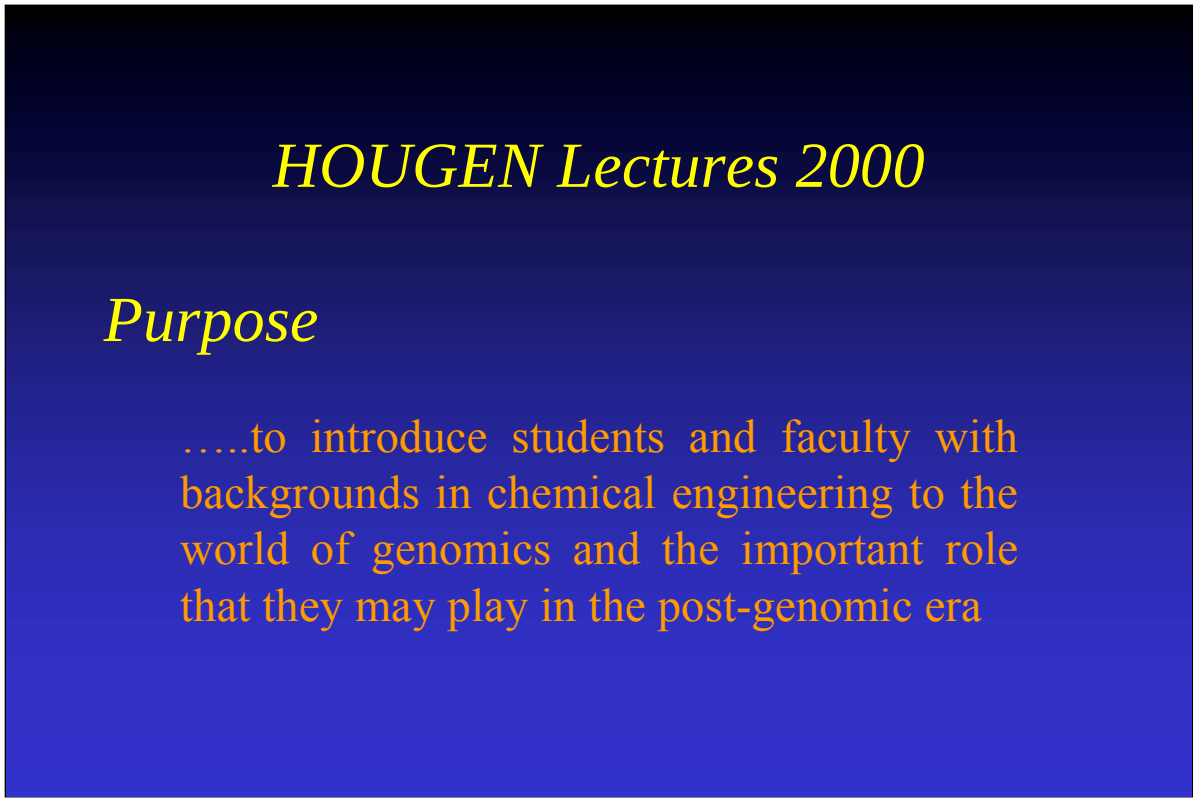
THE HOUGEN 2000 LECTURES

Olaf Hougen
http://www.engr.wisc.edu/che

Bernhard Palsson
http://gcrg.ucsd.edu

The Olaf A. Hougen Professorship in Chemical Engineering is funded by the Hougen Professorship Fund of the University of Wisconsin Foundation. Colleagues and former students of Professor Hougen, other friends, and corporations have contributed to the fund to honor one of the founders of the modern chemical engineering profession. The 2000 award to Bernhard Palsson continues a tradition of providing outstanding individuals with the opportunity, through visiting appointments, to advance chemical engineering by exercising their creative abilities in the congenial and stimulating environment at the University of Wisconsin-Madison.

Bernhard O. Palsson is a Professor of Bioengineering and Adjunct Professor of Medicine at the University of California, San Diego. Professor Palsson is the author of over 140 peer reviewed scientific articles and 18 U.S. patents, many of which are in the area of stem cell transplantation, cell culture technology, bioreactor design, gene transfer, and metabolic engineering. He received his Ph.D. from the University of Wisconsin–Madison Department of Chemical Engineering in 1984. He sits on the editorial boards of several leading peer-reviewed bioengineering and biotechnology journals. Professor Palsson held a faculty position at the University of Michigan for 11 years from 1984 to 1995. He received an Institute of International Education Fellowship in 1977, a Rotary Fellowship in 1979, and a NATO fellowship in 1984. He was named the G.G. Brown Associate Professor at Michigan in 1989, a Fulbright Fellow in 1995, and an Ib Henriksen Fellow in 1996. His current research at UCSD focuses on the construction of genome-scale models of cellular metabolism, and on stem cell fate processes.



PURPOSE

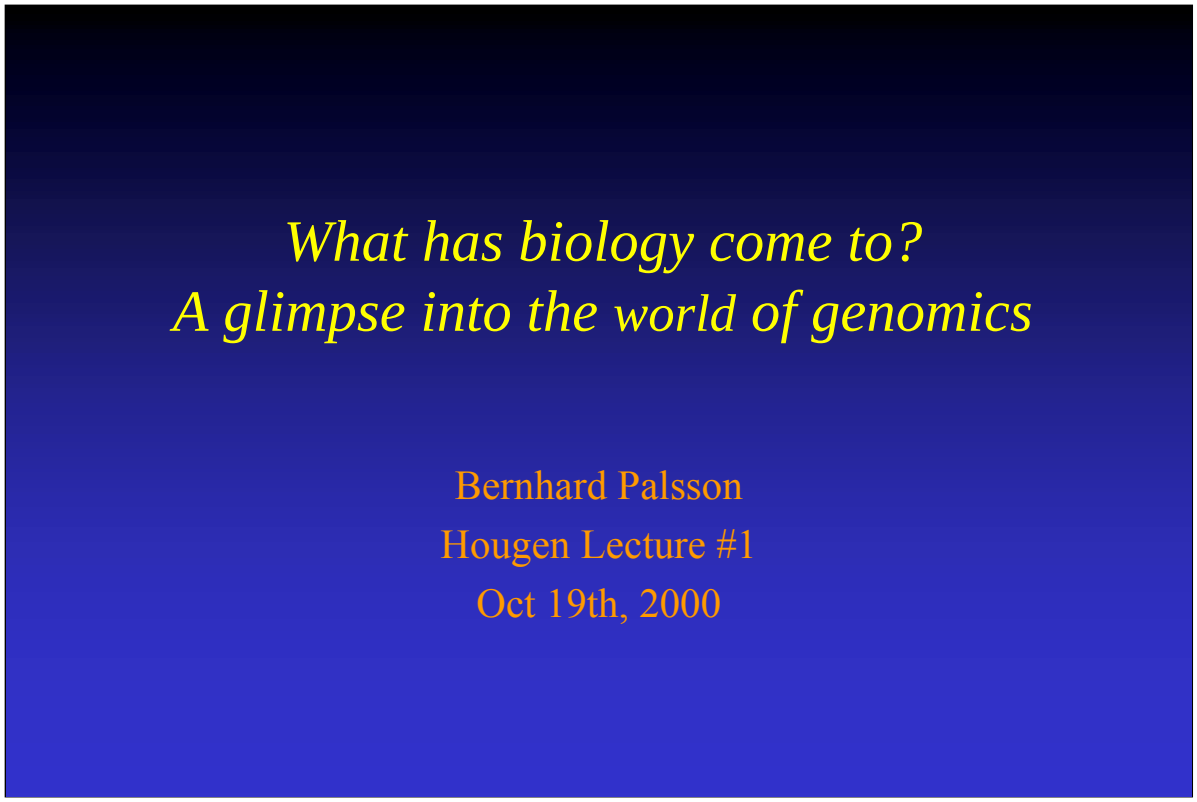
The Hougen visiting professorship was established to enable scholarly and free exchange amongst the visitor, the faculty, and students of chemical engineering. The Wisconsin department has always placed emphasis on fundamental issues and problems that have long-term consequences. It is the opinion of this year's Hougen professor that developments in the post-genomic era will depend heavily on the subjects that are emphasized in the Chemical Engineering Curriculum. Thus, if properly motivated and oriented, Chemical Engineering as a discipline may play a significant role in the historic developments that lie ahead. The purpose of this series of lectures is to illustrate these issues to faculty and students that have a chemical engineering type background.

Tentative Schedule

• October 19th	#1	"Where has biology come to? a glimpse in to the world of genomics"
• October 26th	#2	"Cellular part catalogs; reconstructing biochemical reaction networks"
• November 2nd	#3	"Modeling philosophy: Of single points and solution spaces"
• November 9st	#4	"Operating systems of genomes; Systemically defined pathways"
• November 21th	#5	"Closing the flux cone: imposition of maximum capacities"
• November 30th	#6	"The biological design variables: kinetic and regulatory constraints"
• December 7th	#7	"Entrepreneurship"

SCHEDULE

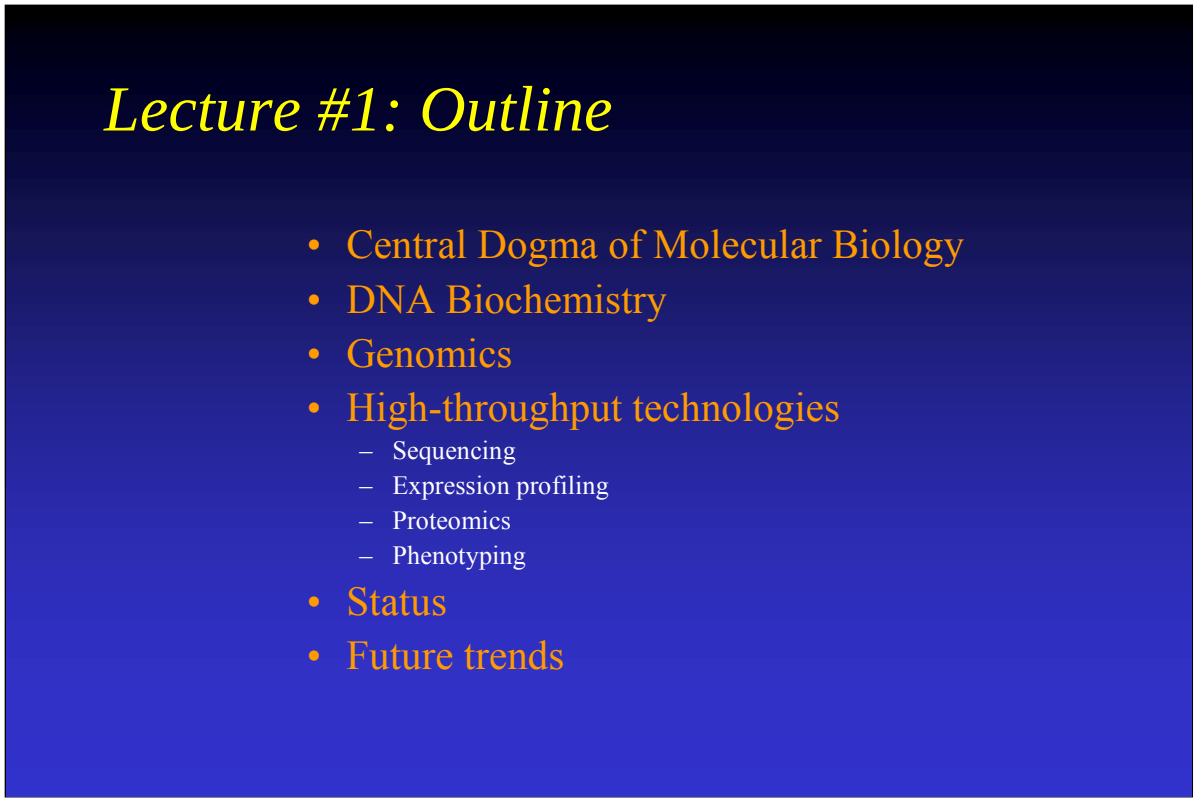
The lectures will be delivered in a very casual setting over lunch on Thursdays. The tentative schedule of topics is given above. We expect that this outline will evolve as the lectures proceed in response to the interest and the expertise of the audience that will attend. An extra time slot on December 7th is included in case more time is need to satisfy higher than anticipated interest in this topic.



INTRODUCTION

High-throughput experimental technologies have been developed to simultaneously analyze a myriad of cellular components. As a result, biology is undergoing a 'phase change' from the classical pure 'in vivo' biology to biology that takes place in a computer, or 'in silico.' This series of lectures will address some of the important issues that are associated with this change and try to illustrate what is to come.

These slides and their accompanying text have been updated since they were presented in the Fall of 2000, and their official publication date is July 1, 2001.

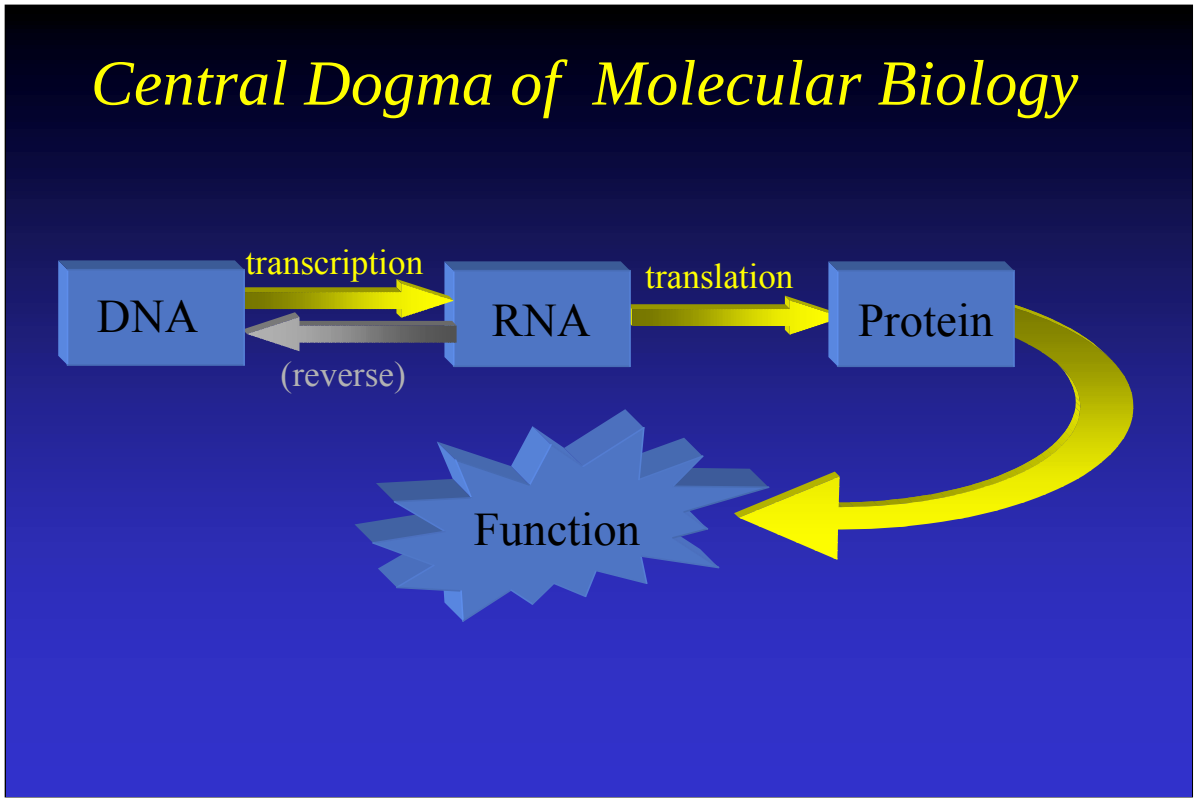


LECTURE #1

This series of lectures will begin with a very brief background on DNA, its biochemistry, and its central role in biology. Then we will introduce the relatively new and rapidly emerging field of genomics. Most of the time will be spent on the impressive high-throughput technologies that have been developed to enable this field and that continue to drive it on.

It should be self-evident to the engineering audience that this field is technology driven, and thus a natural subject for engineering. These technologies are essentially based on automation, miniaturization, and multi-plexing.

The massive amount of ever cheaper and accurate biological information that is resulting from these technologies demand the development of an associated IT infrastructure (collectively called bioinformatics) and mathematical modeling and computer simulation capabilities (currently being referred to as in silico biology).



THE CENTRAL DOGMA

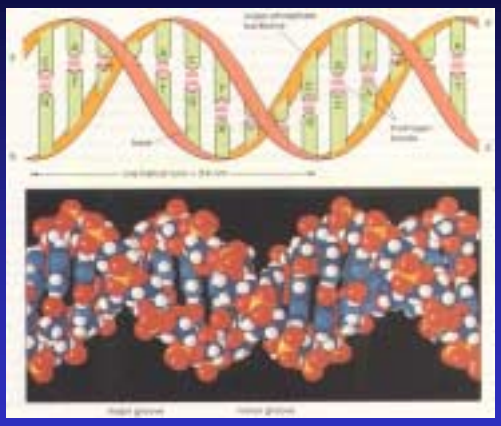
This schema illustrates the central dogma of molecular biology as it was developed about 40 years ago. The DNA, a long thread like molecule of a specific base-pair sequence, carries the inherited information. Short segments of the DNA molecule (called the open reading frames or ORFs) are transcribed into a chemical relative, RNA, in the form of a message. This message is then translated into protein, that in turn carry out individual biochemical functions in the cell.

This dogma has been around for many decades. So what is new? What is new is the fact that we can now characterize the entire DNA molecule(s) of an organisms in detail, measure all the messages coming from the DNA at any given time, and assay for all the different protein molecules in a cell.

This central dogma is now expanding and being revised. No protein functions in isolation, but participates in multi-geneic functions that comprise cellular physiological behavior. This dogma is about to be revised and extended by the elucidation of the networks that the proteins form and their quantitative systemic characterization.

DNA: Structure, discovery, sequencing

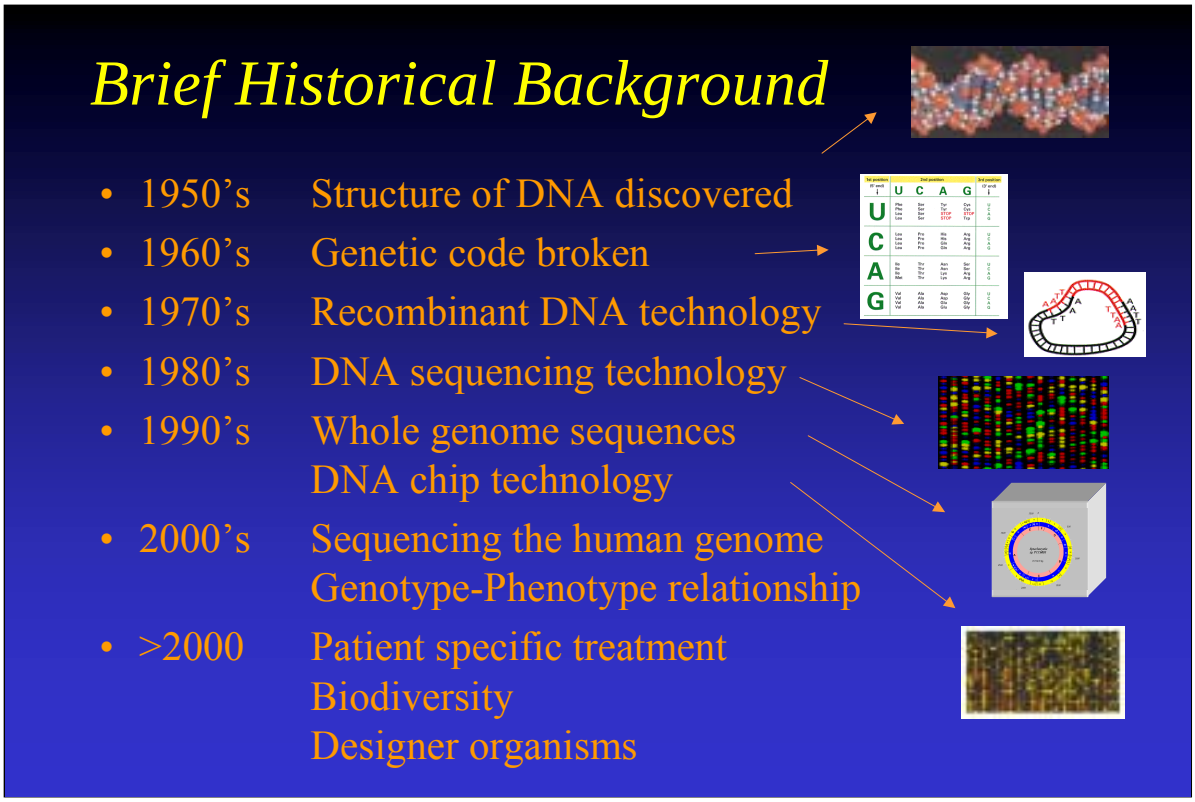
- What is DNA?
  - a linear polymer of nucleotides
  - DNA exists as a molecule of 2 anti-parallel strands that are complementary in their nucleotide sequence.



THE DNA MOLECULE

The DNA molecule is basically a linear atactic polymer of monomers, that are called nucleotides. There are four nucleotides, denoted by A,T,C,G. A complimentary strand can be synthesized based on the A:T and G:C base-pairing. If two strands are complementary they form a double helix with anti-parallel strands.

The length of DNA molecule(s) in living beings varies, but is on the order of 1000 to 10,000 for viruses, a few million for bacterial, a few hundred million for simple multicellular organisms and a few billion for mammals such as the human. It has just become possible to obtain the sequence for the entire set of DNA molecules in complex eukaryotes. There are several such molecules, called chromosomes, in animal cells. In humans there are 23 chromosomes, and every somatic cell carries two sets of each chromosome, one from each parent.



SOME HISTORICAL MILESTONES

The technologies used to study DNA and our knowledge of DNA has grown substantially since the discovery of its structure by Watson and Crick about half a century ago. This slide has just a few of the highlights of this history.

The coding of information on the DNA was broken in the 1960's, the first recombinant DNA was made in 1973, the 1980's saw the development of automated sequencing technology. The 1990's brought the development of DNA chip technology, and the sequencing of entire genomes. And in the new millennium, we have the human DNA sequence virtually completed and are seeing the emergence of quantitative study of the all important genotype-phenotype relationships. There are many milestones omitted in this list, with PCR being perhaps the most prominent omission.

In the coming decades we can expect a rapid continuation of these developments. Although these are hard to forecast, it seems clear that we will develop patient specific treatments that are based on one's particular genotype, study and preservation of 'ecological' genomes, and the design of organisms from scratch.

Genomics: the science of complete genomes

"The complete set of instructions for making an organism is called its genome. Constructed of DNA, the genome contains the master blueprint for all cellular structures and activities for the lifetime of the cell or organism. It orchestrates life from simple bacteria to remarkably complex human beings. Understanding how DNA performs this function requires knowledge of its structure and organization."

- **genome sequencing and assembly**
- **comparative genomics**
- **functional genomics**
- **structural biochemistry**
- **molecular evolution**

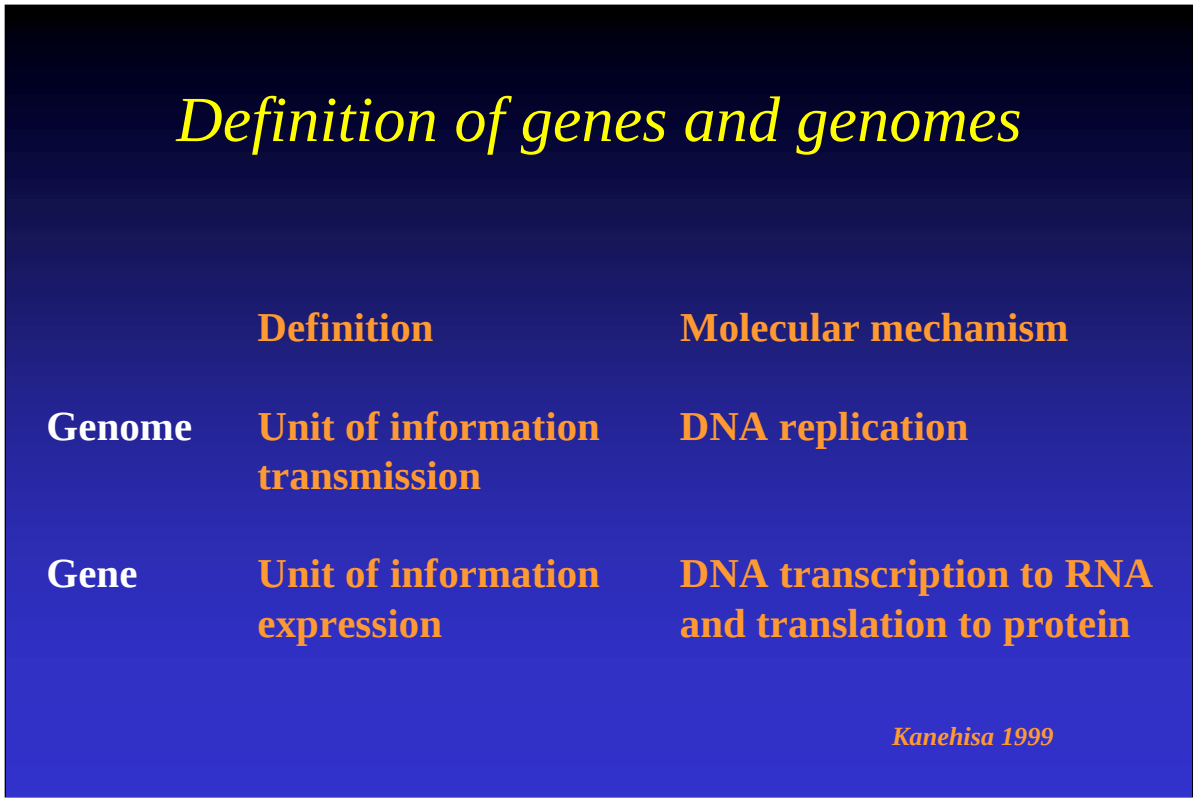
General Genomics Information:

- Genomics: A global resource (www.phrma.org/genomics)
- Primer on Molecular Genetics, DOE (www.bis.med.jhmi.edu/Dan/DOE/intro.html)

GENOMICS

The ability to sequence the entire DNA of an organism has given rise to the field of genomics. The word is a combination of gene and -ome, the latter meaning 'whole.' Thus genomics are the study of the entire composition of the genetic instruction and capabilities that are contained on the chromosomes from a particular cell.

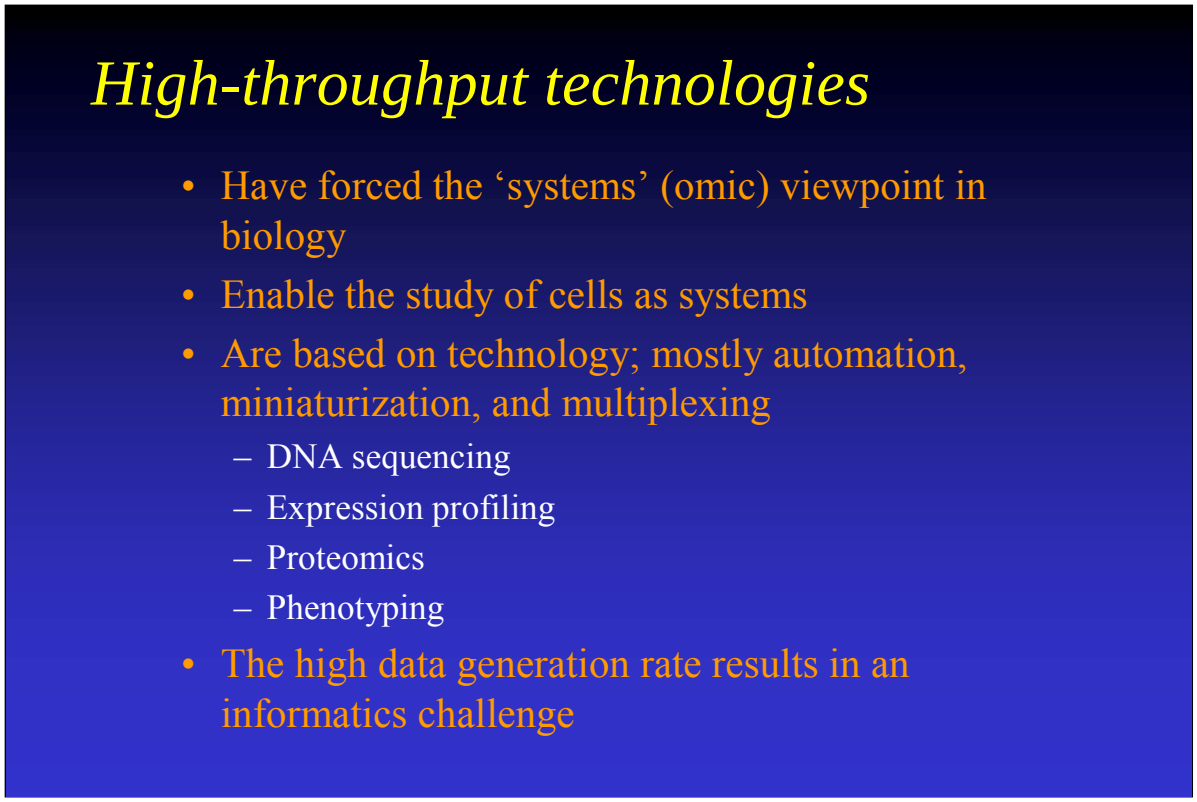
Other 'omics' words are proteome, transcriptome, metabolome, physiome, and phenome, with their obvious meanings.



GENES AND GENOMES

Every gene carries the information that needs to be first transcribed and then translated, per the central dogma, and it represents a unit of information expression.

Genomes, on the other hand, when replicated carry a ‘unit’ of information transmission for a new cell.



HIGH THROUGHPUT TECHNOLOGIES

Several types of high-throughput approaches to the genome-scale analysis of cellular components have been developed. These include sequencing methods that will yield the entire base pair sequence of the genome, DNA chips that allow the analysis of all the mRNA in a cell, and proteomic methods that yield information about the protein portfolio of a cell. Currently, we are seeing rapid developments of cell-based high throughput screening methods that basically amount to high-throughput phenotyping, or allowing us to determine how cells behave under defined circumstances. These methods may eventually remove the 'green thumb' from biology since they are allowing for quantitative and detailed measurements of cellular components and cellular behavior.

The challenges of managing all this information has lead to the rise of bioinformatics.

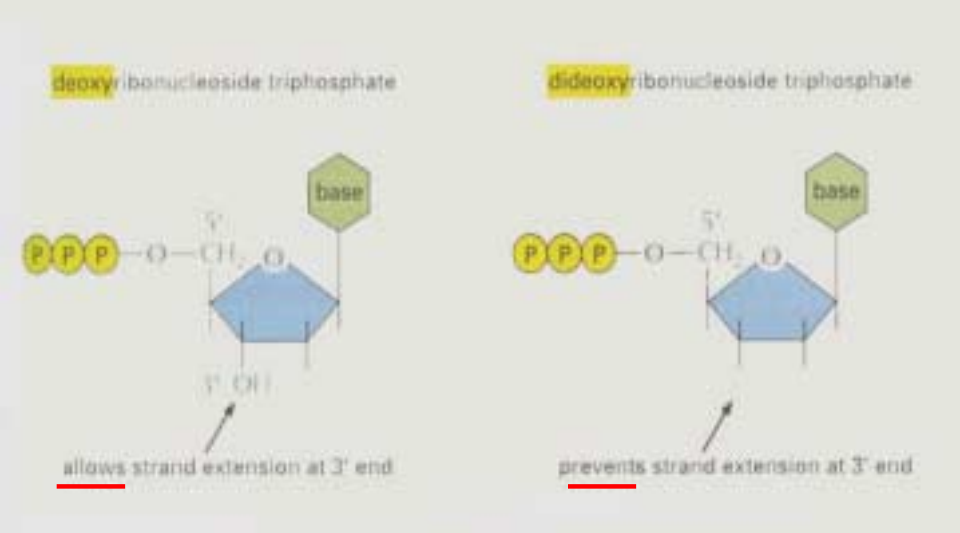
DNA Sequencing

DNA Sequence ACTGTCGAACTGGACTTCAGCTTGATCGGAACGTCAATCGACTACGTAGTCAT

- There are traditionally two different approaches to sequencing DNA.
  - Chemical method
  - Enzymatic method
- The enzymatic method has become the standard procedure for sequencing DNA
- Newer methods are being developed (i.e. DNA chips)

As most of you are aware, the technology exists to completely sequence an entire genome.

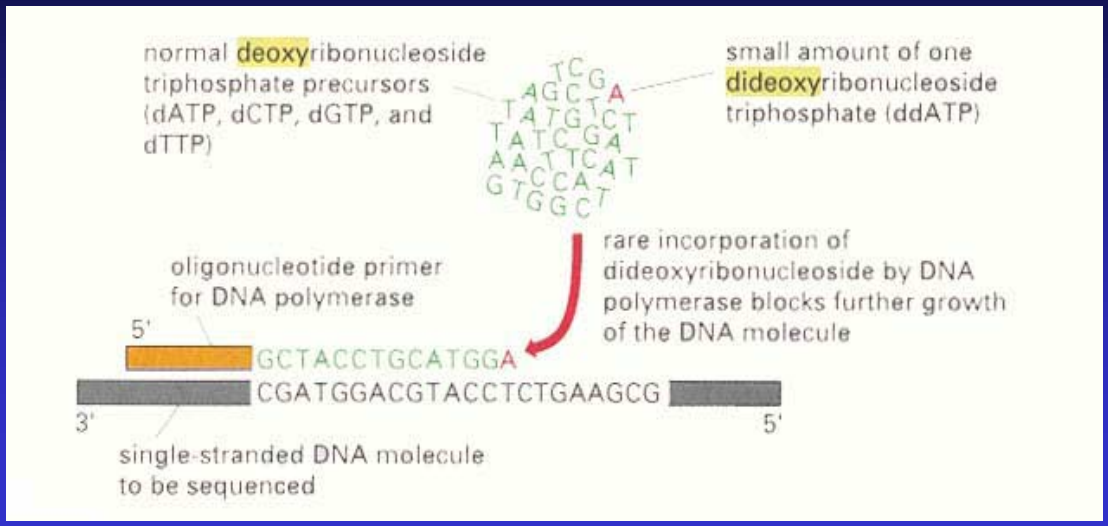
Dideoxy-nucleotides will stop DNA polymerization: they are terminators of polymerization



CHAIN TERMINATION

Nucleic acids are polymers of pentoses tied together with a phosphate diester bond. A base is attached to each pentose giving the sequence specificity. The OH group on the 3' end (third carbon of the pentose) binds the to the 5' (fifth carbon) end via the di phospo-ester bond. Thus a dideoxy- form of the pentose would terminate the polymerization.

Trace amounts of dideoxy-nucleotides will stop DNA synthesis at a defined location



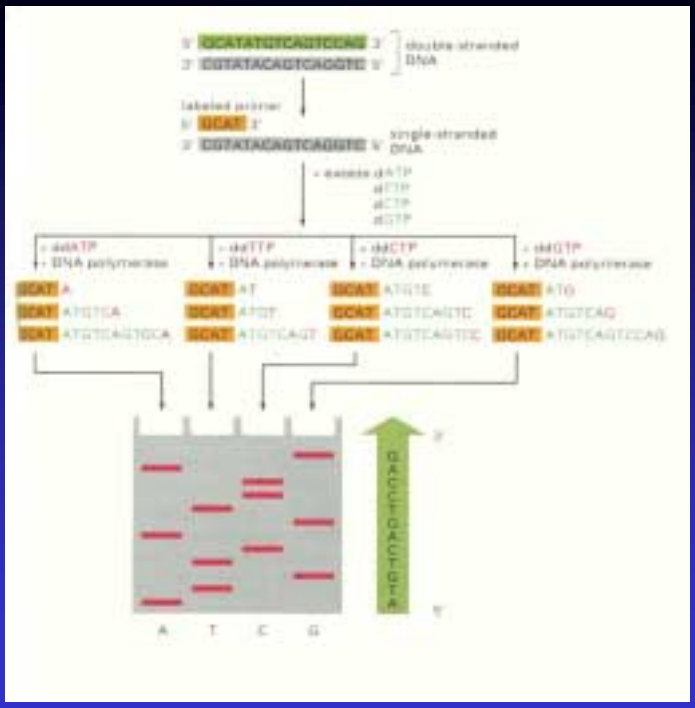
CHAIN TERMINATION

Small amount of a dideoxy form of one of the nucleoside tri-phophates would thus terminate a polymerization reaction in a well defined location. This example shows that a trace amount of ddATP would terminate the reaction at a T base of the original template.

Four mixtures with ddNTP can be used to polymerize from a primer.

Then run each mixture on a size fractionating gel.

Align and call bases to form sequence



A FOUR REACTION PRODUCT CAN BE SIZE FRACTIONATED ON A GEL

Four different reaction mixtures each with a trace amount of a different dideoxy-nucleoside will form a series of fragments each with a defined end. If run on a four lane gel side by side the fragments can be size separated and with the defined termination the base sequence of the original template can be determined as shown.

DNA Sequencing--lab scale

Laboratory scale DNA sequences of 300 bases can be determined in around 1 hour on a standard 32cm gel.

With a triplet codon this allows for the sequencing of a gene that encodes a 100 aa protein.

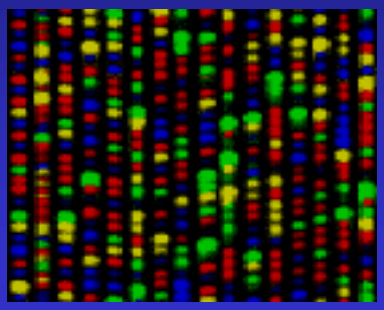
For comparison the alpha and beta chains of hemoglobin are about 150 aa.

LAB SCALE SEQUENCING

DNA sequencing--large scale

A ABI Prism 377XL automated DNA sequencer is capable of:

- running 32 (96) templates simultaneously,
- yielding between 250-400 bases per template,
- run times of 7 to 8 hours allow two to three runs a day,
- yielding a potential 75 kb (*3 =225) of raw sequence per day.



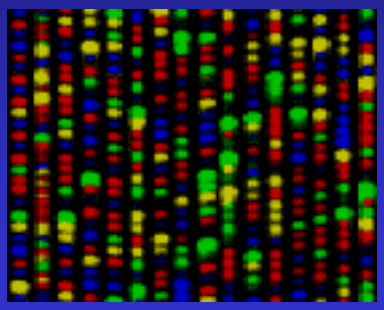
AUTOMATED SEQUENCING

The basic DNA sequencing procedure can be automated. Such developments began in the late ‘80s and by the early to mid-90s, the ABI 277 automated sequencer was made available. Some of the performance characteristics are shown on this slide. Note that high-throughput is achieved by multi-plexing, i.e. running more and more lanes in parallel. Miniaturization of lanes is limited and one sample can slide over one lane, causing a serious error with the automated base calling software. Such lane slides were eliminated with the capillary type sequencer since each sample is physically confined.

DNA sequencing--large scale

Some technical features:

- Slab or capillary gel electrophoresis,
- Laser excitation of fluorescent dyes,
- CCD camera/confocal microscope detection,
- Automated data collection and base calling.



THE TECHNOLOGICAL UNDERPININGS OF AUTOMATED SEQUENCING:

Some of the basic technologies used in automated DNA sequencing are shown in this slide

1. Size separation of fragments
2. Fluorescent probes and laser based activation for signal generation
3. Signal detection using a CCD camera and a confocal microscope
4. Software for automated base calling. This feature turned out to be very important as the large data volumes being generated created a serious informatics challenge

Sequence Databases

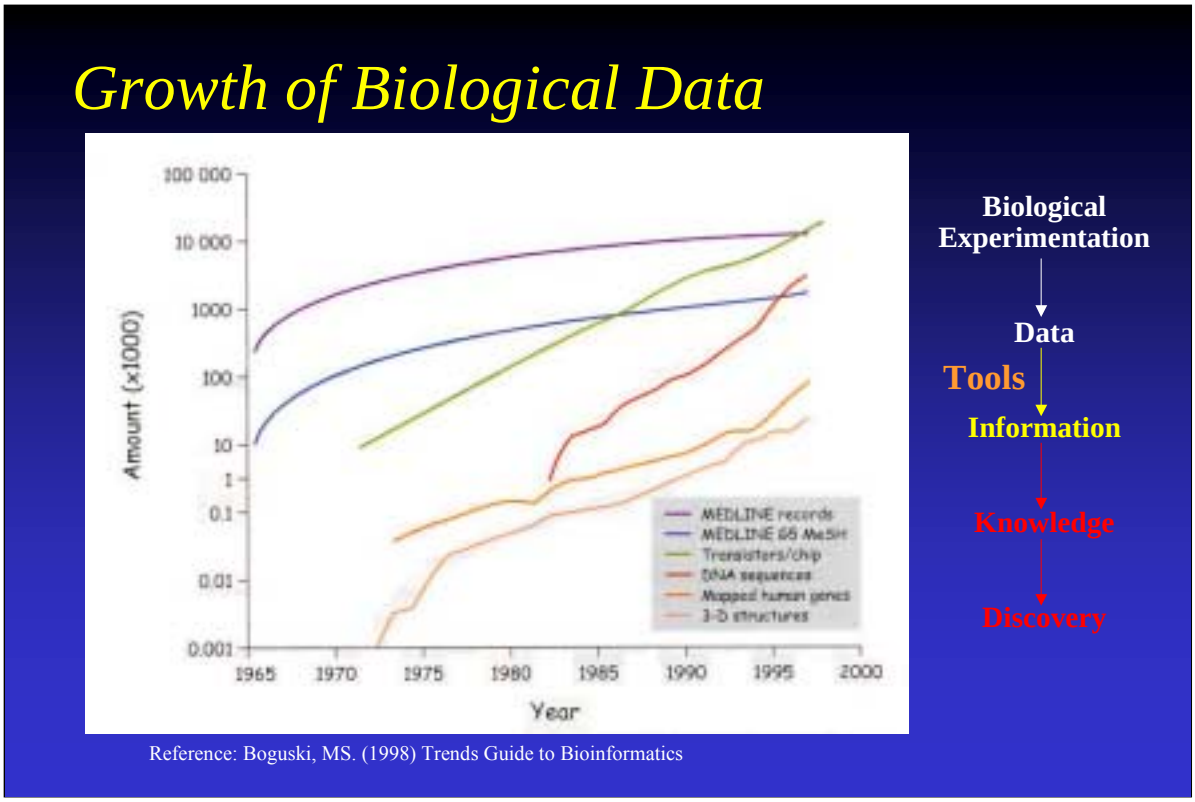
DNASYSTEM (www-biology.ucsd.edu/others/dsmith/dnasys.html) Doug Smith, UCSD Biology Dept.
Provides brief descriptions and links to most bioinformatic databases and web sites

Primary Databases - databases tend to be 'archival', data is submitted with little or no addition of information

- Genbank (NCBI/USA) DNA
- EMBL (EMBO/Europe) DNA
- GSDB (NCGR, USA) genomic DNA
- PIR/NBRF (USA) Protein
- SWISS-PROT (Switzerland) Protein
- PDB (BNL, USA) 3D structure

Secondary Databases - specialized databases with large amounts of additional annotation

- TIGR Microbial Database (Genome sequencing projects and results)
- OMIM (Online Mendelian Inheritance in Man, gene and clinical data)
- KEGG (Kyoto Encyclopedia of Genes and Genomes, metabolic info)
- EcoCyc, HinCyc (E.coli and H. influenzae metabolic databases)

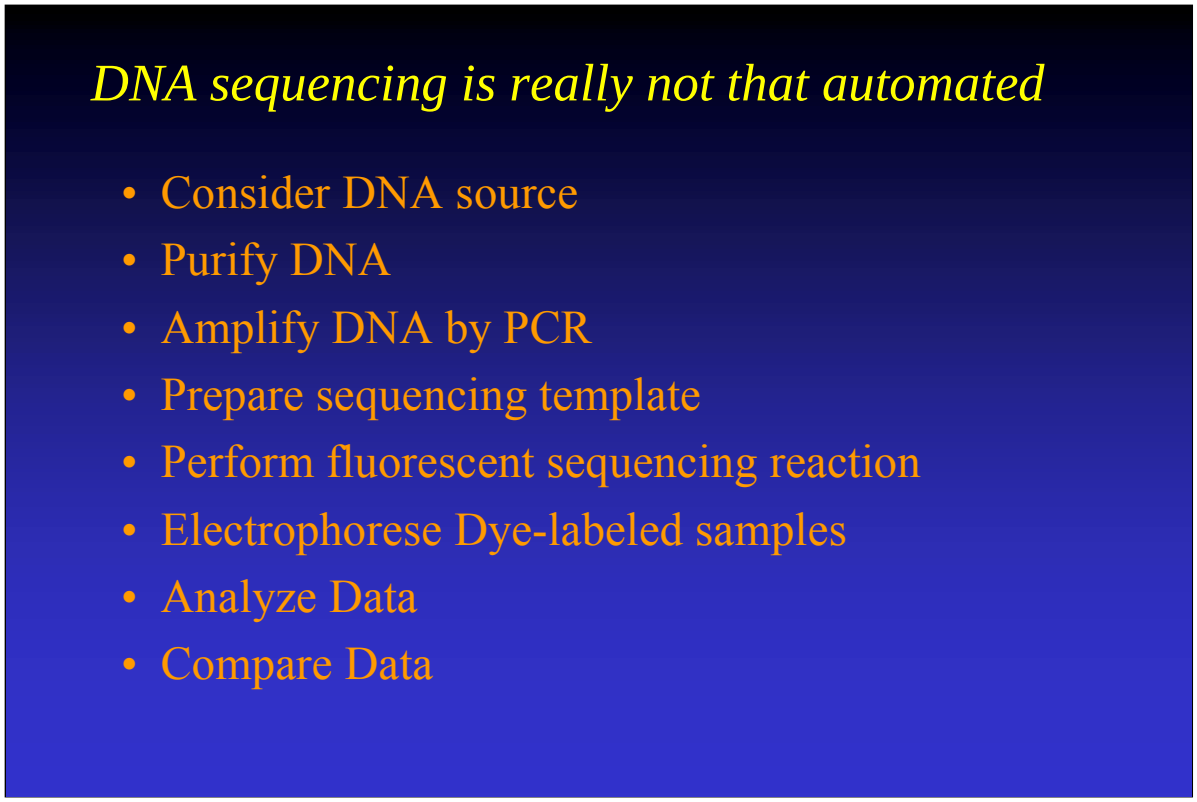


GROWTH OF BIOLOGICAL DATA

This graph, from a special issue of Trends Guide to Bioinformatics in 1998, illustrates the rapid growth of biological information. The size of Genbank (The NIH genetic sequence database) represented by the red line, has been doubling every 18-24 months, and housed over 3 million sequences in 1998. The data generation by high-throughput experimental technologies appears to follow Moore's law, that is doubling approximately every 18 months.

Therefore, we expect that we will soon not be limited by the availability of data, but by our lack of tools available to analyze and interpret this data to generate knowledge and leading to scientific discovery.

The total number of references in the Medline database with headings to molecular biology or genetics, is shown by the blue line and they are not growing at the same rate. This difference has lead some to conclude that less and less knowledge and insight is being generated per unit of information generated. Some are thus boldly claiming that we need to devise ways to increase the knowledge derived from all this information.

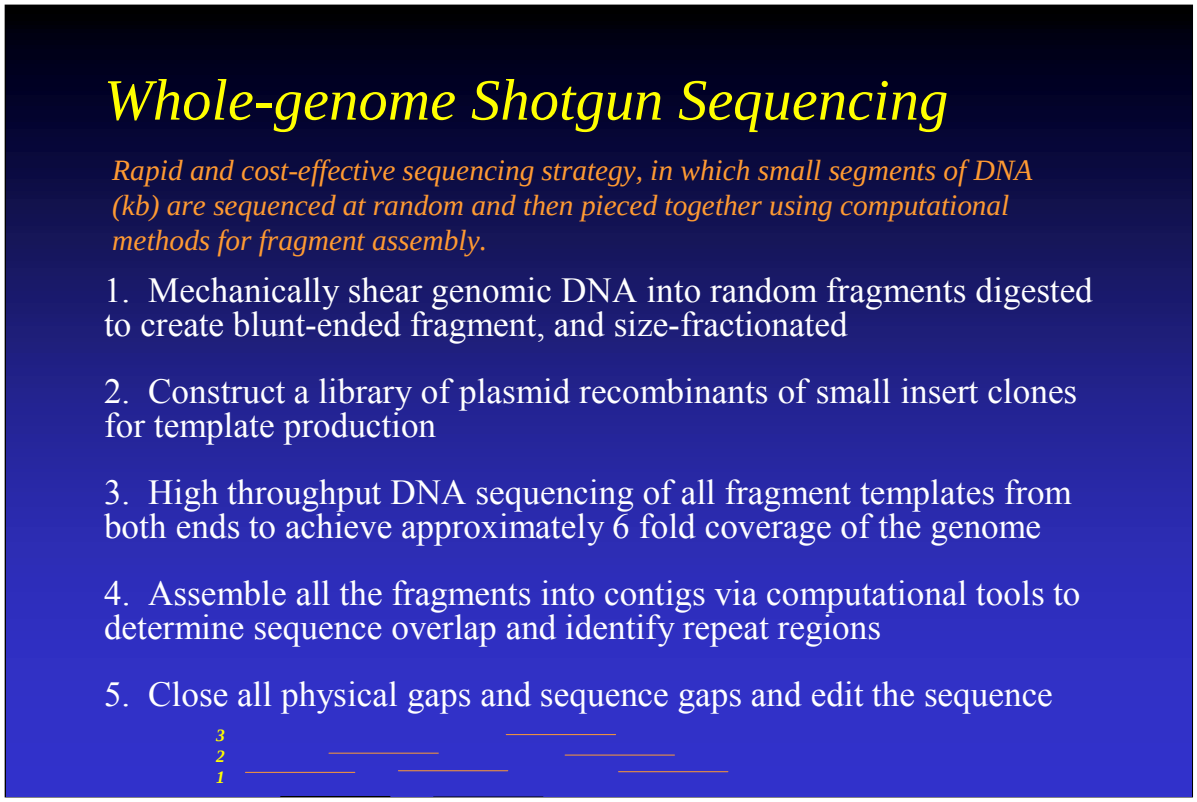


AUTOMATION

Although DNA sequencing has advanced to the stage that it allows for the sequencing of entire genomes, there are still many manual and laborious steps involved in the process. We can anticipate great strides in the full automation of this process and its integration to achieve greater efficiency in the sequencing process.

The cost and availability of sequence data is expected to improve greatly in the near future.

The impressive sequencing capabilities of Celera today, of generating 3 Giga base pairs per month may not seem too spectacular in just a few years.



SHOT GUN SEQUENCING

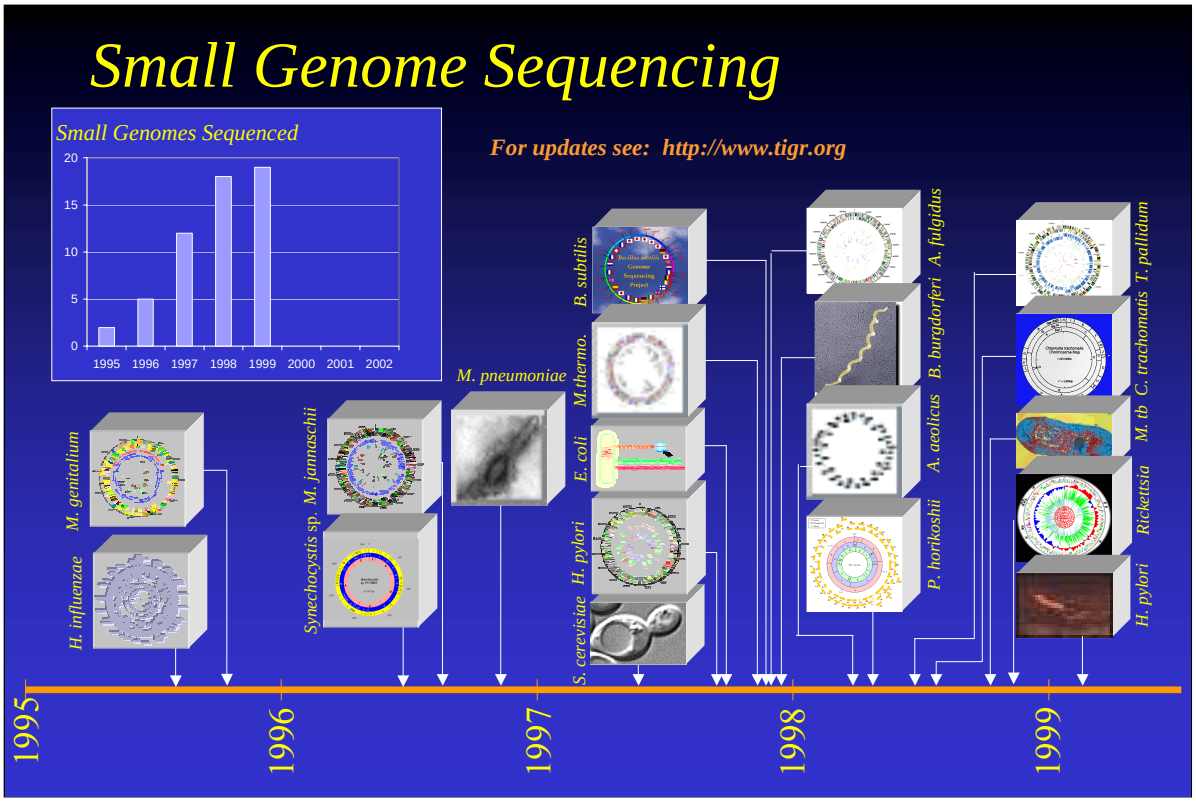
Whole genome sequencing is used to establish a full sequence. The basic idea is to randomly (mechanically) break the DNA into fragments size fractionate these fragments, capture these fragments and sequence them as described. Islands of the whole genome sequence are obtained as shown schematically at the bottom of the slide. This procedure is repeated enough times so that sufficient overlap is obtained between the fragments so that the full sequence is obtained. Due to the Poisson statistics that govern this process, six fold coverage gives more than 99% of the full sequence. The remaining gaps are then specifically sequenced to complete the whole genome sequence.

H. influenzae Sequencing Project (1995)

- Cost of $0.48/base x 1,830,137 bases ~ $880,000
- Actual sequencing took 3 months with 8 people and 14 automated sequencers
- Genome coverage was approximately 6x, thus $e^{-6} = 0.0025$ uncovered
- Estimated error rate of 1/5000 to 1/10000 ~ 0.01%
- Representative numbers:

Sequence fragments in random assembly	24,304
Total base pairs	11,631,485
Contigs	140
Genome size	1,830,137

…and in this way the first genomic sequence was obtained in 1995 for Haemophilus influenzae. Here are some of the interesting numbers associated with this project.



MORE BACTERIAL GENOMES

In the past decade, with the development of automated sequencing technologies, genome sequencing projects have been initiated in which the primary objective is to determine DNA sequences independently of gene function.

In 1995, only five years after the Human Genome Project outlined its initiatives, the first complete genome sequence of an organism (Haemophilus influenzae) was published in Science.

Today, Large-scale DNA sequencing is becoming routine, and the costs have dropped below $0.25/base pair.

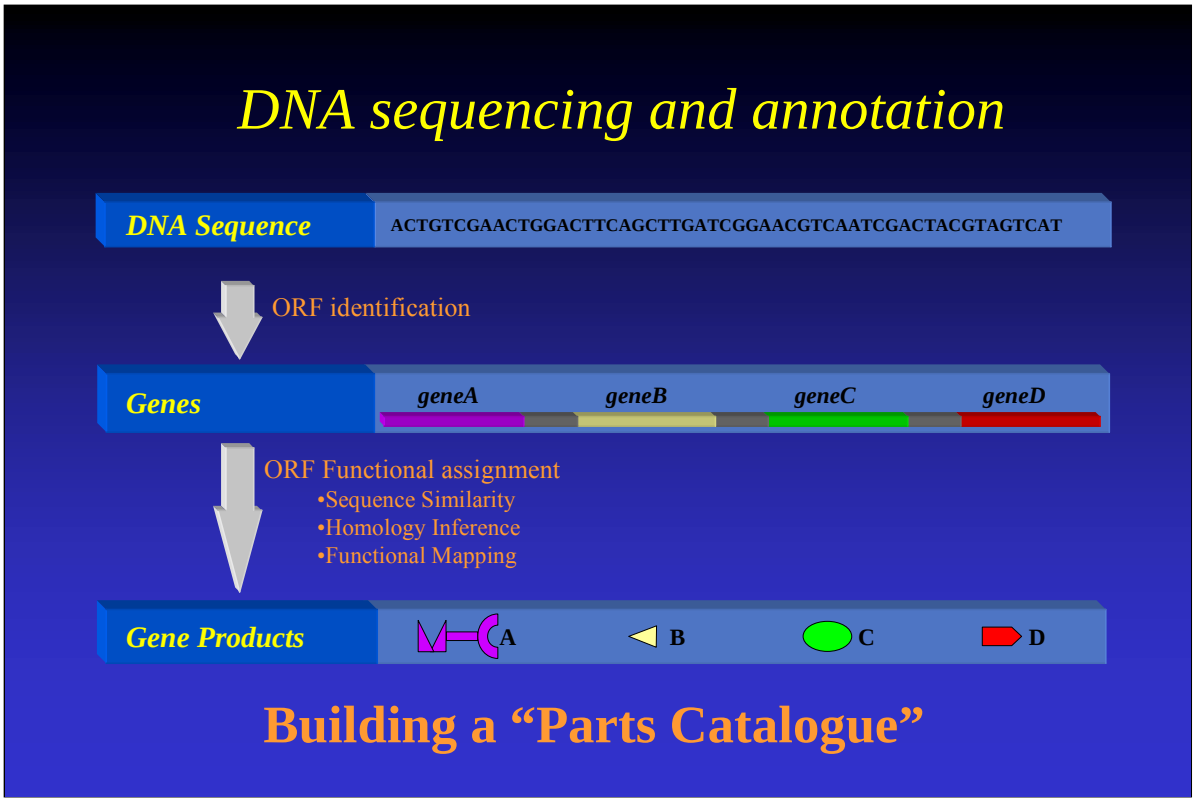
Currently, the complete genome sequence has been determined for hundreds of microorganisms (>30 in public domain), and a handful of multicellular organisms, including human, fruit fly and the nematode C. elegans. The number of these sequences is expected to grow rapidly.

From the inset, it can be seen that the number of completely sequenced genomes is growing rapidly. Many of these organisms are involved in industrial applications (*E coli* and *Bacillus subtilis*), and many are human pathogens causing ailments such as lyme disease, syphilis, tuberculosis, and ulcers.

Bioinformatics: tools for analyzing genomic data

The scientific discipline of computer-based biological information acquisition, processing, storage, distribution, analysis and interpretation.

1. information infrastructure
   - database construction and management
   - sequence databases, genome databases, organism databases
   - information retrieval/database searching
   - analytical capabilities and predictive value
2. computational-based techniques to analyze genomic data
   - sequence analysis (genome annotation, similarity searching)
   - protein function (motif identification, structural modeling)
   - genetic circuit analysis ("emergent properties")
   - new and improved analytical methods



WHAT DO WE DO WITH A SEQUNCE?

The process of going from the genetic content in the cell to cellular physiology will inevitably involve the bioinformatic analysis of genome sequence data

The genome sequencing projects basically provides the base pair sequence of all the DNA in the cell.

Algorithms have been developed to search these DNA sequences for the ORFs, (the genes or coding regions).

These genes can then be searched against databases to look for statistically significant sequence similarity, and when it exists, homology can be inferred.

With the ultimate goal, of then mapping functions of the known genes onto that of the unknown genes.

This basically provides us with a parts catalogue for a given organism.

Finding genes on genomes

- Various computational (in silico) methods now available
- The content of the yeast genome (≈ 6400 ORFs):
  - Previously identified genes 30%
  - Identification by homology analysis 30%
  - Questionable assignments 7%
  - Single orphan ORFs 23%
  - Unidentified members of orphan families 10%

…but this procedure does not give the full gene complement for an organism. Anywhere from 20 to 50% of the identified genes on genomes have no functional assignment--so-called orphan ORFs.

There will be some years before we will be able to define the full gene complement of an organism.

What is found in a genome? Example: E. coli, Blattner et al 1997

• Total protein coding regions	4288
• Biosynthesis no Amino acids	131
• Biosynthesis of cofactors	103
• Biosynthesis of nucleotides	58
• Cell envelop protein	237
• Energy metabolism	243
• Intermediary metabolism	188
• Lipid metabolism	48
• DNA replication, recombination, repair	115
• Protein folding	9
• Regulatory protein	178
• Transcription	55
• Translation	182
• Uptake from environment	427
• Still about 30% of ORFs without assigned function	

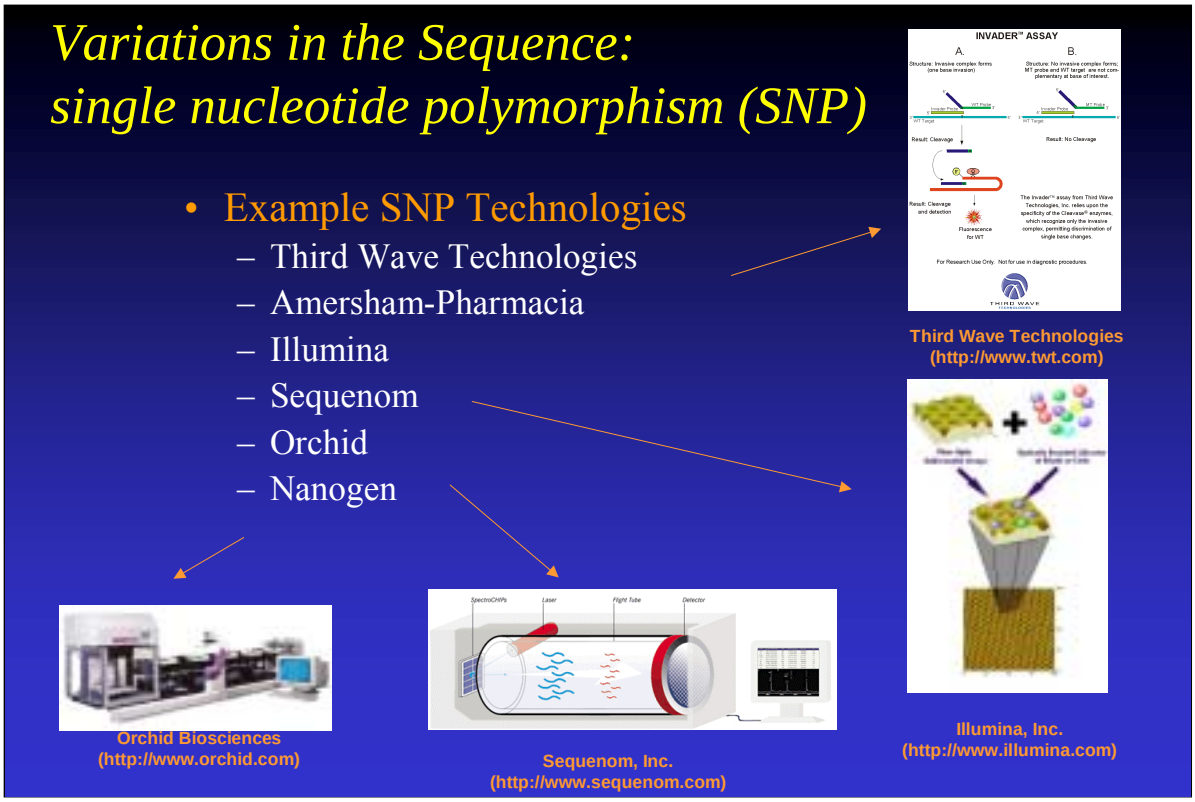
Comparing genomes and sequences

Inter-species

- Genomes can be compared
- Phylogenic trees can be constructed
- Evolutionary implications can be pondered
- Minimal gene sets can be defined
  - (250-300 genes)

Intra-species

- Variations in sequence can be studied
  - Basis for human genotype-phenotype relationship

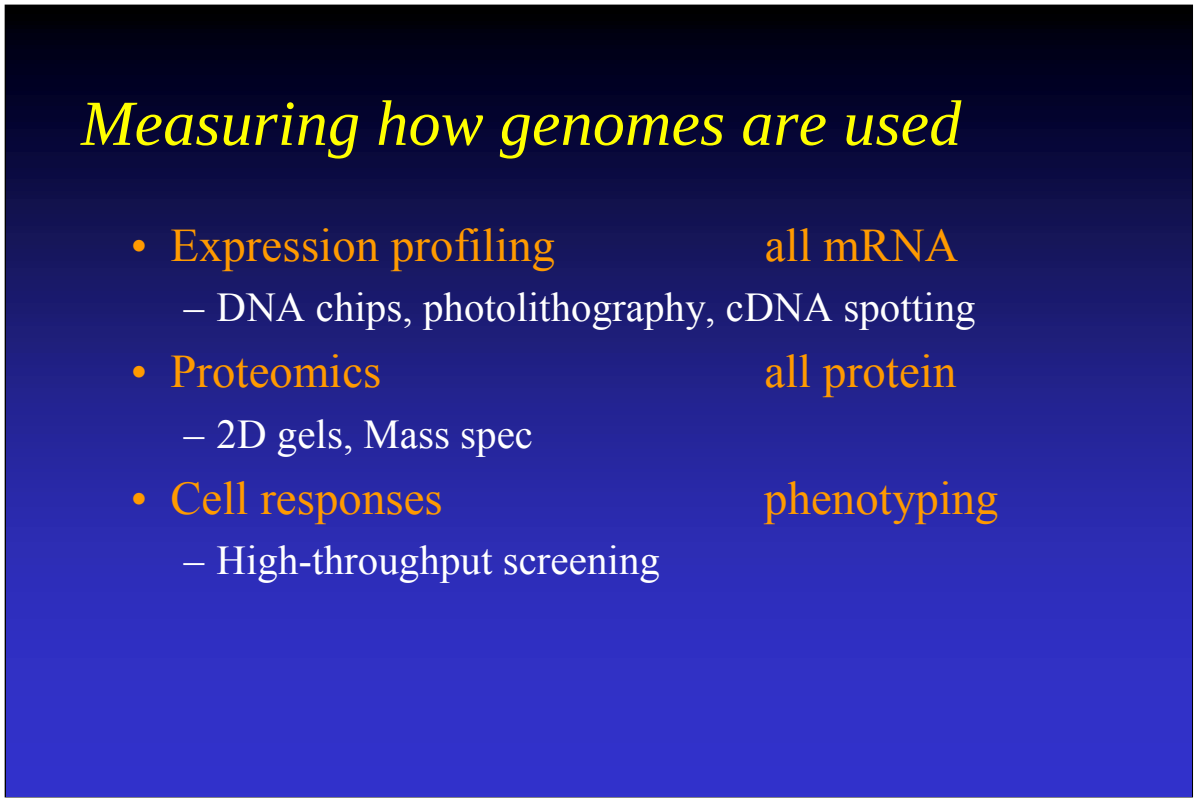


SNPs

With the human sequence in sight, and now in hand, detecting the individual variations in the sequence has come into focus. Although estimates vary there are differences in about 1 per 1000 base pairs between individuals. SNPs are getting most of the interest although deletion and insertions are also an important factor in the genomic differences between individual.

