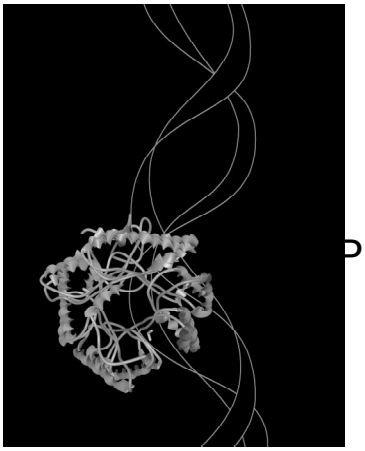


A Very Very Very Short Introduction to Protein Bioinformatics

Patricia Babbitt
University of California, San Francisco
April 24, 2003

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

- 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://www3.ncbi.nlm.nih.gov/Entrez/

- San Diego Supercomputer Center, USA

  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/http://www.ornl.gov/TechResours/
  - 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/

BayGenomics

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

BayGenomics

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

BayGenomics

Introduction to Protein Sequence Analysis

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

BayGenomics

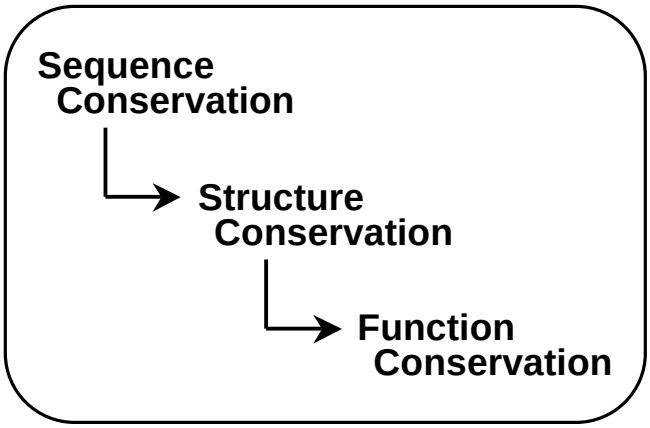
Applications

- 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...



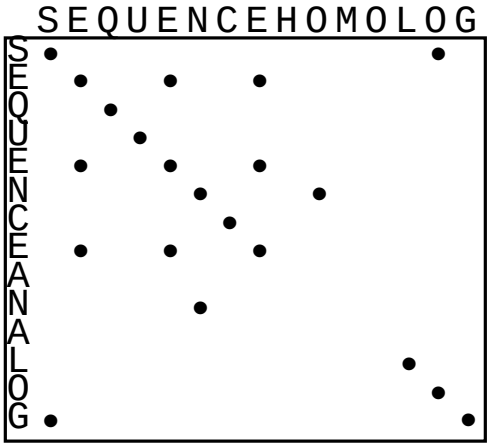
...requires comparison of sequences

- 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

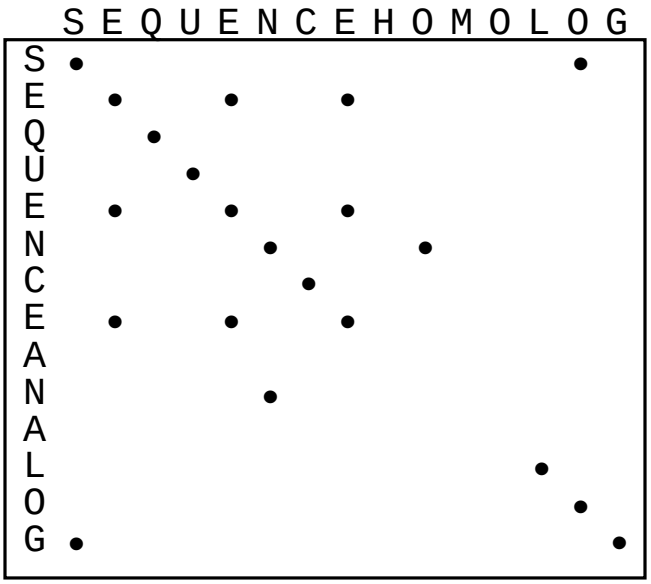
- Dot matrix plots: a simple description of alignment operations illustrating types of relationships between a sequence pair



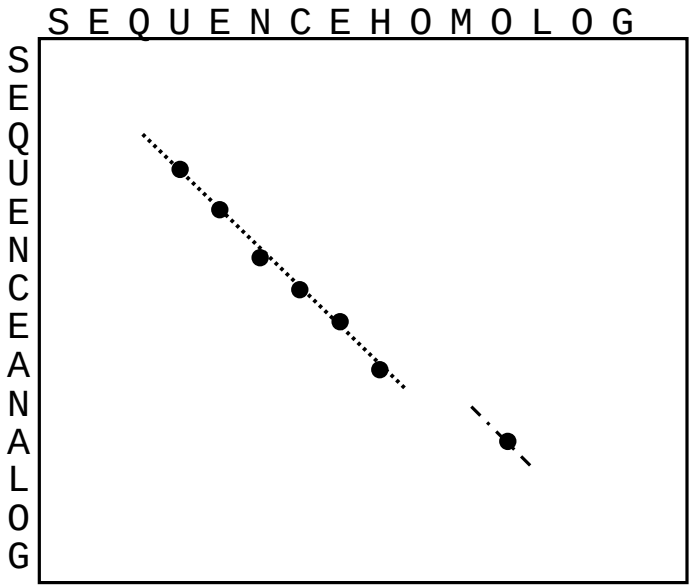
- The signal-to-noise ratio can be improved using filtering techniques designed to minimize the composition-dependent background
- Example of common filters: over-lapping, fixed-length "windows" for sequence comparison
- To be counted, a comparison must achieve a minimum threshold score summed over the 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)

