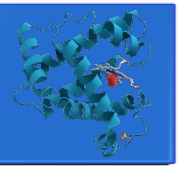
2.04.2020

Curs 7 – Stabilirea funcției unei proteine. Similaritate și omologie la nivel de secvență

Legătura secvență –funcție



Folosind metodele experimentale de secvențiere enumerate anterior, se stabilește secvența unei gene. Aceasta codifică următoarea secvență de aminoacizi:

MAAKYRIGYFVGSLATGSINRVLSQALINLAPEDLEFSEIPIRDLPLYSYDYDADFPPEGR

Care este funcția acestei peptide și implicit a genei codificatoare?

Cunoaşterea secvenţei de nucleotide a unui fragment de ADN și implici a secvenței de aminoacizi a unei proteine nu înseamnă obligatoriu şi cunoaşterea rolului (funcţiei) moleculei respective.

Și totuși, secvența de aminoacizi este cea ce coordonează structura tridimensională a peptidei și deci reacția enzimatică/funcția pe care peptida o are/realizează.

Secvență de nucleotide → **Codul genetic** → Secvență de aminoacizi → **Al II-lea cod genetic** → Funcția proteinei codificate

Manual sau cu metode bioinformatice

Numai cu metode bioinformatice

JOURNAL OF BACTERIOLOGY, May 2006, p. 3431–3432
0021-9193/06/$08.00+0 doi:10.1128/JB.188.10.3431–3432.2006
Copyright © 2006, American Society for Microbiology. All Rights Reserved.

Vol. 188, No. 10

GUEST COMMENTARY

The Difficult Road from Sequence to Function

Robert H. White*

Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0308

Makrythanasis and Antonarakis *Genome Medicine* 2011, 3:21
http://genomemedicine.com/content/3/4/21

Genome Medicine

RESEARCH HIGHLIGHT

From sequence to functional understanding: the difficult road ahead

Periklis Makrythanasis[1] and Stylianos E Antonarakis[1,2*]

Legătura secvență –funcție

Pentru identificarea computerizată a funţiei unei proteine sau gene necunoscute se pleacă de la următoarele **premize**:

1. **toate genele/proteinele au evoluat din alte gene/proteine** prin **mutaţia** secvenţei primare;

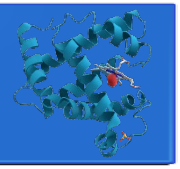
Ce înseamnă și cum funcționează evoluția?

Mutaţiile reprezintă modificări spontane nedirijate a mesajului genetic. Cel mai frecvent mutaţiile apar în procesul de replicarea ADN-ului sau **prin acțiunea factorilor de mediu asupra ADN-ului**. Mutaţiile reprezintă materialul de bază pentru variabilitatea şi evoluţia organismelor vii. **Cum?**

Funcţie de **amploarea** lor mutaţiile se clasifică în:

A. Mutaţii punctiforme
B. Mutaţii de amploare mică
C. Mutaţii de amploare mare

Mutațiile ca sursă de diversificare a informației genetice



A. Mutaţiile punctiforme - mutaţii ce afectează o singură bază azotată din secvenţa acizilor nucleici.

Au fost identicate **trei tipuri de mutaţii punctiforme**:

- **Substituţii** – înlocuirea unei baze azotate cu alta

  5'ACCGTCTA3' → 5'AC*G*GTCTA3'

- **Inserţii** – adăugarea unei baze azotate suplimentare

  5'ACCGTCTA3' → 5'ACC*T*GTCTA3'

- **Deleţii** – pierderea uneia sau a mai multor baze azotate

  5'ACCGTCTA3' → 5'AGTCTA3'

Funcţie de **efectul** lor asupra produsului codificat de gena în care apar, mutaţiile punctiforme se clasifică în:

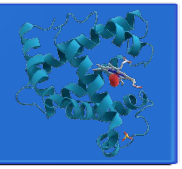
a. **mutaţie non-sens** – modificare unei baze duce la schimbarea mesajului unui codon în STOP, proteina codificată de genă fiind astfel mai scurtă și nefuncțională;

5'AUGGUC**UAU**CUAGGCGAUUAA 3' →5'AUGGUC**UAA**CUAGGCGAUUAA 3'
START V T L G D Stop START V STOP

b. mutaţie cu sens greşit – modificare unei baze duce la schimbarea mesajului unui codon şi duce la încorporarea unui alt aminoacid în molecula proteică care afectează funcţia proteinei codificate