Currently there is a significant effort being put into establishing about 150,000 SNP map of the human chromosomes. Such a map should be unique for each individual on the planet.

Relating these variations to human traits is of significant interest, especially for disease traits and patho-physiology.



USE OF GENOMES

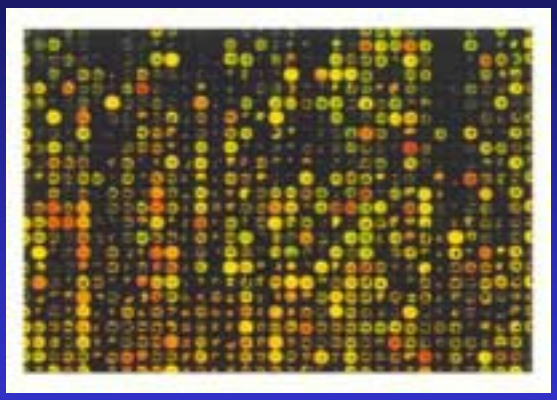
Now that we have full DNA sequences and gene complements, there are a number of approaches emerging that allow us to measure on a genome-scale how these genes are deployed by an organism and what the resulting phenotypic behavior is.

We will only briefly mention expression profiling, namely the measurement of all the messages for protein production that are present in a cell at any given time.

Proteomic methods are aimed at doing the same for the entire protein portfolio of a cell and there are an increasing number of methods being developed for the high-throughput measurement of the physiological responses of a cell.

DNA chips
Photolithography by Affymetrix

Affymetrix, Inc.
(http://www.affymetrix.com)



DNA CHIPS

The so-called DNA chips array a large number of specific oligonucleotides (typically 25 base pairs in length, or 25-mers), at a high density. The feature sizes can be below 50 micron.

Perhaps the best known of these technologies is in situ synthesis using photolithography. Affymetrix makes and sells such chips. They come with a scanner as shown on the left and a read out of the chip as shown on the right.

Other approaches to making arrays include physically arraying oligos using micro-fluidics and in situ synthesis using a large number of stear-able mirrors.

Some examples of expression profiling studies

- Yeast cell cycle
- Yeast sporulation
- Diauxic metabolic shifts
- Fibroblast responses to cell culture
- The aging process
- Classification of cancers into subtypes
- Finding drug targets
- Developmental biology

A number of very insightful studies have been performed using DNA chips. Some of these are shown on the ensuing slides.

DNA chips are still to expensive for routine use. Each array or data point in such studies costs between $1K to $5K depending on sample preparation and other factors.

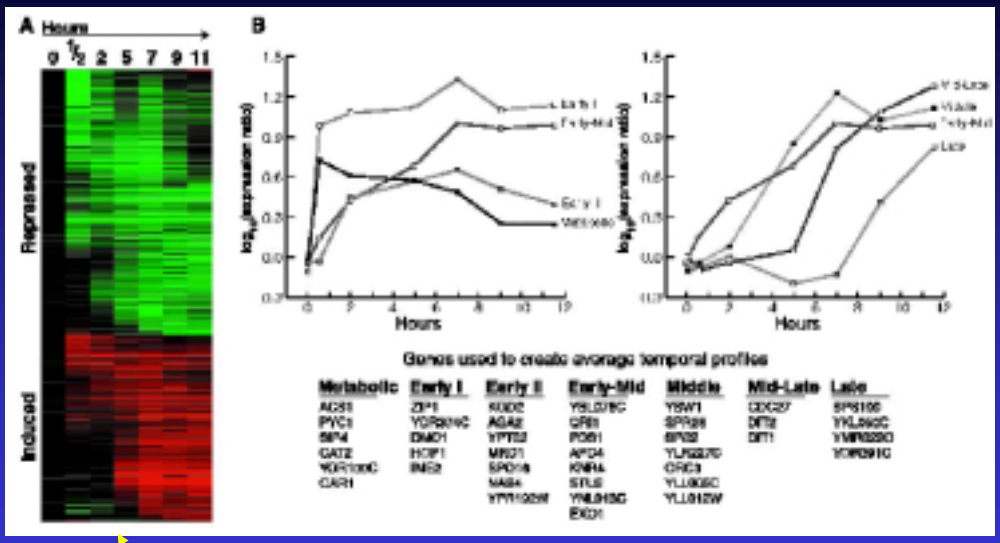
In addition to the slides to follow; here are a couple of studies of great interest:

Ly, D.H., Lockhart, D.J., Lerner, R.A., and Schultz, P.G. "Mitotic mis-regulation and human aging," Science, **287**, 2486.

PT Spellman et al "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization," Molecular Biology of the Cell, Vol 9, 3273-3297, (1998)

Yeast Sporulation

Seven temporal groups of genes
Contain hundreds of unassigned genes
These genes have homologs in other organisms



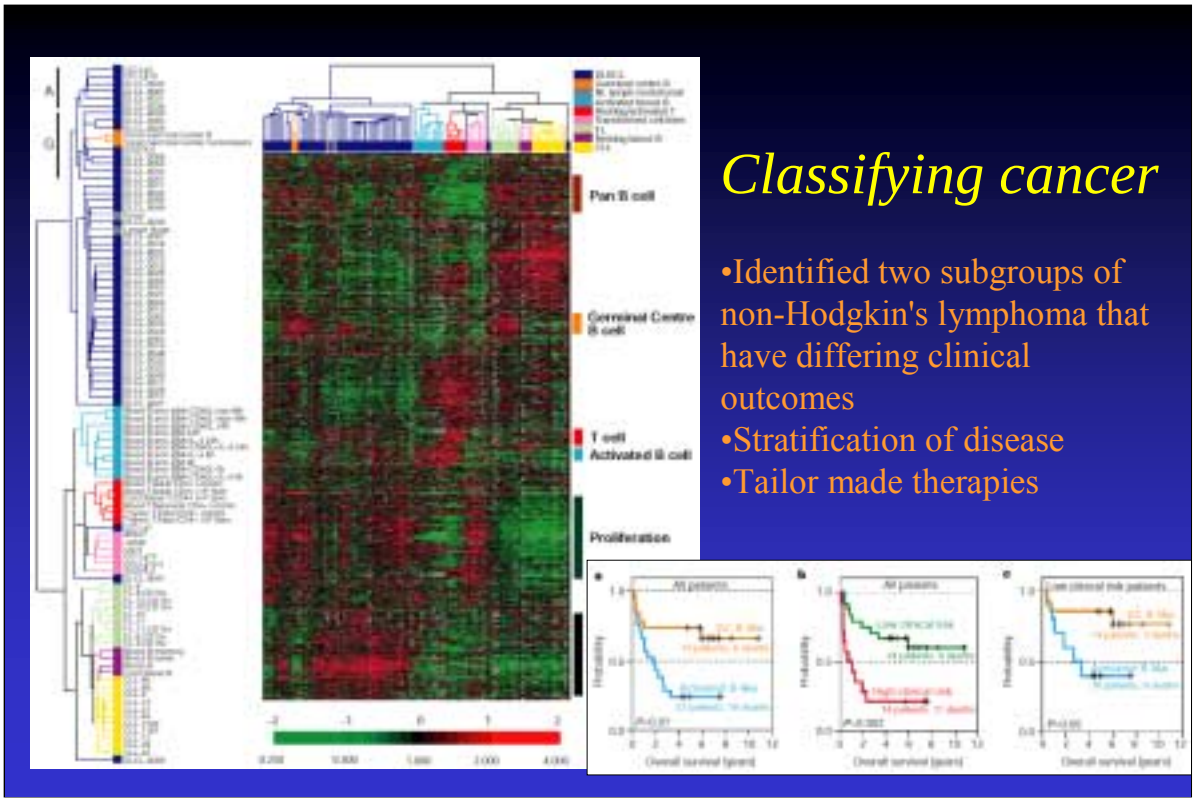
Each array is converted into a column, or a vector.
This vector is a snapshot of the state variables of the cell

From:

The Transcriptional Program of Sporulation in Budding Yeast

S. Chu,* J. DeRisi,* M. Eisen, J. Mulholland, D. Botstein, P. O. Brown,† I. Herskowitz† SCIENCE VOL 282 23 OCTOBER 1998

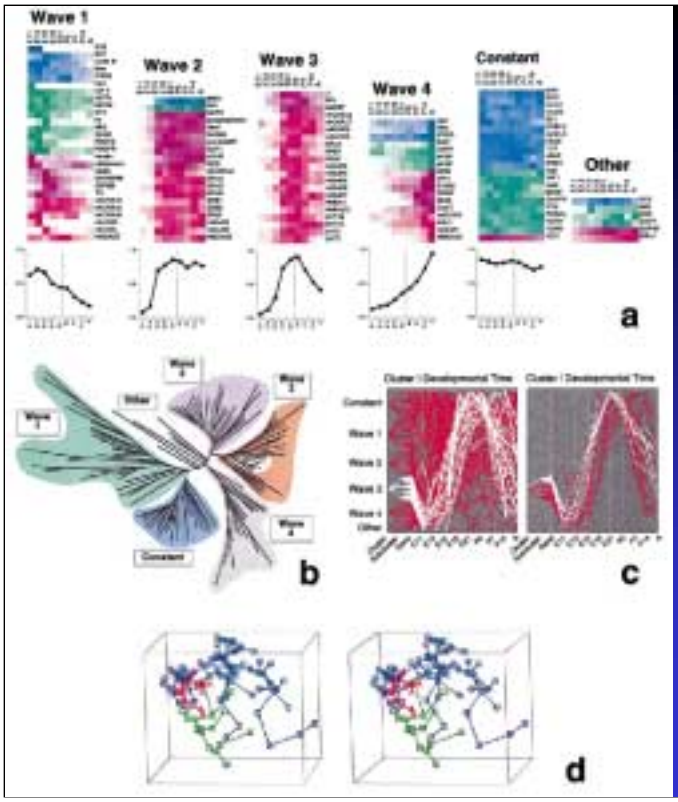
ABSTRACT: Diploid cells of budding yeast produce haploid cells through the developmental program of sporulation, which consists of meiosis and spore morphogenesis. DNA microarrays containing nearly every yeast gene were used to assay changes in gene expression during sporulation. At least seven distinct temporal patterns of induction were observed. The transcription factor Ndt80 appeared to be important for induction of a large group of genes at the end of meiotic prophase. Consensus sequences known or proposed to be responsible for temporal regulation could be identified solely from analysis of sequences of coordinately expressed genes. The temporal expression pattern provided clues to potential functions of hundreds of previously uncharacterized genes, some of which have vertebrate homologs.



Distinct types of diffuse largeB-cell lymphoma identified by gene expression profiling

Ash A. Alizadeh et al, NATURE, VOL 403 : 503 (2000)

Diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma, is clinically heterogeneous: 40% of patients respond well to current therapy and have prolonged survival, whereas the remainder succumb to the disease. We proposed that this variability in natural history reflects unrecognized molecular heterogeneity in the tumours. Using DNA microarrays, we have conducted a systematic characterization of gene expression in B-cell malignancies. Here we show that there is diversity in gene expression among the tumours of DLBCL patients, apparently reflecting the variation in tumour proliferation rate, host response and differentiation state of the tumour. We identified two molecularly distinct forms of DLBCL which had gene expression patterns indicative of different stages of B-cell differentiation. One type expressed genes characteristic of germinal centre B cells (`germinal centre B-like DLBCL'); the second type expressed genes normally induced during in vitro activation of peripheral blood B cells (`activated B-like DLBCL'). Patients with germinal centre B-like DLBCL had a significantly better overall survival than those with activated B-like DLBCL. The molecular classification of tumours on the basis of gene expression can thus identify previously undetected and clinically significant subtypes of cancer.



Developmental Biology

- Fluctuations in mRNA expression of 112 genes during rat central nervous system development
- Classified into consecutive waves of expression
- Identifying coherent patterns and sequences of events in the complex genetic signaling network of development

Large-scale temporal gene expression mapping of central nervous system development

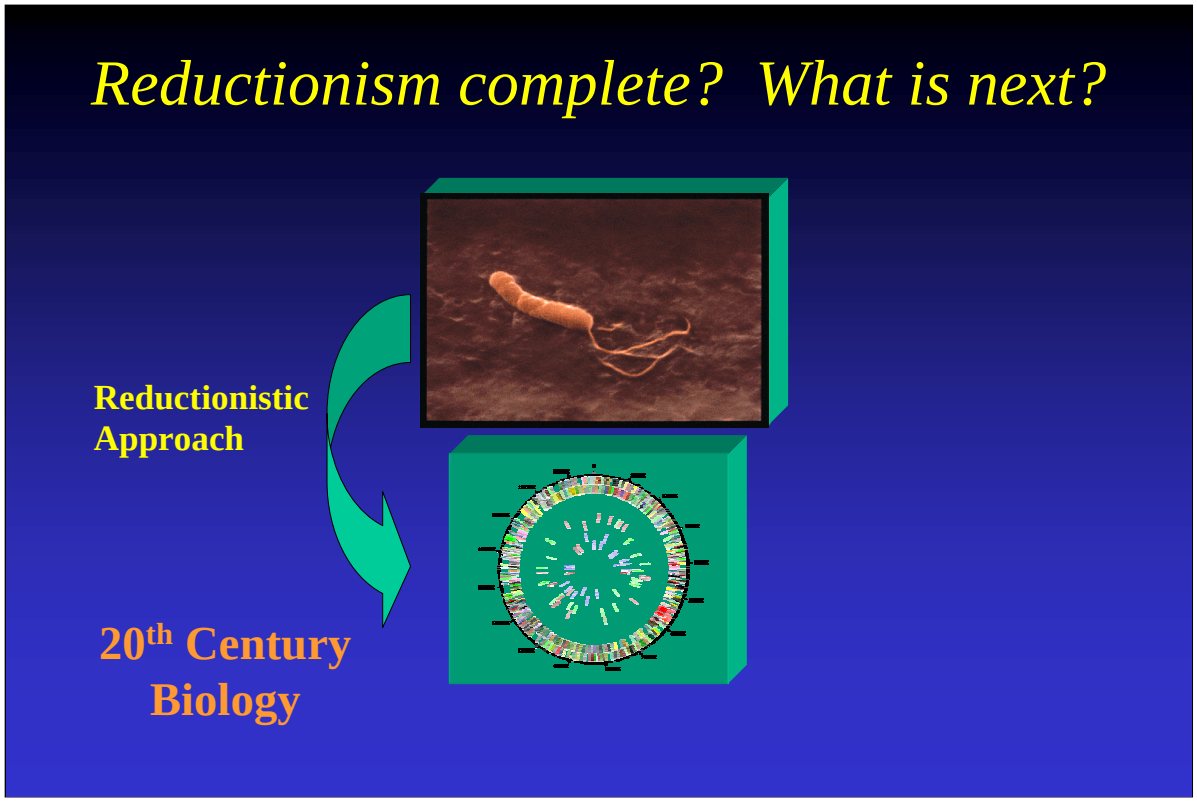
XILING WEN*, STEFANIE FUHRMAN*, GEORGE S. MICHAELS † ,DANIEL B. CARR † , SUSAN SMITH*, JEFFERY L. BARKER*, AND ROLAND SOMOGYI* ‡

Proc. Natl. Acad. Sci. USA Vol. 95, pp. 334–339, January 1998

ABSTRACT We used reverse transcription–coupled PCR to produce a high-resolution temporal map of fluctuations in mRNA expression of 112 genes during rat central nervous system development, focusing on the cervical spinal cord. The data provide a temporal gene expression ‘‘fingerprint’’ of spinal cord development based on major families of inter- and intracellular signaling genes. By using distance matrices for the pair-wise comparison of these 112 temporal gene expression patterns as the basis for a cluster analysis, we found five basic ‘‘waves’’ of expression that characterize distinct phases of development. The results suggest functional relationships among the genes fluctuating in parallel. We found that genes belonging to distinct functional classes and gene families clearly map to particular expression profiles. The concepts and data analysis discussed herein may be useful in objectively identifying coherent patterns and sequences of events in the complex genetic signaling network of development. Functional genomics approaches such as this may have applications in the elucidation of complex developmental and degenerative disorders.

Trends

- Technology is getting faster and cheaper (just like CPUs)
- Seems to follow Moore's law
- Sequencing is becoming pretty cheap $0.1/bp
- SNPs, current $1/marker, $0.01/marker expected in 18 mo
- Expression profiles, very expensive, $1K-$5K/sample
- Proteomics, no good numbers available
- Molecular data will not be limiting,
  - but good physiological responses
  - and mathematical analysis will be



....and in the end

Has the advent of HT-technologies signaled the end of reductionism in biology? Probably. They seem to have closed out last centuries biological research nicely giving us detailed and comprehensive lists of biological components.

We must now figure out how to put the pieces together again. The following lectures will focus on this topic.

How will this be done?

What will the role of Chemical and Bio-engineers be in this process?

References

- Lander, E.S. and Weinberg, R.A., "GENOMICS: Journey to the Center of Biology," *Science*, **287:** 1777
- Palsson, B.O., "What lies beyond bioinformatics?" *Nature Biotechnology*, **15**: 3-4 (1997).
- Strothman, R.C., "The Coming Kuhnian Revolution in Biology," *Nature Biotechnology*, **15**: 194-199 (1997).
- Hartwell, L.H., JJ; Leibler, S; Murray, AW, "From molecular to modular cell biology," *Nature*, **402** (6761 Suppl.):C47-52, (1999)
- Bailey, J.E., "Lessons from metabolic engineering for functional genomics and drug discovery," *Nature Biotechnology*, **17**: 616-8 (1999)
- Aebersold, R; Hood, LE; Watts, JD, "Equipping scientists for the new biology," *Nature Biotechnology*, **18**: 359 (2000).
- Palsson, B.O., "The challenges of *in silico* biology," *Nature Biotechnology*, **18:** 1147-1150 (2000).

Some web sites

- EcoCyc
  - (http://ecocyc.panbio.com/ecocyc/ecocyc.html)
- Kyoto Encyclopedia of Genes and Genomes (KEGG)
  - (http://www.genome.ad.jp/kegg/)
- What Is There (WIT) system
  - (http://216.190.101.28/IGwit/ or http://wit.mcs.anl.gov/WIT/)
- The Munich Information Center for Protein Sequences (mips)
  - (http://www.mips.biochem.mpg.de/)
- Biology Workbench
  - (http://workbench.sdsc.edu/)
- the EMP Project
  - (http://www.empproject.com/)
- SWISS-PROT
  - (http://expasy.cbr.nrc.ca/sprot/)

Thanks to:

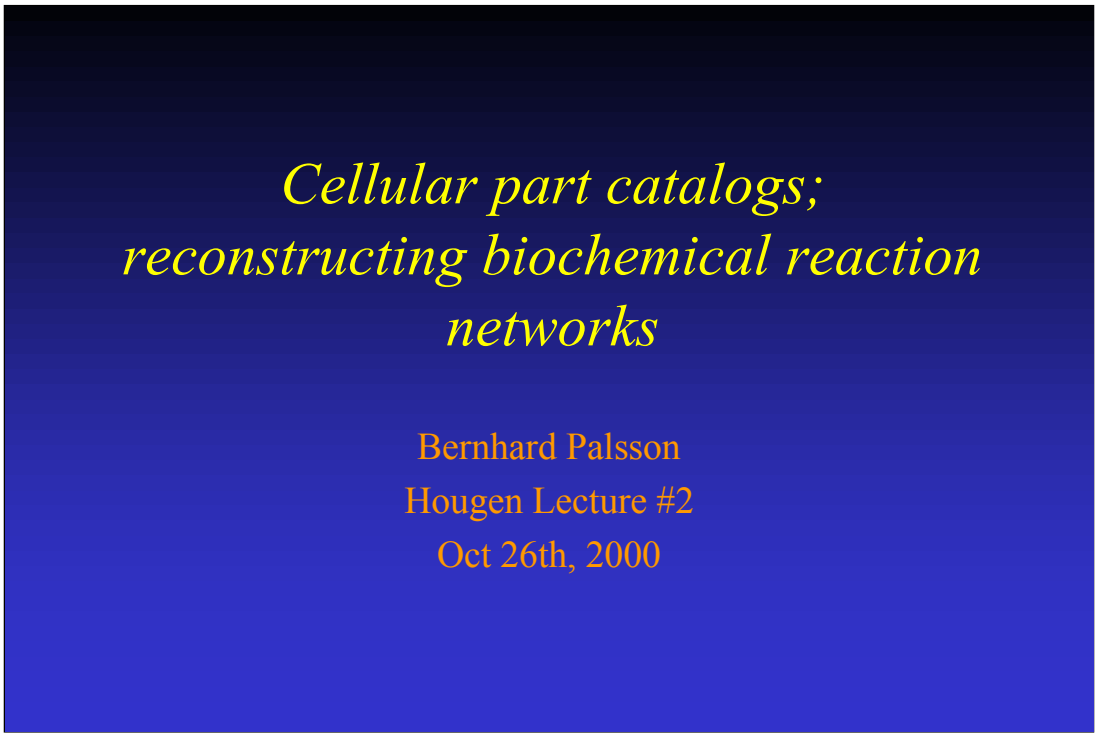
- Marc Abrams
- Markus Covert
- Tom Fahland
- Iman Famili
- Jeremy Edwards
- David Letscher
- Christophe Schilling
- Sharon Smith

For their help making these slides

Special Thanks To:

Ed Lightfoot

for his hospitality and care for my well being while in Madison



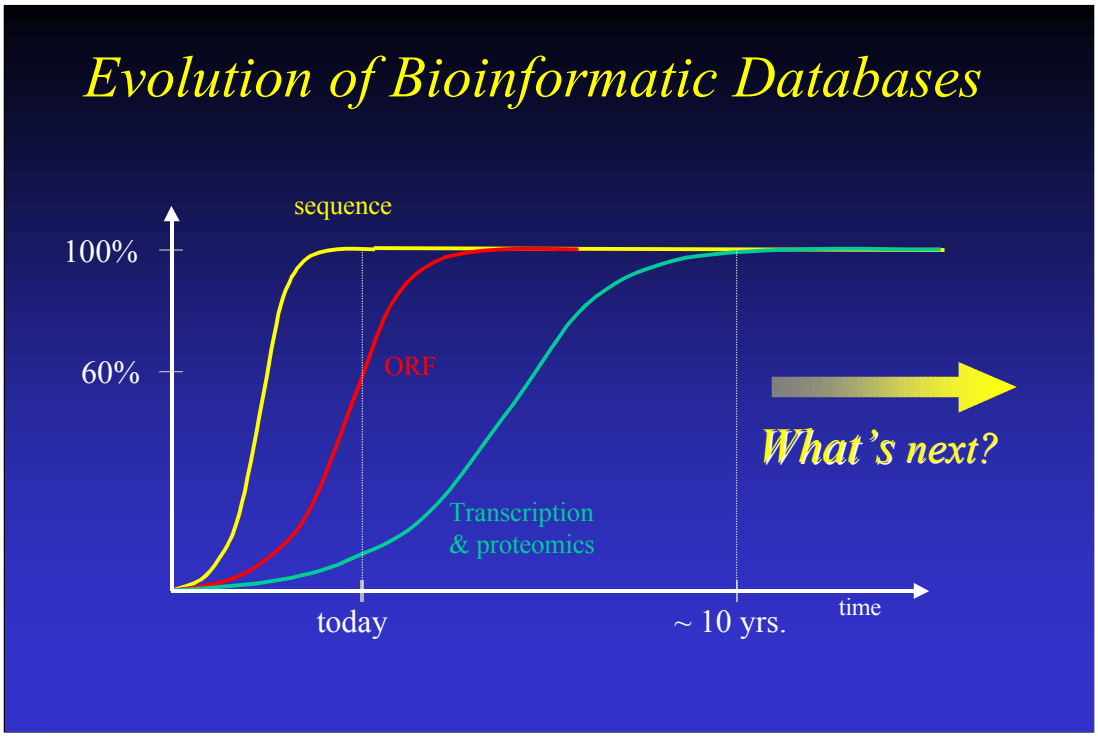
INTRODUCTION

Now that HT experimental approaches give us parts catalogs, we can begin to assess the piece-wise interactions between gene products. These pair-wise interactions will lead to the reconstruction of biochemical reaction networks. This reconstruction process is the subject of lecture #2.

Lecture #2: Outline

- The Dogma of *in silico* Biology
  - Pair-wise interactions
  - Networks
  - Emergent properties and biological function
- Why Bio/Chemical-engineering
- Network reconstruction
  - Genomic data
  - Biochemical information
  - Physiology
- Connectivities
- Why construct mathematical models?

LECTURE #2



PUTTING IT IN PERSPCTIVE

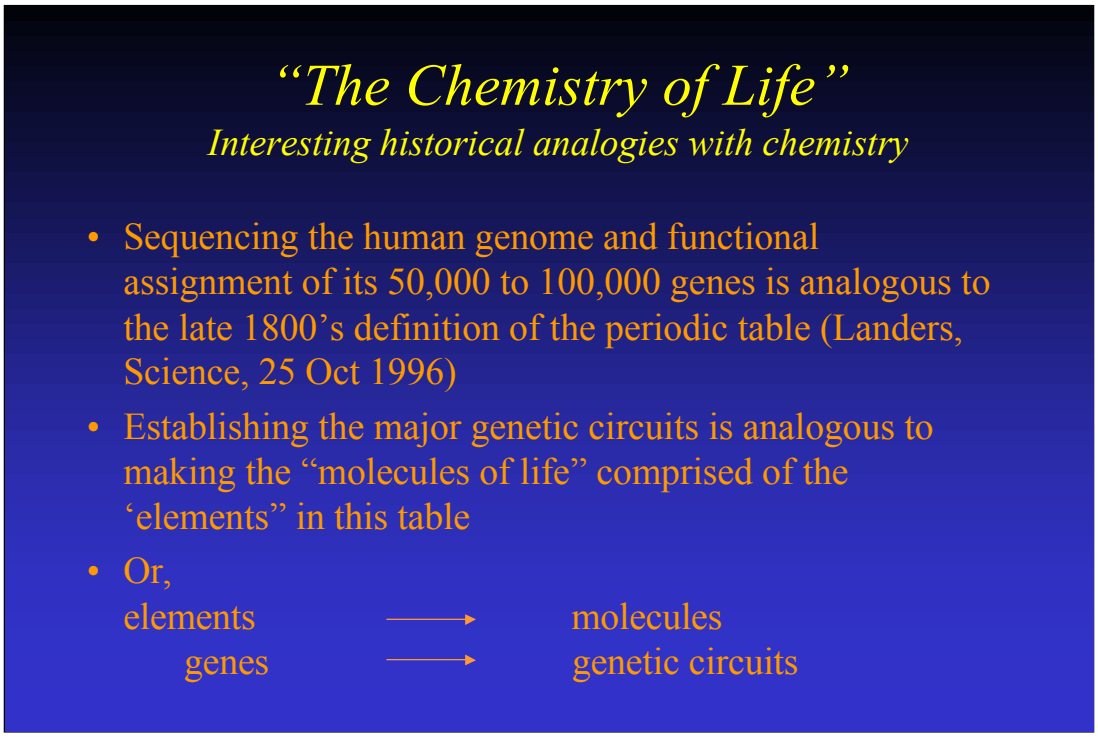
This slide provides just a crude perspective of where we stand today in terms of the evolution of bioinformatic databases and scientific information.

Clearly we have the capability to sequence a complete genome and through genome annotation techniques we can currently assign function to roughly 2/3 of the coding regions in a genome.

And now with the rise of proteomics and expression profiling technologies we are beginning to gain insight on how the genome is utilized by an organism under various environmental conditions, offering us snapshots of the dynamics within the cell.

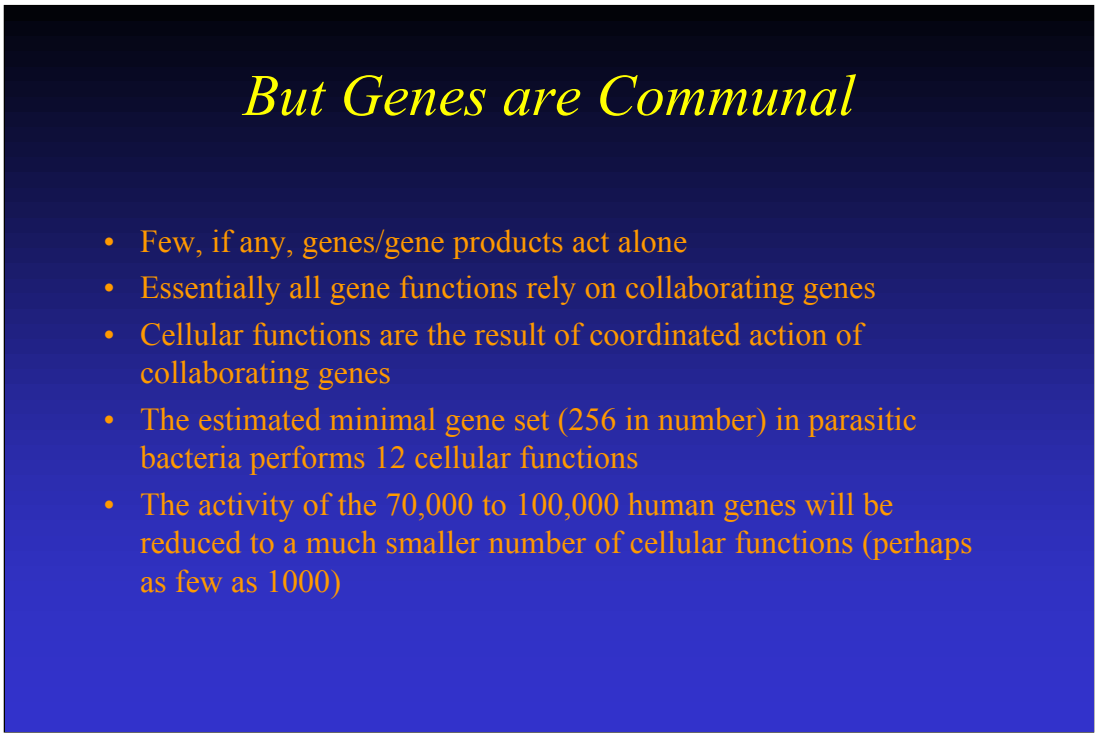
If we look ahead into the not too distant future we can expect to have enormous amounts of information pertaining to the content, structure, and expression of the genotype.

How do we use all of this genomic and biochemical information to gain insight into the relationship between an organism's genotype and its phenotype?



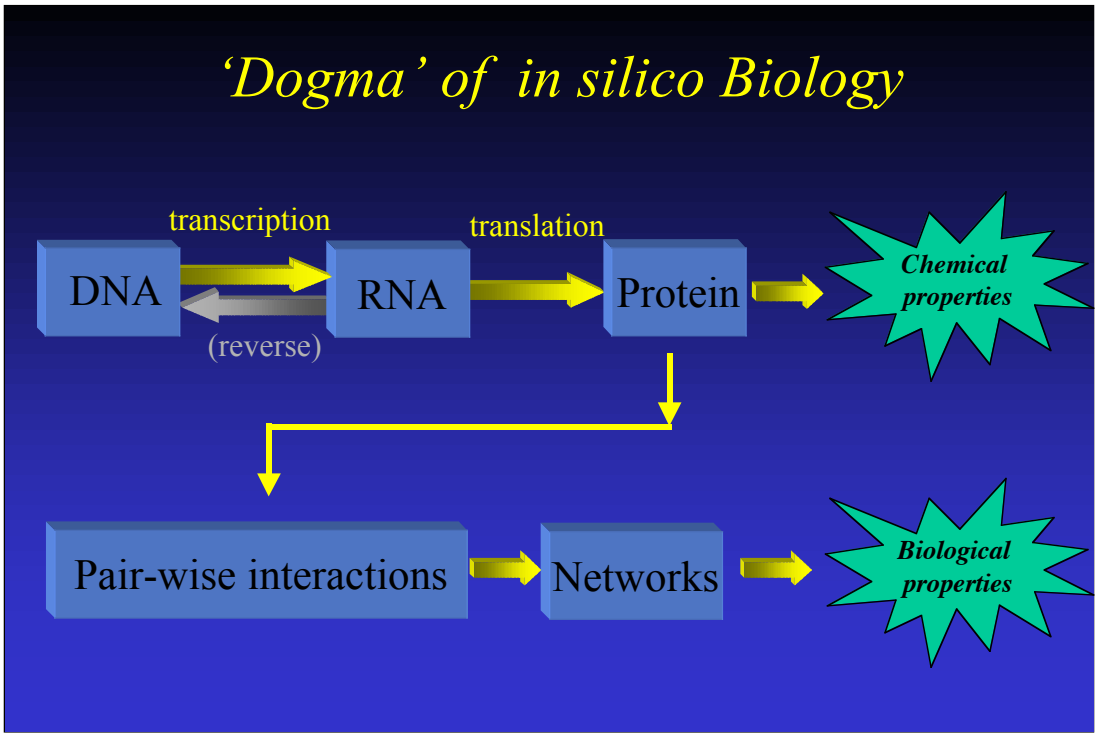
THE LANDER ANALOGY

Eric Landers drew this interesting analogy between the history of chemistry and biology. About one hundred years ago chemists were busy filling in the periodic table. This table represents the atoms that then are build chemical compounds. According to Landers, we are in a large sense constructing an periodic table of life by identifying all the genes that are found in organisms. Then particular combinations of these elements (actually something analogous to isotopes since there are species specific variations in the gene sequences) are put together to build a particular organism.



GENES WORK TOGETHER

With very few exceptions all cellular functions are reliant on multiple gene products. So although the central dogma describes the process of protein molecules from the information encoded on a DNA sequence, the proteins have individual chemical functions. All these chemical functions together form a biological process. It appears that most cellular processes require on the order of 20 to 70 different gene products.

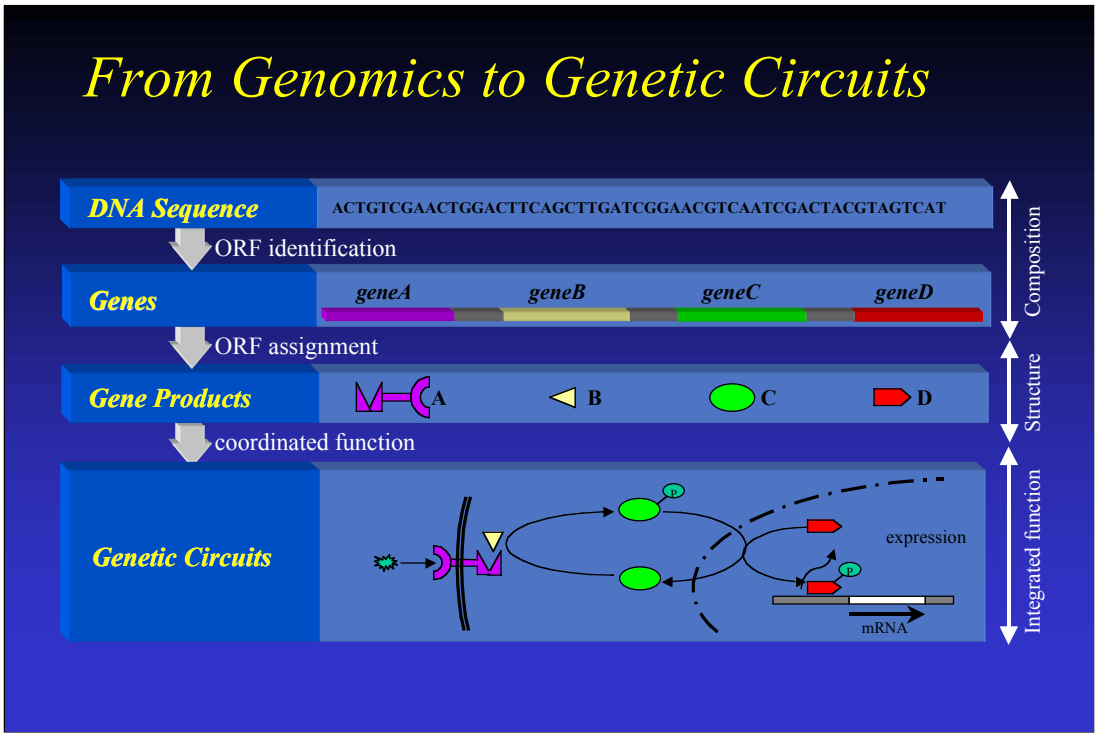


THE DOGMA OF IN SILICO BIOLOGY

Thus we are forced to move beyond the central dogma of molecular biology when trying to reconstruct cellular functions from the component list. First we must identify the pair-wise interactions between the individual gene products. Then we must construct the networks that result from the totality of such pair-wise interactions. There are many in vivo and in silico methods to accomplish this task. We will describe some of these in this lecture.

Then we wish to study the properties of these networks. These properties are those of the whole and represent biological properties. Examples include, redundancy, robustness, built in oscillations, etc. These properties cannot be deduced from the components alone.

Some of the methods available for such analysis will be described in subsequent lectures.



GENETIC CIRCUITS

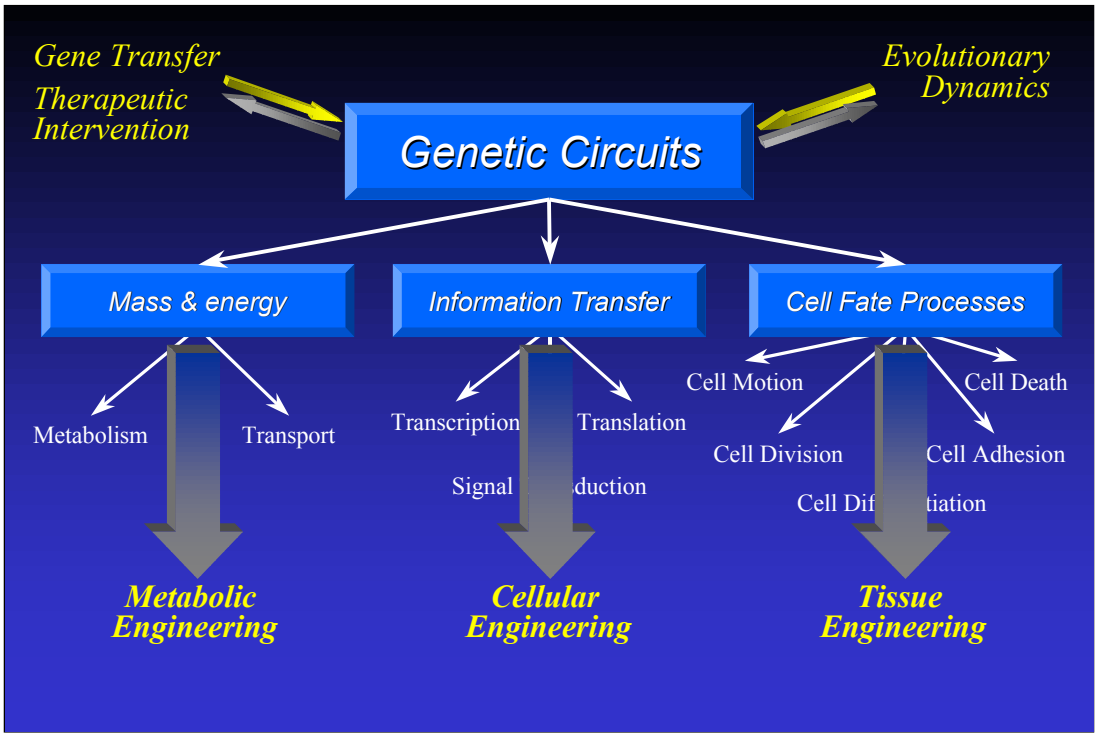
The relationship between the genotype and the phenotype is complex, highly non-linear and cannot be predicted from simply cataloging and assigning gene functions to genes found in a genome.

Since cellular functions rely on the coordinated activity of multiple gene products, the inter-relatedness and connectivity of these elements becomes critical.

The coordinated action of multiple gene products can be viewed as a network, or a "GENETIC CIRCUIT," which is the collection of different gene products that together are required to execute a particular function.

Thus if we are to understand how cellular functions operate, the function of every gene must be placed in the context of its role in attaining the set goals of a cellular function.

This "holistic" approach to the study of cellular function is centered around the concept of a genetic circuit.



CLASSIFICATION OF GENETIC CIRCUITS

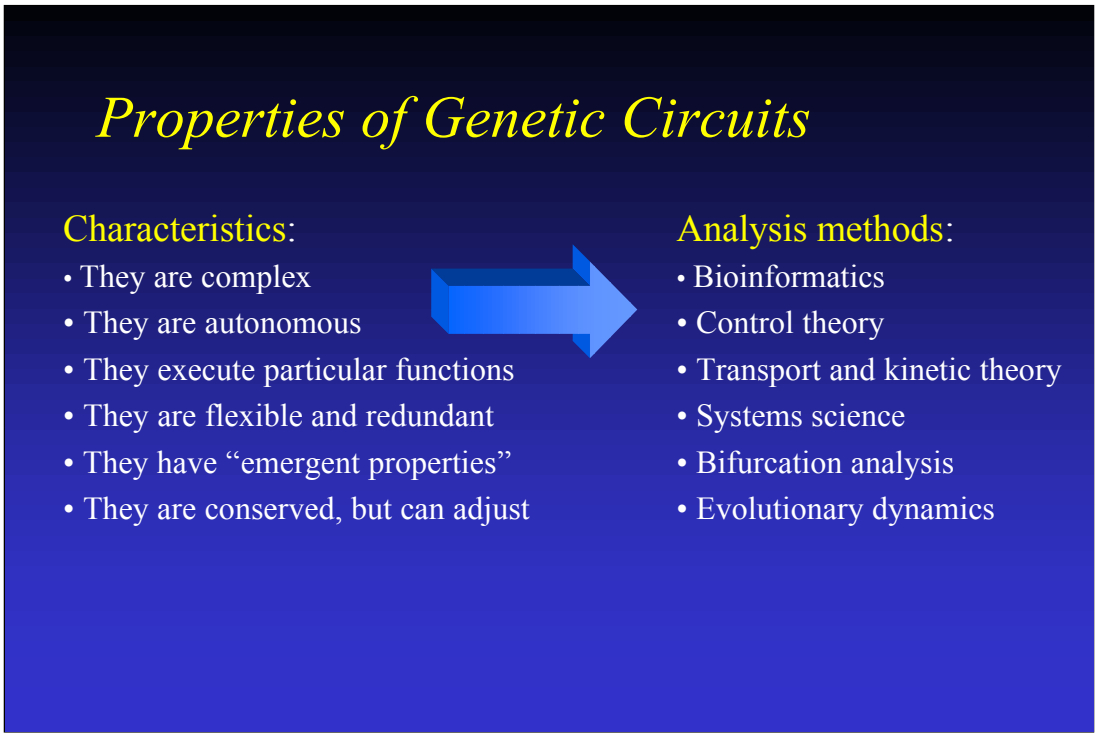
Although we do not know all the genetic circuits found on a genome we can still begin to classify them. A coarse grained classification is illustrated in this slide:

1. Cells allocate their energy and material resources through metabolism. It is universal and can be called the 'chemical engine' that drives the living process. Metabolism consists of a complex set of transforming chemical reactions and associated transport reactions. We know much about metabolism as it has been studied since the 1930s.

2. The processing, maintenance, and transmission of the information carried on the DNA is also universal. All living organisms have processes that carry out these tasks. Again we do know quite a bit about these processes and there are strong similarities amongst different organisms.

3. In multi cellular organisms, the cells must coordinate their activities relative to one-another. These processes are becoming better understood, but are not as well established as 1. and 2. above. For instance many of the gene products associated with programmed cell death (apoptosis) are beginning to be identified but we may not know their biochemical functions

The slide also illustrates how these groups of genetic circuits are fundamental to the bioengineering of various cellular functions and organism properties.



HOW WILL WE STUDY GENETIC CIRCUITS?

The objective of studying genetic circuits is to analyze, interpret, and predict the relationship between the genotype and the phenotype.

Although not all the fundamental properties of genetic circuits are known at present, some important ones can be stated.

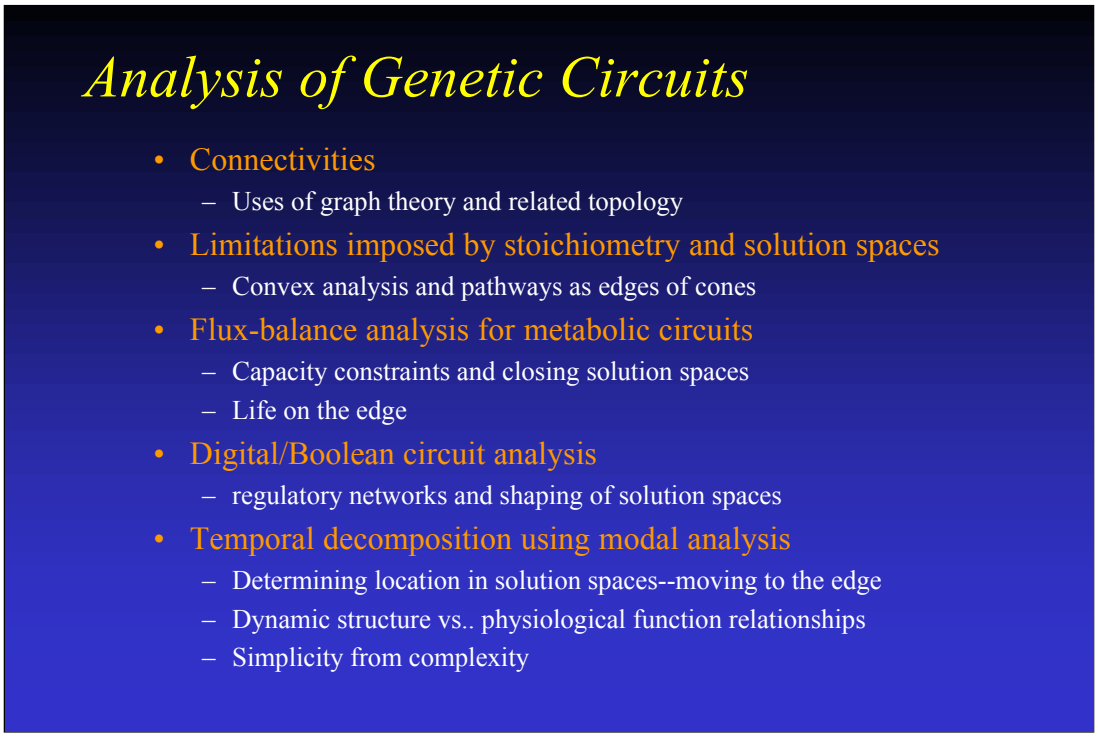
In general they are complex with many components which offer a degree of flexibility in functioning and in evolving. Once genes are expressed, the coordinated function of the gene products is autonomous, and embedded within these built in controls are the capabilities to perform creative functions.

For each of these properties we can look to accompanying theories and analytical tools such as those listed here to help study these circuits.

Of course this only offers a glimpse into the set of existing tools which can be utilized, and the development of novel approaches to study genetic circuits is needed.

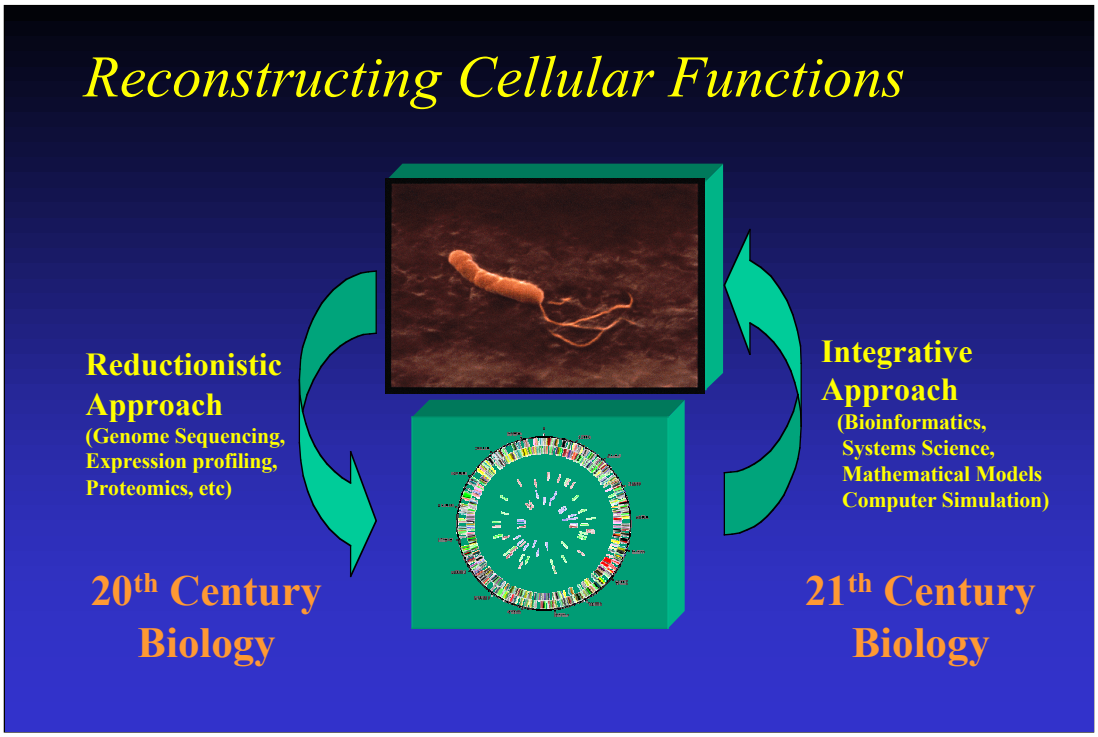
Genetic Circuits; a different point of view

- Bioinformatics: a way to define, classify, and cross-species correlate genetic circuits
- Gene therapy: not replacing a defective gene but fixing a malfunctioning circuit
- View evolution as a process of tuning and acquiring genetic circuits
- Genomic taxonomy based on genetic circuitry
- Bioengineer ex vivo procedures to tune genetic circuits
- Fundamental to applied biology; e.g. metabolic and tissue engineering



ANALYSIS

The following lectures will outline the approach of the successive imposition of governing constraints. This slide illustrates some of these constraints and the order in which we shall ally them.



REDUCTIONISM REVERSED

It is thus becoming clear that we need to reverse the process on the left-hand side, and to study how these components interact to form complex systems.

This poses the question, given the complete genomic sequence, is it possible to reconstruct the functions of a cellular or biological system?

The process of reconstructing the biological system from the reductionist information will rely on bioinformatics to identify the "parts catalogue" if you will.

However, the parts catalogue does not contain functional information. For example, listing all the parts of car, does not even begin to describe the how the the automobile works.

Therefore, to understand multigenic functions, a systems science analysis is required.

Why Bio/chemical-engineering?

- Information intensive-- computer science
- Requires computations
- Each component of circuit obeys P/C principles (chemical kinetics, thermodynamics, biomechanics)
- Simultaneous action of multiple gene products (systems analysis, control theory)
- Most of these issues found in to days BioE/ChE curricula

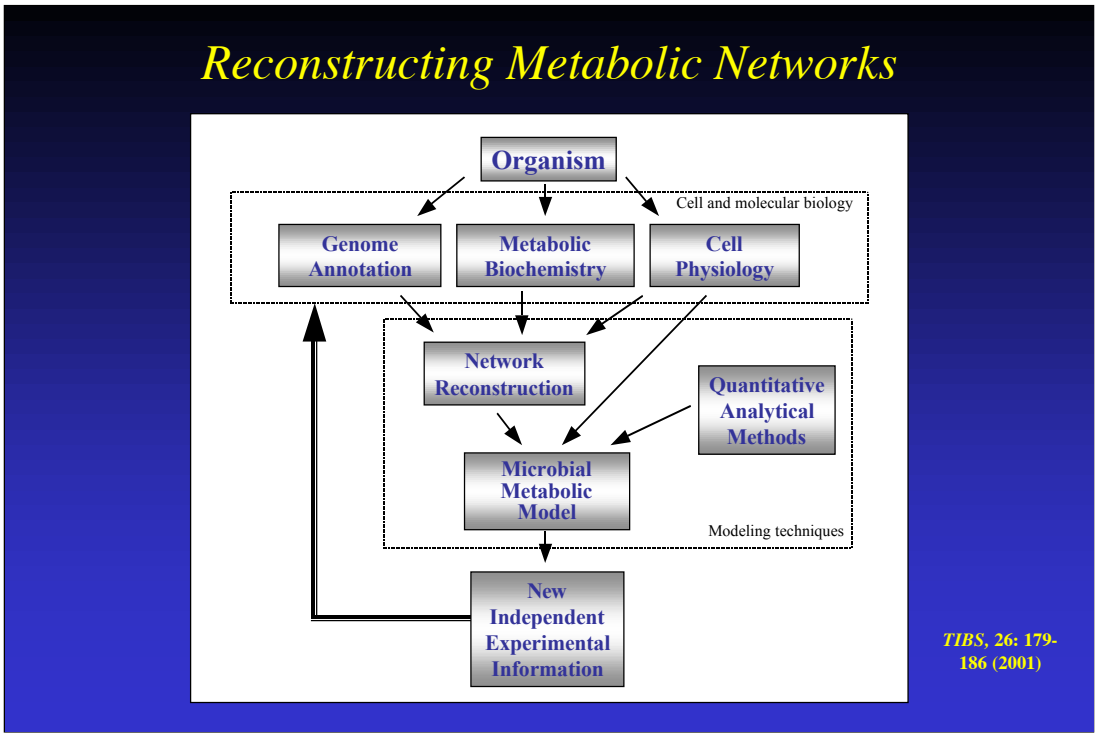
NEW CURRICULA

New degree programs in this area will be primarily comprised of three components. First, fundamental understanding of the under-pinings of the high-throughput experimental technologies. Second, the complex informatics infrastructure that comes with the high volumes of data being generated. Third, we need to be able to mathematically describe all the data generated using the governing P/C principles to construct computer models of complex biological functions.

Upon careful examination of chemical and bioengineering curricula, about 2/3 of what is needed for this new curricula is found therein.

NETWORK RECONSTRUCTION

Given this background and historic perspective we now begin the process of developing systems or in silico biology. We shall first discuss network reconstruction.

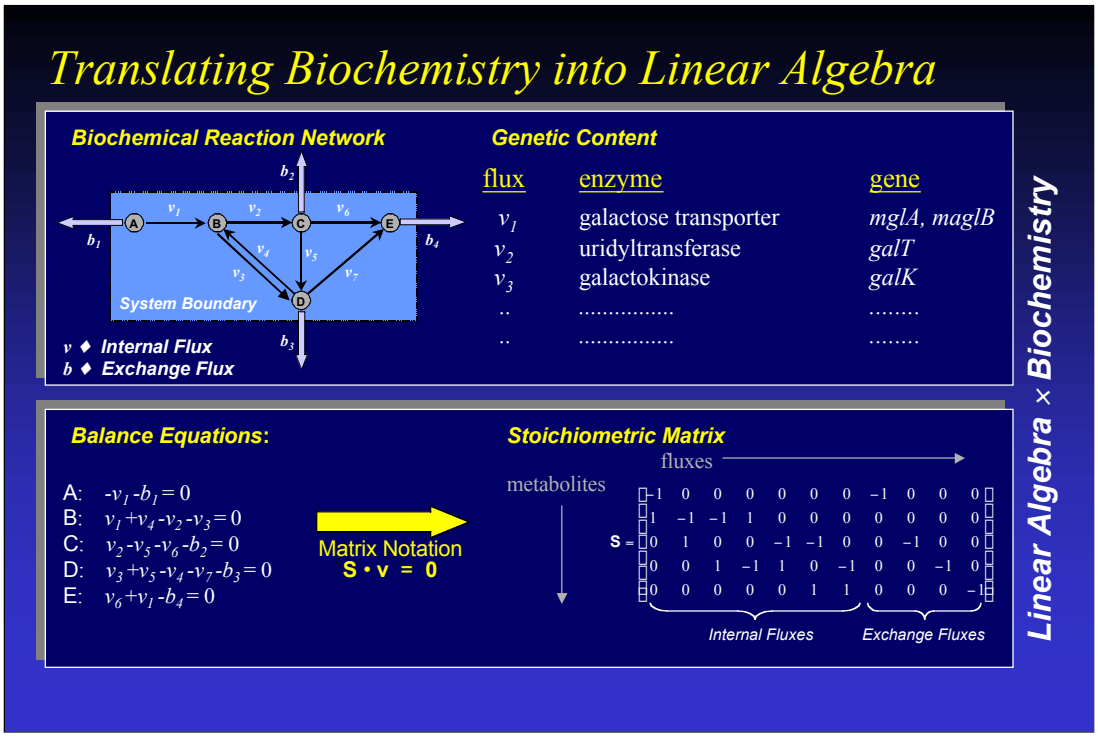


THE RECONSTRUCTION PROCESS

There are three principal types of data for network reconstruction: genomic, biochemical, and physiological. Once the network is formulated, then mathematical methods can be applied to assess its properties. The reconstruction process will be outlined for *H. pylori* in the slides to follow.

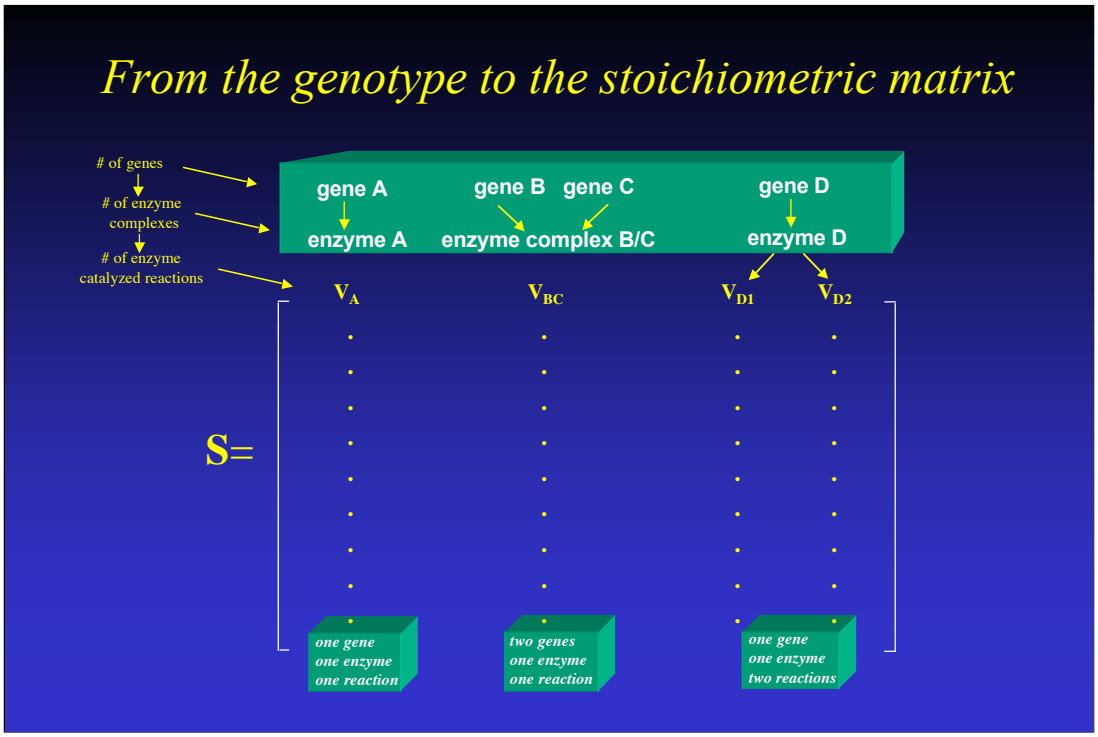
At present this process cannot be automated, and in particular much human input and interpretation is required in reading all the pertinent literature on known biochemical activity reported for the organism in question and to interpret its physiological functions.

At present, this process takes a full time effort for 3 to 6 months for a single individual depending on the complexity of the organism studied and the amount of experimental data that is available.



CASTING GENOMIC INFORMATION INTO CONNECTIVITY MATRICES

Thus we can translate the biochemistry of a reaction network directly into realm of linear algebra in the form of a stoichiometric matrix. Beginning with the gene products of a system we can determine the interconversions of metabolites which occur and then simply take mass balances around each of these metabolites and represent this in the form of a stoichiometric matrix to complete the translation. Within the stoichiometric matrix lies all of the structural information and the architecture of the network. Having the matrix in this form allows for a detailed analysis based on concepts of linear algebra and convex analysis.



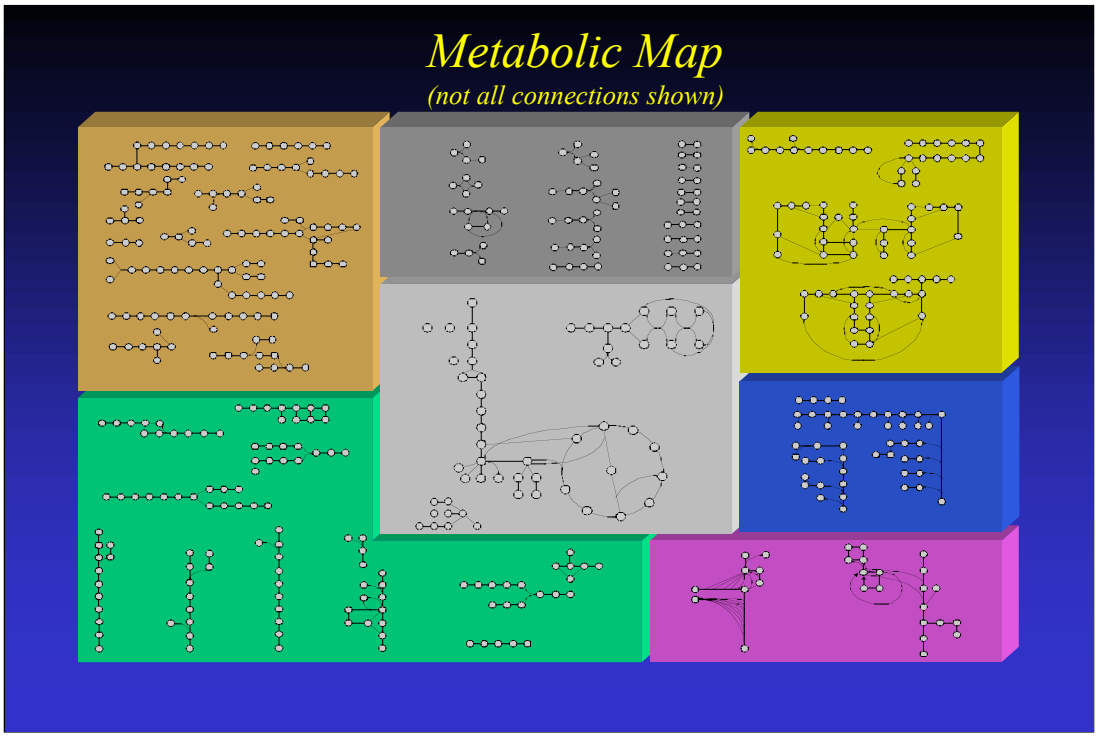
THE NUMBER OF REACTIONS IN A METABOLIC GENOTYPE IS NOT THE SAME AS THE NUMBER OF GENES IN THE GENOTYPE

There is not a one-to-one correspondence between the number of genes that are associated with metabolism and the number of chemical transformations that take place. This difference is due to several factors.

