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.

GEORGE CHURCH: 101 or [? EBIO101, ?] all the other numbers. Usually we'll start these with a very brief, one slide overview of last lecture with a slightly different angle on some of the topics, and then-- and then we'll go over a preview of today's. So last session, was more an emphasis on the biological side of computational biology, and this one will be more computational side of computational biology. But obviously, both of them are-- we're trying to interweave them already.

In the first one, we were specifically looking at the simplest components and the simplest systems in living systems and computational systems, and how self assembly-- here defined broadly to include symbiotic relationships between different living forms and different human inventions. And self assembly is a very critical part of biology. In the mathematics that we use, both symbolic and numeric, you must know about the approximations when we talk about the various theories going from the somewhat subatomic to the population level.

There's an approximation at each level that subsumes previous approximations in it. And when you represent floating point numbers and computers, there are further approximations, which can accumulate as you do calculations. And we saw some of the ways of coping with this, such as using higher precision arithmetic in Mathematica.

The other aspect of this was differential equations as a tool in studying replication, in particular autocatalytic systems which were illustrated by this really simplest possible differential equation where the incremental increase in y is a function of time, the vertical and horizontal axis is a direct function of y. A simple linear function of the-- say the population size, its growth being proportional.

Then we add this extra term, 1 minus y, to-- to represent rather than just the simple exponential curve that you get at the beginning with a population growth in infinite resources. As you get close to the maximum resources here, one, you start to either plateau as you would if this were a ordinary differential equation, or we start to oscillate as you do this with an iterative solution, which we illustrated. And if the rate constant K gets large enough, then you get chaotic behavior where it can go close enough to zero that you're effectively simulating an extinction of the population.

This-- the issues behind replication and approximation come together-- or came together in the concept of mutation, and in particular, the mutations that occur-- that might occur in single molecules, life being full of very important single molecules such as DNA. And here, even though you know there must be stochastics underlying this process because it is a single molecule, there has to be some way of overcoming it, because we went through a calculation that indicated that the 46 chromosomes in every one of your cells in your body has to be replicating quite faithfully, or else you'd be getting cancer.

And that noise is overcome by the single moleculeness of the DNA being compensated by a multi-moleculeness of-- of energy containing molecules such as ATP and associated proteins and so forth. So now we've brought together the approximations and replication, and what this-- what replication leads to are pedigrees. Pedigrees are an example of many that we listed of directed acyclic graphs will have more pedigrees today. In a certain sense, the ultimate pedigree will be shown today.

And the mutations also can be nicely modeled by-- depending on the exact application of your thought about mutations, by the binomial, the Poisson, and the normal distribution, the normal being continuous and the other two being discrete. And selection figured prominently-- and optimality figures prominently in biology. You typically won't lose a bet betting on optimality for at least exactly the circumstances under which the system has been subjected for millions of years.

So this is the outline for today. Again, supposedly more on the biological side. We'll go through how purification has played a central role in the reductionist approach to biology and biochemistry, and how that purification is-- is also the antidote to the reductionism in that it provides a way of creating synthesis in going back up to systems. Systems biology is the second topic, and this is both relevant to the models, applications of models, and making the interconnections of the components ultimately in a synthetic loop of discovery and recreation perturbation.

Then this is the ultimate pedigree that I was talking about. Continuity of life and how it applies to central dogma as the illustration of one of the most robust algorithms that we have in computational biology, which is the genetic code. A truly elegant discovery, and elegant partly because that's what-- the biology underlying it is so amazing. Then we go into the issues behind how we get qualitative models from quantitative data, and then how we go from those quantitative-- qualitative models and fill them again with the quantitation that's required for truth simulation, prediction, and design. And then we end up again on mutations of selection just as we did last time.

OK. Familiar face-- the periodic table. Just like last time, the elements that are involved in the three major biopolymers in the cell. DNA goes to RNA, goes to protein is the central dogma, and all of those can be made of these six elements plus counterions of the polyanionic. So those are typically sodium, potassium, and other counterions. And so the total of 19 elements you can find in almost all life forms, and I would say many of the other elements on here that at least have some stable isotope in nature probably have some story that some living organism could tell if you were interested enough. And that might make a good thesis project or a good project for this class to talk about, what organisms do to either detoxify or to use in some exotic way each of the stable isotopes on this chart.

Now, most organisms do not use these in the elemental form. They typically-- that in the middle here are elemental forms of these elements. Oxygen, hydrogen, and nitrogen are commonly used elemental forms, but there exists life forms for which none of these are really required. For example, oxygen is toxic to a variety of so called constitutive anaerobes, and nitrogen gas is only used by nitrogen fixing microorganisms.

