(Topic 1)

The investigation and discovery of the relationship between the function, evolution, and structure of nucleotide and protein sequences has been a major focus of natural science for the past few decades. In response to the overwhelming burst of data generated by molecular biology initiatives facilitated by improvement of techniques such as microarray, it is essential to develop tools that will be able to expose, identify, store, and analyze effectively large number of biosequences. Sequence alignment is a fundamental part of Bioinformatics research, and in 1970 Needleman and Wunsch published the first pairwise method for optimal global alignment. It allows the best overall score for the comparison of the two sequences to be obtained with the consideration of gaps, and is capable for comparing two sequences expected to share a great deal of similarity over the whole length (Needleman and Wunsch, 1970). About ten years after the global alignment was published, optimal local alignment algorithms were also invented to find the best local similarities between two sequences. However, contrasting to the global alignment, the local alignment searches for regions of local similarity without the need for the complete sequences and it exclude explicit consideration of gaps (Smith and Waterman, 1981; Gotoh 1982). The global and local methods are significant in biological application. With the implication of the dynamic programming and similarity substitution matrix including PAM, Blosum, and Gonnet (Gerstein 2002), global alignment is able to predict the evolutionary distances of any two organisms and construct the phylogenetic trees. Moreover, local alignment methods are very useful for scanning databases to search for conserved amino acid motifs in related protein sequences, and can be used without the initial examination of sequence similarity over the entire lengths.

FASTA, a more sensitive derivative of the FASTP program, was developed by Lipman and Pearson in 1985. It includes an additional step in the calculation of the initial pairwise similarity score that allows multiple regions of similarity to be joined to increase the score of related sequences, and it allows trading speed for precision. It can be used to search protein or DNA sequence databases and can compare a protein sequence to a DNA sequence database by translating the DNA database as it is searched, and it can also show alternative alignment between sequences with periodic structures or duplications (Pearson, 1990; Pearson and Lipman, 1988). BLAST, developed in 1990, is another heuristic that focuses on no gap alignments and attempts to optimize a specific similarity measure without the distance-based procedure (Altschul et al., 1990. 1997). In proteins it is likely to assume regions of functional similarity by using BLAST when there is no gaps with high similarity score.

During the past eight years there has been an increase of the variety of improved sequence alignment techniques mostly based on the fundamentals of the previous methods. A few examples include the more recent version of BLAST which include BLAST 2 and PSI-BLAST in 1997, the WABA algorithm which used the concept of Hidden Markov model as its core for inserting gaps (Kent and Zahler, 2000), and the multiple sequence alignment. Multiple sequence alignment algorithms extract the relationship between many sequences by aligning the key segments together. Other than predicting the secondary or tertiary structure of new sequences, the similarities may

reveal evolutionary history, and are clues about the common biological function of the motifs (Holmes and Bruno 2001; Gerstein 2002). More recently, Super Pairwise alignment (SPA), which is fifteen times faster than the traditional methods, has been developed to reduce computation complexity without sacrificing too much accuracy by adopting algorithm which combines the methods of probabilistic and combinational analysis (Shen et al. 2002). Alignment-free sequence comparisons, with two categories of methods based on word (oligomer) frequency or methods that do not require resolving the sequence with fixed word length segments, were used to complement the alignment methods which did not take into account of genetic recombination and genetic shuffling (Vinga and Almeida 2003). Overall, most methods are used in Biology with the goals of finding homology and evolutionary relation as well as predicting structure and function based on the similarity to the well-researched sequences.

