



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>



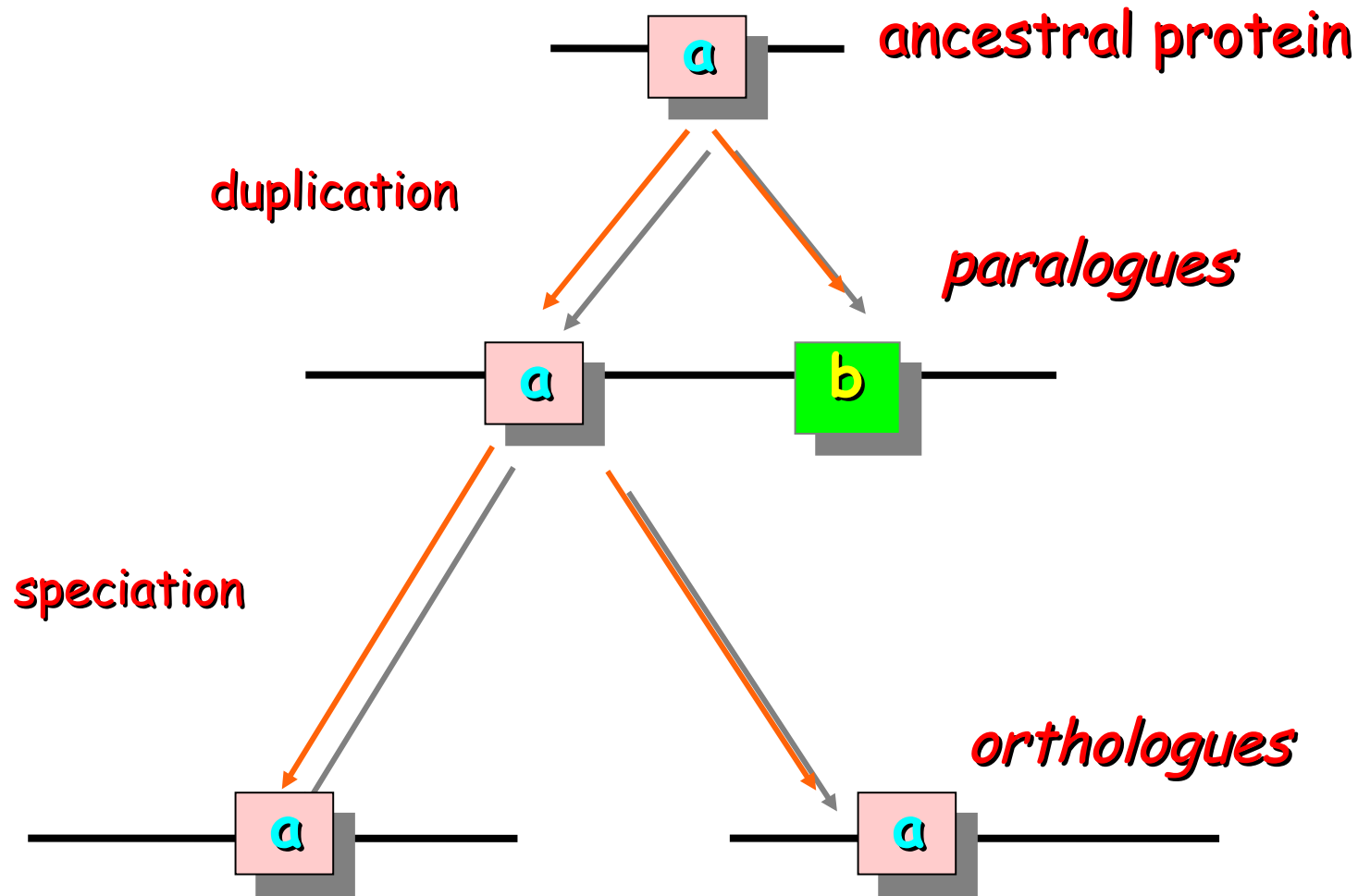
Molecular system-level understanding

- 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

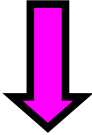


- orthologues - often have very similar functions
- paralogues - may have related functions

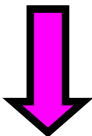
Function Prediction

Molecular function

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



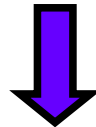
Orthologs identification
HAMAP, EggNogg, COGS,
KOGS



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

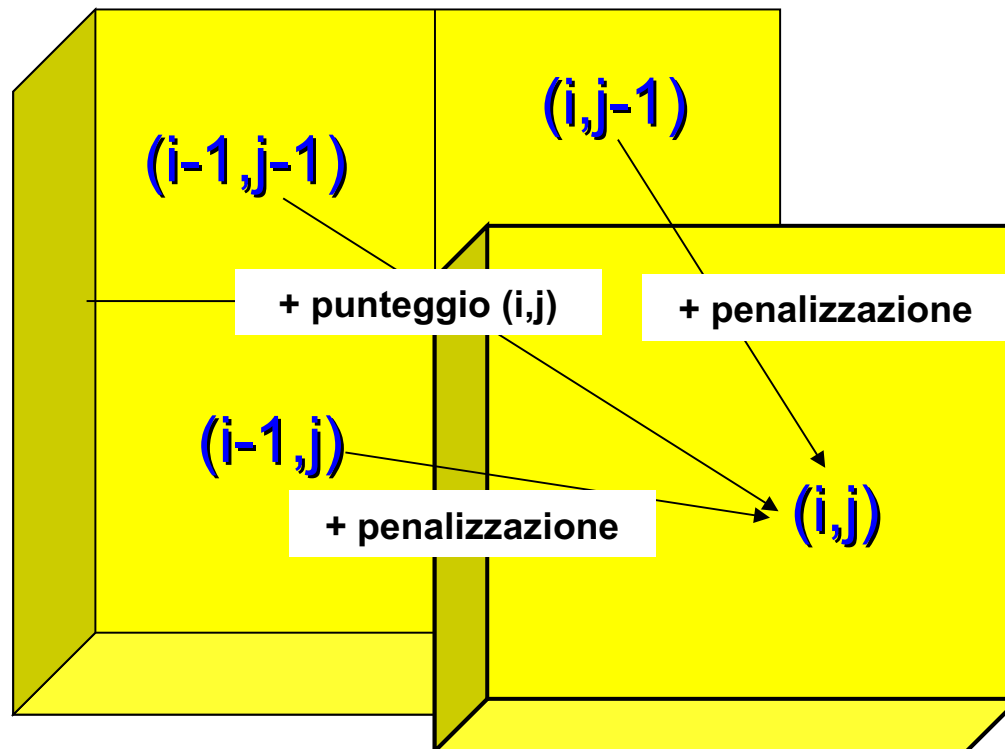
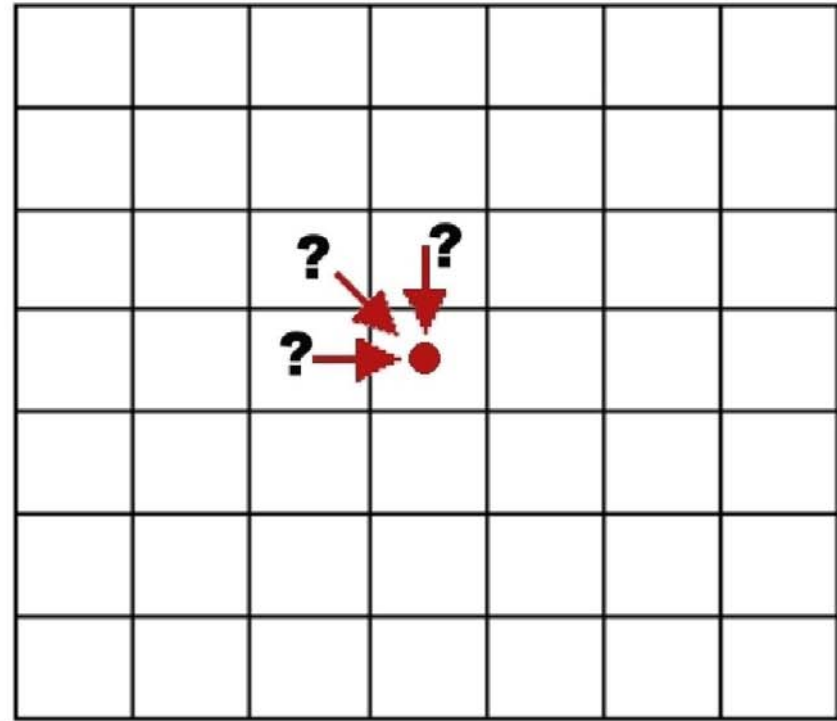
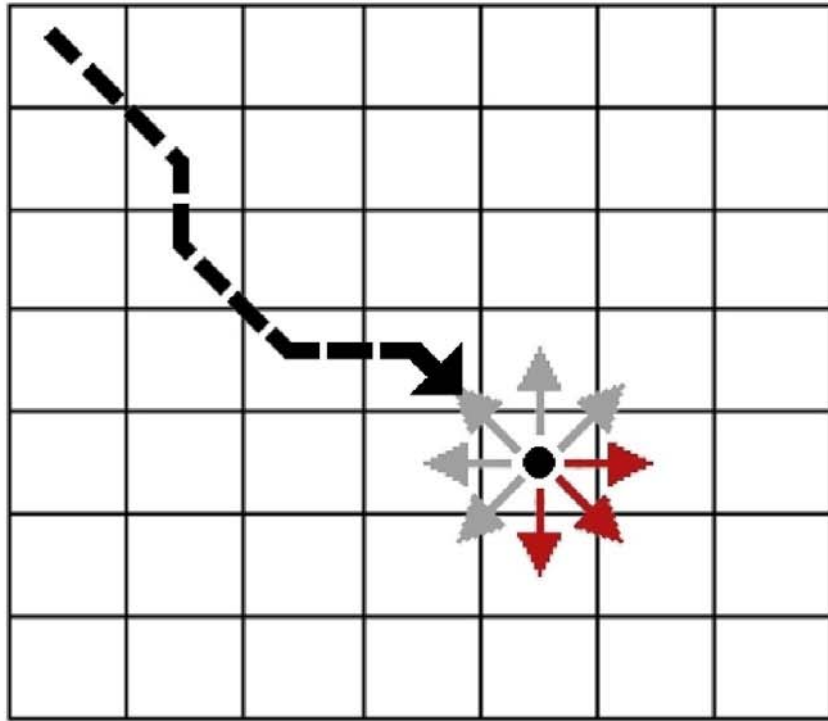
State of art Free-ware bioinformatics programs !!!

