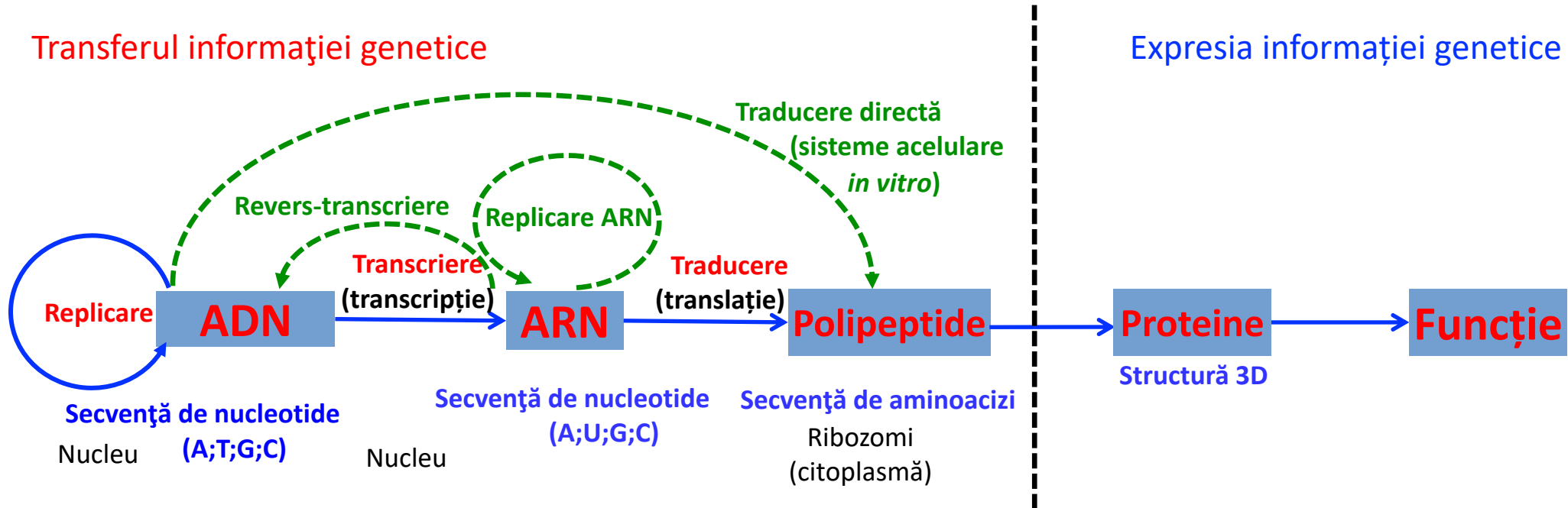
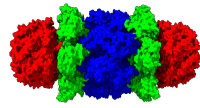
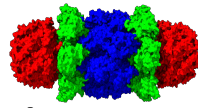


Capitolul V

Modificarea mesajului genetic

II. Secvențierea acizilor nucleici și analiza secvențelor





Secvențierea ADN-ului - procesul de identificare a ordinii precise a nucleotidelor (ATCG) într-o moleculă de ADN – structura primara a ADN-ului.

În prezent există numeroase metode de secvențializare a ADN-ului clasificate în:

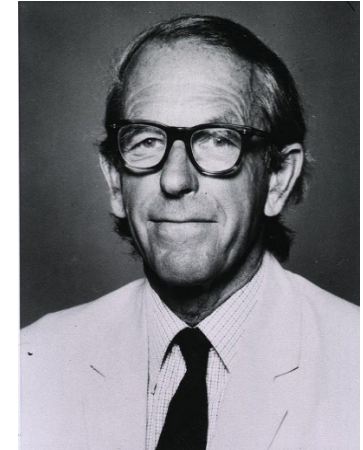
- metode derivate de la **metoda chimică** a lui **Maxam și Gilbert**;
- metode derivate de la **metoda enzimatică** a lui **Sanger**;

Metoda enzimatică Sanger

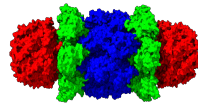
- una din primele metode de secvențializare a fost descrisă în 1977 de Frederick Sanger care a primit cel de-al doilea premiu Nobel pentru descoperirea sa;
- se bazează tot pe reacția de replicare a ADN-ului catalizată de o **polimerază ADN-dependentă**;

Elementele esențiale realizării unei reacții de secvențiere prin metoda Sanger

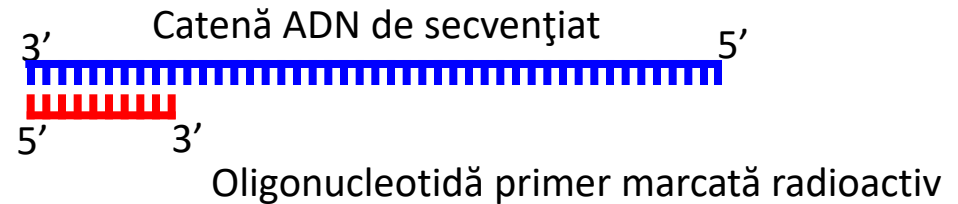
- a) **fragment ADN de secvențiat** – o moleculă de ADN monocatenară;
- b) **polimeraza ADN**;
- c) cele **4 deoxinucleotide** obișnuite (dNTP: A, T, C, G);
- c) o **oligonucleotide amorsă** (primer) complementare cu molecula de secvențiat marcată radioactiv și care oferă un capăt 3' liber;
- d) **4 di-deoxinucleotide** (ddNTPs: ddA, ddC, ddG, ddT) – nucleotide ce nu au gruparea hidroxil în pozițiile **2'** (**deoxi**) și **3'** (**di**).




Frederick Sanger - 1918 - 2013



1) Legarea oligonucleotidului primer




2) **Elongarea** primerului prin acțiunea ADN polimerazei în prezența dNTP și ddNTP cu sinteza unei catene noi. Reacția se realizează în **4 eprubete diferite**, fiecare conținând un singur tip de ddNTP (ddATP, ddGTP, ddCTP sau ddTTP). **Încorporarea unui ddNTP în noua catenă face ca sinteza acesteia să se oprească** (ddNTP sunt "terminatori" de catenă). Încorporarea unui ddNTP de către ADN-polimeraza este un proces statistic ceea ce face ca prin elongare să se genereze în fiecare eprubetă câte o colecție de catene ADN nou sintetizate de dimensiuni diferite.



