

MIT OpenCourseWare  
<http://ocw.mit.edu>

7.012 Introduction to Biology, Fall 2004

Please use the following citation format:

Eric Lander, Robert Weinberg, and Claudette Gardel, *7.012 Introduction to Biology, Fall 2004*. (Massachusetts Institute of Technology: MIT OpenCourseWare). <http://ocw.mit.edu> (accessed MM DD, YYYY). License: Creative Commons Attribution-Noncommercial-Share Alike.

Note: Please use the actual date you accessed this material in your citation.

For more information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>

MIT OpenCourseWare  
<http://ocw.mit.edu>

7.012 Introduction to Biology, Fall 2004  
Transcript – Lecture 12

Good morning. Good morning.

So, I'd like to pick up where we left off last time and just finish off translation and then step back and look at how this central dogma of DNA is replicated into DNA, is read into RNA, and is translated into protein.

Or, actually, as Francis Crick really put it, all information flow from nucleic acid to protein. How that varies amongst organisms. Because first we're going through it and looking at the absolutely common features, DNA replication, so it's five prime to three prime, et cetera, et cetera, transcription, translation. But in a moment I'd like to turn to the variations between different kinds of organisms. But let me briefly finish up, if I may, the bit about translation in general so we can look at its variation.

As we talked about last time, we have a messenger RNA that has been transcribed from a specific region of the chromosome starting at a promoter and going to some stop of transcription. And that messenger RNA will include some particular sequence, and I'll copy one here, A-U-A-C-G-A-U-G-A-A-G-A-G-G-C-C-C, et cetera, et cetera, et cetera, out to a UAG.

And this is the direction five prime to three prime. We'll remember that all nucleic acid polymerization goes five prime to three prime. So, what happens is the cell begins scanning this message. And it does that by this message being exported into the cytoplasm of the cell. The ribosome coming along and glomming onto this message and scanning on for the place to start.

It looks, it looks, it looks, it looks, and it finds the first AUG. Footnote, this isn't 100% true. There are occasional messages that start their translation not at an AUG, and there are even occasional, there are even more messages that don't quite start at the first AUG because the ribosome is really is looking for something a little bit special, but to a first order approximation.

Good enough for the textbooks. It goes along to the first AUG. In reality it's a little more subtle than that. But it starts at the first AUG. And what it does is it builds a protein that corresponds to it according to a three letter genetic code. And you all know the lookup table. It's in your book. AUG, always the first amino acid put in. A methionine.

Then AAG. Lysine, I think. Then arginine. Then a proline. Now, I mean this is this particular sequence. Any other sequence would be different. Et cetera. How does it accomplish this matching between three letters of the genetic code? Oh, and when it gets to AUG, that is one of the three singles for stop, don't put in any more amino acids. There are three such stop signals.

AUG, sorry, UAG, UGG and UGA, oops, what did I just do here? Let's get that right. UAG, UGG and UGA. Those are the three stop codons. So, how many total codons are there? 64 codons. Three of them spell stop. 61 of them spell specific amino acids.

And how many amino acids are there? 20. So, the average redundancy is three. Some are specified by multiple codons. The most extreme is some amino acids are specified by as many as six codons. Did I, oh, thank you. Come back down. Of course. U-A, so it's UAG, right?

Sorry, UAA and UGA and UAG. Thank you. Very good. All right. So, now, how does it accomplish this feat of taking amino acids, of taking nucleotide sequence, RNA sequence and converting it into the sequence of amino acids? As I mentioned last time, there was lots of original somewhat nutty thinking about some looping codes that would make the RNA fold up in such a way to bind the amino acids and all that.

But, as Francis Crick thought up, there had to be some kind of an adapter molecule that would take the RNA sequence and would somehow connect it up to the correct amino acid, and that was UAC.

A particular transfer RNA molecule. And the tRNA molecule is an adapter sequence that has three nucleotides here that match up to the three nucleotides of the codon that we're trying to translate, and it has the appropriate amino acid that's been stuck on the end of it. And how does it get there? How does the right tRNA, the tRNA to match this codon have the right amino acid put on it?