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 !!***

- **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 programming: Needleman 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



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

**Structural domains
of the protein**

**The aminoacids involved in
protein function**

Provides information on:

**Residues buried in the
core of the protein**

Remote homologs search

```

PLMN_HUMAN      EYCNLKKCSSETEASVVVAPPVVLLLPDVEITPSEEDCMFNGGKSYRGRKRAF
PLMN_BOVIN      EFCNLKKCSSETEPEQV--PAAPOAPGVENPPEADDCMI GTGKS YRGRKRAF
PLMN_PIG        EYCNLKKCSSETEQQVTNFPAAIQVPSVEDLS-EDCMFNGGKRYRGRKRAF
UROT_HUMAN      EFCSTPACSEGNSDC-YFNGSAYRGTNLSLTSQSASCLPWN SMILIGKV
UROT_MOUSE      EFCSTPACPKGKSEDCYVSKGVTYRGTNLSLTSQASCLPWN SIVLMGKS
UROT_DESRO      EFCSTPACPKGKSEDCYVSKGVTYRGTNLSLTSQASCLPWN SIVLMGKS
                .460.          .470.          .480.          .490.          .500.

PLMN_HUMAN      VTGTPCQDWAAQEPHRRHSIFTPETNPRAGLEKNYCRNPDGDVGGPWCVYT
PLMN_BOVIN      VAGVPCQEWAAQEPHRRHSIFTPETNPRAGLEKNYCRNPDGDVNGPWCVYT
PLMN_PIG        VAGVPCQEWAAQEPHRRHSIFTPETNPRAGLEKNYCRNPDGDVNGPWCVYT
UROT_HUMAN      TAQNP SAQALGLS-----KHN YCRNPDGD AKPWCHVL
UROT_MOUSE      TAWRTNSQALGLA-----RHN YCRNPDGD ARPWCHVM
UROT_DESRO      TAWRTNSQALGLA-----RHN YCRNPDGD ARPWCHVM
                .510.          .520.          .530.          .540.          .550.

PLMN_HUMAN      NPRKLYDYCDVPCQAAPSFDCGKPKQVEPKKCPGRVVGCCVANPHSWPWQ
PLMN_BOVIN      NPRKLYDYCDVPCQES--SFDGCGKPKQVEPKKCSERIVGGCVSKPHSWPWQ
PLMN_PIG        NPGKLYDYCDVPCQVTS--SFDGCGKPKQVEPKKCPARVVGCCVSIPHSWPWQ
UROT_HUMAN      NRRLLTWEYCDVPCSTCGLRQYSQPQFRIK--GSLFADIASHPWQAAIF
UROT_MOUSE      DRKLLTWEYCDMSFCSTCGLRQYKRPQFRIK--GSLYTDITSNPWQAAIF
UROT_DESRO      ---ATCGLRKYKKEPQLHST--GSLFTDITSNPWQAAIF
                .560.          .570.          .580.          .590.          .600.

PLMN_HUMAN      SLRTRFGMHF-CGGTLISPEWVLTAAACLEKSRPSSYKVVILGAHQEVN
PLMN_BOVIN      SLRRSSR-HF-CGGTLISPEKWVLTAAACLDNIALALSFYKVVILGAHNEKV
PLMN_PIG        SLRYRYRGNF-CGGTLISPEWVLTAKHCLKSSSPSSYKVVILGAHEEYH
UROT_HUMAN      KHRRSPGERFLCAGILISSCVILSAAACFQERFPPHMLTVILGRTYRVV
UROT_MOUSE      KNKRS PGERFLCAGVLISSCWWVLSAAACFLERFPPHMLKVVILGRTYRVV
UROT_DESRO      QNRRSSGERFLCAGILISSCVVLTAAACFQERYPPQHLLRVVILGRTYRVK
                .610.          .620.          .630.          .640.          .650.

PLMN_HUMAN      EPHVQEIIEVSRLEFLEPTRK-----DIALLKLS SPAVITDKVIPACLPS
PLMN_BOVIN      EQSVQEIIPVSRLEFREPSQA-----DIALLKLS RPAIITKEVIPACLPP
PLMN_PIG        GEGVQEIIDVSKLFKEPSQA-----DIALLKLS SPAVITDKVIPACLPT
UROT_HUMAN      GEEEQKFEVEKYIVHKEFD DDDTYDNDIALLQLKSDSSRCAQESSVVRTV
UROT_MOUSE      GEEEQTFEIEKYIVHKEFD DDDTYDNDIALLQLRSQSKQCAQESSVSTA
UROT_DESRO      GKEEQTFEVEKCIVHEEFD DDDTYDNDIALLQLKSGSPQCAQESSVRAI
                .660.          .670.          .680.          .690.          .700.

PLMN_HUMAN      NYVVA---DRT ECFITGSGETQGTFGAGLLKEAQLPVIENKV--CNRV
PLMN_BOVIN      NYMVA---ART ECFITGSGETQGTFGAGLLKEAHL PVIENKV--CNRV
PLMN_PIG        NYVVA---DRT ACYITGSGETKGTFGAGLLKEARLPVIENKV--CNRV
UROT_HUMAN      LPPADLQLPDWTECELSGYGKHEALSPFYSERLKEAHVRLY PSSRCTSQ
UROT_MOUSE      LPDENLQLPDWTECELSGYGKHEALSPFFSDRLKEAHVRLY PSSRCTSQ
UROT_DESRO      LPEANLQLPDWTECELSGYGKHKSSSPFYSEQLKEAHVRLY PSSRCTSK
                .710.          .720.          .730.          .740.          .750.

PLMN_HUMAN      FLNGRVQSTELCASHLAGGT-----DSCQSDSGG PLVCFEKDKYIILQG
PLMN_BOVIN      YLDGRVKPTELCASHLAGGT-----DSCQSDSGG PLVCFEKDKYIILQG
PLMN_PIG        YLGGKVSPNELCASHLAGGI-----DSCQSDSGG PLVCFEKDKYIILQG
UROT_HUMAN      LLNRTVTDNMLCAGDTRSGGQANLHDACQSDSGG PLVCLNDGRMTLVG
UROT_MOUSE      LFNKTVTDNMLCAGDTRSGGNQ-DLHDACQSDSGG PLVCMINKQMTLVG
UROT_DESRO      LFNKTVTKNMLCAGDTRSGEIHFNVDACQSDSGG PLVCRNDNMMTLLS
                .760.          .770.          .780.          .790.          .800.

PLMN_HUMAN      TSWGLSCARP NKPGVYVRVSRFVIWIEGVMRNN
PLMN_BOVIN      TSWGLSCARP NKPGVYVRVSPYVFWIEETMRRN
PLMN_PIG        TSWGLSCALENKPGVYVRVSRFVIWIEEIMRRN
UROT_HUMAN      ISWGLSCGQKDVPGVYTKVTNYLDWIRDNMRR
UROT_MOUSE      ISWGLSCGQKDVPGVYTKVTNYLDWIRDNMRR
UROT_DESRO      ISWGLSCGQKDVPGVYTKVTNYLDWIRDNMRR
                .810.          .820.          .830.          .840.          .850.

```

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.

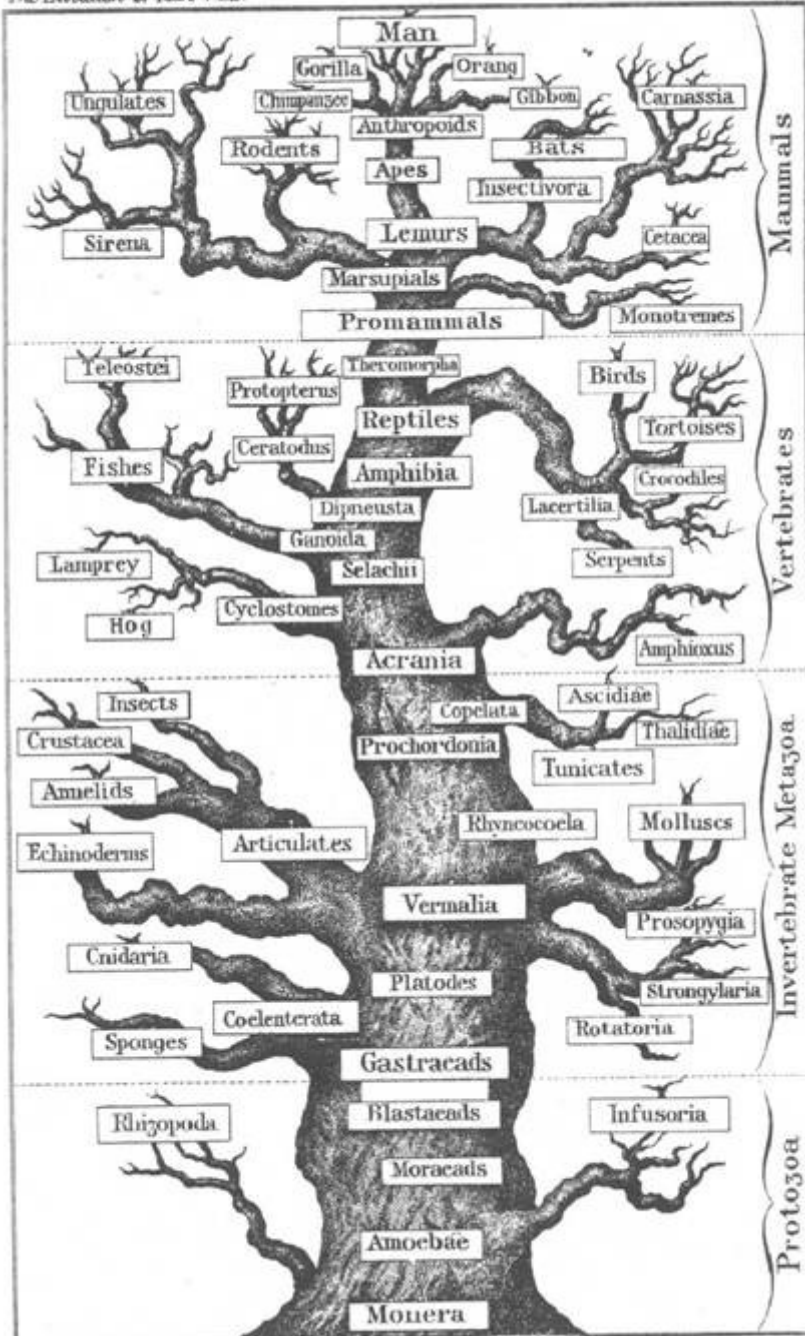
Used in in PSI-BLAST and PSI-search !!!

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

G Genealogical Tree of Humanity.

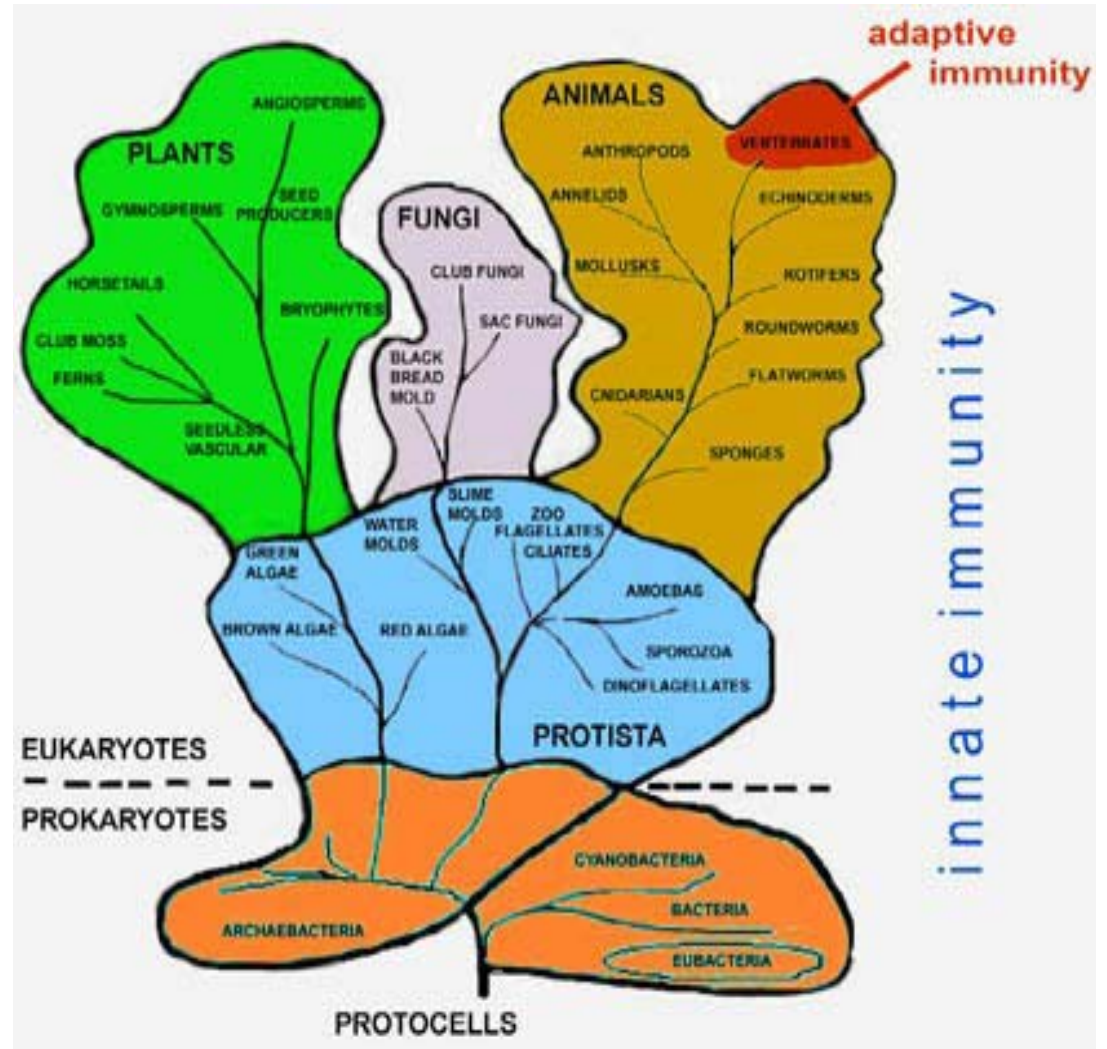
The Evolution of Man V. Ed.

