Computer Applications in Molecular Biology

Russ B. Altman, MD, PhD
Assistant Professor of Medicine,
Section on Medical Informatics
Stanford University Medical School
email: altman@smi.stanford.edu

Informatics for the Human Genome Project

The genetic material that we inherit from our parents, that we use for the structures and processes of life, and that we pass to our children is contained in a sequence of chemicals known as deoxyribonucleic acids (DNA). The total collection of DNA for a single person or organism is referred to as its genome. The Human Genome Project is an international effort with the goal of determining the sequence of DNA for a human being in order to facilitate the use of genetic information in medical settings. Each year, Science magazine devotes an issue to reporting the progress made on this project. In the United States, the two agencies with the greatest interest in this project is the National Institutes of Health (NIH) and the Department of Energy (DOE), each with an annual budget for this project of about $30 million.

The relevance to medicine of the human genome project can be considered in two ways: short term and long term benefits. The short term benefits are principally diagnostic: the availability of sequences of normal and abnormal human genes will allow for the rapid identification of these genes in any patient (using technology that is based primarily on hybridization phenomena discussed below). The long term benefits will include a greater understanding of the proteins produced from the genome: how they interact with drugs, how they malfunction in disease states, and how they participate in the control of development, aging and responses to disease.

The Basic Flow of Information

The three critical entities in the flow of genetic information are DNA, RNA and proteins. They each are specialized to perform a specific set of functions.

DNA

DNA (deoxyribonucleic acid) stores genetic information for long term use. It is a linear sequence of four bases (A,T,G,C). In humans, the DNA is actually divided into 23 separate linear strands, the chromosomes. There are a total of approximately 3,000,000,000 bases in each human genome. A DNA sequence, therefore, might look like this:

```
AGCTAACTGGACTTCCTAGAAATTGACTAGAGACTATAGACATAGCTTTAA
```

The information contained in DNA is linear (one-dimensional) and yet it is able to specify the information necessary for processes that are not just three-dimensional (the human body is three-dimensional, for example) but four-dimensional. The additional dimension is that of time. DNA not only specifies a static structure, but also specifies the sequence of events during embryological development and aging that occur over time. Because of the critical importance of DNA in holding the architectural plans for life, most organisms have developed a number of mechanisms for guarding the integrity of the DNA so that the sequence remains unaltered (except for a few special exceptions). They key to understanding how 1-dimensional information is transformed into 3-dimensional information requires a knowledge of RNA and proteins.

A strand of DNA does not exist in isolation. Instead, it is associated with a sister strand that has a special property: complementarity. It turns out that the four bases have specific preferences for binding each other. A's bind to T's, and G's bind to C's.   The associations between these bases are referred to as base pairing. Base pairing is used in a number of different ways by the cell. For example, within the cell, a DNA strand will exist really as a pair of strands that look like this:

```
AGCTACTGGACTTCCTAGAAATTGACTAGAGACTATAGACATAGCTTTAA
TCGATGACCTGAAGGATCTTTAACTGATCTCTGATATCTGTATCGAAATT
```

These two complementary strands were discovered by Watson and Crick in 1953 (who postulated that they form a double-helix when they are paired up), and are useful for a number of purposes. Specifically, they 1) provide redundancy of sequence so that if part of one strand is damaged, the information in the other strand can be used to repair the damaged one, 2) if the two strands separate, then the complementary strands can be synthesized in order to produce two complete sets of genetic material (precisely the process used to duplicate DNA in dividing cells).

<u>messenger RNA</u>

Whereas DNA is the archival, long term storage form of the genetic information, messenger RNA (mRNA) is a copy of a subsegment of DNA that is meant for temporary use and then recycling.[1] RNA uses the same four-letter alphabet as DNA, except that U is used instead of T. It turns out that T is metabolically more expensive to make, but is more stable over the long term. It is well-suited to the archival DNA, but is unnecessary for RNA.   U is cheaper (in terms of metabolic cost to produce) and less stable, and is perfect for short term use. RNA typically copies a subsegment of DNA and does not usually exist as a base-paired strand (as does DNA in the example above).

Usually, a segment of DNA is copied into a piece of RNA for the purposes of producing a protein. The process of specifying a protein involves 1) identification of a piece of DNA that encodes a protein, 2) copying of that piece into an RNA working-copy, and 3) using the RNA-copy to direct the production of a protein.

