Final Report

Submitted by NIMBioS Sabbatical Visitor: Gary W. Stuart
Department of Biology, Indiana State University, Terre Haute, IN
Sabbatical Dates: August – December 2011

Research:

1. Developing and validating an improved SVD-based approach to phylogenomics

The overall goal of this project was to derive information about why a single fly genome (out of twelve) typically failed to cluster in the expected way when applying a data-rich SVD-based phylogenetic analysis method. Working under the assumption that the unusually large amount of sequence data utilized in the analysis included an identifiable subset of sequences providing a significant homoplasious signal, several different "filters" were developed and deployed in an attempt to identify and remove those sequences. This work was done with the aid of Arun Seetharam, a graduate student working in my laboratory at ISU. Some success was observed using a couple of filtering approaches. In one case, proteins with low but increasingly larger projections on all singular vectors were removed in stepwise fashion in an attempt to see which small subsets of proteins, when removed from the analysis, improved the clustering. With this approach, we found that more than one subset of proteins was capable of doing this and that many of these proteins were uncommonly long (> 1,000 residues). Longer proteins sometimes cluster incorrectly because they tend to possess sequences that can associate with multiple different protein families. They are therefore a likely source of homoplasy. In another filtering approach, we identified candidate gene families to remove that could be a likely source of homoplasy. Our best success with this approach was to remove the odorant receptor gene sequences. These are one of the largest families of genes in flies, and they are also thought to have evolved in "non-clock-like" fashion, including multiple deletions of whole subfamilies in some species of flies. These deletions are likely adaptive: resulting from movement away from generalist feeding practices to specialization for particular kinds of fruit.

Finally, the most useful and surprising observation resulting from this work was that improved clustering could come merely by increasing the number of SVD-derived "dimensions" in the analysis. Whereas roughly 400 dimensions (singular triplets) was apparently insufficient to produce the expected clustering, increasing this to 800 dimensions (thereby producing a much more detailed and complex description of the available data) produced the expected result. This suggests that homoplasious signal within the data could potentially be "reinterpreted" with greater accuracy by increasing the complexity of the SVD-based model.

2. Developing an alternative phylogenomic analysis and visualization method

In this effort, NTF (Non-negative Matrix Factorization) rather than SVD would be used to summarize genomes and produce numerous ranked listings of "words" (i.e. sequence fragments) that contribute to significant "features" within the data set (such as large subsets of sequences shared by particular combinations of individuals). However, an additional goal of this work was to adapt the use of a text-mining tool called "Future Lens" (FL), developed previously by Michael Berry (University of Tennessee's Department of Electrical Engineering and Computer Science & NIMBioS) and his graduate student, Andrey Puretskiy.

FL was originally designed to facilitate the direct browsing of compilations of text (like large corpuses of memos or documents), and through a small set of graphical “visualizations,” improve the ability of the analyst to find and examine subsets and elements of documents germane to particular issues or events. NTF can be used to preprocess the information loaded into FL and thereby provide hypothetical themes, subjects, topics, and/or events for consideration by an analyst. By analogy, we hoped to modify FL to produce genome-appropriate “visualizations” to allow the direct examination and analysis of correlated features within large compilations of genome data. Rather than using whole genome sequences, we elected to use genomic subsamples representing less than 1% of each genome. This should provide sufficient information, but would also facilitate the use of more genomes in the future. Significant changes to the FL software were required in order to adapt it for this purpose, and some of these were undertaken as a class project and under Dr. Berry’s direction by a graduate student at UT, Tiantian Gao. The biggest required change was to replace a graphical visualization of the “% document frequency by time-stamp” with a genomic analog, which was in essence “% allele frequency by geographic coordinate”. In other words, a geographic map capable of summarizing allele frequencies with pie charts. This work mostly focused on same-species populations organisms.

