Transcription

The flow of genetic information

The primary role of DNA is to store heritable information that encodes the instructions for creating an organism. In the past decade we have gotten to be very good at sequencing DNA, but we still don't know how to reliably decode ALL of the information, and we still don't understand ALL of the mechanisms by which it is expressed.

There are however some core principles and mechanisms associated with the reading and expression of the genetic code whose basic steps are understood and that need to be part of the conceptual toolkit for all biologists. Two of these processes are transcription and translation; the copying of parts of the genetic code written in DNA into molecules of the related polymer RNA, followed by and the reading and decoding of the RNA sequence (of nucleotides) into the proteins's sequence (of amino acids).

In BIS2A we focus largely on developing an understanding of the process of transcription (recall that an Energy Story is simply a way of describing a process) and its role in the expression of genetic information. We motivate our discussion of transcription by focusing on functional problems (bringing in parts of our problem solving/design challenge) that must be solved for the process to take place. We then go on to describe how the process is used by Nature to create a variety of functional RNA molecules (that may have various structural, catalytic or regulatory roles) including messenger RNA (mRNA) molecules that carry the information required to synthesize proteins. Likewise, we focus on challenges and questions associated with the process of translation, the process by which the ribosomes synthesize proteins.

The basic flow of genetic information in biological systems is often depicted in a scheme known as "the central dogma" (see figure below). This scheme states that information encoded in DNA flows into RNA via transcription and ultimately to proteins via translation. Processes like reverse transcription (the creation of DNA from and RNA template) and replication
also represent mechanisms for propagating information in different forms. This scheme, however, doesn't say anything about how information is encoded or about the mechanisms by which regulatory signals are move between the various layers of molecule types depicted in the model. Therefore, while the scheme below is a nearly required part of the lexicon of any biologist, perhaps left over from old tradition, students should also be aware that mechanisms of information flow are more complex (we'll learn about some as we go, and that "the central dogma" only represents some core pathways.

![Diagram of the flow of genetic information]

The flow of genetic information.

Attribution: Marc T. Facciotti (original work)
Genotype to Phenotype

An important concept in the following sections is the relationship between genetic information, the **genotype**, and the result of expressing it, the **phenotype**. These two terms and the mechanisms that link the two will be discussed repeatedly over the next few weeks - start becoming proficient with using this vocabulary.

The information stored in DNA is in the sequence of the individual nucleotides when read from 5' to 3' direction. Conversion of the information from DNA into RNA (a process called transcription) produces the second form that information takes in the cell. The mRNA is used as the template for the creation of the amino acid sequence of proteins (in translation). Here two different sets of information are shown. The DNA sequence is slightly different, resulting in two different mRNAs produced, followed by two different proteins, and ultimately, two different coat colors for the mice.

**Genotype** refers to the information stored in the DNA of the organism, the sequence of the nucleotides, the compilation of its genes. **Phenotype** refers to any physical characteristic that you can measure, such as height, weight, amount of ATP produced, ability to metabolize lactose, response to environmental stimuli, etc. Differences in genotype, even slight, can lead to different phenotypes that are subject to natural selection. The figure above depicts this idea. Note also, that while classic discussions of the genotype and phenotype relationships are talked about in the context of multicellular organisms, this nomenclature and the underlying concepts apply to all organisms, even single celled organisms like bacteria and archaea.

Suggested discussion

Can something you can not see "by eye" be considered a phenotype?

Suggested discussion

Can single-celled organisms have multiple simultaneous phenotypes? If so, can you propose an example? If not, why?
Genes

What is a gene? A gene is a segment of DNA in an organism's genome that encodes a functional RNA (such as rRNA or tRNA, etc) or protein product (enzymes, tubulin, etc). A generic gene contains elements encoding regulatory regions (which are often not transcribed) and a region encoding a transcribed unit.