Consider the double-helical, base-paired sequenc

---

[1] RNA also serves a number of other functions. transfer RNA (tRNA) is the molecule that actually recognizes the codon on messenger RNA and brings the appropriate amino acid to be added to the growing polypeptide chain of amino acids.   tRNA has an interesting three-dimensional structure,which relies on base-pairing between different segments of the same RNA--thereby producing a seemingly knotted up structure. Ribosomes are the ensemble of molecules that provide the scaffolding for the production of proteins from the genetic information. They are also made of predominantly RNA (r RNA), and presumably have important three-dimensional shapes.   Finally, there are a number of small RNA molecules within the cell that play a number of roles having to do with transport, control of expression of genes, and other functions.

```
   V                            V
TCGATGACCTGAAGGATCTTTAACTGATCTCTGATATCTGTATCGAAATT
AGCTACTGGACTTCCTAGAAATTGACTAGAGACTATAGACATAGCTTTAA
```

An RNA copy is made from the region specifying a protein (it begins with a special sequence ATG that means "protein starts here", and a special symbol TGA that means "protein ends here.") This results in a working copy of the DNA (remember U = T for RNA):

```
AUGACCUGAAGGAUCUUUAAG
```

This copy is provided to the cellular machinery that makes a protein. The machinery takes the bases as groups of 3 and uses the three bases as the code to determine which amino acid of the protein should be attached next (see discussion below for explanation of proteins and amino acids). Thus, the cellular machinery will "parse" the sequence above as:

AUG- ACC-UGA-AGG-AUC-UUU-AAG

and assign each triplet to a specific amino acid:

ALA-VAL-VAL-TYR-GLU-TYR-etc....

The mapping from base triplets to protein amino acids is referred to as the *genetic code* and is universal for all life-forms on earth (from bacteria to humans). A table with the genetic code is included with this handout (Figure 1: The codon dictionary).

<u>Protein</u>

Proteins are similar to DNA and RNA because they are also a linear sequence of alphabet letters. Unlike RNA and DNA, however, proteins do not have a 4 letter alphabet, they have a 20 letter alphabet. The letters for RNA and DNA are called bases. The letters for proteins are called amino acids. The twenty types of amino acids that make up protein structures have two components: a common backbone component which is hooked up with other backbone components to form a string of amino acids, and a sidechain component which is different for each of the twenty amino acids. These sidechain components have names such as Alanine, Valine, Tryptophan, Tyrosine, etc... which are abbreviated ALA, VAL, TRP, TYR, etc...(there is also a harder to remember one-letter abbreviation for each amino acid). A protein can therefore be drawn schematically like this:

```
 BB--BB--BB--BB--BB--BB...
 |   |   |   |   |   |
Ala Val Val Tyr Glu Tyr...
```

Since the backbone part is common to all amino acids, the sequence is abbreviated just by the sidechain identity:

```
Ala-Val-Val-Tyr-Glu-Tyr...
```

The key insight to how three-dimensional structure is derived from 1-dimensional information is contained in the amino acid sidechains. The sidechains have very different chemical properties (such as charge, size, tendency to interact with water molecules). The protein string folds into a three-dimensional shape based on the chemical properties of the sidechains. For example, some amino acids are positively charged and others are negatively charged. These amino acids will tend to attract. Similarly, amino acids which tend to interact with water molecules will tend to be on the surface of the molecule, those which repel water (similar to how oil and water repel) will tend to associate with each other on the inside of the protein molecule.   Quite surprisingly, a given sequence of amino acids will fold into the same three-dimensional structure every time.   That is, the chemical properties contained within the amino acids are specific enough that they require a certain unique structure in order to match up optimally. Thus, one dimensional information as contained in DNA and translated into protein sequence implies a single unique 3-dimensional structure. The table showing (Figure 2) "Amino Acid Hydrophilic Values" provides information about the one type of physical property that distinguish amino acids. Hydrophilicity is a measure of how much an amino acid tends to associate with water molecules. The ones that are not hydrophilic (such as ILE, TRP, LEU) will tend to associate with one another and exclude water from their contacts. The ones that are hydrophilic (ASP, GLU) will tend to associate freely on water. This will in turn, imply that the hydrophilic residues will be on the surface of a protein, while the hydrophobic residues will be on the interior. These types of interactions, in addition to others such as positive and negative charge, act to drive the unique folding of a protein sequence into a protein 3-dimensional structure.

The fourth dimension of information is specified with the part of the DNA that does not directly encode protein sequence information. Recall our initial sequence:

```
   V                             V
TCGATGACCTGAAGGATCTTTAACTGATCTCTGATATCTGTATCGAAATT
AGCTACTGGACTTCCTAGAAATTGACTAGAGACTATAGACATAGCTTTAA
```

The region of the top strand that starts with ATG and ends with TGA codes for a protein amino acid sequence, as illustrated above. The remainder of the sequence contains *control information* which allows the cellular machinery to encode information about timing.   For example, the region to the right of the TGA may have a sequence that is recognized by a particular protein. Upon recognizing this sequence, the protein may bind to this region, which may in turn block the successful translation of the protein coding region (starting with ATG and ending with TGA) just to the left.   With time, the concentration of this blocking protein (often called a repressor, because it is repressing the production of the protein encoded by this sequence) may decrease and the protein may "fall off" its seat on the DNA.   This would allow production of the gene encoded here. The encoded gene may then bind to another piece of DNA and repress its gene product for a while. In this way, we can imagine a cascade of proteins that are produced, block a piece of DNA for a while, and then fall off. This cascade would have a reproducible chronological sequence of events, and could be used to control which proteins are being made at any given time.

