Scanning Life's Matrix: Genes, Proteins, and Small Molecules (2002)
Lecture Four—Chemical Genomics: New Tools for Medicine
Stuart L. Schreiber, Ph.D.

1. Start of Lecture Four (00:15)

From the Howard Hughes Medical Institute, The 2002 Holiday Lectures on Science. This year's lectures-- "Scanning Life's Matrix: Genes, Proteins, and Small Molecules"-- will be given by Dr. Stuart Schreiber, Howard Hughes Medical Institute investigator at Harvard University; and Dr. Eric Lander, Director of the Whitehead Institute/MIT Center for Genome Research. The fourth lecture is titled: "Chemical Genomics: New Tools for Medicine." And now, to introduce our program, the grants program director of the Howard Hughes Medical Institute, Dr. Dennis Liu.

2. Introduction by HHMI Grants Program Director Dr. Dennis Liu (01:05)

Welcome back to the fourth and final lecture of our 2002 Holiday Lectures on Science. It takes a large team of people to put these lectures together every year, and it involves a lot of collaboration, and actually, even, some friendly competition. Science has increasingly become a team sport that of course features a lot of competition and collaboration. Our next speaker, Stuart Schreiber, is a great collaborator, but also a very keen competitor. For a view of the modern world of research and where industry and academics mix, I recommend this book by Barry Werth called "The Billion Dollar Molecule." It's actually a pretty exciting biotech drama, and Stuart Schreiber is one of the principal characters in this book, and everyone in the book wants to collaborate with Stuart, but nobody wants to compete with Stuart, and the good Stu/ bad Stu conversations make for some very amusing reading, actually. On the previous lecture, Eric Lander talked about using SNPs, Single Nucleotide Polymorphisms, to observe variation between genomes. In his next lecture, Stuart is going to introduce the concept of SMPs, Small Molecule Perturbagens, and describe how those can be used to perturb the genome, so that's why we say that Eric observes with SNPs and Stuart perturbs with SMPs. We have another short video to introduce Stuart, and then Tom Cech will return to close our program and introduce you to next year's speakers.

3. Introductory interview with Dr. Stuart Schreiber (02:55)

When I was a student, there was a very significant distinction between chemistry and biology. They were completely different subjects. That's no longer the case. To really understand biology means to understand the chemistry of living systems. The area of chemistry in which I focus is called organic chemistry. Today, it's especially important to have a grounding in a number of the basic sciences: computer science and information science is absolutely essential to pull together the insights from chemistry and biology. Engineering today is vitally important. Students today join my lab with training in physics, engineering, mathematics, statistics, chemistry, biology--you name it. What does it take to be a good scientist? Without any doubt, I think the number-one ingredient is your curiosity and passion for understanding things. If you have a strong desire to understand how things work, you're gonna be a great scientist. I think one of the myths about science is that you have to be brilliant to do it. Actually, I don't believe that. I think the thing that really makes a difference is that inner drive inside of you to just be so obsessed with understanding something, then you put yourself to the problem, you focus on it, and you can use whatever level of intelligence, creativity, determination you have, and combine them in some magical way that makes it possible to solve tough problems. Very few of my teachers in high school would have ever imagined that I would be the chair of the Chem--Department of Chemistry and Chemical Biology at Harvard University, but when I think back in retrospect, I had a lot of energy, and it was just channeled in a different direction back then. I discovered how exciting it could be to try to

solve problems, and then I just took all that energy and really became very focused on using my energy in that way. Are you an intrinsically curious person? If so, I think you'll do real well as a scientist.

4. Chemical genetics used to explore many biological questions (05:14)

Good morning, and welcome back to the fourth and final lecture of this series. In my previous lecture, I outlined for you some of the key tools of chemical genetics, especially diversity-oriented synthesis and small molecule screening that allow us to explore biology, the life sciences, and perhaps even medicine. These tools are especially useful, as I tried to indicate yesterday, in exploring the very dynamic processes in life. On this slide, I've outlined just a few of the kinds of problems now that are being interrogated by the use of chemical genetics, and these include molecular events, like gene regulation; cellular events, such as program cell death; organismal events, such as the development of a heart in zebrafish; and even those relevant to medicine, such as cancer.

5. Using genomics to investigate glucose sensing and type II diabetes (06:21)

In my lecture today, I'm going to focus on this problem: the problem of glucose and other nutrient sensing by cells, tissues, and organisms like yourself, humans, and I'm going to talk about how when this process goes awry, the outcome can be type II diabetes. Now, we heard yesterday and today about genomics-- the importance of genomics. Genomics has taught us many things. Genomics teaches us to think broadly about a problem-- to think, in fact, system-wide about problems; to take a global approach. We're now able, for the first time, to look at all of the DNA, all of the genes that we heard about, all of the messenger RNAs, and even variations in the DNA sequences and variations in the levels of the messenger RNA. I'm going to use the glucose response in type II diabetes to illustrate that the underlying principles I discussed yesterday in genetics, but also of genomics, are relevant to the way in which we use small molecules to explore biology. And I'm going to, again, do so in the process of looking at glucose response-- either one gene or one protein at a time, but also, in this system-wide approach, by looking at all of the relevant genes or proteins at once.

