OK INBRE Bioinformatics

November 2005

University of Oklahoma Health Sciences Center

# Pairwise Sequence Alignments

*by Edgar Scott, II, M.S.*

One focus of many scientists today is to determine if two sequences are related by the sharing of similar functional domains or sequence motifs. Through analysis DNA and protein sequences, it is possible to determine if two sequences are significantly similar at the sequence level, which suggests that they are homologous. Determining that two sequences are related indicates that they could have a common function, similar 3-D structure and common functional domains, providing many potential uses in genomics, proteomics, pharmacogenomics and other fields.

One area that relies on determining the existence of homology between two sequences is annotation. Annotation is the process of taking the raw sequence data from a genome sequencing project and analyzing them to identify gene products and other key features of the genome. After identifying novel genes from the genome, these genes then can be compared to the genes of other organisms whose genes have been studied and well characterized. Determining that a novel gene and a well known gene are significantly similar implies that both sequences are homologous and are therefore likely to have a common function and structure. This will allow scientists to, temporarily, transfer annotations from the well known gene to the novel gene until experimental techniques have been developed to validate the true function of the novel gene.

Performing sequence comparisons is at the core of many bioinformatics applications. Sequence comparison algorithms compare two sequences by searching for a series of identical characters or character patterns that are in the same order in both sequences. Relatedness can be determined between two protein sequences, two RNA sequences or two DNA sequences. By assessing the degree of similarity between two sequences, one can infer that homology exists between two compared sequences.

The optimal alignment between two sequences is the best evidence for or against the existence of homology between sequences. An optimal alignment is the best of all the possible alignments that could arise between two sequences and will include a maximized number of identical residues appearing in series between both sequences, a minimized number of mismatching residues, and a minimized number of gaps. Gaps are inserted into an alignment to increase the number of matching residues and represent insertion or deletion events that have occurred as the two sequences have evolved from their common ancestor, assuming that they are evolutionarily related.

Pairwise sequence alignments are accompanied by three different scores to provide a quantitative measure of the amount of identity or similarity that exists in an alignment. One such score is the percent identity score, representing the percent of the alignment that involves identical matching residues. A second score is the percent similarity score, which is the percent of the alignment that involves identical and similar matching residues. This is only applicable to protein sequence alignments where similar residues are amino acids that have similar physicochemical properties (i.e., glutamic acid and aspartic acid are both acidic residues). A third score, called the alignment score, uses a scoring matrix and is based on all identical matches, similar matches, mismatches and gaps that arise in the alignment. An alignment score helps to identify an optimal alignment. The higher these three scores, the better, or more optimized, the alignment.

There are two types of scoring matrices that are used to derive alignment scores for protein sequences: percent accepted mutation (PAM) matrices and blocks substitution matrices (BLOSUM). PAM matrices are based on data from the alignment of closely related protein families and they involve the assumption that substitution probabilities for highly related proteins can be extrapolated to probabilities for distantly related proteins. Several PAM matrices - PAM30, PAM50, PAM70, PAM100, and PAM250 - are available for protein analysis. PAM matrices with smaller numbers for suffixes represent matrices to be used when comparing more closely related sequences, while the PAM matrices with larger number suffices are to be used when comparing more distantly related sequences. PAM250, for example should be used when proteins share about 20 percent amino acid identity.

BLOSUM matrices serve the same purpose as PAM matrices but are derived differently. The BLOSUM matrices are based on empirical observation of distantly related protein alignments. There are several BLOSUM matrices available for use as well: BLOSUM40, BLOSUM45, BLOSUM50, BLOSUM62, and BLOSUM80. In contrast to PAM matrices, the lower the suffix number attached to the matrix means that this matrix is to be used to compare more distantly related sequences. For example, BLOSUM62 is useful for aligning proteins that share less than 62 percent identity while BLOSUM40 is useful for aligning proteins that share less than 40 percent identity.

