The Abdus Salam
International Centre for Theoretical Physics

2169-3

Conference on Molecular Aspects of Cell Biology: A Perspective from Computational Physics

11 - 15 October 2010

Hands-on Session on Structural Bioinformatics

A. GIORGETTI
Universita' degli Studi di Verona
Dipt. di Biotecnologie
Verona
Italy

Hands-on session on Bioinformatics

Alejandro Giorgetti

Dept. of Biotechnology

University of Verona

http://molsim.sci.univr.it/bioinfo

Molecular system-level understanding

- Small molecules control a myriad of cellular functions by binding to their target macromolecules: ligands govern processes such as growth, programmed cell death, sensing, and metabolism. This key event triggers complex cellular pathways characterized by reactions, environmental changes, intermolecular interactions, and allosteric modifications

- Ultimately, understanding the molecular basis of ligand-target interactions requires the integration of biological complexes into cellular pathways: "systems biology"

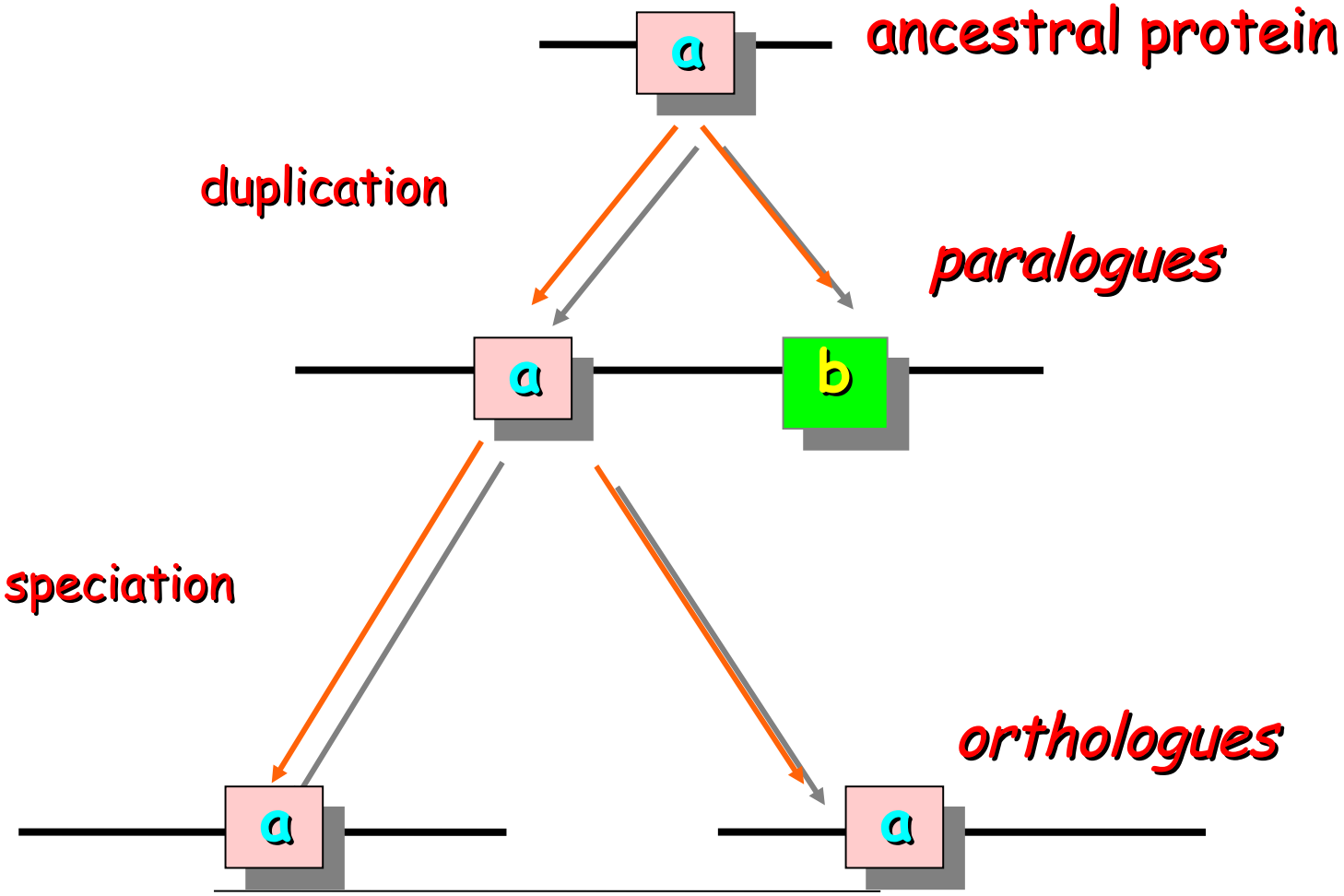
- All of these processes involve molecular recognition

- Nonmolecular modeling, which is advancing tremendously our understanding needs to be paralleled by a quantitative molecular description of pathways

Computational molecular biology methods

- The first strategy, the so-called ***protein bioinformatics***, is aimed at the development of computational tools that enable to decipher the information encoded in the protein sequences, thus enabling the prediction of structure and function
- The second strategy is based on the laws of physics. One of the most important methods here is ***molecular dynamics*** (MD) , which predicts structural, dynamical, and energetic (bio)molecular properties based on Newton laws of motion.

Homology based inference of protein functions



- orthologues - often have very similar functions
- paralogues - may have related functions

Function Prediction

Molecular function

Family characterization
HMM, profiles
SMART, ProtoNet, Everest,
Gene3D, CATH, InterPro
Pfam, TIGR, PRINTS, SCOP

Orthologs identification
HAMAP, EggNogg, COGS,
KOGS

search for conserved
residues
TreeDet, ScoreCons,
GEMMA, ProteinKeys
ETtrace, SCI-PHY,
FunShift

Cellular component

Residues Features
predict transmembrane,
localisation etc

analyse residue features
predict disorder, signal
peptides, localisation
Barcello, DisoPred, FFpred

Biological process

Protein Structure
prediction
Comparative Modeling
Fold recognition
New fold

predict protein
interactions:
ligand-protein
protein-protein

State of art Free-ware bioinformatics programs !!!

Activity presentation

- ***Introduction to bioinformatics*: from sequence to structural models.**
- ***Practical session:* Blast, Psi-blast, PSI-search, ClustalW/Muscle (jalview), PROMALS, TreeDet, Hhpred, Swiss-Model, GeneSilico. Programs for protein visualization.**
- **Protein structure analysis. *Participant cases of interest !!***

Sequence alignment

Problem: optimize the corrispondence between aminoacids from two (or more) different sequences.

Protein Evolution

2 Sequences

```
AGTTTGAATGTTTTGTGTGAAAGGAGTATACCATGAGATGAGATGACCACCAATCATTTC
|||||||||||||||||| ||||||| ||| | |||||| ||||||||||||||||
AGTTTGAATGTTTTGTGTGTGAGGAGTATTCCAAGGGATGAGTTGACCACCAATCATTTC
```

Multiple

```
KFKHHLKEHLRIHSGEKPFECPNCKKRFSHSGSYSSHMSSKKCISLILVNGRNRALLKTl
KYKHHLKEHLRIHSGEKPYECPNCKKRFSHSGSYSSHISSKKCIGLISVNGRMRNNIKT-
KFKHHLKEHVRIHSGEKPFGCDNCGKRFSHSGSFSSHMTSKKCISMGLKLNNNRALLKRl
KFKHHLKEHIRIHSGEKPFECQQCHKRFSHSGSYSSHMSSKKCV----------------
KYKHHLKEHLRIHSGEKPYECPNCKKRFSHSGSYSSHISSKKCISLIPVNGRPRTGLKTsn
```

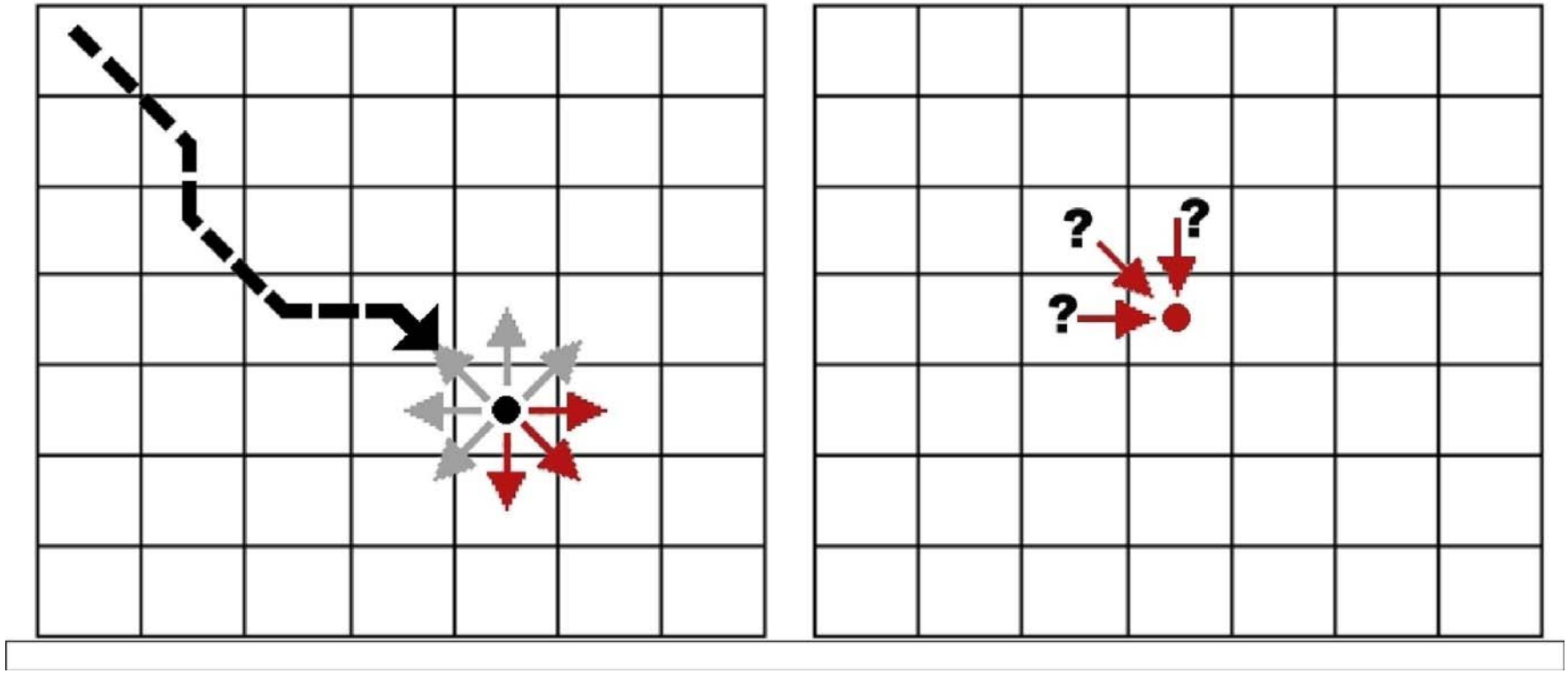
- We need

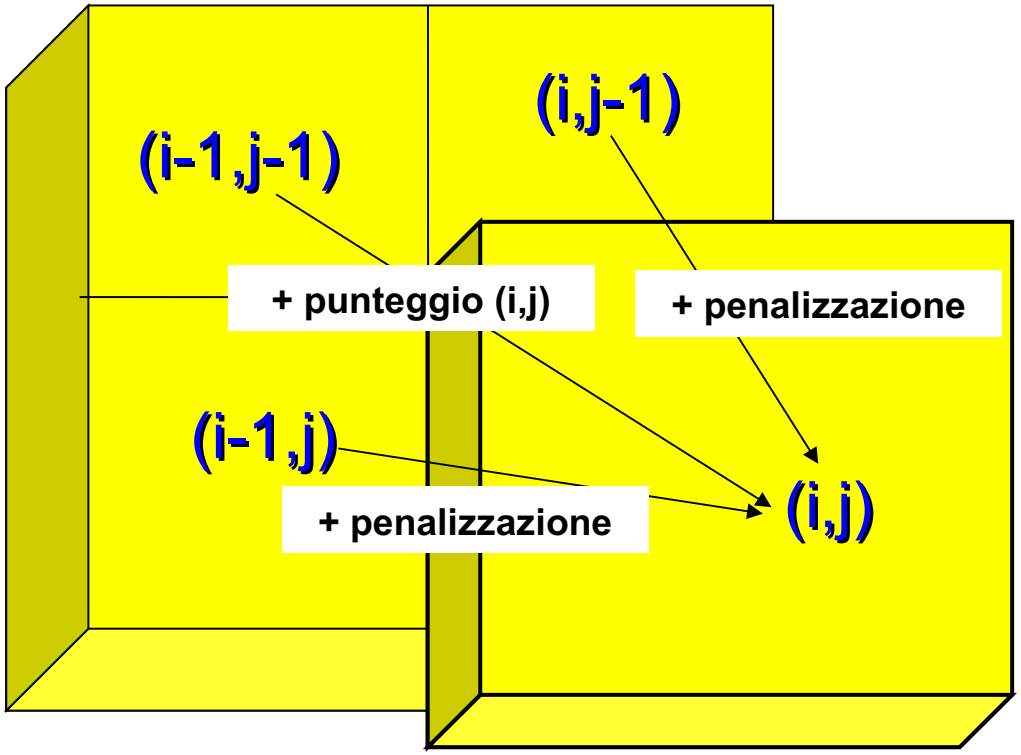
- A method for scoring similarities between aminoacids. Substitution matrices: Blosum, PAM, Gonnet....

