NARRATOR: 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.

DR. GEORGE CHURCH: As promised, we're going into flux balance analysis. The simplifications here are even greater than the ones that we have been making so far. But hopefully, you will see some benefit. We're here. Assuming the time constant metabolic reactions are very fast compared to cell growth.

Cell growth, we've seen the time constants are in the order of seconds or sometimes less, while growth might be on the order of hours. Even though we've seen interesting dynamics that can occur, those phase diagrams that we had in the ordinary differentiable equations, there are also many-- it's very typical for a cell to be in fairly stable conditions. There'll be a transition time, and then they'll be stable again.

And that will be the steady-state assumption. So we'll say there's no net accumulation of metabolites. So even though there may be fluxes into and out of every metabolite, the net change is zero. And so that means this equation we've seen a few times, the change in X will reflect the time is zero. And that means that stoichiometric matrix times the flux rate for that for all relevant reactions, you can think of as a matrix representation.

Sum of all the inputs and outputs of that particular X position, minus transport vector, which is kept separately, is all zero. So what does that mean? That means the stoichiometric matrix times the internal fluxes is equal to transport across the membrane. And this is something-- I'll show you-- we're going to do this in two steps. First, we're going to solve for that as if there was an exact set of equations, just the right number of equations, just the right number of unknowns.

So you can actually get a solve for the unknown fluxes. And then we'll work through it where it's actually underdetermined, but there are inequality constraints. First, let's take the exact case because the stoichiometry is known that's S. These are the zeros and ones, minus ones, twos, and so on. The uptake rates can be known in the sense that you can regulate the amount of input and output by the rate at which you remove substrates and add them.

And the metabolic flux's are what you actually want. This is in contrast to the red blood cell or other ordinary difference equation cases we've been talking about. Where the rate equations were known, but what we were looking for concentrations. Here we're less concerned about concentrations, and we want to learn what it might be the fluxes and later what might be the optimal fluxes.

So Slide 42, let's say that we-- as, in a way, nomenclature, we focus on positive fluxes. If we want to go the opposite direction, we'll make that another reverse flux, separate flux. So focusing on positive fluxes, we might know the flux level in certain reactions that will help us have the equations appropriately constrained. We can control update rates.

We can have maximum values for uptake in internal fluxes. So here's an example. Let's walk through it. It's nice to have a couple of-- one or two examples where we walk through in a given class. So here, we're going to have an input molecule A is transported across the membrane at a rate R sub A. It then has essentially a decision fork here.

Some fraction of the A is going to go through the X1 reaction, and some fraction through the X2. The amount that goes through X1 and X2 is what we want to know. We are going to be given-- we're going to essentially clamp control the rate it goes across membrane RA. So that's going to be a given. You can see the constraint here. RA is going to be 3.

And we're going to also know the rate at which B is removed, and that's going to be kept constant at one. So we're going to solve for the two internal fluxes X, and maybe this external transport flux are C. Now, there's going to be conservation of mass, so we can start setting up the flux balances in the upper right-hand corner here of Slide 43.

Take a look at A. A is created-- essentially, the intracellular concentration of A is dependent on R sub A, and then it splits into X1 and X2. So you know that R sub A minus X1, minus X2 is going to be zero. They're going to cancel out because A, a steady-state assumption, is going to be constant. Even though the fluxes are not zero, there sum is.

And for B, a similar argument can be made X1, which creates B has to sum will also be zero. And same thing for C, all the inputs have to be the outputs. We have these two constraints where we've just said we're going to clamp experimentally the amount of A going in and the amount of B coming out. And we want to solve for the other three kinetic parameters. So we have three equations and three unknowns.

The unknowns are the two internal fluxes in R sub B. And so, another way of stating this equation X1 plus X2 is equal to 3 because we know three going in has to equal X1, X2 by flux balance. This is, essentially, the flux balance for A can be summarized as X1 plus X2 equals 3. Or in matrix formation, it's the top row here 1, 1, 0.

All right, and the 3 is in the transport column vector at the far right hand side. And similarly, you fill up the rest of the symmetric matrix with zeros and ones and twos and minus ones and the transport, the flux vectors are 3, 1, 0, for the constraints that we have.

