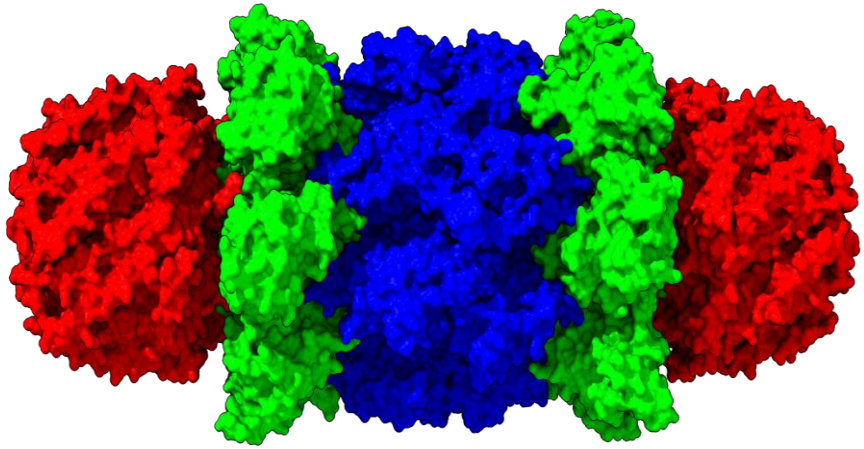
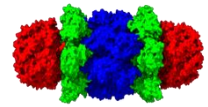
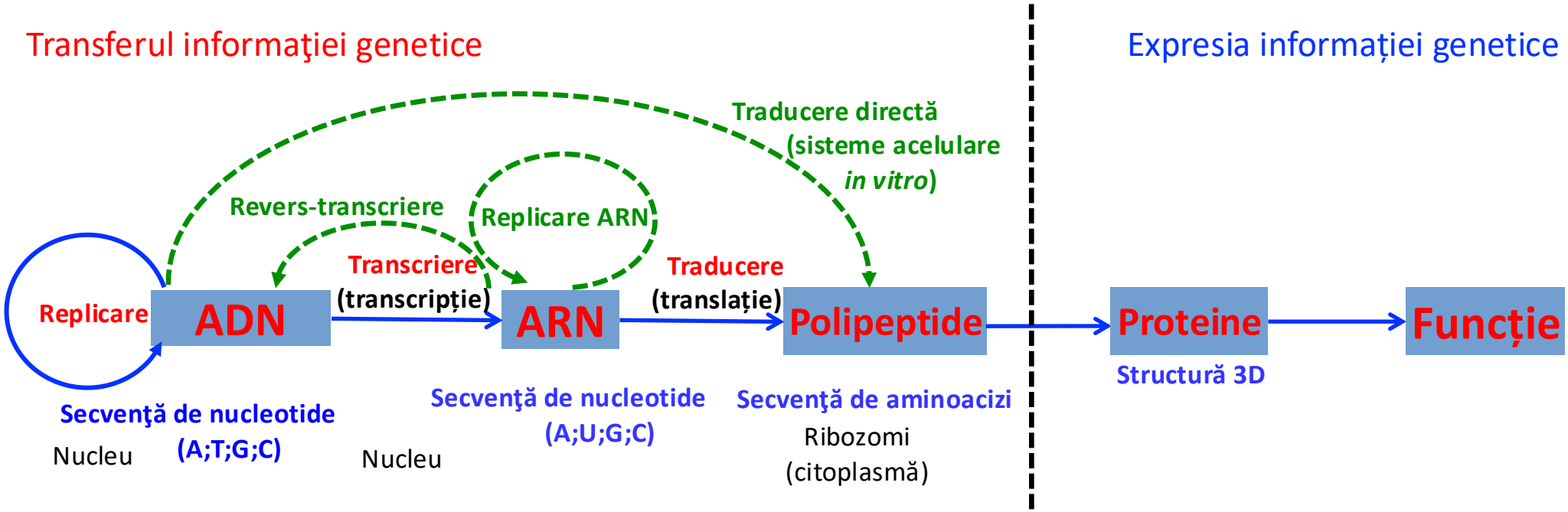


Capitolul V
Modificarea mesajului genetic
II. Secvenţierea acizilor nucleici şi analiza secvenţelor

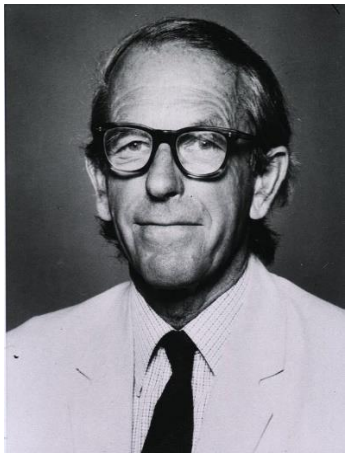
Prof. Dr. Marius Mihășan

Amplasarea cursului în traseul informației genetice





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



Frederick Sanger - 1918 - 2013

În prezent există numeroase metode de secveţializare a ADN-ului clasificate în:
- metode derivate de la **metoda chimică** a lui **Maxam şi Gilbert;**
- metode derivate de la **metoda enzimatică** a lui **Sanger;**

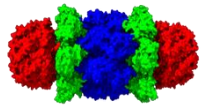
Metoda enzimatică Sanger

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

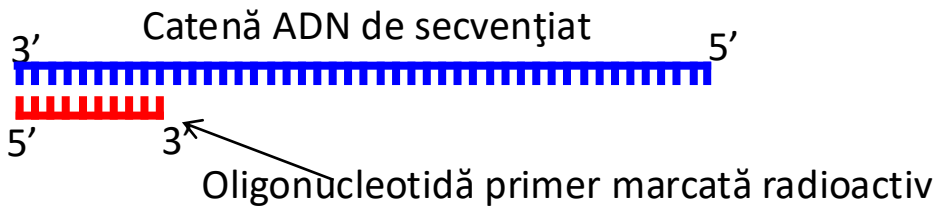
Elementele esenţiale realizării unei reacţii de secvenţiere prin metoda Sanger

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

Etapele reacţiei de secvenţiere



1) Legarea oligonucleotidului primer



2) Elongarea primerului prin acţiunea ADN polimerazei în prezenţa dNTP şi ddNTP cu sinteza unei catene noi. Reacţia se realizează în **4 eprubete diferite**, fiecare conţinând 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 colecţie de catene ADN nou sitetizate de dimensiuni diferite.

