CONTRIBUȚII LA ANALIZA ȘI PRELUCRAREA DATELOR ÎN ANALIZA GENETICĂ

Rezumat al tezei destinate obţinerii
titlului ştiinţific de doctor inginer
la
Universitatea Politehnica Timişoara
în domeniul CALCULATOARE ȘI TEHNOLOGIA INFORMAȚIEI
de către

Dr. Med. Ing. Nicolae Teodor MELIȚĂ

Conducător ştiinţific:	prof.univ.dr.ing. Ștefan HOLBAN
Referenţi ştiinţifici:	prof.univ.dr. Horia CIOCÂRLIE.
	prof.univ.dr.ing. Ionel JIAN.
	conf.univ.dr.ing. Dan PESCARU.

Ziua susţinerii tezei: …31.10.2016…..

1. Introducere

1.1. Motivație

Evoluția fulminantă a tehnologiei din deceniile recente are un impact major asupra activității și cercetării în toate domeniile. Modalități noi, avansate, de achiziție a informației au fost imaginate și implementate pentru cele mai diverse sfere de interes teoretic si practic, iar capacitatea de stocare a datelor e crescut exponențial. O consecință a acestor desfășurări este influxul major de date, datorat achiziției de semnale diverse prin noile metode disponibile. Dacă tehnica de calcul a răspuns pe măsură la imperativul de stocare și organizare a acestei afluențe de informații, metode noi și eficiente de analiză, interpretare și integrare a acestor date sunt necesare pentru a valorifica noile oportunități ale realității actuale.

Metodele inteligenței artificiale (IA) au fost incremental solicitate pentru a analiza și înțelege procese modelate în cele mai diferite domenii de studiu. O disciplină cu impact major în medicină, care a înflorit în mod spectaculos în perioada recentă și cu suportul IA, este bioinformatica. Mai mult, imperativele noi din bioinformatică au apelat adesea la metodele specifice IA și au stimulat evoluția lor în consecință. Analiza genelor diferențial exprimate constituie una dintre direcțiile fundamentale în bioinformatică, iar în acest domeniu, utilitatea metodelor IA s-a atestat prin rezultate spectaculoase, cu impact în activitatea medicală clinică.

Algoritmii evoluționiști (AE) sunt o parte importantă a disciplinei inteligenței artificiale și au fost adesea utilizați pentru a interpreta date achiziționate pentru a descrie modele din cele mai diverse. Algoritmii evoluționiști sunt clădiți pe principii testate timp de miliarde de ani, în care procese evolutive au răspuns cu succes și în mod divers, la schimbările și provocările mediului înconjurător.

Tehnologia ADN microarray este o metodă larg utilizată și bine fundamentată în bioinformatică. Accesul facil la seturi de date reale și rezultatele corespunzătoare, obținute prin metode diverse, de numeroși cercetători, oferă o oportunitate majoră de a dezvolta metodele IA într-un cadru consolidat și extensiv explorat. Posibilitatea de-a evalua performanța unor metode noi de AE în analiza unor date reale, șansa de-a compara comportamentul metodelor propuse cu abordări intens testate și potențialul de-a îmbunătăți metodele de analiză a datelor ADN microarray concomitent, este foarte atractivă.

Această teză de doctorat introduce principii modelate din evoluția biologică naturală cu scopul de-a îmbunătăți performanța și aplicabilitatea algoritmilor genetici (AG).

1.2. Obiective

În realizarea tezei de doctorat de față, am urmărit modelarea și implementarea unor principii care fundamentează evoluția naturală în biologie, cu scopul îmbunătățirii performanței și aplicabilității algoritmilor genetici. Punctual, ne-am propus:

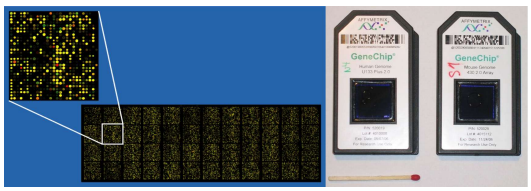
1) **Modelarea dominanței incomplete**,
2) Estimarea oportunității de-a reprezenta genotipul printr-un **număr variabil de cromozomi**,
3) Modelarea principiului **atribuirii aleatorii a cromozomilor din timpul meiozei** și introducerea unui nou operator de recombinare corespunzător,
4) Testarea noilor modele și operatori în contextul selectării atributelor în analiza datelor obținute cu tehnologia ADN microarray,
5) Realizarea unui **pachet software** integrabil în R și Bioconductor, accesibil pentru testare și utilizare de către cercetătorii în domeniul analizei datelor ADN microarray.
6) Modelarea **mutației fără sens**,
7) Modelarea **mutației cu deplasare**,
8) Modelarea **mutației cu ștergerea unui segment de cromozom**,
9) Modelarea **mutației ștergerea unui întreg cromozom**,
10) Modelarea **transpozonilor**.

2. Cadru teoretic

2.1. Tehnologia DNA microroarray

Tehnologia ADN microaray reprezintă un progres major în analiza genetică. Metodologia a fost utilizată în numeroase studii din genetică și biologie moleculară, iar impactul a depășit granițele cercetării fundamentale. Analiza genelor diferențial exprimate cu ajutorul ADN microarray, probabil, aplicația cel mai extensiv tratată în literatura de specialitate, a cunoscut o evoluție și o influență remarcabile în analiza genetică, de când tehnologia a fost introdusă. Teste diagnostice dezvoltate pe fundamentul tehnologiei ADN microarray sunt utilizate în activitatea clinică modernă.

ADN microarray oferă o imagine a condițiilor care circumscriu o instanță la un moment particular și oportunitatea descrierii proceselor biologice complexe plecând de la impactul asupra fenotipului al expresiei genetice. Tehnologia permite evaluarea expresiei genetice a mii de gene imobilizate pe un singur chip (Fig. 2.1). Aceasta abordare aduce informații valoroase pentru:
1) descoperirea unor **marcheri** cu valoare în diagnosticarea unor boli,
2) dezvoltarea de **medicamente** cu eficiență sporită pentru o anumită patologie,
3) descrierea diferitelor **stadii** într-o anumită patologie,
4) schimbările produse de contactul cu anumiți **patogeni**.



a) b)

Fig 2.1 - Exemplu de chip ADN microarray. a) Tehnologie Stanford; b) Tehnologie Affimetrix.
(sursă – Academic Dictionaries and Encyclopedias, www.enacademic.com, domeniu public.)

Experimentele cu ADN microarray sunt complexe, necesită o execuție atentă (Fig. 2.2). În general, câteva etape sunt obligatorii în realizarea cu succes a unui experiment de acest tip:

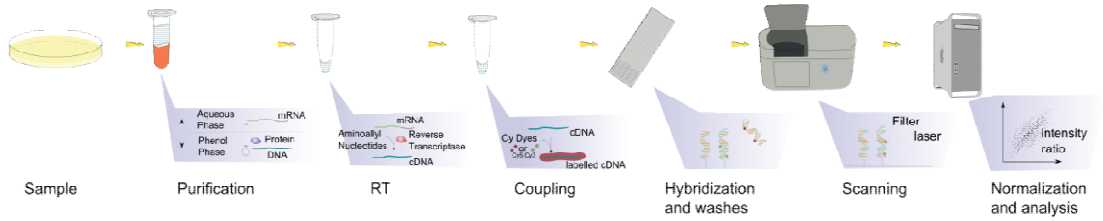


Fig. 2.2 – Etapele unui experiment ADN microarray
(sursa imaginii – wikipedia.com, domeniu public.)

Studiile ADN microarray au ca finalitate descoperirea unui grup restrâns de gene care sunt legate cauzal de evoluţia unei anumite patologii. Statistica este cea care oferă metodele utilizate într-o analiză standard cu date de ADN microarray. Totuși, concluziile și semnificația biologică a rezultatelor se realizează de către biologi și geneticieni supra-specializați în această direcție, depinzând semnificativ de experiența și cunoștințele fiecărui cercetător.

Pentru atingerea obiectivului studiilor de microarray, determinarea unui grup restrâns de gene care pot fi legate cauzal de o anumită patologie, sunt de importanţă capitală metodele de selecţie a atributelor (oligonucleotidelor) semnificative. Aceste metode reprezintă direcţia de dezvoltare a lucrării de faţă şi se concretizează în propunerea metodei noi, evoluate, de selectare a atributelor semnificative.

3. Metodă propusă pentru selectarea unui număr restrâns de atribute, interpretabile din punct de vedere biologic

Algoritmii evoluționiști (AE) utilizează principii din evoluția naturală pentru a oferi răspunsuri la probleme de optimizare. Principiile evoluției naturale sunt testate pe parcursul a miliarde de ani de continuă adaptare la mediul înconjurător și optimizare. Algoritmii genetici (AG) fac parte din domeniul AE și au fost aplicați cu succes în diferite probleme de optimizare. De asemenea, algoritmii genetici au devenit o metodă stocastică de elecție pentru a selecta atribute în diferite domenii.

Obiectivul nostru este selectarea unui număr restrâns de atribute, interpretabile biologic, din date achiziționate cu tehnologia ADN microarray. Propunem o metodă îmbunătățită, fundamentată pe AG și concepută pentru selectarea atributelor în general, dar optimizată pentru aplicațiile cu date microarray. Modelăm fenomene care fundamentează evoluția naturală cu scopul ameliorării performanței AG în acest context.

Metoda descrisă în continuare, gravitează în jurul unui algoritm genetic diploid, dar beneficiază de dezvoltări originale modelate din evoluția naturală:

1) abordare inspirată de **dominanța incompletă** pentru maparea genotipului la fenotip,
2) un operator fasonat după fenomenul **atribuirii aleatorii a cromozomilor** în timpul meiozei,
3) alternative de **operatori pentru mutație**, inspirați din genetica umană, concepuți pentru particularitățile studiilor microarray.

3.1. Algoritmii genetici

Algoritmii genetici au fost teoretizați de către Holland în urmă cu cinci zeci de ani. Tot Holland, a introdus noțiunea de schemă și teorema schemelor [1] pentru a formaliza procesul de evoluție în AG. Abordarea AG simplă a fost extinsă în moduri variate cu reprezentarea diploidă a cromozomilor [2] și diferiți noi operatori, o parte importantă modelați pe observații din evoluția naturală [3]. Finalitatea AG este evoluția înspre soluția care optimizează un criteriu prestabilit.

