Jennifer Lee
12/12/03
MB&B 452a
Final Paper Topic #1

Sequence alignment techniques have been divided into two categories—global and local. For global alignment, the Needleman-Wunsch method is used, generally for a bigger string of amino acid sequence. This method was pivotal because it was the first automatic method to find a global alignment. It was developed in 1970 and uses dynamic programming to determine the optimal alignment between sequences according to the highest score for the whole regardless of regions of local similarity. On the other hand, the Smith-Waterman method is used for local alignments, where segments are the foci of determining the final score, regardless of the overall score. Both employ similarity alignment matrices (see corresponding figure below) that can incorporate insertions and gaps, which introduce penalties that affect the score by a fraction. These techniques may be used in modern biology in order to predict secondary or tertiary protein structure by finding identifying certain motifs simply by analysis of the amino acid sequences. Also, they may be used to discern gene homology among organisms. Throughout sequence alignment and much of bioinformatics, not only is computer technology important, but also statistical principles are used. After these two early methods, more powerful multiple sequence alignment methods followed in 1987.

Researchers today are well acquainted with the more recently developed methods of sequence alignment. Each has its own specialty and knowing these specialties can save researchers a lot of time and energy. One of the most popular ones is FASTA, which was developed by Bill Pearson. With FASTA, there is a hash table of short words in the query sequence. It can be identified by its use of K-tuple, which determines the amino acid string size. For example, “k-tup 5” denotes five amino acids. Using the query sequence, FASTA goes through the database and finds matches for it. BLAST, another popular sequence alignment method, is similar to FASTA, but is slightly different. It employs High Scoring Segment Pairs (HSPs), which are extensions of hash hits that are also found in FASTA. If the total score does not increase, then the extension is terminated. In addition to this form of basic BLAST, there are also variations on the technique, resulting in the development of BLAST2 and Ψ-BLAST. BLAST2 allows for gapped extensions between diagonals of two hash matches. Ψ-BLAST offers yet a different method of sequence alignment and has more “bell and whistles” than its basic counterpart. It is more sensitive, thus taking longer to go through the database at hand, due to variable parameters that define thresholds for the profile that is developed and subsequently used by this particular protein database search program.

Aside from these programs that are widely used throughout biology, there is the less well-known Hidden Markov Model (HMM). This is operated based on probabilities that each position in the sequence will have a particular residue and thus produce a particular sequence. There are also BLOSUM, CLUSTAL alignments, and a few other sequence alignments methods that can be used to analyze amino acid sequences. Granted, each method has its advantages and disadvantages that are often the basis for selection on which is used for data analysis, but each can provide a unique facet of information from the sequence.

