WEBVTT
00:00:18.750 --> 00:00:21.340 - All right, I see more people joining
00:00:31.960 --> 00:00:34.760 Jeff, how long do you have like an hour?
00:00:35.633 --> 00:00:36.466 Less than that?
00:00:36.466 --> 00:00:39.510 - I think I can probably finish in less than an hour.
00:00:39.510 --> 00:00:41.157 - Less than hour, all right.
00:00:58.180 --> 00:01:00.113 I think we should get started.
00:01:01.580 --> 00:01:03.270 So hi, everyone.
00:01:03.270 --> 00:01:06.810 Welcome to our seminar series on COVID-19,
00:01:06.810 --> 00:01:10.260 organized by the Department of Biostatistics.
00:01:10.260 --> 00:01:14.750 I’m very pleased to have here today, Jeff Thompson,
00:01:14.750 --> 00:01:19.750 Professor of biostatistics, Ecology and Evolutionary Biology
00:01:20.330 --> 00:01:22.523 from the Yale School of Public Health.
00:01:23.400 --> 00:01:26.670 Thank you, Jeff, for being here today with us.
00:01:26.670 --> 00:01:29.690 As usual, you’re welcome to write questions
00:01:29.690 --> 00:01:34.573 in the chat box or even unmute yourself, if you can,
00:01:34.573 --> 00:01:38.151 and other people are not talking.
00:01:38.151 --> 00:01:42.191 And, Jeff, why don’t you take it from here?
00:01:42.191 --> 00:01:44.817 - Okay, thank you very much for the introduction, Laura.
00:01:44.817 --> 00:01:45.650 I’m really pleased to have an opportunity to talk
00:01:45.650 --> 00:01:48.267 about the work that we’ve been doing.
00:01:49.164 --> 00:01:51.882 I think like many speakers in this series, you know,
00:01:51.882 --> 00:01:54.013 we’ve been doing a lot of work very hard
00:01:54.013 --> 00:01:56.993 on a short period to try to get some progress on COVID-19.
00:01:57.830 --> 00:01:59.300 Ironically, this is the first work
00:01:59.300 --> 00:02:02.900 I think that I started In response to the COVID-19 epidemic
and it’s turned out to be a lot of work. So it’s actually gotten the least far. We’ve done a little bit of work, for instance, on epidemic modeling of COVID-19. That’s already been submitted, I actually have some other work on quarantine and stuff that turns out to be really interesting and far along in the research. And then this work, which I started early on, which is more evolutionary, and looking at the zoonotic process has gone a little bit slower. What that means is consistent with many other speakers in this series, I’m gonna be talking a lot about the methods that we’re going to be using, which are well developed, and what we’re planning to do, I don’t have a lot of results. But I think that’s consistent with these talks in general. So hopefully, that will be of interest to you and also be illuminating in terms of possible research approaches towards this kind of work. So as Laura mentioned, I use a lot of evolutionary approaches to do my analyses of things. And the title of this talk is model averaged estimation. of molecular evolution and natural selection in SARS coronavirus, one and SARS coronavirus two
two Corona viruses during the zoonotic period.

So what was attracting my interest in this particular case

is that it’s usually very difficult and challenging to find.

And I’ll get to this later in the talk to figure out what’s going on during the zoonotic period,
because you don’t usually get much sampling there.

So, what I wanted to do was apply some techniques that I’ve developed to this problem.

And I will get to those techniques and the application to this problem.

But I first just wanna give a little bit of introduction,

I think, maybe from a statistics point of view towards some of the methodologies that we’re using,

just so everyone can sort of see on board at least how I see this as contributing to interesting statistical questions.

So and in a broad sense, if I can get this to Move forward.

Here we go.

I think one of the most intriguing and interesting and challenging areas of mathematics is understanding this border between the discrete and the continuous.

So these are just some one particular example you can pick out is, if you look at discrete and continuous distributions that are frequently in use in statistical probabilistic analyses,
00:04:21.360 --> 00:04:25.240 we have the geometric and negative binomial distributions.

00:04:25.240 --> 00:04:27.840 And we have the exponential and gamma distributions.

00:04:29.809 --> 00:04:31.906 These are basically essentially waiting for discrete events

00:04:31.906 --> 00:04:33.340 when you have a probability over time.

00:04:33.340 --> 00:04:35.217 We’re waiting for the earth event if you

00:04:35.217 --> 00:04:36.709 have probably over time,

00:04:36.709 --> 00:04:39.160 and they correspond to the distributions on a continuous

00:04:39.160 --> 00:04:42.450 time for the wait for the first event

00:04:42.450 --> 00:04:44.650 or the wait for the alpha event.

00:04:44.650 --> 00:04:46.330 So there’s a real clear correspondence

00:04:46.330 --> 00:04:47.670 between these two distributions.

00:04:47.670 --> 00:04:49.690 And you can actually see in the mathematics,

00:04:49.690 --> 00:05:00.820 how they’re similar as well.

00:05:00.820 --> 00:05:03.840 and the reason why I say it’s interesting is

00:05:03.840 --> 00:05:06.600 because often many of the biggest problems I think

00:05:06.600 --> 00:05:11.288 we wrestle with in statistics are when we’re trying

00:05:11.288 --> 00:05:12.830 to deal with data that is some intermediate

00:05:12.830 --> 00:05:15.290 level between continuous and discrete,

00:05:15.290 --> 00:05:21.010 and where we’re trying to figure out which

00:05:21.010 --> 00:05:21.843 that I think that’s a very interesting area.
And the technique I’m gonna tell you about is definitely wrestling with exactly this kind of question.
What kind of question do I mean? Well, I mean, questions that deal with state spaces,
over time, or over any discrete or continuous axis. And you can see in this diagram just give you a picture of the kinds of problems that one deals with between discrete and continuous measures.
You can have here it’s depicted as time, you could have a discrete state space,
you’re measuring over time, you could have a continuous sorry, you’re gonna have discrete measurements over where you’ve got discrete time in a discrete state space,
you could also have discrete time and a continuous state space.
You can have continuous, continuous or you can have discrete, continuous.
And this two on the bottom are, two on the left, sorry, are the relevant ones for what I wanna talk to you about.
In my research, which is largely focused on informatik data that we can obtain from sequencing
on informatik data that we can obtain from sequencing or other approaches like that.
A lot of what we’re trying to do is look at these discrete linear sequences that have sites DNA sites or amino acid
sites and trying to understand is there some pattern in those sites that allows us to understand something about the biology of the organism or the biology that we want to know something more about?

So what essentially I’m gonna be doing is telling you about an approach that takes essentially discrete items over some X axis, in which case in my case, it’s always going to be sequence space, like the nucleotides or the amino acids of a sequence. And turns it into these kinds of more discrete models. And then in some, in a procedure that I’m going to tell you about actually gives us more of a continuous measure over that space, it’s not completely continuous, but when you work with hundreds of sites, it turns out to look very continuous in terms of how it appears. But it’s done with a discrete model that looks over multiple sites. So well, I’ll tell you how it works in a moment. And I hope it’s of interest to you guys.