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

#ddA
#AGCTGCCCG

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

#AGddC
#AGCTGddC
#AGCTGCddC

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

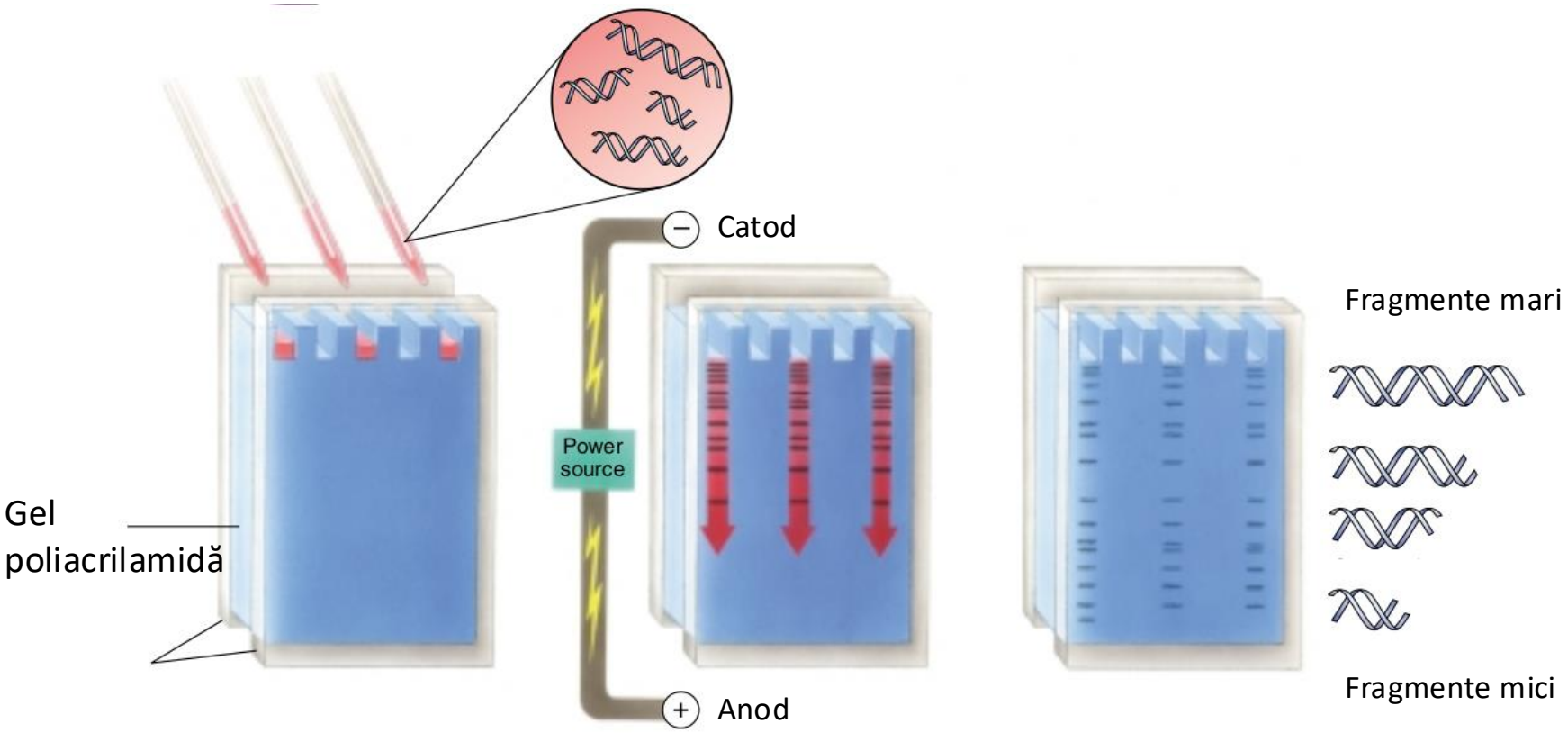
#AddG
#AGCTddG
#AGCTGCCCddG

T
ddT +
A; T; C; G
Amorsa #

#AGCddT
#AGCTGCCCG

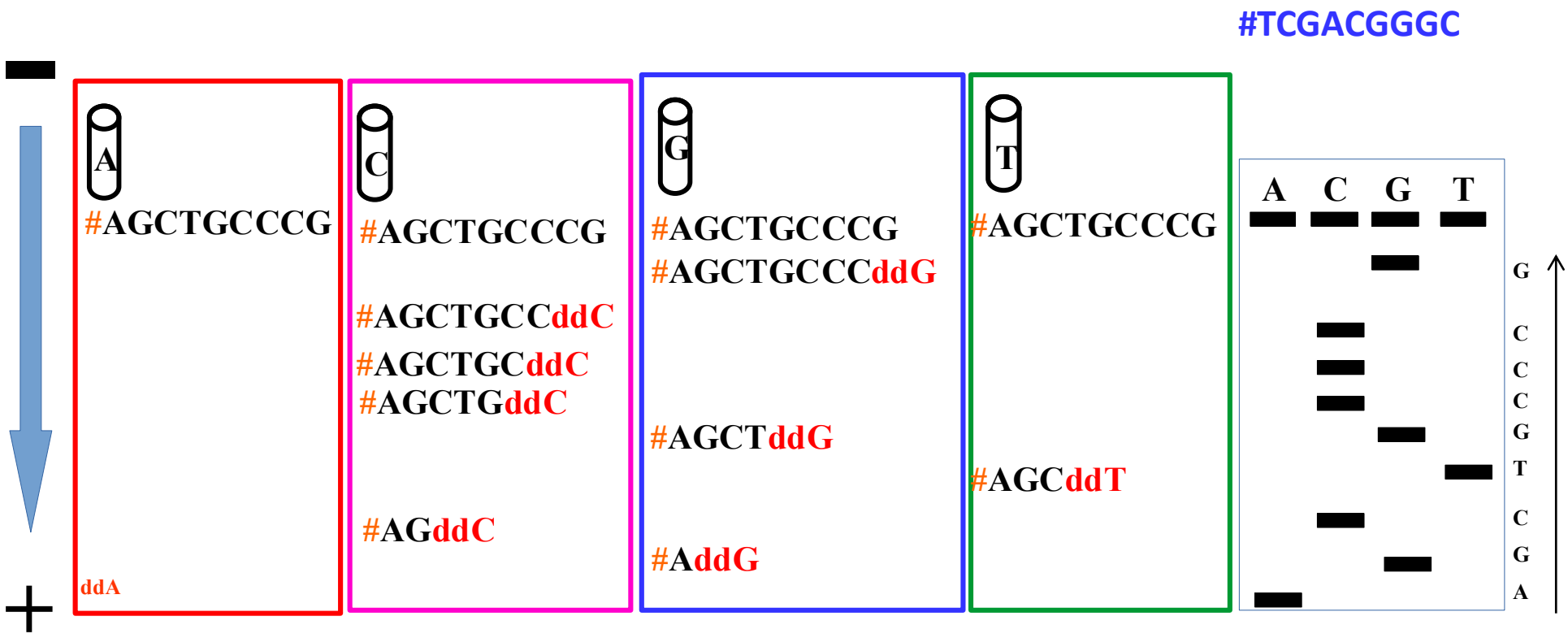
Etapele reacţiei de secvenţiere

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



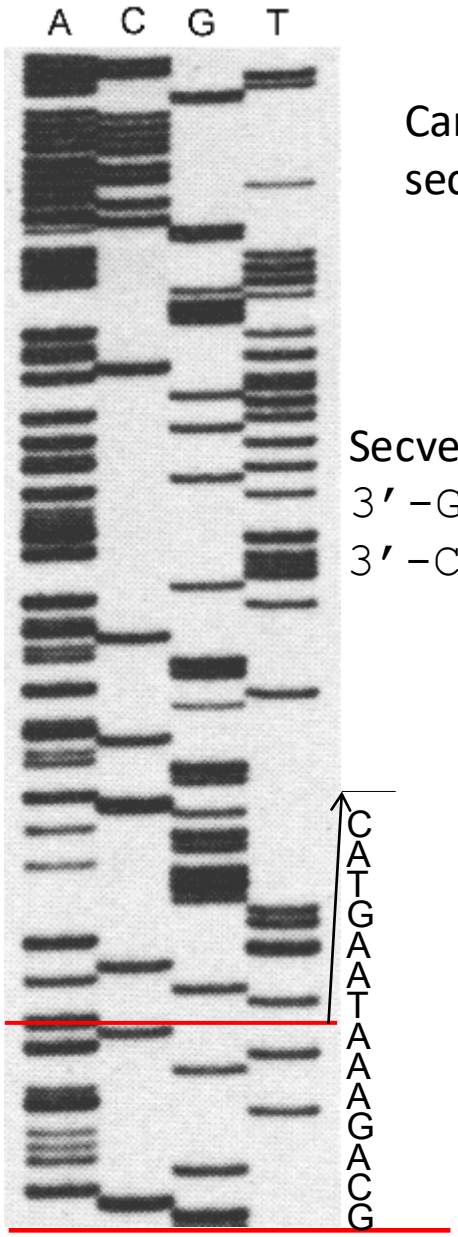
Etapele reacţiei de secvenţiere

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



Schema autoradiografiei unui gel de secvenţializare

Etapele reacţiei de secvenţiere



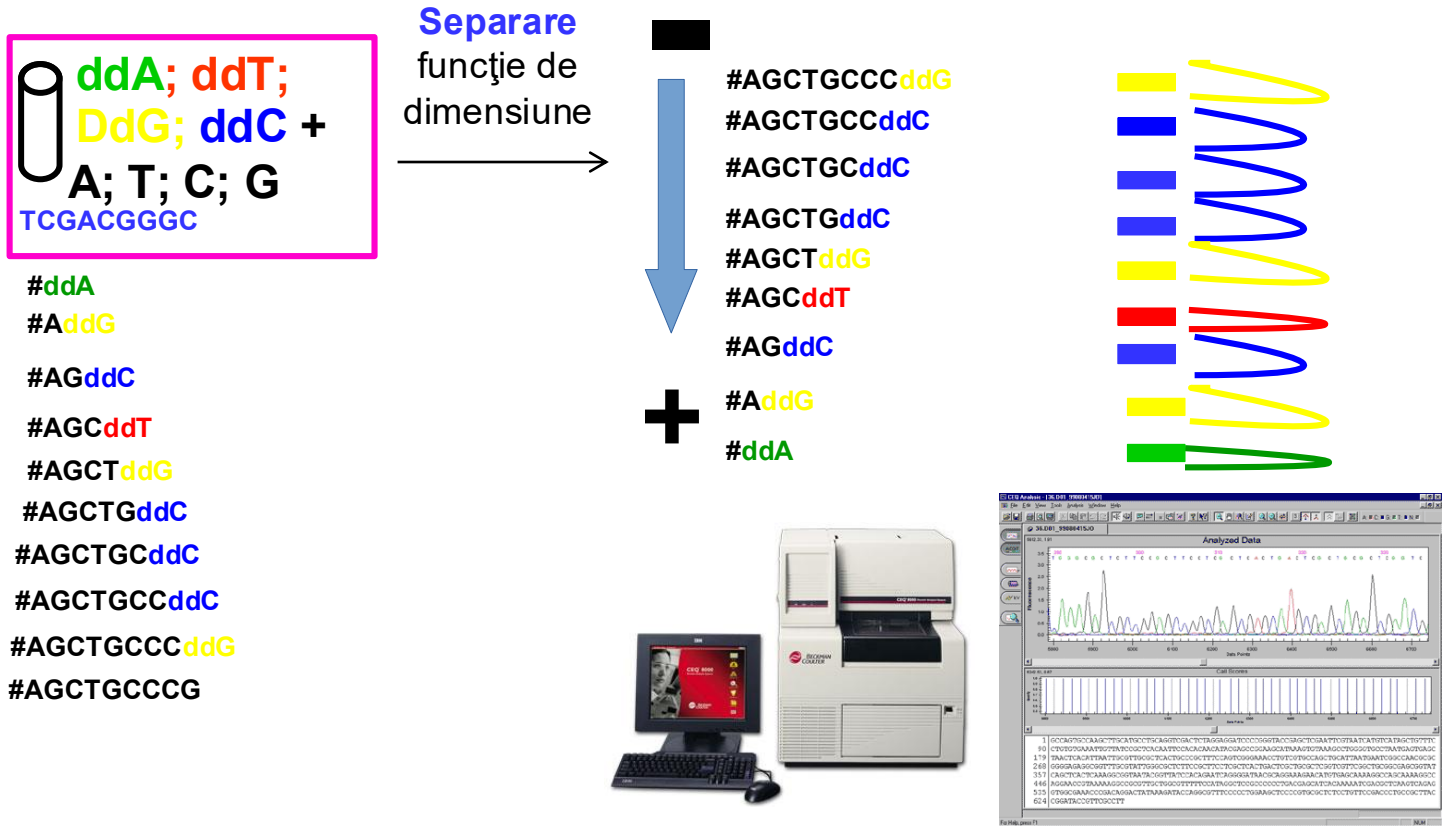