ddA +
A; T; C; G;
#TCGACGGGC
Amorsa #

#ddA

#AGCTGCCCG



ddC +
A; T; C; G
#TCGACGGGC
Amorsa #

#AGddC

#AGCTGddC

#AGCTGCddC




ddG +
A; T; C; G
#TCGACGGGC
Amorsa #

#AddG

#AGCTddG

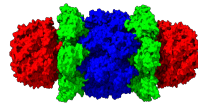
#AGCTGCCddG



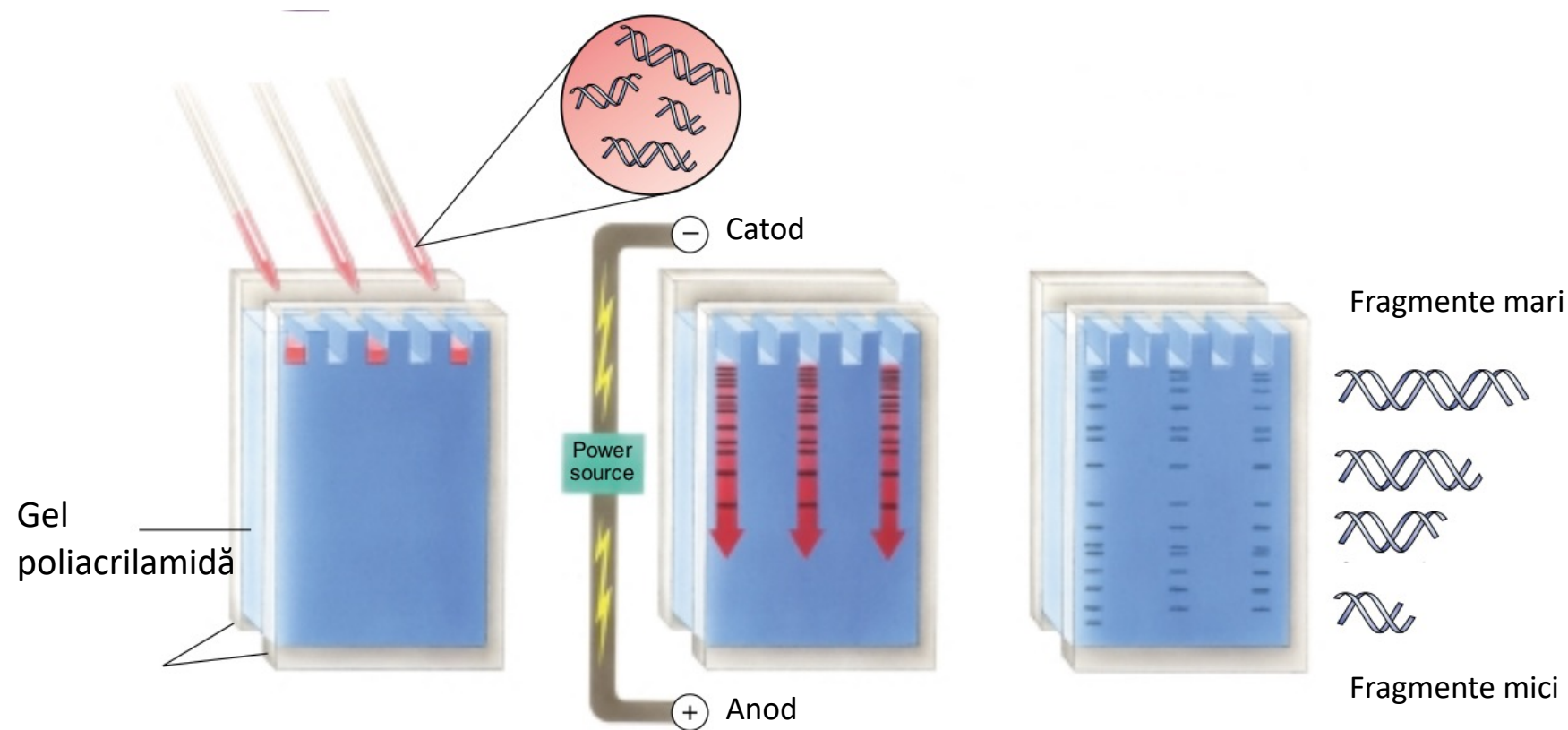
ddT +
A; T; C; G
Amorsa #

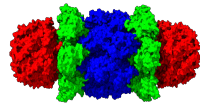
#AGCddT

#AGCTGCCCG

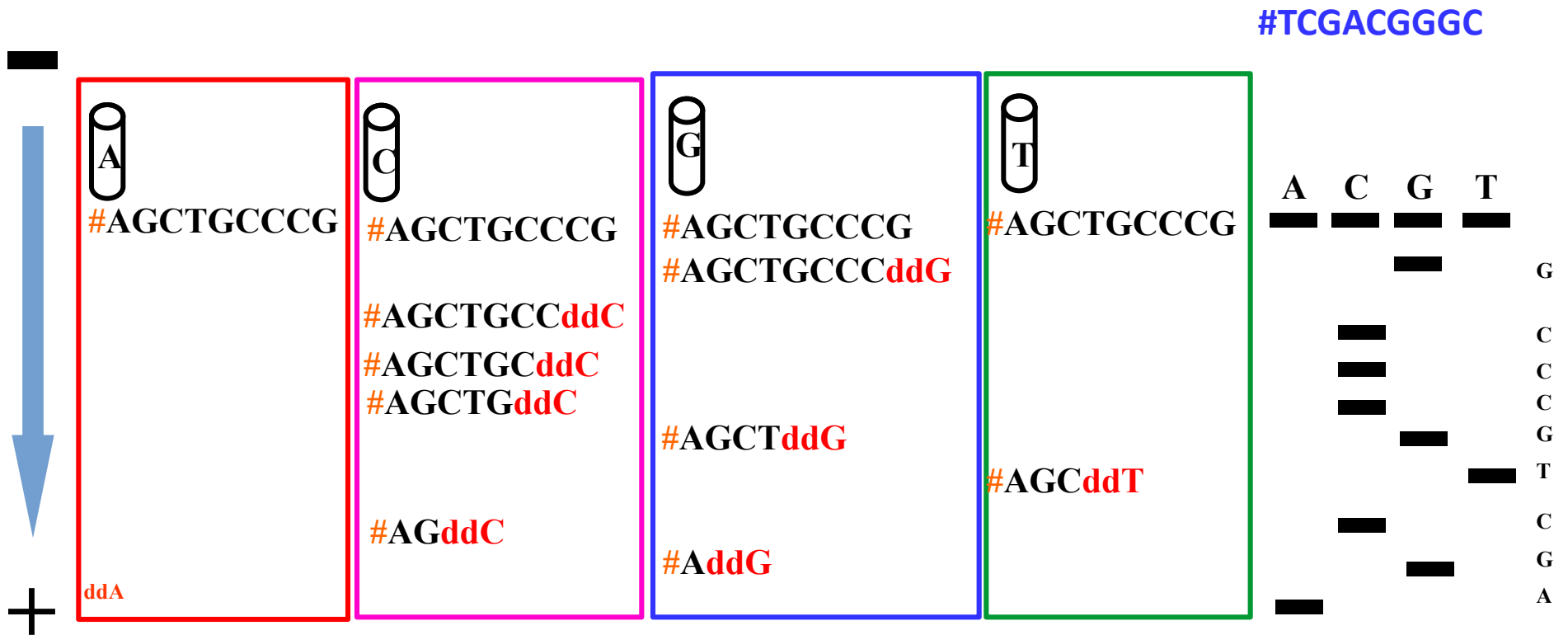


3) Separarea catenelor nou sintetizate funcție de dimensiune – se realizează prin **electroforeză** – moleculele de ADN sunt încărcate negativ și la aplicarea unui curent vor migra spre polul pozitiv. Electroforeza se realizează în geluri de poliacrilamidă ce acționează ca o sită, frânând moleculele mari, astfel încât moleculele mici vor migra mai mult iar cele mari mai puțin.

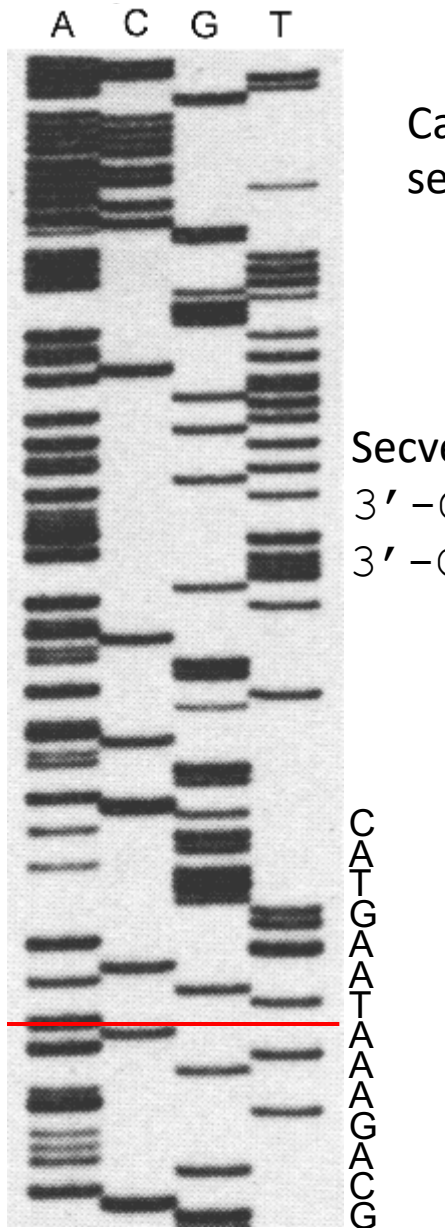
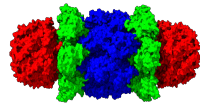




4) **Vizualizarea catenelor nou sintetizate** – se realizează prin **autoradiografie** – pe baza marcajului radioactiv al oligonucleotidului amorsă.



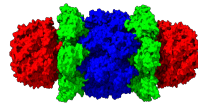
Schema autoradiografiei unui gel de secvențializare



Care este secvența fragmentului de ADN prin a cărui secvențiere Sanger se obține gelul de alături?

Secvența citită de pe gel:

3' -GCAGAAATAAGTAC-5' deci secvența catenei de secvențiat era:
3' -CGTCTTTATTCATG-5' ,
respectiv 5' G T A C T T A T T T C T G C -3'



Etapele reacției de secvențiere

O variantă a metodei Sanger este:

– **utilizarea de "terminatori" de catenă marcați fluorescenți.** Fiecare din cele 4 ddNTP sunt marcate cu câte un fluorocrom diferit (**ddA – verde**, **ddT – roșu**, **ddG – galben**, **ddC – albastru**). Reacția se desfășoară într-o singură eprubetă, iar fragmentele generate sunt separate prin electroforeză. Ordinea nucleotidelor este citită cu ajutorul unui laser și reprezentată ca "peak-uri".

