Tiffany Samaroo
MB&B 452a
December 8, 2003

**Take Home Final**

**Topic 1**

Prior to 1970, protein and DNA sequence alignment was limited to visual comparison. This was a very tedious process; even proteins with global similarity can possess more regions of dissimilarity (eg. introns) than local similarity. In 1970, Needleman and Wunsch introduced the first computerized method for sequence alignment. Their method was based on the concept of dynamic programming. In dynamic programming, a similarity matrix for two sequences is constructed and computed such that each cell in the matrix is given a quantitative score that reflects the best previous alignment less any penalties assigned that result from making the alignment. The optimal alignment is therefore the highest scoring path through the entire matrix. In their original paper describing dynamic programming (Needleman & Wunsch (1971)), they assigned a value to each cell that was reflective of the number of identical matches that preceded that alignment.

The Needleman-Wunsch (NW) method of alignment is a global evaluation of homology. This method is used to compare two sequences in their entirety in an attempt to determine how similar they are overall. This method works well when the sequences are potentially equivalent (eg. human and mouse beta-globin). In 1981, Smith and Waterman derived a dynamic programming-based algorithm for local sequence alignment (Smith & Waterman (1981)). The Smith-Waterman (SW) local alignment algorithm was slightly different than the NW algorithm: they allowed for negative scores in their matrix such that optimal, high-scoring alignments (independent of size) could be identified amongst regions of dissimilarity. In other words, their algorithm allowed for short spans of similarity rather than global similarity. Today, short regions of similarity are critical for genome comparisons---sequence targeting signals, conserved domains, and protein motifs are all routinely identified using local alignment-based algorithms.

NW and SW algorithms are effective for comparing pairwise alignments and for evaluating global and local similarity, respectively. However, multiple alignments are critical for understanding and classifying genomic information. In multiple sequence alignments, clustering approaches are taken because this type of alignment is not amenable to dynamic programming. When a multiple alignment is constructed, the two most similar sequences are aligned pairwise and subsequent sequences are added on to this initial alignment. The ability to align multiple sequences has been critical to algorithm development today.

As sequence databases began to grow exponentially in the last decade, the need to rapidly search and extract information from these databases grew urgently as well. The first two algorithms designed specifically to address rapid database searches were FASTA (Pearson & Lipman (1988)) and BLAST (Altschul, et al. (1990)).

FASTA was the first program to become rapidly and widely utilized. FASTA first searches the database using several short sub-sequences (words) from the query

sequence to find identical matches; FASTA is a local alignment (SW-based) tool with multiple alignment capabilities. It is a more rapid search tool than SW, making it a more desirable tool for biological researchers. Unfortunately, however, sensitivity is sacrificed for speed in the case of FASTA as well as BLAST.

Following closely behind the advent of FASTA was BLAST. BLAST is an even more rapid alignment algorithm, and consequently, not as sensitive as other time consuming methods. BLAST employs a substitution matrix that is used to quantify the alignment such that each possible residue substitution is given a score reflective of the probability that the alignment could not have occurred by chance alone. The algorithm sought out not only high-scoring segment pairs (HSPs), but also HSPs with optimal neighboring ungapped alignments. In other words, the algorithm sought out the best ungapped local alignment that could not be improved with adjustments to the alignment (extensions or trimming). The search method is a two-step process. First, BLAST scans the database for words that possess a given score, thereby obtaining hits. Then second, BLAST verifies each hit by extending the alignment to neighboring pairs in search of HSPs.

One problem with BLAST is that it cannot generate gapped alignments. This problem has been addressed with Gapped BLAST (Altschul, et al. (1997)). Gapped BLAST is a modern counterpart of BLAST that surpasses its predecessor; it is more sensitive and just as fast. Gapped BLAST differs from the original BLAST in three ways. First, the algorithm establishes a cutoff such that sequences scoring above a given E-value threshold are omitted. Second, to increase the speed of the search, the criterion for extensions was modified from the original BLAST algorithm. Third, the ability to form gapped alignments was added. All of these improvements have allowed BLAST to remain one of the more commonly used search algorithms to date.

Another modern counterpart to BLAST is PSI-BLAST. PSI-BLAST is a method of searching for profiles and motifs that utilizes Gapped BLAST in its iteration scheme (Altschul, et al. (1997)). Compared with other BLAST derivatives, PSI-BLAST is slightly more sensitive. In PSI-BLAST, a profile group is created by aligning sequences for the database with the query sequence. Each time a new sequence is aligned to the profile group, the process of iteration is repeated, thereby generating a new multiple sequence alignment. The "new profile" is then reiterated and this process continues until all of the related sequences below a given E-value threshold are extracted from the database (Figure 1). With this algorithm, protein families are easily assembled; for this reason, this application is used when the query sequence contains a position-specific conserved domain. Consequently, PSI-BLAST is an essential tool for uncovering protein relationships.

