

**Universitatea Tehnică a Moldovei**

**DEZVOLTAREA UNEI APLICAȚII pentru alinierea și sortarea secvențelor de ADN**

|  |  |
| --- | --- |
| **Studenți:** | **gr. TI-231,**  **Olari Veronica** **Moglan Anastasia** **Mudrenco Vlada** |
| **Coordonator:** | **Andrievschi-Bagrin Veronica, asist.univ.** |

**Chișinau 2024**

**CUPRINS**

ABREVIERI………………………………………………………………………………………..3

INTRODUCERE……………………………………………………………………………………5

1 ANALIZA DOMENIULUI DE STUDIU…………………………………………………………..6

1.1 Importanța temei………………………………………………………………………………...6

1.2 Sisteme similare cu proiectul realizat……………………………………………………………..6

1.3 Scopul, obiectivele si cerințele sistemului…………………………………………………………6

2 ALEGEREA ALGORITMILOR PENTRU REZOLVAREA PROBLEMEI………………………...10

2.1 Argumentarea alegerii algoritmilor……………………………………………………………....12

2.2 Implementarea algoritmilor într-un limbaj de programare structurat………………………………13

3 MODELAREA ȘI PROIECTAREA SISTEMULUI INFORMATIC……………………………….15

3.1 Descrierea comportamentală a sistemului………………………………………………………..15

3.1.1 Imaginea generală asupra sistemului………………………………………………………..…15

3.1.2 Modelarea vizuală a fluxurilor………………………………………………………………...17  
3.1.3 Stările de tranzacție a sistemului………………………………………………………………18  
3.1.4 Descrierea scenariilor de utilizare a aplicației………………………………………………….19  
**3.1.5 Fluxurile de mesaje și legăturile dintre componentele sistemului**……………………………….20  
**3.2 Descrierea structurii a sistemului**……………………………………………………………..…21

INTERFAȚA……………………………………………………………………………………………...25  
CONCLUZII……………………………………….……………………………………………...28  
BIBLIOGRAFIE…………………………………………………………………………………..29

**ABREVIERI**

### DNA - Materialul genetic al organismelor vii.

### DNA sequences - Secvențe de nucleotide din ADN.

### Reference - O sursă de comparație sau standard.

### Cost - Resursele necesare pentru un proces sau activitate.

### Matching - Procesul de comparare și potrivire a unor elemente.

### Substitution - Înlocuirea unui element cu altul.

### Gaps - Goluri în secvențele de ADN sau proteine.

### Stability - Capacitatea unui sistem de a rămâne constant sau echilibrat.

### Percentage - O valoare exprimată ca fracție din 100.

### Entropy - Măsura aleatorietății sau incertitudinii unui sistem.

### Shannon - Teoretician cunoscut pentru lucrările sale despre entropie și informație.

### Mutation - Schimbarea permanentă a secvenței de ADN.

### G-C - Perechea de baze azotate guanină-citozină în ADN.

### A-T - Perechea de baze azotate adenină-timină în ADN.

### Bioinformatics - Domeniu care folosește tehnologia informatică pentru a analiza date biologice.

### Algorithms - Seturi de pași pentru rezolvarea unor probleme.

### Needleman-Wunsch - Algoritm pentru alinierea globală a secvențelor de ADN sau proteine.

### Smith-Waterman - Algoritm pentru alinierea locală a secvențelor.

### Automated classification - Procesul de clasificare a datelor prin metode automate.

### Similarity - Gradul de asemănare între două obiecte.

### Evolutionary - Relativ la evoluția speciilor sau proceselor biologice.

### Evolutionary trees - Reprezentări grafice ale relațiilor evolutive dintre specii.

### Scoring - Atribuirea unor valori numerice pentru evaluarea calității unui proces.

### BLOSUM62 - Matrice de scoruri pentru alinierea secvențelor de proteine.

### Amino acids - Componentele fundamentale ale proteinelor.

### Proteins - Molecule complexe care îndeplinesc diverse funcții biologice.

### Clustering - Gruparea obiectelor pe baza asemănării lor.

### Molecular biology - Studiul proceselor biologice la nivel molecular.

### Biomedical research - Cercetare științifică în domeniul biomedicinei.

### Research - Procesul de investigare sistematică pentru a obține noi cunoștințe.

### Dynamic programming - Tehnică de programare pentru rezolvarea problemelor complexe prin divizarea acestora în subprobleme.

### Matrix - Structură matematică organizată pe linii și coloane.

### Insertion - Adăugarea unui element într-o secvență.

### Deletion - Înlăturarea unui element dintr-o secvență.

### Sequence alignment - Procesul de aliniere a două sau mai multe secvențe biologice.

### Genetic diseases - Afecțiuni cauzate de anomalii genetice.

### Complementary sequence - Secvență care se potrivește cu alta conform regulilor de complementaritate.

### Fibrosis cystic - Boală genetică care afectează plămânii și sistemul digestiv.

### Huntington's disease - Boală genetică neurodegenerativă.

### Divide and conquer - Tehnică de rezolvare a problemelor prin împărțirea lor în subprobleme mai mici.

### Greedy algorithm - Algoritm care face alegeri locale optime la fiecare pas, în speranța că soluția globală va fi optimă.

### Mutation count - Numărul de mutații într-o secvență.

### Sequence diversity - Diversitatea secvențelor genetice într-o populație.

### Genetic stability - Capacitatea unui organism de a menține caracteristicile genetice constante de-a lungul generațiilor.

### Dynamic programming matrix - Matrice utilizată pentru implementarea tehnicii de programare dinamică.

### Genetic disorders - Afecțiuni cauzate de defecte genetice.

### Message flow - Direcția și secvența mesajelor într-un sistem de comunicație.

### FSM (Finite State Machine) - Model matematic care descrie un sistem cu un număr finit de stări.

### State transitions - Schimbările de stare într-un sistem.

### User input - Datele sau comenzile furnizate de utilizator.

### FASTA - Format de fișier utilizat pentru stocarea secvențelor de ADN.

### Entropy - Măsura aleatorietății unui sistem (din nou menționat, dar definit mai sus).

### 

### **INTRODUCERE**

### Tema acestui proiect, "Eficacitatea șirurilor concatenate: Programare dinamică pentru alinierea și sortarea secvențelor de ADN" adresează o problemă fundamentală în bioinformatică – alinierea optimă a secvențelor ADN. În contextul geneticii, compararea secvențelor este esențială pentru a identifica variații genetice și a înțelege relațiile evolutive. Secvențele ADN, formate din nucleotidele A, T, C și G, poartă informații vitale, iar alinierea lor corectă poate dezvălui diferențe subtile sau similitudini între specii, gene sau probe biologice.

### Proiectul explorează tehnici de programare dinamică și alte paradigme algoritmice pentru a soluționa eficient această problemă. Prin aplicarea unor metode bine stabilite, precum algoritmii Needleman-Wunsch și Smith-Waterman, proiectul nu doar optimizează procesele de aliniere, ci și analizează diversitatea și stabilitatea genetică a secvențelor. Astfel, acest demers contribuie la o mai bună înțelegere a variațiilor genetice și a mecanismelor asociate, cu implicații directe în diagnosticarea bolilor genetice și cercetarea biomedicală.

Problema abordată în cadrul acestui proiect este alinierea secvențelor de ADN, o provocare frecvent întâlnită în bioinformatică. ADN-ul, molecula responsabilă pentru codificarea informației genetice, este format din secvențe de nucleotide care includ patru tipuri de baze azotate: adenină (A), timină (T), citozină (C) și guanină (G). În multe aplicații bioinformatice, este necesar să se compare secvențe ADN diferite pentru a detecta mutații, a identifica similitudini între specii sau a analiza evoluția genelor.

Problema constă în alinierea a două secvențe ADN pentru a maximiza similitudinea lor, utilizând operațiuni de potrivire a caracterelor identice, substituirea caracterelor diferite, inserarea sau ștergerea de caractere (goluri). Această aliniere este esențială în diagnosticul genetic, detectarea mutațiilor și cercetarea biomedicală, deoarece poate dezvălui informații critice despre predispoziția la boli, evoluția genomurilor și stabilitatea genetică.

#### Datele de intrare:

- Secvența ADN analizată (seq): Un șir de caractere format din literele A, C, G, T, ce reprezintă secvența ADN supusă analizei. Aceasta poate proveni din datele genetice ale unui pacient sau dintr-o bază de date genomice.

- Secvența de referință (reference): Un șir de caractere considerat "normal" sau sănătos, utilizat pentru a compara secvența analizată. Poate reprezenta o secvență ideală pentru specia sau genomul în cauză.

**1 ANALIZA DOMENIULUI DE STUDIU**

* 1. **Importanța temei**

Tema proiectului este de o importanță deosebită în domeniul bioinformaticii și cercetării genetice. În contextul actual, în care datele genetice sunt din ce în ce mai accesibile și utilizate pe scară largă în diagnosticarea bolilor, manipularea și analiza eficientă a acestor date devine esențială. Alinierea secvențelor ADN reprezintă un aspect fundamental, având un rol crucial în identificarea variațiilor genetice, în studierea evoluției speciilor și în înțelegerea relațiilor dintre secvențe. Tehnicile de aliniere globală și locală oferă instrumentele necesare pentru a analiza similitudinile și diferențele subtile dintre secvențele ADN, facilitând astfel diagnosticarea genetică timpurie.

Analiza stabilității și diversității secvențelor ADN contribuie, de asemenea, la înțelegerea comportamentului genetic al organismelor. Stabilitatea unei secvențe, determinată prin procentul de perechi G-C, poate dezvălui rezistența acesteia la schimbări, fiind esențială pentru prevenirea bolilor genetice. În același timp, calculul entropiei oferă o măsură a variabilității și distribuției nucleotidelelor, evidențiind potențiale anomalii genetice.

Având în vedere importanța crescută a studiului bolilor genetice și a terapiilor personalizate în cercetarea biomedicală modernă, proiectul aduce o contribuție semnificativă la dezvoltarea unor metode precise și eficiente pentru analiza ADN-ului. Acesta oferă o bază solidă pentru îmbunătățirea metodelor de diagnostic genetic și pentru cercetări mai ample, cum ar fi studiile evoluționare sau identificarea markerilor genetici asociați cu diverse afecțiuni. Astfel, tema proiectului se dovedește a fi nu doar relevantă din punct de vedere științific, ci și esențială pentru progresul în medicină și bioinformatică, având aplicații directe în sănătatea umană și în aprofundarea cunoașterii fundamentelor vieții.