Each of these elements exist in a gaseous form. They tend to be more reduced-- that is to say, more hydrogen containing on the far left hand side of the slide, and they tend to be more oxidized on the right hand side, and in the oxidized form tend to be found often as salts. Here carbonate, for example, being the salt form of carbon dioxide in the oxidized form that's fixed by plants and photosynthetic bacteria in the oceans, and unfortunately in its CO2 form, major global warming gas. And you can see that every of the basic elements here has-- can be obtained as salts. Many organisms require much more complicated versions of this, require eating macromolecules like steak, and-- and maybe have even very much more exotic ones.

As we purify-- when we represent the elements in that periodic table, those were the hard work of chemists that had to purify each of those elements. Not just as molecules, but it wasn't sufficient to get hydrocarbons, it had to get pure carbon. And then-- but that was only the starting point for determining not just the properties-- the elements-- but putting them back together as molecules. And this reduction down to elements and then re-synthesis back to molecules that was part of the proof that they really understood what the molecules were all about.

And the same thing can be obtained in living systems. We typically start with molecules that are a covalent connection in some form or another. And when they non-covalently associate, they're called assemblies. Typically the assemblies that most biologists work with are assemblies of proteins or proteins and nucleic acids. But there are many, many kinds of assemblies and these assemblies blur in their-- in their definition into organelles.

Organelles are basically assemblies that start getting large enough that they can be seen under simple micro-- microscopy scenarios. Organelles are not always, but are often bounded by lipid membranes. These are hydrocarbon containing bilayers or multilayers. Anyway, the distinction between macromolecular assemblies and organelles is soft. And then cells are collections of these assemblies, again typically bounded by a phospholipid bilayer.

And some organisms are cells. Unicellular organisms are vast majority of organisms on Earth, and they are-- many of them are single cellular aggregates of cells. While multicellular organisms such as yours truly and everybody in this room can contain up to 10 to the 14th cells that are direct and regulated descendants.

Now what are examples of purification methods that are actually used in-- on the road to computational biology? We have chromatography, electrophoresis, and sedimentation-- are very common ways of separating molecules, including protein molecules and assemblies. Actually organelles cells as well can be separated by these media. By far, one of the most common ways of separating these protein molecules is via chromatography and electrophoresis, and we'll see some examples in just a couple of slides.

A very incredibly powerful way of separating entities in general is represented here on the far right hand side. This is clonal growth, or-- this is essentially an analysis of single molecules or single organisms. Each of these colonies, which might be growing on a Petri plate about the size of my fist, is a-- represents the growth from one starting cell to 10 to the 8th or so final cells. Finally going through an exponential growth just as we went through in last class until the point that they have depleted the local resources near them, or have produced enough toxic waste products that they have slowed down their growth to form these colonies.

In certain organisms they'll just keep growing until they get to the edges of the Petri dish, but this is much more general than just growing bacterial colonies. You can see it, almost any organism that has a limited ability-- limited motility, these will form little clones such as-- this happens in various tree populations, for example. Also it represents the ultimate purification. In one step, you're getting something that would by combinations of different chromatographic steps and electrophoretic and sedimentation steps-- you might have to serially do several of these steps in a row to get a molecule or assembly purified.

While here in one step by essentially limiting dilution-- you dilute the-- the molecule of interest to the point where it's a single molecule. Well now if they don't undergo clonal growth, this is not terribly useful because it's very hard to study single molecules, even if they're well isolated from all the contaminating molecules. There are ways, and we will talk about them in the course, but you need-- ideally you need some way of amplifying them.

There are ways to amplify nucleic acid molecules such that they exhibit clonal growth like this. And in principle, any-- either by putting them into a bacterium so the bacterium behaves in this manner, carrying along with it the artificial piece of nucleic acid you're interested. Or you can do it entirely with enzymes so that the nucleic acids replicate and make these colony like objects.

Now this is an idiosyncratic view of this purification-- of this process by which we as scientists have gone through purification and then are returning to much less pure systems as a subject of our-- of our research. This is-- we'll call this three revolutions. In the pre-1970s, we had column chromatography-- so called chromatography in that last slide, because literally the substances that were being separated were highly colored as were those bands in the last slide, those two dark bands. Hence the name chromatography.

It's really a separation by the properties of the solid phase and the properties of the mobile phase, and the molecules in the mobile phase being separated by differentiable absorption to the solid phase. Gel electrophoresis and sedimentation in a gravitational field-- these were all part of this amazing revolution that allowed us-- allowed scientists to get molecules, assemblies, and cells purified away from other contaminants.

