Introduction to Bioinformatics (Fall 2003)
Flow of Information from DNA to Protein

Outline:

A. Overview

B. What is DNA?

C. Transcription

D. RNA processing

E. Translation

A. Overview

The information contained within DNA determines the potentialities of a cell. The information within protein present within a cell determines the actuality of a cell. Between the two are enough details to fill several thick textbooks (see Figure 1 for a crude view of the whole). Here we'll confine ourselves to the slice of knowledge that is of greatest interest to the bioinformatician: how the information changes form in going from DNA to protein and what signals regulate the intervening processes.

B. What is DNA?

Just as the properties of proteins are determined by their structure, so it is with DNA. Everyone has heard about DNA, the double helix. What made the structure compelling when it was first conceived 50 years ago was the explanation it provided for the replication of the genetic material (Figure 2). One strand pairs with the other through bonding between nucleotides: adenine (A) with thymine (T) and guanine (G) with cytosine (C). Since the nucleotides on one strand

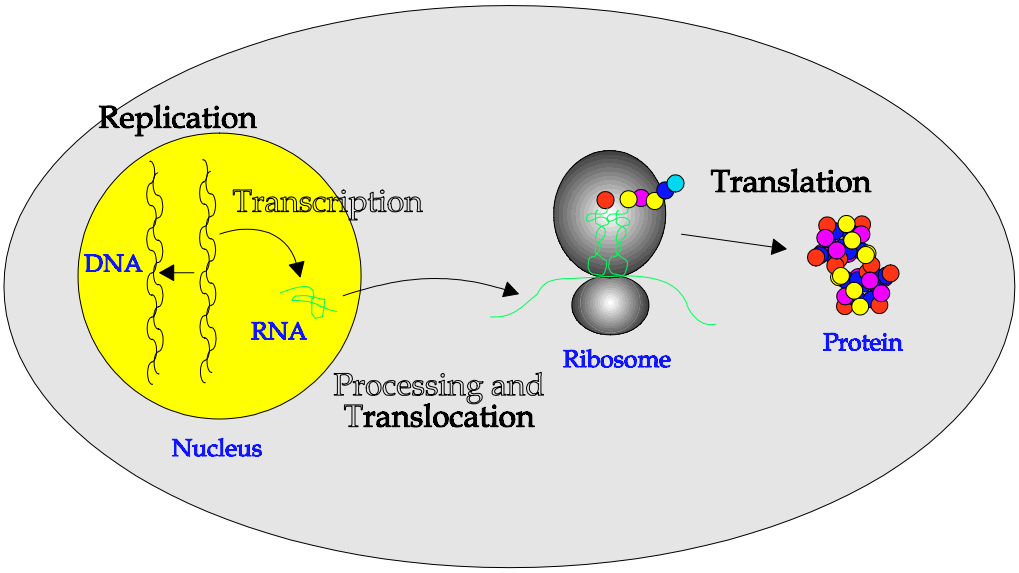

Figure 1: Flow of information from DNA to protein.

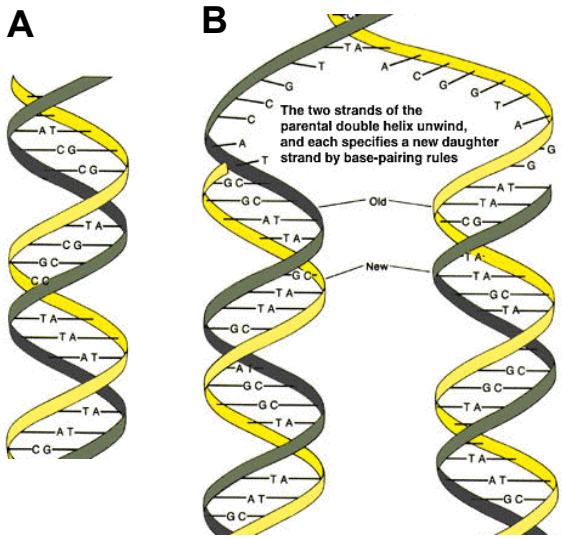

Figure 2: Cartoon of DNA double helix. (**A**) Before replication. (**B**) During replication.

completely determine the nucleotides on the other, the two strands contain the same information. If one strand were destroyed, you'd be able to recreate it from the information contained in the other strand. DNA replicates by separating the two strands and replicating the partner of each.

SQ1. Why is it that some strands in Figure 1B are gray and others yellow? (For example, are the gray strands all old?)

SQ2. If one strand of DNA had the sequence GGACT, what would be the sequence of the second strand? Draw the double helix.

