

Βιοπληροφορική

16



# Περιεχόμενα παρουσίασης

- Βιολογικό υπόβαθρο.
- Το κεντρικό αξίωμα.
- Σύνοψη της Βιοπληροφορικής.
- Ερευνητικές περιοχές.
- Πηγές πληροφοριών.



# Τι είναι η Βιοπληροφορική

- Βιο —————> μοριακή βιολογία
- Πληροφορική —————> επιστήμη των υπολογιστών
- Βιοπληροφορική —————> επιλύει προβλήματα που προκύπτουν από τη βιολογία χρησιμοποιώντας μεθοδολογία από την επιστήμη υπολογιστών

# Βιοπληροφορική και συναφείς επιστήμες



- Επιστήμη των υπολογιστών.
- Μαθηματικά και Στατιστική.
- Βιολογία.
- Ιατρική.
- Χημεία.
- Φυσική.



# Η Βιοπληροφορική σχετίζεται με:

- Την εξαγωγή ακολουθιών DNA από το γενετικό υλικό.
- Τον σχολιασμό ακολουθιών (π.χ. με πληροφορίες από πειράματα).
- Την κατανόηση του ελέγχου της έκφρασης των γονιδίων (δηλαδή κάτω από ποιες συνθήκες δημιουργούνται οι πρωτεΐνες από το DNA).
- Την σχέση μεταξύ της αλληλουχίας αμινοξέων των πρωτεϊνών και της δομής τους.



# Στόχοι της έρευνας στη Βιοπληροφορική

- Κατανόηση της λειτουργίας των ζωντανών όντων.
- Σχεδιασμός φαρμάκων.
- Αναγνώριση γενετικών παραγόντων κινδύνου.
- Γονιδιακή θεραπεία.
- Γενετική τροποποίηση φυτών και ζώων.
- Βελτίωση μέσων βιολογικού πολέμου.

## Βασικά ζητήματα

- Πώς θα ωφεληθεί η ανθρωπότητα.
- Γενετικά μεταλλαγμένα σπαρτά - αποφυγή μόλυνσης.
- Γενετικά μεταλλαγμένη τροφή και επιπτώσεις.
- Γονίδια και επιπτώσεις στη συμπεριφορά.
- Όρια στις δοκιμές στα ζώα.
- Γενετική θεραπεία – ξεπερνούν τα πλεονεκτήματα τους κινδύνους;

# Βασική αρχή της θεωρίας της εξέλιξης



Όλοι οι οργανισμοί προήλθαν από έναν κοινό πρόγονο μέσα από μια διαδικασία που συνεχίζεται και καθορίζεται από τρεις βασικούς παράγοντες:

- Κληρονομικότητα.
- Διαφοροποίηση.
- Επιλογή.

## Γενετικό υλικό

- Όλα τα κληρονομικά χαρακτηριστικά ενός οργανισμού βρίσκονται στο γενετικό υλικό του μέσα σε κάθε κύτταρό του.
- Εκεί αναπαρίστανται σε έναν απλό κώδικα τεσσάρων στοιχείων (γραμμάτων) που ονομάζεται γενετικός κώδικας.



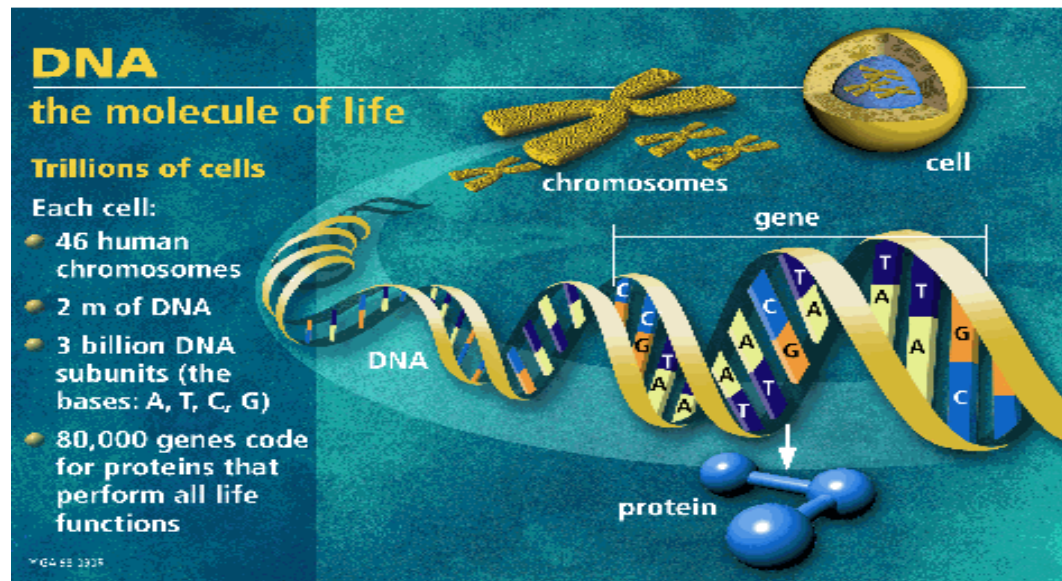
# Κύτταρο και βασικά είδη ζωής

- Το κύτταρο είναι η βασική μορφή ζωής.
- Όλοι οι οργανισμοί αποτελούνται από κύτταρα.
- Ανάλογα με τη δομή των κυττάρων τους, οι βασικές μορφές ζωής κατατάσσονται στις παρακάτω κατηγορίες:
  - Ιοί.
  - Αρχαία.
  - Βακτήρια.
  - Ευκάρυα (τα φυτά, τα ζώα και οι άνθρωποι είναι ευκάρυα).



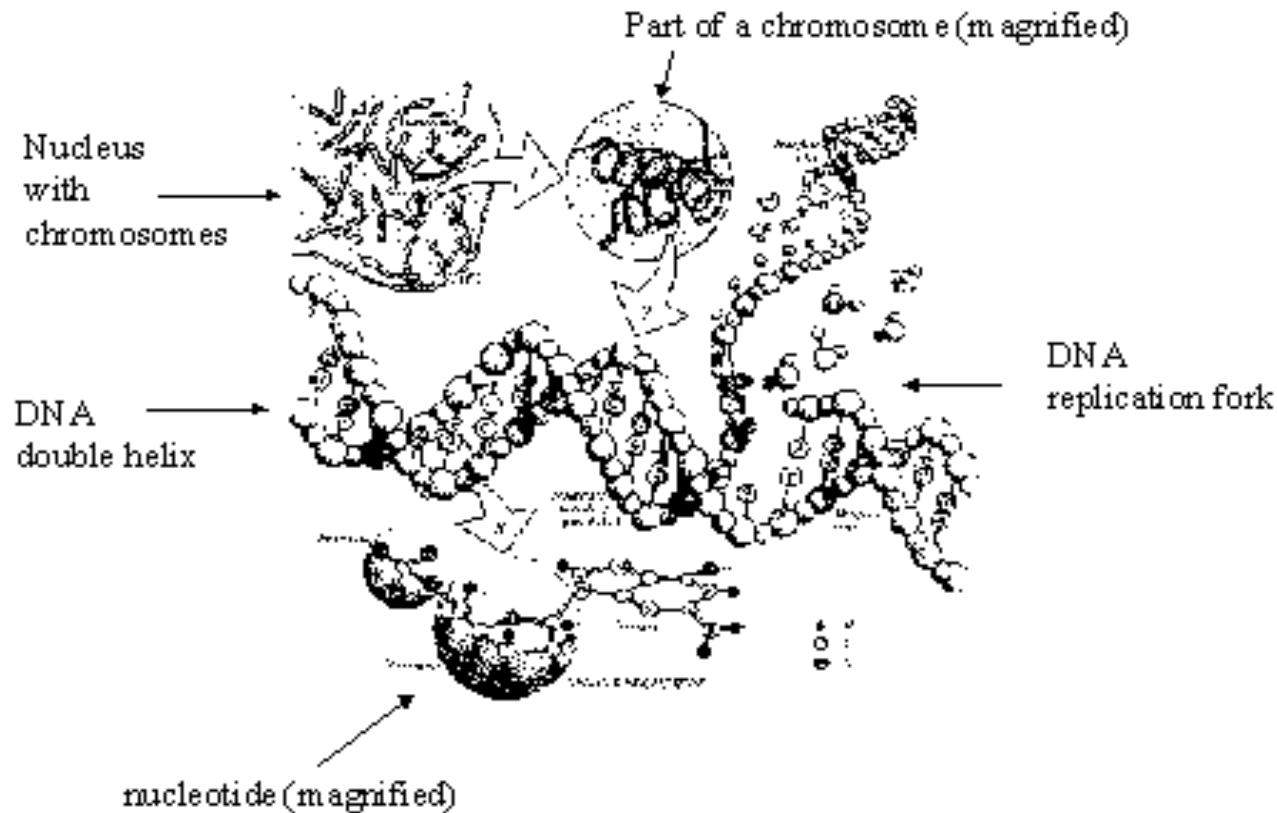
# Δομή του κυττάρου

- Όλα τα κύτταρα διαθέτουν εξωτερική μεμβράνη που τα χωρίζει από το περιβάλλον και εσωτερικό κυτταρόπλασμα.
- Οι ευκαρυωτικοί οργανισμοί διαθέτουν πυρήνα ο οποίος περιέχει το γενετικό υλικό.
- Τα κύτταρα των πολυκύτταρων οργανισμών διαφοροποιούνται αποκτώντας πάνω από 200 είδη.





# The Human Genome at Four Levels of Detail



David Gilbert, 2002

4

## Τι είναι το χρωμόσωμα

- Το DNA αποτελείται από χρωμοσώματα.
- Οι προκαρυώτες, μονοκύτταροι οργανισμοί χωρίς πυρήνα, έχουν μόνο ένα κυκλικό χρωμόσωμα.
- Οι ευκαρυώτες, οργανισμοί με πυρήνα, έχουν αριθμό χρωμοσωμάτων ανάλογα με το είδος.

## Τι είναι το γονιδίωμα

- Είναι το σύνολο του DNA για ένα δεδομένο είδος.
- Ο άνθρωπος έχει 46 χρωμοσώματα σε 23 ζεύγη.
- Κάθε κύτταρο περιέχει το πλήρες γονιδίωμα ενός οργανισμού εκτός από τα κύτταρα που καθορίζουν το φύλο (γαμετικά κύτταρα-έχουν μισό γονιδίωμα) και τα ερυθρά αιμοσφαίρια.