6. Concept of networks of proteins (08:02)

Now, the fact that proteins tend to work in networks as opposed to in isolation is probably most clearly illustrated in life's regulatory processes. These are the mechanisms that life uses to provide the appropriate output given a certain level of input. I'm going to show you how these interacting proteins function as networks, as I said-- network's a term that-- I don't know how familiar you are with the biologists. The biology community is only now beginning to think more and more about concepts of networks from other areas of science, a continuing theme in these lectures. We learn about networks that, for example, when proteins talk to each other, they do so with loops-- feedback loops. They can be positive, and they can be negative, and that when networks are established, redundant pathways are often used. These loops and these redundant pathways yield robust networks, networks that are able to adapt to changes in environmental cues and, very importantly, are able to respond with switch-like, or all-or-nothing behavior, to very tiny changes in environmental cues. So, for example, when your body's detecting 5 millimolar glucose in its bloodstream and suddenly there's a change to 2.5 millimolar glucose, well, that's only a twofold change, but the effect on the physiology, for example, of the brain, is tremendous, and so we need not to just have twofold change in brain physiology, we need a switch-like, all-or-nothing major change in response to that small change in concentration.

7. Two networks are involved in the glucose-sensing system (10:06)

I'm going to be talking about two interacting networks: we call them signaling networks. The first involves the blood that reaches different cells and organs in the body and brings with it the nutrients,

especially glucose. Now, we've learned a little bit about how cells respond to varying levels of glucose, and we've learned about a number of interacting proteins that comprise this nutrient-response signaling network. I've given you 4 such proteins that we have a pretty good handle on right now. Nutrients talk to protein one and then protein two and then protein 3 and then, finally, protein 4. Actually, we know the names of these proteins. I'm not going to give them to you, because they're not necessary for this particular discussion, with the exception of one: I've labeled protein two as the FRAP protein because, some of you may recall, I introduced the FRAP protein as the protein that FKBP12, when bound to rapamycin, is bound to. It is the target of FKBP-rapamycin. Now, there's a second signaling network here-- we call it the insulin-response signaling network. Here, insulin is a signaling protein outside of the cell. It binds to the sensing component of this network, the insulin receptor. It passes its signal through these intermediary proteins, the mediators of the signal-- proteins 5, 6, 7, 8, and then to protein 4.

8. Protein-to-protein interactions in the two networks (11:52)

Now, there's a number of puzzling things about this network-- these two networks. For example, the arrows indicate an activating event. When protein 5 talks to protein 6, it activates it. These T-bars represent an inhibitory event. When protein one talks to protein two, it inhibits. Thus far, what we've seen is that all of the proteins on the first network inhibit each other-- a cascade of inhibitory events. We've seen in the second signaling network that all of the proteins activate each other. Now, we don't know whether there's something important here, whether this is teaching us something or not, and the reason we don't is that these are certainly not all of the components of this network. This is a common case in current studies of signaling networks. We've got pieces of the puzzle, but we don't have the whole picture together, and maybe when we find other components, this simple rule that seems to exist will be broken.

9. Network-to-network interactions at the node protein (12:59)

Now, there's another really important issue in this little picture here that I'd like to focus on. Protein 4-- I've called it a node. One of the things we've learned is that networks talk to networks, and they do so through common proteins--proteins that are part of each of the two networks-- so I'm calling that kind of protein a node. And again, we know the identity of this nodal protein. So, this signaling network, beginning with glucose, sends a set of inhibitory signals. Ultimately, if protein 3 were active, inhibiting protein 4, and this network sends us a set of activating signals, ultimately through protein 8, which, if protein 8 were active, would then activate protein 4. So there's cross talk between signaling networks, and they--this cross talk-- occurs through these nodal proteins.

10. Problems in the glucose-sensing system can cause type II diabetes (14:01)

Now, why should we be interested in these two particular signaling networks? Well, one very interesting element of these networks is that if something goes awry in either of the two or, in fact, in both of the two, type II diabetes results. So the correlation we've seen about this still very mysterious disease, a disease for which we do not know the molecular origins... we know that humans that have acquired type II diabetes have problems, mistakes, in both of these signaling networks.

