

Micah Acinapura
Senior Seminar Fall 2003
Survey Paper

Computational DNA Sequence Analysis

Introduction

While all the sciences help people expand their knowledge of our universe, biology holds a special place because it studies life. In recent years molecular biology the most fundamental aspect of biology has been of particular interest. Being the smallest level of life. Molecular biology covers the inner workings of an organism trying to define its life, however ever molecular biology has its roots planted in DNA. The DNA of an organism, referred to as the genome, contains all the genetic information for that organism's life to be created, and sustained.

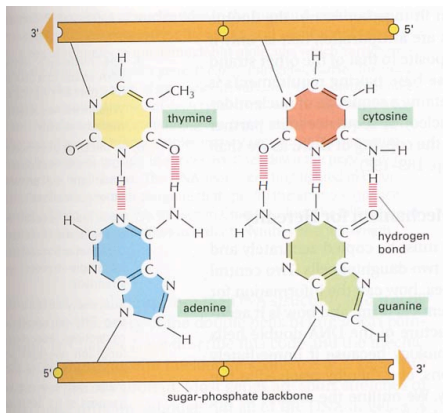
With all the information for an entire organism present in its DNA, the study of DNA can provide a wealth of information in all aspects of life. For instance the study of the DNA of a tiny bacteria can give us great insight into the working of the human body. There are all sorts of applications of DNA analysis that warrant its study, perhaps the most famous these days being in the Judicial system with the DNA testing of blood samples. However there are many practical medical uses, including the study of proteins, which define an organism and its function.

Evolutionary biology is another field where the discovery and study of DNA has become a major factor in its methods. Since the DNA stores all the genetic information for an individual, it is what is passed on when the organism reproduces. An organism becomes a similar life as its parents because all its genetic information was a copy of its

parents. This is the basic idea between evolution, that offspring become different from their parents by mixing DNA, and those that function better create more copies their DNA the those who are less functional. Thus the most functional DNA sequences are replicated the more then others.

DNA to Protein

With the great importance DNA brings to the table, it brings with it a surprisingly simple structure. DNA exists in strands, usually bonded together in a double-helical structure. These strands are comprised of a sugar-phosphate backbone and with nucleotide bases attached to it. A nucleotide base is one of four options, adenine,



thymine, cytosine, or guanine. A,T,C, or G (as show to the left (11)). These are simple molecules, containing one or two Carbon-Nitrogen rings. Each of the four bases will bond to exactly one other base, making what's called a base-pair: A bonds to T and C to G, these base pair interactions (H- bonds) are

what keep two DNA strands together. Two strands that bond to each other complimentary base sequences they are called complementary strands. These DNA strands are stored in the nucleus of a cell in eukaryotes, or are loose in the cytoplasm of a prokaryotic organism.

For eukaryotes, the storage of DNA in the nucleus provides multiple benefits. One is that it keeps the DNA from being damaged, however more importantly it helps with cell division. Cell division, being the main way in which organisms grow, is a large

part of the life cycle, and requires many steps of DNA preparation. In order for a cell to divide, there must be enough DNA for two copies of the cell, this is done through a process of DNA Replication. DNA replication is a complex process regulated by proteins. It begins as a protein binds to a DNA helix and rips the complimentary strands apart. Then other proteins enter and begin to read the two DNA strands. These strands are used as a template for the synthesis of new strands. When reading a base from the template strand, the proper base to bind to the template strand (i.e. T if and A was read), will be added to the growing strand. This process is done repeatedly, and even goes back to check that the last pair it made was correct, minimizing errors in replication. The new strands that are created then bond to their template strands, make two copies of the original DNA helix.