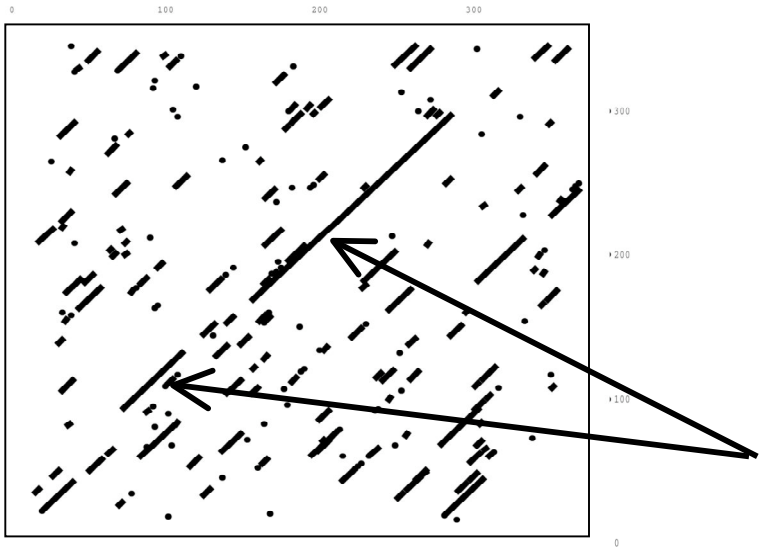
```
SEQUENC
SEQUENCEANALOG          (7/7 matches)

 SEQUENC
SEQUENCEANALOG          (0/7 matches)

   ...

 EQUENCE
SEQUENCEANALOG          (7/7 matches)

   ...

       HOMOLOG
SEQUENCEANALOG          (3/7 matches)
```



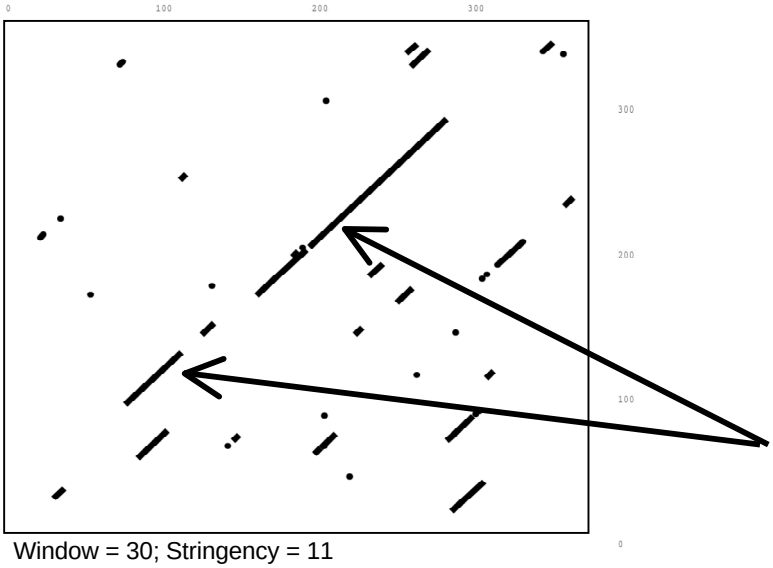
Stringency = 1/7



- 7/7
- 6/7
- 5/7
- 4/7
- 3/7



Window = 30; Stringency = 2



Window = 30; Stringency = 11

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 @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 Ai, Aj, count the # of times Ai aligns with Aj and divide that number by the total # of amino acid pairs in all of the alignments, resulting in the frequency, f(i,j)
- Let fi and fj, respectively, denote the frequencies at which Ai and Aj 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

BayGenomics

Guidelines for using PAM matrices

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

BayGenomics

A	K	W
2	-1	-6
0	1	-5
-2	-5	-8
0	0	-7
0	0	-7
-4	-5	0
1	-2	-7
-1	0	-3
-1	-2	-5
-1	5	-3
-2	-3	-2
-1	0	-4
0	1	-4
1	-1	-6
0	1	-5
-2	3	2
1	0	-2
1	0	-5
0	-2	-6
-6	-3	17
-3	-4	0
0	0	-6

BayGenomics

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

BayGenomics

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

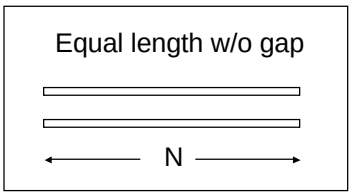
- Many sequences from aligned families are used to generate the matrices
- Sequences identical at >X% are eliminated to avoid bias from proteins over-represented in the database
- Specific matrices refer to these clustering cut-offs, *i.e.,* BLOSUM62 reflects observed substitutions between segments <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

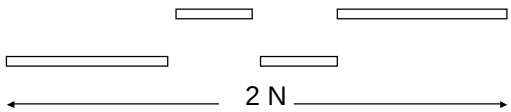
- Considering two sequences, both of length N:
  - If gaps or local alignments are not considered, there is only one optimal solution



  - 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, ~10^{88} comparisons are required

Extreme case: seq 1 aligns with gaps of seq 2



Optimizing the Score

- Dynamic Programming

  (Needleman, S.B. and Wunsch, C.D. J. Mol. Biol. 48 (1970) 443-453)

  – Requires computational time proportional to N^2

  – Addresses the problem for GLOBAL alignments; still has to deal with gaps

- Next step forward: local alignments

  (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

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)

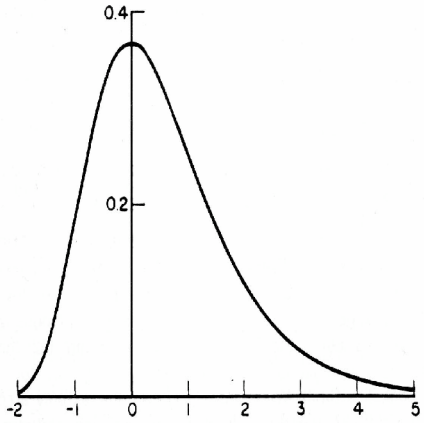


Figure 5. The probability density function of the extreme value distribution (74), with characteristic value $u = 0$ and decay constant $\lambda = 1$.