* 1. **Scopul și obiectivele**

Obiectivele:

Investigarea și implementarea algoritmilor Needleman-Wunsch și Smith-Waterman pentru alinierea globală și locală a secvențelor ADN.Compararea performanței celor două metode în termeni de complexitate temporală și spațială, și aplicabilitatea lor în diverse contexte bioinformatice.Implementarea unui algoritm de calcul al entropiei Shannon pentru a evalua diversitatea secvențelor ADN și distribuția nucleotidelelor A, C, G, T.Evaluarea stabilității secvențelor ADN pe baza conținutului de perechi G-C, analizând importanța acestui parametru în contextul stabilității structurale a ADN-ului.Dezvoltarea unui algoritm pentru detectarea mutațiilor între secvența de ADN analizată și secvența de referință, calculând numărul și procentul de diferențe.Interpretarea rezultatelor pentru a trasa concluzii cu privire la riscurile genetice și posibilele predispoziții la diverse afecțiuni genetice.Îmbunătățirea eficienței algoritmilor pentru a permite analiza rapidă a secvențelor mari de ADN, necesară în cercetarea biomedicală.Asigurarea unei implementări robuste și scalabile care să poată fi aplicată în laboratoare de bioinformatică sau pentru utilizarea în aplicații software de analiză genetică.

Scopul:

Unul dintre scopurile majore este de a dezvolta algoritmi capabili să identifice mutațiile care pot fi corelate cu predispoziția la anumite maladii genetice, contribuind astfel la identificarea timpurie a riscurilor pentru sănătatea umană.Proiectul urmărește să analizeze stabilitatea și diversitatea secvențelor ADN, oferind perspective valoroase asupra evoluției genomurilor și a diferențelor genetice între specii.Scopul pe termen lung este de a oferi un set de instrumente algoritmice și tehnici care să fie utilizabile în studii viitoare de bioinformatică, în special în analiza secvențelor ADN și în cercetarea genetică aplicată.

* 1. **Proiecte similare**

Un sistem similar, dar cu un accent ușor diferit, potrivit pentru o altă aplicație bioinformatică relevantă ar fi: „Eficiența metodelor algoritmice pentru clasificarea și analiza filogenetică a secvențelor proteice”.[1]

Proiectul propune o metodă avansată pentru analiza și interpretarea relațiilor dintre secvențele de proteine, având aplicații extinse în biologie, cercetare medicală și evoluționism. Acest proiect utilizează tehnici avansate de aliniere a secvențelor, bazate pe algoritmi precum Needleman-Wunsch și Smith-Waterman. Algoritmul Needleman-Wunsch este utilizat pentru alinierea completă a secvențelor, oferind o analiză detaliată a acestora pe întreaga lor lungime, fiind potrivit pentru compararea secvențelor de lungimi similare. Algoritmul Smith-Waterman se concentrează pe alinierea unor porțiuni specifice, fiind util pentru identificarea regiunilor de interes care prezintă asemănări relevante. Tabelul 1.1 prezintă o comparație între doi algoritmi fundamentali utilizați în bioinformatică: Needleman-Wunsch, potrivit pentru alinierea globală a secvențelor ADN, și Smith-Waterman, utilizat pentru identificarea regiunilor similare din cadrul secvențelor. Această comparație ajută la înțelegerea aplicabilității fiecărui algoritm în funcție de contextul analizei bioinformatice.

**Tabelul 1.1 - Comparație între Algoritmii Needleman-Wunsch și Smith-Waterman**

|  |  |  |
| --- | --- | --- |
| **Criteriu** | **Needleman-Wunsch** | **Smith-Waterman** |
| **Tip de aliniere** | Globală | Locală |
| **Utilizare** | Secvențe de lungimi similare | Identificarea regiunilor de interes |
| **Complexitate temporală** | O(n × m) | O(n × m) |

Un element important al proiectului este folosirea matricilor de substituție, cum ar fi BLOSUM62, care evaluează gradul de asemănare dintre aminoacizi din secvențele analizate. Aceste matrici oferă o evaluare a gradului de păstrare a caracteristicilor proteinelor și ajută la determinarea funcțiilor lor biologice. Prin aplicarea acestor metode, proiectul analizează în profunzime cât de apropiate sunt proteinele studiate de cele considerate de referință.

Proiectul include și metode de clasificare automată, precum gruparea în funcție de asemănări între secvențe. Această clasificare permite împărțirea proteinelor în familii funcționale sau structurale, oferind informații despre domeniile comune și despre funcțiile acestora. Gruparea secvențelor facilitează analiza familiilor mari de proteine și contribuie la înțelegerea mai clară a relațiilor dintre ele.

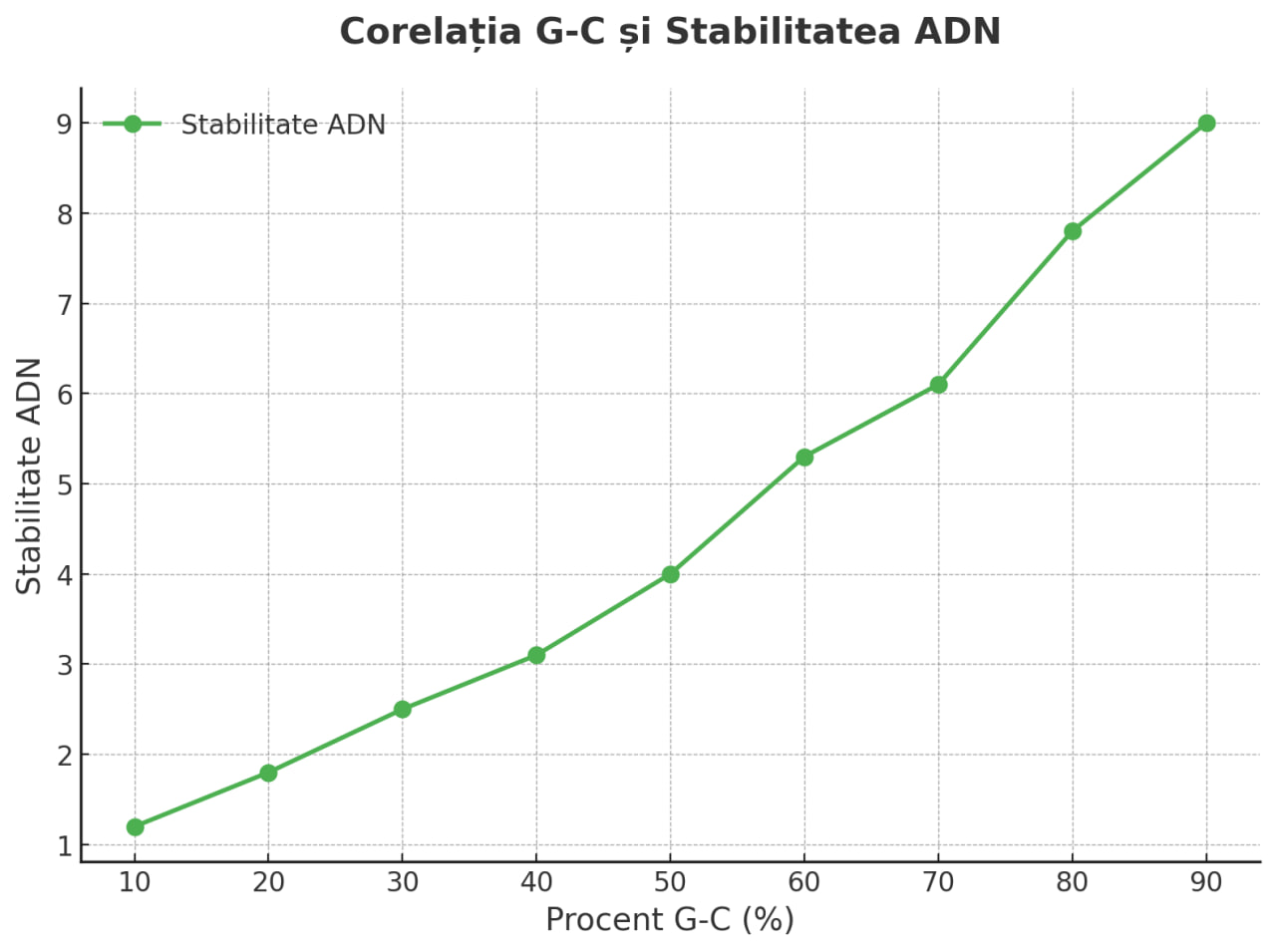
În plus, proiectul folosește metode pentru construirea arborilor evolutivi, care reflectă legăturile dintre proteine din punct de vedere evolutiv. Aceste structuri oferă o imagine clară a gradului de înrudire dintre proteine, evidențiind procesele de evoluție care au dus la diversitatea biologică de astăzi. Analiza relațiilor evolutive este esențială pentru studiul schimbărilor genetice și pentru identificarea trăsăturilor moștenite de-a lungul timpului.[2]

Rezultatele obținute includ scoruri de asemănare rezultate din alinieri, clasificări detaliate ale proteinelor și măsuri ale diferențelor dintre secvențele studiate și cele de referință. În plus, arborii evolutivi și analiza legăturilor genetice oferă perspective esențiale asupra diversității și funcționalității proteinelor.

Proiectul este conceput pentru a sprijini analiza rapidă și corectă a secvențelor de proteine, fiind util în descoperirea trăsăturilor moleculare, studierea variațiilor genetice și identificarea de noi ținte pentru tratamente. Aplicațiile sale practice sunt deosebit de valoroase în cercetarea medicală, descoperirea medicamentelor și biologia moleculară, reprezentând un instrument important pentru progresul științific.

In figura 1.1 este reprezentat un grafic care evidențiază relația dintre procentul de perechi G-C dintr-o secvență ADN (axa X) și stabilitatea acesteia (axa Y). Observăm o tendință ascendentă care arată că, pe măsură ce proporția perechilor G-C crește, stabilitatea secvenței ADN devine mai mare. Acest lucru se datorează legăturilor mai puternice dintre nucleotidele G-C, care formează trei legături de hidrogen, comparativ cu cele două legături ale perechilor A-T.