First, many enzymes are oligomeric complexes that contain more than one protein chain. These complexes are formed by non-covalent binding, or association of several different protein molecules. Hemoglobin, being a tetramer of two alpha and two beta globulins is perhaps the best know example of a protein oligomer.

Second, enzymes can catalyze more than one chemical reaction. This feature is often referred to as substrate promiscuity. These chemical transformations tend to be similar.

These features give rise to a different number of genes from the number of enzymes (or enzyme complexes) and the number of chemical reactions that take place. All of these situations can be accounted for with the stoichiometric matrix as illustrated.

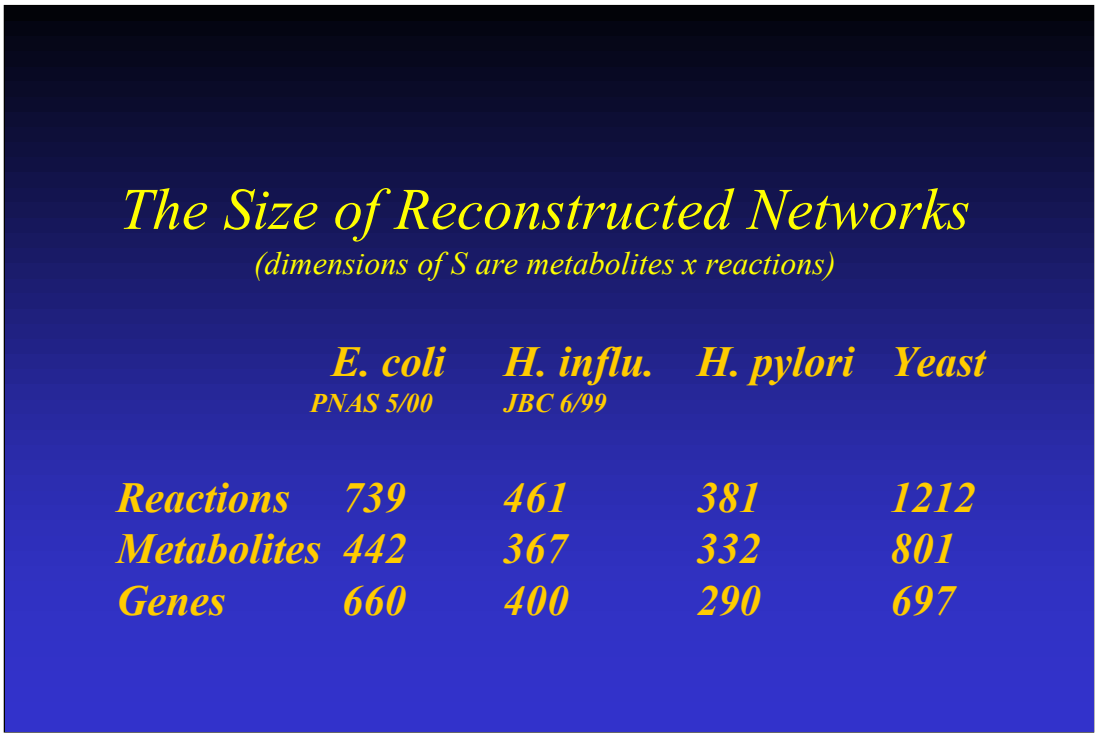


THE METABOLIC MAP REPRESENTATION OF THE *ESCHERICHIA COLI* K-12 METABOLIC GENOTYPE

The metabolic map of the *E. coli* K-12 metabolic genotype divided into metabolic sectors based on a biochemical rationale:

Gray:	Alternative carbon source metabolism
Light gray:	The core metabolic pathways
Orange:	Amino acid biosynthesis
Green:	Vitamin and co-factor metabolism
Yellow:	Nucleotide synthesis
Blue:	Cell wall synthesis
Purple:	Fatty acid synthesis

Not all the 720 reactions are shown. Highly connected metabolites, such as ATP, PEP and pyruvate are linked to dozens of reactions. Showing all of these connections would make this representation visually unattractive. However, these connections should not be overlooked as they play a key role in the stoichiometric characteristics of metabolism.

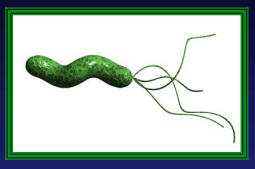


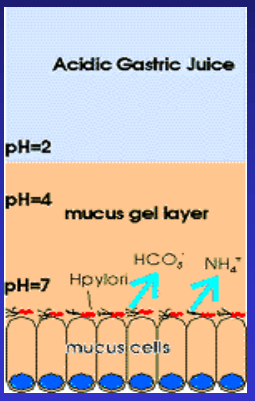
DIMENSIONS OF S

This table shows the size of the reconstructed metabolic networks by our research group. There are 350 to 800 metabolites present and 450-900 reactions depending on the complexity of the organism.

Note that the gene numbers correspond only to those gene products that participate directly in the reactions represented in the network. None of the associated regulatory or structural protein are included. As these models expand to account for regulation of gene expression, transcription and translation, the number of genes represented will increase greatly.

Helicobacter pylori Profile





Pathology

- Gram-negative pathogen colonizes the gastric mucosa
- major causative agent of peptic ulcers and gastric cancer
- inaccessible to human immune system
- survives in 4.0 – 7.0 pH range

Statistics

- Infects 30% of US population & ~50% of World popul.
- 75% of all ulcers are caused by HP (aspirin)
- correlates with socio-economic status

Genome Characteristics

- genome fully sequenced in August '97
- 1.66 Mbp genome length
- 1590 estimated genes

Helicobacter pylori is a spiral shaped bacterium that lives in the stomach and duodenum (section of intestine just below stomach). It has a unique way of adapting in the harsh environment of the stomach.

The inside of the stomach is bathed in about half a gallon of gastric juice every day. Gastric juice is composed of digestive enzymes and concentrated hydrochloric acid, which can readily tear apart the toughest food or microorganism. Bacteria, viruses, and yesterday's steak dinner are all consumed in this deadly bath of chemicals. It used to be thought that the stomach contained no bacteria and was actually sterile, but *Helicobacter pylori* changed all that.

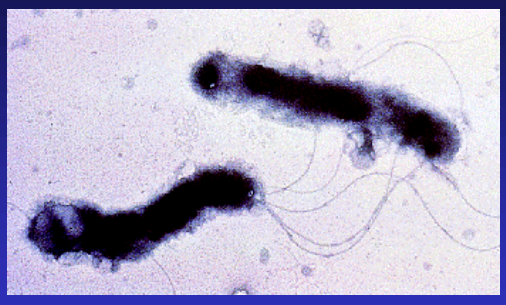
The stomach is protected from its own gastric juice by a thick layer of mucus that covers the stomach lining. *Helicobacter pylori* takes advantage of this protection by living in the mucus lining.

Case Study: H. pylori

- Spiral shaped bacterium
- Found in the stomach and duodenum, in the thick layer of mucus covering the stomach lining
- Protected from gastric juice
- Urease enzyme creates local basic environment

UREA → CO2 +2 NH3

- Causes gastritis and stomach ulcers (Warren and Marshall, 1984)



Once *H. pylori* is safely ensconced in the mucus, it is able to fight the stomach acid that does reach it with an enzyme it possesses called urease. Urease converts urea, of which there is an abundant supply in the stomach (from saliva and gastric juices), into bicarbonate and ammonia, which are strong bases. This creates a cloud of acid neutralizing chemicals around the H. pylori, protecting it from the acid in the stomach. The reaction of urea hydrolysis (urea is broken down to ammonia and carbon dioxide) is shown. This reaction is important for diagnosis of H. pylori by the breath test. (from www.hpylori.com)

Marshall and Warren were able to demonstrate a strong association between the presence of H pylori and the finding of inflammation on gastric biopsy (Marshall & Warren, 1984). People who did not have gastritis did not have the organism, a finding confirmed in a number of studies. Marshall elegantly fulfilled Koch's postulates for the role of *H. pylori* in antral gastritis with self administration of *H. pylori*, and also showed that it could be cured by use of antibiotics and bismuth salts. (from www.jr2.ox.ac.uk)

Another defense H. pylori has is that the body's natural defenses cannot reach the bacterium in the mucus lining of the stomach. The immune system will respond to an *H. pylori* infection by sending white cells, killer T cells, and other infection fighting agents. However, these potential *H. pylori* eradicators cannot reach the infection, because they cannot easily get through stomach lining. Extra nutrients are sent to reinforce the white cells, and the H. pylori can feed on this. Within a few days, gastritis and perhaps eventually a peptic ulcer results. It may not be H. pylori itself which causes peptic ulcer, but rather the inflammation of the stomach lining; i.e. the response to *H. pylori*.

Clinical Significance of *H. pylori*

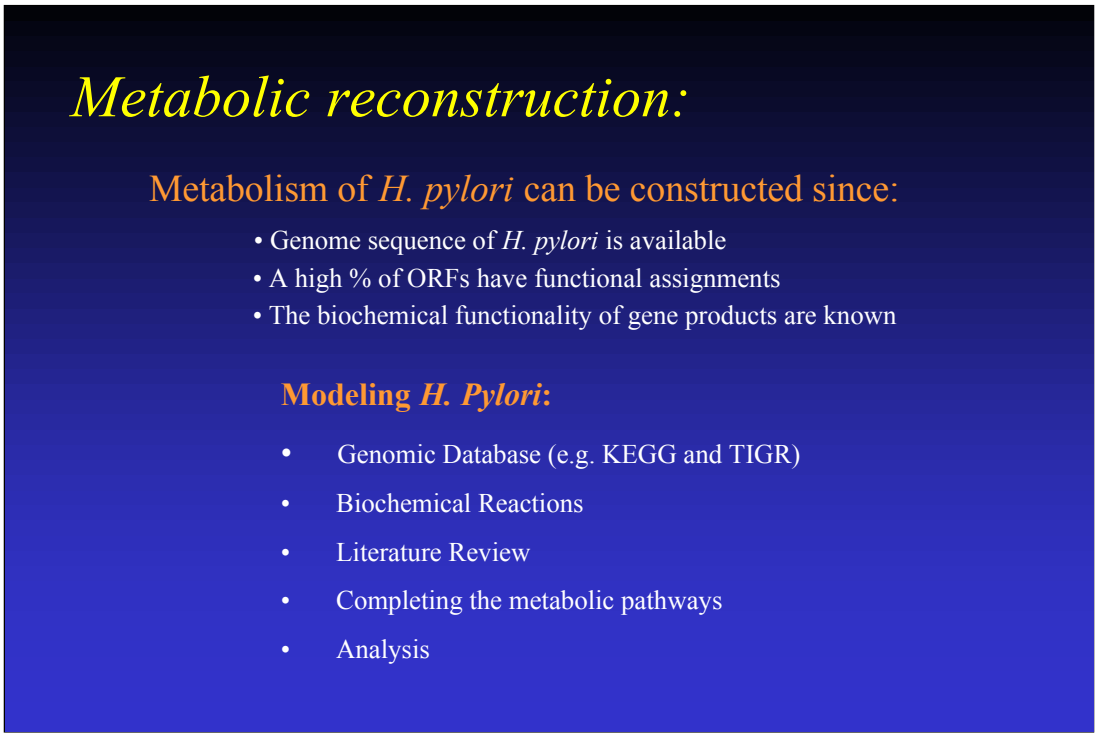
- Immune response cannot reach the infection through stomach lining
- Immune response buildup degrades stomach lining cells (superoxide radicals) – gastritis or peptic ulcers can result within days
- *H. pylori* feeds on nutrients sent to reinforce the white cells
- Carried by >50% of world's population, favoring the poor (Third World countries) and the elderly
- Famous victims: James Joyce , Ayatolla Komheini , George Bush , Pope John Paul II , Imelda Marcos , Stonewall Jackson all had *H.pylori*

H. pylori is believed to be transmitted orally. Many researchers think that *H. pylori* is transmitted orally by means of fecal matter through the ingestion of waste tainted food or water. In addition, it is possible that *H. pylori* could be transmitted from the stomach to the mouth through gastro-esophageal reflux (in which a small amount of the stomach's contents is involuntarily forced up the esophagus) or belching, common symptoms of gastritis. The bacterium could then be transmitted through oral contact.

In general, the following statements can be made to summarize prevalence of *H. pylori* in Western countries:

- *H. pylori* affects about 20% of persons below the age of 40 years, and 50% of those above the age of 60 years.
- *H. pylori* is uncommon in young children.
- Low socio-economic status predicts *H.pylori* infection.
- Immigration is responsible for isolated areas of high prevalence in some Western countries.

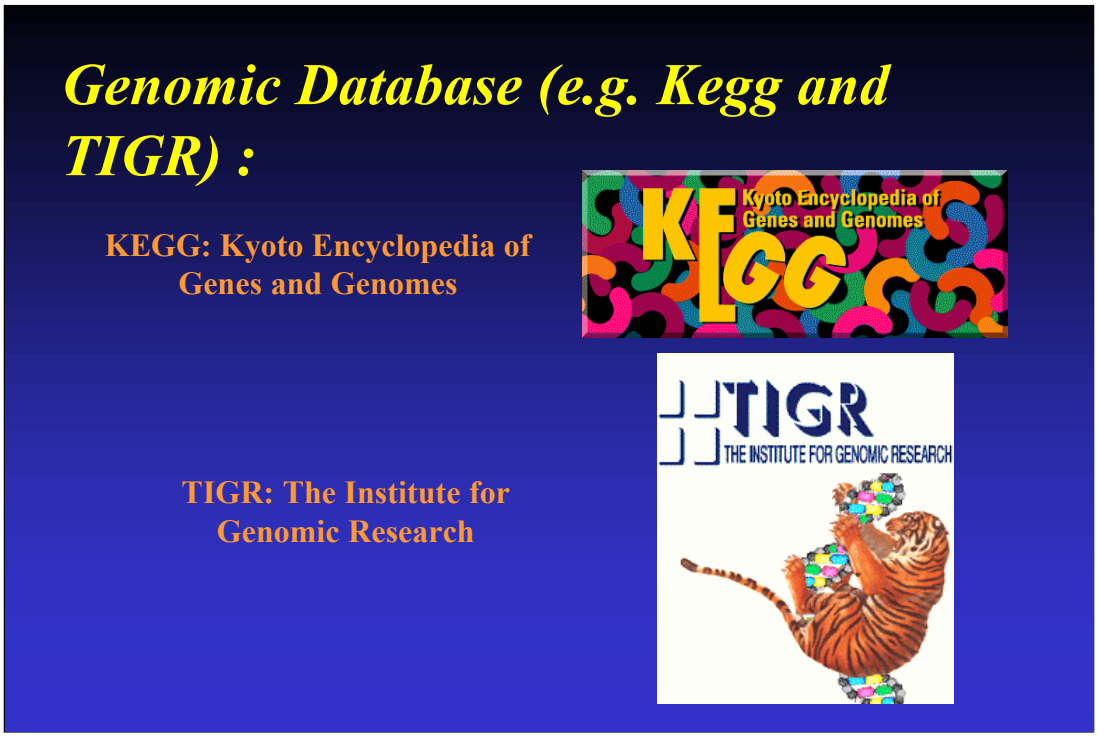
In developing countries most adults are infected. Acquisition occurs in about 10% of children per annum between the ages of 2 and 8 years so that most are infected by their teens. It is evident from careful surveys that the majority of persons in the world are infected with *H.. pylori*. (from www.hpylori.com)



RECONSTRUCTING THE METABOLIC NETWORK

The basis of the metabolic model we will construct for *H. pylori* is genomic data. Constructing this model is only possible if we know most or all of the metabolic reactions which occur in the cell. For *H. pylori*, the genome sequence is finished and available publicly. Furthermore, because most of the open reading frames (ORFs) have been given functional assignments, especially where metabolism is concerned, and because in most cases, we know which reactions are catalyzed by these genes, we are able to make an *in silico* model.

To complete this model will require knowledge of the relevant biochemical reactions in *H. pylori* metabolism and the genes which catalyze these reactions. For this information, we turn to the publicly-available Genomic Databases as well as pertinent literature. Finally, we try to complete the metabolic pathways, inferring the presence of various genes based on experimental data. Each of these steps will be discussed in more detail in the following slides.



MINING DATABASES

Above are details from the home pages of two very useful genomic databases, the Kyoto Encyclopedia of Genes and Genomes (KEGG) and The Institute for Genomic Research (TIGR). Their websites are:

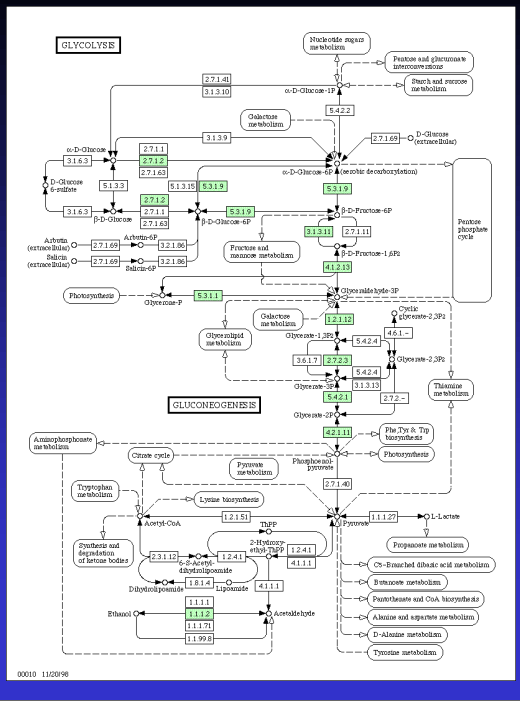
KEGG: www. kegg.com

TIGR: www.tigr.org

It is instructive to surf these sites on your own and become familiar with them. They contain the fully sequenced genomes of many organisms, including *H. pylori*. In many cases, the ORF assignments are also found in these databases, as well as functionality. Both sites organize the known genes by locus number (location on the DNA strand), functionality, and gene name, making it very easy to find genes of interest.

KEGG: Kyoto Encyclopedia of Genes and Genomes:

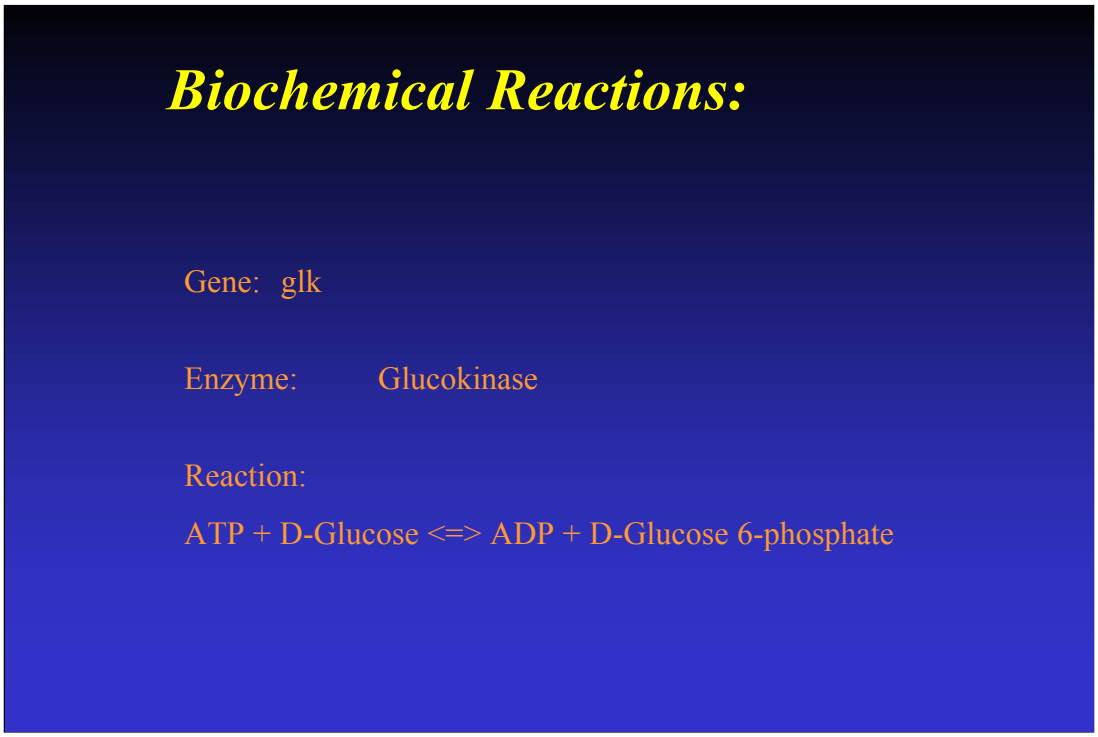
- Genes
- Gene Products
- Metabolic Pathways



THE IMPORTANCE OF METABOLIC MAPS

One interesting way KEGG uses to organize its genomic information is by using these reaction network "maps". The above picture is not so clear, so we recommend that you enter the KEGG website and view it on your own. The above map shows glycolosis. Arrows connect various metabolites to each other, indicating that one metabolite can be converted to another in a reaction. The boxes which stand beside the arrows are the enzymes which catalyze these reactions.

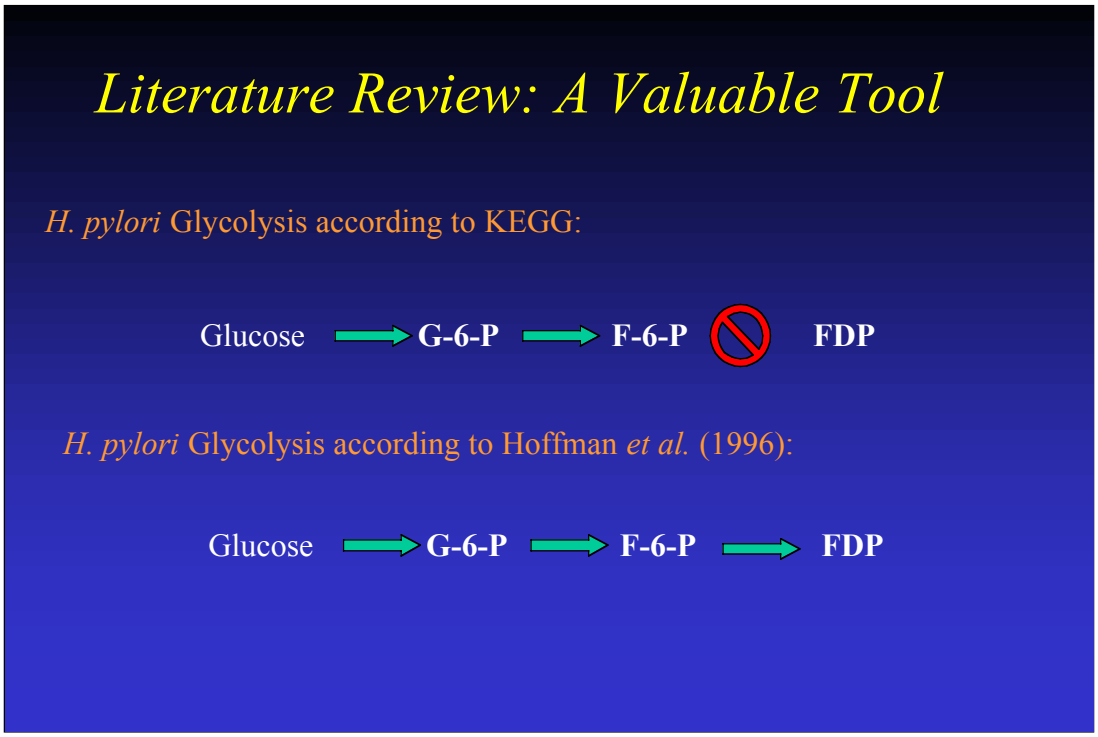
KEGG uses the same maps for many organisms, so not all of the pathways shown in this map are actually available to *H. pylori*. Some are for *E. coli*, for example. The genes actually found in *H. pylori*, according to this map, are the ones which are highlighted in green.



THE CHEMICAL REACTION EQUATION

For example, the enzyme which catalyzes the above reaction, D-Glucose converting to D-Glucose-6-phosphate as ATP is converted to ADP, is called Glucokinase. The gene which encodes this enzyme is commonly called *glk*.

If we were trying to determine whether or not glycolysis occurred in *H. pylori*, we would search in KEGG and TIGR for the relevant genes. The gene *glk* would be found in both of these databases. Once this gene had been positively identified, preferably by both web-based sources, we would add the enzyme that this gene encodes and include its corresponding reaction to our model.

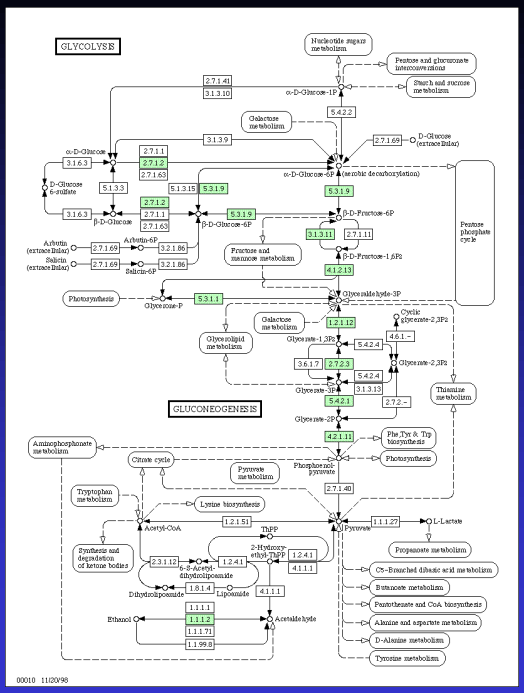


THE NEED FOR USING
PHYSIOLOGY AND INFERRING REACTIONS

Although the model has been mostly determined using various computer databases to find annotated genes, it is not yet complete. Careful study will show the absence of enzymes catalyzing reactions which most likely occur in the thriving organism. In these cases, where the enzyme has not yet been identified, we review the relevant literature to see if various research groups have determined the presence or absence of particular enzymes. For example, in the above case, both KEGG and TIGR give no indication that phosphofructokinase is found in *H. pylori*. This could mean that *H. pylori* is not able to produce 1,6-Fructosebisphosphate (FDP) from Glucose, although there may be other pathways by which FDP is produced.

Careful review of the literature reveals that the Phosphfructokinase enzyme may have been identified by Hoffman et. al. in 1996. Other scientists, however, dispute this claim. After thoroughly examining studies of *H. pylori* metabolism, we will decide whether or not to include this enzyme and the reaction it catalyzes into our model..

Filling in the Gaps



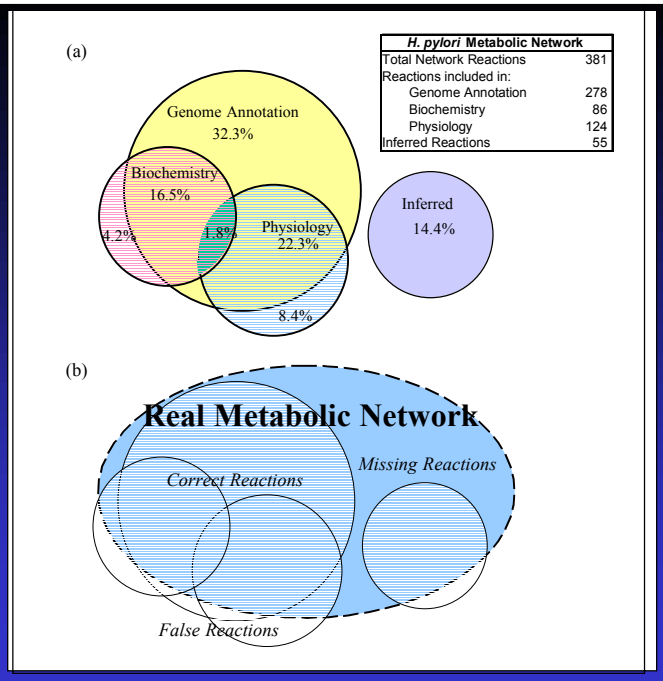
THE NEED FOR USING PHYSIOLOGY AND INFERRING REACTIONS, CONT'D

Finally, even after we have searched the on-line databases and all of the relevant literature, there is still a high probability that several necessary reactions will be missing from the model. This is because the ORFs for the genes in the genome have not yet been identified and/or linked to these reactions. This is one of the most exciting parts of building a model, because we will decide, based on our own knowledge of how *H. pylori* grows, determine that a gene is present simply because it must be present to for *H. pylori* to function as has been determined experimentally.

By "filling in the gaps" in this way, we have the potential to drive further genomic research, determining the presence of genes *in silico*.

The reaction complement of a reconstructed network

Issues of completeness and false members of reaction complement for poorly characterized organisms



Regarding the construction and analysis of microbial metabolic models, the primary issues relating to construction are that first, not all of the reactions suggested by these models are found directly in the databases or the biochemical literature; and second, not all of the metabolic genes actually present in the genotype are accounted for or even noted in the model, because their functions are as yet undiscovered (see part (b) of the figure). For the reconstructed metabolic network (see part (a) of the figure), a "real metabolic network", (i.e. the actual set of all the relevant reactions that occur in *H. pylori* strain 26695) exists. This network, surrounded by a dashed line, is superimposed on the network defined by our model. The lighter area is the set of all reactions that are found both in strain 26695 and in our model, the "correct" reactions. The enclosed area in white represents "false" reactions that were included in the model but do not actually occur in *H. pylori* strain 26695. These reactions represent mistaken assumptions used in creating the model.

The second issue is the inverse problem: many of the proteins synthesized by the organism are not accounted for in the metabolic reconstruction. These "missing reactions" are shown by the darker area in part (b) of the figure. It is likely that some of the metabolic reactions that are catalyzed by the organism are as yet undiscovered. This implies that functionalities open to the organism are neglected by the model.

Finding Orphan ORFs:

Take gene sequences from other organisms and compare them to all *H. pylori* ORFs

Enzymes included in the in silico H. pylori strain without direct evidence, with locus numbers of ORFs with significant similarity to genes encoding these enzymes in other organisms.

Model Name	Organism	HP Locus	Similarity	Identity
Alanine transaminase	*Schizosaccharomyces pombe*	HP0672	35.54%	25.73%
asparagine transport protein	*Salmonella typhimurium*	HP1017	43.86%	32.63%
Cytidylate kinase	*Sus scrofa (Pig)*	HP0618	41.40%	30.65%
Dihydrofolate reductase	*Leishmania tarentolae*	HP0561	39.59%	30.20%
dihydroneopterin aldolase	*Pneumocystis carinii*	HP1232	41.02%	28.15%
Glutaminase	*Pseudomonas sp. (strain 7A)*	HP0723	54.57%	44.51%
Histidine transporter	*Campylobacter jejuni*	HP0940	40.41%	29.80%
Tetraacyldisaccharide 4' kinase	*Francisella novicida*	HP0328	42.34%	29.20%
Lysine transporter/permease	*Escherichia coli*	HP1017	49.25%	37.10%
Malate dehydrogenase	*Corynebacterium glutamicum*	HP0086	36.81%	25.93%
O-Succinylbenzoate-CoA ligase	*Staphylococcus aureus*	HP1045	33.95%	23.66%
Isochorismate synthase 1	*Pseudomonas aeruginosa*	HP1282	32.58%	21.80%
Aspartate oxidase	*Synechocystis sp.*	HP0192	42.08%	30.94%
Ornithine transaminase	*Escherichia coli*	HP0976	39.17%	27.74%
Phenlyalanine transporter	*Escherichia coli*	HP1017	44.20%	30.64%
Sulfate transporter	*Synechococcus sp. (strain PCC 7942)*	HP0474	38.81%	26.48%
Threonine transporter	*Escherichia coli*	HP0133	50.00%	33.33%
Tryptophan transporter	*Saccharomyces cerevisiae*	HP1017	40.68%	31.94%
5'-Nucleotidase	*Escherichia coli*	HP0104	36.71%	25.76%

These metabolic network reconstruction issues can be resolved in part as the model is applied to various analyses. For example, the metabolic *H. pylori* model was used to reexamine the annotation of the metabolic network. All of the genes that were included in the reconstruction of *H. pylori* metabolism without direct genomic or biochemical evidence can be thought of as hypothetical. The presence of these hypothetical genes can be determined by collecting the sequences of other organisms' copies of the hypothetical genes and using BLAST to compare them with the *H. pylori* genome sequence. The genes that are found to be significantly homologous to loci in the *H. pylori* genome sequence can then be studied experimentally to verify their proposed function based on the reconstruction and BLAST analysis.

Network Reconstruction as a Predictive Science

Enzymes included in the *in silico H. pylori* strain without direct evidence, with locus numbers of ORFs with significant similarity to genes encoding these enzymes in other organisms.

HP Locus	Organism	Gene Product Name	Similarity	Identity
HP0086	*Corynebacterium glutamicum*	Malate dehydrogenase	36.81%	25.93%
HP0104	*Escherichia coli*	5'-Nucleotidase	36.71%	25.76%
HP0133	*Escherichia coli*	Threonine transporter	50.00%	33.33%
HP0192	*Synechocystis sp.*	Aspartate oxidase	42.08%	30.94%
HP0328	*Francisella novicida*	Tetraacyldisaccharide 4' kinase	42.34%	29.20%
			38.81%	26.48%
			39.59%	30.20%
			41.40%	30.65%
			35.54%	25.73%
			54.57%	44.51%
			40.41%	29.80%
			39.17%	27.74%
			43.86%	32.63%
			49.25%	37.10%
			44.20%	30.64%
			40.68%	31.94%
			33.95%	23.66%
			41.02%	28.15%
			32.58%	21.80%

in silico Prediction:

The *H. pylori* Network includes a malate dehydrogenase function

L-Malate + NAD^+ V Oxaloacetate + NADH +H^+

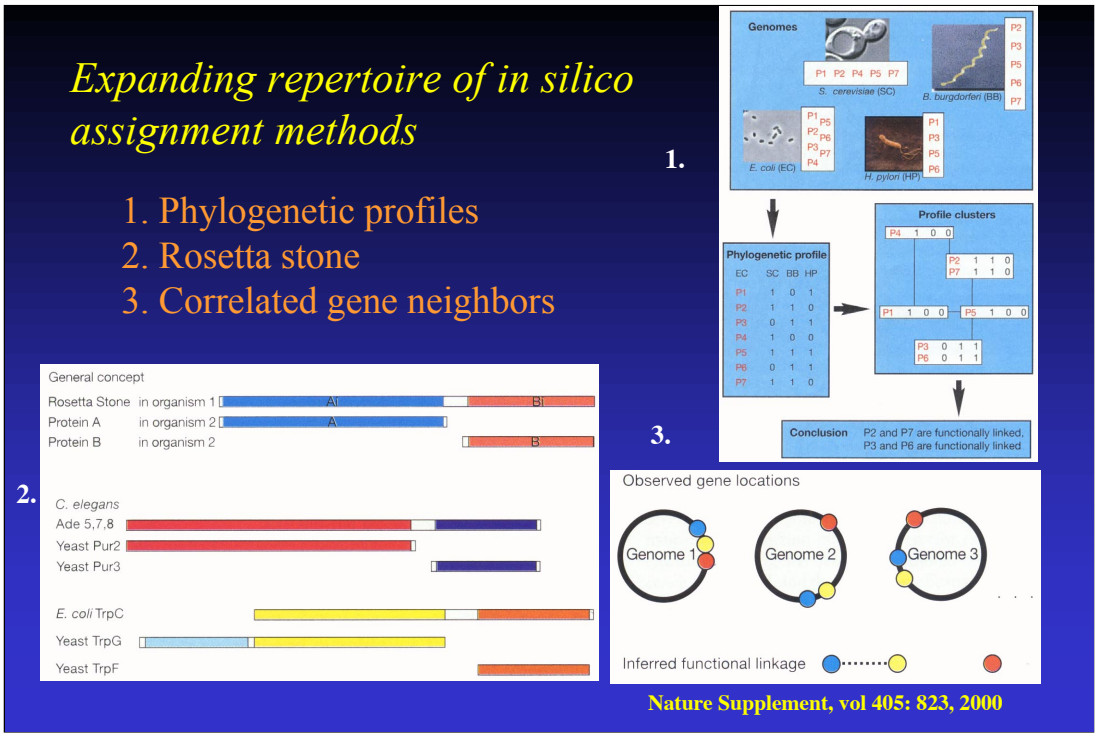
Computational Verification:

BLAST search indicates the presence of a Malate:Quinone Oxidoreductase (MQO) in *C. glutamicum* with significant similarity (36.81%) and identity (25.93%) to locus HP0086 in *H. pylori*.

Biochemical Verification:

Kather et.al. (*J Bact*, June 2000) demonstrate MQO activity of locus HP0086 in *H. pylori*.

One such gene product included in the *H. pylori* model without genomic or biochemical evidence was malate dehydrogenase. A subsequent study indicated that on locus HP0086 of the *H. pylori* genome, an open reading frame was located that showed significant similarity (36.81%) and identity (25.93%) with a malate:quinone oxidoreductase in glutamic acid bacterium *Corynebacterium glutamicum* (ref). Thus, the analysis of microbial metabolic models can also have bioinformatic applications, such as functional assignment of ORFs, in addition to the more obvious experimental applications.

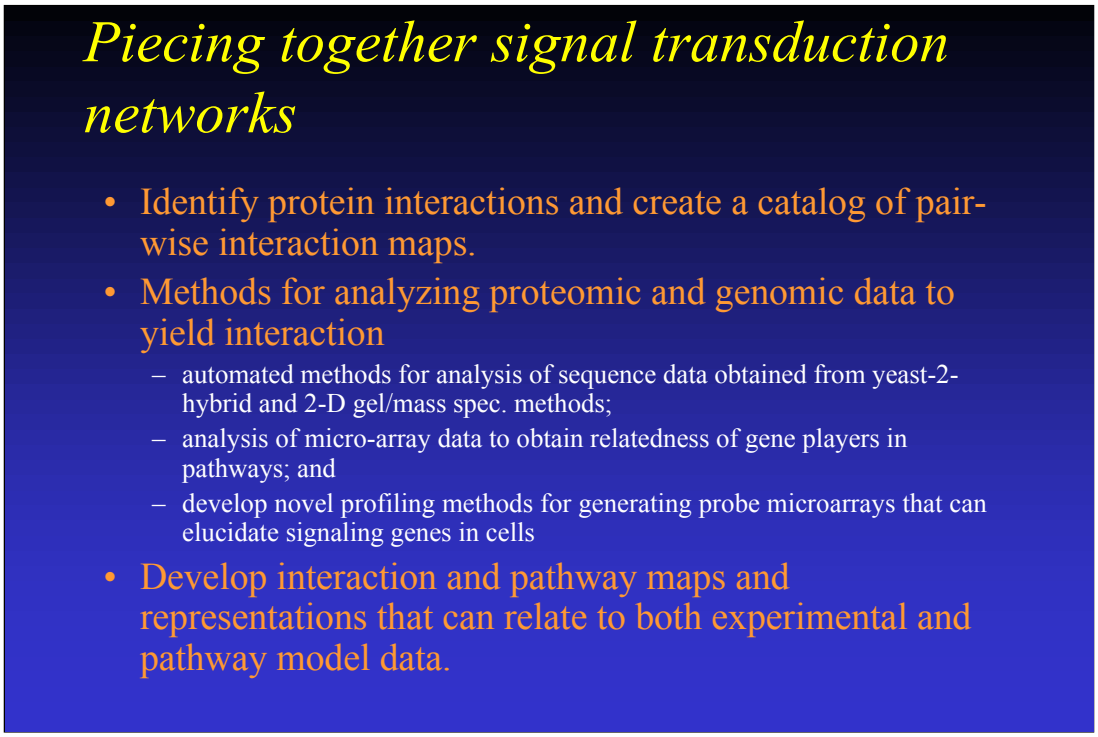


NEW METHODS

Many new methods are now being developed to assign function to ORFs through genome comparison. Some of these methods are illustrated on this slide. They are described in more detail in the reference given in the slide.

Piecing together networks

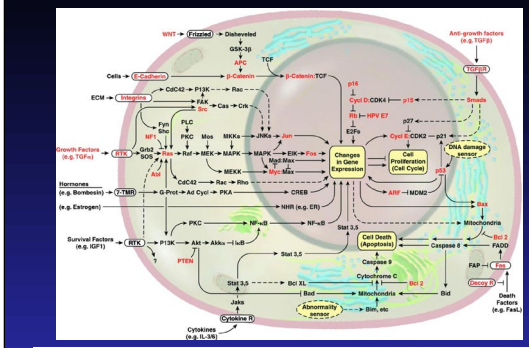
- Make mutants and experimentally determine phenotype
- Expression arrays and cluster analysis
- Computational approach based on co-evolution of protein and analysis of fusion protein (Rosetta Stone)
- Protein-protein interaction maps



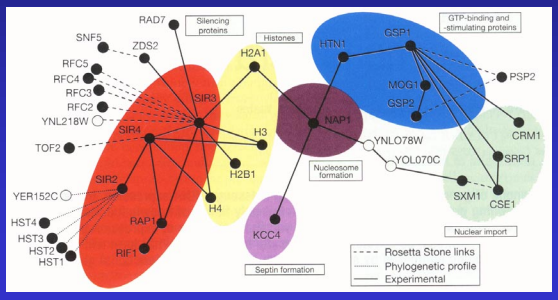
SINGAL TRANSDUCTION NETWROKS

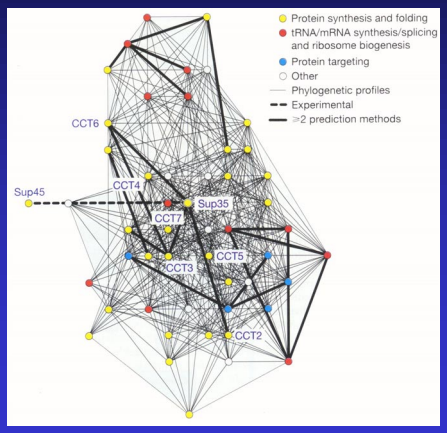
An extremely important step in the construction of signaling pathways in cells is the cataloging of "who talks to whom" vis-à-vis proteins involved in the pathway. The sources of this information are; a) legacy data based on gene knockout and mutant analysis, b) to a small extent gene expression array data, and most importantly c) proteomics data. A large volume of these data exists for *Drosophila*, *C. elegans,* mouse and human and one can create a "validated" catalog of these interactions. Further, one can anticipate increased availability of new genomic and proteomic experimental data that can be mined to obtain protein interaction knowledge. Large-scale study of specific cell types and organisms will likely yield enormous amounts of data pertaining to molecular interaction screens, 2D gel/mass spec experiments, and cDNA expression profiles. Comparative sequence analysis of the proteins identified in the mouse with *Drosophila* is expected to provide a valuable molecular interaction catalog.

Algorithmic methods include: a) extensive schemes to analyze genomic and proteomic data, b) a high throughput pipeline for sequence comparisons across species and c) validation methods to compare diverse sources of data pertaining to specific molecular interactions. Finally, pair-wise interaction data has to be validated in the context of complete pathways and entirely new methods for iterative analysis of interaction pathways can be developed.



Expectation: A combination of in silico and in vitro methods will give arise to network construction



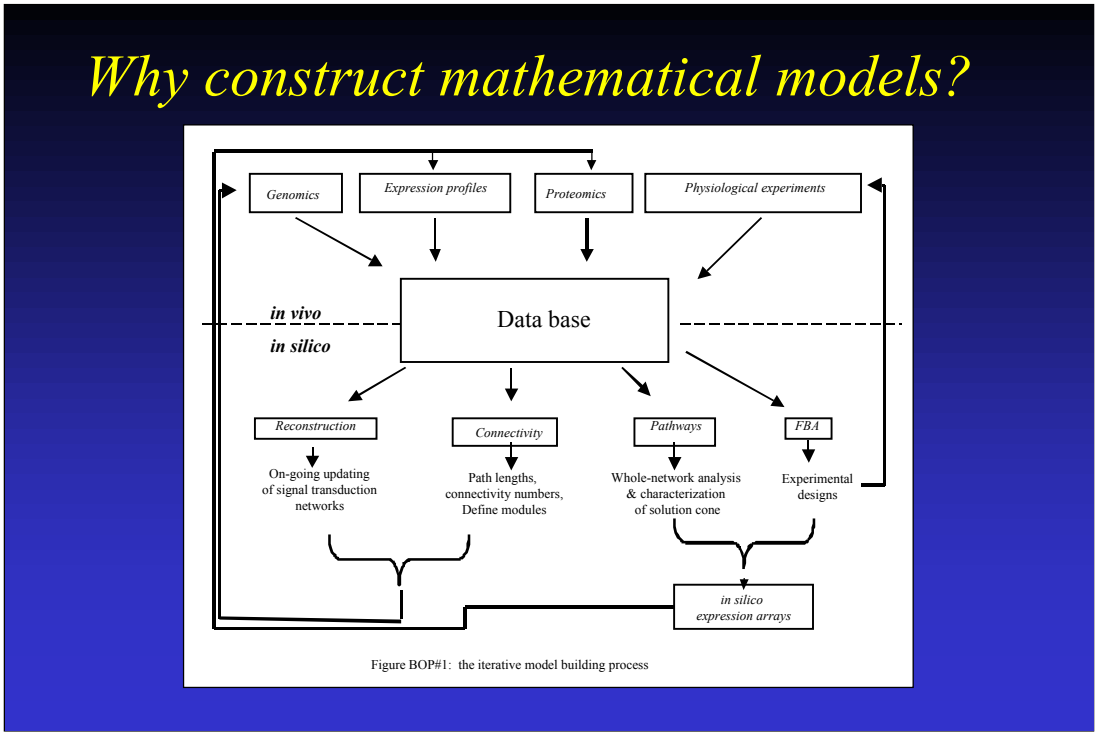


Nature Supplement, vol 405: 823, 2000

TOWARDS RECONSTRUCTED NEWTORKS

The reconstruction of metabolic networks is now at a developed stage. Similar developments are forecasted for signal transduction, and other cellular processes. We can expect that over the coming decade we will develop computer and laboratory methods which will enable us to reconstruct the networks of biochemical interactions that carry out cellular functions.

The challenge is to describe these mathematically.

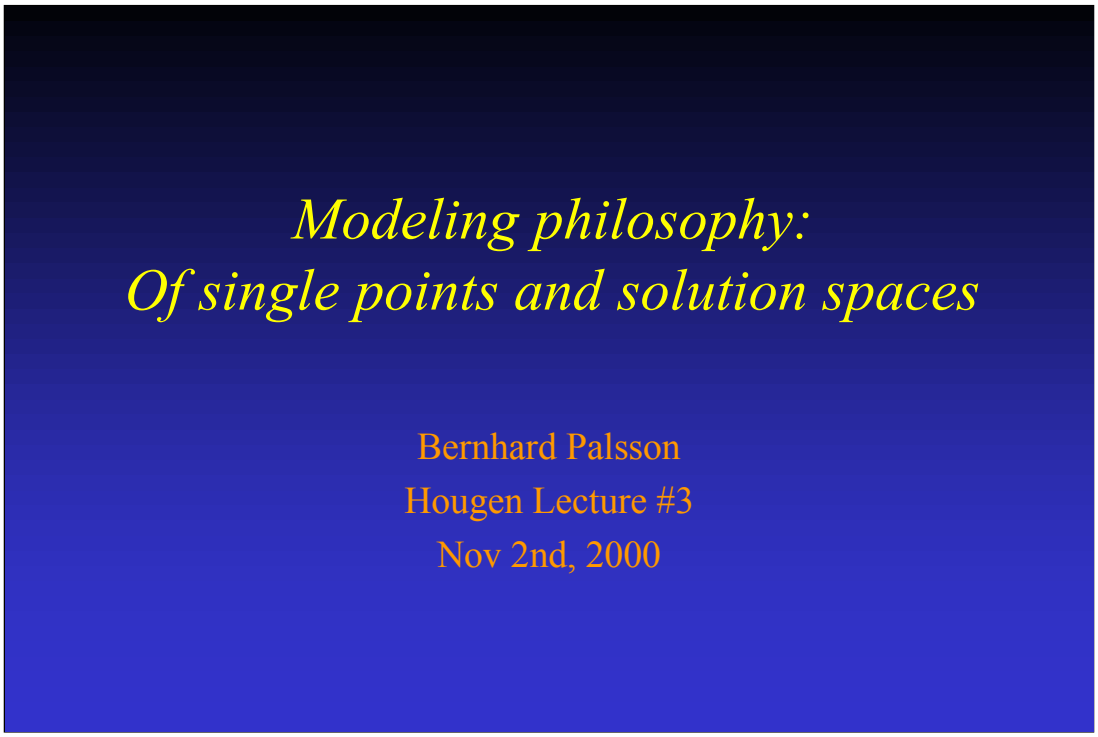


WHY MODEL?

There are many reasons for constructing mathematical models of complex biological processes. Perhaps chief amongst them is to reconcile data and identify missing/incomplete knowledge. This diagram illustrates the iterative process that uses a variety of in vivo and in silico methods to converge on reliable models of cellular and biological activity.

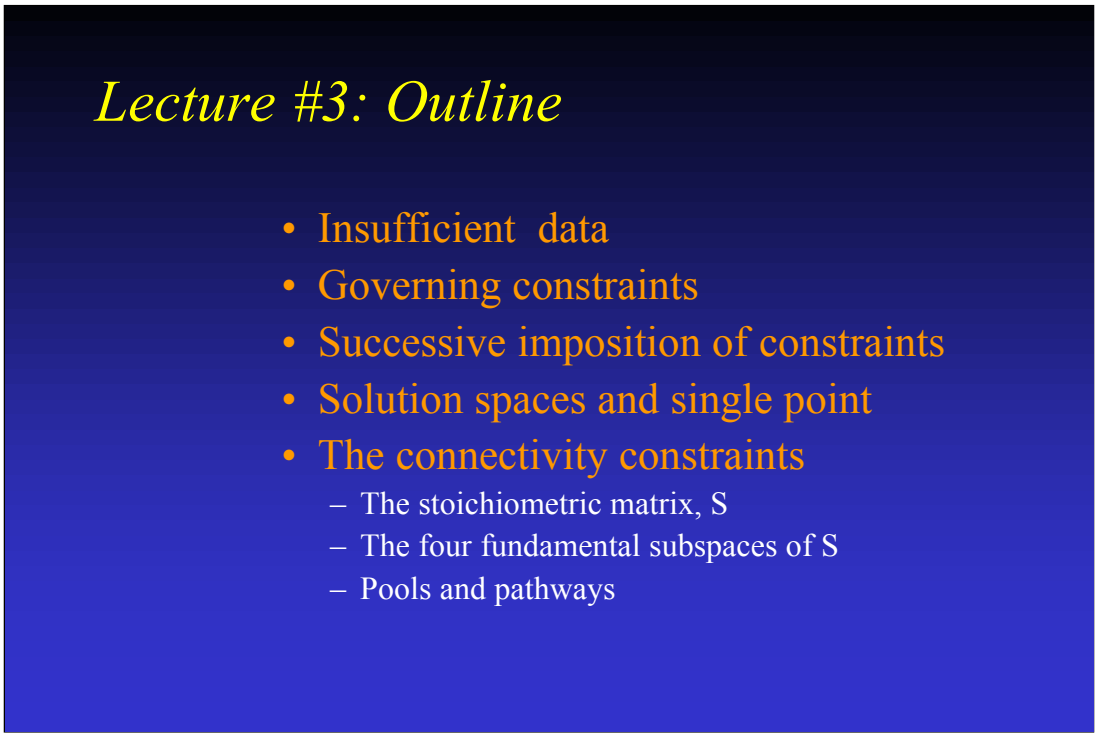
References

- Marshall, B.J. and J.R.Warren, "Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration," *Lancet* **8310**, 1311-1315 (1984).
- Hoffman, PS; Goodwin, A; Johnsen, J; Magee, K; Veldhuyzen van Zanten, SJ. "Metabolic activities of metronidazole-sensitive and -resistant strains of *Helicobacter pylori*: repression of pyruvate oxidoreductase and expression of isocitrate lyase activity correlate with resistance," *Journal of Bacteriology*, **178** :4822-9 (1996).
- Kather, B; Stingl, K; van der Rest, ME; Altendorf, K; Molenaar, D., "Another unusual type of citric acid cycle enzyme in *Helicobacter pylori*: the malate:quinone oxidoreductase," *Journal of Bacteriology*, **182**: 3204-9 (2000).
- Schwikowski, B., Uetz, P., and Fields, S., "A network of protein-protein interactions in yeast," *Nature Biotechnology*, **402**: 1257-61 (2000).
- Uetz, P., Giot, L. Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y.,Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.M. , "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae* ," *Nature*, **403** :623-7 (2000).
- Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S.H. , "Functional discovery via a compendium of expression profiles," *Cell*, **102** :109-26 (2000).
- Eisenberg, D.,Macotte, E.M., Xenarios, I., and Yeates, T.O., "Proteomics in the post-genomic era," *Nature*, **405**: 823-826 (2000).
- Covert, M.W., Schilling, C.H., Famili, I., Edwards, J.S. , Goryanin, I.I., Selkov, E. and Palsson, B.O., "Metabolic modeling of microbial stains in silico," *Trends in Biochemical Sciences*, **26**: 179-186 (2001).



LECTURE #3

The first two lectures discussed the high-throughput technologies and the subsequent determination of cellular component catalogs and reconstruction of biochemical reaction networks. In the third lecture we begin to discuss how one describes the function of such networks in systemic and mathematical terms.



OUTLINE

In spite of the impressive amounts of data that are being generated about cells and their components we do not have all the data that is needed to construct detailed mathematical models of their integrated function. The approaches often used in physicochemical and engineering sciences of stating governing fundamental laws and building detailed mathematical models will thus not work, at least not initially, for the construction of mathematical models of reconstructed biochemical reaction networks. We cannot thus calculate a single solution.

An alternative approach must be developed. The network maps allow us to impose systemic and component constraints on the function of the network as a whole. Thus we can eliminate behaviors but we cannot calculate precise ones. The more governing constraints that we can state the smaller the solution spaces become.

The lecture then ends with a detailed discussion of the consequences of the stoichiometric constraints.

Coping with incomplete constraints: solution spaces vs. single points

--Cannot describe cellular networks in the same detail as we are used to in the P/C sciences

--However, we can subject the networks to known constraints and analyze them given these constraints

3 Problems

•P/C laws may not apply

•Kinetic constraints not known

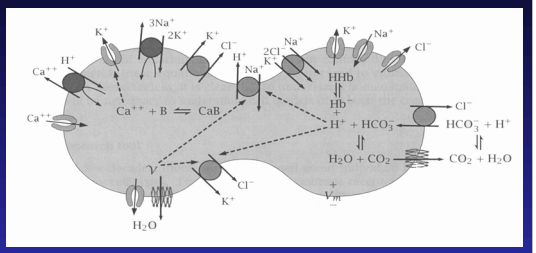
•Even if they were, they

1) Change with time-->evolution

2) Not the same from one individual to the next-->SNPs

Factors Constraining Metabolic Function

- Connectivity:
  - Systemic stoichiometry
- Capacity:
  - Maximum fluxes
- P/C factors:
  - osmotic pressure, electro-neutrality, solvent capacity, molecular diffusion
- Rates:
  - Mass action, Enzyme kinetics, Regulation



CONSTRAINTS

Metabolism is subject to a number of constraints. First, fluxes are balanced in the steady state. For many dynamic metabolic states, the solution does not move far from the steady state. There is an upper limit on the amount of flux that is achievable through every reaction. First there is an upper P/C constraint, a crowding constraint limiting the amount of enzyme present and finally upper limits may be derived from expression and proteomic profiles.

There are a number of physico-chemical constraints that a cell must operate under. These include balancing of osmotic pressure (unless there is a cell wall), maintaining electro-neutrality since charges cannot be separated, the limited solvent capacity of water (i.e. the 30% of cells that is biomass must be divided amongst the thousands of cellular constituents) , and the rate of molecular diffusion limits almost all cellular functions.

Finally, the kinetic parameters that have evolved and the imposed regulatory mechanisms significantly influence the flexibility of the network. These are flexible and the cell can adjust them.

The connectivity and P/C constraints are 'hard' in the sense that the cell cannot manipulate them, the capacity constraints represents fixed upper limit constraints that can be down regulated, while the kinetics may be quite flexible and adjustable by the cell through a evolutionary process.

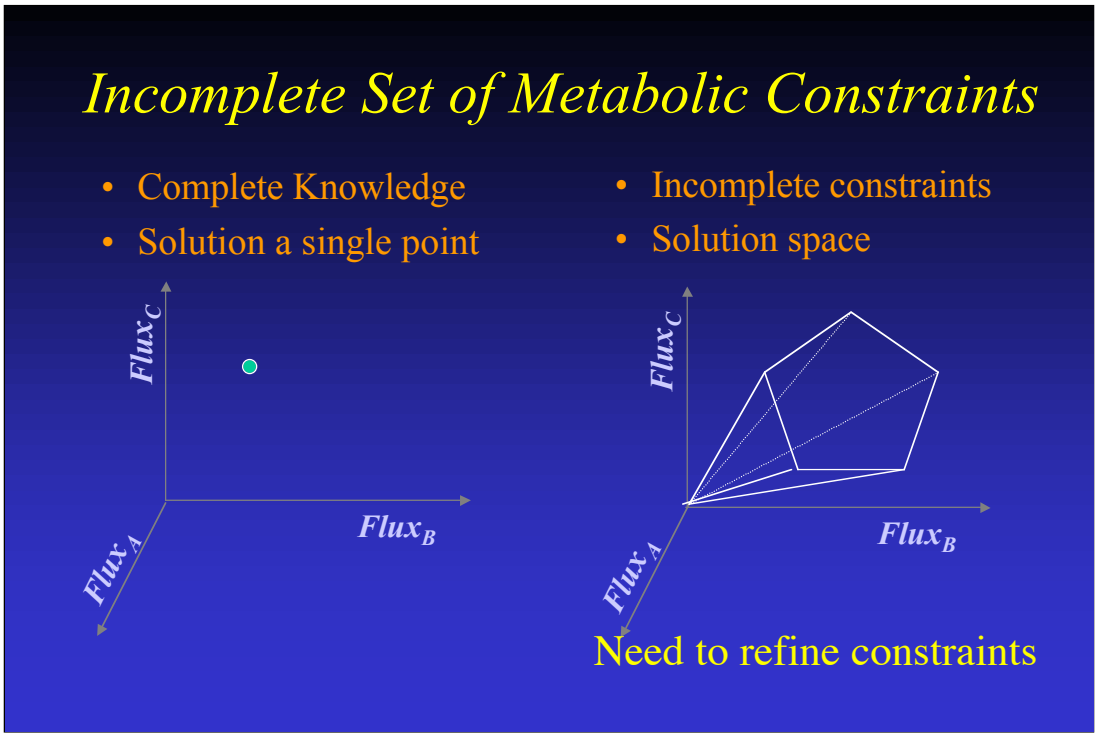
Factors Constraining Metabolic Function

- Connectivity:
  - Systemic stoichiometry
  - Sv=0
- Capacity:
  - Maximum fluxes
  - v_i < maximum value

HARD CONSTRAINTS

In these lectures we will impose two sets of constraints to study possible metabolic functions.

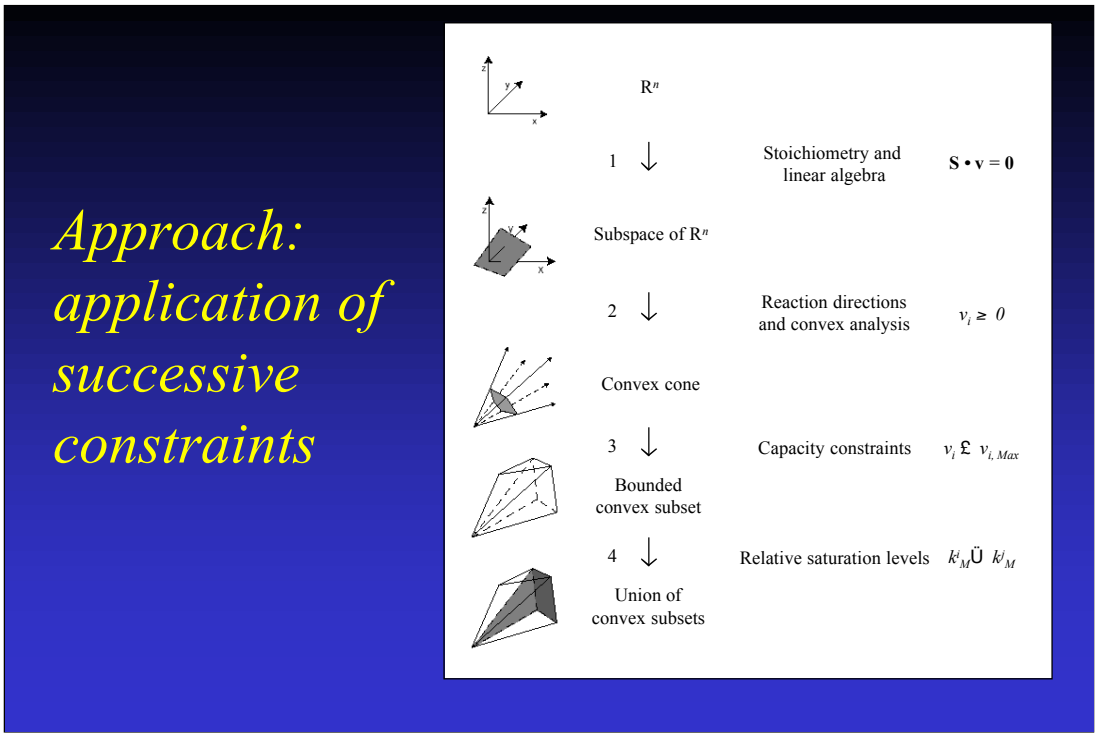
These are the connectivity constraints and the capacity constraints. In Lecture #3 we will cover the consequences of the imposition of the stoichiometric constraints.



CONSEQUENES OF OUR MODELING PHILOSPHY

Normally when we solve a mathematical problem or construct a mathematical model we are looking for 'the solution.' The search for such a solution comes down to a detailed and complete problem statement, and then the use of mathematical or numerical methods to find 'the' solution. It is represented as a single point in the left side of the figure.

However, we are in a situation where we cannot fully define and describe the interior of a cell in all its details. We thus must be content with 'bracketing' the solution. The imposition of governing constraints then eliminates impossible solutions but leaves a range of possible solutions. This range is represented by a solution space that contains all these possible solutions. The more applicable constraints that we find the smaller the solution space.

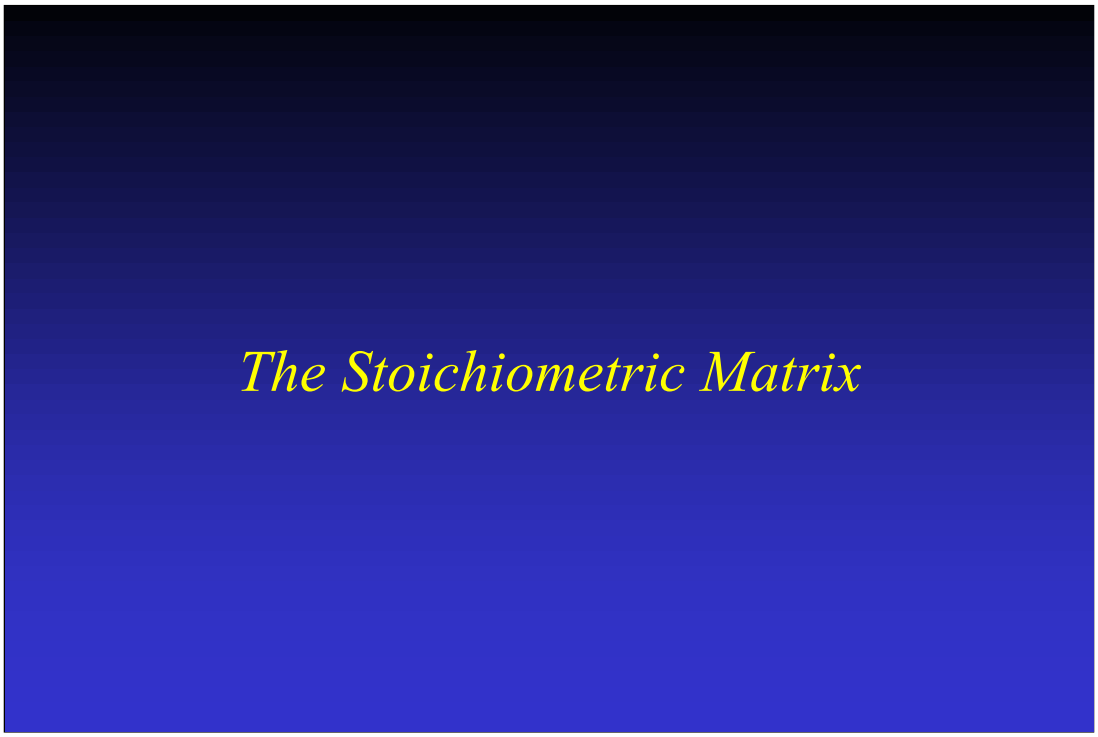


GRAPHICAL ILLUSTRATION OF THE SUCCESSIVE APPLICATION OF GOVERNING CONSTRAINTS

Some years ago it was common to think of each gene/gene product in a cell as an independent element. Genetic engineering came into being and the expectation was that if one would splice a gene into a genome a trait that corresponding to that gene product would be produced. Mathematically, one can represent this as an n-dimensional space (where n is the number of gene products) and any point in this space could be attained.

