PROFESSOR: Why don't we get started? So today we're going to talk about comparative genomics. And first, a brief review of what we did last time. So last time we talked about global alignment of protein sequences, including the Needleman-Wunsch and Smith-Waterman algorithms. And we talked about gap penalties a little bit and started to introduce the PAM series of matrices which are well described in the text.

So what I wanted to do is just briefly go over what I started to talk about at the end, about Markov models of evolution. Because they're relevant, not only for the PAM series, but also for some other topics in the course. A short unit on molecular evolution we're going to do today. And then they also introduce hidden Markov models that will come up later in the course.

So the example that we gave of a Markov model was DNA sequence evolution in successive generations where the observation here is that the base at a particular position at generation n+1 here depends on the base at that generation and the base at generation n. But conditional on knowing the base at generation n, you don't learn anything from knowing what that base was at generation n-1. That's the essence of the Markov properties. So here's the formal definition, as we saw before.

Any questions on this? And I asked you to review your conditional probability if it was rusty, because that's very relevant.

OK so in this example you might, if you had a random variable x that represented the genotype at a particular locus, let's say the apolipoprotein locus, and it had alleles A and a, then you might write something like the probability that Bart's genotype is a homozygous given his grandfather's genotype and his dad's genotype is equal to just the conditional probability given his father's genotype. So those are

the sorts of things that you can do with Markov chains

So when you're working with Markov chains matrices are extremely useful. So another thing that will be helpful in this part of the course and then again in Professor Fraenkel's part, where he's talking-- he'll use also some ideas from linear algebra-- is to review your basics of matrices and vector multiplication.

OK so, if you now make a model of molecular evolution where sn is-- so s is this variable that represents a particular base in the genome and is the generation. And then to describe the evolution of this base over time, we're going to imagine that its evolution is described by a Markov chain.

And a Markov chain can be described by, in this case, a 4 by 4 matrix, since there are four possible nucleotides at generation i, for example, and four possible at generation i plus one. And you simply need to specify what the conditional probability that the base will be, of any possible base, at the next generation, given what it is at the current generation.

So here's the matrix up here. And it describes, for example, the probability of going from a c to an a. So then in general you might know that that base is a g at the first generation. But in general you won't necessarily know what base it is if you're modeling events that may happen in the future. So the most general way of describing what's happening at that base is a vector of probabilities of the four possible bases-- so qa, qc, qg, qt, with those probabilities summing up to 1.

And so then it turns out that with this notation that the content of the vector at generation n plus 1 is equal to simply the vector at generation n multiplied on the right by the matrix, just using the standard vector matrix multiplication. So for example, if we have vectors with four things in them, and we have a 4 by 4 matrix, then to get this term here in this vector you multiply-- you basically take the dot product of this vector times this first column.

The vector times the first column will give you that entry. And this times this column will give you that entry in the vector, and so forth. And you can see that the way this

makes sense, the way the matrix is defined, that first column tells you the probability that you'll have an a at the next generation, conditional on each of the four bases at the previous generation. And so you just multiply by the probabilities of those four bases times the appropriate conditional probability here. And those are all the ways that you can be an a generation, n plus 1.

And so it's also true that if you want to go further in time, so from generation n to generation n plus k-- k is some integer-- then this just corresponds to sequential multiplication by the matrix k-- I'm sorry, by the matrix p. So qn plus 1 equals q times p. And then qn plus 2 will equal q-- I'm sorry. That's a really bad q, but-- qn plus 1 times p, which will equal q times p squared, where p squared means matrix multiplication, again using the standard rules of matrix multiplication that you can look up.

So one of the things you might think about here is what happens after a long time? If you start from some vector q-- for example, q is 0010. That is, it's 100% chance of g. What would happen if you run this matrix on that over a long period of time. And we'll come back to that question a little bit later.

So thinking about the Dayhoff matrices-- and again, I'm not going to go into detail here, because it's well described in the text. Dayhoff looked at these highly identical alignments, these 85% identical alignments, and calculated the mutability of each residue and these mutation probabilities for how often each residue changes into each other one and then scaled them so that on average the chance of mutating is 1% and then took these probabilities, these frequencies, of mutation m, a, b, divided by the frequency of the residue b, took the log, and then just multiplied by two just for scaling purposes, and came up with a-- and then rounded to the nearest integer, again for practical purposes.

And that's how she came up with her PAM 1 matrix. And then you can use matrix multiplication to derive all the successive PAM series. Just multiply the PAM1 matrix times itself to get the PAM2 and recalculate the scores.

So if you actually use PAM matrices in practice there are some issues. And these

are also well described in the text. And the fundamental problem seems to be that the way the proteins evolve over short periods of time and the way they evolve over long periods of time is somewhat different. And basically this model, this Markov model of evolution, is not quite right, that things don't-- what you see in a short periods of time-- it does not match long periods of time.

And why is that? A number of possible reasons. But keep in mind that in addition to proteins simply changing their amino acid sequence, other things can happen in evolution. You can have insertions and deletions that are not captured by this Markov model. And you can also have birth and death of proteins. A protein can evolve according to this model for millions of years. And then it can become unneeded, and just be lost, for example.

So real protein evolution is more complicated. And so about 20 years ago or so Henikoff and Henikoff decided to develop a new type of matrix. And the way they did it was to identify these things called blocks, which are regions of reasonably high similarity, but not as high as Dayhoff required.

So there were many more-- Dayhoff was working the '70s. They were working in the '90s. So there were many more proteins available. And they could identify, with confidence, basically a much larger data set, including more distantly related, but still confidently alignable, protein sequences. And they derived new parameters.

And in the end this matrix they came up with called BLOSUM62 seems to work well in a variety of contexts when comparing moderately distantly related proteins or quite distantly related proteins. If you're comparing very similar proteins it almost doesn't matter. Any reasonable matrix will probably give you the right answer. But when you're comparing the more distant ones, that's where it becomes challenging.

And so this is the BLOSUM62 matrix here. And you can see it's similar to the PAM matrices in that-- I think we showed PAM 250 last time-- in that you have a diagonal with all positive numbers. And it's also similar in that, for example, trytophan down here has a higher positive score than others. It's plus 9. And cysteine is also one of the higher ones. But those are less extreme.

And basically, maybe over short periods of evolutionary time, you don't change your cysteine. But over longer periods there is some rewiring of disulfide bonding, and so cysteines can change. Something like that may be going on.

