Lecture 13

Cistrons, Genes, and Exons

What is a gene?

Seriously... What is a gene? Is a gene defined functionally or structurally?

Some people would define a gene as the segment of DNA that specifies the amino acid sequence of an expressed protein, but that leaves out rRNA "genes" and tRNA "genes". Such a definition would neglect the regulatory regions such as transcriptional promoters that are not transcribed but have functional importance in gene expression. It would also tend to ignore elements that are transcribed but not translated, such as 5' and 3' untranslated regions of mRNAs that may have roles in controlling the stability of the RNA, or intervening sequences (introns).

What is a "cis-trans" test? This is a rather old example, but you should learn it to impress your elder colleagues at scientific meetings (i.e. those who sit at the bar after hours, talking about the good old days and wondering why young people don't know anything anymore).

Suppose you were studying a biochemical pathway that takes a substrate A to product B, and then uses B as a substrate in a second reaction to make product C:

$$A \longrightarrow B \longrightarrow C$$

You isolate two mutations, which will be called x and y, each of which can prevent A from being converted to C. Are they in the same gene? Here's how people used to find out.

It is possible to make bacteria partially diploid for a set of genes, using bacteriophage methods. If loss of function (recessive) mutations x and y are on the same DNA strand we
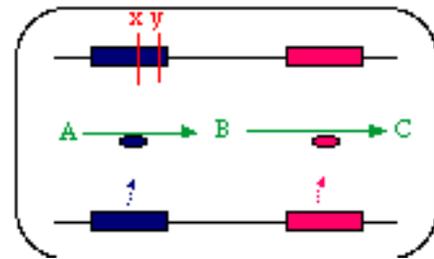
say that they are in "cis" and if they are on different DNA strands we say that they are in "trans."

First, a control: If x and y are placed in cis, we predict that the second copy of the DNA will be wild type, and A will be converted to C.
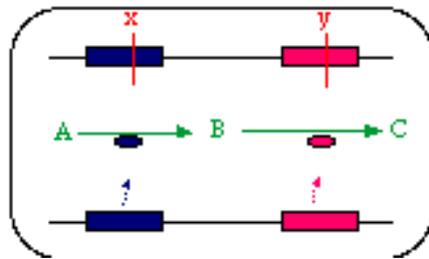
Mutations in different functional units, in cis
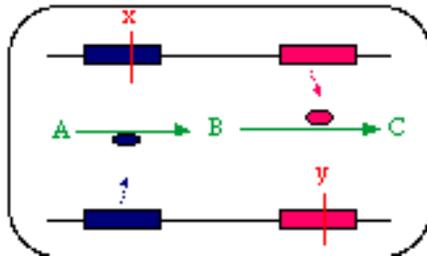
A is converted to C, because both enzymes are made

Mutations in same functional unit, in cis
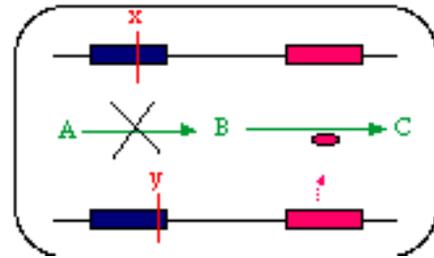
A is converted to C, because both enzymes are made

That is true regardless of whether the x and y mutations are in the same functional unit or not, as the two examples show. The blue and magenta ovals (i.e. enzymes) can be made from the wild type strand.

Now the fun begins: If x and y are placed in trans, different results are possible

| Mutations in different functional units, in trans	Mutations in same functional unit, in trans
A is converted to C, because both enzymes are made. These mutations are said to complement each other.	A is NOT converted to C, because the enzyme to convert A to B is not produced.

This is called the cis-trans test, and mutations that behaved differently in the cis and trans configuration (as shown in the case on the right) were said to lie in the same "cistron" - a word derived from "cis-trans". (For more reading on complementation and cis-trans test, see the web site by Holmes and Jobling)

Remember that if you performed the cis-trans test, you didn't have the benefit of knowing where the blue and magenta boxes were. You only knew whether the biochemical reaction converted A to C, and whether the mutations were in cis or trans. It was the difference in behavior in "cis" and "trans" that led you to know whether x and y were mutations in the same cistron. The cistron was a functional definition (at the time) of a gene, and cistron is a word that remains in our vocabulary.

