

# A Very Very Very Short Introduction to Protein Bioinformatics

developed by  
Patricia Babbitt, Susan Johns, Leslie King & Sean Mooney  
University of California, San Francisco  
July 18, 2002

## Tools for Protein Informatics

- sequence and structure comparison
- multiple alignments
- phylogenetic tree construction
- composition/pI/mass analysis
- motif/pattern identification
- 2° structure prediction/threading
- TMD prediction/hydrophobicity analysis
- homology modeling
- visualization

## Primary Web Resources

- **European Molecular Biology Laboratory, Germany**  
[http:// www.embl-heidelberg.de](http://www.embl-heidelberg.de)
- **ExpASy Molecular Biology Server, Swiss Institute of Bioinformatics, Switzerland**  
<http://ca.expasy.org/>
- **National Center for Biotechnology Information, USA**  
[http:// www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)  
<http://www3.ncbi.nlm.nih.gov/Entrez/>
- **San Diego Supercomputer Center, USA**  
[http:// www.sdsc.edu](http://www.sdsc.edu)

## Other valuable on-line sites

- **Entrez**  
<http://www3.ncbi.nlm.nih.gov/Entrez/>
- **Genome mapping and sequencing**
  - **Human genome project:**  
<http://www3.ncbi.nlm.nih.gov/genome/guide/>
  - **Model organisms:**  
<http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>
  - **Whole genome analysis:**  
<http://www.ncbi.nlm.nih.gov/COG/>
  - **Analysis of polymorphisms:**  
<http://www.ncbi.nlm.nih.gov/SNP/>

- **Functional genomics:**
  - **Online Mendelian Inheritance in Man (OMIM):**  
<http://www.ncbi.nlm.nih.gov/Omim/>
- **Target identification in drug design, agriculture, biocatalysis:**  
<http://www.labmed.umn.edu/umbbd/index.html>
- **Differential digital display (Cancer genome anatomy project):**  
<http://www.ncbi.nlm.nih.gov/ncicgap/>
- **Array technologies:**  
<http://cmgm.stanford.edu/pbrown/>
- **Metabolic pathways:**  
<http://www.ecocyc.org/>  
<http://www.genome.ad.jp/kegg/>

## Primary databases for 3D structure classification/information

- **Entrez**  
<http://www3.ncbi.nlm.nih.gov/Entrez/>
- **Protein Data Bank (PDB)**  
<http://www.rcsb.org/pdb/>
- **Structural Classification of Proteins (SCOP)**  
<http://scop.mrc-lmb.cam.ac.uk/scop/index.html>
- **CATH: Protein Structure Classification**  
[http://www.biochem.ucl.ac.uk/bsm/cath\\_new/index.html](http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html)

## Protein vs. nucleic acid sequence analysis?

- Protein sequence analysis provides greater specificity and less noise than nucleic acid analysis for identification of similarities because of the inherent differences in the message content of nucleic acid and amino acid codes
- Due in part to 4-letter vs. 20-letter code, degeneracy of codon messaging
- But some searches must be done at the nucleotide level...

## Some information properties of messages for sequence analysis

- A sequence can be described in terms of the # of bits needed to specify its message, where one bit distinguishes between two equally likely things.  
Ex: Where base frequencies are equal, one bit distinguishes a purine from a pyrimidine, two bits are required to uniquely specify a single base among A, T, C, G.
- Information content of a random message can be calculated from the set of relevant symbols' frequencies:

$$I = \sum_{i=1}^n P_i \log_2 P_i$$

where  $P_i$  is the probability of finding the symbol  $i$  at any position

- Using a standard measure for overall amino acid frequencies gives the information content of a random protein sequence as 4.19 bits/residue.
- Thus, for an average size protein domain (150 residues), the message length is ~630 bits and the probability that 2 random sequences would specify the same message is  $2^{-630}$  ( $10^{-190}$ ).
  - ∴ Database searching for protein similarities is doable, even for fairly short sequences
- BUT, for a transcription binding site of 8-10 bp, the odds of 2 random sequences arriving at the same message is  $10^{-5}$ .
  - ∴ Database searching for regulatory elements does not work well as databases get larger

## Introduction to Protein Sequence Analysis

- Database searching/pairwise alignments
- Pattern searching and motif analysis
- Multiple alignments and Evaluation using Family/Superfamily Concepts

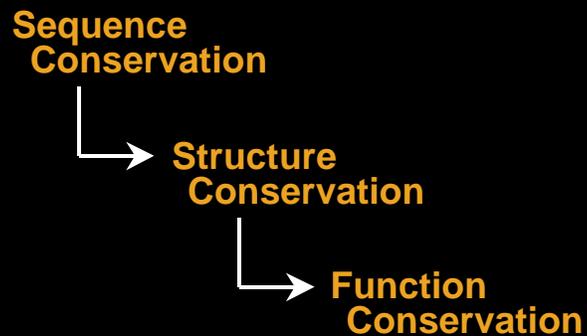
## Applications

- tracing ancestral connections
- deduction/inference of function
- understanding enzyme mechanisms
- clustering of families, superfamilies
- structural analysis of receptors, molecules involved in cell signaling
- identification of molecular surfaces in protein-protein, protein-DNA interactions
- metabolic computing/comparative genome analysis
- guidance for functional genomics, protein engineering

## References: Database searching

- Altschul et al., "Issues in searching molecular sequence databases"
- Pearson, "Comparison of methods for searching protein sequence databases"
- Altschul, "Amino acid substitution matrices from an information theoretic perspective"
- Pearson & Lipman, (the original FASTA paper) "Improved tools for biological sequence comparison"
- Altschul et al., (the original Blast paper) "Basic local alignment search tool"
- Henikoff & Henikoff, "Amino acid substitution matrices from protein blocks"
- Altschul et al., "Gapped Blast and PSI-Blast: A new generation of protein database programs"

## The underlying assumption used in functional inference...



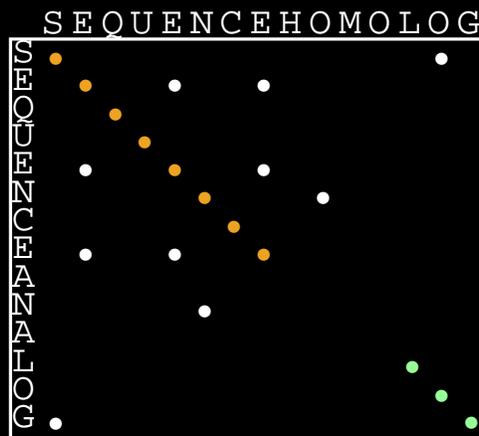
## ...requires comparison of sequences

- The most fundamental operation in protein informatics is finding the best alignment between a query sequence and one or more additional sequences
- Once candidate homologs have been identified, they can be evaluated using statistical methods and structural and biological information
- The correspondence between two aligned sequences can be expressed in a similarity score and/or viewed graphically, *e.g.*, dot plots, alignments, motifs or patterns

## Formalizing the Problem

- Given: two sequences that you want to align
- Goal: find the best alignment that can be obtained by sliding one sequence along the other
- Requirements:
  - a scheme for evaluating matches/mis-matches between any two characters
  - a score for insertions/deletions
  - a method for optimization of the total score
  - a method for evaluating the significance of the alignment

- Dot matrix plots: a simple description of alignment operations illustrating types of relationships between a sequence pair



- The signal-to-noise ratio can be improved using filtering techniques designed to minimize the composition-dependent background
- Example of common filters: over-lapping, fixed-length "windows" for sequence comparison
- To be counted, a comparison must achieve a minimum threshold score summed over the window, derived empirically or from a statistical or evolutionary model of sequence similarity
- The window size and minimum threshold score (often termed "stringency") at which the score is counted can be user-defined

Seq1 = SEQUENCEHOMOLOG  
 Seq2 = SEQUENCEANALOG  
 Window = 7, Stringency = 42% (3/7 matches)

SEQUENC  
 SEQUENCEANALOG (7/7 matches)

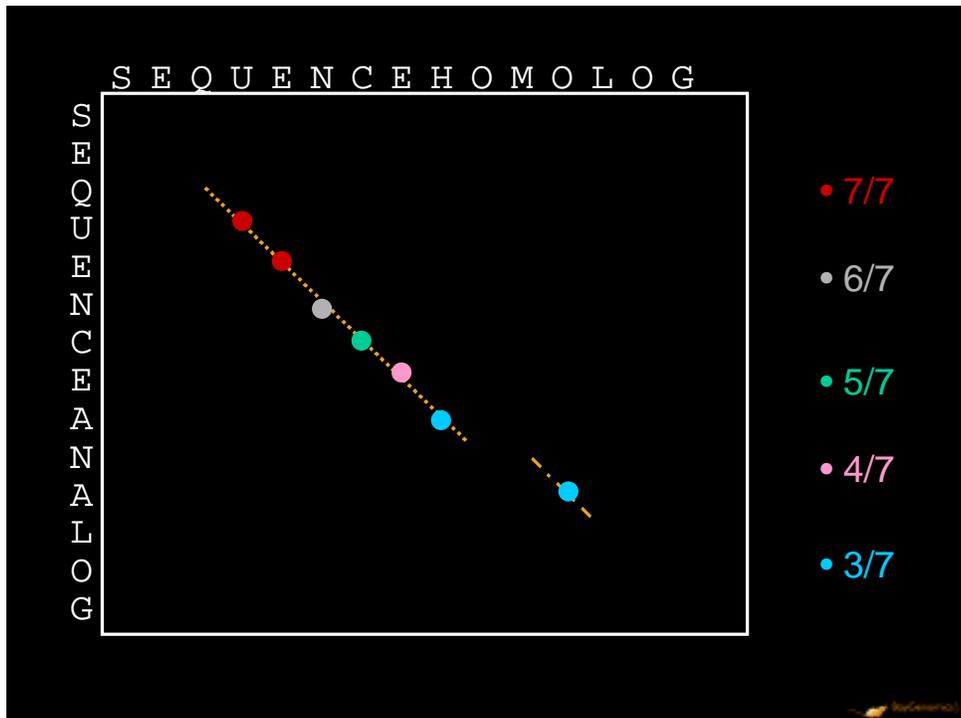
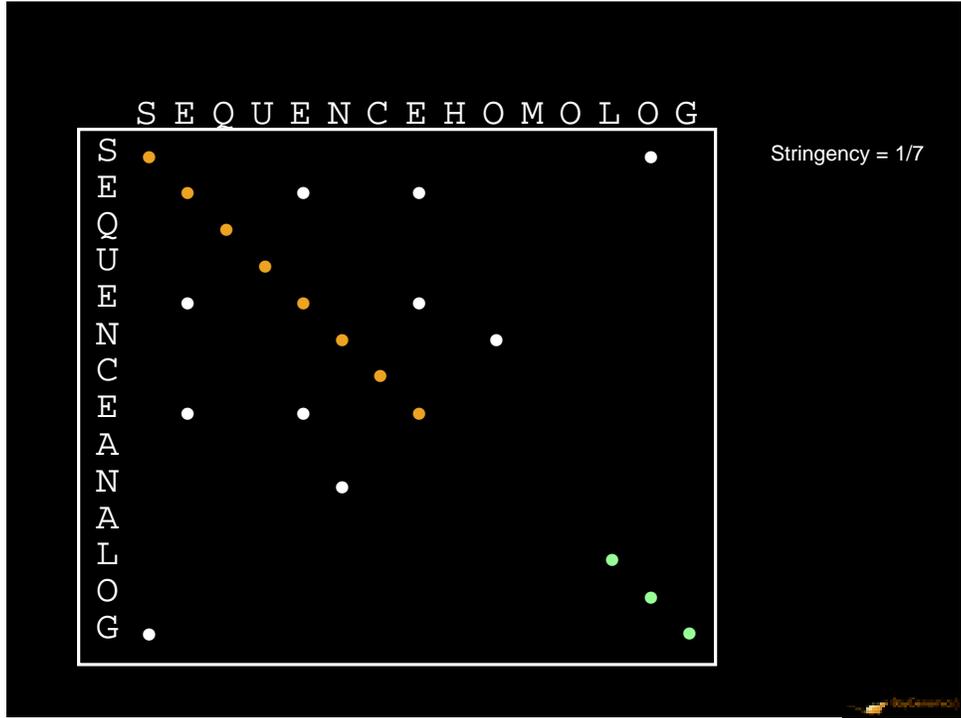
SEQUENC  
 SEQUENCEANALOG (0/7 matches)