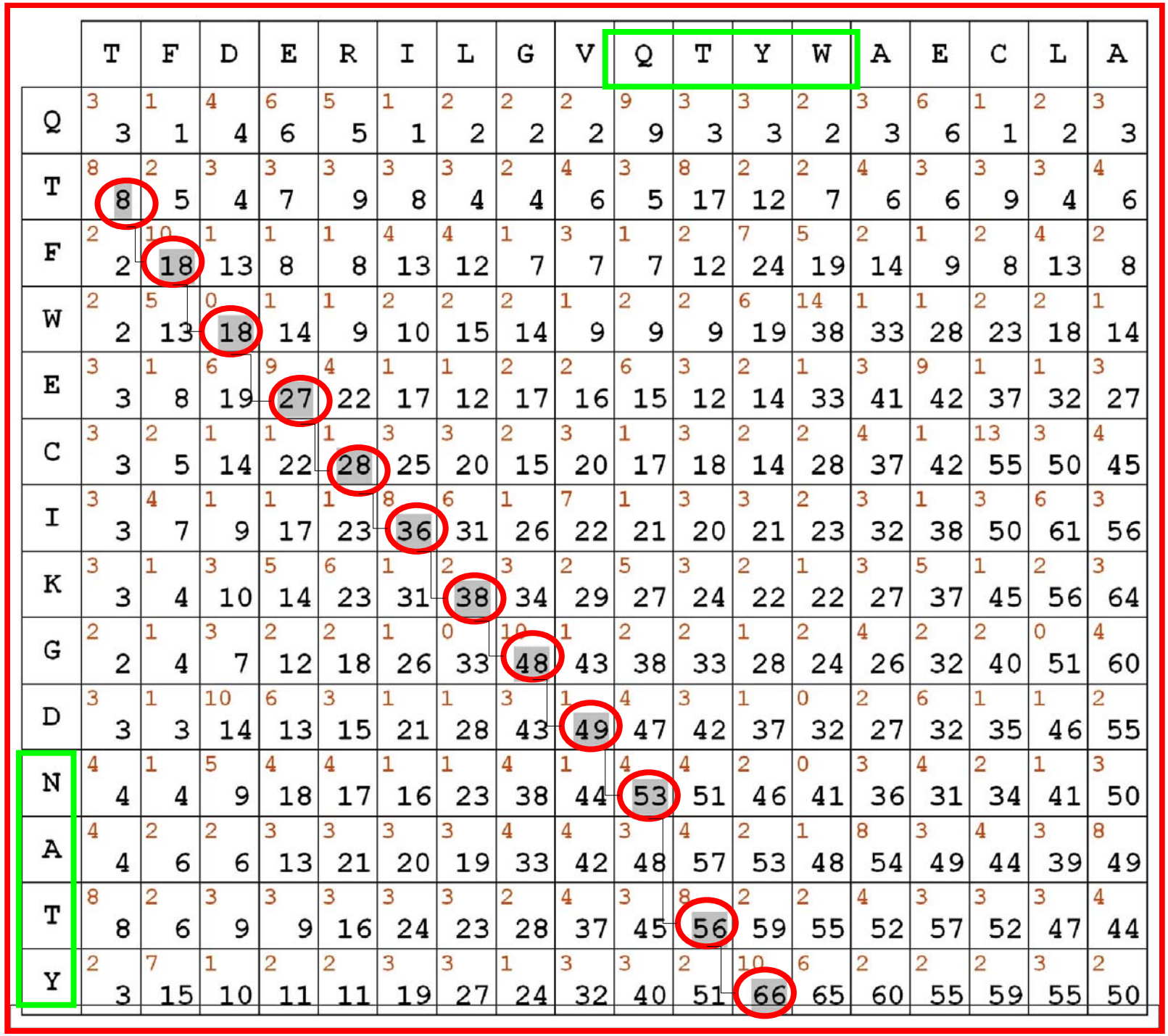
- A penalty value for the insertions and deletions (gaps)

- An algorithm for the alignment. Dynamic programing: Needelman e Wunsch (global alignment) or Smith e Waterman (local alignment)

	T	F	D	E	R	I	L	G	V	Q	T	Y	W	A	E	C	L	A
Q	3 3	1 1	4 4	6 6	5 5	1 1	2 2	2 2	2 2	9 9	3 3	3 3	2 2	3 3	6 6	1 1	2 2	3 3
T	8 8	2 5	3 4	3 7	3 9	3 8	3 4	2 4	4 6	3 5	8 17	2 12	2 7	4 6	3 6	3 9	3 4	4 6
F	2 2	10 18	1 13	1 8	1 8	4 13	4 12	1 7	3 7	1 7	2 12	7 24	5 19	2 14	1 9	2 8	4 13	2 8
W	2 2	5 13	0 18	1 14	1 9	2 10	2 15	2 14	1 9	2 9	2 9	6 19	14 38	1 33	1 28	2 23	2 18	1 14
E	3 3	1 8	6 19	9 27	4 22	1 17	1 12	2 17	2 16	6 15	3 12	2 14	1 33	3 41	9 42	1 37	1 32	3 27
C	3 3	2 5	1 14	1 22	1 28	3 25	3 20	2 15	3 20	1 17	3 18	2 14	2 28	4 37	1 42	13 55	3 50	4 45
I	3 3	4 7	1 9	1 17	1 23	8 36	6 31	1 26	7 22	1 21	3 20	3 21	2 23	3 32	1 38	3 50	6 61	3 56
K	3 3	1 4	3 10	5 14	6 23	1 31	2 38	3 34	2 29	5 27	3 24	2 22	1 22	3 27	5 37	1 45	2 56	3 64
G	2 2	1 4	3 7	2 12	2 18	1 26	0 33	10 48	1 43	2 38	2 33	1 28	2 24	4 26	2 32	2 40	0 51	4 60
D	3 3	1 3	10 14	6 13	3 15	1 21	1 28	3 43	1 49	4 47	3 42	1 37	0 32	2 27	6 32	1 35	1 46	2 55
N	4 4	1 4	5 9	4 18	4 17	1 16	1 23	4 38	1 44	4 53	4 51	2 46	0 41	3 36	4 31	2 34	1 41	3 50
A	4 4	2 6	2 6	3 13	3 21	3 20	3 19	4 33	4 42	3 48	4 57	2 53	1 48	8 54	3 49	4 44	3 39	8 49
T	8 8	2 6	3 9	3 9	3 16	3 24	3 23	2 28	4 37	3 45	8 56	2 59	2 55	4 52	3 57	3 52	3 47	4 44
Y	2 3	7 15	1 10	2 11	2 11	3 19	3 27	1 24	3 32	3 40	2 51	10 66	6 65	2 60	2 55	2 59	3 55	2 50

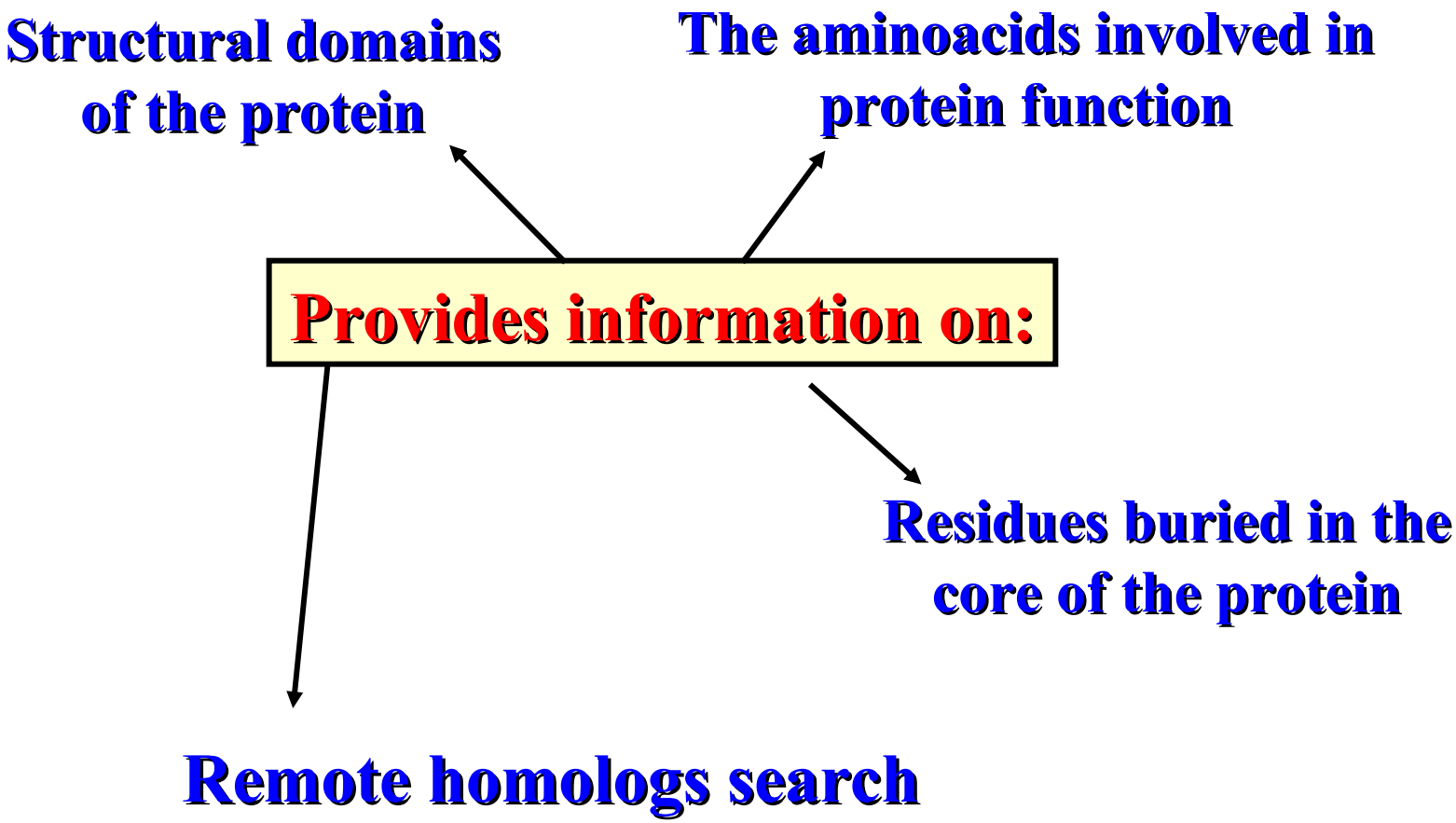

?
?
?

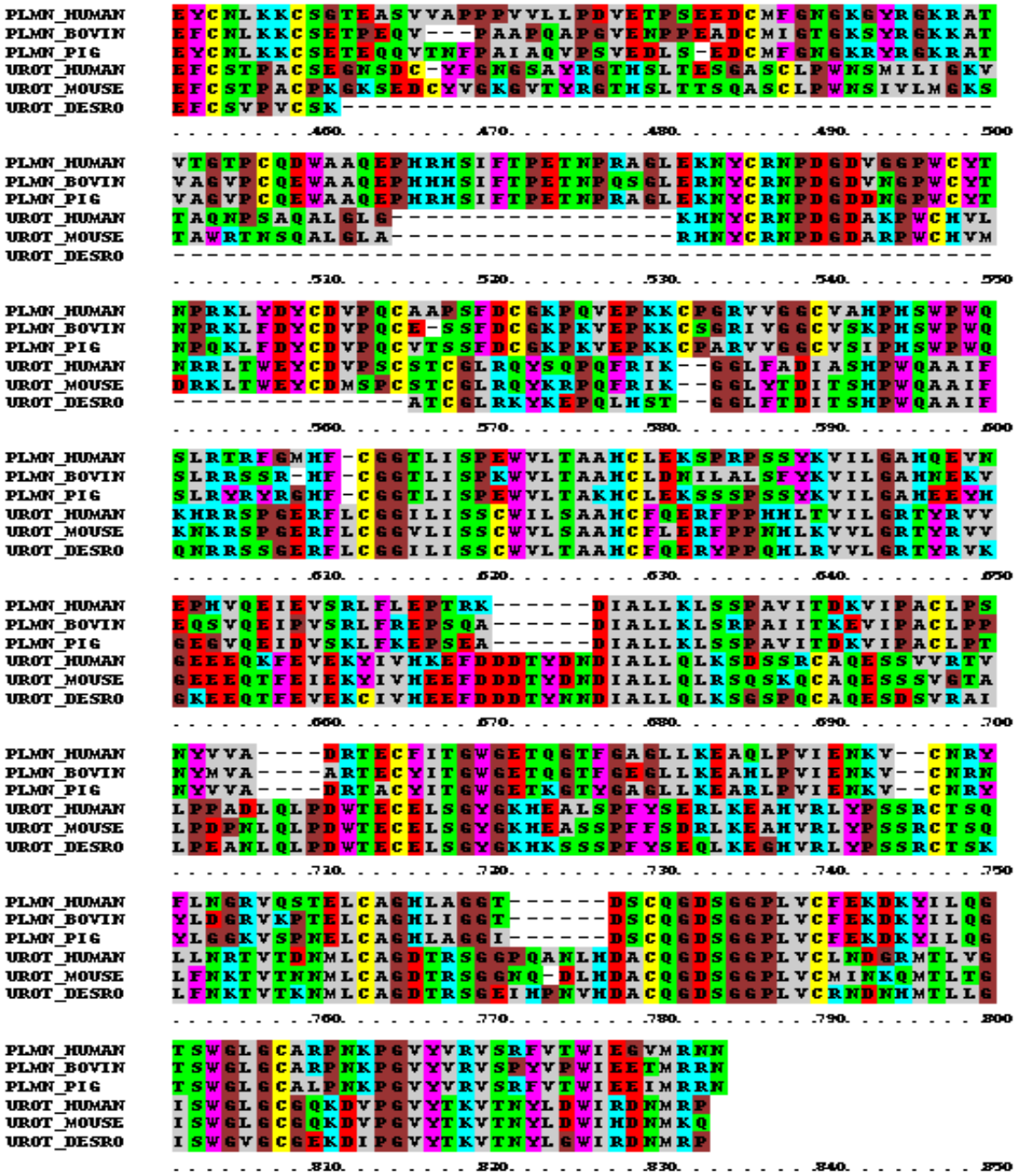

(i-1,j-1)
(i,j-1)
+ punteggio (i,j)
+ penalizzazione
(i-1,j)
+ penalizzazione
(i,j)


T F D E R I L G V Q T Y W A E C L A
Q 3 3 1 1 4 4 6 6 5 5 1 1 2 2 2 2 2 9 9 3 3 3 3 2 2 3 3 6 6 1 1 2 2 3 3
T 8 8 2 5 3 4 3 7 3 9 3 8 3 4 2 4 4 6 3 5 8 17 2 12 2 7 4 6 3 6 3 9 3 4 4 6
F 2 2 10 18 1 13 1 8 1 8 4 13 4 12 1 7 3 7 1 7 2 12 7 24 5 19 2 14 1 9 2 8 4 13 2 8
W 2 2 5 13 0 18 1 14 1 9 2 10 2 15 2 14 1 9 2 9 2 9 6 19 14 38 1 33 1 28 2 23 2 18 1 14
E 3 3 1 8 6 19 9 27 4 22 1 17 1 12 2 17 2 16 6 15 3 12 2 14 1 33 3 41 9 42 1 37 1 32 3 27
C 3 3 2 5 1 14 1 22 1 28 3 25 3 20 2 15 3 20 1 17 3 18 2 14 2 28 4 37 1 42 13 55 3 50 4 45
I 3 3 4 7 1 9 1 17 1 23 8 36 6 31 1 26 7 22 1 21 3 20 3 21 2 23 3 32 1 38 3 50 6 61 3 56
K 3 3 1 4 3 10 5 14 6 23 1 31 2 38 3 34 2 29 5 27 3 24 2 22 1 22 3 27 5 37 1 45 2 56 3 64
G 2 2 1 4 3 7 2 12 2 18 1 26 0 33 10 48 1 43 2 38 2 33 1 28 2 24 4 26 2 32 2 40 0 51 4 60
D 3 3 1 3 10 14 6 13 3 15 1 21 1 28 3 43 1 49 4 47 3 42 1 37 0 32 2 27 6 32 1 35 1 46 2 55
N 4 4 1 4 5 9 4 18 4 17 1 16 1 23 4 38 1 44 4 53 4 51 2 46 0 41 3 36 4 31 2 34 1 41 3 50
A 4 4 2 6 2 6 3 13 3 21 3 20 3 19 4 33 4 42 3 48 4 57 2 53 1 48 8 54 3 49 4 44 3 39 8 49
T 8 8 2 6 3 9 3 9 3 16 3 24 3 23 2 28 4 37 3 45 8 56 2 59 2 55 4 52 3 57 3 52 3 47 4 44
Y 2 3 7 15 1 10 2 11 2 11 3 19 3 27 1 24 3 32 3 40 2 51 10 66 6 65 2 60 2 55 2 59 3 55 2 50

Multiple sequence alignment

The most important residues at the structural and/or functional level are conserved along evolution, and this can be appreciated from a the alignment of enough members of the family.