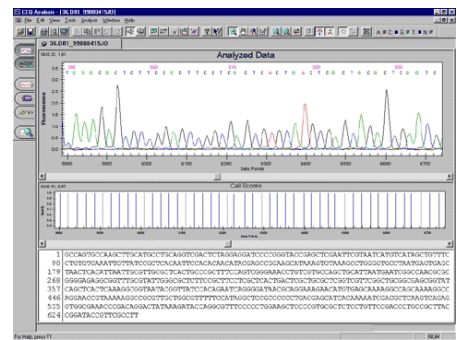
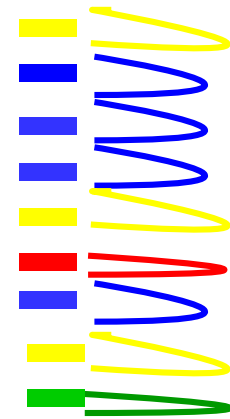


- #ddA
- #A ddG
- #AG ddC
- #AGC ddT
- #AGCT ddG
- #AGCTG ddC
- #AGCTGC ddC
- #AGCTGCC ddC
- #AGCTGCCC ddG
- #AGCTGCCCG

Separare
funcție de
dimensiune



- #AGCTGCCC ddG
- #AGCTGCC ddC
- #AGCTGC ddC
- #AGCTG ddC
- #AGCT ddG
- #AGC ddT
- #AG ddC
- #A ddG
- #ddA



Etapele reacției de secvențiere

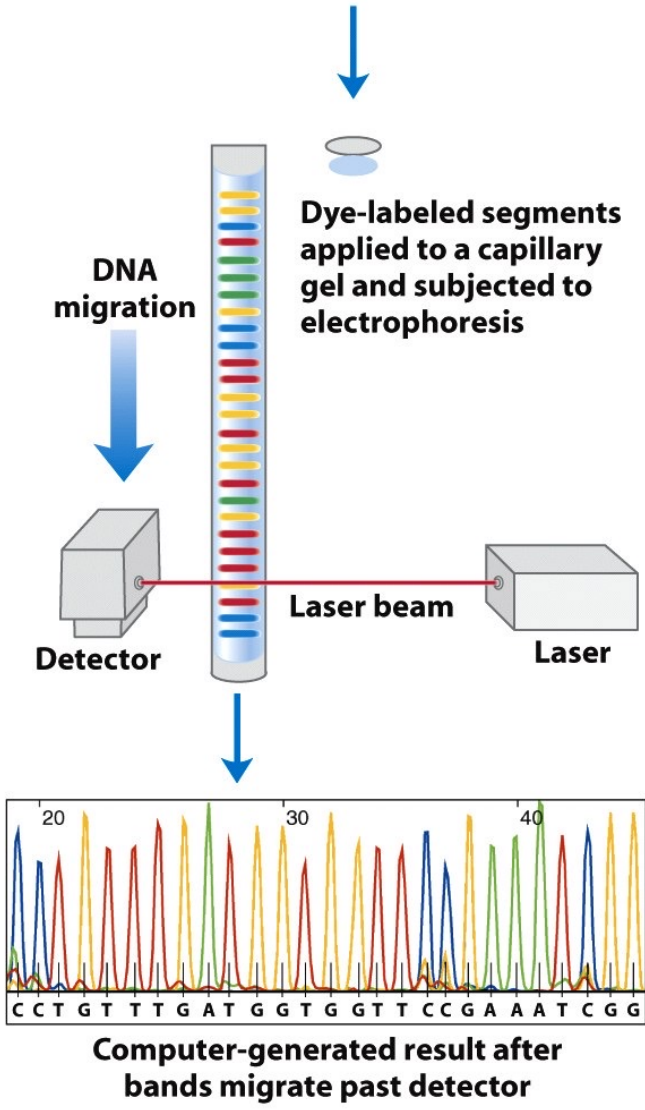
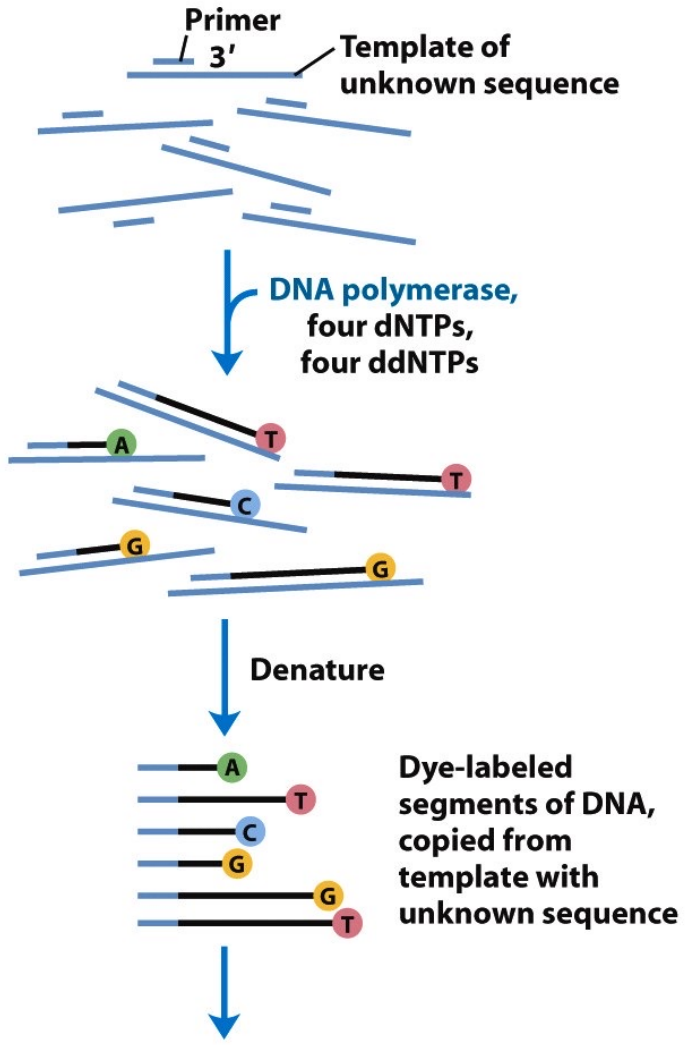
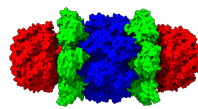
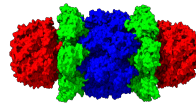


Figure 8-34
Lehninger Principles of Biochemistry, Fifth Edition
 © 2008 W. H. Freeman and Company

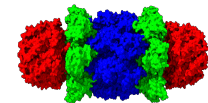


Nevoia de a secvenția cantități mari de ADN (**genomul uman - 6.27 Gpb, un genom bacterian 4-5 Mpb**) rapid și cu costuri reduse a dus la apariția unor noi metode de secvențiere, ce funcționează după principii diferite de metoda Sanger. Aceste metode au fost numite general **next generation sequencing / Nextgen / metode NGS / metode de nouă generație**. Primele metode Nextgen au apărut în ani 80. De atunci s-au dezvoltat succesiv două generații de metode de secvențiere și este de așteptat ca noi metode să fie dezvoltate în viitor. De aceea termenul **Nextgen** a devenit ambiguu și se preferă înlocuirea sa cu:

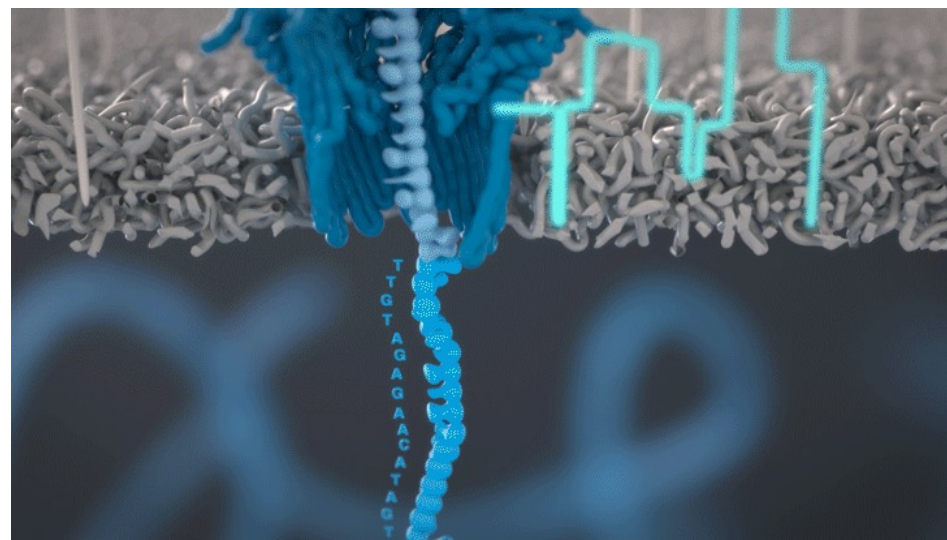
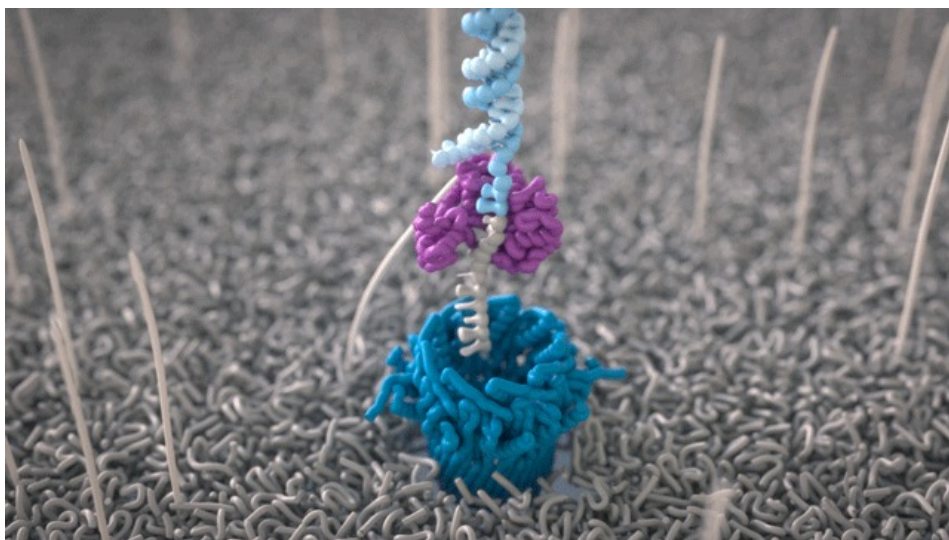
1. Metode de secvenție de generația a **II** / secvențiere masivă paralelă / metode de secvențiere a fragmentelor scurte: **secvențierea prin hibridizare** și **secvențierea prin sinteză** (SBS – **sequencing by synthesis**)
2. Metode de secvenție de generația a **III** / metode de secvențiere a fragmentelor lungi: **SMRT** (Single Molecule Real Time) și secvențiere folosind nanopori proteici **Oxford Nanopore Technologies**

Toate aceste metode permit generarea unui număr semnificativ mai mare de nucleotide secvențiate comparativ cu metoda Sanger, de aceea ele mai sunt denumite și metode **metode extensive (High-throughput)**

Detalii despre **secvențierea prin hibridizare, secvențierea prin sinteză și SMRT** (Single Molecule Real Time) in cursul: Tehnici de Biologie Moleculară, Master Laborator Medical.

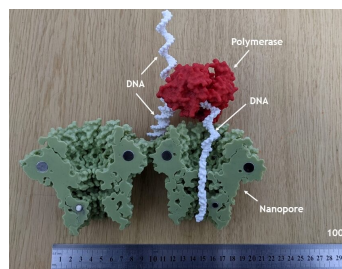


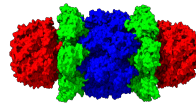
Secvențiere folosind nanopori proteici – folosește proteine transmembranare ce pot fi traversate de molecule de acizi nucleici precum alfa-hemolisină și porina A din *Mycobacterium smegmatis* (MspA). Moleculele de ADN trec prin por cu o viteză de constantă controlată de ADN polimerază phi29 și duc la modificarea diametrului interior al porului. Modificările sunt dependente de forma nucleotidei ce trece și sunt înregistrate sub forma unor curenți transmembranari. Funcție de caracteristicile curentului generat de trecerea unei molecule de ADN, se poate detecta secvența acesteia.



<https://nanoporetech.com/how-it-works>

<https://modele moleculare.ro/nanopor-proteic/>





Spre deosebire de metodele de secvențiere a proteinelor, stabilirea ordinii nucleotidelor în acizii nucleici ADN și ARN este mult mai rapidă și ieftină. În ultimii ani metoda clasică Sanger a fost înlocuită de metode de `generație nouă` (next-gen) ce au un randament și o viteză foarte mare (high-throughput sequencing methods).

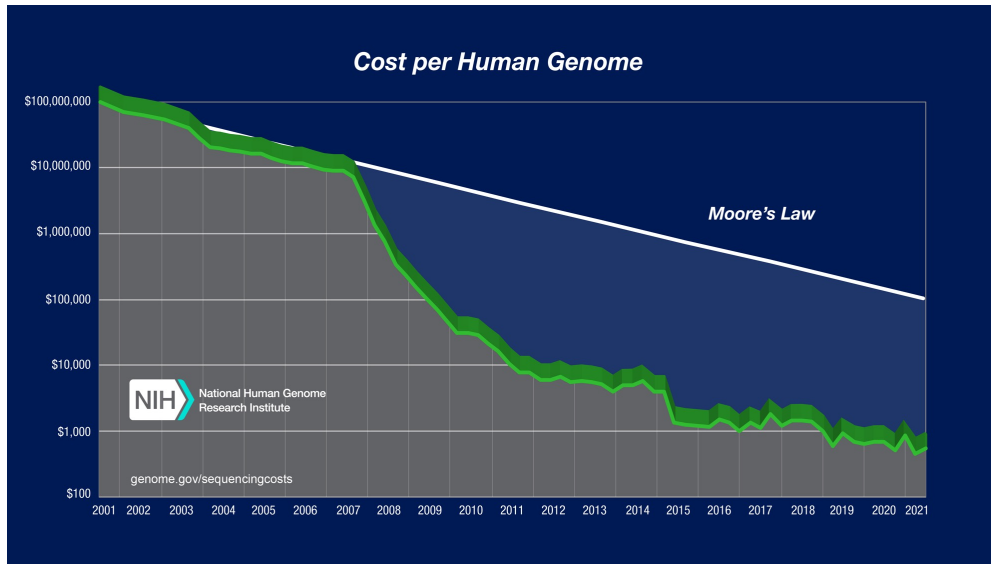
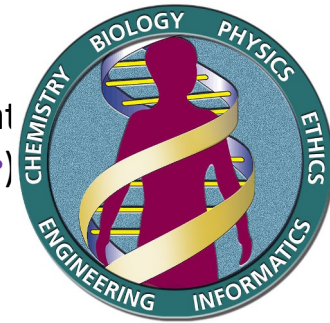
Secvențierea genomului uman - 3 bilioane de baze - **~2.7 bilioane \$**

Demarat în 1990 proiectul internațional Human Genome Project (HGP) ce a fost declarat încheiat în 2003 și a reușit secvențierea unui set haploid de cromozomi umani (*haploid reference genome*)

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 \$**



Mardis Genome Medicine 2010, 2:84
<http://genomemedicine.com/content/2/1/84>



MUSINGS

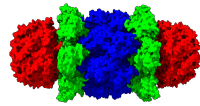
The \$1,000 genome, the \$100,000 analysis?

