**GEORGE CHURCH:** We've come to the fifth lecture, the first of a series, on RNA and expression analysis. First, a quick review of last week, and it has significant connections this week, where we talked to the topic of alignments and different algorithms for obtaining pairwise alignments. In particular dynamic programming, this led to an even harder problem, which is multi-sequence alignment, from which we will draw pretty heavily at the beginning of the discussion today. And then, the issues of getting the motifs, another topic we will touch upon several times today, how you get-- once you have a multi-sequence alignment, how this gives you either an independent weight matrix where different positions are independent or in a hidden Markov model where there is some dependence between, say, adjacent nucleotides in a simple sequence such as CG.

So let us carry over these thoughts about multi-sequence alignment motifs and non independence of positions in sequences to the next level higher. We will eventually talk about protein three-dimensional structure. But a really beautiful intermediate between proteins are the complexities of protein structure and the simplicity of double-stranded DNA is the folding of RNAs because they're based on the same rules of double-stranded DNA, but they have the complicated structures of-- begin to have a complicated structure of proteins.

So we'll start with this integration of multi-sequence alignment of motifs with RNA structure, and then we'll switch to tell about how these RNA structures play their role by achieving different levels in the cell. In other words, we want to start to introduce how we become quantitative about the amounts and localization of RNAs in the cell. Some of the measures that we'll be talking about and the tool-- the computational tools will be more appropriate for individual measures and others will be more, what we call, genomics grade, high throughput and high accuracy.

And we have this-- since this is a new category of biological data, we have to address random and systematic errors just as we did for genotyping and sequence data. This is a new set of them and a new set of solutions about the same themes of random and systematic errors. And then we'll talk about a particular set of interpretation issues that lead to additional considerations. And we'll end on time series data which will be a theme that will connect this talk to much later talks and systems biology where the ability to connect with a time series will help establish causality and connectivity. And we'll tie it to the subject of RNA analysis by looking at messenger RNA to k.

Now slide three is a reminder that we'll use in two different contexts tonight. First, these are the bell curves that you've seen, at least three of them integrated before-- the two discrete binomial Poisson and the normal, which is symmetric around the mean of 20, in this case. And just to connect to the discussion from last time where we were asking what the significance of a match of a single sequence to a database might be, when you're asking for a match of a single-sequence database, you're typically really asking for the maximum match, or the maximum-- the most extreme matches.

And so when you talk about extremes, when you're sampling from a distribution, and you're looking for the most extreme value for finite sampling of that distribution, that tends to be not from the normal distribution, which would be random sampling, but from-- which is this middle magenta curve. But instead, it would be extreme value distribution, which is this blue curve, which you can see in this case, since we're looking for extreme maximum, it's shifted slightly to the right. And so you can see that it comes inside of the other bell curves, on the left-hand side, and goes outside, on the right-hand side. If we were looking for extremely low values then it would've been shifted to the left. And remember, all these continuous functions go off to negative infinity and positive infinity, although at extremely low levels.

And then that's the extreme value distribution. Now in order to connect this nucleotide sequence, which we've been seeing has these wonderful Watson Crick base pairs that were TNA, and so forth, is to this more complicated tertiary structure. We're going to go through an intermediate of secondary structure where we really look at whether what kind of base pairs can form. And I'm going to immediately introduce some complexities so you don't get too complacent right off the bat. Somewhere in this slide of non-Watson-Crick base pairs is a Watson-Crick base pair. Take a moment to find it. Fine.

OK. So since we haven't introduced base pairs, it's going to-- well, we've seen it twice now in double-stranded DNA. Right in the middle here are labeled A is an AU base pair, which is Watson-Crick where the black dot indicates the attachment to the ribose in RNA, or to the oxyribose in DNA. And you'll find three other AT base pairs. One right to the right of it and two down below.

And these four AT base pairs are all different from one another, most easily imagined in terms of the orientation of the riboses, these black dots, relative to one another. And they each have names. But the important thing is that all of these are illustrated such that they maintain the coplanarity of the bases. They typically maintain one, or two, or even three hydrogen bonds. And the planarity allows them to stack on base pairs below and above them, just as you would in normal double-stranded DNA.

But sometimes, the geometry distorts the double helix enough that you might get a penalty in the free energy in the thermodynamic sense, or the kinetics, in a sense. Now here's another. So you can find almost-- you can basically make one of these base pairs for all of the possible combinations of ACG and U in RNA. And all of them will be coplanar, and they will have one or more hydrogen bonds. Probably the most stable and most commonly encountered in otherwise normal RNA double helices is a GU base pair. And you can see that this has fairly similar geometry to the AU base pair, or for that matter, the GC base pair.