And you solve for this. Three equations, three unknowns. You can solve it-- my standard linear algebra tricks S times B equals. So V is equal to inverse matrix S times B. The matrix we had in the previous slide, which was-- first row was 1, 1, 0.

Now, when you do the inverse matrix, you get this upper right-hand portion, and you multiply it out, and you get the column vector solution which is 1, 2, 4, for X1, X2, and R, C, respectively. Now, they're plugged into this diagram, and you can see that the whole masses balance out. 3 is equal to 1 plus 2, 2 times C, 2 times 2 is equal to 4. OK.

That's an example where it's heavily constrained, and so we can solve it exactly. However, very often, there are not [AUDIO OUT] of the measurements that have been made. We can't clamp all these fluxes, and there are many more internal fluxes than external. And so we have an under-determined system. What is a good system biologist to do at this point?

Well, the formal solution is it's no longer a point, a single point, a nice column vector of fluxes. It's now an entire feasible space. This is a kind of a cop-out answer, is that, OK, we have fewer equations than we have unknowns. That means the unknowns can occupy this entire multi-dimensional space, where however many fluxes you have, if you have hundreds of fluxes, you'll have hundreds of dimensions.

You have three fluxes, A, B, and C. Then you'll get this kind of multi-dimensional polyhedron region, and anything inside that polyhedron is the acceptable solution to the under-determined set. Well, this is still progress in that you now know that your solution is in there somewhere. But what if we wanted to find-- add some more constraints, which are not now exact constraints anymore, but they're more an optimization process.

You might have this multi-dimensional polyhedron will be bounded by inequalities. That it's less than-- is greater than zero. Right, that it's positive, or it's less of some maximum flux. But then you want to find some optimal solution, and the optimization is the thing that math has been harnessed for in some other fields outside of ours, such as economics.

When you're doing planning of transport or planning of economic investments, and so forth, you want to know how much of your resources do you want to send to the X1 route? And how much do you want to send to the X2 route? Very analogous to the situation we had in the previous slide. And that can be solved by linear programming. Linear programming often finding economic applications.

But we need to ask what is it we want to optimize? In economics, you want to optimize, typically, the bottom line. You want to lower costs, and increase profits. What we have here is we have this very commonly encountered in many metabolic systems where you have this convex polyhedron cone. Convex just means that there are no little gaps in it.

No little carved-out regions. It's all contained. That convexity allows you to now take a linear objective function, say a multi-dimensional plane or a line that you then move through this convex space. And since there are no gaps in it, you will eventually move off as you move this objective function to-- the objective function gets better and better as it goes off through this feasible space.

It eventually gets to a point where it leaves the feasible space, and that is the maximum value. The optimal value for that objective function. So we can use this feasible space combined with this objective function to find some optimal. Now, what is the objective function that we want to use? And if we were doing red blood cell, the objective function might be ATP or redox or both or delivery of oxygen.

And for a variety of other systems of importance to understanding health or biotechnology production, or other medical and engineering goals, what we want to optimize is biomass. The ability of the cells to grow and produce other cells or produce a particular subset of molecules in the cell, but let's deal mainly with the case of cells producing cells.

And what this is is a sum over all the monomers, all the components of the cell, which represent the body of the cell. So as you move-- as you transport small molecules in from the environment and incorporate it into the cell, as certainly you can think of it as a sync, it's removing those molecules from circulation from some solution.

And this sum over all the components, the ratios, the monomers, you can think of there is a fixed ratio of alanine to glycine to leucine to all the other components of the cell. And this is known. You can know this without knowing lots of other parts of the system we want to know just by taking the cell and doing a chemical composition on it. Very simple experiment. There are tables of this known provide variety of cells.

And it hardly changes depending much on how a cell is growing. What sources of carbon and nitrogen do not greatly affect the ratio of alanine, leucine, and glycine because those are determined by what it takes to run the cell. So this is important, flux you can think of this as a kind of a lumped flux, and this will be our objective function as well. So the objective function Z sometimes called, so Z is going to be equal to the flux of growth.

So you can see that you have, again, the same equation against the chemistry matrix times the internal flux is equal to the uptake fluxes. And we've got this now-- this constraint, this optimization function. So let's take-- just like we had a very simple exact solution before, now let's take this very simple linear programming or LPM solution where now it's underdetermined so we can't get the exact solution.