The cis-trans test was used in bacterial genetics, where genes are represented by a single block of contiguous sequence. As you know, starting in 1977 it was discovered that many eukaryotes have non-contiguous protein-coding genes that are assembled into a single coding region by the process of RNA splicing.

The eukaryote Looking just at the DNA surrounding a gene, we see that in eukaryotes the coding regions are separated into blocks of sequence that are separated by introns. They are shown below as colored blocks. Note also the regulatory sequences that affect expression, such as an enhancer, promoter, and transcription terminator (ter). This is a simplification - there may be multiple promoters with additional surrounding sequences responsible for activation and repression, and there may be multiple terminators.
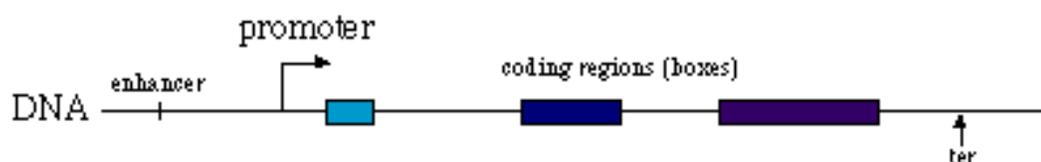

It is a general rule that each gene in a eukaryotic organism needs to have its own transcriptional promoter.
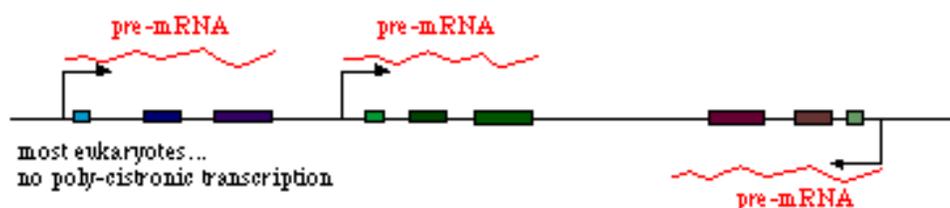

Generally speaking, polycistronic transcription (in which multiple cistrons or genes are included on a single pre-mRNA transcript) is not performed in eukaryotes (there are exceptions to every rule of course).

As you know from our discussion of operons, prokaryotes generally do have polycistronic transcription, and prokaryotic ribosomes are capable of internal initiation in a transcript.
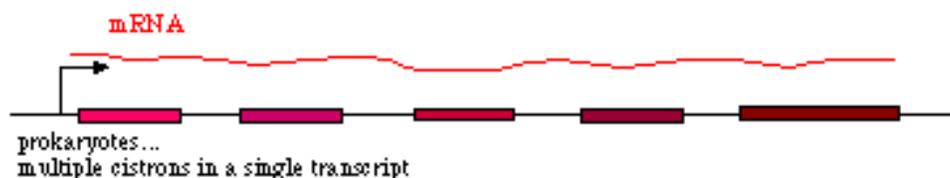
In prokaryotes, a single transcript may encode many different proteins, as genes are organized in operons under the control of a single transcriptional promoter.

When a pre-mRNA transcript is generated in eukaryotes, it may carry a number of regulatory elements (as shown below: SD = splice donor, SA = splice acceptor, PA = cleavage and polyadenylation site).
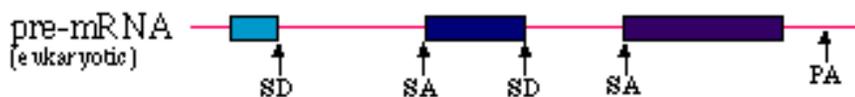
Cleavage and polyadenylation are performed by a polyadenylate polymerase approximately 30 nucleotides after a "AAUAAA" sequence. The poly(A) "tail" that is added is approximately 20-250 adenylate residues in length.