So let's see how these non Watson-Crick base pairs appear. And this is the transfer RNA that we saw a couple of slides ago, spinning around in three dimensions. And the sequence that was behind it was the DNA sequence that corresponds to this unmodified RNA sequence on the right-hand side.

What we have-- you can think of this as four fairly canonical Watson-Crick type double helices for RNA double helices, which are slightly different from the DNA double helices. But in this-- so we have seven base pairs, six of which are Watson-Crick in the top stem loop, starting at position number one with the 5-prime end and ending position 72. So one and 72 is a GC base pair.

And you can see that the anticodon where it meets the messenger RNA is at the bottom right-hand part of the slide. And if you just look at that loop and-- you've got a seven-base loop and a five-base pair anticodon stem. And so each of these-- so that each of these stems has some distinguishing features, a number of base pairs ranges from four to seven. You've got a GU base pair in the middle of the top stem. And you've got little sequence boxes which are fairly conserved, such as this T psi CG The-- is-- actually, in its original form is a UU CG. And the T and the psi, or pseudouridine, are examples of modified bases, which are shown on the left side of the slide.

You can see there's quite a number of them. You can add a methyl group of ch3 groups to either of the bases, such as one methyl group on G. Or you can add them to the riboses, such as the two prime methyl groups, which can go on any of the four bases because it's modifying the ribose which is generic. Most of the other ones are very specific to a base. So for example, dihydrouridine is a modification that can only occur to uridines. The pseudouridine-- similar thing. And so on.

So each of these requires an enzyme. And we will highlight one of the enzymes that's involved in putting the methyl groups onto the sugars, the [? 0,2 ?] [? trimethyl ?] in just a few slides. But right now, what we want to ask is how did we get this folding structure. Now this is not the three-dimensional structure. This is the intermediate, between the primary DNA sequence and the fully modified, fully folded three-dimensional structure that we saw spinning around a couple slides ago.

So the first thing is the way-- you can try folding this, for you oppose each base pair in turn to look for possible matches. And then, what we've done historically-- this is in the mid to late '60s-- was you take each new transfer RNA sequence and ask whether it makes a decent fold in this simple planar representation that's related to the previous ones, under the assumption, the hope that there would be some conservation, not only of sequence of some of these motifs like the PSI CG, but also the way that it folds up.

And you might even hypothesize that maybe it doesn't matter what the sequence is in some of these stems. What's important is that it's capable of forming a stem, that it is that position one is complementary to position 72. If position one were to change from a G to an A, then position 72 should change from a C to a U.

So how do we formalize this? How do we formalize the process by which we generate this so-called cloverleaf structural or any similar folding pattern for small nucleic acids. And what are the limitations of those algorithms. And these dotted lines, you see, are some of the non Watson-Crick based pairs. Some of them will stack. Many of them-- some of them will actually form the hydrogen bonds of the Watson-Crick based pair but they won't otherwise have the rest of the geometry.

And you can see that some of these will provide connections between two loops which are separated by stems. And this kind of folding back means that it's not a simple set of helices. So the way that we formalize this is we say that position number one, if it's bound to position number 72, and the exact sequence maybe isn't as important as their ability to pair well another, then you expect if you take a large number of transfer RNAs and do a multi-sequence alignment, as we did last time, then in that multi-sequence alignment, you expect that when the G changes, the C will change, too. And that's called covariance. And if you look at the vertical axis, the maximum that can be achieved is the same kind of maximum that we had in the motifs last lecture.

The motifs, we've actually had a couple of times. It can get up to two bits. Two bits is the full scale for a base pair, or base, which can have four different values, A, C, G, or T. And we're calling this mutual information if it has the same units, a full scale of zero to two bits. And so we see along the horizontal axis, here-- call them positions I and J, which range from one to 72, which is the core part of the transfer RNA. The last four bases are added by a specialized enzyme. But position number one and position 72 covariates [? assume ?] this peak in the far left-hand region. In effect, there's seven peaks in a row there which correspond to the seven stacked base pairs they covary.

Similarly, in the TC to UC stem that we talked about a couple of times there, that those five nucleotides covary as you would get in a stem. The anti-codon stem is another five. And that D stem, so named after the hydrourodine modifications is four base pairs.