Prin folosirea unor metode eficiente și bine fundamentate, proiectul contribuie la înțelegerea profundă a relațiilor dintre secvențele de proteine și la dezvoltarea de soluții pentru îmbunătățirea sănătății umane.

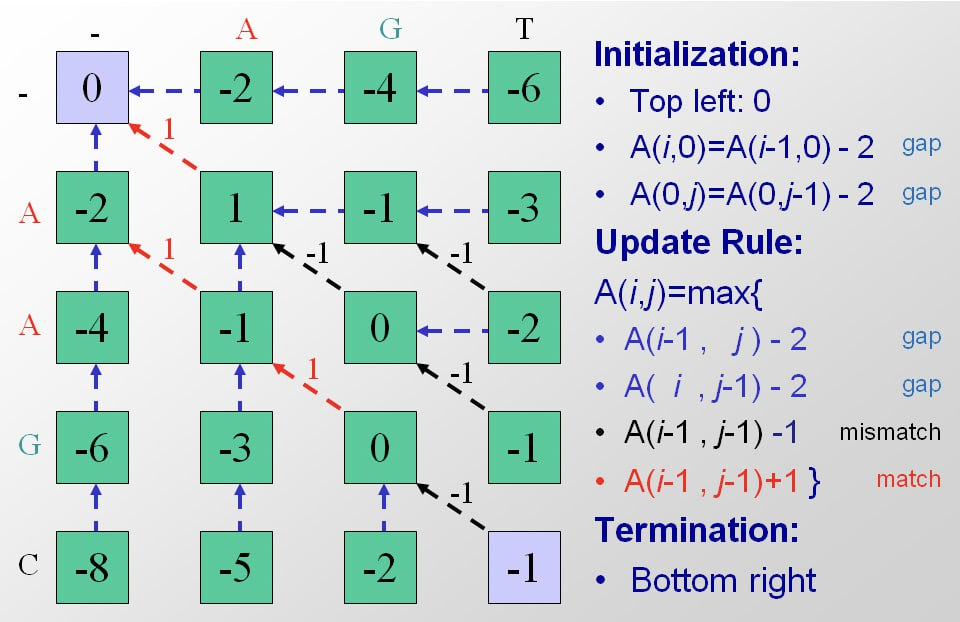


**Figura 1.1 -Coretația G-C si Stabilitatea ADN**

### **2 ALEGEREA ALGORITMILOR PENTRU REZOLVAREA PROBLEMEI**

Pentru a rezolva problema alinierii secvențelor ADN și a efectua analize suplimentare (mutații, stabilitate și entropie), se vor utiliza următorii algoritmi:

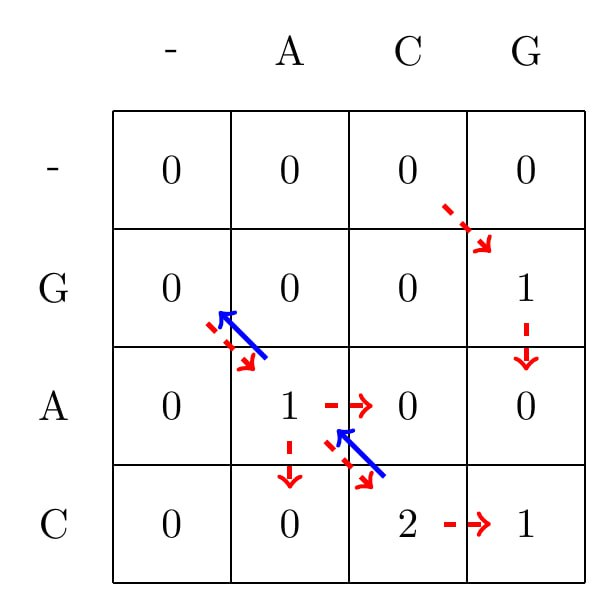
- **Algoritmul Needleman-Wunsch pentru aliniere globală.** Ideal pentru potrivirea completă a două secvențe de lungimi similare. Acesta utilizează o matrice pentru a calcula costul minim al alinierii, luând în considerare potrivirile, substituțiile și golurile. In figura 1.2 este explicat acest lucru:



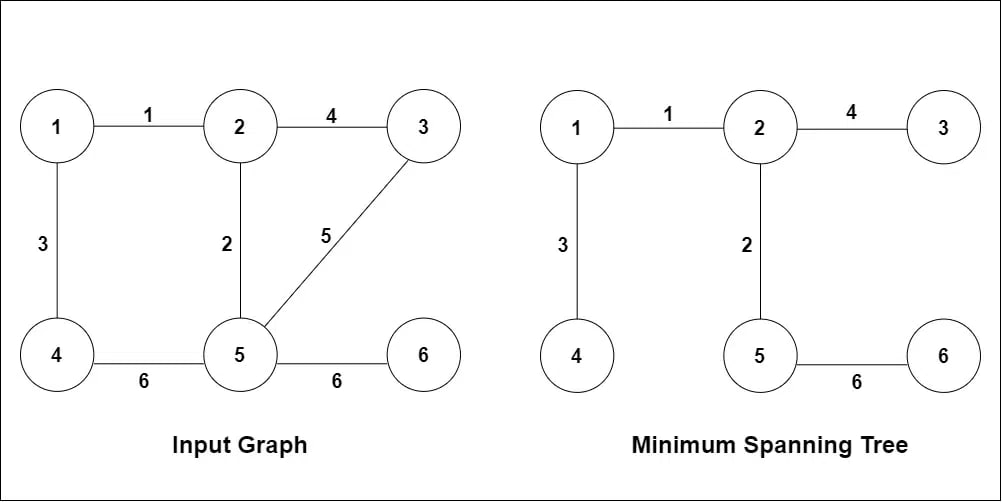
**Figura 2.1 - Algoritmul Needleman-Wunsch pentru aliniere globală**

- **Algoritmul Smith-Waterman pentru aliniere locală.** Folosit pentru a identifica regiunile cele mai similare dintre două secvențe, fiind util în cazurile în care doar anumite segmente sunt relevante pentru analiza mutațiilor sau a similitudinilor. In figura 1.3 este explicat principiul de lucru acestui algoritm.

**- Algoritmul lui Kruskal.** Se va analiza comportamentul bazelor azotate intr-o catea de AND cu ajutorul algoritmului lui Kruskal.Se doreste reprezentarea conexiunilor dintre baze ca un graf structurat. Un graf derivat din structura ADN implică reprezentarea nodurilor ca perechi de baze azotate (de exemplu, G-C, A-T), iar muchiile ca interacțiuni ponderate pe baza energiei de legare sau a stabilității. In figura 1.4 este explicat cum o astfel de reprezentare poate fi utila pentru analiza stabilitatii structurale a ADN-ului si identificarea celor mai eficiente conexiuni, evitând ciclurile și minimizând energia totală a interacțiunilor.[3]



**Figura 2.2 - Algoritmul Smith-Waterman pentru aliniere locală.**



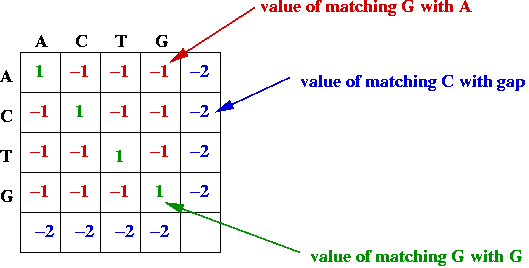
**Figura 2.3 - Algoritmul lui Kruskal**

- **Algoritm pentru calculul entropiei secvenței.** Acest algoritm calculează entropia Shannon pe baza frecvenței fiecărui nucleotid (A, C, G, T), oferind informații despre diversitatea genetică și distribuția nucleotidelelor.[4]

- **Algoritm pentru evaluarea stabilității secvenței AND.** Calculează procentul de perechi G-C și A-T din secvență, stabilind un scor care reflectă stabilitatea chimică și termică a ADN-ului.

### **2.1** **Argumentarea alegerii algoritmilor**

### - **Needleman-Wunsch pentru aliniere globală.** Acest algoritm este ales datorită capacității sale de a găsi alinierea optimă între două secvențe în întregime, aspect necesar pentru analize detaliate în cazurile în care este important să fie luată în considerare fiecare nucleotid. Algoritmul este eficient pentru secvențe de lungimi similare și oferă o soluție globală optimă, având avantajul utilizării unei abordări de programare dinamică pentru a minimiza costurile alinierii. Figură 1.5 reprezintă o matrice de punctaje utilizată în algoritmul Needleman-Wunsch pentru alinierea globală a secvențelor ADN. Valorile din matrice indică scorurile asociate potrivirilor (verde), neconcordanțelor (roșu) și introducerii de goluri (albastru), oferind un ghid vizual pentru calcularea alinierilor optime între nucleotide.[5]



**Figura 2.4 - Needleman-Wunsch pentru aliniere globală**

#### - **Smith-Waterman pentru aliniere locală.** Smith-Waterman este preferat atunci când se dorește identificarea alinierilor optime într-o regiune specifică a secvențelor, fiind util pentru detecția variațiilor locale, cum ar fi mutațiile sau inserțiile/delețiile. Algoritmul penalizează potrivirile slabe și ignoră porțiunile necorespunzătoare, maximizând scorul alinierii pentru regiunile similare.