Just to introduce that, in general, the lab has worked on a lot of different kinds of data, and including things like gene expression data.
that borders this discrete continuous measurement.
The old micro arrays we used to use give us essentially continuous measures of gene expression. Now we get discrete counts from our census sequencing approaches. Then all the sequence data we work with often ends up being essentially clusters of sites and various kinds. And then we also use a lot of phylogenetic inference, which is another kind of just discrete modeling in terms of the topology, but the borders between these two because we have discrete modeling of the topology, there are certain topologies that the taxa that we’re interested in looking at that show their relationship to each other. At the same time, there’s also a continuous measure out of that, which is these branch lengths, or how diverge these different tacks are from each other and constructing the phylogeny. So this sort of border between discrete and continuous measures, always sort of plagues and intrigues me, I guess it would be the question. Okay, so what am I gonna do today? What I wannado today is talk about maximum likelihood model averaging to profile clustering of site types across discrete linear sequences. So at the very base level,
how do we take kind of these discrete sequences of amino acids or nucleotides and understand whether sites are closer to each other or farther apart from each other. this is the question are they just uniformly distributed site types across a sequence? Are they clustered close together or far apart? Secondly, I’m gonna talk about how we can use that approach to understand whether sites are under selection in a gene expressed in a sequence. And what I mean by under selection is that, mutation alone would produce that implies that every time that mutation is happening, it’s spreading across the population. And that’s why you see that uptick in the rate of change of those sites. So we can actually use this clustering approach to identify regions of the gene that have that sort of uptick and I’ll explain how we do that. Now lastly, I’m just going to show you a very few slides
on the title of the talk, which is this model average estimation of the molecular evolution and natural selection in SARS Coronavirus one and SARS Coronavirus two during the zoonosis.

So by the time we refer to these, I'll just let you know we're almost done with the talk.

All right, so to talk about the first one maximum likelihood model averaging five clustering of sites across the street linear sequences. I just want to... (phone ringing)

Sorry, emphasize that we wanna figure out whether site types are clustered within a linear sequence.

This sounds like a very straightforward statistical question seems like something that should have been addressed many, many times in the statistical literature.

Much to my surprise, it's actually not terribly well explored.

You have a linear sequence, it's so long and you have site types of one type another are they clustered next to each other? Well, if you know the bounds of the region of interest, if you can describe oh, it's I'm interested in this domain right here, from site to site 90 or some other description.
If you know the bounds, it’s very simple to analyze that kind of data. You can just quantify the site type proportions within and outside those bounds. Use something like a straightforward Fisher’s exact test for significance extremely simple problem. But what if you don’t actually know those bounds? What if you don’t know even what you’re looking for exactly? You just know you’re interested in concentrations of one site type compared to another site type across some discrete linear sequence, like this series of zeros and ones you see below. There’s one, zero, zeros, there’s one, zero, ones, there’s periods where ones are closer to each other a series of ones are closer or farther apart from each other. How should we figure out whether things are actually clustered in that site? Or are they random? So if you don’t know exactly where to describe, or what size you’re looking for, the most common solution people use is some kind of sliding window, they take a window over the series, and they slide it across and say, "How many are in this window?" And then you can come up with based on the sliding window a sort of diagram of the clustering.
in terms of like you see peaks where there’s a lot of clustering and valleys where there is none. However, significance testing with that kind of approach is often awkward to construct. Due to a strong or autocorrelation among this URL overlapping windows. And of course, if you just sort of take windows arbitrarily from one location to another, then you’re really instituting, (indistinct chatter) then that causes problems. Because what if the cluster is really on a border between two windows, so you have to slide it over and then you have the autocorrelation. And it becomes actually statistically quite challenging to sort of account for all of those auto correlations. Secondly, they need to specify that window size itself presents a user with a procedural ambiguity that almost inevitably leads to post hoc selection of window size. Size and can mislead inference that is just the fact that you have to choose a window size. And if you don’t actually have a good arbitrary outside reason to choose it. It’s very hard not to choose a window size that ends up validating your hypothesis in some way. So it’d be better if we could just have an approach
that does not require us to place in some arbitrary parameter that gives us a window size.

So in order to address this question, a postdoc of mine, John John, who you see below work

with me to address it.

Oh, I wanted to say one other thing, which is that, yes, this has been addressed with some nonparametric methods that people have developed,

including some rather famous people like Sam Carlin.

And these are methods that do not assume prior knowledge.

And they’ve been suggested to detect this clustering and discrete linear sequences.

So you can do runs tests that look for the longest unbroken run, or the variance of the run

links across the entire sequence.

Both of these are indicators of clustering.

Unfortunately, both of those are using are not sufficient tests.

And those they don’t use enough of the information to say that you’re actually have as much power as you’d like to do the analysis.

And that’s because if you use like the longest run link, for instance, of course,

you’re only really using a little bit of information about the entire sequence.
And of course, you’re really missing anything like the cluster of ones that are have a bunch of small clusters that are all next to each other interspersed with a few of the other type, so the longest unbroken run doesn’t work well. If you use the In terms of power, if you use the variance of long run link that gets rid of the fact that you’re looking for just one. But unfortunately, a variance doesn’t tell you anything about the relative position of site that are of the same type across the sequence. So the fact that this one, one, one, one here is close to the one, one here, and the one another is, and this the fact that these are all close to each other, does not give us the power that it should for understanding this region may be under maybe cluster. So variants of run length is also an underpowered approach. The most powerful approach that’s been published out there, aside from the ones we’ve been working on, is the empirical cumulative distribution functions to sick that’s where you sort of go across the sequence and just say, "oh, okay, we’re accumulating ones here, we’re shooting more accumulating more.” And there’s fortunately a number
of highly developed statistical approaches to look at the empirical distribution and figure out whether you see an increase beyond expected during some period during that ECDF, the power is better than either the previous methods, but it’s still not very strong. It’s not clear that it includes all the information that it should. And it can be affected. Research has shown that it can be affected by the location of the cluster, which is not desirable. So if you have a cluster on an end, that has less the ECDF will have less power or more power compared to a cluster in the middle. It’s also challenging to interpret in the end, for reasons I’m not gonna go into right away. So what did we do? What we did was develop a tripartite divide and conquer approach to model variant sites based on iterative sub clustering. And I’ll describe it in detail right now. I’ll just tell you the plus and the minus of this approach at the beginning, which is it’s sort of a bioinformatics approach and that are bioinformatics statisticians approach and it uses intensive computation to solve the problem instead of giving a strict analytical result. And in fact, what it does is it just says,
Well, if we're interested in clustering in any case, clusters should be represented by increases in the probability within some cluster central region compared to some side regions. And if we define CS and CE to be anything from the very beginning to the very end of the sequence, it encompasses all possible single clusters within a sequence.