11. Rapamycin can induce type II diabetes (14:41)

Second reason that we've become interested in this is that we have a useful new tool to begin to pick apart the mistakes that occur in these networks and to learn more about the signaling networks, and that's this small molecule that I introduced to you yesterday named rapamycin. Rapamycin, when bound to FKBP12... binds to FRAP. Remember, the composite surface of FKBP-rapamycin is complementary

to the FRAP binding surface. That was an example of a small molecule activating the FRAP protein-- I mean, the FKBP protein-- to be able to bind to FRAP. Now, when that binding occurs, FRAP's activity is inhibited. So if you inhibit the FRAP protein, what that does is it releases the brakes on the protein to which it regulates, protein 3--releases the brakes. It allows protein 3 to do what it normally does, and what it normally does is to inhibit protein 4. Now, with this little diagram, we can begin to understand why rapamycin causes cells, tissues, organs, and even humans, to be tricked into thinking that-- that there are very low levels of glucose and very low levels of insulin, even when the cell, for example, is swimming in glucose and swimming in insulin, because insulin would normally activate protein 4. FRAP unleashes this inhibitory protein, which inhibits protein 4, sending this confusing message that there must be low levels of insulin, there must be low levels of glucose. Now, what's even more important, as I mentioned to you yesterday, despite the fact that rapamycin is a very powerful, very useful, clinically used drug that saves lives in patients having organ transplants, it has some side effects. One of the side effects is that it's been found, quite unexpectedly recently, to induce type II diabetes in patients that are taking it. Now, that may sound very worrisome, and of course it is a concern, but as I mentioned yesterday, we can dose the rapamycin, and these small molecule effects are reversible, so patients can be taken off of rapamycin and put onto a different immunosuppressant temporarily, but this observation has been shining a powerful new light on this mysterious process, and that's what I'd like to discuss with you.

12. Finding new small-molecule probes with a small-molecule microarray (17:15)

Now, we have 8 different proteins in this network that I've mentioned to you already. We have one small molecule probe of one of those proteins. First thing I'd like to illustrate is a technique that's now being used to generate probes to those other proteins, those other signaling proteins. We saw, in the previous lecture, a beautiful illustration of the miniaturization of arrays, and you saw your DNA chips-- tiny glass microscope slides that have little spots of DNA. Well, scientists today can make spots of small molecules as well as DNA. So the very similar high degree of miniaturization increases the throughput in this analysis. These are called small molecule microarrays. If you remember, I gave you an example yesterday of the synthesis of 88,400 small molecules. In fact, all 88,000 can be arrayed onto about 6 microscope slides. So what I'm gonna show you is an example of a machine that's used to synthesize-- to array--12,000 small molecules on 100 different microscope chips, so giving you vast numbers of small molecules. And I'm gonna show you how we use these arrays to identify new probes of those other signaling proteins. So, here, what we're going to do is take the protein of interest, proteins one through 8, and wash them over these microscope slides. We're gonna do so in a way that the protein is labeled with a fluorophore. So if one of these small molecules is a match, a geometric complementary match, and can recruit the protein then to its spot, we'll see it in the form of light. Can we roll the first movie, please?

13. Animation: Small-molecule microarray (19:12)

So, this is a microarraying robot. You know, it's actually not unlike the robot that's used to print the "New York Times." What we have are 48 quill pens that you see here, and they're spotting one nanoliter solutions of these small molecules. That was making 100 microscope slides. I've honed in on just 8 of them now. They need 8 to run this experiment. This is an example, then, of the microarray with 12,000 different spots per microarray. We wash the protein over these arrays with the label, and if we get lucky and one of those is the complementary shape, then we have--voila-- a new small molecule probe of one of those proteins. So this is one of the ways, conceptually, at least, in which we're finding small molecules to modulate functions, perturbing functions, so that we can then observe the consequences.

14. Limitations of the small-molecule microarray (20:11)

Now, I told you that we didn't have all of the components of the network put together yet, but I've got this new technique that's able to give me new probes of the ones we do have, so are we gonna be able to use this technique alone to completely dissect this network? Does someone out there see a flaw in the technique thus far? Yes? Sometimes the pathways behave differently in cells than out of them. Ah, very good. That's exactly right. This is an in vitro experiment. We're just using small molecules with purified proteins in this experiment. Proteins really function inside of cells, and we've got to make sure that our compounds themselves can function and modulate the function of these target proteins inside of cells. Actually, there's another reason, as well. There may be other components that we haven't figured out that the small molecule can modulate that we could detect if we turned our screening system into a cell-based one, and that's what I'm gonna show you here.

15. Overview of the cell-based screening method (21:21)

When you screen for proteins in their natural cellular setting, you have a new challenge: a lot of technical challenges. First of all, you can't wash the proteins in their cellular context over these microarrays, because proteins don't wash very well. They like to stick to glass. So, first way we solve that problem is we reverse the order. We actually put the proteins in the cells down, and then to those we will add individual small molecules. Now, if you think about the nature of small molecules and cells, it turns out that microscope slides are not as useful anymore, but this simple device, this simple so-called 384 well plate, which has deep wells within it, are subject now-- they're in a standardized form-- and we have robots that can take a stack of them and lay them out in an appropriate order. We have robots that can deliver cells into each individual well very rapidly. We have robots that can deliver stock solutions of small molecules and other reagents that might be useful in this experiment. So this has become a workhorse in this kind of experiment, and I'm gonna show you, in a second, how we do this in the context of this network problem that I've outlined.