5'AUGGUCUAUCUAGGCGAUUAA 3' →5'AUGGUCUAUCUAGGCGA*A*UAA 3'
START V T L G D Stop START V T L G E Stop

Mutațiile ca sursă de diversificare a informației genetice



c. mutaţie neutră – modificare unei baze ce duce la schimbarea mesajului unui codon, încorporarea unui aminoacid echivalent în molecula proteică, dar nu modifică funcţia proteinei codificate; Cum se explică acest lucru?

d. mutaţie silenţioasă – modificare unei baze ce duce la schimbarea mesajului unui codon dar care:

1. are loc într-o zonă ne-tradusă în proteine sau ARN – Ex: introni
2. nu modifică aminoacizii încorporaţi – **de ce? - mutaţie sinonimă**

```
5'AUGGUCUAUCUAGGCGAUUAA 3' →5'AUGGUCUAUCUAGGAGAUUAA 3'
START V  T  L  G  D Stop    START V  T  L  G  D Stop
```

e. mutaţii cu schimbarea cadrului de lectură (frame-shifts) – inserţia sau deleţia unei baze azotate ce duce la modificarea modului în care ribosomul citeşte mesajul genetic de pe molecula de ARNm.

```
5'AUGGUCUAUCUAGGCGAUUAA 3' →5'AUGGUCUUCUAGGCGAUUAA 3'
START V  T  L  G  D Stop    START V  F  STOP
```

B. Mutaţii de amploare mică – sunt asemănătoare mutaţiilor punctiforme d.p.v. al tipurilor şi efectelor, dar cuprind câteva baze azotate

C. Mutaţii de amploare mare – mutaţii de dimesiuni mari ce afectează poziţia unei gene în cadrul cromozomului şi modul de organizarea a materialului genetic – duplicări, inserții de gene, rearanjări cromozomiale.

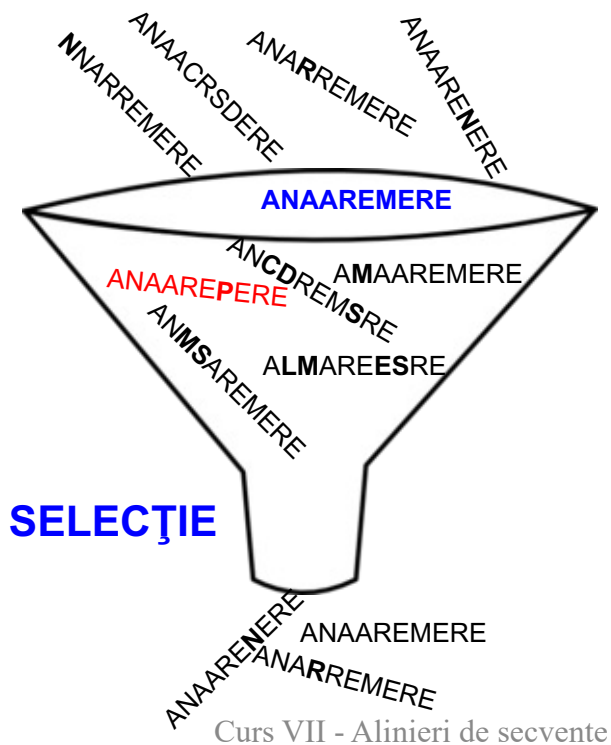
Legătura secvență –funcție

2. înlocuirea unui aminoacid cu altul într-o proteină nu este întoteauna aleatoare ci este corelată cu rolul aminoacidului în cadrul proteinei. Din punct de vedere are frecvenței cu care sunt înlocuiți au fost descrise **3 categorii distincte** de aminoacizi într-o secvență proteică:

a) **aminoacizi înalt conservaţi** - nu sunt înlocuiţi decât extrem rar - sunt aminoacizii din **situsul catalitic** sau **funcţional**, implicați în mod direct de realizarea funcţiei;

b) **aminoacizi conservaţi** - sunt înlocuiţi destul de rar - sunt aminoacizii implicaţi în realizarea structurilor secundare şi terţiare;

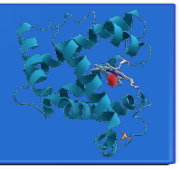
c) **aminoacizi puţin conservaţi** - sunt înlocuiţi frecvent - sunt în general aminoacizii de pe suprafaţa proteinelor, înlocuirea lor nu modifică semnificativ funcția proteinei.



