD, PD, şi TD provin din populații diferite;
  - Nu poate fi respinsă ipoteza că DT şi DP provin din populații identice;
  - Nu poate fi respinsă ipoteza că TT şi TP provin din populații identice;
  - Nu poate fi respinsă ipoteza că PP, PT şi TT provin din populații identice;
  - Cu un risc de a fi în eroare de 5% se respinge ipoteza că PP, PT, TP şi TT ar proveni din populații identice;

Interpretarea statistică prezentată în Tabelul din Anexa 7-1 permite partiția strategiei de selecție şi supraviețuire după populația de proveniență a numărului mediu de genotipuri viabile în cultivar în generațiile ce produc evoluție pe durata evoluției. Tabelul 9-2 sintetizează acest rezultat. Figura 9-1 redă numărul mediu de genotipuri viabile în cultivar în generațiile ce produc evoluție pe durata evoluției pentru fiecare strategie de selecție şi supraviețuire inclusă în studiu.

În Figura 9-1 se observă (şi rezultatele cumulate în Tabelul din Anexa 7-1 demonstrează statistic) că selecția deterministă însoțită de supraviețuirea deterministă (DD) au ca efect menținerea unui număr mediu de genotipuri viabile în cultivar ridicat, a cărei lege de distribuție populațională diferă semnificativ statistic (cu un risc de 5% de a fi în eroare, Tabelul din Anexa 7-1 şi Tabelul 9-2) de celelalte opțiuni de selecție şi supraviețuire.

Tabelul 9-2

Populații pentru numărul mediu de genotipuri viabile în cultivar în generațiile ce produc evoluție pe durata evoluției în funcție de strategia de selecție şi supraviețuire

Populations for the average number of the viable genotypes in cultivar in the generations producing evolution during evolving depending on the selection and survival strategies

Sel\Srv	Proportional	Tournament	Deterministic
Proporțional	PP	PT	PD
Tournament	TP	TT	TD
Deterministic	DP	DT	DD

Sel: Selecție (în linii) vs. Srv: Supraviețuire (în coloane)

Sel: Selection (in rows) vs. Srv: Survival (in columns)

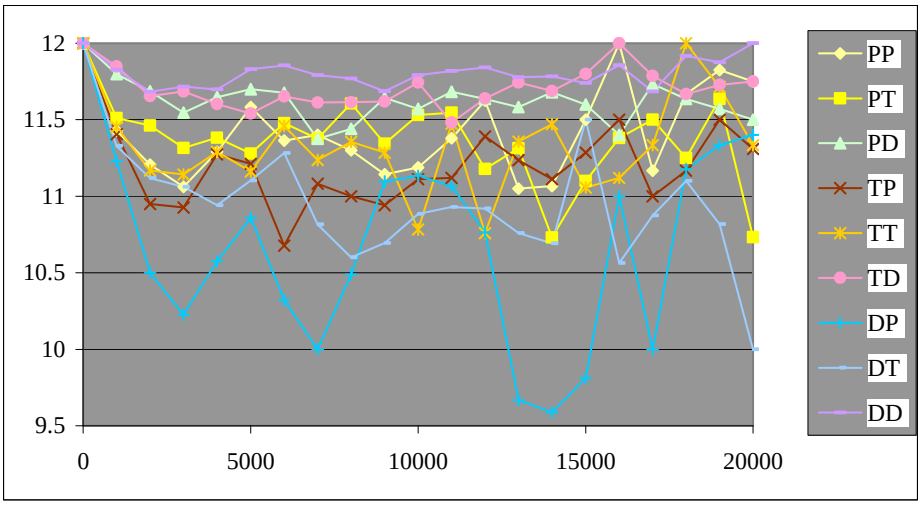


Figura 9-1. Numărul mediu de genotipuri viabile în cultivar la momentele evoluției în funcție de selecție (prima literă) şi supraviețuire (a doua literă)

Average of the number of viable genotypes depending on selection (first letter) and survival (second letter)

Caracterizate de variabilitate ridicată (Figura 9-1), selecția deterministă însoțită de supraviețuirea în turnir (DT) sau proporțională (DP) produc o distribuție populațională ce diferă semnificativ statistic (cu un risc de 5% de a fi în eroare, Tabelul din Anexa 7-1 şi Tabelul 9-2) de populațiile produse de celelalte posibilități de selecție şi supraviețuire.

Supraviețuirea deterministă susținută de o selecție în turnir (TD) produce o distribuție populațională ce diferă semnificativ statistic (cu un risc de 5% de a fi în eroare, Tabelul din Anexa 7-1 şi Tabelul 9-2, Figura 9-1) de populațiile produse de celelalte posibilități de selecție şi supraviețuire ale cărei caracteristici însă, de valoare medie şi variabilitate sunt apropiate de cele ale populației produse de selecția deterministă însoțită de supraviețuirea deterministă (DD) pe de o parte (Figura 9-1) şi de selecția deterministă susținută de supraviețuirea proporțională (DP) de cealaltă

parte (Figura 9-1).

Supraviețuirea deterministă susținută de o selecție proporțională (PD) produce o distribuție populațională ce diferă semnificativ statistic (cu un risc de 5% de a fi în eroare, Tabelul din Anexa 7-1, Tabelul 9-2 şi Figura 9-1) de populațiile produse de celelalte posibilități de selecție şi supraviețuire ale cărei caracteristici însă, de valoare medie şi variabilitate sunt apropiate de cele ale populației produse de selecția deterministă însoțită de supraviețuirea în turnir (DT, Figura 9-1).

Selecțiile şi supraviețuirile care nu sunt nici una deterministe formează un grup relativ compact (Figura 9-1) şi produc populații relativ apropiate (Tabelul 9-2) în ceea ce privește numărul mediu de genotipuri viabile din cultivar pe parcursul evoluției.

Statistica Anderson-Darling aplicată datelor din Tabelul din Anexa 7-1 a urmărit diferențele semnificative statistic în ceea ce privește legile de distribuție ale populațiilor din care provin eşantioanele supuse investigației. Pornind de la acest rezultat, se pot separa populațiile statistice (Figura 9-2, valori medii pe intervale de mii de generații).

Tendințele medii ale evoluțiilor din Figura 9-2 se pot obține calculând pentru fiecare clasă de frecvențe (de la 0 la 20 de mii) valoarea medie a variabilei observate (numărul mediu de genotipuri viabile - media mediilor) de la evoluția generației 0 până la momentul (de timp al) clasei de frecvență. Tendințele medii din Figura 9-2 corespund valorilor care s-ar obține dintr-o simulare Monte-Carlo [4] - clasă de algoritmi înrudită cu algoritmii genetici descoperită în 1947 (Metropolis & Ulam, 1949) şi folosită cu precădere în statistica computațională - a evoluției numărului mediu de genotipuri viabile (tendințe medii reprezentate în Figura 9-3).

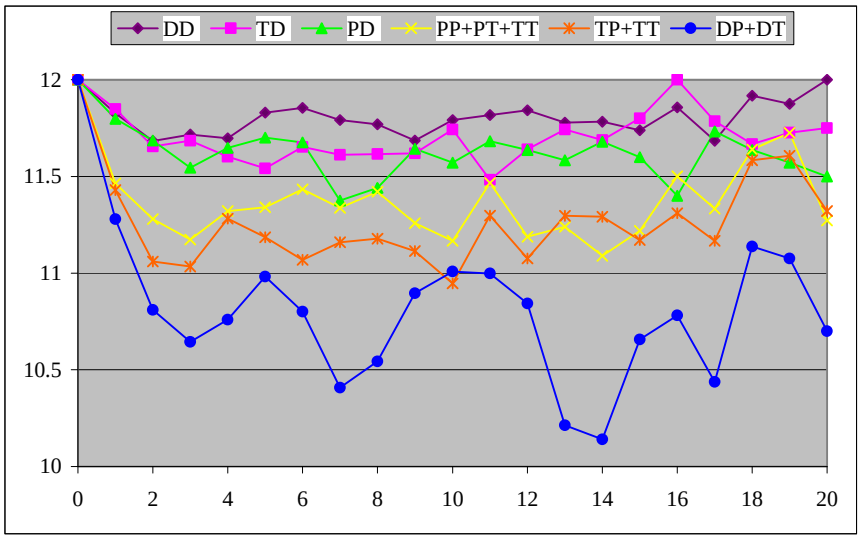


Figura 9-2. Număr mediu de genotipuri viabile reprezentate în cultivar la momentele evoluției: populații distincte semnificativ statistic partajate de selecție şi supraviețuire

Average of the number of viable genotypes: significantly different populations partitioned by selection and survival

[4] Metodele Monte Carlo sunt o clasă de algoritmi de calcul care se bazează pe eşantionarea repetată pentru a obține rezultatul, fiind frecvent folosite pentru simularea sistemelor fizice şi chimice; metodele Monte-Carlo sunt utilizate atunci când nu este fezabil sau este imposibil a se obține un rezultat folosind un algoritm deterministic.

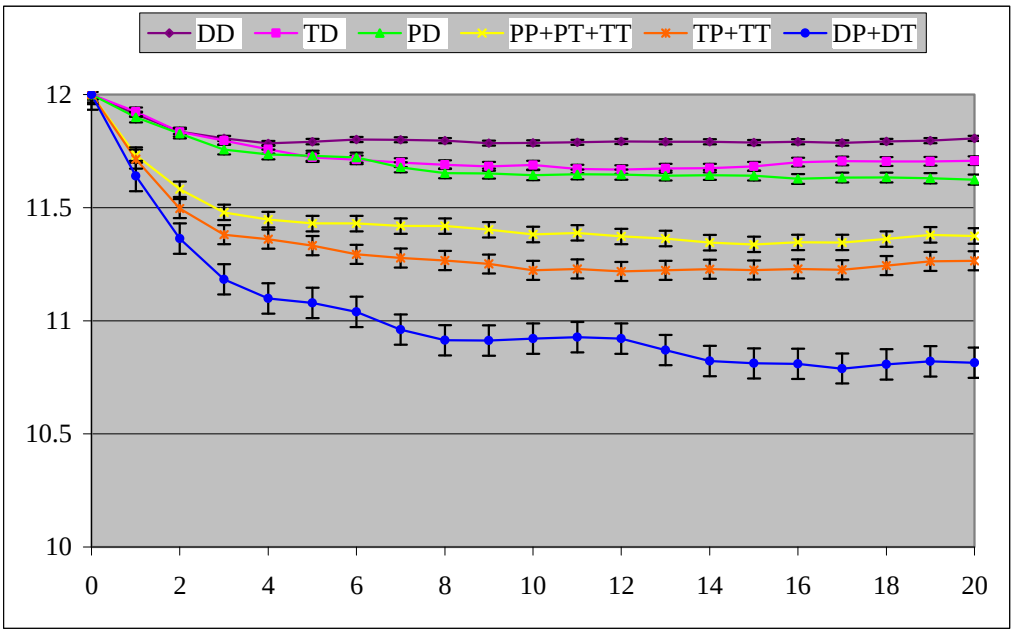


Figura 9-3. Tendințe medii pentru numărul de genotipuri viabile la momentele evoluției: populații distincte semnificativ statistic partajate de selecție și supraviețuire

Average tendencies for the average of the number of viable genotypes: statistically significant different populations partitioned by selection and survival

După cum se observă din Figura 9-3 (în care au fost reprezentate și erorile standard), diferențele care apar în tendințele medii între populațiile produse de DD, TD și PD sunt foarte mici, intervalele de eroare standard fiind suprapuse frecvent pentru perechea (TD, PD); în ciuda acestui fapt, populațiile produse de DD, TD și PD sunt într-adevăr distincte, așa cum statistica Anderson-Darling a pus în evidență (Tabelul din Anexa 7-1).

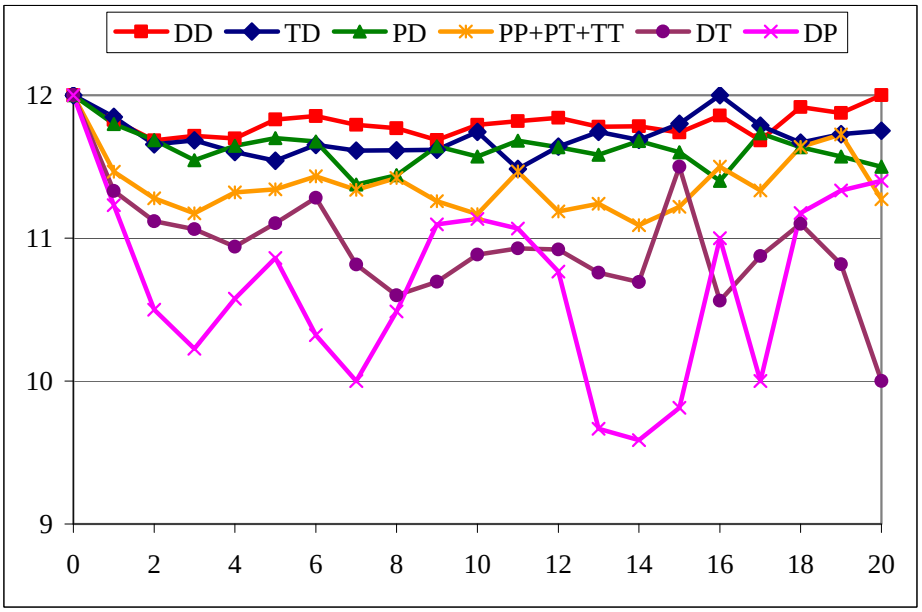


Figura 9-4. Partiția numărului de genotipuri reprezentate în cultivar: soluția cea mai probabilă

Partition of the number of genotypes represented in cultivar: the most probable solution

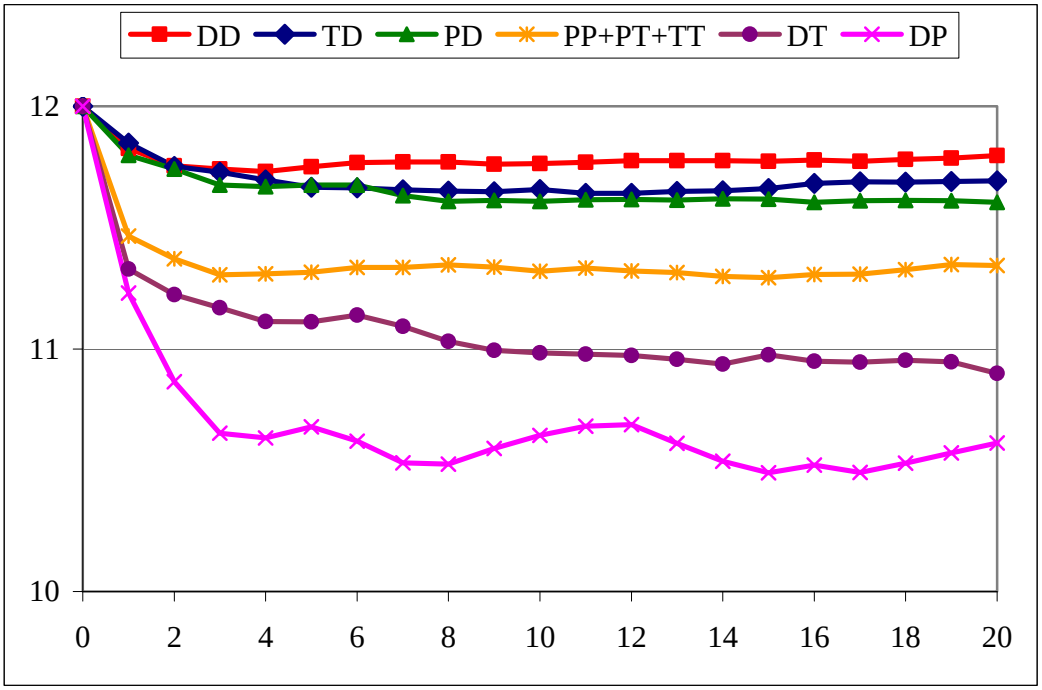


Figura 9-5. Partiția numărului de genotipuri reprezentate în cultivar: media mediilor
Partition of the number of genotypes represented in cultivar: the mean of means

Soluția propusă de Tabelul 9-2 poate fi rafinată suplimentar pentru obținerea soluției celei mai probabile de asociere care este în concordanță cu soluțiile posibile oferite de analiza din Tabelul din Anexa 7-1. Astfel dacă se acceptă că nivelul de încredere de 95% nu a fost suficient pentru asocierile observate în Tabelul din Anexa 7-1 şi se procedează la eliminarea succesivă a celor mai improbabile asocieri până când rezultatul obținut produce o partiție în asocierea metodelor, această eliminare se oprește după eliminarea asocierilor (DP, DT) - cu c/k = 1.2 şi c/k = 1.3, când partiția metodelor este obținută: {DD, TD, PD, DT, DP, (PP, PT, TT)} - aceasta având un singur grup în care se situează trei perechi de selecție şi supraviețuire - PP, PT şi TT. Figura 9-4 redă evoluția numărului mediu de genotipuri reprezentate în cultivar iar Figura 9-5 media mediilor numărului mediu de genotipuri reprezentate în cultivar.

Analiza numărului de fenotipuri viabile din cultivar în generațiile ce au produs evoluție

Tabelul 9-3 cumulează rezultatele obținute pentru numărul de fenotipuri viabile observate (*sum_obs*) şi media numărului acestora (*avg_obs*) pentru acelaşi număr de evoluții observate (*num_obs*) ale căror valori au fost redate în Tabelul 8-1, fiind aici grupate după mii de generații (de la 0..1000 până la 19001-20000).

Ipoteza cu privire la numărul mediu de fenotipuri viabile (valori redate în Tabelul 9-3) şi anume "*Strategia de selecție şi strategia de supraviețuire produc populații distincte pentru numărul mediu de fenotipuri viabile?*" s-a verificat folosind testul Anderson-Darling şi rezultatul analizei este redat în Tabelul din Anexa 7-2.

Se observă din analiza statistică (Tabelul din Anexa 7-2) că interpretarea populațională a numărului mediu de fenotipuri din cultivar în generațiile ce produc evoluție (Tabelul 9-3) este mult mai complexă decât interpretarea populațională (Tabelul din Anexa 7-1) a numărului mediu de genotipuri (Tabelul 9-1) în aceleaşi evoluții.

Tabelul 9-3

Frecvența fenotipurilor viabile din cultivar în momentele evoluției (valori cumulate din 46 de execuții independente) în pentru fiecare pereche de strategie de selecție (Sel = P, T, D) şi strategie de supraviețuire (Srv = P, T, D)

The frequency (cummulated from 46 experiments) of the viable genotypes in cultivar when evolution occurs for every pair of selection (Sel = P, T, D) and survival (Srv = P, T, D) strategy

1000G	SelSrv	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
avg_obs	PP	17.9	17.9	17.3	17.8	17.9	18.2	19.2	17.9	17.4	17.9	19.6	18.1	18.0	17.5	17.9	18.0	19.1	19.0	19.4	18.9
sum_obs	PP	13117	2529	1677	1105	643	799	479	537	610	573	567	289	360	262	143	126	229	400	330	377
avg_obs	PT	18.2	18.5	18.5	19.8	20.7	19.3	19.0	19.3	17.2	18.0	18.8	18.1	18.5	15.9	17.7	18.3	16.5	17.5	16.0	14.6
sum_obs	PT	12968	3157	2464	1766	1118	851	758	636	551	576	207	508	592	414	177	385	99	140	176	219
avg_obs	PD	18.1	17.7	18.4	17.7	17.4	18.5	17.2	18.4	18.1	18.1	17.7	17.3	18.5	18.0	16.0	17.4	18.9	17.1	17.5	15.8
sum_obs	PD	13541	2809	1862	1502	1046	1315	413	461	507	381	389	570	445	449	160	261	284	377	245	285
avg_obs	TP	18.0	18.9	18.9	19.2	19.4	19.2	18.0	18.6	19.8	19.0	19.6	19.0	18.6	20.4	18.6	19.6	17.9	18.1	18.5	19.4
sum_obs	TP	11766	3004	2077	1692	1106	594	666	614	674	514	489	436	317	184	130	157	161	217	185	252
avg_obs	TT	18.5	18.3	17.8	17.9	18.2	18.1	19.1	18.9	19.3	18.6	18.8	17.7	18.1	19.0	18.9	18.2	18.1	19.0	19.9	17.4
sum_obs	TT	13681	3254	1640	1110	1021	671	726	587	406	427	358	443	254	323	340	456	163	19	139	209
avg_obs	TD	18.1	18.5	18.0	18.3	18.5	18.2	18.0	18.0	18.2	18.6	19.1	20.0	17.9	17.8	20.4	19.0	17.8	18.3	17.7	17.1
sum_obs	TD	13715	2037	1762	1522	1131	1200	883	703	382	576	553	500	556	284	102	76	249	165	195	274
avg_obs	DP	16.2	15.1	14.6	15.7	15.7	14.6	13.9	14.3	14.7	15.4	16.7	16.2	14.5	12.6	14.6	16.4	14.2	15.5	15.8	17.8
sum_obs	DP	6817	1418	642	518	676	409	209	613	457	231	251	276	174	215	233	148	213	263	95	89
avg_obs	DT	16.3	15.5	15.3	15.3	15.5	15.6	15.4	14.3	14.9	15.5	13.8	14.0	15.2	15.4	16.4	14.7	15.1	15.6	15.2	12.5
sum_obs	DT	7044	1704	996	780	745	609	585	643	538	665	193	351	441	200	131	235	121	156	167	25
avg_obs	DD	17.8	18.2	19.1	18.1	18.3	17.3	17.7	18.4	19.9	18.6	17.7	19.2	19.6	17.8	19.0	17.9	17.4	19.5	17.6	16.5
sum_obs	DD	8284	2189	1547	1195	969	711	424	717	1015	446	195	364	352	409	436	250	330	234	423	33

avg_obs: media observațiilor (*mean of observations*); sum_obs: suma observațiilor (*sum of observations*);
1000G: mii generații (*thousands of generations*)

Interpretarea rezultatelor din Tabelul din Anexa 7-2 necesită identificarea celor mai mari grupuri de perechi de metode cu populații posibil identice (numită în continuare lista suspecților - suspecți de a proveni din populații identic distribuite) şi eliminarea sub-grupurilor unice ale acestora de ordin inferior; astfel cele mai mari grupuri de ordin maxim sunt grupurile de ordin 5 (PP, PT, TT, TD, DD) - run 400 - şi (PP, PT, PD, TD, DD) - run 448 - care intră în mod automat în lista suspecților; grupurile de ordin inferior sunt după cum urmează din aplicarea algoritmului de incluziune:

÷ Grupuri de ordin 5; lista suspecților: {(PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD)}

- (PP, PT, TD, DD) - run 384 - e simultan în (PP, PT, TT, TD, DD) şi (PP, PT, PD, TD, DD) ceea ce face imposibilă decelarea apartenenței sale; se adaugă la lista suspecților;
- (PT, TT, TD, DD) - run 145 - e doar în (PP, PT, TT, TD, DD); se elimină;
- (PP, TT, TD, DD) - run 272 - e doar în (PP, PT, TT, TD, DD); se elimină;
- (PP, PT, TT, DD) - run 392 - e doar în (PP, PT, TT, TD, DD); se elimină;
- (PP, PT, TT, TD) - run 399 - e doar în (PP, PT, TT, TD, DD); se elimină;
- (PP, PT, PD, DD) - run 440 - e doar în (PP, PT, PD, TD, DD); se elimină;

- (PT, PD, TD, DD) - run 193 - e doar în (PP, PT, PD, TD, DD); se elimină;
- (PP, PT, PD, TD) - run 447 - e doar în (PP, PT, PD, TD, DD); se elimină;

÷ Grupuri de ordin cel puțin 4; lista suspecților: {(PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD), (PP, PT, TD, DD)}

- (PT, TD, DD) - run 129, (PP, PT, DD) - run 376, (PP, TD, DD) - run 256, (PP, PT, TD) - sunt simultan în (PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD) şi (PP, PT, TD, DD); se adaugă;
- (TT, TD, DD) - run 20, (PT, TT, TD) - run 144, (PT, TT, DD) - run 137, (PP, TT, TD) - run 271, (PP, TT, DD) - run 264 şi (PP, PT, TT) - run 391 - sunt doar în (PP, PT, TT, TD, DD); se elimină;
- (PT, PD, DD) - run 185, (PP, PT, PD) - run 439 şi (PP, PD, DD) - run 312 - sunt doar în (PP, PT, PD, TD, DD); se elimină;

÷ Grupuri de ordin cel puțin 3; lista suspecților: {(PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD), (PP, PT, TD, DD), (PT, TD, DD), (PP, PT, DD), (PP, TD, DD), (PP, PT, TD)}

- (PT, DD) - run 121 - în run 376, 129, 384, 448, 400; se adaugă;
- (TT, TD) - run 19, (PT, TT) - run 136, (TT, DD) - run 12, (PP, TT) - run 263 - în run 400; se elimină;
- (PP, DD) - run 248 - în run 256, 376, 384, 448, 400; se adaugă;
- (TD, DD) - 5 în 256, 129, 384, 448, 400; se adaugă;
- (PT, TD) - 128 - în 383, 129, 384, 448, 400; se adaugă;
- (PP, TD) - 255 - în 383, 256, 384, 448 şi 400; se adaugă;
- (DP, DT) - run 3 - nu se regăseşte în grupurile de ordin superior; se adaugă;
- (PP, PT) - 375 - în 383, 376, 384, 448, 400; se adaugă;
- (PT, PD) - run 184 şi (PD, DD) - în run 448; se elimină;
- (TP, TT) - run 42 - nu se regăseşte în grupurile de ordin superior; are însă cea mai mare susceptibilitate situându-se foarte aproape de confidența de 95 pentru respingere (c/k = 1.09); se elimină;

÷ Grupuri de ordin cel puțin 2; lista suspecților: {(PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD), (PP, PT, TD, DD), (PT, TD, DD), (PP, PT, DD), (PP, TD, DD), (PP, PT, TD), (PT, DD), (PP, DD), (TD, DD), (PT, TD), (PP, TD), (DP, DT), (PP, PT)}; se impune continuarea includerii în listă cu populațiile distincte, urmând aceeaşi procedură:

- PP se regăseşte în mai multe grupuri de ordin superior; se adaugă;
- PT: în mai multe grupuri de ordin superior; se adaugă;
- PD: într-un sigur grup de ordin superior: (PP, PT, PD, TD, DD); se elimină;
- TP: nu există în grupuri de ordin superior; se adaugă;

- TT: într-un sigur grup de ordin superior: (PP, PT, TT, TD, DD); se elimină;
- TD: în mai multe grupuri de ordin superior; se adaugă;
- DP: într-un sigur grup de ordin superior: (DP, DT); se elimină;
- DT: într-un sigur grup de ordin superior: (DP, DT); se elimină;
- DD: în mai multe grupuri de ordin superior; se adaugă;

÷ Toate grupurile distincte; lista finală: {(PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD), (PP, PT, TD, DD), (PT, TD, DD), (PP, PT, DD), (PP, TD, DD), (PP, PT, TD), (PT, DD), (PP, DD), (TD, DD), (PT, TD), (PP, TD), (DP, DT), (PP, PT), PP, PT, TP, TD, DD};

Suntem acum în măsură să întocmim lista posibilelor populații fenotipice împreună cu nivelul de încredere asociat. Pentru a obține nivelul de încredere asociat acestora este necesară din nou inspecția valorilor conținute în Tabelul din Anexa 7-2. Rezultatul acestei investigații este redat în Tabelul 9-4, clasificat după nivelul de confidență al rezultatului obținut.

Tabelul 9-4

Numărul de fenotipuri din cultivar pe parcursul evoluțiilor: populații parametrizate de selecție şi supraviețuire cu legi de distribuție distincte

Number of phenotypes in cultivar during evolutions: populations parameterized by selection and survival having distinct distribution laws

Populație *Population*	M(c/k) [n_obs]	Categorie *Category*	Interpretare şi nivel de semnificație *Interpretation and their risk being in error*
(DP, DT)	0.194 [128]	1	Cu un risc de a fi în eroare de 5% produce o populație diferită de toate celelalte perechi de metode.
TP	0.293 [255]		
(PP, PT, PD, TD, DD)	0.368 [16]	2	Nu poate fi respinsă ipoteza (cu o încredere de 95%) că provin din populații identice
(PP, PT, TT, TD, DD)	0.409 [16]		
PP	0.390 [255]	3.1	Nu se poate respinge ipoteza identic distribuției sale (cu o încredere de 95%) cu mai multe populații
TD	0.405 [255]		
DD	0.417 [255]		
PT	0.417 [255]		
(PP, TD)	0.400 [128]	3.2	Nu se poate respinge ipoteza identic distribuției sale (cu o încredere de 95%) cu mai multe populații
(PP, PT)	0.407 [128]		
(PP, DD)	0.417 [128]		
(PT, TD)	0.419 [128]		
(TD, DD)	0.419 [128]		
(PT, DD)	0.444 [128]		
(PP, PT, TD)	0.407 [64]	3.3	Nu se poate respinge ipoteza identic distribuției sale (cu o încredere de 95%) cu mai multe populații
(PP, TD, DD)	0.413 [64]		
(PP, PT, DD)	0.424 [64]		
(PT, TD, DD)	0.430 [64]		
(PP, PT, TD, DD)	0.422 [32]	3.4	Nu se poate respinge ipoteza identic distribuției sale (cu o încredere de 95%) cu mai multe populații
Legendă: M(c/k) [n_obs] - Valoare medie c/k (număr de observații care au intrat în calculul mediei) Legend: M(c/k) [n_obs] - Mean of c/k (number of observations in calculation of the mean) Notă: Cu cât mai mică este valoarea medie c/k, cu atât mai relevant în ceea ce privește interpretarea; Valoarea c/k supraunitară atrage susceptibilitatea respingerii interpretării;			

Rămân următoarele grupuri ale căror evoluție în medie este de urmărit, pentru unele dintre

acestea neputând fi exclusă apartenența la populații identice, în timp ce pentru altele populația distinctă semnificativ statistic fiind evidențiată (Tabelul 9-4).

Tabelul 9-4 ne arată cum să separăm populațiile statistice. Astfel, în categoria 1 a semnificației intră populațiile identificate ca distincte vs. restul populațiilor (Figura 9-6 - valori medii pe intervale de mii de generații; Figura 9-7 - tendințe medii pe intervale de mii de generații).

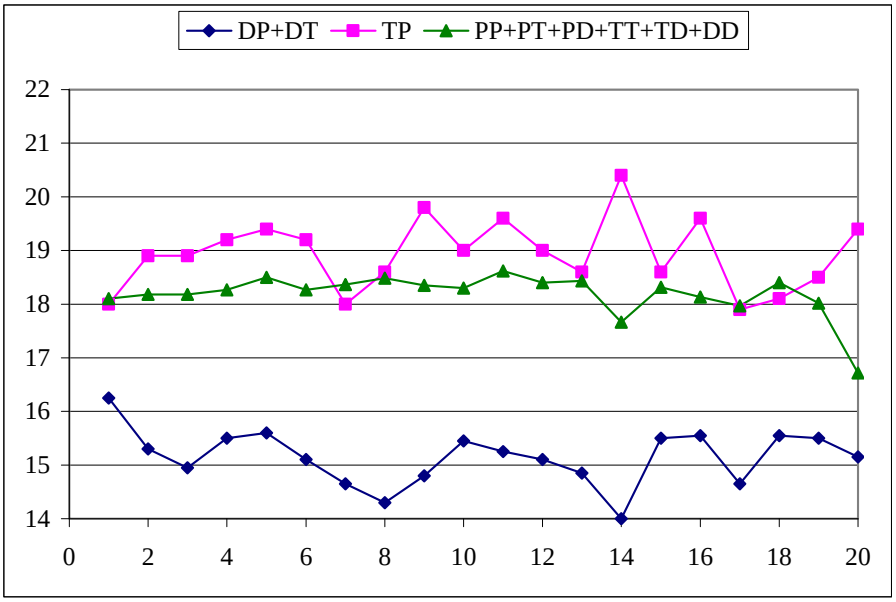


Figura 9-6. Segregarea populațiilor numărului mediu de fenotipuri viabile în momentele evoluției; populații: (DP, DT), TP și restul asocierilor de metode de selecție cu metode de supraviețuire (PP, PT, PD, TT, TD, DD)

Segregation of the populations of average number of viable genotypes in the moments of evolution; populations: (DP, DT), TP and the rest of of selection and survival strategies (PP, PT, PD, TT, TD, DD)

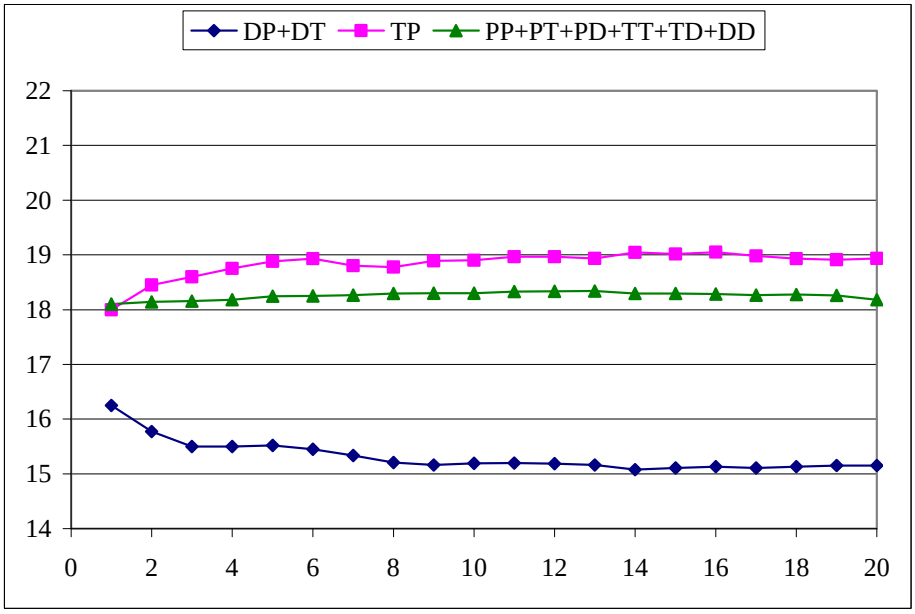


Figura 9-7. Tendințe medii ale populațiilor de fenotipuri viabile

Mean tendencies of the populations of viable phenotypes

Ceea ce evidențiază atât Figura 9-6 și mai ales Figura 9-7 este că numărul mediu de genotipuri se diferențiază nu numai cu privire la legea de distribuție a mediei pe parcursul evoluției

(aşa cum s-a sintetizat în Tabelul 9-4) ci şi sub aspectul valorii absolute a numărului mediu de fenotipuri din cultivar.

Selecția în turnir însoțită de o supraviețuire proporțională (TP) produce o populație distinctă de fenotipuri în cultivar sub aspectul numărului acestora, care în valoare medie se situează aproape în toate momentele evoluției (Figura 9-6) deasupra mediei produse de metodele de selecție şi supraviețuire rămase (PP, PT, PD, TT, TD, DD) şi (Figura 9-7) numărul mediu de fenotipuri ale populației produse de selecția în turnir însoțită de o supraviețuire proporțională (TP) creşte aproape peste tot cu aproape un fenotip față de restul metodelor (medie de 18.94 pentru TP; medie de 18.18 pentru PP+PT+PD+TT+TD+DD), şi cu peste un fenotip decât media generală (17.59), având tendință crescătoare spre 19 fenotipuri pe parcursul evoluției. Numărul mediu de fenotipuri produs de selecția deterministă însoțită de o supraviețuire proporțională (DP) sau în turnir (DT), creează o populație distinctă de fenotipuri în cultivar sub aspectul numărului acestora, care în valoare medie se situează în toate momentele evoluției (Figura 9-6) sub media produsă de metodele de selecție şi supraviețuire rămase (PP, PT, PD, TT, TD, DD) şi (Figura 9-7) numărul mediu de fenotipuri ale populației produse de selecția deterministă însoțită de o supraviețuire proporțională (DP) sau în turnir (DT) produce scăderea cu cel puțin două fenotipuri aproape peste tot față de restul metodelor (medie de 15.15 pentru DP+DT; medie de 18.18 pentru PP+PT+PD+TT+TD+DD) şi cu aproape două fenotipuri față de media generală (17.59), având o tendință descrescătoare spre 15 fenotipuri pe parcursul evoluției.

Al doilea caz analizat este al celor mai numeroase populații pentru care nu s-a putut pune în evidență deosebirea legii de distribuție (Figura 9-8).

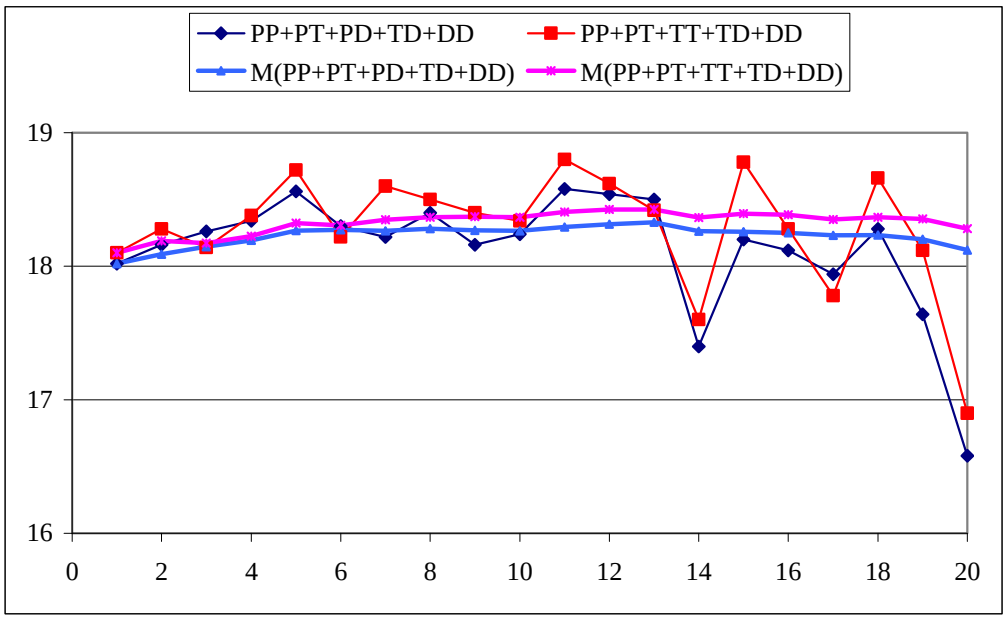


Figura 9-8. Cele mai numeroase populații la nivelul de semnificație de 5% pentru numărul de fenotipuri din cultivar pe parcursul evoluției

Largest populations at 5% risk being in error for the number of phenotypes during evolutions

Referitor la populațiile reprezentate în Figura 9-8 se remarcă prezența a patru perechi în ambele populații (PP, PT, TD şi DD) distincția făcându-se în fiecare dintre acestea de prezența celei de-a 5-a perechi (PD şi respectiv TT). Explicația se regăsește în Tabelul 9-4, populația (PP, PT, TD şi DD) având un scor mediu de ne-discriminare c/k de 0.422 care este chiar accentuat (scăzut) la 0.368 prin cuprinderea sub-populației metodei PD pe de o parte şi accentuat (scăzut) la 0.409 prin cuprinderea sub-populației metodei TT, ambele însă fiind suficient de diferite (scor în pereche de 0.42 - respinsă ipoteza de apartenență la populații identic distribuite cu riscul de 5% de a fi în eroare, Tabelul 9-4) pentru ca includerea lor împreună în grupul (PP, PT, TD, DD) să nu mai fie acceptată ca o singură distribuție de populație, grupul (PP, PT, TD, DD, PD, TT) având (Tabelul din Anexa 7-2, run 464) un scor de respingere a apartenenței la populații identic distribuite cu riscul de 5% de a fi în eroare de 0.92 (aproape de acceptare - valoare supraunitară - dar insuficientă). O remarcă importantă este că PD (selecția proporțională însoțită de supraviețuirea deterministă) şi TT (selecția în turnir însoțită de supraviețuirea în turnir) nu se regăsesc sub formă de populații distincte în conformitate cu schema de analiză descrisă de Tabelul 9-4, ceea ce a făcut imposibilă decizia unică asupra clasificării lor în grupul de ordin 5 ce conține grupurile PP, PT, TD şi DD.

Dacă analiza din Tabelul 9-4 satisface sub aspectul posibilităților, nu aduce răspunsuri sub aspectul probabilităților. Dacă fixăm drept obiectiv obținerea celei mai probabile partiții a populațiilor produse de strategia de selecție şi strategia de supraviețuire, atunci cel mai simplu este ca analiza din Tabelul 9-4 să se continue cu analiza perechilor de câte 2. Tabelul 9-5 descrie obținerea unei astfel de soluții.

Tabelul 9-5

Soluția cea mai probabilă la problema interpretării rezultatelor din Tabelul 9-4

The most probable solution interpreting the results from Table 9-4

Obținerea matricei perechilor; o celulă a matricei conține 0 dacă cu un risc de a fi în eroare de 5% perechea de metode dată de linia celulei şi coloana celulei produc populații cu legi de distribuție diferite şi conține 1 în caz contrar

SS	PP	PT	PD	TP	TT	TD	DP	DT	DD
PP	1	1	0	0	1	1	0	0	1
PT	1	1	1	0	1	1	0	0	1
PD	0	1	1	0	0	0	0	0	1
TP	0	0	0	1	1	0	0	0	0
TT	1	1	0	1	1	1	0	0	1
TD	1	1	0	0	1	1	0	0	1
DP	0	0	0	0	0	0	1	1	0
DT	0	0	0	0	0	0	1	1	0
DD	1	1	1	0	1	1	0	0	1

Se permută (schimbă pozițiile) linii din matricea perechilor şi se permută coloane (schimbă pozițiile) din matricea perechilor (păstrând asocierile) astfel încât să se obțină cele mai mari grupuri compacte de 1; o soluție de permutare este cea alăturată

SS	DD	PP	PT	PD	TP	TT	TD	DT	DP
TP	0	0	0	0	1	1	0	0	0
DP	0	0	0	0	0	0	0	1	1
DT	0	0	0	0	0	0	0	1	1
PP	1	1	1	0	0	1	1	0	0
PT	1	1	1	1	0	1	1	0	0
TT	1	1	1	0	1	1	1	0	0

	DD	1	1	1	1	0	1	1	0	0
	PD	1	0	1	1	0	0	0	0	0
Se marchează toate blocurile; obiectivul este să nu avem cifre de 1 în vecinătatea blocurilor, ceea ce semnifică că unele asocieri identificate la riscul de 5% de a fi în eroare sunt întâmplătoare	SS	DD	PP	PT	PD	TP	TT	TD	DT	DP
	DP	0	0	0	0	0	0	0	1	1
	DT	0	0	0	0	0	0	0	1	1
	TP	0	0	0	0	1	1	0	0	0
	PP	1	1	1	0	0	1	1	0	0
	PT	1	1	1	1	0	1	1	0	0
	TT	1	1	1	0	1	1	1	0	0
	TD	1	1	1	0	0	1	1	0	0
	DD	1	1	1	1	0	1	1	0	0
	PD	1	0	1	1	0	0	0	0	0
Eliminarea cifrelor de 1 din tabelă trebuie să țină seama de șansa de a fi în eroare; astfel inspectând valorile din Tabelul 9-4 se observă că asocierile aflate cel mai aproape de pragul de 5% de eroare sunt (PD, DD) şi (TP, TT), ambele cu o rație c/k de 1.09; astfel aceste asocieri sunt primele care sunt asociate întâmplării şi se elimină	SS	DD	PP	PT	PD	TP	TT	TD	DT	DP
	DP	0	0	0	0	0	0	0	1	1
	DT	0	0	0	0	0	0	0	1	1
	TP	0	0	0	0	1	0	0	0	0
	PP	1	1	1	0	0	1	1	0	0
	PT	1	1	1	1	0	1	1	0	0
	TT	1	1	1	0	0	1	1	0	0
	TD	1	1	1	0	0	1	1	0	0
	DD	1	1	1	0	0	1	1	0	0
	PD	0	0	1	1	0	0	0	0	0
Este uşor de observat acum că doar asocierea (PD, PT) împiedică realizarea partiției de populații de către perechea (metodă de selecție, metodă de supraviețuire); inspectând valoarea acestei asocieri în Tabelul 9-4 se confirmă presupunerea, şi anume după eliminarea asocierilor (PD, DD) şi (TP, TT) asocierea (PD, PT) a rămas asocierea cu cea mai scăzută rație c/k: 1.26; se elimină asocierea (PD, PT)	SS	DD	PP	PT	PD	TP	TT	TD	DT	DP
	DP	0	0	0	0	0	0	0	1	1
	DT	0	0	0	0	0	0	0	1	1
	TP	0	0	0	0	1	0	0	0	0
	PP	1	1	1	0	0	1	1	0	0
	PT	1	1	1	0	0	1	1	0	0
	TT	1	1	1	0	0	1	1	0	0
	TD	1	1	1	0	0	1	1	0	0
	DD	1	1	1	0	0	1	1	0	0
	PD	0	0	0	1	0	0	0	0	0
Liniile şi coloanele tabelului se mai pot permuta pentru convenință	SS	DD	PP	PT	TT	TD	DT	DP	PD	TP
	TP	0	0	0	0	0	0	0	0	1
	PD	0	0	0	0	0	0	0	1	0
	DP	0	0	0	0	0	1	1	0	0
	DT	0	0	0	0	0	1	1	0	0
	PP	1	1	1	1	1	0	0	0	0
	PT	1	1	1	1	1	0	0	0	0
	TT	1	1	1	1	1	0	0	0	0
	TD	1	1	1	1	1	0	0	0	0
	DD	1	1	1	1	1	0	0	0	0

Astfel, asocierea obținută prin procedura descrisă în Tabelul 9-5 a impus eliminarea celor mai slabe 3 asocieri de perechi de populații, şi anume (PD, DD), (TP, TT) şi (PD, PT), caz în care soluția (şi din acest motiv o numim cea mai probabilă) este dată de partiția: {(DD, PP, PT, TT, TD), (DT, DP), PD, TP}. Rezultatul obținut se poate reprezenta grafic pentru distribuția numărului mediu de fenotipuri în cultivar (Figura 9-9) şi pentru media numărului mediu (media mediilor) de

fenotipuri din cultivar (Figura 9-10).

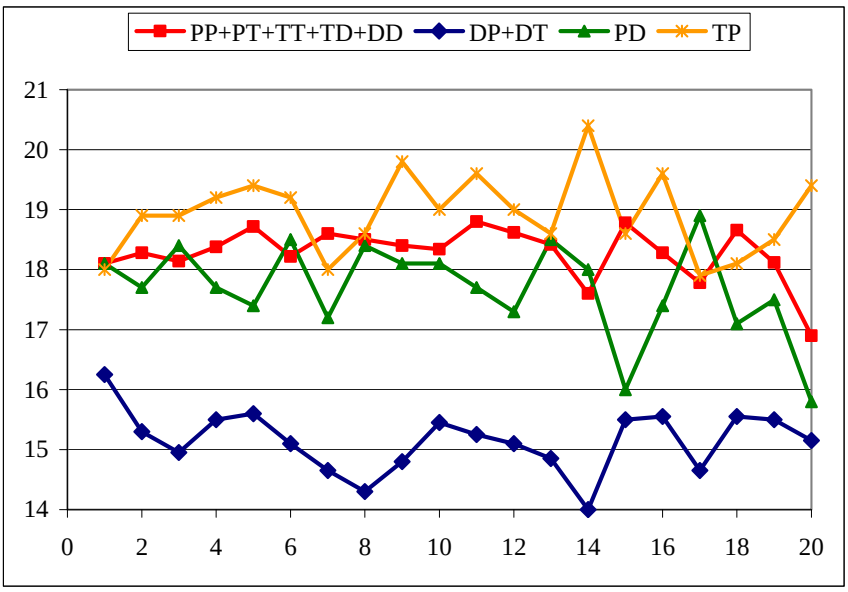


Figura 9-9. Cea mai probabilă partiție a populațiilor numărului de fenotipuri din cultivar parametrizată prin selecție şi supraviețuire: valorii medii pe milenii

The most probable partition of the populations of phenotypes number parameterized by selection and survival

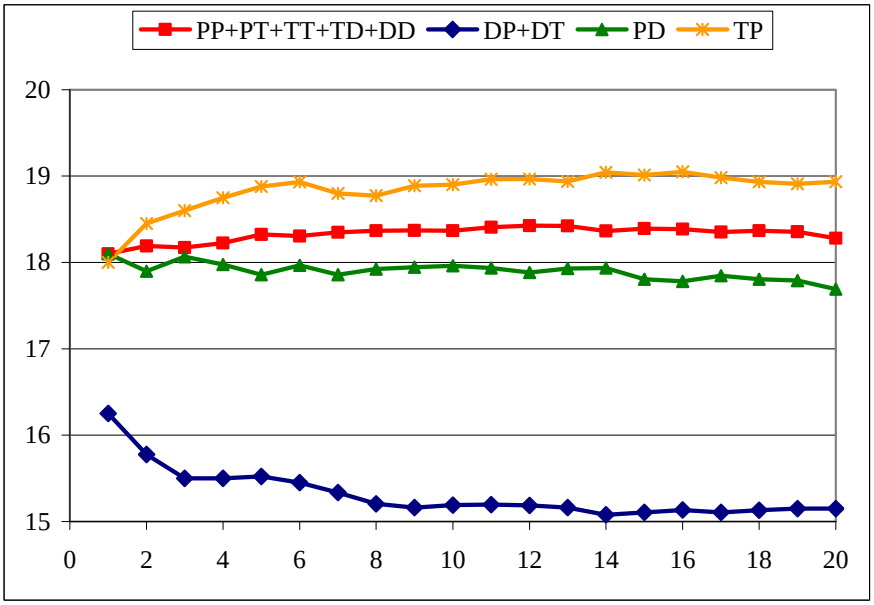


Figura 9-10. Populații distincte ale numărului de fenotipuri din cultivar: Experimentul Monte-Carlo

Distinct populations for the number of phenotypes: The Monte-Carlo experiment

Analiza numărului de asocieri (regresii) viabile din cultivar în generațiile ce au produs evoluție

Tabelul 9-6 cumulează rezultatele obținute pentru numărul de asocieri viabile (regresii cu parametrii semnificativ diferiți de zero statistic cu riscul de cel mult 5% de a fi în eroare) în ceea ce privește media numărului acestora (*avg_obs*) pentru același număr de evoluții observate (*num_obs*) fiind de asemenea grupate după mii de generații (de la 0..1000 până la 19001-20000).

Tabelul 9-6

Frecvența (din 46 de execuții independente) asocierilor viabile în momentele evoluției în funcție de selecție şi supraviețuire Sel, Srv ∈ {P, T, D}

Frequency (from 46 independent runs) of the viable associations in the moments of evolution depending on selection and survival Sel, Srv ∈{P, T, D}

SS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
PP	3793	3965	3532	4539	4569	4531	7444	4030	3599	3917	4594	3414	4301	3936	3788	2614	6620	4885	5054	4602
PT	4144	4872	5088	7798	8018	5068	5272	5347	2957	3959	4215	4104	5058	2404	5051	4403	3037	3448	1557	1574
PD	3264	3514	3800	3140	3024	4028	3343	4282	3440	3564	2919	3174	4354	3540	1868	2994	4747	2912	3501	2192
TP	3743	5506	5361	6637	7540	5593	4168	4687	6611	5077	6093	6260	4630	7280	6301	5013	4301	3704	4531	5459
TT	4909	5279	4958	4569	6587	4426	6074	5967	4917	4348	5513	3446	3762	6979	6163	4539	5769	5910	6100	5029
TD	3456	4023	3321	3905	4011	3655	3453	4034	3443	3641	4683	5390	3758	3905	6060	4418	2621	4099	3070	2590
DP	1786	1448	1026	2005	1698	999	862	1293	1328	1413	2473	1679	1192	772	1205	1788	1250	1422	1777	2800
DT	2172	1595	1513	1524	1532	1878	1726	1180	1512	1992	1117	719	1585	1487	2058	1428	1016	1757	1492	572
DD	2795	3158	4088	3617	3398	2930	3014	3647	4735	3352	2523	4104	5902	3341	4282	2853	2100	4008	2445	1605

Ipoteza cu privire la numărul mediu de asocieri viabile (valori redate în Tabelul 9-6) şi anume dacă metoda se selecție şi strategia de supraviețuire produc populații distincte s-a verificat folosind statistica Anderson-Darling şi rezultatul analizei este redat în Tabelul din Anexa 7-3.

Tabelul din Anexa 7-3 arată că pentru un număr de 10 grupuri de perechi de metode nu s-a putut pune în evidență o diferență statistică semnificativă între legile de distribuție ale acestora. Interpretarea rezultatelor din Tabelul din Anexa 7-3 se face în acelaşi mod în care s-au interpretat datele prezentate din Tabelul din Anexa 7-2.