Transcription works very much like DNA replication, only we are copying the DNA into a different form, RNA. Transcription, like replications, works by splitting the DNA apart and using a strand as a template to create the new growing strand. However there are a few major differences that separate the two. In transcription, a strand of DNA is still used as a template, however wherever we would have put a T in DNA replication we put a U (for Uracil), since RNA is made of A,U,C and G, not A,T,C and G. Also when RNA strands are created, they don't bind to their template strand, they are set loose in the cytoplasm in order to be used. Another main difference is that in replication all of the DNA is copied, where as in transcription only select portions of the DNA are converted into RNA.

RNA acts as an intermediary between DNA and proteins. Select RNA strands are produced from portions of the DNA, which code for the protein that a cell wants to

produce. Since the DNA holds all the information for an organism, each cell doesn't want to produce it all, so it produces RNA for those portions of the genetic data it wants to produce. All the RNA that is produced is in three types: rRNA, which makes up ribosomes (molecules that make proteins), tRNA which aid the ribosomes, and mRNA, the most important one. MRNA strands store the code that is to be produced into a protein through the process of translation.

RNA strands, having a very similar composition to DNA strands, are just a sequence of nucleotide bases, A,U,C and G. In the production of proteins the information is read from the RNA strands in groups of three bases, called codons. These codons are interesting because they represent amino acids, each three letter grouping represents an amino acid. Since there are a lot of possible codons, and only 20 amino acids, some amino acids are coded for by multiple codons. As the ribosomes (the protein production machinery) move along a RNA strand, they read a new codon, which tells it which amino acid to add next to the growing protein chain. There are two special codons, the stop and start codons, which tell a ribosome at which locations in the RNA sequence it should start and stop the codon-to-amino acid conversion, or translation.

A protein is basically an amino acid chain folded up upon itself. This sequence is what defines a protein and distinguishes it from other proteins. While all of the 20 amino acids have the same base structure, they each have a different functional group (a group of atoms where an important interaction takes place). This functional group and the interactions between it and other molecules are what makes each amino acid special, and put together they will interact different ways. It is this property of amino acids that really makes proteins functional. The folding of the protein is based upon these interactions,

and is how the amino acid sequence of a protein can determine its structure. Once a protein is folded in to its proper conformation (or shape), the functional groups of the amino acids on its surface are what determine how the protein will function in a cell. These groups will bond only with the specific molecules that it has been designed to interact with. It is this specificity of proteins that distinguishes and makes possible all life.

Computational Aspects

At first it may be hard to image how we can use computer to calculate biological interaction, however with a certain view it becomes quite simple. For computations involving DNA and RNA one can observe a lot by representing a DNA/RNA sequence as a string. Since the nucleotide bases have historically been represented by the letters A,T,C,G and U this leads to an easy transition to the string form. Proteins as well can be represented in this form by using the sequence of amino acids that form a protein as a linear string representation of the protein. In recent history the genomes of many organisms, including humans, have been mapped out and their sequences have been stored in readily available databases. Using biological data from experiments in laboratories generalizations can be made about types of sequences. With DNA/RNA some examples would be, knowing the properties of a protein coding region, or knowing that promoter and terminator regions always have specific sequence within them. For Protein a lot of computation is put into how specific amino acids will bind to each other and to certain molecules, this can help to determine how proteins fold on themselves (how they make their shape) and what that conformation (shape) means in terms of the protein's function.

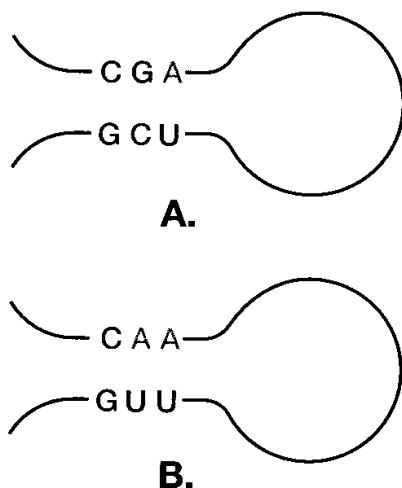
When studying DNA one of the largest problems is the large size of the data set. For example, in 1998, 5 years ago, 100 million base pairs of genomic information were being sequenced per day. (2) The same is true for the study of protein interactions, due to the folding nature of amino acid sequences in a protein, the amount of interaction for each amino acid can get quite large. As the number of amino acids increase the amount of interactions increases exponentially. While this is still a problem today, the increase of popularity of distributed systems should help to alleviate some of the computational challenges of current sequence analysis. Due to the sequential nature of the data of DNA and Proteins, it is easy to turn analysis algorithms in to distributed ones, which provides a solution to the sheer amount of calculations that need to occur during sequence analysis.