Genes can acquire mutations - defined as changes in the in the composition and or sequence of the nucleotides - either in the coding or regulatory regions. These mutations can lead to several possible outcomes: (1) nothing measurable happens as a result; (2) the gene is no longer expressed; or (3) the expression or behavior of the gene product(s) are different. In a population of organisms sharing the same gene, different variants of the gene are known as alleles. Different alleles can lead to differences in phenotypes of individuals and contribute to the diversity in biology that is under selective pressure.

Start learning these vocabulary terms and associated concepts. You will then be somewhat familiar with them when we start diving into them in more detail over the next lectures.
A gene consists of a coding region for an RNA or protein product accompanied by its regulatory regions. The coding region is transcribed into RNA which is then translated into protein. Note that most transcripts to not begin with a start codon and end with a stop codon (unlike this example), there are usually...
Section Summary

All living things must all transcribe genes from their genomes. While the cellular location may be different (eukaryotes perform transcription in the nucleus; bacteria and archaea- lacking a nucleus- perform transcription in the cytoplasm), the mechanisms by which organisms from each of these clades carry out this process are fundamentally the same and can be characterized by three stages: initiation, elongation, and termination.

Transcription: from DNA to RNA

A short overview of transcription

Transcription is the process of creating an RNA copy of a segment of DNA. Since this is a process, we want to apply the Energy Story to develop a functional understanding of transcription. What does the system of molecules look like before the start of the transcription? What does it look like at the end? What transformations of matter and transfers of energy happen during the transcription and what, if anything, catalyzes the process? We also want to think about the process from a Design Challenge standpoint. If the biological task is to create a copy of DNA in the chemical language of RNA, what challenges can we reasonably hypothesize, or anticipate given our knowledge about other nucleotide polymer processes, must be overcome? Is there evidence that Nature solved these problems in different ways? What seem to be the criteria for success of transcription? You get the idea.

Listing some of the basic requirements for transcription

Let us first consider the tasks at hand by using some of our foundational knowledge and imagining what might need to happen during the process of transcription if the goal is to make an RNA copy of a piece of one strand of a double stranded DNA molecule. We'll see that using some basic logic allows us to infer many of the important questions and things that we need to know about to properly describe the process.

Let's imagine that we want to design a nanomachine/nanobot that would conduct transcription. We can use some Design Challenge thinking to identify problems and subproblems that need to be solved by our little robot.

• The first thing that we might want our machine to know is where to start. Along the millions to billions of base pairs, where should the machine be directed?
• Likewise, we need to know where to stop.
• If we have start and stop sites, we will need ways of encoding that information so that our machine(s) can read this information - how will that be accomplished?
• How many RNA copies of the DNA will we need to make?
• How fast do the RNA copies need to be made?
• How accurately do the copies need to be made?
• How much energy will the process take and where is the energy going to come from?

These are, of course, only some of the core questions. One can dig deeper if they wish. However, they are already good enough for us to start getting a good feel for this process. Notice, too, that many of these questions are remarkably similar to those we inferred might be necessary to understand about DNA replication.

### The Building Blocks of Transcription

#### The building blocks of RNA

Recall from our discussion on the structure of nucleotides that the building blocks of RNA are very similar to those in DNA. In RNA the building blocks consists of nucleotide triphosphates that are composed of a ribose sugar, a nitrogenous base and three phosphate groups. The key differences between the building blocks of DNA and those of RNA is that RNA molecules are composed of nucleotides with ribose sugars (as opposed to deoxyribose sugars) and that instead of utilizing thymidine (the thymine containing nucleotide) RNA utilizes uridine (a uracil containing nucleotide). Note below that uracil and thymine are structurally very similar - the uracil is just lacking a methyl (CH$_3$) functional group compared to thymine.

