

Metabolic Engineering
Prof. Amit Ghosh
School of Energy Science and Engineering
Indian Institute of Technology - Kharagpur

Lecture - 22
Flux Sampling, Optknock and Optstrain

Welcome to metabolic engineering course, today we will discuss about flux sampling optknock and optstrain. Flux sampling is very important concept where you can sample the solution space of the metabolic network that you are actually analyzing. It can be E. coli metabolic network or yeast metabolic network and optknock is a strain design algorithm and optstrain is also a strain design algorithm or you can design the strain in silico. And then once we have a better prediction you can go in the lab and perform the experiment.

(Refer Slide Time: 01:02)



So, this involves the sampling flux solution space that we are going to learn today and the strain design algorithm optknock and optstrain. These are very popular algorithm which is used by industries, academia and then we have a lot of success.

(Refer Slide Time: 01:20)

Optimal Solutions

1. FBA
2. Flux Variability

Flux Dependencies

1. Robustness
2. Phase Planes
3. Flux Coupling

All Allowable Solutions

1. Extreme Pathways
2. Elementary Modes
3. Sampling

Altering Phenotypes

1. Genetic Mutations
2. Strain Design

Application of Additional Constraints

1. Regulation
2. Energy Balance

Constraint-Based Methods Price, Reed, and Palsson Nat. Reviews Microbiol.2004

So, this is a constraint based method which we discuss every class for last 2, 3 days, where you can see that we learned about flux variability analysis, which is shown over here and then yesterday we learned about flux sampling and then today we are going to learn about sampling. And flux coupling we learned in previous class robustness analysis also we learned in previous class.

So, dynamic simulation we learned in previous class, gene deletion also we learned in previous class. So, this today we are going to learn about the sampling and also optknock. So, these are the 2 things we are going to learn today.

(Refer Slide Time: 02:09)

Introduction

- Methods exist to characterize metabolic solution spaces such as extreme pathway analysis
- The effect of imposing constraints can also be studied using randomized sampling methods by defining the 'size' of the solution space and how the space changes
- Randomized sampling can be used to characterize flux solution space, concentration space, kinetic solution space
- Randomized sampling can be used to understand network capabilities, enzymopathy, and diseases states (e.g. ischemia, diabetes)
- It can be applied to spaces that are linear or non-linear (convex or non-convex)

Price, Reed, and Palsson Nat. Reviews Microbiol.2004

So, I will just introduce how you can actually characterize the metabolic solution space. So, the solution space you can see is actually bounded by the constraint and how you can sample the solutions with a dot-dot you can see we can have to sample the solutions space that you

can get and get your metabolic network very well defined. That is the generally use a FVA flux variability analysis to get the range of fluxes but sampling is the best way where you can sample the solution space accurately and then characterize your metabolic network.

So there are method to characterize metabolic solution space, such as extreme pathway analysis, this we will learn in subsequent classes and then the effect of imposing constraint can also be study using randomized sampling methods. Defining the size of the solutions space and how the space changes. So, you define the space of the solution space.

And then how the solution space changes with the constraint or, with there is a function of time, you can see how the solution spaces are actually changing with time. And randomize sampling can be used to characterize flux solution space concentration solution space also kinetic solution space and the randomized sampling can be used to understand the network capabilities enzyemopathy disease state that is their diabetes condition of the cell or ischemia condition.

These are the disease condition which you can characterize using the flux sampling method. So it can be applied to spaces that are linear or nonlinear can be applied to convex or non convex spaces, and solution spaces. So mostly we will be applying in convex solution space where the sampling is needed to understand the type of network, the network capabilities in different disease conditions that also can be evaluated.

(Refer Slide Time: 04:14)

Sampling A Space

What is the area of this object?

Absolute Volume Calculation

Size of Space = Hit fraction (p) \times Size of box

Keeping the red points that fall within the shape leaves a uniform sample of points

$$\text{Error} = \frac{p(1-p)}{N}$$

- Goals for shape choice:
 - Easy to generate uniform points within
 - Easy to calculate volume

The slide features a diagram of a blue, irregular shape with red points scattered inside and outside it. The text explains that the size of the space is determined by the hit fraction (p) multiplied by the size of the box. It also provides the formula for error and lists goals for shape choice: easy to generate uniform points within and easy to calculate volume. The NPTEL logo is visible in the bottom left corner.

So let us see what is a sampling a space? What do you mean by sampling a space? What is the area of this object can you calculate the area of the object it is very complicated. To calculate this 2 dimensional object I am showing over here. So a 3 dimensional object can be more complicated. Make it simple, we considered a 2 dimensional object and the 2 dimensional objects you can actually calculate the area just by sampling.

So if you do uniform sampling of this area, then you will be able to calculate the area just by calculating the number of points hitting the area so this I am just doing random sampling. So you can see the points are getting filled up in the designated box, so, I define a red color box and within that box I have my object, I have a very weird object. So you cannot calculate the area of that object.

So if you sample the space, so I am sampling the space I just uniformly I am putting dot so that no 2 dots are actually overlapping each other and they are actually equally spaced and they are equally far apart. So, after some time, you can see that the box is filled with points and then you calculate what is the absolute volume? That is the size of the space, the size of the space is basically hitting fraction.