146

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

BayGenomics

```
Query= /phosphonatase/phosSt.gcg (255 letters) (10/20/99/pcb)
Database: /mol/seq/blast/db/swissprot 78,725 sequences; 28,368,147 total letters
!
                                                                          Score     E
Sequences producing significant alignments:                               (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|YFBT_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|P52183|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_SPIOL 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
```

BayGenomics

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
  - FACS: 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

BayGenomics

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)

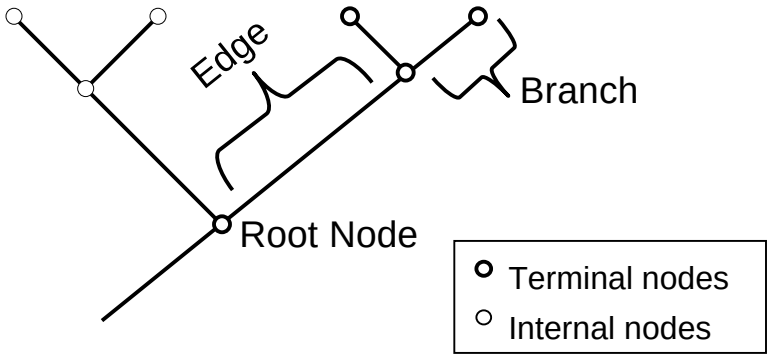
Molecular Systematics, D.M. Hillis & C. Moritz, Eds., Sinauer Associates (1990)

BayGenomics

Web Resources

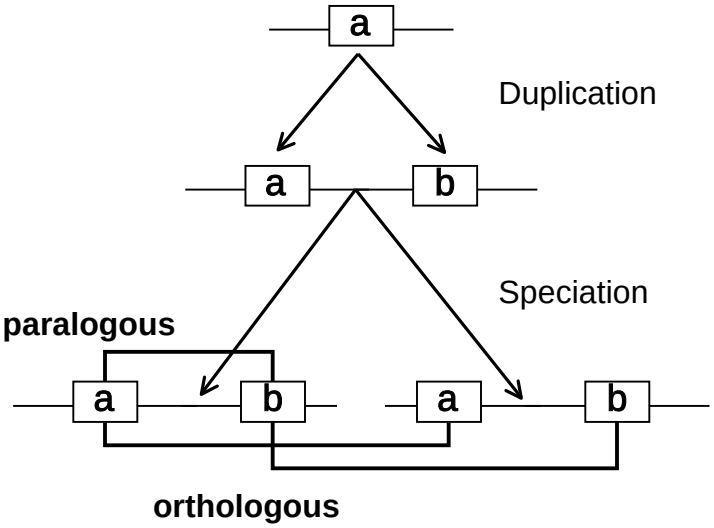
- Useful Lists
  http://www.mcb.harvard.edu/BioLinks/Evolution.html
  http://darwin.eeb.uconn.edu/molecular-evolution.html
- Tree of Life site
  http://tolweb.org/tree/phylogeny.html
- A protocol to get you started
  http://www.infobiogen.fr/docs/MAcours/phylogeny.htmlg

Tree (Network) Nomenclature



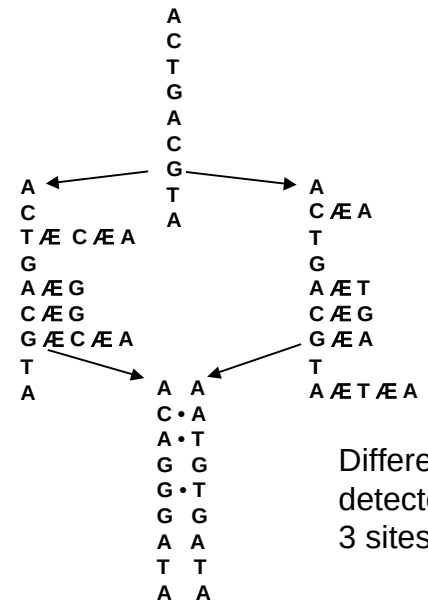
Definitions

- 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

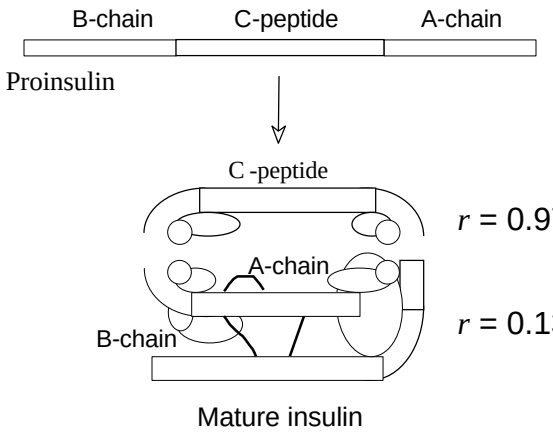


Important principles

- 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 x 10^{-9}/site/year

r = 0.13 x 10^{-9}/site/year

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

BayGenomics

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

BayGenomics

- 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

                                   •
                                   •
                                   •
```

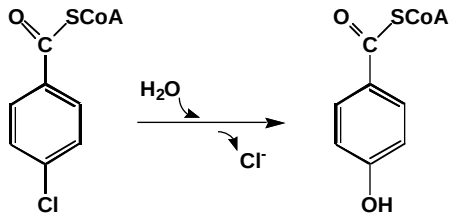
BayGenomics

```
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

                                   •
                                   •
                                   •
```

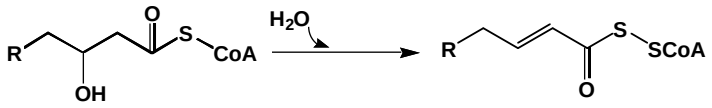
BayGenomics

