



Futures in Biotech, 34: A Great Historical Document – the Human Genome

Leo Laporte

Bandwidth for Futures in Biotech is provided by Cachefly at cachefly.com.

Marc Pelletier

This is Futures in Biotech episode 34. A Great Historical Document – the Human Genome. Futures in Biotech is brought to you by audible.com, the internet's leading provider of spoken word entertainment. Get a free audiobook download of your choice when you sign up today. Log on to audible.com/biotech today for details.

[Music]

Marc Pelletier

Welcome back to Futures in Biotech, I am Marc Pelletier. This week, I think we have a really good example of one of the things that we are trying to do with FiB. That is to give a true first-hand account of some of the great science that's going on today. I really believe that it's by getting this close to the science and to the people doing the science that we can get an understanding or a sense of what lies in our technological future.

Now today's show is on the human genome, and in past shows we've had Dr. Lee Hood, the inventor of the DNA sequencer, we've had Dr. George Church, one of the people who was personally involved in initiating the Human Genome Project, but getting the code, and understanding the human blueprint, are entirely different problems.

Now today's guest, one of his jobs has been to try and make sense of this human blueprint. And his work, among other things, has revealed that the human genome is more than just AC/TG, but a rich historical text about our human past. Our guest is Dr. Mark Gerstein. He is the Albert L. Williams Professor of Biomedical Informatics at Yale University. He is also a Professor in the Department of Molecular Biophysics and Biochemistry, and a Professor of Computer Science. So here's our conversation.

So, what exactly is annotating the human genome? Is it sort of like, well first of all, you guys are the first people to really get a good view of the entire human genome, but what is annotating?

Dr. Mark Gerstein

Well. That's a good question. I mean, actually I've kind of thought – thought about that, you know, question a lot because people talk about annotating this and that and then you say what is that. And I guess what the word really means, of course, is affixing notes to something, you know it comes obviously from annotating like a text, and so in the – you think about what does it mean to annotate text. You know you sort of have this text and you put little notes, whether they're footnotes or side-notes, whatever, on bits of text that sort of describe what is meant out of a particular bit of text. And you know, I think that sort of famous types of annotation would be say related to sort of religious texts – you know the Bible, or the Torah and whatnot, and you know there is this sort of almost tradition of you know people who'd read these texts and then maybe sort of annotate in great detail what this passage meant, and you know often, they'll be like a set of notes, you know there'd be the primary text and there'd be a sort of set of notes below it. And you know I think, actually the – in Judaism, I think, the Talmud is actually almost the – that is what that document is, it's notes that really describe what's in the Torah. And so annotating the genome is kind of the same type of thing. You imagine that you have this canonical primary text, the genome, and then you have these notes that describe, you know, what's going on in this

particular region of the genome, this is a gene, this is a binding site, you know, maybe this is some dead gene, a pseudogene, or this particular region is a sort of block copy of that region, you know so as if you are reading a law and you'd kind of have some sort of interpretation for what you are doing. That's sort of my interpretation of annotation.

Marc Pelletier

Do you see the genome as a historical text?

Dr. Mark Gerstein

Very much so, actually – I think actually the comparison with, a famous – a written text is actually a particularly good thing, because, you know, the human genome, I think, the really interesting thing about it is not only does it actually carry a blueprint for, you know, who we are, and describes how we – how to make a person, but it also has, within it the history of how human beings came to be. And so it has the molecular history of all our genes. And one of the things I am most particularly interested in is pseudogenes or fossil genes where the genome has in it sort of dead copies of genes that used to be functioning or copies of things that were attempted to be constructed, but never really, kind of made it. And the – you know, the interesting thing, about the genome is you might say, oh that seems like a really kind of bizarre inefficient thing, but most of our genome is in fact not this kind of functioning blueprint bits. So the belief is that, out of the, say, three billion bases in the reference genome, only about a percent of them are associated with the actual protein coding exons. And maybe, you know, a couple of more percent are going to be associated with some obvious degree of regulation, or obvious, sort of accompaniments to the protein coding gene bits. And so you have lots of the genome that doesn't have any obvious function, and a lot of that actually is just historical record. I think it's quite impressive actually, how, how much of it is.

Marc Pelletier

[6:30] The annotation has led to identifying segments that encode for proteins, other segments that have regulatory elements that control when a gene is turned on, turned off. Other segments, as you said, where there's duplication and extra genes, and redundancies and things that are no longer used. What are the most interesting fossil genes that you found, and what have they told us about who we are in the present time?

Dr. Mark Gerstein

So, that – that's actually kind of interesting, so the fossil genes in the genome – there's a number of categories. So one category of them, we call unitary pseudogenes, and those are the canonical, I don't how we should say it – what you might think of as a fossil gene. You might think there was this gene and it was happily doing something and at some point, oh I don't know, the needs of the organism shifted and it just died. And so if you lined up the genomes from a number of organisms and found the analogous or what you might properly call the syntenic region of the genome you might see a functioning gene in one organism and lo and behold in another organism the gene will be all messed up – kind of messed up like a fossil or a dead gene. And so, we use the term the unitary pseudogene to describe those things.