The view of DNA depicted in Figure 2 is too simple in a number of ways. For our purposes, the most important defect is that the picture seems to imply that you could rotate a strands 180° and it would work just as well. This turns out not to be true, as illustrated in the somewhat more realistic cartoon in Figure 3. DNA strands have directionality, and the double helical structure is possible only when the two strands are arranged antiparallel to one another. This fact has profound implications for the replication of DNA, which will not concern us very much, and is essential to appreciate when using the DNA polymerase chain reaction (PCR), which probably WILL concern us sometime this semester.

A DNA sequence is meaningless unless you know the direction of the strand. Direction is indicated by the markers "5'" and "3'". A strand with the sequence 5'-AGGCTA-3' has a complementary strand of 3'-TCCGAT-5' or, equivalently, 5'-TAGCCT-3' (it is understood that strands pair in antiparallel orientation). A sequence that is not given these markers is presumed to have been written 5' to 3', left to right.

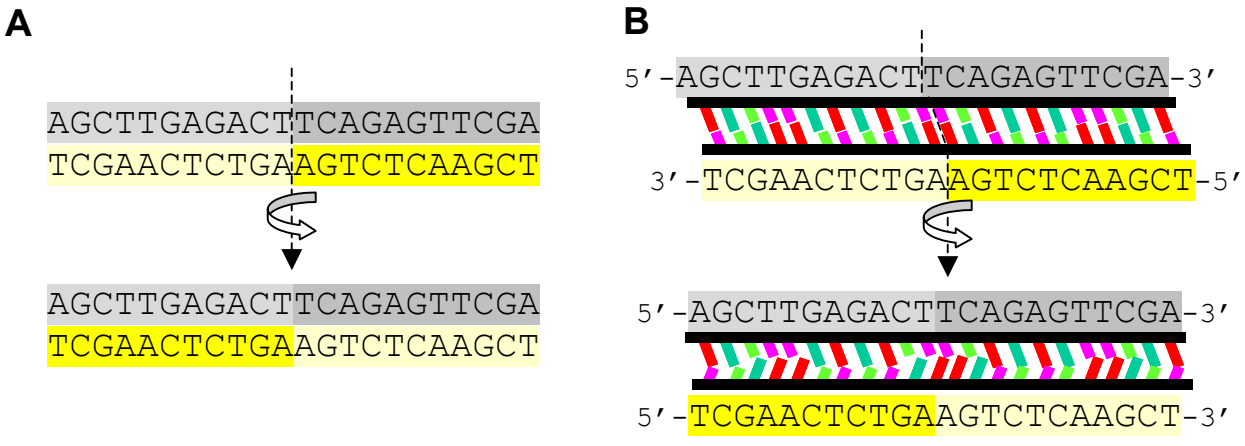

Figure 3: Directionality of DNA. Two views of a DNA sequence that appears to be palindromic (reads the same forwards and backwards). (**A**) False picture of reality, implying that the bottom strand can be rotated 180° and still base pair with the top strand. (**B**) More realistic view, illustrating that rotating the bottom strand 180° prevents the nucleotides from aligning well enough to base pair. The bottom structure thus cannot occur.

SQ3. Which of the following represent two DNA sequences that will pair with each other?

a. **5'-GGAGTT-3' and 5'-CCTCAA-3'**
b. **5'-GGAGTT-3' and 3'-CCTCAA-5'**
c. **5'-GGAGTT-3' and 5'-AACTCC-3'**
d. **5'-GGATCC-3' and 5'-GGATCC-3'**

By the way, DNA sequences that can basepair with themselves are called "palindromes". Such sequences may seem like artificial curiosities, but in fact palindromes are frequent targets of DNA-binding proteins, including proteins that regulate transcription. These sequences are therefore of great importance in the bioinformatic analysis of DNA.

SQ4. Make up an example of a palindromic DNA sequence.

C. Transcription

Transcription is the rewriting of information from DNA format to RNA format. RNA differs from DNA in two regards. First, the backbone of RNA is composed of phosphoribose (hence ***R***iboNucleic Acid) instead of phosphodeoxyribose (hence ***D***eoxyriboNucleic Acid). We don't have to concern ourselves with this difference except to understand that proteins can tell the difference between ribose and deoxyribose. Second, RNA generally contains uracil (U) in place of thymine (T). Uracil and thymine pair equally well with adenine (A). These are not very dramatic differences, and the information in DNA format is readily transformed into RNA format. For example, a DNA sequence of 5'-AGTTCA-3' may be transcribed into the RNA sequence 5'-AGUUCA-3' (note that RNA has directionality as well).

Not all DNA is transcribed. Only a small fraction of a DNA chromosome is transcribed into RNA at any one time because only a fraction of the proteins encoded by the DNA are needed at any given moment (and, in most eukaryotes, because most of the DNA does not encode proteins and is never transcribed at all). Transcription begins at the ***promoter***, the binding site for RNA polymerase. It is the binding of RNA polymerase that determines whether a region is or is not transcribed, and there is often a highly complicated decision making process to determine whether RNA polymerase will or will not bind near a specific gene.