There's a dedicated enzyme that recognizes that tRNA and puts on that amino acid. It's aminoacyl-tRNA synthetase. It sticks the right amino on the right transfer RNA. So, that's how it accomplishes the physical recognition of these three bases and has the right amino acid attached to it. There's an enzymatic machinery that has all of these tRNAs floating around in the cell which can be used for this translation here. How does this actually happen physically?

It happens in this vast machine called the ribosome. In the ribosome, if we have, say, our codon here and we have a tRNA that, well, we'll put that actually in the ribosome that, say, has the first amino acid here, methionine, there's a cavity for this guy and there's a cavity for the next guy.

And other tRNAs come into the cell carrying their next amino acid. Maybe it will be here a lysine that matches up with the codon and the anti-codon. And when the right tRNA fits in the next cavity over, the ribosome itself catalyzes a peptide bond between these amino acids.

Then it chugs over by one, it translocates by one moving this bit of the complex to the left, and the peptide chain continues to grow out this end as each new codon is moved into position, a tRNA comes in bringing the right amino acid until finally a stop codon is hit. And what happens when you hit a stop codon? It stops. And is there a tRNA for a stop?

It turns out there's not. There actually isn't. There's some other factor. There's a protein factor that helps recognize the stops. So, that just continues to chug on. Those of you who are computer scientists or mathematicians will recognize this is a two-tape Turing machine. It is the smallest two-tape Turing machine that I know to exist. If you don't know what that means, you can forget about that comment. In any case, but some of you know what that is. So, that's how it proceeds. That is your basic protein translation.

And, I must say, what I really love about this was that Francis Crick kind of figured out what had to happen just on first principles and was able to think through it much more clearly and direct people to know what to look for in the laboratory. And if people had not had the clarity of thinking that Crick provided by saying, look, there's got to be this kind of adapter, I don't think they would have found it as quickly. But once he said this is what you've got to look for, golly, it was there. You can't do that very often, but Francis Crick seemed to have a very

good track record of doing those things. OK. So, that was just finishing off translation.

Now what I'd like to do is turn to variations on the theme as the major issue for today. How does this central dogma, DNA replicates, is transcribed into RNA and is translated into protein, vary amongst the different kinds of organisms that we might be interested in?

The kinds of organisms we might be interested in, eukaryotes, prokaryotes, viruses. Sample eukaryote, MIT undergraduate. Prokaryote, E. coli. And virus, many possible viruses. The eukaryotes' big nucleated cells. So, in here we're going to have our nucleated cells.

DNA living in there. In our prokaryotes we have no distinct nucleus. The DNA is not in a distinct nucleus, although it's not entirely freely floating around. It tends to be clustered together. In the virus the nucleic acid resides in some kind of a capsid, some kind of a, it could be a protein capsid.

There are some of them that have lipid capsids with lipid particles around them, but some kind of a coat around nucleic acid there. Do they all do exactly the same things with regard to DNA replication, RNA transcription and protein translation? Well, not entirely. So, as a way, in a way to reinforce what we know about these, let's look at how they differ. DNA replication. Eukaryotes.

What's the structure of one of your chromosomes? Is it a long line, a long linear molecule, or is it a circular molecule? How many of you have linear chromosomes? How many of you have circular chromosomes? I heard there were some people with circular. And how many of you are unsure about your chromosomes?

OK. That's good. Well, then I'm pleased to inform you that you have long linear chromosomes. Every human chromosome is a long double-stranded molecule of DNA. Linear double-stranded DNA. They can be extremely long.

You have 23 chromosomes, and together they make up three billion nucleotides of DNA. A typical chromosome could be 150 million bases long as an average size for a chromosome. And it's a single connected molecule. 150 million bases long in the human is a typical chromosome. One tricky little bit about replicating DNA.