So, for instance, if the cluster were on the far left we can just define CS to be at zero, the left hand cluster is nothing and the right hand area that has depressed in variant type intensity would be the other category.

Anyway, so, what we can do is divide any sequence into three sections, just count up the number of site types in each one, estimate the maximum likelihood probability for the site type to be of the variant type of interest, say it's a glycine amino acids within a protein or add mean nucleotides limited gene, whatever it is.

So then you can just come up with a null hypothesis, which is the likelihood under the hypothesis that these things are located at random across the whole sequence. And then an alternate hypothesis that allows that is invoking a model which involves more parameters, which then separate separates into a clustered
versus non-clustered state.
So that would be fine if what we really expected in a sequence was one cluster,
compared to nothing else,
compared to the sort of baseline rate of clustering,
sort of baseline rate of variant types.
And but what we really want is an approach that can take clustering at many, many levels.
So what if there’s a cluster within the cluster or cluster within left?
So what you can do is then take each of these sub clusters you’ve identified and actually do the same process on them looking for whether there’s a higher likelihood of the data given another cluster somewhere within this sequence, et cetera, et cetera.
Now, if you think so this sort of dictates a procedure,
which is that you start, you input the sequence, you start at, you know, the first at the left and move all the way to the right, essentially, you find the most likely cluster among all the possible clusters.
If the cluster is statistically significant,
you then sub sequence each of those three parts,
the left hand part, the central center part and the right hand part, find the most likely clusters within each of them.
And proceed doing this until you reach a point
where you can no longer find any statistical evidence
that there is continued clustering within it.
And that’s the point at which you stop.
And then what you can do.
And this, I think, is sort of a key because
00:18:38.500 -- 00:18:41.780 at the end of that, what you get is one discrete diagram,
kind of like that diagram I showed you initially,
where it proceeds flat, goes up,
proceeds flat goes down, et cetera.
I’ll show you an example of that in a moment.
But what you really wanna do possibly,
right, what I think is really appealing about
this approach is that then you can take
that as one model, the most likely model and you can look
at all the other possible models
that you could have constructed.
And you can use AIC weighting to actually figure
out how much you should believe what is the weight
for every possible model.
And then you can average across those models
to give you a continuous description
of how much clustering you see across the sequence.
And again, the advantage that I mentioned
early on about this,
from my standpoint is I haven’t put in anything
about how big a window how big a cluster,
I put in nothing about what I’m expecting
to see out of the sequence.
I’m just asking, what’s the most likely description of this given the assay penalty for parameterization and what the result gives me.
So then we have a bunch of different weights for all our different models.
And what it gives us something like this.
So on the top, I’ve shown you the AIC model selection which is the first thing I showed you if I just took the most likely description of this particular sequence.
It’s not important what it is it’s PRF ADHD, which has been widely studied in evolutionary biology.
But if you take this model selection would, the most likely description given that sub clustering looks something like this where we have a region with fairly high concentration of polymorphism, in this case, a valley, a region, an intermediate level, a point where we have a lot of polymorphism.
And then it moves and changes across the sequence.
Now, if you then instead take not just that one model, but a series of models and do the AIC model average, you get a much more continuous description across the sequence of what the probability of sight types being different is.
And that enables us to ask a question that’s a little bit more interesting in many cases, and I’ll show you how it enables us to ask questions about natural selection in a moment. So in particular, it allows us to get an estimate, you know, of what the probability is across the entire sequence. Even though we don’t have observations within the central region or this barren region here. We can still estimate what the model average, probably of a change of hearing in different places have this gene are and that enables us to ask questions that we otherwise could not do. All right, so that’s an introduction of MACML. It’s like this is actually published work, so you can find it. But compared to the ECDF statistics, that approach I just showed you has greater power to detect heterogeneous clusters it identifies clusters with greater accuracy and precision based on the Kullback-Liebler divergence between the actual distribution of the observed distribution, sorry, the actual distribution and the inferred distribution. It has better power and accuracy across different levels of clustering, better power and accuracy across.
different sequence links, and better power and accuracy and finding multiple clusters compared to a single cluster. The disadvantage is, it’s extraordinarily computationally intensive, and it is prohibitively so for very long sequences. So for genes a very long length, we can’t actually run it on the full-length gene and we have to do some more heuristic processes to crunch those genes into smaller size. Which we then can analyze and then build them up. Again, I won’t go into those at the moment. But the point is that at certain links, it gets just computationally too intensive to go through all the possible models that could explain the data. Now, I’ve talked about the maximum-likelihood averaging to profile clustering of site types across discrete linear sequences, introduced that methodology to now I’m gonna talk about how we can at apply that methodology to get us a better idea of which sites are under selection using a what’s called a pause on random fields approach. And don’t worry about that terminology. You might know it from statistics, it has to do with a particular observation in molecular evolutionary biology, which is why they’re using it.
and it's not really important for this talk, why it's called that. So let's go on and go ahead and do that talk about the model-averaged site selection using Poisson random fields. So first, I need to give you a little bit of background in the evolutionary biology for those of you who haven't had a lot of biology, so you understand how this fits in with what we tend to do another strategy. Of course, evolutionary biologists are often very interested in understanding what things are under selection. And in the context of this talk, why is that important? Well, we'd really like to know what things are under selection in the COVID epidemic, because we'd like to know what sites are actually causing the COVID epidemic to spread more or not, and what sites may have been important in it prior to zoonosis, MSN, perhaps, especially in the context of this talk, what sites were selected during that zoonotic process that made this virus perhaps able to infect humans in the first place. So what we're doing is, so to give you an introduction, I just wanna mention that they're sort of ways to look at ancient times and understand
whether selection was happening. And that’s this approach that’s called that looks at phylogenetic divergence, looking at multiple sites and saying, "Oh, we have a whole bunch of phylogeny of how these organisms are related.” And then we have a bunch of sites that are for each taxon. When we see sites like this, for instance, that’s having A and then a couple C’s and then a G and another tacks on, we know that this site changed twice on that phylogeny, at least right? So it changed to probably change from C ancestrally to an A in this lineage and to a G in this lineage independently. And so the fact that it changed twice means that it’s got an elevated rate of change. And that elevated rate of change is an indication that there’s been positive selection for change. It’s especially likely in sort of pathogen hosts because pathogens are changing in order to not be recognizable by their hosts. And often the host has recognition proteins that are changing to still recognize the pathogen, even the pathogen is changing. So these high rates of evolution are very strong indicators of selection in host pathogen situations.
So this is one way to study a natural selection. It does depend, though, on having a lot of data going back in time because you’re actually reliant on these changes occurring in multiple places on multiple lineages.

