

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: So yeah, but today, what we're going to do is talk about evolution. And in particular, we're going to complete our discussion of evolution in the presence of clonal interference when multiple mutant lineages are competing in the population at the same time. And then, we'll move on to try to think about evolution on these so-called rugged fitness landscapes.

So such ruggedness occurs when there are interactions between the mutations within the organism. So a so-called epistatic interactions can, perhaps, constrain the path of evolution. So we'll complete our discussion of Roy Kashoney's paper on the equivalence principle and the presence of clonal interference.

Then we'll say something about how clonal interference slows down the rate of evolution. It slows down the rate at which the population can increase in fitness because of this competition between different beneficial lineages. And then, we'll discuss this paper by Daniel Weinreich, which I think, for many of us, had a real important effect in just terms of getting us to think about evolution in a different way. Any questions before? Yes.

AUDIENCE: Question about the exam.

PROFESSOR: Yes. In case you've tried to forget, we have an exam again next week. So research and education indicates that the more exams, the better. So if that makes you feel any better in terms of process, then use that. So next Thursday, 7 o'clock, we will announce the room later. Did you have a question about that?

AUDIENCE: Is it everything from the start or [INAUDIBLE]

PROFESSOR: Yeah, so it will be weighted towards the material that we did not test through exam

one. But you can expect to get two plus or minus 1 questions on material that was covered in the first part of the class. OK. Any other questions? OK.

All right, so coming back to this discussion of the equivalence principle. So the last figure of the paper-- it's a figure of four-- illustrated some alternative underlying distribution. So what we wanted to know is-- all right, so they're going to be some beneficial mutations that these E.coli can get in this new environment. And they're going to be distributed according to some probability distribution.

So what we, in principle, want to know is this distribution of effects of beneficial mutations. So the probability of distribution, $P_{\text{beneficial}}$, is a function of s . And what was, kind of, the scale of beneficial mutations that they observe in the paper. All right, how much better do the populations get after these few 100 generations? Hm?

AUDIENCE: 1%.

PROFESSOR: All right, so [? order of ?] percent. A few percent, right? Precisely. And indeed, they had three different models for some of the underlying distributions and how they might behave. So they had the exponential uniform and the [INAUDIBLE] kind of delta function.

And in all these cases, are those the only three possible underlying distributions? Those are just, kind of, typical distributions. What you want, in this case, is you want these three degree distributions to, somehow, be qualitatively different so you can drive home the point that you can, in principle, describe the results of their evolution experiment via wildly different underlying distributions. OK.

Now in all these cases, you have to specify the mutation rate, as well as the mean selection coefficient. All right, so the mean of this distribution. OK, now the question is, can we get any intuitive insight into why there were the patterns that they observe in terms of the mutation rate that they had to assume and the mean selection coefficient.

There were some region of parameter space for those two distributions, for each of

those distributions, that were, kind of, consistent with their data. So in particular, there was a mutation rate-- μ -- for the exponential distribution, a mutation rate for the uniform, and a mutation rate for the delta function. And also, there was a mean s associated with the exponential, mean s for the uniform and, again, a mean s for the delta.

AUDIENCE: Were they trying to minimize the parameter numbers?

PROFESSOR: Right. OK, so in this case, they wanted to compare these three distributions. Each of them were specified by two parameters. You could come up with other underlying distributions that might, for example, have a larger number that might be specified by a larger number parameters. But then, it's harder to compare the quality of the fit and so forth.

So they chose these three distributions just because this is, somehow, the rate that these new mutations will appear in the population. And then, this tells us something about how good those mutations are. Right. Now some of you have the paper in front of you. And that's OK.

But based on our understanding of how the clonal interference, kind of, manifests itself in terms of leading, eventually, they have the log of the fraction. So the fraction starts out 50-50. $\log F_1$ over F_2 . So like, cyan and yellow, say?

All right. So this thing starts out here. And then, one side gets beneficial mutation. So it, kind of, comes up. So they measure the slope, for example, of the lineage that is taking over the population. So they want to know, well, which of these distributions and associated parameters will be able to explain the range of different trajectories that they saw?

So the question is, can we order these things? And why? All right, which one of these should be the largest, second largest, third largest, and so forth? OK. Now it's OK if you just-- well, if you had the paper in front of you, you could just read it off. But ultimately, you're going to have to be able to explain why it is that one is larger than the other. OK?

So what I want to know is, for example, at what order should these things come in? All right. So what I want to do is let you think about it for a minute. And then, we're going to vote by putting our cards from high mutation rate to low mutation rate among a, b, and c. Yes?

AUDIENCE: So the means are constrained to be the same?

PROFESSOR: So it's really going to be some range of parameters on each of these. So the question is, why is it that in the range of parameters that are consistent with what they observe experimentally, that these things have some order? Are there any other questions about the question?

I'll give you 30 seconds to think about what the mutation rate should, kind of, be in this situation.

AUDIENCE: The answer is highest to lowest mutation rate.

PROFESSOR: Right. So you're going to put the highest mutation rate up high, the lowest mutation rate down there, and the middle one in between. Yeah. All right. Do you need more time? All right, let's see where we are.

And it's OK if you're confused or don't know what I'm trying to ask. But let me see where the group is. Ready? Three, two, one. All right. So I would say it's, pretty much, all over the place, whether people are voting something in reality or not.

OK, right. So the situation is that we have the data, which is shown in figure 3A, which is a bunch of these things that, kind of, look like this. All right. So we have some times. We have some slopes. We want to know how can we understand that data that we get out.