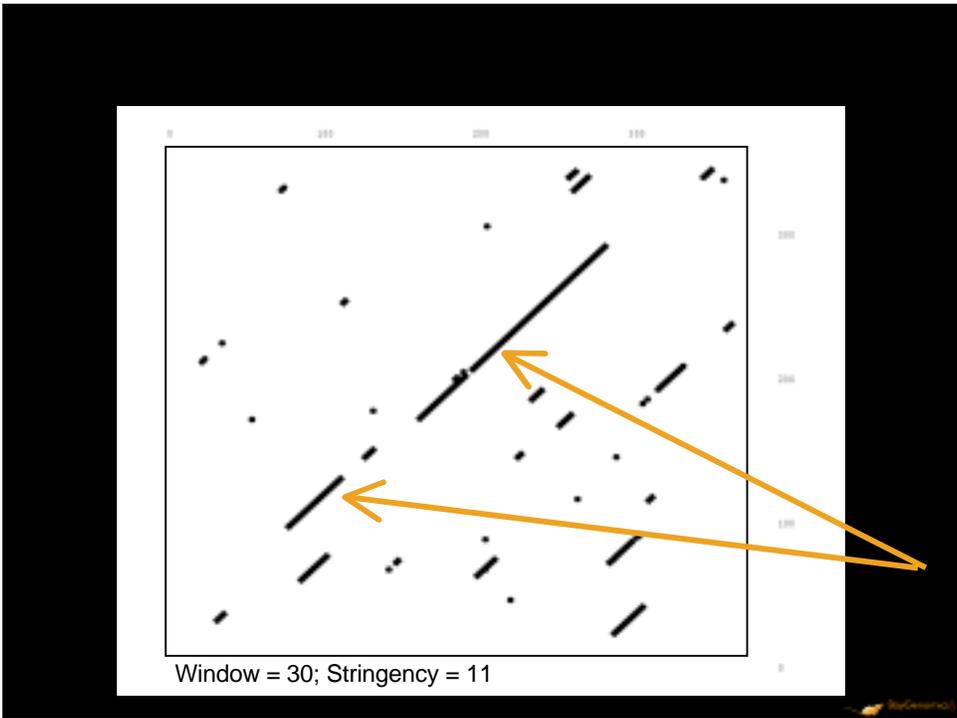
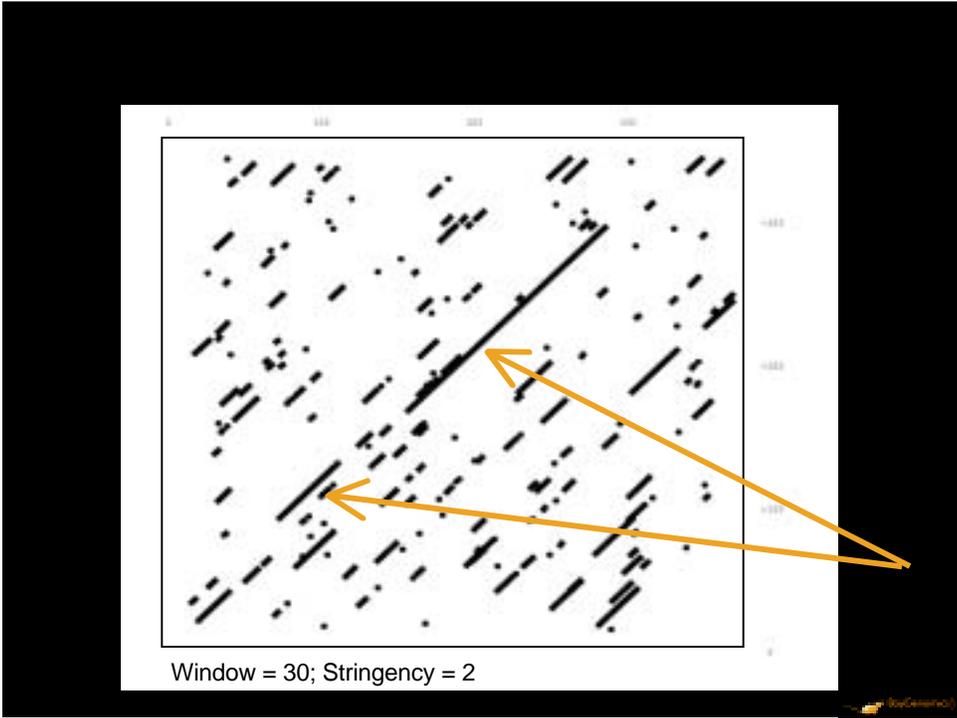
...

SEQUENC  
 SEQUENCEANALOG (7/7 matches)

...

HOMOLOG  
 SEQUENCEANALOG (3/7 matches)





## Scoring Systems

- The degree of match between two letters can be represented in a matrix
- Changing the matrix can change the alignment
  - Simplest: Identity (unitary) matrix
  - Better: Definitions of similarity based on inferences about chemical or biological properties
  - Examples: PAM, Blosum, Gonnet matrices
- The score should have the form:  $p_{ab}/q_a q_b$ , where  $p_{ab}$  is the probability that residue  $a$  is substituted by residue  $b$ , and  $q_a$  and  $q_b$  are the background probabilities for residue  $a$  and  $b$  respectively.
- Handling gaps remains an incompletely solved problem...

## PAM units

- PAM (point accepted mutation) is a unit of evolutionary distance between 2 amino acid sequences\*
  - 1 PAM = 1 accepted point-mutation (no insertions or deletions) event per 100 aa
  - 200 PAM = 200 point mutations/100 aa (assumes mutations occur multiple times at any given position)
  - 2 sequences diverged by 200 PAM  $\cong$  25% identity
- \*PAM is also sometimes defined as "percent accepted mutation"

## PAM matrices

- Substitution matrices used to reflect expected evolutionary change (by point mutations only)
- Given 2 sequences  $i, j$ , for any specific pair of residues  $A_i, A_j$ , the  $(i, j)$  entry in the PAM  $n$  matrix reflects the frequency at which  $A_i$  is expected to replace  $A_j$  in 2 sequences  $n$  PAM units diverged, *i.e.*, use PAM120 matrix to compare 2 protein sequences diverged by 120 PAM units
- Score should be in the form

$$\frac{p_{ij}}{p_i p_j}$$

- Usually presented in *log-odds* form, *i.e.*, probability values are given in logarithmic form

## Derivation of ideal PAM matrices\*

- Using many sets of 2 aligned sequences, for each amino acid pair  $A_i, A_j$ , count the # of times  $A_i$  aligns with  $A_j$  and divide that number by the total # of amino acid pairs in all of the alignments, resulting in the frequency,  $f(i, j)$
- Let  $f_i$  and  $f_j$ , respectively, denote the frequencies at which  $A_i$  and  $A_j$  appear in the sets of sequences
- Then the  $(i, j)$  entry for the ideal PAM matrix is

$$\log \frac{f(i, j)}{f(i) f(j)}$$

\*adapted from *Algorithms on Strings, Trees, and Sequences*, Dan Gusfield, 1997

## Actual Derivation of PAM matrices

- Originally compiled from a group of sequences >85% identical that could be unambiguously aligned  
(M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, in *Atlas of Protein Sequence and Structure*, 5:345-352 (1978))
- These sequences were close in length and the few insertions/deletions could be placed correctly
- A PAM-1 matrix was calculated from these data
- Assumes more distantly related proteins can be described by a series of uncorrelated mutations consistent with the PAM-1 matrix such that a PAM-N matrix is derived by multiplying PAM-1 by itself N times

## Guidelines for using PAM matrices

The relative entropy H of PAM matrices (from Table 1)		
PAM distance	H (bits)	Min. signif length (30 bits)
40	2-26	14
120	0-98	31
250	0-36	83

Ranges of local alignment lengths for which various PAM matrices are appropriate (from Table 3)	
PAM matrix	93% efficiency range for database searching (30 bits)
40	9-21
120	19-50
240	47-123

from Altschul, "Amino acid substitution matrices from an information theoretic perspective"

## PAM250 amino acid substitution matrix\*

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
A	2	0	-2	0	0	-4	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3	0
B	0	2	-4	3	2	-5	0	1	-2	1	-3	-2	2	-1	1	-1	0	0	-2	-5	-3	2
C	-2	-4	12	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0	-5
D	0	3	-5	4	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-7	-4	3
E	0	2	-5	3	4	-5	0	1	-2	0	-3	-2	1	-1	2	-1	0	0	-2	-7	-4	3
F	-4	-5	-4	-6	-5	9	-5	-2	1	-5	2	0	-4	-5	-5	-4	-3	-3	-1	0	7	-5
G	1	0	-3	1	0	-5	5	-2	-3	-2	-4	-3	0	-1	-1	-3	1	0	-1	-7	-5	-1
H	-1	1	-3	1	1	-2	-2	6	-2	0	-2	-2	2	0	3	2	-1	-1	-2	-3	0	2
I	-1	-2	-2	-2	-2	1	-3	-2	5	-2	2	2	-2	-2	-2	-2	-1	0	4	-5	-1	-2
K	-1	1	-5	0	0	-5	-2	0	-2	5	-3	0	1	-1	1	3	0	0	-2	-3	-4	0
L	-2	-3	-6	-4	-3	2	-4	-2	2	-3	6	4	-3	-3	-2	-3	-3	-2	2	-2	-1	-3
M	-1	-2	-5	-3	-2	0	-3	-2	2	0	4	6	-2	-2	-1	0	-2	-1	2	-4	-2	-2
N	0	2	-4	2	1	-4	0	2	-2	1	-3	-2	2	-1	1	0	1	0	-2	-4	-2	1
P	1	-1	-3	-1	-1	-5	-1	0	-2	-1	-3	-2	-1	6	0	0	1	0	-1	-6	-5	0
Q	0	1	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4	1	-1	-1	-2	-5	-4	3
R	-2	-1	-4	-1	-1	-4	-3	2	-2	3	-3	0	0	0	1	6	0	-1	-2	2	-4	0
S	1	0	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2	1	-1	-2	-3	0
T	1	0	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3	0	-5	-3	-1
V	0	-2	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4	-6	-2	-2
W	-6	-5	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	0	-6
Y	-3	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10	-4
Z	0	2	-5	3	3	-5	-1	2	-2	0	-3	-2	1	0	3	0	0	-1	-2	-6	-4	3

\*Version in use by GCG9

B,Z are average values for D/N and E/Q, respectively

## Issues with PAM matrices

- Actually work quite well, with PAM-250 still used routinely for finding distant homologs
- BUT there are some clear problems with the model...
  - PAM model assumes all residues are equally mutable
  - Model devised using the most mutable positions rather than the most conserved positions, *i.e.*, those that reflect chemical and structural properties of importance
  - Derived from a biased set of sequences: small globular proteins available in the database in 1978

## BLOSUM (Blocks Substitution) Matrices

- Derived from the BLOCKS database, which, in turn is derived from the PROSITE library  
<http://blocks.fhcrc.org/blocks/>, <http://www.expasy.ch/prosite/>
- BLOCKS generated from multiply aligned sequence segments without gaps clustered at various similarity thresholds and corrected to avoid sampling bias
- Derived from data representing highly conserved sequence segments from divergent proteins rather than data based on very similar sequences (as with PAM matrices)

## Derivation of BLOSUM matrices

- Many sequences from aligned families are used to generate the matrices
- Sequences identical at >X% are eliminated to avoid bias from proteins over-represented in the database
- Specific matrices refer to these clustering cut-offs, *i.e.*, BLOSUM62 reflects observed substitutions between segments <62% identical
- In analogy to PAM matrices, a log-odds matrix is calculated from the frequencies  $A_{ij}$  of observing residue  $i$  in one cluster aligned against residue  $j$  in another cluster

## BLOSUM vs. PAM Matrices

- BLOSUM matrices have replaced PAM matrices as the default matrices at many database searching sites (Blast, FASTA servers)
- Both PAM-120 and BLOSUM62 work best for moderately diverged proteins and may miss similarities outside their optimum performance windows
- PAM provides the only easily accessible alternative for short sequences (no appropriate version of Blosum available)
- Best solution is to provide a range of scoring systems, which is currently the practice for most primary servers
- Setting appropriate gap penalties can have a large effect on matrix performance

