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

Lorentz JÄNTSCHI

Rezumat al tezei de doctorat

Algoritmi genetici şi aplicaţii ale acestora

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

Cluj-Napoca
2010

Cuprins

Introducere 2

Probleme de optimizare a relațiilor structură-activitate 3

Simularea evoluției cu algoritmi genetici 4

Cadrul cercetării, scop şi obiective 6

Definirea problemei de optimizare a relației structură-activitate 7

Definirea problemei genetice şi proiectarea algoritmului 9

Definirea experimentului de simulare a evoluției 13

Rezultate şi discuții 15

Analiza variabilității şi diversității 19

Interpretarea rezultatelor din observațiile pe observabile calitative 24

Analiza obiectivului evoluției folosind un eşantion întâmplător de generații 27

Analiza legii de distribuție a obiectivului evoluției folosind un studiu sistematic în cadru generalizat 29

Ce lege urmează momentele de apariție a evoluției? 33

Ce distribuție urmează numărul de evoluții? 36

Concluzii şi recomandări 38

Lucrări reprezentative publicate 40

Introducere

Teza “Algoritmi genetici şi aplicaţii ale acestora” este un demers de cercetare fundamentală interdisciplinar, având ca scop simularea evoluției cu algoritmi genetici în probleme de optimizare a relațiilor structură-activitate.

Problematica principală avută în vedere o constituie problemele dificile (cele cu complexitate exponențială) de optimizare a relațiilor cantitative între structura compuşilor chimici şi activitatea lor biologică.

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, 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 soluția “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, a fost conceput cadrul necesar pentru construcția şi aplicarea unui algoritm genetic care să rezolve problemele de optimizare, prin elaborarea unui algoritm genetic într-un cadru definit. Respectivul algoritm genetic a fost implementat într-un program evolutiv şi aplicat pe un set de date experimentale, fiindu-i urmărită evoluția.

S-a realizat un design experimental cu scopul de a face trecerea de la problema de optimizare la o problemă de simulare, şi anume simularea evoluției în diferite strategii de selecție şi de supraviețuire. A fost creată o contingență de 3x3 strategii distincte pentru selecție şi supraviețuire (proporțional, în turnir şi deterministic) şi s-a urmărit evoluția pe parcursul a 20000 de generații în mod repetat de 46 de ori în fiecare strategie în parte. S-au analizat inferențele statistice în observabile calitative şi cantitative ale procesului de evoluție controlată de diferitele strategii de evoluție, cu ajutorul diferitelor variabile pe care programul evolutiv a fost configurat să le înregistreze.

Informațiile şi rezultatele obținute au în mare măsură un caracter fundamental. Analiza statistică a rezultatelor obținute din simularea proceselor de evoluție a permis obținerea unor răspunsuri la întrebări precum: *Care este legea de distribuție a obiectivului evoluției? Care este legea de distribuție a momentelor evoluției? Care este legea de distribuție a numărului de evoluții? Cum este influențată variabilitatea şi diversitatea genotipică de alegerea unei strategii de evoluție? Cât de timpuriu se produc evoluțiile în raport cu strategia de evoluție aleasă? Cât de frecvent se produc evoluțiile în raport cu strategia de evoluție aleasă? Cât de dispers (şi respectiv cât de predictibil) se produc evoluțiile în raport cu strategia de evoluție aleasă? Care sunt similaritățile şi deosebirile între evoluțiile care au loc urmând diferite strategii?*

Au fost obținute şi o serie de rezultate cu caracter aplicativ, cum sunt: implementarea algoritmului genetic într-un program evolutiv capabil să rezolve o problemă dificilă de optimizare a unei relații structură-activitate folosind familii de descriptori de structură; implementarea unor module de calcul automate pentru optimizarea geometriei moleculare; implementarea unor programe pentru calculul statisticii Anderson-Darling de agrement între model şi observație; implementarea procedurii Grubbs de identificare şi eliminare a observațiilor în eroare față de un model.

Teza oferă şi soluții de transfer tehnologic, cuprinzând răspunsuri la o serie de probleme de optimizare în domeniul horticol în care evoluția către un obiectiv de ameliorare sau de planificare este influențată de o serie de parametrii specifici materialului genetic şi/sau ai arealului şi în care obiectivul fixat este influențat de strategia aleasă; în acest cadru experimentul de simulare a evoluției, teza oferă soluții evidențiate statistic cu privire la influența strategiilor de selecție şi supraviețuire.

Probleme de optimizare a relațiilor structură-activitate

Scurt istoric

Abordarea matematică a SAR (Relații Structură-Activitate) pentru BAC (Compuși Biologic Activi), 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 investigați - (Hammett, 1935). Observații de 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), însă au trecut aproape patruzeci de ani de când paradigma relații cantitative structură-activitate (QSAR) şi-a dovedit utilitatea practică în agrochimie, chimie farmaceutică, toxicologie etc. (Hansch & Leo, 1979).

Ce sunt relațiile structură-activitate

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 din Teză). Activitatea biologică este evaluată prin intermediul unor proceduri specifice, supuse standardizării (ex. Tabelul 2 - Anexa 2-1 din Teză redă două astfel de activități biologice). O serie de procese biologice sunt referite distinct atunci când se exprimă activitatea biologică a unui compus chimic (ex. Tabelul 3 - Anexa 2-1 din Teză 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 din Teză.

Compuşii biologic activi (BAC) au o largă utilizare în domeniul agricol şi horticol. Tabelul 5 - Anexa 2-1 din Teză redă o clasificare a regulatorilor de creştere, în acord cu Societatea Americană de Ştiințe Horticole.

Seriile de compuşi sunt alcătuite din compuşi (congeneri) înrudiți, atât în ceea ce priveşte structura, cât şi proprietățile fizico-chimice şi/sau activitățile biologice. 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, trebuie asumată şi 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.

Elaborarea şi valorificarea de relații structură-activitate

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

Simularea evoluției cu algoritmi genetici

Scurt istoric

"Hard inheritance" ("Moştenirea dură") (Weismann, 1893) şi "Soft inheritance" ("Moştenirea uşoară") (Lamarck, 1809), selecția şi supraviețuirea (Darwin, 1859), genele şi recombinarea genică (Morgan & alții, 1915), transmiterea caracterelor (Mendel, 1866) constituie problematici îndelung dezbătute şi disputate de-a lungul secolului al XIX-lea (Fisher, 1954), toate contribuind la fundamentarea geneticii moderne de azi (Ayala & alții, 1994), şi oferind sursele de inspirație ale algoritmilor genetici.

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

În ce situații sunt aplicabili algoritmii genetici

Instrumentul de dezvoltare a algoritmilor genetici îl constituie informatica. Astfel, uzual, în viața de zi cu zi, şi la fel în cercetarea ştiințifică, se operează 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 de cea ilustrată de ***algoritm***. Un algoritm este în esență o rețetă, specificând ce trebuie făcut 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). 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.

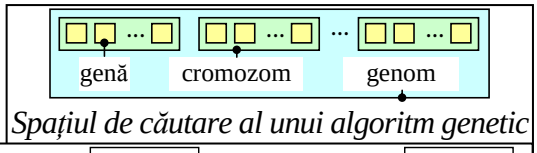
Ce sunt algoritmii genetici

Deoarece întotdeauna cercetătorii s-au confruntat cu mai multe probleme dificile, de foarte mult timp s-a încercat rezolvarea acestora, unul sau mai mulți ***euristici*** fiind de-a lungul anilor concepuți în acest sens. 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 întotdeauna 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 generalității pe care o presupun, 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*** (SA - Simulated Annealing) şi ***căutarea tabu*** (TS - Tabu Search) sunt şi ***algoritmii genetici*** (GA - Genetic Algorithm). Chiar dacă primele studii în care au apărut algoritmii genetici se situează în anul 1954 (Barricelli, 1954), studii de amploare ale acestora au apărut după 1970 (Bosworth & alții, 1972; Holland, 1975), ei fiind re-inventați ceva mai târziu (Davis, 1991; Holland, 1992) odată cu dezvoltarea tehnicii de calcul.

Complexitatea algoritmică

O problemă importantă legată de ***complexitatea algoritmică*** este reprezentată de teorema ***inexistenței mesei pe gratis*** (NFLT - No Free Lunch Theorem; Wolpert & Macready, 1995&1997;

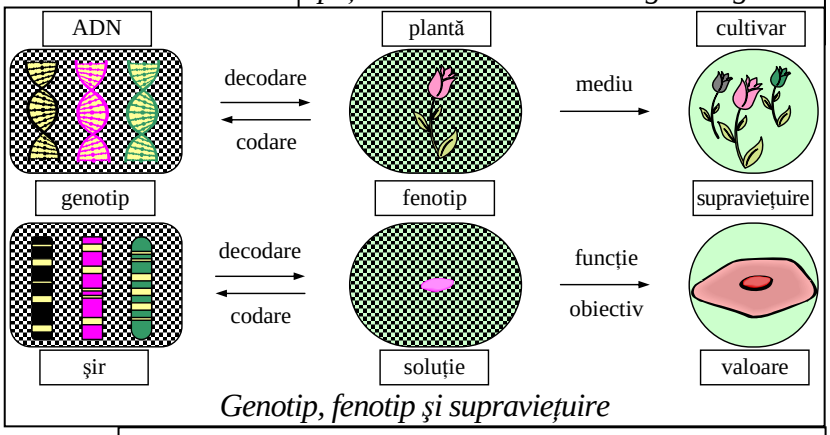
English, 1996), teoremă care, utilizând trei criterii de evaluare a calității unui algoritm (viteză, precizie și scop) sugerează că toți algoritmii sunt strict echivalenți. În esență, aceasta înseamnă că pentru doi algoritmi A și B, pentru fiecare set de date pentru care A performează mai bine decât B, există un set de date pentru care B performează mai bine decât A.



Spațiul de căutare al unui algoritm genetic

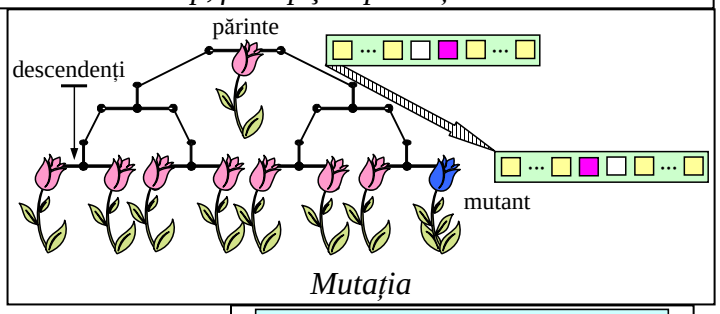
Construcția algoritmilor genetici

÷ Se operează asupra unei populații de reprezentări abstracte 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*** este compusă din ***gene***.



Genotip, fenotip și supraviețuire

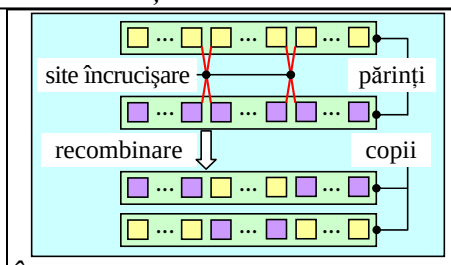
÷ Fiecare ***generație*** este compusă dintr-o populație de șiruri de caractere (sau alte forme de reprezentare abstractă) analog cu cromozomii ADN-ului. Fiecare element al populației reprezintă un punct în spațiul de căutare și în același timp o soluție posibilă.