# DNA και RNA

- DNA  $\longrightarrow$  διοξυριβονουκλεϊκό οξύ
- RNA  $\longrightarrow$  ριβονουκλεϊκό οξύ
- Τα βιολογικά μακρομόρια (DNA/RNA) αποτελούνται από αλυσίδες νουκλεοτιδίων, τα οποία περιλαμβάνουν αζωτούχες βάσεις:

*nucleotides = sugar+phosphate+base*

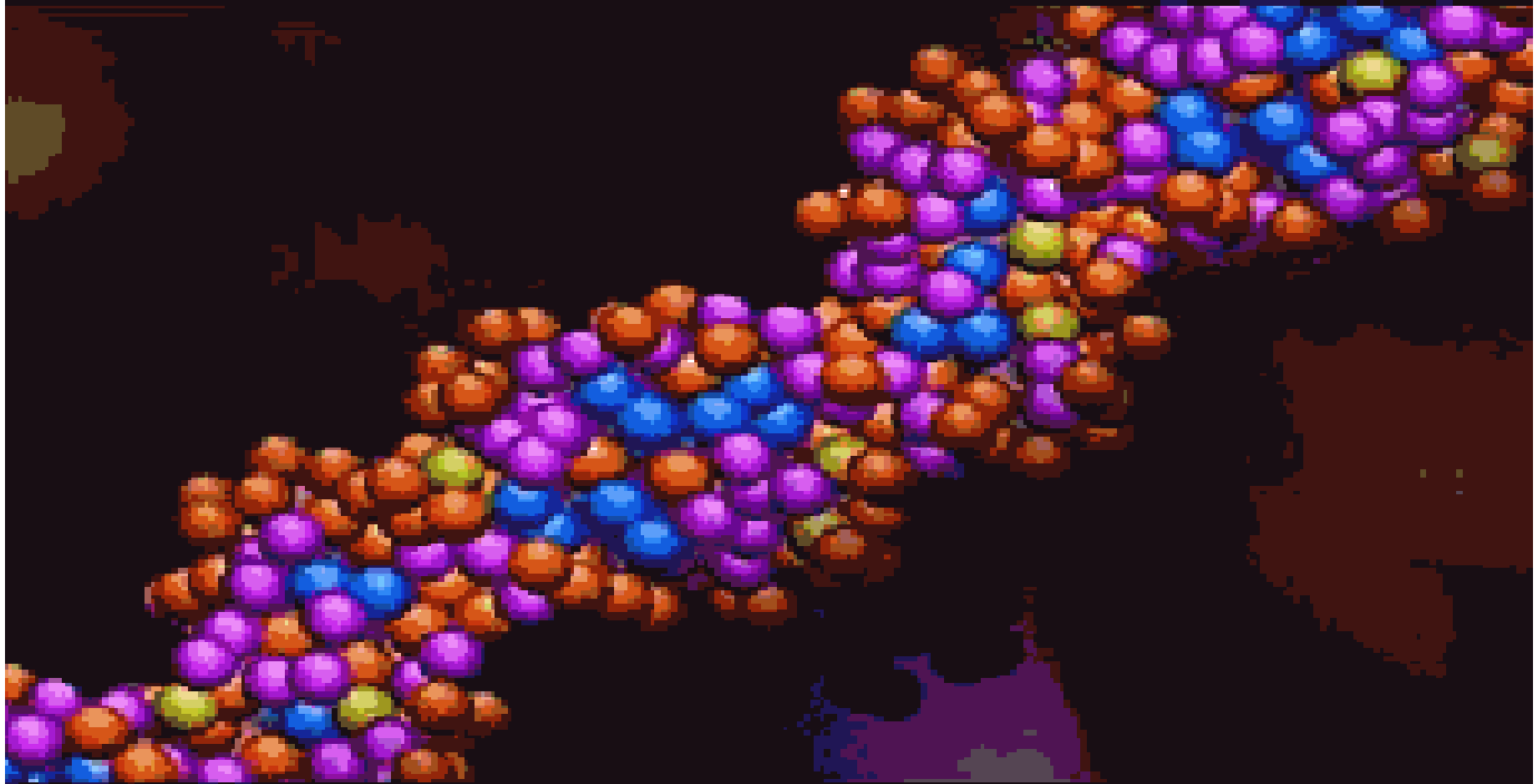


**Bases: adenine (A), cytosine (C), guanine (G), thymine (T), uracil (U)**

4 letter alphabet:

DNA = A C G T (adenine    cytosine    guanine    thymine)

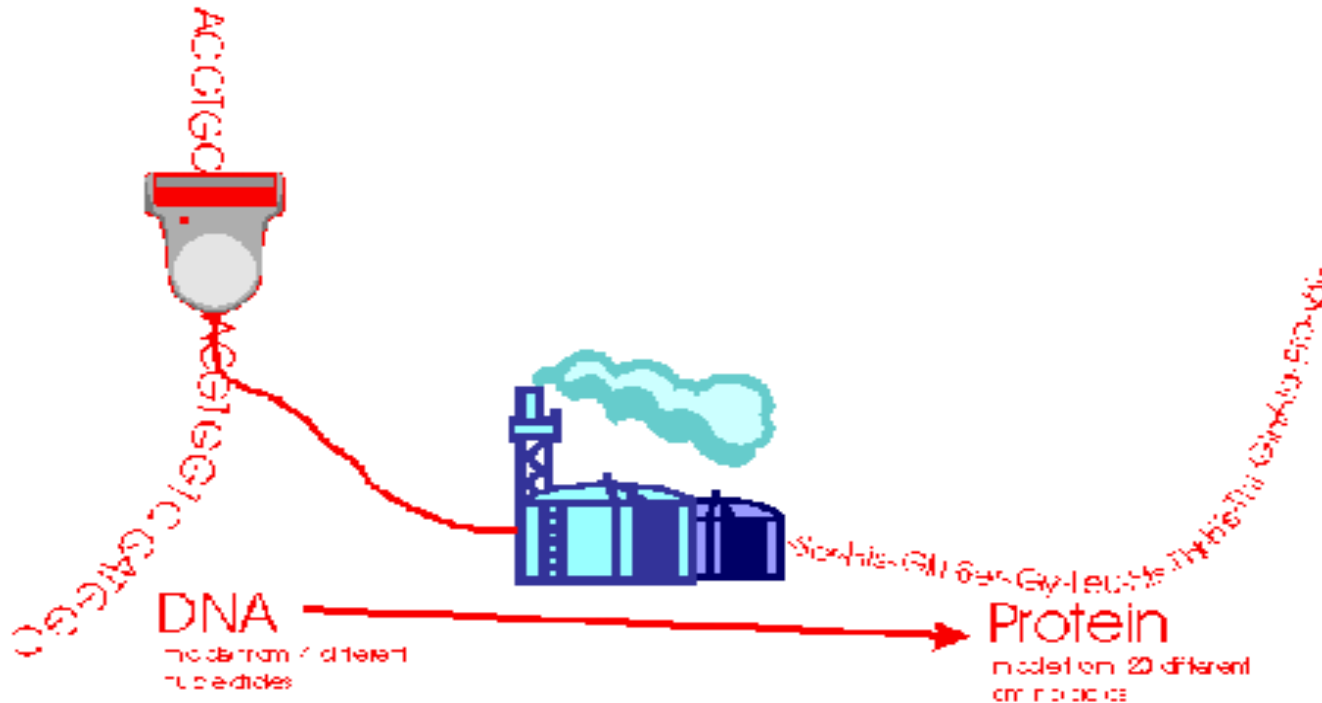
RNA = A C G U (adenine    cytosine    guanine    uracil)



# Ροή πληροφορίας στη μοριακή βιολογία



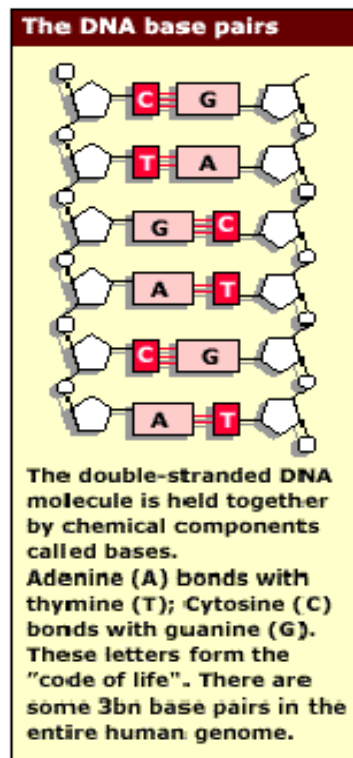
DNA → RNA → Protein → Function





# Δομή αλυσίδας DNA

Η αλυσίδα του DNA αποτελείται από ζεύγη βάσεων που είναι συμπληρωματικές μεταξύ τους:



DNA  
complementarity  
(base-pairing)

A-T  
C-G

David Gilbert, 2002



# Ορισμένα γεγονότα

- Το DNA διαφέρει ανάμεσα σε ανθρώπους κατά 0.2% (1 βάση σε κάθε 500).
- Το ανθρώπινο DNA είναι κατά 98% πανομοιότυπο με αυτό των χιμπατζήδων.
- Μεγάλο μέρος είναι μη κωδικοποιητικό DNA
- Υπάρχουν  $10^{14}$  κύτταρα στο σώμα.
- 12000 γράμματα του DNA αποκωδικοποιούσε κάθε δευτερόλεπτο το Human Genome Project.



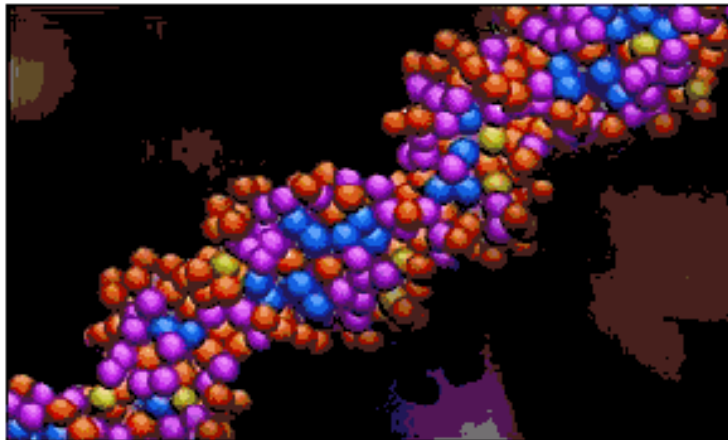
# Human Genome Project

- Παγκόσμια προσπάθεια χαρτογράφησης του ανθρώπινου DNA.
- Το project άρχισε το 1990 και ολοκληρώθηκε το 2003.
- Στόχος να αναγνωριστούν και τα 20000 περίπου γονίδια του ανθρώπινου οργανισμού και να διαβαστούν όλες οι ακολουθίες DNA.
- Στόχος να τοποθετηθεί όλη η πληροφορία σε βάσεις δεδομένων και να βελτιωθούν τα εργαλεία για την ανάλυσή της.



# Ακολουθία βάσεων DNA

```
AAAAAGAAA AGGTT AGAAA GAT GAGAGAT GAT AAAGGGT CCAT TTGAG GTT AGGT AAT AT GGT TT GGT AT CCCTGT AGTT AAAAG  
TTTTTGTCTT ATT TTAG A AT ACTGT GACT ATTTCTTT AGTAT T AATT TTT CCTT CTGTTTT CCTC AT CT AGG GAACCCCAAGAG  
CATCCAMT AGAAG CTGTG CAATT ATGT AAAATTTT CAACTGT CTT CCTC AAAAT AAAGAAGT ATGGT AATCT TTACCTGT ATAC  
AGTGCAGAGCCTT CTCAG AAGCACAGAAT ATTTTT ATATTT C CTTT ATGT GAATTTTT AAGCTGC AA ATCTG MTGGCCTT AATT  
TCCTTTTT GACACTGAAA GTTTTTGT AAAAGAAAATCATGTCCATACAC TTTGTTGCAAGATGTGAATT ATTGACACTGAACTT AA  
TAACTGTGTACTGTTCGG AAGGGGTT CCTCAAATTTTTT GACTTTTT TTGT ATGTGT GTTTTTT CTT TTTTT TT AAGTTCTT AT  
GAGGAGGG AGGGT AATTA AACC ACTGTGCGTCTTGGTGT AAT TTGAA GATTGCCCCATCTAGACT AGCAATCTCTT CATT ATTC  
TCTGCT AT AT AT AAAACG GTGCTGTGAGGGAGGGGAAAAGCATTTTT CAAT AT ATTGAACTTTT GACT GAATTTTTTTGT AAT  
AAGCAATC AAGGT T AT AATTTTTTTTT AAAAT AGAAATTTTGT AAGAA GGC AAT ATT AACCT AATCAC CATGT AAGCACTCTGGA  
TGATGGAT TCCAC AAAACTTGGTTTT ATGGTTACTTCTTCTTAGATTCTT AATT CATGAGGAGGGTGGGG GAGGGAGGTGGA  
GGGAGGGAAGGGT TTTCTCTATT AAAATGCATTGTTGTGTTTT TTAAGATAGTGTAACTTGCTAAATTTCTT ATGTGACATTAA  
CAAAATAAAA AAGCTCTTTT AATATTAGATAA
```



David Gilbert, 2002



# Σύγκριση μεγεθών ακολουθιών

Comparative Sequence Sizes	(Bases)
(yeast chromosome 3)	350 Thousand
Escherichia coli (bacterium) genome	4.6 Million
Largest yeast chromosome now mapped	5.8 Million
Entire yeast genome	15 Million
Smallest human chromosome (Y)	50 Million
Largest human chromosome (1)	250 Million
Entire human genome	3 Billion

# Ανθρώπινες γενετικές διαφοροποιήσεις (Single Nucleotide Polymorphisms)



- Οδηγούν στη γενετική μοναδικότητα κάθε ανθρώπου.
- Κάνουν τους ανθρώπους λιγότερο ή περισσότερο ευάλωτους σε αρρώστιες.
- Μπορεί να επηρεάσουν το αποτέλεσμα των φαρμακευτικών αγωγών.