![Nucleotide Diagram](https://bio.libretexts.org/Courses/University_of_California_Davis/BIS_2A%3A_Introductory_Biology_(Britt)/Readings/20_Lecture_20_reading)

**The basic chemical components of nucleotides.**

*Attribution: Marc T. Facciotti (original work)*

### Transcription Initiation

https://bio.libretexts.org/Courses/University_of_California_Davis/BIS_2A%3A_Introductory_Biology_(Britt)/Readings/20_Lecture

Updated: Sun, 18 Aug 2019 14:39:51 GMT

Powered by
Promoters

Proteins responsible for creating an RNA copy of a specific piece of DNA (transcription) must first be able to recognize the beginning of the element to be copied. A **promoter** is a DNA sequence onto which various proteins, collectively known as the transcription machinery, bind and initiates transcription. In most cases, promoters exist upstream (5’ to the coding region) of the genes they regulate. The specific sequence of a promoter is very important because it determines whether the corresponding coding portion of the gene is transcribed all the time, some of the time, or infrequently.

In the bacterium *E. coli*, at the -10 and -35 regions upstream of the initiation site (the site of the first nucleotide of the transcript), there are two promoter **consensus** sequences, or regions that are similar across many promoters and across various related species. Some promoters will have a sequence very similar to the consensus sequence (the sequence containing the most common sequence elements), and others will look very different. These sequence variations affect the strength to which the transcriptional machinery can bind to the promoter to initiate transcription. This helps to control the number of transcripts that are made and how often they get made.

![Genetic diagram](https://bio.libretexts.org/Courses/University_of_California_Davis/BIS_2A%3A_Introductory_Biology_(Britt)/Readings/20_Lecture_20_reading)

(a) A general diagram of a gene. The gene includes the promoter sequence, an untranslated region (UTR), and the coding sequence. (b) A list of several strong *E. coli* promoter sequences. The -35 box and -10 box are highly conserved sequences throughout the strong promoter list. Weaker promoters will have more base pair differences when compared to these sequences. Source: [http://www.discoveryandinnovation.co...lecture12.html](http://www.discoveryandinnovation.co...lecture12.html)

Suggested discussion

What types of interactions are changed between the transcription machinery and the DNA when the nucleotide sequence of the promoter changes? Why would some sequences create a “strong” promoter and why do others create a “weak” promoter?

**Bacterial vs Eukaryotic Promoters**

In bacterial cells, the -10 region is AT rich, often TATAAT. Sequences just upstream of the -10, s well as the -35 (TTGACA) region, are recognized and bound by the protein σ (“sigma factor”), which is one component of the RNA polymerase holoenzyme. Once this protein-DNA interaction is made, sigma factor facilitates unwinding of the -10 region and loads polymerase onto the template strand. Sigma factor thus assists the polymerase in recognizing promoter sequences and loading the polymerase onto the right spot, pointing in the correct direction. Interestingly, *E. coli*, makes several different sigma factors. In different situations- for example, under a particular stress- the activation of a different sigma factor will change the types of genes most frequently transcribed by RNA polymerase. Some bacteriophage have
taken advantage of this aspect of bacterial transcription, producing their phage-gene specific sigma factors that hijack the host's RNA polymerase and redirect it to the phage genome.

Eukaryotic promoters are much larger and more complex than prokaryotic promoters, but both have an AT-rich region - in eukaryotes it is typically called a "TATA box". For example, in the mouse thymidine kinase gene, the TATA box is located at approximately -30. For this gene, the exact TATA box sequence is TATAAAA, as read in the 5’ to 3’ direction on the non-template strand. This sequence is not identical to the E. coli -10 region, but both share the quality of being A–T rich elements.

Instead of a single bacterial polymerase, the genomes of most eukaryotes encode three different RNA polymerases, each made up of 10 protein subunits or more. Each eukaryotic polymerase also requires a distinct set of proteins known as transcription factors to recruit it to a promoter. The terminology for transcription factors is unfortunately rather inconsistent. Suffice it to say that there are many proteins that are always required to act simultaneously to load, for example, any RNA polymerase II (the polymerase that transcribes mRNAs). These are referred to as "basal" (or "general") transcription factors. In addition, an army of proteins may affect the frequency of the attraction of these basal factors to the promoters, the frequency of loading of RNA pol II, and even the "escape" of the pol II complex from eukaryotic promoters (so that it can actually perform transcription). Enhancers and silencers- both DNA sequences, not proteins- are recognized by these regulatory transcription factors. Basal transcription factors are crucial in the formation of a preinitiation complex on the DNA template that subsequently recruits RNA polymerase for transcription initiation.