PL. XX.



E. Haeckel del.

Phylogenetic Trees

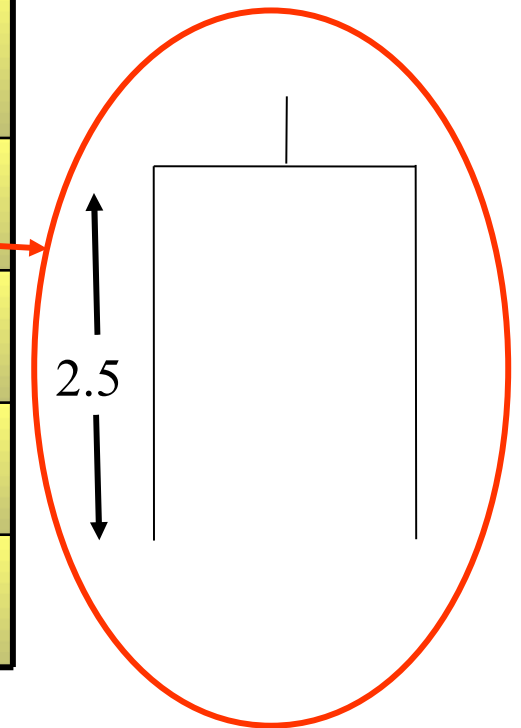


Phylogenetic Trees

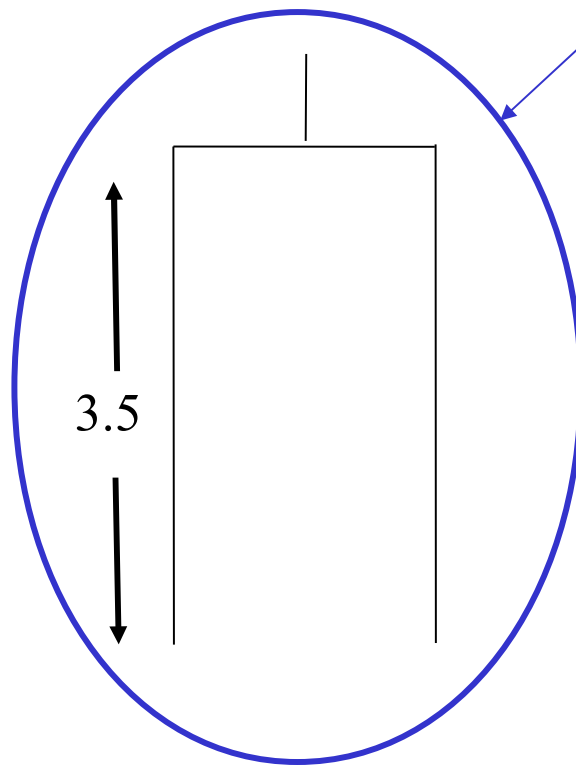
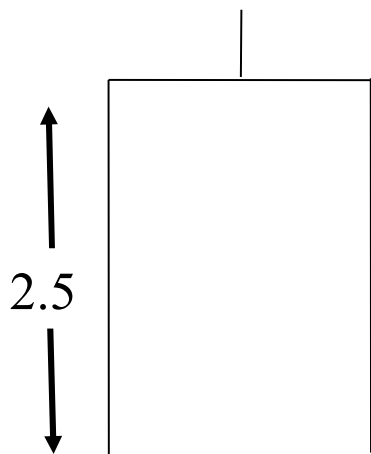
- It's a way of visualizing evolutionary relationships
- Each external node corresponds to a species
- Internal nodes: **speciation event**
- The distance between two nodes is proportional to the divergence time.
- In protein sequences: node → protein
- **The distance between two nodes is proportional to the sequence similarity.**

Phylogenetic Trees

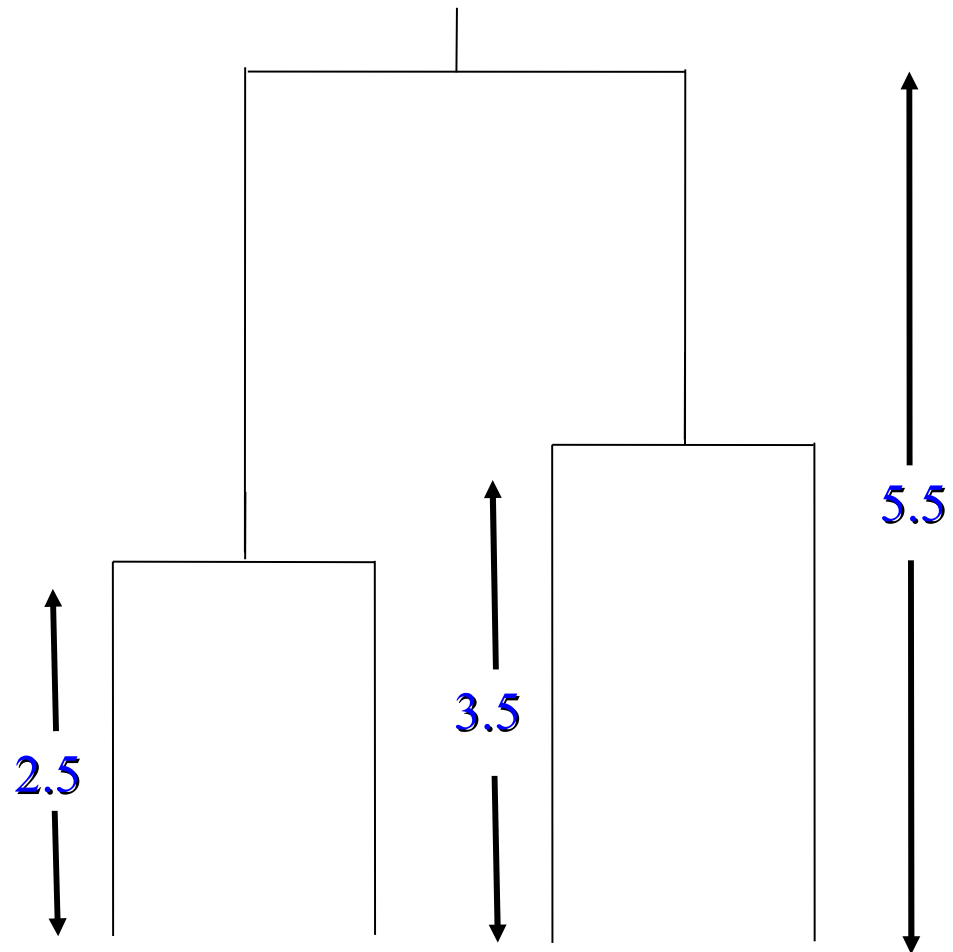
% aa different	Seq1	Seq2	Seq3	Seq4
Seq1	0	5	11	14
Seq2		0	9	10
Seq3			0	7
Seq4				0



% aa different	Cluster 1,2	Seq3	Seq4
Cluster 1, 2	0	$\frac{1}{2}[d(1,3)+d(2,3)]=10$	$\frac{1}{2}[d(1,4)+d(2,4)]=12$
Seq3		0	7
Seq4			0



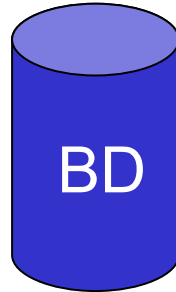
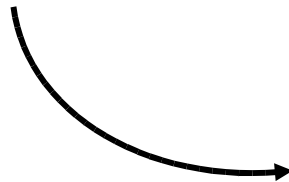
% aa diversi	Cluster 3,4
Cluster 1, 2	$=\frac{1}{2}d[(\text{Cluster } 1,2), 3]+d[(\text{Cluster } 1,2),4]=11$



Data-base search for homology

Fasta (indexing method for a quick alignment production: heuristic)

KRTIDPQ



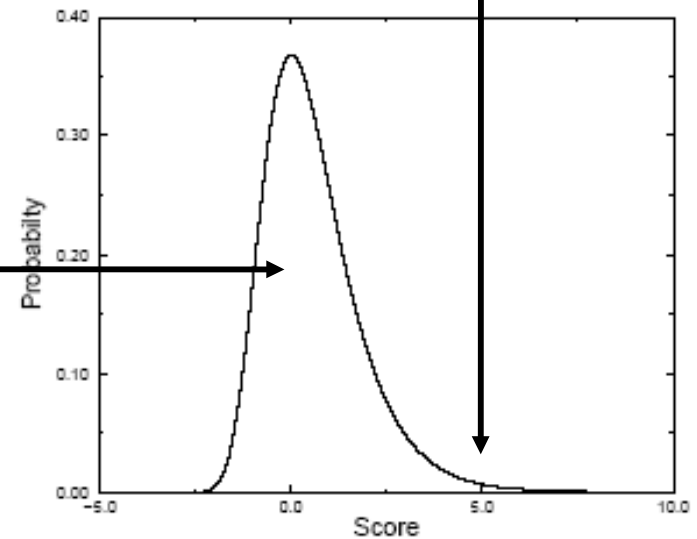
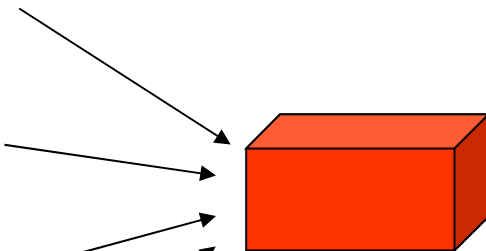
Score S'

