I.5

BAZA DE DATE INTEGRATĂ NCBI. PROGRAMUL BLAST

În aceast material ne vom familiariza cu programul BLAST. Prin intermediul lui vom face căutări in bazele de date după secvențe de nucleotide şi aminoacizi.

1. Introducere

Programul BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) este un program popular de comparare a secvenţelor de ADN, ce face parte dintr-un pachet de programe destinat căutării de secvențe proteice, accesibil în diverse forme la diferiţi furnizori, sau prin intermediul NCBI, care mai oferă şi Entrez, un instrument de meta-căutare care acoperă mare parte a bazelor de date de la NCBI, inclusiv cele care găzduiesc structuri tridimensionale a proteinelor, genoamele complete ale organismelor şi trimiteri la jurnale ştiinţifice care însoţesc intrările din bazele de date.

Asocierea dezvoltărilor tehnologiei de calcul şi moleculare deschide noi oportunităţi cercetărilor genetice. Folosirea combinată a informaţiei oferită de secvenţe, a instrumentelor de calcul, a bazelor de date şi a biologiei tradiţionale creşte speranţa înţelegerii funcţiei şi reglajelor tuturor genelor şi proteinelor, precum şi a descifrării funcţiilor celulei.

2. Prezentarea programului BLAST

BLAST reprezintă instrumentul de căutare a aliniamentului local de bază, fiind un set de programe de căutare a similarităţilor, creat pentru identificarea clasificării şi a omologilor potenţiali pentru o secvenţă dată.

Pentru a înţelege mai bine programele BLAST, trebuie cunoscute aspectele de bază ale aliniamentelor secvenţelor. Acestea sunt folosite în special pentru găsirea potenţialilor omologi ce vor fi folosiţi ulterior pentru prezicerea posibilelor funcţii ale secvenţei necunoscute sau pentru modelarea structurii sale tridimensionale.

Programele BLAST folosesc un algoritm heuristic care identifică aliniamentele locale, găsind omologii cu secvenţele cele mai apropiate, într-un timp eficient.

Serverul BLAST suportă o varietate de programe analitice care sunt fie accesate prin reţeaua Internet, fie instalate în reţele locale pentru a mări viteza de analiză. Programul BLAST bazal nu permite introducerea gap-urilor în aliniamentele sale ceea ce va reduce senzitivitatea căutării. Cu toate acestea, datele de ieşire din program oferă aliniamente regionale multiple, care pot fi folosite pentru a anticipa gap-urile din secvenţa de interes şi cea din baza de date. În continuare sunt enumerate programele BLAST şi utilizarea lor.

a) **BLASTp**: acest program permite utilizatorului să caute similarităţile dintre secvenţa unei proteine necunoscute şi secvenţele proteinelor dintr-o bază de date.
b) **BLASTx**: permite compararea secvenţelor traduse în aminoacizi ale nucleotidelor cu secvenţele proteinelor din bazele de date.
   Secvenţa nucleotidică de interes este tradusă iniţial în toate cele 6 catene de citire **ORF** (**O**pen **R**eading **F**rame) posibile. Acest program este folositor în special pentru găsirea erorilor de secvenţializare a nucleotidelor, prin compararea secvenţei de nucleotide tradusă în aminoacizii săi proteici potenţiali dintr-o bază de date cu secvenţe proteice.
c) **BLASTn**: cu ajutorul acestui program se compară o secvenţă nucleotidică de interes cu secvenţele din bazele de date nucleotidice.
d) **tBLASTn**: permite căutarea similarităţilor dintre o secvenţă proteică şi secvenţele traduse (translatate) ale nucleotidelor dintr-o bază de date.

Secvenţele nucleotidice dintr-o bază de date sunt traduse iniţial în fiecare din cele 6 catene de citire posibile şi sunt apoi comparate cu secvenţa proteinei de interes. Acest program este util pentru găsirea erorilor de secvenţializare în proteine prin compararea secvenţei proteinei respective cu omologii săi potenţiali obţinuţi prin traducerea secvenţelor nucleotidice dintr-o bază de date.