Diverse variante ale aceleași secvențe de aminoacizi obținute prin mutații aleatorii

Variante ale secvenței inițiale în care mesajul (funcția) este păstrată

Alinieri de secvențe



3. Deoarece **secvenţele de aminoacizi ale proteinelor / de nucleotide ale genelor au evoluat una din cealaltă, ele nu au caracter randomic**, ci prezintă mai degrabă un anumit **grad de similaritate** ceea ce permite compararea lor.

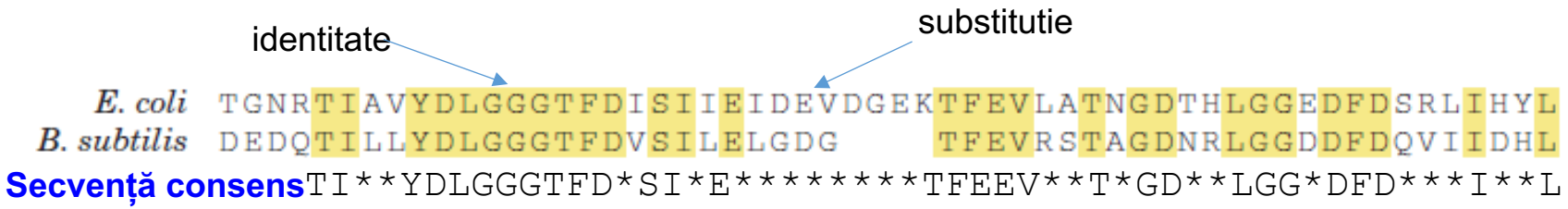
Pentru compararea a două secvențe se introduc noțiunile de:

a. **aliniere a două sau mai multe secvenţe** - fiecare aminoacid (nucleotid) din secvenţa A este comparat cu aminoacidul (nucleotidul) corespunzător din secvenţa B. O corespondenţă între doi aminoacizi (nucleotide) din aceeaşi poziţie pe cele două secvenţe poartă numele de **identitate**, iar o neconcordanţă se numeşte **substituţie;**

Alinierile a două sau mai multe secvențe pot fi:

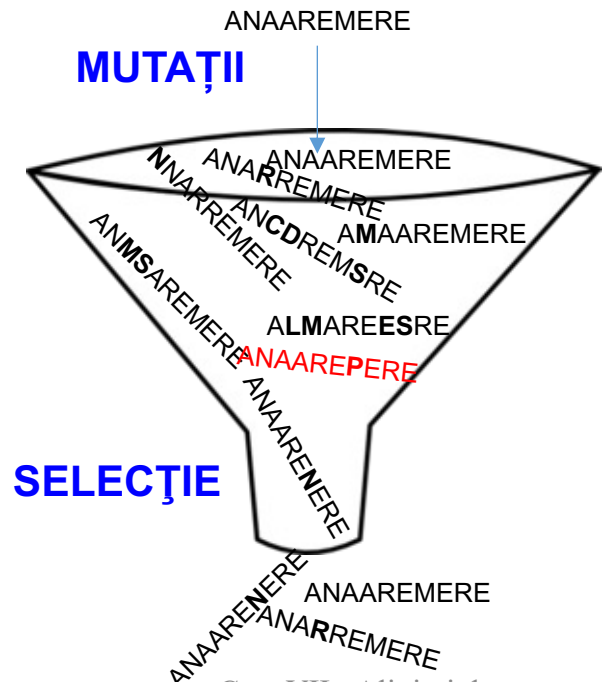
-**alinieri locale** - identifică subregiunile similare dintre două secvenţe

-**alineri globale** - compară două secvenţe pe toată lungimea lor și se utilizează pentru a compara secvenţe de dimensiuni similare dar foarte apropiate evolutiv.



