Βιοπληροφορική

16

MED LAB
MEDICAL TECHNOLOGY & INTELLIGENT INFORMATION SYSTEMS

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

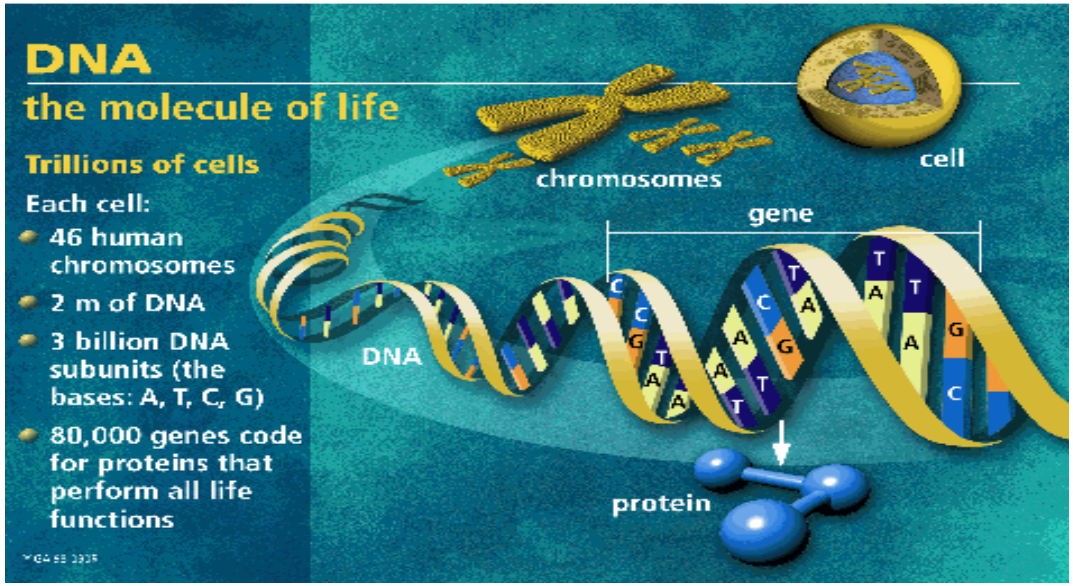
- Όλα τα κληρονομικά χαρακτηριστικά ενός οργανισμού βρίσκονται στο γενετικό υλικό του μέσα σε κάθε κύτταρό του.
- Εκεί αναπαρίστανται σε έναν απλό κώδικα τεσσάρων στοιχείων (γραμμάτων) που ονομάζεται γενετικός κώδικας.

Κύτταρο και βασικά είδη ζωής

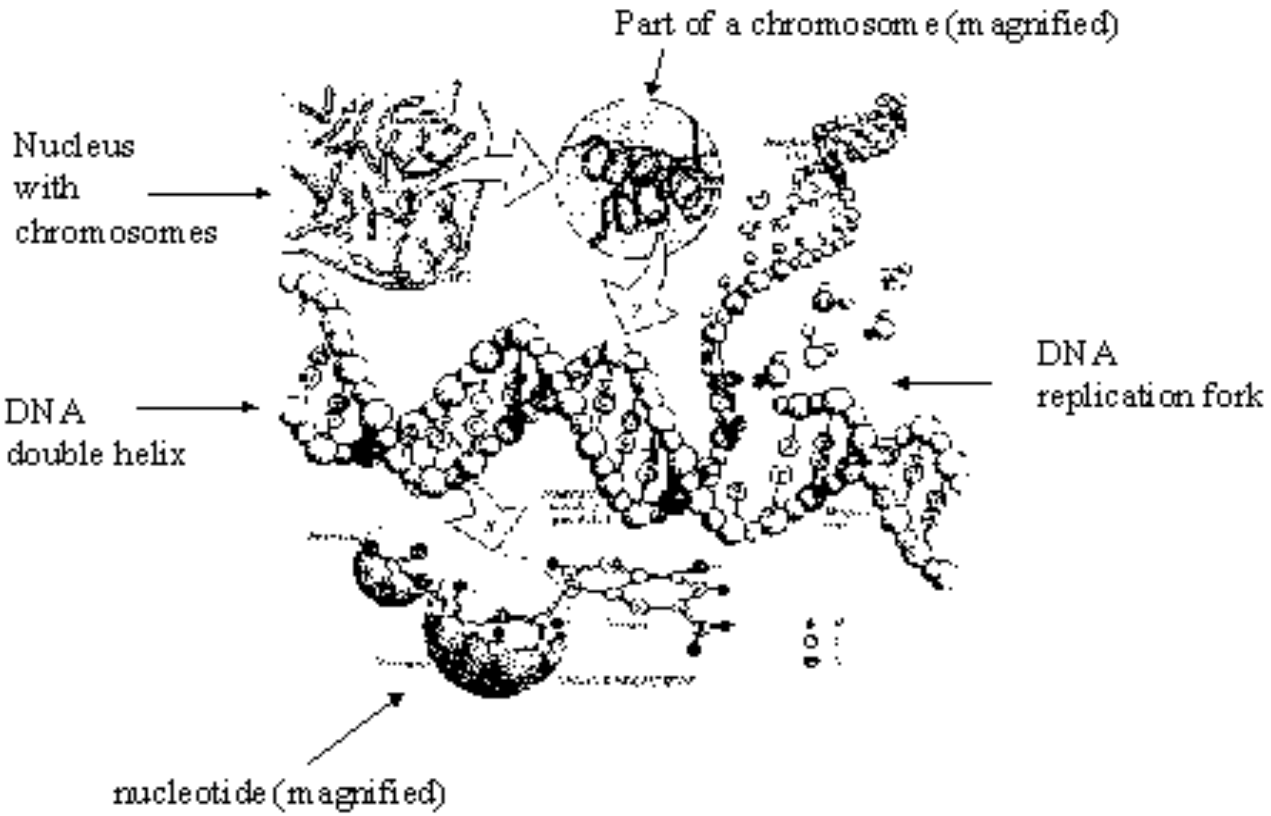
- Το κύτταρο είναι η βασική μορφή ζωής.
- Όλοι οι οργανισμοί αποτελούνται από κύτταρα.
- Ανάλογα με τη δομή των κυττάρων τους, οι βασικές μορφές ζωής κατατάσσονται στις παρακάτω κατηγορίες:
  - Ιοί.
  - Αρχαία.
  - Βακτήρια.
  - Ευκάρυα (τα φυτά, τα ζώα και οι άνθρωποι είναι ευκάρυα).

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

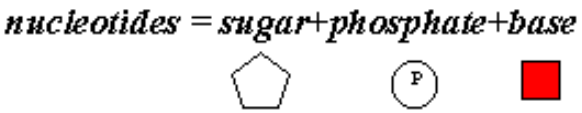
- Το DNA αποτελείται από χρωμοσώματα.
- Οι προκαρυώτες, μονοκύτταροι οργανισμοί χωρίς πυρήνα, έχουν μόνο ένα κυκλικό χρωμόσωμα.
- Οι ευκαρυώτες, οργανισμοί με πυρήνα, έχουν αριθμό χρωμοσωμάτων ανάλογα με το είδος.

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 κυτταρα-έχουν μισό γονιδίωμα) και τα ερυθρά αιμοσφαίρια.

DNA και RNA

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



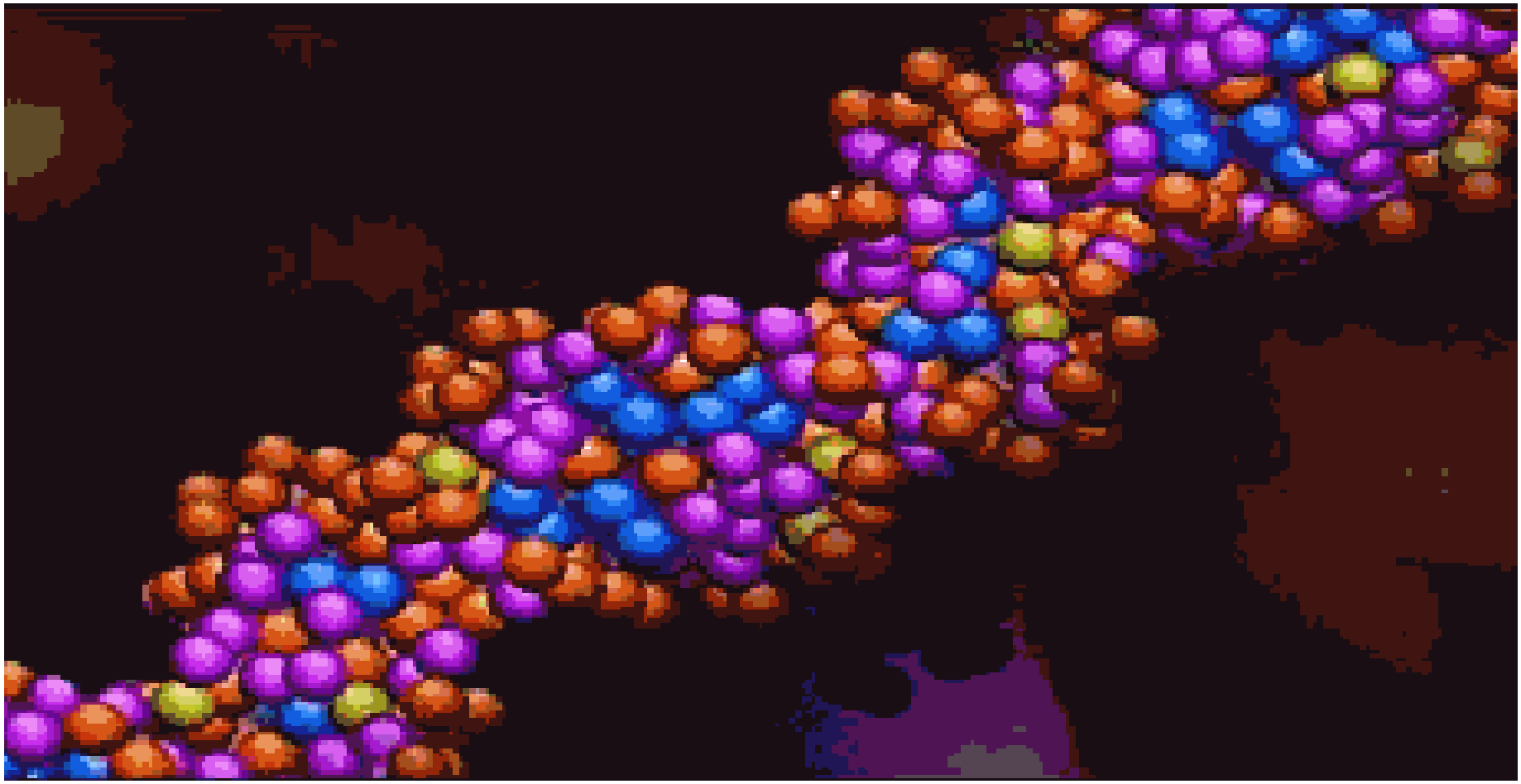
Bases: adenine (A), cytosine (C), guanine (G), thymine (T), uracil (U)

4 letter alphabet:

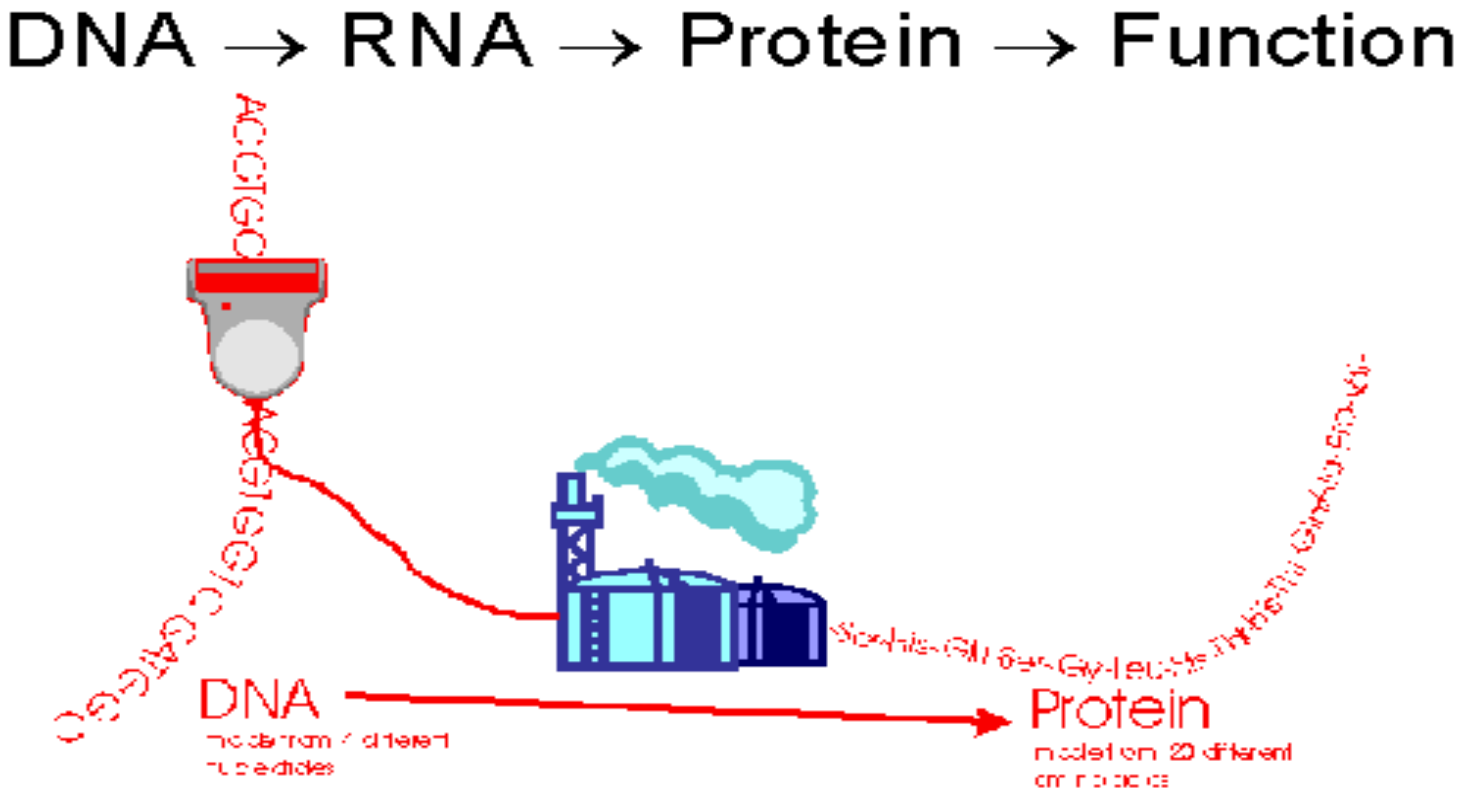
DNA = A C G T (adenine cytosine guanine thymine)

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

Διπλή έλικα DNA

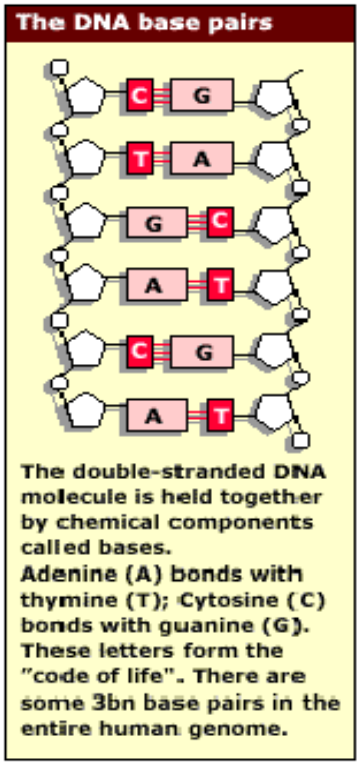


Ροή πληροφορίας στη μοριακή βιολογία



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

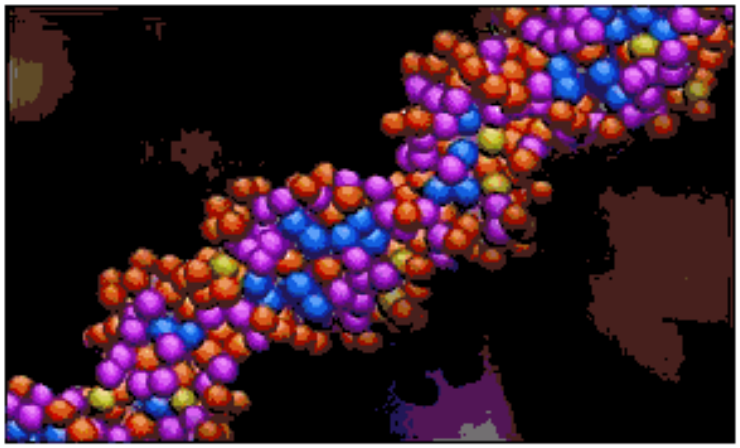
- Το DNA διαφέρει ανάμεσα σε ανθρώπους κατά 0.2% (1 βάση σε κάθε 500).
- Το ανθρώπινο DNA είναι κατά 98% πανομοιότυπο με αυτό των χιμπατζήδων.
- Μεγάλο μέρος είναι μη κωδικοποιητικό DNA
- Υπάρχουν 10^{14} κύτταρα στο σώμα.
- 12000 γράμματα του DNA αποκωδικοποιούσε κάθε δευτερόλεπτο το Human Genome Project.

Human Genome Project

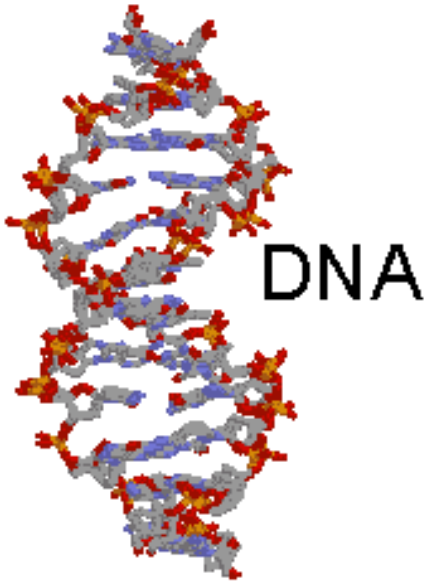
- Παγκόσμια προσπάθεια χαρτογράφησης του ανθρώπινου DNA.
- Το project άρχισε το 1990 και ολοκληρώθηκε το 2003.
- Στόχος να αναγνωριστούν και τα 20000 περίπου γονίδια του ανθρώπινου οργανισμού και να διαβαστούν όλες οι ακολουθίες DNA.
- Στόχος να τοποθετηθεί όλη η πληροφορία σε βάσεις δεδομένων και να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

Ακολουθία βάσεων DNA

AAAAGAAAAGGTTAGAAAGATGAGAGATGATAAAGGGTCCATTTGAGGTTAGGTAATATGGTTTGGTATCCCTGTAGTTAAAAG
TTTTTGTCTTATTTTAGAATACTGTGACTATTTCTTTAGTATTAATTTTTCCTTCTGTTTTCCTCATCTAGGGAACCCCAAGAG
CATCCAATAGAAGCTGTGCAATTATGTAAAATTTTCAACTGTCTTCCTCAAAATAAAGAAGTATGGTAATCTTTACCTGTATAC
AGTGCAGAGCCTTCTCAGAAGCACAGAATATTTTTATATTTCCTTTATGTGAATTTTTAAGCTGCAAATCTGATGGCCTTAATT
TCCTTTTTGACACTGAAAGTTTTGTAAAAGAAATCATGTCCATACACTTTGTTGCAAGATGTGAATTATTGACACTGAACTTAA
TAACTGTGTACTGTTCGGAAGGGGTTCCTCAAATTTTTTGACTTTTTTTGTATGTGTGTTTTTTCTTTTTTTTAAGTTCTTAT
GAGGAGGGAGGGTAAATAAACCACTGTGCGTCTTGGTGTAATTTGAAGATTGCCCCATCTAGACTAGCAATCTCTTCATTATTC
TCTGCTATATATAAAACGGTGCTGTGAGGGAGGGGAAAAGCATTTTTCAATATATTGAACTTTTGTACTGAATTTTTTTGTAAT
AAGCAATCAAGGTTATAATTTTTTTTAAAATAGAAATTTTGTAAGAAGGCAATATTAACCTAATCACCATGTAAGCACTCTGGA
TGATGGATTCCACAAAACTTGGTTTTATGGTTACTTCTTCTCTTAGATTCTTAATTCATGAGGAGGGTGGGGGAGGGAGGTGGA
GGGAGGGAAGGGTTTCTCTATTAAAATGCATTCGTTGTGTTTTTTAAGATAGTGTAACTTGCTAAATTTCTTATGTGACATTAA
CAAATAAAAAAGCTCTTTTAATATTAGATAA



David Gilbert, 2002



Σύγκριση μεγεθών ακολουθιών

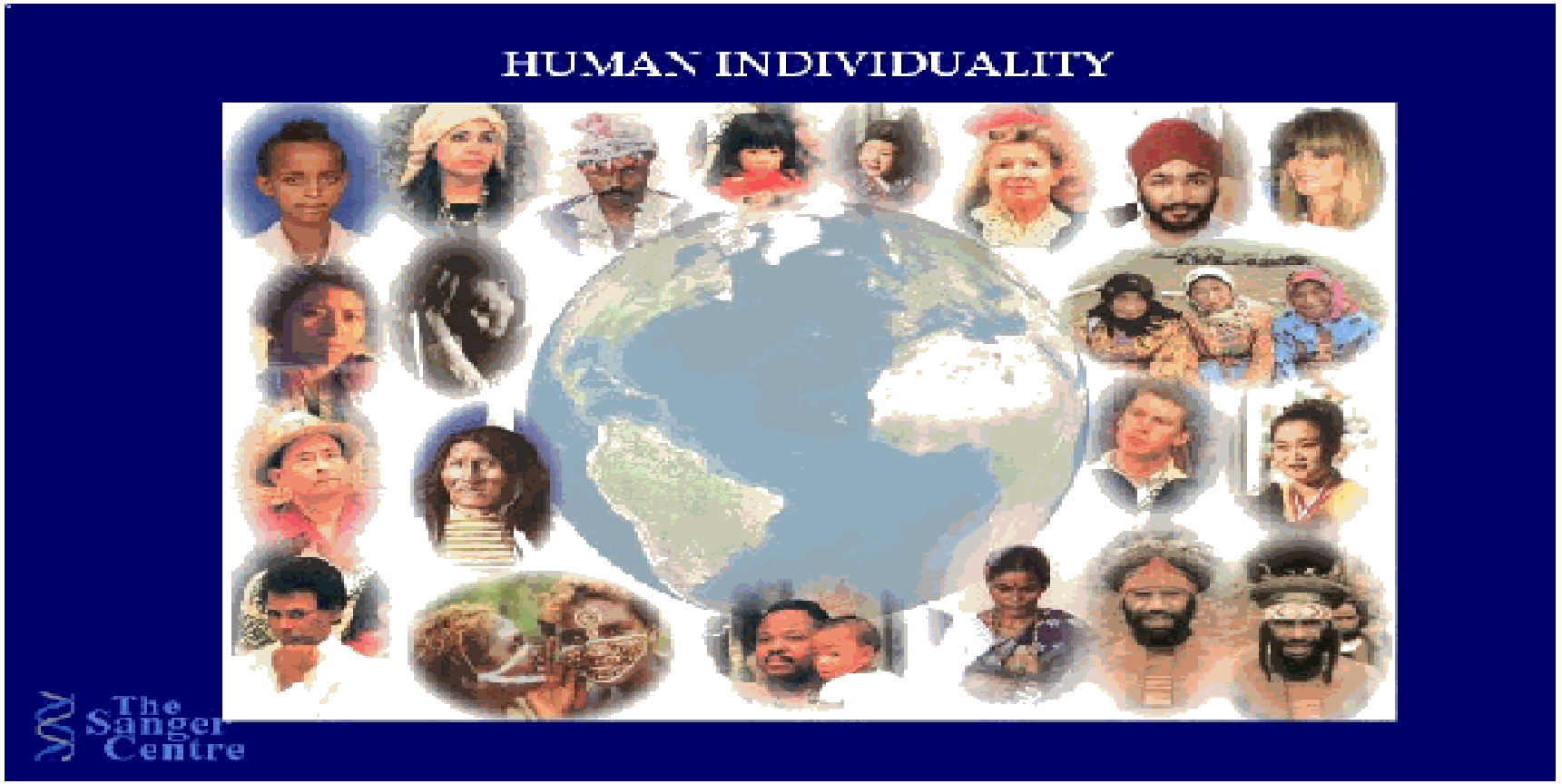
Comparative Sequence Sizes	(Bases)
(yeast chromosome 3)	350 Thousand
Escherichia coli (bacterium) genome	4.6 Million
Largest yeast chromosome now mapped	5.8 Million
Entire yeast genome	15 Million
Smallest human chromosome (Y)	50 Million
Largest human chromosome (1)	250 Million
Entire human genome	3 Billion

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.