Nomenclatura AG este împrumutată din genetică pentru a sublinia sorgintea principiilor în evoluția naturală. Populația este formată din indivizi, reprezentând soluții. Indivizii sunt codificați de un genotip care se exprimă prin fenotip. Genotipul codifică atribute în forma unui șir de gene. În reprezentarea binară clasică, genele codifică atribute, respectă poziții fixe în genotip, numite loci și au alele cu valorile 0 și 1. Testarea adaptabilității unui fenotip la mediul înconjurător, problema propusă, se realizează prin evaluarea unei funcții fitness. Rezultatul acestei evaluări reprezintă adaptabilitatea sau fitness-ul individului. Adaptarea exploatează structura mediului înconjurător, depinde atât de mediul înconjurător cât și de funcția fitness aleasă.

Un context în care algoritmii genetici sunt de preferat altor metode de căutare [4] este conturat de câteva criterii:

1) spațiul soluțiilor este vast și nu se cunosc informații despre configurația lui,
2) funcția fitness este afectată de zgomot,
3) scopul căutării este satisfăcut și prin găsirea unui optim local.

Aceste condiții descriu excelent condițiile din activitatea de selectare a atributelor în experimentele ADN microarray.

Un AG persevereză în două activități: explorare și exploatare [4]. Algoritmul explorează soluții noi pentru a se adapta mai bine la mediul înconjurător. Concomitent, exploatează adaptările deja dobândite pe parcursul căutării. Pentru ca o căutare cu AG să fie eficientă, trebuie să existe un echilibru între aceste două activități. Dacă exploatarea este favorizată excesiv, există riscul de overfitting. Dacă explorarea este favorizată evoluția poate fi împiedicată.

Populația inițială de indivizi este, în general, generată fortuit dintr-o distribuție uniformă discretă. Datorită acestui aspect, replicații ale unei căutări cu un algoritm genetic converg adesea înspre soluții diferite. Acest aspect a fost adresat prin replicări multiple ale căutării sau alternative deterministice la generarea aleatorie a populației inițiale.

Recombinarea modelează un principiu din biologia celulară care stă la baza diversității genetice naturale. În timpul meiozei, are loc un schimb de informație genetică, crossing-over, între cromozomi omologi. În timp, au fost propuși diferiți operatori pentru a susține evoluția în algoritmii genetici. Operatorii clasici, care s-au bucurat de succes major, sunt prezentați în continuare.

1) Recombinarea într-un punct

- un locus este ales aleatoriu,
- Moștenitorul va fi codificat de o parte din genotip:
  - identică unui părinte, până la locus-ul respectiv, combinată cu
  - un segment de genotip de la alt părinte, după același locus.
- limitată în termeni de combinații care pot fi generate
- foarte afectată de pozitional bias, locus-ul genei în genotip:
  - genele cu poziții foarte apropiate sunt moștenite preferențial
  - fragmentele recombinate conțin întotdeauna capete ale lanțului
  - ambele capete ale unui genotip nu vor fi conservate în moștenitor.

În selectărea atributelor din datele microarray, valoare limitată:

- genotipuri foarte lungi, cu puține gene activate în fiecare set haploid de cromozomi
- schemele foarte lungi sunt distruse

2) Recombinarea în două puncte

- două locus-uri sunt alese fortuit:
  - segmentul dintre ele va fi obiectul schimbului de gene
- ameliorează unele dintre dificultățile cu care se confruntă versiunea anterioară:
  - nu mai conferă statut special capetelor lanțului,
  - poate genera scheme inaccesibile recombinării într-un punct.

În selectărea atributelor din datele microarray:

- schemele lungi sunt mai bine conservate
- capetele cromozomilor nu mai sunt favorizate,
- riscul îl reprezintă șansa ridicată ca o recombinare să nu producă efecte în cromozom

Operatorii pentru **mutație** în algoritmii genetici de asemenea, modelează un proces din evoluția naturală. Pentru ca o schimbare în genotip să poată fi considerată mutație, ea trebuie să fie transmisibilă moștenitorilor. Mutația poate transfera generației viitoare caracteristici dezirabile, care suportă adaptarea la mediu, sau proprietăți indezirabile, care dimpotrivă, alterează fitness-ul purtătorului. Datorită acestor incertitudini cu privire la impactul asupra evoluției, probabilitatea ca o mutație să apară este, în general, mult mai mică decât în cazul recombinărilor în algoritmii genetici. Operatorul pentru mutație cel mai frecvent implementat în algoritmii genetice este mutația punctuală. În această abordare, un locus este ales în mod aleatoriu. La acel locus, alela prezentă este înlocuită cu alternativa.

Principiul evoluției prin **selecție** își are, de asemenea, sorgintea într-o lege naturală [5]. Indivizii care se adaptează mai bine la mediul înconjurător, au șanse sporite de supraviețuire și în consecință, au șanse sporite de-a își transmite informația genetică. Astfel, genele lor vor fi mai bine reprezentate în generațiile ulterioare.

3.2. Metodă propusă pentru selectarea unui număr restrâns de atribute

Finalitatea algoritmului genetic implementat o reprezintă selecția atributelor din datele ADN microarray. Scopul unui astfel de experiment nu este găsirea unui clasificator supervizat care poate discrimina perfect între două clase de exemple. Ne propunem să determinăm, dintr-un mare număr de gene diferențial exprimate, un sub-grup care poate caracteriza și determina cauzal cele două clase de exemple. Relația cauzală între un subgrup de gene determinat prin metodele inteligenței artificiale nu poate fi stabilită în mod direct. Validarea biologică ulterioară a rezultatelor obținute prin

metodele IA este obligatorie și depășește scopul acestei lucrări de doctorat. Studiile de microarray sunt realizate în echipe multidisciplinare tocmai datorită acestor exigențe.

Arhitectura AG propus este fundamentată pe reprezentarea atributelor în genotip. Fiecare probă din setul de date micoarray este reprezentată de o genă în genotip. Prin urmare, numărul atributelor din setul de date este egal cu numărul genelor din genotip. Fiecare locus poate fi ocupat de o alelă *1* sau *0*. *Alelele 1* codifică pentru prezența atributului de la acel locus în clasificarea supervizată. *Alelele 0* semnifică ignorarea acelui atribut la discriminarea dintre cele două clase de exemple. Un genotip are aspectul unui șir de valori *0* și *1*, iar atributele din setul de date corespunzătoare fiecărui locus codificat *1* în genotip, sunt utilizate în clasificare.

Utilizăm clasificatori supervizați pentru a evalua adaptabilitatea la mediu (sau fitness-ul) a genotipurilor testate. Acuratețea acestor clasificatori în discriminarea între clasele prezente în date a fost utilizată pentru evaluarea numerică a adaptabilității.

Prin urmare, un genotip, șir de valori *0* și *1* cu lungimea egală cu numărul atributelor din setul de date codifică pentru un clasificator supervizat, a priori stabilit, angajat în a învăța exemplele pe seama sub-grupului de atribute specificat prin alelele cu valoarea=*1*.

Algoritmul propus este fundamentat pe un AG **diploid**. Fiecare individ dispune de două seturi haploide de cromozomi, în consecință, de doi clasificatori care utilizează aceeași tehnică de învățare, dar considerând sub-grupuri diferite de atribute pentru discriminarea între exemple.

Un aspect extrem de important în proiectarea unui algoritm genetic **diploid** îl reprezintă **maparea genotipului la fenotip**, semnificativ diferită față de implementările haploide. Prezența a două alele corespunzător fiecărui locus, în fiecare dintre cele două seturi de cromozomi, necesită o abordare specială. În general, această problemă a fost adresată prin definirea unor scheme de dominare, individualizate pentru un cadru specific de optimizare.

În această teză de doctorat propunem o abordare originală, inspirată din evoluția naturală, pentru maparea genotipului la fenotip în algoritmii diploizi. Modelată după principiul **dominanței incomplete** în genetică, propunerea noastră nu necesită definirea unei scheme de dominare și avantajează explorarea în AG.

Următoarea etapă în AG propus o reprezintă **condensarea genotipurilor în cromozomi**. Utilizatorul poate specifica la inițializarea căutării, numărul de cromozomi din genotip. Distribuirea genelor după numărul de cromozomi solicitat nu se realizează echilibrat. Am ales să utilizăm repartizarea genelor umane pe cei *22* autozomi, prezentată în tabelul 3.1, ca model în acest scop. Evoluția tehnologiei ADN microarray înspre variante adaptabile de către utilizator în termeni de probe imobilizate pe chip, permite o abordare superioară în selectarea atributelor. În general, într-un studiu ADN microarray, numărul mare al genelor fixate pe un chip comercial, conceput pentru o gamă largă de aplicații, face imposibilă utilizarea ordinii genelor pe chip în decizia configurației cromozomilor din algoritmul genetic. Noile variante de biochip-uri adaptabile, deschid această oportunitate. Gruparea unor gene cu roluri similare, cunoscute, în diferite căi deja descrise pe aceiași cromozomi ar reprezenta o abordare foarte dezirabilă, cu efecte potențial remarcabile. Din păcate, nu am avut acces la această tehnologie pentru a studia implicațiile unui astfel de cadru. Testarea algoritmului propus în teza de doctorat în acest context reprezintă o direcție de cercetare pentru viitor.

În etapa următoare, populația inițială este evaluată prin prisma performanței clasificatorilor supervizați aleși, în discriminarea claselor de exemple din setul de date. Media celor două valori pentru acuratețe, corespunzător subgrupurilor considerate în fiecare dintre cele două seturi haploide de cromozomi caracterizează adaptabilitatea la mediu a individului respectiv. Indivizii astfel stratificați, după performanță, sunt ulterior supuși operațiunilor de recombinare, aplicate între cele două seturi haploide de cromozomi ai fiecăruia.

Recombinările sunt realizate cu un operator original, care modelează fenomene din meioză ce stau la baza evoluției în natură. Operatorul de recombinare, implementează **atribuirea aleatorie a cromozomilor**, a priori afectați de recombinări în două puncte, genotipurilor generate în această etapă.

Tabel 3.1 - Repartizarea atributelor pe cromozomi

Cromozom nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
% dintre atribute	9.17 %	7.64 %	5.81 %	4.89 %	5.19 %	5.81 %	5.50 %	4.28 %	4.28 %	4.28 %	6.11 %	4.89 %	2.44 %	3.66 %	3.66 %	3.97 %	4.89 %	1.83 %	5.19 %	2.75 %	1.22 %	2.44 %

Metoda de selecție a seturilor de cromozomi din generației următoare exploatează elitismul, adesea implementat în algoritmii genetici. Pe de o parte seturile de cromozomi care au codificat clasificatorul mai puțin performant în fiecare individ sunt eliminate. O elită din seturile haploide de cromozomi care au fost mai adaptate în fiecare individ la evaluarea generației curente, este păstrată pentru iterația consecutivă. Proporția genotipurilor elitiste, conservate în generația următoare, este aleasă de utilizator, la inițializarea căutării. Din genotipurile obținute prin recombinări sunt eliminate fortuit un număr de instanțe egal cu valoarea stabilită pentru elitism. Acest pas este implementat pentru a perpetua o populație de dimensiune constantă pe parcursul căutării.