There are two main tasks in computational sequence analysis, that of selecting two similar sequences, or sequence alignment, and that of determining the conformation of RNA and proteins from their sequence, or structural prediction. Both have varying issues and solution to their problems, however the conclusions that can be drawn from their results cover the same ground.

Sequence alignment has two different types of classifications. The first is that of alignment of pairs of sequences or alignments of multiple (three or more) sequences. What distinguishes the two, aside from the reasons for their study, is mostly the algorithms involved in determining the alignment, specifically in the scoring involved in matching bases within a sequence. The next distinction is between global alignments and local alignments. A global alignment involves the entire sequence, where an optimal alignment has the most matches across the entire sequence, where as a local alignment involves finding subsequences that have the densest regions of matches. Global

alignments are more suitable for sequences that are similar in composition and length, whereas local alignments work better for sequences that are quite varied.

Structural prediction is also divided into two main categories, RNA structures and protein structures. The prediction of RNA structures is a similar problem to that of sequence alignment; it also has a distinction between single or multiple alignments, however it also varies in significant ways. “RNAs conserve a secondary structure of base-pairing interactions more than they conserve their sequence. ((9) p 260)” This, as



illustrated by the figure(10), is the idea that while two subsequences of RNA can have different bases, these “covarying” sequences can produce the same secondary structure, and thus be functionally equivalent. Obviously this is not always the case, but it is a common occurrence.

Structural prediction of proteins is a much different, while we are still dealing with a sequence, now it is comprised of amino acids instead of nucleotide bases. Here we study the interactions between amino acids, particularly their side chains (since that is what distinguishes them), and how this determines a protein’s conformation . Since the conformation of a protein is the basis of its function, knowledge of the structure of amino acid sequences can help us cluster them into functional groups.

While there are many areas here to study and research, they are all aimed at determining the biochemical functions of specific regions, and how regions of similar function can be grouped into categories that make them easier to classify and understand.

In the case of DNA one main topic is the identification of regulatory regions within a sequences, i.e. the binding site on a DNA strand of a regulatory protein. Another “hot topic” is the identification of protein coding regions within DNA sequences, here the issue being that most of the huge genomes of eukaryotes don't code for proteins. For proteins we can identify specific sequences of amino acids that will “always” fold into certain *motifs* and then identify the significance of these motifs.

Determining the function of a protein, DNA, or RNA sequence helps us to classify that sequence into groups of similar sequences that all share characteristics in terms of their function. This can be very useful in identify the functions of new sequences, by giving more data to reference against, but it can also be very useful to evolutionary biologists in what is known as phylogenetic prediction.

Phylogeny is “the evolutionary development and history of a species or higher taxonomic grouping of organisms (websters.com)” and is directly linked to sequence analysis. Before sequences analysis was a method of study, phylogeny was done observationally, using physical characteristics of a species or group of organisms. With the advent of DNA and sequence analysis, phylogenetic studies can now take a different approach. The evolutionary theory that all complex organisms evolved from few simple ones (or maybe even one base organism) adapts well to the study of DNA sequences and protein structure. DNA, being the storage unit of genetic information, is ideal for phylogeny because the DNA of a complex organism evolved from the DNA of a simpler one. With that, the DNA, and thus the evolution, of complex species can be linked in key areas (I.e. areas the code for a common trait) of their genome, due to their mutual inheritance from a simpler organism. Proteins fall into this model as well because their

specificity combined with their expression define an organism. Along similar lines as with DNA, the proteins of an organism can be sequenced and that are similar and/or perform similar functions can be used to link species in a phylogenetic study or even in a phylogenetic tree.

