

## Lecture 2

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

Patricia C. Babbitt  
Associate Professor  
Department of Biopharmaceutical Sciences  
& Biological and Medical Information Sciences Graduate Group  
August 22, 2002

- 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:

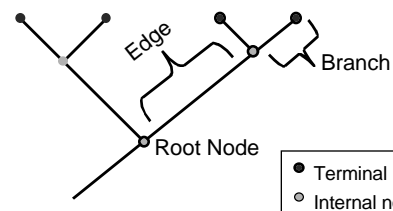
Fundamentals of Molecular Evolution, Li & Graur, Sinauer Associates, 2nd Ed. (1999)

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

## Web Resources

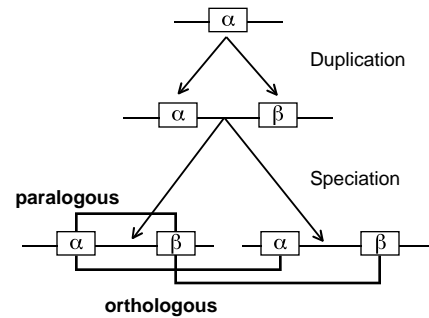
- Useful Lists  
<http://www.mcb.harvard.edu/BioLinks/Evolution.html>  
<http://darwin.eeb.uconn.edu/molecular-evolution.html>
- Tree of Life site  
<http://phylogeny.arizona.edu/tree/phylogeny.html>
- A protocol to get you started  
<http://www.infobiogen.fr/docs/MAcours/phylogeny.html>

## Tree (Network) Nomenclature



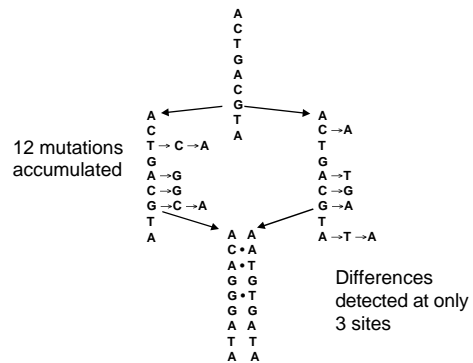
## Definitions

- 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

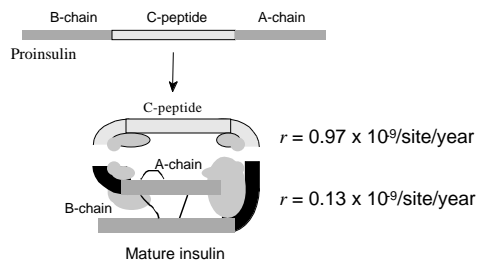


## Important principles

- 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



## 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

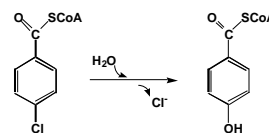
Score = 207 (96.1 bits), Expect = 6.1e-31, Sum P(3) = 6.1e-31  
Identities = 51/137 (37%), Positives = 71/137 (51%)

Query: 89 WHQMIIKIRKRVPLAAINGVAAGGGLGISLASDMAICADSAKFCVWHTTIGINDTAT 148  
+ ++ I FVLAA+NG AAG G ++LA+D+ I A SA F+ A+ IG+ D  
Sbjct: 83 YEPLLQATYSCLPLFLAANNGAAGGANLALAADVVIAQSAAPFQAFTRIGLMDAGG 142

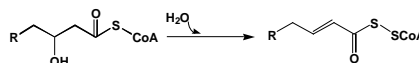
Query: 149 SYSIARTVGMRRAMELMLTNRITLYPEAKDGLVSRVYPKDEPREVAMKVVARELAAPTH 208  
++ L R VGM RAM + L + EEA GL+ P +F  
Sbjct: 143 TWMLPRQVMRAMGMALFAEKIGAEAAARMGLIWEAVPDVDFEHWRAAHLARGPSA 202