Seturile haploide de cromozomi selectate pentru a face parte din generația viitoare sunt supuse alterării prin mutație, cu o incidență specificată la inițializare. Pe parcursul cercetărilor noastre, am constatat că mutația clasică, este insuficientă pentru a susține capacitatea AG de-a părăsi un optim local. Incidențe scăzute ale mutației clasice nu avantajează acest comportat foarte dezirabil, iar incidențe sporite afectează semnificativ exploatarea. Așadar, **am explorat posibilitatea altor operatori pentru mutații**. Evoluția naturală a constituit sursa de inspirație pentru alte implementări, discutate separat în subcapitole dedicate în continuare.

Generația următoare este creată consecutiv, prin asamblarea indivizilor din seturile haploide de cromozomi împerecheate fortuit. Această nouă generație este ulterior evaluată și analizată pe parcursul unei noi iterații. Aceste etape se execută repetat pe parcursul unui număr de iterații specificat la inițializarea căutării, condiția de terminare a algoritmului.

Un număr de replicații al selectării atributelor cu algoritmul genetic este necesar pentru a adresa componenta stocastică a căutării și a obține rezultate semnificative. Interpretarea atributelor selectate cel mai frecvent, trebuie realizată cumulativ din rezultatele obținute în fiecare din repetițiile experimentului.

3.3. Dominanța incompletă

3.3.1. Dominanța incompletă în biologie

Fiecare celulă din organismul uman, cu excepția gameților, are la dispoziție, în nucleu, două copii ale fiecărui autozom. Autozomi sunt toți cromozomii, mai puțin X și Y. O copie a fiecărui autozom provine de la mamă, iar cealaltă este moștenită de la tată. Celulele somatice, au la dispoziție două copii ale fiecărui autozom și sunt prin urmare, diploide. Gameții conțin o singură copie a fiecărui autozom și sunt indicați drept haploizi. Cele două copii ale fiecărui cromozom, poartă numele de omologi. Fiecare dintre cromozomii omologi este moștenit de la unul dintre părinți și are, la aceiași loci, gene pentru aceleași tratamente. În consecință, celulele somatice au, în nucleu, două copii pentru fiecare genă, la același locus, în cromozomi omologi. Genele prezente la același locus în cromozomi omologi poartă numele de alele. Alele identice sunt prezente în cromozomi omologi homozigoți pentru respectivul locus. Cromozomii omologi care au alele diferite sunt numiți heterozigoți pentru locus-ul specificat.

În cazul organismelor diploide, se pune problema modului în care alele diferite, prezente în cromozomi omologi heterozigoți, își găsesc exprimarea în fenotip. În 1865, Mendel a descris un model în care una dintre cele două alele se exprimă în fenotip (caracter dominant), iar cealaltă nu (caracter recesiv). Respectând nomenclatura introdusă de Mendel, această relație poartă numele de dominanță.

În tabelul de mai jos, prezentăm un caz fictiv, în care un organism moștenește gene care-i determină culoarea, de la generația anterioară. Există două alele posibile **R** și **a** pentru gena care determină culoarea organismului. **R** este alela *dominantă* și se exprimă în fenotip prin culoarea roșie. În notația din genetică alela dominantă este reprezentată de o literă majusculă, iar cea recesivă este notată cu litere minuscule. Alela **a** determină culoarea albastră în fenotip. Tabelul 3.2 prezintă combinațiile posibile și efectele în fenotip.

O alternativă la acest model, îl reprezintă dominanța incompletă. În acest tip de interacțiune între alele, adesea întâlnită în natură, nici o alelă nu domină asupra celeilalte și expresia în fenotip a nici uneia nu este suprimată. Fenotipul heterozigot va fi intermediar între variantele de homozigote. Principiul dominanței incomplete este ilustrat în Tabelul 3.3, pentru același exemplu de organism fictiv de mai sus. În acest caz, ambele alele au fost notate cu litere majuscule, deoarece nici una nu este dominată. Se observă în tabel că fenotipul heterozigot RA va avea culoarea violet, o combinație a efectelor celor două alele.

Tabel 3.2 – Dominanță completă

		Moștenire de la TATĂ	
		R	a
Moștenire de la MAMĂ	R	RR	Ra
	a	Ra	aa

Tabel 3.3 – Dominanță incompletă

		Moștenire de la TATĂ	
		R	A
Moștenire de la MAMĂ	R	RR	RA
	A	RA	AA

3.3.2. Dominanța incompletă în algorimii genetici

În proiectarea unui AGD **maparea genotipului la fenotip** este complicată, comparativ cu algoritmii genetici haploizi. Un algoritm genetic evaluează adaptabilitatea la mediul înconjurător pe seama fenotipului, iar codificarea acestuia în genotip este mai elaborată în implementările diploide. Dubla prezență a alelelor la fiecare locus, în fiecare dintre cele două seturi de cromozomi prezente într-un individ, necesită tratament suplimentar.

Dominanța incompletă promovează ideea unui **fenotip intermediar** între variantele homozigote. Într-un algoritm genetic diploid, putem aplica acest concept după cum urmează. Fiecare set de cromozomi prezent în individ, poate fi considerat cu efectele sale particulare. Un individ reprezentat prin două seturi de cromozomi, are la dispoziție doi clasificatori supervizați, diferiți în privința atributelor considerate, pentru a se adapta în același context. Adaptabilitatea unui astfel de individ este apreciată prin prisma **mediei performanțelor** în acomodarea la mediul înconjurător al celor doi clasificatori.

În algoritmul propus în teza de doctorat, pentru analiza datelor de ADN microarray, performanța adaptării la mediul înconjurător este evaluată pe seama acurateței în discriminarea între două clase de exemple cu ajutorului unui tip de clasificator supervizat, la alegere. Fiecare clasificator utilizează un sub-grup dintre atributele prezente în setul de date. Prin urmare, un individ va fi evaluat prin prisma valorii medii a acuratețelor celor doi clasificatori supervizați codificați în cele două seturi intrinseci de cromozomi.

3.4. Atribuirea aleatorie a cromozomilor

3.4.1. Atribuirea aleatorie a cromozomilor în meioză

Procesul de diviziune celulară fundamentează finalitatea supraviețuirii individului și în consecință al speciei. Celulele dintr-un organism trebuie să se dividă pentru atingerea unor obiective diferite. Celule somatice trebuie să se dividă pentru a regenera diferite țesuturi și a susține funcții variate în organism. Celule speciale, numite germinative, se divid pentru a crea gameți, proces care stă la baza reproducerii organismului și, în consecință, a propășirii speciei.

În general, celule somatice diferențiate pentru realizarea optimă a funcții bine stabilite, în țesuturi specifice, se divid cu finalitatea de-a produce celule noi, identice, capabile să susțină aceleași funcții. În acest context, bagajul genetic al celulei care urmează a se divide, trebuie păstrat în integralitatea lui, iar modificările la nivelul ADN nu sunt dezirabile. Acest tip de diviziune celulară poartă numele de mitoză.

Pe de altă parte, celulele germinative se divid într-o manieră fundamental diferită. Acest proces, numit meioză, are ca scop reproducerea individului, iar modul de desfășurare servește finalitatea. Rezultatul meiozei, celule numite gameți, conțin în nucleu un singur set de cromozomi. În timp ce mecanisme de control protejează transmiterea informației genetice prin meioză, **un grad de diversitate genetică este permis**. Scopul toleranței pentru un anumit grad de variabilitate este fără îndoială, evoluția generației următoare în sensul adaptării superioare la mediul înconjurător. Meioza se desfășoară în două etape cu caracteristici diferite, numite meioză I și meioză II. Pe parcursul meiozei I, consecutiv replicării ADN-ului, cromozomii omologi, cu structură dublă, formează sinapse. Cromozomii omologi se asamblează în structuri numite tetrade. În această configurație, are loc *schimbul de informație genetică între cromozomii omologi*. Această comunicare, prin intermediul unor fragmente de ADN, este permisă la nivelul unor poziții specifice de contact, numite chaisme. Fenomenele care au loc în timpul diviziunilor celulare sunt discutate detaliat în [6]. Acest proces a fost pe larg utilizat în algoritmii genetici și s-a concretizat în diferiți **operatori de recombinare** (**crossover** în acord cu nomenclatura din genetică). Ilustrare sugestivă a meiozei este prezentată în Fig. 3.1. Fenomenul de recombinare are loc în profaza I și este o sursă importantă de diversitate genetică, modelat în forme variate în diferite implementări de algoritmi genetici.

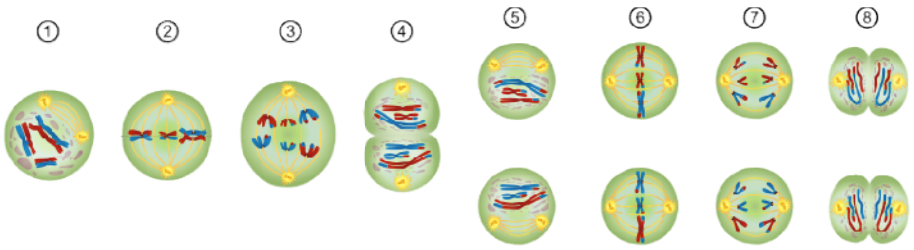


Fig. 3.1 – Etapele meiozei (sursa imaginii: wikipedia.com, cu drepturi libere de utilizare și modificare.)

O altă sursă de variabilitate genetică o reprezintă **atribuirea aleatorie a cromozomilor** în meioză. Alinierea, fortuită în privința sensului, a tetradelor în raport cu cei doi poli are loc în timpul metafazei I și se concretizează consecutiv, în atragerea independentă și întâmplătoare a cromozomilor cu structură dublă spre unul dintre poli, prin acțiunea microtubulelor, pe parcursul anafazei I. Această sursă de diversitate genetică este extrem de importantă, iar cuantumul informației genetice comunicate în timpul acestui proces merită exploatat.

Importanța recombinărilor și atribuirii aleatorii a cromozomilor pentru evoluția naturală este subliniată prin antiteză cu absența lor din mitoză. Meoiza I cu reducerea numărului de cromozomi și procesele care asigură diversitatea genetică au ca finalitate evoluția generației viitoare. De asemenea, mecanismele de control al integrității informației genetice permit evoluția, în contrast cu rigiditatea specifică mitozei.