Care este secvența fragmentului de ADN prin a cărui secvențiere Sanger se obține gelul de alături?

Secvența citită de pe gel:
3’-GCAGAAATAAGTAC-5’ deci secvența catenei de secvențiat era:
3’-CGTCTTTATTCATG-5’,
respectiv 5’GTACTTATTTCTGC-3’

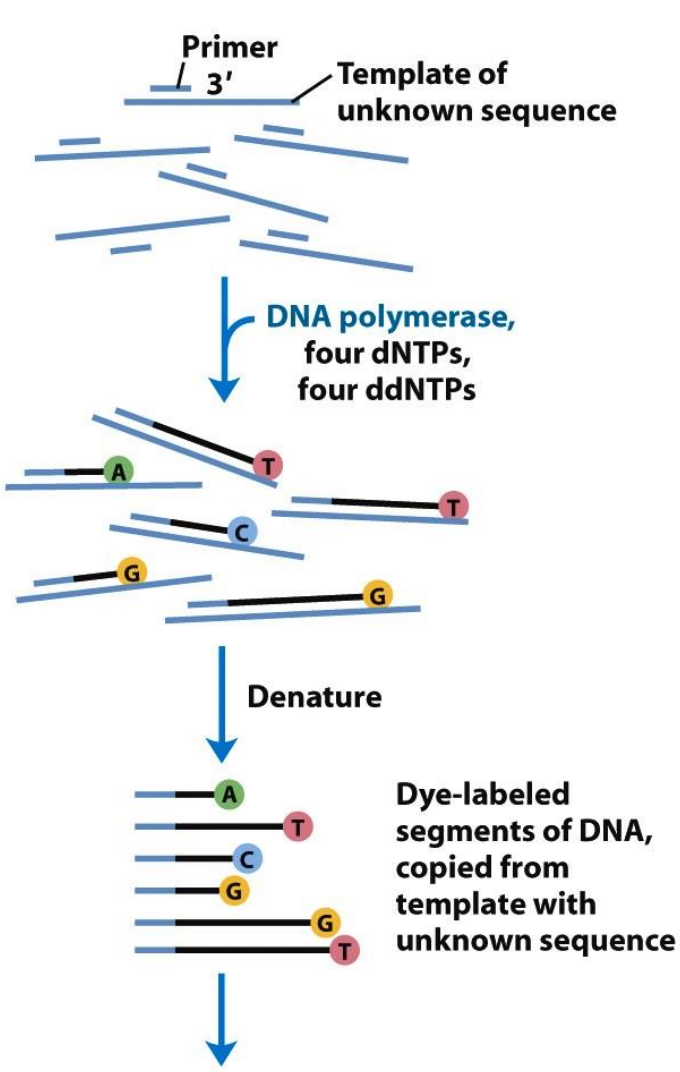
Etapele reacţiei de secvenţiere

O variantă a metodei Sanger este:

– **utilizarea de ''terminatori'' de catenă marcaţi fluorescenţi**. Fiecare din cele 4 ddNTP sunt marcate cu câte un florocrom diferit (ddA – verde, ddT – roşu, ddG – galben, ddC – albastru). Reacţia se desfăşoară într-o singură eprubetă, iar fragmentele generate sunt separate prin electroforeză. Ordinea nucleotidelor este citită cu ajutorul unui laser şi reprezentată ca ''peak-uri''.



