Sequence Analysis

The genome projects are making large amounts of DNA sequence available. The problem now is attempting to understand the information. In most cases, these analyses involve the use of computers. In fact, it has become apparent that without computers, the types of analyses that are necessary to analyze the information being generated would be impossible.

Types of analysis

Several types of computer analyses are frequently performed on DNA sequences. One simple analysis is the search for restriction endonuclease cleavage sites. Finding restriction sites is useful for cloning experiments and is a useful first step in planning to perform restriction mapping of a plasmid.

More complicated analyses, such as looking for genes and looking for sequences similar to those of known genes are discussed further below.

DNA sequence analysis

The genomes of a number of prokaryotic and eukaryotic organisms have been sequenced, including those of widely used experimental organisms such as the enteric bacterium *Escherichia coli*, baker's yeast *Saccharomyces cerevisiae*, and the fruit fly *Drosophila melanogaster*. Sequencing of the human genome is essentially complete, with a draft version of the sequence available for study. As a result of these projects, we have access to a very large amount of DNA sequence information.

In most cases, we are interested in the protein sequence that results from the nucleic acid, not the DNA sequence itself. This means that the DNA sequence must be analyzed for the presence of coding regions or other regions of interest.

Finding genes

Prokaryotic organisms lack introns. This means that any sequence that contains a start codon at one end, and an in-frame stop codon may code for a protein. Such a sequence is called an **Open Reading Frame** (ORF). Differentiating ORFs from actual expressed genes, however, may be somewhat more difficult.

One method for finding a gene involves comparison of the predicted protein sequence (*i.e.* the predicted translation sequence based on the DNA sequence) with the sequences in the database of known protein sequences. The advent of fast computers and improved algorithms has meant that it is possible to compare any sequence against the entire database in a very short time (a few seconds to a few minutes), in spite of the fact that the sequence databases are rapidly growing in size.

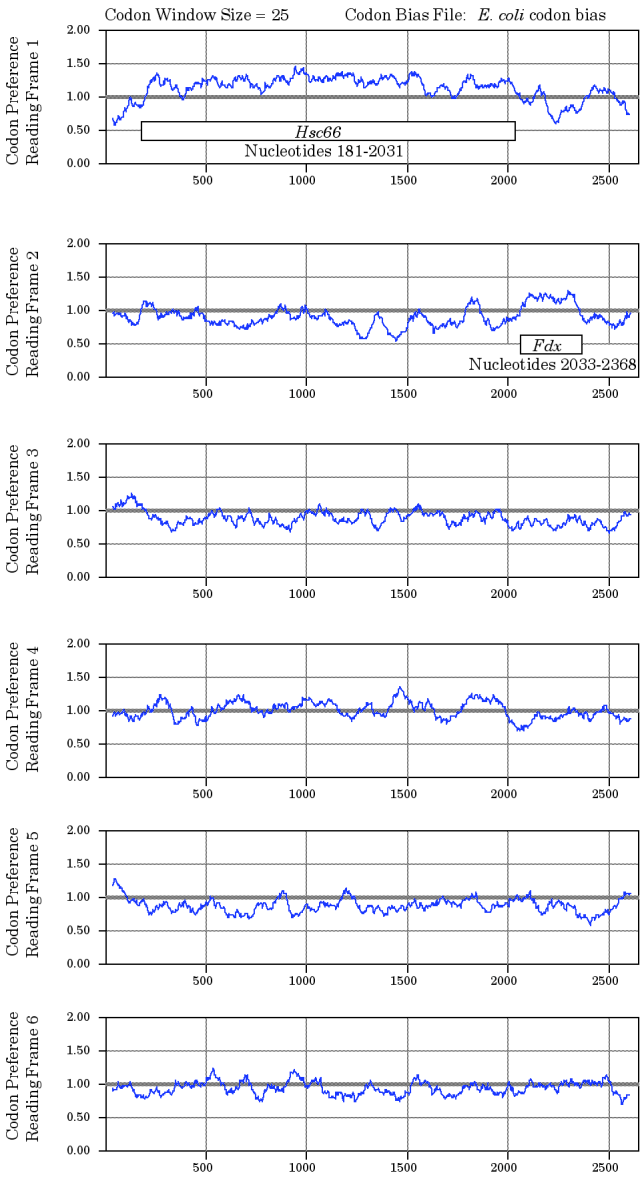
The most frequently used algorithm for comparing one sequence with any other sequence is the **BLAST** algorithm (Basic Local Alignment Search Tool). The NCBI and a number of other web-based servers allow BLAST searches against different databases. In addition, the NCBI website (http://www.ncbi.nlm.nih.gov/) has a wide variety of help and tutorial information that assists in understanding both how to

perform BLAST searches and how to interpret the results.

For prokaryotic organisms it is also, in principle, possible to analyze the DNA sequence for information that may suggest whether a particular sequence is transcribed.

