A bioinformatics approach to the structural and functional analysis of the glycogen phosphorylase protein family

Jieming Shen[1,2] and Hugh B. Nicholas, Jr.[3]

[1]Bioengineering and Bioinformatics Summer Institute, Department of Computational Biology, University of Pittsburgh, Pittsburgh, PA 15261

[2]Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854

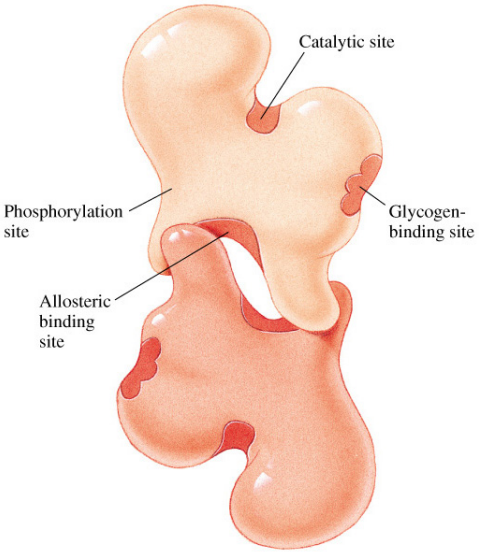
[3]Biomedical Initiative Group, Pittsburgh Supercomputing Center, Pittsburgh, PA 15213

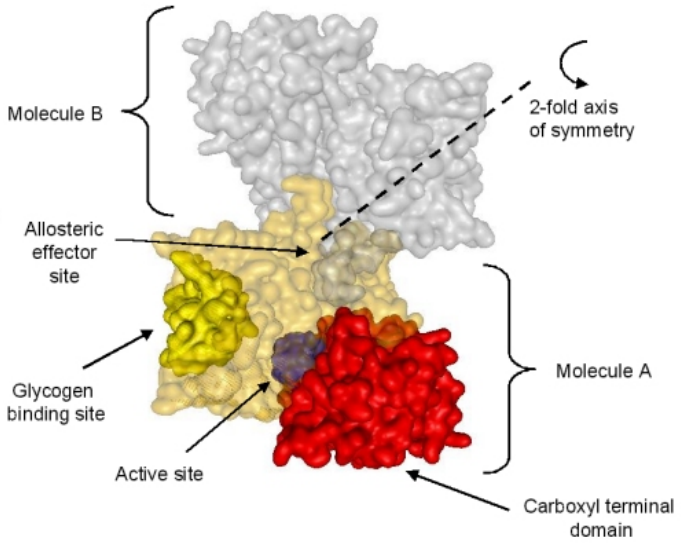
Objectives

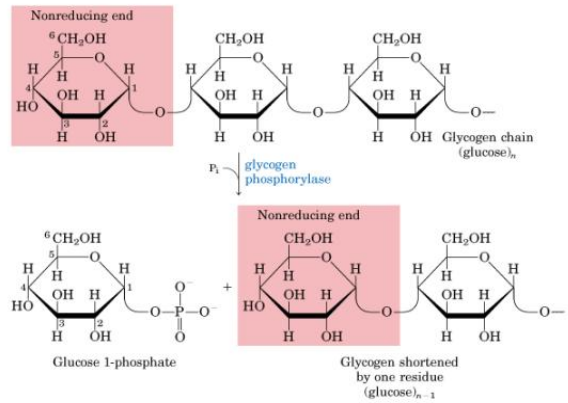
- Build a phylogenetic profile for the glycogen phosphorylase protein family from existing sequences
  - Identify orthologues/paralogues; orthologues have same biochemical/physiological role
  - Proteins involved in the same pathway usually have similar phylogenetic profiles
  - Aids in identification of pathways and physiological processes in which uncharacterized protein appears
- Conduct a validation study of the ProbCons alignment algorithm
  - Prepare multiple sequence alignments in parallel using ProbCons as well as more established methods for obtaining alignment (ClustalW)
  - Compare results

http://www.iscb.org/ismb2004/posters/Thomas.MeinelATmolgen.mpg.de_323.html

Background Information





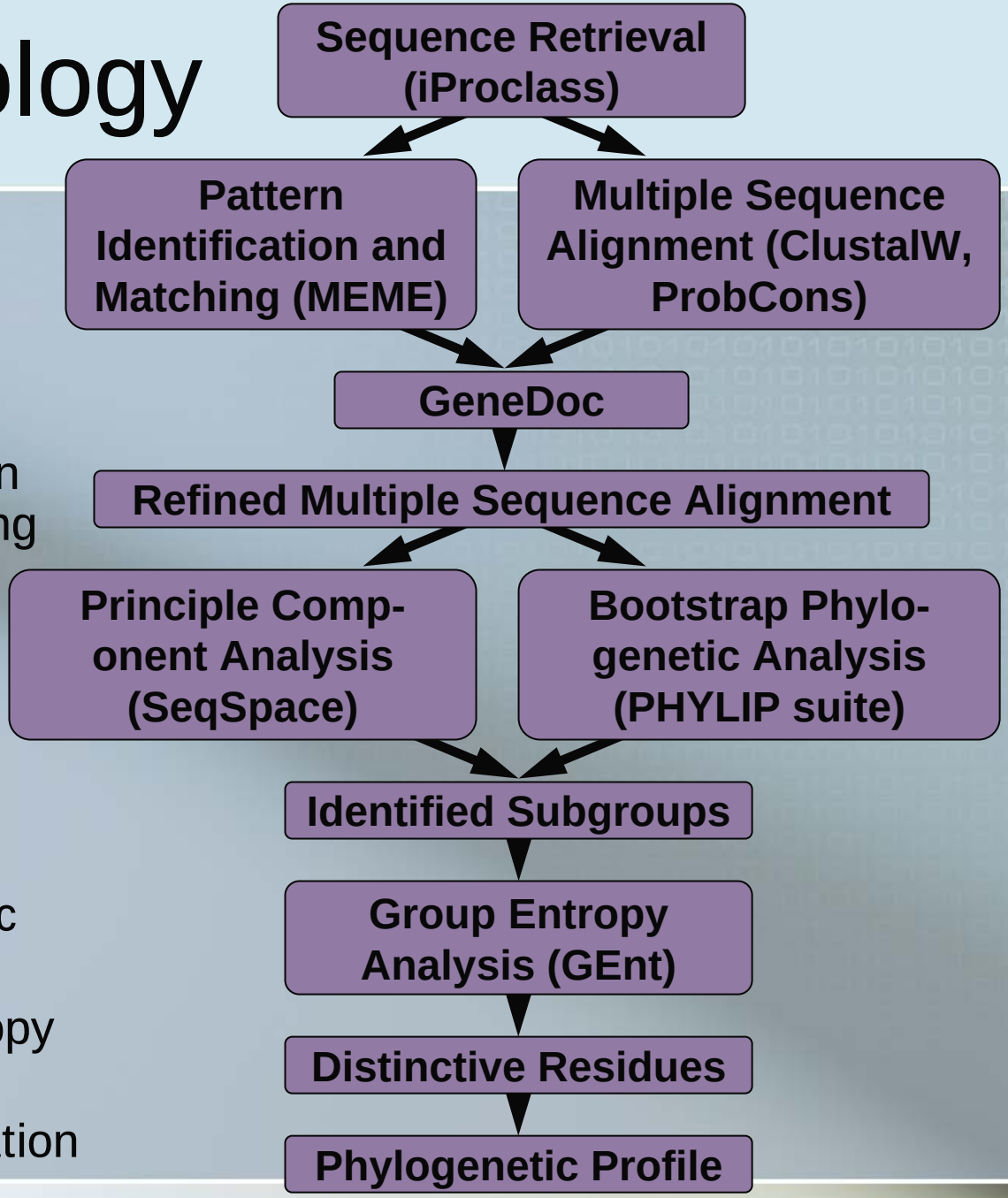


- Glycogen phosphorylase plays a major role in carbohydrate metabolism by catalyzing the breakdown of glycogen into glucose subunits
  - Acts on linear chains of glycogen
  - Glycogen shorted by one residue
  - Dimer of 2 identical subunits

http://wine1.sb.fsu.edu/bch4053/Lecture27/Lecture27.htm
http://cwx.prenhall.com/horton/medialib/media_portfolio/13.html