Now, a more recent level, and I’m going to go back to the middle in a moment. But a very recent time, you may have heard of selective sweep detection, a couple of methods people use are tajima’s D, IHS, there’s a bunch of other methods that are out now. And the idea there is to look at polymorphism. And if you look at an individual, before selection, this is sort of just a idea diagram, not what you look at. But so if you look at an individual who has a variant, and what you see in a population is that one individual with variant, a variant that’s important as somehow swept across the population. So if you see this would be before selection, there’s a lot of variation at a particular locus in the genome after selection, that one individual’s variant which contributed to the reproductive fitness would then imply that they would spread across the population. And if they spread across the population, then the genetic variants that were present
in that original individual spread across the population as well along with this selected site, and so you can look for this kind of partial or speedy. And the selection is going on neither of the approaches that I just talked about or the approach that I’m doing today. So I just wanted to introduce those, so you knew those are different. And they’re different because we’re looking at a more intermediate timescale. That’s like the sweet detection is purely dependent on polymorphism in the population, dependent on this ancient changes that you get from a phylogeny understanding how different species are related to each other at an intermediate level, our methods use that use both the polymorphism and the divergence. And the idea here in the McDonald-Kreitman approach, and the master approach I’m going to tell you about is that the polymorphism what you see generally in the population is sort of consistent with this. Sorry, if I go back to this slide. With this before selection, you know, all of these blue sites are assumed to not be under selection,
and that generally what we believe in evolutionary biology,
because of empirical data that validates it
is that most sites that you find varying in populations
are not under strong selection.
If they were on stronger selection,
they would probably fix it, everyone would have them.
And if they were under negative selection,
they wouldn’t rise to a high frequency.
So generally speaking sites that you actually see
change differences between us and our genetics
typically are not affecting anything.
Of course, we spend in our...
In the media, you only hear about the changes
that actually affect things.
And that’s because those are important to us,
the ones that don’t change anything
we don’t really care about.
So nobody talks about that much.
But most of the changes within population or
differences
within population don’t have much material effect.
So under that hypothesis,
then when you look at polymorphism,
most polymorphism is just an indication
of the underlying mutation rate,
some mutation happened didn’t have any effect.
It’s drifting up and down in the population.
And so the advantage of that is if you know
that polymorphism is signal is a signature
of just random mutation, it gives us an estimate of the underlying mutation rate, which we can then compare to the divergence and using that comparison, we can understand how organisms are related. So whether organisms are under selection or not, if the divergence is high compared to the polymorphism, that indicates a lot of selection.

That means (indistinct chatter) in the timescale of the analysis you’re doing, we have a lot of change the population, and on the other hand, you have a lot of polymorphism and not that much divergence, then that indicates you’ve got a lot of change going on, but it’s not actually being directionally selected because the divergence is much lower. So how does that test work in practice?

Well, just to step back for one moment, so we’re gonna apply that kind of test. In this talk I’m applying that test to the emergence of COVID-19. I’m actually applying it but also to SARS, which is closely related the SARS coronavirus one because we have similar data and can apply the same test in the same way to that data set. And we’re using in addition the SARS like Coronavirus in a sample that had been sequence basically collected from bats. Over the past 20 years or so,
so what you can see here is a phylogeny, which includes COVID-19 epidemic ongoing now in humans,
the SARS epidemic, which caused some 400 deaths or so back in the early 2000s.
And what we’re doing is analyzing both and looking at,
in particular, the very short internode here were between the most closely related non human infections
and the human infection set that we can see.
And this internode here, also, between these non human infections and the human
infections we can see, because the changes may have enabled, we don’t know,
there may be no changes that enabled it,
maybe this virus throughout its entire history could have infected humans,
but it just never managed to or never did.
But if there are changes that are unique to this virus
that happened during zoonosis, enabling it to infect us,
they happened on this lineage,
and so we’re interested in seeing what those changes are.
And so that’s what we’re gonna do is we’re gonna run this polymorphism and divergence approach on this lineage.
And what I just want to make (indistinct chatter) clear to you is the reason
why the polymorphism divergence approach is important is
the phylogenetic approach, the ancient approach relies on a large clade of data, which we don’t have
for that particular lineage here,
we just have the human infection, which is no longer zoonotic.
And we have this one lineage.
And so what we can do is ancestrally reconstruct the ancestor of this lineage, which is right here, actually on the phylogeny,
and we have this one lineage.
And so what we can do is ancestrally reconstruct the ancestor of this lineage, which is right here, actually on the phylogeny,
and also the ancestor right here,
and then use mass PRF, this approach that’s based
on polymorphism in the room, so I’ll explain to you
on the divergence between that ancestor and the first ancestor of all the human infections.
And we can take that as the near zoonosis time and figure out what mutations might have happened during that time.
All right, so we’re gonna do that in both the COVID-19 and SARS cases.
Now, how does this work in principle?
Well, there’s an old approach,
which is not what we’re using.
But I have to compare it to in order to sort of reference it in terms of the literature.
And that is that when you assume that polymorphism is neutral,
we expect a different proportion of replacement
to synonymous divergence compared to replacement.

to synonymous polymorphism in a gene.

So it’s just a two by two table here, again,

very simple statistics, where we look at

the number of replacement sites that are divergent.

the number of synonymous sites replacement.

again, is when an amino acid change

occurs in a DNA sequence.

DNA sequence changes can either change the amino acid

or not depending on what the sequence of the code on

the three base pair code on in the DNA sequences.

So if there’s a replacement, we tally it here,

if it’s a synonymous change, that doesn’t change the amino

acid, we tally it here, these ones are preserved.

Sometimes changes are presumably neutral because

don’t change anything about your protein.

And then the if it’s a polymorphic replacement,

then we see it here.

And if it’s a synonymous polymorphism we see it here.

So under the hypothesis that I mentioned,

all three of these cells should occur, it should be sort of changing in exactly the same way.

because polymorphic sites, whether they’re replacement.

are synonymous, we’re assuming are neutral,

synonymous sites, whether the divergent

we’re assuming is neutral.
The only one that apparently that under assumption is not neutral are these replacement changes at replacement divergence sites. So, if this replacement divergence, if the marginals add up so that this replacement divergence is sort of in line with all these others, then we assume nothing important. If this divergence is higher, though, then we might conclude that it’s under selection for changes at a rapid pace. So neutrality yields a DN over DS that’s equal to the PN over PS positive selection means that changes are actually being selected against at a high level indicates the DN DS is gonna be less than PN PS. All right now Let’s get to a little bit of the complexity on this thing that I mentioned that’s called Poisson random field theory, quantitatively estimates gene-wide selection intensity. So what you can do is take that same two by two table, and you can say under a model of selection, what do we actually think is happening here.
And that gives us the ability to estimate the selection coefficient, which is the rate at which change allows the virus to increase its reproductive ability.