Cele mai mari grupuri de ordin maxim ne-discriminate sunt grupurile de ordin 3 (PD, TD, DD) - run 66 - şi (PP, PT, TD) - run 383 - care intră în mod automat în lista suspecților; grupurile de ordin inferior sunt după cum urmează din aplicarea algoritmului de incluziune:

÷ Grupuri de ordin 2; lista suspecților: {(PD, TD, DD), (PP, PT, TD)}
  - (PD, DD) - run 58, (TD, DD) - run 5, şi (PD, TD) - run 65 - sunt doar în (PD, TD, DD); se elimină;
  - (TP, TT) - run 42 şi (DP, DT) - run 3 - nu se regăsesc în nici unul din grupurile de ordin superior; se adaugă;
  - (PP, PT) - run 375, (PP, TD) - run 255, şi (PT, TD) - run 128 - sunt doar în (PP, PT, TD); se elimină;

÷ Grupuri de ordin cel puțin 2; lista suspecților: {(PD, TD, DD), (PP, PT, TD), (TP, TT), (DP, DT)}; se impune continuarea includerii în listă cu populațiile distincte, urmând aceeaşi procedură:
  - PP, PT se regăsesc într-un singur grup (PP, PT, TD); se elimină;
  - PD, DD: într-un sigur grup (PD, TD, DD); se elimină;
  - TP, TT: într-un sigur grup (TP, TT); se elimină;
  - DP, DT: într-un singur grup (DP, DT); se elimină;
  - TD: în două de ordin superior; se adaugă;

÷ Toate grupurile distincte; lista finală: {(PD, TD, DD), (PP, PT, TD), (TP, TT), (DP, DT), TD};

Suntem acum în măsură să întocmim lista posibilelor asocieri fenotipice împreună cu nivelul de încredere asociat. Pentru a obține nivelul de încredere este necesară din nou inspecția valorilor conținute în Tabelul din Anexa 7-3. Rezultatul acestei investigații este redat în Tabelul 9-7, clasificat după nivelul de confidență al rezultatului obținut, care conține acele grupuri pentru care apartenența la populații identic distribuite nu a fost respinsă statistic cu o confidență de 95%.

Tabelul 9-7

Numărul de asocieri fenotipice din cultivar pe parcursul evoluțiilor: populații cu legi de distribuție distincte în funcție de strategia de selecție şi strategia de supraviețuire

The number of phenotypes associations during evolution: populations with distinct distribution depending on selection and survival strategies

Populație *Population*	c/k	Interpretare la nivelul de semnificație de 5% *Interpretation at risk being in error of 5%*
(PD, TD, DD)	1.29	Nu poate fi respinsă ipoteza că provin din populații identice.
(PP, PT, TD)	1.12	Nu poate fi respinsă ipoteza că provin din populații identice.
(TP, TT)	4.96	Nu poate fi respinsă ipoteza că provin din populații identice.
(DP, DT)	3.15	Nu poate fi respinsă ipoteza că provin din populații identice.
TD	-	Se distribuie identic cu două populații.

O primă remarcă asupra informațiilor redate în Tabelul 9-7 este că valorile c/k ale asocierilor (TP, TT) şi (DP, DT) sunt departe de a putea fi respinsă ipoteza apartenenței la populații identic distribuite. Nu în aceeaşi situație se află grupurile (PD, TD, DD) şi (PP, PT, TD) a căror rație de semnificație c/k este foarte aproape (1.29 şi 1.12) de a fi respinsă apartenența la populații identic distribuite (1.00). Mai mult, numărul de posibilități de partiție pe care îl implică analiza sintetizată în Tabelul 9-7 este uşor de calculat. Sunt exact 3 posibilități (la care s-a calculat şi suma rațiilor c/k):

÷ {(PD, TD, DD), (TP, TT), (DP, DT), (PP, PT)}, $\Sigma c/k = 1.29 + 4.96 + 3.15 + 2.04 = 11.44$;

÷ {(PD, DD), (TP, TT), (DP, DT), (PP, PT, TD)}, $\Sigma c/k = 5.72 + 4.96 + 3.15 + 1.12 = 14.95$;

÷ {(PD, DD), (TP, TT), (DP, DT), (PP, PT), TD}, $\Sigma c/k = 5.72 + 4.96 + 3.15 + 2.04 = 15.87$;

Astfel, cea mai probabilă separație a populațiilor se produce după schema {(PD, DD), (TP, TT), (DP, DT), (PP, PT), TD}, schemă (Tabelul 9-8) care va fi analizată în continuare.

Tabelul 9-8

Cea mai probabilă partiție în selecție vs. supraviețuire după numărul mediu de asocierilor fenotipice din cultivar; grupuri notate crescător în ordinea probabilității de asociere

The most probable partition in selection vs. survival on average number of phenotypes associations observed in cultivar; groups ranked by the probability of association

	D	P	T
D	1	3	3
P	1	4	4
T	5	2	2

Un alt indicator al faptului că soluția obținută este cea mai probabilă este că asocierea de ordinul 3 (PP, PT, TD) este (vezi Tabelul 9-7) cea mai slabă asociere de ordinul 3 (c/k = 1.12) şi

simultan cu aceasta asocierea de ordinul 2 (PT, TD) este cea mai slabă asociere de ordinul 2 (c/k = 1.04) - vezi Tabelul din Anexa 7-3, ambele asocieri fiind astfel (singurele) respinse de soluția propusă {(PD, DD), (TP, TT), (DP, DT), (PP, PT), TD}. Tabelul 9-8 ne arată cum să separăm populațiile statistice în timp ce Figura 9-11 prezintă valorile medii pe intervale de mii de generații pentru fiecare populație identificată iar Figura 9-12 dă tendințele medii pe intervale de mii de generații.

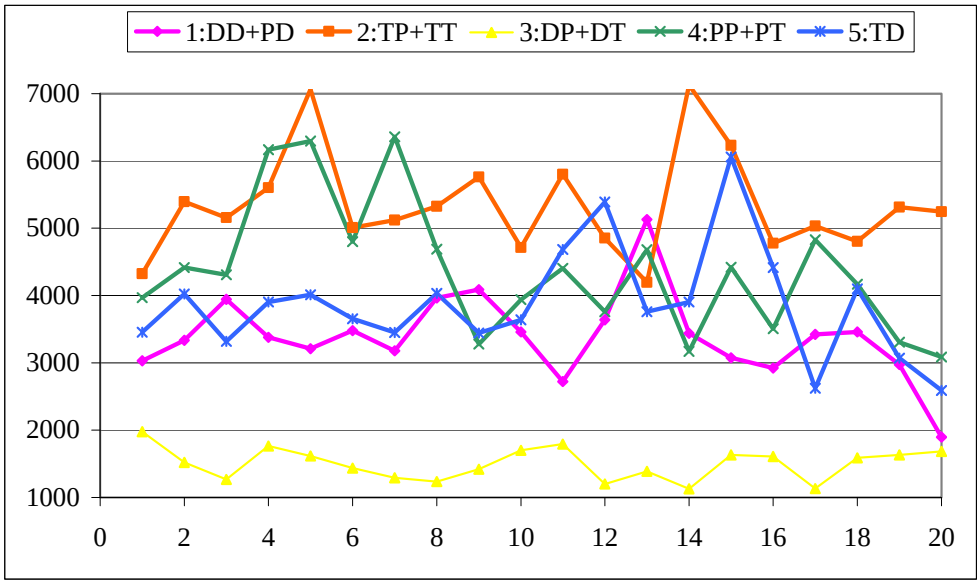


Figura 9-11. Număr de asocieri fenotipice în cultivar: populații distincte semnificativ statistic în raport cu strategia de selecție şi strategia de supraviețuire

Number of phenotypes associations in cultivar: significanly distinct populations in relation with selection and survival strategies

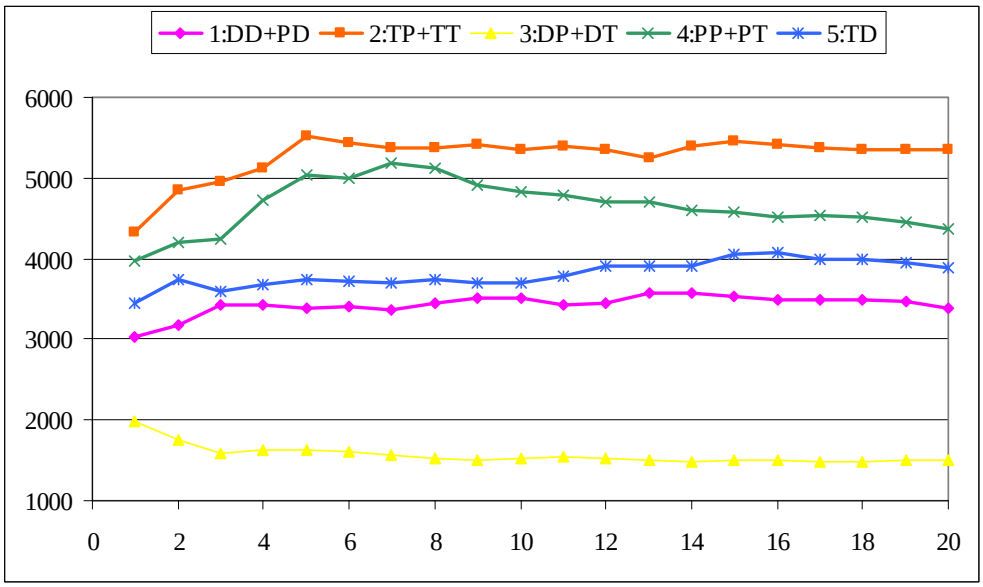


Figura 9-12. Tendințe medii ale asocierilor de fenotipuri reprezentate în Figura 9-11

Mean tendencies of phenotypes associations represented in Figure 9-11

Ceea ce evidențiază atât Figura 9-11 şi mai ales Figura 9-12 este că numărul mediu de genotipuri se diferențiază nu numai cu privire la legea de distribuție (așa cum s-a sintetizat în Tabelul 9-7) ci şi sub aspectul valorii absolute a numărului mediu de fenotipuri din cultivar.

Selecția în turnir însoțită de o supraviețuire proporțională (TP) produce o populație distinctă de fenotipuri în cultivar sub aspectul numărului acestora, care în valoare medie se situează aproape în toate momentele evoluției (Figura 9-11) deasupra mediei produse de metodele de selecție şi supraviețuire rămase (PP, PT, PD, TT, TD, DD) şi (Figura 9-12) numărul mediu de fenotipuri ale populației produse de selecția în turnir însoțită de o supraviețuire proporțională (TP) crește aproape peste tot cu aproape un fenotip față de restul metodelor (medie de 18.94 pentru TP; medie de 18.18 pentru PP+PT+PD+TT+TD+DD), şi cu peste un fenotip decât media generală (17.59), având tendință crescătoare spre 19 fenotipuri pe parcursul evoluției. Numărul mediu de fenotipuri produs de selecția deterministă însoțită de o supraviețuire proporțională (DP) sau în turnir (DT), creează o populație distinctă de fenotipuri în cultivar sub aspectul numărului acestora, care în valoare medie se situează în toate momentele evoluției (Figura 9-11) sub media produsă de metodele de selecție şi supraviețuire rămase (PP, PT, PD, TT, TD, DD) şi (Figura 9-12) numărul mediu de fenotipuri ale populației produse de selecția deterministă însoțită de o supraviețuire proporțională (DP) sau în turnir (DT) produce scăderea cu cel puțin două fenotipuri aproape peste tot față de restul metodelor (medie de 15.15 pentru DP+DT; medie de 18.18 pentru PP+PT+PD+TT+TD+DD) şi cu aproape două fenotipuri față de media generală (17.59), având o tendință descrescătoare spre 15 fenotipuri pe parcursul evoluției.

Al doilea caz analizat este al celor mai numeroase populații pentru care nu s-a putut pune în evidență deosebirea legii de distribuție (Figura 9-13).

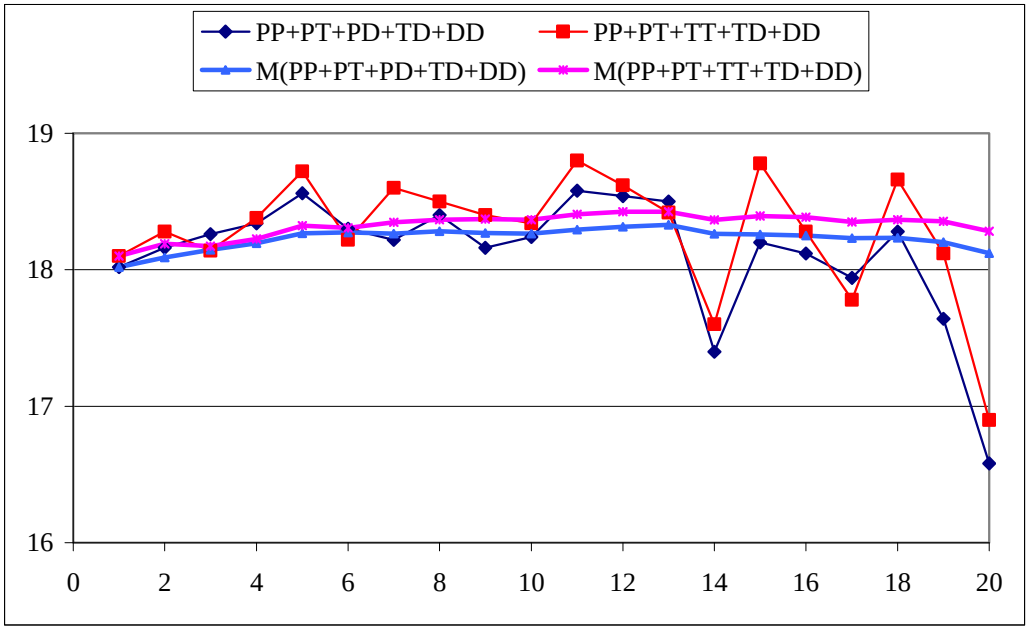


Figura 9-13. Numărul de fenotipuri din cultivar pe parcursul evoluției: cele mai numeroase populații nediferențiabile la nivelul de semnificație de 5%

Number of phenotypes during evolution: largest populations not differentiable at 5% risk being in error

Referitor la populațiile reprezentate în Figura 9-13 se remarcă prezența a patru perechi în ambele populații (PP, PT, TD şi DD) distincția făcându-se în fiecare dintre acestea de prezența celei

de-a 5-a perechi (PD şi respectiv TT). Explicația se regăseşte în Tabelul din Anexa 7-3, populația (PP, PT, TD şi DD) având un scor mediu de ne-discriminare c/k de 0.422 care este chiar accentuat (scăzut) la 0.368 prin cuprinderea sub-populației metodei PD pe de o parte şi accentuat (scăzut) la 0.409 prin cuprinderea sub-populației metodei TT, ambele însă fiind suficient de diferite (scor în pereche de 0.42 - respinsă ipoteza de apartenență la populații identic distribuite cu riscul de 5% de a fi în eroare, Tabelul din Anexa 7-3) pentru ca includerea lor împreună în grupul (PP, PT, TD, DD) să nu mai fie acceptată ca o singură distribuție de populație, grupul (PP, PT, TD, DD, PD, TT) având (Tabelul din Anexa 7-3, run 464) un scor de respingere a apartenenței la populații identic distribuite cu riscul de 5% de a fi în eroare de 0.92 (aproape de acceptare - valoare supraunitară - dar insuficientă). O remarcă importantă este că PD (selecția proporțională însoțită de supraviețuirea deterministă) şi TT (selecția în turnir însoțită de supraviețuirea în turnir) nu se regăsesc sub formă de populații distincte în conformitate cu schema de analiză descrisă de Tabelul din Anexa 7-3, ceea ce a făcut imposibilă decizia unică asupra clasificării lor în grupul de ordin 5 ce conține grupurile PP, PT, TD şi DD.

10. Analiza obiectivului evoluției folosind un eşantion întâmplător de generații

Pentru această statistică fiind variabilă continuă şi producând variații doar la evoluție s-a înlocuit ordonata *evoluție* (generație în care se produce o îmbunătățire a scorului funcției obiectiv) cu ordonata *generație*, cu ajutorul căreia salturile de evoluție se produc în medie mult mai lent, atenuând astfel variațiile observate, ceea ce permite observarea mult mai bună a funcției de distribuție a variabilei de interes (coeficientul de determinare). Procedura de transformare a datelor de la evoluții la generații este simplă, în generațiile intermediare evoluției statistica de interes având aceeaşi valoare cu cea pe care a avut-o în ultima sa evoluție.

Obținerea eşantionului din distribuția uniformă discretă

O primă ipoteză care trebuie verificată în acest context este dacă generatorul de numere aleatoare ce provin din distribuția uniformă discretă urmează într-adevăr această distribuție, fiind cunoscute problemele care pot apărea din generarea pseudoaleatoare [1]. În acest sens simulări Monte-Carlo recente (Steele & alții, 2005) au arătat că cele mai potrivite teste statistice sunt Pearson-Fisher χ^2 şi Anderson-Darling AD, fiecare dintre acestea obținând performanțe superioare celeilalte pentru diferite cazuri de legi de distribuție, dar amândouă fiind superioare celorlalte teste statistice. Din acest motiv verificarea ipotezei statistice folosind ambele teste este necesară. Tabelul 10-1 sumarizează această procedură.

Tabelul 10-1

Ipoteza de distribuție uniformă a numerelor aleatoare generate cu Rand(·,·) în PHP

Hypothesis of uniform distribution of the random numbers generated with Rand(·,·) in PHP

9 valori generate aleatoriu între 0 şi 20000	9221; 4182; 14283; 15329; 8875; 4599; 994; 8620; 7404
Valorile ordonate crescător	994; 4182; 4599; 7404; 8620; 8875; 9221; 14283; 15329
Probabilități cumulate observate - pco	1/9; 2/9; 3/9; 4/9; 5/9; 6/9; 7/9; 8/9; 9/9
Probabilități cumulate aşteptate (teoretice) - pca	.11; .22; .33; .44; .56; .67; .78; .89; 1.0
Testul Anderson-Darling	1AD = 0.9687; $cAD_{teoretic}$ = 2.5024; c/k = 2.58 > 1 (5%)
Testul Pearson χ^2	La 10 clase de frecvență X^2 = 8.5; df = 7; χ^2 (8.5, 7) = 29% > 5%
Testul Kolmogorov - Smirnov	D = 0.317 (0.778-0.461); $D\sqrt{9} = 0.95 = K^{-1}(9, 13.31\%)$

Concluzia simultană a testelor Anderson-Darling, Pearson χ^2 şi KS: există o asociere probabilă care depăşeşte 5% între cele două distribuții (cea teoretică şi cea observată) şi nu poate fi respinsă ipoteza că cele 18 valori provin din distribuția uniformă discretă 0..20000.

[1] Mads HAAHR. 1998-2009. ©. http://www.random.org/randomness/

Folosind programul DataPlot [2] s-a verificat dacă şirul de numere generate ar putea proveni din altă lege de distribuţie; din nou această procedură impune utilizarea testului Anderson-Darling. Rezultatele sunt redate în Tabelul 10-2.

Tabelul 10-2

Din ce lege de distribuţie provin numerele generate cu Rand(·,·) în PHP

From which distribution law comes the numbers gererated with Rand(·,·) in PHP

Lege de distribuţie *Distribution law*	kAD	AD*	cAD				$r_{10\%}$	Concluzie asupra eşantionului (nivel de semnificaţie)
			90%	95%	97.5%	99%		
Normală	.261	.297	.578	.683	.779	.926	1.9	Provine din distribuţia normală (10%)
Log-normală	.577	.655	.578	.683	.779	.926	0.9	Provine din distribuţia log-normală (5%)
Exponenţială	.867	.925	1.02	1.26	1.51	1.89	1.1	Provine din distribuţia exponenţială (10%)
Valori extreme	.275	.293	.637	.757	.877	1.04	2.2	Provine din distribuţia valori extreme (10%)
Logistică	.273	.280	.563	.660	.769	.906	2.0	Provine din distribuţia logistică (10%)
Dublu exponenţială	.362	.362	.714	.869	1.02	1.23	2.0	Provine din distribuţia dublu exponenţială (10%)
Uniformă	.833	.833	1.93	2.49	3.07	3.86	2.3	Provine din distribuţia uniformă (10%)
Pareto generalizată	.302	.302	.641	.771	.905	1.09	2.1	Provine din distribuţia Pareto generalizată (10%)

AD* - statistica kAD ajustată corectează numărul de grade de libertate; cAD: valori critice (critical values) scăzând din volumul eşantionului numărul de parametrii ce caracterizează legea de distribuţie

$r_{10\%}$ - valoarea raţiei de semnificaţie între valoarea de prag de 10% şi valoarea statisticii

Aşa cum se observă în Tabelul 10-2, acceptarea ipotezei că datele ar putea proveni dintr-o lege de distribuţie (se respinge ipoteza că cele două legi de distribuţie sunt diferite cu un risc de a fi în eroare de 10%, 5%, 2.5% sau 1%) nu exclude posibilitatea ca aceleaşi date să provină din altă lege de distribuţie. În acelaşi timp, raportul între statistica ajustată şi valoarea sa la pragul de 10% arată că cea mai probabilă lege de distribuţie este distribuţia uniformă (cu 2.3) urmată de distribuţia valorilor extreme (cu 2.2) şi Pareto generalizată (cu 2.1).

Căutarea legii de distribuţie a coeficientului de determinare al unei generaţii

A doua ipoteză care trebuie verificată este că valorile coeficientului de determinare obţinut în (cele 46 de) execuţii independente ale algoritmului genetic pentru fiecare pereche de strategie de selecţie şi supravieţuire provin dintr-o populaţie distribuită după o lege de distribuţie (având la dispoziţie o serie de alternative). Pentru aceasta fiecare din cele 9 generaţii extrase din distribuţia uniformă a generaţiilor 0..20000 au constituit subiectul investigaţiei şi rezultatele sunt redate în Tabelele 10-3-DD..10-3-TT (DE - dublu exponenţială; LG - logistică; LN - log-normală; NO - normală; UN - uniformă (0,1); EX - exponenţială; GU - valori extreme de tip I) pentru fiecare asociere de strategie de selecţie şi supravieţuire incluse în studiu.

[2] DataPlot v.2/2005. National Institute of Standards and Technology.

Tabelul 10-3-DD

Distribuția a coeficientului de determinare într-o generație pentru selecția deterministă însoțită de supraviețuirea deterministă (DD)

The distribution for the determination coefficient in a generation for deterministic selection and deterministic survival (DD)

DD	994	4182	4599	7404	8620	8875	9221	14283	15329	90	95	99	Σsig
DE	1.05	2.47	1.62	1.68	1.97	1.78	1.67	1.38	1.11	0.79	0.97	1.39	0;0;3
LG	0.94	1.74	1.19	1.13	1.16	1.14	1.10	0.94	0.88	0.56	0.66	0.91	0;0;1
LN	1.29	1.87	1.25	1.10	1.18	1.13	1.09	0.92	0.88	0.62	0.74	1.02	0;0;2
NO	1.30	1.87	1.26	1.10	1.17	1.13	1.09	0.92	0.88	0.62	0.74	1.02	0;0;2
UN	60.3	61.2	61.3	61.7	61.8	61.8	61.8	62.2	62.3	1.93	2.49	3.86	0;0;0
EX	21.3	21.3	21.3	21.2	21.2	21.2	21.2	21.2	21.2	1.06	1.32	1.95	0;0;0
GU	0.78	1.53	0.93	1.22	1.42	1.44	1.34	1.26	1.28	0.64	0.76	1.04	0;0;2
Σsig: număr de eşantioane care acceptă ipoteza distribuției la risc de eroare de 10, 5 şi 1%													

Tabelul 10-3-DP

Distribuția a coeficientului de determinare într-o generație pentru DP

The distribution for the determination coefficient in a generation for DP

DP	994	4182	4599	7404	8620	8875	9221	14283	15329	90	95	99	Σsig
DE	1.93	2.39	1.47	0.84	0.72	0.71	0.70	1.04	1.38	0.79	0.97	1.39	3; 4; 6
LG	1.03	1.69	1.68	0.87	0.67	0.68	0.69	0.74	0.68	0.56	0.66	0.91	0; 0; 6
LN	1.39	2.57	2.76	1.55	1.24	1.24	1.27	1.16	1.03	0.62	0.74	1.02	0; 0; 0
NO	1.40	2.59	2.78	1.57	1.26	1.27	1.29	1.17	1.04	0.62	0.74	1.02	0; 0; 0
UN	59.8	60.3	60.4	60.6	60.7	60.7	60.7	61.0	61.0	1.93	2.49	3.86	0; 0; 0
EX	21.3	21.2	21.2	21.2	21.2	21.2	21.2	21.2	21.2	1.06	1.32	1.95	0; 0; 0
GU	0.57	0.98	0.93	0.45	0.42	0.47	0.45	0.46	0.51	0.64	0.76	1.04	7; 7; 9
Σsig: număr de eşantioane care acceptă ipoteza distribuției la risc de eroare de 10, 5 şi 1%													

Tabelul 10-3-DT

Distribuția a coeficientului de determinare într-o generație pentru DT

The distribution for the determination coefficient in a generation for DT

DT	994	4182	4599	7404	8620	8875	9221	14283	15329	90	95	99	Σsig
DE	1.22	1.17	1.66	1.05	1.39	1.45	2.01	1.12	1.15	0.79	0.97	1.39	0; 0; 5
LG	1.01	0.87	0.75	0.90	0.76	0.73	0.97	0.85	0.92	0.56	0.66	0.91	0; 0; 6
LN	1.17	0.99	0.81	1.05	0.94	0.96	1.19	1.08	1.19	0.62	0.74	1.02	0; 0; 4
NO	1.18	1.00	0.82	1.06	0.95	0.97	1.20	1.09	1.20	0.62	0.74	1.02	0; 0; 4
UN	60.0	60.8	60.8	61.0	61.1	61.1	61.2	61.7	61.7	1.93	2.49	3.86	0; 0; 0
EX	21.3	21.3	21.3	21.3	21.2	21.2	21.3	21.3	21.3	1.06	1.32	1.95	0; 0; 0
GU	0.81	0.66	0.65	0.71	0.58	0.58	0.77	0.89	0.91	0.64	0.76	1.04	2; 5; 9
Σsig: număr de eşantioane care acceptă ipoteza distribuției la risc de eroare de 10, 5 şi 1%													

Tabelul 10-3-PD

Distribuția a coeficientului de determinare într-o generație pentru PD

The distribution for the determination coefficient in a generation for PD

PD	994	4182	4599	7404	8620	8875	9221	14283	15329	90	95	99	Σsig
DE	0.95	0.70	1.23	0.55	0.74	0.81	0.81	0.77	0.74	0.79	0.97	1.39	5; 8; 9
LG	0.78	0.36	0.39	0.50	0.42	0.52	0.52	0.68	0.68	0.56	0.66	0.91	6; 6; 9
LN	0.98	0.49	0.56	0.68	0.59	0.71	0.71	0.86	0.88	0.62	0.74	1.02	3; 6; 9
NO	0.99	0.50	0.57	0.69	0.60	0.72	0.72	0.88	0.89	0.62	0.74	1.02	3; 6; 9
UN	61.3	62.5	62.5	62.8	62.9	62.9	62.9	63.0	63.1	1.93	2.49	3.86	0; 0; 0
EX	21.2	21.2	21.2	21.2	21.2	21.2	21.2	21.2	21.2	1.06	1.32	1.95	0; 0; 0
GU	0.62	0.60	0.70	0.63	0.73	0.69	0.69	0.48	0.42	0.64	0.76	1.04	5; 9; 9
Σsig: număr de eşantioane care acceptă ipoteza distribuției la risc de eroare de 10, 5 şi 1%													

Tabelul 10-3-PP

Distribuția a coeficientului de determinare într-o generație pentru PP

The distribution for the determination coefficient in a generation for PP

PP	994	4182	4599	7404	8620	8875	9221	14283	15329	90	95	99	Σsig
DE	0.99	1.06	1.16	0.83	1.37	0.84	0.63	0.62	0.59	0.79	0.97	1.39	3; 5; 9
LG	0.59	0.67	0.62	0.41	0.44	0.41	0.38	0.37	0.35	0.56	0.66	0.91	6; 8; 9
LN	0.55	0.73	0.69	0.39	0.43	0.38	0.34	0.36	0.33	0.62	0.74	1.02	7; 9; 9
NO	0.56	0.74	0.70	0.39	0.43	0.39	0.35	0.36	0.33	0.62	0.74	1.02	7; 8; 9
UN	61.1	62.1	62.1	62.3	62.5	62.5	62.6	62.9	62.9	1.93	2.49	3.86	0; 0; 0
EX	21.2	21.2	21.2	21.2	21.2	21.2	21.2	21.2	21.2	1.06	1.32	1.95	0; 0; 0
GU	0.41	0.52	0.40	0.47	0.41	0.48	0.46	0.74	0.69	0.64	0.76	1.04	7; 9; 9
Σsig: număr de eşantioane care acceptă ipoteza distribuției la risc de eroare de 10, 5 şi 1%													

Tabelul 10-3-PT

Distribuția a coeficientului de determinare într-o generație pentru PT

The distribution for the determination coefficient in a generation for PT

PT	994	4182	4599	7404	8620	8875	9221	14283	15329	90	95	99	Σsig
DE	1.11	0.80	0.70	0.98	0.94	0.94	1.05	1.19	0.98	0.79	0.97	1.39	1; 4; 9
LG	0.74	0.69	0.55	0.88	0.81	0.81	0.88	0.70	0.59	0.56	0.66	0.91	1; 2; 9
LN	0.67	0.86	0.70	1.05	0.98	0.98	1.03	0.95	0.81	0.62	0.74	1.02	0; 2; 7
NO	0.67	0.86	0.71	1.06	0.98	0.98	1.03	0.96	0.82	0.62	0.74	1.02	0; 2; 7
UN	61.0	61.9	61.9	62.4	62.4	62.4	62.4	62.7	62.7	1.93	2.49	3.86	0; 0; 0
EX	21.3	21.2	21.2	21.2	21.2	21.2	21.2	21.2	21.2	1.06	1.32	1.95	0; 0; 0
GU	0.90	1.24	1.12	1.33	1.39	1.39	1.43	0.86	0.73	0.64	0.76	1.04	0; 1; 3
Σsig: număr de eşantioane care acceptă ipoteza distribuției la risc de eroare de 10, 5 şi 1%													

Tabelul 10-3-TD

Distribuția a coeficientului de determinare într-o generație pentru TD

The distribution for the determination coefficient in a generation for TD

TD	994	4182	4599	7404	8620	8875	9221	14283	15329	90	95	99	Σsig
DE	0.72	0.94	0.95	0.81	1.07	1.07	1.12	1.23	1.30	0.79	0.97	1.39	1; 4; 9
LG	0.67	0.93	0.72	0.63	0.82	0.82	0.76	0.85	0.89	0.56	0.66	0.91	0; 1; 8
LN	1.06	1.30	0.96	0.65	0.86	0.86	0.78	0.85	0.90	0.62	0.74	1.02	0; 1; 7
NO	1.07	1.32	0.97	0.66	0.86	0.86	0.79	0.86	0.91	0.62	0.74	1.02	0; 1; 7
UN	61.8	62.4	62.6	63.0	63.1	63.1	63.1	63.5	63.5	1.93	2.49	3.86	0; 0; 0
EX	21.3	21.2	21.2	21.2	21.2	21.2	21.2	21.2	21.2	1.06	1.32	1.95	0; 0; 0
GU	0.75	0.63	0.56	0.46	0.61	0.61	0.63	0.72	0.70	0.64	0.76	1.04	6; 9; 9
Σsig: număr de eşantioane care acceptă ipoteza distribuției la risc de eroare de 10, 5 şi 1%													

Tabelul 10-3-TP

Distribuția a coeficientului de determinare într-o generație pentru TP

The distribution for the determination coefficient in a generation for TP

TP	994	4182	4599	7404	8620	8875	9221	14283	15329	90	95	99	Σsig
DE	1.43	0.73	1.21	0.96	0.90	0.94	0.99	0.59	0.59	0.79	0.97	1.39	3; 6; 8
LG	0.59	0.56	0.49	0.73	0.57	0.55	0.47	0.54	0.54	0.56	0.66	0.91	5; 8; 9
LN	0.56	0.80	0.67	0.87	0.68	0.68	0.59	0.76	0.76	0.62	0.74	1.02	2; 5; 9
NO	0.57	0.82	0.68	0.89	0.69	0.69	0.60	0.77	0.77	0.62	0.74	1.02	2; 5; 9
UN	61.1	62.1	62.1	62.4	62.4	62.4	62.5	62.7	62.7	1.93	2.49	3.86	0; 0; 0
EX	21.2	21.2	21.2	21.2	21.2	21.2	21.2	21.2	21.2	1.06	1.32	1.95	0; 0; 0
GU	0.61	0.36	0.36	0.43	0.32	0.29	0.28	0.45	0.45	0.64	0.76	1.04	9; 9; 9
Σsig: număr de eşantioane care acceptă ipoteza distribuției la risc de eroare de 10, 5 şi 1%													

Tabelul 10-3-TT

Distribuția a coeficientului de determinare într-o generație pentru TT

The distribution for the determination coefficient in a generation for TT

TT	994	4182	4599	7404	8620	8875	9221	14283	15329	90	95	99	Σsig
DE	0.74	0.89	0.91	0.95	1.42	1.42	0.94	0.83	0.88	0.79	0.97	1.39	1; 7; 7
LG	0.46	0.58	0.62	0.44	0.46	0.46	0.50	0.51	0.56	0.56	0.66	0.91	6; 9; 9
LN	0.46	0.52	0.57	0.48	0.54	0.54	0.52	0.53	0.56	0.62	0.74	1.02	9; 9; 9
NO	0.47	0.52	0.57	0.49	0.55	0.55	0.52	0.54	0.57	0.62	0.74	1.02	9; 9; 9
UN	61.0	61.9	62.0	62.1	62.2	62.2	62.2	62.4	62.4	1.93	2.49	3.86	0; 0; 0
EX	21.3	21.2	21.2	21.2	21.2	21.2	21.2	21.2	21.2	1.06	1.32	1.95	0; 0; 0
GU	0.49	0.78	0.87	0.35	0.28	0.28	0.36	0.44	0.44	0.64	0.76	1.04	7; 7; 9
Σsig: număr de eşantioane care acceptă ipoteza distribuției la risc de eroare de 10, 5 şi 1%													

Tabelele 10-3-DD..10-3-TT arată genul de rezultat care nu ar fi fost de dorit de obținut; practic aproape fiecare tip de selecție şi încrucişare determină propria lege de distribuție a determinării. Tabelul 10-3-SS cumulează aceste observații.

Tabelul 10-3-SS

Distribuția a coeficientului de determinare într-o generație pentru diferite strategii de selecție şi supraviețuire

Distribution of the determination coefficient in a generation for different selection and survival strategies

SS	%	DE	LG	LN	NO	UN	EX	GU	Maxim de apariții *Max. of occurences*	Distribuție selectată *Selected distribution*
PP	10	3	6	7	7	0	0	7	(7): EV; NO; LN	Valori extreme (GU) sau Log-normală (LN)
	5	5	8	9	8	0	0	9	(9): EV; LN	
	1	9	9	9	9	0	0	9	(9): EV; NO; LN; LG; DE	
PT	10	1	1	0	0	0	0	0	(1): DE; LG	Dublu exponențială (DE)
	5	4	2	2	2	0	0	1	(4): DE	
	1	9	9	7	7	0	0	3	(9): DE; LG	
PD	10	5	6	3	3	0	0	5	(6): LG	Valori extreme (GU)
	5	8	6	6	6	0	0	9	(9): EV	
	1	9	9	9	9	0	0	9	(9): DE; LG; LN; NO; EV	
TP	10	3	5	2	2	0	0	9	(9): EV	Valori extreme (GU)
	5	6	8	5	5	0	0	9	(9): EV	
	1	8	9	9	9	0	0	9	(9): LG; LN; NO; EV	
TT	10	1	6	9	9	0	0	7	(9): LN; LG	Log-normală (LN) sau Logistică (LG)
	5	7	9	9	9	0	0	7	(9): LG; LN; NO	
	1	7	9	9	9	0	0	9	(9): LG; LN; NO; EV	
TD	10	1	0	0	0	0	0	6	(6): EV	Valori extreme (GU)
	5	4	1	1	1	0	0	9	(9): EV	
	1	9	8	7	7	0	0	9	(9): DE; EV	
DP	10	3	0	0	0	0	0	7	(7): EV	Valori extreme (GU)
	5	4	0	0	0	0	0	7	(7): EV	
	1	6	6	0	0	0	0	9	(9): EV	
DT	10	0	0	0	0	0	0	2	(2): EV	Valori extreme (GU)
	5	0	0	0	0	0	0	5	(5): EV	
	1	5	6	4	4	0	0	9	(9): EV	
DD	10	0	0	0	0	0	0	0	(0): -	Dublu exponențială (DE)
	5	0	0	0	0	0	0	0	(0): -	
	1	3	1	2	2	0	0	2	(3): DE	

Tabelul 10-4 sintetizează rezultatele din Tabelul 10-3-SS. Se observă că pentru selecția deterministă însoțită de supraviețuirea deterministă nici o lege nu a întrunit acceptarea.

Tabelul 10-4

Legi de distribuție ale coeficientului de determinare într-o generație pentru diferite strategii de selecție şi supraviețuire

Distribution laws for determination coefficient in a generation for different selections and survivals

Lege de distribuție	PP	PT	PD	TP	TT	TD	DP	DT	DD
Valori extreme de tip I (GU)	X		X	X		X	X	X	
Log-normală (LN)	X				X				
Dublu exponențială (DE)		X							X*
Logistică					X				

Din Tabelul 10-4 se observă că legea valorilor extreme de tip I (GU) este legea de distribuție cea mai frecventă a coeficientului de determinare într-o generație a algoritmului genetic rulat cu diferite strategii de selecție şi supraviețuire.

Distribuția valorilor extreme de tip I (sau distribuția Gumbel) este legea limită de distribuție a minimumului unui număr mare variabile întâmplătoare distribuite identic. Dacă variabilele întâmplătoare sunt mărginite (cum este cazul coeficientului de determinare, mărginit de 0 şi 1) distribuția valorilor extreme devine (atinge cazul particular al) distribuției Weibul. Dacă numărul de variabile întâmplătoare distribuite identic devine unitar, atunci distribuția valorilor extreme devine (atinge cazul particular al) distribuției normale Gauss.

Există o explicație pentru care distribuția valorilor extreme de tip I este acceptată de marea majoritate a cazurilor şi această explicație provine din designul experimental: cele 46 de execuții sunt chiar independente, ceea ce înseamnă mai exact că fiecare dintre acestea îşi creează propriul eşantion inițial (cel al generației 0) în mod întâmplător, şi evoluția ulterioară este în mod evident influențată de compoziția generației inițiale, deci cel puțin pentru o bună bucată de timp începând cu momentul inițial este de aşteptat ca populația coeficientului de determinare să nu poată fi aproximată ca provenind din aceeaşi variabilă întâmplătoare. Mai mult, pe parcursul evoluției au loc procese de mutație şi încrucişare, selecție şi supraviețuire care au la bază alegeri mai mult (mutație şi încrucişare) sau mai puțin sau deloc întâmplătoare (selecție şi supraviețuire).

Această secțiune a avut ca scop obținerea legii de distribuție a observabilei coeficient de determinare într-un moment al evoluției (o anumită generație).

Rezultatele sintetizate în Tabelul 10-4 sugerează drept lege de distribuție pentru coeficientul de determinare legea de distribuție Gumbel, şi respinge fără drept de apel ipoteza de distribuție normală (Gauss) a acestuia (cu nici măcar o singură intrare în Tabelul 10-4).

Media nu este o statistică suficientă

O problemă statistică importantă şi-a găsit răspuns în secțiunea anterioară, şi anume:

÷ *Valoarea medie a coeficientului de determinare pentru fiecare generație (ca valoare medie din cele 46 de execuții independente) este o* ***statistică suficientă*** *pentru acest număr? Varianța sau*

abaterea standard este o statistică suficientă?

Răspunsul la această întrebare este NU, motivul fiind faptul că aproape pentru nici o pereche de strategii selecție - supraviețuire nu s-a putut accepta ipoteza distribuției normale a valorilor, care ar fi acceptat valoarea medie din eșantionul de 46 de valori drept estimator nedeplasat al mediei populației tuturor execuțiilor independente iar varianța drept statistică suficientă.

Analiza valorilor coeficientului de determinare într-un număr de 9 generații întâmplătoare a arătat (capitolul anterior) că cel mai frecvent legea de distribuție acceptată pentru valorile coeficientului de determinare ale unei generații relativ la cele 46 de execuții independente este legea valorilor extreme de tip I (dată în Tabela 16 din Anexa 3). Această concluzie ne permite să obținem o serie de estimatori ai evoluției coeficientului de determinare (de la o generație la alta), în acord cu legea de distribuție urmată în cadrul generației. Întrucât nici această lege nu a fost acceptată în totalitate de toate strategiile de selecție și supraviețuire, și astfel nici unul dintre parametrii legii valorilor extreme nu reprezintă la rândul său o statistică suficientă, ci doar cea mai bună statistică. Pentru fiecare pereche de strategie de selecție și strategie de supraviețuire s-au estimat parametrii populației valorilor coeficientului de determinare al fiecărei generații (de la 0 la 20000) pentru legea de distribuție a valorilor extreme de tip I. Statisticile obținute au constituit subiectul investigației în vederea obținerii ecuației de evoluție în timp. Acestea pot fi clasificate ca în următorul tabel (Tabelul 10-5).

Tabelul 10-5

Statistici ale evoluției către obiectiv (*Statistics of evolution to the objective*)

Clasă *Class*	Statistică *Statistic*	Variabilă și estimatorul acesteia din ecuația de timp *Variable and their estimator from time dependence*
Tendință centrală	Valoarea medie	MEAN; EMEAN
	Locație	din momente centrale MELOC*; EMELOC* din maximizarea ratei șansei LELOC*; ELELOC*
Cazuri limită	Valoarea minimă	MINIMUM; EMINIMUM*
	Valoarea maximă	MAXIMUM; EMAXIMUM*
Împrăștiere	Deviația standard	STD; ESTD*
	Scală	din momente centrale MESCA*; EMESCA*
		din maximizarea ratei șansei LESCA*; ELESCA*
		cu corecție de abatere BCESCA*; EBCESCA*
Erori standard	SE Locație	din momente centrale SEMELOC*; ESEMELOC*
		din maximizarea ratei șansei SELELOC*; ESELELOC*
	SE Scală	din momente centrale SEMESCA*; ESEMESCA*
		din maximizarea ratei șansei SELESCA*; ESELESCA*
*Estimatori obținuți în ipoteza de distribuție Gumbel (valori extreme de tip I) folosind programul DataPlot (freeware, NIST, USA)		

O primă analiză a rezultatelor obținute a referit urmărirea evoluției tendinței centrale în ipoteza de distribuție Gumbel având ca scop obținerea de modele de evoluție care să caracterizeze evoluția în strategia dată de asocierea strategiei de selecție cu metoda de supraviețuire.

Distribuția Gumbel nu este suficient de generală

Au fost făcute estimările parametrilor din Tabelul 10-5 pentru fiecare generație (0...20000) folosind aplicația DataPlot şi au fost salvate în fişiere distincte pentru fiecare asociere de selecție şi supraviețuire. Rezultatele au fost apoi interpolate folosind aplicația SlideWrite.

A fost definită o funcție putere

$$\hat{Y}(X) = \frac{\frac{X^{\gamma} - 1}{\gamma} - \alpha}{\beta}$$

şi parametrii de interpolare (α, β, şi γ), intervalele de încredere şi semnificația statistică t sociată au fost obținute pentru fiecare caz în parte în ceea ce privește tendința centrală.

A fost definită o funcție rațională:

$$\hat{Y}(X) = 0.907\left(1 - \frac{X^2 + a_3X + a_2}{X^2 + a_1X + a_0}\right)$$

şi parametrii de interpolare (α, β, şi γ), intervalele de încredere şi semnificația statistică t sociată au fost obținute pentru fiecare caz în parte în ceea ce privește valorile extreme (minimul şi maximul). Această din ultimă analiză, a valorilor extreme a arătat că distribuția Gubmel nu poate fi acceptată drept lege de distribuție, probabilitățile de apariție ale valorilor extreme calculate fiind inaceptabil de departe de valorile aceloraşi probabilități observate.

Analiza cuprinzând valoarea medie observată ca estimator al tendinței centrale, valorile minime observate şi valorile maxime observate este redată în Anexa 8. În Anexa 8 nu a fost cuprinsă analiza parametrilor estimați folosind distribuția Gumbel, şi anume cei menționați cu asterix în Tabelul 10-5 dintr-un motiv foarte simplu: concluzia studiului cuprins în Anexa 8 a fost respingerea ipotezei de distribuție după legea valorilor extreme de tip I.

11. Analiza legii de distribuție a obiectivului evoluției folosind un studiu sistematic în cadru generalizat

Legea de distribuție Fisher-Tippett a obiectivului evoluției

Întrucât analiza din Tabelul 10-4 relevă că totuşi pentru majoritatea cazurilor observate legea de distribuție Gumbel este totuşi acceptată, rezultă consecința că legea de distribuție urmată de observații este o lege de distribuție mai generală decât legea de distribuție Gumbel dar care o are pe aceasta (Gumbel) drept caz limită. Într-adevăr există această lege de distribuție, numită Legea Generală a Valorilor Extreme, sau legea Fisher-Tippett (Fisher & Tippett, 1928), lege care are drept caz limită legea Gumbel (a valorilor extreme de tip I) ca frontieră de separație între alte două familii de legi de distribuție, legea Weibull şi legea Fréchet. Expresiile distribuției de probabilitate (PDF) şi funcției cumulative de probabilitate (CDF) ale legii Fisher-Tippett (FT) sunt:

$$FT_{PDF}(X)=\begin{cases}\frac{1}{\beta}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)(1+k\frac{x-\lambda}{\beta})^{-1-1/k}, & k<0 \quad \text{Weibull}\\ \frac{1}{\beta}\exp\left(-\frac{x-\lambda}{\beta}-\exp(-\frac{x-\lambda}{\beta})\right), & k=0 \quad \text{Gumbel}\\ \frac{1}{\beta}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)(1+k\frac{x-\lambda}{\beta})^{-1-1/k}, & k>0 \quad \text{Fréchet}\end{cases}$$

$$FT_{CDF}(X)=\left\{\begin{array}{ll}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right), & k<0 \quad \text{Weibull}\\ \exp\left(-\exp(-\frac{x-\lambda}{\beta})\right), & k=0 \quad \text{Gumbel}\\ \exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right), & k>0 \quad \text{Fréchet}\end{array}\right\}\text{Fisher}-\text{Tippett}$$

O primă problemă care se cere rezolvată este verificarea ipotezei de distribuție după legea Fisher-Tippett a observațiilor experimentale. Pentru a realiza această problemă s-a folosit aplicația EasyFit [3]. Distribuția Fisher-Tippett a fost o alternativă dintr-un număr de peste 55 de legi de distribuție în total, număr care conține atât legi de distribuție continue cât şi discrete, distribuții mărginite şi respectiv nemărginite. O serie numeroasă dintre aceste distribuții nu s-au calificat din start sau după primele testări ca distribuții posibile. Setul a fost restrâns la distribuții mărginite şi distribuții generalizate, set din care în studiu au fost cuprinse Beta, Johnson, Kumaraswamy, Pert, Putere, Reciprocă, Triangulară, Uniformă (distribuții mărginite) şi Fisher-Tippett, Pareto, Log-Pearson tip III (distribuții generalizate). Trei statistici: Pearson-Fisher Chi-Square, Anderson-Darling şi Kolmogorov-Smirnov au evaluat agrementul între observație şi model. Analiza este redată în Anexa 9.

[3] EasyFit Professional v.50. 2008. Software. MathWave Technologies.

O a doua problemă care se cere rezolvată este identificarea parametrilor distribuției FT şi anume forma (k), locația (λ) şi scala (β) din cele 46 de observații independente ale valorii coeficientului de determinare pentru fiecare din cele 20000 de generații de evoluție supuse observației. Pentru a realiza această problemă s-a folosit din nou aplicația EasyFitXL [3].

Rezultatele comparației distribuției observate cu distribuțiile teoretice listate în Tabelul 1 - Anexa 9 suportă analize multicriteriale, dintre care după generație şi apoi după legea de distribuție, clasificare menită să pună în evidență influența generației asupra abaterii observate de la legea de distribuție teoretică, şi după metoda de selecție şi de supraviețuire şi apoi după legea de distribuție să se pună în evidență dacă asocierea de selecție şi supraviețuire generează o lege proprie de distribuție.

Însă cel mai important criteriu este respingerea statistică la unul sau mai multe nivele de semnificație, şi această analiză dată în Tabelul 1 - Anexa 9 este sintetizată în Tabelul 11-1.

Tabelul 11-1

Respingeri semnificative statistic în comparația distribuțiilor observate şi distribuțiile teoretice Jonson SB, Pert şi Fisher-Tippett în cele 81 de eşantioane

Rejections statistically significant in comparing observed distributions with Jonson SB, Pert and Fisher-Tippett in the 81 samples

Respingere *Rejection*	Distribuție *Distribution*	Număr *Number*	Generații *Generations*
La risc de a fi în eroare de 1% şi toate statisticile utilizate (K-S, A-D şi C-S)	Pert	17	PT(6): 994, 7404, 8620, 8875, 9221, 15329 TP(4): 4182, 4599, 14283, 15329 DD(2): 994, 15329 DT(2): 4599, 7404 PD(2): 4599, 15329 TT(1): 994
	Johnson SB	11	PT(6): 4182, 4599, 7404, 8620, 8875, 9221 PD(3): 4182, 4599, 8620 TD(1): 994 TP(1): 994
	Fisher-Tippett	0	-
La risc de a fi în eroare de 1% şi două din trei statistici	Johnson SB	9	DD(6): 4182, 4599, 7404, 8620, 8875, 9221 DP(3): 4182, 4599, 7404
	Pert	1	DT(1): 994
	Fisher-Tippett	0	-
La risc de a fi în eroare de 1% şi o statistică din trei	Pert	7	K-S(1): 9221(DT); A-D(2): 14283, 15329 (DT) C-S(4): 4182(PT), 4599(PT,DP), 8620(DT)
	Fisher-Tippett	4	C-S(4): TD(2): 4182, 4599; DT(2): 994, 8620
	Johnson SB	1	C-S(1): 7404(TT)
Probabilitățile de a observa statisticile C-S pentru distribuția Fisher-Tippett în cele 4 cazuri în care s-a coborât sub riscul de eroare de 1% sunt: DT, 994: C-S = 16.667; df = 3; p = 0.83‰; TD, 4182: C-S = 14.275; df = 4; p = 6.47‰; DT, 8620: C-S = 14.236; df = 4; p = 6.58‰; TD, 4599, C-S = 13.917; df = 4; p = 7.56‰;			

Rezultatele din Tabelul 11-1 arată că distribuția Fisher-Tippett este singura care nu este respinsă semnificativ statistic (cu riscul de a fi în eroare de 1%) de cel puțin două din cele trei teste

statistice utilizate, în timp ce celelalte două distribuții înregistrează un număr semnificativ de respingeri din totalul de 81 de eşantioane: Jonson SB - 20 (~25%), Pert - 18 (~22%). Analizând rezultatul obținut la respingerea unei statistici din trei, se observă în primul rând că probabilitatea are în 3 din 4 cazuri acelaşi ordin de magnitudine cu nivelul de semnificație ales (10‰) şi în al doilea rând că respingerea se face pentru statistica C-S (Chi Square). Relativ la această din urmă observație, o remarcă se cere făcută: acuratețea testului Chi-Square este scăzută pentru volume de eşantioane mici sau medii (cum e cazul celor 46 de execuții independente cuprinse în observabilă) datorită faptului că împărțirea în clase de frecvență pe care o impune aplicarea testului Chi Square este (mai mult sau mai puțin) arbitrară.

O altă posibilă explicație pentru cele patru observații respinse de testul Chi Square în compararea distribuției teoretice Fisher-Tippett cu distribuția observată în cele 81 de cazuri este legată de estimarea parametrilor distribuției teoretice Fisher-Tippett, care aşa cum sugerează şi relațiile de definire, nu este deloc simplă şi/sau cu nivel foarte ridicat de certitudine. Mai mult, s-a făcut presupunerea că variabila timp (sau generație) nu interferează (nu condiționează) cu legea de distribuție observată, ceea ce se va demonstra în continuare că nu este întocmai adevărat (având influență asupra valorilor parametrilor distribuției şi în acest mod asupra sub-tipului distribuției) şi acesta influențând deci decisiv scorul observat la testul Chi Square.