Let's just think back to our little model of replicating DNA. Let's come to the chromosome end here. It's five prime to three prime. Five prime to three prime. We're going to start replicating. We're getting to the end of chromosome number one. We've got a primer here, and the primer is going to be used to extend, extend, extend. We get right to the end. That's good. Tell me how we're going to replicate back.

We need a little primer to start it, right? And where's that primer going to land? Maybe over here it will start replicating back. Oh, boy, we haven't done this figured. So, what do we have to do there? So, we need to primer a little further back. OK. But, you know what, the chance that we're going to get that right at the end, that we're going to get a primer exactly at the end is pretty low.

And if we don't have a primer exactly at the end, what's going to be wrong with that copy of the chromosome? Too short. Now, big deal. So, it's short by maybe 20 bases. But that's just this cell division. What about next cell division? It will be short on average by a little bit, and then the next cell division and the next cell division. It's actually pretty tricky to replicate a linear chromosome on the lagging strand, unless you can land the primer in exactly the right place, which doesn't happen.

So, a special little solution is used. The ends of chromosomes here are called telomeres, telo meaning end. These telomeres have very specific structures. In the human they repeat, T-T-A-G-G-G, again and again and again.

At the end of the chromosome there's a special enzyme that will come along and add some extra telomere to the chromosome. That, sorry? Did I say leading strand? It's the, oh, yeah, sorry. It's the lagging, sorry. It's the leading strand. No, no, no, this is the lagging strand. This is the leading strand because it's running along happily not having to make a primer. The Okazaki fragment should be here.

I'll stick by that. We'll debate it later. Anyway, they, we get the point. But it's lagging because you've got the Okazaki fragments there. So, anyway, we have a problem of replication. And the way the cell solves it is the actual replication is shorter, but since it manages to stick some repeat at the end of the chromosome it adds back some more T-T-A-G-G-G, T-T-A-G-G-G, T-T-A-G-G-G, and it keeps dynamically adding more.

What do you think would happen if you didn't, or what's the enzyme that adds telomeres? Telomerase. Telomerase adds that. What cells do you think need to have active telomerase? Rapidly dividing cells would need to have telomerase. Cells that are not rapidly dividing, cells that have stopped dividing can shut off their telomerase. But if a cell is going to go through lots and lots of cell divisions it's got to, it's got to tidy up its telomeres each time because they're getting too short.

You've got to have an enzyme that's adding back ends of chromosomes. What cells do you think particularly care about having telomerase on them? Cancers. It turns out that this is not a trivial point. More than 90% of cancers turn on actively the telomerase gene, which would be a shut off in normal cells because the cell is not dividing anymore.

Part of becoming a cancer is having to turn on this repair mechanism for the ends, this extension mechanism for the ends of your chromosomes. And so, various people are trying to make drugs to inhibit cancers by inhibiting this telomerase enzyme. So, understanding just your linear replication of chromosomes is a kind of useful thing even in dealing with things like cancer. Genome sizes. I mentioned, how big was the human genome? Three times ten to the ninth bases.

The mouse genome? It's almost as big, about 2.7 times ten to the ninth bases, 2.7 billion bases. The elephant genome? I actually just found this out last week because we just finished sequencing elephant DNA, and I can now tell you I think it's 3.1. The dog is 2.5 times ten to the ninth. Anyway, it's about, for most mammals it's pretty close to three billion bases.

And there is some fluctuation. Some are a little bigger. Some are a little smaller. It doesn't scale with sizing the animal, though, because the dog has a smaller genome, for example, than the mouse does, but the elephant is a bit bigger than us. And check in later in the term, I'll tell you about the aardvark. We should know in a little while. But here are, for example, fruit flies. The fruit fly, it has a genome of two times ten to the eighth.

I'm giving, I'm being quite approximate. In fact, I'll make it, I'll give you 1.5 times ten to the eighth. 150 million bases. Yeast, by contrast, has a genome of 1.2 times ten to the seventh. So, that's 12 million, 150 million give or take, and about three billion, so 3,000

million. So, genome sizes can vary quite dramatically amongst different eukaryotes.

