Hi everyone, I'm delighted to welcome to our Fall Colloquium Series. Our speaker today, happy to introduce, Dr. Todd Gibson from the Department of Computer Science here at Chico State. Todd is an assistant professor in our Department at Chico State. His research interests are in biologically plausible models of biological network evolution and reverse engineering the evolution of empirically derived protein interaction networks. Ooh, you picked an easy one, huh? He graduated with a PhD in Computational Bioscience from the University of Colorado, Denver in 2009. He tells me, however, that he earned his Bachelor of Science in Music right here at Chico State. So I don't know if that means you saw the light or you ran from the light. But anyway, he ended up with a PhD in Computational Bioscience. Prior to pursuing his graduate degrees, he spent a dozen years in industry as a software engineer, manager and executive. We are very pleased to have him here today. He's going to talk about the role of mathematics, statistics and computer science in bioinformatics. Todd, take it away.

Thanks. Thank you.

[ Applause ]
Yes. So the view of the music degree was that I actually started here in computer engineering. The computer engineering program was just starting up and I was one of the few to actually be accepted and not be probationary accepted. But I had been through 12 years of school. I decided I didn't want to do anymore math which is ironic because I'm here in front of you now, and so I switched to music. Of course--

[ Inaudible Remark ]

Sorry, they're right next to each other in the catalog. Yes. So, I imagine this is probably going to be rather off the beaten path compared to some of the presentations that have probably been here earlier in the series. And that I'm really not going to be showing you a lot of math. I really am here to give you an idea of the breath of the field. Just pick some quirky things out of it. Some--yeah, I guess quirky is the right word and really just talk about it. So, you aren't going to see any Greek symbols. You might see one or two, but for the most you're not going to see a lot of Greek symbols or me going through a lot of formulas. It's really a high level presentation. So if you're expecting low level, then my apologies in advance.
And where I wanted to start this was with the Human Genome Project, probably among the best known of bioinformatics project. And this was begun in 1984, I believe it was. It was actually started out off the Department of Energy. And the Department of Energy of course has experience with nuclear power and radiation. And they have done some work with using radiation to break up genomes and to learn something about them. And so that was the motivation for them to begin this process. Of course, NIH thought they should be the ones to be doing the Human Genome Project because, by golly, it's a problem of health. And so there is actually in like Nature and Science and these journals, there is actually quite a discussion going, back and forth debate on who should actually be doing the Human Genome Project. And it ended up that they jointly were going to participate in this task. And so it kicked off in 1990. And as you can see here, it is predicted to be about three billion dollars to take 15 years. And amazingly enough, unlike so many projects in the world, it came in under budget and earlier than expected. And this is certainly due in part to the technology maturing as the years went by. The human has, what I've written here, three billion nucleic acid bases. I'm going to talk just briefly about that in case you know nothing about biology. I'm actually I'm going to sneak a little--I'm going to sneak more biology in this talk than math actually. So I'll talk about that in a little bit. And so its completion was announced 13 years later, in 2003.
Now, to paint a very simple picture of what the process was to do this is they started with the whole genome, cloned many times over.
And it's actually quite expensive to sequence portions of the genome. So they wanted—they, in order to bring down the cost, wanted to do a lot of work upfront in deciding through more traditional experimental methods, chunking up the genome, and then just sticking small chunks into the sequencing tool, if you will. And the nature of the sequencing tool is it chunks up whatever it's given and then they have to reassemble it. So they're given—they are actually given—and down here, you can kind make out a series of letters, so this is the end product. And they are just given random pieces and they have to figure out how these random pieces fit together. So you can immediately see that this is where the computational aspect comes in where they're going to need some algorithms to figure out where some sections maybe overlapping 10 times and there maybe even some gaps where nothing's overlapping, right? It's all a statistical roll of the dice to figure out what they end up with. And in fact—I forgot to look it up in preparing for this talk. But I certainly remember that there is talk of Perl, the program language Perl saving the Human Genome Project because they swooped in and really help bring this bit together. I don't have any details on that or the veracity of the statement.
Anyway, so things then finished and they were actually so proud of their effort, they printed it out. And this is over 100 tomes [phonetic], over 1,000 pages each, and typed that's almost too small to read of the human genome, OK? And--

[ Inaudible Remark ]