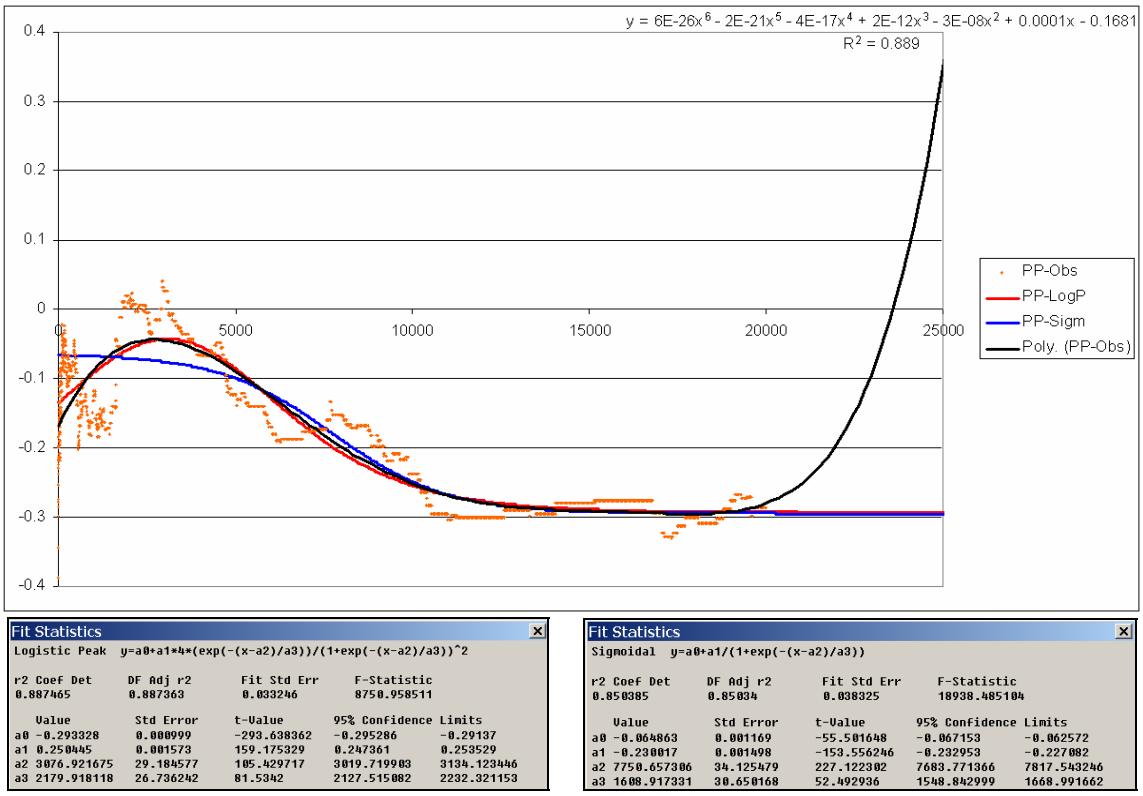
Evoluția nu este polinomială

Analiza expusă în continuare are ca scop identificarea expresiilor în funcție de timp ale parametrilor statistici ai legii de distribuție Fisher-Tippett.

Ideea de a obține expresiile de timp ale parametrilor statistici ale legilor de distribuție nu este nouă. Astfel studiul literaturii de specialitate arată preocupări susținute de cercetare în acest sens, cu aplicații în special în domeniul meteorologic (Kallache & alții, 2005).

Ceea ce însă diferă semnificativ între abordarea propusă în (Kallache & alții, 2005) şi abordarea propusă aici este legat de ipotezele de lucru.

Astfel, în (Kallache & alții, 2005) se propune utilizarea de serii de timp polinomiale pentru a modela expresia parametrilor distribuției Fisher-Tippett, ceea ce poate asigura o bună estimare (practic se poate defini un polinom de orice ordin astfel încât suma reziduurilor să fie minimizată în măsură suficientă) dar predicția bazată pe expresii polinomiale este dacă nu un eşec total cu siguranță nu este de folos decât în imediata vecinătate a intervalului de timp supus observației. Edificator în acest sens este un exemplu bazat pe parametrul formă (k) al distribuției teoretice Fisher-Tippett pentru coeficientul de determinare al evoluției bazate pe selecție şi supraviețuire proporționale (PP) - exemplu redat în Figura 11-1.



Fit Statistics

Logistic Peak y=a0+a1*4*(exp(-(x-a2)/a3))/(1+exp(-(x-a2)/a3))^2

r2 Coef Det	DF Adj r2	Fit Std Err	F-Statistic
0.887465	0.887363	0.033246	8750.958511

	Value	Std Error	t-Value	95% Confidence Limits	
a0	-0.293328	0.000999	-293.638362	-0.295286	-0.29137
a1	0.250445	0.001573	159.175329	0.247361	0.253529
a2	3076.921675	29.184577	105.429717	3019.719903	3134.123446
a3	2179.918118	26.736242	81.5342	2127.515082	2232.321153

Fit Statistics

Sigmoidal y=a0+a1/(1+exp(-(x-a2)/a3))

r2 Coef Det	DF Adj r2	Fit Std Err	F-Statistic
0.850385	0.85034	0.038325	18938.485104

	Value	Std Error	t-Value	95% Confidence Limits	
a0	-0.064863	0.001169	-55.501648	-0.067153	-0.062572
a1	-0.230017	0.001498	-153.556246	-0.232953	-0.227082
a2	7750.657306	34.125479	227.122302	7683.771366	7817.543246
a3	1608.917331	30.650168	52.492936	1548.842999	1668.991662

Fig. 11-1. Observație vs. Model în 3 cazuri de tendință: Logistic, Sigmoidal şi Polinomial
Observation vs. Model in three cases of tendency: Logistic, Sigmoidal and Polynomial

Ceea ce se observă în Figura 11-1 este că deşi modelul polinomial are cea mai mare determinare statistică (88.9%) în raport cu regresia după un model logistic (88.7%) şi sigmoidal (85%), practic modelul polinomial este inutilizabil pentru estimarea observațiilor dincolo de intervalul de observație (0..20000), pentru simplul motiv că nu conține nimic din modelul real al fenomenului observat.

Evoluția locației, scalei şi formei distribuției Fisher-Tippett a obiectivului evoluției

S-a folosit macro-ul pentru Excel al aplicației EasyFit (EasyFitXL) pentru calcularea parametrilor legilor de distribuție Fisher-Tippett pentru fiecare generație şi fiecare asociere de selecție şi supraviețuire. Estimările parametrilor locație (λ), scală (β) şi formă (k) au fost obținute în fiecare caz în parte folosind principiul maximizării ratei şansei dintr-un volum de 46 de observații, şi anume cele 46 de execuții independente ale algoritmului genetic. S-a folosit aplicația Statistica pentru uniformizări exponențiale.

Rezultatul pentru parametrul formei (k) este redat în Figura 11-2. Invarianții formelor din Figura 11-2 s-au obținut prin uniformizare exponențială (tehnică aplicată în mod curent seriilor de timp). O curbă este trasată la coordonatele Y (parametrul formei, k) în seria de date în acord cu o procedură de uniformizare exponențială simplă. Fiecare punct este calculat ca medie ponderată a

tuturor observațiilor precedente (lanț Markov) când pondere mai mare este acordată observațiilor mai recente, când ponderea descreşte geometric cu vârsta observației anterioare. Algoritmic acest fapt este satisfăcut de relația matematică:

$$S(t) = \alpha \cdot Y(t) + (1-\alpha) \cdot S(t-1)$$

în care S(t) este valoarea seriei transformate la momentul t, S(t-1) este valoarea seriei transformate la momentul (t-1), Y(t) este valoarea seriei inițiale la momentul t, iar α este constanta de uniformizare exponențială.

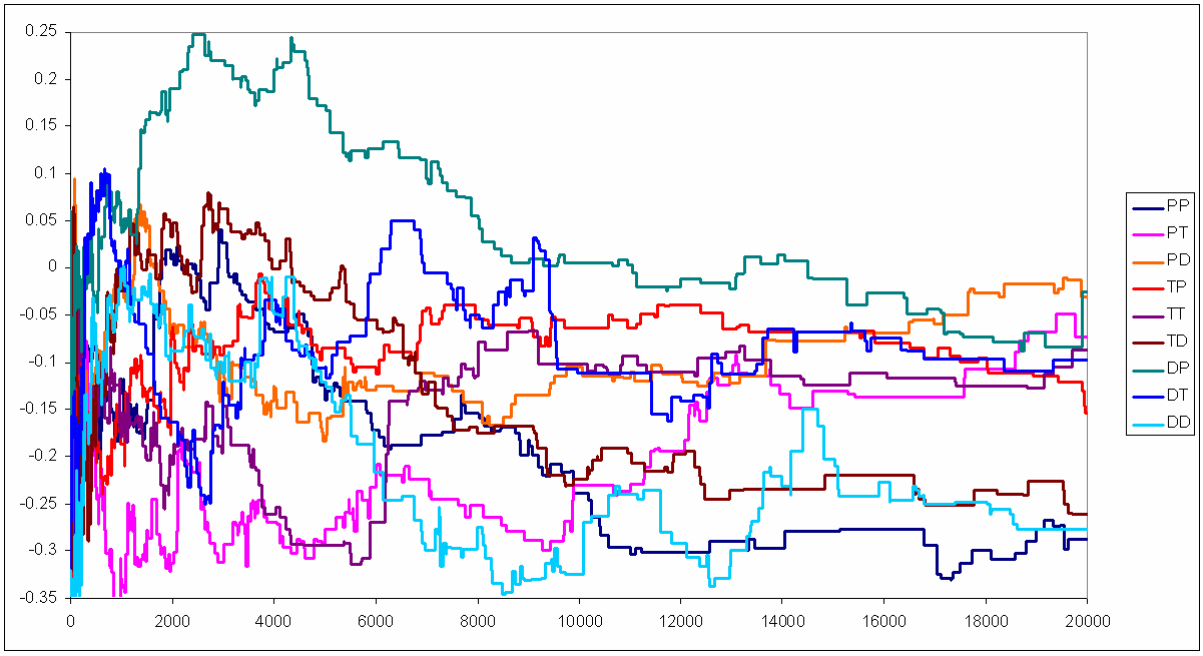


Fig. 11-2. Parametrul formă (k) al distribuției Fisher-Tippett: estimare din observații
Shape parameter (k) of Fisher-Tippett distribution: estimation from oservations

Prin simpla înlocuire a valorilor în ecuația de uniformizare de mai sus se obține că:

$$S(t) = \alpha \cdot Y(t) + (1-\alpha) \cdot S(t-1) = \alpha \cdot Y(t) + (1-\alpha) \cdot (\alpha \cdot Y(t-1) + (1-\alpha) \cdot S(t-2)) = \ldots$$

$$= \alpha \cdot Y(t) + \alpha(1-\alpha) \cdot Y(t-1) + \alpha(1-\alpha)^2 \cdot Y(t-2) + \ldots + \alpha(1-\alpha)^{t-1} \cdot Y(1) + (1-\alpha)^t \cdot Y(0)$$

şi astfel ponderile în expresia lui S(t) acordate observațiilor Y(t), Y(t-1), ... evoluează în progresie geometrică: $\{(1-\alpha)^0, (1-\alpha)^1, (1-\alpha)^2, \ldots\}$.

Dacă α este apropiat de zero (0) atunci mai multă importanță este acordată observațiilor anterioare unui moment de timp t, rezultând o curbă mai uniformă (Figura 11-3); dacă α este mai aproape de unu (1), mai multă importanță este acordată observației de la momentul t rezultând o curbă mai puțin uniformă dar mult mai apropiată de datele observate (Figura 11-3).

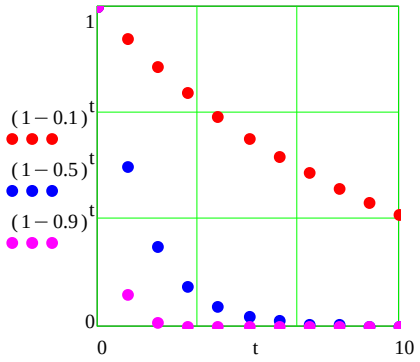


Fig. 11-3. Primele 10 ponderi în uniformizarea exponențială la α = 0.1, 0.5 şi 0.9
First ten weights in exponential smoothing at α = 0.1, 0.5 and 0.9

Pentru serii de date suficient de mari, transformarea lui Y(t) la S(t) lasă neschimbată relația între timp şi observabilă. Demonstrația acestui fapt este dificil de făcut, însă este uşor de exemplificat.

Astfel, dacă se transformă forma pentru PP se obțin relațiile:

$$S_{PP}(G) = 0.1 \cdot k_{PP}(G) + (1-0.1) \cdot S_{PP}(G-1);\ S_{PP}(0) = -0.206$$

Corelațiile care se stabilesc sunt următoarele:

$$r(G,k_{PP}) = -0.85844... = r(G,S_{PP});\ r(\ln(G),k_{PP}) = -0.72577... = r(\ln(G),S_{PP})$$

adică transformata S_{PP} a variabilei k_{PP} păstrează relația (liniară, logaritmică, etc.) care există între k_{PP} şi G.

Pentru o mai bună uniformizare s-a ales şi s-a fixat valoarea parametrului de uniformizare α la 0.1 urmând ca valoarea S(0) - care este denumită forma seriei (în cazul de față forma formei) să se obțină prin minimizarea erorilor. Tabelul 11-2 conține analiza de reziduuri aplicată transformatelor S ale formelor k din legile de distribuție Fisher-Tippett pentru seria de metode de selecție şi supraviețuire supuse analizei, Figura 11-4 conține formele uniformizate, în timp ce Figura 11-5 reprezintă ecuațiile de tendință.

Tabelul 11-2

Transformarea formelor legilor de distribuție Fisher-Tippett şi ecuația de tendință
Transformation of the shapes of the Fisher-Tippett distributions and tendency equation

SS	S0	MAE	SSE	MSE	$S(G) = a_0 + a_1 \cdot G$	r	F	t_0	t_1
PP	-0.206	1.3E-3	0.45	2.3E-5	$a_0 = -0.1912;\ a_1 = -1.47 \cdot 10^{-6}$	0.858	56017	-2661	-237
PT	-0.200	1.6E-3	0.64	3.2E-5	$a_0 = -0.2108;\ a_1 = 1.08 \cdot 10^{-6}$	0.809	38004	-3287	195
PD	-0.093	1.5E-3	1.38	6.9E-5	$a_0 = -0.0961;\ a_1 = 3.12 \cdot 10^{-7}$	0.374	3245	-1519	57
TP	-0.082	1.4E-3	1.54	7.7E-5	$a_0 = -0.0833;\ a_1 = 1.24 \cdot 10^{-7}$	0.173	619	-1444	25
TT	-0.142	1.4E-3	0.97	4.8E-5	$a_0 = -0.1476;\ a_1 = 5.58 \cdot 10^{-7}$	0.510	7044	-1924	84
TD	-0.150	1.6E-3	0.93	4.6E-5	$a_0 = -0.1352;\ a_1 = -1.47 \cdot 10^{-6}$	0.831	44775	-1684	-212
DP	-0.041	1.2E-3	1.19	6.0E-5	$a_0 = -0.0193;\ a_1 = -1.32 \cdot 10^{-6}$	0.800	35649	-238	-189
DT	-0.081	1.6E-3	0.96	4.8E-5	$a_0 = -0.0797;\ a_1 = -1.35 \cdot 10^{-7}$	0.132	354	-961	-19
DD	-0.216	1.8E-3	1.01	5.1E-5	$a_0 = -0.0207;\ a_1 = -9.52 \cdot 10^{-7}$	0.592	10779	-1949	-104

Figura 11-2 pune în evidență forme similare (dar semnificativ diferite, Tabelul 11-2) ale distribuției Fisher-Tippett pentru diferitele metode de selecție şi supraviețuire. O altă observație

importantă este cu privire la valoarea parametrului de formă care este în toate cazurile negativ (Tabelul 11-2), ceea ce particularizează distribuția Fisher-Tippett la distribuție Weibull. Valorile negative mici ale parametrului formei, variind între -0.25 şi 0 cu uşoare tendințe (Figura 11-5) crescătoare (pentru PT, PD, TP şi TT) sau descrescătoare (pentru PP, TD, DP, DT şi DD) explică de ce în testarea inițială a ipotezei distribuției Gumbel în marea majoritate a cazurilor s-a putut accepta ipoteza distribuției Gumbel dar nu s-a putut accepta pentru toate cazurile (Tabelul 10-4) şi în acelaşi timp de ce în calcularea probabilităților de apariție a valorilor extreme (Figura 11 - Anexa 8) s-a eşuat în verificarea ipotezei de eşantionare din distribuția Gumbel (Tabelul 11 - Anexa 8).

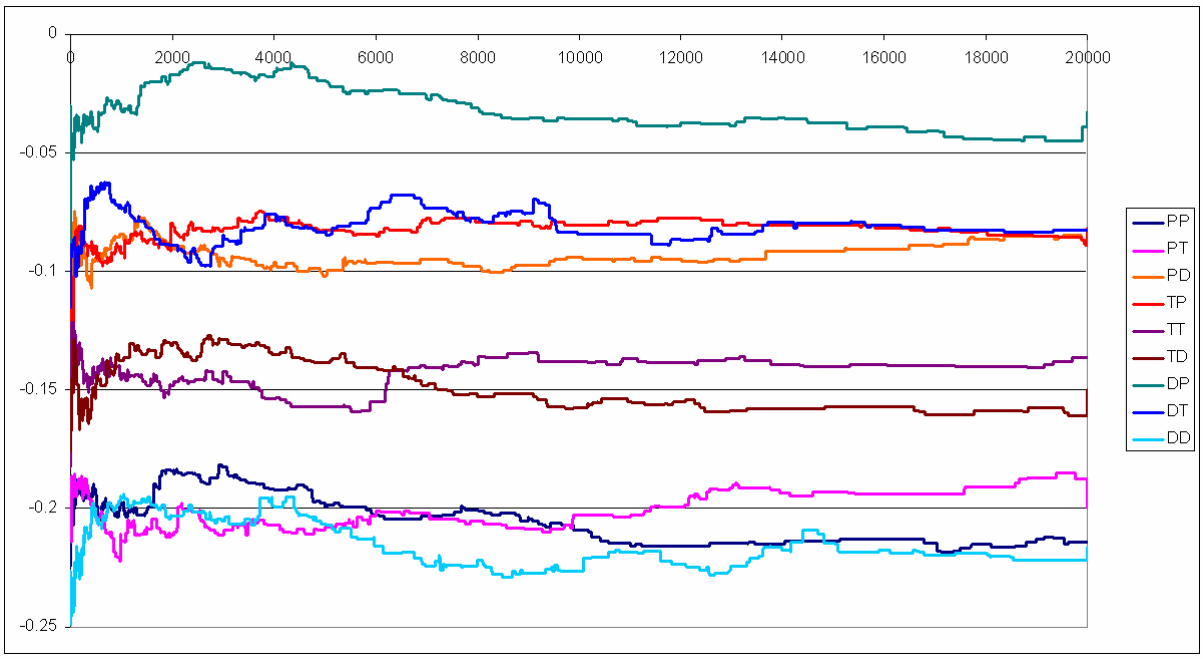


Fig. 11-4. Parametrul formă (k) al distribuției Fisher-Tippett: uniformizat
Shape parameter (k) of Fisher-Tippett distribution: uniformed

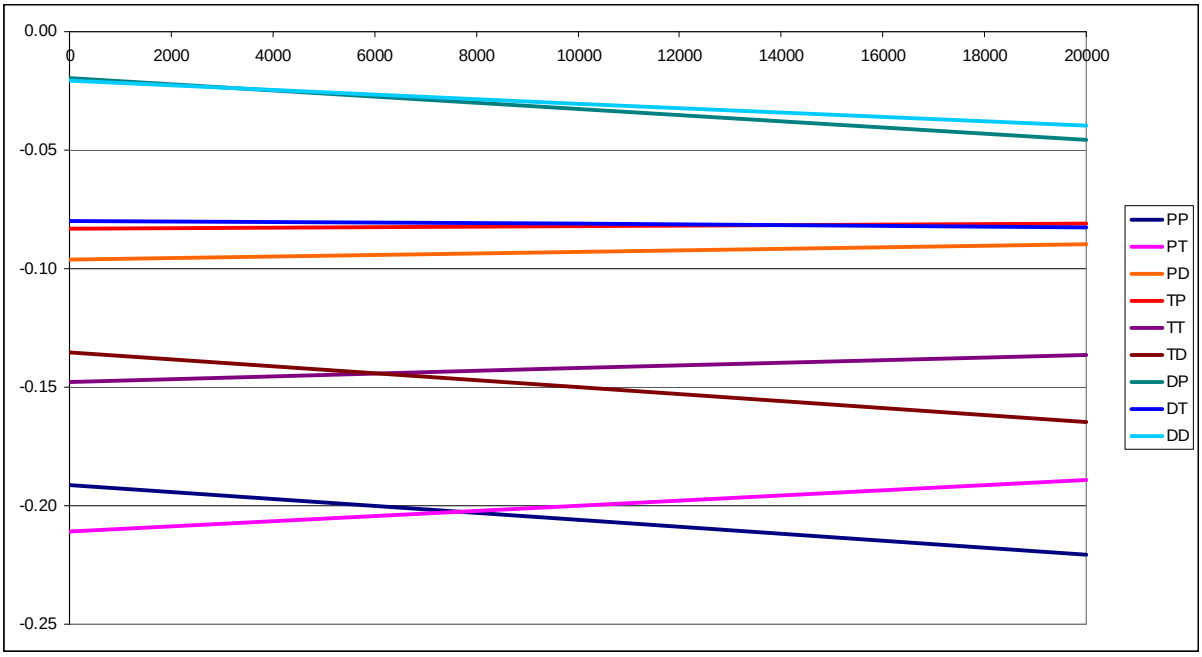


Fig. 11-5. Parametrul formă (k) al distribuției Fisher-Tippett: tendințe
Shape parameter (k) of Fisher-Tippett distribution: tendences

Rezultatul pentru parametrul scalei (β) este redat în Figura 11-6. Pentru o mai bună uniformizare s-a ales şi s-a fixat valoarea parametrului de uniformizare α la 0.1 urmând ca valoarea S(0) - care este denumită forma seriei (în cazul de față forma scalei) să se obțină prin minimizarea erorilor. Tabelul 11-3 conține analiza de reziduuri aplicată transformatelor S ale formelor k din legile de distribuție Fisher-Tippett pentru seria de metode de selecție şi supraviețuire supuse analizei, Figura 11-7 conține scalele uniformizate în timp ce Figura 11-8 reprezintă ecuațiile de tendință.

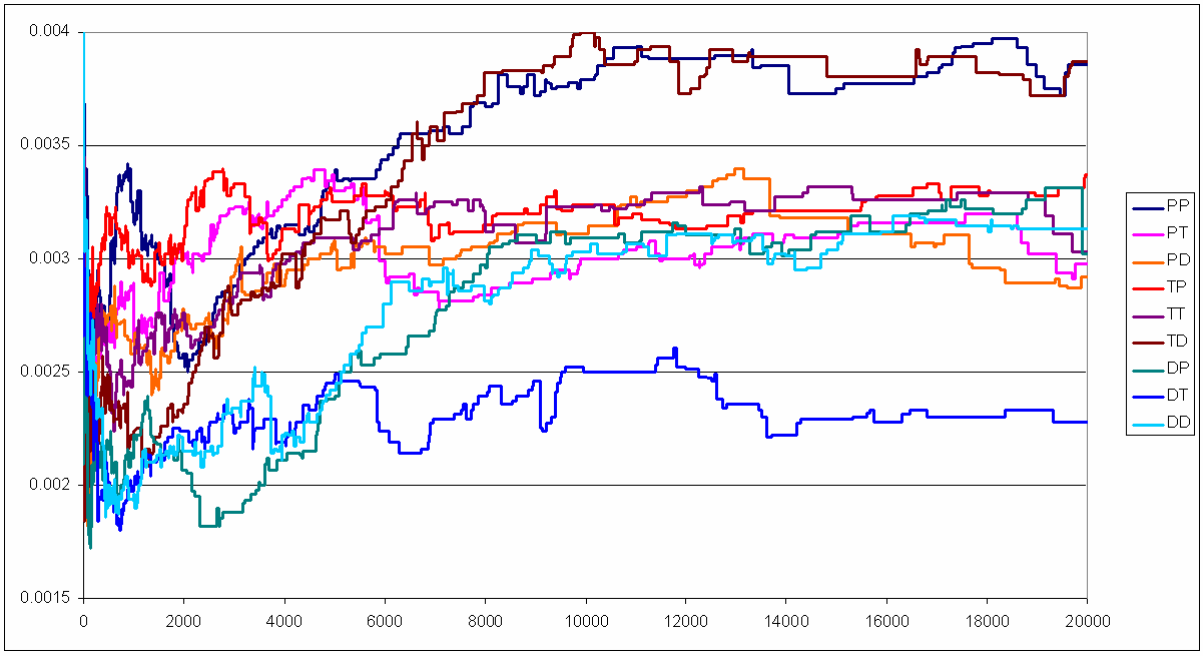


Fig. 11-6. Parametrul scală (β) al distribuției Fisher-Tippett: estimare din observații
Scale parameter (β) of Fisher-Tippett distribution: estimation from oservations

Tabelul 11-3

Transformarea scalelor legilor de distribuție Fisher-Tippett şi ecuația de tendință
Transformation of the scales of Fisher-Tippett distributions and the tendency law

SS	S0	MAE	SSE	MSE	$S(G) = a_0 + a_1 \cdot G$		t_0	t_1	r	F
PP	0.0036	4.6E-6	8.6E-6	4.3E-10	3.541E-3	5.5E-9	11599	210	0.829	43915
PT	0.0030	4.3E-6	1.1E-5	5.4E-10	2.996E-3	8.2E-10	13116	42	0.283	1739
PD	0.0030	4.2E-6	9.7E-6	4.9E-10	2.983E-3	1.9E-9	11534	84	0.513	7134
TP	0.0032	3.9E-6	5.9E-6	3.0E-10	3.192E-3	8.9E-10	25597	82	0.503	6786
TT	0.0031	4.1E-6	1.3E-5	6.4E-10	3.072E-3	2.9E-9	14164	157	0.743	24578
TD	0.0035	5.1E-6	1.2E-5	6.2E-10	3.419E-3	7.9E-9	7700	205	0.823	41884
DP	0.0028	4.0E-6	9.0E-6	4.5E-10	2.730E-3	7.1E-9	9043	271	0.887	73442
DT	0.0023	3.8E-6	6.5E-6	3.3E-10	2.296E-3	6.1E-10	12484	38	0.263	1482
DD	0.0028	4.5E-6	1.3E-5	6.5E-10	2.745E-3	5.6E-9	10150	241	0.862	58091

O observație importantă este cu privire la tendința parametrului de scală care este în toate cazurile crescătoare, ceea ce arată tendința de creştere a variabilității pentru toate asocierile de metode supuse observației.

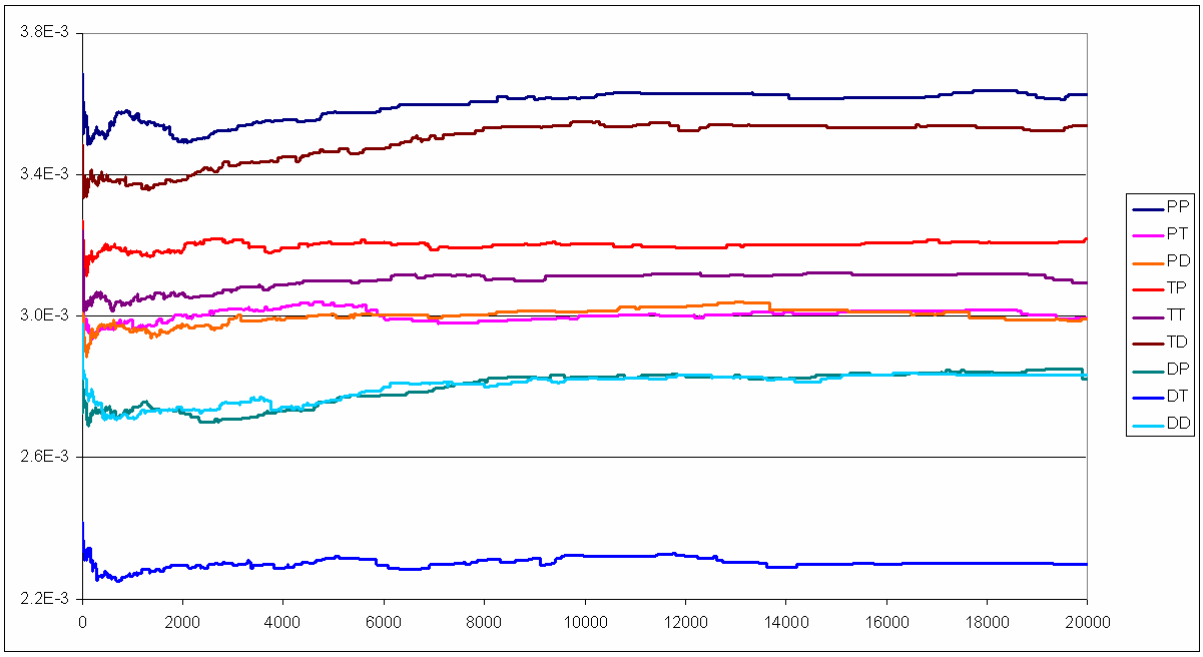


Fig. 11-7. Parametrul scală (β) al distribuției Fisher-Tippett: uniformizat
Scale parameter (β) of Fisher-Tippett distribution: uniformed

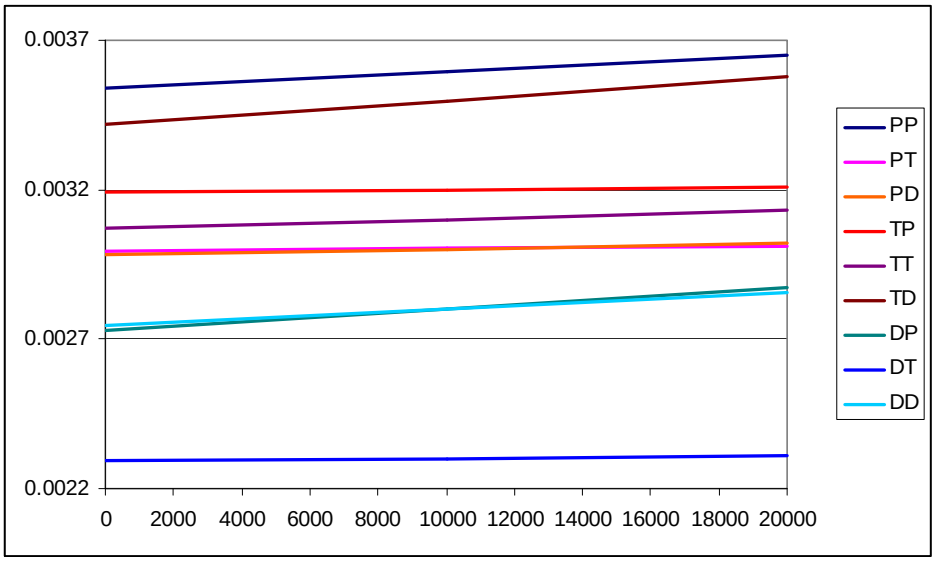


Fig. 11-8. Parametrul scală (β) al distribuției Fisher-Tippett: tendințe
Scale parameter (β) of Fisher-Tippett distribution: tendences

Figura 11-5 și Figura 11-8 pun în evidență grupuri distincte de forme și scale între asocierile de metode, grupuri observabile vizual în figuri. Astfel Tabelul 11-4 redă grupurile ce se formează după forma (k) și scala distribuției.

Tabelul 11-4

Grupuri de formă și scală în selecție și supraviețuire
Groups of shape and scale in selection and survival

Parametrul distribuției Fisher-Tippett *Parameter of the Fisher-Tippett distribution*	Grupuri de selecție și supraviețuire *Groups of selection and survival*
Forma distribuției (k) - *Shape*	DD, DP; PP, TP
Scala distribuției (β) - *Scale*	PT, PD; DD, DP

Rezultatul pentru estimarea parametrului locație (λ) este redat în Figura 11-9. Pentru o mai bună uniformizare s-a ales şi s-a fixat valoarea parametrului de uniformizare α la 0.1 urmând ca valoarea S(0) - care este denumită forma seriei (în cazul de față forma locației) să se obțină prin minimizarea erorilor.

Figura 11-10 conține locațiile uniformizate în funcție de generație.

Tabelul 11-5 conține analiza de reziduuri aplicată transformatelor S ale formelor k din legile de distribuție Fisher-Tippett pentru seria de metode de selecție şi supraviețuire supuse analizei.

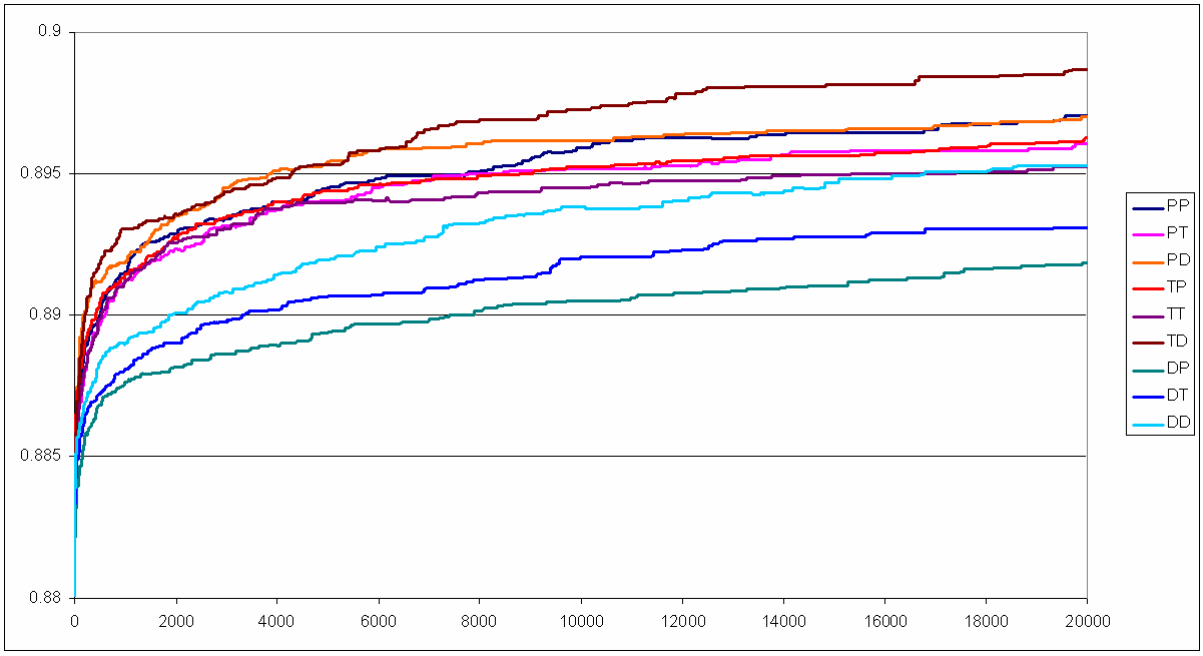


Fig. 11-9. Parametrul locație (λ) al distribuției Fisher-Tippett: estimare din observații
Location parameter (k) of Fisher-Tippett distribution: estimation from oservations

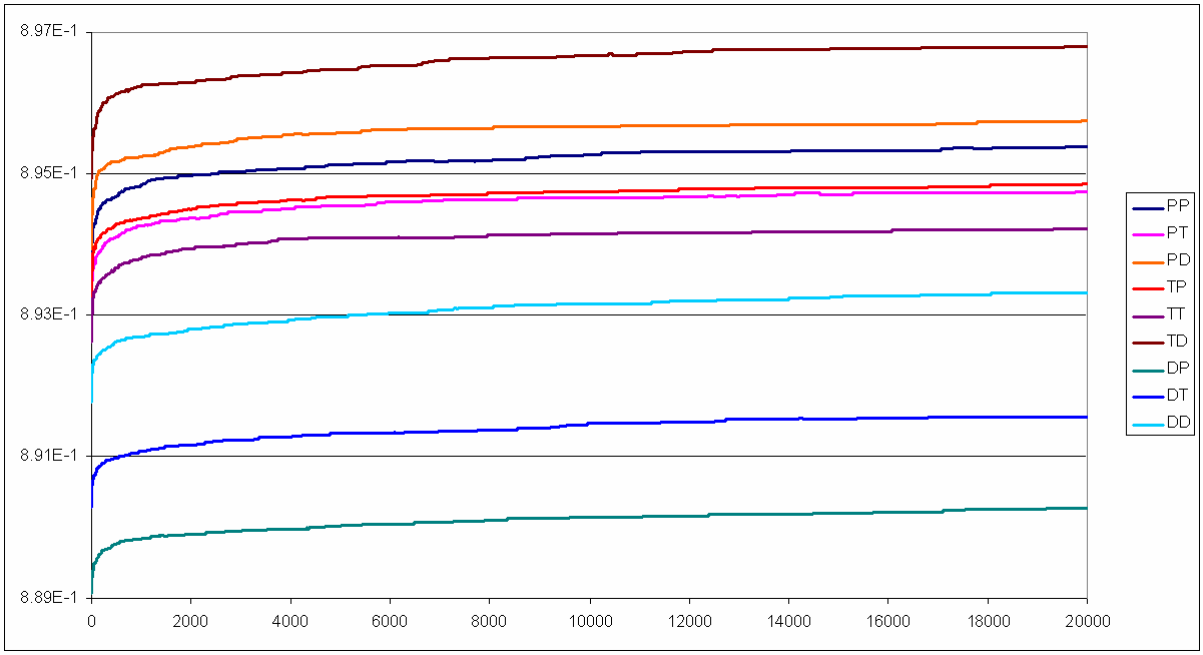


Fig. 11-10. Parametrul locație (λ) al distribuției Fisher-Tippett: uniformizat
Location parameter (k) of Fisher-Tippett distribution: uniformed

Tabelul 11-5

Transformarea locațiilor legilor de distribuție Fisher-Tippett

Transformations of the locations of Fiher-Tippett distributions

SS	S0	MAE	SSE	MSE
PP	0.8952	1.2E-5	9.4E-4	4.7E-8
PT	0.8946	1.1E-5	9.4E-4	4.7E-8
PD	0.8956	1.1E-5	1.0E-3	5.0E-8
TP	0.8947	1.1E-5	9.0E-4	4.5E-8
TT	0.8941	1.0E-5	8.0E-4	4.0E-8
TD	0.8966	1.3E-5	1.1E-3	5.5E-8
DP	0.8901	8.4E-6	4.2E-4	2.1E-8
DT	0.8914	8.5E-6	4.6E-4	2.3E-8
DD	0.8931	9.8E-6	6.2E-4	3.1E-8

Pentru a pune în evidență dependența locației (λ) din distribuția Fisher-Tippett de numărul de generații s-a efectuat o analiză de regresie folosind un eşantion de 16 modele de regresie frecvent observate (Figura 11-11, în care User-Defined a fost o regresie logaritmică în forma $y = a_0 + a_1 \cdot \ln(x + a_2)$). Dependența sugerată de analiza de regresie ilustrată în Figura 11-11 este fie o funcție putere fie o funcție logaritmică:

$$\lambda(G) = a_0 + a_1 \cdot G^{a_2} \text{ sau } \lambda(G) = a_0 + a_1 \cdot \ln(G + a_2)$$

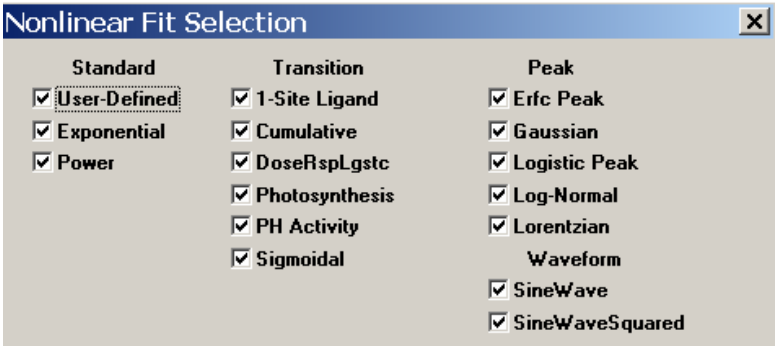


Fig. 11-11. Eşantion de 16 modele de regresie testate în obținerea dependenței locației

Sample of 16 regression models tested for obtaining the dependence of the location

Tabelul 11-6 prezintă analiza de regresie şi ecuațiile obținute în timp ce Figura 11-12 prezintă ecuațiile locației pentru un interval al generațiilor G = 1..50000.

Tabelul 11-6

Analiza de regresie pentru locațiile legilor de distribuție Fisher-Tippett

The regression analysis for locations of the Fisher-Tippett distributions

SS	Model	a_0	a_1	a_2	t_0	t_1	t_2	r	F
PP	$\lambda(G) = a_0 + a_1 \cdot \ln(G + a_2)$	0.89357	$1.82 \cdot 10^{-4}$	0.867	640993	1174	9	0.9966	737924
PD		0.89422	$1.55 \cdot 10^{-4}$	-0.344	344366	536	-12	0.9833	146308
TP		0.89333	$1.54 \cdot 10^{-4}$	-0.213	666193	1027	-8	0.9954	539368
TT		0.89286	$1.40 \cdot 10^{-4}$	-0.348	361929	507	-12	0.9814	130838
PT	$\lambda(G) = a_0 + a_1 \cdot \ln(G)$	0.89309	$1.69 \cdot 10^{-4}$	-	502833	853	-	0.9932	727191
TD	$\lambda(G) = a_0 + a_1 \cdot G^{a_2}$	0.89465	$6.84 \cdot 10^{-4}$	0.117	38625	39	76	0.9923	321942
DP		0.88916	$2.02 \cdot 10^{-4}$	0.171	132950	51	124	0.9947	467811
DT		0.89016	$3.19 \cdot 10^{-4}$	0.151	100793	56	125	0.9957	574382
DD		0.89173	$2.93 \cdot 10^{-4}$	0.172	134833	75	183	0.9975	1010738

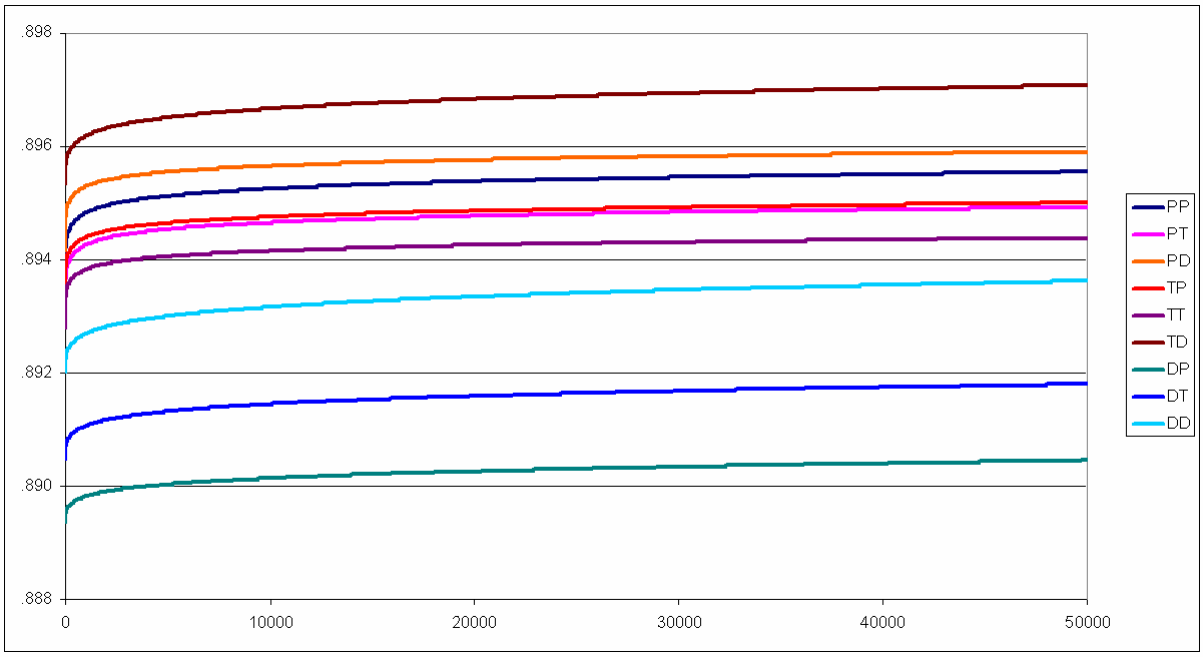


Fig. 11-12. Parametrul locație (λ) al distribuției Fisher-Tippett: tendințe
Location parameter (λ) of Fisher-Tippett distribution: tendences

Reprezentarea din Figura 11-12 pune în evidență capacitatea modelului obținut pentru a face predicții, în spre deosebire de contra-exemplul din Figura 11-1.

Obținerea modelului matematic al dependenței de timp a parametrului locație (λ) al distribuției Fisher-Tippett încheie modelarea matematică a coeficientului de determinare (funcția obiectiv a evoluției).

Se pot astfel exprima ecuațiile de distribuție Fisher-Tippett ale obiectivului evoluției în forma dependentă de timp (generație a evoluției). Ceea ce trebuie efectuat este înlocuirea expresiilor obținute pentru formă (k) - Tabelul 11-2, scală (β) - Tabelul 11-3 şi locație (λ) - Tabelul 11-6 în expresiile pentru densitatea de probabilitate (PDF) şi repartiția de probabilitate (CDF):

$$FT_{PDF}(X)=\frac{1}{\beta}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)(1+k\frac{x-\lambda}{\beta})^{-1-1/k},\ FT_{CDF}(X)=\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)$$

Tabelul 11-7 sumarizează rezultatele obținute pentru ecuațiile de timp ale determinării.

Tabelul 11-7

Ecuațiile formei (k), scalei (β) şi locației (λ) din distribuția Fisher-Tippett
The equations for shape (k), scale (β) and location (λ) of Fiher-Tippett distributions

SS	$k(G)=a_0+a_1\cdot G$		$\beta(G)=a_0+a_1\cdot G$		Tendință λ(G) *Tendency*	a_0	a_1	a_2
PP	-0.1912	$-1.47\cdot10^{-6}$	3.541E-3	5.5E-9	$\lambda(G)=a_0+a_1\cdot\ln(G+a_2)$	0.89357	$1.82\cdot10^{-4}$	0.867
PD	-0.0961	$3.12\cdot10^{-7}$	2.983E-3	1.9E-9		0.89422	$1.55\cdot10^{-4}$	-0.344
TP	-0.0833	$1.24\cdot10^{-7}$	3.192E-3	8.9E-10		0.89333	$1.54\cdot10^{-4}$	-0.213
TT	-0.1476	$5.58\cdot10^{-7}$	3.072E-3	2.9E-9		0.89286	$1.40\cdot10^{-4}$	-0.348
PT	-0.2108	$1.08\cdot10^{-6}$	2.996E-3	8.2E-10	$\lambda(G)=a_0+a_1\cdot\ln(G)$	0.89309	$1.69\cdot10^{-4}$	
TD	-0.1352	$-1.47\cdot10^{-6}$	3.419E-3	7.9E-9	$\lambda(G)=a_0+a_1\cdot G^{a_2}$	0.89465	$6.84\cdot10^{-4}$	0.117
DP	-0.0193	$-1.32\cdot10^{-6}$	2.730E-3	7.1E-9		0.88916	$2.02\cdot10^{-4}$	0.171
DT	-0.0797	$-1.35\cdot10^{-7}$	2.296E-3	6.1E-10		0.89016	$3.19\cdot10^{-4}$	0.151
DD	-0.0207	$-9.52\cdot10^{-7}$	2.745E-3	5.6E-9		0.89173	$2.93\cdot10^{-4}$	0.172

Expresiile care dau tendința pentru funcția repartiției de probabilitate Fisher-Tippett ($SSFT_{CDF}$, unde SS - Selecție și Supraviețuire) sunt următoarele:

$$PPFT_{CDF}(r^2,G)=\exp\left(-\left(1+(-0.1912-1.47\cdot10^{-6}\cdot G)\frac{r^2-0.89357-1.82\cdot10^{-4}\cdot\ln(G+0.867)}{3.541\cdot10^{-3}+5.5\cdot10^{-9}\cdot G}\right)^{1/(0.1912+1.47\cdot10^{-6}\cdot G)}\right)$$

$$PDFT_{CDF}(r^2,G)=\exp\left(-\left(1+(-0.0961+3.12\cdot10^{-7}\cdot G)\frac{r^2-0.89422-1.55\cdot10^{-4}\cdot\ln(G-0.344)}{2.983\cdot10^{-3}+1.9\cdot10^{-9}\cdot G}\right)^{1/(0.0961-3.12\cdot10^{-7}\cdot G)}\right)$$

$$TPFT_{CDF}(r^2,G)=\exp\left(-\left(1+(-0.0833+1.24\cdot10^{-7}\cdot G)\frac{r^2-0.89333-1.54\cdot10^{-4}\cdot\ln(G-0.213)}{3.192\cdot10^{-3}+8.9\cdot10^{-10}\cdot G}\right)^{1/(0.0833-1.24\cdot10^{-7}\cdot G)}\right)$$

$$TTFT_{CDF}(r^2,G)=\exp\left(-\left(1+(-0.1476+5.58\cdot10^{-7}\cdot G)\frac{r^2-0.89286-1.4\cdot10^{-4}\cdot\ln(G-0.384)}{3.072\cdot10^{-3}+2.9\cdot10^{-9}\cdot G}\right)^{1/(0.1476-5.58\cdot10^{-7}\cdot G)}\right)$$

$$PTFT_{CDF}(r^2,G)=\exp\left(-\left(1+(-0.2108+1.08\cdot10^{-6}\cdot G)\frac{r^2-0.89309-1.69\cdot10^{-4}\cdot\ln(G)}{2.996\cdot10^{-3}+8.2\cdot10^{-10}\cdot G}\right)^{1/(0.2108-1.08\cdot10^{-6}\cdot G)}\right)$$

$$TDFT_{CDF}(r^2,G)=\exp\left(-\left(1+(-0.1352-1.47\cdot10^{-6}\cdot G)\frac{r^2-0.89465-6.84\cdot10^{-4}\cdot G^{0.117}}{3.419\cdot10^{-3}+7.9\cdot10^{-9}\cdot G}\right)^{1/(0.1352+1.47\cdot10^{-6}\cdot G)}\right)$$

$$DPFT_{CDF}(r^2,G)=\exp\left(-\left(1+(-0.0193-1.32\cdot10^{-6}\cdot G)\frac{r^2-0.88916-2.02\cdot10^{-4}\cdot G^{0.171}}{2.73\cdot10^{-3}+7.1\cdot10^{-9}\cdot G}\right)^{1/(0.0193+1.32\cdot10^{-6}\cdot G)}\right)$$

$$DTFT_{CDF}(r^2,G)=\exp\left(-\left(1+(-0.0797-1.35\cdot10^{-7}\cdot G)\frac{r^2-0.89016-3.19\cdot10^{-4}\cdot G^{0.151}}{2.296\cdot10^{-3}+6.1\cdot10^{-10}\cdot G}\right)^{1/(0.0797+1.35\cdot10^{-7}\cdot G)}\right)$$

$$DDFT_{CDF}(r^2,G)=\exp\left(-\left(1+(-0.0207-9.52\cdot10^{-7}\cdot G)\frac{r^2-0.89173-2.93\cdot10^{-4}\cdot G^{0.172}}{2.745\cdot10^{-3}+5.6\cdot10^{-9}\cdot G}\right)^{1/(0.0207+9.52\cdot10^{-7}\cdot G)}\right)$$

Similar cu ecuațiile de mai sus se pot obține și expresiile care dau tendința funcției de densitate de probabilitate (PDF). Acestea sunt însă mult mai complicate pentru a fi redate sub formă de expresii matematice, însă mult mai sugestivă este reprezentarea tridimensională a acestora. Figura 1 - Anexa 10 redă reprezentările tridimensionale ale funcției de densitate de probabilitate în care în locul variabilei generație s-a folosit logaritmul în baza 10 al acesteia (scară logaritmică).

Folosind ecuațiile de tendință pentru formă (k), scală (β) și locație (λ) se pot obține frontierele de probabilitate 95% (în care șansa de a fi atinsă o valoare superioară acestei frontiere îi este rezervată la numai 5% dintre cazurile care ar putea fi observate - Figura 11-13) și respectiv de probabilitate 5% (în care șansa de a fi atinsă o valoare inferioară acestei frontiere îi este rezervată la numai 5% dintre cazurile care ar putea fi observate - Figura 11-14).