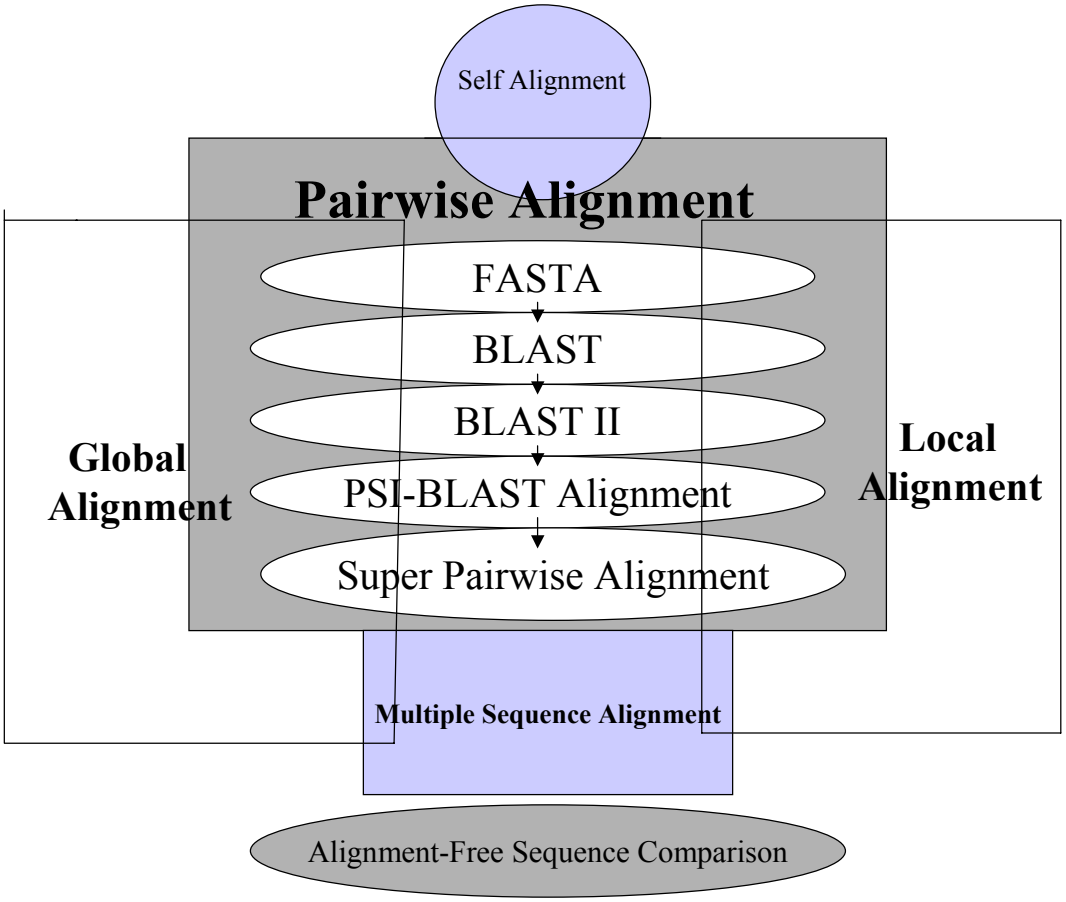

Figure 1: The interaction of different sequence alignment techniques. The first two alignment methods, global and local alignment, are the foundations which other improved later alignment methods build on. The arrows indicated order of invention, in which Super Pairwise Alignment is the latest and fastest and FASTA is the earliest among them (after global and local). The alignment-free sequence comparison method is an alternative technique dealing with special genetic events such as genetic recombination and shuffling, and it does not interact with others.

(Topic 2)

With the information of the complete sequence and open reading frames as the only source of information on the microbial genome, it is very difficult to study this organism and draw any conclusions. However, I would start by searching for the most similar complete sequence from another microbial organism with the method of global alignment with the help of BLAST against the published database such as NCBI. This is a key step to find the homology among other microbial organisms. The degrees of similarity of the whole sequence is significant in showing how related those two organisms are. The higher scoring on homology indicates that the two microbes were separated later evolutionarily and may contain more functional similarities. In addition, it is an important start to categorize the unknown microbial organism into species. I would choose a few other microbes with the highest scoring homology and determine what kind of organisms or species they are, and the unknown microbe is most likely to be close to them on the evolutionary tree. Using only one organism with the highest similarity is a little bit risky due to the possible small genome size, therefore it might be more applicable to compare with five organisms together.

After finding the closest related microbes, the next step will be to find the similar motifs between those few organisms by performing the multiple sequence alignment. It is one of the most essential tools in molecular biology for phylogenetic analysis and has the potential of predicting protein secondary and tertiary structure. By comparing the common motifs between these highly related organisms and searching for the documented functions and structures of the proteins within the motifs, it is possible to make accurate predictions of the gene families with the information of the open reading frames. In addition, since the open reading frame is known, it is possible to generate the protein sequences of the unknown organism with the correct reading frame and predict and the secondary protein structures by using GOR and confirm the prediction with the result we have from the multiple sequence alignment. However, all of the previous steps are just hypothetical computational prediction by searching the database, and with no further information we will not be able to perform structure alignment and other techniques. Nevertheless, it is possible to move a step further on the prediction based on the data we have. For example, with the motif comparison, we might be able to generate probable subcellular localization based on the sequence pattern, the level of expression, and the Bayesian system, assuming that a significant amount of information was obtained from the previous computational analysis experiments.

However, all of the predictions will have to be backed up by experiments for at least a portion of it in order to assess the accuracy of the prediction. There are a few kinds of experiments that will give us a large amount of molecular data with relatively small amount of time and effort committed. First of all, I would perform a microarray experiment with the same treatments and conditions which had been conducted on other closely related organisms. By comparing the expression similarity and investigation of individual genes, it is rather easy to confirm our previous predictions. Nevertheless, the individual gene analysis should be compared if there are large amount of data available

because it is unreliable to reach any conclusion of a single gene based on one microarray experiment. Second, I would perform a whole genome transposon knockout experiment to investigate the number of essential genes in the genome and to see the function of each knockout gene.