That's right, four letters, right? And probably the only place you're going to see additional letters is in like the attribution in the front or something like that. Maybe the page numbers, right? And my thought when I saw this is they ought to take those wiz kids who memorized thousands of digits of pi and throw them at this.
So anyway, now I'm going to slip in to biology mode for just a moment. Your DNA is composed of four nucleic acids, adenine, thymine, guanine and cytosine. So it's just these four letters repeating in various combinations. One thing--so you'll always hear that DNA is the blueprint for an organism, OK. But what--I want to explain what they mean by the blueprint and the realization of that. So these blueprint, if you will, there are portions of it, not all of it. There are small portions of it which actually contain genes. And it's the genes which is the blueprint that at least I'm concern about talking about today. And those genes go through a conversion process. So they take a portion of that DNA and use that as a blueprint. And then what they refer to as codons are sets of three nucleic acids, so ATA, CAG. And each of these codons maps to a single amino acid and there are 20 of these, OK. And--so it's these amino acids which chain together, so you have a whole bunch of ATCs and Gs. They're broken down into triples. Each triple turns into a different amino acid of which there are 20. So you have an alphabet of 20 instead of an alphabet of four. And then as these things all string together, which was what you call the gene, then it kind of globs together into a protein. And this is the actual thing that's doing work in your body, OK. And one of the predominant ways in which they're doing work is that they stick to each other and they interact together. And through that, some sort of biomolecular process is able to occur, and we digest and get oxygen into our blood and all sorts of fun things, right, thousands of these proteins.
There are 20,000 to 25,000 genes in the human genome. Now there is certainly a huge amount that goes on with DNA that has nothing to do with genes turning into proteins, but that's certainly beyond the scope of what I want to talk about.
So, these are going to add in some background and some of that information will come in coming up. This is a more accurate picture of what's happening now, is that sequencing has become inexpensive. It's dirt cheap now compared to how it used to be. And we'll look at some numbers here in a little bit. And--so they can generate a whole lot in a short period of time. However, these pieces are actually quite a bit smaller. So in the old day, having 200 base pairs, bases from your DNA to be sequenced, that would be considered actually quite large. Here--excuse me--would be considered on the small side. Here, it would be considered huge, OK. Much, much smaller, which means that it is a lot harder to figure out computationally how all these things fit together because you got so many more pieces and they're so much smaller. But that's the way they're doing it. They're--you know, you can--I have a vision of the biologist with the bucket of DNA slop [phonetic] throwing it into trough [phonetic] and out it comes, OK. No disrespect to biologists because I'm feeding at the trough. So, you don't have as much of that scaffolding upfront, kind of all that hard work upfront. Now, it can be done computationally. Our computers are better to handle this harder problem. So they're going to let the computers handle it and let the sequencing go wild. So what is--to look at a comparison to the Human Genome Project, what they have going now, kind of where--they are pretty much at the analysis phase at this point and I believe--and I could be mistaken, but I believe most, if not, all the sequencing is done is what they call the 1000 Genomes Project. And that is just what it sounds like, to sequence 1,000 genomes.
Actually it's the sequence more than 1,000, it's to sequence over 2,600 people from 26 populations from around the world, OK. And the idea here is now you can start comparing human genome to human genome and look at differences in the genes to be able to isolate diseases, what genes are causing diseases and so forth. Note the cost, only 30 to 50 million dollars compared to almost three billion dollars to do, you know, many times as many of the number of genomes. Also, you're going to want to buy that second thumb drive if you want to download the stuff, OK. The datasets are huge, absolutely huge. And this particular one is actually being housed by Amazon and of course it's not intended to be downloaded but you would use Amazon to cluster services to do useful processing on a dataset this large.
Now—so we've got all these genomes sequenced out and the question is what useful things can we do with them and understand what useful things we can do with this information? We have to understand a little bit about evolution. So we--I've already mentioned several times genes and differences in genes, and you can imagine words like mutations and stuff like this that you've heard and genes that have gone bad or causing disease and so on and so forth. But let me describe that more biological terms. So, whenever you're creating an offspring, we know that the DNA is getting duplicated, OK. And the fidelity of this duplication mechanism is absolutely amazing. So if you look at duplication mechanism just for the cells in your body, every single cell in your body has a copy of your DNA and you've got somewhere approaching a trillion cells in your body. And the fact that it all holds together in a trillions times over, it's able to duplicate your DNA without significant error to my mind is amazing. Same thing, when you're creating offspring, the errors that are going to be introduced in creating an offspring are quite low. But occasionally things do go awry with this biomechanical machinery in where it's trying to create a copy of CATT, it may screw up and go CAGG, OK. There are a number of ways it could screw up. It can "forget" to do a portion of it in which case something has been deleted that isn't in the child that was in the parent. There are times where it can get duplication, so it kind of slips. And there are whole chunks of DNA where you get two, three, four, multiple copies of them.