But we can ask what's the maximum for a particular objective function. The objective function here is not going to be the biomass production of the entire cell. It's going to be either maximizing the production of D or C. Now, this is a slightly different diagram than we had before. We still have the rate of uptake of A on the left side of this kind of circular pseudo cell. A goes in.

It makes the same binary decision-- two-way decision of X1 versus X2. It's not binary. It's these quantitative real numbers determine how much X and how much X2. And then, if it takes the upper X1 route, then it splits into B and C. If it takes the X2 route, it turns into B and D. Both cases it produces a B molecule. You can already see a constraint coming up here, which is the RA is going to be for RB.

However you go from A to the outside again, it's going to produce one molecule of B for every molecule A. This is a perfect conservation reaction. And we've said that we're only interested in positive fluxes, so you get this little triangle here of X1 and X2 are greater than zero. And they're constrained, and we've said that RA we're going to clamp so that it can't have more than one arbitrary molds per liter, per minute.

And so X1 plus X2 are constrained to be less than RA, so they are less than 1 as well. So you get this feasible space, which is this diagonally cross-hatched region, and that's the exact solution. That's the set of all exact solutions. But now, if we want to maximize a particular objective function Z, in this case, let's maximize the production of D, and we're not too concerned about anything else.

Then you get this line. You can think this is a hyper-plane going-- a line basically going up through feasible space starting at the bottom of the slide going up and up and up until it just barely leaves the feasible space. And when it does, the last point it gets to is the maximum, and the maximum happens here to be X1 equals 0, X2 equal to the max RD. Maximum rate of production of the molecule we're interested in.

If we had an objective function that went off the other axis, it would be X2 zero, and X1 equal to maximum RC. So you can see how this works. We create the feasible space with this design and objective function Z, and we run the Z off the edge of this convex space. It's important that it be convex. OK, so how applicable is this linear programming and so-called flux balance analysis?

It works when the stoichiometry is well known. For E coli, it's well known. For newly sequenced genomes, we can make connections to previous enzymatic reactions. And we can guess at what the stoichiometric matrix would be. But in that case, the stoichiometry is less well known, and in which case, you're going to have to really embrace your outliers. When you get to the N, you're going to see all the errors and go back and figure out what was wrong with our stoichiometric matrix derived from our genome.

You don't need much experimental information to run this. But you need some, and we'll explore two of them or at least to test how it's going to do data. So what are the precursors to cell growth that we are monitoring as a Z function? We want to come up with-- define this growth function in terms of the biomass.

And there's like I said there's tables of composition we'll show on in a couple of slides. But you can use this as a part of the complete metabolic network. You can also use this as observation function. It can be described as some small number of biosynthetic precursors, plus the energy and redox cofactors. Now there are many ways of doing in silico cells.

We show the red blood cells ordinary equation. The kind of in silico cells here where the optimization is fairly limited, the stoichiometric matrices have only been worked out for three, maybe five cells, three published. Yeast is on its way. And these are mostly hand curated. There is definitely a need for getting a more automated input from genomic models to these flux analyzes.

Here's some references here. We'll be talking in a moment. First, we'll talk about the wild-type case for each of these cells under a variety of different growth conditions. Then we'll move on to mutants and ask whether mutants we expect to be optimal or not. And that will be called minimization of metabolic adjustment. Now, where do these stoichiometric matrices come from?

Parenthetically the kinetic parameters that we are talking about are known for some of these complicated biochemical systems, not just for red blood cell. And where both the stoichiometric matrices and the kinetic parameters come from, of course, as a vast literature is an unwieldy literature in the sense it was done at a time before anybody thought that we were going to be-- that they were going to be responsible for getting this into computer passable databases.

So it's mostly been re-entered by technicians into databases, and you'll end up with these diagrams, such as the central one where each of these boxes contains is a node that contains a substrate. And the lines have inside information numbers this dotted-- this multiple decimal point showing a hierarchical classification going from left to right getting more and more detailed in the enzymatic reaction.

And this is built up in a database. And from this, you can access such things as effects of kinetic constants, effects of pH, and other details. And also, from this, you can get the stoichiometric matrices in principle because here, you can see that you've got, say, an input of ADP plus glucose 6 phosphate going into this reaction or coming out depending on which direction you're going.