When aligning DNA sequences, there are only a couple of scoring matrices from which to choose. An identity matrix is often used to align DNA sequences where an identical match receives a +1 score and a mismatch receives a -1 score. Some matrices will provide more complexity by considering transitions and transversions, where the likelihood of finding a transition (a purine aligned with a different purine or a pyrimidine aligned with a different pyrimidine) is greater than the likelihood of finding a translation (a purine aligned to a

The University of
Oklahoma Health
Sciences Center

Laboratory for Genomics and
Bioinformatics

Edgar Scott II, M.S.
Multicampus Bioinformatics
Education Specialist

Phone:
405-271-2133 Ext. 32511
E-Mail:
edgar-scott@ouhsc.edu

We're on the Web!

*See us at:*

**microgen.ouhsc.edu/inbre/**

**Websites**

1. www.ebi.ac.uk/emboss/align/index.html
2. www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi
3. pir.georgetown.edu/pirwww/search/pairwise.html
4. www.expasy.ch/tools/sim-prot.html

# Bulletin (continued)

pyrimidine). These two types of scoring matrices will help to identify the optimal alignment between two DNA sequences.

These scoring matrices are applied to two types of sequence alignments: global alignments and local alignments. (A more simple identity matrix is used for DNA alignments instead of PAM and BLOSUM matrices.) The first sequence alignment was developed by Needleman and Wunsch, called the global alignment. This alignment algorithm searches for an optimal alignment over the entire length of both sequences. Global alignments are better suited for determining if both sequences share a significant degree of sequence identity along the entire length of both sequences. Following the Needleman-Wunsch algorithm, Smith and Waterman developed an algorithm for finding local alignments. Local alignment algorithms compare two sequences to find local regions of high identity or similarity. Local alignment tools are useful for highlighting functional domains or sequence motifs shared by two proteins regardless of whether they share a common function.

How does one determine if two sequences are significantly related? A rule of thumb is that if two proteins share 35 percent or more amino acid identity over a span of 100 or more amino acids, they probably are significantly related. It usually is more informative to compare protein sequences because they can identify homologous sequences, indicating shared protein function, from organisms that last shared a common ancestor more than 1 billion years ago. DNA sequence comparisons typically can look back only 600 million years. DNA sequence alignments are important in other areas, such as searching for point mutations, analyzing the identity of cloned cDNA fragments or confirming the identity of a DNA sequence in a database search.

Many sites give the public the opportunity to perform pairwise sequence alignments. One site that provides the user with options to perform either a local or global alignment with either a protein or DNA sequence is the EMBL-EBI sequence analysis site[1]. This site also gives you the option to choose between several different scoring matrices, where the default matrix is BLOSUM62. NCBI also has a site[2] that will perform pairwise alignments, but these only are local alignments. In this case you also have the option to choose from several matrices, where the default matrix also is BLOSUM62. Note that there are several alignment programs that are available on this site: *blastp* for comparing two protein sequences, *blastn* for comparing two DNA sequences, *blastx* for comparing a translated DNA sequence to a protein sequence, *tblastn* for comparing a protein sequence to a translated DNA sequence and *tblastx* for comparing a translated DNA sequence to a translated DNA sequence.

The pairwise alignment algorithm is one of the basic operations of bioinformatics and is the foundation of many other tools, such as gene prediction and protein structure prediction. This algorithm allows scientists to determine the relationship between any two sequences, and the degree of relatedness that is observed helps one to form a hypothesis about whether they are homologous.

## Related bioinformatics papers of interest…

- http://www.fcc.chalmers.se/~marina/files/BioI_PairAlign_2005.pdf
- Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970 Mar; 48(3):443-53
- Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981 Mar 25; 147(1):195-7

*About Us…*

The OUHSC Laboratory for Genomics and Bioinformatics is a full-service genomics facility offering DNA sequencing (small- and large-scale projects), microarray design and hybridization and other services, including bioinformatics support. Edgar Scott is the INBRE Multicampus Bioinformatics Education Specialist, responsible for fostering the development of bioinformatics education on 14 undergraduate campuses in the state of Oklahoma, and coordinating INBRE-related bioinformatics activities with the INBRE Bioinformatics Core. The University of Oklahoma is an equal opportunity institution. This publication is printed and issued by the University of Oklahoma; the cost of $150 was paid by OK INBRE.