The quite common version of it would be referred to as a gene duplication where actually a whole section of DNA which holds a gene gets duplicated. And that'll come into play a little bit later. Most of these aren't viable. These mutations quite often result in an organism that can possibly live with these mutations, and so you don't see them at all. Other times, it creates some horrendous disease and--so you don't tend to see that arise in the population at all. But on rare occasion, there is some mutation that either differentiates the child from the parent or provide some sort of advantage in the environment that the parent didn't have. And so you draw this out over millions of years and this is the model that we have for evolution, OK.
So, now that we know that the difference between us and—between many of us or between us and a different organism like a cow is these mutations occur over evolutionary time, to differentiate us. So what we can do is we can compare the cow's DNA to our own DNA and find out where they're the same and where they're different. And one of the very common tasks to do and one of the reasons for wanting to do this comparison between the two and see where the changes have occurred is the example and there are many, many things that you may want to do, but inferring the function of unknown genes. So, maybe I've worked in yeast for quite a while and yeast is relatively easy, certainly compared to human because they won't hold still, to figure what the various functions of those genes are, OK, in yeast. And then what you can do is you can take a—identify a gene in a human that you don't know what that does. And if you can find which gene it's similar to in yeast, then you can start to make inferences. Well, it's doing this in yeast. It must be doing something similar in human. So, that's kind of the—one of the rationales for wanting to line these up and find out where they're the same and different. And so the—at a more computational level, you can come up with some sort of scoring matrix and this is the most absurdly simple kind of scoring system you can have. It gets a lot more realistic and complicated.
But it’s the most basic level. If I want to compare this sequence on my head, so this sequence here, here’s one possible way of aligning it. The ones with the lines here are where they align just fine. You could say that what happened was there was never anything here and it was added to this one. This--I guess I should go down this way. The G got deleted. A G got turned into an A. These are the same. A T was inserted, a G was deleted. OK, so you get what these insertions and deletions and modifications. And you can apply a score to each of those happening. And then it becomes a computational problem on what is going to be the best way of aligning these given these scores, all right? So an example of a horrible, horrible alignment would be to take this entire thing and put it right and then there is a bunch of lesions and then a bunch of insertions, right? And so obviously you can find out--find something more optimal than that.
Now the--an established technique for finding how these align is dynamic programming. For those that aren't familiar with the term, it doesn't imply computer programming. It's just a methodology. And it will find an--given your scoring, it will find an optimal alignment, OK. However, the complexity is basically the--multiplying the length of the two sequences by each other. Now for two relatively small sequences, that's not a big deal. However, we look at this table here, and I apologize, it's five years old. It's the most recent one I can find. But what we are looking at is the number of sequences in million. So as of 2008, we have 100 million sequences. If you're looking at individual basis, 100 billion, OK. So now algorithms such as this are just going to take too long to do on so much data. And so my only reason for bringing that up is to show that so much of bioinformatics is having to let go of the optimal and finding heuristic solutions which are going to be acceptable.
The next stage in aligning these things is multiple sequence alignment. So what you’re looking at here is each row corresponds to a gene or part of a gene, it’s not even a whole gene, a gene in an organism. So as many roses I have or as many organisms as I’m comparing to each other. If you can at all read the alphabet, you can see there’s a variety of letters. So we're looking at the amino acids which comprise of protein. And you basically see colors where things are conserved. So everything here is the same. The bits of yellow are the same. And you can kind of get a picture for how things are diverging. And if you can read it, you can see at the very top bovine, human, mouse, rat, chicken, so all the usual suspects are up here, OK. And once you start putting these sequences together this way of the manifold things that you can do with them, one of them is to build a phylogenetic tree. How are these organisms related to each other? And those which are more similar are going to be closer to each other on the tree and those which are more distant obviously are going to be further away on the tree. It is hard. You have to establish where the branches are coming out. It is in mathematics terms that is empty heart problem so you end up having heuristic methods to do this as well when you get to any realistic size number of organisms you want to compare. Compounding that is if we figure in this--by the way, this is not connected to this, OK. If I was to want to draw--so I take this and I go ahead and I create a tree.