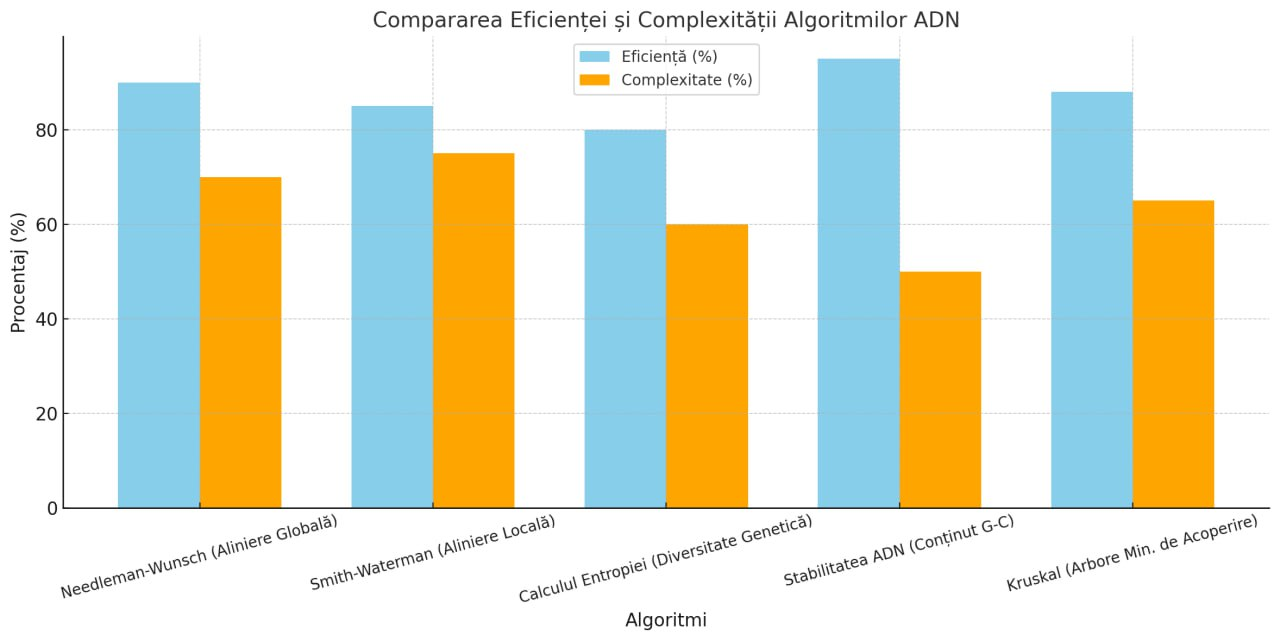
- **Calculul entropiei pentru diversitatea genetică.** Calcularea entropiei oferă o măsură matematică pentru distribuția nucleotidelelor. O valoare ridicată indică o distribuție uniformă și echilibrată, în timp ce o valoare scăzută poate semnala anomalii genetice sau presiuni selective. Algoritmul de calcul al entropiei Shannon este simplu de implementat și eficient pentru detectarea dezechilibrelor în secvență.

#### - **Evaluarea stabilității pe baza conținutului G-C.** Stabilitatea genetică este strâns legată de proporția perechilor G-C în ADN, deoarece acestea formează trei legături de hidrogen, comparativ cu două în cazul perechilor A-T. Algoritmul care evaluează stabilitatea pe baza procentului G-C ajută la identificarea regiunilor predispuse la denaturare și la analiza riscului asociat unor boli genetice.

### **2.2 Implementarea algoritmilor într-un limbaj de programare structurat**

#### În această secțiune, sunt prezentate detaliile implementării algoritmilor pentru alinierea secvențelor ADN, calculul entropiei și evaluarea stabilității secvenței. Implementarea se va face în limbajul C++, cunoscut pentru eficiența sa în manipularea datelor și execuția rapidă. Codurile oferă soluții clare pentru realizarea alinierilor globale și locale, calcularea diversității genetice și evaluarea stabilității ADN-ului. Algoritmul Needleman-Wunsch este utilizat pentru a găsi alinierea globală optimă între două secvențe ADN, folosind o abordare de programare dinamică. Matricea dp este utilizată pentru a stoca costurile minime pentru alinierea parțială a secvențelor.[6]

In figura 1.6 este reprezentat garficul care ilustrează eficiența (procentaj de performanță) și complexitatea (dificultatea de implementare) a cinci algoritmi utilizați pentru analiza secvențelor ADN. Algoritmii Needleman-Wunsch și Smith-Waterman sunt evaluați pentru alinierea secvențelor, în timp ce calculul entropiei măsoară diversitatea genetică, iar evaluarea stabilității ADN analizează conținutul de perechi G-C. Relația dintre acești factori ajută la identificarea celui mai potrivit algoritm pentru fiecare tip de analiză.

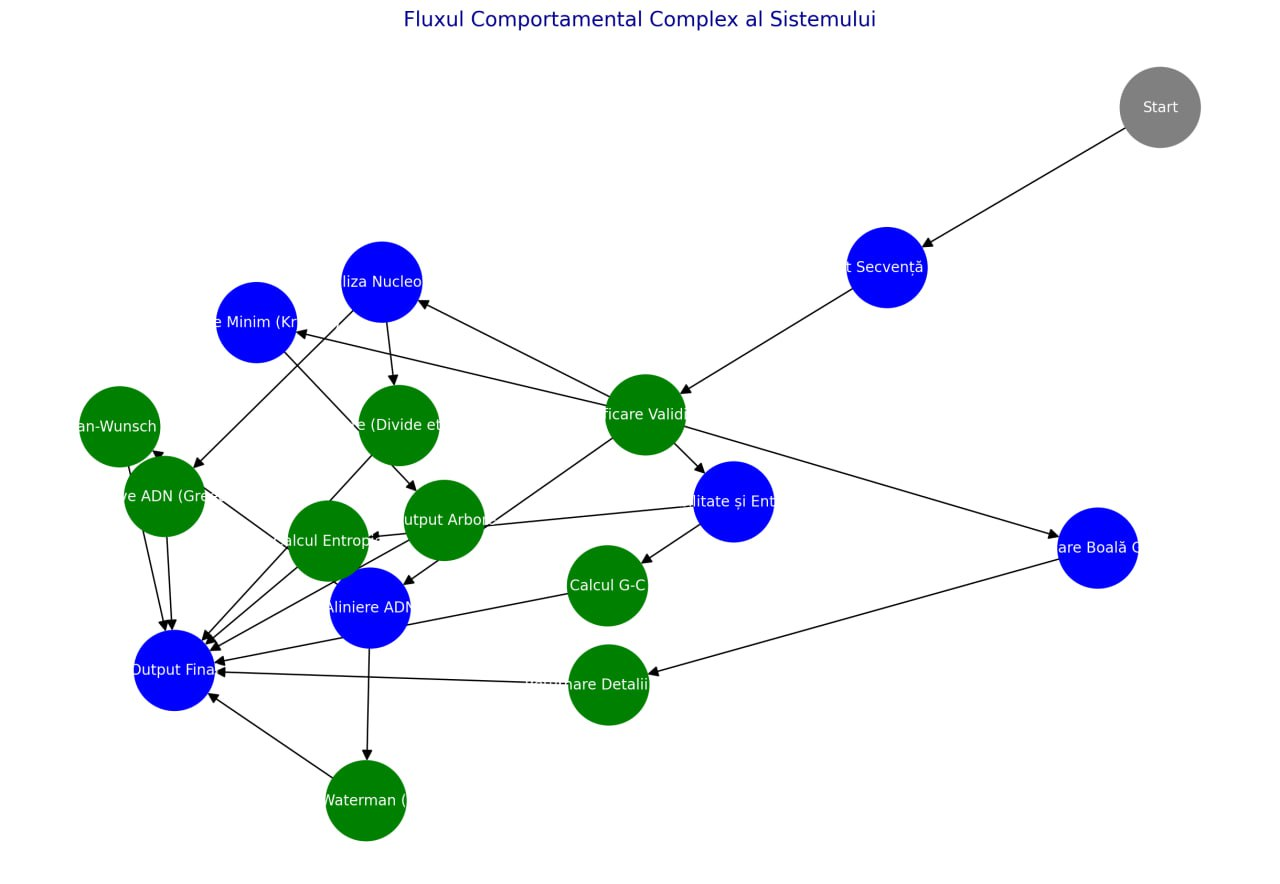


**Figura 2.5 - Compararea Eficienței și Complexității Algoritmilor ADN**

**3 MODELAREA ȘI PROIECTAREA SISTEMULUI INFORMATIC**

#### **3.1 Descrierea comportamentală a sistemului**

Figura 3.1 ilustrează interacțiunea dintre componentele programului. Procesul începe cu validarea inputului utilizatorului și continuă cu aplicarea algoritmilor selectați pentru a genera rezultate precise și detaliate. În fiecare etapă, sunt folosite metode avansate pentru a analiza ADN-ul din mai multe perspective: genetică, structurală și statistică. Această modelare asigură un sistem robust și scalabil.



**Figura 3.1 - Interacțiunea dintre componentele programului**

##### **3.1.1 Imaginea generală asupra sistemului**

Sistemul dezvoltat este o aplicație bioinformatică dedicată analizei secvențelor ADN, care include mai multe module algoritmice pentru manipularea și procesarea datelor genetice. În esență, sistemul permite analiza secvențelor de ADN introduse de utilizator, detectarea posibilelor boli genetice asociate și aplicarea de tehnici avansate de aliniere a secvențelor, analiza frecvenței nucleotidelor și căutarea motivelor ADN. Arhitectura sistemului este modulară și se bazează pe algoritmi fundamentali din domeniul bioinformaticii și informaticii teoretice. Acesta este structurat în patru module principale:

**- Modulul de introducere a datelor.** Permite utilizatorului să introducă o secvență ADN, fie manual, fie printr-un fișier. Sistemul preia această secvență și o verifică pentru a se asigura că este validă și că nu conține caractere de ARN (precum „U”). Generarea complementului ADN pentru secvența introdusă. Acesta se realizează folosind o hartă a complementului ADN, care leagă fiecare bază de complementul său: „A” se leagă cu „T”, „T” cu „A”, „G” cu „C” și „C” cu „G”.**Detectarea bolilor genetice**: Sistemul verifică secvența de ADN pentru posibile secvențe asociate cu boli genetice cunoscute, cum ar fi Fibroza chistică sau Boala Huntington. Fiecare secvență ADN este comparată cu o bază de date internă a bolilor genetice, iar bolile sunt identificate pe baza unor secvențe complementare specifice.[7]

- **Modulul de aliniere a secvențelor.Alinierea globală (Needleman-Wunsch)**: Se aplică algoritmul Needleman-Wunsch pentru a realiza alinierea globală a secvenței de ADN cu complementul său. Acest algoritm folosește o matrice de programare dinamică pentru a calcula cele mai bune punctaje de aliniere între două secvențe.**Alinierea locală (Smith-Waterman)**: Algoritmul Smith-Waterman se folosește pentru a identifica cele mai bune regiuni de similaritate între secvențele comparate. Acest algoritm este utilizat atunci când vrem să aliniem doar părțile relevante ale secvențelor, fără a forța o aliniere completă a întregii secvențe.