And that term right here in these terms, these look complicated, but essentially, these formulas are just saying that the expectation for a synonymous change is essentially the mutation rate. These terms are about the sampling properties. Similarly, the polymorphic sequence is just basically dependent on the mutation rate. How the replacement sequences are a little bit more complicated in that they have to account for kinds of selection that may be going on. For reasons that I don’t wanna get into, the polymorphic selection, so both of them are depending on the mutation rate for replacement sites, and both of them depend on how much each variant is selected.
Selection doesn’t pack the polymorphism to a certain degree in the sense that if variants are moving through the population very fast, that can change how much polymorphism you see. But then if you use these sampling formulas, and the formula for the estimate of the strength of selection, given how many variants we see changing, you get these formulas for how much replacement divergence and polymorphism you expect to see. So this is a population genetics that was worked out by Stan Sawyer and Dan Hurley in 1992. The only change I’m making in this is pure F, instead of using a year which was how many grants that you see in the McConnell Craven uses it, I’m taking the probabilities of replacement divergence and the probabilities of some polymorphism and putting them in here. And the advantage here is that what I can do with that is what I mentioned earlier, I can go back to the old mass MACML approach sequence clustering approach that I mentioned before, estimating those probabilities across the entire gene, I can then estimate action across the entire gene by using these probability single site, I don’t have changes for single site. So what this allows us to estimate this gamma, minimizing likelihood of what.
gamma is to blame those problems exist, see.
So this is a very complex diagram of how this all works,
again, is a pretty elaborate method of computation.
But again, has the nice properties that I’m not putting
in any I’m not using assumptions
and not putting in any parameters.
They go in.
I just take the polymorph at the end analyze it for
weather sites are clustered into four different categories.
Again, replacement polymorphism.
That’s this arc here.
So polymorphisms anonymous divergence, placement divergence,
we cluster within all four of those categories.
We calculate the model average probability,
all those clusters and merge the data together.
I’m not going to go through the details.
But just if you were to do essentially the KML,
like clustering on those four categories
for a particular gene polymorphisms
and Ana’s polymorphisms, monster and placement divergence
if you plug those in, to the formulas I showed you before,
you’re basically plugging into these categories,
you can estimate those formulas.
And in the end, what you get is
an estimate of gamma across nucleotide positions in a gene.
I won’t go into what this result here, it’s an interesting result for reasons that are only of interest mostly to evolutionary biologist, but you can see here in this particular gene.

There’s a lot of variation in the selection intensity across the gene.

Now, that is actually really consistent with what we’d expect.

From a sort of basic biology standpoint.

Different parts of a gene are gonna either be very strongly selected to stay the same or they’re gonna change, you shouldn’t really expect.

That all parts of gene are equally likely to change.

And this gives a very nice diagram that allows you to understand how it’s different across the gene.

So if we compare this kind of approach to the McDonald kreitman tests, which again, are just putting in the DN DS, PN PS values into this two by two table, and I went through that, the important difference is that.

The Mk test assumes this intergenic homogeneous selection that in fact, a gene has the same selection across the entire sequence.

The problem with that is if you have one small region that’s under selection, the averaging out process across that entire gene can mean that you don’t detect the selection there,
even though it may be very strong for that small region.

And so the hope is that mastery graph can identify those regions much better than MK for instance, would.

And in fact, I went through this already. I'll just skip past this because I went through it already.

And this does do that. So this is an example of McDonnell Craven tests here applied to a Drosophila gene, which turns out to indicate high selection.

And you can see here that the DN DS ratio is about eight to one word as the PN PS ratio is almost even.

So this is a gene that’s under very strong selection based on the McDonald Kreitman test. Now, interestingly, so this one works with a homogeneity.

And then if you analyze the ACP 26 AA gene and look for the probability of all four categories. These are the four categories and of course, the replacement divergence here is the one that’s most likely to drive selection.

What do you get when you estimate gamma using this?

Well, interestingly, what you see is not something that’s under very strong selection across the entire gene,
but something that’s on moderately strong selection,
basicall in the second half of the gene, and then one peak of very strong
selection right around the middle of the gene. And this is visible in currents because
of a number of changes that occur in one particular domain of the gene here.
Now, if you look at just the replacement divergence,
you wouldn’t be able to figure this out. Because you see there are other
peaks along here. Those don’t turn out to be so important.
And the reason why they don’t turn out to be so important
is that the synonymous divergence synonymous by morphism
replacement polymorphism.
Tell us more about the underlying mutation rate
that says those elevations are probably have
something to do with mutation rate, and not necessarily
do with added divergence.
You can sort of see this elevation
on the right hand side over here compared
to the small dip right here and up here
and the way it all works out mathematically
is we can really see that there’s strong selection here.
We can also get what I call model intervals for this.
If you look across all the models,
what are the estimates of selection?

Possibly, what do we get is the 95% model interval for this?

And that’s what these very faint gray lines you may be able to see are those allow us to detect whether these are significant, least significant, statistically significant differences in selection.

All right, I’m gonna skip through this just because I want to spend the time but the point is, you can do this for other genes, and it shows similar results that allow us to understand where sites are under selection in that gene.

I’ll just cover a few more examples of how we’ve used this to give you an idea of what it can look like in a comparison between humans and chimpanzees where we’ve run this just to understand how we’ve diverged from chimpanzees.

We see a bunch of different examples here. Again, doing a little bit of comparison to that traditional McDonald Kreitman test and the mass PRF test.

Here you see a gene, which is statistically significant people’s point of view.

Based on the Mk tests, the four categories of the four tallies of which are indicated here. Here’s the mass -PRF profile, and it shows us again a particular region within this SLC AA
one gene that is under selection. There are interesting stories behind all of these, but I’m not gonna take the time to go through them.

Here’s another example where and this is an example where the McDonald pregnant test comes out is not significant. There’s just not that much divergence compared to the other categories. But if you do this, spatially with the MASS-PRF test, you actually see that a very central region there has very strong selection, and then the rest of the gene is under almost zero selection or almost no selection.

So this is an example I talked about, where you could have some very small portion of the gene under very strongest selection. And McDonald-Kreitman test wouldn’t detect it because it’s averaging over the entire gene. Similarly, you’ll get some genes.

Some jeans, here’s M gamma over here, where there’s a... Well, let me turn to that one last. Actually, let me look at TPH First,

there’s no statistical selection according to the Mk tests. And in fact, in our MASS-PRF, there’s no specific selection either.

the error bars are entirely overlapping zero here,
which indicates no selection.
Lastly, here’s M gamma.
This is the one of the very few examples
we were able to find where McDonald test did detect
selection, where MASS-PRF didn’t.
As you can see, there’s quite high tallies here,
which means there’s a lot of power
to detect selection if it’s there,
but it’s probably not very strong,
because the numbers are not all that different
from each other.
And McDonald-Kreitman says it’s statistically significant.
Now the reason why McDonald Kreitman is telling
it’s statistic’s nothing compared to mass PRF
is that actually, I didn’t explain this in detail to you.
But McDonald-Kreitman doesn’t actually assume
that there’s an elevation of rate here.
And so the significance here is actually driven by
the high polymorphic replacement level.
So there’s a lot of polymorphic replacements in there.
And what that means is there’s some other
kind of selection that isn’t a directional selection.
I won’t go into the details there.
But the nice thing is that in the examples
where we find that McDonald kreitman is statistically
significant and MASS-PRF isn’t examples
where in fact MASS-PRF is not designed to detect
that kind of selection and MK test is. In general MASS-PRF turned out to be significant in almost every case math MK tests were not.