Splicing depends on recognition of splice donor and splice acceptor sites, which are specified by a variety of nearby sequences. By and large, splice donors have a GU dinucleotide at the 5' end of the intervening sequence and splice acceptors have an AG dinucleotide at the 3' end of the intervening sequence.

A pre-mRNA might look like this:

Where the green nucleotides represent sequences in exons, and the aforementioned GU and AG are in red, then the spliced product would have the following structure:

I've overlayed the green mRNA sequence with red lines to indicate how a ribosome might read the triplet code. Note that splice junctions need not be placed right at the junctions between codons - there is really no relationship between the two. A codon at a splice can be assembled from two pieces.

Finding a splice junction is more difficult than just looking for GU and AG dinucleotides. While those are nearly universal consensus sequences, not every GU is a splice donor and not every AG is a splice acceptor. There are also many examples of "alternative" splicing in which a gene may be spliced in many different ways. Predicting splice junctions is an art (see several prediction programs on line: pasteur and cshl) and it is critical in making sense of the large genome sequencing projects.

What happens to the intron RNA? It is initially a byproduct that contains a 2'-5' branch as a result of the nucleophilic attack in the first step of splicing.

2'-5' branch

The overall reaction may look something like this:

(Reload to start animation)
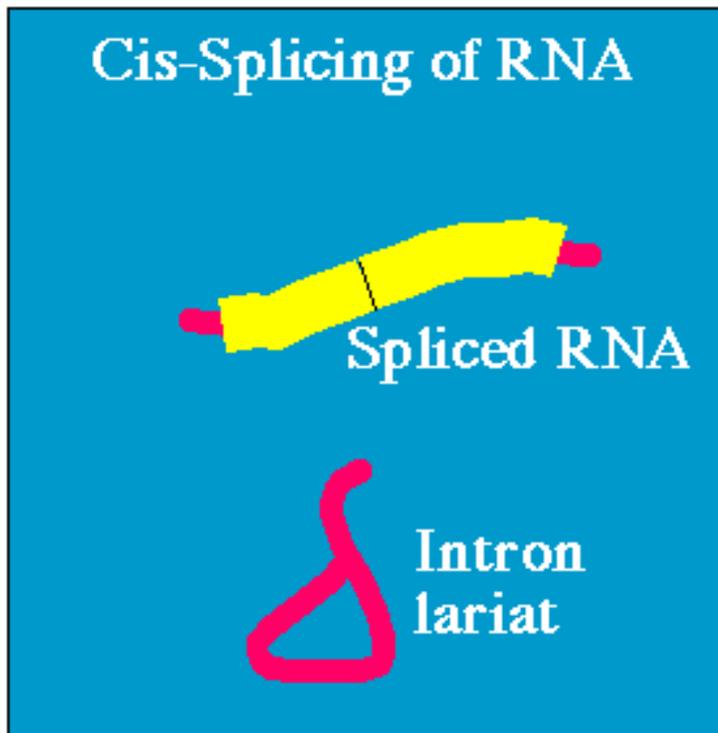
...Except that there is something wrong with the picture. The process requires a set of small nuclear ribonucleoprotein particles (snRNP):
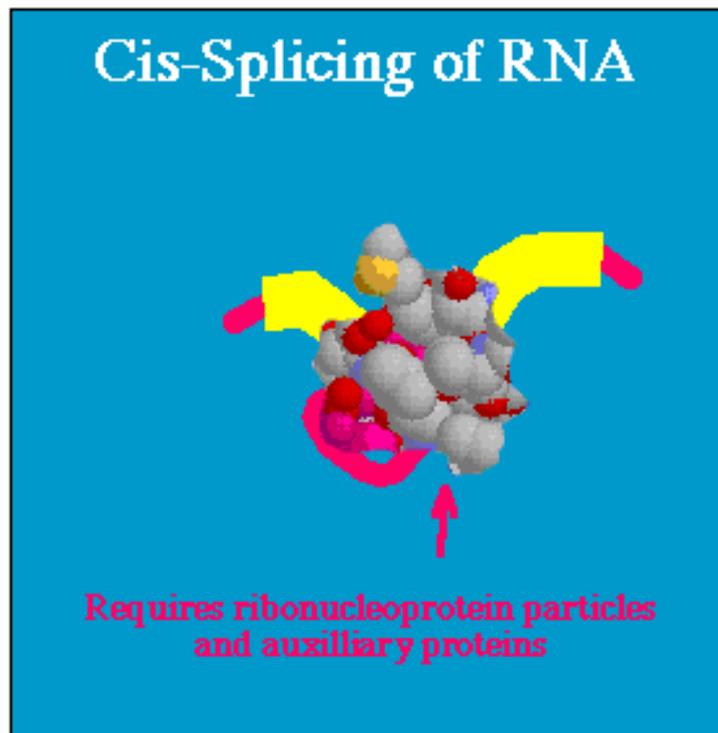
(For more information on splicing and other regulatory devices, visit Michael King's site in Indiana or the Oregon State

