Crystal Ann dela Torre   Bioinformatics
December 12, 2003   Final Topic 1

On June 26, 2000, the first complete draft of the human genome was announced by Celera and the US government (National Institute of Health, Department of Energy, etc.)—it was ~3.2x10^9 base pairs (Lesk, 2002). This has added a tremendous amount to a continually growing database of sequences comprised of many organism genomes. There are related databases with amino acid sequences of proteins. To compare a test sequence to a database of sequences several things had to be developed: organized databases, alignment algorithms, scoring matrices, gap penalties and methods to evaluate statistical significance. Here I will discuss the algorithms and scoring matrices developed.

There have been many algorithms developed over the years to compare sequences and to generate sequence alignments. The first were dynamic algorithms which are mathematically rigorous and determine the best alignment of a pair of sequences given certain parameters. These take up computational memory and time that is proportional to the product of the length of the sequences being compared. Two examples of dynamic algorithms for pair-wise alignments are the ones developed by Needleman and Wunsch in 1970 and one developed by Smith and Waterman in 1981. The Needleman-Wunsch algorithm creates a global alignment by comparing a pair of sequences through a two-dimensional matrix, creating the best alignment of both sequences over their entire lengths. The Smith-Waterman algorithm is designed to find short segments of high sequence similarity. For a multiple sequence alignment, MSA, a matrix of greater dimensions is required. Often Hidden Markov Models are used to generate a fully probabilistic representation of these more complex alignments.

Dynamic algorithms are very computer intensive, so heuristic algorithms have been developed which are faster and require less computer memory. Heuristic algorithms are much more practical for comparing a test sequence to an entire database of sequences—however it does sacrifice in sensitivity. Instead of creating a full similarity matrix, these programs compare short words or “tuples” (pairs or more of sequence) in a hash table to find areas of local alignment. Heuristic algorithms include basic local alignment search tool (BLAST) created in 1990 and fast alignment (FASTA) created in 1988 by William Pearson.

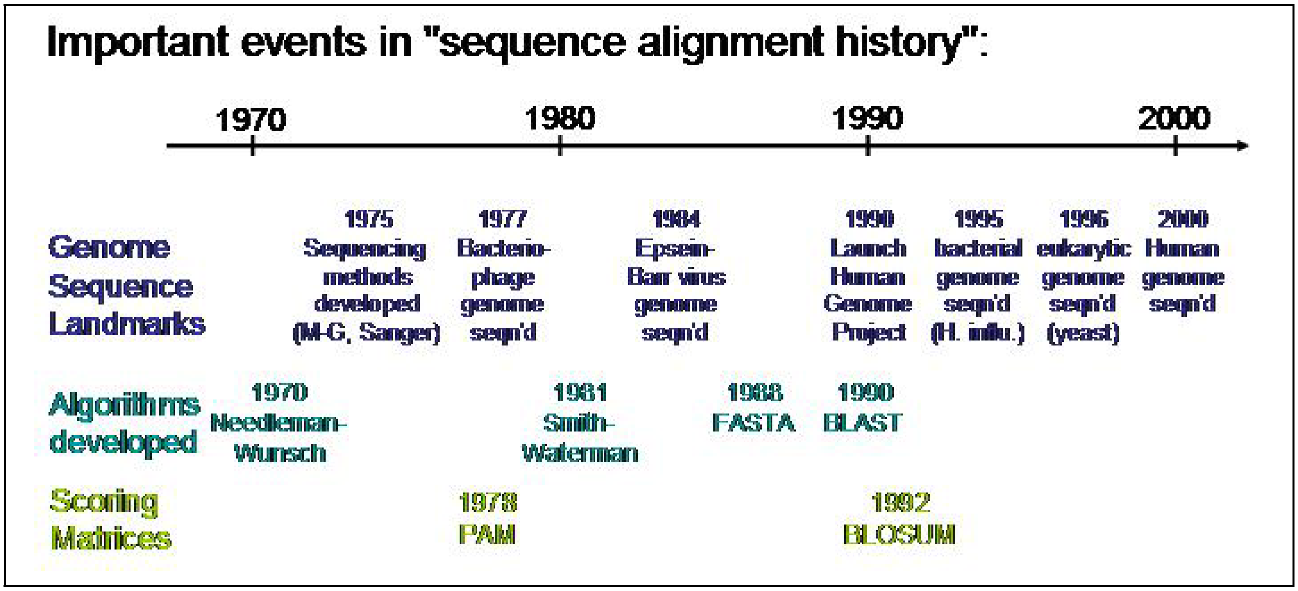
There are several scoring matrices developed for sequence comparisons—particularly for protein sequences. In 1978, Dayhoff created a series of scoring matrices that are evolution-based—based on mutational rates found in paralogous protein sequences. Different matrices represent the percent acceptable mutation for a given evolutionary distance. For example, the PAM-78 matrix represents the percent acceptable mutations for sequences more evolutionarily related (compared to PAM-250). In 1992, the Henikoff’s created the BLOSUM matrices for different percent identity levels (1%-100%). These matrices take into account the fact that not all regions of a protein have a uniform evolutionary rate of mutation.

