Robert Carrillo
12_12_03: MBB452 Final Paper

TOPIC 1

The first methods of sequence alignment use dynamic programming algorithms and focus on the comparison of a limited number of sequences. One of the first methods was that developed by Needleman and Wunsch (NW) in 1970. This method concentrates on the global alignment of two sequences. NW is based on an iterative matrix technique. The two sequences are represented in a 2-dimensional array with one sequence along the columns and the other along the rows. The matrix is filled with similarity values at the appropriate positions and a pathway is created that represents the optimal alignment. In contrast, the Smith-Waterman (SW) method, created in 1981, sacrifices the global score and finds local regions of significant similarity. It results in the best and longest local sequence pairs that maximize similarity.

Another set of methods rely on substitution matrices for protein alignments, whereby similarity and not just identity are scored. These matrices take into account the likelihood of change from one amino acid to another as a result of evolution. The Dayhoff "Percent Accepted Mutation" (PAM) matrix, developed in 1978, is used for this purpose. BLOSUM in another method based on scoring similarities, but this technique is based on a large set of diverse proteins.

These aforementioned methods, however, provide a means for computing alignments with a limited number of sequences since they can be computationally complex with regard to speed and memory storage. There is another group of techniques referred to as heuristic alignment algorithms. These methods may be less sensitive, but their accuracy is still quite good and their speed is increased greatly. The FASTA method created by Pearson and Lipman in 1985, introduced the comparisons of groups of letters instead of analyzing individual pair-wise sequences. The length of the "word", or sequence, that is used for comparison, directly correlates with the speed at which the computations are performed: more letters in the word, less speed, but greater sensitivity and vice versa. This method, like SW, results in local alignment optimization. BLAST, created in 1989, is another of these techniques and is based on the BLOSUM62 amino acid substitution matrix. There are also various updated versions of both FASTA (FASTA3) and BLAST (BLAST2, PSI-BLAST, and ψ-BLAST) that have been further optimized.

In addition, an abundance of other programs exist aimed at multiple sequence alignment based on various methods such as hidden Markov models (HMMs), profiles, and motifs. The HMM, for example, is a statistical model that takes into account every possible combination of matches, mis-matches, and gaps in order to produce an alignment. This model is created by using a known set of data as training, and then applying it to the database to produce multiple sequence alignments. The Sequence Alignment and Modeling Software System (SAM) is a program based on HMMs.

Figure 1

Sequence A:	L	•	A	R	D
Sequence B:	T	•	A	R	D
Sequence C:	L	I	A	R	-
Sequence D:	L	I	A	R	D

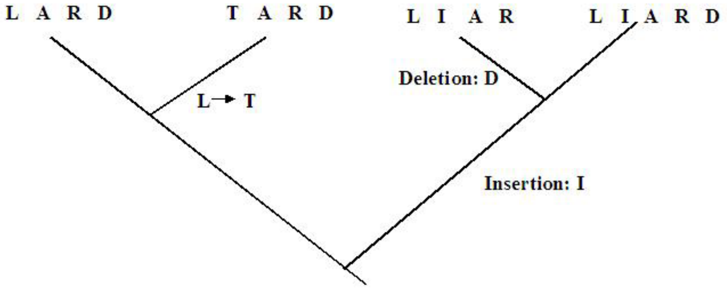

Fig. 1: Homologous proteins with various mutations. Through evolution insertion, deletions, or substitutions can occur, but the function of the protein may remain the same, depending on the extent of mutation.

One function of BLAST is to search a database for sequences that share a high similarity. For example, say I would like to gain some insights into the function of a novel protein by identifying homologous proteins (Figure 1). If my protein of interest is very similar to another protein that has previously been identified, it is likely they also share similar functions, though this is not always the case. Another use for multiple alignment programs such as BLAST or SAM is for the prediction of specific probes that can be used for PCR to identify other members of the same family of genes in either the same or other organisms.

TOPIC 2

Now that we have completely sequenced the genome and identified the open reading frames (ORFs) of our organism, the next step will be to investigate the functionality of the ORFs. The classical way would be to perform various experiments including mutational analysis. However, with the advent of rigorous computational techniques in functional genomics, the most logical procedure to follow will be to use the data collected from other organisms and apply specific techniques to infer information from the ORFs of the new organism.

