The following content is provided by MIT OpenCourseWare under a Creative Commons license. Additional information about our license and MIT OpenCourseWare in general is available at ocw.mit.edu.

[SIDE CONVERSATION]

GEORGE CHURCH: OK, welcome back to the second half of the second lecture. Here, we're going to take this beautiful example of an algorithm, which the cellular machinery applies to get from DNA to RNA protein.

And so a simple example where that is incorporated into a program very similar to what you'll be doing on your problem set. So here, we've got it nicely color coded. This is unrelated, of course, to the color coding and the genetic code we've been using so far, but it includes the genetic code.

You can see the comments here are preceded by this little number sign. We go from the genome, a DNA sequence, which is a string. Strings are one of the things that Perl does very well. It's not the only string-manipulation programming language, but it's a particularly easy one.

And so here is how easy it is to enter a DNA sequence, one of the many ways of entering a DNA sequence. You can bring it in from a file. Here it is as part of the code. You transcribe it into RNA in silico here, by the simple command where you say RNA sequence is equal to DNA sequence, and then you substitute all the Ts for Us globally. That's what this 12th line is on slide 23.

Now, off to the side here, is a reminder that it's really much more complicated than that. There's all these proteins involved and doing it accurately and with regulations, so forth. But for the sake of this Perl program, this is quite sufficient to get us to an RNA sequence, which we can then translate, and here, the translation process uses-- it's going to be a cycle. So in a sense, the cycle in the RNA m where you put in one nucleotide at a m is all compact here. It's all just substitute every T for U.

Here, we're going to have a more explicit loop, this while loop. You can see it's indented. Everything within the loop, that's going to be iterated, is offset a bit. And what you're going to be doing, is looking in groups of 3. That's what this 3 is on line 17 is looking through the position you are in the RNA, and taking a chunk of 3 at a time. The chunk here is a substring, substring. And you pull out a codon as that substring. And then you do a translation.

So now this is, in a sense, representing the modularity of biology and the modularity of good programming code. It's you put this whole business of translating in a separate part of the program, so you don't have to embed the code everywhere that it's going to be used. And the translation here, is that simple table. So this subroutine, S-U-B, line 22, again, indented all the code in the subroutine. It goes off the bottom of the slide and down through the floor all the different cases.

Now, we could list 64 cases for the 64 trinucleotides, or we can use the more compact string manipulation that you can do in Perl to indicate-- a dot means any kind of character after G and C, would return alanine so GC, ACG, or U.

And then for cysteine, which has two possibilities, so that's four possibilities represented by that dot. And cysteine has two possibilities. It's either UGC, or UGU, where the vertical line means or. And you get the idea. There's a whole, very compact, syntax here for doing the translation.

And that's how we do one of the cleaner, more simple algorithms in computational biology. And now I'm going to make it complicated again. But first, to set the stage for how the genetic code is not universal, we have to explain by what we might mean by it being universal.

This is the ultimate pedigree on slide 24 here. It's, in principle, some very simple organism, possibly an RNA-based organism that may have made a proto-ribosome that may have done protein synthesis or may have done some other chemical reaction at what is now [INAUDIBLE] transferase site.

Something along this line, it's speculated by some biologists, was the common ancestor of all living species, all living cells. And certainly, by the time we started getting branching of the three major branches bacteria, the archaea bacteria, and the eukaryotes, by the time we get to that point, there was, probably a ribosome set of genes that encodes all the proteins and RNAs or the ribosome, and these were shared. And then at each cell division, you split off two cells were slightly different from one another. As they would mutate and differentiate and be selected, they would generate this huge diversity of organisms, which you see here.

Now, this is a [? directed ?] basically graph in the sense that you can't have a descendant in this process being an ancestor of one of its parents. So time has, as an axis, going up in this case, unlike some of the more physics based diagrams we had before.

And as you branch out to existing species, you see that things like plants, have actually inheritance, not just along this direct tree-like structure, but you've got more of a network-like structure, where some of the genomic material came in from one of the bacterial branches long ago.

And this has recently been put on very firm footing on a genomic scale, in a recent article this month, where the bacterial genome of cyanobacteria, these are blue-green algae that fix carbon in all the oceans and plant. And the simplest plant that's been sequenced is Arabidopsis, a weed.

