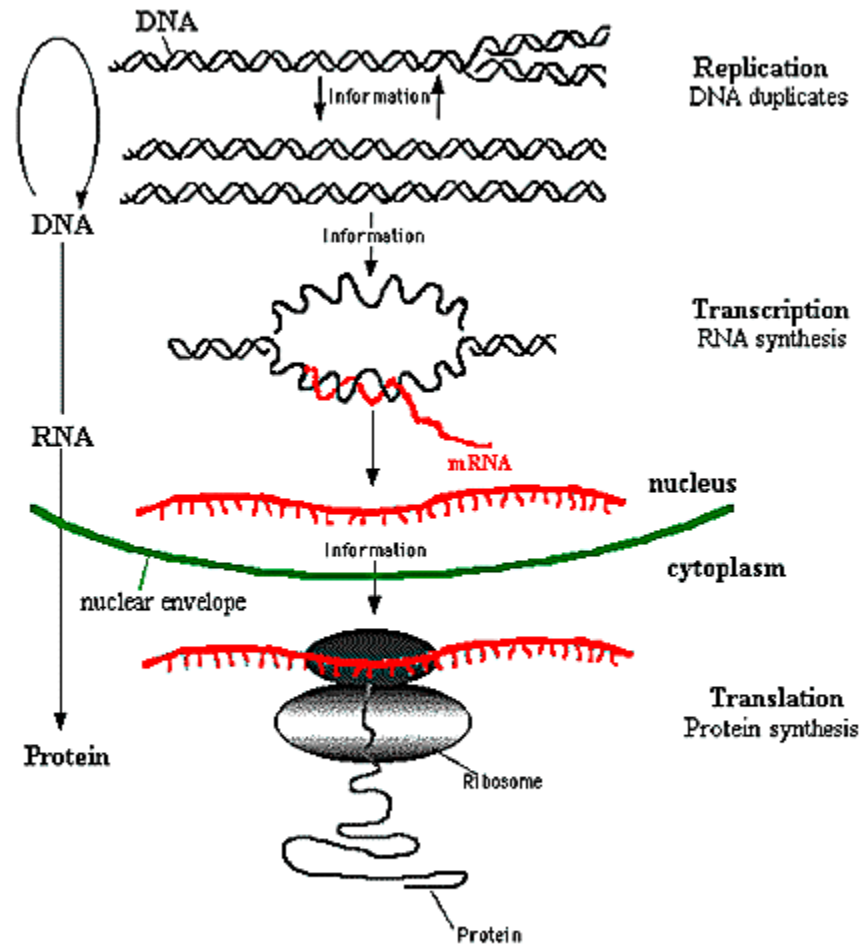


Βιοπληροφορική Ι

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Παν/μιο Θεσσαλίας
Λαμία 2015

Το Κεντρικό Δόγμα της Μοριακής Βιολογίας ...

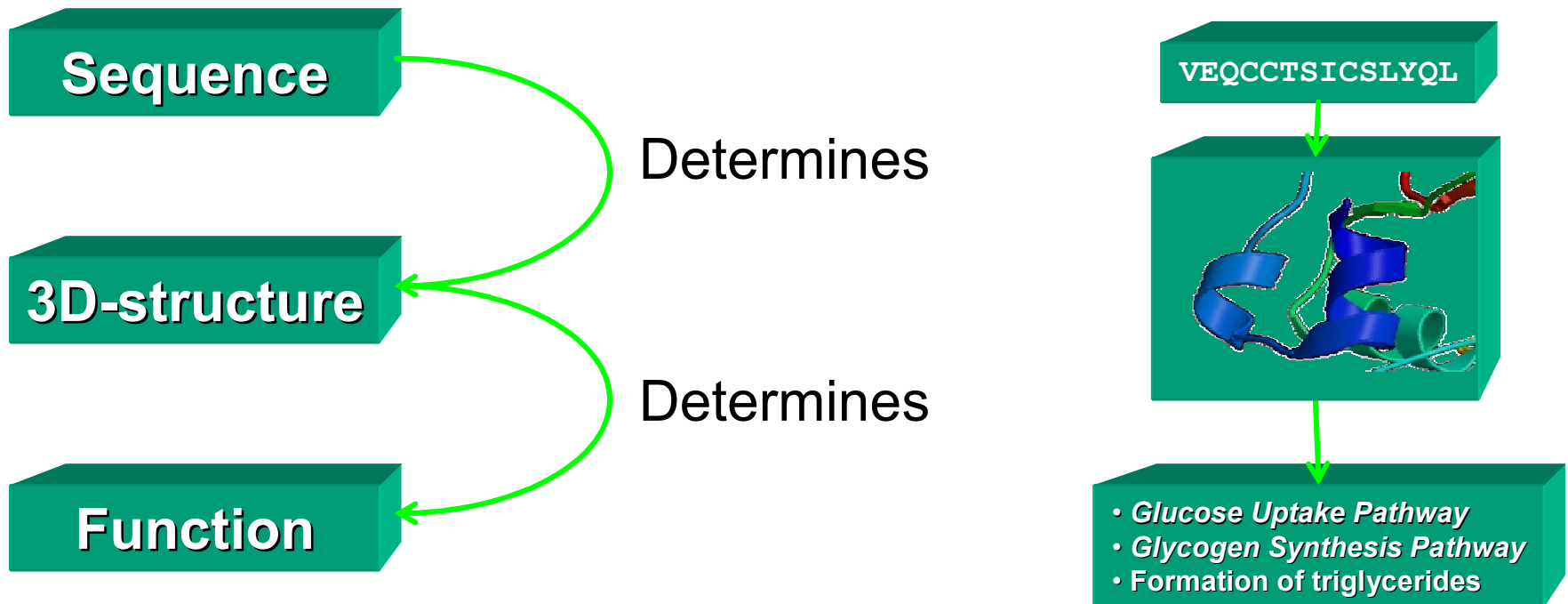


The Central Dogma of Molecular Biology

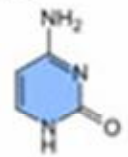
<http://www.accessexcellence.org/AB/GG/central.html>

Βιολογικές Βάσεις Δεδομένων

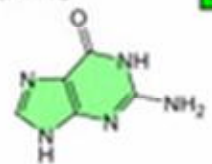
... και μια φυσική προέκτασή του ...



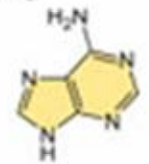
Cytosine **C**



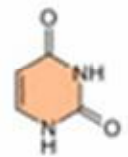
Guanine **G**



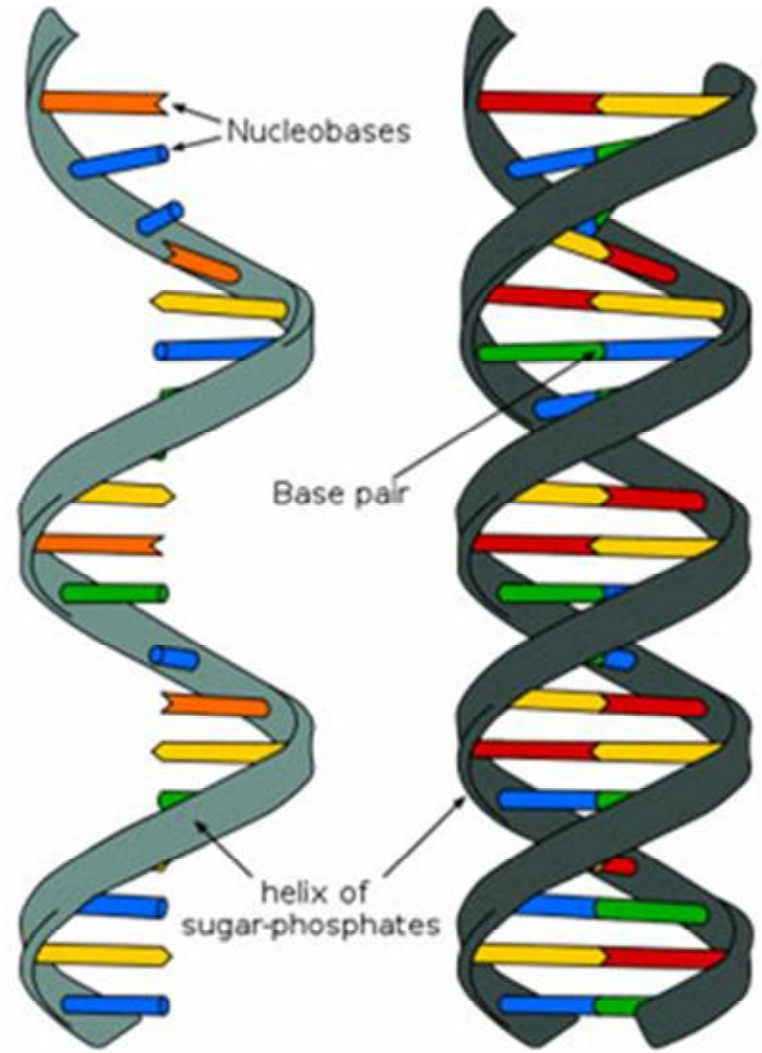
Adenine **A**



Uracil **U**



Nucleobases of RNA



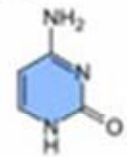
RNA

Ribonucleic acid

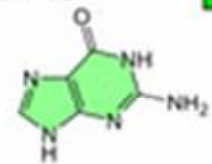
DNA

Deoxyribonucleic acid

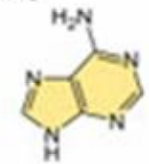
Cytosine **C**



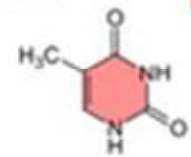
Guanine **G**



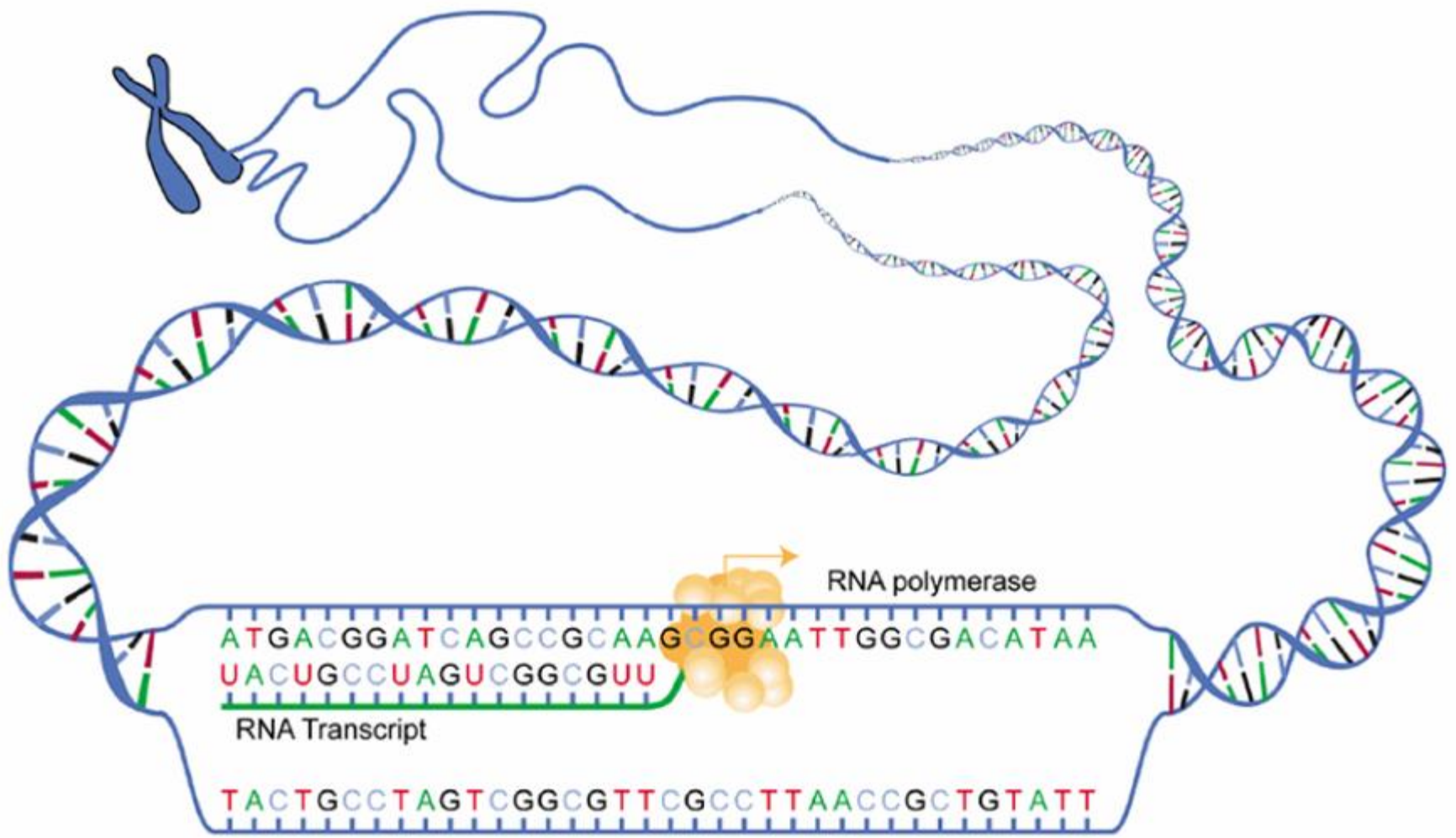
Adenine **A**



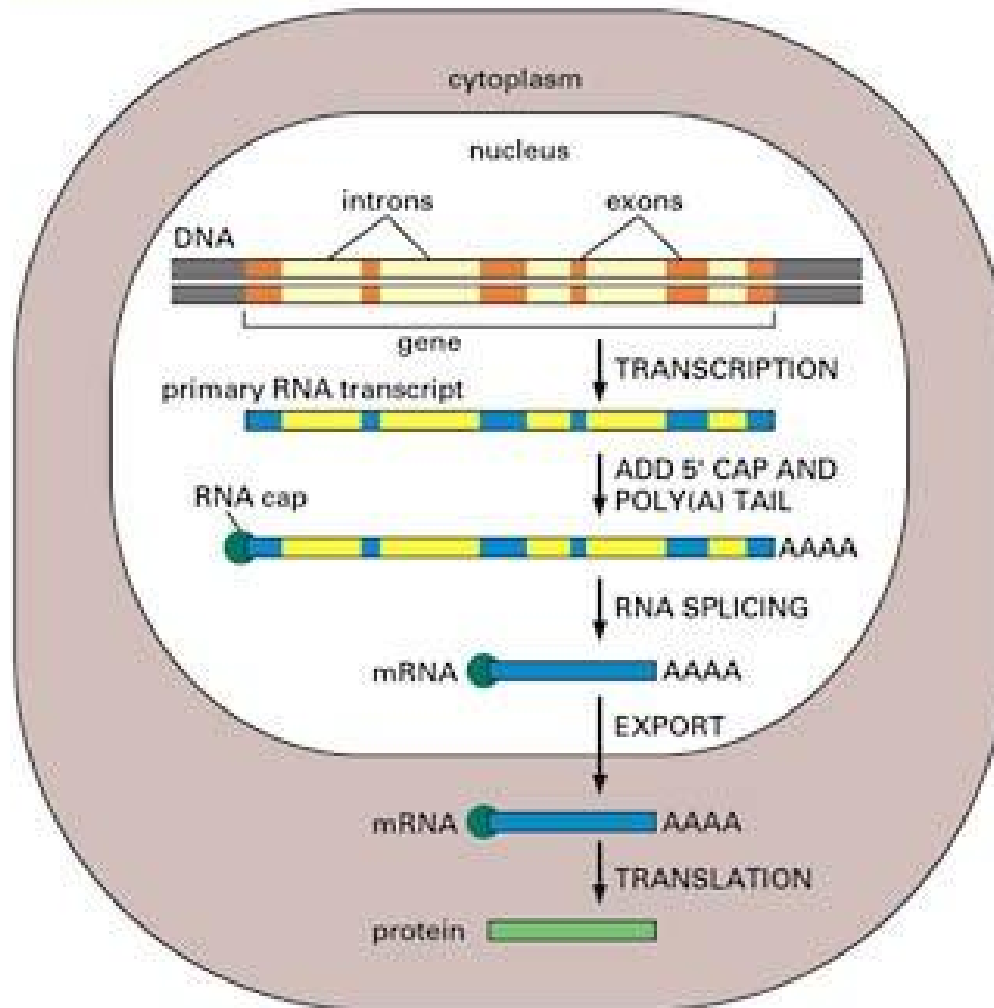
Thymine **T**



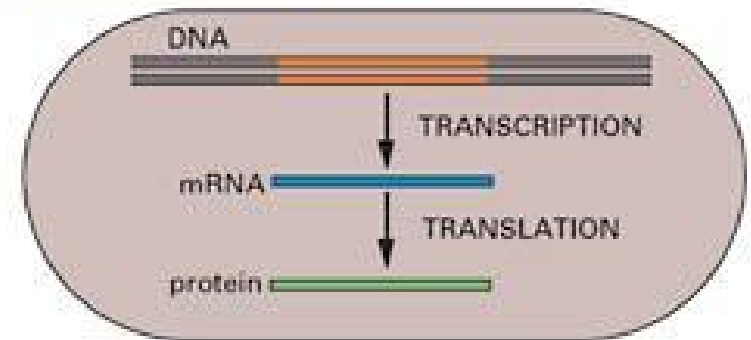
Nucleobases of DNA



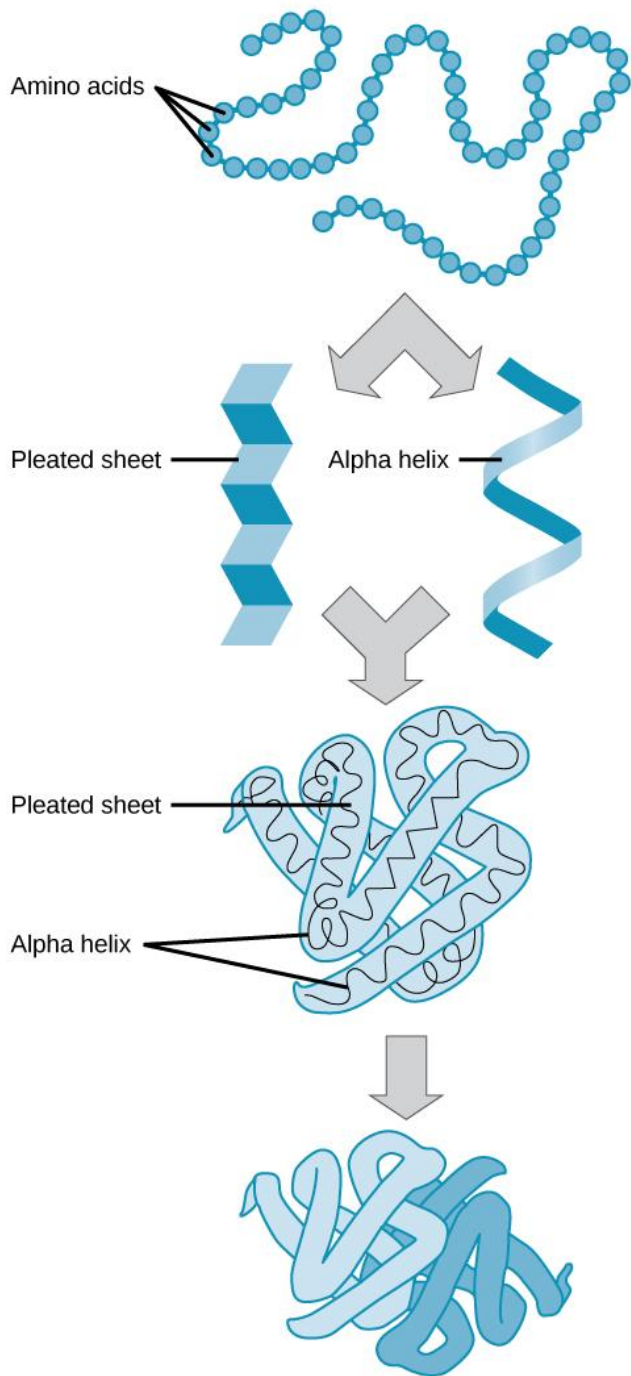
(A) EUCARYOTES



(B) PROCARYOTES



http://www.accessexcellence.org/AB/GG/steps_to_Prot.html



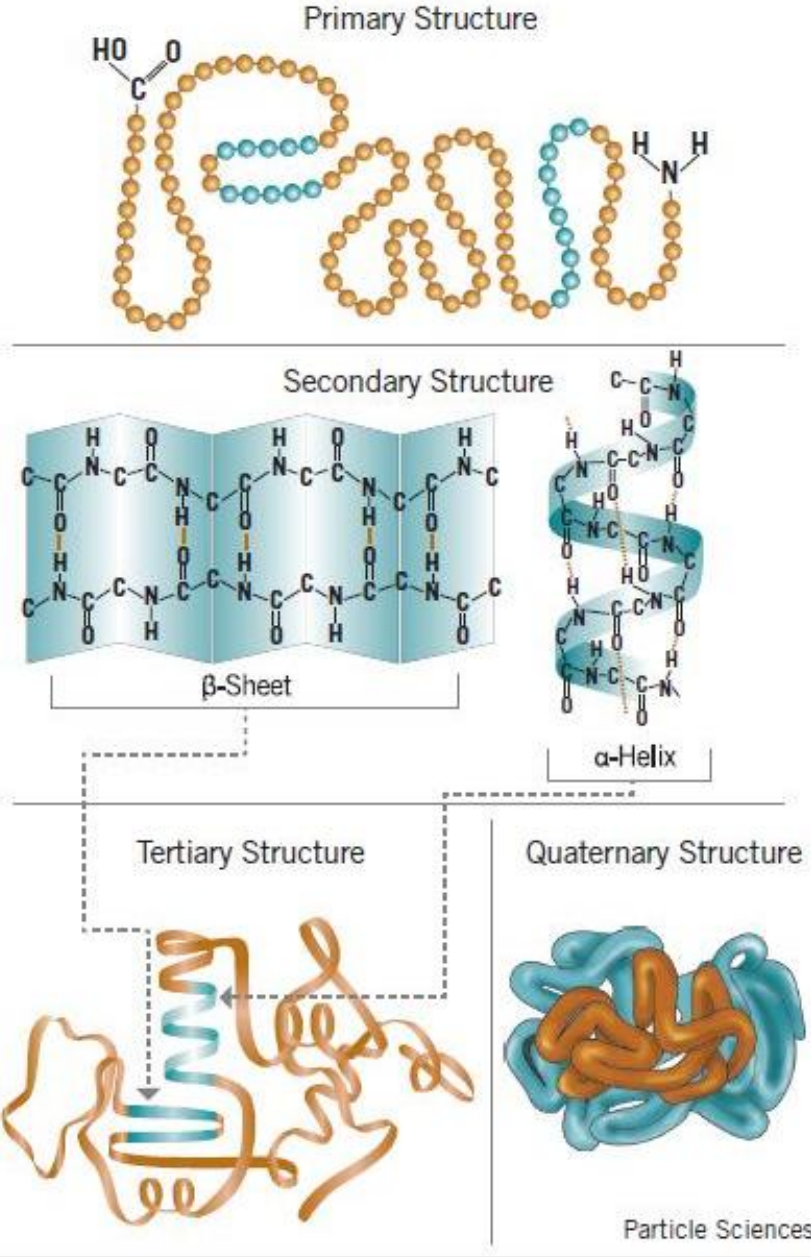
Primary Protein structure
sequence of a chain of amino acids

Secondary Protein structure
hydrogen bonding of the peptide backbone causes the amino acids to fold into a repeating pattern

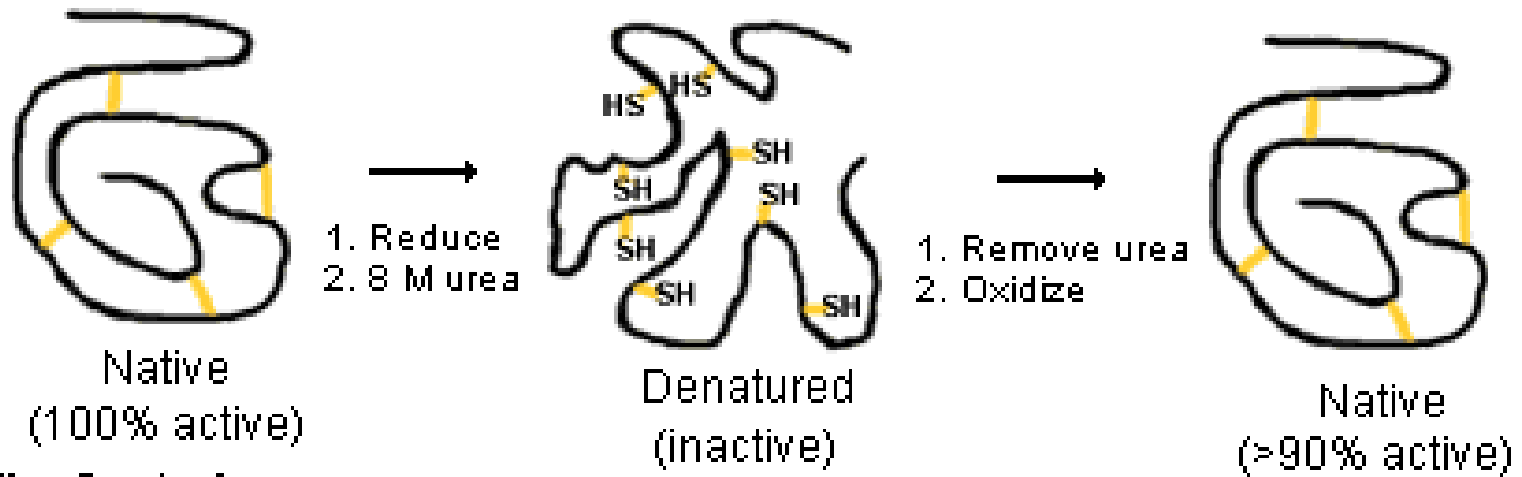
Tertiary protein structure
three-dimensional folding pattern of a protein due to side chain interactions

Quaternary protein structure
protein consisting of more than one amino acid chain

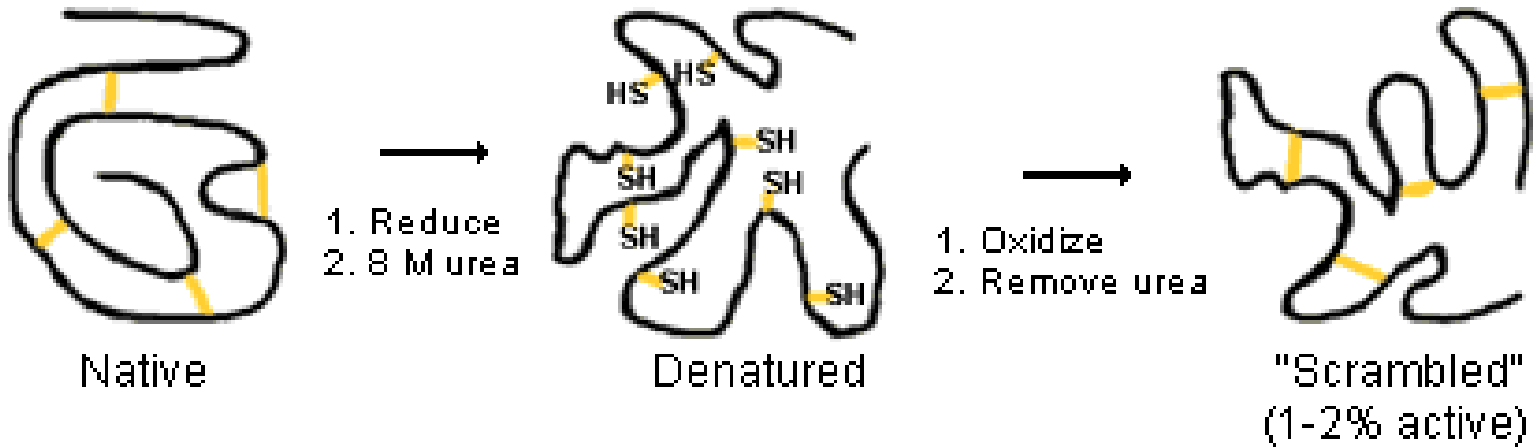
LEVELS OF PROTEIN STRUCTURE

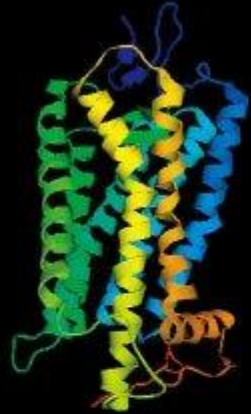


The Observation:



The Control:





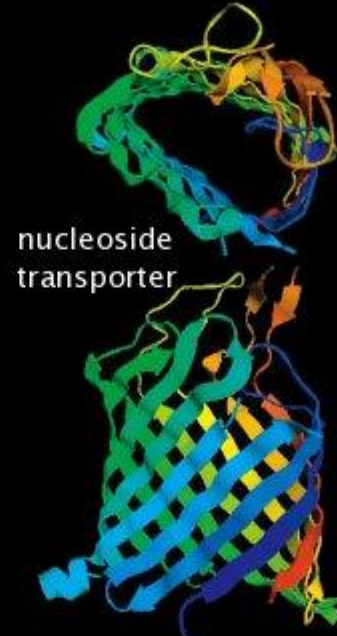
bovine rhodopsin



human telomere protein



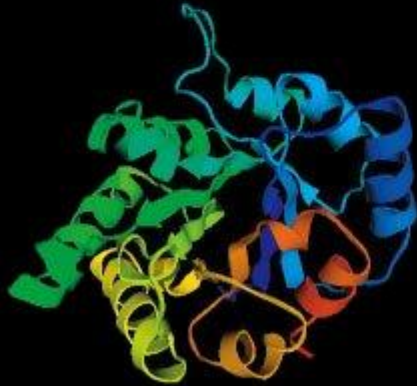
leucine rich repeat protein



nucleoside transporter



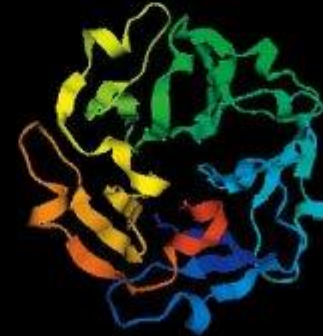
mouse cadherin



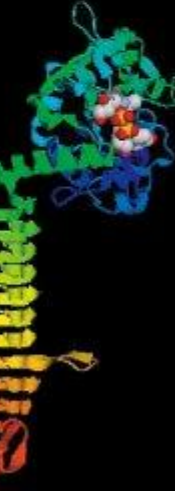
β/α -barrel form (TIM)



α/β superhelix (ribonuclease inhibitor)



5-propeller form (lectin)



solenoid form (transferase)

Βάσεις Βιολογικών Δεδομένων

- Πρωτογενείς βάσεις δεδομένων, οι οποίες περιέχουν τα πρωτογενή δεδομένα όπως αυτά προσδιορίζονται από τους πειραματικούς
 - Βάσεις δεδομένων ακολουθιών νουκλεοτιδικών ακολουθιών
 - Βάσεις δεδομένων ακολουθιών πρωτεϊνικών ακολουθιών
 - Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών
 - Βάσεις δεδομένων γονιδιακής έκφρασης
 - Βάσεις δεδομένων γενετικής ποικιλομορφίας
 - Βάσεις δεδομένων βιβλιογραφίας
- Δευτερογενείς βάσεις δεδομένων, στις οποίες υπάρχουν κυρίως ταξινομήσεις των πρωτογενών δεδομένων, χρήσιμες για αναλυτικούς σκοπούς
 - Βάσεις δεδομένων οικογενειών (κυρίως πρωτεϊνών)
 - Εξειδικευμένες βάσεις δεδομένων

Βάσεις δεδομένων ακολουθιών νουκλεοτιδικών ακολουθιών

- GenBank
- EMBL
- DDBJ

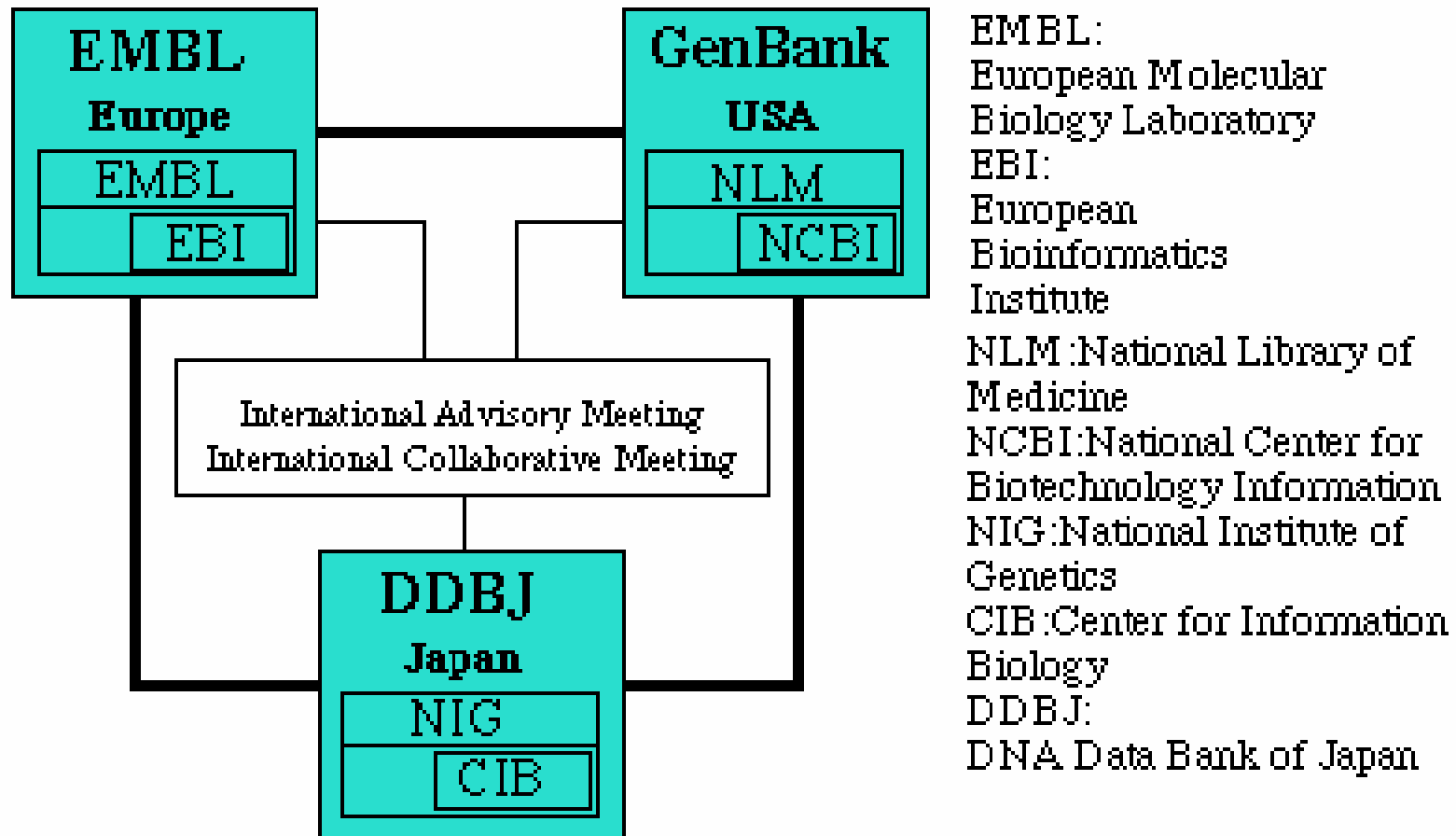
GENBANK: Η GENBANK

(<http://www.ncbi.nlm.nih.gov/Genbank/index.html>) είναι μια βάση νουκλεοτιδικών αλληλουχιών, διατίθεται ελεύθερα στην επιστημονική κοινότητα και βρίσκεται και υπό την αιγίδα του Εθνικού Ινστιτούτου Υγείας των Η.Π.Α (National Institutes of Health). Τα δεδομένα της βάσης προέρχονται από υποβολές δεδομένων διαφόρων ερευνητικών ομάδων όπως αυτά προκύπτουν από πειραματικές διεργασίες. Η διαδικασία υποβολής γίνεται με την συμπλήρωση κατάλληλης φόρμας μέσω διαδικτύου. Τα δεδομένα που υποβάλλονται στην βάση επεξεργάζονται, σχολιάζονται (annotate) από τους υπεύθυνους της βάσης και στη συνέχεια δημοσιοποιούνται σε αυτήν. Σε συχνά χρονικά διαστήματα τα δεδομένα που έχουν καταχωρηθεί στη βάση επανεξετάζονται και διορθώνονται σε περίπτωση που έχουν προκύψει νέα δεδομένα. Ο αριθμός των νουκλεοτιδικών βάσεων που περιέχονται στην GENBANK διπλασιάζεται κάθε 14 μήνες με αποτέλεσμα η τελευταία έκδοση (Rel. 206, Φεβρουάριος 2015) να περιέχει 181.336.445 ακολουθίες και 187.893.826.750 συνολικό αριθμό βάσεων.

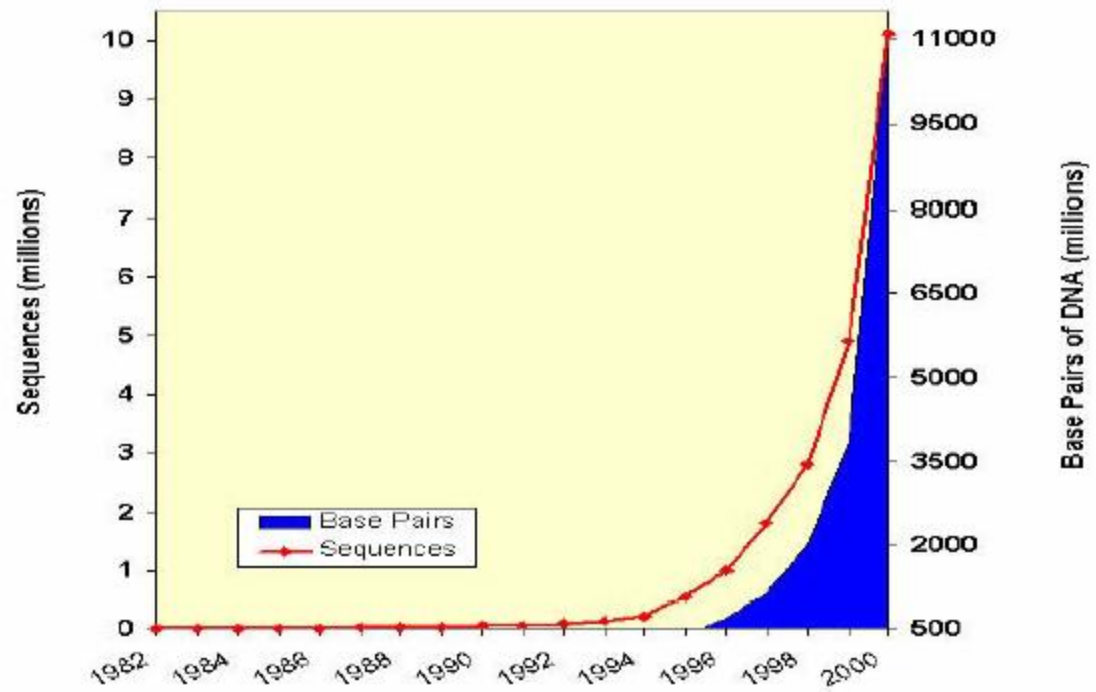
•**EMBL-Bank:** Η EMBL Nucleotide Sequence Database2 (<http://www.ebi.ac.uk/embl/>) αποτελεί τη μεγαλύτερη βάση νουκλεοτιδικών αλληλουχιών στην Ευρώπη, βρίσκεται υπό την αιγίδα του Ευρωπαϊκού Εργαστηρίου Μοριακής Βιολογίας (EMBL) ενώ εδράζεται και συντηρείται από το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EBI) στο Cambridge, UK. Οι ακολουθίες κατατίθενται στην EMBL-Bank μέσω διαδικτύου, ακολουθώντας μία απλή διαδικασία από ανεξάρτητα ερευνητικά εργαστήρια ή ομάδες που ασχολούνται με τον προσδιορισμό των γονιδιωμάτων διαφόρων οργανισμών. Αντίστοιχα με την GENBANK, οι νέες καταχωρήσεις ακολουθιών επεξεργάζονται, σχολιάζονται από τους υπεύθυνους της βάσης και δημοσιοποιούνται. Παράλληλα διατίθενται διάφορα εργαλεία ανάλυσης ακολουθιών όπως το Fasta και το BLAST. Η παρούσα έκδοση της EMBL-Bank (Rel. 122 - Νοέμβριος 2014) περιέχει 510.014.239 εγγραφές. Ο συνολικός αριθμός νουκλεοτιδίων φτάνει τα 1.094.969.877.589

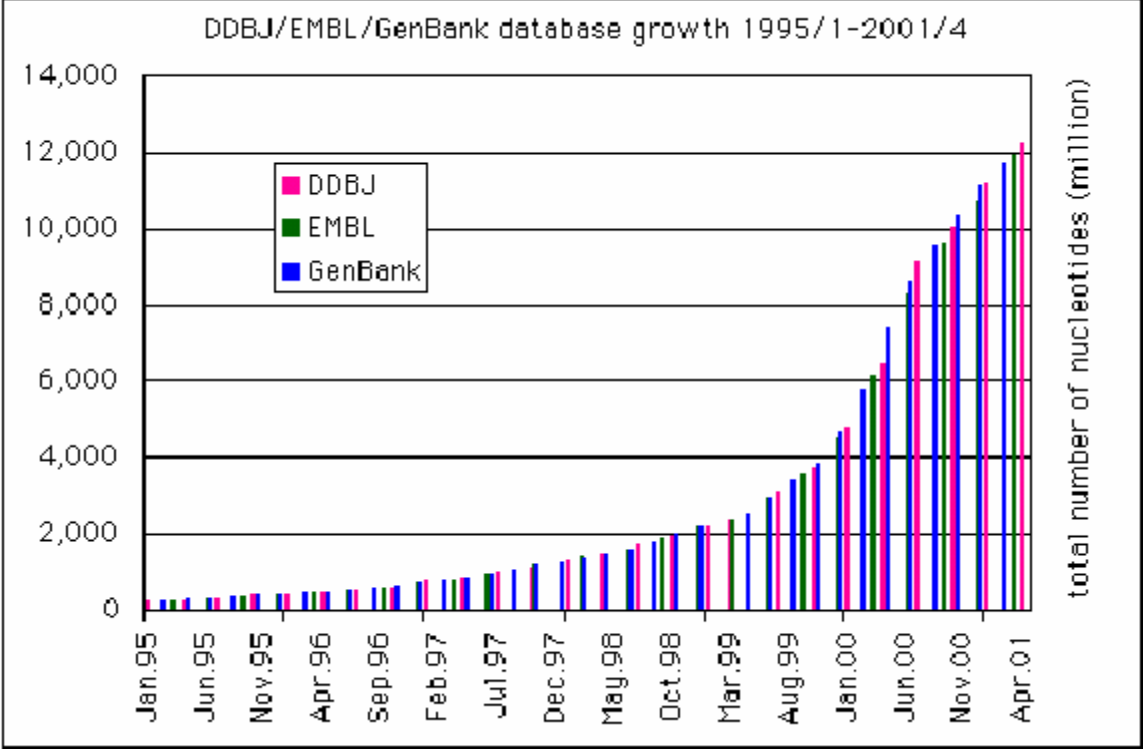
•**DDBJ:** Η DNA Databank of Japan (DDBJ - <http://www.ddbj.nig.ac.jp/>) είναι η μοναδική διεθνώς αναγνωρισμένη βάση νουκλεοτιδικών αλληλουχιών στην Ιαπωνία. Ιδρύθηκε το 1986 στο Εθνικό Ινστιτούτο Γενετικής (NIG) και βρίσκεται υπό την αιγίδα του Υπουργείου Παιδείας, Επιστημών και Αθλητισμού της Ιαπωνίας. Βασική πηγή δεδομένων της βάσης αποτελούν οι εργασίες των Ιαπώνων ερευνητών. Επιπλέον στην DDJB είναι διαθέσιμα διάφορα εργαλεία ανάλυσης νουκλεοτιδικών αλληλουχιών. Η παρούσα έκδοση της DDJB (Rel. 99, Δεκέμβριος 2014) περιέχει 178.825.615 εγγραφές και συνολικά 184.410.381.191 νουκλεοτιδικές βάσεις που περιέχονται στις ακολουθίες.

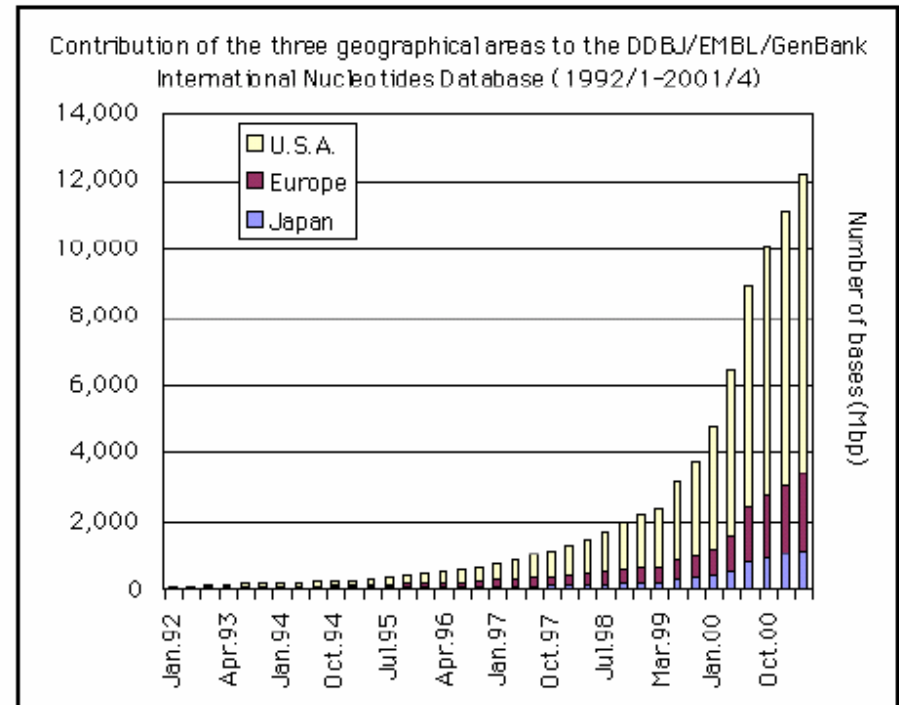
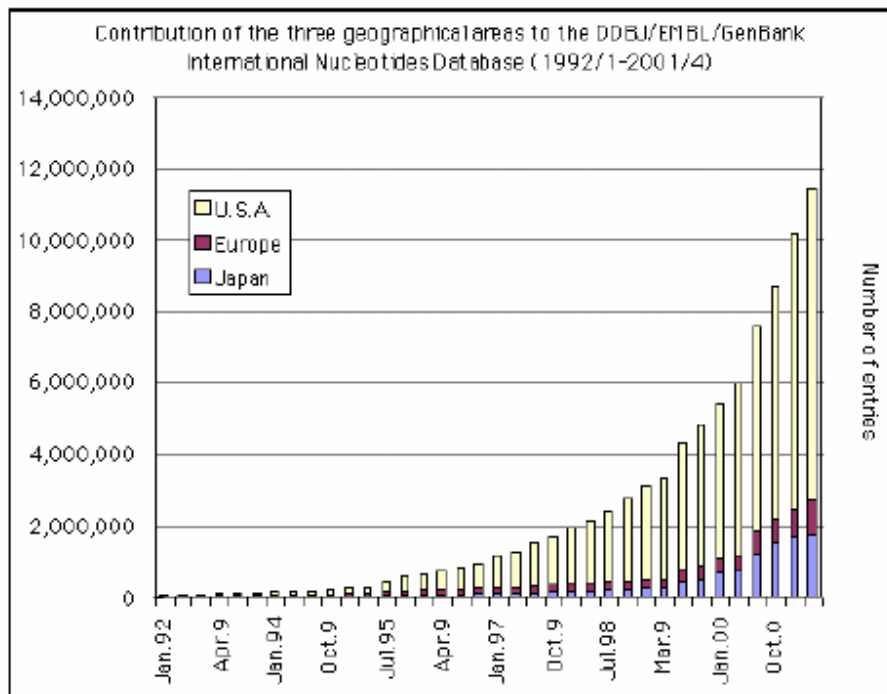
International Nucleotide Data Banks

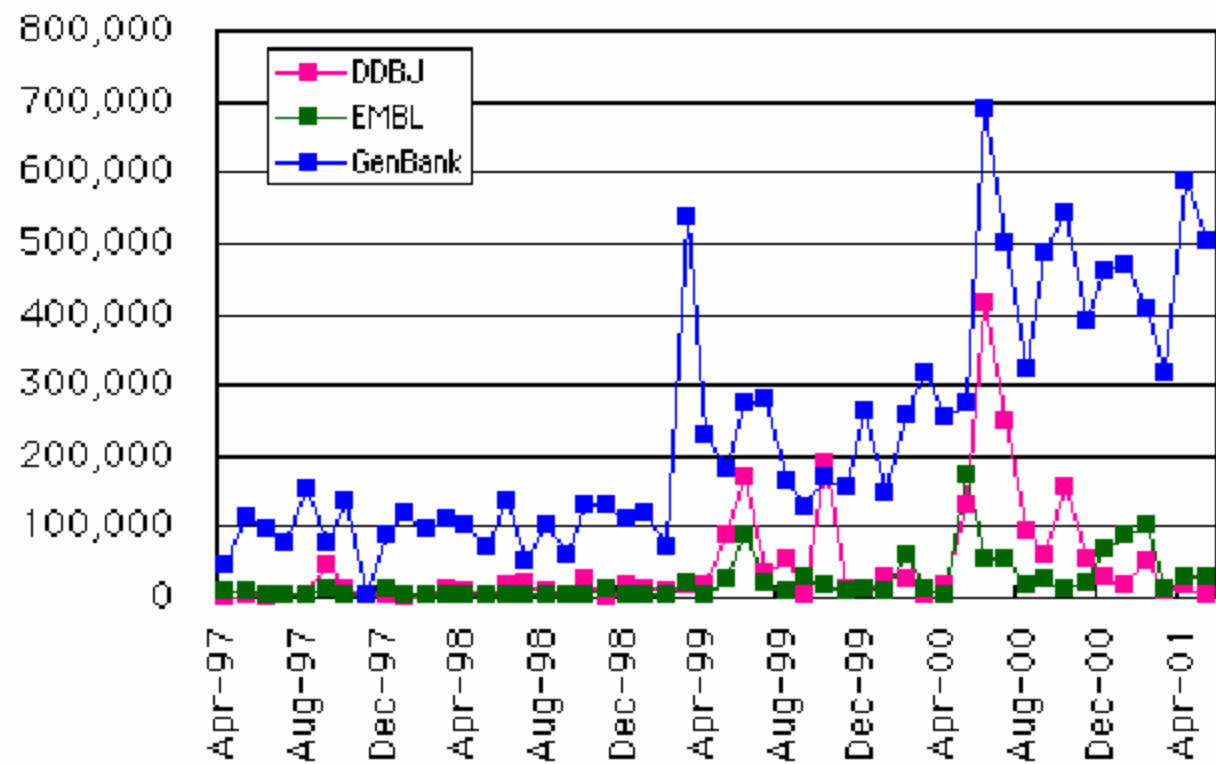


Growth of GenBank









LOCUS SCU49845 5028 bp DNA PLN 21-JUN-1999
 DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Ax12p
 (AXL2) and Rev7p (REV7) genes, complete cds.
 ACCESSION U49845
 VERSION U49845.1 GI:1293613
 KEYWORDS .
 SOURCE Saccharomyces cerevisiae (baker's yeast)
 ORGANISM Saccharomyces cerevisiae
 Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
 Saccharomycetales; Saccharomycetaceae; Saccharomyces.
 REFERENCE 1 (bases 1 to 5028)
 AUTHORS Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
 TITLE Cloning and sequence of REV7, a gene whose function is required for
 DNA damage-induced mutagenesis in Saccharomyces cerevisiae
 JOURNAL Yeast 10 (11), 1503-1509 (1994)
 PUBMED 7871890
 REFERENCE 2 (bases 1 to 5028)
 AUTHORS Roemer,T., Madden,K., Chang,J. and Snyder,M.
 TITLE Selection of axial growth sites in yeast requires Ax12p, a novel
 plasma membrane glycoprotein
 JOURNAL Genes Dev. 10 (7), 777-793 (1996)
 PUBMED 8846915
 REFERENCE 3 (bases 1 to 5028)
 AUTHORS Roemer,T.
 TITLE [Direct Submission](#)
 JOURNAL Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
 Haven, CT, USA
 FEATURES Location/Qualifiers
 source 1..5028
 /organism="Saccharomyces cerevisiae"

Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
 Saccharomycetales; Saccharomycetaceae; Saccharomyces.

