Sarah Fiorentino
MBB 752
12/12/03

Sequence Alignment Techniques and Their Uses

Since rapid sequencing technology and whole genomes sequencing, the amount of sequence information has grown exponentially. With all of this data, it is possible to do comparisons where one can learn about the structure, function, and evolutionary relationships of different parts of organisms. Sequence alignment techniques have been developed to do comparisons. The goal of alignment is to obtain the optimal alignment of sequences. Pairwise alignment techniques, where two sequences are studied at a time, were developed first with multiple alignment techniques, where many sequences are compared at once, coming later (Fig. 1).

Dot plots were developed by W.M. Fitch (1969) as a way to visualize similarities and differences between two sequences. Regions of similarity appear as diagonal lines on the matrix. The Needleman-Wunsch (1970) algorithm globally aligns two sequences, meaning the path starts at one edge and runs continuously to the other edge. This method does not penalize the score of the alignment for the insertion of gaps into the sequence (Vingron, 2002). The Smith-Waterman (1981) algorithm was developed to detect local alignments. This method allows paths to begin and end inside the matrix (Schuler, 1998). In 1983, the concept of using "words" to look for local similarities was developed (Wilbur and Lipman, 1983). Next, modifications to the Smith-Waterman algorithm to detect the best nonintersecting, suboptimal, local alignment were added (Altschul and Erickson, 1986; Waterman and Eggert 1987). The first substitution matrix, PAM (Percent Accepted Mutation), was developed to measure evolutionary distances (Dayhoff, et al, 1978). Another substitution matrix, BLOSUM, was developed that compared sequences based on their maximum level of identity (Henikoff and Henikoff, 1992). In the mid-1980s, techniques to search similarities within databases were created. The first was FASTA, which used substitution matrices to match words (Lipman and Pearson, 1985). The next technique was BLAST (Basic Local Alignment Search Tool), which uses neighborhoods of words and Karlin-Altschul statistics, but does not allow gaps (Altschul, at al, 1990). BLAST was modified to BLAST 2.0, which allowed gaps, and PSI-BLAST, which creates and iteratively refines profiles (Altschul, et al, 1997).

The first multiple alignment method was creating profiles, which iteratively applied pairwise alignments with a fixed alignment of a subgroup, thus allowing the determination of conserved patterns and creation of hierarchical trees (Gribskov, et al, 1987). The next method developed was CLUSTAL, which also uses profiles, and has been modified over time (Higgins, et al, 1992; Higgins et al, 1996: Higgins, et al, 1997). Two other variations of profiles are MultAlin by Corpet (1988) and generalized profiles by Bucher and Karplus (1996). The Hidden Markov Model (HMM) is a powerful way to align and search sequences that "learns" the characteristic traits of the sequence sets (Krogh, et al, 1994).

PAM matrices can be used to determine whether an amino acid substitution would be favored or avoided over time (Vingron, 2002). PAM matrices with lower numbers, e.g. PAM120, are closer in time than PAMs with higher numbers, and have a higher degree of similarity. Evolutionarily related proteins have biased amino acid frequencies that represent changes that have been accepted over time (Shuler, 1998).

Functions of unknown proteins can be determined by doing BLAST searches. Sequences that align best are most likely to have similar functions. If the function of the

best aligned protein is known, the function of the unknown protein can be inferred by "guilt by association", i.e. if it looks like it, it probably acts like it. Further biochemical experiments can then be performed to confirm the function.

Sarah Fiorentino
MBB 752
12/12/03

FIGURE 1. Timeline of Sequence Alignment Techniques

1960	Dot Plot (Fitch)
1970	Needleman-Wunsch Algorithm
1978	PAM (Dayhoff, et al)
1981	Smith-Waterman Algorithm
1983	"Words" used in Searches (Wilbur & Lipman)
1985	FASTA (Lipman & Pearson)
1986	Modifications to Smith-Waterman (Altschul & Erickson)
1987	Modification to Smith-Waterman (Waterman & Eggert)
1987	Profiles (Gribskov, et al)
1988	FASTA 2.0 (Pearson & Lipman)
1988	MultAlin (Corpet)
1990	BLAST (Altschul, et al)
1992	BLOSUM (Henikoff &Henikoff)
1992	CLUSTAL V (Higgins, et al)
1994	HMM (Haussler, et al)
1996	CLUSTAL W (Higgins, et al)
1996	Generalized Profiles (Bucher &Karplus)
1997	CLUSTAL X (Higgins, et al)
1997	BLAST 2.0 (Altschul, et al)
1997	PSI-BLAST (Altschul, et al)

Black, Green, and Red are Pairwise Comparisons

Green are Substitution Matrices

Red are involved with Database Searching

Blue are Multiple Sequence Alignments