Typical output from the ALIGN database – residues are coloured according to the following criteria: red, acidic; blue, basic; green, polar neutral; grey, aliphatic; purple, aromatic; brown, proline or glycine; and yellow, cysteine.

Multiple sequence alignment profile

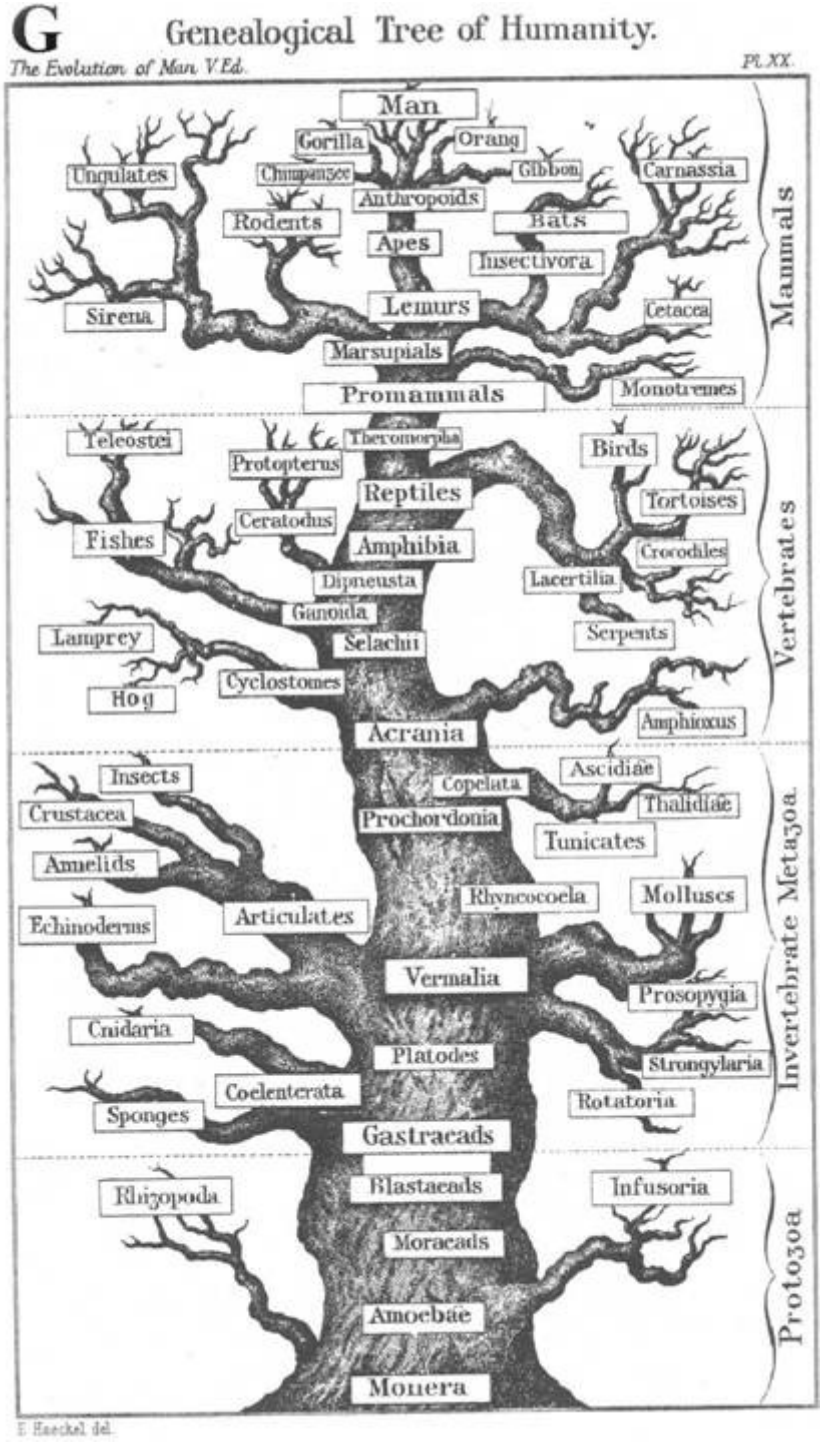
A **profile** shows all the information contained in a multiple sequence alignment

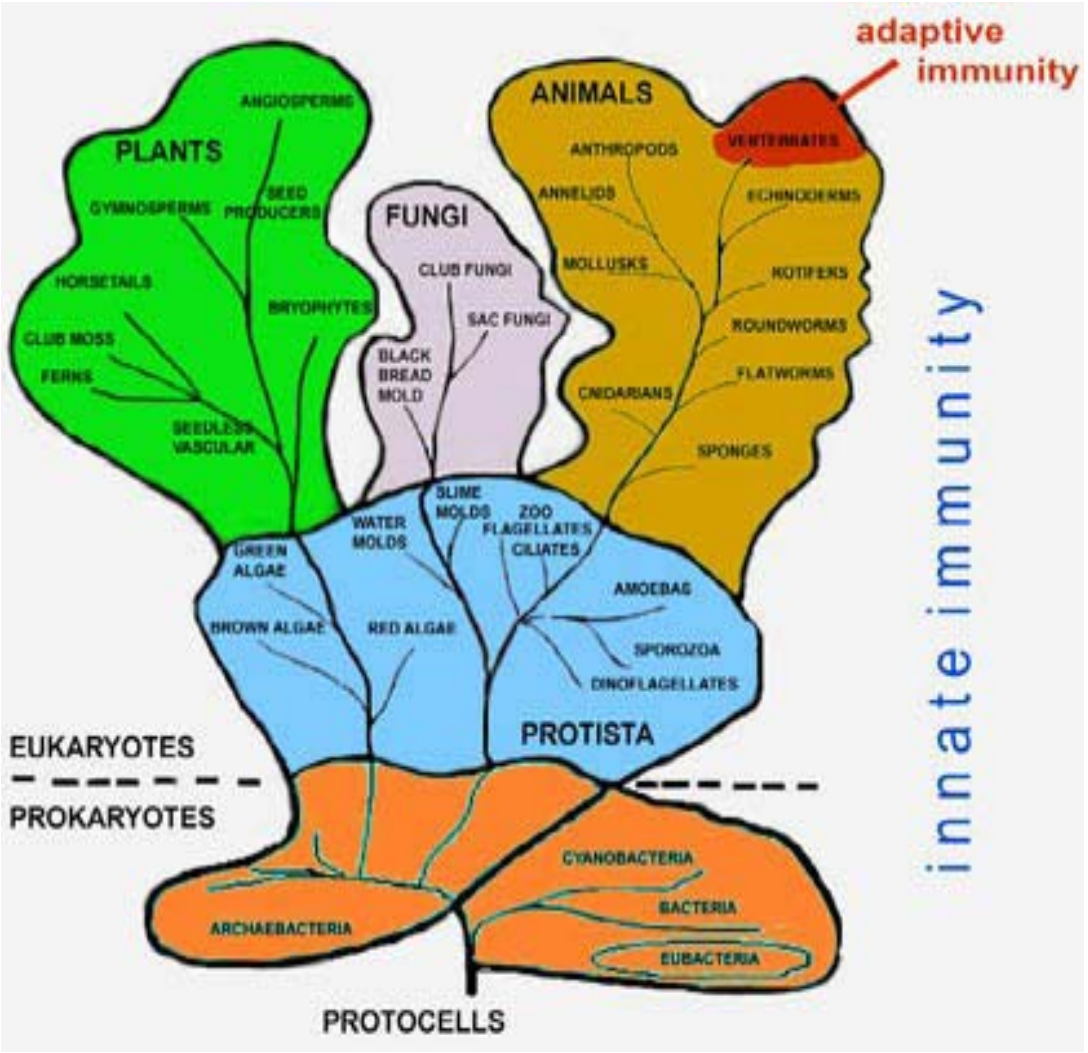
The **profile** is generated by the calculation of the frequencies of each of the aminoacids in all the positions of the alignment.

Used in in PSI-BLAST and PSI-search !!!

The latter uses as algorithm the 'exact' Smith and Waterman alignment protocol

Phylogenetic Trees



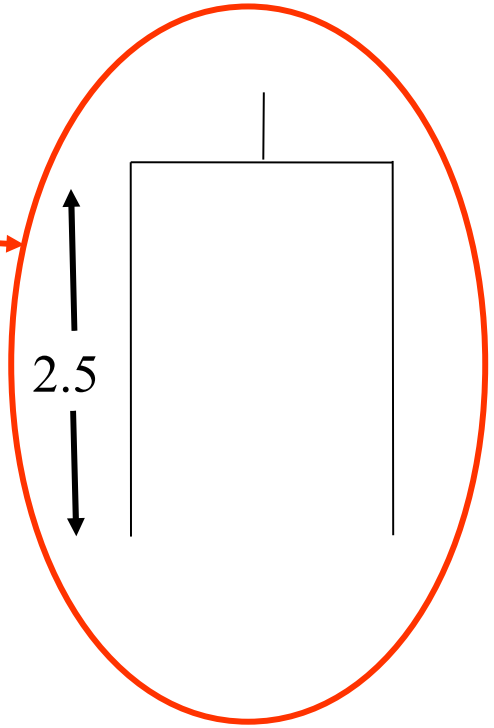


Phylogenetic Trees

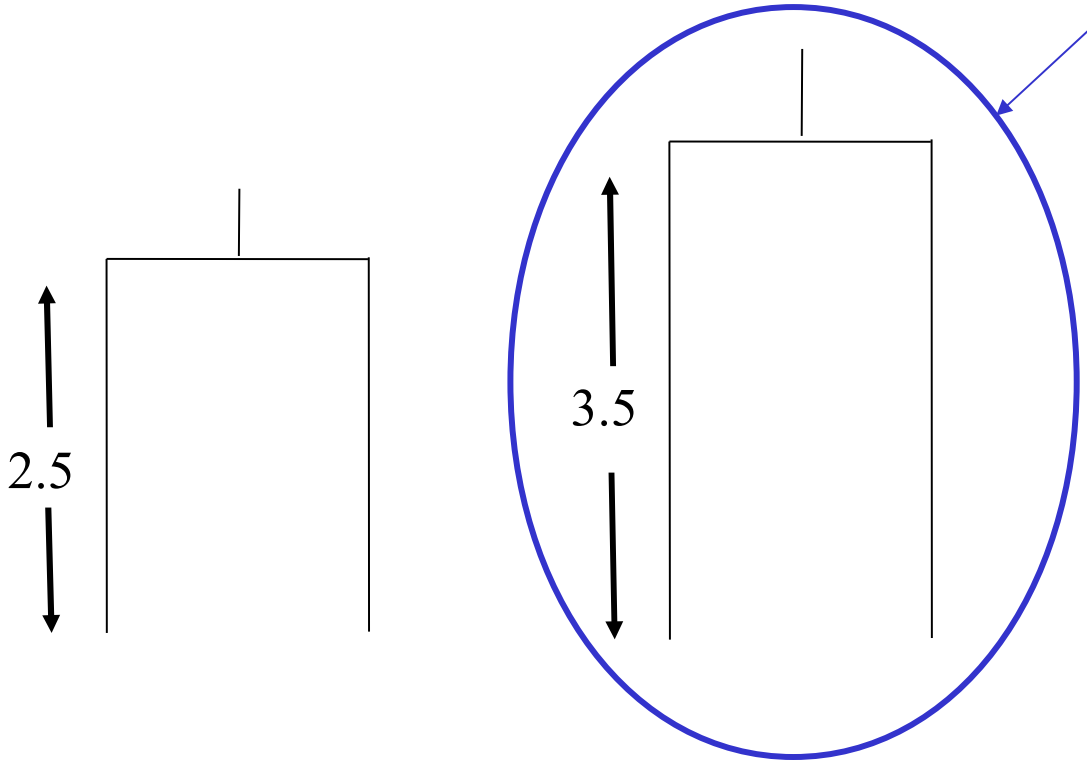
- It's a way of visualizing evolutionary relationships
- Each external node corrisponds to an species
- Internal nodes: speciation event
- The distance between two nodes is proprtional to the divergence time.
- In protein sequences: node -> protein
- The distance between two nodes is proprtional to the sequence simillarity.