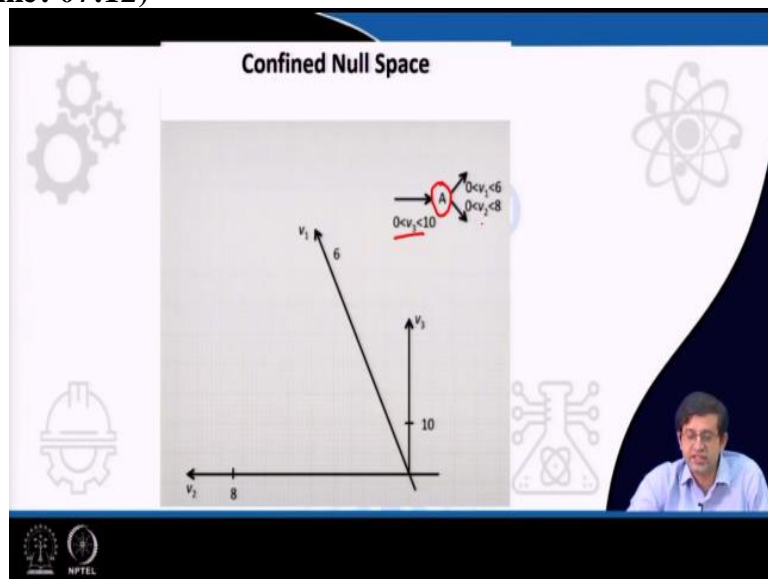
So, how many dots is actually hitting the blue area that is your hitting pressure is the number of points hitting and divided by the total number of points. Total number of points, you know the total number of points you are supplying and the how many points are actually hitting the area that you calculate by counting the number of points and that becomes the fraction of points, which is hitting the area of object and then the size of the box then you multiplied by the size of the box that becomes the size of the space.

So, this is the size of the space that and the object you have defined and that become the size of the space keeping the red point that fall within the shape leaves a uniform sample of points. So, you want to make sure that it is a uniform sampling and the error is given by the p you have calculate that is heat fraction $1 - p$ and multiplied by p divided by N . So, this N is the number of points so, this way you can calculate the error and also the size of the space that you are actually considering here.

So, goals for shape choice: Easy to generate a uniform point within an easy to calculate volume. So any shape you can generate and any and the volume of the shape can be

calculated by sampling that space. So, this is widely used in various fields, where you can calculate the surface area of any object.

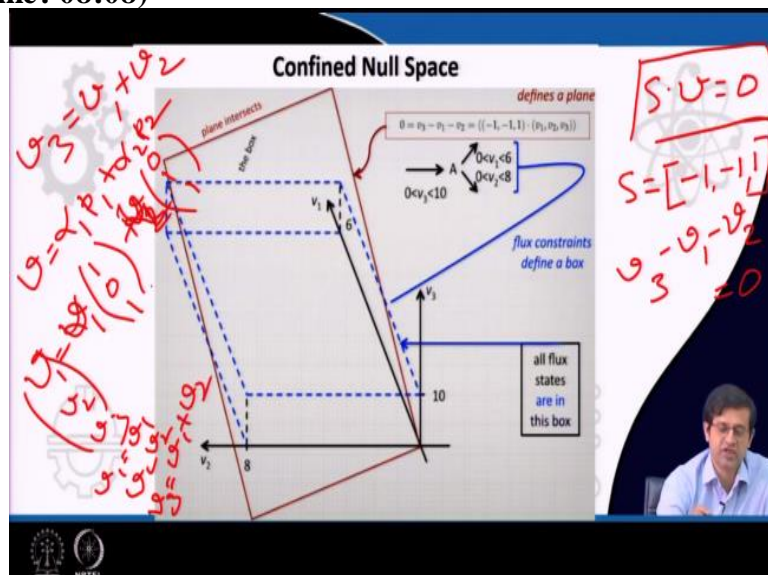
(Refer Slide Time: 07:12)



So, this to make it more clear in the metabolic network I can consider a confined null space that is a that I have a network with 1 metabolite so we have a 1 metabolite network A. And which is actually the reaction which is producing metabolite A is v_3 and v_3 is goes from 0 to 10, so it is greater than 0, so v_3 is greater than 0 and less than 10. And then it is going to reaction v_1 and v_2 .

So v_1 and v_2 is actually goes from 0 to 6, and v_2 goes from 0 to 8. Now, I want to plot this flux in 3 dimensions so how it looks like that. And in 3 dimensional, you want to plot these 3 fluxes, so v_3 goes from 0 to 10, v_1 goes from 0 to 6, and v_2 goes from 0 to 8.

(Refer Slide Time: 08:08)



If you put a rectangular box, then what we will see, that all the fluxes states are allowed in this box. So any point in this box is actually a solution for this small network, which has 3 reaction. So the boxes defined by you can see this is v_2 this is v_1 and then we have v_3 . So these are the 3 axis. For the 3 fluxes, we have 3 axes, which you can see from the network. So I am just removing v_1, v_2, v_3 .

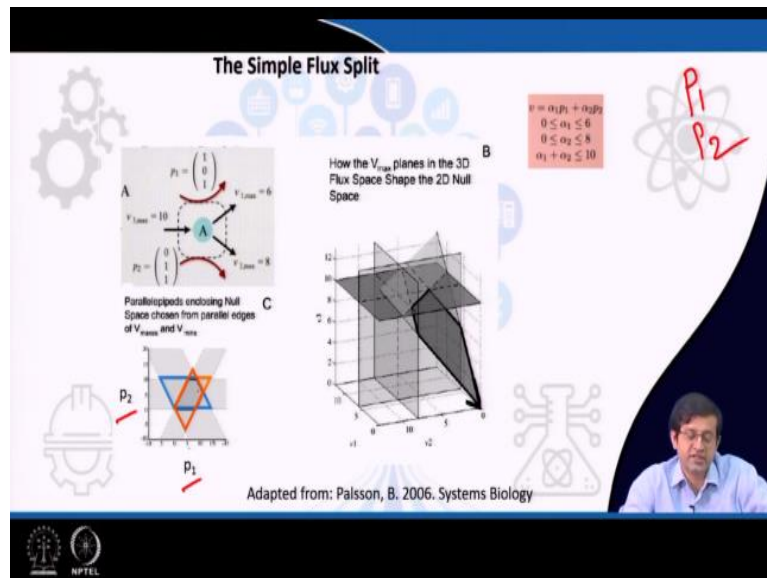
Now, you can see that any point in this box is a solution. So, it is not a steady state solution. So this is a solution total solution space, which is given by v_1, v_2, v_3 and then I define a steady state condition or the null space vector. So I consider a plane so plane which is intersecting the box that is $S \cdot v$ the plane is defined by $S \cdot v = 0$ that is the steady state approximation.

