Sara Nichols

Fall 2003 Take Home Final

TOPIC I

Many biologists cite the first prominent step in sequence alignment to be the paper published in 1970 by S.B Needleman and C. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins"; but even before amino acid sequence strings could be aligned, text comparison was emphasized by Vladimir Levenshtein in 1966 by introducing a distance term for strings, in "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." As many of the advancements in bioinformatics have been spearheaded by algorithmic ingenuity, Needleman and Wunsch's took the concepts of "edit distance" of Levenshtein and put it to practical biological use (Pevzner, 2000).

Needleman and Wunsch's dynamic programming algorithm, which avoids the exhaustive search of trying all paths by storing previously computed values, is a global alignment algorithm. While Needleman and Wunsch's technique is a valid method for matching the entire sequence, often looking at conserved subsequences is more helpful. T. Smith and M. Waterman published an alternative local alignment algorithm in 1981 (Smith, 1981).

Along with these dynamic programming approaches, the idea of 'filtering' surfaced in the early seventies. This notion is built up on the assumption that shorter sequences are primarily exact, or nearly exact, matches (Pevzner, 2000). Efficiency advances made with filtration by Lipman and Pearson in 1983 and 1985 lead to the FASTA algorithm. Later Basic local alignment search tool by Altschul et al., (1990) which focused on high scoring pair alignment or highest scoring local alignment, was derived from FASTA (Schuler, 1998) and it is currently one of the most popular database searches on the internet. Current research deals mostly with the compromise between speed and optimization (Baxevanis, 1998).

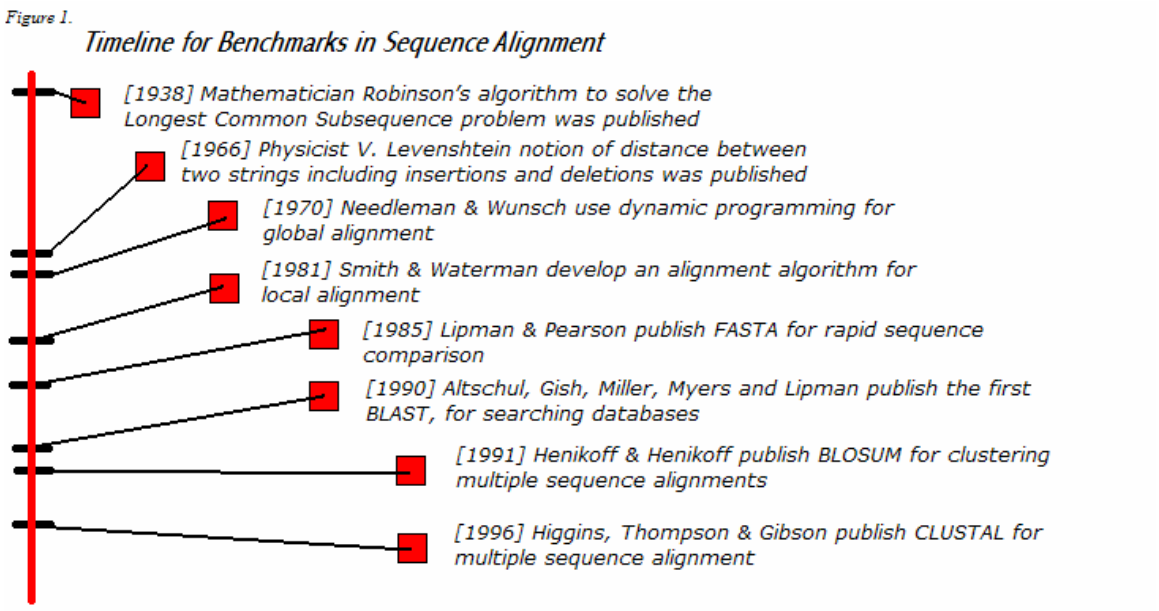
In addition to considering local alignment, biologist must take into account substitution and gaps. Substitution and gap scoring has been factored into variations of the previously mentioned methods. As biochemistry dictates in the case of amino acid sequence, some amino acids can be substituted with higher probability. Such probabilities are factored in with substitution scoring matrices. Dayhoff published the first widespread matrix in 1978, and it dictated point-accepted-mutations (PAM), using log odds (Schuler, 1998). PAM250 and BLOSUM (Henikoff, 1992) are matrices commonly used for optimal alignment.