And anyway, the DNA from the bacteria has not only gone into the chloroplast, which is an organelle, which has a very reduced genome, but it's been spread through thousands of genes in the nucleus, which is the major place where all the chromosomes are for the plants.

And so this, possibly symbiotic relationship, has resulted in a complicated inheritance. And this is not unique. Another one coming in from the purple bacteria, has been incorporated into a separate membrane-bound organelle, which provides the ATP for both plants and non-photosynthetic multicellular organisms, animals.

And these two arrows, that could be there could be thousands of arrows going over long periods of time at deep branches, and we know there are many interconnecting arrows in recent times in almost all these organisms, or certain representatives from all over this tree, can take up DNA in various ways, can even mate with organisms of various species and exchange DNA and incorporate it.

And so it's not this simple tree, but it is, certainly, going forward in time. It is directed and acyclic in that regard. So how many living species are there? We're building to the point, connecting back to the central dogma, of how many different genetic codes are there. So we need to know how many species there are.

If you take a gram of some thumbnail of soil, from any of a variety of different soils that have been tested, you can find about 5,000 bacterial species. Well, what does it mean to be a species?

In animals, that typically means that they don't produce fertile offspring when they interbreed two different species. But there are books full of exceptions of this, even in animals. And bacteria, of course, where they exchange DNA all over the place, as I said, it breaks down even more.

So working definition that many biologists adopt, is that if two microbial species share 20% of their DNA, if you take their DNA and align it by algorithms, such as the ones we we'll be using this course, and you find that 70% of the base pairs are conserved, then they're the same species. Otherwise, they're different species.

And there are millions of non-microbial species, many of which harbor microbial species. This number is dropping slightly because of our inability to restrain our growth and other activities that cause species extinction.

And the number of whole genomes is getting closer to 100, and the number in the pipeline is probably 600 or so, maybe in the thousands, with new technologies. And there's over 80,000 species defined by one or more nucleic acid sequences in the NCBI, National Center for Bioinformatics, which is one of the three major nucleic-acid databases in the world.

Why do we study more than one species? The comparison between species allows subtle and not so subtle analyses of what are the important positions to stay constant because they provide some very fundamental biochemical activity. But what are the important ones to vary because they provide some important variants, for example, escaping immune surveillance and so on. So there's reasons to be constant, there's reason to be variable, and reasons to be neutral.

So let's go back now and apply this to the genetic code, this particularly simple and elegant, nearly-universal code. This is how genetic codes are represented in CBI, one of the ways. And here, the three bases of the codon, bases 1, 2, and 3 remember, we said that you UUU was phenylalanine encoded by DNA TTT. So going down from the top on the leftmost column, it's TTT single letter code F for phenylalanine.

So the amino acids go along the bottom row of this table. And you can see all the amino acids are represented. Stars represent stop codons which are not recognized by transfer RNAs but by proteins called release factors that simulate the function of the transfer RNA and cause the release of this cyclic incorporation of amino acids into polypeptides.

Now, this is the standard code, so-called, where you have one methionine here in the middle here, encoded by ATG, and three stop codons in all the rest. Here's where it gets complicated. There are over 22 different genetic codes. Some of the changes from the standard code are indicated here in blue.

You have here, for example, for the [INAUDIBLE] mitochondrial code, this is the code actually used in every cell in your body for the subset of the cell that is the powerhouse that makes ATP, the mitochondria that we talked about before, which was part of the horizontal transmission of information from purple bacteria long ago.

But anyway, the normal stop codon is now tryptophan, abbreviated W. There's an extra methionine. And there's two extra stop codons, which are replacing what would have been an arginine in the standard code. And you can see there's little blues all over the place. Changes in the number where you can start. The start sites that are indicated by 3, 1 to 3 starts in the standard code.

And when you talk about the starts, you start getting to how much do you favor that particular start? What other signals are required to start at that particular position? It's not as simple as just having an ATG trinucleotide to get a start of protein synthesis. You need other nucleic acid components.

Anyway, still, it's a slightly more complicated algorithm. You have to know exactly what organelle and what organism you're dealing with, but you can apply the same kind of computer codes that we had, a couple of slides back.

But now we get into even more complicated. And part of the reason I'm showing you this early on, some of these things would not be in your textbooks for the first biology course. And they would not be in a first computational biology lecture or two.