```
>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 WHQMIHKIIRVKRPVLAAINGVAAGGGLGISLASDMAICADSAKFVCAWHTIGIGNDTAT 148
             +  ++  I     PVLAA+NG AAG G  ++LA+D+ I A SA F+ A+  IG+  D
Sbjct:    83 YEPLLQAIYSCPLPVLAAVNGAAAGAGANLALAADVVIAAQSAAFMQAFTRIGLMPDAGG 142

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

Query:   209 LNVMAKERFHAGWMNPV 225
                  K+ FHAG  NP+
Sbjct:   203 AFAAVKKAFHAGLSNPL 219
```

4-Chlorobenzoate Dehalogenase



Enoyl-CoA Hydratase



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

```
                 GGGLGISLASDMAICADSAKFVCAWHTIGIGNDTAT   Dehalogenases
                 GGGLGMSLACDLAVCTDRATFLPAWMSIGIANDASS
                 GGGCELAMMCDIIYAGEKAQFGQPEILLGTIPGAGG   Enoyl CoA
                 GGGNELAMMCDIIYAGEKARFGQPEINIGTIPGAGG   Hydratases
                 GGGCELAMMCDFIIASETAKFGLPEITLGVIPGMGG
                 GAGCELALLCDVVVAGENARFGLPEITLGIMPGAGG
```

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= /phosphonatase/phosBc.gcg        (302 letters)

Database:  swissprot
           77,273 sequences; 27,815,109 total letters.
                                                                    Smallest
                                                                      Sum
                                                              High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

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
```

BayGenomics

~ 21% identical

```
PGPhos    -----MPGVVFDLDGTLVHSAPDIHAAVNK
Phosphon  MDRMKIEAVIFDWAGTTVDYGCFAPLEVFM

PGPhos    ALAEEGGAPFTLAEITGFIG--NGVPVLIQ
Phosphon  EIFHKRGVAITAEEARKPMGLLKIDHVRVT

PGPhos    RVLAARGEAPDAHRQAELQGRFMAHYEADP
Phosphon  EMPRIASEWNRVFRQLPTEADIQEMYEEFE

PGPhos    ATLTSVYP-------GAEAAIRHLRAEGWR
Phosphon  EILFAILPRYASPINGVKEVIASLREGIK

PGPhos    IGLCTNKPVGASRQILSLF---GLLELF--
Phosphon  IGSTT----GYTREMMDIVAKEAALQGYKP

PGPhos    DAIIGGESLPQRKPDPAPLRATAAALN---
Phosphon  DFLVTPDDVPAGRPYPWMSYKNAMELGVYP

PGPhos    EEVVLYVGDSEVDAATAEAAGLRFALFTEG
Phosphon  MNHMIKVGDTVSDMKEGRNAGMWTVGVILG

PGPhos    YRHAPV--HELPHHGLFSHHDELQDLLRRL
Phosphon  SSELGLTEEEVENMDSVELREKIEVVRNRF
```

BayGenomics

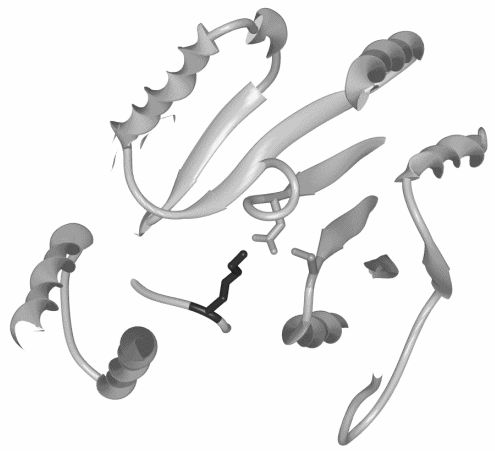
```
                        10                      151                     176
                         *                       *                       *
Cu++ATPase.Ec       LDTVVFDKTGTLTEG    VIAGVLPDGKAEAIKHL    AMVGDGINDAPAL
Cu++ATPase.Hs       VKVVVFDKTGTITHG    VFAEVLPSHKVAKVKQL    AMVGDGINDSPAL
Ca++ATPase.At       ATTICSDKTGTLTTN    VMARSSPMDKHTLVRLL    AVTGDGTNDAPAL
Urf.Mj              KVAIVFDSAGTLVKI    E---AHQELKRDLIRNL    IMVGDGANDVPAM
PhosSerPhos.Hs      ADAVCFDVDSTVIRE    TAE-SGGKGKVIKLLKE    IMIGDGATDMEAC
2-DO-6-PPhos.Sc     VDLCLFDLDGTIVST    ITGFDVKNGKPDPEGYS    VVFEDAPVGIKAG
DL-Gly-3-Phos.Sc    INAALFDVDGTIIIS    ITANDVKQGKPHPEPYL    VVFEDAPAGIAAG
Phosphon.Pa         LQAAILDWAGTVVDF    ATDEV-PNGRPWPAQAL    VKVDDTWPGILEG
Phosphon.St         IHAVILDWAGTTVDF    ATDDLAAGGRPGPWMAL    VKVDDAAPGISEG
Phosphon.Bc         IEAVIFDWAGTTVDY    TPDDV-PAGRPYPWMSY    IKVGDTVSDMKEG
PhosGlycolPhos.Rs   MPGVVFDLDGTLVHS    IGGESLPQRKPDPAPLA    LYVGDSEVDAATA
NtermDom.IGPD.Pp    VQALLLDMDGVMAEV    LEDCPP---KPSPEPIL    AMVGDTVDDIIAG
B-PhosGlucoMut.Ll   FKAVLFDLDGVITDT    AEVAAS---KPAPDIFI    IGLEDSQAGIQAI
HaloAcidDehal.PspYL IKGIAFDLYGTLFDV    LSVDPVQVYKPDNRVYE    LFVSSNAWDATGA
NtermDomEpoxHyd.Hs  LRAAVFDLDGVLALP    IESCQVGMVKPEPQIYK    VFLDDIGANLKPA
EnolasePhos.Ko      IRAIVTDIEGTTSDI    FD--TLVGAKREAQSYR    LFLSDIHQELDAA
```