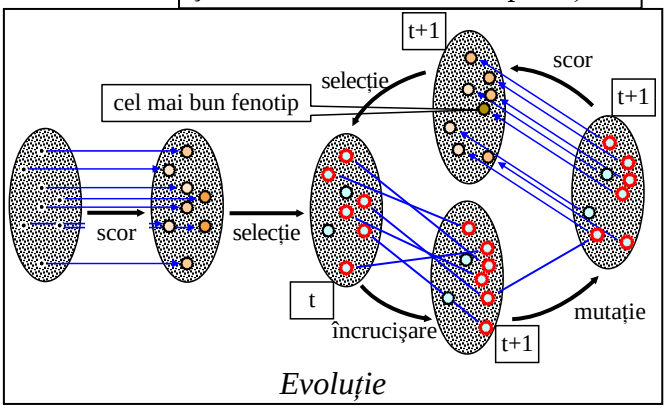
Mutația

÷ Un scor sau ***șansă de supraviețuire*** a fiecărei soluții este calculată pentru fiecare genotip cu ajutorul unei funcții, numită și ***funcție obiectiv***. Valoarea acestei funcții este asociată cu abilitatea individului să supraviețuiască și definește astfel ***fenotipul*** asociat genotipului.

÷ Scorul este asociat fiecărui fenotip (soluție) reprezentând abilitatea acestuia de a concura pentru resurse de mediu, pentru ***supraviețuire*** (selecție). Scopul algoritmului genetic este de a aplica ***încrucișarea*** și ***mutația*** selectivă a fenotipurilor (prin intermediul decodării lor în genotipurile din care provin), pentru a produce descendenți mai buni decât părinții lor.



Încrucișare dublă, implicând ruperea și reunirea cromozomilor părinților



Evoluție

÷ Într-un algoritm genetic clasic, pentru a rezolva o problemă, se generează întâmplător sau se inițiază cu valori predefinite o populație de un volum dat de genotipuri și evoluția se iterează prin repetiția selecției, mutației și încrucișării până când cel mai bun fenotip al populației satisface o condiție impusă (condiție care reprezintă condiția de sfârșit a algoritmului).

Elaborarea și valorificarea algoritmilor genetici

Algoritmii genetici servesc în clasificarea filogenetică (Jäntschi & alții, 2008-PTA), analiza

secvențelor de gene (Jäntschi & alții, 2009-GSA), probleme dificile de dinamica proceselor (Jäntschi & alții, 2009-DPA) şi în orice altă categorie de probleme dificile de decizie, clasificare, optimizare sau simulare (Falkenauer, 1998).

Cadrul cercetării

Dezvoltarea continuă a depozitelor de cunoştinţe de genul celor administrate de NIH, cum sunt PubMed, PubChem, Genome etc., accentuează necesitatea de a poseda instrumente eficace de a relaționa aceste cunoştinţe, 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 al cărui subiect sunt algoritmii genetici.

Studii asupra altor operatori esențiali pentru evoluție sunt axate pe eficiența algoritmică (viteza cu care se atinge obiectivul şi apropierea de maximul global). O colecție de lucrări de acest tip este reprezentativă în acest sens (Martin & Spears, 2001). Astfel subiectul îl constituie diferiții operatori de încrucişare (Prügel-Bennett, 2001), mutația şi încrucişarea (Spears, 2001), sau alți parametrii dinamici (Droste & alții, 2001).

Studiile sunt adesea concentrate spre rezolvarea problemelor dificile cu ajutorul algoritmilor genetici, uneori abordându-se direcționat eficiența acestora (ca timp de execuție, resurse de memorie necesare), dar foarte puțin influența diferitelor strategii de evoluție asupra obiectivului urmărit. În acest din urmă caz, se are în vedere în special eficiența algoritmului, şi aproape niciodată parametrii ce caracterizează eşantionul supus evoluției.

Datorită potențialului de valorificare a rezultatelor pe care îl au, algoritmii genetici au depăşit demult granițele domeniului informatică. Teze de doctorat având ca obiectiv proiectarea de algoritmi genetici, implementarea de programe evolutive şi realizarea de studii cu ajutorul lor, se regăsesc practic în toate domeniile de cercetare. Astfel, în domeniul agricultură şi-au găsit utilizarea la planificarea culturilor (Matthews & Kraw, 2001), 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 domeniul biologie, se desprind două direcții principale în ceea ce priveşte elaborarea şi utilizarea algoritmilor genetici: în probleme de evoluție (Suzuki & Iwasa, 1998) şi în studii filogenetice (Zwickl & Hills, 2006). În privința caracterului practic, de utilizare a algoritmilor genetici în domeniul agricol şi horticol, algoritmii genetici au o largă aplicabilitate, de la studii de creştere (Venard & Vaillancourt, 2006), la clasificări taxonomice (Sarmiento-Monroy & Sharkey, 2006) sau analiza diversității genetice (Zhang & Ghabrial, 2006).

Scop şi obiective

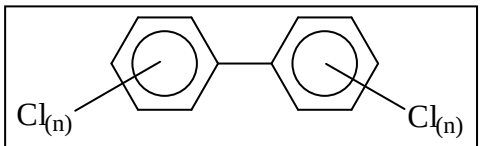
Scop: *Simularea evoluției cu algoritmi genetici în probleme de optimizare a relațiilor structură-activitate*. Proiectarea, implementarea şi analiza statistică a influenței diferitelor metode de selecție şi de supraviețuire asupra evoluției unui algoritm genetic utilizat pentru obținerea de relații structură-activitate în serii de compuşi chimici biologic activi.

Obiective:

1. Elaborarea algoritmului genetic (Definirea problemei dificile, cu complexitate exponențială; Formularea problemei în termeni genetici; Proiectarea algoritmului genetic; Implementarea şi documentarea programului evolutiv bazat pe algoritmul genetic);
2. Simularea evoluției (Evoluție: definirea observabilelor; Definirea contingenței selecție vs. supraviețuire; Proiectarea experimentului statistic; Realizarea experimentului statistic);
3. Analiza şi interpretarea rezultatelor: Observabile calitative; Obiectivul evoluției - r2 - observabilă cantitativă);

Definirea problemei de optimizare a relației structură-activitate

Setul de molecule ales pentru studiu este seria bifenililor policloruraţi (PCBs), formată din 209 compuşi, a cărui studiu este de o deosebită importanţă pentru impactul acestora asupra ecosistemului.



PCBs: Seria bifenililor policloruraţi (structuri & activităţi în Anexa 2-2)

Chiar dacă producţia de PCB a fost stopată în 1970 datorită toxicităţii ridicate a celor mai mulţi congeneri ai familiei PCB, efectele acestora sunt încă prezente în mediu, datorită faptului că PCBs sunt poluanţi organici persistenţi - clasificaţi ca atare, care se acumulează plante şi animale. Coeficientul de partiţie octanol/apă (K_{ow}) este raportul concentraţiilor unui compus chimic între octanol şi apă aflate în contact la o anumită temperatură. Este un parametru adimensional (fiind un raport de concentraţii) 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 magnitudinii de bioacumulare în peşti (U.S. Geological Survey, 2008).

O serie de statistici au fost aplicate pentru verificarea ipotezei de normalitate pe seria de activităţi biologice observate (pe 206 din 209 compuşi).

Ipoteza de normalitate a lkow măsurate, 206 compuşi (descrierea statisticilor în Anexa 3-3 din Teză)

Statistică	Valoare	Probabilitate de observaţie	Concluzie
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			

Ipoteza de normalitate a lkow măsurate, 205 compuşi (206\{PCB209(lkow=9.603)}

Statistică	Valoare	Probabilitate de observaţie	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

Valoarea eliminată (9.603) a fost identificată folosind procedura Grubbs elaborată pe baza testului Grubbs (Grubbs, 1969). În setul format de cei 205 compuşi, se remarcă nu numai agrementul între măsurat şi model (Gauss) dar şi agrementul între statistici.

Modelarea structurii moleculare este condiţia obligatorie pentru o analiză structură-activitate. Realizarea unui model tridimensional (3D) se impune în situaţiile în care descriptorii calculaţi uzează de geometria moleculară, aspect valabil pentru cazul prezent. Obţinerea 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 (licenţă v. 8.0/2007). Obţinerea informaţiei de structură 3D presupune parcurgerea unei serii de paşi, care cuprinde definirea unui model de mecanică moleculară (a fost definit AMBER), optimizarea geometriei moleculare până la convergenţă folosind un algoritm de optimizare (a fost folosit POLAK-RIBIERE), definirea unei metode semi-empirice de calcul energetic (a fost definită AM1) şi obţinerea unei serii de parametrii energetici ce

caracterizează optimul împreună cu sarcinile electrice parțiale.

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 dintr-o 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.

Algoritmul genetic a fost elaborat având ca model patru familii de descriptori de concepție proprie. Descrierea detaliată a familiilor de descriptori moleculari este redată în Anexa 4 din Teză. Pentru analiza structură-activitate implicând PCBs şi activitatea măsurată a acestora lkow s-a ales familia MDF (în Anexa 4-2 din Teză) să definească informația structurală în asociere cu activitatea biologică, această familie fiind cea mai bine documentată şi susținută de rezultate (cu ajutorul acesteia au fost analizați peste 50 de seturi de compuși).

Relația structură-activitate căutată este compusă din patru descriptori membrii MDF şi definește o problemă dificilă de optimizare datorită volumului populației MDF de descriptori (787968 membri).

Definirea problemei genetice şi proiectarea algoritmului

Fiecare ***genă*** [1] codifică câte un operator folosit în construcția ***cromozomului*** [2] unui descriptor molecular (Tabelul 5-2 în Teză). Fiecare descriptor al unei familii de descriptori este un ***genotip*** [3] şi toți împreună constituie ***materialul genetic*** [4] al familiei respective.

Secvență din Tabelul 5-2. Spațiul de căutare al MDF

Familie	Gene	Genom																											
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																						

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, Tabelul 5-3 sumarizând aceste variații:

Tabelul 5-3. Volumele familiilor de descriptori moleculari

Familie	Gene								Volum (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

Metodologia de lucru a algoritmilor genetici presupune prelevarea inițială (întâmplător sau deterministic) a unui ***eşantion*** [5] de cromozomi din materialul genetic format dintr-un şir de descriptori X_1, ..., X_p care este supus procesului de ***evoluție*** [6] în ***cultivar*** [7]. ***Algoritmul genetic*** [8] operează astfel asupra eşantionului care suferă modificări în fiecare ***generație*** [9]. Fiecare mulțime de `*n*` descriptori distincți reprezintă un punct în ***spațiul de căutare*** [10] şi în acelaşi timp o ***soluție posibilă*** [11]. Operatorii de bază ai unui algoritm genetic sunt încrucişarea şi mutația. ***Încrucişarea*** [12] a două genotipuri presupune alegerea unei porțiuni de încrucişat de-a lungul

---

[1] genă = una din valorile de pe coloana Gene a Tabelului 36; ex. I_M pentru FPIF

[2] cromozom = secvența de gene a unei familii în Tabelul 36; ex. $D_MA_PI_DI_MF_CS_ML_O$ pentru MDF

[3] genotip = o concretizare posibilă a valorilor fiecărei gene a unui cromozom; ex. TCJtAAfDI pentru MDFV

[4] 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

[5] 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

[6] evoluție = proces genetic complex care implică procese de selecție, încrucişare şi mutație

[7] 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)

[8] algoritm genetic = algoritm care descrie prin instrucțiuni procesul de evoluție asupra eşantionului

[9] generație = una din iterațiile algoritmului genetic

[10] spațiul de căutare = mulțimea posibilităților de selecție a `n` descriptori din `$V_S$` posibilități (relația 20)

[11] soluție posibilă = o ecuație de regresie cu `n` descriptori distincți definită de relația (1) sau (2)

[12] 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

şirului de gene (întâmplător sau deterministic), valorile celor două porţiuni de şiruri fiind schimbate între ele la descendenţi. ***Mutaţia*** [13] unui genotip presupune modificarea unei valori a unei gene a cromozomului cu o altă valoare din lista valorilor posibile pentru gena respectivă. Rezultatul încrucişării şi al mutaţiei este obţinerea de ***descendenţi*** [14] sau urmaşi, cu genotipuri diferite. ***Selecţia*** [15] genotipurilor este operaţia prealabilă necesară încrucişării şi mutaţiei şi se face pe baza unui ***scor de selecţie*** [16]. Cel puţin o parte a descendenţilor reprezintă descriptori ***viabili*** [17] 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*** [18] aplicat indivizilor din eşantion pe baza unui ***scor de supravieţuire*** [19]. Un alt parametru al algoritmului genetic îl reprezintă ***obiectivul evoluţiei*** [20] care este urmărit pe baza unei ***funcţii obiectiv*** [21]. 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 fac parte din cea mai bună ecuaţie de regresie obţinută pe baza indivizilor din cultivar). Se poate opta ca indivizii care ating obiectivul evoluţiei într-o generaţie să fie păstraţi în eşantion, caz în care acestora nu li se mai aplică procesul de supravieţuire, ei fiind automat incluşi în eşantionul generaţiei următoare.