So what we're going to do is we're going to take a model in which we say, all right, we're going to start with this population that's all identical. And then, we're going to allow some mutations to accumulate. And we're going to let them compete against each other. And then, see what happens to the other. So there's going to be some, again, distribution of slopes and so forth.

All right. To what degree does this sort of data constrain that underlying distribution? Between something that looks like an exponential, something that has a uniform distribution, and something that is a delta function. So that's the exponential. This is the delta. And this is the uniform. Yeah.

AUDIENCE: So we measure the slope at what time [INAUDIBLE]?

PROFESSOR: Yeah, it could. That's right. So this thing could turn around at various times. So I think that there are a number of different ways that you could argue about the right way to do this. In practice, I think it's not going to be very sensitive because there's a minority of them that will actually be turning around, for example.

So you could, for example, just say all of the trajectories that cross some point, I mean, measure the slope. And I think that would be sufficient.

AUDIENCE: But if you don't see this fraction turn over, you could still have clonal interference?

PROFESSOR: If you don't see the fraction.

AUDIENCE: Like, in the sense--

PROFESSOR: That's right. So even if you don't see these things like, the flatten out, for example, then you could still have clonal interference because the slope might still be steeper than it would be in the absence of clonal interference.

All right, what I'm going to do is I'm going to let you discuss with a neighbor for one minute. And then, we'll, maybe, discuss as a group just because I want to make sure that everybody gets a chance to try and verbalize their thought process. And if we discuss in a group, then only a few of us get to. All right, so one minute. Try to discuss it with your neighbor. And then we'll reconvene.

[SIDE CONVERSATIONS]

Yeah, and we're going to discuss the means in a moment. So indeed, these distributions will not end up having the same mean s.

AUDIENCE: What are you controlling?

PROFESSOR: What we're controlling is that we're asking about what range of parameters for each distribution will adequately fit the data.

AUDIENCE: I know [INAUDIBLE]. Does it depend on what you get?

PROFESSOR: The data will be, basically, the initial slopes here and when they deviated from a 50-50 mixture.

AUDIENCE: OK.

PROFESSOR: All right, so that's what [INAUDIBLE] is those histograms.

AUDIENCE: Yeah. OK. [INAUDIBLE].

PROFESSOR: All right, so it seems like we've quieted down, which means that we all agree on the answer. Is that-- no? OK, well I think that this is, actually, pretty tricky. So that's fine. I just want to see where we are, though. All right. Reconfigure your cards. Your best guess for the orders of the mutation rates between exponential uniform and delta.

All right, ready? Three, two, one. OK, so we're migrating towards some things. OK, great. And can somebody verbalize the answer that their group got?

AUDIENCE: So our answer is A, B, and C.

PROFESSOR: OK.

AUDIENCE: [INAUDIBLE] exponential [INAUDIBLE]. We can not see most of the [INAUDIBLE] lower selection coefficient mutations.

PROFESSOR: OK.

AUDIENCE: So we're actually underestimating the mutation rate from the data.

PROFESSOR: Underestimate. OK, no, I can see what you're saying. OK, yeah, so the idea is that you're saying that we don't see an awful lot of the mutations here, which means that

the true mutation rate, the underlying mutation rate is, somehow, much larger than you would have thought based on the mutations that you actually see here or something.

And there's maybe another. OK, so they're different. All right. OK, it's certainly along-- yeah, sometimes it's true. And then, of course, there are different ways of saying this. Yes?

AUDIENCE: Yeah, same answer but a slightly different way of thinking about it. If you're just randomly sampling any of these distributions, then your sample drawn from the exponential distribution. It's going to be low selection [INAUDIBLE] more often than it's going to be for the other ones. Like the delta is [INAUDIBLE].

PROFESSOR: That's right. Yeah, yeah, yeah.

AUDIENCE: [INAUDIBLE] every time the uniform. It's going to be equally likely to be a high selection coefficient as opposed to a low selection coefficient. But with the exponential distributions, your most likely to be a low selection coefficient. So you want more mutations.

PROFESSOR: That's right. You, somehow, need more mutations of that exponential in order to sample out there. Right? So which one is going to have a more clonal interference? Which of these distributions will end up having the most clonal interference after you fit the data? Yeah?

AUDIENCE: The one with the highest mutation.

PROFESSOR: The one with the highest mutation, right? Kind of has to. Of course, and even though some of those mutations are going to be loss, still it's going to have the most clonal interference there. And indeed, if the underlying distribution were modeling as a delta function and, in their [? fit ?], what they got was that this might be around 5 and 1/2%, I think.

Yeah, so [? 5 to 5 and 1/2 ?]. OK. So here, this guy was around 0.055. Between 5 and 5 and 1/2%. So what they're saying is, all right, well you could explain all of our

data just by assuming that there's some mutation rate where, periodically, some individual gets a beneficial mutation that is a 5, 5 and 1/2%.

And that could, in principle, be used to explain the base features here, how long have to wait before anything happens, and the slope when something starts happening. So the histogram that they plot is actually, somehow, this initial this initial slope once you start seeing it deviate from 50/50. OK? All right.

But their point is it that that does not prove that the underlying distribution is a delta function with some mutation. And indeed, to explain the data with a delta function, you don't actually don't need any clonal interference. Right? You just say, OK, well somebody gets a mutation. It's 5%. And eventually, it's going to spread. And that's what we see.

If you want to explain the later dynamics of flattening out and so forth, then you have to allow the other lineage to get a mutations, as well to cause a flattening. But as far as the base dynamics of when you leave the 50-50 in the initial slope, you don't even really need to have any clonal interference to explain their data with a delta function underlined. And that's why you also can get by with a very low mutation rate because you don't really need much in the way of competing lineages. Yeah?

