Introduction to Bioinformatics

Fundamental principles for inference of molecular function

3rd Penn State Bioinorganic Workshop

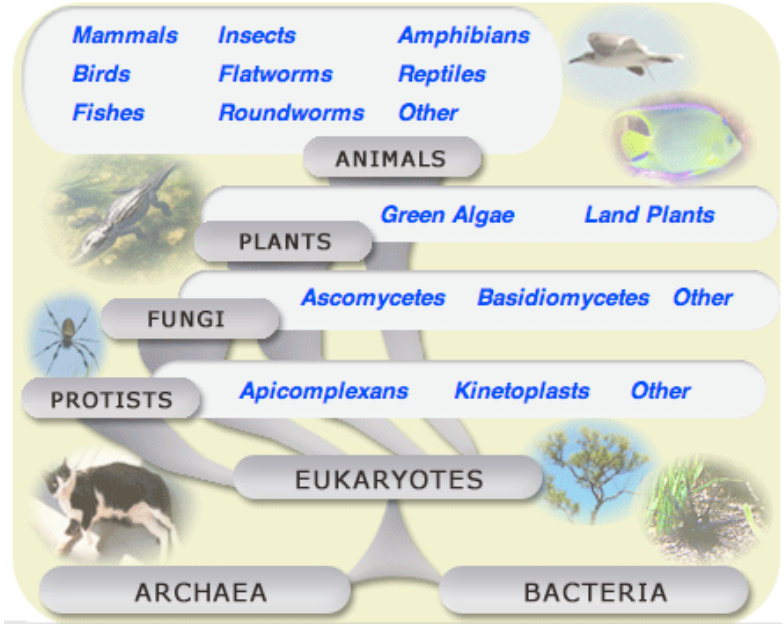
Patsy Babbitt
babbitt@cgl.ucsf.edu

Workshop Leaders: Gemma Holliday, Shoshana Brown

Function prediction & the annotation problem

>50 million protein sequences in UniProt/TrEMBL (5/14)

- Billions of metagenomic sequence reads already available

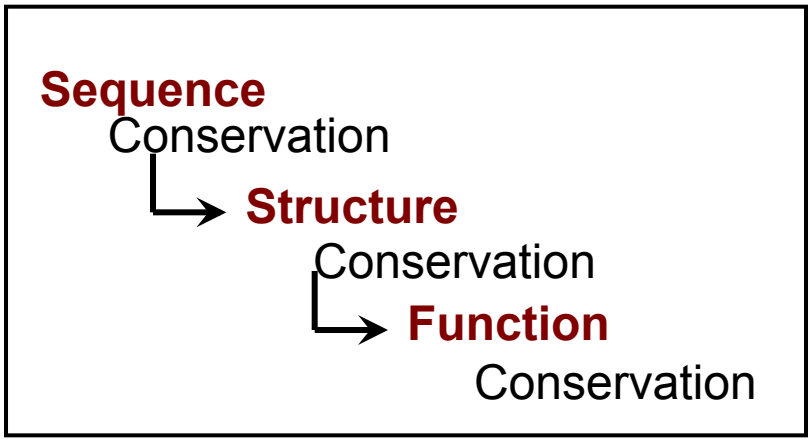


NCBI

- Effective use of these data requires bioinformatics & automated analysis
- Proportion for which annotations can be experimentally validated
- Proportion of ORFs that are annotated [correctly] with a molecular function

Today's main topic: Homology-based analysis

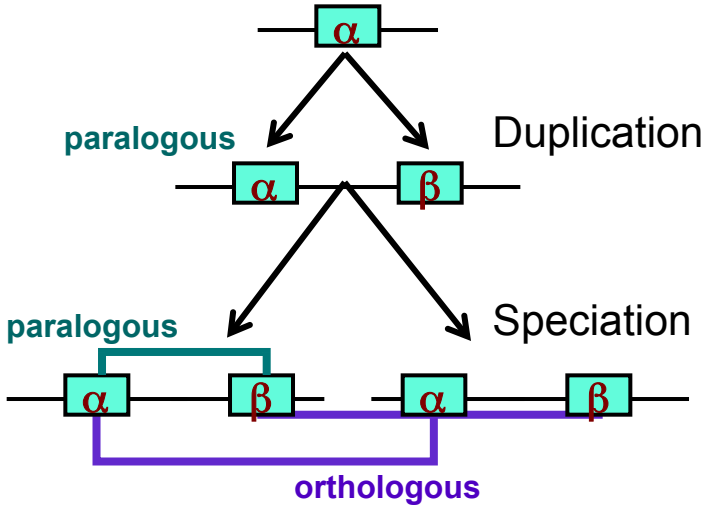
- Functional annotation of all of the gene products of genomes cannot be achieved experimentally or even from first principles
  - Requires the use of [computational] comparative genomics



- Assumption: What is conserved in a gene [protein] family over evolutionary timescales is functionally important
  - due to purifying selection driven by functional constraints observable in a background described by the theory of neutral evolution
- Application: Comparison of sequence/structures can identify homologous relationships, allowing inference of [molecular] functional properties

A very short primer on molecular evolution

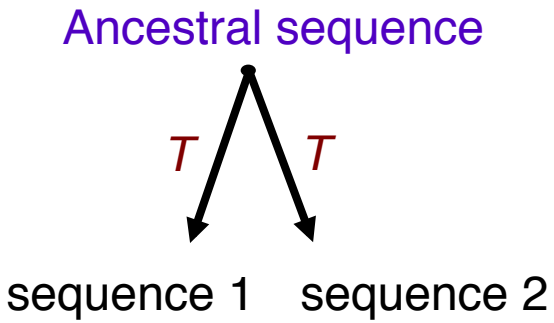
- Homologs: Share a common ancestor by descent or recombination; may have similar or dissimilar molecular functions
  - Orthologs: gene/protein similarity as a consequence of a speciation event
    - Often infers similar function but the concept only defines a path of evolution
  - Parologs: similarity between the descendants of a duplicated ancestral gen
    - Often infers dissimilar function but paralogs can also have similar function
  - Xenologs: appearance in an organism due to horizontal gene transfer

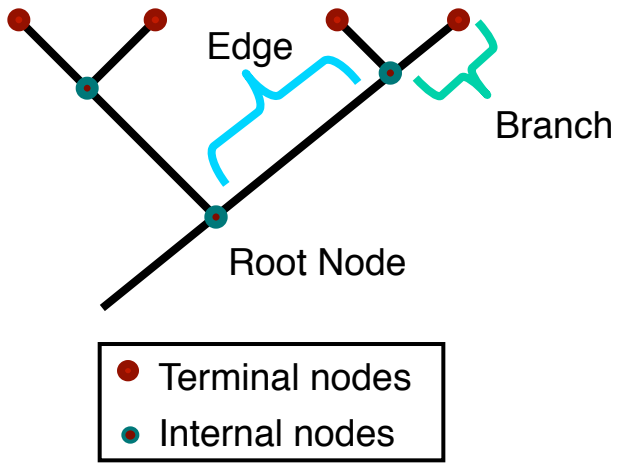


- Convergent evolution: Similar function, some similar structural characteristics, but do not share a common ancestor

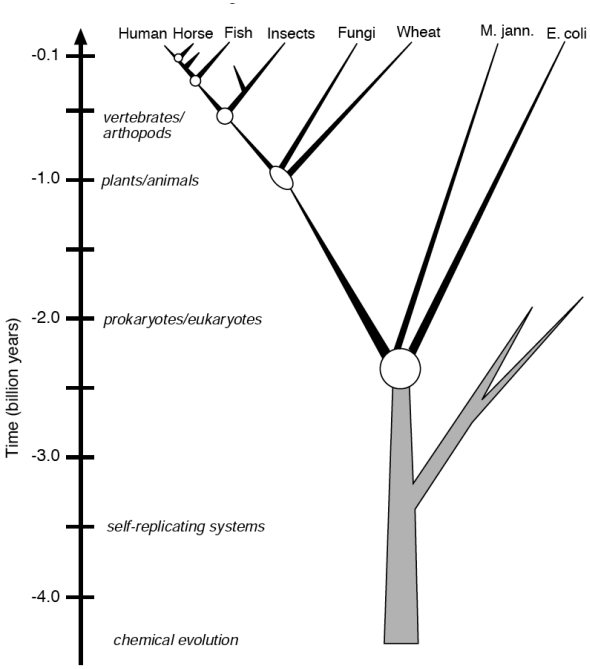
Assumptions & Caveats

- Phylogenetic reconstruction assumes homology across the set of sequences to be evaluated
- Differences among these sequences are assumed to derive from gene duplication followed by mutations and/or rearrangements
- Both the mechanisms and rates of divergence are critical issues for phylogenetic reconstruction
- A multiple alignment of homologous sequences provides the necessary data for counting nucleotide or amino acid substitutions used in phylogenetic reconstruction



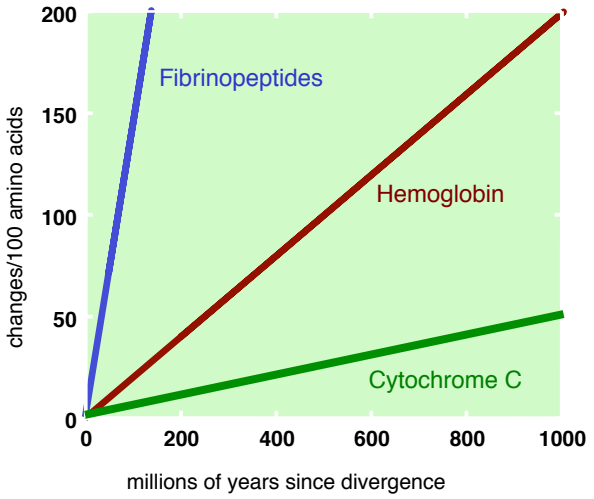


- Evolutionary history is accessed only through contemporary species and molecules
- For highly diverse sequences, obtaining a high confidence multiple alignment may be challenging
  - The quality of a phylogenetic tree is no better than the quality of the underlying alignment

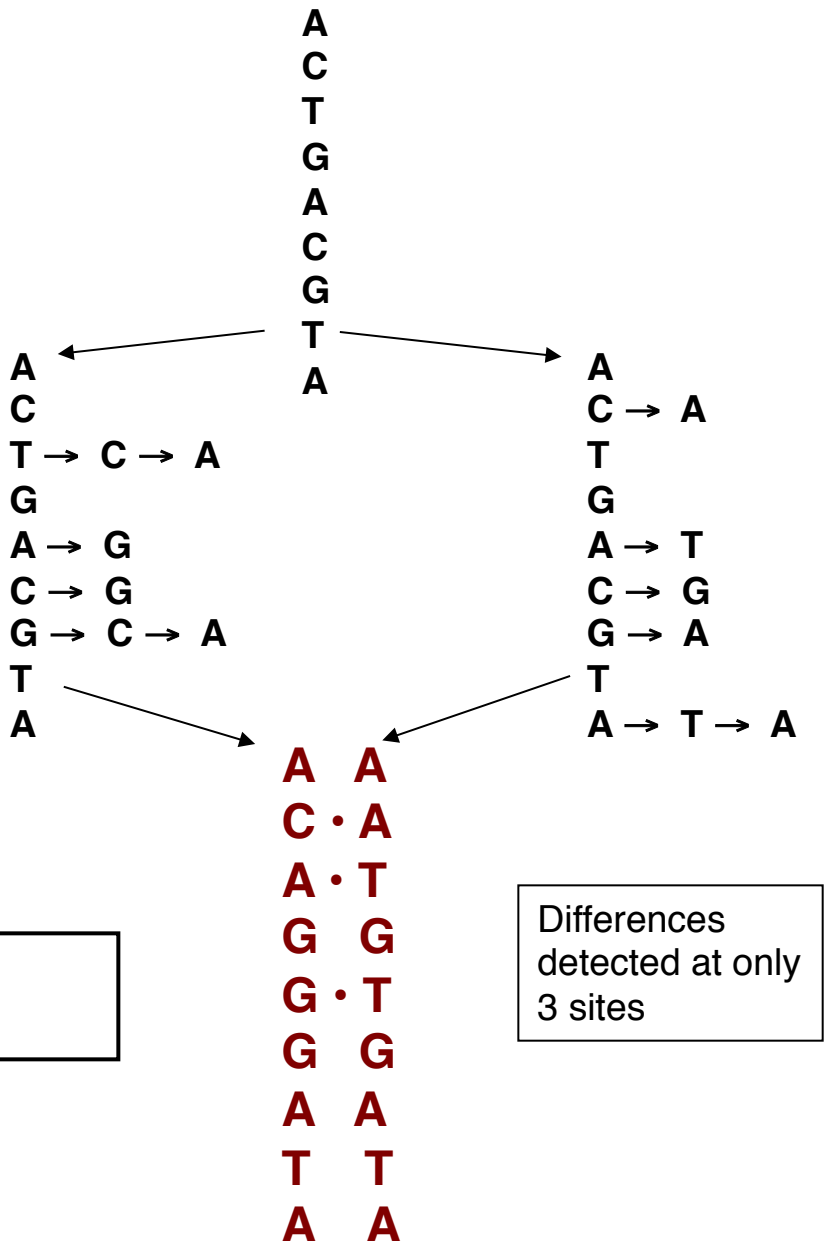


(from 2004 Mol.Evolution Workshop at Woods Hole: Pearson)

- Simplifying assumptions don't always hold
  - Independence of changes at each site or in different copies of a gene?
  - Do all sites changes at the same rate?
  - Do all genes evolve at the same rate?



Another complication:



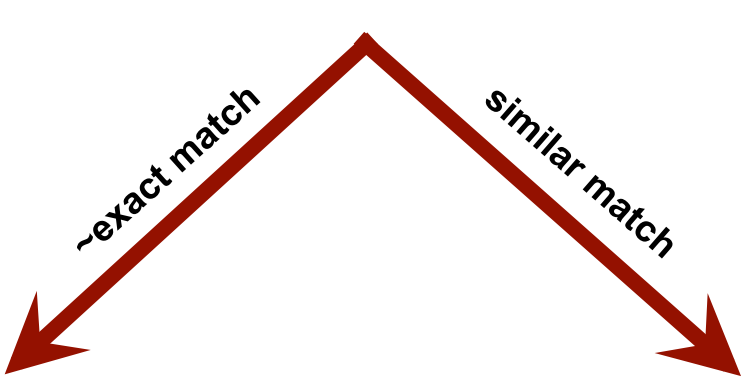
12 mutations accumulated

Differences detected at only 3 sites

PAM*250 = 250 nucleotide changes/100 amino acids

*Point accepted mutation

- Sequence analysis provides access to genomics/related information
  - Because it represents the primary data, accessing genomics web sites via sequence comparison bypasses problems associated with searching using key words, gene names, various types of accession #s



Access to many types of info about a gene/protein

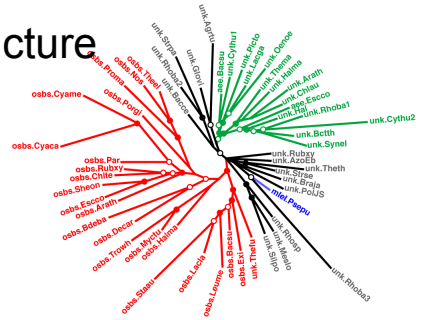
- localization (genome browsers)
- organism DBs
- specialty DBs (ex: Mouse knockouts, expression data)
- proteomics data

Near & distant homologs in multiple species

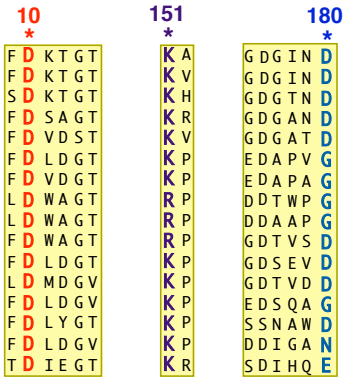
- primary sequence DBs
- precalculated family DBs
- gene families within or across species of interest (ex: Kinases)

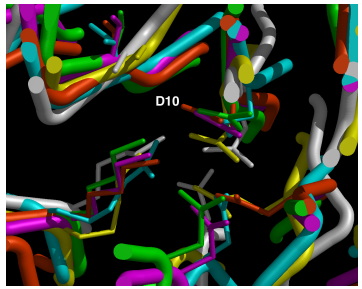
Protocol for [non-trivial] functional inference

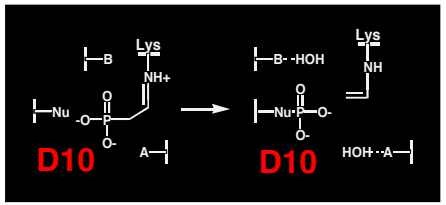
0. Sources of sequence & structure data & annotation
1. Identify [divergent] homologs using sequence & structure



2. Cluster the data at appropriate granularity
3. Map conserved elements of sequence & structure to conserved elements of function







4. Orthogonal approaches: genome localization & metabolic context, comparative genomics, in silico docking
5. Experiment!

0. Sources of sequence data & annotation

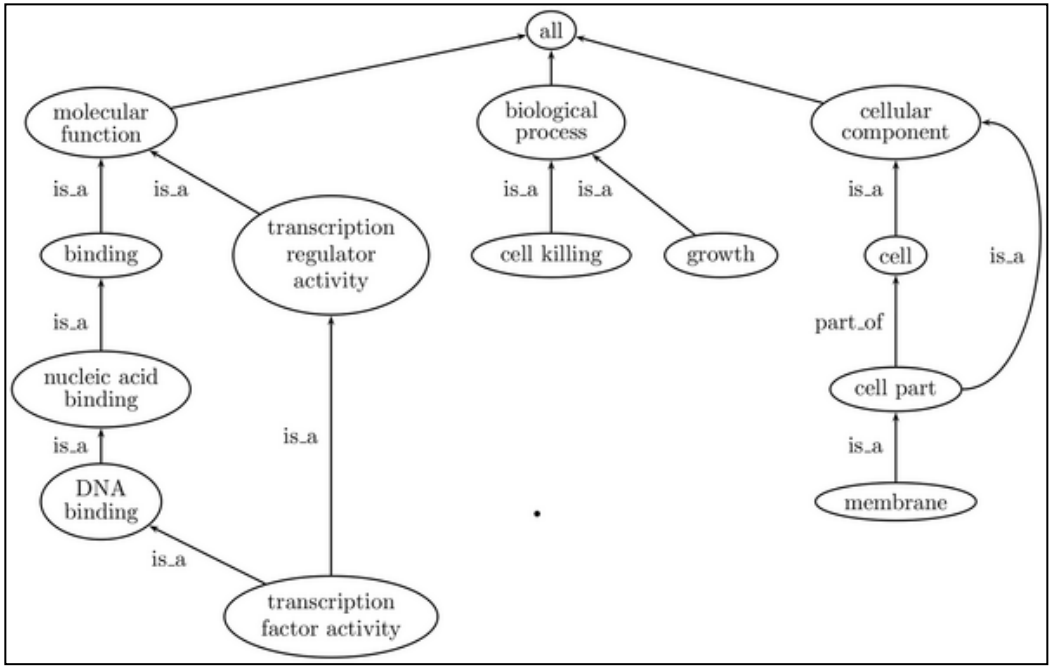
- Archive sites - comprehensive
  - Genbank (NCBI)
    - library; Genbank does't own its annotations
  - UniProt/TrEMBL
    - computer annotated translations of EMBL annotated nucleotide sequences
- Curated resources
  - UniProt/SwissProt ~500,000 sequences
    - manually reviewed knowledge base w Gene Ontology (GO) terms, functions, refs, post-translational modifications, domains & functional sites, similarities to other proteins, diseases associated w variants, etc.
  - Organism DBs
  - Metagenomics DBs
  - Thematic DBs
    - MEROPS (proteases)
    - GPCRDB (G-protein coupled receptors)
    - Structure-Function Linkage Database (SFLD), Catalytic Site Atlas, Mechanism, Annotation and Classification in Enzymes (MACiE) (enzymes)

Classification of protein sequence information

- **InterPro**
  - Functional classification into “families” (updated frequently)
  - Site & domain prediction using predictive models (signatures/ profiles) from member databases of Consortium
  - Used by the Gene Ontology Annotation group to automatically assign Gene Ontology terms
  - Member databases
    - CATH/Gene3D at University College, London, UK
    - PANTHER at University of Southern California, CA, USA
    - PIRSF at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, USA
    - Pfam at the Wellcome Trust Sanger Institute, Hinxton, UK
    - PRINTS at the University of Manchester, UK
    - ProDom at PRABI Villeurbanne, France
    - PROSITE and HAMAP at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland
    - SMART at EMBL, Heidelberg, Germany
    - SUPERFAMILY at the University of Bristol, UK
    - TIGRFAMs at the J. Craig Venter Institute, Rockville, MD, US

Classification & analysis using Gene Ontology (GO)

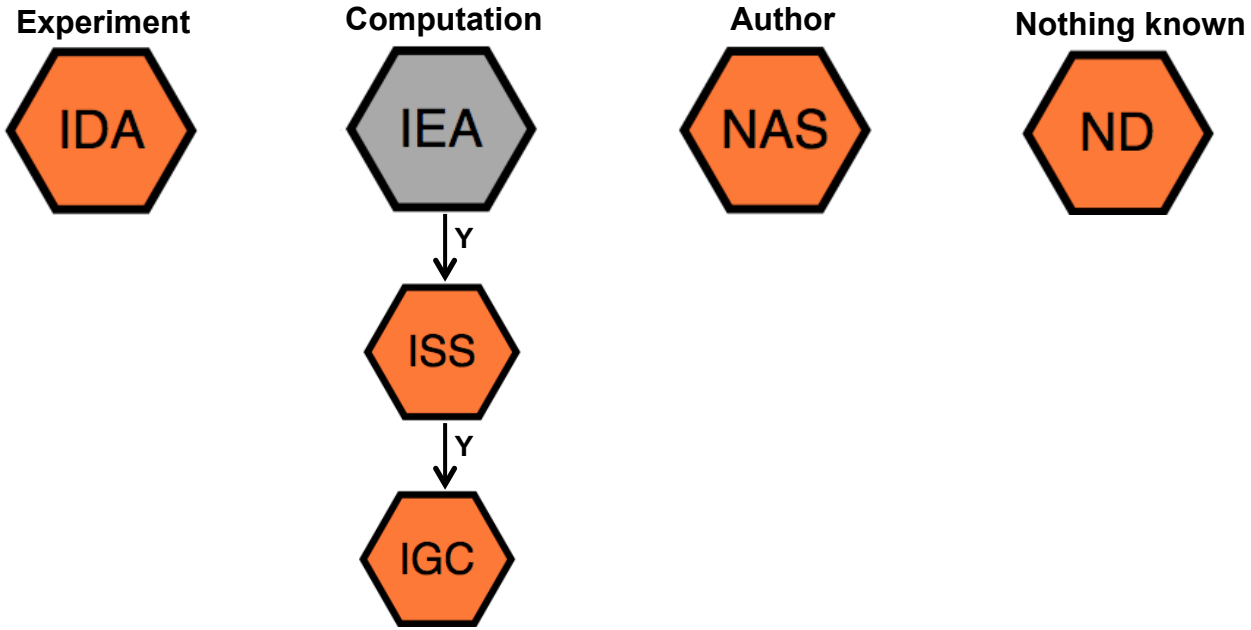
- Community collaboration for consistent descriptions of gene products in different databases
  - 3 structured vocabularies (ontologies)



- Main tasks
  - Development/maintenance of the ontologies (incomplete for enzymes, for example)
  - Annotation of gene products integrating data from collaborating databases
  - Tools development

Gene Ontology Annotation

- UniProt-GOA database aims for high quality GO annotation for UniProt Knowledge Base
- Many member & other databases use GO terms to facilitate uniform queries across them
- Evidence codes describe the types of evidence associated with annotations & are used by SwissProt, SFLD, many other DBs



1. Identification of divergent homologs

- Sequence-based methods
  - pairwise alignment & database searching
  - profile & other probabilistic methods
    - pre-computed solutions: Pfam
  - motif analysis
- Structure-based methods
  - structural alignment
  - fold-based comparisons
  - motif analysis

Pairwise alignment

- How to find a relationship between a query sequence & one or more additional sequences?
  - correspondence between 2 aligned sequences
    - similarity score
    - graphical views: dot plots, alignments, motifs, patterns
  - candidate homologs evaluated using statistics

- 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 (gaps)
  - a method for optimization of the total score
  - a method for evaluating the significance of the alignment

Scoring systems

- The degree of match between two letters can be represented in a matrix and 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, profiles, HMMs, etc.
  - Score has 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
  - Scores historically derived empirically from an evolutionary model describing expected evolutionary change by point mutations (scoring gaps not accommodated in the models)
- Example: PAM* = unit of evolutionary distance between 2 sequences
  - 1 PAM = 1 accepted point mutation event/100 residues
  - 200 PAM = 200 point mutations/100 residues
  - JTT, updated version of PAM used in many phylogenetic programs

BLOSUM (BLOcks SUbstitution Matrices)

(Henikoff & Henikoff, "Amino acid substitution matrices from protein blocks," PNAS USA 89:10915-19, 1992)

- generated from data from multiple aligned sequence segments from families, clustered at various similarity thresholds & corrected to avoid sampling bias
- statistical rather than an evolutionary model (can't use directly in phylogenetic inference)
- derived using highly conserved segments from divergent proteins rather than using the most mutable positions among highly similar proteins (PAM matrices)
- for a given matrix, sequences identical at >X% eliminated to avoid bias due to over-representation, e.g., Blosum 62 reflects observed substitutions between segments <62% identical
- for short sequences (9-21 residues), you have to use a different matrix such as PAM30

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

Statistical Significance

- For DB searching, the ONLY criteria available to judge the likelihood of an evolutionary relationship between 2 sequences is an estimate of statistical significance
  - to determine if the alignment score has statistical meaning, compare it with the score generated from the alignment of random sequences
  - a model of random sequences is needed
    - simplest: choose the amino acid residues in a sequence independently, with background probabilities
    - comparing a query sequence to a set of random sequences of uniform length results in scores that obey an extreme value distribution rather than a normal distribution, the use of which can lead to overestimation of an alignment's significance

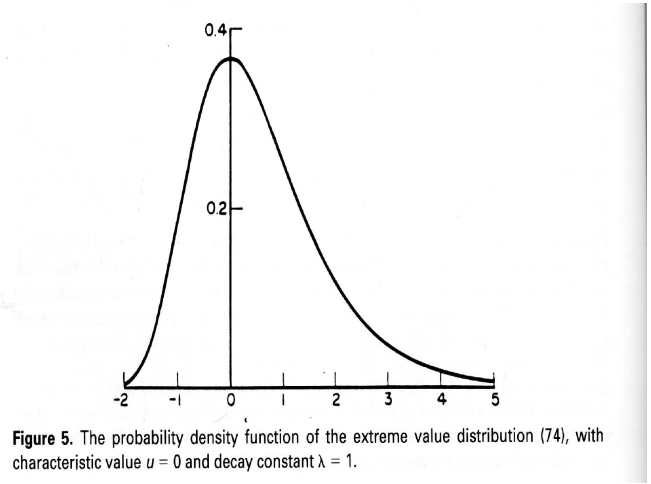


Figure 5. The probability density function of the extreme value distribution (74), with characteristic value $u = 0$ and decay constant $\lambda = 1$.

(Altschul et al. "Issues in searching molecular sequence databases," Nat Gen 6: 119-129, 1994)

Finding putative homologs: Database searching

- BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=Blastdocs)
  - Major heuristic algorithms speed up searching but compromise sensitivity (slightly)
  - Current version (Gapped Blast) incorporates gaps
  - Generates alignments & estimates of statistical significance
  - Different scoring matrices/other parameters provide tuning
  - As data increases, variations enable more focused searches

blastp protein vs protein

blastn DNA vs DNA

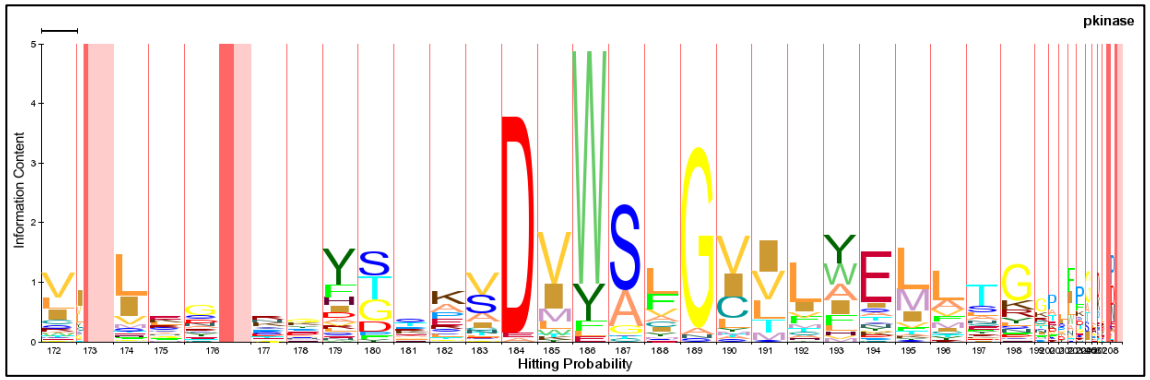
blastx 6-frame conceptual translation products of a nucleotide query (both strands) against a protein database

tblastn protein query against a nucleotide sequence database dynamically translated in all six reading frames (both strands)

tblastx 6-frame translations of a nucleotide query sequence against the 6-frame translations of a nucleotide sequence database

Extending our reach: Probabilistic models

- Psi-BLAST: Altschul et al, Nuc Ac Res, 17:3389 (1997)
  - Generalizes BLAST algorithm to use a position-specific score matrix (PSSM) in place of a query sequence & associated substitution matrix for searching the databases
  - Position-specific score matrix generated from the output of a gapped Blast search, *i.e.,* profile defined in the initial Blast search in place of a single query sequence
  - Results in a database search tuned to the specific sequence characteristics of interest
- Hidden Markov models (HMMs): Eddy, Bioinform, 14:755 (1998)
  - Probabilistic models now broadly used for database searching, classification of new sequences to known families, many other applications



  - HMM Logo (Schuster-Böckler et al, BMC Bioinfor, 5:7 (2004)

- Probabilistic methods have become the primary approach for identification of divergent homologs & classification of new sequences into superfamilies, families

Pre-computed solutions at UniPro & elsewhere

› Pfam: Multiple sequence alignments (MSAs) and HMMs for many protein domains http://pfam.sanger.ac.uk/

› Tigerfams: Families defined by MSAs & HMMs for classification,mapped to Gene Propeties http://www.tigr.org/TIGRFAMs/

› Superfamily: structural and functional protein annotations for >2400 completely sequenced genomes

– covering all proteins of known structure (SCOP superfamilies) using HMMs http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/

Motif analysis (sequence)

- Many variants of motif searching [in proteins] available
  - Consensus-based, e.g., Prosite http://prosite.expasy.org/
  - Manually annotated motifs, distant relationships, e.g., PRINTS http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/
  - Statistical, e.g., MEME http://meme.nbcr.net/meme/
- Motif-finding algorithms developed for finding transcription factor binding sites, RNA motifs, other problems related to finding short sequence signals
- Applications
  - Identification of functionally important residues
  - Identification of distant homologs
  - Anchoring multiple alignments

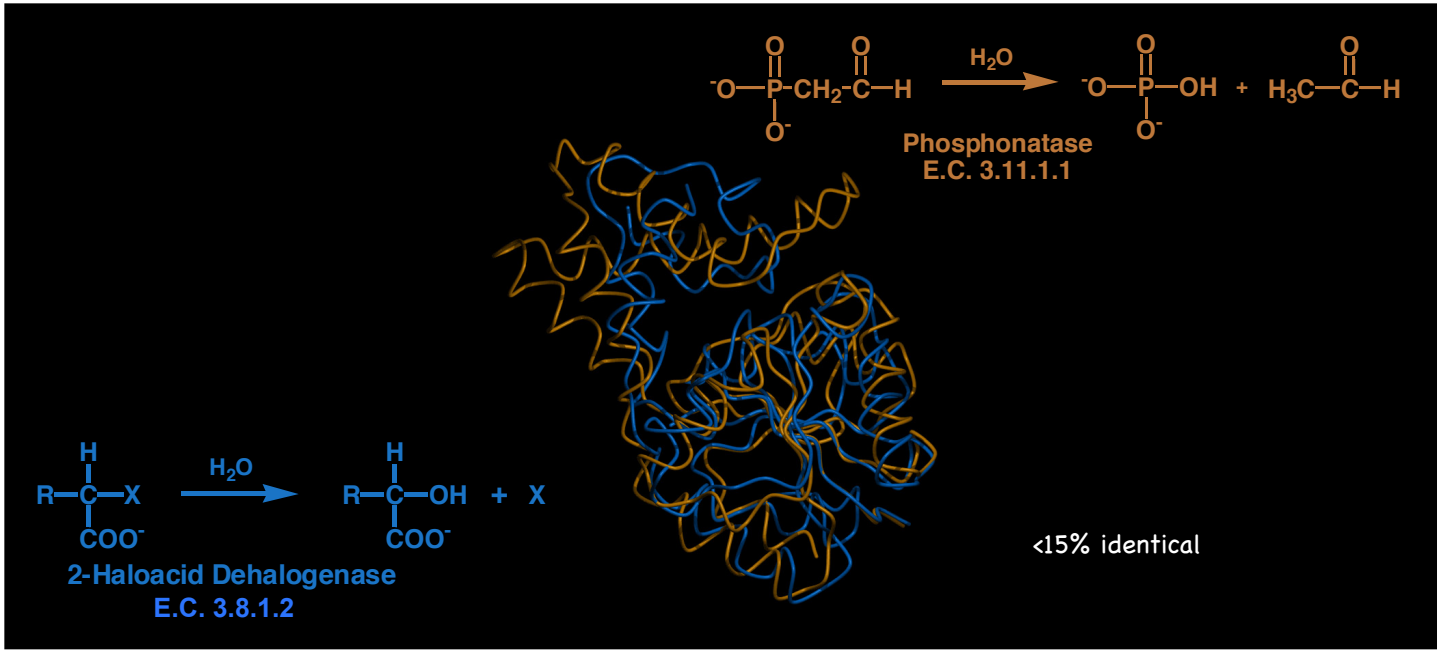
Protein Structure Comparison

- Requirements are similar to those for sequence comparisons
  - a score for insertions/deletions
  - a method for optimization of the total score
  - a method for evaluating the significance of the alignment
  - mismatches tend to be ignored & most algorithms report a score over the # of Cα that can be aligned
- Structure-Structure comparison
  - structural alignment
  - fold-based comparison
  - “fine structure” comparison (3D motifs)
- Sequence-Structure comparison
  - threading
  - active site profiling

Structure comparison

We can “see” further with structure than we can with sequence

(Chothia & Lesk, Embo J, 5:823 (1986)



- BLASTP, BLAST 2seq find no statistically significant similarities
- Using phosphonatase as a query, Delta BLAST finds
  - Swiss-Prot: 2-haloacid dehalogenase as the 106th hit in a search of Swiss-Prot, E-value = 2 x10^{-28}, 17% ID
  - Non-redundant Genbank: 0 hits against non-redundant Genbank

Structure alignment

- Focused on superposing 2 or more structures
  - Major methods produce pairwise alignments based on fold
  - Specialized methods to align 2 structures around active sites, other features (Ex: UCSF Chimera)
- Problem: Find the best way to superimpose 2 structures in 3D
- Requirements
  - an alignment providing a list of correspondences between residues in 2 or more proteins
  - a measure of structural [dis]similarity
    - RMSD (root mean square deviation)
    - intramolecular distance
    - orientation in 2° structure, overall 2° content
- Methods are heuristic
  - primary methods don't generate optimal solutions
- **Given current methods, structural alignment has nothing to do with evolutionary history**

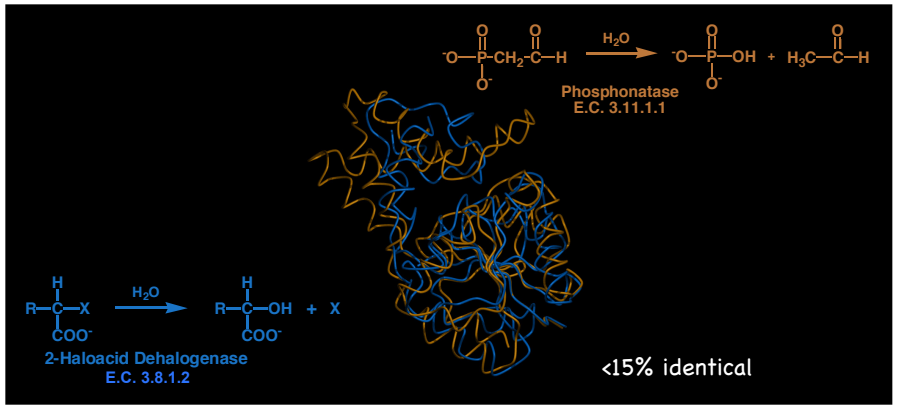
Fold-based comparisons

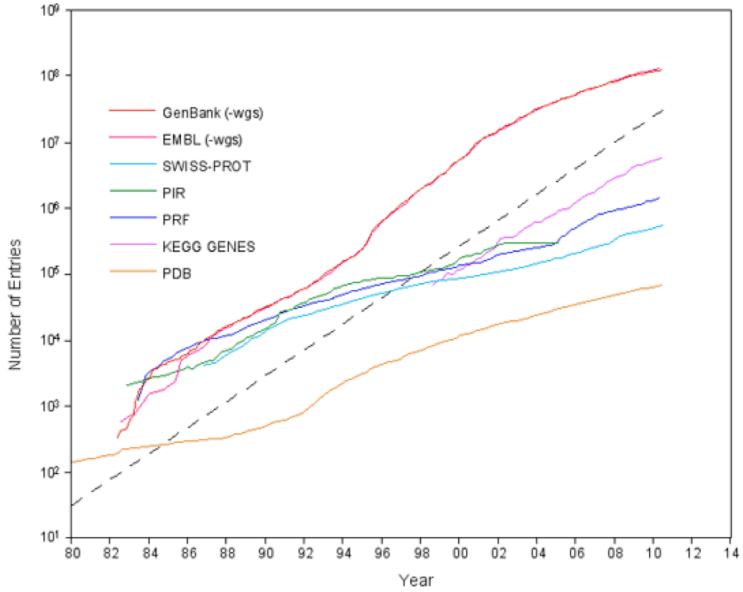
- Major methods produce pairwise alignments
  - Dali (Distance Alignment Matrix) (Holm & Sander)
    - breaks the input structures into hexapeptide fragments and calculates a distance matrix by evaluating the contact patterns between successive fragments
    - used to generate the FSSP database (families of structurally similar proteins)
  - SSAP (Sequential Structure Alignment Program) (Orengo)
    - using β-carbons, inter-residue distance vectors used to generate a series of matrices for all pairs & optimal local alignments calculated from these, then summed to produce overall alignment
    - used to construct the CATH protein structure classification database
  - VAST (Vector Alignment Search Tool) (Bryant - NCBI)
    - represents structures as a set of vectors representing secondary structural elements (SSEs)
    - topology inferred from type, directionality & connectivity of SSEs
  - TM-Align (Zhang & Skolnick), FAST (Zhu & Weng)
    - fast methods producing good alignments
- None of these methods give the same answer, even when proteins are relatively similar
- See: "Advances & pitfalls of protein structural alignment" Hasegawa & Holm, Curr Op Struct Biol 19:341 (2009)

Classification of the [protein] structural universe

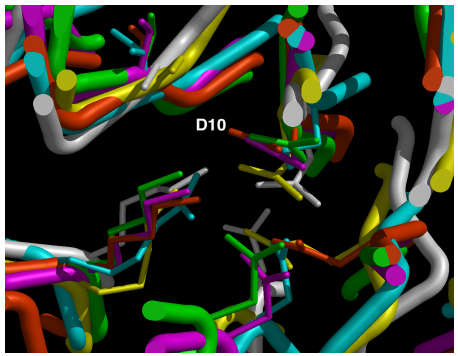
- Major resources for hierarchical classification
  - Structural Classification of Proteins (SCOP) (http://scop.mrc-lmb.cam.ac.uk/scop/)
    - manual classification
    - structure & sequence
    - superfamilies classified using inference of homology & similar function
    - primary gold standard for structural comparison & algorithm development
  - SUPERFAMILY (http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/)
    - protein domain assignments at SCOP superfamily, family levels
    - generated using expert curated profile HMMs
    - superfamilies defined based on structural, functional, evolutionary data
  - CATH protein structure classification (http://www.cathdb)
    - major levels: Class(C), Architecture(A), Topology(T) & Homologous superfamily (H)
    - semi-automated
    - structure & sequence
  - Dali Fold Index - FSSP (families of structurally similar proteins)

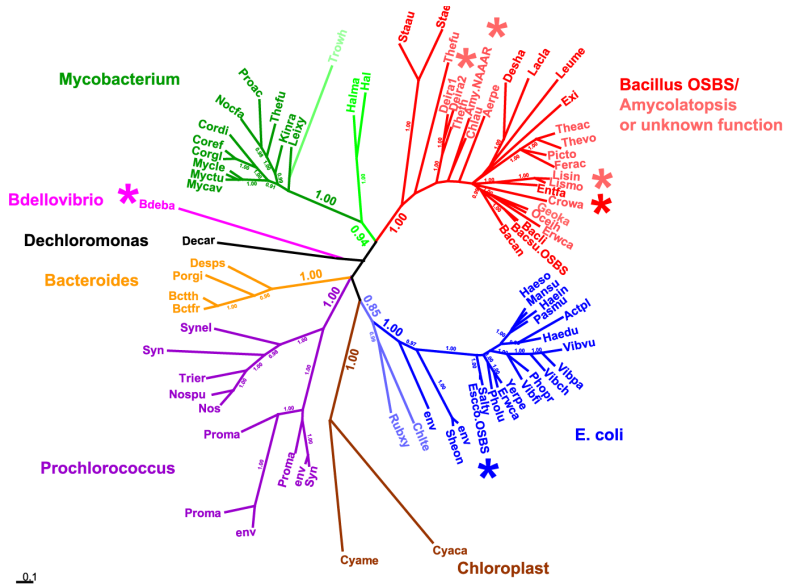
Why use sequence information when structure works better?





10	151	180
F D KTGT	K A	GDGIN D
F D KTGT	K V	GDGIN D
S D KTGT	K H	GDGTN D
F D SAGT	K R	GDGAN D
F D VDST	K V	GDGAT D
F D LDGT	K P	EDAPV G
F D VDGT	K P	EDAPA G
L D WAGT	R P	DDTWP G
L D WAGT	R P	DDAAP G
F D WAGT	R P	GDTVS D
F D LDGT	K P	GDSEV D
L D MDGV	K P	GDTVD D
F D LDGV	K P	EDSQA G
F D LYGT	K P	SSNAW D
F D LDGV	K P	DDIGA N
T D IEGT	K R	SDIHQ E



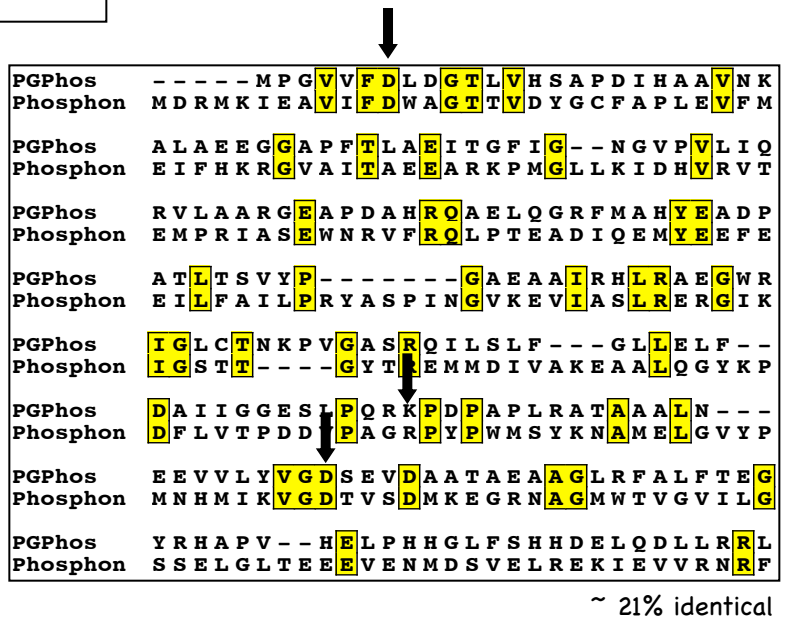


2. Clustering

Using multiple alignments for evaluation of sequence & structure relationships

- Multiple alignments provide more information than pairwise alignments
  - Screening for membership in a family/superfamily
  - Identification of conserved elements important to function
  - Determination of the level and sites of variability across the members of subgroups/families/ superfamilies
  - “Speciation” over multiple alignment space helps to connect and confirm widely degenerate motifs
  - Issues & caveats, especially for divergent sequences (<30% ID)
    - 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

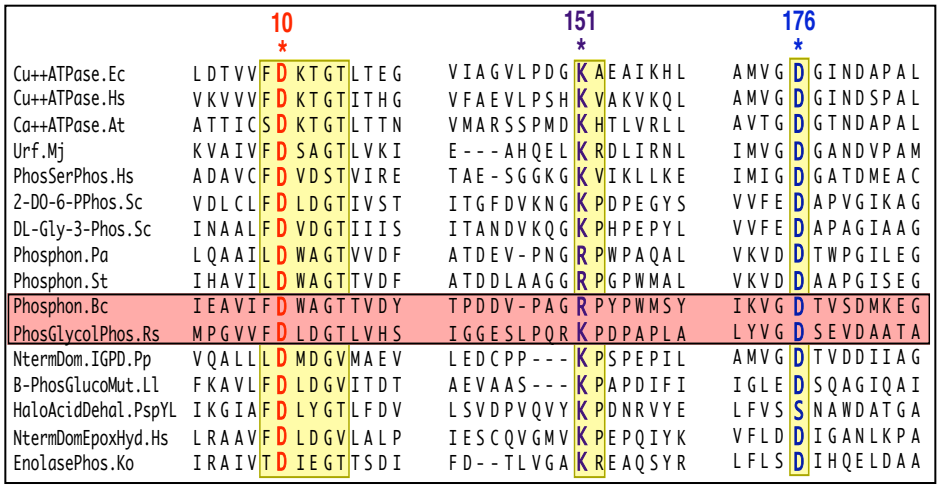
```
BLASTP 2.0a19MP-WashU

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

Database:  swissprot

                                                                  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
```



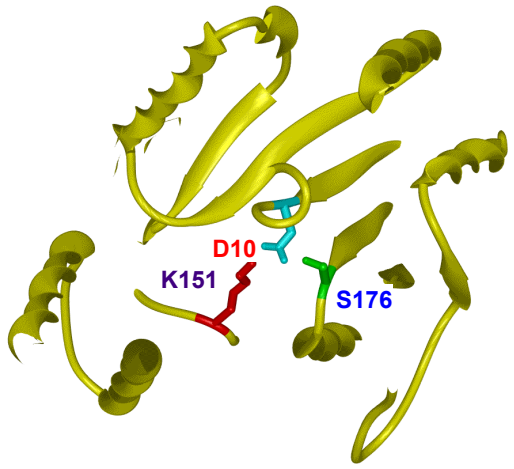
```
PGPhos    -----MPGVVFDLDGTLVHSAPDIHAAVNK
Phosphon  MDRMKIEAVIFDWAGTTVDYGCFAPLEVFM

PGPhos    ALAEEGGAPFTLAEITGFIG--NGVPVLIQ
Phosphon  EIFHKRGVAITAEEARKPMGLLKIDHVRVT

PGPhos    RVLAARGEAPDAHRQAELQGRFMAHYEADP
Phosphon  EMPRIASEWNRVFRQLPTEADIQEMYEEFE

PGPhos    ATLTSVYP-------GAEAAIRHLRAEGWR
Phosphon  EILFAILPRYASPINGVKEVIASLRERGIK

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

PGPhos    DAIIGGESLPQRKPDPAPLRATAAALN---
Phosphon  DFLVTPDDYPAGRPYPWMSYKNAMELGVYP

PGPhos    EEVVLYVGDSEVDAATAEAAGLRFALFTEG
Phosphon  MNHMIKVGDTVSDMKEGRNAGMWTVGVILG

PGPhos    YRHAPV--HELPHHGLFSHHDELQDLLRRL
Phosphon  SSELGLTEEEVENMDSVELREKIEVVRNRF
```

~ 21% identical

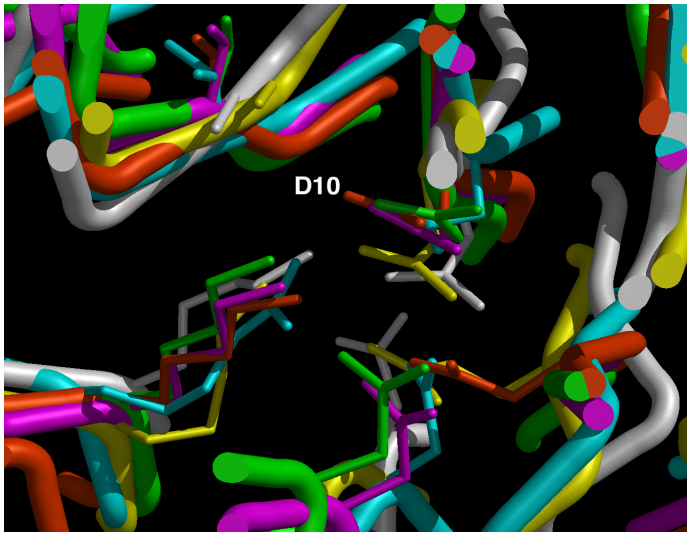
```
                    10                  151                 176
                     *                    *                   *
Cu++ATPase.Ec     LDTVVFDKTGTLTEG   VIAGVLPDGKAEAIKHL   AMVGDGINDAPAL
Cu++ATPase.Hs     VKVVVFDKTGTITHG   VFAEVLPSHKVAKVKQL   AMVGDGINDSPAL
Ca++ATPase.At     ATTICSDKTGTLTTN   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
```



	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

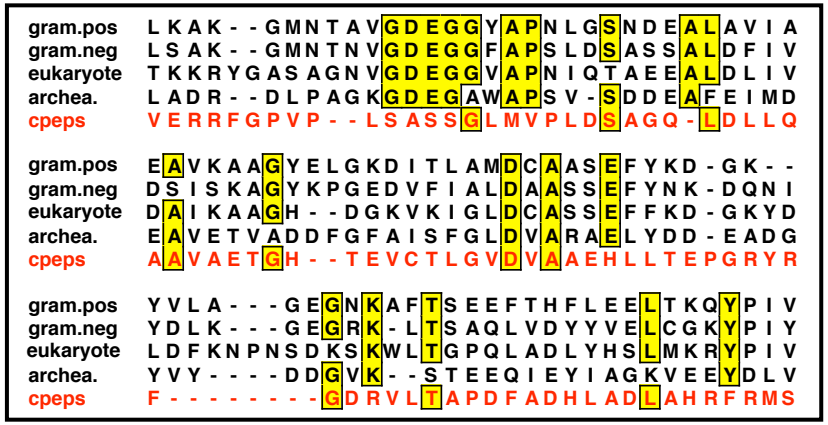


	10	151	176
	*	*	*
Cu++ATPase.Ec	LDTVVFDKTGTLTEG	VIAGVLPDGKAEAIKHL	AMVGDGINDAPAL
Cu++ATPase.Hs	VKVVVFDKTGTITHG	VFAEVLPSHKVAKVKQL	AMVGDGINDSPAL
Ca++ATPase.At	ATTICSDKTGTLTTN	VMARSSPMDKHTLVRLL	AVTGDGTNDAPAL
Urf.Mj	KVAIVFDSAGTLVKI	E---AHQELKRDLIRNL	IMVGDGANDVPAM
PhosSerPhos.Hs	ADAVCFDVDSTVIRE	TAE-SGGKGKVIKLLKE	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



Other issues for multiple alignments

- Statistics for multiple alignments not as well developed compared to the theory for pairwise alignments
  - Algorithmic issues
    - computational complexity: a true multiple alignment of N sequences requires an N-dimensional matrix
    - no single “correct” multiple alignment can be achieved except in trivial cases
    - all methods assume sequences are independent rather than related by a phylogenetic tree in which different branches evolve at different rates (sensitivity to the # of sequences used
    - sensitivity to the specific sequences used for a multiple alignment
      - filtering to remove redundancy difficult to do without losing information



all of these sequences = 35-50% identical

filtering w 35% ID cut-off requires us to choose one sequence to represent the group

using a less stringent filtering cutoff to include more sequences produces a different alignment

Multiple Alignment Methods

Profile-based progressive alignment: early errors propagate throughout

- Thompson, JD et al., "ClustalW": Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," NAR 22:4673, 1994
- Sievers, F et al., "Clustal Omega": Fast, scalable generation of high-quality protein MSAs," Mol Sys Biol 7:539, 2011 (improved guide tree, final MSA computed using profile-profile alignment)

Consistency-based methods: pair-wise alignments improved via comparison to an intermediate sequence; allows incorporation of heterogeneous data

- Notredame et al, "T-Coffee: A novel method for fast and accurate multiple sequence alignment," JMB 302: 205, 2000

Progressive iterative: re-align groups of sequences within an MSA

- Edgar, RC "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," BMC Bioinformatics 19:113, 2004

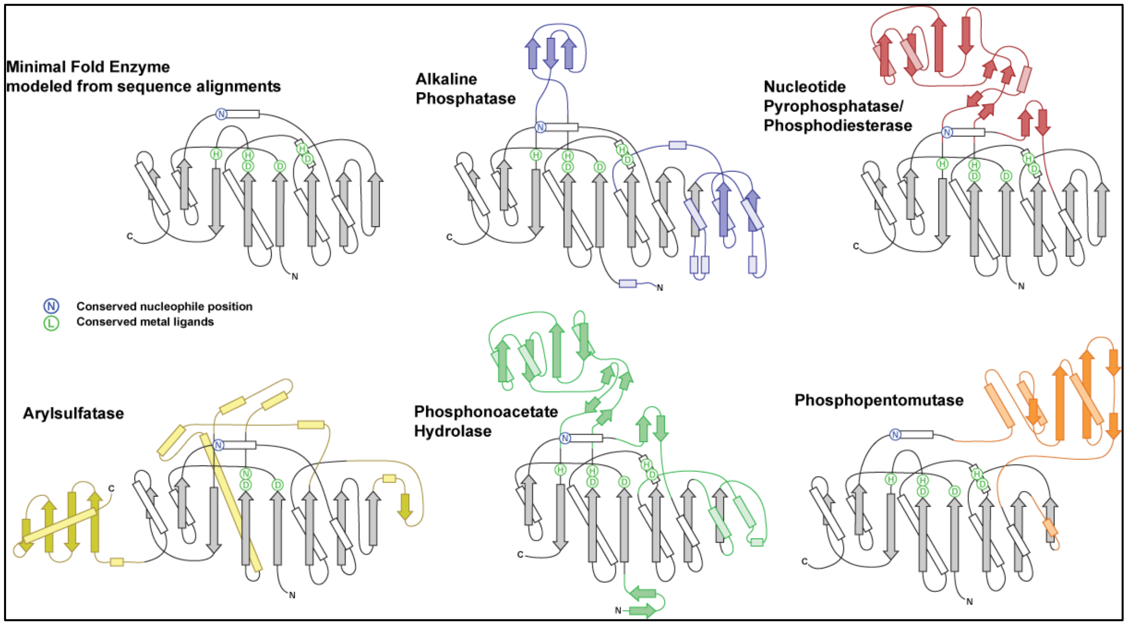
Template-based: adds information from structural/other template type

- O'Sullivan et al, "3DCoffee: combining protein sequences & structures within MSAs," JMB 340:385, 2004
- Pei et al, "PROMALS3D: a tool for multiple protein sequence & structure alignments," NAR 36:2295, 2008

Others of note:

- Sadreyev et al, "COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance," JMB 326:317, 2003
- Katoh et al, "MAFFT: a novel methods for rapid multiple sequence alignment based on fast Fourier transform," NAR 30: 3059, 2002

Structure-guided multiple alignment is also important

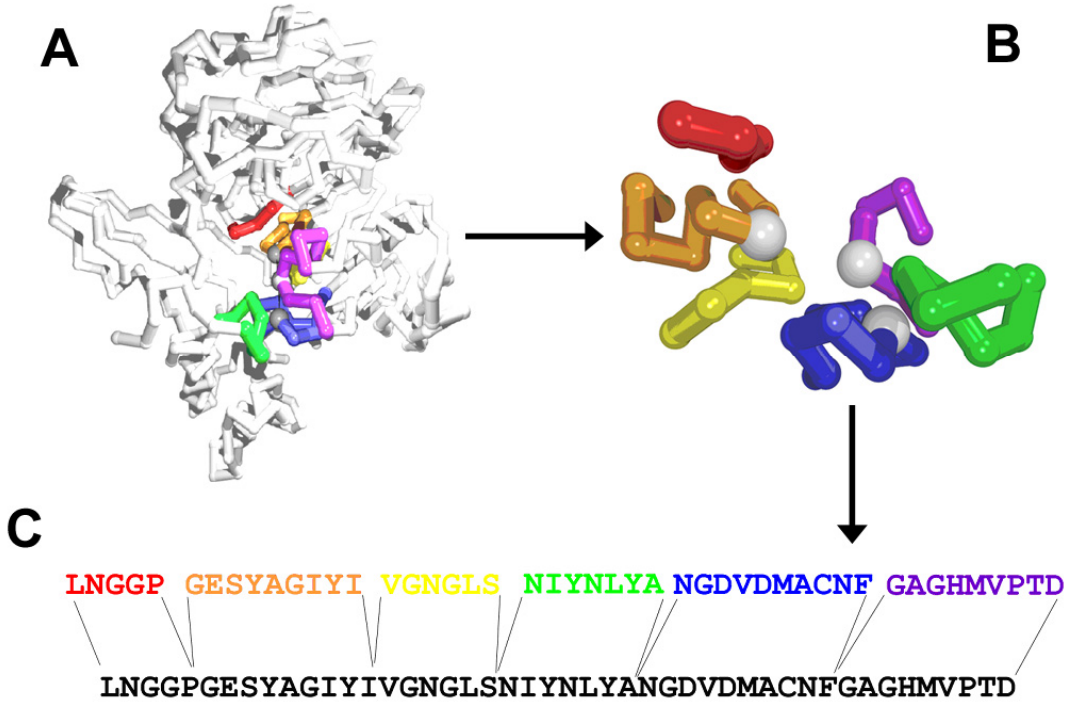


			L L		N		L L		L L		L		
Minimal scaffold enzyme	Q5LG14	[30]	L I A L D G	[33]	S S A I N W A	[109]	D H - - - T G H	[39]	M T A D H G	[5]	H	[61]
Acid phosphatase	2d1g	[38]	I I F Q E N	[129]	P S T P G A L	[171]	H - G G Y S - -	[28]	I I Y D D S	[21]	-	[89]
Sulfatase-like	2w8d	[35]	Y V S L E S	[41]	K T S D A E F	[111]	H - F - - P - -	[53]	M Y G D H Y	[23]	-	[153]
	2vqr	[36]	L I V V D Q	[42]	P C G P A R A	[157]	H P - - A F - -	[99]	F T S D H G	[10]	K	[177]
	1p49	[9]	L V M A D D	[37]	L C T P S R A	[211]	H T - - A L - -	[45]	F T S D Q G	[11]	-	[228]
	1fsu	[9]	L V M A D D	[37]	L C T P S R A	[211]	H T - - P L - -	[45]	F T S D Q G	[11]	-	[228]
	1hdh	[8]	V I V A D D	[35]	T C S P T R S	[156]	H W - - P L - -	[99]	F M S D N G	[47]	-	[170]
	1auk	[8]	V I V A D D	[35]	T C S P T R S	[156]	H W - - P Q - -	[99]	F M S D N G	[47]	-	[170]
NPP-like	3q3q	[44]	A I A V D Q	[37]	E T C P G H S	[207]	D Y - - - V G H	[37]	L T A D H G	[143]	H	[74]
	3t01	[27]	A V C L D G	[36]	F T N P N N L	[139]	D Y - - - V Q H	[31]	V T A D H G	[124]	H	[47]
	3nkr	[131]	I F S V D G	[35]	K T F P N L Y	[98]	D F - - - S G H	[39]	F V G D H G	[113]	H	[392]
	2gsu	[10]	L I S I D G	[33]	L T F P N H Y	[116]	D E - - - A G H	[39]	V V S D H G	[103]	H	[69]
Mutases	1ejj	[7]	L I I L D G	[47]	N S E V G H L	[337]	D M - - - V G H	[33]	I T A D H G	[15]	H	[49]
	3m8z	[13]	L V V M D S	[69]	D T M T G H W	[197]	D A - - L F G H	[32]	I T A D H G	[9]	H	[55]
	2zkt	[8]	L I I L D G	[44]	G S D T A H L	[241]	D A - - - A G H	[33]	I T G D H S	[7]	H	[56]
Alkaline phosphatase	3tg0	[38]	L L I G D G	[48]	D S A A S A T	[221]	D K - - - Q D H	[34]	V T A D H A	[40]	H	[40]

3D motifs

- "Fine structure" motifs for database searching & family classification
  - Manually generated or automated methods to capture conserved 3D signals of side-chains & functionally important regions
  - Residue conservation in sequence alignments & spatial clustering often considered
    - Active Site Profiling (Fetrow/Skolnick)
    - ASSAM (Artymiuk/Willett)
    - TESS/JESS (Thornton)
    - DRESPAT (Wangikar)

Combined approaches: Active site profiling



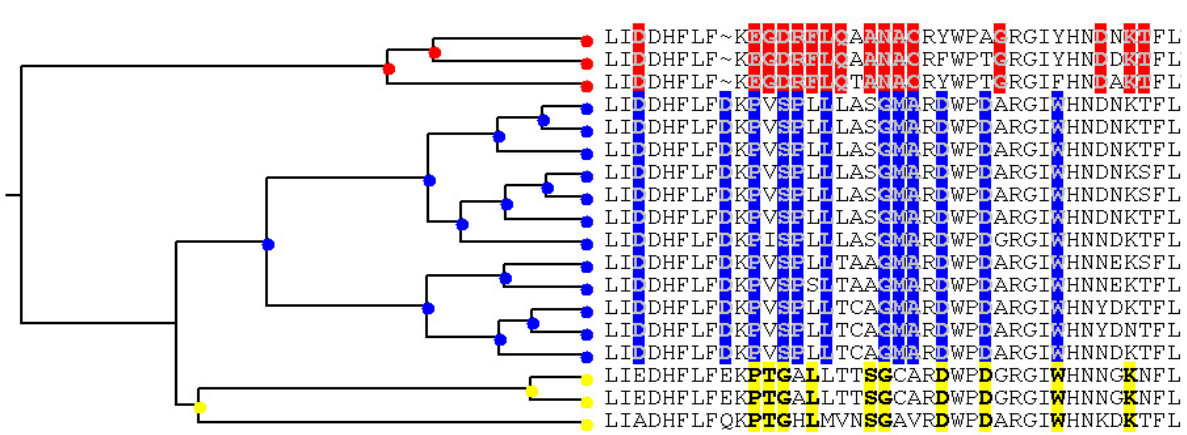
- Identify local structural features around a functional site using a set of distances between Cα
- Project motifs onto sequence to create a signature
- Identify homologs from signature(s) & align generate profile(s)
- Cluster profiles to identify functionally distinct families

Cammer, SA et al, J Mol Biol (2003) 334:387-401

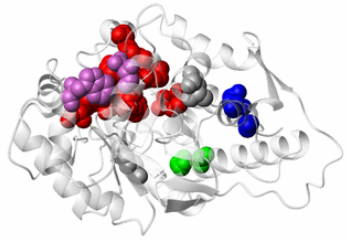
Combined approaches: Evolutionary trace

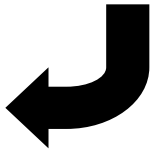
(see papers by O. Lichtarge)

- Takes advantage of the larger context provided by a family-based view of proteins to improve the accuracy of binding site determination
- Uses sequence-based clustering of related proteins to distinguish class-specific differences in ligand binding determinants across a particular family or superfamily of proteins.
- The sequence information derived from multiple alignments and phylogenetic clustering is leverage by mapping class specific patterns likely to be functionally important onto crystal structures and structural models



2. Select class conserved regions

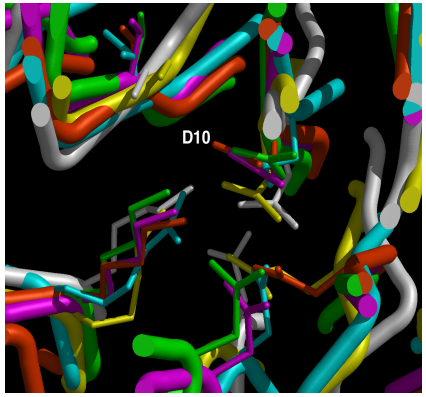


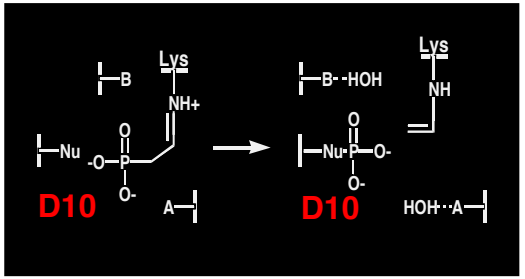


3. Map selections to structure

3. Mapping conserved elements of sequence & structure to conserved elements of function

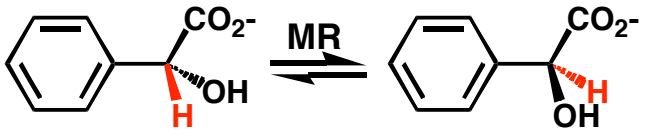
10	151	180
*	*	*
F D K T G T	K A	G D G I N D
F D K T G T	K V	G D G I N D
S D K T G T	K H	G D G T N D
F D S A G T	K R	G D G A N D
F D V D S T	K V	G D G A T D
F D L D G T	K P	E D A P V G
F D V D G T	K P	E D A P A G
L D W A G T	R P	D D T W P G
L D W A G T	R P	D D A A P G
F D W A G T	R P	G D T V S D
F D L D G T	K P	G D S E V D
L D M D G V	K P	G D T V D D
F D L D G V	K P	E D S Q A G
F D L Y G T	K P	S S N A W D
F D L D G V	K P	D D I G A N
T D I E G T	K R	S D I H Q E

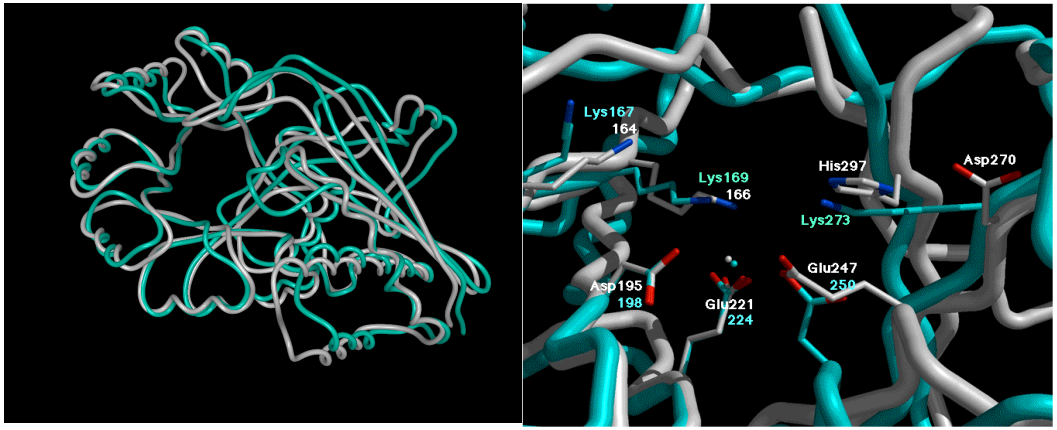




Chemistry-constrained evolution

Structural elements involved in mediating a fundamental aspect of chemistry in functionally diverse superfamilies are conserved across divergent proteins while the ability to bind new substrates & perform different overall reactions evolves

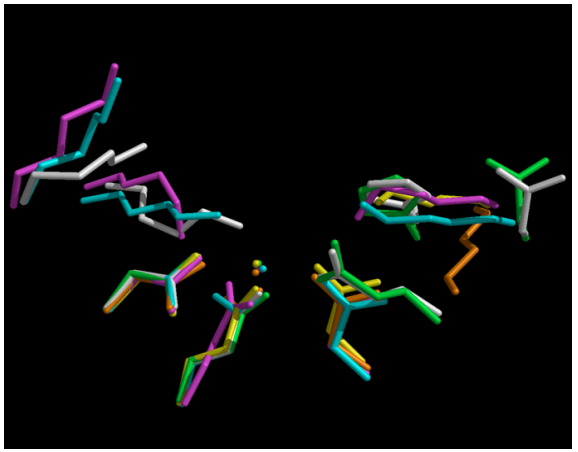


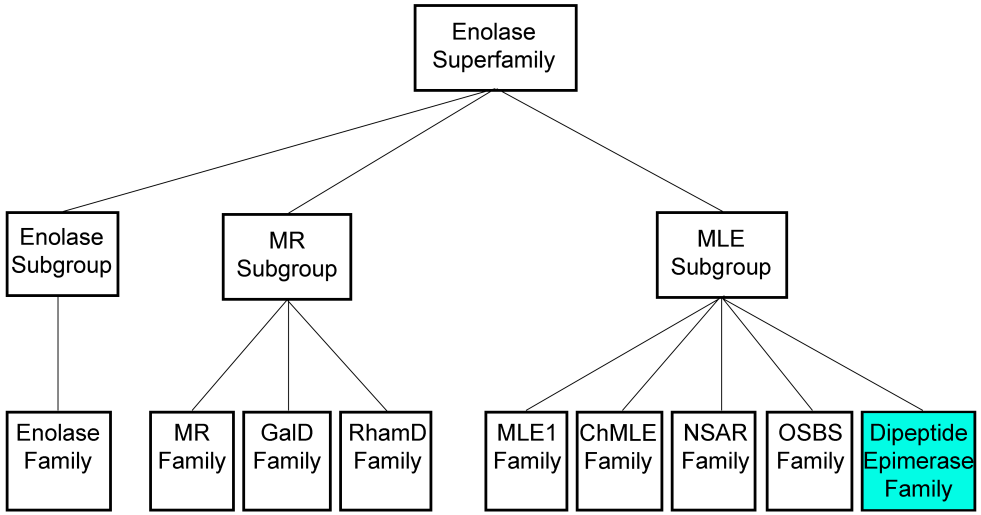


Jensen, R.A. Ann. Rev. Microbiol 30:409-425(1976)
Petsko, G. A. et al., TIBS 18: 372-376 (1993)
Babbitt & Gerlt, JBC 272: 30591 (1997)
Gerlt & Babbitt, Ann Rev Biochem 70: 209 (2001)

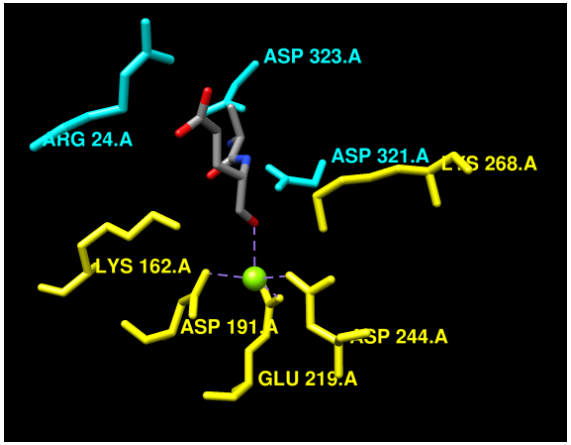
The chemistry-constrained superfamily model suggests a strategy for clustering & annotation transfer

Superfamily: Subset of active site residues conserved in all members of a superfamily catalyze a conserved partial rxn





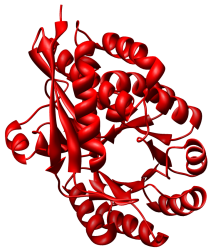
Family: Family-specific subsets of residues mediate reaction & substrate specificity



Annotation transfer only at the level of granularity at which good supporting evidence exists

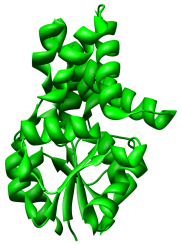
Clustering for large-scale data

Functionally diverse enzyme superfamilies represent more than a third of the enzyme superfamily universe



Enolase: 40 rxns; 27K sequences

Amidohydrolase: ~100 rxns



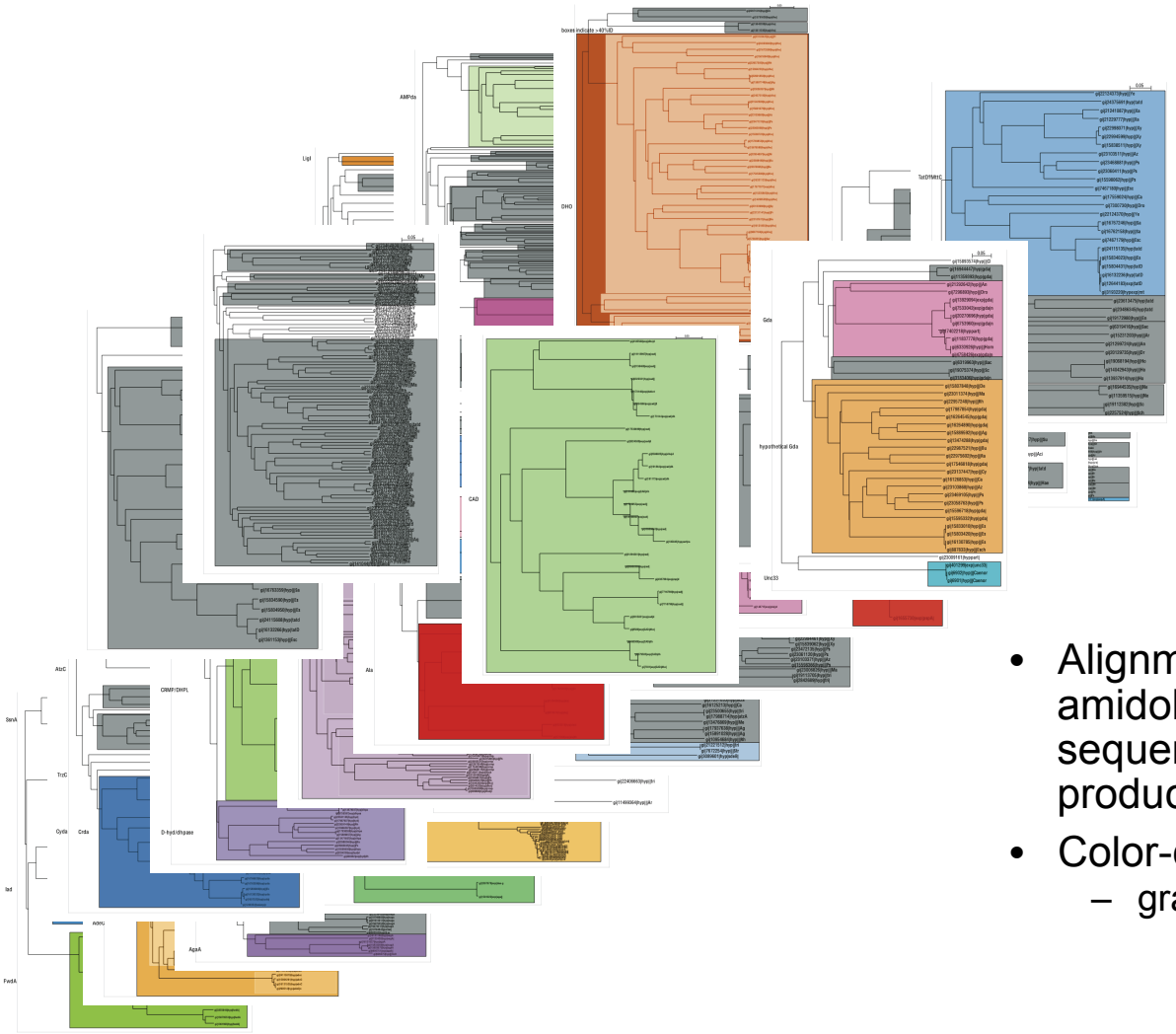
Haloalkanoic acid dehalogenase: 30-50 rxns; 80K sequences

Glutathione transferase: 44 major subgroups: ?? Rxns;13K sequences

Isoprene synthase I: ?? Rxns, >50,000 products; 16K sequences

Crotonase: 20 rxns; 48K sequences

Each superfamily links a conserved active site architecture to a fundamental catalytic capability



- Alignments for ~5,000 amidohydrolase superfamily sequences clustered at 40% identity produces 17 major subgroupss
- Color-coded by known function
  - gray = unknown functions

We needed a better way for clustering large-scale data & mapping functional properties to sequence & structure

Protein similarity networks: Large-scale analysis & visualization

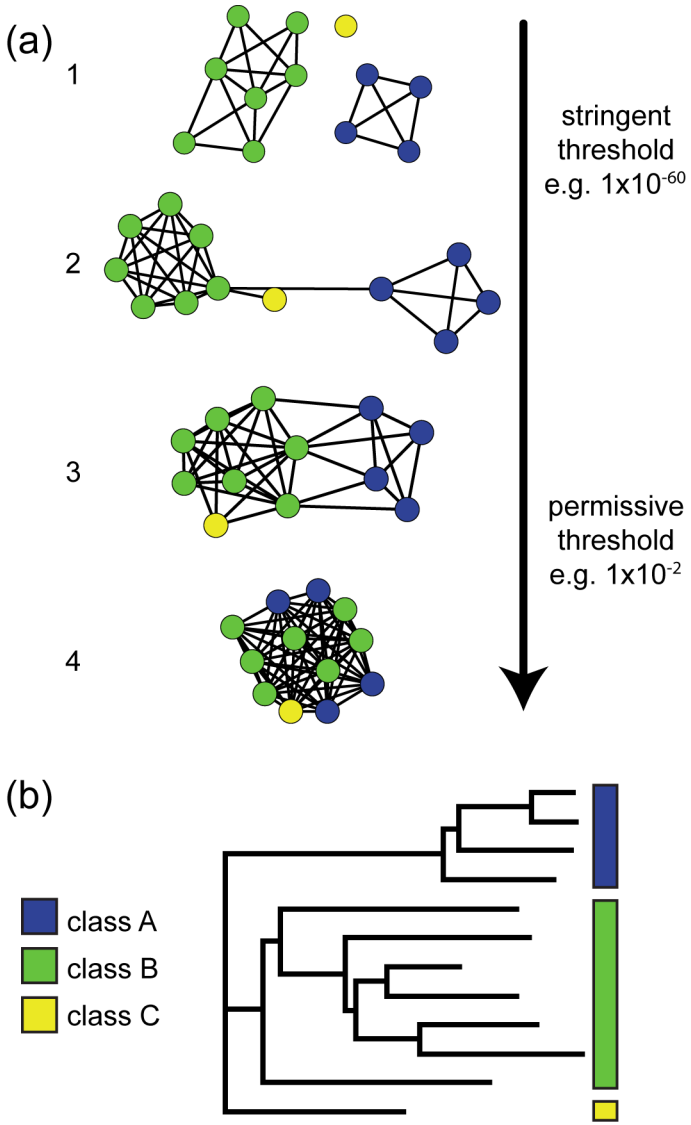
- Clustering sequence & structure data for probing relationships in functionally diverse enzyme superfamilies
- Interactive visualization of functional trends in the context of sequence or structural similarity

Networks are available for download for a small number of enzyme superfamilies our Structure-Function Linkage Database (SFLD) http://sfld.rbvi.ucsf.edu/

At least 7 new servers for generating your own networks have been published in the last couple of years

For an introduction, see Atkinson HJ & Babbitt PC. *PLoS ONE* (2009) 4: e4345

Exploring similarity at many levels of detail



- Interactive
- Fast
- Easy visualization (Cytoscape)
- Handles thousands of sequences, structures, etc.
- Pairwise comparisons don't require a multiple alignment
- Track well with known structure-function relationships
- Capture sequence, structure, ligand, relationships
- Many metrics & algorithms for comparison & clustering of similarity data
- Networks NOT a substitute for phylogenetic trees

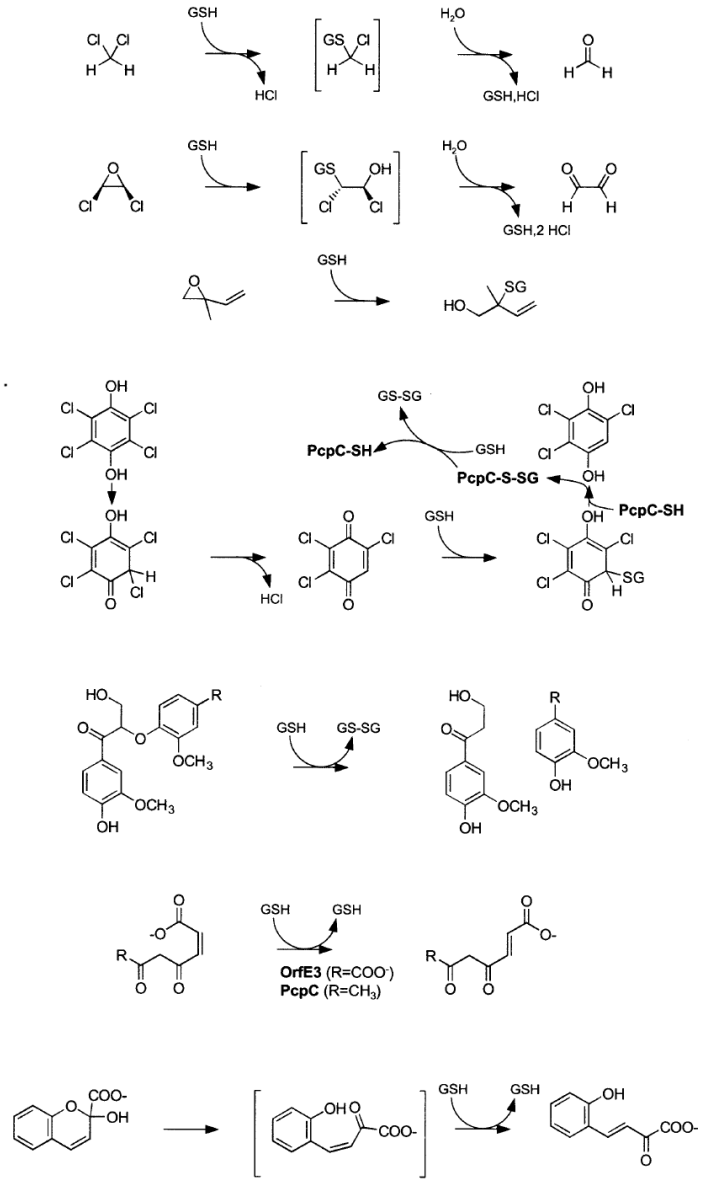
Node = sequence (or structure)

Edge = connections between sequences w scores as good as the E-value cutoff threhold

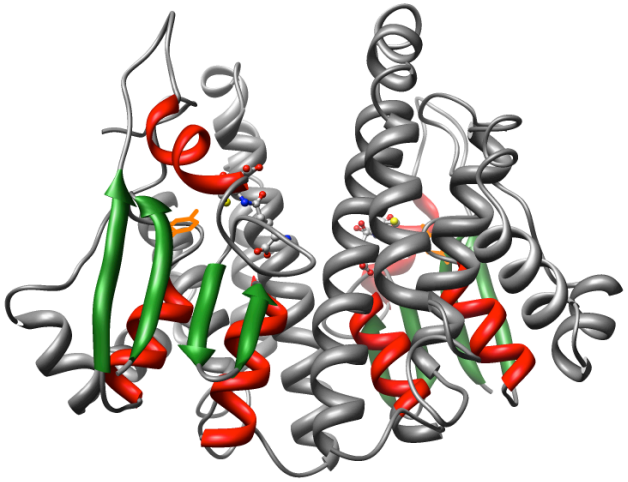
Sequence networks in this presentation: all-by-all BLAST (*E*-values as scores); structure: all-by-all FAST scores

Example: Glutathione transferases (GSTs)

>13,000 nonredundant sequences



from Vuilleumier & Pagni, Appl Microbiol Biotech 58: 138 (2002)



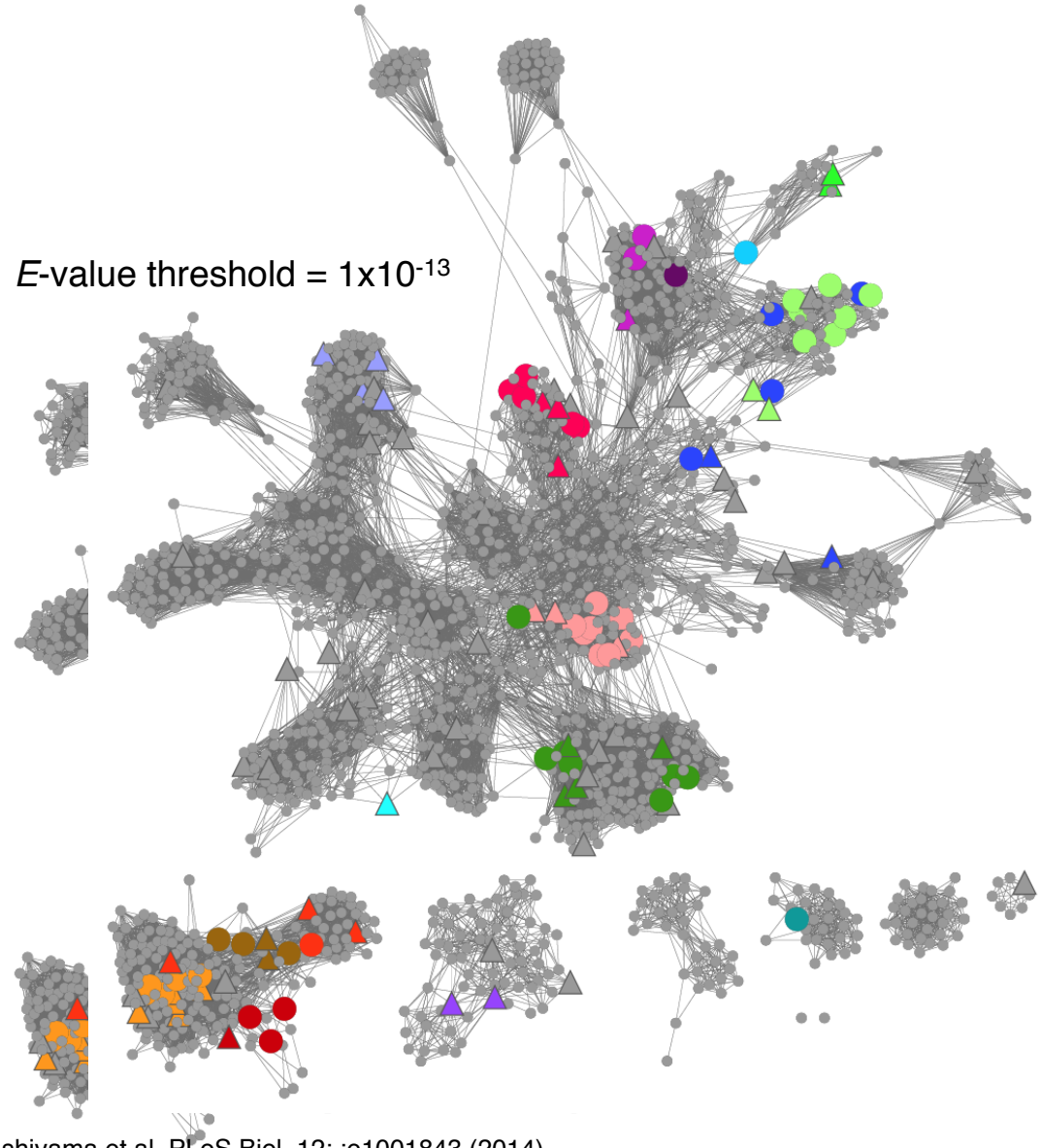
Fundamental catalytic step: nucleophilic attack on an electrophilic substrate using GSH

Structural variations across the superfamily distinguish classes within it

Sequence identity ~20-30% across the most structurally similar mammalian classes, alpha, mu, pi

Major roles in detoxification of xenobiotics, defense against oxidative stress, roles in cell signaling, biosynthesis of sex steroids & prostaglandins...

How little we know



Node color: Swiss-Prot family annotation

- Alpha
- Beta
- Omega
- Mu
- DHAR
- Phi
- Pi
- FAX
- Tau
- Sigma
- HSP26
- Theta
- Lambda
- Zeta
- chloride channel CLIC

"New" classes

- Nu
- Xi

1,568 nodes represent 13,000 sequences (50% ID filtered)

Median edge *E*-value = 4x10^{-25}

Median alignment length = 210 residues

Large nodes colored by SwissProt classification if >50% of sequences in each node belong to that class

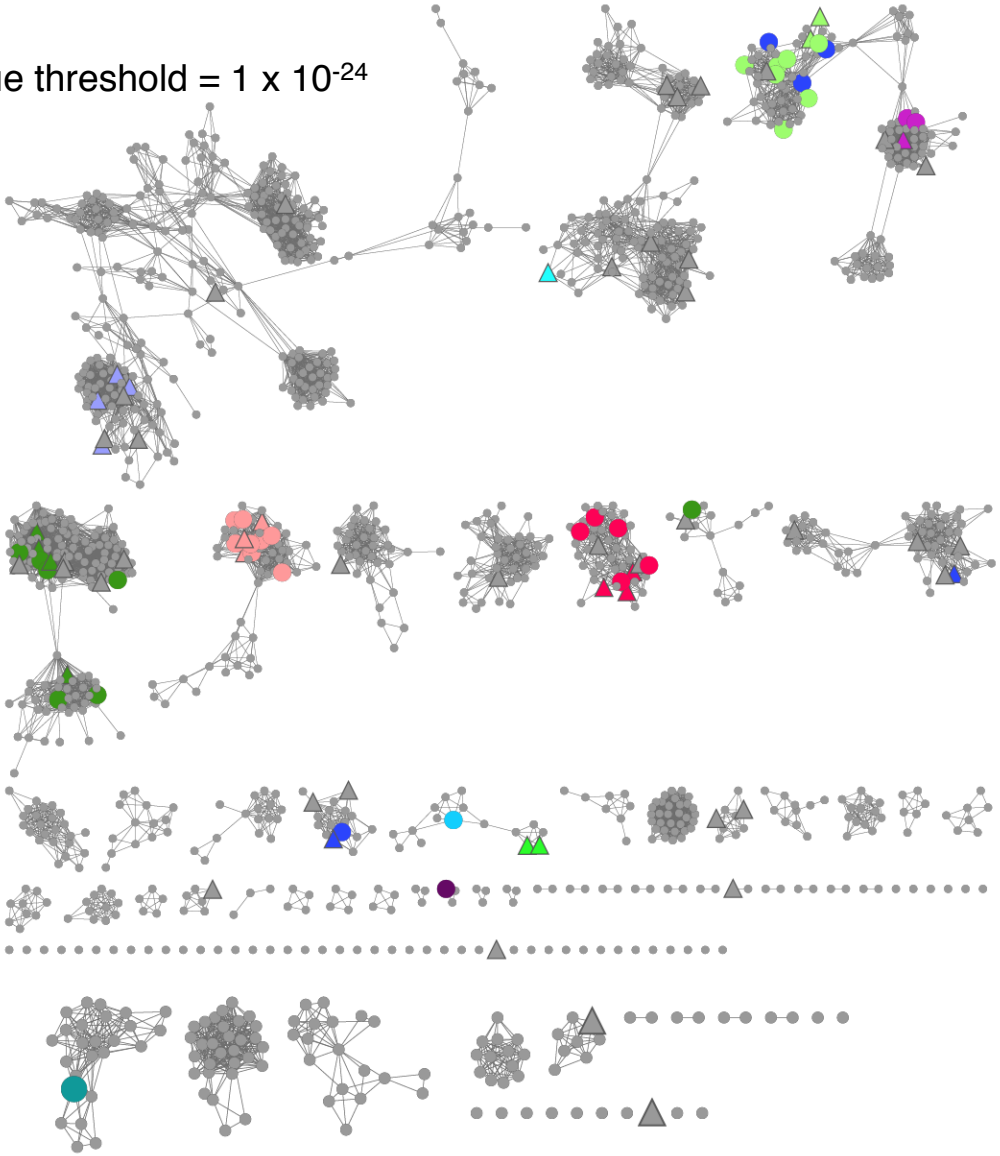
Gray: Not classified

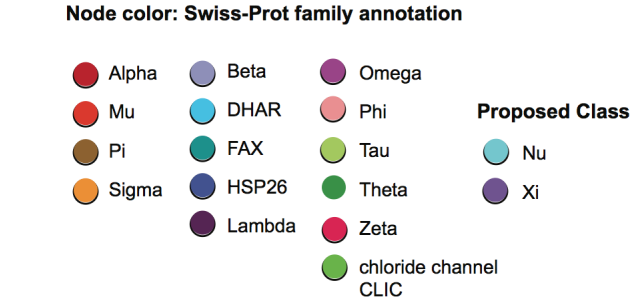
Triangle: PDB structure

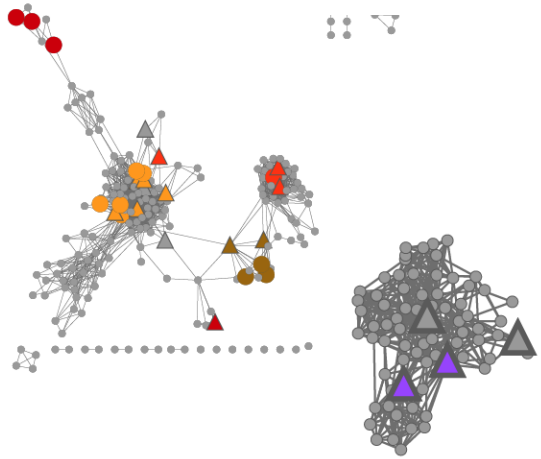
Mashiyama et al, PLoS Biol, 12: :e1001843 (2014)

Many new classes yet to be discovered

E-value threshold = 1 x 10^{-24}

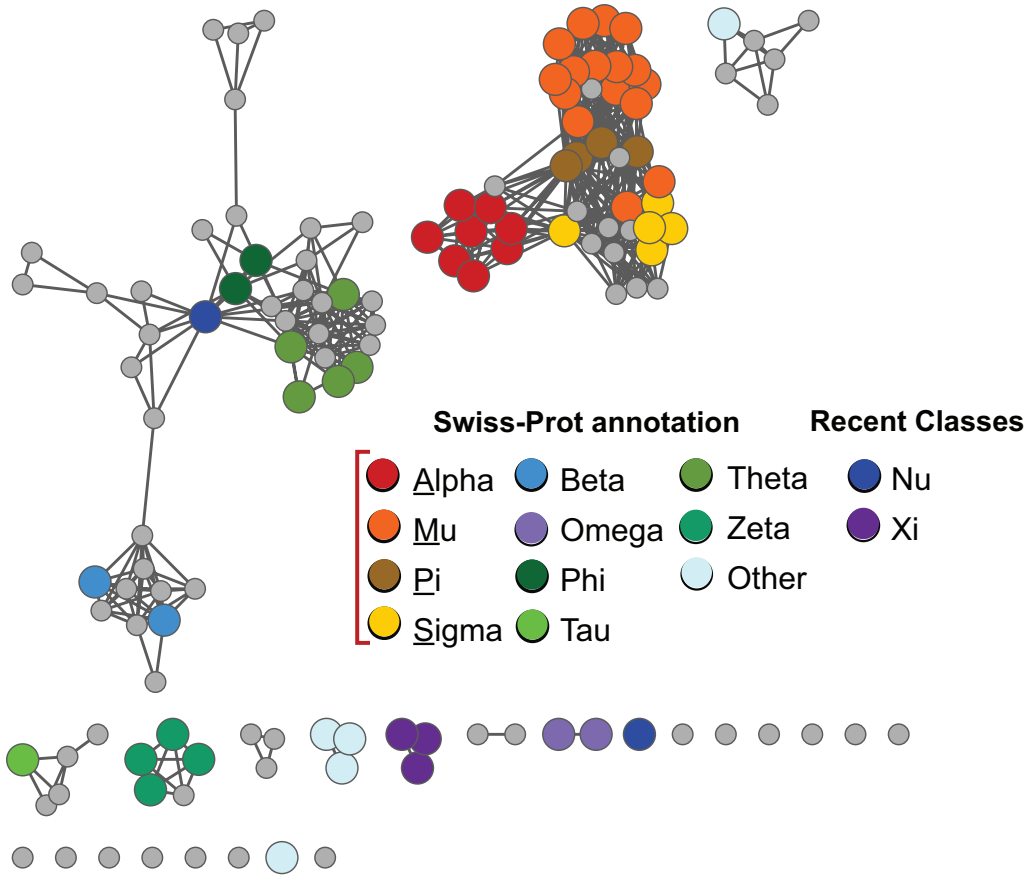






- Median *E*-value = 6 x 10^{-35}
- Median alignment length = 212

GST structure similarity networks



FAST score threshold = 20
131 structures filtered to 95%ID

Pros & Cons of Sequence Similarity Networks

– Advantages
  - Very fast to generate many networks in the same time it takes to generate one tree
  - Each sequence has multiple connections – identifying hubs, linkers not accessible in trees
  - Relative divergence captured by viewing networks across multiple E-value cutoffs
  - Simple metrics (BLAST) capture strong conservation signals from pairwise comparisons

– Caveats
  - Networks are not trees
    – no evolutionary model
    – lacks mathematical rigor of phylogenetic inference
  - Interpretation can be distorted due to multiple E-values represented in one network or between networks
  - As E-values get less significant, network connections become increasingly spurious

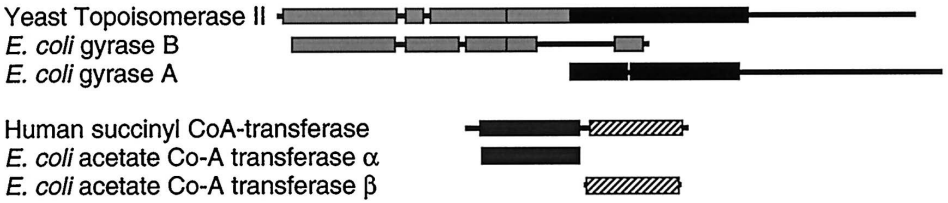
4. Orthogonal approaches

- Homology-based annotation transfer alone isn't good enough anymore
  - high levels of misannotation
  - fails to capture higher orders of biological function such as pathway membership, fusion/assembly partners, etc.
  - blind to experimental data from expression arrays, proteomics & other types of profiling
- Many types of biological data available to improve function prediction
  - Gene fusion
  - Gene order/co-localization on a genome
  - Phylogenetic profiling
  - Comparative genome/proteome analysis
  - Interaction networks

Gene fusion

- Premise: If the different domains of a composite protein (gene fusion) in one species is similar to two component proteins in another species, the component proteins are associated by function

  Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO & Eisenberg, D (1999) Science 285: 751-753

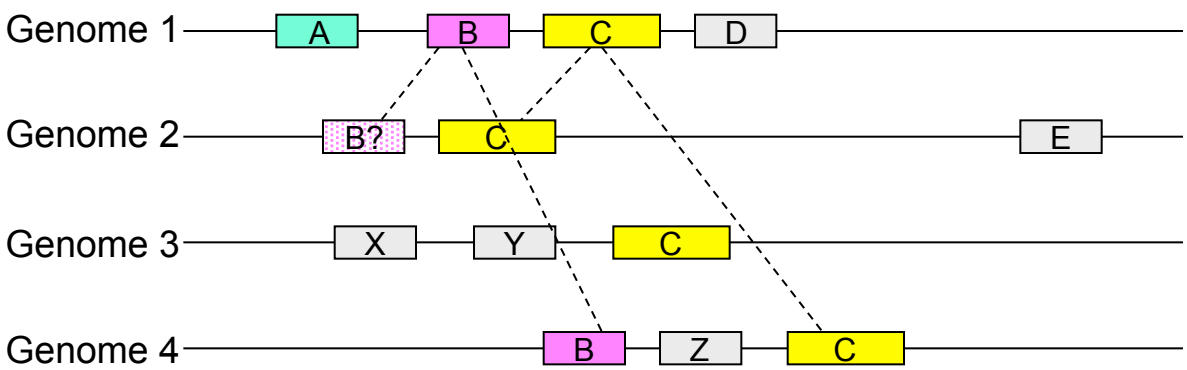


- Fusion events also used to detect protein-protein interactions on a genomic scale

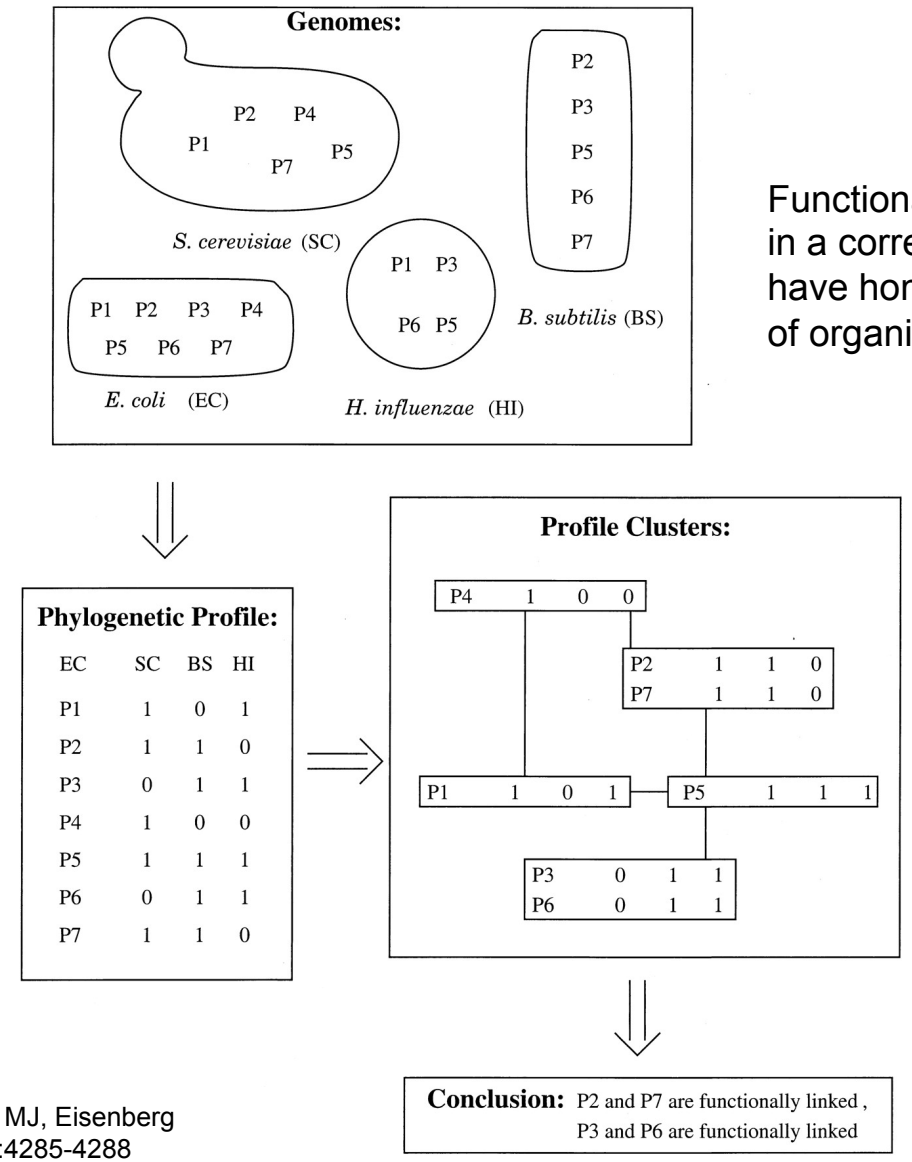
  Enright AJ, Iliopoulos I, Kyrpides NC & Ouzounis CA (1999) Nature 402: 86-90

Gene order/cluster analysis

- Functional inference from conserved gene neighbor/gene order relationships on a chromosome
- This works best in prokaryotes where operon/regulon organization may co-localize genes together on the genome



Phylogenetic Profiling

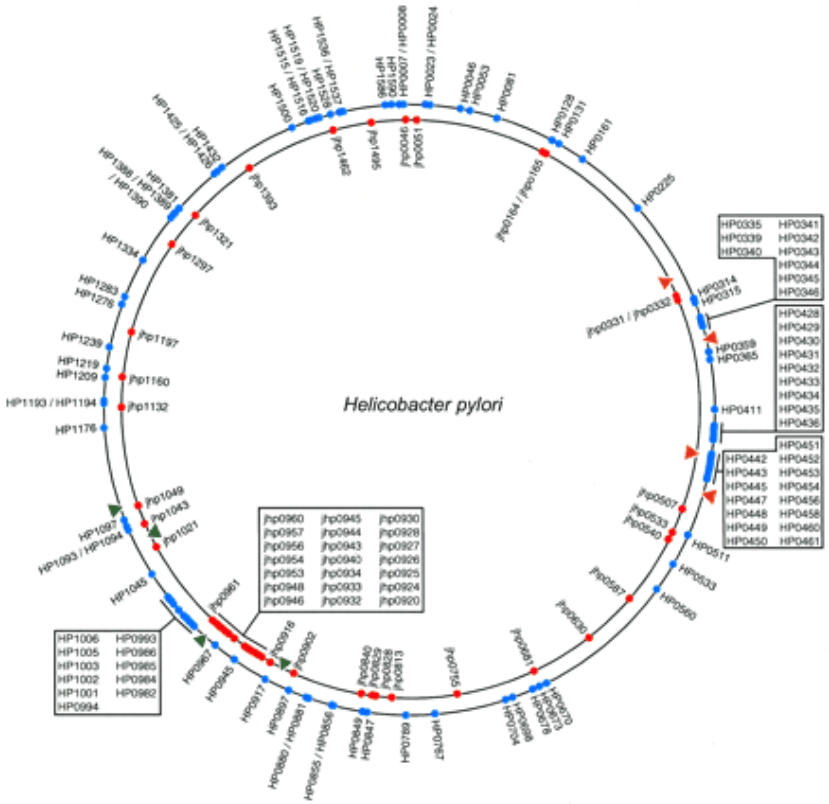


Functionally related proteins evolve in a correlated fashion & therefore have homologs in the same subset of organisms

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D & Yeates TO. (1999) PNAS USA 96:4285-4288

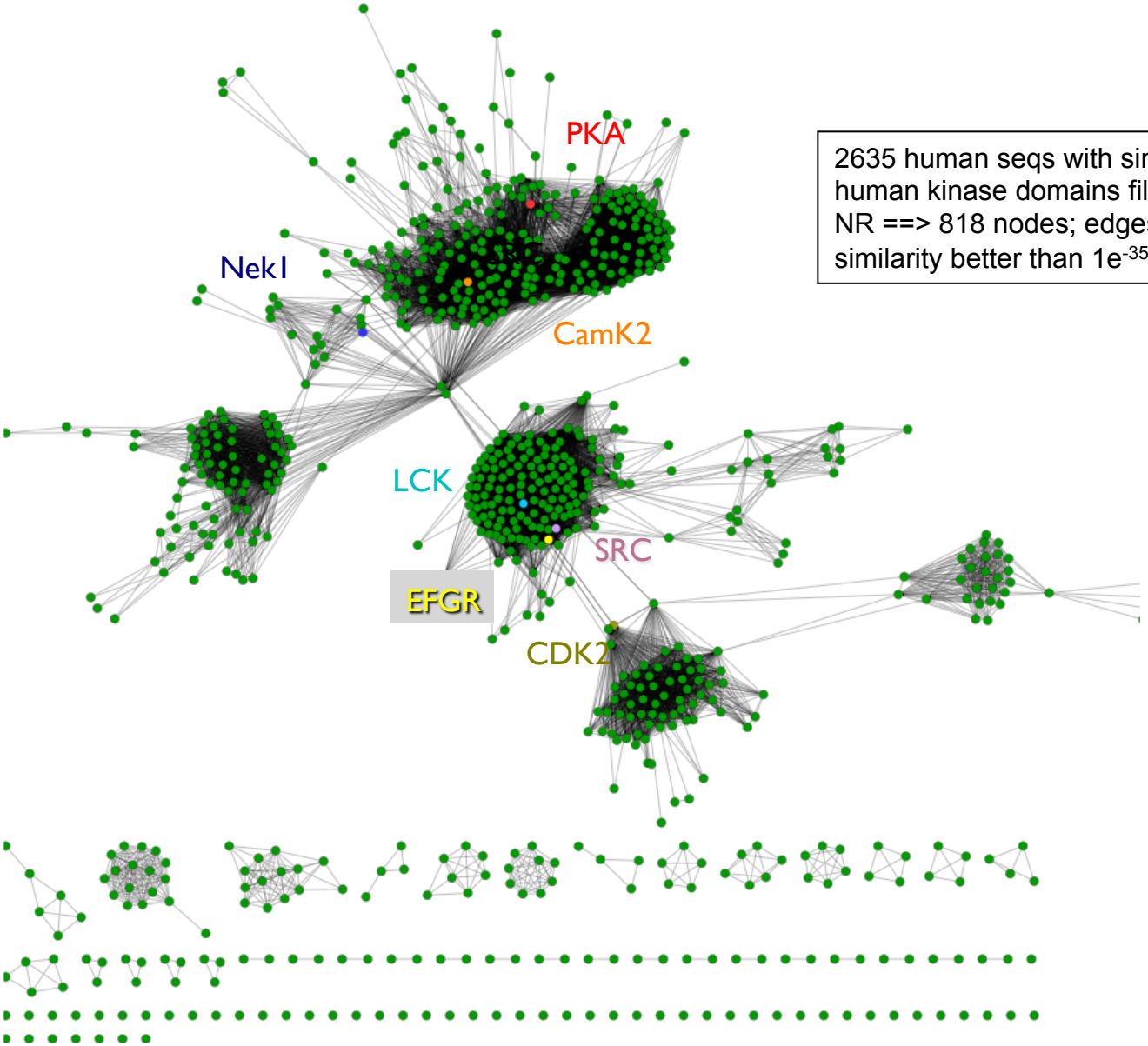
Differential genome analysis

- Analysis of genes common to several genomes vs. genes unique to a particular genome
  - Identification of genes that differ critically between related species and application of orthogonal biological knowledge, phenotypic variation can be correlated to a specific gene/set of genes
  - Analysis of classes of genes conserved across strains but altered/deleted from more divergent species, allow inference of function



Alm RA et al. (1999) Nature 397: 176-180
Huynen M, Dandekar R & Bork, P (1998) FEBS Lett. 426: 1-5
Janssen PJ, Audit B & Ouzounis CA (2001) NAR 29: 4395-404

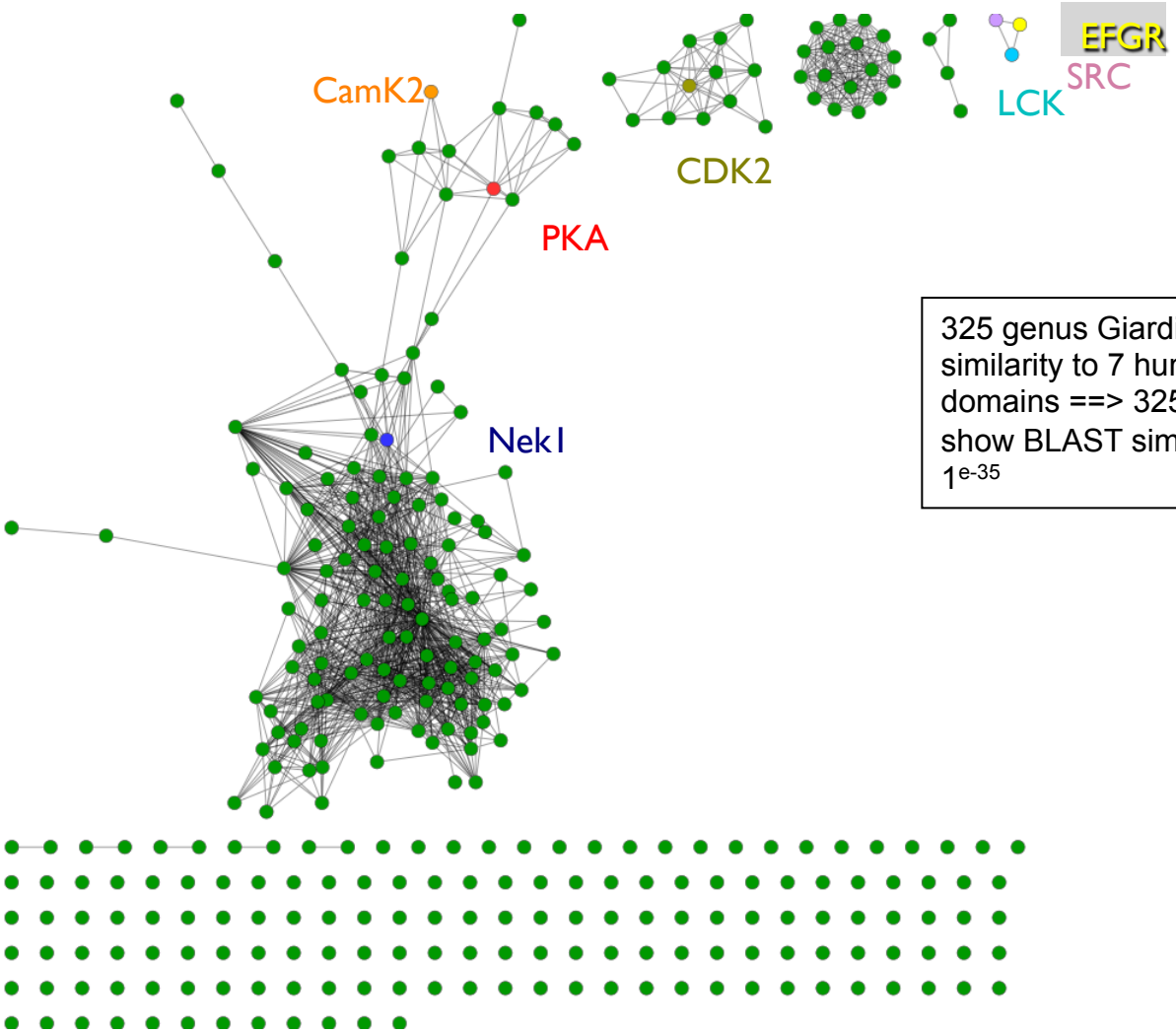
Comparative proteomics: Human vs parasite kinomes



2635 human seqs with similarity to 7 human kinase domains filtered to 90% NR ==> 818 nodes; edges show BLAST similarity better than $1e^{-35}$

Holly Atkinson

Parasite kinomes are simpler — and different

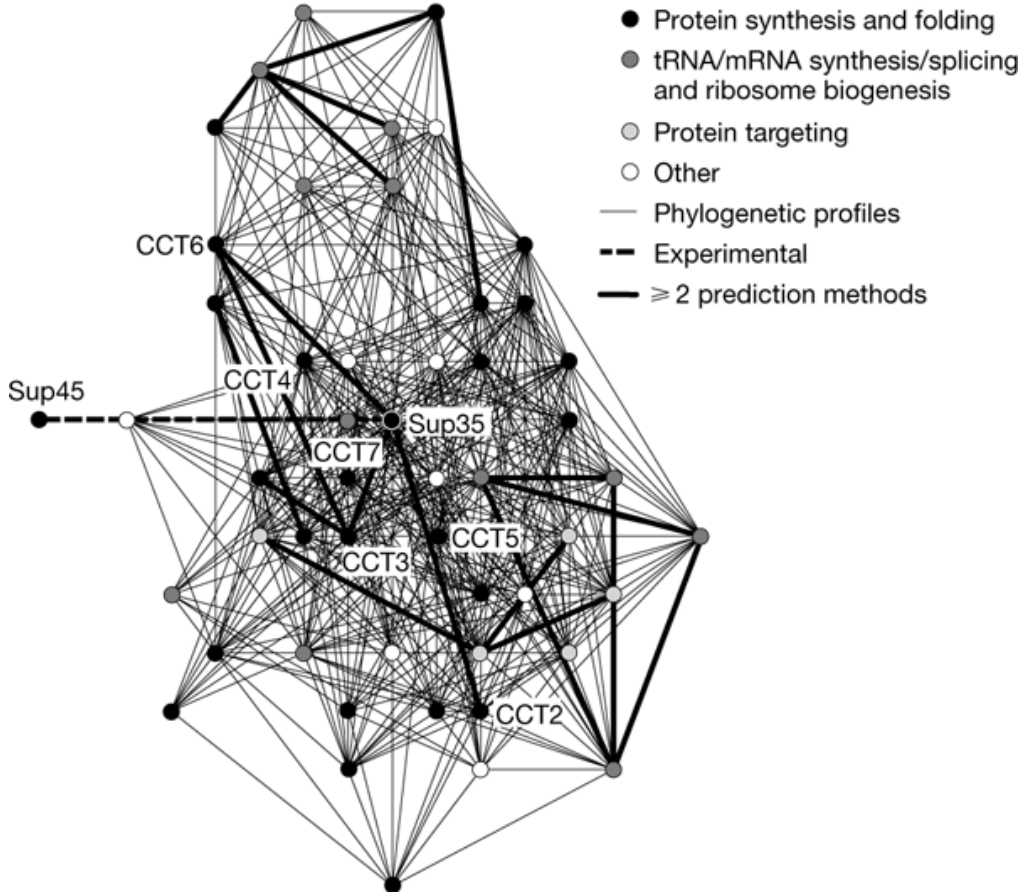


325 genus Giardia seqs with similarity to 7 human kinase domains ==> 325 nodes; edges show BLAST similarity better than 1^{e-35}

Holly Atkinson

Interaction networks

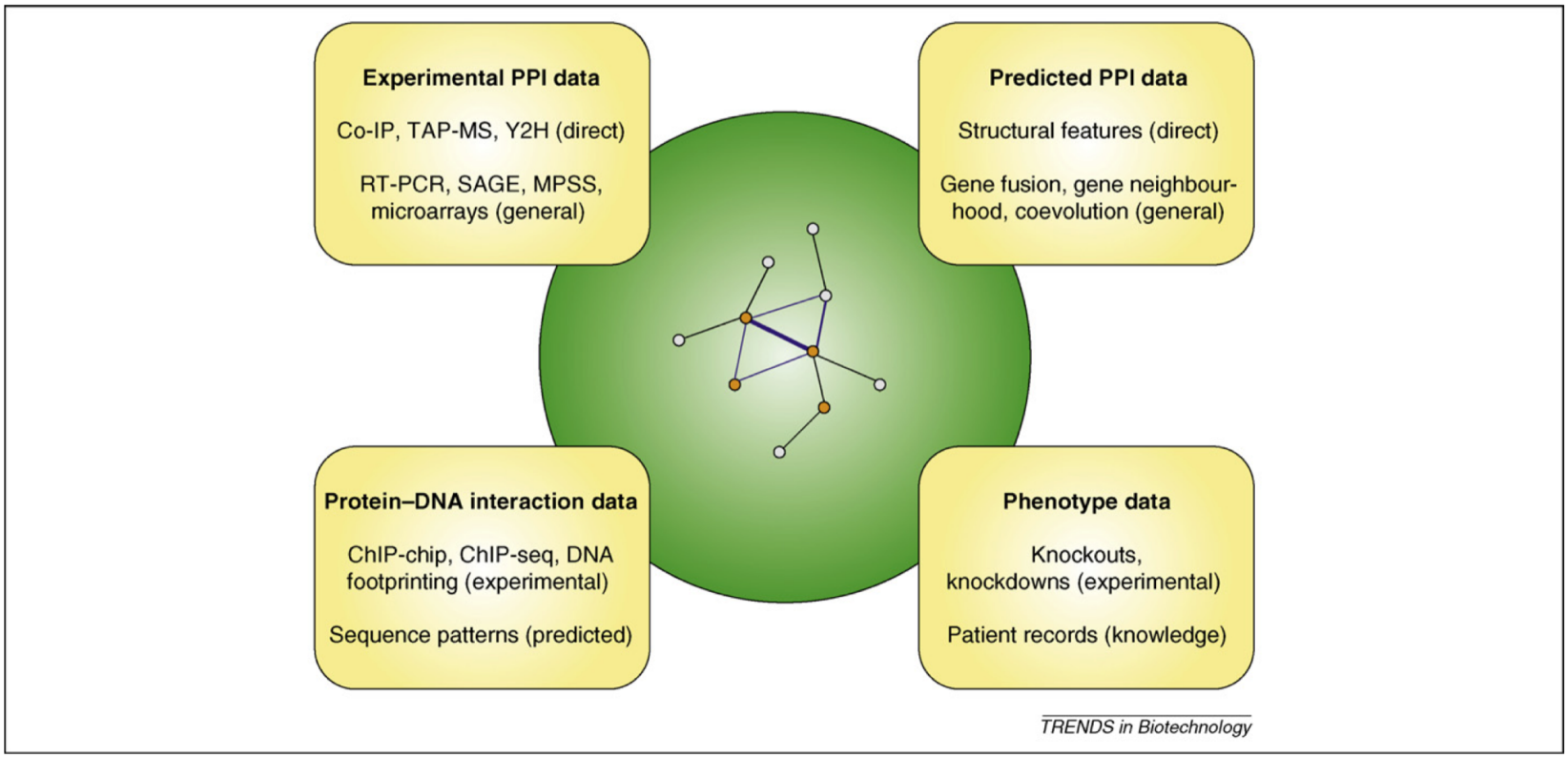
(Marcotte et al, A combined algorithm for genome-wide prediction of protein function, Nature 402:83 (1999))



Protein interactions among proteins linked to the yeast prion Sup35 leads to functional assignment related to mitochondrial protein synthesis

Functional Linkage Networks

from Figure 3, Rentzsch & Orengo. Tr Biotech 27:210 (2009)

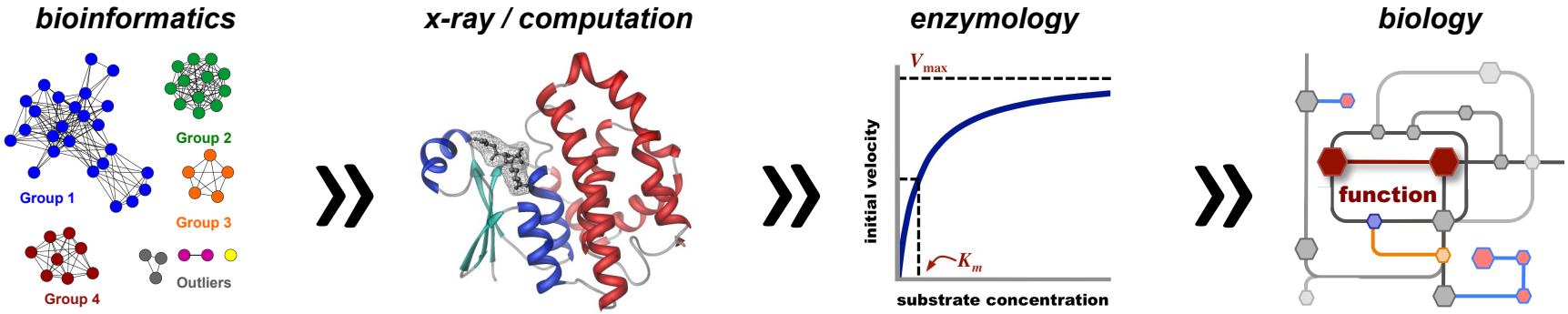


Ultimately, bioinformatics cannot afford to be disconnected from experiment

- Loses the checks that come from expert information
- Experiment & analysis are no longer integrated and sometimes have suffered an ugly divorce...
  - Experimental artifacts
    - measuring low abundance proteins in proteomics
  - Functional inference
    - experimental validation of some predictions provides a critical control & helps to define an appropriate range of variation for annotation transfer
  - Lack of structured vocabularies across biological fields leads to enormous confusion, loss of critical information, error
    - Parsing a database that changes its format...
    - Importance of evidence codes
    - Does everyone mean the same thing in describing a protein family?

5. Experiment!

- What does it take to identify new functions in functionally diverse enzyme superfamilies?



- Example: Enzyme Function Initiative
  - development of a multidisciplinary sequence / structure based strategy for facilitating discovery of *in vitro* enzymatic and *in vivo* metabolic / physiological functions of unknown enzymes discovered in genome projects

Some useful bioinformatics texts

- *Sequence - Evolution - Function*. Koonin & Galperin, ISBN-13: 978-1402072741
- *An Introduction to Bioinformatics Algorithms*. Jones & Pevzner, ISBN-13: 978-0262101066
- *Algorithms on Strings, Trees and Sequences*: Computer Science and Computational Biology. Dan Gusfield, ISBN-13:978-0521585194
- *Statistical Methods in Bioinformatics*. Ewens & Grant, ISBN-13: 978-0387952291