Elaine R Mardis*

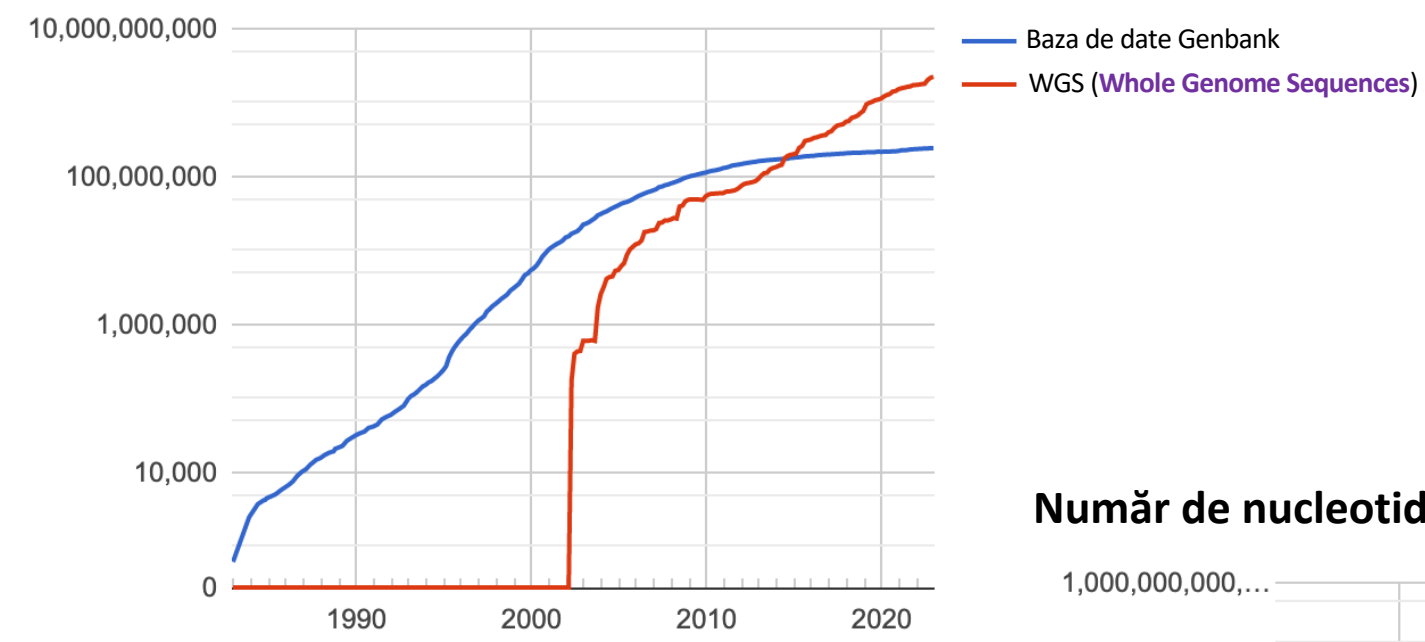
Bioinformatica – ansamblul de metode și tehnici ce permit stocarea, manipularea și interpretarea informației genetice – definiția restrânsă !!!!

<https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>

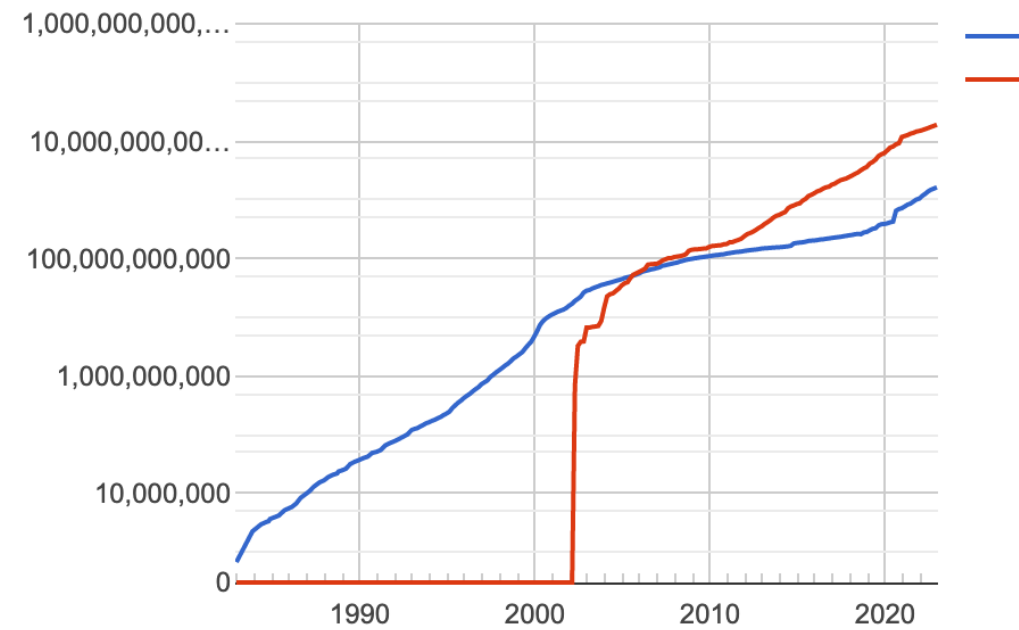
Câte secvențe sunt disponibile?

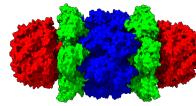


Număr de secvențe cunoscute



Număr de nucleotide în secvențe cunoscute





O **secvență**, fie că este ADN, ARN sau proteine **reprezintă o înșiruire de litere** (A; T; G;C pentru ADN, A;U;G;C pentru ARN, cei 20/22 de aminoacizi notați cu o literă pentru proteine) **ce poate fi foarte ușor stocată într-un fișier text**.

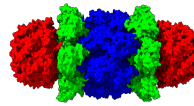
În general o secvență este însoțită de o serie de **informații accesorii** precum **specia de la care provine secvența, gena, cromozomul, lucrarea în care este descrisă secvența**. **Modul în care sunt organizate și stocate aceste informații alături de secvența propriu-zisă reprezintă formatul sau standardul 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.

Fișierul FASTA – Fast-All

```
>AJ507836.1 Arthrobacter nicotinovorans pA01 megaplasmid sequence, strain ATCC 49919
GATCCGGCGGTCCGCCGTCGTGGCGGCGGGCGGAGGTTGCCGGCGGCGGACCGCCCGCCGACAAGAA
GGCCTTCGGGTTCCGGCAGGTGGCGCGGCCCGGACACTGGTCCTCGCCTGGGGTGGAAACGTGGGGTGCT
GGGTGTGGGCGGTTCCGCCGGCGAACC GGAAAGGCGTCCACTCCTCTTCGCTTCCGTGGCCGGCGGTTG
GGGCCGGTCCCGTCGTTCGAGAGCTCCGCCGTGCCCGGCCCGCCCGTTGTCTACGACGTCTTTTTGTGG
CTGTCCTTCGGATCATAACGGTTCGGCTCCAGCTCAAGCTCCTGGCAGTCCGCAAAGCTCCGGTCTAGCAC
ACAGCGATGCGGGTAATGATGGCGACGAATTCGTCCAGAGCGCTGGTGGGATTTCGTGGCCCATACCTG
```

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILLLLLLLLLLALLSPDMLGDPDNHMPADPLNTPHMKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFPLPIAGX
IENY
```



Fișierul FASTA

- un fișier text ce poate **avea sau nu** extensia .fasta;
- conține secvențe de nucleotide sau de aminoacizi stocate conform **standardului FASTA**:

Elementele formatului FASTA:

A.primul rând de text, marcat cu “>” (mai mare) conține o serie de **informații cu caracter opțional**, - specia sau denumirea genei (proteinei);

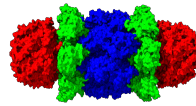
B.urmatoarele rânduri conțin secvența propriu-zisă, în care nucleotidele/aminoacizii sunt reprezentați folosind codul standard IUPAC cu o singură literă;

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 intermediară
Y	T/C (pYrimidine)		

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

Standardul IUPAC pentru notarea nucleotidelor și aminoacizilor



Fișierul FASTA

- 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 PO₄ liberă, ultima literă de pe ultimul rând are gruparea OH 3' liberă. Moleculele circulare se reprezintă linear, prima nucleotidă fiind cel mai frecvent originea de replicare;
- E. O secvență polipeptidică 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, numărarea 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**;

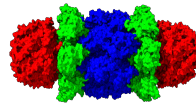
```
>secventa peptidica necunoscuta 1 fara numere
QIKDLLVSSSTDLDLTLVLVNAIYFKGMWKTAFAEDTREMPPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPPFASGDLMLVLLPDEVSDLERIEKTINFEKLTWETNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPSESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

```
>secventa peptidica necunoscuta cu numere
1   qikdllvsss tlddtltrlv naiyfkgmwk tafnaedre mpfhvtkqes kpvqmmcmnn
61  sfnvatlpae kmkilelpfa sgdlsmlvll pdevsdleri ektinfeklt ewtnpntmek
121 rrvkvylpqm kieekynlts vlmalgmtdl fipsanltgi ssaeslksq avhgafmels
181 edgiemagst gviedikhsp eseqfradh flflihnpt ntivyfgryw sp
```

Cum scriu secvențele?
- Cu font **monospațiat**

TNR **Proportional**
Courier New **Monospace**

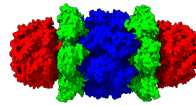
De unde pot obține secvențe?