Etapele reacţiei de secvenţiere



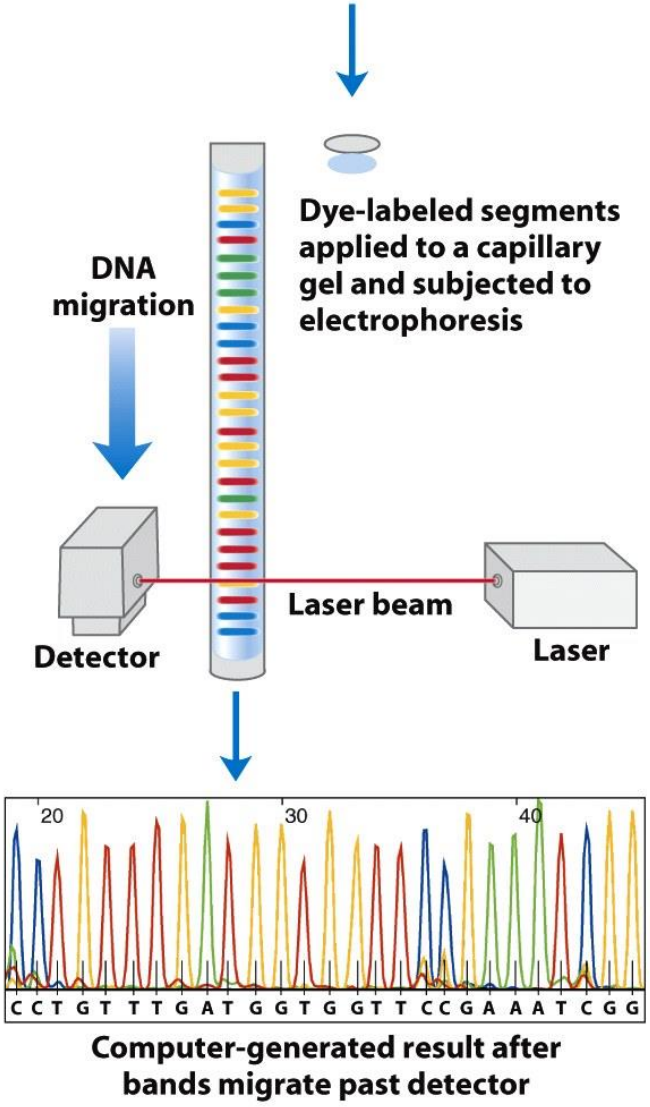
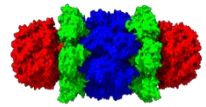


Figure 8-34
Lehninger Principles of Biochemistry, Fifth Edition

Metode de secvențiere “de nouă generație”



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)

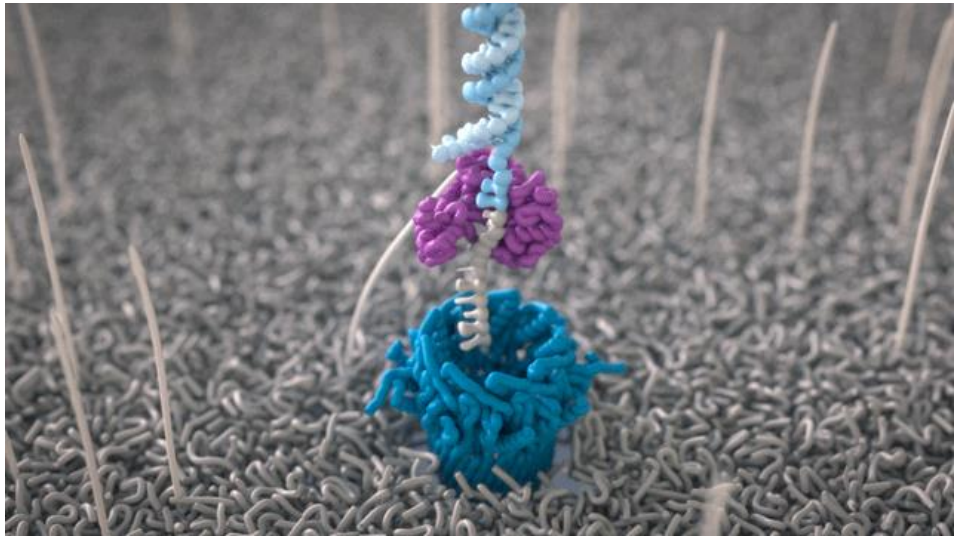
2. Metode de secvenție de generația a **III** / metode de secvențiere a fragmentelor lungi: **SMRT** (Single Molecule Real Time) și secvențiere folosind nanopori proteici **Oxford Nanopore Technologies**

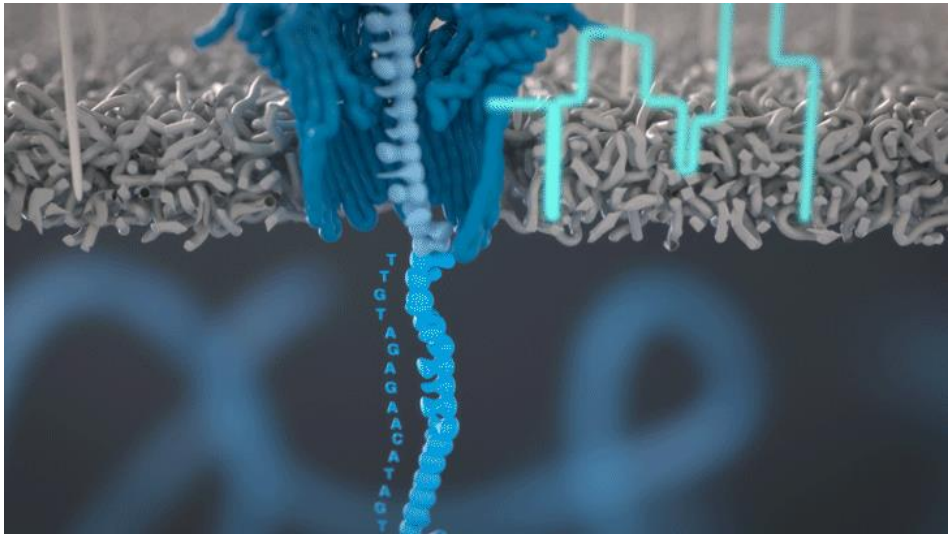
Toate aceste metode permit generarea unui număr semnificativ mai mare de nucleotide secvențiate comparativ cu metoda Sanger, de aceea ele mai sunt denumite și metode **metode extensive** (**High-throughput**)

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

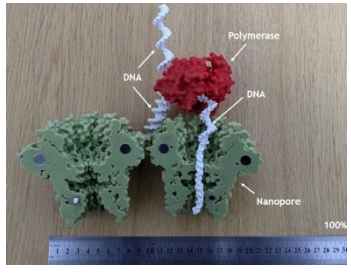
Metode de secvențiere “de nouă generație”

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.



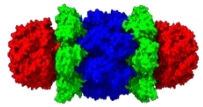


https://nanoporetech.com/how-it-works



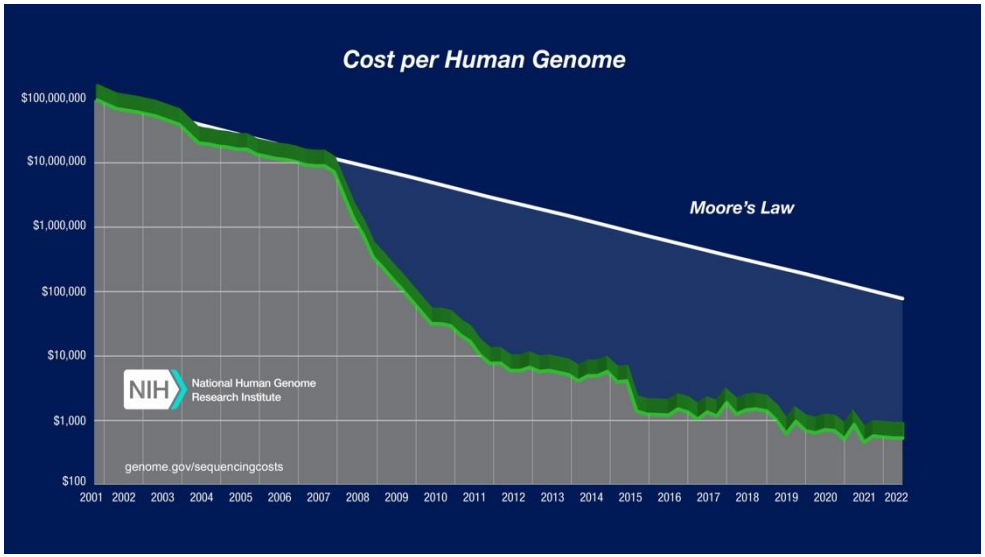
https://modelemoleculare.ro/nanopor-proteic/

