Edmund Burke

Genomics Final Project

*Topic 1*

While researchers attempted *ad hoc* to align sequences earlier, the first widely used alignment algorithm is attributed to Needleman and Wunsch, who published in 1970 their "general method applicable to the search for similarities in the amino acid sequence of two proteins." In their groundbreaking algorithm, two sequences are compared in a two dimensional matrix. A "one" is inserted in boxes marking the coincidence of the same amino acid in both sequences, and then to the value in each cell (starting at lower left) is added the maximum value of the set of numbers constituted by the cell to its lower right and every cell past that cell in both column and row. The best alignment is the path through the matrix in numerical order.

This was made more complicated by the idea that simple percent identity was not biologically significant enough. In the late 70's Dayhoff and Eck perfected the Percentage Accepted Mutation (PAM) matrices, which is a matrix of weights derived from the frequency with which amino acids replace each other in natural evolution; the weights were based on a study of amino acid changes in closely related proteins. Concerns about sample bias impelled Henikoff and Henikoff (1992) to come up with the BLOSUM matrices, based on data from highly conserved segments from divergent proteins. This system uses a log-odds matrix calculated from the frequencies A(*ij*) of residue *i* in one cluster replacing residue *j* in another.

In 1981, Smith and Waterman built on to the work of Needleman and Wunsch; this algorithm, rather than looking at the sequences in total, finds an alignment that gives the longest and best *sub*sequence. Functionally, this means that the beginning and end need not be in the last column or row of the matrix, as it finds the regions of highest homology and builds outward.

Eventually, sequence databases started growing at an exponential rate, and it was impossible to query all the available data with the most sensitive algorithms (Smith-Waterman especially); the computers of the early 80s could not handle these calculations. In 1983 Wilbur and Lipman came up with an shortcut algorithm (FASTA), which finds

many small matching sequences, creating diagonals, then attaches these diagonals (with gaps); then it finds the optimal local alignment of these.

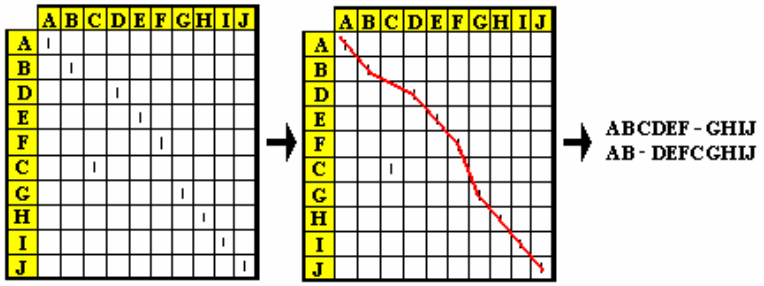
This type of shortcut programming gave way to the BLAST series (Altschul *et al*., 1990). Like FASTA, BLAST is a heuristic program, sacrificing complete sensitivity for time, but was designed for aligning protein sequences. The original BLAST program did not allow gaps, but more recent incarnations (gapped BLAST, WU-BLAST) do. Position Specific Iterated BLAST uses a profile created from multiple alignments of the individual highest-scoring hits from an initial BLAST search; conserved positions receive high scores and the program repeats (iterates); each round refines the search.

Multiple alignment programs allow some of the most useful applications of sequence alignment. Among the first such programs was Clustal (Higgins, 1988), built on the algorithm of Doolittle and Taylor; it is a recursive program which aligns the easiest sequences first and the more divergent ones last.

Multiple alignment is of tremendous importance in evolution biology. Carl Woese revolutionized the system of classification when his alignment-based study on the divergence of 16s rRNA of ribosomes between different species suggested that there are three clusters of similarity. Science (for the most part) now accepts these three domains of life, rather than the five kingdoms which are based on more superficial and phenotypic differences.

Sequence alignment also figures to be of the utmost importance in designing rational drugs, which is seen as the future of medicine. By studying how differences in amino acid sequence affect 3D protein structure, it may be possible to design extremely specific and effective inhibitors.

**Important Events in the History of Sequence Alignment**

| Year | Inventor | System Features |
|---|---|---|
| 1970 | Needlman and Wunsch | Introduced Scored Matrix   |
| 1967-78 | Dayhoff | Percent Accepted Mutation Matrices<br>-- gave the coefficients used in the matrix analysis biological signficance |
| 1981 | Smith and Waterman | Their algorithm favors local alignment over global alignment. The difference is small mathematically (essentially the 0 is the only difference), but a sea change in considering sequence divergence. $H_{ij} = \max\left\{H_{i-1,j-1} + s(a_i, b_j), \max_{k\geq 1}\{H_{i-1-k,j-1} - g_k\}, \max_{l\geq 1}\{H_{i-1,j-1-l} - g_l\}, 0\right\}$. the difference! (more or less) |
| 1983 | Wilbur and Lipman | Their shortcut algorithm, which sacrifices complete sensitivy for efficiency was a response to the slow computing of the 1980s (the CDC 4000, e.g.). Gave way to FASTA and then BLAST, today's preeminent alignment system. |

*Topic 2*

Once the genome of a new organism has been sequenced and the open reading frames identified, the challenge is to find the function of all the genes or gene products. It is estimated that about 40% of the open reading frames of all sequenced organisms have no known biochemical function, so this, the territory of functional genomics, is no simple task. Functional genomics has been defined (Hietor and Boguski, 1997) as "the development and application of global experimental approaches to assess gene function by making use of the information and reagents provided by genome sequencing and mapping."

Functional genomic analysis employs a multitude of methods, requiring varying degrees of computational assistance, to decipher gene function. Many of these would be appropriate in this hypothetical situation, including expression profiling, reverse genetics, mutagenensis screens, random mutagenesis and mutant screening, PSI-BLAST homology searches, and computer-assisted function assignment from amino acid sequence.

Once a comprehensive cDNA library of an organism is in place, expression profiling is a tremendously useful tool of functional genomics. Knowing when (and where) the mRNA transcript of a given gene is found gives myriad information about the gene's function. The expression pattern of a gene product can suggest that it is involved in metabolism (undergoes significant change when nutritional compliment altered), that it is involved in reproduction (expression patterns temporally match reproductive cycles) or any other cellular function. Hierarchical clustering will then group genes that respond similarly across varied experimental conditions. The computational aspect comes in trying to ultimately use these data to decipher the underlying regulatory network of the organism.

Alignment programs often can shed light on the function of a gene or a protein. Once the ORFs of the mystery organism are known, the predicted protein can be translated and PSI-BLAST can be used to search sequence databases for known sequences with which the mystery protein has high homology. It is likely that similar sequences will have similar functions; once function is suspected, experimentation can confirm the functional homology. A main developing goal of functional genomics is the elucidation of protein function from amino acid sequence without the help of known sequences with known function. This is a nascent field, so for now function from alignment comparison is far more useful.

The use of phylogenetic profiling can assist in the understanding of the protein compliment of an organism. A phylogenetic profile follows the presence or absence of the same proteins (or closely homologous ones) across many organisms. The similarity of two protein's profiles is related to their functional interaction; the implication is that if two proteins are almost always found together or are almost always both absent across many organisms, they probably function together.

Another computationally based tool is the fusion method of protein analysis

across species. Essentially, if two proteins from one organism are present in other organisms as one larger, fused protein there is a high likelihood that the separate proteins interact. Functional links between genes can also be found by investigating if proximity/location of genes is conserved across genomes.

Other less computational-based or less genomics-based methods for functional assignment include: protein-interaction studies, which highlight possible functional interactions; post-translational modification mapping; and genetic screens for suppressors or enhancers.