1. Un instrument de secvențiere a ADN-ului – cel mai frecvent;
2. Un instrument de secvențiere a proteinelor bazat pe degradarea Edman (cel mai puțin frecvent) sau un spectrometru de masă – destul de rar;
3. **O bază de date cu secvențe** – cel mai ușor de accesat; create de utilizatori ce au acces la instrumentele 1 și 2 și deci pot determina experimental o secvență;

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**. 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, accesul la bazele de date cu secvențe este gratuit.

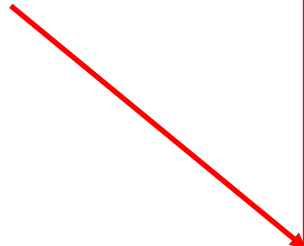
Bază de date	Site web	Dimensiune *
Baze de date cu secvențe		
INSDC	http://www.insdc.org/	206 293 625 secvențe
DDBJ	http://www.ddbj.nig.ac.jp	Aceste baze de date sunt sincronizate zilnic între ele și conțin aceleași secvențe.
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl/	
GenBank	http://www.ncbi.nlm.nih.gov/genbank/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



Operații simple cu secvențe

<http://www.bioinformatics.org/sms2/>

Operații cu secvențe
selectabile în SMS2



SMS

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

Sequence Manipulation Suite:

Translate

Translate accepts a DNA sequence and converts it into a protein in the reading frame you specify. Translate supports the entire IUPAC alphabet and several genetic codes.

Paste a raw sequence or one or more FASTA sequences into the text area below. Input limit is 200000 characters.

```
>Arthrobacter nicotinovorans orf388 sequence
ctactaacctcagcgccccccttcaaccttttctctgaggaaaatagactaaaca
gcgattgtccgctgcctatgaatggcatcaccggccggatgggctaccgccagcacctg
ccgcgacccggcggcttcacctcgaagacggcaccgaagtcagatcgaaccgatcct
cgtaggccgaacgaagcaagatcccggaactcggcgaagcacaaggtgcccagtg
gagcacggacctggactcggctcgtcaacgacccccaccgtcgacatcatcttcgacgctc
catgaccagcctccgcgccaccctgaagaaggcgtgctggccggaagcacatctt
caccgagaagccaccggaaaacctggaagaggccattgaactggcccgcatcggcaa
gcaggcaggcgtcaccgcaggcgtcgtacacgacaagctgtacctccgggctggtaa
gctccgcccctggtagcgaaggtctcttcggccgcatcctgtccatcccggtgagtt
cggctactgggtcttgaaggtgacgttcaggcagcacagcgccgctcctggaactaccg
caaggaagcggcgggtggaatgaccacggacatgtctgccactggaactacgtccttga
aggcatcatcggcaaggtcaagagcgtcaacgccaagaccgccacacatccccaccg
ctgggacgaagccggaagagtagaaggaacggctgatgacgcttctacggcatctt
cgagcttgaaacccccggcggcagcagcgtcatcggccagatcaactcttctgggcccgt
ccgctctaccgcagcaactcgtcgaattccaggtggacggcaccacggctccgcccgt
tgccggcctgaacaagtgcgtcgcaccagcagcgcacacaccccccaagccggtctggaa
ccctgacctgccctcaccgaatccttccgcgaccantgcaagaaatccccccaacgc
```

Please check the browser compatibility page before using this program.

- Translate in reading frame 1 on the direct strand.
- Use the standard (1) genetic code.

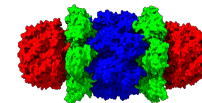
*This page requires JavaScript. See [browser compatibility](#).
*You can [mirror this page](#) or use it off-line.

[new window](#) | [home](#) | [citation](#)

Fri Mar 30 17:46:57 2012

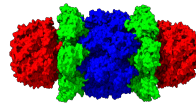
Căsuță text pentru
secvență FASTA





Denumire operație*	Descriere
Convertire între diverse formate	
Combine FASTA	– permite combinarea a două sau a mai multe secvențe în format FASTA și obținerea unei singure secvențe
EMBL to FASTA	– permite convertirea unei secvențe din formatul EMBL în formatul FASTA; funcția este utilă în situația în care se dorește eliminarea rapidă a informațiilor care nu au legătură cu secvența de ADN dintr-un fișier EMBL
Filter DNA	– elimină caracterele care nu corespund codului standard IUPAC pentru ADN-ul dintr-o secvență (inclusiv spații, numere, etc)
Filter Protein	– elimină caracterele ce nu corespund cu codul standard IUPAC de o literă pentru aminoacizi dintr-o secvență (inclusiv spații, numere etc.)
GenBank to FASTA	– permite convertirea unei secvențe din formatul GenBank în formatul FASTA; funcția este utilă în situația în care se dorește eliminarea rapidă a informațiilor care nu au legătură cu secvența de ADN dintr-un fișier GenBank
One to Three	– permite convertirea unei secvențe de aminoacizi din codul standard IUPAC de o literă în cel de trei litere
Analiza secvențelor	
Codon Plot	– analizează frecvența de utilizare a codonilor dintr-o secvență de ADN și generează un grafic cu aceste frecvențe
Codon Usage	– analizează frecvența de utilizare a codonilor dintr-o secvență de ADN și poate fi utilizat pentru a identifica preferința pentru un anumit codon sinonim
DNA Molecular Weight	– calculează masa moleculară a uneia sau a mai multor secvențe de ADN
DNA Stats	– calculează frecvența fiecărei nucleotide într-o secvență dată, rezultatele fiind exprimate în procente
PCR Primer Stats	– analizează și calculează proprietățile specifice ale unui set indicat de primeri, inclusiv temperatura de topire și procentul de GC
PCR Products	– realizează amplificarea virtuală prin PCR a unei secvențe date folosind un set de primeri indicat de utilizator
Protein Isoelectric Point	– calculează valoarea teoretică a punctului izoelectric pentru o secvență dată
Protein Molecular Weight	– calculează valoarea teoretică a masei moleculare pentru una sau mai multe secvențe de aminoacizi
Protein Stats	– calculează frecvența fiecărui aminoacid într-o secvență dată, rezultatele fiind exprimate în procente
Translate	– realizează translația și transcripția virtuală a unei secvențe de ADN date, utilizatorul având posibilitatea de a alege cadrul de citire specific
Reverse Translate	– realizează procesul invers față de Translate, acceptând o secvență de aminoacizi și generând secvența ADN corespunzătoare

* corespunzătoare cu denumirea din suita de programe SMS2



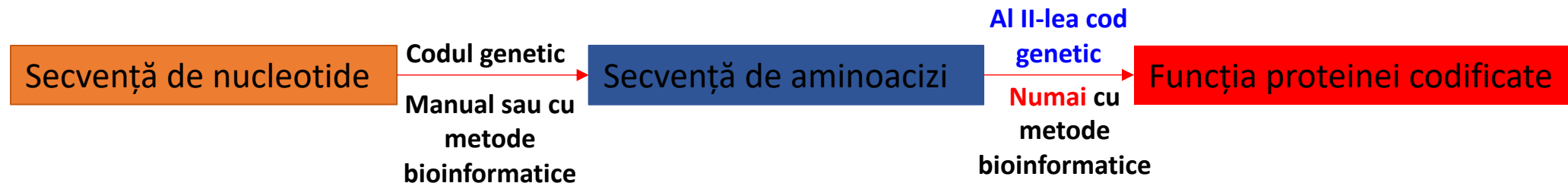
Folosind metodele experimentale de secvențiere enumerate anterior, se stabilește secvența unei gene. Aceasta codifică următoarea secvență de aminoacizi:

MAAKYRIGYFVGSLATGSINRVLSQALINLAPEDLEFSEIPIRDLPLYSDYDADFPPEGR

Care este funcția acestei peptide și implicit a genei codificatoare?

Cunoașterea secvenței de nucleotide a unui fragment de ADN și implicii a secvenței de aminoacizi a unei proteine nu înseamnă obligatoriu și cunoașterea rolului (funcției) 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.