Regardless of the details of bacterial vs. eukaryotic polymerase localization and orientation, initiation of transcription begins with the binding of RNA polymerase to the promoter. Transcription requires the DNA double helix to partially unwind such that one strand can be used as the template for RNA synthesis. Note that unwinding occurs within the polymerase; RNA polymerase, unlike DNA polymerase, has an intrinsic helicase activity. Double stranded DNA is sucked into the enzyme (as are NTPs), and double stranded DNA, plus RNA, emerges. The region of unwinding is called a transcription bubble.
During elongation, RNA polymerase tracks along the DNA template, synthesizes mRNA in the 5' to 3' direction, and unwinds then rewinds the DNA as it is read. The "nontemplate" strand illustrated here is often called the "coding stand", simply because its sequence will
match
the
sequence
of
the
transcript.

Elongation

Transcription always proceeds from one of the two DNA strands, which is called the **template strand**. The RNA product is complementary to the template strand and is almost identical to the non-template strand, called the **coding strand**, with the exception that RNA contains a uracil (U) in place of the thymine (T) found in DNA. During elongation, an enzyme called **RNA polymerase** proceeds along the DNA template adding nucleotides by base pairing with the DNA template in a manner similar to DNA replication, with the difference that an RNA strand is being synthesized that does not remain bound to the DNA template. As elongation proceeds, the DNA is continuously unwound ahead of the core enzyme and rewound behind it. Note that the direction of synthesis is identical to that of synthesis in DNA - 5' to 3'.

During elongation, RNA polymerase tracks along the DNA template, synthesizes mRNA in the 5' to 3' direction, and unwinds then rewinds the DNA as it is read.
A) 

The addition of nucleotides during the process of transcription is very similar to nucleotide addition in DNA replication. The RNA is polymerized from 5' to 3' and with each addition of a nucleotide, a phosphoanhydride bond is hydrolyzed by the enzyme resulting in a longer polymer and the release of two inorganic phosphates.
Suggested discussion

Compare and contrast the energy story for the addition of a nucleotide in DNA replication to the addition of a nucleotide in transcription.

Bacterial vs Eukaryotic Elongation

In bacteria, elongation begins with the release of the σ subunit of the RNA polymerase holoenzyme. The dissociation of σ allows the core enzyme to proceed along the DNA template, synthesizing mRNA in the 5’ to 3’ direction at a rate of approximately 40 nucleotides per second. As elongation proceeds, the DNA is continuously unwound ahead of the core enzyme and rewound behind it. The base pairing between DNA and RNA is not stable enough to maintain the stability of the mRNA synthesis components. Instead, the RNA polymerase acts as a stable linker between the DNA template and the nascent RNA strands to ensure that elongation is not interrupted prematurely.

In eukaryotes, following the formation of the preinitiation complex, the polymerase is released from the other transcription factors, and elongation is allowed to proceed as it does in prokaryotes with the polymerase synthesizing pre-mRNA in the 5’ to 3’ direction. As discussed previously, RNA polymerase II transcribes the major share of eukaryotic genes, so this section will focus on how this polymerase accomplishes elongation and termination.

Termination

In Bacteria

Once a gene is transcribed, the bacterial polymerase needs to be instructed to dissociate from the DNA template and liberate the newly made mRNA. Depending on the gene being transcribed, there are two kinds of termination signals. One is protein-based and the other is RNA-based. **Rho-dependent termination** is controlled by the rho protein, which tracks along behind the polymerase on the growing mRNA chain. Near the end of the gene, the polymerase encounters a run of G nucleotides on the DNA template and it stalls. As a result, the rho protein collides with the polymerase. The interaction with rho releases the mRNA from the transcription bubble.