AUDIENCE: But what if the slopes are [INAUDIBLE]?

PROFESSOR: Yeah, yeah. No, right. So you're not actually going to get the true distribution of slopes. But their argument is that a lot of that could just be noise and measuring the slopes and so forth because, if everything is a delta function, then you would start out by just getting one slope, unless you offer multiple mutations on a lineage.

And then, things could get more complicated. But yeah. In this case, all of these guys would have the same slope. But that's, at least, a reasonable first order approximation to the data. However, as you move to these distributions in uniform and exponential, you're going to need more and more clonal interference to, kind of, explain the data. So you'll need higher and higher mutation rate. What's interesting

is that you also have a lower and lower mean s . OK?

AUDIENCE: Can you just explain why you need to explain the data?

PROFESSOR: Yeah, sure. And I think that drawing these underlying distributions is really helpful. So first, we're going to draw the delta function. That, kind of, makes sense that you can fit everything just by assuming 5, 5 and 1/2%, right.

So what we're going to do is draw the various P . So I drew those distributions. But they weren't necessarily to scale. i.e, they didn't necessarily have the proper mean selection coefficient. What we can do here is we can draw this is the mean s of the delta function, which was 5, 5 and 1/2%. All right, we got this guy here.

Now the question is can we describe the data using a uniform distribution with the same mean selection coefficient? All right, so we're going to have you vote yes and no. And if you say no, then you have to say what's going to go wrong. All right? The question is can we just use the same mean selection coefficient for our uniform distribution.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Right. So what we're going to do is we're going to measure the experiment where we would have, say, 96 different evolutionary trajectories where we measure the time that it takes for something to happen and then the initial slope. And we're going to take a histogram of those things, and compare it between what we would get in the model with what we got experimentally.

All right, the question is can we use the same mean s for a uniform as we did for the delta function. And if you say no, you have to say why not. A is yes. B is no. Ready? Three, two, one.

All right, so we have a bunch of no's. Maybe a few yes's. All right but then some of the no's, it's incumbent on you [INAUDIBLE]. So I don't know. So yes, so one of the no's, why not?

AUDIENCE: So if we have [INAUDIBLE] if it has the same average, then there are going to be

outliers that are more better for the [INAUDIBLE].

PROFESSOR: That's right. So the problem here is that, if we use the same mean selection coefficient pretty for the uniform, then what we're going to end up with is something that comes out twice as far.

AUDIENCE: Wait. But the delta function is just a uniform distribution with 0.

AUDIENCE: Yeah.

AUDIENCE: So you could always just fit it with [INAUDIBLE].

PROFESSOR: Yeah, we're assuming it's a uniform distribution that starts at 0 and then goes out to some [? amount ?]. Yeah. So the idea is that-- I mean, whatever. That's the model. And I think it's a reasonable model because we know that there are a lot of mutations that have little effect. So it makes sense for a distribution to start at 0, if you're going to have something like a uniform. Right?

And the problem with such a uniform distribution is that because there's going to be some clonal interference, what that means is that you're going to be, kind of, weighted out here. And that means that you'll, kind of, see mutations that are out here around 10%, instead of around 5%. So what you actually want, then, is something where the mean selection coefficient is around half of what you had as the delta function. So you want something that really looks more like this.

And that's actually why, if you look at the data or if you look at this figure, then the area that works for the uniform distribution has a mean/coefficient of 3%. So this thing comes out to around 6% here. So just beyond the s correspond to the delta function. So this is the delta. And this is the uniform.

So here is around 6%. And the mean of this is [? at half ?] of that. Right? What's happening is that there's some clonal interference. And you only need a modest amount of clonal interference because if you sample from a uniform distribution just a few then, the most fir one will be around here.

And it'll already be relatively peaked. And of course, there's also this issue that you have to survive stochastic extinction. So that amplifies the effect further. So really, if you just have, say, two mutations sampled from the uniform that survives stochastic extinction, you're already going to get something that's peaked around there. Does that make sense? Yeah.

AUDIENCE: But what if we [INAUDIBLE]?

PROFESSOR: So then, the question is what exact mutation rate do you need. And basically, you need a high enough mutation rate that you get some mutations and that you can have some clonal interference. And the question is, what prevents you from having a mutation rate that's too high, maybe?

AUDIENCE: My question was why do you need clonal interference to explain the data.

PROFESSOR: Ah, right.

AUDIENCE: I mean, of course, you have these [INAUDIBLE] focused on what happened early on.

PROFESSOR: That's right. Yeah. Right, OK. So in there it comes down to how peaked this distribution of slopes is because there are very few shallow slopes. And of course, then, it gets into questions about quality of data and so forth. And that's more subtle.

But certainly in principal, in the absence of clonal interference, you would have some fair number of shallow slopes. It would be under represented just because of the stochastic extinction business. But still, you need some to explain the, sort of, peakiness of that distribution of the slope distribution.

Now the exponential is interesting because it's in a very, very different regime. So I just want to show this is mean s for the uniform. And actually, if you look at the figure of the mean s for the exponential, it's down there around 1%, which is a little bit surprising because what this is saying is that, if you look at this distribution, this initial slope, kind of, extends down here.

And you have something that falls off, dramatically, though, right. OK. So how is it possible that you could use such a distribution that's peaked over here that's so far over on the left and still explain the same data?

AUDIENCE: [INAUDIBLE].

PROFESSOR: That's right. So it's true that I've drawn this, kind of, around 0. But the exponential and principle goes to [? infinity ?]. It's just that it falls off exponentially. But what this means is that you're sampling pretty far out on the exponential in order to get the same mean effect.