For multiple alignment methods, often the sequences being aligned are predetermined to be related. This includes Higgins' CLUSTAL W (1996), which also calculates distance matrices, uses pair wise comparisons which are then in turn used to make a tree wise comparison. Motifs and patterns of the sequences, which can lead to families of proteins for example, are often used to gather together the sequences before a multiple sequence alignment (Baxevanis, 1998).

Computational molecular biology, including techniques for sequence alignment, is making a large impact on many aspects of biology. In the case of DNA homology, phylogenetic analysis is one example where sequence comparison has completely changed the field. Previously phylogenetic trees were based on phenotypes, and with the sequencing of DNA of organisms, we get closer to understanding how organisms evolved. Some organisms that were thought to be closer to one organism have now been shown to be closer to others. The aforementioned CLUSTAL can be used to cluster the data and fit into a phylogenetic tree (Hershkovitz & Leipe 192).

Another application of sequence alignment, this instance dealing with protein sequence alignment, is protein function inference. If a protein has a similar structure to a second protein, say for example they are in different organisms, or species, then past a certain homology, the proteins may function the same way. If a biologist has aligned two proteins against each other he may be able to infer that their function is related.

Figure 1.



TOPIC II

Computational functional genomics implies the use of computers to have high throughput methods for determining the function of proteins generated by a genome. In my proposal I will discuss how I will use a database of the unknown microbe's protein sequences to match homologous proteins, of which the function is predetermined. This is only the first step in working with proteins, which is a bridge to determining interactions, networks and ultimately a better understanding of the organism.

As dictated by the information given, we are provided with a previously unknown genome, with the open reading frames already determined; let's call it *M. genomeunknownus*. Since we know the ORFs, the DNA sequence can be translated into its corresponding amino acid sequence. We then have a database consisting of the DNA sequences and the corresponding protein sequences. We will name this MGDB for *M. genomeunknownus* Data Bank. We also know the parts of the genome which are nonprotein-coding sequences, which also can be stored for functional determination at a later date.

The first step I will take is to try to find homology of the entries in MGDB in other organism. To search the sequences against all of the microbes, we will first BLAST against the known microbial genomes. The University of Birmingham's ViruloGenome site has this subset, and the resources to do both nucleotide and protein sequence comparison at www.vge.ac.uk.

First we would compare the amino acid sequence, because as Dr. Gerstein said in class, the "protein search is more sensitive". Both nucleotide and amino acid sequence comparisons will benefit our genomic research for M.gen because some proteins may have been mis-annotated or missed completely, and comparing against the nucleotide sequence can only supplement our study. If indeed nucleotide sequences match that are not verified by an amino acid sequence which has been annotated, it provided further avenues to explore. In addition to Blasting against the microbial genomes, I would also look for sequence homology in other genomes. Although this data may be secondary to microbial genomes, the possibility of lateral gene transfer genes from microbial genomes may be prevalent in more complex organisms, again supplementary information.

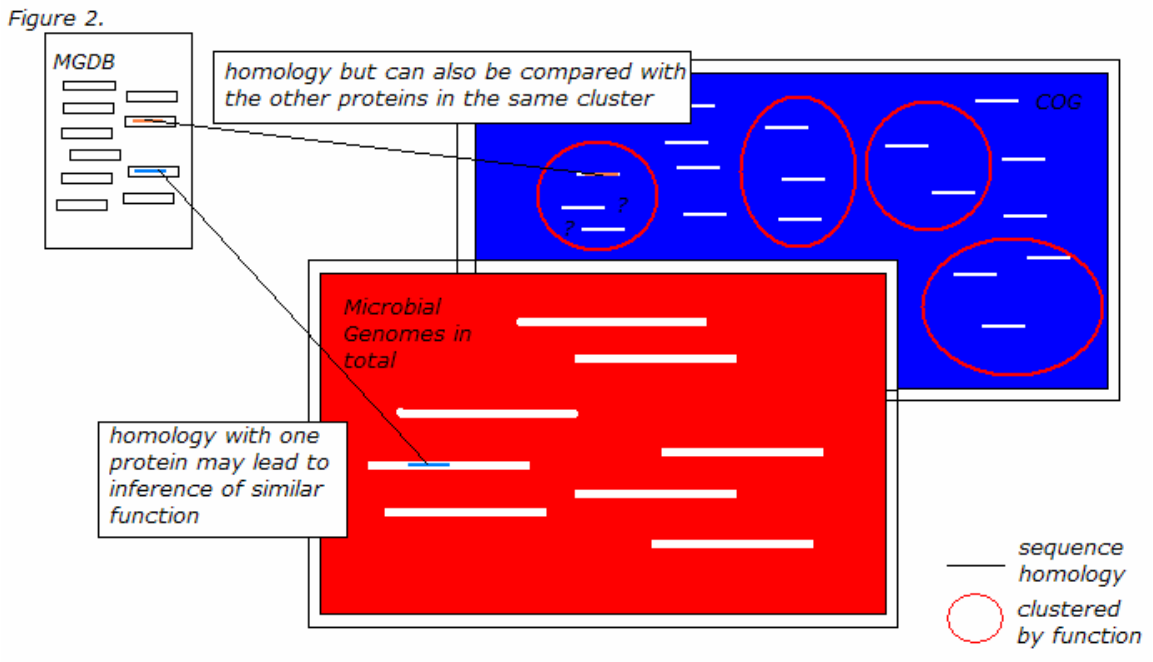
For each of the protein sequences listed in MGBD, I also propose searching against the COG database which, as of December 11, 2003, has 74,059 (Tatusoc, 2001) known protein sequences. This database is organized into a hierarchy of function. This will allow for