And so those convert to zeros and ones in the matrix. That's the source of the stoichiometric matrix, which tells you what reactions are allowed. What two things can come together, or one thing can be converted by the various enzymes that actually exist in cells. And you can toggle those on and off, either by regulation or by evolutionary change or by genetic manipulation or mutagenesis. You can basically have a universal matrix, and you toggle on only the ones that you think happen in your organism.

That's how you get the stoichiometric matrix. How you get the optimization function, the Z function, which is how you want to make those decisions of how much X1 and how much X2. This is dependent on the biomass composition. And the biomass composition, say we have on the vertical axis here, is the coefficient in the growth reaction, which ranges here over almost eight logs from the lowest compositional fraction, which are some of these coenzymes.

So just like enzymes, they're only needed in small amounts to these that are needed both for major participants in hundreds of reactions, like ATP. Some of which contribute heavily to biomass. So ATP also contributes to large biomass like ribosomal RNA. And then the various amino acids, which tend to be many of these stars at the higher levels, like lysine, leucine, and the other 18 amino acids.

OK, so these are examples of the numbers that would go into the optimization curve here. These red hyper planes would be that linear some in the ratios in the previous slide. Now, we've already seen, we've already worked through an example where we slid one of these hyperplanes off the edge of in a two-dimensional model and got the optimum production of a particular substance D.

But now, we're trying to optimize this linear sum of all the metabolites that go into the body of the organism. And just focus on the green feasible space. It's actually-- some of it's hidden behind the yellow, but imagine this whole convex polyhedron of green which is feasible space, FS for wild type, WT, and you can get an optimum.

You assume that this optimum, whatever the conditions that you're looking at is likely to be achieved because for millions of years, this organism has been living through all the different growth scenarios. Glucose on glucose, galactose, various nitrogen sources, and so forth, so no matter what reasonable set of conditions you throw at it, it's seen something like that before or some other combinations, and so it's optimal for it.

And that's what that top right red dot means is that's the optimum for the wild type. Moving this hyperplane, which is the sum of all the different growth components in the correct ratio. So if you're going to optimize all the decisions, X1s and X2s so that you get the right ratios, so you don't get way too much, say, of some rare amino acid, like tryptophan and not enough of the glycine, say as alanine is most common. That's great for a wild type under a variety of different growth conditions.

But what if you throw a real curveball and say not give it a condition, but give it a perturbation it really doesn't see very often, if at all, which would be to knockout of gene completely by deletion or conceivably knock in a gene. Add a gene that it hasn't seen before very often. Well, now you could say we're going to do the same optimization.

We'll run the same red hyperplane through the new-- so the new feasible space is reduced if it's a knockout, could be increased if it's a knock in, where we're adding a gene. But in this reduced space, the original optima is no longer accessible. And if you rerun the optimization running this plane off the edge, you'll find a new red dot in the yellow space, which is a knockout optimum.

But that could take-- after you do the knockout, it could take evolutionary time or at least long lb times to allow all the other genes to mutate and be selected so that they accommodate this new knockout. That the wild type had plenty of time to do that had millions of years. The yellow feasible space for the knockout may not have had that, so you want to know what's this immediate response.

And for its immediate response, you might imagine is showing here as an orthogonal. A closest distance think of this is a Euclidean distance in this multi-dimensional space between the wild type optimum and the projection onto the feasible space of the mutant. Now that distance, you should already be thinking that this is no longer a linear programming. This may be quadratic because the distance is a quadratic function.

And we'll see in particular since there are certain pathologies where you really are forced to take-- you can't just take a projection. Because sometimes, the projections can end up in a part of space, which is not feasible. This projection does not land on the yellow feasible space, FS of the knockout. So what you need to do is nudge this purple symbol up just to the nearest point of the feasible space, which minimizes the distance to the wild type optimum, and still falls in feasible space.

Now, this is a quadratic programming algorithm. It's really just-- the linear is very simple to think of as just moving this plane off the edge of the convex space. We'll go up-- quadratic programming is a bit more complicated, but I think you get the idea you're minimizing that distance. OK, now, with any good modeling exercise, there should be some data lurking in the wings.