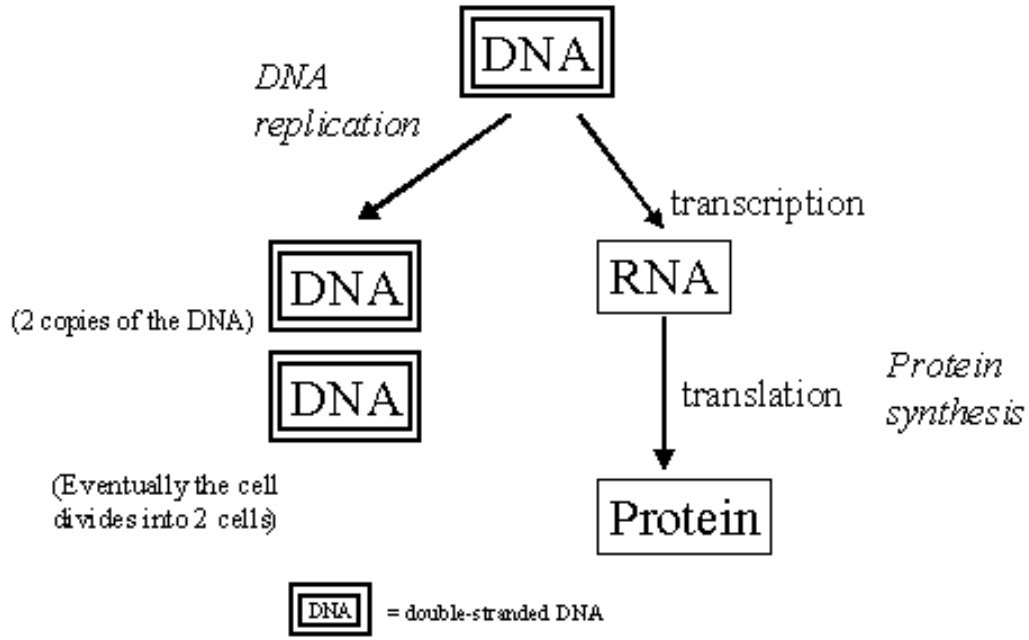
TTTGCTCCGTTTTCA
TTTGCTCYGTTTTCA
TTTGCTCTGTTTTCA