So we've just talked about pairwise sequence alignments. But in practice you often have, especially these days you often have, many proteins though. So you want to align three or five or 10 different proteins together to find out which residues are most conserved, for example. And so basically the principles are similar to pairwise alignment.

But now you want to find alignments that bring the greatest number of single characters into register. So if you're aligning three proteins, you really want to have columns where all three are the same residue, or very similar residues. And you need to then define scoring systems, define gap penalties, and so forth.

This is also reasonably well described in the text. I just wanted to make one comment about the sort of computational complexity of multiple sequence alignment. So if you think about pairwise sequence alignment, say with Needleman-Wunsch or Smith-Waterman, with a sequence of length-- let's say you're aligning one protein of sequence length n to another of life n, what is the computational complexity of that calculation in using this big O notation that we've talked about? Let's just say standard gap penalties, linear gap penalties. Anyone? Or does it matter? Yeah, go ahead.

STUDENT: n squared.

PROFESSOR: It's n squared. So even though this has gaps, with local-- with ungapped it was also n squared, or n times n, So why is it that gaps don't make it worse? Or do they? Any thoughts on that?

STUDENT: You put a constant number of gaps in the sequence. So it's just stating the essence of the complexity should still be n squared.

PROFESSOR: You put a constant number of gaps? The-- I mean, yeah-- let's just hear a few

different comments. And then we'll try to summarize. Go ahead.

STUDENT: So we're still only filling out an n by n matrix at any given time.

PROFESSOR: You're still filling out an n by n matrix, right. There happen to be a few more things. The recursion is slightly more complicated. But there's a few more things you have to calculate to fill in each. But it's like three things, or four things. It's not-- so it doesn't grow with the size. So it's just still n squared, but with a larger constant. OK, good.

And then if you did affine gap penalty, remember where you had a gap opening penalty and a gap extension, what then? Does that make it worse? Or is it still n squared?

STUDENT: I think it's still n squared.

PROFESSOR: Why is that?

STUDENT: Computing the affine gap penalty is no more than o of n, right?

PROFESSOR: Yeah, basically with the affine you have to keep track of two things at each place. So yeah, it is. You're right. It's still n squared. It's just you got to keep track of two numbers in each place there. OK, good.

And so what about when we go to three proteins? So how would you generalize, let's say, the Needleman-Wunsch algorithm to align three proteins? Any ideas? What structure would you use, or what-- analogous to a matrix-- yeah, in the back.

STUDENT: Another way to do this would be have a 3D matrix.

PROFESSOR: OK, a 3D matrix, like a cube. And can everyone visualize that? So yeah, basically you could have a version of Needleman-Wunsch that was on a cube. And it started in the 0, 0, 0 corner and went down to the n, n, n corner, filling in in 3D. OK so what kind of computational complexity do you think that algorithm would have?

STUDENT: n cubed?

PROFESSOR: n cubed. Yeah, makes sense. There would be a similar number, a few operations to fill in each element in the cube. And there's n cubed. So the way that the problem grows with n is as n cubed.

And what about in general, if you have k sequences?

STUDENT: n to the k?

PROFESSOR: n to the k. So is this practical? With three proteins and modern computers you could do it. You could implement Needleman-Wunsch on a cube. But what about with 20 proteins? Is that practical?

So it's really not. So if proteins are 500 residues long and there's 500 to the 20th, right. It starts to explode. So that approach really only works in two dimensions and a little bit in three dimensions. And it becomes impractical. So you need to use a variety of shortcuts. And so this is, again, described pretty well in chapter six of the text.

And a commonly used-- if you're looking for a default multiple sequence aligner, CLUSTALW is a common one. There's a web interface if you just need to do one or two alignments. That works fine. You can also download a version called CLUSTALX and run it locally.

And it does a lot of things with pairwise alignments and then combining the pairwise alignments. It aligns the two closest things first and then brings in the next closest, and so forth. And it does a lot of tricks that are-- they're basically heuristics. They're things that usually work, give you a reasonable answer, but don't necessarily guarantee that you will find the optimal alignment if you were to do it on a 20 dimensional cube, for example. So they work reasonably well in practice. And then there's a variety of other algorithms.

OK, good. So that's a review of what we've mostly been talking about. And now I want to introduce a couple of new topics. So we're going to briefly talk a little bit more about Markov models of sequence evolution. And these are closely related to some classic evolutionary theory from Jukes-Cantor and Kimura. So we'll just briefly

mention that. And we'll talk a little bit about different types of selection that sequences can undergo-- so neutral, negative, and positive-- and how you might distinguish among those for protein coding sequences.

And this will basically serve as an intro into the main topic today, which is comparative genomics. And comparative genomics-- it's not really a field, exactly. It's more of an approach. But I wanted to give you some actual concrete examples of computational biology research, successful research that has led to various types of insights into gene regulation, in this case, mostly to emphasize that computational biology is not just a bag of tools. We've mostly been talking about tools. We introduced tools for local alignment and multiple alignment and statistics and so forth. But really it's a living, breathing field with active research.

And even using-- comparative genomics is one of my favorite areas within this field. Because it's very powerful. And you can often use very simple ideas. And simple algorithms can sometimes give you a really interesting biological result, if you have the right sequences and ask the question the right way. So I have posted a dozen of my favorite comparative genomics papers in a special section on the website.

Obviously I'm not asking you to read all of these. But I'm going to give you a few insights and approaches that were used in each of these papers here, just to give you a flavor of some of the things that you can do with comparative genomics, in the hopes that this might inspire some of your projects. So hopefully you're going to start thinking about finding teammates and thinking about projects. And this will hopefully help in that direction.

Of course, they don't have to be comparative genomics projects. You could do anything in computational biology or systems biology in this class. But that's just one area to start thinking about. Yeah, I'll also-- I'm sorry, I think I haven't posted this yet. But I will also post this review by Sabeti that has a good discussion of positive selection a little bit later. Again, not required.

All right, so let's go back to this question that I posed earlier. We have a Markov model of DNA sequence evolution. And we-- sn is the base at generation n. And

then what happens after a long time? If you take any vector-- q, to start with, might be a known base, for example-- and apply that matrix many times, what happens as n goes to infinity. And so it turns out that there's fairly classical theory here that gives us an answer. This is not all the theory that exists, but this describes the typical case.