BayGenomics


```
                        10                      151                     176
                         *                       *                       *
Cu++ATPase.Ec       LDTVVFDKTGTLTEG    VIAGVLPDGKAEAIKHL    AMVGDGINDAPAL
Cu++ATPase.Hs       VKVVVFDKTGTITHG    VFAEVLPSHKVAKVKQL    AMVGDGINDSPAL
Ca++ATPase.At       ATTICSDKTGTLTTN    VMARSSPMDKHTLVRLL    AVTGDGTNDAPAL
Urf.Mj              KVAIVFDSAGTLVKI    E---AHQELKRDLIRNL    IMVGDGANDVPAM
PhosSerPhos.Hs      ADAVCFDVDSTVIRE    TAE-SGGKGKVIKLLKE    IMIGDGATDMEAC
2-DO-6-PPhos.Sc     VDLCLFDLDGTIVST    ITGFDVKNGKPDPEGYS    VVFEDAPVGIKAG
DL-Gly-3-Phos.Sc    INAALFDVDGTIIIS    ITANDVKQGKPHPEPYL    VVFEDAPAGIAAG
Phosphon.Pa         LQAAILDWAGTVVDF    ATDEV-PNGRPWPAQAL    VKVDDTWPGILEG
Phosphon.St         IHAVILDWAGTTVDF    ATDDLAAGGRPGPWMAL    VKVDDAAPGISEG
Phosphon.Bc         IEAVIFDWAGTTVDY    TPDDV-PAGRPYPWMSY    IKVGDTVSDMKEG
PhosGlycolPhos.Rs   MPGVVFDLDGTLVHS    IGGESLPQRKPDPAPLA    LYVGDSEVDAATA
NtermDom.IGPD.Pp    VQALLLDMDGVMAEV    LEDCPP---KPSPEPIL    AMVGDTVDDIIAG
B-PhosGlucoMut.Ll   FKAVLFDLDGVITDT    AEVAAS---KPAPDIFI    IGLEDSQAGIQAI
HaloAcidDehal.PspYL IKGIAFDLYGTLFDV    LSVDPVQVYKPDNRVYE    LFVSSNAWDATGA
NtermDomEpoxHyd.Hs  LRAAVFDLDGVLALP    IESCQVGMVKPEPQIYK    VFLDDIGANLKPA
EnolasePhos.Ko      IRAIVTDIEGTTSDI    FD--TLVGAKREAQSYR    LFLSDIHQELDAA
```

BayGenomics

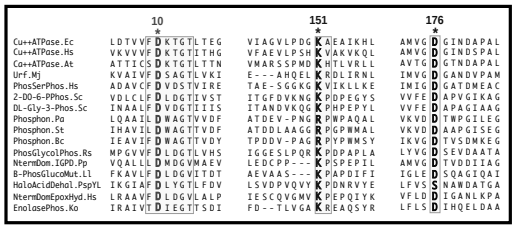
Active site of haloacid dehalogenase

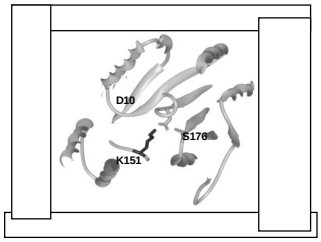


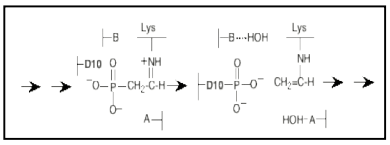
BayGenomics

	10	151	176
	*	*	*
Cu++ATPase.Ec	LDTVVFDKTGTLTEG	VIAGVLPDGKAEAIKHL	AMVGDGINDAPAL
Cu++ATPase.Hs	VKVVVFDKTGTITHG	VFAEVLPSHKVAKVKQL	AMVGDGINDSPAL
Ca++ATPase.At	ATTICSDKTGTLTTN	VMARSSPMDKHTLVRLL	AVTGDGTNDAPAL
Urf.Mj	KVAIVFDSAGTLVKI	E---AHQELKRDLIRNL	IMVGDGANDVPAM
PhosSerPhos.Hs	ADAVCFDVDSTVIRE	TAE-SGGKGKVIKLLKE	IMIGDGATDMEAC
2-DO-6-PPhos.Sc	VDLCLFDLDGTIVST	ITGFDVKNGKPDPEGYS	VVFEDAPVGIKAG
DL-Gly-3-Phos.Sc	INAALFDVDGTIIIS	ITANDVKQGKPHPEPYL	VVFEDAPAGIAAG
Phosphon.Pa	LQAAILDWAGTVVDF	ATDEV-PNGRPWPAQAL	VKVDDTWPGILEG
Phosphon.St	IHAVILDWAGTTVDF	ATDDLAAGGRPGPWMAL	VKVDDAAPGISEG
Phosphon.Bc	IEAVIFDWAGTTVDY	TPDDV-PAGRPYPWMSY	IKVGDTVSDMKEG
PhosGlycolPhos.Rs	MPGVVFDLDGTLVHS	IGGESLPQRKPDPAPLA	LYVGDSEVDAATA
NtermDom.IGPD.Pp	VQALLLDMDGVMAEV	LEDCPP---KPSPEPIL	AMVGDTVDDIIAG
B-PhosGlucoMut.Ll	FKAVLFDLDGVITDT	AEVAAS---KPAPDIFI	IGLEDSQAGIQAI
HaloAcidDehal.PspYL	IKGIAFDLYGTLFDV	LSVDPVQVYKPDNRVYE	LFVSSNAWDATGA
NtermDomEpoxHyd.Hs	LRAAVFDLDGVLALP	IESCQVGMVKPEPQIYK	VFLDDIGANLKPA
EnolasePhos.Ko	IRAIVTDIEGTTSDI	FD--TLVGAKREAQSYR	LFLSDIHQELDAA