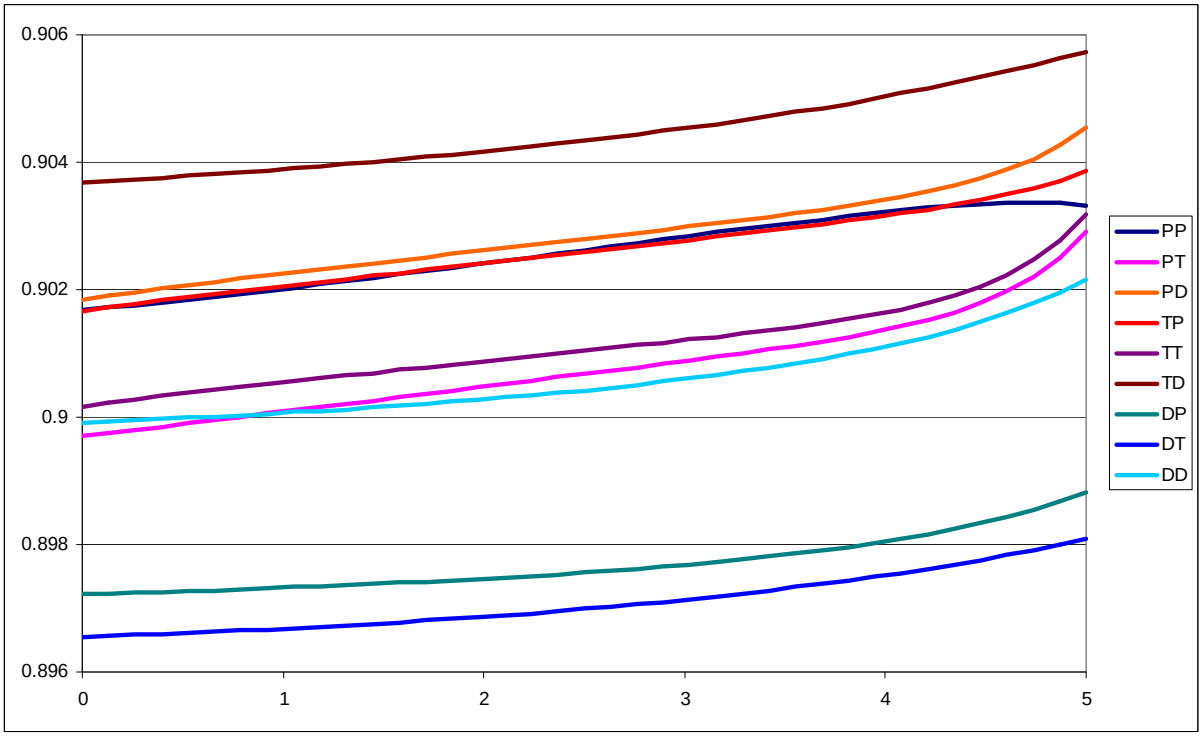


Fig. 11-13. Loteria norocoşilor (CDF = 95%) *Lucky lottery*

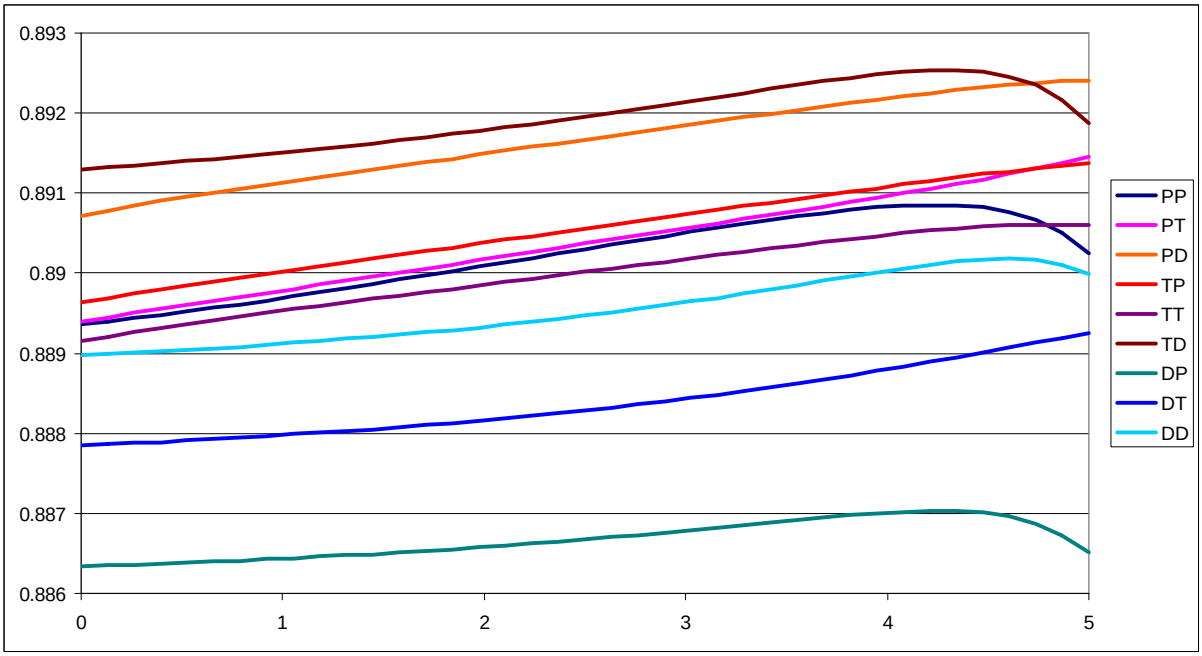


Fig. 11-14. Loteria ghinioniştilor (CDF = 5%) *Unlucky lottery*

Se remarcă în Figurile 11-13 şi 11-14 comportarea selecției deterministe însoțite de supraviețuirea în turnir (DT) care este singura asociere de selecție şi supraviețuire al cărui interval de normalitate (fără noroc şi fără ghinion) rămâne aproximativ acelaşi pe parcursul evoluției.

Tendința generală observată (Figurile 11-13 şi 11-14) este că norocul creşte ceva mai accentuat decât în tendință liniară odată cu evoluția - doar selecția proporțională însoțită de supraviețuire proporțională (PP) se abate de la această regulă în timp ca ghinionul creşte ceva mai puțin accentuat decât în tendință liniară odată cu evoluția (doar PT şi DT se abat de la această

regulă). Pentru a ilustra cum variază intervalul de normalitate s-a reprezentat diferența celor două frontiere (Figura 11-15).

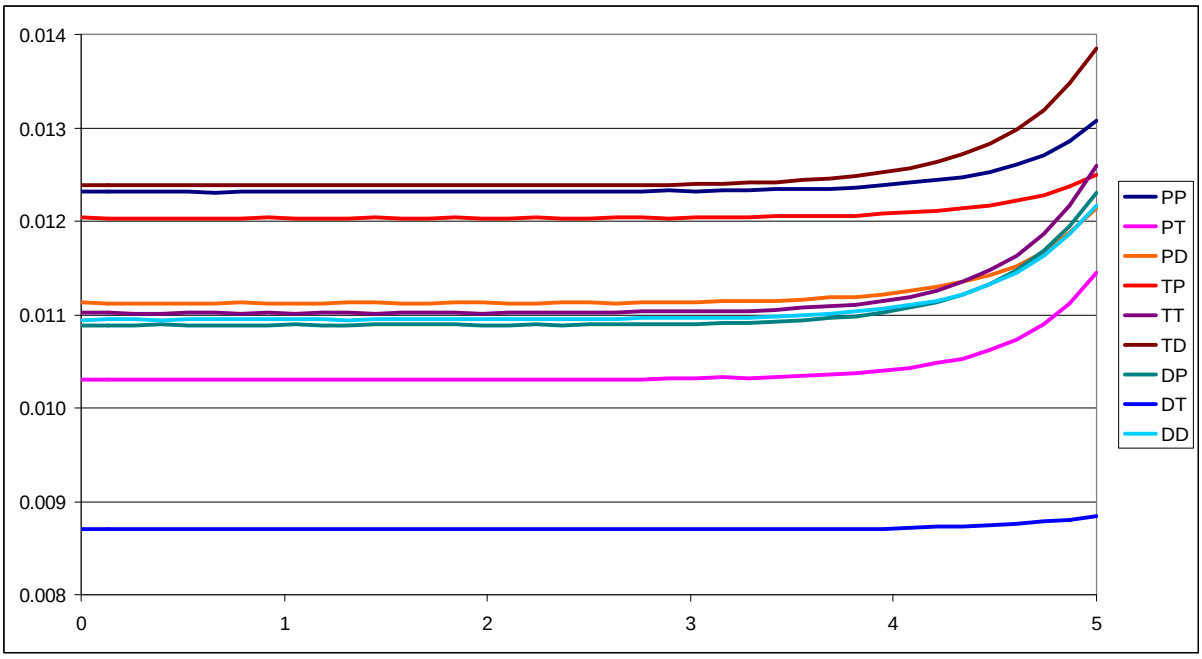


Fig. 11-15. Lărgimea intervalului de normalitate în selecție şi supraviețuire (fără noroc, fără ghinion)
The weight of the normality interval in selection and survival (no lucky, no unlucky)

Figura 11-15 evidențiază un grup compact evoluție similară în normalitate (TT, DP, DD şi PD) şi două disimilarități extreme: DT - aproape constantă în normalitate de-a lungul evoluției şi TD - cu cea mai mare creştere a lărgimii intervalului de normalitate pe parcursul evoluției.

Expresiile densităților de probabilitate Fisher-Tippett se pot obține şi din estimările directe ale parametrilor de formă (k), scală (β) şi locație (λ) din observații (Figura 2 în Anexa 10) şi respectiv din formele uniformizate ale acestora (Figurile 3-5 în Anexa 10) caz în care se observă că uniformizarea modifică semnificativ distribuția Fisher-Tippett doar în porțiunea sa inițială aproape indiferent de alegerea constantei de uniformizare.

Compararea reprezentărilor Figurile 3-5 în Anexa 10 din cu reprezentările din Figura 2 - Anexa 10 evidențiază ceea ce era de aşteptat, şi anume că tendința centrală ilustrată în Figura 1 - Anexa 10 poate fi cazul limită al repetării observației de un număr infinit de ori, însă ceea ce din observații putem extrage este o distribuție mai mult (Figura 2 în Anexa 10) sau mai puțin oscilantă (gradual mai puțin în Figurile 3, 4 şi 5 din Anexa 10) prin valorile parametrilor distribuției. Avantajul reprezentării din Figura 1 - Anexa 10 este că pune în evidență tendința centrală, dar dezavantajul este că nu poate fi folosită în estimări mai mult decât ale tendinței centrale, eşuând aproape de fiecare dată în estimarea unui anume caz în parte, deoarece atât forma cât şi ceilalți parametrii ai distribuției observate conțin o variabilitate intrinsecă ridicată, cel puțin în prima parte a evoluției (Figura 2 din Anexa 10).

Având obținute funcțiile de densitate de probabilitate (PDF) şi de repartiție de probabilitate

(CDF) ale distribuției Fisher-Tippett se poate obține (formulele de mai jos) şi apoi reprezenta grafic hazardul (HDF) - direct din observații în Figura 6 din Anexa 10 şi respectiv uniformizat profund marcat de istorie în Figura 7 din Anexa 10:

$$HDF(X) = \frac{PDF(X)}{1 - CDF(X)}$$

Ceea ce în special Figura 7 din Anexa 10 evidențiază (şi anume hazardul profund marcat de istorie) este că cu o singură excepție, aceasta fiind selecția şi supraviețuirea deterministe (DD), turnirul (când este folosit atât în selecție cât şi/sau în supraviețuire) măreşte hazardul.

În Figura 8 din Anexa 10 s-a reprezentat supraviețuirea (celor mai dotați indivizi), funcție dată de relația:

$$SDF(X) = \frac{1}{1 - CDF(X)}$$

Se remarcă că în ceea ce priveşte supraviețuirea (Figura 8 din Anexa 10) nu se remarcă diferențe deosebite între evoluțiile date de diferitele modalități de selecție şi de supraviețuire, fapt aşteptat deoarece evoluția s-a bazat pe păstrarea celor mai dotați indivizi (cei ce dau cea mai bună regresie) în cultivar în toate cazurile, metoda de supraviețuire permițând prin intermediul celor 3 modalități (determinist, în turnir şi proporțional) alegerea indivizilor care vor fi înlocuiți (omorâți) doar dintre cei rămaşi (care nu sunt dintre cei mai dotați).

Caracterizarea statistică a distribuției evoluției

Estimarea parametrilor distribuției Fisher-Tippett pornind de la observațiile coeficientului de determinare în cele 46 de execuții independente face posibilă urmărirea valorilor parametrului statistic k (forma distribuției) pentru fiecare generație observată (0..20000) şi pentru fiecare asociere de selecție şi supraviețuire aşa cum este redat în Tabelul 11-8.

Tabelul 11-8

Tipul formei distribuției Fisher-Tippett (*Type of the shape of Fisher-Tippett distributions*)

Valori extreme de tip *Extreme values of type*	PP	PT	PD	TP	TT	TD	DP	DT	DD
I ($\|k\|<10^{-2}$) ≈ Gumbel	778 (3.9%)	0 (0%)	317 (1.6%)	63 (0.3%)	23 (0.1%)	992 (5%)	3237 (16.2%)	1091 (5.5%)	292 (1.5%)
II ($k>10^{-2}$) = Fréchet	324 (1.6%)	0 (0%)	299 (1.5%)	0 (0%)	36 (0.2%)	2158 (10.8%)	9012 (45.1%)	1619 (8.1%)	0 (0%)
III ($k<-10^{-2}$) = Weibull	18899 (94.5%)	20001 (100%)	19385 (96.9%)	19938 (99.7%)	19942 (99.7%)	16851 (84.3%)	7752 (38.8%)	17291 (86.5%)	19709 (98.5%)

Se remarcă în Tabelul 11-8 selecția deterministă însoțită de supraviețuire proporțională (DP) care are în compoziția sa o cantitate însemnată (45.1%) de valori extreme de tip II şi selecția proporțională însoțită de supraviețuire în turnir (PT) care are în compoziția sa numai valori extreme de tip III.

Obținerea parametrilor distribuțiilor Fisher-Tippett face posibilă estimarea parametrilor statistici ai populațiilor. Astfel, pentru legea de distribuție Fisher-Tippett exprimată în forma:

$$FT_{PDF}(X) = \frac{1}{\beta}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)(1+k\frac{x-\lambda}{\beta})^{-1-1/k}$$

se pot calcula parametrii statistici ai populațiilor, când se obține:

÷ media:

$$\mu = \lambda - \frac{\beta}{k}\left(1-\Gamma(1-k)\right);\ \Gamma(z) = \int_0^{\infty} e^{-t}t^{z-1}dt$$

÷ mediana:

$$\hat{\mu} = \lambda - \frac{\beta}{k}\left(1-(\ln 2)^{-k}\right)$$

÷ moda:

$$\tilde{\mu} = \lambda - \frac{\beta}{k}\left(1-(1+k)^{-k}\right)$$

÷ deviația standard:

$$\sigma = \frac{\beta}{|k|}\sqrt{\Gamma(1-2k)-\Gamma^2(1-k)}$$

÷ asimetria:

$$\gamma_1 = \frac{2\Gamma^3(1-k)-3\Gamma(1-k)\Gamma(1-2k)+\Gamma(1-3k)}{\left(\Gamma(1-2k)-\Gamma^2(1-k)\right)^{3/2}}$$

÷ excesul de boltire:

$$\gamma_2 = \frac{-6\Gamma^4(1-k)+12\Gamma^2(1-k)\Gamma(1-2k)-3\Gamma^2(1-2k)-4\Gamma(1-k)\Gamma(1-3k)+\Gamma(1-4k)}{\left(\Gamma(1-2k)-\Gamma^2(1-k)\right)^2}$$

Valorile calculate ale parametrilor de mai sus au fost reprezentate grafic (Figurile 11-16..11-24.

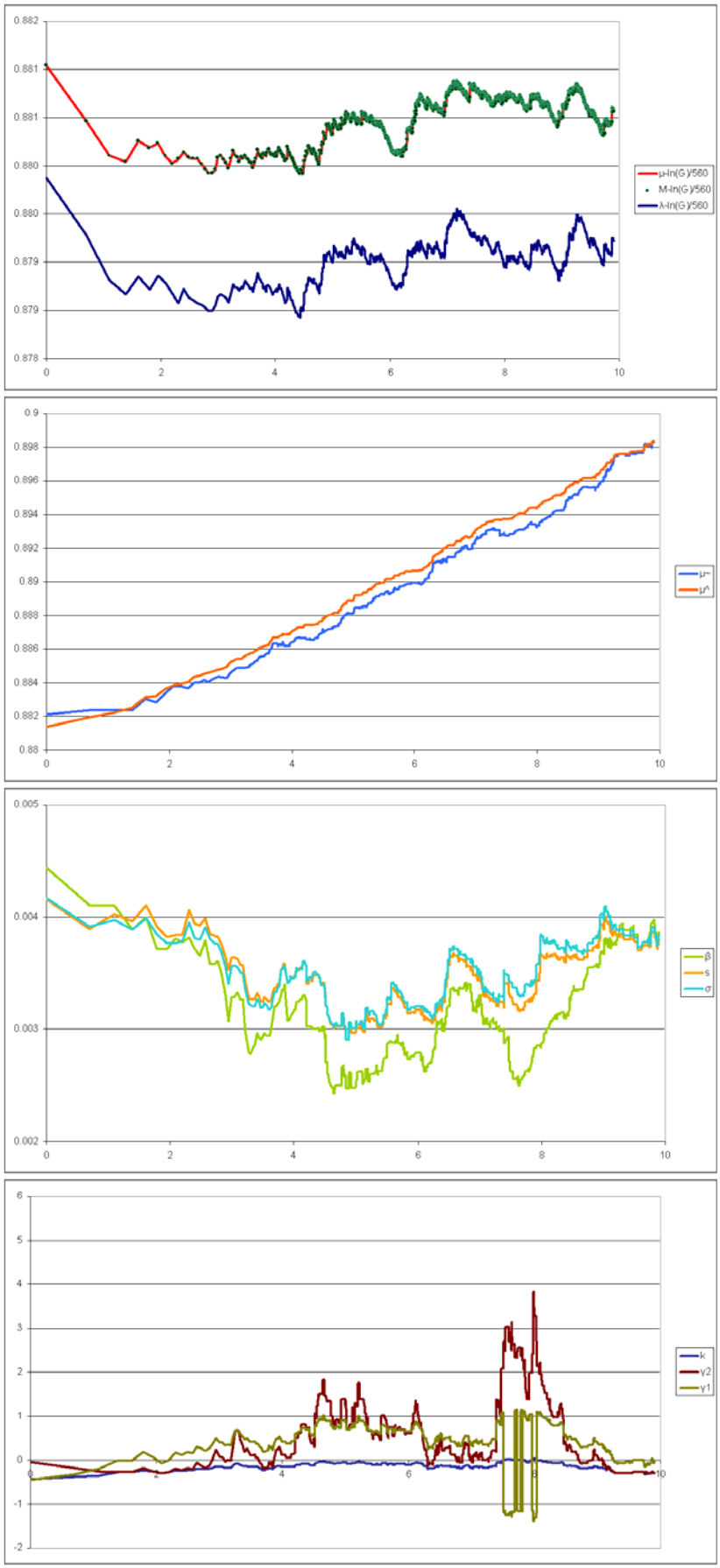


Fig. 11-16. Evoluția parametrilor statistici ai distribuției Fisher-Tippett pentru selecția proporțională însoțită de supraviețuire proporțională (PP)

Evolution of the statistical parameters of Fisher-Tippett distribution for proportional selection and proportional survival (PP)

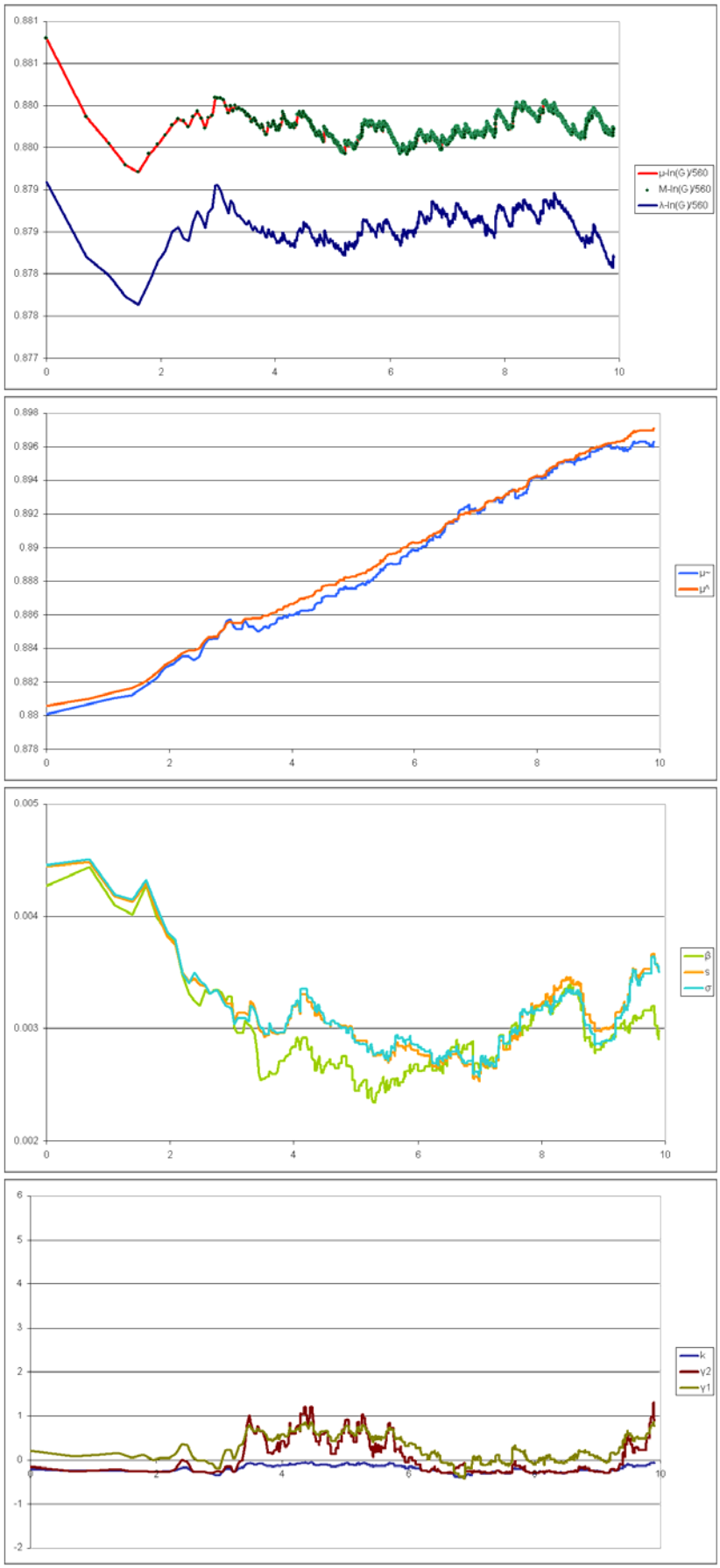


Fig. 11-17. Evoluția parametrilor statistici ai distribuției Fisher-Tippett pentru selecția proporțională însoțită de supraviețuire în turnir (PT)

Evolution of the statistical parameters of Fisher-Tippett distribution for proportional selection and tournament survival (PT)

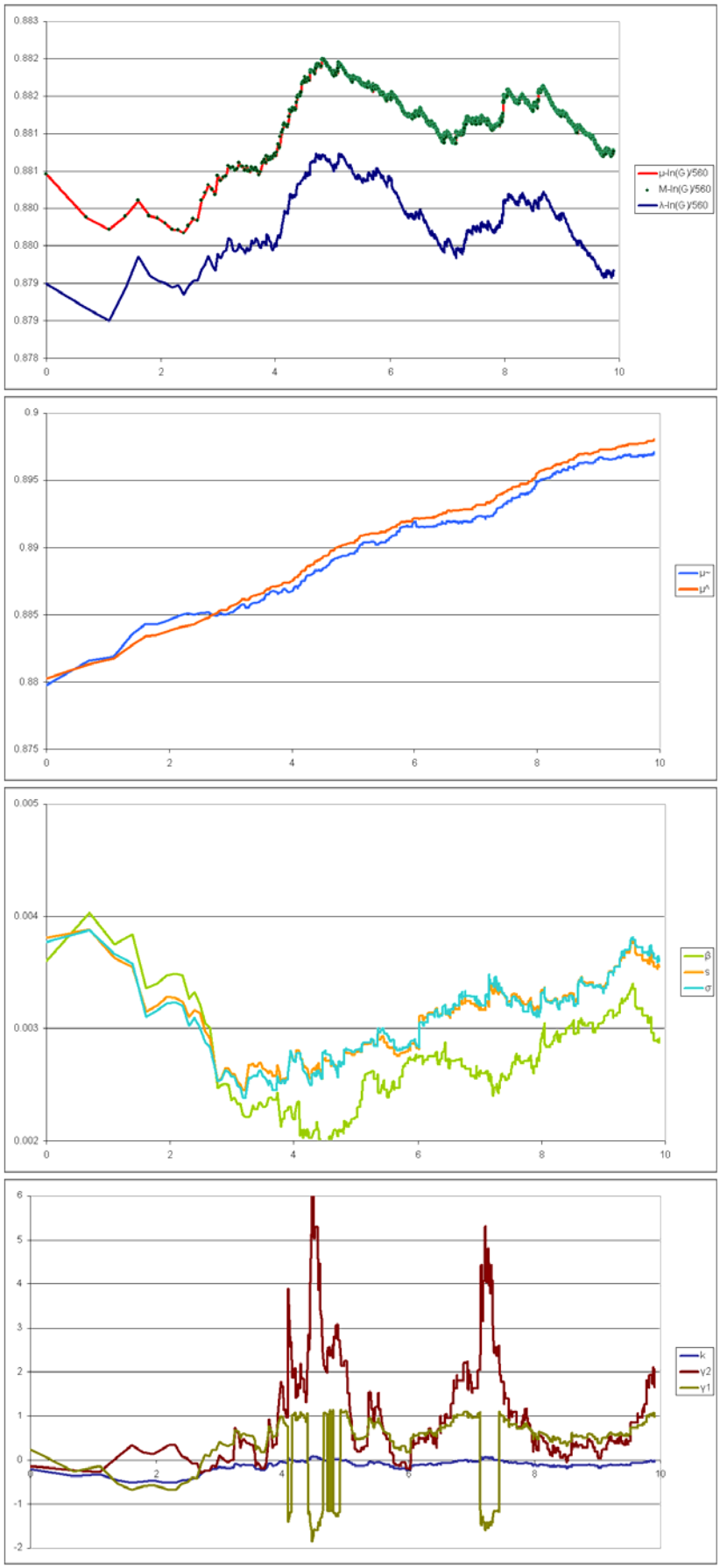


Fig. 11-18. Evoluția parametrilor statistici ai distribuției Fisher-Tippett pentru selecția proporțională însoțită de supraviețuire deterministă (PD)

Evolution of the statistical parameters of Fisher-Tippett distribution for proportional selection and deterministic survival (PD)

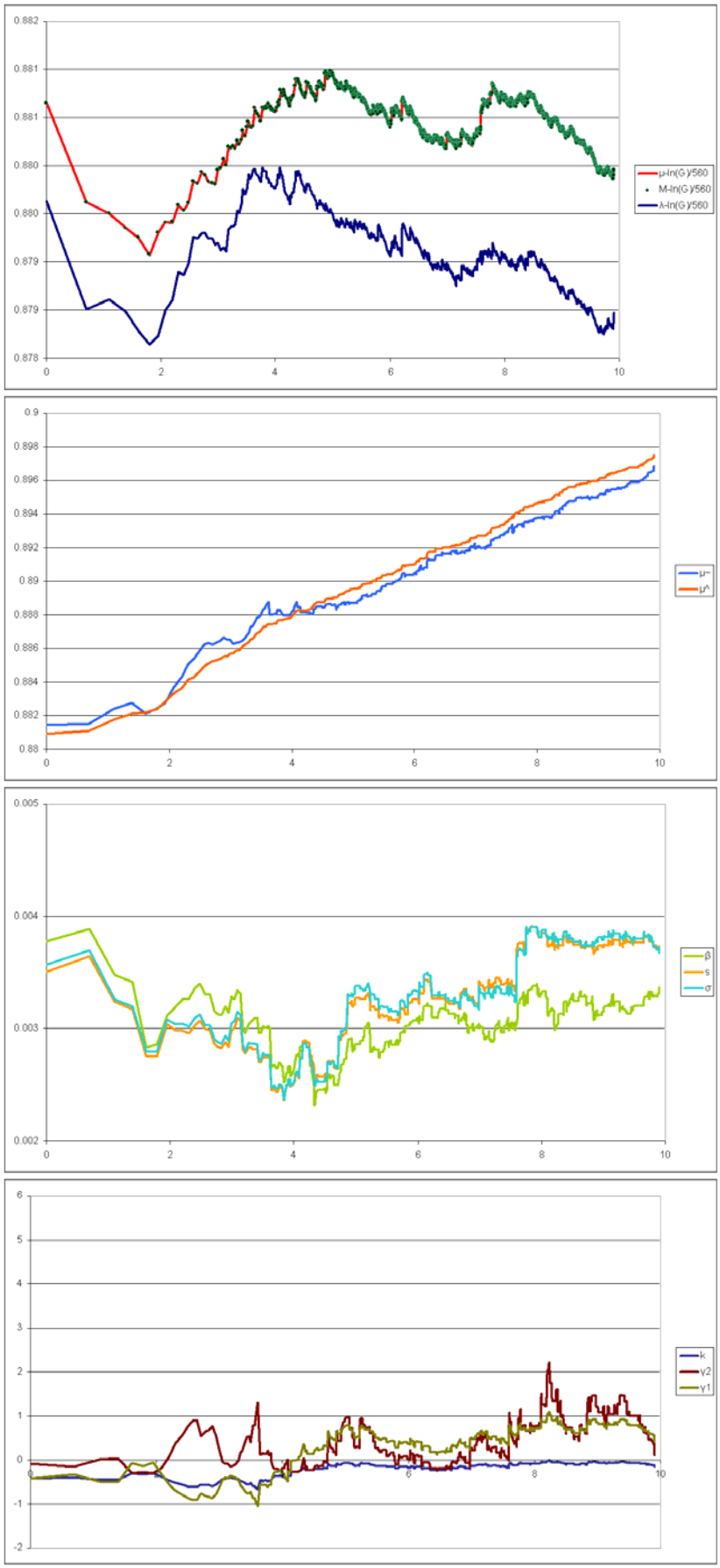


Fig. 11-19. Evoluția parametrilor statistici ai distribuției Fisher-Tippett pentru selecția în turnir însoțită de supraviețuire proporțională (TP)

Evolution of the statistical parameters of Fisher-Tippett distribution for tournament selection and proportional survival (TP)

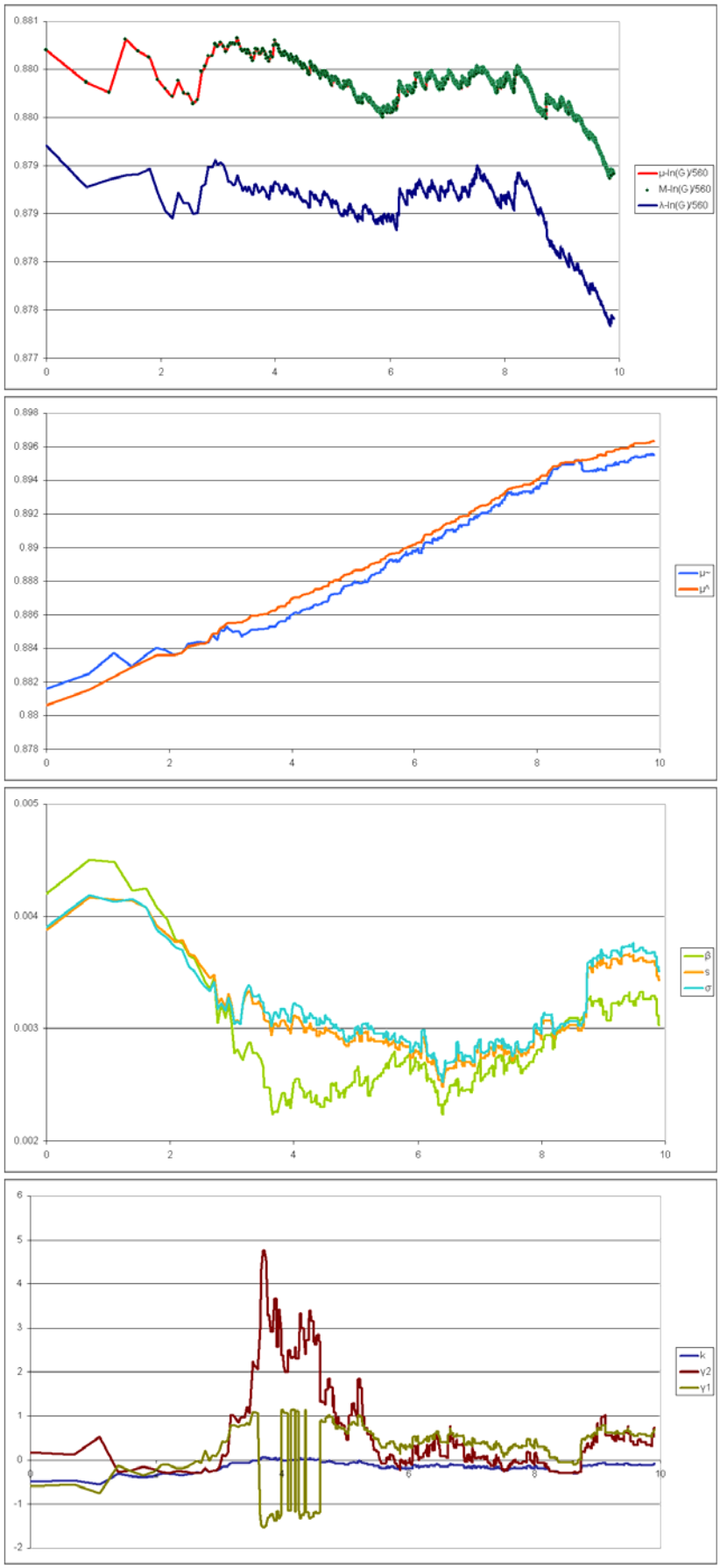


Fig. 11-20. Evoluția parametrilor statistici ai distribuției Fisher-Tippett pentru selecția în turnir însoțită de supraviețuire în turnir (TT)

Evolution of the statistical parameters of Fisher-Tippett distribution for tournament selection and tournament survival (PT)

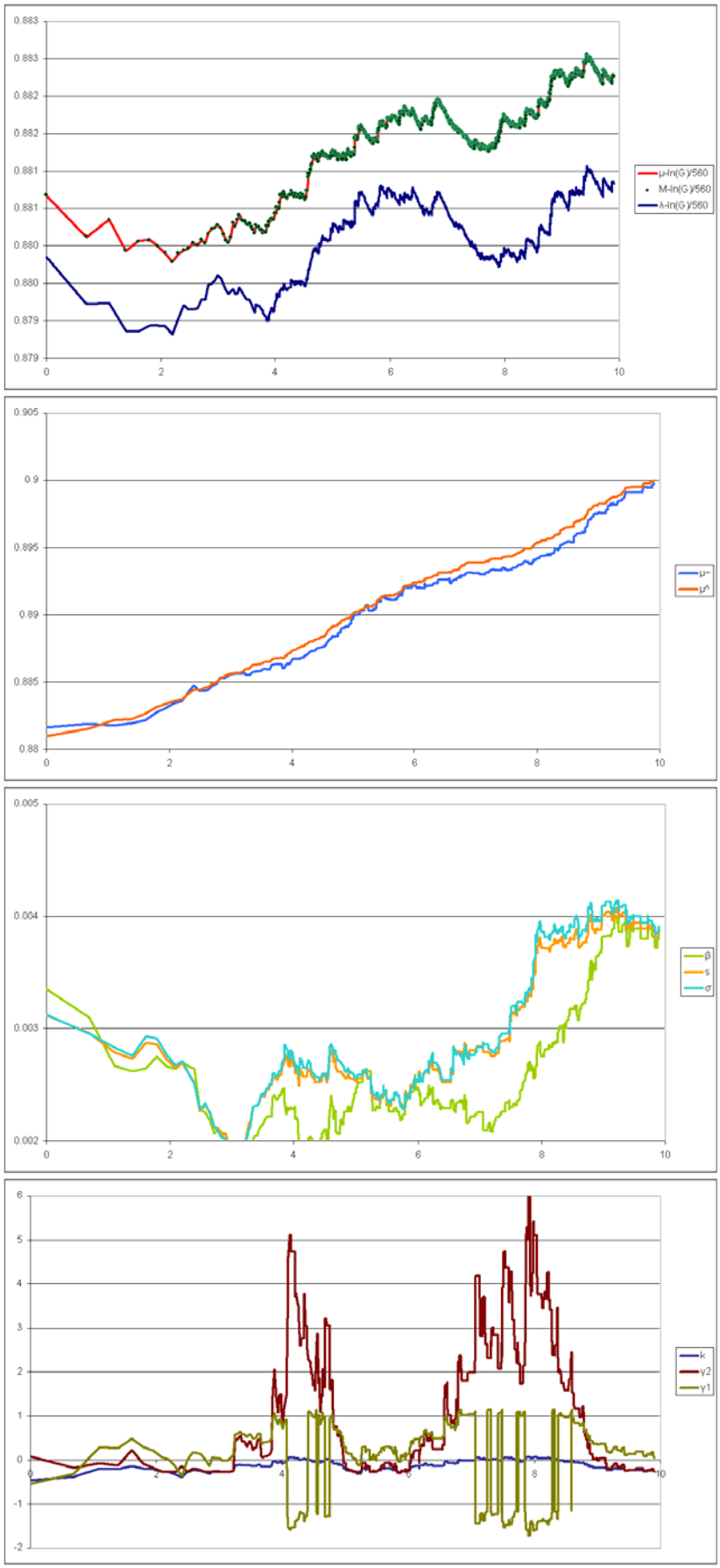


Fig. 11-21. Evoluția parametrilor statistici ai distribuției Fisher-Tippett pentru selecția în turnir însoțită de supraviețuire deterministă (TD)

Evolution of the statistical parameters of Fisher-Tippett distribution for tournament selection and deterministic survival (TD)

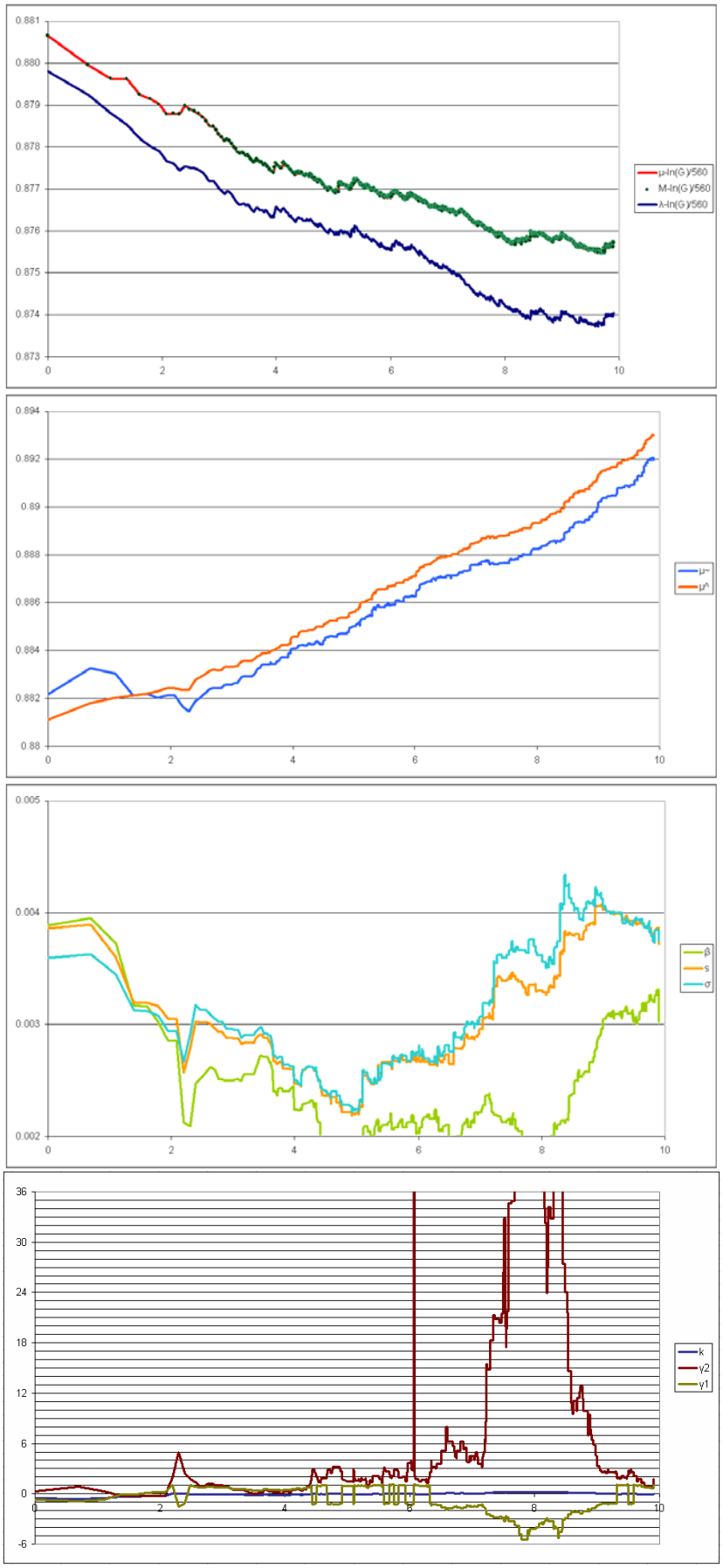


Fig. 11-22. Evoluția parametrilor statistici ai distribuției Fisher-Tippett pentru selecția deterministă însoțită de supraviețuire proporțională (DP)

Evolution of the statistical parameters of Fisher-Tippett distribution for deterministic selection and proportional survival (DP)

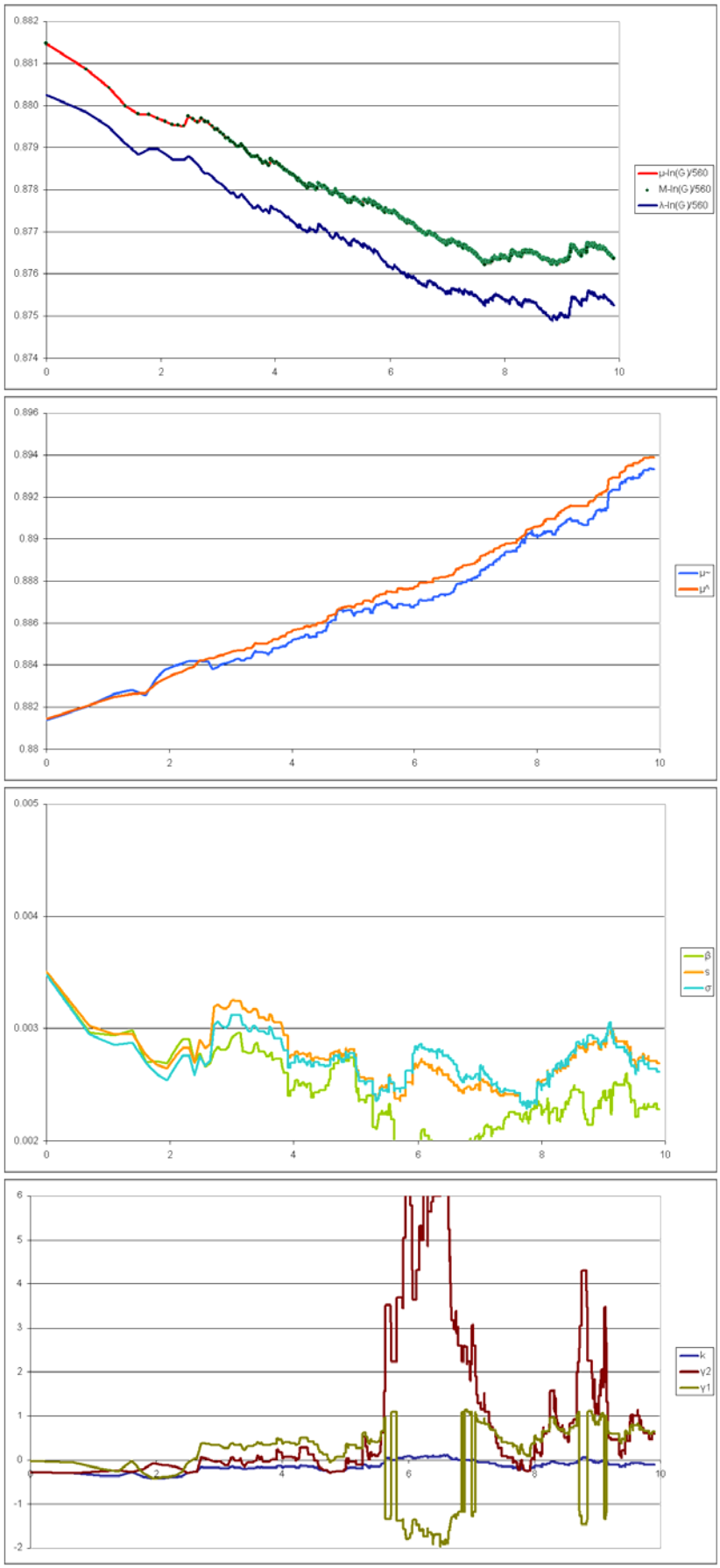


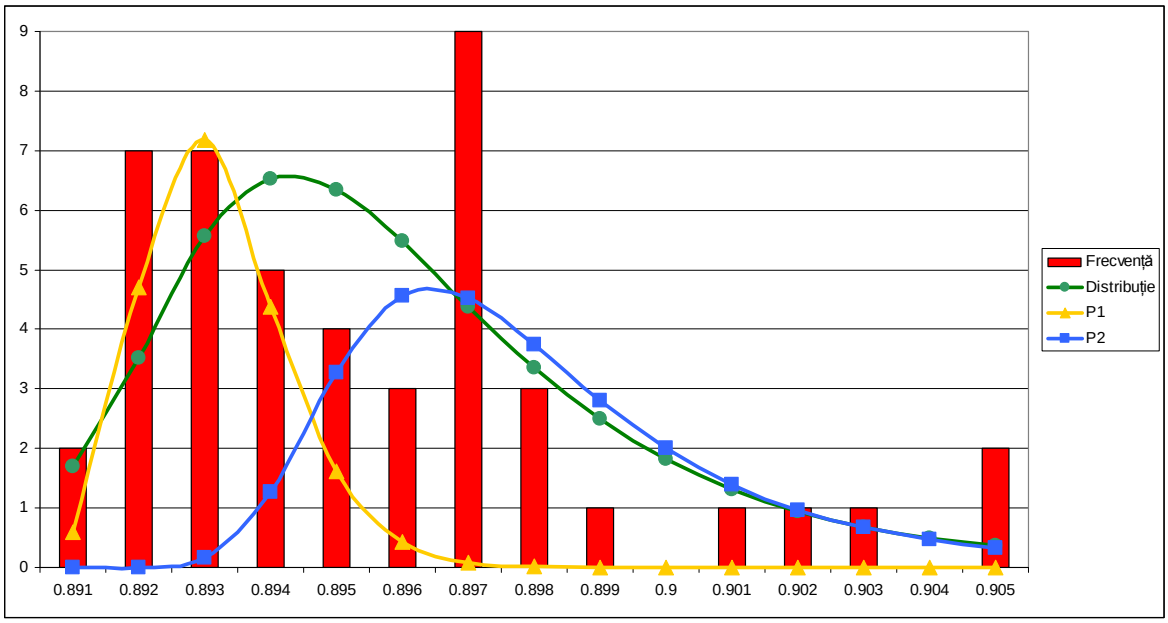
Fig. 11-23. Evoluția parametrilor statistici ai distribuției Fisher-Tippett pentru selecția deterministă însoțită de supraviețuire în turnir (DT)

Evolution of the statistical parameters of Fisher-Tippett distribution for deterministic selection and tournament survival (DT)

Figura 11-24. Evoluția parametrilor statistici ai distribuției Fisher-Tippett pentru selecția deterministă însoțită de supraviețuire deterministă (DD)

Evolution of the statistical parameters of Fisher-Tippett distribution for deterministic selection and deterministic survival (PT)

Oscilațiile pronunțate (Figurile 11-16..11-24) ale asimetriei spre valori mici (γ_1) şi boltirii (γ_2) spre valori foarte mari arată că varianța este datorată unor evenimente rare de deviații extreme, în timp ce Figurile 11-25 şi 11-26 conțin două astfel de cazuri extrase din datele observate pentru exemplificare.



Distribuție = FT(k:0.06441, β:0.00256, λ:0.89394)
P1 = {0.891(2), 0.892(7), 0.893(6), 0.894(4), 0.895(1)}; FT(k: -0.1011, β: 9.28e-4, λ: 0.8922)
P2 = {0.893(1), 0.894(2), 0.895(3), 0.896(3), 0.897(9), 0.898(3), 0.899(1), 0.901(1), 0.902(1), 0.903(2), 0.905(2)}; FT(k: 0.1158, β: 0.00211, λ: 0.8962)

Fig. 11-25. O posibilă segregare populațională în selecția în turnir însoțită de supraviețuirea deterministă (DT, Generația: 2605..2631)

A possble segregation in the population when tournament selection and deterministic survival were used (DT, Generation: 2605..2631)

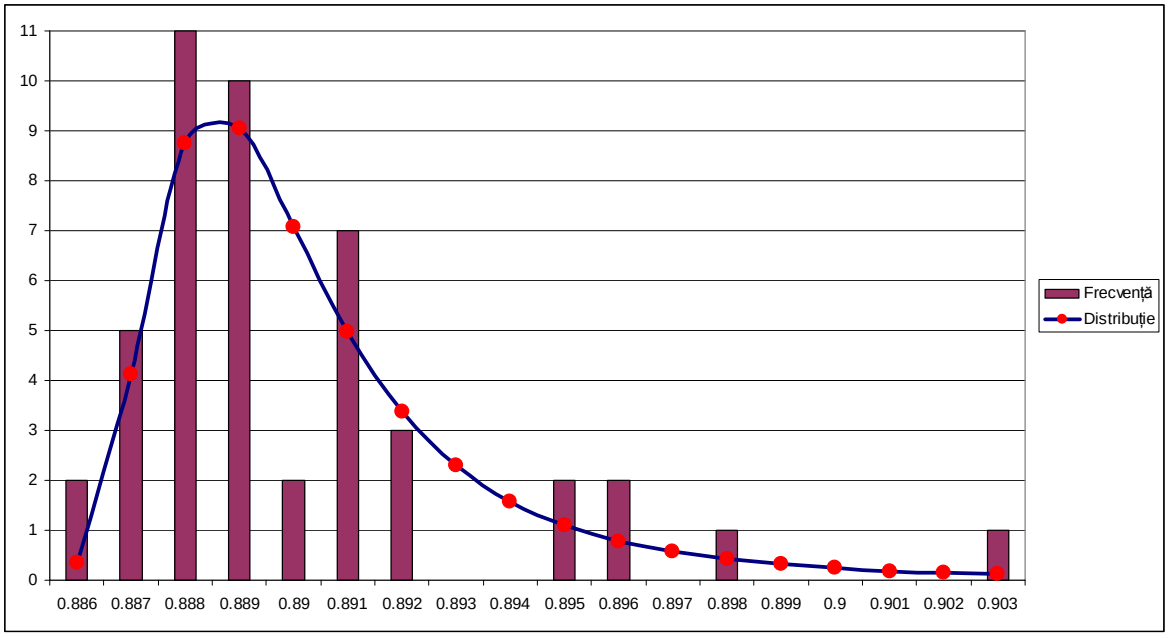


Fig. 11-26. Distribuția determinării în intervalul de extremă boltire (Generația: 2415..2642; γ_2 = 1163) şi varianță ridicată al selecției deterministe însoțite de supraviețuire proporțională (DP)

Distribution of the determination in the interval of highest kurtosis (Generation: 2415..2642; γ_2 = 1163) and high variance when deterministic selection and proportional survival were used

Mai mult (Figurile 11-16..11-24) cu o singură excepție (selecția deterministă însoțită de supraviețuirea proporțională - DP - Figura 11-22) aceste deviații extreme sunt însoțite de scăderea statisticilor variabilității (s, σ şi β) ceea ce face ca aceste momente să poată fi considerate momente critice ale evoluției, momente în care boltirea arată existența temporară a unei segregări (Figura 11-25, împărțire în două populații distincte, dar care pe parcursul evoluției datorită recombinărilor se diluează până la dispariție).

Figurile 11-16..11-24 arată că aceste momente critice ale evoluției se petrec începând de la generația $e^4 \approx 60$ până la generația $e^8 \approx 3000$ şi o posibilă explicație este saturarea variabilității materialului genetic din eşantinul supus evoluției în cultivar. Surprinde desigur (Figura 11-26) extrema boltire (segregare) şi de durată (Figura 11-22) a selecției deterministe însoțite de supraviețuirea proporțională (DP), făcută pe fondul creşterii variației, dar nu mai mare decât varianța observată la începutul evoluției, dar care şi aceasta se diluează în timp.

12. Ce lege urmează momentele de apariție a evoluției?

Familia de curbe log-Pearson de tipul III

În această secțiune s-a căutat răspuns le următoarea întrebare: *Ce lege urmează momentele de apariție a evoluției?*

S-au încercat diferite modalități de a corela momentele de apariție a evoluției (generațiile în care se aduce o îmbunătățire pentru coeficientul de determinare al regresiei în 4 variabile cu descriptori membrii ai familiei MDF şi activitatea observată lkow a PCBs). Rezultate relevante s-au obținut aplicând o procedură de transformare asupra datelor primare, procedură exemplificată în Tabelul 12-1 (exemplu din run 1, DP).

Tabelul 12-1

Transformarea momentelor evoluției la durate relative (exemplu din run 1, DP)
Transformation of the moments of evolution to relative (example from run 1, DP)

Numărul generației	0	15	136	188	246	528	5423	11887
Momentele evoluției	1	16	137	189	247	529	5424	11888
Durata până la viitoarea evoluție	15	121	52	58	282	4895	6464	
Durata relativă la momentul evoluției	1.5E+1	7.6E+0	3.8E-1	3.1E-1	1.1E+0	9.3E+0	1.2E+0	1.7E-4

Un prim răspuns la întrebarea "*Ce lege urmează momentele de apariție a evoluției?*" se poate obține complectând întrebarea astfel "*Ce lege urmează momentele de apariție a evoluției independent de strategia de evoluție urmată?*".