16. Using rapamycin-treated “diabetic” cells to find new small-molecule probes (22:45)

Next problem is how do we focus in on the network? So what we're gonna do here is focus in on the diabetic network, I'll call it, the composite of those two, by--well, we can't find diabetic cells, because diabetes is a disease of a whole organism, but we can do something that's a pretty good approximation. I just told you rapamycin affects this pathway. I actually forgot to mention that not only does rapamycin affect this pathway, but it actually induces this disease, so let's take the rapamycin molecule and treat cells with it and then consider that a diabetic-like cell, and now we're gonna screen for small molecules that can reverse this effect, that can convert the diabetic cell into the healthy cell. Can we roll this animation, please?

17. Animation: Cell-based screening finds small-molecule inhibitor of rapamycin (SMIR) (23:39)

So here's our 384 well plate. We're gonna zoom in on it now. We've already robotically delivered cells, you can see, into the individual wells, and another robot is delivering pure samples of different small molecules. Here you see rapamycin-treated cells. Now, we have techniques from optical physics today, for example, that can detect optically whether a cell is in a diabetic-like state, whether it's been treated with rapamycin or not, and likewise, we have techniques that distinguish those from the healthy cells. So when we ran this experiment, we searched for many, many, many small molecules. Remember, I told you we synthesized thousands of them-- 88,000 in one particular case-- and it was only one rare one-- this one we call "SMIR" for "Small Molecule Inhibitor of Rapamycin"-- that was able to reverse that diabetic-like state of rapamycin and convert those cells to healthy cells. Now, we don't know how this happens. We don't understand yet how it happens, but SMIR is now a new probe. We don't know the protein to which SMIR binds, but whatever it's binding to must be part of this network and must be a really interesting part if it's able to flip the switch back to the healthy state. How we find the target of

SMIR and furrowstatin that I mentioned yesterday will be the subject of the next part of my lecture, but at this point I'd like to stop and take some questions.

18. Q&A: How do you know SMIR isn't directly affecting rapamycin? (25:21)

A question in the back. Right there. Do you know if the SMIR is affecting directly rapamycin or if it's directing one of the proteins in the pathway, and how do you--I guess you're getting to that... So, that's a great question, because one thing we could be doing, which would be a little less interesting-- the molecules directly affecting the rapamycin molecules. So, when you don't know the target, one of the things you can do is start making guesses, and that's a really good first guess, and that's the one guess that we did make because we wanted to make sure that it wasn't something as simple as that. So we did about every experiment we could imagine to show that, in fact, it was not affecting directly rapamycin nor the FKBP protein or FRAP proteins to which rapamycin binds. So it's somewhere else out there, and actually, when I follow up in my final segment here, I'll try to show you how we're using computers to get at this question. Now, yesterday we had so many good questions that I actually ran out of T-shirts, and we have some Hughes Institute's-- these aren't Harvard-- Hughes Institute T-shirts, not some Harvard T-shirts, so I'm going to bide you with one of those. I'm gonna double up my gifts, and these are my little reminders for those of you asking good questions out here. This is like the holiday lectures' moon rock sample for you. We're gonna disassemble this little small molecule. This is your little token for remembering the holiday lectures.

19. Q&A: Does SMIR counteract the immunosuppressive effects of rapamycin? (27:00)

OK. Let's see. Let's go all the way over-- let's work over on this side. How about in the very corner back there? Has the SMIR molecule been found to counteract the immunosuppressing effect of rapamycin, as well? Very good question again. We haven't checked it, but my feeling is that the immunosuppressing effect of rapamycin is intimately linked with its diabetes-inducing one, and the reason I think that is that this signaling network happens to be extremely active in immune cells, lymphocytes, and that very tiny perturbations in it render the immune cells inactive. And I think that's why rapamycin is clinically so useful as an immunosuppressive agent. It's distributing throughout the body, but it's hitting cells and this pathway that are hypersensitive to it, so that's the first thing you see. But we need to do the experiments and confirm that hypothesis. Here's a T-shirt for you. Oops... And here's a piece of furrowstatin for you. OK. Great. Let's see.