## Optimizing the Score: Brute-force Approach

- Considering two sequences, both of length N:
  - If gaps or local alignments are not considered, there is only one optimal solution

Equal length w/o gap



- The computational time required to compute the optimal alignment =  $N^2$

- But when gaps or local alignments are considered, things get complicated because we have to repeat the calculation  $2N$  times to allow for the possibility of gaps at each position of each sequence
- Requires time proportional to  $N^{4N}$
- Even when nonsensical alignments are removed (aligning gaps with gaps), for  $N = 300$  residues,  $\sim 10^{88}$  comparisons are required

Extreme case: seq 1 aligns with gaps of seq 2



## Optimizing the Score: Dynamic Programming

- Requires computational time proportional to  $N^2$
- Original version often termed the “Needleman-Wunsch” algorithm  
(Needleman, S.B. and Wunsch, C.D. *J. Mol. Biol.* 48 (1970) 443-453)
- Addresses the problem for GLOBAL alignments; still has to deal with gaps

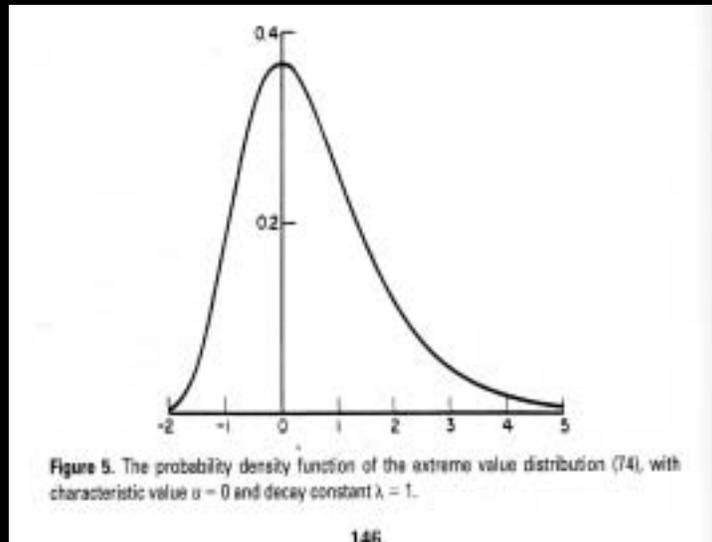
## Next step forward: local alignments

- Implemented by Smith & Waterman  
(Smith & Waterman. J. Mol. Biol. 147 (1981) 195-197)
- Finds the two “most similar” segments to generate an alignment from parts of the two sequences
- Modifications of dynamic programming algorithm:
  - The scoring system must include negative scores for mismatches
  - 0 = the minimum score recorded in the score matrix
  - The end of the optimal path can be anywhere in the matrix, not just in the last row or column

## Statistical Significance

- A good way to determine if an alignment score has statistical meaning is to compare it with the score generated from the alignment of two random sequences
- A model of ‘random’ sequences is needed. The simplest model chooses the amino acid residues in a sequence independently, with background probabilities

(Karlin & Altschul (1990) Proc. Natl. Acad. Sci. USA, 87 (1990) 2264-2268)



## A most important caveat...

- For database searches, the **ONLY** criteria available to judge the likelihood of a structural or evolutionary relationship between 2 sequences is an estimate of statistical significance
- For a medium-sized protein using default parameters (Blosum62,  $E = 10$ ), the cut-off for statistical significance is  $P = 10^{-7} - 10^{-5}$   
( for the relationship between E and P, see <http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>)
- Statistical significance and biological significance are **NOT** necessarily the same

Query= /phosphonatase/phosSt.gcg (255 letters) (10/20/99/pcb)  
 Database: /mol/seq/blast/db/swissprot 78,725 sequences; 28,368,147 total letters

Sequences producing significant alignments:			Score	E
			(bits)	Value
sp O06995 PGMB_BACSU	Begin: 93 End: 204	PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BETA-PGM)	38	0.020
sp P31467 YIEH_ECOLI	Begin: 1 End: 180	HYPOTHETICAL 24.7 KD PROTEIN IN TNAB-BGLB I...	36	0.10
sp O14165 YDX1_SCHPO	Begin: 34 End: 201	HYPOTHETICAL 27.1 KD PROTEIN C4C5.01 IN CHR...	31	2.6
sp P41277 GPP1_YEAST	Begin: 133 End: 200	(DL)-GLYCEROL-3-PHOSPHATASE 1	30	4.4
sp Q39565 DYHB_CHLRE	Begin: 3911 End: 4032	DYNEIN BETA CHAIN, FLAGELLAR OUTER ARM	29	7.6
sp P77625 YFET_ECOLI	Begin: 143 End: 187	HYPOTHETICAL 23.7 KD PROTEIN IN LRHA-ACKA I...	29	10.0
sp Q40297 FCPA_MACPY	Begin: 146 End: 176	FUcoxANTHIN-CHLOROPHYLL A-C BINDING PROTEIN...	29	13
sp P40853 GPHP_ALCEU	Begin: 94 End: 188	PHOSPHOGLYCOLATE PHOSPHATASE, PLASMID (PGP)	29	13
sp Q40296 FCPB_MACPY	Begin: 146 End: 176	FUcoxANTHIN-CHLOROPHYLL A-C BINDING PROTEIN...	29	13
sp P52186 ANNU_SCHAM	Begin: 119 End: 168	ANNULIN (PROTEIN-GLUTAMINE GAMMA-GLUTAMYLTR...	29	13
sp P40106 GPP2_YEAST	Begin: 133 End: 200	(DL)-GLYCEROL-3-PHOSPHATASE 2	28	17
sp P37934 MAY3_SCHCO	Begin: 435 End: 552	MATING-TYPE PROTEIN A-ALPHA Y3	27	29
sp O06219 MURE_MYCTU	Begin: 255 End: 371	UDP-N-ACETYLMURAMOYLALANYL-D-GLUTAMATE--2,6...	27	29
sp P08419 EL2_PIG	Begin: 182 End: 245	ELASTASE 2 PRECURSO	27	38
sp Q11034 Y07S_MYCTU	Begin: 163 End: 218	HYPOTHETICAL 69.5 KD PROTEIN CY02B10.28C	27	38
sp P00577 RPOC_ECOLI	Begin: 1290 End: 1401	DNA-DIRECTED RNA POLYMERASE BETA' CHAIN (T	27	38
sp P32662 GPH_ECOLI	Begin: 20 End: 49	PHOSPHOGLYCOLATE PHOSPHATASE (PGP)	27	38
sp P32662 GPH_ECOLI	Begin: 116 End: 224	PHOSPHOGLYCOLATE PHOSPHATASE (PGP)	27	28
sp P32282 RIR1_BPT4	Begin: 239 End: 266	RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE ALPHA C...	27	50
sp P17346 LEC2_MEGRO	Begin: 36 End: 121	LECTIN BRA-2	27	50
sp P54947 YXEH_BACSU	Begin: 24 End: 51	HYPOTHETICAL 30.2 KD PROTEIN IN IDH-DEOR IN...	27	50
sp P77366 PGMB_ECOLI	Begin: 95 End: 190	PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BETA-PGM)	27	50
sp P30139 THIG_ECOLI	Begin: 43 End: 79	THIG PROTEIN	27	50
sp P95649 CBBY_RHOSH	Begin: 96 End: 189	CBBY PROTEIN	27	50
sp Q43154 GSHC_SPIOI	Begin: 228 End: 327	GLUTATHIONE REDUCTASE, CHLOROPLAST PRECURSO...	26	66
sp P34132 NT6A_HUMAN	Begin: 191 End: 215	NEUROTROPHIN-6 ALPHA (NT-6 ALPHA)	26	66
sp P34134 NT6G_HUMAN	Begin: 115 End: 144	NEUROTROPHIN-6 GAMMA (NT-6 GAMMA)	26	66
sp P95650 GPH_RHOSH	Begin: 48 End: 114	PHOSPHOGLYCOLATE PHOSPHATASE (PGP)	26	66

## Database searching

- The first and most common operation in protein informatics...and the only way to access the information in large databases
- Primary tool for inference of homologous structure and function
- Improved algorithms to handle large databases quickly
- Provides an estimate of statistical significance
- Generates alignments
- Definitions of similarity can be tuned using different scoring matrices and algorithm-specific parameters

## BLAST and FASTA

- The rigorous Needleman-Wunsch and Smith-Waterman algorithms are too slow for large database searches
- There are two major heuristic algorithms (BLAST and FASTA) to speed up the searching
- However, these compromise speed and sensitivity and neither of them guarantees to find the best alignment
- BUT, these are the primary search engines used by the majority of scientists today and their excellent performance justifies such use
- NOTE: Pairwise comparisons limit information content

## FASTA suite

- "Fast" search algorithm generates global alignments, allows gaps
- Good documentation (Pearson)  
<http://www2.ebi.ac.uk/fasta3/help.html>; <http://fasta.bioch.virginia.edu/>
- Extensively updated since first release
  - more rigorous statistical analysis has been added
  - multiple variants available
  - FASTA3 is the current implementation

## BLAST suite

- Original "fast" search algorithm generates local alignments without gaps (Blast 1.4)
- Newer versions (Blast 2.0x) accommodates gaps
- Documentation
  - **Manual:** [http://www.ncbi.nlm.nih.gov/BLAST/blast\\_help.html](http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html)
  - **FACS:** [http://www.ncbi.nlm.nih.gov/BLAST/blast\\_FAQs.html](http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.html)
  - **Tutorial:** <http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>
- Other subtypes recently available for aligning 2 sequences, motif searching, domain matching, short sequences

## BLAST flavors

- **blastp** compares an amino acid query sequence against a protein sequence database
- **blastn** compares a nucleotide query sequence against a nucleotide sequence database
- **blastx** compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database
- **tblastn** compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands)
- **tblastx** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

## Psi-Blast: Extending our reach...

- Generalizes BLAST algorithm to use a position-specific score matrix in place of a query sequence and associated substitution matrix for searching the databases
- Position-specific score matrix is generated from the output of an initial Gapped Blast search, *i.e.*, uses a profile or motif defined in the initial Blast search in place of a single query sequence and matrix for subsequent searches of the database
- Results in a database search tuned to the specific sequence characteristics representative of the sequence set of interest

## Steps in a Psi-Blast search\*

- Constructs a multiple alignment from a Gapped Blast search and generates a profile from any significant local alignments found
- The profile is compared to the protein database and PSI-BLAST estimates the statistical significance of the local alignments found, using "significant" hits to extend the profile for the next round
- PSI-BLAST iterates step 2 an arbitrary number of times or until convergence

\*Adapted from the PSI-BLAST tutorial at NCBI

## PSI-BLAST information at NCBI

- **Access**

<http://www.ncbi.nlm.nih.gov/BLAST/>

- **Tutorial**

<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-2.html>

- **A short explanation of PSI-BLAST statistics**

<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html>

- **See also:**

Park J; Karplus K; Barrett C; Hughey R; Haussler D; Hubbard T; Chothia C. "Sequence comparisons using multiple sequences detect three times as many remote homologs as pairwise methods." (1998) *J. Mol. Biol.*, 284:1201-10

## Part 2

Beyond database searching: How do we turn our results into knowledge?