So this is $S \cdot v$ is nothing but a plane. So the stoichiometric mathematics for this network is basically minus 1 minus 1, 1. We have only 1 metabolite that is why I have 1 row. So the s is basically minus 1, that is the v_1 , then minus 1 for v_2 and then 1 for v_3 so this I have 3 components. So this if I multiplied s with v , then I get $v_3 - v_1 - v_2 = 0$ this is the equation of a plane.

And if I draw the plane and the plane which is intersecting the box and the plane and that is the intersection region you can see the plane while intersect the box is the plane where the steady state condition is valid that is, the steady state fluxes are defined on that plane only. So, this plane actually cutting the box and on that plane there is also a steady state solution lies which obeys $S \cdot v = 0$. So, this obeys $S \cdot v = 0$ as the steady state solution space.

And now, using this to the steady you know steady state solution space is basically a null space $S \cdot v$ is nothing but a null space. Any matrix S you can calculate the null space any matrix you can see from any other literature where you can see that how we can calculate the null space of a matrix. So, S is a matrix and that call the matrix you are calculating the null space and $S \cdot v = 0$ gives you the null space.

(Refer Slide Time: 11:27)



So, what is the null space for this equation $S \cdot v = 0$, which is nothing but the null space vectors are actually given by p_1 and p_2 . So, we can see that the p_1 if I define p_1 and p_2 are the null space vector and the span covered by these null space vector actually you will see this a solution lies. So, here we can see the null space vector p_1 and p_2 it can easily be calculated.

So, because I have an equation of v_3 equal to $v_1 + v_2$, previous question you can see that $v_3 = v_1 + v_2$. I rearrange this equation $v_3 = v_1 + v_2$. So, if I assume that the null space vector it says that v_1 if I define the vector $v = \alpha_1 p_1 + \alpha_2 p_2$ then your total flux will be α_1 multiplied by $(1, 0, 1)$ plus α_2 multiplied by $(0, 1, 1)$ and this p_1 will be because v is $(v_1, v_2$ and $v_3)$. So, here, since $v_3 = v_1 + v_2$ and $v_3 = v_1 + v_2 = (1, 0, 1)$ and this one will be $(0, 1, 1)$ so, that if you multiply that, you will get $v_3 = v_1 + v_2$.

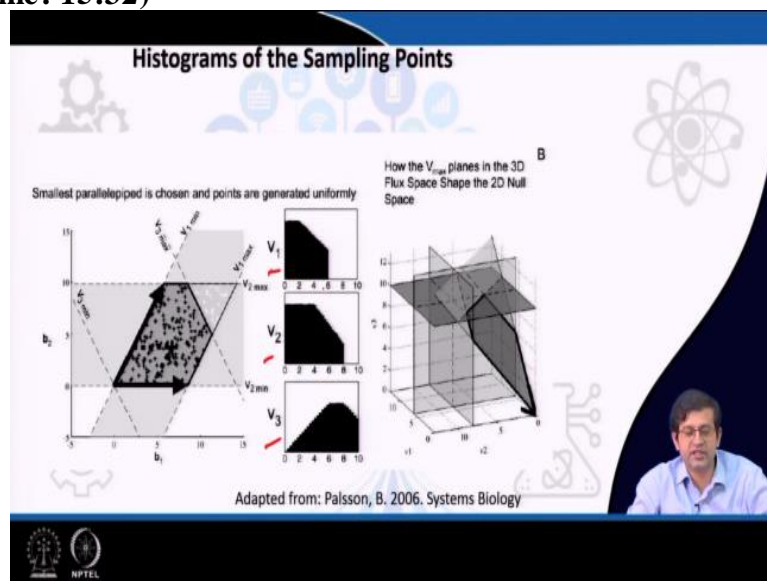
So, here you see that this is v_1 and then this is v_2 . So, $v_1, v_2 = v_1 + v_3$ and then $v_1 = v_1$ and $v_2 = v_2$ and $v_3 = v_1 + v_2$ according to this equation. So, this is the null space vector which is given by p_1 and p_2 which is basically you can express as a linear combination. So, the total flux which is given by the net flux is actually given by $\alpha_1 p_1 + \alpha_2 p_2$, p_1 and p_2 are the null space vector and α_1 goes from 0 to 6 and α_2 goes from 0 to 8 says that $\alpha_1 + \alpha_2$ is less than or equal to 10.

Now, you consider this small flux split that the simple flux split. Now your flux vector changes from v_1, v_2, v_3 to p_2, p_1 . So, these are the vector the basis vector for this matrix and the basis vectors that you get by calculating the null space you can calculate the basis

vector p_1 and p_2 and if you plot then what is your final solution space is governed by is actually govern by this p_1 and p_2 .

So, if you draw the small 2 dimensional flux split as a function of p_1 and p_2 you see that because of the constraint the first box in the yellow color you can see there that rectangular boxes form and then the other rectangular boxes because of the constraint of p_2 , and which is in blue color and then the other box, then if you intersect this all 3 box and what you get? The common region of all the 3 boxes that is there is a region the flux plate which is actually feasible or the steady state solution space lies. So, that is the region your steady state solution lies.

(Refer Slide Time: 15:32)



And in 3 dimensions, you can see this is shown over here. So, this is a 2 dimensional map and this is a 3 dimensional map. The more clear you can see if you plot the smallest parallelepiped is chosen and the points are generated uniformly now we sample this space shape. After applying constraint, you can actually sample the space and you can see that; your, final v_1 , the v_1 , v_2 , v_3 looks like this after sampling.

So, it goes from 0 to 6, and the maximum the values are shown over here, it goes from 0 to 10 and 0 to 8 and 0 to 6. And how the maximum value changes you also can be visible from this histogram. So, this is a histogram plot 3D histogram where you can know how much fluxes are actually going through the reaction. So the null space actually represents all the possible function or a phenotype state of a network at a particular point in the phenotype represent one function or one particular phenotypic state.

And then how do you sample now the next point and once you find the flux split, so this is the way you can calculate the flux speed by applying constraint. And then how do you sample the solution space?