Actually, most of the pseudogenes in the human genome are not in that category. So, a lot of the other pseudogenes or a lot of the other dead genes, they come about from a number of sort of fundamental processes of genome evolution. So, the way you think about how the genome evolves, or how the genome changes is very much the way you might imagine a text evolves. So, how is that? Well, if you have a text, you might say well one operation is you can just basically copy it. You can copy it and you can mutate it, that's the sort of operations.

And there are sort of two main types of copying that you can do in the genome. One is kind of that you literally take a chunk and you duplicate it and you copy it again, and you might want to do this, for instance – say in the context of thinking about sort of how you might change the human text, you might want to kind of duplicate something if you're going to kind of reuse it with a slight modification. So, if you were, say, writing a computer program, often people might take,

say, a subroutine and copy it and then slightly change it and then they'd have a variant routine. And one imagines a kind of similar thing happening in relation to duplication of the genome.

Now sometimes, this process of duplication is going to lead to a variant gene and that's why a lot of the genes we have are in gene families. But sometimes it's going to lead to either a messed-up gene or it's going to give rise to a gene that's only going to be used for a little bit and then it's going to die. So, that gives rise to what's called duplicated pseudogenes, okay?

And then the third process, which is a little bit more convoluted but which is kind of interesting, is there's another copying process in the genome called retro-transposition where you have a bit of the genome that's transcribed, right? So it's sort of on and it's copied in that sort of transcription way and then that is retro-transcribed back into the genome and that creates what's called a processed pseudogene. And that might sound very, oh how should one say, convoluted but it turns out that lots of the pseudogenes in the human genome, lots of the dead genes, are actually created that way where you'll have the sort of – how should I say, shadow of a highly transcribed gene would be all these processed pseudogenes splattered over the genome.

So, one of the biggest families of pseudogenes in the human genome comes from processed pseudogenes of highly transcribed genes such as metabolic enzymes or ribosomal proteins which are highly transcribed. They have all these kind of messed-up copies splattered into the genome. In the extreme...

Marc Pelletier

Wow! That's scary.

Dr. Mark Gerstein

Yeah, no, no, it's actually kind of interesting. So, in the extreme, you might have, say, one functioning copy of the gene and you might have, I don't know, 50, 100 plus copies of messed-up pseudogene and that's kind of, it's – and just on a very practical level if you're annotating the genome and trying to annotate the actual copy of the gene – you know, say this is where it is – you have to discombobulate this from all the pseudogenes. So, that's a big process in annotation.

Marc Pelletier

So, when you have a lot of copies – so you have your DNA, let's bring it down to the real basics. You have DNA. DNA gets printed into a molecule called RNA – ribonucleic acid. That RNA then gets processed to specific forms to then be translated into a protein, right?

Dr. Mark Gerstein

Correct.

Marc Pelletier

[11:36] So, and that then gets reverse copied back into the genome once it's been cleaned up and shortened and sort of like a compiler, like a computer compiler, I suppose, then thrown back into the genome and you get – and you're saying the housekeeping genes, the things that maintain – things that are always turned on have the most copies of mistake – or not make mistakes but the most copies back returned to the genome.

Dr. Mark Gerstein

Yep, no, no, that's correct. Look, the analogy with computers is kind of a useful one. So you imagine the genome is sort of like the fundamental copy that sits on the disk, of like the operating system of a person right? So, you could imagine one process is just you find the region on the disk and you just copy it and then you might mutate or mess up some of that region. But another thing that happens of course with computer programs we're all accustomed to is the program that actually executes gets copied into memory right? It's sort of – and this is I think sort of similar in some sense to the transcriptional process. Things are turned on and they're transcribed and they're put in kind of a more active form, and then you could imagine in this analogy to retro-

transposition the copy that has been put into memory, maybe as you were saying, has been shortened or, in the words of RNA processing, has been spliced. It's then that copy in memory is then recopied back onto the disk, right, in its slightly changed version and that's of course what gives rise to these processed pseudogenes.

Marc Pelletier

But they should almost be functional. I mean they should be the easiest ones to turn on, right? If they're – if they've been processed into the shortest forms, they don't require a whole level of machinery to do the modifications that are typical...

Dr. Mark Gerstein

Well, that's interesting. You might think that but actually it's sort of the opposite because even though – you're correct in what you're saying, it actually is the opposite and it's because when they're copied back into the genome, they're missing all of the upstream regulatory elements, right?

Marc Pelletier

The dipswitches...