REFERENCE 1 (bases 1 to 5028)
 AUTHORS Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
 TITLE Cloning and sequence of REV7, a gene whose function is required for
 DNA damage-induced mutagenesis in *Saccharomyces cerevisiae*
 JOURNAL Yeast 10 (11), 1503-1509 (1994)
 PUBMED 7871890

REFERENCE 2 (bases 1 to 5028)
 AUTHORS Roemer,T., Madden,K., Chang,J. and Snyder,M.
 TITLE Selection of axial growth sites in yeast requires Axl2p, a novel
 plasma membrane glycoprotein
 JOURNAL Genes Dev. 10 (7), 777-793 (1996)
 PUBMED 8846915

REFERENCE 3 (bases 1 to 5028)
 AUTHORS Roemer,T.
 TITLE [Direct Submission](#)
 JOURNAL Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
 Haven, CT, USA

FEATURES Location/Qualifiers
 source 1..5028
 /organism="Saccharomyces cerevisiae"
 /db_xref="taxon:4932"
 /chromosome="IX"
 /map="9"
 CDS <1..206
 /codon_start=3
 /product="TCP1-beta"
 /protein_id="AAA98665.1"
 /db_xref="GI:1293614"
 /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLRVSSASEA
 AEVLLRVDNIIRARPRTANRQHM"
 gene 687..3158
 /gene="AXL2"
 CDS 687..3158
 /gene="AXL2"
 /note="plasma membrane glycoprotein"
 /codon_start=1
 /function="required for axial budding pattern of *S.*
cerevisiae"

```

/gene="AXL2"
/note="plasma membrane glycoprotein"
/codon_start=1
/function="required for axial budding pattern of S.
cerevisiae"
/product="Axl2p"
/protein_id="AAA98666.1"
/db_xref="GI:1293615"
/translation="MTQLQISLLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTL YFN
VILEGTDSDSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTINGKNALKLDPNE
VFNVTFRDRSMFTNEESIVSYGRSQLYNAPLPNWLF FDSGELKFTGTAPVINSAlAPE
TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLI INVTD TGNVSYDLPLNYV
YLDDDPISD KLGSINLLDAPD WVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
DVIYFNFEVWSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTVNSLEF TNSQ
DHDWVKFQSSNLT LAGEVPKNFDKLSLGLKANQGSQSQEL YFNIIGMDSKITHSNHSA
NATSTRSSHSTSTSSYTSSTYTAKISSTSAATSSAPAALPAANKTSSHNKKAVALA
CGVAIPLGVILVALICFLIFWRRRRENPD DENLPHAI SGPDLN NPANKPNQENATPLN
NPFDDDASSYDDTSIARRLAALN TLKLDNHSATESDISSVDEKRDSLGMNTYNDQFQ
SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTG NLSPVSDIVRDS
YGSQKTVDTEKLF DLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
HRNRHLQNIQDSQSGKNGITPTT MSTSSSDDFVPVKDGENFCWVHSM EPDRRPSKKRL
VDFSNKSNVNVGQVKDIHGRIPEML"

```

gene [complement](#) (3300..4037)

CDS [complement](#) (3300..4037)

```

/gene="REV7"
/codon_start=1
/product="Rev7p"
/protein_id="AAA98667.1"
/db_xref="GI:1293616"
/translation="MNRWVEKWL RVYLKCYINLILFYRNVYPPQSF DYT TYQSFNLPQ
FVPINRHPALIDYIEELILDVLSKLTHVYRFSIC I INKKNDLCIEKYVLD FSELQHVD
KDDQII TETE VFDEFRSSLNSLIMHLEKLPKVND D TITFEAVINAIELELGHKLDNRN
RVDSLEEKAEIERDSNVVKCQEDENLPD NNGFQPPKIKL TSLVGS DVGPLI IHQFSEK
LISGDDKILNGVYSQYEEGESIFGSLF"

```

ORIGIN

```

1 gatcctccat atacaacggt atctccacct cagggttaga tctcaacaac ggaaccattg
61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
121 ctgcatctga agccgctgaa gttctactaa ggggtggataa catcatccgt gcaagaccaa
181 gaaccaccaa taacaaacat atataacata ttttaggatat acctcgaaaa taataaaccc

```

GenBank Format

LOCUS LISOD 756 bp DNA BCT 30-JUN-1993
 DEFINITION L.ivanovii sod gene for superoxide dismutase.
 ACCESSION X64011 S78972
 NID g44010
 VERSION X64011.1 GI:44010
 KEYWORDS sod gene; superoxide dismutase.
 SOURCE Listeria ivanovii.
 ORGANISM Listeria ivanovii
 Bacteria; Firmicutes; Bacillus/Clostridium group; Bacillaceae;
 Listeria.
 REFERENCE 1 (bases 1 to 756)
 AUTHORS Haas,A. and Goebel,W.
 TITLE Cloning of a superoxide dismutase gene from Listeria ivanovii by
 functional complementation in Escherichia coli and characterization
 of the gene product
 JOURNAL Mol. Gen. Genet. 231 (2), 313-322 (1992)
 MEDLINE 92140371
 REFERENCE 2 (bases 1 to 756)
 AUTHORS Kreft,J.
 TITLE Direct Submission
 JOURNAL Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologie,
 Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzburg, FRG
 FEATURES Location/Qualifiers
 source 1..756
 /organism="Listeria ivanovii"
 /strain="ATCC 19119"
 /db_xref="taxon:1638"
 RBS 95..100
 /gene="sod"
 gene 95..746
 /gene="sod"
 CDS 109..717
 /gene="sod"
 /EC_number="1.15.1.1"
 /codon_start=1
 /transl_table=11
 /product="superoxide dismutase"
 /protein_id="CAA45406.1"
 /db_xref="SWISS-PROT:P28763"
 /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVS
 GHAEIASKPGEELVANLDSVPEEIRGAVRNHGGGHHHTLFWSSLSPNGGGAPTGNLK
 AAIESEFGTFDEFKEKFNAAAARFGSGWAWLVVNNKLEIVSTANQDPSLSEKTPV
 LGLDVWEHAYLKFQNRPRPEYIDTFWNVINWDERNKRFDAAK"
 terminator 723..746
 /gene="sod"
 BASE COUNT 247 a 136 c 151 g 222 t
 ORIGIN
 1 cgttatntaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat
 61 gtaatttctt ttcacataaa taataaacaa tccgaggagg aatttttaat gacttacgaa
 121 ttaccaaaaat taccttatac ttatgatgct ttggagccga attttgataa agaacaatg
 181 gaaattcact atacaaagca ccacaatatt tatgtaacaa aactaaatga agcagtcctc
 241 ggacacgcag aacttgcaag taaacctggg gaagaattag ttgctaactc agatagcggt
 301 cctgaagaaa ttctgtggcg agtacgtaac cacgggtggtg gacatgctaa ccatacttta
 361 ttctgggtcta gtcttagccc aaatgggtgt ggtgtccaa ctggttaactt aaaagcagca
 421 atcgaaaagcg aattcggcac atttgatgaa ttcaaagaaa aattcaatgc ggcagctgcg
 481 gctcgttttg gttcaggatg ggcattggcta gtagtgaaca atggtaaac agaaattggt
 541 tccactgcta accaagattc tccacttagc gaaggtaaaa ctccagttct tggcttagat
 601 gtttgggaac atgcttatta tcttaaattc caaaaccgtc gtcctgaata cattgacaca
 661 ttttgggaatg taattaactg ggatgaacga aataaacgct ttgacgcagc aaaataatta
 721 tcgaaaggct cacttaggtg ggtcttttta tttcta

//

DDBJ Format

LOCUS LISOD 756 bp DNA BCT 30-JUN-1993
DEFINITION L.ivanovii sod gene for superoxide dismutase.
ACCESSION X64011 S78972
NID g44010
VERSION X64011.1
KEYWORDS sod gene; superoxide dismutase.
SOURCE Listeria ivanovii.
ORGANISM Listeria ivanovii
Bacteria; Firmicutes; Bacillus/Clostridium group; Bacillaceae;
Listeria.
REFERENCE 1
AUTHORS Haas,A. and Goebel,W.
TITLE Cloning of a superoxide dismutase gene from Listeria ivanovii by
functional complementation in Escherichia coli and characterization
of the gene product.
JOURNAL Mol. Gen. Genet. 231, 313-322(1992)
MEDLINE 92140371
REFERENCE 2 (bases 1 to 756)
AUTHORS Kreft,J.
JOURNAL Submitted (21-APR-1992) to the EMBL/GenBank/DDBJ databases. J.
Kreft, Institut f. Mikrobiologie, Universitaet Wuerzburg, Biozentrum
Am Hubland, 8700 Wuerzburg, FRG
FEATURES Location/Qualifiers
source 1..756
/organism="Listeria ivanovii"
/db_xref="taxon:1638"
/strain="ATCC 19119"
RBS 95..100
/gene="sod"
terminator 723..746
/gene="sod"
CDS 109..717
/db_xref="SWISS-PROT:P28763"
/transl_table=11
/gene="sod"
/EC_number="1.15.1.1"
/product="superoxide dismutase"
/protein_id="CAA45406.1"
/translation="MTYELPKLPYTYDALEPNFDKETMEIHYTEKHHNIYVTKLNEAVSG
HAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLKAA
IESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNNGKLEIVSTANQDSPLSEGKTPVLGL
DVWEHAYYLKFNRRRPEYIDTFWNVINWDERNKRFDAAK"
BASE COUNT 247 a 136 c 151 g 222 t
ORIGIN
1 cggtatttaa ggtgttacaat agttctatgg aaatagggtc tatacctttc gccttacaat
61 gtaattttctt ttcacataaa taataaacaaa tccgaggagg aatttttaat gacttacgaa
121 ttaccaaaaat taccttatac ttatgatgct ttggagccga attttgataa agaaacaatg
181 gaaattcact atacaaaagca ccacaatatt tatgtaacaa aactaaatga agcagttctca
241 ggacacgcag aacttgcaag taaacctggg gaagaattag ttgctaactc agatagcggt
301 cctgaagaaa ttctgtggcgc agtacgtaac cacggtggtg gacatgctaa ccatacttta
361 ttctgtgcta gtcttagccc aaatggtggt ggtgctccaa ctggttaact aaaagcagca
421 atcgaaaagcg aattcggcac atttgatgaa ttcaaagaaa aattcaatgc ggcagctgcg
481 gctcgttttg gttcaggatg ggcatggcta gtagtgaaca atggtaaact agaaattggt
541 tccactgcta accaagattc tccacttagc gaaggtaaaa ctccagttct tggcttagat
601 gtttgggaac atgcttatta tcttaaatc caaaaccgct gtcctgaata cattgacaca
661 ttttgaatg taattaactg ggatgaacga aataaacgct ttgaccgagc aaaataatta
721 tcgaaaggct cacttaggtg ggtcttttta ttteta

//

EMBL format

```

ID LISOD standard; DNA; PRO; 756 BP.
XX
AC X64011; S78972;
XX
SV X64011.1
XX
DT 28-APR-1992 (Rel. 31, Created)
DT 30-JUN-1993 (Rel. 36, Last updated, Version 6)
XX
DE L.ivanovii sod gene for superoxide dismutase
XX
KW sod gene; superoxide dismutase.
XX
OS Listeria ivanovii
OC Bacteria; Firmicutes; Bacillus/Clostridium group;
OC Bacillus/Staphylococcus group; Listeria.
XX
RN [1]
RX MEDLINE; 92140371.
RA Haas A., Goebel W.;
RT "Cloning of a superoxide dismutase gene from Listeria ivanovii by
RT functional complementation in Escherichia coli and characterization of the
RT gene product.";
RL Mol. Gen. Genet. 231:313-322(1992).
XX
RN [2]
RP 1-756
RA Kreft J.;
RT ;
RL Submitted [21-APR-1992] to the EMBL/GenBank/DBJ databases.
RL J. Kreft, Institut f. Mikrobiologie, Universitaet Wuerzburg, Biozentrum Am
RL Hubland, 8700 Wuerzburg, FRG
XX
DR SWISS-PROT; P28763; SODM_LISIV.
XX
FH Key Location/Qualifiers
FH
FT source 1..756
FT /db_xref="taxon:1638"
FT /organism="Listeria ivanovii"
FT /strain="ATCC 19119"
FT RBS 95..100
FT /gene="sod"
FT terminator 723..746
FT /gene="sod"
FT CDS 109..717
FT /db_xref="SWISS-PROT:P28763"
FT /transl_table=11
FT /gene="Sod"
FT /EC number="1.15.1.1"
FT /product="superoxide dismutase"
FT /protein_id="CAA45406.1"
FT /translation="MTYELPKLPVYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVSG
FT HAEELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGAPTGNLKA
FT IESEFDTFDFEKFENAAAARFGSGMAWLVVNNKLEIVSTANQDSPLSEKTFVLGL
FT DVMEHAYLKFQNRPEYIDTFWVIVNWDERNKRFDAAK"
XX
SQ Sequence 756 BP; 247 A; 136 C; 151 G; 222 T; 0 other;
cggtatttaa ggtgttaacat agttctatgg aaatagggtc tataccttcc gccttacaat 60
gtaatttcct ttacacataa taataaacaa tccgaggagg aatttttaat gacttaacaa 120
ttaccaaaat taccttatac ttatgatgct ttggagccga attttgataa agaaaacaatg 180
gaaattcaat atacaaagca ccacaatatt tatgtaacaa aactaaatga agcagctctca 240
ggacacgcag aacttgcaag taaacctggg gaagaattag ttgctaactc agatagcgtt 300
cctgaagaaa ttctgtgggc agtacgtaac cacgggtggt gacatgctaa ccatacttta 360
ttctggtcta gtcttagacc aaatgggtgt ggtgctccaa ctgtaactt aaaaagcagca 420
atcgaagagc aattcggcac atttgatgaa tccaaagaaa aattcaatgc ggcagctgoc 480
gctctgtttg gttcaggatg gccatggcta qtagtgaaca atggtaaact agaaaattgtt 540
tccactgcta accaagattc tccacttagc gaaggtaaaa ctccagttct tggetttagt 600
gtttgggaac atgcttatta tcttaaatc caaaacogtc gtcctgaata cattgacaca 660
ttttggaatg taattaaact ggatgaaaga aataaacgct ttgaagcagc aaaaataatta 720
tcgaaaagct cacttaggtg ggtcttttta tttcta 756

```

//

Fasta Format

```
>gi|29848|emb|X61622.1|HSCDK2MR H.sapiens CDK2 mRNA
ATGGAGAACTTCCAAAAGGTGGAAAAGATCGGAGAGGGCACGTACGGAGTTGTGTACAAAGCCAGAAACA
AGTTGACGGGAGAGGTGGTGGCGCTTAAGAAAATCCGCCTGGACACTGAGACTGAGGGTGTGCCCAGTAC
TGCCATCCGAGAGATCTCTCTGCTTAAGGAGCTTAACCATCCTAATATTGTCAAGCTGCTGGATGTCATT
CACACAGAAAATAAACTCTACCTGGTTTTTTGAATTTCTGCACCAAGATCTCAAGAAATTCATGGATGCCT
CTGCTCTCACTGGCATTCTCTTCCCCTCATCAAGAGCTATCTGTTCCAGCTGCTCCAGGGCCTAGCTTT
CTGCCATTCTCATCGGGTCTCCACCGAGACCTTAAACCTCAGAACTGCTTATTAACACAGAGGGGGCC
ATCAAGCTAGCAGACTTTGGACTAGCCAGAGCTTTTGGAGTCCCTGTTTCGTACTTACACCCATGAGGTGG
TGACCCTGTGGTACCGAGCTCCTGAAATCCTCCTGGGCTCGAAATATTATTCCACAGCTGTGGACATCTG
GAGCCTGGGCTGCATCTTTGCTGAGATGGTGACTCGCCGGGCCCTGTTCCCTGGAGATTCTGAGATTGAC
CAGCTCTTCCGGATCTTTCGGACTCTGGGGACCCCAGATGAGGTGGTGTGGCCAGGAGTTACTTCTATGC
CTGATTACAAGCCAAGTTTTCCCAAGTGGGCCCCGGCAAGATTTTAGTAAAGTTGTACCTCCCCTGGATGA
AGATGGACGGAGCTTGTATCGCAAATGCTGCACTACGACCCTAACAAGCGGATTTTCGGCCAAGGCAGCC
CTGGCTCACCCCTTTCTTCCAGGATGTGACCAAGCCAGTACCCCATCTTCGACTCTGATAGCCTTCTTGAA
GCCCCGACCCTAATCGGCTCACCCCTCTCCTCCAGTGTGGGCTTGACCAGCTTGGCCTTGGGCTATTTGG
ACTCAGGTGGGCCCTCTGAACTTGCCTTAAACACTCACCTTCTAGTCTTAACCAGCCAACTCTGGGAATA
CAGGGGTGAAAGGGGGGAACCAAGTGAAAATGAAAGGAAGTTTCAGTATTAGATGCACTTAAGTTAGCCTC
CACCACCCTTTCCCCCTTCTCTTAGTTATTGCTGAAGAGGGTTGGTATAAAAATAATTTTAAAAAAGCCT
TCCTACACGTTAGATTTGCCGTACCAATCTCTGAATGCCCCATAATTATTATTTCCAGTGTTTGGGATGA
CCAGGATCCCAAGCCTCCTGCTGCCACAATGTTTATAAAGGCCAAATGATAGCGGGGGCTAAGTTGGTGC
TTTTGAGAATTAAGTAAAACAAAACCACTGGGAGGAGTCTATTTTAAAGAATTCGGTTAAAAAATAGATC
CAATCAGTTTATAACCCTAGTTAGTGTTTTCTCACCTAATAGGCTGGGAGACTGAAGACTCAGCCCGGGT
GGGGGT
```

Nucleic Acid Code	Meaning
A	Adenosine
C	Cytidine
G	Guanine
T	Thymidine
U	Uracil
R	G A (pu R ine)
Y	T C (p Y rimidine)
K	G T (K etone)
M	A C (a Mino group)
S	G C (S trong interaction)
W	A T (W eak interaction)
B	G T C (not A) (B comes after A)
D	G A T (not C) (D comes after C)
H	A C T (not G) (H comes after G)
V	G C A (not T, not U) (V comes after U)
N	A G C T (a Ny)
-	gap of indeterminate length

Amino Acid Code	Meaning
A	Alanine
B	Aspartic acid or Asparagine
C	Cysteine
D	Aspartate
E	Glutamate
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine
K	Lysine
L	Leucine
M	Methionine
N	Asparagine
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threonine
U	Selenocysteine
V	Valine
W	Tryptophan
Y	Tyrosine
Z	Glutamate or Glutamine
X	any
*	translation stop
-	gap of indeterminate length

Βάσεις δεδομένων πρωτεϊνικών ακολουθιών

- Οι βάσεις δεδομένων πρωτεϊνικών ακολουθιών, αποτελούν το δεύτερο μεγαλύτερο σε όγκο τμήμα του συνόλου των βιολογικών βάσεων δεδομένων (μετά τις ακολουθίες DNA), αλλά αποτελούν ίσως το σημαντικότερο τμήμα, καθώς οι πρωτεϊνικές ακολουθίες παρουσιάζουν μεγάλη ποικιλομορφία τόσο στη δομή όσο και στη λειτουργία. Κατά συνέπεια, μεγάλο μέρος της σύγχρονης βιοπληροφορικής ανάλυσης, αναφέρεται σε πρωτεϊνικές ακολουθίες και υπάρχει τεράστιος όγκος λειτουργικών δεδομένων που παράγονται συνεχώς πειραματικά, και τα οποία αποτελούν ή θα έπρεπε να αποτελούν μέρος της πληροφορίας που περιέχεται σε αυτές τις βάσεις.
- **UniprotKB** (Uniprot Knowledgebase <http://www.uniprot.org/>)

UniprotKB

- Η **UniprotKB** (Uniprot Knowledgebase <http://www.uniprot.org/>), αποτελεί την κύρια, σε παγκόσμιο επίπεδο βάση δεδομένων πρωτεϊνικών ακολουθιών ([Uniprot, 2013](#)).
- Αποτελείται από δύο υποσύνολα, την Uniprot/SwissProt η οποία περιέχει τις καλά σχολιασμένες πρωτεϊνικές ακολουθίες, και την Uniprot/TrEMBL η οποία περιέχει τις πρωτεϊνικές ακολουθίες που έχουν προκύψει από αυτόματη (ηλεκτρονική) μετάφραση γονιδιωματικών αλληλουχιών.
- Η UniprotKB/SwissProt περιέχει 547.599 ακολουθίες (Rel. 2015_02 – Φεβρουάριος 2015) οι οποίες έχουν περάσει από κάποιου είδους έλεγχο και συνοδεύονται από συμπληρωματικά σχόλια όπως, βιβλιογραφικές αναφορές, γενικά στοιχεία δευτεροταγούς δομής, σύνδεσμοι σε άλλες βάσεις δεδομένων σχετικές με κάθε εγγραφή καθώς και σημειώσεις για τη βιολογική λειτουργία (αν είναι γνωστές) καθώς και άλλες χρήσιμες πληροφορίες.
- Η Uniprot/TrEMBL περιέχει σήμερα (Rel. 2015_02 – Φεβρουάριος 2015) 92.124.243 ακολουθίες η οποίες όμως δεν έχουν υποστεί ανθρώπινο σχολιασμό. Περιοδικά, οι σχολιαστές της UniprotKB εντοπίζουν δεδομένα από τη βιβλιογραφία αλλά και με χρήση αυτοματοποιημένων εργαλείων, αλλάζουν το σχολιασμό των καταχωρήσεων και έτσι μια πρωτεϊνική αλληλουχία ενδέχεται να "περάσει" από την Uniprot/TrEMBL στην Uniprot/SwissProt. Το είδος, το εύρος και η μεγάλη ποικιλομορφία του σχολιασμού που μπορεί να υπάρχει σε επίπεδο πρωτεϊνικής αλληλουχίας είναι τεράστιο (σε ποιο κυτταρικό οργανίδιο υπάρχει, σε ποιον ιστό εκφράζεται, ποια είναι η δευτεροταγής δομή της, ποιος ο βιολογικός της ρόλος, ποια τα μονοπάτια στα οποία εμπλέκεται κ.ο.κ.), και κατά συνέπεια, ο όγκος της πληροφορίας στην Uniprot/SwissProt είναι τεράστιος, όπως επίσης και η πιθανότητα (παρόλες τις προσπάθειες), η πληροφορία αυτή να είναι λαθεμένη ή απλά ελλιπής..

Η ιστορία...

- Ιστορικά, αξίζει να αναφερθεί ότι η **Uniprot** προέκυψε το 2002 από μια συνένωση των δύο μεγαλύτερων τότε βάσεων δεδομένων, της SwissProt και της PIR.
- Η **SwissProt** Ιδρύθηκε το 1986 στο Ελβετικό Ινστιτούτο Βιοπληροφορικής (Swiss Institute of Bioinformatics) και λειτουργούσε σε συνεργασία με το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (European Bioinformatics Institute).
- Η **Protein Information Resource** (PIR - <http://pir.georgetown.edu/>) ήταν η αντίστοιχη Αμερικάνικη βάση δεδομένων. Η έδρα της ήταν στο Πανεπιστήμιο του Georgetown και αποτελούσε τμήμα του Εθνικού Ιδρύματος Βιοϊατρικής Έρευνας (NBRF) των Η.Π.Α. Η κυριότερη βάση που περιέχει είναι η PIR-International Protein Sequence Database (PSD), της οποίας τα δεδομένα προκύπτουν από την συνεργασία της PIR με το Munich Information Center for Protein Sequences (MIPS) και την Japanese International Protein Information Database (JIPID).
- Το 2002, η PIR σε μια κοινή προσπάθεια με το EBI (European Bioinformatics Institute) και το SIB (Swiss Institute of Bioinformatics) σχημάτισαν το UniProt consortium. Με αυτόν τον τρόπο οι ακολουθίες της PIR-PSD αλλά και ο σχολιασμός τους ενσωματώθηκαν στην UniProt Knowledgebase. Προστέθηκαν διασυνδέσεις μεταξύ των καταχωρήσεων της UniProt και της PIR-PSD για να διευκολυνθεί ο εντοπισμός παλαιών καταχωρήσεων της PIR-PSD. Πρωτεΐνες που ήταν μοναδικές στην PIR-PSD όπως και οι αναφορές τους αλλά και τα πειραματικά δεδομένα που υπήρχαν στις σχετικές καταχωρήσεις μπορούν πλέον να βρεθούν στις αντίστοιχες καταχωρήσεις της UniProt.

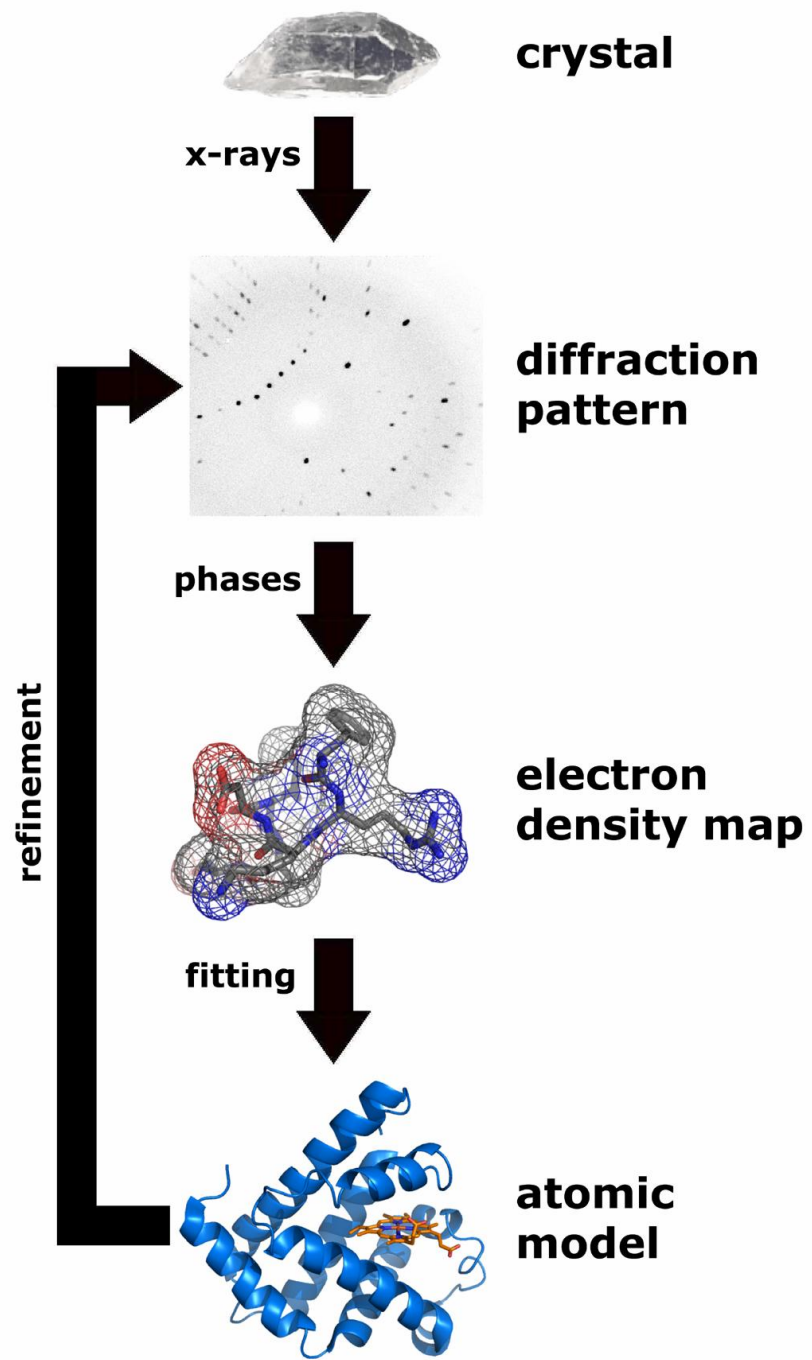
PIR/NBRF Format

```
>P1;CRAB_ANAPL ALPHA CRYSTALLIN B CHAIN (ALPHA(B) -  
CRYSTALLIN) .
```

```
MDITIHNPLI RRPLFSWLAP SRIFDQIFGE HLQESELLPA  
SPSLSPFLMR SPIFRMPSWL ETGLSEMRLE KDKFSVNLDV  
KHFSPPEELKV KVLGDMVEIH GKHEERQDEH GFIAREFNRK  
YRIPADVDPL TITSSLSLDG VLTVSAPRKQ SDVPERSIPI  
TREEKPAIAG AQRK*
```


Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών

- Οι βάσεις αυτές περιέχουν δεδομένα που έχουν να κάνουν με την τρισδιάστατη δομή βιολογικών μακρομορίων.
- Οι τρισδιάστατες δομές αποτελούν το τελικό στάδιο μιας επίπονης διαδικασίας η οποία μετά τη χρήση μοριακών τεχνικών (κλωνοποίηση, απομόνωση, κρυστάλλωση κ.ο.κ.), οδηγεί τελικά στην υπολογιστική επίλυση της δομής μέσω της διαδικασίας της κρυσταλλογραφίας ακτίνων Χ, ή, σε πιο σπάνιες περιπτώσεις με φασματογραφία NMR.
- Το μεγαλύτερο ενδιαφέρον, βέβαια, έχουν οι δομές πρωτεϊνών, καθώς οι πρωτεΐνες είναι τα μακρομόρια των οποίων η μεγάλη ποικιλομορφία της δομής συνδέεται άμεσα με την βιολογική δράση. Η μοναδική βάση αυτού το είδους παγκοσμίως, είναι η PDB, η οποία και αναλύεται παρακάτω.
- Protein Data Bank (PDB - www.rcsb.org)



Protein Data Bank

- Η Protein Data Bank (PDB - www.rcsb.org) είναι παγκοσμίως η μοναδική βάση στην οποία περιέχονται τρισδιάστατες δομές βιολογικών μακρομορίων ([Kouranov, et al., 2006](#)).
- Ιδρύθηκε το 1971 στα εργαστήρια Brookhaven National Laboratories (BNL) των ΗΠΑ. Αρχικά αποτελούνταν από 7 δομές μακρομορίων οι οποίες προέκυψαν από κρυσταλλογραφικές μελέτες ενώ είχε μικρό ρυθμό αύξησης εγγραφών μέχρι τα τέλη της δεκαετίας του '70.
- Την δεκαετία του '80 παρατηρήθηκε σημαντική αύξηση του ρυθμού προσθήκης δεδομένων λόγω της τεχνολογικής εξέλιξης σε κάθε στάδιο του προσδιορισμού των δομών, ενώ πλέον η PDB περιέχει και δομές που έχουν προκύψει με φασματοσκοπία Πυρηνικού Μαγνητικού Συντονισμού (NMR).
- Σήμερα (Φεβρουάριος 2015) η PDB περιλαμβάνει 106.858 δομές βιομορίων.
- Οι εγγραφές της PDB περιλαμβάνουν εκτός από τις συντεταγμένες των ατόμων που απαρτίζουν τη δομή και επιπρόσθετα βοηθητικά στοιχεία όπως βιβλιογραφικές αναφορές, λεπτομέρειες για τον προσδιορισμό της δομής καθώς και άλλα στοιχεία που προκύπτουν από τη συγκεκριμένη δομή.
- Κάθε δομή πριν δημοσιευθεί στην βάση ελέγχεται για την ορθότητα της με τη χρήση ειδικού λογισμικού. Στη συνέχεια εφόσον περάσει τις δοκιμές με επιτυχία αποκτά ένα χαρακτηριστικό κωδικό και προστίθεται στη βάση.