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

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



GUEST COMMENTARY

The Difficult Road from Sequence to Function

Robert H. White*

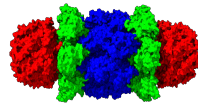
Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0308

U3.U6.2024

RESEARCH HIGHLIGHT

From sequence to functional understanding: the difficult road ahead

Periklis Makrythanasis¹ and Stylianos E Antonarakis^{1,2*}



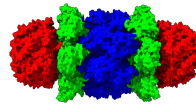
Pentru identificarea computerizată a funcției unei proteine sau gene necunoscute se pleacă de la următoarele **premize**:

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



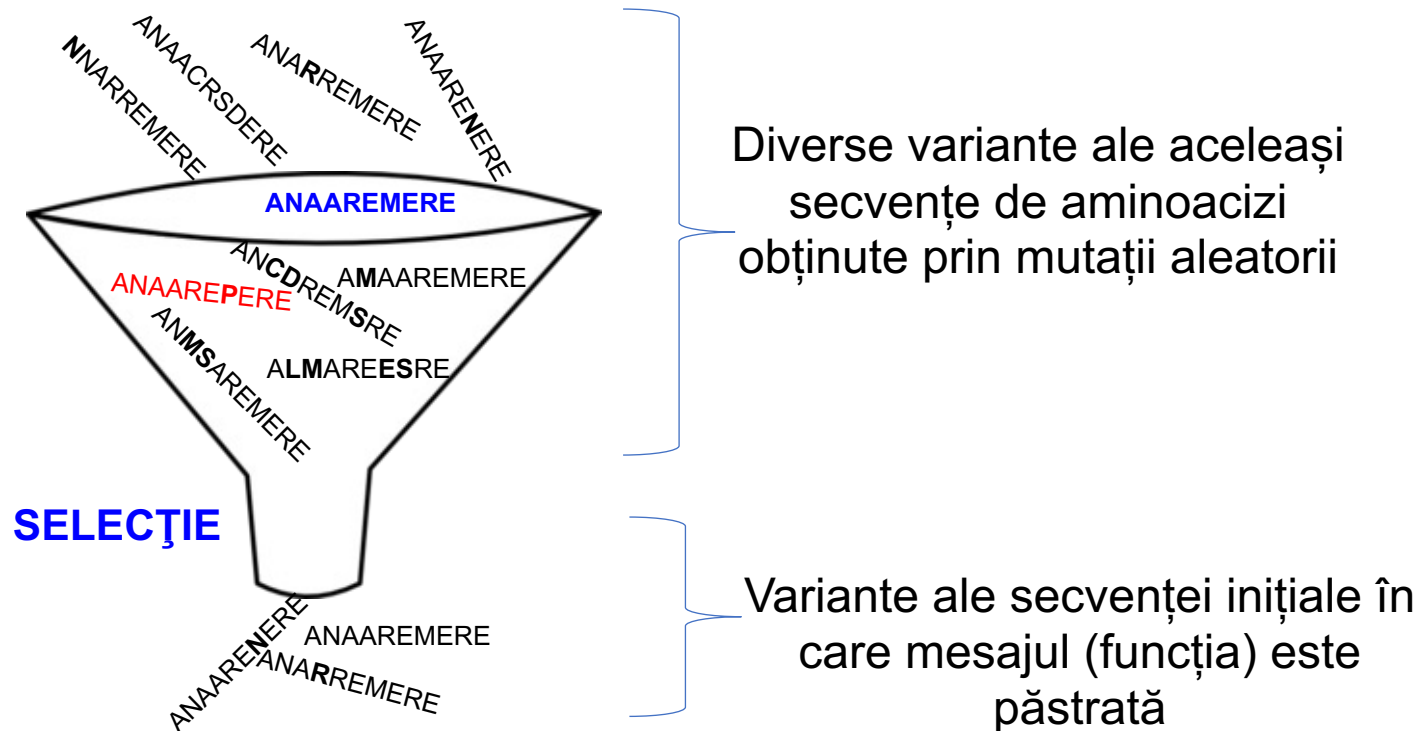
Ce înseamnă și cum funcționează evoluția?

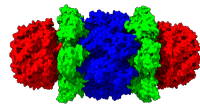
Mutațiile reprezintă modificări spontane nedorite a mesajului genetic. Cel mai frecvent mutațiile apar în procesul de replicarea ADN-ului sau **prin acțiunea factorilor de mediu asupra ADN-ului.** Mutațiile reprezintă materialul de bază pentru variabilitatea și evoluția organismelor vii. **Cum?**



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

- a) **aminoacizi înalt conservați** - nu sunt înlocuiți decât extrem rar - sunt aminoacizii din **situsul catalitic** sau **funcțional**, implicați în mod direct de realizarea funcției;
- b) **aminoacizi conservați** - sunt înlocuiți destul de rar - sunt aminoacizii implicați în realizarea structurilor secundare și terțiare;
- c) **aminoacizi puțin conservați** - sunt înlocuiți frecvent - sunt în general aminoacizii de pe suprafața proteinelor, înlocuirea lor nu modifică semnificativ funcția proteinei.





3. Deoarece **secvențele 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. **alinieră a două sau mai multe secvențe** - fiecare aminoacid (nucleotid) din secvența A este comparat cu aminoacidul (nucleotidul) corespunzător din secvența B. O corespondență între doi aminoacizi (nucleotide) din aceeași poziție pe cele două secvențe poartă numele de **identitate**, iar o neconcordanță se numește **substituție**;

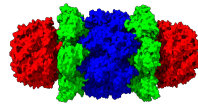
Alinierea a două sau mai multe secvențe pot fi:

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

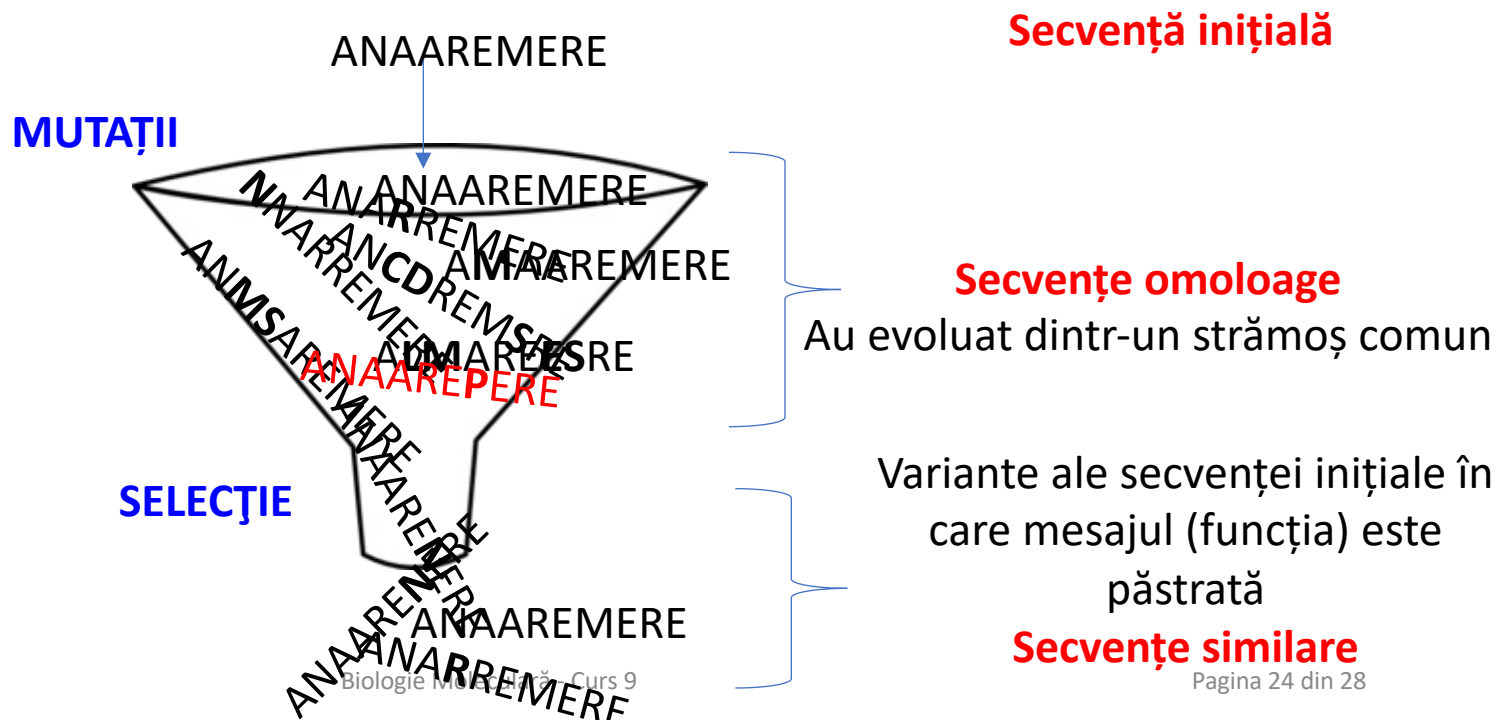
-**alinieri globale** - compară două secvențe pe toată lungimea lor și se utilizează pentru a compara secvențe de dimensiuni similare dar foarte apropiate evolutiv.