Okay, so how can we use this, apply this to instances like COVID-19, the point of this whole talk,

and I'm just gonna give you one example first to justify why we think it's a good idea, because we don't have results on doing it, at least not many results on doing it to COVID-19 yet, and that is that we applied this influenza before,

which has some similarities to COVID-19, as everyone knows and in influenza, again, we're interested in looking across the gene are there sites that are under selection because those sites that are under selection are candidates where we need to be aware that in fact, vaccines need like for every year they design a new influenza vaccine, right?

And what they're trying to do is accommodate the fact that these changes occur on the sites that are actually susceptible to your immune system recognizing the influenza virus.

So we need to understand those sites that are changing and where they are in in order to design more universal vaccines that maybe could target sites that won’t change rapidly because they can’t change
because they’re structurally constrained in the virus.

So what we did was apply this MASS-PRF approach to influenza similarly on a phylogeny. I don’t have the phylogeny in the slide set, but the point is just looking at the ancestral influenza and its divergent sites within a particular region. And what we were able to do is identify a set of sites that are under selection using mass PRF that are beyond what people had prophesied as positive selection sites in the past. There’s a paper by Westgeest et al. 2012 which is essentially the gold standard for this and they found a bunch of sites that are all these circled sites in gray MASS-PRF. Also found those the orange diagram here is the MASS-PRF for this gene. And it also identified other sites that are under selection as well. And we’re in the process of understanding better how those can be validated. But the ultimate point is that these are important selected sites that may be relevant to the design of vaccines for influenza. So similarly, we’d like to illuminate which sites might be changing rapidly and under positive selection in Coronavirus,
not only during the human epidemic, but again during the zoonotic time period. And so now we’re finally coming to the final part of my talk, which is what we’re doing in terms of the model average estimation the mcos and natural selection in SARS coronavirus, one and SARS coronavirus two, Corona viruses during zoonosis. But the whole point here is really to explain to you what I’ve done because the results I have as I said are just a few plots of some of the longest selection we were able to check because we have to process through a lot more data before we get a more in depth look at the lesser selected sites that are on these genes. And so we looked at this for the Coronavirus. This is just a Coronavirus, Getty image that Yale has used looking at Coronavirus. This is just a Coronavirus, Getty image that Yale has used looking at Coronavirus. And again, as I mentioned, we’re looking at these two sites of where COVID-19 emergence occurred, and where SARS emergence occurred. And the question is, are there changes that happen there that are specifically responsible perhaps for those zoonosis and the only results I have are just a few results again, highlighting some of the strongest selection we saw.
This is actually a diagram of the spike protein which if you’ve heard much about COVID-19 molecular biology, you probably have heard about the spike protein, it’s what sticks out from the virus. It’s what grabs onto the AC receptor, and essentially is what most vaccines that one might design for the virus would target. And the point is that the recombination binding domain, which has gotten a lot of press already turns out to have the selected sites. You can see them here, here, and here. These are sites that are selected, meaning they’re changing rapidly during the pr zoneotic phase. So these are sites that are changing, not in humans, but in the bats in the pangolins. And whatever other animals that this virus is spreading among, or has been spreading among before the zoonosis to humans. So then the question is, are similar sites under selection during zoonosis? And during post zoonosis? It seems kind of similar, although we don’t get the same sites. So we have to do a little bit more molecular, you know, staring at this and understanding
it because these results are literally
I got these results today, actually.
So we have to sort of do more of this
and we actually can actually look at more depth
and get more sites with other approaches
that we haven’t implemented at this moment.
But during near zoonosis what you see is again,
the selected sites which are in bright red
are also on the sort of the visible side
of the recombination binding domain
of the spike protein, which is the tip
the outside portion of this gene.
Lastly, if we look post-zoonosis that’s in
the evolution of humans, we again see that
the selected sites are sites that are at this tip
region.
Again, none of this is terribly surprising.
The interesting thing is that it kind of indicates
that the zoonosis it kind of indicates consistency.
Again, there’s a lot more to do before
we can conclude anything like this,
but the idea we have right now indicates
a good deal of consistency between the selection
that’s ongoing in humans during zoonosis and pre
zoonosis.
And what that implies is that this may
well have been as I said, very briefly,
during this talk an instance where there’s a virus
just circulating around in bats and penguins
that could have caused this disease at any time,
it’s just a matter of whether or not we actually
have exposure to, to those organisms that allows the transmission to happen. Consistent with this, I’ll just mention a couple like verbal points, which is that all the evidence that we have indicates that this virus spread extremely quickly from the moment that it zoonosis into humans. And in fact, in most cases of zoonosis, we find that that’s true, which is somewhat counterintuitive. Obviously, it hasn’t adapted to humans, it has adapted to the amount of mammalian immune system. And so to the extent that our immune system is not tremendously different from that of bats or pangolins, it may be not surprising that it can infect us. But one of the things that is true is that if it did not spread very quickly, very easily from the very moment it transmitted to someone, it would probably lead to a dead end. In other words, if you don’t have an ability to transmit and spread just from the get go, the first person who gets infected is very unlikely to transmit it to someone else. So it sort of has to be well pre adapted for a zoonotic event to actually spread in humans. Now there’s, we need more zoonotic events,
to really get a better picture of that.

But the general result and the scientific literature does seem to show that zoonosis happens.

disease’s already well set to cause problems. And the examples that we don’t have where it happens like that, like MERS or like, well, MERS is a good example.

It’s a really deadly disease, but it doesn’t transmit well among humans. And so that’s an example where maybe it’s transmitting to humans, but it’s not transmitting among humans.

And it’s very hard for that disease to catch on within the human population and do human transmission as opposed to zoonotic events. And that’s because it doesn’t transmit and it doesn’t usually evolve that ability to transmit over the short time that individuals might get infected.

when they get it usually from camels. Okay, so I’ve showed you those examples. I just wanna to mention what else we’re gonna be doing.

So I what I just showed you was actually the sort of SARS coronavirus to some sites that are under selection in search for Coronavirus two genes. This is the S gene right here.