Dr. Mark Gerstein

So...Yeah, the things that would say turn them on. So your normal gene has say the things upstream, that sort of signal for it to go on and it has a promoter and then they have the region that's actually transcribed, the gene itself, right? Now when it gets reinserted you just have the gene itself. You don't have these upstream regions and so in a sense they're kind of like copies of genes that don't have, you're right, the switches to turn them on. And now it turns out and you know in the process of tinkering in nature, sometimes...

Marc Pelletier

Who's tinkering? Who's tinkering?

Dr. Mark Gerstein

Say that again.

Marc Pelletier

Who's tinkering?

Dr. Mark Gerstein

It's... No, I don't know – that's a...

Marc Pelletier

Okay, you don't have to answer.

Dr. Mark Gerstein

That's certainly a good question. Or watch making or tinkering – whatever you want to describe that – the process of I guess evolutionary change. But sometimes you can acquire a upstream regulatory sequence or a promoter on to some of these – well actually technically they're processed genes initially when they are reinserted back in the genome. And they could become functional. But usually, what happens is they are not functional and then they decay and decay means they acquire these mutations that are very un-gene-like or very deleterious. And then they're seen as processed pseudogenes.

Marc Pelletier

So, this raises two questions for me. One, are there events that happened in the genome and have you seen this where an oncogene or a – so a pre-oncogene – something that if it's then mutated because of sunlight, or whatever environmental threat, you get your mutation – some kind of carcinogenic, that mutates into – so if you've got a highly active metabolic gene, can you

turn a cell into cancer? I mean are proto-oncogenes this or are those mutations in regulatory elements most commonly?

Dr. Mark Gerstein

So let me just go through that again? So you're asking if...

Marc Pelletier

Or am I showing my ignorance here?

Dr. Mark Gerstein

No, no, no, no, no problem so – I just want to make sure I'm clear on what you said. So you're saying if you tend to get a lot of mutations in a particular gene is that something that can give rise to a cancer? Is that what you're asking?

Marc Pelletier

If you have housekeeping genes, and those are the ones that are most commonly put back into the genome somewhere else at the wrong place, are those candidates for as proto-oncogenes or as genes that will turn into cancer causing genes?

Dr. Mark Gerstein

Oh, I know...

Marc Pelletier

Because they're housekeeping metabolic.

Dr. Mark Gerstein

Yeah, I don't know about that. Now I don't know if you can see that lots of processed genes, or processed pseudogenes that they would be associated with cancers and things like that. I don't know if there's – people have really found such a connection. I mean it's certainly an interesting thing. You know, one thing that people have speculated a lot on is maybe the copies of the genes that turn to pseudogenes could have some sort of regulatory role or maybe they have an evolutionary purpose because they're so similar to the functioning gene.

And so there's been a lot of recent discussions about how some pseudogenic copies might actually be somewhat active or they might also themselves be transcribed. And the transcription of these things gives rise to an RNA but it doesn't necessarily give – that RNA doesn't necessarily give rise to a protein. And maybe the RNA could, since it's so similar to the functioning gene, regulate the gene through some processes.

In fact, there's been lot of discussions just of late of a new type of RNAi called endogenous siRNA that provides a kind of mechanism for the transcribed pseudogene to actually regulate its parent gene. And also people have talked a lot about how since the pseudogene is so similar in sequence to its parent gene, but sort of slightly different, maybe it provides a way for the underlying sequence of the parent gene to kind of slightly change or be modified and evolve in certain ways without effecting the parent. And then for it to kind of resurface at some other point in time as a variant function when it's necessary. So there's a lot of speculation on that particular topic.

Marc Pelletier

[18:13] So, do – this is crazy stuff because it can get so complicated and yet all we really have to remember here is that we have As, Cs, Gs, and Ts, right? And that the code is pretty straightforward. There's 64 different possible combinations, right. Is it 64? I'm going to do some editing.

Dr. Mark Gerstein

Are you talking about for codons? Three...

Marc Pelletier

Yeah.

Dr. Mark Gerstein

Well that's correct. Yes. That's – that's correct.

Marc Pelletier

I'm actually probably going to get kicked out of the department for asking you these questions. But I also wanted to sort of bring it down to the simplicity side of the genome too. But before we do that and I'd like to know about some of those – the history – the human history through these pseudogenes. Have we or do we have, for example, genes that encode for wings or do we have fins or are there elements in our metabolism that show that we were anaerobic for a long period of time? You know living in dirt. What can you tell us?

Dr. Mark Gerstein

No, that's a very good question. I mean that's sort of the – I won't use the word the dream of genome archaeology, but in extreme you might imagine you could find the molecular fossils that actually correspond to the real fossils that we see right? I mean in theory maybe somewhere in our genome is the molecules that almost give rise to a dinosaur. Right? Now I don't think people have actually found that you know – such a direct association. They have found however the remnants of genes that give rise – that were associated with recent changes in the human lineage.