- Some Basic Principles of Molecular Evolution
- Evaluation using Multiple Alignments
- Finding and Analyzing Motifs
- New Directions in Bioinformatics

## Molecular Evolution

Highly relevant but we only have time to mention some very basic issues

## References

Saier, M.H. Jr. "Phylogenetic approaches to the identification and characterization of protein families and superfamilies"

Labedan, B. & Riley, M. "Gene products of E.coli: Sequence comparisons and common ancestries"

Green, P. et al. "Ancient conserved regions in new gene sequences and the protein databases"

Murzin, A.G. "How far divergent evolution goes in proteins"

Textbooks:

Fundamentals of Molecular Evolution, Li & Graur, Sinauer Associates, 2nd Ed. (1999)

Molecular Systematics, D.M. Hillis & C. Moritz, Eds., Sinauer Associates (1990)

## Web Resources

- **Useful Lists**

<http://www.mcb.harvard.edu/BioLinks/Evolution.html>  
<http://darwin.eeb.uconn.edu/molecular-evolution.html>

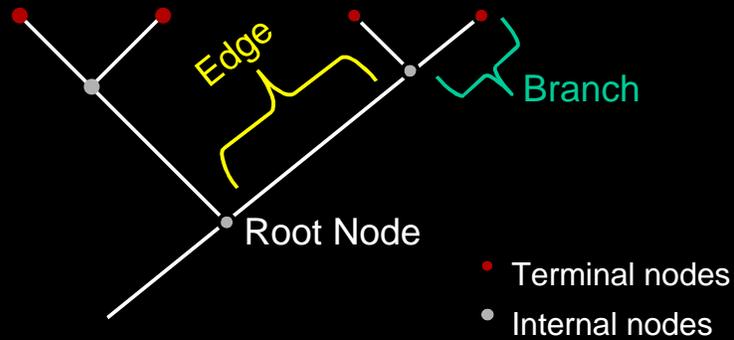
- **Tree of Life site**

<http://phylogeny.arizona.edu/tree/phylogeny.html>

- **A protocol to get you started**

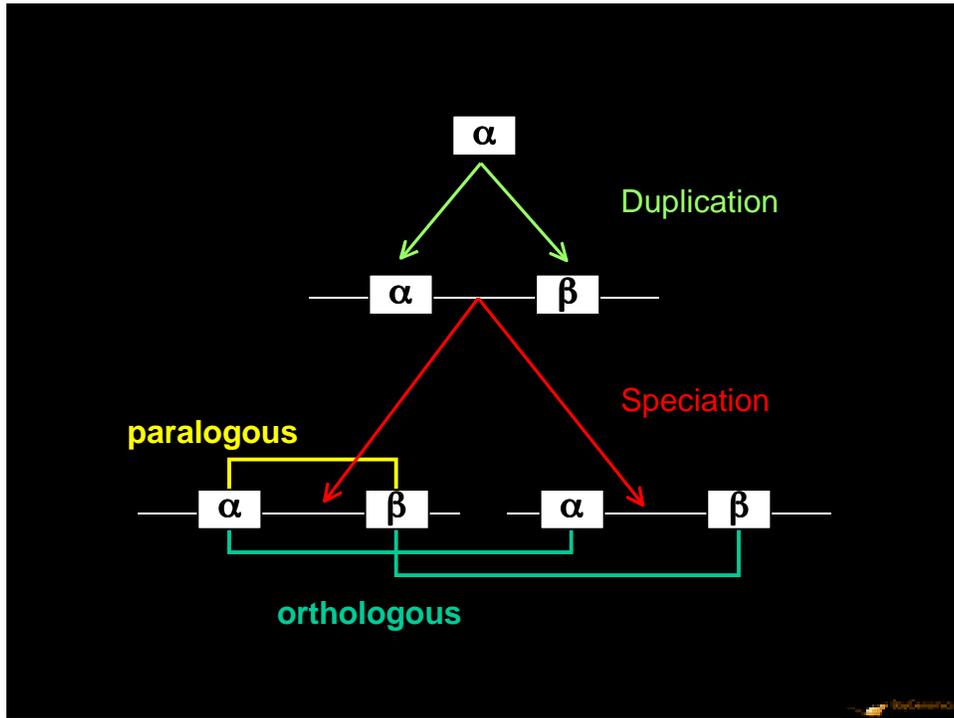
<http://www.infobiogen.fr/docs/MAcours/phylogeny.html>

## Tree (Network) Nomenclature



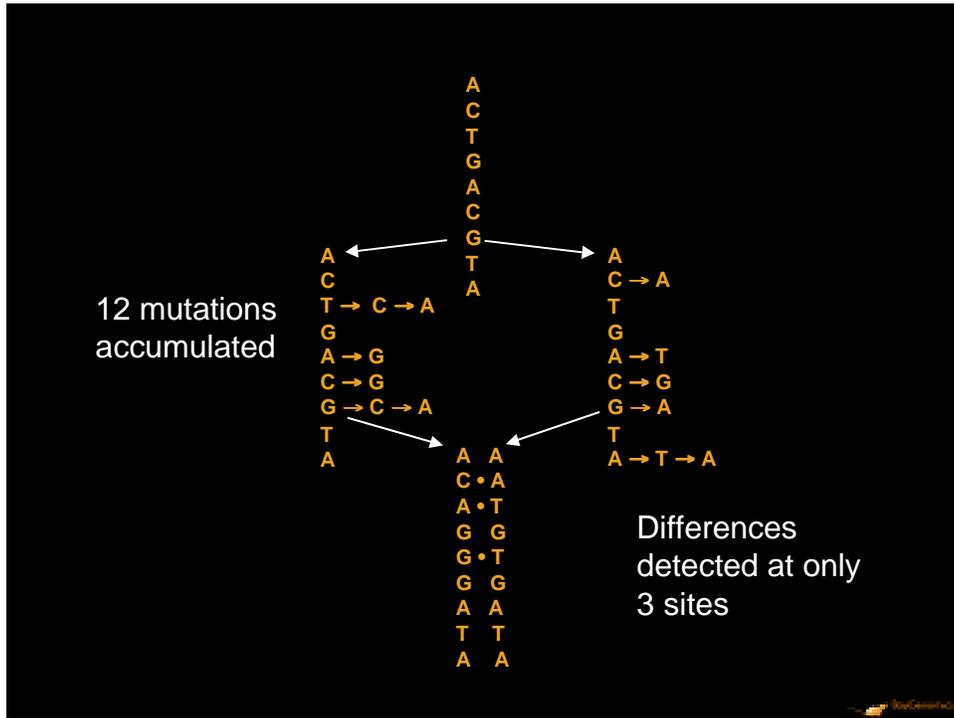
## Definitions

- Homology: Sharing a common ancestor, may have similar or dissimilar functions
- Analogy: Performing a common function but no common ancestry
- Convergence: Performing the same function, having similar structural characteristics, but do not share a common ancestor
- Paralogy: Sequence similarity between the descendants of a duplicated ancestral gene
- Orthology: Sequence similarity as a consequence of a speciation event

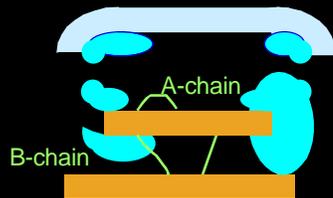
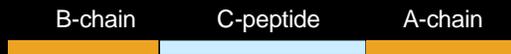


## Important principles

- Evolutionary history is accessed only through contemporary species and molecules
- The basic models for substitution are generally robust for sequences 80% identical (nucleotide level), e.g., not highly diverged
- General assumptions of the models
  - Changes in different copies of genes are independent
  - Changes at each site are independent
  - All sites change at the same rate
  - All bases occur at equal frequencies (corrected in later models to come a little closer to reality)



- Different domains within a single protein evolve at different rates



$$r = 0.97 \times 10^{-9}/\text{site}/\text{year}$$

$$r = 0.13 \times 10^{-9}/\text{site}/\text{year}$$

Mature insulin

## Evaluation using Multiple Alignments

## References on multiple alignment tools

McClure, "Comparative analysis of multiple protein sequence analysis methods"

Thompson et al., "ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice"

(MSA) Lipman et al., "A tool for multiple sequence alignment"

Notredame & Higgins, "SAGA: Sequence alignment by genetic algorithm"

(PIMA) Smith & Smith "Automatic generation of primary sequence patterns from sets of related protein sequences"

See also:

(MACAW) Schuler, G.D., Altschul, S.F., Lipman, D.J. (1991) "A workbench for multiple alignment construction and analysis," *Proteins* 9, 180-90

(PILEUP) Feng, D.F. & Doolittle, R.F. "Progressive sequence alignment as a prerequisite to correct phylogenetic trees" (1987) *J. Mol. Evol.* 25, 351-60

## Evaluation of sequence relationships using multiple alignments

- Screening for membership in a family/superfamily
- Identification of conserved elements important to function
- Distinguishing global vs. local patterns of similarity characteristic of the structural scaffold
- Determination of the level and sites of variability across the members of subgroups/families/superfamilies

- Multiple alignments are more informative than pairwise comparisons

BLASTP 1.4.9

Query= TITLE: URF  
(269 letters)

Database: Non-redundant SwissProt sequences  
49,825 sequences; 17,390,645 total letters.

	Smallest Sum Probabil.
sp P24162 ENOYL-COA HYDRATASE HOMOLOG (ORF257)...	6.1e-31
sp P34559 PROBABLE ENOYL-COA HYDRATASE, MITOCH...	5.2e-29
sp P14604 ENOYL-COA HYDRATASE, MITOCHONDRIAL P...	3.1e-28
sp P30084 ENOYL-COA HYDRATASE, MITOCHONDRIAL P...	1.3e-24
sp P23966 NAPHTHOATE SYNTHASE (DIHYDROXYNAPHTH...	2.3e-21

•  
•  
•

BLASTP 1.4.9

Query= TITLE: Urf  
(269 letters)  
Database: Non-redundant SwissProt sequences  
49,825 sequences; 17,390,645 total letters.

	Smallest Sum Probabil.
sp P24162 ENOYL-COA HYDRATASE HOMOLOG (ORF257)...	6.1e-31
sp P34559 PROBABLE ENOYL-COA HYDRATASE, MITOCH...	5.2e-29
sp P14604 ENOYL-COA HYDRATASE, MITOCHONDRIAL P...	3.1e-28
sp P30084 ENOYL-COA HYDRATASE, MITOCHONDRIAL P...	1.3e-24
sp P23966 NAPHTHOATE SYNTHASE (DIHYDROXYNAPHTH...	2.3e-21
	•
	•
	•

>sp|P24162|ECHH\_RHOCA ENOYL-COA HYDRATASE HOMOLOG (ORF257). >pir||s19026  
enoyl-CoA hydratase homolog - Rhodobacter capsulatus >gi|45984  
(X60194) enoyl-CoA hydratase homologue [Rhodobacter capsulatus]  
Length = 257

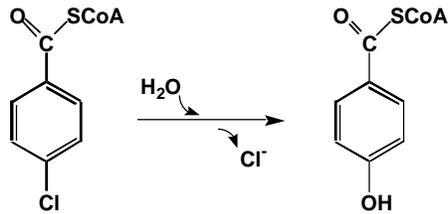
Score = 207 (96.1 bits), Expect = 6.1e-31, Sum P(3) = 6.1e-31  
Identities = 51/137 (37%), Positives = 71/137 (51%)

Query: 89 WHQMIHKIIRVKRPVLAANGVAAGGGLGISLASDMAICADSAKFVCAWHTIGINDTAT 148  
+ ++ I PVLAA+NG AAG G ++LA+D+ I A SA F+ A+ IG+ D  
Sbjct: 83 YEPLLQAIYSCPLPVLAAVNGAAAGAGANLALAADVIAAQSAAFMQAFTRIGLMPDAGG 142

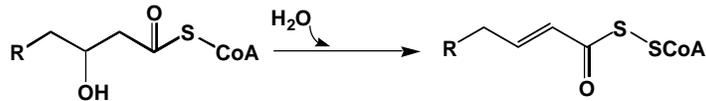
Query: 149 SYSLARIVGMRRAMEMLTNRRTLYPEEAKDWGLVSRVYPKDEFREVAVKQVARELAAAPTH 208  
++ L R VGM RAM + L + EEA GL+ P +F A LA P+  
Sbjct: 143 TWWLPRQVGMARAMGMALFAEKIGAEAAARMGLIWEAVPDVDFEHHWRARAHLARGPSA 202

Query: 209 LNVMAKERFHAGWMPV 225  
K+ FHAG NP+  
Sbjct: 203 AFAAVKKAFFHAGLSNPL 219

### 4-Chlorobenzoate Dehalogenase



### Enoyl-CoA Hydratase



- A multiple alignment distinguishes the dehalogenase from the enoyl Co-A hydratase family

Dehalogenases	G	G	L	G	I	S	L	A	S	D	M	A	I	C	A	D	S	A	K	F	V	C	A	W	H	T	I	G	I	G	N	D	T	A	T	
Enoyl CoA Hydratases	G	G	L	G	M	S	L	A	C	D	L	A	V	C	T	D	R	A	T	F	L	P	A	W	M	S	I	G	I	A	N	D	A	S	S	
	G	G	G	C	E	L	A	M	M	C	D	I	I	Y	A	G	E	K	A	Q	F	G	Q	P	E	I	L	L	G	T	I	P	G	A	G	G
	G	G	G	N	E	L	A	M	M	C	D	I	I	Y	A	G	E	K	A	R	F	G	Q	P	E	I	N	I	G	T	I	P	G	A	G	G
	G	G	G	C	E	L	A	M	M	C	D	F	I	I	A	S	E	T	A	K	F	G	L	P	E	I	T	L	G	V	I	P	G	M	G	G
	G	A	G	C	E	L	A	L	L	C	D	V	V	V	A	G	E	N	A	R	F	G	L	P	E	I	T	L	G	I	M	P	G	A	G	G

## Multiple alignments provide more information than pairwise alignments

- Useful to confirm distant relationships
- Provides a context for interpreting patterns of similarity and difference
- "Speciation" over alignment space helps to connect and confirm widely degenerate motifs

Query= /phosphonataase/phosBc.gcg (302 letters)

Database: swissprot

77,273 sequences; 27,815,109 total letters.