So the theory says that if all of the elements in the matrix are greater than 0, and then of course all of the-- pij's, when you sum over j, they have to equal 1. That's just for it to be a well-defined Markov chain. Because you're going from i to j. And so from any base you have to go-- the probability of going to one of those four bases has to sum to 1.

And so if those conditions hold, then there is a unique vector r such that r equals r times p. And the limit of q times p to the n equals r, independent of what q was. So basically, wherever you were starting from-- you could have been starting from 100% g, or 50% a, 50% g, or 100% c-- you apply this matrix many, many times, you will eventually approach this vector r.

And the theory doesn't say what r is, exactly. But it says that r equals r times p. And that turns out to basically implicitly define what r is. That is, you can solve for r using that equation. And r, for this reason, because the matrix doesn't move r, r is called the stationary distribution. And it's often also called the limiting distribution, for obvious reasons. And if you want to read more, like where this theory comes from, here's a reasonable reference. So any questions about this theory?

All the elements in the matrix have to be strictly greater than 1-- I'm sorry, strictly greater than 0. Otherwise, really no conditions. All right, question? Yeah, go ahead.

STUDENT: Does the [INAUDIBLE] distribution ever change, based on the sequence, or are we assuming that it doesn't?

PROFESSOR: The theory says it only depends on p. It doesn't depend on q. So it depends on the model of how the changes happen, the conditional probability of what the base will be at the next generation given what it is at the current generation. It doesn't

depend where you start. q is what your starting point is, what base you're initially at. Does that make sense?

And this is obviously a very simplified case, where we're just modeling evolution of one base, and we're not thinking about whether the rates vary at different positions or within-- this is the simplest case. But it's important to understand the simplest case before you start to generalize that.

OK, so let's do some examples here. So here are some matrices. So it turns out the math is a lot easier if you limit yourself to a two-letter alphabet instead of four. So that's what I've done here. So let's look at these matrices and think about what they mean. So we have two-letter alphabet. R is purine. Y is pyrimidine.

These matrices describe the conditional probability that, at the next generation, you'll be, for example-- oops, here we go. That, for example, if you start at purine, that you'll remain purine at the next generation. That would be 1 minus P. And the probability that you'll change to pyrimidine is P. And the probability of pyrimidine will remain as a pyrimidine is 1 minus P.

So what is the stationary distribution of this matrix? OK, so if p is small, this describes a typical model, where most of the time you remain-- DNA replication and repair is faithful. You maintain the same base. But occasionally a mutation happens with probability p.

Anyone want to guess what the stationary distribution is or describe a strategy for finding it? Like what do we know about this distribution? Or imagine you start with a purine and then you apply this matrix many times to that vector that's 1 comma 0, what will happen? Yeah, Levi.

STUDENT: Probably 50-50 because any other that way you skew it it would be pushed towards the center because there's more [INAUDIBLE] the other.

PROFESSOR: OK, everyone get that? So Levi's comment was that it's probably 50-50. Because mutation probabilities are symmetrical. Purine-pyrimidine and pyrimidine-purine are the same. So if you were to start with say, lots of purine, then there will be more

mutation toward pyrimidine in a given generation.

So if you think about this is your population of R and that's your population of Y, then if this is bigger than that, you'll tend to push it more that way. And there will be less mutation coming this way, until they're equal. And then you'll have equal flux going both directions.

So that's a good way to think about it. And that's correct. Can you think of how would you show that? What's a way of solving for the stationary distribution? Anyone? So remember, we'll just get back one. The theory says that R equals RP. That's the key. R equals RP. So what is R?

Well we don't know R. So we let that be a general vector. So notice there's only one free parameter. Because the two components have to sum to 1. It's a frequency vector, so x and 1 minus x.

And we just multiply this times the matrix. So you take x comma 1 minus x. And you multiply it by this matrix. The matrix is 1 minus P P. I'm using too much space here. I'll just make it a little smaller-- P 1 minus P. And that's going to equal R. And so we'll get x times 1 minus P plus-- remember, it's dot product of this times this column, right? So x times 1 minus P plus 1 minus x times P.

That's the first component. And the second component will be xp plus 1 minus x times 1 minus p. OK, everyone got that? So now what do we do?

STUDENT: r.

PROFESSOR: What's that?

STUDENT: Make that equal to the initial r.

PROFESSOR: Yeah, make that equal to the initial r. So it's two equations and-- well, you really only need one equation here. Because we've already simplified it. In general there will be two equations. There will be one equation that says that the components of the vector sum to 1.

And there will be another equation coming from here. But we can just use either one, either term. So we know that the first component of a vector-- if this vector is equal to that vector, then the first components have to be equal, right? So x equals x times-- times what? Times 1 minus p, just combining these two.

And then plus what are all the-- I'm sorry, that's 1 minus p-- 1 minus p here. And then there's another term here, minus another p. And then there's a term that's just p. And so then what do you do? You just solve for x.

And I think when you work this out you'll get two p x equals p, so x equals 1/2. Right, everyone got that? OK, so yeah. So if x is 1/2, then the vector is 1/2 comma 1/2, which is the unbiased.

All right, what about this next matrix, right below-- 1 minus p 1 minus q. p and q are two positive numbers that are different. So now there's actually a different probability of mutating purine to pyrimidine and pyrimidine to purine. So Levi, can we apply your approach to see what the answer is?

STUDENT: Not exactly.

PROFESSOR: Not exactly? OK, yeah, it's not as obvious. It's not symmetrical anymore. But can anyone guess what the answer might be? Yeah, go ahead Diego.

STUDENT: It'll go either all the way to one side or depending on q and d.

PROFESSOR: All the way to one side or all the way to the other? So meaning it'll be all purine or all pyrimidine again.

STUDENT: Yeah, depending on which--

PROFESSOR: Which is bigger? OK, anyone else have an alternative theory? Yeah, go ahead. What was your name again?

STUDENT: Daniel.

PROFESSOR: Sorry, Daniel?

STUDENT: Daniel, yeah.

PROFESSOR: Daniel. OK, go ahead.

STUDENT: It'll reach some intermediate equilibrium once they balance each other out. And that would be exactly-- I'm not sure-- some ratio of q to p.