KITRQDP

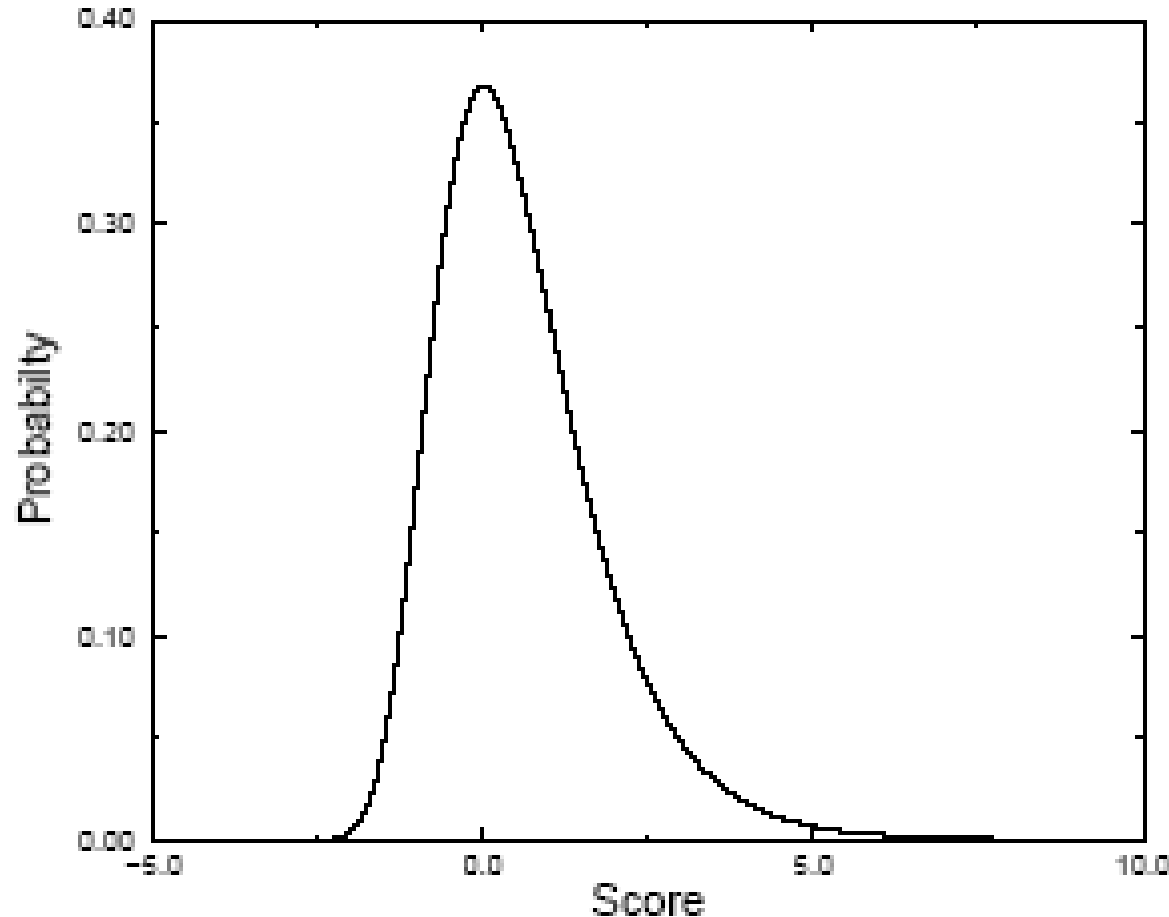
PDQKRIT

DPQTKRI

DPQTKRI



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 maximum 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:

$$E = Kmne^{-\lambda S}$$

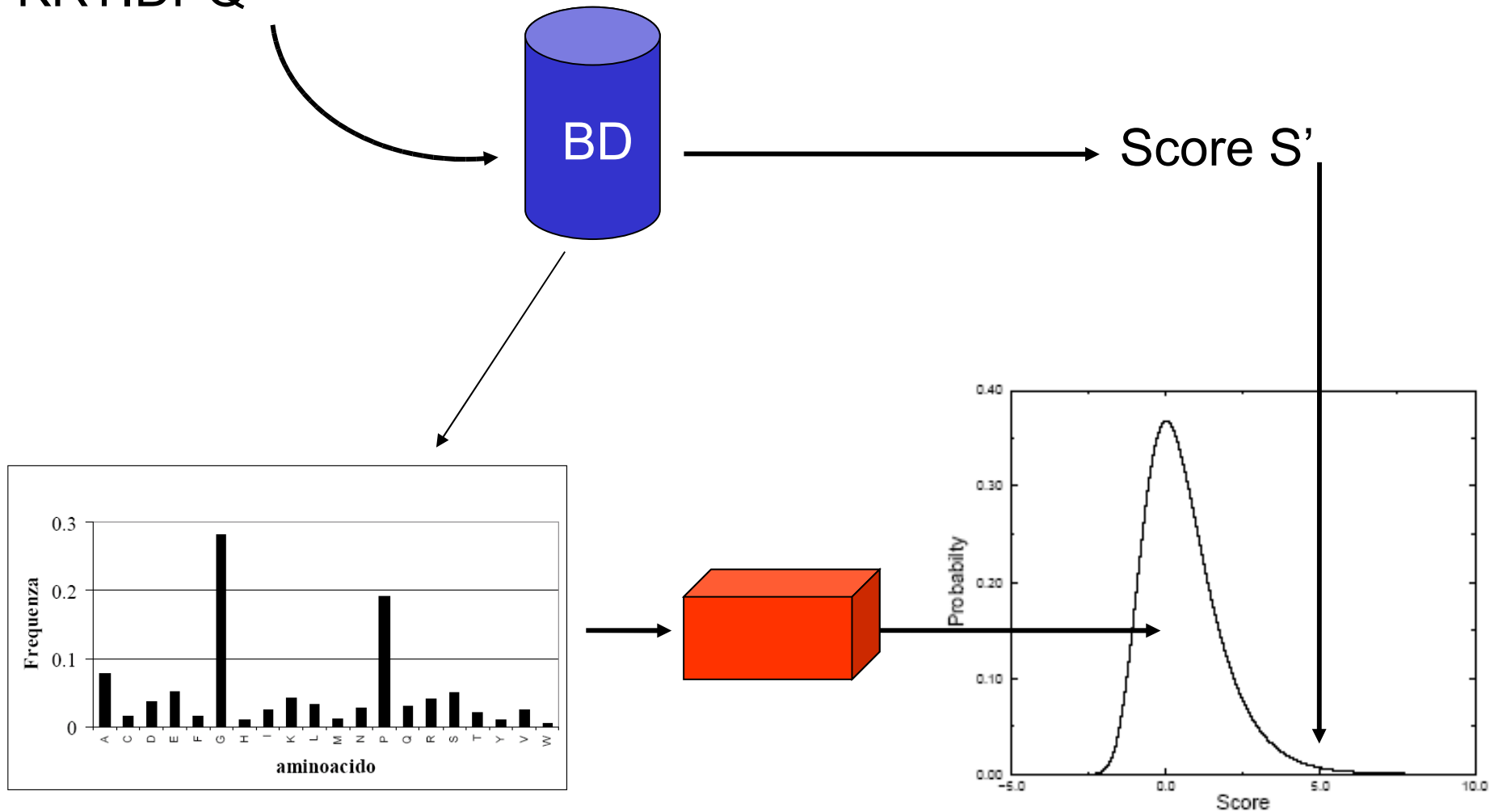
S is generally normalized: $S' = (\lambda S - \ln K) / \ln 2$

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 ?

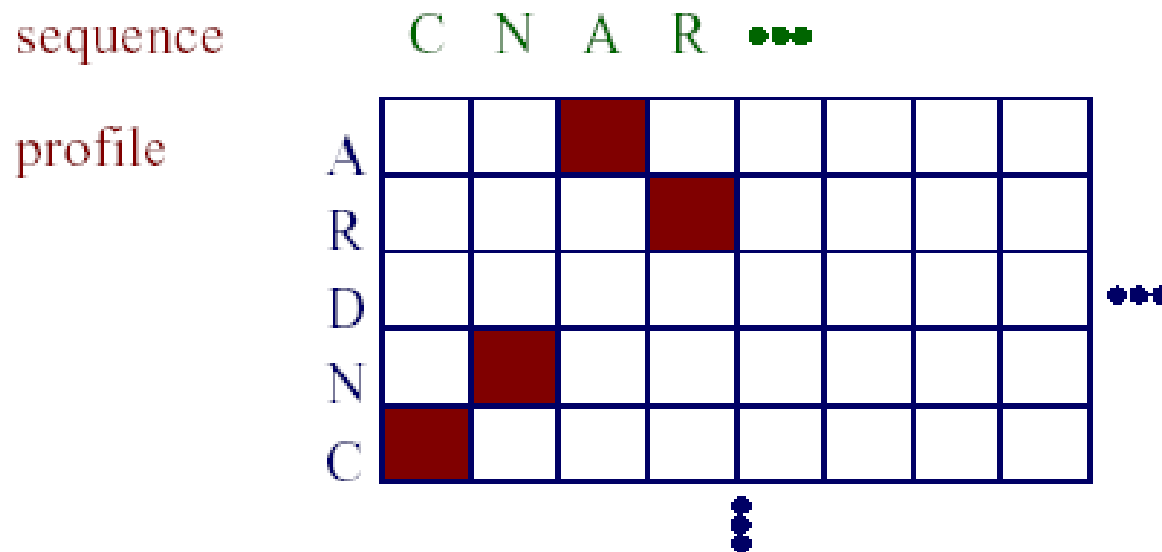
Profile Matrix (Position Specific Scoring Matrix – PSSM)

sequence positions

	1	2	3	4	5	6	7	8	
A			-2.4						
R			1.2						
D			0.5						...
N			-0.2						
C			-3.1						
					...				

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

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

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

Hidden Markov models

- Representation of the multiple sequence alignments through the '**transition**' probability.
- We can use an alignment for calculating for each position, together with the profile matrix (HMM of 0th 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.