Alinieri de secvențe

b. **identitate -** la nivel de secvenţă - se referă la două secvenţe care prezintă asemănări una în raport cu cealaltă datorită unui număr mai mare sau mai mic de **aminoacizi identici;** similaritatea se exprimă ca **procente identitate** - % de aminoacizi ce sunt identici între 2 secvențe;
c. similaritate la nivel de secvenţă - se referă la două secvenţe care prezintă asemănări una în raport cu cealaltă datorită unui număr mai mare sau mai mic de **aminoacizi identici** dar ia în calcul și **semnificația substituțiilor dintre aminoacizi ;**
c. omologie - se referă la faptul că două secvenţe se aseamănă una cu cealaltă deoarece au evoluat dintr-un strămoş comun, dar nu au obligatoriu aceeași funcție;
d. **secvența de aminoacizi identici înalt conservați** dintr-o aliniere a două sau mai multe secvențe se numește **secvență consens (consensus)**;

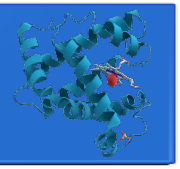


Secvență inițială

Secvențe omoloage
Au evoluat dintr-un strămoș comun

Variante ale secvenței inițiale în care mesajul (funcția) este păstrată
Secvențe similare

Alinieri de secvențe



Gradul de similaritate a două proteine la nivel de secvenţă este dictat, pe de o parte, de **numărul de mutații ce le diferențiază** (distanţa evolutivă) şi, pe de altă parte, de **structurile lor tridimensionale** şi **de funcţiile specifice** pe care cele două proteine le îndeplinesc.

Două secvențe de nucleotide similare vor codifica un mesaj genetic similar și deci vor avea funcții similare.

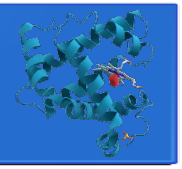
Două proteine similare vor avea structuri similare și deci funcții similare.

Întrebarea inițială: `Care este funcția acestei peptide și implicit a genei codificatoare?´

MAAKYRIGYFVGSLATGSINRVLSQALINLAPEDLEFSEIPIRDLPLYSYDYDADFPPEGR

devine: **cu ce peptidă cunoscută este similară această secvență?**

BLAST – identificarea de secvențe similare



BLAST - Basic Local Alignment Search Tool

1. identifică, dintr-o bază de date, secvenţele similare cu o **secvenţă ţintă (tinta analizei, experimentului)**. Aceste secvenţe identificate poartă numele de **secvenţe "subiect",** iar identificarea lor se bazează pe **alinieri locale**. Secvenţa „subiect este „suprapusă" peste cea țintă la nivelul alinierilor locale astfel încât secvenţele comparate vor fi alcătuite din zone perfect aliniate şi zone nealiniate (aşa numitele **GAP**'s) care formează bucle între o aliniere locală şi următoarea aliniere locală.

2. cuantifică nivelul de similaritate dintre secvenţele "subiect" şi secvenţa ţintă prin utilizarea unor **matrici de substituţie.** O matrice de substituţie arată frecvenţa cu care un aminoacid este înlocuit cu altul şi are la bază observaţiile experimentale.

	A	C	D	E	F	G	H →
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	
G	0	-3	-1	-2	-3		
H	-2	-3	-1	0			
↓							

BLOSUM 62

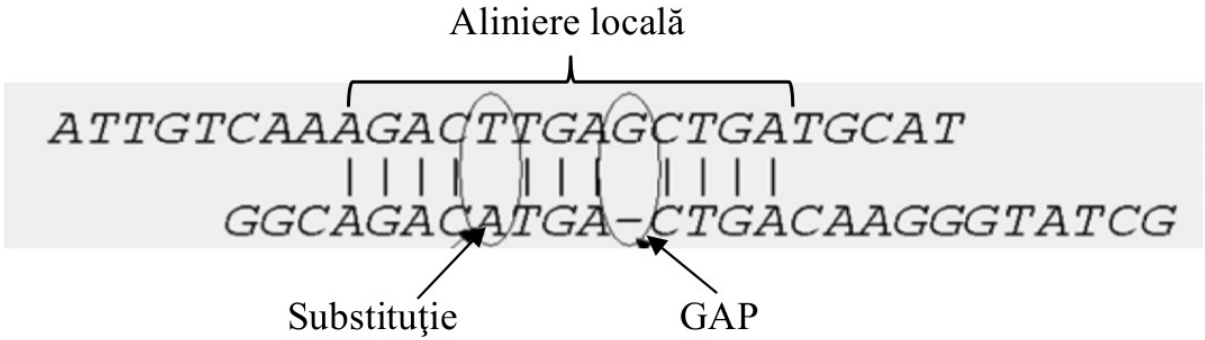
FIGURA 35. Exemplu de matrice BLOSUM.
Cu litere mari sunt reprezentaţi aminoacizii. Cel mai mare punctaj au identităţile, iar funcţie de frecvenţa de substituţie (observată practic în laborator) primesc un anumit punctaj şi substituţiile. Se poate observa şi existenţa unor punctaje negative, alocate pentru substituţiile cel mai puţin întâlnite.