Aşa cum rezultă din aplicarea procesului de supravieţuire, nu toţi indivizii unei generaţii supravieţuiesc şi sunt incluşi în generaţia următoare. Motivul acestui fapt este păstrarea unui număr constant de genotipuri în eşantionul dezvoltat în cultivar, astfel încât numărul de indivizi înlocuiţi din eşantion este egal cu numărul de descendenţi viabili obţinuţi în urma selecţiei, încrucişării şi mutaţiei.

Întrucât selecţia şi supravieţuirea au la bază scorurile de selecţie şi de supravieţuire, se realizează folosind o ***modalitate de selecţie şi supravieţuire*** [22].

---

[13] Mutaţie = operatorul care introduce modificări noi (inexistente în eşantionul unei generaţii); ceea ce este caracteristic în general mutaţiei 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

[14] Descendenţi = genotipurile obţinute din încrucişarea şi eventual mutaţia indivizilor din eşantion

[15] Selecţie = operatorul cu ajutorul căruia se extrage din eşantion mai mulţi indivizi care participă la înmulţire

[16] Scor de selecţie = valoare numerică asociată individului din eşantion calculată pe baza (sau exprimată din) tăriei fenotipului în cultivar

[17] 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).

[18] Supravieţuire = operatorul cu ajutorul căruia se extrage din eşantion mai mulţi indivizi care vor fi înlocuiţi în eşantion de către descendenţi

[19] 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)

[20] 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)

[21] Funcţia obiectiv = algoritmul (procedura) de calcul al parametrului sau caracteristicii ce constituie obiectivul evoluţiei eşantionului.

[22] modalitate de selecţie şi modalitate de supravieţuire = metodă de extragere a unui individ din eşantion ce foloseşte drept parametru valorile scorurilor (de selecţie şi respectiv de supravieţuire) ale indivizilor ce compun eşantionul; ceea ce Tabelul 38 prezintă în mod formal exprimă faptul că se folosesc trei alternative de selecţie (proporţional, deterministic şi turnir) care se pot aplica valorilor scorurilor sau rangurilor scorurilor (când selecţia 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 şi alta de maxim) care se actualizează global în fiecare generaţie pe parcursul întregii evoluţii a eşantionului în cultivar

Strategiile de selecție şi supraviețuire implementate sunt sintetizate în tabelul de mai jos.

Selecție (și supraviețuire) în algoritmul genetic implementat

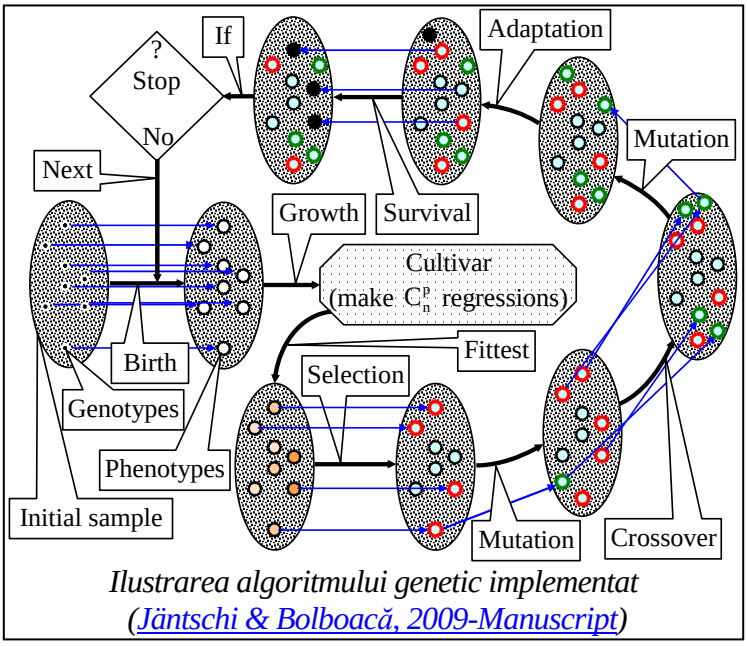
Metodă	Expresia funcției de scor	Selecție	Comentarii
Proporțional ÷ SL_Pro ÷ SV_Pro	f_i=Fitness(Cromozom_i)	$p_i=f_i/\Sigma_i f_i$	Șansa de selecție este proporțională cu scorul (utilizând probabilitatea p_i în selecție)
Deterministic ÷ SL_Det ÷ SV_Det		i | f_i = max. sau min.	Selecția indivizilor este 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

Căutarea MLR cu MDF definește o ***problemă dificilă*** [23], verificată (demonstrația fiind redată în Teză). Pentru a rezolva o ***problemă dificilă***, un algoritm genetic generează întâmplător (sau inițiază cu valori predefinite) un eșantion de un volum dat de genotipuri (volum ce va fi păstrat constant pe toată durata evoluției).

Algoritmul genetic (detaliat în Algoritm 4 - Anexa 5 din Teză) 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 eșantion ce întrunesc obiectivul în generație şi eventual include în generația următoare automat genotipurile din acest grup;
- 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;



Ilustrarea algoritmului genetic implementat (Jäntschi & Bolboacă, 2009-Manuscript)

[23] problemă dificilă = problemă a cărei rezolvare de către cel mai bun algoritm (care ar putea fi imaginat) necesită un timp de execuție care crește 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)

÷ 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).

Programul evolutiv [24] rezultat a fost gândit astfel încât să poată lucra cu oricare familie de descriptori moleculari (FPIF, MDF, MDFV, SAPF date în Anexa 4 din Teză) 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 din Teză.

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 miciDe corectat/reformulat) şi de predicție (pentru molecule care aparțin de aceaşi clasă de compuşi, dar care nu au fost incluşi în analiză).

Identificarea celui 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).

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

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

Definirea experimentului de simulare a evoluției

Fişierul de configurare a execuției algoritmului genetic, excluzând opțiunile de afişare, numără 30 de parametri (redați în Tabelul 5 - Anexa 5 din Teză), dintre care 19 parametrii ordinali şi 12 parametrii cu valori dintr-o listă de valori impuse (finită şi definită).Atenție: 19+12=31! Performanța algoritmului genetic a fost evaluată prin prisma rezultatelor colectate în fişierele de ieşire. Configurația parametrilor de ieşire este redată în Tabelul 6-2 din Teză.

În evaluarea algoritmului genetic s-a considerat de importanță teoretică şi 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 [25] este esența care stă la baza asocierii *Optimizare (SAR) → Evoluție (Observabile)*. Principiul a fost aplicat în simularea evoluției controlate către obiectivul evoluției, folosind diferite strategii de selecție şi supraviețuire. Principiul a avut ca scop evidențierea diferențelor în valorile observabilelor pe parcursul evoluției.

Conform acestui principiu, pe toată durata experimentului s-au definit şi păstrat aceeaşi parametrii intrinseci execuției programului evolutiv şi evoluției algoritmului genetic.

Parametri de configurare în experimentul de execuție a programului evolutiv

Clasă	Parametru	Valoare
Topologia familiei de descriptori moleculari	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	Mydb	MDFSARs
	TabE	PCB_lkow_data
	TabM	PCB_lkow_tmpx
Material genetic şi 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	cn0_CROSSOVER_Pairs	2
	m_m_MUTATION_Genes	2
	mpp_MUTATION_Parent_probability	5%
	mcp_MUTATION_Child_probability	5%
Obiectivul evoluției	rn0_REGRESSION_Multiple	4
	b_p_SELECTION_parameter	r2
	b_o_SELECTION_objective	max
Evoluție	e1n_GENERATIONS_max	20000
	e0n_RUNS_number	46
Selecție	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	v_p_SURVIVAL_phenotyping_p	1.0
	v_g_SURVIVAL__genotyping_p	1.0
	vfr_SURVIVAL_ranks	No

Doi parametri: strategia de selecție (*sfs_FITNESS_strategy* în fişierul de configurare - vezi

[25] 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*.

Tabelul 6-1 în Teză) şi strategia de supravieţuire (*vfs_SURVIVAL_strategy* în fişierul de configurare - vezi Tabelul 6-1 în Teză) au luat pe rând valorile: *proportional* (pentru strategie proporţională), *deterministic* (pentru strategie deterministă) şi *tournament* (pentru strategie în turnir). S-au proiectat astfel nouă execuţii ale programului evolutiv, în fiecare din cele nouă execuţii dând valori diferite celor doi parametrii ce definesc cele două strategii, ceilalţi parametrii rămânând cu aceleaşi valori pe toată durata experimentului.

Modalităţi de selecţie şi supravieţuire: design experimental de execuţie program

Supravieţuire / Selecţie	Proporţional (P) SV_Pro	Deterministic (D) SV_Det	Turnir (T) SV_Tur
Proporţional (P) - SL_Pro	P:P (1)	P:D (2)	P:T (3)
Deterministic (D) - SL_Det	D:P (4)	D:D (5)	D:T (6)
Turnir (T) - SL_Tur	T:P (7)	T:D (8)	T:T (9)

Rezultate şi discuţii

Strategiile implementate pentru selecţie şi supravieţuire

Pentru a realiza o selecţie bazată pe un scor de selecţie se impune parcurgerea unei serii de paşi. Algoritmul ***FS*** (Algoritm 1 în Teză - pentru scorurile de selecţie) parcurge o astfel de serie de paşi şi realizează astfel premizele pentru aplicarea unei strategii de selecţie. Algoritmul ***PS*** (Algoritm 2 - pentru strategia proporţională) realizează o selecţie proporţională folosind un şir de scoruri de selecţie şi dă o şansă de selecţie proporţională cu scorul de selecţie. Algoritmul ***DS*** (Algoritm 3 - pentru strategia deterministă) realizează o selecţie deterministă folosind un şir de scoruri de selecţie extrăgând cele mai mari N_Sel scoruri de selecţie. Algoritmul ***TS*** (Algoritm 4 - pentru strategia în turnir) realizează o selecţie în turnir folosind un şir de scoruri de selecţie extrăgând acele valori care se califică în urma unui turnir între două valori candidate repetat de un număr de ori dat de N_Sel. Scorul de supravieţuire (***VS***) este o valoare compozită, menită să asigure deopotrivă diversitatea genotipică şi cea fenotipică în cultivar. Pentru aceasta, două măsuri de similaritate între doi descriptori intră în expresia scorului de supravieţuire: o măsură de similaritate fenotipică (***VSP*** - dată de diferenţa între valorile scorurilor de selecţie) şi o măsură de similaritate genotipică (***VSG*** - dată de diferenţa între codurile genetice ale celor doi descriptori). Tabelul 7-1 din Teză redă expresiile de calcul folosite. Odată obţinute valorile ***VS*** ale scorurilor de supravieţuire în modalitatea expusă în tabelul de mai sus (în care scara de similaritate individuală are acum acelaşi obiectiv - cele mai mari valori dau cei mai potenţi candidaţi pentru înlocuire în materialul genetic) pentru şirul de indivizi reprezentaţi în cultivar, algoritmul proiectat pentru pregătirea scorurilor de selecţie ***FS*** este legitimat a fi folosit şi pentru scorul de supravieţuire ***VS***, ceea ce a şi fost făcut (motiv pentru care nu există descris un algoritm separat pentru ***VS***). În mod similar, algoritmii ce implementează strategiile proporţională, deterministă şi în turnir sunt perfect legitimaţi pentru a fi folosiţi şi pentru şirul scorurilor de supravieţuire (***PV*** = ***PS***, ***DV*** = ***DS***, ***TV*** = ***TS***); nici pentru aceştia nu există descriere separată.