The most complex of several initial data sets included 31 D. melanogaster genomes originating from six specific locations within sub-Saharan Africa (http://www.dpgp.org). Our very recent initial observations have tended to reinforce evidence described by others indicating that these six populations are well mixed, even with populations outside of Africa, and hence provide little location-based “distinctiveness” discernible as “features” of significance. Although a bit premature relative to current progress, we hope to improve upon both our data sets and software so as to potentially present a “prototype” at a future conference perhaps as early as the end of the academic year (April-May).

3. Developing a “microinversion” detector to uncover rare genomic changes (RGC’s) for phylogenetic analysis

RGC’s in general and microinversions in particular are currently of interest because they are best candidates for “near perfect” phylogenetic characters due to multiple alternative states and very slow conversion rates. There is a very broad size range definition for microinversions (5bp to over 50kb), and it is currently believed that the smallest inversions are generally the most frequent class. I have decided to look systematically for microinversions on the smaller end of the scale – those between 10 and 30 bp in length. I have decided to take a mostly “non-alignment” approach. The downside of this approach is that many “weak” microinversions would likely have to be ignored due to the lack of statistical support for the inversion (vs. an indel/mutation process). With advice from Brian O’Meara (UT’s Department of Ecology and Evolutionary Biology & NIMBioS), I am building a prototype pairwise microinversion detector (in PERL) and am beginning to use it to examine various fly genomes in pairwise combination (Dmel, Dsim, Dsec). Dsim and Dsec diverged roughly 5 MYA, while Dmel diverged from them roughly 10 MYA, hence they are closely related and share enough sequence to presumably make pairwise microinversion detection possible. In an early trial on only small subfractions of these genomes, roughly 3-5 examples of microinversions shared between these species were identified with the unfinished tool. As the tool improves, we will scale up to full genome comparisons and begin to provide more accurate estimates of the frequency of observable microinversions within genomes. It will also be of interest to uncover large numbers of microinversions and to determine where in the genome microinversions tend to be found and in what sequence context.

Meeting and Working Group Participation:

1. NIMBioS Working Group on Species Delimitation: Sept. 22-24

This working group met for the second time during my stay at NIMBioS, and its members graciously agreed to my request to participate. I participated in several high-level discussions, and I was exposed to some extremely valuable ideas and approaches. In the near future, I intend to increase my involvement with large population genomic data sets, and hence I very much appreciated the opportunity to learn about the advantages and limitations of the latest software tools used in this line of research (e.g. Structure and BPP), as well current plans for improving these tools.

2. Oak Ridge National Laboratory CMS-Phylogenomics Cross-fertilization Meeting: Nov. 2

This meeting was organized and hosted by Tom Potok, ORNL-Applied Software Engineering. Dr. Potok's CMS group (about 15 researchers) had an interest in large scale data-mining of medical documents, and was interested in learning more about ongoing collaborative work between Dr. Berry and myself. The intention was to potentially uncover "overlap" areas and "spur some new ideas." The meeting began with back-to-back slide presentations by Dr. Berry and me, followed by both explanatory and exploratory discussions.

3. Weekly Research Meetings - Gavrilets/O'Meara/Gilchrist Group: Tuesdays @ 2 p.m.

This group considered many topics of relevance to the research of its principle participants, which revolved primarily around mathematical modeling of organisms and their behavior. Examples ranged from allele frequency changes in asexual populations, the evolution of robustness and adaptability, the behavioral utility of overconfidence in populations, and the effect of horizontal gene transfer on the coupling of organelle and nuclear genomes.

4. PDG (Phylogeny Discussion Group) Meeting – O'Meara/Matheny/Hulsey/Fordyce: Fridays @ 10 a.m.

This group focused on the fundamental principles and theory of phylogenetic analysis. A high fraction of the discussion revolved around coalescent theory and maximum likelihood methods, particularity as these related to the estimation of phylogenetic relatedness. These sessions were very useful for bringing me up to date on the theory and practice of phylogenetic estimation.