Samuel Flores
Genomics and Bioinformatics
11 December 2003

Instr.: Mark Gerstein
TAs: Mike Seringhaus, John Rinn

An Historical Summary of Sequence Alignment Techniques (1970-present)

Early attempts to use protein sequence in the study of evolutionary development were hampered by the absence of fast and sensitive methods for sequence alignment. Practical sequence alignment can be said to have been born with Needleman and Wunsch's seminal paper in 1970. Methods such as those proposed by Braunitzer in 1965 were computer adaptable and showed some degree of sensitivity but were very expensive computationally, which limited their usefulness in light of the state of computers of the day.

The central contribution of the NW method was the application of dynamic programming to the sequence alignment problem. The key idea of dynamic programming is that the problem is divided into stages. A partial alignment is frozen and not revisited as the algorithm proceeds to align further residues.

A maximum match in the NW method was defined in this method as the largest number of amino acids that could be matched while allowing gaps in either the query or target sequence. The allowance of gaps results in many matches, but the method excludes those that do not lead to a maximum alignment. A matrix is constructed in which matching amino acids are indicated with a 1 in the corresponding matrix element. Comparisons are made by finding pathways through the matrix and automatically choosing the one that results in the highest sum.

Waterman and Smith in 1981 introduced an algorithm that identified subsequences of high similarity and assigned a penalty to gaps.

One obvious problem with these methods is that they do not assign a partial score for conservative substitutions. The first steps toward addressing this problem were based on Dayhoff's 1968 statistical study of pairwise substitution probabilities culled from 1,572 changes in 71 groups of related proteins. From this work were born the PAM (Percent Accepted Mutation) matrices.

Henikoff and Henikoff in 1992 introduced an alternative substitution matrix based on blocks of aligned sequence segments corresponding to over 500 groups of related proteins and obtained improved sensitivity.

Wilbur and Lipman in their 1983 article introduced a lookup table to the field and obtained a leap forward in processing speed. They also employed the PAM250 matrix to extend the matching word pairs along their diagonals. The highest scoring diagonals were then connected without gap penalties.

Altschul and Lipman in 1990 introduced BLAST, which used a lookup table and then extended the high scoring pairs left and right. The single best highest scoring pair was reported as a match -- this was very fast but a big disadvantage was an intolerance of gaps. In 1997 this algorithm was improved by implementing a two-hit method for quickly identifying the best HSP's, coupled with a gap-tolerant method for joining HSP's with dynamic programming. They also introduced PSI-BLAST, which performed multiple alignments by starting with a set of BLAST matches and constructing an HMM to then find further matches. These methods are currently considered the state of the art.

As demonstration of the utility of PSI-BLAST, a human HIT protein was used as a BLAST query, which turned up six other HIT proteins. These were in turn used to generate a scoring matrix which

was then used to find homology with GalT proteins known to be similar. It also turned up a previously undiscovered homology with a yeast phosphorylase protein.

A multiple alignment procedure has been used to discover Kazal-type protease inhibitor domains in mammalian anion transporters. This appears to be the first discovery of these domains in transmembrane proteins.

	A	C	N	Q	C	L	R	A	W
A	**6**	4	4	4	3	3	2	2	0
C	4	**5**	3	3	4	3	2	1	0
N	3	3	**4**	3	3	3	2	1	0
Y	3	3	3	3	3	3	2	1	0
K	3	3	3	3	3	3	2	1	0
R	2	2	2	2	2	2	**3**	1	0
A	1	1	1	1	1	1	1	**2**	0
W	1	1	1	1	1	1	1	0	**1**
A	1	0	0	0	0	0	0	1	0

Sample alignment by Needleman-Wunsch method. The aligned residues are highlighted in red.

1. Needleman, S.B., Wunsch, C.D. (1970) J.Mol.Biol. 48, 443.
2. Braunitzer,G. Evolving Genes and Proteins, ed. by V.Bryson and H.J. Vogel, p. 183. New York: Academic Press
3. Waterman, Smith (1981) J.Mol.Biol. 147, 195.
4. Dayhoff, M.O., ed. (1968) Atlas of Protein Sequence and Structure, Vol.5
5. Altschul,S., Gish, W., Mill, W., Myers, E.W., and Lipman, D.J. (1990) J.Mol.Biol. 215,403.
6. Altschul, S., Lipman, D. (1997) Nucleic Acids Research 25,3389.
7. Henikoff, S., Henikoff, J.G. (1992) PNAS 22, 10915.

Samuel Flores
Genomics and Bioinformatics
11 December 2003

Instr.: Mark Gerstein
TAs: Mike Seringhaus, John Rinn

Functional Genomics of a New Microbial Genome

A newly sequenced genome always presents a challenge and opportunity for functional characterization. Such a wide-open task admits a wide array of approaches. Here we present a series of steps that may be taken to obtain probable functional information on genes using purely computational techniques, by means of comparison to other, well annotated genomes. We also propose DNA and protein microarray experiments to generate data that will contribute to a computational survey.

Our first task is to identify COGS formed by proteins from X (the microbe of interest) and genomes from a number of well characterized microbes believed to be related to some extent. A proven approach is that due to Tatusov [1]. In this method, all pairwise comparisons (e.g. BLAST) are made between genes in the various genomes. The best hit (BeT) between a gene and each of the other genomes is detected and a link is drawn between the two. In order to eliminate spurious matches and matches to paralogs, it is argued that If a agene from one genome has BeTs in two of the other genomes, it is unlikely that the two are also BeTs for one another unless the three are true orthologs. Thus a triangle of BeTs is the simplest unit indicating an orthologous group.

Further information which can be gained from sequence is given by comparison of whole-genome structure [2]. Recombination events that occurred in the organism's distant past move segments from place to place but conserved the sequence of genes within a segment. Thus, if local similarity can be shown between certain segments of X and segments in another, well characterized genome, it is likely that many of the genes will have similar structure and function. In the best case, this permits automated transfer of annotation, at least for some segments of the genome.

As mentioned, the appeal of microarray experiments is undeniable for gathering of data complementary to the above. Of high importance will be coexpression data gathered by DNA microarrays under a variety of growth conditions and stimuli. Expression patterns should be clustered by an unsupervised method unless the COGs are deemed a reliable basis for inferring function.[3] Local clustering should be performed to detect time-shifted or inverted correlations. [4] Interacting groups can then be examined in detail to determine whether the automated annotation obtained above is consistent. All possible pairwise interactions between proteins can be assayed with protein microarrays to develop families of interacting proteins which can be similarly checked against annotation.

If all of the above methods prove to be often contradictory and of limited reliability as is most often the case, then a consensus voting method should be employed to determine whether a given pair of proteins interact or not.. A Bayesian network [5] can be used for this purpose but note that there are some key differences between this problem and that treated by Jansen and Gerstein. Here we would be assigning proteins to families first theoretically, by comparison to orthologs in better characterized microbes, and then experimentally, by clustering expression data and by forming relationship networks from protein-protein microarray data. The algorithm should be trained by feeding it data from a reasonably related organism whose functional protein interactions have been characterized. The output of the Bayes network is a yes-no decision as to whether the proteins interact or not. This information can be used in conjunction with the automated annotation to reconstruct families of interacting proteins and infer function for each gene.