Methodology

- Overview
  - Sequence retrieval
  - Pattern identification and matching
  - Multiple sequence alignment
  - Principle component analysis
  - Phylogenetic analysis
  - Group entropy analysis
  - 3D visualization



Sequence Retrieval

- iProclass database
  - Obtained 355 glycogen phosphorylase sequences of about 800+ residues each
  - Eliminated duplicates or near duplicates—left with 282 sequences
  - Final dataset included top most sequenced genomes:
    - honeybee, chicken, sea squirt, cow, dog, drosophila, Japanese puffer fish, human, mosquito, mouse, and C. elegans

Pattern Identification and Matching

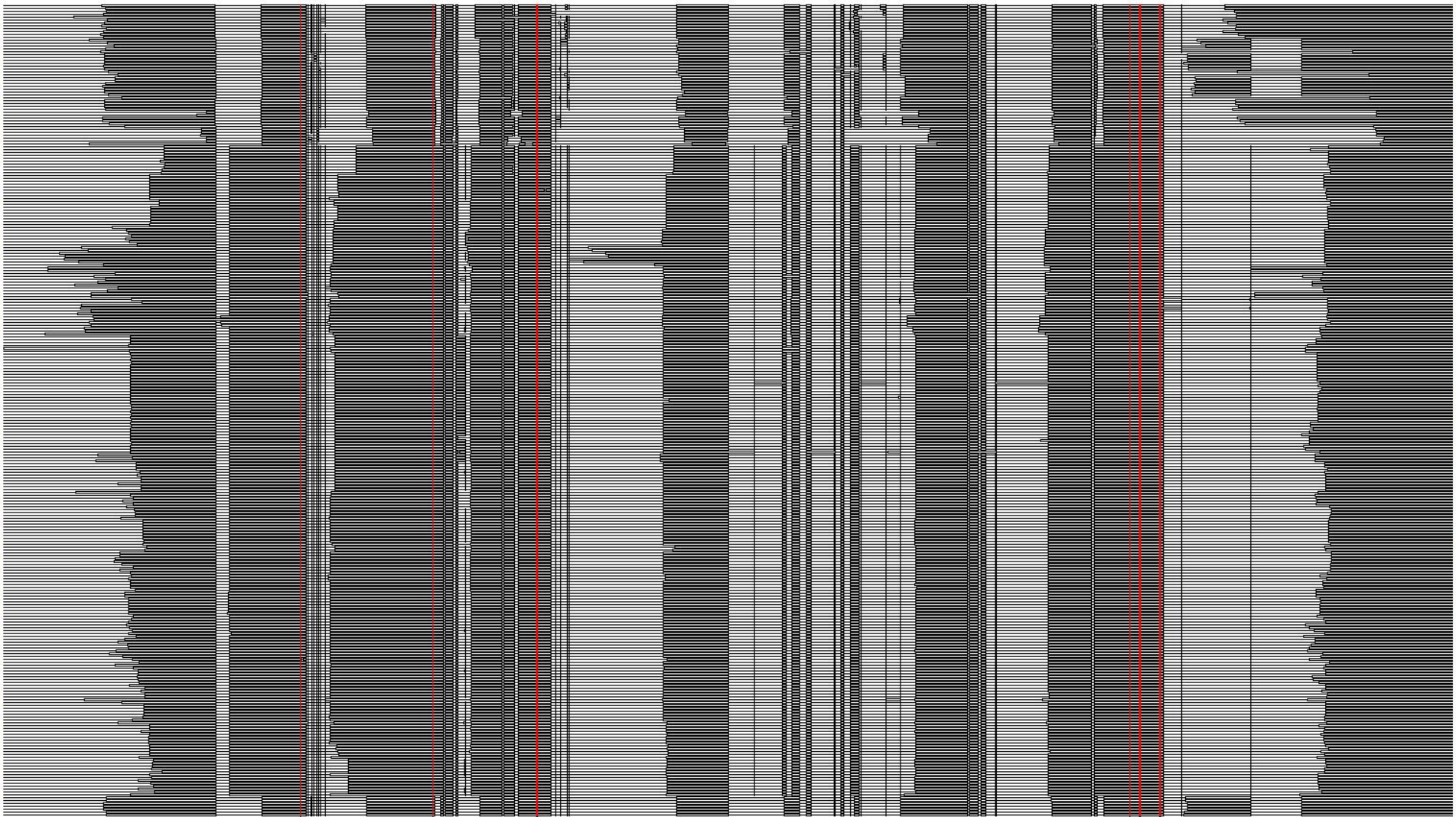
- MEME (Multiple EM for Motif Elicitation)
  - Discovers motifs (highly conserved sequence patterns) in group of related protein sequences
  - Sorts through sequences and finds and reports motif alignments regardless of their placement along the protein
  - Obtained 20 highly conserved motifs
  - Conserved residues and motifs essential to protein structure and function

Multiple Sequence Alignment

- ClustalW
  - Progressive method
  - ~1 hour
- ProbCons
  - Consistency-based method
  - ~10 hours
- T-COFFEE, another consistency-based method, was attempted but did not complete the MSA within a reasonable amount of time (2 weeks)
  - may have had to do with large size of dataset and long sequence length
- GeneDoc program used to highlight MEME patterns in the MSAs