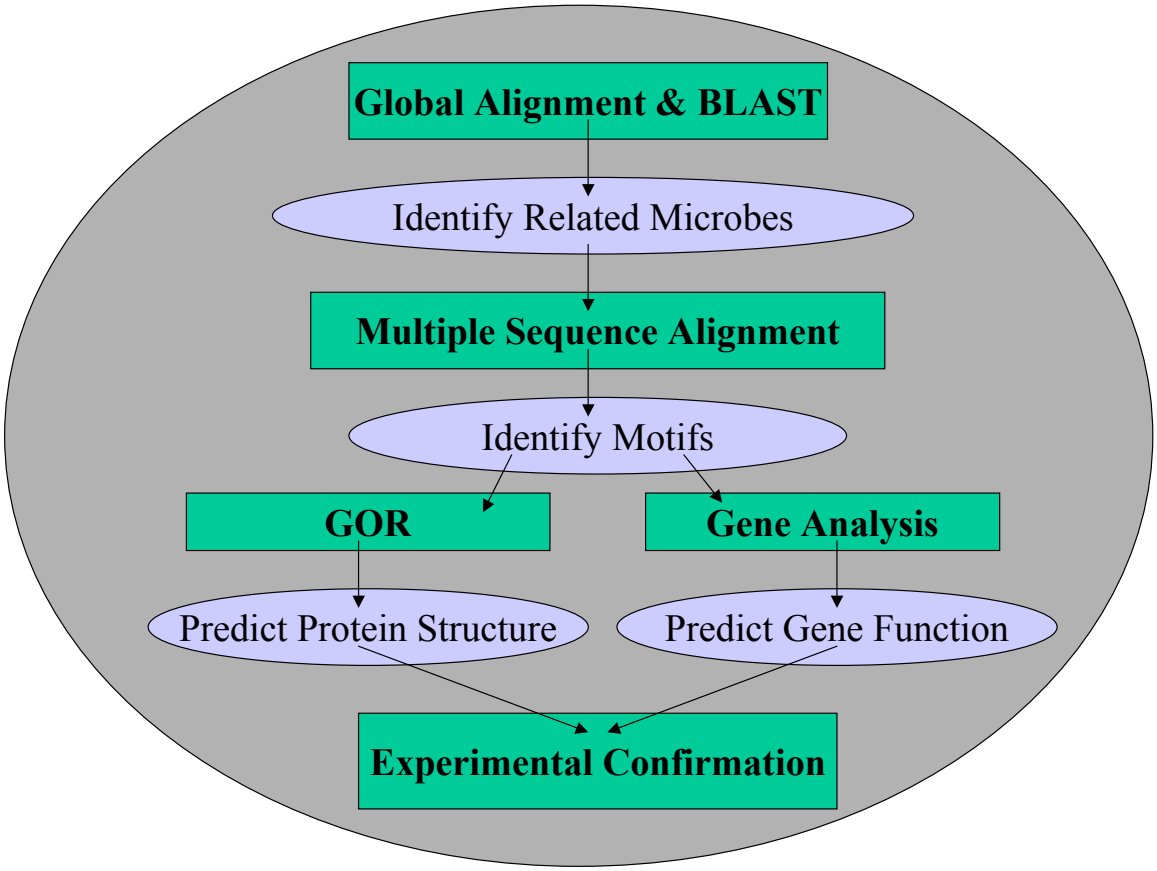

Figure 2: Computationally based Analysis and procedures for learning unknown organisms. Experimental procedures are highlighted in green squares and functions to be investigated are in blue circles.

Reference:

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Gerstein, M. (2002). Class Material. Yale, Bioinfo.mbb.yale.edu/mbb452a.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. J. Mol. Biol. 162:705-708.

Holmes, I. and W.J. Bruno. (2001) Evolutionary HMMs: a Bayesain approach to multiple alignment. Bioinformatics. 17:803-820.

Kent, W.J. and A.M. Zahler. (2000). Conservation, regulation, synteny, and introns In a large-scale C. briggsae-C. elegans genome alignment. Genome Res. 10:1115-1125.

Needleman, S.B., and C.D. Wunsch. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48: 443-453.

Pearson, W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 183:63-98.

Pearson, W.R., and D.J. Lipman. (1998). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA. 85:2444-2448.

Shen, S.Y., J. Yang, A. Yao, and P.I. Hwang. Super Pairwise Alignment (SPA): An efficient approach to global alignment for homologous sequence. J. Computational Bio. 9:477-486.

Smith, T.F. and M.S. Waterman. (1981). Identification of common molecular subsequences. J. Mol. Biol., 147:195-197.
Vinga, S. and J. Almeida. (2003). Alignment-free sequence comparison-a review. Bioinformatics. 19:513-523.