So you're actually going out to five or six times this characteristic s . So you're talking e to the minus 5. Right? That means there's a lot of clonal interference that has to be happening in order to explain this, right? But why is it that it's way over here?

I mean, why not just use an exponential with an s that's more like 3, 4, 5, 6%? What's that? Well OK. But, you know, I mean, this is just a model. I can do whatever I want because, actually, they fit their data with these models. So realistic just means that it explains their data.

So there's nothing a priori wrong with saying, oh, here's an exponential with a characteristic fall off of 5%. That's, in principle, fine. I mean, it doesn't work, for some reason, but we have to figure out why. Yeah.

AUDIENCE: [INAUDIBLE].

PROFESSOR: OK, right. So you're likely going to get a low selection coefficient, OK. Is that the problem?

AUDIENCE: [INAUDIBLE].

PROFESSOR: OK, that's true. But I guess the question is why can't we use?

AUDIENCE: [INAUDIBLE].

PROFESSOR: Two mutations? Two clusters? What do you mean?

AUDIENCE: [INAUDIBLE].

PROFESSOR: OK, the time between two. OK, that's an interesting statement. Although, the region or parameter space that they claim works is actually a region where we have a really high mutation rate. Orders of magnitude higher than the other two distributions.

So in that sense, the time between mutations being established is really small because they're saying that, oh, if you want to fit the [INAUDIBLE] exponential, then you have to assume lots of clonal interference. So that means the time between successive establishments or mutations is really, actually, very short. So that's actually the regime where they claim works.

So the question is why is it that we can't go to this other regime? In this figure that they make, why is it that the allowed mean coefficient doesn't extend out to over there? Yeah.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Right.

AUDIENCE: Because the exponential is [INAUDIBLE] that [INAUDIBLE].

PROFESSOR: That's right. Yeah, that's right. The point is that, to explain their data, you need something to cause this distribution to get more peaked over here, which means you need to have a fair amount of clonal interference. But that means you need to have a high mutation rate.

But once you have that high mutation rate, if you have an exponential that comes out here, then you would actually sample way out here, as well. So the only way to get a peaks distribution around here is to have it so that the exponential is really suppressing those really good mutations. But you have a lot of clonal interference that, kind of, pulls things out.

Now there still is a fair range of parameters that work here. You know, it goes from,

say, half a percent up to, maybe, 1 and 1/2% in terms of this mean selection coefficient. And the mutation rate that works then changes. So as you got a larger mean selection coefficient, the mutation rate that is compatible goes down because you need less clonal interference.

So that's why, if you look at this figure for the region that works in terms of the mutation rate for the beneficial mutation and the mean s , there's some region that looks, kind of, like this that works for the exponential. And this is, actually, a big range. So this is, actually, a factor of 100 in mutation rate that would be compatible.

And then, a factor of, maybe, three in mean selection coefficient. So there's some range of parameters that work. And you can understand why it is that this thing has to be shaped the way it is. Does that make sense? Yeah.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Right because the uniform is out here. And then, the delta function is here. So this is delta, uniform, and exponential. And you're just saying that it all, [? kind of ?], following. Is that-- yeah.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Yeah, well maybe some power law fall off would do that.

AUDIENCE: [INAUDIBLE].

PROFESSOR: There's nothing there's nothing magic about these three regions. And it's not that we're claiming that the mean selection coefficient cannot be in here. It's just that, if you pick these three underlined distributions, you get this range of different values that would work. So if you chose other underlying distributions, you could get other blogs.

But it's true that there is a general trend that, the higher the mean selection coefficient, the less clonal interference you need or want to explain the data.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Right. Why can't it go higher or lower? Yeah. I don't know. A factor of 100 isn't enough for you? Yeah, it's a good question. I'd have to think about it to figure out which-- because there's going to be a different effect on each side, presumably.

Yeah, and all of these distributions, there's some floor just because we know that you have to get these beneficial mutations within the first few tens of generations. Otherwise, the mutation wouldn't have gotten a chance to spread when it did. So that means we know that there's going to be a lower bound on the mutation always because, if it's too low, then we wouldn't have gotten the mutations in time.

And now, in terms of why it can't be higher, yeah, I'd have to think about it. Are there any other questions about this paper? I think it's a challenging paper, kind of, conceptually/mathematically. But I think it's interesting because it does get you to think about this process of clonal interference in new ways.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Yeah.

AUDIENCE: So is that enough? Like, is this condition rate known [INAUDIBLE]?

PROFESSOR: Oh, I see. Oh, that's a good question. Yeah. Yeah, so I think that the mutation rates that-- first of all, this is not a per base pair mutation rate. This is the rate that you get beneficial mutations.

So I'd say, the numbers are not ridiculous. But it's not that you can just take the known per base [? pair of ?] mutation rate and say, oh, well it has to be here because an awful lot of them are deleterious.

And it's very sensitive because the mutations that are occurring around here don't really matter for evolution because they tend not to survive stochastic extinction. They're not going to survive clonal interference. Yet, it could be a big part of the distribution. Right?

It could be that the majority of the mutations-- for example, in the exponential-- is

there. So you can actually have very different distributions that don't really change the evolutionary process but would change the rate of beneficial mutations. So I think the numbers are not ridiculous. But it's hard to constrain, actually.

What I want to do is, before talking about these rugged fitness landscapes and the Weinreich paper, let's just say something about the rate of evolution. And when we say rate, we're referring to the change and the mean fitness of the population with respect to time.

All right, so this is the rate of evolution. So this is the change in the mean fitness. Delta. Delta mean fitness divided by delta time. So what we want to do is just start by thinking about a situation where we assume that we're not all running out of new mutations that are good for us.