TTTGCTCCGTTTTCA  
TTTGCTCYGTTTTCA  
TTTGCTCTGTTTTCA

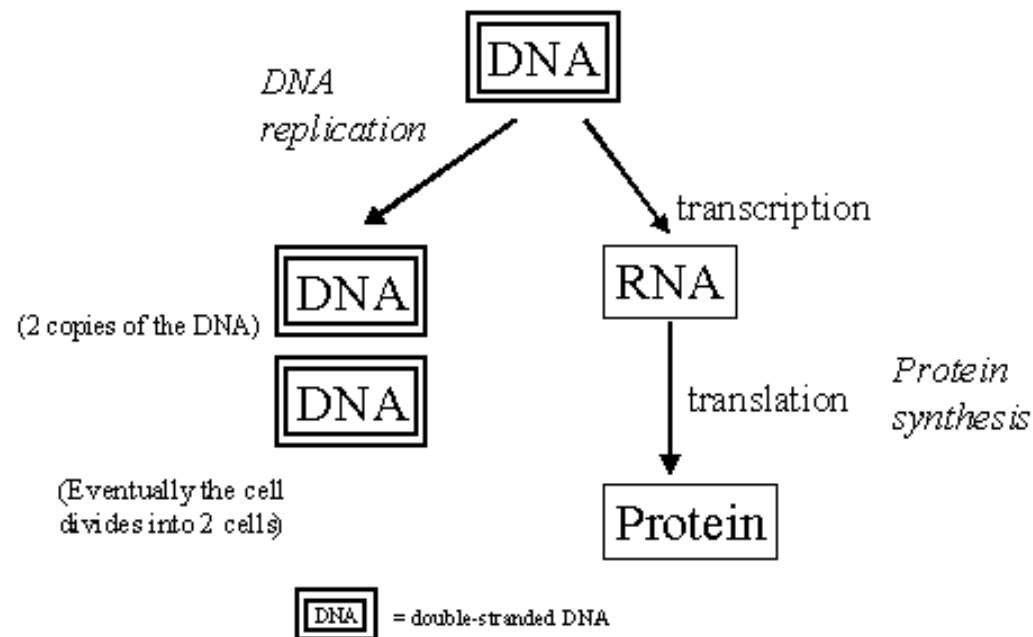
## HUMAN INDIVIDUALITY





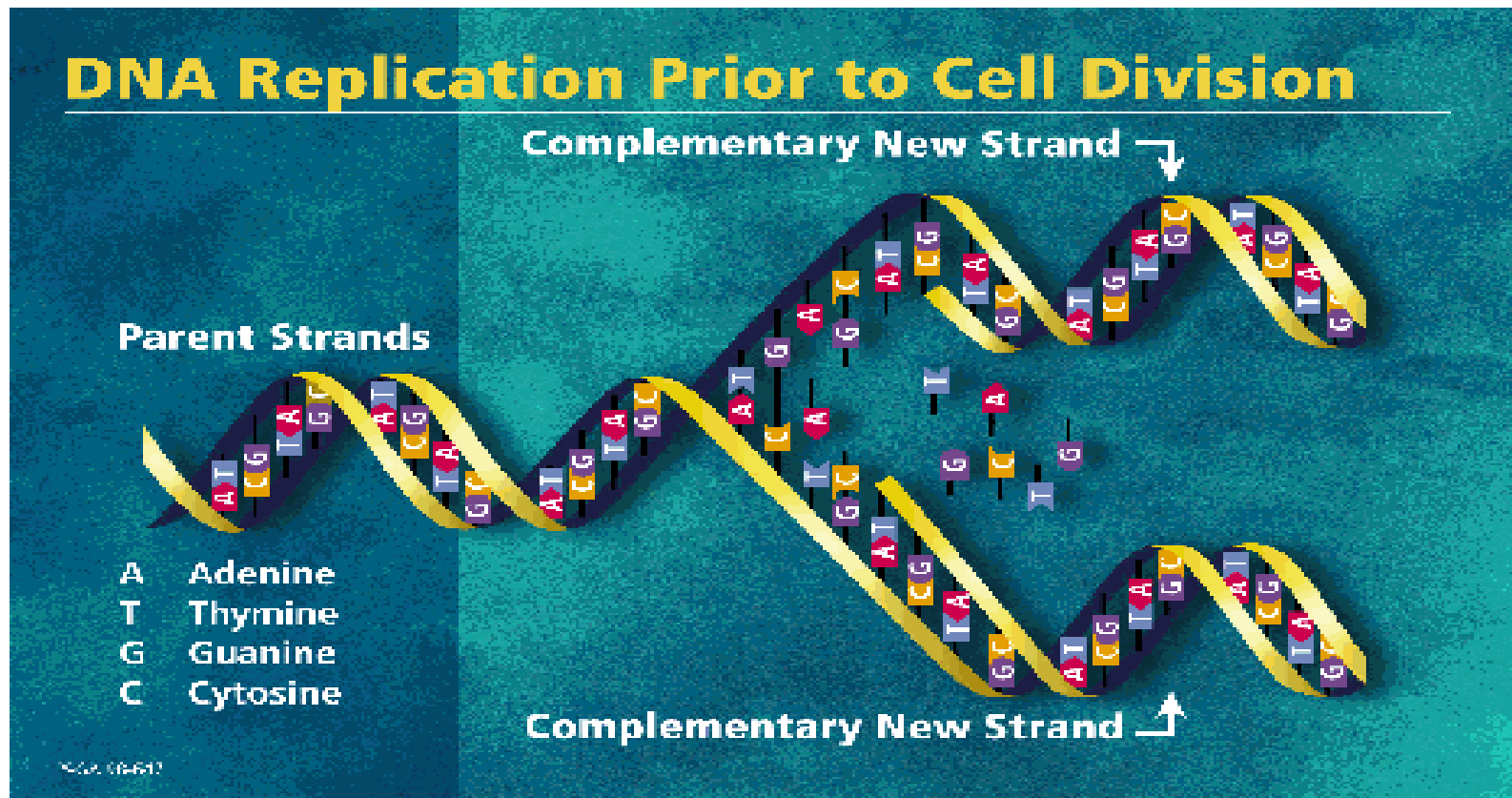
# Τι μπορεί να συμβεί στο DNA

- Αντιγραφή.
- Μεταγραφή στο RNA και μετάφραση σε πρωτεΐνη.





# Αναπαράσταση αντιγραφής DNA





# Μεταγραφή και μετάφραση

- Μεταγραφή:  
Η διαδικασία της αντιγραφής του DNA στο RNA.
- Μετάφραση:  
Σύνθεση της πρωτεΐνης από το mRNA.
- Ριβοσώματα ονομάζονται τα σωματίδια του κυτταροπλάσματος που συνθέτουν πρωτεΐνες από το mRNA.

## Τι είναι γονίδιο

- Βασική μονάδα κληρονομικότητας.
- Ακολουθία βάσεων η οποία μεταφέρει την πληροφορία που απαιτείται για να δημιουργηθεί μια συγκεκριμένη πρωτεΐνη.
- Ένα γονίδιο κωδικοποιεί μια πρωτεΐνη ή ένα μόριο RNA.

## Αριθμοί γονιδίων

Άνθρωποι και ποντίκια:	19000 – 25000
C.elegans (worm):	19000
S.cerevisiae (yeast):	6000
Tuberculosis microbe:	4000



# Πρωτεΐνες

- Αποτελούν ~60% της μάζας ενός κυττάρου.
- Γραμμικά ετεροπολυμερή.
- Αποτελούνται από αλυσίδα αμινοξέων (20 διαφορετικά είδη).
- Η λειτουργία των πρωτεϊνών (και του RNA) καθορίζεται από τη δομή τους.
- Η δομή καθορίζεται από την αλληλουχία των αμινοξέων (των νουκλεοτιδίων για το RNA).



# Ακολουθία αμινοξέων (πρωτεΐνη)

```
DR   EMBL; U22530; AAA64218.1; -.
DR   HSSP; P03122; 2BOP.
KW   Oxidoreductase; Dioxygenase; Flavoprotein; FMN.
FT   PROPEP           1      15      POTENTIAL.
FT   CHAIN            16     378     2-NITROPROPANE DIOXYGENASE.
SQ   SEQUENCE      378 AA;  39916 MW;  E453EB43FD23E441 CRC64;
MHFPGHSSKK EESAQAALTK LNSWFPTTKN PVIISAPMYL IANGTLAAEV SKAGGIGFVA
GGSDFRPGSS HLTALSTELA SARSRLGLTD RPLTPLPGIG VGLILTHTIS VPYVTDTVLP
ILIEHSPQAV WLFANDPDFE ASSEPGAKGT AKQIIEALHA SGFVVFFQVG TVKDARKAAA
DGADVIVAQG IDAGGHQLAT GSGIVSLVPE VRDMLDREFK EREVVVVAAG GVADGRGVVG
ALGLGAEGVV LGTRFTVAVE ASTPEFRRKV ILEINDGGLN TVKSHFHDQI NCNTIWHNVY
DGRAVRNASY DDHAAGVPFE ENHKKFKEAA SSGDNSRAVT WSGTAVGLIK DQRPAGDIVR
ELREEAKERI KKIQAFAA
```