Evoluția costului secvențierii

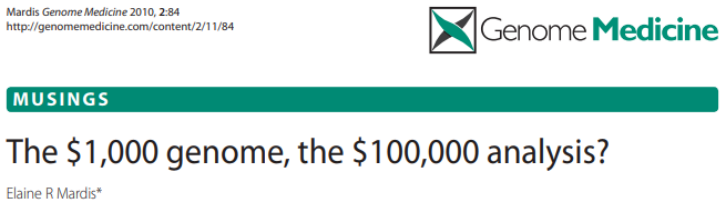


Spre deosebire de metodele de secvențiere a proteinelor, stabilirea ordinii nucleotidelor în acizii nucleici ADN și ARN este mult mai rapidă și ieftină. În ultimii ani metoda clasică Sanger a fost înlocuită de metode de `generație nouă`` (next-gen) ce au un randament și o viteză foarte mare (high-throughput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

Secvențierea genomului uman - 3 bilioane de baze - **~2.7 bilioane $**
Demarat în 1990 proiectul internaţional Human Genome Project (HGP) ce a fost declarat încheiat în 2003 şi a reuşit secvenţierea unui set haplod 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 $**



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


Mardis *Genome Medicine* 2010, **2**:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

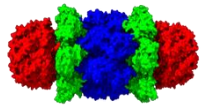
MUSINGS

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

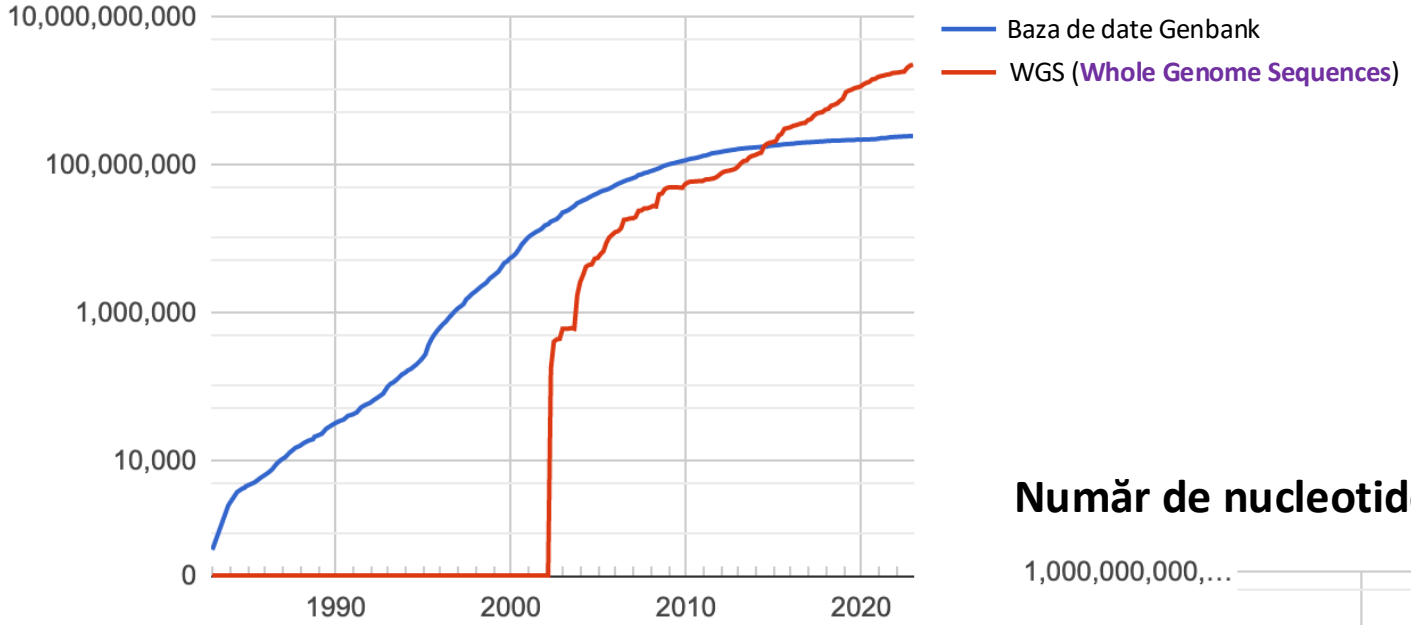
Elaine R Mardis*

Bioinformatica – ansamblul de metode şi tehnici ce permit stocarea, manipularea şi interpretarea informaţiei genetice – definiţia restrânsă !!!!

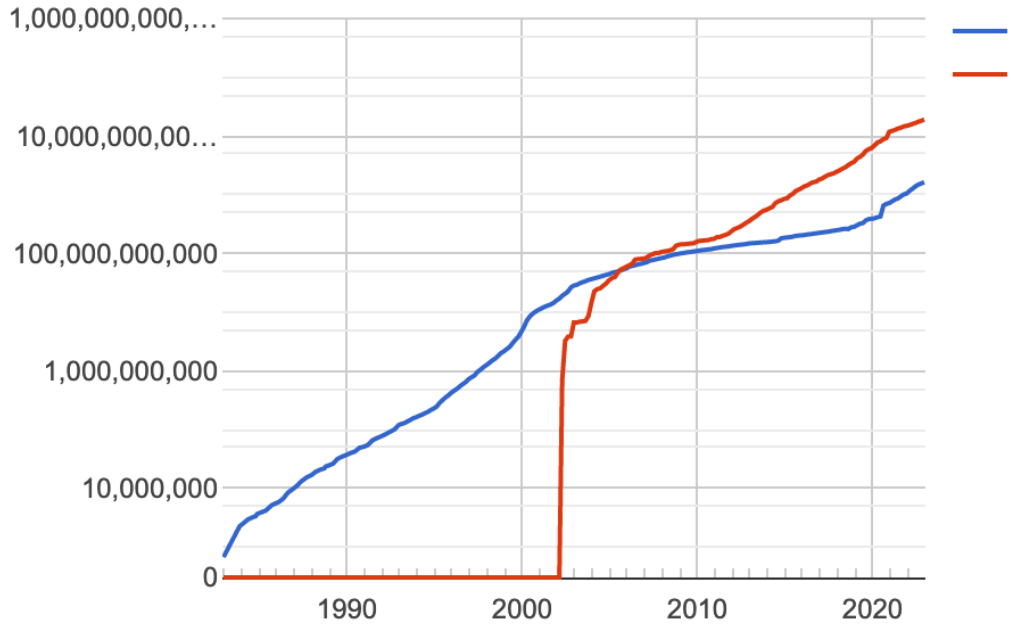
Câte secvențe sunt disponibile?



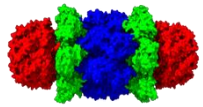
Număr de secvențe cunoscute



Număr de nucleotide în secvențe cunoscute



Cum și unde sunt stocate secvențele



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**.

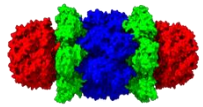
De-a lungul timpului au existat un număr mare de variante de fișiere/formate create pentru a înregistra secvențe, însă fișierul/formatul FASTA este cel ce s-a impus.

Fișierul FASTA – Fast-All

```
>AJ507836.1 Arthrobacter nicotinovorans pAO1 megaplasmid sequence, strain ATCC 49919
GATCCGGCGGTCCGCCGTCGTGGCGGCGCGGGCGGAGGTTGCCGGCGGCGGACCGCCGCCCGCACAAGAA
GGCCTTCGGGTTCCGGCAGGTGGCGCGGCCCCCGACACTGGTCCTCGCCTGGGGTGGAACGTGGGGTGCT
GGGTGTGGGCGGTTCGCCGGGCGAACCGGGAAAGGCGTCCACTCCTCTTCGCTTCCGTGGCCGGCGGTTG
GGGCCGGTCCCGTCGTCGCAGAGCTCCGCCGTGCCCGGCCCCGCCCGTTGTCTACGACGTCTTTTTGTGG
CTGTCCTTCGGATCATACGGTTCCGCTCCAGCTCAAGCTCCTGGCAGTCCGCAAAGCTCCGGTCTAGCAC
ACAGCGATGCGGGTAATGATGGCGACGAATTCGTCCCAGAGCGCTGGTGCGATTTCGTGGCCCATACCTG

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

Fișiere FASTA - fișiere cu secvențe



Fișierul FASTA

- un fișier text ce poate **avea sau nu** extensia .fasta;
- conține secvențe de nuvleotide 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;

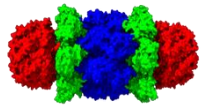
C.fiecare rând al secvenței are în general 80 de caractere (nu mai mult de 120)

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 intermediară
Y	T/C (pYrimidine)		

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



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 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**;

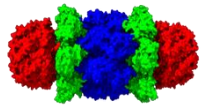
```
>secventa peptidica necunoscuta 1 fara numere
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP

>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

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 putin 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.

