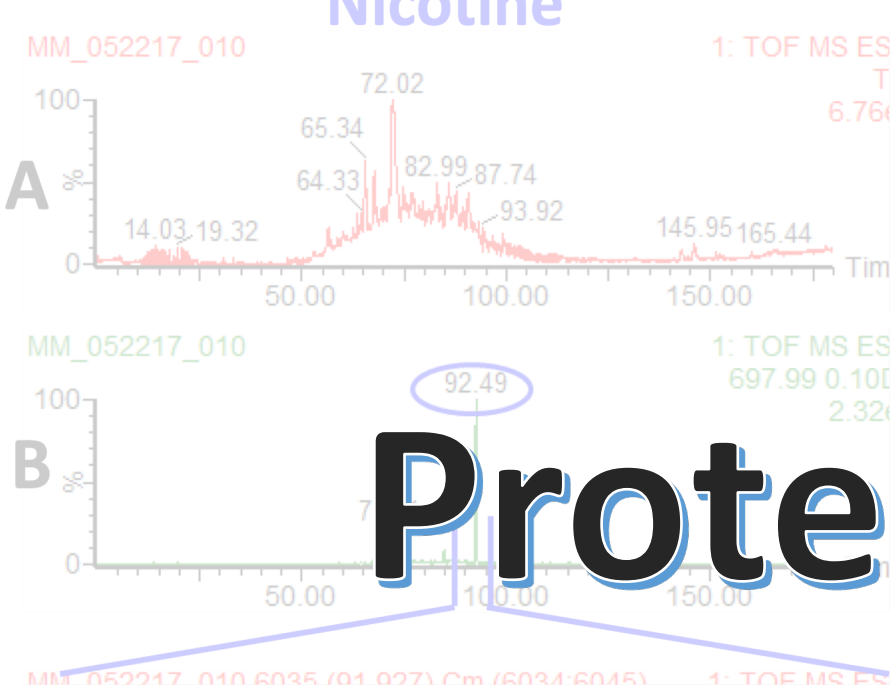
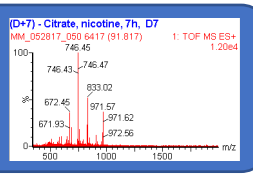


Proteomică

19.03.2026

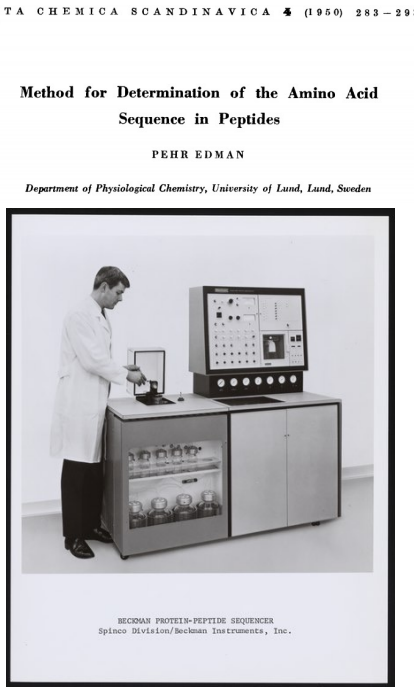
Curs VI – Secvențe și baze de date cu secvențe

Secvențe de aminoacizi și baze de date cu secvențe



Multă vreme s-a considerat că stabilirea secvenței numărului, naturii şi succesiunii resturilor de aminoacizi, sau, pe scurt a secvenței de aminoacizi a unei proteine este imposibilă. Apariția în **1950** a lucrării lui **Edman P.** a pus bazele metodelor de secvențiere a catenelor de aminoacizi și apariția **secvențiatoarelor de peptide**. Frederick Sanger este cel care, în **1953** reușește să secvențieze pentru prima dată în istorie o proteină – **insulin**a și primește premiul Nobel pentru acest lucru.

În prezent, metodele de de stabilire a secvenței de aminoacizi a unei proteine sau peptide se clasifică în două două categorii distincte:


ACTA CHEMICA SCANDINAVICA 4 (1950) 283–293

Method for Determination of the Amino Acid Sequence in Peptides

PEHR EDMAN

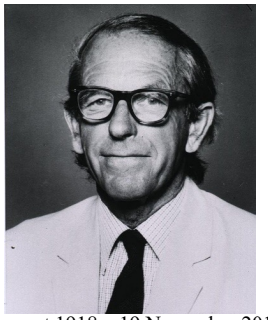
Department of Physiological Chemistry, University of Lund, Lund, Sweden

BECKMAN PROTEIN-PEPTIDE SEQUENCER
Spinco Division/Beckman Instruments, Inc.

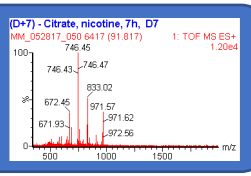
1. **Metode experimentale** ce implică izolarea/purificarea proteinei de interes – **degradarea Edman** (Lehninger, editia 7, pag. 98-100) și **spectrometria de masă;**

2. **Metode *in-silico,*** bazate pe cunoașterea prealabilă a secvenței de nucleotide și realizarea cu ajutorul calculatorului a procesului de transcriere/traducere a informației genetice. Acizii nucleici pot fi secvențiați în prezent prin numeroase metode ce se clasifică în:

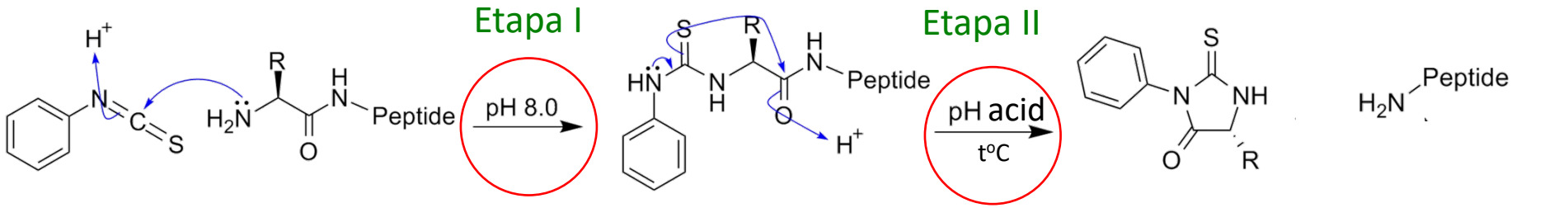
- metode derivate de la **metoda chimică** a lui **Maxam şi Gilbert;**
- metode derivate de la **metoda enzimatică** a lui **Sanger** (același Frederick Sanger ca cel de sus, primește al doua oară premiul Nobel în 1980)**;**
- **metode de "nouă generaţie (next-gen)"** sau mai corect **metode de secvențiere în masă (high throughput)** – pirosecvenţiere, ion-torrent, etc.


Frederick Sanger - 13 August 1918 – 19 November 2013

Degradarea Edman



Metoda Edman sau **metoda degradării Edman** - are la bază reacția Edman ce constă în marcarea aminoacidului N terminal cu un cromofor specific, clivarea și identificarea lui.



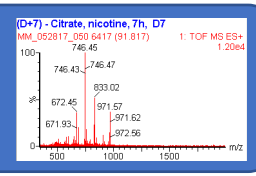
Fenil-izotiocianat

Aac N terminal este marcat

Aac N terminal este clivat și poate fi identificat prin cromatografie

Peptida are un aminoacid în minus

Degradarea Edman



Prin repetarea reacției Edman pe peptida scurtată în prima reacție se poate stabili aminoacidul următor. Reacția Edman poate fi astfel folosită în mod secvențial, pentru a identifica unul câte unul fiecare aminoacid dintr-o secvență peptidică**. Randamentul reacției nu este însă 100%, astfel încât prin folosirea repetată a reacției Edman se pot stabili primi aproximativ 30-40 aminoacizi dintr-o proteină**.

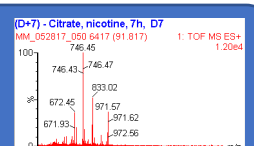
În cazul proteinelor de dimensiuni mari, procesul de **secvențiere Edman** constă în parcurgerea **următoarelor etape**:

1. **Stabilirea aminoacizilor** ce intră în alcătuirea proteinei precum și a raportului molar dintre aceștia – legăturile peptidice se hidrolizează cu HCL și aminoacizi rezultați sunt identificați prin HPLC sau cromatografie în strat subțire;
2. **Stabilirea numărului de peptide** din structura proteinei – se realizează o reacție Edman folosind 1-fluoro-2,4-dinitrobenzen (FDNB) și se hidrolizează proteina cu HCL, numărul de aminoacizi derivați rezultați indică numărul de capete N-terminale existente;
3. **Reducerea legăturilor S-S** și **clivarea cu agenți proteolitici cu specificitate cunoscută** - se generează peptide de dimensiuni convenabile **într-o manieră predictibilă**;
4. **Secvențierea folosind reacția Edman cu fenil-izotiocianat** a tuturor peptidelor obținute;
5. **Asamblarea secvenței peptidelor rezultate** pentru a obține secvența proteinei inițiale.

TABLE 3–7 The Specificity of Some Common Methods for Fragmenting Polypeptide Chains

Reagent (biological source)*	Cleavage points†
Trypsin (bovine pancreas)	Lys, Arg (C)
Submaxillarus protease (mouse submaxillary gland)	Arg (C)
Chymotrypsin (bovine pancreas)	Phe, Trp, Tyr (C)
Staphylococcus aureus V8 protease (bacterium *S. aureus*)	Asp, Glu (C)
Asp-*N*-protease (bacterium *Pseudomonas fragi*)	Asp, Glu (N)
Pepsin (porcine stomach)	Leu, Phe, Trp, Tyr (N)
Endoproteinase Lys C (bacterium *Lysobacter enzymogenes*)	Lys (C)
Cyanogen bromide	Met (C)

Degradarea Edman



Stabilirea aminoacizilor

Polypeptide

Procedure: hydrolyze; separate amino acids

Result:

A 5	H 2	R 1
C 2	I 3	S 2
D 4	K 2	T 1
E 2	L 2	V 1
F 1	M 2	Y 2
G 3	P 3	

Conclusion: Polypeptide has 38 amino acid residues. Trypsin will cleave three times (at one R (Arg) and two K (Lys)) to give four fragments. Cyanogen bromide will cleave at two M (Met) to give three fragments.