University's site)

In the end, here's what is made: a mature mRNA ready for translation by the ribosome.

This being a course in which we discuss techniques, this would be a good time to talk about how you work with RNA in the laboratory.

When we study the expression of genes, it is important that we analyze the structure of the mRNA. Many eukaryotes perform RNA splicing on transcripts, so the distinction between a genomic DNA sequence and an mRNA sequence is very important.

RNA gels

Determining the size of RNAs is often an important component of analysis. RNA can be analyzed on a gel in much the same way as DNA, but there are a few modifications in approach:

- Since RNases (RNA degrading enzymes) are common and difficult to inhibit, all of your reagents and equipment must be kept clean and free of contamination. Workers usually wear gloves when working with RNA, not to protect themselves from their sample, but rather to protect their sample from themselves.

- RNA stains poorly with ethidium bromide, but fortunately now there are fantastic dyes available for the staining of RNA in a gel. A "**silver stain**" can be used to detect very small quantities of RNA.

- RNA tends to form secondary structures (by hydrogen bonding between different parts of the same molecule), and that will alter its mobility on a gel. Since the shape of the molecule is less predictable than it is with double stranded DNA (because of the hydrogen bonding), the RNA bands will not be well resolved under hydrogen-bonding conditions. The solution is to prepare the sample and run the gel under conditions where the secondary structures are minimized. That is, make the RNA entirely single stranded and keep it that way!

How does one "flatten out" the RNA so that it will run in a predictable fashion on a gel?

- Denaturing chemicals (Urea, formamide, formaldehyde, glyoxal)

- Heat

- All of the above

The denaturation of RNA is usually started prior to loading of a gel, and then steps are also taken to continue the denatured state during running of the gel.

When formamide is added to an aqueous solution, in a final concentration of approximately 40-50% v/v, it causes the hydrogen bonds in nucleic acids to be less stable. That is, the strands of nucleic acid (whether DNA or RNA) fall apart more easily, or we say that the "melting point has dropped." In the case of an RNA molecule, the prevention of hydrogen bonding between different parts of the same molecule prevents secondary structure formation.

A typical treatment for RNA before it is loaded on a gel, is resuspension in a buffer containing 50% formamide, heating it to 95-100 C for 3 to 5 minutes, then plunging the sample tube into ice. Why ice? Because we do not want the sample to

"reanneal" and a rapid drop in temperature discourages that from happening.

The gel often contains 7 to 8 M urea as a denaturant, if the matrix of the gel is polyacrylamide (i.e. a vertical gel). In that case the gel may be run "hot" as well -- approximately 50 to 55 C. The wattage of the gel electrophoresis provides the heat, and the front or back of the gel plates may have a large heat sink or other mechanism for distributing heat evenly over the surface. The heat and urea both prevent hydrogen bonding in the sample, so the bands are "sharp" and well-resolved.

If the gel is prepared with agarose, formaldehyde is often used as a denaturant (though that requires running the gel in a hood to avoid exposing your nose and eyes to the fumes). Formaldehyde is added to the agarose mix after it is boiled and cooled to approximately 60 C (you really don't want to melt your agarose with the formaldehyde mixed in!).

An alternate method is to pre-treat your RNA with glyoxal, which chemically modifies the bases to prevent secondary structure formation. After glyoxal treatment, the sample can be run on a more traditional gel that lacks denaturants.