Răspunsul la această întrebare a fost obținut imediat astfel:

÷ S-au obținut momentele relative aşa cum descrie Tabelul 12-1 pentru fiecare run (1..46) şi fiecare strategie (PP, PT, PD, TP, TT, TD, DP, DT, DD);

÷ Fiecare şir de genul celui ce rezultă din linia "Durata relativă la momentul evoluției" a Tabelului 12-1 reprezintă o observație a evoluției; două astfel de linii reprezintă două observații ale aceluiaşi fenomen, dar în acelaşi timp poate fi privită ca o singură observație din punctul de vedere al distribuției valorilor, una venind să o complecteze pe cealaltă în ceea ce priveşte observarea momentelor; prin extensie, toate la un loc (în număr de 414) sunt cea mai largă observație a momentelor evoluției care poate fi constituită din datele înregistrate din execuția algoritmului genetic;

÷ Fişierul rezultat descris din procedura de mai sus cumulează 11347 momente de evoluție; acestea au intrat în analiza de distribuție;

÷ Analiza de distribuție a fost realizată cu EasyFit având ca alternative un număr de 65 de legi de distribuție continue; au fost folosite măsurarea agrementului între observație şi model un număr de 3 statistici (C-S, A-D şi K-S).

÷ Rezultatele obținute au fost edificatoare cu privire la ipoteza de distribuție care poate fi formulată

asupra momentelor evoluției; prima porțiune a acestei analize este redată în Tabelul 12-2 (cuprinzând primele 3 legi de distribuție în ordinea agrementului între observație și model).

Tabelul 12-2

Legi de distribuție pentru momentele relative ale evoluției
Distribution laws for relative moments of evolution

Dist\Stat	K-S	p(K-S)	Rang	A-D	p(A-D)	Rang	C-S(df)	p(C-S)	Rang
Log-P-3	0.01197	0.07683	1	2.4264	0.05617	1	41.731(13)	7.3E-05	1
Burr	0.01635	4.57E-03	3	6.7901	3.23E-04	3	46.345(13)	1.25E-05	2
Burr-4P	0.01592	6.27E-03	2	6.0813	7.48E-04	2	51.408(13)	1.71E-06	3
Dist: Lege de distribuție; Stat: Statistică; Rang: Rangul statisticii în lista celor 65 de alternative *Dist: Distribution law; Stat: Statistic; Rang: Rank of the statistic from 65 alternatives* Log-P-3: log-Pearson de tipul III (*Log-Pearson of type III*)									

Rezultatele din Tabelul 12-2 arată următoarele:

1. O singură lege de distribuție din cele 65 de alternative se califică ca ipoteză de distribuție pentru momentele relative ale evoluției; este singura pentru care se obțin riscuri de a fi în eroare la respingerea distribuției mai mari de 1%, care sunt de fapt în cazul distribuției Log-Pearson 3 mai mari de 5% (7.68% K-S și 5.62% A-D în Tabelul 12-2);
2. Din modalitatea de calcul se poate evidenția că statistica K-S măsoară agrementul între rangurile observațiilor, în timp ce statistica C-S măsoară agrementul între valorile observațiilor, iar statistica A-D chiar dacă folosește ranguri (ca și K-S) este totuși o măsură care o apropie de C-S;
3. Statistica C-S pentru Log-Pearson 3 are valoarea $7.3 \cdot 10^{-5}$, adică observații mai defavorabile agrementului cu modelul dat de Log-Pearson 3 se obțin în mai puțin de 0.08‰ din cazuri; valoarea probabilității de observare scade de la 7.7% pentru K-S, la 5.6% pentru A-D ca să ajungă la 0.08% pentru C-S;
4. Având în vedere ce măsoară statisticile C-S, A-D și K-S (remarca 2 de mai sus) și valorile obținute pentru aceste statistici (remarca 3 de mai sus) se desprinde concluzia că agrementul între rangurile observate ale momentelor relative ale evoluției și rangurile din distribuția teoretică Log-Pearson 3 este mai probabil decât agrementul între valorile observate ale momentelor relative ale evoluției și valorile din distribuția teoretică Log-Pearson 3;
5. Există o explicație pentru remarca 4, și anume eșantionul de evoluții a cuprins observații din toate cele nouă asocieri de strategii; este astfel așteptat ca agrementul să fie mult mai mare pentru ranguri (statisticile K-S și A-D) decât pentru valori (statistica C-S).
6. Având în vedere remarcile 4 și 5 de mai sus, concluzia care se obține în urma analizei este că se poate accepta cu riscul de a fi în eroare de 5% (superior probabilităților de a greși din Tabelul 12-2 pentru Log-Pearson 3 folosind statisticile K-S & A-D) că Log-Pearson 3 este legea de distribuție a momentelor relative ale evoluției obținute prin procedura descrisă în Tabelul 12-1.

Concluzia că "*Log-Pearson 3 este legea de distribuție a momentelor relative de evoluției independent de strategia de selecție și de supraviețuire*" se poate verifica. În acest sens pentru

fiecare pereche de strategie de selecție și supraviețuire din cele 46 de execuții independente au fost cumulate momentele relative ale evoluției descrise de procedura din Tabelul 12-1. Rezultatele sunt disponibile online:

http://l.academicdirect.org/Horticulture/GAs/MLR_MDF_selection_vs_survival/evolut_SS.txt

Rezultatele au intrat în aceași analiză a agrementului între observație și model cuprinsă în Tabelul 12-2, și fișierele rezultat se află la aceași adresă ca mai sus, denumirile fișierelor fiind date în Tabelul 12-3, împreună cu semnificațiile statistice ale agrementului între observații și modelul dat de distribuția Log-Pearson 3 (asupra căreia deja s-a decis că dacă ea este singura alternativă posibilă ca lege de distribuție din cele 65 investigate).

Tabelul 12-3

Agrementul între observație și Log-Pearson 3 pentru distribuția momentelor relative ale evoluției
The agreement between observations and Log-Pearson 3 for distribution or relative moments of evolution

Stra\Stat	nr.Obs	K-S	$p_{K\text{-}S}$	A-D	$p_{A\text{-}D}$	C-S(df)	$p_{C\text{-}S}$
TT	1379	0.02284	0.46	0.63251	0.47	12.3(10)	0.27
TD	1429	0.01224	0.98	0.23477	0.75	3.3064(10)	0.97
TP	1318	0.02691	0.29	1.2118	0.24	14.35(10)	0.16
DT	996	0.02845	0.39	0.73496	0.41	10.628(9)	0.30
DD	1084	0.01919	0.81	0.34184	0.66	8.1401(10)	0.62
DP	851	0.02416	0.69	0.6234	0.47	6.8598(9)	0.65
PT	1463	0.0203	0.58	0.70531	0.43	12.512(10)	0.25
PD	1474	0.03055	0.13	0.93998	0.33	8.6564(10)	0.56
PP	1353	0.01212	0.99	0.23201	0.75	3.5574(10)	0.97
Stra (DD, DP, DT, PD, PP, PT, TD, TP, TT): strategie (*strategy*) Stat (nr.Obs, K-S, $p_{K\text{-}S}$, A-D, $p_{A\text{-}D}$, C-S(df), $p_{C\text{-}S}$): statistică (*statistic*)							

Tabelul 12-3 demonstrează că ipoteza "*Log-Pearson 3 este legea de distribuție a momentelor relative de evoluției independent de strategia de selecție și de supraviețuire*" formulată pe baza analizei din Tabelul 12-2 se verifică pentru fiecare strategie de selecție și supraviețuire în parte.

Inspectând valorile din Tabelul 12-3 se evidențiază că:

÷ nu există nici o respingere a ipotezei formulate la un risc de a fi în eroare de 10% sau mai mic;

÷ există două respingeri (din 27 de cazuri) la un risc de a fi în eroare de 20% sau mai mic (PD pentru K-S & TP pentru C-S) care este o situație așteptată (fixându-se un risc de a fi în eroare de 20% s-a făcut în fapt o eroare inferioară lui 20% de 2/27 = 7.4%);

÷ dacă statistica C-S respingea ipoteza cu un risc de a fi în eroare mai mic de 0.08‰ pentru eșantionul reunind observațiile de la toate strategiile, atunci când se analizează separat fiecare strategie în parte valoarea riscului de a fi în eroare în respingerea ipotezei urcă dramatic, cea mai mică valoare a acestuia fiind 16% și având o valoare medie de 53%;

÷ explicația pentru dezagrementul măsurat de statistica C-S între observație și modelul Log-Pearson 3 remarcat în Tabelul 12-2 și explicația de mai sus pentru agrementul măsurat de statistica C-S între observație și modelul Log-Pearson 3 remarcat în Tabelul 12-3 aduce pe cale de consecință

că:

- ipoteza "*Log-Pearson 3 este legea de distribuție a momentelor relative de evoluției independent de strategia de selecție şi de supraviețuire*" este susținută de observații;
- parametrii legii de distribuție Log-Pearson 3 sunt însă dependenți de strategia de selecție şi de supraviețuire (dezagrement în Tabela 12-2, agrement în Tabela 12-3);
- nu există nici un motiv pentru a presupune că parametrii legii de distribuție Log-Pearson 3 sunt dependenți de altceva decât de strategia de selecție şi de supraviețuire (pentru experimentul desfăşurat) - motivul - agrementul din Tabela 12-3 înregistrat în toate cele 9 asocieri de strategii de selecție şi supraviețuire;

Sintetizând, o concluzie importantă a fost extrasă din studiul momentelor relative ale evoluției, şi anume că Log-Pearson 3 este legea de distribuție a momentelor relative de evoluției iar parametrii distribuției depind de strategia de selecție şi de supraviețuire.

Această din urmă concluzie deschide calea analizei parametrilor distribuției Log-Pearson 3.

Distribuția Pearson de tipul III este o familie de legi de distribuție (fiind parametrizată în consecință ca şi distribuția Fisher-Tippett de 3 parametrii şi nu doar de 2) este un caz particular al unei familii de legi de distribuție (numite astăzi Pearsoniene) dezvoltate şi clasificate de Karl PEARSON în 1895 într-o lucrare intitulată - în traducere aproximativă - "Contribuții la teoria matematică a evoluției. II. ..." (Pearson, 1895). Distribuția Pearson de tipul III a fost ulterior adusă în actualitate de RA Fisher în lucrarea sa celebră despre fundamentele matematice ale statisticii teoretice, acordându-i un amplu spațiu în discuție - pag. 332-337 (şi nu numai) în (Fisher, 1922-Estimation). Distribuția Log-Pearson 3 a fost propusă inițial ca o modalitate de calcul al parametrilor distribuției Pearson de tipul III prin logaritmarea prealabilă a datelor, modalitate care este consistentă cu estimarea parametrilor distribuției (Phien & Hira, 1983; Arora K & Singh, 1989) din momente centrale (Fisher, 1922-Estimation), pentru ca ulterior să devină întâi un ghid [1] şi apoi lege de distribuție ca atare [2].

În literatura de specialitate se evidențiază frecvent utilizarea distribuției Log-Pearson 3 în studiul frecvenței inundațiilor (Durrans & Tomic, 1996; Khan, 2009) - în timp ce alți autori contestă eficiența modelului Log-Pearson 3 pentru frecvența inundațiilor (Vogel & alții, 1993).

Distribuția Log-Pearson 3 este definită de 3 parametrii continui: $\alpha > 0$, $\beta \neq 0$, şi $\gamma \in \Re$. Domeniul de valori ce pot fi observate este dat de valoarea lui β astfel:

÷ pentru $\beta < 0$ domeniul de valori este $0 < x \leq e^{\gamma}$;

---

[1] U.S. Water Resources Council, 1967. Guidelines for determining flood flow frequency. Hydrol. Comm., Bull. 15, Washington, D.C. (also revised versions, Bull. 15, 1975, Bull. 17A, 1977, and Bull. 17B, 1981) http://www.fema.gov/library/viewRecord.do?id=2360

[2] (Donthamsetti, 1980): Donthamsetti VR, 1980. Technical Publication SJ 80-8 Log Pearson Type 3 Distribution: Tables Of Quantiles. Water Resources Department, Project Number 2003233.

÷ pentru $\beta > 0$ domeniul de valori este $e^{\gamma} \le x < +\infty$;

Funcția de densitate de probabilitate $LP3_{PDF}$ este:

$$LP3_{PDF}(x;\alpha,\beta,\gamma) = \frac{1}{x|\beta|\Gamma(\alpha)}\left(\frac{\ln(x)-\gamma}{\beta}\right)^{\alpha-1}\exp\left(-\frac{\ln(x)-\gamma}{\beta}\right)$$

iar funcția cumulativă de distribuție $LP3_{CDF}$ este:

$$LP3_{CDF}(x;\alpha,\beta,\gamma) = \frac{\Gamma_{(\ln(x)-\gamma)/\beta}(\alpha)}{\Gamma(\alpha)}$$

unde $\Gamma(z)$ este funcția Gamma, iar $\Gamma_y(x)$ este funcția Gamma incompletă:

$$\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}dt\ ;\ \Gamma_y(z) = \int_y^{\infty} t^{z-1}e^{-t}dt$$

Parametrii α, β şi γ sunt numiți - în accord cu (Phien & Hira, 1983) - forma (α), scala (β) şi locația (γ) distribuției. Folosind aplicația EasyFitXL, au fost estimați parametrii distribuțiilor Log-Pearson 3 prin maximizarea ratei şansei [3] pentru fiecare strategie de selecție şi supraviețuire (date pe baza cărora s-au obținut valorile statisticilor din Tabelul 12-3), valori, date în Tabelul 12-4.

Tabelul 12-4

Forma, scala şi locația distribuției Log-Pearson 3 a momentelor relative de evoluție*
Shape, scale and location of Log-Pearson 3 for relative moments of evolution (**)

SS	Forma (α) *Shape*	Scala (β) *Scale*	Locația (γ) *Location*
TT	27.666	-0.41137	9.0948
TD	140.85	-0.18338	23.409
TP	37.404	-0.37736	11.725
DT	81.525	-0.26053	19.298
DD	211.9	-0.16693	33.136
DP	294.76	-0.1443	40.665
PT	92.711	-0.22803	18.666
PD	48.108	-0.31208	12.615
PP	73.5	-0.25278	16.289
*: din maximizarea ratei şansei **: *from Maximum Likelihood Estimation*			

Cu ajutorul parametrilor dați în Tabelul 12-4 s-a obținut seria de statistici redată în Tabelul 12-5 pe baza formulelor:

---

[3] Maximizarea ratei şansei (Maximum Likelihood Estimation)

÷ Referințe:

- (Fisher, 1912): Fisher RA, 1912. On an Absolute Criterion for Fitting Frequency Curves. Messenger of Mathematics 41:155-160.
- (Jäntschi, 2009): Jäntschi L, 2009. Distribution Fitting 1. Parameters Estimation under Assumption of Agreement between Observation and Model. Bulletin UASVM Horticulture, 66(2):684-690.

÷ Model matematic:

- Pentru un şir de n observații (X_1, ..., X_n) agrementul între observații şi un model de distribuție dat de o lege de distribuție cunoscută dar cu parametrii necunoscuți (param), maximizarea ratei şansei referă acele valori ale parametrilor necunoscuți pentru care:

$$\text{MLE}: \quad \sum_{i=1}^{n}\log\left(f(X_i;\text{param})\right) \to \max$$

÷ media: $\mu = \int_0^{e^{\gamma}} LP3_{PDF}(z)zdz$;

÷ mediana: $LP3_{CDF}(\tilde{\mu}) = 0.5$ ($\tilde{\mu} = LP3_{InvCDF}(0.5)$);

÷ moda: $\frac{\partial LP3_{PDF}(\hat{\mu})}{\partial \hat{\mu}} = 0$;

÷ deviația standard: $\sigma = \int_0^{e^{\gamma}} (z-\mu)^2 LP3_{PDF}(z)dz$;

÷ asimetria: $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$, unde $\mu_k = \int_0^{e^{\gamma}} (z-\mu)^k LP3_{PDF}(z)dz$, k>1

÷ excesul de boltire: $\gamma_2 = \frac{\mu_4}{\mu_2^{2}} - 3$;

Tabelul 12-5

Statistici ale distribuției Log-Pearson 3 a momentelor relative de evoluție
Statistics of Log-Pearson 3 distribution for relative moments of evolution

SS	μ	$\hat{\mu}$	$\tilde{\mu}$	σ	γ_1	γ_2
TT	0.645	7.18E-5	0.117	2.11	15.6	6.31E+2
TD	0.736	3.37E-4	0.095	4.01	74.6	3.98E+4
TP	0.778	3.24E-5	0.104	3.26	26.8	2.32E+3
DT	1.526	1.15E-4	0.156	8.91	69.6	2.78E+4
DD	1.528	1.09E-4	0.113	13.57	258.1	9.22E+5
DP	2.543	1.40E-4	0.162	26.88	443.6	3.94E+6
PT	0.685	2.19E-4	0.091	3.42	53.3	1.55E+4
PD	0.636	1.58E-4	0.101	2.50	26.5	2.44E+3
PP	0.759	2.65E-4	0.110	3.41	39.8	7.22E+3
μ: Media; $\hat{\mu}$: Moda; $\tilde{\mu}$: Mediana; σ: Deviația standard; γ_1: Asimetria; γ_2: Excesul de boltire *μ: Mean; $\hat{\mu}$: Mode; $\tilde{\mu}$: Median; σ: Standard deviation; γ_1: Asymmetry; γ_2: Kurtosis excess*						

Asocieri în familia de curbe de evoluție

Există asocieri între parametrii formă, scală şi locație ai distribuțiilor Log-Pearson 3 obținuți pentru momentele relative ale evoluției urmând diferite strategii (Tabelul 12-4). Analiza de corelație din Tabelul 12-6 susține această afirmație.

Tabelul 12-6

Asocieri (liniare) între parametrii distribuțiilor de evoluție (Pearson r)
(linear) Associations between parameters of the evolution distribution

r	γ	α	β
γ	1		
α	0.994	1	
β	0.885	0.857	1
r: (Pearson, 1896)			

Intrarea (α, γ) în Tabelul 12-6 arată că există asociere între formă (α) şi locație (γ). Există

posibilitatea ca, în procesul de maximizare a ratei şansei această relație să se fi atenuat, motiv pentru care coeficientul de corelație obținut - r(α, γ) a coborât în valoare de la 1 la 0.994. Pornind cu această ipoteză, Figura 12-1 reprezintă valorile obținute pentru formă (α) şi locație (γ) - valori din Tabelul 12-4.

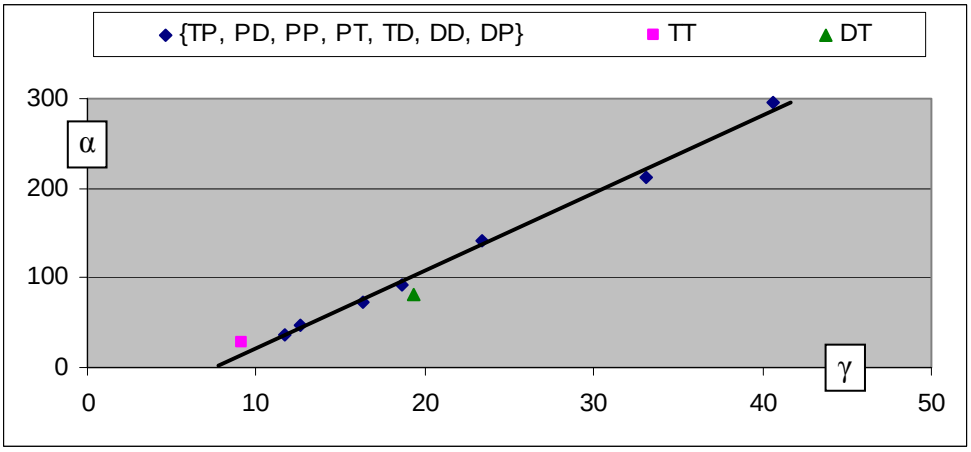


Fig. 12-1. Asocierea între formă (α) şi locație (γ) în familia de curbe Log-Pearson 3 de evoluție
Association between shape (α) and location (γ) in the family of Log-Pearson of type III curves

Analiza de corelație folosind coeficientul de corelație semi-cantitativă λ (Bolboacă & Jäntschi, 2006-PSK) a pus în evidență că:

÷ pentru {TP, PD, PP, PT, TD, DD, DP} $\lambda = 0.998$, $t(7,2,\lambda) = 51$, $p_t(t,5) = 5\cdot10^{-8}$

÷ pentru {TP, PD, PP, PT, TD, DD, DP, TT, DT} $\lambda = 0.988$, $t(9,2,\lambda) = 17$, $p_t(t,7) = 5\cdot10^{-7}$

Analiza de regresie în seria de date {TP, PD, PP, PT, TD, DD, DP} date în Tabelul 12-4 şi reprezentate în Figura 12-1 s-a făcut folosind un model de regresie liniară simplă în formă implicită (Sârbu & Jäntschi, 1998; Jäntschi & Bolboacă, 2007-Linear):

$$b_1\cdot\alpha + b_2\cdot\gamma = 1;\ b_1 = \frac{S(\alpha)\cdot S(\gamma^2) - S(\gamma)\cdot S(\alpha\gamma)}{S(\alpha^2)\cdot S(\gamma^2) - S^2(\alpha\gamma)};\ b_2 = \frac{S(\gamma)\cdot S(\alpha^2) - S(\alpha)\cdot S(\alpha\gamma)}{S(\alpha^2)\cdot S(\gamma^2) - S^2(\alpha\gamma)}$$

Semnificațiile statistice ale parametrilor b_1 şi b_2 din modelul de regresie se pot calcula folosind formula dată în (Fisher, 1922-Regression) p. 611, unde $\varepsilon = 1 - b_1\cdot\alpha - b_2\cdot\gamma$:

$$\Delta = \begin{vmatrix} S(\alpha^2) & S(\alpha\gamma) \\ S(\alpha\gamma) & S(\gamma^2) \end{vmatrix};\ t(b_1) = \frac{b_1}{\sqrt{S(\varepsilon^2)}}\frac{\sqrt{\Delta}}{\sqrt{\Delta_{11}}};\ t(b_2) = \frac{b_2}{\sqrt{S(\varepsilon^2)}}\frac{\sqrt{\Delta}}{\sqrt{\Delta_{22}}};\ \Delta_{11} = S(\gamma^2);\ \Delta_{22} = S(\alpha^2)$$

Rezultatele analizei de regresie sunt:

÷ ecuația: $-1.463\cdot10^{-2}\cdot\alpha + 1.285\cdot10^{-1}\cdot\gamma = 1$;

÷ semnificații statistice: $t(-1.463\cdot10^{-2}) = -4.98$; $t(1.285\cdot10^{-1}) = 6.89$;

÷ probabilități de observare întâmplătoare: $p_t(4.98,5) = 4‰$; $p_t(6.89,5) = 1‰$;

Ecuația $-1.463\cdot\alpha + 12.85\cdot\gamma = 100$ a fost transformată la $\gamma = 0.11386\cdot\alpha + 7.7824$ şi folosită pentru a alinia parametrii distribuțiilor Log-Pearson în acord cu maximizarea ratei şansei pentru fiecare şir de observații $SS_j \in$ {TP, PD, PP, PT, TD, DD, DP, TT, DT}:

$$\sum_{i=1}^{n_{SS_j}} \log\left(LP3_{PDF}(x;\alpha_j,\beta_j,0.11386\alpha_j+7.7824)\right)$$

Tabelul 12-7 prezintă rezultatele obținute pentru formă (α), scală (β) şi locație (γ) redate comparativ cu valorile din Tabelul 12-4 (fiind incluse aici şi valorile maximizate ale ratei şansei (MLE), iar în Tabelul 12-8 valorile statisticilor K-S, A-D şi C-S obținute pe aceeaşi cale ca cele din Tabelul 12-3.

Tabelul 12-7

Forma, scala şi locația distribuției Log-Pearson 3 a momentelor relative de evoluție în ipoteza de asociere liniară între formă şi locație

Shape, scale and location of Log-Pearson 3 distribution of relative moments of evolution under assumption of linear association between shape and location

	LP3(x; α_j, β_j,γ_j)				LP3(x; α_j, β_j, 0.011386α_j + 7.7824)			
SS	Forma (α)	Scala (β)	Locația (γ)	MLE	Forma (α)	Scala (β)	Locația (γ)	MLE
TT	27.666	-0.41137	9.0948	150.1	58.463	-0.28599	14.439	147.2
TD	140.85	-0.18338	23.409	324.0	135.98	-0.18892	23.265	323.7
TP	37.404	-0.37736	11.725	192.9	41.507	-0.35880	12.508	192.5
DT	81.525	-0.26053	19.298	-328.5	33.879	-0.40104	11.640	-335.4
DD	211.90	-0.16693	33.136	-72.80	249.43	-0.15403	36.183	-72.80
DP	294.76	-0.1443	40.665	-387.4	277.43	-0.14866	39.370	-387.4
PT	92.711	-0.22803	18.666	401.1	83.812	-0.23611	17.325	400.9
PD	48.108	-0.31208	12.615	317.7	62.578	-0.27655	14.907	316.9
PP	73.500	-0.25278	16.289	140.4	73.132	-0.25152	16.109	140.3

Tabelul 12-8

Agrementul între observație şi modelul Log-Pearson 3 pentru distribuția momentelor relative ale evoluției în ipoteza de asociere liniară între formă şi locație

Agreement between observation and Log-Pearson 3 distribution of relative moments of evolution under assumption of linear association between shape and location

Stra\Stat	nr.Obs	K-S	$p_{K\text{-}S}$	A-D	$p_{A\text{-}D}$	C-S/df	$p_{C\text{-}S}$
TT	1379	0.02991	0.17	1.14190	0.26	14.3/10	0.16
TD	1429	0.01108	0.99	0.19340	0.79	1.5/10	1.00
TP	1318	0.02707	0.28	1.24890	0.23	14.2/10	0.16
DT	996	0.02688	0.46	0.52884	0.53	7.6/9	0.58
DD	1084	0.01865	0.84	0.34461	0.66	7.9/10	0.64
DP	851	0.02449	0.68	0.61738	0.48	7.2/9	0.62
PT	1463	0.02255	0.44	0.94039	0.33	16.3/10	0.09
PD	1474	0.02962	0.15	0.98832	0.31	11.5/10	0.32
PP	1353	0.01343	0.96	0.27687	0.71	3.7/10	0.96
Stra (DD, DP, DT, PD, PP, PT, TD, TP, TT): strategie Stat (nr.Obs, K-S, $p_{K\text{-}S}$, A-D, $p_{A\text{-}D}$, C-S/df, $p_{C\text{-}S}$): statistică							

Analiza sintetizată în Tabelul 12-8 nu oferă nici un motiv de a respinge ipoteza formulată de asociere liniară între parametrul formă (α) şi parametrul locație (γ) ai distribuțiilor Log-Pearson pentru momentele relative ale evoluției în seria de strategii de evoluție investigată (DD, DP, DT, PD, PP, PT, TD, TP, TT).

Valorile locației (γ) şi scalei (β) din Tabelul 12-7 evidențiază existența şi celei de-a doua asocieri între parametrii distribuțiilor. Ecuația de dependență este însă una după o funcție putere.

Analiza de regresie între locație (γ) şi scală (β) este dată în Figura 12-2, analiza fiind efectuată cu aplicația SlideWrite. Pe baza analizei de regresie neliniare din Figura 12-2 şi pe baza asocierii liniare din Figura 12-1 s-au obținut ecuațiile de asociere între formă (α), scală (β) şi locație (γ) pentru distribuțiile Log-Pearson 3 ale distribuției momentelor relative ale evoluției:

÷ $\alpha = 8.77 \cdot \gamma - 68.3$ (s-a restrâns numărul de cifre semnificative la 3 față de ecuația exprimată anterior şi s-a exprimat invers, adică γ=γ(α) în loc de α=α(γ));

÷ $\beta = -0.14 - 144 \cdot \gamma^{-2.57}$;

c	-b	est-b	Regresie neliniară (*Nonlinear regression*)
14.439	0.28599	0.290678	
23.265	0.18892	0.184222	
12.508	0.3588	0.357915	
11.64	0.40104	0.402157	
36.183	0.15403	0.154214	
39.37	0.14866	0.151442	
17.325	0.23611	0.234335	
14.907	0.27655	0.278819	
16.109	0.25152	0.253735	
	r=0.999488		r^2=0.998976

```
Power  y=a0+a1*pow(x,a2)

r2 Coef Det     DF Adj r2       Fit Std Err     F-Statistic
0.998976        0.998634        0.003194        2925.775603

   Value          Std Error     t-Value       95% Confidence Limits
a0 0.140084       0.003239      43.253108     0.132159        0.14801
a1 143.910225     31.511624     4.566893      66.801281       221.019169
a2 -2.572438      0.090744      -28.348238    -2.794489       -2.350387
```

Fig. 12-2. Asocierea între locație (γ) şi scală (β) în familia de curbe Log-Pearson 3 de evoluție
The association between location and scale in the log-Pearson 3 curves of evolution

Noile ecuații au intrat în procedura de maximizare a ratei şansei (MLE) şi Tabelul 12-9 conține noile valori ale parametrilor distribuțiilor Log-Pearson 3, în timp ce Tabelul 12-10 conține semnificațiile statistice ale modelelor obținute.

Tabelul 12-9

Forma, scala şi locația distribuției Log-Pearson 3 a momentelor relative de evoluție în ipoteza de asociere liniară între formă şi locație şi asociere neliniară între scală şi locație
Shape, scale and location of relative moments of evolution Log-Pearson 3 distribution under assumption of linear association between shape and location and nonlinear association between scale and location

	$LP3(x;\alpha_j,\beta_j,\gamma_j)$				$LP3(x;8.77\cdot\gamma_j-68.3,-0.14-144\cdot\gamma_j^{-2.57},\gamma_j)$			
SS	Forma (α)	Scala (β)	Locația (γ)	MLE	Forma (α)	Scala (β)	Locația (γ)	MLE
TT	27.666	-0.41137	9.0948	150.1	82.293	-0.23659	17.171	146.3
TD	140.85	-0.18338	23.409	324.0	72.113	-0.25562	16.011	323.9
TP	37.404	-0.37736	11.725	192.9	43.589	-0.34725	12.758	192.4
DT	81.525	-0.26053	19.298	-328.5	33.783	-0.40234	11.640	-335.3
DD	211.90	-0.16693	33.136	-72.80	250.61	-0.15404	36.364	-72.80
DP	294.76	-0.1443	40.665	-387.4	222.51	-0.15780	33.160	-390.5
PT	92.711	-0.22803	18.666	401.1	66.295	-0.26890	15.347	401.3
PD	48.108	-0.31208	12.615	317.7	73.919	-0.25189	16.216	316.8
PP	73.500	-0.25278	16.289	140.4	82.373	-0.23645	17.180	140.2
SS	Shape (α)	Scale (β)	Location (γ)	MLE	Shape (α)	Scale (β)	Location (γ)	MLE

Tabelul 12-10

Agrementul observație - model Log-Pearson 3 pentru distribuția momentelor relative ale evoluției în ipoteza de asociere liniară între formă şi locație şi neliniară între scală şi locație

Observation - log-Pearson 3 model agreement for distribution of relative moments of evolution under assumption of linear association between shape and location and nonlinear association between scale and location

SS	nr.Obs	$p_{K\text{-}S}$		$p_{A\text{-}D}$		$p_{C\text{-}S}$	
TT	1379	0.46	0.09	0.47	0.17	0.27	0.12
TD	1429	0.98	0.98	0.75	0.74	0.97	0.77
TP	1318	0.29	0.30	0.24	0.19	0.16	0.10
DT	996	0.39	0.47	0.41	0.52	0.3	0.55
DD	1084	0.81	0.88	0.66	0.66	0.62	0.47
DP	851	0.69	0.14	0.47	0.15	0.65	0.21
PT	1463	0.58	0.68	0.43	0.46	0.25	0.36
PD	1474	0.13	0.08	0.33	0.24	0.56	0.44
PP	1353	0.99	0.90	0.75	0.64	0.97	0.80

÷ $p_{K\text{-}S}$, $p_{A\text{-}D}$, $p_{C\text{-}S}$: probabilități de observare întâmplătoare (*probabilities of a random observation*)

÷ prima valoare p (*first p value*): din MLE parametrii independenți (*from MLE independent parameters*)

a doua valoare p (*second p value*): din MLE cu un parametru independent - γ (*from MLE with a independent parameter - γ)*

Analiza sintetizată în Tabelul 12-10 nu oferă nici un motiv de a respinge ipoteza formulată de asociere liniară între parametrul formă (α) şi parametrul locație (γ) şi de asociere neliniară între parametrul scală (β) şi parametrul locație (γ) ai distribuțiilor Log-Pearson pentru momentele relative ale evoluției în seria de strategii de evoluție investigată (DD, DP, DT, PD, PP, PT, TD, TP, TT).

Dacă se acceptă ipoteza celor două dependențe, atunci valorile statisticilor din Tabelul 12-5 se modifică, fiind redate în Tabelul 12-11, în care s-a adăugat şi valoarea parametrului independent din Tabelul 12-9 (γ - locația).

Tabelul 12-11

Statistici ale distribuției Log-Pearson 3 a momentelor relative de evoluție în ipoteza celor două asocieri între parametrii distribuțiilor

Statistics of Log-Pearson 3 distribution for relative moments of evolution under assumption of the two associations between distribution parameters

SS	γ	μ	$\hat{\mu}$	$\tilde{\mu}$	σ	γ_1	γ_2
TT	17.171	0.737	3.28E-4	0.109	3.34	42.1	8.53E+3
TD	16.011	0.668	2.23E-4	0.097	3.00	39.4	7.01E+3
TP	12.758	0.791	5.03E-5	0.104	3.45	30.6	3.32E+3
DT	11.640	1.242	2.96E-5	0.162	5.15	25.3	1.97E+3
DD	36.364	1.583	1.13E-4	0.112	14.98	319.9	1.66E+6
DP	33.160	1.746	2.38E-4	0.150	13.95	206.9	5.51E+5
PT	15.347	0.643	1.72E-4	0.092	2.88	38.0	6.25E+3
PD	16.216	0.677	2.40E-4	0.098	3.04	39.9	7.27E+3
PP	17.180	0.738	3.29E-4	0.109	3.34	42.1	8.54E+3

γ: Locația; μ: Media; $\hat{\mu}$: Moda; $\tilde{\mu}$: Mediana;
σ: Deviația standard; γ_1: Asimetria; γ_2: Excesul de boltire

Valorile din Tabelul 12-11 evidențiază o serie de asocieri între valorile statisticilor. Astfel, valori pereche au: DP şi DD pentru locație, medie, abatere standard, oblicitate şi exces de boltire, DT şi TP pentru locație şi modă, şi aşa mai departe. Pentru a evidenția aceste asocieri de valori, s-a construit o analiză de componente principale, dată în Figura 12-3.

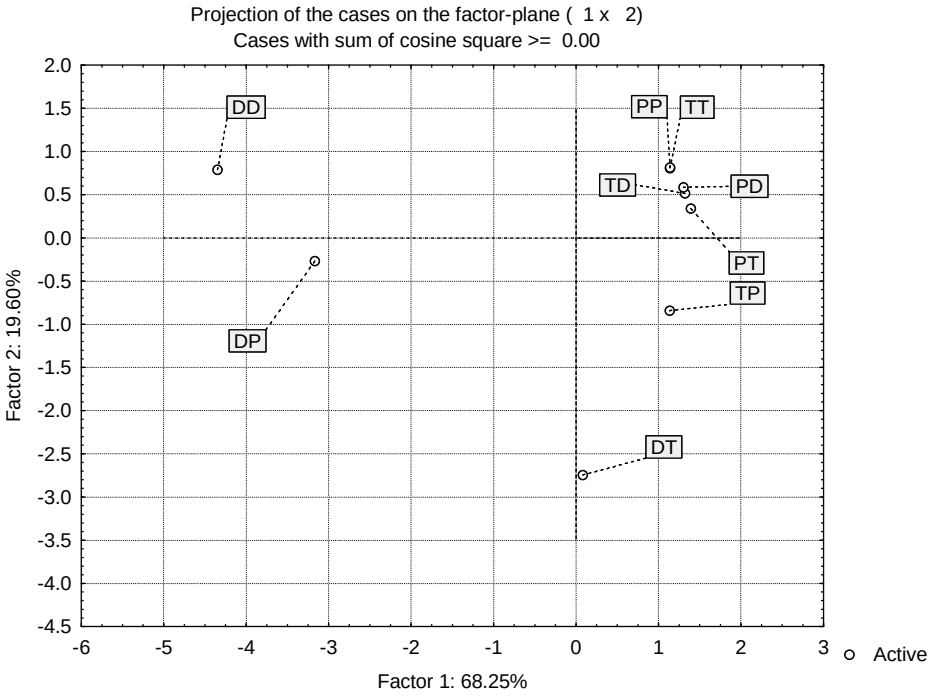


Fig. 12-3. Proiecțiile primilor doi factori în statisticile din Tabelul 12-10
Projections of first two factors in statistics from Table 12-10

Figura 12-3 evidențiază:

÷ Strategiile DD şi DP definesc şi se află aproape în lungul unuia din cei doi factori principali - cel cu 68.25% explicare cantitativă - datorită motivelor evidențiate în analiza valorilor din Tabelul 12-11 şi anume valorile mari pe care aceste strategii le produc pentru majoritatea statisticilor;

÷ Strategia DT se află aproape în lungul unuia din cei doi factori principali - cel cu 19.6% explicare cantitativă - în valorile din Tabelul 12-11 evidențiindu-se doar cu locație, medie şi abatere standard relativ mari în raport cu celelalte strategii, dar cu valori mici - de fapt cele mai mici - pentru asimetrie şi exces de boltire;

÷ Un grup de strategii - PP, TT, TD, PD şi PT este poziționat în planul primilor doi factori principali compact cu diferențe extrem de mici;

÷ Strategia TP este situată în lungul celui de-al doilea factor principal la distanță relativ egală de grupul compact de strategii (format din PP, TT, TD, PD şi PT) şi strategia DT, situându-se în cadranul acesteia din urmă.

Diferențe în familia de curbe de evoluție

O măsură a diferențelor între curbele log-Pearson de tipul III după care se distribuie momentele relative ale evoluției ce urmează diferite strategii este diferența de probabilitate de observare. Pentru a obține această măsură s-au calculat diferențele de probabilitate. Folosind notația:

$$LP3_i(x) = LP3(x; 8.77 \cdot \gamma_i - 68.3, -0.14 - 144 \cdot \gamma_i^{-2.57}, \gamma_i),\ i = 1..9$$

pentru expresiile distribuțiilor de probabilitate ale strategiilor de evoluție TT (i=1), TD (i=2), TP (i=3), DT (i=4), DD (i=5), DP (i=6), PT (i=7), PD (i=8) şi PP (i=9) diferențele de probabiliate sunt date de:

$$diff_{i,j}(x) = \max\left(0, LP3_i(x) - LP3_j(x)\right);\ Diff_{i,j} = \int_0^{\min(e^{\gamma_i}, e^{\gamma_j})} diff_{i,j}(x)dx$$

Funcția $diff_{i,j}(x)$ defineşte evenimentul "la momentul relativ de timp `x` se produce evoluție în strategia `i` şi nu se produce evoluție în strategia `j`". Spațiul de probabilitate al acestei funcții este $(0, \min(\exp(\gamma_i), \exp(\gamma_j)))$ în timp ce cantitatea $Diff_{i,j}$ cumulează probabilitatea de observare pentru "se produce evoluție în strategia `i` şi nu se produce evoluție în strategia `j` pe un interval de timp cu durată nedeterminată".

Diferența maxim posibilă dată de cantitatea $Diff_{i,j}$ este 1 şi corespunde pentru două spații de probabilitate (în cazul de față intervale de timp în care se petrec evenimentele evoluției) complect diferite. Diferența minim posibilă dată de cantitatea $diff_{i,j}$ este 0 şi corespunde pentru două spații de probabilitate (în cazul de față intervale de timp în care se petrec evenimentele evoluției) complect identice.

În altă formulare, cantitatea $Diff_{i,j}$ este probabilitatea ca să se observe (pe o durată de timp nedefinită - de la 0 la infinit) două procese de evoluție, unul urmând strategia de evoluție `i` şi celălalt strategia de evoluție `j` în care cel urmând strategia de evoluție `i` produce evoluții într-o serie de momente relative de evoluție în timp ce cel urmând strategia de evoluție `j` nu produce evoluții în aceleaşi momente. Tabelul 12-12 conține valorile calculate pentru cantitățile $Diff_{i,j}$.

Tabelul 12-12

Diferențe în probabilitățile de observare ale evoluțiilor în diferite strategii

Differences in the probabilities of observation of evolution in different strategies

Diff	TT	TD	TP	DT	DD	DP	PT	PD	PP
TT	0	0.021	0.030	0.083	0.063	0.070	0.031	0.018	0.000
TD	0.020	0	0.030	0.099	0.061	0.084	0.010	0.003	0.020
TP	0.021	0.023	0	0.075	0.034	0.058	0.027	0.022	0.021
DT	0.070	0.087	0.070	0	0.068	0.024	0.095	0.084	0.069
DD	0.042	0.042	0.021	0.060	0	0.043	0.042	0.041	0.042
DP	0.052	0.068	0.049	0.020	0.047	0	0.075	0.065	0.052
PT	0.028	0.008	0.033	0.106	0.061	0.090	0	0.011	0.028
PD	0.017	0.003	0.030	0.097	0.062	0.082	0.013	0	0.017
PP	0.000	0.021	0.030	0.083	0.063	0.070	0.031	0.018	0

Pe baza valorilor din Tabelul 12-12:

÷ Cel mai improbabil este să se observe că un proces de evoluție urmând strategia TT produce evoluții în timp ce un proces de evoluție urmând strategia PP nu produce evoluții în aceleași momente relative și viceversa, probabilitatea acestui eveniment fiind de sub 1‰ ($Diff_{TT,PP}$ = $Diff_{PP,TT}$ = 0.000 în Tabelul 12-12; evidențiate ca atare și în Figura 12-3 - fiind aproape suprapuse - în grupul compact de cinci strategii din care mai fac parte TD, PD și PT);

÷ Cel mai probabil se poate observa că un proces de evoluție urmând strategia PT produce evoluții în timp ce un proces de evoluție urmând strategia DT nu produce evoluții în aceleași momente relative, probabilitatea acestui eveniment fiind de 10.6% ($Diff_{PT,DT}$ = 0.106 în Tabelul 12-12; evidențiate ca atare și în Figura 12-3 - PT cu cea mai mare valoare pozitivă a primei componente principale și DT cu cea mai mică valoare pozitivă a primei componente principale în lungul primei componente principale) în timp ce evenimentul contrar are probabilitatea de 9.5% ($Diff_{PT,DT}$ = 0.095 în Tabelul 12-12).

În Figura 12-4 s-au reprezentat primele porțiuni din densitatea de probabilitate (PDF, prima imagine) și funcția cumulativă de probabilitate (a doua imagine) a momentelor relative ale evoluției (intervalele de timpi relativi de la 0 la 0.002 pentru PDF și la 0.006 pentru CDF) iar în Figura 12-5 s-a reprezentat pentru un interval de timp relativ mai larg (de la 0 la 1) funcția cumulativă de probabilitate pentru cele două strategii selectate pentru comparație și anume DT și PT.

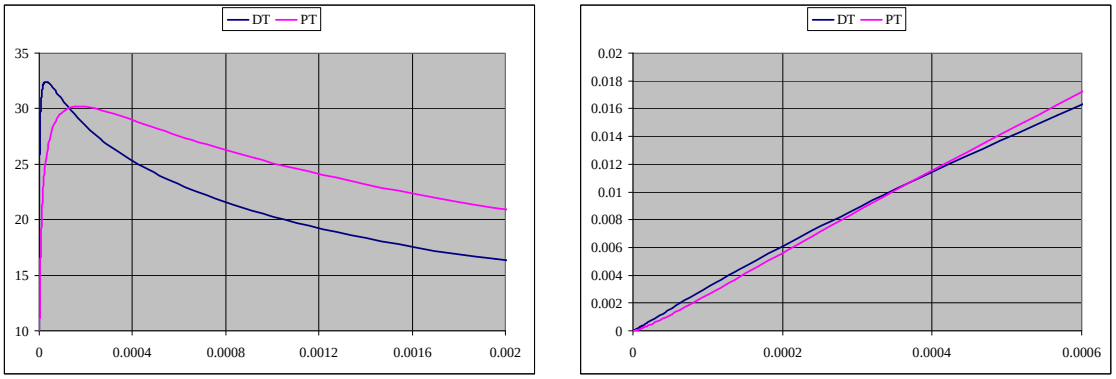


Fig. 12-4. Probabilități pentru momentele relative de evoluție în DT & PT - PDF & CDF
Probabilities for the relative moments of evolution in DT & PT - PDF & CDF

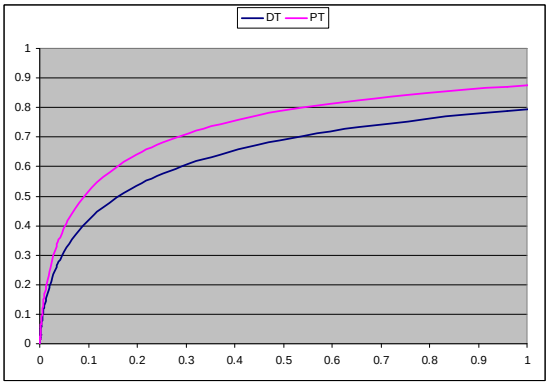


Fig. 12-5. Probabilități cumulate pentru momentele relative de evoluție în DT și PT
Cummulated probabilities for relative moments of evolution in DT & PT

Diferența între aparițiile momentelor relative de evoluție la strategiile DT şi PT evidențiată pe baza datelor din Tabelul 12-12 se explică din inspecția graficelor legilor de distribuție (Figurile 12-4 şi 12-5 de mai sus).

De ce există diferența de probabilitate de 10.6% ($Diff_{PT,DT}$ = 0.106 în Tabelul 12-12) în favoarea lui PT față de DT la evenimentul "un proces de evoluție urmând strategia PT produce evoluții în timp ce un proces de evoluție urmând strategia DT nu produce evoluții în aceleaşi momente relative" se explică pe baza locației γ: γ_{PT} = 15.347 (Tabelul 12-11) în timp ce γ_{DT} = 11.640 (Tabelul 12-11) astfel încât domeniul de valori probabile ale lui PT este mai larg (cu de la $e^{11.640}$ până la $e^{15.347}$, în care doar în PT mai apar momente relative de evoluție de această magnitudine). În acelaşi timp evenimentul contrar are o probabilitate de 9.5% ($Diff_{PT,DT}$ = 0.095 în Tabelul 12-12), ceea ce se explică pe baza reprezentării din Figura 12-4-PDF - şi anume strategia DT este mai productivă în evoluție la momente relative de timp mai mici.

Există un moment relativ de timp (şi acesta este 0.00038) până la care cele în cele două strategii se produc un număr egal de evoluții (Figura 12-4-CDF). Din acest moment relativ de timp este mai probabil să se fi produs mai multe evoluții în strategia PT decât în strategia DT (Figura 12-4-CDF şi Figura 12-5), strategia PT producând astfel evoluții mai încet decât strategia DT.

Ce distribuție urmează numărul de evoluții?

Într-o lucrare anterioară (Jäntschi & alții, 2010-Evolution) având la dispoziție un număr redus de repetiții (30) şi de generații (1000) nu a putut fi pusă în evidență respingerea semnificativ statistică a legii de distribuție Poisson pentru numărul de evoluții.

Pentru a da un răspuns mai complect la problema distribuției numărului de evoluții s-a folosit întreg volumul de date din cele 46 de repetiții ale evoluției, fiecare pe parcursul a 20000 de generații.

Datele au fost grupate pe strategii şi apoi toate la un loc, obținând astfel 10 eşantioane reprezentative pentru numărul de evoluții. Volumul fiecărui eşantion din primele 9 este de 46, numărul de execuții independente, iar pentru cel de-al zecelea 414 (9·46) în timp ce o valoare din eşantion reprezintă numărul de evoluții observate într-o execuție.

Pentru cele 10 eşantioane au fost folosite statisticile K-S, A-D şi C-S pentru a măsura agrementul între model şi observație având la dispoziție o serie de peste 60 de alternative de model de lege de distribuție. A fost folosită aplicația EasyFit pentru a obține estimările parametrilor distribuțiilor şi a da măsurile agrementelor.

S-au extras din aplicația EasyFit pentru fiecare eşantion (în număr de 10) şi fiecare statistică (în număr de 3) lista acelor legi de distribuție pentru care ipoteza de agrement între observație şi

model nu a putut fi respinsă la riscul de a fi în eroare de 1%. Un număr de 60 de legi de distribuție au apărut cel puțin odată în aceast clasament, iar un număr de 27 dintre acestea au exact 30 de apariții, şi anume pentru fiecare eşantion (din cele 10) şi fiecare statistică (din cele 3) nu a putut fi respinsă ipoteza agrementului observației cu modelul distribuției.

Pentru aceste 27 de legi, un nou clasament a fost întocmit, şi anume conținând rangul (începând cu rang 0 pentru prima) statisticii obținute în şirul statisticii pentru toate alternativele de distribuție ale unui eşantion.

Rezultatul obținut este sintetizat în Tabelul 12-13.

O primă observație care referă datele din Tabelul 12-13 este că toate legile de distribuție sunt legi continue. Toate legile discrete (Uniformă, Bernoulli, Binomială, Poisson, Negativ binomială, Geometrică, Logaritmică, Hipergeometrică au fost respinse).

Tabelul 12-13

Ragurile statisticilor în alternativele de distribuție (*min e cel mai bun*)
Ranks of the statistics between the alternatives of distribution (mini s the best)

Nr	Alternativă de distribuție (*Distribution alternative*)	Rang (*Rank*)*
1	Gen. Extreme Value	284
2	Dagum	422
3	Burr	441
4	Log-Logistic (3P)	452
5	Pearson 5 (3P)	453
6	Pearson 6 (4P)	455
7	Fatigue Life (3P)	465
8	Lognormal (3P)	470
9	Log-Pearson 3	481
10	Gamma (3P)	501
11	Weibull (3P)	510
12	Gen. Logistic	587
13	Gamma	600
14	Error	606
15	Gen. Gamma	614
16	Normal	620
17	Pearson 6	620
18	Frechet (3P)	642
19	Nakagami	653
20	Inv. Gaussian (3P)	655
21	Weibull	699
22	Chi-Squared (2P)	706
23	Lognormal	773
24	Erlang (3P)	783
25	Rice	834
26	Logistic	842
27	Gumbel Max	904
Rang(*Rank*)*: începând cu rang = 0 pentru prima (beginning with rank 0)		

O a doua observație asupra rezultatelor din Tabelul 12-13 este că 10 din primele 11 în ordinea rangurilor sunt distribuții definite de 3 parametrii.

În final, ultima observație referă distribuția generalizată a valorilor extreme (GEV, sau Fisher-Tippett), cea care este şi distribuția pentru coeficientul de determinare al unei generații: ea obține cele mai bune scoruri după rangul celor trei statistici folosite, rang care este cu 32% mai mic

decât rangul distribuției imediat următoare, în timp ce diferențele între cele care îi urmează nu depășesc 5%.

Ipoteza naturală cu privire la legea de distribuție a numărului de evoluții care se petrec într-o execuție a algoritmului genetic în evoluția sa către modelul MDF SAR optim este deci legea Fisher-Tippett. Măsurile agrementului între numărul de evoluții observat și modelul dat de distribuția Fisher-Tippett prin intermediul celor trei statistici (Pearson-Fisher Chi-Square, Anderson-Darling & Kolmogorov-Smirnov) sunt sintetizate în Tabelul 12-14.

Tabelul 12-14

Agrementul între Fisher-Tippett și numărul de evoluții ale unei execuții a algoritmului genetic

The agreement between Fisher-Tippett and the number of evolutions from execution of the genetic algorithm

Strategie (*Strategy*)	Obs	K-S	P_{K-S}	A-D	P_{A-D}	C-S/df	P_{C-S}
TT	46	0.0924	0.7931	0.4183	0.6028	5.17/5	0.3956
TD	46	0.1199	0.4859	0.5976	0.4877	3.57/4	0.4671
TP	46	0.0454	0.9999	0.0818	0.8972	0.96/5	0.9661
DT	46	0.0632	0.9873	0.2303	0.7527	1.27/5	0.9381
DD	46	0.0615	0.9906	0.215	0.7665	0.72/5	0.9816
DP	46	0.0954	0.7612	0.2766	0.7127	3.76/4	0.4389
PT	46	0.0712	0.9608	0.2052	0.7754	4.23/5	0.5171
PD	46	0.0634	0.9869	0.1693	0.8090	0.99/5	0.9632
PP	46	0.0665	0.9787	0.2428	0.7417	0.69/5	0.9835
Toate (*All*)	414	0.0342	0.7066	0.307	0.6875	7.14/8	0.5218

Agrementul pe care îl realizează distribuția Fisher-Tippett cu observațiile experimentale ale numărului de evoluții este de-a dreptul remarcabil (Tabelul 12-14). Nu numai că nu se semnalează respingeri semnificativ statistice la nici un risc de a fi în eroare uzual între 1% și 20%, ci mai mult, agrementul între model și date este în medie de 86.51% în acord cu statistica Kolmogorov-Smirnov, de 72.33% în acord cu statistica Anderson-Darling și de 71.73% în acord cu statistica Chi-Square. În cea mai mică măsură agrementul a fost înregistrat de statistica Chi-Square (39.56%) pentru strategia TT și de statisticile Kolmogorov-Smirnov (48.59%) și Anderson-Darling (48.77%) pentru strategia TD. Tabelul 12-15 redă statisticile comune cu privire la legile de distribuție obținute.