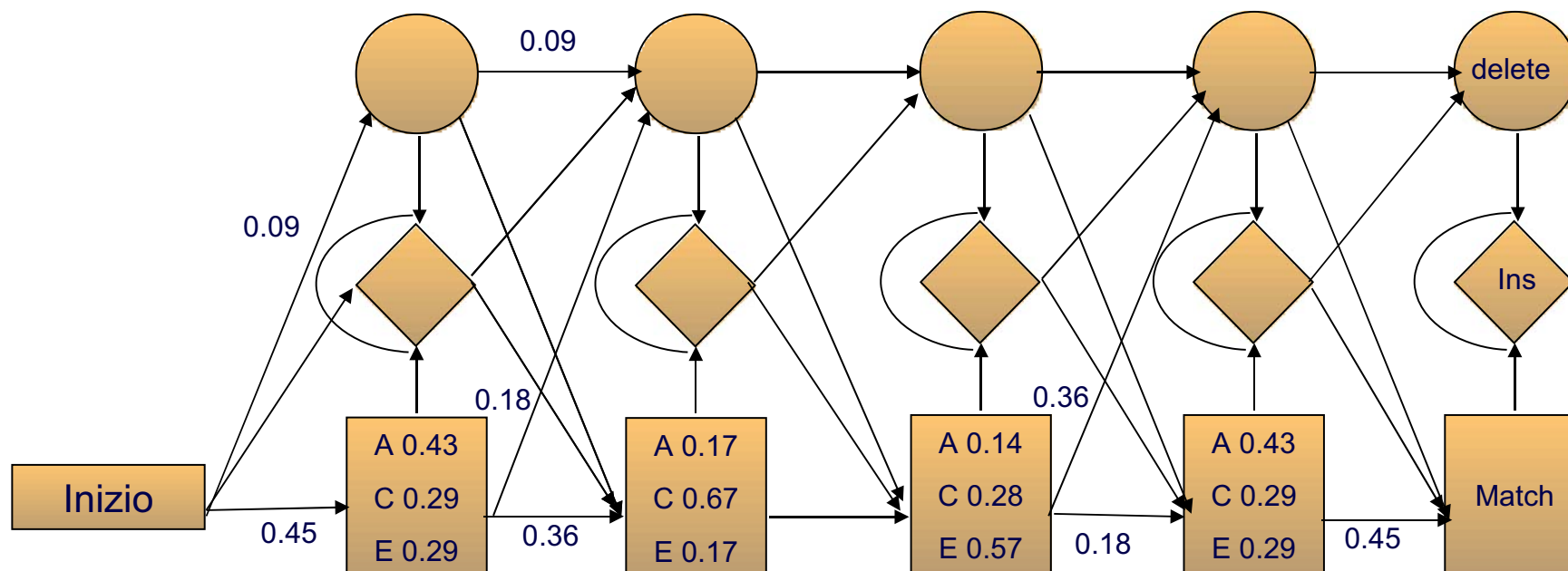
Seq1: A C C - E

Seq2: E C E - A

Seq3: A C E A A

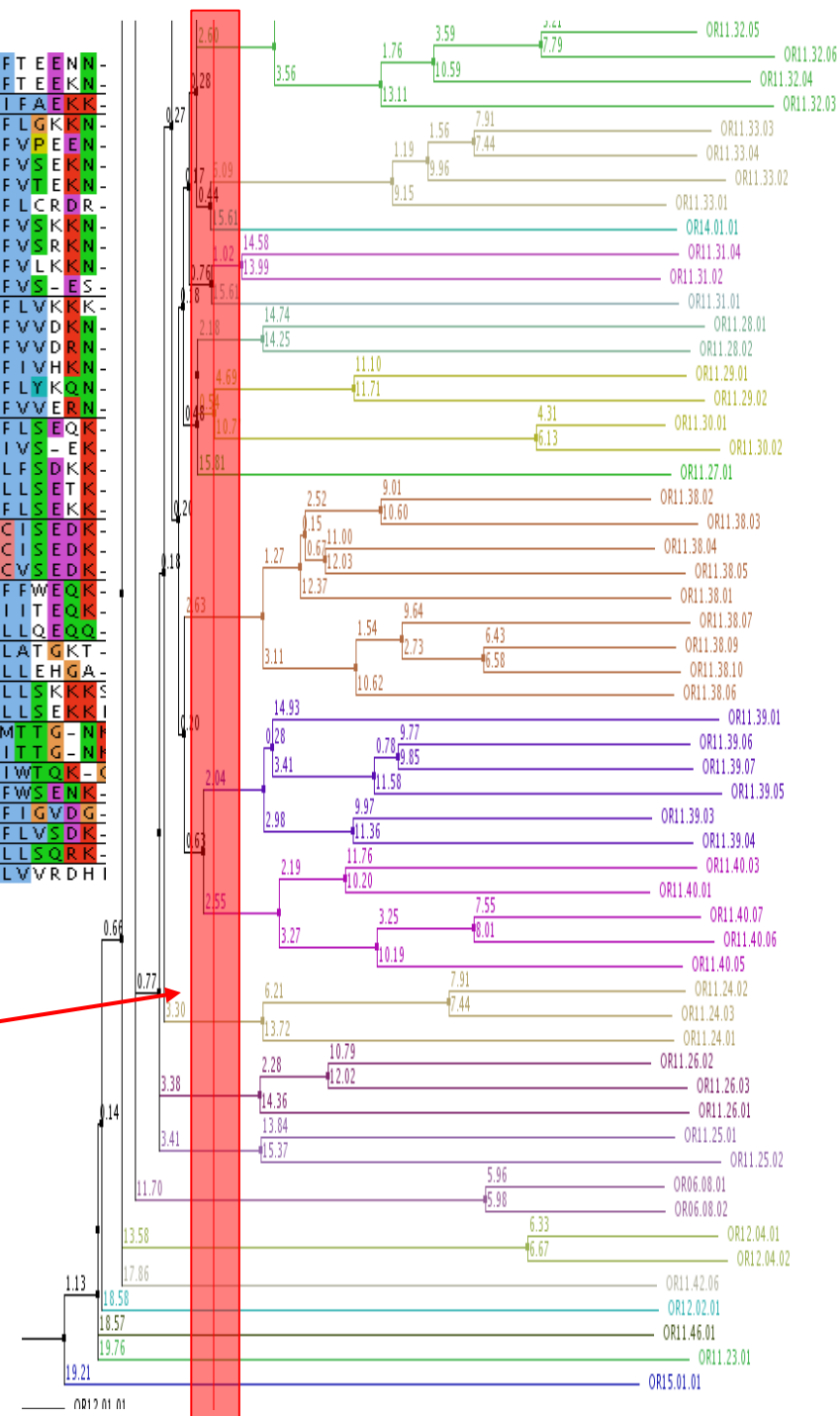
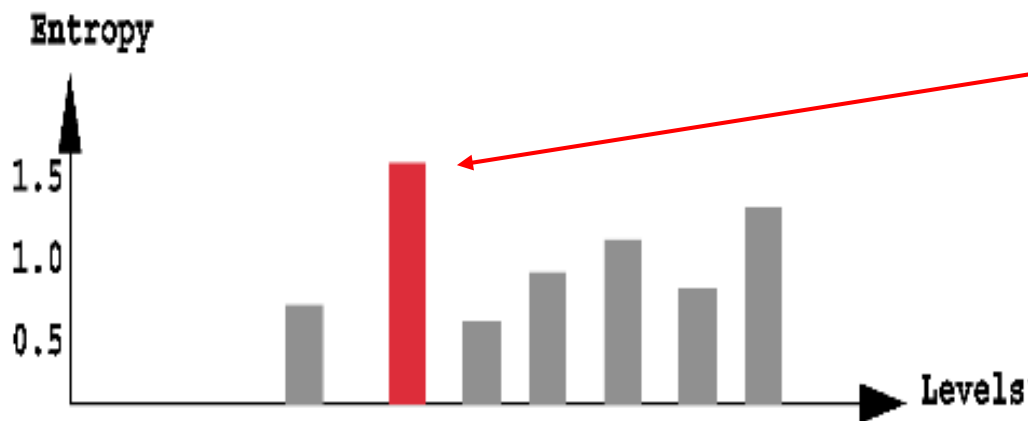
Seq4: C - E - E

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



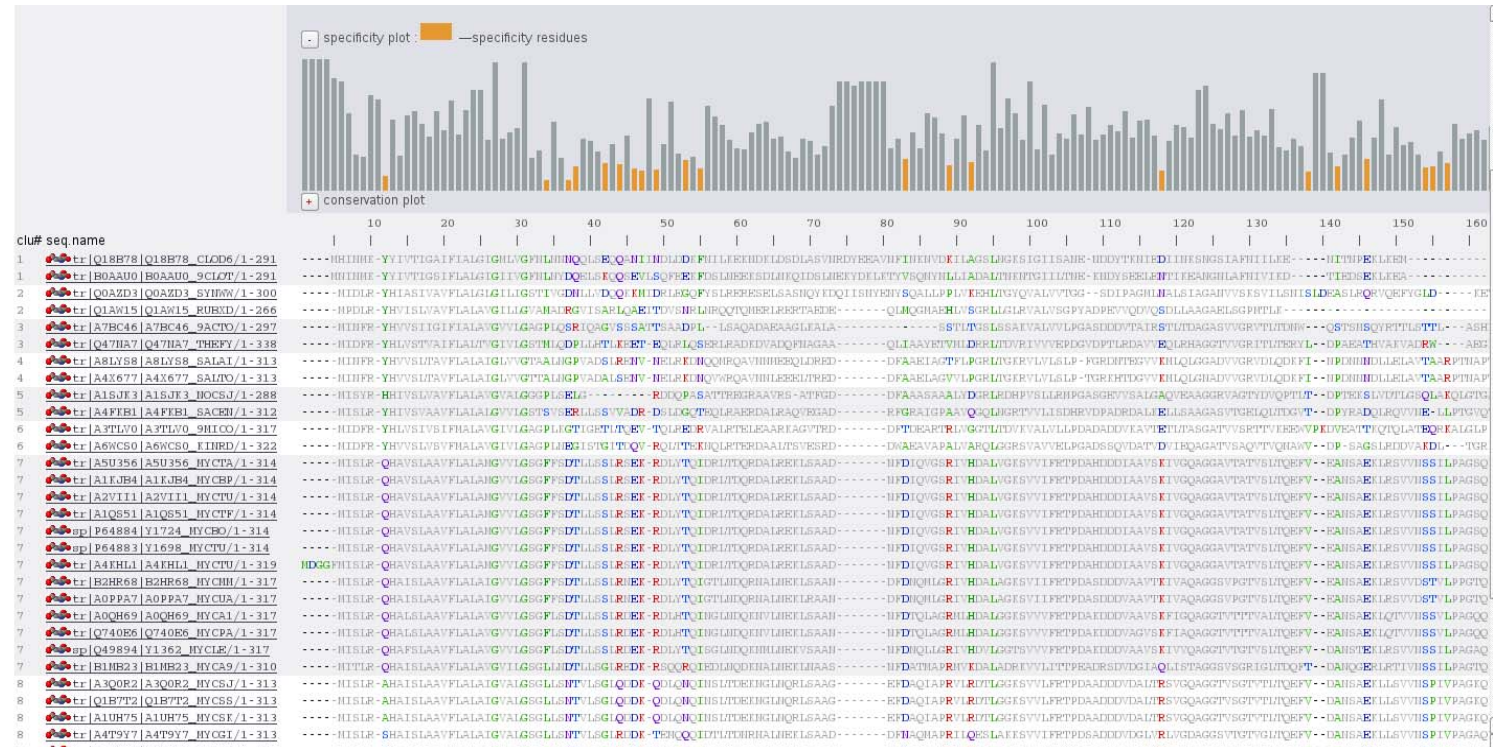
Identification of functional subfamilies