La adresa https://blast.ncbi.nlm.nih.gov/Blast.cgi se găsesc opţiunile programului BLAST:

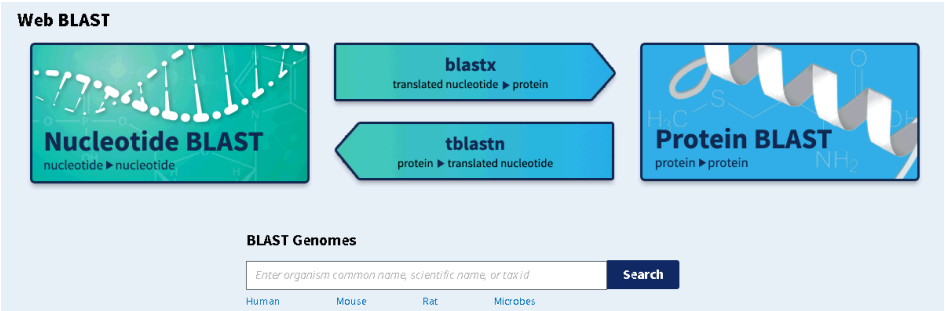

Figura 1. Interfaţa BLAST

3. Exemplu de căutare

Dacă dorim să căutăm secvenţa de gene care codează receptorul pentru endotelină la om (figura 2) introducem în căsuţa search următoarele: **Endothelin receptor**.

Din mulţimea de răspunsuri vom selecta varianta receptorului pentru specia umană (homo sapiens). Va fi afişat cromozomul pe care se află secvenţa de nucleotide (cromozomul 40) – figura 3, localizarea genei pe cromozom, secvenţa de nucleotide şi secvenţa codantă.