Now I get a different gene for all of these and I map its similarity, and I generate a tree for that. Even if I'm using the same deterministic method, I may very well end up with different trees because different genes are not necessarily going to separate from each other at the same rate, OK. So, again, the purpose is just to show really, really frustratingly hard problems to come up with a definitive solution. And I would say that's what makes the field so exciting is that you just take two steps and you immediately get off the beaten path and you have a gravel road and you can actually do some real contributions.
So now just a little sidestep down an alley here is once you're starting to align these sequences, then you can start to do some interesting comparisons. So you can--given a lot of organisms, you want to see how it is that they're diverging. So a lot of those are changes from one amino acid to another. You could--one thing that seems obvious to do is, well, what's the frequency of that, OK. And what this is, is the [inaudible] table. I'm looking at the transition frequencies to go from one amino acid to another. And then what you can do, this then becomes a lot more informative for doing future alignments down the road because now you can roll in the information on the probability of one set of differences versus another, OK. The reason this one delights me so much, it's BLOSUM62. It's been around for, I don't know, 20 years, used all over the place, really, really useful. Five years ago in 2008, some researchers looked at the source code for it which us of course freely available. And they found out there is some mistakes in the source code and that they actually all this time had been doing this table a little bit wrong. It's not actually--in every case, it's not actually showing the correct frequencies, OK.
To add insult to injury, the wrong table, the incorrect matrices performed better than the intended matrices. And so that's just some of the wackiness and weirdness about this field. It's preflexing [phonetic].
To throw in another example, hidden Markov models which I imagine many, if not, all of you mathematicians are familiar with. To set up a scenario for where you would use this is these lines are representing DNA and these chunks here would be—that DNA which eventually gets converted to amino acids and then to a protein, so a gene. And these genes are converted into proteins not constantly, right? There's regulation going on. Obviously, you're not sweating all the time which requires a set of proteins to do something and you're not freezing everything else. So depending on where the cell is in your body and what's happening to you at the moment, different proteins need to be expressed to do different tasks. And so to turn these on and on, and they refer to it as gene expression, so to express them and start that wonderful process to create proteins, there is something called a CG island, so taking those four base pairs. A CG island is literally just—ephemerally described as those areas of DNA that tend to have more Cs and Gs than else where. And it turns out that in about 40 percent of those locations, those end up being the locations which serve as what are called promoters where some molecules latches on to this island and then starts turning that into a protein, OK. So it's useful to find these. It's informative because it also helps you discover where genes are. If you find one of these CG islands, likely enough far downstream, there's going to be a promoter or a gene. And so you can set up these state transitions like this. So you have—the state—so these are transition, right?
You have a--every one of these arrows represents a probability and it's your probability of going from an A to a C or a C to a G, in any combination at all. You have them two times because there's a certain set of probabilities for the next nucleic acid being one of these if you're in an island and there's a different set of probabilities if you're not in an island. And what's hidden is--in hidden Markov model is you don't know what the path is that's being taken to produce that set of DNA you're examining. So you build this model and you run it along the DNA and it'll tell you the areas where it's finding the CG islands, OK.
Now to take the discussion in a totally different direction, just again kind of paint a picture of the breadth of this field, natural language processing. PubMed is the repository for finding information about articles related to biology and medicine, and has abstracts for them all and information data publication and so forth. And it's got 23 million of these on record right now. I did this earlier today actually, so that's this morning. Last week, there were 14,000--nearly 15,000 articles published. So that many articles are being published every week, OK. It's mind-boggling. If 90--