Phylogenetic Trees

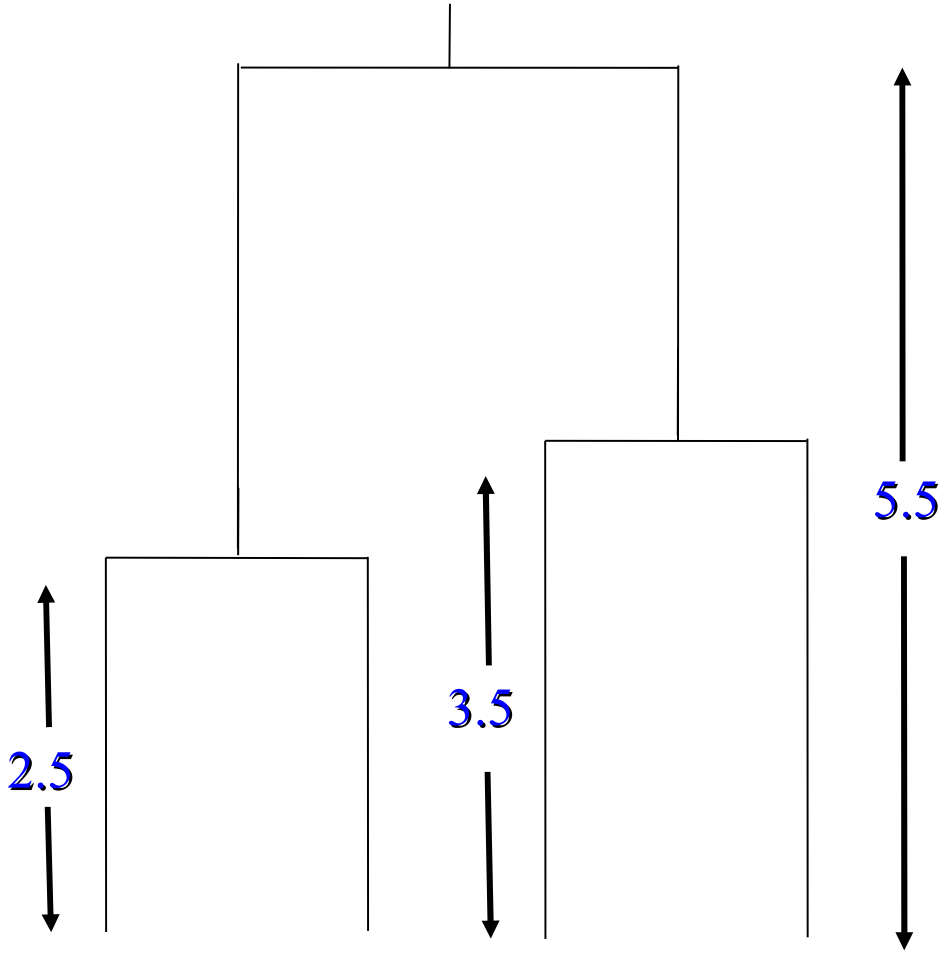
% aa different	Seq1	Seq2	Seq3	Seq4
Seq1	0	5	11	14
Seq2		0	9	10
Seq3			0	7
Seq4				0



% aa different	Cluster 1,2	Seq3	Seq4
Cluster 1, 2	0	½[d(1,3)+d(2,3)]=10	½[d(1,4)+d(2,4)]=12
Seq3		0	7
Seq4			0

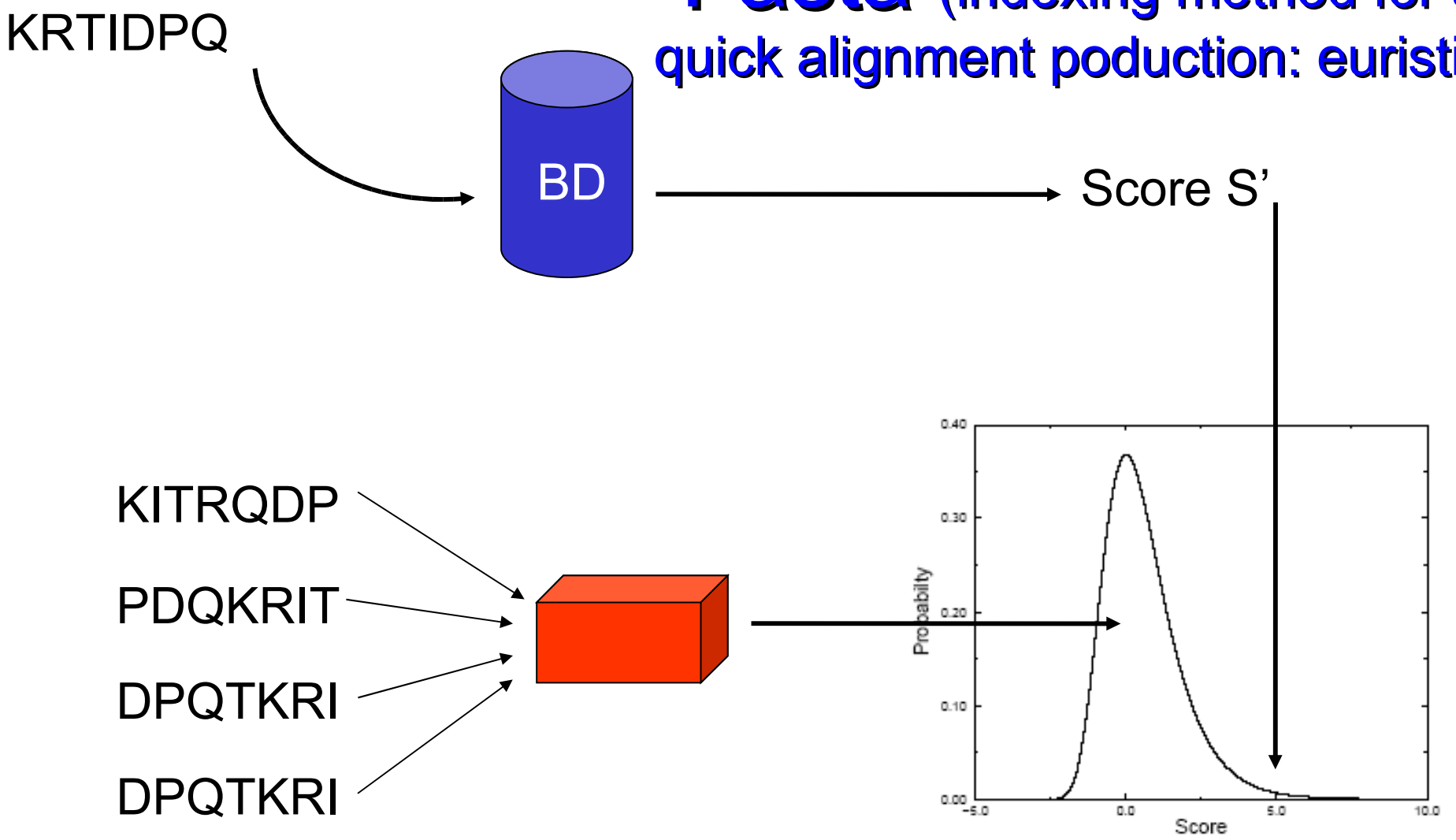


% aa diversi	Cluster 3,4
Cluster 1, 2	=½d[(Cluster 1,2), 3]+d[(Cluster1,2),4)]=11

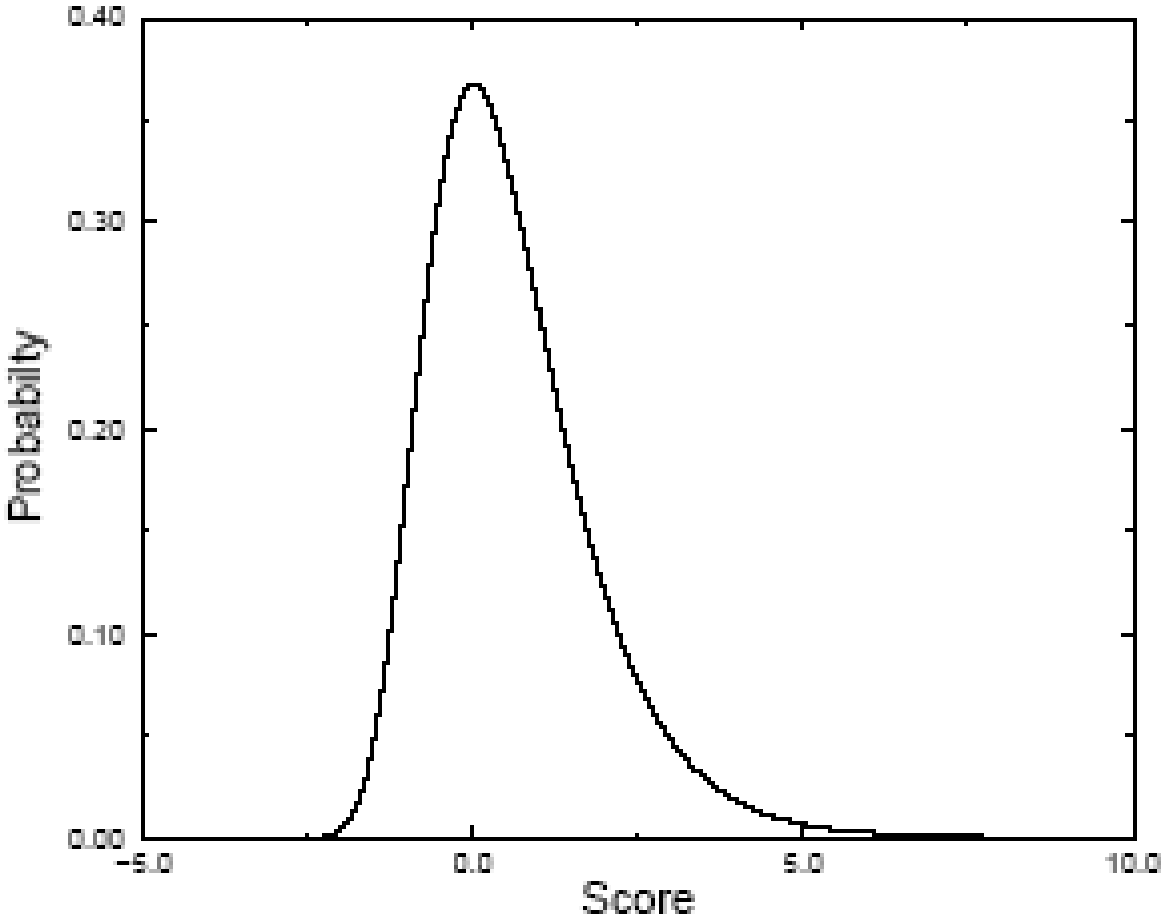


Data-base search for homology

Fasta (indexing method for a quick alignment poduction: euristic)



Extreme value distribution



$$P(>x) = 1 - \exp(-Ke^{-\lambda x})$$

Probability of finding an alignment with score greater than x

K and λ are parameters related to the maximun value position and to the width of the distribution

Extreme value distribution

E-value: expected value !!! is the number of random hit that we expect to obtain with an score S:

$$E = Kmne(-\lambda S)$$

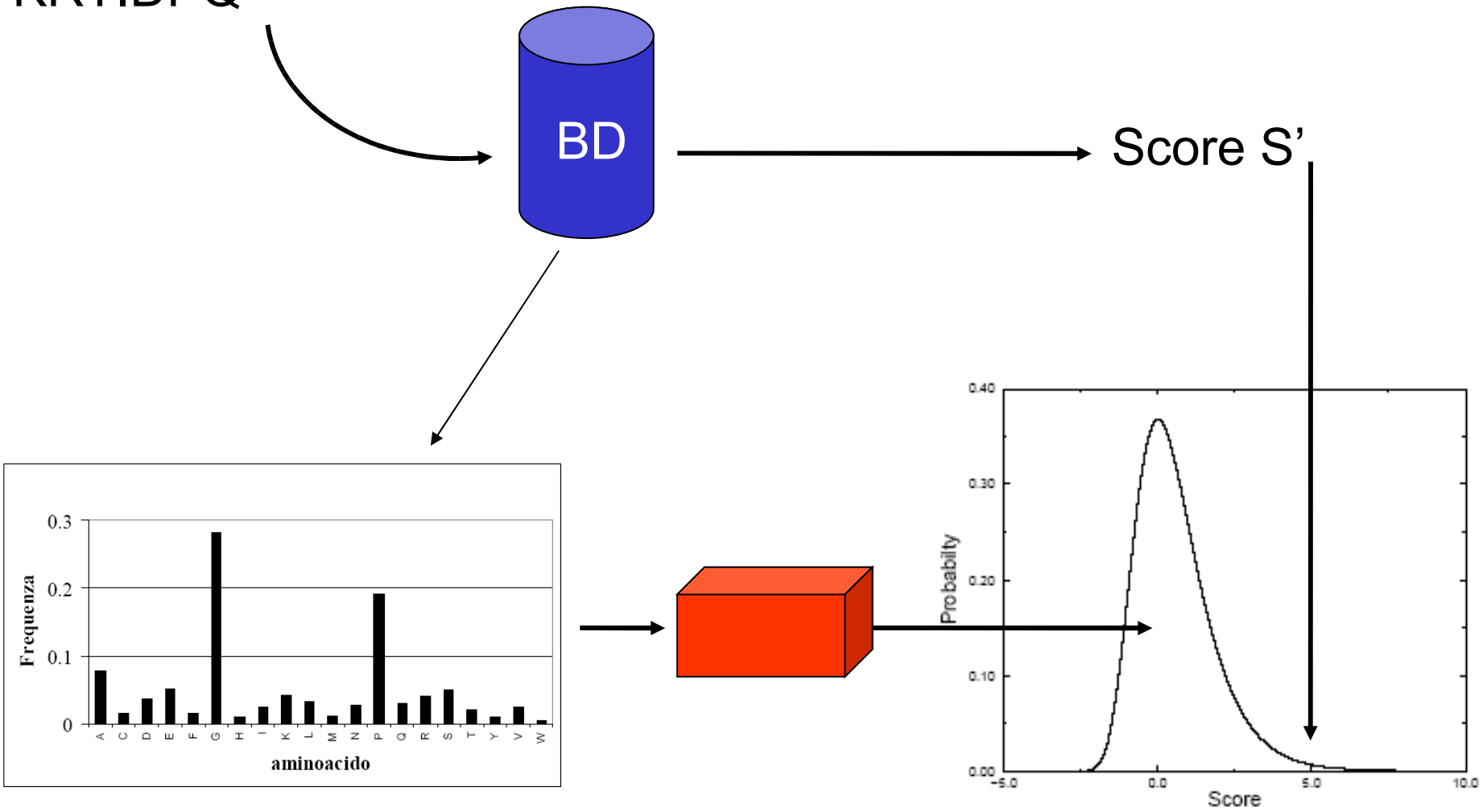
S is generally normalized: $S' = (\lambda S - \ln K)/\ln 2$