Stabilirea numărului de peptide

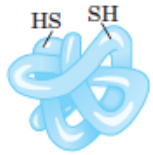
react with FDNB; hydrolyze; separate amino acids

Result: 2,4-Dinitrophenylglutamate detected

Conclusion: E (Glu) is amino-terminal residue.

Reducerea legăturilor S-S

reduce disulfide bonds (if present)



Clivarea cu agenți proteolitici

cleave with trypsin; separate fragments; sequence by Edman degradation

- T-1 GASMALIK
- T-2 EGAAYHDFEPIDPR
- T-3 DCVHSD
- T-4 YLIACGPMTK

T-2 placed at amino terminus because it begins with E (Glu).
T-3 placed at carboxyl terminus because it does not end with R (Arg) or K (Lys).

Secvențierea folosind reacția Edman cu fenil-izotiocianat

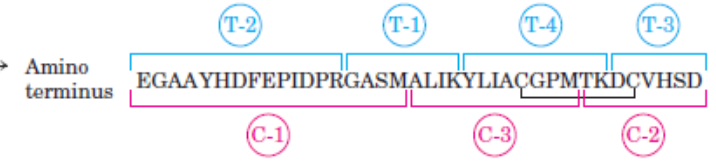
cleave with cyanogen bromide; separate fragments; sequence by Edman degradation

- C-1 EGAAYHDFEPIDPRGASM
- C-2 TKDCVHSD
- C-3 ALIKYLIACGPM

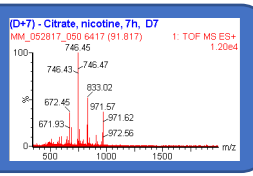
C-3 overlaps with T-1 and T-4, allowing them to be ordered.

Asamblarea secvenței peptidelor rezultate

establish sequence



Metode in-silico stabilire a secvenței proteinelor (peptidelor)



Dogma centrală a biologie moleculare postulează că **fiecare tripletă de nucleotide (codon) din ADN codifică câte un aminoacid** (exceptând codonii STOP), corespondența codon-aac fiind dată de **codul genetic**. În principiu, dacă se cunoaște secvența unei gene, stabilirea secvenței de aminoacizi codificată este o chestiune simplă – se găsește corespondența aminoacizilor pe baza codonilor din molecula de ARNm generată.

Informație **ADN** Informație

Secvență de nucleotide (A;T;G;C)

Transcriere (transcripție)

Complementaritatea nucleotidelor

ARN

Informație

Secvență de nucleotide (A;U;G;C)

Traducere (translație)

Codul genetic

Polipeptide

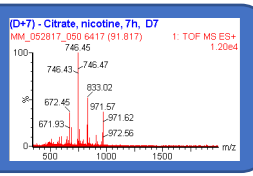
Secvență de aminoacizi

Nucleic Acid	Nucleobases	Base complement
DNA	adenine(A), thymine(T), guanine(G), cytosine(C)	A=T, G≡C
RNA	adenine(A), uracil(U), guanine(G), cytosine(C)	A=U, G≡C

1st base	2nd base U		2nd base C		2nd base A		2nd base G		3rd base
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine	UCA		UAA[B]	Stop (*Ochre*)	UGA[B]	Stop (*Opal*)	A
	UUG		UCG		UAG[B]	Stop (*Amber*)	UGG	(Trp/W) Tryptophan	G
C	CUU		CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	AUG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

Proprietățile codului genetic?

Metode in-silico stabilire a secvenței proteinelor (peptidelor)



Pentru a putea stabili secvența unei proteine plecând de la secvența de ADN a genei codificatoare este necesar elucidarea următoarelor aspecte:

1. Care din cele 2 catene ale moleculei de ADN codifică informația genetică?

Catenă **sens** (+) - **direct**

ADN

5' ATGGCTAGGGCCGCAAGGGAAATGGAGAGGGAATAA 3'

3' TACCGATCCCGGCGTTCCCTTTACCTCTCCCTTATT 5'

Transcriere

Catenă **antisens** (-) - **reverse**

ARNm 5' AUGGCUAGGGCCGCAAGGGAAAUGGAGAGGGAAUAA 3'

Traducere

Catenă polipeptidică **M** A R A A R E M E R E **STOP**

START

ARNm **complementar** cu **catena antisens**, **aceeași secvență cu catena sens** conține informație - are **sens**

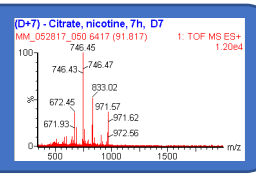
5' UUAUUCCCUCUCCAUUUCCCUUGCGGCCCUAGCCAU 3'

Catenă polipeptidică L F P L H F P C G P S H

ARNm **complementar** cu **catena sens (inversata directia!!!)**, **aceeași secvență cu catena non-sens** nu conține informație bună– este **non-sens**

Mesajul genetic este codificat pe catena sens din ADN, dar catena copiată în ARNm este cea antisens. ARNm are aceeași secventă cu catena sens, dar T este inlocuit de U.

Metode in-silico stabilire a secvenței proteinelor (peptidelor)



2. Care este prima nucleotidă a primului codon – cadrul de lectură – de unde începe traducerea informației genetice?

Pe molecula de ARNm (și corespunzător catena sens) există 3 posibilități de a citi informația genetică – 3 cadre de lectură diferite:

Incepând cu nucleotida 1:

5' AUGGCUAGGGCCGCAAGGGAAAUGGAGAGGGAAUAA 3'

M A R A A R E M E R E STOP

START

Incepând cu nucleotida 2:

5' AUGGCUAGGGCCGCAAGGGAAAUGGAGAGGGAAUAA 3'

W L G P Q G K W R G N

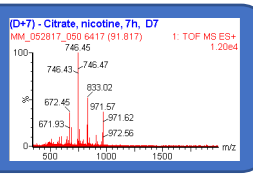
Incepând cu nucleotida 3:

5' AUGGCUAGGGCCGCAAGGGAAAUGGAGAGGGAAUAA 3'

G * G R K G N G E G I

Pe o molecula de ADN dublu catenar există 6 posibilități de traducere a informației genetice: 3 cadre de lectură pe o catenă și 3 cadre de lectură pe cealaltă catenă

Transcriere si traducere in-silico - Sequence Manipulation Suite 2



SMS

Sequence Manipulation Suite:

Version 2

- The Sequence Manipulation Suite is a collection of JavaScript programs for generating, formatting, and analyzing short DNA and protein sequences. It is commonly used by molecular biologists, for teaching, and for program and algorithm testing.
- See the about the Sequence Manipulation Suite page for more information about individual Sequence Manipulation Suite programs.
- You can easily mirror the Sequence Manipulation Suite on your own web site, or you can use it off-line.
- This version of the Sequence Manipulation Suite represents a complete re-write of the previous version. The new version is much faster and has many new features. The previous version of the Sequence Manipulation Suite can still be accessed.
- Send questions and comments to stothard@ualberta.ca.

new window | home | citation

Mon Nov 6 02:56:29 2017

Valid XHTML 1.0; Valid CSS

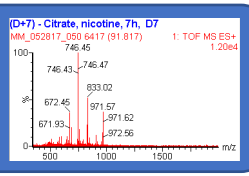
Format Conversion
-Combine FASTA
-EMBL to FASTA
-EMBL Feature Extractor
-EMBL Trans Extractor
-Filter DNA
-Filter Protein
-GenBank to FASTA
-GenBank Feature Extractor
-GenBank Trans Extractor
-One to Three
-Range Extractor DNA
-Range Extractor Protein
-Reverse Complement
-Split Codons
-Split FASTA
-Three to One
-Window Extractor DNA
-Window Extractor Protein

Sequence Analysis
-Codon Plot
-Codon Usage
-CpG Islands
-DNA Molecular Weight
-DNA Pattern Find
-DNA Stats
-Fuzzy Search DNA
-Fuzzy Search Protein
-Ident and Sim
-Multi Rev Trans
-Mutate for Digest
-ORF Finder
-Pairwise Align Codons
-Pairwise Align DNA
-Pairwise Align Protein
-PCR Primer Stats
-PCR Products
-Protein GRAVY
-Protein Isoelectric Point
-Protein Molecular Weight
-Protein Pattern Find
-Protein Stats
-Restriction Digest
-Restriction Summary
-Reverse Translate
-Translate

Sequence Figures
-Color Align Conservation
-Color Align Properties
-Group DNA
-Group Protein
-Primer Map
-Restriction Map
-Translation Map

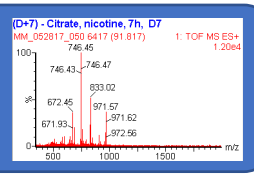
Random Sequences
-Mutate DNA
-Mutate Protein
-Random Coding DNA
-Random DNA Sequence
-Random DNA Regions
-Random Protein Sequence
-Random Protein Regions
-Sample DNA
-Sample Protein
-Shuffle DNA
-Shuffle Protein

https://mail.uaic.ro/~marius.mihasan/research/mirrored_sites_tools/sms2/index.html