BayGenomics





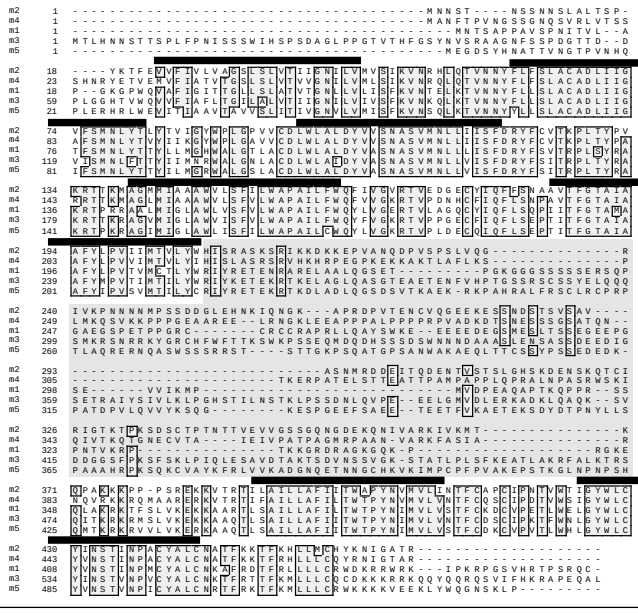


Issues in using multiple alignment information

- 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

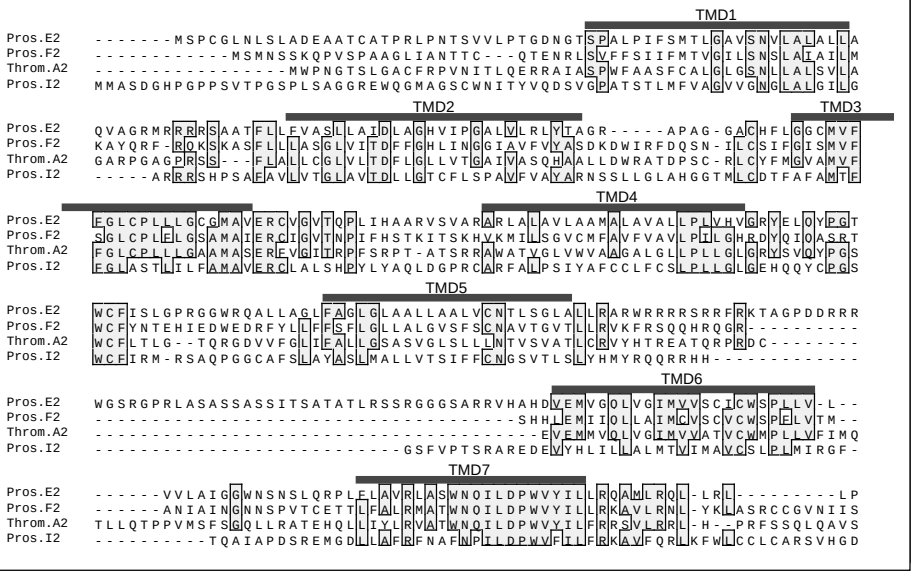
Muscarinic Receptor Sub-types (45-60% identical)



I3 loop

BayGenomics

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



BayGenomics

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

```
                    TMD1                                TMD2
  1  ----MLDHITPFFALVAFFLVLMPFAWHIKSKNVGLIMLSIWLMLGNLDNFVNSMVWWK----TT
  1  MFSGKENVSFGVLCLLAGCISTSSCLIHLQAKNIGVLLMMFWCFTGLVNKGINALAFNN----SL
  1  ---MSYKSAIIGLCLLAVILLAPPLAWHSHTKNIPAIILITWLLTMNLTCIVDAAIWSDDDFLT-
  1  --MLPIGIFYQFYAYFALVLSIPILYMQLRARNIPCLLLLFWLTLTTLIYVVESAIWSNPYAETI
                    TMD3                                TMD4
 58  ADLAPAYCELSVRLRHLLFIAIPASNLAIARKLESIASTRQVRAGPGDHRRAVIIDLLICLGIPI
 62  RLAWTLGCDLSAIIERTWQFGLCCSALCVLQRLEGIASLRQAHSTVWDRKRRLLIDFGVGLGLPA
 62  RWDGKGWCDIVIKLQVGANIGISCAVTNIIYNLHTILK-ADSVLPDLSSWTKIVKDLVISLFTPV
 64  RWMGYGLCDITSRIVTCSSIGIPASAFTLVLYLDTVIR-RDHPLKRYENW---IWHVCLSILLPL
                                          TMD5                    I3
123  IYTSLMIVNQSNRYGILEEAGCWPMMVFSWLWVLLVAAPVIVVSLCSAVYSALAFRWFWVRRRQF
127  LQIPMFFIVQPYRLNVIENIGCSAPLYASVPALFIYHLWRLLVSLVCAVYAVLVLRWFMLRRRQF
126  MVMGFSYLLQVFRYGIARYNGCQNLLSPTWITTVLYTMWMLIWSFVGAVYATLVLFVFYKKRKDV
125  IIMAMMVPLESNRYVVICMNGCYSSFYQTWYTLLFFYIPPCLLSFGGLFFVSRIVVLYWRRQREL
                    TMD6
188  QAVLASSASTINRSHYVRLLLLTAIDMLLFEPIYVGTIAAQI-KSSISIPYGSWSSVHTGFNQIP
192  TAALSSQHSGLSQKKYFRLFALAICERVLVSAGQFYVIIQSL-QIGGLLPYTSWAEVHTNFNRIL
191  RDILHCTNSGLNLTRFARLLIFCFIIILVMFPFSVYTFVQDLQQVEGHYIFKNTHSSTIWNTIIK
190  QQFFQ-RDSQLTSKRFLRLLCLAAVFFLGYFPLTIFMVVAN-GKLQQFLPFNHELVEAWHQESIT
                    TMD7
252  QYPASLVLMENTFQRNLILARLVCPLSAYIFFAMFGLGLEVRQGYKEAFHRA-LLFCRLRKEPKA
256  FVPVDTIAHSSLL--SLSILRWFSLTPAMALFVFFGLTEEAQSVYKARWKAL-INLC----SSKG
256  FDPGRPI-YNI----------WLYVLMSYLVFLIFGLGSDALHMYSKFLRSIKLGFVLDMWKRFI
253  YYPTTKVGLND----------WVPPTVLYLMSLFFSTSGGWTEKVALILWSLLVWLPFTK-----

316  SALQHVVADIEVVTFRSHDTFDANTSTKSEKSDIDMRGSEAA-----------------------
314  KKQTDGRESLDLEAFESHG---------SKFSVLVQRDTVIC-----------------------
310  DKNKEKRVGILLNKLSSRKESRNPFSTDSENYISTCTENYSPCVGTPISQAHFYVDYRIPDDPRK
303  NTALGRHAQFKLDCCKSIESTMAGKTLDSTDFKEKC-----------LVLERQWSKSSIPSDNSS
```