S': bit score and therefore: $E = mn2^{-S'}$

Data-base search for homology

Blast

KRTIDPQ

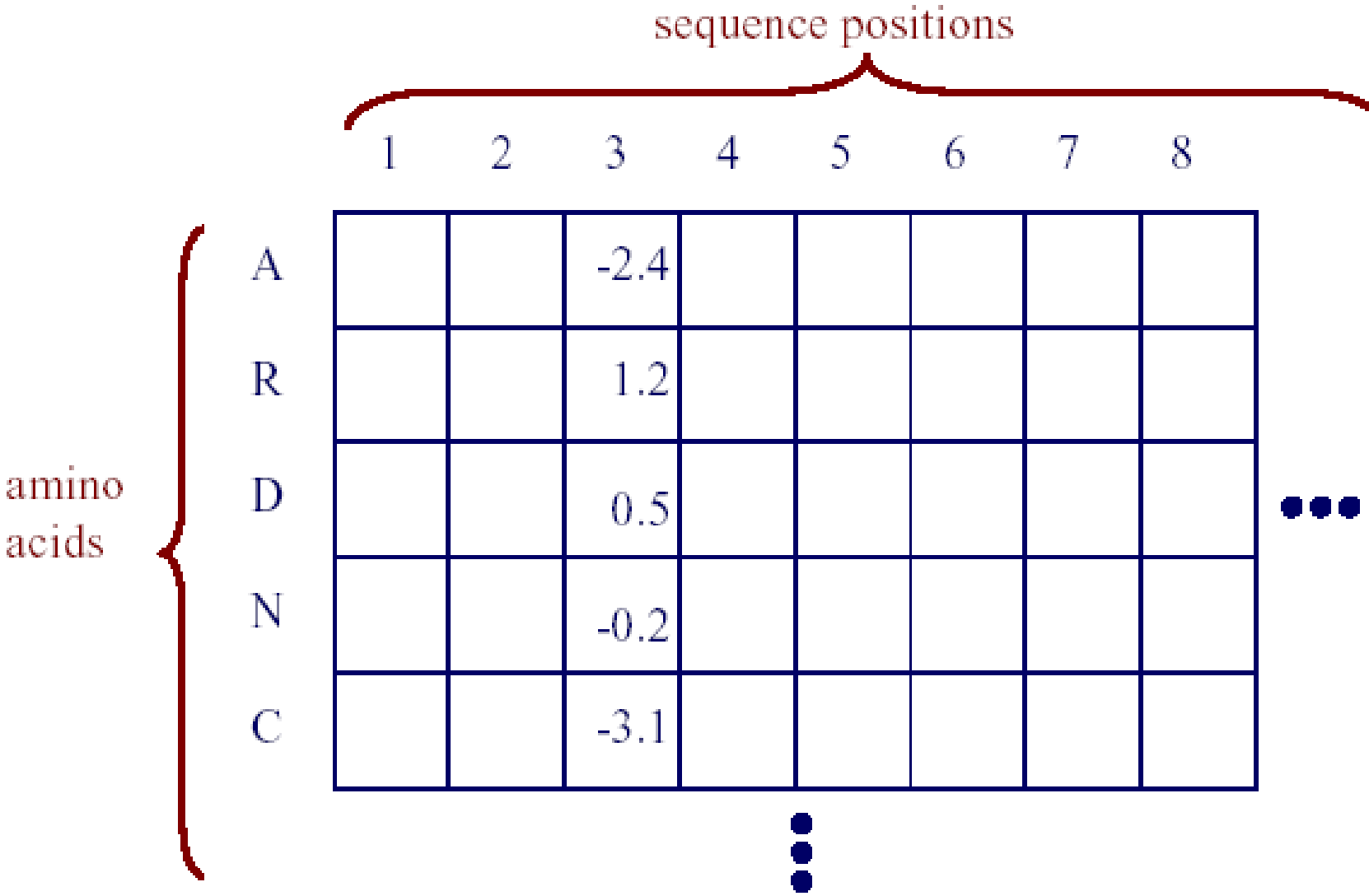


PSI (*Position Specific Iterated*) BLAST

- Idea:
  - to use BLAST results for generating a profile matrix.
  - Search in database using the profiles, instead of the sequence.
- Iterative process:

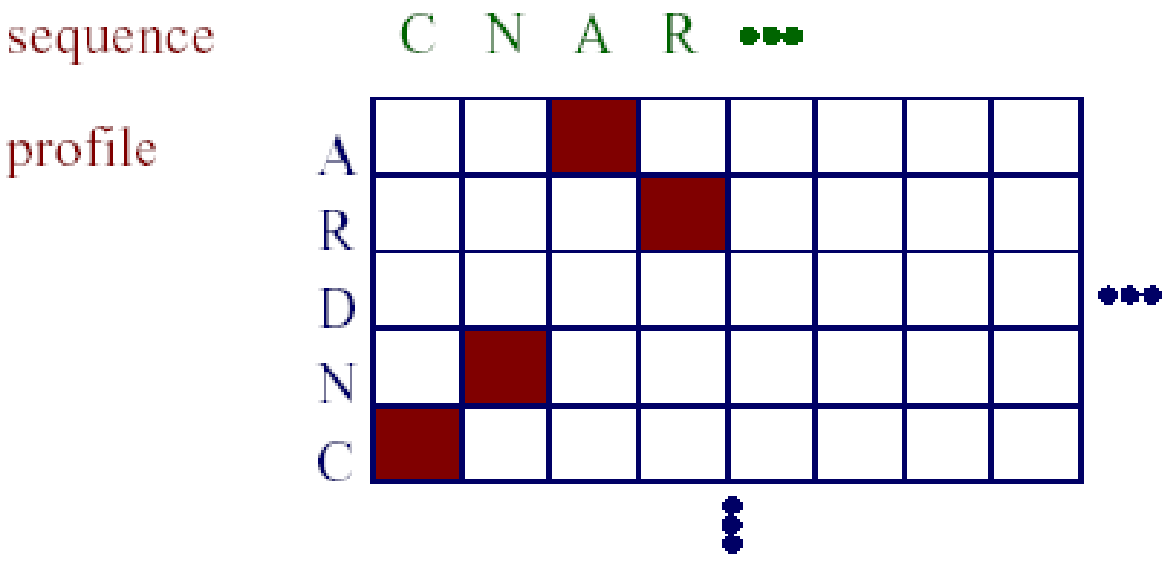
  until convergence is reached ?

Profile Matrix (Position Specific Scoring Matrix – PSSM)



PSI (*Position Specific Iterated*) BLAST

- Profile search
- Alignment of a position specific matrix with a sequence:
  - Is the same as aligning sequences.
  - The score of aligning one character of the position with a character from the matrix, is given by the matrix.
  - There is no need of a substitution matrix



PSI (*Position Specific Iterated*) BLAST

- Value calculation

$$matrix(i,j) = \log\left(\frac{\Pr(a_i \mid \text{col} = j)}{\Pr(a_i)}\right)$$

- Dove $\Pr(a_i|\text{col}=j)$ probability of finding the aminoacid a_i *in column j by chance*
- $\Pr(a_i)$ frequence of finding a_i in the alignment.

Hidden Markov models

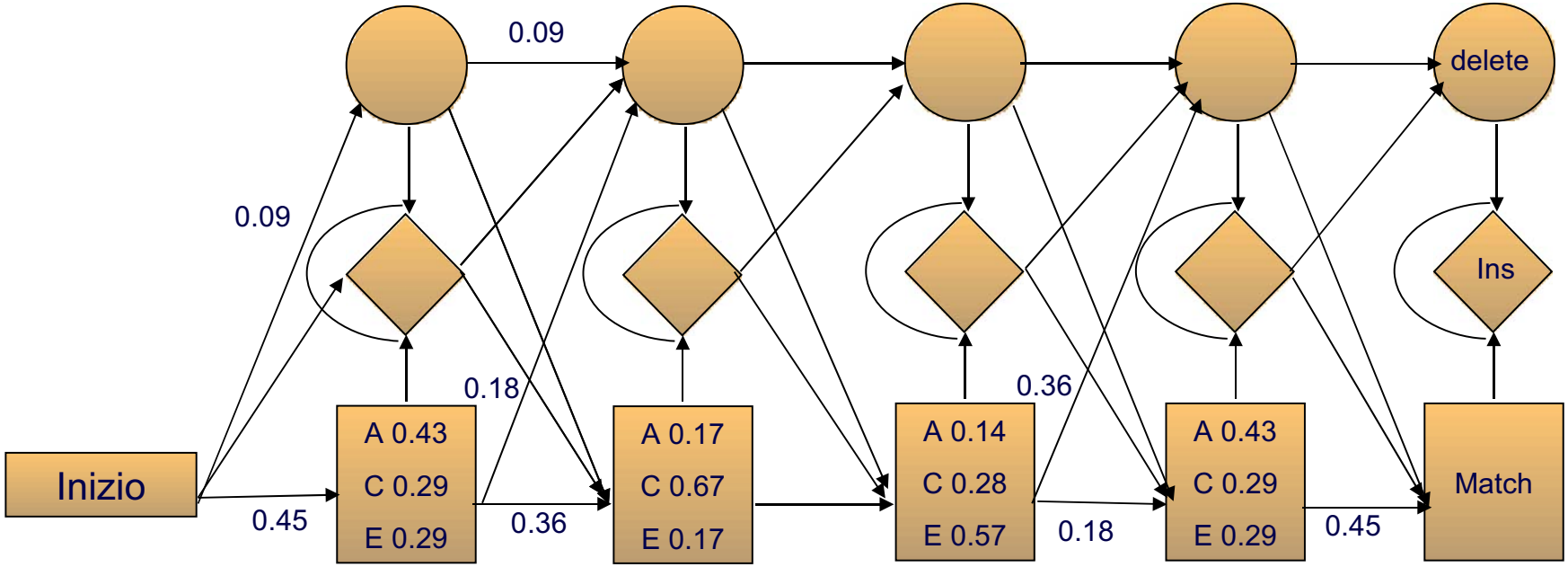
- Representation of the multiple sequenc alignments through the 'transition' probability.

- We can use an alignment for calculating for each position, together with the profile matrix (HMM of 0^{th} order), the probability of finding, after a particular position, an insertion, a deletion or a match state.

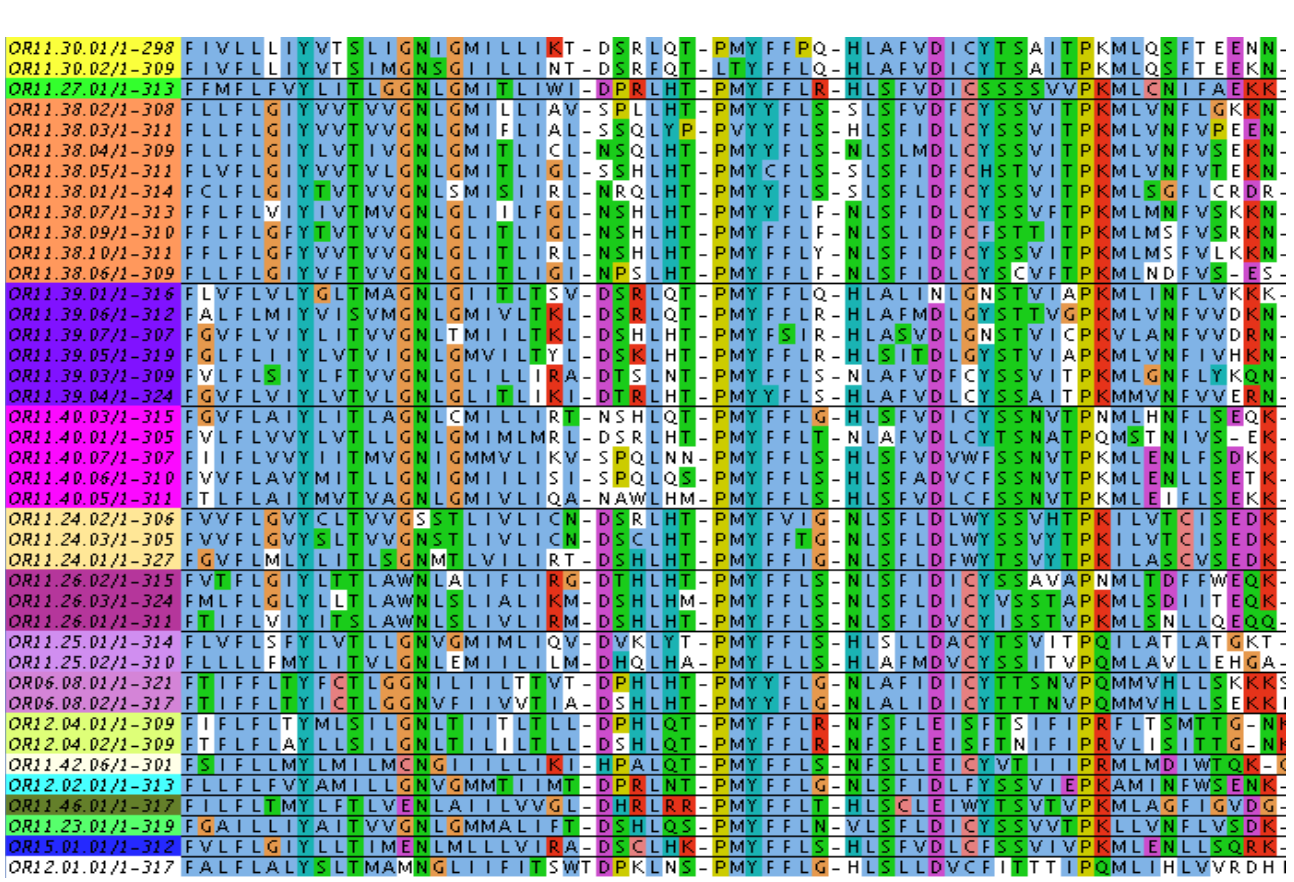
- This permits a more complete characterization of a family, and allows better and more sensible searches for remote homologs, by aligning HMM profiles against databases.

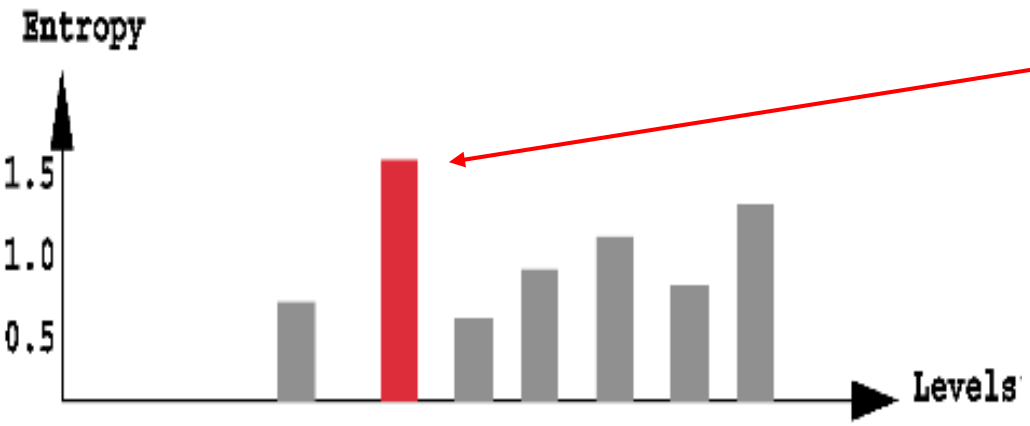
```
Seq1: A C C - E
Seq2: E C E - A
Seq3: A C E A A
Seq4: C - E - E
```