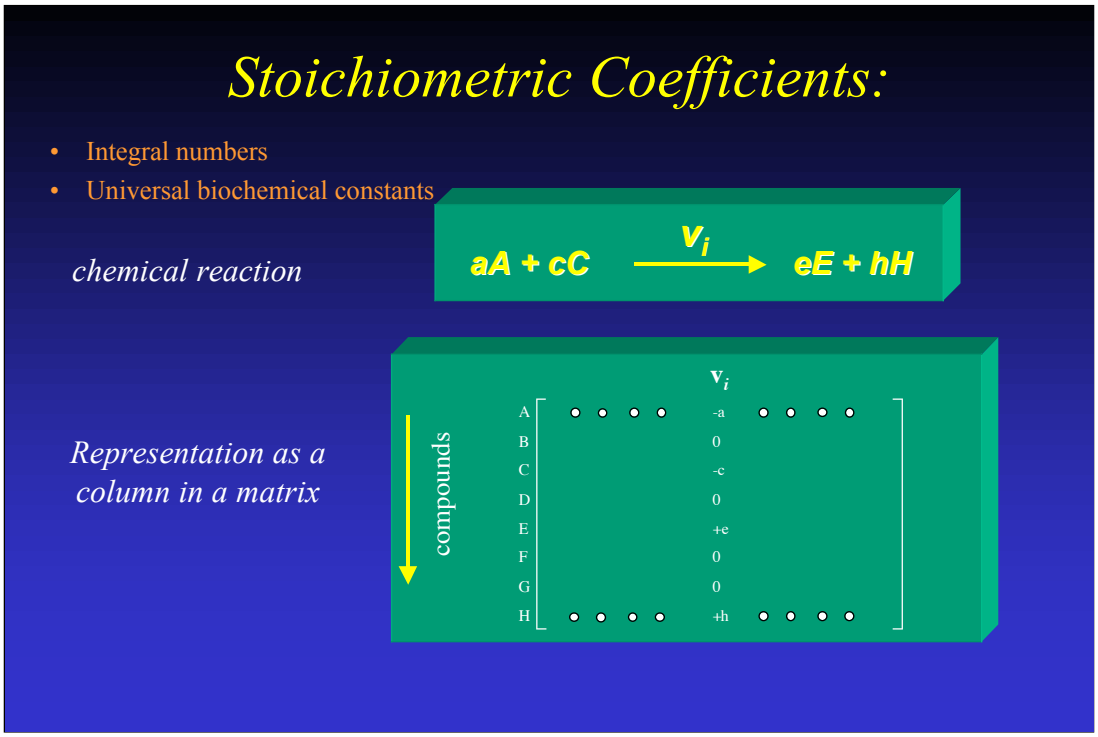
However, every gene product works in the context of many others and is thus constrained in its activity. For instance once can over express an enzyme in a linear pathway and get no increase in flux down the pathway since the flux through all the steps has to be the same. Such connectivity, or stoichiometric constraints reduce the accessible space to a subspace, or a 'hyper-plane' as illustrated. The 'size' of this hyper plane is substantially smaller than the n-dimensional space. Thus these constraints limit the attainable behaviors.

A hyper-plane is infinite in all directions. If we consider all reactions to have positive fluxes (so reversible reactions are represented as two irreversible reactions) the hyper-plane is converted to a semi-finite conical solution space. If we then impose the maximum flux constraints then the solution space is 'capped off' and becomes a 'lock-box' for the solution. This lock box is formed based on hard constraints. Certain kinetic constraints drive the solution to the edge as shown later. These represent adjustable constraints.



THE MATRIX S

For the rest of this lecture we shall discuss the consequences of the connectivity constraints in metabolism, namely stoichiometry

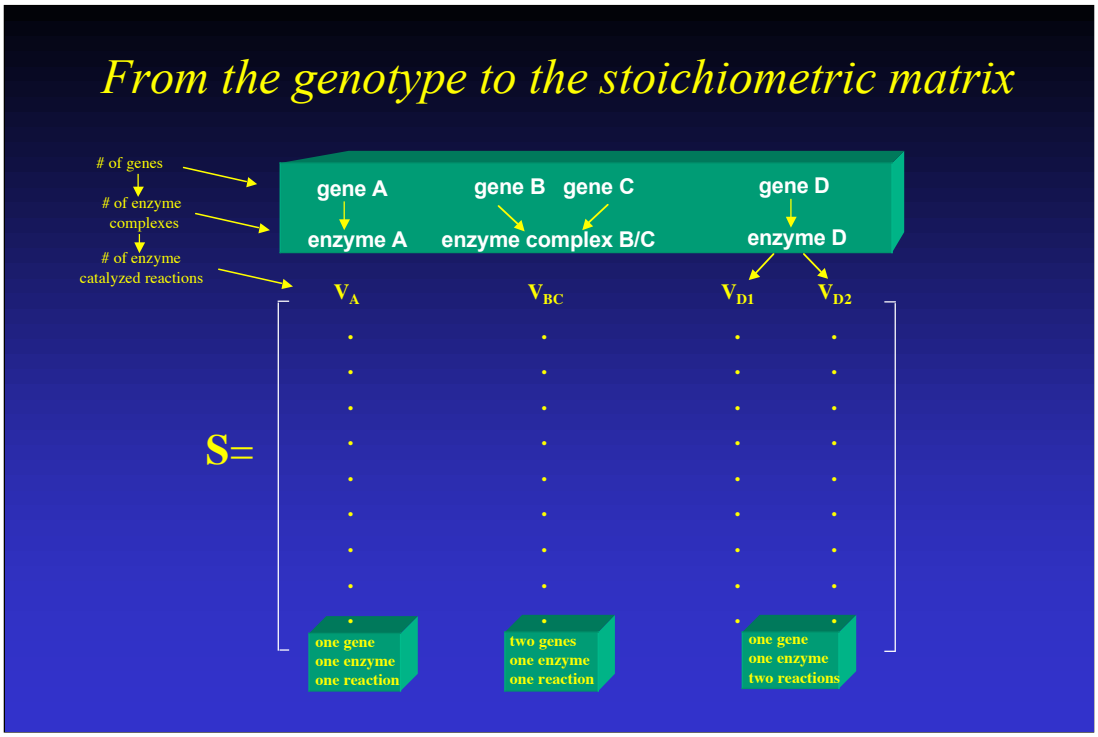


EACH COLUMN IN THE STOICHIOMETRIC MATRIX CORRESPONDS TO A PARTICUALR METABOLIC BIOCHEMICAL REACTION

The stoichiometric coefficients: They are integers (a,c,e,h in the example given) that represent the number of molecules of chemical species (A,C,E,H in the examples) that are transformed in this particular chemical reaction. These coefficients are constants (i.e. are not condition dependent, that is functions of temperature, pressure, pH, etc). Further they are biologically universal, that is the same metabolic reaction proceeds the same way in all cells; for instance hexokinase always catalyzes the reaction:

Glucose + ATP --> Glucose-6-phosphate + ADP

Formation of a column in S: Each metabolite has a row in the stoichiometric matrix, and each reaction has a column. The stoichiometric coefficients are used to form a column, with the stoichiometric coefficient that corresponds to a particular metabolite appearing in the row that it corresponds to. If a metabolite is formed by the reaction the coefficient has a positive sign, if it is consumed by the reaction the stoichiometric coefficient appears with a negative sign. All other rows (corresponding to metabolites that do not participate in the reaction) are zero.



THE NUMBER OF REACTIONS IN A METABOLIC GENOTYPE IS NOT THE SAME AS THE NUMBER OF GENES IN THE GENOTYPE

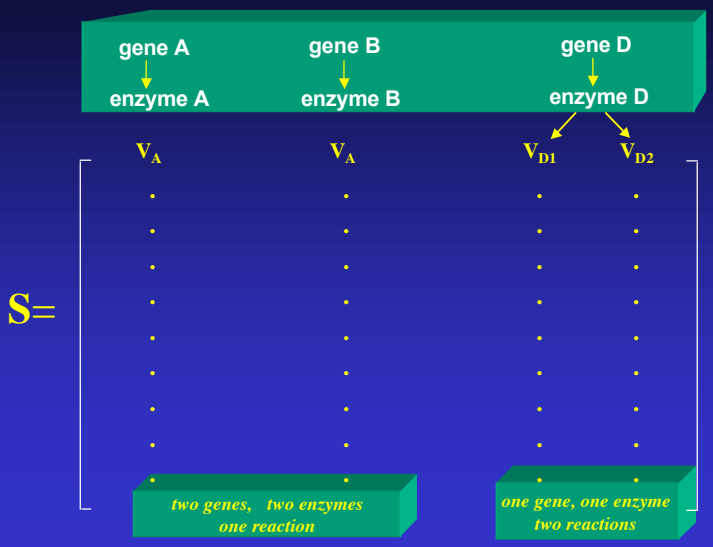
There is not a one-to-one correspondence between the number of genes that are associated with metabolism and the number of chemical transformations that take place. This difference is due to several factors.

First, many enzymes are oligomeric complexes that contain more than one protein chain. These complexes are formed by non-stoichiometric binding, or association of several different protein molecules. Hemoglobin, being a tetramer of two alpha and two beta globins is perhaps the best know example of a protein oligomer.

Second, enzymes can catalyze more than one chemical reaction. This feature is often referred to as substrate promiscuity. These chemical transformations tend to be similar.

These features give rise to a different number of genes from the number of enzymes (or enzyme complexes) and the number of chemical reactions that take place. All of these situations can though be accounted for with the stoichiometric matrix as illustrated.

Redundancy and pleiotrophy in the stoichiometric matrix



Partitioning of the flux vector into internal and external fluxes

- External fluxes are those fluxes that flow across the cellular boundary.
  - These are denoted by b_i. These fluxes are often accessible to measurement or can be estimated based on experimental data. The sign convention adopted for these fluxes is that they are positive if mass is flowing out of the cell.
- Internal fluxes are those that take place with in the cell (within our system boundary).
  - These fluxes are hard to measure, but often we will know their maximum value.

PARTITIONING THE FLUX VECTOR

We draw a systems boundary around the metabolic system that we are interested in. Thus there will be reactions that take place within the system and those that exchange molecules with the surroundings. We partition the flux vector accordingly.

Normally the system boundary is drawn such that the metabolic system being considered is the entire metabolic system in a cell. Then the system boundary effectively becomes the cell membrane. In other cases we may be interested in an organelle, such as the mitochondrion, and we will draw our system boundary around it. In yet other cases we draw system boundaries around certain sectors of metabolism, such as the fueling reactions, or the amino acid synthetic pathways. In such cases the system boundary is a conceptual one and not a physical one.

The concept of a 'system boundary' is frequently used in the physical and engineering sciences, while for life scientists reading these notes may be a new one. It may take some getting used to.

Chemical Reactions vs. Fluxes through them

- The columns of the stoichiometric matrix represent the reactions (n in number)
- The actual reaction rates, or the fluxes that take place through these reaction are denoted by a **v**
- The assignment of a flux through a reaction can be performed by a simple matrix multiplication

$$\mathbf{S} \cdot \mathbf{v} \qquad \begin{vmatrix} S_{11} & \cdots & S_{1n} \\ \vdots & & \vdots \\ S_{m1} & \cdots & S_{mn} \end{vmatrix} \begin{vmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n_i} \\ b_j \\ \vdots \\ b_n \end{vmatrix}$$

METABOLIC REACTIONS AND THE FLUXES THROUGH THEM

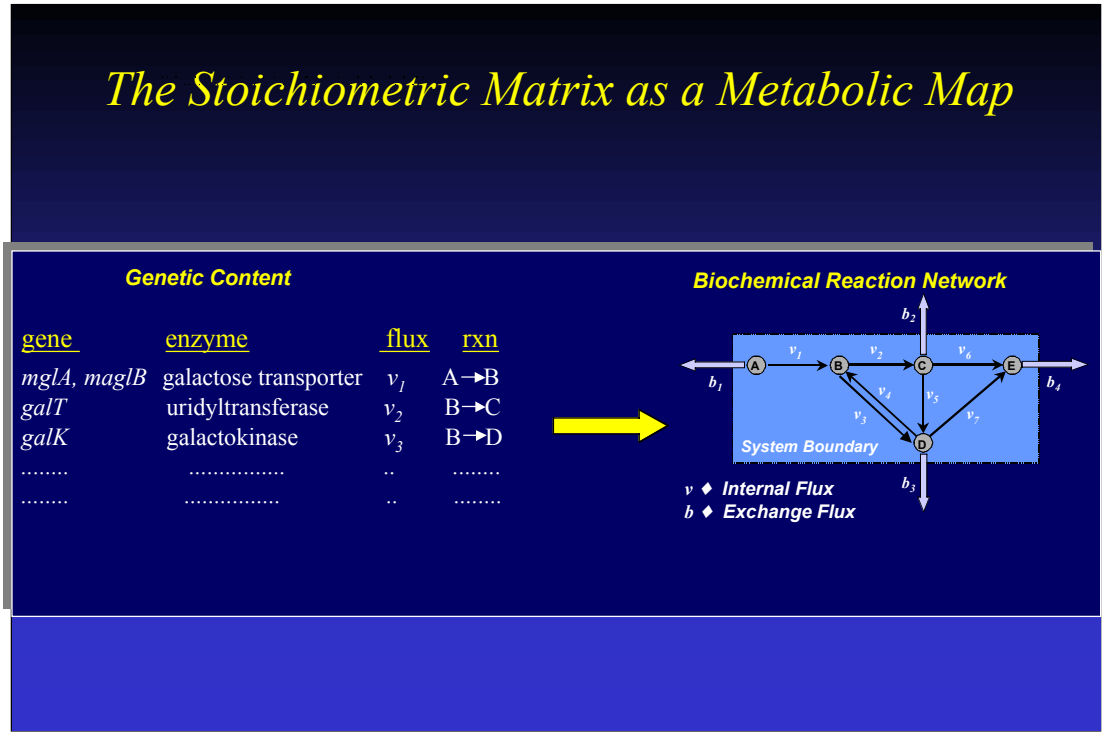
The annotated sequence and biochemical knowledge of the metabolic enzymes lead to the definition of the stoichiometric matrix. Each column in this matrix represents a particular metabolic reaction. However, the flux though a reaction is highly dependent on what the cell is doing. For instance, if an amino acid is available to the cell, it will import it and not synthesize it. Although the cell is capable of carrying out all the reactions that lead to the synthesis of the amino acid they are not used. The flux through them is zero. Later we will see how the cell regulates flux (either by kinetic means or by regulation of gene expression), but for now we introduce the product of the stoichiometric matrix and the flux vector**.** The matrix is a constant, while the flux vector is a variable.

The number of genes, enzymes and metabolic reactions for some gram-negative bacteria

	E. coli	*H. influ.*	*H. pylori*	*Yeast*
Total # of Genes	4288	1743	1590	6259
# of metabolic genes	660	400	290	697
# of metabolic enzymes	697	412	272	626
# of metabolic reactions	739	461	381	1212
# of metabolites	442	367	332	569

ACTUAL NUMBERS FOR ACTUAL ORGANISMS

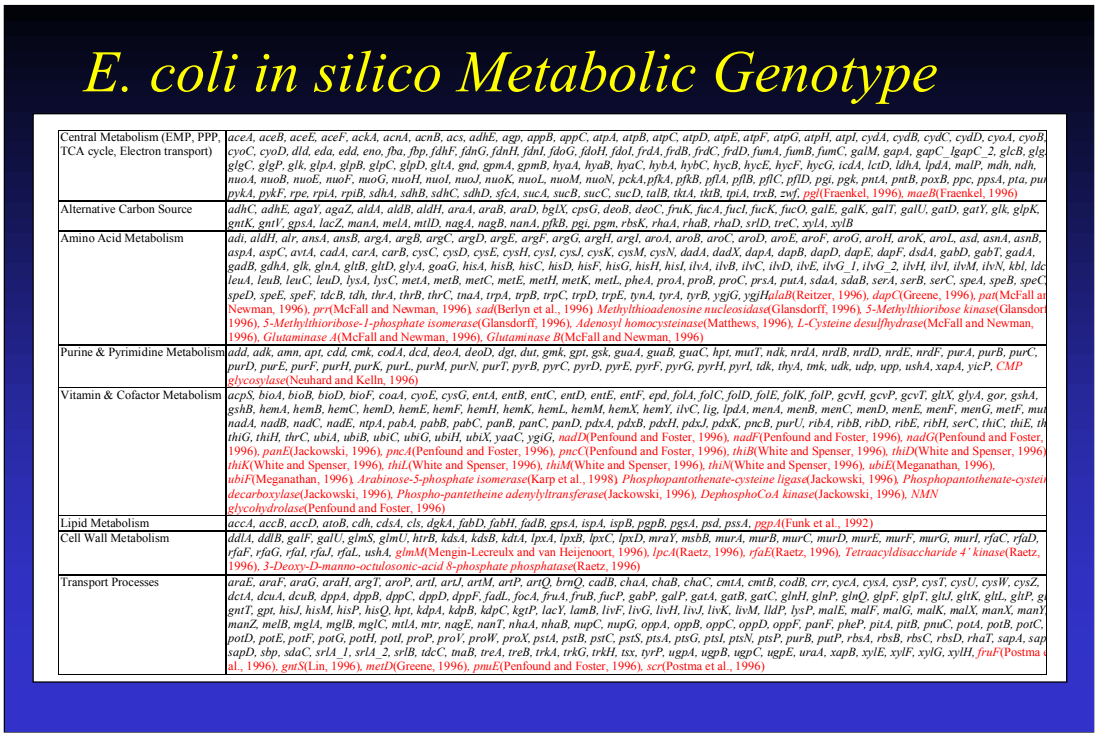
Several in silico genome-scale metabolic maps have been reconstructed. This slide shows actual numbers for three gram-negative bacteria. *E. coli* is a free living organism that can live off of several different individual carbon sources. *E. coli* has thus been called the 'complete' organic chemist as it can synthesize all the chemical structures that it needs for its biomass synthesis. In sharp contrast, *H. influenzae* and *H. pylori* are human pathogens that require several different substrates to grow.



TRANSLATION OF THE STOICHIOMETRIC MATRIX INTO A METABOLIC MAP

Thus we can translate the genomics and biochemistry of a metabolic reaction network directly into the realm of linear algebra in the form of a stoichiometric matrix. Beginning with the gene products of a system we can determine the inter-conversions of metabolites which occur and represent this in the form of a stoichiometric matrix to complete the translation. Within the stoichiometric matrix lies all of the structural information and the architecture of the network. The word structure here is not used to denote the physical structure but the structure of a network.

The stoichiometric matrix is a connectivity matrix that ties all the metabolites, the 'nodes,' in the network together, where the 'edges' or 'connections,' are the metabolic reactions. The stoichiometric matrix is thus a compact mathematical representation of a metabolic map. These maps give us a visual, and easier to understand, representation of the metabolic network in a cell.

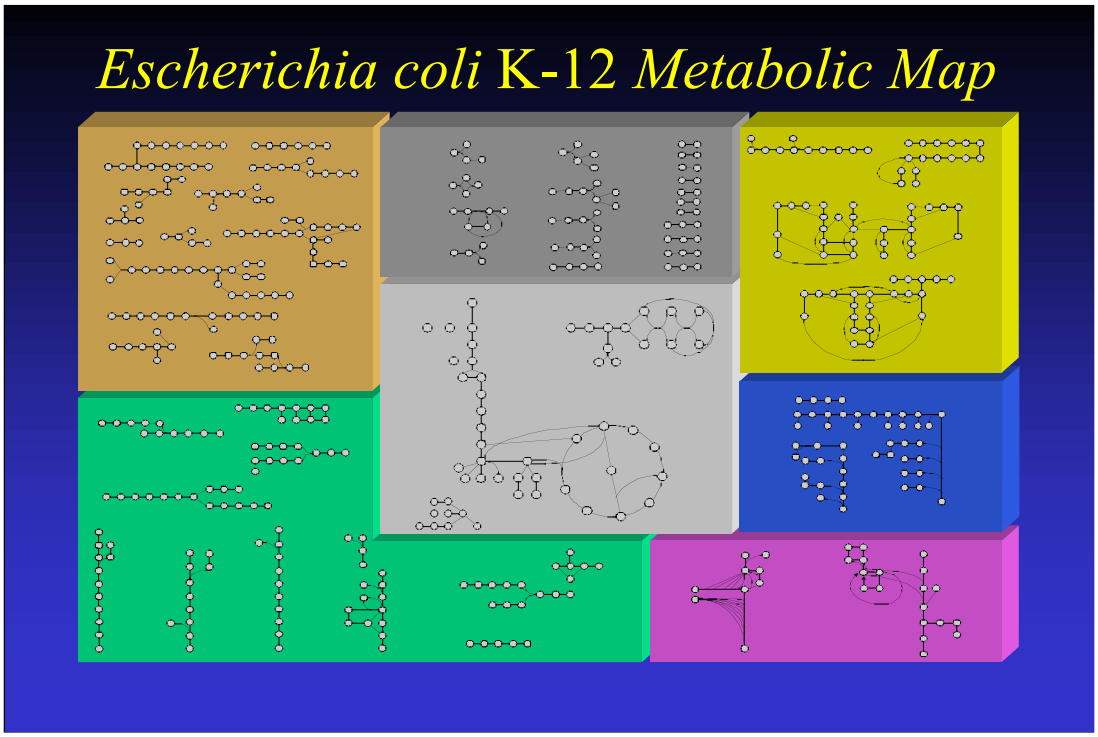


THE E. COLI METABOLIC GENOTYPE

The list of metabolic genes that are found on the E. coli K-12 genome is shown. These genes are categorized into different classes depending on the part of metabolism in which they participate.

The lowest group represents the transporters, or the **b** fluxes.

The enzymes indicated in red represent enzymatic activities that have been found to be in E. coli by biochemical means but the corresponding gene, or ORF, has not been located on the genome.

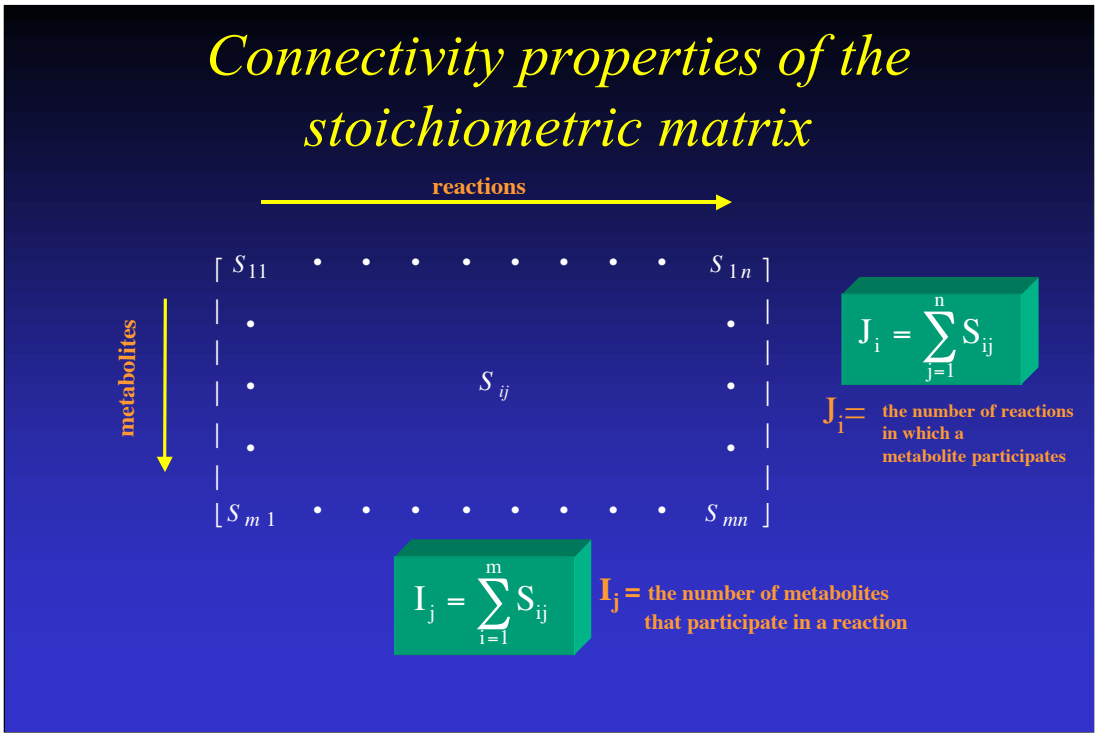


THE METABOLIC MAP REPRESENTATION OF THE *ESCHERICHIA COLI* K-12 METABOLIC GENOTYPE

The metabolic map of the *E. coli* K-12 metabolic genotype divided into metabolic sectors based on a biochemical rationale:

Gray:	Alternative carbon source metabolism
Light gray:	The core metabolic pathways
Orange:	Amino acid biosynthesis
Green:	Vitamin and co-factor metabolism
Yellow:	Nucleotide synthesis
Blue:	Cell wall synthesis
Purple:	Fatty acid synthesis

Not all the 720 reactions are shown. Highly connected metabolites, such as ATP, PEP and pyruvate are likened to dozens of reactions. Showing all of these connections would make this representation visually unattractive. However, these connections should not be overlooked as they play a key role in the stoichiometric characteristics of metabolism.

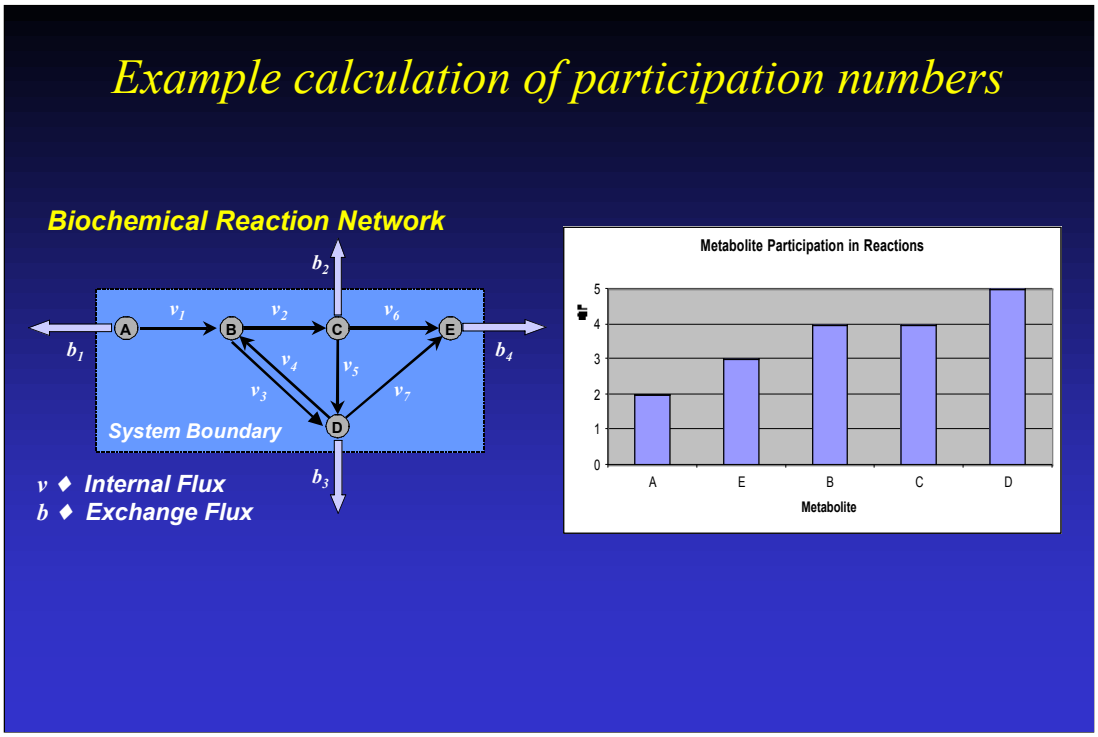


SOME CONNECTIVITY PROPERTIES OF THE STOICHIOMETRIC MATRIX

As illustrated above the stoichiometric matrix is a connectivity matrix that connects all the metabolites in a defined metabolic system. We now introduce some of its connectivity properties:

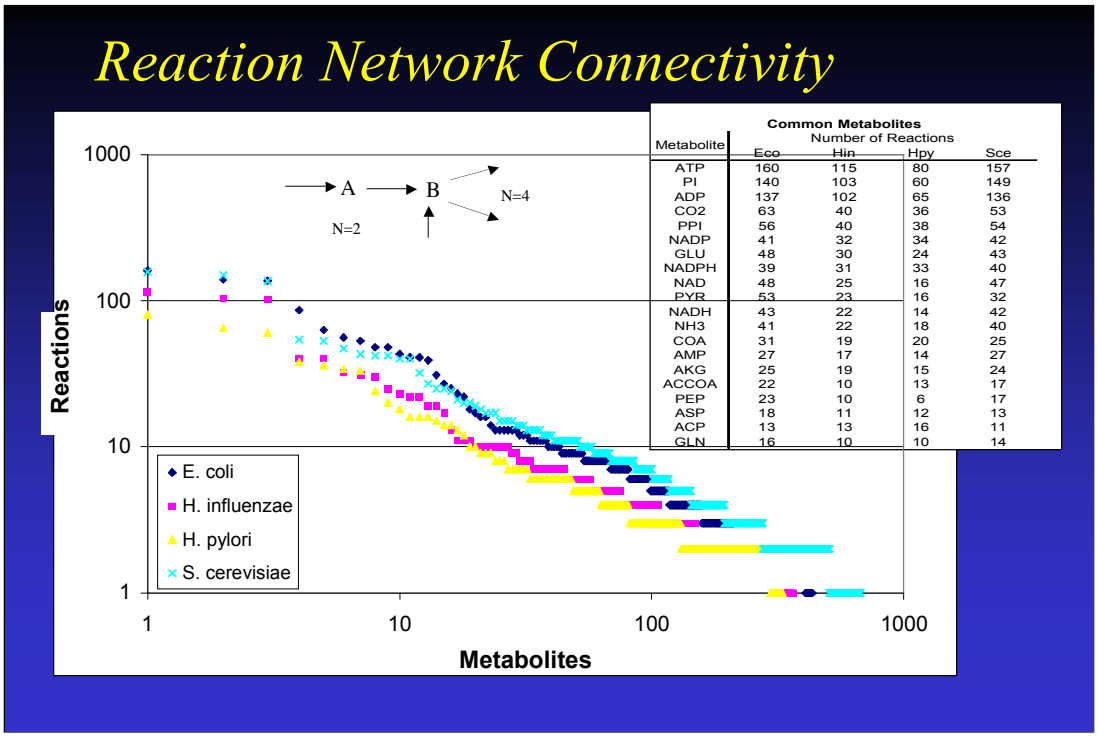
1. The participation number. Metabolites can participate in several metabolic reactions. The number of metabolic reactions that a metabolite participates in can be obtained by simply summing up the number of non- zero elements in the row that corresponds to the metabolite. Note that all internal metabolites must have a participation number of two or more. If not there is a dead end in the network. This feature can be used to curate and diagnose genome annotation, as being either incomplete or erroneous. External metabolites typically will have only a single reaction associated with them, namely membrane transport.

2. The number of molecules participating in a particular metabolic reaction can be obtained by simply summing up the absolute value of all the stoichiometric coefficients that appear in a column. The most frequent number is 4.



CALCULATION OF PARTICIPATION NUMBERS

This slide shows a calculation of the participation number for the simple reaction schema that we have been using. D is the most highly connected metabolite participating in five reactions, while A is the least, participating in the minimum number of two reactions.

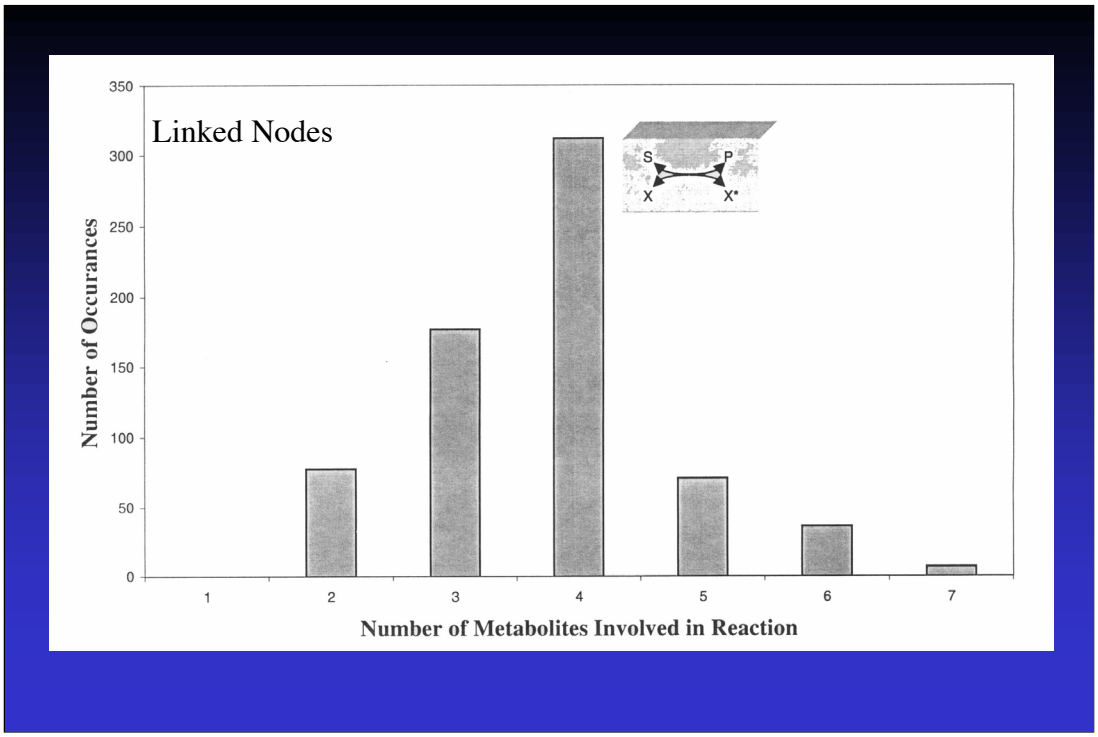


THE PARTICIPATION NUMBERS FOR FOUR METABOLIC MAPS

The 436 metabolites in the *E. coli* K-12 metabolic genotypes all have a participation number associated with them. Here we have calculated them all and rank ordered the metabolites according to the number of reactions that they participate in. This data must be plotted on a log-log scale in order to see the entire range of participation numbers.

ATP is the most highly connected metabolite in *E. coli* K-12. It participates in 161 of the 720 reactions, about one in five reactions. Similarly, ADP and P_i participate in a similar number of reactions. Thus high-energy phosphate metabolism tightly connects the entire metabolic network. Glutamine, the central metabolite of nitrogen metabolism participates in 40+ reactions, and PEP and pyruvate are also highly connected at 25 and 55 respectively. The redox carriers participate in a few dozen reactions. It is therefore not surprising that metabolic regulation must be focused on maintaining the concentrations of these metabolites within a narrow range. Otherwise the entire system would be influenced.

The majority of the metabolites, 198 of the 426, participate in only two reactions, one that forms them and one that degrades them.



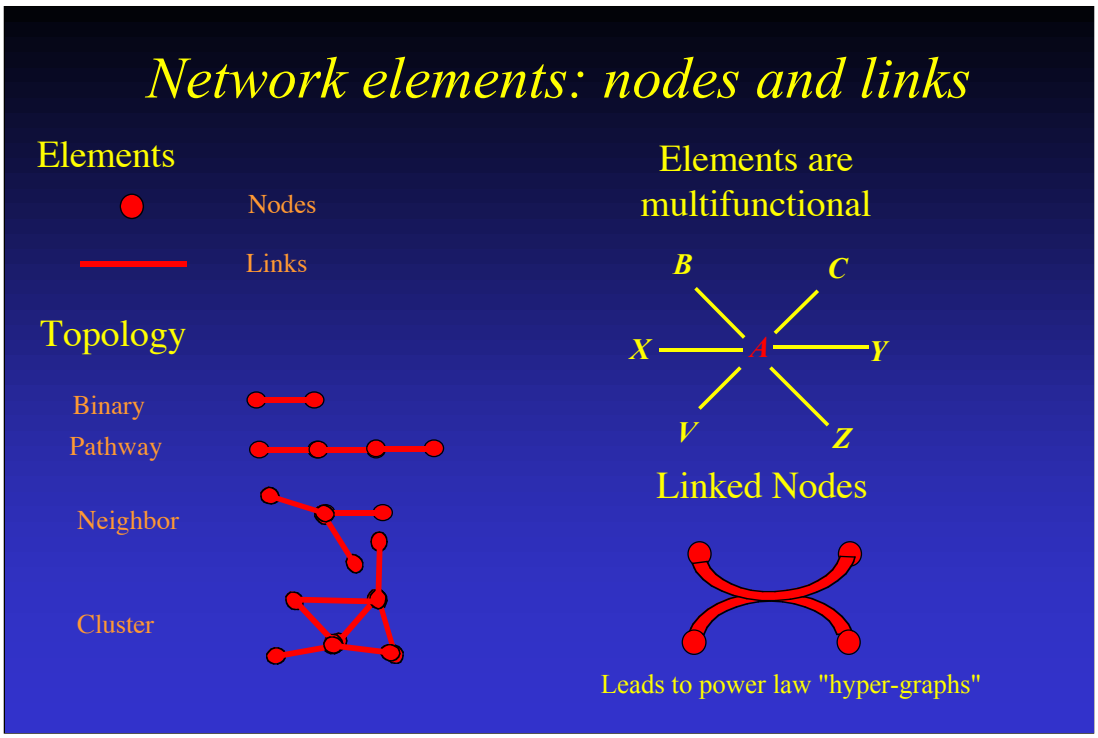
THE NUMBER OF MOLECULES THAT PARTICIPATE IN THE REACTIONS IN THE *ESCHERISCHIA COLI* K-12 METABOLIC GENOTYPE

This histogram shows the number of reactions in E. coli that have 2,3,4, etc molecules participating in the reaction. The most common reaction is of the form:

$$A\text{-}x + B \iff A + B\text{-}x$$

In other words an exchange of a moiety, group, or electrons among molecules. As we saw above, most commonly A-x would be ATP and A would be ADP, and the moiety x is a high energy phosphate group.

We shall see below that this feature has a significant influence on metabolic dynamics. Also this feature makes the map a power-law hyper-graph.



GRAPH THEORY

Much work will be needed to study the structural features of biochemical reaction networks. A few of the issues are illustrated here:

1. There are elements and links in networks. In metabolism, these two correspond to metabolites and the enzymatically catalyzed reactions between them

2. The topological features will be studied.

3. Each element in a network will have many function and potentially many types of links

4. In metabolism, there are linked nodes, i.e. one link will tie together more than two nodes (see previous slide). This changes the nature of the network substantially.

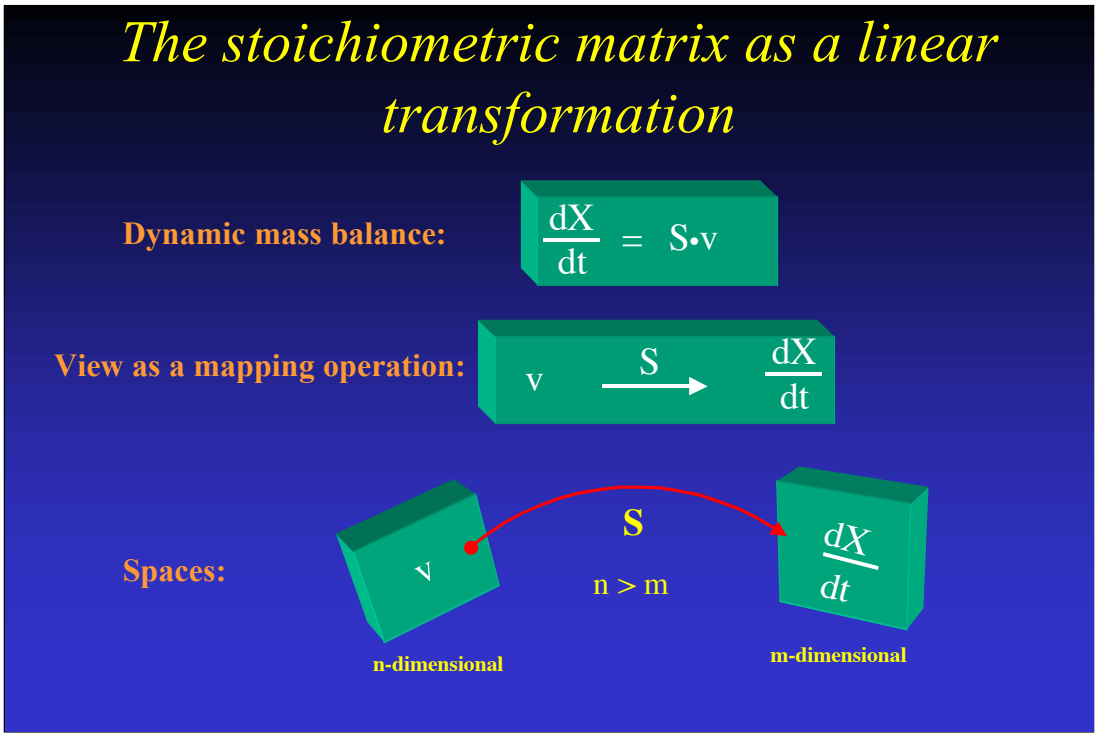
Dynamic Mass Balance: Matrix Form

$$\frac{d}{dt}\begin{bmatrix} X_1 \\ . \\ . \\ . \\ X_m \end{bmatrix} = \begin{bmatrix} S_{11} & . & . & . & . & . & . & S_{1n} \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ S_{m1} & . & . & . & . & . & . & S_{mn} \end{bmatrix} \begin{bmatrix} v_1 \\ . \\ . \\ . \\ v_n \end{bmatrix}$$

THE GENERAL DYNAMIC MASS BALANCE EQUATIONS

This slide shows the details of the general mass balance equations. The time derivatives of the metabolite concentrations (X) is the matrix multiplication of the stoichiometric matrix (S) and the flux vector.

Multiply one row times the vector to see how the summation of fluxes forms the RHS of the differential equation for that metabolite.



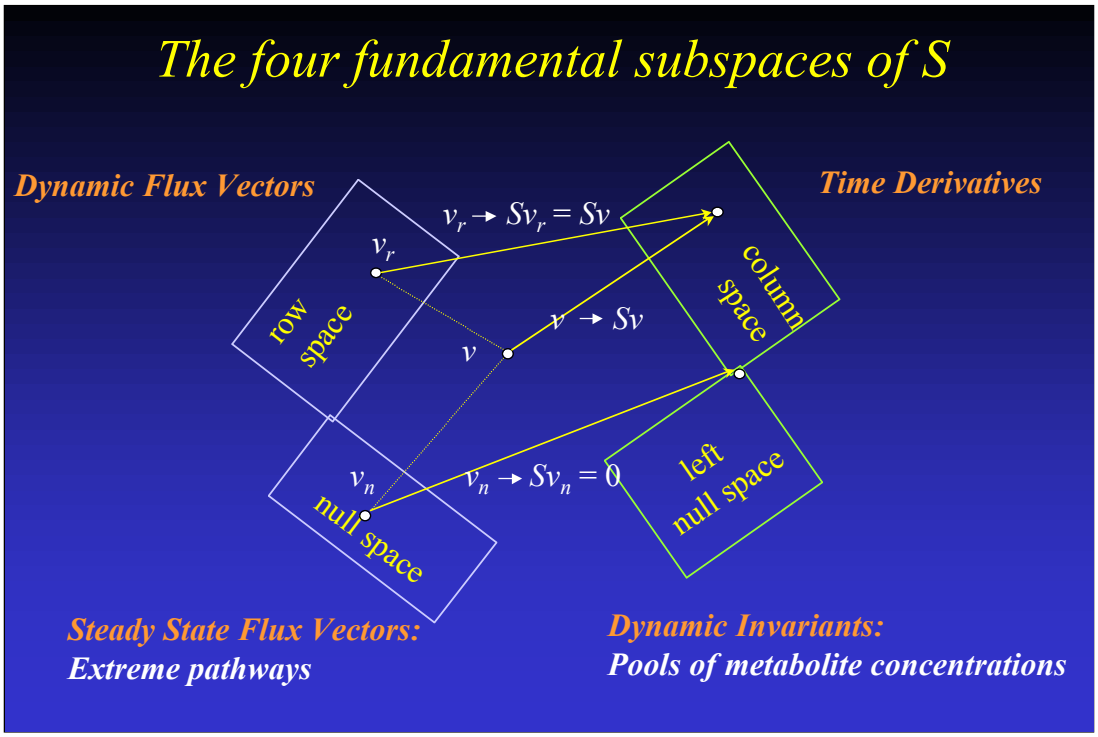
ANY MATRIX MAPS AN ELEMENT FROM ONE VECTOR SPACE INTO ANOTHER; THAT IS IT TRANSFORMS ONE VECTOR INTO ANOTHER

A matrix is a linear transformation;

$$y = A\,x,$$

simply is x mapped into y by the matrix A. The stoichiometric matrix maps the flux vector into the time derivatives. As noted, and as we will discuss in much more detail later, the flux vector is a function of the metabolite concentrations, denoted by the vector X in this slide.

The stoichiometric matrix 'S' thus takes the flux vector at any instant and calculates the time derivative of the concentrations; or how the system will move away from the point that it was located at. The next slide illustrates this point.



A SCHEMATIC DEPICTION OF THE ACTION OF A MATRIX AND THE FOUR SUBSPACES ASSOCIATED WITH IT

Every matrix can be thought of as a mapping operation or a linear transformation. It takes a vector in one space and transforms into a vector in another space. The four fundamental spaces are the row, column, null, and the left null space. These spaces are further described on the next slide.

The four fundamental subspaces of ***S***

- The null space **S**•v=0,
  - contains all the steady state solutions to the flux balance equations
- The column space of **S** (range);
  - contains the time derivatives resulting from the mapping
- The row space of **S**;
  - contains the dynamic flux vectors on which **S** operates
- The left null space of **S**;
  - contains all the dynamic invariants of **S**

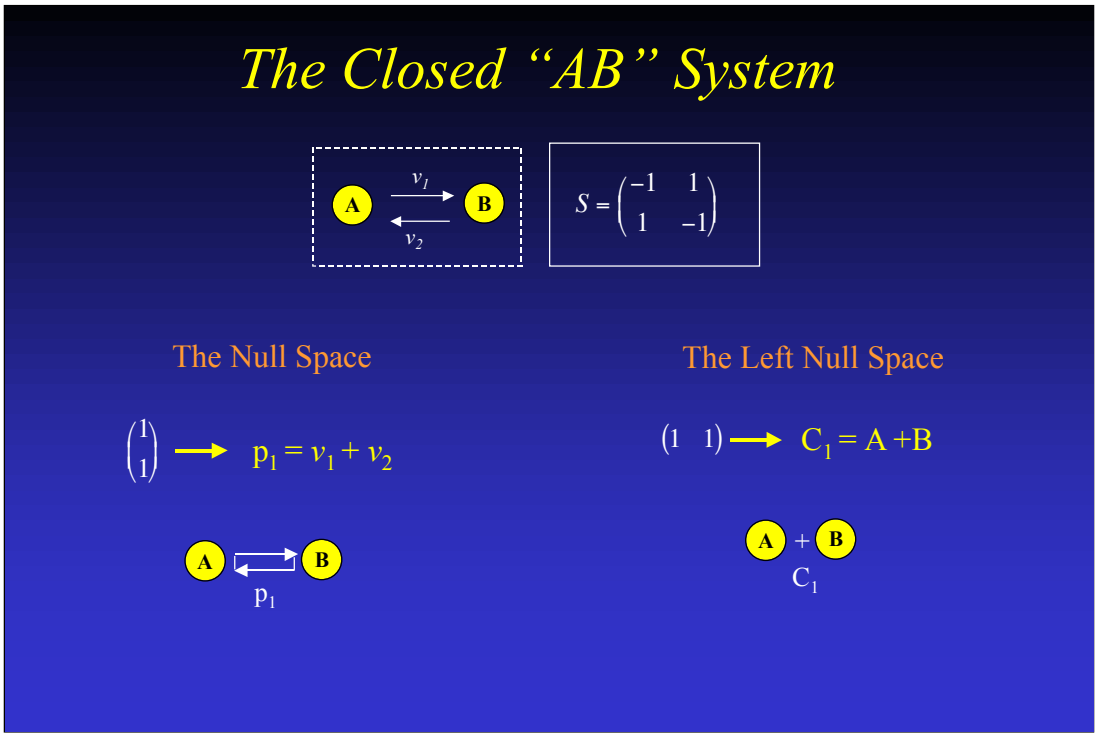
THE FOUR SUBSPACES OF THE STIOCIOMETRIC MATRIX

All the four fundamental subspaces of **S** will be of interest to us. The first spaces that we will study are the right and left null space of **S**, since it contains all the steady state solutions;

$$\mathbf{S}\mathbf{v} = 0$$

And the pooled variables

$$\Sigma_i \, (d\mathbf{X}_i/dt) = 0$$



THE SIMPLE 'AB' EXAMPLE:

Let's consider a reversible reaction. The stoichiometric matrix S is shown and it is rank deficient.

The addition of the two columns gives zero. This can be seen by multiplying the stoichiometric matrix with the column vector $(1,1)^t$. Thus this column vector spans the null space. This vector represents the pathway

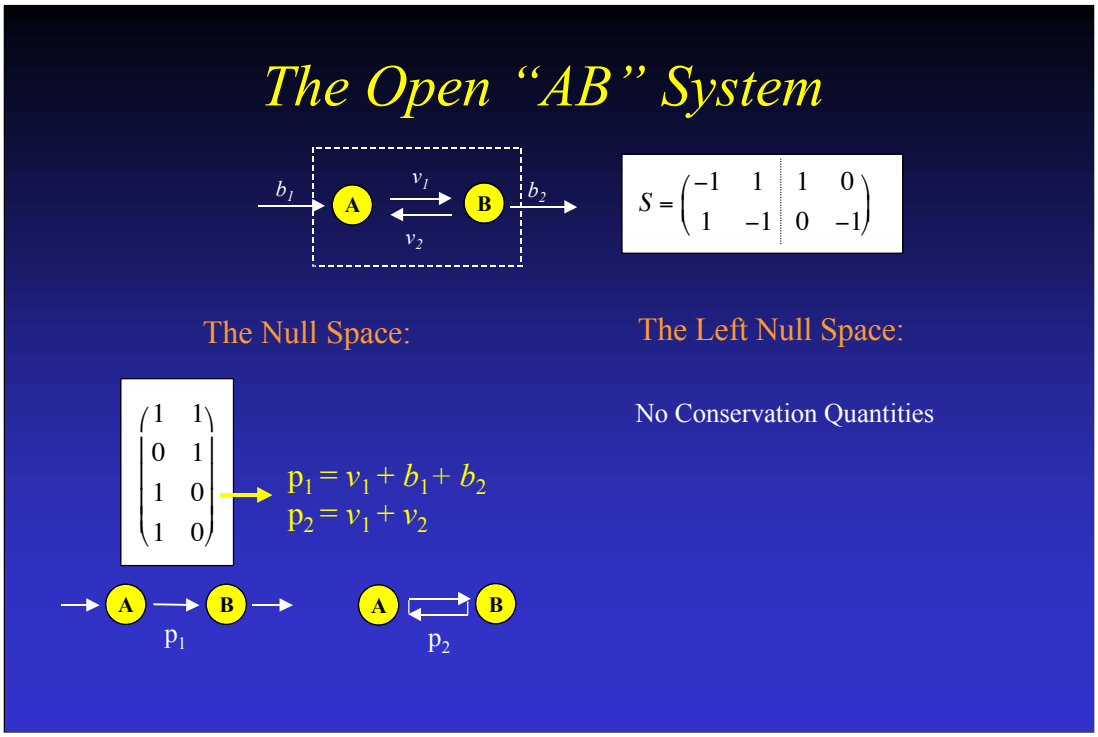
$$v_1+v_2$$

or the reversible back and forth reaction.

The addition of the rows gives a zero. This can be seen by multiplying from the left with the vector (1,1). Thus (1,1) spans the left null space and represents the summation of

$$A+B.$$

It is obvious in this case that this sum is time invariant.



THE OPEN 'AB' EXAMPLE

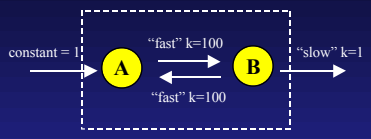
If we now add exchange fluxes the stoichiometric matrix for the closed system is 'appended' with the exchange reactions. The matrix no longer rank deficient. Thus the left null space is of zero dimension and there are no conserved quantities. The sum of A and B will vary with time depending on the exchange fluxes.

The null space is now two dimensional. It is spanned by two pathways. The same pathway as existed for the closed system, corresponding to the reversible reaction, is still there. Later we shall classify this pathway, as Type III.

There is a new pathway vector. It ties the input and the output via a straight pass through the system. Later we shall classify this pathway as Type I.

Any steady state flux distribution in this simple open 'AB' system is a linear combination of these two basis pathways.

Overlaying Order of Magnitude Kinetics



Rate Equations:

$v_1 = 100[A]$

$v_2 = 100[B]$

$b_1 = 1$ (constant)

$b_2 = [B]$

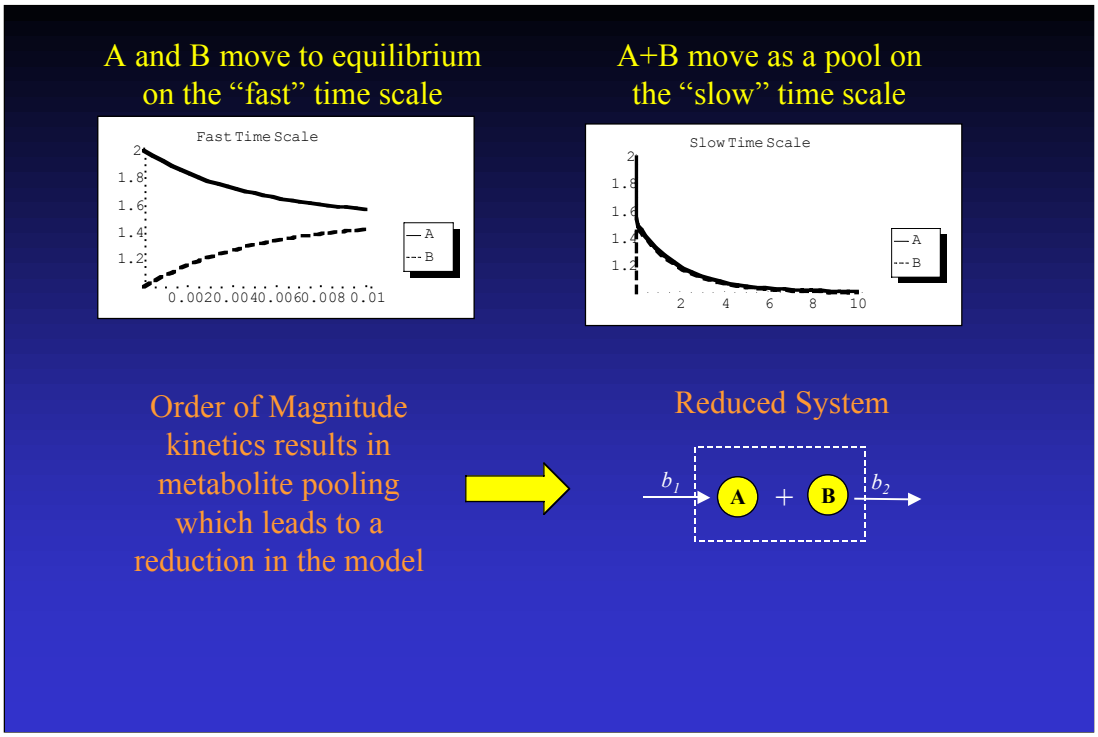
Gradient of the flux vector

$$\frac{\partial v_i}{\partial C_j} = \begin{array}{c} \\ v_1 \\ v_2 \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} 100 & 0 \\ 0 & 100 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

INCLUDING CHEMICAL KINETICS

If one wants to simulate the dynamic states of this system, the kinetics of the reactions need to be known. This means that algebraic expressions for the rate laws must be provided. Here we show a simple mass action type representation of these rate laws assuming that the reaction is first order.

The Jacobian matrix that describes the dynamics is a product of S and the gradient matrix shown at the bottom of the slide.



INTRODUCTION TO TIME SCALE SEPARATION

If the reversible reaction is fast compared to the exchange fluxes, there is very little net exchange with the environment as the reaction equilibrates. Thus the system will behave like a closed system on the fast time scale. Thus a 'pool' of A+B will be formed quickly, and the total inventory in the pool will change slowly as dictated by the exchange fluxes. The state of the system is thus described by only one variable 'A+B'.

This pool formation procedure will be a key element in the massive model reduction challenge that faces us.

Geometric Representation in the Null Space

The steady state solution can be decomposed into weightings on the extreme pathways.

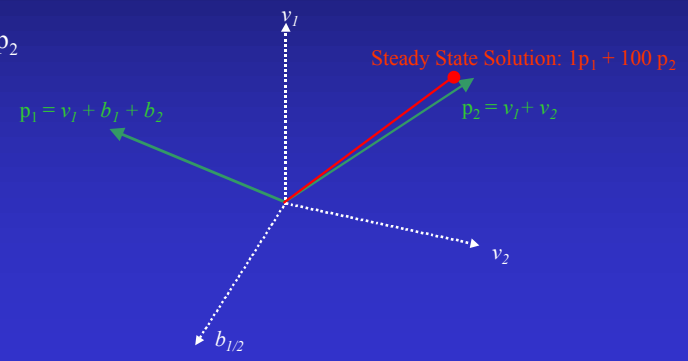
For the sample system above, the steady state solution is:

v_1=101; v_2=100; b_1=1; b_2=1

This can be broken down into weightings on the pathways:

$1p_1 + 100p_2$

The fast kinetics of v_1 and v_2 "push" the steady state solution to the p_2 edge of the cone/plane.

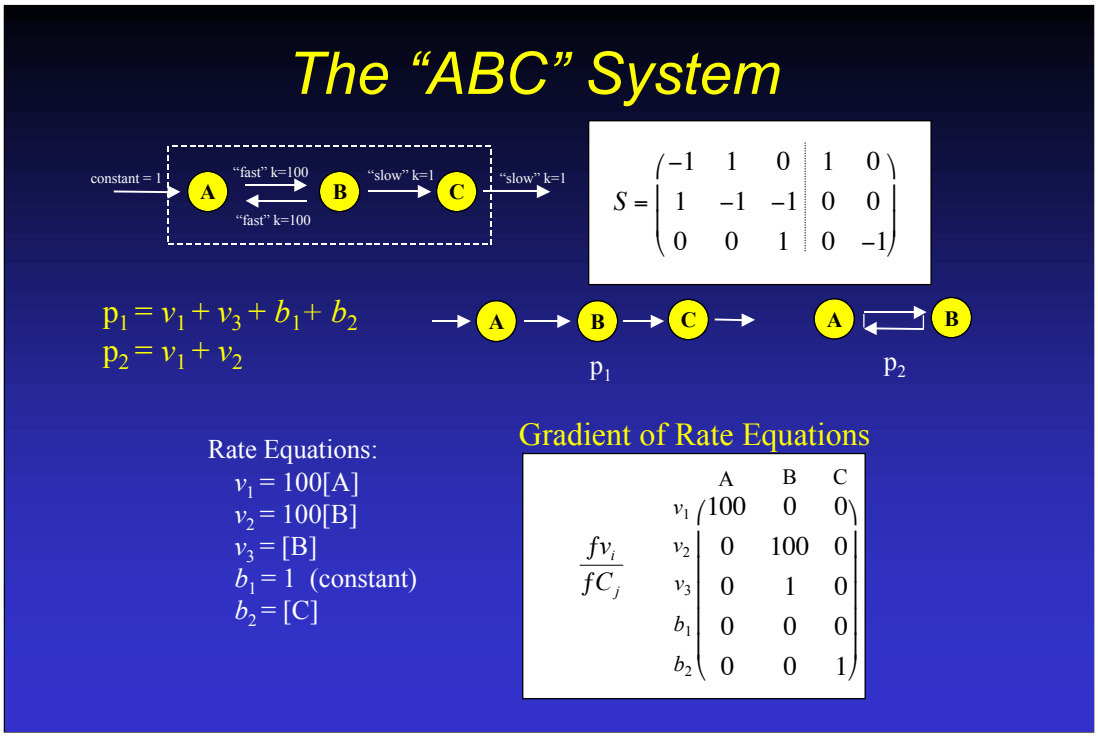


INSIGHTS FROM GEOMETRY

The null space is a cone, as we shall see in much more detail later, and the two pathways are the edges of this cone. If the kinetic parameters are well separated as indicated the steady state flux distribution is

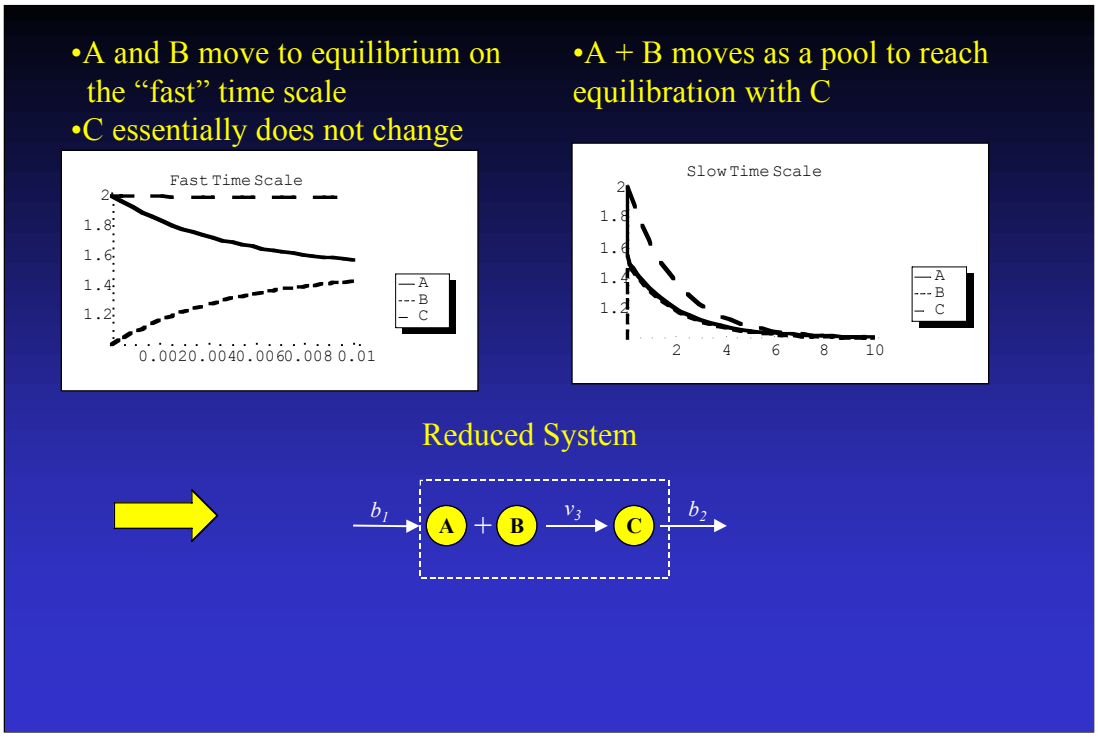
$$v_{ss} = p_1 + 100\ p_2$$

This solution lies 'close to the edge' of the solution space. We shall see this feature emerge as a principle later on



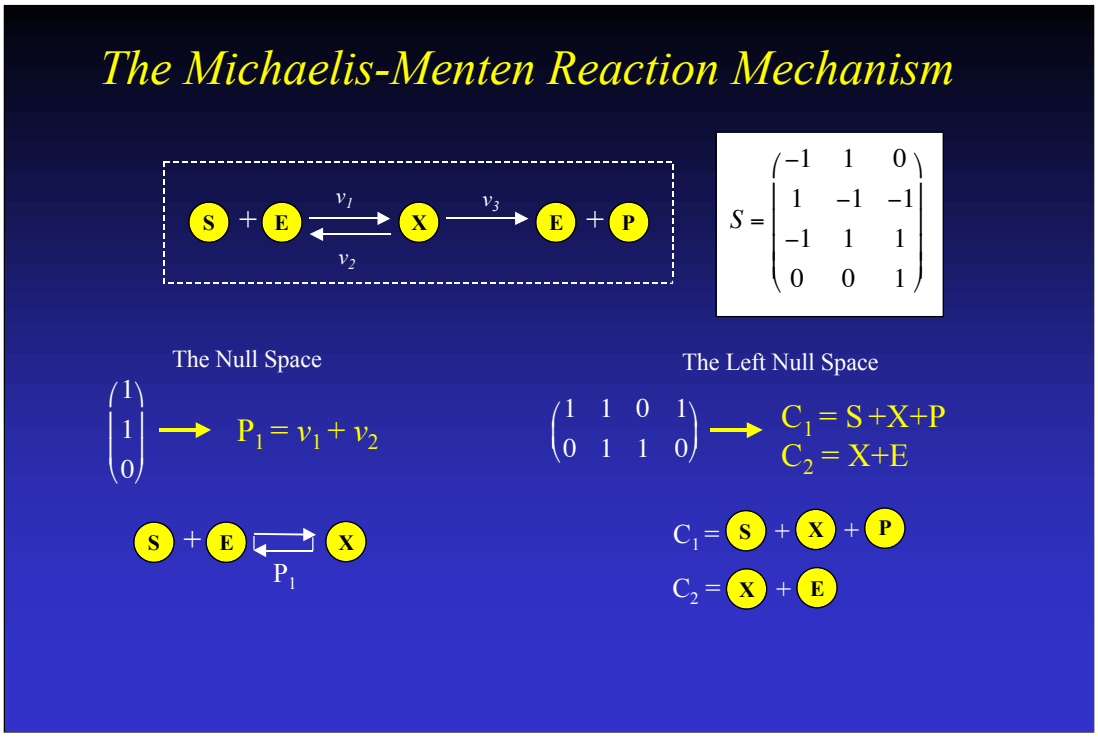
A SLIGHTLY MORE COMPEX EXAMPLE

The next two slides have a slight variation on the previous example. Now we are examining a 3 component system but the analysis is the same. A and B equilibrate on the fast time scale forming a pool (A+B). On the slower time scale the the pool (A+B) is filled via the input reaction and drained via the conversion to C.