*Algoritm 2. Algoritmul **SP** ce realizează o strategie de selecţie proporţională*

Scorul de selecţie se calculează astfel:
- Date de intrare:
  - ÷ FS_Array - şirul scorurilor de selecţie pentru fiecare genotip (n_of_sample valori)
  - ÷ FSD_Array - şirul valorilor distincte de scor de selecţie (cel mult n_of_sample valori)
  - ÷ FSC_Array - şirul numărului de apariţii pentru fiecare scor distinct de selecţie
  - ÷ N_Sel - număr de selecţii pentru încrucişare şi mutaţie;
- Iniţializează Selected_Genotypes_Array la ∅ (mulţimea vidă);
- Pentru fiecare selecţie (de la 1 la N_Sel)
  - Calculează suma scorurilor pentru genotipurile încă neselectate în FS_Sum;
  - Generează întâmplător (distribuţie uniformă) un număr FS_Freq între 0 şi FS_Sum (inclusiv);
  - Găseşte primul (indice) Group din FSD_Array pentru care FS_Freq ≤ $\Sigma_{i \le Group}$FSD_Array[i]*FSC_Array[i];
  - Generează întâmplător (distribuţie uniformă) un număr FSD_Next între 1 şi FSC_Array[Group] (inclusiv)
  - Adaugă a FSD_Next-a valoare egală cu FSD_Array[Group] neselectată încă din FS_Array în Selected_Genotypes_Array şi scade o unitate din FSC_Array[Group];
- Sfârşit 'Pentru'
- Date de ieşire: Selected_Genotypes_Array - şirul genotipurilor selectate (în număr de N_Sel)

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

Scorul de selecţie se calculează astfel:
- Date de intrare:
  - ÷ FS_Array - şirul scorurilor de selecţie pentru fiecare genotip (n_of_sample valori)
  - ÷ FSD_Array - şirul valorilor distincte de scor de selecţie (cel mult n_of_sample valori)
  - ÷ FSC_Array - şirul numărului de apariţii pentru fiecare scor distinct de selecţie
  - ÷ N_Sel - număr de selecţii pentru încrucişare şi mutaţie;
- Iniţializează Selected_Genotypes_Array la ∅ (mulţimea vidă);
- Iniţializează Already_Selected la 0;
- Iniţializează 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);
├ Sfârşit 'Cât timp'
├ Cât timp Already_Selected ≤ N_Sel (au mai rămas câteva scoruri identice într-un ultim grup din care trebuie făcută o
      selecţie):
        ├ Generează întâmplător (distribuţie uniformă) un număr FSD_Next între 1 şi FSC_Array[Group] (inclusiv)
        ├ Adaugă a FSD_Next-a valoare egală cu FSD_Array[Group] neselectată încă din FS_Array în
            Selected_Genotypes_Array şi scade o unitate din FSC_Array[Group];
├ Sfârşit 'Cât timp'
┌ Date de ieşire: Selected_Genotypes_Array - şirul genotipurilor selectate (în număr de N_Sel)
```

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