My Study

I plan to build a tool for DNA sequence analysis using multiple sequence alignment. My hope is to structure the tool so that the type of analysis can be module based, this adding more modules in the future would increase the functionality of the tool. For maximum computational effectiveness this should be a distributed tool and will most likely be implemented in C using the PVM environment. The results of the analysis should be presented to the user in some visual fashion, most likely web-based (CGI?) with a graphical representation, even if it is as simple as displaying aligned portions of sequences.

The analysis module I plan to implement will try to identify binding sites on eukaryotic DNA strands for transcription factors that bind in the dimmer motif. Based on an algorithm for identifying binding sites for transcription factors in bacterial DNA by Li et al. (4), I hope to extend to extended this algorithm to eukaryotic DNA sequences. While Li et al's algorithm is based on the fact that most bacterial transcription factors bind in the dimmer motif, yet eukaryotic transcription factors bind in a variety of motifs, there are still enough eukaryotic transcription factors that bind in the dimmer motif that I feel this endeavor is worthwhile.

While this project does involve a lot of work in the biological field, a significant portion, if not most of the work, will be involved in the design and implementation of the

software system, requiring my knowledge of software engineering, parallel programming, algorithm implementation, and perhaps some OpenGL for the rendering of a visualization of the results.

Bibliography

1. Peter Friedland and Laurence H. Kedes. "Discovering the secrets of DNA." Communications of the ACM, November 1985. vol.28 no.11 pp.1164-1186.
2. T. Head-Gordon and J.C. Wooley. "Computational challenges in structural and functional genomics." IBM Systems Journal, 2001. vol 40. no 2. pp 265-296.
3. T. Inman, H.R. Flores, G.D. May, J.W. Weller, C.J. Belo "A high-throughput distributed DNA sequence analysis and database system." IBM Systems Journal, 2001. vol. 40. no.2. pp. 464-486.
4. Hao Li, Virgil Rhodius, Carol Cross, and Eric D. Siggia. "Identification of the binding sites of regulatory proteins in bacterial genomes." Proceedings of the National Academy of Sciences of the United States of America, September 2002. vol. 99. no. 18. pp11772-11777.
5. Burkhard Morgenstern, Andreas Dress, and Thomas Werner "Multiple DNA and protein sequence alignment based on segment-to-segment comparison." Applied Mathematics October 1996. vol 93. pp. 12098-12103.
6. Eran Segal and Yoseph Barash and Itamar Simon and Nir Friedman and Daphne Koller. "From promoter sequence to expression: a probabilistic framework" Proceedings of the sixth annual international conference on Computational biology, 2002. ACM Press. pp. 263-272
7. Erez Hartuv and Armin Schmitt and Jürg Lange and Sebastian Meier-Ewert and Hans Lehrach and Ron Shamir "An algorithm for clustering cDNAs for gene expression analysis." Proceedings of the third annual international conference on Computational molecular biology, 1999. ACM Press. Pp. 188-197.

8. Zheng Zhang and William R. Pearson and Webb Miller “Aligning a DNA sequence with a protein sequence” Proceedings of the first annual international conference on Computational molecular biology, 1997. ACM Press. pp.337-343.
9. Durbin, Eddy, Krogh, Mitchison. “Biological sequence analysis.” Cambridge University Press, 1998.
10. David Mount. “Bioinformatics: Sequence and Genome Analysis.” Cold spring Harbor Laboratory Press, 2001.
11. Alberts, Bray, Johnson, Leviw, Raff, Roberts, Walter. “Essential Cell Biology.” Garland Publishing, 1998.