PROFESSOR: OK. How many people think that might happen? OK, some people. OK Daniel has maybe slightly more supporters. So let's see. So how are we going to solve this? How do we figure out what the stationary distribution is? You just use that same approach. So you can do-- you have x 1 minus x times that matrix, which is got the 1 minus p p q 1 minus q. OK, and so now you'll get x 1 minus p.

Anyway, go through the same operations. Solve for x. And you will get-- I think I put the answer on the slide here. You will get q over p plus q. So as Danny predicted, some ratio involving q's and p's. And does this make sense? Seeing what the answer is, can you rationalize why that's true?

STUDENT: It's like a kind of equilibrium. You have one mode of force play pushing one way and another different one in this case pushing the other.

PROFESSOR: Yeah, that's basically the same idea. And so they have to be in balance. So the one that has less, where the mutation rate is a lower, will end up being bigger, so that the amount that flows out will be the same as the amount that flows in. You can apply Levi's idea of thinking about how much flux is going in each way. So there's going to be some flux p in one direction, q in the other direction. And you want x times p to equal 1 minus x times q. And this is the value of that works.

OK, good? What about this guy down here? So this is a very special matrix called the identity matrix. And what kind of model of evolution is this?

STUDENT: There's no mutation.

PROFESSOR: There's no evolution. This is like a perfect replication repair system. The base never changes. So what's a stationary distribution?

STUDENT: It's all--

PROFESSOR: What's that?

STUDENT: It'll just stay where it is.

PROFESSOR: It'll stay where it is. That's right. So any vector is stationary for this matrix. Remember that the theory said there's a unique stationary distribution. This seems to be inconsistent. Why is it not inconsistent? Sally?

STUDENT: We defined all of the variables to be greater than 0. So when you have anything that's [INAUDIBLE] that is equal to 0.

PROFESSOR: Right, so a condition of the theorem is that all the entries be strictly greater than 0. And this is why. If you have 0s, in there then crazy things can happen. Wherever you start, that's where you end up with this matrix. So every vector is stationary.

And what about this crazy matrix over here, matrix q? What does it do? Joe.

STUDENT: It's going to swap them back and forth.

PROFESSOR: It swaps them back and forth. So this is like a hyper mutable organism that has such a high mutation rate that it always mutates every base to the other kind. It's never happy with its genome. It always wants to switch it, get something better. And so what can you say about the stationary distribution for this matrix? Jeff?

STUDENT: There isn't going to be one.

PROFESSOR: There isn't going to be one? Anyone else?

STUDENT: Well, actually, I guess 1, 1, like 0.5, 0.5.

PROFESSOR: 0.5, 0.5 would be stationary. Because you're--

STUDENT: But you won't converge to it.

PROFESSOR: But you won't converge to it. That's right. it's stationary, but not limiting. And again, the theory doesn't apply. Because there's some 0s in this matrix. But you can still

think about that. OK, everyone got that? All right, good.

OK so let's talk now about Jukes-Cantor. So Jukes-Cantor is very much a Markov model of DNA sequence evolution. And it simply has-- now we've got four bases. It's got probability alpha of mutating from each base to any other base. And so the overall mutation rate, or probability of substitution, at one generation is three alpha. Because from the base G there's an alpha probability mutate to A, an alpha probability to C, an alpha to T, so the three alpha.

And you can basically write a recursion that describes what's going on here. So if you start with a G at time 0, the probability of a G at time 1 is 1 minus 3 alpha. It's a probability that you didn't mutate. But then, at generation two, you have to consider two cases really.

First of all, if you didn't mutate, that's PG1. Then you have a 1 minus alpha probability of not mutating again, so remaining G. But you might have mutated. With probability 1 minus PG 1 you mutated. And then whatever you were-- might be a C-- you have an alpha probably of mutating back to G. Does that make sense? Everyone clear why there's a 3 in one place and only a 1 alpha in the other?

All right, so you can actually solve this recursion. And you get this expression here, P G of t equals 1/4 plus 3/4 E to the minus 4 alpha t. OK so what does that tell you about-- we know from our previous discussion what the stationary distribution of this Markov chain is going to be. What will it be? What's the stationary distribution?

STUDENT: 1/4 of each.

PROFESSOR: 1/4 of each. And why, Daniel, is that?

STUDENT: Because the probability of them moving to any base is the same?

PROFESSOR: Right, it's totally symmetrical. So that has to be the answer by symmetry. And you could solve it. You could use this same approach with defining a value-- the theory applies if alpha is greater than 0 and less than 1-- or less than-- I think it has to be less than a quarter, actually, or something like that.

And you can apply the theory. So there will be a stationary distribution. You can set up a vector. Now you have to have four terms in it and multiplication. And then you'll get a system of basically four equations and four unknowns. And you can solve that system using linear algebra and get the answer. And yeah, the answer will be 1/4, as you guessed.

And so what this Jukes-Cantor expression tells you is how quickly does it get to that equilibrium. We're thinking about G. You can start at 100% G. And it will then approach 1/4. You can see 1/4 is clearly what's going to happen in the limit. Because as t gets big that second term is going to 0. And so what does the distribution look like? How rapidly do you approach 1/4?

You approach it exponentially. So you start at 1 here. And this is 0. This is 1/4. You'll start here. And you'll go like that. You go rapidly at the beginning. And then you get just very gradual approach 1/4.

So you can do a little bit more algebra with this expression. And here's where the really useful part comes in. And you can show that K, which we'll define as the true number of substitutions that have occurred at this particular base that we're considering, is related to D, where D is the fraction of positions that differ when you just take say the parental sequence and the daughter sequence, the eventual sequence that you get to.

You just match those two. And you count up the differences. That's D. And then K is the actual number of substitutions that have occurred. And those are related by this equation, K equals minus 3/4, natural log, 1 minus 4/3 d.

So let's try to think about, first of all, what is the shape of that curve? What does that look like? Here's 0. I'll put 1 over here. So we all know that log-- if it was just simply log of something between 0 and 1, it would look like what-- look like that. Starts from negative infinity and comes up to 0 at 1. But it's actually not log of D. It's log of 1 minus D, or 1 minus a constant times D. So that will flip it. So the minus infinity will be there. It will come in like that.

And then we also have minus 3/4. There's a minus in front of this whole thing. So all these logs are of numbers that are less than 1. So they're all negative. But then it'll get flipped. So it'll actually look like that.

And it will go to infinity where? Where does this go to infinity? So if this is now K is on this axis. And yeah, sorry if that wasn't clear. D is here.