The ability to do sequence alignments has transformed methods of biological research. For example, when studying a protein of interest through a yeast two-hybrid assay I find that it interacts with an unknown protein X. I can find the sequence of protein X then BLAST against the database to identify characterized proteins with similar sequence. Similar sequence often implies similar function. Another important use of sequence alignments is in determining conserved regions or motifs within related sequences. In many cases, the residues that are most highly conserved tend to be important for function. For example, in serine proteases, there is a triad of highly conserved residues (including serine) for this enzyme family because they are involved in catalysis. The Smith-Waterman algorithm is best for determining conserved motifs which again can be compared to motifs in the database to infer function.

Crystal Ann dela Torre
December 12, 2003

Bioinformatics
Final Topic 1

Figure:



Important events in "sequence alignment history":

1970
1980
1990
2000
Genome Sequence Landmarks
1975 Sequencing methods developed (M-G, Sanger)
1977 Bacterio-phage genome seqn'd
1984 Epsein-Barr virus genome seqn'd
1990 Launch Human Genome Project
1995 bacterial genome seqn'd (H. influ.)
1996 eukarytic genome seqn'd (yeast)
2000 Human genome seqn'd
Algorithms developed
1970 Needleman-Wunsch
1981 Smith-Waterman
1988 FASTA
1990 BLAST
Scoring Matrices
1978 PAM
1992 BLOSUM

Reference List:

Lesk, A. (2002) Introduction of Bioinformatics. Oxford University Press, NY

Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol*. **48**: 443-453.

Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol*. **147**: 195-197

Altschul, S., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol*. **215**: 403-410.

Pearson, W. R., & Lipman, D. J. (1988) *P.N.A.S.* 85: 2444-2448

Dayhoff, M.O. et al. (1978) Atlas of Protein Sequence and Structure. Vol. 5, Suppl. 3 National Biomedical Research Foundation, Washington D.C.

Henikoff, S. & Henikoff, J.G. (1992) *PNAS* **89**: 10915-10919.

Crystal Ann dela Torre
December 12, 2003

Bioinformatics
Final Topic 2

If a new prokaryotic genome has been identified, there are many questions that can be asked regarding the genome. The main goal in a computational functional genomics analysis would be to determine the function of each predicted ORF's gene-product within the genome. Such a task is daunting, but I propose the following study of this new genome. Information obtained from sequence analysis can be 'combined' with information from experimental data (expression and interaction assays) into Bayesian networks to determine probabilities that an ORF gene-product belongs in a given functional category (e.g. metabolism, biosynthesis, etc.).

Several clues can be found within the sequence of each ORF as to its functional category. The sequence can be scanned for any known functional or structural motifs or localization signals in the genomic database. For better identification of protein motifs/signals for protein gene-products, it is often better to translate the given DNA sequence first (due to degeneracy in the triplet codons, protein sequence is better conserved evolutionarily). Also, whole-ORF comparisons to the database can be done to find the most similar characterized gene-product (protein or RNA). Finally, unique to prokaryotic sequences are the high incidence of operons, or genes transcribed under the same regulational control that have related functions. Approximately a quarter of all *E. coli* genes are in operons. Therefore, it is worth taking into account the predicted function of neighboring genes within the genome.