So, for instance, one of the famous examples is lots of our smell genes or olfactory receptors are pseudogenes and you can really see how in the human lineage a lot of the – our ability to smell has kind of died off. And presumably, you can look at some of those pseudogenes and see, oh, geez, that's the gene or that's the sequence that codes for a functioning thing in say a mouse or a even a monkey that's no longer necessary in the human. And there's been actually some famous examples of this where people have noticed the prevalence of olfactory receptor pseudogenes or smell pseudogenes and found their prevalence really associated with the appearance of three-color vision in the primate lineage.

So the idea being that now that we're able to see more clearly than our immediate relatives, we don't need as much smell ability and that kind of ability is a bit vestigial and has died off.

Marc Pelletier

Wow! That's pretty wild stuff. That's really crazy. I'm wondering with the Neanderthal genome being sequenced, are we going to see at what breakpoint maybe other lineages of hominids have – would have maintained that expression? Wonder if that's the kind of thing we'll be able to tell from the Neanderthal genome?

Dr. Mark Gerstein

Well I definitely think people are going to be really interested in when they look at the Neanderthal genome, when they look at just the variation of genome amongst the different humans or they look at other primates. They are going to be particularly interested in the genes that are immediately variant in the human lineage. I mean obviously that's of great interest to people. What are the things that are different among the different people or even different races of humans? What are the differences that immediately set us apart from other primates and so forth? I think that's going to be particularly interesting to people.

Marc Pelletier

I think with the fossils now that we – what's his name? Svante Pääbo and his colleagues have developed their ability to grind up bones from museums, then decipher what was contaminating DNA and what's real DNA from the organism. I think won't it be fun to go back to all the fossil

records that we have and then do an evolutionary comparison, a comparative genomics on organisms that lived 200 million years ago?

Dr. Mark Gerstein

I mean in the extreme, I think that would be extremely interesting. I mean I don't know – I'm not an expert in it. I don't know the degree to which you can really get bits of DNA from those things. I mean I think that's a very complex...

Marc Pelletier

You can...

Dr. Mark Gerstein

You can but I think it's – I expect it's very complicated being sure that that bit of DNA really has been preserved over that long period of time or isn't contaminated and whatnot. I'm sure this was obviously a big thing obviously.

Marc Pelletier

He works it out, right, looking for quality bone samples that have a high level of our original intact DNA. He practiced by going to New York City, to a deli picking up some salami, beef salami, and then sequencing the beef genome. That's crazy.

Dr. Mark Gerstein

[23:45] Yeah, I know, that's very – it's very impressive and I think now the really amazing thing is with the new high throughput sequencing that's coming online, the 454, the Solexa, SOLiD whatnot. You know people are going to be able to just sequence everything as a matter of course. I mean within a fairly short period of time, it's going to be possible for any sample, if someone can get their hands on a sample, for them to just sequence it like crazy and really determine what's in that sample in a very molecular way. And so I do think people are going to be starting to think about just the most bizarre types of sequencing projects, in this archaeological sense and whatnot, for sure.

Marc Pelletier

Well they're two guys we've had on. We've had Jeff Gordon who's doing the human metabolome – what is it? Not the human metablome. The human meta-genomic project where they sequence the – microbiome, there it is. He's sequencing the entire genome of all the bacteria that live on the human body and I think there's tenfold more bacteria on the human body than there are human cells, right, so the – he's going in and sequencing every single genome he can get his hands on. And then we had – I mean have to edit here – our good friend from MIT...

Dr. Mark Gerstein

DeLong?

Marc Pelletier

DeLong. Ed DeLong. He's fantastic, at Episode 9, he's going out into the Pacific, north of Hawaii and just going into the war and sequencing every genome that he can get his hands on. So indeed we're going to have enormous amounts of the entire earth genome right? But how hard is it to annotate? Because we know as a genome...

Dr. Mark Gerstein

I think it's going to be a – I think that's the thing, it's going to be a huge challenge really properly organizing and analyzing all this information. And I think that's really the thing that's driving bioinformatics because on a sort of a very simpleminded level you might be, well this is all very interesting, we're going to be sequencing all these things and there's sort of a simple level of, well – a lot of the questions are immediately understandable. We're going to sequence all of those gut microbes and we're going to sequence organisms from seawater and whatnot and we're going to kind of compare the sequences up. And that in a sense is not very hard but when you're

confronted with the technical reality of these are gigabases of sequences, and you want to assemble them into some coherent genomes and then you want to kind of quantitatively start to compare them, that's where I think you really start to say, geez it's useful to have some bioinformatics people to really figure out how to deal with this information properly. Because it's actually very hard on a just a purely technical or informatic level to manage very large amounts of information.

So, I think one of the projects that's kind of driving this stuff forward now is the big – the two big NHGRI projects the – the Ed and co project and the 1000 Genomes Project – both of those projects I'm participating in and the data files that they are producing in these projects are just – they're just gargantuan. Like to give you an idea, in one of the most – the recent freeze of the 1000 Genomes Project they've sequenced now well over a terabase of sequences, right. And just simply taking a terabase of sequence and moving that around the world and getting it on your computer and starting to look at it is a massive challenge. I mean just the downloading of that amount of information on to your computer just takes hours, if not days, sometimes, to simply move it onto the computer. And then the sort of analyze – indexing it really requires a very high level of organization on the computer.