Many of the network models that we will do are getting more ambitious than even the massive amounts of data that are coming in. There are two types of data that might leap to mind as being appropriate for this. Remember, we said that what the organism has done is optimize the use of these networks so that you can maximize growth.

So the two sources of data that you'd want to test this with would be the flux data itself because you're saying-- you're predicting how much is going to be going in each flux direction. And the growth data, so you're making predictions that you'll optimally use the fluxes in the network in order to maximize growth for various mutants and various different conditions of growth, different carbon and nitrogen sources.

So here you have a group in Zurich, which is among the very few groups that can actually measure the internal fluxes for metabolic pathways that begin with, say, isotopically labeled glucose and end in the various amino acids. You can measure these-- we prepared ourselves for this a little bit in the last class where we talked about the isotopomers where you can have different chemical compounds whose only difference-- they have basically identical chemically.

But in different carbon positions, you have a carbon-13 in one place, carbon-12 in another. And the exact arrangement of carbon-12 and carbon-13s in this will determine the isotopomer. And if you have mixtures of carbon-12 and carbon-13 glucose-- in the upper left-hand corner of this diagram includes glycolysis and pentose phosphate and TCA cycle. This isotopically labeled glucose goes through here.

And then you need a little bit of modeling that we won't go into. But that does require a stoichiometry matrix, just like our optimization modeling does. But now, very independently of that, you need to ask how the isotopes of glucose would make their way into the amino acids. And then once you have that, then given the amino acid ratios by mass spect in NMR, you can go back and calculate what the fluxes must have been in here.

Once you have those fluxes, then you can ask how close to optimal are they for the wild type under one condition. For the mutants, under the same condition. For the wild type and mutant under different conditions. OK, so this is the first class of data that we'll use, and it's the internal fluxes for wild-type and mutants. Look in the upper right-hand corner where-- so you've got in the upper left-hand corner is the icon color-coded from the previous slide.

The upper right-hand corner is the predictions for wild type using the linear programming or FBA model. Get predicted fluxes on the vertical axis and experimental fluxes on the horizontal axis. You see, here is a good correlation coefficient of about 90 plus percent. And a probability that this would not be random.

That this is a positive linear correlation is better than 10 to the minus 7. Very unlikely to happen at random. Then you can focus in on the outliers like number 18 here, but overall, it's a good-- I mean, this can allow you to ask what experimental or model problems you might have. But let's ask-- that's the wild type under common growth conditionns. What about mutant under the same growth conditions?

Same stoichiometric modeling, same measurements, now the experimental fluxes versus predicted fluxes in the lower left-hand quadrant is almost completely random. There's no positive or negative correlation that's significant, as we might have expected for the mutant. Remember, it isn't optimal. So you don't expect linear programming method which produces the optimum to be appropriate.

Unless we had allowed this to evolve in the laboratory for hundreds of generations or a sufficient number, however many that is, to get all the other genes or enough of the other genes to adapt so that you get near an optimum. In any case, this one is random. If you now use the quadratic programming approach, the MOMA or minimization of metabolic adjustment.

Now get the very exciting result that it now becomes statistically significant again. Still a few outliers, 17 and 16 here in magenta, and these are things where you can now make very specific tests. Because now you know which ones are most discrepant between the experimental and predicted fluxes. And you can go in and ask what part of that model or what part of that data collection might be accounting for the poor fit.

But overall, this starts to give one the impression that maybe the mutants will not be eligible and that they can be somewhat better served by this quasi-non-optimal solution. And if you walk through various different conditions with different knockouts, you can see multiple examples of this. So here in the upper left-hand quadrant of Slide 59, where we have the carbon limitation condition on the far left and then the comparison of the method for flux balance on wild type.

Here's that 0.9 correlation coefficient and good P-value, and then comparing MOMA and FBA. And again, the significance of switching from FBA to MoMA is 3 times 10 to the minus 3, indicating that the quadratic programming is the more appropriate application here. And as you walk through this, you'll see many good correlation coefficients and many improvements when you go to the quadratic or MOMA, not every case, but certainly many of them.

Now that's one class of test, which is the internal fluxes. The idea is that you adjust the internal fluxes until they're optimal to produce growth. Well, how about measuring growth itself on a set of mutants. Again, these mutants you expect will follow the MOMA prediction slightly better. And so Slide 60 is a particular way of measuring a large genome-wide set of mutants.