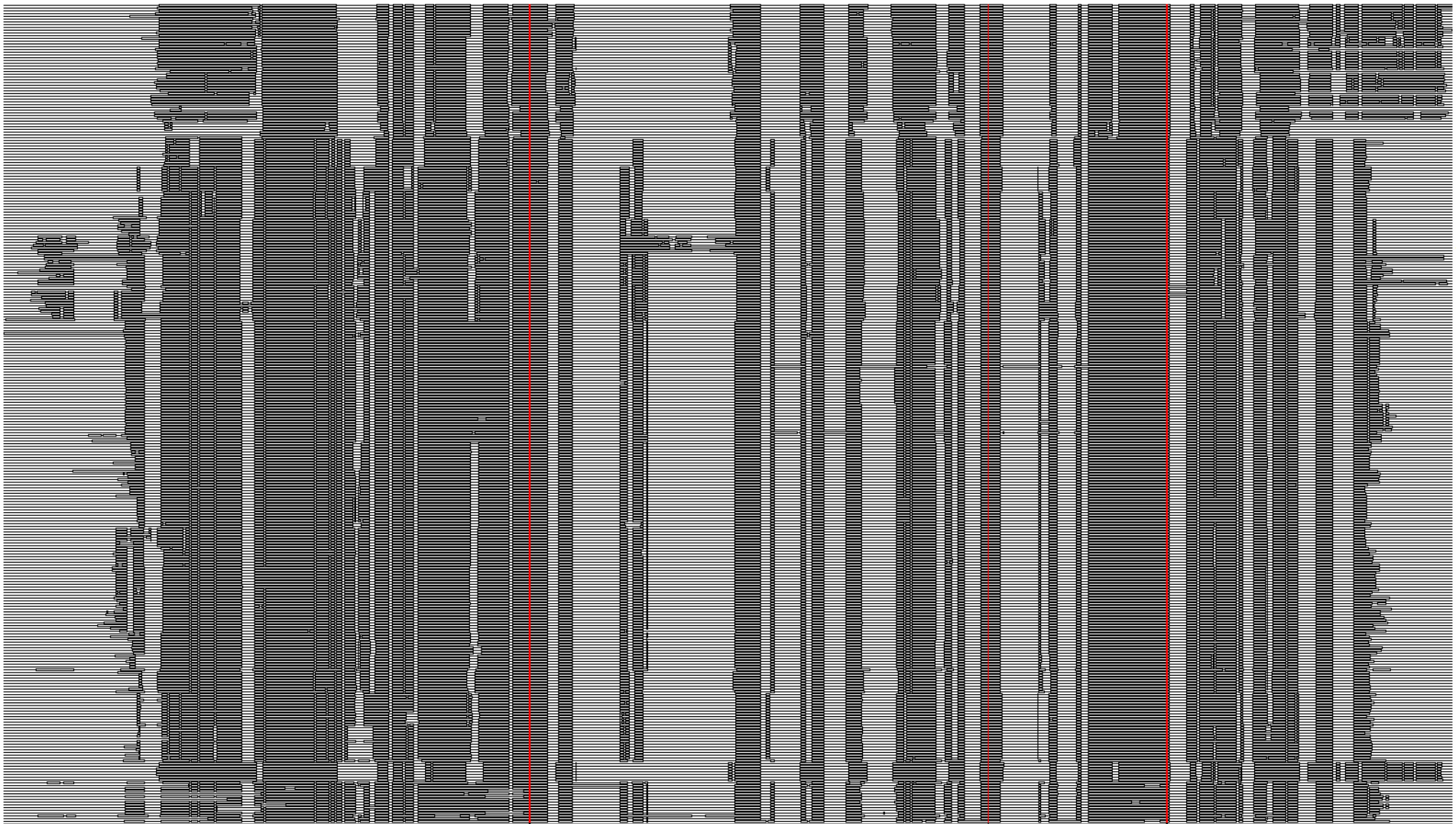
Multiple Sequence Alignment

- GeneDoc: ProbCons conserved residues



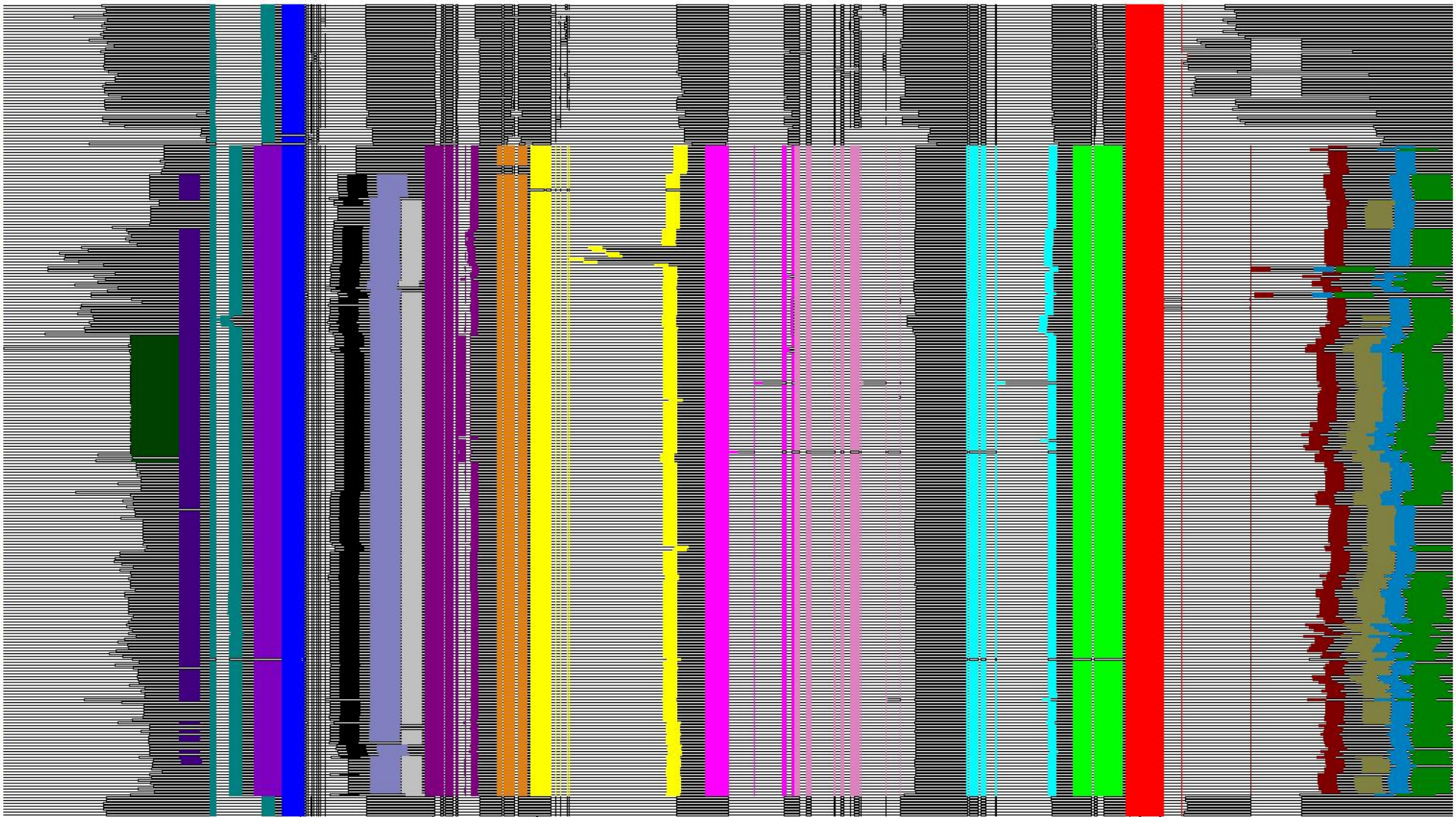
Multiple Sequence Alignment

- GeneDoc: ClustalW conserved residues



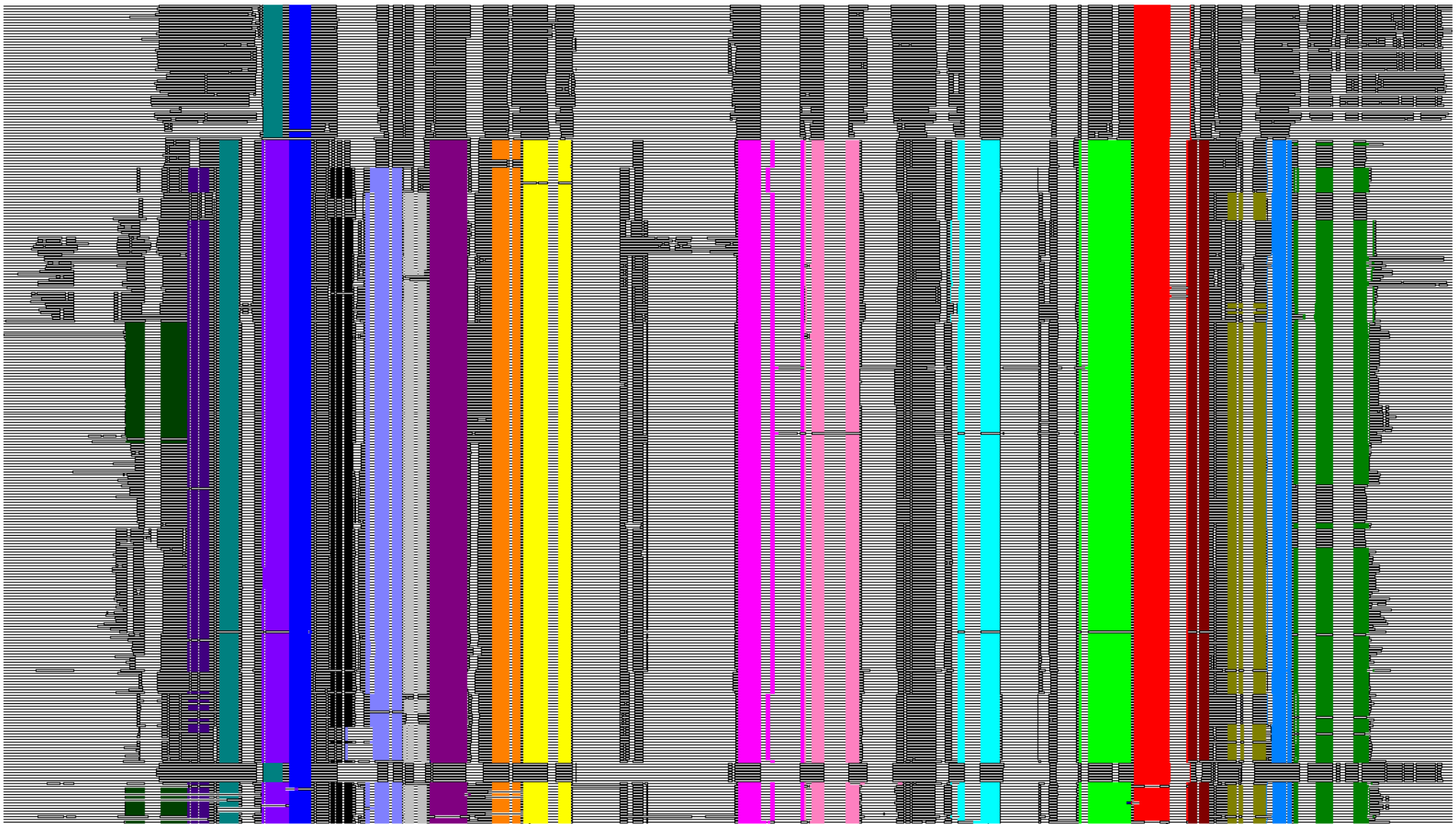
Multiple Sequence Alignment

- GeneDoc: ProbCons with highlighted MEME motifs