And the way that this is derived-- and we're going to work through this, an example, in the next slide-- but just as a labeling of this axis here, the mutual information between the Ith base and the jth base, that is to see, for example, between I equals 1 and J equals 72, is simply the sum of the frequency of getting that particular I, N, J-- F is the frequency of getting, say, a G as position one and a C at position 72-- times the log base 2. Remember, when we're talking about information content of nuclear-- or of information in general, bits in a computer or nucleotides in a sequence, we do log base 2, and-- as introduced by Shannon and others.

So now it's going to be the log of that same frequency. So the frequency of getting that particular I and that particular J type of base normalized now to how frequently those two bases occur throughout those positions. In other words, you know how often they co-occur at those two positions.

Now, how often do they occur independently of one another? And that's what the denominator is here. So when you take this ratio, you put it on a conventional scale, and then you have something that's analogous to the P law P of information theory. And you sum over all of the observed bases at positions I and J.

And that that's for a particular I and J, you sum over all the X's that occur at position, say, one and 72. And then you repeat that. You can get this Nth of IJ for every matrix element going from one to 72 in a symmetric square matrix. So let's work through this for two extreme examples. The extreme case where have perfect covariance and the extreme case where you have no real association.

So we're going to illustrate this with a toy multi-sequence alignment here. This is just the same way we did the multi-sequence alignment in the last class. Here there are no insertions or deletions, but it's the same thing. You could derive a weight matrix for this. And you would see that the first column-- the far left-hand column, I equals one, has all four possibilities, and the so does the rightmost column of that multi-sequence alignment J equals six.

And so let's calculate, are these covarying in this simple multi-sequence alignment of four [INAUDIBLE]. So we calculate the mutual information for I equals one, J equals six Nth of one sixth. It's going to be equal to sum. The first term in the sum is for the AU. And then we're going to walk through the CG, GC, UA. So there will be four terms in the sum. Each of the terms will have, coincidentally in this case, have the same frequency for that particular pairing of AU. And remember this is not a base pair. This is a covariant pair of nucleotides that could have been anywhere in the sequence. We happen to pick the first and the last base.

So they all have the same frequency. That frequency is 1/4. So the AU occurs one quarter of the for sequences in the multi-sequence alignment, so that's one quarter. And then, remember it's the same frequency inside the logarithm, but now in the denominator, we're going to normalize it to the frequency that the A occurs in the I equals one position, which is one quarter, and the frequency that U occurs in the J equals six position, which is one quarter. So that's one quarter over one quarter squared, or four. So 0.25 times log base to a of four is going to be two. And 0.25 times two is going to be 0.5.

And that's the first term. That's for the AU pairing. If you go down through all four terms, they all end up being the same form. The frequency is always going to be 0.25 for the pair and 0.25 for each of the individual bases. So you end up with four of those, four examples of those for each of the four cases. And so four times 0.5 is two. So that's consistent, hopefully, with what you would have expected for perfect covariance. You're getting the full information content, the full range of two bits, and so that's what we achieved.

So now, as a controller is just further gratification that we actually understand this, as we're working through the example of comparing I equals 1 with J equals 2, so the two far columns. And here, you're familiar with I of 1. J equals 2 is always C, and so it's not covarying with the first column as in the previous example. So let's just work through it the same way.

So the first term in the series is 0.25 again because the AC pair, not base pair, but pair of bases, occurs only once in the four of the multi-sequence alignment so that's 0.25. Then you have the logarithm base two of that same 0.25 now normalized to the frequency of the A in its column, the one column, which is 0.25. And the C in the J equals 2 column, which is-- it's always there as unity, so it's one.

So now that's the big change here, is instead of having 0.25 in both of the denominator terms, it's now 0.25 times one. It's now it's 0.25 over 0.25, so you have the log base 2 of one, which will be zero. And that zeros out the whole term. And so you have mutual information of zero, as you would expect from this particular toy example where columns one and two do not covary.

And so this is the same formulas that was in the previous slide. And a generalization of the one that we walked through term by term. And here's the reference for that. So now how do we go-- so that was-- we've taken now hundreds, possibly thousands, of transfer RNAs.

We've done a multi-sequence alignment. We've produced that mutual information pattern that we saw before with the one by 72-- one to 72 by one to 72 comparison where you got the spikes at each of the double helices. Now how do we turn that into more general practice. How do we generate secondary structures which are kind of at this intermediate in between the primary sequence and the three-dimensional structure using a particular class of experimental data combined with the sequence data.