So this is just again, this is if we did log of D it would look like this. If we do log of 1 minus something times D, that'll flip it. And then if we do minus that, it'll flip it again that way. OK so now K, as a function of D, is going to look like this.

Sometimes people like to put-- anyway, but let's just think about this. So it's going to go to up to infinity somewhere. And where is that?

STUDENT: 3/4.

PROFESSOR: 3/4. So does that make sense? Can someone tell us what's going on and what is the use of this whole thing here? Yeah, in the back. What's your name?

STUDENT: Julianne.

PROFESSOR: Yeah, Julianne. Go ahead.

STUDENT: [INAUDIBLE] 0. So part, it would give you negative infinite. And so you just solve for D in there.

PROFESSOR: OK, so when D is 3/4 you'll get 1 minus 1. You get 0. That'll be negative infinity. And then there's a minus in front, so it'll be constant infinity. So that's true. And does that intuitively make sense to you? We have a sequence. It's evolving randomly, according to this model.

And then we have that ancestral sequence. And then we have a modern descendant of that sequence, millions of generations-- or maybe thousands of generations, or some large number of generations away. We line up those two sequences. We count how many matches and how many mismatches. What's the fraction of mismatches, of differences we have? Basically if that-- let's look at a

different case. What if d is very small? What if it's like 1%. Then what happens?

If d is small, turns out k is pretty much like d. It grows linearly with d in the beginning. So does that make sense? That makes sense. Because k is the true number of substitutions that happen. When you go one generation, the true number of substitutions and the measured number of substitutions is the same.

Because there's no back mutations. But when you go further, there's an increasing chance of a back-- there's an increasing chance of a mutation, therefore increasing chance that you also have a back mutation. And so this is what happens at long time.

So basically this is linear here and then goes up like that. And so what this allows you to do is d something that you can measure. And then k is something that you want to know. The point is, if I measure the difference between human and chimp sequence, it might be only 1% different. And if I have an idea of mutation rate per generation, I configure out how many generations apart, or how much time has passed, since humans split from chimp.

But if I go to mouse, where the average base might be-- there might be only a 50% matching-- if that's true, there have been a lot of changes there. There will be a lot of bases that have changed once, as well as a lot that may have changed twice, and may have actually changed back.

And so that let's say human and mouse are 50% identical. That 50% identical-- I can't just compare it to let's say the 1% with chimp and say it's 50 times longer. That 50% will be an underestimate of the true difference. Because there's been some back mutations as well. And so you have to use this formula to figure out what the true evolutionary time is, the true number of changes that happened. Yeah, go ahead.

STUDENT: Does simple count refer to just the difference in the amount of mutations? Or what's--

PROFESSOR: The simple count is what you actually observe. So you have a stretch of sequence--

let's say the beta globin genomic locus in human. You line it up to the beta globin locus in chimp. You count what fraction of positions differ? What fractions are different? That's d.

And then k is-- actually, it's slightly complicated here. Because if this is human and that's chimp, then k is more like-- because you don't actually observe the ancestor. You observe chimp. So you have to go back to the ancestor and then forward. So that's the relevant number of generations. And so k will tell you how many changes must have occurred to give you that observed fraction of differences. And for short distances, it's linear. And then for long, it's logarithmic, basically. Yeah, question.

STUDENT: So I'm guessing all of [INAUDIBLE] that selection is absent.

PROFESSOR: Right, right. This is ignoring selection. That's a good point.

So think about this. And let me if other questions come up.

So this actually came up the other day when we were talking about DNA substitution models. So Kimura and others have observed that transitions occur much more often than transversions, maybe two to three times as often, and so proposed a matrix like this. And now you can use what you know about stationary distributions to solve for the limiting or stationary distribution of this matrix. And actually, you will find it's still symmetrical. It's a little bit more complicated now, but you'll still get that 1/4, 1/4.

But then more recently others have observed that really, dinucleotides matter in terms of mutation rates, particularly in vertebrates So what's special about vertebrates is that they have methylation machinery that methylates CPG dinucleotides on the C.

And that makes those C's hypermutable. They mutate at about 10 times the rate of any other base. And so you can give a higher mutation rate to C, but that doesn't really capture it. It's really a higher mutation rate of C's that are next to G's. And so you can define a model that's 16 by 16, which has dinucleotide mutation rates. And that's actually a better model of DNA sequence evolution. And it's just the math gets

a little hairier if you want to calculate stationary distribution. But again, it can be done.

And it's actually pretty easy to simulate. Knowing that it will converge to the stationary, you can just run the thing many times. And you'll get to the answer. And there's even been strand-specific models proposed, where there are some differences between how the repair machinery treats the two DNA strands that are related to transcription coupled repair. So you actually get some asymmetries there. And this is a reasonably rich area. And you can look at some of these references.

All right, so one more topic, while we're on evolution-- this is very classical. But I just wanted to make sure that everyone has seen it. If you are looking specifically at protein coding sequences, exons, and you know the reading frame, you can just align them.

And then you can look at two different types of substitutions. You can look at what are called the nonsynonymous substitutions, so changes to the codons that change the underlying amino acid, the encoded amino acid. And you define often a term that's either called Ka or dN, depending who you read, that is the fraction of nonsynonymous substitutions divided by nonsynonymous sites.

And in this case let's do synonymous first. So you can also look at the other changes. So these are now synonymous changes which are base changes to triplets that do not change the encoded amino acid. So in this case, there are three of those.

And a lot of evolutionary approaches are just based on calculating these two numbers. You count synonymous changes. You divide by synonymous sites, count non-synonymous substitutions, divide by non-synonymous sites. And so what do we mean synonymous site?

Well if you have only amino acids that are fourfold, that have fourfold degenerate codons, which is all of them are like that in this case, then for example GG-- or let's see what's up here. Yeah, CC anything codes for proline. Do we have any of those?

Actually, these are not all fourfold degenerate. I apologize.

But glycine, for example-- so GG anything is glycine. So in this triplet, this triplet here, there's one synonymous site. The third side is a synonymous site. You can change that without changing the amino acid. But the other two are non-synonymous.

So to do first approximation, you take non-synonymous substitutions and divide by the number of codons-- I'm sorry, the number of codons times 2, since there are two non-synonymous positions in each codon. And you take synonymous substitutions, divide by the number of codons. OK, does that make sense? One per codon.