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

### 4-Chlorobenzoate Dehalogenase

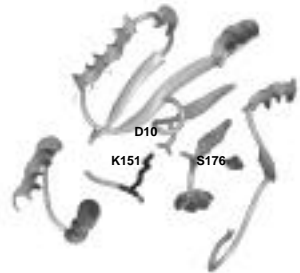


### Enoyl-CoA Hydratase



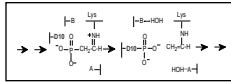
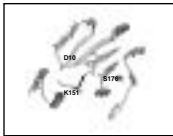


### Active site of haloacid dehalogenase



	10	151	176
Co→ATPase: Ec	LDTVVFDKATGTLTEG	VIAGVLPDCKKAEIKHL	AMVGDCINDAPAL
Co→ATPase: Bs	VKVVVFDATGTLTHG	VFAEVLPSHKVAKVKQL	AMVGDGINDSPAL
Co→ATPase: At	ATTICSDATGTLTTN	VMARSSPMDKHTLVRL	AVTGDGTNDAPAL
Trf: M	KVAIVFDKAGTLVKI	E--AHQELKRDILRNL	IMVGDGANDVPAM
PhosSerPhos: Bs	ADAVCFDYDSTVIRE	TAE-SGGKCKVILKKE	IMIGDGATDMEAK
2-De-6-Phos: Sc	VDLCLFDLDGTLIVST	ITGFVDKNGCKPDPGYS	VVFE DAPVGIKAG
DL-Gly-3-Phos: Sc	INAALFDVDGTLIIS	ITANDVKQCKPDPPEYL	VVFE DAPAGIAAG
Phosphon: Pa	LQAALFDWAGTVVDF	ATDEV-PNGKPPWAQAL	VKVD DTPWGLEEG
Phosphon: St	IHAVLFDWAGTVVDF	ATDOLAAGRPPGPMAL	VKVD D AAPGISEG
Phosphon: Bc	I EAVLFDWAGTVVDF	TPDDV-PAGRPPWMSY	IKVD DTVSDMKEG
PhosGlycol Phos: Bs	MPGVVFDLDGTLVHS	IGGESLPQRKDPAPLA	LYVG DSEVDAATA
Nteratom: I GPD: Pp	VQALLFDMDGVMAEV	LEDPPP--KHPPEPLI	AMVGD TVDDIAG
B-PhosGlucoMn: Ll	FKAALFDLDGTLTDT	AEVAAS--KHPDPTFI	IGLE DSAQGIQAI
Bal oclidihal: PspYl	IKGIAFDLYGTLFDV	LSVDPVQVYKPDNRVYE	LFVSDSNAWDTGA
Nteratom: poxid: Bs	LRAAVFDLDGVLALP	IESCQVGMVKEPQIYR	VFLD D I GANLKPA
EnolasePhos: Ko	IRAIVTDIEGTSDDI	FD--TLVGAKEAQSYR	LFLS D I RQELDAA

	10	151	176
Co→ATPase: Ec	LDTVVFDKATGTLTEG	VIAGVLPDCKKAEIKHL	AMVGDCINDAPAL
Co→ATPase: Bs	VKVVVFDATGTLTHG	VFAEVLPSHKVAKVKQL	AMVGDGINDSPAL
Co→ATPase: At	ATTICSDATGTLTTN	VMARSSPMDKHTLVRL	AVTGDGTNDAPAL
Trf: M	KVAIVFDKAGTLVKI	E--AHQELKRDILRNL	IMVGDGANDVPAM
PhosSerPhos: Bs	ADAVCFDYDSTVIRE	TAE-SGGKCKVILKKE	IMIGDGATDMEAK
2-De-6-Phos: Sc	VDLCLFDLDGTLIVST	ITGFVDKNGCKPDPGYS	VVFE DAPVGIKAG
DL-Gly-3-Phos: Sc	INAALFDVDGTLIIS	ITANDVKQCKPDPPEYL	VVFE DAPAGIAAG
Phosphon: Pa	LQAALFDWAGTVVDF	ATDEV-PNGKPPWAQAL	VKVD DTPWGLEEG
Phosphon: St	IHAVLFDWAGTVVDF	ATDOLAAGRPPGPMAL	VKVD D AAPGISEG
Phosphon: Bc	I EAVLFDWAGTVVDF	TPDDV-PAGRPPWMSY	IKVD DTVSDMKEG
PhosGlycol Phos: Bs	MPGVVFDLDGTLVHS	IGGESLPQRKDPAPLA	LYVG DSEVDAATA
Nteratom: I GPD: Pp	VQALLFDMDGVMAEV	LEDPPP--KHPPEPLI	AMVGD TVDDIAG
B-PhosGlucoMn: Ll	FKAALFDLDGTLTDT	AEVAAS--KHPDPTFI	IGLE DSAQGIQAI
Bal oclidihal: PspYl	IKGIAFDLYGTLFDV	LSVDPVQVYKPDNRVYE	LFVSDSNAWDTGA
Nteratom: poxid: Bs	LRAAVFDLDGVLALP	IESCQVGMVKEPQIYR	VFLD D I GANLKPA
EnolasePhos: Ko	IRAIVTDIEGTSDDI	FD--TLVGAKEAQSYR	LFLS D I RQELDAA

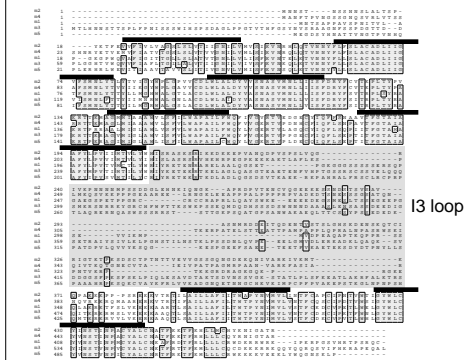


### Issues in using multiple alignment information

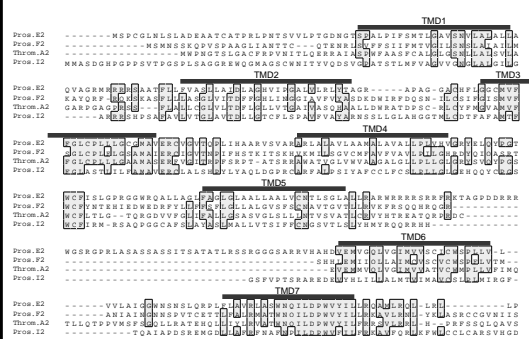
- What question are you asking when you create a multiple alignment?
  - Example: GPCRs

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)



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

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

enol1	EAMKMGAEVYHHLKSVIKKKYGDATNVGDEGGFAPNIQENKKEGL
enol2	EAMKMGCEVYHHLKAVIKKKYGDATNVGDEGGFAPNIQENKKEGL
enol3	EALRIQSEVYHNLKSLTKKKYQDAGNVGDEGGFAPNIQENKKEGL
enol4	EAMKMGVEVYHNLKSLIKKKYQDAGNVGDEGGFAPNIQENKKEGL
enol5	EALRIQSEVYHNLKSLTKKKYGAAGNVGDEGGFAPNIQENKKEGL
enol6	EALKMGSEVYHHLKSVIKKKYQDAGNVGDEGGFAPNIQENKKEGL
enol7	EAMKMGSEVYHHLKSVIKKKYQDAGNVGDEGGFAPNIQENKKEGL
enol8	DAMRVGAEVYHSLKGVIKKKYQDAGNVGDEGGFAPNIQENKKEGL
cpeps	D-IEVADRIVFTAAHRNVERRFQVPLS-ASGLMVLH--LDSAGQL

enol1	ELLKTALIAKAGYTGKVVIGMDVAASEFYG-SDKTYDLNFKENND
enol2	ELLKTALIEKAGYTGKVVIGMDVAASEFYG-KDKSYDLNFKESND
enol3	DLLMDAIDKAGYTGKVVIGMDVAASEFYG--KDKYDLDLDFKNPESD
enol4	ELLKAAIEKAGYTGKVVIGMDVAASEFYG-KDKTYDLNFKENND
enol5	DLLVDAIAKAGHDKGVKIGDCAISEFF--KDKYDLDLDFKNPESD
enol6	ELLNBAIAKAGTGKVVIGMDVAASEFYG--KDKYDLDLDFKNPESD
enol7	NLLSDAIAKAGYTGKVVIGMDVAASEFYG--KDKYDLDLDFKNPESD
enol8	ELLKAAIAKAGYTGKVVIGMDVAASEFYG--RDKYDLDLDFKNPESD
cpeps	DLLQAIAVAETGRIEVCITLGVDAALHLLTERGRVRF-----

enol1	EAMKMGAEVYHHLKSVIKKKYQDATNVGDEGGFAPNIQENKKEGL
enol2	EAMKMGCEVYHHLKAVIKKKYQDAGNVGDEGGFAPNIQENKKEGL
enol3	EALRIQSEVYHNLKSLTKKKYQDAGNVGDEGGFAPNIQENKKEGL
enol4	EAMKMGVEVYHNLKSLIKKKYQDAGNVGDEGGFAPNIQENKKEGL
enol5	EALRIQSEVYHNLKSLTKKKYGAAGNVGDEGGFAPNIQENKKEGL
enol6	EALKMGSEVYHHLKSVIKKKYQDAGNVGDEGGFAPNIQENKKEGL
enol7	EAMKMGSEVYHHLKSVIKKKYQDAGNVGDEGGFAPNIQENKKEGL
enol8	DAMRVGAEVYHSLKGVIKKKYQDAGNVGDEGGFAPNIQENKKEGL
cpeps	D-IEVADRIVFTAAHRNVERRFQVPLS-ASGLMVLH--LDSAGQL

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

gram.pos	LKAK--GMNTAVIGDEGGVAPNLGSDNDEALAVIA
gram.neg	LSAK--GMNTAVIGDEGGVAPNLDSSASSALDFIV
eukaryote	TKRYGASAGNVGDEGGVAPNIQTAEELDLIV
archaea	LADR--DLPAGKDEGGVAPNVSDDAEFAELMD
cpeps	VRRFGVPP--LSASSGLMVLPLSDAGQLDLLQ

gram.pos	EAVKAAQYELGKIDITLAMDCAASEFYKD-GK--
gram.neg	DSISKAGYKPGEDVFIALDCAASEFYFNK-DQNI
eukaryote	DIAKAGH--DGKVIKGLDCASSEFFKD-GKYD
archaea	EAVETVADDFGFAISFLDVARAEELLYD- EADG
cpeps	AAVAETGH--TEVCTLGVDAALHLLTERGRVRF

gram.pos	YVLA--GEGNKAFTSEEFTHFLEELTKQYPIV
gram.neg	YDLK--GEGRIK-LTSAQLVDYVELCGKQYPIV
eukaryote	LDFKNPNSDKSKWLLTGSQLADLYHSLMKRQYPIV
archaea	YVY--DDGVH--STEQEIYIAGKVEEYDLD
cpeps	F-----GDRVLTAPDFADHLADLAHRFRMS

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

### How do you handle internal repeats?

hyp. protein (S55115)

glyoxalase

q09751.N	VERSKRIGILELTYNFGTEKKEGPVYIN
s55115.N	PDVFSAHGVLELTHNWGTEKNPDYKINN
q09751.C	-----EGLLELTHNWGTEKESGPPVYHN
s55115.C	---VFSCESVLELTHNWGTEKENDPNFHYHN
glyox.	-----BAVIELTYNNGVDK-----YREL

q09751.N	GNTEPKRGFGHICFIVDNIESACAYLE-
s55115.N	GNBEPHRGFGHICFVSDINRKTCEELE-
q09751.C	GNDGDEKGYGHVCIISVDNINAACSKFIE-
s55115.C	GN-SEPOQYGHICISCDNAGALCKEIEV-
glyox.	GT-----AYGHIALESVDNAEAACEKIRQ

q09751.N	--SKGVSFKKLSLDEGKMKHIAF-----
s55115.N	--SQGVFKKRLSEGRQKDIADF-----
q09751.C	--AEGLPFKKLTDEGRMKDIAF-----FLLD
s55115.C	KYGDKIQWSPKFNQGRMKNIADF-----FLKD
glyox.	NGGNVTR EAGPVKGE-----TTVIAFVED

## 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 (dendrogram) 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) (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>
  - also at UCSF SACS

## 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/systershlp.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 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)



## 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