Tabelul 12-15

Statistici ale distribuțiilor Fisher-Tippett ale numărului de evoluții către optim

Statistics of Fisher-Tippett distributions of number of evolutions to the optimum

Strategia	Distribuția F-T(α; β; γ)	μ	$\hat{\mu}$	$\tilde{\mu}$	σ	γ_1	γ_2
TT	F-T(-0.0771; 8.0028; 26.929)	31.0	28	29.8	9.38	0.739	0.849
TD	F-T(-0.19367; 8.9378; 28.367)	32.1	30	31.5	9.44	0.276	-0.095
TP	F-T(0.04267; 8.7648; 24.208)	29.7	24	27.4	11.93	-1.420	3.975
DT	F-T(-0.0309; 7.0811; 18.775)	22.7	19	21.4	8.74	0.966	1.635
DD	F-T(-0.30349; 9.3813; 21.38)	24.6	25	24.6	9.26	-0.079	-0.289
DP	F-T(-0.27344; 8.0192; 16.622)	19.5	19	19.4	8.05	0.013	-0.280
PT	F-T(-0.15998; 8.6245; 29.02)	32.8	31	32.1	9.35	0.398	0.074
PD	F-T(-0.12837; 9.3279; 28.721)	33.0	30	32.1	10.39	0.520	0.299
PP	F-T(-0.24824; 9.8865; 26.7)	30.4	29	30.2	10.07	0.093	-0.249
Toate	F-T(-0.16044; 9.6882; 24.161)	28.4	26	27.6	10.50	0.396	0.072
μ: Media; $\hat{\mu}$: Moda; $\tilde{\mu}$: Mediana; σ: Deviația standard; γ_1: Asimetria; γ_2: Excesul de boltire							

Rezultatele din Tabelul 12-15 evidențiază că legile de distribuție sunt foarte apropiate una de cealaltă. În fapt nu a fost respinsă nici ipoteza că toate împreună au o singură lege de distribuție şi strategia nu creează populații distincte în numărul de evoluții. Se poate remarca însă că valori la modă sub moda globală (26) au strategiile de selecție deterministă (DP: 19; DT: 19; DD: 25) şi selecția în turnir cuplată cu supraviețuirea proporțională (TP: 24) în timp ce valori la modă superioare modei globale (26) au strategiile de selecție proporțională (PT: 31; PD: 30; PP:29) şi în turnir (TD: 30; TT: 28) - cu o excepție, cea de mai sus (TP: 24).

Tot rezultatele din Tabelul 12-15 evidențiază că o singură lege de distribuție este a valorilor extreme de tipul II (Fréchet) - TP ($\alpha > 0$), toate celelalte, inclusiv distribuția globală fiind a valorilor extreme de tipul III (Weibull). Aceeaşi strategie (TP) produce şi cel mai mare exces de boltire (3.975) fiind aproape dublu decât în cazul oricărei alte strategii. Dacă se admite că valoarea asimetriei pentru strategia DD (-0.079) este aproape nulă, atunci strategia TP rămâne singura cu asimetrie negativă (-1.42), fiind în acelaşi timp distribuția cu cea mai mare deviație standard (11.93).

Un calcul simplu asupra deviației standard date în Tabelul 12-5 în ipoteza de independență între eşantioanele extrase din strategii diferite ($\sigma_{\Sigma}^2 = (\sigma_{TT}^2 + ... + \sigma_{PP}^2)/9$) ne permite să separăm varianța totală ($\sigma_T^2 = 10.5^2$) în varianță în interiorul strategiilor ($\sigma_{\Sigma}^2 = 9.68^2$) şi varianță între strategii (4.07^2). Acest din urmă rezultat pune în evidență o măsură în care alegerea unei strategii influențează evoluția, contribuția cantității variației între strategii fiind importantă în valoarea variației totale.

Concluzii şi recomandări

Prezenta lucrare demonstrează că algoritmii genetici, ca tehnici adaptive de căutare euristică, bazate pe principiile geneticii şi selecției naturale, pot fi eficient utilizați în simularea procesului biologic al evoluției şi în cel de ameliorare a plantelor.

Modelele informatice elaborate prin intermediul algoritmilor genetici, emulează modelele biologice evoluționiste, asigurând rezolvarea unor probleme concrete de optimizare sau căutare în experiențele de genetica şi ameliorarea plantelor. Prin intermediul elementelor individuale, reprezentate sub forma şirurilor binare, şi a operatorilor de natură biologică definiți asupra populației şi a modelului molecular, algoritmii genetici manipulează cele mai promițătoare şiruri, evaluate conform unei funcții obiectiv, căutând soluții mai bune, tinzând în esență spre cea "optimă", dar acceptând în final una apropiată de optim.

În lucrare, eficiența aplicării algoritmilor genetici în optimizarea relației structură-activitate în seria de compuşi PCB, care prezintă potență biologică distructivă asupra mediului vegetal şi animal, a fost probată într-un experiment de evoluție, folosind diferite strategii pentru estimarea efectului procesului de selecție şi supraviețuire asupra indivizilor în cadrul populațiilor.

S-a constatat că eşantionul de genotipuri supus evoluției tinde relativ rapid către optim, probabilitățile din funcțiile cumulative de distribuție obținute asigurând obținerea a 99% din optim în 1000 de generații (analizându-se mai puțin de $2 \cdot 10^{-11}$ din numărul de regresii posibile în întreaga populație) pentru strategia TD în 55% din cazuri, PD - 67%, PP - 68%, TP - 73%, PT - 78%, TT - 80%, DD - 87%, DP - 95% şi DT în 97% din cazuri.

(***recomandare***) Asigurarea condițiilor optime pentru plantele cultivate, cu privire la ansamblul cerințelor tehnologice, incluzând necesarul de nutrienți şi, în particular, a necesarului de apă, este esențială pentru reuşita culturilor horticole. Acestea pot contribui semnificativ la productivitatea şi eficiența economică a culturilor horticole, precum şi la calitatea materialului biologic şi, în mod deosebit, a calității fructelor. Controlul optimal al proceselor fiziologice de creştere, dezvoltare, fructificare etc. în asociere cu genotipul (cultivarul), necesarul de apă şi nutrienți pentru culturile de câmp sau seră (spații protejate), în asociere cu factorii tehnologici, de cultură (ex. suplimentul de apă furnizat prin sistemele de irigare) poate fi asistat de calculator folosind algoritmii genetici, care, după cum s-a demonstrat în teză, sunt capabili să ofere soluții optimale la problemele complexe de evoluție în condiții specifice de mediu, în timp scurt.

Experiența acumulată în dezvoltarea algoritmului genetic şi apoi a programului evolutiv au permis desprinderea următoarei serii de concluzii şi recomandări:

÷ Modelarea şi apoi simularea unei evoluții impune parcurgerea următorilor paşi:
  - definirea materialului genetic (cod genetic; genotipuri);

- definirea cultivarului (spațiul de căutare; fenotipuri);
- proiectarea şi implementarea operatorilor genetici:
  - mutație;
  - încrucişare;
  - selecție;
  - supraviețuire;
- definirea obiectivului evoluției şi condiției de stop;
- proiectarea şi implementarea algoritmului genetic;
- definirea unei serii de parametrii de monitorizare a evoluției.

÷ În general studiul proceselor de evoluție este un proces extensiv consumator de spațiu de calcul (memorie), spațiu de stocare (disc fizic) şi timp de lucru (viteză de calcul), ceea ce face ca aplicațiile care se doresc destinate în acest scop trebuie ca în faza de transpunere a algoritmului (genetic) în program (evolutiv) să țină seama de următoarele în alegerea platformei (mediului) de implementare (programare):

- posibilitate de conectare la baze de date pentru transferul datelor şi/sau al rezultatelor; MySQL s-a dovedit a fi un bun sistem de gestiune a datelor de intrare şi având o bună reprezentare în ceea ce priveşte suportul de conectivitate pentru diferite limbaje de programare;
- viteză mare de procesare a codului generat (mediul de programare trebuie să ofere posibilități de optimizare a codului generat, eventual de generare de cod maşină dedicat pentru sistemul de operare, tipul de calculator şi tipul de procesor folosit); FreePascal s-a dovedit a fi un mediu propice de programare din acest punct de vedere;
- capabilitatea de a adresa spații mari de memorie internă; programul executabil realizat a necesitat adresarea a până la 2G de memorie internă în execuție; FreePascal s-a dovedit a fi un mediu propice de programare din acest punct de vedere;
- capacitate mare de abstractizare a tipurilor de date oferite programatorului, întrucât complexitatea de programare a unui algoritm genetic care să realizeze operații complexe cu volume mari de date este maximă; algoritmul şi programul realizate reprezintă o situație atipică în algoritmică şi programare şi anume în general (tipic) problemele informatice fie necesită operații simple cu volume mari de date - aplicații economice şi de gestiune - fie necesită operații complexe cu volume mici de date - aplicații ştiințifice şi de cercetare; din acest punct de vedere FreePascal s-a dovedit doar soluție relativ optimală - nota bene: cu cât mediul de programare oferă mai multe libertăți de programare cu atât viteza de procesare a codului generat de limbaj scade, deoarece necesită instrucțiuni suplimentare de tratare (sau analiză) a contextului (sau situației);

- obiectivul de generalitate şi portabilitate maximă a aplicației realizate se urmăreşte prin:
  - flexibilitate în definirea codului genetic; soluția finală rezultată pentru aceasta este specificarea topologiei genetice într-un fişier separat de configurare;
  - flexibilitate în definirea parametrilor de conectivitate la baza de date de intrare: din nou soluția finală rezultată pentru aceasta este specificarea acestora într-un fişier separat de configurare;
  - flexibilitate în definirea parametrilor de evoluție (modalități de selecție, supraviețuire, probabilități de mutație, mărime cultivar, număr de încrucişări per evoluție, ş.a.m.d.); din nou soluția finală rezultată pentru aceasta este specificarea acestora într-un fişier separat de configurare;
  - generarea de cod dedicat pentru mai multe tipuri de calculatoare, procesoare şi sisteme de operare permite maximizarea resurselor de calcul, şi execuția aplicației în paralel pe mai multe calculatoare într-un regim de eficiență maximă, cât şi utilizarea la maxim a facilităților oferite de producatorul calculatorului - cod optimizat pentru maşini de calcul *dual core*, ş.a.m.d.

Definirea unui design de experiment corect este esențială în obținerea, prelucrarea şi interpretarea datelor experimentale; designul de experiment realizat în prezenta lucrare, a permis studiul evoluției mai multor parametri ce definesc materialul genetic şi cultivarul, cu ajutorul algoritmului genetic; analiza evoluției a cuprins atât evoluția în ansamblu, cât şi evoluția bazată pe diferite strategii de selecție şi supraviețuire.

(***recomandare***) Procesele de ameliorare care, prin selecții repetate, au ca scop amplificarea unui caracter, cum este de exemplu mărimea fructului la o specie horticolă oarecare, trebuie realizate într-un design experimental urmărit cu atenție pe întreg parcursul proceselor de încrucişare repetată. Conform datelor obținute, asemenea caractere nu urmează pe parcursul selecției o lege de distribuție normală; astfel, media care se obține din observațiile pe întreg eşantionul unei generații supuse înmulțirii repetate reprezintă o statistică suficientă pentru ilustrarea evoluției. În explicarea unui asemenea fenomen, selecția naturală sau cea artificială, ultima cu scop de ameliorare, este necesară o analiză mai detaliată, complexă, în care variabilitatea şi heritabilitatea caracteristicilor specifice trebuie abordată diferențiat.

S-au obținut concluzii importante cu caracter fundamental privind influența metodei de selecție şi metodei de supraviețuire asupra evoluției. Analiza compoziția materialului genetic în cultivar a dus la următoarele concluzii:

÷ Selecția deterministă face ca indiferent de strategia de supraviețuire numărul total de genotipuri distincte reprezentate în cultivar să scadă semnificativ statistic;

÷ Supraviețuirea deterministă asociată cu selecția turnir sau proporțională mărește semnificativ statistic (fapt evidențiat de toate cele trei statistici ale agrementului între observație și model folosite) numărul total de genotipuri reprezentate în cultivar;

÷ Supraviețuirea deterministă îmbogățește semnificativ statistic grupul celor mai frecvente genotipuri din generațiile ce produc evoluție în timp ce selecția deterministă sărăcește semnificativ statistic numărul total al genotipurilor din generațiile care produc evoluție;

÷ Fiecare strategie de selecție definește câte o populație genotipică în generațiile care produc evoluție; argumentul este că oricare ar fi statistica urmărită al numărului total (*Total*) de genotipuri (*Num*, *Apar* sau *Part*), fiecare dintre variațele produse de variația metodei de supraviețuire în grupul în care metoda de selecție este aceeași sunt semnificativ statistic mai mici decât varianța totală. Nu aceeași concluzie se poate trage cu privire la strategia de supraviețuire, pentru care se produce segregare populațională doar pentru supraviețuirea deterministă, care creează o populație cu un număr mediu de genotipuri semnificativ statistic mai mare decât supraviețuirea proporțională și respectiv în turnir;

Urmărind generațiile care produc evoluție (prin îmbunătățirea scorului obiectiv) pe baza numărului și poziției acestor evoluții pe axa temporală a execuției s-a observat că:

÷ ***Media generațiilor care produc evoluții*** este o măsură a ***cât de timpuriu / târziu se produc evoluțiile***. Selecția deterministă (și cu atât mai mult însoțită de supraviețuirea deterministă) produce cele mai târzii evoluții, în timp ce selecția turnir (și cu atât mai mult însoțită de supraviețuirea proporțională sau turnir) produce cele mai timpurii evoluții.

÷ ***Numărul de evoluții dintr-un număr dat de generații*** este o măsură a ***cât de frecvent se produc evoluțiile***. Selecția proporțională (și cu atât mai mult însoțită de supraviețuirea deterministă sau turnir) produce cele mai dese evoluții, în timp ce selecția deterministă (și cu atât mai mult însoțită de supraviețuirea proporțională sau turnir) produce cele mai rare evoluții.

÷ ***Variabilitatea poziției evoluțiilor pe parcursul generațiilor*** este o măsură a ***cât de compact / dispers se produc evoluțiile***. Cea mai bună compactitate de evoluție o au selecția proporțională asociată cu supraviețuirea proporțională și selecția în turnir asociată cu supraviețuirea în turnir. Cea mai mare dispersie a evoluției se observă la selecția proporțională asociată cu supraviețuirea deterministă urmată îndeaproape de selecția deterministă asociată cu supraviețuirea proporțională.

÷ ***Variabilitatea numărului evoluțiilor*** este o măsură a ***cât de predictibil / impredictibil se produc evoluțiile***. Patru asocieri de selecție și supraviețuire au rezultate similare sub aspectul predictibilității evoluției: (selecție deterministic, supraviețuire deterministic), (selecție în turnir, supraviețuire deterministic), (selecție în turnir, supraviețuire în turnir) și (selecție proporțională, supraviețuire în turnir). Predictibilități extreme ale evoluției obțin selecția deterministă asociată

cu supraviețuirea proporțională având cea mai mare predictibilitate (cea mai mică variabilitate) şi selecția în turnir asociată cu supraviețuirea proporțională având cea mai mică predictibilitate (cea mai mare variabilitate) de evoluție.

Numărul de genotipuri reprezentate în cultivar a fost subiectul măsurării variabilității genotipice a materialului genetic prezent în cultivar; au fost transformate frecvențele de apariție la proporții relative în asocierile dintre metoda de selecție şi cea de supraviețuire; astfel, ***proporția relativă a numărului de genotipuri reprezentate în cultivar în generațiile ce produc evoluție*** este o măsură a ***diversității pe parcursul evoluției***; s-a observat că:

÷ Evoluția bazată pe selecția deterministă însoțită de o supraviețuire deterministă induce o înmulțire a celor mai adaptate genotipuri.

÷ Evoluția bazată pe selecția deterministă însoțită de o supraviețuire proporțională sau în turnir induce o înmulțire a numărului de genotipuri diferite.

÷ Selecția proporțională însoțită de supraviețuirea proporțională face ca în evoluție genotipurile să fie prezente în mod uniform (distribuție uniformă) - fiecare având o aceeaşi şansă de apariție.

Ordinul de mărime al vitezei de evoluție (ca raport dintre numărul de evoluții şi numărul de generații) a fost de 1.37‰ adică în medie 1.37 evoluții la 1000 de generații.

Analiza independenței dintre numărul mediu (*avg_obs*) de genotipuri viabile (calculat din împărțirea la numărul de generații ce au produs evoluție observate *num_obs* a numărului de genotipuri viabile observate în aceste generații *sum_obs*) şi metoda de supraviețuire a dus la următoarele concluzii:

÷ Cu un risc de a fi în eroare de 5% DD, PD, şi TD provin din populații diferite.

÷ Nu poate fi respinsă ipoteza că DT şi DP provin din populații identice.

÷ Nu poate fi respinsă ipoteza că TT şi TP provin din populații identice.

÷ Nu poate fi respinsă ipoteza că PP, PT şi TT provin din populații identice.

÷ Cu un risc de a fi în eroare de 5% se respinge ipoteza că PP, PT, TP şi TT ar proveni din populații identice.

÷ Selecția deterministă însoțită de supraviețuirea deterministă (DD) au ca efect menținerea unui număr mare de genotipuri viabile în cultivar, a cărei lege de distribuție populațională diferă semnificativ statistic (cu un risc de 5% de a fi în eroare) de celelalte opțiuni de selecție şi supraviețuire.

÷ Caracterizate de variabilitate ridicată, selecția deterministă însoțită de supraviețuirea în turnir (DT) sau proporțională (DP) produc o distribuție populațională ce diferă semnificativ statistic (cu un risc de 5% de a fi în eroare) de populațiile produse de celelalte posibilități de selecție şi supraviețuire.

÷ Supraviețuirea deterministă susținută de o selecție în turnir (TD) produce o distribuție populațională ce diferă semnificativ statistic (cu un risc de 5% de a fi în eroare) de populațiile produse de celelalte posibilități de selecție şi supraviețuire ale cărei caracteristici însă, de valoare medie şi variabilitate sunt apropiate de cele ale populației produse de selecția deterministă însoțită de supraviețuirea deterministă (DD) pe de o parte şi de selecția deterministă susținută de supraviețuirea proporțională (DP) de cealaltă parte.

÷ Supraviețuirea deterministă susținută de o selecție proporțională (PD) produce o distribuție populațională ce diferă semnificativ statistic (cu un risc de 5% de a fi în eroare) de populațiile produse de celelalte posibilități de selecție şi supraviețuire ale cărei caracteristici însă, de valoare medie şi variabilitate sunt apropiate de cele ale populației produse de selecția deterministă însoțită de supraviețuirea în turnir (DT).

÷ Selecțiile şi supraviețuirile care nu sunt nici una deterministe formează un grup relativ compact şi produc populații relativ apropiate în ceea ce priveşte numărul mediu de genotipuri viabile din cultivar pe parcursul evoluției.

Analiza statistică a arătat că interpretarea populațională a numărului mediu de fenotipuri din cultivar în generațiile ce produc evoluție este mult mai complexă decât interpretarea populațională a numărului mediu de genotipuri în aceleaşi evoluții. Următoarele concluzii au fost obținute:

÷ Selecția în turnir însoțită de o supraviețuire proporțională (TP) produce o populație distinctă de fenotipuri în cultivar sub aspectul numărului acestora, care în valoare medie se situează aproape în toate momentele evoluției deasupra mediei produse de metodele de selecție şi supraviețuire rămase (PP, PT, PD, TT, TD, DD) şi numărul mediu de fenotipuri ale populației produse de selecția în turnir însoțită de o supraviețuire proporțională (TP) creşte aproape peste tot cu aproape un fenotip față de restul metodelor (medie de 18.94 pentru TP; medie de 18.18 pentru PP+PT+PD+TT+TD+DD), şi cu peste un fenotip decât media generală (17.59), având tendință crescătoare spre 19 fenotipuri pe parcursul evoluției. Numărul mediu de fenotipuri produs de selecția deterministă însoțită de o supraviețuire proporțională (DP) sau în turnir (DT), creează o populație distinctă de fenotipuri în cultivar sub aspectul numărului acestora, care în valoare medie se situează în toate momentele evoluției sub media produsă de metodele de selecție şi supraviețuire rămase (PP, PT, PD, TT, TD, DD) şi numărul mediu de fenotipuri ale populației produse de selecția deterministă însoțită de o supraviețuire proporțională (DP) sau în turnir (DT) produce scăderea cu cel puțin două fenotipuri aproape peste tot față de restul metodelor (medie de 15.15 pentru DP+DT; medie de 18.18 pentru PP+PT+PD+TT+TD+DD) şi cu aproape două fenotipuri față de media generală (17.59), având o tendință descrescătoare spre 15 fenotipuri pe parcursul evoluției.

÷ S-a obținut soluția cea mai probabilă de partiție a numărului mediu de fenotipuri în raport cu strategiile de selecție și supraviețuire prin diferențe semnificative statistic: {(DD, PP, PT, TT, TD), (DT, DP), PD, TP}.

Analiza statistică a arătat că interpretarea populațională a numărului mediu de asocieri de fenotipuri în cultivar în generațiile ce produc evoluție a permis obținerea cea mai probabile segregări a populațiilor de asocieri fenotipice care se produce după schema {(PD, DD), (TP, TT), (DP, DT), (PP, PT), TD} - unde ordinea în această listă este dată de probabilitatea de asociere.

(***recomandare***) În procesele de ameliorare a plantelor horticole, care au ca scop general îmbunătățirea productivității, a calității producției precum şi creşterea rezistenței plantelor la atacul diferitelor boli şi dăunători, utilizarea algoritmilor genetici poate reduce considerabil durata procesului de obținere a unor genotipuri superioare.

Analiza de distribuție a observabilelor continue a permis evidențierea unui rezultat fundamental, şi anume: acceptarea ipotezei că datele ar putea proveni dintr-o lege de distribuție (se respinge ipoteza că cele două - cea observată şi cea teoretică - legi de distribuție sunt diferite cu un risc de a fi în eroare de 10%, 5%, 2.5% sau 1%) nu exclude posibilitatea ca aceleaşi date să provină din altă lege de distribuție (teoretică). Eşantionarea întâmplătoare din legea de distribuție uniformă (care stă la baza construcției de eşantioane întâmplătoare din orice lege de distribuție cunoscută) a evidențiat dificultatea extremă cu care se confruntă motoarele deterministe de generare a numerelor aleatoare; practic orice generare de numere aleatoare folosind o maşină de calcul este deterministă şi induce un bias considerabil în ipoteza de întâmplare cu privire la numerele generate; raportul între statistica Andreson-Darling ajustată şi valoarea sa la pragul de semnificație de 10% a arătat pentru eşantionul uniform generat că chiar dacă cea mai probabilă lege de distribuție este distribuția uniformă (cu 2.3), ea este urmată îndeaproape de alte două legi de distribuție: distribuția valorilor extreme (cu 2.2) şi Pareto generalizată (cu 2.1).

Analiza legii de distribuție a coeficientului de determinare folosind o serie de alternative (DE - dublu exponențială; LG - logistică; LN - log-normală; NO - normală; UN - uniformă (0,1); EX - exponențială; GU - valori extreme de tip I) a evidențiat că legea valorilor extreme de tip I (GU) este legea de distribuție cea mai frecventă a coeficientului de determinare într-o generație a algoritmului genetic rulat cu diferite metode de selecție şi supraviețuire. Distribuția valorilor extreme de tip I (sau distribuția Gumbel) este legea limită de distribuție a minimumului unui număr mare variabile întâmplătoare distribuite identic. Dacă variabilele întâmplătoare sunt mărginite (cum este cazul coeficientului de determinare, mărginit de 0 şi 1) distribuția valorilor extreme devine (atinge cazul particular al) distribuției Weibul. Dacă numărul de variabile întâmplătoare distribuite identic devine unitar, atunci distribuția valorilor extreme devine (atinge cazul particular al) distribuției normale Gauss.

O întrebare fundamentală a căpătat răspuns, şi anume: *Valoarea medie a coeficientului de determinare pentru fiecare generație (ca valoare medie din cele 46 de execuții independente) este o un **estimator nedeplasat** pentru acest număr? Varianța sau abaterea standard sunt de asemenea **statistici suficiente**?* - răspunsul este NU, iar motivul este faptul că aproape pentru nici o pereche de metode selecție - supraviețuire nu s-a putut accepta ipoteza distribuției normale a valorilor, care ar fi acceptat valoarea medie din eşantionul de 46 de valori drept estimator ***nedeplasat*** al mediei populației tuturor execuțiilor independente iar varianța drept ***statistică suficientă***.

Valoarea maximă observată a coeficientului de determinare în cele 46 de execuții independente ale algoritmului genetic pe parcursul a 20000 de generații a fost de 0.907. Această valoare serveşte în comparația performanțelor algoritmului genetic în ipoteza execuției pe timp nelimitat dar în cadrul impus de o anumită valoare a funcției obiectiv ($r^2 = 0.907$).

Analiza evoluției tendinței centrale a arătat că există o foarte strânsă dependență între numărul generației şi tendința centrală a coeficientului de determinare obținut. Această dependență sugerează posibilitatea obținerii unei ecuații care să exprime generația în care se obține o anumită valoare a coeficientului de determinare - mai exact să se facă estimarea şi posibil predicția evoluției algoritmului genetic în funcție de obiectivul de optimizare propus. Pentru a atinge acest scop, expresia de dependență trebuie inversată şi anume să se exprime numărul de generații în funcție de determinarea obținută. Logaritmând şi analizând rezultatele s-a obținut că:

÷ Pentru medie (MEAN) există cel puțin cinci grupuri semnificativ statistic distincte:
  - Un grup de trei cazuri, şi anume selecția deterministă însoțită de supraviețuirea proporțională (DP), selecția deterministă însoțită de supraviețuirea în turnir (DT) şi selecția în turnir însoțită de supraviețuirea în turnir (TT) au retardări care le disting semnificativ statistic față de medie şi unul față de celălalt în ceea ce priveşte ordinul de mărime al numărului de generații necesar în medie pentru atingerea idealului optimizării (coeficientul de determinare maxim posibil).
  - Un grup de cinci cazuri - TP (selecție în turnir, supraviețuire proporțională), PD (selecție proporțională, supraviețuire deterministă), DD (selecție deterministă, supraviețuire deterministă), PT (selecție proporțională, supraviețuire în turnir), PP (selecție proporțională, supraviețuire proporțională) - ating în medie idealul optimizării într-un ordin de mărime al numărului de generații inferior mediei fiind foarte aproape situate una față de cealaltă şi față de medie.
  - Un caz se distinge semnificativ statistic prin precocitate în atingerea idealului optimizării şi anume selecția în turnir urmată de supraviețuire deterministă (TD).

Analiza coeficientului de determinare relevă pe lângă precocitate vs. retardare în atingerea idealului optimizării existența a încă unei măsuri importantă ce caracterizează evoluția algoritmului

genetic în raport cu asocierea de selecție și supraviețuire: puterea cu care se produce evoluția către valoarea maximă a optimului observat.

În cazul selecției proporționale se observă că pe toată durata evoluției supraviețuirea în turnir (PT) produce cei mai retardați descendenți în raport cu supraviețuirea proporțională (PP) și supraviețuirea deterministă (PD) în timp ce supraviețuirea deterministă (PD) produce cei mai precoce descendenți și decalajul de precocitate între supraviețuirea deterministă (PD) și supraviețuirea proporțională (PP) se micșorează odată cu trecerea generațiilor astfel încât în generația 100000 ajung să aibă precocitate foarte apropiată. Se poate spune astfel că în selecția proporțională o supraviețuire în turnir (PT) nu este favorabilă evoluției.

În cazul selecției în turnir se observă că precocitatea supraviețuirii deterministe (TD) crește odată cu trecerea generațiilor și dacă la începutul evoluției produce cei mai retardați descendenți (generația 10), la scurt timp după (generația 1000) ajunge să producă cele mai valoroase progenituri. Poziția relativă de medie (mediocritate) a supraviețuirii proporționale (TP) nu se schimbă pe durata evoluției în timp ce precocitatea supraviețuirii în turnir (TT) se alterează odată cu evoluția, ajungând din cea mai precoce la începutul evoluției în cea mai retardată către finalul evoluției (către completa adaptare la mediul menținut constant pe durata evoluției). Se poate spune astfel că în cazul selecției în turnir supraviețuirea proporțională (TP) asigură o evoluție mai rapidă în primele generații dar care nu este durabilă în timp ce supraviețuirea deterministă (TD), deși timidă în primele generații ajunge destul de repede să furnizeze cheia succesului în evoluție.

În cazul selecției deterministe se observă că precocitatea supraviețuirii deterministe (DD) crește odată cu trecerea generațiilor și dacă la începutul evoluției produce cei mai retardați descendenți și cea mai mică variabilitate (generația 10), evoluția face ca atât variabilitatea cât și precocitatea să crească rapid și să devină cea mai evoluționistă (generația 1000), în timp ce supraviețuirea în turnir (DT) pierde teren pe parcursul generațiilor devenind (generația 100000) cea mai puțin variabilă iar supraviețuirea proporțională (DP) devenind cea mai retardată (generația 100000). Se poate spune astfel că în cazul selecției deterministe supraviețuirea proporțională (TP) asigură o evoluție mai rapidă în primele generații dar care nu este durabilă în timp îndelungat în timp ce supraviețuirea deterministă (DD), deși timidă în primele generații ajunge destul de repede să furnizeze cheia succesului în evoluție.

În cazul supraviețuirii proporționale se observă că selecția deterministă (DP) produce sistematic cei mai retardați descendenți, în timp ce selecția proporțională (PP) cu o precocitate mai redusă a descendenților la începutul evoluției devine în timp cea mai precoce (generația 100000), fără însă ca diferența între selecția proporțională (PP) și selecția în turnir (TP) să fie semnificativă, existând un moment în evoluție (generația 1000) când evoluția în cele două modalități de selecție este aproape identică. Se poate spune astfel că în cazul supraviețuirii proporționale selecția

deterministă (DP) este cea mai nefericită soluție pentru evoluție în timp ce selecția în turnir (TP) şi selecția proporțională (PP) sunt comparabile în capacitatea de evoluție, cu un uşor ascendent pe care selecția proporțională (PP) îl dobândeşte în timp mai îndelungat în raport cu selecția în turnir (TP) care avantajează la începutul evoluției.

În cazul supraviețuirii în turnir se observă că selecția deterministă (DT) produce sistematic cei mai retardați descendenți în timp ce selecția în turnir (TT) are o evoluție rapidă doar la început (generația 10) fiind depăşită în precocitatea evoluției în timp (generația 100000) de către selecția proporțională (PT) existând un moment (generația 1000) când cele două selecții produc evoluții relativ apropiate. Se poate spune astfel că în cazul supraviețuirii în turnir selecția deterministă (DT) este cea mai nefericită alegere pentru evoluție în timp ce selecția proporțională (PT) oferă cele mai mari şanse ca în timp evoluția să producă cei mai precoce descendenți.

În cazul supraviețuirii deterministe selecția deterministă (DD) produce sistematic cei mai retardați descendenți în timp ce selecția în turnir (TD) cu o evoluție lentă la început (generația 10) devine în timp (generația 100000) cea mai precoce pentru evoluție, iar selecția proporțională (PD), având o evoluție rapidă la început pierde din vivacitate cu trecerea timpului (generația 100000). Se poate spune astfel că în cazul supraviețuirii deterministe alegerea nefericită pentru evoluție o reprezintă o selecție de asemenea deterministă (DD) în timp ce selecția în turnir (TD) promite cele mai bune evoluții în timp mai îndelungat iar selecția proporțională (PD) asigură evoluții spectaculoase doar în prima parte a timpului.

Viteza evoluției nu este constantă şi nu este nici măcar liniară, şi asocierile de selecție şi supraviețuire care dau performanțe ridicate în primele generații ale evoluției (cum este cazul selecției în turnir însoțită de supraviețuirea în turnir - TT) produc întârzieri semnificative odată cu apropierea de limita maximă (TT ajungând din prima în performanță în prima generație în penultima în generația 10^8) în timp ce metode cu performanță scăzută (sau chiar foarte scăzută în cazul selecției deterministe însoțită de supraviețuire deterministă - DD) la începutul evoluției creează descendenți foarte valoroşi după un timp îndelungat (selecția deterministă însoțită de supraviețuire deterministă - DD ajunge a doua în performanță în generația 10^8).

Analiza a arătat că într-un număr inacceptabil de mare de cazuri media probabilității de a observa valorile extreme observate nu este conformă cu experimentul elaborat (observația minimului din 46 de execuții independente). O singură explicație este posibilă, şi ea este că legea de distribuție a valorilor observate în 46 de execuții independente într-o anumită generație a algoritmului genetic se abate de la legea valorilor extreme de tip I (Gumbel). Întrucât totuşi pentru majoritatea cazurilor observate legea de distribuție Gumbel este acceptată, rezultă consecința că legea de distribuție urmată de observații este o lege de distribuție mai generală decât legea de distribuție Gumbel dar care o are pe aceasta (Gumbel) drept caz limită. Într-adevăr această lege de

distribuție există şi este numită Legea Generală a Valorilor Extreme, sau legea Fisher-Tippett lege care are drept caz limită legea Gumbel (a valorilor extreme de tip I) ca frontieră de separație între alte două familii de legi de distribuție, legea Weibull şi legea Fréchet. S-au identificat parametrii acestei legi de distribuție pentru fiecare generație (din primele 20000) pornind de la cele 46 de execuții independente ale algoritmului genetic.

Ipoteza de distribuție după legea Fisher-Tippett a fost verificată la alternativă cu alte legi de distribuție mărginite (Beta, Johnson, Kumaraswamy, Pert, Putere, Reciprocă, Triangulară, Uniformă) şi generalizate (Pareto, Log-Pearson tip III) pentru valorile medii (mediile mediilor) dintre metode în eşantionul aleator de generații construit în acest sens când trei dintre acestea (Jonson SB - forma mărginită, Pert şi Fisher-Tippett) au obținut cea mai mare relevanță statistică prin testare cu Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D) şi Chi Square (C-S). Acestea au fost incluse în analiza individuală pentru cele 81 de extrageri (9 eşantioane aleatoare de generații la 9 asocieri de selecție şi supraviețuire). Rezultatele arată că distribuția Fisher-Tippett este singura care nu este respinsă semnificativ statistic (cu riscul de a fi în eroare de 1%) de cel puțin două din cele trei teste statistice utilizate, în timp ce celelalte două distribuții înregistrează un număr semnificativ de respingeri din totalul de 81 de eşantioane construite: Jonson SB - 20 (~25%), Pert - 18 (~22%). Analizând rezultatul obținut la respingerea unei statistici din trei, se observă în primul rând că probabilitatea are în 3 din 4 cazuri (0.83‰; 6.47‰; 6.58‰; 7.56‰) acelaşi ordin de magnitudine cu nivelul de semnificație ales (10‰) şi în al doilea rând că respingerea se face pentru statistica C-S (Chi Square). Relativ la această din urmă observație, o remarcă se cere făcută: acuratețea testului Chi-Square este scăzută pentru volume de eşantioane mici sau medii (cum e cazul celor 46 de execuții independente cuprinse în observabilă) datorită faptului că împărțirea în clase de frecvență pe care o impune aplicarea testului Chi Square este (mai mult sau mai puțin) arbitrară.

S-au pus în evidență forme similare (dar semnificativ statistic diferite) ale distribuției Fisher-Tippett pentru diferitele metode de selecție şi supraviețuire. O altă observație importantă este cu privire la valoarea parametrului de formă care este în toate cazurile negativ, ceea ce particularizează distribuția Fisher-Tippett la distribuție Weibull. Valorile negative mici ale parametrului formei, variind între -0.25 şi 0 cu uşoare tendințe crescătoare (pentru PT, PD, TP şi TT) sau descrescătoare (pentru PP, TD, DP, DT şi DD) explică de ce în testarea inițială a ipotezei distribuției Gumbel în marea majoritate a cazurilor s-a putut accepta ipoteza distribuției Gumbel dar nu s-a putut accepta pentru toate cazurile şi în acelaşi timp de ce în calcularea probabilităților de apariție a valorilor extreme s-a eşuat în verificarea ipotezei de eşantionare din distribuția Gumbel.

O observație importantă este cu privire la tendința parametrului de scală care este în toate cazurile crescătoare, ceea ce arată tendința de creştere a variabilității pentru toate asocierile de metode supuse observației.

Au fost puse în evidență grupuri distincte de forme şi scale între asocierile de metode, grupuri observabile vizual în figuri. Astfel grupurile ce se formează după forma (k) şi scala (β) distribuției sunt:

÷ Forma distribuției (k): DD, DP; PP, TP

÷ Scala distribuției (β): PT, PD; DD, DP

Pentru a pune în evidență dependența locației (λ) din distribuția Fisher-Tippett de numărul de generații s-a efectuat o analiză de regresie folosind un eşantion de 16 modele de regresie frecvent observate; dependența sugerată de analiza de regresie este fie o funcție putere fie o funcție logaritmică astfel:

÷ PP, PD, TP, TT au expresia locației în funcție de generație în forma: $\lambda(G) = a_0 + a_1 \cdot \ln(G + a_2)$

÷ PT are expresia locației în funcție de generație în forma: $\lambda(G) = a_0 + a_1 \cdot \ln(G)$

÷ TD, DP, DT, DD au expresia locației în funcție de generație în forma: $\lambda(G) = a_0 + a_1 \cdot G^{a_2}$

Loteria norocoşilor (valoarea de prag 95% din distribuția Fisher-Tippett a coeficientului de determinare în funcție de logaritmul generației) arată că susținând o selecție în turnir şi o supraviețuire deterministă indivizii din cultivar sunt sistematic mai norocoşi în găsirea determinării maxime decât cei din alte strategii de selecție şi supraviețuire, fiind urmați de cei proveniți din selecție proporțională şi supraviețuire deterministă. Codaşi în noroc sunt cei proveniți din selecție deterministă şi supraviețuire în turnir.

Loteria ghinioniştilor (valoarea de prag 5% din distribuția Fisher-Tippett a coeficientului de determinare în funcție de logaritmul generației) arată că există aceeaşi doi competitori la ghinion mic în găsirea determinării maxime: selecția în turnir şi supraviețuirea deterministă şi selecția proporțională şi supraviețuirea deterministă, ghinionul inversându-se însă între acestea la apropierea de generația 10^5. Codaşi în noroc sunt cei proveniți din selecție deterministă şi supraviețuire proporțională.

Tendința generală observată este că norocul creşte ceva mai accentuat decât în tendință liniară odată cu evoluția - doar selecția proporțională însoțită de supraviețuire proporțională (PP) se abate de la această regulă în timp ca ghinionul creşte ceva mai puțin accentuat decât în tendință liniară odată cu evoluția (doar PT şi DT se abat de la această regulă).

Evoluția în normalitate (intervalul 5%-95% în distribuția determinării în generație) evidențiază un grup compact de evoluție similară în normalitate (TT, DP, DD şi PD) şi două disimilarități extreme: DT - aproape constantă în normalitate de-a lungul evoluției şi TD - cu cea mai mare creştere a lărgimii intervalului de normalitate pe parcursul evoluției.

Turnirul (când este folosit atât în selecție cât şi în supraviețuire) măreşte hazardul alături de asocierea de strategii DD (selecție şi supraviețuire deterministe).

În ceea ce priveşte funcţia de supravieţuire nu se remarcă diferenţe deosebite între evoluţiile date de diferitele modalităţi de selecţie şi de supravieţuire, fapt aşteptat deoarece evoluţia s-a bazat pe păstrarea celor mai dotaţi indivizi (cei ce dau cea mai bună regresie) în cultivar în toate cazurile, metoda de supravieţuire permiţând prin intermediul celor 3 modalităţi (determinist, în turnir şi proporţional) alegerea indivizilor care vor fi înlocuiţi (omorâţi) doar dintre cei rămaşi (care nu sunt dintre cei mai dotaţi).

Se remarcă că selecţia deterministă însoţită de supravieţuire proporţională (DP) are în compoziţia sa o cantitate însemnată (45.1%) de valori extreme de tip II (Fréchet) şi că selecţia proporţională însoţită de supravieţuire în turnir (PT) are în compoziţia sa numai valori extreme de tip III (Weibull).

Oscilaţiile pronunţate ale asimetriei (γ_1) spre valori mici şi boltirii (γ_2) spre valori foarte mari imediat înainte de momentele de evoluţie este datorată unor evenimente rare de deviaţii extreme. Cu o singură excepţie (selecţia deterministă însoţită de supravieţuirea proporţională - DP) aceste deviaţii extreme sunt însoţite de scăderea statisticilor variabilităţii (s, σ şi β) ceea ce face ca aceste momente să poată fi considerate momente critice ale evoluţiei, momente în care boltirea arată existenţa temporară a unei segregări (împărţire în două populaţii distincte, dar care pe parcursul evoluţiei datorită încrucişării se diluează până la dispariţie).

Se estimează existenţa unei singure reţete către atingerea obiectivului (determinarea maximă) în cele mai favorabile cazuri, şi anume selecţia în turnir însoţită de supravieţuirea proporţională (TP), această combinaţie părând a fi singura ce permite (la o determinare de 94%) accesul elitelor în cultivar (având astfel cea mai scăzută retardare). Explicaţia găsită este că succesul acestei asocieri se face cu costuri pe seama precocităţii, şi anume având asigurată atingerea a 98% din valoarea coeficientului de determinare în generaţia 4150 (la o determinare de 88.8%), vine abia pe poziţia a 7-a (din totalul de 9) în ceea ce priveşte precocitatea. Cea mai precoce combinaţie, selecţia deterministă însoţită de supravieţuirea proporţională (DP), atinge în cele mai defavorabile cazuri ale sale 98% din valoarea maximă până în generaţia 47 (la o determinare de 75.6%), atinge în cele mai favorabile cazuri ale sale 99% din valoarea maximă până în generaţia 700 (şi vine pe poziţia a 5-a în clasamentul retardării în cele mai favorabile cazuri ale sale la 99% din valoarea maximă), şi fără a se putea produce estimare cu privire la atingerea valorii maxime în cele mai favorabile cazuri ale sale.

La întrebarea *Ce lege urmează momentele relative de apariţie a evoluţiei?* răspunsul găsit este că acestea urmează o distribuţie uniparametrică degenerată din log-Pearson de tipul III. S-au evidenţiat o serie de asocieri care apar între parametrii distribuţiilor log-Pearson de tipul III ale momentelor relative de evoluţie, reducând din complexitatea distribuţiei din distribuţie tri-parametrică în distribuţie uni-parametrică, distribuţie la care s-a estimat parametrul independent

rămas (locația) şi pentru care statisticile Perason-Fisher, Anderson-Darling şi Kolmogorov-Smirnov nu au putut respinge ipoteza de distribuție pentru distribuțiile obținute la un risc de a fi în eroare de 5%.

La întrebarea "*Ce lege urmează numărul de evoluții?*" s-a găsit ca răspuns legea de distribuție Fisher-Tippett.

Algoritmii genetici, considerați de specialişti o aplicație a inteligenței artificiale, au conform studiului realizat perspective de a deveni şi apanajul geneticienilor şi amelioratorilor, respectându-se condiția ca, în procesul de selecție al celor mai bune genotipuri destinate înmulțirii, folosirea unei anumite strategii de selecție (proporțională, în turnir şi deterministă fiind cele analizate în lucrare) trebuie asociată cu obiectivele generale de ameliorare urmărite.

Un asemenea deziderat se suprapune aplicațiilor practice prin care algoritmii genetici contribuie la rezolvarea problemelor de optimizare, planificare ori căutare în chimie, informatică, matematică, modelare moleculară etc., dar şi în genetica şi ameliorarea plantelor. În acest sens, prezenta lucrare deschide noi perspective aplicative în domeniu, şi se constituie totodată într-o cercetare fundamentală originală, inedită în horticultură, deschizătoare de noi direcții de cunoaştere a fenomenelor şi proceselor biologice, capabilă să permită formularea şi verificarea de noi ipoteze, modele conceptuale şi teorii.

Referințe

(Aickelin & Dowsland, 1999): Aickelin U, 1999. Genetic Algorithms for Multiple-Choice Optimisation Problems. PhD Thesis (European Business Management) - Supervisor Prof. Dowsland K. University of Wales Swansea, Swansea, UK. PhD_Aickelin&Dowsland_1999.pdf

(Altschul, 1989): Altschul SF, 1989. Gap costs for multiple sequence alignment. *J Theoretical Biology* 138:297-309. http://dx.doi.org/10.1016/S0022-5193(89)80196-1

(Anandhi & alții, 2008): Anandhi A, Srinivas VV, Nanjundiah RS, Nagesh Kumar D, 2008. Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. *Int J Climatol* 28(3):401-420. Anandhi&others_2008.pdf

(Anderson & alții, 2003): Anderson RP, Lew D, Peterson AT, 2003. Evaluating predictive models of species distributions: criteria for select optimal models. *Ecol Model* 162(3):211-232. http://dx.doi.org/10.1016/S0304-3800(02)00349-6

(Anderson & Darling, 1952): Anderson TW, Darling DA, 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics* 23(2):193-212. Anderson&Darling_1952.pdf

(Anisimova & Liberles, 2007): Anisimova M, Liberles DA. 2007. The quest for natural selection in the age of comparative genomics. *Heredity* 99(6):567-579. Anisimova&Liberles_2007.pdf

(Annevelink, 1992): Annevelink E, 1992. Operational planning in horticulture: Optimal space allocation in pot-plant nurseries using heuristic techniques. *J Agr Eng Res* 51(C):167-177. http://dx.doi.org/10.1016/0021-8634(92)80035-Q

(Arora K & Singh, 1989): Arora K, Singh VP, 1989. A comparative evaluation of the estimators of the log Pearson type (LP) 3 distribution. J Hydrol 105:19-37. Arora&Singh_1989.pdf

(Auerbach & alții, 1947): Auerbach C, Robson JM, Carr JG, 1947. The chemical production of mutations. *Science* 105(2723):243-247. http://dx.doi.org/10.1126/science.158.3805.1141

(Ayala & alții, 1994): Ayala FJ, Escalante A, O'Hugin C, Klein J, 1994. Molecular genetics of speciation and human origins. *Proc Natl Acad Sci USA* 91(15):6787-6794. www.pnas.org/content/91/15/6787.full.pdf

(Bak & Sneppen, 1993): Bak P, Sneppen K, 1993. Punctuated equilibrium and criticality in a simple model of evolution. *Phys Rev Lett* 71(24):4083-4086. http://dx.doi.org/10.1103/PhysRevLett.71.4083

(Bălan & alții, 2008): Bălan MC, Damian M, Jäntschi L, 2008. Preliminary results on design and implementation of a solar radiation monitoring system. *Sensors* 8(2):963-978. www.mdpi.org/sensors/papers/s8020963.pdf

(Baldwin & Sanderson, 1998): Baldwin BG. Sanderson MJ, 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae). *Proc Natl Acad Sci USA* 95(16):9402-9406. http://www.pnas.org/content/95/16/9402.full

(Banzhaf & alții, 1997): Banzhaf W, Nordin P, Keller RE, Francone FD, 1997. Genetic Programming: An Introduction: On the Automatic Evolution of Computer Programs and Its Applications. San Francisco: Morgan Kaufmann Publishers (450 p.). http://lccn.loc.gov/97051603

(Baranczewski & alții, 2006): Baranczewski P, Stanczak A, Sundberg K, Svensson R, Wallin A, Jansson J, Garberg P, Postlind H, 2006. Introduction to in vitro estimation of metabolic stability and drug interactions of new chemical entities in drug discovery and development. *Pharmacol Rep* 58(4):453-472. www.if-pan.krakow.pl/pjp/pdf/2006/4_453.pdf

(Barricelli, 1954): Barricelli N, 1954. Esempi numerici di processi di evoluzione. Methodos 1954:45-68. www.citeulike.org/user/richpb7/article/5092082

(Bateson, 1907): Bateson W. 1907. Facts limiting the theory of heredity. *Science*

26(672):649-660. http://dx.doi.org/10.1126/science.26.678.928-a
(Benoit & alții, 2005): Benoit B, Fleurey F, Jézéquel J-M, Le Traon Y, 2005. Automatic Test Case Optimization: A Bacteriologic Algorithm. *IEEE Software* 22(2):76-82. http://dx.doi.org/10.1109/MS.2005.30
(Bleicher & alții, 2003): Bleicher KH, Böhm H-J, Muller K, Alanine AI 2003. Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov* 2:369-378. Bleicher&others_2003.pdf
(Boer & alții, 2005): de Boer P-T, Kroese DP, Mannor S, Rubinstein RY, 2005. A Tutorial on the Cross-Entropy Method. *Ann Oper Res* 134(1):19-67. Boer&others_2005.pdf
(Boken & alții, 2008): Boken VK, Hoogenboom G, Williams JH, Diarra B, Dione S, Easson GL, 2008. Monitoring peanut contamination in Mali (Africa) using AVHRR satellite data and a crop simulation model. *Int J Remote Sens* 29(1):117-129. Boken&others_2008.pdf
(Bolboacă & Jäntschi, 2006-PSK): Bolboacă SD, Jäntschi L, 2006. Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds. *Leonardo Journal of Sciences* 5(9):179-200. http://ljs.academicdirect.ro/A09/179_200.pdf
(Bolboacă & Jäntschi, 2007-DoE): Bolboacă SD, Jäntschi L, 2007. Design of Experiments: Useful Orthogonal Arrays for Number of Experiments from 4 to 16. *Entropy* 9(4):198-232. http://dx.doi.org/10.3390/e9040198
(Bolboacă & Jäntschi, 2007-HowGood): Bolboacă SD, Jäntschi L. How Good the Characteristic Polynomial Can Be for Correlations? *International Journal of Molecular Sciences* 2007:8(4):335-345. www.mdpi.org/ijms/papers/i8040335.pdf
(Bolboacă & Jäntschi, 2009-Comparison): Bolboacă SD, Jäntschi L, 2009. Comparison of QSAR Performances on Carboquinone Derivatives, *TheScientificWorldJOURNAL* 9(10):1148-1166. http://dx.doi.org/10.1100/tsw.2009.131
(Bosworth & alții, 1972): Bosworth J, Norman F, Zeigler BP, 1972. Comparison of Genetic Algorithms with Conjugate Gradient Methods. *NASA Contractor Reports* CR-2093. Bosworth&others_1972.pdf
(Bouktir & Slimani, 2005): Bouktir T, Slimani L. 2005. Optimal Power Flow of the Algerian Electrical Network using an Ant Colony Optimization Method, *Leonardo J Sci* 4(7):43-57. http://ljs.academicdirect.org/A07/43_57.pdf
(Bozorg-Haddad & alții, 2009): Bozorg Haddad O, Moradi-Jalal M, Mirmomeni M, Kholghi MKh, Marino MA, 2009. Optimal Cultivation Rules in Multi-Crop Irrigation Areas. *Irrig Drain* 58(1):38-49. Bozorg-Haddad&others_2009.pdf
(Brady & alții, 2006): Brady SG, Schultz TR, Fisher BL, Ward PS, 2006. Evaluating alternative hypotheses for the early evolution and diversification of ants. *Proc Natl Acad Sci USA* 103(48):18172-18177. http://www.pnas.org/content/103/48/18172.full
(Brauer, 1937): Brauer R, 1937. On algebras which are connected with the semisimple continuous groups. *Ann Math* 38(4):857-872. www.jstor.org/pss/1968843
(Bravi & alții, 1997): Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A, 1997. MS-WHIM., New 3D Theoretical Descriptors Derived from Molecular Surface Properties: A Comparative 3D QSAR Study in a Series of Steroids, J Comput-Aided Mol Des 11:79-92. Bravi&others_1997.pdf
(Brown & alții, 2000): Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares MJr, Haussler D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 97(1):262-267. http://www.pnas.org/content/97/1/262.full
(Burckhardt & Basset, 2000): Burckhardt D, Basset Y, 2000. The jumping plant-lice (Hemiptera, Psylloidea) associated with Schinus (Anacardiaceae): Systematics, biogeography and host plant relationships. *J Nat Hist* 34(1):57-155. Burckhardt&Basset_2000.pdf
(Busch & alții, 2008): Busch H, Camacho-Trullio D, Rogon Z, Breuhahn K, Angel P, Eils R,