[ Inaudible Remark ]

It didn't slow them down any. [Inaudible] are impressed. Even if you assume that 99 percent of them garbage, the 1 percent remaining would far more than any one person could read or keep track of. So you know that there's a ton of information out there that isn't getting out, connections that aren't being made, right? You have three articles next to each other, if you had the right three articles and read them together, you could identify clever things so that no single paper described but that you can hallucinate by reading the whole thing. So obviously natural language processing plays a key role here. Very, very challenging, so they--you know, there are genes we mentioned in here and what they do and what they don't do. And just--well, there's nothing more to say about it other than it's a very rich area for exploration and developing algorithms and so forth, [inaudible] this information out of these texts.
Now, granted that you don't have the full article for all of these. You do have the abstract for all of them, OK. But that said, we have a relatively recent trend in the last 10 years of things going open. Yeah, don't need a subscription to get out of it. So there are hundreds of thousands. I don't know what the current count. It could well be over a million of articles that are currently available in this general field where you have the full text for it. So very, very rich.
So to now step back and look at all this, we have to understand why it is we're doing this. It is because we--I think just in principle, we want to deepen our understanding of the world and I think that's an honorable motivation for entering this or any field of research for that matter, but also to improve human health. And certainly that's one that I need to keep track of because my research predominantly has to do with looking at data on yeast and it's hard to think how grandma is going to get out of the hospital early because I'm looking at this evolutionary network but I keep convincing myself it is. Now, is bio--bioinformatics is big. I think everyone will agree on that. How big is it getting?
I'll let you read that quote on your own. So--yeah. So biology, molecular biology, bioinformatics is definitely taking hold of folk's imagination and is definitely making it into the educational establishment. Now another issue that certainly doesn't necessarily come up in this talk but I want to bring it up since it's an interesting to me is the role of biologist versus computer scientist, mathematicians and statisticians. And there can be an uneasy relationship. I think it's not nearly as bad as it used to be. But I think the culture is there, not that it needs to be explored. Let me just show this slide.
There's no doubt that mathematics, statistics, computer science is on its knees serving biology, OK. We are applying these domains in the service of biology so that biologists can do their thing and create wonderful discoveries. However, you don't want to let that attitude take hold as far as the roles. So if you are doing cutting-edge research, really the mathematicians, the statisticians and computer scientists are not in the service of biologists. Their peer is working together on problems. But there's definitely no doubt that the technologies, if you will for lack of a better term, these sciences are in the service of biology. So I thought that was an interesting dichotomy and I just wanted to throw that out there to you. All right. So now I'm going to totally shift gears. I wanted to get a little bit deeper into something, so I thought I would kind of trim at the edges of my research and talk about that a little bit.
Again, it's--to tell the truth, it's not very deep as far as needing to understand it, so I think you'll do just fine. I always get on the mathematician's good side when I show Paul Erdos. So, 19--the 1950s, early 1960s, he published a set of seminal papers which kind of kicked off graph theory and the study of these networks, their properties and so forth. And for the next four years, really, that's the way research on graphs went.
And one of the interesting things to note, and there are always exceptions, so acknowledging that there are exceptions, what was studied predominantly were two types of graphs or two types of networks, lattices which are regular or completely random. And everyone should understand completely random means throw up a bunch of dots and with some probability connect each, all right? All right.
And then to throw in the only Greek symbols I have in this talk, or symbols, the clustering coefficient. The clustering coefficient is merely how well a network is connected to it. And the formula is simply to count up the number of triangles in your network, so I got a triangle here, I've got a triangle here, I've got a triangle here, so on and so forth. You count up all the triangles, multiply it by 3 and you divide it by the number of triples, meaning two edges with three nodes. And that's going to give you a number between zero and one. And the closer to one it is, the more highly connected that network is. The other one is characteristic path length which is simply to find the shortest path between all pairs of nodes. And if you average them out, that's your characteristic path length, OK. Keeping those two measures in mind,
a feature of each of these is if you look at the characteristic path length, the average distance between any two nodes in a random graph is quite short and it's quite long if you're in a lattice, takes you quite a bit longer to get from one node to another. And you have the opposite kind of effect with clustering coefficient. Clustering tends to be quite low in random graphs and you can get quite high clustering coefficients in lattices.
Now in 1998, Watts and Strogatz published a paper and they had an interesting observation. They said, "Let's take a regular lattice and with some probability rewire some of those edges." So if you rewire with 100 percent probability, you're going to end up with something random and then you have everything in between. Some low percentage probability of rewiring is going to give you something in between. And this is the money plot, if you will, that they had in this paper and that is to note that you end up with a combination of features from both of these. So, here's the clustering coefficient for a lattice. If you took the clustering coefficient--vary the clustering coefficient--excuse me. As you vary this probability from a lattice, you'll note the clustering coefficient stays quite high and this path length drops quite low. So you're getting features of both of these networks occurring at the same time. And what's remarkable is that probability of rewiring edges right here is just 1 percent to get that change. So rewire 1 percent of it and you totally changed the way that network works. And the other observation that go along with this was that, "Hey, we're not finding that the networks that we're now looking in the real world, they're neither like lattices nor like random graphs, they have this property right here," OK. And so that kicked off--[inaudible] even put a little circle there.
Collective dynamics of ‘small-world’ networks