20. Q&A: How can you find a common adhesive for all small molecules? (28:16)

How about now we'll take one in the middle here? How about this gentleman right here? Your first method of screening for the small micromolecules, you use a microscope slide. If you synthesize these small molecules that have different binding patterns, how are you finding a common adhesive to immobilize the small molecules onto these microscope slides? Boy, this is an amazingly good question. So, the question is, you've got so many different compounds, how do we get them to stick to this microscope slide? That's a really important problem, and so this is a good example of when you develop new technologies like diversity-oriented synthesis and small-molecule microarrays. You actually need to develop them in parallel, as partners. You need to think about this hand-off mechanism. So what we do is we have 3 different handles. We make sure that every one of our 88,400 compounds have at least one of these handles. And then we develop 3 different complementary surface chemistries. There's a whole area of chemistry called surface science, where you learn how to learn how to do chemistry right on the surface, in this case, of a microscope slide. So we have 3 different kinds of microscope slides and 3 different kinds of handles. And thus far, we're able to keep track of each of the compounds that we make. Ok. How about a question

21. Q&A: Could you circumvent genetic defects by using small molecules? (29:42)

over on this side? I understand that a lot of genetic diseases are caused by one inactive protein that doesn't work, that affects an entire pathway. Could you then, by understanding the product of each protein, almost circumvent the pathway? By giving, like, a cell, like, the product? By artificially creating the product and giving the cell the product and thus circumventing the pathway? Oh, OK, very good. In fact, that's exactly what one, in fact, does in many instances. But the pathway has an input, and it has these intermediary events, molecules, and it has an output. So when we look for a perturbation now, let's say, of the SMIR protein, we might add to the cell the input from each of the various steps and see at which point in time we can lose the effect of the SMIR. So that's actually a genetic principle. It's called epistasis analysis, or ordering of events along a cell. And it's a very powerful one in genetics, and therefore, it turns out to be a very powerful one in chemical genetics. So, thank you very much. That was another great question, and let's see--where's my furrowstatin molecule? I give you a T-shirt, I give you a piece of furrowstatin. And I think I owe another one over here. You get an unusually large chunk of furrowstatin. OK. Thank you.

22. Importance of information science and GenBank (31:26)

So I'm going to turn now to the final section of the 2002 Holiday Lectures on Science. I'm going to tell you about something I've been alluding to throughout these lectures and something, certainly, Eric has mentioned a number of times. That's the importance of information science and computer science on the life sciences today. I'm going to, in particular, tell you about a project that is information-science-based that we call the Chembank Project. But first, let me remind you of an element of Eric's previous lecture shown on the screen. Eric talked about a matrix of genes and phenotypes and the way in which genes and phenotypes can be linked by studying variation. And the variation in the DNA sequences, for example, are called SNPs, which you've heard about. So, linking gene 2 to diabetes through a variation is a very powerful way of studying disease processes. This information and related information concerning genes and genomes is available on the World Wide Web through this database called Genbank. This is sort of the incredibly powerful tool that's sort of the ultimate creation of the ultimate level playing field within science now because it means anywhere in the world, everyone has all the same access to all the same information as long as you can gain access to a modem and a PC.

23. Overview of the ChemBank project (33:14)

So, as I work towards the end of my lecture, I'm going to increasingly be talking about things that are in the future. The first project I'm going to tell you about is one that's called Chembank, and it's sponsored by the National Cancer Institute. We've been working on this Chembank project now for a couple of years, and we're hoping that by the summer of 2003, we'll have our first launch on the World Wide Web of Chembank. What is Chembank? Chembank is conceptually like Genbank. In fact, we select the name as an homage to the venerable Genbank. There's a big difference, though. Chembank, like Genbank, is a suite of informatics tools. You can access them, and you can analyze data with these tools. They can access a set of databases that themselves are linked, so they can talk to each other. But the matrix of Chembank has, instead of genes and phenotypes, proteins and phenotypes. So we're going to look at all the different proteins that one can imagine and all the different phenotypes that one can imagine, including some of those phenotypes that were in Genbank. Analyzed with SNPs. What is the glue that connects the two axes? Well, they are, as Dennis said, not SNPs, but SMPs: Small-Molecule Perturbagens.

24. How ChemBank works: Furrowstatin as an example (34:40)

How does this work? Well, let's start with one example that we discussed yesterday. Remember I told you that Tim Mitchison at Harvard searched for small molecules that perturbed the cleavage furrow process in cell division. That's a topic of great interest to him. And he found this molecule furrowstatin. Well, it turns out, if you go to the proto-Chembank, the precursor of it, not quite ready for prime time, it's already sufficiently functioning. But if you type in "furrowstatin" and you ask, "Did any other researcher ever find anything interesting about this same small molecule?" one thing you might find is that it may have induced other phenotypes-- for example, in an organism. Then you might link cleavage furrow to that process in an organism. But even more illuminating is, as Tim found, you find-- actually, furrowstatin was found to bind to a particular protein, like in one of those small- molecule microarray experiments. It binds to a particular protein, and the protein he found that furrowstatin was already known to bind to is non-muscle myosin II. So just by going to the computer, after seeing, observing, this effect, going to the computer, you got a link-- a link between myosin II and the furrow formation. Now, once again, like I said yesterday, this is not guilt by association-- it does not prove that myosin II is involved in the cleavage furrow-- but it's a very interesting and reasonable hypothesis. So we look at Chembank as a kind of hypothesis-generating tool. And of course, what we want to do is find other small molecules that link other proteins to other phenotypes.