These graphs show the pooling effect over the 2 times scales. On the fast time scale it is evident that A and B are equilibrating while C is unchanging (the concentration does not change under the fast "window of observation"). On the slower time scale, you can see that A and B move as a pool to equilibrate with C.

The reduced network that is a result of the pooling is diagramed at the bottom of the slide.



THE MICHAELIS-MENTEN REACTION MECHANISM

The classical MM mechanism can be studied along the lines introduced. For the closed system there is one pathway and two conserved moieties: the total enzyme and the total substrate species. The latter disappears as we open the system on the next slide.

A detailed kinetic analysis of the irreversible MM mechanism is found in:

B. O. Palsson (1987), "On the Dynamics of the Irreversible Michaelis-Menten Reaction Mechanism", *Chem. Eng. Sci*., **42**, 447-458.

Stoichiometric Matrix

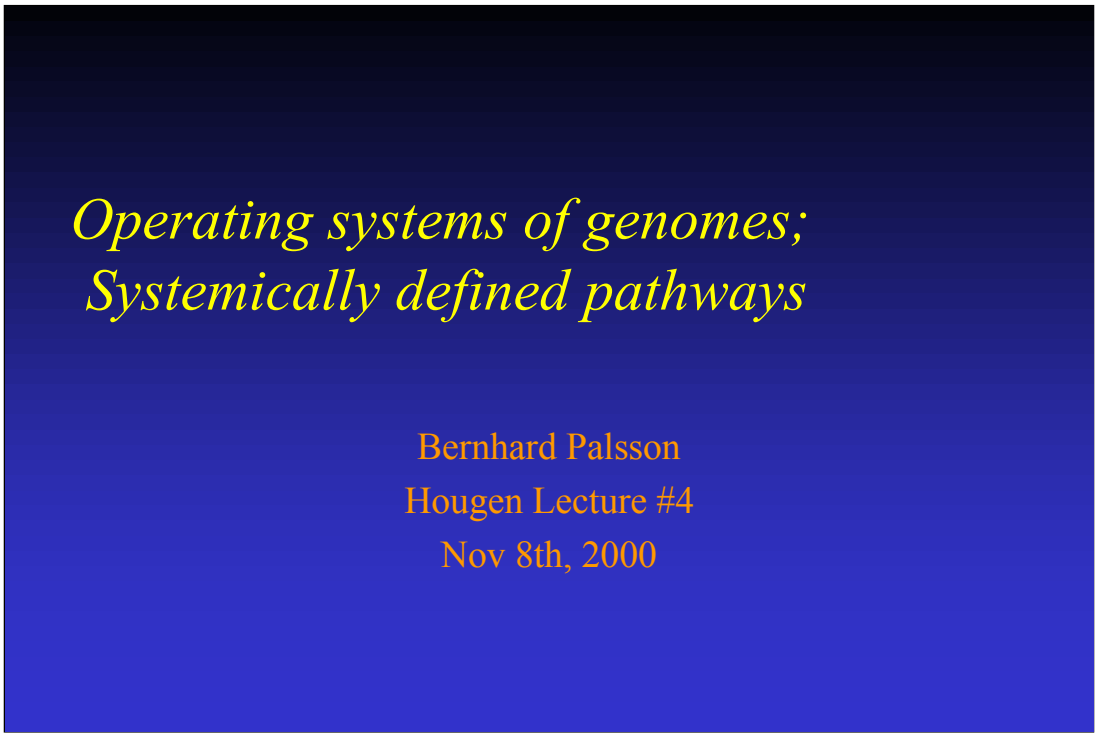
- Can now be derived from annotated genomes given knowledge of enzyme stoichiometries
- A mathematically compact description of metabolic maps
- Has characteristic connectivity and graph properties
- Its size for simple prokaryotic cells is (300-450) X (500-750)
- The stoichiometric matrix is 'sparse', i.e. few non-zero elements
- It has well defined associated fundamental sub-spaces
- These subspaces are keys to understanding pool and pathway formation, and thus model reduction and conceptual simplification

SUMMARY OF POINTS MADE ABOUT THE STOICHIOMETRIC MATRIX

This list is a summary of the points made about the matrix. Next we shall look into linear algebra and examine the matrix properties of S.

References

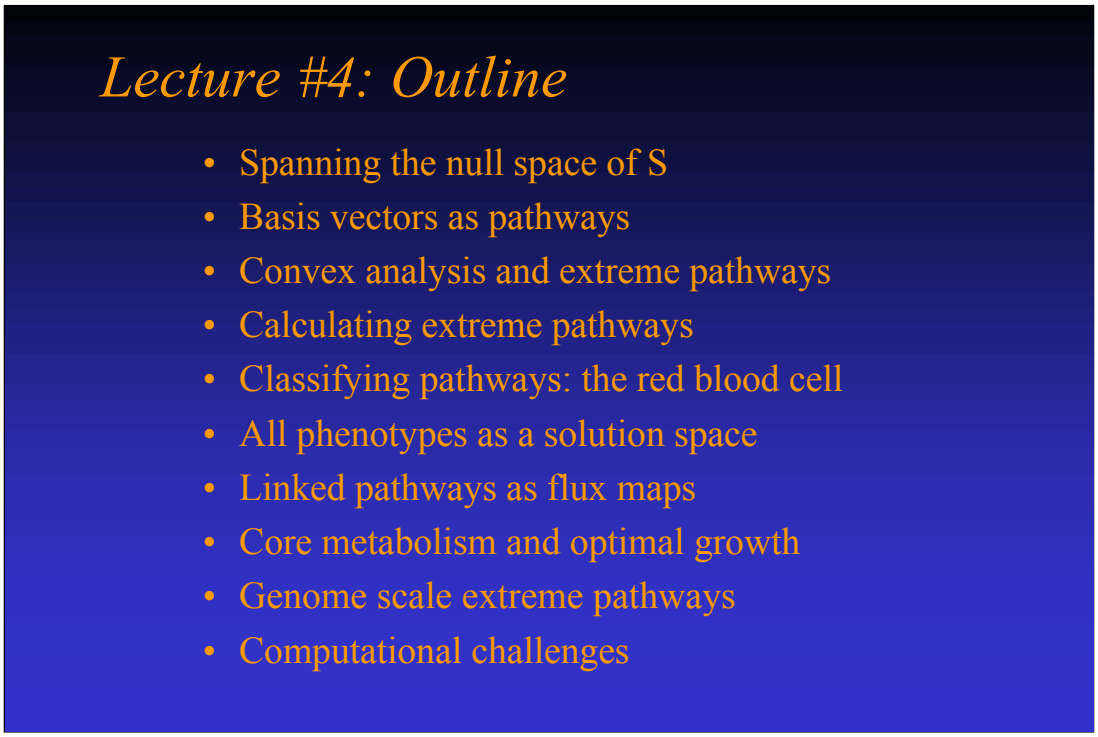
- Gilbert Strang, Linear Algebra and Its Applications, Academic Press, New York, 1981.
- B. O. Palsson, "On the Dynamics of the Irreversible Michaelis-Menten Reaction Mechanism", *Chem. Eng. Sci*., **42**, 447-458 (1987).
- C.H. Schilling, S. Schuster, B.O. Palsson, and R. Heinrich, "Metabolic Pathway Analysis: Basic Concepts and Scientific Applications in the Post-Genomic Era," *Biotechnology Progress*, **15**: 296-303 (1999).
- J.S. Edwards and B.O. Palsson, "The *Escherichia coli* MG1655 *in silico* metabolic genotype; Its definition, characteristics, and capabilities," *Proc. Natl Acad Sci (USA)*, **97**: 5528-5523 (2000).
- B.O. Palsson, "The challenges of *in silico* biology," Nature Biotechnology, **18:** 1147-1150 (2000).



INTRODUCTION

In the previous three lectures we surveyed the world of genomics, how this information is giving us the biochemical reaction networks that operate in cells, and how we can approach the mathematical modeling of these networks and their simulation in a computer.

We now begin the mathematical modeling process in earnest and analyze the consequences of connectivity constraints and thermodynamics, i.e. the irreversibility of some reactions



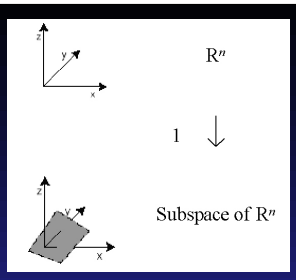
LECTURE #4

This lecture will cover the definition and use of systemic pathways. We begin with the vectors that span the null space of S. We show that these are biochemical pathways. However the basis for linear spaces are not unique but by imposing irreversibility constraints of thermodynamics we leave the domain of linear analysis and enter that of convex analysis. Now the solution space is conical in shape and the edges of the cone become the spanning vectors. These vectors are unique and are the 'extreme pathways.'

We then cover the algorithm that is used to calculate these extreme pathways, and compute them for red cell metabolism and investigate their biochemical significance. We then introduce linked outputs that respond to physiological functions such as growth and show that the extreme pathways are now metabolic maps.

We end the lecture by discussing some of the computational challenges associated with calculating these maps on a genome wide scale.

The Null Space of the Stoichiometric Matrix



- Contains all the solutions to **Sv=0**
- These are the steady state solutions to the dynamic mass balances
- The time constants of metabolic transients are typically very fast, i.e. shorter than about 1 to 5 minutes, especially in bacteria
- Thus for most practical purposes metabolism is in a steady state
- The null space contains all the steady state flux distributions and is thus of special importance to us
- The dimension of the null space is the number of columns in the matrix minus the number of independent rows (the rank of the matrix)

METABOLIC TRANSIENTS AND THE NULL SPACE

The concentrations of metabolites tend to be very low, in the order of miro-molar, or about 60,000 molecules per *E. coli* cell. Yet the metabolic fluxes are about 100,000 molecules per sec per cell. Thus, the average response time of a metabolic concentration is about 1 second. These transients are too fast for essentially all practical purposes.

Metabolites that are in higher concentrations, such as ATP, can have time constants that are on the order of minutes. Nevertheless, compared to the progression of an infection or bioprocessing this time is very short and metabolism is very fast and can be effectively considered to be in a steady state.

In some highly specialized mammalian cells, metabolic transients can be slower. For instance in the human red blood cell, the ATP turnover time is about a hour, and transient changes in 2,3DPG are on a 12 to 24 hr time scale. 2,3DPG binds to hemoglobin to regulate its affinity for oxygen. This time constant is responsible for the time that it takes us to adjust to higher altitudes.

Finding the basis for the null space

Any matrix A:

$$A = \begin{bmatrix} -3 & 6 & -1 & 1 & -7 \\ 1 & -2 & 2 & 3 & -1 \\ 2 & -4 & 5 & 8 & -4 \end{bmatrix}$$

Can be row reduced using Gaussian elimination:

$$\begin{bmatrix} 1 & -2 & 0 & -1 & 3 & 0 \\ 0 & 0 & 1 & 2 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \begin{aligned} x_1 - 2x_2 - x_4 + 3x_5 &= 0 \\ x_3 + 2x_4 - 2x_5 &= 0 \\ 0 &= 0 \end{aligned}$$

FINDING A BASIS FOR THE NULL SPACE

A basis for the null space can be found by a so-called parameterization procedure. First any matrix A is row reduced by Gaussian elimination into the echelon form of the matrix (typically denoted by U). The pivot columns are identified (columns one and three in the example given). These are the columns with the pseudo-diagonal elements. These columns represent the fixed variables. The free variables are in the columns between the pivot columns (the second, fourth and fifth in the example given)

The matrix A has two pivot columns (1 and 3) and three free variables (2,4,5). All the variables can be expressed in terms of the free variables (parametric form)

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 2x_2 + x_4 - 3x_5 \\ x_2 \\ -2x_4 + 2x_5 \\ x_4 \\ x_5 \end{bmatrix} = x_2 \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 1 \\ 0 \\ -2 \\ 1 \\ 0 \end{bmatrix} + x_5 \begin{bmatrix} -3 \\ 0 \\ 2 \\ 0 \\ 1 \end{bmatrix}$$

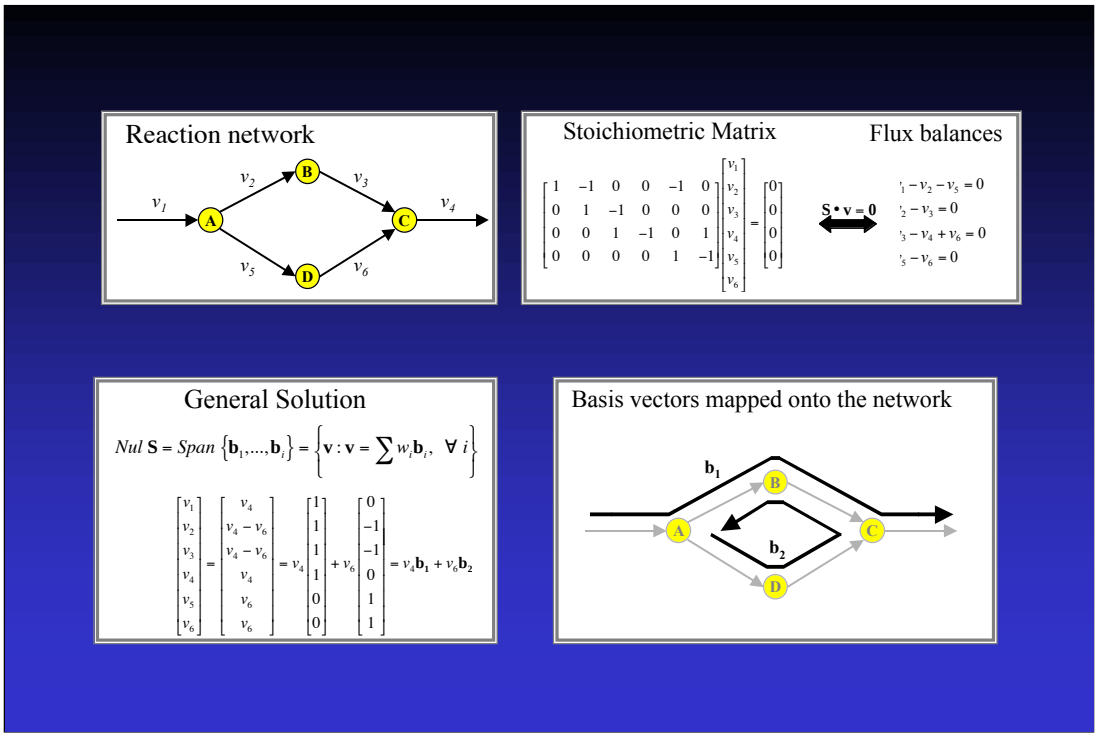
$$= x_2\mathbf{u} + x_4\mathbf{v} + x_5\mathbf{w}$$

Then the vectors u, v, w span the 3-dimensional null space

FINDING A BASIS FOR THE NULL SPACE--CONTINUED

All the variables are then written in terms of the free variables. The equations are then written in vector form by factoring our the free variables individually. The columns that form constitute a spanning set--a basis--for the null space of **A**. The free variables can take on any numerical values to form additions of these basis vectors.

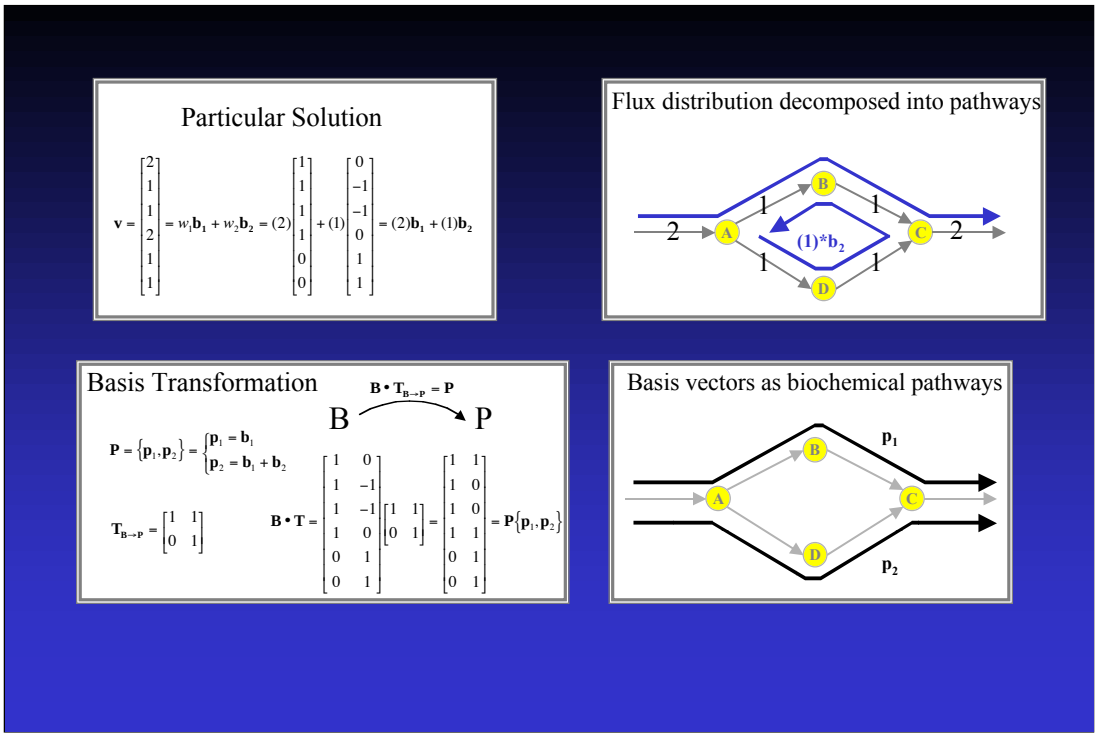
Verify that **Au=Av=Aw=0**, and that **u,v,w** is an independent set of vectors.



FINDING A BASIS FOR AN EXAMPLE STOICHIOMETRIC MATRIX

A simple reaction network is presented in the ULH panel. The corresponding stoichiometric matrix and its flux balances are written in the URH panel. The parameterization method, of the previous two slides, is applied to find a basis for this stoichiometric matrix, as shown in the LLH panel. These basis vectors can be graphically represented on the metabolic map (LRH Panel).

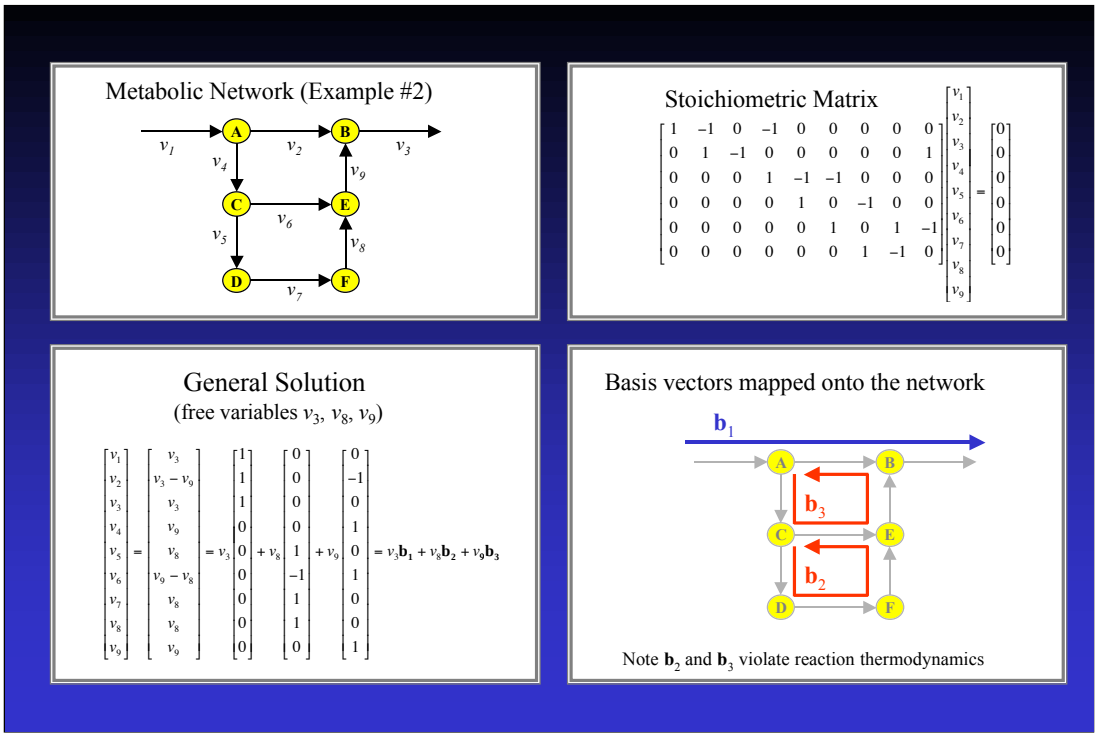
Note that the two basis vectors form a string of connected reactions--effectively pathways. The first basis vector, b_1, is a straight through pathway through the upper part of this small network. The second basis vector, b_2, is a circular path. It has steps in it that run opposite to the direction of two irreversible reactions. Although the basis vectors form mathematically acceptable pathway, biochemically they are not acceptable. However, as we saw previously, we can form another equivalent basis.



FINDING A BASIS FOR AN EXAMPLE STOICHIOMETRIC MATRIX--CONTINUED

•Every flux distribution, **v**, can be *uniquely* described by a combination of the particular set of basis vectors chosen to describe the null space. (Unique Representation Theorem). An example is given in the ULH panel and shown on the metabolic map in the URH panel

•A basis for a vector space imposes a coordinate system on the space. However, this coordinate system is not unique, which implies that other sets of vectors can be used as a basis for the same vector space. One basis can be transformed into another using a basis transformation, as shown in the LLH panel. We seek to find basis vectors whose elements are all positive. Such vectors will form biochemically acceptable pathways as shown in the LRH panel.



FINDING A BASIS FOR A STOICHIOMETRIC MATRIX

EXAMPLE #2

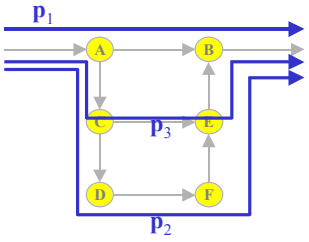
Basis Transformation (Example #2)

$$B \xrightarrow{B \bullet T_{B\to P} = P} P$$

$$P = \{p_1, p_2, p_3\} = \begin{cases} p_1 = b_1 \\ p_2 = b_1 + b_2 + b_3 \\ p_3 = b_1 + b_3 \end{cases} \qquad T_{B\to P} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

$$B \bullet T = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} = P\{p_1, p_2, p_3\}$$

Basis vectors as biochemical pathways



All of the pathways now obey the thermodynamic constraints if they are positively weighted.

The selection of basis vectors is not unique. Therefore it is irrelevant that any flux distribution can be uniquely represented by a set of basis vectors. *We need to try and find a unique "basis" or set of pathways to describe the solution space.*

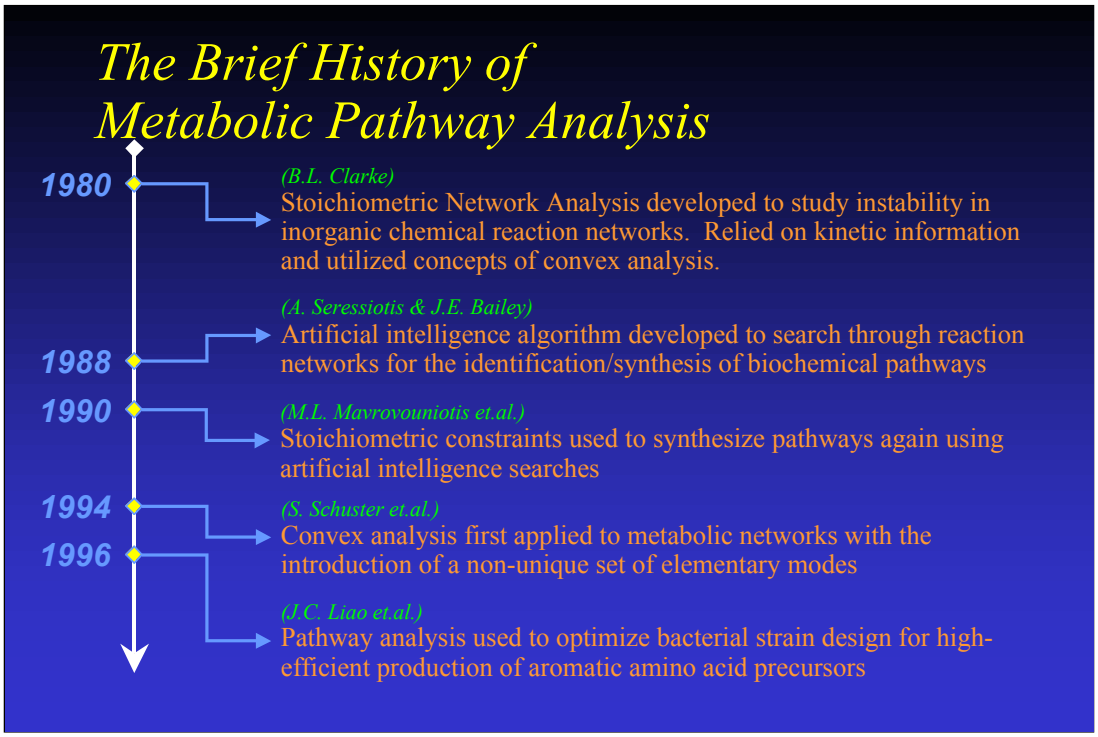
FINDING A BASIS FOR A STOICHIOMETRIC MATRIX
EXAMPLE #2--CONTINUED

The Null space of S

- The matrix has dimension of n columns, and m rows, representing the number of reactions and metabolites, respectively.
- Has dimensions of n-r (r is rank of S, r=m if matrix is full rank).
- Can be found by the parametric approach
- A linear basis is not unique
- An equivalent basis can be found by replacing a member of the spanning set with a linear combination of the other members of the set
- Basis for the null space may be found that contain only positive weights on the elements of v
- Such bases have members of the spanning set which are biochemically meaningful pathways

SOME FACTS ABOUT THE NULL SPACE OF S

We have now established that we can find vectors that span the null space of **S** that represent biochemically acceptable pathways. Note that these pathways are properties of the matrix itself, as they are the basis for one of its fundamental subspaces.



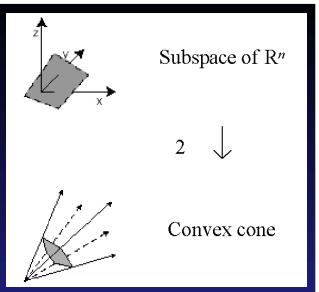
A BRIEF HISTORY OF THE FIELD OF PATHWAY ANALYSIS.

The first work on pathways can be traced back to 1980 with the development of SNA by Bruce Clarke. The theory was developed to study instability in inorganic chemical networks. This was the first attempt to apply convex analysis to reaction networks but was never extended to living systems. This was followed by some work using AI to search through reaction networks following along the lines of graph theory. This was taken another step by Mavro with the introduction of stoichiometric constraints. Both of these approaches lacked a sound theoretical basis.

In 1994 Schuster became the first to apply convex analysis to metabolic networks with the introduction of a non-unique set of elementary modes. This theory was applied a few years later by Liao to optimize bacterial strain design for the high-efficient production of aromatic amino acids.

So at this point in time pathway analysis is just beginning to be applied but still lacks a unified theoretical foundation, which is where the present work comes in.

Convex Analysis



- The study of systems of linear equations and inequalities

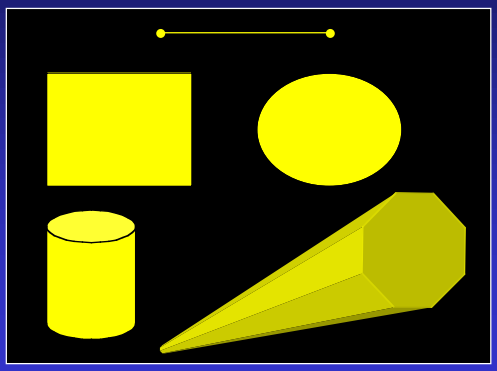
- Convex analysis is used to study metabolic networks where the linear equations are derived from the mass balances and the inequalities are generated from thermodynamic information on the reversibility of reactions.

- From linear algebra a null space is defined which contains all of the solutions to the set of linear homogenous equations. When we add inequality constraints (such as all variables must be positive) the solution space becomes restricted by these inequalities (the portion of the null space in the positive orthant)

What is Convexity?

Definition of a Convex Space: For every two points in the space, the line connecting these two points lies entirely within the space.

Convex Shapes



Non-Convex Shapes

Polyhedral Cones and Pathways

• Region determined by a linear homogeneous equation/inequality system is a convex polyhedral cone (***C***)

$$\mathbf{0} = \mathbf{S} \bullet \mathbf{v}, \qquad v_i \geq 0, \quad i = 1,\ldots,n$$

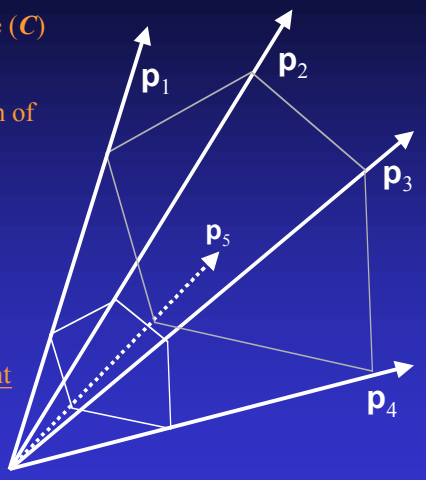
• Every point in the cone is a non-negative combination of the generating vectors (Extreme Pathways) of the cone

$$C = \left\{ \mathbf{v} \in \mathrm{R}^n \;\middle|\; \mathbf{v} = \sum_{i=1}^{k} \alpha_i \mathbf{p}_i, \quad \alpha_i \geq 0 \right\}$$

• The number of generating vectors can exceed the dimensions of the cone (i.e. linearly dependent)

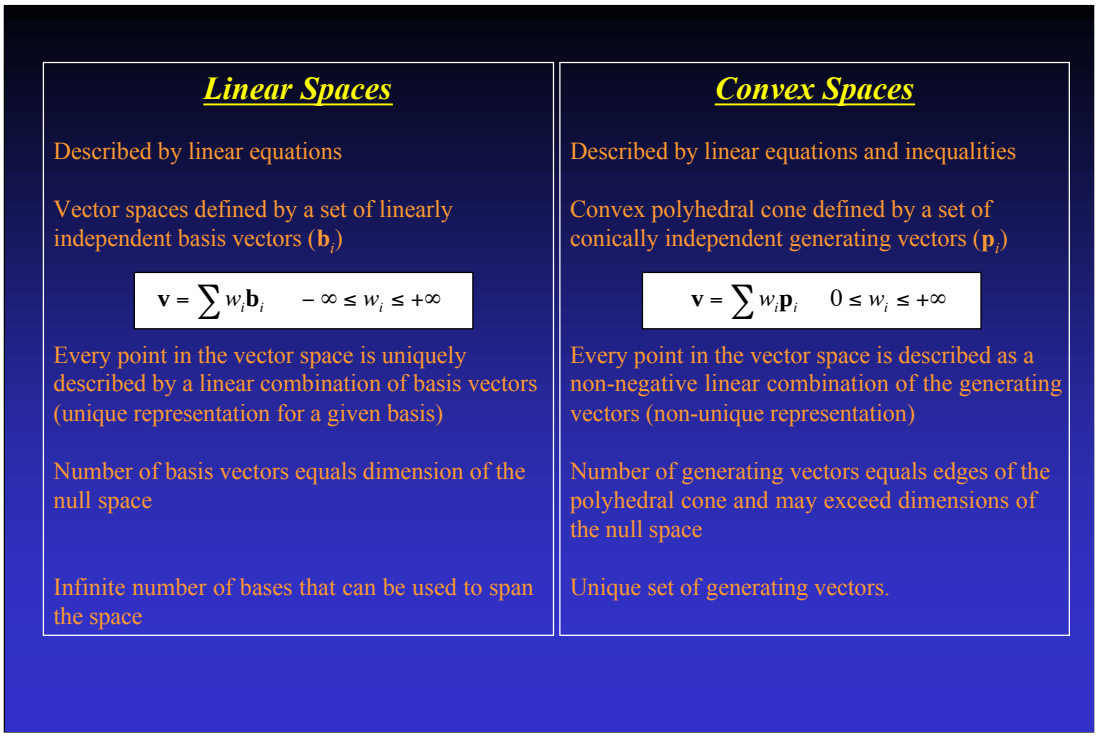
• Generating vectors represent systemically independent pathways which can theoretically be "switched" on or off

• Generating vectors are unique for a system



THE FLUX CONES

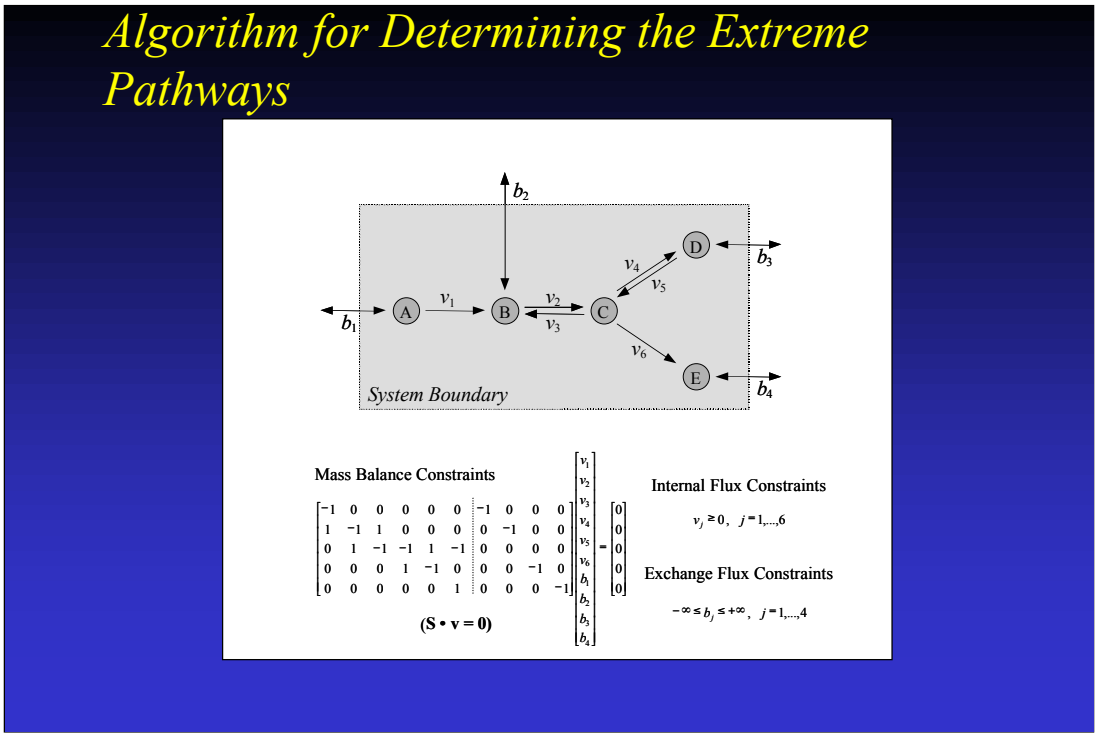
Through the principles of convex analysis it turns out that the shape of the null space for a set of linear equation with positive flux values such as the systems which we are concerned with is a convex polyhedral cone such as the one depicted here on the right. The perspective of the cone is supposed to look as if it is going into the plane of the slide. What is nice about cones is the condition that every point within the cone can be described as a non-negative combination of the generating vectors where the generating vectors are the edges of the cone. If we can determine these generating vectors which are biochemically feasible then we can describe every point within the cone. Additionally the number of generating vectors can exceed the dimensions of the cone which has the mathematical consequence that all of these pathways are not linearly independent. The best analogy for some of these concepts is to think of an Egyptian pyramid which has 4 edges and exists in three dimensional space. Algorithms exist for the determination of these generating vectors and the set of generating vectors represents what may be referred to as genetically independent pathways. This means that each pathway utilizes a unique set of reactions and gene products utilizing a different genotype. Also extremely important is the fact that the set of generating vectors is unique. So to best describe the null space and navigate through the metabolic map of an organism we have to determine this unique set of genetically independent pathways.



COMPARING LINEAR SPACES AND CONVEX ANALYSIS

The number of generating vectors can exceed the dimensions of the cone which has the mathematical consequence that all of these pathways are not linearly independent. The best analogy for some of these concepts is to think of an Egyptian pyramid which has 4 edges and exists in three dimensional space. While not linearly independent these pathways are systemically independent in that they cannot be decomposed into a combination of other pathways in a convex manner. An important fact is that the set of generating vectors is unique and below we represent algorithms to solve for these generating vectors. Thus, to describe the flux space and navigate through the metabolic map of an organism we have to determine that it is best to use this unique set of systemically independent pathways.

This set of pathways can be thought of as the "operating system" for a defined metabolic genotype, since the control over these pathways will enable the attainment of any state (phenotype) allowable by the constraints placed on the metabolic system.

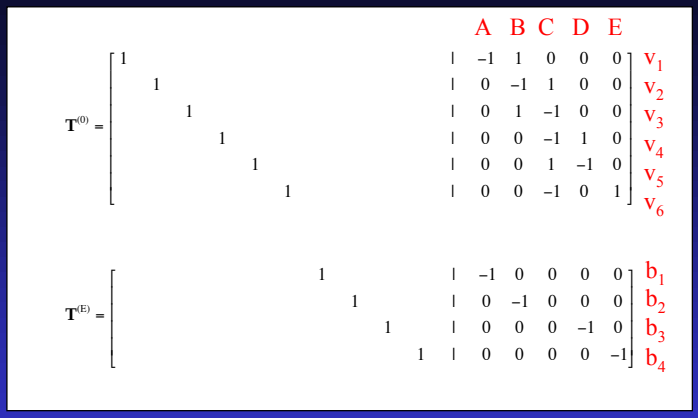


FINDING EXTREME PATHWAYS

Consider the example metabolic system shown above. The stoichiometric matrix is given and so are the constraints placed on the system. How do we now determine the extreme pathways?

The algorithm that is implemented to determine the set of extreme pathways for a reaction network follows the principles of algorithms for finding the external rays/generating vectors of convex polyhedral cones. This algorithm will give unique, biochemically feasible pathways which define the edges of the flux cone.

Initialization of the First Tableau



- Append identity matrix to S^T.
- Separate out unconstrained exchange fluxes into temporary $T^{(E)}$ matrix.

THE ALGORITHM

The algorithm begins with the formulation of an initial matrix consisting of an *n* x *n* identity matrix (**I**) appended to the transpose of the stoichiometric matrix, $\mathbf{S}^T$. Then we examine the constraints on each of the exchange fluxes. If the exchange flux is constrained to be positive nothing is done, however, if the exchange flux is constrained to be negative then we multiply the corresponding row of the initial matrix by –1. If the exchange flux is unconstrained then we move the entire row to a temporary matrix, $\mathbf{T}^{(E)}$. This completes the initialization of the first tableau, $\mathbf{T}^{(0)}$.

The above tableau has been constructed for the example network. In this network, all of the exchange fluxes are unconstrained hence they are all in the temporary matrix $\mathbf{T}^{(E)}$.

Formulating the Next Tableau: $T^{(1)}$

$$\mathbf{T}^{(1)} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & & & & & \end{array} \\ \left[\begin{array}{ccccccc} 1 & & & & & & \\ & 1 & 1 & & & & \\ & 1 & & 1 & & & \\ & 1 & & & & & 1 \\ & & 1 & & 1 & & \\ & & & 1 & 1 & & \\ & & & & 1 & 1 & \end{array}\right. \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \left|\begin{array}{ccccc} -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{array}\right. \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \\ \left.\begin{array}{c} v_1 \\ v_2+v_3 \\ v_2+v_4 \\ v_2+v_6 \\ v_5+v_3 \\ v_5+v_4 \\ v_5+v_6 \end{array}\right] \end{array}$$

- Copy all rows from $T^{(0)}$ into $T^{(1)}$ that have a zero in the column that corresponds to the first metabolite without an unconstrained flux (in this case C)
- Of the remaining rows, add together all possible combinations of rows which contain values of the opposite sign in the C column such that the addition produces a zero in the column and add to $T^{(1)}$.

Each element of the matrix is designated by T_{ij}. Starting with x equal to 1 and $\mathbf{T}^{(0)}$ equaling $\mathbf{T}^{(x-1)}$ the next tableau is generated in the following manner:

1. Identify all of the metabolites that do not have an unconstrained exchange flux associated with them. The total number of such metabolites is denoted by μ. In this example only metabolite C does not have such an unconstrained exchange flux so $\mu=1$.

2. Begin forming the new matrix $\mathbf{T}^{(x)}$ by copying all rows from $\mathbf{T}^{(x-1)}$ which contain a zero in the column of $\mathbf{S}^T$ that corresponds to the first metabolite identified in step 1, denoted by the index c. (This will be the third column of the transposed stoichiometric matrix, with only the first row containing zero in this column).

3. Of the remaining rows in $\mathbf{T}^{(x-1)}$ add together all possible combinations of rows which contain values of the opposite sign in column c, such that the addition produces a zero in this column. Given two rows, $\mathbf{r}_1$ and $\mathbf{r}_2$, whose elements will be denoted by $r_{1,j}$ and $r_{2,j}$ for $j=1,\ldots,(n+m)$, combine the rows using the following equation to generate a new row $\mathbf{r}'$ to be added to $\mathbf{T}^{(i)}$:

$$\mathbf{r}' = (|r_{2,c}| * \mathbf{r}_1) + (|r_{1,c}| * \mathbf{r}_2)$$

For the example, these steps result in the above $\mathbf{T}^{(1)}$ matrix.

Formulation Continued

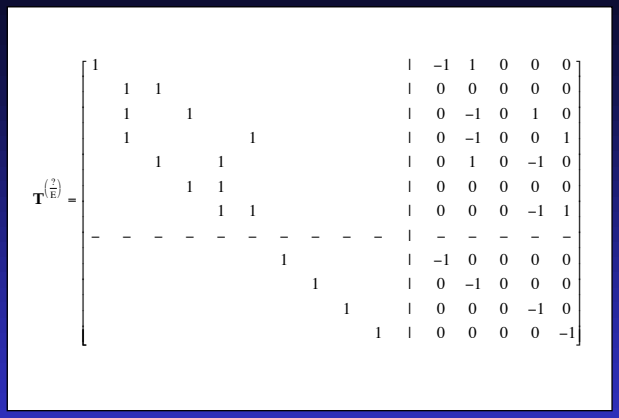
- Check for rows that are non-negative combinations of other rows and eliminate
- Repeat tableau formulation procedure and non-negative combination check for all unconstrained metabolites ending with $T^{(\mu)}$

In the end, the number of rows in $T^{(\mu)}$ will equal the number of extreme pathways

4. For all of the rows added to $\mathbf{T}^{(x)}$ in steps 2 and 3 check to make sure that no row exists that is a non-negative combination of any other sets of rows in $\mathbf{T}^{(x)}$. One method used is as follows: let $A(i)$ equal the set of column indices, j, for which the elements of row i equal zero. Then check to determine if there exists another row (h) for which $A(i)$ is a subset of $A(h)$.

5. With the formation of $\mathbf{T}^{(x)}$ complete repeat steps 2 through 4 for all of the metabolites that do not have an unconstrained exchange flux operating on the metabolite, incrementing x by one up to m. The final tableau will be $\mathbf{T}^{(\mathsf{m})}$. (In this example there is only one such metabolite so we do not need to iterate through steps 2-4 again. Therefore $\mathbf{T}^{(\mathsf{m})}$ equals $\mathbf{T}^{(1)}$ as in Eq.B.3.) Note that the number of rows in $\mathbf{T}^{(\mathsf{m})}$ will be equal to (k), the number of extreme pathways.

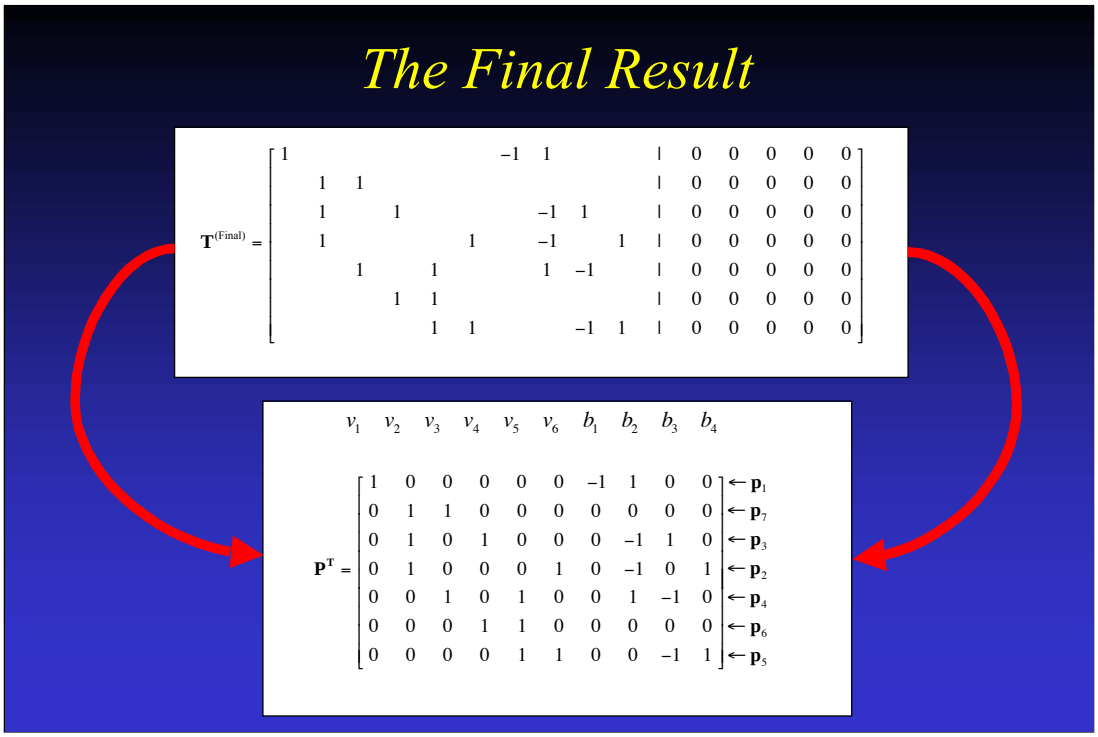
The Final Step



- Use $T^{(E)}$ to "zero" out the right hand side of $T^{(\mu)}$ by adding or subtracting rows from $T^{(E)}$.

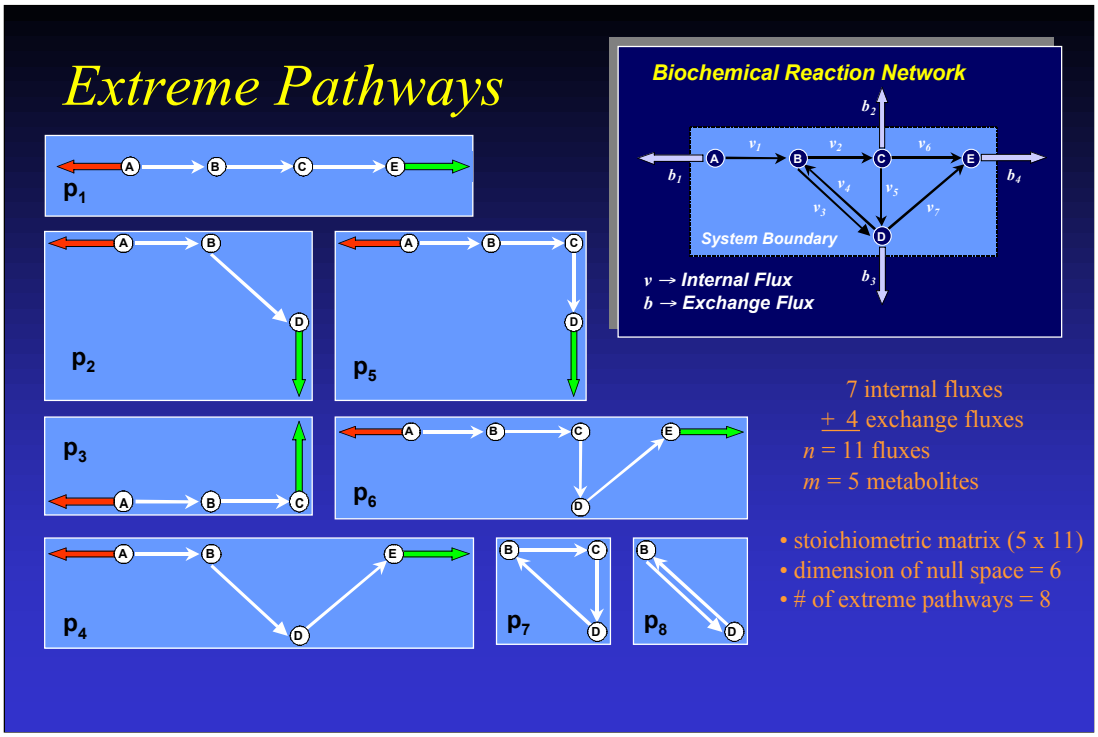
6. Starting in the n+1 column (or the first non-zero column of the right side), if $T_{i,(n+1)}$ does not equal zero, then add the corresponding nonzero row from $\mathbf{T}^{(E)}$ to row i so as to produce a zero in the (n+1) column. This is done by simply multiply the corresponding row in $\mathbf{T}^{(E)}$ by $T_{i,(n+1)}$ and adding this row to row i. Repeat this procedure for each of the rows in the upper portion of the tableau so as to create zeros in the entire upper portion of the (n+1) column. When finished remove the row in $\mathbf{T}^{(E)}$ corresponding to the exchange flux for the metabolite just balanced.

7. Follow the same procedure in step 7 for each of the columns in the right portion of the tableau contain non-zero entries. (In this example we need to perform step 7 for every column except the middle column of the right side which corresponded to metabolite C).



THE FINAL RESULT

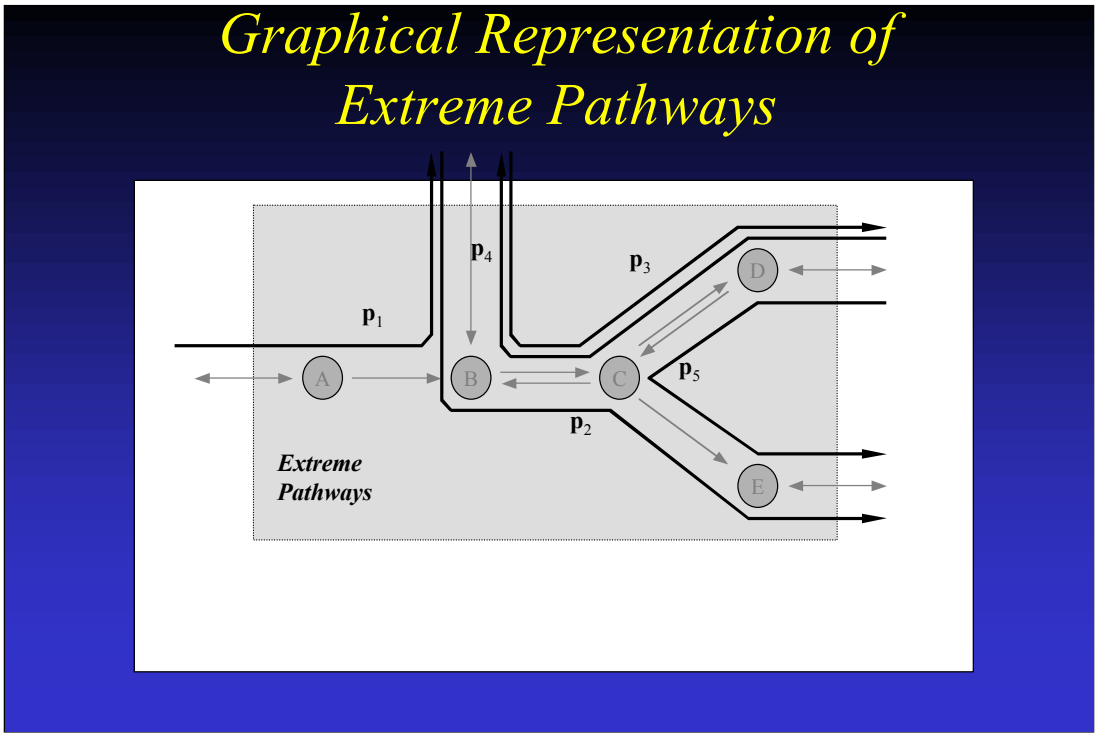
The final tableau, $\mathbf{T}^{(\text{Final})}$, will contain the transpose of the matrix **P** containing the extreme pathways in place of the original identity matrix.



WHAT DO THE EXTREME PATHWAYS LOOK LIKE?

Here is an example of what extreme pathways look like for the hypothetical reaction network shown above. In this case the stoichiometric matrix is 5 by 11 with the dimensions of the null space equaling 6 and the number of extreme pathways equaling 8. We can see here that 6 of these pathways are actually performing net reactions which consume a metabolite to produce another, however there are two pathways here that are only internal cycles within the network. So we see the necessity for a classification scheme for these pathways.

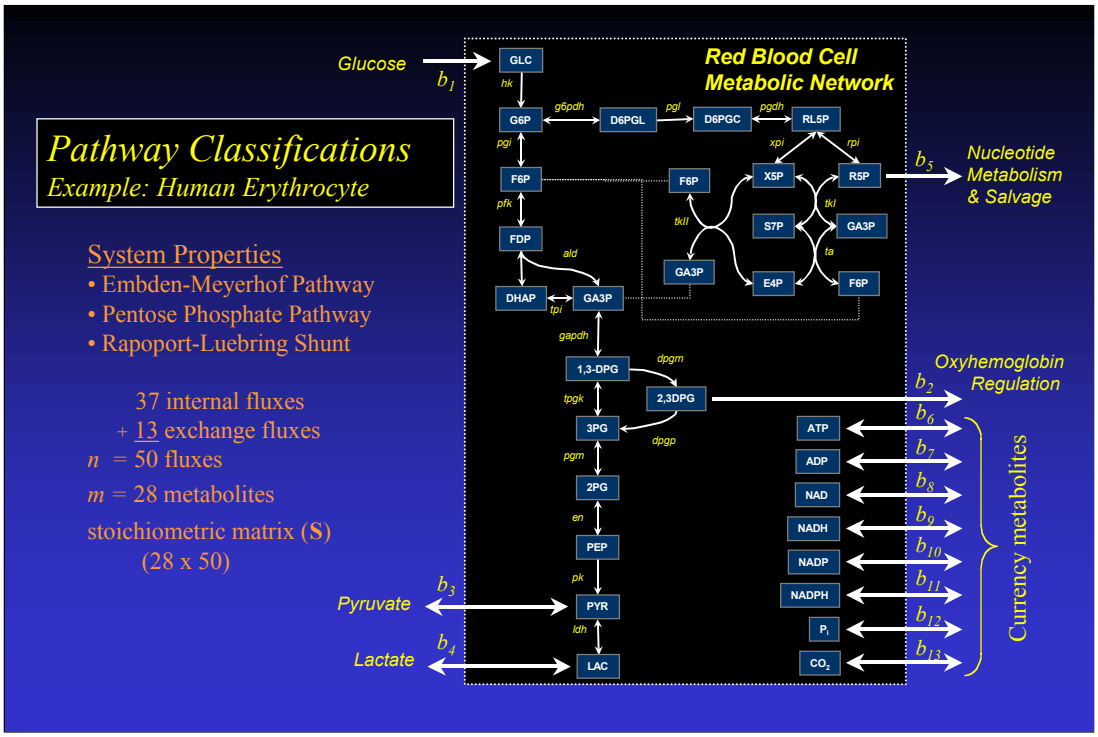
Compare these pathways to the linked output pathways that appear on a later slide.



GRAPHICAL REPRESENTATION OF EXTREME PATHWAYS

This slide shows the extreme pathways evaluated for a sample system. These pathway vectors will be the edges of a 7-dimensional cone. All admissible steady state solutions lie in this cone.

We need to work on software for good representation of these pathways.



THE RED BLOOD CELL

To illustrate the different classifications of pathways I will use the red blood cell as a limited but biologically realistic example. Here we have a partial metabolic reaction system for the red blood cell which is composed of the

• Embden-Meyerhof pathway,

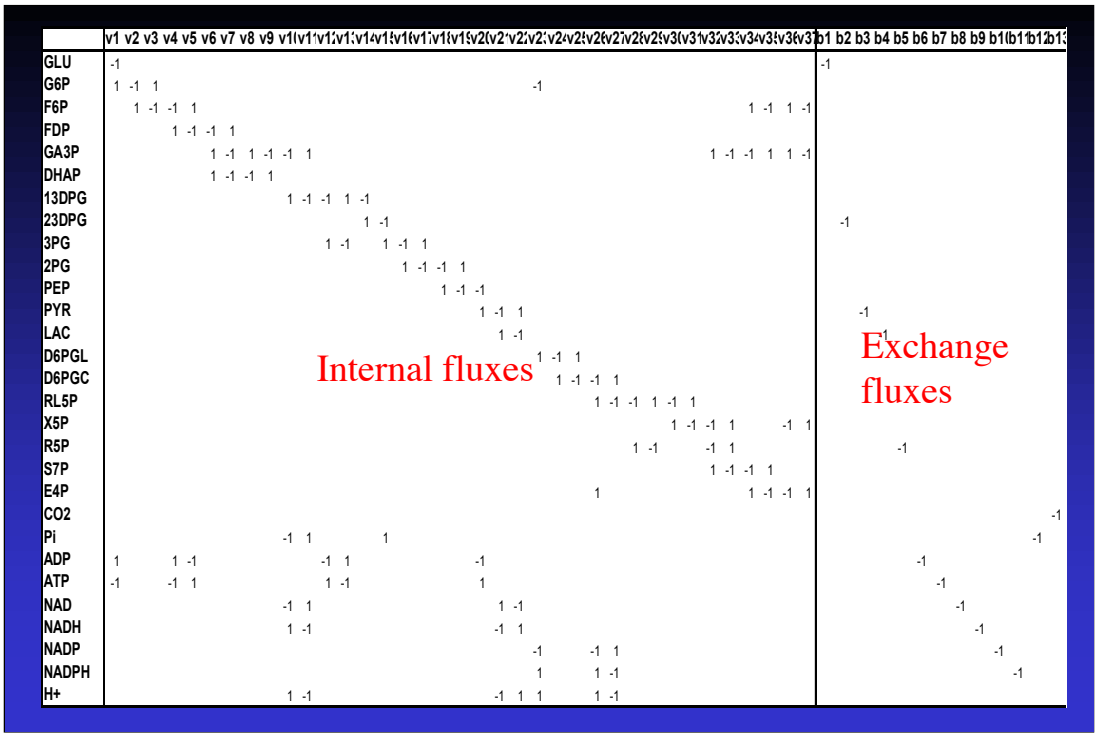
•Pentose phosphate pathway, and

•Rapoport-Luebring shunt.

The characteristics of this metabolic system are shown in this slide.

Note that there is a distinction made between primary metabolites and currency metabolites or those which are mainly involved in energy & redox exchange in the cell. So once we construct the stoichiometric matrix for this system we simply find the independent pathways which define the edges of the flux cone.

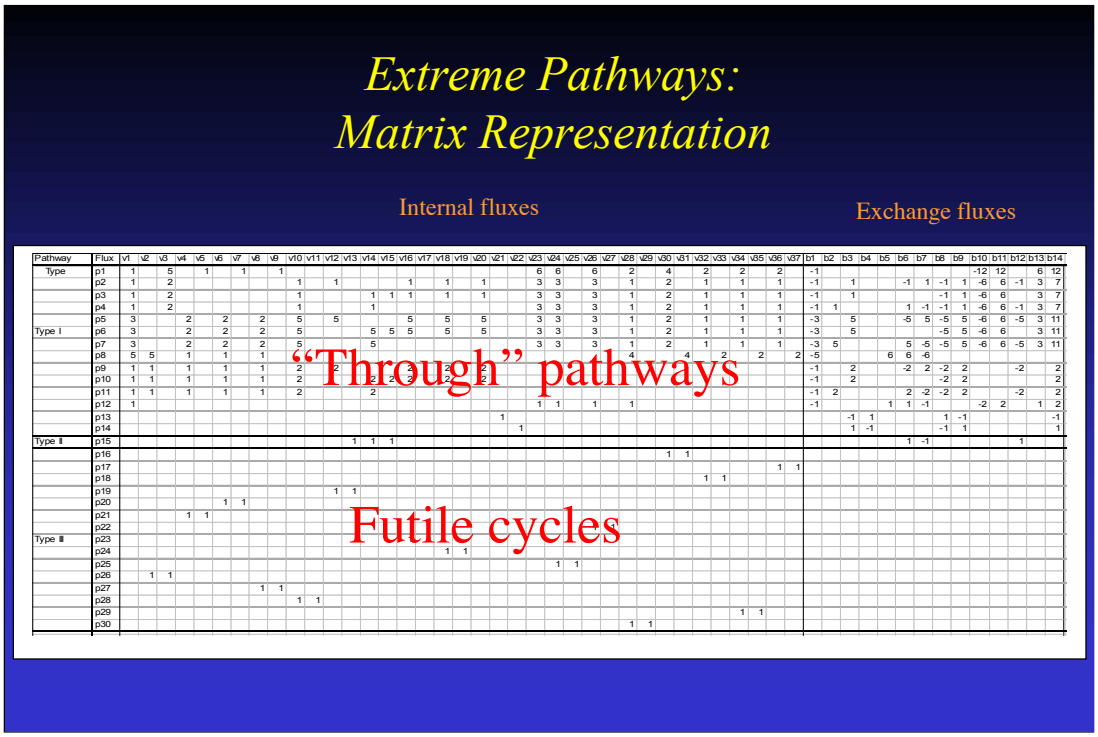
These extreme pathways then together can be used to describe every possible state which this system can operate in.



THE RED CELL STOICHOMETRIC MATRIX

There is the stoichiometric matrix for the red blood cell. Notice that the matrix is sparse and has only 1 and -1 non-zero entries.

It is partitioned based on the internal flux/exchange flux distinction as shown above.



EXTREME PATHWAYS FOR THE RED BLOOD CELL

The calculated extreme pathways for the red cell system are shown.

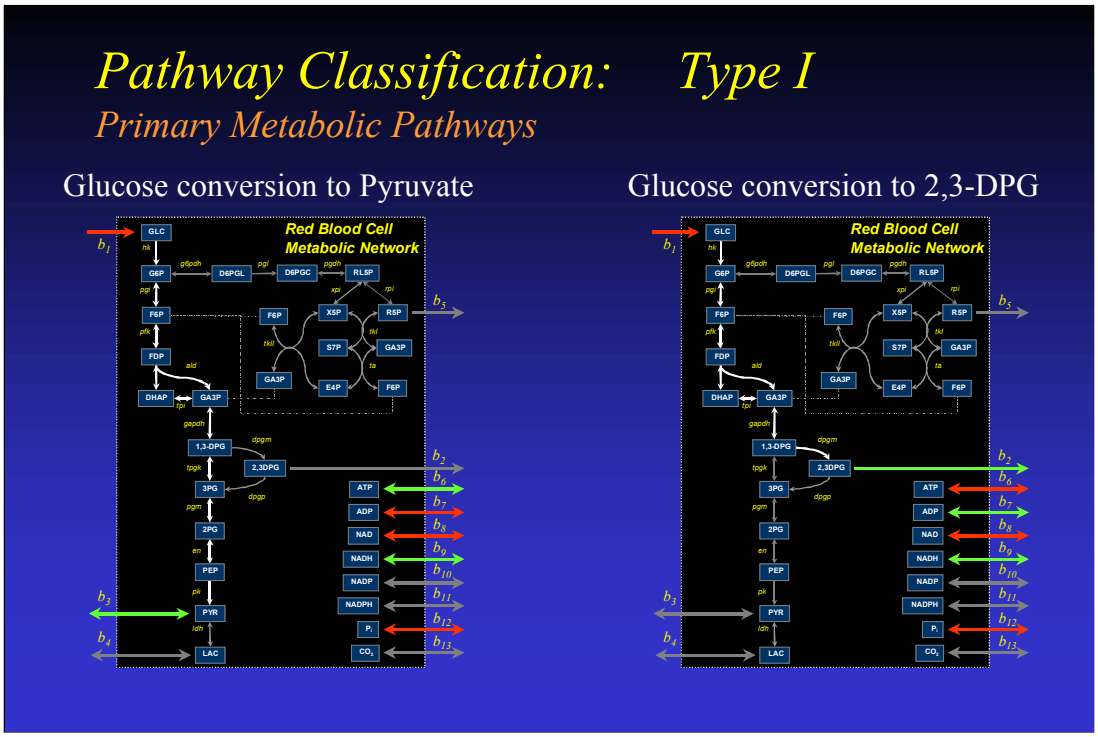
Note that there is a group of pathways (at the bottom) that have no exchange reactions. They are thus internal to the cell with no connections to the outside. We call these pathways type III pathways.