```
Scorul de selecţie se calculează astfel:
└ Date de intrare:
÷ FS_Array - şirul scorurilor de selecţie pentru fiecare genotip (n_of_sample valori)
÷ FSD_Array - şirul valorilor distincte de scor de selecţie (cel mult n_of_sample valori)
÷ FSC_Array - şirul numărului de apariţii pentru fiecare scor distinct de selecţie
÷ N_Sel - număr de selecţii pentru încrucişare şi mutaţie;
├ Iniţializează Selected_Genotypes_Array la o permutare întâmplătoare (distribuţie uniformă) a mulţimii
    {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 iteraţiei 'Pentru';
            ├ Permută în FS_Array valorile de pe poziţiile i_Sel & i_Sel-1;
        ├ Sfârşit 'Dacă'
├ Sfârşit 'Pentru'
├ Dacă N_Sel < n_of_sample atunci (ultimul nu a participat în turnir şi încă mai sunt genotipuri cu care să concureze în
      eşantion)
        ├ Generează întâmplător (distribuţie uniformă) un număr i_Sel între N_Sel + 1 şi 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 poziţiile N_Sel & i_Sel;
        ├ Sfârşit 'Dacă'
├ Sfârşit 'Dacă'
┌ Date de ieşire: Selected_Genotypes_Array - şirul genotipurilor selectate (în număr de N_Sel)
```

Fişiere rezultat

Execuţia programului evolutiv s-a făcut pe calculatoare din generaţia P6 (Dual P5) în perioada Ianuarie - Februarie 2009 şi rezultatele au fost salvate de program în fişierele date în tabelul de mai jos.

Fişiere rezultat (configurare şi evoluţie) după designul experimental

Selecţie	Supravieţuire	Configurare	Evoluţie
Proporţional	Proporţional	PCB_4044_cfg.txt	PCB_4044_evo.txt
Proporţional	Deterministic	PCB_2441_cfg.txt	PCB_2441_evo.txt
Proporţional	Turnir	PCB_9878_cfg.txt	PCB_9878_cfg.txt
Deterministic	Proporţional	PCB_5108_cfg.txt	PCB_5108_evo.txt
Deterministic	Deterministic	PCB_6369_cfg.txt	PCB_6369_evo.txt
Deterministic	Turnir	PCB_6690_cfg.txt	PCB_6690_evo.txt
Turnir	Proporţional	PCB_5828_cfg.txt	PCB_5828_evo.txt
Turnir	Deterministic	PCB_4872_cfg.txt	PCB_4872_evo.txt
Turnir	Turnir	PCB_1758_cfg.txt	PCB_1758_evo.txt

Disponibile pentru descărcare de la adresa:
http://l.academicdirect.org/Horticulture/GAs/MLR_MDF_selection_vs_survival/

Verificarea datelor: Testul Benford

S-au cumulat frecvențe din 46 de execuții independente ale numărului (num_obs) și aparițiilor (sum_obs) 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} (Tabelul 7-3 în Teză).

S-au aplicat testele Chi-Square și Kolmogorov-Smirnov pentru a verifica dacă numerele urmează legea Benford. Tabelele de mai jos redau această analiză.

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

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: $X^2=7<15.5=\chi^2(10\text{-}2, 5\%)$;

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

Cifra			Frecvență cumulată așteptată și observată						Diferență			| Diferență |		
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 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 primare (Tabelul 7-4 din Teză), 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 şi 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 şi suma diferențelor modulelor constituie o statistică suficientă. Pentru diferențele din Tabelul 7-4 din Teză 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 este că nu se poate respinge ipoteza distribuției Benford cu o probabilitate de 95% (la nivelul de semnificație de 5%). Mai mult, valorile probabilităților din distribuția Kolmogorov arată că un număr relativ scăzut de date experimentale concordă mai bine cu legea de distribuție Benford (9.18% pentru d_0; 4.84% pentru d_1; 5.4% pentru d_2).

Analiza variabilității

Frecvența de apariție a genotipurilor în eşantion de-a lungul evoluțiilor permite aprecieri cu privire la capacitatea de adaptare a acestora, şi în acelaşi timp reprezintă o măsură a variabilității materialului genetic al eşantionului pe care o induce metoda de selecție şi metoda de supraviețuire.

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

Pentru testarea independenței între metodele de selecție şi supraviețuire în ceea ce priveşte numărul de genotipuri s-a folosit testul $\chi 2$ aplicat la o tabelă de contingență de 3x3 pentru fiecare serie de valori numerice din Tabelul 8-1 din Teză şi rezultatele sunt în Tabelele 8-2, ...8-7 din Teză.

(Tabelul 8-2 în Teză): Sunt independente metoda de selecție față de metoda de supraviețuire în ceea ce priveşte numărul de genotipuri distincte din cultivar în generațiile ce produc evoluție? - NU

(Tabelul 8-3 în Teză): Sunt independente metoda de selecție față de metoda de supraviețuire în ceea ce priveşte numărul total de genotipuri din cultivar în generațiile ce produc evoluție? - NU

(Tabelul 8-4 în Teză): Sunt independente metoda de selecție față de metoda de supraviețuire în ceea ce priveşte genotipurile participante la regresiile ce produc evoluție? - NU

(Tabelul 8-5 în Teză): Sunt independente metoda de selecție față de metoda de supraviețuire în ceea ce priveşte numărul de genotipuri distincte din Top 23 în generațiile ce produc evoluție? - NU

(Tabelul 8-6 în Teză): Sunt independente metoda de selecție față de metoda de supraviețuire în ceea ce priveşte numărul total de genotipuri din Top 23 în generațiile ce produc evoluție? - NU

(Tabelul 8-7 în Teză): Sunt independente metoda de selecție față de metoda de supraviețuire în ceea ce priveşte genotipurile din Top 23 participante la regresiile ce produc evoluție? - NU

Confidența în dependența de strategia de selecție şi 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 observații nu creşte în aceeaşi ordine. În baza dependenței remarcate în strategia de selecție şi de supraviețuire la toți parametrii ce caracterizează numărul de genotipuri pe parcursul evoluției, s-a impus o caracterizare a acestei dependențe.

La întrebarea "Există legătură între cele trei serii de numere de genotipuri?" se răspunde calculând coeficientul de corelație (calcul dat în tabelul de mai jos).

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

Serii	Coeficient de determinare	Valoare F; probabilitate de a greşi
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}

S-au calculat media şi deviația standard (în ipoteza că distribuția de eşantionare induce o distribuție normală a acestor statistici ale eşantionului) în jurul parametrului statistic (al populației) asociat. Teorema Limită Centrală asigură faptul că se pot folosi în această analiză statistică valoarea medie (m) şi abaterea standard (s), prezentate în Tabelul 3 & Tabelul 4 - Anexa 3-1 din Teză, şi pe baza cărora se poate exprima intervalul de încredere al acestora din distribuția Student t. S-a folosit aplicația Statistica pentru a calcula aceste valori.

Analiza valorilor medii a permis formularea următoarelor concluzii:

÷ 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 scăderea 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 creşterea

semnificativă statistic numai în ceea ce privește cele mai frecvente genotipuri pentru toți parametrii (Distincți, Apariții și 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 și Participări în regresii).

Analiza abaterilor standard a permis formularea următoarelor concluzii:

÷ 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 și Participări în regresii), și luând pentru exemplificare numărul de genotipuri distincte (Num), se obține:
  - 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 aceeași 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) și respectiv turnir (T).

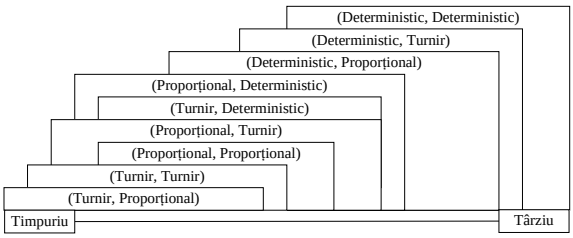
Pe baza rezultatelor experimentale obținute, poate fi interpretat și un alt parametru important al evoluției: numărul de generații care produc evoluție în cursul execuției cu număr impus de generații (20000), ca măsură a capacității de adaptare definită de combinația celor două metode (de selecție și supraviețuire), precum și valoarea medie a numerelor generațiilor care produc evoluție ca măsură a vitezei de adaptare.

Rezultatele au fost prelucrate (în Tabelul 8-11 din Teză) astfel: pentru fiecare execuție independentă a algoritmului genetic s-au consemnat numerele consecutive ale generațiilor care au produs îmbunătățirea valorii funcției obiectiv, și 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 întotdeauna este mai mic decât numărul maxim de generații al unei execuții) și numărul de evoluții distincte (numărul de valori), informații care sunt prezentate în Tabelul 8-11 din Teză. Întrucât ambele valori (media și 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 eșantionare, astfel încât s-a putut presupune aproximația la normalitate a acestora (atât cele ca valori ale eșantionului de 46 de observații) cât și populația din care provin, care așa 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 din Teză au fost supuse unei analize statistice descriptive, care a inclus calcularea valorilor medii și a deviațiilor standard, împreună cu intervalele de încredere la un nivel de semnificație de 95%, rezultate care sunt redate în tabelul de mai jos. Datele din tabelul de mai jos au servit pentru caracterizarea genotipurilor reprezentate în cultivar (numărul și media acestora). Caracterizarea s-a făcut pentru fiecare parametru statistic calculat, atribuind semnificațiile fizice măsurilor folosite și interpretând valorile obținute. Rezultatele acestei caracterizări sunt redate în continuare.

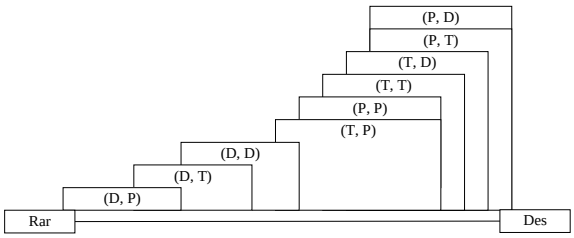
Media generațiilor care produc evoluții este o măsură a ***cât de timpuriu / târziu se produc evoluțiile***. Figura (Figura 8-2 în Teză) reprezintă valorile mediei și intervalului său de încredere; analizând, se remarcă: selecția deterministă (și cu atât mai mult însoțită de supraviețuirea deterministă) produce cele mai târzii evoluții; selecția turnir (și cu atât mai mult însoțită de supraviețuirea proporțională sau turnir) produce cele mai timpurii evoluții.

Statistici ale generațiilor ce produc evoluție în funcție de strategia de selecție și supraviețuire

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

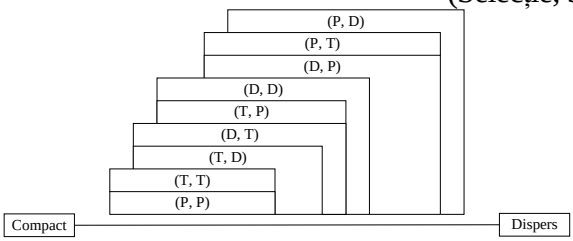


[CI(95%,Medie$_{n=46}$(generație medie))]
Cât de timpuriu se produc evoluțiile?

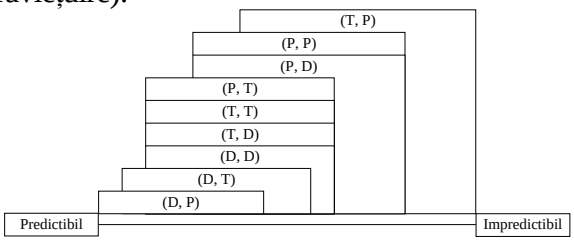


[CI(95%,Medie$_{n=46}$(număr evoluții))]
Cât de frecvent se produc evoluțiile?

(Selecție, Supraviețuire):



[CI(95%,Deviație$_{n=46}$(generație medie))]
Cât de dispers se produc evoluțiile?



[CI(95%,Deviație$_{n=46}$(număr evoluții))]
Cât de predictibil se produc evoluțiile?