25. How ChemBank may help explain SMIR's action (36:30)

So let's do this with SMIR. Let's go to the computer. I just told you that SMIR-- this interesting small molecule that came about by the phenotypic observational screens screening for small molecules that reversed the diabetic-like state of a cell converted to the healthy state-- has a very interesting effect on cells, but we don't know the protein to which it binds, and therefore we can't suggest or hypothesize what protein component might be a part of the signaling networks. Well, in this case, unlike furrowstatin, we find that the database is not sufficiently populated. We don't know the answer yet. We need more data. We need more information to go into Chembank-- not unlike the time line that Eric discussed, where the SNPs hopefully will be completed in two or 3 years down the road. But incomplete at this point means you don't always get the answer.

26. Like genomics, ChemBank shows the importance of global measurements (37:31)

I think the main point that I'm trying to make here is that taking clues again from the genomic principles that we heard about earlier, we get this sense of the importance of global measurements, of doing everything to fill out the matrix, the various matrices that you've seen in the various presentations. Or another way to put it is that genomics teaches us that the outcome of all of the experiments is sometimes more important than the outcome of any individual one because if you get the outcome of all the experiments, you use a computer and a database and search tools to interrogate it, you might make connections between seemingly disparate experiments that suddenly generate your new hypothesis. So information science is important today to link chemistry, biology, and medicine. And what I've given you is an illustration of Chembank in its early form. Informing bi-- Chembank, through chemistry, is informing biology. Those furrowstatin molecules and SMIR molecules coming from chemistry are making links between biology and teaching us about networks and pathways and regulation.

27. Concept of chemical space: Measuring small-molecule diversity (38:52)

I'm going to finish this presentation with an illustration of how Chembank, multi-faceted as we hope it to be, can allow biology to now inform chemistry. So here's the problem that we are trying to solve through the use of modern information science and computers. Yesterday, I told you about basic outline of diversity-oriented synthesis. And I suggested that diversity-oriented synthesis--DOS-- aims to make highly complex and diverse, structurally diverse molecules-- molecules having many different shapes-- ultimately, all the possible shapes necessary to interact with all of the proteins of life. If you think about

what I told you, the driving force behind diversity-oriented synthesis is really a lot of chemists' knowledge of reactions and chemists' intuition. We didn't actually think about which compound should we be making. Of course, we don't really know which compounds we should be making, but we'd at least like to know which set of compounds are maximally different from each other so we might populate chemical space. This is a concept I'm trying to indicate to you in this cube. Populate chemical space in some dense and random way-- complete way. We want to complete the matrix of chemical space. Now, what do I mean by this abstract term "chemical space"? Well, just imagine if these 3 axes represented 3 computable qualities, characteristics of small molecules. For example, we can certainly calculate the molecular weight of every compound we make. And we don't want to synthesize 88,000 compounds that all have a molecular weight of 522, because what if the magical number is 523? We'd like to get a range from very low molecular weight to very high molecular weight. Consider another axis here is, let's say, solubility. Some compounds are very water-soluble, some compounds are very lipid-soluble. If we make compounds that are only one or the other, we might miss a whole lot of interesting opportunities to perturb biology. And likewise, this third axis could be the polarity of a small molecule, which you can compute. It's computed dipole. We don't want to make molecules that all have exactly the same dipole.

28. How diverse were the molecules synthesized by DOS in Lecture Two? (41:26)

However, if you think about what I told you about diversity-oriented synthesis, we didn't take any of these considerations into mind when we synthesized those 88,000 compounds. So let's retrospectively go back and see how did we do. What I'm plotting for you are 3 computable properties of small molecules. And in blue, in fact, are the 88,400 compounds that I told you about that Ohyun Kwon made. Doesn't look too good, does it? They are pretty clumpy. They're not distributing themselves throughout chemical space the way we'd like to. And I've got, in yellow, another DOS pathway; in white, another diversity-oriented synthesis pathway; in red; and in each instance, they're pretty clumpy. They are different clumps, and that's good, but this isn't the way to go about this, I think, in the future.