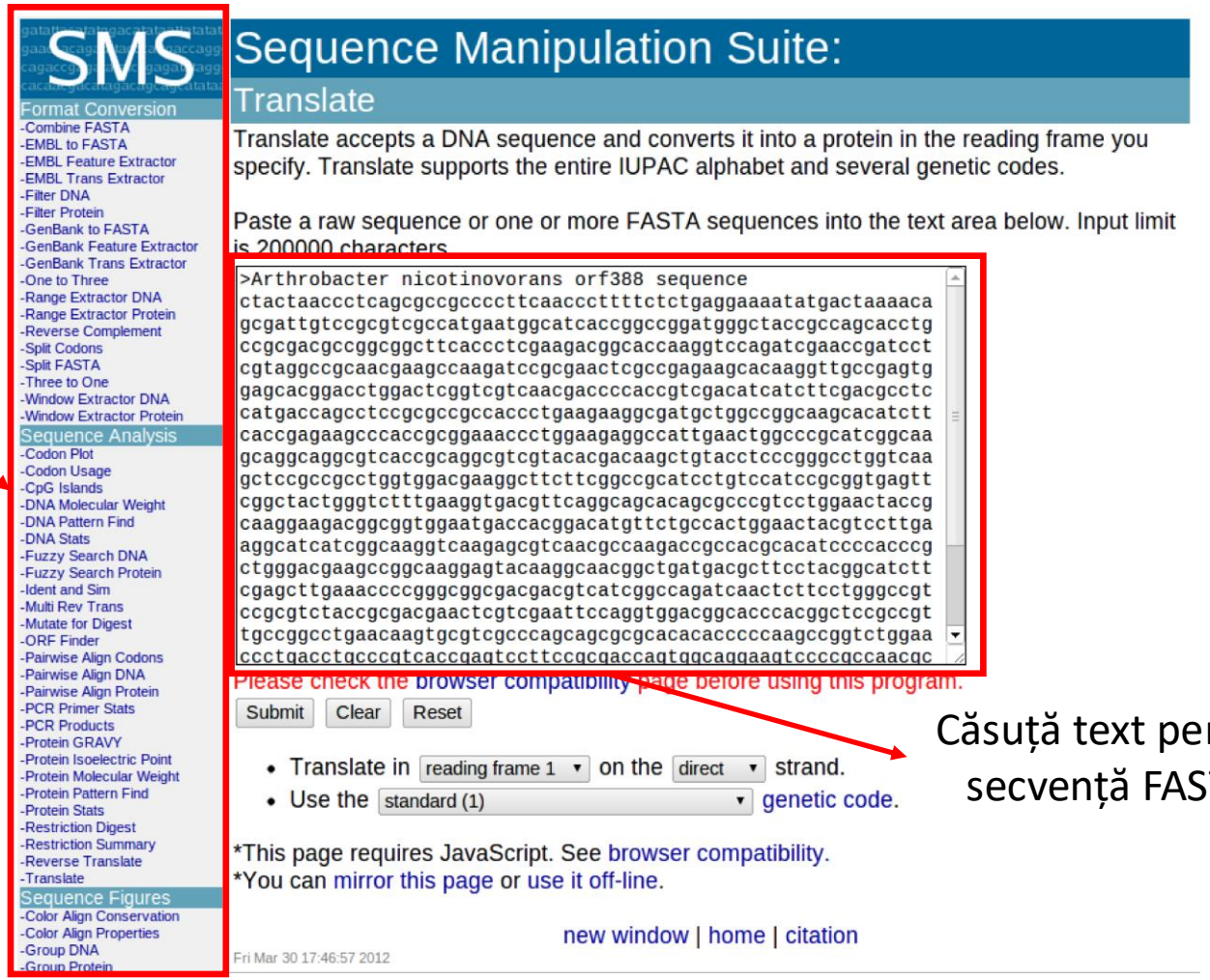
Bază de date	Site web	Dimensiune *
Baze de date cu secvenţe		
INSDC	http://www.insdc.org/	206 293 625 secvenţe Aceste baze de date sunt sincronizate zilnic între ele şi conţin aceleaşi secvenţe.
DDBJ	http://www.ddbj.nig.ac.jp	
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl/	
GenBank	http://www.ncbi.nlm.nih.gov/genbank/	
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

Operații simple cu secvențe

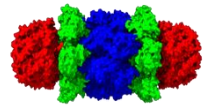
http://www.bioinformatics.org/sms2/

Operații cu secvențe selectabile în SMS2



Căsuță text pentru secvență FASTA

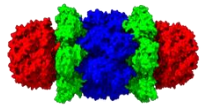
Operații simple cu secvențe



Denumire operaţie*	Descriere
Convertire între diverse formate	
Combine FASTA	– permite combinarea a două sau a mai multe secvenţe în format FASTA şi obţinerea unei singure secvenţe
EMBL to FASTA	– permite convertirea unei secvenţe din formatul EMBL în formatul FASTA; funcţia este utilă în situaţia în care se doreşte eliminarea rapidă a informaţiilor care nu au legătură cu secvenţa de ADN dintr-un fişier EMBL
Filter DNA	– elimină caracterele care nu corespund codului standard IUPAC pentru ADN-ul dintr-o secvenţă (inclusiv spaţii, numere, etc)
Filter Protein	– elimină caracterele ce nu corespund cu codul standard IUPAC de o literă pentru aminoacizi dintr-o secvenţă (inclusiv spaţii, numere etc.)
GenBank to FASTA	– permite convertirea unei secvenţe din formatul GenBank în formatul FASTA; funcţia este utilă în situaţia în care se doreşte eliminarea rapidă a informaţiilor care nu au legătură cu secvenţa de ADN dintr-un fişier GenBank
One to Three	– permite convertirea unei secvenţe de aminoacizi din codul standard IUPAC de o literă în cel de trei litere
Analiza secvenţelor	
Codon Plot	– analizează frecvenţa de utilizare a codonilor dintr-o secvenţă de ADN şi generează un grafic cu aceste frecvenţe
Codon Usage	– analizează frecvenţa de utilizare a codonilor dintr-o secvenţă de ADN şi poate fi utilizat pentru a identifica preferinţa pentru un anumit codon sinonim
DNA Molecular Weight	– calculează masa moleculară a uneia sau a mai multor secvenţe de ADN
DNA Stats	– calculează frecvenţa fiecărei nucleotide într-o secvenţă dată, rezultatele fiind exprimate în procente
PCR Primer Stats	– analizează şi calculează proprietăţile specifice ale unui set indicat de primeri, inclusiv temperatura de topire şi procentul de GC
PCR Products	– realizează amplificarea virtuală prin PCR a unei secvenţe date folosind un set de primeri indicat de utilizator
Protein Isoelectric Point	– calculează valoarea teoretică a punctului izoelectric pentru o secvenţă dată
Protein Molecular Weight	– calculează valoarea teoretică a masei moleculare pentru una sau mai multe secvenţe de aminoacizi
Protein Stats	– calculează frecvenţa fiecărui aminoacid într-o secvenţă dată, rezultatele fiind exprimate în procente
Translate	– realizează translaţia şi transcripţia virtuală a unei secvenţe de ADN date, utilizatorul având posibilitatea de a alege cadrul de citire specific
Reverse Translate	– realizează procesul invers faţă de Translate, acceptând o secvenţă de aminoacizi şi generând secvenţa ADN corespunzătoare

* corespunzătoare cu denumirea din suita de programe SMS2

Legătura secvență –funcție



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

MAAKYRIGYFVGSLATGSINRVLSQALINLAPEDLEFSEIPIRDLPLYSYDYDADFPPEGR

Care este funcția acestei peptide și implicit a genei codificatoare?