Protein Data Bank

- Πρέπει να τονιστεί, ότι η καταχώρηση στην PDB είναι η τρισδιάστατη δομή, και όχι η πρωτεΐνη.
- Κατά συνέπεια, είναι δυνατόν να υπάρχει μια καταχώρηση της PDB η οποία να περιέχει περισσότερες από μία (ακόμα και μερικές δεκάδες) πρωτεϊνικές ακολουθίες, όπως για παράδειγμα όταν αναφερόμαστε σε πολυενζυμικά σύμπλοκα τα οποία περιέχουν πολλές υπομονάδες.
- Επίσης, είναι δυνατόν να υπάρχουν περισσότερες από μία δομές μιας συγκεκριμένης πρωτεΐνης, καθώς είναι δυνατόν να έχουν γίνει διαφορετικά πειράματα είτε σε διαφορετικές συνθήκες, είτε παρουσία άλλων παραγόντων, είτε και απλά με άλλη τεχνική για να επιτευχθεί καλύτερη ανάλυση.
- Φυσικά, όπως είναι αναμενόμενο, μόνο ένα μικρό υποσύνολο των γνωστών πρωτεϊνών έχουν γνωστή τρισδιάστατη δομή, γιατί η διαδικασία επίλυσης της δομής είναι χρονοβόρα και δύσκολη. Αυτό φαίνεται ξεκάθαρα αν συγκρίνουμε τον αριθμό των καταχωρήσεων της UniProt με αυτόν της PDB.
- Ειδικότερα δε, για κάποιες ειδικές κατηγορίες πρωτεϊνών όπως οι διαμεμβρανικές πρωτεΐνες, τα πράγματα είναι ακόμα πιο δύσκολα από πειραματικής πλευράς και οι τρισδιάστατες δομές τους, είναι ακόμα πιο σπάνιες.
- Τέλος, αξίζει να αναφερθεί, ότι παρόμοια βάση (MMDB) συντηρείται και στις ΗΠΑ στα πλαίσια του NCBI, με συνεχή όμως επαφή και ενημέρωση από την PDB.

- Home
- Tutorial About This Site
- Getting Started
- Download Files
- Deposit and Validate
- Structural Genomics
- Dictionaries & File Formats
- Software Tools
- Educational Resources
- BioSync
- General Information
- Acknowledgements
- Frequently Asked Questions
- Known Problems
- Report Bugs/Comments

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the [wwPDB](#) whose mission is to ensure that the PDB archive remains an international resource with uniform data.

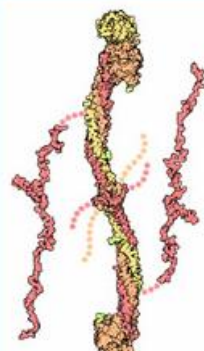
This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found [here](#).

A [narrated tutorial](#) illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the [Macromedia Flash player download](#).]

Comments? info@rcsb.org

Molecule of the Month: Fibrin



When you cut yourself, you bleed, but the bleeding rapidly stops. Blood has a built-in emergency repair system that quickly blocks any damage to the circulatory system, creating a temporary patch that allows time for more permanent repairs. Three basic mechanisms are at work. First, platelets (small fragments of blood cells that circulate in the blood) clump at the site of the wound, forming a weak plug. Second, neighboring blood vessels constrict, reducing the amount of blood flowing into the area. Finally, the protein fibrin assembles into a tough network that clots the blood and forms an insoluble blockage. Together, these methods stop the loss of blood and create a sturdy scab to protect the area as you heal.

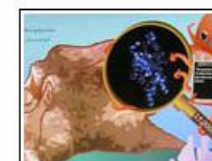
More ...

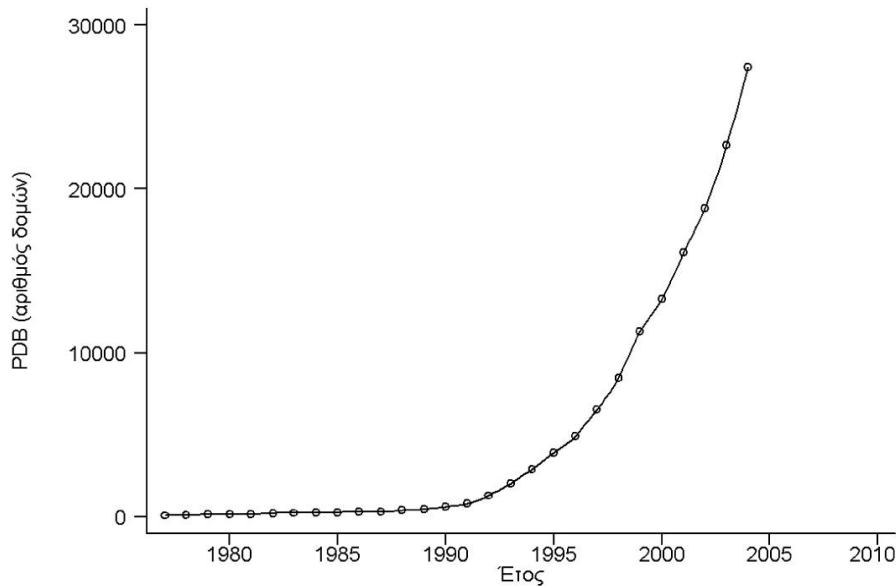
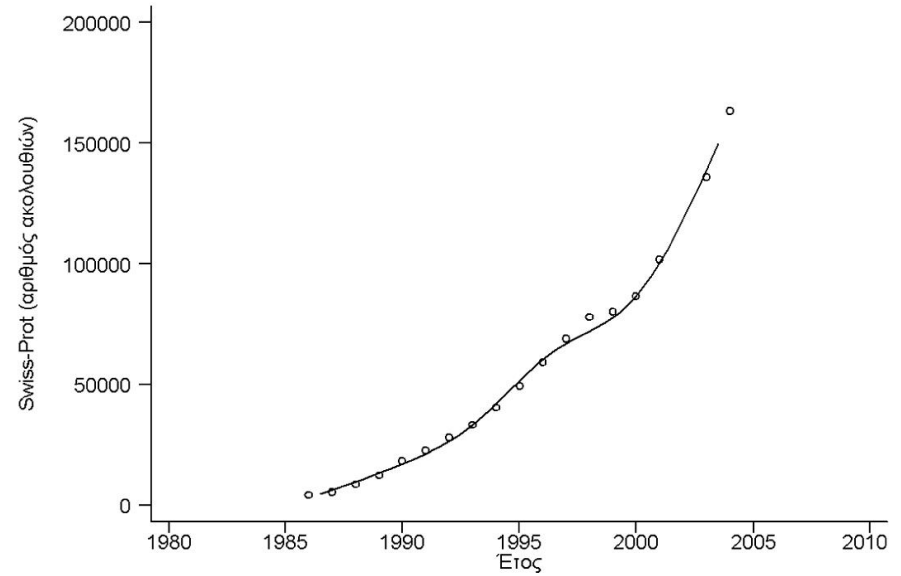
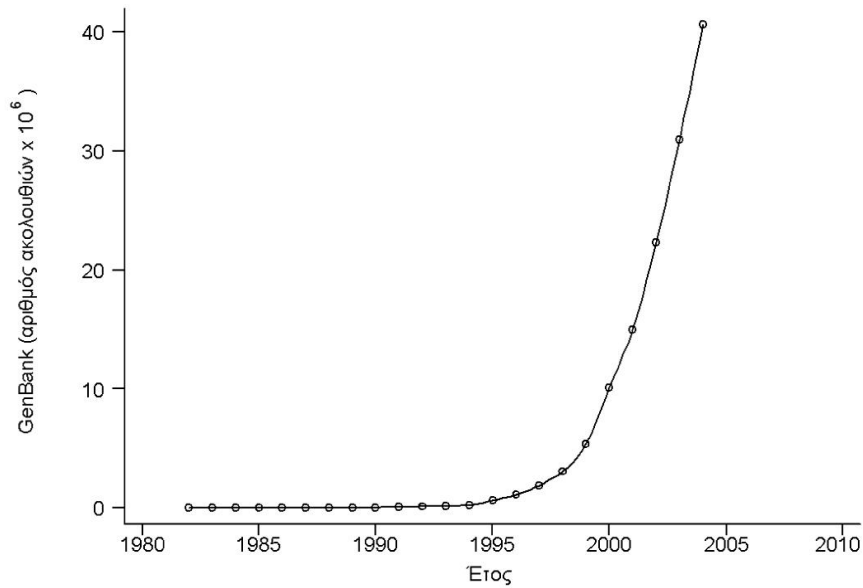
NEWS

- Complete News
- Newsletter
- Discussion Forum

31-October-2006 PDB Structures on Exhibit at the Birch Aquarium

The *Sea of Genes* exhibit at the Birch Aquarium (La Jolla, CA) helps to unravel the genetic secrets of life in the ocean through interactive displays that highlight the exciting discoveries of Scripps researchers. One feature of this exhibition is an interactive kiosk that invites the user to display information about specific proteins found in marine organisms — and in the PDB.





GenBank, SwissProt, PDB

Κοινό χαρακτηριστικό:

Η Εκθετική Αυξηση
(διπλασιασμός κάθε 1,5
– 2 χρόνια)

Βάσεις δεδομένων γονιδιακής έκφρασης

- Εκτός από τις βάσεις δεδομένων ακολουθιών και δομών, σημαντική είναι τα τελευταία χρόνια και η ανάπτυξη των βάσεων δεδομένων γονιδιακής έκφρασης.
- Με την εξέλιξη της τεχνολογίας και τη δημιουργία νέων οικονομικότερων τσιπ μικροσυστοιχιών, αλλά και με την εμφάνιση των τεχνολογιών Next Generation Sequencing, τα πειράματα ανάλυσης γονιδιακής έκφρασης πραγματοποιούνται με μεγαλύτερο ρυθμό και έτσι υπάρχει ανάγκη αποθήκευσης και ανάλυσης όλων αυτών των δεδομένων.
- Τη λύση στο παραπάνω πρόβλημα έδωσαν οι βάσεις δεδομένων οι οποίες περιέχουν δεδομένα από χιλιάδες πειράματα μικροσυστοιχιών. Οι βάσεις δεδομένων αυτές επιτρέπουν την καταχώρηση αποτελεσμάτων από πειράματα μικροσυστοιχιών, ενώ κάποιες από αυτές προσφέρουν και επιπλέον εργαλεία ανάλυσης. Επίσης, παρέχουν πληροφορίες σχετικά με το είδος των δεδομένων, την πλατφόρμα μικροσυστοιχιών που χρησιμοποιήθηκε στο πείραμα, τα γονίδια τα οποία μελετώνται καθώς επίσης και πληροφορίες σχετικά με τα είδη των δειγμάτων τα οποία χρησιμοποιήθηκαν.
- Η βασική δομή αυτών των αρχείων, διαφέρει πολύ από αυτά που αναφέραμε μέχρι τώρα, καθώς έχουμε να κάνουμε με έναν πίνακα, στον οποίο αναγράφονται τιμές "έκφρασης" ενός γονιδίου για κάθε άτομο. Συνήθως τα πειράματα αυτά αφορούν λίγα άτομα, αλλά ανάλογα με την πλατφόρμα μπορούμε να έχουμε δεδομένα έκφρασης για μερικές εκατοντάδες έως μερικές δεκάδες χιλιάδες γονίδια.
- Επειδή ο όγκος των δεδομένων γονιδιακής έκφρασης είναι μεγάλος και πολύπλοκος, για να καταχωρηθούν τα δεδομένα των μικροσυστοιχιών στις δημόσιες βάσεις δεδομένων θα πρέπει να ακολουθούν ένα συγκεκριμένο πρωτόκολλο με βάση το οποίο καταχωρείται η ελάχιστη πληροφορία που περιγράφει ένα πείραμα μικροσυστοιχιών (MIAME: Minimum Information About a Microarray Experiment).

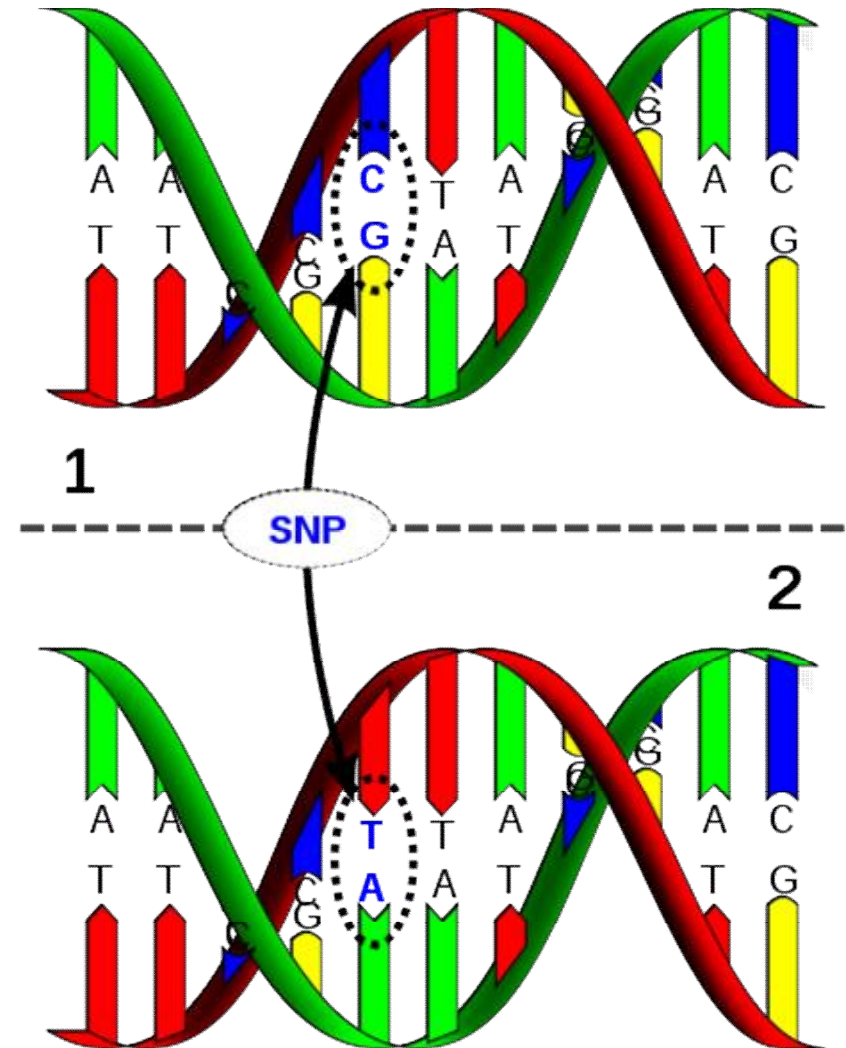
Βάσεις δεδομένων γονιδιακής έκφρασης

- **GeneExpression Omnibus (GEO):** Βάση δεδομένων του NCBI που παρέχει δεδομένα γονιδιακής έκφρασης, τόσο από μικροσυστοιχίες όσο και από αλληλούχιση (next generation sequencing). Είναι διαθέσιμη στην ιστοσελίδα <http://www.ncbi.nlm.nih.gov/geo/> ενώ στην ίδια διεύθυνση υπάρχουν διαθέσιμα και κάποια διαδικτυακά εργαλεία που επιτρέπουν απλές αναλύσεις των δεδομένων της βάσης. Τα δεδομένα υπάρχουν τόσο σε ακατέργαστη (raw) όσο και σε επεξεργασμένη μορφή (με κανονικοποιήσεις κ.ο.κ.).
- **Array Express:** Δημόσια βάση δεδομένων μικροσυστοιχιών η οποία διατηρείται στο Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής, EBI, διαθέσιμη στην ιστοσελίδα <http://www.ebi.ac.uk/arrayexpress/>. Είναι της ίδιας λογικής με την GEO, την οποία περιέχει ως υποσύνολο βάσει της συνεργασίας των ιδρυμάτων. Στην ιστοσελίδα υπάρχουν επίσης διαθέσιμα εργαλεία για ανάλυση, οδηγίες για προγραμματιστική πρόσβαση στις υπηρεσίες και tutorials).
- **Stanford Microarray Database (SMD):** Βάση δεδομένων που κατασκευάστηκε αρχικά για να καλύπτει τις ανάγκες διαμοιρασμού αρχείων των ερευνητών του Stanford, αλλά μετεξελίχθηκε σταδιακά σε ένα δημόσιο αποθετήριο δεδομένων για μικροσυστοιχίες. <http://smd.stanford.edu>

ID_REF	GSM480382	GSM480383	GSM480384	GSM480385	GSM480386	GSM480387	GSM480388
1007_s_at	51,62	33,44	29,82	30,08	21,33	37,61	36,71
1053_at	103,47	108,09	162,45	70,62	81,96	73,55	56,09
117_at	1666,73	1159,20	1019,50	671,72	2504,60	1049,85	372,75
121_at	19,63	18,63	25,06	42,95	16,99	12,24	17,76
1255_g_at	4,23	4,20	6,02	5,30	4,53	3,22	7,26
1294_at	185,64	169,91	289,11	159,58	278,09	422,08	446,76
1316_at	26,70	14,56	25,29	70,09	37,19	34,78	23,34
1320_at	12,93	14,02	15,93	17,07	14,14	11,09	9,29
1405_i_at	1415,67	889,13	2333,74	640,93	806,76	1440,97	3351,53
1431_at	5,56	8,26	5,46	6,32	5,35	4,84	8,32
1438_at	9,05	9,69	12,97	8,88	7,52	6,86	8,17
1487_at	121,30	120,59	138,16	99,10	76,56	77,27	85,89
1494_f_at	16,05	14,05	13,39	15,29	11,47	12,21	8,40
1552256_a_at	10,61	11,64	16,11	32,43	16,99	10,78	8,02
1552257_a_at	47,15	61,56	38,35	66,74	41,12	63,73	23,72
1552258_at	16,44	13,67	27,31	28,03	31,03	19,75	17,82
1552261_at	4,74	6,97	7,47	6,17	5,17	3,72	4,56
1552263_at	294,75	310,94	327,27	62,27	325,14	169,60	277,95
1552264_a_at	597,81	538,79	695,00	236,29	988,23	371,01	396,51
1552266_at	5,08	5,28	4,84	9,10	6,02	3,44	4,87
1552269_at	3,09	3,20	4,11	3,16	3,36	3,29	2,82

Βάσεις δεδομένων γενετικής ποικιλομορφίας

- Οι βάσεις αυτές, αν και συνδέονται στενά με τις βάσεις δεδομένων ακολουθιών DNA, δεν αποτελούν ευθέως παράγωγα τους, αλλά μάλλον ανεξάρτητες οντότητες.
- Τούτο είναι κατανοητό αν σκεφτούμε ότι σε μια δεδομένη θέση ενός γονιδιώματος ενός είδους (πχ του ανθρώπου), τα διαφορετικά άτομα είναι δυνατόν να έχουν διαφορετική γενετική πληροφορία (πχ A αντί για T, κ.ο.κ.).
- Η βάση η οποία καταγράφει τους πολυμορφισμούς και τις συχνότητες τους στους διάφορους πληθυσμούς είναι η dbSNP, ενώ η βάση που καταγράφει πρωτογενώς τουλάχιστον τις αλληλοσυσχετίσεις των πολυμορφισμών αυτών, είναι η HarMap.



Βάσεις δεδομένων γενετικής ποικιλομορφίας

- **dbSNP:** Η dbSNP είναι η δημόσια βάση για τους νουκλεοτιδικούς πολυμορφισμούς <http://www.ncbi.nlm.nih.gov/snp> Εκτός από νουκλεοτιδικούς πολυμορφισμούς (single nucleotide polymorphisms - SNPs), περιέχει και δεδομένα για πολυμορφικές θέσεις που αφορούν απαλοιφές ή εισαγωγές βάσεων (deletion insertion polymorphisms -DIPs), καθώς και για ένθετα μεταθετά στοιχεία και μικροδορυφορικές επαναλήψεις (short tandem repeats - STRs). Κάθε καταχώρηση στην dbSNP περιέχει πληροφορίες για το που βρίσκεται ο πολυμορφισμός (δηλαδή την περιβάλλουσα αλληλουχία), τη συχνότητα του πολυμορφισμού σε διάφορους πληθυσμούς, αλλά και για την πειραματική μέθοδο, τα πρωτόκολλα και τις συνθήκες με τις οποίες μετρήθηκε η ποικιλομορφία. Η dbSNP δέχεται επίσης υποβολές για καταχωρήσεις πολυμορφισμών από κάθε είδος, αλλά και από διαφορετικά σημεία του γονιδιώματος. Λεπτομερής περιγραφή της βάσης δεδομένων υπάρχει στο ελεύθερο διαδικτυακό βιβλίο του NCBI στη διεύθυνση <http://www.ncbi.nlm.nih.gov/books/NBK3848/>.
- **HapMap:** Το International HapMap Project (<http://hapmap.ncbi.nlm.nih.gov/>) είναι το αποτέλεσμα μια διεθνούς συνεργασίας σε μια προσπάθεια να εντοπισθούν και να καταγραφούν οι γενετικές διαφορές αλλά και οι ομοιότητες των ανθρώπινων πληθυσμών. Ο σκοπός του προγράμματος είναι να συγκρίνει τις γενετικές αλληλουχίες διαφορετικών ατόμων (από διαφορετικούς πληθυσμούς) και να εντοπίσει με αυτόν τον τρόπο χρωμοσωμικές περιοχές στις οποίες οι γενετικές παραλλαγές (συνήθως, νουκλεοτιδικοί πολυμορφισμοί), κληρονομούνται μαζί. Στην αρχική φάση του προγράμματος, έγινε χρήση γενετικών δεδομένων από 4 πληθυσμούς Αφρικανικής, Ασιατικής και Ευρωπαϊκής καταγωγής. Σε μεταγενέστερες εκδόσεις, προστέθηκαν και άλλοι πληθυσμοί, σε μια προσπάθεια να υπάρχει όσο το δυνατό μεγαλύτερη κάλυψη παγκοσμίως. Τα τελικά δεδομένα που είναι διαθέσιμα από τη βάση αυτή, είναι οι απλότυποι, δηλαδή οι συνδυασμοί πολυμορφισμών που συνκληρονομούνται, και ακριβέστερα οι συντελεστές ανισορροπίας σύνδεσης (Linkage Disequilibrium), των διαφόρων πολυμορφισμών του ίδιου χρωμοσώματος, μεταξύ τους. Με τη χρήση αυτής της πληροφορίας, είναι δυνατόν να σχεδιαστούν μέθοδοι και αλγόριθμοι στατιστικής γενετικής με τους οποίους θα επιχειρείται να απαντηθούν ερωτήματα σχετικά με τη γενετική προδιάθεση σε ασθένειες και την ανταπόκριση σε φάρμακα. Επιπλέον, τέτοια δεδομένα είναι πολύ χρήσιμα στη μελέτη της γενετικής δομής των ανθρώπινων πληθυσμών

Βάσεις δεδομένων βιβλιογραφίας

- Παρόλο που οι βάσεις αυτές δεν είναι με την στενή έννοια «βιολογικές βάσεις δεδομένων», ιστορικά, αλλά και για λόγους που θα φανούν στην πορεία, είναι καλό να γίνεται αναφορά και σε αυτές. Οι βάσεις αυτές, έχουν σαν «καταχώρηση» τα στοιχεία μιας επιστημονικής δημοσίευσης (συγγραφέας, περιοδικό, περίληψη κ.ο.κ.). Η κυριότερη βάση του είδους, είναι η **PubMed** (<http://www.ncbi.nlm.nih.gov/pubmed>) η οποία στεγάζεται στο NCBI και περιλαμβάνει περισσότερα από 24 εκατομύρια καταχωρήσεις επιστημονικών άρθρων από τη βιοϊατρική βιβλιογραφία (έχοντας κάλυψη της MEDLINE, άλλων περιοδικών των επιστημών της ζωής αλλά και από κάποια online βιβλία). Οι αναφορές μπορεί να περιέχουν συνδέσμους στο πλήρες κείμενο των εργασιών, είτε μέσω της PubMed Central (το υποσύνολο με τις ελεύθερα διαθέσιμα δημοσιεύσεις πλήρους κειμένου), είτε απευθείας μέσω των ιστοσελίδων των εκδοτικών οίκων. Παρόλο που τα στοιχεία της PubMed είναι δημόσια διαθέσιμα, το να έχει πρόσβαση κανείς στο πλήρες κείμενο μιας εργασίας, εξαρτάται από την πολιτική του εκδοτικού οίκου. Στον ίδια ιστοσελίδα, υπάρχουν διαθέσιμα και tutorials για τη χρήση της υπηρεσίας (<http://www.nlm.nih.gov/bsd/disted/pubmed.html>).
- Άλλες βάσεις δεδομένων, παρόμοιας φύσης, είναι το **SCOPUS** (<http://www.scopus.com/>) και το **Web of Science** (<http://webofknowledge.com/>).



A service of the National Library of Medicine
and the National Institutes of Health

All Databases PubMed Nucleotide Protein Genome Structure

Search PubMed for Go Clear

Limits Preview/Index History Clipboard Details

About Entrez
NCBI Toolbar

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation
Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI

- ◆ To get started, enter one or more search terms.
- ◆ Search terms may be [topics](#), [authors](#) or [journals](#).