OK and so what do you then do with this? You can correct this value using-- basically this is the Jukes-Cantor correction that we just calculated, this 3/4 log 1 minus 4/3. That applies to codon evolution as well as individual base evolution.

And what people often do with this is they calculate Ka and Ks for a whole gene. Let's say you have alignments of all human genes to their orthologs in mouse-- that is, the corresponding homologous gene in mouse. And you calculate Ka Ks. And then you can look at those genes where this ratio is significantly less than 1, or around 1, or greater than 1. And that actually tells you something about how that-- the type of selection that that gene is experiencing.

So what would you expect to see-- or if I told you we've got two genes and the Ka/Ks ratio is much less than 1. It's like 0.2. What would that tell you? Or what could you infer about the selection that's happening to that gene? Ka/Ks is much less than 1. Any ideas? Julianne, yeah.

STUDENT: The protein sequence is important-- or the amino acid sequence.

PROFESSOR: Yeah, exactly. The amino acid sequence is important. Because you assume that those synonymous sites and non-synonymous sites-- they're going to mutate at the same rate, right? The mutation processes don't know about protein coding. So what you're seeing is an absence, a loss, of the non-synonymous changes.

80% of those non-synonymous changes have been kicked out by evolution. You're only seeing 20%. And you're using, assuming the non-synonymous are neutral-- I'm sorry. I seem to have trouble with these words today. But you assume that the synonymous ones are neutral. And then that's calibrates everything. And then you see that the non-synonymous are much lower. Therefore you must have lost-- these ones must have been kicked out by evolution.

So the amino acid sequence is important. And it's optimal in some sense. The protein works-- the organism does not want to change it. Or changes to that protein sequence make the protein worse. And so you don't see them. And that's what you see for most protein coding genes in the genome-- a Ka/Ks ratio that's well below one. It says we care what the protein is. And it's pretty good already. And we don't want to change it.

All right, what about a gene that has a Ka/Ks ratio of around 1? Anyone have an idea what would that tell you about that gene? There are some-- Daniel?

STUDENT: The sequence is-- it doesn't particularly matter. Maybe it's a non-coding, non-regulatory patch of DNA. I assume there must be something.

PROFESSOR: Yeah, so it could be that it's not really protein coding after all. It's non-coding. Then this whole triplet thing we were doing to it is arbitrary. So you don't expect any particular distribution. That's true. Any other possibilities? Yeah, Tim.

STUDENT: Could be that there are opposite forces that are equilibrating. For example, we're taking the unit of the G. But maybe in one half of the G there's a strong selective pressure for non-synonymous and in the other half it's strong selective pressure for synonymous. Alternatively, it could be in the same par of the gene, but it's involved in two different processes. It's diatropic. So in one process it's selecting this one thing.

PROFESSOR: Yeah, or one period of time, if you're looking at 10 million years of evolution, it could have been for this first five million years it was under negative selection, and then it was under positive. And it averages out. Yes, all those things are possible, but kind

of unusual.

And so maybe if you saw that the-- if you plotted Ka/Ks along the gene and you saw that it was high in one area and low in another, then that would tell you that you probably shouldn't be taking the average across the gene. And that would be a good thing to look for.

But what if-- again, so we said if Ka/Ks is near 1 it could be that it's not really a protein coding gene at all. That's certainly possible. It could also be though that it's a pseudogene. Or it's a gene that is no longer needed by the organism. It still codes for protein, but the organism just could care less about its function.

It's something that maybe evolved in some other time. It helps you adapt to when the temperature gets below minus 20. But it never gets below minus 20 anymore. And so there's no selection on it, or something like that. So neutral indicates-- this is called neutral evolution.

And then what about a gene which has a Ka/Ks ratio significantly greater than 1? Any thoughts on what that might mean and what kind of genes might happen to-- yes, what's your name?

STUDENT: Simona.

PROFESSOR: Simona, go ahead.

STUDENT: It might be a gene that's selected against, so something that's detrimental to the cell or the organism.

PROFESSOR: It's detrimental-- so the existing protein is bad for you, so you want to change it. So it's better to change it to something else. That's true. Can you think of an example where that might be the case?

STUDENT: A gene that produces a toxin.

PROFESSOR: A gene that produces toxin. You might just lose the gene completely if it produced a toxin. Any other examples you can think of or other people? Yeah, Jeff.

STUDENT: Maybe a pigment that makes the organism more susceptible to being eaten by a predator.

PROFESSOR: OK, yeah if it was a polar organism and it happened to have this gene that made the fur dark and it showed up against the snow, or something like that. And you can imagine that. Or a very common case is, for example, a receptor that's used by a virus to enter the cell. It probably had some other purpose.

But if the virus is very virulent, you really just want to change that receptor so that the virus can't attack it anymore. So you see this kind of thing is much rarer. It's only less than 1% of genes probably are under positive selection, depending on how you measure it and what time period you look at. But it tends to be really recent, really strong selection for changing the protein sequence. And the most common-- well, probably the most common-- is these immune arms races between a host and a pathogen.

But there are other cases too. You can have very strong selection where-- well, I don't want to-- basically where a protein is maladapted, like the organism moves from a very cold environment to a very warm environment. And you just need to change a lot of stuff to make those proteins better adapted. Occasionally you can get positive selection there. Yeah, go ahead.

STUDENT: So the situation where K or Ks is 1-- could it be possible that the mRNA is under selection?

PROFESSOR: Yeah, so that basically we have always been implicitly assuming that the synonymous substitution rate was neutral. But it could actually be it's not neutral. That's under negative selection too. And it happens that they balance. That's also possible.

So for that, to assess that, you might want to compare the synonymous substitution rate of that gene to neighboring genes. And if you find it's much lower, that could indicate that the coding sequences-- the third base of codons is under selection-- could be for splicing, maybe. It could be for RNA secondary structure, translation,

different other-- that's a good point.

So yeah, you guys have already poked holes in this. This is a method. It gives you something. You'll see it used. It gives you some inferences. But there are cases where it doesn't fully work. OK, good.

So in the remaining time I wanted to do some examples of comparative genomics. So as I mentioned before, these are chosen to just give you some examples of types of things you can learn about gene regulation by comparing genomes again, often by using really simple methods, just blasting all the genes against each other or things like this. And also, if you do choose to read some of these papers, it can give you some experience looking at this literature in regulatory genomics.