# Γενετικός κώδικας

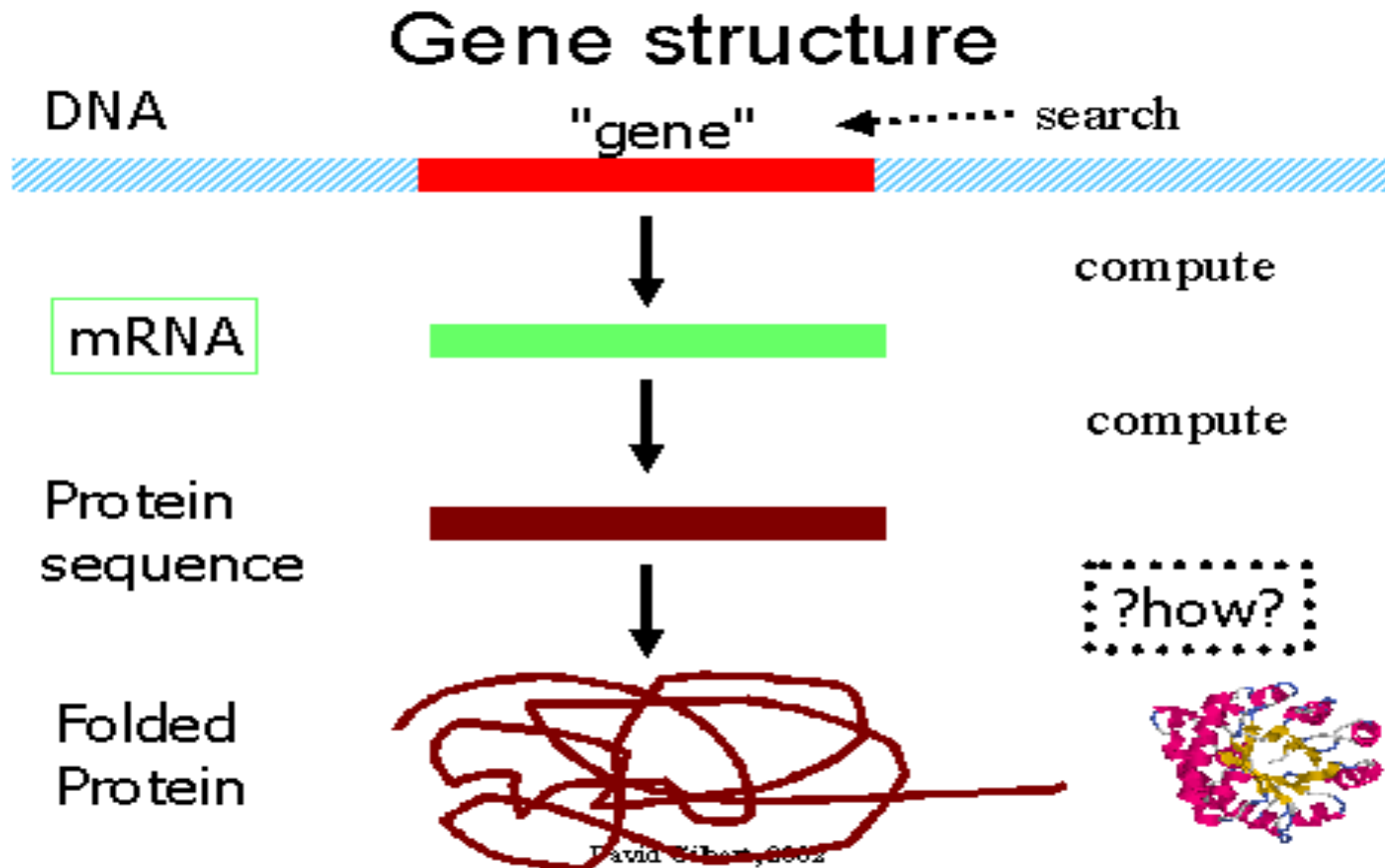
## The genetic code

First Position (5' end)	Second Position								Third Position (3' end)
	T		C		A		G		
T	<b>TTT</b>	Phe	<b>TCT</b>	Ser	<b>TAT</b>	Tyr	<b>TGT</b>	Cys	T
	<b>TTC</b>	Phe	<b>TCC</b>	Ser	<b>TAC</b>	Tyr	<b>TGC</b>	Cys	C
	<b>TTA</b>	Leu	<b>TCA</b>	Ser	<b>TAA</b>	<i>Stop</i>	<b>TGA</b>	<i>Stop</i>	A
	<b>TTG</b>	Leu	<b>TCG</b>	Ser	<b>TAG</b>	<i>Stop</i>	<b>TGG</b>	Trp	G
C	<b>CTT</b>	Leu	<b>CCT</b>	Pro	<b>CAT</b>	His	<b>CGT</b>	Arg	T
	<b>CTC</b>	Leu	<b>CCC</b>	Pro	<b>CAC</b>	His	<b>CGC</b>	Arg	C
	<b>CTA</b>	Leu	<b>CCA</b>	Pro	<b>CAA</b>	Gln	<b>CGA</b>	Arg	A
	<b>CTG</b>	Leu	<b>CCG</b>	Pro	<b>CAG</b>	Gln	<b>CGG</b>	Arg	G
A	<b>ATT</b>	Ile	<b>ACT</b>	Thr	<b>AAT</b>	Asn	<b>AGT</b>	Ser	T
	<b>ATC</b>	Ile	<b>ACC</b>	Thr	<b>AAC</b>	Asn	<b>AGC</b>	Ser	C
	<b>ATA</b>	Ile	<b>ACA</b>	Thr	<b>AAA</b>	Lys	<b>AGA</b>	Arg	A
	<b>ATG</b>	Met*	<b>ACG</b>	Thr	<b>AAG</b>	Lys	<b>AGG</b>	Arg	G
G	<b>GTT</b>	Val	<b>GCC</b>	Ala	<b>GAT</b>	Asp	<b>GGT</b>	Gly	T
	<b>GTC</b>	Val	<b>GCC</b>	Ala	<b>GAC</b>	Asp	<b>GGC</b>	Gly	C
	<b>GTA</b>	Val	<b>GCA</b>	Ala	<b>GAA</b>	Glu	<b>GGA</b>	Gly	A
	<b>GTG</b>	Val	<b>GCG</b>	Ala	<b>GAG</b>	Glu	<b>GGG</b>	Gly	G

David Gilbert, 2002



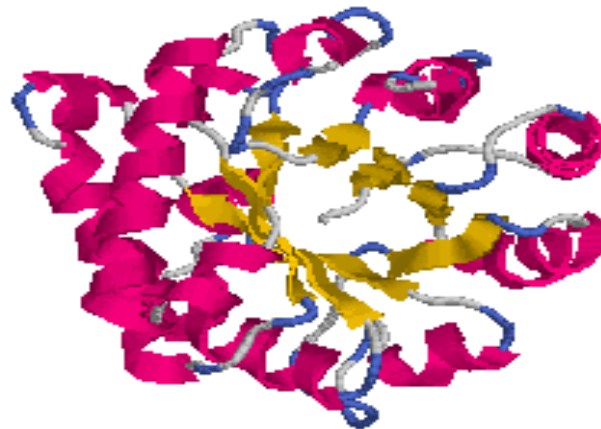
# Από το γονίδιο στην πρωτεΐνη





# Δομή πρωτεΐνης

## Protein structure



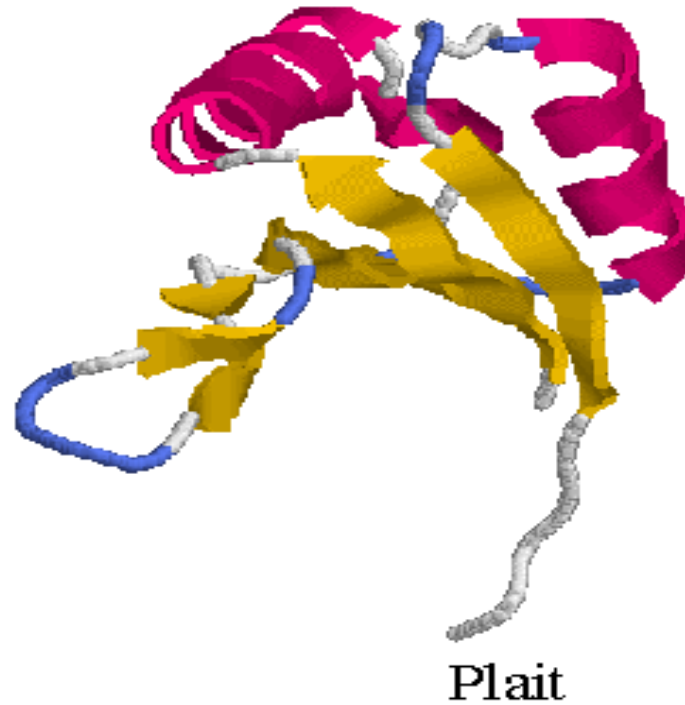
Sizes: 50 - 500 amino-acids (Globular proteins)

(c) David Gilbert, Aik Choon Tan, Gilleain Torrance and Mallika Veeramalai 2002



# Δομή πρωτεΐνης

## Protein structure



Πρωτογενής δομή (γραμμική ακολουθία αμινοξέων)  
Δευτερογενής δομή  
Τριτογενής δομή  
Τεταρτογενής δομή

δομές αμινοξέων στο χώρο ολόένα και πιο σύνθετες

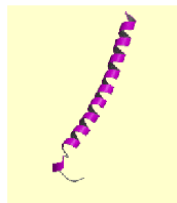
## Primary Structure

```
VHLTPEEKSAVTALWGKVVNDE  
VGGEALGRLLVYPWTQRFES  
FGDLSTPDVAMGNPKVKAHGKK  
VLGAFSDGLAHLNLIKGTATLS  
ELHCDKLHVDPENFRLLGNVLV  
CVLAHHFGKEFTPPVQAAYQKV  
VAGVANALAHKYH
```

amino acids (1hbb - Haemoglobin B)

Ask Choon Tan & David Gilbert, 2001

## Secondary Structure



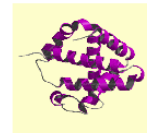
alpha helix



beta sheets

Ask Choon Tan & David Gilbert, 2001

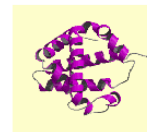
## Tertiary Structure



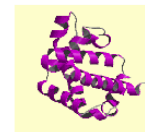
1hbbA0



1hbbB0



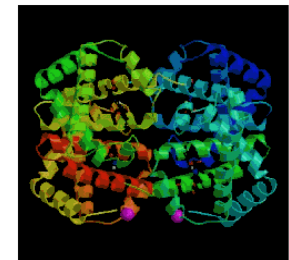
1hbbC0



1hbbD0

Ask Choon Tan & David Gilbert, 2001

## Quaternary Structure



Hemoglobin (1hbb)

Ask Choon Tan & David Gilbert, 2001



# Κεντρικό αξίωμα ροής πληροφορίας στη βιολογία

Η ακολουθία αμινοξέων που αποτελεί μια πρωτεΐνη και επιπλέον προσδιορίζει τη δομή και τη λειτουργία της καθορίζεται από μεταγραφή του DNA μέσω του RNA.

**DNA → RNA → Πρωτεΐνη**

## Ένας σημαντικός στόχος

Ανάπτυξη υπολογιστικών μεθόδων ώστε να καθοριστεί η δομή των πρωτεϊνών από την αλληλουχία των αμινοξέων.