OR11.30.01/1-298	FIVLLLIYVTSLIGNIGMILLIKT-D	SRLQT-PMYFFPQ-HLAFVDICYTSAITPKMLQSFTENN-			
OR11.30.02/1-309	FIVLLLIYVTSIMGNSGIILLINT-D	SRLFQ-LTYFFLQ-HLAFVDICYTSAITPKMLQSFTFEKN-			
OR11.27.01/1-313	FFMFLFVYVYLVITLGGNIGMILTI	LIWI-DPRLHT-PMYFFLR-HLSFVDICSSSVVTPKMLCNIFAEEKK-			
OR11.38.02/1-308	FLFLFLGIYVVTVVGNLGMIILLI	AV-SPLLHT-PMYFFLS-SLSFVDFCYSSVITPKMLVNFVGKKN-			
OR11.38.03/1-311	FLFLFLGIYVVTVVGNLGMIILLI	AL-SSQLY-PVYFFLS-HLSFIDL	LCYSSVITPKMLVNFVPEEN-		
OR11.38.04/1-309	FLFLFLGIYVVTVVGNLGMIILLI	CL-NSQLY-PMYFFLS-NLSLMD	LCYSSVITPKMLVNFVSEKN-		
OR11.38.05/1-312	FLVFLGIYVVTVVGNLGMIILLI	GL-S	SHLHT-PMYCFLS-SLSFID	FCHSTVITPKMLVNFVTEKN-	
OR11.38.01/1-314	FCLFLGIYVVTVVGNLGMIILLI	SIIRL-NRQLHT-PMYFFLS-SLSFID	FCHSTVITPKMLVNFVTEKN-		
OR11.38.07/1-313	FLLFLGIYVVTVVGNLGMIILLI	GLIIFGL-NSHLHT-PMYFFLR-NLSFID	LCYSSVITPKMLVNFVSEKN-		
OR11.38.09/1-310	FFLFLGIYVVTVVGNLGMIILLI	GLIIFGL-NSHLHT-PMYFFLR-NLSFID	LCYSSVITPKMLVNFVSEKN-		
OR11.38.10/1-311	FLLFLGIYVVTVVGNLGMIILLI	LR-NSHLHT-PMYFFLR-NLSFID	LCYSSVITPKMLVNFVSEKN-		
OR11.38.06/1-309	FLLFLGIYVVTVVGNLGMIILLI	GLIIFGL-NSHLHT-PMYFFLR-NLSFID	LCYSSVITPKMLVNFVSEKN-		
OR11.39.01/1-316	FLVFLVLYGLTMAGNLGMIILLI	SV-DSRLQT-PMYFFLQ-HLALIN	LGNSVIA	PKMLINFLVKKK-	
OR11.39.06/1-312	FALFLMIYVITVVGNLGMIILLI	KL-DSRLQT-PMYFFLR-HLAFMD	LVG	STVGP	PKMLVNFVVDKKN-
OR11.39.07/1-307	FGVFLVIYVITVVGNLGMIILLI	KL-DSHLHT-PMYFFLR-HLAS	LDLGNSTVICPKVLANFVVDNRN-		
OR11.39.05/1-319	FGLFLIYVITVVGNLGMIILLI	YL-DSKLHT-PMYFFLR-HLSIT	DLGYSTVIA	PKMLVNFIVHKN-	
OR11.39.03/1-309	FVLFLSIYVITVVGNLGMIILLI	RA-DTSLNT-PMYFFLS-NLAFVD	FCYSSVITPKMLGNFLYKQN-		
OR11.39.04/1-324	FGVFLVIYVITVVGNLGMIILLI	LIKI-DT	RLHT-PMYFFLR-NLAFVD	LCYSSVITPKMLVNFVSEKN-	
OR11.40.03/1-315	FGVFLAIYVITVVGNLGMIILLI	LR-NSHLHT-PMYFFLQ-HLSFVD	ICYSSVITPKMLVNFVSEKN-		
OR11.40.01/1-305	FVLFLVYVITVVGNLGMIILLI	MLR-DSRLHT-PMYFFLR-NLAFVD	LCYSSVITPKMLVNFVSEKN-		
OR11.40.07/1-307	FIIIFLVYVITVVGNLGMIILLI	KV-S	PQLNN-PMYFFLS-HLSFVD	VWFSNVTPKMLNLFSDKK-	
OR11.40.06/1-310	FVVFLAVYMITVVGNLGMIILLI	SI-S	PQLQS-PMYFFLS-HLSFAD	VCFSSNVTPKMLNLFSETK-	
OR11.40.05/1-311	FTLFLAIYVITVVGNLGMIILLI	QA-NAWLHM-PMYFFLR-NLSFVD	LCYSSVITPKMLNLFSETK-		
OR11.24.02/1-306	FVVFLGVYVITVVGNLGMIILLI	ICN-DSRLHT-PMYFFLR-NLSFVD	LWYSSVHTPKILVTCISEDK-		
OR11.24.03/1-305	FVVFLGVYVITVVGNLGMIILLI	ICN-DSCLHT-PMYFFLR-NLSFVD	LWYSSVHTPKILVTCISEDK-		
OR11.24.01/1-327	FGVFLMLYVITVVGNLGMIILLI	IR-DSHLHT-PMYFFLR-NLSFVD	LWYSSVHTPKILVTCISEDK-		
OR11.26.02/1-315	FVTFLGIYVITVVGNLGMIILLI	RG-DT	HLHT-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-	
OR11.26.03/1-324	FMLFLGLYVITVVGNLGMIILLI	KM-DS	HLHM-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-	
OR11.26.01/1-311	FTIFLVYVITVVGNLGMIILLI	RM-DS	HLHT-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-	
OR11.25.01/1-314	FLVFLSIYVITVVGNLGMIILLI	QV-DVKLYT-PMYFFLR-NLSL	LDACYSVITPQILATLATGKT-		
OR11.25.02/1-310	FLLLLFLMYVITVVGNLGMIILLI	LM-DH	QLHA-PMYFFLR-NLAFMD	VCYSSVITVPMQLAVLLEHGA-	
OR06.08.01/1-321	FTIFFLTYVITVVGNLGMIILLI	VT-D	PHLHT-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-	
OR06.08.02/1-317	FTIFFLTYVITVVGNLGMIILLI	IA-DS	HLHT-PMYFFLR-NLAL	IDICYSSAVAPNMLTDFFWEQK-	
OR12.04.01/1-309	FIFFLTYVITVVGNLGMIILLI	LL-D	PHLHT-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-	
OR12.04.02/1-309	FTIFFLAYVITVVGNLGMIILLI	LL-D	SHLHT-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-	
OR11.42.06/1-301	FSIFLLMYVITVVGNLGMIILLI	GIILLIKI-HPALQT-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-		
OR12.02.01/1-313	FLLFLFVYVITVVGNLGMIILLI	MMTIIMT-DPRLHT-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-		
OR11.46.01/1-317	FILFLMYVITVVGNLGMIILLI	VVGL-DH	RLRR-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-	
OR11.23.01/1-319	FGAIIYVITVVGNLGMIILLI	FT-DS	HLQS-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-	
OR15.01.01/1-312	FVLFLGIYVITVVGNLGMIILLI	RA-DS	CLHK-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-	
OR12.01.01/1-317	FALFLALYSITVVGNLGMIILLI	FTSWTDPKLNLS-PMYFFLR-NLSFVD	ICYSSAVAPNMLTDFFWEQK-		

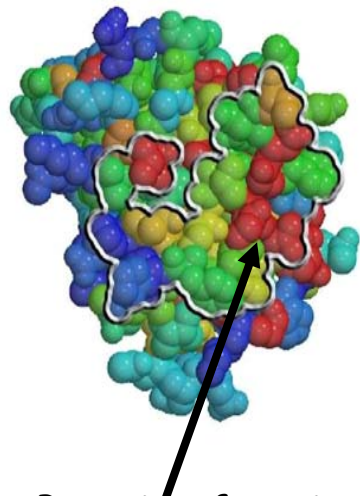


Identification of functional subfamilies

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



Structural model



Putative functional site

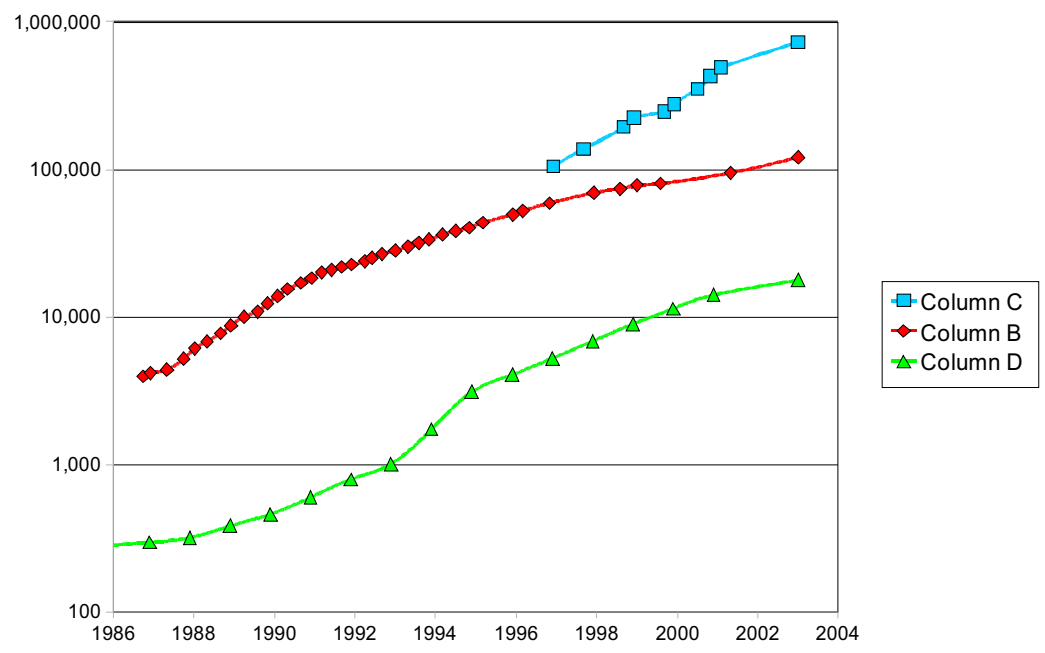
1 = highly conserved



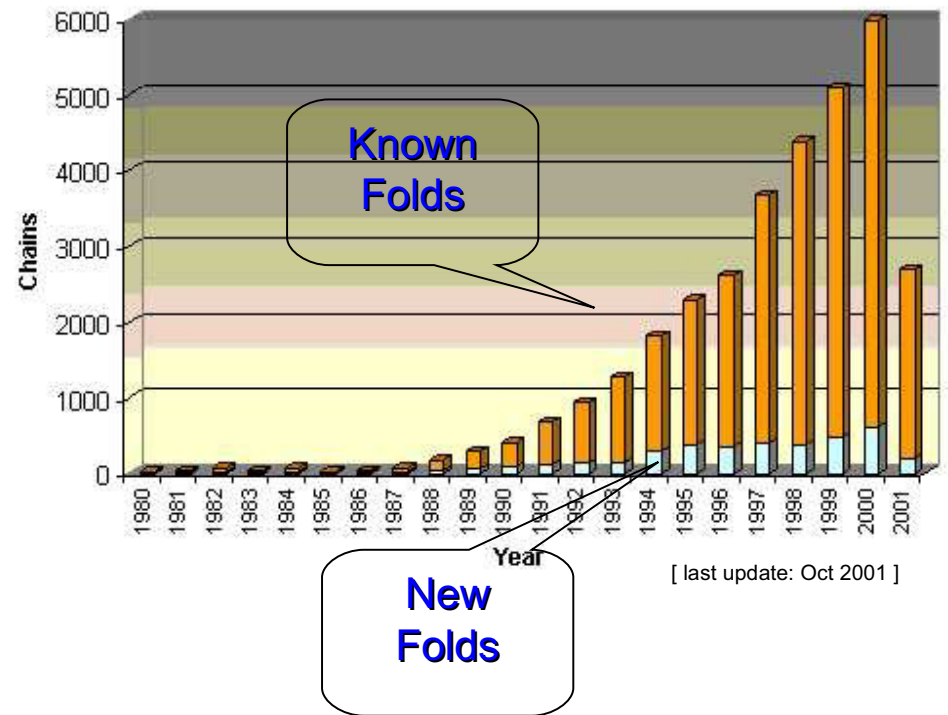
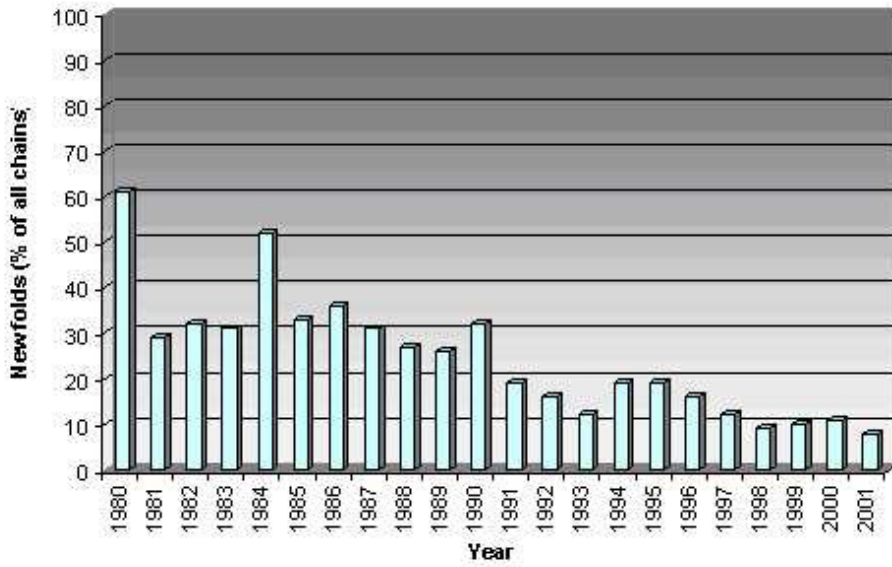
0 = unconserved

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

Public Database Holdings:



The number of different protein folds is limited:

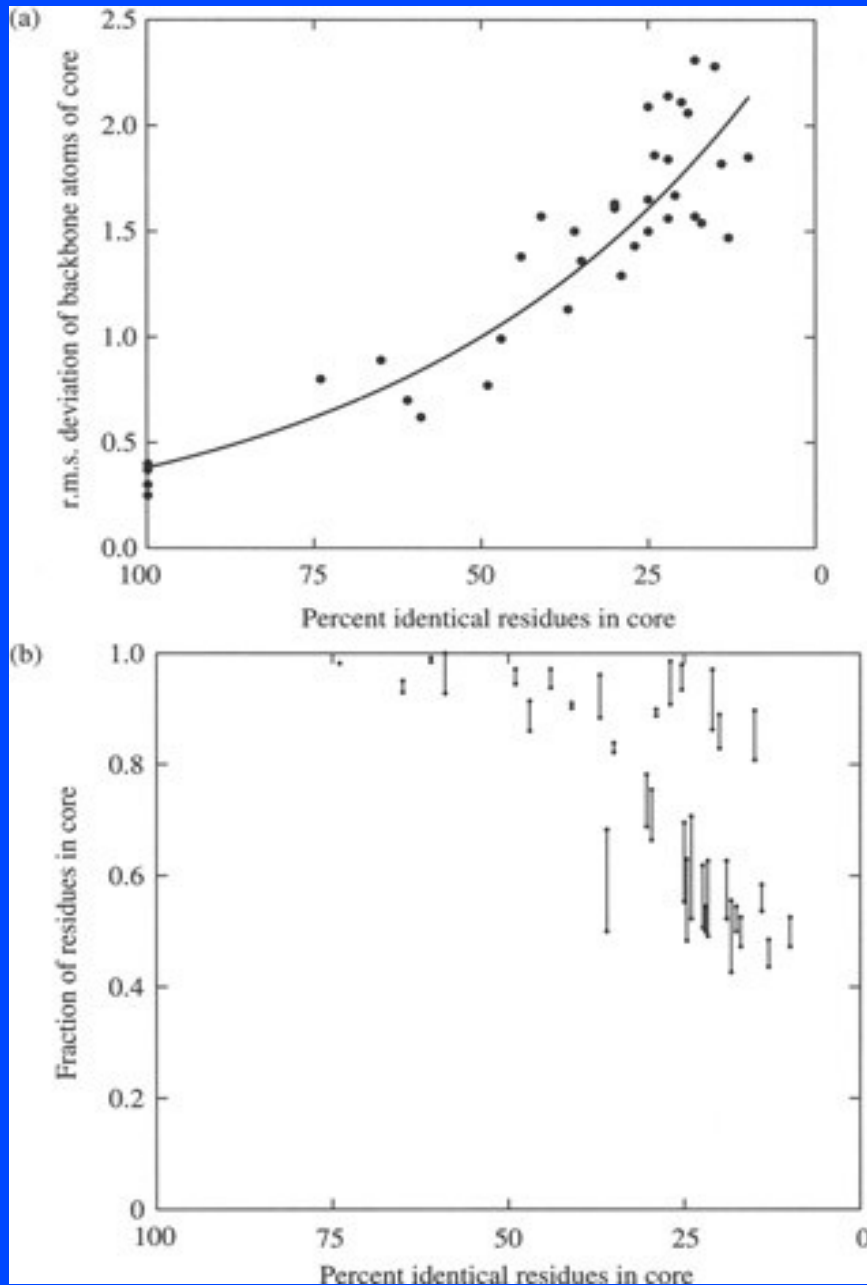


Protein evolution

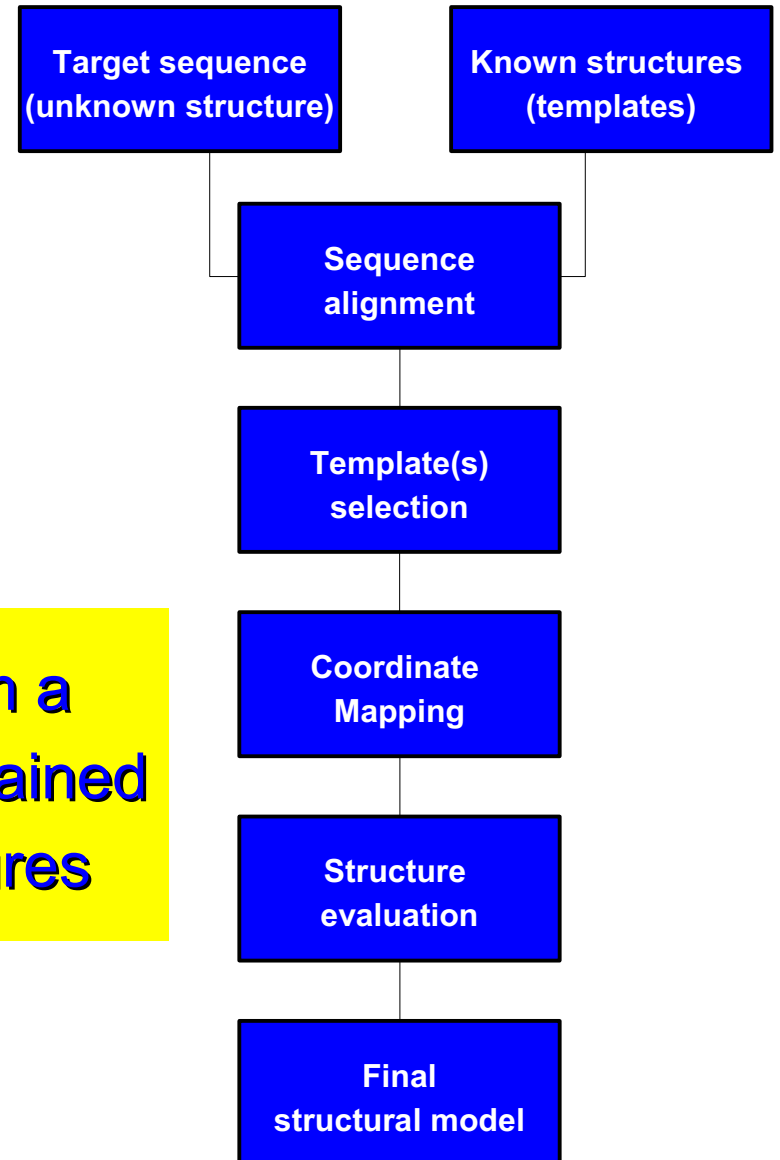
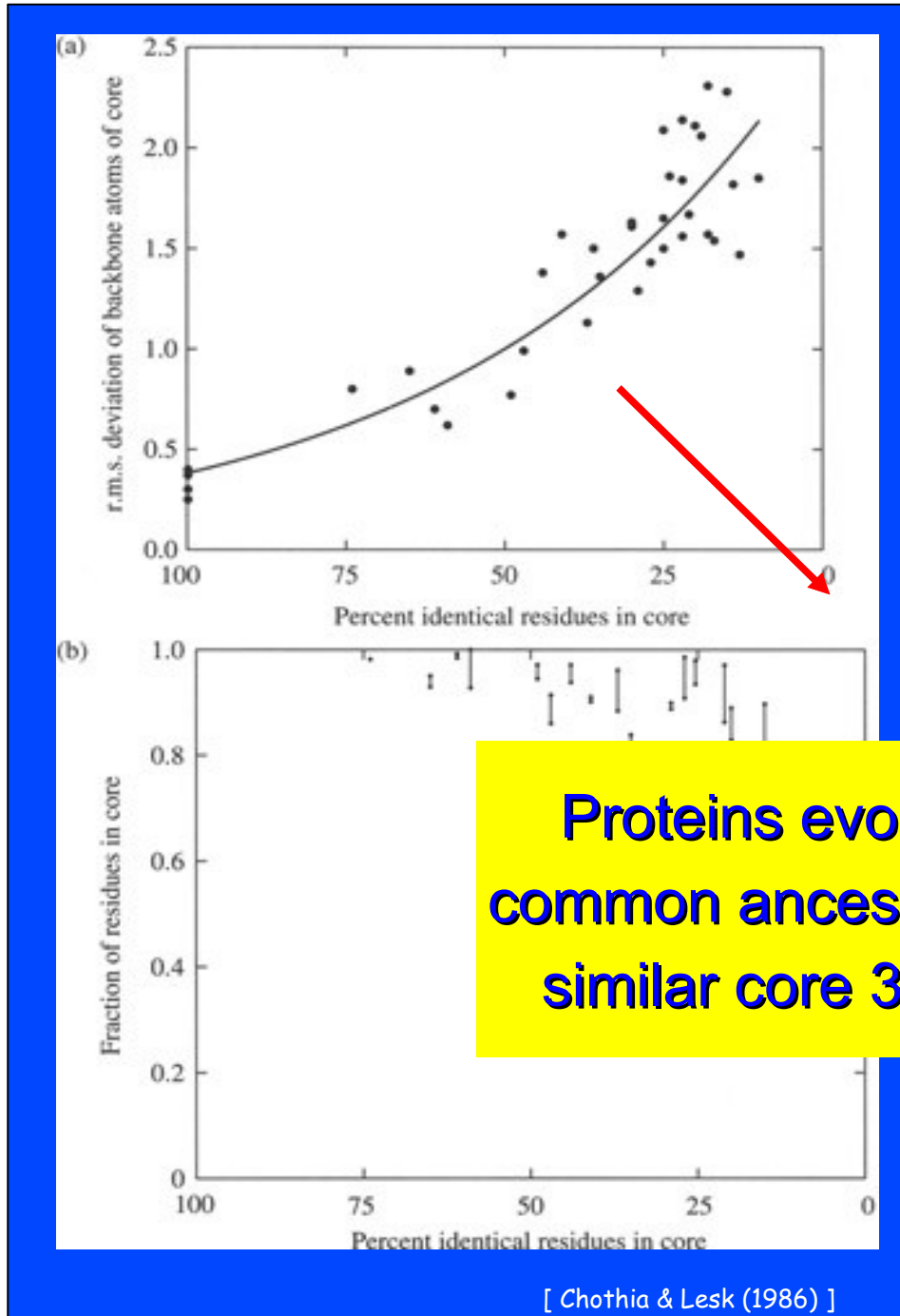
Comparative modelling

