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

- Protein Sequences
- All-vs-All Comparison
- Sequence Alignment
- Phylogenetic Tree Construction
- Ancestral Sequence Prediction
- Determine Effect of Ancestral Sequences
Sequences for the desired species are extracted from NCBI’s FASTA formatted nonredundant protein database (nr).
Protein Sequences

All-vs-All Comparison

Sequence Alignment

Phylogenetic Tree Construction

Determine Effect of Ancestral Sequences

Ancestral Sequence Prediction

findgene

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

Protein Sequences

formatdb

All-vs-All Comparison

blastclust

Sequence Alignment

clustalw

Phylogenetic Tree Construction

Ancestral Sequence Prediction

Determine Effect of Ancestral Sequences

**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

Protein Sequences

All-vs-All Comparison

Sequence Alignment

Phylogenetic Tree Construction

Determine Effect of Ancestral Sequences

Ancestral Sequence Prediction

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

Protein Sequences

All-vs-All Comparison

Sequence Alignment

Phylogenetic Tree Construction

Determine Effect of Ancestral Sequences

Ancestral Sequence 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.
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

Protein Sequences

All-vs-All Comparison

Sequence Alignment

Phylogenetic Tree Construction

Ancestral Sequence Prediction

Determine Effect of Ancestral Sequences

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

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

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

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

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

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

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 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
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 . . .
Olfactory Receptor Proteins
Initial Clustering

313 – 320 amino acids

309 – 313 amino acids

312 – 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

Total of 31 sequences
233 linkages (not all shown!!)
## Comparative Analysis

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

<table>
<thead>
<tr>
<th>Numbers</th>
<th>1050</th>
<th>951</th>
<th>4569</th>
<th>956</th>
<th>7</th>
<th>71</th>
<th>20592</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numbers</td>
<td>1815</td>
<td>7375</td>
<td>1815</td>
<td>11</td>
<td>133</td>
<td>30805</td>
<td></td>
</tr>
<tr>
<td>Numbers</td>
<td>1050</td>
<td>5822</td>
<td>1051</td>
<td>7</td>
<td>59</td>
<td>33974</td>
<td></td>
</tr>
<tr>
<td>Numbers</td>
<td>951</td>
<td>4569</td>
<td>956</td>
<td>7</td>
<td>71</td>
<td>20592</td>
<td></td>
</tr>
<tr>
<td>Numbers</td>
<td>183</td>
<td>520</td>
<td>183</td>
<td>0</td>
<td>10</td>
<td>2040</td>
<td></td>
</tr>
<tr>
<td>Numbers</td>
<td>112</td>
<td>784</td>
<td>110</td>
<td>2</td>
<td>17</td>
<td>2147</td>
<td></td>
</tr>
<tr>
<td>Numbers</td>
<td>34</td>
<td>101</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>420</td>
<td></td>
</tr>
<tr>
<td>Numbers</td>
<td>53</td>
<td>157</td>
<td>53</td>
<td>0</td>
<td>0</td>
<td>829</td>
<td></td>
</tr>
</tbody>
</table>
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 (e.g. 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


[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.


[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.


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


[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.

Questions?

Thank you for attending!
<table>
<thead>
<tr>
<th>Input</th>
<th>Software</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI's nr database</td>
<td>spFilter.pl</td>
<td>speciesDB(FASTA)</td>
</tr>
<tr>
<td>speciesDB(FASTA)</td>
<td>dbFormatter.pl(\textit{formatdb})</td>
<td>speciesDB(BLAST)</td>
</tr>
<tr>
<td>speciesDB(BLAST)</td>
<td>bClust.pl(\textit{blastclust})</td>
<td>cluster list \newline neighbor/hit-list</td>
</tr>
<tr>
<td>cluster list</td>
<td>countClust.pl</td>
<td>summary of cluster sizes</td>
</tr>
<tr>
<td>cluster list</td>
<td>collectClust.pl</td>
<td>cluster files (FASTA)</td>
</tr>
<tr>
<td>cluster files (FASTA)</td>
<td>buildClustalBat.pl</td>
<td>bat file(\textit{clustalw})</td>
</tr>
<tr>
<td>cluster files (FASTA)</td>
<td>bat file (\textit{clustalw})</td>
<td>alignment files (PHYLIP-C) \newline tree files (PHYLIP-C)</td>
</tr>
<tr>
<td>alignment files (PHYLIP-C)</td>
<td>modifyAlignFiles.pl</td>
<td>alignment files (PHYLIP-P)</td>
</tr>
<tr>
<td>tree files (PHYLIP-C)</td>
<td>modyifyTreeFiles.pl</td>
<td>tree files (PHYLIP-P)</td>
</tr>
<tr>
<td>alignment files (PHYLIP-P)</td>
<td>buildCtlFiles.pl</td>
<td>control files \newline bat file(\textit{codeml})</td>
</tr>
<tr>
<td>tree files (PHYLIP-P)</td>
<td>bat file(\textit{codeml})</td>
<td>ancestral sequences files</td>
</tr>
<tr>
<td>ancestral sequence files</td>
<td>extractAnSeq.pl</td>
<td>ancestral sequences (FASTA)</td>
</tr>
<tr>
<td>speciesDB(FASTA)</td>
<td>concatenate files together</td>
<td>speciesDB + AnSeq(FASTA)</td>
</tr>
<tr>
<td>ancestral sequences (FASTA)</td>
<td>dbFormatter.pl(\textit{formatdb})</td>
<td>speciesDB + AnSeq (BLAST)</td>
</tr>
<tr>
<td>speciesDB + AnSeq(FASTA)</td>
<td>bClust.pl(\textit{blastclust})</td>
<td>+ AnSeq cluster list \newline + AnSeq neighbor/hit-list</td>
</tr>
<tr>
<td>speciesDB + AnSeq (BLAST)</td>
<td>compareClust.pl</td>
<td>summary of clusters combined and sequences maintained and added</td>
</tr>
<tr>
<td>cluster list + AnSeq cluster list</td>
<td>xnbr.c</td>
<td>neighbor/hit-list(text)</td>
</tr>
</tbody>
</table>

37
Olfactory Receptor Proteins

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