	Avantaje	Dezavantaje
1. Metode experimentale de elucidare a secvenței proteinelor (peptidelor)	**Secvențiază strict proteina exprimată, poate identifica și aminoacizi mai putini comuni**	**Necesită purificarea prealabilă a proteinei de interes**
a. Degradarea Edman		
b. Spectrometria de masă	**Timp de lucru scurt**	**Costul de achiziție al instrumentelor**
2. Metode *in-silico* stabilire a secvenței proteinelor (peptidelor)	**Extrem de rapidă**	**Necesită cunoașterea prealbilă a secvenței de ADN** **Nu întotdeauna secvența unei gene este complet tradusă în catena polipeptidică (exoni vs introni)** **Nu identifică aminoacizii modificați și cei mai putin comuni.**

Secvențe de aminoacizi și baze de date cu secvențe



Majoritatea covârșitoare a secvențelor proteice cunoscute în prezent provin din secvența de nucleotide prin treducere cu ajutorul calculatorului, adică *in-silico.* Acest lucru se explică prin faptul că:

A. **Acizii nucleici pot fi foarte ușor și relativ ieftin produși în laborator.** Dacă molecula ADN de secvențiat izolată este într-o cantitate foarte mică, ea poate fi amplificată *in vitro* cu ajutorul unor oligonucleotide amorsă potrivite și a ADN-polimerazi prin reacția PCR. În prezent, o asemenea reacție *in vitro* nu este posibilă pentru peptide sau proteine. De asemenea, procesul de purificare și izolare a moleculelor de ADN este aplicabil pentru pentru orice moleculă de ADN, indiferent de secvență sa, dar în cazul proteinelor nu există o metodă de purificare „universală", ci trebuie adaptată funcție de proprietățile proteinei de secvențiat.
B. **Costurile de secvențiere** a acizilor nucleici sunt semnificativ **mai mici** decât pentru secvențierea peptidelor.

Secvențierea genomului uman - 3 bilioane baze – 10 ani (1990 -2000) - **~2.7 bilioane $**
Secvențierea unui genom uman în 2006 - ~6 bilioane baze - **~14 milioane $**
Secvențierea unui genom uman în 2015 - ~6 bilioane baze - **~4000 $**
Secvențierea unui genom uman în 2016 - ~6 bilioane baze - **~1000 $**

Whole GenomeZ - Whole Genome for Advanced Analysis (130X + 30X) - 2 week turnaround time

€849.00 EUR

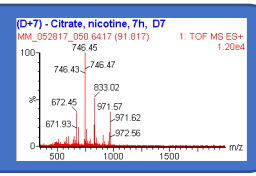
TURNAROUND TIME

2 weeks

Add to cart

Buy it now

Secvențe și fișiere cu secvențe

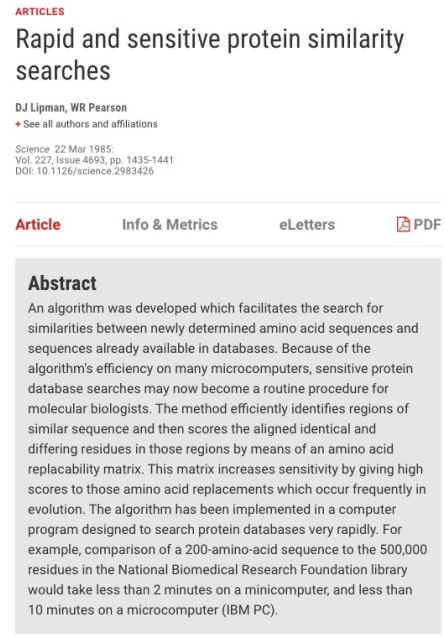


Scăderea continuă a costurilor de secvențiere a condus la acumularea unor cantități uriașe de informații sub formă de secvenţe de nucleotide şi, bineînțeles, de aminoacizi. A fost astfel necesară dezvoltarea unor modalități de stocare computerizată a acestor informații, în așa fel **încât datele să fie ușor prelucrate atât de cercetători cât și de computere**. Au apărut astfel **fișiere cu secvențe** precum și **baze de date** specializate în colectare și stocarea acestora.

Fișierul FASTA

O secvență, fie că este ADN, ARN sau aminoacizi **poate fi reprezentată ca o înșiruire de litere** (cei 20/22 de aminoacizi notați cu o literă pentru proteine). Pentru a se putea manipula și procesa ușor cu ajutorul computerelor, secvențele de nucleotide sau aminoacizi se stochează sub forma **unui fișier text**, în care, alături de secvența propriu-zisă, sunt incluse și o serie de **informații accesorii** precum specia de la care provine secvența, gena, cromozomul, lucrarea în care este descrisă secvența. **Modul în care sunt incluse aceste informații alături de secvența propriu-zisă reprezintă formatul sau standardul unui fișier cu secvențe**.

De-a lungul timpului au existat un număr mare de variante de fișiere/formate create pentru a înregistra secvențe, însă fișierul/formatul **FASTA** este cel ce s-a impus. Inițial, acest format de fișiere a fost propus cu denumirea **FASTA/Pearson** ca parte a suitei de programe de aliniere și comparare a secvențelor **FASTA**. Denumirea programului provine de la **FAST-All**, autorii dorind să atragă atenția că programul propus poate alinia foare repede toate tipurile de secvențe, atât de aminoacizi, cât și de baze azotate.


ARTICLES

Rapid and sensitive protein similarity searches

DJ Lipman, WR Pearson

+ See all authors and affiliations

Science 22 Mar 1985:
Vol. 227, Issue 4693, pp. 1435-1441
DOI: 10.1126/science.2983426

Article | Info & Metrics | eLetters | PDF

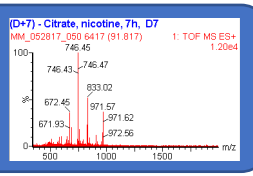
Abstract

An algorithm was developed which facilitates the search for similarities between newly determined amino acid sequences and sequences already available in databases. Because of the algorithm's efficiency on many microcomputers, sensitive protein database searches may now become a routine procedure for molecular biologists. The method efficiently identifies regions of similar sequence and then scores the aligned identical and differing residues in those regions by means of an amino acid replacability matrix. This matrix increases sensitivity by giving high scores to those amino acid replacements which occur frequently in evolution. The algorithm has been implemented in a computer program designed to search protein databases very rapidly. For example, comparison of a 200-amino-acid sequence to the 500,000 residues in the National Biomedical Research Foundation library would take less than 2 minutes on a minicomputer, and less than 10 minutes on a microcomputer (IBM PC).

Un fisier FASTA tipic

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
```

Fișiere FASTA - fișiere cu secvențe



Fișierul FASTA

- un fișier text **ce poate avea sau nu** una din extensiile .fasta .fna, .ffn, .faa, .frn;
- conține secvențe de proteine sau de aminocizi stocate conform **standardului FASTA**:

Elementele formatului FASTA:

A.primul rând de text, marcat cu "**>**" (mai mare) conţine o serie de **informaţii cu caracter opţional**, - specia sau denumirea genei (proteinei);

B.următoarele rânduri conţin secvenţa propriu-zisă, în care nucleotidele/aminoacizii sunt reprezentați folosind codul standard IUPAC cu o singură literă;

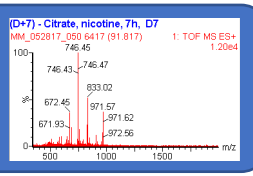
C.fiecare rând al secvenței are în general 80 de caractere (nu mai mult de 120)

A adenozină	**M** A/C (amino)
T timină	**W** A/T (weak)
C citidină	**R** G/A (puRine)
U uracil	**B** G/T/C
G guanină	**D** G/A/T
N A/G/C/T (oricare)	**H** A/C/T
K G/T (keto)	**V** G/C/A
S G/C (strong)	**-** spatiu de dimensiune
Y T/C (p**Y**rimidine)	intermediară

Standardul IUPAC pentru notarea nucleotidelor și aminoacizilor

A alanină	**P** prolină
B aspartat/asparagină	**Q** glutamină
C cistină	**R** arginină
D aspartat	**S** serină
E glutamat	**T** threonină
F fenilalanină	**U** selenocisteină
G glicină	**V** valină
H histidină	**W** triptofan
I isoleucină	**Y** tirosină
K lisină	**Z** glutamat/glutamin
L leucină	**X** oricare
M metionină	***** stop
N asparagină	**-** spațiu

Fișiere FASTA - fișiere cu secvențe



D. O secvență de nucleotide este notată în direcția 5' → 3‘ : prima literă de pe primul rând corespunzător secvenței este nucleotida 1 și are o grupare PO_4 liberă, ultima literă de pe ultimul rând are gruparea OH 3´ liberă. Moleculele circulare se reprezentate linear, prima nucleotidă fiind cel mai frecvent originea de replicare;

E. O secvență polipeptdică este notată în sensul sintezei, de la capătul N terminal spre cel C terminal; prima literă din primul rând reprezintă aminoacidul 1 din secvență - aminoacidul N-terminal; ultima literă reprezintă aminoacidul C-terminal;

F. Poziția literelor poate fi sau nu numerotată; în cazul în care literele din secvență sunt numerotate, numeroatarea se face la începutul fiecărui rând și se include un spațiu după fiecare a 10-a literă.

G. **Literele ce desemnează secvența pot fi sau nu scrise cu majuscule**; indiferent de tipul de scriere, semnificația este aceeași;

H. Unele programe nu acceptă caracterul ‘**–**’ (spațiul în secvență) se indică cu un sir de N pentru nucleotide sau X pentru aminoacizi; **spațiul între litere este ignorat**;

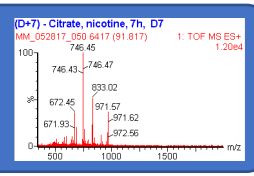
```
>secventa peptidica necunoscuta 1 fara numere
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP

>secventa peptidica necunoscuta cu numere
1   qikdllvsss tdldttlvlv naiyfkgmwk tafnaedtre mpfhvtkqes kpvqmmcmnn
61  sfnvatlpae kmkilelpfa sgdlsmlvll pdevsdleri ektinfeklt ewtnpntmek
121 rrvkvylpqm kieekynlts vlmalgmtdl fipsanltgi ssaeslkisq avhgafmels
181 edgiemagst gviedikhsp eseqfradhp flflikhnpt ntivyfgryw sp
```

Cum scriu secventele?
- Cu font **monospatiat**

TNR Proportional
Courier New Monospace

Baze de date cu secvențe



Baze de date cu secvențe

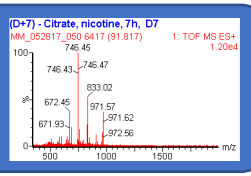
O bază de date cu secvențe reprezintă **o colecție de secvențe de acizi nucleici sau aminoacizi ce au fost stabilite experimental și care au fost depozitate în formă digitalizată pe un server central într-un format tip specific**. Secvențele sunt trimise în bazele de date de către cercetători ce au la dispoziție instrumente de secvențiere. Fiecărei secvențe i se alocă un **identificator unic** – **ID** – **o combinație unică de litere și cifre ce poate fi folosită pentru a regăsi fără echivoc secvența în respectiva bază de date**.

În general, înregistrate unei secvențe în bazele de date precum și accesarea bazelor de date cu secvențe este gratuit.

Bază de date	Site web	Dimensiune *
Baze de date cu secvenţe		
INSDC	http://www.insdc.org/	206 293 625 secvenţe Aceste baze de date sunt sincronizate zilnic între ele şi conţin aceleaşi secvenţe.
DDBJ	http://www.ddbj.nig.ac.jp	
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl/	
GenBank	http://www.ncbi.nlm.nih.gov/genbank/	
European Nucleotide Archive (ENA)	http://www.ebi.ac.uk/ena/	
KEGG	http://www.genome.jp/kegg/	25 679 056 secvenţe
nr	http://blast.ncbi.nlm.nih.gov	145 296 712 secvenţe
UniProtKB/Swiss-Prot	http://www.uniprot.org/	556 568 secvenţe
UniProtKB/TrEMBL	http://www.uniprot.org/	107 627 435 secvenţe

* În Februarie 2018

Redundanță și incertitudine în bazele de date



Secvenţele sunt înregistrate în bazele de date de către o multitudine de utilizatori, independenți unul de celălalt. În momentul înregistrării unei secvențe aceasta primește un ID unic și este tratată ca entitate de sine stătătoare. Pot apare astfel 2 situații:

I

aceeaşi secvență este înregistrată cu nume diferite

aceeaşi proteină/genă este secvențiată de la specii diferite

aceeaşi secvență este înregistrată de autori diferiţi

II

Mai multe ID pentru aceeaşi secvenţă

II

V

Redundanță

2 secvențe marcate ca diferite pot avea aceeși funcție

II

Datele de secvențe sunt înregistrate automat. Metodele moderne de secvențiere produc o cantitate mare de date ce sunt analizate și înregistrate în bazele de date automat.

II

Funcția secvențelor este stabilită automat și nu este verificată experimental.

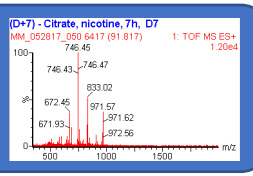
II

V

Nesiguranță

1 secvență poate fi identificată ca având o anumită funcție în mod eronat.

Principalele baze de date cu secvențe



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

- o bază de date cu secvențe ADN creată și întreținută de National Institutes of Health (NIH), USA;
- conține **secvențe adnotate** – pe lângă secvența propriu-zisă, fiecare ID în baza de date conține și o serie de informații accesorii precum:

1. Denumirea
2. GenBank ID
3.
4.
5. Incadrarea taxonomica
6. Lucrări științifice

```
LOCUS       AJ507836              165137 bp    DNA     linear   BCT 24-JUL-2016
DEFINITION  Arthrobacter nicotinovorans pAO1 megaplasmid sequence, strain ATCC
            49919.
ACCESSION   AJ507836
VERSION     AJ507836.1
KEYWORDS    .
SOURCE      Paenarthrobacter nicotinovorans
  ORGANISM  Paenarthrobacter nicotinovorans
            Bacteria; Actinobacteria; Micrococcales; Micrococcaceae;
            Paenarthrobacter.
REFERENCE   6  (bases 1 to 165137)
  AUTHORS   Igloi,G.L. and Brandsch,R.
  TITLE     Sequence of the 165-kilobase catabolic plasmid pAO1 from
            Arthrobacter nicotinovorans and identification of a pAO1-dependent
            nicotine uptake system
  JOURNAL   J. Bacteriol. 185 (6), 1976-1986 (2003)
   PUBMED   12618462
REFERENCE   18 (bases 27948 to 117790)
  AUTHORS   Mihasan,M. and Brandsch,R.
  TITLE     pAO1 of Arthrobacter nicotinovorans and the spread of catabolic
            traits by horizontal gene transfer in gram-positive soil bacteria
  JOURNAL   J. Mol. Evol. 77 (1-2), 22-30 (2013)
   PUBMED   23884627
REFERENCE   19 (bases 1 to 165137)
  AUTHORS   Igloi,G.L. and Brandsch,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (12-SEP-2002) Institute of Biochemistry and Molecular
            Biology, University of Freiburg, Hermann-Herder-Str. 7, Freiburg
            D-79104, GERMANY
REFERENCE   20 (bases 1 to 165137)
  AUTHORS   Mihasan,M. and Brandsch,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (23-DEC-2013) Biochemistry and Molecular Biology
            Laboratory, Faculty of Biology, Alexandru Ioan Cuza University,
            Copou Bvd, No.22B, Iasi 700506, ROMANIA
REMARK      revision by M. Mihasan, authorised by R. Brandsch.
```

7. Autorul ce a solicitat înregistrarea și cel responsabil de secvență

D36–D42 Nucleic Acids Research, 2013, Vol. 41, Database issue *Published online 27 November 2012*
doi:10.1093/nar/gks1195

GenBank

Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and Eric W. Sayers*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

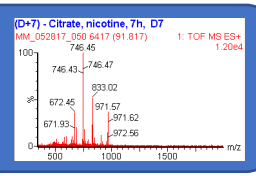
Received September 28, 2012; Revised and Accepted October 29, 2012

8. Gene identificate cu:

```
FEATURES             Location/Qualifiers
     source          1..165137
                     /organism="Paenarthrobacter nicotinovorans"
                     /mol_type="genomic DNA"
                     /strain="ATCC 49919"
                     /culture_collection="ATCC:49919"
                     /db_xref="taxon:29320"
                     /plasmid="pAO1"
     CDS             complement(344..1138)
                     /experiment="non-experimental evidence, no additional
                     details recorded"
                     /note="ORF2"
                     /codon_start=1
                     /transl_table=11
                     /product="putative hydrolase"
                     /protein_id="CAD47860.1"
                     /db_xref="GOA:Q8GAP3"
                     /db_xref="InterPro:IPR000073"
                     /db_xref="UniProtKB/TrEMBL:Q8GAP3"
                     /translation="MPTTPSGIYWESQGPVAAPAVLLIEGYTGQLIGWRDAFCDLLLA
                     QGLRVLRMDNRDVGLSRREDGNYMIADMADDVIDVIADADLGKITIVGQSMGGLIAQH
                     TALGYPNMVTGLVLFYTTPVLDDIDPGILTAEIPRPQHREDAIASFLEGDRATASPAW
                     GYDEAWKRELAGRMFDRAPDRSGLSRQRNAVALLPDLRPRLTELTMPVALIHGRNDAL
                     IRARGSLRIAEVVPQAELHLYPGMGHEIAPALWDEFVAIITRIAVC"
ORIGIN
        1 gatccggcgg tccgccgtcg tggcggcgcg ggcggaggtt gccggcggcg gaccgccgcc
       61 cgcacaagaa ggccttcggg ttccggcagg tggcgcggcc cccgacactg gtcctcgcct
      121 ggggtggaac gtggggtgct gggtgtgggc ggttcgccgg gcgaaccggg aaaggcgtcc
      181 actcctcttc gcttccgtgg ccggcggttg gggccggtcc cgtcgtcgca gagctccgcc
      241 gtgcccggcc ccgcccgttg tctacgacgt ctttttgtgg ctgtccttcg gatcatacgg
      301 ttccgctcca gctcaagctc ctggcagtcc gcaaagctcc ggtctagcac acagcgatgc
      361 gggtaatgat ggcgacgaat tcgtcccaga gcgctggtgc gatttcgtgg cccatacctg
      421 ggtagaggtg cagttcagct tgggggacga cttcggcgat cctcagggat ccacgggctc
      481 ggatgagggc gtcgttccgg ccatgtatga gtgcgaccgg catcgtgagt tcagtaagtc
      541 tcggccgcag atcaggcaac agtgcaactg cgttcctttg ccgggataag ccgctgcggt
      601 cgggggcacg gtcaaacatg cggcctgcca gttcccgctt ccaggcctcg tcgtaacccc
      661 atgccggtga ggctgtggca cggtcgcctt ccaggaacga ggcgatggca tcctcccggt
      721 gttgaaggcg cggaatctcc gcggtcagga tcccgggatc gatgtcgtcg agtacgggcg
```

a. Poziția pe secvență și catena codificatoare
b. denumire genei
c. secvența proteinei codificate
9. Secvența de ADN propriu-zisă

Principalele baze de date cu secvențe



UniProtKB/TrEMBL - http://www.uniprot.org/

- conţine toate secvenţele de nucleotide din GenBank translate în secvenţe de aminoacizi;
- are un nivel foarte mare de redundanță;

UniProtKB/SwissProt - http://www.uniprot.org/

-toate secvenţele conţinute sunt adnotate manual – siguranță crescută legată de funcțiile secvențelor conținute;

-nivelul de redundanţă este minim;

-adăugarea unei noi secvenţe în SwissProt este un proces care implică prelucrarea manuală a datelor corespunzătoare secvenţei respective => SwissProt nu conține cele mai noi secvențe.

TrEMBL este completă dar incorectă, SwissProt este mai corectă, dar incompletă.

Protein Identification Resource PIR

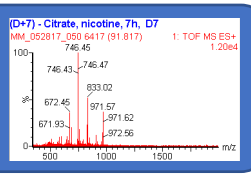
-o colecţie completă de secvenţe lipsită de redundanţă;

-secvenţele identice sau foarte similare sunt gruparea într-o singură intrare în baza de date;

nr

- o bază de date non-redundantă completă, cuprinzând intrări din GenBank, SwissProt, PIR, Protein Research Foundation (PRF) şi Protein data Bank (PDB).

Ce informații pot fi obținute din secvențele de aminoacizi?



MTKTAIVRVAMNGITGRMGYRQHLLRSILPIRDAGGFTLEDGTKVQIEPILVGRNEAKIRELAEKHKVAEWSTDLDSVVNDPTVDIIFDASMTSLRAATLKK
AMLAGKHIFTEKPTAETLEEAIELARIGKQAGVTAGVVHDKLYLPGLVKLRRLVDEGFFGRILSIRGEFGYWVFEGDVQAAQRPSWNYRKEDGGGMTTDMFC
HWNYVLEGIIGKVKSVNAKTATHIPTRWDEAGKEYKATADDASYGIFELETPGGDDVIGQINSSWAVRVYRDELVEFQVDGTHGSAVAGLNKCVAQQRAHTP
KPVWNPDLPVTESFRDQWQEVPANAELDNGFKLQWEEFLRDVVAGREHRFGLLSAARGVQLAELGLQSNDERRTIDIPEITL

1. Stabilirea proprietăților fizico-chimice simple: masă moleculară, pI, indice de stabilitate

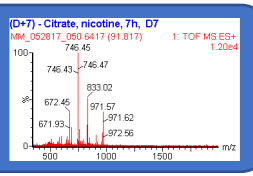
2. Identificarea proteinelor pe baza secvenței lor parțiale

3. Identific alte proteine similare (BABS)

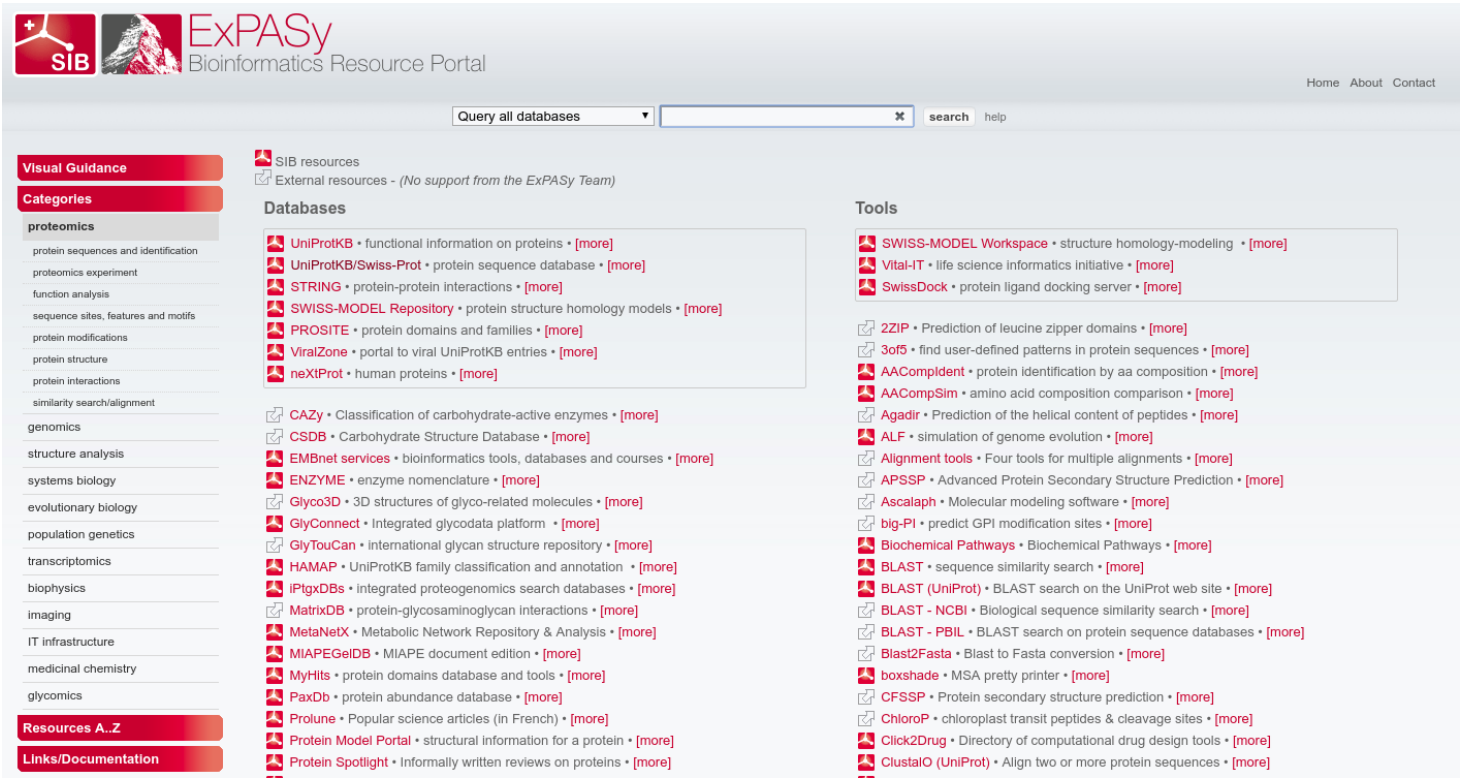
4. Modelarea structurii tridimensionale complete (BABS)

Indiferent de tipul lor, toate aceste informații **au caracter predictiv** și oferă un punct de plecare ce trebuie verificat experimental. Informațiile sunt însă extrem de utile deoarece **permit prioritizarea și orientarea demersului experimental.**

1. Calcularea proprietăților fizico-chimice simple: **ProtParam**

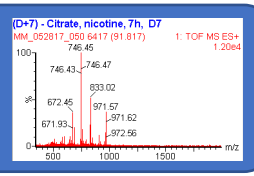


Calcularea proprietăților fizico-chimice simple are la bază o serie de operații matematice simple ce țin cont de numărul și tipurile de aminoacizi dintr-o secvență polipeptidică precum și de proprietățile cunoscute ale acestora. Site-ul **ExPASy (https://www.expasy.org)** al Swiss Institute of Bioinformatics oferă numeroase programe de investigare a secvențelor printre care și programul **ProtParam (https://web.expasy.org/protparam/)**.



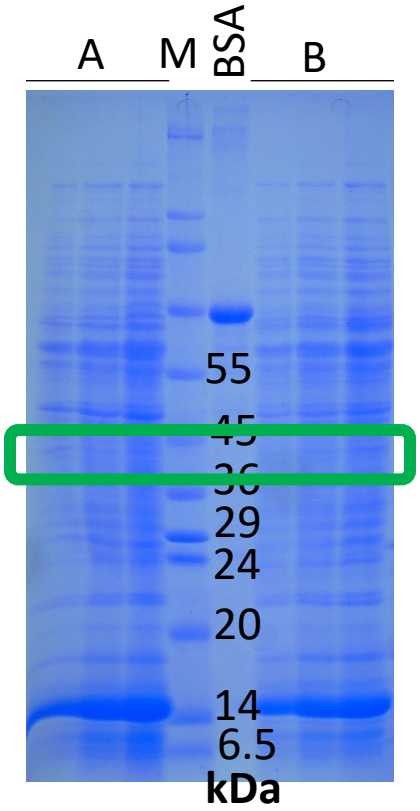
Protein Identification and Analysis Tools on the ExPASy Server;
Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.; (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005); pp. 571-607

1. Calcularea proprietăților fizico-chimice simple: **ProtParam**



Parametrii calculați de **ProtParam**:

1. Masa moleculară Mw funcție de secvență **– suma maselor moleculare a tuturor aminoacizilor din secvența dată**; este exprimată în Daltoni (Da); utilă pentru localizarea unei proteine pe un gel SDS-PAGE;

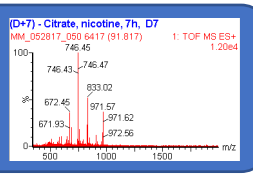


Unde mă aștept să fie localizată pe gelul SDS-PAGE din imagine proteina cu secvența:

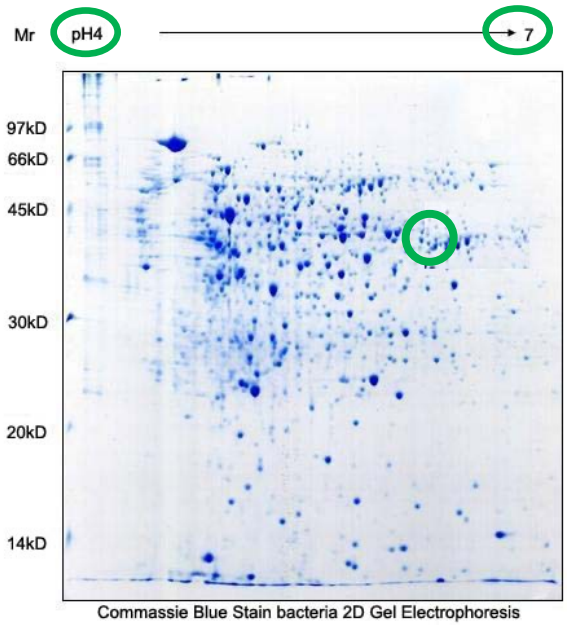
MTKTAIVRVAMNGITGRMGYRQHLLRSILPIRDAGGFTLEDGTKVQIEPILVGRNEAK
IRELAEKHKVAEWSTDLDSVVNDPTVDIIFDASMTSLRAATLKKAMLAGKHIFTEKPT
AETLEEAIELARIGKQAGVTAGVVHDKLYLPGLVKLRRLVDEGFFGRILSIRGEFGYW
VFEGDVQAAQRPSWNYRKEDGGGMTTDMFCHWNYVLEGIIGKVKSVNAKTATHIPTRW
DEAGKEYKATADDASYGIFELETPGGDDVIGQINSSWAVRVYRDELVEFQVDGTHGSA
VAGLNKCVAQQRAHTPKPVWNPDLPVTESFRDQWQEVPANAELDNGFKLQWEEFLRDV
VAGREHRFGLLSAARGVQLAELGLQSNDERRTIDIPEITL ??????

ProtParam: Molecular weight: 43189.17 Da
Theoretical pI: 5.6

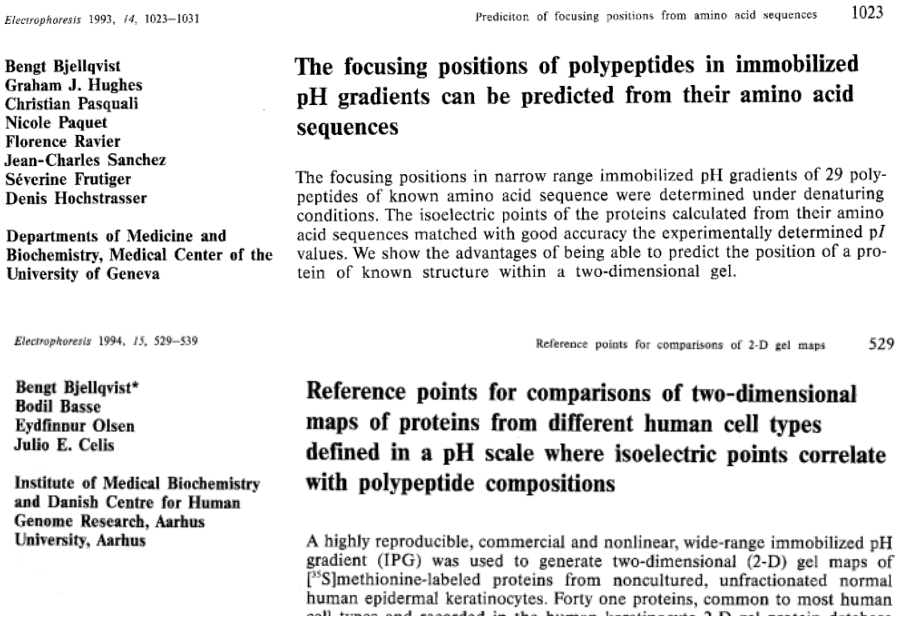
1. Calcularea proprietăților fizico-chimice simple: **ProtParam**



2. pI – punctul izoelectric – valoarea de pH a unei soluții în care molecula proteică dată nu se deplasează în câmp electrostatic deoarece sarcina sa electrică de suprafață este zero; **ProtParam calculează valoare pI luând în considerare valorile pKa ale aminoacizilor determinate practic la o concentrație de 9.2M sau 9.8M uree și 15°C sau 25°C;** util pentru localizarea unei proteine pe un gel 2D**;**



Commassie Blue Stain bacteria 2D Gel Electrophoresis


Electrophoresis 1993, *14*, 1023–1031

Prediciton of focusing positions from amino acid sequences 1023

Bengt Bjellqvist
Graham J. Hughes
Christian Pasquali
Nicole Paquet
Florence Ravier
Jean-Charles Sanchez
Séverine Frutiger
Denis Hochstrasser

Departments of Medicine and Biochemistry, Medical Center of the University of Geneva

The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences

The focusing positions in narrow range immobilized pH gradients of 29 polypeptides of known amino acid sequence were determined under denaturing conditions. The isoelectric points of the proteins calculated from their amino acid sequences matched with good accuracy the experimentally determined p*I* values. We show the advantages of being able to predict the position of a protein of known structure within a two-dimensional gel.

Electrophoresis 1994, *15*, 529–539

Reference points for comparisons of 2-D gel maps 529

Bengt Bjellqvist*
Bodil Basse
Eydfinnur Olsen
Julio E. Celis

Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, Aarhus

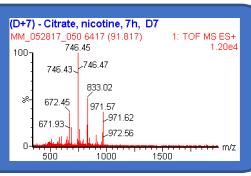
Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions

A highly reproducible, commercial and nonlinear, wide-range immobilized pH gradient (IPG) was used to generate two-dimensional (2-D) gel maps of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human epidermal keratinocytes. Forty one proteins, common to most human

Unde ar fi proteina anterioară cu pI 5,6 amplasată în gelul de mai sus?

În calcularea Mw și pI programul ProtParam **nu ține cont de eventualele modificări post-traducere** ale proteinelor (PTM: acetylation, myristoylation; palmitoylation, prenylation, alkylation, glycosylation, hydroxylation, etc). Proteinele ce conțin aceste modificări vor avea mase molare și valori pI diferite față de cele teoretice calculate de ProtParam și deci vor migra diferite pe gelurile de poliacrilamidă.

1. Calcularea proprietăților fizico-chimice simple: **ProtParam**



3. Timpul de înjumătățire estimat (In vivo half-life) – timpul necesar pentru ca jumătate din cantitatea de proteină de interes dintr-o celulă să dispară după ce a fost sintetizată;

ProtParam calculează acest parametru pornind de la așa numita **regulă a capătului N-terminal** (N-end rule) – timpul de înjumătățire a unei proteine depinde de **natura aminoacidului N-terminal**. Regula s-a stabilit pe baza unor observații experimentale ce au arătat că viteza de inactivare metabolică a unor beta-galactozidaze artificiale diferă funcție de aminoacidul terminal și de organismul în care sunt exprimate în limite foarte mari – 100 h până la mai puțin de 2 minute.

REVIEW

The N-end rule pathway of protein degradation

Alexander Varshavsky*

Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA

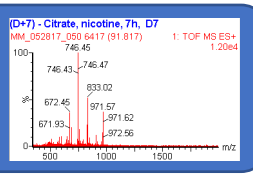
The N-end rule relates the *in vivo* half-life of a protein to the identity of its N-terminal residue. Similar but distinct versions of the N-end rule operate in all organisms examined, from mammals to fungi and bacteria. In eukaryotes, the N-end rule pathway is a part of the ubiquitin system. Ubiquitin is a 76-residue protein whose covalent conjugation to other proteins plays a role in many biological processes, including cell growth and differentiation. I discuss the current understanding of the N-end rule pathway.

Table 1 The N-end rule in *E. coli* and *S. cerevisiae*

	Half-life of X-βgal	
Residue X	*E. coli*	*S. cerevisiae*
Arg	2 min	2 min
Lys	2 min	3 min
Phe	2 min	3 min
Leu	2 min	3 min
Trp	2 min	3 min
Tyr	2 min	10 min
His	> 10 h	3 min
Ile	> 10 h	30 min
Asp	> 10 h	3 min
Glu	> 10 h	30 min
Asn	> 10 h	3min
Gln	> 10 h	10 min
Cys	> 10 h	> 30 h
Ala	> 10 h	> 30 h
Ser	> 10 h	> 30 h
Thr	> 10 h	> 30 h
Gly	> 10 h	> 30 h
Val	> 10 h	> 30 h
Pro	?	> 5 h
Met	> 10 h	> 30 h

Approximate *in vivo* half-lives of X-βgal proteins in *E. coli* at 36 °C (Tobias *et al.* 1991) and in *S. cerevisiae* at 30 °C (Bachmair & Varshavsky 1989). A question mark at Pro indicates its uncertain status in the N-end rule (see the main text).

1. Calcularea proprietăților fizico-chimice simple: **ProtParam**