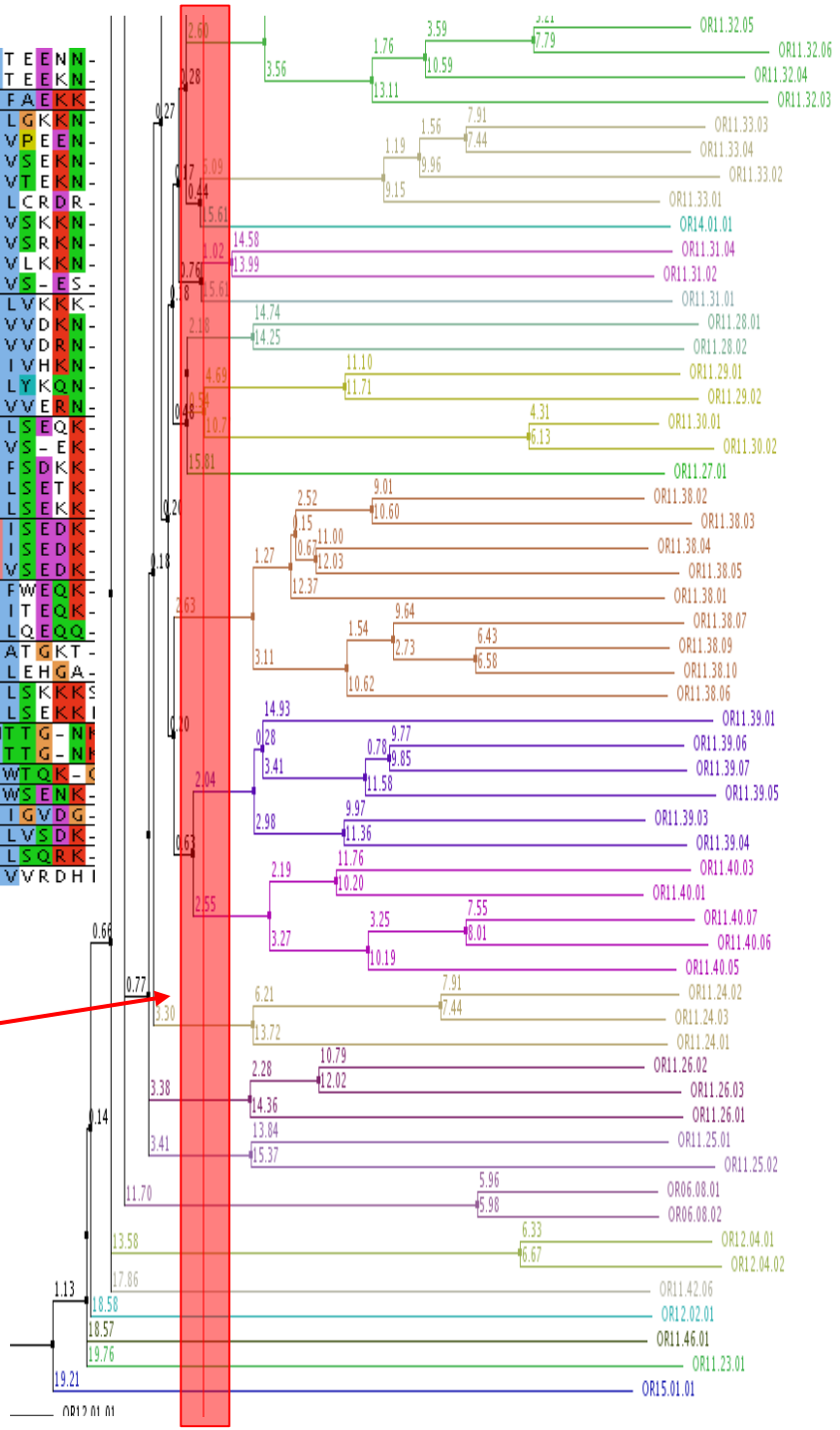
Quantità	Inizio - 1	1 – 2	2 – 3	3 - 4	4 - 5	5 - end
Match - match	4 + 1	3 + 1	3 + 1	1 + 1	1 + 1	4 +1
Match - del	0 + 1	1 + 1	0 + 1	3 + 1	0 + 1	0 + 1
Ins – Del	0 + 1	0 + 1	0 + 1	0 + 1	0 + 1	0 + 1
Del – match	0 + 1	0 + 1	1 + 1	0 + 1	3 + 1	0 + 1
Frequenze						
Match - match	0.45	0.36	0.36	0.18	0.18	0.45
Match - del	0.09	0.18	0.09	0.36	0.09	0.09
Ins - Del	0.09	0.09	0.09	0.09	0.09	0.09
Del – match	0.09	0.09	0.18	0.09	0.36	0.09



Identification of functional subfamilies

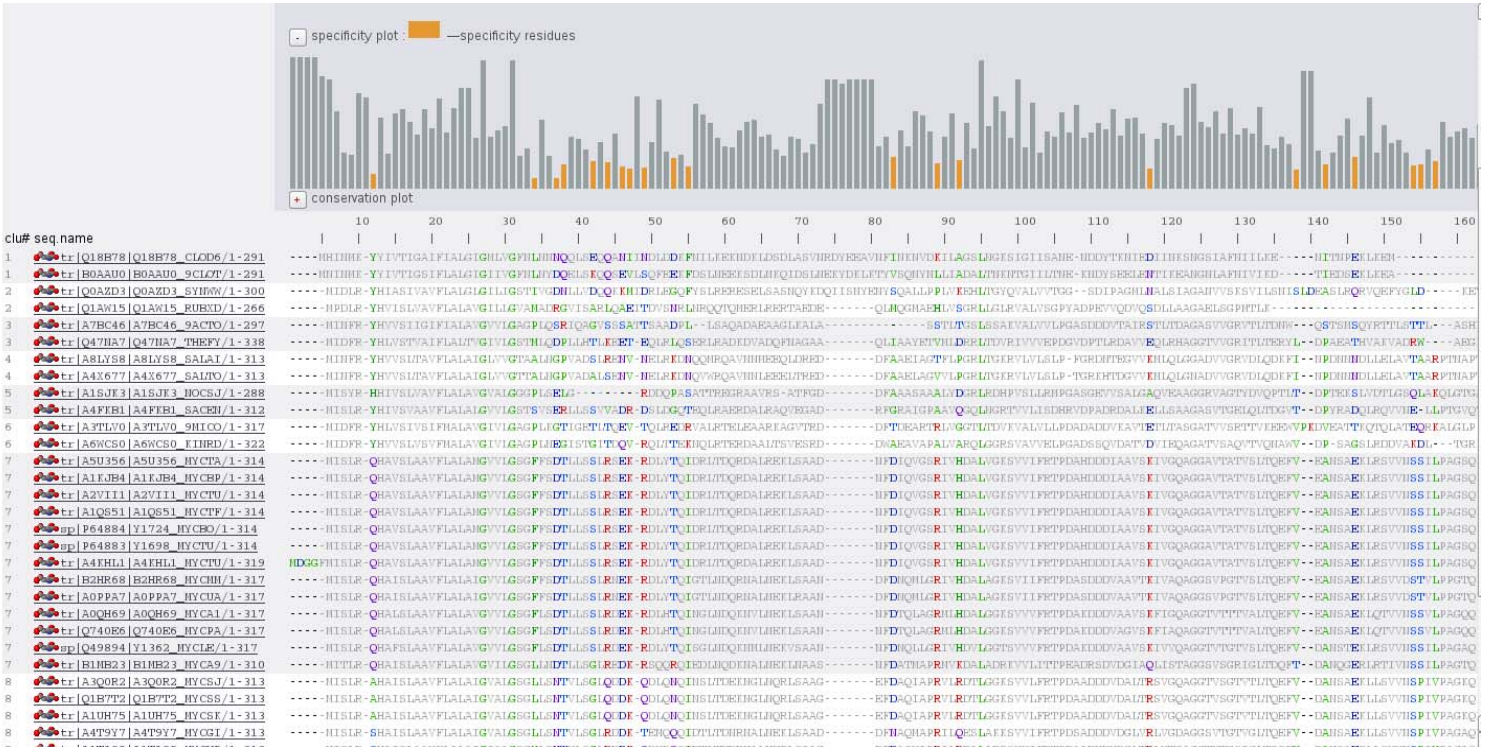




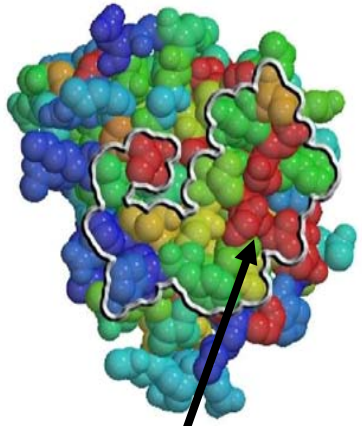


Identification of functional subfamilies

multiple sequence alignment, clustering and amino acid identification from functional subfamily

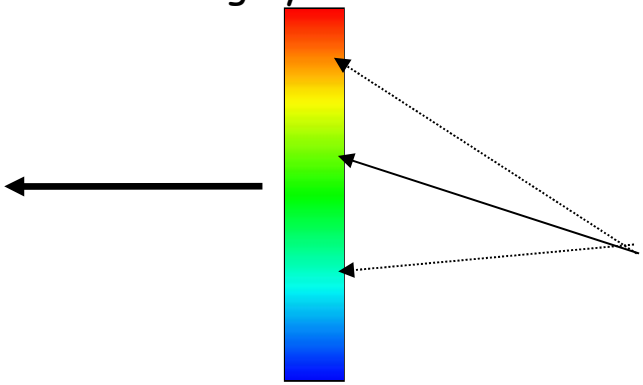


Structural model



Putative functional site

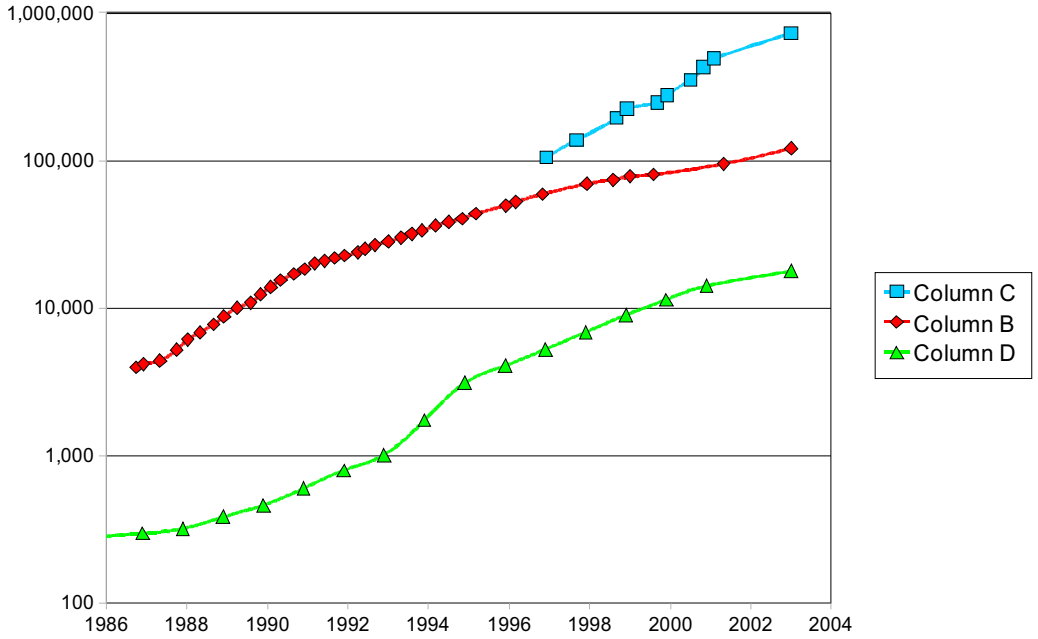
1 = highly conserved



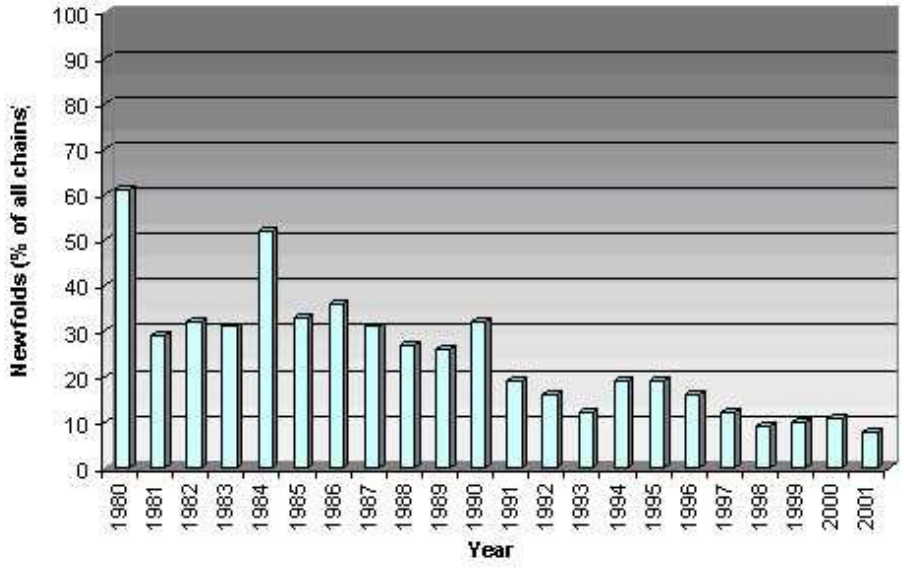
0 = unconserved

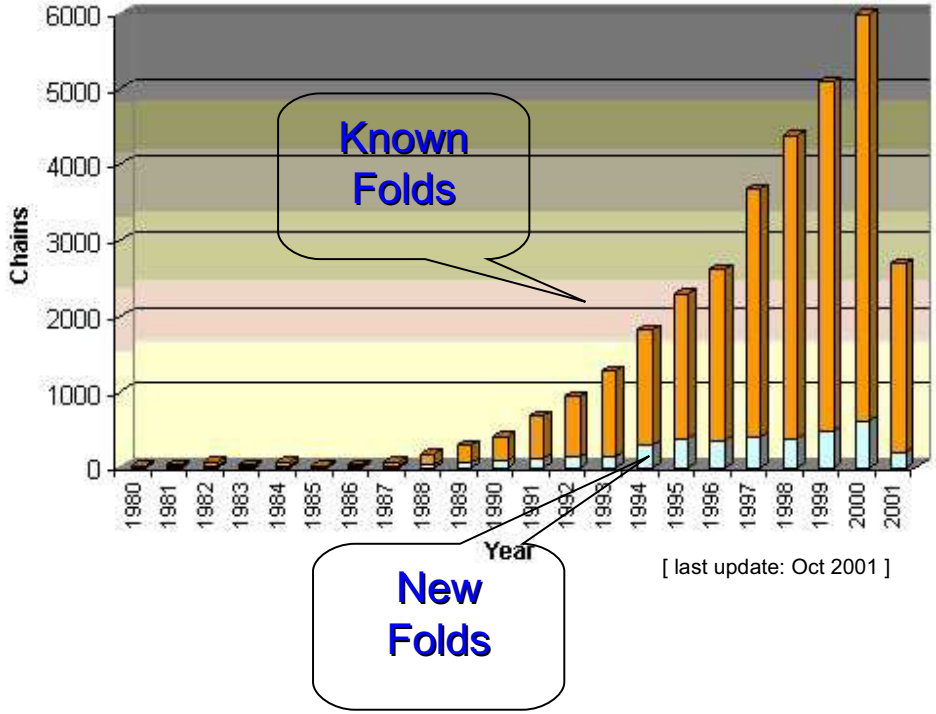
$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