Figure

Similarity matrices: The basis of NW and SW methods

	A	D	R	M
A	1			
D		1		
F				
R			1	

	A	D	R	M
A	3	1	0	0
D	1	2	0	0
F	1	1	0	0
R	0	0	1	0

Optimal alignment:
A D – R M
A D F R –

Jennifer Lee
12/12/03
MB&B 452a
Final Paper Topic #2

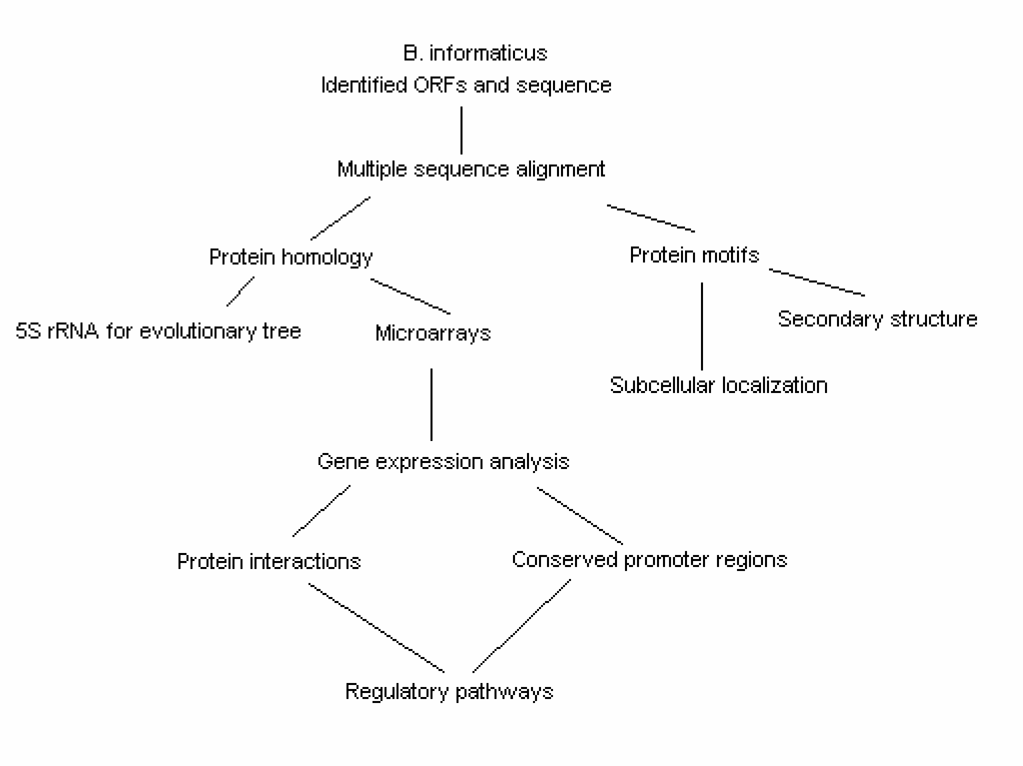
With limited information such as an organism's sequence and ORFs computational functional genomics allows for the inference and prediction of information that would otherwise take a large amount of time to determine experimentally. In the case of this new archaeal organism, *B. informaticus*, methods that we use could allow us to identify and understand cellular proteins and functions.

In order to identify proteins within the organism, we could use multiple sequence alignment programs on the sequence of *B. informaticus* against those of other known organisms. We could look for global alignment so that we could identify homologous proteins across a variety of organisms so the amino acid sequences would be similar. Perhaps we could use BLAST to do this search and the results would not only allow us to see general protein homology, but also the amount of homology between the 5S rRNA of *B.informaticus* and of other organisms. This information would place it on the evolutionary tree, so we can further make guesses on what proteins we should study first. Using local sequence alignment methods, like Smith-Waterman, we could identify protein motifs that would not only give preliminary information for secondary structure, but also sequence patterns that could give insight to subcellular localization.

With the information on essential or unique proteins and assuming that we have the ability to grow and harvest *B.informaticus* in the lab, we could do a series of experiments with the objective of determining protein interactions within the cell. Similar to sequence alignments, we could do clustering analysis to find interacting proteins upon doing microarray experiments. Throughout the cell cycle of the organism, we could collect fractions from the cell culture and isolate the mRNA for different proteins. The mRNA could then be spotted onto microarray slides and the fluorescence intensity would denote the mRNA expression levels of those proteins. If their intensities are similar throughout the different phases of the cell cycle, they would indicate that their protein functions are highly correlated. There could also be a time-shifted, inverted, or random relationship between the proteins, which would all prove to be useful information. With such protein interaction information, we could also determine cellular pathways of *B. informaticus.* Also, with this expression data and sequence information, we could also find conserved promoter regions that would also be important in finding the different regulatory pathways found in *B. informaticus*.

Since the information on protein homology would give us a fair idea of what corresponding genes were essential or not, in conjunction with the ORF and mapping information, we could proceed to knock out each gene in the organism. Even though this process would be tedious, we could obtain more concrete information on each gene of *B. informaticus*. Depending on what works best in the organism, we could use transposons to randomly inactivate gene throughout the genome. After that, we could sequence and BLAST each of the resulting strains to see where the transposons integrated into the genome. If the transposons avoided a certain gene, it can be assumed that that particular gene is essential.

Each step of the way presents its caveats that can easily be generalized into three things: 1) reproducibility; 2) accuracy; and 3) data normalization. However, at the center of computational functional genomics is prediction, so these risks are certainly worth the information that can be gleaned.


B. informaticus
Identified ORFs and sequence
Multiple sequence alignment
Protein homology
Protein motifs
5S rRNA for evolutionary tree
Microarrays
Secondary structure
Subcellular localization
Gene expression analysis
Protein interactions
Conserved promoter regions
Regulatory pathways