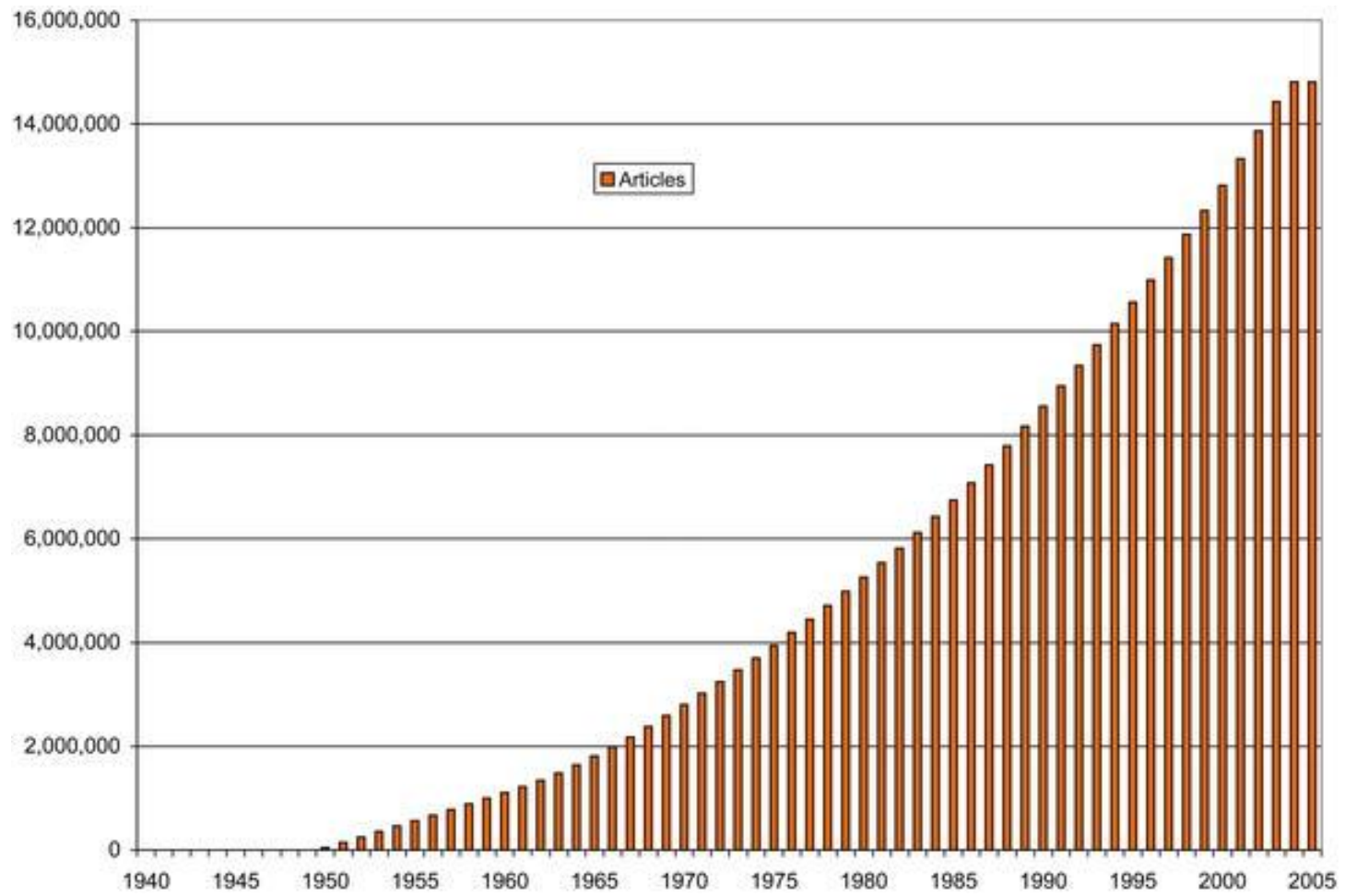


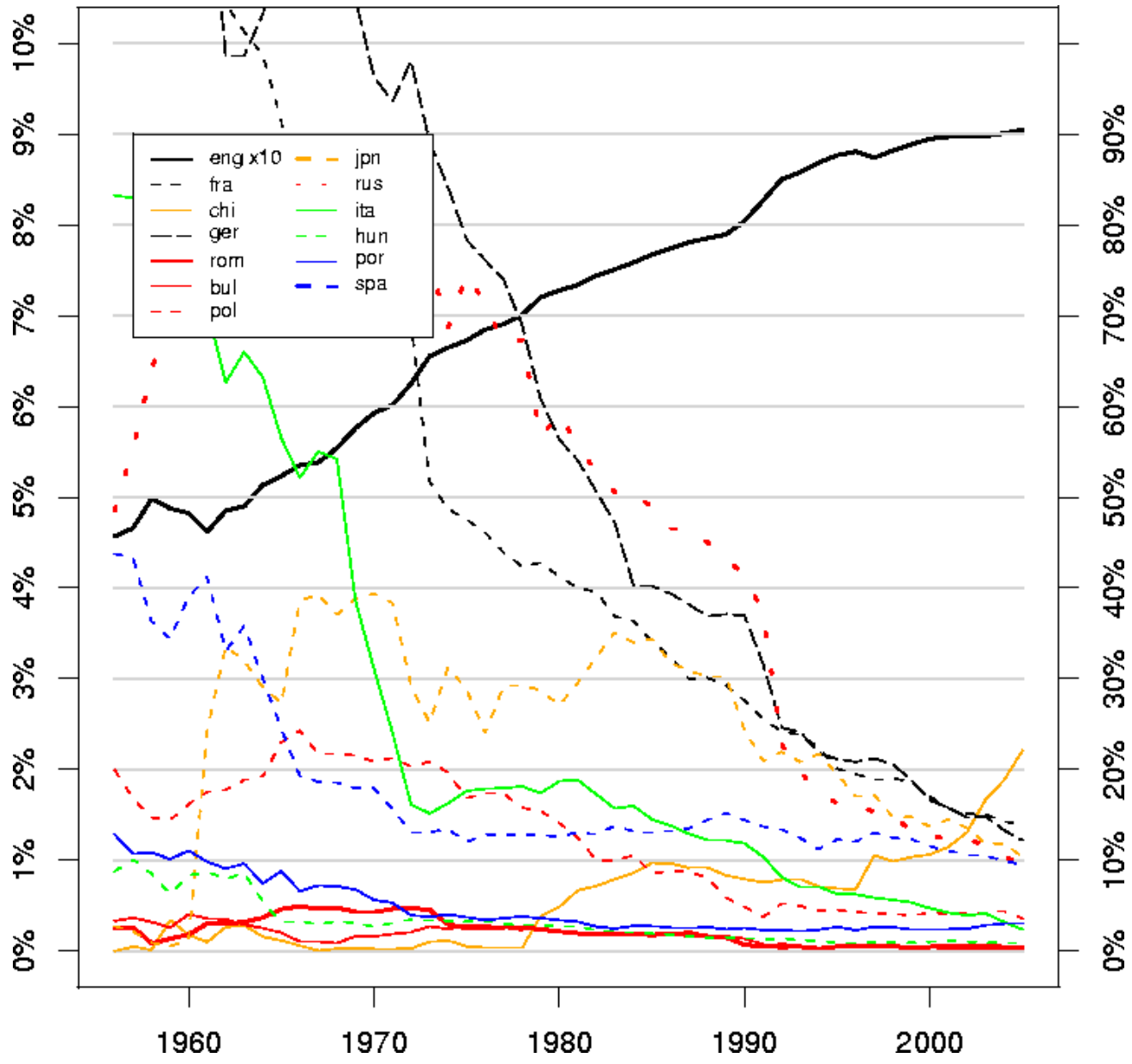
Need to find a specific citation or journal issue?

- (1) Click [Single Citation Matcher](#) on the PubMed sidebar.
- (2) Enter what you know and click Go.

The journal and author boxes include an autocomplete feature that suggests titles and names as you type. Read the [PubMed Help](#) for additional information.

PubMed is a service of the [U.S. National Library of Medicine](#) that includes over 16 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources.





PubMed Job results

Abstracts for Search ID 27.
Possible connection CAMK2A to CAMK2A (A0011 - A0011)

A0011


(*Calcium/calmodulin-dependent protein kinase type II alpha chain(EC 2.7.1.123) (CaM-kinase II alpha chain) (CaM kinase II alphasubunit) (CaMK-II alpha subunit).* CAMKA CAMK2A
Calcium/calmodulin-dependent protein kinase type II alpha chain KIAA0968 *CaMKII alpha*)

A0011

(*Calcium/calmodulin-dependent protein kinase type II alpha chain(EC 2.7.1.123) (CaM-kinase II alpha chain) (CaM kinase II alphasubunit) (CaMK-II alpha subunit).* CAMKA CAMK2A
Calcium/calmodulin-dependent protein kinase type II alpha chain KIAA0968 *CaMKII alpha*)

Count Select Score PubmedId

Abstract

1  0.027 [7931307](#) [Rapid translocation of cytosolic Ca2+/calmodulin-dependent protein kinase II into postsynaptic density after decapitation.](#)

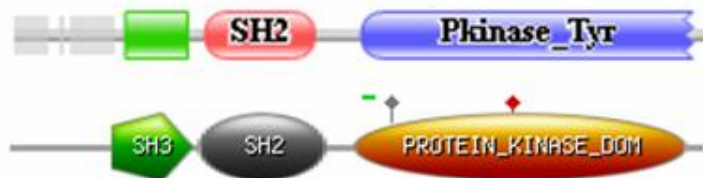
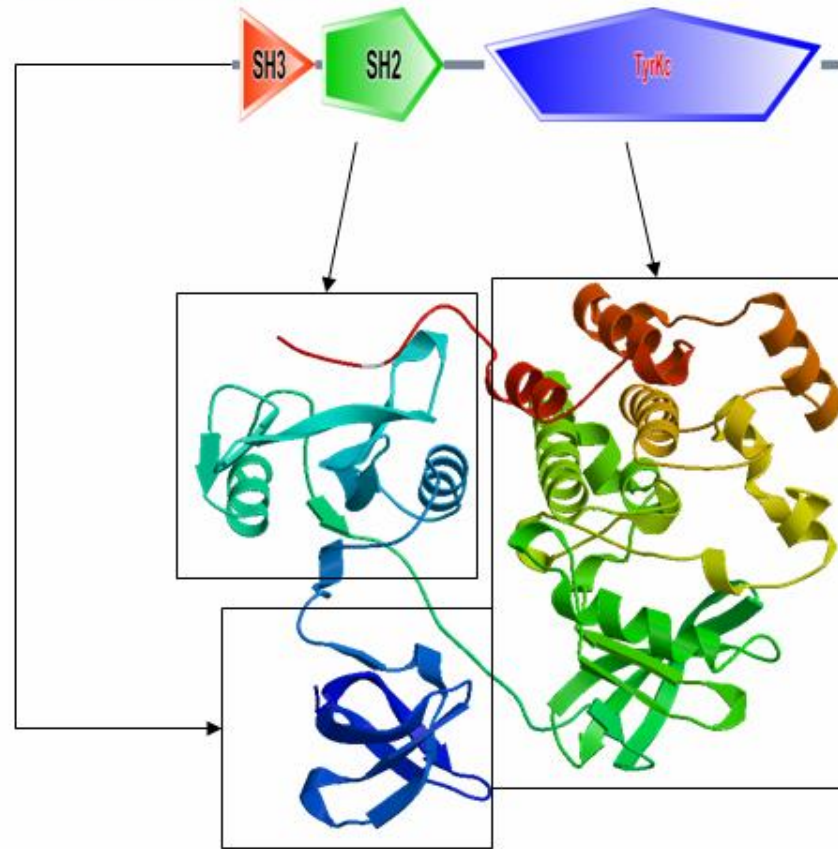
The postsynaptic density (PSD) fraction prepared from rat forebrains frozen with liquid nitrogen immediately after dissection (within 30 s after decapitation) contained major postsynaptic density protein (mPSDp), alpha subunit of Ca2+/calmodulin-dependent protein kinase II (CaMKII) at a level of merely 2.7% of the total protein. The content of the protein in the fraction was increased to approximately 10% by placing the forebrains on ice for a few minutes. Accumulation, but to a lesser extent, of the protein after placement was also observed in the particulate, synaptosome, and synaptic plasma membrane fractions with its concomitant decrease in the cytosolic fraction. The distribution change may be translocation of the protein, because the amounts of the losses of the protein in the cytosolic fraction were balanced by the gains in the particulate fractions. By translocation, CaMKII became Triton X-100 insoluble and partially inactivated. The amount of CaMKII transferred from the cytosol to particulate fractions at 0 degrees C was about the same as that contained in the conventional PSD fraction. Furthermore, the thickness of the PSD was increased by the treatment of the forebrains at 37 degrees C, by which the content of CaMKII alpha in the PSD fraction was increased to twofold. These results suggest that most of the CaMKII alpha subunit associated with the PSD fraction (mPSDp) is translocated from cytosol after decapitation. We also showed similar translocation of CaMKII beta/beta'.

Δευτερογενείς βάσεις δεδομένων

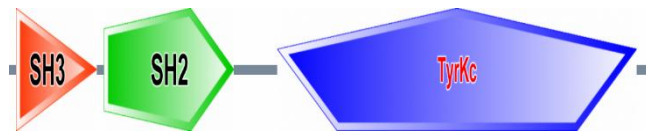
- **Βάσεις δεδομένων οικογενειών**
- Όπως είναι γνωστό, οι πρωτεΐνες γενικά αποτελούνται από μία ή περισσότερες διακριτές λειτουργικές περιοχές (domains), οι οποίες πολλές φορές είναι και δομικά αυτοτελής.
- Οι περιοχές αυτές, θεωρείται ότι μπορούν να λειτουργήσουν αλλά και να εξελιχθούν ανεξάρτητα από το υπόλοιπο τμήμα της πρωτεΐνης. Διαφορετικοί συνδυασμοί τέτοιων περιοχών οδηγούν σε μια μεγάλη ποικιλία των πρωτεϊνών στη φύση.
- Συνεπώς, η ανίχνευση τέτοιων περιοχών είναι σημαντική στην προσπάθεια λειτουργικής ταξινόμησης των πρωτεϊνών.

```

>sp|P08631|HCK_HUMAN
MGGRSSCEDPGCPRDEERAPRMGCMKSKFLQVGGNTFSKTETSASPHCPVYVPDPTSTIK
PGPNSHNSNTPGIREAGSEDIIVVALYDYEAIHHEDLSFQKGDQMVVLEESGEWVKARSL
ATRKEGYIPSNYVARVDSLETEEWFKGISRKDAERQLLAPGNMLGSFMIRDSETTKGSY
SLSVRDYDPRQGDTVKHYKIRTLDNNGFYISPRSTFSTLQELVDHYKKGNDGLCQKLSVP
CMSSKPQKPWEKDAWEIPRESLKLEKKGAGQFGEVWMATYNKHTKVAVKTMKPGSMSVE
AFLAEANVMKTLQHDKLVKLHAVVTKEPIYIITEFMAKGSLLDFLKSDEGSKQPLPKLID
FSAQIAEGMAFIEQRNYIHRDLRAANILVASLVCKIADFGLARVIEDNEYTAREGAKFP
IKWTAPEAINFGSPTIKSDVWSFGILLMEIVTYGRIPYPGMSNPEVIRALERGYRMPRPE
NCPEELYNIMRCWKNRPEERPTFEYIQSVLDDFYTATES
  
```



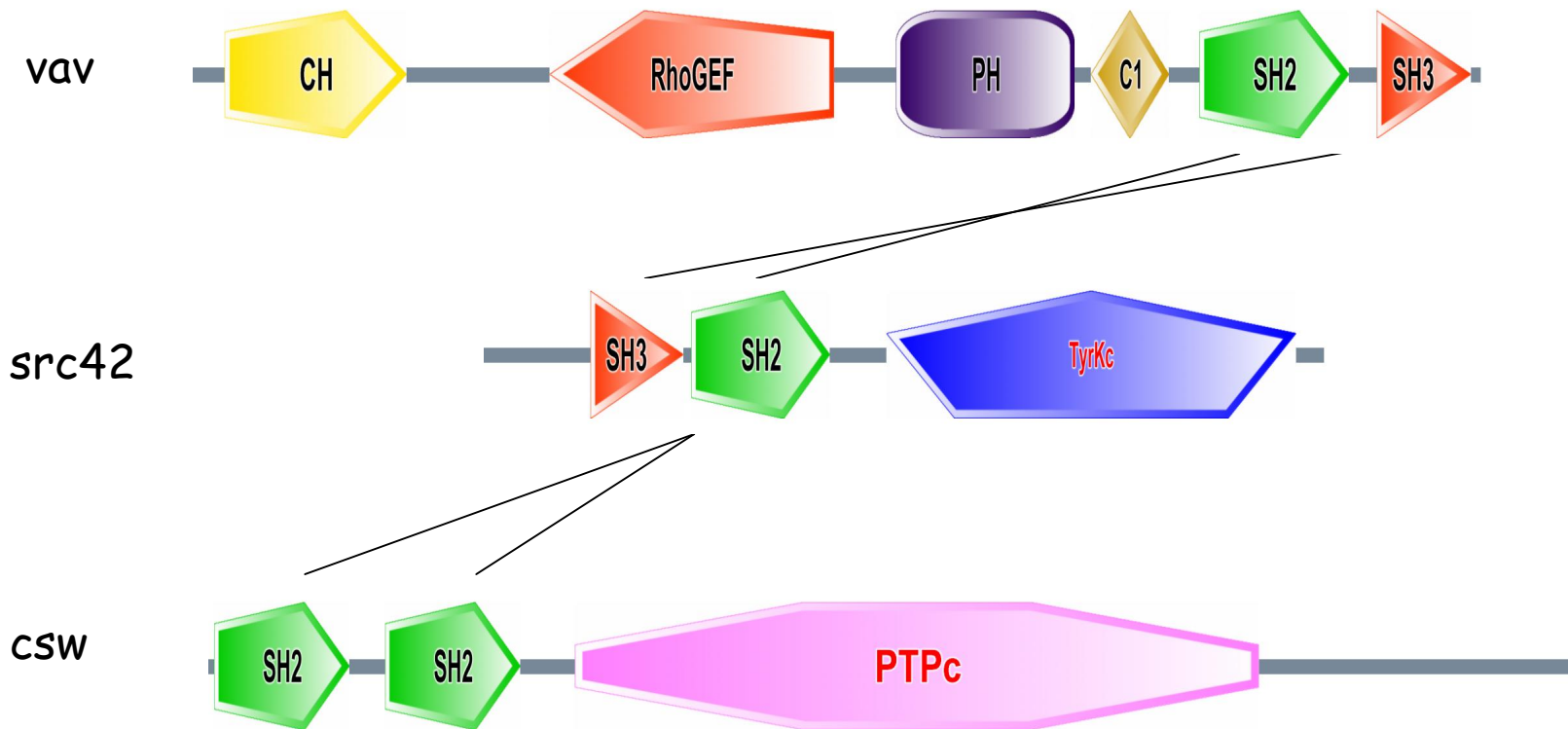
Protein Families



src-like protein tyrosine kinase - 5 in *Drosophila* proteome

38 tyrosine kinases
43 SH2 domain containing
110 SH3 domain containing

Local Similarity



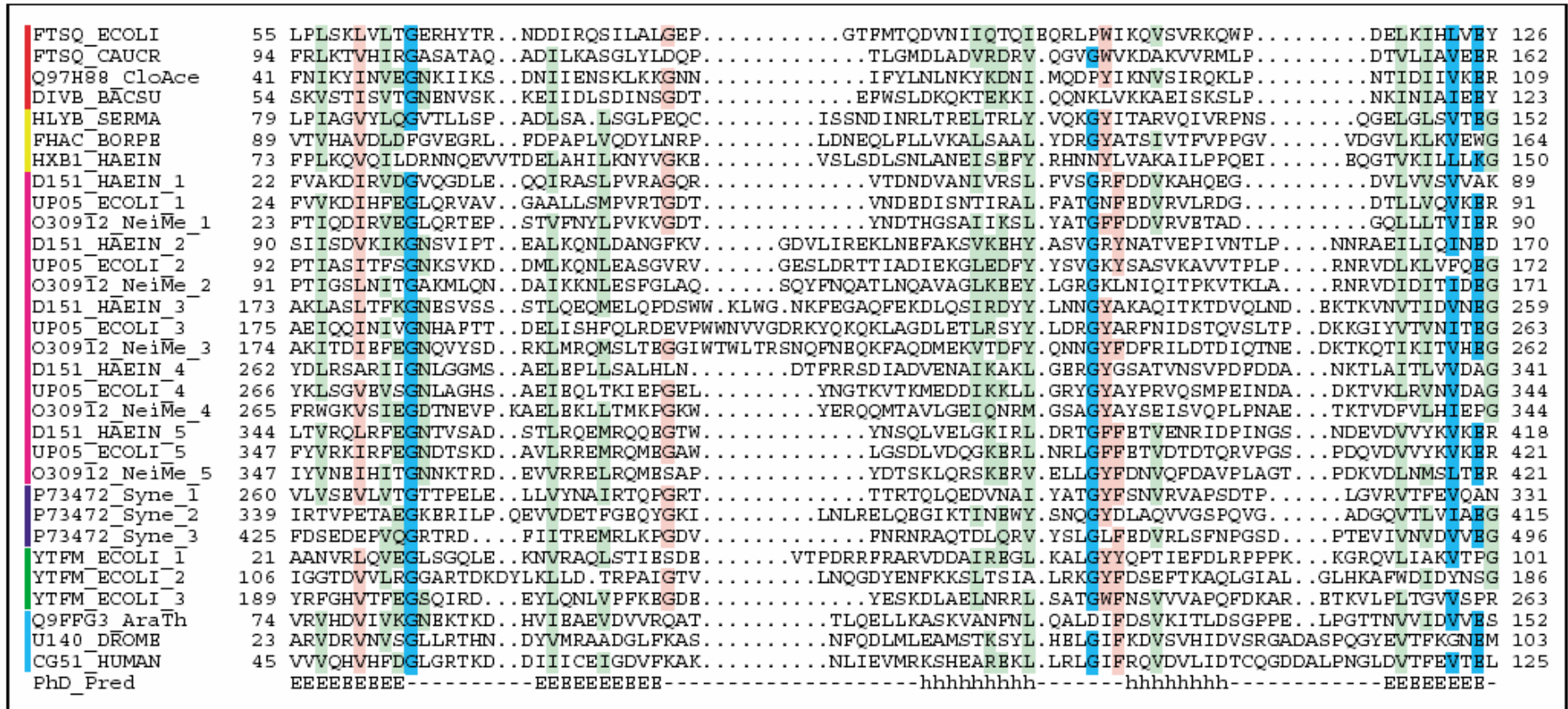


Figure 1. Representative multiple alignment of the POTRA (for polypeptide-transport-associated) domain. The alignment was produced with HMMer [10] and T-Coffee [23] using default parameters and was slightly refined manually. It is viewed with the Belvu program (http://www.sanger.ac.uk/Software/Pfam/help/belvu_setup.shtml). The colour scheme indicates the average BLOSUM62 score (correlated to amino-acid conservation) in each alignment column: cyan, > 1.6; light red, 1–1.6; and light green, 0.3–1. The boundaries of the domains are indicated by the residue positions on each side. Consensus PHD secondary-structure prediction [11] is shown below the alignment, with E indicating a β strand and H an α helix, in upper and lower case for high and low accuracy, respectively. The sequences are named with their SWISSPROT or SPTREMBL identifications. Multiple alignments and trees for each family, profiles and other information are accessible at: <http://www.pdg.cnb.uam.es/POTRA>. A larger version of this multiple sequence alignment (alignment number ALIGN_000590) has been deposited at the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000590.dat). The species abbreviations are: AraTh, *Arabidopsis thaliana*; BACSU, *Bacillus subtilis*; BORPE, *Bordetella pertussis*; CAUCR, *Caulobacter crescentus*; CloAce, *Clostridium acetobutylicum*; DROME, *Drosophila melanogaster*; ECOLI, *Escherichia coli*; HAEIN, *Haemophilus influenzae*; NeiMe, *Neisseria meningitidis*; SERMA, *Serratia marcescens*; Syne, *Synechocystis sp.* The numbering after the protein name indicates the domain-repeat number when more than one is detected in the sequence. Different groups identified by sequence similarity are shown by coloured lines to the left of the alignment: red, FtsQ; yellow, ShIB; violet, D15; blue, Toc75; green, YTFM; and cyan, CGI51.

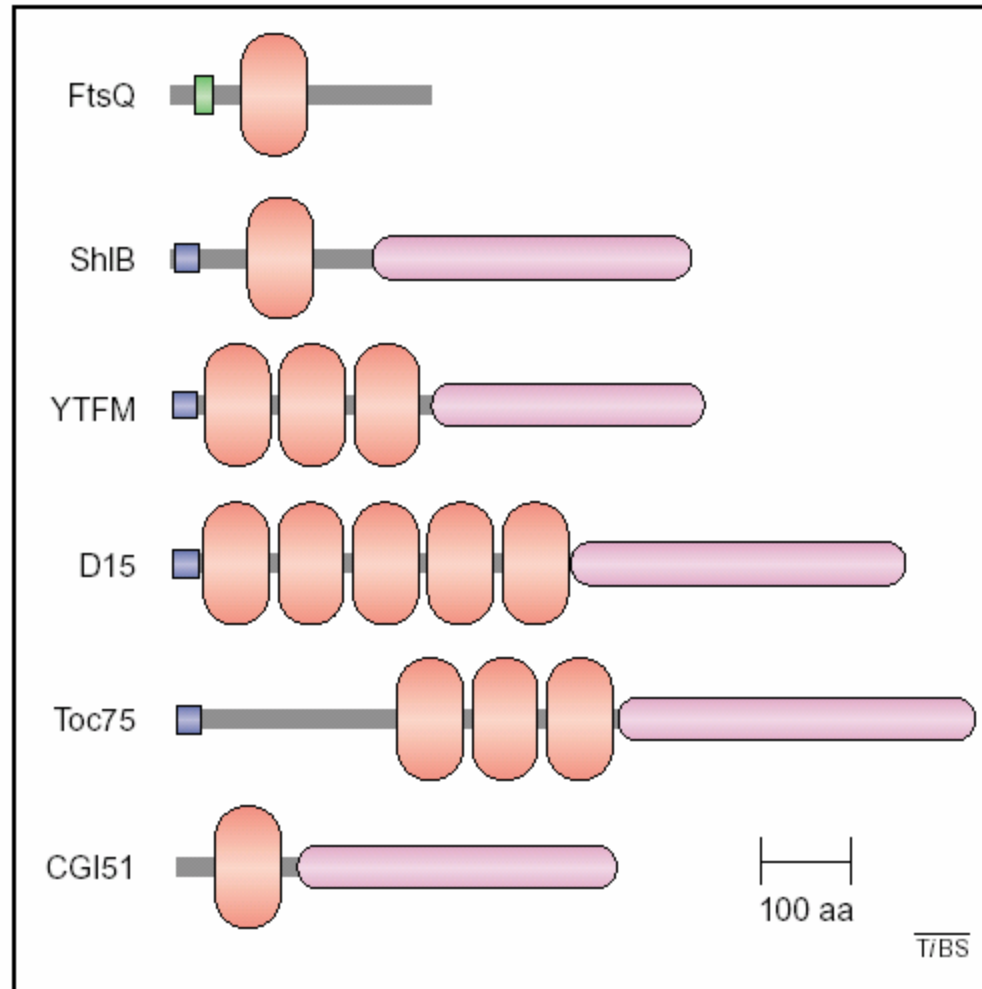


Figure 2. Schematic representation of the domain architectures of a representative set of POTRA (for polypeptide-transport-associated domain) domain-containing proteins. Corresponding to SWISSPROT identifiers: CGI51, SW:Q9Y512:CG51_HUMAN; D15, SW:P46024:D151_HAEIN; FtsQ, SW:P06136:FTSQ_ECOLI; ShIB, SW:P15321:HLYB_SERMA; Toc75, SPTREMBL:P73472; YTFM, SW:P39320:YTFM_ECOLI. The proteins are drawn approximately to scale and colour coded as follows: transmembrane region, green; signal peptide, blue; POTRA domain, pale red; β barrel, pink. Hypothetical signal peptides were predicted with Signal P [24]. Transmembrane β -sheets were predicted using B2TMPRED [25].