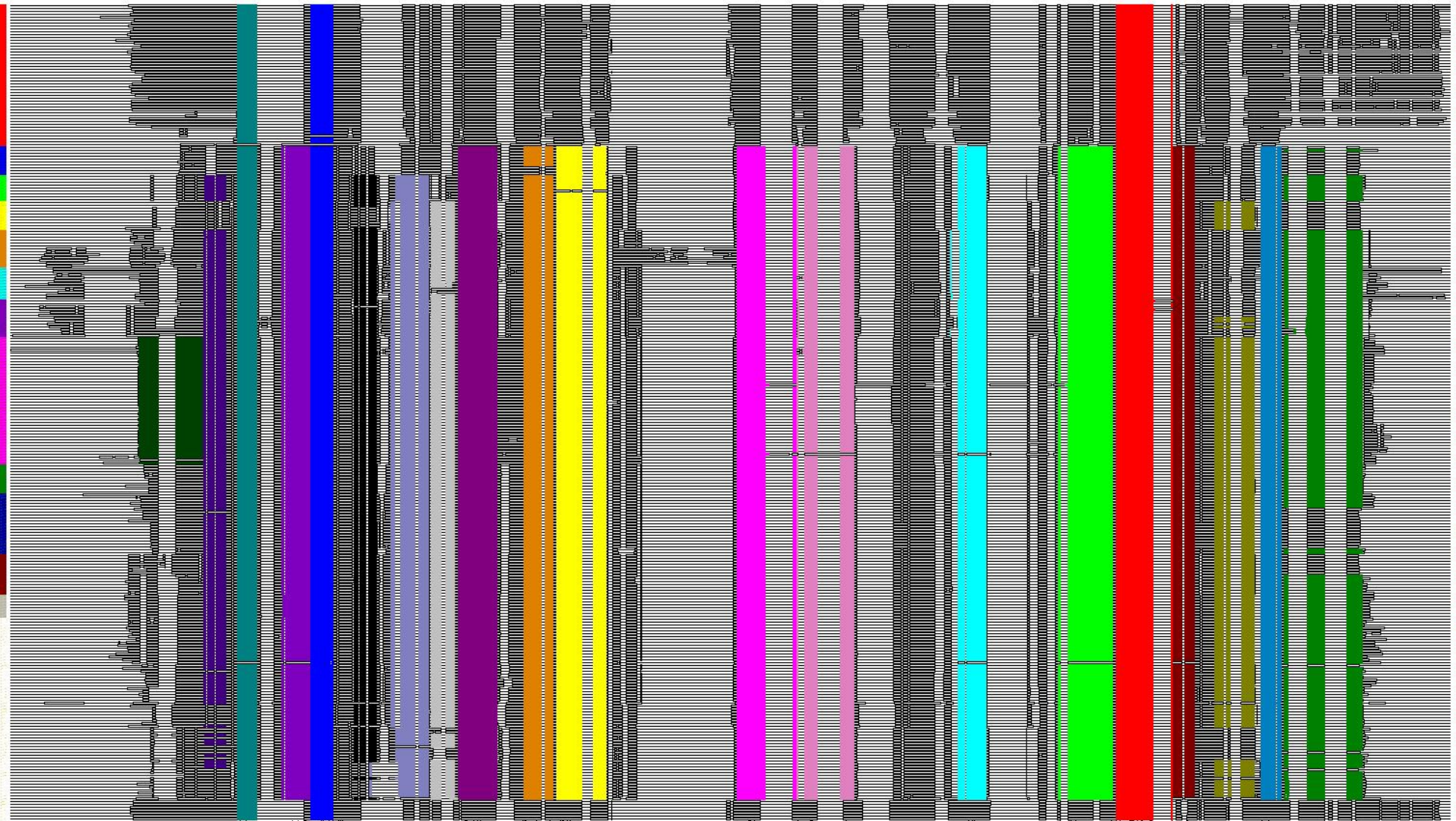
Multiple Sequence Alignment

- GeneDoc: ClustalW with highlighted MEME motifs



Refined MSA

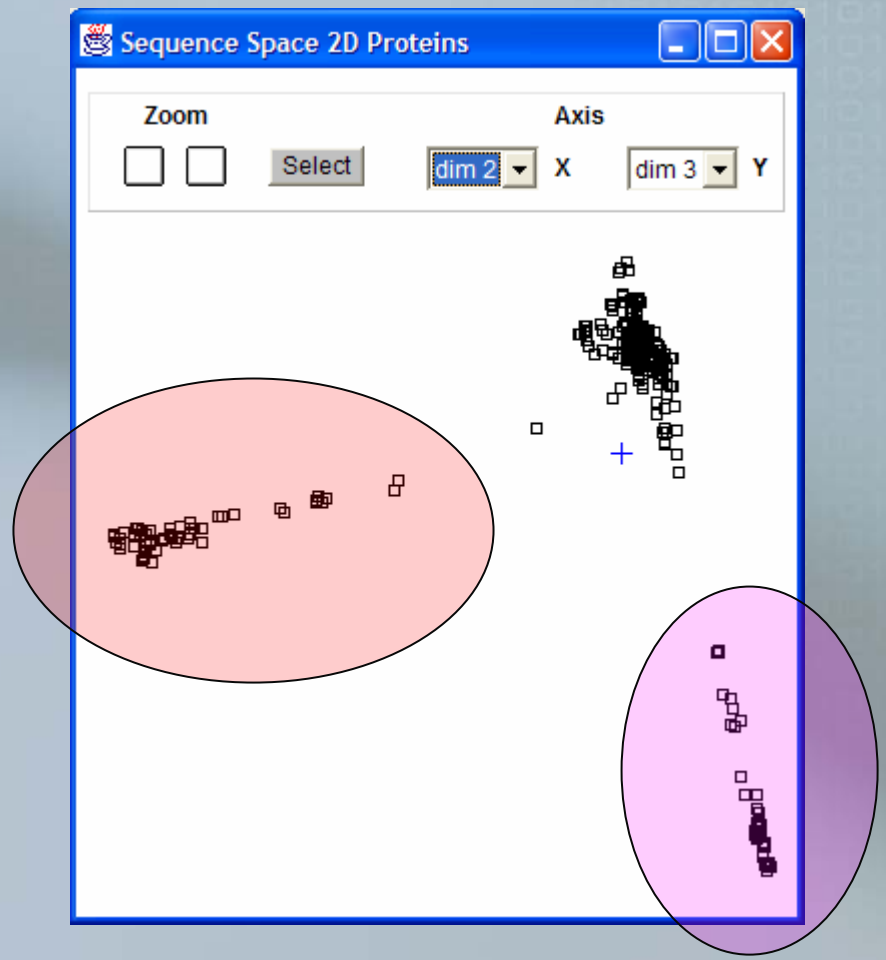
- Incorporated information from both the ClustalW and ProbCons outputs
- Refined alignment by hand using MEME motifs as guide