So the papers I've chosen-- we'll start with Bejerano et al from 2002, who basically sought to identify regulatory elements that are things that are under evolutionary constraint. That's all he was trying to find. Didn't know what their functions were. But they turned out to be interesting nonetheless, which is maybe a little surprising.

And then this other work from Eddy Rubin's lab and others-- Steve Brenner's lab-- actually characterized some of these extremely conserved regions and assessed their function. And then Bejerano came back a few years later and actually had a paper about where these extremely conserved regions actually came from.

So we'll talk about those. Then we'll look at some papers that have to do with inferring the regulatory targets of a transacting factor. And the factors that we'll consider here will be microRNAs, mostly, Either trying to understand what the rules are for microRNA targeting and these Lewis et al papers, or trying to identify the regulatory targets in the genome.

And then, time permitting, we'll talk about a few other examples of slightly more exotic things. Graveley identified a pair-- or pairs-- of interacting regulatory elements through a clever comparative genomic approach. And then I'll talk about these two examples at the end if there's time, where a new class of transacting factors was inferred from the locations of the encoded genes in the genome. And

also an inference was made about the functions of some repetitive elements from, again, looking at the matching between these elements and another genome.

All right, so first example-- Bejerano "Ultraconserved elements." So they defined, in a fairly arbitrary way, ultraconserved elements as unusually long segments that 100% identical between human, mouse, and rat. This was in 2000-- I'm sorry, I might have the wrong-- it's either 2004 or 2002. I forget.

This was basically when the first three mammalian genomes had been sequenced, which were human, mouse, and rat. And there were whole genome alignments. So they basically said let's try to use these whole genome alignments to find what's the most conserved thing in mammals. So they wanted to see if there's anything 100% conserved. And so they did statistics to say what's an unusually long region of 100% identity.

Any ideas how you would do that calculation, what kind of statistics you would use? They used a really simple approach. What they did was they took one megabase segments of the genome, assuming it might vary across the genome. They took ancestral repetitive elements-- so repetitive elements that were inserted, that were present in mouse, rat, and human-- and assumed that they were neutrally evolving, they were not under selection.

And then therefor you could look at the number of differences and get an idea what the background rate of mutation is. And they use that. And they found that that rate was-- this is from their supplementary data-- that was never greater than 0.68. And so they just said well, if we have a probability of-- I'm sorry. One is heads.

So if they're all three the same-- yeah, so if we have a probability of 0.7 of heads, meaning that they're all three the same, then the chance that you have 200 heads in a row would be 1 minus P P to the 200, just like [INAUDIBLE] trials. And you can just multiply that times the size of the genome. And you say it's extremely unlikely that you'll ever see anything where there's 200 identical nucleotides in a row. So that's what they defined as an ultraconserved element.

So it all seems very silly for now, until you actually get to what they find. So they looked at where are these elements around the genome. They found about 100 overlapped exons of known protein coding genes, 100 are in introns, and the remainder are in intergenic regions.

So then they looked at well what kind of genes contain exons with overlapping-- or contain ultraconserved elements that overlap exons? Those are type 1 genes. And what kind of genes are next to the intergenic ultraconserved elements, to try to get some clues about the function of these elements.

And so they did this early gene ontology analysis. And what they found was that the ultraconserved elements that overlapped exons tended to fall in genes that encoded RNA-binding proteins, particular splicing factors, by an order of magnitude more frequent. And then the type 2 genes, the ones that were next to these intergenic ultraconserved regions, tended to be transcription factors. In particular, homeobox transcription factors were the most enriched class.

So this gave them some clues about what might be going on. Particularly the second class was followed up by Eddy Rubins's lab at Berkeley. And they tested 167 extremely conserved sequences. So some of them were these ultraconserved elements. And some of them were just highly conserved, but not quite 100% conserved.

And they had an assay where they have a reporter. It's a lacZ with a-- you take a minimal promoter, fuse in to lacZ, and then you take your element of interest and fuse it upstream. And then you do staining of whole mount embryos. And you say what pattern of gene expression does this element drive, or does it drive a pattern of gene expression?

And so 45% of the time it drove a particular pattern of gene expression. So it functioned as an enhancer. And these are the types of patterns that they saw. So they saw often forebrain, sometimes midbrain, neural tube, lim, et cetera. So many of these things are enhancers that drive particular developmental patterns of gene expression. So that out to be actually-- that was a pretty good way to identify

developmental enhancers.

So they wondered, is there anything special about these ultraconserved regions, these 100% identical regions, versus others that are 95% identical. And so they tested a bunch of each. And they found absolutely no difference there. They drive similar types of expression.

And you can even find individual instances of them that drive pretty much exactly the same pattern of expression. So this whole 100% identical thing was just a purely-- it was purely arbitrary. But still, it's useful. These things are among the most interesting enhancers that have been identified.

So what about the-- oh yeah, so where did they come from? OK, so this is totally from left field. Bejerano was looking at some of these ultraconserved elements, probably just blasting them against different genomes as they came out, and noticed something very, very strange. And that was there had recently been some sequencing from coelacanth. So for those of you who aren't fish experts, this is a lobed fin fish, where they found fossils from dating back to 400 million years.

And they noticed that these fossils-- the morphology never changed. From 400 million, 300 million years, you could see this fish. It was exactly like this. And it has lobed fins. That was why they're interested in it. Because the fins-- they have a round structure. They look almost like limbs, like maybe this guy could have evolved into something that would eventually live on land.

Anyway, but they thought it was extinct. And then somebody caught one. In the '70s, in the West Indian Ocean, from deep water fishing, they pulled one up, and it looked exactly like these fossils from 400 million years before. And so then of course somebody took some DNA and did some sequencing.

And what Bejerano noticed is that this one megabase or so coelacanth sequence had a very common repeat in it that was around 500 bases or so, that looked like a SINE element. SINE elements-- short, interspersed nuclear element, like Alus, if you're familiar with those, so some sort of repetitive element.

And this repetitive element was very similar to these ultraconserved enhancers in mammals. So something that we normally think of as the least conserved of all, like a repetitive element that inserts itself randomly in the genome, had become-- some of these elements had become among the most conserved sequences later in evolution.