Expression data of each ORF obtained from microarrays are quite useful in predicting function. First of all, it can provide experimental evidence that a predicted ORF is indeed expressed *in vivo*. If a unique sequence from each ORF is placed on a microarray, the array can be probed with fluorescently labelled total RNA from the cell to monitor expression of that ORF. When total RNA samples are tested at different times in the cell life cycle, an expression profile over time can be generated for each ORF. Clustering of the expression profiles can categorize ORFs temporally—if certain ORFs are expressed at the same time they may have similar function. Likewise if the gene products are localized to the same cellular compartment, they may function together. Specific functions can be tested under specific growth conditions. For example, if the cells are grown on media lacking particular nutrients, then expression of proteins involved in biosynthesis should differ. It is highly likely that proteins that interact *in vivo* will function in related pathways. Thus experimental evidence for *in vivo* interactions, obtained through yeast two-hybrid or immunoprecipitation assays conducted between proteins from predicted ORFs also help in functional categorization of ORFs.

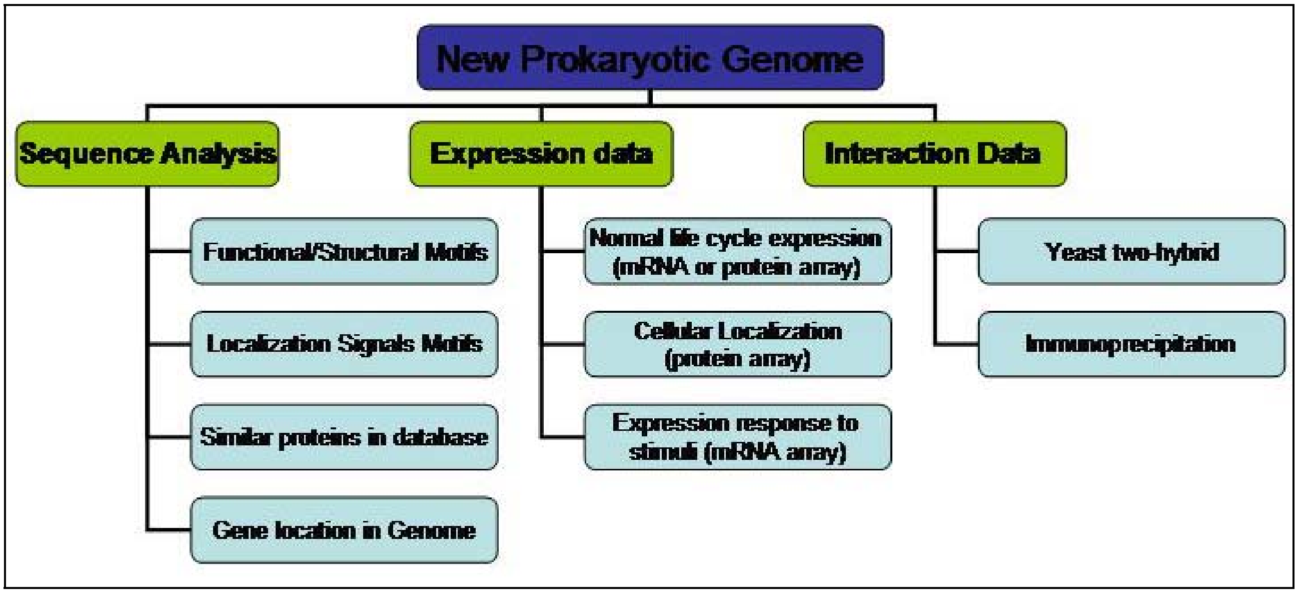
To take into account all the data to categorize the function of each ORF, Bayesian networks are most useful. Bayes rule for combining uncorrelated bits of information allows for differential weighting of each type of data. This is important, because some assays are more reliable than others. For example, yeast two-hybrid assays often give false positives and are not as accurate as immunoprecipitation experiments. Previous work with Bayesian networks combining similar data was able to predict function with ~50% accuracy (Jansen, 2003), which is much better than any data set taken individually.

When the function of each ORFs is predicted, one can start to answer more interesting questions regarding this new organism. Comparisons to other known organisms may identify unique protein functions that may illuminate other areas of research.

Crystal Ann dela Torre
December 12, 2003

Bioinformatics
Final Topic 2

Figure:


New Prokaryotic Genome
Sequence Analysis
Expression data
Interaction Data
Functional/Structural Motifs
Localization Signals Motifs
Similar proteins in database
Gene location in Genome
Normal life cycle expression (mRNA or protein array)
Cellular Localization (protein array)
Expression response to stimuli (mRNA array)
Yeast two-hybrid
Immunoprecipitation

References for paper and figure:

Lesk, A. (2002) Introduction of Bioinformatics. Oxford University Press, NY
Jansen, R., et al. (2003) *Science* **302**: 449-454.