But I do this so that you'll have a healthy distrust of everything that you read and everything you hear, including everything you hear from me. And this should really make you distrust the genetic code.

Because what these ribosomes do, in this particular sequence, which has been well-documented, eight years ago, is that they will hop over 50 nucleotides. They're not going 3 nucleotides at a time as they should, in fact, it's not even an integral multiple of 3 nucleotides. It is literally coming to a stop codon, and rather than stopping, if it has just the right sequence context, including this complicated RNA secondary structure called a pseudoknot.

The RNA folds up the messenger, which really should just be a messenger, which the computer should just be recognizing-- or the biological, biochemical computer, as is the ribosome, should be recognizing three nucleotides at a time, instead it's recognizing a morphology. This thing folds up, and no longer is an informational molecule. It is a morphological recognition element.

Anyway, when it finds that, it skips over 50 nucleotides, skips over the stop codon, and makes a, otherwise, perfectly normal protein. So don't even trust dogma, especially don't trust dogma, central dogma included. Plenty of counter-examples.

Now, we're going to move from this very, very simple example of an algorithm, where we can model proteins directly from the nucleic acids that come out of DNA sequences wholesale. We now want to ask, how do we get the more quantitative data, which comes out of functional genomics rather than classical sequencing? How do we get that into quantitative models, and then get the quantitative models then repopulated with additional quantitative data to make a full model?

Now, in order to say, came up earlier as a question, what is the function of a gene product? We dip into qualitative statements which are made in the literature, which have various ways of representing the evidence for this. Some of them very convoluted arguments, some of them very casual.

But when an attempt is made to put these into a database or a data structure, as representative, gross oversimplification of the literature, this is what often comes out, something like this, where you'll have a hierarchical table. Here, I've blown up one of the levels of the hierarchy. You can think of as a list, where the list may not be in a particularly logical order, but the hierarchy is, so that under metabolism would be some covalent change in substrates, which enzymes would catalyze.

And then you'd have information transfer we've been talking about, like the DNA to RNA protein, these biopolymers. Regulation of information transfer or of metabolism would have all these four subheadings type of regulation, trigger, and so on. And then transport in these various other processes.

Each of these functions, such as illustrated by these references here, can be used in a way of connecting all the new information we get to some of systematic best guess, best estimate, of encapsulation of the literature.

Another example of this, in addition to the MIPS for yeast, is gene ontology, which is derived from the word ontology, or nature of being. And the objective f GO, abbreviation of Gene Ontology, is to provide a controlled vocabulary. My vocabulary during this lecture has been uncontrolled, as you've probably guessed.

But when you start talking about-- I have pointed out the problems that you get into when you casually refer to gene expression when you really mean RNA expression, and refer to genes as protein-coding entities, when you really mean protein or RNA-encoding entities. That process of being more precise about our use of terms, at least when we're communicating with computers, is very important.

We communicate with each other, you guys will give me a little slack, some of you but computers won't. They will misinterpret every chance they get. And so that's what control of vocabulary is all about. And you'll have three different-- Dave, the inventors of gene ontology, have a hierarchy including molecular function, biological processes, and the cellular component, which we'll expand upon in the next slide.

Cautionary note, whenever you do modeling or you will be assumptions in this case. Some of the assumptions exclude vast parts of biology, which are listed here as part of their documentation. They have things that are not modeled in the gene ontology are domain structure and three-dimensional structure, which, obviously, has played a big role in the two lectures so far.

The evolution and gene expression, we've already talked about the phylogenetic tree of evolution, and the gene expression will be a big topic in the RNA and proteomics part of this course. The small molecules we've illustrated today. Almost everything in this course, seems to be excluded from the gene ontology.

Nevertheless, here we go with just one slide talking about the functions. We have molecular function. What the gene product can do without specifying where or when. A broad example of this, would be enzyme, something that catalyzes. And then a very specific example of an enzyme would be, a adenylate cyclase, something that makes a cycle in the ribose of a adenylate.

So both of these fall under the net molecular function when you're describing a function of a protein in describing a genome. A biological process has to have more than one step. If it's one step, that's not a process. It has to have a time component, typically, and there's a transformation that occurs. Examples of signal transduction is a broad biological process. An example of signal transduction is cyclic AMP biosynthesis.