Duncan J. Watts & Steven H. Strogatz
Department of Theoretical and Applied Mechanics, Cornell Hall,
Cornell University, Ithaca, New York 14853, USA

Networks of coupled dynamical systems have been used to model biological oscillators\textsuperscript{a}, Josephson junction arrays\textsuperscript{b}, excitable media\textsuperscript{c}, neural networks\textsuperscript{d}, spatial games\textsuperscript{e}, genetic control networks\textsuperscript{f} and many other self-organizing systems. Ordinarily, the connection topology is assumed to be either completely regular or completely random. But many biological, technological and social networks lie somewhere between these two extremes. Here we explore simple models of networks that can be tuned through this middle ground: regular networks ‘rewired’ to introduce increasing amounts of disorder. We find that these systems can be highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs. We call them ‘small-world’ networks, by analogy with the small-world phenomenon\textsuperscript{g} (popularly known as six degrees of separation\textsuperscript{h}). The neural network of the worm \textit{Caenorhabditis elegans}, the power grid of the western United States, and the collaboration graph of film actors are shown to be small-world networks. Models of dynamical systems with small-world coupling display...
That kicked off a change in how network research was going to occur. And there are a number of things that it acknowledged. No longer is it purely theoretical analysis and, again, I said there are exceptions to social sciences where we're good at doing some clever stuff with graphs in social networks. Theoretical analysis, it's now coupled with empirical evidence from real world networks. Also, we're talking about the turn of the century and we have a ton of data coming, OK. So now finally it's no longer studying small networks but real systems where we have a lot of data for them. They are dynamic systems which change and particularly they change over time. So it isn't like taking one graph and then studying all--everything about it but studying how it changes over time. So I think that characterizes really how this field change around the turn of the century.
Now using that as preamble, this is just showing how I would combine my work in protein interaction networks and certainly I didn't invent it, it's done long before me, with graphs. Protein is a node. If two proteins interact, you put an edge between them. And also this quirkiness which is fairly common in biology is if you have--when you have multiple instances of a protein, it's possible it'll interact with itself and you want to represent that as a self-loop.
This is old. This is 2004 I believe. But this is to give you an idea of what a protein interaction network might look like if you took interaction data.
I've already talked about evolution but I think this picture is good so I want to reprise it in the context of networks. So we have these genes and I said one of the errors that can occur is that the machinery slips and it ends up creating two copies of a gene. So now, when you imagine this gene being converted to a protein, you have initially twice as much of it. But what will happen is over evolutionary time, you will have those mistakes occurring where you get mutations in these. And a lot of times what's going to happen is you're going to get some mutation in here that makes this protein totally nonfunctional. No problem, you still get the offspring living because we've got a good one right here, OK. And in fact, if you look at--many genomes are certainly--it's no exception, there is a ton of this kind of stuff. We have a lot of duplications that are totally nonfunctional because they've been mutated out of all recognition. However, one other possibility is that you can get some sort of mutation that gives this some sort of new functionality.
Or, alternatively, I could say that this original has functions A, B and C, it duplicates ABC and A, B and C. Over time, I lose AB in here and I lose C over here, so they're dividing up functionality. And there are a couple of terms to use for that which is neofunctionalization for the acquisition of a new function and subfunctionalization for kind of dividing up those functions amongst the duplicates. Also, these aren't dealing with functions. They're dealing with interactions. So generally, we draw an analogy where interaction corresponds to the function. So here, we have the duplicated--I refer to it as the proteins, it's the proteins that interact and they both have the same interactions as the progenitor does. And it's those interactions that can go away or new interactions can arise due to these mutations. And there we go.
I'm just accustomed using my hands. So what might that look like if you were to animate it?
Not very informative, there are labels on it. Really there's no biological aspect to these. These are truly just generic but to give you an idea of what would happen over evolutionary time with these mistakes and evolution causing genes to duplicate, creating multiple proteins and then they diverge and their interactions change. So that's enough information for us to look at some of the interesting problems that can occur when you examine these things. So there are a couple--yeah.

