Lecture 2

Basics of Molecular Biology (continued)

January 6, 2000
Notes: Tory McGrath

2.1. Course Projects

A typical course project might be to take some existing biological sequences from the public databases on the web, and design and run some sequence analysis experiments, using either publicly available software or your own program. For example, there is reason to believe that some of the existing bacterial genomes may be misannotated, in the sense that the identified genes are not actually located exactly as annotated. There is existing software to identify gene locations. We will discuss more such suggested projects as the course proceeds, but the choice of topic is quite flexible, and is open to suggestion, provided there is a large computational aspect. You will be required to check your project topic with the instructor before embarking.

You may work on the project in groups of up to four people. For maximum effectiveness, it is recommended to have a mix of biology and math/computer participants in each group.

The project will entail a short write-up as well as a short presentation of your problem, methods, and results.

2.2. Translation (continued)

In prokaryotes, which have no cell nucleus, translation begins while transcription is still in progress, the $5'$ end of the transcript being translated before the RNA polymerase has transcribed the $3'$ end. (See Drlica [1, Figure 4-4].) In eukaryotes, the DNA is inside the nucleus, whereas the ribosomes are in the *cytoplasm* outside the nucleus. Hence, transcription takes place in the nucleus, the completed transcript is exported from the nucleus, and translation then takes place in the cytoplasm.

The ribosome forms a complex near the $5'$ end of the mRNA, binding around the *start codon*, also called the *translation start site*. The start codon is most often $5'$-AUG-$3'$, and the corresponding anticodon is $5'$-CAU-$3'$. (Less often, the start codon is $5'$-GUG-$3'$or $5'$-UUG-$3'$.) The ribosome now brings together this start codon on the mRNA and its exposed anticodon on the corresponding tRNA, which hybridize to each other. (See [1, Figure 4-4].) The tRNA brings with it the encoded amino acid; in the case of the usual start codon $5'$-AUG-$3'$, this is methionine.

Having incorporated the first amino acid of the synthesized protein, the ribosome shifts the mRNA three bases to the next codon. A second tRNA complexed with its specific amino acid hybridizes to the second codon via its anticodon, and the ribosome bonds this second amino acid to the first. At this point the ribosome releases the first tRNA, moves on to the third codon, and repeats. (See [1, Figure 4-5].) This

process continues until the ribosome detects one of the STOP codons, at which point it releases the mRNA and the completed protein.

2.3. Prokaryotic Gene Structure

Recall from Section 1.6 that a gene is a relatively short sequence of DNA that encodes a protein or RNA molecule. In this section we restrict our attention to protein-coding genes in prokaryotes.

The portion of the gene containing the codons that ultimately will be translated into the protein is called the *coding region*, or *open reading frame*. The transcription start site (see Section 1.6.1) is somewhat *upstream* from the start codon, where "upstream" means "in the $5'$ direction". Similarly, the transcription stop site is somewhat *downstream* from the stop codon, where "downstream" means "in the $3'$ direction". That is, the mRNA transcript contains sequence at both its ends that has been transcribed, but will not be translated. The sequence between the transcription start site and the start codon is called the *$5'$ untranslated region*. The sequence between the stop codon and the transcription stop site is called the *$3'$ untranslated region*.

Upstream from the transcription start site is a relatively short sequence of DNA called the *regulatory region*. It contains *promoters*, which are specific DNA sites where certain regulatory proteins bind and regulate expression of the gene. These proteins are called *transcription factors*, since they regulate the transcription process. A common way in which transcription factors regulate expression is to bind to the DNA at a promoter and from there affect the ability (either positively or negatively) of RNA polymerase to perform its task of transcription. (There is also the analogous possibility of *translational regulation*, in which regulatory factors bind to the mRNA and affect the ability of the ribosome to perform its task of translation.)

2.4. Prokaryotic Genome Organization

The *genome* of an organism is the entire complement of DNA in any of its cells. In prokaryotes, the genome typically consists of a single chromosome of double-stranded DNA, and it is often circularized (its $5'$ and $3'$ ends attached) as opposed to being linear. A typical prokaryotic genome size would be in the millions of base pairs.