Then recombinant DNA did that trick that was in the lower right hand slide of the previous slide, which was going directly to purification by going to single molecule isolation by dilution to the point where you had less than one molecule per cell, and less than one cell per square centimeter on the Petri plate. That gives you single step purity of the gene, which in effect allows you to get single step purity of whatever is encoded by the gene-- the RNA, the protein, or the enzymatic activity.

This was a huge change. Suddenly everybody was spending a lot of time so called cloning of DNA and sequencing it, and almost every thesis and paper of that time-- everybody was turning into molecular biologists in order to do this, and it became very routine and very time consuming and expensive. And so the third revolution was automating this and using economies of scale so that all the genes were obtained at once and sequenced at once, rather than going through here where you would have to go through the entire library of all the genes just to find your favorite one, and then you would sequence that one working hard on isolating it away from everything else.

But with sequencing genomes, it was more a process of everything you came upon was interesting, and so you didn't have to do quite as much time selecting, you just made it more of a production effort. But the subtext of this was not just automation and-- and economies of scale. It also started to return us to thinking about whole systems and doing things systematically. And this was particularly valuable in the sequelae of genome sequencing, which was functional genomics, which we'll talk about quite a bit in this course-- applying the same attitude to other biological measurements, and this returns us to whole systems.

Now that leads us to the discussion of whole systems modeling-- systems biology and the models therein. So we have-- this is just one of the earlier papers. There are many now that are trying to-- we're trying to grope our way towards what we mean by system biology, but this is paraphrasing from that paper. We want to follow these four steps as a protocol to find all the components of the system. Systematically perturb and monitor the components of the system so that we can do this either genetically or environmentally, meaning changing the small and large molecules that program the-- the biological system from the outside.

Then refine the model which you had maybe before perturbing it such that the predictions most closely agree with observations. Listen carefully to that statement-- refine the model so that predictions agree with observations, and then do new perturbation experiments to distinguish among the model hypotheses. We do this in a cyclic fashion, basically going back up to item two so that we're perturbing and monitoring.

Now what is the-- what's the critique of this systems biology manifesto? We have-- those of you who have read books that predate the genome project and systems biology say, hey, this-- what's new here? This is the way biologists were doing it even before recombinant DNA. So it is an old approach, but the new spin on it is that-- is the word "all components." Typically before, the components would be chosen and the perturbations would be chosen based on the latest biological fad or what was available technically at the time based on the history of the component studies before that.

So it's a significant deviation to now even set as a goal all components. A very challenging goal is been met in the case of certain genome sequences, it has not been met in any functional genomics that I'm aware of. But it certainly is a goal, and we're getting asymptotically close to it just as we got asymptotically close to the genome sequences. To systematically perturb has the conceit that we can list all the perturbations we would want to do and then walk through them in a systematic way rather than a-- a more whimsical way.

So this is new spins, but what's missing from this manifesto in the previous slide that-- in systems biology? For one thing, and I cautioned you in the previous slide that when you start fitting your model to the data and refining your model, there's a problem of overfitting. And this will come up a couple of times in-- in this course. If you have enough adjustable parameters, you can fit almost anything, and so you have to be careful that as you refine your model that you state exactly how many adjustable parameters you have and-- and how many data points you have that are truly independent for fitting that.

So we have methods to recapture unautomated data. There is a step implicit in the previous one-- actually explicitly stated elsewhere in some of these papers-- that when you have-- as you're developing the model, you will draw not only upon the systematically collected data, but also upon the literature. And the literature, as we'll see in the next couple of slides, not only has unautomated data, but it has models that are derived in a variety of-- of somewhat undisciplined-- or a different discipline. It's not an electronically compatible discipline.

And so there's a process by which one captures this unautomated data and integrates it with the automated data, which can be either challenging or pathological. So we need to make more explicit these-- the logical connections that are used for deriving these systems biology diagrams and quantitative models.

Finally, when you make these perturbation experiments, if they're done using-- there's a new optimization that needs to be made in order to integrate them with the systems biology loop. As I mentioned in the previous talk, the thing that makes the killer applications in computational biology so far are searching, merging, and checking. If you can find ways to-- to search, merge, and check large data sets via models, then you've made a great deal of progress. And that should be the goal here too.

So what the systems biology will do, I think possibly more well illustrated by this slide and the last talk than by some of the examples so far in the literature on system biology-- but the goal is to be able to work with very simple parts, this reduction down to the basic parts, and then move up through models that are hierarchical and include all these intermediate steps to very high level ways of describing, understanding a system.