BayGenomics

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

```
enol1  EAMKMGAEVYHHLKSVIKKKYGQDATNVGDEGGFAPNIQENKEGL
enol2  EAMKMGCEVYHHLKAVIKKKYGQDATNVGDEGGFAPNIQENKEGL
enol3  EALRIGSEVYHNLKSLTKKKYGQSAGNVGDEGGVAPDIKTPKEAL
enol4  EAMKMGVEVYHNLKSIIKKKYGQDATNVGDEGGFAPNIQENKEGL
enol5  EALRIGSEVYHNLKSLTKKRYGASAGNVGDEGGVAPNIQTAEEAL
enol6  EALKMGSEVYHALKSVIKAKYGQDACNVGDEGGFAPNIQDNKEGL
enol7  EAMKMGSEVYHHLKNVIKAKFGLDATAVGDEGGFAPNIQSNKEAL
enol8  DAMRVGAEVYHSLKGVIKAKYGKDATNVGDEGGFAPNILDNHEAL
cpeps  D-IEVADRVFTAAHRNVERRFGPVPLS-ASSGLMVP--LDSAGQL

enol1  ELLKTAIAKAGYTGKVVIGMDVAASEFYG-SDKTYDLNFKEENND
enol2  ELLKTAIEKAGYTGKVVIGMDVAASEFYG-KDKSYDLNFKEESND
enol3  DLIMDAIDKAGYKGKVGIAMDVASSEFY--KDGKYDLDFKNPESD
enol4  ELLKAAIEKAGYTGKVVIGMDVAASEFFGEKDKTYDLNFKEENND
enol5  DLIVDAIKAAGHDGKVKIGLDCASSEFF--KDGKYDLDFKNPNSD
enol6  ELLNEAIAKAGYTGKVKIGMDVASSEFY--KDGKYDLDFKNPNSD
enol7  NLISDAIAKAGYTGKIEIGMDVAASEFY--KDGQYDLDFKNEKSD
enol8  ELLKAAIAQAGYTDKVVIGMDVAASEFC--RDGRYDLDFKSP-PD
cpeps  DLLQAAVAETGHTEVCTLGVDVAA-EHLLTEPGRYRF--------
```

BayGenomics

```
enol1  EAMKMGAEVYHHLKSVIKKKYGQDATNVGDEGGFAPNIQENKEGL
enol2  EAMKMGCEVYHHLKAVIKKKYGQDATNVGDEGGFAPNIQENKEGL
enol3  EALRIGSEVYHNLKSLTKKKYGQSAGNVGDEGGVAPDIKTPKEAL
enol4  EAMKMGVEVYHNLKSIIKKKYGQDATNVGDEGGFAPNIQENKEGL
enol5  EALRIGSEVYHNLKSLTKKRYGASAGNVGDEGGVAPNIQTAEEAL
enol6  EALKMGSEVYHALKSVIKAKYGQDACNVGDEGGFAPNIQDNKEGL
enol7  EAMKMGSEVYHHLKNVIKAKFGLDATAVGDEGGFAPNIQSNKEAL
enol8  DAMRVGAEVYHSLKGVIKAKYGKDATNVGDEGGFAPNILDNHEAL
cpeps  D-IEVADRVFTAAHRNVERRFGPVPLS-ASSGLMVP--LDSAGQL

enol1  ELLKTAIAKAGYTGKVVIGMDVAASEFYG-SDKTYDLNFKEENND
enol2  ELLKTAIEKAGYTGKVVIGMDVAASEFYG-KDKSYDLNFKEESND
enol3  DLIMDAIDKAGYKGKVGIAMDVASSEFY--KDGKYDLDFKNPESD
enol4  ELLKAAIEKAGYTGKVVIGMDVAASEFFGEKDKTYDLNFKEENND
enol5  DLIVDAIKAAGHDGKVKIGLDCASSEFF--KDGKYDLDFKNPNSD
enol6  ELLNEAIAKAGYTGKVKIGMDVASSEFY--KDGKYDLDFKNPNSD
enol7  NLISDAIAKAGYTGKIEIGMDVAASEFY--KDGQYDLDFKNEKSD
enol8  ELLKAAIAQAGYTDKVVIGMDVAASEFC--RDGRYDLDFKSP-PD
cpeps  DLLQAAVAETGHTEVCTLGVDVAA-EHLLTEPGRYRF--------
```