This is not a trivial undertaking. We know how to measure a transcriptome set of RNAs. Where on microarrays, maybe there's an analogous way to measure genomes worth of mutants. Ideally, you would have a mutant, not just one per gene. One knockout per gene, which is maybe a very expensive handcrafted deletion, but you'd have a little targeted mutation in every domain that contributes to that gene function.

Mutations in the DNA regulatory elements, various protein domains, RNA stability domain, and so on, not quite at that level of ideal, but getting close to it. It's transposon mutagenesis, where you insert small bits of DNA randomly throughout the genome. The modern set of commonly used transposons are pretty random in their insertion site choice. And then, you need a way of turning this collection of transposons into a readout of growth rates.

Well, one way to do this is to have a population of cells, each having their own distinctive transposon and growing them as a mixture. And as they grow as a mixture, the ones that grow better will dominate the population. And so their transposon hit the junction between the transposon-- the transposons is universal. It's present in every cell.

But where it sits is unique to that cell, or nearly so. So it sits in-- the junction between transposon and genome is unique, and you want a way of assaying that. The way you assay that is you take the complete DNA from the entire cell mixture, and you want to say how much of each transposon exists? So you cut the entire DNA mixture with a restriction enzyme that cuts frequently. So called 4 cutter, cuts with about every 236 base pairs on average.

And you like it on this very special kind of linker, which is a linker which will not amplify with the corresponding primer until some other DNA synthesis has gone through because you see this little Y linker is not perfectly logically based paired. It requires DNA polymerase to make a complement. Complement then will bind the primer.

It's a long way of saying that only in cases where you have a primer in the transposon near one of these Y linkers will you get amplification. Transposon alone, you won't get it. Y linker alone, you won't get it. That's one step of enrichment, and the reason you have to enrich is if you just threw this whole thing on the microarray, you'd get everything lighting up because every genome has every piece of the genome in it, and the transposons are present in trace amounts.

So you need to amplify the junction fragment by first this trick, followed by a T7 promoter. This is a phage promoter a very commonly used. Very clean RNA polymerase background that's been incorporated into any transpose on your favorite transposon. So now you have two steps. First, this ligation-mediated PCR, and second the T7 in vitro transcription to make RNA. Now, you've basically reduced this to a problem similar to transcriptome microarrays that we've done before.

Now the RNAs are surrogates for the amount of each strain. We're not measuring RNA. We've made an RNA that represents the transposon junction fragment and hence allows you to quantitate the amount of that particular mutant in a population. As that mutant goes up or down in the population, so does the RNA from this assay. So you might ask at this point, well, this is great, but I don't trust it. Just like some people might not have trusted mass spectrometry as being reproducible enough to be quantitative.

So the way that you determine whether something is sufficiently reproducible to be quantitative, one way to do it is to do two independent selection experiments. Evolutionists might say, oh, if we reran evolution all over again, we would not get the same set of organisms we have on Earth today. That may be true, but that's because there are all sorts of bottlenecks in historical events.

In this case, we specifically tried to make it reproducible by avoiding bottlenecks by having every mutation, every transposon type represented by 1,000 different independent events. As well as 1,000 copies of every transposon hit, so you keep the population size large, and so it becomes more reproducible.

And the way you measure the reproducibility is you do the two selection experiments. The two ideally independent sets of transposons subjected to exactly the same selection procedure. And you see a very gratifying curve here with an R-squared regression measure in excess of 98%. And the kind of scatter that you would be pleased to get in an RNA microarray experiment.

So this is reproducible and, therefore, quite capable of producing quantitative data. Well, now let's go back and compare it to the two different methods of modeling the flux optimization. There's the FBA or linear programming, which assumes that the mutant is immediately optimal. And then there's the MOMA or minimization of metabolic adjustment, which doesn't assume immediate ophthalmology, but assumes that it's close to wild-type optimal.

So starting at the top with the linear flux balance, you can classify the predictions of the model for each gene. And here we have almost 400-- almost 500 different genes mutated, and first you run through in silico predictions you can classify them as being essential for a particular growth condition. They aren't necessarily essential for every growth condition.