This does not necessarily require the large set of line sequences, but it obviously benefits from it. You could do a secondary structure for each element in the aligned sequence in order to-- and use the mutual information, if you have it. But let's just talk about just the simple application of these thermodynamic parameters to the prediction of secondary structure. And what are our expectations before we go through the algorithm. How good is it?

In this fairly close to state-of-the-art paper, looking through over 700 generated structures, they have-- in each set, it contains one structure that, on average, has 86% of its known base pairs. That's not saying that it's necessarily identified as the top, the best, structure. It's saying that it has one structure by the criterion that they're using. This as a weak self praise.

But let's walk through how that works. When someone says that they're going to predict a secondary structure or a three-dimensional structure from a primary sequence, more or less, from scratch, they really typically mean that there's going to be a variety of other chemical data that they take into account, but it will be generic data. It will not be specific chemical data for this particular molecule. And the generic data, in this case, are measurements of the thermodynamics of melting of model oligonucleotides, usually large amounts of them, monitored spectral photometrically.

And from the temperatures of melting, basically, at equilibrium where you're getting half melted structures, you can determine the free energies where the negative free energies are the desirable ones, the ones that are likely to happen if you let the system go to equilibrium. And this is a kind of interesting application of the free energies for nucleic acids. Here, the algorithm that one uses is concerned mostly with adjacent base pairs of base pairs. So it's not a base pair, as you might think that the hydrogen bonds that determine the Watson-Crick and non Watson-Crick base pairs would dominate.

Instead, it's the stacking interactions that dominate. And since it's the stacking interactions that dominate, the hydrogen bonds are basically exchanging a water hydrogen bond for a base pair hydrogen bond. It looks very specific but in terms of free energy, it's fairly weak.

The free energy is determined more by the stacking of pi orbitals as depending on the geometry that you get, say, when you have a GC-- a CG base on top of an AU base pair here at the bottom of this helix. And that stack, it gives you a -2.1 kilocalories per mole. All the units on this are kilocalories per mole. And by going along and taking each of these stacks a pair of base pairs is what you're measuring. You can get all the negative free energy so they're stacking.

Then you have some penalties, some things that are less favorable, that would not happen spontaneously if they did not have these mitigating negative free energies already accumulated, which would be the loop and the bulge here. The base pairs on either side of that bulge will stack up on one another. And that bulge will kind of flip out of the, otherwise, regular double helix. Similarly, bases at the end have a slight penalty. And so then you can add it all up and you can calculate it overall delta G for the entire structure. And if you do enough of these things, you can get a feeling for which ones are likely to be occurring in your RNAs.

Now this should trigger in your mind, as the third example we've had where the conceit of a motif analysis, that you can do a multi-sequence alignment and each column and multi-sequence alignment is independent. This is the third example where that's not true. We have here the three energies are dependent on pairs of base pairs.

The previous examples were the very distant connections that you can get in folding up a transfer RNA. And the earlier example was the CG dinucleotides. The assumption of independence of columns in a multi-sequence alignment is a very powerful one. I don't want to undermine it too much. But it doesn't hurt you to have three examples this early on in the course.

Question the independence of columns in multi-sequence alignments-- very important thing to question. We've got mutual information theory that we had a couple of slides ago as one of the most powerful ways of questioning that when you see it. Now, that's the way that this particular base pairing, that we see here, is one example-- now you could take each of these and shift the right-hand half of the molecule relative to the left-hand by one base pair. That would give you a much poorer set of energies and much, much more bulges and more-- longer loops and so forth. And you end up with a poor delta G.

And what you can do is you can rank and do one of these maximum-value searches by going through that. This should trigger in your mind this is another way of thinking about that search. You take the entire sequence, whether it's transfer RNA, or in this case, a 400 nucleic acid sequence, and you draw lines between every base where you have a favorable free energy. And you look for a set of lines which do not overlap one another, because these would represent short sequences of local-- you can think of this as a local sequence alignment between one half of the nucleic acid and the other half.

Now this is not a sequence identity, remember, this is a sequence complementarity. That is to say, a reverse complement where complement means you've substituted As for Us and Cs for Gs. So you're looking for-- but in many other ways, this is analogous to the dynamic programming where we took two independent sequences and slid them along one another and allowed for insertions and deletions. In that-- in the dynamic equilibrium before, we did that formally, all possible such slippages by setting them as the two axes of a table and then filled in the squares for all the matches. Here, we would fill in the squares, not for the matches, but for the free energy of the stacking for these short subsequences.