That’s the spike gene.
We’re gonna be looking at that in SARS coronavirus, one and two, we’re also going to be looking at other genes in the genomes. These have other functions. The M gene, for instance, is a membrane gene. So it might be relevant to and the gene as well might be relevant to vaccine generation. Like if we could generate a vaccine that targeted those, maybe they would be unable to change at the same pace that spike protein would they might be more conserved. And that might be one approach towards developing a vaccine. That would be a longer term vaccine because one thing we have to worry about, of course with this Coronavirus, is and I have other research that we’re doing on this question, which I’d love to talk about if anyone’s curious, but you can estimate what the actual waning immunity of it is, even though we don’t have data on that by Looking at other related species and using the phylogeny to understand how the waning immunity has evolved across the species and what the projected or most likely waning immunity of SARS coronavirus is, it’s, it tends to be it looks like it’s around 80 weeks or so.
So if we get about 8 weeks of waiting a period of immunity from this, that's not that much in terms of every two years or so we're gonna have Coronavirus coming around and in terms of we're going to be susceptible again to Coronavirus. Not that we're going to get it every two years. And what that would mean is that it's likely to persist as a circulating virus. And if it remains as deadly as it is that's a serious issue. So we're gonna really want to buy a vaccine. And we're not necessarily going to wanna have another flu vaccine that we have to get every year. So what we really want to do is target some genes that may be under more constraint than the recombination binding protein gene, the spike gene. And if you actually design a therapeutic
that interferes with the sites that are under selection
sort of in an opposite way, from vaccines, vaccines,
we really want to target something that just doesn’t change.
With therapeutics, we may want to target the changing regions, if we can design something that generically does, because those changing regions are functional.
In other words, those sites at the end of the spike protein are clearly ones that do bind the ACE gene.
It’s just that they’re flexible about what they are in order to bind it.
So we need to include all of those changing sites, if we wanna dissolve develop a therapeutic that for instance, would somehow interfering with the binding ofAce to receptors from the spike genes.
So thank you very much for listening to the ongoing work we’re doing on COVID-19.
I would love to entertain any questions that you have.
Let me just take one moment to acknowledge some of the people that I should acknowledge in this work,
I already showed you a picture of John John who was earlier the picture and the associated with the Mac ml approach that we developed many years ago 10 years ago basically
Yinfei Wu has been taking the lead on this project. She's a master student. Yano os Wang was an assistant was in visiting Assistant Professor Stephen Gaugham, is in the Evie department. Haley Hassler is in my lab, has been helping out with phylogenetics. Jayveer Singh is an undergrad who’s been doing some of the research work. Mofeed Najib produced those diagrams of the spike protein with the sites that we have identified as under selection so far. Zheng Wang is a long term collaborator of mine. Mofeed Najib has been doing the work we’re doing. Zheng Wang is a long term collaborator of mine, who works on nearly all the phylogenetic projects that I do, who’s works with me. And then Alex Thornburg is a long term collaborator of mine, now in North Carolina. He was while he’s currently at the North Carolina Museum of sciences, but he works on a lot of phylogenetic projects with me as well. And by the way, all of this, fortunately was recently awarded one of the NSF rapid grants to do this research. So we’re very pleased to have funding to continue to work on this as time goes on, which is good.
because it’s taking quite a lot of work
do the sequence wrangling.
And the analyses themselves.
As I mentioned, they’re computationally intensive.
So Alex and I were the PI’s on that particular
grant from the NSF.
So we’re excited to continue to do that work.
And with that, I think I would
like to entertain any questions you might have.
Thank you, Jeff, this was great.
I’m sure we have a lot of questions
who gets first?
Again, you can type the questions on the
chat box or just mute.
I have a quick question.
Okay.
- You mentioned or you touched a bit on this
before,
but how would this compare to cite wise estimates
of omega that you would get from Pamel
or similar program?
- So I’m sorry, I sort of was rushing at the end,
I didn’t explain that, in fact, I’m using pamel for
some,
So I’m using Pamela
for the pre zoonosis analysis, and for the post
zoonosis
analysis, because as I mentioned during the talk,
if you have a large phylogeny
with multiple branches, et cetera, et cetera,
where you can look over that entire phylogeny
then you
can get multiple changes at individual sites, which is what pamel actually uses to infer selection, right? You have to have the site change not just once but twice or three times. And then it says all that’s under selection because it keeps changing again and again and again. So, so Pamela allows you to do that if you have this sort of deep time or large amount of time and multiple lineages that you’re looking at, the master of approach that I’m using, enables you to do that on just a single lineage without needing multiple changes, I mean, multiple changes on a single language you can’t even detect because it just looks like one change if you have the ancestral sequence, which is what we do.

And if you have the descendant sequence, a changes to T, you don’t know if it changed to A to G to C to T again or if it just changed a to T, you have no idea you can just say it changed once. And so there’s no real way to run pants, there is a way but it’s really it’s statistically really underpowered terrible thing to try to run pamel on a single lineage.
00:57:44.164 --> 00:57:46.731 and figure out whether something’s under selection.
00:57:46.731 --> 00:57:49.320 The advantage of this approach is because it can use that polymorphism data, the data of like what’s just circulating in within populations as a metric for how much mutation is occurring.
00:57:51.382 --> 00:57:54.072 You can essentially divide out by that and again, because we’re integrating over all these models of how these things change, we’re essentially borrowing information from neighboring sites for what their rates of change are, et cetera et cetera.
00:57:59.390 --> 00:58:02.680 to estimate what the possible amount of selection is on all these sites.
00:58:03.544 --> 00:58:05.850 So by using the polymorphism data, and by doing this model averaging approach, we’re actually able to take individual lineages and estimate the selection on them.
00:58:06.879 --> 00:58:08.930 And that’s what we’re doing in the near zonosis analysis.
00:58:10.488 --> 00:58:12.837 that I showed you in the middle here.
00:58:12.837 --> 00:58:13.670 So there are different ways of doing the analysis. And it’s necessitated by the fact that we just have this one lineage and there’s no way it won’t be a single lineage.
00:58:14.770 --> 00:58:16.122 And that’s what we’re doing in the near zonosis analysis.
00:58:16.122 --> 00:58:19.263 that I showed you in the middle here.
00:58:19.263 --> 00:58:21.445 In any dataset we look at because for zoonosis, we’re going to have human sequences, we’re gonna have some animal sequences, we’re not going to know we’re not going
to have any information about the actual zoonosis. Even if we knew the first human, we could just take that as an estimate. We still probably need some data here. Maybe you could have the first human and the first animal that you got it from. That just doesn’t exist. We don’t have that data for any zoonosis. How would we would never be there at the moment. So we have to assume that there’s a number of transmissions among humans and a number of transmissions among animals during that near zoonotic period. And it’s just a single lineage. So we can’t really run pamel on that, in summary, because panel requires multiple changes multiple lineages to have power to actually infer evolutionary change. MASS-PRF fortunately, can do that, because you can look on single lineages. So you can use MK tests as well on single lineage is basically designed to look at single lineages. But the problem with MK tests, as I mentioned, is that they’re assuming the entire gene is under selection, which means it doesn’t give you the scope or understanding about recombination binding gene sites under selection or something like that. It often will just give you a result of the genes not under
selection, which is not true.

- Does that answer your question?
- Yes.

Great.