(Refer Slide Time: 17:05)

Hit-and-Run Sampling
Markov Chain Monte Carlo Sampler

- One initial point x_0 ✓
- Random direction and step size choice to go to the next point

Crucial point: reducing number of steps needed to converge to a uniform distribution

Key parameters include:

- How the random direction is chosen?
- Number of steps 'bounced' between saving points?

Hit-and-Run Sampling

u2

u1

u3

x_0

And then how do you calculate the solution space that is one of the routine algorithm is the hit and run sampling. So you start with a Monte Carlo sampler Markov chain Monte Carlo sampler to one with one point. So you start with point $x = 0$ and the random direction and step size choice to go to the next point. So, randomly you choose the direction and go, it will go to the next one and the crucial reducing number of steps needed to converge to a uniform distribution.

So here, you have to make sure that you should converge to a uniform distribution and the key parameter including the in the sampling is basically how the random direction is chosen and the number of steps you take. You have to save those points and so these, number of step bounced, so you actually some of these steps are bounced that you have to save so that you do not have to repeat those bounced step.

So, this is an example where you can hit run on sampling, we have 3 flux vector u_1, u_2, u_3 and then you want to sample the solution space. So, you start with the initial point X_0 which is shown in green color, and then you start the sampling and randomly choose the direction. So randomly, you can see that you go to a blue point and then the next day we go to another and another point randomly.

And then you can see that at this point it is getting bounced so it cannot go further direction it cannot go in that line. So and then it will go bounce back and go into another direction in this way randomly simple sample the solution space.

(Refer Slide Time: 18:43)

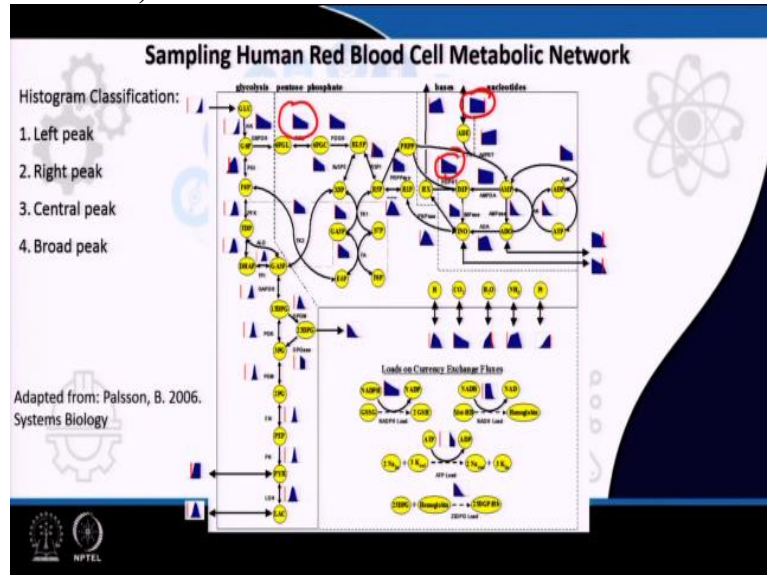
The slide is titled "Artificially Centered Hit-and-Run (ACHR)". It contains three bullet points: "First estimates the center of the solution space", "Then the estimated center is used to inform the direction of further sampling such that the full solution space is covered in fewer steps", and "Overcomes the edge-trapping limitation of the standard hit-and run algorithm". Below the text is a 3D diagram of a tetrahedron in a coordinate system with axes labeled u_1 , u_2 , and u_3 . A point x_0 is marked on one of the edges. A dashed line represents a path starting from x_0 , moving towards the center, and then reflecting off the edges. The NPTEL logo is visible in the bottom left corner of the slide.

And then artificially centre hit run another sampling method, where you can fast estimate the centre of the solution space and then estimate a centre is used to inform the direction of further sampling such that the full solution space is covered in few steps. Then the estimated centre is used to inform the direction of further sampling such that so you can cover it in a fewer number of steps and overcomes the edge trapping limitation of the heat and run sampling.

So, in the heat and run sampling it may trap it is found that the edge trapping limitation is not there in ACHR and this is how it is done. So, you first estimate the centre of the solution space and then estimate the centre, estimated centre is used to uniform the direction of the sampling. So using that centre you try to sample uniformly these flux solution space in a few number of steps.

So the number of steps will be lesser then standard heat and run sampling, we start with X_0 point and then you calculate your centre of this solution space and try to connect it and then you define your next point and this way you choose another point and then so on you calculate the direction. So, using the centre you calculate the direction what should be the next iteration step and this way you can actually sample the entire solution space.

(Refer Slide Time: 20:18)



So, using the sampling of the human red blood cell metabolic layer network has been simulated. And here you can see the human cell metabolic network is shown over here and then the steady state flux solution space of the human red blood cells have been studied through randomized sampling, which was previously studied. And the probability distribution of each flux in the network can be shown on the reaction map. The probability distribution is shown over here.

So, this is the probability distribution the y axis is the probability and the x axis is the flux value, how much probability it may have that is shown in the y axis. This allow one to visualize all the allowable flux values for all the reaction in the network simulation these studies have led to several notable results 3 of which I will discuss briefly. The histogram provide information about shape of the solutions space.

So, the histogram you can see that each flux has the glycolysis and the PPV pathway, you can see that the histogram. These simple histogram actually provide the shape of the solution space and how likely the flux are to fall into certain numerical values. For instance, some of the histograms are flat implying that every numerical value for a flux with a particular reaction is equal likely.