Now, what about prokaryotes? How do the prokaryotes differ? Prokaryotes differ because their genomes are typically not linear chromosomes. The typical prokaryotic chromosome is a double-stranded circle. It's a double-stranded circular DNA.

Now, the double-stranded circular DNA doesn't have this problem of telomeres. You just keep replicating around and you get to the end. So, there you have a much simpler replication system than having to worry about your ends of chromosomes. You also have much smaller genomes. The typical prokaryotic genome size, it's on the order of a few million bases. *E. coli*, 4.6 million bases.

There are, for example, mycobacteria, such as the mycobacteria that caused tuberculosis or leprosy, have on the order of, well, actually, not quite them, but other mycobacteria have on the order of about a million bases or so. Mycobacteria, *M. genitalia* has actually slightly less than a million bases. So, these are basically several million bases. So, there's a huge variation in genome size.

Your genome is about a thousand times bigger than *E. coli*'s genome. Now, you do actually have one circular chromosome. Do you know what it is? I speak about the 23 pairs of human chromosomes. There's actually one more human chromosome. The mitochondria have their own chromosome. It's a circle. That's very odd that you would have one chromosome that's a circle that looks like a bacterial chromosome. Do you know why that is?

The mitochondria arose as a symbiotic bacterium that became a symbiont of eukaryotic cells about 1.5 billion years ago. It was a bacterium taken up into another cell, and that's how eukaryotes evolved. And we can even see that little signature of it having been a prokaryote from the fact that it's got one of these circular prokaryotic looking chromosomes. Now, it, because it's living in your cells, has thrown out all sorts of genes that it doesn't need anymore because the main, the nucleus supplies most of the proteins.

So, your mitochondrial genome is a circle that's a mere 16,000 bases long. It's a very small circle encoding a very limited number of genes, but it's, in fact, the residue of the bacterial symbiont that led to the formation of euks. Now, viruses, what do viruses have? Do they have double-stranded linear chromosomes? Which is it?

Is it double-stranded linear DNA or is it double-stranded circular DNA? Circular DNA. So, who votes for linear? Who votes for circular? Who's undecided? Ah, the undecided are very larger here. So, the answer is both.

Some viruses have double-stranded linear DNA. Some viruses have double-stranded circular DNA. It's worse than that, though. Some viruses have single-stranded linear, circular DNA. Ha? How does that work? Some viruses actually infect the cell injecting DNA, and it's just single-stranded.

As soon as it gets into the cell, however, it's replicated to make a double-stranded DNA which can then be transcribed, et cetera, et cetera, et cetera. But it travels around as a single-stranded piece of DNA. And it's actually weirder than that. Some viruses, viruses being very small can experiment with all sorts of things.

Some viruses actually consist not of DNA at all but of RNA, single-stranded RNA. How does it do that? So, in other words, in the capsid there's a single strand of RNA. When it gets into the cell, what does it do? Sorry?

It creates DNA. How does it create DNA? From the RNA. How's it going to do that? Well, how is it going to turn itself into DNA? It needs an enzyme to do that? Reverse transcriptase. You would like to reverse the transcription process, and you would like to name that reverse transcriptase. And where are you going to get this reverse transcriptase from? Laying around. Laying around where?

I mean the cell is just sitting there with reverse transcriptase waiting to obligingly reverse transcribe this virus? Your RNA. So make it how? With ribosomes. So, in other words, if I'm an RNA, why don't I encode the sequence for reverse transcriptase and actually translate myself. So, if you were really clever, you might decide to put in the genetic code for reverse transcriptase.

And when that message gets into the cell, it will first act as an mRNA, a messenger RNA, translate, make, here's the reverse transcriptase enzyme, which is then going to go, and it's going to reverse transcribe this thing into, say, DNA. So, wow. Now, that's a good one. This is a plus strand virus.