I mean you simply – files that are that size you can't just open it in the Microsoft Word or sort through them in Excel or something, you really have to have some special ways of organizing the files or searching through them, of querying them to do that. And those are really problems that fundamentally computer scientists have dealt with. I mean, those are the problems – those are the types of questions of course that lead to some of the amazing things that Google can do. In a certain sense, well, you go to the Google website and you type in – oh, I don't know, "diesel truck" or something and it finds lots of hits in a sort of simple way, it's not that complicated, right? It's finding lots of other web pages that have a sort of similar word and it's ranking them in certain sense by how many other pages link to them and so forth. But doing that very quickly and having indexed all the pages in that thing really requires a very high level of computer organization and really almost the exact same thing is required for doing that thing for lots of genomic sequences. And actually what we do now with – when we blast sequences against the database or when we...

Marc Pelletier

You might want to explain that a bit.

Dr. Mark Gerstein

[29:11] So, one – a lot of the basic operations people are doing in bioinformatics or in sequence analysis are very similar to the type of operations that people are doing with text that, you know, in the framework of say Google or web search. I mean, so, the most simple operation is you have a little query string that you are interested in and if one is talking about human text, you might have a phrase, you know, "pink pansies" or "diesel cars" or "computer printer is broken" and you want to simply take that phrase and you want to look at its occurrence in all of the other documents on the web, right? And then you want to rank the matches in some intelligent way.

In a high level conceptual way it doesn't seem that hard to do, but of course, if you literally went to every single document and kind of slid that string across and looked for matches it would take just, not even hours, it would take weeks to do that probably on – if we just did that one-off there for a search. So, people have to think about how to do that more efficiently and the similar issue comes about with – in genomics, I mean, often one will take a sequence of a protein of interest. A good example may be the hemoglobin protein is the one that always come to mind as the first protein really – one of the first proteins we really studied intensively.

And one would want to, say, take that sequence and look at the sequences from many, many different organisms all through the databases and see where it occurs and again, conceptually, not a hard thing to do. But to get – to be able to do that calculation fairly quickly and then to do it for many, many different queries does require a substantial amount of computer infrastructure,

and one of the nice pieces of computer infrastructure that's been developed is a number of programs that use a particular speed-up technique called hashing and one of the most famous is this program called BLAST which was developed by – partially by David Litman, who is now the director of the NCBI in Washington, and people use that program all the time to figure out what's in the big DNA data banks.

Marc Pelletier

By the way, it is my favorite tool on the internet. And for fun what I'm going to do for the audience is I'm going to put a gene up there and we're not going to – I'll put the sequence in the show notes so that they can copy and paste it into the BLAST site where I'll provide a link to it. And then they will be able to find out as if they were just doing random sequence out in the field what organism and what protein it encodes. So, this will be fun. But, go ahead. So the audience will be able to go back to the show notes, copy and paste the DNA sequence, you will see ACGT and then do this BLAST search.

Dr. Mark Gerstein

Okay, that sounds cool.

Marc Pelletier

It's fun, it's a fun thing to do.

Dr. Mark Gerstein

I mean, the impressive thing about that, I guess, it's sort of hard when you immediately see it for the first time, you know like, "Wow, it's found all these matches" and you – the impressive thing is it's so fast. That's really – and that's the impressive thing about Google, right? I mean, If you type "diesel truck" in and you get all these matches, immediately you're not, "What's so impressive about that?" You're like, "What's the high technology going on there?" And the high technology is the ability to have gotten that answer so quickly. And that's the thing that, of course, makes Google this essentially amazing company and also it's the technology that makes BLAST or a lot of these rapid sequence matching programs important and impressive. Because if you do it in a very naïve way, it just takes a gargantuan amount of time.

Marc Pelletier

Just a tidbit of information, I heard and I'm not sure about it, maybe you can validate it. I heard that the National Center for Biological Information (NCBI) website, the one that I'll link the BLAST site in the show notes, it apparently gets 30 million hits a day.

Dr. Mark Gerstein

Oh, wow. No, I didn't know that exact statistic but I would not doubt that. They have a tremendous amount of useful services that they have. I mean, BLAST being one of them, but they also have PubMed which of course lets people do lots of queries against the biological literature and they have Genbank which is really this repository of all the sequences and so forth. And the analogous one to the NCBI in Europe is called the EBI, European Bioinformatics Institute and they have many useful services too. Those services maybe a little bit more oriented towards proteins than DNA.

Marc Pelletier

So, do you do all your work, your annotation work, local, or do you do it online or do you have like a client-server infrastructure?