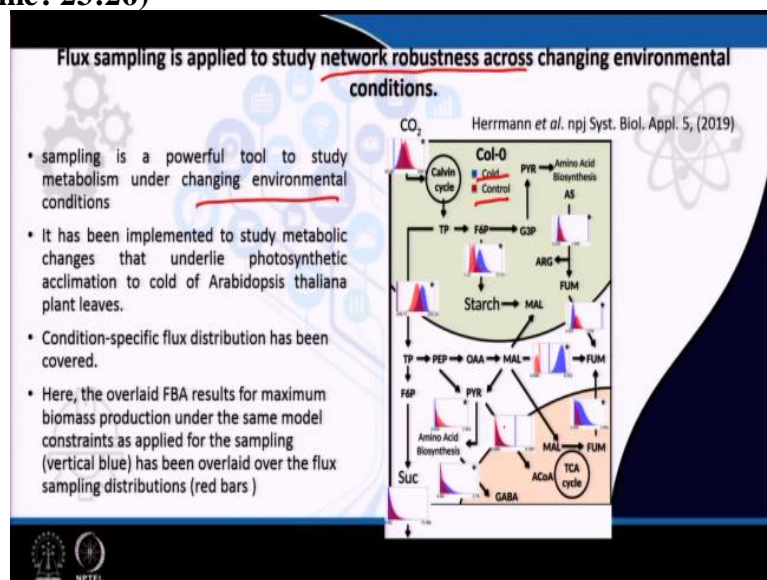
So, some reaction fluxes are actually flat in this case, in this case, you can see that it is flat. So, all these values are equally probable. So many of them are actually flat. So, all these value on this range is actually equally probable and the cross correlation between any pair of

flux can be computed. So, you can go also correlate 2 fluxes and correlation of values also can be computed.

And you can see how the flux values are correlating and this is another way you can actually say how 2 reactions are correlated. The measurement of poorly correlated flux is likely to be more informative than measuring highly correlated fluxes. So, some reactions are very highly correlated, but there are many reactions which are very less correlated. So, those poorly correlated reactions are also have values for higher importance.

The pairs of fluxes that are correlated can be studied further. So, the pair of fluxes which are actually correlating each other can also be studied further. So, this way you can actually sample the solution space and for the red blood cell which is reported earlier and then the histogram classification also you can do and the some histograms are left peak some histograms are right peak some histogram are central peak and some histograms are actually broad peak. So, based on that, you can know how the probability changes for a left peak reaction histogram and then identify which of the reactions are actually highly probable.

(Refer Slide Time: 23:26)



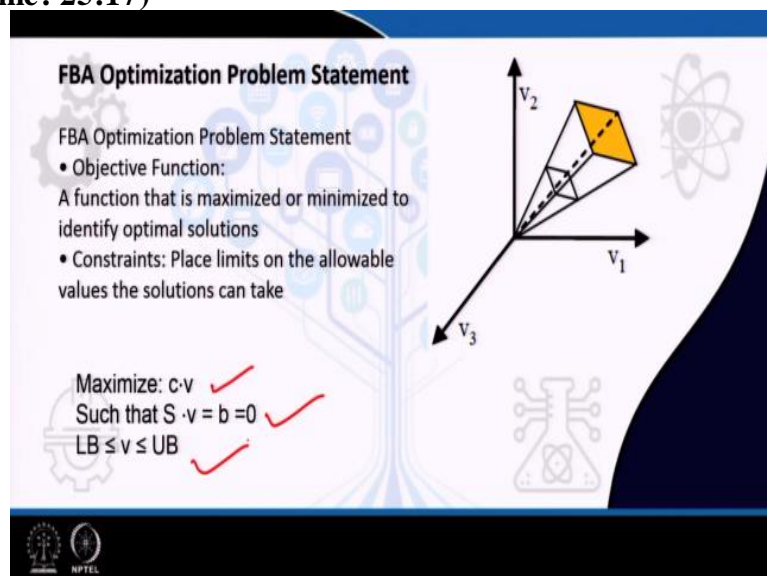
And then the flux sampling is also applied to study the network robustness. So, this is another example where network robustness is actually applied in a network where the environment is changing. So, environmental conditions are changing where robustness have been calculated. The sampling as I told is a powerful tool to study metabolism under changing environment.

So, sampling can be used to actually study the metabolic network, where the environments are actually changing the environmental condition, the moment the environmental conditions, the regulation changes and the network changes and the network especially the genetic circuit. The wiring diagram of the life of the cell is also changing. It has been implemented to study metabolic changes that underlie photosynthetic condition to a cold.

So, we have a control condition and have a condition cold. So, these are the 2 condition where the sampling, have been compared. So, the blue ones are actually in the cold condition and the red ones are actually in the control. So, the flux histogram that has been calculated in 2 different conditions and then you compare how the fluxes are actually the probability of the fluxes are changing with time in photosynthetic organism.

Here the overlaid FBA result for maximum biomass condition under the same model, as applied for the sampling has been shown over here where are the blue colors are actually represent the, the change in the condition while the cold condition is applied and the red bars is actually the control. And you compare how we the flux values are changing by applying a different environmental condition.

(Refer Slide Time: 25:17)



The slide is titled "FBA Optimization Problem Statement" and contains the following text:

FBA Optimization Problem Statement

- Objective Function: A function that is maximized or minimized to identify optimal solutions
- Constraints: Place limits on the allowable values the solutions can take

Maximize: $c \cdot v$ ✓
Such that $S \cdot v = b = 0$ ✓
 $LB \leq v \leq UB$ ✓

The slide also features a 3D coordinate system with axes labeled v_1 , v_2 , and v_3 . A yellow polyhedron is shown within this space, representing the feasible region. There are also faint background graphics of a cell and a molecular structure.

So, now, we will move into another concept that is the strain design algorithm as I told the strain design algorithm, this is the FBA problem statement, you can see that you maximize the objective function and then apply the steady state condition and then boundary condition. This is a FBA problem statement, which we have discussed earlier. Now, how you can design this strain?

(Refer Slide Time: 25:43)

Computational Design of Mutant Strains: OptKnock

- Finds reactions, that if removed, couple of biomass production and metabolite production (ie. Higher growth requires higher metabolite production levels)

OptKnock: Find gene deletions needed such that maximizing biomass is coupled with maximizing metabolic engineering objective

Strain Designs for:

- Lactate Production ✓
- Succinate Production ✓
- 1,3 Propanediol Production ✓
- Chorismate Production ✓
- Alanine Production ✓
- Serine Production ✓
- Aspartate Production ✓
- Glutamate Production ✓

Optimization Problem:

- maximize bioengineering objective (through gene knockouts) ✓
- subject to maximize cellular objective (over fluxes) ✓
- subject to
 - fixed substrate uptake
 - network stoichiometry
 - blocked reactions identified by outer problem
 - number of knockouts \leq limit

Graph: A graph with V_{chemical} on the y-axis and V_{biomass} on the x-axis. Point A is the origin. Point B is on the x-axis. Point C is the top vertex of a shaded triangle. A dashed line from A to C is labeled (WOMA). The text 'Complete E. Coll Network' is near point B.