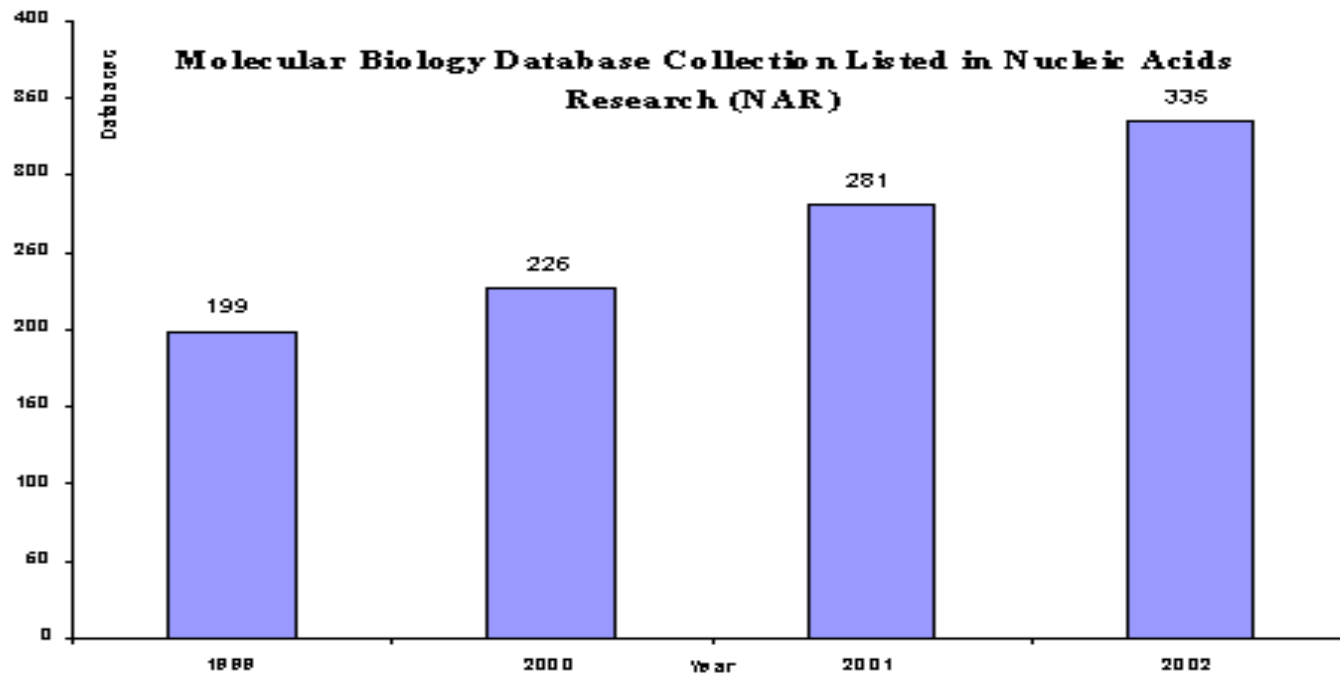
# Κατηγορίες περιοχών προβλημάτων

- Περιοχές σχετιζόμενες με το κεντρικό αξίωμα: ακολουθία, δομή ή λειτουργία.
- Περιοχές σχετιζόμενες με τα δεδομένα: αποθήκευση, ανάκτηση και ανάλυση (γιγάντωση της γνώσης στη μοριακή βιολογία).
- Περιοχές σχετιζόμενες με την προσομοίωση των βιολογικών διαδικασιών (protein folding or metabolic pathways).

# Ραγδαία αύξηση των βάσεων δεδομένων



## Growth of Molecular biological databases



(c) David Gilbert & Aik Choon Tan, 2002

11



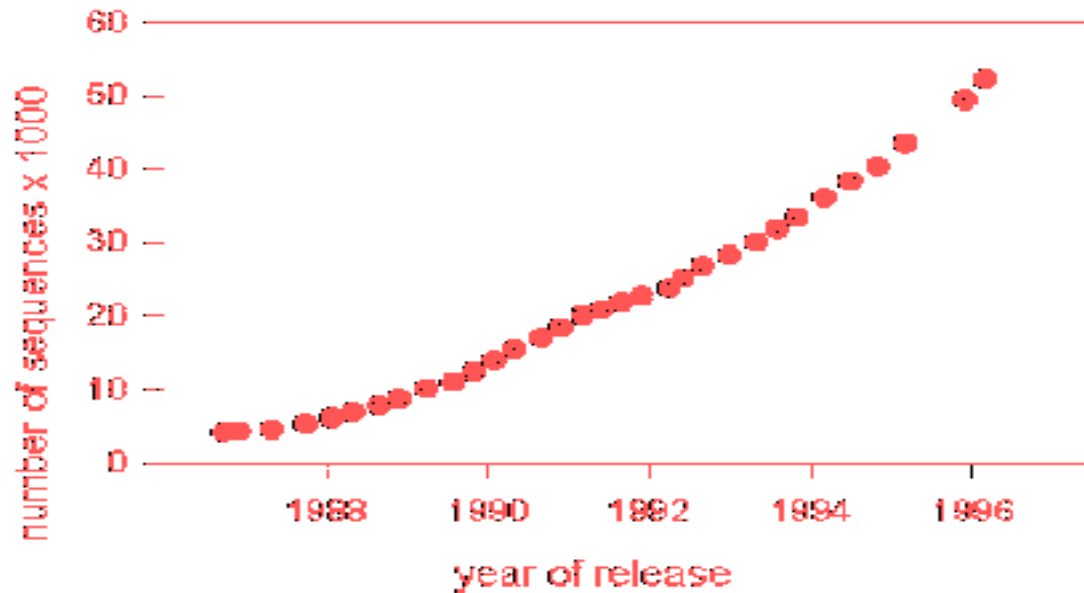
# Μεγάλος όγκος δεδομένων

Βάση δεδομένων	Καταχωρήσεις δεδομένων	Μέγεθος δεδομένων σε GB
EMBL (ακολουθίες DNA)	15.242.454 ακολουθίες (14/1/02)	7,5
PIR (ακολουθίες πρωτεϊνών)	250.417 ακολουθίες (30/9/01)	0,2
PDB (ακολουθίες πρωτεϊνών)	17.022 ακολουθίες (8/1/02)	17



# Αύξηση όγκου δεδομένων βάσης

## Flood of data! (SWISSPROT)





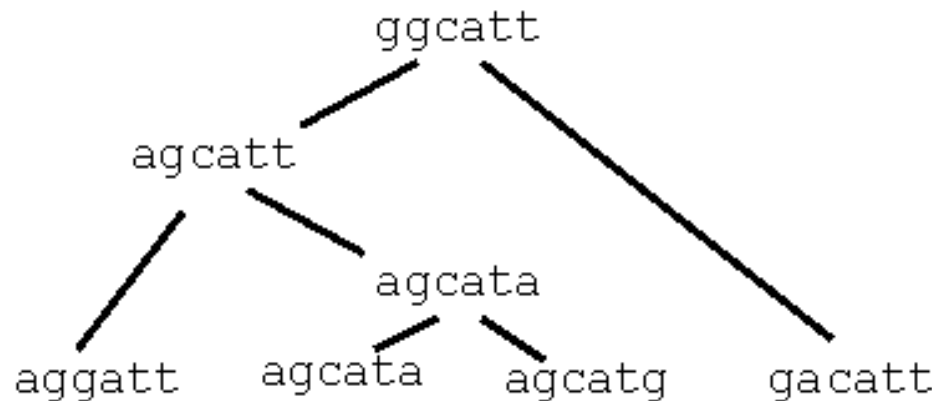
# Ανάλυση του όγκου δεδομένων

- Εκτός από την αποθήκευση μας ενδιαφέρει και η ανάλυση των δεδομένων.
- Με σύγκριση των ακολουθιών, προκύπτουν στοιχεία όπως τα παρακάτω:
  - Πρόγονοι οργανισμών.
  - Φυλογενετικά δέντρα.
  - Δομές πρωτεϊνών.
  - Λειτουργία πρωτεϊνών.



# Φυλογενετικά δέντρα

- Οπτικοποιούν την εξέλιξη από έναν κοινό πρόγονο.
- Η εξέλιξη προχωράει μέσω αναπαραγωγής και σταδιακής μετάλλαξης.
- Η Βιοπληροφορική ανιχνεύει μακρινές ομοιότητες ανάμεσα σε οργανισμούς .



# Πρόβλεψη δομής πρωτεΐνης (protein folding problem)



Ακολουθία  $\longrightarrow$  Δομή  $\longrightarrow$  Λειτουργία

Προσεγγίσεις:

- Βιοχημική προσέγγιση.
- Προσομοίωση (μοριακά δυναμικά).
- Πρόβλεψη με ευρεστικές μεθόδους/απλοποιημένα μοντέλα.
- Σχεδιασμός φαρμάκων και άλλες ιατρικές χρήσεις.

# Ελαχιστοποίηση ενέργειας πρωτεΐνης

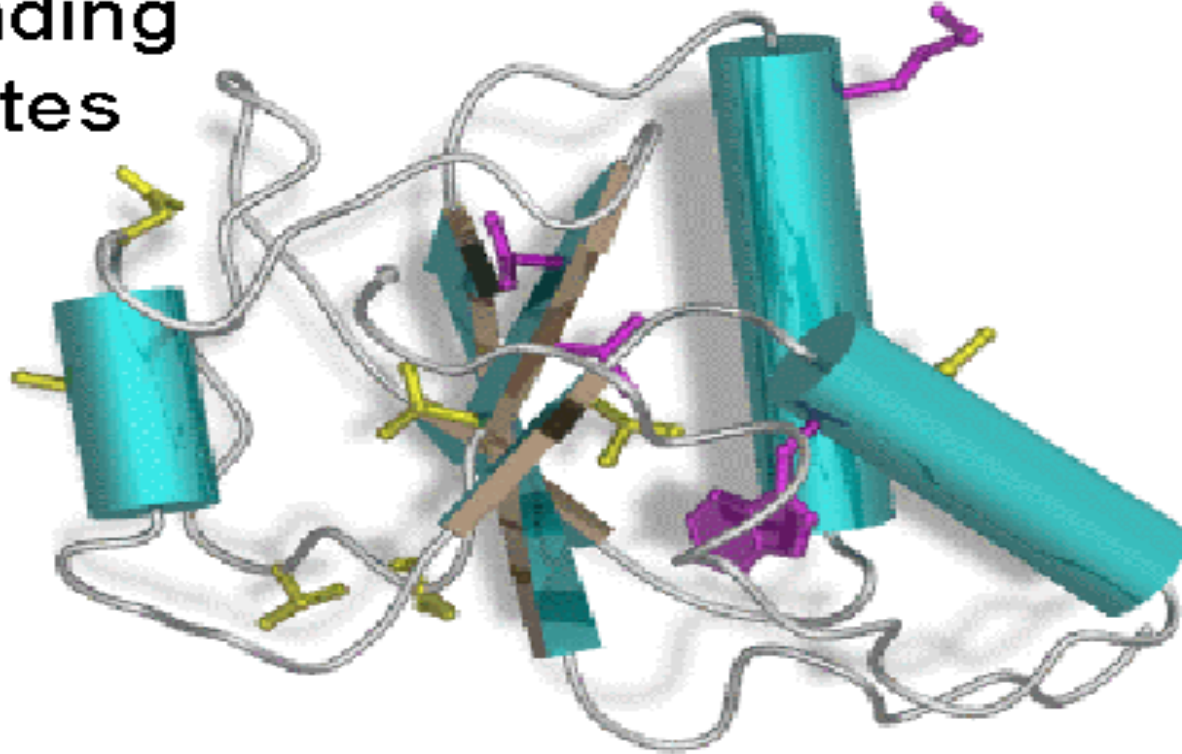


- Το πρόβλημα συνίσταται στην εύρεση της πιο σταθερής κατάστασης σύνδεσης μεταξύ δυο μορίων πρωτεΐνης, ξεκινώντας από τις ατομικές συντεταγμένες των δυο πιο απομονωμένων στοιχείων.
- Στόχος είναι η ελαχιστοποίηση της ενέργειας λόγω της αλληλεπίδρασης με το διαλυτικό μέσο.
- Υπάρχουν διάφορες προσεγγίσεις για να βρεθούν οι θέσεις σύνδεσης (binding sites).



# Θέσεις σύνδεσης

Binding sites



## Άλλα προβλήματα

- Έρευνα και ανακάλυψη προτύπων σε ακολουθίες.
- Βιολογικές προσομοιώσεις.
- Πίνακες γονιδιακής έκφρασης.
- Πηγές πληροφοριών και ανάγνωσης.