It encodes its own reverse transcriptase in its instructions. There actually are minus-strand viruses that don't, but what they do is instead in their own code, in their own package bring along a reverse transcriptase. So, either you can encode your own reverse transcriptase or in the package you can include your own reverse transcriptase. Do you know any viruses?

And then the reverse transcriptase is then used to transcribe the DNA, the RNA into DNA, and eventually into a double-stranded DNA which, in some of the viruses, can then be slammed into and inserted into your own chromosomes. So, a DNA copy of the virus can be installed into your own chromosomes, which is somewhat insidious.

What viruses do you know that do this? HIV. More generally retroviruses are the class of these viruses that can, in fact, run this replication process from RNA to DNA and install DNA copies of them in your genome. And how do you then get the DNA copy out of your genome? You don't.

It doesn't come out. Retroviral insertions don't come out. That's one of the issues in dealing with AIDS is once this DNA copy is in a cell it's not coming out. We have no way to remove it. We have to make sure that the virus is shut down by other mechanisms that might inhibit its products, et cetera, but once its stuck a DNA copy into your chromosomes, you know, there's no way of getting it out.

So, if we had to try to inhibit the action of the AIDS virus, we might wish to make inhibitors of this aspect of replication, inhibitors or reverse transcription. And, of course, as probably many of you know, some of the important AIDS drugs are reverse transcriptase inhibitors, very important to limiting the replication of the AIDS virus. And there are many other kinds of weirdnesses. Viruses pretty much explore, everything you possibly can do, viruses come up with ways to do.

Let's take now the process of transcription. We have replication up there. Let's look at transcription. And this time let's start with prokaryotes. For the simple aspect of transcribing genes, the prokaryotic genome looks just like the simple model I gave you. There is some kind of a promoter that tells RNA polymerase to come sit down here.

RNA polymerase hops on, RNA polymerase begins to copy in RNA, and eventually it hits the signal that says to terminate transcription. OK.

This is not a stop codon which is about translation. This is a termination of transcription.

And this RNA then goes off. A perfectly happy thing, a messenger RNA, mRNA. So, there's nothing weird about proks compared to the simple description that we gave before. But eukaryotes are different. There are some funny things that happen in the eukaryote. Well, first off it starts the same. There's a promoter.

RNA polymerase sits down there, it starts transcribing, it makes an mRNA, it hits the transcriptional termination signal, it stops, and then this RNA gets processed in interesting ways. The first thing that happens is three modifications happen.

The first one is at the five prime end, remember five prime to three prime, a funny modification is put on. It's a, if the message, say, were A-U-C-U-G-G-C et cetera, a G triphosphate is put on backwards. It's actually a methyl G triphosphate is put on backwards, so going in the other direction. You have the triphosphate bond there, a methyl G.

And the only thing that you should care about that, I don't care if you know the structure, is that there's a funny cap. This thing is called a cap that is put on this message. And that cap is very important to signaling to the cell this is a messenger RNA to be dealt with in a certain way, to get the ribosome to hop on, to get this thing processed properly, et cetera. At the other end of the message a long string of As is added to messenger RNAs.

This long string of As is called, very sensibly, a poly A tail. The poly A tail is added to the messenger RNA, and very often, I mean it's, if you wanted to purify messenger RNAs from your own human cells, you can actually use poly T as a reagent because it turns out, because messenger RNAs have a poly A tail, they'll bind to and stick to poly T.

So, people actually purify messenger RNAs by binding them to poly T and they get the poly A tail. But it is broadly believed that the reason for this poly A tail is not to make things convenient for molecular biologists to purify messages. To the contrary, it is an important function for the cell. And it turns out that this is important in regulating the stability of messages. If, in fact, you don't have a poly A tail, if you contrive to make the same message without the poly A tail, the message will be degraded rather rapidly.

And the lengths of the poly A tails control aspects of the degradation, et cetera. So, in a complex eukaryotic cell, already it's how to attach a little signal at the front, some signals at the back that says process me in a certain way, et cetera, don't degrade me yet. You could even imagine that this poly A tail could serve as a little bit of a clock for how long that message sticks around. It's not quite that simple but there are ways to do it. But all of these pale in comparison to the third way in which eukaryotic messages differ from prokaryotic messages.