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

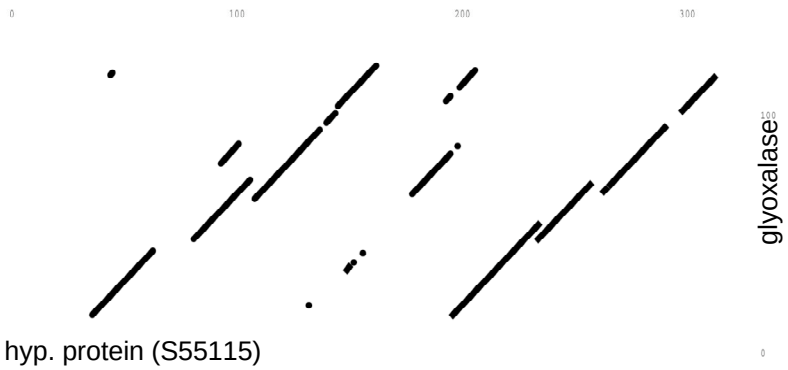
BayGenomics

```
gram.pos   LKAK--GMNTAVGDEGGYAPNLGSNDEALAVIA
gram.neg   LSAK--GMNTNVGDEGGFAPSLDSASSALDFIV
eukaryote  TKKRYGASAGNVGDEGGVAPNIQTAEEALDLIV
archea.    LADR--DLPAGKGDEGAWAPSV-SDDEAFEIMD
cpeps      VERRFGPVP--LSASSGLMVPLDSAGQ-LDLLQ

gram.pos   EAVKAAGYELGKDITLAMDCAASEFYKD-GK--
gram.neg   DSISKAGYKPGEDVFIALDAASSEFYNK-DQNI
eukaryote  DAIKAAGH--DGKVKIGLDCASSEFFKD-GKYD
archea.    EAVETVADDFGFAISFGLDVARAELYDD-EADG
cpeps      AAVAETGH--TEVCTLGVDVAAEHLLTEPGRYR

gram.pos   YVLA---GEGNKAFTSEEFTHFLEELTKQYPIV
gram.neg   YDLK---GEGRK-LTSAQLVDYYVELCGKYPIY
eukaryote  LDFKNPNSDKSKWLTGPQLADLYHSLMKRYPIV
archea.    YVY----DDGVK--STEEQIEYIAGKVEEYDLV
cpeps      F--------GDRVLTAPDFADHLADLAHRFRMS
```

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

BayGenomics

How do you handle internal repeats?




```
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)

BayGenomics

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

BayGenomics

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 (Gribskov, 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
- Systers: http://www.dkfz-heidelberg.de/tbi/services/documentation/systershelp.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

BayGenomics

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 Æ 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

BayGenomics

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

BayGenomics

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)

BayGenomics

- 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

BayGenomics

New Directions in Protein Bioinformatics

BayGenomics

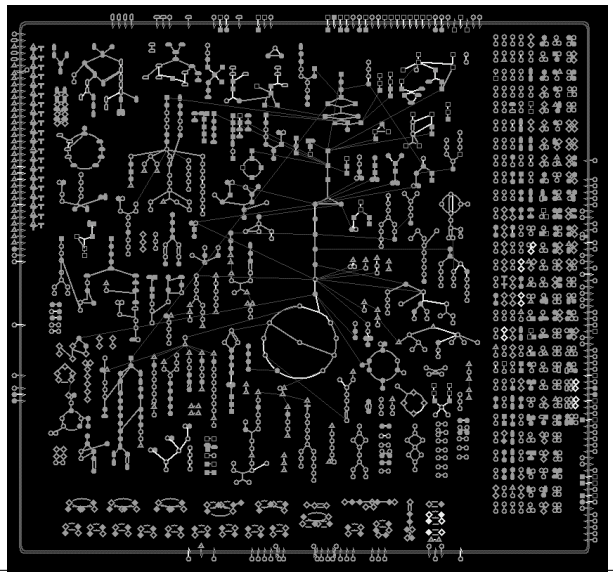
Using Protein Informatics for Really New Insight into Biology

- Comparative genomics
  - Metabolic computing: EcoCyc & MetaCyc
    http://ecocyc.org/ecocyc/index.html
  - **C**lusters of **O**rthologous **G**roups (COGS)
    http://www.ncbi.nlm.nih.gov/COG/
- Genetic circuits/Systems analysis
  http://gobi.lbl.gov/~aparkin/index.html
- Protein-Protein Interactions
  - Co-evolution

BayGenomics

Overview of E. coli metabolic systems

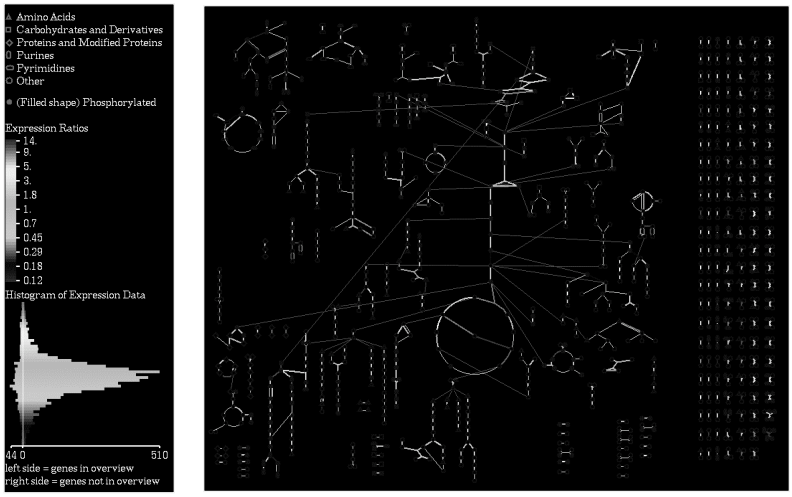
used with permission: Peter D. Karp (EcoCyc)



BayGenomics

MetaCyc: Yeast Expression Data

used with permission: Peter D. Karp (EcoCyc)



BayGenomics

Nature 1999 Nov 4;402(6757):83

6

"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.

BayGenomics

A few important topics we didn't even mention

- Mapping Sequence Æ Structure Æ 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. , Jan. issue each year
  - Description and useful information on 112 databases of interest to the genomics/proteomics/bioinformatics communities