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

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

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

Manual sau cu metode bioinformatice

Numai cu metode bioinformatice

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

Vol. 188, No. 10

GUEST COMMENTARY

The Difficult Road from Sequence to Function

Robert H. White*

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

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

Genome Medicine

RESEARCH HIGHLIGHT

From sequence to functional understanding: the difficult road ahead

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

Legătura secvență –funcție

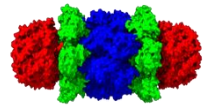
Pentru identificarea computerizată a funţiei unei proteine sau gene necunoscute se pleacă de la următoarele **premize**:

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

Ce înseamnă și cum funcționează evoluția?

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

Legătura secvență –funcție

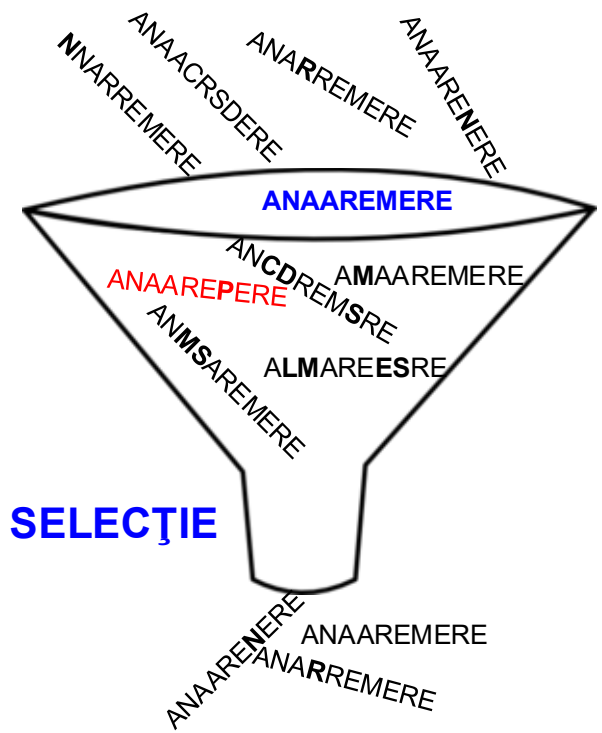


2. înlocuirea unui aminoacid cu altul într-o proteină nu este întoteauna aleatoare ci este corelată cu rolul aminoacidului în cadrul proteinei. Din punct de vedere are frecvenței cu care sunt înlocuiți au fost descrise **3 categorii distincte** de aminoacizi într-o secvență proteică:

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

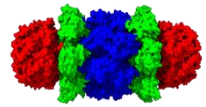
b) **aminoacizi conservaţi** - sunt înlocuiţi destul de rar - sunt aminoacizii implicaţi în realizarea structurilor secundare şi terţiare;

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



Diverse variante ale aceleași secvențe de aminoacizi obținute prin mutații aleatorii

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



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

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

Alinierile a două sau mai multe secvențe pot fi:
-**alinieri locale** - identifică subregiunile similare dintre două secvenţe
-**alineri globale** - compară două secvenţe pe toată lungimea lor și se utilizează pentru a compara secvenţe de dimensiuni similare dar foarte apropiate evolutiv.

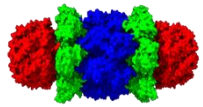
identitate

substitutie

E. coli TGNRTIAVYDLGGGTFDISIIEIDEVDGEKTFEVLATNGDTHLGGEDFDSRLIHYL
B. subtilis DEDQTILLYDLGGGTFDVSILELGDG TFEVRSTAGDNRLGGDDFDQVIIDHL

Secvență consens TI**YDLGGGTFD*SI*E********TFEEV**T*GD**LGG*DFD***I**L

Identitate – similaritate - omologie

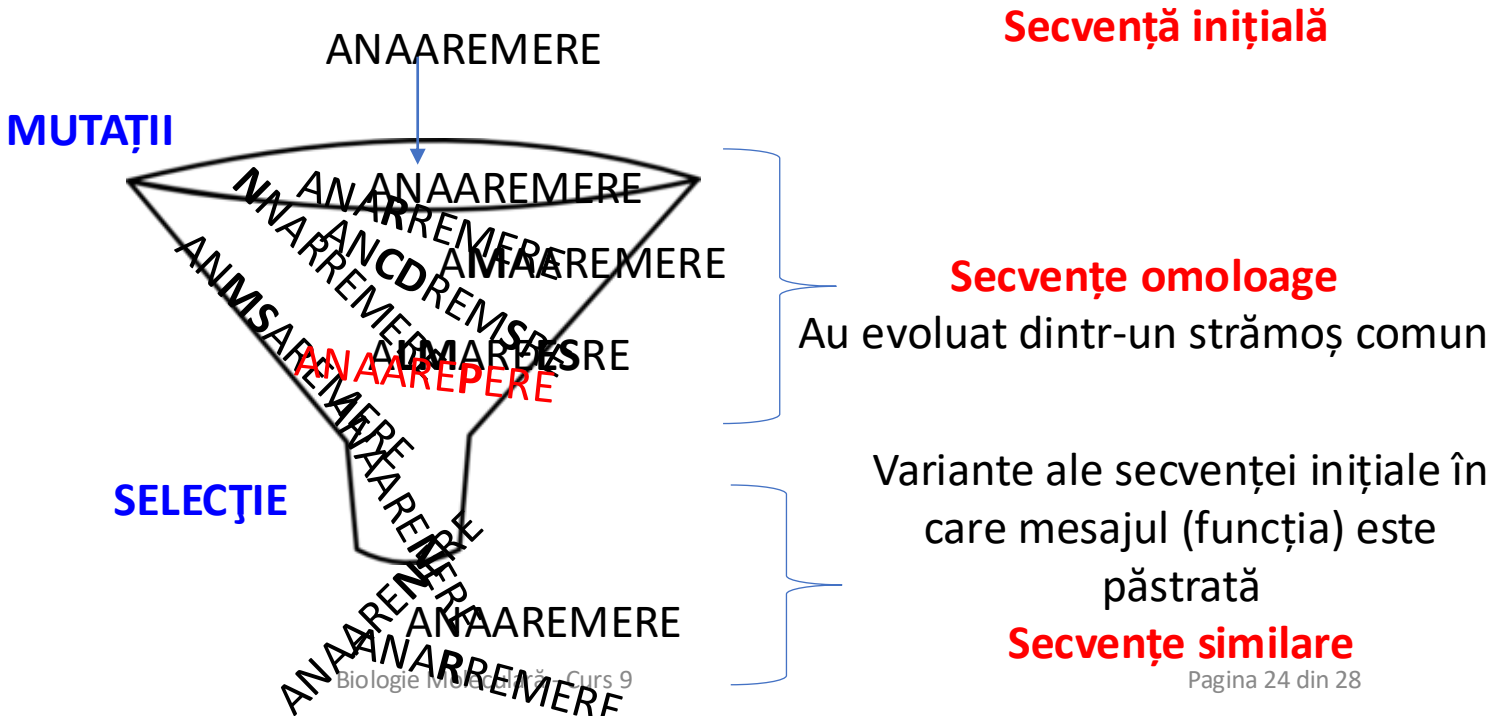


b. **identitate -** la nivel de secvenţă - se referă la două secvenţe care prezintă asemănări una în raport cu cealaltă datorită unui număr mai mare sau mai mic de **aminoacizi identici;** similaritatea se exprimă ca **procente de similaritate sau procente identitate** - % de aminoacizi ce sunt identici între 2 secvențe;

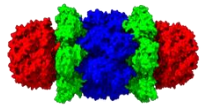
c. similaritate la nivel de secvenţă - se referă la două secvenţe care prezintă asemănări una în raport cu cealaltă datorită unui număr mai mare sau mai mic de **aminoacizi identici** dar ia în calcul și **semnificația substituțiilor dintre aminoacizi ;**

c. omologie - se referă la faptul că două secvenţe se aseamănă una cu cealaltă deoarece au evoluat dintr-un strămoş comun, dar nu au obligatoriu aceeași funcție;

d. **secvența de aminoacizi identici înalt conservați** dintr-o aliniere a două sau mai multe secvențe se numește **secvență consens (consensus)**;



Legătura secvență –funcție



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

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

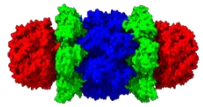
Două proteine similare vor avea structuri similare și deci funcții similare.