So how does that make any sense at all? Anyone have a theory on that? I can tell you how they interpreted it. So their theory-- here's some text from their-- anyway, you can look at the paper for the details here. But their theory is basically that once you have a repetitive element-- initially it's a parasitic element, inserts itself randomly in the genome, doesn't actually do anything.

But once you have hundreds of them, by chance there will be perhaps a set of genes that have this element next to them, where you'd like to control them coordinately. You'd like to turn all those genes on or all those genes off in a particular circumstance-- a stress response, during development, something like that.

And so then it's relatively easy to evolve a transcription factor, for example, that will bind to some sequence in that element. And then it'll turn on all those genes. Of course, it'll turn out all the genes that have the elements near them. So it'll probably turn on some extra genes that you don't want. But you can then-- selection will then tune these elements. It gives you a quick way of generating a large-scale gene expression response. Because you've got so many of these things scattered across the genome.

And so this-- that's as good as an explanation as we have, I would say, for what is going on here. And there's been some theories about this. And they point out that actually something like 50% of our genome actually comes from transposons, if you go back far enough. Some are recent, some are ancient. And that maybe a lot of the regulatory elements-- not just these ultraconserved enhancers, but others-- may have evolved in this way.

So basically you insert a bunch of random junk throughout. And then the fact that

it's all identical, because it derived from a common source, you use-- that fact actually turns it into something that's useful, a useful regulatory element.

All right, just wanted to throw that out. So what about the exonic ultraconserved elements? So here's one. This is a 600 18 nucleotide region that's 10% identical between human, mouse, and rat. It's one of the longest in the genome. And where is it? It's in a splicing factor gene called SRp20. And it's actually not in the protein coding part. It's in a essentially non-coding exon of this splicing factor. So it's this yellow exon here.

And what you'll notice is there's this little red thing here. That's a stop codon. So this gene is spliced-- produces two different isoforms. The full length is the blue, when you just use all the blue exons. But when you include this yellow exon, there's a premature termination codon that you hit. So you don't make full-length protein. Instead, that mRNA is degraded in a pathway called nonsense mediated mRNA decay.

So the purpose of this exon appears to be so that this gene can regulate expression of the protein at the level of splicing. And others have shown that this protein, the protein product, actually binds to that exon and promotes the splicing of that exon. So it's basically a form of negative auto regulation. The gene-- when the protein gets high, it comes back and shifts the splicing of its own transcripts to produce a non-functional form of the message and reduce the protein expression.

So the theory is that this helps to keep this splicing factor at a constant level throughout time and between different cells, which might be important for splicing. But that's only a theory. It could be something else. And it does not explain why you need 600 nucleotides perfectly conserved in order to have this function. So I think these exonic ones are still fairly mysterious and worth investigating.

A couple examples from microRNAs-- you probably it's just a brief review on microRNAs. They are these small, non-coding RNAs, typically 20 to 22 nucleotides or so. They have a characteristic RNA secondary structure in their precursor, often called miRNAs.

And they're produced from primary transcripts typically, or introns, or protein coding genes, which are then processed in the nucleus of an enzyme called drosha into a hairpin structure, like so. And then that is exported to the cytoplasm, where it's further processed by an enzyme called dicer to produce the mature microRNA, which enters the risk complex, and which then pairs the microRNA with mRNA targets, usually in the 3'-UTR. And that either inhibits their translation or triggers the decay of those messages.

So microRNAs can do-- they can be really important. Weird animation-- but for example, this bantam microRNA in flies inhibits a proapoptotic gene hid. If you delete bantam, apoptosis goes crazy. And you can see this is a normal fly. There's a little fly in there with red eyes and so forth. In this guy there's just a sack of mush. All the cells-- most of the cells actually died.

So microRNAs play important roles in developmental pathways. And so we wanted to figure out the rules for their targeting. And so this was an early study from Ben Lewis, where he looked for conserved instances of segments, short oligonucleotides, that match perfectly to different parts of the microRNA, using again these human, mouse, rat alignments, which were what was available at the time.

And what he found was that if you took the set of microRNAs which were known, and you identified targets of these defined as 7-mers that are perfectly conserved in 3'-UTRs of mammalian messages, and then you looked at how many you got and you compared that to the number of targets of shuffled microRNA-- so where you take the whole set of microRNAs, randomly permute their sequences so you generate random stuff, look at how many conserve targets they have-- that there was a significant signal above background, in the sense of real conserved targets, specifically only for the 5'-end of the microRNA. Especially, bases 2 to 8 of the microRNA gave a signal. And no other positions in the microRNA gave a significant signal above background. And so that led to the inference that the 5'-end of the microRNA is what matters, specifically these bases.

And then later, alignments of actually paralogous microRNA genes, shown here-- so

these are different let-7 genes. You can actually see that the 5'-end of the microRNA, which the microRNA's shown here in blue-- this is the fold-back. So you get conservation of the microRNA and of the other arm of the fold-back, which is complimentary. Little conservation of the loop, but the most conserved part of the microRNA is the very 5'-end, consistent with that idea.

Just one more example, because it's so cool-- so this is the dscam gene in drosophila. And this gene has four different alternative spliced regions which are each spliced by mutually exclusive splicing. So there are actually 12 copies of exon 4 and 48 different copies of exon 6. And messages from this gene only ever contain one of those particular exons.

And so Brent Graveley asked how does this gene get spliced in a mutually exclusive way? How do you only choose one of those 48 different versions of exon 6? And so what he did was did some sequencing from various fly and other insect species of this locus, did some alignments. And he noticed that there was this very conserved sequence just stream of exon 5, right upstream of this cluster.

And then, looking more carefully, he saw that there is another sequence, just immediately upstream of each of the alternative exons, that was very similar between all those exons, and also conserved across the insects. And then he started at these for a while, and recognized that actually this sequence up at the 5'-end is-- its consensus is perfectly complimentary to the sequence that's found upstream of all of the other exons.

And so what that suggested, immediately, is that splicing requires the pairing of this sequence from exon 5 to one of those downstream sequences. And then you'll splice to the next exons that's immediately downstream and skip out all of the others. And that's been subsequently confirmed, that that's the mechanism.

So this just shows you that to figure this out by molecular genetics would have been extremely difficult. But sometimes comparative genomics, when you ask the right question, you get a really clear-- you can actually get mechanistic insights from sequences.

So that's it. And I'm actually passing the baton over to David, who will be-- take over next week.