All right, so what we can do is just assume that, at some rate, μ -- so there's at rate μ beneficial, we'll say. We sample from some probabilities distribution of beneficial mutations. And then, something happens to cast an extinction. Maybe clonal interference and whatnot. But someone of them will fix. And then, that increases the fitness.

And then, for now, we'll just assume that the fitness is add. But for small S's, it doesn't matter whether we're thinking about [? fitness as ?] adding or multiplying. You guys understand what I just said there?

So if you get a mutation that has affect S1 and maybe the right way to think about this is that, if you get a mutation s2, then the [? fitnesses ?] perhaps, should multiply as the [? μ ?] model. But this is, of course, for a small s1 and s2, this is around 1 plus s2.

So for small S's for short times, maybe we don't need to worry about this because this is, for s1 and s2, much less than 1. Yes?

AUDIENCE:

My intuition would be that, if you already had a 5% increase i [? fitness ?], it would be harder to get [INAUDIBLE].

PROFESSOR: That's right. So eventually, that certainly is going to be the case that we're going to run out of these beneficial mutations. But for the first few thousand generations, it's roughly linear. So eventually, it does start curving over. But maybe not as fast as you would have thought.

And at the very least, this is a good [? no ?] model. And then, we can, of course, complicate things later. But for now, we'll just assume that you always sample from the same probability distribution of beneficial mutations, just for simplicity. The question is, how fast will the fitness of the population increase with time? OK?

Do you guys understand the question? All right, let's start by thinking about the regime where $\mu b N$ is much less than 1. So very low rates of mutation relative to the population size.

Now you might recall-- for clonal interference. So clonal interference not relevant. What we found last time was it required that the time that it took for mutation to fix had to be much less than the time between successive establishments of these beneficial mutations.

And this was 1 over $s \log N$'s. You should be able to [? derive ?] both of these. This and that and this, and the next step, as well. All right, so you can ignore clonal interference if this is true.

So let's say that we can ignore clonal interference. So for small population sizes or in the limit of low mutation rates. What we want to know is the rate of evolution. How will it scale with various things?

I'll go ahead and have us vote for-- OK. How does it scale with, [? particular ?], both, the mutation rate and the population size? It's proportional to what?

Holding another thing is constant. Holding, for example, the distribution and n constant. OK? Do you understand the question? Yes.

AUDIENCE: [INAUDIBLE].

PROFESSOR: OK. So this is just the rate of beneficial mutations. And this is just to the 0th power.

i.e, it doesn't defend.

AUDIENCE: Oh, OK.

PROFESSOR: Linearly, it's squared. All right, ready?

AUDIENCE: [INAUDIBLE].

PROFESSOR: Yeah, assuming that clonal interference is not relevant. So assuming, for a small mutation rate, how does it scale in mutation rate? OK, ready. Three, two, one.

All right, so we got a lot of B's, which is nice. So this is saying that, indeed, if mutation rate is small, then there's going to be some rate that mutations enter into the population. They may or may not survive stochastic extinction. But if they do, and that doesn't depend on mutation rate, then they can fix.

And in this low mutation rate regime, they don't compete with each other. In which case, if you double the rate of these things entered into the population, you'll double the rate that they get established. And you'll double the rate that these beneficial mutations will fix in the population So you double the rate. OK?

All right, now as a function of n . Does it go as N to the 0 or other? We got A, B, C, D. All right, I'll give you 15 seconds to think about it.

Ready? Three, two, one. OK, so now we're getting more disagreement. All right. And I'd say, largely, A's and B's. All right, there's enough disagreement. Let's go ahead and spend just 30 seconds. Turn to your neighbor.

[SIDE CONVERSATIONS]

Per individual, yeah.

[SIDE CONVERSATIONS]

All right, let's go ahead and re-vote just so I can see where we are. How is it that the

rate of evolution in this regime, no clonal interference, how is this going to scale the population size? Ready? Three, two, one. OK

So, I'd say that we have not really convinced each other of anything. All right, so A's B's? Somebody explain the reasoning. Yeah.

AUDIENCE: [INAUDIBLE] mutation rate [INAUDIBLE] larger the population, the larger we're going to get [INAUDIBLE] population.

PROFESSOR: OK. All right, so if you double the size of the population, you'll double the rate the new mutations enter into the population.

AUDIENCE: And the other important thing is that the fixation time just goes, like, [INAUDIBLE].

PROFESSOR: OK, right. So then there's the fixation time business. So how is that relevant? Yeah, somebody that said A, what was your partners reasoning? Or where you convinced by this argument? Or confused.

AUDIENCE: Well if it's [INAUDIBLE], then the [INAUDIBLE].

PROFESSOR: Ah. Right, so if it's a nearly new [INAUDIBLE] mutation, then it's true that the [INAUDIBLE] would be $1/N$ but--

AUDIENCE: But if N is very large then, sometimes, s will not [INAUDIBLE].

PROFESSOR: OK, right. But now you're invoking N being large, which I don't think we necessarily want to do. I mean, I guess there are a couple things to say there. One is that the mutations that are really nearly neutral will not have a very significant effect on the fitness.

And within that regime, I think, yeah, you'd have to check what happens there, right? But I think that, in most of these cases, small population, we're often saying population [? must be ?] 10^4 or so. In which case, the nearly neutral mutations are not very relevant. Right?

So indeed, over a broad range of conditions, in this situation, it's going to scale as

N. So the rate, so far, it's going to equal to there's a μb times an N. And that's basically because the rate the new mutations enter into the populations is μN . And the rate that they get established is just μN 's. This is, indeed, what this calculation is telling us.