**- Modulul de analiză a frecvenței nucleotidelor. Divide și Stăpânește**: Folosește tehnica „divide and conquer” pentru a calcula frecvența nucleotidelor în secvența de ADN. Această abordare este eficientă pentru procesarea secvențelor mari și se bazează pe recursivitatea divizării secvenței în subsecvențe mai mici, care sunt analizate individual și combinate ulterior.

- **Modulul de căutare a motivelor AND. Căutare Greedy**: Sistemul folosește o abordare „greedy” pentru a căuta secvențe de motiv ADN de lungime fixă, în mai multe secvențe ADN. Algoritmul caută cele mai frecvente motive care se potrivesc în secvențele date și selectează cele mai bune din punct de vedere al scorului de similaritate.[8]

- **Interfața utilizatorului. Input de secvență ADN**: Utilizatorul poate introduce o secvență ADN și poate vizualiza rezultatele analizei.

**- Prezentarea rezultatelor.** Rezultatele sunt prezentate într-o formă accesibilă, care include informații despre complementul ADN-ului, alinierea secvențelor, analiza frecvenței nucleotidelor, și identificarea bolilor genetice, alături de o descriere detaliată a fiecărei boli identificate.

La nivel conceptual, fluxul de lucru al aplicației este următorul:Utilizatorul introduce o secvență AND, secvența este procesată pentru a genera complementul AND, secvența introdusă este comparată cu baza de date a bolilor genetice, algoritmii de aliniere (Needleman-Wunsch și Smith-Waterman) sunt utilizați pentru a compara secvențele de AND, se efectuează analiza frecvenței nucleotidelor folosind tehnica divide-and-conquer, algoritmul Greedy pentru căutarea motivelor ADN caută cele mai bune secvențe de motiv.

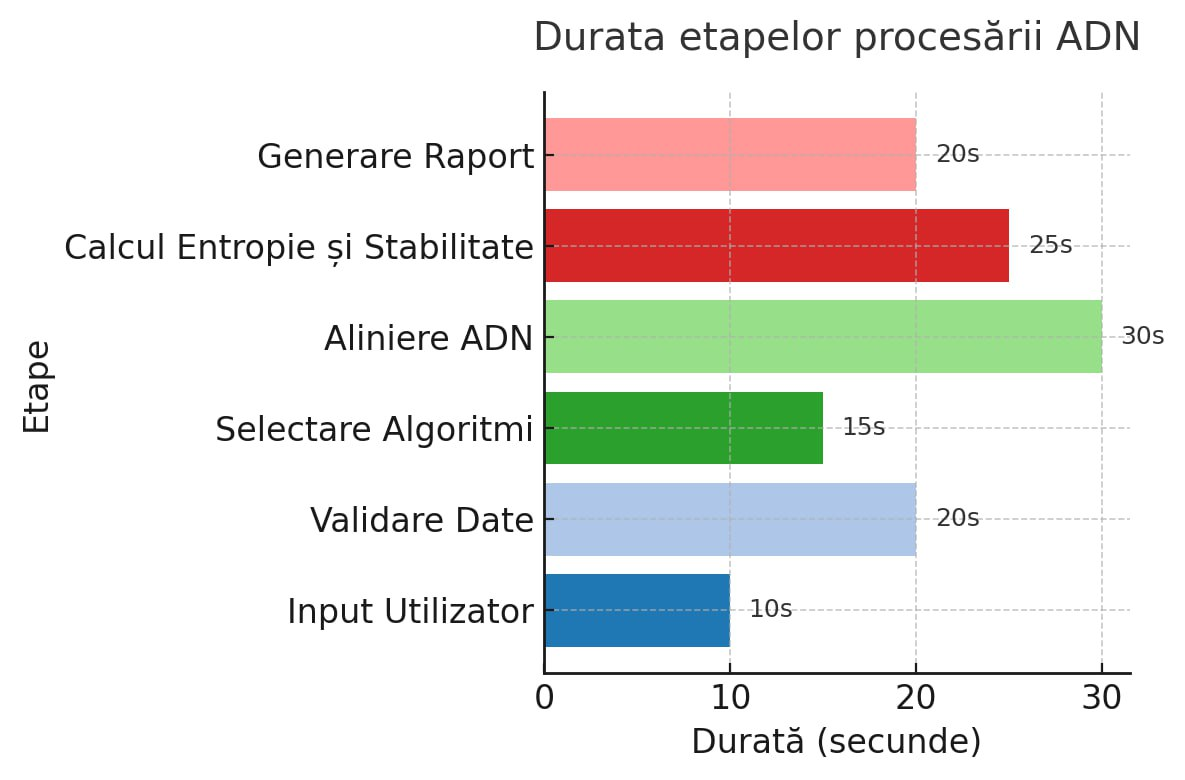
Această arhitectură permite utilizatorului să obțină o imagine completă asupra caracteristicilor secvenței ADN introduse, să identifice eventuale afecțiuni genetice și să efectueze analize detaliate ale secvenței la nivel molecular.

### **3.1.2 Modelarea vizuală a fluxurilor**

Modelarea vizuală a fluxurilor de date este crucială pentru a înțelege clar cum interacționează componentele aplicației în timpul procesului de analiză a secvenței ADN. Fluxul de date descrie cum sunt manipulate informațiile de la intrarea utilizatorului, trecând prin diverse etape de procesare (inclusiv validare, selecție de algoritmi și aliniere a secvențelor), până la generarea și prezentarea rezultatelor finale.

În acest sens, se va folosi o diagramă de flux pentru a vizualiza pașii principali ai procesului, interacțiunile dintre algoritmi, precum și fluxul de mesaje între diferitele componente ale sistemului, în special între interfața utilizatorului și modulele de procesare a datelor. Fiecare pas din fluxul de date este legat de o acțiune specifică, iar rezultatele obținute sunt transmise la etapele următoare ale analizei.[9]

Graficul din figura 3.2 ilustrează durata fiecărei etape din procesul de analiză ADN, începând cu introducerea datelor utilizatorului și continuând cu validarea, selectarea algoritmilor, alinierea ADN, calculul entropiei și stabilității, până la generarea raportului final. Acesta evidențiază pașii critici și timpul necesar fiecărei etape pentru o procesare eficientă.



### **Figura 3.2 - Modelarea vizuală a fluxurilor**

Utilizatorul introduce secvența ADN în interfața grafică a aplicației. Această acțiune declanșează procesul de validare a datelor și pregătirea acestora pentru procesare. Sistemul verifică validitatea secvenței ADN. Această verificare include asigurarea că lungimea secvenței este adecvată pentru procesare și că aceasta conține doar caractere valide ale ADN-ului: A, T, G, C. Dacă secvența nu este validă, utilizatorul primește un mesaj de eroare și i se solicită să introducă o secvență corectă.

După validarea secvenței, utilizatorul selectează algoritmii de aliniere. Aplicația oferă opțiuni precum

alinierea globală prin algoritmul Needleman-Wunsch sau alinierea locală prin algoritmul Smith-Waterman.

După selectarea algoritmilor, aceștia sunt aplicați pe secvențele introduse. În cazul alinierii globale, algoritmul Needleman-Wunsch aliniază întreaga secvență ADN cu complementul său pentru a obține cel mai bun scor de potrivire. În cazul alinierii locale, algoritmul Smith-Waterman identifică regiunile cu cea mai bună similaritate între secvențe. În continuare, se calculează entropia secvenței, care măsoară diversitatea nucleotidelor și variabilitatea acestora, precum și stabilitatea secvenței, evaluând proporția perechilor G-C pentru a determina stabilitatea structurii ADN.[10]

Rezultatele finale sunt afișate în interfața grafică a aplicației. Acestea includ numărul de mutații, care reprezintă variabilitatea dintre secvența introdusă și o secvență de referință, scorul de stabilitate calculat pe baza perechilor G-C și entropia secvenței, care indică diversitatea nucleotidelor. Rezultatele sunt prezentate sub forma unui raport detaliat care poate include grafice, tabele și explicații ale proceselor și algoritmilor utilizați.

### **3.1.3 Stările de tranzacție a sistemului**

### Sistemul va implementa un **model de mașină de stări finite (FSM)** pentru a descrie evoluția tranzacției de la primirea inputului până la generarea raportului final. Fiecare stare reprezintă o etapă importantă a procesului, iar tranzițiile între stări sunt determinate de acțiuni ale utilizatorului sau de rezultatele procesării. În acest sens, FSM va asigura că procesul de prelucrare al secvențelor ADN se face într-un mod organizat și logic.

Sistemul începe procesul în starea inițială, așteptând introducerea secvențelor ADN de către utilizator. Aceasta este prima interacțiune a utilizatorului cu aplicația, iar inputul este esențial pentru procesare. Utilizatorul poate introduce manual secvențele, încărca fișiere sau selecta formatele de input disponibile. Tranziția către următoarea stare are loc doar dacă secvențele introduse sunt valide. În cazul în care acestea sunt incomplete sau conțin erori de formatare, sistemul rămâne în această stare și solicită corectarea inputului.

După ce secvențele au fost introduse corect, sistemul trece în starea de validare, unde verifică formatul, completitudinea și conformitatea acestora cu regulile ADN. Verificările includ lungimea secvenței, absența erorilor și respectarea codificării standard. Dacă toate criteriile sunt îndeplinite, se face tranziția către starea de aliniere. În caz contrar, utilizatorul este notificat despre erorile detectate și trebuie să furnizeze date corecte înainte de a continua.

În starea de aliniere, sistemul aplică algoritmi pentru a compara secvențele ADN introduse. Alinierea poate fi globală, utilizând algoritmul Needleman-Wunsch, locală, folosind algoritmul Smith-Waterman, sau multiplu, pentru mai multe secvențe simultan. Acest proces identifică regiunile similare sau diferite între secvențe, oferind o bază pentru analize ulterioare. După finalizarea alinierii, sistemul trece în starea de calcul.