Now the reason that they don't cross over, and the reason for this little note in the lower left-hand corner of slide 11, that does not handle pseudo knots. Psuedo knots we'll explain in the next-- we'll show graphic examples in the next slide. But it basically means that if you allow such sequences to occur willy-nilly throughout the sequence, then you'll get these tangles that for a while people weren't sure whether they occurred or not. The one or two non Watson-Crick base pairs that you might find connecting up tRNA in these tangles were not considered long stretches that would connect loops. But since then, they've been proven to be of great biological significance.

In any case, to do this without the pseudo knots, without allowing any crosses is still a challenging problem. It's basically the dynamic programming where N is the length of the primary sequence then takes on the order of N squared and compute time and space in order to figure out all the possible pairings that can occur. And then you go through and you rank them, which one gives the best free energy, and then you do the trace back and you get the top scores for that molecule.

Now let's talk about pseudo notes. We excluded that, but now we'll reinvite it. We had those little ones, a couple of base pairs in the transfer RNA.

But a much more dramatic one we alluded to in the second lecture. We talked about the genetic code. And in order to introduce you to exceptions to the genetic code, I gave an example where the ribosome jumped over 50 base pairs if presented with the right context. It didn't follow the normal code of having a triplet and another triplet right in a row with no punctuation. Here we had the punctuation that required, what may have slipped by at the time, a pseudo knot. And this is an example of once such pseudo knot in the best that we can do of a two dimensional schematic. And then something slightly better where we have a more three-dimensional and another three-dimensional view of this.

And this is the RNA pseudo knot, which is-- one of them, which is responsible for frame shifting in the-- that breaks this genetic code. And so let's just follow how this goes. You've got basically a normal helix here at the bottom starting at the five prime in position one to seven.

It would go through a normal five-base-- sorry-- six-base loop, from eight to 13, and then finish the stem 14 through 18. That would be a normal stem loop where the loop is six long, eight to 13. But at the end of eighteen, you have this little green loop that goes back and now makes a nice perfect four base Watson-Crick stem.

Now in the middle of what would have been a loop-- and so this fold back is what we meant by a pseudo knot, and what would have been represented by a crossover of those red lines in the previous slide-- something that makes it much harder to compute. In fact, it makes it so much harder, in the next slide, that it goes up from an order of N squared, which is your typical dynamic programming in a pairwise alignment to order in sixth in CPU time and order into the fourth power in memory space that you need to set aside for storing up the table of possible pseudo knots that can occur in the context of the otherwise normal circle with the non-overlapping connections.

This is a relatively recent innovation where a dynamic-- it's still a dynamic programming algorithm, it just has more possibilities, more complex-- higher algorithmic complexity. And the combination of the biological discovery to pseudo knots are important for frame shifting a variety of other biological phenomenon. And the now three-dimensional structure and now an algorithm puts pseudo knots well within the sort of things that you should feel comfortable. Now we're going to go back to hidden Markov models in a slightly more complicated context here, that we take the simplest one we could, which was a dinucleotide, simple, unfolded straight DNA.

And the part that was hidden, as you will recall, was whether the the CG dinucleotides-- or sorry-- the dinucleotides, which could be any of the possible dinucleotides, including AA, CG, and so on, whether it was present in a CG island, or whether it was in a region of the chromosome which was likely to have CGs, or whether it was in a CG ocean which was low in CG dinucleotide content. So the hidden part was the plus-minus whether it was in an island or not. Now what we're going to do now is take this and transfer this over to the kinds of motifs we are finding in RNAs, like transfer RNA and another class of RNA, and say OK, now, whether it's the hidden part of the Markov model is these transition probabilities. The hidden part is whether it's in a particular secondary structure or not, not whether it's in an island or not, but a secondary structure.

In this particular case we're going to talk about is a very interesting biological illustration where the hidden Markov models will be modeling these boxes, these motifs, that are involved in base pairing or recognition that forms the secondary structure of-- if it's necessary for guiding a particular enzyme. Now, remember we had all these modified bases that we used, that we saw in transfer RNA. Some of those are simple protein interactions with the transfer RNA that adds a methyl group here or there.