	High Score	Probability P(N)	Smallest Sum N
Sequences producing High-scoring Segment Pairs:			
sp P77247 YNIC_ECOLI HYPOTHETICAL 24.3 KD PROTEIN IN PFKB...	116	2.2e-05	1
sp O67359 GPH_AQUAE PHOSPHOGLYCOLATE PHOSPHATASE (PGP)	106	0.00030	1
sp O06995 PGMB_BACSU PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BE...	97	0.0039	1
sp P31467 YIEH_ECOLI HYPOTHETICAL 24.7 KD PROTEIN IN TNAB...	94	0.0082	1
sp P44755 GPH_HAEIN PHOSPHOGLYCOLATE PHOSPHATASE (PGP)	93	0.011	1
sp P54607 YHCW_BACSU HYPOTHETICAL 24.7 KD PROTEIN IN CSPB...	89	0.030	1
sp P32662 GPH_ECOLI PHOSPHOGLYCOLATE PHOSPHATASE (PGP)	87	0.067	1

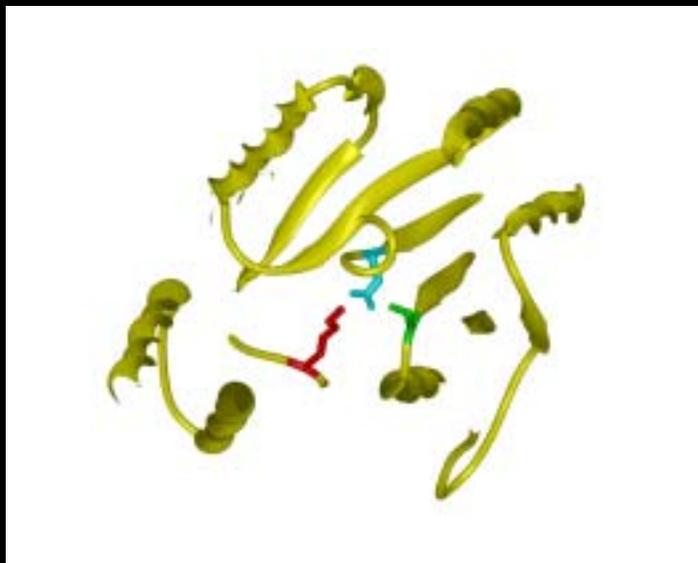
~ 21% identical

PGPhos	- - - - M P G V V F D L D G T L V H S A P D I H A A V N K
Phosphon	M D R M K I E A V I F D W A G T V D Y G C F A P L E V F M
PGPhos	A L A E E G G A P F T L A E I T G F I G - - N G V P V L I Q
Phosphon	E I F H K R G V A I T A E E A R K P M G L L K I D H V R V T
PGPhos	R V L A A R G E A P D A H R Q A E L Q G R F M A H Y E A D P
Phosphon	E M P R I A S E W N R V F R Q L P T E A D I Q E M Y E E F E
PGPhos	A T L T S V Y P - - - - - G A E A A I R H L R A E G W R
Phosphon	E I L F A I L P R Y A S P I N G V K E V I A S L R E R G I K
PGPhos	I G L C T N K P V G A S R Q I L S L F - - - G L L E L F - -
Phosphon	I G S T T - - - - G Y T R E M M D I V A K E A A L Q G Y K P
PGPhos	D A I I G G E S L P Q R K P D P A P L R A T A A A L N - - -
Phosphon	D F L V T P D D V P A G R P Y P W M S Y K N A M E L G V Y P
PGPhos	E E V V L Y V G D S E V D A A T A E A A G L R F A L F T E G
Phosphon	M N H M I K V G D T V S D M K E G R N A G M W T V G V I L G
PGPhos	Y R H A P V - - H E L P H H G L F S H H D E L Q D L L R R L
Phosphon	S S E L G L T E E E V E N M D S V E L R E K I E V V R N R F

	10 *	151 *	176 *
Cu++ATPase.Ec	L D T V V F D K T G T L T E G	V I A G V L P D G K A E A I K H L	A M V G D G I N D A P A L
Cu++ATPase.Hs	V K V V V F D K T G T I T H G	V F A E V L P S H K V A K V K Q L	A M V G D G I N D S P A L
Ca++ATPase.At	A T T I C S D K T G T L T T N	V M A R S S P M D K H T L V R L L	A V T G D G T N D A P A L
Urf.Mj	K V A I V F D S A G T L V K I	E - - - A H Q E L K R D L I R N L	I M V G D G A N D V P A M
PhosSerPhos.Hs	A D A V C F D V D S T V I R E	T A E - S G G K G K V I K L L K E	I M I G D G A T D M E A C
2-DO-6-PPhos.Sc	V D L C L F D L D G T I V S T	I T G F D V K N G K P D P E G Y S	V V F E D A P V G I K A G
DL-Gly-3-Phos.Sc	I N A A L F D V D G T I I I S	I T A N D V K Q G K P H P E P Y L	V V F E D A P A G I A A G
Phosphon.Pa	L Q A A I L D W A G T V V D F	A T D E V - P N G R P W P A Q A L	V K V D D T W P G I L E G
Phosphon.St	I H A V I L D W A G T T V D F	A T D D L A A G G R P G P W M A L	V K V D D A A P G I S E G
Phosphon.Bc	I E A V I F D W A G T T V D Y	T P D D V - P A G R P Y P W M S Y	I K V G D T V S D M K E G
PhosGlycolPhos.Rs	M P G V V F D L D G T L V H S	I G G E S L P O R K P D P A P L A	L Y V G D S E V D A A T A
NtermDom.IGPD.Pp	V Q A L L L D M D G V M A E V	L E D C P P - - - K P S P E P I L	A M V G D T V D D I I A G
B-PhosGlucMut.Ll	F K A V L F D L D G V I T D T	A E V A A S - - - K P A P D I F I	I G L E D S Q A G I Q A I
HaloAcidDehal.PspYL	I K G I A F D L Y G T L F D V	L S V D P V Q V Y K P D N R V Y E	L F V S S N A W D A T G A
NtermDomEpoxyHyd.Hs	L R A A V F D L D G V L A L P	I E S C Q V G M V K P E P Q I Y K	V F L D I G A N L K P A
EnolasePhos.Ko	I R A I V T D I E G T T S D I	F D - - T L V G A K R E A Q S Y R	L F L S D I H Q E L D A A

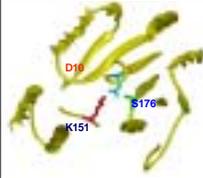
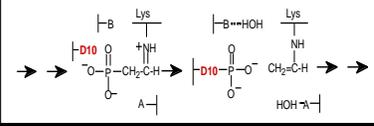
	10 *		151 *		176 *			
Cu++ATPase.Ec	LDTVV	F D K T G T	L T E G	VIAGVLPDG	K A E I K H L	AMVG	D	G I N D A P A L
Cu++ATPase.Hs	VKVVV	F D K T G T	I T H G	VFAEVLPSH	K V A K V K Q L	AMVG	D	G I N D S P A L
Ca++ATPase.At	ATTIC	S D K T G T	L T I N	VMARSSPMD	K H T L V R L L	AVTG	D	G T N D A P A L
Urf.Mj	KVAIV	F D S A G T	L V K I	E--AHQEL	K R D L I R N L	IMVG	D	G A N D V P A M
PhosSerPhos.Hs	ADAVC	F D V D S T	V I R E	TAE-SGGKG	K V I K L L K E	IMIG	D	G A T D M E A C
2-DO-6-PPhos.Sc	VDLCL	F D L D G T	I V S T	ITGFDVKNG	K P D P E G Y S	VVFE	D	A P V G I K A G
DL-Gly-3-Phos.Sc	INAAL	F D V D G T	I I I S	ITANDVKQG	K P H P E P Y L	VVFE	D	A P A G I A A G
Phospon.Pa	LQAAL	L D W A G T	V V D F	ATDEV-PNG	R P W P A Q A L	VKVD	D	T W P G I L E G
Phospon.St	IHAVI	L D W A G T	T V D F	ATDDLAAGG	R P G P W M A L	VKVD	D	A A P G I S E G
Phospon.Bc	IEAVIF	F D W A G T	T V D Y	TPDDV-PAG	R P Y P W M S Y	IKVG	D	T V S D M K E G
PhosGlycolPhos.Rs	MPGVVF	F D L D G T	L V H S	IGGESLPOR	K P D P A P L A	LYVG	D	S E V D A A T A
NtermDom.IGPD.Pp	VQALL	L D M D G V	M A E V	LEDCPP---	K P S P E P I L	AMVG	D	T V D D I I A G
B-PhosGlucMut.Ll	FKAVL	F D L D G V	I T D T	AEVAAS---	K P A P D I F I	IGLE	D	S O A G I Q A I
HaloAcidDehal.PspYl	IKGIAF	F D L Y G T	L F D V	LSVDPVQVY	K P D N R V Y E	LFVS	S	N A W D A T G A
NtermDomEpoxyHyd.Hs	LRAAV	F D L D G V	L A L P	IESCQVGMV	K P E P Q I Y K	VFLD	D	I G A N L K P A
EnolasePhos.Ko	IRAIV	T D I E G T	T S D I	FD--TLVGA	K R E A Q S Y R	LFLS	D	I H Q E L D A A

### Active site of haloacid dehalogenase



	10	151	176
	*	*	*
Cu+ATPase.Ec	LDTVVFDKTGTLTEG	VIAGVLPDGKAEAIKHL	AMVGDGINDAPAL
Cu+ATPase.Hs	VKVVFVFDKTGTLTHG	VFAEVLPSHKVAKVKQL	AMVGDGINDSPAL
Ca+ATPase.At	ATTICSDKTGTLTNN	VMARSSPMDKHTLVRL	AVTGDGTNDAPAL
Urf.Mj	KVAIVFD SAGTLVKI	E--AHQELKRDLIRNL	IMVGDGANDVPAM
PhosSerPhos.Hs	ADAVCFD VDS TVIRE	TAE-SGGKGVIKLLKE	IMIGDGATDMEAC
2-DO-6-PPhos.Sc	VDLCLFDLDGTIVST	ITGFDVKNGKPDPEGYS	VVFE DAPVGIKAG
DL-Gly-3-Phos.Sc	INAALFDVGTIIIS	ITANDVKQKKPHPEPYL	VVFE DAPAGIAAG
Phosphon.Pa	LQAAILDWAGTVVDF	ATDEV-PNGRPWPAQAL	VKVD DTPWGILEG
Phosphon.St	IHAVILDWAGTVDF	ATDDLAAGRRPGPMAL	VKVD DAAPGISEG
Phosphon.Bc	IEAVILDWAGTVVDY	TPDDV-PAGRPYPWMSY	IKVGD TVSDMKEG
PhosGlycolPhos.Rs	MPGVVFDLDGTLVHS	IGGESLPQRKPDPAPLA	LYVGD SEVDAATA
NtermDom.IGPD.Pp	VQALLDMDGVMAEV	LEDCCP--KPSPEPIL	AMVGD TVDDIAG
B-PhosGlucMut.Ll	FKAVLFDLDGVITDT	AEVAAS--KPAPDIFI	IGLE DSQAGIQAI
HaloAcidDehal.PspYL	IKGIADF LYGTLFDV	LSVDPVQVYKPDNRVYE	LFVSS NAWDATGA
NtermDomEpoxyHyl.Hs	LRAAVFDLDGVLALP	IESCQGMVKPEPQIYK	VFLD DIGANLKPA
RhoLasePhos.Ko	IRAIVTDIEGTTSDI	FD--TLVGAKREAQSYR	LFLSD IHQELDAA