In the good old days (pre 1980), people used to run "methyl mercury gels" as a denaturing gel electrophoresis method, but that was extremely dangerous.

Northern blots This is really the same sort of idea as Southern blots, in the sense that a nucleic acid is being run on a gel and transferred to a membrane. What is different is that there are no restriction enzymes involved in the analysis. One is usually looking at an intact structure. The issues of probes and hybridization are largely the same, although one must remember that only one strand of RNA is usually present on the gel, unlike DNA in which both strands may be present after denaturation.

Read the following mirrored sites:	
Cornell University site	[Northern blots (page 1)](#) [Northern blots (page 2)](#)

What are cDNAs?

It is not feasible to clone RNA sequences directly. As the next best thing, we prepare a complementary DNA (cDNA) that contains the same genetic information but is in a form that can be ligated into a plasmid.

The method of cDNA synthesis

In order to copy the RNA sequence into complementary DNA sequence, we need to use an enzyme called "reverse transcriptase" which is an RNA dependent DNA polymerase. This enzyme requires a primer, and for the generation of cDNAs from a collection of polyadenylated RNAs, we often use oligo(dT); a short stretch of deoxythymidylate residues.

First Strand cDNA synthesis

Extension of an oligo(dT) primer using reverse transcriptase, an RNA-dependent DNA polymerase of retroviruses

There are several problems with generating a full-length cDNA strand:

- RNA is often digested by ubiquitous RNases.

- Reverse transcriptase stops at certain secondary structure features of RNA.

The reaction shown above is called the "**first strand** synthesis," because it yields a heteroduplex of DNA and RNA base paired to each other. There are three common methods for initiating synthesis of a **second strand** to make the cDNA double stranded. Since one is now copying DNA, it is not necessary to use reverse transcriptase -- the issue at hand is how to prime the synthesis of the second strand?

The second strand synthesis

In the first method, the RNA still annealed to the cDNA is nicked by the enzyme RNase H. Each nick (free 3' hydroxyl) can be used as a primer by DNA polymerase I.

Priming synthesis of the second strand

1. Random nicking of the RNA in a heteroduplex by RNase H

mRNA
5' AAAAAAAAAAAAAAAAAAAAAAAA
TTTTTTTTTTTT-5'
cDNA
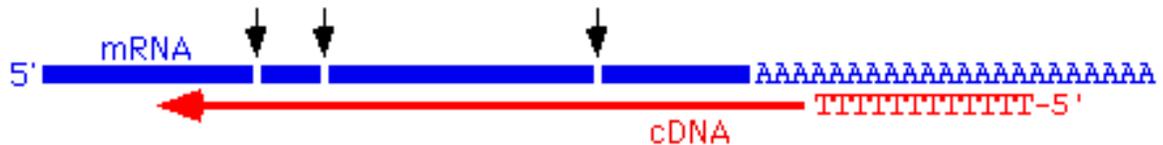
2. Extension of internal 3' sites by "nick translation" using DNA pol. I

mRNA   second strand
5' AAAAAAAAAAAAAAAAAAAAAAAA
TTTTTTTTTTTT-5'
cDNA

The DNA polymerase step also suffers from problems of extension, in that secondary structures can cause the enzyme to stop prematurely. If RNase H is used to generate free hydroxyls as primers, the 5'-most random nick in the RNA is as far as one can go towards copying the 3' end of the first cDNA strand. That is, one cannot expect complete copying of a second strand because the primers are randomly created.

In a second method of priming synthesis, the RNA template is

destroyed (by alkaline hydrolysis) and the 3' end of the first strand is allowed to form a "**hairpin**" structure, priming the second strand.

A second method of priming synthesis of the second strand

Hydrolyze the RNA with alkali and allow the cDNA to "self-prime".

A third method is to use the enzyme terminal deoxynucleotidyl transferase (which we usually just call "**terminal transferase**") to add a homopolymer of dC to the 3' end of the RNA (this is an unusual DNA polymerase that does not require a template). An oligo(dG) primer can then be used to initiate second strand synthesis of the cDNA.