Principle Component Analysis

- SeqSpace
  - One method to identify groups of subfamilies
  - “Top-down” analysis
  - Takes distance measures among set of sequences and converts into self-consistent set of coordinates in some arbitrary number of dimensions
    - Coordinates plotted and examined for clusters of sequences
    - Clustered protein sequences share similar pattern of substitutions—presumed to reflect some common biochemical/physiological property or function

SeqSpace: Dimensions 2 x 3



Bootstrap Phylogenetic Analysis with PHYLIP suite

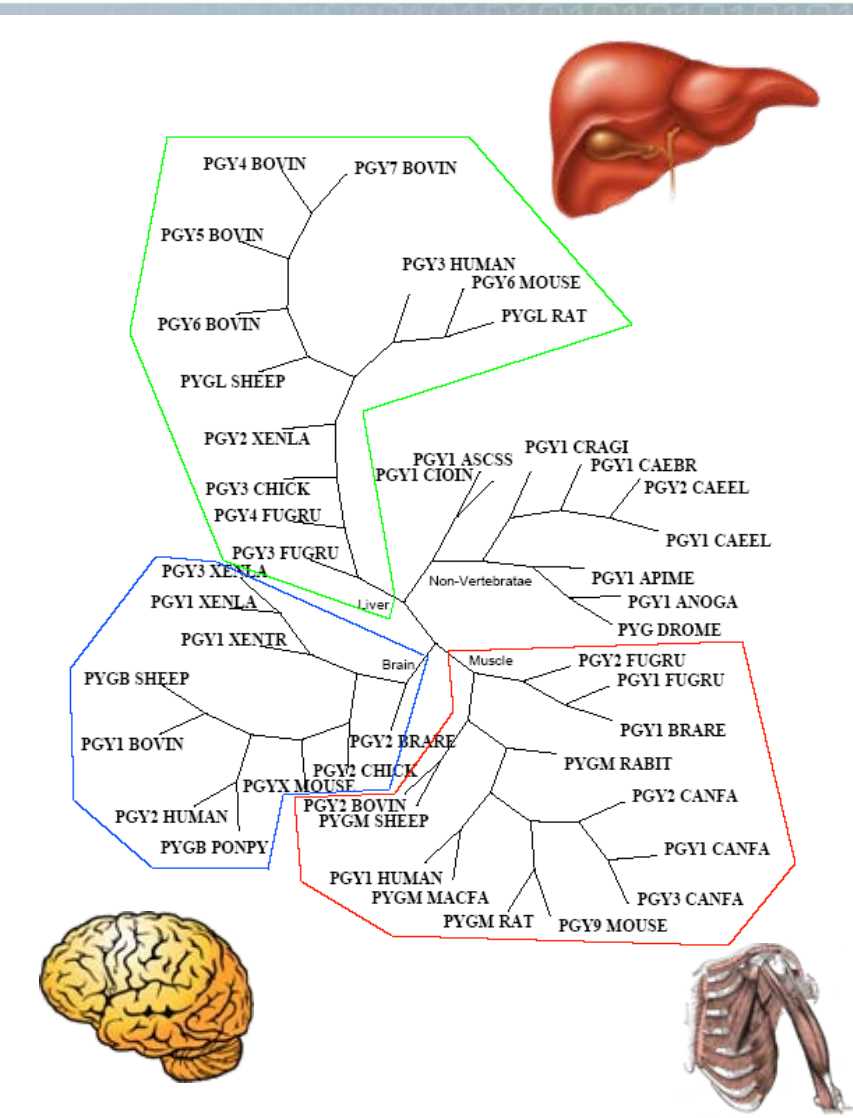
- ClustalW alignment with gap columns eliminated
- Gblocks: eliminate poorly aligned positions and divergent regions of sequence of protein alignment
- Seqboot: generate multiple data sets (1000) that are resampled versions of the input data set, randomly sample columns with replacement
- Protdist: analyzes the multiple data sets and computed a distance measure for protein sequences, using maximum likelihood estimates based on the Dayhoff PAM matrix in this case

Bootstrap Phylogenetic Analysis with PHYLIP suite

- Neighbor: constructs a tree by successive clustering of lineages, setting branch lengths as the lineages join
- Consense: computes consensus trees by majority-rule method
- Phylogenetic consensus tree viewed in ATV viewer and TreeView
- Groups of subfamilies compared to those obtained from Seqspace for further refinement

TreeView

- Unrooted tree for Metazoa subfamily recalculated from distinct subset of overall tree
- Reveals enzyme isozyme groups for brain, muscle, and liver tissue within the vertebrates



http://science.howstuffworks.com/brain.htm
http://digilander.libero.it/BodyMindCare/kapil/moremedi.htm
http://en.wikipedia.org/wiki/Biceps_brachii

GEnt Group Entropy Analysis

- Identify distinctive features of mutually exclusive subsets determined by phylogenetic analysis through cross entropy analysis of the columns in MSA
- Within single column, contrast:
  - Amino acid composition within defined group of sequences
  - Amino acid composition in all sequences outside of group
- Alignment positions where residue composition inside groups is very different from that outside group are expected to indicate positions associate with distinctive properties of sequences of that subgroup

Group Entropy Equations

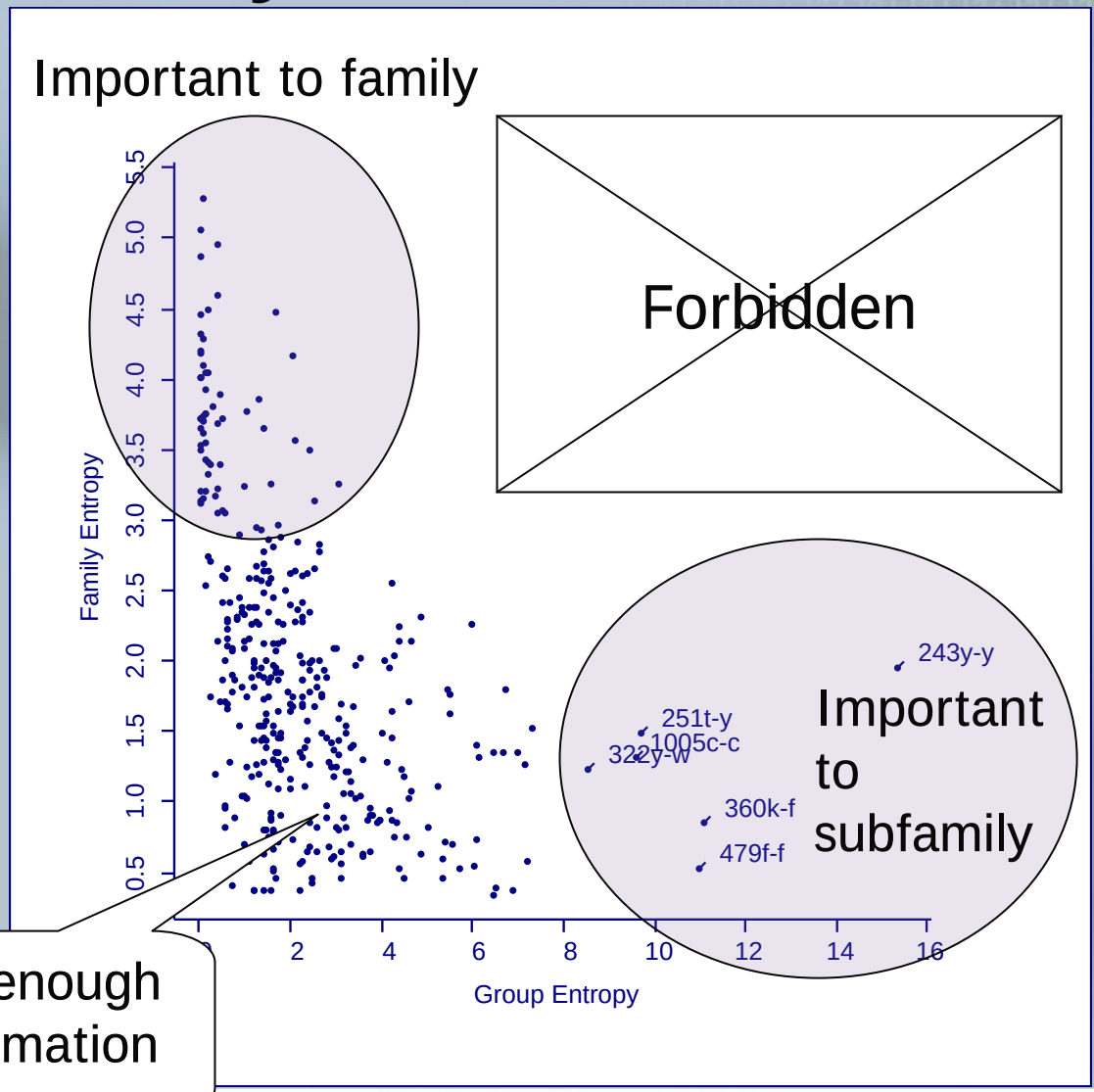
$$\text{Family Entropy Distance} = \sum \left[(p_i) \times \log_2 \left(\frac{p_i}{q_i} \right) \right]$$