	10	151	176
	*	*	*
Cu+ATPase.Ec	LDTVVFDKTGTLTEG	VIAGVLPDGKAEAIKHL	AMVGDGINDAPAL
Cu+ATPase.Hs	VKVVFVFDKTGTLTHG	VFAEVLPSHKVAKVKQL	AMVGDGINDSPAL
Ca+ATPase.At	ATTICSDKTGTLTNN	VMARSSPMDKHTLVRL	AVTGDGTNDAPAL
Urf.Mj	KVAIVFD SAGTLVKI	E--AHQELKRDLIRNL	IMVGDGANDVPAM
PhosSerPhos.Hs	ADAVCFD VDS TVIRE	TAE-SGGKGVIKLLKE	IMIGDGATDMEAC
2-DO-6-PPhos.Sc	VDLCLFDLDGTIVST	ITGFDVKNGKPDPEGYS	VVFE DAPVGIKAG
DL-Gly-3-Phos.Sc	INAALFDVGTIIIS	ITANDVKQKKPHPEPYL	VVFE DAPAGIAAG
Phosphon.Pa	LQAAILDWAGTVVDF	ATDEV-PNGRPWPAQAL	VKVD DTPWGILEG
Phosphon.St	IHAVILDWAGTVDF	ATDDLAAGRRPGPMAL	VKVD DAAPGISEG
Phosphon.Bc	IEAVILDWAGTVVDY	TPDDV-PAGRPYPWMSY	IKVGD TVSDMKEG
PhosGlycolPhos.Rs	MPGVVFDLDGTLVHS	IGGESLPQRKPDPAPLA	LYVGD SEVDAATA
RhoLase.Dom.IGPD.Pp	VQALLDMDGVMAEV	LEDCCP--KPSPEPIL	AMVGD TVDDIAG
B-PhosGlucMut.Ll	FKAVLFDLDGVITDT	AEVAAS--KPAPDIFI	IGLE DSQAGIQAI
HaloAcidDehal.PspYL	IKGIADF LYGTLFDV	LSVDPVQVYKPDNRVYE	LFVSS NAWDATGA
RhoLase.DomEpoxyHyl.Hs	LRAAVFDLDGVLALP	IESCQGMVKPEPQIYK	VFLD DIGANLKPA
RhoLasePhos.Ko	IRAIVTDIEGTTSDI	FD--TLVGAKREAQSYR	LFLSD IHQELDAA

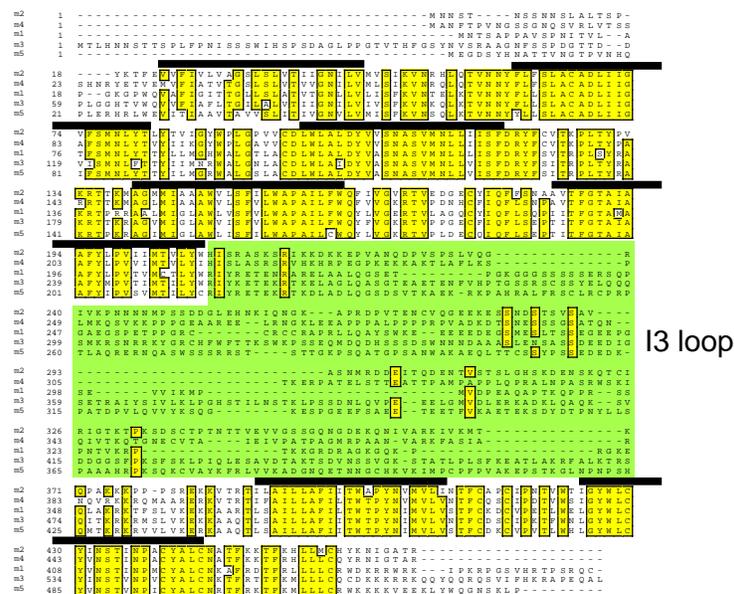
  

  


## Issues in using multiple alignment information

- What question are you asking when you create a multiple alignment?
  - Example: GPCRs

Close relationships: Muscarinic receptors  
 Intermediate relationships: Prostaglandin receptors  
 Distant relationships: Fungal pheromone receptors

### Muscarinic Receptor Sub-types (45-60% identical)



## Prostaglandin Receptors: Family 1 (23-40% identical)

**TMD1**  
 Pros.E2 -----MSPCGLNLSLDAEAAATCATPRLPNTSVVLPTGDNGT**S**ALPFI~~SM~~TLGAVSNVLAALALAA  
 Pros.F2 -----MSMNSKQPVSPAAGLIANTTC---Q~~T~~ENRLLVFFSIIFMTVGLISNLSLALAIEM  
 Throm.A2 -----MWFNGTSLGACFRPNITLQERRIA**S**PWFAASFCALGLGNSLLALSVHA  
 Pros.I2 MMASDGHPPGPPSVTPGSPLSAGGREGWQGMAGSCNITVYQDSV**G**ETSTLMFVAGVVGNGGLALGLG

**TMD2** **TMD3**  
 Pros.E2 QVAGRMR**R**RSAA**T**FLLEVAS**L**LAI**D**LACHV**I**P**G**AL**L**RL**V**T**A**GR-----APAG-GA**C**H**F**L**G**CC**M**V**F**  
 Pros.F2 KAYQRF**R**Q**K**SK**A**S**F**LL**L**AS**G**LV**I**T**D**FF**G**H**L**IN**G**GI**A**V**F**V**A**S**D**K**D**W**I**R**F**D**Q**S**N**-I**L**C**S**I**F**G**I**S**M**V**F**  
 Throm.A2 GARPGAG**R**RS**S**---F**L**AL**L**C**L**LV**L**T**D**FF**L**L**V**T**G**A**I**V**A**S**O**H**A**LL**D**W**R**A**T**D**P**S**C**-R**L**C**Y**F**M**G**V**A**M**V**F**  
 Pros.I2 -----A**R**R**S**HP**S**A**F**AV**L**V**T**G**A**V**T**D**L**L**G**T**C**F**L**S**P**A**I**V**A**Y**A**R**N**S**S**L**L**G**L**A**H**G**G**T**M**L**C**D**T**F**A**F**A**M**T**G

**TMD4**  
 Pros.E2 **E**GL**C**P**L**L**L**G**C**G**M**A**V**E**R**C**V****T****T**Q**L**I**H**A**A**R**V**S**V**A**R**A**R**L**A**L**A**V**L**A**A**M**A**L**A**V**A**L**L**P**P**V**H**V**A**R**Y**E**L**O**V**E**D**T  
 Pros.F2 **S**GL**C**P**L**L**L**G**S**A**M**A**L**E**R****I****T****T**N**P**I**F**H**S**T**K**I**T**S**K**H**V**K**M**I**L**S**G**V**C**M**F**A**V**F**V**A**V**P**L**L**G**H**A**D**Y**Q**I**O**A**S**K**T  
 Throm.A2 **F**GL**C**P**L**L**L**G**A**M**A**S**E**R**E**V**I****T****R**P**S**R**P**T**-**A**T**S**R**R**A**M**A**T**V**G**L**V**M**V**A**G**A**L**G**L**L**P**L**L**G**L**R**H**S**V**I**P**G**S  
 Pros.I2 **E**L**I**A**S**T**L**I**L**F**A**M**A**V**E**R**E**L**A**L**S**H**Y**L**Y**A**Q**L**D**G**P**R**C**L**R**F**A**L**P**S**I**Y**A**F**C**L**F**C**S**L**P**L**L**G**L**E**H**Q**Y**C**C**

**TMD5**  
 Pros.E2 **W**C**F**I**S**L**G**P**R**G**G**W**R**Q**A**L**L**A**G**I**F**A**G**I**G**L**A**A**L**L**A**A**L**V**C**N**T**L**S**G**L**A**L**L**R**A**R**W**R**R**R**R**R**S**R**F**R**K**T**A**G**P**D**D**R**R  
 Pros.F2 **W**C**F**Y**N**T**H**E**I**D**W**E**D**R**F**Y**L**L**F**F**S**F**L**G**L**L**A**L**G**V**S**F**S**C**N**A**V**T**G**V**T**L**L**R**V**K**F**R**S**Q**O**H**R**Q**R**-----  
 Throm.A2 **W**C**F**L**T**L**G**--**T**O**R**G**D**V**V**F**G**L**I**F**A**L**L**G**S**A**S**V**G**L**S**L**L**L**N**T**V**S**V**A**T**L**C**R**V**Y**H**T**R**E**A**T**O**R**P**R**D**C-----  
 Pros.I2 **W**C**F**I**R**M**-**R**S**A**Q**P**G**G**C**A**F**S**L**A**Y**A**S**L**M**A**L**L**V**T**S**I**F**F**C**N**G**S**V**T**L**S**L**Y**H**M**Y**R**Q**R**R**H**H**-----

**TMD6**  
 Pros.E2 **W**S**R**G**P**R**L**A**S**A**S**A**S**S**I**T**S**A**T**A**T**L**R**S**R**G**G**S**A**R**R**V**H**A**H**D**E**M**V**G**O**L**V**G**I**M**V**S**C**I**C**W**S**P**L**L**V**-**L**--  
 Pros.F2 -----S**H**L**E**M**I**I**O**L**L**A**I**M**V**S**C**V**C**W**S**P**L**L**V**T**M**--  
 Throm.A2 -----E**V**E**M**M**V**O**L**V**G**I**M**V**A**T**V**C**M**P**L**L**V**F**I**M**Q**  
 Pros.I2 -----G**S**F**V**P**T**S**R**A**R**E**D**E**V**H**L**I**L**A**L**M**T**V**I**M**A**V**C**S**L**P**L**M**I**R**G**F--

**TMD7**  
 Pros.E2 -----V**L**A**I**A**G**W**N**S**N**S**L**Q**R**P**L**E**L**A**V**R**L**A**S**W**N**O**I**L**D**P**W**V**V**I**L**L**R**Q**A**L**D**R**Q**L**-**L**L**L**-----**L**P**  
 Pros.F2 -----A**N**I**A**I**N**G**N**N**S**P**V**T**C**E**T**T**L**F**L**R**M**A**T**W**N**O**I**L**D**P**W**V**V**I**L**L**R**L**V**L**R**N**L**-**Y**K**L**A**S**R**C**C**G**V**N**I**I**S  
 Throm.A2 T**L**L**Q**T**P**P**V**M**S**F**S**Q**L**L**R**A**T**E**H**Q**L**L**I**V**I**R**M**A**T**W**N**O**I**L**D**P**W**V**V**I**L**F**R**S**V**L**R**L**-**H**-----**P**R**F**S**S**Q**L**Q**A**V**S  
 Pros.I2 -----T**Q**A**I**A**P**D**S**R**E**M**G**L**L**F**S**P**N**A**F**L**L**I**D**P**W**V**V**I**L**F**R**K**A**V**F**Q**R**L**K**F**M**L**C**C**L**C**A**R**S**V**H**G**D**

## Fungal Pheromone Receptors from Several Species (17-25% identical)

**TMD1** **TMD2**  
 1 ---MLD**H**I**T**P**F**F**A**L**V**A**F**F**L**V**L**M**P**F**A**W**H**I**K**S**K**N**V**G**L**I**M**L**S**I**W**I**M**L**G**N**L**D**N**F**V**N**S**M**V**W**K**---**T**T  
 1 M**F**S**G**K**E**N**V**S**F**G**V**L**C**L**L**A**G**C**I**T**S**S**C**L**H**L**Q**A**K**N**I**G**V**L**L**M**P**F**C**F**T**G**L**V**N**K**G**I**N**A**L**A**E**N**---**S**L**  
 1 ---M**S**Y**K**S**A**I**I**G**L**C**L**A**V**I**L**A**P**P**L**A**W**H**S**H**T**K**N**I**P**A**I**I**L**I**T**W**L**L**T**M**N**L**T**C**I**V**D**A**A**I**W**S**D**D**D**F**L**T**-**  
 1 --M**L**P**I**G**I**F**Y**Q**F**Y**A**F**A**L**V**L**S**I**P**I**L**Y**M**Q**L**R**A**R**N**I**P**C**L**L**L**F**W**I**L**T**L**T**L**I**Y**V**V**E**S**A**I**W**S**N**P**Y**A**E**T**I