These small modifications are, as I say, small. The most striking way in which they differ is that only a small portion often of the gene, here's my gene, matters for the protein that is made.

So, my mRNA gets made. It includes the whole long sequence. And then the cell comes along and splices this message together. So, this is the immature RNA.

It is processed by clipping out this, clipping out this, clipping out this. And what you get is a splice where the mature message throws this stuff out, splices between here and here, splices here, splices here, splices here, and you get a much shorter mRNA.

And this is a mature mRNA. This splicing is a remarkable phenomenon. In fact, it was discovered by Phil Sharp here, for which he won a Nobel prize.

This splicing is a very complex operation. First off, how does, well, actually, what accomplishes splicing? It should be splicase, right? But it turns out it's not a single enzyme. It's a big body of stuff. So, instead it's the spliceosome, OK. Everything is either ase or some or something like that. So, it turns out it's the spliceosome that does that. It's just wonderful how all those names work out. The spliceosome.

The spliceosome comes along and splices it. How does the spliceosome know how to do this? Well, there are kind of codes. It turns out that there are some information encoded along in these messages. It turns out that there is, you know, slight biases. Typically the sequence just after where the splice starts here is a GU and the sequence here is an AG, but that's obviously not enough information, right?

It's not enough bases of information to get this right. And so there's a little more preferences for what bases use, but the truth is we don't

fully know. Our best picture right now involves some cellular factors help recognizing the parts that are supposed to stay in some sequences here. But the truth is we don't have the simple codes.

Because if we had the simple codes, we'd be able to take a long stretch of DNA and figure out exactly where the splices go based on just computer analysis. And we can't do that so well. These bits that stay in are called exons. The bits that go out are called introns. This is the source of extraordinary confusion for students because you might think that the bits that are excised are the exons, but they're not.

The bits that stay in are the exons. Why are they called exons if they stay in and ex is a prefix meaning out? Well, because the introns are named because they're intervening sequences. Once the introns, the intervening sequences were named as intervening sequences or introns, you were stuck then having to name the things that stay in as exons. This was all done by a Harvard professor, don't blame me. In any case, a good friend Harvard professor.

But, nonetheless, I'm not sure that this was the best way to name them. But you're stuck with it. So, for a typical human gene, typical human gene, the length of the gene itself might be 30,000 bases. But the mature RNA, the mature mRNA might be one and a half, 1,500 bases.

That's remarkable. Out of 30,000 letters in the initial transcript that is made, the genes start, the promoter, and the transcription will stop 30,000 bases away. The cell goes through the trouble of making an RNA of 30,000 bases long, and then it trims it down by throwing out 28,500 of the bases, keeping only 1,500 bases at the end.

Now, this may seem profligate but it ain't nothing compared to some extreme cases. The clotting factor gene, the factor 8 gene, the gene that has mutated in individuals with hemophilia, that gene is 200,000 bases long, and it gets spliced down to a mere 10,000 bases. 190,000 bases are thrown away.

But that's nothing compared to Duchene muscular dystrophy. The Duchene muscular dystrophy is the all time winner. That gene makes an immature initial RNA of 2 million bases. RNA polymerase hops on at the promoter and it gets off at the end of the Boston Marathon on here 2 million bases later having made an RNA of 2 million bases long.

Calculate the speed of RNA polymerase and you'll find out that it's at it for hours. It hops on and it stays on for hours until it gets to the other end. And then for all its troubles this gene is spliced down to 16,000 bases in the mature message. Yup? How would it increase the chance of mutations? Yup.

So, splicing mutations could be a problem. Some diseases could arise from errors in splicing. Do you think that happens? Sure does. There could be mutations that create, that change a splicing, or mutations that create a new splicing, and all of that could screw up the gene. Why do this? What in the world is going on?