Burgard, Pharkya, Maranas. Biotechnology & Bioengineering. 84(6): 647-657

And for strain design algorithm the optknock is very popular as I told you. Using optknock, you can actually design a strain. So, the strain design has been applied for lactate production, succinate production, propanediol production, chorismate production, alanine, serine, aspartate, glutamate these are the molecules it has been already applied for a while you can design the strain for better output.

So, optknock actually find gene deletion so, what it does? Optknock actually gives you a set of gene which you need to delete such that so, it does by the algorithm works in a way that it maximize the biomass and that is coupled with the metabolic engineering objective function, which is exactly given over here. So, we have a metabolic bioengineering objective function that is the product you are looking for suppose you are looking for succinate.

Suppose you are looking for lactate, that becomes your objective function that the flux for those reaction becomes your objective function and then we have a cellular objective function. So, we have 2 objective functions, it is a bi level optimization, where 2 objective functions were considered and the cellular objective function this is nothing but the FBA. So, this is nothing but the FBA and but the maximization of the biomass, bioengineering objective function is actually controlled by the number of knockout we were doing.

So, the number of knockout that is the limit you can put that is number of knockout you want to perform by doing that limit. Suppose I choose I just want to have 3 knockouts. So then you can put your limit equal to 3 less than equal to 3. So, that actually it looks for the network, the

combination of the gene that to be knocked out. So that your, chemical production goes high. So here you can see that this is a mutant network.

So, here this is the optimal condition with the originally have the optimal quantity, the maximum biomass is here, but the maximum biomass you decrease the biomass, so that your production become high. So initially, there is no production of the chemical. So v chemical can be any compound you are targeting. So, initially at point A, for the wild type strain, there is no way you can produce a chemical that you are looking for.

But when once you apply the knockout, the gene deletion, the combination of gene deletion says that you are the v chemical increases and at the same time the biomass decreases. So, the biomass decreases, and your chemical increases, this is a play between the biomass growth rate and then the chemical you are producing.

(Refer Slide Time: 28:42)

Succinate			OptKnock	MOMA
ID	Reactions	Enzyme	Biomass (1/hr)	Succinate (mmol/hr)
Wild	"Complete network"		0.38	0.12
A	1. COA + PYR → ACCOA + FOR 2. NADH + PYR → LAC + NAD	Pyruvate formate lyase Lactate dehydrogenase	0.31	10.70
B	1. COA + PYR → ACCOA + FOR 2. NADH + PYR → LAC + NAD 3. ACCOA + 2 NADH → COA + ETH + 2 NAD	Pyruvate formate lyase Lactate dehydrogenase Acetylaldehyde dehydrogenase	0.31	4.79
C	1. ADP + PEP → ATP + PYR 2. ACTP + ADP → AC + ATP or ACCOA + Pi → ACTP + COA 3. G6C + PEP → G6P + PYR	Pyruvate kinase Acetate kinase Phosphoenolpyruvate carboxylase Phosphoenolpyruvate carboxykinase	0.19	13.11

Lactate			OptKnock	MOMA
ID	Reactions	Enzyme	Biomass (1/hr)	Lactate (mmol/hr)
Wild	"Complete network"		0.38	0
A	1. ACTP + ADP → AC + ATP or ACCOA + Pi → ACTP + COA 2. ACCOA + 2 NADH → COA + ETH + 2 NAD	Acetate kinase Phosphoenolpyruvate carboxylase Acetylaldehyde dehydrogenase	0.28	10.45
B	1. ACTP + ADP → AC + ATP or ACCOA + Pi → ACTP + COA 2. ATP + PEP → ADP + PEP or PEP → GAP + DHAP	Acetate kinase Phosphoenolpyruvate carboxylase Phosphoenolpyruvate carboxykinase Phosphoenolpyruvate decarboxylase	0.13	18.00
C	1. ACTP + ADP → AC + ATP or ACCOA + Pi → ACTP + COA 2. ATP + PEP → ADP + PEP or PEP → GAP + DHAP 3. ACCOA + 2 NADH → COA + ETH + 2 NAD 4. G6C + ATP → G6P + PEP	Acetate kinase Phosphoenolpyruvate carboxylase Phosphoenolpyruvate carboxykinase Phosphoenolpyruvate decarboxylase Acetylaldehyde dehydrogenase Glucose-6-phosphate dehydrogenase	0.12	18.13

So, this optknock has been applied for succinate production. So for succinate production the optknock has been applied and you can see that the wild type strain is actually having a growth rate of 0.38 and the succinate production in the wild is only 0.12 millimole per hour. And the moment you apply 2 gene knockout 2 gene in the sense they have removed 2 reactions.

So they have remove 2 reactions and that is the pyruvate formase lyase and the lactate dehydrogenase and so they found that the biomass has reduced 1.38 to 0.31 and then the production of succinate has really improved 0.1 to 10.7. And then in the set, second set of

calculation, they found that they delete 3 genes. So, these 2 genes are actually already deleted and they have added one more gene deletion that is acetaldehyde dehydrogenase they found that the production of this succinate has not improved or not the biomass has improved.

But when they go went for 3 reaction deletion, then they found that the production of succinate has improved to 15.15 and the biomass reduced significantly. So biomass has reduced, become half less than half and this succinate has actually really improved. Similarly, the optknock has been applied for the production of lactate. So, lactate they removed 3 reactions and then 4 reactions.

So, this many number of reaction they have removed and they found that when they remove acetate kinase and phospho transacetylase, and acetaldehyde dehydrogenase and they found the production of lactate has improved to 10.46 whereas in the wild type it was 0. And then they remove another set of reaction and then they found that the production of lactate has improved from 10.46 to 80.

