CMSC858 Project Proposal

Poorani Subramanian

April 22, 2008

1 Introduction

Biologists, medical researchers, and other scientists routinely use BLAST to do protein sequence alignment. In order to do protein alignments, BLAST uses a substitution matrix which holds the values of the scores (or penalties) associated with aligning one amino acid with another. These substitution matrices can be calculated in various ways. One of the most popular family of matrices is the BLOSUM family. These were developed by Steven and Jorja Henikoff in the early 1990s [1]. The matrices were based on the BLOCKS database which contains sequences of groups of shared motifs (which are the blocks). They used the alignments in this database to calculate the frequency of various amino acid substitutions which became the elements of the matrix. Each matrix is constructed based on a minimum required level of sequence identity match for the sequences in the database. For example, BLOSUM62, which is the default matrix for BLAST on the NCBI website, is based on sequences that match with at least 62% identity.

In a recent paper, researchers at MIT and the Broad Institute found that the source code published with the original BLOSUM paper had some errors [2]. The matrices that were published and computed along with the original paper were not what we would compute if we followed the Henikoffs' algorithm precisely. They were actually quite different. For the past 15 years, NCBI-BLAST (and probably many others) have been using these "incorrect" matrices. The researchers wanted to know if there was a significant performance difference between the original matrices and the ones that they recalculated by following the algorithm exactly. They, in particular, studied the BLOSUM62 matrix because it is the most widely used. They tested both matrices using both BLASTp and a Smith-Waterman algorithm to do a search for protein homologues from various protein superfamilies. Surprisingly, the original matrices outperformed the recalculated ones. In fact, if one was to follow the algorithm exactly, certain columns in the matrix could be shuffled without a change in results. The original matrix (not correctly calculated) can not be shuffled in this way, however the matrix as published is the best of all the possible column combinations. This accumulation of good luck has lead to BLOSUM62's popularity.

The question still remains as to why the original matrices perform better and whether or not this is a characteristic of the whole family of BLOSUM matrices. The purpose of my work would be to do additional comparisons among other matrices in the family to aid in finding the answer.

2 Proposed Work

I would like to do a similar study on the BLOSUM45 and BLOSUM80 matrices, both of which are popular among researchers. I would recompute the matrices according to the original algorithm, and then use them with BLASTp. As my comparison dataset, I will use the ASTRAL database which is maintained by the Brenner Computational Genomics Research Group at Berkelely. It is a human curated set of proteins which are categorized according to structure, function, and sequence into classes, superfamilies, and families. We would expect proteins which are closely related to be in the same family, so a search done with the BLOSUM80 matrix should yield homologues that are in the same family. Similarly, we would expect more distantly related proteins to be in the same superfamily or class, so for a search with the BLOSUM45 matrix, the results should be proteins that are at least in the same class. I would then compare the search results given by both the original matrices and the recalculated ones. I would also like to compare characteristics of the BLOSUM matrices such as their eigenvalues and eigenvectors as many researchers use these characteristics

to infer evolutionary conclusions. For example, one can compare eigenvectors across different substitution matrices with different percent identity to determine which protein traits are most conserved. It could be interesting to see if there is a difference, and if scientists should be paying attention to these values at all.

I think this project would be of interest to biologists who often use these matrices, and give some insight into what we know (or do not know) about the evolution of proteins.

References

[1] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *National Academy of Sciences of the United States of America*, 89(22):10915–10919, 1992.

[2] Mark P. Styczynski, Kyle L. Jensen, et al. Blosum62 miscalculations improve search performance. *Nature Biotechnology*, 26(3):274–275, 2008.