There is one pathway whose only exchange fluxes involve cofactors. This is an internal futile cycle. We call these type II pathways.

The rest are 'through' pathways, that connect an input to an output. These are the type I pathways.

Classifying Pathways

- Type I: Primary Metabolic Pathways
- Type II: "Futile" Cycles
  - only currency exchange fluxes active
- Type III: Reaction Cycling
  - no active exchange fluxes

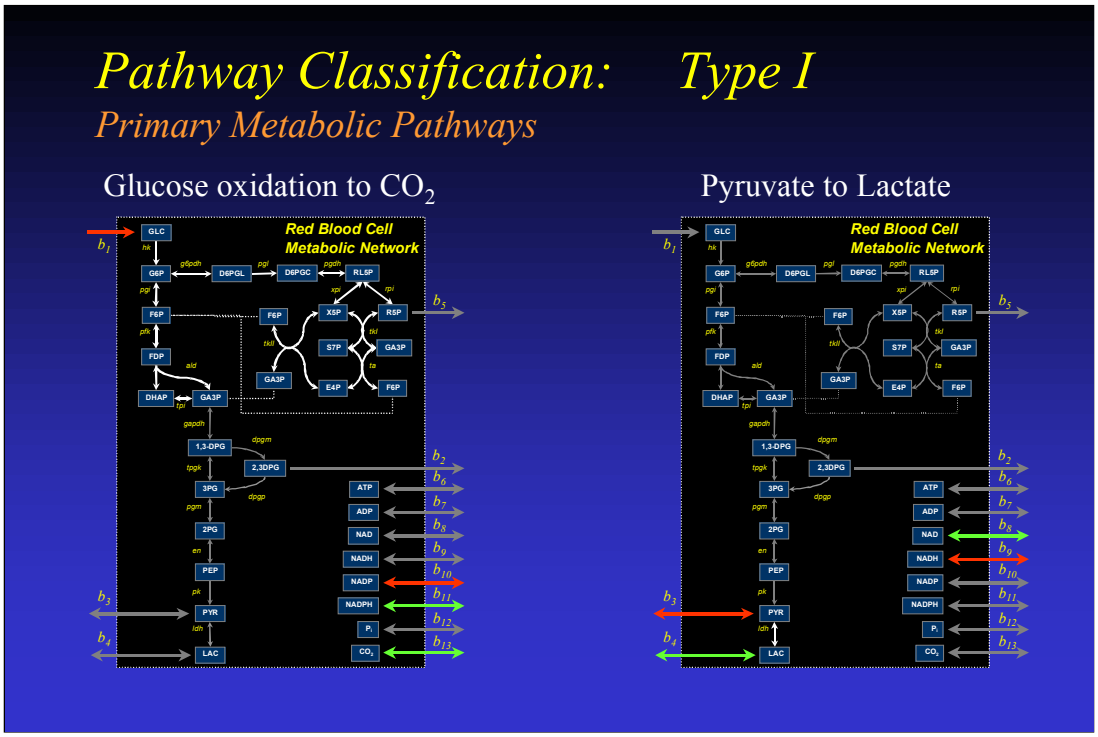


TYPE I PATHWAYS

The first type of pathways that are generated are what we refer to as primary pathways and these are the types of pathways that first come to mind when thinking about a metabolic map. These are simply pathways that connect an input to an output. The only requirement of the pathway is that one of the primary exchange fluxes must be active. Here are two examples of primary metabolic pathways that are extreme pathways on the cone. The green arrow denote the production of a metabolite by the pathway and the red arrows indicate the consumption of a metabolite while the white arrows indicate the internal fluxes which are operating.

The first example is simply the conversion of glucose to pyruvate using the glycolytic pathway . This is basically the glycolytic pathway that picks up glucose and secretes pyruvate, and producing both ATP and NADH.

The second pathway is the production of 2,3DPG from the Rapoport-Luebering shunt. This pathway becomes active when more 2,3DPG needs to be produced such as when one goes through changes in altitude and the oxygen binding characteristics of hemoglobin need to be changes. This will always be a low flux pathway.



MORE TYPE I PATHWAYS

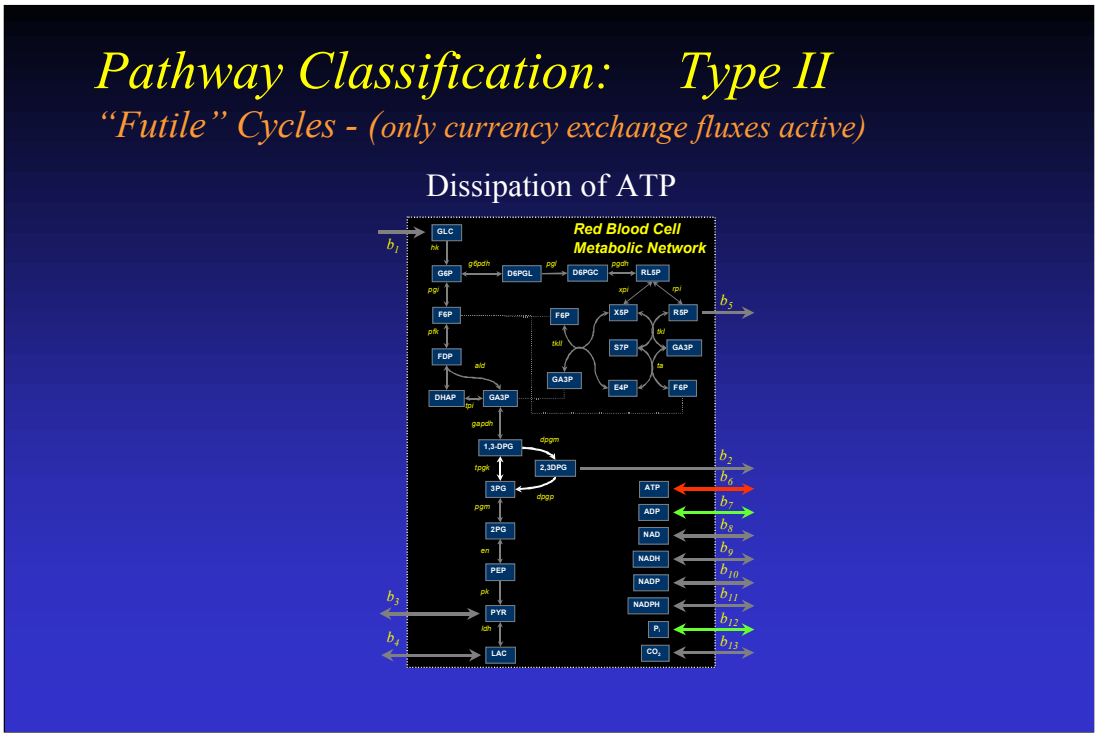
Here are two more examples of type I extreme pathways pathways:

•the complete oxidation of glucose to CO2 through the cycling of the pentose phosphate pathway producing NADPH and

•primary pathway which consists of only one reaction converting pyruvate into lactate used to balance the NAD/NADH ratio of the cell.

As you can see each of these primary metabolic pathways has a functional role in the cell and therefore these pathways can be used to interpret the functional attributes and activities of red cell metabolism in this case.

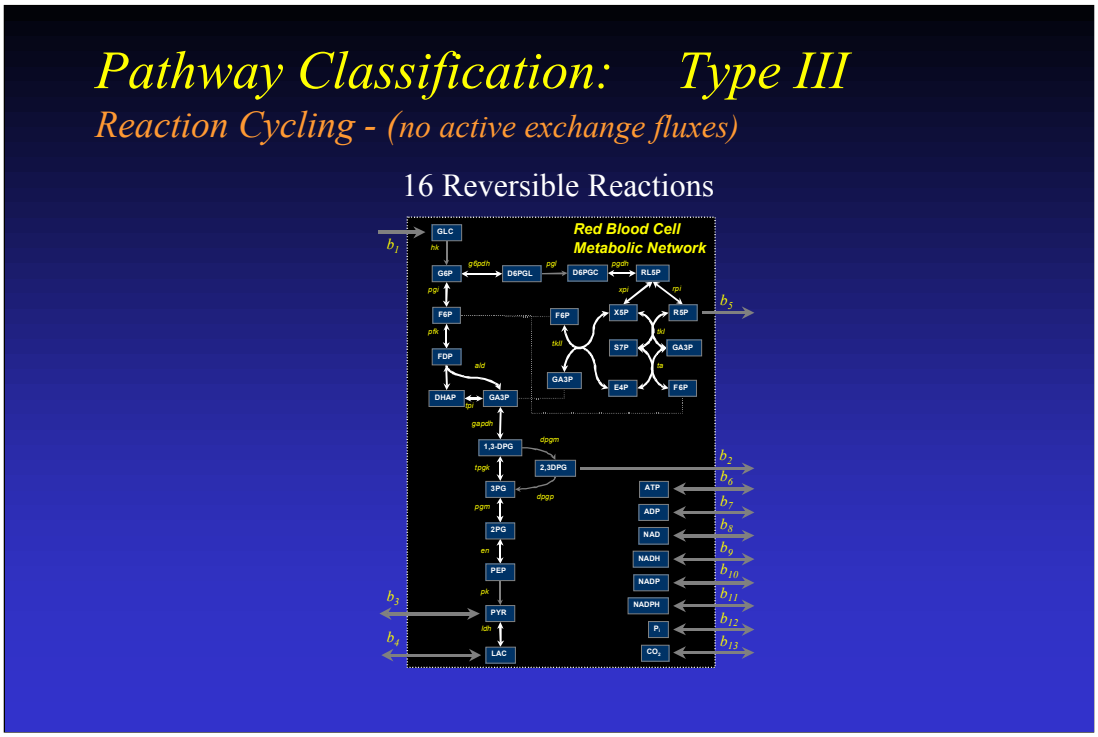
There are a total of 14 type I pathways in this simple red cell model and the others will not be discussed in detail.



TYPE II PATHWAYS

The second type of pathway is what is commonly referred to as a futile cycle in the truest sense of the word futile. In these pathways only the exchange fluxes for the currency metabolites are active. In this system there exists one futile cycle which occurs around the Rapoport-Luebering shunt. The net result of this pathway is the conversion of ATP into ADP and releasing inorganic phosphate which is obviously dissipating metabolic energy.

There is one futile cycle that operate in this system.



TYPE III PATHWAYS

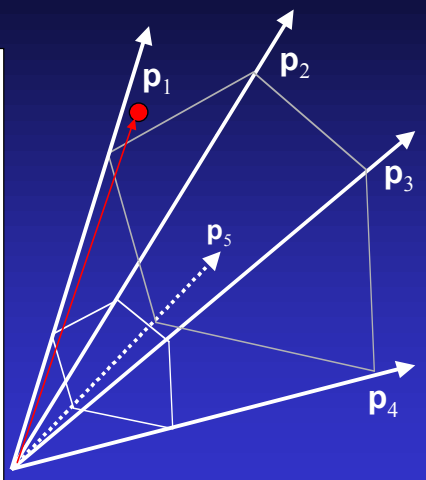
The third type of pathway consists of reversible cycles which are mainly the result of reversible reactions be characterized by a forward reaction and a separate reverse reaction. These pathways show no activity in any of the exchange fluxes. In this map all 16 reversible reactions are highlighted white. While these pathways are essentially generating vectors they can effectively be dismissed in any further analysis of the system as they have no net effect on the production capabilities of the system as they influence none of the exchange fluxes.

These pathways will become important later when we examine temporal decomposition of this system. A fast internal pathway leads to an 'equilibration' or the 'tying together' of two or more concentrations that then can be 'pooled' together to form an aggregate dynamic variable. Some simple examples of this feature were show in the last lecture.

Together all of the extreme pathways in a system fall under one of these three classifications of pathways and they all are edges of the cone determining the flux space.

Pathway Utilization in the Red Cell:
Geometric representation

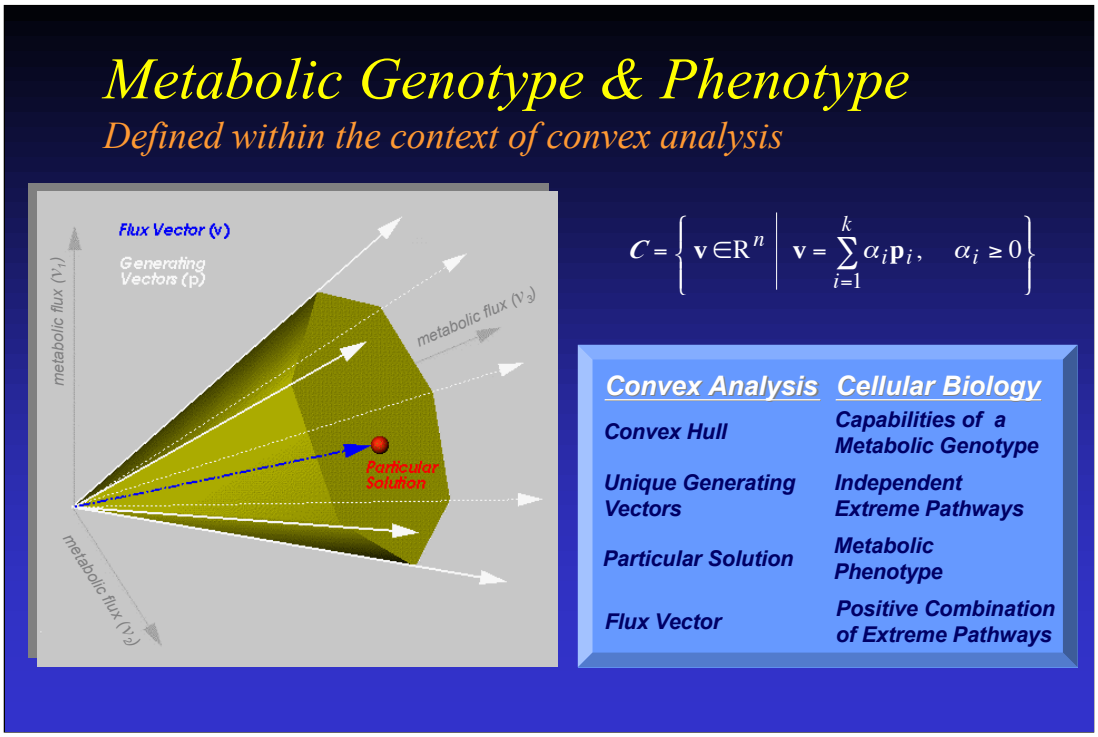
Basis Pathway	Net reaction Equation	Primary Functional Attribute
p1	glucose + 2 Pi +2 ADP -> 2 lactate +2 ATP	ATP energy production for metabolic energy
P2	glucose + 2 NAD+ -> 2 pyruvate + 2 NADH + 2 H+	NADH energy production for methemoglobin reduction
P3	glucose + 2 Pi + 2 ADP + 2 NAD+ -> pyruvate +2 ATP 2 NADH+ + 2 H+	ATP production for metabolic energy and NADH production for methemoglobin reduction
P4	glucose + 2 ATP +2 NAD+ -> 2 2,3DPG +2 Pi +2 ADP +2 NADH +2 H+	2,3DPG production for oxyhemoglobin modulation
P5	glucose + ATP + 2 NADP+ -> R5P +CO2 + ADP + 2 NADPH + 2 H+	R5P production for adenosine salvaging
P6	glucose +12 NADP+ -> 6 CO2 + 12 NADPH + 12 H+	NADPH energy production for glutathione reduction and subsequent antioxidant activity
p7-p22	(no net reaction)	



NORMAL OPERATION OF THE RED CELL MATABOLIC NETWORK

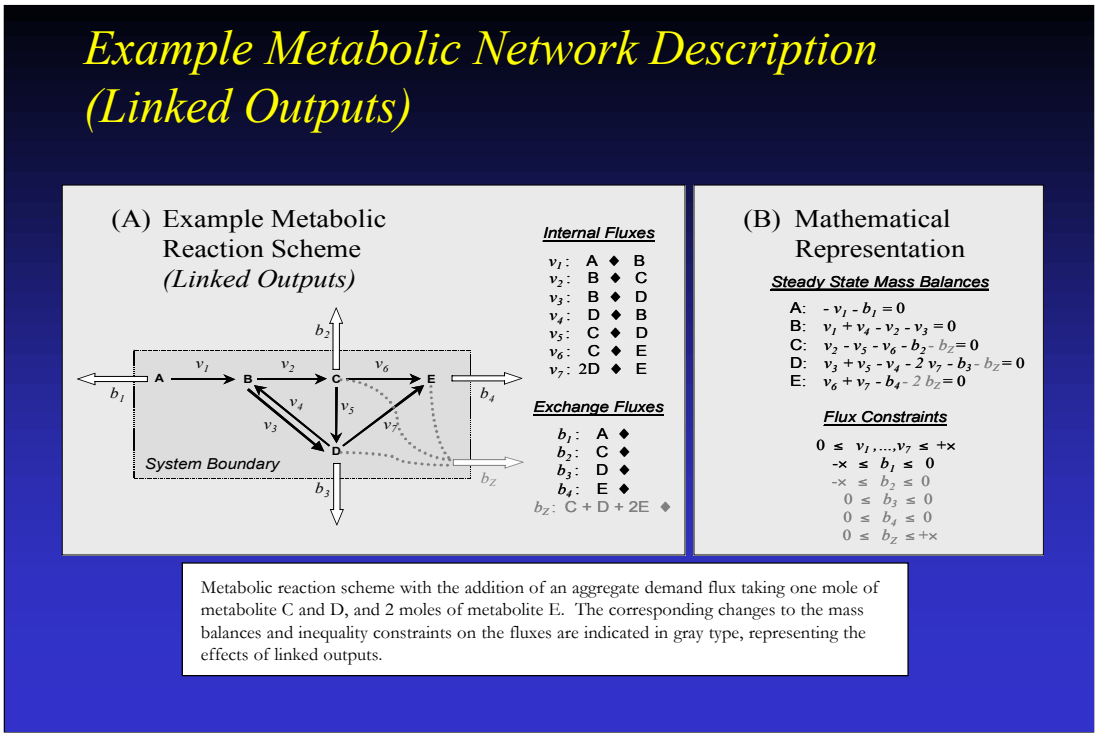
The nominal physiological steady state of the red cell metabolism is to produce ATP to run the Na/K pump to maintain the osmotic balance across the cells membrane. If other pathways are activated to produce, say 2,3 DPG with altitude change, two of these pathways would contribute to the flux map and the solution could 'creep' towards an edge of the flux cone.

This flux solution can be obtained from the full dynamic red cell model (downloadable in a MATHEMATICA form from http://gcrg.ucsd.edu) or from a flux balance model where the demands of the pump are stated and the uptake rate of glucose in minimized.



NIFTY INTERPRETATION OF THE FLUX CONE

So what does this all mean from a biological point of view. Here is the geometric interpretation of the flux cone in which every point described by the equation given. The entire flux cone actually corresponds to the capabilities of a reaction network and hence the defined metabolic genotype. What can the reconstructed network do, what can it not do? Each one of the generating vectors corresponds to an extreme pathway which the cell could theoretically control to reach every point in the flux cone. Now a particular point within this flux cone corresponds to a given flux distribution which represents a particular metabolic phenotype. The actual flux vector describing that point can be thought of as a positive combination of these extreme pathways. So you may think of these pathways as being theoretically turned on and off to reach a particular metabolic phenotype. Once again this means that every phenotype which the system can exhibit is a combination of these pathways which are then turned on or off. It's that simple. With these pathways we can describe all of the capabilities of the metabolic system and so we may say that these pathways represent the underlying pathway structure of the system.



TOWARDS PHYSIOLOGICAL FUNCTIONS: LINKED OUTPUTS

Under changing substrate/supply conditions metabolic networks are continuously faced with balanced sets of biosynthetic demands (i.e. production of amino acids, nucleotides, phospho-lipids, as well as metabolic energy and redox potential). Effectively this means that the network must generate a balanced rate through the exchange fluxes for the particular metabolites required to meet these demands. To assess the systemic performance of a network in meeting balanced biosynthetic demands, an exchange flux is introduced into a network. Additional constraints must also be added to the network to effectively close the material balances on the metabolites participating in the biosynthetic demand (or growth) flux. The introduction of a new flux and the associated restriction of existing fluxes will alter the mass balances and linear inequalities of the network, and subsequently alter the pathway structure. To distinguish between the two different forms (all material balances closed with a growth flux are included versus no growth flux and material balances not closed on biosynthetic precursors) we consider a system without a biosynthetic demand flux to have free outputs, and the consideration of balanced network demands defines a linked output system. For the example system, we introduce the growth exchange flux b_Z which is described in the drawing below. This flux must then be included into the mass balances. Additionally we change the constraints on the specific exchange fluxes for metabolites C, D, and E so as not to allow them to exit the system. All of the changes from the open to the closed system are highlighted in the figure.

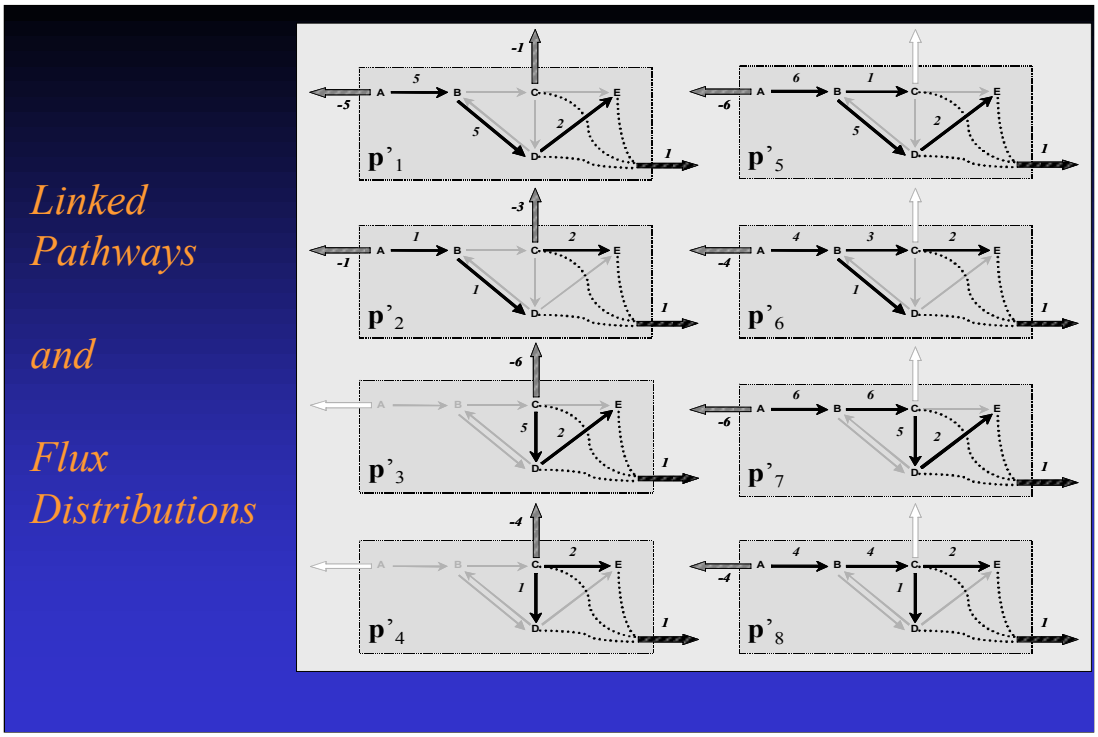
Extreme Pathways (Linked Outputs)

The 10 extreme pathway vectors for the linked output description of the example network. The first eight pathways correspond to type I pathways while the last two pathways are of type III. Pathway equivalencies between the free and linked output systems are provided for each pathway

Pathway	***Internal Fluxes***							***Exchange Fluxes***					***Pathway Equivalences***
Number	v_1	v_2	v_3	v_4	v_5	v_6	v_7	b_1	b_2	b_3	b_4	b_z	***Linked ~ Free***
$\mathbf{p'}_1$	5	0	5	0	0	0	2	-5	-1	0	0	1	$\mathbf{p'}_1 \sim \mathbf{p}_2 + 2\,\mathbf{p}_3$
$\mathbf{p'}_2$	1	0	1	0	0	2	0	-1	-3	0	0	1	$\mathbf{p'}_2 \sim \mathbf{p}_2 + 2\,\mathbf{p}_6$
$\mathbf{p'}_3$	0	0	0	0	5	0	2	0	-6	0	0	1	$\mathbf{p'}_3 \sim \mathbf{p}_4 + 2\,\mathbf{p}_5$
$\mathbf{p'}_4$	0	0	0	0	1	2	0	0	-4	0	0	1	$\mathbf{p'}_4 \sim \mathbf{p}_4 + 2\,\mathbf{p}_6$
$\mathbf{p'}_5$	6	1	5	0	0	0	2	-6	0	0	0	1	$\mathbf{p'}_5 \sim \mathbf{p}_1 + \mathbf{p}_2 + 2\,\mathbf{p}_3$
$\mathbf{p'}_6$	4	3	1	0	0	2	0	-4	0	0	0	1	$\mathbf{p'}_6 \sim 3\,\mathbf{p}_1 + \mathbf{p}_2 + 2\,\mathbf{p}_6$
$\mathbf{p'}_7$	6	6	0	0	5	0	2	-6	0	0	0	1	$\mathbf{p'}_7 \sim 6\,\mathbf{p}_1 + \mathbf{p}_4 + 2\,\mathbf{p}_5$
$\mathbf{p'}_8$	4	4	0	0	1	2	0	-4	0	0	0	1	$\mathbf{p'}_8 \sim 4\,\mathbf{p}_1 + \mathbf{p}_4 + 2\,\mathbf{p}_6$
$\mathbf{p'}_9$	0	0	1	1	0	0	0	0	0	0	0	0	$\mathbf{p'}_9 \sim \mathbf{p}_7$
$\mathbf{p'}_{10}$	0	1	0	1	1	0	0	0	0	0	0	0	$\mathbf{p'}_{10} \sim \mathbf{p}_8$

LINKED PATHWAYS ARE COMPINATIONS OF SINGLE OUTPUT PATHWAYS

For the linked output system there are 10 extreme pathways (8-type I and 2-type III pathways). The complete pathway vectors are provided in this table for pathway #1 through #8 (pathway #7 and #8 are type III pathways that exhibit no activity for the exchange fluxes, i.e. internal cycles).



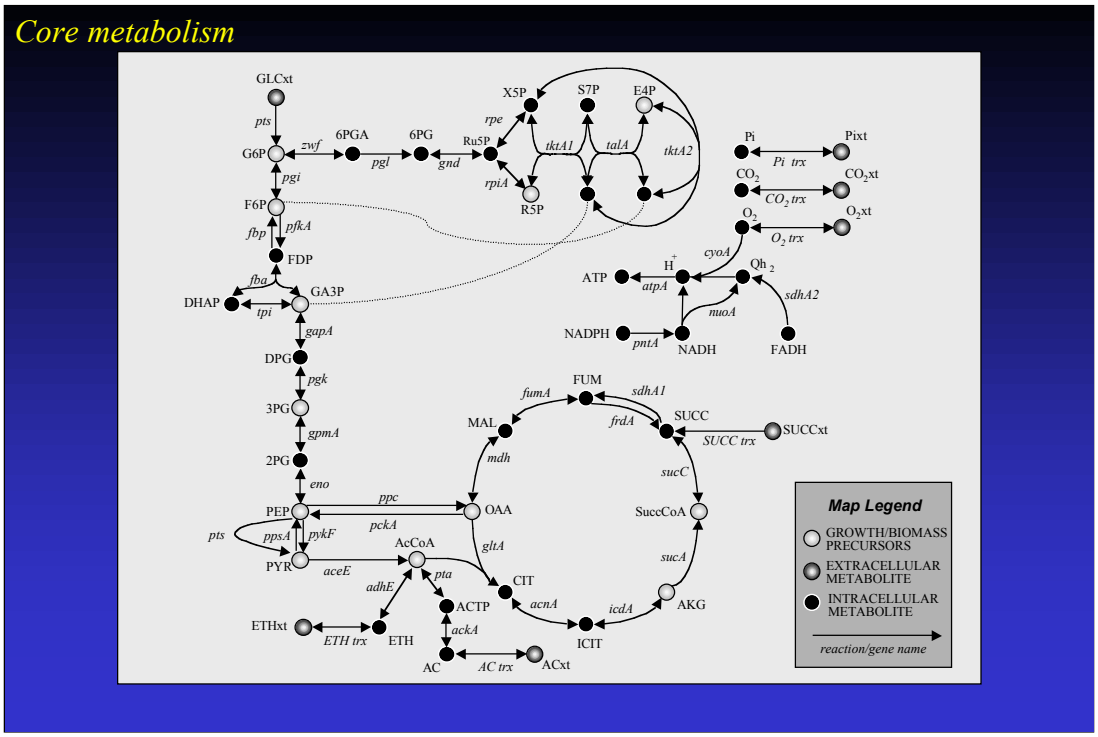
GRAPHICAL REPRESENTATION OF LINKED PATHWAYS

The pathway distributions are also illustrated graphically in this figure. Note that the extreme pathways for the linked outputs are systemic flux distributions that meet the balanced set of demands represented by the growth flux (b_z). These extreme pathways are non-negative combinations of the extreme pathways for the free output system. This leads to the definition of pathway equivalences that relate the free output system pathways to the linked output system.

The first two produce the required output using two inputs (A and C), the next two only from C, and the last four from A alone.

Compare these to the single output pathways shown on an earlier slide for the same network.

The linked pathways are no longer 'linear' or 'one-dimensional' entities, but actual flux maps.



A PSEUDO-REALISTIC METABOLIC NETWORK

A schematic of the central metabolic network of *E. coli* is depicted in this diagram. The network is comprised of glycolytic reactions, pentose phosphate shunt, and the tricarboxcylic acid cycle without the glyoxylate shunt along with the necessary transport reactions. The genes whose gene products carry out the reactions are used as the reaction names in most cases. The necessary electron transport chain reactions are included with the P/O ratio of 4/3. The system is comprised of 53 metabolites, 78 internal fluxes, and 8 exchange fluxes.

Note that this network does not completely describe central metabolism in *E. coli*. This representation has been chosen as a compromise between successfully representing the basic aspects of central metabolism and providing a useful example of the combined approach to study metabolic systems.

This system and its linked pathways are very insightful as we shall see on the ensuing slides.

Functional characteristics of the reduced set of 12 extreme pathways calculated for succinate as the sole carbon source for the *E. coli* metabolic network with linked outputs. Pathways are ordered based on the activity of the growth flux normalized by the succinate uptake. Pathway numbers coincide with the original numbers of the pathway vectors retained from the complete set. All of the exchange flux values are normalized to the succinate intake level (negative values are relative uptake ratios, positive values are production ratios). Exchange flux abbreviations: SUCC-succinate, ETH-ethanol, AC-acetate, PI-inorganic phosphate, CO_2-carbon dioxide, O_2-oxygen, GRO-biomass/growth flux .

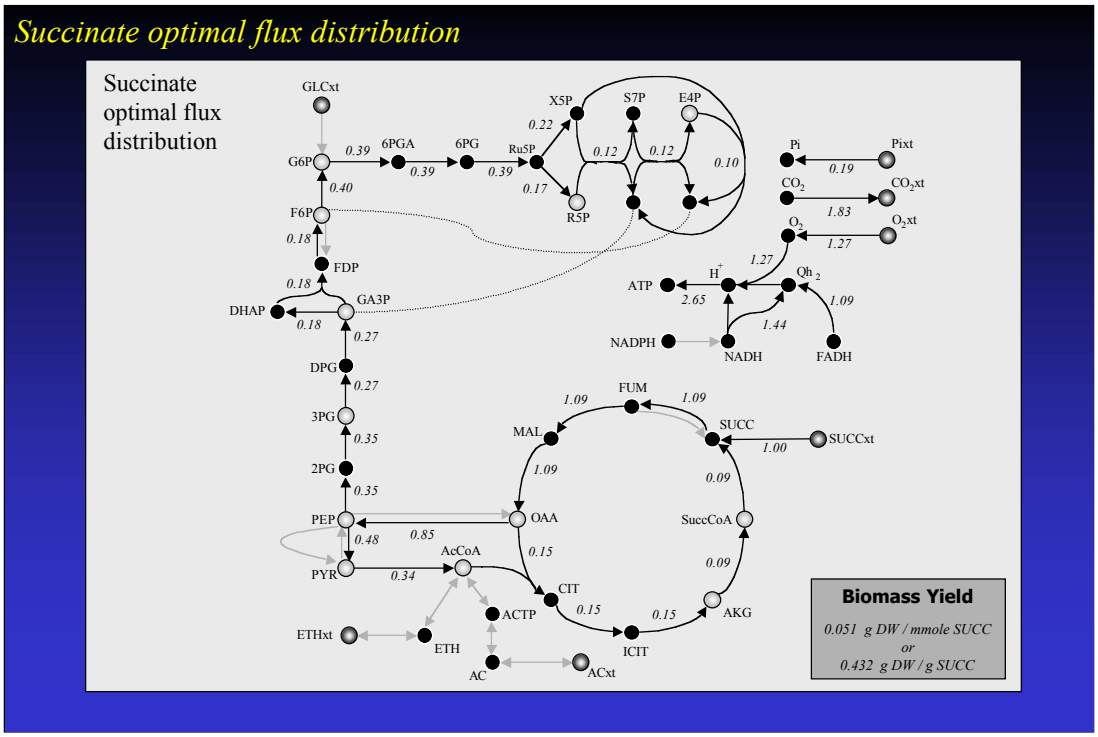
Pathway Number	*Exchange Fluxes* SUCCxt/ SUCCxt	ETHxt/ SUCCxt	ACxt/ SUCCxt	GRO/ SUCCxt	PIxt/ SUCCxt	CO2xt/ SUCCxt	O2xt/ SUCCxt	*Net Pathway Reaction Balance*
1	-1.000	0	0	0.051	-0.188	1.825	-1.267	SUCCxt + 0.188 PIxt + 1.267 O2xt ? 0.051 GRO + 1.825 CO2xt
10	-1.000	0	0	0.049	-0.182	1.895	-1.338	SUCCxt + 0.182 PIxt + 1.338 O2xt ? 0.049 GRO + 1.895 CO2xt
20	-1.000	0	0	0.047	-0.172	1.696	-1.142	SUCCxt + 0.172 PIxt + 1.142 O2xt ? 0.047 GRO + 1.696 CO2xt + 0.158 ACxt
3	-1.000	0	0.000	0.034	-0.125	2.553	-2.014	SUCCxt + 0.125 PIxt + 2.014 O2xt ? 0.034 GRO + 2.553 CO2xt
7	-1.000	0	0.000	0.033	-0.121	2.600	-2.062	SUCCxt + 0.121 PIxt + 2.062 O2xt ? 0.033 GRO + 2.6 CO2xt
12	-1.000	0	0	0.032	-0.117	2.644	-2.108	SUCCxt + 0.117 PIxt + 2.108 O2xt ? 0.032 GRO + 2.644 CO2xt
16	-1.000	0.000	0	0.031	-0.114	2.679	-2.144	SUCCxt + 0.114 PIxt + 2.144 O2xt ? 0.031 GRO + 2.679 CO2xt
19	-1.000	1	0.000	0.025	-0.092	1.837	-0.759	SUCCxt + 0.092 PIxt + 0.759 O2xt ? 0.025 GRO + 1.837 CO2xt + 0.549 ETHxt
23	-1.000	0.000	0.000	0.000	0.000	4.000	-3.500	SUCCxt + 3.5 O2xt ? 4.0 CO2xt
27	-1.000	0.000	1	0.000	0.000	2.000	-1.500	SUCCxt + 1.5 O2xt ? 2.0 CO2xt + 1.0 ACxt
31	-1.000	1.000	0	0.000	0.000	2.000	-0.500	SUCCxt + 0.5 O2xt ? 2.0 CO2xt + 1.0 ETHxt
35	-1.000	0.000	0	0.000	0.000	4.000	-3.500	SUCCxt + 3.5 O2xt ? 4.0 CO2xt

PATHWAYS FOR GROWTH ON SUCCINATE

The pathway analysis was performed with succinate as the sole carbon source for the system, generating the complete set of 66 extreme pathways (36 type I and 30 type III). To generate a reduced set of pathways that represents the full capabilities of the network, we retain only pathways from these sets that utilized the ATP drain flux instead of one of the three futile cycles; i.e. (*pfkA*/*fbp*), (*pckA*, *ppc*), (*pykF*,*ppsA*,*adk*). The type III pathways, which are mainly a consequence of the decomposition of reversible reactions into a forward and a reverse reaction, are also removed from consideration as they show no activity in the exchange fluxes. Following this simplification, a reduced set of 12 pathways is generated from the complete set.

The pathways in the table are ordered by the biomass yield that they generate (mg/mol Succinate). The best pathways produces 0.051 biomass units. This represents the optimal use of the network to produce biomass. Note that the next-best pathway produces 0.049 and is in general very similar to the best one. This is a feature that one observes. There are 'bundles' of extreme pathways located 'close' in this high dimensional conical space, which biologically is a reflection of the redundancy of the system.

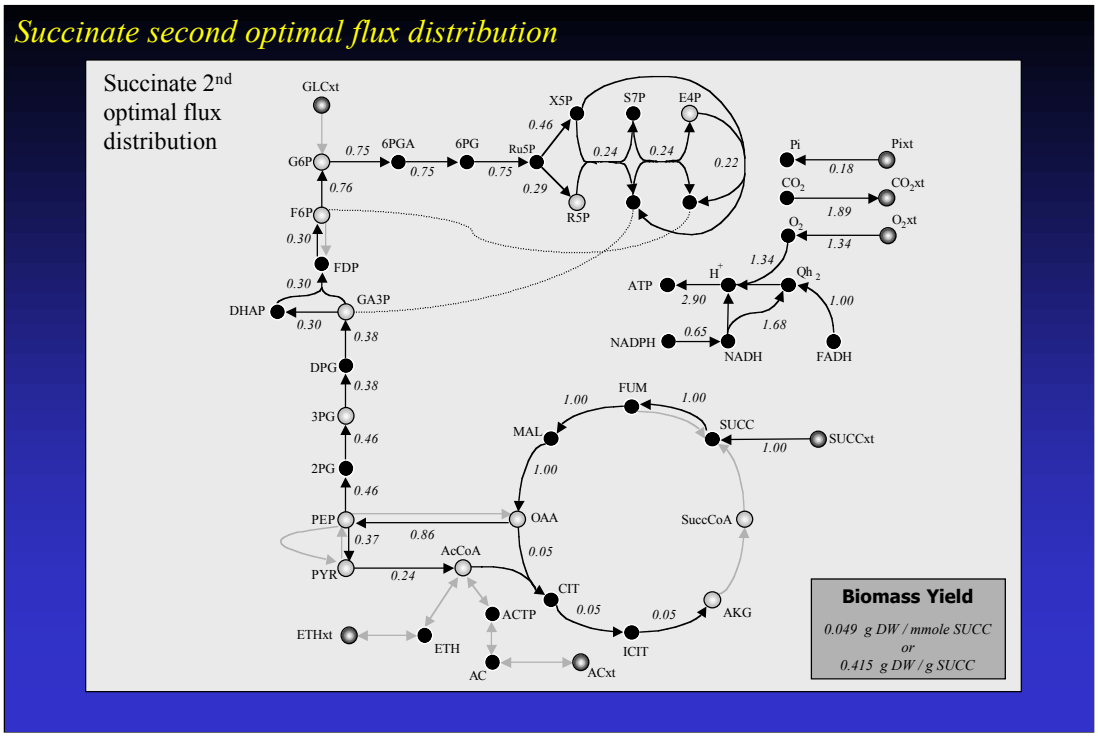
The third pathway represents partially aerobic growth and the secretion of acetate. As we shall see later, if oxygen is limiting, then the growth becomes a combination of this pathway and pathway #1. Note that there are two purely fermentative pathways producing acetate and ethanol respectively.



THE OPTIMAL PATHWAY (FLUX MAP)

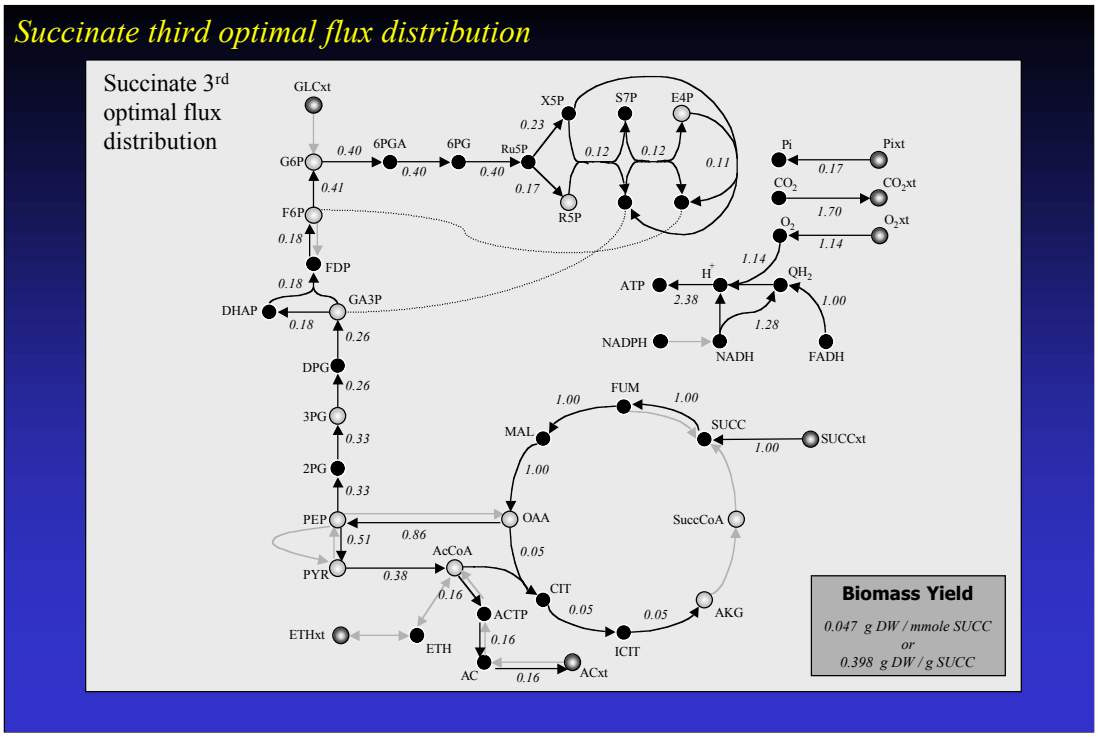
Flux balance analysis can be used to quantitatively examine the system with linked outputs. Geometrically, the constraints imposed on the input values of the exchange fluxes will bound the flux cone by the extreme pathways into a bounded polyhedron. Optimal solutions within this space will then lie on a vertex of the polyhedron. These are the bounded feasible solutions of the linear programming problem.

The flux distributions for growth are calculated on succinate normalized to 1 mmol of substrate. The optimal flux distributions are illustrated in this figure. The optimal biomass yield is 0.051 g DW/mmol succinate, which is identical to the optimal yield calculated from the pathways (pathway #1). This result reveals that the optimal solution lies directly on the vertex of the polyhedron that is defined by the extreme pathway.



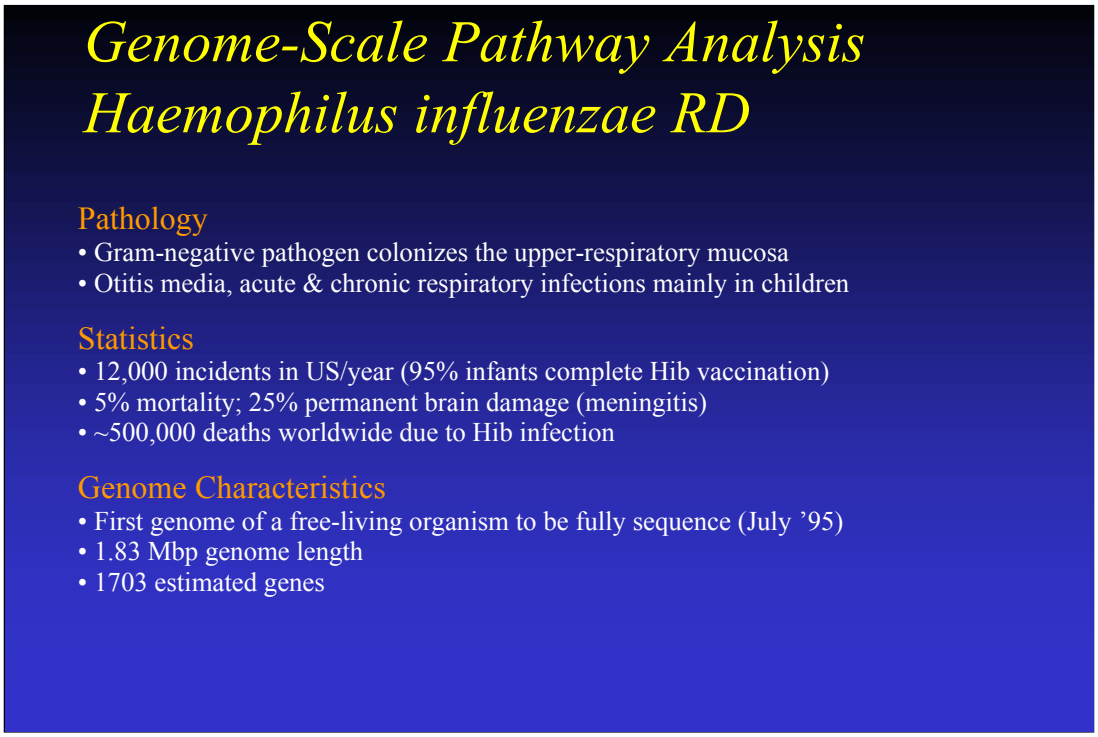
A PSEUDO-OPTIMAL PATHWAY (FLUX MAP)

The second highest optimal flux distribution may also be depicted graphically (pathway #10 in the table). The fluxes in the 2nd optimal flux distribution are identical to the optimal flux distribution as shown in the previous diagram except for the reactions catalyzed by the following enzymes: 2-ketoglutarate dehydrogenase (converting AKG to SuccCoA), succinyl-CoA synthetase (SuccCoA to Succ), and pyridine nucleotide transhydrogenase.The flux values of glycolytic and pentose phosphate pathways are higher and tricarboxylic acid cycle fluxes are lower than the optimal flux distribution.

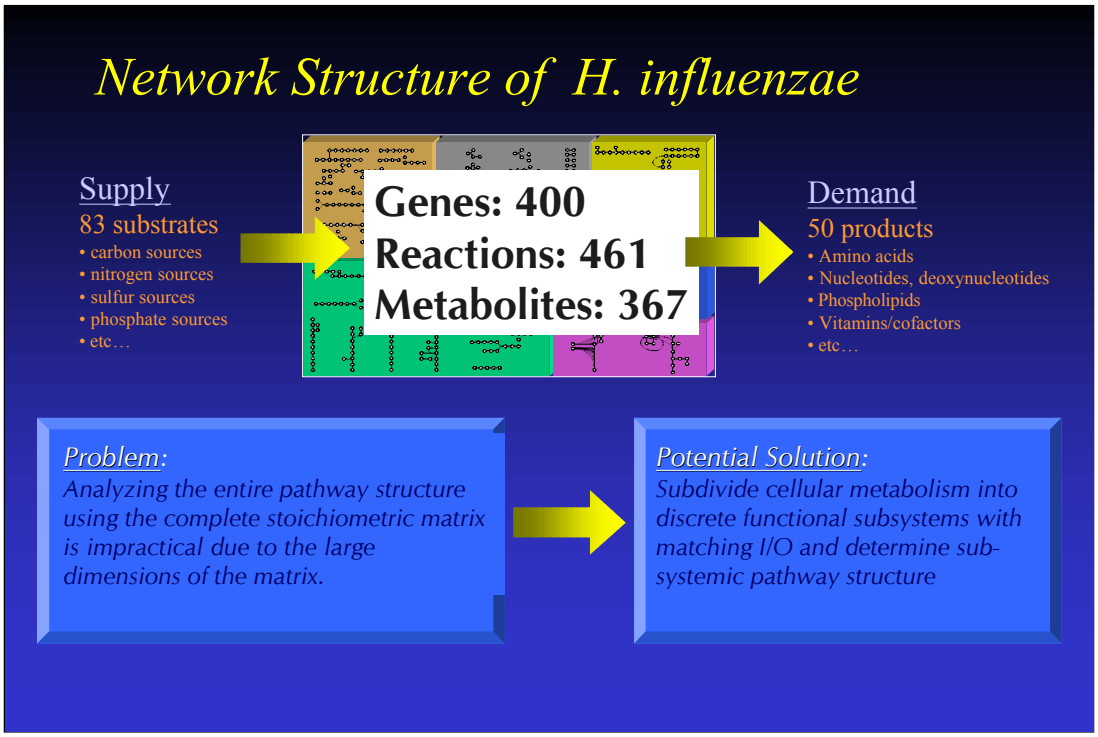


A PATHWAY FOR PARTIALLY AEROBIC GROWTH (ACETATE SECRETION)

The third optimal flux distribution of the core metabolic network of *E. coli* with succinate as the sole carbon source is illustrated on this figure (corresponding to pathway #20). In comparison with the optimal flux distribution, acetate is secreted here and enzymatic activities of 2-ketoglutarate dehydrogenase and succinyl-CoA synthetase are reduced to zero.



H. Influenzae is a gram-negative pathogen which colonizes the upper respiratory track and leads to acute and chronic respiratory infections primarily in children. While this pathogen used to be a serious threat to the health of children, the implementation of effective vaccination programs has significantly reduced the incidences of H. influenzae infections. Although its prominence as a pathogen has decreased, it gained recognition in 1995 as being the first free living organism to have its genome completely sequenced. There genome itself is ~1.8Mbp and contains over 1700 genes. We used the genome along with biochemical and physiological data on the organism to reconstruct metabolism and determine the pathway structure of the metabolic network so as to assess the organism's capabilities and fitness under various simulated conditions.

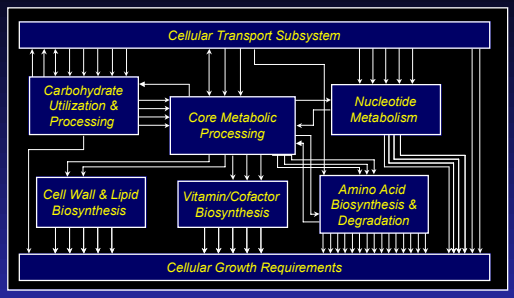


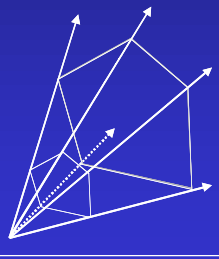
Using the algorithm to determine the stoichiometric matrix and hence metabolic genotype of an organism we assembled the metabolic network for H.influenzae and some of the important numbers are listed here. The network is supplied by 83 potential substrates and requires 50 products to be generated using a series of 461 reactions. As you can see the number of reactions and metabolites which exist within the system is quite large to no surprise. If we were to apply an algorithm to determine all of the genetically independent pathways operating in such systems the number of pathways would be on the order of tens of thousands. Obviously this quite impractical and so a potential solution is to divide and conquer. We can divide cellular metabolism into discrete function subsystems with matching inputs and outputs and determine the pathway structure of each subsystem.

Pathway Structure of H.influenzae

Subsystems of Cellular Metabolism:

- (T) Transport [Electron Trans, Assim./Dissim.]
- (A) Amino Acid Metabolism
- (N) Nucleotide Metabolism
- (V) Vitamin & Cofactor Biosynthesis
- (C) Central Metabolic Processing
- (L) Lipid and Cell Wall Biosynthesis





Subsystem	System Characteristics: Reactions	Intr. Fluxes	Exch. Fluxes	Extreme Pathways: Type I	Type II
T	144	207	152	150	16
A	87	111	59	44	1
C	34	53	35	67	3
N	84	124	45	702	13
V	58	63	39	15	1
L	56	74	34	16	2

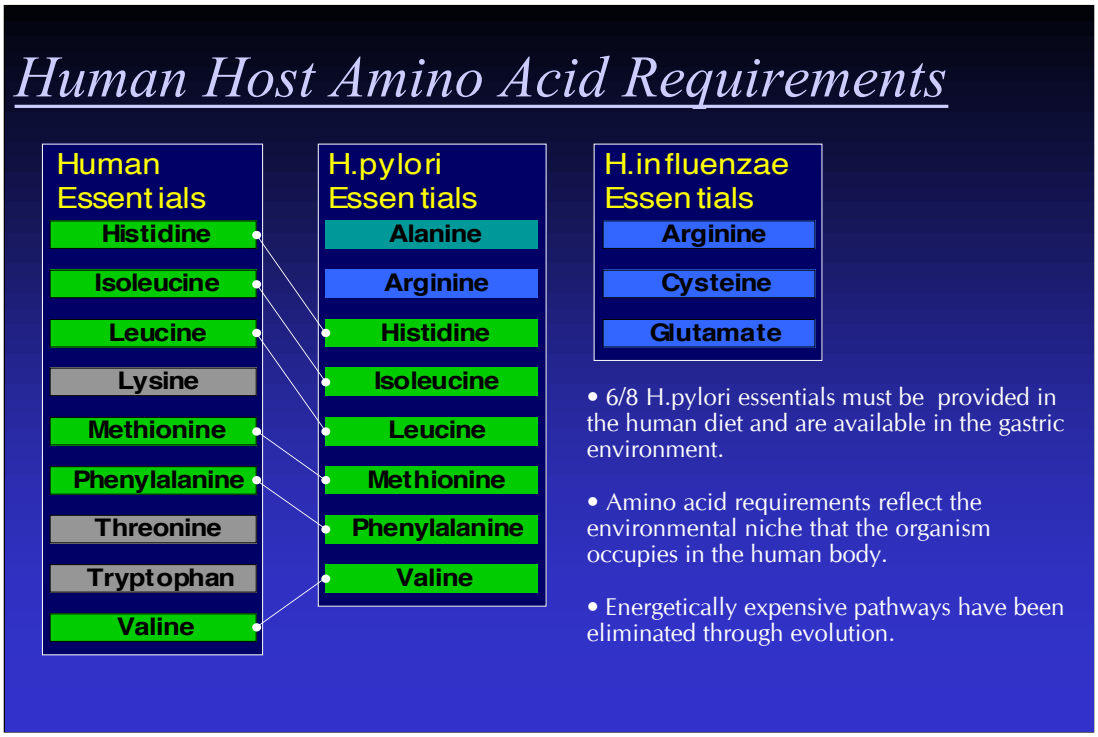
Using the computer algorithms to determine these generating vectors or extreme pathways for the system we can find all of the pathways in each subsystem and classify them. The table above indicates the number of pathways calculated in each subsystem along with the systems characteristics of each of the subsystems. The same exact analysis was performed on Helicobacter pylori which is comprised of a metabolic network that is roughly the same size as the H. influenzae model.

Uses of Pathway Analysis

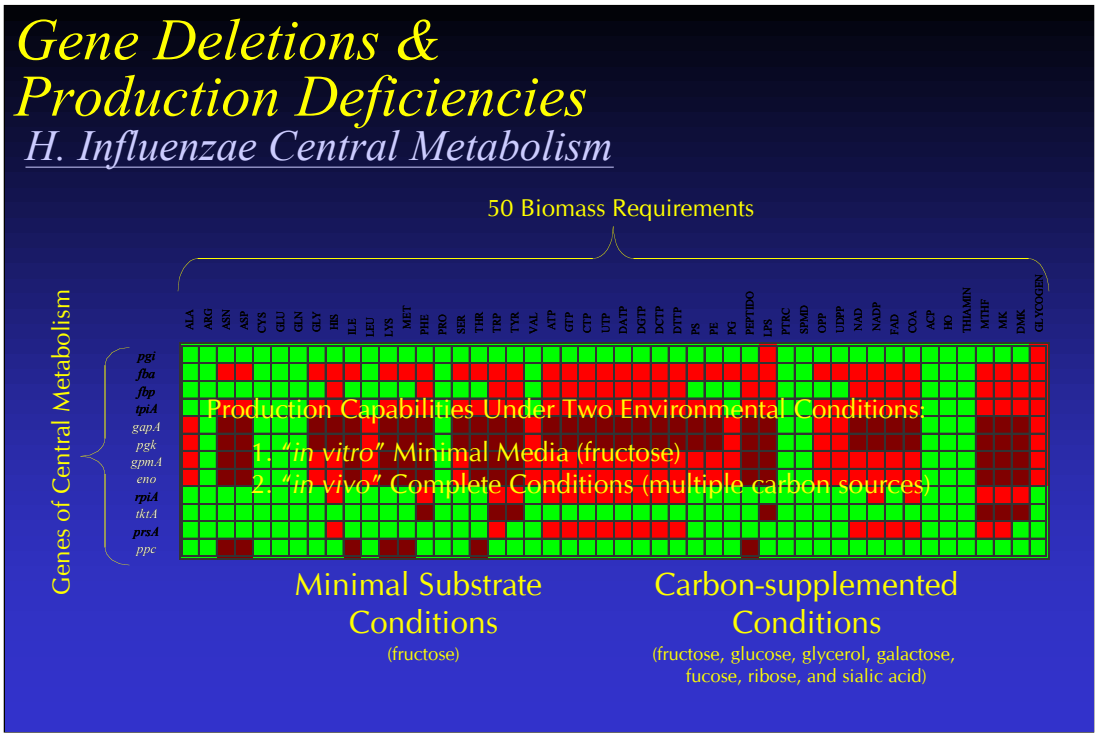
- Complete network divided into six subsystems and extreme pathways calculated in each system
- Applied to H. pylori & H. influenzae Metabolism

- Reaction subsets imply limited regulation
- Minimal Substrate Requirements
- Essential Amino Acid Requirements
- Gene deletions and loss of capability
- 7 Global Entry Points into central metabolism

From the detailed analysis of these sets of pathways a number of interesting results can be generated. Reactions that do not appear in any of the pathways can be used to reconcile possible gaps in the genome annotation. Enzyme subsets can be identified which indicate groups of reactions that always occur in the same pathways in the same flux ratios indicating potential regulons. Also minimal substrate requirements and alternative substrate can be identified by assessing which pathways can be combined under different conditions to produced the required demands on the system.



For Helicobacter pylori the model reveals that there are 8 essential amino acid which the organism must acquire from the environment. Of these 8, 6 of them are essential amino acids which are required by the human host, meanwhile in H. influenzae there is no overlap between required amino acids and human essentials. This leads to the conclusion that H. pylori has removed expensive biosynthetic pathways for amino acid production in favor of acquiring the amino acid from the gastric environment where amino acids should be plentiful as this is the site of proteolysis in the human digestive system.



Using the pathways we can begin to identify potential antimicrobial drug targets by examining the consequences of removing genes under different environmental conditions. In H. influenzae we found 12 genes in central metabolism which were essential to the networks ability to produce the biomass requirements. The red boxes indicate those requirements which could not be produced when the gene on the left was deleted and the minimal substrate conditions were presented. The dark red boxes indicate products which could not be produced even with the addition of all the possible carbon sources made available to the network.

Computation of genome-scale pathways for H. pylori

- The metabolic network for H. pylori consists of 583 reactions and 381 metabolites
- Currently computing the pathways on the San Diego supercomputer cluster (alpha machines)
  - Algorithm can not be parallelized easily and requires a fast processor with large memory
  - Currently calculation of the full pathway structure is infeasible (time and memory requirements are too large)
  - Can restrict the outputs of the metabolic network to smaller subsets for useful studies

Computation of pathways, continued

- Instead of looking at pathway structure of entire network with all the outputs, restrict the network to subsets (amino acid production, nucleotide production)
- Can examine a number of different issues for this limited set of outputs
  - Correlation of pathways
  - Biochemical yields
  - How these pathways correspond to physiological function
  - Degree of robustness and duplication of the network

Immer immer…..

in dem computer zimmer

Jim Rawlings, 1982

Summary

- Basis vectors of the null space are pathways
- Convex analysis by using positive fluxes only
- Extreme pathways as edges of cones--there are three basic types
- These pathways give much physiological insight
- Linked outputs lead to flux distributions
- Linked pathways cannot yet be calculated on a genome scale

References

- Clarke, B.L., "Stability of complex reaction networks," *Advances in Chemical Physics* **43**: 1-215 (1980).
- Gilbert Strang, Linear Algebra and Its Applications, Academic Press, New York, 1981.
- Seressiotis, A., and Bailey, J.E., "MPS: An artificially intelligent software system for the analysis and synthesis of metabolic pathways," *Biotech. And Bioeng.* **31**: 587-602 (1988).
- Mavrovouniotis, M.L. Stephanopolous, G. and Stephanopoulous, G.,"Computer-aided synthesis of biochemical pathways," *Biotech. And Bioeng.*, **36**: 1119-1132 (1990).
- Schuster, S., and Hilgetag, C., "On elementary flux modes in biochemical reaction systems at steady state," *J. Biological Systems* **2**: 165-182 (1994).
- Reinhart Heinrich and Stefan Schuster, The Regulation of Cellular Systems, Chapman and Hall, New York, 1996.
- Liao, J.C., Hou, S-Y., and Chao, Y-P., "Pathway analysis, engineering, and physiological considerations for redirecting central metabolism," *Biotech. and Bioeng.*, **52**: 129-140 (1996).
- David Lay, Linear Algebra and its Applications, Addison-Wesley, Menlo Park, 1997.
- C.H. Schilling and B.O. Palsson, "The underlying pathway structure of biochemical reaction networks," *Proc. Natl. Acad. Sci (USA),* **95**: 4193-4198, (1998)
- C.H. Schilling, S. Schuster, B.O. Palsson, and R. Heinrich, "Metabolic Pathway Analysis: Basic Concepts and Scientific Applications in the Post-Genomic Era," *Biotechnology Progress*, **15**: 296-303 (1999).

References

- C.H. Schilling, D. Letscher, and B.O. Palsson, "Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective," *J. Theoret. Biol.*, **203:** 229-248 (2000)**.**
- C.H. Schilling and B.O. Palsson, "Assessment of the Metabolic Capabilities of *Haemophilus influenzae* Rd through a Genome-Scale Pathway Analysis," *J. Theoret. Biol.*, **203:** 249-283 (2000).
- Schwikowski, B., Uetz, P., and Fields, S., "A network of protein-protein interactions in yeast," *Nature Biotechnology*, **402**: 1257-61 (2000).
- Eisenberg, D.,Macotte, E.M., Xenarios, I., and Yeates, T.O., "Proteomics in the post-genomic era," *Nature,* **405**: 823-826 (2000).

Closing the Flux Cone: imposition of maximal capacities

Bernhard Palsson

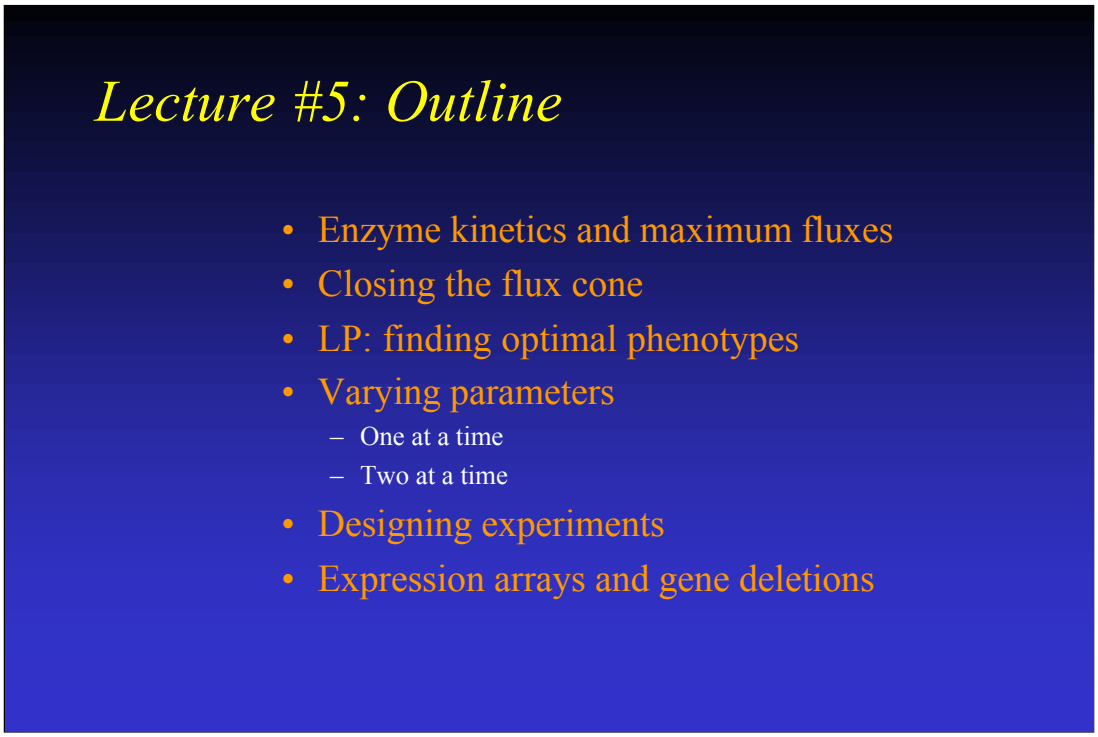
Hougen Lecture #5

Nov 21th, 2000

INTRODUCTION

In the previous lecture we looked at the combined stoichiometric and thermodynamic constraints that cells must obey. These led to the formation of a conically shaped solution space--called the flux cone. The edges are vectors that in a positive linear combination span the cone. These edges where shown to be extreme pathways. The flux through these pathways is limited by a maximum value.