Szabowski A. 2008. Gene network dynamics controlling keratinocyte migration. *Molec Syst Biol* 4:199(16). Busch&others_2008.pdf
(Buwalda & alții, 2006): Buwalda F, van Henten EJ, de Gelder A, Bontsema J, Hemming J, 2006. Toward an Optimal Control Strategy for Sweet Pepper Cultivation - 1. A Dynamic Crop Model. *ISHS Acta Horticulturae* 718(III):367-374. www.actahort.org/books/718/718_42.htm
(Cabeza & Moilanen, 2001): Cabeza M, Moilanen A, 2001. Design of reserve networks and the persistence of biodiversity. *Trends Ecol Evol* 16(5):242-248. http://dx.doi.org/10.1016/S0169-5347(01)02125-5
(Cains & alții, 1988): Cairns J, Overbaugh J, Miller S. 1988. The origin of mutants. *Nature* 335(6186):142-145. http://dx.doi.org/10.1038/335142a0
(Califano, 2000): Califano A, 2000. SPLASH: Structural pattern localization analysis by sequential histograms. *Bioinformatics* 16(4):341-357. Califano_2000.pdf
(Caroppo & alții, 2006): Caroppo C, Stabili L, Aresta M, Corinaldesi C, Danovaro R, 2006. Impact of heavy metals and PCBs on marine picoplankton. *Environ Toxicol* 21(6):541-551. Caroppo&others_2006.pdf
(Carpio-Muñoz & alții, 2002-Bioinformatic): Del Carpio-Muñoz CA, Hennig T, Fickel S, Yoshimori A. 2002. A combined bioinformatic approach oriented to the analysis and design of peptides with high affinity to MHC class I molecules. *Immun Cell Biol* 80(3):286-298. Carpio-Munoz&others_2002-B.pdf
(Carpio-Muñoz & alții, 2002-Miax): Del Carpio-Muñoz CA, Ichiishi E, Yoshimori A, Yoshikawa T. 2002. MIAX: A new paradigm for modeling biomacromolecular interactions and complex formation in condensed phases. *Prot Struct Funct Genet* 48(4):696-732. Carpio-Munoz&others_2002-M.pdf
(Chen & alții, 2003): Chen Z, Huang GH, Chan CW, Geng LQ, Xia J, 2003. Development of an Expert System for the Remediation of Petroleum-Contaminated Sites. *Environmental Modeling and Assessment* 8(4):323-334. Chen&others_2003.pdf
(Chen & alții, 2008): Chen C-S, Liu C-H, Su H-C. 2008. A nonlinear time series analysis using two-stage genetic algorithms for streamflow forecasting. *Hydrol Process* 22(18):3697-3711. Chen&others_2008.pdf
(Chen & MacKerell, 2000): Chen I-J, MacKerell Jr AD, 2000. Computation of the influence of chemical substitution on the pKa of pyridine using semiempirical and ab initio methods. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* 103(6):483-494. Chen&MacKerell_2000.pdf
(Chen, 1997): Chen Y-M, 1997. Management of water resources using improved genetic algorithms. *Comput Electron Agr* 18(2-3):117-127. Chen_1997.pdf
(Chtioui & alții, 1996): Chtioui Y, Bertrand D, Barba D, 1996. Reduction of the size of the learning data in a probabilistic neural network by hierarchical clustering. Application to the discrimination of seeds by artificial vision. *Chemometr Intell Lab* 35(2):175-186. http://dx.doi.org/10.1016/S0169-7439(96)00065-2
(Chtioui & alții, 1997): Chtioui Y, Bertrand D, Devaux M-F, Barba D, 1997. Comparison of multilayer perceptron and probabilistic neural networks in artificial vision. Application to the discrimination of seeds. *J Chemometr* 11(2):111-129. Chtioui&others_1997.pdf
(Chtioui & alții, 1998): Chtioui Y, Dominique B, Dominique B, 1998. Feature Selection by a Genetic Algorithm. Application to Seed Discrimination by Artificial Vision, *J Sci Food Agr* 76:77-86. Chtioui_others_1998.pdf
(Chu & alții, 2009): Chu CW, Holliday JD, Willett P, 2009. Effect of Data Standardization on Chemical Clustering and Similarity Searching. Journal of Chemical Information and Modeling, 49(2): 155-161. http://dx.doi.org/10.1021/ci800224h
(Chuman & alții, 1998): Chuman H, Karasawa M, Fujita T, 1998. A Novel Three-Dimensional QSAR Procedure: Voronoi Field Analysis, *QSAR* 17(4):313-326. Chuman&others_1998.pdf

(Corpet & Michot, 1994): Corpet F, Michot B, 1994. RNAlign program: alignment of RNA sequences using both primary and secondary structures. *Comput Applicat Biosci (Bioinformatics)* 10(4):389-399. http://bioinformatics.oxfordjournals.org/cgi/content/short/10/4/389
(Crum-Brown & Fraser, 1868): Crum-Brown A, Fraser TR, 1868. On the Connection between Chemical Constitution and Physiological Action. Part I. On the Physiological Action of the Salts of the Ammonium Bases, derived from Strychnia, Brucia,Thebaia, Codeia, Morphia, and Nicotia. *Philosophical Transactions of the Royal Society of London* 25:151-203. www.ncbi.nlm.nih.gov/pmc/articles/PMC1318606/pdf/janatphys00216-0024.pdf
(Cui & Wu, 2005): Cui Y, Wu R, 2005. Mapping genome-genome epistasis: A high-dimensional model. *Bioinformatics* 21(10):2447-2455. Cui&Wu_2005.pdf
(Cunningham & Day, 2005): Cunningham SL, Cunningham AR, Day BW, 2005. CoMFA, HQSAR and molecular docking studies of butitaxel analogues with b-tubulin, *Journal of Molecular Modeling* 11(1):48-54. Cunningham&Day_2005.pdf
(Cunningham & Day, 2005): Cunningham SL, Cunningham AR, Day BW, 2005. CoMFA, HQSAR and molecular docking studies of butitaxel analogues with b-tubulin, *Journal of Molecular Modeling* 11(1):48-54. Cunningham&Day_2005.pdf
(Czaplinski & alții, 1995): Czaplinski K-H, Haensel W, Wiese M, Seydel JK, 1995. New benzylpyrimidines: Inhibition of DHFR from various species - QSAR, CoMFA and PC analysis. *Eur J Med Chem* 30:779-787. http://dx.doi.org/10.1016/0223-5234(96)88297-3
(Czaplinski & Grunewald, 1994): Czaplinski K-H, Grunewald GL, 1994. A comparative molecular field analysis derived model of the binding of taxol analogues to microtubules. *Bioorg Med Chem* 4:2211-2216. http://dx.doi.org/10.1016/S0960-894X(00)80073-6
(Dagdeviren & Eraslan, 2008): Dagdeviren M, Eraslan E, 2008. Priority determination in strategic energy policies in Turkey using analytic network process (ANP) with group decision making. *Int J Energ Res* 32(11):1047-1057. Dagdeviren&Eraslan_2008.pdf
(Dai & Lodder, 2007): Dai B, 2007. Simulations-guided design of process analytical sensor using molecular factor computing. PhD Thesis (Chemistry) - Supervisor Prof. Lodder RA. University of Kentucky, Lexington, KY, USA. PhD_Dai&Lodder_2007.pdf
(Daniell & alții, 2008): Daniell H, Wurdack KJ, Kanagaraj A, Lee S-B, Saski C, Jansen RK, 2008. The complete nucleotide sequence of the cassava (Manihot esculenta) chloroplast genome and the evolution of atpF in Malpighiales: RNA editing and multiple losses of a group II intron. *Theor Appl Genet* 116(5):723-737. Daniell&others_2008.pdf
(Darwin, 1859): Darwin CR, 1859. The origin of species by means of natural selection or the preservation of favoured races in the struggle for life, London: John Murray (546 p.). Darwin_1859.pdf
(Darwin, 1872): Darwin CR, 1872. The origin of species by means of natural selection or the preservation of favoured races in the struggle for life, London: John Murray (482 p.). Darwin_1872.pdf
(David & alții, 1999): David NP, Darryl JCP, David MC, John SC, 1999. Probability based protein identification by searching sequence database using mass spectrometry data. *Electrophoresis* 20(18):3551-3567. David&others_1999.pdf
(Davis, 1987): Davis L, 1987. Genetic Algorithms and Simulated Annealing. San Francisco: Morgan Kaufmann (216p.). http://lccn.loc.gov/87021357
(Davis, 1991): Davis L, 1991. Handbook of Genetic Algorithms. New York: van Nostrand Reinhold. (385 p.). http://lccn.loc.gov/90012823
(Debnath, 1994): Debnath AK, Jiang S, Strick N, Lin K, Haberfield P, Neurath AR, 1994. Three-dimensional structure-activity analysis of a series of porphyrin derivatives with anti-HIV-1 activity targeted to the V3 loop of the gp120 envelope glycoprotein of the human immunodeficiency virus type 1. *J Med Chem* 37:1099-1108. http://dx.doi.org/10.1021/jm00034a007
(Denisov & alții, 2003): Denisov E, Sarkisov O, Likhtenshtein G, 2003. Chapter 3 (Methods

for Studying Elementary Reactions) In: *Chemical Kinetics: Fundamentals and New Developments*, Amsterdam: Elsevier, p. 67-97. http://lccn.loc.gov/2003051385
(Diaconis & Holmes, 1998): Diaconis PW, Holmes SP, 1998. Matchings and phylogenetic trees. *Proc Natl Acad Sci USA* 95(25):14600-14602. http://www.pnas.org/content/95/25/14600.full
(Diaz & alții, 2005): Diaz B, Ribeiro A, Bueno R, Guinea D, Barroso J, Ruiz D, Fernadez-Quintanilla C, 2005. Modelling Wild-Oat Density in Terms of Soil Factors: A Machine Learning Approach. *Precis Agric* 6:213-228. Diaz&others_2005.pdf
(Diudea & alții, 2001): Mircea DIUDEA, Ivan GUTMAN, Lorentz JÄNTSCHI, Molecular Topology. New York: Nova Science, (332 p.). http://lccn.loc.gov/2001031282
(Donthamsetti, 1980): Donthamsetti VR, 1980. Technical Publication SJ 80-8 Log Pearson Type 3 Distribution: Tables Of Quantiles. Water Resources Department, Project Number 2003233. Donthamsetti_1980.pdf
(Doweyko, 1994): Doweyko AM, 1994. Pharmacophores from binding data. *J Med Chem* 37(12):1769-1778. http://dx.doi.org/10.1021/jm00038a006
(Droste & alții, 2001): Droste S, Jansen T, Wegene I, 2001. Dynamic Parameter Control in Simple Evolutionary Algorithms. In: *Foundations of genetic algorithms 6* (Martin WN, Spears WM). San Francisco: Morgan Kaufmann, p. 275-294. http://lccn.loc.gov/2001275761
(Duchowicz & alții, 2008): Duchowicz PR, Talevi A, Bruno-Blanch LE, Castro EA, 2008. New QSPR study for the prediction of aqueous solubility of drug-like compounds. *Bioorganic & Medicinal Chemistry* 16:7944-7955. Duchowicz&others_2008.pdf
(Duda-Seiman & alții, 2006): Duda-Seiman C, Duda-Seiman D, Dragoş D, Medeleanu M, Careja V, Putz MV, etc, 2006. Design of anti-HIV ligands by means of minimal topological difference (MTD) method. *International Journal of Molecular Sciences* 7(11):537-555. www.mdpi.org/ijms/papers/i7110537.pdf
(Durrans SR & Tomic S, 1996): Durrans SR, Tomic S, 1996. Regionalization of low-flow frequency estimates: An Alabama case study. Water Resources Bulletin 32(1):23-37. http://dx.doi.org/10.1111/j.1752-1688.1996.tb03431.x
(Durst, 1998): Durst GL, 1998. Comparative Molecular Field Analysis (CoMFA) of Herbicidal Protoporphyrinogen Oxidase Inhibitors using Standard Steric and Electrostatic Fields and an Alternative LUMO Field. *Quant Struct-Act Relat*, 17:419-426. Durst_1998.pdf
(Eisler & Belisle, 1996): Ronald EISLER, André A. BELISLE. 1996. Planar PCB Hazards to Fish, Wildlife, and Invertebrates: A Synoptic Review. Contaminant Hazard Reviews. Biological Report 31. http://www.pwrc.usgs.gov/infobase/eisler/chr_31_planar_pcbs.pdf
(English, 1996): English TM, 1996. Evaluation of Evolutionary and Genetic Optimizers:No Free Lunch. In: *Evolutionary Programming V: Proceedings of the Fifth Annual Conference on Evolutionary Programming* (Fogel LJ, Angeline PJ & Bäck T), Cambridge, Mass: MIT Press, p. 163-169. English_1996.pdf
(Eriksson & alții, 2003): Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P, 2003. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environmental Health Perspectives* 111(10):1361-1375. www.ncbi.nlm.nih.gov/pmc/articles/PMC1241620/pdf/ehp0111-001361.pdf
(Falkenauer, 1998): Falkenauer E, 1998. Genetic Algorithms and Grouping Problems. New York:Wiley. (220p.). http://lccn.loc.gov/97031527
(Filimonov & alții, 2009): Filimonov DA, Zakharov AV, Lagunin AA, Poroikov VV, 2009. QNA-based 'Star Track' QSAR approach. SAR and QSAR in Environmental Research 20(7-8):679-709. http://dx.doi.org/10.1080/10629360903438370
(Fink, 1986): Fink WL, 1986. Microcomputers and phylogenetic analysis. *Science* 234(4780):1135-1139. http://dx.doi.org/10.1126/science.234.4780.1135

(Fisher & Tippett, 1928): Fisher RA, Tippett LHC, 1928. Limiting Forms of the Frequency Distribution of the Largest of Smallest Member of a Sample. *Proceedings of the Cambridge Philosophical Society* 24:180-190. Fisher&Tippett_1928_GEVd.pdf
(Fisher, 1918): Fisher RA, 1918. The Correlation between Relatives on the Supposition of Mendelian Inheritance, *Transactions of the Royal Society of Edinburgh* 52:399-433. Fisher_1918_Correlation.pdf
(Fisher, 1920-Accuracy): Fisher RA, 1920. A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error. *Monthly Notices of the Royal Astronomical Society* 80(S):758-770. Fisher_1920_Accuracy.pdf
(Fisher, 1922): Fisher RA, 1922. On the Dominance Ratio, *Proceedings of the Royal Society of Edinburgh* 42:321-341. Fisher_1922_Dominance.pdf
(Fisher, 1922-Estimation): Fisher RA, 1922. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A* 222:309-368. Fisher_1922_Estimation.pdf
(Fisher, 1922-X2): Fisher RA, 1922. On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society* 85:87-94. Fisher_1922_X2.pdf
(Fisher, 1924): Fisher RA, 1924. The Conditions Under Which $\chi 2$ Measures the Discrepancy Between Observation and Hypothesis. *Journal of the Royal Statistical Society* 87:442-450. Fisher_1924_ Discrepancy.pdf
(Fisher, 1926): Fisher RA, 1926. The arrangement of field experiments. *J Minist Agr Great Brit* 33:503-513. Fisher_1926_Experiments.pdf
(Fisher, 1935): Fisher RA, 1935. The Mathematical Distributions Used in the Common Tests of Significance. *Econometrica* 3:353-365. Fisher_1935_ Distributions.pdf
(Fisher, 1954): Fisher RA, 1954. Retrospect of the Criticisms of the Theory of Natural Selection, In: *Evolution as a Process* (Huxley JS, Hardy AC, Ford EB), London: Allen&Unwin, p. 84-98. Fisher_1954_Retrospect.pdf
(Fogel & alții, 1966): Fogel LJ, Owens AJ, Walsh MJ, 1966. Artificial Intelligence Through Simulated Evolution. New York: John Wiley & Sons (170 p.). http://lccn.loc.gov/66021067
(Fowler & alții, 2007): Fowler HJ, Blenkinsop S, Tebaldi C, 2007. Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *Int J Climatol* 27(12):1547-1578. Fowler&others_2007.pdf
(Fraser, 1957-1970): Fraser AS. 1957. Simulation of genetic systems by automatic digital computers. I. Introduction. *Australian J Biol Sci* 10:484-491. Fraser AS. 1957. Simulation of genetic systems by automatic digital computers. II. Effects on linkage on rates of advance under selection. *Australian J Biol Sci* 10:492-499. Barker JSF. 1958. Simulation of genetic systems by automatic digital computers. III. Selection between alleles at an autosomal locus. *Australian J Biol Sci* 11:603-612. Barker JSF. 1958. Simulation of genetic systems by automatic digital computers. IV. Selection between alleles at a sex-linked locus *Australian J Biological Sciences* 11:613-625. Fraser AS. 1960. Simulation of genetic systems by automatic digital computers. V. Linkage, dominance, and epistasis, In: *Biometrical Genetics* (Kempthorne O). New York: Pergamon Press, p. 70-83. Fraser AS. 1960. Simulation of genetic systems by automatic digital computers. VI. Epistasis. Australian *J Biol Sci* 13(2):150-162. Fraser AS. 1960. Simulation of genetic systems by automatic digital computers. VII. Effects of reproductive rate and intensity of selection on genetic structure. Australian *J Biol Sci* 13(3):344-350. Fraser AS. 1962. Simulation of genetic systems VIII. *J Theor Biol* 2(3):329-346. Fraser AS, Hansche PE. 1965. Simulation of genetic systems. IX. Major and minor loci. In: *Genetics Today, Proc. XI International Congress of Genetics* (Geerts SJ), Oxford: Pergamon Press, 3:507-516. Fraser AS, Burnell D, Miller D. 1966. Simulation of genetic systems. X. Inversion polymorphism. *J Theor Biol* 13:1-14. Fraser AS, Burnell D. 1967. Simulation of genetic systems. XI. Inversion

polymorphism. *Amer J Human Genetics* 19:270-287. Fraser A, Burnell D. 1967. Simulation of genetic systems. XII. Models of inversion polymorphism. *Genetics* 57:267-282. Fraser A, Burnell D, 1970. Computer Models in Genetics. New York: McGraw-Hill, p. 206. http://l.academicdirect.org/Horticulture/GAs/Refs/Fraser_1954_1970/

(Gantovnik & Gürdal, 2005): Gantovnik VB, 2005. An Improved Genetic Algorithm for the Optimization of Composite Structures. PhD Thesis (Engineering Mechanics) - Supervisor Prof. Gürdal Z. Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. PhD_Gantovnik&Gurdal_2005.pdf

(Gilks & Roberts, 1996): Gilks WR, Roberts GO, 1996. Strategies for improving MCMC, In: Markov Chain Monte Carlo in Practice (Gilks WR, Richardson S, Spiegelhalter D). London: Chapman & Hall, p. 75-88. http://lccn.loc.gov/98033429

(Glover & alții, 1993): Glover F, Taillard E, Taillard E. 1993. A user's guide to tabu search. *Ann Oper Res* 41(1):1-28. http://dx.doi.org/10.1007/BF02078647

(Glower, 1977): Glover F, 1977. Heuristics for Integer Programming Using Surrogate Constraints. *Decision Sci* 8(1):156-166. http://dx.doi.org/10.1111/j.1540-5915.1977.tb01074.x

(Glower, 1986): Glover F, 1986. Future Paths for Integer Programming and Links to Artificial Intelligence. *Comput Oper Res* 13(5):533-549. Glover_1986.pdf

(Greenberg & Ziff, 1984): Greenberg ME, Ziff EB. 1984. Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene. *Nature* 311(5985):433-438. http://dx.doi.org/10.1038/311433a0

(Gribskov & alții, 1987): Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA* 84(13):4355-4358. http://www.pnas.org/content/84/13/4355.full.pdf+html

(Grubbs, 1969): Grubbs F, 1969. Procedures for Detecting Outlying Observations in Samples. *Technometrics* 11(1):1-21. http://www.jstor.org/pss/1266761

(Guangzhu & alții, 2006): Guangzhu F, Fasong L, Heping L, Hai Q, Yingde C, 2006. A Hybrid Genetic Algorithm for the Estimation of Polyesterification Kinetic Parameters. *Chem Eng Technol* 29(6):740-743. Guangzhu&others_2006.pdf

(Hammett, 1935): Hammett LP, 1935. Some Relations between Reaction Rates and Equilibrium Constants. *Chemical Reviews* 17:125-136. http://dx.doi.org/10.1021/cr60056a010

(Hansch & Leo, 1979): Hansch C, Leo A 1979. Substituent Constants for Correlation Analysis in Chemistry and Biology. New York: John Wiley & Sons. http://lccn.loc.gov/78026721

(Hao & alții, 2009): Hao DC, Huang B, Yang L. 2009. Molecular evolution of paclitaxel biosynthetic genes TS and DBAT of Taxus species. *Genetica* 135(2):123-135. Hao&others_2009.pdf

(Hartley, 1928): Hartley RVL, 1928. Transmission of Information. Bell Syst Tech J 1928:535-563. Hartley_1928.pdf

(Hashimoto, 1997): Hashimoto Y, 1997. Applications of artificial neural networks and genetic algorithms to agricultural systems. *Comput Electron Agr* 18(2-3):71-224. Hashimoto_1997.pdf

(Hemmateenejad, 2006): Hemmateenejad B. Chemometrics in Iran. *Chemometrics and Intelligent Laboratory Systems* 81(2):202-208. Hemmateenejad_2006.pdf

(Henderson & alții, 2000): Henderson CE, Potter WD, McClendon RW, Hoogenboom G, 2000. Predicting aflatoxin contamination in peanuts: A genetic algorithm/neural network approach. *Appl Intell* 12(3):183-192. Henderson&others_2000.pdf

(Henten & alții, 2006): van Henten EJ, Buwalda F, de Zwart HF, de Gelder A, Hemming J, Bontsema J, 2006. Toward an Optimal Control Strategy for Sweet Pepper Cultivation - 2. Optimization of the Yield Pattern and Energy Efficiency. *ISHS Acta Horticulturae* 718(III):391-398. www.actahort.org/books/718/718_45.htm

(Higuchi & alții, 2000): Higuchi T, Tsutsui S, Yamamura M, 2000. Theoretical analysis of simplex crossover for real-coded genetic algorithms. *Lecture Notes in Computer Science* 1917:365-374. http://dx.doi.org/10.1007/3-540-45356-3_36
(Hlavacek & Savageau, 1996): Hlavacek WS, Savageau MA, 1996. Rules for coupled expression of regulator and effector genes in inducible circuits. *J Mol Biol* 255(1):121-139. http://dx.doi.org/10.1006/jmbi.1996.0011
(Holland, 1975): Holland JH, 1975. Adaptation in Natural and Artificial Systems: an introductory analysis with applications to biology, control, and artificial intelligence. Michigan: University of Michigan Press (183 p.) http://lccn.loc.gov/74078988
(Holland, 1992): Holland JH, 1992. Adaptation in Natural and Artificial Systems: an introductory analysis with applications to biology, control, and artificial intelligence. Cambridge: MIT Press (228 p.). http://lccn.loc.gov/91041044
(Horwitz & alții, 1993): Horwitz JP, Massova I, Wiese T, Wozniak J, Corbett TH, Sebolt-Leopold JS, Capps DB, Leopold WR, Comparative Molecular Field Analysis of in Vitro Growth Inhibition of L1210 and HCT-8 Cells by Some Pyrazoloacridines. *J Med Chem* 36(23):3511-3516. http://dx.doi.org/10.1021/jm00075a004
(Hruschka & alții, 2003): Hruschka Jr ER, Hruschka ER, Ebecken NFF, 2003. A feature selection Bayesian approach for extracting classification rules with a clustering genetic algorithm. *Appl Artif Intell* 17(5-6):489-506. http://dx.doi.org/10.1080/713827176
(Huelsenbeck & alții, 2001): Huelsenbeck JP, Ronquist FR, Nielsen R, Bollback JP, 2001. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* 294(5550):2310-2314. http://dx.doi.org/10.1126/science.1065889
(Husmann & Tantau, 2001): Husmann H-J, Tantau H-J, 2001. Integrated optimization of energy supply systems in horticulture using genetic algorithms. *Comput Electron Agr* 31(1):47-59. Husmann&Tantau_2001.pdf
(Hvidsten & alții, 2001): Hvidsten TR, Komorowski J, Sandvik AK, Laegreid A, 2001. Predicting Gene Function from Gene Expressions and Ontologies. *Pac Symp Biocomput* 6:299-310. Hhvidsten&others_2001.pdf
(Iyer & alții, 1999): Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson JJr, Boguski MS, Mark S. Boguski, Lashkari D, Shalon D, Botstein D, Brown PO, 1999. The Transcriptional Program in the Response of Human Fibroblasts to Serum. *Science* 283(5398):83-87. Iyer&others_1999.pdf
(Jähner & Hunter, 1991): Jähner D, Hunter T, 1991. The ras-related gene rhoB is an immediate-early gene inducible by v-Fps, epidermal growth factor, and platelet-derived growth factor in rat fibroblasts. *Mol Cell Biol* 11(7):3682-3690. Jahner&Hunter_1991.pdf
(Jäntschi & alții, 2000): Jäntschi L, Katona G, Diudea MV, 2000. Modeling Molecular Properties by Cluj Indices, *Commun. Math. Comput. Chem. (MATCH)* 41:151-188. Jantschi&others_2000.pdf
(Jäntschi & alții, 2007-Chromatogr): Jäntschi L, Bolboacă SD, Diudea MV. 2007. Chromatographic Retention Times of Polychlorinated Biphenyls: from Structural Information to Property Characterization. *Int J Mol Sci* 8(11): 1125-1157. http://dx.doi.org/10.3390/i8111125
(Jäntschi & alții, 2008-PTA): Jäntschi L, Bolboacă SD, Sestraş RE, 2008. On about what Can Be Done and what Cannot Be Done with Genetic Algorithms in Phylogenetic Tree and Gene Sequence Analyses. Bulletin UASVM, Horticulture 65(1):63-70. Jantschi&others_2008_PTA.pdf
(Jäntschi & alții, 2008-Tetrahymena): Jäntschi L, Popescu V, Bolboacă SD, 2008. Toxicity caused by para-substituted phenols on Tetrahymena pyriformis: The structure-activity relationships. *Electron J Biotechnol* 11(3):9(12). http://dx.doi.org/10.2225/vol11-issue3-fulltext-9
(Jäntschi & alții, 2009-DPA): Jäntschi L, Bolboacă SD, Sestraş RE, 2008. Classical Approaches of Genetic Algorithms and their Suitability. Asian Journal of Chemistry

22(3):2275-2284, 2010. Jantschi&others_2009_DPA.pdf
(Jäntschi & alții, 2009-GSA): Jäntschi L, Bolboacă SD, Sestraş RE, 2008. Hard Problems in Gene Sequence Analysis: Classical Approaches and Suitability of Genetic Algorithms. Biotechnology & Biotechnological Equipment 23(2):1275-1280, 2009. http://www.diagnosisp.com/dp/journals/view_pdf.php?journal_id=1&archive=1&issue_id=23&article_id=765
(Jäntschi & alții, 2010-Evolution): Jäntschi L, Bolboacă SD, Sestraş RE, 2010. A Study of Genetic Algorithm Evolution on the Lipophilicity of Polychlorinated Biphenyls. *Chemistry and Biodiversity* DOI: 10.1002/cbdv.200900356. http://dx.doi.org/10.1002/cbdv.200900356
(Jäntschi & Bolboacă, 2006-MDF-SAR-6): Jäntschi L, Bolboacă SD, 2006. Molecular Descriptors Family on Structure Activity Relationships 6. Octanol-Water Partition Coefficient of Polychlorinated Biphenyls, *Leonardo Electronic Journal of Practices and Technologies* 5(8):71-86. http://lejpt.academicdirect.org/A08/71_86.pdf
(Jäntschi & Bolboacă, 2007-DoE): Bolboacă SD, Jäntschi L. 2007. Design of Experiments: Useful Orthogonal Arrays for Number of Experiments from 4 to 16. *Entropy* 9(4):198-232. www.mdpi.org/entropy/papers/e9040198.pdf
(Jäntschi & Bolboacă, 2007-Linear): Jäntschi L, Bolboacă SD, 2007. The Jungle of Linear Regression Revisited, *Leonardo Electronic Journal of Practices and Technologies* 6(10):169-187. http://lejpt.academicdirect.org/A10/169_187.pdf
(Jäntschi & Bolboacă, 2007-Modeling): Jäntschi L, Bolboacă SD, 2007. Modeling the Octanol-Water Partition Coefficient of Substituted Phenols by the Use of Structure Information. *International Journal of Quantum Chemistry* 107(8):1736-1744. http://dx.doi.org/10.1002/qua.21292
(Jäntschi & Bolboacă, 2007-Results): Jäntschi L, Bolboacă SD, 2007. Results from the Use of Molecular Descriptors Family on Structure Property/Activity Relationships. *International Journal of Molecular Sciences* 8(3):189-203. www.mdpi.org/ijms/papers/i8030189.pdf
(Jäntschi & Bolboacă, 2008-Marine): Jäntschi L, Bolboacă SD, 2008. A structural modelling study on marine sediments toxicity. *Mar Drugs* 6(2):372-388. www.mdpi.com/1660-3397/6/2/372/pdf
(Jäntschi & Bolboacă, 2008-rET108): Jäntschi L, Bolboacă SD. 2008a. Processes at phases interface: Mathematical modelling, Numerical optimization, Web implementation, with applications for Separation and characterization of Biological active compounds series (MEC/5892/18.09.2006) - Research Results Synthesis (in Romanian). UEFISCSU-ET108-2008. http://lori.academicdirect.org/research/grants/Raport_ET108_2008.pdf
(Jäntschi & Bolboacă, 2009-Manuscript): Jäntschi L, Bolboacă SD. 2009. A genetic algorithm for structure-activity relationships: software implementation, ArXiv manuscript, 26 Jun 2009 (21p.). http://arxiv.org/abs/0906.4846
(Jäntschi & Diudea, 2000): Jäntschi L, 2000. Predicția proprietăților fizico-chimice şi biologice cu ajutorul descriptorilor matematici. Teză de doctorat (Chimie) - coordonator Prof. Diudea MV. Universitatea "Babeş-Bolyai" din Cluj-Napoca, Cluj-Napoca, Ian.2000. http://aleph.bibnat.ro:8991/F/?func=full-set-set&set_number=132110&set_entry=000015
(Jäntschi & Diudea, 2009): Jäntschi L, Diudea MV. 2009. Subgraphs of Pair Vertices. *J Math Chem* 45(2):364-371. Jantschi&Diudea_2009.pdf
(Jäntschi & others, 2009-DistFit1): Jäntschi L, 2009. Distribution Fitting 1. Parameters Estimation under Assumption of Agreement between Observation and Model, Bulletin UASVM Horticulture, 66(2):684-690. Jantschi_2009_DistFit1.pdf
(Jäntschi şi Bolboacă, 2008-Posters): Jäntschi L, Bolboacă SD, 2008. Embedded Molecular Geometry and Molecular Topology Approach for SARs. Strasbourg Summer School on Chemoinformatics, Louis Pasteur University, June 22-25, Obernai, France (Poster #8). Jäntschi L, Bolboacă SD, 2008. Embedded Molecular Geometry and Molecular Topology Approach for SARs. Summer School on Neural Networks in Classification, Regression and Data Mining, Polytechnic School of Engineering of Porto, 07-11 July, Porto, Portugal

(Poster #10). http://lori.academicdirect.org/conferences/work_list.php?user=lori&id=188
(Jäntschi, 1999): Jäntschi L, 1999. Fragmental Property Indices In: *Molecular Topology* (Diudea MV, Gutman I, Jäntschi L), New York: Nova Science, p. 197-232. http://lccn.loc.gov/2001031282
(Jäntschi, 2003): Jäntschi L, 2003. Kinetic Biochemistry. *Leonardo J Sci* 2(2):1-40. http://ljs.academicdirect.org/A02/01_40.pdf
(Jäntschi, 2004): Jäntschi L, 2004. MDF - A New QSAR/QSPR Molecular Descriptors Family. *Leonardo Journal of Sciences* 3(4):68-85. http://ljs.academicdirect.org/A04/68_85.pdf
(Jäntschi, 2004): Jäntschi L, 2004. MDF - A New QSAR/QSPR Molecular Descriptors Family. *Leonardo Journal of Sciences* 3(4):68-85. http://ljs.academicdirect.org/A04/68_85.pdf
(Jäntschi, 2005): Jäntschi L, 2005. Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology. *Leonardo Electronic Journal of Practices and Technologies* 6:76-98. http://lejpt.academicdirect.org/A11/163_180.pdf
(Jarque & Bera, 1980): Jarque CM, Bera AK, 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ Lett* 6(3):255-259. Jarque&Bera_1980.pdf
(Jarque & Bera, 1981): Jarque CM, Bera AK, 1981. Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Econ Lett* 7(4):313-318. http://dx.doi.org/10.1016/0165-1765(81)90035-5
(Jong & Holland, 1975): de Jong KA, 1975. An analysis of the behaviour of a class of genetic adaptive systems. PhD Thesis (Computer and Communication Sciences) - Supervisor Prof. Holland JH. University of Michigan, Ann Arbor, MI, USA. PhD_Jong&Holland_1975.pdf
(Kadirvelraj & alții, 2006): Kadirvelraj R, Gonzalez-Outeirino J, Foley BL, Beckham ML, Jennings HJ, Foote S, Ford MG, Woods RJ, 2006. Understanding the bacterial polysaccharide antigenicity of Streptococcus agalactiae versus Streptococcus pneumoniae. *Proc Natl Acad Sci USA* 103(21):8149-8154. http://www.pnas.org/content/103/21/8149.full
(Kallache & alții, 2005): Kallache M, Rust HW, Lange H, Kropp J, 2005. Extreme Value Analysis considering Trends: Methodology and Application to Runoff Data of the River Danube Catchment. R&D Research Project: Scaling analysis of hydrometeorological time series data. *Potsdam Workshop*, December 12-13. Kallache&others_2005.pdf
(Kapou & alții, 2008): Kapou A, Benetis N-P, Durdagi S, Nikolaropoulos S, Mavromoustakos T, 2008. 3D QSAR/CoMFA and CoMSIA Studies on Antileukemic Steroidal Esters Coupled with Conformationally Flexible Nitrogen Mustards. *Journal of Chemical Information and Modeling* 48(11):2254-2264. http://dx.doi.org/10.1021/ci800240m
(Karlin & Altschul, 1990): Karlin S, Altschul SF, 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc Natl Acad Sci USA* 87(6):2264-2268. http://www.pnas.org/content/87/6/2264.full.pdf+html
(Kellogg & alții, 1991): Kellogg GE, Semus SF, Abraham DJ, 1991. HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J Comput-Aided Mol Des* 5(6):545-552. http://dx.doi.org/10.1007/BF00135313
(Khan, 2009): Khan LA, 2009. Selecting Flow Frequency Distributions for Designing a Small Low-Head Dam Removal Cofferdam. Journal of Hydrologic Engineering 14(9):935-943. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000073
(Kikuchi & alții, 2003): Kikuchi S, Tominaga D, Arita M, Takahashi K, Tomita M, 2003. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics* 19(5):643-650. Kikuchi&others_2003.pdf
(Kjellström, 1991): Kjellström G, 1991. On the Efficiency of Gaussian Adaptation. *J Optim Theor Appl* 71(3):589-597. http://dx.doi.org/10.1007/BF00941405
(Knowles & Corne, 2002): Knowles JD, 2002. Local Search and Hybrid Evolutionary Algorithms for Pareto Optimization. PhD Thesis (Computer Science) - Supervisor Prof.

Corne D. University of Reading, Reading, UK. PhD_Knowles&Corne_2002.pdf
(Koba & alții, 2010): Koba M, Baczek T, Macur K, Bober L, Frackowiak T, Bucinski A, Rystok-Grabska D, Stasiak J, Koba K, 2010. Factor analysis of microbiological activity data and structural parameters of antibacterial quinolones. *Journal of Molecular Modeling* 16(2):327-335. http://dx.doi.org/10.1007/s00894-009-0549-3
(Kobti & alții, 2004): Kobti Z, Reynolds RG, Kohler T, 2004. Agent-Based Modeling of Cultural Change in Swarm Using Cultural Algorithms, *SwarmFest* 8:8p. http://l.academicdirect.org/Horticulture/Gas/Refs/Kobti_others_2004.pdf
(Kolmogorov, 1941): Kolmogorov A, 1941. Confidence Limits for an Unknown Distribution Function. *The Annals of Mathematical Statistics* 12(4):461-463. Kolmogorov_1941.pdf
(Krogh & alții, 1994): Krogh A, Brown M, Mian IS, Sjolander K, Haussler D, 1994. Hidden Markov Models in computational biology applications to protein modeling. *J Mol Biol* 235(5):1501-1531. Krogh&others_1994.pdf
(Kumar şi alții, 2007): Kumar M, Thurow K, Stoll N, Stoll R, 2007. Robust fuzzy mappings for QSAR studies. *European Journal of Medicinal Chemistry* 42(5):675-685. Kumar&others_2007.pdf
(Kuo & Liu, 2003): Kuo S-F, Liu C-W, 2003. Simulation and optimization model for irrigation planning and management. *Hydrol Process* 17(15):3141-3159. Kuo&Liu_2003.pdf
(Kysely & Dubrovsky, 2005): Kyselý J, Dubrovský M, 2005. Simulation of extreme temperature events by a stochastic weather generator: Effects of interdiurnal and interannual variability reproduction. *Int J Climatol* 25(2):251-269. http://direct.bl.uk/research/1C/17/RN163638367.html
(Laarhoven & Aarts, 1987): van Laarhoven PJM, Aarts EHL, 1987. Simulated Annealing: Theory and Applications. Dordrecht: D. Reidel. (198 p.). http://lccn.loc.gov/87009666
(Lamarck, 1809): Lamarck JBPAM, 1809. Philosophie zoologique, ou Exposition des considérations relatives à l'histoire naturelle des animaux; à la diversité de leur organisation et des facultés qu'ils on obtiennent; aux causes physiques qui maintiennent en eux la vie et donnent lieu aux mouvemens qu'ils exécutent; enfin, à celles qui produisent les unes le sentiment et les autres l'intelligence de ceux qui en sont douse, Paris: Dentu; 2 volumes (422 p.; 450 p.). Lamarck_1809.pdf
(Larkin & alții, 2007): Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG, 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947-2948. Larkin&others_2007.pdf
(Larsen & alții, 2006): Larsen FH, van den Berg F, Engelsen SB, 2006. An exploratory chemometric study of 1H-NMR spectra of table wines. *J Chemometr* 20(5):198-208. Larsen&others_2006.pdf
(Lathrop, 1994): Lathrop RH, 1994. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Engng* 7(9):1059-1068. Lathrop_1994.pdf
(Lau & Nathans, 1987): Lau LF, Nathans D. 1987. Expression of a set of growth-related immediate early genes in BALB/c 3T3 cells: coordinate regulation with c-fos or c-myc. *Proc Natl Acad Sci USA* 84(5):1182-1186. http://www.pnas.org/content/84/5/1182.full.pdf+html
(Lee & alții, 1999): Lee B, Yen J, Yang L, Liao JC, 1999. Incorporating Qualitative Knowledge in Enzyme Kinetic Models Using Fuzzy Logic. *Biotechnol Bioeng* 62(6):722-729. Lee&others_1999.pdf
(Lemmon & Milinkovitch, 2002): Lemmon AR, Milinkovitch MC, 2002. The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. *Proc Natl Acad Sci USA* 99(16):10516-10521. http://www.pnas.org/content/99/16/10516.full
(Letort & alții, 2008): Letort V, Mahe P, Cournède P-H, De Reffye P, Courtois B, 2008. Quantitative genetics and functional-structural plant growth models: Simulation of

quantitative trait loci detection for model parameters and application to potential yield optimization. *Annals Botany* 101(8):1243-1254. Letort&others_2008.pdf
(Lewis, 1998): Lewis PO, 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol Biol Evol* 15(3):277-283. Lewis_1998.pdf
(Lewis, 2001): Lewis PO, 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol* 50(6):913-925. Lewis_2001.pdf
(Li & alții, 2007): Li J, Liu H, Yao X, Liu M, Hu Z, Fan B, 2007. Quantitative structure-activity relationship study of acyl ureas as inhibitors of human liver glycogen phosphorylase using least squares support vector machines. *Chemometrics and Intelligent Laboratory Systems* 87(2):139-146. Li&others_2007.pdf
(Lindenmayer, 1968): Lindenmeyer A, 1968. Mathematical models for cellular interaction in development. *J Theor Biol* 18:280-315. http://dx.doi.org/10.1016/0022-5193(68)90079-9
(Linnaei, 1735): Linnaei C, 1735. Systema Naturae. The Netherlands (11p.). http://lccn.loc.gov/66045627
(Lipinski & alții, 1997&2001): Lipinski CA, Lombardo F, Dominy BW, Feeney PJ, 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Del Rev* 23(1-3):3-25. http://dx.doi.org/10.1016/S0169-409X(96)00423-1 & Lipinski CA, Lombardo F, Dominy BW, Feeney PJ, 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings[1]. *Adv Drug Del Rev* 46(1-3):3-26. http://dx.doi.org/10.1016/S0169-409X(00)00129-0
(Liu & alții, 2001): Liu J, Goering CE, Tian L, 2001. A neural network for setting target corn yields. *Trans Am Soc Agric Eng* 44(3):705-713. http://asae.frymulti.com/abstract.asp?aid=6097&t=2
(Luis & alții, 2010): Luis P, Garea A, Irabien A, 2010. Quantitative structure-activity relationships (QSARs) to estimate ionic liquids ecotoxicity EC50 (Vibrio fischeri), *Journal of Molecular Liquids* 152(1-3):28-33. http://dx.doi.org/10.1016/j.molliq.2009.12.008
(Lundmark & alții, 2007): Lundmark AM, Corfu F, Spürgin S, Selbekk RS, 2007. Proterozoic evolution and provenance of the high-grade Jotun Nappe Complex, SW Norway: U-Pb geochronology. *Precambrian Res* 159(3-4):133-154. Lundmark&others_2007.pdf
(Marichal, 2000): Marichal J-L, 2000. On an axiomatization of the quasi-arithmetic mean values without the symmetry axiom. Aequationes Mathematicae, 59(1-2):74-83. Marichal_2000.pdf
(Martin & Spears, 2001): Martin WN, Spears WM, 2001. Foundations of genetic algorithms 6. San Francisco: Morgan Kaufmann (342 p.). http://lccn.loc.gov/2001275761
(Masi & alții, 2005): de Masi L, Castaldo D, Galano G, Minasi P, Laratta B, 2005. Genotyping of fig (Ficus carica L) via RAPD markers. *J Sci Food Agr* 85(13):2235-2242. Masi&others_2005.pdf
(Matthews & Kraw, 2001): Matthews KB, 2001. Applying Genetic Algorithms to Multi-objective Land-Use Planning. PhD Thesis (Agriculture) - Supervisor Prof. Kraw S. Robert Gordon University, Craigiebuckler, Aberdeen, NSW, Australia. PhD_Matthews&Kraw_2001.pdf
(Mayer & alții, 1999): Mayer DG, Belward JA, Widell H, Burrage K, 1999. Survival of the fittest - Genetic algorithms versus evolution strategies in the optimization of systems models. *Agr Syst* 60(2):113-122. http://dx.doi.org/10.1016/S0308-521X(99)00022-0
(McGaughey & Mewshaw, 1998): McGaughey GB, Mewshaw RE, 1998. Molecular Modeling and the Design of Dopamine D2 Partial Agonists. *Network Science* Feature 20 (May 1998). http://www.netsci.org/Science/Compchem/feature20.html
(Mendel, 1866): Mendel JG, 1866. Versuche über Plflanzenhybriden Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr, Abhandlungen:3-47(english translation: Druery CT & Bateson W, 1901. Experiments in plant hybridization, *Journal of the Royal Horticultural Society* 26: 1-32. Mendel_1865.pdf

(Metropolis & Ulam, 1949): Metropolis N, Ulam S, 1949. The Monte Carlo Method. *Journal of the American Statistical Association* 44(247):335-341. www.jstor.org/pss/2280232
(Micheli & alții, 2007): Micheli A, Sperduti A, Starita A, 2007. An introduction to recursive neural networks and Kernel methods for cheminformatics. *Current Pharmaceutical Design* 13(14):1469-1495. Micheli&others_2007.pdf
(Montazar & alții, 2008): Montazar A, Rahimikob A, 2008. Optimal Water Productivity of Irrigation Networks in Arid And Semi-Arid Regions. *Irrig Drain* 57(4):411-423. Montazar&others_2008.pdf
(Moore & alții, 2003): Moore AB, Morris KP, Blackwell GK, Jones AR, Sim PC, 2003. Using geomorphological rules to classify photogrammetrically-derived digital elevation models. *Int J Remote Sens* 24(13):2613-2626. Moore&others_2003.pdf
(Moore & alții, 2007): Moore MJ, Bell CD, Soltis PS, Soltis DE, 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc Natl Acad Sci USA* 104(49):19363-19368. http://www.pnas.org/content/104/49/19363.full
(Morgan & alții, 1915): Morgan TH, Sturtevant AH, Muller HJ, Bridges CB, 1915. The mechanism of Mendelian heredity, New York: Henry Holt & Co (270 p.). Morgan_others_1915.pdf
(Morimoto & alții, 1997): Morimoto T, Purwanto W, Suzuki J, Hashimoto Y, 1997. Optimization of heat treatment for fruit during storage using neural networks and genetic algorithms. *Comput Electron Agr* 19(1):87-101. http://dx.doi.org/10.1016/S0168-1699(97)00037-9
(Mouchlis & alții, 2010): Mouchlis VD, Mavromoustakos TM, Kokotos G, 2010. Design of new secreted phospholipase A2 inhibitors based on docking calculations by modifying the pharmacophore segments of the FPL67047XX inhibitor. *Journal of Computer-Aided Molecular Design* 24(2):107-115. http://dx.doi.org/10.1007/s10822-010-9319-7
(Muller, 1928): Muller HJ, 1928. The Production of Mutations by X-Rays. *Proc Natl Acad Sci USA* 14(9):714-726. Muller_1928.pdf
(Mullins & alții, 1984): Mullins MD, Pochini CM, McCrindle S, Romkes M, Safe SH, Safe LM, 1984. High resolution PCB analysis: synthesis and chromatographic properties of all 209 PCB congeners. *Environ Sci Technol* 18:468-476. http://dx.doi.org/10.1021/es00124a014
(Myers & alții, 1994): Myers AM, Charifson PS, Owens CE, Kula NS, McPhail AT, Baldessarini RJ, Booth RG, Wyrick SD, 1994. Conformational Analysis, Pharmacophore Identification, and Comparative Molecular Field Analysis of Ligands for the Neuromodulatory .sigma.3 Receptor, *J Med Chem* 37(24):4109-4117. www.ncbi.nlm.nih.gov/pubmed/7990111
(Narasimhan & alții, 2006): Narasimhan B, Mourya V, Dhake A, 2006. Design, synthesis, antibacterial, and QSAR studies of myristic acid derivatives. *Bioorganic and Medicinal Chemistry Letters* 16(11):3023-3029. Narasimhan&others_2006.pdf
(Nelder & Mead, 1964): Nelder JA, Mead R, 1964. A simplex method for function minimization. *Comput J* 7:308-313. Nelder&Mead_1964.pdf
(Nguyen & Liu, 2008): Nguyen N-K, Liu M-Q, 2008. An algorithmic approach to constructing mixed-level orthogonal and near-orthogonal arrays. *Comput Stat Data Anal* 52(12):5269-5276. Nguyen&Liu_2008.pdf
(Notredame & alții, 1997): Notredame C, O'Brien EA, Higgins DG, 1997. RAGA: RNA sequence alignment by genetic algorithm. *Nucleic Acid Res* 25(22):4570-4580. Notredame&others_1997.pdf
(Notredame & Higgins, 1996): Notredame C, Higgins DG. 1996. SAGA: Sequence alignment by genetic algorithm. *Nucleic Acids Research* 24(8):1515-1524. Notredame&Higgins_1996.pdf
(Opriș & Diudea, 2001): Opriș DM, Diudea MV, 2001. Peptide Property Modeling by Cluj Indices. *SAR and QSAR in Environmental Research* 12(1-2):159-179.