There are four classes of RNA (Table 1). You'll be concerned primarily with messenger RNA (mRNA), the class that is translated to protein. However, about 97% of the RNA in a cell consists of ribosomal RNA (rRNA) and transfer RNA (tRNA), which is not translated but rather takes part in the translation apparatus. It may seem strange that so much of the cell's RNA is devoted to the translation machinery and so little of it to that which is translated, but so it is in macroscopic machines. Consider how little of the weight of an active tape recorder is the tape itself!

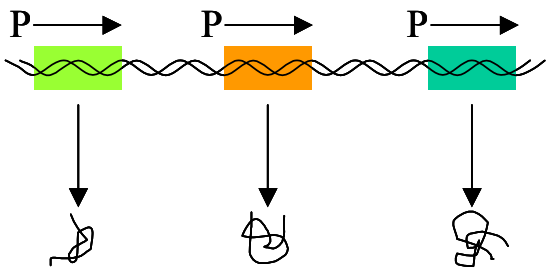

Figure 4: Selective transcription of DNA. Each region of DNA is transcribed separately into RNA, each molecule of which is very small compared to the size of the chromosome. **P** represents the promoter, i.e. the binding site on the DNA for RNA polymerase, the enzyme that catalyzes transcription.

Table 1: Comparison of classes of RNA in a typical bacterium (*E. coli*)

Class	Types	Fraction	Sizes	Stability	Function
Ribosomal RNA (rRNA)	3	80%	2904 bp, 1542 bp, 120 bp	Stable	Scaffold for ribosomal proteins. Participates in translation.
Transfer RNA (tRNA)	37	12%	76 bp to 91 bp	Stable	Connects codon to amino acid in translation.
Messenger RNA (mRNA)	1000's	5%	~300 bp to ~10,000 bp	Unstable	Carries information on amino acid sequence of specific protein.
Other RNA	???	???	???	???	Other structural RNA; regulatory RNA

SQ5. Go further with the tape recorder analogy, identifying what is analogous to each component of the translation process.

SQ6. Give three differences that generally distinguish RNA from DNA (think sugar content, base composition, and usual structure)

D. RNA processing

Prokaryotes lack nuclei, and the mRNA that is transcribed is immediately ready for translation. In fact, the leading end of prokaryotic mRNA is often translated even ***while*** the hind end is still being transcribed! The situation is quite different in eukaryotes. There, transcription takes place in the nucleus, while translation is confined to the cytoplasm (see Fig. 1). In between, the RNA transcript may be extensively processed (Figure 5) before it is translocated into the cytoplasm to serve as mRNA.

Eukaryotic RNA that will eventually be translated is capped on the 5' end with a modified guanosine. This serves as a handle recognized by ribosomes to initiate translation. The other end of the transcript is modified with the addition of hundreds of adenosine residues, which helps governs the lifetime of the RNA in the cytoplasm. Finally, the RNA is spliced, removing portions of the RNA that is not part of the gene. A region of the transcript that is excised is called an intron, while a region that is retained is called an exon. The presence of introns in eukaryotic DNA greatly complicates automated efforts to recognize genes, because the information related to a single protein is disconnected.

SQ7. Why are introns problematic for auto-mated gene-finders? (This may be easier to answer after the next section)

SQ8. So often, Nature chooses a seemingly chaotic strategy quite different from what we would have done if we were running the show. No one anticipated the existence of introns. Why do you think they exist?

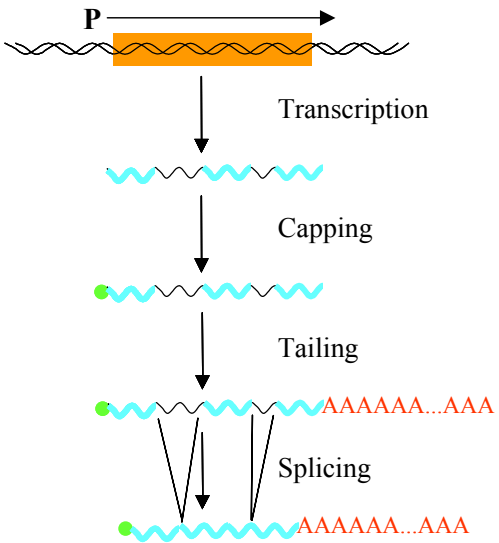

Figure 5: Processing of RNA transcript. Thick blue-green lines represent exons, thin black lines introns. Green balls represent m^7G-caps. The polyadenosine tail is typically hundreds of nucleotides in length.