Dr. Mark Gerstein

So, that's an interesting question. I am very interested in the kind of computer technology for doing this type of stuff. It's always a kind of push and pull literally in terms of what you are going to do on your own computer versus what is going to be, kind of, distributed over the web. And, you know, people go back and forth, I mean, sometimes it is very advantageous to try to get as much information on to one's computer where it is kind of under their control and easily

manipulable. But other times, of course, the information is so large or requires such a degree of computation that it's advantageous to leave a lot of the information on a remote computer or a number of remote computers and kind of distribute the calculation over many things or to distribute the query that one is doing over many different computers.

And one interesting issue from – it's more of a technical issue, but it is how the kind of information architecture for the biological sciences should be set up in the future. I mean, should it be something that there's a number of central databases that have everything and you download a little thing on to your computer to look at it. Or is there a, kind of, this vast web of different sites that have bits of information that you would, kind of, link a little bit to to do things. And the latter, in a technical way is always referred to as a federated information architecture as opposed to a very highly centralized one. People talk about the merits or demerits of these different architectures for storing lots of information.

Marc Pelletier

[36:05] Well Mike, what about security, right? What if there's two centers that have the centralized information. All the genomes getting sequenced; they get put into the repository. I mean these are extremely valuable – in that as people are discovering or using techniques to monitor gene expression, right, with basically tricorders telling us what genes are turned on and turned off, and then comparing a disease state, right? Somebody who has cancer versus a non-cancerous kidney, for example. What genes are turned on in a kidney cancer that aren't turned on in cancer we now have, you cross-reference them against that database and then you have a – basically a way to diagnose with a simple quick DNA test, right? But what happens if that the database is gone? When happens to modern medicine when there is no more – if, and I'm not saying some kind of threatening failure or somebody is going to hack it, I just spent three days reinstalling Windows, right! So...

Dr. Mark Gerstein

I think that you are correct that that's a big issue and one of the things I am interested in, actually, is practically how do you set up a very distributed network of information resources. And, you know, again on a very high level, like, well it is not very hard, you know. You put a little bit of information on one computer, a little bit of information on another computer, but you sort of get into a lot of problems of – if the computers need to interact and connect with each other, they have to have a certain interface to each other and to have to expose a certain degree of services and whatnot. And when they do that, of course they leave themselves open to malicious people who are hackers, attackers, and whatnot on the computer.

And so as more and more people are doing biological science or doing biology who are interacting with a lot of computers they are depending on, there being a lot of computer services up there on the web for them to use. And this sounds all great, but there is a lot of people that are just trying to endlessly hack into computers and they are creating a big problem for this type of vision. And this is of course not a problem that's exclusive to biological science, it is a general problem that you see in society now as we more and more move to kind of this web-centric or Internet-dependent type of world. But it is a big, a big thing that people are thinking more and more about.

And there's sort of this – when the people originally created the Internet or they created the Web, they didn't really maybe think about to what degree this was going to catch on. It's not – this isn't their fault, but they just didn't think about it and there is a lot of protocols and standards and ways things are done that are not – that do scale to a very large degree, but are really not completely optimal from a security or stability type of framework, and a lot of times we are suffering from that now. I mean the people who get viruses on their Windows computer or – have a web site that gets a bit overwhelmed that they're trying to use. They're really suffering from design decisions that were made long ago where people didn't realize that they would have these implications, but it's a very interesting thing I think.

Marc Pelletier

I would like to take a minute to thank audible.com for sponsoring Futures in Biotech. Currently, they have over 50,000 audio books, speeches, radio shows, children's audio books, they have a lot of great stuff. And this week's pick is Legacy of Ashes, the history of the CIA by Tim Weiner. Here is a clip.

[Audio clip from Legacy of Ashes]

The mission of the CIA above all was to keep the President forewarned against surprise attack: a second Pearl Harbor. The agency's ranks were filled with thousands of patriotic Americans in the 1950s. Many were brave and battle hardened. Some had wisdom; few really knew the enemy. Where understanding failed, Presidents ordered the CIA to change the course of history through covert action. The conduct of political and psychological warfare in peace time was a new art wrote Gerald Miller, then the CIA's covert operations chief for Western Europe. Some of the techniques were known, but doctrine and experience were lacking.

The CIA's covert operations were, by and large, blind stabs in the dark. The agencies only course was to learn by doing, by making mistakes in battle. The CIA then concealed its failures abroad lying to Presidents Eisenhower and Kennedy. It told those lies to preserve its standing in Washington. The truth said Donald Gregg, a skilled cold war station chief, was that the agency at the height of its powers "had a great reputation and a terrible record".

Marc Pelletier

This pick was actually a pick on TWiT a few weeks back and, well, I guess it's my pick too this week.

So, if you want to download a free copy of Legacy of Ashes, go to audible.com/biotech and sign up for a 14-day free trial. If you decide not to keep the membership, well you can simply cancel, but you get to keep the free book. Either way it's a win-win. But if you have a long commute or you travel a lot audio books are really great.