It turns out that all of the methyl groups, the O2 prime methyl groups, these are on the sugar of the ribose. In ribosomal RNA, a few, just a small number of a few dozen of the ribosomes in this multikilobase ribosomal RNA are methylated O2 position. How does the enzyme know, or the enzymes, know to get exactly those bases?

The way it knows is it doesn't use pure protein brute force to make a complimentary surface of protein nucleic acid. It actually uses this elegance of base pairing to make a guide sequence. And so what it's looking for is-- the protein cooperates with a small RNA, so-called snow RNA, or small nucleolar RNA, to find a place where the snow will recognize the place that you want to methylate. And then the protein methylates the base in the middle of that guide sequence. So then the game, the computational biology game, that these authors played was how can we find all the small RNAs, the snow RNAs, present in a genome when we very little about that the genome?

What they knew was they knew the genome sequence. This is for yeast. They had a few examples of snow RNAs in humans-- almost none in yeast. They had the subsequence of the ribosomal RNA, of course.

And they could-- what they wanted to do then is ask where in the genome do we have little guide sequences flanked by some of these other motifs and characteristics like a base pair, 4-8 base pair stem, that will match the ribosomal RNA. So you basically march along the algorithms, you march along the ribosomal RNA looking for matches elsewhere in the genome. And then ask whether those matches elsewhere in the genome have some of these other contexts, features.

You can see this is going to be a more complicated algorithm than just looking for CGs. So this is how it works. That stem that we had is now item number one. The various boxes which were basically sequences are now turned into ungapped hidden Markov models. The hidden part of it is whether it is present or not in the context that adds up to this guide sequence. The guide sequence itself is a hidden Markov model which has to be an imperfect, probably imperfect, duplex with the ribosomal RNA. So that's how that's modeled.

The most complicated is that terminal stem number one, which is a so-called Stochastic Context-Free Grammar. That's what the SCFG stands for. And that just means that it is even less constrained than the HMM. The HMM is less constrained than a simple motif, which is less constrained than, say, a consensus sequence. It is constrained, it has the grammar, if you will, or the particular rules for the base pairing that have to occur over a certain region in a certain part of this putative snow RNA.

So anyway, you apply each of these criteria and you have transition probabilities which come in from a learning set, such as the human snow RNAs. You have a learning set that tells you what these transition probabilities will be. And you and you now apply this to the entire yeast genome, and you get a bunch of candidates, snow RNA encoding genes.

Now you can't use things like the long open reading frames that you normally would use for finding genes. So this is a very valuable tool. But now how do you convince yourself that this is a gene, that this actually encodes a snow RNA, and that those are responsible for guiding the methylation at particular positions in the ribosomal RNA. The way you do that is-- well, before we get to how you do that, we want to ask how does this algorithm perform relative to other-- the few other algorithms there are for finding genes which do not encode proteins.

And the first of these actually dates well before 1991. But there were ways of looking for transfer RNAs in sequence. They would use everything we know about transfer RNAs-- the little boxes that are conserved as sequences, the regions that are conserved only-- not as sequences, but as base pairing potential, et cetera.

The loop lengths look-- are constrained. All the constraints that you can muster back in '91 were applied. And it was fairly slow. It would only do 400 base pairs of genome chunk per second. And when you have genomes on the order of many mega bases, this is slow. And it had-- it missed about 5% of the true positives. It had 95%.

And false positives sounds impressive-- only 10 to the minus 6. But when you think of a double stranded-- both strands of E. coli being about 10 million bases, then this is about four false positives. And bigger genomes, of course, would be even a larger number in an absolute scale.

So then, six years later, the speed is now 100 times faster. You're now only missing 0.5% of the true positives instead of 5%. And the false positives is now vanishingly small. So very often you can just arbitrarily trade off the number of true positives you miss with the number of false positives you get and make one-- take one advantage of the other. But here, it was a win-win situation. They both went in a favorable direction.

So how do the snow RNAs compare with that? Here, another two years passed. We have the snow RNAs are just starting out. They have, probably, a little better than 93% true positives. This is not as good as transfer RNAs. This may be-- this may improve or it may not. The false positive rate is acceptable. So then the question becomes, how do you track down-- after you track down these genes, how do you then prove that they do what you think they do, that they actually are responsible for methylating the riboses or the bases in question?

So it turns out that the technology we set up in the sequencing and genotyping lecture where you extend with DNA polymerase a primer on a template so the primer binds to the template and you extend by either many base pairs, as in the conventional dideoxy sequencing, or one or two base pairs in some of the more up-and-coming genotyping methods. Those extension methods, those DNA polymerase-based extension methods will stall when you run into this particular kind of modified base where a bulky group is introduced onto the two prime position of the ribose on the template. So you're extending the primer, sitting on the template, and it will stall here.