Thus, in the development of an embryo, there is a protein for the early sequence of events that blocks the production of machinery needed for later events. When the early sequence of events are finished, this protein may be degraded, destroyed or otherwise "fall off" the DNA, and allow the proteins associated with later events to be produced. This process can proceed indefinitely to allow for complex sequences of events required for constructing a complicated organism. Thus, the 1-dimensional information contained in the DNA is able to code for not only the three-dimensional information about the size and shape of an organism, but also can code for events that are separated by time.

Some Details of the Human Genome

Each human genome contains 23 chromosomes. There are 22 chromosomes common to both sexes. Women have two copies of the X (one is inactivated randomly early in development). Men have an X chromosome and a Y chromosome (the Y chromosome contains the information necessary to be a male, e.g. increased synthesis and responsiveness to testosterone). The XX of women or XY of men, count as a single chromosome for the purposes of calculating the total of 23, but are actually different so that men contain 24 chromosomes in reality.

The size of chromosomes ranges from 60 megabases (MB) to 260 MB. There is a total of about 3000 MB, or 3 billion bases in each human genome. The 23 chromosomes are thought to encode for 100,000 different proteins of average length 1000 amino acids (see discussion of amino acids and bases below). Proteins are the molecules that take on the three-dimensional structure that gives each organism form, and are also responsible for many of the chemical processes that contribute to life: muscle contraction involves the concerted contraction of ensembles of proteins, vision involves the response of proteins to photons as they land on the surface of the eye, digestion involves the degradation of food products by proteins that have the enzymatic ability to break down complicated molecules, intelligence and memory is not yet understood but involves the interactions brain cells which are constructed with and by proteins.

Physical and genetic map construction. The chromosomes that are currently being sequenced are too long to sequence all at once. In order to make their size manageable, most sequencing technologies require that the long sequences be cut up into smaller pieces and then sequenced. This leaves the problem of reconstructing the proper order of the sequences. Overlapping fragments are identified and used to reconstruct the order of fragments that have been cut randomly. This task is made non-trivial by the fact that there are errors in the sequence. There are therefore many efforts afoot at intelligent editors that will try to combine fragments, but allow human intervention for difficult cases.

If the sequences are not cut randomly, then they can be cut by restriction enymes. These enymes cut DNA at points where specific 4-10 base sequences occur. Usually, in order to get fragments that are small enough, a DNA molecule will be exposed to a series of restriction enzymes (as well as subsets of enzymes), producing a set of fragments for sequencing. As can be imagined, there is an interesting (and difficult) post-processing task of reconstructing the connectivity of the fragments by looking for possible pairings of fragments, and then searching for fragments from other series of restriction cuts that contain the overlapping region. For example, if you have the sequence:

```
AGCTACTGGACTTCCTAGAAATTGACTAGAGACTATAGATTCATAGCTTTAA
```

and you have two restriction enzymes that cut at all (#1) TTC sites and all (#2) TTG sites, then you would have the following fragments:

(from #1 alone)
```
AGCTACTGGACTTC
CTAGAAATTGACTAGAGACTATAGATTC
ATAGCTTTAA
```

(or from #2 alone)
AGCTACTGGACTTCCTAGAAATTG
ACTAGAGACTATAGATTCATAGCTTTAA

If you have the result of both enzymes in series, then you have:

AGCTACTGGACTTC
CTAGAAATTG
ACTAGAGACTATAGATTC
ATAGCTTTAA

The results of any single of these experiments is not sufficient to reconstruct the original sequence, but all three together provide enough information. When you add the issue of errorsin the sequencing, you can see that we have a combinatoric, noisy, constraint satisfaction problem.

The problems of reconstruction the order of linear segments comes up in other areas within molecular biology. For example, genetic information often comes in the form of "linkage" information which provides an effective distance between two genes. Given a large set of such distances, there is a requirement for algorithms that can construct linear sequences of these genetic markers that are compatible with the genetic "distance" information.

Further Possibilities for Reading

Lubert Stryer's textbook entitled "Biochemstry" discusses the basics of protein structure and DNA sequence. 4th ed. New York, N.Y. : W.H. Freeman, 1995.

The course web page will also have pointers to some primers.