One method involves searching for promoter elements and ribosome binding sites in the sequence near potential ORFs. Although both promoters and ribosome binding sites are highly variable, appropriate algorithms can search for the presence of consensus sequences. Unfortunately, these algorithms have their limitations, and tend both to find sequences that are not actually expressed, and to miss sequences that are expressed.

A second method involves analyzing for coding regions by analyzing **codon bias**. Because organisms tend to use different codons for the same amino acid with different frequencies, it is possible to look for patterns within the codons present in a DNA sequence. For example, of the six codons for leucine, *E. coli* tends to use CTG very frequently, and to use CTA very rarely. A sequence that contains a high proportion of frequently used codons is likely to be expressed.



An example of an analysis for coding regions in a portion of the *E. coli* genome is shown at right. The combination of codon bias and ORF analysis reveals the present of two genes within this DNA sequence shown. Note that in reading frame 1, a long sequence has a high relative codon preference, with a high proportion of common codons. At the end of the ORF, the codon preference decreases in reading frame 1. The codon preference in reading frame 2 has a peak that corresponds to an ORF. None of the other possible reading frames have significant peaks on these plots in this 2.6 kb DNA sequence. Experiments designed to test for transcription confirmed that these predicted sequences were actually expressed.

These methods work for prokaryotes because prokaryotes have relatively small

genomes, and because the vast majority of their genomes are devoted to coding for proteins. For higher eukaryotes, however, the only effective algorithms currently available for finding genes involve similarity searches.

Similarity searches are used to compare DNA sequences to known protein coding sequences and to Expressed Sequence Tags (ESTs). ESTs are sequence fragments discovered by sequencing cDNA libraries. Their presence in the cDNA library is an indication that the sequence is expressed, although in many cases, the function of the sequence in the EST is unknown.

Why is sequence similarity useful?
All known life appears to be related. All organisms use the same genetic code, and all organisms use the same set of amino acids and nucleotides. They also tend to have proteins with similar sequences performing similar functions.

Two proteins that exhibit a significant degree of sequence similarity are "**homologous**", meaning that the two proteins are considered to have a common ancestor.[3] Homologous proteins contain similar sequences. While this could be a coincidence, a great deal of evidence has accumulated to indicate that proteins with similar sequences also exhibit similar structures and similar functions.

Percent Sequence **Identity** = [# of identical residues]/[total # of residues] x 100

Percent Sequence **Similarity** = [# of similar residues]/[total # of residues] x 100

Note that the term "similar" is open to interpretation, but in general describes a residue with similar chemical properties (*e.g.* aspartate and glutamate, which both contain carboxylate side chains).

Structure and function

For any protein to have a function, it must have a three-dimensional structure. However, structural data is difficult to obtain. Structures are available for only a relatively small number of proteins. A check of the Protein Databank (the main repository for structural data) website indicates that more than 20,000 structures are deposited in the database. However, only about 20% of these represent unique proteins that have been solved; the remaining files contain information concerning relatively minor variants of previously solved proteins. While the variants are frequently useful, such as different conformations of the same protein, or different proteins that have similar folds but different functions, many of the structural variants, such as those of proteins with point mutations or species variants of a particular protein, have relatively little new information.

---

[3] Homology is "yes or no"; some papers discuss "percent sequence homology", but this is an incorrect usage. The term "homology" is derived from the evolutionary biology term to refer to the property of evolutionary relationship, and was misused by molecular biologists; we are (slowly) attempting to correct this usage so that molecular biologists and evolutionary biologists will both use the term correctly.

In contrast, because determining the three-dimensional structure of a protein is much more difficult than sequencing the DNA that codes for the protein, we have sequence information available for a great many proteins which have not had their three dimensional structure determined.