Numărul de evoluții dintr-un număr dat de generații este o măsură a ***cât de frecvent se produc evoluțiile***. Figura (Figura 8-3 în Teză) reprezintă valorile mediei şi intervalului său de încredere; analizând, se remarcă: 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 privește frecvența evoluțiilor, 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 aceeași selecție proporționalăDe verificat/corectat! 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).

Variabilitatea momentului evoluției pe parcursul generațiilor este o măsură a ***cât de compact / dispers se produc evoluțiile***. Figura (Figura 8-4 în Teză) reprezintă valorile deviației

standard şi intervalului său de încredere; analizând, se remarcă: cea mai bună compactitate de evoluție o au selecția proporțională asociată cu supraviețuirea proporțională şi selecția în turnir asociată cu supraviețuirea în turnir; 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.

Variabilitatea numărului evoluțiilor este o măsură a ***cât de predictibil / impredictibil se produc evoluțiile***. Figura (Figura 8-5 în Teză) reprezintă valorile deviației standard şi intervalului său de încredere; analizând, se remarcă: 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.

Analiza diversității

Diversitatea genotipurilor reprezentate în cultivar în momentele evoluției se poate cuantifica folosind entropia informațională, având la dispoziție o familie de măsuri entropice, date de expresia entropiei generalizate (sau Rényi) - (Rényi, 1961) - $H(p_1, p_2, ..., p_n; \alpha) = H_\alpha(p_1,p_2, ...,p_n)$ şi unde are loc relația:

$H(\{p\};\infty)$ $=H_\infty(\{p\})$	$\leq$	$H(\{p\};2)$ $=H_2(\{p\})$	$\leq$	$H(\{p\};1)$ $=H_1(\{p\})$	$\leq$	$H(\{p\};0)$ $=H_0(\{p\})$	$\leq$	$2H(\{p\};\infty)$ $=2H_\infty(\{p\})$
H_∞: entropia min; H_2: logaritmul cu semn schimbat al 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 ştiințele 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:

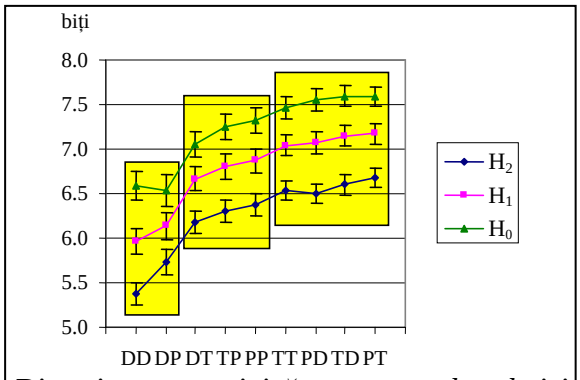
÷ Au fost extrase pentru fiecare execuție, fiecare asociere (selecție, supraviețuire) şi fiecare evoluție lista genotipurilor reprezentate în cultivar în momentul evoluției (coloanele Gen0..Gen11 în fişierele de rezultate);

÷ 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;



Diversitatea genotipică pe parcursul evoluției

÷ 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 din Teză şi s-au reprezentat grafic în ordinea crescătoare a valorii medii împreună cu intervalul de încredere în figura alăturată.

Graficul arată că cele nouă strategii de evoluție se grupează în 3 grupe după diversitatea genotipică care o produc în cultivar:

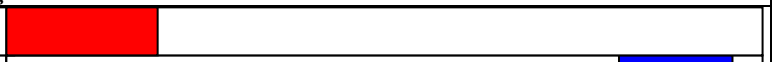
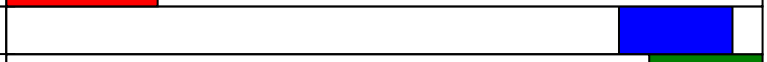
÷ Grupul (DD, DP) - selecție deterministă, supraviețuire deterministă sau proporțională este semnificativ statistic diferit în diversitate de toate celelalte strategii în toate trei măsurile entropice; produc cea mai scăzută diversitate genotipică;

÷ Pentru grupul (DT,TP,PP) - (selecție deterministă, supraviețuire în turnir) împreună cu (supraviețuire proporțională, selecție în turnir sau proporțională) se poate evidenția 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 şi supraviețuire în turnir), (selecție proporțională şi supraviețuire deterministă), (selecție în turnir şi supraviețuire deterministă), şi respectiv selecție proporțională şi supraviețuire în turnir.

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

Entropia Shannon (H_1) pentru selecție şi pentru supraviețuire

Strategie	Medii şi intervale de încredere la riscul de a fi în eroare de 5%	
Selecție	D(6.26±0.10)	
	T(7.00±0.07)	
	P(7.04±0.07)	
Supraviețuire	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		

Analiza statistică din tabel 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 şi cea în turnir nu este semnificativă statistic la riscul de a fi în eroare de 5%. Strategia de supraviețuire are, de asemenea, 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 şi 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.

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 din Teză) trebuie investigate un număr de 2^9-C_9^0-C_9^1 = 502 inferențe statistice. Rezultatul analizei este redat în Tabelul din Anexa 7-1.

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

÷ Analiza grupurilor de câte două perechi de metode evidențiază statistic că:
  - o 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:
  - o Cu un risc de a fi în eroare de 5% următoarele provin din populații diferite: DD, PD, TD;
  - o Nu poate fi respinsă ipoteza că DT şi DP provin din populații identice;
  - o 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:
  - o Nu poate fi respinsă ipoteza că PP, PT şi TT (c/k = 2.2) provin din populații identice;
  - o 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:
  - o Cu un risc de a fi în eroare de 5% următoarele provin din populații diferite: DD, PD, TD;
  - o Nu poate fi respinsă ipoteza că DT şi DP provin din populații identice;
  - o Nu poate fi respinsă ipoteza că TT şi TP provin din populații identice;
  - o 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:
  - o 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:
  - o Cu un risc de a fi în eroare de 5% DD, PD, şi TD provin din populații diferite;
  - o Nu poate fi respinsă ipoteza că DT şi DP provin din populații identice;
  - o Nu poate fi respinsă ipoteza că TT şi TP provin din populații identice;
  - o Nu poate fi respinsă ipoteza că PP, PT şi TT provin din populații identice;
  - o Cu un risc de a fi în eroare de 5% se respinge ipoteza că PP, PT, TP şi TT ar proveni din populații identice;

Analiza numărului de fenotipuri viabile din cultivar în generațiile ce au produs evoluție

Tabelul 9-3 din Teză 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 din Teză, 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 din Teză) 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 generaţiile ce produc evoluție este mult mai complexă decât interpretarea populațională (Tabelul din Anexa 7-1) a numărului mediu de genotipuri în aceleaşi evoluții.

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 regăsește 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 regăsește 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 regăsește 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};

La finele procesului, se poate întocmi 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 din Teză, clasificat după nivelul de confidență al rezultatului obținut.

Analiza numărului de asocieri (regresii) viabile din cultivar în generațiile ce au produs evoluție

Tabelul 9-6 din Teză 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).

Ipoteza cu privire la numărul mediu de asocieri viabile (valori redate în Tabelul 9-6 din Teză) 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 același mod în care s-au interpretat datele prezentate din Tabelul din Anexa 7-2.

Cele mai mari grupuri de ordin maxim nediscriminate 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};

La finele procesului, se poate întocmi 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 din Teză, 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%.

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 verificată în acest context a fost 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 [26]. 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 comparativ cu alte teste statistice. Din acest motiv, verificarea ipotezei statistice folosind ambele teste este necesară. Tabelul de mai jos (Tabelul 10-1 în Teză) sumarizează această procedură.

Ipoteza de distribuție uniformă a numerelor aleatoare generate cu Rand(·,·) în PHP (numerele vor fi generații)

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.

S-a investigat dacă numerele ar putea proveni din altă distribuție. Analiza este redată în Tabelul 10-2 în Teză, când s-a constatat că acceptarea ipotezei că datele ar putea proveni dintr-o lege de distribuție (se respinge ipoteza că cele două 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. În acelaşi timp, raportul între statistica ajustată (Anderson-Darling) şi valoarea sa la pragul de 10% arată că cea mai probabilă lege de distribuție este distribuția uniformă (cu 2.3) urmată de distribuția valorilor extreme (cu 2.2) şi Pareto generalizată (cu 2.1).

Căutarea legii de distribuție a coeficientului de determinare al unei generații

A doua ipoteză care trebuie verificată este dacă valorile coeficientului de determinare obținut în cele 46 de execuții independente ale algoritmului genetic pentru fiecare pereche de strategie de selecție şi supraviețuire provin dintr-o populație distribuită după o lege de distribuție (având la dispoziție o serie de alternative). Pentru aceasta, fiecare din cele nouă generații extrase din distribuția uniformă a generațiilor 0..20000 a constituit subiectul investigației, rezultatele fiind redate

[26] Mads HAAHR. 1998-2009. ©. http://www.random.org/randomness/

în Tabelele 10-3-DD..10-3-TT din Teză (folosind alternativele de distribuție: DE - dublu exponențială; LG - logistică; LN - log-normală; NO - normală; UN - uniformă (0,1); EX - exponențială; GU - valori extreme de tip I) pentru fiecare asociere de strategie de selecție şi supraviețuire incluse în studiu. Tabelele 10-3-DD..10-3-TT din Teză 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 din Teză cumulează aceste observații. Tabelul de mai jos (Tabelul 10-4 din Teză) re-sintetizează rezultatele din Tabelul 10-3-SS din Teză. Se observă că pentru selecția deterministă însoțită de supraviețuirea deterministă nici o lege nu a întrunit acceptarea.

Legi de distribuție ale coeficientului de determinare într-o generație pentru diferite strategii de selecție şi supraviețuire

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				

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. Rezultatele sintetizate 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).

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 eşantionul de 46 de valori drept estimator nedeplasat al mediei populației tuturor execuțiilor independente iar varianța drept statistică suficientă.

Distribuția Gumbel nu este suficient de generală

Au fost făcute estimările parametrilor 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.

$\hat{Y}(X) = \dfrac{\dfrac{X^{\gamma}-1}{\gamma} - \alpha}{\beta}$	eq.1	