Sequence alignment methods have progressed greatly in the last 30 years, each new improvement building on previous algorithms and meeting the growing demands of bioinformatics. As databases continue to grow, future advances should focus on creating algorithms that increase the speed with which alignments are carried out, without sacrificing sensitivity in the process.

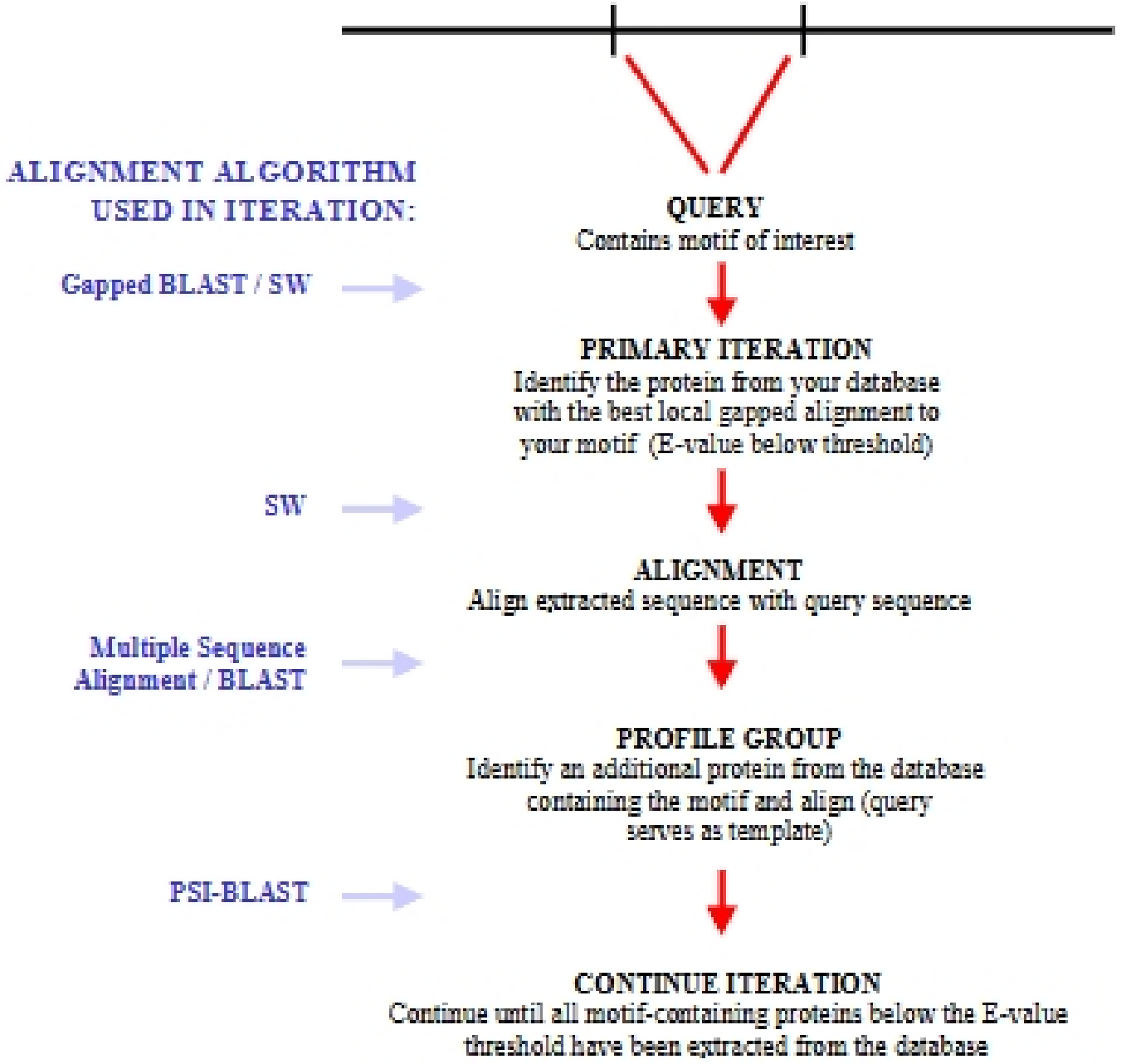


**FIGURE 1**
PSI-BLAST is a modern counterpart of BLAST that incorporates many sequence alignment algorithms into its iteration scheme.