The cellular component, would be somehow reflecting this assembly to organelle that we were talking about earlier. And here, an example, you have a ribosomal protein being part of a ribosome. So it gets you some idea of the component, some molecular-function biological process and components.

Now, as I said, this gene ontology is based on facts. The facts that are included, it's not-- ideally, there would be a direct logical connection between the facts that are summarized in the hierarchical gene ontology and the raw data that came out of some instrument. That is not the case.

This is all from the literature, and it's done on a low budget, wow. And examples of how they summarize it, is it's inferred from a mutant phenotype or a genetic interaction. So these two are genetic. Or physical interaction, this passes for biophysics.

Sequence similarity, now we're starting-- as we go down this list, we're starting to get into murkier and murkier evidence. Sequence similarity, as you'll see in a subsequence slide, has problems. A direct assay could be a physical interaction, or it could be some other biochemical assay.

Expression pattern might be evidence of some of the associations that were mentioned in the gene ontology. Now we get to electronic annotation. In a certain sense, all of these things are electronic annotation. Sequence similarity might be a way that you automatically get electronic annotation.

Then you get to a traceable author statement. This means that someone said something is true, without saying how he or she knows it's true, so we're getting really murky. And the murkiest of all is non-traceable authors' data. You don't even know who said something might be true, OK.

Let's go back up to the top of that, in fact, go beyond the top of it, where we now will start tracking the data from the instruments to statements. And hopefully, in this course, you will see how we will, in the future and present, make models in a rigorous way where you can track it all the way back, to data.

So one class, the most obvious class, of data collection, is what I would call direct observation, typically through a microscope. And here's a particularly powerful case. I promised you earlier that we would talk about how you have 959 cells in the non-gonadal cell lineage of the worm.

It starts as a single cell up here at the top middle, a fertilized egg, as this egg-shaped thing at the top. And then it splits off, way off to the left and off to the right. And that makes two cells, two stem cells, that are capable of differentiating and dividing further. And they each make two more, and it keeps going.

But you can see it starts getting symmetry breaking, almost immediately. In fact, the egg itself is an asymmetric entity. And you start getting lineages that will either die as they terminate, or they will just stop dividing. And eventually, you end up with, after about 1,000 cell divisions or so, you end up with these 959 non-gonadal cells.

And this lineage has been completely mapped out by direct microscopic observation, where a series of photographs you can show that this single cell turns into these two cells, so you have a time axis, and you have a lineage axis, which is one of these directed acyclic graphs.

In addition, and even more amazing, to me, anyway, is that you have a complete neural connection for this multicellular organism. It has a fairly simple brain if you've ever had a conversation with one of these things. But each neuron can have dozens to hundreds of connections. And these have been mapped by a serial section through the entire worm, very thin sections and electron microscopy. And then checking out the whole wiring diagram. This is really a tour de force.

And part of the reason it's possible, this would even be hard to do in a variety of organisms, but this is another case where biology cooperates, just like with the genetic code, in this particular organism, that lineage happens the same way every time.

In even slightly more complicated organisms, like the Drosophila fruit fly or humans. The lineages are not so strict, and a cell can take on a number of different directions depending on the exact physical environment it finds itself in. But nevertheless, for this one, the neural connections are reproducible, and the cell lineage is reproducible. And so you can map this all out.

For other organisms, it doesn't mean you shouldn't try it, it just means you'll have to represent it in a slightly less-fixed pattern. You'll have to represent it as probabilistic set of divisions and probabilistic set of neural connections. And maybe even, conditional, on various conditions. OK, that's direct observation as a class of source of data for modeling.

Here's three other sources of data. In each case, I've shown pretty raw representations of the data. You can think of these all as representing an intensity readout, with some sort of separation, as the horizontal axis or is some cases, both axes.

So the intensity here, is indicated by a line plot of four different color fluorophores in an electrophoretic separation, which is the basis for the genomic sequencing that we're so proud of here. Then so here, the detection of this fluorescence of the terminated chains of DNA, we'll get to that later in the course.

Here you have mass spectrometry where you're measuring differences in masses, even more accurately than in sequencing. You're separating nucleic acids by their differences in mass, so about 1 part in 1,000. The mass spectrometry is more like 1 part in 10,000, or even better.