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

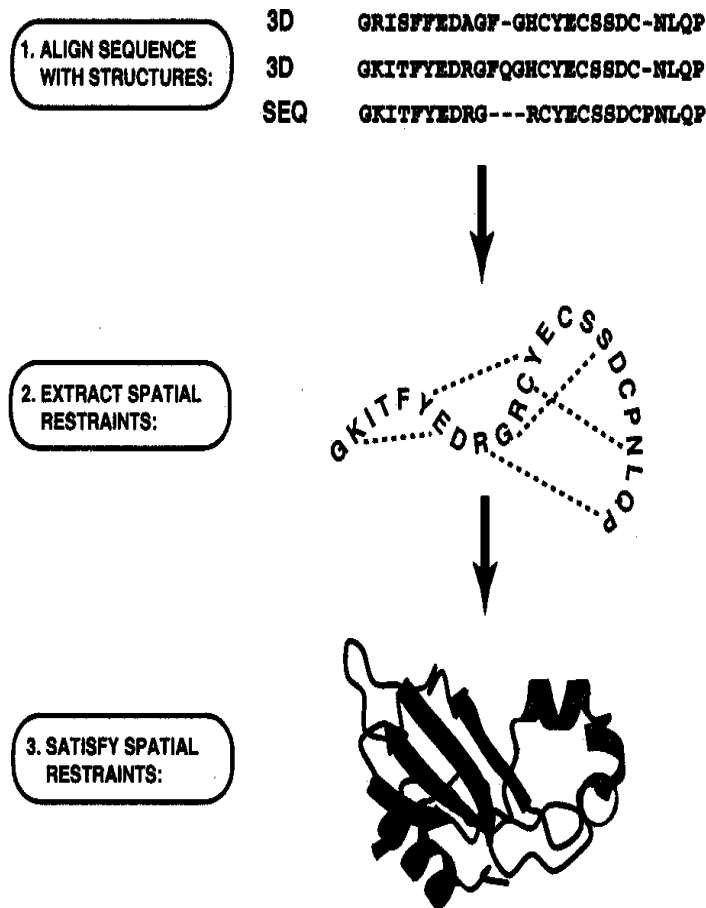


Homology modeling



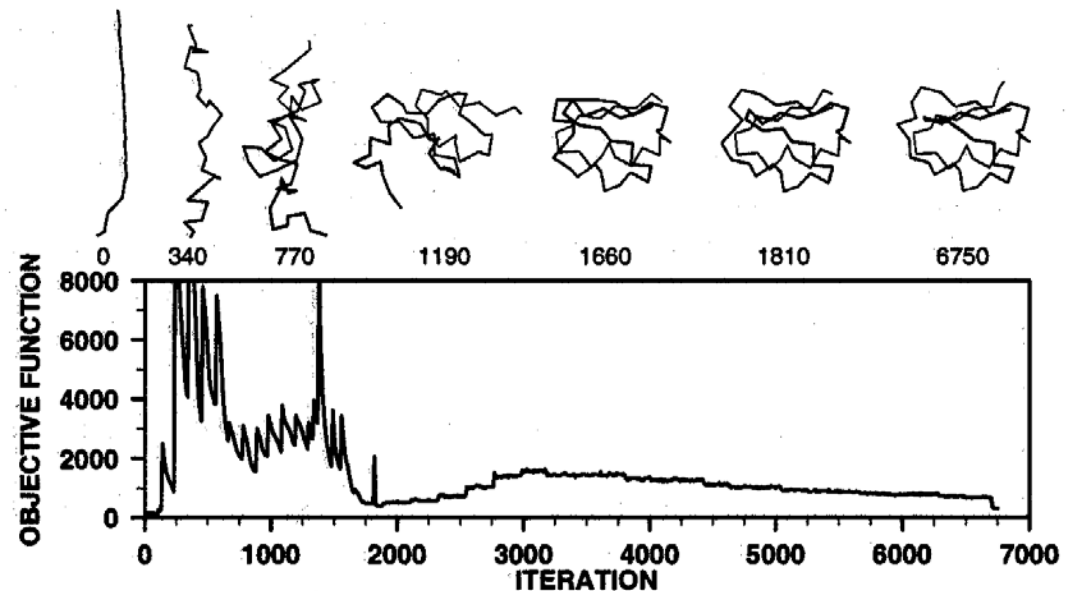
II. Modeling by Satisfaction of Spatial restraints

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



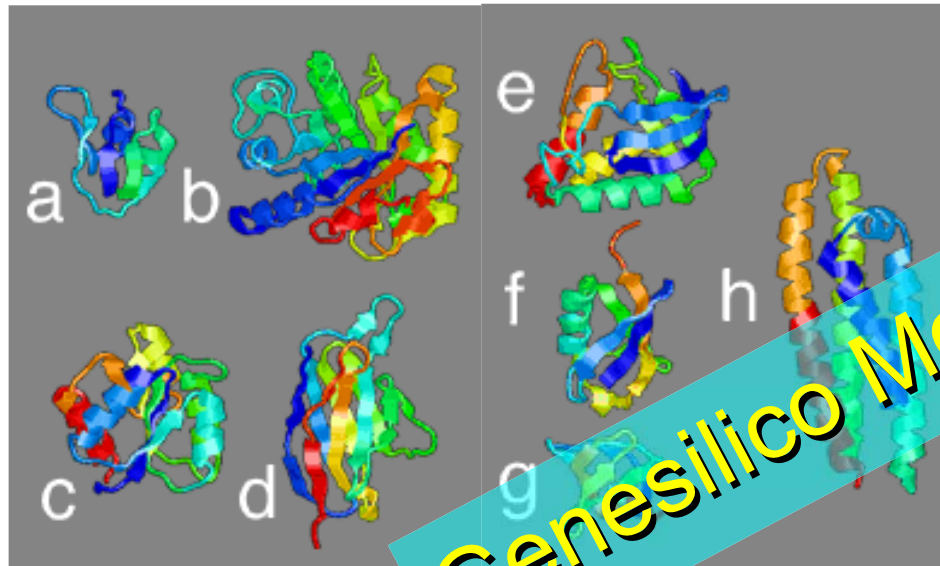
Comparative protein modeling by satisfaction of spatial restraints.

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

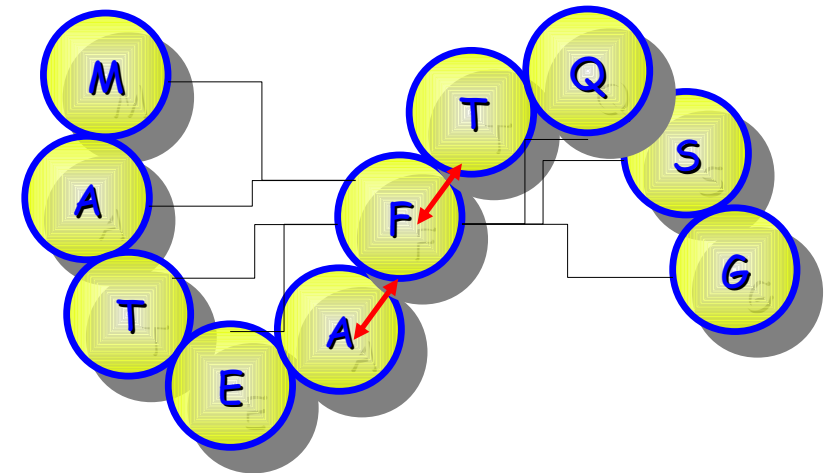
For each aa we can calculate its frequency:

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

Each aa will be substituted by an intrinsic character

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

Threading

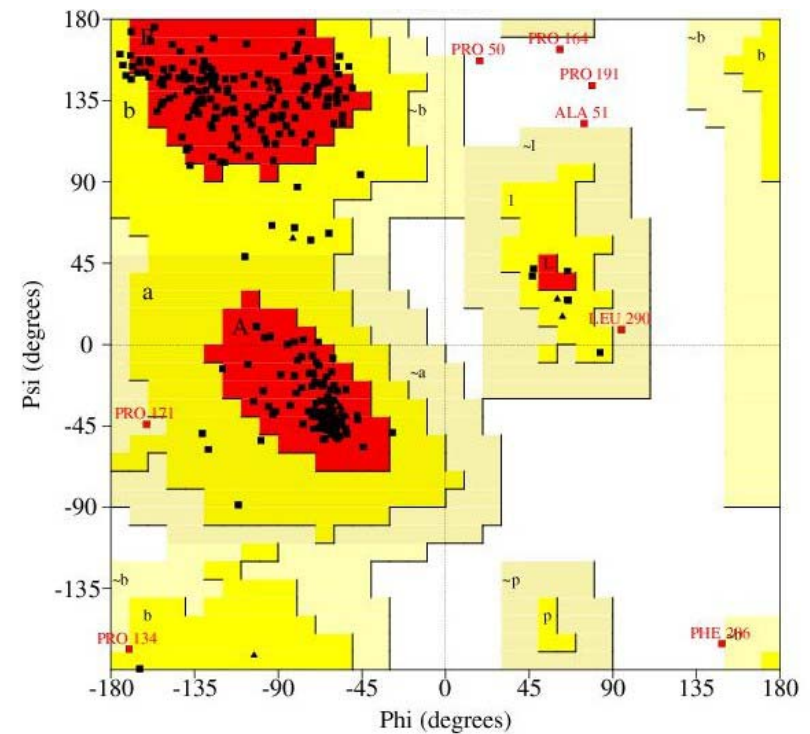


$$\Delta E (Ala - Ala) = -kT \ln \left(\frac{P_{folded} (Ala - Ala)}{P_{unfolded} (Ala - Ala)} \right)$$

Model Evaluation ?

Topics:

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



Plot statistics

Residues in most favoured regions [A,B,L]	224	88.2%
Residues in additional allowed regions [a,h,l,p]	27	10.6%
Residues in generously allowed regions [-a,-b,-l,-p]	1	0.4%
Residues in disallowed regions	2	0.8%
Number of non-glycine and non-proline residues	254	100.0%
Number of end-residues (excl. Gly and Pro)	5	
Number of glycine residues (shown as triangles)	17	
Number of proline residues	21	
Total number of residues	297	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.


NIH MBI Laboratory for Structural Genomics and Proteomics
[\[Servers Home\]](#)

People Seminars
 Lectures Webmail
 Links Facilities
 Software Home



STRUCTURE VALIDATION

Help!

PDB File upload:

Structure Factor upload (Optional) (mtz format):

➤ Structure /

<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: output = 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 Secondary structure 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)

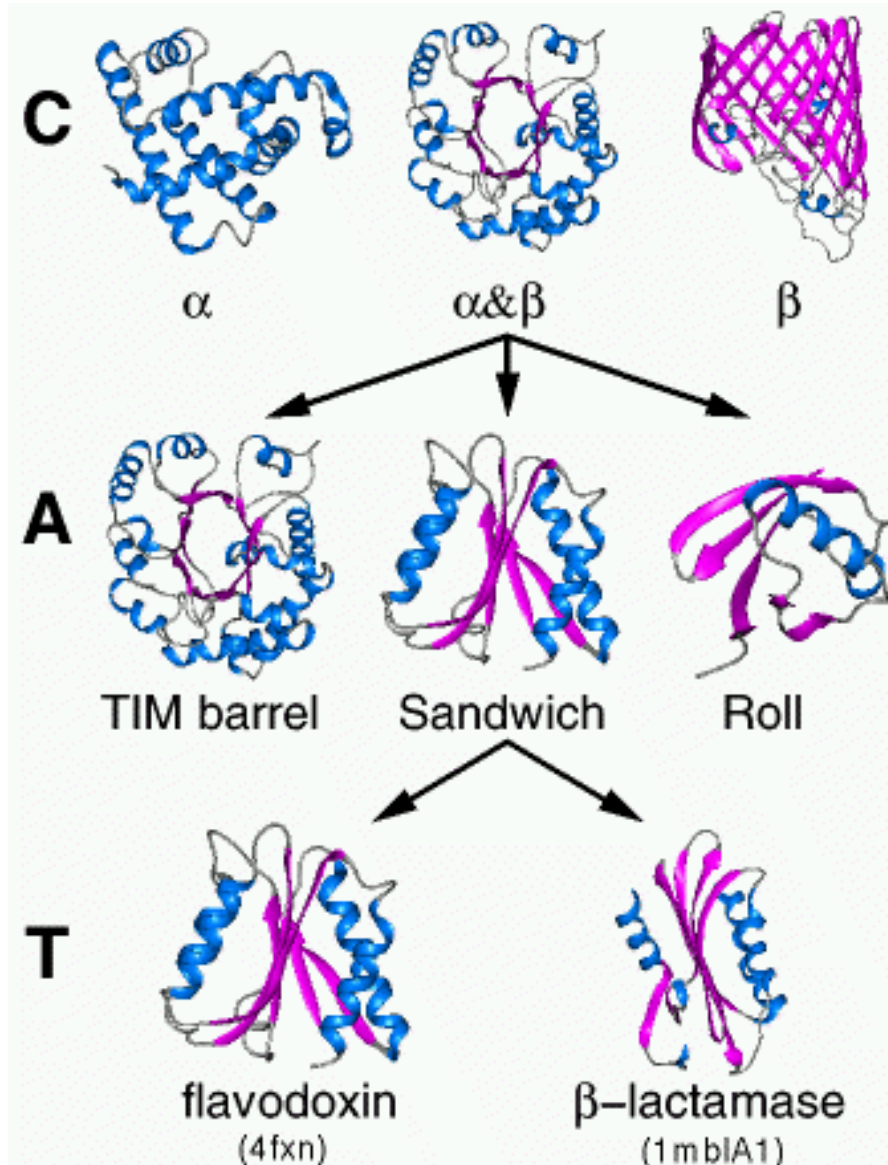
SCOP - Structural Classification of Proteins

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/>