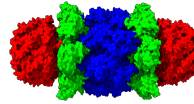
	identitate		substituție			
<i>E. coli</i>	TGNRTI	AVYDLGGGTFD	ISII	EIDEVDG	EKTTFEVLATNGDTHLGGEDDFDSRLIHYL	
<i>B. subtilis</i>	DEDQTI	LLYDLGGGTFD	VSILEL	GDG	TFEVRS	TAGDNRLGGDDDFDQVIIDHL
Secvență consens	TI**	YDLGGGTFD*	SI*	E*****	TFEEV**T*GD**LGG*DFD***I**L	

Identitate – similaritate - omologie



- b. **identitate** - la nivel de secvență - se referă la două secvențe care prezintă asemănări una în raport cu cealaltă datorită unui număr mai mare sau mai mic de **aminoacizi identici**; similaritatea se exprimă ca **procente de similaritate sau procente identitate** - % de aminoacizi ce sunt identici între 2 secvențe;
- c. **similaritate** la nivel de secvență - se referă la două secvențe 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 secvențe se aseamănă una cu cealaltă deoarece au evoluat dintr-un strămoș 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)**;





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

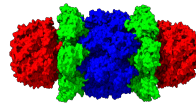
Două secvențe de nucleotide similare vor codifica un mesaj genetic similar și deci vor avea funcții similare.

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

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

MAAKYRIGYFVGSLATGSINRVLSQALINLAPEDLEFSEIPIRDLPLYSDYDADFPPEGR

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



BLAST - Basic Local Alignment Search Tool

1. identifică, dintr-o bază de date, **secvențele similare cu o secvență țintă (tinta analizei, experimentului)**. Aceste secvențe 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 și zone nealiniat (așa numitele **GAP’s**) care formează bucle între o aliniere locală și 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 observațiile experimentale.

298

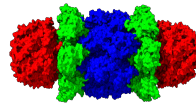
Biologie moleculară. Metode experimentale

	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	-1
G	0	-3	-1	-2	-3	4	-2
H	-2	-3	-1	0	-1	-2	3

BLOSUM 62

FIGURA 35. Exemplu de matrice BLOSUM.

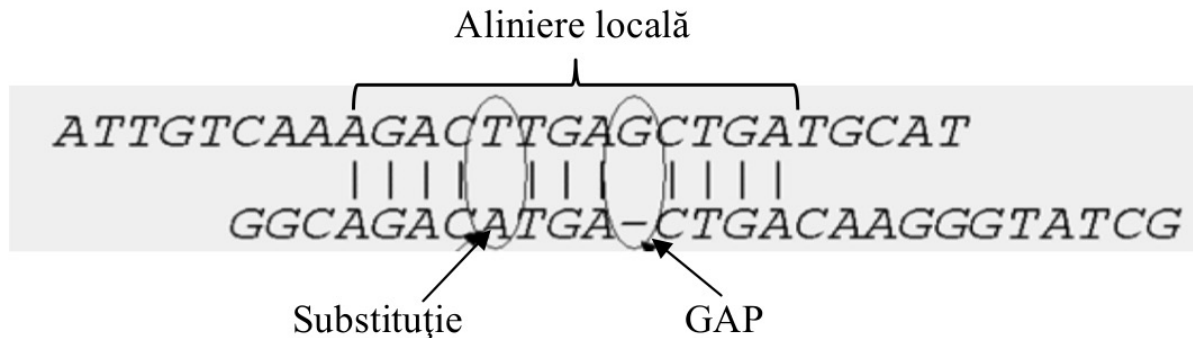
Cu litere mari sunt reprezentați aminoacizii. Cel mai mare punctaj au identitățile, iar funcție de frecvența de substituție (observată practic în laborator) primesc un anumit punctaj și substituțiile. Se poate observa și existența unor punctaje negative, alocate pentru substituțiile cel mai puțin întâlnite.



3. Calculează un **scor de similaritate** prin însumarea punctelor pentru fiecare pereche aminoacid-aminoacid și **ierarhizează secvențele** țintă funcție de valoarea acestui scor.

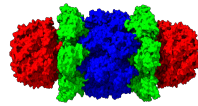
Scoruri de similaritate calculate de BLAST:

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



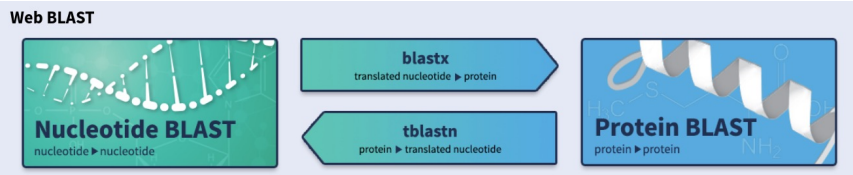
$$S = \sum_{(\text{identități, substituții})} - \sum_{(\text{penalizări GAP})}$$

- **scorul în biți notat cu S'** - se calculează prin normalizarea lui S în funcție de diverse variabile statistice care depind, la rândul lor, de tipul de matrice utilizat. **Cu cât punctajul S' obținut este mai mare cu atât asemănarea dintre cele două secvențe este mai mare;**
- **parametru statistic E** - care se definește ca număr de potriviri care apar doar datorită șansei 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 semnificație (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:

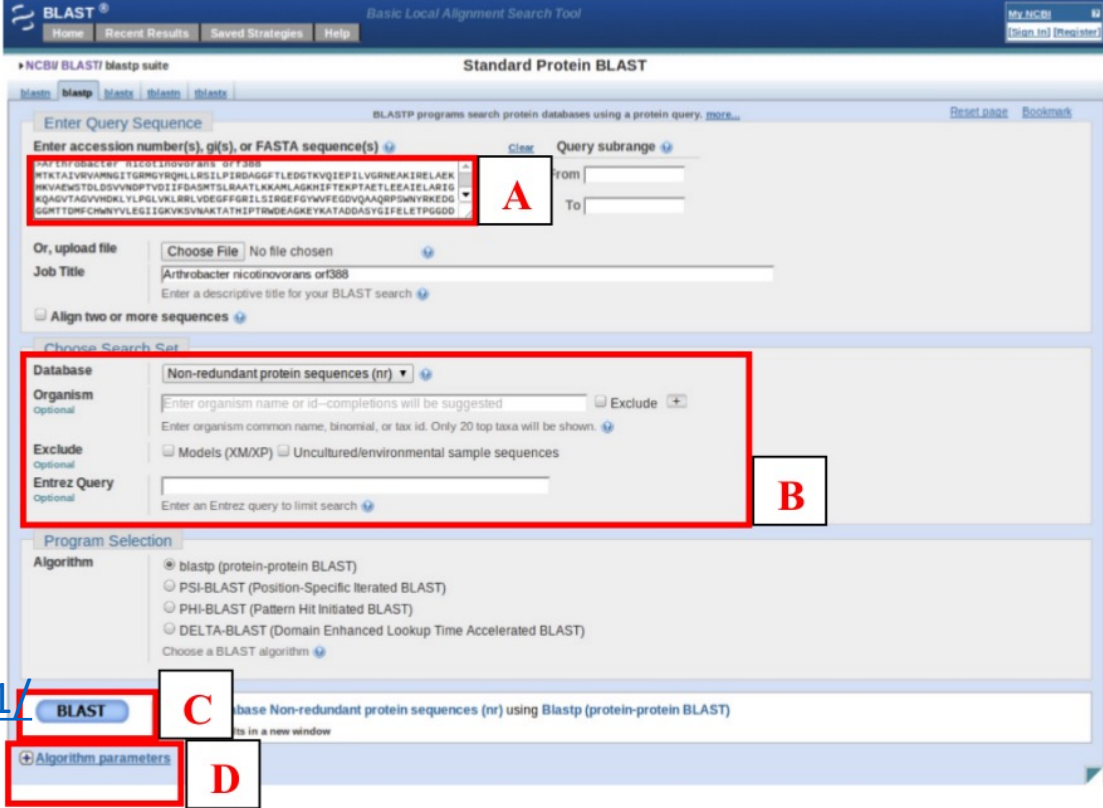


Metode computaționale

303

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

- A – căsuța text în care a fost inserată secvența țintă în format FASTA;
- B – zona cu parametrii utilizați pentru restrângerea spațiului de căutare;
- C – buton pentru lansarea investigației;
- D – zona cu parametrii algoritmului de căutare



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

Mai multe detalii și aplicații practice la cursul Bioinformatică aplicată în Biologia Structurală, Opțional, An III.