But the growth conditions here the essential or nonessential or something in between. We'll just focus on the extremes of essential and nonessential. And then, the experiment can be classified as to whether there's significant negative selection or no noticeable selection in the particular growth conditions. And so you expect the ones to be essential in these growth conditions to be heavily negative selectively, and you have 80 examples of that out of 142 predicted.

However, you wouldn't expect any to be missing in selection since they're essential. So the 62 should be a zero for the 62. This is an example of how we're going to try to explain or use this as a way of generating interesting hypotheses about the exceptions or outliers. Similarly, the nonessential genes should have no selection. This 180 in the lower right-hand part of the upper FBA, but the 119 should be zero.

So the first pass explanation is OK. We knew that these weren't going to be optimal right away. So let's use the other model, the sub-optimal, nearly optimal model MOMA and see how well it does. Well, on the far right-hand side of Slide 62, you see the chi-square P-value is how well the predictions agree with the observations. And it is significant for the linear programming.

It's 4 times 10, then minus 3, fairly significant. And the MOMA is much more significant than the minus 5. It has improved some of those upper right 66 and lower left 108. Those should still be zero, and they're not. And the extent that they deviate from the ideal expectation means that either the data-- or somehow the way we're collecting the data not the way we expected it.

Or that the model, more likely the model has some problems in it. And examples of problems for those two classes, the predicted essential genes which show those-- which show no selection, might be examples of redundancies that we failed to model when we put together the stoichiometric matrix. The stoichiometric matrix we take every known biochemical reaction.

We might take convincing homologs sequenced homology level, and say those might be examples of redundancy. We might take analogous or parallel roots that are known documented biochemically, where you can get to the same point by a series of possibly non-sequence homologous enzymatic steps. So those are novel redundancies. This is an example of potential discovery, and can be pursued in a very directed way because these are 66 very specific predictions.

108 essential gene knockouts, which nevertheless show negative selection, could be examples of position effects. Where a mutation in one gene affects other genes that are close along the DNA position, meaning along the DNA, an example of this is that a variety of mutations in a gene which is upstream from another gene in an operon, have poorer effects on downstream genes due to the coupling of transcription and translation.

You can have other position effects as well. These are examples of possibly many explanations and discoveries that can contribute to the wonderful nature of-- it's a win-win situation. Either you get great correlation, or you get great exceptions and discoveries, or both. OK, now, when we talked about redundancy as one of the possible explanations in the previous slide, that brings up really important conceptual component of the post-genomic functional genomics world.

Which is what do we do about multiple homologous domains? When you go through and you sequence genomes, you find parallels. You find either whole genes or pieces of genes which have high sequence homology, which we talked about at the beginning of the course. And here's examples of three protein-coding regions. We encode enzymes involved in the biosynthesis of amino acids.

So you can imagine that when you grow the cells on minimal media that you are not providing the amino acids, so the cell has to make the amino acids. If it's got a mutation, a transposon hit in one of these genes that's key in the biosynthesis synthesis of lysine 3 and methionine. And you might expect that it won't grow well. It'll be at a selective disadvantage in minimal media unless one of the other genes covers for it. As possibly as green-- so these have two or three domains here color coded.

So the green domain is shared by these three biosynthetic proteins that are otherwise very different in their metabolic contributions. And you might imagine that some of these green domains might cover for others when one of them is mutated. One of them, however, the one in lysine when it's got a transposon hit in that domain, the replication rate selective disadvantage in minimal media is severe. It's a fact. It's an order of magnitude, while the others are more subtle.

It could be that the lysine covers for the other two because it's made in large amounts, is very active. But the other two can't cover for the lysine, or there's a variety of other possible explanations for this observation. The point is that this generates hypotheses that allow us to follow up on the flux balance or MOMA type of modeling.

And I think the reason I spent so much time on it is I think the concept of optimization is something that's important both in the sense that we're in engineering, as engineers, we're optimizing living systems. And also, as students of evolutionary systems where to understand what those systems are optimized to do, we need to look at from this perspective.

So this is just what we've covered today. We've covered both continuous and discrete ways of modeling molecular systems. And, in particular, the red blood cell and the copy number control, as a way of dealing with metabolism and biopolymers separately, and the flux balance has brought these together, and brought in the notion of optimization. OK, so until next time.