And further they have removed reaction and it has not improved significantly, but at the same time it is compensated by the biomass. So, we can see that biomass is reducing to actually improve the production of the chemical that you are looking for. And these on the parallel you can see these are the MOMA production, how much MOMA is predicting based on this gene deletion is also shown over here.

(Refer Slide Time: 31:07)

OptKnock Problem Statement

maximize $v_{chemical}$ ✓ (OptKnock)

subject to maximize $v_{biomass}$ (Primal)

subject to $\sum_{j=1}^M S_{ij} v_j = 0$

$v_{in} + v_{pk} = v_{pk_output}$

$v_{in} \geq v_{pk_output}$

$v_{biomass} \geq v_{output}$

$v_j^{min} \cdot y_j \leq v_j \leq v_j^{max} \cdot y_j, \forall j \in \mathcal{R}$

$y_j \in \{0, 1\}$ ✓ $\forall j \in \mathcal{R}$

$\sum_{j \in \mathcal{M}} (1 - y_j) \leq K$

Cells have to grow

If a reaction (j) is removed, set $y_j=0$ so that v_j has to equal 0.

Specify the maximum number of reactions reactions you want to delete, this is K.

To solve this problem, you transform it by using the dual constraints for the primal problem, in addition to the primal constraints

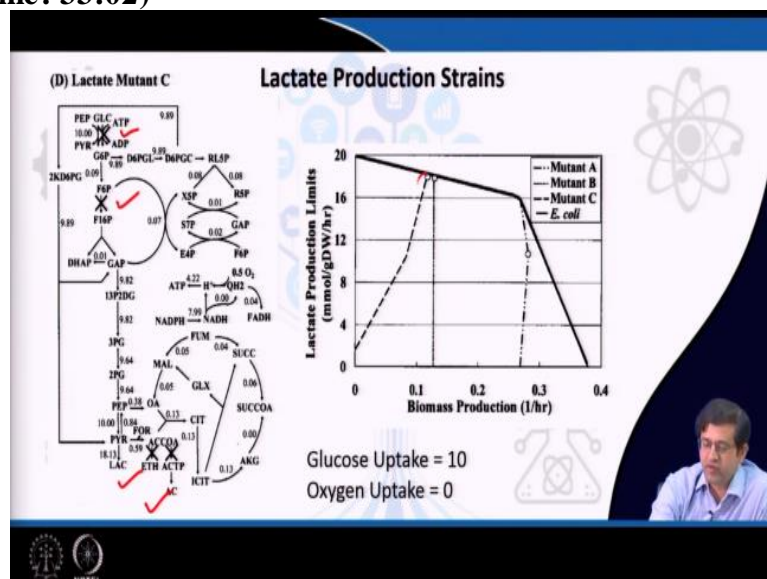
NPTEL

So, the optknock problem is actually is a maximization of the chemical subject to maximize the biomass. So, you maximize chemical at the same time, you actually delete some of the genes. So, these are the genes you want to delete. So, the cell you give a condition that this target biomass would be greater than the biomass you want to target so, we can provide that target biomass.

So, in that biomass range, it will calculate the number of gene deletion. So, if the reaction is to remove then y_j will be 0. So that v_j has to be equal to 0. So, this way, y_j is actually the binary number that is goes from 0 to 1. So, the binary number and y_j is actually 0, that means that reaction is actually removed. So, this way, you can check for all the reaction which combinations of the gene deletions are actually helpful for improving the chemical production.

And also specify the maximum number of reactions that need to be deleted that is given by K . So, first, you allow the cell to grow that is giving a target biomass and then specify the number of reaction that you want to remove from the network, which is given by K . So, to solve this problem, you actually need to transform it is because this is a bilevel optimization, which you have to use the dual constraint from the primal problem. So, in addition to the primal constraints, so, there is algorithm which actually take care of this bilevel problem can be addressed by considering dual constraint.

(Refer Slide Time: 33:02)

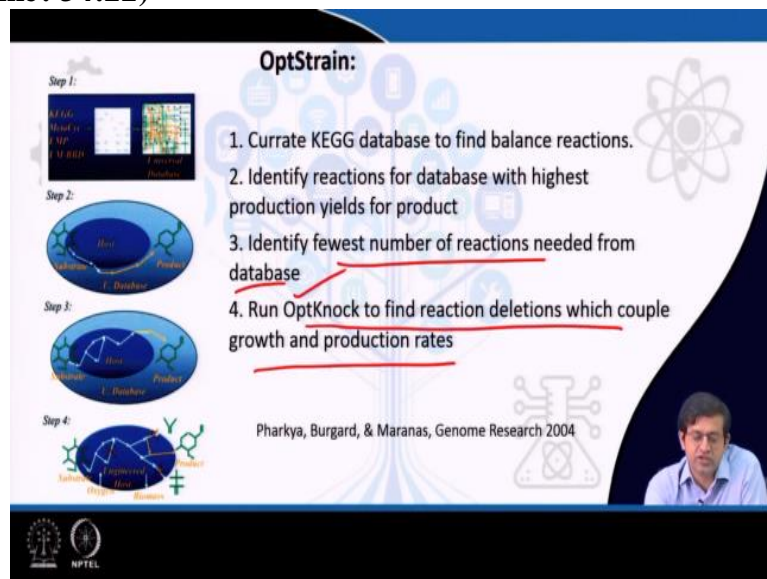


So, this succinate production strain, you can see that the mutant A, B, C they have produced mutant A, B, C and this mutant A the production is actually less and then mutant B. their the

production has improved and so on then they found that these are the modifications they have done in the network. This is one knockout, this is another knockout and then they removed the competing pathways for acetate they have removed in order to improve the production of succinate.

So, the 1, 2, 3; 3 genes they have removed 3 reactions, they have removed in order to improve the production of succinate the glucose uptake rate is 10 and the oxygen uptake rate is 0. And then in the lactate production, they removed 1 reaction that is from glucose to glucose 6 phosphate and then there will be competing pathway that is ethanol and acetate they and then also the remove fructose 6 phosphate. 1, 2, 3, 4; 4 reactions they have removed. So they keep on increasing the number of reaction, deletion of reaction and then they found this is the point while they get the maximum lactate.