BLAST – identificarea de secvențe similare

3. Calculează unu **scor de similaritate** prin însumarea punctelor pentru fiecare pereche aminoacid-aminoacid și **ierarhizează secvenţele** ţintă funcţie de valoarea acestui scor.

Scoruri de similaritate calculate de BLAST:

- **punctaj brut** (engl. Raw score) notat cu S, este calculat prin însumarea punctelor pentru fiecare pereche aminoacid-aminoacid, aminoacid-nimic şi penalizărilor pentru GAP; nu permite ierarhizarea secvențelor, valoare lui depinde de lungimea secvențelor analizate;



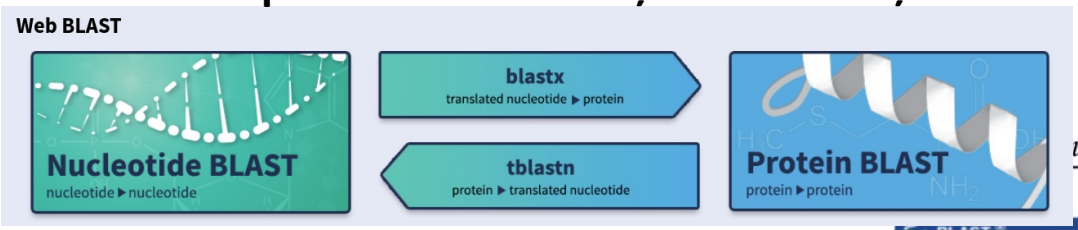
$$S = \Sigma_{(\text{identităţi, substituţii})} - \Sigma_{(\text{penalizări GAP})}$$

- **scorul în biţi notat cu S'** - se calculează prin normalizarea lui S în funcţie de diverse variabile statistice care depind, la rândul lor, de tipul de matrice utilizat. **Cu cât punctajul S' obţinut este mai mare cu atât asemănarea dintre cele două secvenţe este mai mare;**
- **parametru statistic E** - care se defineşte ca număr de potriviri care apar doar datorită şansei într-o bază de date de o anumită dimensiune. **Cu cât valorile lui E sunt mai mici, cu atât rezultatele sunt considerate ca având un înalt grad de semnificaţie (alinierea fiind deci statistic semnificativă).**

Cum se realizează o analiză BLAST?