Typically 90% of the prokaryotic genome consists of coding regions. For instance, the *E. coli* genome has size about 5 Mb and approximately 4300 coding regions, each of average length around 1000 bp. The genes are relatively densely and uniformly distributed throughout the genome.

2.5. Eukaryotic Gene Structure

An important difference between prokaryotic and eukaryotic genes is that the latter may contain "introns". In more detail, the transcribed sequence of a general eukaryotic gene is an alternation between DNA sequences called *exons* and *introns*, where the introns are sequences that ultimately will be spliced out of the mRNA before it leaves the nucleus. Transcription in the nucleus produces an RNA molecule called *pre-mRNA*, produced as described in Section 1.6.1, that contains both the exons and introns. The introns are spliced out of the pre-mRNA by structures called *spliceosomes* to produce the *mature mRNA* that will be transported out of the nucleus for translation. A eukaryotic gene may contain numerous introns, and each intron may

be many kilobases in size. One fact that is relevant to our later computational studies is that the presence of introns makes it much more difficult to identify the locations of genes computationally, given the genome sequence.

Another important difference between prokaryotic and higher eukaryotic genes is that, in the latter, there can be multiple regulatory regions that can be quite far from the coding region, can be either upstream or downstream from it, and can even be in the introns.

2.6. Eukaryotic Genome Organization

Unlike prokaryotic genomes, many eukaryotic genomes consist of multiple linear chromosomes as opposed to single circular chromosomes. Depending on how simple the eukaryote is, very little of the genome may be coding sequence. In humans, less than 3% of the genome is believed to be coding sequence, and the genes are distributed quite nonuniformly over the genome.

2.7. Goals and Status of Genome Projects

Molecular biology has the following two broad goals:

1. Identify all key molecules of a given organism, particularly the proteins, since they are responsible for the chemical reactions of the cells.

2. Identify all key interactions among molecules.

Traditionally, molecular biologists have tackled these two goals simultaneously in selected small systems within selected model organisms. The genome projects today differ by focusing primarily on the first goal, but for *all* the systems of a given model organism. They do this by *sequencing* the genome, which means determining the entire DNA sequence of the organism. They then perform a computational analysis (to be discussed in later lectures) on the genome sequence to identify (most of) the genes. Having done this, (most of) the proteins of the organism will have been identified.

With recent advances in sequencing technology, the genome projects have progressed very rapidly over the past five years. The first free-living organism to be completely sequenced was the bacterium *H. influenzae* [2], with a genome of size 1.8 Mb. Since that time, 18 bacterial, 6 archaeal, and 2 eukaryotic genomes have been sequenced. Presently there are approximately an additional 95 prokaryotic and 27 eukaryotic genomes in the process of being sequenced. (See, for example, the Genomes On Line Database at `http://geta.life.uiuc.edu/~nikos/genomes.html` for the status of ongoing genome projects.)

The human genome is expected to be sequenced within the next two years or so. Although every human is a unique individual, the genome sequences of any two humans are about 99.9% identical, so that it makes some sense to talk about sequencing *the* human genome, which will really be an amalgamation of a small collection of individuals. Once that is done, one of the interesting challenges is to identify the common *polymorphisms*, which are genomic variations that occur in a nonnegligible fraction of the population.

2.8. Sequence Analysis

Once a genome is completely sequenced, what sorts of analyses are performed on it? Some of the goals of *sequence analysis* are the following:

1. Identify the genes.
2. Determine the function of each gene. One way to hypothesize the function is to find another gene (possibly from another organism) whose function is known and to which the new gene has high sequence similarity. This assumes that sequence similarity implies functional similarity, which may or may not be true.
3. Identify the proteins involved in the regulation of gene expression.
4. Identify sequence repeats.
5. Identify other functional regions, for example *origins of replication* (sites at which DNA polymerase binds and begins replication; see Section 1.5), *pseudogenes* (sequences that look like genes but are not expressed), sequences responsible for the compact folding of DNA, and sequences responsible for nuclear anchoring of the DNA.

Many of these tasks are computational in nature. Given the incredible rate at which sequence data is being produced, the integration of computer science, mathematics, and biology will be integral to analyzing those sequences.

References

[1] Karl Drlica. *Understanding DNA and Gene Cloning*. John Wiley & Sons, second edition, 1992.

[2] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae rd. Science*, 269:496–512, July 1995.