4. Coeficientul de extincție – **cantitatea de lumină pe care proteina de interes o absoarbe la 280 nm**; este utilă pentru a putea măsura direct cu precizie concentrația unei proteine purificate folosind un spectrofotometru;

ProtParam calculează acest parametru ținând cont de faptul că absorbția la 280 nm a unei catene polipeptidice este dependentă de numărul de resturi de **tirozină** (Y), **triptofoan** (W) și **cistină** (două resturi de C legate printr-o punte disulfurică) după fromula:

Coef Extinctie Prot (280 nm) = Nr(Y)*Ext(Y) + Nr(W)*Ext(W) + Nr(Cistină)*Ext(Cistină); unde

Coef Extinctie Prot (280 nm) – coeficientul de extincție molară la 280 nm exprimat în M^{-1} cm^{-1}; **Nr(...)-** numărul de resturi de aminoacizi; **Ext(Y)** = 1490; **Ext(W)** = 5500; **Ext(Cistină)** = 125;

Exemplu de rezultate din ProtParam:

Extinction coefficients:

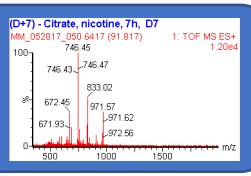
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 79700
Abs 0.1% (=1 g/l) 1.842, assuming all pairs of Cys residues form cystines

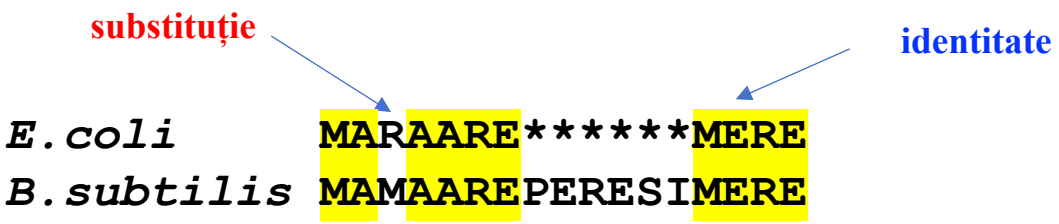
Când programul va afișa un singur coeficient de extincție?

Ext. coefficient 78950
Abs 0.1% (=1 g/l) 1.825, assuming all Cys residues are reduced

2. Identificarea proteinelor pe baza secvenței lor parțiale

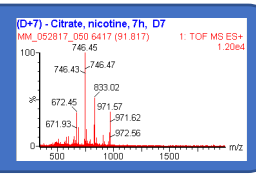


Bazele de date cu secvențe proteice reprezintă unul din factorii esențiali ce au permis dezvoltarea proteomicii pentru că **permit identificarea propriu-zisă a proteinelor plecând de la secvențele de peptide stabilite cu ajutorul spectrometriei de masă MS/MS**. În identificarea proteinelor pe baza secvenței lor parțiale un rol esențial îl are noțiunea de **aliniere de două sau mai multe secvențe.** Programul care s-a impus pentru identificarea și cuantificarea gradului de identitate și similaritate dintre două secvențe este **BLAST.**



Aliniere a două sau mai multe secvențe presupune așadar compararea celor două secvențe aminoacid cu aminoacid, aminoacidul 1 din secvența A este comparat cu aminoacidul corespunzător din secvența B. O corespondență între doi aminoacizi din aceeași poziție pe cele două secvențe poartă numele de **identitate**, iar o neconcordanță se numește **substituție.**

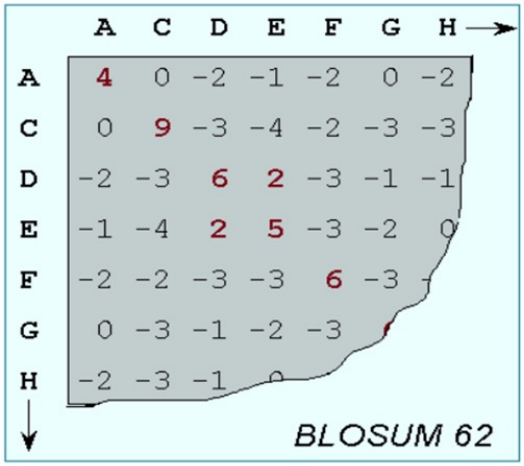
BLAST – identificarea de secvențe similare



BLAST - Basic Local Alignment Search Tool

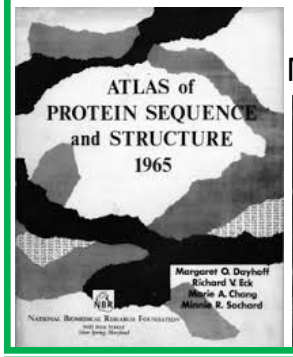
1. **identifică**, dintr-o bază de date, secvenţele similare cu o **secvenţă ţintă (tinta analizei, experimentului)**. Aceste secvenţe identificate poartă numele de **secvențe "subiect",** iar identificarea lor se bazează pe **alinieri locale**. Secvența „subiect este „suprapusă" peste cea țintă la nivelul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

2. **cuantifică nivelul de similaritate** dintre secvențele "subiect" și secvența țintă prin utilizarea unor **matrici de substituție.** O matrice de substituție arată frecvența cu care un aminoacid este înlocuit cu altul și are la bază observații experimentale (frecvențe de substituție măsurate din colecții de secvențe proteice).



298 Biologie moleculară. Metode experimentale

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.



Matrici PAM – **p**oint **a**ccepted **m**utation
Margaret Oakley Dayhoff, 1925-1983

Matrici BLOSUM - **BLO**cks **SU**bstitution **M**atrix

Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 10915–10919, November 1992
Biochemistry

Amino acid substitution matrices from protein blocks

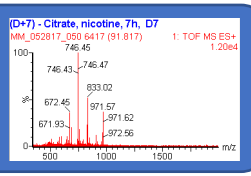
(amino acid sequence/alignment algorithms/data base searching)

STEVEN HENIKOFF* AND JORJA G. HENIKOFF

Howard Hughes Medical Institute, Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98104

Communicated by Walter Gilbert, August 28, 1992 (received for review July 13, 1992)

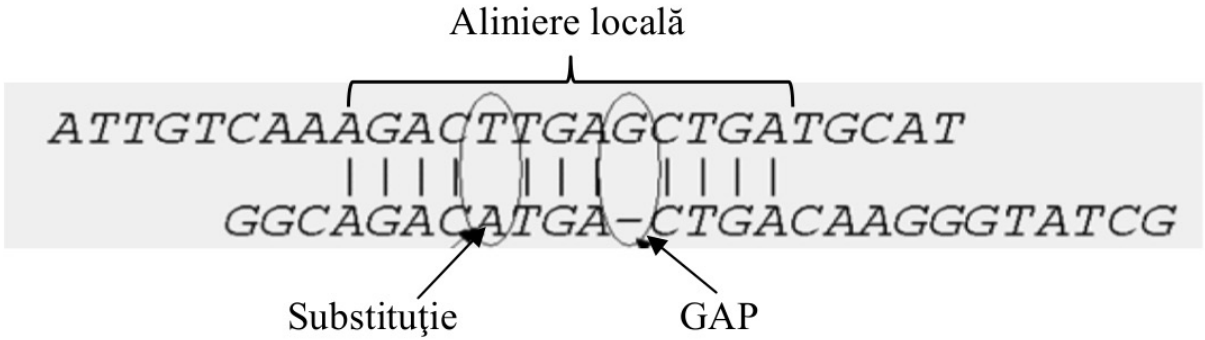
BLAST – identificarea de secvențe similare



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

Scoruri de similaritate calculate de BLAST:

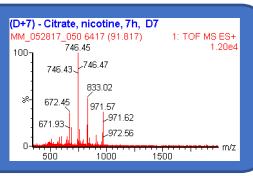
- **punctaj brut** (engl. Raw score) notat cu S, este calculat prin însumarea punctelor pentru fiecare pereche aminoacid-aminoacid, aminoacid-nimic şi penalizărilor pentru GAP; nu permite ierarhizarea secvențelor, valoare lui depinde de lungimea secvențelor analizate; cel mai frecvent nu este utilizat;



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

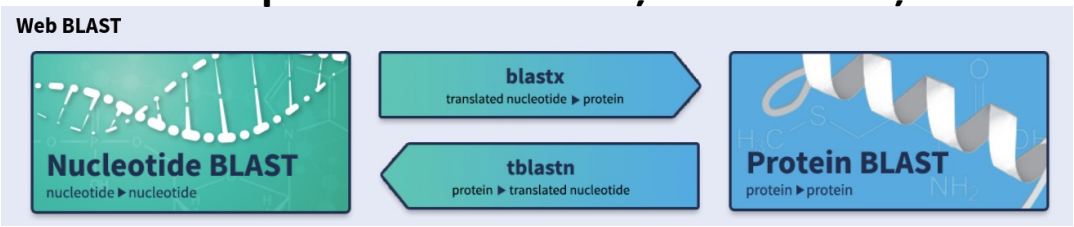
- **scorul în biţi notat cu S'** - se calculează prin normalizarea lui S în funcţie de diverse variabile statistice care depind, la rândul lor, de tipul de matrice utilizat. **Cu cât punctajul S' obţinut este mai mare cu atât asemănarea dintre cele două secvenţe este mai mare;**
- **parametru statistic 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

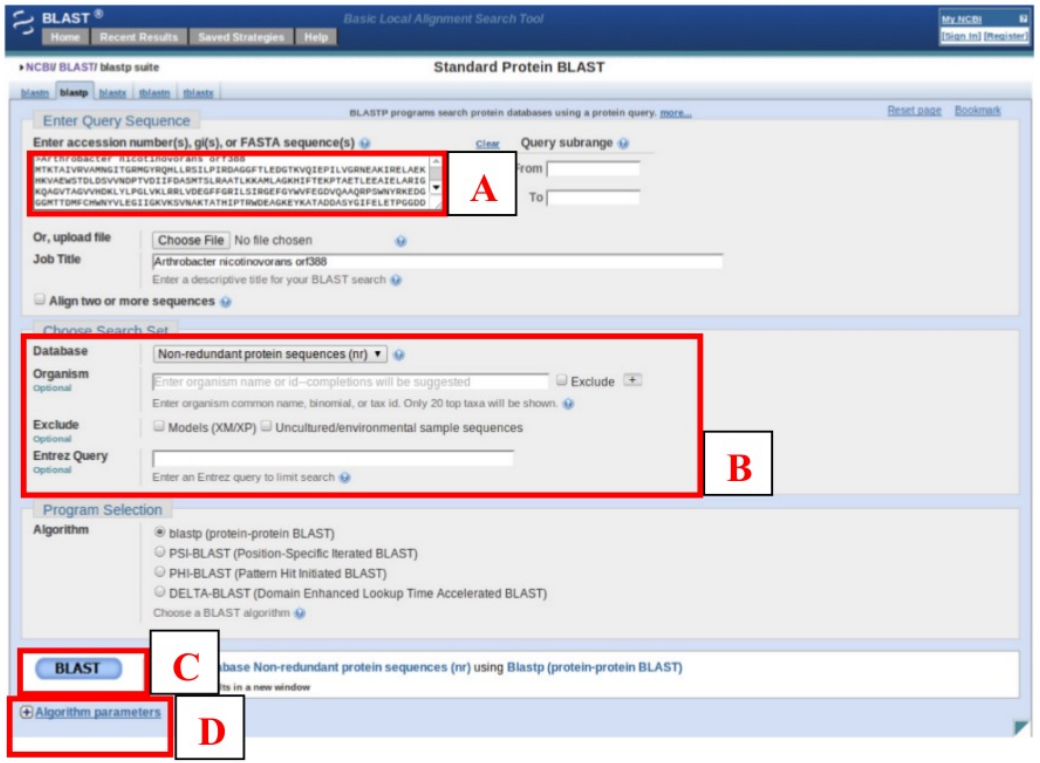
2. Selectează tipul de analiză funcție de secvența de interes:



Metode computaţionale 303

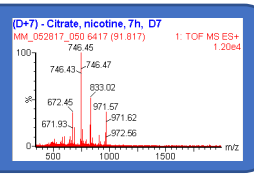
3. Copie secvența în căsuța pentru secvența țintă (query), setează parametrii căutării și apasă BLAST

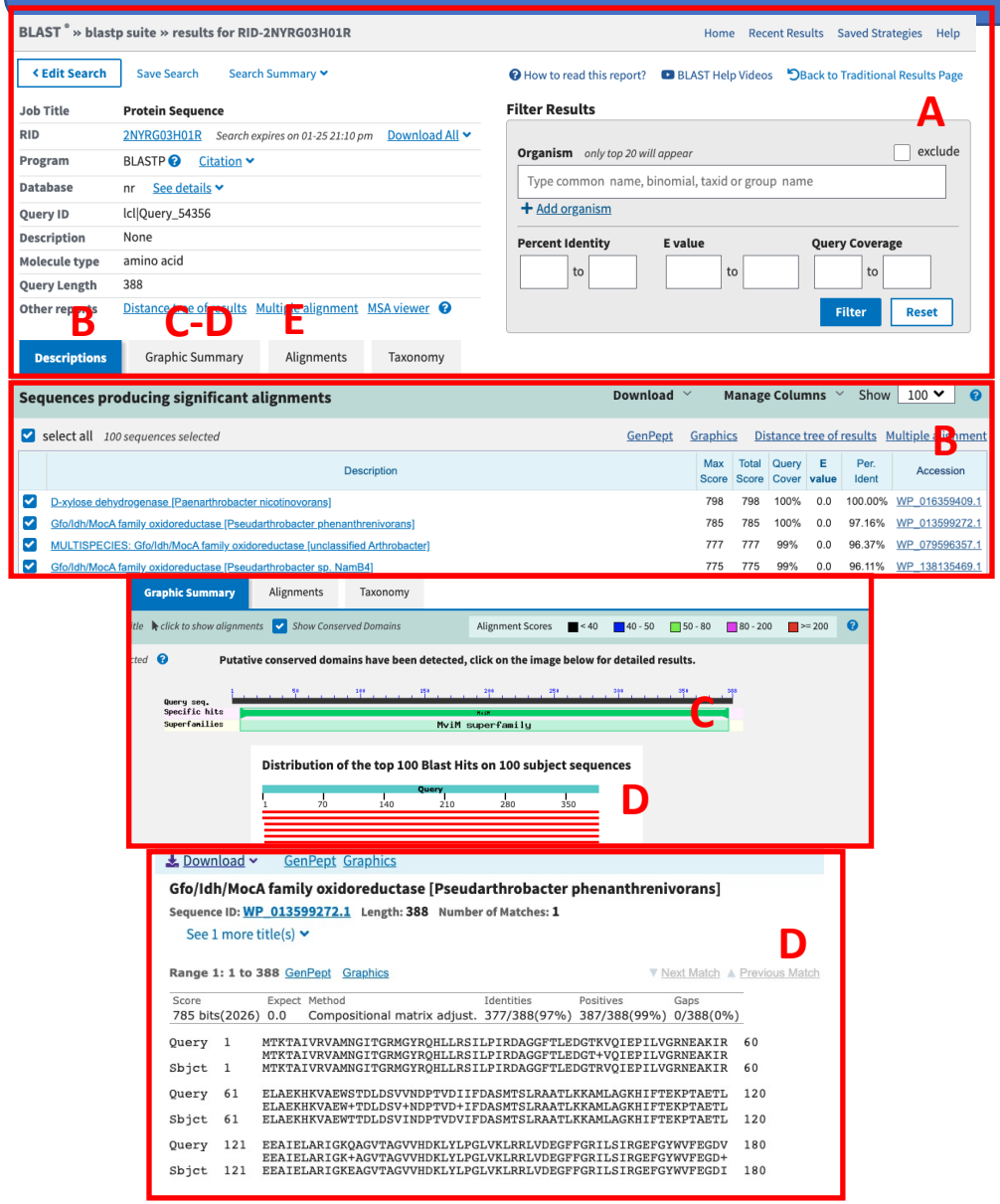
A – căsuţa text în care a fost inserată secvenţa ţintă în format FASTA;
B – zona cu parametrii utilizaţi pentru restrângerea spaţiului de căutare;
C – buton pentru lansarea investigării;
D – zona cu parametrii algoritmului de căutare



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

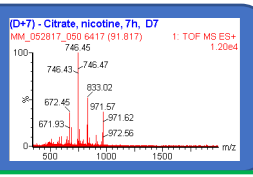
Rezultate BLAST





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

Rezultate BLAST și semnificația lor