29. Computation could guide synthesis is evenly distribute molecules in chemical space (42:13)

So one element of Chembank that we're trying to incorporate into the program is application of mathematical algorithms that can tell us in advance which combinations of which building blocks and which combinations of which reactions-- should we select the minimal set of building blocks and reactions? Should we select in advance-- that would then yield compounds that distributed themselves more evenly in chemical space. So, what you have here is an illustration of virtual compounds. These actually have not yet been synthesized, but the program is teaching us how to distribute the compounds in this three-dimensional chemical space, in a Gaussian distribution. OK We can ensure the lack of clumpy-ness, at least, in this process. OK? So, one thing we hope to do with Chembank in the future is be able to allow chemists to go to their computer and, rather than just running off in the lab and using that blue cow that I showed you yesterday to run all these reactions and parallel, run them in a way where the computer has told us in advance, what's the magical combination to distribute these small molecules in chemical space?

30. Could distinct regions of chemical space affect specific biological functions? (43:36)

Now, is that what we really want to do? I think, for the time being, that's a pretty good goal. We want to be able to occupy all of chemical space with these small molecules. If we don't occupy all of chemical space, how are we ever going to find out what the really important parts of chemical space are as it relates to biological space? OK? So we want to fill out chemical space so that we can test which areas of chemical space, which region of this three-dimensional plot might be most relevant to different swaths of biology space. What do I mean by that? I mean, is it going to be the case that these compounds that

affect diabetes, for example-- the subset, the tiny subset of these small molecules that affect diabetes-- will themselves be distributed in a Gaussian distribution, or is it possible that there are certain regions of chemical space that are well suited for interacting with that network of proteins in diabetes space?

31. Concept of biology space: The blood-brain barrier as an example (44:48)

I have a strong sense that there are going to be subsets of chemical space that will serve as the sweet spot for certain swaths of biology space. And I can give you one really simple example. If you want to study memory and cognition, the function of the brain... we've already learned something about the brain in higher organisms, and that is that it has a filter that's called the blood-brain barrier. And we've learned a little bit about this filter. It's a filter that's probably there to make sure that chemicals in the environment that might have a deleterious effect on brain function don't get to the brain. We've learned about the physical chemistry properties of this filter with regard to small molecules. So, it turns out that the net charge of small molecules is critical in order to pass through the blood-brain barrier. The net charge should be zero. It doesn't matter if you have two positive charges, as long as you balance them with two negative charges. Then the molecules get into the brain. So if one of our axes is charged, net charged, everything that's strongly positive or strongly negative is going to be off-limits for the swath of biology space called memory and cognition. So there's one concrete example. I don't think it's going to be easy to make this connection, but

32. Animation: Future research may link chemical space and biology space (46:14)

I want to finish with a little glimpse to the future. I admit this is very much in the future, but this is the challenge for all of you and many life scientists, or those of you who might move into the life sciences for the future. So if we can roll the next and final video... Now, I've plotted for you-- this is just hypothetical. This is where we'd like to go in the future. This is the subset of various small molecules, for example, within chemical space-- the SMIR and the rapamycin. I am suggesting that diabetes space may be like memory and cognition space: not uniformly distributed throughout all of chemical space. We already have hints of this, not only in the memory and cognition space, but in terms of other aspects of biology-- furrowstatin, for example, interacting with cell-division space. So, the molecules that are being discovered that affect certain swaths of biology are clustering together. They're not completely random from each other. And one of the very exciting goals in the future will be to link, to understand, the relationship of chemical space to biology space so that scientists will be able to, in advance, synthesize molecules that take them right to the sweet spot of biology they seek to explore. So with that, I think I'll take, hopefully, some more questions from the audience.

33. Q&A: How long would it take to describe biology space? (47:54)

And let's see. I'm going to try and distribute here. How about one over there? How long do you think that will take, hypothetically, to figure out these sweet spots and stuff? I think that, like everything in science, it's not going to be an on/off switch. It's going to be a slow progression, and just to calibrate you, where we are right now, what we want to do first is get this tool out there. So we're aiming for this public launch of Chembank in the summer of 2003, but what that really means is that Chembank has to be interactive. See, Chembank's already working right now, but we need to make it interactive so that everyone--anyone out there in the world who has something to contribute to it-- can contribute it and increase its value. So that's a first big step, summer of 2003. Then I think it's, like, maybe not unlike the SNP measurements. It's going to be a period of time, of populating the matrix to get an increasingly dense matrix. So I think the next 3 or 4 years are going to be a rich period of populating that matrix. Now, what happens after that? How effective will Genbank and its variations of SNPs and expression levels and Chembank and its relationship of chemical space and biology space-- how effective will that be? How much time after that will we be able to put all this into place to actually create the new

medicines to be able to treat disabling disease? That's a difficult one. I think there's an awful lot of work out there, and one of the reasons Eric and I are here today is that we're really hoping that some of you in this audience will find some of this to be exciting because we need you to ensure the progress of science. So that's a very stimulating question. There's a Howard Hughes Medical Institute T-shirt and a little piece of furrowstatin for you.