Such maximal constraints close the solution space. In this lecture we explore the characteristics of the closed space.



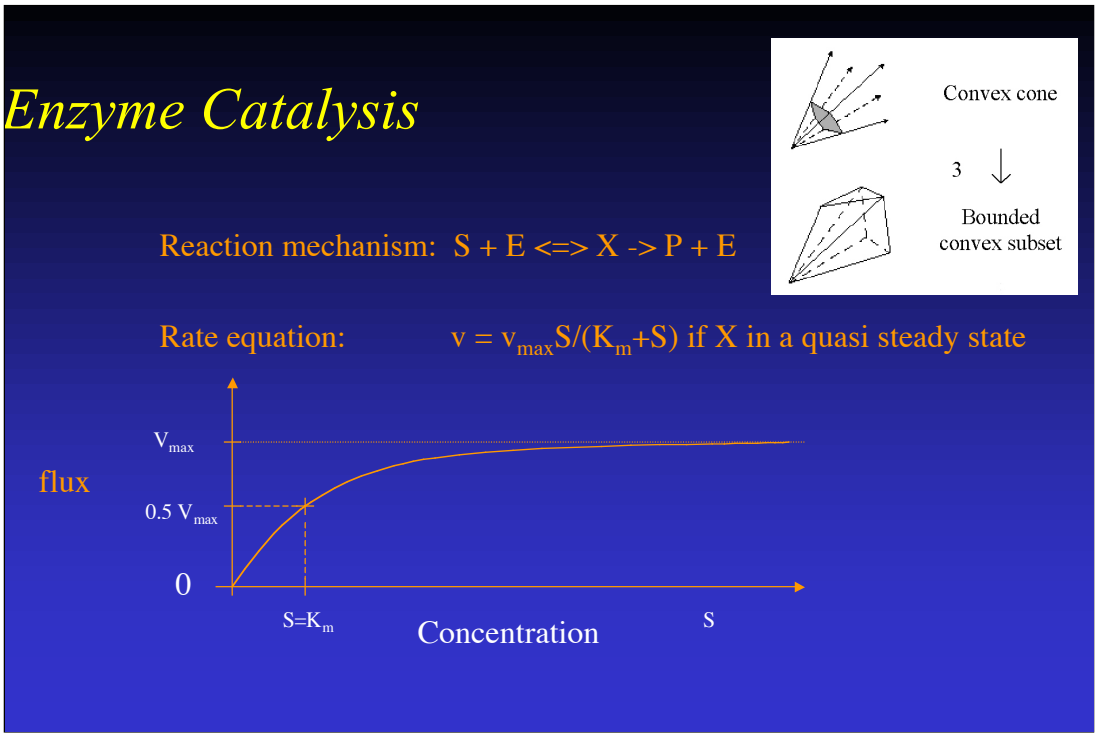
LECTURE #5

This lecture begins with an introduction to the origin of the maximal fluxes that are achievable through an enzymatic reaction and how these limitations cap off and close the flux cone. Although there are infinitely many possible flux distributions found within this closed solutions space, if an objective is stated one can find the 'best' solution by that criteria within this solution space. Linear programming or optimization is used to find this solution. The optimal solution will always lie at the edge of the cone or on one of its surfaces.

A single solution is rarely of interest. We thus explore the optimal solution as a function of an environmentally varying parameter. There are 'kinks' found in the piece-wise linear solutions. At these discontinuities we discover that the shadow price structure of the basal solution changes. These changes will thus correspond to a change in the phenotype. Thus there are a limited number (a discrete number) of phenotypes found within the solution space.

We then explore the simultaneous variation of two environmental variables and introduce the concept of a phase plane. These phase planes can then be used to design insightful experiments.

Finally, we show how flux-balance analysis can be used to interpret and predict the consequences of gene deletions and metabolic shifts as measured by expression arrays.



CONSTRAINTS ON METABOLIC FLUXES

Linear spaces are characterized by a basis set where any linear combination of the basis vectors is found in the space, i.e.;

$$\mathbf{v} = \Sigma_i w_i \mathbf{p_i}$$

Where $\mathbf{p_i}$ are the conical basis vectors, as introduced in the last lecture. The weights, w_i, used to multiply the basis vectors in the summation are positive.

Since the individual reaction steps (v_i) in a pathway vector are carried out by an enzyme there are limitations placed on the numerical values that w_i can take in a real system:

•Minimum: the reactions are irreversible, thus the weights are positive

•Maximum: there is maximum flux through an enzymatic reaction, thus there are maximum weights; thus

$$0 < w_i < w_{max}$$

Since a pathway vector is comprised of a series of individual reactions, the step with the lowest capacity will limit the flux through a linear pathway.

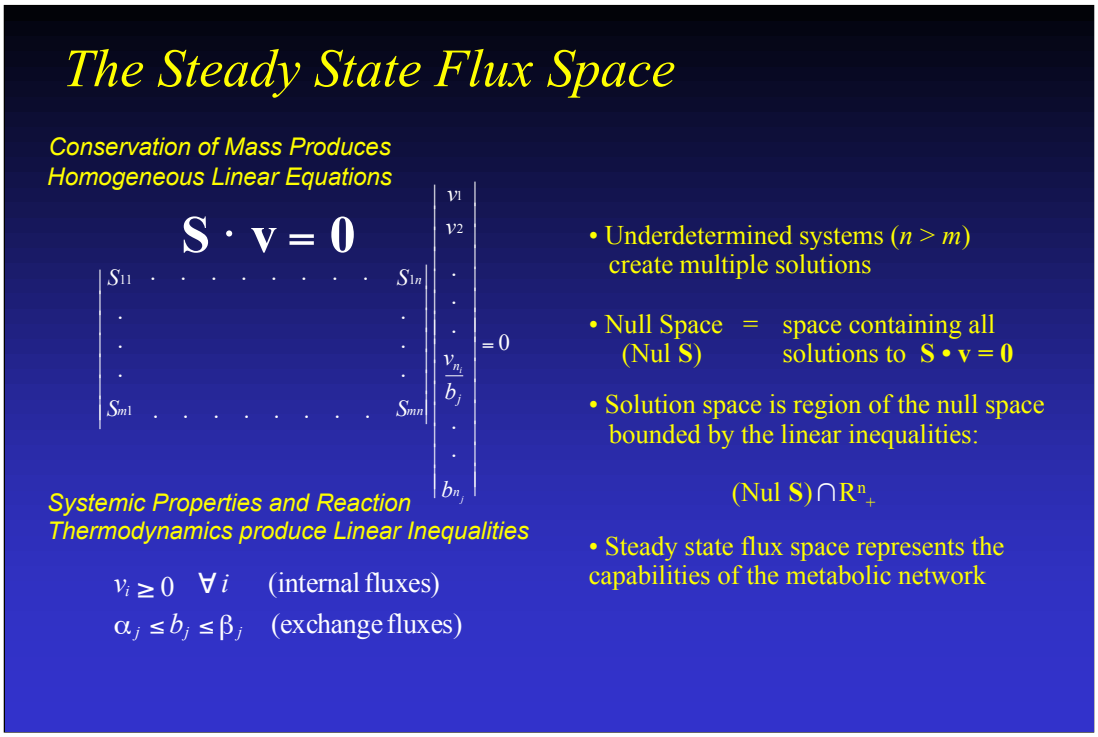
If a reaction is reversible we will write each direction as a separate irreversible reaction.

Estimation of maximal fluxes

- Using typical numerical values for:
  - concentrations for enzymes (4μM) and
  - metabolites (100μM), and
  - theoretical maximal bimolecular association rate constants and
  - data on enzyme turnover numbers, we estimate that:

  V_{max} to be one million molecules per cubic micron per second

- The maximal measured fluxes are about half that value



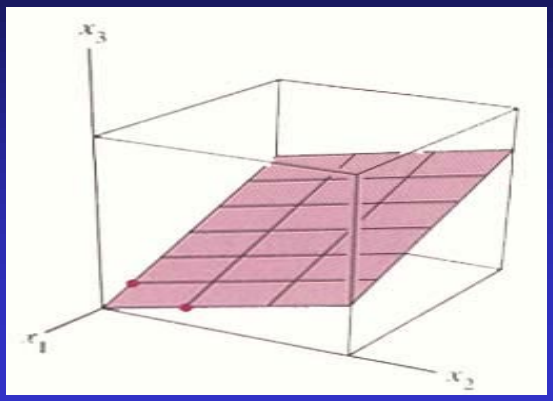
THE CONSTRAINED FLUX SOLUTION SPACE

With the stoichiometric matrix constructed, how do we determine metabolic pathways and analyze them?

As we have seen in previous lectures, the principles of conservation of mass produce a system of homogeneous linear equations, **Sv = 0**. Additionally there are constraints placed on the direction of flow under which each flux can operate creating a set of linear inequalities, $0 < w_i < w_{max}$.

This defines our conditions which in most cases creates an underdetermined system. This means that there are more fluxes operating within the system than there are metabolites which leads to multiple solutions or flux distributions which satisfy all of the stoichiometric constraints, and all the capacity constraints.

A geometric representation of the null space and constraints imposed through inequalities: it is the intersection of the null space and the positive orthant in the n-dimensional space: (Nul **S**) $\cap \mathbf{R}^n_+ \cap \mathbf{V}_{max}$

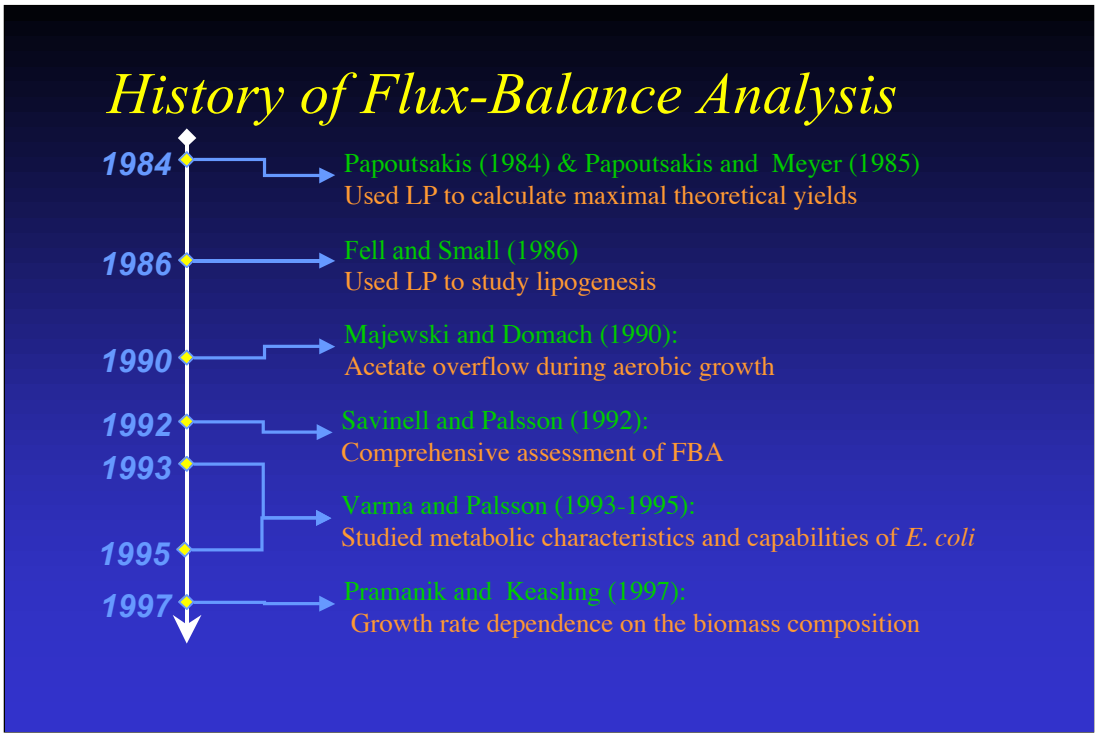


THE CONFINED SOLUTION SPACE AS AN INTERSECTION

Nul S $\cap$ $\mathbf{R}^n_+ \cap \mathbf{V}_{max}$

In linear algebra the term null space is used to describe the space which contains all of the solutions to a system of homogeneous linear equations. The solution space of interest to us is actually the intersection of this null space with the region bounded by the inequalities placed on the weights. This space represents and defines the boundaries and capabilities of a metabolic genotype describing all of the possible flux distributions and routes which can theoretically operate through the system, clearly defining what an organism can and cannot do.

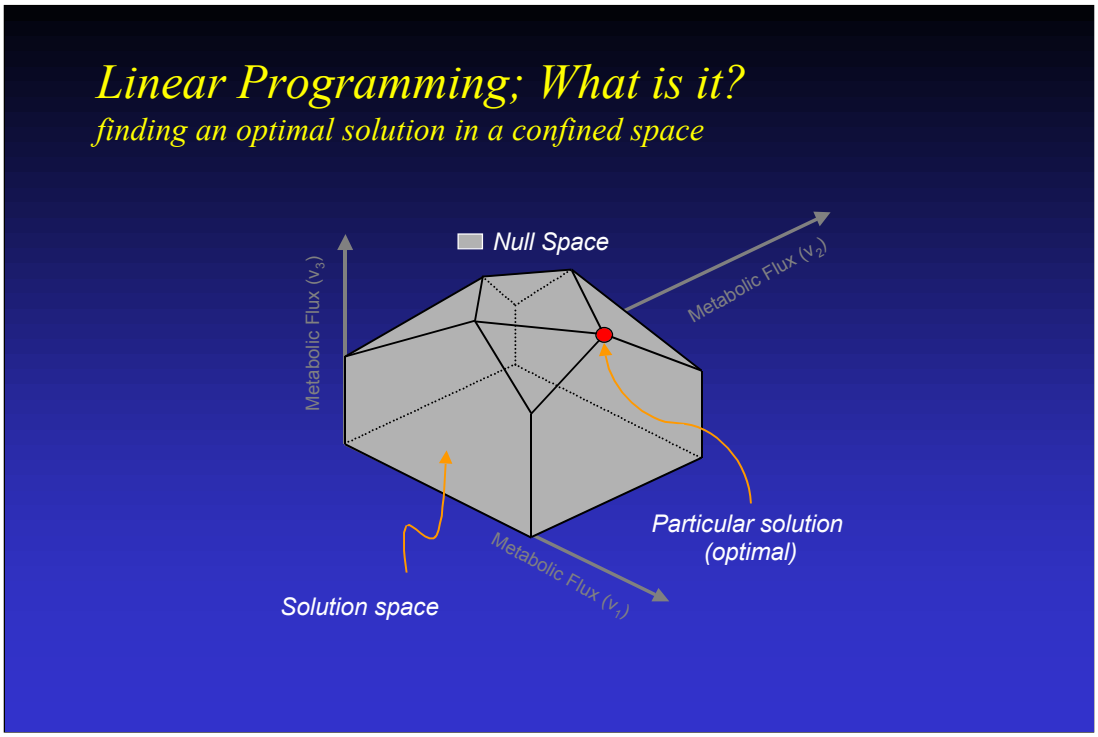
In the solution space we can find the answers to any and all of our questions which pertain to the structure and production capabilities of an organism.



SOME HISTORICAL EVENTS IN THE DEVELOPMENT OF FBA

This slides shows some of the historical events in the development of FBA of under-determined systems. A detailed historical review is found in:

Edwards, et al *Metabolic flux balance analysis* in Metabolic Engineering, Lee and Papoutsakis Editors



LP: What is it?

This diagram depicts a bounded polytope in 3 dimensions. Imagine that it is the space of possible solutions to a set of linear equalities with constraints, such as the flux balance equations and the capacity constraints. Each point in this space satisfies these conditions. However, the nature of the solutions differs. We can choose a particular solution in this space that is the 'best' in some sense.

This idea underlies LP. We state an objective function that measures what we are interested in. Then we try to find the best value for this objective function under the given constraints. The best value normally means the maximum value. Minimization can be performed by simply finding the maximum of the negative of the objective function.

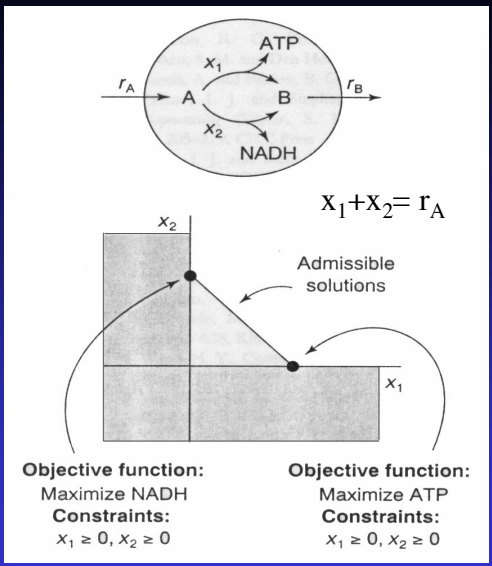
The optimal solution normally lies in a corner of the polytope. Occasionally the objective function has the same value along a whole edge and all the points on that edge are optimal values. In this rare case the objective function is 'parallel' to the edge of the polytope.

How does LP work? A very simple example

The solution space is the line of admissible in the positive orthant.

If we maximize ATP production the solution lies on the x-axis where all the flux would be through reaction x_1. Conversely, maximizing NADH production would give the point at the y-axis, where only reaction x_2 is active.

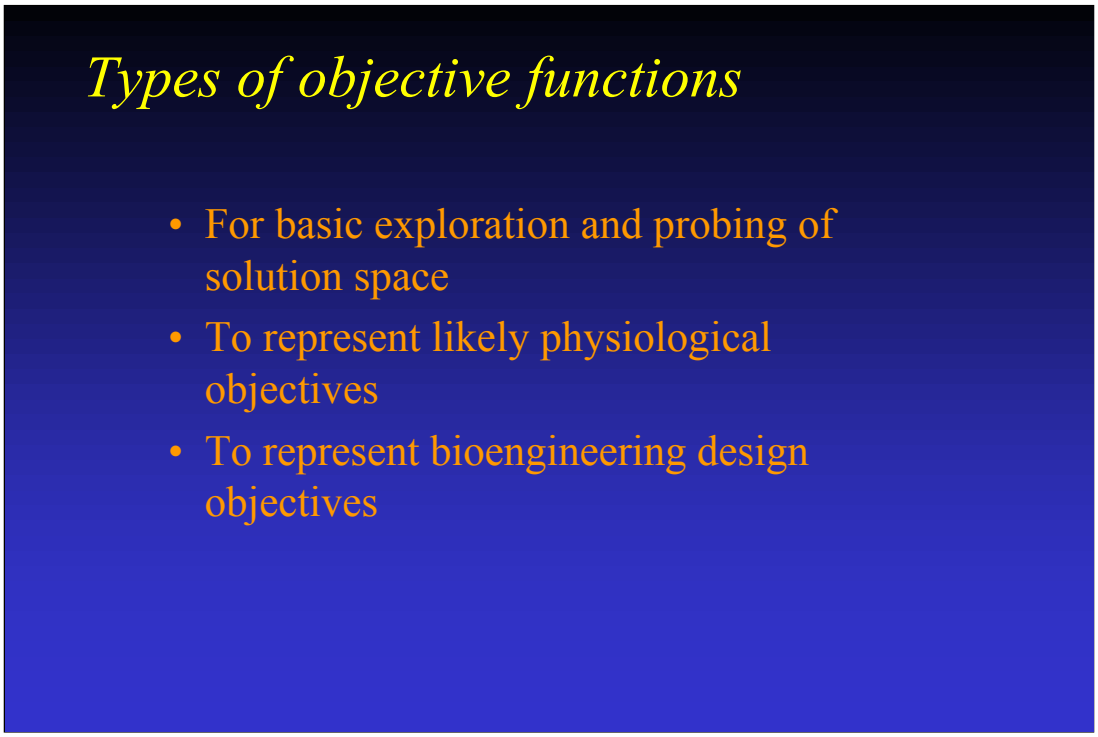
Note that the optimal solutions lie at the boundary of the admissible space.



Bonarius,et al TIBTECH vol 15:308 (1997)

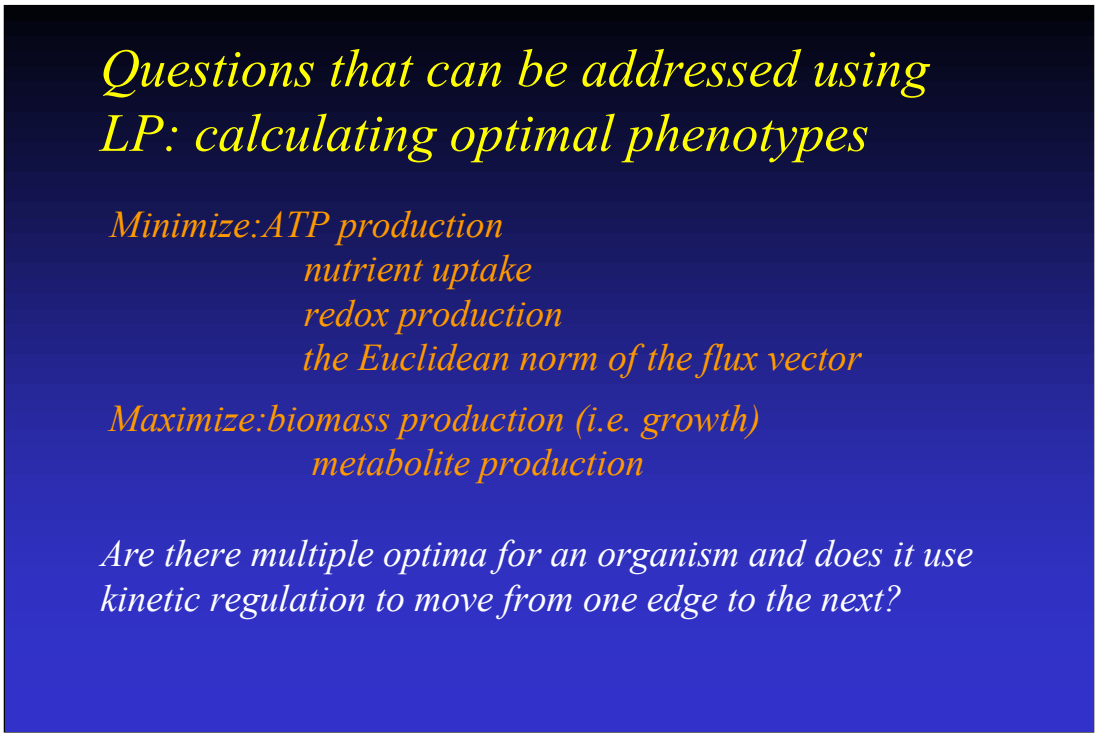
This readily understandable example shows a space of admissible solutions and the optimal phenotypes lying at the edges of this space.

Q: What happens if you optimize x_1+x_2 ?



The Objective Function

Within the solution space defined by the connectivity and capacity constraints, we can search for the best solution using linear optimization. What we search for is determined by the objective function stated. There are several types of objective functions that can be used. First, we can use objective functions to explore the properties of the solution space, and the capabilities of an organism. These objective functions include things like maximizing the ATP from a given substrate, or maximizing the amount of an amino acid produced from a given substrate. These types of objective functions are non-physiological, but can be used to probe the properties of a network. A second class of objective functions would represent objectives that we believe are physiologically relevant. For microbial cells, the belief is that they maximize their growth rate given the constraints under which they operate. In this case, and as shown, below the objective is the balanced exit from the network of all the precursors needed for the synthesis of the cellular mass. The third type of objective function may relate to an intentional engineering objective of a metabolic system. We may wish to maximize a product like Lysine, for instance, and try to figure out what the best flux maps are that lead to the production of Lysine. We can add or delete reactions from the network to determine how those changes affect the yield of the desired product.



OPTIMAL PHENOTYPES

A number of different objective functions have been used for metabolic analysis, these include

Minimize ATP production: This objective is stated to determine conditions of optimal metabolic energy efficiency.

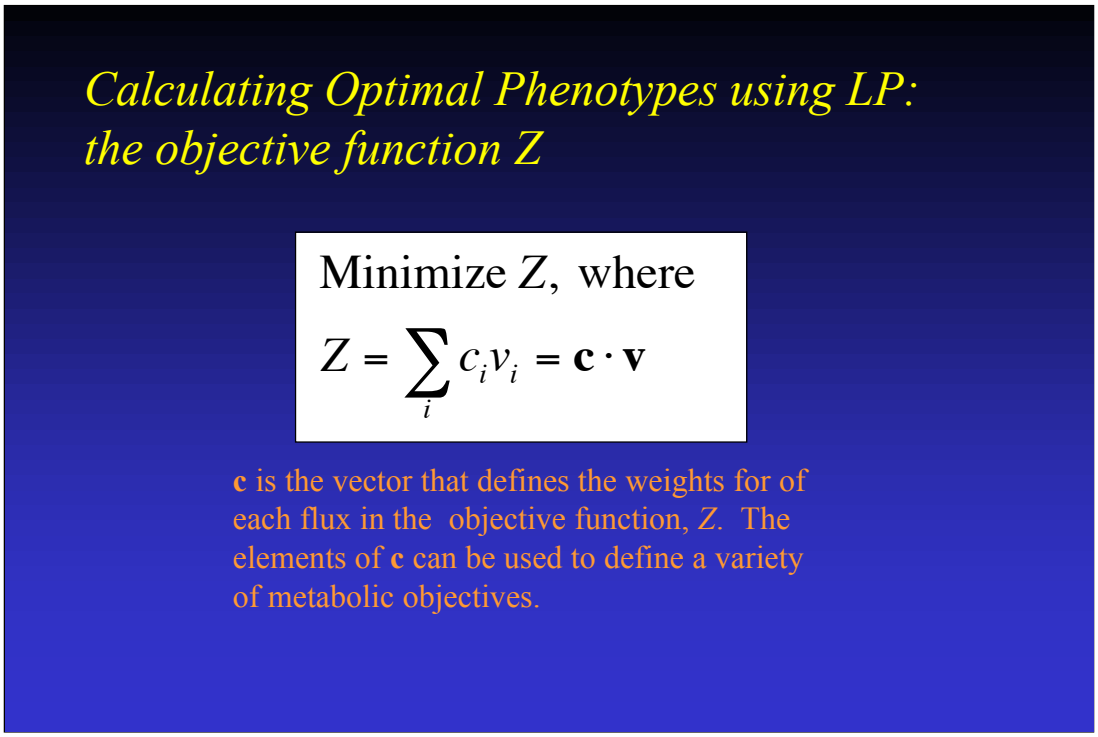
Minimize nutrient uptake: This objective function is used to determine the conditions under which the cell will perform its metabolic functions while consuming the minimum amount of available nutrients.

Minimize redox production: This objective function finds conditions where the cells operate to generate the minimum amount of redox potential.

Minimize the Euclidean norm: This objective has been applied to satisfy the strategy of a cell to minimize the sum of the flux values, or to channel the metabolites as efficiently as possible through the metabolic pathways.

Maximize metabolite production: This objective function has been used to determine the biochemical production capabilities of *Escherichia coli*. In this analysis the objective function was defined to maximize the production of a chosen metabolite (i.e. lysine or phenylalanine).

Maximize biomass and metabolite production: By weighing these two conflicting objectives appropriately, one can explore the tradeoff between cell growth and forced metabolite production in a producing strain.

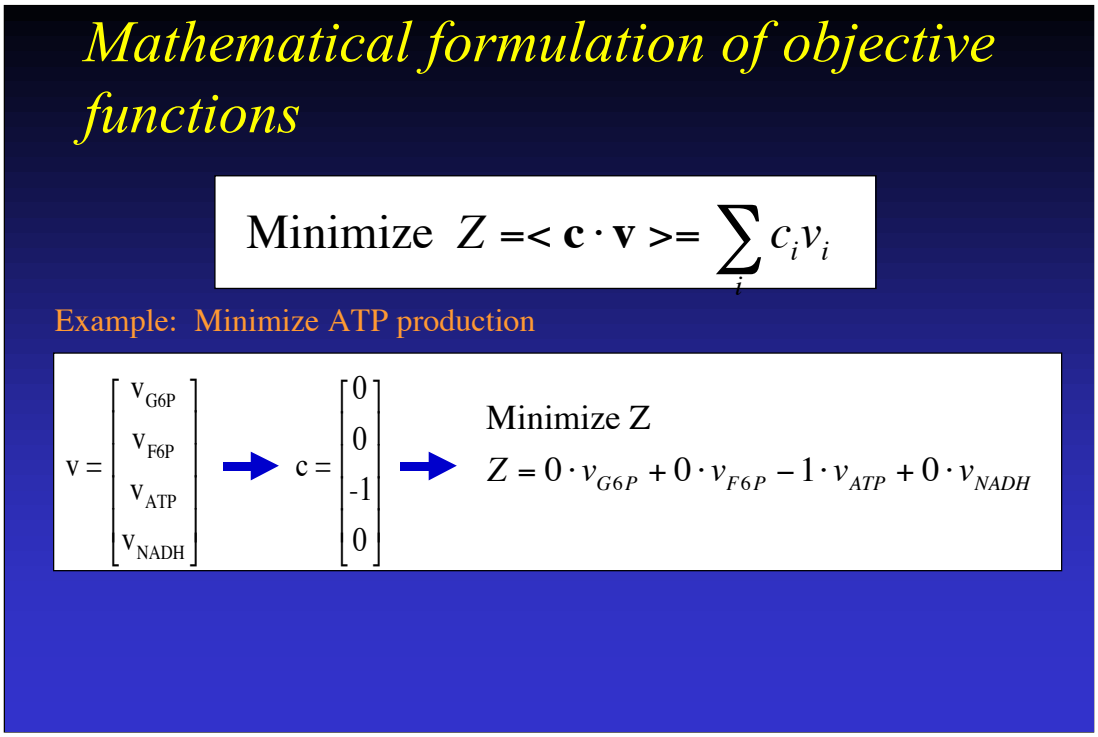


THE OBJECTIVE FUNCTION

Numerous questions about metabolic capabilities can be answered using LP. The stoichiometric and capacity constraints define a range of allowable behavior. We can then find the best value within these constraints. Biologically, we have defined the space of all phenotypes (that is particular solutions) that can be derived from a genotype. We can calculate the best phenotype from a particular standpoint. For instance we can calculate the maximum number of ATP molecules that can be generated from a particular substrate.

The next slide lists a number of important phenotypic behaviors that can be calculated using LP. The maximum growth function is perhaps the one of greatest interests from an evolutionary standpoint.

This general representation of *Z* enables the formulation of a number of diverse objectives. These objectives can be design objectives for a strain, exploitation of the metabolic capabilities of a genotype, or physiologically meaningful objective functions, such as maximum cellular growth.



MATHEMATICAL FORMULATION OF OBJECTIVE FUNCTIONS

This slide illustrates the formation of the objective function using a simple example. In the example there are 4 metabolite fluxes. The objective is to minimize ATP production therefore the **c** matrix has a zero "weight" on all fluxes except v_{ATP} which has a -1 The coefficient on the ATP flux is negative since it is being minimized.

The growth requirements

Metabolic demands of precursors and cofactors required for 1 g of biomass of *E. coli.*

These precursors are removed from the metabolic network in the corresponding ratios.

Thus, the objective function is:

$$Z = 41.2570\, v_{ATP} - 3.547 v_{NADH} + 18.225 v_{NADPH} + \ldots.$$

Metabolite	Demand (mmol)
ATP	41.2570
NADH	-3.5470
NADPH	18.2250
G6P	0.2050
F6P	0.0709
R5P	0.8977
E4P	0.3610
T3P	0.1290
3PG	1.4960
PEP	0.5191
PYR	2.8328
AcCoA	3.7478
OAA	1.7867
AKG	1.0789

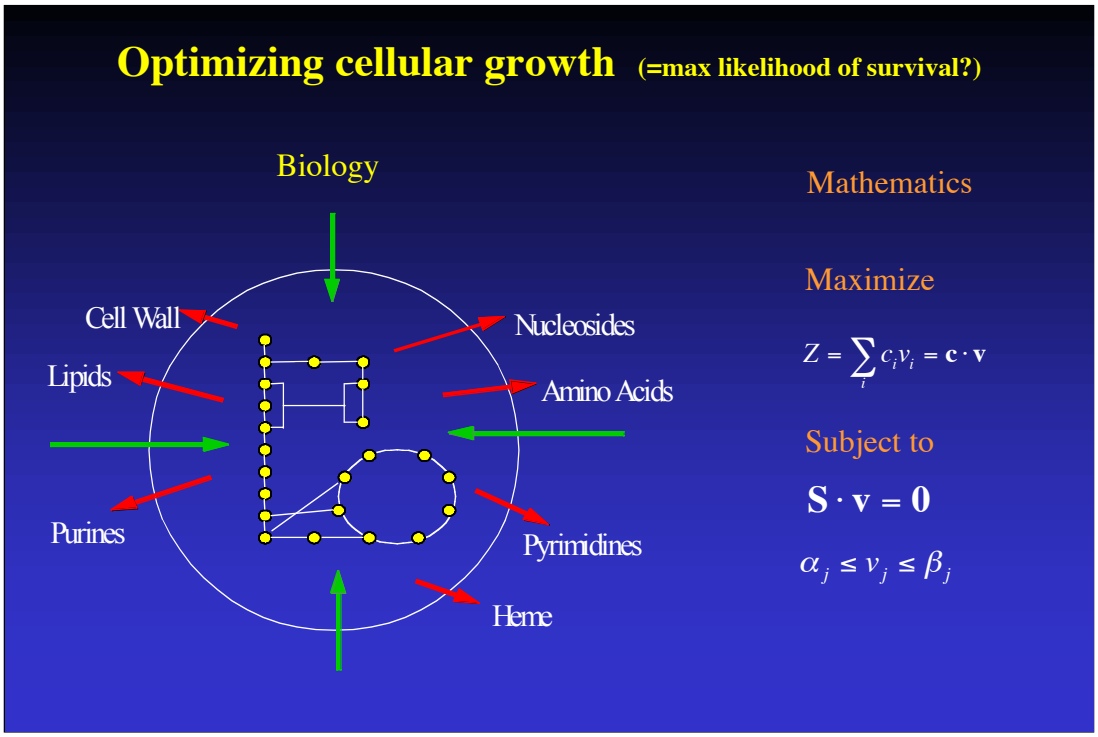
THE GROWTH FUNCTION

This table shows the requirements for making one gram of *E. coli*. This means that for the cell to grow all these components must be provided in these amounts. Thus, a balanced set of metabolic demands makes up the growth objective function:

$$Z = 41.257 v_{ATP} - 3.547 v_{NADH} + 18.225 v_{NADPH} + 0.205 v_{G6P} + 0.0709 v_{F6P} + 0.8977 v_{R5P} + 0.361 v_{E4P} + 0.129 v_{T3P} + 1.496 v_{3PG} + 0.5191 v_{PEP} + 2.8328 v_{PYR} + 3.7478 v_{AcCoA} + 1.7867 v_{OAA} + 1.0789 v_{AKG}$$

The biomass composition thus serves to define the weight vector **c.**

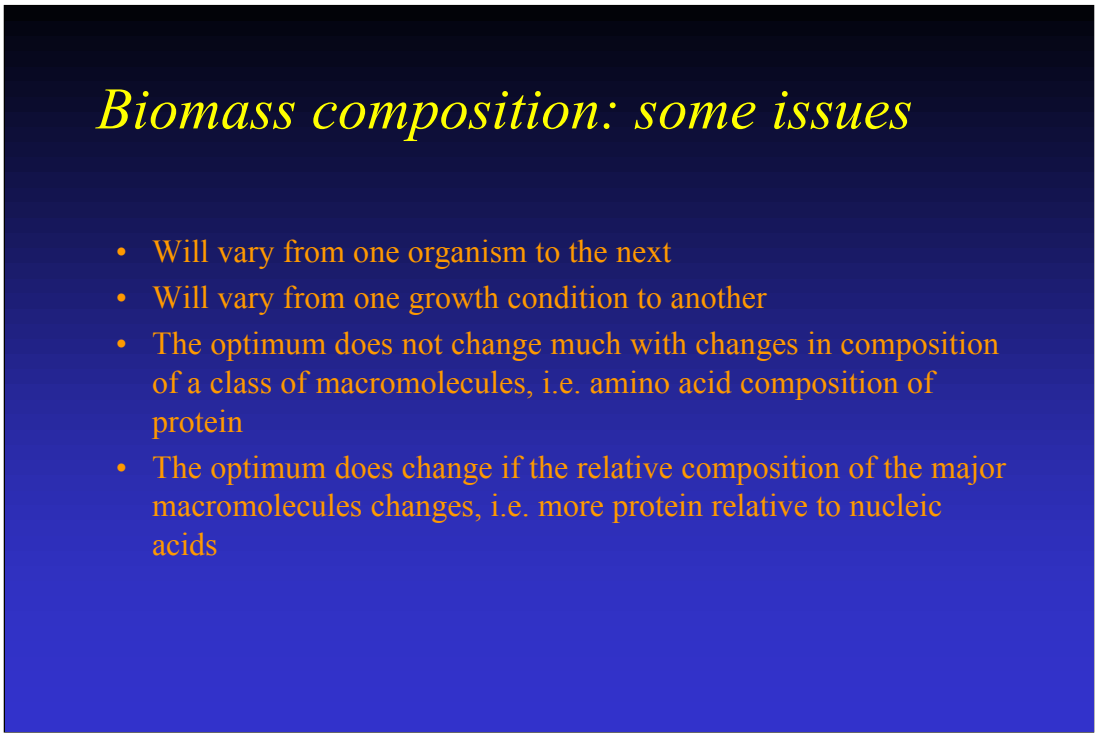
The full growth function for *E. coli* is more complicated than the one given above, since various maintenance functions need to be considered.



THE MAXIMAZATION OF BIOMASS FORMATION

This slide shows schematically on the left the idea of maximizing biomass formation. There can be one or more inputs (the green arrows) and a balanced (linked) output that corresponds to the biomass composition.

On the right we show the mathematical formulation of the problem. We wish to maximize the objective function under the stated constraints. These constraints form a closed cone as explained earlier.



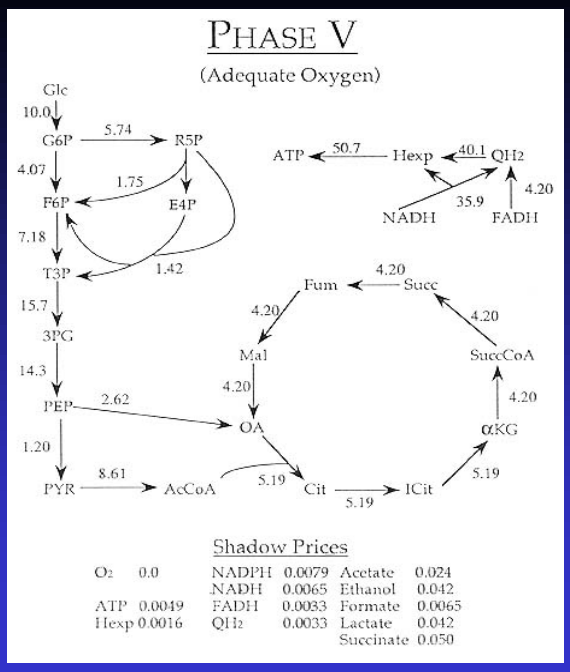
Biomass Composition

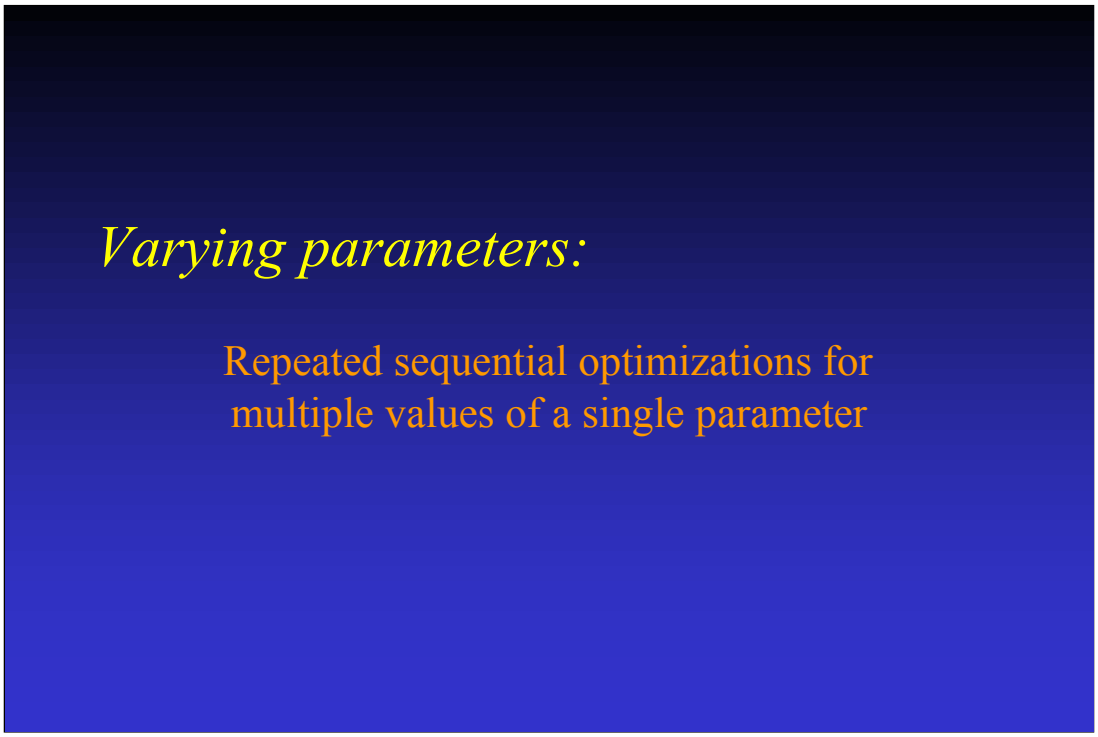
The physiologically interesting objective that we wish to study throughout these notes is the maximization of biomass yield. The definition of the solution space has few ambiguities associated with it, but the statement of the objectives has a few uncertainties built into it. First, the biomass composition is variable. It is different from one organism to another. It varies from one growth condition to another, and both of those may potentially be important issues and change the predicted optimum behavior. Legacy databases of biomass composition are needed.

The limited calculations that have been performed show that the optimum solutions do not change significantly with the monomeric composition of the major macromolecules. For instance, if the Valine to Alanine ratio is varied in the protein of a cell, the optimal growth rate does not significantly change. Conversely, if the protein relative to lipid composition in a cell changes, the optimum solution tends to be affected.

As will be shown, one can invert this problem and look at an edge of the solution space and then calculate all the objective functions that are maximized under those conditions. This might give better insight into the objectives that cells are trying to accomplish.

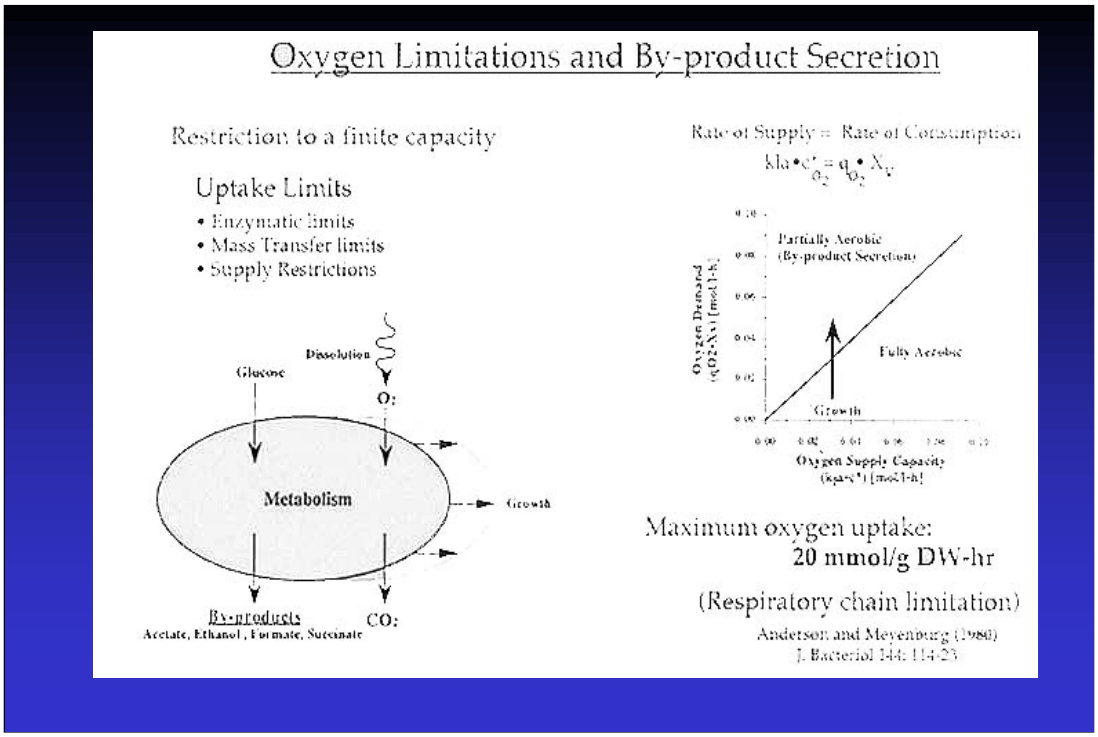
The solution displayed as a flux map: example, aerobic growth on glucose





PARAMETER VARIATION

We looked at one optimal flux map for different substrates and for constraints on several internal fluxes. These are calculations for a discrete set of conditions. We may however be interested in the a range of numerical values for a particular parameter. Thus, we can calculate a series of optimal solutions for small incremental changes in a parameter in the system. If the increments are small enough, we effectively get a continuous variation in the parameter of interest.



EXAMPLE: REDUCING OXYGEN AVAILABILITY

When cells grow in the laboratory with an abundance of substrate they grow into high densities eventually outstripping the ability for oxygen to be supplied rapidly enough to support fully aerobic growth. As oxygen becomes limiting, the cells must partially oxidize their substrate and secrete a metabolic by-product.

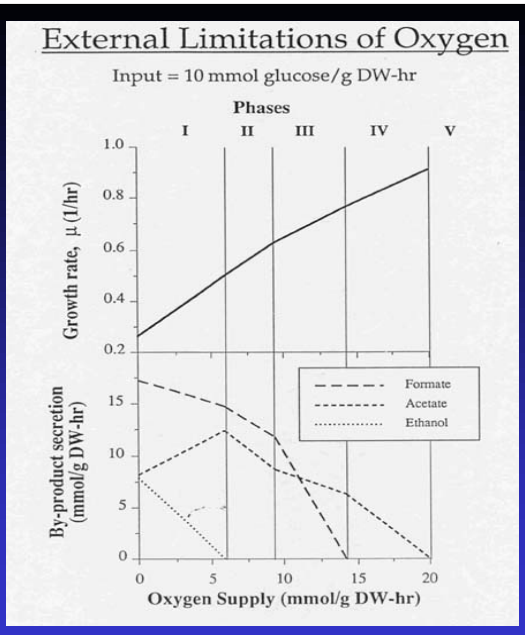
The panel on the left illustrates this problem at the cellular level. On the right this problem is illustrated from a bioprocess viewpoint.

The following slides were prepared with a reduced E. coli model in 1993 (Varma, A&EM), but it illustrates how parameter variations can be used to study problems of fundamental physiological relevance, and those that are of practical importance.

Example:

In this example we vary the maximum allowable uptake rate of oxygen. The whole range of oxygenation is shown, from fully aerobic conditions to fully anaerobic conditions.

The growth rate is graphed in the upper panel and the by-product secretion rates in the lower.



anaerobic ⟶ aerobic

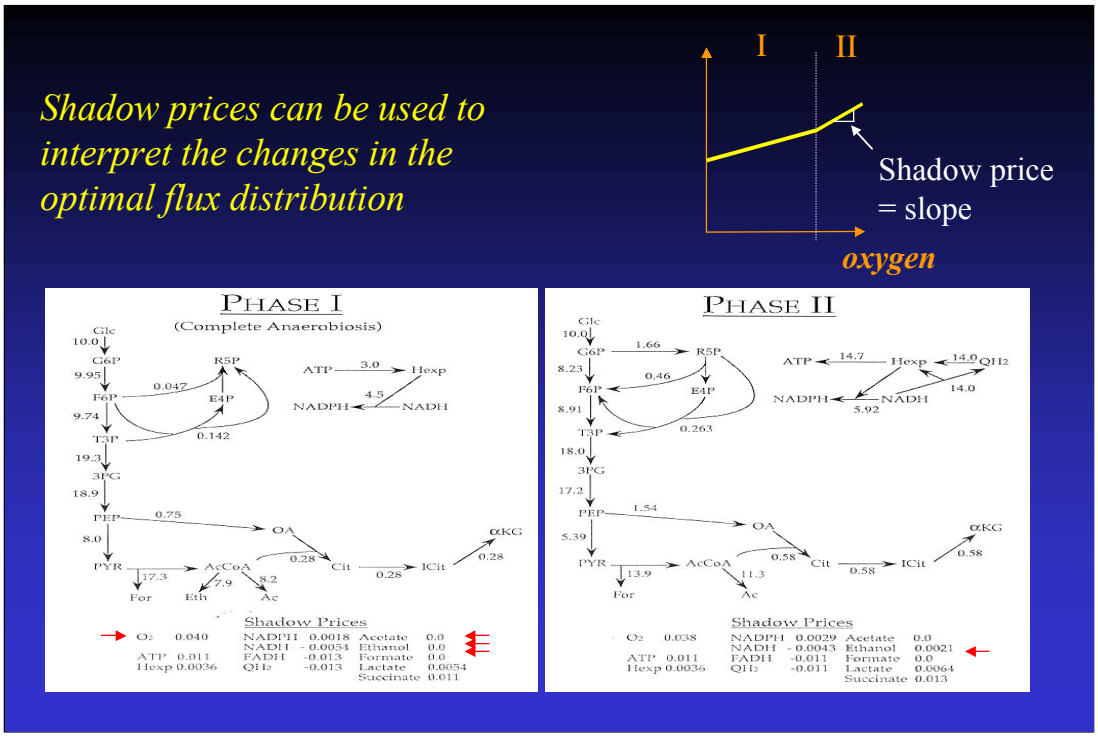
VARYING OXYGEN AVAILABILITY

As the dissolution of oxygen cannot keep up with the high volumetric consumption rates at high cell density, the amount available per cell is reduced. Computationally this is represented by lowering the capacity constraints on the oxygen uptake rate.

The results from a series of LP calculations with varying b_{O2} is shown in this slide. The optimal growth rate drops as the oxygen uptake rate is reduced, as shown in the upper panel. It does so in piece-wise linear fashion where changes in the slope occur at well defined oxygen uptake rates. This feature naturally divided the range of oxygen uptake rates into distinct phases.

The lower panels shows the secretion rates of metabolic by-products; formate, ethanol and acetate. Each one of these by-products is secreted in a fundamentally different way in each phase. As oxygen is reduced, incomplete oxidation of glucose takes place and metabolic by-products are secreted; acetate is first secreted, then formate followed by ethanol.

The LP solution in each phase is fundamentally different and the transition from one to another can be interpreted using shadow prices.

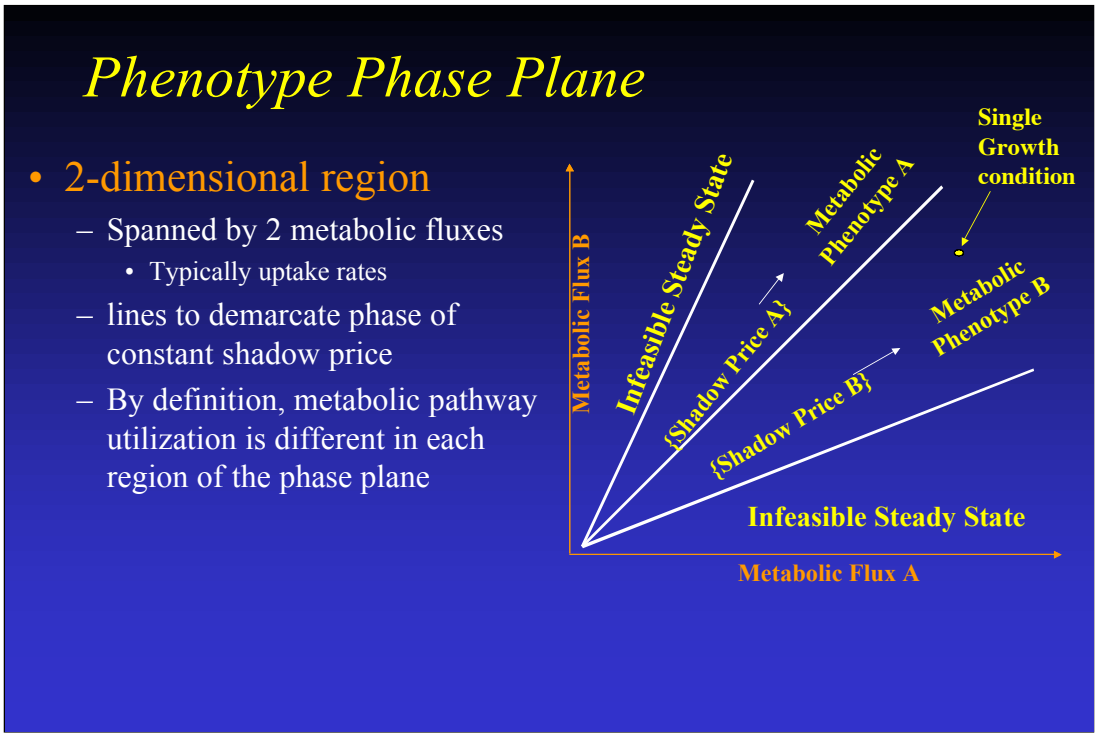


CHANGES IN SHADOW PRICES AT PHASE BOUNDARIES

The shadow price changes discontinuously at the boundary from one phase to the next. In fact the change in the shadow price defines the boundary between the phases. The shadow prices basically tell us how the governing constraints on the objective function change and how the base optimal LP solution changes. This change is reflected in a shift in the flux map.

Phase I shown above is for completely anaerobic growth. The shadow price for oxygen and ATP is positive, indicating that these are constraining factors, since the objective function would increase if more of these compounds were provided to the cell. Some of the redox carriers have negative shadow prices indicating that the cell has a problem with excess redox potential. The latter is characteristic of anaerobic metabolism.

In Phase I, acetate, ethanol, and formate, all have zero shadow prices, indicating that these intermediates are useless to the cell. Thus they are secreted. Notice that in Phase II, ethanol has a positive shadow price. It thus has value to the cell and is not secreted. In fact the defining difference between the optimal flux maps in phase I and II is the secretion of ethanol. The shadow prices are thus key in interpreting the optimal flux maps and changes in the maps as parameters vary.



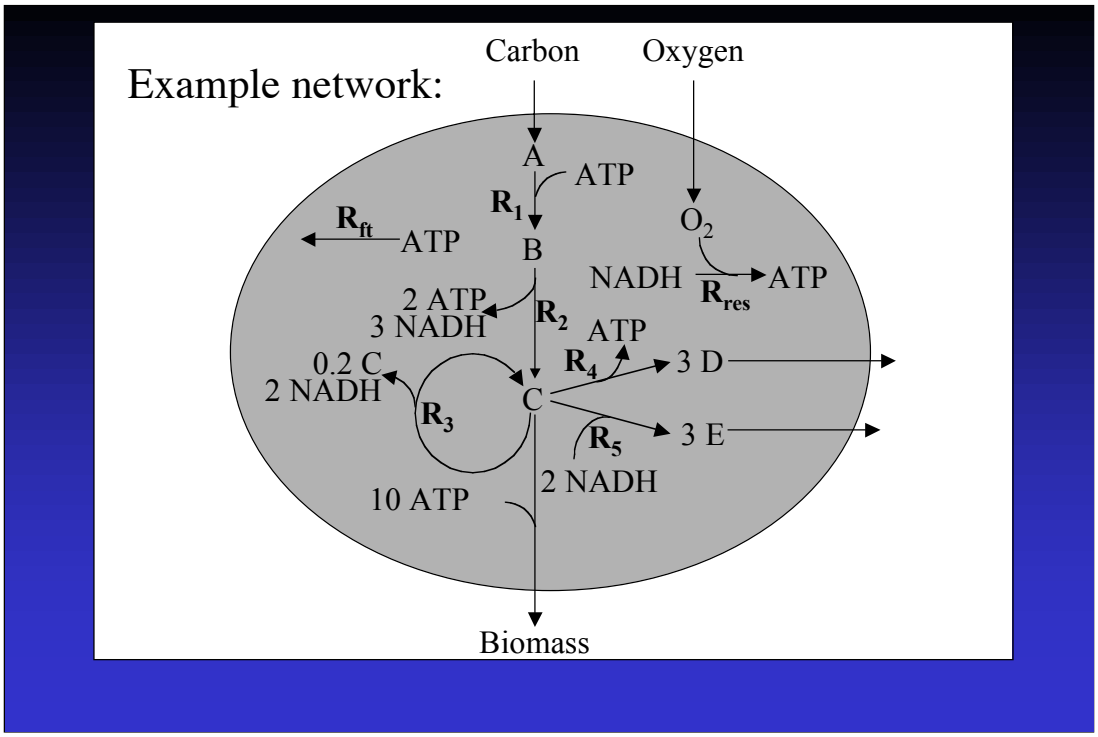
VARYING TWO PARAMETERS: THE PHENOTYPIC PHASE PLANE

A phase plane is a two dimensional region that is spanned by 2 metabolic fluxes. These fluxes are often uptake rates, but this isn't required. The shadow prices for the metabolites are calculated for all the points within this space, and lines are drawn to demarcate regions of constant shadow prices.

The shadow prices are constant within each region and are different in the other regions.

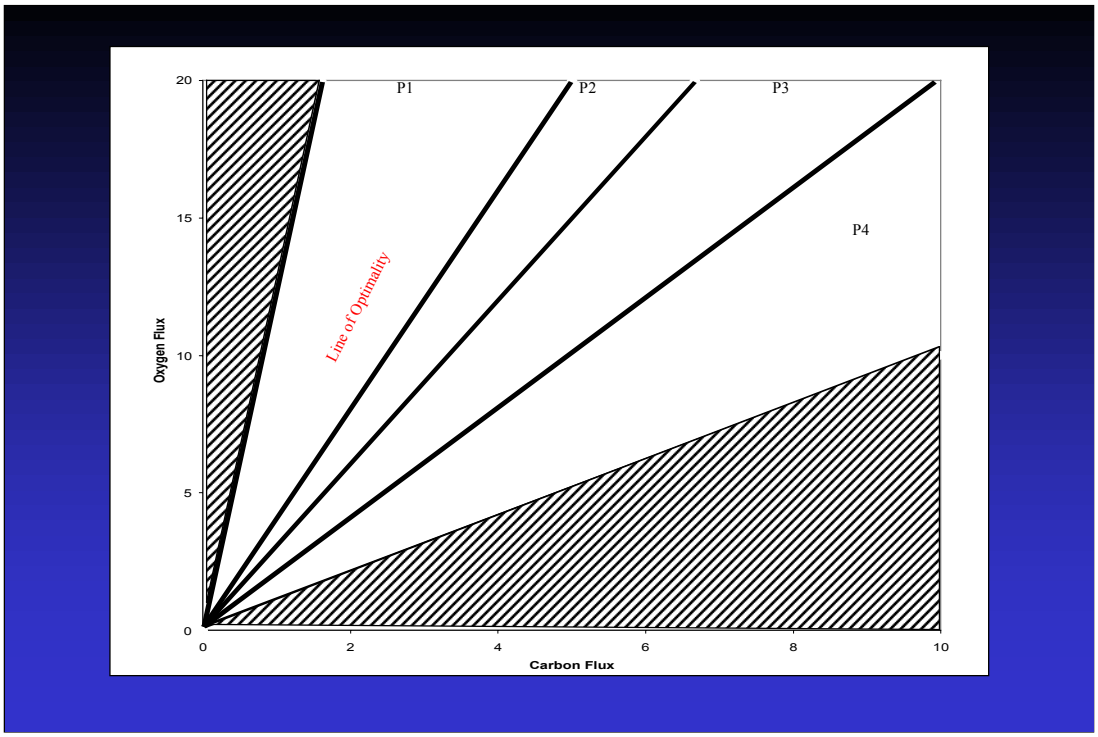
Each region refers to a different basis solution, which implies a different utilization of the metabolic pathways.

Thus, the utilization of the metabolic pathways will be qualitatively different depending on the region of operation within the phase plane.

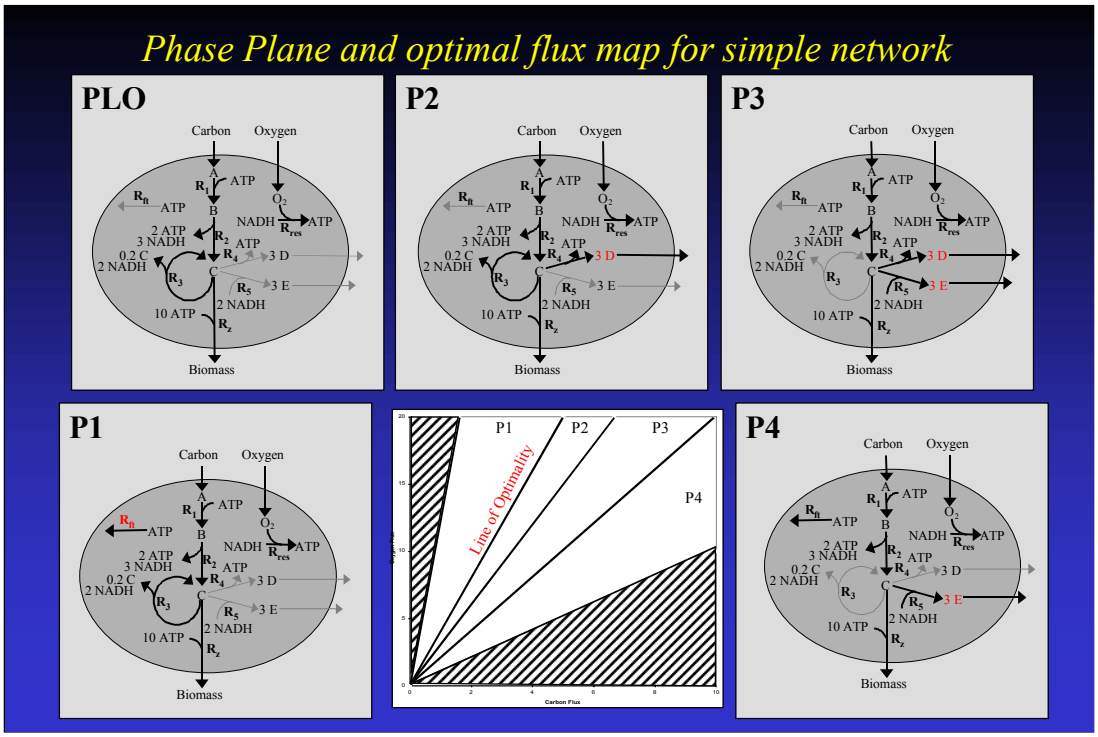


EXAMPLE

To illustrate these concepts, we now present an example of a hypothetical metabolic system. This network utilizes a single carbon source, which it metabolizes to a single biosynthetic precursor, **C**. This precursor is converted into biomass, via $\mathbf{R}_z$ (the objective function), and to two different metabolic by-products, **D** and **E**. An electron acceptor, oxygen, is also included in this example. This electron acceptor can be used to convert redox potential into high-energy phosphate bonds, $\mathbf{R}_{res}$. Additionally, there is a reaction, $\mathbf{R}_3$, which consumers 0.2 **C** to generate NADH. Finally, one reaction, $\mathbf{R}_{ft}$, represents futile cycles that hydrolyzes ATP.



The methods presented in the previous slides were used to calculate the PhPP for this hypothetical metabolic system. The PhPP and the qualitative flux maps for each phase are shown the next slide. P1 is the futile region where the electron acceptor is provided in excess. The metabolic network dissipates the excess electron acceptor taken up by the cell by increasing the flux in $\mathbf{R_3}$, which generates NADH but also oxidizes the precursor, **C**. Additionally, the futile cycle reaction $\mathbf{R_{ft}}$ is utilized to eliminate the excess ATP produced. The upper limit of P1 occurs when the entire biosynthetic precursor produced is oxidized to eliminate the excess electron acceptor, and thus no biomass can be generated.