Now the question is, how much of a fitness gain will we get? How is it going to scale with this probability distribution. So this probability distribution will give us some function of s . So this distribution has to be relevant. Do we agree?

So we want to know in what way is it relevant. Is it relevant via the mean, via the mean squared, the mean cubed-- I don't know-- the mean squared, or other? This one is harder. So it's worth spending, I'll give you, a full 30 seconds to think about it.

So the question is, how is it that the probability distribution will enter into the rate of evolution in this situation? OK? Question. No?

AUDIENCE: Can [INAUDIBLE]?

PROFESSOR: I'm sorry, multiple what?

AUDIENCE: Is it possible that it's entered in multiple situation?

PROFESSOR: Oh, yeah. You know, this is why I gave you flashcards. You can put up any combination you want. Another 15 seconds. This one is trickier. Let's go ahead and vote. Ready? Three, two, one. All right, so we got a lot of A's and B's. Some, maybe, C's and D's. All right, so yeah. We're pretty far. OK, we're all over the place.

All right, so there's going to be some distribution. There might be, for example, probability distribution function of s , function of s . We want to know it's going to something that's going to fall off in some way. Maybe exponential. Maybe something else.

Now to keep track of this, what we need to do is remember that the probability of establishment goes as s . The probability of establishment. And the [INAUDIBLE] model is equal to s . In other models, it might be $2s$. But it's around s .

And then, the question is, how much of a benefit will you get if you do establish? All right. And that is, again, going to go as-- and whatever the delta fitness is actually, again, equal to whatever acid is that you sampled. And the given mutation that appeared, you sample somewhere.

It has to, both, survive and well- if it survives, then it gives you an s . So what that means is that, actually, you end up averaging s squared of the distribution to determine the rate of evolution. If it was just a delta function at s , then it's just an s squared.

So then you think, oh, it's going to be mean of x squared. Right? But if it's a delta function, then the mean of s squared and the mean s squared are the same thing. But in this situation, it's just useful to play with some different distributions and see how it plays out.

But in this case, it is, indeed, mean s squared. So the rate of evolution in the limit of no clonal interference is, actually, rather simple. It just goes as μN . But then, you have to take the expectation of s squared of whatever this distribution is of underlying mutations. OK?

But of course, this is going to break down at some point, as everything does. And can somebody remind us why it is that it's going to break down? Clonal interference. Perfect. And we know exactly where clonal interference is going to start being relevant here.

In particular, what we can imagine drawing is something that's [INAUDIBLE] the rate of evolution as a function of N . So rate is a function of N . And we might even want to plot in a log, log scale. So we'll say the log of the rate. The log of N .

And at the beginning, what do I draw here for small n ? A line with slope 1. OK. If it had depended on n squared, then what would I be drawing here? A line with slope 2.

Yep. Don't mess that up because it's really easy to. OK, so it's a line with slope 1. Right? So at the beginning, for small population size, if you double the population

size, you should double the rate of evolution. But this can't go on forever, and it won't. So it's going to curve over somewhere. all right.

AUDIENCE: How can the rate go down [INAUDIBLE]?

PROFESSOR: the rate goes down relative to what it would have if it had continued. And that's because you're wasting some beneficial mutations. With clonal interference, what's happening is that you have a great mutation over here but also a great mutation over here. And only one of those great mutations can win. So with clonal interference, you're, somehow, wasting some of the beneficial mutations that you acquired.

AUDIENCE: Even though you're always taking the maximum because what you're doing is you're [INAUDIBLE].

PROFESSOR: Right because the alternative would have been to take the sum of them, which is what happens over here. If they're not competing, then you got one. And you get the other. And you just get higher and higher.

With clonal interference, you indeed take the maximum. But that's lower than sum. And that affect just gets worse and worse as you get to larger population sizes.

So although it's relatively simple to calculate the rate of evolution in the limit of small populations, the rate of evolution in the case of clonal interference is, actually, a very hard problem. Hard, well, experimentally, theoretically, and in all ways.

I just want to say a few things. There have been some really, I think, interesting studies occurring over the last 10 years trying to get at this regime. So the question is how should it behave. And there's a paper by Desai, Fisher, and Murray when they were all at Harvard.

Since then, Daniel Fisher, the [INAUDIBLE], has moved to Stanford. Desai went to Princeton but then came back to Harvard. So he's now Harvard faculty. So this is current biology in 2006, '07, '08. It's 2007.

So they have, kind of, a simplification of this where they asked let's just assume that we don't have a probability distribution of beneficial mutations. Instead, let's just assume that there's some mutation rate to acquire beneficial mutation that's exactly s . OK?

So you say, all right, well that sounds super simple. Of course, it's not true. But even that problem is hard. But what they can do in this regime is that then it's nice because everything's, kind of, discrete because then the population is going to be described by a series of-- so this is abundance as a function of the fitness.

So this is, maybe, the bulk minus fitness relative 0. Here, this is s , $2s$, $3s$, $4s$, and maybe there's a little bit here at $5s$. OK? So what happens is that there's going to be some equilibrium distribution of this front or the nose of the population.

And at some rate, these guys get mutations where some individual, kind of, comes like this because it gets a mutation that's beneficial. But that doesn't actually affect the dynamics very much because all these guys are growing exponentially. What's really relevant are when individuals that have several or more of these beneficial mutations than the rest of the population, when these guys get a mutation that they can move forward.

And it's this dynamic that is, kind of, pulling the population here. So you can actually do analytic calculations on this model that you would not be able to do if you did a full distribution of mutations here. But it's actually still complicated and hard. And I do not claim to have gone through the full derivation because the full thing is they get that the velocity in this model approximately equal to. By then, it's an s squared.