functionally grouping the proteins in M.gen with other clusters of previously determined function (Lio, 2003), but in this clustering case the functions can be more generalized, and slightly more diverse, allowing for a higher probability of a similarity in function (Figure 2).

The genes that are determined to be homologous to functionally known genes in other organisms are a first step in inferring function. If there are a few multiple matching sequences for each alignment, then those can serve as a base for a more stringent motif search. The motif can be searched in a database such as PROSITE (Sigrist, 2002). The sequences of these motifs, which intern match to secondary structure of previously can also be used to infer structure of M. gen proteins. When we start to get into secondary structure comparison, function is more relevant. Many avenues of secondary structure comparison can be explored.

In addition to sequence comparison, and secondary structure comparison, chip expression can be explored. DNA chip expression can be compared to DNA chip expression of other functionally predetermined proteins. Expression profiles of homologous proteins under the same growth conditions should be similar, and infer common functionality. Of course there are always exceptions, and misclassification, but leads determined through homology can be further tested with experimental data.

Figure 2.



Altschul S.F. et al. (1990), Basic local alignment search tool. *J Mol Biol*. Oct 5; 215(3):403-10.

Baxevanis, A. (1998), 'Practical aspects of multiple sequence alignment', in: Baxevanis, A.D. & Ouellette, B.F.F. (eds) *Methods of Biochemical Analysis: Bioinformatics, A Practical Guide to the Analysis of Genes and Proteins*, vol. 39, Wiley-Interscience, Inc., New York

Dumas J. & Ninio J.(1982) Effective algorithms for folding and comparing nucleic acid Sequences. *Nucleic Acids Reseach*, 10:197-206.

Henikoff S. Henikoff J.G. (1992), Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*. Nov 15; 89(22):10915-9.

Hershkovitz M.A. and Leipe D.D. (1998), 'Phylogenetic analysis', in : Baxevanis, A.D. & Ouellette, B.F.F. (eds) *Methods of Biochemical Analysis: Bioinformatics, A Practical Guide to the Analysis of Genes and Proteins*, vol. 39, Wiley-Interscience, Inc., New York

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Daklady*, 10:707-710.

Lio, P. (2003), Statistical bioinformatics methods in microbial genome analysis. *BioEssays* 25:266-273.

Lipman D.J. & Pearson W.R. (1985), Rapid and sensitive protein similarity searches. *Science*. Mar 22; 227(4693):1435-41.

Pevzner, P.A. (2000), *Computational Molecular Biology: An Algorithmic Approach*, The MIT Press, Cambridge, Massachusetts, pp 93-132

Robinson, G.de E. (1938) On Representations of the Symmetric Group. *American Journal of Mathematics*, 60:745-760.

Schuler, G.D. (1998), Sequence Alignment and Database Searching, in : Baxevanis, A.D. & Ouellette, B.F.F. (eds) *Methods of Biochemical Analysis: Bioinformatics, A Practical Guide to the Analysis of Genes and Proteins*, vol. 39, Wiley-Interscience, Inc., New York

Sigrist, C.J et al. (2002), PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*. 3:265-274.

Smith T.F. & Waterman M.S. (1981), Identification of common molecular subsequences. *J Mol Biol*. Mar 25; 147(1):195-7.

Tatusoc, R.L. et al. (2001) The COG database: new developments in phylogenetic classification of proteins for complete genomes. *Nucleic Acids Res*; 29:22-28.