```
MTVTSQVKPEDEMLNWGRLILDGVSYSDMVGARDRPKEITWFDYWMSLANEYEQEAERKVALGHDLSAGELLMSAALCAQYAQFLWFDERRQKGQARKVELY
QKAAPLLSPPAERHELVVDGIPMPVYVRIPEGPGPHPAVIMLGGLESTKEESFQMENLVLDRGMATATFDGPGQGEMFEYKRIAGDYEKYTSAVVDLLTKLE
AIRNDAIGVLGRSLGGNYALKSAACEPRLAACISWGGFSDLDYWDLETPLTKESWKYVSKVDTLEEARLHVHAALETRDVLSQIACPTYILHGVHDEVPLSF
VDTVLELVPAEHLNLVVEKDGDHCCHNLGIRPRLEMADWLYDVLVAGKKVAPTMKGWPLNG
```

Sequences producing significant alignments Download Manage Columns Show 100

select all *29 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
RecName: Full=D-xylose dehydrogenase; AltName: Full=NADP-dependent D-xylose dehydrogenase [Paenarthrobacter nicotinovorans]	798	798	100%	0.0	100.00%	Q8GAK6.1
RecName: Full=Uncharacterized oxidoreductase sll0816 [Synechocystis sp. PCC 6803 substr. Kazusa]	75.9	75.9	85%	6e-14	24.56%	P74041.1
RecName: Full=Uncharacterized oxidoreductase YrbE [Bacillus subtilis subsp. subtilis str. 168]	72.8	72.8	69%	5e-13	28.00%	O05389.2
RecName: Full=Uncharacterized oxidoreductase YjhC [Escherichia coli K-12]	68.9	68.9	32%	1e-11	32.00%	P39353.2
RecName: Full=Uncharacterized oxidoreductase YhhX [Escherichia coli K-12]	66.2	66.2	31%	7e-11	27.64%	P46853.1
RecName: Full=1,5-anhydro-D-fructose reductase; Short=Anhydrofructose reductase; AltName: Full=1,5-anhydro-D-fructose reductase (1,5-an	64.7	64.7	36%	2e-10	29.86%	Q2I8V6.1
RecName: Full=scyllo-inositol 2-dehydrogenase (NADP(+)) IolW; AltName: Full=NADP(+)-dependent scyllo-inositol dehydrogenase 1; Short=N	60.5	60.5	45%	7e-09	25.42%	O32223.1
RecName: Full=1,5-anhydro-D-fructose reductase; Short=Anhydrofructose reductase; AltName: Full=1,5-anhydro-D-fructose reductase (1,5-an	60.1	60.1	36%	8e-09	28.47%	Q92KZ3.1
RecName: Full=scyllo-inositol 2-dehydrogenase (NAD(+)) [Bacillus subtilis subsp. subtilis str. 168]	59.7	59.7	42%	1e-08	30.06%	P40332.2
RecName: Full=Uncharacterized oxidoreductase YdgJ [Escherichia coli K-12]	56.2	56.2	39%	2e-07	25.95%	P77376.2
RecName: Full=Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase; AltName: Full=D-xylose 1-dehydrogenase; AltName: Full=D-xylose-NADP (	53.5	53.5	75%	1e-06	20.07%	Q7JK39.1
RecName: Full=Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase; AltName: Full=D-xylose 1-dehydrogenase; AltName: Full=D-xylose-NADP (	53.1	53.1	75%	1e-06	18.64%	Q148L6.1
RecName: Full=Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase; AltName: Full=D-xylose 1-dehydrogenase; AltName: Full=D-xylose-NADP (	52.4	52.4	75%	2e-06	21.09%	Q9UQ10.1
RecName: Full=Inositol 2-dehydrogenase/D-chiro-inositol 3-dehydrogenase; AltName: Full=Myo-inositol 2-dehydrogenase/D-chiro-inositol 3-de	48.9	48.9	29%	3e-05	32.48%	A5YBJ7.1
RecName: Full=Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase; AltName: Full=Can2DD; AltName: Full=D-xylose 1-dehydrogenase; AltNam	48.9	48.9	53%	3e-05	19.71%	Q9TV68.1
RecName: Full=Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase; AltName: Full=D-xylose 1-dehydrogenase; AltName: Full=D-xylose-NADP (	48.5	48.5	39%	4e-05	24.05%	Q642M9.2
RecName: Full=Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase; AltName: Full=D-xylose 1-dehydrogenase; AltName: Full=D-xylose-NADP (	48.1	48.1	75%	5e-05	19.39%	Q5R5J5.1