1. Accesează: https://blast.ncbi.nlm.nih.gov/Blast.cgi

2. Selectează tipul de analiză funcție de secvența de interes:

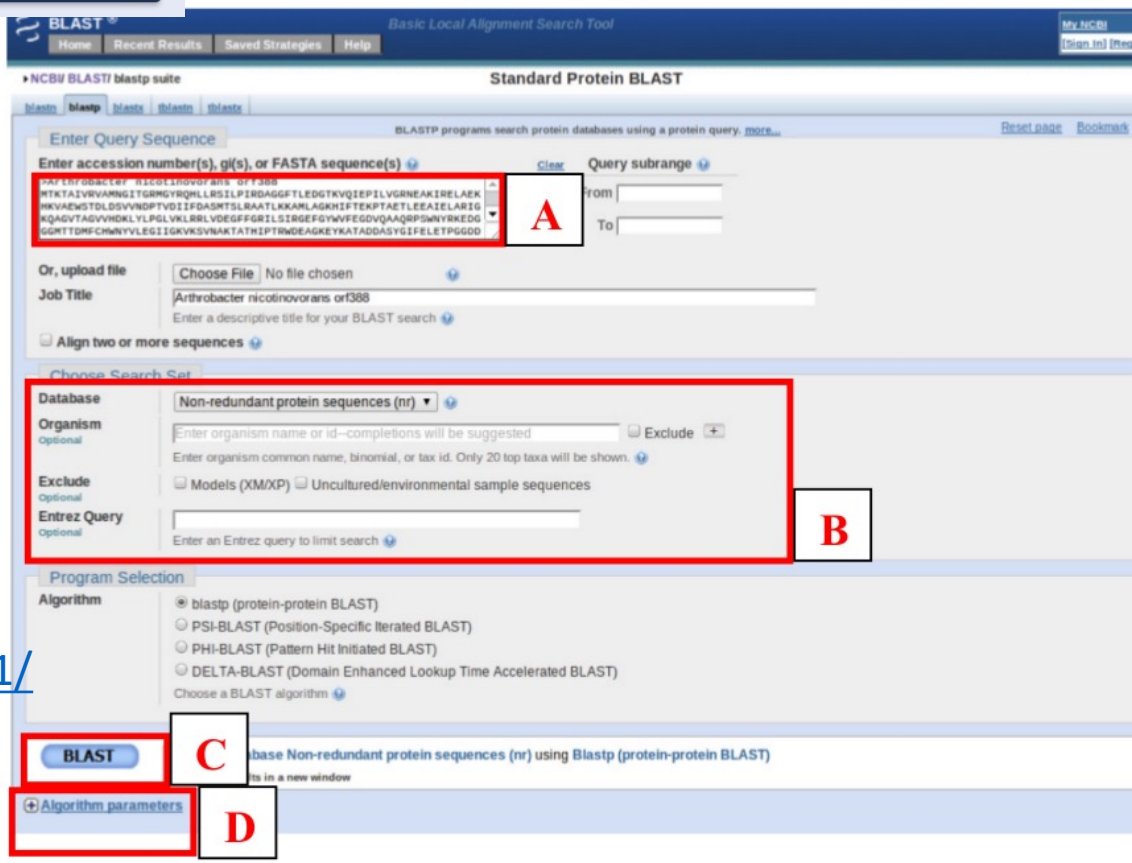


utaționale 3

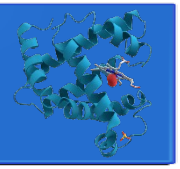
3. Copie secvența în căsuța pentru secvența țintă (query), setează parametrii căutării și apasă BLAST

A – căsuţa text în care a fost inserată secvenţa ţintă în format FASTA;
B – zona cu parametrii utilizaţi pentru restrângerea spaţiului de căutare;
C – buton pentru lansarea investigării;
D – zona cu parametrii algoritmului de căutare

http://www.ncbi.nlm.nih.gov/books/NBK21101/



Parametri pentru restrângerea spațiului de căutare



„Query subrange”– se utilizează pentru a reduce secvenţa investigată doar la un anumit fragment precizat prin poziţia de început şi de sfârşit;

„Or, upload file” – permite încărcarea unui fişier text în format FASTA; în acest caz nu mai este necesară precizarea secvenţei în căsuţa text;

„Database” – permite selecţia bazei de date unde se va realiza investigarea; funcţia este utilă pentru a restrânge investigarea în diverse direcţii; cea mai mare bază de date utilizabilă este **„**non-redundant protein sequences (nr)”; cea mai mică este **„**Protein Data Bank proteins (pdb)” care are avantajul de a fi alcătuită doar din proteine cu structură determinată experimental şi funcţie cunoscută;

„Algorithm” – permite selectarea diverşilor algoritmi de identificare şi aliniere a secvenţelor similare; de asemenea, tot în această pagină, apăsând semnul **„**+” din faţa textului **„**Algorithm parameters”, se pot modifica o serie de parametri ai algoritmului de căutare, precum:

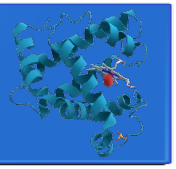
„Max target sequences” – numărul maxim de secvenţe subiect ce va fi luat în calcul;

„Expect” – vor fi afişate doar rezultatele care au o valoare a lui E mai mică decât cea specificată;
„Matrix” – tipul de matrice care va fi utilizat pentru calcularea punctajului similarităţii (BLOSUM62, PAM70 etc.)
„Gap Costs” – schema de punctaj corespunzătoare prezenţei unui GAP.

Rezultate BLAST

A – informaţii generale privind interogarea realizată;
B – domeniile înalt conservate identificate;
C – prezentarea grafică de ansamblu a rezultatelor;
D – tabel cu secvenţele identificate;
E – exemplu de aliniere între secvenţa de interes şi o secvenţă subiect identificată prin BLAST.