All right, so you might sneer at this model and say, oh, well, you know, this is an oversimplification, blah, blah, blah. But even this is hard. And you get a complicated expression. And it's describing something fundamental that's happening in these populations with clonal interference.

The important thing is that you can-- more or less, this term is going to be the dominant one in many cases, especially because we're interested in the large end

regime here. So what you see is that, for large N in this model and they did experiments in this paper that are consistent with it, they find that the velocity-- the rate of evolution-- this is the rate.

You see her the s squared term that we already talked about. Right? And there's no probability distribution, so it's just an s squared. But what you see is that it ends up going as $\log N$ for large N . So this is when there's a lot of clonal interference. So maybe this thing goes as $\log N$.

But I'd say, maybe, this is not an open and shut case because real populations are more complicated than this because it's not that every mutation has magnitude s . But I think it's a reasonable first order model. And I think this is a nice set of calculations to make sense of things. Are there any questions about that calculation or why this thing curves off [INAUDIBLE]? Yes.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Yeah.

AUDIENCE: [INAUDIBLE]?

PROFESSOR: Sure. OK, so the question is why is it that it's average s squared instead of average s squared? So what is it that's useful to say? I'm trying to think of a nice probability distribution where those things will have very different--

AUDIENCE: [INAUDIBLE]?

PROFESSOR: For a delta distribution, they're the same thing. So if we wanted to do an intuitive explanation, maybe, [? probably ?]-- I'm trying to think if we had two delta distributions. So the mean s would be here. Yeah, OK.

So imagine that-- I know that this is not a real beneficial mutation so maybe but OK. But it's something that's small. You know, negligently small here and magnitude s over here. So the mean s is, indeed, equal to this s_0 over 2. Right?

Whereas, the mean of s squared is, then, s^2_0 over 4. Are we looking at

something different? Oh, no. OK, I think I'm going to have to come up with a better example to answer. So maybe after class we can come up with one. I don't want to take five minutes finding a good explanation. Yeah?

AUDIENCE: [INAUDIBLE]?

PROFESSOR: Yeah, so this is the rate of evolution in the regime of large population size.

AUDIENCE: OK.

PROFESSOR: And I don't know how small you have to go before [? funny ?] [INAUDIBLE].

AUDIENCE: That's only for the [? large ?] [INAUDIBLE].

PROFESSOR: Yep. Yeah, and, of course, you could imagine that taking various limits is complicated here. But the important thing is that the rate goes as $\log N$ for, in the case, a clonal interference because that's the key thing to remember, besides the fact that they actually had to work to analyze this model.

OK, so what I want to do is, in the last 10, 15 minutes, to talk about this [INAUDIBLE] paper because I think it's pretty. And I mean, it's an elegant example of how, if you look at a problem in a different way, then you can get, I think, really interesting insights using a minimal amount of, like, measurements.

I mean, basically, how many measurements are in this paper?

AUDIENCE: 32.

PROFESSOR: Basically, 32. Right? So what they are doing is they're analyzing mutations in the enzyme beta lactamase or the gene encoding beta lactamase, which confers resistance to beta lactam drugs like, in this case, cefataxime

OK, so this guy gives resistance to these beta lactam drugs that are like ampicillin or penicillin. In this case, cefataxime. but not all of the versions of this gene or the enzyme, actually, can break down this new drug. So all 32 versions of the enzyme that they study break down, for example, ampicillin. But they had widely varying

levels of resistance or ability to break down this drug, cefataxime.

So they wanted to try to understand something about what happens if you start out with this base version of the enzyme. You know, if you just look up what's the sequence for beta lactamase. What's going to be the sequence? And that we're going to call the minus, minus, minus, minus, minus.

And we might, reasonably, want to know how does it get to the version of the enzyme that has all five of these [? point ?] mutations? Does anybody remember what those five mutations were? Like, what kind of mutations are they?

AUDIENCE: [INAUDIBLE].

PROFESSOR: They're all [? point ?] mutations. And are they all protein coding mutations? No. So actually, this one here is actually a promoter mutation that increases expression by a factor of two or three. Whereas, these things here are indeed protein coding and change the amino acid. Protein coding. The amino acids in the end.

Well they each change one amino acid in the resulting protein. So what Weinreich in this paper was trying to understand is what is the shape of these fitness landscapes? And what does that mean about the course of evolution or the repeatability or predictability of evolution?

And I just want to stress this is the Weinreich 2006 because this version of the gene/enzyme is, essentially, unable to break down cefataxime at all. So E-coli that has this version of enzyme, it's almost as if they don't have any enzyme at all.

Whereas, this version of the enzyme is able to break down this drug, cefataxime, at very high rates. And indeed, the way that these things are quantified in this paper is we have what's known as the MIC or the Minimum Inhibitory Concentration.

And basically, you just ask-- oops, inhibitory concentration-- what's the minimum amount of the antibiotic that you have to add to prevent growth of the bacterial population after 20 hours starting from some standard cell density, OK? So it's a very easy experiment to do because, 96 well [? plate ?], you just have many wells.

And you just go down a concentration, maybe, by a factor of a root 2 each time. So then, you go across 12 or 24. And you get over a broad range of antibiotic concentrations. And what you should see is that this is, maybe, dividing by route 2 each time.

So you get growth here. You get growth here. Growth here but then no growth, no growth, no growth. And the concentration and that you added here is the MIC. What you'll see is that, depending on the version of the enzyme that the bacteria have, the growth will occur up to different concentrations.