	eq.2	$\hat{Y}(X) = 0.907\left(1 - \dfrac{X^2 + a_3X + a_2}{X^2 + a_1X + a_0}\right)$

A fost definită o funcție putere (eq. 1) 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 priveşte tendința centrală. A fost definită o funcție rațională (eq. 2) şi parametrii de interpolare (α, β, şi γ), intervalele de încredere şi semnificația statistică *t* sociată De verificat: e corect t sociata? au fost obținute pentru fiecare caz în parte în ceea ce priveşte valorile extreme (minimul şi maximul). Această ultimă analiză, a valorilor extreme, a arătat că distribuția Gumbel nu poate fi acceptată drept lege de distribuție, probabilitățile de apariție ale valorilor extreme calculate fiind inaceptabil de depărtate 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 prezentat dintr-un motiv foarte simplu: concluzia studiului cuprins în Anexa 8 a fost respingerea ipotezei de distribuție după legea valorilor extreme de tip I.

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 de mai sus (Tabelul 10-4 în Teză) relevă faptul că pentru majoritatea cazurilor observate legea de distribuție Gumbel este totuşi acceptată, consecința rezultată este 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 s-a cerut rezolvată a fost verificarea ipotezei de distribuție după legea Fisher-Tippett a observațiilor experimentale. Pentru a realiza această problemă, s-a folosit aplicația EasyFit [27]. 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.

O a doua problemă care s-a cerut rezolvată a fost 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 [27].

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.

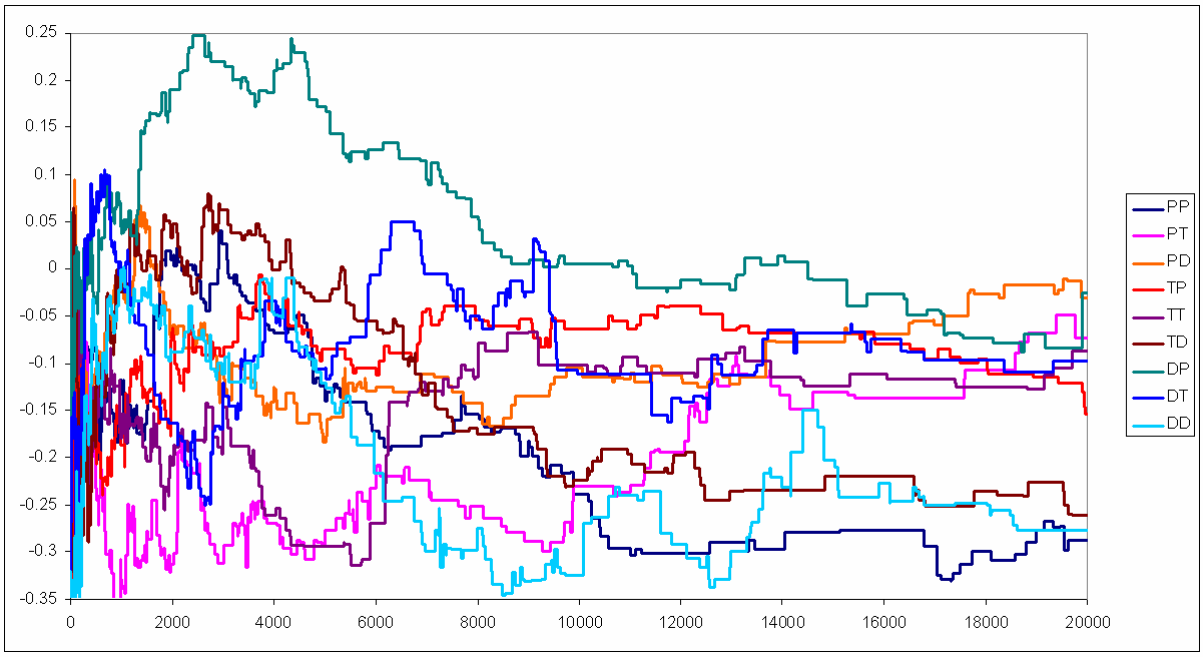
Însă cel mai important criteriu este respingerea statistică la unul sau mai multe nivele de

[27] EasyFit Professional v.50. 2008. Software. MathWave Technologies.

semnificație, şi această analiză dată în Tabelul 1 - Anexa 9 este sintetizată în Tabelul 11-1 din Teză. 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: Jonson SB - 20 (~25%), Pert - 18 (~22%).

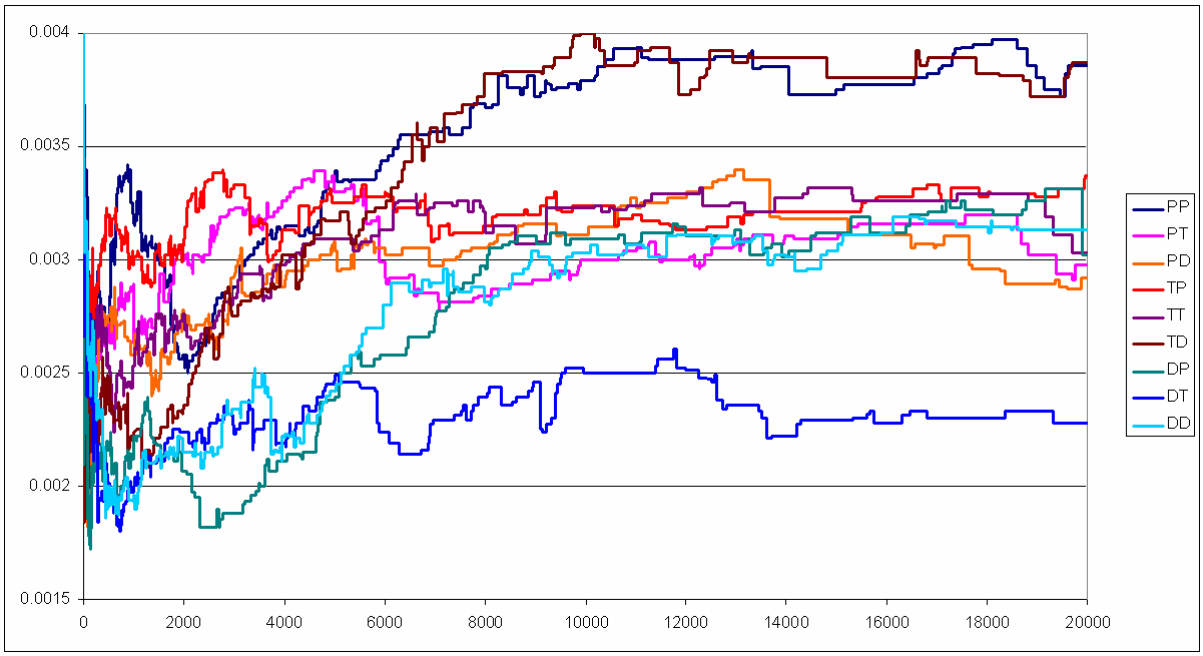
Evoluția locației, scalei şi formei distribuției Fisher-Tippett a obiectivului evoluției

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 de mai jos (Figura 11-2 din Teză).



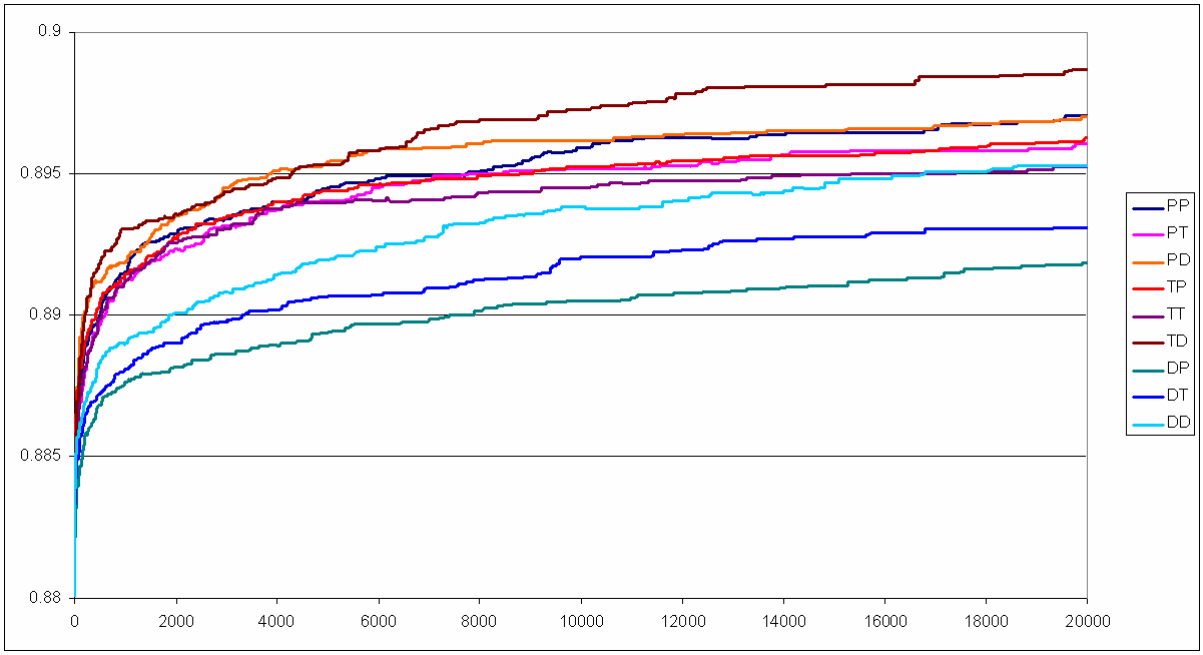
Parametrul formă (k) al distribuției Fisher-Tippett: estimare din observații

Rezultatul pentru parametrul scalei (β) este redat în figura de mai jos (Figura 11-6 din Teză).



Parametrul scală (β) al distribuției Fisher-Tippett: estimare din observații

Rezultatul pentru estimarea parametrului locație (λ) este redat în figura de mai jos (Figura 11-9 din Teză).



În valorile parametrilor distribuțiilor Fisher-Tippett s-au găsit ecuații de tendință, care sunt redate în tabelul de mai jos (Tabelul 11-7 din Teză).

Ecuațiile de tendință pentru formă (k), scală (β) și locație (λ) ale distribuțiilor Fisher-Tippett

SS	$k(G) = a_0 + a_1 \cdot G$		$\beta(G) = a_0 + a_1 \cdot G$		Tendință λ(G)	a_0	a_1	a_2
PP	-0.1912	$-1.47 \cdot 10^{-6}$	3.541E-3	5.5E-9	$\lambda(G) = a_0 + a_1 \cdot \ln(G + a_2)$	0.89357	$1.82 \cdot 10^{-4}$	0.867
PD	-0.0961	$3.12 \cdot 10^{-7}$	2.983E-3	1.9E-9		0.89422	$1.55 \cdot 10^{-4}$	-0.344
TP	-0.0833	$1.24 \cdot 10^{-7}$	3.192E-3	8.9E-10		0.89333	$1.54 \cdot 10^{-4}$	-0.213
TT	-0.1476	$5.58 \cdot 10^{-7}$	3.072E-3	2.9E-9		0.89286	$1.40 \cdot 10^{-4}$	-0.348
PT	-0.2108	$1.08 \cdot 10^{-6}$	2.996E-3	8.2E-10	$\lambda(G) = a_0 + a_1 \cdot \ln(G)$	0.89309	$1.69 \cdot 10^{-4}$	
TD	-0.1352	$-1.47 \cdot 10^{-6}$	3.419E-3	7.9E-9	$\lambda(G) = a_0 + a_1 \cdot G^{a_2}$	0.89465	$6.84 \cdot 10^{-4}$	0.117
DP	-0.0193	$-1.32 \cdot 10^{-6}$	2.730E-3	7.1E-9		0.88916	$2.02 \cdot 10^{-4}$	0.171
DT	-0.0797	$-1.35 \cdot 10^{-7}$	2.296E-3	6.1E-10		0.89016	$3.19 \cdot 10^{-4}$	0.151
DD	-0.0207	$-9.52 \cdot 10^{-7}$	2.745E-3	5.6E-9		0.89173	$2.93 \cdot 10^{-4}$	0.172

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).

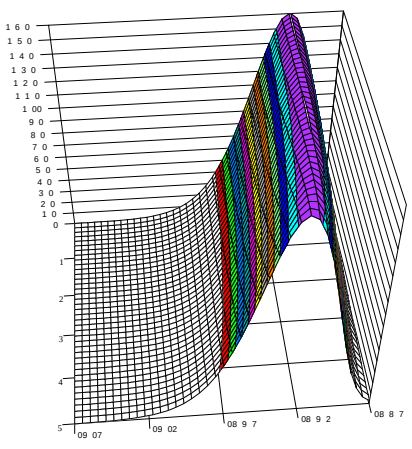
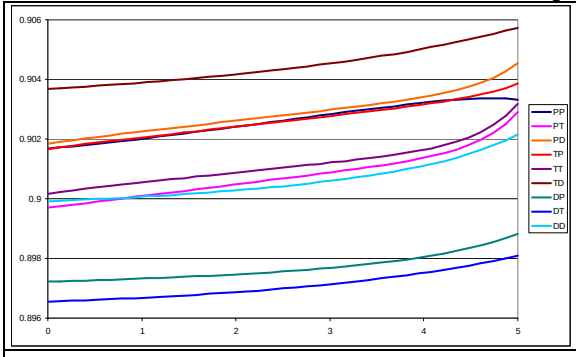


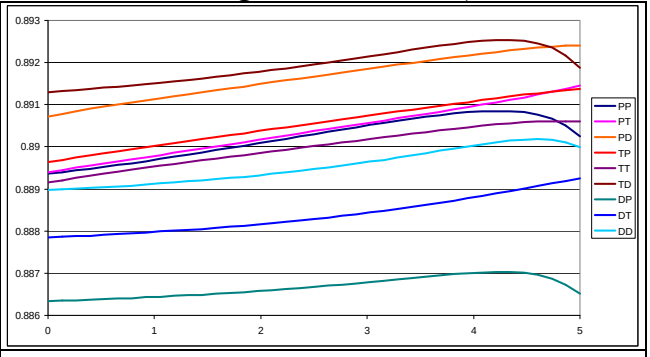
Figura alăturată este extrasă din Figura 1 - Anexa 10 și reprezintă tendința în densitatea de probabilitate Fisher-Tippet a coeficientului de determinare pentru strategia de selecție în turnir și strategia de supraviețuire deterministă - $DTFT_{PDF}(r^2, \log_{10}G)$.

Caracterizarea statistică a distribuției evoluției

Folosind ecuațiile de tendință pentru formă (k), scală (β) şi locație (λ) se pot obține frontierele de probabilitate 95% (în care şansa 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 în Teză) şi respectiv de probabilitate 5% (în care şansa 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 în Teză).



Loteria norocoşilor (CDF = 95%)



Loteria ghinioniştilor (CDF = 5%)

Se remarcă în figurile de mai sus comportarea selecției deterministe însoțite de supraviețuirea în turnir (DT) care este singura asociere de selecție şi supraviețuire al cărui interval de normalitate (fără şansă/noroc şi fără ghinion) rămâne aproximativ acelaşi pe parcursul evoluției. 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).

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. În tabelul următor sunt redate frecvențele observate în raport cu cele 3 categorii definite (Tabelul 11-8 din Teză).

Tipul formei distribuției Fisher-Tippett

Valori extreme de tip	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%)

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?*

În acest scop, s-a aplicat o procedură de transformare asupra datelor primare, procedură exemplificată în tabelul de mai jos (Tabelul 12-1 în Teză, exemplu din run 1, DP).

Transformarea momentelor evoluției la durate relative

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 astfel: s-au obținut momentele relative aşa cum descrie tabelul de mai sus pentru fiecare run (1..46) şi fiecare strategie (PP, PT, PD, TP, TT, TD, DP, DT, DD); fiecare şir de tipul celui din linia "Durata relativă la momentul evoluției" a tabelului de mai sus 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 completeze 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 pentru? măsurarea agrementului între observație şi model un număr de trei 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 de mai jos (Tabelul 12-2 în Teză, cuprinzând primele trei legi de distribuție în ordinea agrementului între observație şi model).

Cele mai probabile legi de distribuție pentru momentele relative ale evoluției

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 Log-P-3: log-Pearson de tipul 3									

Rezultatele din tabelul de mai sus 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 şi 5.62% A-D).
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 foloseşte ranguri (ca şi K-S) este totuşi 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 şi K-S (remarca 2 de mai sus) şi valorile obținute pentru aceste statistici (remarca 3 de mai sus) se desprinde concluzia că agrementul între rangurile observate ale momentelor relative ale evoluției şi rangurile din distribuția teoretică Log-Pearson 3 este mai probabil decât agrementul între valorile observate ale momentelor relative ale evoluției şi valorile din distribuția teoretică Log-Pearson 3.
5. Există o explicație pentru remarca 4, şi anume eşantionul de evoluții a cuprins observații din toate cele nouă asocieri de strategii; este astfel aşteptat ca agrementul să fie mult mai mare pentru ranguri (statisticile K-S şi A-D) decât pentru valori (statistica C-S).
6. Având în vedere remarcile 4 şi 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 greşi 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 menționată.

Concluzia că "*Log-Pearson 3 este legea de distribuție a momentelor relative ale evoluției, independent de strategia de selecție şi de supraviețuire*" se poate verifica. În acest sens, pentru fiecare pereche de strategie de selecție şi supraviețuire din cele 46 de execuții independente au fost cumulate momentele relative ale evoluției. Rezultatele sunt disponibile online:

http://l.academicdirect.org/Horticulture/GAs/MLR_MDF_selection_vs_survival/evolut_SS.txt

Rezultatele au intrat în aceeaşi analiză a agrementului între observație şi model, şi fişierele rezultat se află la aceeaşi adresă de mai sus, denumirile fişierelor fiind date în tabelul de mai jos (Tabelul 12-3 în Teză), împreună cu semnificațiile statistice ale agrementului între observații şi modelul dat de distribuția Log-Pearson 3.

Agrementul între observație şi modelul Log-Pearson 3 pentru distribuția momentelor relative ale evoluției

Stra\Stat	nr.Obs	K-S	$p_{K\text{-}S}$	A-D	$p_{A\text{-}D}$	C-S(df)	$p_{C\text{-}S}$
DD	TT	1379	0.02284	0.46	0.63251	0.47	12.3(10)
DP	TD	1429	0.01224	0.98	0.23477	0.75	3.3064(10)
DT	TP	1318	0.02691	0.29	1.2118	0.24	14.35(10)
PD	DT	996	0.02845	0.39	0.73496	0.41	10.628(9)
PP	DD	1084	0.01919	0.81	0.34184	0.66	8.1401(10)
PT	DP	851	0.02416	0.69	0.6234	0.47	6.8598(9)
TD	PT	1463	0.0203	0.58	0.70531	0.43	12.512(10)
TP	PD	1474	0.03055	0.13	0.93998	0.33	8.6564(10)
TT	PP	1353	0.01212	0.99	0.23201	0.75	3.5574(10)
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							

Rezultatele din tabelul de mai sus demonstrează că ipoteza "*Log-Pearson 3 este legea de distribuție a momentelor relative ale evoluției, independent de strategia de selecție şi de supraviețuire*" formulată pe baza analizei cuprinzând cumulat toate observațiile şi se verifică pentru fiecare strategie de selecție şi supraviețuire în parte.

Valorile din tabelul de mai sus 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 aşteptată (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 eşantionul 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% şi având o valoare medie de 53%; explicația pentru dezagrementul măsurat de statistica C-S între observație şi modelul Log-Pearson 3 remarcat în experimentul cu observațiile cumulate şi explicația de mai sus pentru agrementul măsurat de statistica C-S între observație şi modelul Log-Pearson 3 remarcat în experimentul cu observațiile

separate pe strategii aduce pe cale de consecință că:

÷ ipoteza "*Log-Pearson 3 este legea de distribuție a momentelor relative ale evoluției independent de strategia de selecție şi de supraviețuire*" este susținută de observațiile asupra fiecărei strategii în parte;

÷ parametrii legii de distribuție Log-Pearson 3 sunt însă dependenți de strategia de selecție şi de supraviețuire (dezagrement la X^2 la cumularea observațiilor, agrement la X^2 individual);