3. Rezultate BLAST și semnificația lor



```
LOCUS       YP_002488236            389 aa            linear   BCT 27-JAN-2012
DEFINITION  oxidoreductase [Arthrobacter chlorophenolicus A6].
ACCESSION   YP_002488236
VERSION     YP_002488236.1  GI:220912927
DBLINK      Project: 58969
DBSOURCE    REFSEQ: accession NC_011886.1
KEYWORDS    .
SOURCE      Arthrobacter chlorophenolicus A6
  ORGANISM  Arthrobacter chlorophenolicus A6
            Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
            Micrococcineae; Micrococcaceae; Arthrobacter.
REFERENCE   1  (residues 1 to 389)
  AUTHORS   Lucas,S., Copeland,A., Lapidus,A., Glavina del Rio,T., Tice,H.,
            Bruce,D., Goodwin,L., Pitluck,S., Goltsman,E., Clum,A., Larimer,F.,
            Land,M., Hauser,L., Kyrpides,N., Mikhailova,N., Jansson,J. and
            Richardson,P.
  CONSRTM   US DOE Joint Genome Institute
  TITLE     Complete sequence of chromosome of Arthrobacter chlorophenolicus A6
  JOURNAL   Unpublished
REFERENCE   2  (residues 1 to 389)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (09-JAN-2009) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
```

A

Secvenţă nepublicată, funcție teoretică

```
LOCUS       YP_002541450            385 aa            linear   BCT 01-DEC-2011
DEFINITION  oxidoreductase [Agrobacterium radiobacter K84].
ACCESSION   YP_002541450
VERSION     YP_002541450.1  GI:222082085
DBLINK      Project: 58269
DBSOURCE    REFSEQ: accession NC_011983.1
KEYWORDS    .
SOURCE      Agrobacterium radiobacter K84
  ORGANISM  Agrobacterium radiobacter K84
            Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales;
            Rhizobiaceae; Rhizobium/Agrobacterium group; Agrobacterium.
REFERENCE   1  (residues 1 to 385)
  AUTHORS   Slater,S.C., Goldman,B.S., Goodner,B., Setubal,J.C., Farrand,S.K.,
            Nester,E.W., Burr,T.J., Banta,L., Dickerman,A.W., Paulsen,I.,
            Otten,L., Suen,G., Welch,R., Almeida,N.F., Arnold,F., Burton,O.T.,
            Du,Z., Ewing,A., Godsy,E., Heisel,S., Houmiel,K.L., Jhaveri,J.,
            Lu,J., Miller,N.M., Norton,S., Chen,Q., Phoolcharoen,W., Ohlin,V.,
            Ondrusek,D., Pride,N., Stricklin,S.L., Sun,J., Wheeler,C.,
            Wilson,L., Zhu,H. and Wood,D.W.
  TITLE     Genome sequences of three agrobacterium biovars help elucidate the
            evolution of multichromosome genomes in bacteria
  JOURNAL   J. Bacteriol. 191 (8), 2501-2511 (2009)
   PUBMED   19251847
```

B

Secvenţă publicată, funcție teoretică

```
LOCUS       2GLX_A                  332 aa            linear   BCT 24-SEP-2008
DEFINITION  Chain A, Crystal Structure Analysis Of Bacterial 1,5-Af Reductase.
ACCESSION   2GLX_A
VERSION     2GLX_A  GI:114794099
DBSOURCE    pdb: molecule 2GLX, chain 65, release Aug 27, 2007;
            deposition: Apr 5, 2006;
            class: Oxidoreductase;
            source: Mol_id: 1; Organism_scientific: Ensifer Adhaerens;
            Organism_common: Bacteria; Strain: S-30.7.5; Gene: Afr;
            Expression_system: Escherichia Coli; Expression_system_common:
            Bacteria; Expression_system_strain: Bl21(De3);
            Expression_system_vector_type: Plasmid; Expression_system_plasmid:
            Pet24a;
            Exp. method: X-Ray Diffraction.
KEYWORDS    .
SOURCE      Ensifer adhaerens (Sinorhizobium morelense)
  ORGANISM  Ensifer adhaerens
            Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales;
            Rhizobiaceae; Sinorhizobium/Ensifer group; Ensifer.
REFERENCE   1  (residues 1 to 332)
  AUTHORS   Kuhn,A., Yu,S. and Giffhorn,F.
  TITLE     Catabolism of 1,5-anhydro-D-fructose in Sinorhizobium morelense
            S-30.7.5: discovery, characterization, and overexpression of a new
            1,5-anhydro-D-fructose reductase and its application in sugar
            analysis and rare sugar synthesis
  JOURNAL   Appl. Environ. Microbiol. 72 (2), 1248-1257 (2006)
   PUBMED   16461673
```

C

Secvenţă publicată, funcție demonstrată experimental