3.4.2. Atribuirea aleatorie a cromozomilor în AG

Algoritmii genetici utilizează în mod tradițional operatori de recombinare pentru a asigura explorarea și susține evoluția. Cele mai populare propuneri au fost recombinările într-unul sau două puncte (Fig. 3.2 și Fig. 3.3). Al doilea fenomen care asigură diversitatea genetică, **atribuirea aleatorie a cromozomilor** în timpul meiozei *nu a fost exploatat corespunzător în AG*.

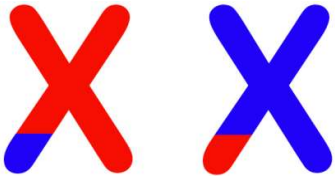


Fig. 3.2 – Recombinarea într-un punct.

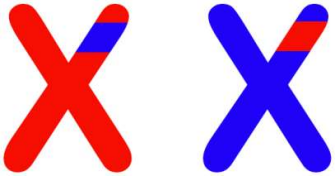


Fig. 3.3 – Recombinarea în două puncte.

Teza de doctorat de față propune un algoritm care beneficiază de modelarea atribuirii aleatorii a cromozomilor în timpul meiozei. În abordarea noastră, genotipul este a priori configurat într-un *număr variabil de cromozomi*. Tratarea genotipului ca seturi de cromozomi este o condiție obligatorie în vederea modelării acestui fenomen. Impactul utilizării modelului atribuirii aleatorii a cromozomilor în AG cu un număr variabil de cromozomi este ilustrat în Fig. 3.4. Tratăm un genotip cu informația genetică distribuită pe trei cromozomi pentru claritate. Odată cu utilizarea unui număr sporit de cromozomi, cresc și efectele operatorului propus de noi. Figura reprezintă și recompensele în comparație cu operatorii clasici de recombinare într-unul sau două puncte. Primul nivel surprinde o celulă diploidă cu materialul genetic organizat pe trei cromozomi, a priori replicați. Cromozomii moșteniți pe linie maternă și paternă sunt ilustrați cu culori diferite, roșu și respectiv albastru. Este ilustrat fenomenul de formare al sinapselor. Efectele recombinărilor într-unul sau două puncte sunt reprezentare pe nivelul al doilea, iar rezultatul obținut prin atribuirea aleatorie a cromozomilor este evident în nivelul inferior al figurii. Ultimul nivel al figurii prezintă impactul asupra materialului genetic pregătit pentru a fi transmis generației viitoare, corespunzător setului haploid de cromozomi rezultat în urma meiozei II din natură. Consecutiv meiozei II, patru astfel de seturi ar trebui figurate, dar ilustrația a fost simplificată pentru claritate. Este evidentă recombinarea superioară din punct de vedere al diversității genetice în abordarea cu atribuirea aleatorie a cromozomilor, figurată la nivelul inferior al figurii.

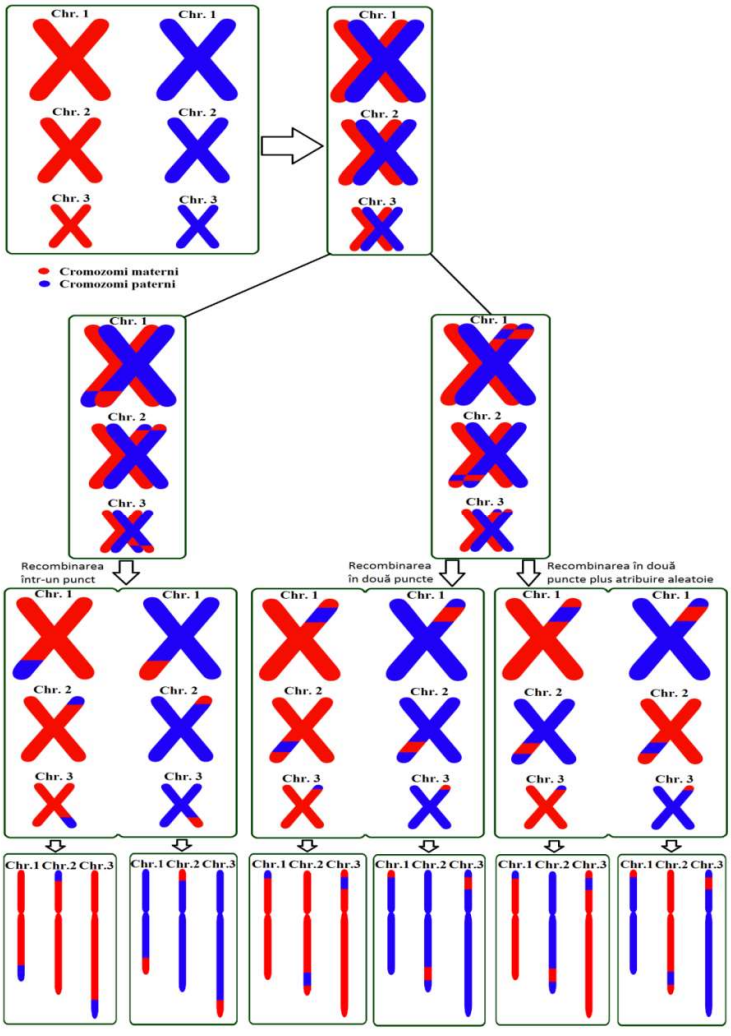


Fig. 3.4 - Diversitatea genetică (a)recombinarea într-un punct; (b)recombinarea în două puncte; (c)recombinarea în două puncte urmată de atribuirea aleatorie a cromozomilor.

3.5. Operatori pentru mutații

Operatorii pentru mutație își au sorgintea în biologie. În selectarea atributelor din datele de ADN microaray, se lucrează cu mii de atribute și se urmărește selectarea unui sub-grup restrâns dintre acestea. Utilitatea mutației într-un punct, în acest context este limitată. Cu toate că **mutația într-un punct** tulbură eficient seturile haploide de cromozomi din populație, fenotipul indivizilor nu este afectat așa cum am dori. Numărul genelor active într-un individ, sporește progresiv în generațiile evoluate. Utilizarea unei șanse prea mari de a apărea o mutație, are efect negativ asupra explorării.

Există deosebiri fundamentale între codul genetic în natură și principiile utilizate în algoritmii genetici. Mecanismele de apariție ale mutațiilor în genetică nu pot fi modelate fidel în operatori pentru algoritmii genetici. Cu toate acestea, principii învățate din genetică pot fi utilizate pentru îmbunătățirea unor astfel de operatori. Foarte multe variante de operatori pentru mutații au fost propuși diacronic de diferiți autori, iar genetica a fost adesea izvorul de inspirație pentru acele metode. Nu considerăm că propunerile descrise în continuare sunt originale în totalitate sau că principiile utilizate nu au fost abordate anterior. Este însă foarte interesantă utilizarea acestor abordări în contextul algoritmului propus de noi pentru selectarea atributelor în datele de ADN microarray.

3.5.1. Mutația fără sens

În genetică, mutația fără sens (eng. nonsense mutation) este un caz special de perturbare a ADN. Tulburarea apare la nivelul unei singure nucleotide în ADN. Modificarea respectivă, prin transripție, devine un codon stop în ARN. Consecutiv, translația este terminată prematur, iar proteina care ar fi codificată nu mai este sintetizată complet.

În algoritmii genetici, mutația fără sens nu poate fi modelată fidel situației din biologie. Fenomenul rezultat în urma acestui tip de mutație poate fi utilizat însă în conceperea unui operator de mutație valoros în AG. Operatorul pentru mutația fără sens anulează toate alelele prezente într-un cromozom, consecutiv unui locus selectat aleator.

3.5.2. Mutația cu deplasare

Codul genetic este descris în biologie de proprietăți care determină modul în care mutațiile se întâmplă în natură. Pe de o parte, codul genetic nu se suprapune, pe de altă parte este continuu. Prin urmare, când o nucleotidă din șir este întâmplător eliminată sau adăugată, tot subșirul consecutiv este decodificat în mod eronat.

Operatorul propus pentru mutația cu deplasare utilizează principiul din genetică, dar nu modelează întocmai fenomenul biologic. Mutația alterează cromozomul în sensul deplasării la stânga a șirului începând cu o poziție generată aleatoriu. Ultima poziție de pe cromozom este completată ulterior cu alela 0. Un singur cromozom dintr-un set haploid este afectat de mutație. Șansa ca o mutație să se producă este specificată la inițializarea algoritmului. Cromozomii afectați și locus-urile interesate sunt alese la întâmplare.

3.5.3. Ștergerea unui segment

În timpul meiozei, au loc recombinări, schimburi de informație genetică între cromozomii omologi organizați în tetrade. Este posibil să apară erori în această etapă. Segmente din cromozomi pot fi șterse complet dintr-un cromozom și adăugate excesiv în omolog. Astfel, ambii omologi sunt anormali, iar fenotipul este afectat consecutiv.

În implementarea noastră pentru mutația cu ștergerea unui segment de cromozom, cu o probabilitate specificată la inițializare, cromozomii care suferă mutația sunt selectați întâmplător. Ulterior, pentru fiecare cromozom astfel ales, sunt generate aleatoriu marginile unui interval și toate alelele din acel interval sunt anulate.

3.5.4. Ștergerea unui cromozom

Erori pot apărea și la separarea și atribuirea cromozomilor recombinați în timpul meiozei. Pot rezulta astfel celule cu un număr incorect de cromozomi, mai mare sau mai mic.

Operatorul pentru mutație prin ștergerea unui cromozom anulează toate alelele de pe cromozomi aleși fortuit. Șansa ca o mutație cu ștergerea întregului cromozom să se producă este stabilită la lansarea algoritmului.

3.5.9. Transpozonii

Transpozonii sunt secvențe care își pot schimba poziția în lanțul ADN. Se consideră că existența transpozonilor este datorată unor fragmente de ADN viral care s-au inserat în ADN-ul uman.