Public Database Holdings:



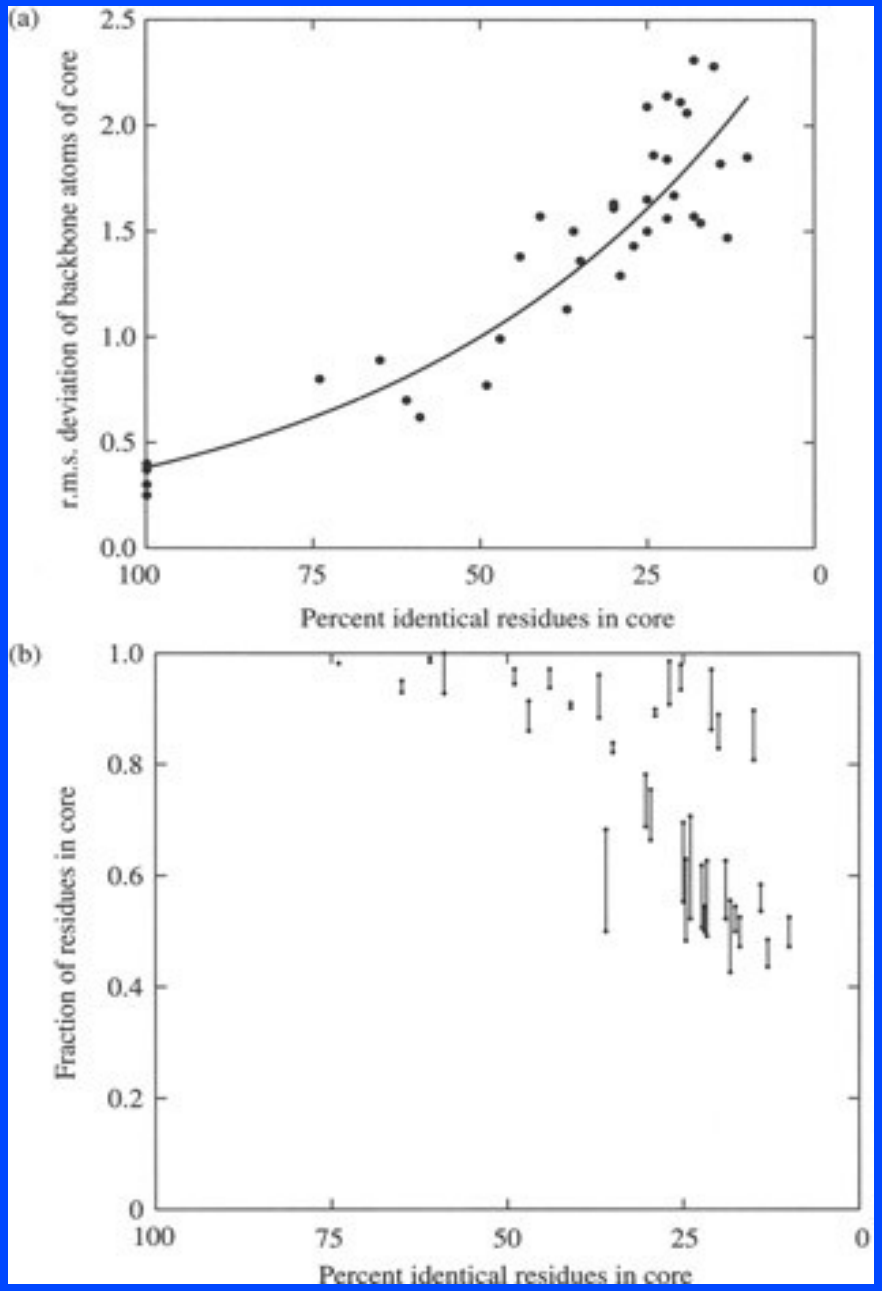
The number of different protein folds is limited:





[last update: Oct 2001]

Protein evolution



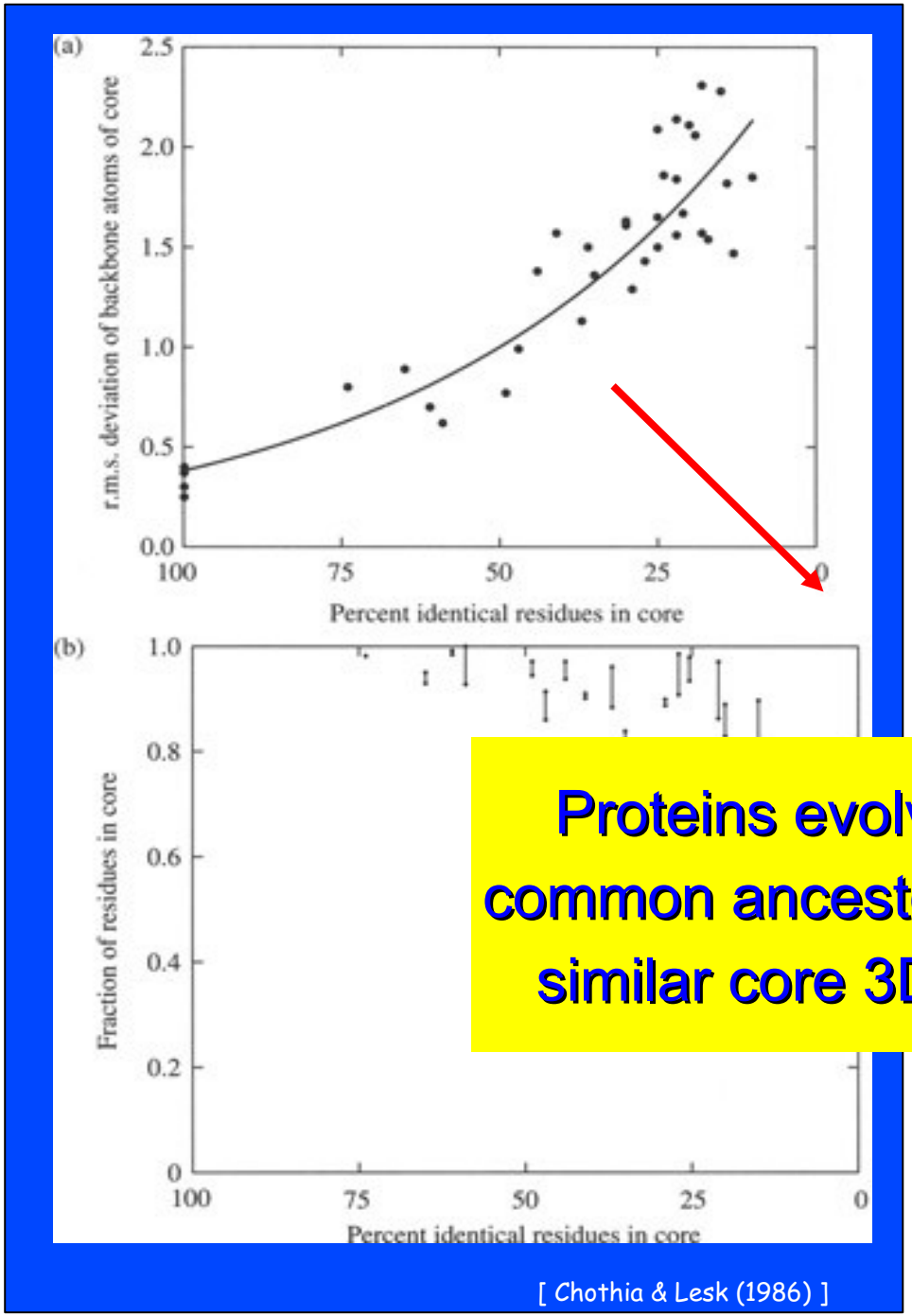
[Chothia & Lesk (1986)]

Comparative modelling

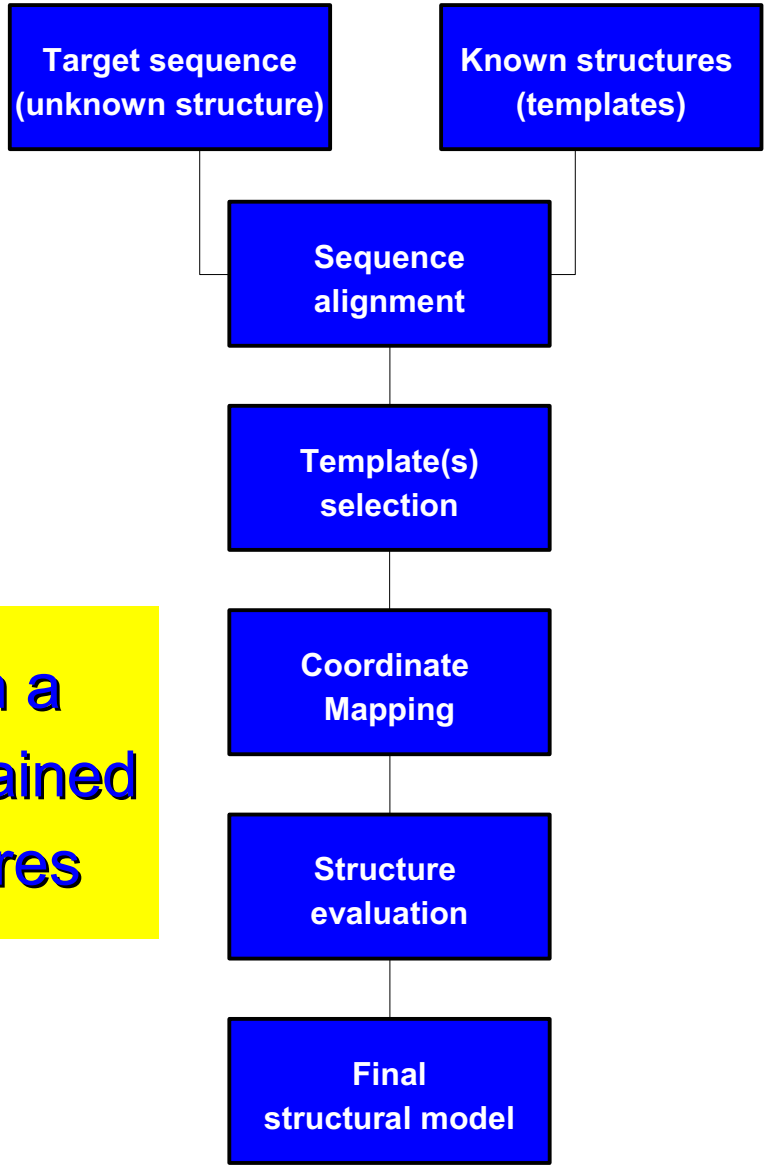
mutation of one amino acid:

- the protein does not fold any more
- the protein accommodates the replacement with minor modifications: evolutionary pressure!
- the protein folds in a completely different structure

Homology modeling

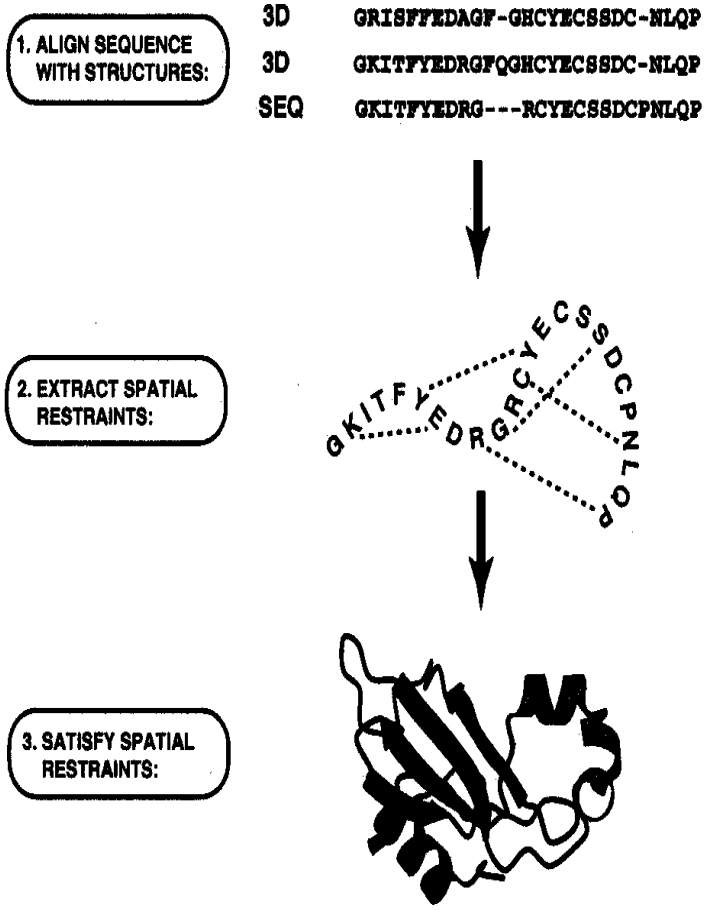


Proteins evolving from a common ancestor maintained similar core 3D structures

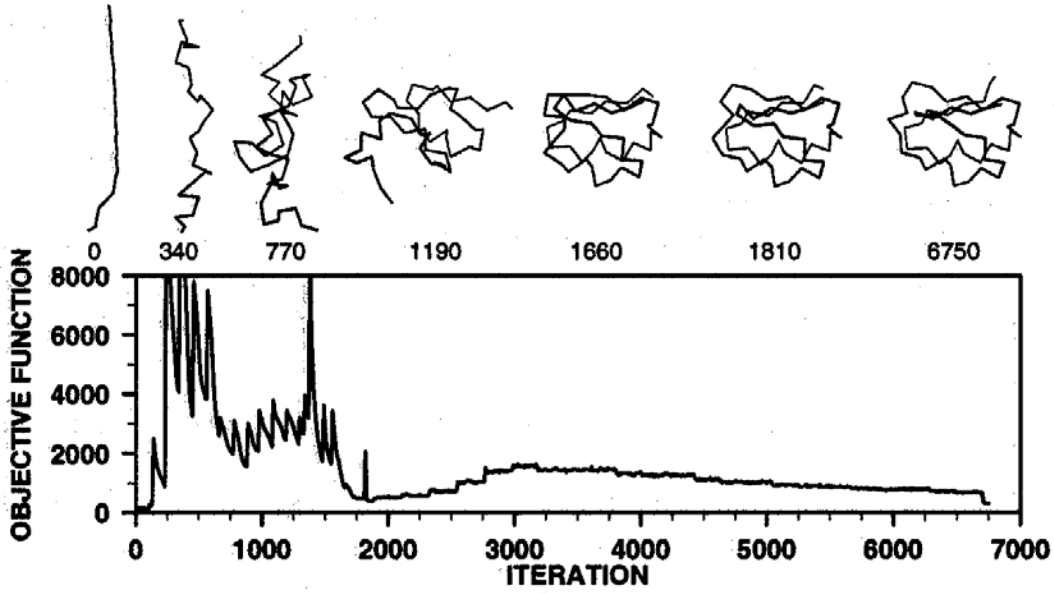


II. Modeling by Satisfaction of Spatial restraints

- Find the most probable structure given its alignment
- Satisfy spatial restraints derived from the alignment.
- Uses probability density functions.
- Minimizes violations on restraints.