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

MAAKYRIGYFVGSLATGSINRVLSQALINLAPEDLEFSEIPIRDLPLYSYDYDADFPPEGR

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

BLAST – identificarea de secvențe similare



BLAST - Basic Local Alignment Search Tool

1. **identifică**, dintr-o bază de date, secvenţele similare cu o **secvenţă ţintă (tinta analizei, experimentului)**. Aceste secvenţe identificate poartă numele de **secvenţe "subiect",** iar identificarea lor se bazează pe **alinieri locale**. Secvenţa „subiect este „suprapusă" peste cea țintă la nivelul alinierilor locale astfel încât secvenţele comparate vor fi alcătuite din zone perfect aliniate şi zone nealiniate (aşa numitele **GAP'**s) care formează bucle între o aliniere locală şi următoarea aliniere locală.
2. **cuantifică nivelul de similaritate** dintre secvenţele "subiect" şi secvenţa ţintă prin utilizarea unor **matrici de substituţie.** O matrice de substituţie arată frecvenţa cu care un aminoacid este înlocuit cu altul şi are la bază observaţiile experimentale.

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

BLOSUM 62

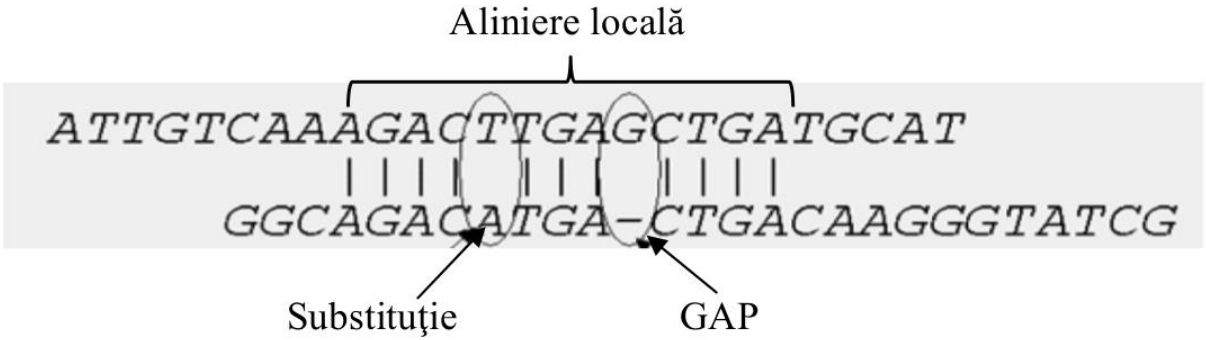
FIGURA 35. Exemplu de matrice BLOSUM.
Cu litere mari sunt reprezentaţi aminoacizii. Cel mai mare punctaj au identităţile, iar funcţie de frecvenţa de substituţie (observată practic în laborator) primesc un anumit punctaj şi substituţiile. Se poate observa şi existenţa unor punctaje negative, alocate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

BLAST – identificarea de secvențe similare

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

Scoruri de similaritate calculate de BLAST:

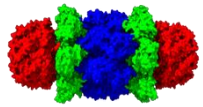
- **punctaj brut** (engl. Raw score) notat cu S, este calculat prin însumarea punctelor pentru fiecare pereche aminoacid-aminoacid, aminoacid-nimic şi penalizărilor pentru GAP; nu permite ierarhizarea secvențelor, valoare lui depinde de lungimea secvențelor analizate;



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

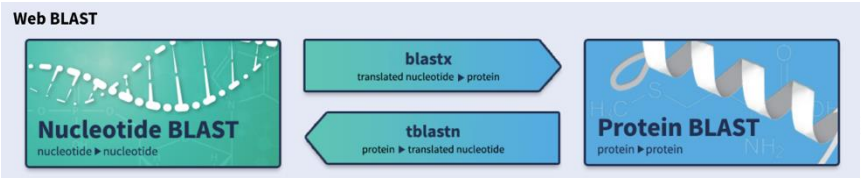
- **scorul în biţi notat cu S'** - se calculează prin normalizarea lui S în funcţie de diverse variabile statistice care depind, la rândul lor, de tipul de matrice utilizat. **Cu cât punctajul S' obţinut este mai mare cu atât asemănarea dintre cele două secvenţe este mai mare;**
- **parametru statistic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

Cum se realizează o analiză BLAST?

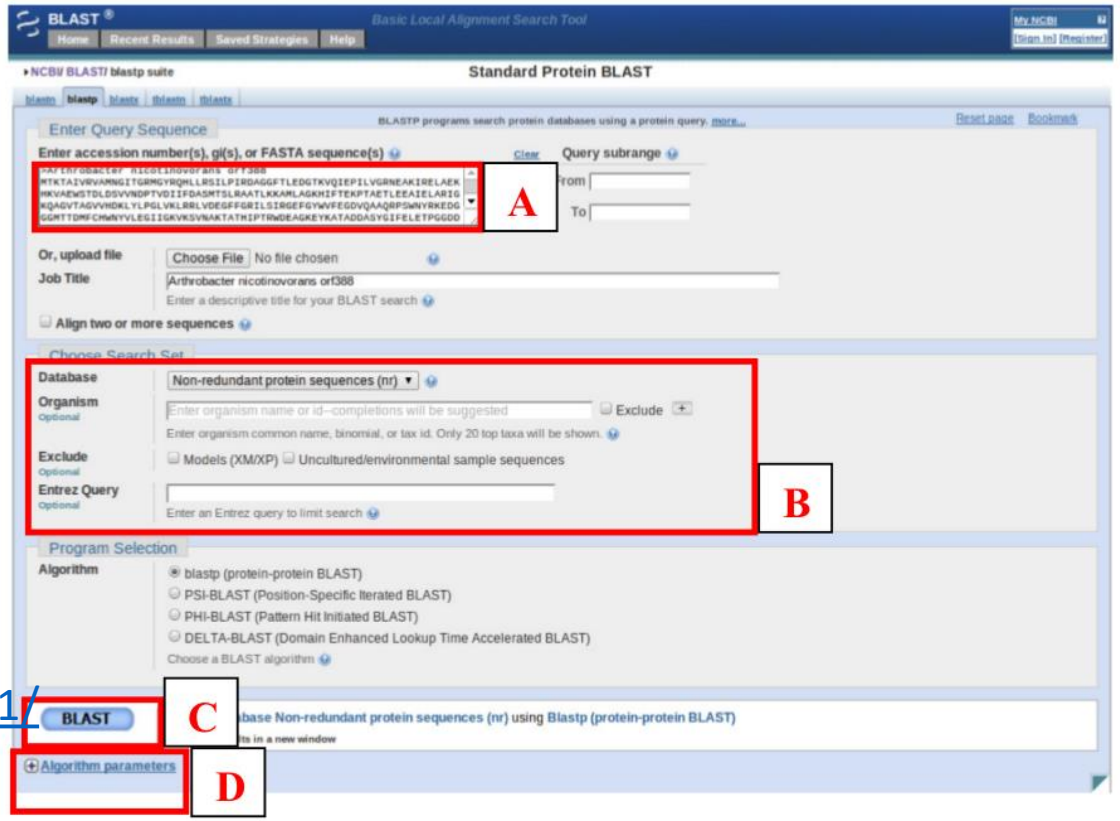


1. Accesează: https://blast.ncbi.nlm.nih.gov/Blast.cgi

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



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/

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