Operatorul pentru mutație inspirat din caracteristicile transpozonilor, selectează aleatoriu, cu o șansă prestabilită, cromozomi care vor suferi mutația. Ulterior, sunt generate fortuit un locus cu alelă 1 și o valoare pentru distanța deplasării. Distanța poate rezulta în valori negative, specificând o migrație la stânga, sau pozitive pentru deplasarea la dreapta, cu numărul de poziții specificat. Funcționarea operatorului de mutație inspirat de transpozoni este ilustrat în Fig. 3.5.

```
> chrConf
[1] 1 1 1 1 2 2 2 2
> individualsOriginal
Id 1005_at 1007_s_at 1008_f_at 1009_at 1020_s_at 1030_s_at 1038_s_at 1039_s_at
1  1       1         1         1       0         0         0         0         1
2  1       1         1         1       0         1         0         0         0
3  2       1         1         1       0         1         0         0         0
4  2       1         0         0       1         1         0         0         1
5  3       0         0         1       1         0         1         0         1
6  3       1         0         0       0         1         0         1         1
7  4       0         1         0       0         0         1         1         1
8  4       0         1         1       1         0         0         0         1
> individuals
Id 1005_at 1007_s_at 1008_f_at 1009_at 1020_s_at 1030_s_at 1038_s_at 1039_s_at
1  1       1         1         1       0         0         0         0         1
2  1       1         1         1       0         1         0         0         0
3  2       1         1         1       0         0         1         0         0
4  2       1         0         0       1         1         0         0         1
5  3       0         0         1       1         1         0         0         1
6  3       1         0         0       0         1         0         1         1
7  4       0         1         0       0         0         1         1         1
8  4       0         1         1       0         0         0         0         1
```

Fig. 3.5 – Efectele transpozonilor.

4. Pachetul R dGAselID

Abordarea propusă pentru selectarea atributelor în datele ADN microarray este implementată în pachetul software **dGAselID**. Deși este conceput pentru a selecta atribute în datele de ADN microarray, algoritmul poate fi aplicat unei game largi de probleme care impun selectarea unui număr variabil de atribute în date cu un număr mare de dimensiuni. Toate metodele propuse și testările efectuate pe parcursul tezei de doctorat sunt desfășurate utilizând dGAselID.

Între multiplele pachete software orientate pe IA şi analiză statistică, un avânt remarcabil l-a cunoscut limbajul de programare **R** [7]. R a fost dezvoltat din limbajul de programare S, de către Robert Gentleman şi Ross Ihaka la Auckland University din Noua Zeelandă. Un efort conjugat, care sprijină cercetătorii în bioinformatică, este **BioConductor** [8], ce oferă implementări foarte bine documentate pentru metode necesare cercetătorilor, în special celor care se concentrează pe analiza genetică. BioConductor a fost demarat în 2001 la Fred Hutchinson Cancer Research Center şi

actualmente este dezvoltat de echipa Bioconductor core team formată din cercetători de la multiple institute şi universităţi din întreaga lume.

Am ales implementarea algoritmului propus de noi în mediul R, integrat cu Bioconductor. Proiectul R înglobează o gamă variată de metode de analiză statistică, iar Bioconductor integrează instrumente foarte valoroase în analiza genetică. În plus ambele proiecte sunt deschise şi beneficiază de contribuțiile numeroșilor cercetători implicați activ în domeniile respective. Un argument semnificativ pentru alegerea noastră a fost accesul liber la implementări, surse, documentație și comunitatea pasionată și sociabilă de cercetători care utilizează și contribuie la cele două proiecte. Bioconductor oferă acces liber la numeroase seturi de date ADN microarray și facilitează posibilitatea de-a dezvolta și compara metode pentru analiza genetică. R și Bioconductor oferă o multitudine de pachete special implementate pentru a facilita fiecare pas în analiza microarray, de la achiziție și preprocesare, până la inferențe și interpretarea semnificației biologice a rezultatelor obținute. Prezentarea și evaluarea diferitelor metode disponibile în acest sens a fost abordată pe larg în jurnale [9]. Metodologia de cercetare cu ADN microarray cu emfază pe uneltele oferite în R și Bioconductor [10] a fost tratată extins în literatura de specialitate.

În timpul rulării algoritmului, informații despre evoluție sunt disponibile utilizatorului. Date cu privire la acuratețea minimă, medie și maximă în populația curentă sunt afișate pe ecran după fiecare evaluare. De asemenea, cercetătorul este informat în legătură cu numărul mutațiilor efectuate la fiecare iterație și etapa curentă în desfășurarea algoritmului.

Algoritmul afișează, de asemenea, o reprezentare grafică a evoluției după fiecare iterație. Evoluțiile acurateței maxime și medii acompaniază o histogramă a genelor cel mai frecvent selectate, în reprezentarea grafică prezentată după evaluarea fiecărei generații. O captură de ecran din timpul desfășurării algoritmului sunt prezentate în Fig. 4.1.

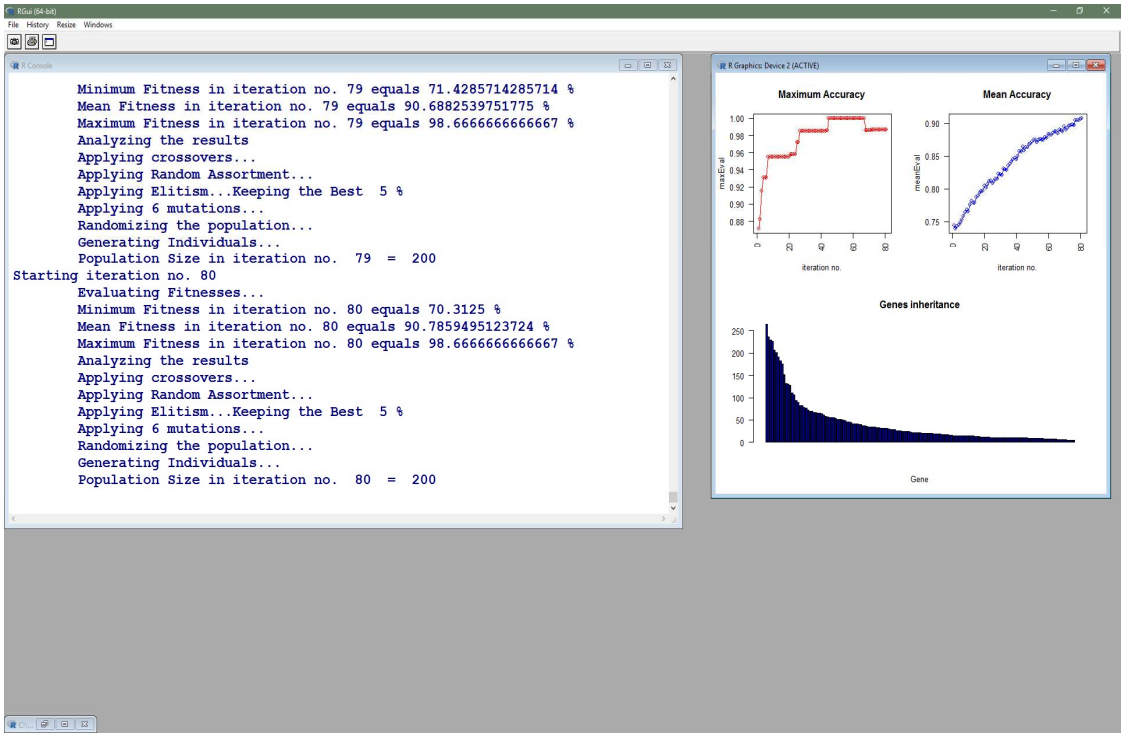


Fig. 4.1 - Captură de ecran după 80 de generații.

Pachetul software **dGAselID** propune o soluție originală și eficientă la problema selectării atributelor în datele ADN microarray. Pachetul este perfect integrabil și beneficiază de toate metodele de analiză și vizualizare din Bioconductor și deopotrivă, completează metodele de analiză a datelor implementate în acel mediu cu o alternativă pentru selectarea atributelor în acest tip de experiment. Metoda propusă este flexibilă și aplicabilă unei game largi de probleme care impun selectarea atributelor în date vaste. *Operatorul original pentru atribuirea aleatorie a cromozomilor* și modelul *dominanței incomplete pentru maparea genotipului la fenotip* favorizează explorarea și

produc rezultate superioare metodelor implementate în MLInterfaces pentru selectarea atributelor. Diferitele opțiuni de operatori pentru mutații oferă flexibilitate în tratarea a diverse seturi de date sau aplicații de selectare a atributelor. Elasticitatea în utilizarea clasificatorilor supervizați și a multiplelor variante de validare încrucișată sunt foarte dezirabile în contextul selectării atributelor în date microarray sau alte contexte. Pentru rezultate foarte consistente, atât cunoștințele a priori despre datele analizate, cât și testarea unora dintre opțiunile disponibile în pachet pot determina cadrul optim pentru algoritmul diploid propus.

5. Experimente

Popularitatea de care tehnologia microarray s-a bucurat în deceniul recent, permite accesul la date și rezultate reale și fundamentează o relație mutual avantajoasă între IA și bioinformatică. IA oferă metode îmbunătățite pentru abordarea problematicilor din bioinformatică. Pe de altă parte, bogăția datelor și rezultatelor disponibile în bioinformatică asigură un cadru propice progresului metodelor IA, aplicabile ulterior în domenii variate.

Metodele propuse de noi care vor fi testate și evaluate în continuare sunt:

1) **dominanța incompletă** pentru maparea genotipului la fenotip,
2) **operatorul atribuirii aleatorii a cromozomilor** în contextul unui număr variabil de cromozomi cu dimensiuni diferite,
3) operatorii pentru mutații.

5.1. Setul de date Acute Lymphoblastic Leukemia (ALL)

Experimentele din acest capitol analizează setul de date Acute Lymphoblastic Leukemia (ALL) [11]. Am ales să utilizăm acest set de date ADN microarray în experimentul nostru deoarece rezidă în date reale, care au fost în prealabil prelucrate. Setul ALL constă în 128 de exemple, pacienți suferinzi de leucemie și pentru fiecare dintre ele, 12625 de atribute reprezentând probe de pe chipuri Human Genome U95 Set produse de compania Affymetrix.

Ne propunem să descoperim un sub-grup de gene care poate fi utilizat pentru a discrimina optim între exemple ale unor categorii bine cunoscute și descrise. Datele de fenotip despre biologia moleculară BCR/ABL a pacienților incluși în studiul ALL oferă această oportunitate. Ne așteptăm ca o metodă validă de selectare a atributelor să descopere genele cauzal responsabile de clasificare pe chip-urile considerate. După eliminarea din setul de date a pacienților cu leucemie T și a celorlalte clasificări de biologie moleculară, 79 exemple sunt păstrate pentru analiză suplimentară. Dintre acestea, 42 reprezintă exemple negative, iar 37 pozitive de BCR/ABL, într-o proporție dezirabilă pentru analiza de recunoașterea a formelor. Setul de date ALL, este analizat cu finalitatea selectării unui subgrup restrâns de gene diferențial exprimate, între pacienți cu clasificare BCR/ABL pozitivă sau negativă, care pot discrimina între cele două categorii.