And it stalls more when you decrease the concentration of the deoxynucleotide triphosphates in the extension reaction. So that's what these little wedges mean at the top of each of these columns. They've done an extension with all four triphosphates present, either in high amounts at the big end of the wedge or low amounts in the small ends of the wedge. And to tell where you are in the sequence-- this is what using reverse transcriptase polymerase on a ribosomal RNA template-- to find out where you are, you do this dideoxy, which is basically the conventional DNA sequencing. Where you terminate, it either is used Us, Gs, Cs, or As in the template.

This allows you to get oriented. And basically, your sequencing on the far left-hand set of lanes. And these pause sites our present, say, the wild-type is the first pair of lanes next to the sequence lanes on the left-hand side of this display. And you can see there's a pause at every single known methylated base. You can determine methylated bases by other methods as well.

But so now then, the computational biology predicted a set of snow genes. In fact, ultimately, all of the snow genes in the yeast genome we think, explaining all of the methyl groups, at least. And one by one, these were knocked out cleanly so that there's no gene there anymore for the small RNA.

And then you ask, well, how does this affect the methylation as detected by this extension assay? And if you look on the far right-hand side for deleted number 40, and you can see that position number 596, near the bottom, circled in red, which is present in the wild type and all the other mutants, is absent from that particular mutant number 40. So there's no pause there. One infers there's no methylation.

And that was the specific site that that snow RNA guide sequence was predicted to bind. It's aligned with the position in the guide where you expect there to be a methylation occurring. And you can see in each lane there's a different circled red missing black pause site. And until we get to the one in the middle, Newton numbers for snow RNA number 60-- and here's actually two missing bands in the same lane. And how that can occur, there's two different ways that a snow RNA can-- knocking out a single gene, a single snow RNA can have an effect on two different methyl groups.

One is if the guide sequence can bind to two different places in the ribosomal RNA. And the other is if there are two guides sequences within the same snow RNA. So now that we have at least some grounding in the kind of structures that can occur, now we're going to ask how we monitor and measure the amounts of these structures in biological systems. And we will also see how these structures impact the methods that we use for the quantitation of the structures. So we have choices of molecule that we're going to measure when we're monitoring the various molecules in the cell.

Why are we focusing on RNA? Well, part of it is because of its nice structural continuity between the simple DNA and the very complicated proteins. But the other is that if we want to study different points in the regulatory and metabolic networks that we'll be talking about at the end of this course having to do with systems biology, if we choose-- we want-- every part of it is subject to some kind of control. Transcriptional control is one of the early stages. And then there are many stages subsequence to that lead to the protein and ultimate phenotypes that result in proliferation of the species.

If you want to look at transcriptional control, it would not do well to study protein because the closest thing to transcriptional control that you can measure is the RNA products. You can study the transcriptional control itself directly, as well, by studying the DNA protein interactions. But if you want to measure, a defusible molecule RNA is the thing to do. And there are multiple different methods for getting at a co-regulated set of genes, co-regulated at the transcriptional level.

We'll illustrate a few here. And why do we need multiple methods? Well, we've talked about random and systematic errors. Random errors you can compensate by repeating the experiment. The random errors will average out, ultimately. Systematic errors will happen the same way over and over again.

So you want to have something out of the box to allow you to check it, or to model it, or to allow you to do integration, as you might want to have in complicated systems. So here, just to start us thinking about integration and checking different ways of getting transcriptional co-regulation, let's think about-- if you look through all the proteins that occur, you will find proteins that occur together, frequently, as either as fusions or as separate proteins. Where in operands, they'll occur as coding regions that are clustered together in some species or maybe less clustered in other species.

When we have metabolic pathways where a small molecule will be shared by-- as the product of one will be the substrate of another, and so on, you'll have this chain of events as in the lower left-hand corner. And these sets of enzymes that need to be working together, need to be co-expressed. They need to come up together and go down together when they're not needed.

They need to come up when they're suddenly needed. And so, you might have an entire pathway, or a set of pathways, that are co-expressed. And one way to do that is to cluster them in the genome. When they are co-expressed, you will sometimes find upstream of them, motifs, such as this.