## Έρευνα και ανακάλυψη προτύπων

- Λειτουργικά σημαντικές περιοχές, που επαναλαμβάνονται σε διαφορετικές ακολουθίες, συνήθως περιγράφονται από πρότυπα.
- Μπορεί να γίνει έρευνα σε βάσεις δεδομένων για καταχωρήσεις που ταιριάζουν στο πρότυπο.
- Τα βιολογικά δεδομένα έχουν πολύ θόρυβο. Μπορούμε να χρησιμοποιήσουμε γλώσσες συμβολοσειρών ή στοχαστικές προσεγγίσεις.

# Διερεύνηση ακολουθιών και ευθυγράμμιση (sequence alignment)



## Sequence alignment problem

T C A T G  
/ / / |  
C A T **T** G

T C A T G  
/ / | |  
C A **T** T G



# Multiple alignment

		3	5	7	9	1	1	1	1	1	2	2	2	2
F37262				A E	L M K	P G	A S	V K I	S C	K A	T G	Y K	F S	S
B27563		Q P	G A E	L V K	P G	A S	V K L	S C	K A	S G	Y T	F T	N	
C30560		Q S	G A E	L V K	P G	A S	V K I	S C	K A	S G	Y T	F T	S	
G1HUDW		E S	G P A	L V R	P T	Q T	L T L	T C	T F	S G	F S	L S	G	
S09711	l l v r a p r w c l s q v q l q	E S	G P G	L V K	P S	E T	L S V	T C	T V	S G	G S	V S	S	g
B36006								K I	S C	K G	S G	Y S	F T	S
F36005		E S	G G G	V V Q	P G	R S	L R L	S C	A A	S G	F T	F S	S	
A36194	l f l l s v t a g v h s e v q l q	Q S	G A E	L V R	A G	S S	V K M	S C	K A	S G	Y T	F T	N	
A31485		E T	G G G	L V Q	P G	R P	M K L	S C	V A	S G	F T	F S	D	
D33548		Q S	G A E	V K K	P G	A S	V K V	S C	E A	S G	Y T	F T	G	
AVMSJ5		E S	G G G	L V Q	P G	G S	L K L	S C	A A	S G	F D	F S	K	
D30560		Q S	G P S	L V Q	P S	Q S	L S I	T C	T V	S D	F S	L T	N	
S11239	i f l l a i l k g v q c e v q l v	E S	G G G	L V Q	P G	R S	L R L	S C	A A	S G	F T	F N	D	
G1MSAA		Q S	G A E	L V K	A G	S S	V K M	S C	K A	T G	Y T	F S	S	
I27888		E S	G G G	L V K	P G	G S	L R L	S C	A A	S G	F T	F S	S	
PL0118		E S	G S G	L V K	P S	Q T	L S L	T C	A V	S G	G S	I S	S	g g
PL0122		E S	G G G	L V Q	P G	G S	L K L	S C	A A	S G	F T	F S	G	
A33989		Q S	E S V	V I K	P G	G S	L K L	S C	T A	S G	F T	F S	S	
A30502		Q S	G P E	L V K	P G	A S	V K M	S C	K A	S G	D T	F T	S	
PH0097		E S	G G G	L V K	P G	G S	L K L	S C	A A	S G	F T	F S	S	

# Προγράμματα για συνδυαστική διερεύνηση (pair wise searching)



- FASTA
- BLAST (Basic Local Alignment Search Tool)
- Γίνεται διερεύνηση μία προς μία ακολουθιών σε βάσεις δεδομένων για να βρεθούν ποιες μοιάζουν με τη δεδομένη ακολουθία και σε ποιο βαθμό

## Εισαγωγή δεδομένων στο BLAST

NCBI Blast - Microsoft Internet Explorer

Διερεύνηση: ROGRAM=blastp&SERVICE=plain&SET\_DEFAULTS.x=41&SET\_DEFAULTS.y=5&SHOW\_OVERVIEW=on&END\_OF\_HTTPGET=yes&SHOW\_LINKOUT=yes

NCBI **protein-protein BLAST**  
Nucleotide Protein Translations Retrieve results for an RID

Search: MSIDIDIIKAKARNEYRLSKVRGEAMISVRIIPGGILPAHLLTVARDIAETWNGQIHLTT

Set subsequence From: To:

Choose database: nr

Do CD-Search:

Now: **BLAST!** or [Reset query](#) [Reset all](#)

Options for advanced blasting

Limit by entire query:  or select from: (none)

Composition-based statistics:

Evangelos Bioinformatic... SRSWWW... Microsoft Po... NCBI Bla... Magenta E... 1. Samantha... 12:55 μμ



# Αποτελέσματα BLAST

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi - Microsoft Internet Explorer

Αρχείο Επεξεργασία Προβολή Αγαπημένα Εργαλεία Βοήθεια

Πίσω Εμπρός Διακοπή Ανανέωση Αρχική Αναζήτηση Αγαπημένα Μέσα Ιστορικό Αλληλογραφ... Μέγεθος Εκτύπωση

Δευθυνα http://www.ncbi.nlm.nih.gov/blast/Blast.cgi Μετάβαση Συνδέσεις

### Distribution of 14 Blast Hits on the Query Sequence

AAC17122 (AF065403) nitrite reductase [Synechococcus sp. WH 8103]..S=31.6 E=1.8

Hit	Score (bits)	E Value
1	123	3e-28
2	123	3e-28
3	64	4e-10
4	56	7e-08
5	50	5e-06
6	38	0.025
7	33	0.53
8	33	0.59
9	33	0.68
10	33	0.69
11	32	1.8
12	31	2.2
13	31	2.4
14	31	2.4

Sequences producing significant alignments:

Query	Hit	Score (bits)	E Value
qi 114283 sp P26476 ASRC_SALTY	ANAEROBIC SULFITE REDUCTASE ...	123	3e-28
qi 16761464 ref NP_457081.1	(NC_003198) anaerobic sulfite ...	123	3e-28
qi 18310518 ref NP_562452.1	(NC_003366) anaerobic sulfite ...	64	4e-10
qi 18310420 ref NP_562354.1	(NC_003366) anaerobic sulfite ...	56	7e-08
qi 15894793 ref NP_348142.1	(NC_003030) Anaerobic sulfite ...	50	5e-06
qi 16329461 ref NP_440189.1	(NC_000911) ferredoxin-sulfite...	38	0.025
qi 16080396 ref NP_391223.1	(NC_000964) similar to sulfite...	33	0.53
qi 267002 sp P30008 SIR_SYWP?	SULFITE REDUCTASE (FERREDOXIN...	33	0.59
qi 7007444 db BAA90857.1	(AB031214) YvgQ [Bacillus halodu...	33	0.68
qi 15613173 ref NP_241476.1	(NC_002570) sulfite reductase ...	33	0.69
qi 13152707 gb AAC17122.1	(AF065403) nitrite reductase [Syn...	32	1.8
qi 15893390 ref NP_346739.1	(NC_003030) Ferredoxin-nitrite...	31	2.2
qi 7432744 pir JJE0260	sulfite reductase (ferredoxin) (EC 1...	31	2.4
qi 3721540 db BAA33531.1	(D83583) Sulfite Reductase [Nico...	31	2.4

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi#3152707

Evapξη ICQ Meeting... NCBI Blast... Τα έγγραφα... Microsoft Po... http://ww... 10:57 πμ

# Αποτελέσματα BLAST



Alignments

>[gi|114283|sp|P26476|ASRC\\_SALTY](#) ANAEROBIC SULFITE REDUCTASE SUBUNIT C  
[gi|196707|pir|IC38453](#) anaerobic sulfite reduction protein C - Salmonella typhimurium  
[gi|153884|gb|AAA99277.1](#) (M57706) anaerobic sulfite reductase [Salmonella typhimurium]  
Length = 337

Score = 123 bits (309), Expect = 3e-28  
Identities = 60/60 (100%), Positives = 60/60 (100%)

Query: 1 MSIDIDIKARAKNEYRLSKVRGEAMISVRIPGGILPAHLLTVARDIAETWGNQIHLTT 60  
MSIDIDIKARAKNEYRLSKVRGEAMISVRIPGGILPAHLLTVARDIAETWGNQIHLTT  
Sbjct: 1 MSIDIDIKARAKNEYRLSKVRGEAMISVRIPGGILPAHLLTVARDIAETWGNQIHLTT 60

>[gi|16761464|ref|NP\\_457081.1](#) (NC\_003198) anaerobic sulfite reductase subunit C [Salmonella enterica subsp. enterica serovar Typhi]  
[gi|16765870|ref|NP\\_461485.1](#) (NC\_003197) anaerobic sulfide reductase [Salmonella typhimurium LT2]  
[gi|16421096|gb|AAL21444.1](#) (AE008816) anaerobic sulfide reductase [Salmonella typhimurium LT2]  
[gi|16503764|emb|CAD02753.1](#) (AL627275) anaerobic sulfite reductase subunit C [Salmonella enterica subsp. enterica serovar Typhi]  
Length = 337

Score = 123 bits (309), Expect = 3e-28  
Identities = 60/60 (100%), Positives = 60/60 (100%)

Query: 1 MSIDIDIKARAKNEYRLSKVRGEAMISVRIPGGILPAHLLTVARDIAETWGNQIHLTT 60  
MSIDIDIKARAKNEYRLSKVRGEAMISVRIPGGILPAHLLTVARDIAETWGNQIHLTT  
Sbjct: 1 MSIDIDIKARAKNEYRLSKVRGEAMISVRIPGGILPAHLLTVARDIAETWGNQIHLTT 60

>[gi|18310518|ref|NP\\_562452.1](#) (NC\_003366) anaerobic sulfite reductase subunit C [Clostridium perfringens]  
[gi|18145198|dbj|BAB81242.1](#) (AP003190) anaerobic sulfite reductase subunit C [Clostridium perfringens]



# Εισαγωγή δεδομένων στο FASTA

**FASTA : Sequence database search (version 3) (W. Pearson)**

Reset Run  your e-mail  
(● = required, ● = conditionally required)

fasta: protein or DNA query vs similar db (FASTA release 3.0) Fasta program

● Query sequence File : please enter either :

1. the name of a file:  Αναζήτηση...

```
MS ID I D I I K A R A K N E Y R L S K V R G E A M I S V R I P G G I L P A H L L T V A R D I A E T
W G N G Q I H L T T
```

2. or the **actual data** here:

(sequence format)

● Is it a DNA or protein sequence (-n)?  DNA  protein

Protein Database

Nucleotid Database

Break long library sequences into blocks (-N)

[Selectivity options](#)

[Scoring options](#)