Since ORFs represent the DNA sequence of the gene, we will first convert the DNA to amino acids using the ExPASy translation tool. We will then begin the search for other proteins that have high homology to the translated ORFs. Several techniques can be applied for this purpose, including FASTA and BLAST. Both of these methods allow multiple sequence alignments to search through the vast sequence databases. If we find that the novel proteins have high homology to previously identified proteins, we can conclude, with certain confidence, that the proteins are homologous and share similar function. However, functional assignment based merely on homology can lead to problems, thus, other functional tools need to be applied in parallel.

An important piece of data that will be investigated to give insights not only to the function of the proteins, but also to the analysis of the organism in general, is protein folds, or secondary structure. Various databases exist with this type of structural information, including SCOP, FSSP, and CATH. It has previously been shown that organisms use various protein folds in their proteins and that the usage of folds can be a characteristic of that organism. The folds of the novel protein sequences can be identified by using the Protein Sequence Analysis Server (PSA) or GOR IV method, which predicts the secondary structure of unknown proteins with no known homologs. In addition, another program can be applied to the folds in the individual proteins to determine possible function. A supervised clustering or a HHM can be created with previous data that correlates specific folds and combination of folds with specific function. This program can then be used to predict the function of the ORFs based on the predicted folds. The types of proteins present will allow further analysis of the new organism.

There are a plethora of other databases that contain functional classes such as proteins involved in transcription, translation, DNA repair, cellular metabolism, phosphorylation, and many others. Other databases can also serve as predictors of function, such as cellular localization, protein-protein interactions, and expression profiles. The amino acid composition correlating to the specific class has also been identified. All of these biologically relevant features obtained from other organisms and studies, will then be used to train a neural network or HHM. A small subset of the data will be used to validate the method, and after confirmation, the model will be applied to the new sequences. The predicted functional class of each ORF will be reported and the functional protein composition of other organisms will be compared to our new organism.

Figure 2

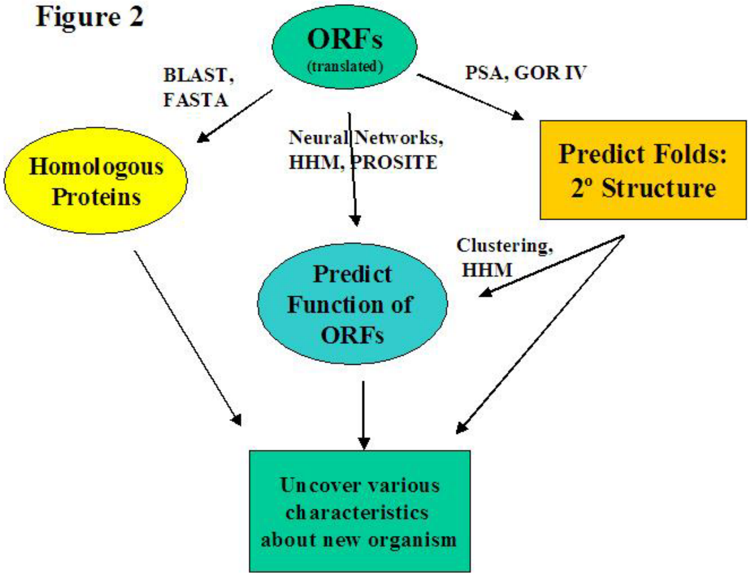

Fig. 2: Flow-chart of steps in identifying functions of ORFs and characterizing a new organism

Figure 2 illustrates the procedure taken herein to conduct a functional analysis of a recently sequenced genome and learn about the organism from which the genome originated. Through the parallel prediction of homologous proteins, protein secondary structure, and functional classification of the ORFs, a wealth of knowledge will be obtained with regard to the new organism. We will be able to infer possible types of metabolism, natural habitat, method of movement, and many other characteristics about the new organism that have been identified previously in other organisms.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol. 215:403-410

Hadley C, Jones DT (1999) A systematic comparison of protein structure classifications: SCOP, CATH, and FSSP. Structure Fold Des. Sep 15;7(9):1099-112

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. 48:443-53

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci. Apr;85(8):2444-8

Smith TF and Waterman MS (1981) Identification of common molecular substances. J. Mol. Biol. 147:195-97