**Rho-independent termination** is controlled by specific sequences in the DNA template strand. As the polymerase nears the end of the gene being transcribed, it encounters a region rich in C–G nucleotides. The mRNA folds back on itself, and the complementary C–G nucleotides bind together. The result is a stable **hairpin** that causes the polymerase to stall as soon as it begins to transcribe a region rich in A–T nucleotides. The complementary U–A region of the mRNA transcript forms only a weak interaction with the template DNA. This, coupled with the stalled polymerase, induces enough instability for the core enzyme to break away and liberate the new mRNA transcript.
In Eukaryotes

The termination of transcription is different for the different polymerases. Unlike in prokaryotes, elongation by RNA polymerase II in eukaryotes takes place 1,000–2,000 nucleotides beyond the end of the gene being transcribed. This pre-mRNA tail is subsequently removed by cleavage during mRNA processing. On the other hand, RNA polymerases I and III require termination signals. Genes transcribed by RNA polymerase I contain a specific 18-nucleotide sequence that is recognized by a termination protein. The process of termination in RNA polymerase III involves an mRNA hairpin similar to rho-independent termination of transcription in prokaryotes.

In Archaea

Termination of transcription in the archaea is far less studied than in the other two domains of life and is still not well understood. While the functional details are likely to resemble mechanisms that have been seen in the other domains of life the details are beyond the scope of this course.

Cellular Location

In bacteria and archaea

In bacteria and archaea, transcription occurs in the cytoplasm, where the DNA is located. Because the location of the DNA, and thus the process of transcription, is not physically segregated from the rest of the cell, translation often starts before transcription has finished. This means that mRNA in bacteria and archaea is used as the template for a protein before the entire mRNA is produced. The lack of spacial segregation also means that there is very little temporal segregation for these processes. The image below shows the processes of transcription and translation occurring simultaneously.
Here the pale blue circles represent RNA polymerases proceeding along a DNA template (from left to right). Note that multiple polymerases can load sequentially onto a single gene. Because this is a prokaryote, ribosomes can begin to use a transcript to synthesize protein before the transcript is complete. Note also that multiple ribosomes can
In Eukaryotes....

In eukaryotes, the process of transcription is physically segregated from the rest of the cell, sequestered inside of the nucleus. This results in two things: the mRNA is completed before translation can start, and there is time to "adjust" or "edit" the mRNA before translation starts. The physical separation of these processes gives eukaryotes a chance to alter the mRNA in such a way as to: extend the lifespan of the mRNA or even alter the protein product that will be produced from the mRNA.

**mRNA Processing**

**5' G-Cap and 3' Poly-A tail**

When a eukaryotic gene is transcribed, the primary transcript is processed in the nucleus in several ways. Eukaryotic mRNAs are modified at the 3' end by the addition of a poly-A tail. This run of A residues is added by an enzyme that does not use genomic DNA as a template. Additionally, the mRNAs have a chemical modification of the 5' end, called a 5'-cap. Data suggests that these modifications both help to increase the lifespan of the mRNA (prevent its premature degradation in the cytoplasm) as well as to help the mRNA initiate translation.

*Figure: An example of an almost-completely eukaryotic transcript and the signals required for addition of the polyA tail (not yet added!). These signal will be cleaved from the transcript upon addition of the tail.*
Figure: the 5’ cap of eukaryotic transcripts. Often the first 2 nucleotides (here purple) of the mRNA are modified also. Note the odd 5’ to 5’ linkage.

Splicing

"Splicing" of eukaryotic mRNAs refers to the removal of noncoding sequences (referred to as introns) embedded within the coding region of the transcript (the exons) (see below). Dr. Britt's section will not discuss splicing at length, except to refer to it as a process that must occur before the mRNA is fully mature and can be shipped to the cytoplasm for translation.