CS 75 Final Project Proposal

Varun Ravishankar and Alex Welton

October 22, 2013

Background: An interesting problem that has arisen in the field of bioinformatics is the desire to classify biological similarities across species. One methodology for determining a potential biological connection is phylogenetic profiling. This technique involves using homology to determine presence or absence of specific genes across the set of genomes. The intuition is that genes located (or not located) in the same pattern across genomes indicate a biological connection between these genomes.

Main Concept: we intend to utilize a two-step process to predict what features of n given genomic sequences evolved under similar conditions. This is accomplished by building a phylogenetic profile for each protein in a given sequence (up to a parameter m proteins), then using a weighted distance algorithm to identify similarities between proteins. These similarities are used as a proxy for similarity between genomes, and, based on the conjecture that similar environmental conditions yield similar evolutionary adaptations, used to make a rough prediction of the likelihood that the n genomes evolved under similar conditions. Results are then output using a pleasing and easy-to-follow GUI.

Dataset: we plan to implement phylogenetic profiling on a test set of genomes. We will establish a base set of genomes across several species using the databases provided by the National Center for Biotechnology Information (NCBI)[1]. We will evaluate the accuracy of this initial profiling largely based on the known biological relatedness of the species in question, and plan on selecting a small set of organisms (initial n around 5) with relatively small genomic sequences for initial testing as done in Psomopoulos et al.[2].We will then store this initial dataset and select several new genomes BLASTed against genomes in the initial test set. This way we will be able to evaluate the effectiveness of the profiling against known similar sequences. We plan on running this with n around 20 based on [2] and [3].

Implementation: We will utilize the basic profiling model established by Pellegrini et al.[3]. We plan on making some general improvements to distance calculations and potentially implement some of the normalization techniques implemented in Psomopoulos[2]. We intend to use Python 3 as the core basis of our application due to its extensive third-party graphics libraries and mature object model. However, we worry that the algorithms required may take a suboptimally long time to complete their computations, especially when m is increased past a low value. We believe that using a higher m value will yield far more accurate results at the expensive of drastically higher computation time. If this turns out to be the case, we plan to use the Cython module to precompile the computationally intensive portions of the algorithm using C types, which research into average speed gains suggests will improve running time by a factor of at least 3x. Additionally, we intend to have a recompute-with button that allows users to change parameters for the computation without restarting the entire program. Once we have the more basic functionality running, we intend to optimize this re-computation process by caching results instead of re-computing everything.

Testing: we intend to test the functionality of our program using genomic sequences known to have evolved under similar conditions as well as genomic sequences known to have evolved under drastically different conditions. Once we are sure that the most basic functionality of the program is correct (i.e. closer genomes get higher scores), then we will focus on tweaking for maximum accuracy and speed optimization.

Work Distribution:

Varun: has begun by working on the code for creating a correlation matrix corresponding to individual proteins across each genome, and is currently attempting to handle I/O to best represent the genomes. Varun will then move on to the matrix generation algorithm. Once the program functionality is working in the basic sense, Varun will begin working on GUI components for the final program.

Alex: has begun investigating metholodolgies for implementing the phylogenetic profiling code to define connections between proteins. Alex has begun coding the main project framework as well. Once the program functionality is working in the basic sense, Alex will begin working on optimization (potentially using Cython) and re-caching results for the re-compute functionality.

Co-Developed: once correlation matrices and phylogenetic profiles are created, a jointly developed algorithm will combine these results into a final coevolution prediction for the set.

Bibliography: (1) http://www.ncbi.nlm.nih.gov/
(2) http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0052854
(3) http://www.pnas.org/content/96/8/4285.long