Opris&Diudea_2001.pdf
(Ortiz-Martinez & alții, 2008): Ortiz-Martınez T, Rico-Gray V, Martınez-Meyer E, 2008. Predicted and verified distributions of Ateles geoffroyi and Alouatta palliata in Oaxaca, Mexico. *Primates* 49(3):186-194. Ortiz-Martinez&others_2008.pdf
(Osman & McManus, 2007): Osman NY, 2007. The Development of a Predictive Damage Condition Model of Light Structures on Expansive Soils using Hybrid Artificial Intelligence Techniques. PhD Thesis (Engineering and Industrial Sciences) - Supervisor Prof. McManus KAM. Swinburne University of Technology, Melbourne, VIC, Australia PhD_Osman&McManus_2007.pdf
(Pan & alții, 2002): Pan K-H, Lih C-J, Cohen SN. 2002. Analysis of DNA microarrays using algorithms that employ rule-based expert knowledge. *Proc Natl Acad Sci USA* 99(4):2118-2123. http://www.pnas.org/content/99/4/2118.full
(Panait & Luke, 2006): Panait L, 2006. The Analysis and Design of Concurrent Learning Algorithms for Cooperative Multiagent Systems. PhD Thesis (Computer Science) - Supervisor Assist. Prof. Luke S. George Mason University, Fairfax, VA, USA. PhD_Panait&Luke_2006.pdf
(Patterson, 1928): Patterson JT, 1928. The effects of X-rays in producing mutations in the somatic cells of Drosophila melanogaster. *Science* 68(1750):41-43. http://dx.doi.org/10.1126/science.68.1750.41
(Pearson, 1895): Pearson K, 1895. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society of London. A* 186:343-414.
(Pearson, 1896): Pearson K, 1896. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Phil Trans R Soc Lond A* 187:253-318. http://dx.doi.org/doi:10.1098/rsta.1896.0007
(Pearson, 1900): Pearson K, 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine 5th Ser* 50:157-175. Pearson_1900.pdf
(Phien & Hira, 1983): Phien HN, Hira MA, 1983. Log Pearson Type-3 Distribution: Parameter Estimation. *Journal of Hydrology* 64:25-37. Phien&Hira_1983.pdf
(Pizarro & alții, 2001): Pizarro C, Gonzalez-Saiz JM, Garrido-Vidal D, 2003. Kinetic modelling of acetic fermentation in an industrial process by genetic algorithms with a desirability function. *J Chemometr* 17(8-9):453-462. Pizarro&others_2001.pdf
(Plant, 2004): Plant N, 2004. Strategies for using in vitro screens in drug metabolism. *Drug Discov Today* 9:328-336. Plant_2004.pdf
(Plauchithiu & alții, 1978): Plauchithiu MG, Vlad A, Simon Z. 1978. Mapping of the alpha adrenergic receptor by the minimal steric difference technique. *Revue Roumaine de Biochimie* 15(3):229-235. Plachitiu&others_1978.pdf
(Posada & Crandall, 1998): Posada D, Crandall KA, 1998. MODELTEST: Testing the model of DNA substitution. *Bioinformatics* 14(9):817-818. Posada&Crandall_1998.pdf
(Potter & Jong, 1997): Potter MA, 1997. The Design and Analysis of a Computational Model of Cooperative Coevolution. PhD Thesis (Computer Science) - Supervisor Assoc. Prof. de Jong KA. George Mason University, Fairfax, VA, USA. PhD_Potter&Jong_1997.pdf
(Prügel-Bennett, 2001): Prügel-Bennett A, 2001. The Mixing Rate of Different Crossover Operators. In: *Foundations of genetic algorithms 6* (Martin WN, Spears WM). San Francisco: Morgan Kaufmann, p. 261-274. http://lccn.loc.gov/2001275761
(Putz & Lacrămă, 2007): Putz MV, Lacrămă A-M, 2007. Introducing Spectral Structure Activity Relationship (S-SAR) Analysis. Application to Ecotoxicology. *International Journal of Molecular Sciences* 8(5):363-469. http://www.mdpi.com/1422-0067/8/5/363/pdf
(Randic & Pompe, 2005): Randic M, Pompe M, 2005. Retro-regression - A way to resolve

multivariate regression ambiguities. Acta Chimica Slovenica 52(4):408-416. Randic&Pompe_2005.pdf
(Rao & alții, 2009): Rao HB, Yang GB, Tan NX, Li P, Li ZR, Li XY, 2009. Prediction of HIV-1 Protease Inhibitors Using Machine Learning Approaches. *QSAR & Combinatorial Science* 28(11-12):1346-1357. http://dx.doi.org/10.1002/qsar.200960021
(Rault & alții, 1992): Rault S, Bureau R, Pilo J-C, Robba M, 1992. Comparative molecular field analysis of CCK-A antagonists using field-fit as an alignment technique. A convenient guide to design new CCK-A ligands. *Journal of Computer-Aided Molecular Design* 6(6):553-568. http://dx.doi.org/10.1007/BF00126214
(Rényi, 1961): Rényi A, 1961. On measures of entropy end information. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960*. p. 547-561. Renyi_1961.pdf
(Rienzo & alții, 2004): de Rienzo F, Gabdoulline RR, Wade RC, Sola M, Menziani MC, 2004. Computational approaches to structural and functional analysis of plastocyanin and other blue copper proteins. *Cellular and Molecular Life Sciences* 61(10):1123-1142. Rienzo&others_2004.pdf
(Robbins, 1918): Robbins RB, 1918. Some applications of mathematics to breeding problems. III. Genetics 3:375-389. Robbins_1918.pdf
(Rodkaew & alții, 2004): Rodkaew Y, Chuaiaree S, Siripant S, Lursinsap C, Chongstitvatana P, 2004. Animating plant growth in L-system by parametric functional symbols. *Int J Intell Syst* 19(1-2):9-23. Rodkaew&others_2004.pdf
(Ronquist & Huelsenbeck, 2003): Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572-1574. Ronquist&Huelsenbeck_2003.pdf
(Ryder & alții, 1988): Ryder K, Lau LF, Nathans D, 1988. A gene activated by growth factors is related to the oncogene v-jun. *Proc Natl Acad Sci USA* 85(5):1487-1491. http://www.pnas.org/content/85/5/1487.full.pdf
(Sacan & alții, 2009): Saçan MT, Çeçen F, Ertürk MD, Semerci N, 2009. Modelling the relative toxicity of metals on respiration of nitrifiers using ion characteristics. *SAR and QSAR in Environmental Research* 20(7):727-740. http://dx.doi.org/10.1080/10629360903438594
(Saliner & alții, 2006): Saliner AG, Netzeva TI, Worth AP, 2006. Prediction of estrogenicity: Validation of a classification model. *SAR and QSAR in Environmental Research* 17(2):195-223. Saliner&others_2006.pdf
(Salomez & Hofman, 2007): Salomez J, Hofman G, 2007. A Soil Temperature/Short-Wave Radiation Growth Model for Butterhead Lettuce Under Protected Cultivation in Flanders. *J Plant Nutr* 30(3):397-410. Salomez&Hofman_2007.pdf
(Sanderson, 2002): Sanderson MJ, 2002. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. *Mol Biol Evol* 19:101-109. Sanderson_2002.pdf
(Sanderson, 2003): Sanderson MJ, 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301-302. Sanderson_2003.pdf
(Sârbu & alții, 2004): Sârbu C, Casoni D, Darabantu M, Maiereanu C, 2004. Quantitative structure-retention and retention-activity relationships of some 1,3-oxazolidine systems by RP-HPTLC and PCA. Journal of Pharmaceutical and Biomedical Analysis 35(1):213-219. http://dx.doi.org/10.1016/j.jpba.2004.01.002
(Sârbu & Jäntschi, 1998): Sârbu C, Jäntschi L, 1998. Statistic Validation and Evaluation of Analytical Methods by Comparative Studies. I. Validation of Analytical Methods using Regression Analysis, Revista de Chimie, 49(1):19-24. Sarbu&Jantschi_1998.pdf
(Sarmiento-Monroy & Sharkey, 2006): Sarmiento-Monroy CE, 2006. Taxonomic revision of zelomorpha ashmead, 1900 and hemichoma enderlein, 1920 (braconidae: agathidinae) with

a phylogenetic analysis of color patterns. PhD Thesis (Entomology) - Supervisor Prof. Sharkey MJ. University of Kentucky, Lexington, KY, USA. PhD_Sarmiento-Monroy&Sharkey_2006.pdf

(Saski & alții, 2007): Saski C, Lee S-B, Fjellheim S, Guda C, Jansen RK, Luo H, Tomkins J, Rognli OA, Daniell H, Clarke JL, 2007. Complete chloroplast genome sequences of Hordeum vulgare, Sorghum bicolor and Agrostis stolonifera, and comparative analyses with other grass genomes. *Theor Appl Genet* 115(4):571-590,591. Saski&others_2007.pdf

(Sastry & alții, 2007): Sastry KM, 2007. Genetic Algorithms and Genetic Programming for Multiscale Modeling: Applications in Materials Science and Chemistry and Advances in Scalability. PhD Thesis (Systems and Entrepreneurial Engineering) - Supervisor Prof. Goldberg DE & Johnson DD. University of Illinois at Urbana-Champaign, Urbana, IL, USA. PhD_Sastry&altii_2007.pdf

(Savageau, 1976): Savageau MA, 1976. Biochemical System Analysis: a Study of Function and Design in Molecular Biology. Reading: Addison-Wesley, (199 p.). http://lccn.loc.gov/76042469

(Scalliet & alții, 2008): Scalliet G, Piola F, Douady CJ, Rety S, Raymond O, Baudino S, Bordji K, Bendahmane M, Dumas C, Cock JM, Hugueney P, 2008. Scent evolution in Chinese roses. *Proc Natl Acad Sci USA* 105(15):5927-5932. http://www.pnas.org/content/105/15/5927.full

(Scholz & Stephens, 1986): Scholz FW, Stephens MA, 1986. K-sample Anderson-Darling Tests of Fit, for Continuous and Discrete Cases. *Technical Report* No:81 (May 1986). Scholz&Stephens_1986.pdf

(Scholz & Stephens, 1987): Scholz FW, Stephens MA, 1987. K-sample Anderson-Darling Tests. *Journal of the American Statistical Association* 82(399):918-924. http://www.jstor.org/pss/2288805

(Schultz & alții, 2006): Schultz A-K, Zhang M, Leitner T, Kuiken C, Korber B, Morgenstern B, Stanke M, 2006. A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. *BMC Bioinform* 7:265(15). Schultz&others_2006.pdf

(Schultz & Brady, 2008): Schultz TR, Brady SG. 2008. Major evolutionary transitions in ant agriculture. *Proc Natl Acad Sci USA* 105(14):5435-5440. http://www.pnas.org/content/105/14/5435.full

(Schuster & alții, 2006): Schuster D, Steindl TM, Langer T, 2006. Predicting drug metabolism induction in silico. *Curr Top Med Chem* 6(15):1627-40. http://dx.doi.org/10.2174/156802606778108924

(Schwefel, 1995): Schwefel H-P, 1995. Evolution and Optimum Seeking. New York: Wiley & Sons (456 p.). http://lccn.loc.gov/94022270

(Seginer & alții, 2007): Seginer I, Ioslovich I, Bux M, 2007. Optimal Control of Solar Sludge Dryers. *Dry Technol* 25(2):401-415. http://dx.doi.org/10.1080/07373930601184577

(Seginer & Ioslovich, 1999): Seginer I, Ioslovich I, 1999. Optimal spacing and cultivation intensity for an industrialized crop production system. *Agr Syst* 62(3):143-157. http://dx.doi.org/10.1016/S0308-521X(99)00057-8

(Shannon, 1948): Shannon CE, 1948. A Mathematical Theory of Communication. The Bell System Technical Journal, 27:379-423 & 27:623-656. Shannon_1948.pdf

(Sharma & alții, 2010): Sharma BK, Sarbhai K, Singh P, 2010. A rationale for the activity profile of arylpiperazinylthioalkyls as 5-HT1A-serotonin and alpha(1)-adrenergic receptor ligands. *European Journal of Medicinal Chemistry* 45(5):1927-1934. http://dx.doi.org/10.1016/j.ejmech.2010.01.034

(Shevchenko & alții, 2001): Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, Standing KG, 2001. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time of flight mass spectrometry and BLAST homology searching. *Anal Chem* 73(9):1917-1926. http://dx.doi.org/10.1021/ac0013709

(Simon & Bohl, 1992): Simon Z, Bohl M, 1992. Structure-activity relations in gestagenic steroids by the MTD method. The case of hard molecules and soft receptors, *Quantitative Structure-Activity Relationships* 11(1):23-28. http://dx.doi.org/10.1002/qsar.19920110104
(Simpson, 1949): Simpson HE, 1949. Measurement of diversity. *Nature* 163:688. Simpson_1949.pdf
(Sirdas & Sahin, 2008): Sirdas S, Sahin AD, 2008. Relationship between drought and solar irradiation variables. *Hydrol Process* 22(10):1460-1472. Sirdas&Sahin_2008.pdf
(Skolicki & Jong, 2007): Skolicki ZM, 2007. An Analysis of Island Models in Evolutionary Computation. PhD Thesis (Computer Science) - Supervisor Prof. de Jong KA. George Mason University, Fairfax, VA, USA. PhD_Skolicki&Jong_2007.pdf
(Smirnov, 1948): Smirnov NV, 1948. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics* 19(2):279-281. Smirnov_1948.pdf
(Smith & alții, 1994): Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM, 1994. The genetic data environment and GUI for multiple sequence analysis. *Comput Appl Biosci* 10(6):671-675. Smith&others_1994.pdf
(Smith, 2001): Smith P, 2001. Priority setting in agricultural research: Beyond economic surplus methods. *Public Admin Devel* 21(5):419-428. Smith_2001.pdf
(Smith, 2007): Smith JE, 2007. Coevolving Memetic Algorithms: A Review and Progress Report. *IEEE Trans Syst Man Cy B* 37(1):6-17. http://dx.doi.org/10.1109/TSMCB.2006.883273
(Solan, 2009): Solan E, 2009. Stochastic Games. In: *Encyclopedia of Database Systems*. Berlin: Springer (p. 8698-8708). Solan_2009.pdf
(Spears & Jong, 1998): de Spears WM, 1998. The role of mutation and recombination in evolutionary algorithms. PhD Thesis (Computer Science) - Supervisor Assoc. Prof. de Jong KA. George Mason University, Fairfax, VA, USA. PhD_Spears&Jong_1998.pdf
(Spears, 2001): Spears WM, 2001. The Equilibrium and Transient Behavior of Mutation and Recombination. In: Foundations of genetic algorithms 6 (Martin WN, Spears WM). San Francisco: Morgan Kaufmann, p. 241-260. http://lccn.loc.gov/2001275761
(Srivastava & alții, 2008): Srivastava V, Kumar A, Mishra BN, Siddiqi MI, 2008. Molecular docking studies on DMDP derivatives as human DHFR inhibitors. *Bioinformation* 3(4):180-188. www.bioinformation.net/003/004000032008.pdf
(Steele & alții, 2005): Steele M, Chaseling J, Hurst C, 2005. Simulated Power of the Discrete Cramer-von Mises Goodness-of-Fit Tests. *International Congress on Modelling and Simulation - Advances and Applications for management and decision making MODSIM* 2005:1300-1304. Steele&others_2005.pdf
(Steele & alții, 2005): Steele M, Chaseling J, Hurst C, 2005. Simulated Power of the Discrete Cramer-von Mises Goodness-of-Fit Tests. *International Congress on Modelling and Simulation. Advances and Applications for management and decision making. MODSIM* 2005:1300-1304. Steele&others_2005.pdf
(Steiger, 1980): Steiger JH, 1980. Tests for comparing elements of a correlation matrix. *Psychol Bull* 87:245-251. Steiger_1980.pdf
(Stender & alții, 1994): Stender J, Hillebrand E, Kingdon J, 1994. Genetic algorithms in optimisation, simulation, and modelling. Amsterdam & Washington DC: IOS Press (261p.). http://lccn.loc.gov/94077520
(Stoenoiu & alții, 2008): Stoenoiu CE, Bolboacă SD, Jäntschi L, 2008. Model Formulation & Interpretation - From Experiment to Theory. *Int J Pure Appl Math* 47(1):9-16. Stoenoiu&others_2008.pdf
(Suzuki & Iwasa, 1998): Suzuki H, 1998. Evolution of a Novel Function Fascilitated by Genetic Recombination in Genetic Algorithms. PhD Thesis (Biology) - Supervisor Prof. Iwasa Y. Kyushu University, Fukuoka, Fukuoka, Japan. PhD_Suzuki&Iwasa_1998.pdf
(Taguchi, 1995): Taguchi G, 1995. Quality engineering (Taguchi methods) for the development of electronic circuit technology. *IEEE Transactions on Reliability* 44(2):225-

229. http://dx.doi.org/10.1109/24.387375
(Tao & Wang, 2007): Tao J, Wang N, 2007. DNA computing based RNA genetic algorithm with applications in parameter estimation of chemical engineering processes. *Comput Chem Eng* 31(12):1602-1618. Tao&Wang_2007.pdf
(Tiptara & alții, 2008): Tiptara P, Petsom A, Roengsumran S, Sangvanich P. 2008. Hemagglutinating activity and corresponding putative sequence identity from Curcuma aromatica rhizome. *J Sci Food Agr* 88(6):1025-1034. Tiptara&others_2008.pdf
(Todeschini & alții, 1994): Todeschini R, Lasagni M, Marengo E, 1994. New Molecular Descriptors for 2D and 3D Structures. *Theory Journal of Chemometrics* 8:263-272. http://dx.doi.org/10.1002/cem.1180080405
(Todeschini & alții, 1995): Todeschini R, Gramatica P, Provenzani R, Marengo E, 1995. Weighted Holistic Invariant Molecular (WHIM) descriptors. Part2. There Development and Application on Modeling Physico-chemical Properties of Polyaromatic Hydrocarbons. *Chemometrics and Intelligent Laboratory Systems* 27:221-229. http://dx.doi.org/10.1016/0169-7439(95)80026-6
(Todeschini & alții, 1996): Todeschini R, Vighi M, Provenzani R, Finizio A, Gramatica P, 1996. Modeling and Prediction by Using WHIM Descriptors in QSAR Studies: Toxicity of Heterogeneous Chemicals on Daphnia Magna, *Chemosphere* 8:1527-1545. http://dx.doi.org/10.1016/0045-6535(96)00153-1
(Tominaga & alții, 1999): Tominaga D, Okamoto M, Maki Y, Watanabe S, Eguchi Y, 1999. Nonlinear Numerical Optimization Technique Based on a Genetic Algorithm for Inverse Problems: Towards the Inference of Genetic Networks. *GCB 99: German Conference on Bioinformatics* 1999:Talk14. Tominaga&others_1999.pdf
(Tominaga & alții, 2000): Tominaga D, Koga N, Okamoto M, 2000. Efficient numerical optimization algorithm based on genetic algorithm for inverse problem. *Proc Genet Evolut Comput Conf* 2000:251-258. http://sciencelinks.jp/j-east/article/200417/000020041704A0526376.php
(Tominaga & Okamoto, 1998): Tominaga D, Okamoto M, 1998. Design of canonical model complex nonlinear dynamics. In: *Computer applications in biotechnology 1998: (CAB7): horizon of bioprocess systems engineering in 21st century: a proceedings volume from the 7th IFAC International Conference, Osaka, Japan, 31 May-4 June 1998* (Yoshida T & Shioya S), p. 85-90. http://lccn.loc.gov/98046029
(Tsutsui & alții, 1999): Tsutsui S, Yamamura M, Higuchi T, 1999. Multi-parent recombination with simplex crossover in real coded genetic algorithms. *Proc Genet Evolut Comput Conf* 1999:657-664. Tsutsui&others_1999.pdf
(Tufail & Ormsbee, 2006): Tufail M, 2006. Optimal water quality management strategies for urban watersheds using macro-level simulation models linked with evolutionary algorithms. PhD Thesis (Civil Engineering) - Supervisor Prof. Ormsbee LE. University of Kentucky, Lexington, KY, USA. PhD_Tufail&Ormsbee_2006.pdf
(Tzimopoulos & alții, 2010): Tzimopoulos D, Czapik A, Gdaniec M, Bakas T, Isab AA, Varvogli A-C, Akrivos PD, 2010. Synthesis and study of triorganostannyl esters of 3-,4- and 3,5-pyridinylimino substituted aminobenzoic acids: Crystal structures of dimorphs of aqua-trimethyltin 3-pyridinyliminobenzoate, *Journal of Molecular Structure*, 965(1-3):56-64. http://dx.doi.org/10.1016/j.molstruc.2009.11.038
(Ungureşan & Niac, 2007): Ungureşan ML, Niac G, 2007. PART I. Visible Spectrum of the Copper(II)-Thiosuphate Complex. *Phys Chem Indian J* 2(1):13(4). Unguresan&Niac_2007.pdf
(Uray & alții, 2003): Uray K, Price MR, Majer Z, Vass E, Hollósi M, Hudecz F, 2003. Identification and solution conformation of multiple epitopes recognized by a MUC2 mucin-specific monoclonal antibody. *Arch Biochem Biophys* 410(2):254-260. http://dx.doi.org/10.1016/S0003-9861(02)00693-8
(Ursu & alții, 2006): Ursu O, Costescu A, Diudea MV, Parv B, 2006. QSAR modeling of

antifungal activity of some heterocyclic compounds. *Croatica Chemica Acta* 79(3):483-488. Ursu&others_2006.pdf

(Veies, 1902): de Veies H, 1902. The origin of species by mutation. *Science* 15(384):721-729. http://dx.doi.org/10.1126/science.15.384.721

(Veith & Wolfe, 2002): Veith TL, 2002. Agricultural BMP Placement for Cost-Effective Pollution Control at the Watershed Level. PhD Thesis (Biological Systems Engineering) - Supervisor Prof. Wolfe ML. Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. PhD_Veith&Wolfe_2002.pdf

(Venard & Vaillancourt, 2006): Venard CMP, 2006. The development of colletotrichum graminicola inside maize stalk tissues. PhD Thesis (Plant Pathology) - Supervisor Assoc. Prof. Vaillancourt L. University of Kentucky, Lexington, KY, USA. PhD_Venard&Vaillancourt_2006.pdf

(Vingron & Argos, 1990): Vingron M, Argos P, 1990. Determination of reliable regions in protein sequence alignments. *Protein Engng* 3(7):565-569. http://dx.doi.org/10.1093/protein/3.7.565

(Vogel & alții, 1993): Vogel RM, Thomas WO, Chiew FHS, 1993. Flood-Flow Frequency Model Selection in Southwestern United-States. Journal of Water Resources Planning and Management-ASCE 119(3):353-366. Vogel&others_1993.pdf

(Waller & alții, 1994): Waller CL, Wyrick SD, Park HM, Kemp WE, Smith FT, 1994. Conformational Analysis, Molecular Modeling, and Quantitative Structure-Activity Relationship Studies of Agents for the Inhibition of Astrocytic Chloride Transport. *Pharm Res* 11(1):47-53. http://dx.doi.org/10.1023/A:1018937425823

(Waller & Marshall, 1993): Waller CL, Marshall GR, 1993. Three-Dimensional Quantitative Structure-Activity Relationship of Angiotensin-Converting Enzyme and Thermolysin Inhibitors. II. A Comparision of CoMFA Models Incorporating Molecular Orbital Fields and Desolvation Free Energy Based on Active-Analog and Complementary-Receptor-Field Alignment Rules, *J Med Chem* 36:2390-2403. http://dx.doi.org/10.1021/ja00066a004

(Walter & Kellogg, 1996): Walter CL, Kellogg GE, 1996. Adding Chemical Information of CoMFA Models with Alternative 3D QSAR Fields. *Network Science* Feature 10 (January 1996). http://www.netsci.org/Science/Compchem/feature10.html

(Wardlaw & Bhaktikul, 2004): Wardlaw R, Bhaktikul K, 2004. Application of genetic algorithms for irrigation water scheduling. *Irrig Drain* 53(4):397-414. Wardlaw&Bhaktikul_2004.pdf

(Waterman & Vingron, 1994): Waterman MS, Vingron M, 1994. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc Natl Acad Sci USA* 91(11):4625-4628. http://www.pnas.org/content/91/11/4625.full.pdf+html

(Weismann, 1893): Weismann FLA, 1893. The Germ-Plasm. A theory of heredity, New York: Charles Scribner's Sons (496 p.). Weismann_1893.pdf

(Wiegand & Jong, 2003): Wiegand PR, 2003. An Analysis of Cooperative Coevolutionary Algorithms. PhD Thesis (Computer Science) - Supervisor Prof. de Jong KA. George Mason University, Fairfax, VA, USA. PhD_Wiegand&Jong_2003.pdf

(Wiegmann & alții, 2003): Wiegmann BM, Yeates DK, Thorne JL, Kishino H, 2003. Time Flies, a New Molecular Time-Scale for Brachyceran Fly Evolution Without a Clock. *System Biol* 52(6):745-756. Wiegmann&others_2003.pdf

(Wiese & Pajeva, 1998): Wiese M, Pajeva IL, 1998. A Comparative Molecular Field Analysis of Propafenone-type Modulators of Cancer Multidrug Resistance, *Quant Struct-Act Relat* 17(4):301-312. Wiese&Pajeva_1998.pdf

(Wikler & alții, 1998) Wikler DA, Burden FR, Watkins AJR, 1998. Atomistic Topological Indices Applied to Benzodiazepines using Various Regression Methods. *Quantitative Structure Activity Relationships* 17(1), p. 14-19. Wikler&others_1998.pdf

(Wollman & alții, 2008): Wollman R, Civelekoglu-Scholey G, Scholey JM, Mogilner A, 2008. Reverse engineering of force integration during mitosis in the Drosophila embryo.

Molec Syst Biol 4:195. Wollman&others_2008.pdf
(Wolpert & Macready, 1995&1997): Wolpert DH, Macready WG, 1995. No Free Lunch Theorems for Search. The Santa Fe Institute: Technical Report, SFI-TR-95-02-010. Wolpert DH, Macready WG, 1997. No Free Lunch Theorems for Optimization, *IEEE T Evolut Comput* 1(1):67-82. Wolpert_Macready_1995.pdf & Wolpert_Macready_1997.pdf
(Yeates, 2002): Yeates DK, 2002. Relationships of extant lower Brachycera: A quantitative synthesis of morphological characters. *Zool Scripta* 31(1):105-121. Yeates_2002.pdf
(Zaliani & Gancia, 1999): Zaliani A, Gancia E, 1999. MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies. *J Chem Inf Comput Sci* 39(3):525-533. http://dx.doi.org/10.1021/ci980211b
(Zhang & alții, 2009): Zhang J, Xiong B, Zhen XC, Zhang A, 2009. Dopamine D-1 Receptor Ligands: Where Are We Now and Where Are We Going. *Medicinal Research Reviews* 29(2):272-294. http://dx.doi.org/10.1002/med.20130
(Zhang & Ghabrial, 2006): Zhang C, 2006. Genetic diversity of bean pod mottle virus (bpmv) and development of bpmv as a vector for gene expression in soybean. PhD Thesis (Plant Pathology) - Supervisor Prof. Ghabrial SA. University of Kentucky, Lexington, KY, USA. PhD_Zhang&Ghabrial_2006.pdf
(Zhu & alții, 2008): Zhu C, Yan G, Liu Z, Gao L, 2008. An Optimization Genetic Algorithm for Image Databases in Agriculture. In: *Computer and Computing Technologies in Agriculture* (Li D), Berlin: Springer-Verlag, II:1283-1289 (736p.-I&744p.-II). http://dx.doi.org/10.1007/978-0-387-77253-0_63
(Zitzler & Thiele, 1999): Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications. PhD Thesis (Technical Sciences) - Supervisor Prof. Thiele L. Swiss Federal Institute of Technology, Zurich, Elveția. PhD_Zitzler&Thiele_1999.pdf
(Zwickl & Hills, 2006): Zwickl DJ, 2006. Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion. PhD Thesis (Biology) - Supervisor Prof. Hills DM. University of Texas at Austin, Austin, TX, USA. PhD_Zwickl&Hills_2006.pdf

Publicații

P.1

On about what Can Be Done and what Cannot Be Done with Genetic Algorithms in Phylogenetic Tree and Gene Sequence Analyses

Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ

Bulletin UASVM, Horticulture 65(1):63-70, 2008

÷ Lucrarea este de tip review din stadiul actual al cunoaşterii;

÷ Sunt sintetizate rezultatele raportate în literatura de specialitate de utilizare a algoritmilor genetici în probleme dificile de clasificare filogenetică;

÷ Sunt sintetizate principalele programe care implementează algoritmi genetici folosite în clasificare filogenetică împreună cu descrierea caracteristicii principale pe care o posedă;

÷ Este sintetizată metodologia urmată de numeroşi autori în construcția spațiului de căutare specific aplicării unui algoritm genetic şi în construcția strategiei evolutive;

P.2

Hard Problems in Gene Sequence Analysis: Classical Approaches and Suitability of Genetic Algorithms

Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ

Biotechnology & Biotechnological Equipment 23(2):1275-1280, 2009

÷ Lucrarea este de tip review din materialul şi metoda cercetării;

÷ Sunt sintetizate rezultatele raportate în literatura de specialitate de utilizare a algoritmilor genetici în probleme dificile de aliniere a secvențelor de gene;

÷ Sunt sintetizate principalele categorii predilecte de probleme dificile de aliniere a secvențelor de gene care fac apel la algoritmi genetici;

÷ Sunt prezentate caracteristicile unui algoritm genetic (secvență de cod şi spațiu de căutare, operatorii de încrucişare şi mutație, funcții de scor şi selecție bazată pe acestea);

÷ Sunt prezentate caracteristicile cantitative ale unei alinieri de secvențe de gene;

÷ Sunt sintetizate principalele programe care implementează algoritmi genetici folosite în alinierea secvențelor de gene împreună cu descrierea caracteristicii principale pe care o posedă;

P.3

Classical Approaches of Genetic Algorithms and their Suitability

Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ

Asian Journal of Chemistry 22(3):2275-2284, 2010

÷ Lucrarea este de tip lucrare originală din materialul şi metoda cercetării;

÷ Este prezentată o problemă dificilă de model de cinetică a unei reacții chimice;

÷ Este reformulată problema în termenii specifici aplicării unui algoritm genetic;

÷ Este construit spațiul de căutare, sunt definiți operatorii de mutație şi încrucişare, funcția de scor şi operatorul de selecție, este elaborată strategia de evoluție către obiectivul optimizării valorilor parametrilor de cinetică, sunt prezentate rezultatele obținute;

÷ Sunt discutate rezultatele obținute în raport cu alte strategii de rezolvare a problemei dificile de cinetică a unei reacții chimice;

P.4

Distribution Fitting 1. Parameters Estimation under Assumption of Agreement between Observation and Model

Lorentz JÄNTSCHI

ArXiv electronic library permanent link(July 16, 2009): http://arxiv.org/abs/0907.2829
(Subject: Statistics - Methodology)

Bulletin UASVM, Horticulture 66(2):684-690, 2009

÷ Lucrarea este de tip lucrare originală din materialul, metoda şi rezultatele cercetării;

÷ Seria de date formată din cei 206 din 209 bifenili policloruraţi (PCB) asupra cărora s-au făcut măsurătorile coeficientului de partiţie octanol/apă (lkow) au servit drept studiu de caz în ceea ce priveşte ipoteza de normalitate cu privire la distribuţia valorilor observate într-un cadru mai general în care s-a definit şi folosit o generalizare a legilor de distribuţie Gauss şi Laplace (Gauss-Laplace este o distribuţie cu 3 parametrii, al 3-lea parametru definind forma distribuţiei şi având drept două cazuri particulare cele două distribuţii menţionate);

÷ Subiectul lucrării este măsura în care seturile de date observate asupra cărora facem presupunerea de normalitate a datelor în scopul de a fi incluse într-o analiză ale cărei căi sunt bătătorite (cea a ipotezei de normalitate) sunt într-adevăr normal distribuite;

÷ Problematica estimării parametrilor distribuţiei (forma generalizată Gauss-Laplace utilizată în ipoteza - verificată - a absenţei oblicităţii) este discutată în detaliu (cum anume o anumită presupunere - cum ar fi asupra valorii unui parametru - influenţează estimarea; cum anume alegerea unei metode de estimare - momente centrale, regresii, statistici, maximizarea ratei şansei - influenţează estimarea);

P.5

Distribution Fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Kramer-von-Misses and Jarque-Bera statistics

Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ

ArXiv electronic library permanent link (July 16, 2009): http://arxiv.org/abs/0907.2832
(Subject: Statistics - Methodology)

Bulletin UASVM, Horticulture 66(2):691-697, 2009

÷ Lucrarea este de tip lucrare originală din materialul, metoda şi rezultatele cercetării;

÷ Seria de date formată din cei 206 din 209 bifenili policloruraţi (PCB) asupra cărora s-au făcut măsurătorile coeficientului de partiţie octanol/apă (lkow) au servit drept studiu de caz în ceea ce priveşte valorile suspecte în ipoteza de normalitate cu privire la distribuţia valorilor observate, fiind identificată o valoare suspectă de abatere grosolană de la model;

÷ O baterie de teste statistice au servit în analiză;

÷ Un studiu critic cu privire la ipotezele şi modelele pe care sunt construite statisticile analizate este cuprins în lucrare;

÷ Un set de date similar cu lkow al PCB raportat în literatura de specialitate a servit drept contraexemplu în ceea ce priveşte acceptarea ipotezei de normalitate cu privire la datele observate;

÷ Statistica Grubbs este propusă drept instrument de eliminare a datelor îndoielnice şi este arătat cum pentru un set de date normal distribuit (lkow al PCB) această statistică identifică corect datele îndoielnice în timp ce pentru un set de date care nu este normal distribuit (contraexemplul) aceeaşi statistică fie nu identifică nici o dată îndoielnică (cazul investigat) fie eliminarea presupusei date îndoielnice nu duce la îmbunătăţirea agrementului între observaţie şi model;

P.6

Distribution Fitting 3. Analysis under Normality Assumption

Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI

Bulletin UASVM, Horticulture 66(2):698-705, 2009

÷ Lucrarea este de tip lucrare originală din materialul şi metoda cercetării;

÷ Cercetarea întreprinsă în "Distribution Fitting 2. ..." cu privire la calitatea reprezentării agrementului între observație şi model dată de o serie de statistici şi cu privire la statistica Grubbs ca instrument de eliminare a datelor îndoielnice a fost aplicată sistematic folosind metodologia elaborată în "Distribution Fitting 2. ..." pentru patru seturi de date similare raportate în literatura de specialitate în scopul evaluării agrementului între observații şi modelul de normalitate a datelor care a fost ipoteza de lucru în studiile menționate;

÷ Rezultatele obținute au confirmat totdeauna consistența metodologiei elaborate şi nu întotdeauna consistența ipotezei de normalitate formulate asupra datelor măsurate;

P.7

Distribution Fitting 4. Benford test on a sample of observed genotypes number from running of a genetic algorithm

Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Carmen E. STOENOIU, Mihaela IANCU, Monica M. MARTA, Elena M. PICĂ, Monica ŞTEFU, Adriana F. SESTRAŞ, Marcel M. DUDA, Radu E. SESTRAŞ, Ştefan ȚIGAN, Ioan ABRUDAN, Mugur C. BĂLAN

Bulletin UASVM, Agriculture 66(1):82-88, 2009

÷ Lucrarea este de tip lucrare originală din materialul, metoda şi rezultatele cercetării;

÷ Rezultatele în urma observației execuției algoritmului genetic de optimizare a SAR cu 4 membrii MDF pe lkow a seriei de 206 din 209 PCB au servit în verificarea ipotezei de valididate a datelor;

÷ Testul Benford a servit în acestă analiză fiind aplicat pentru prima, a doua şi a treia cifră semnificativă ale numerelor obținute folosind pentru măsura agrementului între observație şi modelul dat de legea Benford două statistici: Chi-Square şi Kolmogorov-Smirnov;

÷ Diferențele observate între măsurile date de cele două statistici agrementului au fost puse pe seama formei legii de distribuție a diferențelor; astfel a fost investigat agrementul în ipoteza de distribuție Gauss şi în ipoteza de distribuție Laplace; a fost explicată disconcordanța între semnificațiile statistice ale statisticilor Chi-Square şi Kolmogorov-Smirnov cu ajutorul testului de normalitate Jarque-Bera;

P.8

(Distribution Fitting 5.) Knowledge Assessment: Distribution of Answers to an Online Quizzed System

Lorentz JÄNTSCHI, Carmen E. STOENOIU, Sorana D. BOLBOACĂ, Mugur C. BĂLAN, Liviu C. BOLUNDUȚ, Violeta POPESCU, Horea I. NAȘCU, Ioan ABRUDAN

Bulletin UASVM, Horticulture 66(2):680-683, 2009

÷ Lucrarea este de tip lucrare originală din materialul şi metoda cercetării;

÷ Metodologia elaborată în cadrul cercetării cu privire la studiul legii de distribuție a datelor observate şi agrementului între observație şi model au servit în această analiză asupra unor date de cu totul altă natură: numărul de răspunsuri corecte şi timpii de răspuns la un chestionar de examinare la încheierea parcursului unei discipline de către studenți aflați în semetrul II al anului I de studiu;

÷ Cele trei statistici Pearson-Fisher Chi-Square (C-S), Anderson-Darling (A-D) şi Kolmogorov-Smirnov (K-S) au dat măsura agrementului între observații (atât cu privire la număr de răspunsuri cât şi cu privire la timpii de răspuns) şi model;

÷ Un număr de 61 de alternative de model au fost incluse în studiu;

÷ Un număr de cinci distribuții s-au calificat (nefiind respinse semnificativ statistic) pentru o analiză mai amănunțită, redată în lucrare: Birnbaum-Saunders (sau a vieții obosite), Gausiana Inversă, Log-Normala, Gauss (sau Normală) şi Weibull;

÷ Două dintre acestea (Birnbaum-Saunders şi Log-Normala) - probabil datorită volumului insuficient de date - au necesitat analiză suplimentară, nefiind respinse la nici un nivel de semnificație sub 10%;

÷ Pentru a face distincție, s-a construit un experiment în jurul modei, care a arătat - de departe - că doar distribuția Birnbaum-Saunders (a vieții obosite) poate fi acceptată drept lege de distribuție la chestionarul de examinare aplicat (cu răspunsuri multiple la 30 de întrebări în sistem contra-timp în care timpul are pondere în notă la fel ca şi numărul de răspunsuri);

÷ Este important de subliniat că legea de distribuție Birnbaum-Saunders a fost verificată atât pentru numărul de răspunsuri corecte date cât şi pentru timpii de răspuns.

P.9

(Distribution Fitting 6.) Issues in Choosing a University

Monica M. MARTA, Carmen E. STOENOIU, Lorentz JÄNTSCHI, Ioan ABRUDAN

Bulletin UASVM, Horticulture 66(2):816-822, 2009

÷ Lucrarea este de tip lucrare originală din materialul şi metoda cercetării;

÷ Metodologia elaborată în cadrul cercetării cu privire la studiul legii de distribuție a datelor observate şi agrementului între observație şi model au servit în această analiză asupra unor date de cu totul altă natură: măsura percepției studenților asupra mediului academic. În lucrare au fost analizate statistic datele dintr-un chestionar aplicat la 403 studenți cu privire la motivele pentru care au ales Universitatea Tehnică din Cluj-Napoca; datele au fost colectate pe cinci grade de relevanță (de la nimic la maxim) pentru o serie de şase factori care au contribuit la alegere;

÷ Contingența între factori şi grade a fost măsurată cu statistica Chi-Square;

÷ Atât forma distribuției observate cât şi metodologia de obținere (care presupune a alege dintr-o serie de mai mulți factori pe aceia care sunt mai importanți şi să dai o măsură a importanței) au sugerat că legea de distribuție este de valori extreme; astfel s-a făcut presupunerea că legea de distribuție este legea Generalizată a Valorilor Extreme (sau Fisher-Tippett, F-T);

÷ Analiza statistică a arătat că pentru patru din cei şase factori ipoteza de distribuție după legea F-T nu poate fi respinsă (în fapt obținându-se probabilități de observare întâmplătoare pentru eşantionul analizat ca selecție din populație F-T mai mari de 40% pentru fiecare din cei patru factori, în timp ce pentru ceilalți doi probabilitățile de observare au penalizat sever ipoteza - sub 3% şi respectiv sub 0.01‰);

÷ Între cei patru factori care urmează legea F-T s-a măsurat asocierea în probabilitate, ştiut fiind că în mod intrinsec factori ca calitatea infrastructurii şi calitatea mediului academic sunt corelați; analiza suprapunerii în probabilitate a acestor factori a pus în evidență asocieri interesante care au loc în percepția studenților;

÷ Cu privire la unul dintre factori, şi anume locația geografică a universității (care răspunsuri nu sunt distribuite F-T, fiind respinsă ipoteza la 0.01‰) s-a făcut presupunerea că cauza este designul experimental; astfel, studenții pot fi din oraş, din județ sau din afara județului (deci trei categorii) şi din acest motiv cerința de a clasifica după cinci categorii importanța este în primul rând greşit din punctul de vedere al designului de experiment şi în al doilea rând dificil de făcut din calitatea de student; s-a presupus astfel că există doar trei categorii de răspunsuri, răspunsurile din categoriile intermediare provenind de fapt din una din cele două categorii învecinate; s-a optimizat clasificarea în acord cu acest principiu în ipoteza de distribuție Binomial(2) reducându-se valoarea X^2 până la $6 \cdot 10^{-3}$ (înalt semnificativ statistic);

P.10

(Distribution Fitting 7.) Analysis of the genotypes number in different selection and survival strategies

Lorentz JÄNTSCHI, Mugur C. BĂLAN

Bulletin UASVM, Horticulture 66(1):58-65, 2009

÷ Lucrarea este de tip lucrare originală din materialul, metoda şi rezultatele cercetării;

÷ Rezultatele în urma observației execuției algoritmului genetic de optimizare a SAR cu 4 membrii MDF pe lkow a seriei de 206 din 209 PCB au servit în verificarea ipotezei de valididate a datelor;

÷ Analiza de variabilitate a genotipurilor reprezentate în cultivar în momentele a fost studiată folosind o contingență de trei observabile: număr de genotipuri distincte, număr de apariții şi număr de asocieri în ecuațiile de regresie valide rezultate şi două măsuri: număr total şi număr de din cele mai frecvente 23 de genotipuri în spațiul de observare format din cele 46 de execuții independente pentru fiecare strategie de evoluție;

÷ La fiecare din cele 6 rezultate ale observării s-a evaluat independența de contingența strategiilor de selecție cu cele de supraviețuire, şi pentru fiecare caz în parte ipoteza de independență a fost respinsă, adică aceste rezultate ale observării sunt o particularitate a strategiei de selecție şi de supraviețuire alese pentru evoluție;

÷ Asocierea în scară logaritmică între câte două rezultate ale observării din cele trei referitoare la numărul total au evidențiat că cel puțin din punctul de vedere al numărului de genotipuri distincte, al numărului de genotipuri şi al numărului de asocieri ale acestora în regresii o strategie (selecție deterministă şi supraviețuire deterministă) se abate semnificativ statistic de la regula liniară de asociere între aceste numere impusă de celelalte opt strategii;

÷ Au fost analizate datele şi sub aspectul valorilor medii, abaterilor standard şi al intervalelor de încredere ale acestora, când o serie de concluzii importante cu privire la distincția semnificativ statistică a datelor au putut fi extrase;

P.11

(Distribution Fitting 8.) Weather Monitoring: Wind Analysis (May, 2009; GPS: Lat. N46°45'35"; Long. E23°34'19")

Mugur C. BĂLAN, Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI, Radu E. SESTRAŞ

Bulletin UASVM, Horticulture 66(1):7-9, 2009

÷ Lucrarea este de tip lucrare originală din materialul şi metoda cercetării;

÷ Rezultatele monitorizării vântului folosind o stație meteo amplasată în spațiul Universității de Științe Agricole și Medicină Veterniară din Cluj-Napoca au fost prelucrate matematic pentru a obține informații importante cu privire la deplasarea curenților de aer - activitatea vântului (timp) direcție, distanță și viteză - folosind datele colectate de la stația meteo în luna Mai 2009 și stocate într-o bază de date din minut în minut;

P.12

(Distribution Fitting 9.) Student perception of degree of academic community involvement in academic life

Carmen E. STOENOIU, Sorana D. BOLBOACĂ, Ioan ABRUDAN, Lorentz JÄNTSCHI

Actual Problems of Economics 11(101):276-283, 2009

÷ Metodologia elaborată în cadrul cercetării cu privire la studiul legii de distribuție a datelor observate şi agrementului între observație şi model au servit în această analiză asupra unor date de cu totul altă natură: măsura percepției studenților asupra mediului academic. În lucrare au fost analizate statistic datele dintr-un chestionar aplicat la 403 studenți cu privire la percepția studenților asupra gradului de implicare al comunității academice în viața academică;

÷ Datele au fost colectate pe cinci grade de relevanță (de la "deloc" la "în foarte mare măsură") pentru o serie de şase factori care au contribuit la alegere;

÷ Testul X^2 a fost aplicat pentru verificarea ipotezei de independență a observabilei grad de implicare de categoriile de relevanță şi de populațiile comunității academice (în număr de cinci);

÷ S-au asociat valori gradelor de relevanță astfel: 0 = "deloc"; 1 = "în mică măsură"; 2 = "în măsură medie"; 3 = "în mare măsură"; E_i = "în foarte mare măsură" (i = 0..4) - asupra valorii asociate gradului de relevanță "în foarte mare măsură" s-a formulat ipoteza că nu se situează la valoarea 4 pe scara definită de celelalte grade de relevanță ci la o valoare necunoscută, de determinat, care defineşte expectanța;

÷ Asupra variabilei "grad de implicare" s-a formulat ipoteza de distribuție normală; în ipoteza de normalitate s-a calculat formal valoarea statisticii Jarque-Bera pentru fiecare eşantion de observații (JB_i, i = 0..4) ca expresie de valorile necunoscute ale mediei (μ_i) şi expectanței (E_i);

÷ S-a optimizat (minimizat) valoarea statisticii Jarque-Bera (maximizat agrementul între observații şi modelul de distribuție normală), obținându-se valorile mediei (μ_i) şi expectanței (E_i) fiecărei populații academice (Conducere academică, Conducerea facultății, Cadre didactice, Personal administrativ, Studenți) în percepția studenților;

÷ Minimizarea statisticii JB a evidențiat o observabilă (JB_4=5.6) cu depărtare consistentă ($p_{\chi 2}$(5.6,2) = 6%) de normalitate în timp ce toate celelalte au valori de agrement foarte bune ($JB_0..JB_3 < 10^{-3}$) - populația în cauză fiind studenții; rezultatul a fost pus pe seama subiectivității - în acest caz cerându-se studenților să dea măsura implicării lor înşişi în viața academică;

÷ O singură asociere a fost evidențiată între populații folosind suprapunerea de probabilitate între legile de distribuție obținute ale gradului de implicare în viața academică în percepția studenților - între conducerea academică şi personalul administrativ (21% suprapunere în probabilitate);

P.13

(Distribution Fitting 10.) The Role of the Academic Community in Defining the Professional Route: Students' Perception

Carmen E. STOENOIU, Sorana D. BOLBOACĂ, Ioan ABRUDAN, Lorentz JÄNTSCHI

Actual Problems of Economics 103(1):277-285, 2010

÷ Metodologia elaborată în cadrul cercetării cu privire la studiul legii de distribuție a datelor observate şi agrementului între observație şi model au servit în această analiză asupra unor date de cu totul altă natură: măsura percepției studenților asupra mediului academic. În lucrare au fost analizate statistic datele dintr-un chestionar aplicat la 403 studenți cu privire la rolul pe care îl are fiecare comunitate academică în definirea propriei rute profesionale; datele au fost colectate pe cinci grade de relevanță (de la "deloc" la "în foarte mare măsură") pentru o serie de şase factori care au contribuit la alegere;

÷ Testul X^2 a fost aplicat pentru verificarea ipotezei de independență a observabilei grad de implicare de categoriile de relevanță şi de populațiile comunității academice (în număr de şase);

÷ S-au asociat valori gradelor de relevanță astfel: 0 = "deloc"; 1 = "în mică măsură"; 2 = "în măsură medie"; 3 = "în mare măsură"; 4 = "în foarte mare măsură"; s-a verificat că ipoteza de distribuție binomială a observațiilor nu este acceptată în nici una din populațiile asupra cărora s-au făcut observații;

÷ S-a formulat ipoteza $4 \neq$ "în foarte mare măsură" $= E_i$ $(i = 0..5)$ - valoarea asociată gradului de relevanță "în foarte mare măsură" s-a formulat ipoteza că nu se situează la valoarea 4 pe scara definită de celelalte grade de relevanță ci la o valoare necunoscută, de determinat, care defineşte expectanța; asupra variabilei "grad de implicare" s-a formulat ipoteza de distribuție normală; în ipoteza de normalitate s-a calculat formal valoarea statisticii Jarque-Bera pentru fiecare eşantion de observații (JB_i, $i = 0..5$) ca expresie de valorile necunoscute ale mediei (μ_i) şi expectanței (E_i);

÷ S-a optimizat (minimizat) valoarea statisticii Jarque-Bera (maximizat agrementul între observații şi modelul de distribuție normală), obținându-se valorile mediei (μ_i) şi expectanței (E_i) fiecărei populații academice (Conducere academică, Conducerea facultății, Conducerea departamentului, Cadre didactice, Asociații studențești, Colegi) în percepția studenților;

÷ Minimizarea statisticii JB a evidențiat că în toate cazurile agrementul cu normalitatea este excelent ($JB_0..JB_5 < 2 \cdot 10^{-3}$) - rezultatul a fost pus pe seama inexistenței de aceată dată față de situația anterioară - lucrarea 13.12 - a subiectivității în apreciere;

÷ Testul Student t a servit la compararea mediilor obținute; probabilitățile din distribuția Student t au evidențiat existența unor asocieri între populațiile supuse observației în percepția studenților - între Conducerea facultății, Conducerea departamentului, şi Asociațiile studențeşti;

P.14

Meta-heuristics on quantitative structure-activity relationships: study on polychlorinated biphenyls

Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ

Journal of Molecular Modeling DOI: 10.1007/s00894-009-0540-z 16(2):377-386, 2010

÷ Lucrarea este de tip lucrare originală din materialul, metoda şi rezultatele cercetării;

÷ Lucrarea descrie parametrii de configurare ai procesului de evoluție studiat (volumul eşantionului de genotipuri, numărul de variabile în modelul de regresie multiplă, criteriile impuse pentru adaptarea la mediu, probabilitatea de mutație pentru părinți şi pentru descendenți, procedura de încrucişare, parametrul (observabila) supus optimizării în procesul de evoluție, calculul scorului parametrului supus optimizării, obiectivul evoluției;

÷ În lucrare se comunică rezultatele obținute din evoluția folosind o strategie de selecție în turnir şi o strategie de supraviețuire deterministică pe un număr de 50000 de generații, caz în care optimul s-a obținut în generația 17277;

÷ Modelul obținut s-a comparat cu modelul obținut din căutări sistematice folosind testul Steiger; compararea a evidențiat statistic că cele două modele nu se disting semnificativ statistic la riscul de a fi în eroare de 5%;

÷ Modelul obținut a permis interpretări fenomenologice cu privire la natura relației între structură şi activitatea biologică observată; modelul obținut s-a confirmat în natura fenomenologică de modelul obținut anterior prin căutări sistematice;

P.15

A Study of Genetic Algorithm Evolution on the Lipophilicity of Polychlorinated Biphenyls

Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ

Chemistry and Biodiversity DOI: 10.1002/cbdv.200900356, 2010

÷ Lucrarea este de tip lucrare originală din materialul, metoda şi rezultatele cercetării;

÷ Lucrarea descrie parametrii de configurare ai aplicației realizate ce implementează algoritmul genetic (limbajul de programare, conectivitatea cu baza de date şi structura bazei de date şi a tabelelor ce stochează datele primare), parametrii de configurare a evoluției (volumul eşantionului de genotipuri, numărul de variabile în modelul de regresie multiplă, criteriile impuse pentru adaptarea la mediu, probabilitatea de mutație pentru părinți şi pentru descendenți, procedura de încrucişare, parametrul (observabila) supus optimizării în procesul de evoluție, calculul scorului parametrului supus optimizării, obiectivul evoluției;

÷ Au fost analizate datele obținute din evoluția pe un număr de 1000 de generații în strategia de selecție proporțională şi strategia de supraviețuire deterministică;

÷ S-a analizat legea de distribuție a numărului de evoluții şi a coeficientului de determinare având la dispoziție rezultatele de evoluție dintr-un număr de 30 de execuții independente;

÷ Nu a putut fi pus în evidență statistic că ipoteza de distribuție Poisson a numărului de evoluții este respinsă;

÷ Nu a putut fi pus în evidență statistic că ipoteza de distribuție normală a coeficientului de determinare este respinsă;

÷ S-a evidențiat statistic că cel mai bun model MDF SAR obținut din evoluția algoritmului genetic pe parcursul a 1000 de generații în strategia de selecție proporțională şi supravietuirea deterministă din 30 de execuții independente nu diferă semnificativ de modelul obținut din căutarea sistematică;

P.16

A genetic algorithm for structure-activity relationships: software implementation

Lorentz JÄNTSCHI

ArXiv electronic library permanent link (June 26, 2009): http://arxiv.org/abs/0906.4846
(Subject: Neural and Evolutionary Computing)

A genetic algorithm for structure-property relationships: software implementation

Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ

Algorithms for Molecular Biology, Manuscript ID 1935642659299401, August 20, 2009

÷ Lucrarea este de tip lucrare originală din materialul, metoda şi rezultatele cercetării;

÷ Lucrarea cuprinde:

- descrie spațiul de căutare, construcția famililor de descriptori, secvența de cod genetic asociată unei familii de descriptori, complexitatea problemei de căutare a unei regresii multiple folosind familii de descriptori;
- descrie implementarea algoritmului genetic (implementarea operatorilor genetici, a strategiilor de evoluție, a scorurilor obiectiv, a scorurilor de selecție şi de supraviețuire, bucla de evoluție;
- prezintă algoritmii care realizează strategiile de evoluție (proporțional, turnir, deterministic);
- prezintă programul evolutiv realizat, structura şi specificațiile fişierelor de configurare pentru topologia genetică a familiei de descriptori, pentru conectivitatea la baza de date ce stochează datele primare, şi configurarea parametrilor de evoluție;
- este prezentată modalitatea în care programul evolutiv stochează rezultatele obținute în urma procesului de evoluție şi configurațiile care sunt posibile cu privire la ce şi cum să fie stocate rezultatele;
- este prezentat designul experimental - contingența strategiilor de selecție cu cele de supraviețuire - care a fost fixată în cele 46 de execuții independente ale algoritmului genetic pus să producă evoluții pe un parcurs de câte 20000 de generații;
- sunt sintetizate rezultatele cu privire la frecvența genotipurilor;
- este aplicată statistica X^2 pentru a dovedi independența numărului de genotipuri de contingența strategiilor de selecție cu cele de supraviețuire;

P.17

Recording Evolution Supervised by a Genetic Algorithm for Quantitative Structure-Activity Relationship Optimization

Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ

Applied Medical Informatics 26(2):89-100, 2010

÷ Lucrarea este de tip lucrare originală din materialul, metoda şi rezultatele cercetării;

÷ Lucrarea cuprinde:

- specificațiile de implementare ale programului evolutiv realizate cu scopul de a observa evoluția materialului genetic pe parcursul generațiilor algoritmului genetic;
- observabilele care se înregistrează pe baza compoziției materialului genetic prezent în cultivar;
- modalitatea prin care este apreciată capacitatea de adaptare a fenotipurilor obținute în urma proceselor de mutație şi încrucişare, instrument cu ajutorul căruia s-a realizat clasificarea de adaptabilitate la mediu a descendenților obținuți în urma procesului de înmulțire;
- opțiunile definite pentru înregistrarea parametrilor de stare ai materialului genetic în fişierele de ieşire pe care programul evolutiv le creează;
- valorile care au fost fixate pentru datele de intrare ale programului evolutiv în experimentul de simulare a evoluției urmând diferite strategii de selecție şi supraviețuire, cu privire la topologia familiei de descriptori, la eşantionul supus evoluției, la parametrii proceselor de selecție, supraviețuire, mutație şi încrucişare;
- valorile care au fost variate în datele de intrare ale programului evolutiv, şi anume strategia de selecție şi strategia de supraviețuire;
- designul experimental în experimentul de simulare a evoluției, şi anume numărul de generații pe parcursul cărora se fac observațiile asupra evoluției şi numărul de execuții independente în acelaşi cadru al datelor de intrare, în scopul asigurării semnificației statistice în rezultatele obținute;

P.18

Rezultate în premieră - posibil de publicat

÷ Teze de doctorat susținute pe subiectul algoritmi genetici - analiză sintetică - lucrare tip review;

÷ Metodologia optimizării geometriei în modelarea moleculară - lucrare tip notă tehnică;

÷ Statistici ale evoluției algoritmului genetic în funcție de selecție și supraviețuire din "Generații ce produc evoluții (număr și medie) pentru 46 de execuții independente" (Cât de timpuriu se produc evoluțiile?; Cât de frecvent se produc evoluțiile?; Cât de dispers se produc evoluțiile?; Cât de predictibil se produc evoluțiile?) - lucrare tip articol original;

÷ Diversitatea genotipurilor reprezentate în cultivar în momentele de evoluție (diversitatea Simpson, entropia Shannon, entropia Hartley, entropia Renyi) - lucrare tip articol original;

÷ Tendințe medii ale numărului de genotipuri, fenotipuri și asocieri pe parcursul evoluției (din 1000 în 1000 de generații) & experimentul Monte-Carlo (media mediilor) - lucrare tip articol original;

÷ Distribuția Gumbel nu este suficient de generală pentru a caracteriza distribuția pe parcursul evoluției - lucrare tip articol original;

÷ Ecuații de tendință în valorile parametrilor distribuțiilor Fisher-Tippett pentru coeficientul de determinare în evoluțiile urmând diferite strategii - lucrare tip articol original;

÷ Loteria norocoșilor, loteria ghinioniștilor și intervalul de normalitate în evoluție - lucrare tip articol original;

÷ Ce lege urmează momentele de apariție a evoluției? - lucrare tip articol original;

÷ Asocieri în familia de curbe de evoluție - lucrare tip articol original;

÷ Ce distribuție urmează numărul de evoluții? - lucrare tip articol original;

Anexe

1. Secvențe reprezentative pentru subiectul abordat din lucrările lui Lamarck, Darwin, Mendel, Weismann, Morgan şi Fisher (15 pagini)

http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Annex_1.pdf

2. Măsurarea bioactivității (12 pagini)

http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Annex_2.pdf

3. Statistică descriptivă şi inferențială (28 pagini)

http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Annex_3.pdf

4. Familii de descriptori moleculari (4 pagini)

http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Annex_4.pdf

5. Programul evolutiv implementat (18 pagini)

http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Annex_5.pdf

6. Entropii genotipice (3 pagini)

http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Annex_6.pdf

7. Testul Anderson-Darling pentru asocieri în distribuție (29 pagini)

http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Annex_7.pdf

8. Respingerea distribuției Gumbel (47 pagini)

http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Annex_8.pdf

9. Alternativa Fisher-Tippett (8 pagini)

http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Annex_9.pdf

10. Fisher-Tippett 3D (8 pagini)

http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Annex_10.pdf