This one you have the unfair advantage that the entire thing was designed from scratch by humans, but in biological systems you want-- you want to reverse engineer it to the point where it has some of the same flavor of [? board ?] engineered systems, and then develop ways of simulating the systems so that you can design new versions of it.

Whenever you find yourself doing an experiment or computational biology, you're going to be asking yourself, why am I doing this? Why am I in this classroom? And whenever you do that, you should ask yourself why a bunch of times until you get down to the real core reasons. And so for example, we were in effect sequencing the genome prior to the genome project. We were spending hundreds of millions of dollars of NIH money and other-- every funding organization's money doing it in a very inefficient way.

And we knew why we were doing it, I think. We wanted to map variation in sequences, variation within a species like humans that make us different from one another, variation between different species, which is comparative genomics, item three. And in between item two, we wanted to have a complete set of human RNAs, proteins, and regulatory elements, and for every other organism too. I just use human as an illustration because it was called the Human Genome Project. And we wanted this complete set so that we could go back and measure them systematically.

Although this was not articulated in any-- in any way, it was-- we never did-- we didn't use the words complete or systematic very much prior to the Genome Project. And if we said these were the reasons, then we could ask why do we want to map variation, why do we want a partial or complete set of these various molecules and regulatory elements? And why do we want to compare them in different organisms and in different environments?

And the answer to that would be that we would want to make quantitative biosystems models, such as the ones we were describing the last couple of slides, of the molecular interactions at all the levels extending from atoms to cells to organisms and populations of organisms, because it's the population upon which selection acts, and it's the population that allows us to understand and make useful products.

And so when we ask, why do we want to make these biosystems models we have three reasons, some of which we touched-- all of which we touched upon last time when we asked, why do we model? Same thing, why are we collecting all this data to model? We model so that we can share information and so that we can construct a test of understanding. One of the tests of understanding is making useful products.

So on that theme of why and making useful products, I will put this in the context of, say, the projects that you'll be doing for this class. I like to stimulate you to think about this early on, In every class I'll mention something. And you might say, well grand challenges-- grand and useful challenges are really not a great place to be doing a short term project for a course. But actually I think that the piece that you choose should be a piece of a grand challenge because it really gives you the context. And so I'm just going to walk through these three classes of challenge, not so you will feel limited but so that you will feel broadened in your mandate for what you can do.

So at the simplest level, kind of reflecting last lecture, simple going from atoms to small cells with small genomes, maybe even minimal or miniature. Just like you want to downscale electronic components, we have, if we really want to show that we understand the biosynthetic route that will be the topic of today, this really key mechanism by which we go from DNA to proteins, we should be able to take it apart, put it back together again. That's the way-- that's one way of proving that we really understand it.

That has not been done in a purely synthetic route. We have taken apart protein synthetic apparatus and put it back together again, but we have not completely synthesized the synthetic apparatus. That sounds odd, but that would be one step, but the impact of that would not be just proving that we can do it. The impact would be that we can now make a simple biological system that is self-replicating, uses proteins, and allows us to link the atomic changes to population evolution.

In populations of humans, this would be daunting computationally. But when you think of populations of self-replicating molecules such as the simplest one last week, which is these trinucleotides being ligated into hexanucleotides, you can start to conceive of actually connecting the atomic modeling to the population modeling, which is basically the breadth of this course. It covers the whole breadth, it was collapsed down to that simple model.

But more-- even more importantly, we can start engineering smart materials. Materials that have important properties that in a certain sense compute in chemistry. We can make a whole alternative chemistry of stereospecific meaning sensitive to the actual handedness of the molecules, which is so critical in pharmaceuticals and in enzymatics in general. This can be engineered by getting a handle on the synthetic machinery of life.

OK, so that's one that's at the simple end of the spectrum. What about going from-- that's going from atoms to cells, how about cells to tissues? When we typically-- and many of you are either already in the biotech pharmaceutical industry or feel that the research that you might be doing as a graduate student would contribute to that in some way, however indirect. The way that you would program a computer, we might fill up this room full of laptops and go pouring random chemicals on them to see if they then produce the graphical user interface that we desire.

This would be the drug screening approach to programming. And obviously the way we program actually is we work in the natural biopolymers of computers, which is the strings of zeros and ones represented in the computer. And we program those as long strings, and so the equivalent in cells if you-- might be to manipulate the genome itself. We're doing genome level programming either at DNA, RNA level via nucleic acids. This is not an either/or, this is probably something where one is augmenting by studying and programming at this detail level.

And to do this manipulation of stem cells is a growing avenue of research. This gives us access to STEM cells and cells that are capable of replicating and differentiating into almost any cell in your body, and rather than dosing your entire body with a drug, you can now specifically deliver a particular cell to a particular place, and have it take its-- its role as a replacement. And so this is the kind of programming I think we should be thinking about as a grand challenge. Remember, grand challenges are not going to happen tomorrow. You have to do some piece of it. Question?