Because here, you're separating in a gas phase, based on the electrical and magnetic properties. Here, you're separating by charge in a liquid phase, liquid and gel phase. Each of these, you can specify the throughput per day or the throughput per unit dollar. This becomes important in planning these structures.

The third category here, is arrays. These can be arrays of nucleic acids for quantitating RNAs, or arrays of antibodies, proteins, small-molecule chemicals, which we can quantitate the binding of one kind of molecule to an array of other molecules.

In both the top and the bottom, you can have multiple colors. And these can be used quantitatively as internal standards so that you can monitor this process. See, we're going to go into this in great detail, later on. But I wanted to give you a feeling for where the source of these things are.

This array analysis, in a sense, is another example of microscopy. Just like in the previous slide, we used direct observation microscopy to monitor cell lineages. So too we can do it/ We can make a-- it's just wonderful. The battery's charged, OK.

We can take the microscopy of artificial patterns, such as arrays. Just as we have separation here by mass, we can also have separation on a variety of other properties, sometimes called multidimensional separation. This gets back to the first slide of the lecture, which was the purification aspect.

Now, how do we jump from that kind of raw data to this common way that biologists communicate in journals, where they have circles and arrows, where the circles might be some kind of protein molecule such as a stat, and arrow indicates some of interaction, or regulation, or quantitative influence that one protein has on another, or a protein.

So in alternative diagrams, nodes could be small molecules, and the edges, the links between the nodes, could be an enzymatic reaction catalyzed by a protein. There are about 500 biological databases that we'll talk about in the database talk.

How the data and models were entered into these databases, is a huge issue. Many of them have been done very casually. For DNA sequencing and crystallography, I think the process by which you go from the raw data to the models, is very well understood, very well communicated, for this sort of thing. It will take this whole course for us to even to scratch the surface.

Here's another example-- that one was protein-protein interactions. This is an example where the nodes now are not proteins but small molecules. And they're connected by an enzymatic pathway.

This is another example of an application of ordinary differentiable equations, just like the one last class. We had exponential growth. Here, you have simple fluxes, where a catalytic reaction occurs, not autocatalytic, but catalytic. There's no there's no exponential growth occurring in this cell. It doesn't have any biopolymers in it, biopolymer synthesis.

But these catalytic reactions that form this network, and you can model the influx of fresh molecules. And its utilization within the cell and the efflux. We'll come back to that.

Inside there, are a set of kinetic equations. We need to figure out how to get from the raw data types that I've shown you to this kind of equation. This will be one of the goals of the course.

Here, you have a velocity on the far-left hand side of the top equation, which is related to a maximum velocity on the numerator. And then a series of linear sums and quotients. Now, some of the terms will be nonlinear.

Here's an exponent of 4 that enters in, because you have one of these [? propertivities ?] that gives you that kind of sigmoidal curve that we showed for transistors and will enter to a number of biological consequences, where the steepness of that sigmoidal curve is determined by this exponent, sometimes called a Hill coefficient.

But other than that, you'll get these simple, linear sums and quotients. And we'll come back to that.

What actually constitutes these networks? I want you to feel less limited than you might get in a simple textbook. A simple, textbook definition of a catalytic enzyme-catalyzed process you might have, A is a substrate that turns into B as a product. This is a process that A could go to B spontaneously, but in the presence of enzyme, it goes faster. Or it could be, that for all intents and purposes, A never turns into B. It's so slow that you need this enzyme here to even detect it. The enzyme will form a complex with A. This could be a non-covalent complex or a covalent one. It then produces a covalent change in A. Ant it becomes an enzyme-bound B. B is released. Enzyme E is regenerated.

And so in a certain sense, in this process of turning A to B, E is not consumed. But let's think about an increasingly important class of biochemistry, such as signal transduction, where the enzyme now has a new role. It changes places with the substrate. It becomes a substrate.

The E now is a substrate in which a small molecule, ATP, which might have been the A up here, combines with the E. And the E could either catalyze its own phosphorylation or in context with another enzyme, but in any case, it becomes covalently modified to produce a phosphorylated enzyme, a phosphorylated protein.

And then the ATP is regenerated by a simple enzymatic process. And so in a certain sense, formally, it's very similar to this process, except you now flip the enzyme and the substrate. ATP is not consumed, the small molecule is not consumed. The enzyme is consumed.