A third method of priming synthesis of the second strand

1. Hydrolyze the RNA with alkali and "tail" the DNA with dC, using the enzyme "terminal transferase".
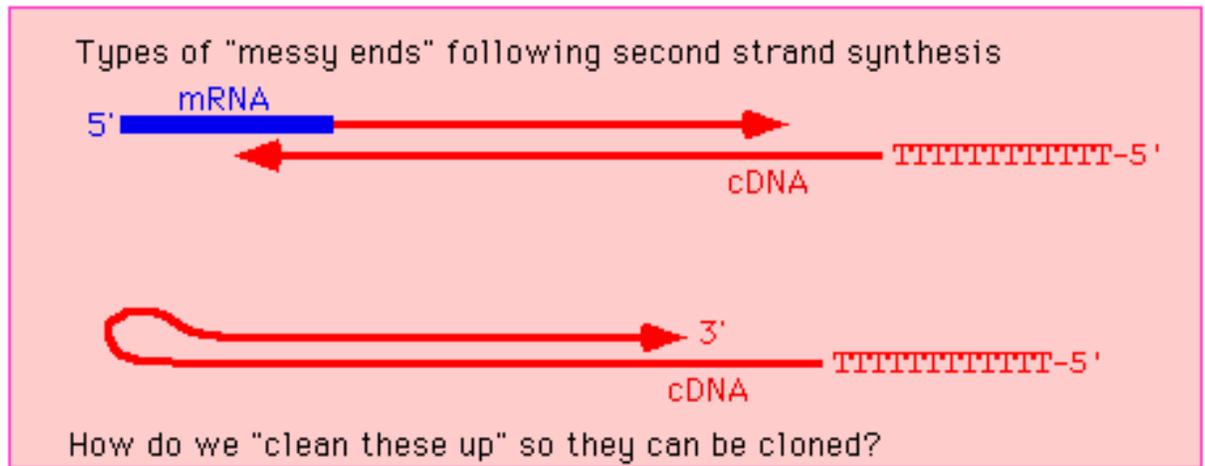

2. Use oligo(dG) as a primer for second strand synthesis.

Cloning a messy duplex DNA

If the hope is to clone the products of cDNA synthesis, the ends must be fixed. Here are common problems with the cDNA ends: 5' and 3' overhanging ends and hairpins.

Types of "messy ends" following second strand synthesis

How do we "clean these up" so they can be cloned?

A variety of enzymes may be used during the "clean-up" phase, including **S1 nuclease** (to digest single stranded regions such as hairpin structures) and T4 DNA polymerase (to digest 3' overhangs and complete synthesis of the second strand).

Once the cDNA is double stranded and has blunt ends, it can be cloned.

RT-PCR

A modern twist on the story is that if you know some of the sequence in your gene of interest, you can simply perform a PCR reaction on the first strand cDNA. This is called RT-PCR (for Reverse Transcription-Polymerase Chain Reaction).
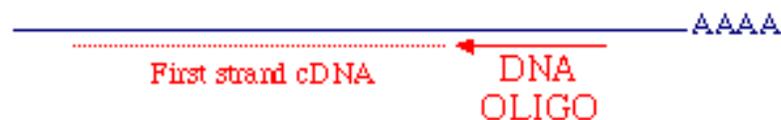
The first strand would typically be made using a specific primer:
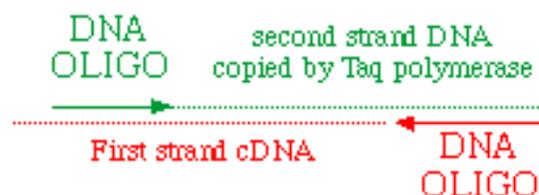
The second strand would be made using a second primer that could anneal to the cDNA.

From this point, a PCR reaction can be used to make additional product, and the net result is that an RNA sequence can be determined from very little template.

Stan Metzenberg
Department of Biology
California State University Northridge
Northridge CA 91330-8303
stan.metzenberg@csun.edu

© 1996, 1997, 1998, 1999, 2000, 2001, 2002