AUDIENCE: In the first bullet in B, since it appears we don't really yet know what the function of a protein is going to be unless we know how it folds, right?

GEORGE CHURCH: Right.

AUDIENCE: So we might be able to model the sequence, but how would we model anything other than that?

GEORGE CHURCH: So the question is, what do we know about-- or, rephrasing the question. What do we know about a protein before we know its fold? And actually historically we knew more about protein-- much more about protein function than we knew about their folding because the-- and we'll get to some of the definitions of functions in just a moment. But you can study it biochemically in terms of what it binds to, what place it-- it holds in the replication of the cell, and so on.

We do know the folding of most proteins, and that will be-- part of the post genomic era will be producing the three dimensional structure and biochemical function of all the proteins.

AUDIENCE: But here you're asking for changing that, changing the genome programming, right? And then try to--

GEORGE CHURCH: Of the parts that you understand.

AUDIENCE: OK.

GEORGE CHURCH: I mean, obviously we do engineer a variety of physical and biological systems without full understanding. An even grander challenge would be full understanding. Here we're trying to take subsystems that we do-- it could even be a highly integrated system where you do model the entire system, but there will always be gaps in your knowledge, just like there are gaps in the human genome sequence.

And the final illustration, number C, is going up to the most complex systems that we're dealing with, which would be morphological systems and even-- and the populations that result from that. And here we will be-- you can deal with morphology in a way that-- at the molecular level all the way up through the morphology of assemblies of cells, how cells aggregate. And all this can be modeled and used to great effect, whether it's smart materials or replacements in human systems.

So let's talk a little bit more about these components and how they're interconnected. Whether we are taking these components apart or putting them back together again, we need to understand how many of them there are and-- and how we're going to access them in databases. I'm illustrating this with three organisms that are nicely poised to show the extremes on the left and right and something in between. So this is mycoplasma pneumoniae-- sorry, mycoplasma genitalium, one of the smallest living organisms and smallest genome.

Its genome size is a little over half a million base pairs. The worm caenorhabditis elegans, one of the first metazoan multicellular organisms sequenced, was a little less than 100 million bases, and the human at 3 billion bases. Neither the worm nor the human is completely sequenced, despite some possible indications to the contrary. There are quite a number of gaps in each. The many bacterial genomes are completely sequenced, including mycoplasma. The number of DNAs in each of these-- you have one circular genome in many bacteria. Some have multiple chromosomes. The worm has seven chromosomes and human has 25.

Those of you who have studied biology or just listened in the last lecture where I said we had 23 pairs of chromosomes to segregate, you nevertheless, there are 25 different kinds of molecules. I'll leave that as an exercise for you. You can ask me in just a couple of minutes.

The number of genes encoded in these DNA depends on your definition of gene. If we define it somewhat arbitrarily as a piece of inherited material that encodes one or more RNAs where those one or more RNAs share some of the same inherited material-- so in principle there's inherited material, which is not nucleic acid. But for most intents and purposes, these-- the genes that we'll be interested in do pass through RNA on their way to protein.

And the number of genes in mycoplasma is roughly one gene per kilobase-- it's about 500 genes or so. The number in worms is estimated at around 20,000, and in humans it ranges from 30,000 to 150,000. And there are betting pools on exactly how many there are, and there probably should be a betting pool on when it is we will know how many there are. It's been announced at various times, but to some extent the exact number will have some softness to it because some of the genes will be of marginal utility to-- to humans. They will have been of some consequence maybe many generations ago, but on a day to day basis they will be hard to detect the importance of whether that gene is present or not.

In terms of RNAs, in bacteria you have a tendency to have more genes than you have RNAs because the genes will be constructed in a series such that one RNA can make it through multiple genes in an operon, then that operon will then make multiple proteins. So you might have slightly fewer RNAs than you have genes. And worms are an example of a multicellular organism that also has operons where genes will be strung together. They tend to be shorter and fewer in number, but then they-- but that doesn't just reduce the number of RNAs.

You can then increase them because you have alternative splicing where the RNAs will be made up of multiple pieces called exons which are stitched together by a specific biochemical machine-- splicing machinery. And that could happen in more than one way. They tend to be in a linear order in the genome, but there's exotic mechanisms like trans-splicing where you pull up an exon from completely different part of the genome and then splice it in together. So this number is larger than the number of genes because one gene can produce multiple RNAs by alternative splicing.