34. Q&A: How are scientists sorting out different proteins? (49:58)

How about-- let's see--in the blue sweater? How are scientists able to sort out all the different proteins to be used in your Chembank? Well, how do we sort out the different proteins? Boy, this human genome and mouse genome sequence is very, very helpful. I'll tell you, it's changed the way we think about the science because, just 10 years ago, there was this feeling of vast, unexplored, unending protein space out there. It's no longer that way. It's bounded. We know what all the proteins should be because at least we think we know what all the genes are, and the genes give rise to the proteins. Now, there's a lot of variations of the proteins that are very important, and we have to work all that out. But so, first part of my answer to your question is that the human genome, genome sequencing, gives us a kind of finite quality to the number of proteins out there. Now, what good does that do us? Well, there's a whole new area called proteomics. It's like genomics, but now all of the proteins. And one of the things-- it's like these new fields are going to require development of new technology platforms, new techniques, new higher throughput ways to gain access to the mechanism of purifying these proteins, or expressing these proteins, or getting them to turn on in a cell or turn off in a cell. And that again is-- many labs now, many consortium of labs working around the world, trying to develop those new tools. It's a great time to be very inventive in sort of developing new technology because we can see conceptually how to do it, we just don't know mechanically the best way to do it yet. OK. So let's... give you a T-shirt here... and dwindling supply of furrowstatin here. OK. Let’s

35. Q&A: how widespread is ChemBank right now? (52:01)

see. I want to be even here. Come back over to the question up top there. I was wondering how widespread is Chembank right now in terms of how many people can access it? And how international do you expect it to be, do you hope-- Great questions. Chembank right now is being used-- I should say proto-Chembank because, again, not quite ready. It's not user-friendly enough. Chembank right now is being used by a group of labs, about 140 of which are sharing technology platforms, some of which I've been describing for you. And we've gotten together and decided and agreed upon common ways of expressing our experimental outcomes. That's a real key when you want to pool together resources. So we're thinking that we're nucleating a group. It's probably of those, 80% are in the United States right now. The remaining are outside the United States. We are really trying to promote more and more activity outside of the United States. So, first step has been a sort of user group to road-test this, a beta version, and work out all of the glitches. The second part is that launching on the World Wide Web. Once you do that, it's available everywhere, absolutely no restrictions. As I said, you need a PC and a modem, and you gain access to it. So it's really just those two steps-- working out the bugs right now and-- although this first launching of it, I think, will still be in a somewhat primitive stage. It's got a lot of new components of Chembank that we need to keep building, so this requires that we bring more and more computer scientists and mathematicians into our world of science. I

36. Q&A: Does rapamycin affect everyone’s cells in the same manner? (53:57)

may have time for one more question. Let's see. Question... let's see. The very back. Yeah. I was wondering if rapamycin affects everybody's cells in the same manner, and if there are other small molecules besides it, if it doesn't affect everyone in the same manner, that could affect... You could not have asked a better closing question. That's a great question because you know what it does? Just like

the previous last question in there, you've just bridged our two lectures. I'm talking about small molecule perturbations in cells in humans as if they're all the same, or at least all humans are the same. Eric told us that we're pretty similar, but there's these one-in-a-thousand differences, and those one-in-a-thousand differences do make a difference in drug response. We don't know yet how those differences relate to the case of rapamycin itself, but we can draw on analogy to other drugs where there are clear differences. This is another frontier. It's, in fact-- I'll give you another "omics." it's called pharmacogenomics-- the way in which different drugs will respond to different people based on subtle differences in their genomes. Another frontier, once again. I'm going to have to wrap up with that question. It was a great question to finish. I want to thank you all. You have been a terrific audience. No, you have been an amazing audience. For two days here, the level of discussion has just been astonishing. It's been very exciting, and, again, we are very hopeful that some of you will find something about this world of life science that would excite you in your future because we really do depend on you for progress in science in the future. Thank you.

37. Closing remarks by HHMI President Dr. Thomas Cech (55:58)

I want to thank everyone who contributed to these Holiday Lectures on Science, both the team here at Howard Hughes Medical Institute and especially our two speakers, Stuart Schreiber and Eric Lander, for some really stimulating lectures. Before we part, let me give a little bit of a foreshadowing of what we have in mind for you next year. The 2003 holiday lecturers have already been chosen, who will be giving their lectures next December. These are two HHMI investigators-- Huda Zoghbi from the Baylor College of Medicine and Bert Vogelstein from Johns Hopkins University. Dr. Vogelstein is an expert in the genetics of cancer, and Dr. Zoghbi and her laboratory have made breakthroughs understanding the genetic basis of various neurological disorders. The common theme between these two lecturers is going to be: "What I Learned From My Patients." In other words, how do you get the hypotheses for the research you do in the laboratory? In these cases, these medical researchers get the hypotheses from observing the anomalies that are brought to them in the clinic, in their medical research. So, at this point, from all of us here at HHMI, please have a happy holiday season and a healthy new year.