Regular Expressions

RU1A_HUMAN	SRSLKMRGQAFVIEKEVSSAT
SXLF_DROME	KLTGRPRGVAFVRYNKREEAQ
ROC_HUMAN	VGCSVHKGFAFVQYVNERNAR
ELAV_DROME	GNDTQTKGVGFIREDKREEAT

PROSITE Syntax:

[RK] -G- {EDRKHPCG} - [AGSCI] - [FY] - [LIVA] -x- [FYM]

Regular Expression:

[RK] G[^EDRKHPCG] [AGSCI] [FY] [LIVA] . [FYM]

Motifs, Profiles και Patterns σε πολλαπλές στοιχίσεις

```

1  MPAF-YLSCGVIWQAGA---WEGHYD
2  -PAFHYLSCREQWQE-----VEGHYP
3  GPAFHYLSCCVEHQGAGGGAYEGHY-
4  APAF-YVSCIEIYQQRGV---QGHYG
5  SPAFHYLSCDFYWQAAD---IEGHYA
6  -PAWIWLSCIERWQGAD---EGHY-
consensus -PAf-yLAc---wQ-----eGHY-

```

PROSITE Syntax:

P-A-[FW]-X-[YW]-[LV]-S-C-X(3)-[WYH]-Q-X(1-7)-[EQ]-G-H-Y

Regular Expression:

PA[FW].[YW][LV]SC.{3}[WYH]Q.{1,7}[EQ]GHY

EGF domain

AGRI_CHICK/1-3	P	C	D	S	H	--	P	C	L	H	G	G	T	C	E	D	D	-----	G	R	E	F	T	C	R	C	P	A	G	K	G	A	V	C	E					
GLP1_CAEL/2-0	P	C	D	S	D	--	P	C	N	N	G	-	L	C	Y	P	F	Y	-----	G	G	F	Q	C	I	C	N	N	G	Y	G	G	S	Y	C	E				
NTC3_MOUSE/25-	P	C	F	S	R	--	P	C	L	H	G	G	I	C	N	P	T	H	-----	P	G	F	E	C	T	C	R	E	G	F	T	G	S	Q	C	Q				
NTC3_MOUSE/19-	A	C	E	S	Q	--	P	C	Q	A	G	G	T	C	T	S	D	G	-----	I	G	F	R	C	T	C	A	P	G	F	Q	G	H	Q	C	E				
NTC3_MOUSE/32-	P	C	E	S	Q	--	P	C	Q	H	G	G	Q	C	R	H	S	L	G	R	G	G	-	L	T	F	T	C	H	C	V	P	P	F	W	G	L	R	C	E
CRB_DROME/14-0	E	C	D	S	N	--	P	C	S	K	H	G	N	C	N	D	G	I	-----	G	T	Y	T	C	E	C	E	P	G	F	E	G	T	H	C	E				
NTC4_MOUSE/25-	L	C	Q	S	Q	--	P	C	S	N	G	G	S	C	E	I	T	T	G	P	---	P	G	F	T	C	H	C	P	K	G	F	E	G	P	T	C	S		
NTC4_MOUSE/17-	A	C	H	S	G	--	P	C	L	N	G	G	S	C	S	I	R	P	-----	E	G	Y	S	C	T	C	L	P	S	H	T	G	R	H	C	Q				
FAT_DROME/2-0	V	C	Y	S	K	--	P	C	R	N	G	G	S	C	Q	R	S	P	D	G	---	S	S	Y	F	C	L	C	R	P	G	F	R	G	N	Q	C	E		
NOTC_BRARE/3-0	A	C	M	N	S	--	P	C	R	N	G	G	T	C	S	L	L	T	L	-----	D	T	F	T	C	R	C	Q	P	G	W	S	G	K	T	C	Q			
NOTC_BRARE/6-0	P	C	L	P	S	--	P	C	R	S	G	G	T	C	V	Q	T	S	D	-----	T	T	H	T	C	S	C	L	P	G	F	T	G	Q	T	C	E			
DLK_HUMAN/4-0	N	C	A	S	S	--	P	C	Q	N	G	G	T	C	L	Q	H	T	Q	-----	V	S	Y	E	C	L	C	K	P	E	F	T	G	L	T	C	V			
NTC4_MOUSE/1-3	L	C	G	G	S	P	E	P	C	A	N	G	G	T	C	L	R	L	S	Q	---	G	Q	G	I	C	Q	C	A	P	G	F	L	G	E	T	C	Q		
NOTC_BRARE/9-0	D	C	A	S	A	--	A	C	S	H	G	A	T	C	H	D	R	V	-----	A	S	F	F	C	E	C	P	H	G	R	T	G	L	L	C	H				
NTC4_MOUSE/18-	H	C	V	S	A	--	S	C	L	N	G	G	T	C	V	N	K	P	-----	G	T	F	F	C	L	C	A	T	G	F	Q	G	L	H	C	E				
DLL1_MOUSE/6-0	D	C	A	S	S	--	P	C	A	N	G	G	T	C	R	D	S	V	-----	N	D	F	S	C	T	C	P	P	G	Y	T	G	K	N	C	S				
DL_DROME/7-0	L	C	L	I	R	--	P	C	A	N	G	G	T	C	L	N	L	N	-----	N	D	Y	Q	C	T	C	R	A	G	F	T	G	K	D	C	S				
FBP1_STRPU/2-0	D	C	D	P	N	--	L	C	Q	N	G	A	A	C	T	D	L	V	-----	N	D	Y	A	C	T	C	P	P	G	F	T	G	R	N	C	E				

* * * *

-C-x-C-x(5)-G-x(2)-C

Patterns

- Απλά στην κατασκευή και διαισθητικά
- Εύκολα στην υλοποίηση και αναζήτηση
- Δύσκαμπτα σε πιο πολύπλοκες καταστάσεις
- Πολλά false positive/false negative
- Τα μικρά patterns έχουν μεγάλη πιθανότητα τυχαίας εμφάνισης
- Motifs: small patterns

Sequence profiles are a condensed representation of multiple alignments

master sequence →

HBA_human	...	W	G	K	V	G	A	-	-	H	A	G	E	...
HBB_human	...	W	G	K	V	-	-	-	-	N	V	D	E	...
MYG_phyca	...	W	G	K	V	E	A	-	-	D	V	A	G	...
LGB2_luplu	...	W	K	D	F	N	A	-	-	N	I	P	K	...
GLB1_glydi	...	W	E	E	I	A	G	A	D	N	G	A	G	...

Each column of the profile $p_j(a)$ contains the amino acid frequencies in the multiple sequence alignment

A	...	0	0	0	0	0.25	0.75			0	0.2	0.4	0	...
C	...	0	0	0	0	0	0			0	0	0	0	...
D	...	0	0	0.2	0	0	0			0.2	0	0.2	0	...
E	...	0	0.2	0.2	0	0.25	0			0	0	0	0.4	...
F	...	0	0	0	0.2	0	0			0	0	0	0	...
G	...	0	0.6	0	0	0.25	0.25			0	0.2	0.2	0.4	...
H	...	0	0	0	0	0	0			0.2	0	0	0	...
I	...	0	0	0	0.2	0	0			0	0.2	0	0	...
K	...	0	0.2	0.6	0	0	0			0	0	0	0.2	...
L	...	0	0	0	0	0	0			0	0	0	0	...
M	...	0	0	0	0	0	0			0	0	0	0	...
N	...	0	0	0	0	0.25	0			0.6	0	0	0	...
P	...	0	0	0	0	0	0			0	0	0.2	0	...
Q	...	0	0	0	0	0	0			0	0	0	0	...
R	...	0	0	0	0	0	0			0	0	0	0	...
S	...	0	0	0	0	0	0			0	0	0	0	...
T	...	0	0	0	0	0	0			0	0	0	0	...
V	...	0	0	0	0.6	0	0			0	0.4	0	0	...
W	...	1.0	0	0	0	0	0			0	0	0	0	...
Y	...	0	0	0	0	0	0			0	0	0	0	...

PROSITE

- Η **PROSITE** (<http://www.expasy.ch/prosite/>) αποτελεί μια βάση ταξινόμησης πρωτεϊνικών ακολουθιών και αυτοτελών περιοχών ακολουθιών (sequence domains) σε οικογένειες ([Sigrist, et al., 2010](#)).
- Υπάρχουν γενικά δύο τρόποι για τη δημιουργία των 'αποτυπωμάτων'. Ο ένας βασίζεται στη χρήση μιας γλώσσας παρόμοιας με αυτής των "κανονικών εκφράσεων" (regular expressions), και είναι ο πιο παλιός και εύκολος στη δημιουργία, ενώ ο άλλος βασίζεται στην κατασκευή profiles (πίνακες με ειδικές ανά θέση πιθανότητες εμφάνισης αμινοξέων), μέθοδος η οποία είναι πιο σύνθετη αλλά και πιο ευαίσθητη.
- Μέχρι σήμερα η PROSITE περιέχει 'αποτυπώματα' για περίπου 1716 οικογένειες για καθεμία από τις οποίες συμπεριλαμβάνεται λεπτομερής ανάλυση για τη δομή και τη λειτουργία των πρωτεϊνών που την αποτελούν. Συνολικά, υπάρχουν στη βάση 1308 patterns, 1107 profiles και 1105 "κανόνες" (αφορούν κυρίως πληροφορίες για το που θα πρέπει να βρίσκεται το pattern για να θεωρηθεί έγκυρο αλλά και πληροφορίες για συνδυασμούς από patterns). Προφανώς, υπάρχουν οικογένειες για τις οποίες υπάρχουν διαθέσιμα και patterns και profiles (συνήθως, η παλαιότερες καταχωρήσεις αφορούσαν το pattern).
- Στην βάση υπάρχουν επίσης, αναλύσεις για τις πρωτεΐνες της UniProt που ανήκουν σε κάθε οικογένεια όσο και για τις πρωτεΐνες στις οποίες εμφανίζεται ένα "αποτύπωμα" (κυρίως όταν έχουμε να κάνουμε με pattern) αλλά είναι γνωστό ότι αυτές δεν ανήκουν λειτουργικά στην οικογένεια αυτή. Τέλος, υπάρχουν εργαλεία για την αναζήτηση των patterns και των profiles σε ακολουθίες, όσο και εργαλεία αναπαράστασης της "σπονδυλωτής" δομής των πρωτεϊνών, δηλαδή της αναπαράστασης των περιοχών αυτών και την αποτύπωση τους πάνω σε μια δεδομένη ακολουθία.

PFAM

- **PFAM:** Η βάση Pfam (<http://pfam.xfam.org/>) αποτελεί μια μεγάλη συλλογή πρωτεϊνικών οικογενειών (Finn, et al., 2014). (Andreeva, et al., 2004) Βασίζεται στην ίδια λογική με την PROSITE (ειδικά με το υποσύνολο της που βασίζεται σε profiles), αλλά η μεγάλη διαφορά είναι ότι εδώ οι οικογένειες χαρακτηρίζονται από ένα hidden Markov model (HMM), μέθοδος η οποία είναι πιο ευαίσθητη στον εντοπισμό μακρινών ομόλογων, χωρίς όμως να υστερεί σε ταχύτητα και αποτελεσματικότητα.
- Στην τρέχουσα έκδοση (2013), η βάση περιέχει δεδομένα για 14.831 οικογένειες παρέχοντας κάλυψη για πάνω από το 80% των πρωτεϊνικών καταχωρήσεων της UNIPROT.
- Η PFAM αποτελείται από δύο υποσύνολα, την PFAM-A, και την PFAM-B. Η PFAM-A αποτελείται από καταχωρήσεις (οικογένειες) υψηλής «ποιότητας», καθώς έχουν όλες υποστεί σχολιασμό από ειδικούς, ενώ υπάρχουν αναφορές σε άλλες βάσεις δεδομένων και κυρίως σε βιβλιογραφία. Η PFAM-B είναι το υποσύνολο, το οποίο προκύπτει με αυτοματοποιημένο τρόπο εντοπίζοντας τις ομοιότητες ανάμεσα στις πρωτεϊνικές περιοχές που απομένουν όταν αφαιρεθούν οι περιοχές που αντιστοιχούν στις καταχωρήσεις της PFAM-A. Η PFAM-B είναι ιδιαίτερα χρήσιμη, γιατί με στοχευμένη ανάλυση αυτών των «οικογενειών», μπορούν να προκύψουν οικογένειες που μετέπειτα θα «προαχθούν» στην PFAM-A.
- Το βασικό χαρακτηριστικό της PFAM, και αυτό που την κάνει τόσο δημοφιλή, είναι ότι με τη χρήση του HMM (και ειδικά του πακέτου HMMER), μπορεί να επιλεγεί για κάθε οικογένεια μία τιμή διαχωριστικού κατοφλίου στο σκορ, και κατά συνέπεια κάθε πρωτεΐνη ταξινομείται μόνο σε μία οικογένεια (σε αυτή που σκοράρει πάνω από το κατώφλι).
- Παρόλα αυτά, χαμηλότερη ομοιότητα μπορεί να υπάρχει μεταξύ πρωτεϊνών που ανήκουν σε διαφορετικές οικογένειες, γιαυτό και η βάση περιέχει και μια ανώτερη κατηγορία οργάνωσης, την υπερ-οικογένεια (clan).

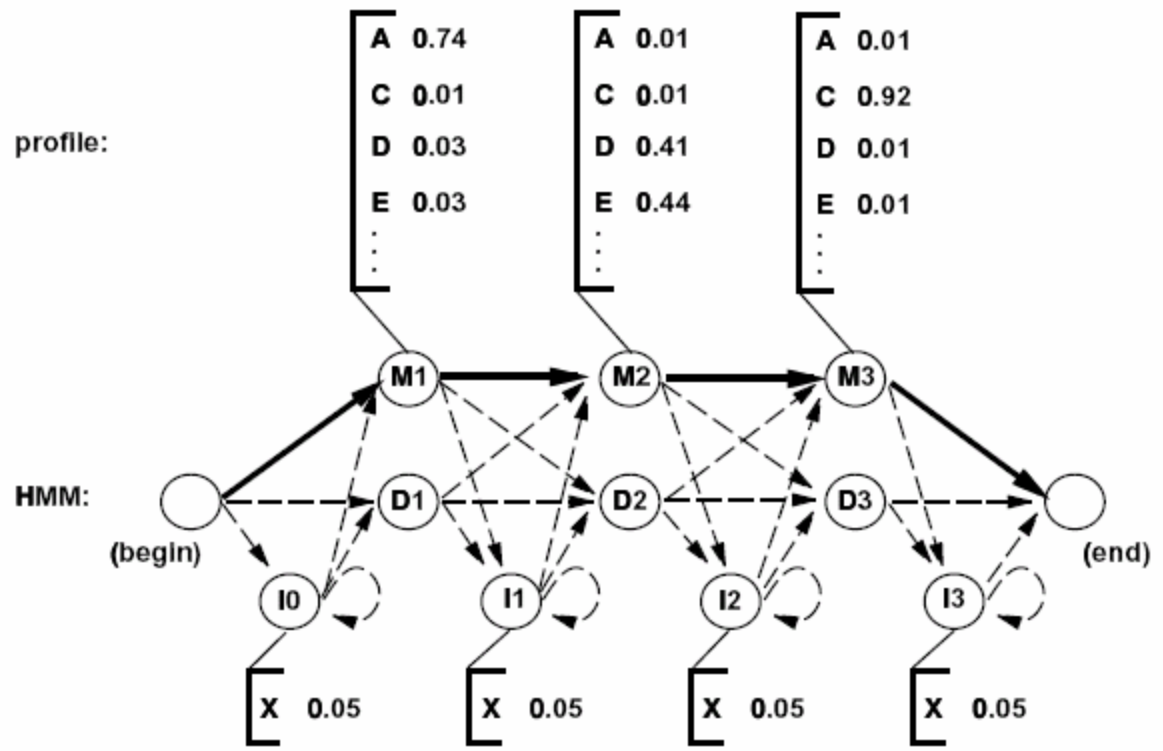
multiple alignment:

```

- A D T C
W A E - C
- V E - C
- A D - C
- A E - C

```

consensus: A D/E C

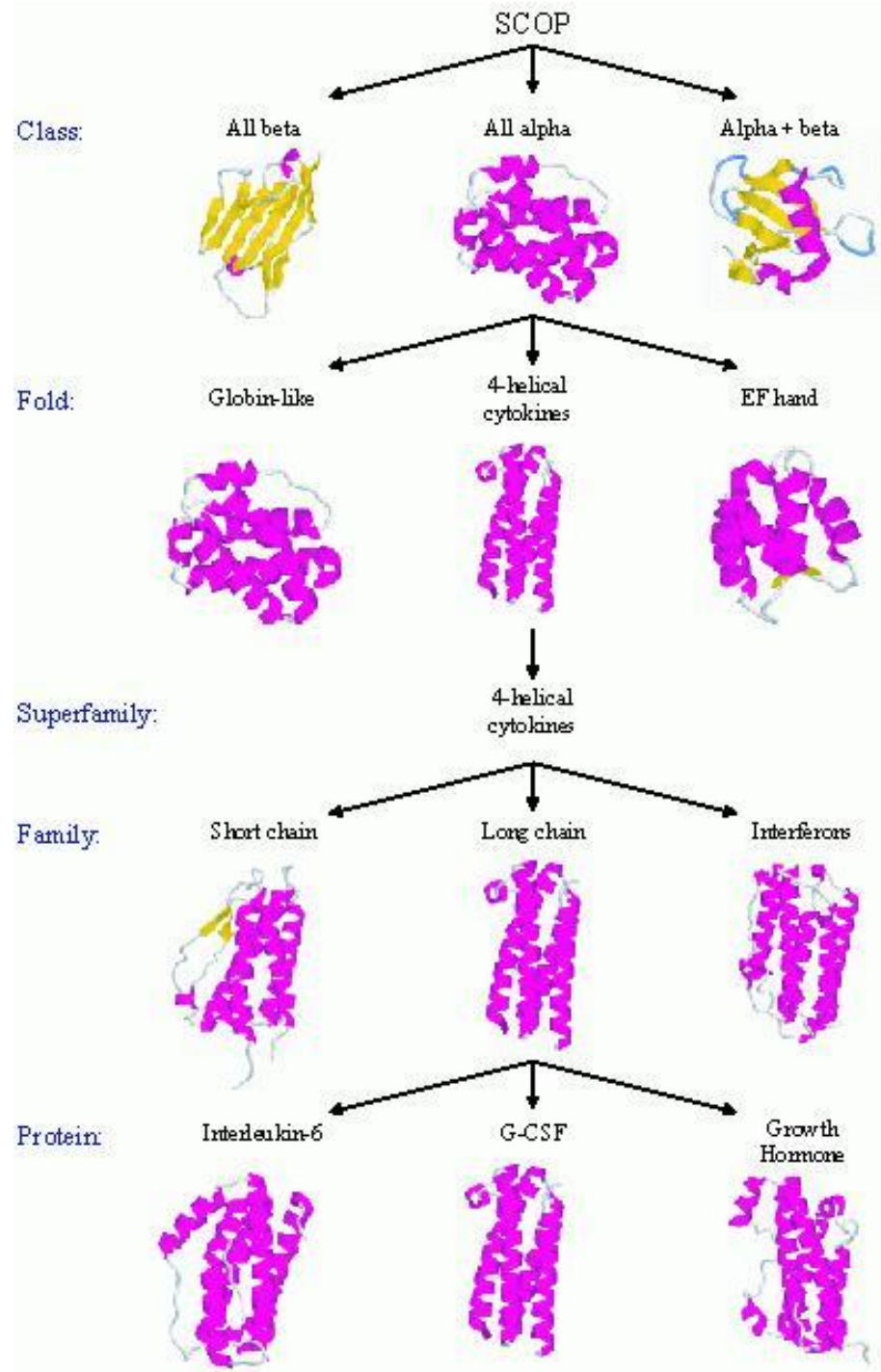


CATH

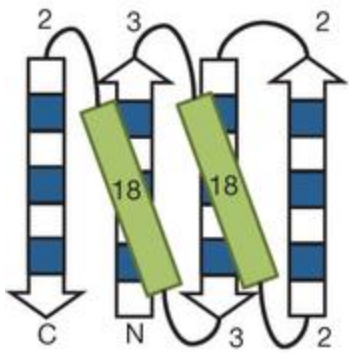
- **CATH:** Η CATH (http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html) είναι μια βάση ιεραρχικής ταξινόμησης πρωτεϊνικών δομών που αποτελούν εγγραφές της PDB με βάση τις αυτοτελείς δομικές περιοχές (domains) που τις απαρτίζουν ([Knudsen and Wiuf, 2010](#)).
- Η CATH περιέχει αποκλειστικά πρωτεϊνικές δομές που είναι προσδιορισμένες σε διακριτικότητα καλύτερη των 3 Angstroms και χρησιμοποιεί κυρίως αυτοματοποιημένες μεθόδους για την ταξινόμησή τους. Σε ειδικές περιπτώσεις και όταν αυτό κρίνεται απαραίτητο χρησιμοποιούνται και ανθρώπινα κριτήρια.
- Η ιεραρχία αποτελείται κυρίως από τέσσερα επίπεδα: 1) την Τάξη (Class), 2) την Αρχιτεκτονική (Architecture), 3) την Τοπολογία (Οικογένεια διπλώματος) (Topology (fold family)) και 4) την Ομόλογη Οικογένεια (Homologous superfamily). Οι πρωτεΐνες που αποτελούνται από πάνω από μία αυτοτελείς δομικές περιοχές (domains), αναλύονται στα επιμέρους στοιχεία αυτόματα με βάση ειδικούς αλγόριθμους αναγνώρισης των περιοχών.
- Η αυτόματη αυτή διαδικασία κατατάσσει το 53% των δομών. Οι υπόλοιπες διαχωρίζονται στις επιμέρους αυτοτελείς δομικές περιοχές με παρατηρήσεις που προκύπτουν είτε από τους αλγόριθμους αυτόματου διαχωρισμού είτε από τη βιβλιογραφία. Η ταξινόμηση πραγματοποιείται μόνο στις αυτοτελείς δομικές περιοχές.

SCOP

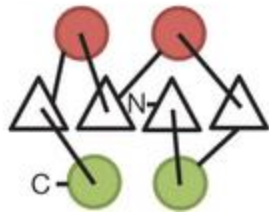
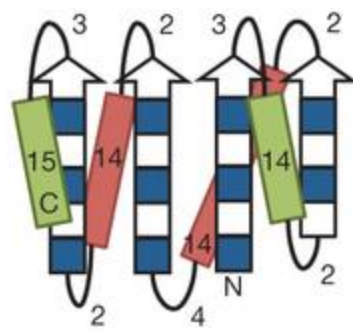
- **SCOP:** Ο βασικός στόχος της βάσης SCOP (<http://scop.mrc-lmb.cam.ac.uk/scop/index.html>) είναι η ανάλυση των δομικών και εξελικτικών σχέσεων που παρατηρούνται μεταξύ όλων των πρωτεϊνών γνωστής δομής καταχωρημένων στην PDB (Andreeva, et al., 2004). Η ταξινόμηση των πρωτεϊνών πραγματοποιείται βάσει αυτών των δομικών και εξελικτικών σχέσεων. Τα βασικά επίπεδα ταξινόμησης είναι τέσσερα: 1) η οικογένεια (Family), 2) η υπερ-οικογένεια (Superfamily), 3) το δίπλωμα (Fold) και 4) η τάξη (Class).
- Οικογένεια (Family): Μεταξύ των μελών της οικογένειας παρατηρείται ξεκάθαρη εξελικτική σχέση. Η ομοιότητα σε επίπεδο ακολουθίας είναι ίση ή μεγαλύτερη του 30%. Παρόλα αυτά υπάρχουν περιπτώσεις στις οποίες οι δομές και η λειτουργία είναι παρόμοιες υποδηλώνοντας κοινό πρόγονο ενώ η ομοιότητα σε επίπεδο ακολουθίας είναι μικρότερη του 30% (σφαιρίνες, 15%).
- Υπερ-οικογένεια (Superfamily): Οι πρωτεΐνες που κατατάσσονται στις υπερ-οικογένειες εμφανίζουν πολύ μικρή ομοιότητα στο επίπεδο της ακολουθίας αλλά τα δομικά τους χαρακτηριστικά και η λειτουργία τους υποδηλώνουν ότι πιθανά προέλθει από κοινό πρόγονο.
- Δίπλωμα (Fold): Σε αυτό το επίπεδο κατατάσσονται πρωτεΐνες που παρουσιάζουν ομοιότητα σε επίπεδο δομής. Οι πρωτεΐνες που εμφανίζουν το ίδιο δίπλωμα έχουν τα ίδια σε μεγάλο βαθμό χαρακτηριστικά δευτεροταγούς δομής, με κοινό προσανατολισμό και τις ίδιες τοπολογικές συνδέσεις μεταξύ τους. Πρωτεΐνες που έχουν το ίδιο δίπλωμα αλλά δεν είναι όμοιες από άποψη αμινοξικής ακολουθίας έχουν ορισμένα περιφερειακά στοιχεία της δευτεροταγούς τους δομής και στροφές ανόμοια και όσον αφορά στο μέγεθος και όσον αφορά στη διαμόρφωση. Πρωτεΐνες που εμφανίζουν κοινό δίπλωμα δεν είναι απαραίτητο να έχουν κοινή εξελικτική προέλευση.
- Τάξη (Class): Η ταξινόμηση γίνεται με βάση το δίπλωμα των στοιχείων δευτεροταγούς δομής των πρωτεϊνών σε τέσσερις κύριες δομικές κατηγορίες: 1) την all- α , όπου η δομή σχηματίζεται από α -έλικες, 2) την all- β , όπου η δομή αποτελείται από β -πτυχωτές επιφάνειες, 3) την α/β , όπου στην δομή της πρωτεΐνης εναλλάσσονται α -έλικες και β -πτυχωτές επιφάνειες και 4) την $\alpha+\beta$, όπου σε διακριτές περιοχές της δομής βρίσκονται α -έλικες και β -πτυχωτές επιφάνειες.
- Η αναγνώριση των σχέσεων καθώς και η ταξινόμηση βάσει των σχέσεων μεταξύ των πρωτεϊνών πραγματοποιείται αποκλειστικά από ειδικούς επιστήμονες μετά από λεπτομερή μελέτη και σύγκριση των πρωτεϊνικών δομών. Αυτοματοποιημένες μέθοδοι χρησιμοποιούνται μόνο για την ομοιογένεια των δεδομένων που περιέχονται στη βάση



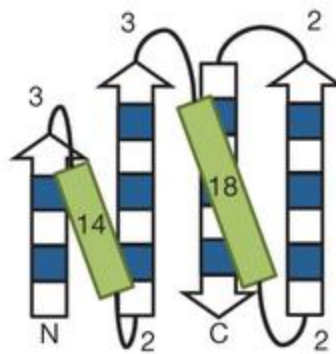
Fold-I: $\beta\alpha\beta\beta\alpha\beta$



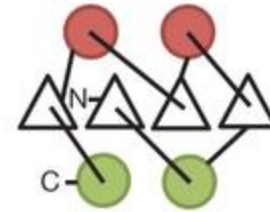
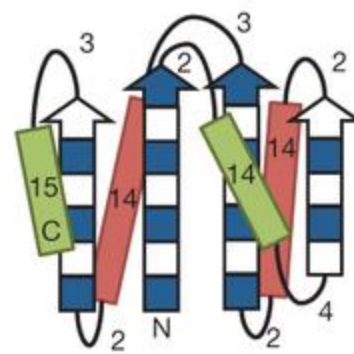
Fold-II: $\beta\alpha\beta\alpha\beta\alpha\beta$



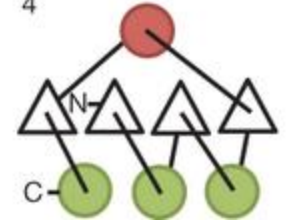
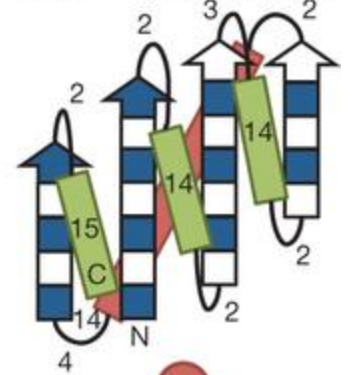
Fold-III: $\beta\alpha\beta\alpha\beta\beta$