Starea de calcul implică analiza statistică a secvențelor ADN aliniate, prin calcularea entropiei și stabilității. Entropia reflectă diversitatea informației genetice din fiecare secvență, iar stabilitatea este evaluată pe baza variațiilor și a perechilor de nucleotide G-C. Aceste măsurători oferă informații despre diversitatea, evoluția și robustețea secvențelor. La finalizarea calculelor, sistemul trece în starea finală, unde sunt generate și afișate rezultatele.[11]

În starea finală, raportul de rezultate este prezentat utilizatorului. Acesta conține detalii despre secvențele ADN aliniate, valorile de entropie și stabilitate, procentul de similaritate și alte statistici relevante. Rezultatele pot fi vizualizate sub formă de grafice sau tabele și pot fi descărcate de utilizator. După această etapă, procesul poate fi reluat cu un nou set de date sau finalizat.

Dacă pe parcursul oricărei stări apare o problemă, sistemul intră în starea de eroare. Aceasta este destinată gestionării situațiilor în care apar date incorecte sau algoritmi care nu pot fi aplicați. În această stare, utilizatorului i se afișează un mesaj de eroare, oferindu-i indicații pentru remedierea problemei. Sistemul poate reveni la starea inițială sau poate oferi alte opțiuni pentru continuarea procesului.

**Tabelul 3.1 - Starile și tranzițiile**

|  |  |  |
| --- | --- | --- |
| **Stare curentă** | **Acțiune** | **Stare următoare** |
| **Stare inițială** | Utilizator introduce secvențele ADN | **Stare de validare** |
| **Stare de validare** | Date valide | **Stare de aliniere** |
| **Stare de validare** | Date invalide | **Stare inițială** (cu mesaj de eroare) |
| **Stare de aliniere** | Aliniere completă | **Stare de calcul** |
| **Stare de calcul** | Calcul finalizat | **Stare finală** |
| **Stare finală** | Generare raport | **Stare inițială** (pentru reluarea procesului) |

**3.1.4 Descrierea scenariilor de utilizare a aplicației**

Scenariile de utilizare a aplicației pentru procesarea secvențelor ADN sunt esențiale pentru a înțelege interacțiunea utilizatorilor cu sistemul și desfășurarea proceselor de analiză. Utilizatorul începe prin încărcarea secvențelor ADN fie într-un fișier, fie prin introducerea manuală în sistem. Aplicația validează datele, verificând dacă acestea sunt corecte și conforme cu formatele așteptate. În cazul unor erori, utilizatorul este notificat și trebuie să corecteze secvențele înainte de a trece la etapa următoare.

După validarea datelor, sistemul aplică algoritmi de aliniere, cum ar fi Needleman-Wunsch sau Smith-Waterman, pentru a compara și alinia secvențele ADN. Acest proces este esențial pentru identificarea regiunilor similare și a diferențelor dintre secvențe, constituind baza analizelor bioinformatice ulterioare.[12]

În etapa următoare, utilizatorul poate solicita calcularea entropiei și stabilității secvențelor ADN aliniate. Entropia oferă informații despre diversitatea informației genetice, în timp ce stabilitatea reflectă robustețea secvențelor, fiind evaluate pe baza variațiilor și a perechilor de nucleotide. Aceste analize sunt importante în cercetările genetice și studiile evolutive.

După finalizarea calculului, aplicația generează un raport cu rezultatele analizei. Raportul include detalii despre alinierea secvențelor, măsurile de entropie și stabilitate, precum și reprezentări vizuale, cum ar fi grafice sau tabele comparative. Utilizatorul poate descărca raportul sau îl poate vizualiza într-un format interactiv, finalizând astfel procesul de analiză.

**3.1.5 Fluxurile de mesaje și legăturile dintre componentele sistemului**

Fluxurile de mesaje și legăturile dintre componentele sistemului sunt esențiale pentru a asigura comunicarea corectă între părțile implicate în procesul de analiză al secvențelor ADN. Aceste fluxuri definesc cum datele sunt transmise între utilizator, aplicație, modulele de procesare și rezultatele finale. Iată o descriere detaliată a fluxurilor de mesaje și legăturilor dintre componentele principale ale sistemului:

**- Interacțiunea cu utilizatorul. Inputul utilizatorului**: Utilizatorul începe prin a introduce secvențele ADN în sistem printr-un fișier de tip text sau printr-un formular interactiv. Acest mesaj conține datele ADN care urmează să fie procesate. Fluxul de date între utilizator și aplicație începe cu trimiterea unui mesaj de tip "input ADN" din interfața utilizatorului către modulul de validare a datelor. Dacă secvențele ADN conțin erori, aplicația trimite mesaje de eroare către utilizator (de exemplu, "Datele sunt incorecte", "Formatul ADN nu este valid"), iar utilizatorul trebuie să corecteze intrările.[13]

- **Procesarea datelor AND.** După ce datele sunt primite, fluxul de mesaje se transmite de la interfața utilizatorului la modulul de validare a datelor. Modulul de validare analizează secvențele ADN pentru a verifica corectitudinea și conformitatea acestora cu formatele specifice. Dacă validarea este reușită, mesajul "validare reușită" este trimis la modulul de aliniere; dacă nu, un mesaj de eroare este trimis înapoi către utilizator. **Alinierea secvențelor**: După validare, fluxul de mesaje trimite datele către modulul de aliniere, care folosește algoritmi de aliniere pentru a compara secvențele ADN și a le organiza într-un format unificat. La finalul acestui proces, rezultatul aliniamentului (o matrice de aliniere) este trimis mai departe către modulul de calcul al entropiei și stabilității. **Calculul entropiei și stabilității**: Modulul de calcul primește datele aliniate și efectuează calculele statistice. După finalizare, rezultatele sunt trimise către modulul de generare a raportului.

**- Generarea și vizualizarea rezultatelor.** După efectuarea calculului, fluxul de mesaje transmite datele procesate către modulul de generare a raportului, care creează un document structurat ce conține rezultatele analizei. Acest raport poate include grafice vizuale, statistici și tabele, care sunt trimise înapoi către interfața utilizatorului pentru vizualizare sau descărcare. După generarea raportului, un mesaj final este trimis utilizatorului cu detalii despre rezultate, oferindu-i posibilitatea de a descărca raportul sau de a vizualiza informațiile în format interactiv. **Legăturile dintre componentele sistemului** sunt realizate prin intermediul unor interfețe bine definite (APIs sau module de comunicare internă), care permit schimbul de informații între fiecare parte a aplicației. Aceste componente sunt interconectate într-un flux logic, asigurându-se că fiecare etapă de procesare este completă înainte de a trece la următoarea.

**3.2Descrierea structurii statice a sistemului**

Structura statică a sistemului descrie modul în care sunt organizate și interconectate componentele și modulele acestuia, fără a ține cont de fluxurile dinamice de mesaje și interacțiunile de la un moment dat. Aceasta se concentrează pe definirea arhitecturii la un nivel înalt, clarificând cum sunt dispuse componentele fundamentale și relațiile dintre ele pe parcursul vieții sistemului.

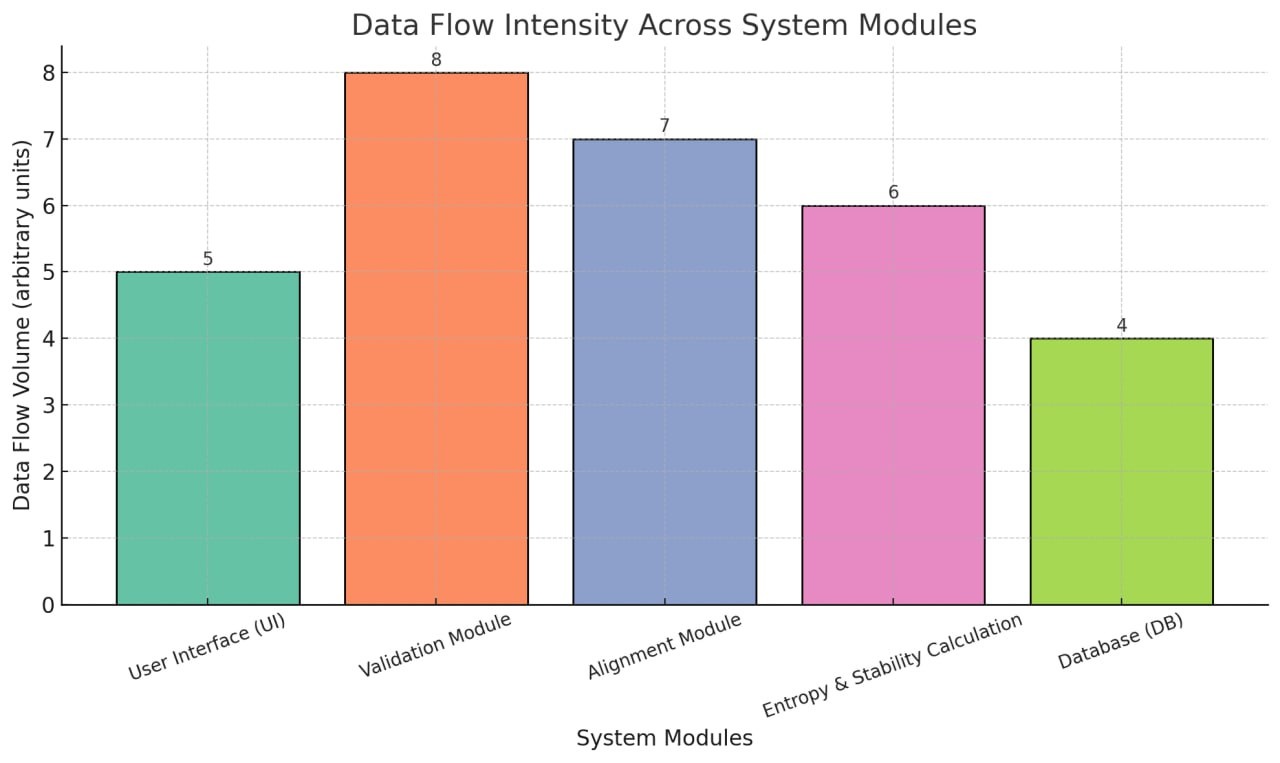
**- Interfața Utilizatorului (UI).** Modulul de Interfață Utilizator (UI) este responsabil pentru interacțiunea directă a utilizatorului cu aplicația. Acesta oferă un mediu vizual și intuitiv în care utilizatorii pot introduce secvențele ADN necesare pentru procesare și pot vizualiza rezultatele obținute în urma calculelor. UI-ul include elemente grafice care facilitează navigarea și înțelegerea fluxului de lucru al aplicației. Utilizatorul poate să inițieze procesul de aliniere ADN, să valideze datele introduse și să acceseze rapoartele generate într-un format clar și ușor de înțeles. De asemenea, UI-ul permite feedback imediat asupra erorilor sau corectitudinii datelor introduse, ajutând utilizatorul să corecteze eventualele greșeli înainte de a avansa la etapele ulterioare.[14]