Again, here's the two bits for the vertical scale, where this might be enriched. And so this would be another indication-- so when you find these in front of genes, you might expect them to be co-regulated. When you find a set of proteins that are consistently together in different organisms, so-called phylogenetic profiles, you will find that this set of proteins that is involved in a common enzymatic pathway, metabolic pathway, are not only co-regulated and found together but in-- along the chromosome, but they're found together when you go through many different species. They will be deleted or inserted as a block, or they'll be found scattered around the genome. But you'll find that when one disappears, they all disappear, in general, statistically speaking.

It's this phylogenetic co-occurrence is another clue that you might expect them to be co-regulated in those genomes in which there they do co-occur. Anyway, and microarrays will be-- and variations on that theme will be the main thing that we'll talk about. But I wanted to put it in the context. And I'll just expand on one of these at the bottom here in slide 22. This is an algorithm for reconstructing likely combinations where in some organisms you might have the entire biosynthetic pathway as a series of genes which encode one by one, all the proteins in this case that are involved in purine biosynthesis from simpler molecules.

But in other organisms, you might have them scattered all over the genome, but they might be co-regulated. Their RNAs might go up and down together. And so, if you look at enough genomes, you can reconstruct the likely combination of enzymes. And here's how it might work.

In any one of these, for example, E. coli, you might see that they're scattered about-- a pair here, and a pair there. Singletons don't help much. But if you take all the pairs from a lot of different organisms, you can reconstruct this network where you say, oh, this gene will call L, Q, Y, C-- all these are probably involved in the same process. If you get a hint for what any one of them does, say, one of them is involved in purine biosynthesis, so then you find that they all are. And if you guess they might be co-regulated very tightly.

So now let's figure out how we actually measure that they're co-regulated very tightly. And the way we can do that-- as we do that, whatever method we use, we want to ask are we interested in ratios, relative changes, or are we interested in absolute values? There are various things that we can do with absolute amounts that are very hard to do with ratios. In particular, if we want to ask, is a particular protein level high because its translation is efficient or is it high because its transcription is efficient, if you find that it's full of abundant codons as if it wants to be efficiently translated, is it also have a high-level promoter as if it wants to be transcriptionally active?

These sorts of questions really benefit from having absolute amounts, meaning so many molecules of RNA per cell, so many molecules of protein per cell. But we get that to direct causality, we want to get at the motifs. This would be one of the objectives of doing the RNA quantitation, to allow us to cluster RNAs that are co-expressed, and then to start looking for motifs and direct causality. Another thing that we might want to do is to classify. We can ask whether small molecules or mutations, such as occur in cancers, cause enough of a signature that you can then use as to say, OK, this self state that we see is a recognizable small molecule effect, or stress effect, or mutational effect, cancer.

Now, when we-- we will be talking about microarray and related methods, but I want you to question the advantages and disadvantages of these methods. And so I'll compare it to a number-- but let's start with the most dramatic comparison, which is with in situ hybridization. So in array hybridization you'll have tens of thousands of different gene probes mobilized on a solid surface. And you'll label up the RNA from a mixture of different cells-- different mixture of different RNAs within a cell. But you'll be able to ask questions about 10,000 genes at a time.

In an in situ experiment, it's the other way around. You take a cell in it's fairly natural environment, usually fixed, but fixed with maintaining the spatial aspects. Then, if you look within the cell, with a single gene at a time, or maybe two or three at a time, a very small number, not tens of thousands, you can look to see whether the RNA is uniformly spread throughout the cell and uniformly spread throughout all the cells in the tissue, or in, say, you've got a mixed population of yeast cells, whatever. And you can find cases in the literature where it is not uniformly present in all the cells and not even uniformly within a cell.

Here is one of the more dramatic cases where the two X chromosomes in mammals behave differently from one another. The female mammals will have one RNA-- one chromosome expressing most of its RNAs at normal levels, and the other chromosome expressing almost no RNAs. It is expressing at least one RNA, and that RNA is-- which is XIST and it's covering that whole chromosome or is localized over that chromosome and not the rest of the cell. So this is an extreme case of localization that you can monitor with microscopic methods, fluorescent microscopic methods.

Instead, we'll-- keep this in the back of your mind as you look through the microarray and other experiments where you're mushing together a variety of cells that might be in different stages of the cell cycle, might have slightly different environments, and even within the cell, the RNA-- you're losing the information about the RNA localization. Let's take a short break and then come back and connect on-- finish up in situ hybridization, and connect to other methods for quantitation of RNA.