Ανθρώπινη ποικιλομορφία

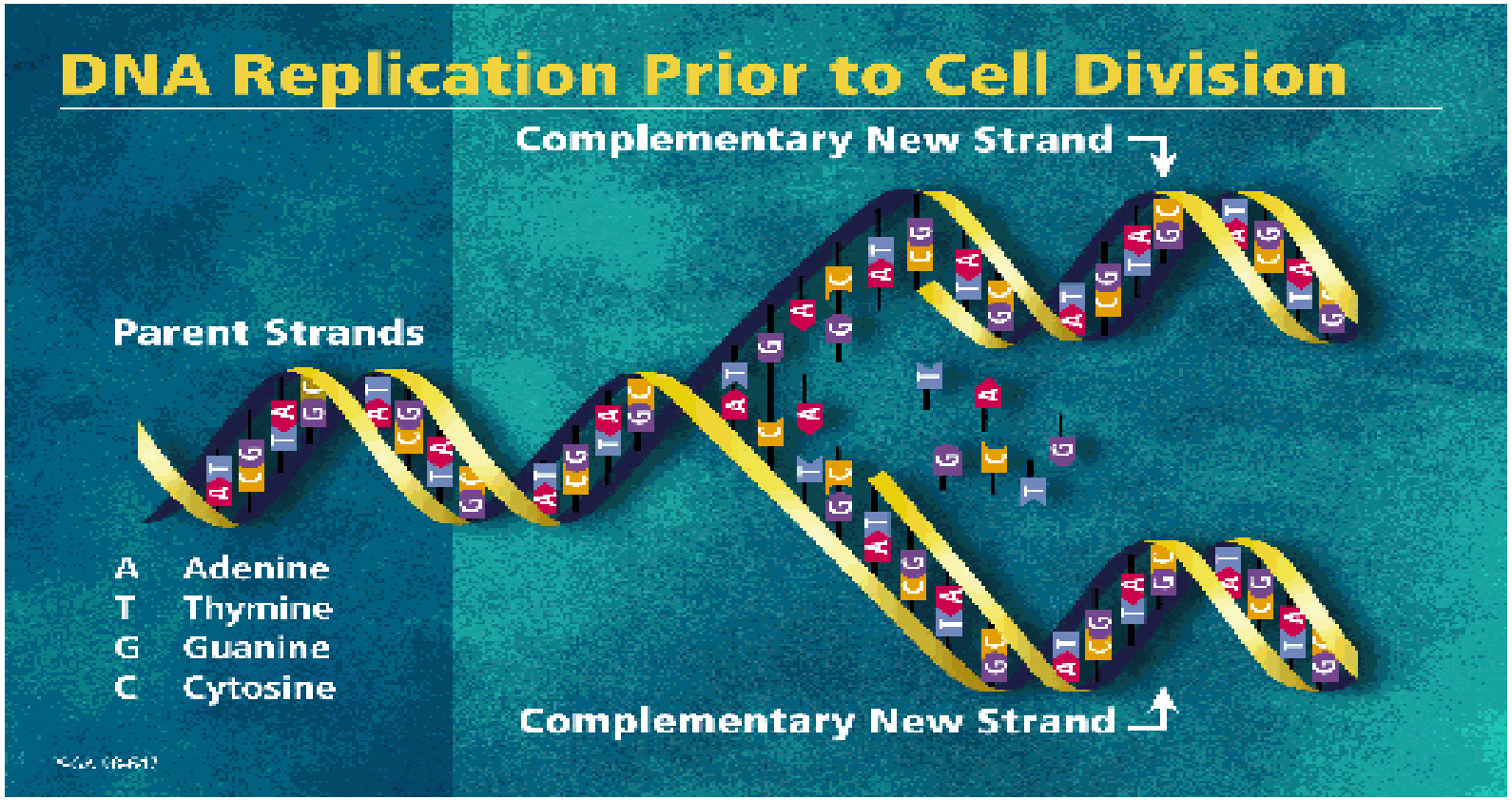


Τι μπορεί να συμβεί στο DNA

- 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 ILETNDGGLN TVKSHFHDQI NCNTIWHNVY
DGRAVRNASY DDHAAGVPFE ENHKKFKEAA SSGDNSRAVT WSGTAVGLIK DQRPAGDIVR
ELREEAKERI KKIQAFAA
```

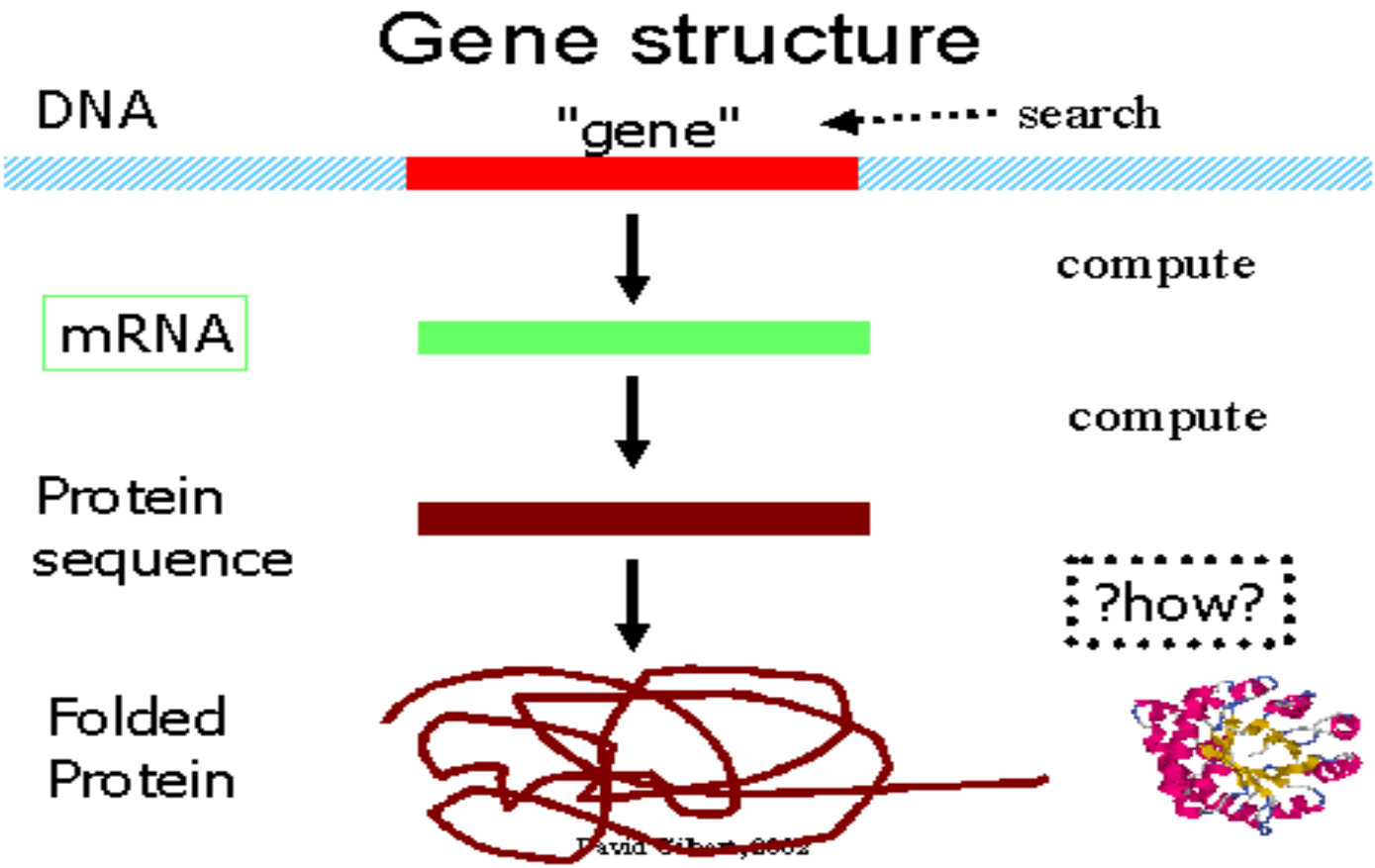
Γενετικός κώδικας

The genetic code

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

David Gilbert, 2002

Από το γονίδιο στην πρωτεΐνη

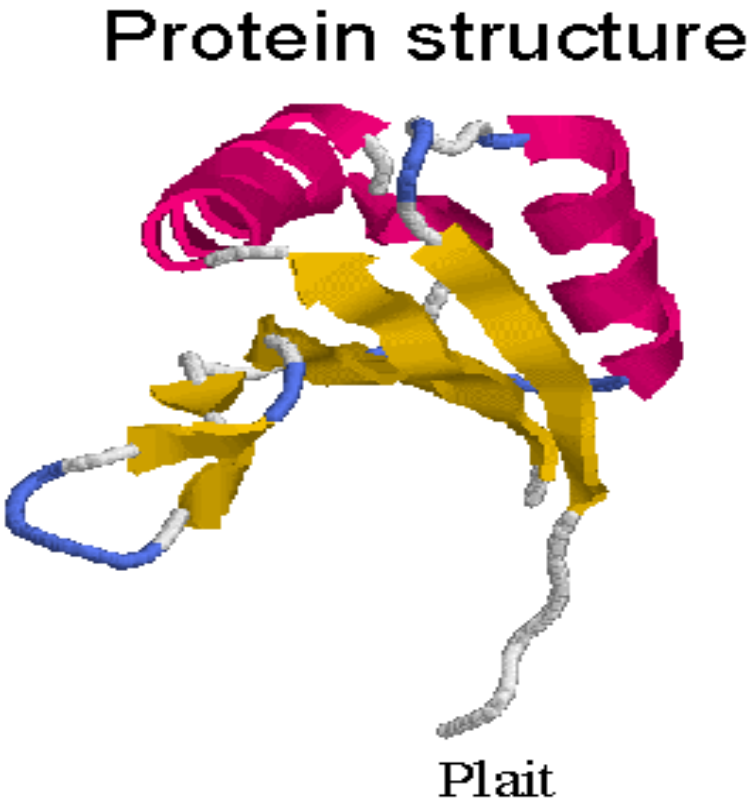


Δομή πρωτεΐνης

Protein structure

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

Δομή πρωτεΐνης



Επίπεδα δομής πρωτεϊνών

Πρωτογενής δομή (γραμμική ακολουθία αμινοξέων)

Δευτερογενής δομή

Τριτογενής δομή

Τεταρτογενής δομή

→ δομές αμινοξέων στο χώρο ολοένα και πιο σύνθετες

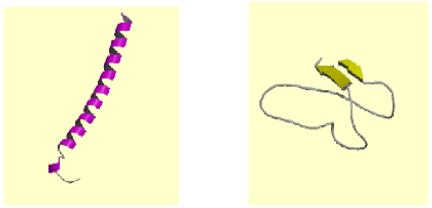
Primary Structure

VHLTPEEKSAVTALWGKVNVDE
VGGEALGRLLVVYPWTQRFFES
FGDLSTPDAVMGNPKVKAHGKK
VLGAFSDGLAHLDNLKGTFATLS
ELHCDKLHVDPENFRLLGNVLV
CVLAHHFGKEFTPPVQAAYQKV
VAGVANALAHKYH

amino acids (1hbb - Haemoglobin B)

Aik Choon Tan & David Gilbert, 2001

Secondary Structure

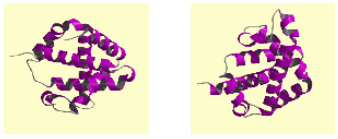


alpha helix

beta sheets

Aik Choon Tan & David Gilbert, 2001

Tertiary Structure

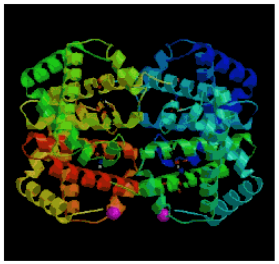


1hbbA0 1hbbB0

1hbbC0 1hbbD0

Aik Choon Tan & David Gilbert, 2001 12

Quaternary Structure



Hemoglobin (1hbb)

Aik 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

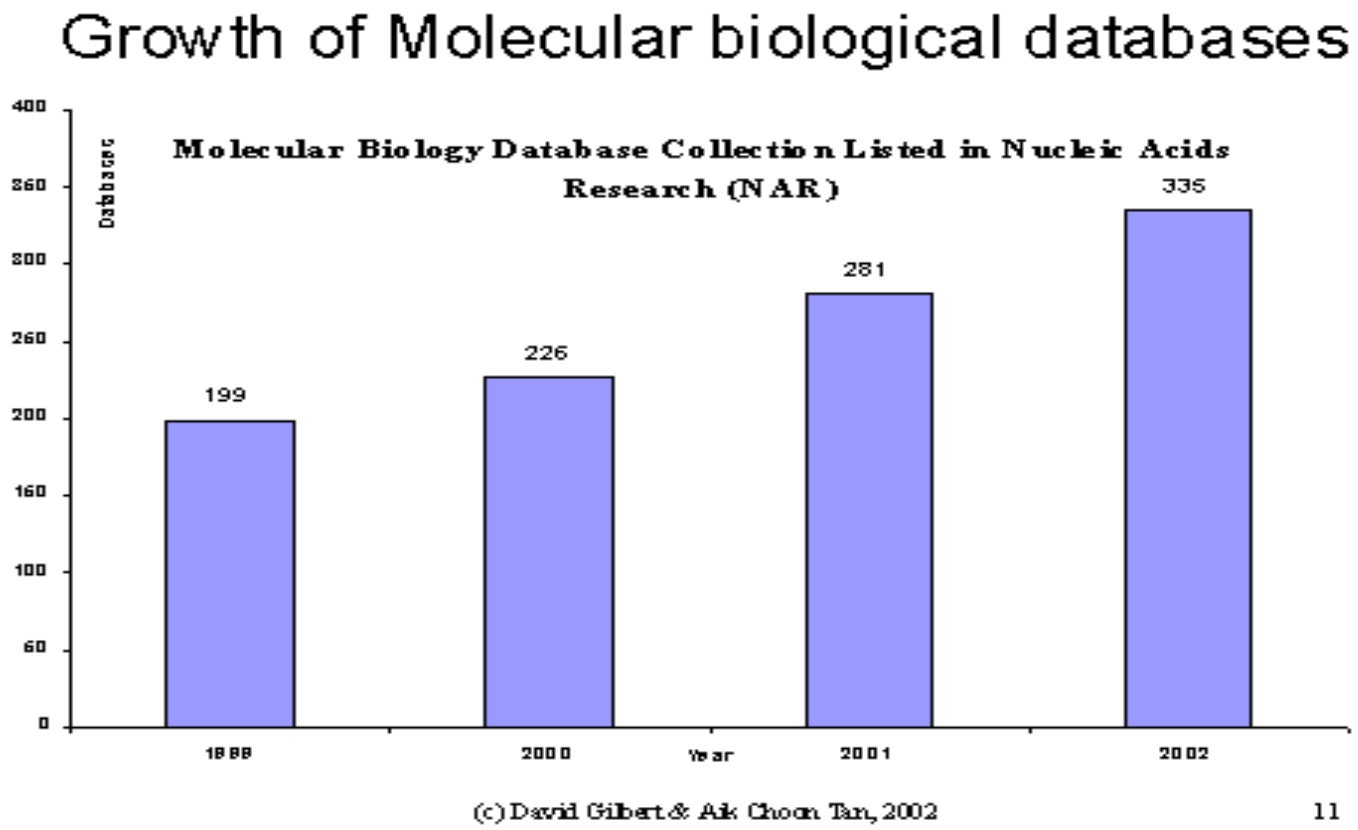
Ένας σημαντικός στόχος

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).

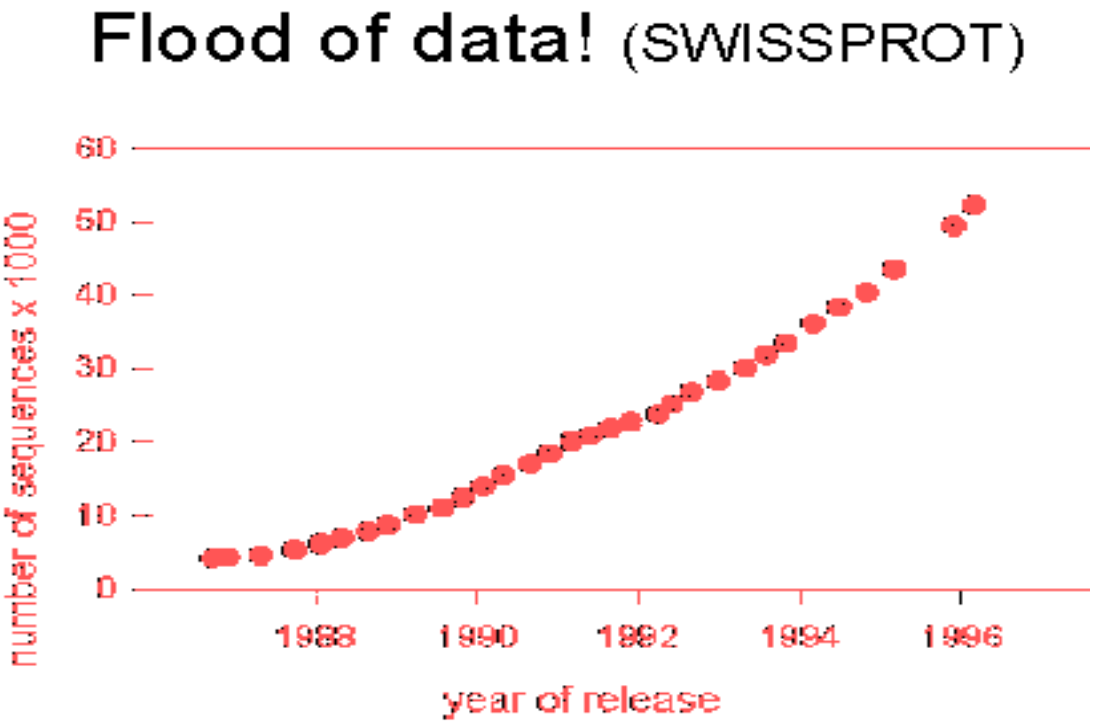
Ραγδαία αύξηση των βάσεων δεδομένων



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

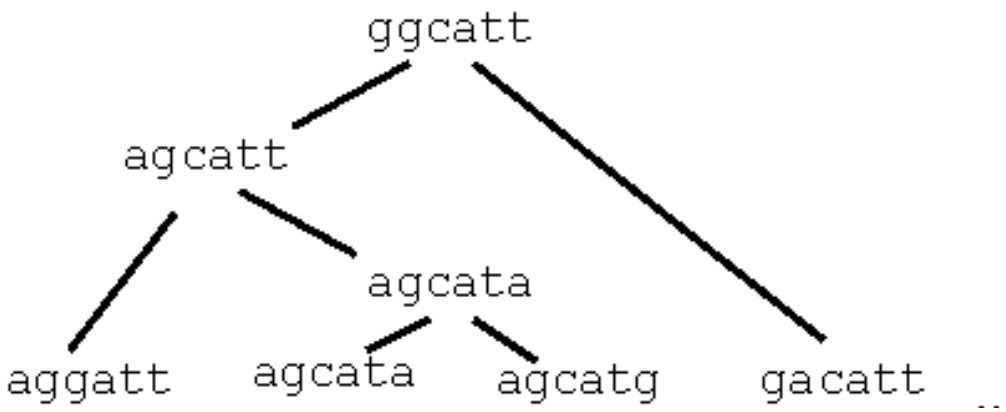


Ανάλυση του όγκου δεδομένων

- 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 → Δομή → Λειτουργία

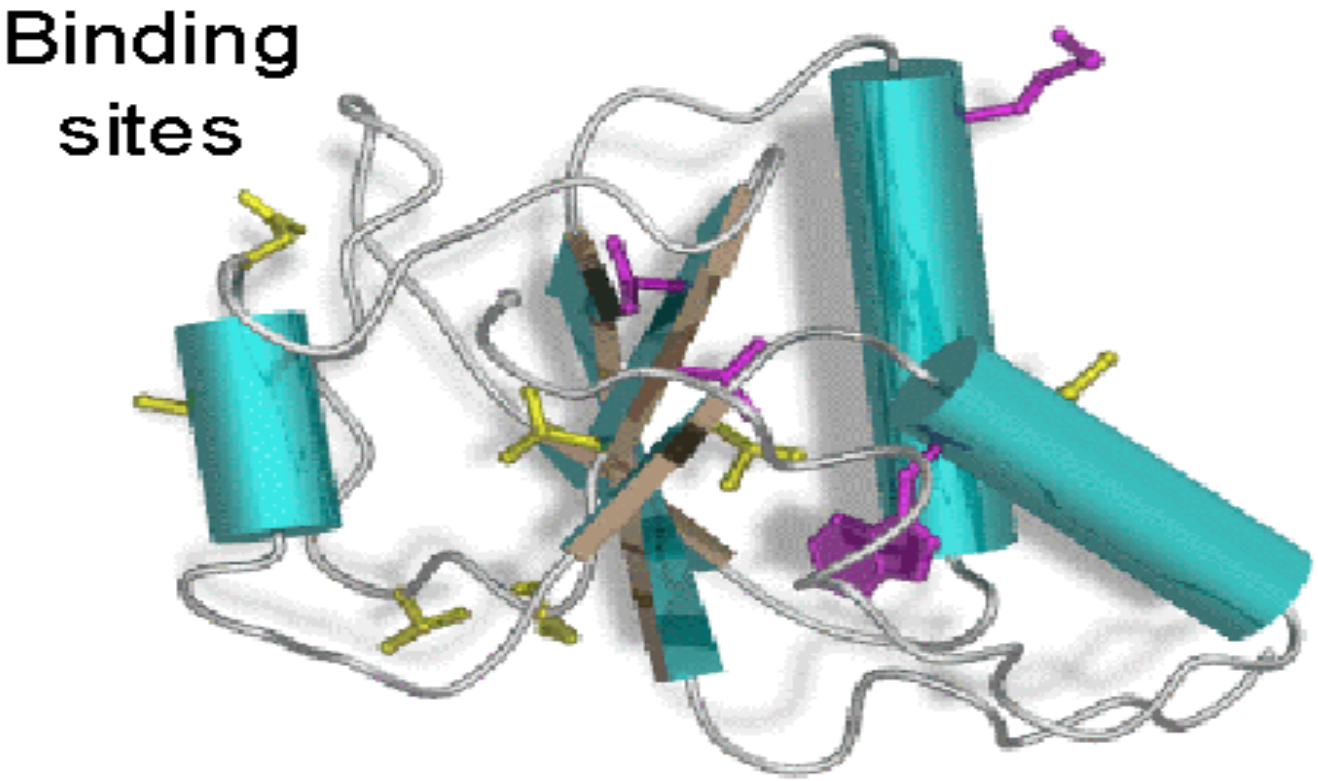
Προσεγγίσεις:

- 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



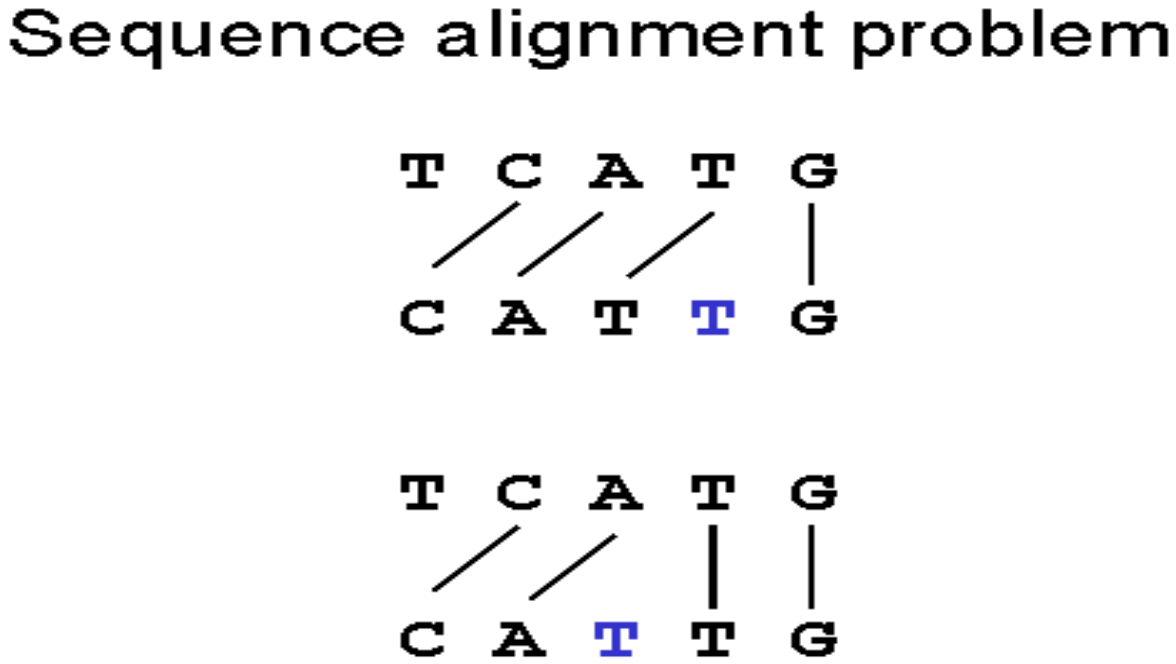
Άλλα προβλήματα

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

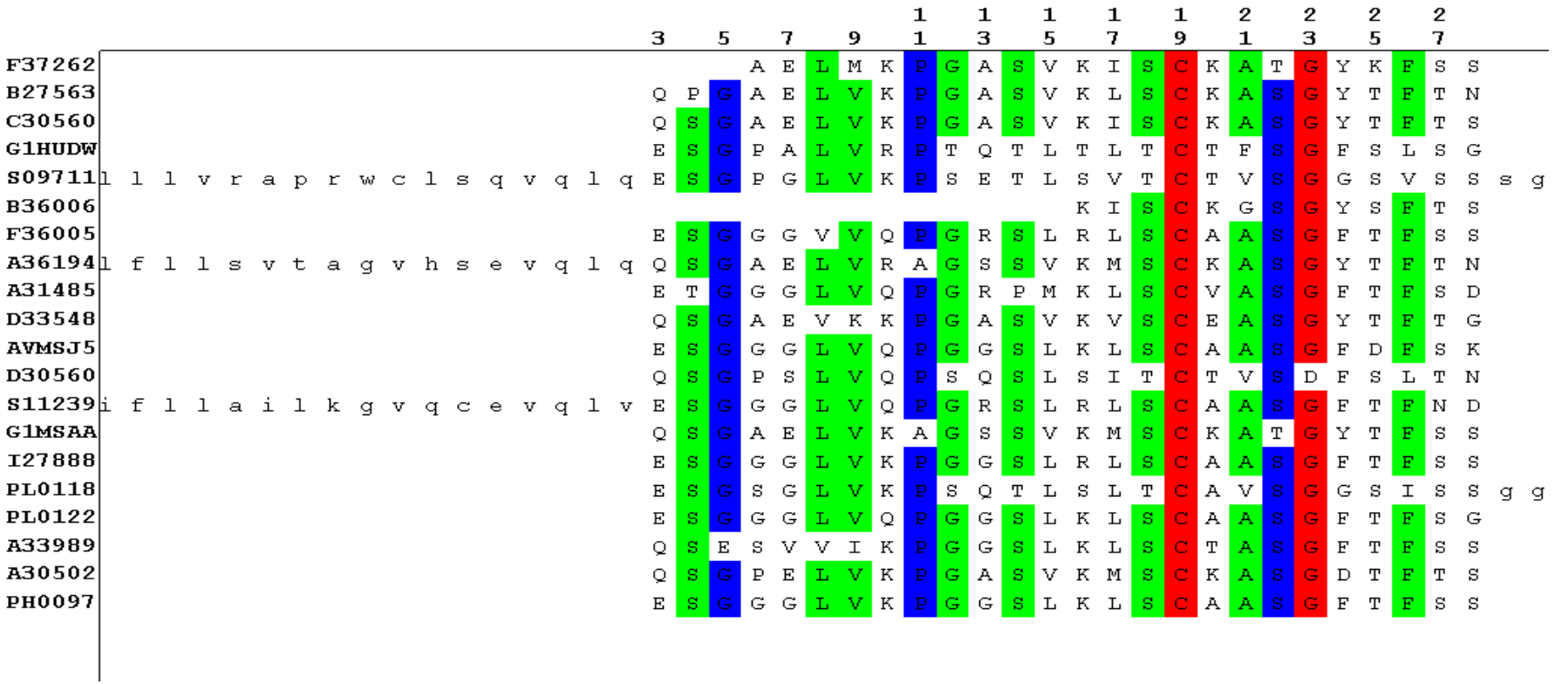
Έρευνα και ανακάλυψη προτύπων

- 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)



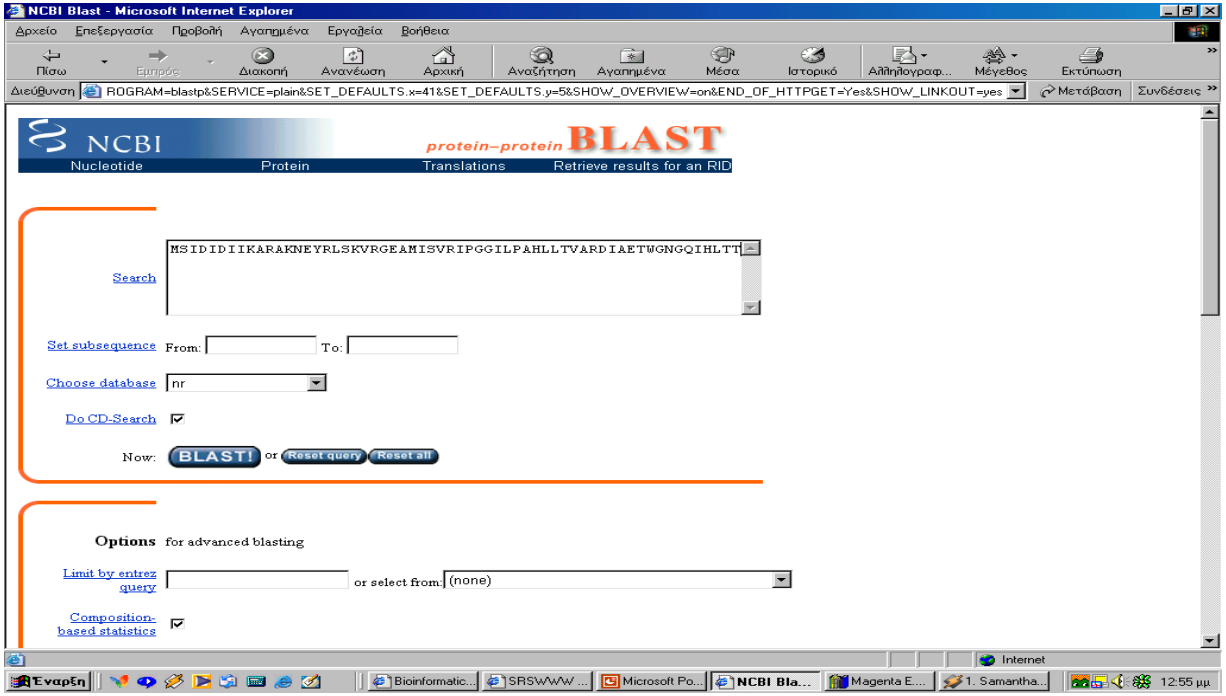
Multiple alignment



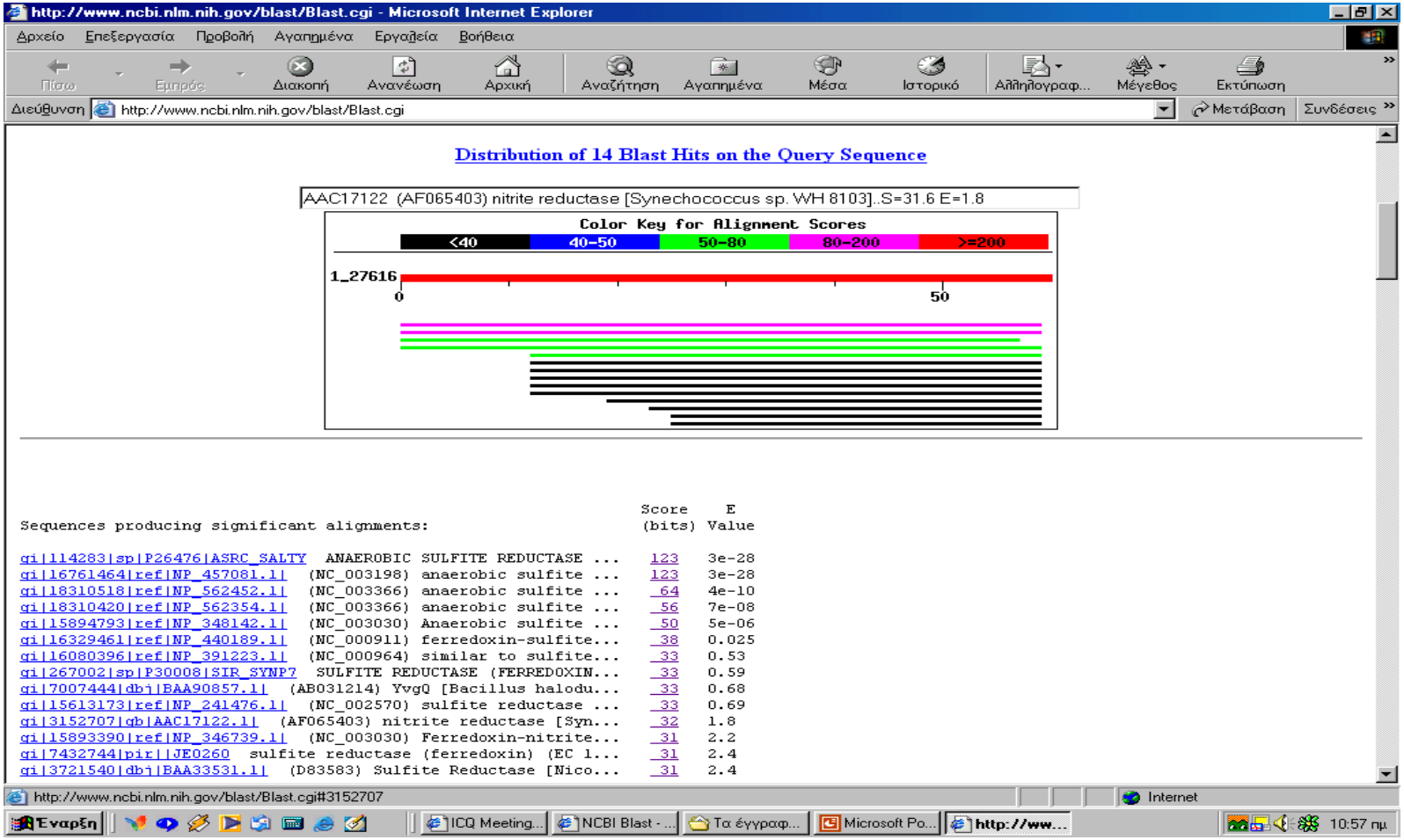
Προγράμματα για συνδυαστική διερεύνηση (pair wise searching)

- 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

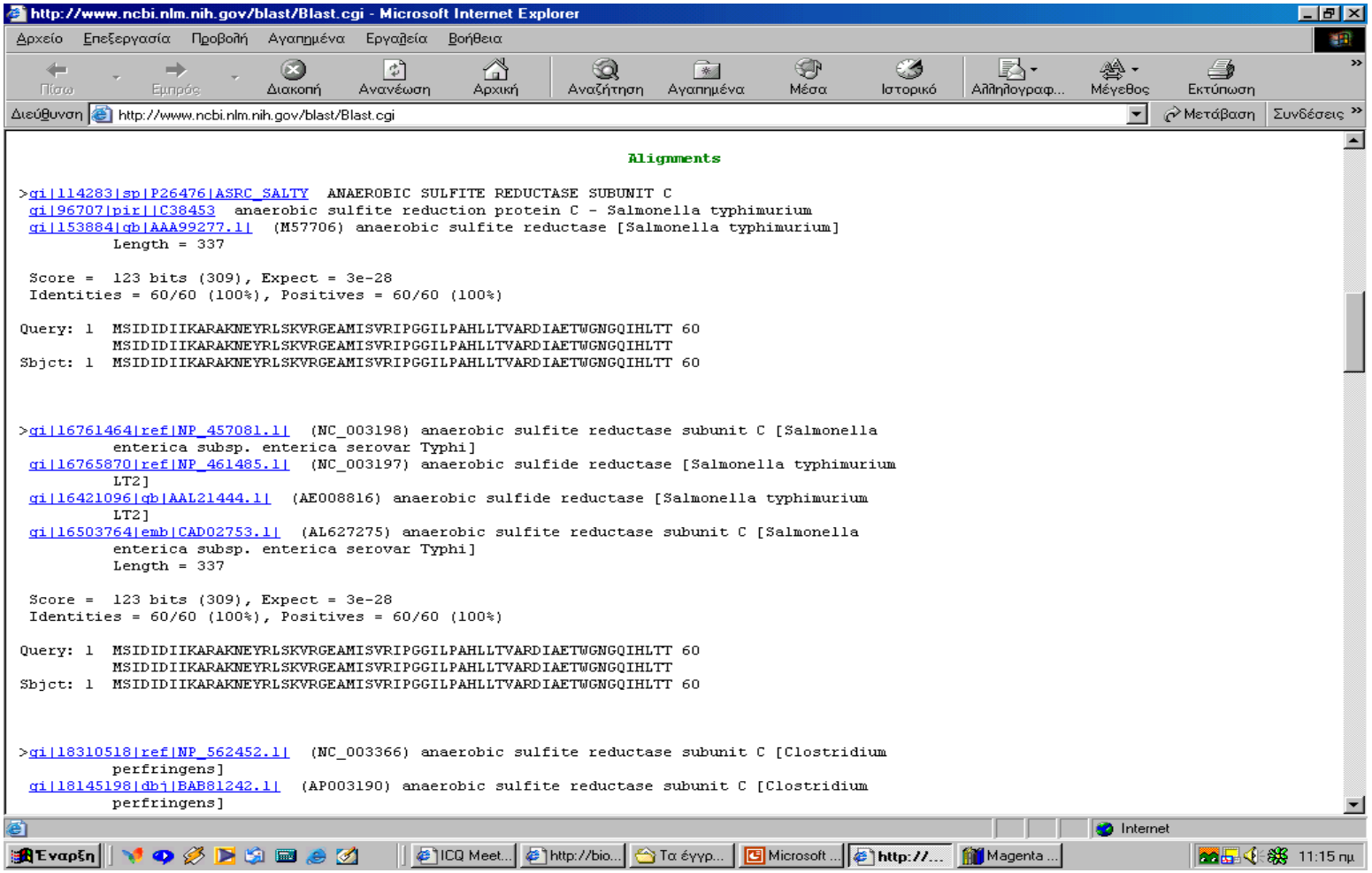
Εισαγωγή δεδομένων στο BLAST



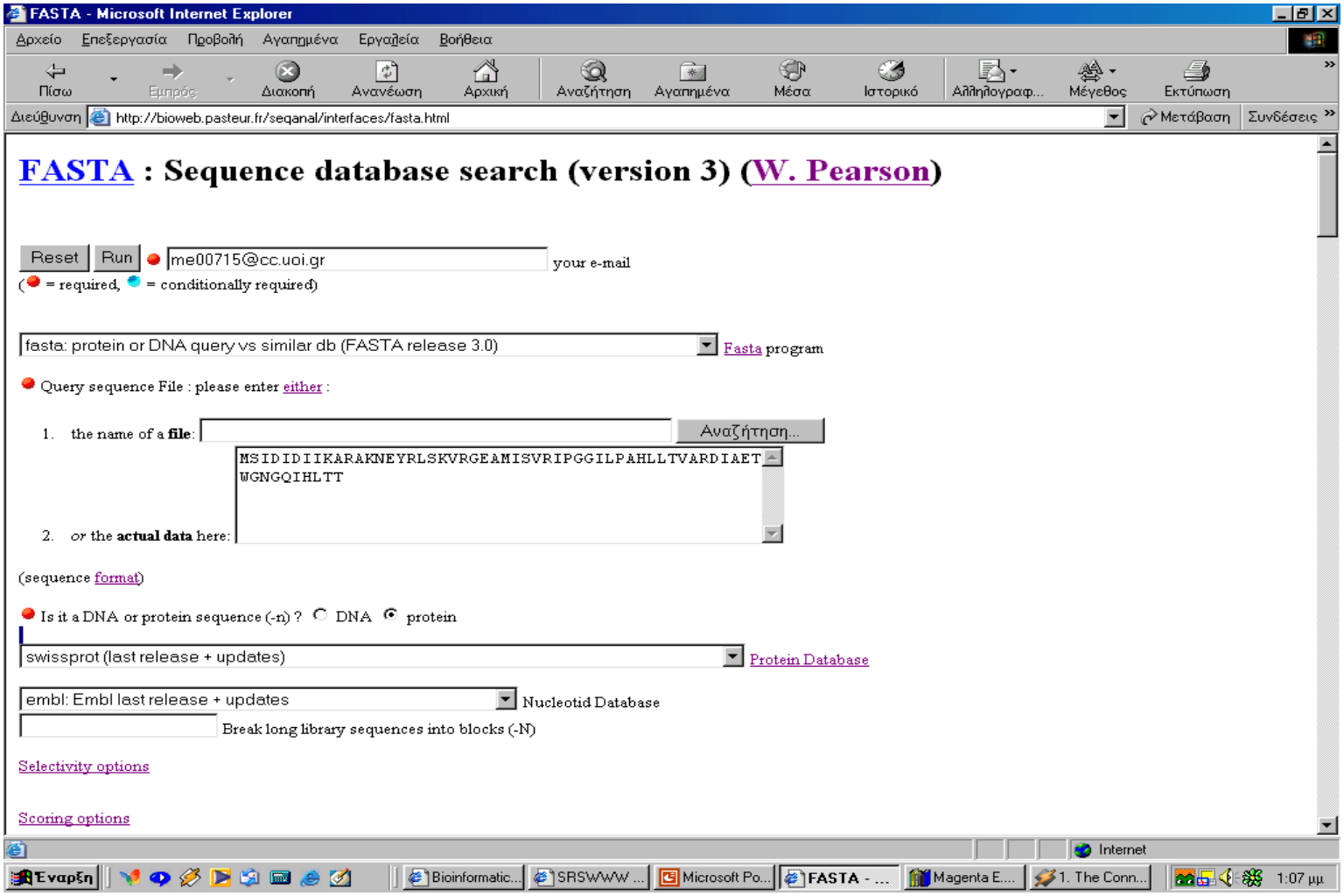
Αποτελέσματα BLAST



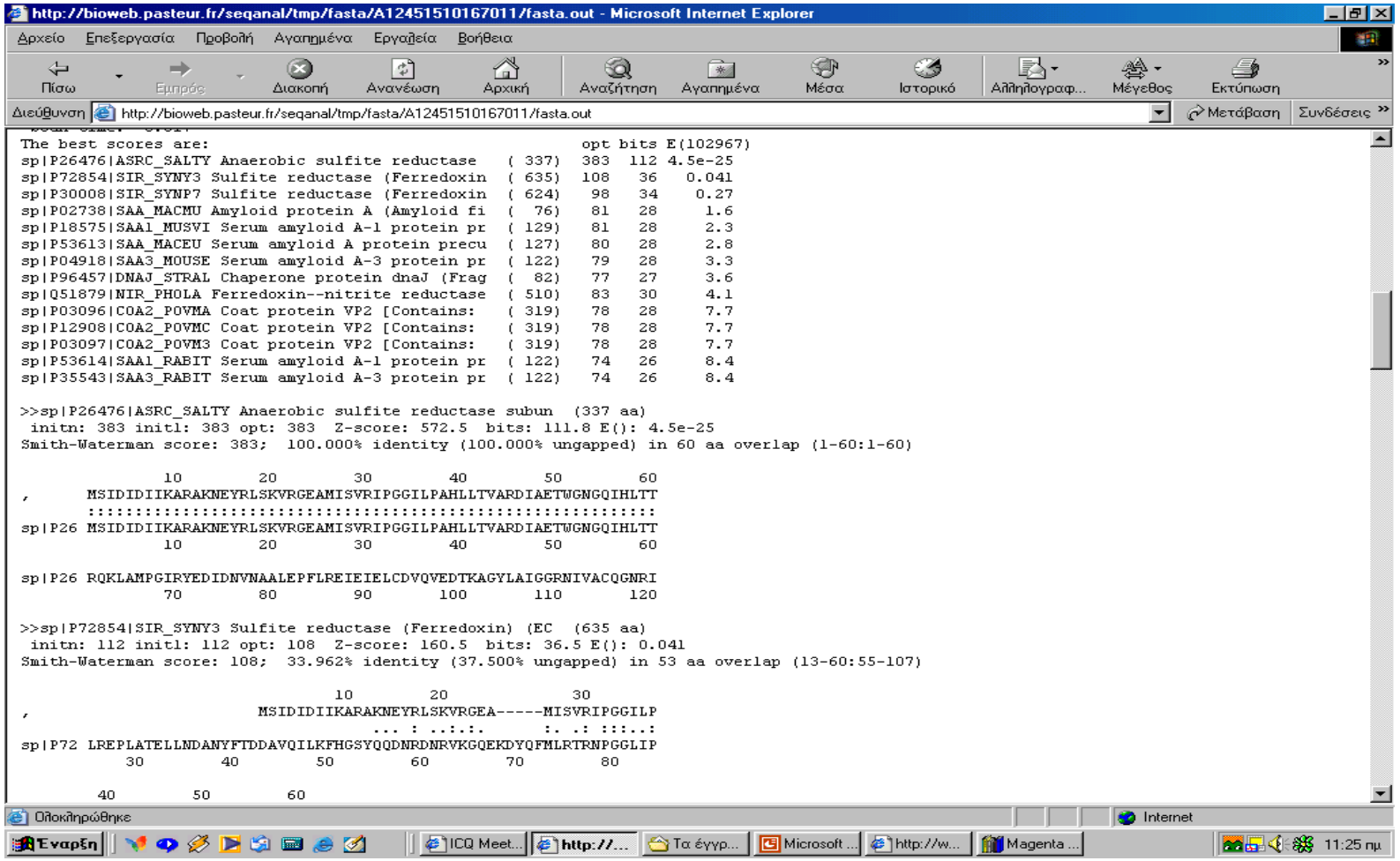
Αποτελέσματα BLAST



Εισαγωγή δεδομένων στο FASTA



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:                                      opt bits E(102967)
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

Έναρξη | ICQ Meet... | http://... | Τα έγγρ... | Microsoft ... | http://w... | Magenta ... | 11:25 πμ

MED LAB MEDICAL TECHNOLOGY & INTELLIGENT INFORMATION SYSTEMS

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

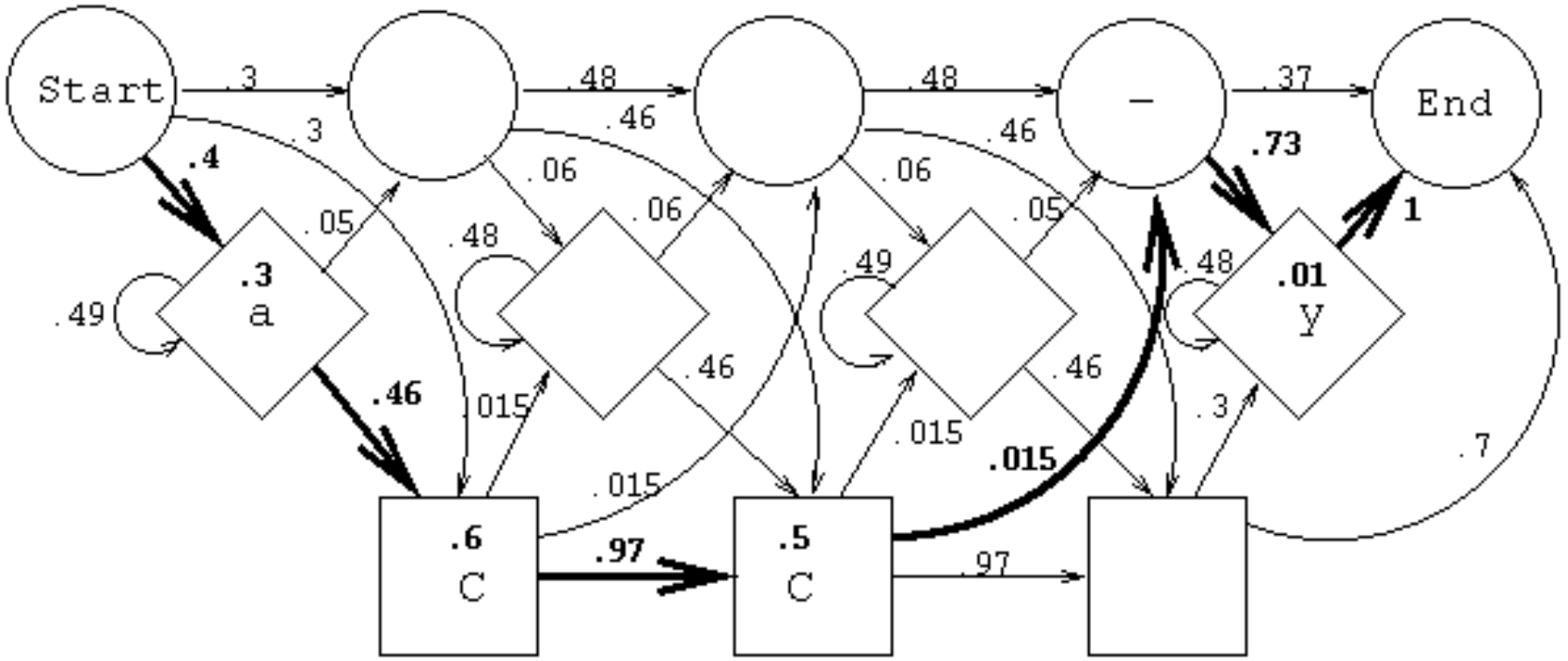
- ✓ Πρότυπα
  - ενότητες προτύπων,
  - δέντρα απόφασης,
  - ακριβής / προσεγγιστική αντιστοίχιση.
- ✓ Alignments, weight matrices, profiles, HMMs, neural networks,…

HMM profiles

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

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 αναγνωριστεί σε ποια γνωστή οικογένεια πρωτεϊνών ανήκει (αν ανήκει) μια νέα ακολουθία.

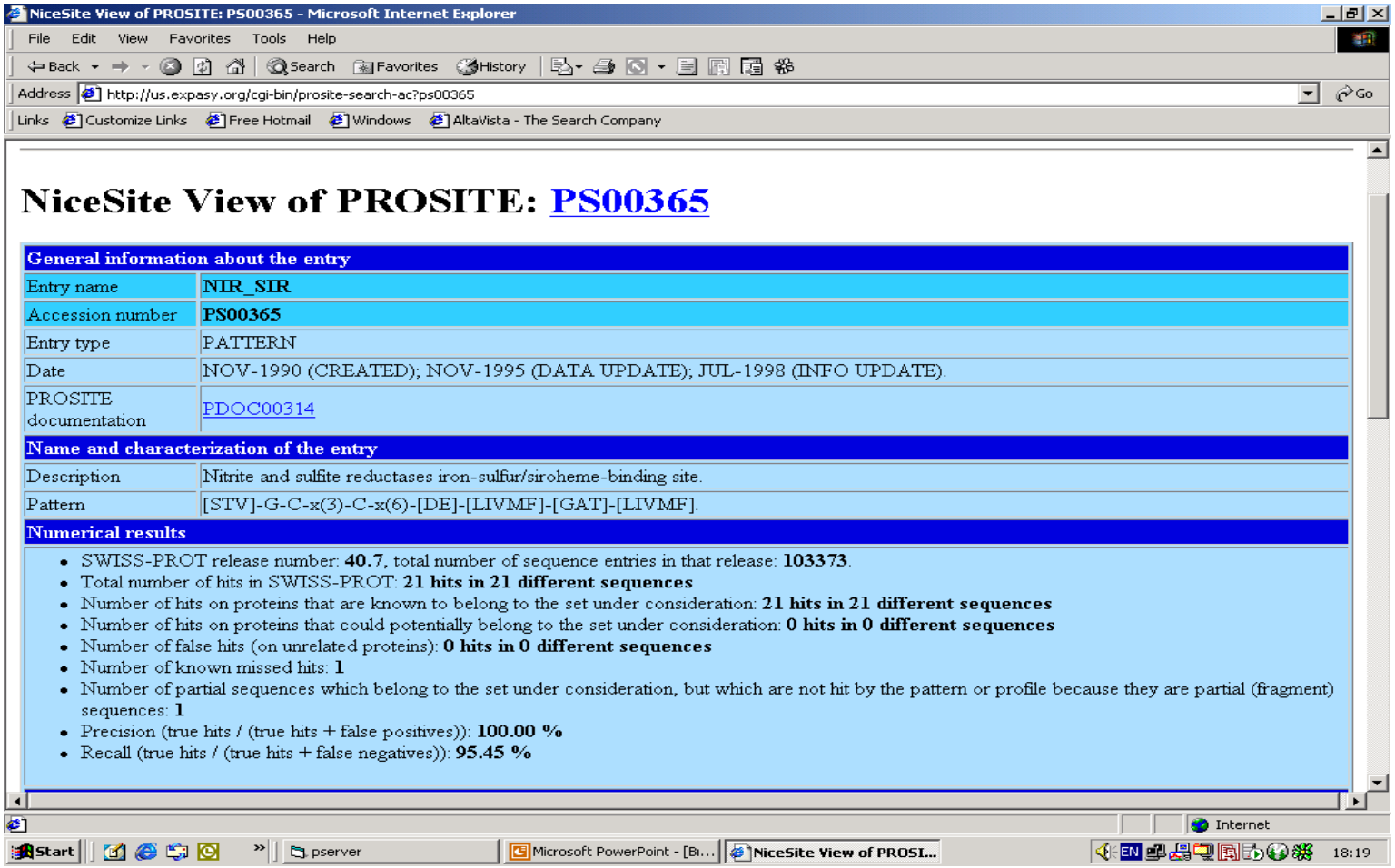
Πρότυπα PROSITE

- ‘x’ οποιοδήποτε αμινοξύ
- Ασάφειες:

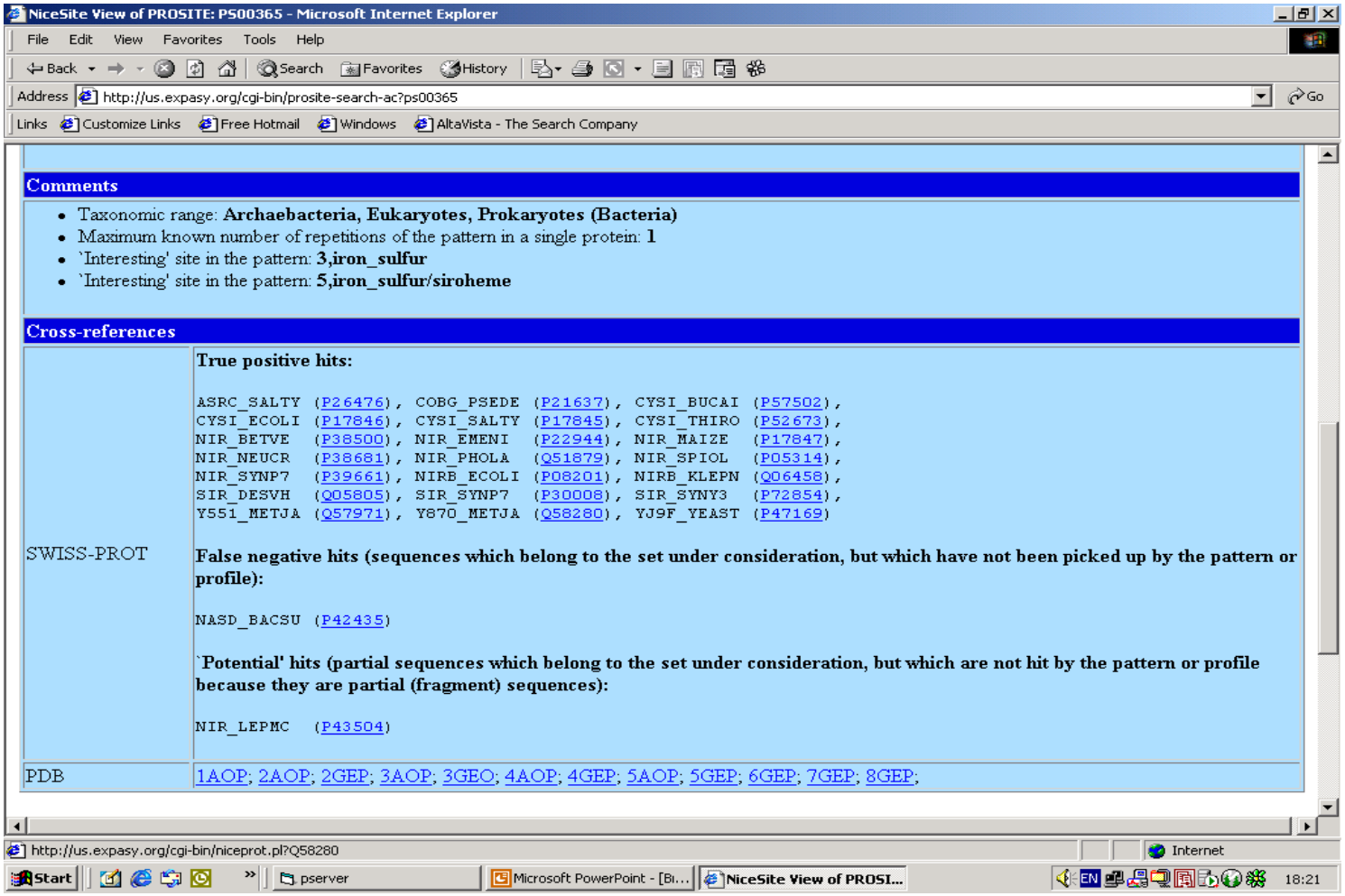
  [ALT] = Ala ή Leu ή Thr

  {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

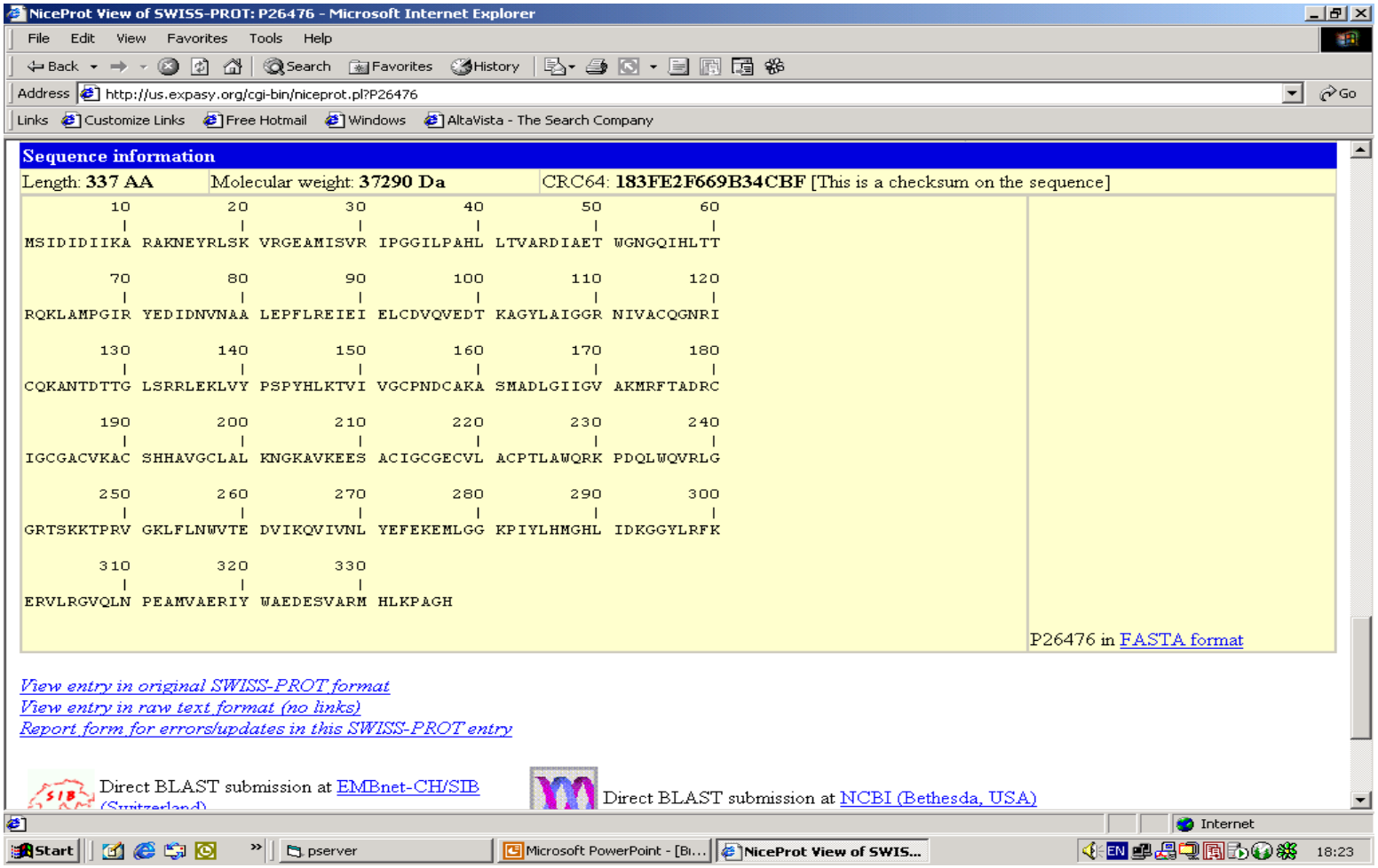
Παραδείγματα PROSITE



Παραδείγματα PROSITE



Παραδείγματα PROSITE

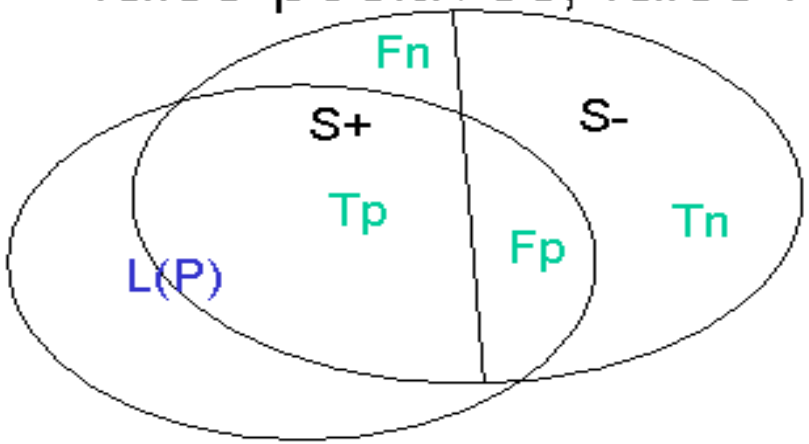


Πρόβλημα μάθησης

- 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.

Ύπαρξη 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

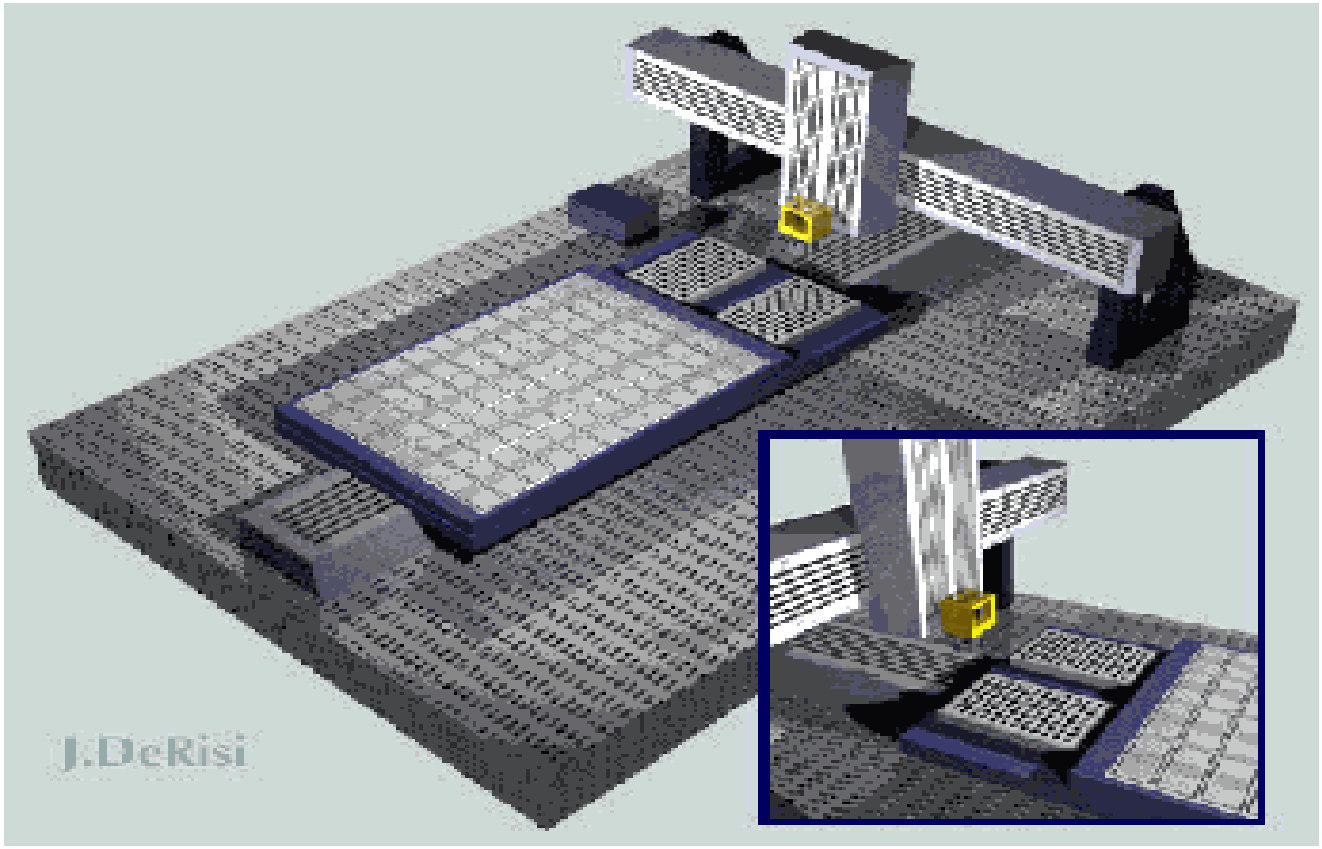
Microarray technology

- Η τεχνολογία micro array έχει σκοπό να επιταχύνει τη γενετική ανάλυση με τον ίδιο τρόπο που οι μικροεπεξεργαστές επιτάχυναν τους υπολογισμούς.

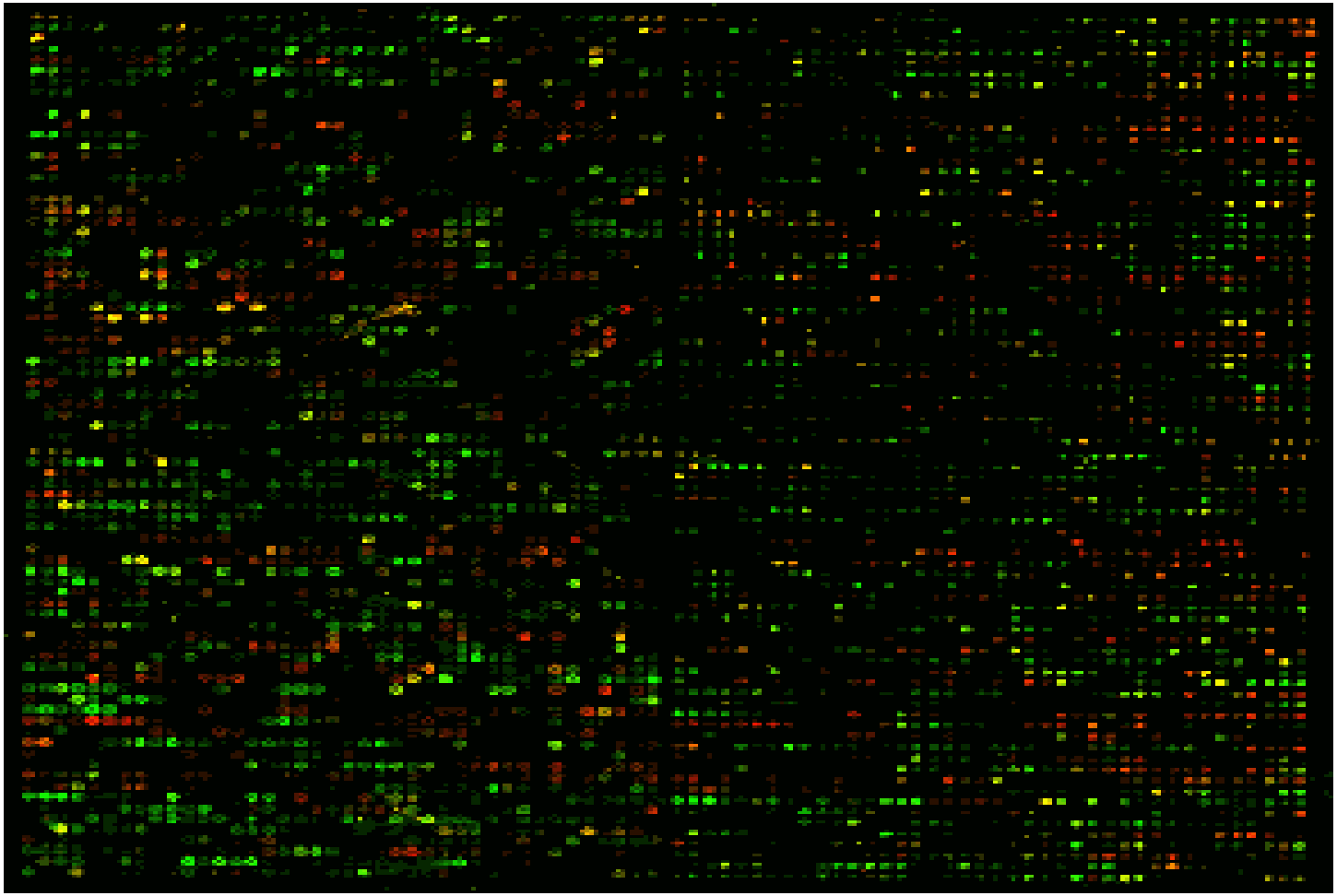
- Πρόκειται για μικροσκοπικούς πίνακες (arrays) από τμήματα του γονιδιώματος προσδεδεμένα σε γυάλινα chips.

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.