**- Modulul de Validare a Datelor.** Modulul de Validare a Datelor este responsabil pentru asigurarea corectitudinii și integrității secvențelor ADN introduse de utilizator. În această etapă, datele sunt verificate pentru a se asigura că respectă formatele necesare și nu conțin erori care ar putea afecta rezultatele procesării. Acest modul efectuează verificări de sintaxă și validări pe baza unor reguli prestabilite, cum ar fi lungimea secvențelor, prezența caracterelor incorecte sau lipsa datelor esențiale. De asemenea, poate identifica secvențele ADN care sunt incomplete sau greșit formate și poate oferi utilizatorului informații detaliate despre erorile detectate, permițând astfel corectarea lor înainte de a trece la procesul de aliniere.

**- Modulul de Aliniere AND.** Modulul de Aliniere ADN este responsabil pentru procesul de aliniere al secvențelor ADN, esențial pentru compararea acestora în scopul identificării similarităților sau diferențelor. Acest modul utilizează algoritmi de aliniere, precum alinierea globală sau locală, pentru a potrivi secvențele într-un mod care să maximizeze asemănările între ele. Algoritmii pot fi adaptați în funcție de cerințele aplicației, iar selecția acestora depinde de specificul secvențelor ADN analizate. Alinierea corectă a secvențelor ADN este crucială pentru etapa următoare de calcul, deoarece erorile de aliniere pot afecta acuratețea rezultatelor finale.

**- Modulul de Calcul Entropie și Stabilitate.** După realizarea alinierii, secvențele aliniate sunt procesate de modulul de calcul pentru a determina entropia și stabilitatea acestora. Entropia este o măsură a diversității secvențelor, iar stabilitatea reflectă capacitatea acestora de a menține structuri constante în cadrul analizei. Aceste metrici sunt esențiale pentru a evalua variabilitatea genetică a secvențelor ADN și pentru a înțelege comportamentele acestora într-un context mai larg. Modulul folosește formule matematice și statistice pentru a calcula aceste valori și pentru a le interpreta în funcție de scopul analizei ADN, oferind informații valoroase pentru cercetători și specialiști în domeniu.[15], [16]

**- Baza de Date (DB).** Baza de Date (DB) este un sistem de gestionare a bazelor de date care stochează informațiile și rezultatele obținute în urma proceselor de aliniere și calcul ale secvențelor ADN. Aceasta asigură persistența datelor, permițând stocarea pe termen lung a tuturor datelor de intrare și ieșire generate de aplicație. Baza de date permite accesul rapid la informațiile necesare pentru validare, analiză și generarea rapoartelor, iar aceasta poate fi structurat în tabele dedicate pentru stocarea secvențelor ADN, rezultatelor calculelor și istoricului proceselor. Graficul din figura 3.3 ilustrează volumul fluxului de date procesat de fiecare modul al sistemului, evidențiind intensitatea operațiunilor din etapele cheie: validarea datelor, alinierea ADN, calculul entropiei și generarea raportului final. Culorile distincte facilitează identificarea fiecărui modul.



**Figura 3.3 - Structura sistemului**

Relațiile de dependență dintre elementele sistemului reflectă modul în care componentele interacționează și depind unele de altele pentru a asigura buna funcționare a proceselor. În cadrul acestui sistem, fiecare modul joacă un rol esențial și interacționează cu celelalte module pentru a îndeplini sarcinile specifice.

Interfața Utilizatorului (UI) depinde de Modulul de Validare a Datelor, deoarece utilizatorul furnizează secvențele ADN prin intermediul UI, iar aceste date trebuie validate înainte de a putea continua procesul. Fără validarea corectă a datelor, sistemul nu poate trece la pașii următori. Modulul de Validare a Datelor depinde de Interfața Utilizatorului (UI) pentru a obține datele introduse de utilizator. De asemenea, acest modul este conectat la Baza de Date (DB) pentru a verifica integritatea datelor în raport cu informațiile deja stocate. Validarea corectă a datelor este crucială pentru evitarea erorilor de procesare în etapele următoare.

Modulul de Aliniere ADN depinde de Modulul de Validare a Datelor, deoarece procesul de aliniere nu poate începe decât după ce datele au fost validate. De asemenea, acest modul interacționează cu Baza de Date (DB) pentru a salva secvențele ADN aliniate și pentru a accesa informațiile istorice necesare pentru comparații. Modulul de Calcul Entropie și Stabilitate depinde de Modulul de Aliniere ADN, deoarece calculele pentru entropie și stabilitate se bazează pe secvențele ADN aliniate. Fără alinierea corectă a secvențelor, aceste calcule nu ar fi valide.

Modulul de Generare a Raportului depinde de Modulul de Calcul Entropie și Stabilitate, deoarece raportul final va include valorile calculate pentru entropie și stabilitate. De asemenea, acest modul poate interacționa cu Baza de Date (DB) pentru a salva raportul generat și pentru a-l face accesibil utilizatorilor. Baza de Date (DB) joacă un rol central în întreaga structură a sistemului. Aceasta asigură persistenta datelor la fiecare pas: stochează secvențele ADN, rezultatele validării, secvențele aliniate, calculele de entropie și stabilitate, și rapoartele finale. De asemenea, baza de date poate fi accesată de toate celelalte module pentru a salva și a recupera datele necesare.

Astfel, relațiile de dependență sunt esențiale pentru fluxul corect al proceselor din sistem, fiecare componentă având un rol bine definit și interacționând cu celelalte pentru a asigura finalizarea cu succes a procesului de analiză a secvențelor ADN. Fiecare modul depinde de ieșirea corectă a modulului anterior pentru a continua procesul și a produce rezultatele finale.

Tabelul 1.2 ilustrează componentele principale ale sistemului de procesare ADN, rolurile acestora și fluxurile de mesaje între ele. Fiecare componentă contribuie la realizarea etapelor specifice ale procesului: introducerea datelor, validarea, alinierea, analiza statistică și generarea raportului. Fluxurile de mesaje descriu modul în care informațiile sunt transmise și procesate pentru a asigura o comunicare eficientă între părți. Această structură asigură coerența și completitudinea procesului.