For proteins, there's additional diversification that you can modify the RNAs in various ways. In prokaryotes the number of RNA modifications is relatively limited, but the number of protein modifications starts to go up. You can have proteolytic modifications, phosphorylation of various amino acids. And-- and in multicellular organisms like worms and humans, the number of modifications reaches up into about 250 different amino acids-- modified amino acids of the basic 20 amino acids, which we'll talk about in just a moment.

The number of cell types in a very simple organism in a very simple environment might be as little as one. We don't really know how many cell types there are, but basically all the cell types for an organism like mycoplasma look fairly similar morphologically and probably functionally. On the other hand, the worm has 595, 500, 959 cells, and those three-- this is three significant figures. This is pretty good for biology. And these are non-gonadal cells, and the reason we know this so precisely is the entire lineage, the entire division of all the cells, have been mapped out for this worm. And we'll show this a few slides from now.

Humans on the other hand, not only is the number of-- the lineages is very poorly defined for most of the cell types in the human body, and even the number of cell types is unknown. Some people will estimate as few as 200 cell types. This is just a soundbite that is made up as far as I can tell. Some people say 200,000. It's probably a safe bet that a given-- at any given time point there are fewer than 10 to the 14th cell types. This is probably not very reassuring for those of you who would like to, say, measure expression in all the different cell types, because 10 to the 14th expression patterns would be quite a number.

In addition, you have various developmental stages where let's say you have-- as you grow from single cell to 10 to the 14th cells, you pass through stages, and what may be-- appear to be the same cell type at one stage at an earlier time point may have completely different gene expression. That is known, even though the total number of cell types is not.

AUDIENCE: [INAUDIBLE] the RNA, do you know how many [? extra ?] [INAUDIBLE] to [? include ?] RNA [INAUDIBLE]?

GEORGE CHURCH: OK, yes. I meant to caution you that-- just that the terminology here is used quite loosely. Gene expression is often used interchangeably with RNA expression, clearly. And then in almost the same paper, they will refer to as genes-- as protein encoding genes, completely sidestepping a large number of RNAs which really are never translated into proteins, such as ribosomal RNAs, tRNAs, small nuclear RNAs, and a whole variety of regulatory RNAs, RNAis and so forth. It's becoming very important, this class of RNAs which stay as RNAs, so be careful wher people use genes and RNAs interchangeably, or genes and proteins interchangeably.

So this is an example of molecular morphology, a particularly elegant example that illustrated last time. In a certain sense, the morphology of these two strands of DNA greatly-- go a long way towards explaining the inheritance and fidelity of the basic macromolecule which stores the information. What we're going to do is expand on the-- look outside of these bases which form the base pairs down the-- that has stacked up along the core of the DNA to look at how they're actually covalently attached, what the precursors are when we go from monomers to polymers.

We're going to talk about polymer synthesis for the next few slides. So in order to get this-- this exquisite base pairing here, which a recent article has argued is optimal. Of all the different base pairs that could have formed in the prebiotic times, these are some of the optimal alignment of hydrogen bonds. But the hydrogen bonds just guides the base pairing, the polymerization occurs in something not shown on this slide, but shown on the next slide.

Here are the two examples, two very similar bases for DNA and RNA, the monomers that are polymerized by enzymes-- polymerases-- to make DNA and RNA. So on the top is the deoxy-ATP and below is the ribo-ATP. Ribo-ATP is distinguished not only as a precursor for polymer, but it's one of the key biomolecules providing energy or transmitting energy from one part of the cell to another, one machine to another. If you look at a network diagram of the metabolism, ribo-ATP would probably be one of the most-- is one of the most highly connected nodes in that graph. It's connected hundreds of times in a graph where many things are connected once or twice.

And this is the structure, this is the base here in skeletal form and space filling form. The space filling form, aside from the colors, is-- is getting to be a more accurate representation of the electron density-- sort of the electron density you might observe in crystallography or in quantum calculations. But the skeletal form allows you to see some of the hidden atoms a little bit better. You can see the nitrogens are color coded for-- by blue, and the phosphates and sulfurs-- in this case, phosphates by yellow and the oxygens by red. Carbons gray or black. And so what you see is the only real difference between the deoxy-ATP and the ribo-ATP refers to this oxygen at the two prime position, which is the deoxy.

They-- they both share the-- this ribose and phosphate which were what was not represented on the previous slide are the repeating backbone. You go from this 3 prime hydroxyl-- the numbering here, by the way, for the ribose has primes after it to distinguish it from the numbering of the basis. These were studied chemically by chemists and numbered independently, and so when they were found in the same molecule, you had to have a separate.