AUDIENCE: Why would you [INAUDIBLE]

PROFESSOR: Why would you? Why is that--

AUDIENCE: [INAUDIBLE].

PROFESSOR: Oh, I'm just telling you what they actually did experimentally in this paper. You could do a factor of 2. Or it's just a question how fine of a resolution you want. How to do root 2? OK, so this is like a mathematicians question, all right because your point is going to be that--

AUDIENCE: [INAUDIBLE] route 2.

PROFESSOR: OK, so it's true. Root 2, it's an irrational number. It's the first proof in a analysis textbook. It doesn't matter, OK? Our error in pipe heading is a percent, which means that if you do 1.41, that's fine. Yes. So don't be paralyzed by petting a root 2. OK?

AUDIENCE: Is it always very sharp, this changing in growth?

PROFESSOR: You know, biology and the word always should never be used in the same sentence. I'd say that it's a reasonable [? assay. ?] It's, typically, sharp. It happens, though, that you get growth here. And then, you get stressed out because you don't know what to do.

I mean, the important thing is that you do the experiment multiple times and you

have some reasonable rule for treating these things. Yeah, it can be more complicated though. All right, what we're going to use in the context of this paper, though, is we're just going to assume that this MIC is a measure for fitness.

The mapping from MIC to fitness is, actually, very nontrivial. Something my group has spent a long time thinking about. But for the purpose of this paper, just when you hear MIC, you can just say think of it as fitness. OK, [? higher ?] MIC, he assumes that it could be selected for by evolution.

So there are, in principle, 2 to the five different states. Different versions of this gene. So what he did is he constructed each 2 to the 5 versions of the gene, put them into the same strain of E-coli, and then measured the MIC of each of those 32 strains.

And that was all the measurements in this paper, basically because everything else is just analysis of that resulting fitness landscape. But what's exciting about this is just the ability to have an experimental fitness landscape because we've talked about fitness landscapes for years.

But then, it tends to be much more like what you saw in Martin Novak's book, right? That you can think about these fitness landscapes. And you can do calculations of what should happen on them. But this is a case where we can actually just measure something akin to a fitness landscape, and to try to say what it means.

So you can ask questions about how rugged is the landscape. How many different paths can you take from this version to this version? So first of all, how many peaks were in this landscape that he measured?

AUDIENCE: [INAUDIBLE].

PROFESSOR: One peak. This is important. This was the one and only peak. And when you read this paper, you come away thinking, oh yeah, this is a really rugged landscape, right? Many of the paths were not allowed by so-called Darwinian or selective evolution.

But it's easy to forget that, actually, it's not that. It's a moderately rugged landscape because there was still one peak. In particular, if you just assume that the population starts at that minus, minus, minus, state, starts travelling uphill in fitness, gets a mutation and goes uphill, is there any possibility for it to get trapped in a non-optimal?

No because there are no other peaks. So there's only one peak. That's the same thing as saying that you can take any path you like going uphill, and you will always get to the same final location. You will never get stuck anywhere.

I mean, it does not mean that you can take any old path that you want. Many of the paths may be blocked in the sense that they may go downhill and so forth. But at any location you're at, there's always, at least, one path going up in fitness, up in MIC.

Doesn't matter which path you take. You will always be able to get to the peak of this landscape. OK? So it's not too rugged of a landscape, in that sense. Yeah?

AUDIENCE: Wasn't it, [INAUDIBLE]?

PROFESSOR: Right. So some paths are not allowed in the sense that some paths decrease fitness, locally. But what I'm saying is that you can take a different path that goes up in fitness. And you'll still get to the same peak.

AUDIENCE: Right, but you started in that [? wrong path ?] and you [INAUDIBLE].

PROFESSOR: Well no, no. The thing is that you can't take that path because that path goes down in fitness is the claim. So the statement from this paper is that, if we just assume that the only mutations that can fix in a population are mutations that increase fitness, then it does not matter which of those beneficial mutations you take because you will always end up reaching the peak. So there are 120 possible trajectories. Can somebody say how we got 120?

AUDIENCE: [INAUDIBLE].

PROFESSOR: Right. So this is 5 factorial. What he found by analyzing the resulting fitness

landscape is that 102 were selectively inaccessible. Oh, I don't know how many-- is that the right way to spell that? Yeah.

And what he's assuming is that the only mutations that can fix are beneficial mutations. OK, right. In particular, if you have two states, there's a mutation that you could acquire. From here, let's say you get this mutation, and that just leads to the same MIC. He assumes that that is inaccessible. All right?

You know, and of course, like all things, you can argue about it. What he's saying is that it won't fix in reasonable times, which is fair if it's really a neutral mutation. In particular, if you have 10^6 bacteria, and if the mutation rates are the same everywhere and some mutations lead to a significant increase in fitness, then a neutral mutation would be unlikely to fix.

So what he found is that there were only, then, 18 of the 120 paths off this fitness landscape that had monotonically increasing fitness values. And indeed, if you analyze those trajectories, what he found is that, actually, 18 [? isn't ?] maybe even over [INAUDIBLE] because only a few of those trajectories would likely occupy majority of what you might call the actually observed paths just because of the statistics of when those paths branch and so forth.

So the argument from this paper is that we can measure fitness landscapes. And from it, we can say something about the path of evolution, perhaps. Other people have since gone and done, actually, laboratory evolution on a different antibiotic resistance gene-- again, Roy [? Kashoney ?], actually-- to confirm that these landscapes, at least in some cases, can inform laboratory evolution.

So there is a sense that maybe evolution is more predictable than you would have thought. We're out of time. But if you have any questions, please, go ahead and come on up and ask them. All right? Thanks.