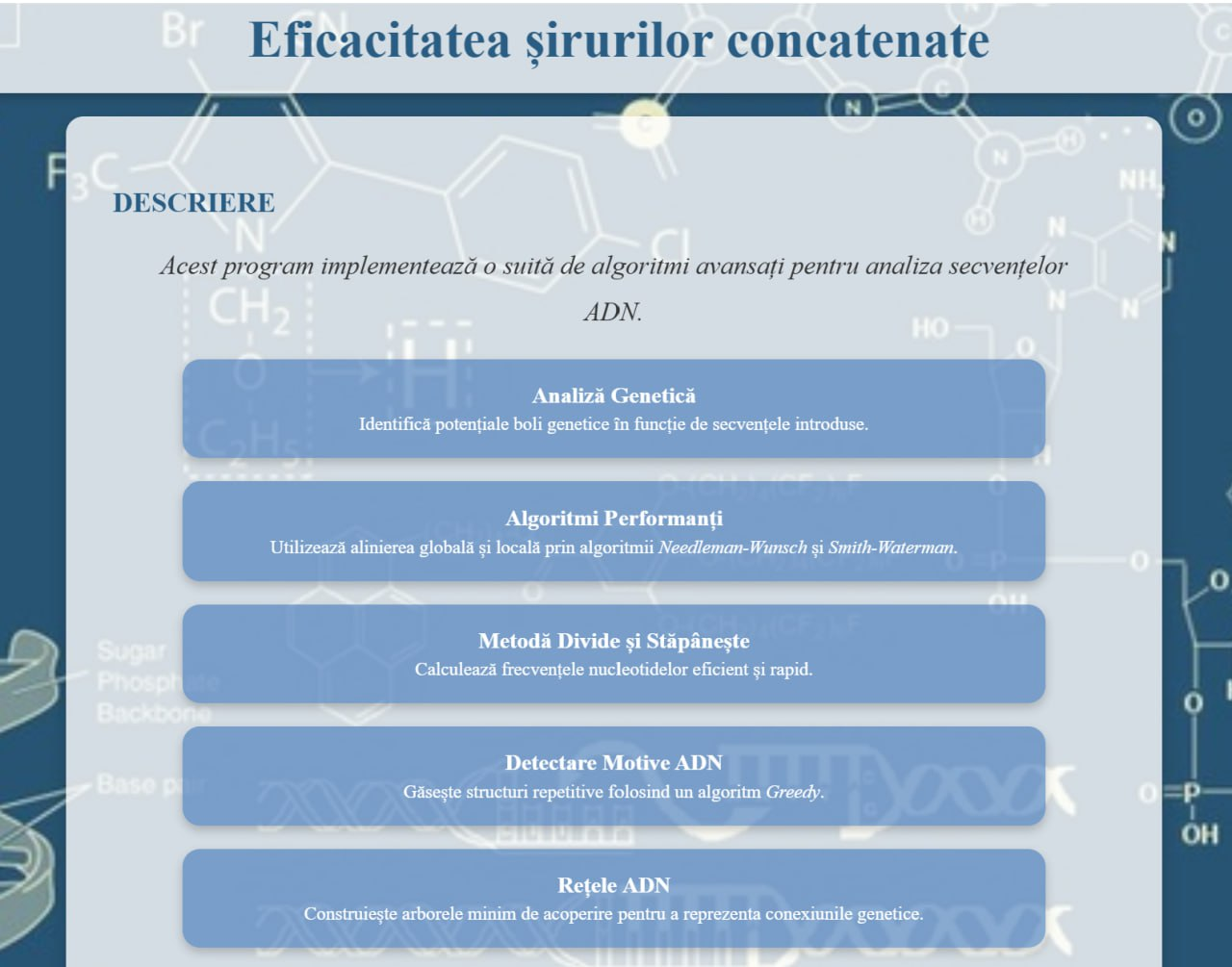
**Tabelul 3.2 - Fluxuri in Sistem ADN**

|  |  |  |
| --- | --- | --- |
| **Componentă** | **Rol** | **Flux de mesaje** |
| Interfața Utilizatorului | Permite utilizatorului să introducă secvențele ADN și să vizualizeze rezultatele analizei. | Utilizatorul trimite secvențe ADN → Primește feedback (erori sau succes) → Vizualizează raportul generat. |
| Modulul de Validare a Datelor | Verifică corectitudinea și integritatea secvențelor ADN introduse de utilizator. | Primește date de la interfață → Verifică validitatea → Trimite mesaje de eroare sau succes către interfață și modulul de aliniere. |
| Modulul de Aliniere | Aplică algoritmi pentru compararea și organizarea secvențelor ADN. | Primește date validate → Efectuează alinierea secvențelor → Trimite matricea aliniată către modulul de calcul. |
| Modulul de Calcul | Realizează analize statistice, cum ar fi calculul entropiei și stabilității secvențelor. | Primește matricea aliniată → Efectuează calcule statistice → Trimite rezultatele către modulul de generare a raportului. |
| Modulul de Generare a Raportului | Creează un raport structurat cu rezultatele analizei, inclusiv tabele, grafice și statistici. | Primește rezultatele analizei → Formatează raportul → Trimite raportul final către interfață pentru vizualizare sau descărcare. |

Modelarea echipamentelor mediului de intrare presupune definirea și descrierea dispozitivelor care furnizează datele de intrare pentru sistemul de procesare a secvențelor ADN. Aceste echipamente pot include calculatoare desktop, laptopuri, servere sau chiar dispozitive mobile care permit utilizatorilor să introducă secvențele ADN în aplicație. De asemenea, ar putea include aplicații externe care pot interacționa cu sistemul, precum baze de date biologice sau servicii web care furnizează secvențe ADN deja preprocesate. Echipamentele din mediul de intrare sunt conectate la sistemul principal prin intermediul unor interfețe de comunicație, fie prin rețele locale, fie prin rețele de internet, asigurându-se astfel transferul corect și rapid al datelor. În funcție de natura echipamentelor, aceste date pot fi preluate direct din fișiere (de exemplu, fișiere FASTA sau FASTQ) sau prin interacțiune directă cu utilizatorul, care le introduce prin interfața grafică.

**4 INTERFAȚA**

Figura 4.1 prezintă structura principală a interfeței „Eficacitatea șirurilor concatenate”. Elementele principale includ un titlu animat și o secțiune de descriere, care evidențiază funcțiile principale ale aplicației, cum ar fi analiza genetică și alinierea ADN.



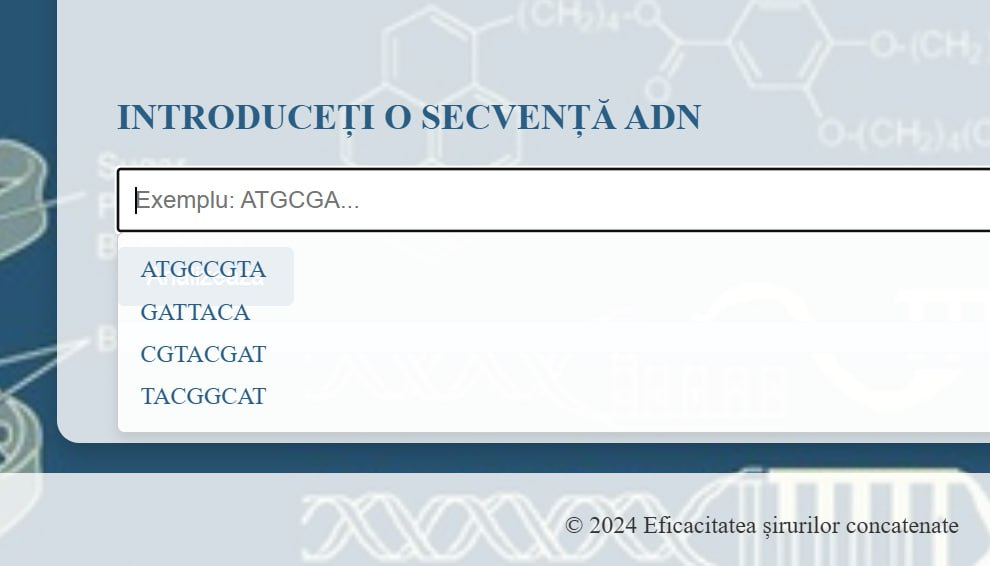
**Figura 4.1 - Pagina principală**

In figura 4.2 este ilustrată secțiunea dedicată introducerii secvenței ADN de către utilizator. Exemplul sugerat (ATGCGA...) oferă ghidaj vizual, iar utilizatorul poate accesa lista de secvențe ADN predefinite prin focalizarea câmpului de introducere.



**Figura 4.2 - Secțiunea de introducere a secvențelor AND**

Figura 4.3 evidențiază meniul de sugestii care apare automat atunci când utilizatorul începe să introducă sau selectează câmpul ADN. Utilizatorul poate face clic pe una dintre opțiunile disponibile pentru a completa rapid o secvență validă.



**Figura 4.3 - Selectarea exemplelor de secvențe AND**

În figura 4.4, sunt afișate rezultatele generate după analizarea unei secvențe ADN. Datele includ boli genetice detectate, alinierea globală și locală, frecvențele nucleotidelor, motivele ADN identificate și detalii despre arborele minim de acoperire generat prin algoritmul Kruskal.



**Figura 4.4 - Rezultatele analizei ADN**

**CONCLUZII**

Tema proiectului propus, referitoare la analiza secvențelor ADN și aplicarea tehnicilor de aliniere și evaluare a diversității genetice, are o semnificație deosebită în domeniul bioinformaticii și al cercetării genetice. În contextul în care secvențele ADN devin tot mai accesibile și utile în medicina personalizată și în diagnosticarea bolilor genetice, importanța unei analize precise și rapide a acestora este mai evidentă ca niciodată. Proiectul propune contribuții semnificative la dezvoltarea unor instrumente algoritmice eficiente, care să permită analiza detaliată a ADN-ului și să sprijine cercetările fundamentale în domeniul geneticii, precum și studiile aplicative în domeniul medical.

Prin implementarea și compararea algoritmilor Needleman-Wunsch și Smith-Waterman pentru alinierea globală și locală, proiectul își propune nu doar aprofundarea înțelegerii metodelor de aliniere, dar și evaluarea eficienței acestora în aplicarea lor în diverse contexte bioinformatice. Un alt obiectiv esențial al proiectului este calculul entropiei Shannon pentru evaluarea diversității secvențelor ADN, precum și determinarea stabilității acestora pe baza perechilor G-C. Aceste abordări vor oferi perspective valoroase asupra structurii și rezistenței ADN-ului, facilitând predicția comportamentului genetic al organismelor și înțelegerea evoluției acestora.

De asemenea, un aspect semnificativ al proiectului este analiza mutațiilor genetice și a potențialului acestora de a influența predispozițiile la afecțiuni genetice. Acest obiectiv nu doar că abordează o problemă fundamentală în diagnosticarea timpurie a bolilor, dar contribuie și la dezvoltarea unor strategii de prevenție, un element important al cercetărilor biomedicale moderne.

Implementarea acestui proiect într-un cadru academic sau profesional poate sprijini progrese importante în domeniul bioinformaticii, atât în ceea ce privește eficiența și precizia analizelor genetice, cât și în dezvoltarea unor instrumente scalabile care să poată fi utilizate în laboratoare de cercetare genetică. Rezultatele obținute pot fi folosite ca fundament pentru studii evoluționare, identificarea markerilor genetici sau în cercetarea personalizată a terapiilor pentru diverse afecțiuni genetice.

În ansamblu, proiectul propus aduce o contribuție semnificativă atât în domeniul cercetării fundamentale, cât și în cel aplicativ, având implicații importante în medicina personalizată, în dezvoltarea de noi tratamente pentru bolile genetice și în aprofundarea cunoașterii despre genomurile viețuitoarelor. Acesta demonstrează potențialul bioinformaticii de a transforma domeniul sănătății prin instrumente inovative care pot revoluționa modul în care sunt diagnosticate și tratate afecțiunile genetice.

**BIBLIOGRAFIE**

[1] Bioinformatics: Sequence and Genome Analysis de David W. Mount

[2] Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids de Richard

Durbin, Sean R. Eddy, Anders Krogh și Graeme Mitchison

1. Nucleic Acids Research  
   [4] Bioinformatics (Oxford Academic)  
   [5] UCSC Genome Browser(https://genome.ucsc.edu/)  
   [6] **Geneious** (<https://www.geneious.com/>)  
   [7] **Bioinformatics: Sequence Alignment and Analysis** de Arthur M. Lesk  
   [8] **Algorithms on Strings, Trees, and Sequences** de Dan Gusfield  
   [9] **Programming for Bioinformatics** de P. P. K. Vasantharaj David  
   [10] **Pattern Recognition and Machine Learning** de Christopher M. Bishop  
   [11]WunsAlgorithmhttps://bio.libretexts.org/Bookshelves/Computational\_Biology/Book%3A\_Computational\_Biology\_Genomes\_Networks\_and\_Evolution\_%28Kellis\_et\_al.%29/02%3A\_Sequence\_Alignment\_and\_Dynamic\_Programming/2.05%3A\_The\_Needleman-Wunsch\_Algorithm

[12 ] https://www.public.asu.edu/~cbaral/cse591-s03/classnotes/seq-align.pdf  
[13] <https://www.kaell.se/bibook/pairwise/waterman.html>

[14] http://elf.cs.pub.ro/sda-ab/wiki/laboratoare/laborator-07