- Any other questions?
- I have one more if no one else wants to.

Great.

- Any other questions?
- I have one more if no one else wants to.

Sure, go ahead.

So in B cells, we have mechanisms that have mutation that specifically bias towards replacement mutations.

So in the absence of selection, the mutation mechanisms actually cause an Omega greater than one.

Would this have any way of correcting for that?

So the tricky part is, and I don’t know how it might,

the tricky part is not so much running the software,

which you could certainly do on that.

The tricky part would be identifying what polymorphism is, in the case of those cells.

So if you could identify sets of cells that are undergoing the mutation but aren’t under selection in some way, then you could use that as the proxy for the way we use it here.

is polymorphism within population polymorphism,

and then estimate that.

I just don’t know whether you have a way of doing that if you want to discuss it with me, we could.
That’s sort of always the key for detecting selection.
And it’s, you know, many of you may be familiar that I work on cancer and some of the work that I do.
It’s the same problem that I’m working on there all the time, I’m trying to understand what the baseline mutation rates of cancer in cancer and somatic evolution of cells are. Because if I understand the baseline rates, how often those things change, just the mutation alone, then I can always estimate selection. And that’s the thing we almost always want to know about in the analog analysis of sequence data.
So, again, it’s all about figuring out if there’s some piece of the data that can be used to estimate that polymorphism and an approach like this, the benefit of an approach like this would be, you know, maybe you can estimate that for some portions of the gene, but not others, you know, maybe then there’s a way that you could use this sort of model averaging approach to get at the underlying rate that it’s happening, even if you can’t estimate for that particular site, for instance. So I think the Might be potential to do it,
but it just depends, you know, about on whether there’s a critical, you know, set of data in what you’re looking at which I haven’t spent much time looking at back in the day.

So I wouldn’t know whether there’s some way of baseline getting that baseline polymorphism or baseline mutation rate, which essentially amounts to the same thing.

It just depends on whether, you know, you’re assuming the population is sort of has, you know, it’s just whether you’re looking at at a population level, or you have some sort of covariance matrix to better understand the mutation rates itself.

-I think there is a similar population B cells, Great, so I encourage you to look into that.

-Jeff, I have a quick question. I’m not too familiar with genome sequencing. But I think the Clustering Problem, the issue and the solution you have can be applied to many types of data.

So I’m kind of confused. So you start In the diagram where you describe the different steps, you said that you first pick the most likely cluster and then you essentially keep splitting the clusters, right?

How do you get the first clusters? Like there is some randomness in how you split the first?
Oh, so I'm sorry, I apologize. I didn't explain it in enough detail. The reason why it's so computationally intensive is we look at all possible exhaustedly. Now, I actually spent a year of my life trying to find a way to develop a Bayesian approach or some approach that would allow me to not look at all possible, you know, like to make this because because if you could do that, this would be a great way for doing tons of different things on very large data sets, right, large, like, and what amazed me is, I found that it was just an impenetrable problem. If I didn't look at every possible model, I could not, I always caused these biases, like there was no way to sample them. I even have ways of sampling the models according to their probability. But even that causes a bias because sometimes
There’s a large number.

So if you look at the, if you think about the set of models, it’s a very large set of models.

And there isn’t actually a huge amount of likelihood differences between these models.

That’s the thing.

So when you don’t exhaustively sample the models,

if you just sample some of the most likely models,

you actually are sampling just one corner of the space.

And it’s possible for a bunch of not quite so likely models, but reasonable models that are not in that corner to sort of be actually highly influential on the model average.

And so the bottom line is like sampling by trying to pick in the you know, most likely space doesn’t work.

And I could go into more detail about it.

But it turned out that I couldn’t do it any way other than exhaustive sampling.

So, I say that Sorry, I missed that mistake.

Actually, there are two ways of doing it.

One is to sample stochastically, according to likelihood, and the other is to sample exactly across all exhausted sampling significantly works.
In fact, it’s implemented in the approach that I was just showing, I’m sorry, I just sort of jumped too fast to say what I was saying.

So sampling stochastically works and sampling exhaustively work sampling stochastically is still very computationally intensive. But there’s no way to sort of, you know, important sample or do some sort of approach that would allow me to get a smaller set of models, which would then if we could do that, that could be really important, because then you could do this on more than like 2000 site, it’s somewhere around 2000 sites. So you start running into real problems with just too much computing computation time to make it worthwhile.

So we could extend this to 10,000, 100,000, you know, potentially really, really large numbers of sites, and really sparse sets of sites. If only we could find a way to bias the sampling towards models that are more likely without causing biases in the results. I couldn’t find any way to do.

This seems very much related to tree based
methods where essentially you’ve got, like split the space
and then you model of geology models,
like the random forest, for example,
or is very much related to that right.
Yeah, I have to say I was now familiar
with those approaches.
But when I was completely unfamiliar with it, yeah, I sort
of thought about it that way.
But you’re absolutely right.
Yeah, I guess the difference but here
you have a sequence like one sequence,
there you have a space.
So you just split in
different dimensions, but it is really good.
And I can mention, just to speculate,
I’m kind of interested in a number of
other ways of applying this.
So for instance, if the one I’ve been thinking about
and actually worked on a little
bit haven’t gotten very far with, but it’s like,
when you’re dealing with event spaces over time,
like if you have days, and you have individuals like,
prominent us in public health,
like individuals who are undergoing events
you end up with a very sparse matrix of events.
And so we use these approaches like survival plots
all these approaches that we use to sort of understand
how these rare events are happening,
and how people are changing over this, that event space is actually really sparse. But it’s kind of a matrix. And you could do this in two dimensions, not just one, right? So you could model average across two dimensions, and then you could get something that the thing that really appeals to me about that is that again, it’s really this approach is really, it only builds up from the this binomial event No, no event, stuff, a picture that’s very continuous over over the space and involves no assumptions about distribution whatsoever. So I’m just wondering if there aren’t instances where, you know, we could come up with a better understanding of what’s going on with individuals in a matrix such as that by using this approach. And it’s an approach that is that still works even with these sparse spaces, because you can model average over these tremendously large number of models that all have fairly likely fairly equal likelihood to get a result. So I don’t know that’s just a sort of a speculation that there might be some interesting approaches, ways to approach those problems using this kind of kind, of model averaging technique.
01:08:46.360 --> 01:08:48.870 - Great, I think we should wrap up.
01:08:48.870 --> 01:08:52.200 Thank you, Jeff, for this great presentation was great.
01:08:52.200 --> 01:08:54.843 And thank you all for joining today.
01:08:56.604 --> 01:08:57.930 See you next seminar
01:08:57.930 --> 01:09:01.283 is gonna be I think, July 14.
01:09:01.283 --> 01:09:05.430 So we'll send out invites.
01:09:05.430 --> 01:09:07.331 All right, thank you, Jeff.
01:09:07.331 --> 01:09:08.223 Thank you all, bye, bye.