# Αποτελέσματα FASTA

```
http://bioweb.pasteur.fr/seqanal/tmp/fasta/A12451510167011/fasta.out - Microsoft Internet Explorer
Αρχείο Επεξεργασία Προβολή Αγαπημένα Εργαλεία Βοήθεια
Πίσω Εμπρός Διακοπή Ανανέωση Αρχική Αναζήτηση Αγαπημένα Μέσα Ιστορικό Αλληλογραφ... Μέγεθος Εκτύπωση
Διεύθυνση http://bioweb.pasteur.fr/seqanal/tmp/fasta/A12451510167011/fasta.out Μετάβαση Συνδέσεις
The best scores are:
sp|P26476|ASRC_SALTY Anaerobic sulfite reductase ( 337) 383 112 4.5e-25
sp|P72854|SIR_SYNY3 Sulfite reductase (Ferredoxin ( 635) 108 36 0.041
sp|P30008|SIR_SYNP7 Sulfite reductase (Ferredoxin ( 624) 98 34 0.27
sp|P02738|SAA_MACMU Amyloid protein A (Amyloid fi ( 76) 81 28 1.6
sp|P18575|SAA1_MUSVI Serum amyloid A-1 protein pr ( 129) 81 28 2.3
sp|P53613|SAA_MACEU Serum amyloid A protein precu ( 127) 80 28 2.8
sp|P04918|SAA3_MOUSE Serum amyloid A-3 protein pr ( 122) 79 28 3.3
sp|P96457|DNAJ_STRAL Chaperone protein dnaJ (Frag ( 82) 77 27 3.6
sp|Q51879|NIR_PHOLA Ferredoxin--nitrite reductase ( 510) 83 30 4.1
sp|P03096|COA2_POVMA Coat protein VP2 [Contains: ( 319) 78 28 7.7
sp|P12908|COA2_POVMC Coat protein VP2 [Contains: ( 319) 78 28 7.7
sp|P03097|COA2_POVM3 Coat protein VP2 [Contains: ( 319) 78 28 7.7
sp|P53614|SAA1_RABIT Serum amyloid A-1 protein pr ( 122) 74 26 8.4
sp|P35543|SAA3_RABIT Serum amyloid A-3 protein pr ( 122) 74 26 8.4

>>sp|P26476|ASRC_SALTY Anaerobic sulfite reductase subun (337 aa)
  initn: 383 initl: 383 opt: 383 Z-score: 572.5 bits: 111.8 E(): 4.5e-25
  Smith-Waterman score: 383; 100.000% identity (100.000% ungapped) in 60 aa overlap (1-60:1-60)

      10      20      30      40      50      60
  /
  MSIDIDIIKARAKNEYRLSKVRGEAMISVRIPGGILPAHLLTVARDIAETWGNGQIHLTT
  ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
sp|P26 MSIDIDIIKARAKNEYRLSKVRGEAMISVRIPGGILPAHLLTVARDIAETWGNGQIHLTT
      10      20      30      40      50      60

sp|P26 RQKLAMPGIRYEDIDNVNAALEPFLREIEIELCDVQVEDTKAGYLAIGGRNIVACQGNRI
      70      80      90      100     110     120

>>sp|P72854|SIR_SYNY3 Sulfite reductase (Ferredoxin) (EC (635 aa)
  initn: 112 initl: 112 opt: 108 Z-score: 160.5 bits: 36.5 E(): 0.041
  Smith-Waterman score: 108; 33.962% identity (37.500% ungapped) in 53 aa overlap (13-60:55-107)

      10      20      30
  /
  MSIDIDIIKARAKNEYRLSKVRGEA----MISVRIPGGILP
  ... : ..... : . : .....
sp|P72 LREPLATELLNDANYFTDDAVQILKFHGSYQQDNRDNRVKGQEKDYQFMLRTRNPGGLIP
      30      40      50      60      70      80

      40      50      60

Ολοκληρώθηκε
Internet
Εναρξη ICQ Meet... http://... Τα έγyp... Microsoft ... http://w... Magenta ... 11:25 πμ
```

# Ανακάλυψη προτύπων σε βιολογικές ακολουθίες

Κίνητρα:

- Πρόβλεψη λειτουργικής κατηγορίας γονιδίων.
- Συγκόλληση RNA.
- Δομή και λειτουργία πρωτεΐνης.
- Ρύθμιση γονιδίων (πρόβλεψη θέσης πρόσδεσης παράγοντα μεταγραφής).

## Οικογένειες πρωτεϊνών

- Πρόβλεψη λειτουργίας πρωτεΐνης με βάση την ακολουθία της.
- Διερεύνηση ομοιοτήτων σε βάσεις δεδομένων ακολουθιών.
- Σύγκριση με περιγραφές οικογενειών.
- Προγράμματα πρόβλεψης δομής.

# Ανάλυση οικογενειών πρωτεϊνών

- ✓ Συλλογή ακολουθιών στην οικογένεια.
- ✓ Ανάλυση.
  - Τοπικό (local) multiple alignment.
  - Συνολικό (global) multiple alignment.
  - Ανακάλυψη προτύπων.
- ✓ Περιγραφή οικογένειας.
- ✓ Επιλογή επιπλέον μελών για την οικογένεια.

# Πολλαπλή & συνδυαστική σύγκριση

Η πολλαπλή σύγκριση ακολουθιών:

- Είναι πιο ευαίσθητη.
- Δίνει περισσότερες πληροφορίες.
- Είναι σαφώς πιο δύσκολη σε σχέση με τη συνδυαστική.

# Πρότυπα και εναλλακτικές αναπαραστάσεις

- ✓ Πρότυπα
  - ενόητες προτύπων,
  - δέντρα απόφασης,
  - ακριβής / προσεγγιστική αντιστοίχιση.
- ✓ Alignments, weight matrices, profiles, HMMs, neural networks,...

## HMM profiles

Χρησιμοποιούν Hidden Markov Models και μπορούν να χαρακτηρίσουν μια ολόκληρη οικογένεια ακολουθιών.





# PROSITE

- Βάση δεδομένων οικογενειών πρωτεϊνών.
- Αποτελείται από σημαντικά βιολογικά πρότυπα που βοηθούν στο να αναγνωρισθεί σε ποια γνωστή οικογένεια πρωτεϊνών ανήκει (αν ανήκει) μια νέα ακολουθία.

## Πρότυπα PROSITE

- 'x' οποιοδήποτε αμινοξύ
- Ασάφειες:  
[ALT] = Ala ή Leu ή Thr  
{AM} = οποιοδήποτε αμινοξύ εκτός από Ala και Met
- '-' διαχωριστικό, '<' N-terminal, '>' C-terminal
- '.' τέλος προτύπου
- Επανάληψη: x(3) = x-x-x
- X(2,4) = x-x ή x-x-x ή x-x-x-x



# Παραδείγματα PROSITE

NiceSite View of PROSITE: PS00365 - Microsoft Internet Explorer

Address: <http://us.expasy.org/cgi-bin/prosite-search-ac?ps00365>

## NiceSite View of PROSITE: [PS00365](#)

General information about the entry	
Entry name	NIR_SIR
Accession number	PS00365
Entry type	PATTERN
Date	NOV-1990 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
PROSITE documentation	<a href="#">PDOC00314</a>

Name and characterization of the entry	
Description	Nitrite and sulfite reductases iron-sulfur/siroheme-binding site.
Pattern	[STV]-G-C-x(3)-C-x(6)-[DE]-[LIVMF]-[GAT]-[LIVMF]

Numerical results	
<ul style="list-style-type: none"><li>• SWISS-PROT release number: <b>40.7</b>, total number of sequence entries in that release: <b>103373</b>.</li><li>• Total number of hits in SWISS-PROT: <b>21 hits in 21 different sequences</b></li><li>• Number of hits on proteins that are known to belong to the set under consideration: <b>21 hits in 21 different sequences</b></li><li>• Number of hits on proteins that could potentially belong to the set under consideration: <b>0 hits in 0 different sequences</b></li><li>• Number of false hits (on unrelated proteins): <b>0 hits in 0 different sequences</b></li><li>• Number of known missed hits: <b>1</b></li><li>• Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: <b>1</b></li><li>• Precision (true hits / (true hits + false positives)): <b>100.00 %</b></li><li>• Recall (true hits / (true hits + false negatives)): <b>95.45 %</b></li></ul>	

Start | pserver | Microsoft PowerPoint - [Bl... | NiceSite View of PROSI... | Internet | 18:19



# Παραδείγματα PROSITE

NiceSite View of PROSITE: P500365 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://us.expasy.org/cgi-bin/prosite-search-ac?ps00365>

Links [Customize Links](#) [Free Hotmail](#) [Windows](#) [AltaVista - The Search Company](#)

### Comments

- Taxonomic range: **Archaeobacteria, Eukaryotes, Prokaryotes (Bacteria)**
- Maximum known number of repetitions of the pattern in a single protein: **1**
- 'Interesting' site in the pattern: **3,iron\_sulfur**
- 'Interesting' site in the pattern: **5,iron\_sulfur/siroheme**

### Cross-references

	<b>True positive hits:</b> ASRC_SALTY ( <a href="#">P26476</a> ), COBG_PSEDE ( <a href="#">P21637</a> ), CYSI_BUCAI ( <a href="#">P57502</a> ), CYSI_ECOLI ( <a href="#">P17846</a> ), CYSI_SALTY ( <a href="#">P17845</a> ), CYSI_THIRO ( <a href="#">P52673</a> ), NIR_BETVE ( <a href="#">P38500</a> ), NIR_EMENI ( <a href="#">P22944</a> ), NIR_MAIZE ( <a href="#">P17847</a> ), NIR_NEUCR ( <a href="#">P38681</a> ), NIR_PHOLA ( <a href="#">Q51879</a> ), NIR_SPIOL ( <a href="#">P05314</a> ), NIR_SYNP7 ( <a href="#">P39661</a> ), NIRB_ECOLI ( <a href="#">P08201</a> ), NIRB_KLEPN ( <a href="#">Q06458</a> ), SIR_DESVH ( <a href="#">Q05805</a> ), SIR_SYNP7 ( <a href="#">P30008</a> ), SIR_SYNY3 ( <a href="#">P72854</a> ), Y551_METJA ( <a href="#">Q57971</a> ), Y870_METJA ( <a href="#">Q58280</a> ), YJ9F_YEAST ( <a href="#">P47169</a> )
SWISS-PROT	<b>False negative hits (sequences which belong to the set under consideration, but which have not been picked up by the pattern or profile):</b>  NASD_BACSU ( <a href="#">P42435</a> )  <b>'Potential' hits (partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences):</b>  NIR_LEPMC ( <a href="#">P43504</a> )
PDB	<a href="#">1AOP</a> ; <a href="#">2AOP</a> ; <a href="#">2GEP</a> ; <a href="#">3AOP</a> ; <a href="#">3GEO</a> ; <a href="#">4AOP</a> ; <a href="#">4GEP</a> ; <a href="#">5AOP</a> ; <a href="#">5GEP</a> ; <a href="#">6GEP</a> ; <a href="#">7GEP</a> ; <a href="#">8GEP</a> ;

<http://us.expasy.org/cgi-bin/niceprot.pl?Q58280>

Start | pservers | Microsoft PowerPoint - [Bl... | NiceSite View of PROSI... | Internet | 18:21



# Παραδείγματα PROSITE

NiceProt View of SWISS-PROT: P26476 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print Copy Paste

Address <http://us.expasy.org/cgi-bin/niceprot.pl?P26476> Go

Links Customize Links Free Hotmail Windows AltaVista - The Search Company


### Sequence information


Length: **337 AA**      Molecular weight: **37290 Da**      CRC64: **183FE2F669B34CBF** [This is a checksum on the sequence]

10	20	30	40	50	60
MSIDIDIIKA	RAKNEYRLSK	VRGEAMISVR	IPGGILPAHL	LTVARDIAET	WGNGQIHLTT
70	80	90	100	110	120
RQKLAMPGIR	YEDIDNVNAA	LEPFLREIEI	ELCDVQVEDT	KAGYLAIGGR	NIVACQGNRI
130	140	150	160	170	180
CQKANTDTTG	LSRRLEKLVY	PSPYHLKTVI	VGCPNDCAKA	SMADLGIIGV	AKMRFTADRC
190	200	210	220	230	240
IGCGACVKAC	SHHAVGCLAL	KNGKAVKEES	ACIGCGECVL	ACPTLAWQRK	PDQLWQVRLG
250	260	270	280	290	300
GRTSKKTPRV	GKFLFNWVTE	DVIKQVIVNL	YEFEKEMLGG	KPIYLMHGH	IDKGGYLRPK
310	320	330			
ERVLRGVQLN	PEAMVAERII	WAEDESVARM	HLKPAGH		