Investigăm 3 seturi de date pe parcursul testărilor:

1) setul de date complet (79 exemple și 12625 de atribute)
2) subset cu 79 exemple și 2391 de atribute obținut prin filtrarea nespecifică a setului complet după condițiile IQR(x) > 0.5 și cel puțin 25% dintre valori > log2(100).
3) subset cu 79 exemple și 628 de atribute obținut prin filtrarea specifică după *testul t*, cu valoarea de tăiere p=0.1

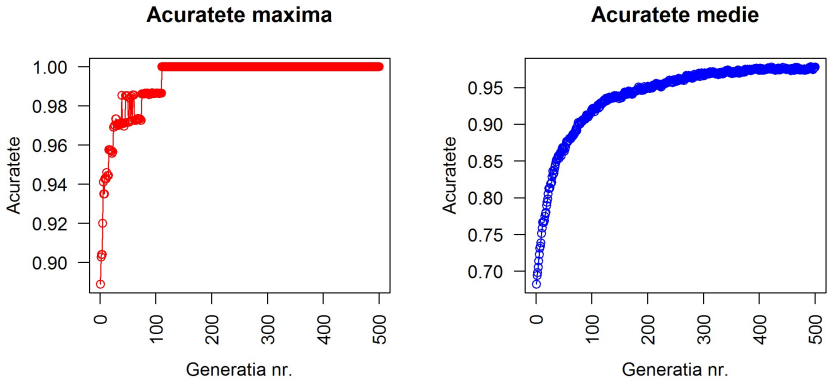
5.2. Evaluarea dominanței incomplete

Clasificatorul supervizat utilizat pentru evaluarea adaptabilității indivizilor este *kNN (k=8)*. Algoritmii genetici au fost inițializați în mod diferit pe fiecare dintre cele trei seturi de date considerate. Particularitățile la inițializare pentru fiecare dintre cele trei seturi de date analizate sunt ilustrate în tabelul 5.1. Condiția de terminare a fost aceeași în toate situațiile, 500 de generații de evoluție. De asemenea, valori uniforme peste toate seturile de date pentru elitism și rata mutațiilor punctuale au

fost stabilite empiric la valorile 5% și respectiv 0.005. Am realizat 20 de replicații în acest cadrul experimental cu fiecare set de date analizat și am interpretat rezultatele obținute.

Tabel 5.1 - Inițializare AG pe diferitele seturi de date.

ID Set	Lungimea Genomului (nr. atribute)	Gene active	Indivizi în populație	Clasificatori kNN
1	12625	30	500	1000
2	2391	20	200	400
3	628	12	100	200



a) b)

Fig. 5.1 – Evoluția AG a) Evoluția acurateței maxime în 500 generații; b) Evoluția acurateței medii în 500 generații.

Tabel 5.2 - Cel mai frecvent selecționate cinci probe din fiecare set

Nr.	Setul de date cu **12625** de probe		Setul de date cu **2391** de probe		Setul de date cu **628** de probe	
	Affymetrix ID	Gene ID	Affymetrix ID	Gene ID	Affymetrix ID	Gene ID
1	$`39730_at`	"ABL proto-oncogene 1, non-receptor tyrosine kinase"	$`39730_at`	"ABL proto-oncogene 1, non-receptor tyrosine kinase"	$`39730_at`	"ABL proto-oncogene 1, non-receptor tyrosine kinase"
2	$`1635_at`	"ABL proto-oncogene 1, non-receptor tyrosine kinase"	$`37027_at`	"AHNAK nucleoprotein"	$`38052_at`	"coagulation factor XIII A chain"
3	$`1636_g_at`	"ABL proto-oncogene 1, non-receptor tyrosine kinase"	$`38052_at`	"coagulation factor XIII A chain"	$`39338_at`	"S100 calcium binding protein A10"
4	$`38052_at`	"coagulation factor XIII A chain"	$`33440_at`	"zinc finger E-box binding homeobox 1"	$`40480_s_at`	"FYN proto-oncogene, Src family tyrosine kinase"
5	$`38968_at`	"SH3-domain binding protein 5"	$`1635_at `	"ABL proto-oncogene 1, non-receptor tyrosine kinase"	$`1636_g_at`	"ABL proto-oncogene 1, non-receptor tyrosine kinase"

Între atributele cel mai frecvent selectate de algoritmul genetic drept semnificative în discriminarea dintre clasele BCR/ABL pozitiv și negativ apar adesea probe reprezentând "*ABL proto-oncogene 1, non-receptor tyrosine kinase*". Acest rezultat sugerează o legătură puternică între cele două clase studiate și nivelul expresiei pentru această genă. Chip-urile comerciale utilizează copii multiple ale aceleiași gene tocmai pentru a oferi o măsură a calității măsurătorilor și rezultatelor obținute. Selectarea a mai multe copii ale aceleiași probe subliniază eficiența AG în selectarea atributelor din datele de ADN microarray. Prezența probei `38968_at`, reprezentând "*SH3-domain binding protein 5*", poate fi de asemenea, semnificativă. Alterarea domeniului SH3 al genei "ABL proto-oncogene 1, non-receptor tyrosine kinase" este asociată cu efecte oncogene, dar interpretarea acestui rezultat în situația dată, depășește scopul și posibilitățile studiului de față. Proba `38052_at`, reprezentând *"coagulation factor XIII A chain"* a fost deopotrivă selectată de algoritmul genetic propus. Această genă a fost implicată în etiologia unor tipuri de leucemie acută, dar nu a fost asociată cu clasificarea biologică BCR/ABL, iar apariția ei în rezultatele noastre necesită interpretare biologică.

5.3. Evaluarea dominanței incomplete versiunea 2

Operatorul pentru elitism oferă o șansă majoră de-a adresa tendința AG de-a converge într-un optim local. În aplicarea elitismului, putem considera adaptabilitatea unui *genotip* sau a *individului* pentru ordonarea după performanță. Selectarea celor mai performante genotipuri pentru a fi păstrate în generația viitoare avantajează exploatarea. Perpetuarea în iterația următoare a genotipurilor care au făcut parte din cei mai adaptați indivizi, poate susține explorarea cu AG.

Ne propunem să testăm impactul asupra evoluției obținut prin abordarea elitismului la nivelul individului, în comparație cu evaluarea genotipurilor. Algoritmul genetic diploid conceput cu dominanță incompletă evoluează cu sau fără un operator pentru elitism, datorită eliminării implicite a genotipului mai puțin adaptat din fiecare individ după evaluare și al recombinărilor succesive. Așadar, o selecție este intrinsecă în modelul propus. Am decis totuși să numim algoritmul genetic care beneficiază de elitismul la nivelul individului AG cu dominanță incompletă versiunea 2 (DI2), pentru a sublinia această flexibilitate. În realitate, este același algoritm cu dominanță incompletă și abordare diferită a operatorului pentru elitism. Pentru claritate ne vom referi la implementarea inițială a dominanței incomplete în algoritmul genetic diploid propus cu denumirea de versiunea 1 (DI1).

Am testat evoluțiile pe subsetul de date ALL cu 79 exemple, filtrat specific și nespecific la 628 de atribute. Șansa ca o mutație să apară a fost anulată, pentru a elimina aceste efecte asupra explorării și a obține o evaluare mai corectă a impactului obținut prin utilizarea celor două tipuri de operatori pentru elitism. De asemenea, am utilizat valoarea de 2% pentru elitism față de 5% considerată anterior. Rezultatele obținute cu dominanță incompletă versiunea 2 și implementarea inițială, cu populații generate din același random seed, după 400 de generații, sunt prezentate în Fig. 5.2.

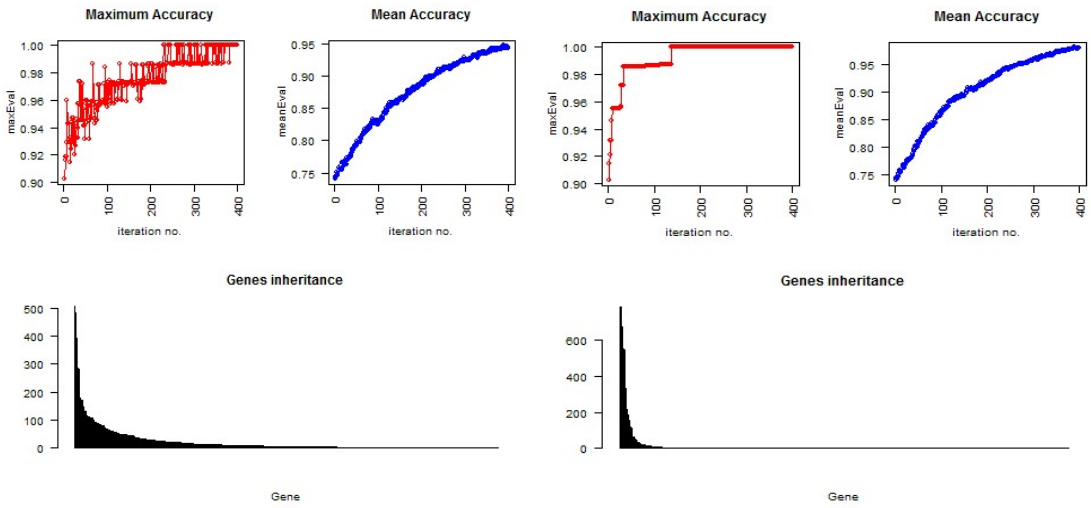


Fig. 5.2 – Rezultatele obținute cu dominanță incompletă versiunea 2 versus versiunea 1.

Pornind de la aceeași populație inițială versiunea 2 se dovedește superioară predecesoarei în privința consistenței rezultatelor obținute. O imagine sugestivă a diferenței în evoluția celor două implementări poate fi obținută prin vizualizarea evoluțiilor acurateței medii și maxime (Fig. 5.3) în populațiile generate din random seed-uri identice. Se remarcă ușurința cu care AG părăsește un optim local în implementarea DI2, cu prețul evoluției mai lente a acurateței medii în populație față de DI1.