$$\text{Group Entropy Distance} = \sum \left[(p_i - q_i) \times \log_2 \left(\frac{p_i}{q_i} \right) \right]$$

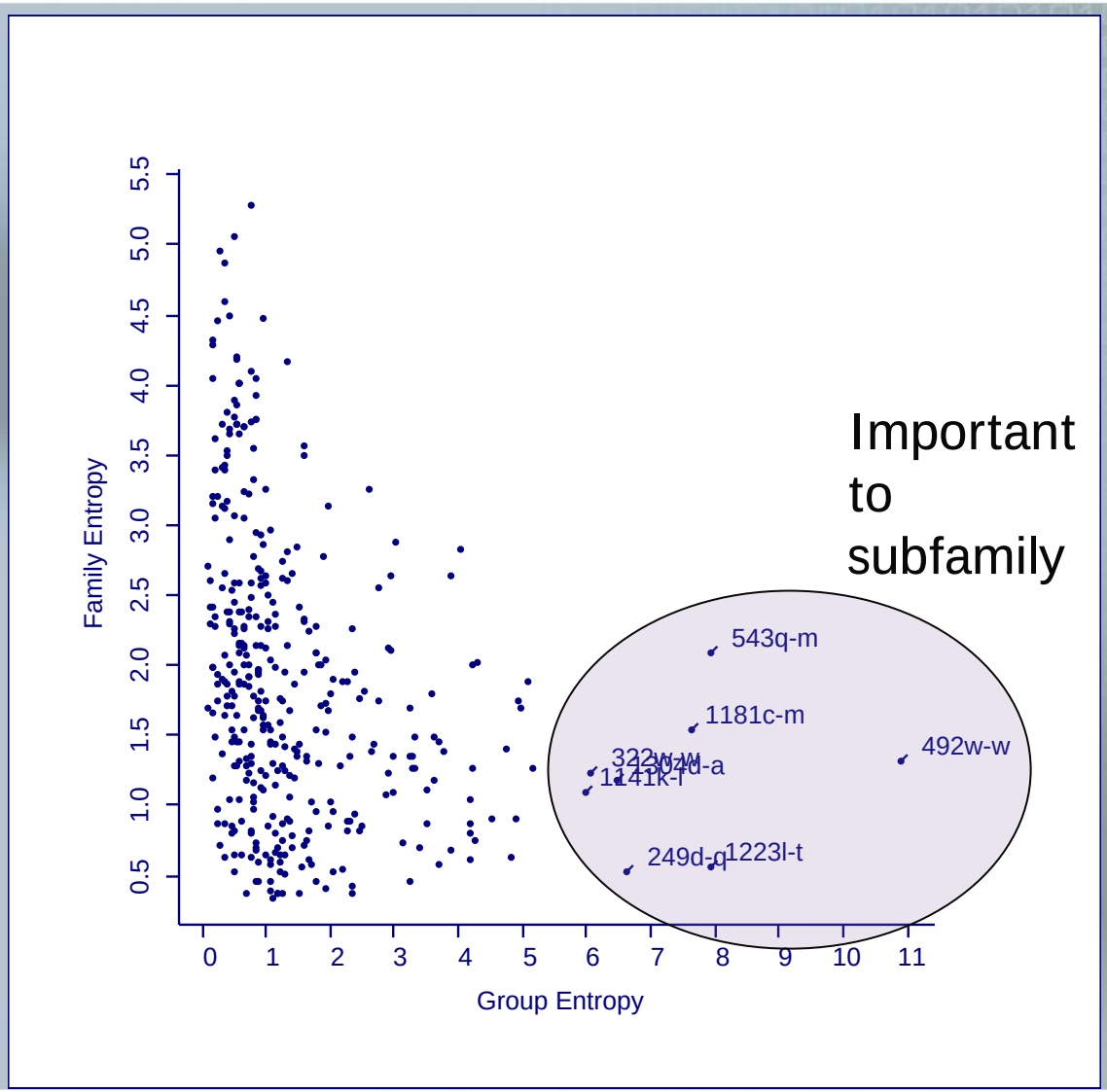
p_i = foreground residue frequency

q_i = background residue frequency

GEnt Results for Metazoa Subfamily

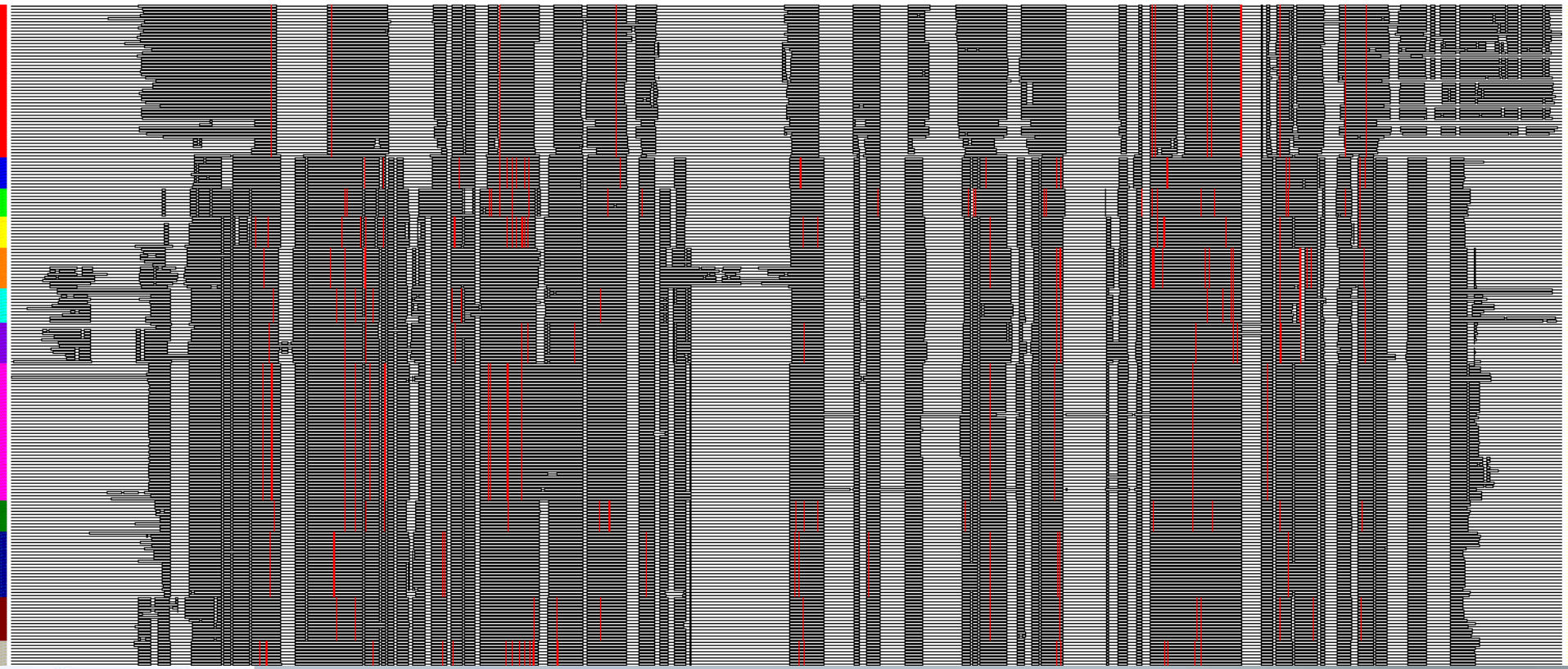


GEnt Results for Fungi Subfamily



GEnt

- Distinctive subfamily residues highlighted

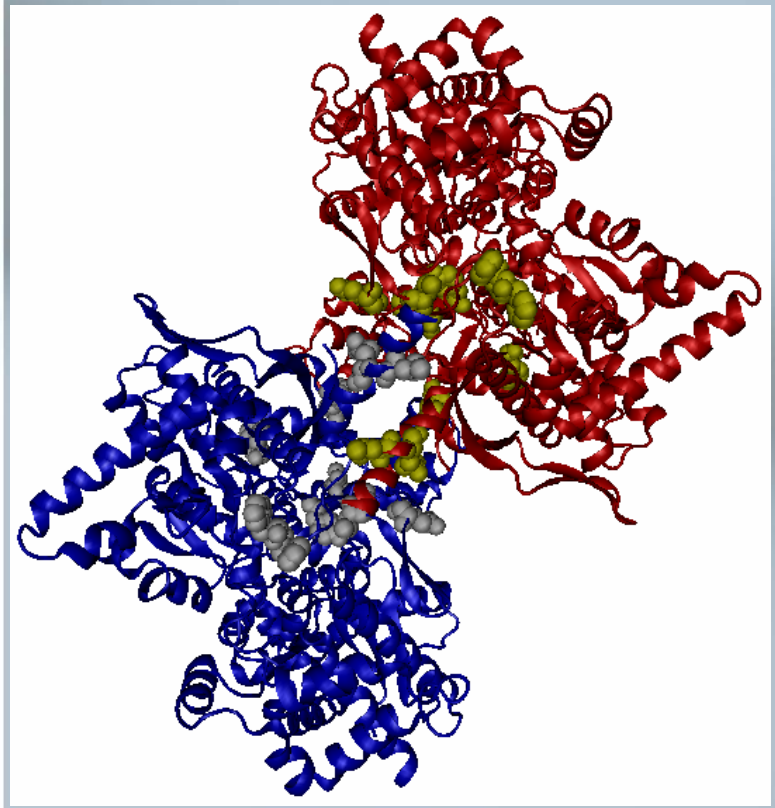


Visualization

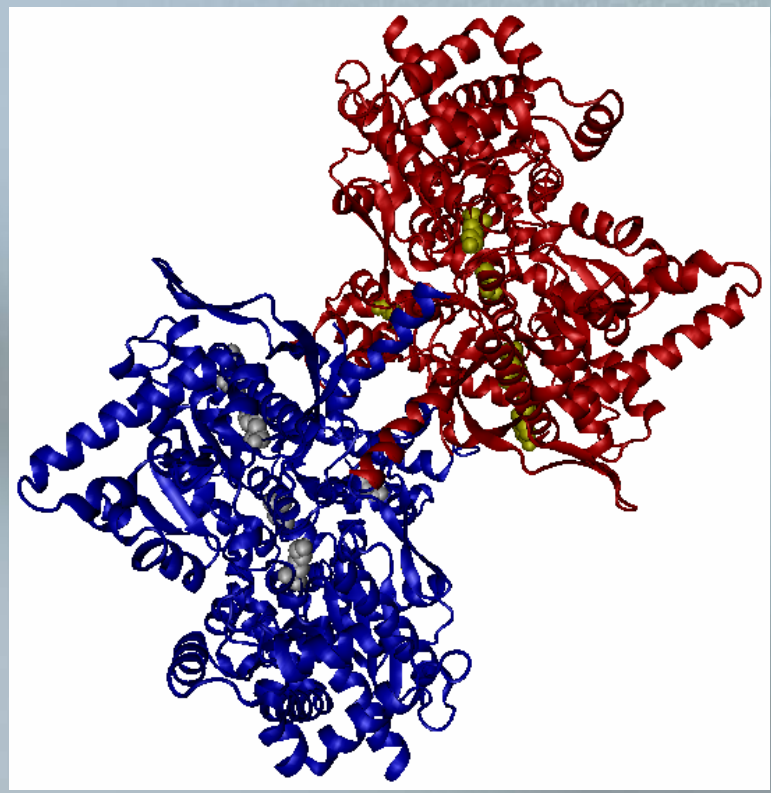
- 3D structures for human liver isoform and yeast obtained from Protein Data Bank
- Residues of interest highlighted in VMD

Visualization: Human Liver Isoform

GEnt: distinctive residues

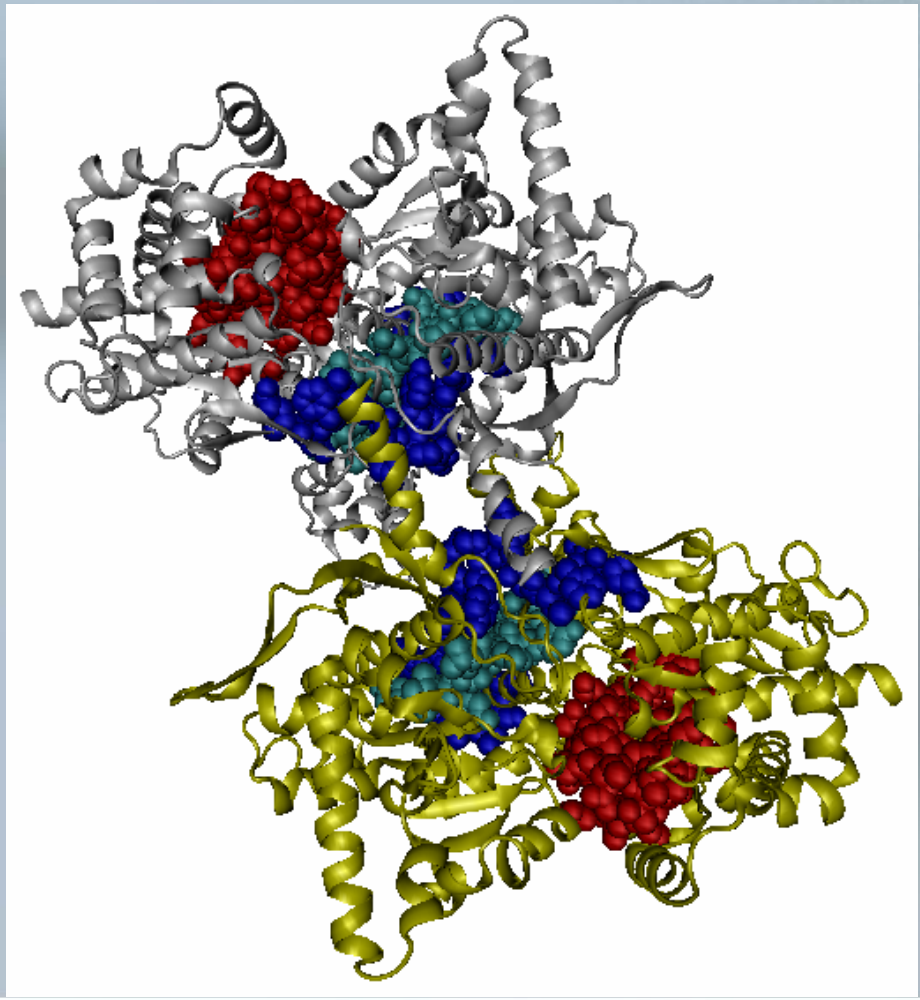


UniProt: binding/association sites



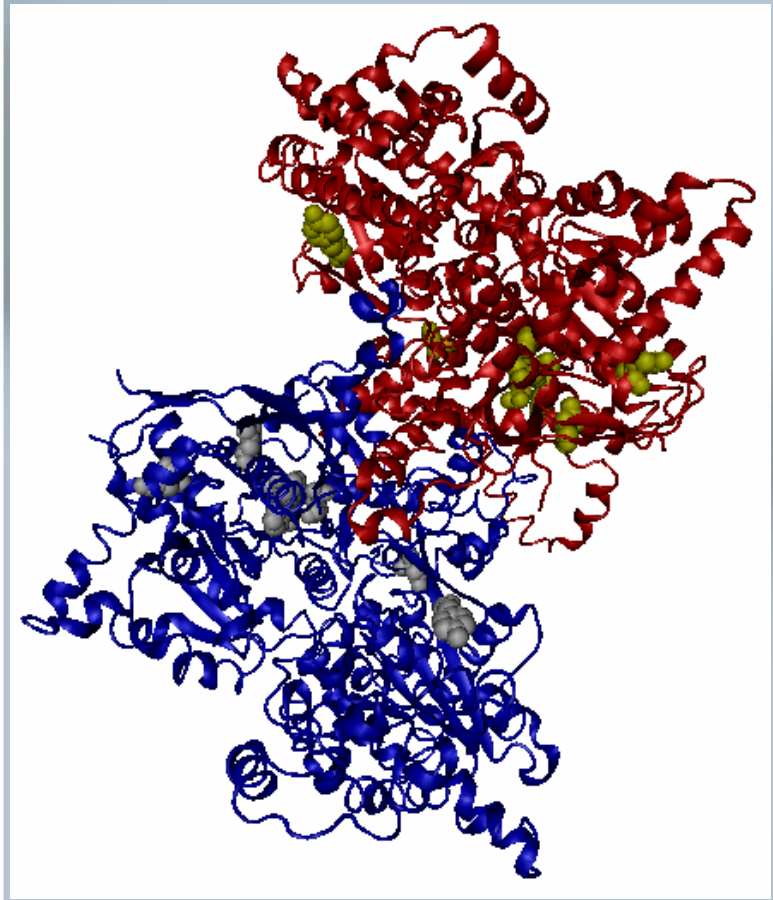
Visualization: Human Liver Isoform

- Globally conserved MEME motifs



Visualization: Yeast

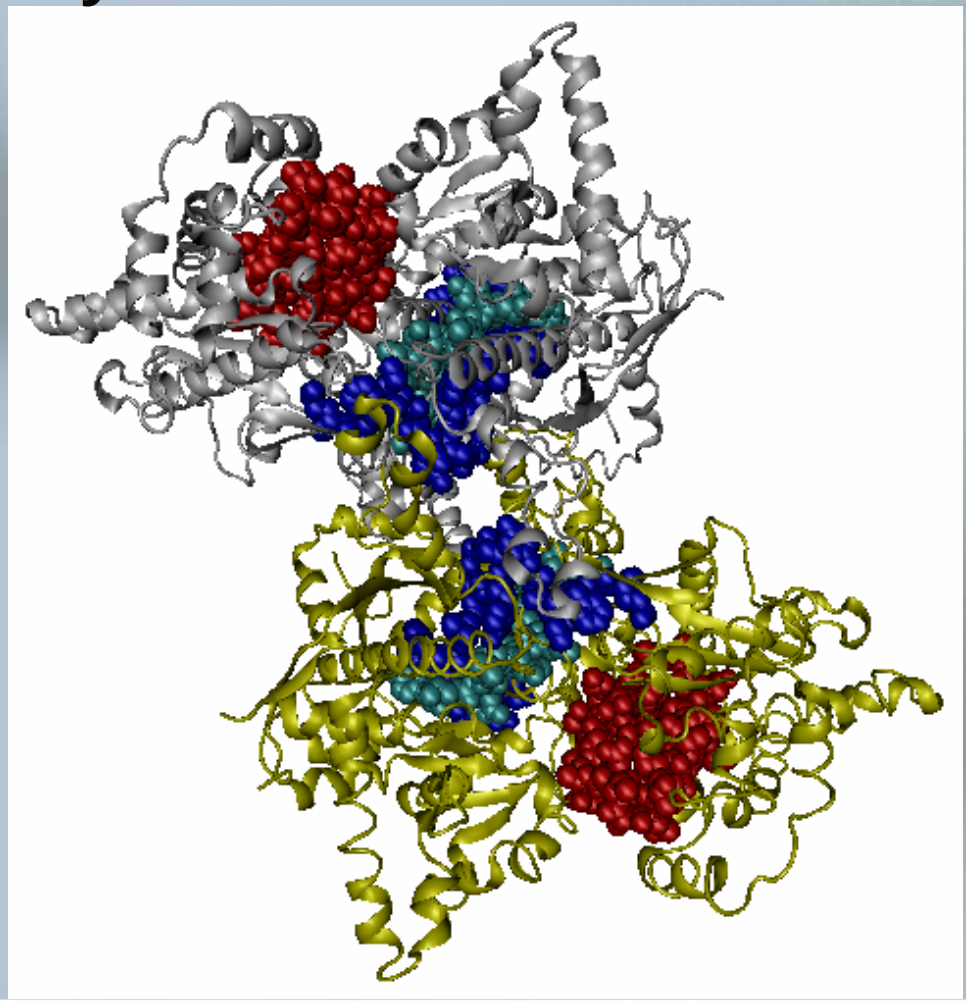
GEnt: distinctive residues



UniProt: binding sites

Visualization: Yeast

- Globally conserved MEME motifs



Conclusions

- GEnt
  - In metazoans: identified residue (155) known to be important in allosteric control
  - Other predicted residues have high probability of reflecting important adaptations of enzyme to environment
- Determined possible evolutionary path of glycogen phosphorylase
- MSA methodology
  - ProbCons and T-COFFEE did not perform well due to the large number of sequences and long sequence length of the dataset
  - In this case, the ClustalW alignment method gave the best results

Future Work

- Experimentally test GEnt-predicted residues (site-directed mutagenesis)
- Additional computation work with quantum mechanics and molecular mechanics to identify roles of conserved residues, motifs, and GEnt-predicted residues

Acknowledgements

- Hugh B. Nicholas, Jr.
- Alexander J. Ropelewski
- Ricardo G. Mendez
- Pittsburgh Supercomputing Center
- Bioengineering & Bioinformatics Summer Institute

References

- PSC website tutorials (http://www.psc.edu/nrbsc/education/tutorials/)
- Michael M. Crerar, Ph.D. (http://www.biol.yorku.ca/grad/faculty/michaelm.htm)
- Buchbinder et al. (2001) Structural relationships among regulated and unregulated phosphorylases. Annu. Rev. Biophys. Biomol. Struct. 30: 191-209.
- Nicholas Jr., H.B. Glutathione S-transferase subfamily differences: remodeling the subunit and domain interfaces.