So think of these things, these networks, as symmetrically as you can. Try not to get too embedded in the names, this is an enzyme, this is a substrate, and think more about the concepts. The concept here, is some things are consumed and some things are catalytic and regenerated.

So again, we are going to integrate this metabolic processes we were talking about, in the last couple of slides, with the information flow, which was the topic of central dogma, in order to get functional genomics, which measures those information molecules mostly and produces quantitative modeling.

You need to have the qualitative models to know what's connected to what. You need to have the raw data as illustrated in slide 41. Again, to remind you, the source of quantitative data here, you can measure RNA, or proteins, or peptides in the mass spectrometry. RNA in the arrays connected to the DNA provided by the DNA sequencing.

I warned you that one of the gene ontology data type sources of data was sequenced-- electronic sequence annotation by sequence similarity. I want to elaborate on this warning with this slide, where we say, we have various justifications for looking for distant homologs, examples of gene products, which are related by, on that ultimate pedigree tree of life, by very long distances.

It's been a long time since those things were present as a common ancestor. And we want to find those because they help us limit the number of hypotheses that we need to test whenever we find a new molecule. If we can connect it to another molecule, however distant, then we feel that we don't have to test every possible hypothesis. We just have to test that little narrow one.

But what happens when we do that? Let's say, instead of some distant homology, where we have, say 20% amino acid identity. You line up the sequences by methods that we'll talk about later. And you have 20% of the positions that are the same, or even less, can sometimes be meaningful. But how good is that?

There's going to be some kind of curve that relates how close two proteins are with a probability that they will have the same biochemical, or cell biological, or genetic function. And here's some worst case scenarios. And I don't mean to represent these as typical, but they get you doubting again so that you don't trust anything.

100% sequence identity. This should be a best case scenario, but it's not. The amylase enzyme, which catalyzes a carbon metabolism in most cells when it's expressed to high levels in a vertebrate like our friend this marine tortoise, turtle, it turns into the major eye lens protein.

And actually, this is true of most vertebrates. They have some kind of enzyme, like a glycolic enzyme, which is overproduced and aggregates and makes a clear lens more morphologically interesting feature, which just focuses light. Completely new function by all those definitions of function. No longer does the enzymatic activity, does an optical activity instead.

Another example, we have 100% sequence identity. Not something really distant homolog like 20% or 10%, but 100% sequence identity. [? Thyroxine, ?] which is involved in redox reactions involving [INAUDIBLE] and other things.

In the right context with other proteins, it can now be part of a DNA polymerase, when it globs on the DNA, it goes really without stopping with [? thyroxine, ?] but it falls off if [? thyroxine ?] isn't around. That's not a redox function completely different biochemical function.

But like I say, there will be a curve. Sometimes, there will be very great hypothesis limitation that can come from very distant relatives. These are more examples of the quantitative data that we will use to get hints at relationships among genes that go up and down together.

They form the basis of asking, what is function not based just on sequence homology, but based on a variety of quantitative data, such as the RNA data and the microarrays.

This is three more ways of looking at how we define functions. Function definition number one, is the effects of mutation on fitness. This is, in a certain sense, what the organism cares about the function of a gene product. It's how many grandchildren am I going to have? That's what it cares about.

And that's what shaped the function over time. And so if we're going to understand any of our other definition functions, we have to at least give some attention to what shaped it over billions of years and over many different environments. We need to have some feeling for the ecology of these organisms.

The second definition is the more commonly used one, which is what is actually its function in a machine-like sense. In the cogs, in the wheels, how does it function structurally? What's the three-dimensional structure? What's the mechanism?

The third function, is more forward-looking, not what good has it been to organisms in the past, but what good can it be to us in the future or to other organisms in the future? This may not involve reproducing the organism, making copies of it. It could be that there's some other engineering goal or objective function.

When we say that we've proven something. We've proven a biological hypothesis. What we mean is, given the assumption, it's a statistical statement that the odds of the hypothesis being wrong are less than 5% of the time keeping in mind hidden hypotheses and multiple hypotheses.

In genomics, it's all too easy to collect a lot of data, and therefore, when you mine the data, you can make a lot of hypotheses. And you test them, and you find you will find thousands of things, which by themselves would be significant at the 5% level, the standard statistical test, but you've got to correct for the number of hypotheses you implicitly or explicitly test. We'll mention this time and again in specific cases as we go forward.