Just think about the energetic cost. I mean count up the ATPs involved in synthesizing a nucleotide, and then the ATPs involved in adding nucleotides up. You know, think about this totally wasted energy. What is the point? I might be able to encode multiple proteins with the same gene. How would I do that? Ooh, wouldn't that be clever?

I might be able to take a single gene and make a mix and match product. It might be, do you mean like one type of cell might splice up that message one way to produce a certain protein, but a different cell type might splice the same gene another way to produce a different protein? Ooh. So, you're proposing, if I understand you correctly, alternative splicing.

Alternative splicing could create multiple proteins, multiple distinct proteins. It might be, for example, that you might make one protein that has a cytoplasmic tail and another protein that doesn't have cytoplasmic tail or a different tail or, or, this is true. This actually happens. It's very clever. Anything that can happen does happen somewhere, and it's fairly regularly used. A typical gene in the human being has at least two alternative splice forms, on average.

Most, many don't, but there are some that have large numbers. The most extreme is there's a gene known, drosophila, that has more than a thousand alternative splice forms. How does it know, how does the cell know whether to splice it one way in the liver and one way in a heart or something? We don't fully know but there's machinery and signals people are trying to work out for that. Now, I don't want to confuse you too much about it. You know, mostly, when we give you a gene, you should think about it spliced out introns, exons.

But the truth is it is more complicated than that. There can be alternative splicing that allows genes to be used in multiple ways. Sometimes they don't make multiple proteins. They may splice into portions of the mRNA that are not translated, but, there is that, but, boy, it's a huge amount of overhead here just to do that. Is it justified? Yes? That is by computer if I just gave you the sequence? Not quite. Almost. Maybe.

Sort of. It turns out that the computer programs for automatically recognizing the matter of the human genome are sort of, they're mediocre, not very good. We have some idea of the signals, and various people have trying to write better and better algorithms for doing that, but the cell knows what it's doing and we don't fully know, as evidenced by the fact that we can't write a clean computer program to do it yet. We need to get information from the cell or from evolution or various other things like that, and that's the ultimate test. If we knew what we were talking about we'd just be able to write a computer program and splice it out.

And we don't. There's another reason why people think these big introns and exons, these big introns are helpful, and that is an evolutionary reason. The evolutionary reason is a little bit harder to follow, but let me try it on you. Suppose a random event happens and a chromosome breaks, that happens, and suppose a random breakage sticks one part of a chromosome to some other part of the chromosome.

If it lands smack dab in the middle of the coding sequence of a gene that's bad news. But it turns out that if it lands in the introns of two different genes and sticks them together it could make a new gene that would still work. By having a random break between two genes in their introns and slamming them together, you could make a gene that had a bunch of exons from one gene and a bunch of exons from another gene.

And this intervening sequence in the middle and it would get spliced up. Evolution might like that because it would be a very easy way for evolution to build new genes that had a portion of one protein and a portion of another protein. This kind of mix and match domain swapping could be very useful. And when we look across genomes, we see lots and lots of examples of genes that have a similar first half but different second halves.

Or have some portion in the middle, a domain that we recognize, that we see in multiple proteins. And so, in fact, an argument for why we have all of this intronic DNA, one that's impossible to prove but is an argument is that from an evolution point of view, this allows a great deal of evolutionary innovation. You have to be careful that you say those organisms that have this extra space are able to mix and match and create more new kinds of combination proteins, et cetera, et cetera, and therefore survived better, et cetera, et cetera, et cetera.

Why don't bacteria have this? Sorry? They're not as complicated. That's one thought is that we can take a sort of condescending attitude to these bacteria. They're not very, they're just not so complicated. There's another point of view which is bacteria are far more sophisticated than we are because they're under incredibly rigorous evolutionary selection.

You might argue that if I'm a bacterium, can I really afford all this extra DNA? Now, the metabolic cost of all that extra DNA is huge to a bacterium which competes on replication. It's got to divide every 20 minutes, and trying to put in all these extra bases would be very news. So, you might imagine, just to be, you know, stand things on its head, that early life all had introns and bacteria, in the process of competing to be more and more efficient got rid of their introns.