(Refer Slide Time: 34:22)



The diagram illustrates the OptStrain process in four steps:

- Step 1:** A screenshot of a software interface showing a list of reactions and their yields.
- Step 2:** A metabolic map showing a substrate being converted to a product through a series of reactions.
- Step 3:** A metabolic map similar to Step 2, but with some reactions highlighted in red, indicating they have been selected or modified.
- Step 4:** A metabolic map showing the final optimized pathway with specific reactions highlighted in red.

OptStrain:

1. Currate KEGG database to find balance reactions.
2. Identify reactions for database with highest production yields for product
3. Identify fewest number of reactions needed from database
4. Run OptKnock to find reaction deletions which couple growth and production rates

Pharkya, Burgard, & Maranas, Genome Research 2004

The slide also features a small video inset of a man in the bottom right corner and the NPTEL logo in the bottom left corner.

So opstrain is another algorithm, where you can design the strain. First you get the KEGG database to find balance reaction, and identify the reaction from database with highest production of yield. In this case in optstrain, you add gene rather than in optknock, you have seen that you have removed the gene or remove the reaction but here you add the reaction from the database.

Identify the fewest number of reactions that is needed from the database and you add it and then you run optknock to find the reaction deletion which couple both growth and production rate. So, here also you can apply optknock algorithm to find in a reaction deletion, which could couple the production rate of the chemical that you are looking for, but you first you

have to add the reaction. The fewest number of reactions that need to be added from the database, and then you run the optknock again to find the reaction which needs to be removed and for improvement of the production rate.

(Refer Slide Time: 35:26)

CONCLUSION

- Solution spaces can be studied by randomly sampling points contained within them
- Flux sampling generates a sequence of feasible solutions (called a chain) that satisfy the network constraints, until the entire solution space is analyzed.
- It can be run with or without a specified objective function and can provide a range of feasible solutions space
- The dynamic metabolic properties can be estimated by capturing all the feasible solutions of metabolic model in a condition specific manner.
- Identifying such key parameters may be useful in directing experimental efforts to select critical parameters that must be measured with precision, as well as determining the most sensitive constraints for computational modeling and pathway simplification.

NPTEL

So, in summary, we learned about the solution space that can be studied by randomly sampling points contained within them the solution space, you need to be randomly sample and then found the random sample what you get is basically the histogram of fluxes and that the probability of the fluxes can be calculated and the flux sample is generally a sequence of feasible solution called a chain that satisfy the network constraint until the entire solution space is analyzed.

So actually this is a chain of feasible solution that satisfy the network constraint. So, all those points which are generated actually obey the network constraint, and when you sample those points, those points are actually satisfying the metabolic network until the entire solution space is analyzed, it will resampling and will keep on running. So, sometimes it takes a lot of time to sample the solution space.

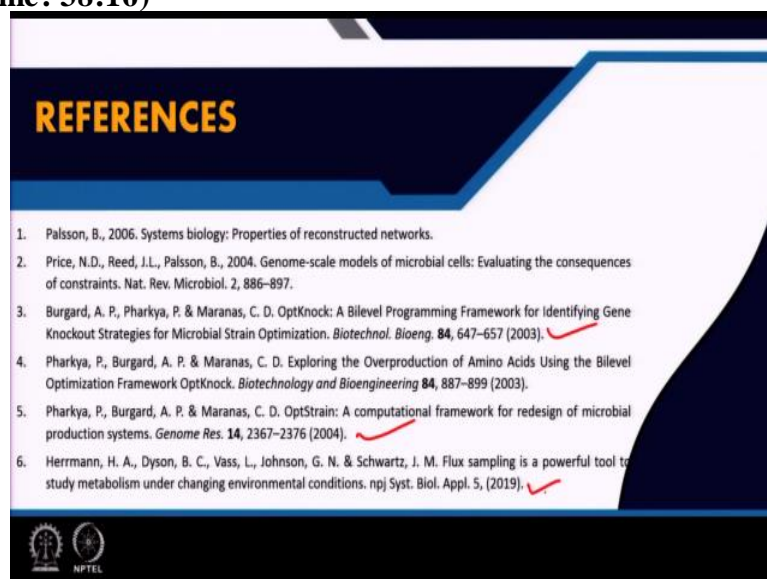
So, you have to wait for several hours to sample your network depending on the size of the network if it is a very big network, then it takes several hours to sample the solution space, it can be run with or without specified objective function and can provide a range of feasible solution space. So it can be run without an objective function, it is the beauty of sampling. So, because the microorganism you are good with objective function of growth or biomass.

But if you come to mammalian cell then you cannot consider the biomass as the objective function. Because multicellular organism, the mammalian cells, you may or may not be good to actually consider the objective function as biomass. Because the cell does not grow after some time it does not grow. So, that is the problem with the multicellular organism. So there the multicellular sampling is one of the most powerful tools to actually get the fluxes.

So, there are dynamic metabolic properties can be estimated by capturing the feasible solutions of metabolic model in a condition specific manner. The metabolic properties can be estimated by capturing all feasible solutions in a conditioned space. Because in the previous example, where the photosynthetic organism which I considered, in that they have used it control and also a cold condition.

So, 2 different conditions they have calculate the flux solution and they try to make a difference and identify see such key parameters may be useful in and directing experimental efforts. So, this important parameter that you are identifying from this solution space can be used to design the experiment or to select the parameter that must be measured with precision. And as well as determining the most sensitive constraint for computational modeling and pathway simplification. This way, you can determine the most sensitive constraint of the computational modeling.

(Refer Slide Time: 38:16)



So, these are the references you can follow you can further read about the optknock algorithm. The bilevel algorithm which is given in this paper and then we have other references of strain differences given over here. And then the photosynthetic organism why

did they calculate the robustness of the photosynthetic organism by flux sampling is also given here. You can read in more detail. Thank you, thank you for listening will meet you on next lecture.