**TMD3** **TMD4**  
 58 **A**D**L**A**P**A**Y**C**E**L**S**V**R**L**R**H**L**L**F**A**I**P**A**S**N**L**A**I**A**R**K**L**E**S**I**A**S**T**R**Q**V**R**A**G**P**G**D**H**R**R**A**V**I**D**L**L**I**C**D**G**I**P**I**  
 62 **R**L**A**W**T**L**G**C**D**L**S**A**I**E**R**T**W**Q**F**L**C**C**S**A**L**C**V**L**Q**R**L**E**G**I**A**S**L**R**Q**A**H**S**T**V**W**D**R**K**R**R**L**L**I**D**F**G**V**G**L**G**L**P**A**  
 62 **R**W**D**G**K**G**W**C**D**I**V**I**K**L**Q**V**G**A**N**I**G**I**S**C**A**V**T**N**I**Y**N**L**H**T**L**K**-**A**D**S**V**L**P**D**L**S**S**W**T**K**I**V**K**D**L**V**I**S**L**F**T**P**V**  
 64 **R**W**M**G**Y**G**L**C**D**I**T**S**R**I**V**T**C**S**S**I**G**I**P**A**S**A**F**T**L**V**L**Y**L**D**T**V**I**R**-**R**D**H**P**L**K**R**Y**E**N**W**---**I**W**H**V**C**L**S**I**L**L**P**L**

**TMD5** **TMD6**  
 123 **I**Y**T**S**L**M**I**V**N**C**S**N**R**Y**G**I**L**E**E**A**G**C**Q**P**M**M**V**F**S**W**L**W**V**L**V**A**A**P**V**I**V**S**L**C**S**A**V**Y**S**A**L**A**F**R**W**F**W**R**R**R**Q**P  
 127 **L**Q**I**P**M**F**F**I**V**Q**P**Y**R**L**N**V**I**E**N**I**G**C**S**A**P**L**Y**A**S**V**P**A**L**F**I**Y**H**L**W**R**L**L**V**S**L**V**C**A**V**Y**A**V**L**R**W**F**M**L**R**E**R**Q**P**  
 126 **M**V**M**G**S**Y**L**L**Q**V**F**Y**G**I**A**R**Y**N**C**Q**N**L**S**P**T**W**I**T**V**L**Y**T**M**W**M**L**I**W**S**F**V**G**A**Y**A**T**L**V**L**F**V**E**Y**K**K**E**D**V  
 125 **I**I**M**A**M**M**V**P**L**E**S**N**R**Y**V**V**I**C**M**N**G**Y**S**S**F**Y**Q**T**W**Y**T**L**L**F**F**Y**I**P**P**C**L**L**S**P**G**G**L**F**F**V**S**R**I**V**L**V**Y**W**R**C**E**L

**TMD7**  
 252 **Q**Y**P**A**S**L**V**M**E**N**T**F**Q**R**N**L**I**L**A**R**L**V**C**P**L**A**W**I**F**F**A**M**F**L**G**L**E**V**R**Q**G**Y**K**E**A**F**H**R**-**L**L**F**C**L**R**K**E**P**K**A  
 256 **F**V**P**D**T**I**A**H**S**S**L**--**S**L**S**I**L**R**W**F**S**L**T**P**A**M**A**L**V**F**F**L**T**E**E**A**Q**S**V**Y**K**A**R**K**A**L**-**I**N**L**C**---**S**S**G**  
 256 **F**D**P**G**R**P**I**-**Y**N**I**-----**W**L**V**L**M**S**W**L**V**L**I**F**L**G**S**D**A**L**H**M**S**K**F**L**R**S**I**K**L**G**F**V**L**D**M**W**K**R**F**I  
 253 **Y**Y**E**T**T**K**V**G**L**N**D**-----**W**V**P**P**T**V**L**L**M**S**L**F**E**S**T**S**G**G**W**T**E**K**V**A**L**I**L**W**S**L**L**V**L**W**L**P**F**T**K**-----

316 **S**A**L**Q**H**V**V**A**D**I**E**V**V**T**F**R**S**H**D**T**F**D**A**N**T**S**T**K**S**E**K**S**D**I**D**M**R**G**S**E**A**-----  
 314 **K**K**Q**T**D**G**R**E**S**L**D**E**A**F**E**S**H**G-----**S**K**F**S**V**L**V**Q**R**D**T**V**I**C-----  
 310 **D**N**K**K**R**R**V**G**I**L**N**K**L**S**R**K**E**S**R**N**P**F**S**T**D**S**E**N**Y**I**S**T**C**T**E**N**S**P**C**V**G**T**P**I**S**Q**A**H**F**V**D**Y**R**I**P**D**D**P**R**K  
 303 **N**T**A**L**G**R**H**A**Q**F**K**L**D**C**C**K**L**E**S**T**M**A**G**K**T**L**D**A**T**D**F**K**E**K**-**-----**L**V**L**E**R**Q**W**S**K**S**I**P**S**D**N**S**S**

- What is the range of sequence divergence among the sequences you plan to align?

```

eno11 EAMKMGAEVYHHLKSVIKKKYQDATNVGDEGGFAPNIQENKEGL
eno12 EAMKMGCEVYHHLKAVIKKKYQDATNVGDEGGFAPNIQENKEGL
eno13 EALRIGSEVYHNLKSLTKKKYQDSAGNVGDEGGVAPDIKTPKEAL
eno14 EAMKMGVEVYHNLKSLTKKKYQDATNVGDEGGFAPNIQENKEGL
eno15 EALRIGSEVYHNLKSLTKKRYGASAGNVGDEGGVAPNIQTAEAL
eno16 EALKMGSEVYHALKSVIKAKYQDACNVGDEGGFAPNIQDNKEGL
eno17 EAMKMGSEVYHHLKSNVIAKAKFGLDATAVGDEGGFAPNIQSNKEAL
eno18 DAMRVGAEVYHSLKGVIAKAKYKDATNVGDEGGFAPNILDNHEAL
cpeps D-IEVADRVFTAAHRNVERRFGPVPLS-ASSGLMVP--LDSAGQL

eno11 ELLKTAIAKAGYTGKVVIGMDVAASEFYG-SDKTYDLNFKKEENND
eno12 ELLKTAIEKAGYTGKVVIGMDVAASEFYG-KDKSYDLNFKKEESND
eno13 DLIMDAIDKAGYKGKVGIAMDVASSEFY--KDGKYDLDFKKNPESD
eno14 ELLKAAIAEKAGYTGKVVIGMDVAASEFFGKDKTYDLNFKKEENND
eno15 DLIVDAIKAAGHDGKVKIGLDCASSEFF--KDGKYDLDFKKNPNSD
eno16 ELLNEAIAKAGYTGKVKIGMDVAASEFY--KDGKYDLDFKKNPNSD
eno17 NLISDAIAKAGYTGKIEIGMDVAASEFY--KDGQYDLDFKNEKSD
eno18 ELLKAAIAQAGYTDKVVIGMDVAASEFC--RDGRYDLDFKSP-PD
cpeps DLLQAAVAETGHTEVCTLGVDVAA-EHLLTEPGRYRF-----

```

```

eno11 EAMKMGAEVYHHLKSVIKKKYQDATNVGDEGGFAPNIQENKEGL
eno12 EAMKMGCEVYHHLKAVIKKKYQDATNVGDEGGFAPNIQENKEGL
eno13 EALRIGSEVYHNLKSLTKKKYQDSAGNVGDEGGVAPDIKTPKEAL
eno14 EAMKMGVEVYHNLKSLTKKKYQDATNVGDEGGFAPNIQENKEGL
eno15 EALRIGSEVYHNLKSLTKKRYGASAGNVGDEGGVAPNIQTAEAL
eno16 EALKMGSEVYHALKSVIKAKYQDACNVGDEGGFAPNIQDNKEGL
eno17 EAMKMGSEVYHHLKSNVIAKAKFGLDATAVGDEGGFAPNIQSNKEAL
eno18 DAMRVGAEVYHSLKGVIAKAKYKDATNVGDEGGFAPNILDNHEAL
cpeps D-IEVADRVFTAAHRNVERRFGPVPLS-ASSGLMVP--LDSAGQL

eno11 ELLKTAIAKAGYTGKVVIGMDVAASEFYG-SDKTYDLNFKKEENND
eno12 ELLKTAIEKAGYTGKVVIGMDVAASEFYG-KDKSYDLNFKKEESND
eno13 DLIMDAIDKAGYKGKVGIAMDVASSEFY--KDGKYDLDFKKNPESD
eno14 ELLKAAIAEKAGYTGKVVIGMDVAASEFFGKDKTYDLNFKKEENND
eno15 DLIVDAIKAAGHDGKVKIGLDCASSEFF--KDGKYDLDFKKNPNSD
eno16 ELLNEAIAKAGYTGKVKIGMDVAASEFY--KDGKYDLDFKKNPNSD
eno17 NLISDAIAKAGYTGKIEIGMDVAASEFY--KDGQYDLDFKNEKSD
eno18 ELLKAAIAQAGYTDKVVIGMDVAASEFC--RDGRYDLDFKSP-PD
cpeps DLLQAAVAETGHTEVCTLGVDVAA-EHLLTEPGRYRF-----

```

Enols 1-8: all >60% identical to each other  
Cpeps: <35% identical to Enols 1-8

gram.pos L K A K - - G M N T A V **G D E G G Y A P** N L G S N D E **A L** A V I A  
 gram.neg L S A K - - G M N T N V **G D E G G F A P S L D S** A S S **A L** D F I V  
 eukaryote T K K R Y G A S A G N V **G D E G G V A P N I Q T A E E** **A L** D L I V  
 archae. L A D R - - D L P A G K **G D E G A W A P S V - S** D D E **A F E I M D**  
 cpeps **V E R R F G P V P - - L S A S S G L M V P L D S A G Q - L D L L Q**

gram.pos E **A** V K A A **G** Y E L G K D I T L A M **D C A A S E** F Y K D - G K - -  
 gram.neg D S I S K A **G** Y K P G E D V F I A L **D A A S S E** F Y N K - D Q N I  
 eukaryote D **A** I K A A **G** H - - D G K V K I G L D C A S S E F F K D - G K Y D  
 archae. E **A** V E T V A D D F G F A I S F G L D V A R A E L Y D D - E A D G  
 cpeps **A A V A E T G H - - T E V C T L G V D V A A E H L L T E P G R Y R**

gram.pos Y V L A - - - G E **G N K A F T** S E E F T H F L E E **L T K Q Y** P I V  
 gram.neg Y D L K - - - G E **G R K - L T** S A Q L V D Y Y V E **L C G K Y** P I Y  
 eukaryote L D F K N P N S D K S **K W L T** G P Q L A D L Y H S **L M K R Y** P I V  
 archae. Y V Y - - - - D D **G V K - -** S T E E Q I E Y I A G K V E E Y D L V  
 cpeps **F - - - - - - - - - G D R V L T A P D F A D H L A D L A H R F R M S**

Enls: 37-62% identical to each other  
 Cpeps: 35-50% identical to Enls

## How do you handle internal repeats?

hyp. protein (S55115)

glyoxalase

```

q09751.N  V E R S K R E G I L E L T Y N F G T E K K E G P V Y I N
s55115.N  P D V F S A H G V L E L T H N W G T E K N P D Y K I N N
q09751.C  - - - - - E G L L E L T H N W G T E K E S G P V Y H N
s55115.C  - - V F S C E S V L E L T H N W G T E N D P N F H Y H N
glyox.    - - - - - E A V I E L T Y N W G V D K - - - - - Y E L

```

```

q09751.N  G N T E P K R G F G H I C F T V D N I E S A C A Y L E -
s55115.N  G N E E P H R G F G H I C F S V S D I N K T C E E L E -
q09751.C  G N D G D E K G Y G H V C I S V D N I N A A C S K F E -
s55115.C  G N - S E P Q G Y G H I C I S C D D A G A L C K E I E V
glyox.    G T - - - - - A Y G H I A L S V D N A A E A C E K I R Q

```

```

q09751.N  - - S K G V S F K K K L S D G K M K H I A F - - - - -
s55115.N  - - S Q G V K F K K R L S E G R Q K D I A F - - - - -
q09751.C  - - A E G L P F K K K L T D G R M K D I A - - - F L L D
s55115.C  K Y G D K I Q W S P K F N Q G R M K N I A - - - F L K D
glyox.    N G G N V T R E A G P V K G G - - - - T T V I A F V E D