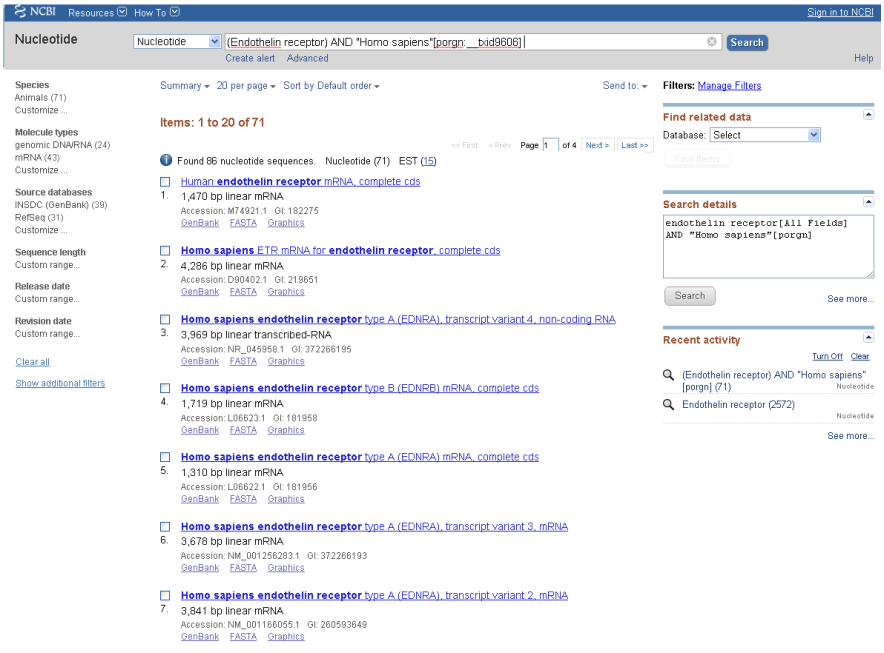

Figura 2. Exemplu de căutare a unei secvenţe de nucleotide

```
FEATURES             Location/Qualifiers
     source          1..3841
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="4"
                     /map="4q31.22-q31.23"
     gene            1..3841
                     /gene="EDNRA"
                     /gene_synonym="ET-A; ETA; ETA-R; ETAR; ETRA; hET-AR; MFDA"
                     /note="endothelin receptor type A"
                     /db_xref="GeneID:1909"
                     /db_xref="HGNC:HGNC:3179"
                     /db_xref="MIM:131243"
     exon            1..460
                     /gene="EDNRA"
                     /gene_synonym="ET-A; ETA; ETA-R; ETAR; ETRA; hET-AR; MFDA"
                     /inference="alignment:Splign:1.39.8"
     misc_feature    30
                     /gene="EDNRA"
                     /gene_synonym="ET-A; ETA; ETA-R; ETAR; ETRA; hET-AR; MFDA"
                     /note="alternative transcription initiation site"
     STS             46..236
                     /gene="EDNRA"
                     /gene_synonym="ET-A; ETA; ETA-R; ETAR; ETRA; hET-AR; MFDA"
                     /standard_name="SHGC-67921"
                     /db_xref="UniSTS:38684"
     exon            461..950
                     /gene="EDNRA"
                     /gene_synonym="ET-A; ETA; ETA-R; ETAR; ETRA; hET-AR; MFDA"
                     /inference="alignment:Splign:1.39.8"
     misc_feature    501..503
                     /gene="EDNRA"
                     /gene_synonym="ET-A; ETA; ETA-R; ETAR; ETRA; hET-AR; MFDA"
                     /note="upstream in-frame stop codon"
     CDS             531..1487
                     /gene="EDNRA"
                     /gene_synonym="ET-A; ETA; ETA-R; ETAR; ETRA; hET-AR; MFDA"
                     /note="isoform b precursor is encoded by transcript
                     variant 2; endothelin receptor subtype A;
```

Datele despre secventa de nucleotide cromozomul, bratul cromozomului, gena

Figura 3. Rezultatele căutării

Ca secvență de nucleotide și aminoacizi, găsim următoarea formulare (fig. 4).

Figura 4. Formatul BLAST de vizualizare

4. Exerciții practice

NOTA: pentru secvențele-text accesați fișierul *Bioinfo-2018-practic.txt.*

Exerciţiul 1

National Center for Biotechnology Information:
https://www.ncbi.nlm.nih.gov/
BLAST - Basic Local Alignment Search Tool
Nucleotide BLAST - Enter Query Sequence:

1 ctagaaaccg tatgctatat aattatatac tataaagtaa taatgtatac agtgtaatgg
61 atcatgggcc atgtgctttt caaactaatt gtacataaaa caagcatcta ttgaaaatat
121 ctgacaaact catcttttat ttttgatgtg tgtgtgtgtg tgtgtgtgtg ttttttaac
181 aggggatttgg gg

(căutarea este comandată cu butonul BLAST)

Observați scorul de potrivire - investigați/căutați informațiile referitoare la gena Homo sapiens Human cystic fibrosis transmembrane conductance regulator (CFTR), Locus AH002646.

Observați adnotările/referințele bibliografice – coloana din dreapta.

Exerciţiul 2

Căutați proteina ***insulina.***

Soluţii

National Center for Biotechnology Inform ation: https://www.ncbi.nlm.nih.gov/

Protein - Search for "insulin" - alegeți "insulin[Homo sapiens]"

Observați adnotările/referințele bibliografice – coloana din dreapta
Observați secvența de aminoacizi ai proteinei – ORIGIN:

1 malwmrllpl lallalwgpd paaafvnqhl cgshlvealy lvcgergffy tpktrreaed
61 lqvgqvelgg gpgagslqpl alegslqkrg iveqcctsic slyqlenycn

```
     Site            order(32,35,37,108)
                     /site_type="other"
                     /note="putative receptor binding surface"
                     /db_xref="CDD:239833"
     CDS             1..110
                     /gene="INS"
                     /coded_by="join(AH002844.2:2424..2610,
                     AH002844.2:3397..3542)"
                     /note="precursor"
                     /db_xref="GDB:G00-119-349"
ORIGIN
        1 malwmrllpl lallalwgpd paaafvnqhl cgshlvealy lvcgergffy tpktrreaed
       61 lqvgqvelgg gpgagslqpl alegslqkrg iveqcctsic slyqlenycn
//
```

Figura 5. Secvenţa de aminoacizi ai proteinei insulina umană

Exerciţiul 3

Căutaţi după secvenţa de aminoacizi **malwmrllpl.**

Soluţii

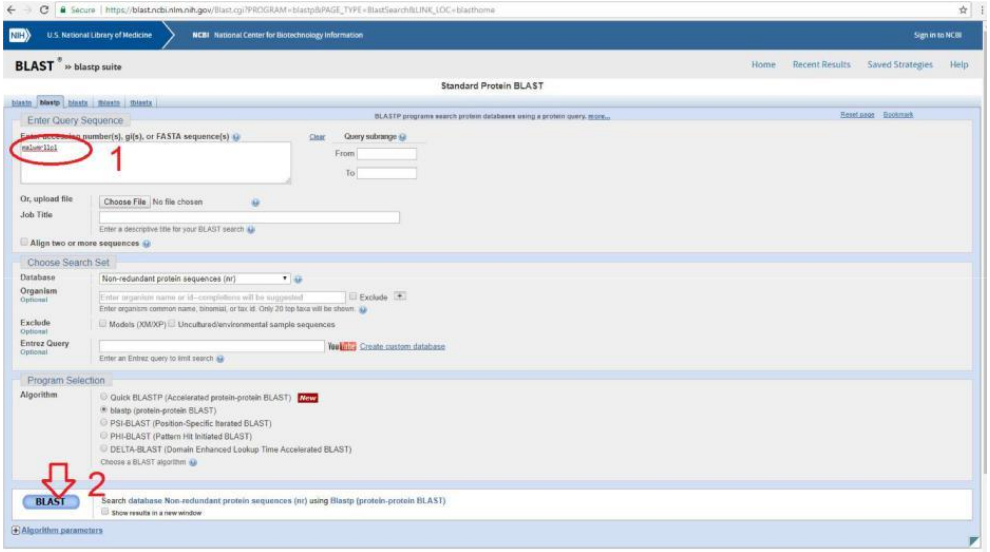

Figura 6. Interogarea BD după o secvenţă de aminoacizi

Observaţi scorul de potrivire....

Exerciţiul 4

Căutaţi secvenţa de aminoacizi:

1 mqngagasrt stiflngnre rplnvfcdme tdgggwlvfq rrmdgqtdfw rdwedyahgf
61 gnisgefwlg nealhsltqa gdysirvdlr agdeavfaqy dsfhvdsaae yyrlhlegyh
121 gtagdsmsyh sgsvfsardr dpnsllisca vsyrgawwyr nchyanlngl ygstvdhqgv
181 swyhwkgfef svpftemklr prnfrspagg g

Şi vizualizaţi structura 3D.

Soluţii

National Center for Biotechnology Information:
https://www.ncbi.nlm.nih.gov/

Protein BLAST - Enter Query Sequence….(copy/paste secvenţă)

Pentru *fibrinogen [Homo sapiens]* accesaţi structura 3D.

Figura 7. Vizualizarea 3D a structurii fibrinogenului uman

Exercițiul 5

Căutați informații avansate despre fibrinogenul uman.

Soluții

National Center for Biotechnology Information:
https://www.ncbi.nlm.nih.gov/

Protein - Advanced Search for "human fibrinogen".