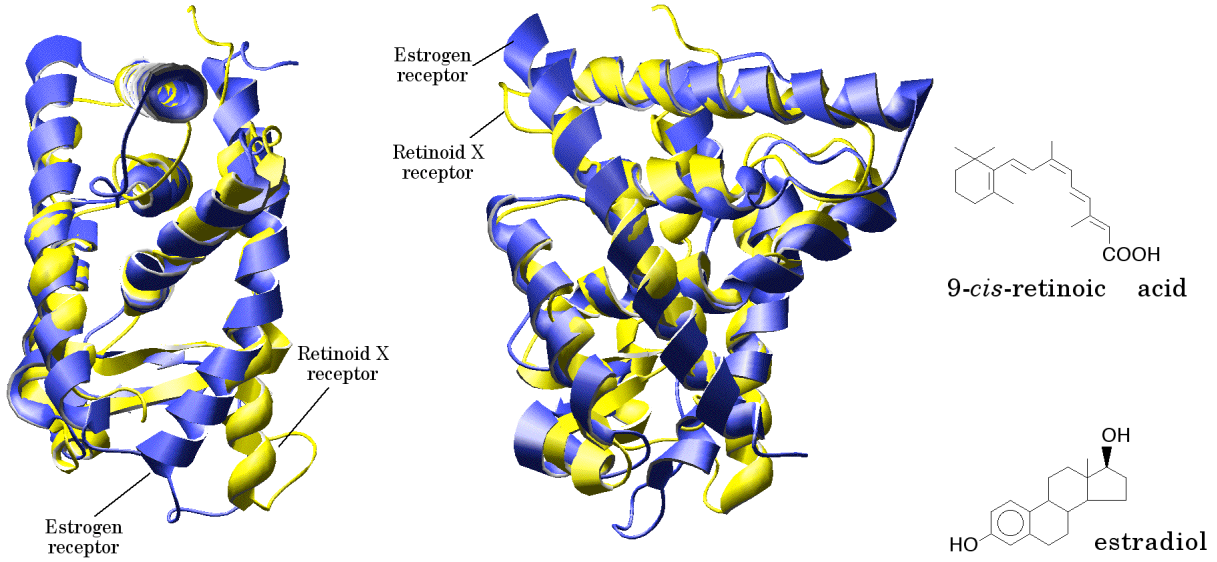
In the early 1960s, C.B. Anfinsin and F. White unfolded a protein called ribonuclease. Removing the agent that caused the ribonuclease to unfold allowed the protein to refold into conformation indistinguishable from the original. This means that, even in absence of cellular components, the protein can fold properly, and therefore all of the three-dimensional structural information must be contained within the amino acid sequence.

In theory, since the three-dimensional information is included in the linear amino acid sequence, it should be possible to predict the structure based entirely on the sequence. However, the three-dimensional structure of any given protein is the result of very large numbers of weak interactions. Because current theory cannot calculate the strength of each interaction with sufficient accuracy, protein structures cannot be predicted from sequence information alone. This leaves us with the question: "what **can** we do with all of this sequence information?"

Examination of the three-dimensional structures that have been solved has revealed that similar sequences nearly always fold into similar three-dimensional structures. This means that:

1) The "same" protein from different species has the same overall structure, in spite of differences in sequence.
2) Proteins can be "related" – proteins of slightly, or in some cases greatly, different function may exhibit sequence similarity, and similar overall structures.
3) It is possible to model a structure of one protein based on the structure of a similar one; the usefulness of these models is somewhat variable, but often results in insights into aspects of the new protein.

The superimposed structures shown below are an example. The blue structure is a portion of the estrogen receptor; the yellow structure is the retinoid-X receptor. These proteins both bind small molecule (drawings of which are shown on the right). The proteins share some similar functions (such as binding small molecules), and clearly have similar structures. However, these proteins share only about 25% sequence identity.



Proteins exhibiting a high degree of sequence similarity (or greater than ~10 to 15% sequence identity) are usually considered to be evolutionarily-related. Evolutionarily-related proteins exhibit similar structures and similar functions. Thus, by comparison of sequences of newly discovered proteins with sequences of known proteins, it is frequently possible to predict the function of the new protein with some degree of accuracy.

Structure prediction

In difficulties associated with a proper theoretical approach to solving the protein-folding problem, many attempts have been made to predict protein structure for proteins that have not been solved.

One attempt is based on an **analysis of secondary structure**. In 1978, Chou and Fasman published an estimate of the likelihood that any given amino acid would be in a type of secondary structure. They determined the likelihood based on an analysis of the protein structures that had then been solved. These values can be used to predict the secondary structure for other proteins that have been sequenced.

The Chou-Fasman prediction is of some use, because it tends to be correct about 70% of the time. Unfortunately, this means that for any given protein, it is likely to predict the incorrect secondary structure for 30% of the protein. In addition, the Chou-Fasman approach (and later, similar approaches) does not predict how the secondary structural elements are arranged in the protein.

A second attempt is based on the fact that similar sequences have similar structures. The technique of **homology modeling** uses the fact that the relative position of secondary structural elements from related proteins is usually very similar. The modeling procedure then attempts to locate the secondary structural elements in the new protein based on a sequence alignment with the known protein. This is followed by an attempt to find minimal energy positions for the loop regions (*i.e.* the sequences that connect the secondary structural elements). Homology models usually have an overall structure that approximates the actual structure.

Difficulties arise, however, when sequence differences make the sequence alignment uncertain; in addition, our current ability to accurately model non-regular secondary structure remains extremely limited.

Molecular Evolution

Proteins with a high degree of sequence similarity are usually considered to have a common ancestor. Invariant residues within such proteins often are critical for proper folding, for activity, or both. The variant residues are assumed to have fewer constraints, and therefore to be less critical for function; mutations in these residues may have no effect, or may be part of the evolutionary process that creates new functions in old proteins. Careful analysis of the sequence deviations between variants of the same protein in different species yields information about both protein structure/function relationships, and about the evolutionary changes that occur as species diverge.