So that's the reason, throughout the rest of this course and the rest of your life probably, you'll be referring to things going from 5 prime to 3 prime. It's because the people studying the bases won over the people studying the riboses. Anyway, so the last two phosphates just provide a higher energy bonds, and [? these ?] by equilibrium is pushed so that this whole splitting in polymerization is a very favorable in terms of free energy.

OK. Now as we discussed-- so those were the nucleic acid components, and then the proteins that they encode and the proteins that are required for the replication of the nucleic acid components are made up of simple derivatives of glycine, which can be represented-- full name, three letter code, one letter code. You should learn the 20 one letter codes because they're very valuable in this course and in bioinformatics in general.

Again, the same color coding here. You have nitrogen car-- this is the central carbon, it's the alpha, and this is a carboxyl group. So this is-- as an amino [? acid, this ?] is a [INAUDIBLE] ion with a positively charged nitrogen and a negatively charged carboxylate. And the way you represent this in a computer, you can either represent it as a pretty picture here, either skeletal or space filling. That's of course represented by zeros and ones, but it's not a very useful way for searching, merging, or checking, right? It's going to-- if you rotate this slightly in three dimensions, it's going to [? give ?] you a completely different image and it'll be hard to search.

You could represent it as the three-- the coordinates, the x, y, and z coordinates of each of these atoms. And that is something that you can search, but you also need to represent it in a-- in a way that represents the hierarchical structure by which these things form covalent bonds, and so that you can recognize groups-- groupings of atoms into polymers, polymers into assemblies, and so all the way up. And this is an example of such a hierarchical description which would be recognizable to all the computer scientists here if they were comfortable with some of the biochemical terms.

So here's the configuration of this thing we're calling glycine. By the way, there'll be a lot of jargon in this [? court ?] for those of you that are computer scientists. The point of this course is not to give you an encyclopedic knowledge, it's more to flesh out the concepts [? that ?] supplies both the computer and biologists in the group. And so if you learn a lot of facts, then we'll hold it against you, but-- but every time you see a piece of jargon and you think I haven't defined it, just call it Fred or George or something like that. It's an arbitrary name. It will be defined in the databases, and that's what the database is for-- keeping tract of this.

But you will have to understand the concepts, and the concept here we're trying to illustrate [? are ?] different ways of representing the molecular definitions, here by describing the-- the syntax as you would in breaking up a English sentence into its structure. So you have here-- it has a substituent of a backbone. Here in order to try to tie together all amino acids, you've made something that's a little nonsensical, which is you talk about the L backbone of amino acid. All the amino acids except for glycine actually have a handedness. That is to say, if you hold them up in a mirror it looks different from what the thing that you're holding your hand if you take, say, a space filling [? model and hold ?] [? it up to a ?] mirror.

And the reason is that you can have-- these two hydrogens here in glycine will have an actual side chain coming off in amino acids. And if it comes off here, then it's [? a D ?] amino acid. If it comes off of the other hydrogen, it's a L amino acid. So here you're saying natural amino acids are L amino acids, and so you want to take this L backbone and put a substituent on it. That substituent is HYD for hydrogen, and it's linked through carbon [? 1 ?] to another hydrogen and so forth. Nil means nothing.

And here's another way of representing it slightly more compact. You can think of this as just one long string even though it's on multiple lines. Here's one that's definitely [? at ?] the bottom line [? is ?] you've got this CH2 group, this methylene group, right in the middle bounded by this positively charged nitrogen and negatively charged carboxyl group. So you can see these nested, parenthetical ways of indicating the hierarchy. This allow you to search through complicated databases of compounds looking for shared properties, say, of all the drugs that bind to a receptor whether or not the structure of the receptor. If you know the structure of the drugs, you can do a structure activity relationship.

AUDIENCE: Question.

GEORGE CHURCH: Yeah?

AUDIENCE: What's the significance of the fact that you've got the amino group and the carboxyl group both at the end, and what is usually in the middle is brought up to the left?

GEORGE CHURCH: You're asking why [? in ?] a particular order? Well you can think of this [? likely ?] that your calculators. You can either enter them in the natural way that you do it, or you can do reverse [? Polish ?] notation. And different-- different ways of setting up syntax have this different thing. If you really wanted to research this, you'd look into the SMILES definition. This is a particular chemical definition, and they could justify this much better than I could.

So there are 20 amino acids in the simple genetic code. There are 280 that are post-synthetic modifications of these simple 20 amino acids. We've been talking about glycine as the basic backbone shown here in black on this slide, and in blue are these side chains that lended its-- its chiral nature, its nature that has a mirror image. And each of them provides the properties which are color coded here. Orange have the property that the side chains, and hence the amino acid and the protein, are hydrophobic. They try to get out of water, they are not-- they try to bury themselves in other hydrophobic moieties, like other amino acids like this in the core of proteins or in lipids which are hydrophobic as well.