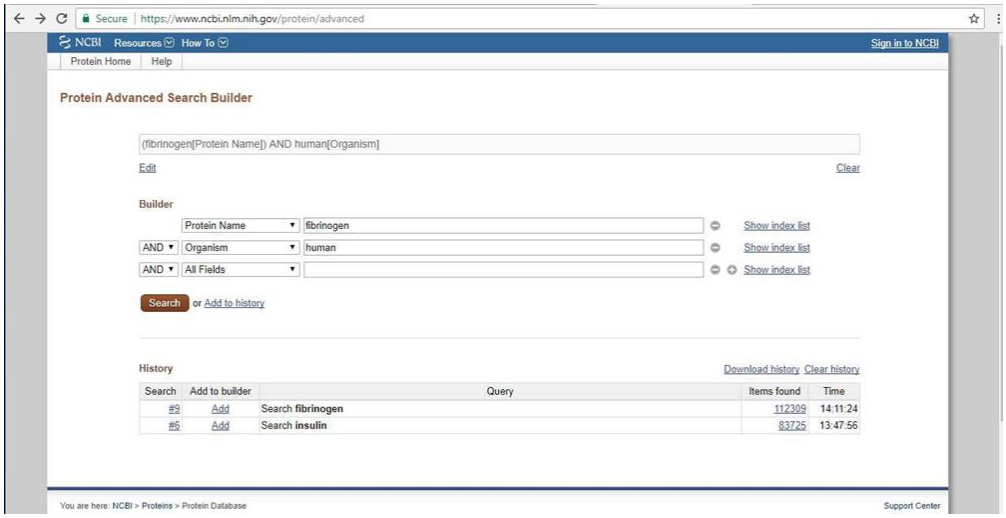

Figura 8. Filtru avansat de căutare

Cautați informații și pe Wikipedia:
https://en.wikipedia.org/wiki/Fibrinogen

Exercițiul 6

Căutați secvența:

1 mkwvtfisll flfssaysrg vfrrdahkse vahrfkdlge enfkalvlia faqylqqcpf
61 edhvklvnev tefaktcvad esaencdksl htlfgdklct vatlretyge madccakqep
121 ernecflqhk ddnpnlprlv rpevdvmcta fhdneetflk kylyeiarrh pyfyapellf
181 fakrykaaft eccqaadkaa cllpkldelr degkassakq rlkcaslqkf gerafkawav

241 arlsqrfpka efaevsklvt dltkvhtecc hgdllecadd radlakyice nqdsissklk
301 eccekpllek shciaevend empadlpsla adfveskdvc knyaeakdvf lgmflyeyar
361 rhpdysvvll lrlaktyett lekccaaadp hecyakvfde fkplveepqn likqncelfe
421 qlgeykfqna llvrytkkvp qvstptlvev srnlgkvgsk cckhpeakrm pcaedylsvv
481 lnqlcvlhek tpvsdrvtkc cteslvnrrp cfsalevdet yvpkefnaet ftfhadictl
541 sekerqikkq talvelvkhk pkatkeqlka vmddfaafve kcckaddket cfaeegkklv
601 aasqaalgl

Observați și analizați rezultatele.

Accesați opțiunea "Analize your query with SmartBLAST".

5. Concluzii

Din acest material am învațat să facem:

- căutări în bazele de date după secvenţe de gene;
- determinăm cromozomul pe care se află o secvenţă de nucleotide;
- căutări în bazele de date după secvenţe de nucleotide;
- căutare în BD după secvenţa de aminoacizi;
- găsim gradul de potrivire al secvenţei de amioacizi introdusă.

Referințe

Gheorghe-Ioan Mihalas, Anca Tudor, Sorin Paralescu – Bioinformatica, Colectia “Stiintele exacte in cercetarea medicala”, Timisoara: Editura Victor Babes, 2011, ISBN 978-606-8054-33-9

https://www.ncbi.nlm.nih.gov/