[ Inaudible Remark ]

Yeah, the--there are rates that are described and it changes wildly. So it very much depends on the kind of thing that's getting duplicated, but yeah. And you're--I know your next question is going to be, "Well, give me a sample rate." And I would have to open up my paper because I don't remember the number.

>> So your animation always was making new nodes. But it seems like you have slow chance that sometimes you were under the direction [inaudible] things too.
Yes, absolutely. So that was definitely only showing gene duplication. Another possibility is that a gene would be silenced and no organism would still be viable in which case I would just simply remove the node from the animation. So you're absolutely right that it would be there but it isn't showing. OK. So in the--this work done by Wagner was looking at these duplicates and noting that there are two possibilities for a gene. And again, this would be the expressed protein of a gene. So I'm using the terms gene and protein interchangeably, OK. Either my progenitor could be self-interacting or my progenitor is not self-interacting.
The Role of Mathematics in Bioinformatics
The Role of Mathematics in Bioinformatics
The Role of Mathematics in Bioinformatics
Mathematics and Statistics Colloquiums

The Role of Mathematics in Bioinformatics
The Role of Mathematics in Bioinformatics
The Role of Mathematics in Bioinformatics

Mathematics and Statistics Colloquiums
The Role of Mathematics in Bioinformatics
Paralogous interactions and neofunctionalization

A  
B  

Wagner 2001, 2003
And if my progenitor is self-interacting, then when a duplication occurs, there should be an interaction between them because certainly immediately after duplication, they are identical so they should be interacting with each other as well as themselves, OK.
The other possibility is if that original progenitor was not self-interacting, then when a duplication occurs, there would be no interaction between them. And if there was an interaction to arise, this would be through this concept of neofunctionalization, a new function that didn't exist suddenly appeared.
Now, what Wagner observed in the data was that there was-- Oh, notice 34 in the data, notice 34 pairs of these gene duplicates, of these protein duplicates. They're called paralogs. He saw 34 pairs--paralogs and none of them had the self-interaction. So 34 of these were witnessed in the network. So there are two possible conclusions. Either this interaction arose de novo 34 times or these interactions were lost 68 times. And then it just becomes a matter of, well, what’s more plausible? What’s more plausible is what requires the less disturbing to what was there before. So the conclusion is the--there’s quite a bit of de novo interaction occurring.
But--and this is where us as computational folk get hung up. This is a picture of an assay in--that the biologists use to decide if two proteins are interaction with each other, OK. So we're getting now into the biologist's lab to see where they're coming up with all these data that Wagner was using there and then I use and so on and so forth. And this is one of those experiments. Here are two of those proteins floating around and we want to find out if they're hooking up together, OK. And there is this interesting protein right here, this red one. And whenever it binds to the DNA, it will get expressed. And this is actually a phosphorescent protein, meaning that you could look at a microscope and see the magic happening, the glowing happening. And so another thing that was discovered is that you can actually separate this out into two separate parts. And then using a little bit of magic that only biologists know, they're able to hook one end of it up to one protein and then other end up to the other protein, and then you set the proteins loose and you look to see if things glow. If things glow, then the only reason things are glowing is because these two proteins actually interact with each other. A-ha, one data point, right? We can throw that into our data and include it. All right. And this is kind of the picture you see when you get into Biology 101 texts. However, this is called GAL4. This stuff in red here, this would be called the binding domain.
If you were to look at a more realistic picture of it, you can clearly see this is the DNA. Now the whole thing looks messy in a couple of ribbons. But one thing you notice is that there are two of them, OK.
So really this diagram here isn't accurate. The diagram that should've been shown to us is
this, all right? Now what's the problem with that? What happens if my little Pac-Man protein is self-interacting?
They're going to bind to each other. Now, this thing floating around has nothing to grab on to and nothing lights up. So what that means is if you have a protein that's self-interacting, you're more likely than not going to report that as an interaction.
So getting back to it, this may very well--it doesn't discount this is a possibility. But this may very well be a possibility simply because you can't discount the fact that these interactions are actually there. We're just not seeing it in the data because of the quirks of the biological assay, all right? So I think the computational work was spot on, pretty simple, straightforward conclusion. However, we are at the mercy of the data and I think that's an important part of being a bioinformatician is to be cognizant of where the data is coming from, what's being used to generate it, piling up with biologists so that you learn all this clever stuff.
And then what I want to conclude with is my ultimate slide, is just to throw this out. I feel a little bit like a salesperson, not in a [inaudible] way because I want to sell you on the idea that you all want to run to your car and you go drive down to the bioinformatics shop and buy all the cool costumes. The language is I think everyone should be familiar with. I've heard of Python and Perl, they're very prominent in bioinformatics research. You can do a lot of good work with them. There are a lot of libraries that support your use of analyzing biological data in these languages. Also, statistics and graphics for doing plots are--if you're not familiar that, it's a--well, it's for doing statistics and it's open source, so it's free, and it's absolutely fantastic and wonderful. Conferences, there are--the two--two of the largest are RECOMB and ISMB. RECOMB I think is mathematicians might find that more appealing. So it tends to be on the more theoretical side, whereas ISMB is a little bit more applied. Now I use these terms relatively. That's not to say that there isn't some hardcore, really good statistics in math being done. It's full of it here, OK. It's full of it in both places. And likewise that there's nothing applied here. There are certainly plenty applied here. But if you had to put them on a scale, you will find RECOMB more theoretical and ISMB more applied. So using that as the end of my pitch, I thank you for hosting me.
Do we have any questions?