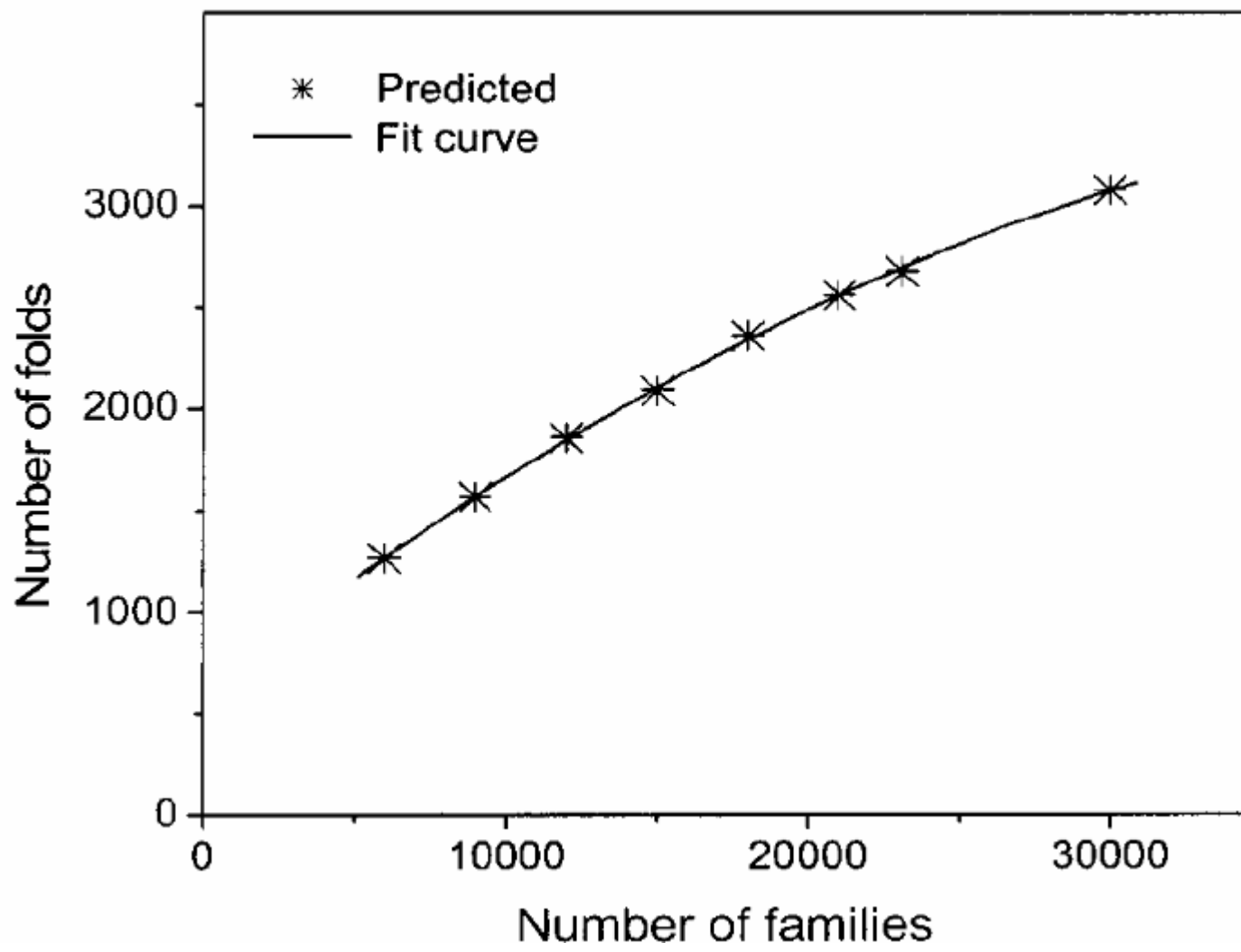


Fold-IV: $\beta\alpha\beta\alpha\beta\alpha\beta$



Fold-V: $\beta\alpha\beta\alpha\beta\alpha\beta$





Εξειδικευμένες βάσεις δεδομένων

- Εκτός από τις μεγάλες, δημόσια διαθέσιμες και ευρέως χρηματοδοτούμενες βάσεις δεδομένων που αναφέρθηκαν παραπάνω, σημαντικό ρόλο στην πρόοδο της βιοπληροφορικής παίζουν και οι εξειδικευμένες βάσεις δεδομένων.
- Συνήθως, αλλά όχι πάντα, αφορούν τις πρωτεϊνικές ακολουθίες (γιατί για αυτές υπάρχει μεγάλη πληθώρα λειτουργικών δεδομένων, σε μεγάλη λεπτομέρεια, που δεν μπορεί να καλυφθεί από τις βάσεις όπως η Uniprot), και τις περισσότερες φορές, συντηρούνται από μικρές ή μεσαίου μεγέθους ερευνητικές ομάδες.

Specialized Protein Resources Network

- Στις 11 - 12 Αυγούστου 2014 πραγματοποιήθηκε με την χρηματοδότηση του Wellcome Trust, στο Hinxton της Αγγλίας, μία συνάντηση είκοσι ενός κύριων ερευνητών που ο καθένας διατηρεί μια εξειδικευμένη πρωτεϊνική βάση δεδομένων ή διεξάγει έρευνα σχετικά με την διατήρηση ενός τέτοιου αποθετηρίου (Specialized Protein Resources Network).
- Το θέμα της συνάντησης ήταν η χάραξη πολιτικής για την δημιουργία και διατήρηση πρωτεϊνικών βάσεων δεδομένων και αποτελούνταν από πέντε ενότητες:
 - (1) βασικές προκλήσεις,
 - (2) εισαγωγή δεδομένων,
 - (3) βέλτιστες πρακτικές για τη διατήρηση και την επιμέλεια,
 - (4) ροή πληροφοριών προς και από τα μεγάλα κέντρα δεδομένων, και
 - (5) επικοινωνία και χρηματοδότηση.

Specialized Protein Resources Network

- Στην συνάντηση συμμετείχαν ερευνητές που διατηρούν «εξειδικευμένες» ηλεκτρονικές βάσεις δεδομένων συγκεκριμένων ειδών πρωτεϊνών (όπως αυτές ορίζονται από τα ενζυματικά, λειτουργικά ή δομικά χαρακτηριστικά τους) αλλά και διαχειριστές μεγάλων πρωτεϊνικών αποθετηρίων (συμπεριλαμβανομένων των Pfam, RefSeq, Swiss-Prot, και UniProt).
- Αυτά τα μεγάλα κέντρα δεδομένων χρησιμοποιούν διάφορες προσεγγίσεις για να συντηρήσουν το περιεχόμενο των δεδομένων τους, όπως η υπολογιστική ανάλυση, η συνεργασία, η ενοποίηση δεδομένων από πολλαπλές πηγές, και η επιμέλεια από ειδικούς σχολιαστές.
- Όλες οι βάσεις δεδομένων υποστηρίζονται από ειδικό σχολιασμό ώστε να εξασφαλιστεί η ακρίβεια και η πληρότητα των στοιχείων που παρουσιάζονται σε κάθε μικρή ή μεγάλη πρωτεϊνική βάση δεδομένων.
- Ένα κοινό πρόβλημα όλων των συμμετεχόντων της συνάντησης ήταν η επιμέλεια και η ανανέωση των βάσεων, δεδομένου ότι είναι δύσκολη η ανάκτηση πληροφοριών από δημοσιευμένα άρθρα επειδή συχνά δεν αναφέρουν αναλυτικές συγκεκριμένες πληροφορίες για τον υπό μελέτη οργανισμό (ειδικά για τα strains), ή τις ακριβείς πληροφορίες της αλληλουχίας που αναλύθηκε (πχ ο κωδικός πρόσβασης στη UniProt ή το gi).
- Η διεύρυνση των συνεργασιών για τη διόρθωση λαθών στις βάσεις δεδομένων και η διάδοση της γνώσης αναγνωρίστηκαν από όλους ως βασικοί τρόποι δράσης, που θα ωφελήσουν όλες τις πρωτεϊνικές πηγές αλλά και τους χρήστες τους.



Group photo of the participants at the Protein Bioinformatics and Community Resources Retreat. The name of each participant is followed by the short name of their protein resource or resources in parentheses. Back row: David Landsman (Histone database), Dan Haft (TIGRFAMS), Bernard Henrissat (CAZy), Rob Finn (InterPro and Pfam), David Craik (ConoServer and CyBASE), Arnaud Chatonnet (ESTHER), Neil Rawlings (MEROPS); Middle row: Amos Bairoch (neXtProt), Gerard Manning (Kinase.com), Michael Spedding (IUPHAR), Gert Vriend (GPCRDB), Milton Saier (TCDB), Pantelis Bagos (OMPdb); Front row: Narayanaswamy Srinivasan (KinG), Ramanathan Sowdhamini (PASS2), Alex Bateman (Pfam & UniProt), Patsy Babbitt (SFLD), Kim Pruitt (RefSeq), Claire O'Donovan (UniProt), Gemma Holliday (MACiE), Nozomi Nagano (EzCatDB).

Best practices (Gert Vriend)

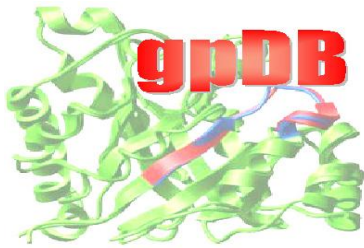
1. Longevity - The one rule to rule them all. Gert asks that unless you can maintain your database for at least 10 years, then do not start.
2. Users - All databases need users and citations. To gain and keep users, you need to provide query and browsing interfaces as well as someone who answers emails.
3. Befriend Nucleic Acids Research and DATABASE journals - The descriptions of your database are essential to inform new users. But it is also essential to target publications to the readership.
4. Collaborate - Your collaborators may offer an exit strategy in the future.
 - 4a. Be open - Nobody is going to steal your resource.
5. Give credit - There is more than 100% to go around.
6. Automate - Too much manual intervention makes for an unsustainable database leading to premature death. You need to automate roughly 90% of everything every year.
7. No new standards – Don't invent a new standard. Use what exists.
8. Keep it simple - Google is a model interface.
9. Visibility - Be at the right conferences and be recognizable. Use the same logo and present a poster.
10. Exit strategy - At some point you will retire. Start planning early to ensure your database continues.

Σφάλματα

- Κύρια πρόκληση είναι η αξιοπιστία και όχι η ποσότητα των δεδομένων. Εξαρτάται εξολοκλήρου από το πεδίο εφαρμογής και τη λειτουργία της βάσης δεδομένων.
- Υπάρχουν πολλοί διαφορετικοί τύποι σφαλμάτων που μπορούν να βρεθούν στις πηγές δεδομένων. Κάποια είναι σχετικά εύκολο να εντοπιστούν μέσω αυτοματοποιημένων διαδικασιών, όπως για παράδειγμα τα ορθογραφικά λάθη στο σχολιασμό. Σφάλματα όμως που σχετίζονται με επιστημονικές πληροφορίες είναι πολύ πιο δύσκολο να βρεθούν, ειδικά αφού η γνώση εξελίσσεται πολύ γρήγορα
- Υπάρχουν πολλοί ακόμα τύποι σφαλμάτων, για παράδειγμα ένα συχνό σφάλμα στην ανάλυση πρωτεϊνικών αλληλουχιών σχετίζεται με την σπονδυλωτή (modular) δομή πολλών πρωτεϊνών
- Ένας άλλος τύπος σφάλματος προκαλείται από την υπερεκτίμηση που συνάγεται από την "απόδειξη μέσα από την επανάληψη".

Διόρθωση

- Πολλές πηγές, όπως η UniProtKB, διαθέτουν μηχανισμούς ώστε οι χρήστες να αναφέρουν πιθανά προβλήματα. Άλλες βάσεις, όπως η PDB, δεν επιτρέπουν την διόρθωση των δεδομένων
- Όταν το σφάλμα διορθωθεί, πως μπορούμε να ενημερώσουμε όλες τις βάσεις δεδομένων που χρησιμοποιούν την αρχική εγγραφή;
- Με τον συνεχώς αυξανόμενο όγκο διαθέσιμων δεδομένων, πώς θα μπορούσε να διατηρηθεί ή ακόμα και να ενισχυθεί η αξιοπιστία των πηγών των βάσεων;
- Απαίτηση πριν από μια δημοσίευση. Εμπλοκή των χρηστών. Wikipedia.
- Τυποποιημένη γλώσσα (Αυτό που εννοεί μία βάση με τον όρο superfamily μπορεί να μην σημαίνει το ίδιο σε μία άλλη βάση)
- Οντολογίες
- Καταγραφή των βάσεων δεδομένων και διατήρηση (dead URL)

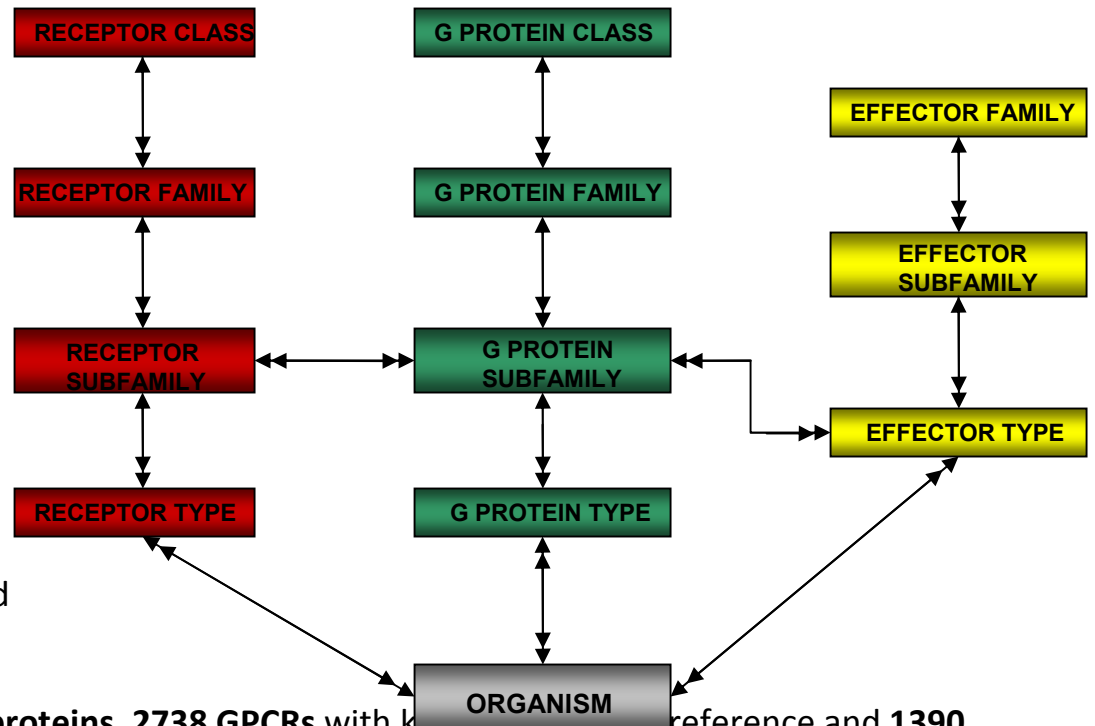


Margarita C Theodoropoulou, Pantelis G Bagos,
Ioannis C Spyropoulos and Stavros J
Hamodrakas.

"gpDB: A database of GPCRs, G-proteins,
Effectors and their interactions."

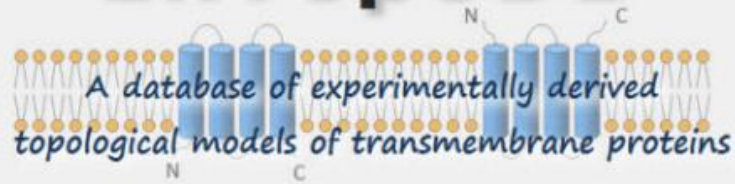
Bioinformatics. 2008 Jun 15;24(12):1471-

- A publicly accessible, **relational** database of **G-proteins** and their interactions with **GPCRs** and **effector** molecules
- gpDB currently contains data concerning **391 G-proteins**, **2738 GPCRs** with known coupling preference and **1390 effectors**, knowing to interact with specific G-proteins.
- Classification according to a hierarchy of different **classes, families, subfamilies** and types, based on extensive literature search
- The **relational model** of the database describes the known **coupling specificity** of the GPCRs to their respective alpha subunit of G-proteins and, also, the **interaction between G-protein subfamilies and specific effector types**, a unique feature not available in any other database
- Full sequence information with cross-references to publicly available databases
- Advanced text search, BLAST search against the database and a pattern search tool.
- Approx. 50 unique visitors per month



Availability: <http://bioinformatics.biol.uoa.gr/gpDB/>

ExTopoDB



Introduction

Navigation

Text search

Blast search

User manual

Download

Availability: <http://bioinformatics.biol.uoa.gr/ExTopoDB/>

- **Experimental** information collected from studies in the **literature** that report the use of biochemical methods.
- Topological models of **alpha-helical** transmembrane proteins.
- **2143** transmembrane proteins from **1833** studies.
- Topological information is combined with transmembrane topology prediction (**constrained predictions** using **HMM-TM**) resulting in more reliable topological models.
- Signal peptide annotation using **SignalP**.
- Interface that allows user-defined constrained topology prediction

Tsaousis G.N., Tsirigos K.D., Andrianou X.D., Liakopoulos T.D., Bagos P.G., Hamodrakas S.J. **ExTopoDB: A database of experimentally derived topological models of transmembrane proteins.** 2010, *Bioinformatics*,26(19): 2490–2497



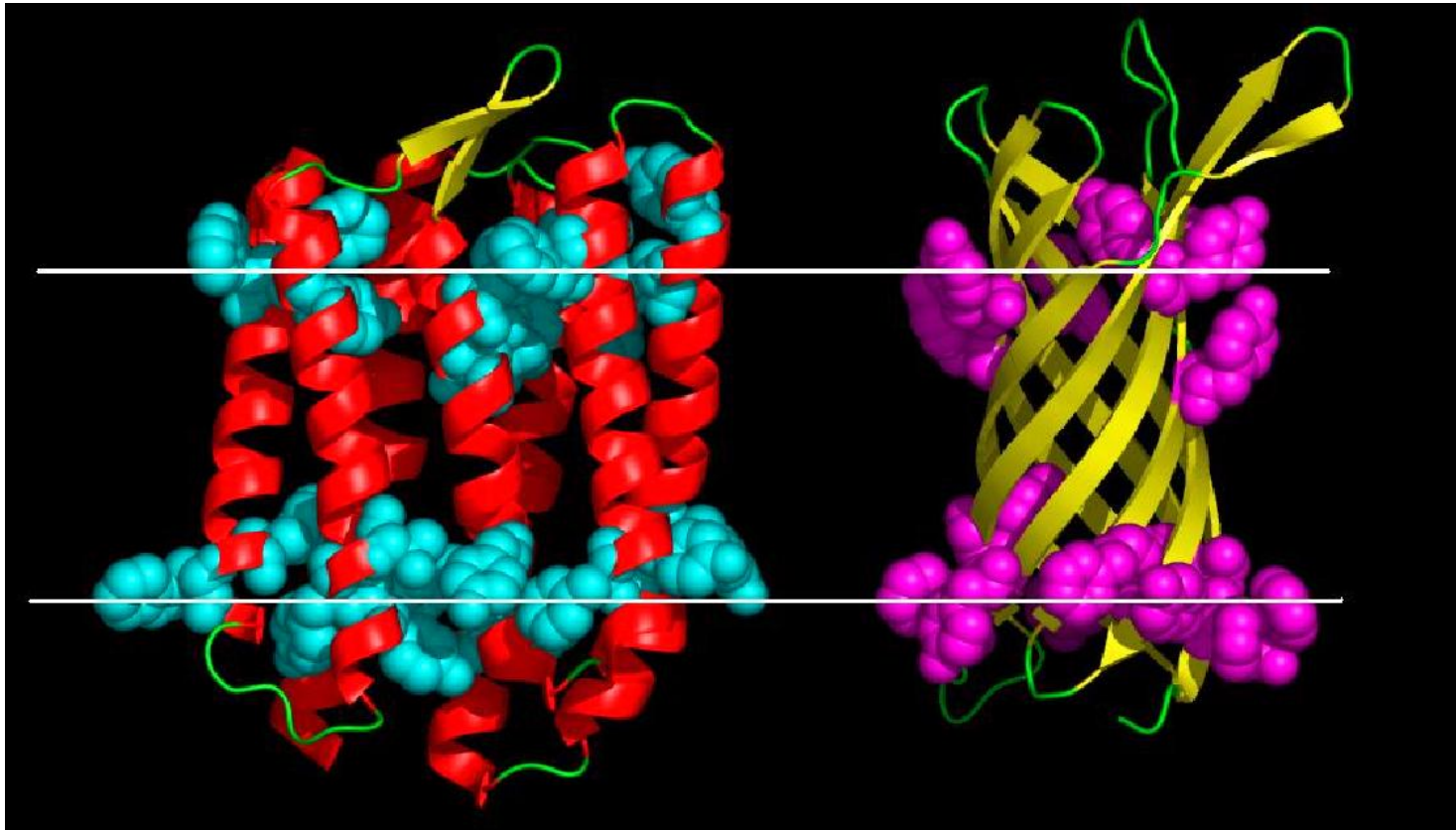
A database of β -barrel outer membrane proteins
from Gram-negative bacteria

Availability: <http://www.ompdb.org>

- The biggest collection of beta barrel proteins currently available.
- Started off with **85** families and **70,000** protein sequences and currently contains **91** families and more than **400,000** proteins.
- Out of the 91 families, **15** families were built completely from scratch, **16** do not belong to the respective clan of Pfam, while **6** of them are annotated as DUF in Pfam
- Each **family** entry contains extensive information (function of protein members, literature references, list of proteins with 3D-structure, seed and full protein alignments)
- Each database entry contains the following fields: OMPdb name, OMPdb id, Uniprot accession number, protein description and classification, sequence, species, organism name, taxonomy, links to other databases, accompanied with annotation for TM segments and signal peptides.
- OMPdb follows the monthly updates of Uniprot through an semi-automated procedure.
- **Domain and Blast Search** against OMPdb is available.
- The database can be downloaded in several **formats (text, FASTA, XML)** through the **Download** page.
- Approx. 350 unique visitors per month

Tsirigos K.D, Bagos P.G., Hamodrakas S.J. **OMPdb: a database of β -barrel outer membrane proteins from Gram-negative bacteria.** 2011, *Nucleic Acids Research*,39 (Database Issue): 324–331.

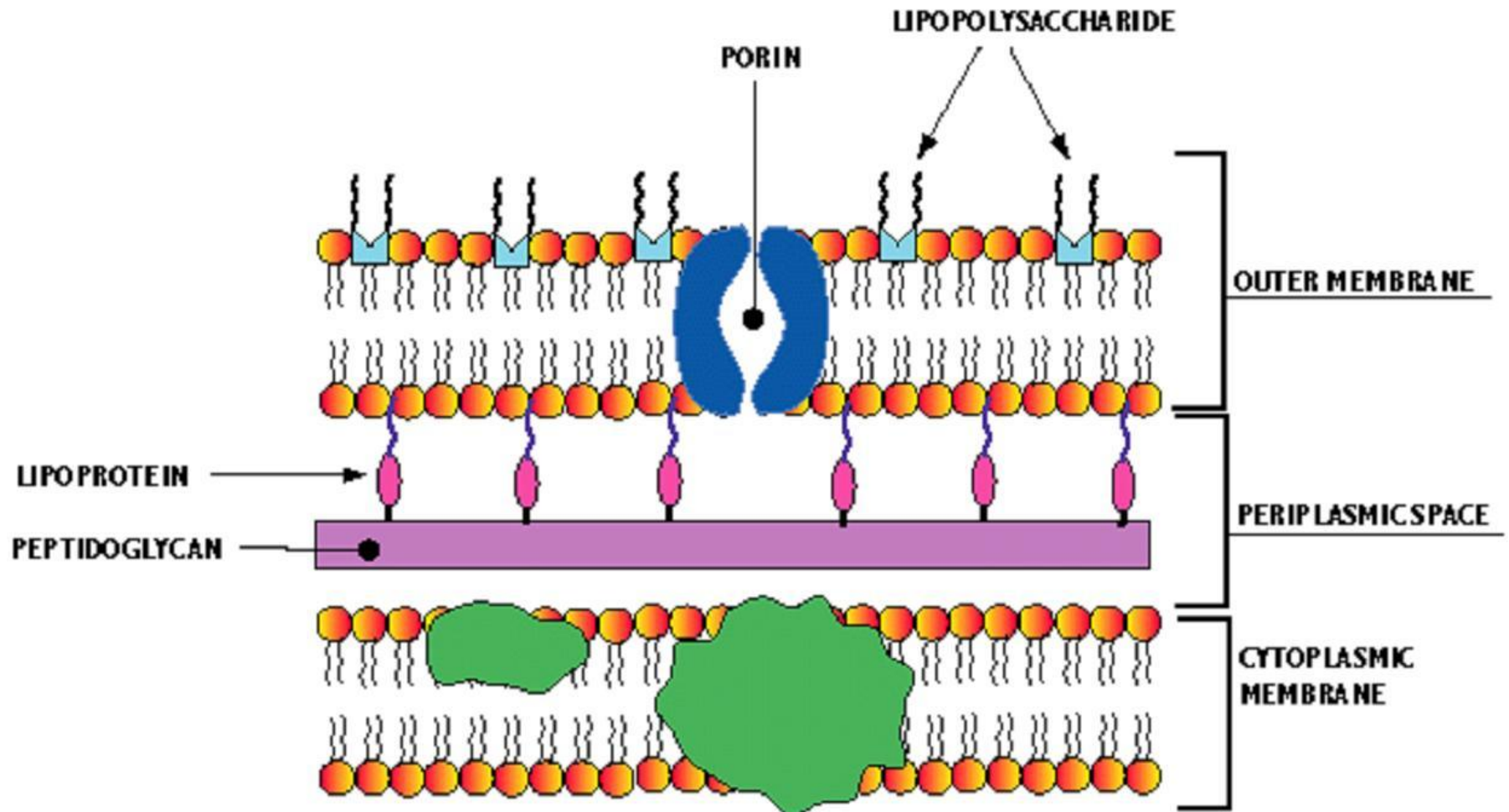
The two structural classes



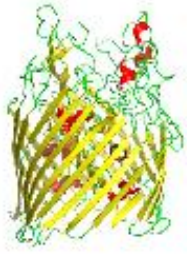
α -helical membrane proteins

β -barrel membrane proteins

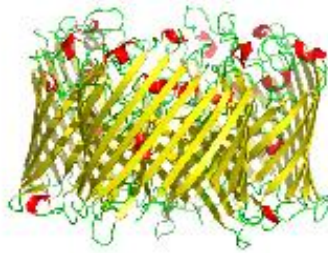
Gram-negative bacteria



Variety of structures...