E. Translation

There are three bioinformatic problems in translating mRNA to protein: where to initiate the process, how to translate one type of information into another, and where to end. The mechanism of initiation differs between prokaryotes and eukaryotes, but the rest of the process is common to all of life. Understanding the basis of translation, the genetic code, marks one of the most spectacular advances in the history of science. We don't have time, unfortunately, to pause to appreciate the manner in which this advance was achieved. Instead, we will focus on how to ***use*** the genetic code. After that, we'll return to the question of initiation.

The genetic code is the solution that Nature found to transforming information spelled in a four-letter code (A, C, G, T) into information spelled in a twenty-letter code (the twenty amino acids).The problem is analogous to transforming numbers spelled in base two into numbers spelled in base sixteen. There the solution is straightforward:

Translation of number in base 2 to number in base 16

0000 → 0	0100 → 4	1000 → 8	1100 → C
0001 → 1	0101 → 5	1001 → 9	1101 → D
0010 → 2	0110 → 6	1010 → A	1110 → E
0011 → 3	0111 → 7	1011 → B	1111 → F

There are sixteen possible 4-digit numbers in base two and sixteen possible 1-digit numbers in base sixteen – all very neat.

The translation from RNA to protein is more problematic, because the number of letters of one code is not a power of the number of letters of the other.

SQ9. How many amino acids can be determined from a code consisting of one nucleotide (either A, C, G, or T)? How about a code of two nucleotides? Three nucleotides?

As it happens, Nature chose a triplet code. This gives 64 possible triplets, whose amino acid assignments constitute the genetic code. You can see this code online by going to the Links section in the course web page. It is essential that you know how to read this table. If you want to find the amino acid encoded by the triplet codon AGC, look at the left side for A, the top for G, and that gets you to the box where you can scan for AGC. You should find the amino acid to be serine (ser).

Note that there are three special codons for which there are no amino acids given. UAG, UAA, and UGA are the STOP codons, marking the end of an encoded polypeptide chain. There is another special codon, AUG. Most but not all genes begin with AUG. It is sometimes called the START codon, though there are genes that start with GUG or UUG. Besides marking the beginning of the protein, AUG is also used to encode the amino acid methionine. In this respect, it is just a normal codon. AUG may be found not only at the beginning but also in the middle of genes. How ribosomes distinguish between these two situations will be taken up momentarily.

SQ10. Starting from the first conventional start codon, translate the RNA strand given below:

GAAGCAUGUCCGAGCAAUGAGCCGA

Important features of the code

1. The genetic code is degenerate: There are 64 possible triplet codons but only 20 possible codons. Nonetheless, 61 of the 64 codons are assigned amino acids. Most amino acids are encoded by more than one codon. This many-to-one relationship is what defines "degeneracy".

SQ11. The three amino acids most commonly found in human protein are leucine, glycine, and serine. The three amino acids least commonly found in human protein are tryptophan, methionine, and histidine. Draw a conclusion about how degeneracy relates to the natural frequencies of amino acids.

2. Not all amino acid changes are possible from a single basepair mutation: Virtually all mutations found in nature are single events: single basepair changes or single insertions or deletions. This fact places a strong limitation on what amino acid changes are observed. For example, leucine, encoded by CUA, can mutate in one of three positions: C in the first position to U, A, or G; U in the second position to C, A, or G; and in the third position to U, C, or G. In principle, then, a one-base change in the codon can lead to any one of nine other codons and any one of nine amino acids. Immediately, one sees that it is impossible with a single base change to get to triplets encoding all 20 amino acids.

Actually, the situation is even more extreme. Of the nine codons listed above, five of them also encode leucine. Changes to one of these are called silent mutations, mutations that do not affect the amino acid sequence of the protein. Two other of the nine codons encode isoleucine or valine, amino acids very similar to leucine. These are called conservative changes. Changes to these amino acids very well may not affect the structure of the protein. Only two of the nine possible changes lead to a hydrophobic-to- hydrophilic amino acid change. The failure of many mutations to produce functional changes in protein is no accident. The genetic code appears to be built with that in mind.

SQ12. List the changes that can be produced by a single basepair mutation in the AGA codon encoding arginine and label each silent, conservative, hydrophobic-to-hydrophilic, hydrophilic-to-hydrophobic, or other.

Initiation of translation

The AUG codon serves two purposes: (1) it encodes the amino acid methionine, and (2) in some cases it marks the beginning of the protein. How do ribosomes tell which action to take? In eukaryotes, the solution is relatively simple. Ribosomes bind to the m^7G caps put on mRNA and slide over until they encounter the first AUG. ***Most*** of the time, this AUG serves to initiate translation. There are obviously other cues, because some messages reproducibly begin at later AUG triplets.

Prokaryotes solve the problem in a very different fashion. Most start codons are preceded by a special sequence, variations on AGGAG, to which ribosomes bind. If there is an AUG triplet close by, translation is initiated. Here too, the signals perceived by ribosomes are more complex than this description implies, and we have only limited ability to predict initiation sites from the sequence of mRNA.