P26476 in [FASTA format](#)

[View entry in original SWISS-PROT format](#)  
[View entry in raw text format \(no links\)](#)  
[Report form for errors/updates in this SWISS-PROT entry](#)

 Direct BLAST submission at [EMBLnet-CH/SIB](#) (Switzerland)

 Direct BLAST submission at [NCBI \(Bethesda, USA\)](#)

Start | pserver | Microsoft PowerPoint - [Bl... | NiceProt View of SWIS... | Internet | 18:23



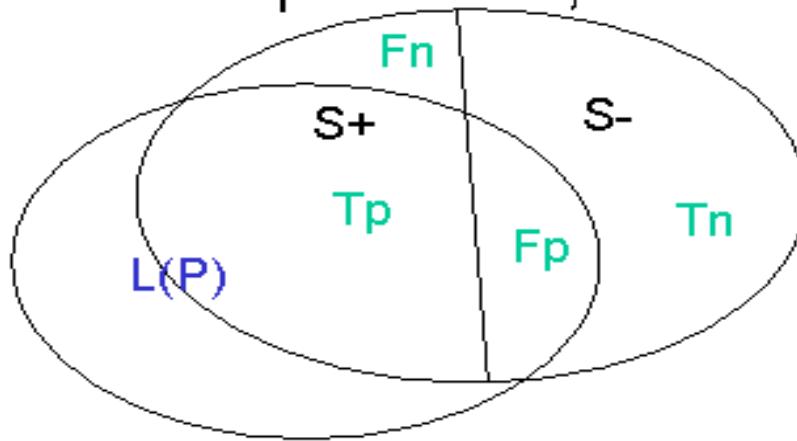
# Πρόβλημα μάθησης

- Αυτόματη εύρεση του προτύπου δεδομένων κάποιων ακολουθιών.
- Χαρακτηρισμός προτύπων που περιγράφουν ιδιότητες που καθορίζουν μια οικογένεια.
- Κατάταξη ακολουθιών ανάλογα με την ύπαρξη ή μη του κατάλληλου προτύπου στην κατάλληλη οικογένεια και περιορισμός των σφαλμάτων.



# Υπαρξη σφαλμάτων

True positives, true negatives,  
false positives, false negatives



Tp - true pos  
Tn - true neg  
Fp - false pos  
Fn - false neg

$$Tp = L(P) \cap S+$$

$$Tn = \neg L(P) \cap S-$$

$$Fp = L(P) \cap S-$$

$$Fn = \neg L(P) \cap S+$$

$L(P)$  - the set of sequences matched by the pattern  $P$



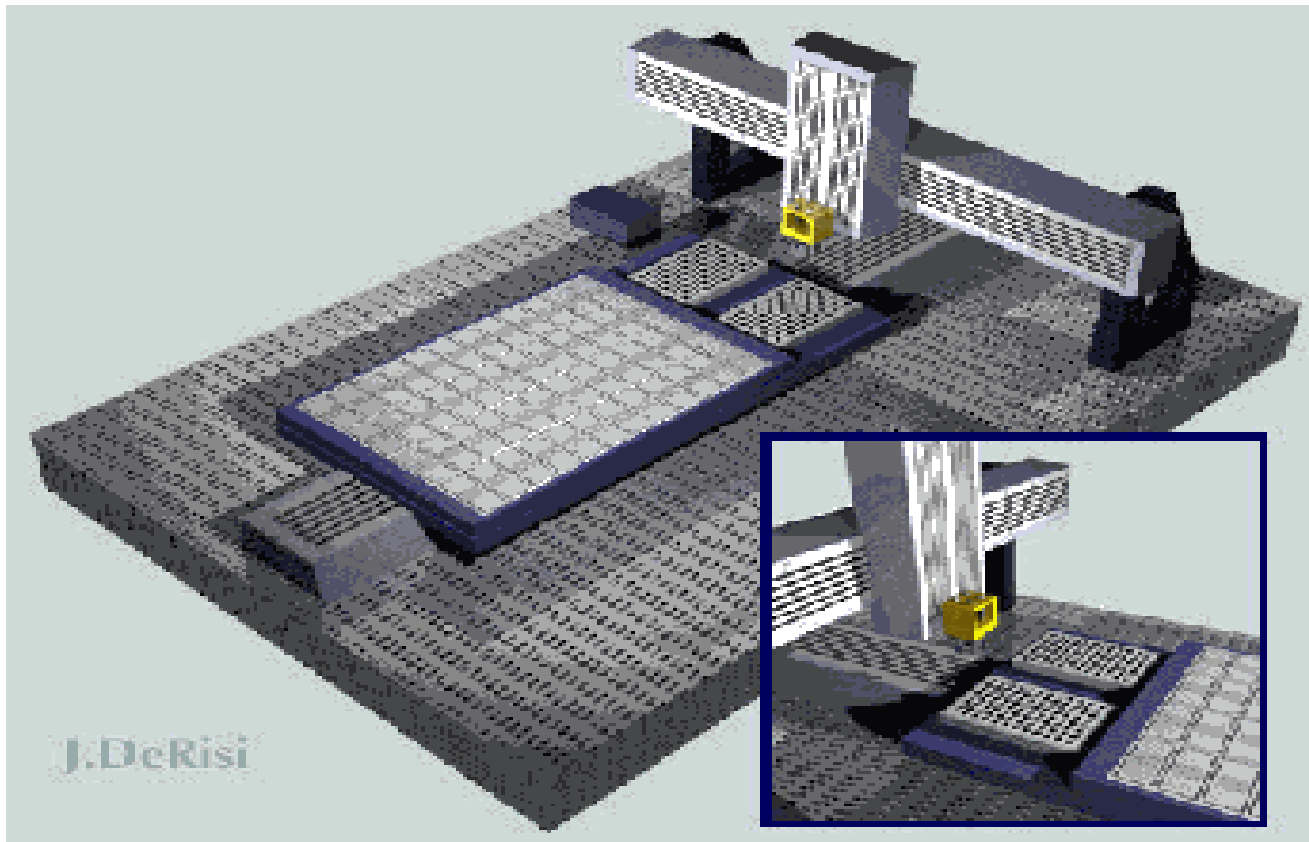
# Microarray technology

- Η τεχνολογία micro array έχει σκοπό να επιταχύνει τη γενετική ανάλυση με τον ίδιο τρόπο που οι μικροεπεξεργαστές επιτάχυναν τους υπολογισμούς.
- Πρόκειται για μικροσκοπικούς πίνακες (arrays) από τμήματα του γονιδιώματος προσδεδεμένα σε γυάλινα chips.

Με κατάλληλο τρόπο μπορεί να διαβαστεί το περιεχόμενο των ακολουθιών DNA και μετά να υποστεί επεξεργασία με εργαλεία της υπολογιστικής βιολογίας.

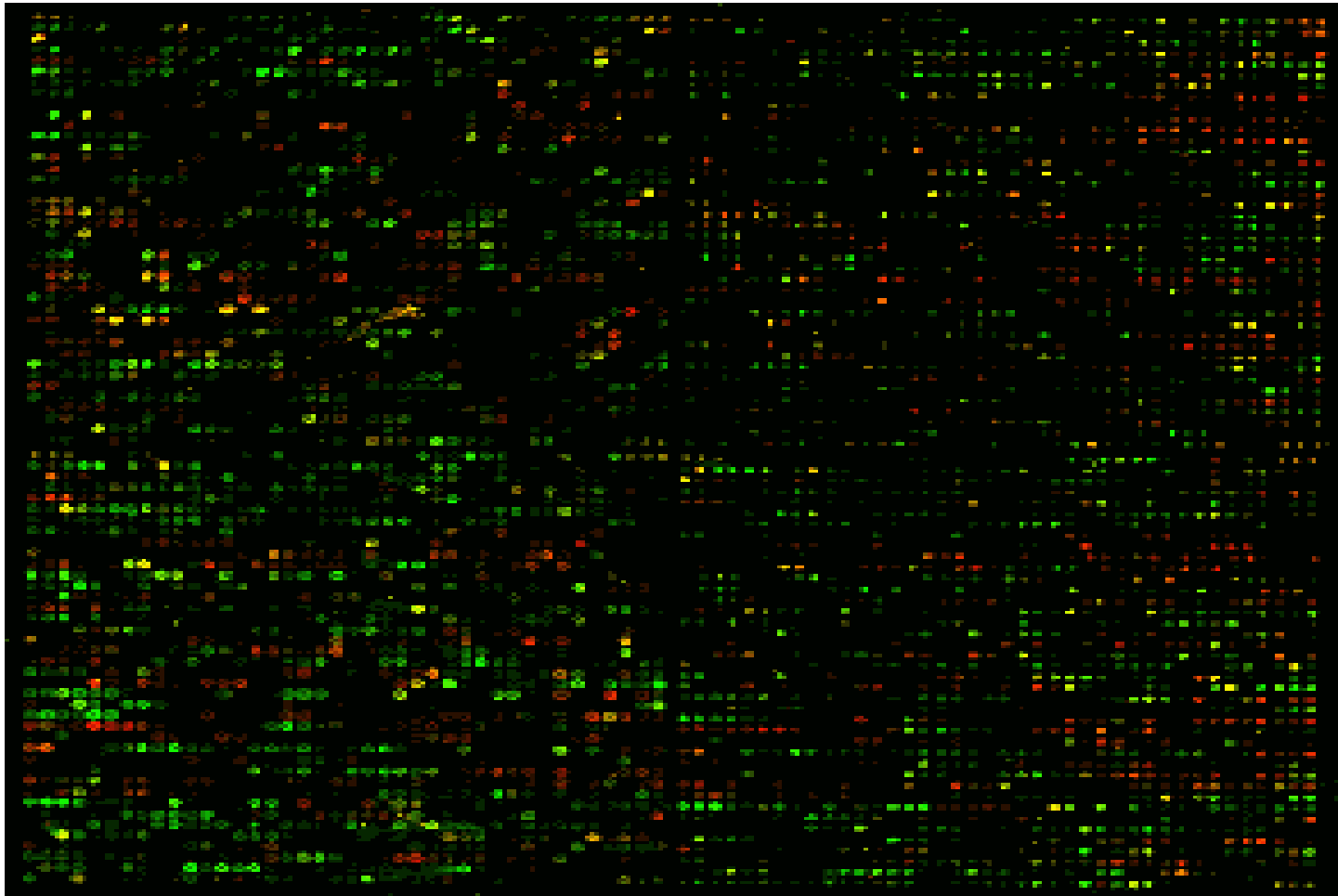


# Array technology





# Yeast array



## Επιπλέον ζητήματα για τα δεδομένα

- Σχεδιασμός βάσεων δεδομένων για βιολογικές πηγές.
- Αναπαράσταση και οπτικοποίηση της βιολογικής γνώσης.
- Εφαρμογή μεθόδων ανάλυσης δεδομένων.

## Απαιτούμενη εμπειρία

- Συνδυασμένη προσπάθεια από ερευνητές και των δύο πεδίων.
- Χρήση κοινής γλώσσας.
- Μάθηση σχετικά με ζητήματα της άλλης πλευράς.
- Εξειδικευμένη γνώση σε μαθηματικά και στατιστική.
- Αποτελεσματικός σχεδιασμός αλγορίθμων.



# Στάδια ερευνητικής διαδικασίας

- Συνεργασία με μοριακούς βιολόγους.
- Διερεύνηση ως προς το τι έχει γίνει.
- Επικύρωση αποτελεσμάτων.
- Επίτευξη ταχύτητας και ακρίβειας.