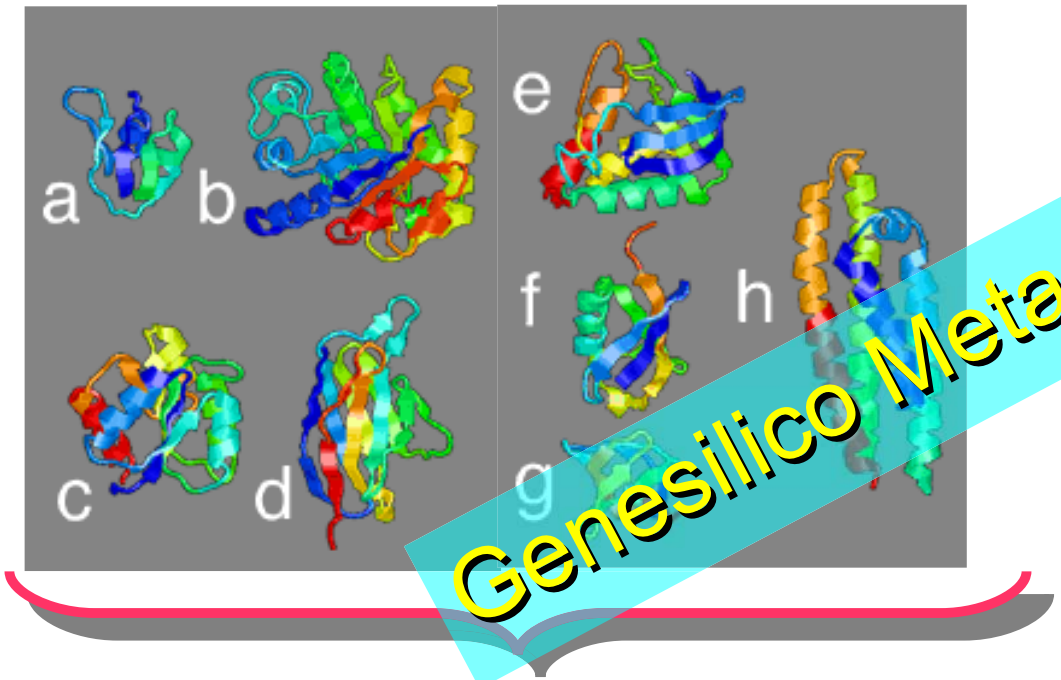


Comparative protein modeling by satisfaction of spatial restraints.
A. Šali and T.L. Blundell. J. Mol. Biol. ***234****, 779-815*



Fold Recognition

MSTLYEKLGGTTAVDLAVAAVAGAPAHKRDVLNQ



Genesilico Meta-server

Build model of target protein based on each known fold

Rank models according to

SCORE or ENERGY

Profile methods

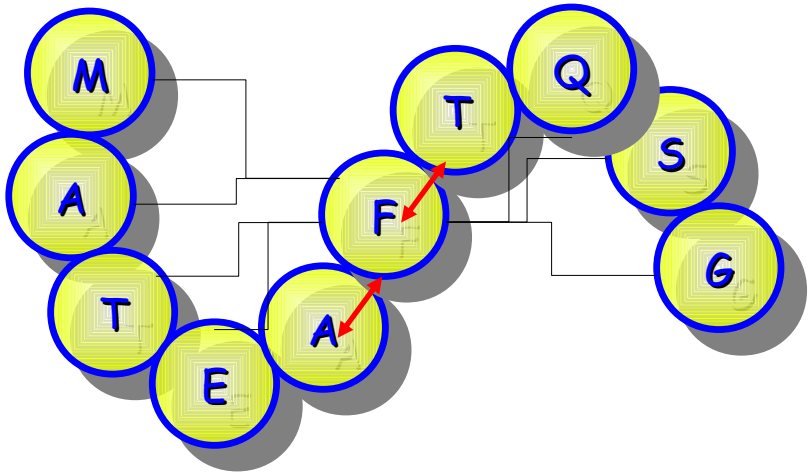
For each aa we can calculate its frequency:

- **Found in secondary structure elements**
- **Found in the surface**
- **found in an hydrophobic environment**

Each aa will be substituted by an intrinsic characteristic

We can do the same for each of the most representative folds.

Threading

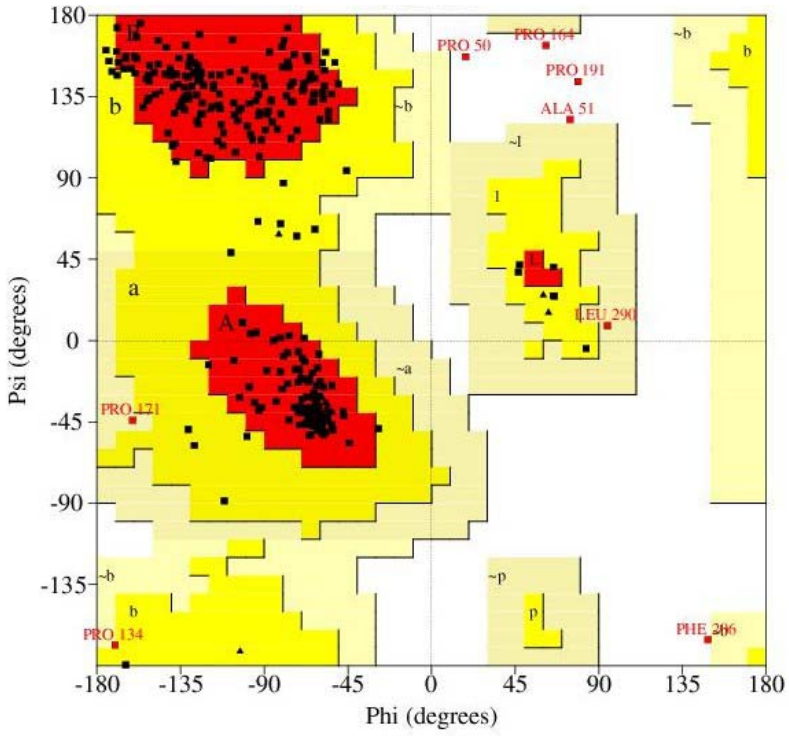


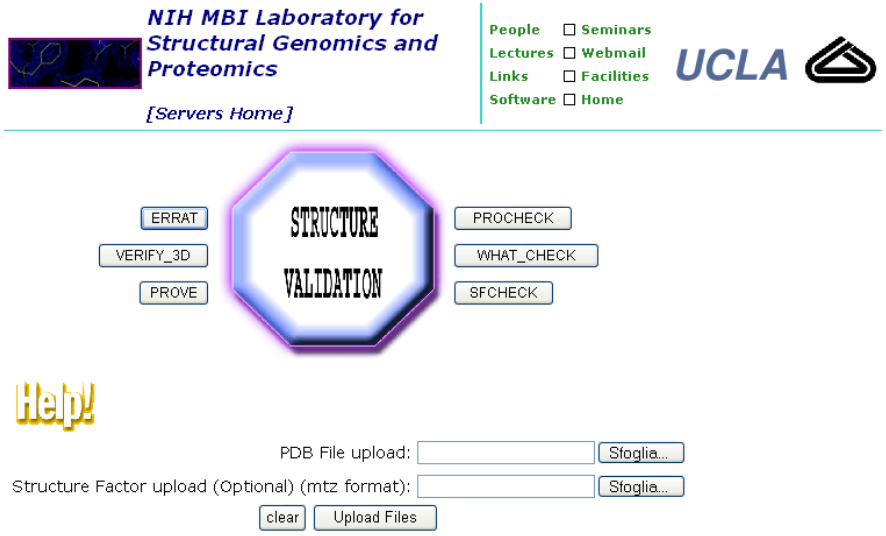
$$\Delta E(Ala-Ala) = -kT \ln \left(\frac{P_{folded}(Ala-Ala)}{P_{unfolded}(Ala-Ala} \right)$$

Model Evaluation ?

Topics:

- correct fold
- model coverage (%)
- Cα - deviation (rmsd)
- alignment accuracy (%)
- side chain placement





- Structure A

http://nihserver.mbi.ucla.edu/SAVS/

Day activities:

Search the human Thioesterase 8 from UNIPROT data base

Analyse the annotated data: cross-links, literature, structures ....

Perform a Blast search (from UNIPROT) and retrieve about 30 members of the family

Generate a Multiple sequence alignment using CLUSTALW (remember: ouput = input)

Analyse the alignment using the program 'jalview' and save the alignment.

Upload your alignments on the TreeDet server.

Generate the structural model using Hhpred.

What is CASP?

Analyse the structure using the different methods.

Visualise it using VMD and compare it with its template: 1C8U

Calculate electrostatic potential. File model.pqr from pdb2pqr server

Protein Structure Resources

PDB http://www.pdb.org
PDB – Protein Data Bank of experimentally solved structures (RCSB)

CATH http://www.biochem.ucl.ac.uk/bsm/cath
Hierarchical classification of protein domain structures

SCOP http://scop.mrc-lmb.cam.ac.uk/scop
Alexey Murzin's Structural Classification of proteins

DALI http://www2.ebi.ac.uk/dali
Lisa Holm and Chris Sander's protein structure comparison server

SS-Prediction and Fold Recognition

PHD http://cubic.bioc.columbia.edu/predictprotein
Burkhard Rost's Secondary Structure and Solvent Accessibility Prediction Server

PSIPRED http://bioinf.cs.ucl.ac.uk/psipred/
L.J McGuffin, K Bryson & David T. Jones Secndary struture prediction Server

3DPSSM http://www.sbg.bio.ic.ac.uk/~3dpss
Fold Recognition Server using 1D and 3D Sequence Profiles coupled.

THREADER: http://bioinf.cs.ucl.ac.uk/threader/threader.html
David T. Jones threading program

Protein Structure Classification

CATH - Protein Structure Classification

UCL, Janet Thornton & Christine Orengo

Class (C), Architecture(A), Topology(T), Homologous superfamily (H)

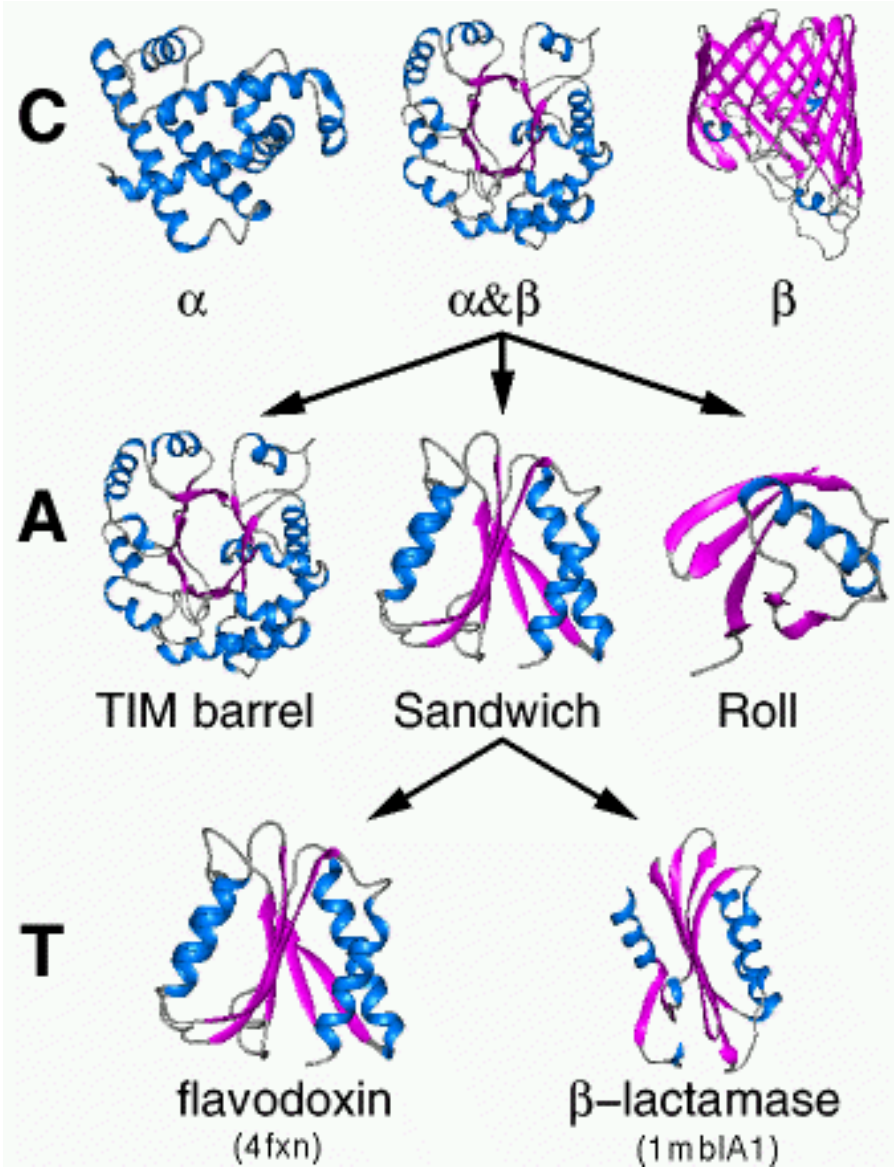
SCOP - Structural Classification of Proteins

MRC Cambridge (UK), Alexey Murzin, Brenner S. E., Hubbard T., Chothia C.

created by manual inspection

comprehensive description of the structural and evolutionary relationships

[http://scop.mrc-lmb.cam.ac.uk/scop/



- Class(C)
  derived from secondary structure content is assigned automatically

- Architecture(A)
  describes the gross orientation of secondary structures, independent of connectivity.

- Topology(T)
  clusters structures according to their topological connections and numbers of secondary structures

- Homologous superfamily (H)

Protein Homology Modeling Resources

SWISS MODEL:
http://www.expasy.org/swissmod/SWISS-MODEL.html

Deep View - SPDBV:
homepage: http://www.expasy.ch/spdbv
Tutorials http://www.expasy.org/spdbv/text/tutorial.htm

WhatIf http://www.cmbi.kun.nl:1100/
Gert Vriend's protein structure modeling analysis program WhatIf

Modeller: http://guitar.rockefeller.edu/modeller
Andrej Sali's homology protein structure modelling by satisfaction of spatial restraints

ROBETTA: http://robetta.bakerlab.org/
Full-chain Protein Structure Prediction Server

Programs and www servers very useful in Comparative modeling:
http://salilab.org/tools/