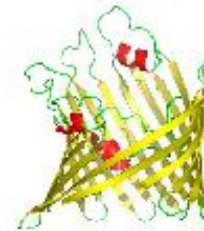
FepA (1FEP)



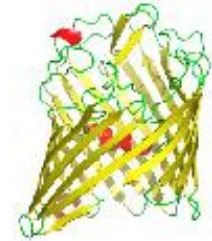
Sucroporin (1A0S)



Omp32 (1E54)



Porin (2POR)



OmpF (2OMF)



OmpX (1QJ8)



OmpA (1QJP)



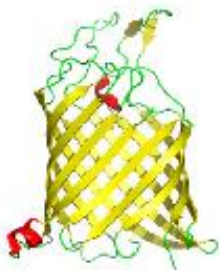
OmpT (1I78)



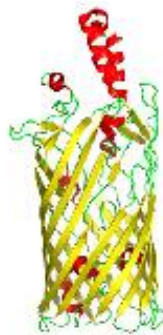
OpcA (1K24)



Omp1A (1QD5)



Tsx (1TLY)



FADL (1T1L)



NspA (1P4T)

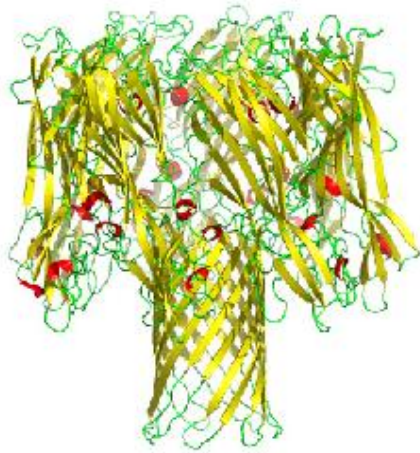


NaIP (1UYN)



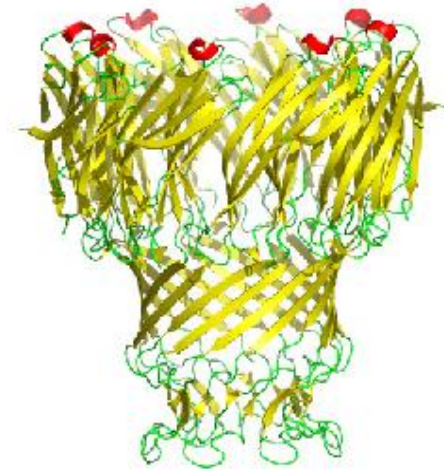
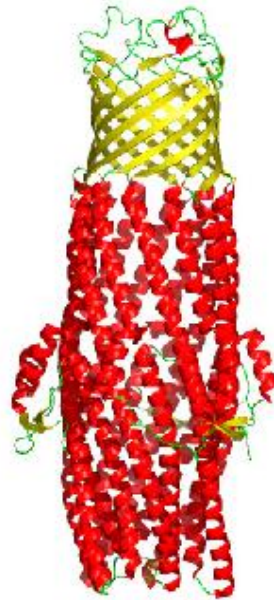
PagP (1MM4)

Variety of structures...



alpha-hemolysin (7AHL)

TolC (1EK9)



MspA (1UUN)

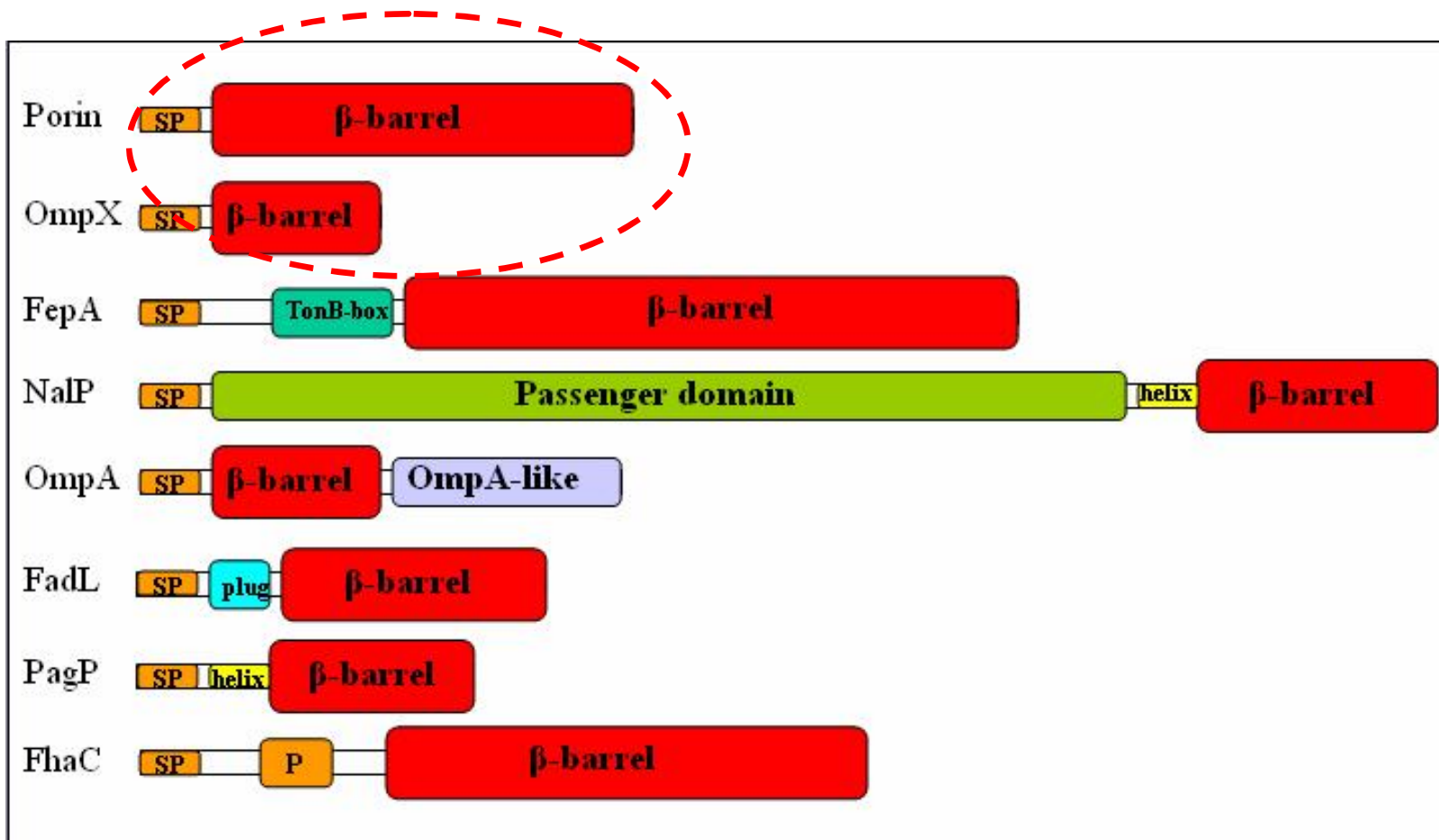
and functions...

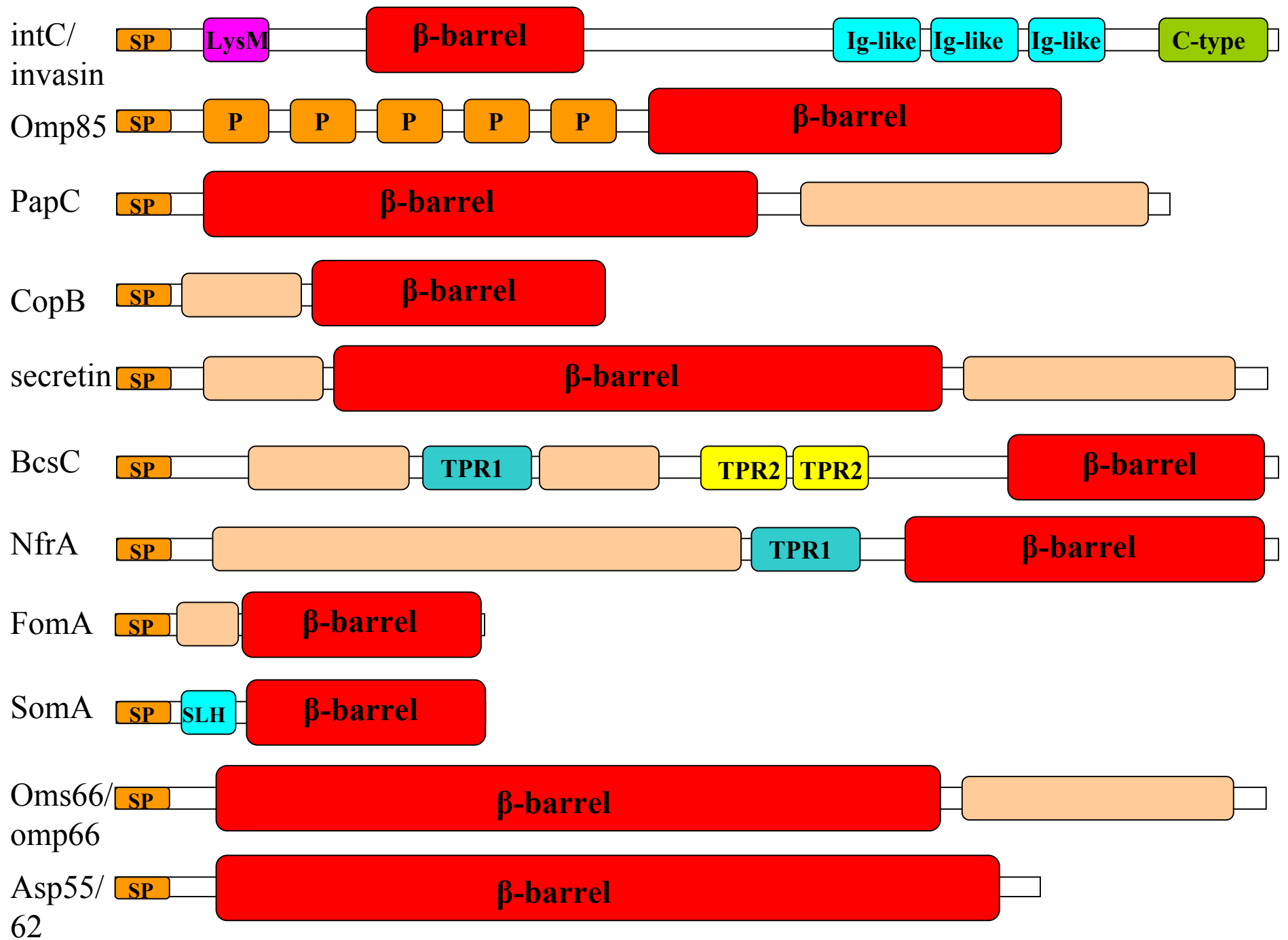
- Specific and non-specific channels (porins)
- Receptors for passive and active intake (TonB-dependent receptors, FadL, Tsx etc)
- Adhesion molecules (OmpX, NspA, OpcA)
- Structural proteins-interactions with peptidoglycan (OmpA)
- Outer membrane enzymes (OmpT, OmpLA, PagP, PagL)
- Protein secretion in nearly all secretory pathways (Secretins, Ushers, autotransporters, TPS etc)
- Folding and assembly of membrane proteins (Omp85/Sam50)
- Assembly of the outer membrane-LPS delivery (Imp/OstA)

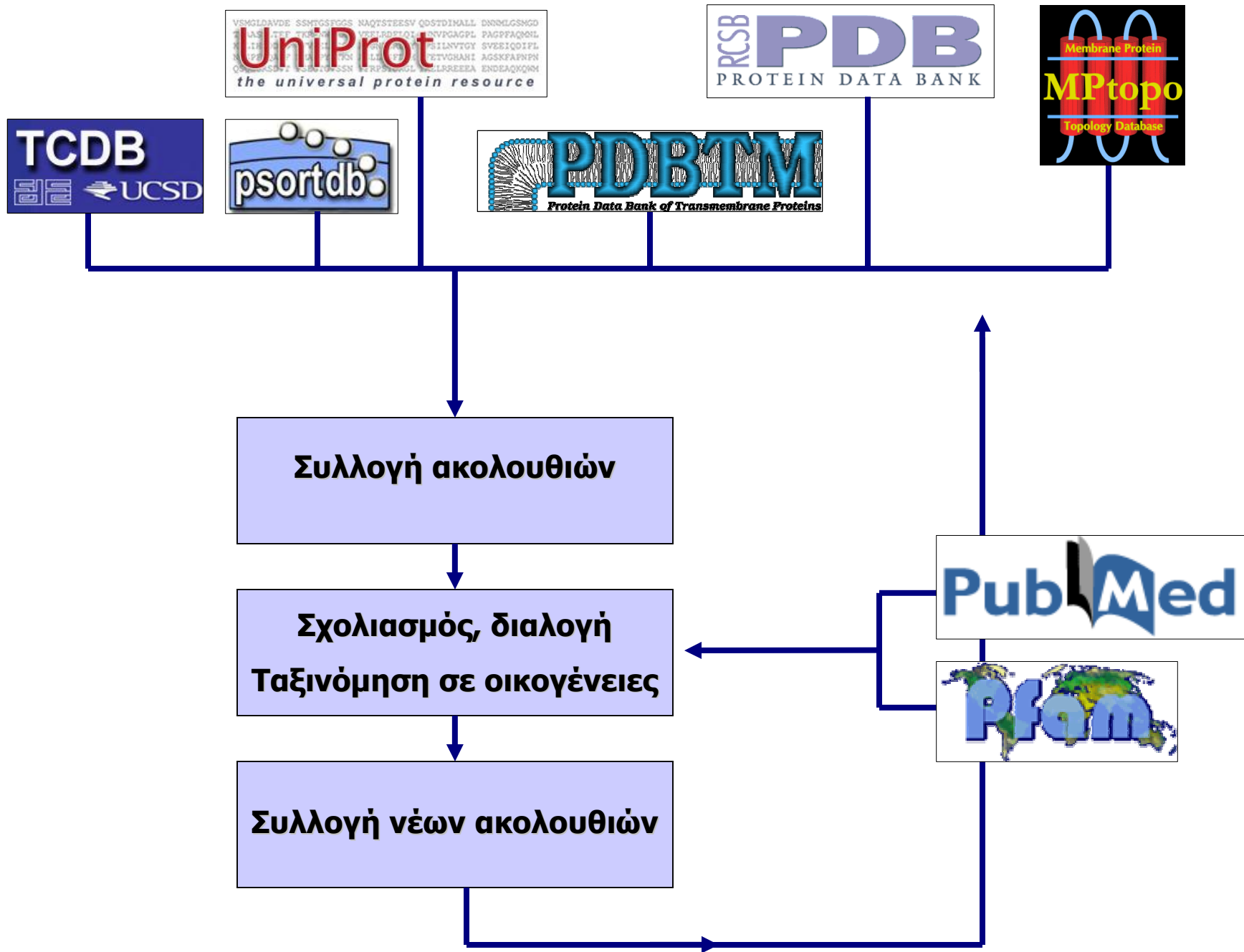
Representative TM beta-barrels of known structure

Protein name	function	Number of β -strands	PDB code	PFAM code	Organism
OmpA	Structural protein	8	1QJP	PF01389	<i>Escherichia coli</i>
OmpX	Adhesion	8	1QJ8	PF06316	<i>Escherichia coli</i>
NspA	Adhesion	8	1P4T	PF02462	<i>Neisseria Meningitidis</i>
PagP	Enzyme	8	1MM4	PF07017	<i>Escherichia coli</i>
PagL	Enzyme	8	2ERV	PB038312 *	<i>Pseudomonas aeruginosa</i>
OmpW	General Porin	8	2F1T	PF03922	<i>Escherichia coli</i>
OmpT	Enzyme	10	1I78	PF01278	<i>Escherichia coli</i>
OpcA	Adhesion	10	1K24	PF07239	<i>Neisseria Meningitidis</i>
OmpLA	Enzyme	12	1QD5	PF02253	<i>Escherichia coli</i>
NalP	Autotransporter	12	1UYN	PF03797	<i>Neisseria Meningitidis</i>
Tsx	Transporter	12	1TLY	PF03502	<i>Escherichia coli</i>
OmpG	General Porin	14	2F1C	PB051875 *	<i>Escherichia coli</i>
FadL	Transporter	14	1T1L	PF03349	<i>Escherichia coli</i>
OprP	General Porin	16	2O4V	PF07396	<i>Pseudomonas aeruginosa</i>
OmpF	General Porin	16	2OMF	PF00267	<i>Escherichia coli</i>
FhaC	Transporter (TPS)	16	2QDZ	PF03865	<i>Bordetella pertussis</i>
Porin	General Porin	16	2POR	PB028487 *	<i>Rhodobacter capsulatus</i>
Maltoporin	Specific Porin	18	2MPR	PF02264	<i>Salmonella typhimurium</i>
FepA	TonB-dependent Receptor	22	1FEP	PF00593	<i>Escherichia coli</i>

**Bagos PG, Hamodrakas SJ.
2007, submitted**







a) single-domain OMPs



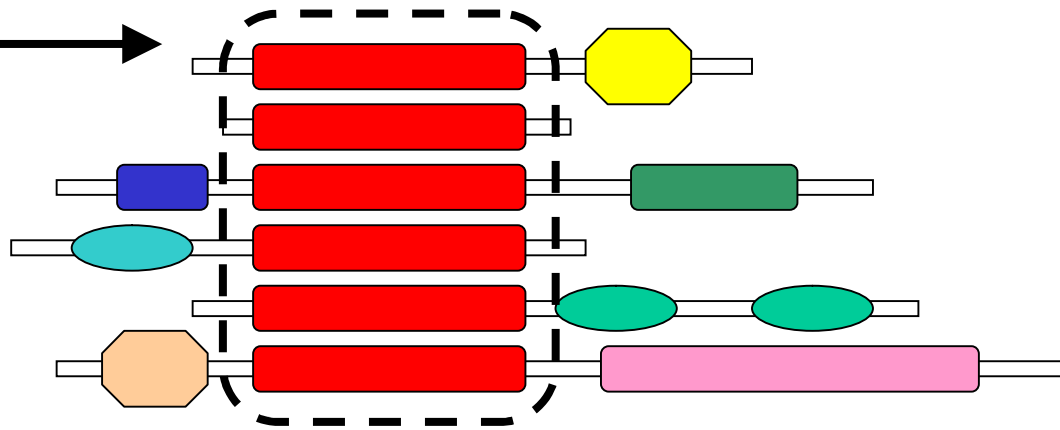
or

b) multi-domain OMPs



beta-barrel domain

database search (BLAST)



Until no
new family
members
are found

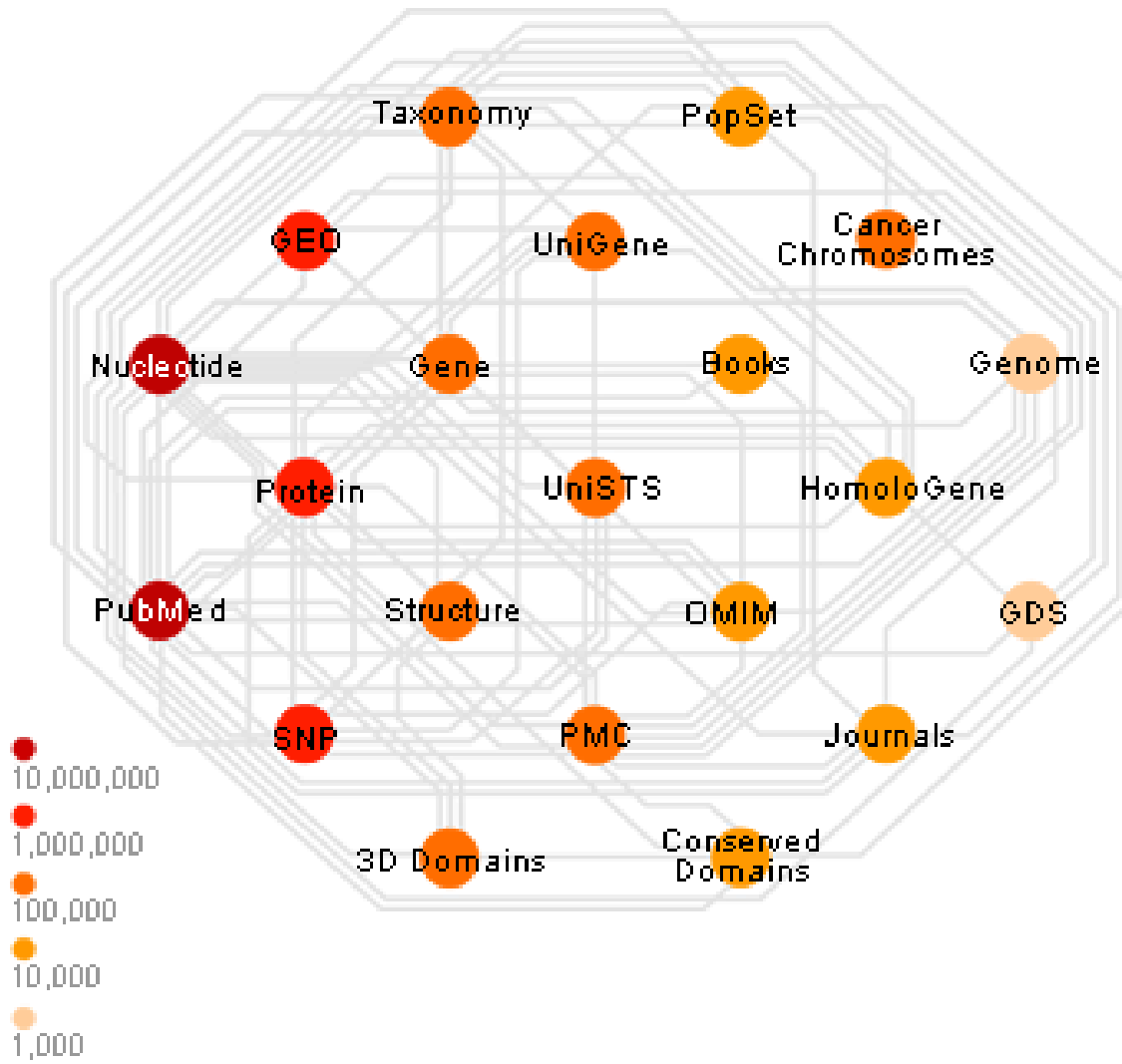
multiple alignment (ClustalW)

profile HMM (HMMER)



Entrez

- Το **Entrez** αποτελεί ένα σύστημα διαχείρισης δεδομένων για την αναζήτηση και ανάκτηση πληροφοριών όλων των βάσεων δεδομένων που περιέχονται στο NCBI (National Center for Biotechnology Information) των ΗΠΑ.
- Το Entrez είναι ανάλογο του SRS και παρέχει στον χρήστη τη δυνατότητα αναζήτησης σε βάσεις δεδομένων νουκλεοτιδικών και πρωτεϊνικών ακολουθιών, δομές βιομορίων και γονιδιωμάτων.
- Επιπλέον, μέσω του ίδιου γραφικού περιβάλλοντος, παρέχει την δυνατότητα αναζήτησης στη βάση βιβλιογραφίας PUBMED καθώς και πιο πολύπλοκες αναζητήσεις ανάμεσα στα στοιχεία τους.
- Βασικό μειονέκτημα αποτελεί το γεγονός ότι περιορίζεται μόνο στις βάσεις δεδομένων του NCBI και ότι δεν επιτρέπει ιδιαίτερα πολύπλοκες αναζητήσεις.
- Παρόλα αυτά, αποτελεί για χρόνια τώρα την διεπαφή όλων των βάσεων δεδομένων του NCBI, και επιτρέπει με τον ίδιο απλό τρόπο ο χρήστης να πραγματοποιήσει αναζητήσεις σε τελείως διαφορετικές βάσεις δεδομένων



SRS (Sequence Retrieval System)



Quick Search Library Page Query Form Tools Results Projects Views Databanks **HELP**

SRS **LION**

Start a Permanent Project

Tips

★ *Want to know more about using SRS?*

- go to the [Help Center](#) where you'll find all the searchable online help you need.

★ *Where is the old library page?*

- Click on the 'Library Page' tab on the menu bar.

★ *Linking to SRS?*

- Please read our [Linking to SRS](#) guide for important information regarding linking to our SRS server.

★ [Public SRS servers worldwide](#)

Quick Text Search

[Search Tips](#)

Get matching :

Searches Databanks: EMBL Nucleotides

Search

News and Announcements

[Search Tips](#)

Important notes to all users.

Note: from EMBL release 88 entries are identified only by their accession number. To find the EMBL accession corresponding to a pre-release 88 EMBL identifier please use the [EMBL ID/Accession Mapping](#) databank.

25.07.06 [EMBL WGS Masters](#), a databank describing all the sequences derived from a whole genome shotgun project, is now available. If you have any problems please let us know via the [EBI Support form](#).

12.07.06 [EMBL MGA](#), a databank of sets of short sequences useful for genome annotation, is now available. If you have any problems



SRS is a product of Lion Bioscience AG

Search Options

1. Select the **databanks** you want to search
2. Enter your **search terms** in the **Quick Search** box, or choose a **query form** from below

[Standard Query Form](#)

[Extended Query Form](#)

You can **browse** through all the **entries** in any **databanks**. First, **select** the **databanks** you want to browse, then click:

[Browse Entries](#)

Tips

- ▶ bookmark this [link](#) to return to your project
- ▶ [Linking to SRS?](#)
- Please read our [Linking to SRS](#) guide for important information regarding linking to our SRS server.

BookMarkLets

[About BookmarkLets](#)

- ◆ [Protein Seq](#)
- ◆ [DNA/RNA Seq](#)
- ◆ [Structures](#)

Available Databanks

[Expand all](#) [Collapse all](#)

Show databanks tooltips:

Literature, Bibliography and Reference Databases

[all](#) [TAXONOMY](#) [GENETICCODE](#) [OMIM](#) [MEDLINE](#)

[Patent Abstracts](#) [Karyn's Genomes](#)

Literature, Bibliography and Reference Databases - subsections

[all](#) [MEDLINE \(Updates\)](#) [MED2PUB](#) [MEDLINE \(Main Release 2006\)](#)

Gene Dictionaries and Ontologies

Nucleotide sequence databases

[all](#) [EMBL](#) [Patent DNA](#) [IMGTLIGM-DB](#) [IMGT/HLA](#)

[IPD-KIR](#) [EMBL \(Contig\)](#) [EMBL \(Contigs expanded\)](#) [EMBL \(Annotated Cons\)](#)

[EMBL \(Coding Sequences\)](#) [Genome Reviews](#) [RefSeq Genome](#) [LiveLists](#)

[EMBL ID/Accession Mapping](#) [EMBL MGA](#)

Nucleotide sequence databases - subsections

[all](#) [EMBL \(Updates\)](#) [EMBL \(Release\)](#) [EMBL \(Whole Genome Shotgun\)](#)

[EMBL \(Whole Genome Shotgun release\)](#) [EMBL \(Whole Genome Shotgun updates\)](#) [EMBL \(Contig release\)](#)

[EMBL \(Contig updates\)](#) [EMBL \(Contigs expanded release\)](#) [EMBL \(Contigs expanded updates\)](#)

[EMBL \(Annotated Cons release\)](#) [EMBL \(Annotated Cons updates\)](#) [RefSeq Genome \(Release\)](#)

[RefSeq Genome \(Updates\)](#) [EMBL \(Whole Genome Shotgun Masters\)](#)

Nucleotide related databases

UniProt Universal Protein Resource

[all](#) [UniProtKB](#) [UniProtKB/Swiss-Prot](#) [UniProtKB/TrEMBL](#) [UniRef100](#) [UniRef90](#)

[UniRef50](#) [UniParc](#)

Other protein sequence databases

Protein function, structure and interaction databases

Enzymes, reactions and metabolic pathway databases

Mutation and SNP databases

Biological Resources Catalogues (CABRI)

Mapping databases

Other databases

User owned databases

Application result databases

FMOSS result databases

Reset

search [UniProtKB/Swiss-Prot](#)
[UniProtKB/TrEMBL](#)

Search Options

Combine search terms

with:

Use wildcards

Get results of type:

Result Display Options

View results using:

or

Create a view

Show

results per page

Tips

Fields you can search

Your search terms

In a single field, you can separate multiple values by &, |, !

<input type="checkbox"/> ID	<input type="text"/>
<input type="checkbox"/> Description	<input type="text"/>
<input type="checkbox"/> Gene Name	<input type="text"/>
<input type="checkbox"/> AllText	<input type="text"/>

Create a view

Select the fields you want displayed in your view and choose the format

Choose 1 or more fields:

ID
 EntryName
 AccessionNumber
 Creation Date
 Seq Mod Date
 Annot Mod Date
 Description

Display As: Table List

Sequence Format: