Extending Gene Families via Predicted Ancestral Sequences

Loretta Goldberg
Computational Biosciences
Arizona State University
April 28, 2006

This internship project is presented in partial fulfillment of the requirement of the Professional Science Master's in Computational Biosciences,

Arizona State University

Research conducted from

September 2005 – April 2006

Acknowledgements

- Thanks to my Advisor, Dr. Michael Rosenberg, who proposed this internship project, and has provided guidance throughout.
- Thanks to my committee: Dr. Rosenberg, Dr. Renaut and Dr. Touchman, for the knowledge they have shared throughout my coursework and for review of this final project.

Personal Background

- B.A. Chemistry
  - Russell Sage College, Troy NY
  - 4 years as a Process Development Chemist in the Pharmaceutical Industry
- M.S. Computer Sciences
  - Arizona State University
  - 14 years as a Software Engineer in the Telecommunications Industry
- P.S.M. Computational Biosciences
  - TBD

Extending Gene Families via Predicted Ancestral Sequences

- Gene families
  - What are they?
  - How do we determine them?
- Ancestral Sequences
  - What are they?
  - How do we determine them?
- How does this project propose to utilize ancestral sequences in the context of gene families?

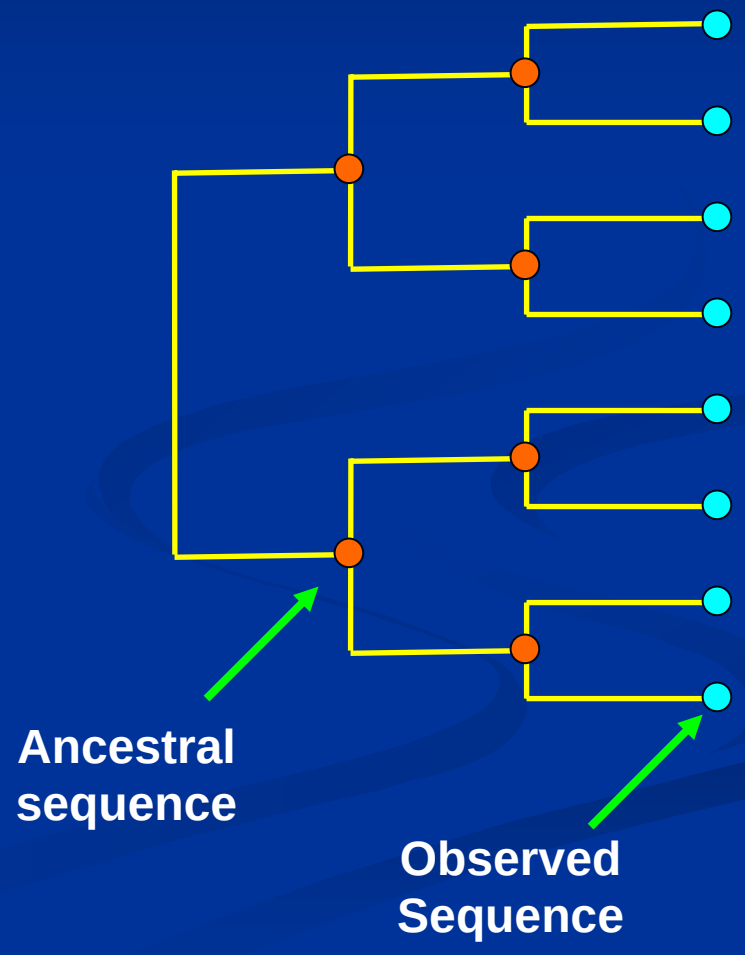
Gene Families

- A gene family is the grouping together of genes based on the similarity of the products or proteins that they produce.

- A gene family is a set of related genes occupying various loci in the DNA, almost certainly formed by duplication of ancestral genes, and having a recognizably similar sequence.

Ancestral sequences

- A phylogenetic tree is a tool used to show the evolutionary relationship between biological objects
- The tips represent the observed genes (sequences) in living organisms
- The nodes or branching points represent the extinct ancestors



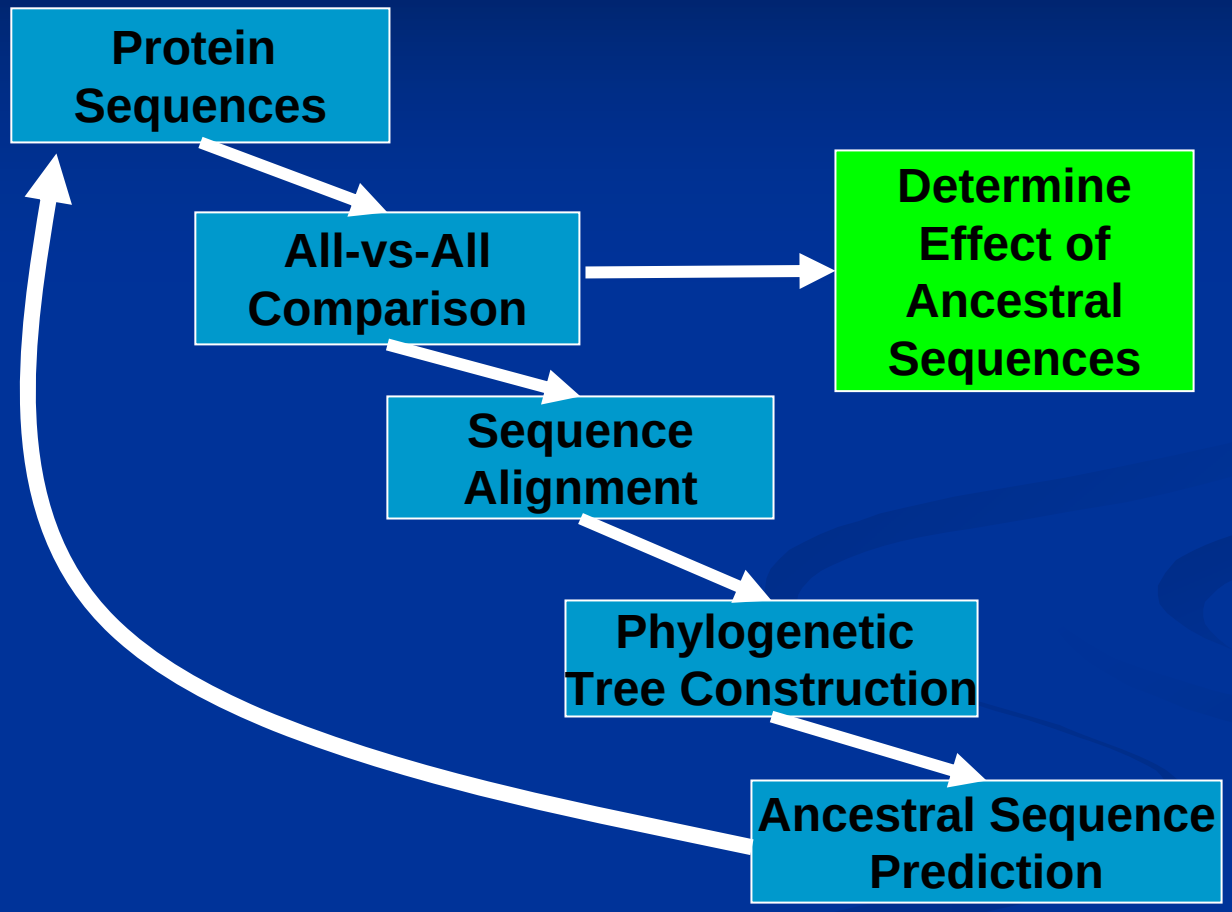
Project Goal

To determine if the inclusion of ancestral sequences in an all-vs-all comparison of protein sequences will extend the number of homologous sequences that can be extracted from this comparison.

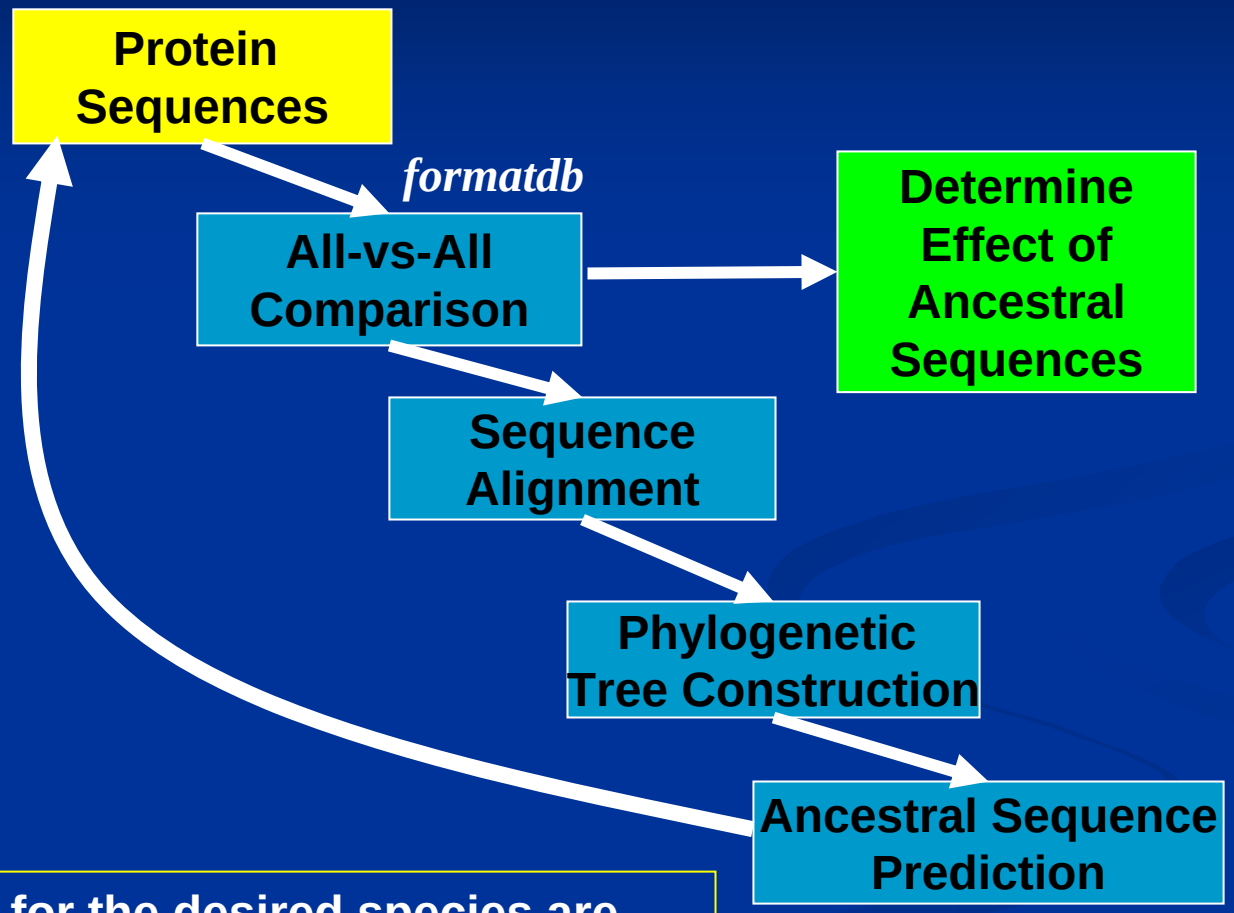
Proposed Work Flow

- Identifying protein sequences of interest
- All-vs-all sequence comparison
- Alignment of sequences
- Generation of Phylogenetic trees
- Prediction of ancestral sequences
- All-vs-all sequence comparison (including ancestral sequences)

Proposed Work Flow

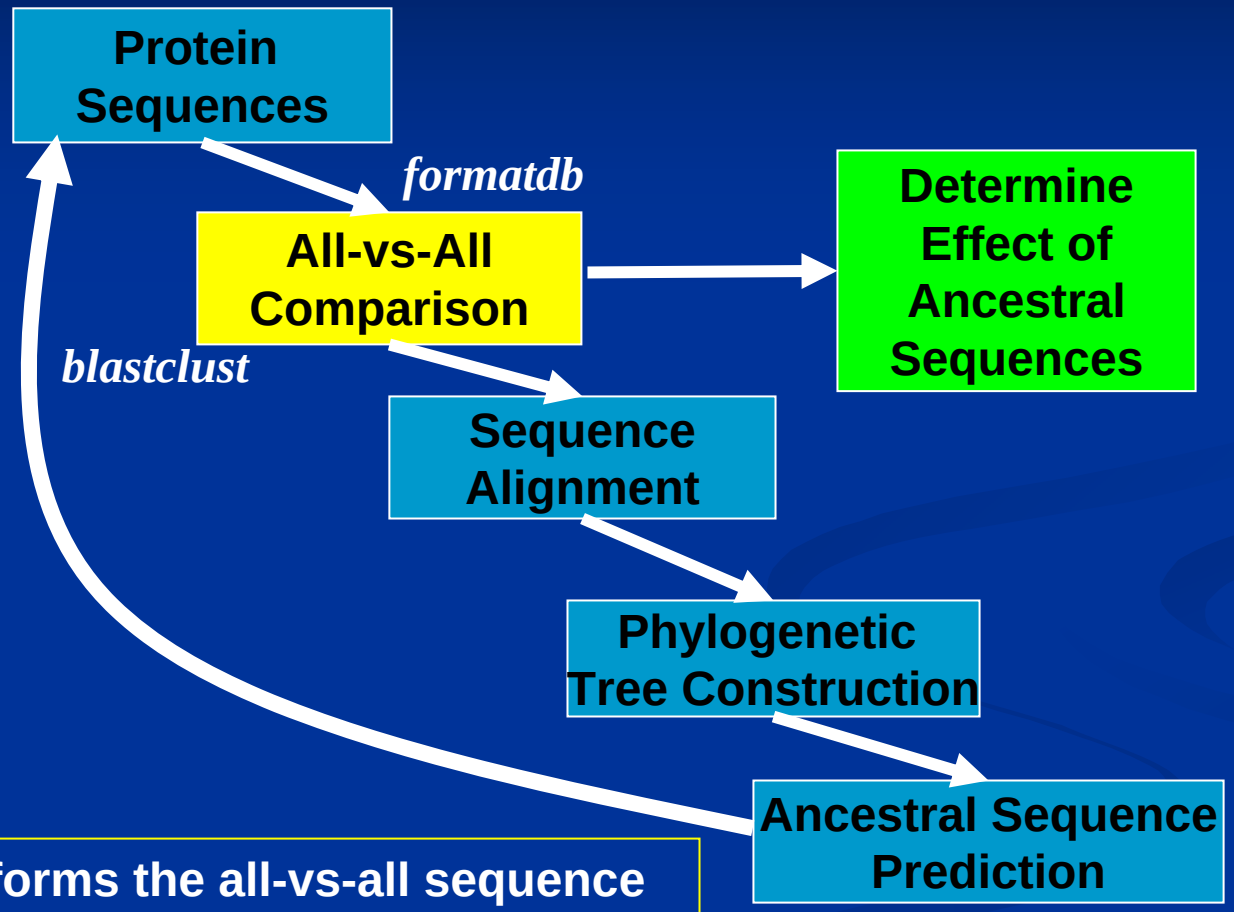


Obtaining Protein Sequences



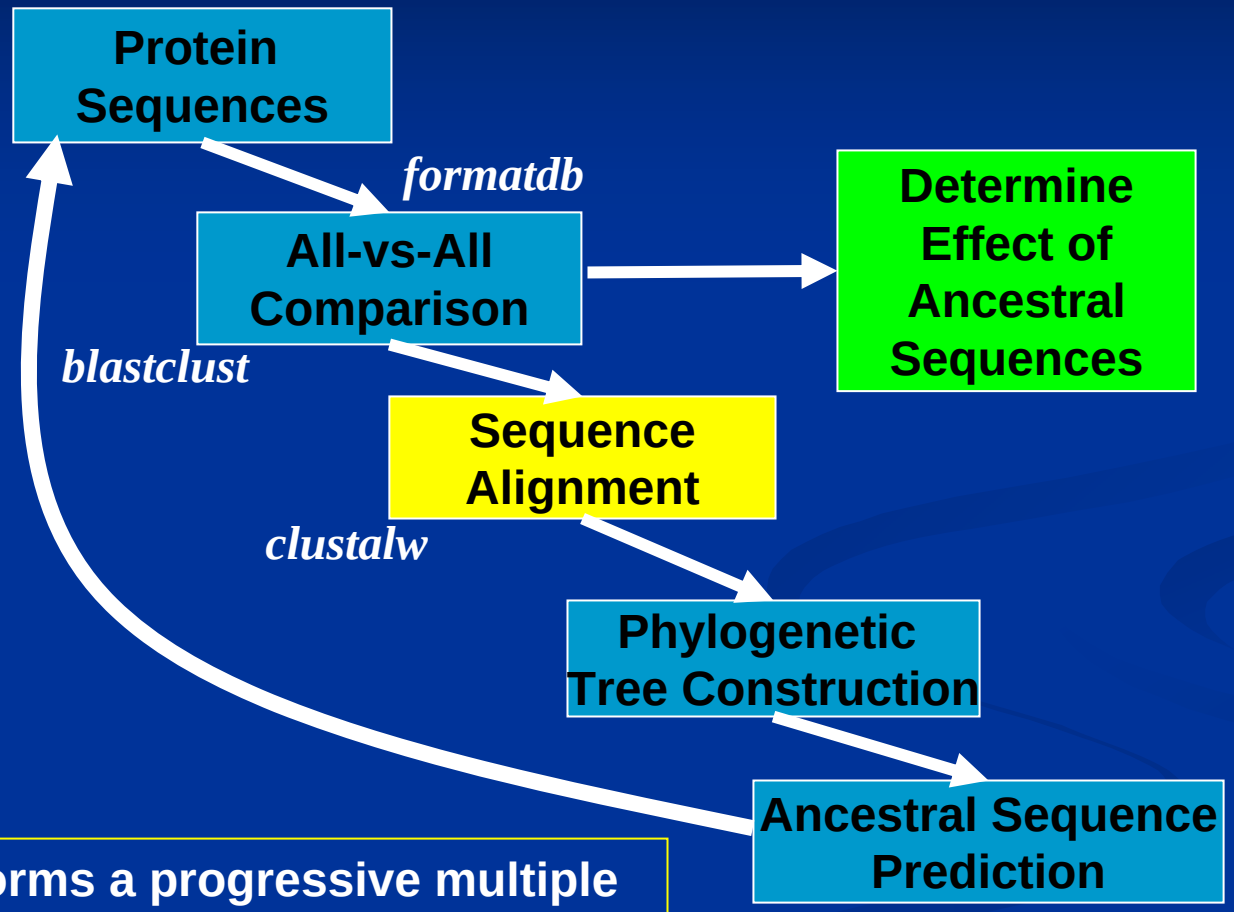
Sequences for the desired species are extracted from NCBI's FASTA formatted nonredundant protein database (nr).

Finding Gene Families



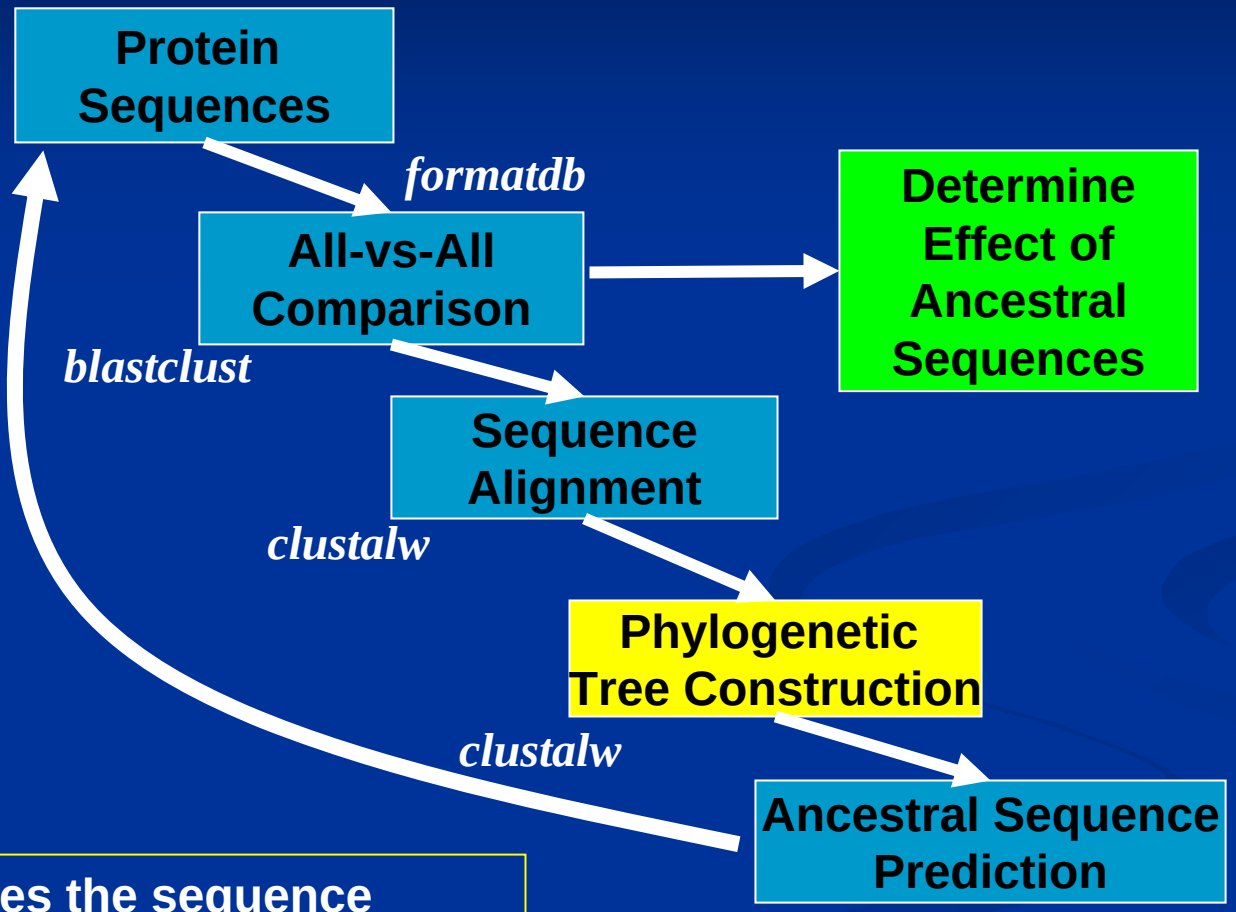
blastclust performs the all-vs-all sequence comparison, returning a cluster list (each cluster represents a gene family or group of homologous sequences).

Sequence Alignment



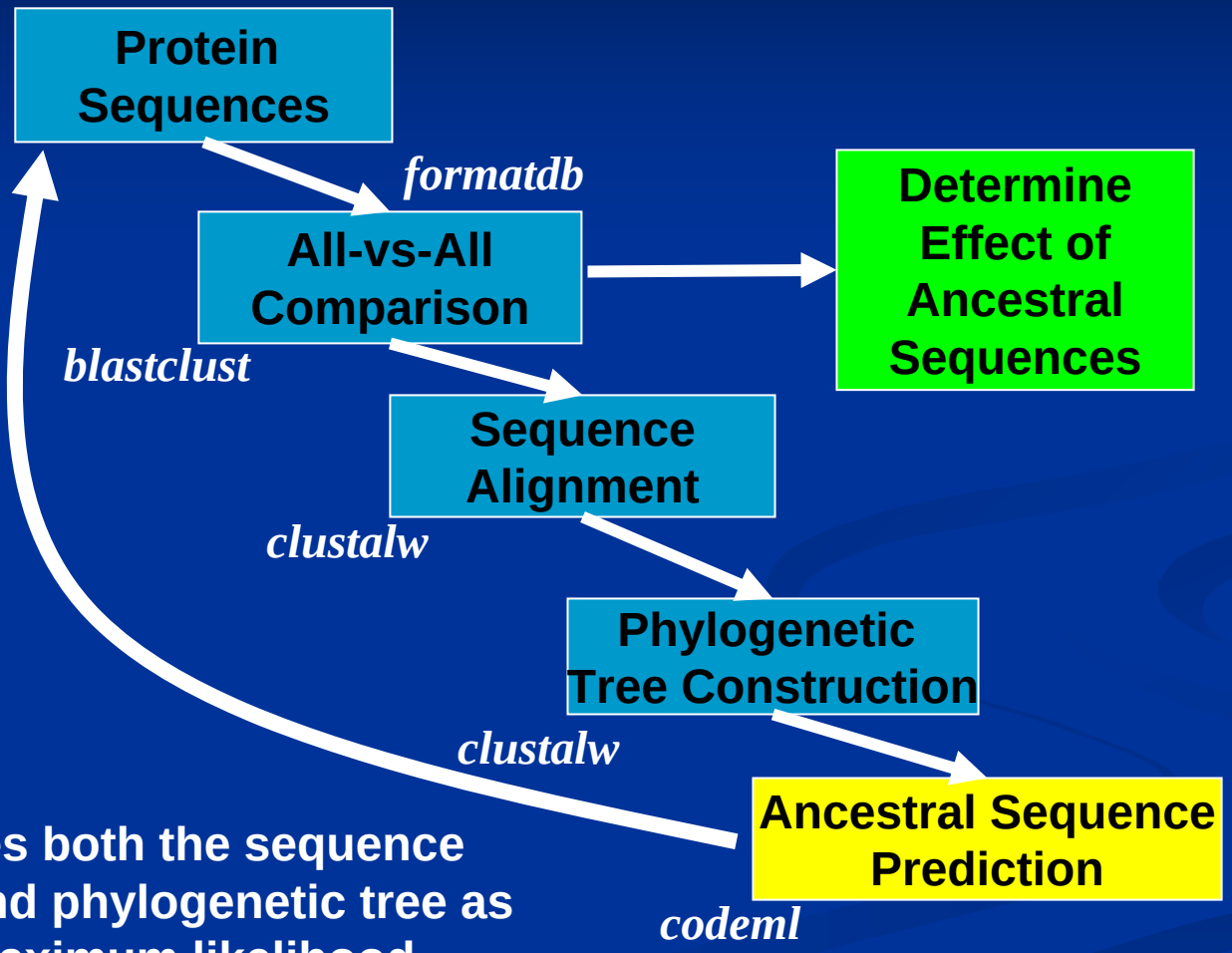
clustalw performs a progressive multiple sequence alignment for each cluster of sequences. *clustalw* is the command line version of web-based CLUSTAL W.

Phylogenetic Tree Construction



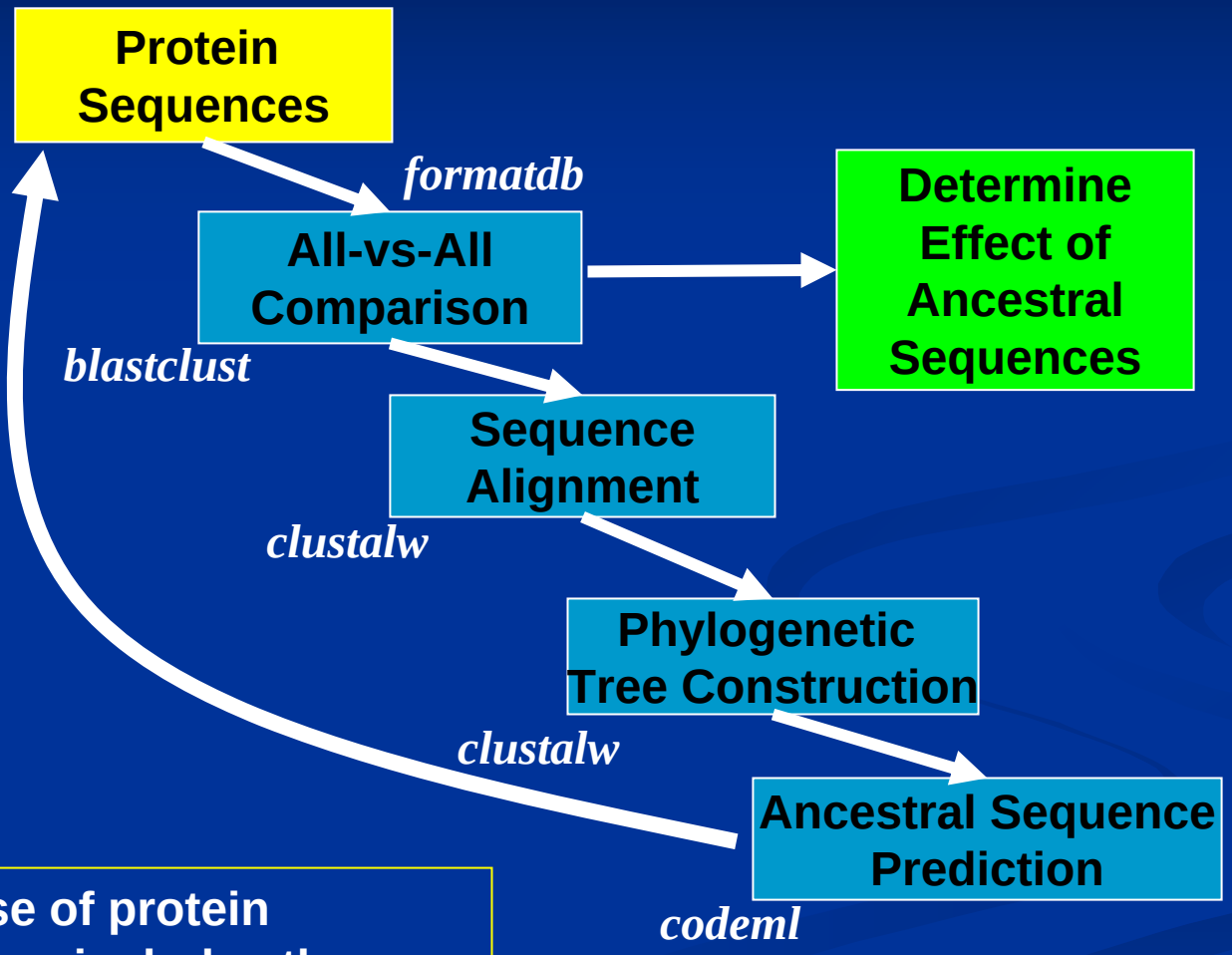
clustalw utilizes the sequence alignment to build a phylogenetic tree via neighbor joining.

Ancestral Sequences Prediction



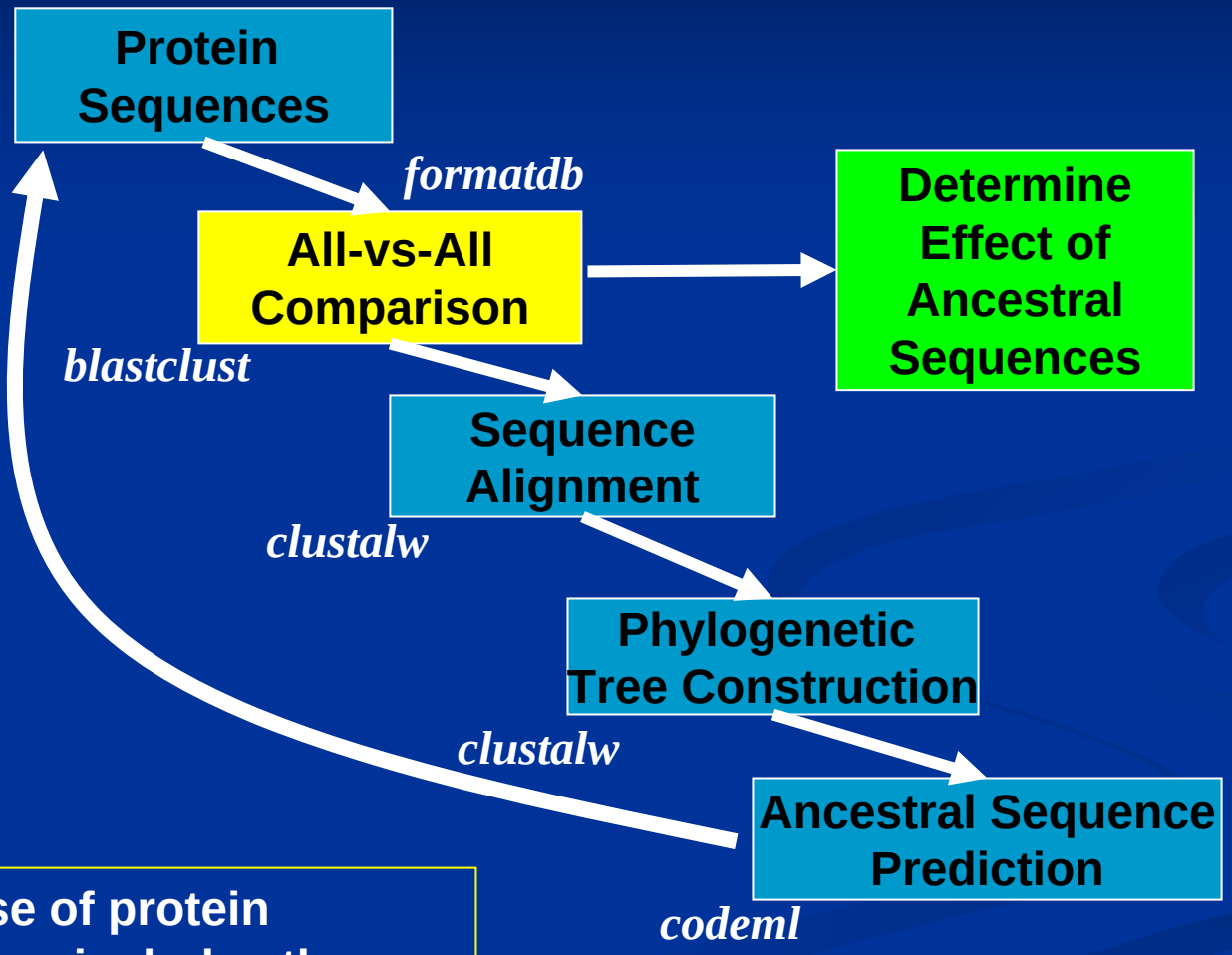
codeml **utilizes both the sequence alignment and phylogenetic tree as input. The maximum likelihood method is used in the reconstruction of the ancestral sequences.**

Extending Sequence Database



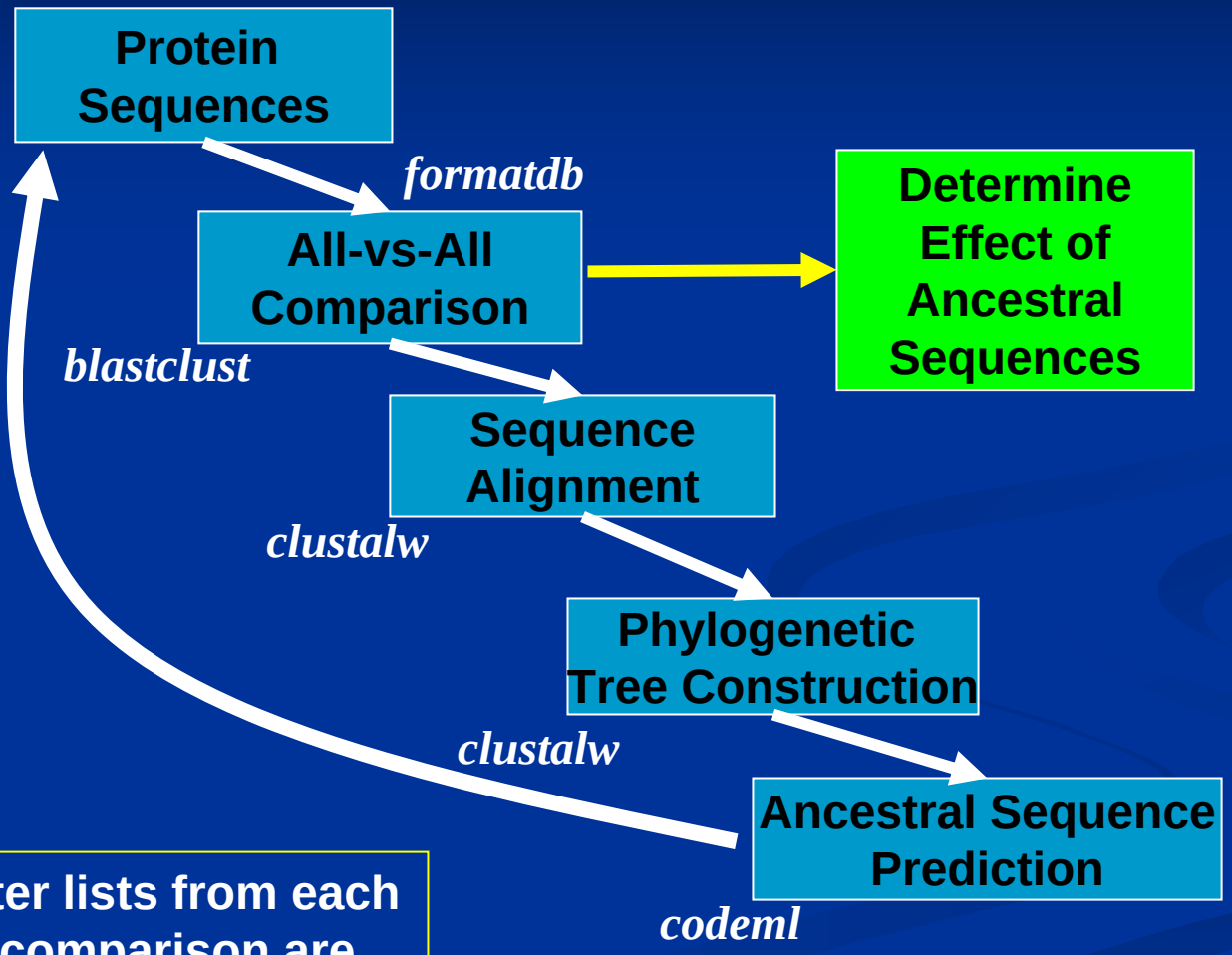
The database of protein sequence now includes the predicted ancestral sequences.

Finding Gene Families (again)



The database of protein sequence now includes the predicted ancestral sequences.

Gene Family Analysis



The cluster lists from each all-vs-all comparison are the basis of this analysis.

Programming Environment

- All of the software developed for this project included:
  - Perl
  - C
  - Batch files executed from the DOS prompt
- All of the development was done under Windows XP. The development environment included:
  - Active State Perl
  - MinGW C
  - Source Edit
- All processing was performed locally
- The external programs utilized (formatdb, blastclust, clustalw, and codeml) were command line versions, compiled for Windows, and executed from the DOS prompt.

Sequences/Species

FASTA sequences from the nr protein database	Sequences	Rel Size
All species as of 01/11/06	**3203752**	
Mus musculus and Rattus norvegicus – rodents	**136403**	**110.35%**
Homo sapiens – human	**123609**	**100.00%**
Mus musculus – mouse	**106205**	**85.92%**
Arabidopsis thaliana - thale cress	**52159**	**42.20%**
Bos taurus – cattle	**36855**	**29.82%**
Rattus norvegicus – rat	**30881**	**24.98%**
Pan troglodytes – chimp	**22873**	**18.50%**
Escherichia coli	**12225**	**9.89%**
Saccharomyces cerevisiae and Schizosaccharomyces pombe – yeast	**15484**	**12.53%**
Saccharomyces cerevisiae – baker's yeast	**9158**	**7.41%**
Zea mays – corn	**3410**	**2.76%**

Processing Time Required

Species Database (nr 1/11/06)	FASTA recs	formatdb	blastclust	clustalw	codeml	Ancestral Seq	blastclust
rodents	136403	<2m	4d 8h 3m	3hr 38m	>4 d	7375	4d 15h 1m
human	123609	46 s	3d 1h 22m	3h 45m	>3 d	5822	3d 10h 36m
mouse	106205	41 s	2d 7h 55m	1h 27m	>3d	4569	2d 14h 19m
thale cress	52159	22 s	10h 6m	<5m	9m	520	10h 01m
rat	30881	10 s	6h 23m	<5m	2h 57m	784	6h 40m
chimp	22873	6 s	3h 48m				
yeast	15484	8 s	1h 25m	<5m	4m	101	1h 26m
baker's yeast	9158	6 s	35m				
corn	3410	1 s	2m	<5	9m	157	2m

Initial Clustering

Species Database	FASTA recs	Clusters size ≥ 6	Largest cluster	Largest Cluster processed	Sequences processed	Avg. # Seq. / cluster
rodents	136403	1815	4051	404	23314	13
human	123609	1050	3210	406	28103	27
mouse	106205	951	4049	402	15958	17
thale cress	52159	183	38	38	1513	8
rat	30881	112	195	195	1346	12
chimp	22873	17	92	NA	NA	NA
yeast	15484	34	34	34	319	9
byeast	9158	33	34	34	313	9
corn	3410	53	48	48	672	13

Second Clustering

Species Database	Ancestral Sequences added	Clusters size ≥ 6	Largest cluster	Sequences processed	Avg. # Seq. / cluster
rodents	7375	1815	4051	30805	17
human	5822	1051	3210	33974	32
mouse	4569	956	4049	20592	22
thale cress	520	183	75	2040	11
rat	784	110	403	2147	20
yeast	101	34	60	420	12
corn	157	53	52	829	16

Comparative Analysis

```
DEBUG: Collecting clusters from "corn_db\clust.out"
DEBUG: Collecting clusters from "corn_db2\aClust.out"
-------------------------------------------------------
Initial clustering (size 6+) included:  672 Sequences
# maintained in reclustering:           672
# lost during reclustering:               0
-------------------------------------------------------
Reclustering (size 6+) included:          829 sequences
# maintained during reclustering:         672
# ancestral added for reclustering:       157
# ancestral maintained during reclustering:  157
# other seq added during reclustering:      0
```

```
  672
+157
  829
```

For this relatively small sequence database (corn), the presence of the ancestral sequences did not change the sequences that were clustered together

Comparative Analysis

```
DEBUG: Collecting clusters from "thale_db\clust.out"
DEBUG: Collecting clusters from "thale_db2\aclust.out"
-----------------------------------------------------------
Initial clustering (size 6+) included:  1513 Sequences
# maintained in reclustering:          1513
# lost during reclustering:            0
-----------------------------------------------------------
Reclustering (size 6+) included:            2040 sequences
# maintained during reclustering:           1513
# ancestral added for reclustering:         520
# ancestral maintained during reclustering:  517
# other seq added during reclustering:       10
   Sequence gained during reclustering: cluster 00019_0001 id 15231449
   Sequence gained during reclustering: cluster 00027_0001 id 15809938
   Sequence gained during reclustering: cluster 00027_0001 id 16648811
   Sequence gained during reclustering: cluster 00046_0001 id 2832540
   Sequence gained during reclustering: cluster 00046_0001 id 2832572
   Sequence gained during reclustering: cluster 00027_0001 id 30690085
   Sequence gained during reclustering: cluster 00075_0001 id 32364494
   Sequence gained during reclustering: cluster 00075_0001 id 32364496
   Sequence gained during reclustering: cluster 00075_0001 id 32364523
   Sequence gained during reclustering: cluster 00010_0018 id 7671458
```

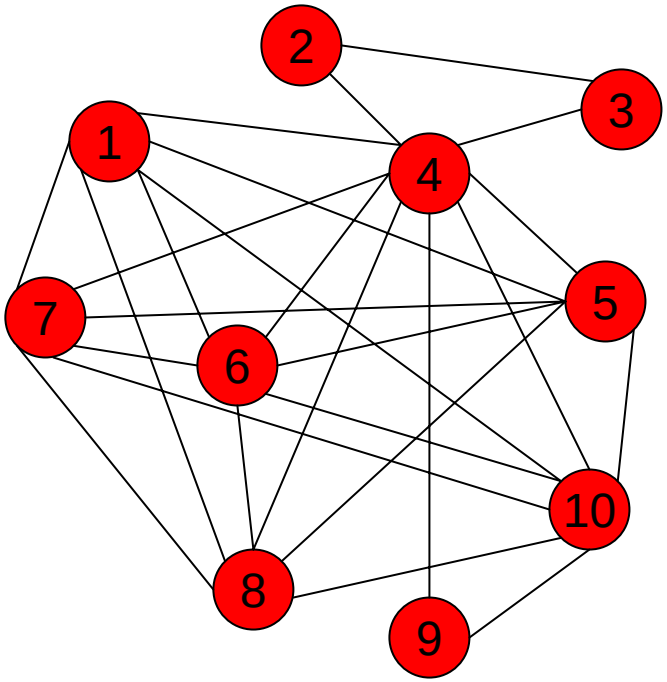
For this sequence database, the presence of the ancestral sequences caused 10 additional sequences to be pulled into gene families. The number of clusters size 6+ was unchanged.

Comparative Analysis

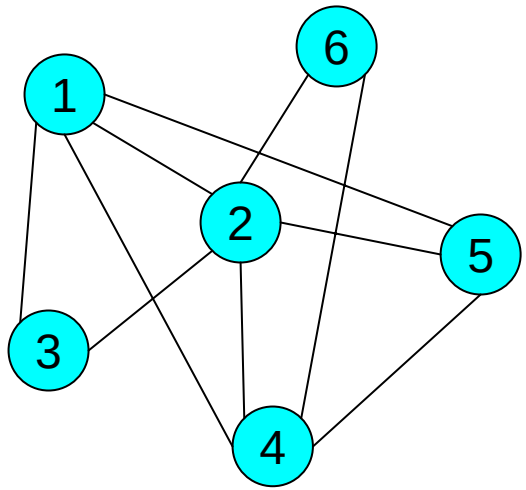
```
DEBUG: Collecting clusters from "rat_db\clust.out"
DEBUG: Collecting clusters from "rat_db2\aclust.out"
-----------------------------------------------------------
Found composite cluster 00403_0001 in "rat_db2\aclust.out"
        Sequences from 00195_0001
                and 00037_0001 in "rat_db\clust.out"
-----------------------------------------------------------
Found composite cluster 00031_0001 in "rat_db2\aclust.out"
        Sequences from 00010_0003
                and 00006_0018 in "rat_db\clust.out"
-----------------------------------------------------------
Initial clustering (size 6+) included:  1346 Sequences
# maintained in reclustering:           1346
# lost during reclustering:                0
-----------------------------------------------------------
Reclustering (size 6+) included:          2147 sequences
# maintained during reclustering:         1346
# ancestral added for reclustering:         784
# ancestral maintained during reclustering:  784
# other seq added during reclustering:      17
   Sequence gained during reclustering: cluster 00032_0001 id 27692470
   Sequence gained during reclustering: cluster 00022_0001 id 27695749 . . .
```

For this sequence database, the presence of the ancestral sequences caused 17 additional sequences to be pulled into gene families. In addition, 2 pairs of clusters were combined to form larger clusters.

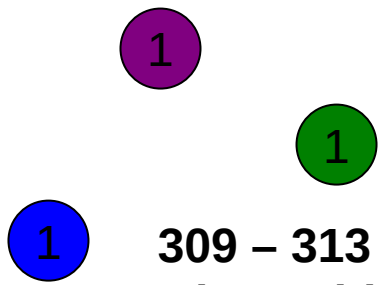
Olfactory Receptor Proteins
Initial Clustering



313 – 320
amino acids

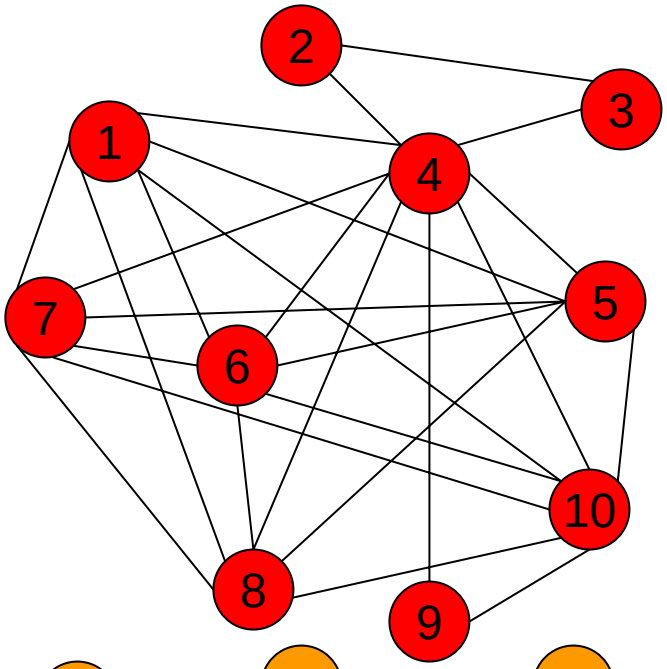


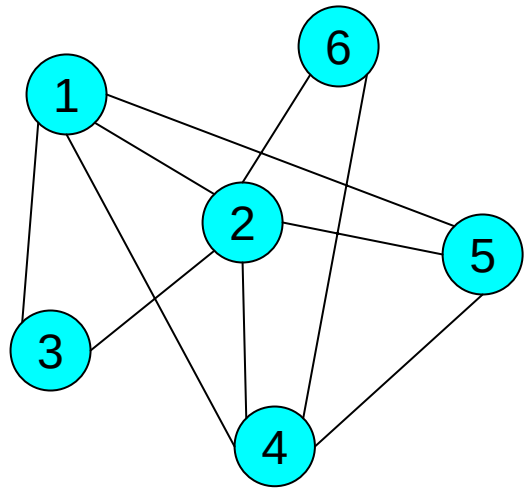
312 – 313
amino acids

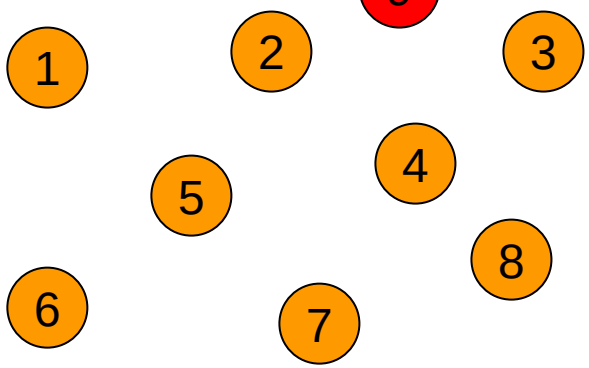


309 – 313
amino acids

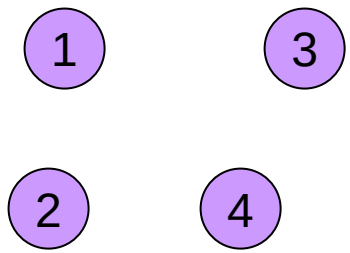
Olfactory Receptor Proteins + Predicted Ancestral Sequences



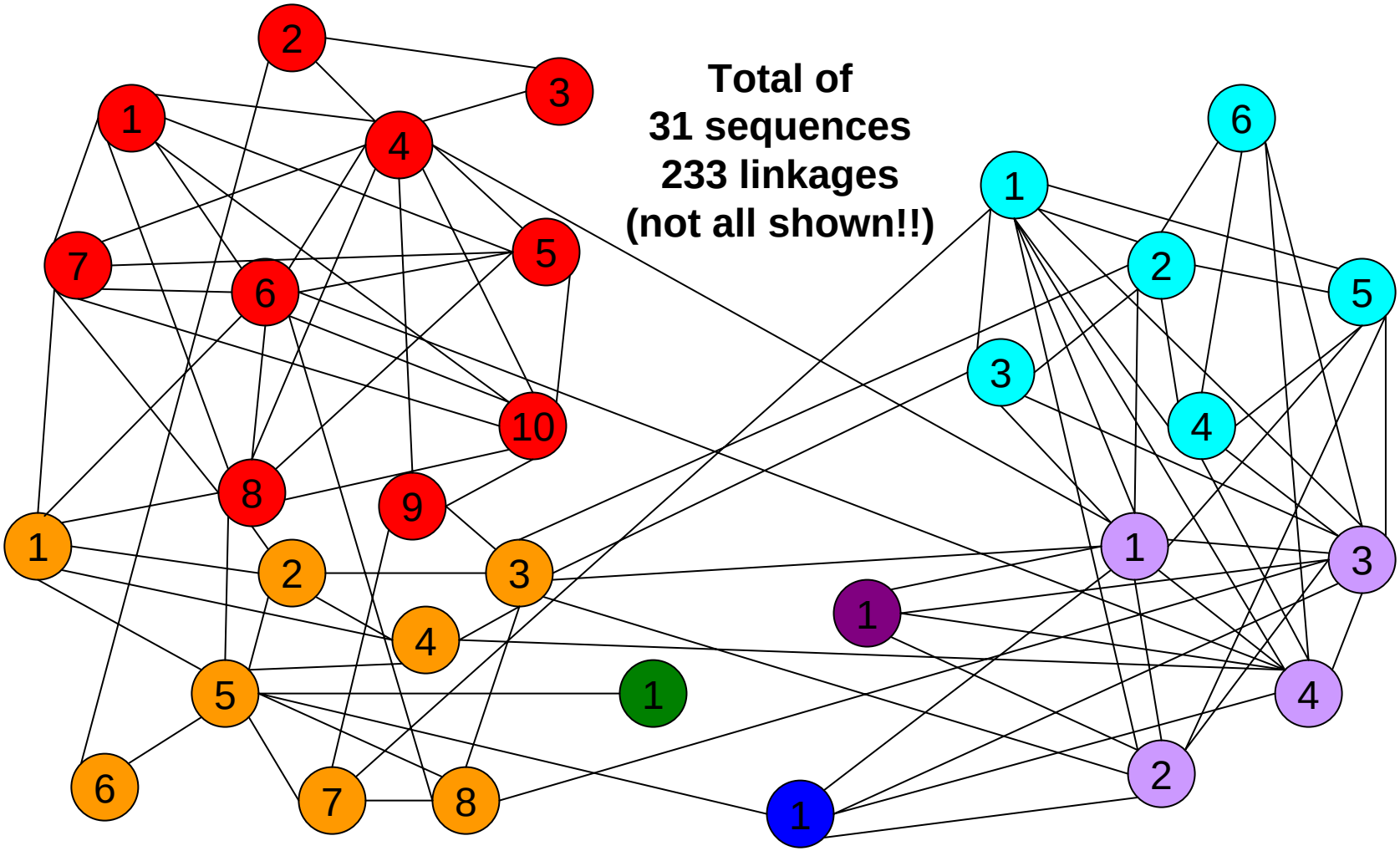




(n – 2) unique ancestral sequences generated for each cluster



Olfactory Receptor Proteins Clustering with Ancestral Sequences



Comparative Analysis

Species Database	Clusters size ≥ 6 Observed Sequences	Ancestral Sequences added	Clusters size ≥ 6 Observed+ Ancestral Sequences	Number Of Composite Clusters	Additional Sequences Clustered	Sequences processed
rodents	1815	7375	1815	11	133	30805
human	1050	5822	1051	7	59	33974
mouse	951	4569	956	7	71	20592
thale cress	183	520	183	0	10	2040
rat	112	784	110	2	17	2147
yeast	34	101	34	0	0	420
corn	53	157	53	0	0	829

Conclusions

- It appears that using ancestral sequences to extend gene families is a viable approach.
- The work presented here is just a starting point, a demonstration of "proof of concept".
- There is much more work that can be done to determine how effective this approach is.

Future Work

- Default parameters were used for *blastclust* resulting in fairly stringent similarity criteria, thus all of the sequences clustered together are annotated similarly in Genbank (eg. Olfactory receptors, immunoglobulin chains, etc.)

- A true test of this approach would be to determine the criteria that clusters all of sequences perceived as homologous in the first pass, then cluster again with ancestral sequences.

Future Work

For each of the steps described: alignment, tree building, and prediction of ancestral sequences we have tried one algorithm.

- How does the accuracy of each of these step effect the result?
- Would another method/algorithm be better?

  eg. Bayesian methods for ancestral sequences

  Parsimony or Likelihood methods for tree construction

References

[Atls 90] Altschul S F, Gish W, Miller W, Myers E W, Lipman D J; Basic Local Alignment Search Tool; *Journal of Molecular Biology* 1990; 215:403-410.

[Atls 97] Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W, Lipman D J; Gapped BLAST and PSI-BLAST: a new generation of protein database search programs; *Nucleic Acids Res* 1997; 25:3389-3402.

[Brid 06] Bridgham J T, Carroll S M, Thornton J W; Evolution of Hormone-Receptor Complexity by Molecular Evolution; Science 2006; 312:97-101.

[Geer 02] Geer L Y, Domrachev M, Lipman D J, Bryant S H; CDART: Protein Homology by Domain Architecture; *Genome Research* 2002; 12(10):1619-1623.

[Heni 92] Heinkoff S, Heinkoff J G; Amino Acid Substitution Matrices for Block Proteins; *Proc. Nat. Acad. Sci. USA* 1992; 89(22):10915-9.

[Jean 98] Jeanmougin F, Thompson J D, Gouy M, Higgins D G, and Gibson, T J; Multiple Sequence Alignments with Clustal X; Trends Biochem Sci 1998; 23:403-5.

[Jone 92] Jones D T, Taylor W R, Thornton J M; The Rapid Generation of Mutation Data Matrices from Protein Sequences; *CABIOS*[1] 1992; 8:275-282.

[Kosi 05] Kosiol C, Goldman N; Different Versions of the Dayhoff Rate Matrix; *Molecular Biology and Evolution* 2005; 22:193-199.

[Lewi 04] Lewin B; *Genes VIII;* Pearson Prentice Hall, NJ 2004.

[1] CABIOS is the former name of the Bioinformatic Journal, Oxford University Press.

References

[Li 01] Li W, Jaroszewski L, Godzik A; Clustering of Highly Homologous Sequences to Reduce the Size of Large Protein Databases; *Bioinformatics* 2001; 17(3):282-283.

[Pevs 03] Pevsner J; Bioinformatics *and Functional Genomics*; John Wiley & Sons, Inc., NJ 2003.

[Remm 00] Remm M, Sonnhammer E; Classification of Transmembrane Protein Families in the *Caenorhabditis elegans* Genome and Identification of Human Orthologs; *Genome Research* 2000; 10(11):1679-1689.

[Tatu 97] Tatusov R L, Koonin E V, Lipman D J; A Genomic Perspective on Protein Families; *Science* 1997; 278(5338):631-637.

[Tigr 05] TIGR; Domain Based Paralogous Protein Families; www.tigr.org ; Annotation Workshop July 13, 2005.

[Thom 94] Thompson J D, Higgins D G, Gibson T J; Clustal W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice; *Nucleic Acids Research* 1994; 22(22):4673-4680.

[Yang 94] Yang Z, Goldman N, Friday A; Comparison of Models for Nucleotide substitution used in Maximum-Likelihood Phylogenetic Estimation; *Molecular Biology and Evolution* 1994; 11:316-324.

[Yang 95] Yang Z, Kumar S, Nei M; A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences; *Genetics* 1995; 141:1641-1650.

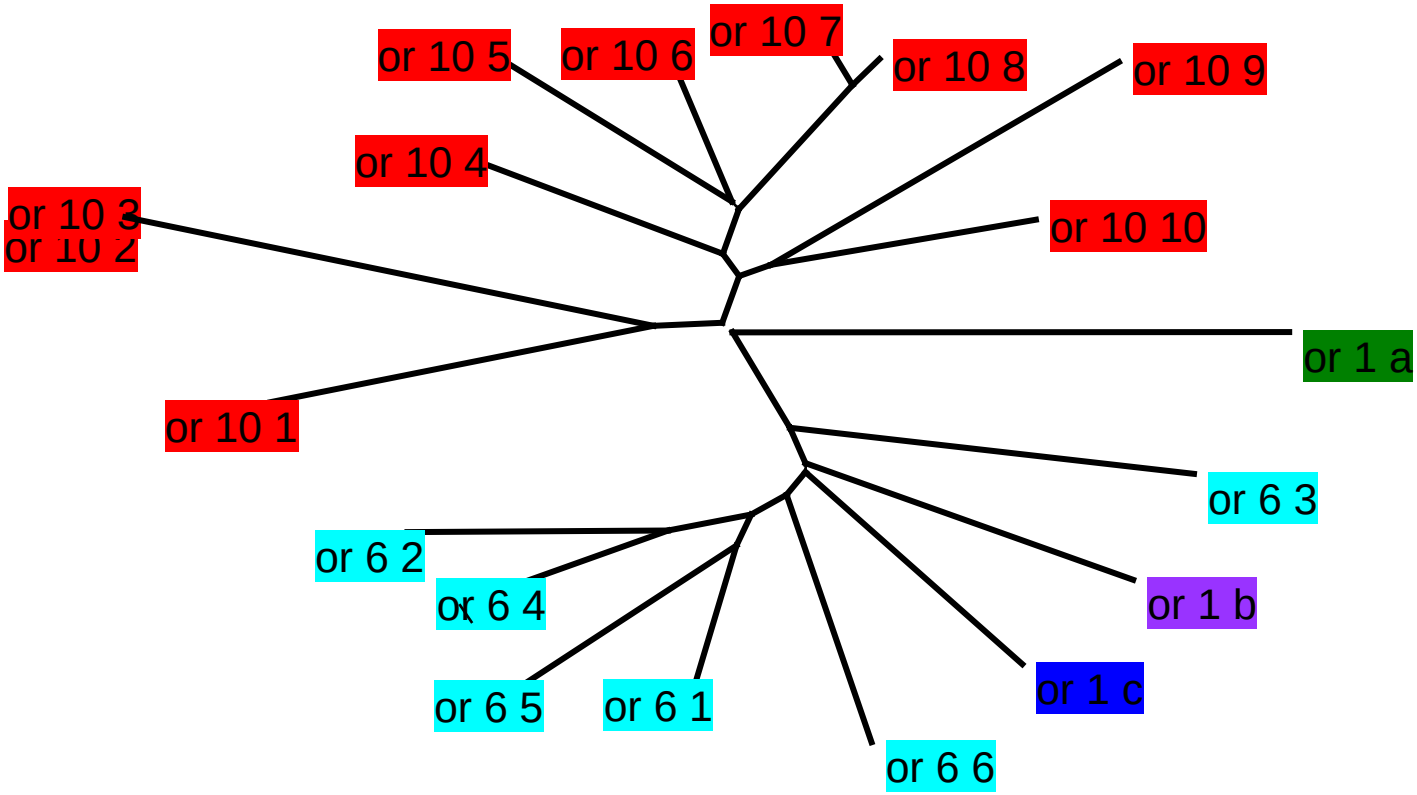
[Yang 97] Yang Z; PAML: a Program Package for Phylogenetic Analysis by Maximum Likelihood; *CABIOS* 1997; 12(7):555-556.

Questions?

Thank you for attending!

Input	Software	Output
NCBI's nr database	spFilter.pl	speciesDB(FASTA)
speciesDB(FASTA)	dbFormatter.pl(*formatdb*)	speciesDB(BLAST)
speciesDB(BLAST)	bClust.pl(*blastclust*)	cluster list neighbor/hit-list
cluster list	countClust.pl	summary of cluster sizes
cluster list	collectClust.pl	cluster files (FASTA)
cluster files (FASTA)	buildClustalBat.pl	bat file(*clustalw*)
cluster files (FASTA)	bat file (*clustalw*)	alignment files (PHYLIP-C) tree files (PHYLIP-C)
alignment files (PHYLIP-C)	modifyAlignFiles.pl	alignment files (PHYLIP-P)
tree files (PHYLIP-C)	modyifyTreeFiles.pl	tree files (PHYLIP-P)
alignment files (PHYLIP-P) tree files (PHYLIP-P)	buildCtlFiles.pl	control files bat file(*codeml*)
alignment files (PHYLIP-P) tree files (PHYLIP-P) control files	bat file(*codeml*)	ancestral sequences files
ancestral sequence files	extractAnSeq.pl	ancestral sequences (FASTA)
speciesDB(FASTA) ancestral sequences (FASTA)	concatenate files together	speciesDB + AnSeq(FASTA)
speciesDB + AnSeq(FASTA)	dbFormatter.pl(*formatdb*)	speciesDB + AnSeq (BLAST)
speciesDB + AnSeq (BLAST)	bClust.pl(*blastclust*)	+ AnSeq cluster list + AnSeq neighbor/hit-list
cluster list + AnSeq cluster list	compareClust.pl	summary of clusters combined and sequences maintained and added
neighbor/hit-list	xnbr.c	neighbor/hit-list(text)

Olfactory Receptor Proteins



Unrooted tree, constructed by Clustal W and displayed via Tree View
Colors show positioning of sequences from each initial cluster.