Now, back to the interview with Mark Gerstein.

Dr. Mark Gerstein

[41:31] Well I think one thing that is useful just to sort of get across is, I guess, we talked a little about this sort of process of annotation as kind of affixing meaning to a bit of sequence. But I think it's sort of useful to flesh that out a little bit more and just sort of say well, the first step in annotation is kind of to go through the genome and say this little bit is a gene, this bit is a pseudogene, this is a regulatory element. But then the next bit is to sort of take those genes or regulatory elements and to start to link them together into networks, and to say all of these genes are controlled by this master transcription factor, or all of these genes produce proteins that then interact in a complex or part of a pathway.

And so a big part of my lab and what I'm interested in looks at the structure of these resulting networks. And so that network structure, I think, is actually a really kind of interesting thing. I mean if you – you can just think about a piece of paper that has a bunch of dots and you connect those dots with lines to create this network and you might say to yourself, well those – that patterning has a lot of interesting structures. Sometimes they can have hubs in the network that seem to be very central positions, other times you have little bottlenecks through the network or you can sometimes take a network that's a directed network where you've arrows on it and you can find hierarchies or things that are at the top, things that are at the bottom, or little motifs.

And so I'm particularly interested in this issue of network structure. And then another final aspect of annotation is when you start to really say to yourself, well I've got this network and I've got these nodes in the network. You really drill into them and you say what are those nodes. And of course the nodes in all of these networks, they're not little marks on a piece of paper, they're actually molecules. And the molecules on a very fundamental level have a three-dimensional

structure and they interact as an entity with other molecules and so another level of annotation is where you start to take all the molecules in the network and group the molecules into families of proteins and start thinking of them as proteins and then eventually you start to think of those proteins as having complex three-dimensional structures that can potentially move...

Marc Pelletier

They're machines, right? They're basically molecular machines.

Dr. Mark Gerstein

Exactly, as machines. And I think that's kind of neat because that's where you really make the transition, I think, between genetics and chemistry. I mean when – genetics is the world of AGCT, the world of this code, the world of the genome and chemistry is the world of the periodic table and the world of molecules and so forth and I think the really impressive thing about molecular biology is the linkage between these two things. The linkage of the abstract world of the gene or the abstract world of genetics where you just have kind of things associated with traits, to actually associating them with real molecules that have a three-dimensional reality.

And the impressive thing about Watson and Crick's original double helix in 1953 was here you have the basis of genetics. It's got all the AGCTs and all that type of stuff but it really states it in a molecular way as a real molecule. You can understand how the actual mechanics of the molecule is carrying out the functions of genetics and so I think that making that linkage in an annotation and also the conceptualization is I think really interesting.

Marc Pelletier

I think it's unbelievable, isn't it. You're sort of creating the infrastructure and piecing the data together to assemble a Google human, right? So that you could go in, look at, take – have a picture of a human but then zoom in and there is a fingernail and zoom in down to the atomic level, right? I think it's unbelievable stuff.

Dr. Mark Gerstein

Definitely. I think that – and actually what people, the vision people have is sort of a, I don't know if you've seen this book Powers of Ten where they kind of – you look at a person, you zoom in a factor of ten, you keep zooming in and you see cells eventually. You see organelles, you see molecules and then of course they keep on going down into protons, neutrons, electrons whatnot.

Marc Pelletier

Strings!

Dr. Mark Gerstein

But it is interesting at each of these different levels there are different types of organization that you see and different things kind of come into play and I think that what's, I think, very impressive is trying to kind of link up between levels. So what people are really interested in now is there's kind of this idea of molecular biology where you think of the molecules in biology and then there is also this idea of cell biology which you sort of think about cells as a unit. And now people are starting to think of, well, really big molecules that start to be players on the cellular landscape and they start to think of these molecular networks as now how do they map into the organization of this cell and so forth. And kind of spanning between these levels, these scales.

Marc Pelletier

You – when you look at the genome, right, you're seeing ACGT and the pages go on and on and on and on. It's tremendous. You can scroll down for 19,000 times – 3 gigabytes, right, 3 or 4 gigabytes?

Dr. Mark Gerstein

That's correct.

Marc Pelletier

So when you look through just the ACGTs and you're looking at that screen, do you see like, I hate to bring it up, I always bring up the analogy to the Matrix. So with Eric Kandel, I was wondering if you could plug in a USB into the back of your head and then learn Kung Fu, if you ever see that happening. He said, no, not right away but maybe in a hundred years, right. So do you – when you look at the genome and you've spent so much time and so much effort dissecting it and figuring out the little parts and piecing it together, do you see stuff in almost like the three-dimension, do you see just by glancing at it, look at regulatory elements, decipher where genes are encoded, introns stuff like that. Do you see the algorithms, are they built into your eyesight?