Green are hydrophilic. Blue and red are also hydrophilic, but they're not only that. So they're also charged, the red ones being negatively charged as in the red of oxygen and blue being positively charged as in the blue of nitrogen. And the yellow being the sulfur containing, moderately hydrophilic amino acids. You can have more than one chiral center of symmetry, like this mirror image. You can have two such centers as in three [? inning. ?]

So now we're going to put these amino acids in the context of a-- the central dogma, going from DNA to RNA to protein. We want to illustrate this in a case of a very complicated machine and a very elegant and simple algorithm. This algorithm is simple because biology cooperates largely with us. The code, unlike many codes in biology, is fairly universal, found in almost all organisms in the same form. And it's very strict, and with relatively few exceptions where you have three nucleotides which encode one amino acid.

So there are 64 possible trinucleotides, 4 to the third power, and most of those trinucleotides encode some amino acid. The exceptions are stop codons indicated here, the little dashes in this table. And so let's just go through the table because that's the algorithm part of it. We have the color coded amino acids in here in a single [? little ?] code. Remember orange is hydrophobic, green is hydrophilic, blue is positive, and red is negative amino acids. And what we have [? as ?] an example would be AUG on the messenger RNA-- this is going from DNA to RNA-- is decoded by a complementary trinucleotide on this transfer RNA.

Here it's unfolded, but in reality and in-- in last lecture and in the next couple of slides, you will see it more folded up. But this is unfolding it to show the 76 nucleotides or so of the transfer RNA, which has been preloaded with an enzyme which is truly the miraculous part of the geniculate code, which is the aminoacyl tRNA synthetases which recognize the transfer RNA, recognize this methionine away from all the other 20 many acids, and put it on the right transfer RNA.

Once that's done, then the rest is base pairing. Mostly very Watson-Crick base pairing or something like it where the first two positions dominate and the second one can wobble. Here a G and a U is not part of the ATGC cannon of Watson and Crick, but it's close enough and it allows some ambiguity at this third position. So for example, UUU not only encodes phenylalanine, but UUC can also-- here the triplet in this table is UXU where the X is U, C, A, or G along the top of the table.

And so you can look this-- you basically look up the trinucleotide in a table like this in the computer and you can find the corresponding amino acid. This allows you to go basically from a DNA sequence to an RNA sequence to a protein in the computer. And what's-- so OK, that sounded too simple. Well, I'm going to give you a couple of slides that illustrate why it's more complicated.

First of all, why it's more complicated biochemically is not only do you have that amazing protein molecule, sometimes one or two subunit proteins are sufficient to take these tRNAs here encoded by red and green where one amino acid is going to be added to the growing peptide chain on the other, and the two business ends of the molecule that are responsible for handing off amino acids where they get coupled together in this polymerization reaction.

These two transfer RNAs have been properly charged by the amino [INAUDIBLE], but then they require this truly huge apparatus, one of the largest molecular machines in the cell, arguably similar or larger than any other one, which-- which allows the messenger RNA not shown to bind to these two trinucleotides in the tRNAs. And then a catalytic reaction occurs where the amino acid is shifted from one tRNA to the other, making a growing peptide chain.

And this chemical reaction here shown with all these circles and arrows for the chemical bond reformation is actually catalyzed by an RNA. This is the second RNA catalyst that we've talked about in this course. The first one was briefly mentioned on the subject of replication in last class, where you can find RNAs that can be engineered and probably existed in other scenarios which can replicate using small molecule precursors.

So by far, most catalysts that we will be dealing with will be proteins. But in order to get the proteins we need this really complicated ribosime-- RNA enzyme to catalyze this. Here the white are the base pairs, the Watson Crick and non-Watson Crick base pairs of the RNA, the gold is the backbone ribose and phosphates of all of those RNAs, and then the blue are the proteins, which you can see are mostly out of the periphery, not involved in either the enzymatic reaction or the recognition reaction which does the decoding at the trinucleotide codon level.

This is made up of total of three RNAs. Here you see two of the RNAs in the large subunit. There's a small subunit that fits on top of this which would mask the reaction we were interested in. Over 50 different proteins and the complete three dimensional structure is known from a variety of different organisms now. OK. Now after the break, we will then take this table that we had, and the complex biochemical machinery we [? had, ?] and turn it into a program that does the central dogma from DNA to RNA to protein. So take a brief break and come back and we'll talk about this.