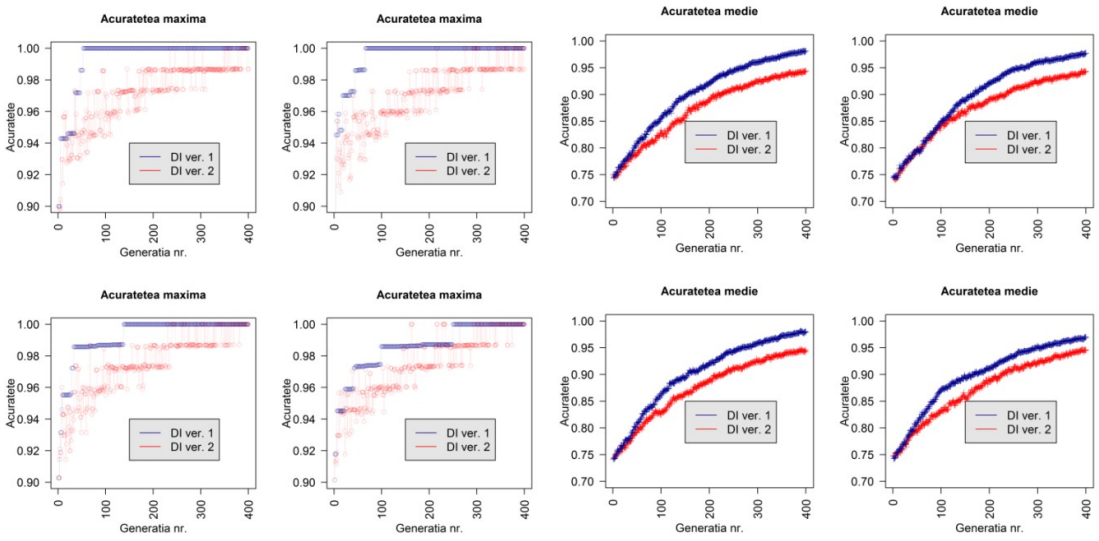


Fig. 5.3 – Evoluția comparativă acurateței maxime și medii pe patru populații inițiale cu DI1 și DI2.

Interpretarea diferențelor dintre cele două abordări trebuie abordată coroborând observațiile cu privire la evoluția algoritmilor și rezultatele obținute. Deși este evident avantajul în privința evoluției acurateței medii în implementarea DI1 față de DI2, rezultatele mai consistente și uniforme obținute prin replicarea experimentelor pe diferitele populații inițiale înclină balanța în avantajul DI2. În mod evident, implementarea DI2 avantajează explorarea pe seama exploatării, dar în contextul dat, al selectării atributelor în datele de ADN microarray, oferă un raport profitabil între cele două laturi ale căutării.

5.4. Evaluarea operatorului pentru atribuirea aleatorie a cromozomilor

Pentru a obține o imagine cuprinzătoare a impactului operatorului pentru atribuirea aleatorie a cromozomilor am decis să testăm diferite configurații de seturi de cromozomi pe subsetul ALL cu 79 de exemple și 628 de atribute. Ne-am îndreptat atenția asupra a 3 scenarii cu seturi de cromozomi, obținute prin distribuția genelor pe unul, cinci sau douăzeci și doi de cromozomi.

Algoritmul genetic utilizat pentru testare beneficiază de dominanța incompletă versiunea 1, cu elitism la nivelul seturilor de cromozomi. AAC este conceput pentru a susține evoluția prin sporirea explorării, așadar DI1 oferă un cadru propice pentru a evalua acest impact independent de efectele ID2. Pentru a stimula exploatarea și a evidenția mai clar efectele AAC asupra explorării, valoarea aleasă pentru elitismul de tip ID1 a fost 5% în experimentele noastre.

Algoritmii genetici testați au beneficiat de aceeași metodă de evaluare a adaptabilității indivizilor, `knn.cvI(k=8)`, în context de validare încrucișată leave-one-out și aceeași șansă ca o mutație punctuală să se producă, 0.005. Rezultatele au fost evaluate după ce fiecare algoritm a evoluat pe parcursul a 500 de generații. Cele trei scenarii considerate, cu unul, cinci și douăzeci și doi de cromozomi, au fost evaluate după 20 de replicări, pornind de la 20 de random seed-uri prestabilite și prin urmare 20 de populații inițiale diferite. Aceleași random seed-uri și respectiv populații inițiale au fost testate în fiecare dintre cele trei scenarii. Populațiile inițiale au constat din 200 de indivizi generați fortuit, cu genotipuri conținând 628 de gene și 12 gene activate în fiecare set de cromozomi.

Algoritmul cu AAC și implicit un număr de cromozomi diferit de 1 a excelat în privința evoluției acurateței maxime (Fig. 5.4). În 19/20 replicări, AG cu AAC a atins soluții cu acuratețe maximă de 100%, ceea ce s-a întâmplat în 50% dintre replicările cu gentotipul reprezentat pe un singur cromozom.

Relația explorare – exploatare este afectată pozitiv de implementarea AAC.

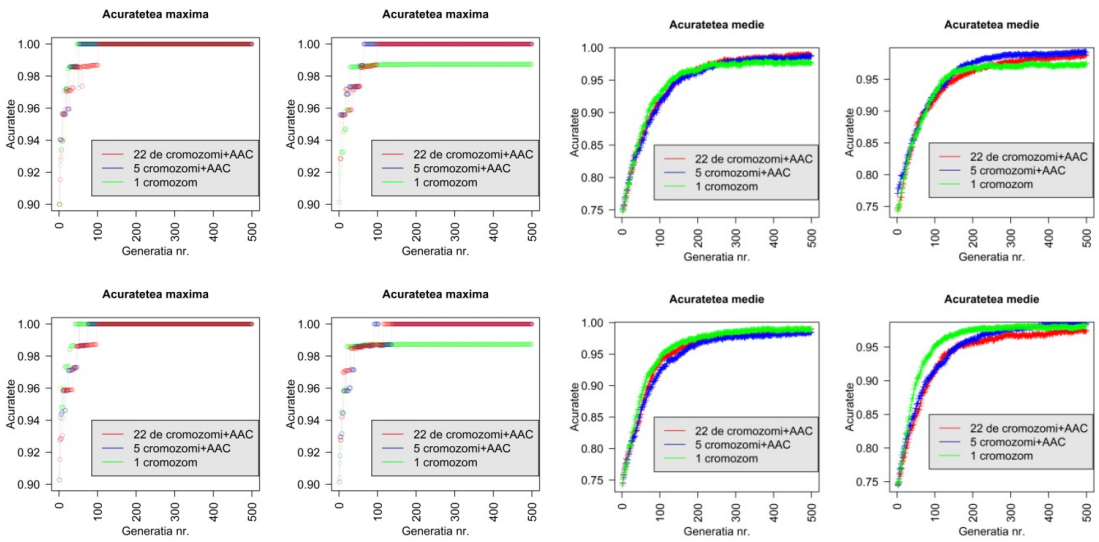


Fig. 5.4 – Evoluția acurateței maxime și medii pe patru populații inițiale.

5.9. Evaluarea operatorului pentru transpozoni

Pentru a obține comparații cât mai semnificative, am testat fiecare tip de operator pentru mutație într-un cadru uniform. Pe de o parte, am considerat că diversitatea introdusă cu operatorul de atribuire aleatorie al cromozomilor este foarte dezirabilă și va fi utilizată standard în selectarea atributelor cu AG diploizi din datele de microarray. Așadar, am efectuat testele pentru fiecare tip de operator pe setul de date cu 79 exemple și 628 de atribute distribuite pe 22 de cromozomi și AAC activat. Considerăm că superioritatea abordării DI2 față de DI1 impune această metodă pentru contextul de cercetare considerat. Totuși, am decis să utilizăm DI1 în testele noastre deoarece diversitatea genetică menținută de DI2, deși foarte dezirabilă pentru rezultate substanțiale, ar masca efectele diferitelor tipuri de mutații efectuate. În toate experimentele efectuate, am pornit de la populații inițiale cu 200 de indivizi, generate aleatoriu din random seeds prestabilite, pentru a asigura uniformitatea inițializării și o consistență sporită a rezultatelor. Un număr de 12 atribute a fost activat în populația inițială în toate situațiile. Valoarea utilizată pentru elitismul în context DI1 a fost 2%. Condiția de terminare utilizată a fost aceeași, 400 de generații. Rezultatele sunt prezentate prin comparație cu AG-ul diploid identic specificat, cu singura deosebire constând în tipul mutației aplicate. Mutații punctuale cu șansa apariției de 0.05%, s-au concretizat în 62 de operațiuni pe populație la fiecare generație. Am ales o șansă de apariție a mutației la o valoare superioară celei dezirabile într-un studiu real, pentru a vizualiza mai bine efectele mutațiilor studiate.

Aspectul calității evoluției la utilizarea operatorului pentru transpozoni este zugrăvit de ansamblul ilustrațiilor din Fig. 5.5-5.7. Această propunere, deservește în mod superior explorarea cu algoritmul genetic propus, în comparație cu mutația într-un punct. Deși capacitatea AG de-a părăsi un optim local este doar parțial afectată, evoluțiile acurateței maxime și medii sunt susținute de această abordare, iar numărul genelor active în populațiile înaintate nu crește semnificativ, așa cum se întâmplă în cazul mutației într-un singur punct.

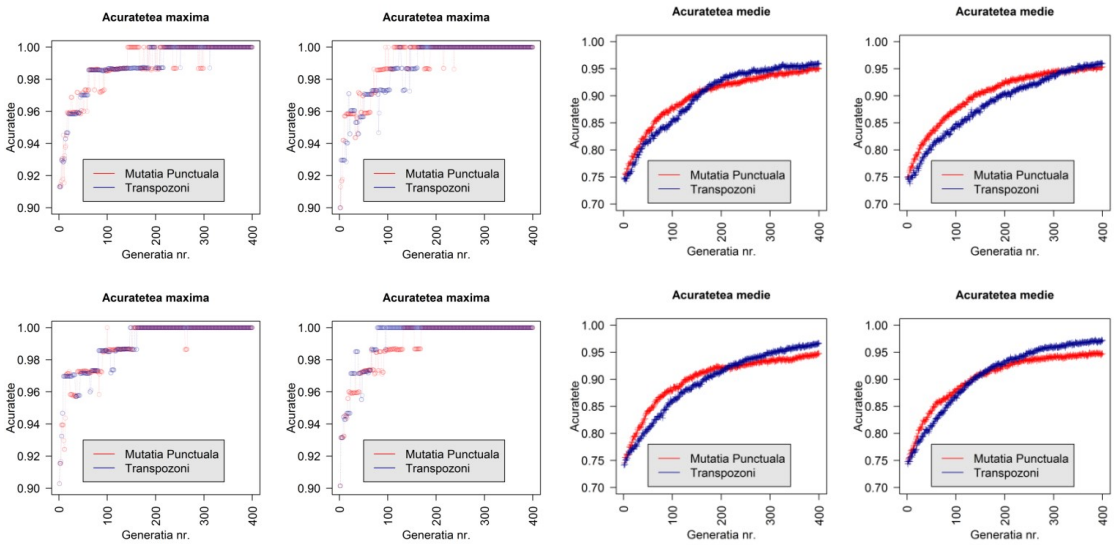


Fig. 5.5 - Evoluția acurateței maxime și medii la utilizarea transpozonilor comparativ cu mutația punctuală.