Dr. Mark Gerstein

[48:09] Yeah, no, that's actually kind of interesting. I think the neat thing about a lot of this genome annotation is, I would actually say the – almost the opposite in the sense that all the people who do this like myself included, I'm almost sure if you showed them a section of their genome, you just looked at the raw AGCTs it would be just like gibberish, right, it wouldn't mean anything to them, right? And the interesting thing really is that actually through the application of all of these computer matching techniques, of connecting things up into networks, suddenly this gibberish has this hidden structure that you could not see and that's really the essence of data mining, right?

When you have all this data and the – it is right they are in front of you, but only through kind of mining and looking for these hidden patterns can you see what is really going on and that is kind of impressive. I think when you – what's neat actually is to see when people show you a screen of just all of the letters, right, and you look at it and you're like, "Oh, doesn't look like anything". And then what they do is they color in, "Oh yes, these are letters associated with exons, and these are the letters associated with, say, a promoter and then these are the letters associated with another gene". And suddenly, you can see that there's this kind of underlying structure that was not apparent and what is neat, too is if you really look carefully at the letters that have been highlighted, sometimes you can kind of see that there is this hidden pattern that you could not see before. Sometimes you can see the triplet code for instance for genes or you can see some kind of conserved motif whatnot for a binding site, and then it very much has this idea of cracking a little code and seeing this hidden structure.

Marc Pelletier

Or a glutamic acid stretch or where you keep seeing repeats and repeats and repeats.

Dr. Mark Gerstein

Correct.

Marc Pelletier

So I bet you're really good at Boggle! Have you tried Boggle?

Dr. Mark Gerstein

No, I am actually not, I mean that is the funny thing like, I like all this type of stuff but I do not really think that many people in this business are actually good at looking at a bunch of letters and right there making sense of it. It is really through the computer. It is not – the other thing that is kind of neat about this is it is not really meaningful for a person to look at this stuff and make sense of it. The average Joe can look at a few thousand letters and maybe if they spend a day at it they can circle little patterns, but it is just not meaningful for a human to look at that many letters.

You really have to have a computer doing it. And there are some people that – there is a niche of people in bioinformatics who are really sequence gazers in the most literal sense who – a new sequence would come out and they would actually sit there and look at it, and they would highlight little features of the sequence and these people are actually quite amazing. Some of them have really special minds and they can see really interesting patterns in them and there

actually are people who can do this for protein structures too because protein structures also have very complicated structure. And when you look at it, again, it is completely gibberish, it doesn't mean anything to you when you look at it.

Marc Pelletier

Absolutely.

Dr. Mark Gerstein

But there are some people that can see some interesting patterns, but my personal view is that even though that is a powerful thing, it is impressive what some of these people can do. In the long run, that is not the way that science is going to move forward or that is not really what bioinformatics is about in the sense that I think it is not going to really scale to a really large scale nor is it going really provide objective interpretations of things. Fundamentally, this stuff has to take place through computer algorithms.

And this I think is actually a real eye opener to more of your traditional biologists, because I think a lot of traditional biologist are really people that are highly steeped in an observational tradition. They are people who are like, "Jeez, I like to look at the data, I want to look at it really carefully, I want to understand what is going on". And to have to interpose into that process of looking at something, this whole idea of computer algorithms and kind of computation, sometimes this is very foreign to people, frankly. And it is a kind of whole different process and I think that for a lot of people it is a cultural change, but I do think it is kind of inevitable.

Marc Pelletier

I'm, I guess, one of those gene gazers who likes to look at the sequence that just came out of the machine. I don't know, maybe I am always in a little bit of awe when I get a sequence back from a DNA sequencer, even if it's genes that I have just copied and pasted and put into new carriers and vectors and stuff. But it always seems almost mystical when you go down to the basic essential chemical of life, which is those four bases.

I will let you go. Thank you very much for being on the show. This has been a fascinating tour of the human genome and genomes as a whole.

Dr. Mark Gerstein

Sure, sure. So we are all set.

Marc Pelletier

I think so.

Dr. Mark Gerstein

Okay, and you got the, I don't know how I should say, my recording quality of my voice has held up through – you're all set with the acoustics?

Marc Pelletier

I am looking right now at the gain; it says 63%.

Mark's concerns are legitimate here. We actually did the interview twice. The first time we did it the gain was way too high and the over modulation just made it very hard to listen to. So many, many thanks to Mark Gerstein for coming back on the show twice, I really, really appreciate it.

I would also like to thank Tom Price and his team at Pods in Print for the transcripts, they are available at podsinprint.com.

I would like to mention Lori LeBeau Walsh has TWiT TV T-shirts that look really, really good. She is taking on the responsibility of producing T-shirts for this WEEK in TECH and you can get them at artandtechees.com, I'll put a link in our show notes.

I would like to also thank Phil Pelletier and Will Hall for the opening and closing themes. It is also very important to mention here that this podcast is supported in part by the listeners, thank you for your donation. For Futures in Biotech, I'm Marc Pelletier.