÷ 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 la observațiile ce provin dintr-o anumită strategie înregistrat în toate cele nouă 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 ale evoluției, iar parametrii distribuției depind de strategia de selecție şi de supraviețuire*".

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 susține această afirmație: r(α,β)=0.857; r(α,γ)=0.994; r(β,γ)=0.885. r(α,γ) = 0.994 arată că există o asociere liniară între formă (α) şi locație (γ). S-a evidențiat de asemenea existența unei dependențe după o funcție putere (r(β),pow(γ))>0.999) între scală (β) şi locație (γ). Astfel familia de curbe log-Pearson de tipul III este degenerată în caracterizarea evoluției:

$$LP3(x;\alpha_j,\beta_j,\gamma_j) \text{ degenerează în } LP3(x;8.77\cdot\gamma_j-68.3,-0.14-144\cdot\gamma_j^{-2.57},\gamma_j)$$

unde j ∈ {TT, TD, TP, DT, DD, DP, PT, PD, PP}.

Pentru a verifica ipoteza de degenerare sugerată de regresiile parametrilor, s-a estimat din nou parametrul locație pentru expresiile degenerate ale distribuției log-Peasron de tipul III cu un singur parametru independent, şi tabelul următor (Tabelul 12-10 în Teză) prezintă pentru comparație valorile probabilităților de observare ale distribuțiilor din eşantioane în ipoteza că ele provin din legea de distribuție log-Pearson de tipul III.

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

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

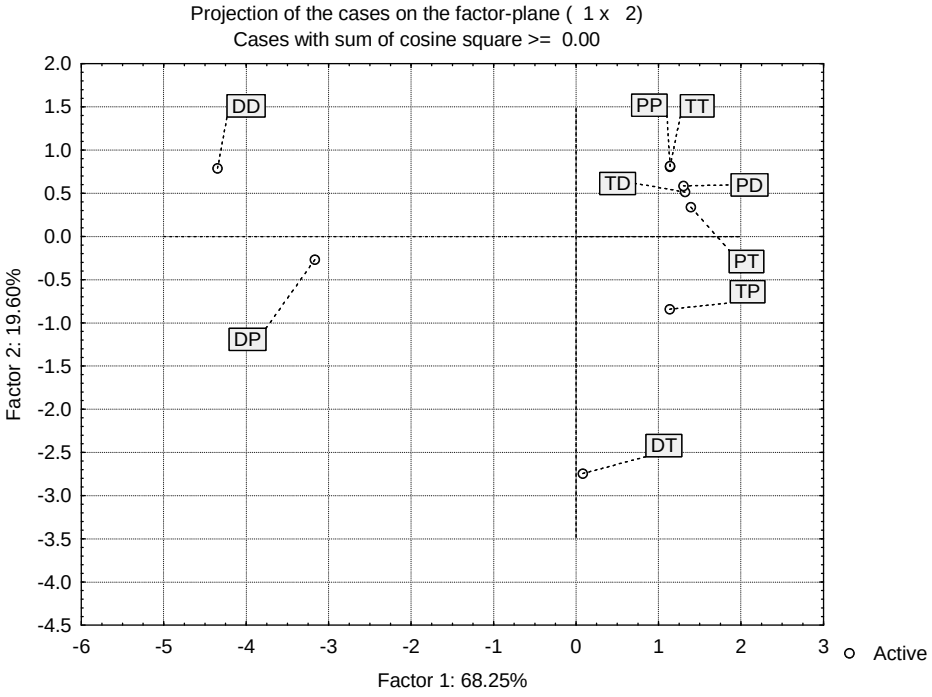
÷ $p_{K\text{-}S}$, $p_{A\text{-}D}$, $p_{C\text{-}S}$:
probabilități de observare întâmplătoare
÷ prima valoare p:
din MLE parametrii independenți
÷ a doua valoare p:
din MLE cu un parametru independent (γ)

Analiza sintetizată în tabelul de mai sus 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 de tipul III pentru momentele relative ale evoluției în seria de strategii de evoluție investigată (DD, DP, DT,

PD, PP, PT, TD, TP, TT).

Diferențe în familia de curbe de evoluție

S-au pus în evidență o serie de asocieri între valorile statisticilor distribuțiilor degenerate. 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 între statisticile distribuțiilor (locație, medie, modă, mediană, deviație standard, asimetrie, exces de boltire), s-a construit o analiză de componente principale (figura de mai jos, în Figura 12-3 în Teză, obținută folosind aplicația Statistica).



Proiecțiile primilor doi factori (principali) în valorile (γ, μ, $\hat{\mu}$, $\tilde{\mu}$, σ, γ_1, γ_2)

Figura de mai sus (Figura 12-3 în Teză) evidențiază următoarele: 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 din Teză - 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 din Teză - 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.

Ce distribuție urmează numărul de evoluții?

S-au derulat 10 analize, 9 dintre ele cu datele obținute în fiecare strategie în parte din fiecare execuție (46 observații în eşantionul numărului de evoluții) şi a 10-a cu toate la un loc (414 observații în eşantionul numărului de evoluții). A fost folosită aplicația EasyFit pentru a obține estimările parametrilor distribuțiilor şi a da măsurile agrementelor. S-a întocmit un clasament pe ranguri după statistici (C-S, A-D, K-S) care a dat cea mai probabilă distribuție ca fiind Fisher-Tippett (rang 284 față de peste 420 restul alternativelor în număr de peste 60).

Tabelul următor (Tabelul 12-14 în Teză) conține agrementul între observații şi modelul de distribuție Fisher-Tippett pentru fiecare din cele 10 eşantioane care au fost supuse analizei.

Agrementul pe care îl realizează distribuția Fisher-Tippett cu observațiile experimentale ale numărului de evoluții este de-a dreptul remarcabil. Nu numai că nu se semnalează respingeri semnificativ statistice la nici un risc de a fi în eroare uzual între 1% şi 20%, ci mai mult, agrementul

între model şi date este în medie de 86.51% în acord cu statistica Kolmogorov-Smirnov, de 72.33% în acord cu statistica Anderson-Darling şi de 71.73% în acord cu statistica Chi-Square.

Agrementul între Fisher-Tippett şi numărul de evoluţii dintr-o execuţie a algoritmului genetic

Strategie	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	414	0.0342	0.7066	0.307	0.6875	7.14/8	0.5218

În cea mai mică măsură agrementul a fost înregistrat de statistica Chi-Square (39.56%) pentru strategia TT şi de statisticile Kolmogorov-Smirnov (48.59%) şi Anderson-Darling (48.77%) pentru strategia TD. Tabelul următor (Tabelul 12-15 în Teză) redă statisticile comune cu privire la legile de distribuţie obţinute.

Statistici ale distribuţiilor Fisher-Tippett ale numărului de evoluţii către optim

Strategia	Distribuţia 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: Deviaţia standard; γ_1: Asimetria; γ_2: Excesul de boltire							

Rezultatele din tabelul de mai sus evidenţiază că legile de distribuţie sunt foarte apropiate una de cealaltă. În fapt, nu a fost respinsă nici ipoteza că toate împreună au o singură lege de distribuţie şi strategia nu creează populaţii distincte în numărul de evoluţii. Se poate remarca însă că valori la modă sub moda globală (26) au strategiile de selecţie deterministă (DP: 19; DT: 19; DD: 25) şi selecţia în turnir cuplată cu supravieţuirea proporţională (TP: 24) în timp ce valori la modă superioare modei globale (26) au strategiile de selecţie proporţională (PT: 31; PD: 30; PP:29) şi în turnir (TD: 30; TT: 28) - cu o excepţie, cea de mai sus (TP: 24). Tot rezultatele din tabelul de mai sus evidenţiază că o singură lege de distribuţie este a valorilor extreme de tipul II (Fréchet) - TP ($\alpha > 0$), toate celelalte, inclusiv distribuţia 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 distribuţia cu cea mai mare deviaţie standard (11.93). Un calcul simplu asupra deviaţiei standard date în tabelul de mai sus în ipoteza de independenţă între eşantioanele extrase din strategii diferite ($\sigma_\Sigma^2 = (\sigma_{TT}^2 + ... + \sigma_{PP}^2)/9$) ne permite să separăm varianţa totală ($\sigma_T^2 = 10.5^2$) în varianţă în interiorul strategiilor ($\sigma_\Sigma^2 = 9.68^2$) şi varianţă între strategii (4.07^2). Acest din urmă rezultat pune în evidenţă o măsură în care alegerea unei strategii influenţează evoluţia, contribuţia cantităţii varianţei între strategii fiind importantă în valoarea varianţei 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.

(***Concluzie***)

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.

(***Concluzie***)

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.

(***Concluzie***)

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/cultivarul, cu ajutorul algoritmului genetic; analiza evoluției a cuprins atât evoluția în ansamblu, cât şi evoluția bazată pe diferite strategii de selecție şi supraviețuire. Astfel, s-a constatat că obiectivul fixat al evoluției (coeficientul de determinare) se distribuie după o lege de distribuție generalizată: distribuția Fisher-Tippett. Numărul de evoluții către optim se distribuie după aceeaşi lege de distribuție. Momentele relative ale evoluției se distribuie după o lege degenerată (punându-se în evidență două dependențe între cei trei parametrii) din familia de curbe log-Pearson de tipul III.

(***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 eșantionul 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.

(***Concluzie***)

Analiza compoziției eșantionului de material genetic pe parcursul evoluției a dus la următoarele concluzii:

÷ Selecția deterministă, indiferent de strategia de supraviețuire a indivizilor în populație, are drept consecință o scădere semnificativă din punct de vedere statistic a numărului total de genotipuri distincte reprezentate în cultivar;

÷ Supraviețuirea deterministă, asociată cu selecția turnir sau proporțională, favorizează o creştere semnificativă din punct de vedere statistic (fapt evidențiat de către toate cele trei statistici folosite: Pearson-Fisher, Anderson-Darling şi Kolmogorov-Smirnov) a numărului total de genotipuri reprezentate în cultivar;

÷ Supraviețuirea deterministă favorizează o creştere semnificativă din punct de vedere statistic a numărului de genotipuri în grupul celor mai frecvente 23 din 46 de execuții independente în generațiile ce produc evoluție, în timp ce selecția deterministă determină scăderea semnificativ statistică a acestora.

(***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.

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.

Lucrări reprezentative publicate

÷ 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. (Jäntschi & alţii, 2008)

÷ 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. (Jäntschi & alţii, 2009)

÷ 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. (Jäntschi & alţii, 2010)

÷ Distribution Fitting 1. Parameters Estimation under Assumption of Agreement between Observation and Model. Lorentz JÄNTSCHI, *Bulletin UASVM, Horticulture* 66(2):684-690, 2009. ArXiv electronic library permanent link (July 16, 2009): http://arxiv.org/abs/0907.2829 (Subject: Statistics - Methodology).

÷ Distribution Fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Kramer-von-Misses and Jarque-Bera statistics. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ. *Bulletin UASVM, Horticulture* 66(2):691-697, 2009. ArXiv electronic library permanent link (July 16, 2009): http://arxiv.org/abs/0907.2832 (Subject: Statistics - Methodology).

÷ Distribution Fitting 3. Analysis under Normality Assumption. Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI. *Bulletin UASVM, Horticulture* 66(2):698-705, 2009. (Bolboacă & alţii, 2009)

÷ 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 ŢIGAN, Ioan ABRUDAN, Mugur C. BĂLAN. *Bulletin UASVM, Agriculture* 66(1):82-88, 2009. (Jäntschi & alţii, 2009)

÷ 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. (Jäntschi & alţii, 2010)

÷ Meta-heuristics on quantitative structure-activity relationships: study on polychlorinated biphenyls. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ. *Journal of Molecular Modeling* 16(2):377-386, 2010, DOI: 10.1007/s00894-009-0540-z.

÷ A Study of Genetic Algorithm Evolution on the Lipophilicity of Polychlorinated Biphenyls. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ. *Chemistry and Biodiversity*, 2010, DOI: 10.1002/cbdv.200900356.

÷ 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).