The systems biology manifesto that I mentioned earlier, had this little loop where you would generate perturbations and test things and so forth. But an alternative way rather than doing additional experiments, is if you really have bought-in fully into systems biology and you really have all the components and a systematic perturbations, then you might be able to test the hypothesis generated by data mining one data set by going into another data set.

You need to ensure that they are independent. And you need to ensure that the hypothesis itself, came from the first data set and not the second when you go out and test it. But that would be a pure data mining loop, systems biology loop.

Now, just like when we say we have a proof, you should be distrusting of anybody that says I have an absolute proof. What they really mean, is a statistical statement. So too, when someone says, when they refer to the quality of their data, is this is the answer at the raw data level, what they really mean, is that they have some error level that they can quantitate.

And you should be especially distrustful if someone doesn't attempt to give you any feeling for that. Not to say that everybody that gives error bars or error estimates is to be trusted, but you get the idea.

So for DNA sequencing, there's a standard of practice. It was not always such, but a meeting in Bermuda, it's called the Bermuda standard, this is the best place to establish standards, is 99.99% accurate. You can see they have very high standards in Bermuda. But that's across the Genome Project.

These are aspects that, I think, we got from genomics, in addition to the raw data, we've got kind of an attitude. The attitude is, we can start looking at whole systems again, less on the individual gene-hypothesis driven standard NIH grant proposals that predated the Genome Project. Now you can do less hypothesis driven, you can do data mining, and so on.

We've also inherited the concept of automation the modeling and completion. Completion is something which still is not reduced to practice for functional genomics, but it has been reduced to practice for sequencing. And there is hope that we can approach it for functional genomics.

Be careful using the word impossible. There certainly are things that appear not to be cost effective at any given moment, but technology is moving quickly enough. Remember those greater than exponential curves in the last lecture. There are technologies arriving that make things suddenly become cost effective.

And that's a particularly important warning when you're designing a computational method that will compete with an experimental method, the experimental method suddenly becomes cost effective, then you need to revise your computational goals.

We have types of mutations that we've talked about. We have a null mutation, for example, phenylketonuria, which is tested in newborns, in almost all newborns, that are born in the United States and certainly Massachusetts. This is a very serious source of mental retardation completely wiping out that gene.

Small dosage effects, like a 1.5-fold effect that we talked about in trisomies, like Down syndrome, are important. You have conditional mutants. Classically temperature sensitivity of a mutation, meaning the protein unfolds. Or more recent enthusiasm for chemicals, a mutation which depends upon a chemical for producing its phenotype.

You can not only have these things that affect dosage or condition or complete knockout, you can have a new function to obtain for changing the ligand specificity or changing the aggregation of a protein. Here in the background, are how a change in the hemoglobin, which normally transports oxygen, can change the morphology of a cell and hence the function in transporting oxygen.

I just want to end on two slides on how you can represent the competition among cells or among organisms, which represents the Darwinian function, function number one, a few slides back. Here you have mutants in a population. Selection acts on populations, and mutations are tagged, by definition, by their nucleic acid.

If you can use the tags, you make a pool of such mutants, these are naturally occurring population. And when these pools are subjected to selections natural, or complex, or simple, or in the laboratory you can now read out these tags in many of the quantitative ways we talked about, for instance, mass spectrometry, arrays, so on.

And as you go through more rounds of selection, you'll eventually pick the winner, which is the most highly selected of the mutations, the winner. Or you might have a mixture if you through a very limited number of rounds. This will follow the exponential curve that we had here, whether it's exponential decay or exponential growth.

You can have a very subtle difference in growth, due to the function of that mutated gene product, but that small, say 1% turns into complete all or none replacement if you have enough generations. This is the awesome power of the exponential that we talked about last time.

And in real world, and also in the laboratory, you can think of this as going over a variety of environments, E, over different times. So the time you spend in each of these different environments, has some unit that happens. In natural environment, you'll spend say, more time in one condition than another. And the selection coefficients are a simple sum, and this exponential gives you the ratio of the organisms.

Here's some references on this. And I urge you to take a look at these, where actual experiments have been done getting these. And we'll come back to this later in the course. So this is the end of this lecture number two. Thank you very much.