```

## General Issues in Multiple Alignment

- Computational complexity: a true multiple alignment of N sequences would require an N-dimensional matrix
- No single "correct" multiple alignment can be achieved except in trivial cases
- Methods assume sequences are independent rather than related by a phylogenetic tree in which the "branches" may evolve at different rates and with different positions being important to function

## Some Primary Algorithms for Multiple Alignment

- Global alignment methods construct an alignment throughout the length of the entire sequence
  - Examples: Pileup, Clustal family, MSA
- Local alignment methods identify ordered series of motifs, then aligns the intervening regions
  - Examples: MACAW, PIMA
- 1D profile analysis

## PILEUP (in GCG package\*)

- 1) Calculates a diagonal matrix of  $N(n-1)/2$  distances between all sequence pairs of  $N$  sequences using Needleman-Wunsch algorithm
- 2) Constructs a guide tree (dendogram) from the distance matrix to direct the order of addition of subsequent pairwise alignments
- 3) Progressively aligns each cluster to the next most related sequence or cluster of sequences, adjusting the position of indels in all sequences

\*Genetics Computer Group, Madison, WI (available through UCSF SACS)

## Issues in the use of PILEUP

- Fast, generates reasonable alignments
- Current implementation in GCG handles up to 500 sequences
- All alignments determined from pairwise alignments, losing the information contained in the multiple alignment for position-specific scoring
- Overrepresentation of a subset of sequences to be aligned may bias the inference of an ordered series of motifs

## ClustalW\*

- From a family of programs using profile-based progressive alignment
- Access: <http://www2.ebi.ac.uk/clustalw/>
- Permits user adjustment of many parameters for both the pairwise and multiple alignment stages
- Computes position-specific gap opening and extension penalties as the alignment proceeds, *e.g.*, varies parameters at different positions

\*"W" stands for "weighting" the sequences to correct for unequal sampling of sequences from different evolutionary distances

## Steps in a ClustalW alignment

- 1) Constructs a distance matrix of all  $N(N-2)/2$  pairs using dynamic programming and converts scores to distances
- 2) Generates a "guide tree" using the neighbor-joining clustering algorithm of Saitou & Nei
- 3) Progressively aligns sequences in order of decreasing similarity using variable parameters and position-specific gap penalties

## The Bottom Line... \*

- For multiple alignments of divergent proteins, e.g., <30% identity, none of these methods is very satisfactory, suffering from 3 types of problems:
  - Inability to produce a single multiple alignment from correctly aligned subsets of the input sequences
  - Sensitivity to the number of sequences used
  - Sensitivity to the specific sequences used for multiple alignment

\*from the McClure paper listed in the lecture references

## 1-D Profile analysis

- Access: GCG package at SACS and at [http://www.sdsc.edu/projects/profile/new/help\\_main.html](http://www.sdsc.edu/projects/profile/new/help_main.html)  
(Gribnikov, M., McLachlan, E.D., Eisenberg, D. (1987) *PNAS USA*, 84:4355-4358)
- Information in a multiple alignment is represented quantitatively as a table of position-specific symbol comparison values and gap penalties
- All information in the alignment is used
- Implementations available for both for database searching/sequence alignment

## Hidden Markov Models

- Probability-based models for database searching, multiple alignments, family generation (Pfam)
- Software and tools sites:  
<http://hmmer.wustl.edu/>  
<http://www.cse.ucsc.edu/research/compbio/HMM-apps/HMM-applications.html>

## Precomputed Multiple Alignments of Protein Families

- **Pfam:** <http://pfam.wustl.edu/>
  - Multiple sequence alignments and HMMs for many protein domains (3071 models as of 8/01)
- **Prodom:** <http://protein.toulouse.inra.fr/prodom.html>
  - Families generated automatically using PSI-BLAST with a profile built from the seed alignments of Pfam
- **Systems:** <http://www.dkfz-heidelberg.de/tbi/services/documentation/systemshelp.html>
  - Families clustered from SW-Prot/PIR using sequence walks and aligned via ClustalW
- **MetaFam:** <http://metafam.ahc.umn.edu/>
  - Functional assignments and a tool for comparison of how other family databases have made the classification

## Finding and Analyzing Motifs

## Applications for Motif Analysis

- Identification of very distant homologs
- May point to important functional units in a protein
- Can be used to "anchor" a multiple alignment
- Databases of motifs can be used to develop other informatics applications

Example: BLOCKS → Blosum

See: Bork, P. & Gibson, T. J. "Applying Motif and Profile Searches," in *Methods in Enzymology* 266: Computer methods for macromolecular sequence analysis, pp. 162-184

## Prosite: Protein Family Signatures

<http://tw.expasy.org/prosite/>

- Contains signatures for ~1500 families/domains
- Can be accessed using description, accession number, author, citation, full text search
- Provides several useful tools allowing a user to
  - Scan a sequence against a PROSITE pattern
  - Scan a pattern generated by a user or from PROSITE against the Swiss-Prot database
  - Scan a sequence against Profile databases, e.g., generalized profiles derived from multiple alignments
  - Many other specialized tools for motif/pattern generation and analysis
  - Includes substantial meta data: experts on each system, references, some statistical analysis

## Meme & Mast

<http://meme.sdsc.edu/meme/website/>

- **Meme: motif discovery tool**  
(Grundy, W. M. et al. 1997. CABIOS 13, 397)
  - motifs represented as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern
  - output can be converted to BLOCKS which can then be converted to PSSMs (position-specific scoring matrices)

- **Mast: database searching tool using one or more motifs as queries**
  - provides a match score for each sequence in the database compared with each of the motifs in the group of motifs provided represented as P-values
  - provides probable order and spacing of occurrences of the motifs in the sequence hits

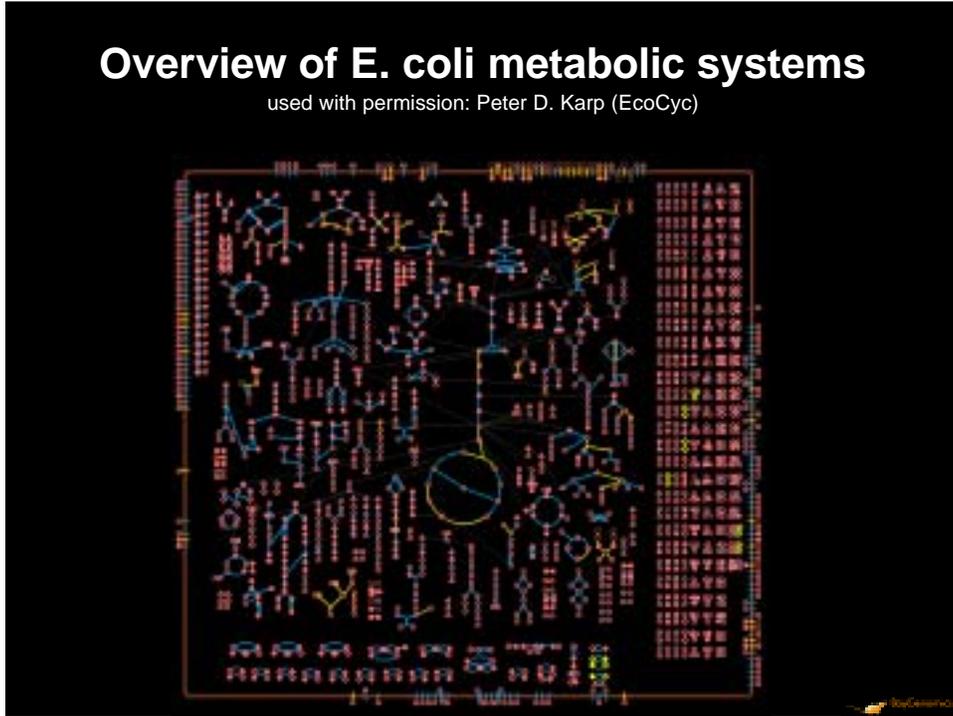
## New Directions in Protein Bioinformatics

### Using Protein Informatics for Really New Insight into Biology

- **Comparative genomics**
  - **Metabolic computing: EcoCyc & MetaCyc**  
<http://ecocyc.org/ecocyc/index.html>
  - **Clusters of Orthologous Groups (COGS)**  
<http://www.ncbi.nlm.nih.gov/COG/>
- **Genetic circuits/Systems analysis**  
<http://gobi.lbl.gov/~aparkin/index.html>
- **Protein-Protein Interactions**
  - **Co-evolution**

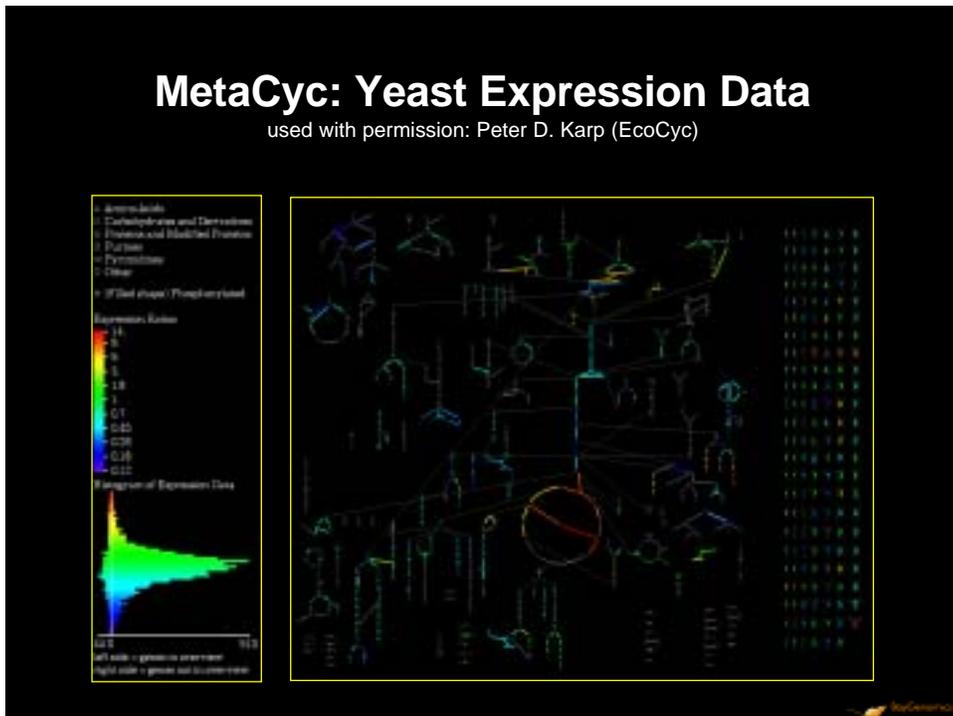
# Overview of E. coli metabolic systems

used with permission: Peter D. Karp (EcoCyc)



# MetaCyc: Yeast Expression Data

used with permission: Peter D. Karp (EcoCyc)



## Nature 1999 Nov 4;402(6757):83

“A combined algorithm for genome-wide prediction of protein function”

EDWARD M. MARCOTTE, MATTEO PELLEGRINI, MICHAEL J. THOMPSON,  
TODD O. YEATES & DAVID EISENBERG

The availability of over 20 fully sequenced genomes has driven the development of new methods to find protein function and interactions. Here we group proteins by correlated evolution, correlated messenger RNA expression patterns and patterns of domain fusion to determine functional relationships among the 6,217 proteins of the yeast *Saccharomyces cerevisiae*. Using these methods, we discover over 93,000 pairwise links between functionally related yeast proteins. Links between characterized and uncharacterized proteins allow a general function to be assigned to more than half of the 2,557 previously uncharacterized yeast proteins. Examples of functional links are given for a protein family of previously unknown function, a protein whose human homologues are implicated in colon cancer and the yeast prion Sup35.

## A few important topics we didn't even mention

- Mapping Sequence → Structure → Function
- Structural superposition and 3D motif finding
- The 3D genome project
- Mapping the protein universe
- Census studies (Gerstein)
- Informatics for Proteomics
  - post-translational modifications
  - investigating protein machines

## See also:

- **Nucleic Acids Res. 2002 30**
  - Description and useful information on 112 databases of interest to the genomics/proteomics/bioinformatics communities