Yes.

[ Inaudible Remarks ]

How much of that is understood in the sense that how much of that do we have sense of what those sequences of those genes do?

We have--I don't know. Do you know what the proportion is of genes that we have functions for in the human genome?

[ Inaudible Discussion ]

The follow-up question for that is do we understand how genes interact? Like if we [inaudible] and changed one gene, do we have a sense of how that gene interacts with the rest of the genome?

Yeah, there are certainly genetic experiments where you'll knock them out and see how their--what happens to the organism when you knock one out compared to another one. You want to share more of the assays for that?
In humans, I mean they're very complex and proteins are interacting inside the cell [inaudible] interaction network, they're binding to one another and they bind all sorts of things of which you don't know [inaudible] or what's not.

>> Right.

>> So [inaudible] we've been studying for a long time, [inaudible] about biochemical pathways. We know a lot about a lot of biochemical pathways and how [inaudible] organize and how they work [inaudible]. And then there are other things that we didn't even know existed.

[ Inaudible Remark ]
One more—another confounding thing to think about is that when I was a researcher in using this protein interaction networks, the simplification that's going on is just incredible because one thing—just to throw out a piece of information that isn't in that network is time. Not evolutionary time, but if you're looking, I don't know, at a yeast organism, from one portion of the lifecycle to the next, there are going to be different proteins interacting, right? And none of that information is there. That's just saying they interacting, we don't know when or under what conditions. So to step it up and make my research more valuable, I need to start integrating that kind of information from the experiments into the network. And now we go from simple math established graph theory to stuff that's a lot more ephemeral and hard to get your head around. But that's what makes it fun. And I think that's what—that's where we really show our value as computational folks is that we've got all these brilliant ideas of how we can apply the things we know computationally to this really hard problem. Anything else?

So I thought that there's big chunks of the genome that are like [inaudible], is that correct?

[ Inaudible Remark ]
Right. And certainly there was a time where--and there was a time where everyone referred to that as the garbage. Now everyone's taking a step back and going, "Ooh, it's not garbage."

[ Inaudible Discussion ]

[ Inaudible Remark ]

So like where are they getting these matrices? How are they dividing those [inaudible]? They're based on [inaudible].

[ Inaudible ] chicken and egg problem because they have to align the sequences in order to get the matrix. And then from the matrix, they can align sequences. So there are certainly some of that. The different--and then they are basically--

[ Inaudible Discussion ]

Right, right.

[ Inaudible Remark ]
Right. And so--right. And so you group them by similarity. And so you're going to get different matrices depending on how spread apart evolutionarily the sequences are.

[ Inaudible Remark ]

Yeah.

>> Has there been any [inaudible] looking at trade-offs between genomic stability [inaudible] with [inaudible] like--well how to make genome more stable by having more genes [inaudible]? Something will make them less stable--

[ Inaudible Remark ]

>> I don't have an answer for that. I'll again defer to the biologists.

[ Laughter ]

>> I imagine the GC islands, for instance, there. And those actually tend to be [inaudible] and basically more likely mutated.
Hey, thank you very much.

[ Inaudible Remarks ]

>> Hey, thank you very much.

[ Applause ]