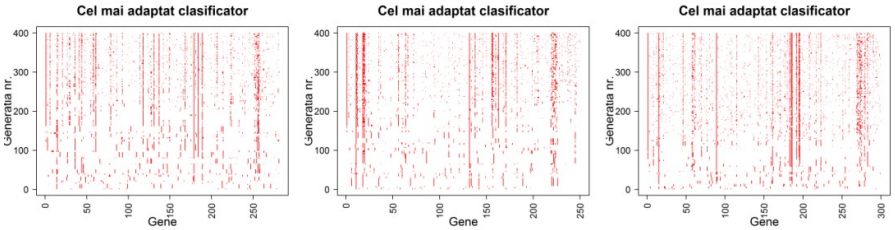


Fig. 5.6 - Evoluția celui mai adaptat clasificator în trei dintre populațiile inițiale, la utilizarea transpozonilor.

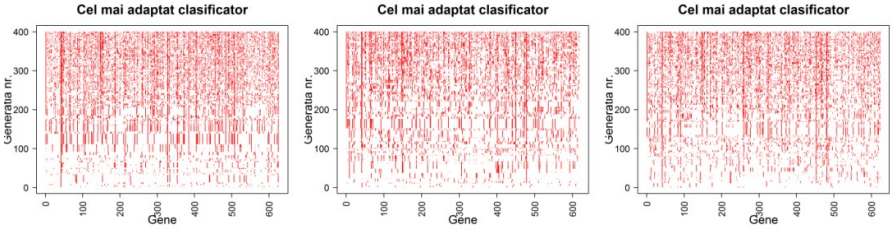


Fig. 5.7 - Evoluția celui mai adaptat clasificator în trei dintre populațiile inițiale, în abordarea cu mutația punctuală.

Argumente pro- și contra- utilizării mutațiilor descrise în acest context au fost determinate pentru fiecare operator în parte. Efecte dezirabile au fost observate în cazurile operatorilor pentru transpozoni și al ștergerii unui întreg cromozom. Cu toate acestea, efectele lor depind semnificativ de șansa ca o mutație să apară, care trebuie determinată empiric. Nu putem concluziona despre nici unul dintre operatorii pentru mutație propuși că este recomandabilă utilizarea lor pentru selecția atributelor în orice set de date microarray. Putem îndemna la testarea lor în contextul unui experiment similar, pentru evaluarea a priori a efectelor asupra calității evoluției.

5.10. Evaluarea efectelor cumulate ale DI2 și AAC

Ne propunem să testăm efectul combinat al celor două abordări, DI2 și AAC, pentru selectarea atributelor cu AG. Vom utiliza același context experimental ca și în cazul testării mutațiilor, AG cu

specificații identice cu o excepție: șansa ca o mutație de orice tip să apară a fost anulată. În acest cadru, testăm doi algoritmi genetici, unul implementat cu DI2 și AAC, iar celălalt cu DI1 și AAC.

Evoluțiile acurateței maxime (Fig. 5.8) relevă două aspecte extrem de importante. Pe de o parte, algoritmul cu DI2 și AAC prezintă tendința de-a părăsi un optim local, proprietate observată ocazional la AG cu DI1 și AAC. Implementarea cu AAC și DI1 reușește în marea majoritate a replicărilor experimentale efectuate să atingă o soluție cu acurateță de 100%. În 19 dintre cele 20 de experimente un grup de atribute cu care un clasificator discriminează perfect între exemple a fost determinat. Totuși, într-o situație această abordare nu a găsit un astfel de subgrup. Pe aceeași populație inițială, generată din același random seeds AG beneficiind de DI2 și AAC a determinat un subgrup de atribute care permit clasificarea cu o acurateță de 100%.

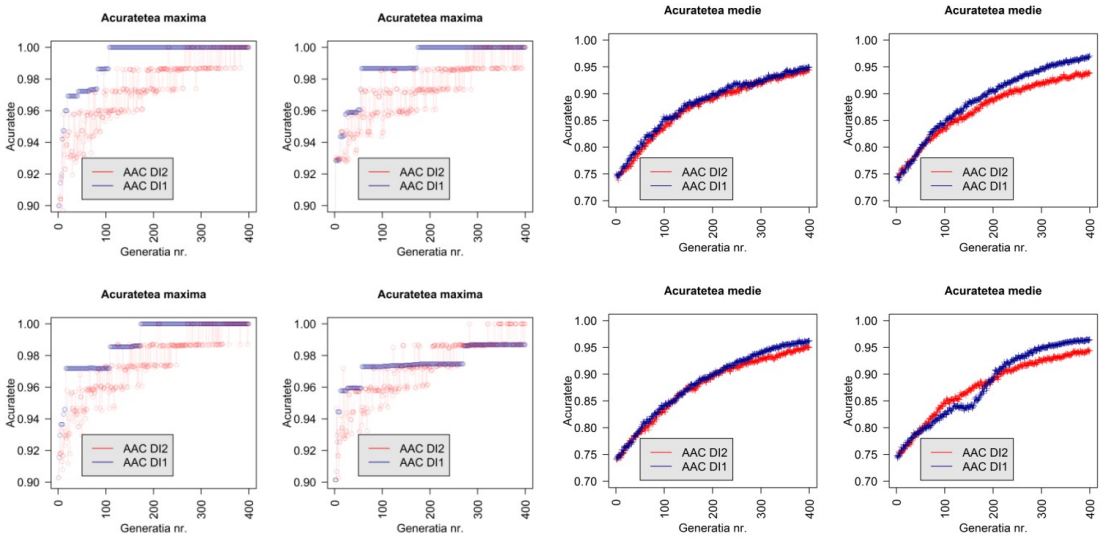


Fig. 5.8 - Evoluția acurateței maxime și medii la utilizarea DI2 plus AAC comparativ cu DI1 combinat cu AAC.

În ceea ce privește evoluția acurateței medii în populație pe 400 de generații, remarcăm o ușoară superioritate a implementării cu DI1. Tendința sporită de-a părăsi un optim local în favoarea explorării de noi soluții observată cu DI2 și AAC, înclină balanța în această direcție, în ciuda avantajului minor al acurateței medii în populațiile înaintate obținut cu metoda alternativă.

6. Concluzii și contribuții personale

Pentru realizarea metodei propuse, am studiat fenomene care fundamentează evoluția naturală. Am pornit de la premiza că, principiile care susțin evoluția naturală sunt validate de succesul pe parcursul a miliarde de ani. Elucidarea acestor legi și implementarea lor în algoritmii evoluționiști va continua să amelioreze metodele existente și vor impulsiona conceperea de abordări noi.

1) **Dominanța incompletă.** Modelul propus reprezintă o alternativă la definirea unei scheme de dominanță, particulare unui cadru individual, pentru maparea genotipului la fenotip în implementarea algoritmilor genetici diploizi. Pe parcursul dezvoltării experimentelor s-a dovedit utilă elaborarea a două variante, **DI1** și **DI2**, cu proprietăți semnificativ diferite și aplicabile de elecție în contexte diferite. Experimentele efectuate recomandă implementarea DI2 în cazul selectării atributelor în datele ADN microarray.
2) **Operatorul pentru atribuirea aleatorie a cromozomilor (AAC)**. Operatorul propus modelează un fenomen care susține diversitatea genetică și are loc în timpul meiozei. Testele efectuate confirmă utilitatea implementării acestui model în algoritmii genetici.
3) **Pachetul software dGAselID**. Perfect integrat în R și Bioconductor, pachetul dGAselID facilitează utilizarea metodei propuse pentru selectarea atributelor în contextul analizei genetice

și alte domenii de cercetare. Metoda este astfel accesibilă unei comunități foarte diverse de investigatori din mediul academic.

4) **Alternative la mutația punctuală**. Am modelat și evaluat impactul unor operatori pentru mutații în analiza datelor ADN microarray. Operatorii pentru mutații implementați în pachetul software dGAselID sunt:
    1. Operatorul pentru **mutația fără sens**,
    2. Operatorul pentru **mutația cu deplasare**,
    3. Operatorul pentru **mutația cu ștergerea unui segment**,
    4. Operatorul pentru **mutația ștergerea unui cromozom**,
    5. Operatorul pentru **mutația de tip transpozon**.

6.1. Perspectivă de dezvoltare

Pentru viitor, ne propunem următoarele direcții de dezvoltare a direcțiilor urmate pe parcursul realizării tezei de doctorat:

1) testarea efectului *separării genomului într-un număr variabil de cromozomi* în selectarea atributelor din date ADN microarray cu **biochip-uri adaptabile** unei cercetări particulare,
2) evaluarea modelului *dominanței incomplete* asupra evoluției algoritmilor genetici angajați pentru selectarea atributelor din date aparținând **altor domenii de cercetare**, în afara spectrului analizei genetice,
3) validarea *operatorului AAC* în algoritmii genetici implicați în problematici variate, din **domenii diferite de investigare**,
4) *elaborarea unui operator pentru mutații mai eficient* în susținerea evoluției AG pentru selectarea atributelor din date microarray.

BIBLIOGRAFIE

[1] J.H. Holland. Adaptation in natural and artificial systems. University of Michigan Press, 1975.
[2] W.D. Hillis. Co−evolving parasites improve simulated evolution as an optimization procedure. Physica D 42:228–234, 1990.
[3] J.R. Levenick. Inserting introns improves genetic algorithm success rate: taking a cue from biology. Proceedings of the Fourth International Conference on Genetic Algorithms. Morgan Kaufmann, 1991.
[4] M. Mitchell. An introduction to genetic algorithms. The MIT Press, Cambridge, Massachusetts, London, England, 1999.
[5] J.M. Baldwin. A new factor in evolution. American Naturalist 30: 441–451, 536–553, 1986.
[6] R. Lewis. Human genetics, Ediţia a XI-a. McGraw-Hill Science/Engineering/Math, pag. 46, 2014.
[7] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[8] W. Huber, V.J. Carey, R. Gentleman, ..., M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods, 2015:12, 115.
[9] A. Koschmieder, K. Zimmermann, S. Trissl, T. Stoltmann și U. Leser. Tools for managing and analyzing microarray data. Brief Bioinform 13(1):46–60, 2012.
[10] W. Gregory Alvord, J.A. Roayaei,O.A. Quiñones și K.T. Schneider. A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R. Brief Bioinform. 8(6):415-31, 2007.
[11] X. Li. ALL: a data package. Pachet R versiunea 1.14.0, 2009.