There's actually a large camp of people who think it went that way, that early life evolved with introns, and then bacteria, in the pressure to compete, got rid of them. And there's some evidence to support that. Bacteria don't have introns. Small eukaryotes like yeast that sort of do compete on replication have some introns, but a small number. There are only about 250 introns in yeast. Only about 5% of the genes have an intron and they're small. Bigger eukaryotes have bigger introns. And the bigger you get, often on average the bigger the genome sizes are the more you can tolerate it.

And so I actually think, I actually probably favor this notion that introns were the original state and they've been gotten rid of. And the more pressure you're under to replicate rapidly the less you can tolerate this interesting and complicated innovation. Anyway, that's another way that things differ. And then, finally, viruses can do it either way. Viruses, depending on whether they are prokaryotic viruses or eukaryotic viruses, are able to replicate, are able to either do or don't have splicing.

Last topic. Translation. Here eukaryotes are relatively simple. You get a message, you get a gene, you get an mRNA. The mRNA goes to the ribosome. Here's a ribosome. The ribosome goes to the mRNA, actually, and it starts turning out one protein as it chugs along.

Prokaryotes differ in an interesting way. I get a promoter here that is transcribed into my mRNA, but it turns out that the mRNA can encode multiple independent proteins, protein one, protein two, protein three on the same mRNA.

And a ribosome will hop on here and synthesize this one. A ribosome will hop on here and synthesize this one, and a ribosome will hop on here and synthesize that one. And you have what is called a polycistronic message. Poly, many. Cistronic, cistrons were an old name for coding regions of genes here. Polycistronic messages.

Why would you want to do that, have a single mRNA that encodes multiple distinct proteins, each starting with its own ribosome start site there? Efficiency. Maybe, in fact, these would be, how about, oh, this would be clever, make them multiple steps in a biochemical pathway? Have them coded on a single messenger so then you'd only have to worry about regulating that once. If you have the regulatory machinery to turn on, you'll make all the enzymes for the pathway. And that's exactly what bacteria do. They tend to put all the enzymes for a pathway on a single message so when they want to call up, let's digest hexose this morning, they have a whole thing that will let them be able to do that, poly-cystronic.

That's because they're small genomes. They're pressed for space. And, because of that, they have to slam a lot into a single unit. And this single unit that has multiple genes encoded in a single message is called an operon, and we'll talk more about that. Last of all viruses. Viruses. Viruses have very little room.

Their genomes can be tiny. A typical virus might have a genome of 5,000 bases to 10,000 bases to, in some cases, 200,000 bases, but it hasn't got a lot of room. It wants to pack a lot of protein-coding information in. And some viruses have come up with the most extraordinary way of doing that. Some viruses have gone to the extreme of having RNAs that get made from them that have a sequence --

I'm just going to pick up in the middle of the sequence here. A-C-U-A-C-U-A-C-U-A-C-U. You might decide to read the sequence like this,

that those are the codons, and you'd get a certain protein. But I might also decide to read that sequence C-U-A-C-U-A-C-U-A.

And, of course, I'm giving this in a repeating form because it's easy to note. I could give you any sequence and I could read it in this reading frame, I could read it in this reading frame, or I could read it as U-A-C-U-A-C-U-A-C. In other words, there are three reading frames that, in principle, you could translate a protein from. In a typical prokaryotic gene or eukaryotic gene only one of those is used.

You start at the first AUG and that sets the reading frame. But some viruses are so pressed for space and are so clever and are so efficient that they make messages that have tricks that they actually use two or, in some cases, all three reading frames, which is an extraordinary packing of information density into a simple message. So, the basic point. We have a simple model. DNA is replicated.

Transcribed into RNA. Translated into protein. But there are a lot of important variations between eukaryotes, prokaryotes and viruses. And understanding them can be useful for treating cancer, for treating AIDS, and for treating viral and bacterial infections. Next time.