The metabolic flux map of this system is also shown for conditions on the line of optimality (LO). The LO is a special case of P1, this is the point where the electron acceptor is no longer in excess and the futile cycle flux is zero (Table 1). The LO represents the optimal utilization of this example metabolic network to produce biomass. The qualitative flux map indicates that under conditions defined by the LO there is no metabolic by-product production and futile cycle flux equals zero.

The next distinct flux map for this hypothetical metabolic network is found in region P2. In P2 a reduced metabolic by-product (**D**) is secreted from the cell. The shadow price for the metabolite **D** in this system is zero in region P2, and the utilization of the metabolic pathways in this region is fundamentally different than in P1, Plo, P3, and P4. The metabolic pathway for the production and secretion ($\mathbf{R_4}$) of **D** is turned on under the conditions defined in this region, and the excess redox potential is eliminated through the secretion of **D**.

The utilization of the metabolic network in P3 is fundamentally different than in P2. In P3, the cyclic reaction $\mathbf{R_3}$ is not utilized, and thus redox potential production is reduced. Both of the reduced metabolic by-products are secreted (**D** and **E)** as sinks for redox potential. Thus, in this region, both of these metabolites will have a shadow price equal to zero.

Shadow prices for simple network

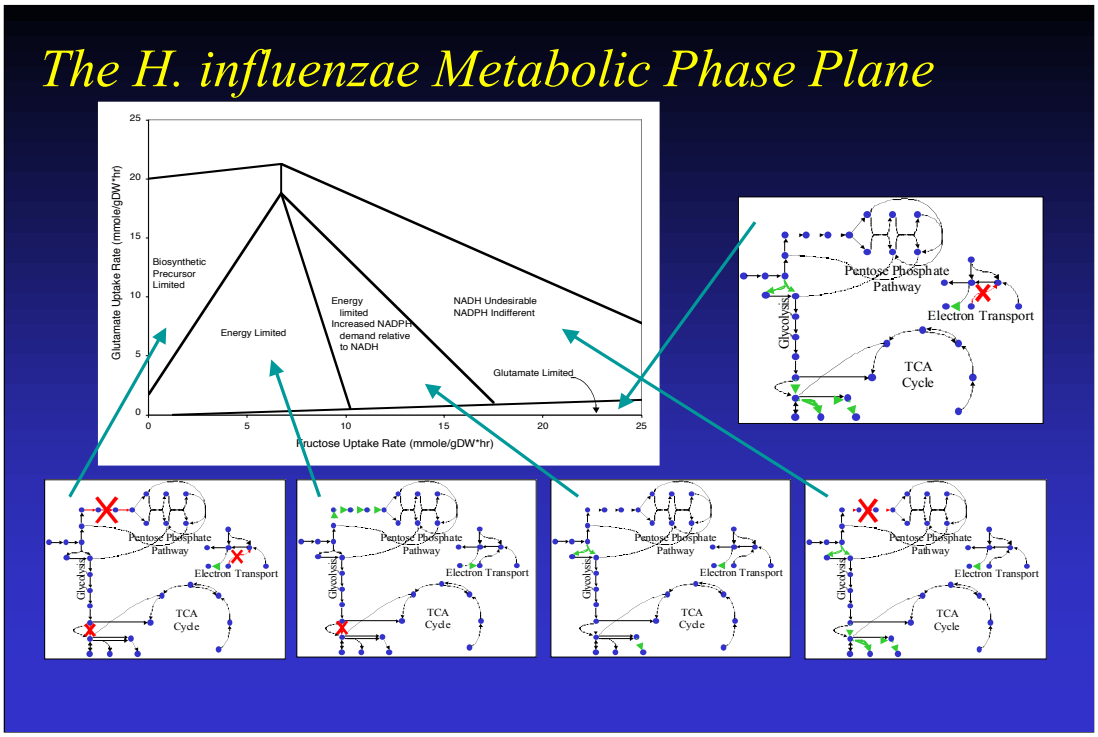
Table 1: Shadow price of the metabolites from the example shown in Figure 1.

	Carbon	A	B	C	D	E	O_2	NADH	ATP
P1	-1.30	-1.30	-1.30	-1.00	-0.33	-0.40	**0.10**	-0.10	
Plo	-0.90	-0.90	-0.93	-0.67	-0.21	-0.27		-0.07	-0.03
P2	-0.21	-0.21	-0.30	-0.09		-0.04	-0.17	-0.01	-0.09
P3	-0.05	-0.05	-0.14	-0.09			-0.23	**0.05**	-0.09
P4	**0.50**	**0.50**	**0.50**	-1.00	-0.33		-0.50	**0.50**	

Finally, in P4, the futile cycle reaction is utilized, and all the metabolic by-product formation is directed toward the formation of the more reduced by-product, **E**. When the oxygen uptake and the carbon uptake define a point on the lower boundary of P4, all the carbon source is directed toward the formation of metabolite **E**, and no biomass is generated. Thus, below this line (the crosshatched region) is another region of unobtainable steady states of the metabolic network.

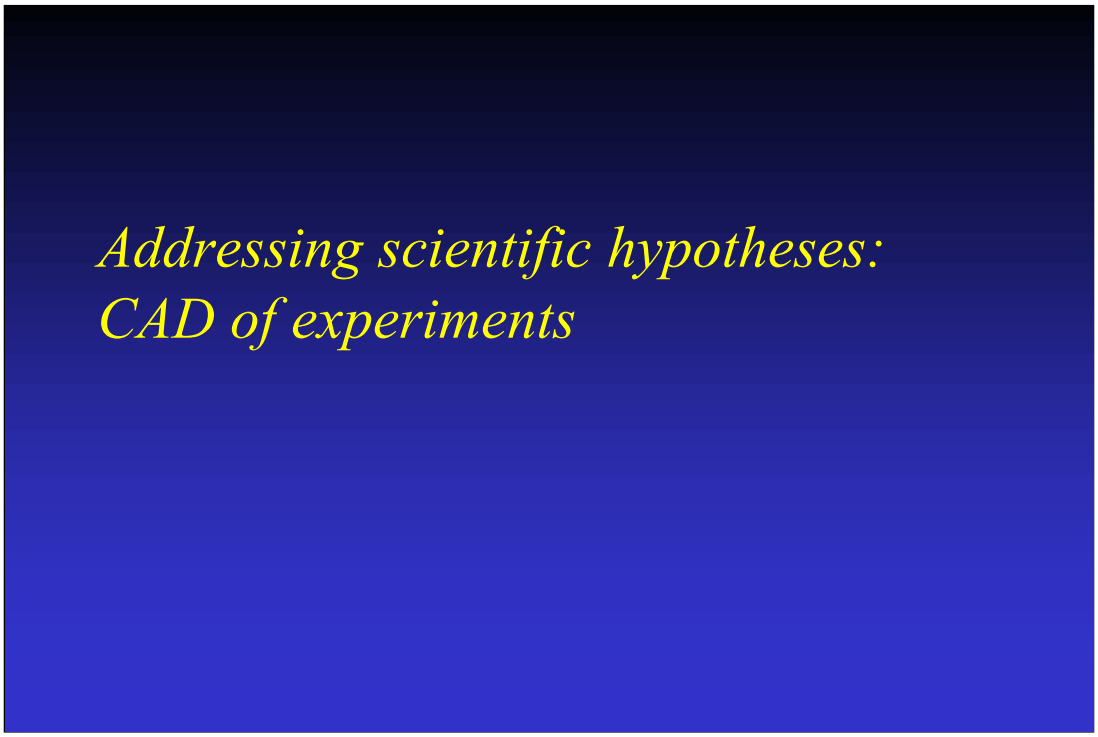
This simple example illustrates the utility of the PhPP in the interpretation of the metabolic physiology of the system. It clearly shows that the optimal phenotypes are condition dependent, and that a finite number of qualitatively different optimal phenotypes can be derived from a single genotype.

End of example

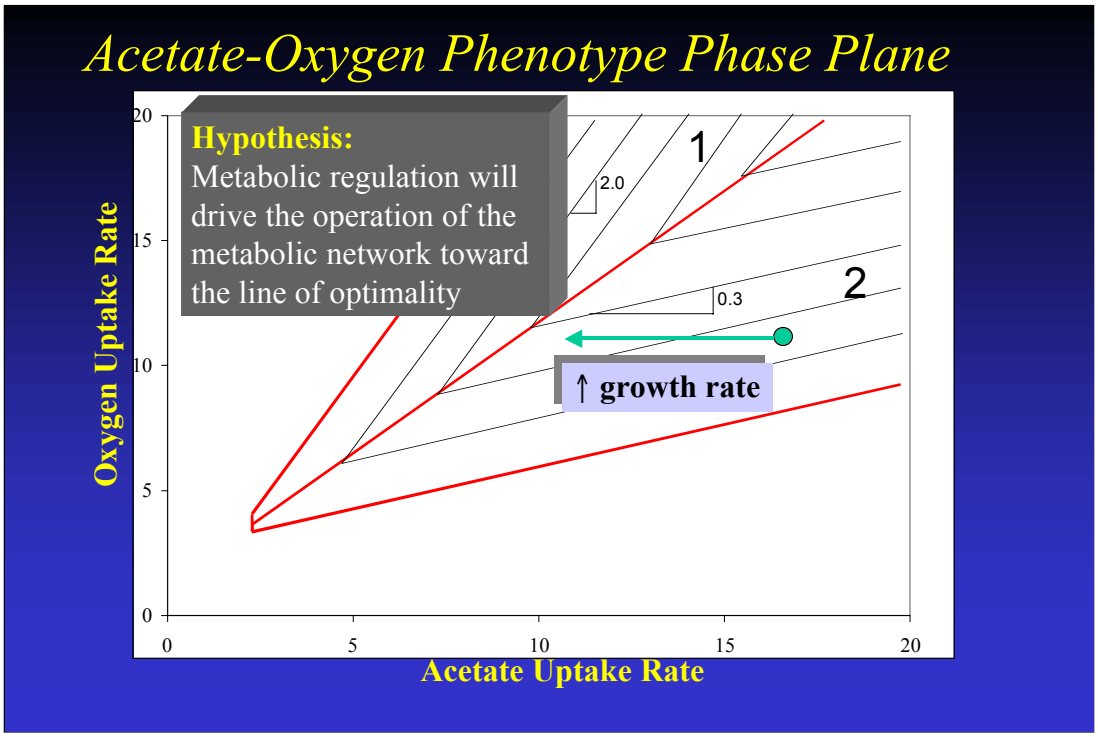


An example of a phase plane for a genome scale metabolic map.

From J.S. Edwards and B.O. Palsson (1999), "Systems Properties of the *Haemophilus influenzae* Rd Metabolic Genotype," *The Journal of Biological Chemistry*, **274**: 17410-17416.



Perhaps the most useful application of in silico strains is to design meaningful experiments. Agreement confirms the model, while failure indicates that the model is missing features. Therefore we like failure, so that the model can be continually improved.



INTERPRETING THE PHASE PLANE:

Using isoclines

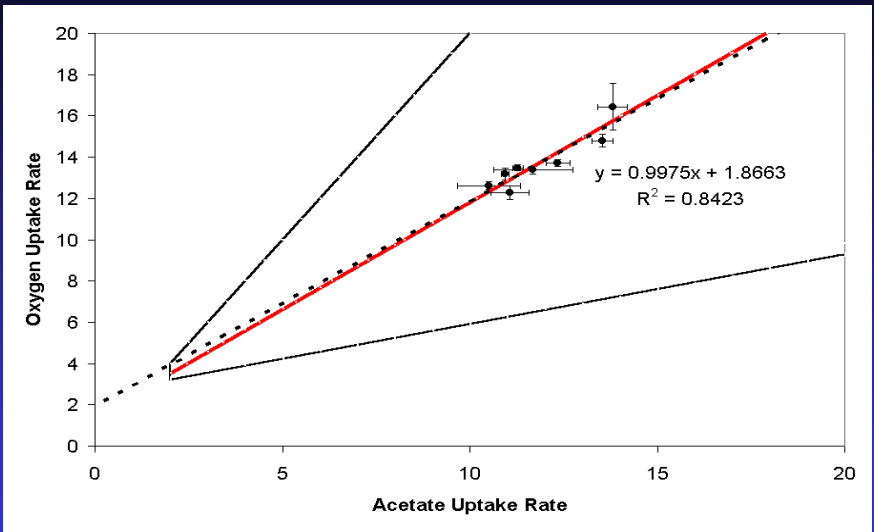
This slide describes the acetate-oxygen phenotype phase plane for E. coli.

It can be seen that there are 2 distinct regions. We have also drawn the isoclines on this figure, and it can be seen that the isoclines have a positive slope in both regions. This means that they are unstable -- it is advantageous for the organism to move the the edge of the region

The optimal growth occurs at the line separation the two phases, the so-called line of optimality.

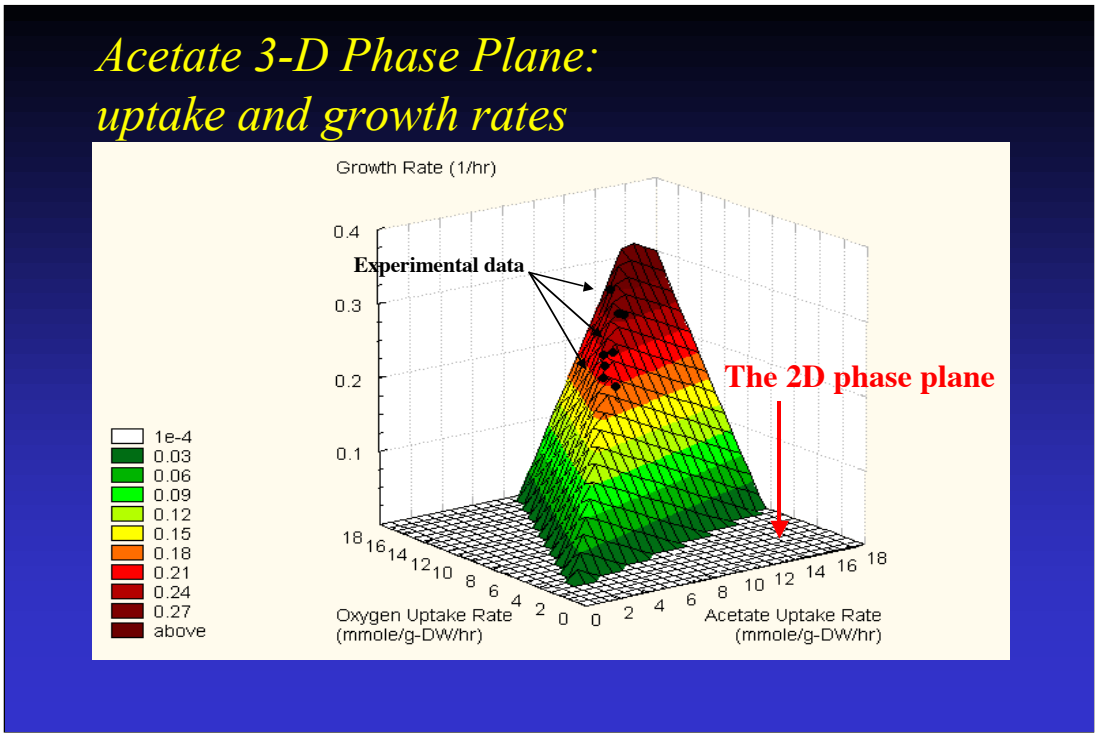
The thinner lines in each feasible phase plane are called isoclines. They denote a constant growth rate.

Acetate Phase Plane:Experimental Data



THE EXPERIMENTAL DATA:

Right on the line!



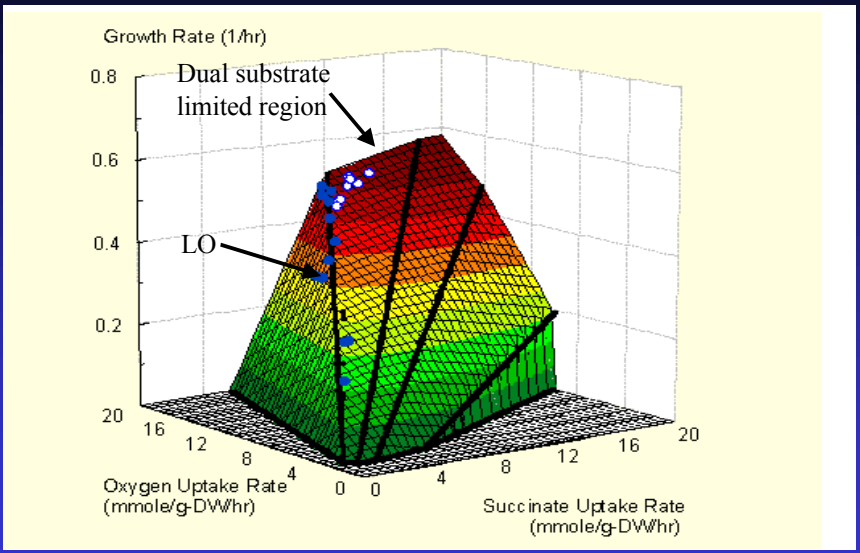
3D REPRESENTATION:

Including growth rate as a dependent variable

This slide shows how the maximal growth rates can be graphed above the phenotypic phase plane. We see the outline of a cone. For a given maximal uptake rate of either acetate or oxygen, the best (highest growth rate) solution is on the edge of the cone.

The experimental data falls there, indicating that the E. coli strain has optimized its growth rate on acetate.

Succinate 3-D Phenotype Phase Plane



This also works for other substrates!

The case of succinate

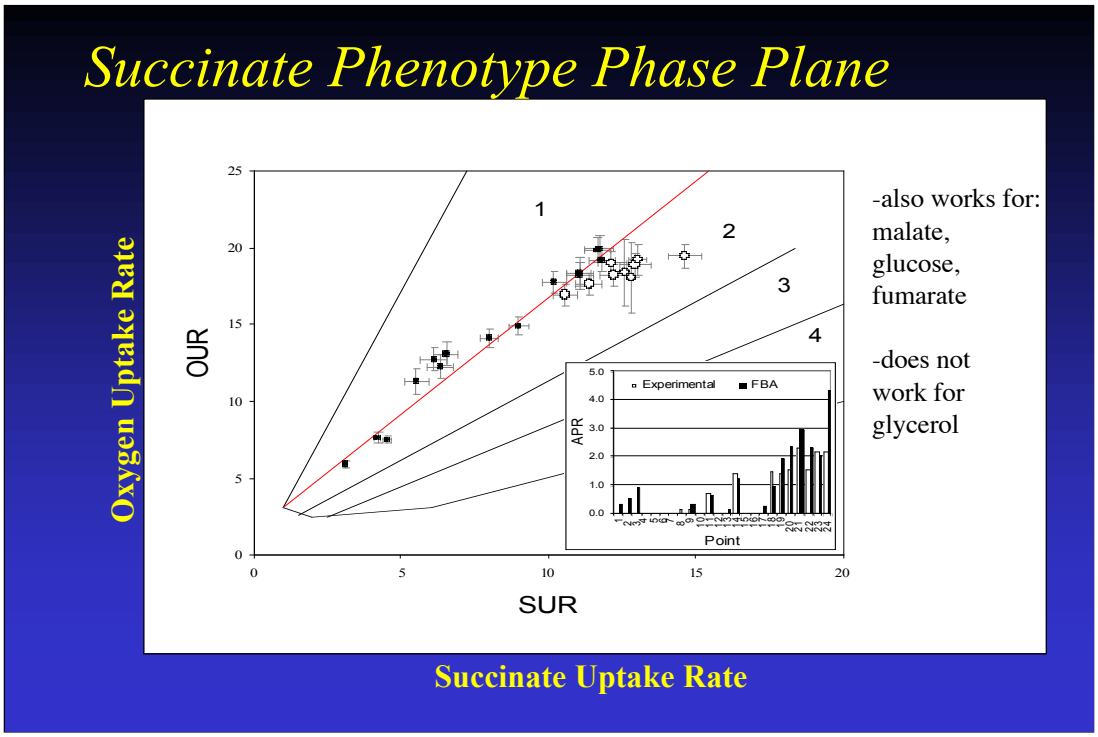
This figure, shows the succinate-oxygen PhPP in three dimensions.

•The formalism is similar to the 3-D acetate PhPP

•Here the effect of the carbon source on the structure of the PhPP can be seen.

•The LO is shown here, and the data points with reduced succinate uptake rates all lie on (or near) the LO,

•However, when the succinate uptake rate was increased, the experimental data followed the LO until the oxygen mass transfer constraint was reached. At this point, the growth rate and the succinate uptake were increased by moving into region 2 of the phase plane (white data points).



Succinate in 2D

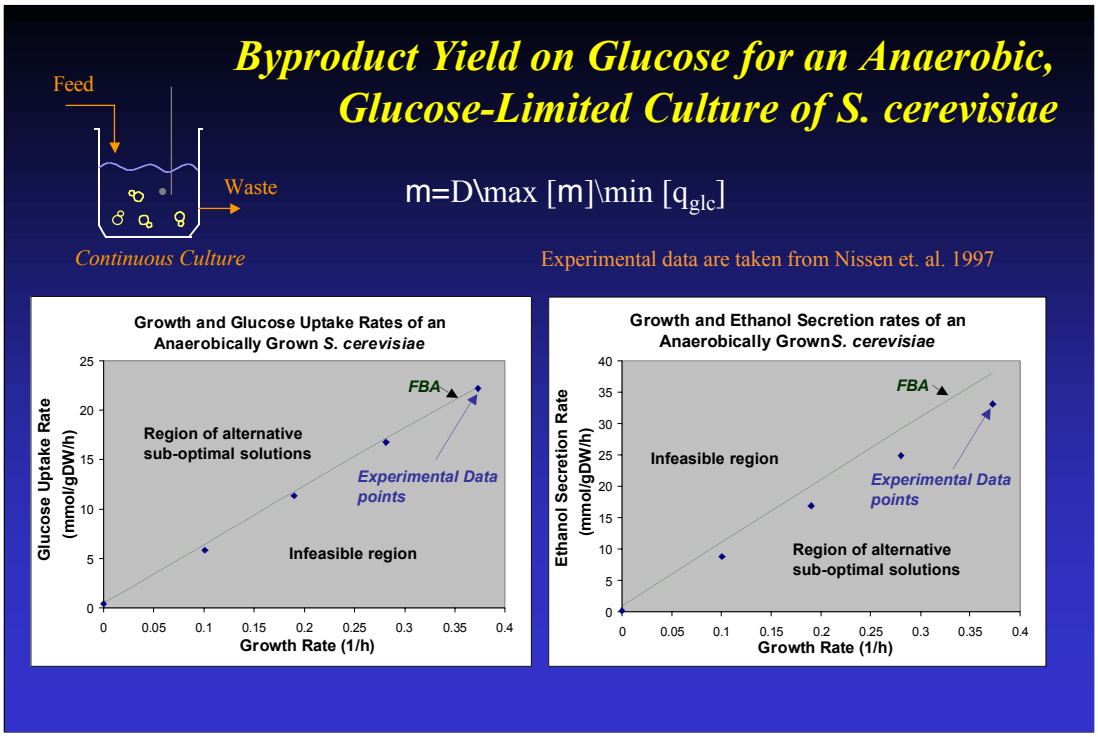
All the experimental data points were plotted onto the succinate-oxygen PhPP. And the results are shown on this slide.

Consistent with the maximal growth hypothesis, all the data points were constrained to region 2 of the PhPP.

•Within region 2, all the points were restricted to two different regions.

- •either they were on the LO, or
- •they were at a maximal oxygen uptake rate with the succinate uptake rate defining points within region 2.

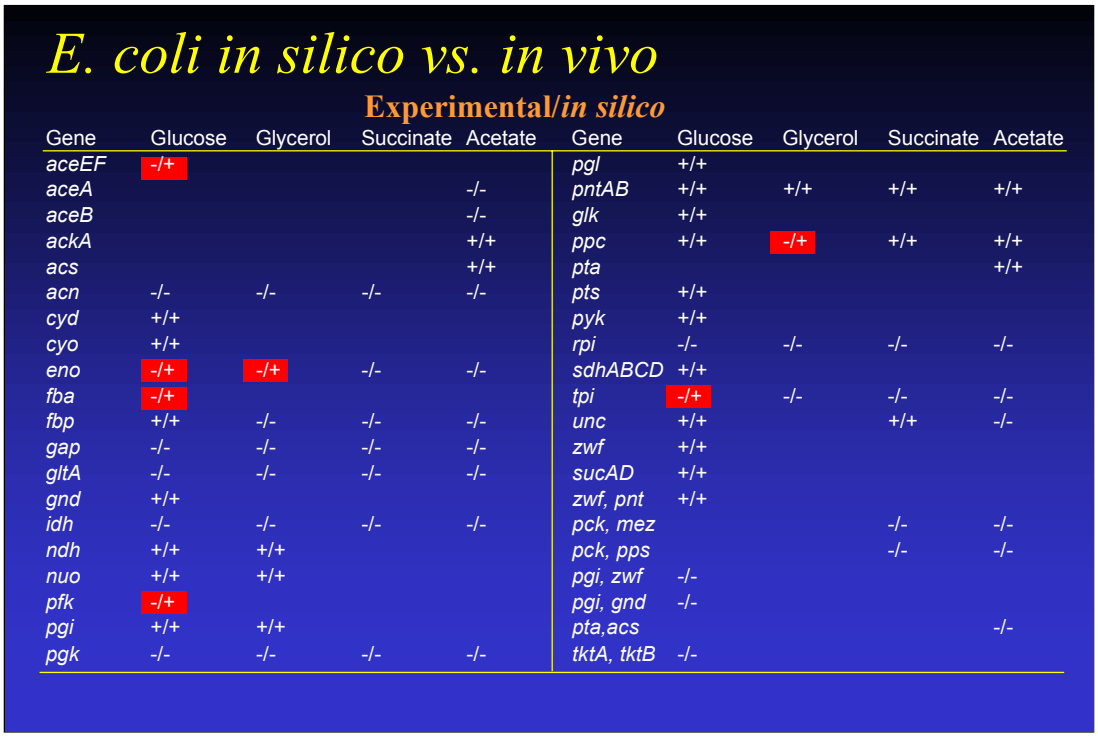
•The insert shows the calculated and measured acetate secretion rate in within region 2



CONTINUOUS CULTURE OF YEAST

In continuous culture the dilution rate specifies the growth rate. In the previous slides we maximized the growth rate (output) for a given uptake rate (input). Here in contrast, we fix the growth rate (output) and thus minimized the input (the uptake rate). The in silico solution and the measured uptake rates are shown and they agree reasonably well. Yeast seems to operate close to the edge of its allowable solution cone.

This formalism can be used to examine changes in the genotype. Genes can be added or deleted and the consequences on the ability to grow, or to generate other phenotypes, can be calculated and compared to the wild type in silico strain.



DELETION STUDY

•An important question arises as to how well these in silico predictions represent the actual metabolic behavior.

•The plus/minus nomenclature represents the ability of the respective mutant cell to grow. The first being the experimental determination, and the second being the in silico prediction.

•We have compared our in silico results to the growth of mutants in about 80 different conditions reported in the literature, and the results are summarized on this slide. The in silico strain correctly predicted the ability to grow in all but 7 cases.

•The inaccuracies are highlighted here by the red boxes.

EMBRACING AND ANALYZING FAILURE

•There are 7 inaccuracies, and they are explained by 2 basic reasons.

- •1. toxic intermediate production
- •2. metabolic regulation

•For example, when these gene products are removed, it is thought that the cell produces a toxic intermediate, and this prevents the cell from growing. This can not be predicted using the methodology that I have introduced.

•Also, when the enolase gene is removed from the system, the experimental data suggests that this cell is unable to grow, whereas the *in silico* cell is able to grow, and upon further examination, it is seen that the *in silico* cell is able to grow by synthesizing and degrading an amino acid, something that the cell is unlikely to do.

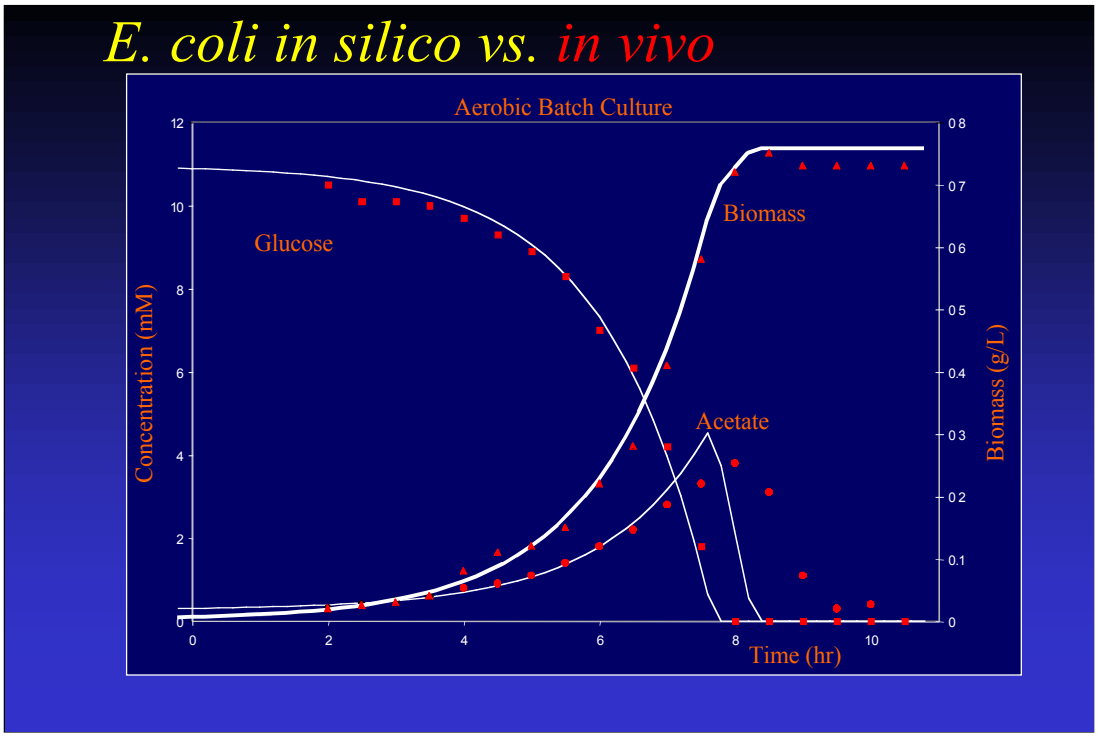
•However, it has been observed that revertants can spontaneously arise with altered expression.

•For example, ATPase mutants have been shown not to grow on succinate, however, this metabolic model predicts that they theoretically can. It was recently reported a couple of months ago that the ATPase deletion strains were unable to grow due to a transport deficiency, and revertants arose after about a week that do grow on succinate, at yields near the theoretical maximum.

Predicting expression arrays

Metabolic maps show the phenotype. Expression arrays also show the phenotypes. One is a flux phenotype whereas the other is the expression phenotype. The two cannot be directly and quantitatively compared.

However, the two can be qualitatively compared for a transition from one state to another. Pathways need to be up and down regulated. The patterns of the two can be compared qualitatively, i.e. in an off/on sense,



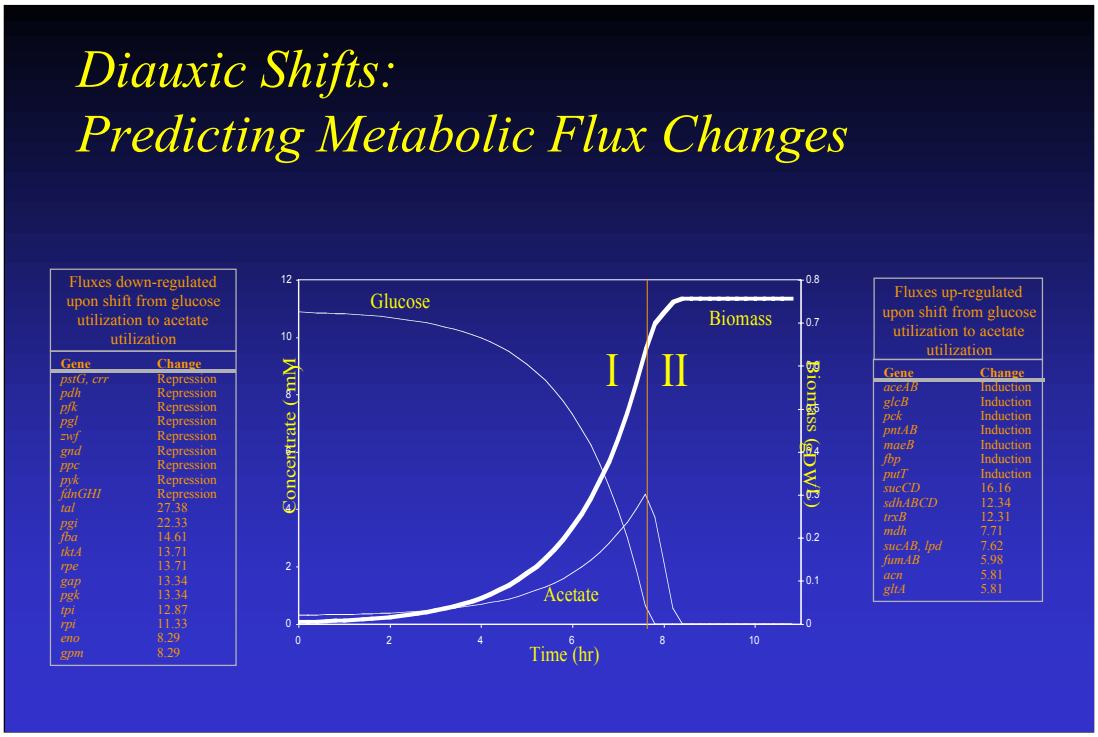
METABOLIC SHIFTS

This slide demonstrates an aerobic batch culture in glucose minimal media. The lines are the FBA predictions from a quasi-steady state simulation in a batch culture, and the points are experimentally determined.

This line represents the glucose concentration in the media, and it can be seen, as the glucose is utilized, the cells grow, and produce acetate. At this point, the glucose is completely utilized from the media, and the simulation predicts the reutilization of the acetate, and this is also experimentally observed.

However, it is at this point that the in silico predictions deviate from the experimental data. Due to the steady state assumption, the in silico strain is able to immediately reutilize the acetate. However, the experimental data lags behind by about 40 minutes.

This lag is due to the time required to to adjust the metabolic network for acetate utilization.

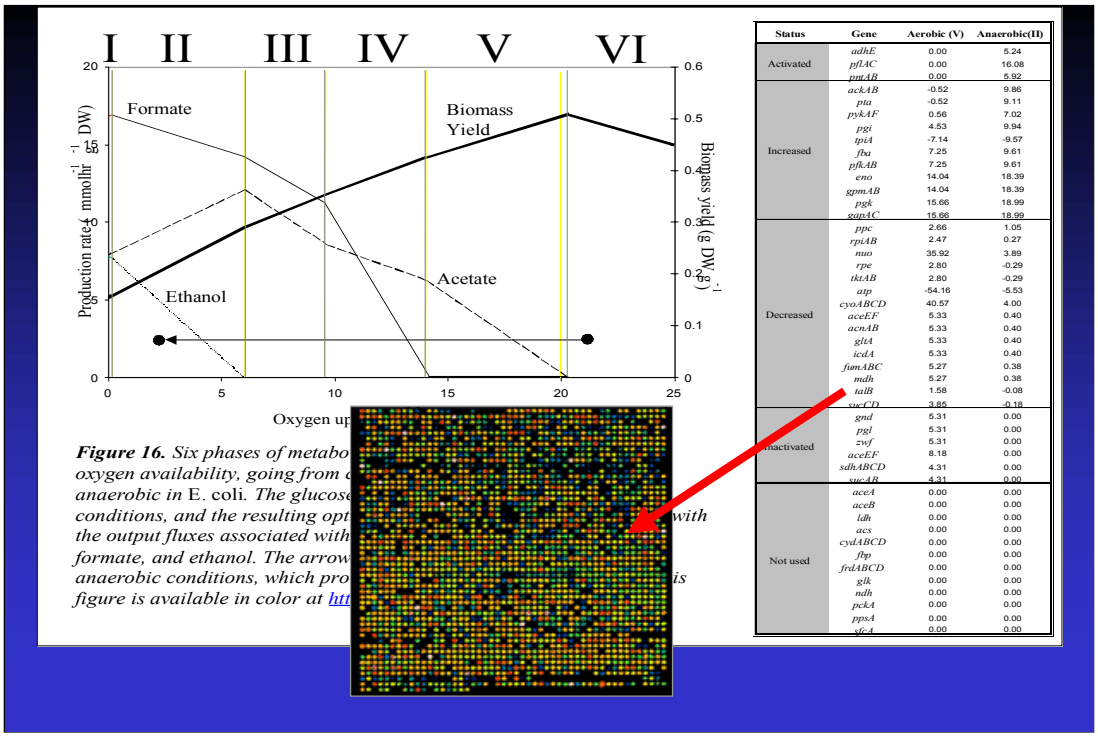


THE DIAUXIC SHIFT

This figure reproduces the data from an earlier slide, where batch growth on glucose was observed with the secretion of acetate. Then the acetate was re-consumed. The flux maps for growth on acetate and glucose are quite different. The relative flux levels through all the steps can be compared. Based on such comparisons relative fluxes through the different metabolic steps can be estimated.

If the expression levels are proportional to the needed flux levels then the indicated (predicted) up- and down-regulation of genes should be observed.

This result is a testable experimental hypothesis.



Status	Gene	Aerobic (V)	Anaerobic(II)
Activated	*adhE*	0.00	5.24
	pflAC	0.00	16.08
	pntAB	0.00	5.92
Increased	*ackAB*	-0.52	9.86
	pta	-0.52	9.11
	pykAF	0.56	7.02
	pgi	4.53	9.94
	tpiA	-7.14	-9.57
	fba	7.25	9.61
	pfkAB	7.25	9.61
	eno	14.04	18.39
	gpmAB	14.04	18.39
	pgk	15.66	18.99
	gapAC	15.66	18.99
Decreased	*ppc*	2.66	1.05
	rpiAB	2.47	0.27
	nuo	35.92	3.89
	rpe	2.80	-0.29
	tktAB	2.80	-0.29
	atp	-54.16	-5.53
	cyoABCD	40.57	4.00
	aceEF	5.33	0.40
	acnAB	5.33	0.40
	gltA	5.33	0.40
	icdA	5.33	0.40
	fumABC	5.27	0.38
	mdh	5.27	0.38
	talB	1.58	-0.08
	sucCD	3.85	-0.18
Inactivated	*gnd*	5.31	0.00
	pgl	5.31	0.00
	zwf	5.31	0.00
	aceEF	8.18	0.00
	sdhABCD	4.31	0.00
	sucAB	4.31	0.00
Not used	*aceA*	0.00	0.00
	aceB	0.00	0.00
	ldh	0.00	0.00
	acs	0.00	0.00
	cydABCD	0.00	0.00
	fbp	0.00	0.00
	frdABCD	0.00	0.00
	glk	0.00	0.00
	ndh	0.00	0.00
	pckA	0.00	0.00
	ppsA	0.00	0.00
	sfcA	0.00	0.00

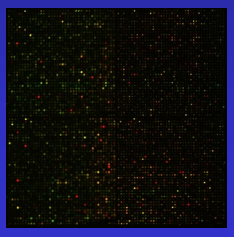
Figure 16. *Six phases of metabo*
oxygen availability, going from a
anaerobic in E. coli. *The glucose*
conditions, and the resulting opt *with*
the output fluxes associated with
formate, and ethanol. The arrow
anaerobic conditions, which pro *is*
figure is available in color at htt

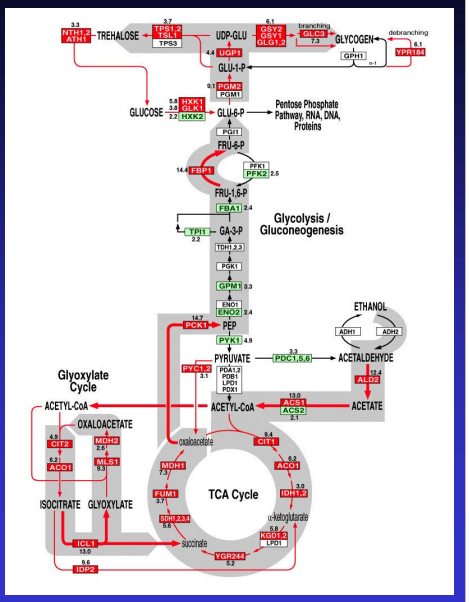
The relative *in silico* calculated fluxes can be compared to the relative expression levels under the two conditions considered. Only qualitative comparisons can be made since the flux is not proportional to the expression levels.

Gene expression on a genomic scale

- Analyzing the gene expression patterns during growth in different conditions
  - oxygenation, carbon sources
- Comparing the gene expression patterns to the FBA predictions - wild-type & knockout strains
- Goal is to relate gene expression patterns to metabolic pathway utilization

DeRisi et.al.,
Science, 278:680-686, 1997





DIAUXIC SHIFT IN YEAST FOR GROWTH ON GLUCOSE

It has been shown by Patrick Brown's group at Stanford, that the shift in metabolic pathway utilization can be determined from genomic scale measurements of gene expression.

They have generated cDNA micro-arrays with probes for virtually every gene in the yeast genome, and used these micro-arrays to study the changes in gene expression on a genome scale during a diauxic shift from glucose to ethanol utilization.

Shifts in expression levels that correspond to pathway usage were observed.

Summary

- Maximum capacity constraints close the flux cone
- LP can be used to find optimal solutions in the so formed closed solution space
- There are many types of objectives that can be studied; perhaps the maximal growth rate is the most appropriate
- Methods can be developed to show all optimal solutions as a function of environmental parameters
- The phase plane analysis shows that there is a finite number of optimal phenotypes
- This analysis can be used to interpret and predict the consequences of losing genes and the expression changes during shifts from one growth condition to another

References

- Papoutsakis, E.T., "Equations and calculations for fermentations of butyric acid bacteria," *Biotechnol Bioeng*, **26**: 174-187 (1984).
- Papoutsakis, E. and Meyer, C.,"Equations and calculations of product yields and preferred pathways for butanediol and mixed-acid fermentations," *Biotechnol Bioeng*, **27**: 50-66 (1985).
- Fell, D.A. and Small, J.A., "Fat synthesis in adipose tissue. An examination of stoichiometric constraints," *J. Biochem*, **238**: 781-786 (1986).
- Majewski, R.A., and Domach, M.M., "Simple constrained optimization view of acetate overflow in *E. coli*," *Biotechnol Bioeng,* **35**: 732-738 (1990).
- Savinell, J.M., and Palsson, B.O., "Network analysis of intermediary metabolism using linear optimization. II. Interpretation of hybridoma cell metabolism," *J Theor Biol*, **154**: 455-473 (1992).
- Varma, A. and Palsson, B.O., "Metabolic capabilities of *Escherichia coli*. I. Synthesis of biosynthetic precursors and cofactors," *J Theor Biol*, **165**:477-502 (1993).
- Varma, A. and Palsson, B.O., "Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110," *Appl Environ Microbiol*, **60**: 3724-3731 (1994).

References

- Varma, A. and Palsson, B.O., "Predictions for oxygen supply control to enhance population stability of engineered production strains," *Biotechnol Bioeng*, **43**: 275-285 (1994).
- Varma, A. and Palsson, B.O., "Parametric sensitivity of stoichiometric flux balance models applied to wild-type *Escherichia coli* metabolism," *Biotechnol Bioeng*, **45**: 69-79 (1995).
- Pramanik, J. and Keasling, J.D., "Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and and mechanistic energy requirements," *Biotechnol Bioeng*, **56**: 398-421 (1997).
- Bonarius, H. P. J., Schmid, G. & Tramper, J. (1997) *Trends in Biotechnology* **15,** 308-314.
- J.S. Edwards, R. Ramakrishna, C.H. Schilling, and B.O. Palsson, "Metabolic Flux Balance Analysis," *Metabolic Engineering* **2**: 13-57, (1999), S.Y. Lee, Papoutsakis, E.T., Eds; Springer-Verlag: New York.
- J.S. Edwards and B.O. Palsson, "Systems Properties of the *Haemophilus influenzae* Rd Metabolic Genotype," *The Journal of Biological Chemistry*, **274**: 17410-17416 (1999).
- J.S. Edwards and B.O. Palsson, "The *Escherichia coli* MG1655 *in silico* metabolic genotype; Its definition, characteristics, and capabilities," *Proc. Natl Acad Sci (USA)*, **97**: 5528-5523 (2000).

The biological design variables: kinetic and regulatory constraints

Bernhard Palsson

Hougen Lecture #6

Nov 21th, 2000

INTRODUCTION

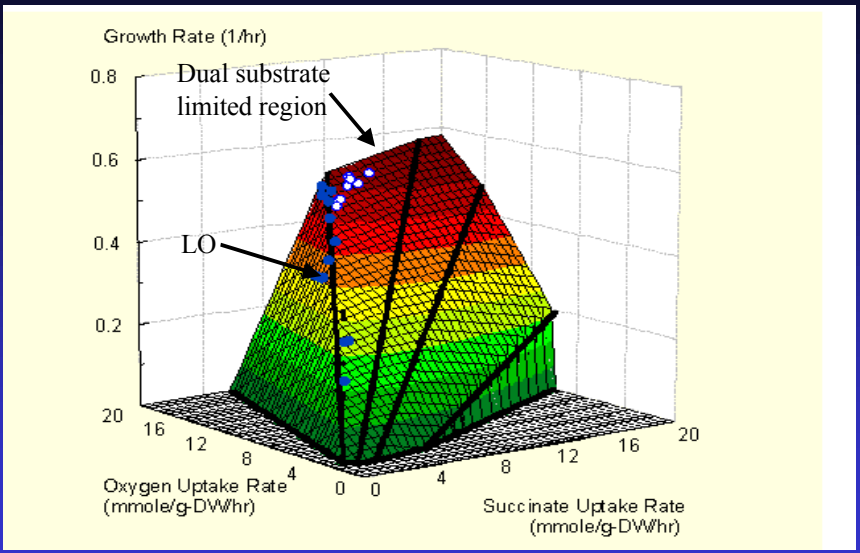
We have up to this point imposed the constraints that arise from basic physico-chemical considerations. Now we look at biological, "self-imposed" constraints.

Lecture #6: Outline

- Brief recap of Lecture #5
- E. coli as an optimizer
- Engineering vs. biological design procedures
- Accounting for regulation of gene expression:
  - Logistical and flux balance representation
  - Examples: multiple substrates
- Dealing with kinetics
  - Numerical values of kinetic constants
  - Relative values
  - Temporal decomposition
- Numerics

LECTURE #6

Succinate 3-D Phenotype Phase Plane



This also works for other substrates!

The case of succinate

This figure shows the succinate-oxygen PhPP in three dimensions.

•The formalism is similar to the 3-D acetate PhPP

•Here the effect of the carbon source on the structure of the PhPP can be seen.

•The LO is shown here, and the data points with reduced succinate uptake rates all lie on (or near) the LO,

•However, when the succinate uptake rate was increased, the experimental data followed the LO until the oxygen mass transfer constraint was reached. At this point, the growth rate and the succinate uptake were increased by moving into region 2 of the phase plane.

•How do cells find this optima?

Engineering Design

- Objective
  - separation of protein, building a bridge, designing a car, etc
- Constraints:
  - geometry, materials, diffusion constants, cost, time
- Design envelope
- Optimize design using free design variables
  - optimal engineering designs do evolve

Engineering design begins with a statement of an objective; i.e. separating a protein or building a bridge. The constraints on the design are then defined. Cost and time are always important, but so are material properties (strength, elasticity, etc), physical constants (diffusivities, thermal conductivities), and geometric considerations. These constraints then define a design envelope within which the design must fall. Optimization of the design is then carried out within the allowable ranges to produce the 'best' design.

Constraints on biological networks

- Stoichiometry
- Maximum Capacities
- P/C constraints
  - Diffusion, electroneutrality
- Kinetics/Regulation

→ • Non-adjustable
  - Horizontal gene transfer

→ • Upper limit
  - Downwardly adjustable by gene expression

→ • Non-adjustable

→ • Highly adjustable
  - Evolutionary design

Engineering vs. Biological Design

- Objective
  - Separation of protein
- Constraints:
  - Geometry
  - Materials
  - Diffusion constants
- Design envelope
- Optimize design using free design variables

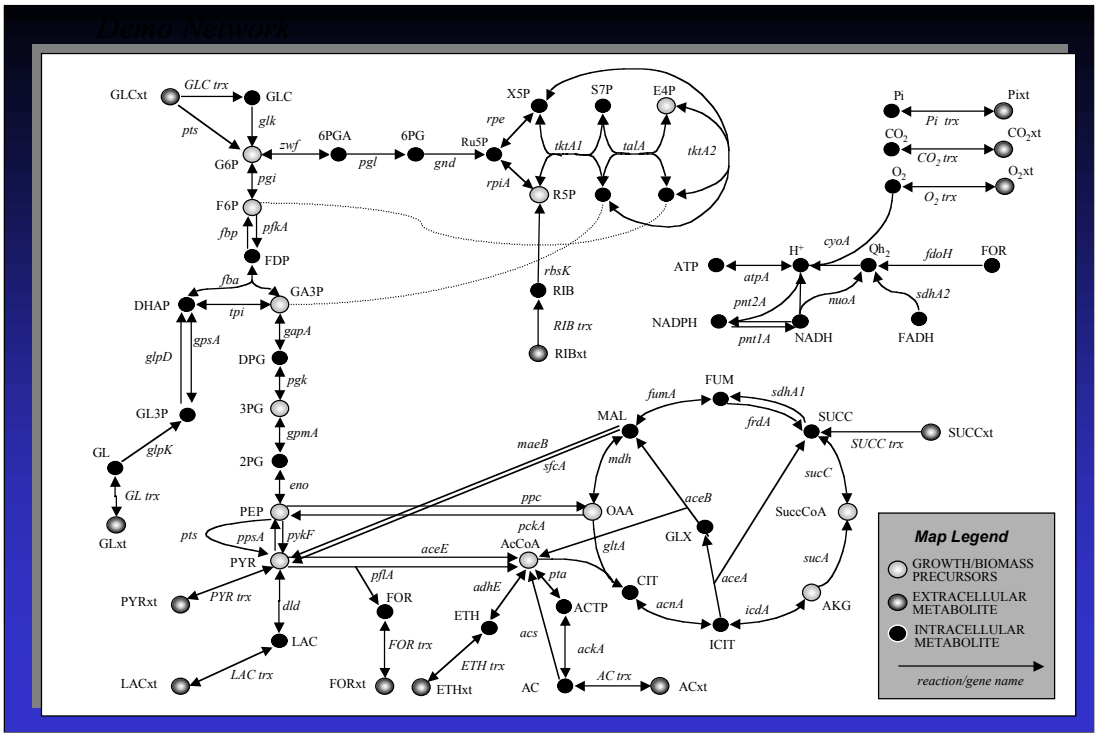
- Objective
  - Survival, growth
- Constraints:
  - Max fluxes
  - Connectivity
  - P/C factors
- Solution space
- Optimize design using kinetic and regulatory variables

There is some uncertainty about how to apply the basic physical laws in the intra-cellular milieu and even if we knew how, we would not have numerical values for the myriad of constants that appear in such equations. The alternative approach relies on the successive imposition of constraints that govern biochemical reaction networks. Such constraints include the maximum flux achievable through a reaction, the connectivity of the network and so forth. The imposition of these constraints defines a solution space, similar to the design envelope discussed above. The 'best' solution in the allowable solution space is then determined based on an optimization procedure. The optimization is based on an assumed objective that the cell is striving to meet. A match has been obtained between measured growth and metabolic by-product secretion of E. coli K-12 for growth on acetate and succinate and the calculated optimal performance based on the constraint-based approach.

Biological Design

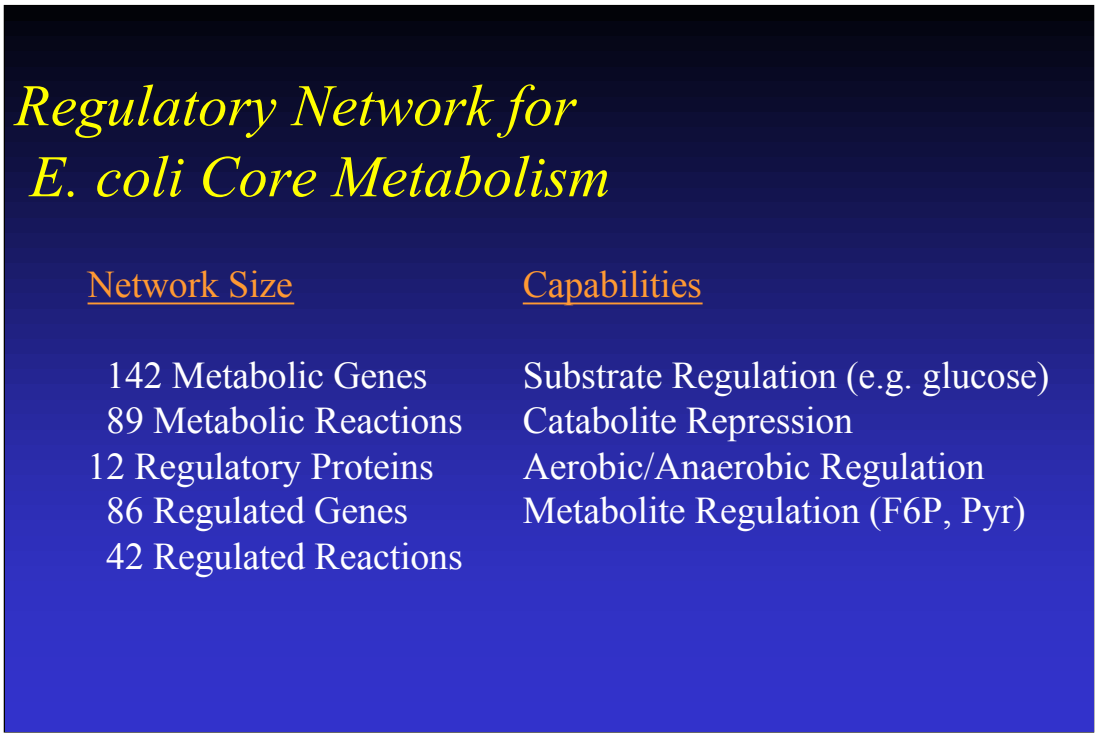
Regulation of expression:
shaping solution spaces
Regulation of activity:
location within a solution space

Given the solution space that is determined in part by hard physicochemical constraints, the exact solution is determined by the kinetic and regulatory parameters that the cell can alter. Thus, we can now view the kinetic and regulatory parameters as 'biological design' variables, based on an analogy with the engineering design procedure. In order for this analogy to hold and to view the kinetics as biological design variables, we must be able to observe the evolutionary motion of a suboptimal design towards an optimal under the given constraints.



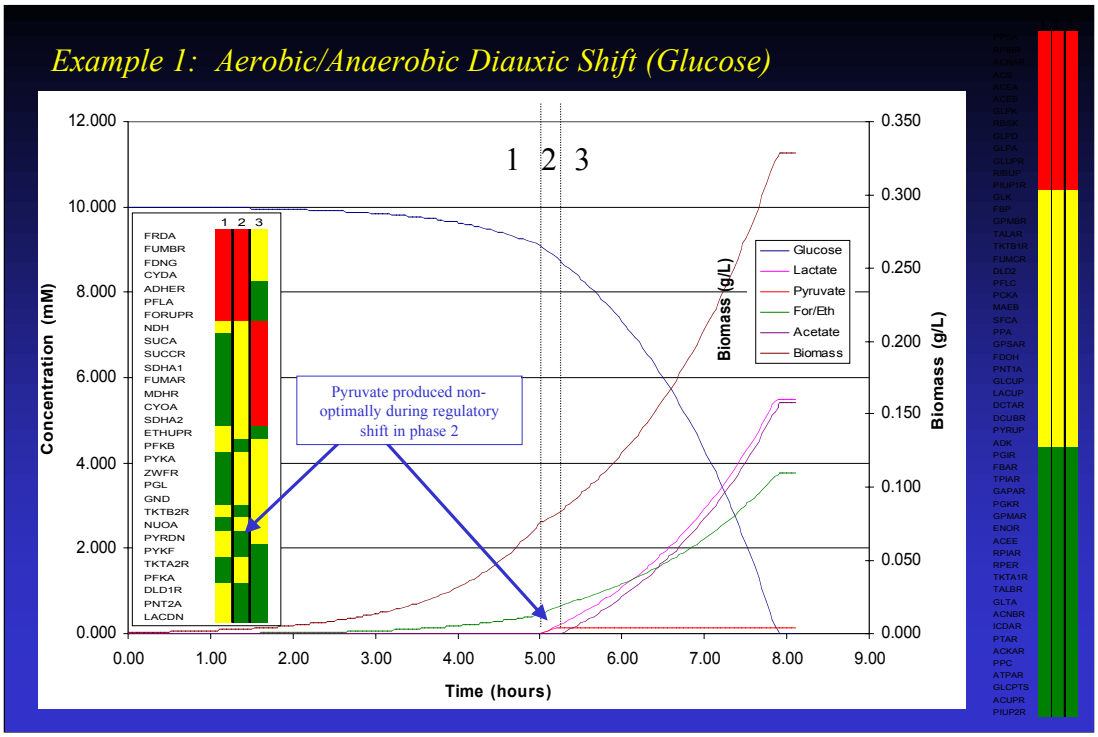
Logistical -FBA Models

Known regulatory effects can be used to close off or open links in the network. The known operon structure for *E. coli* can be used to implement a condition-dependent map available to the cell.

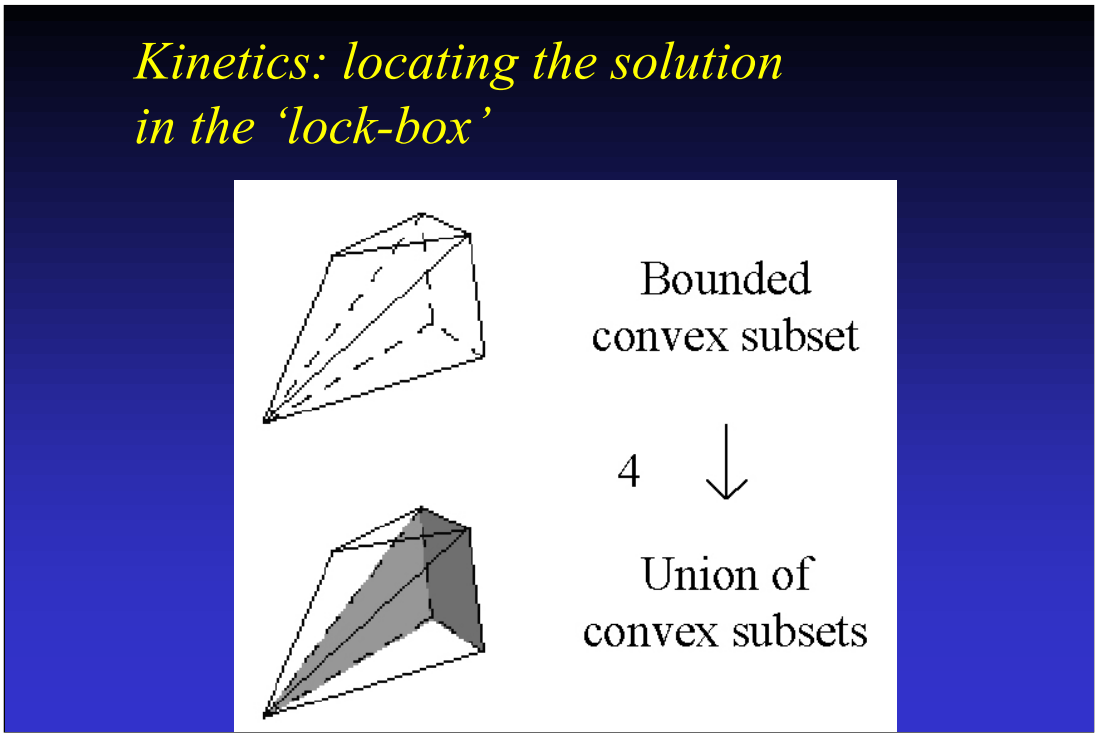


SPECS

These are the specifications on the regulated core *E. coli* metabolic model.



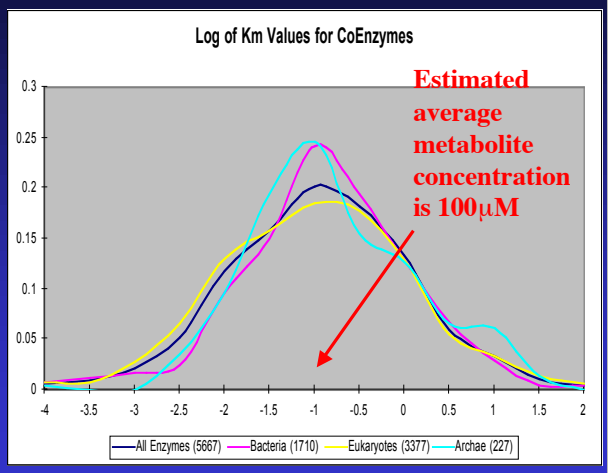
Dynamic simulations of the regulated *E. coli* model. The bar to the left shows changes in gene expression, while the expression of the genes described in the bar on the right does not change.



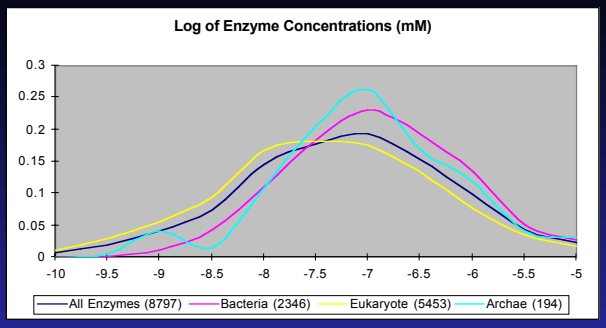
Regulation of gene expression and maximal flux constraints close-off a solution space. The exact location of the solution in the 'lock-box' will be determined by the numerical values of the kinetic constraints.

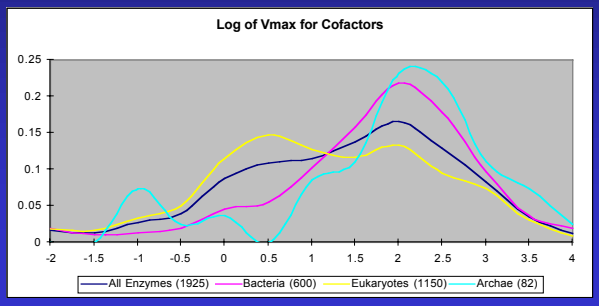
Numerical values of kinetic constants

- Compilations of legacy data
  - i.e. EMP data base
- Determine how well we need to know the kinetic parameters
  - Order-of-magnitude



Enzymes

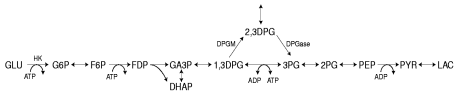

Log of Enzyme Concentrations (mM)
0.3
0.25
0.2
0.15
0.1
0.05
0
-10
-9.5
-9
-8.5
-8
-7.5
-7
-6.5
-6
-5.5
-5
All Enzymes (8797)
Bacteria (2346)
Eukaryote (5453)
Archae (194)


Log of Vmax for Cofactors
0.25
0.2
0.15
0.1
0.05
0
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
All Enzymes (1925)
Bacteria (600)
Eukaryotes (1150)
Archae (82)

Orders of Magnitude: Kinetics and edges of solution cones: Use of dimensionless groups

RBC Network:

(a)



Extreme Pathways:

(b)

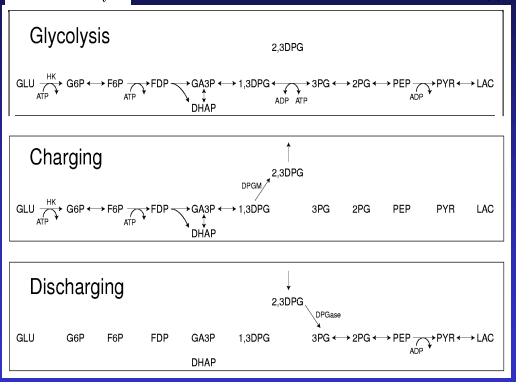


Figure 9. *(A) The simplified red blood cell reation network comprised of only glycolysis and Rapoport-Leubering shunt. (B) The three extreme pathways for this network; glycolysis, charging, discharging.*

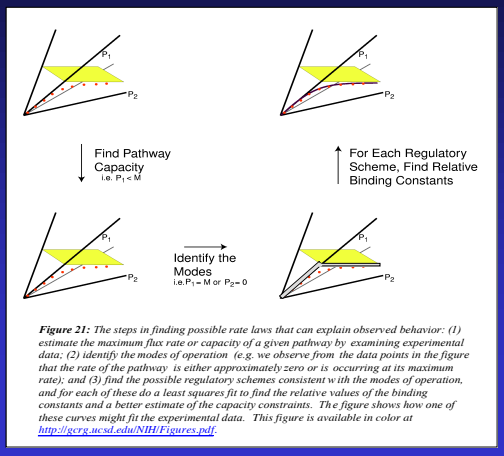
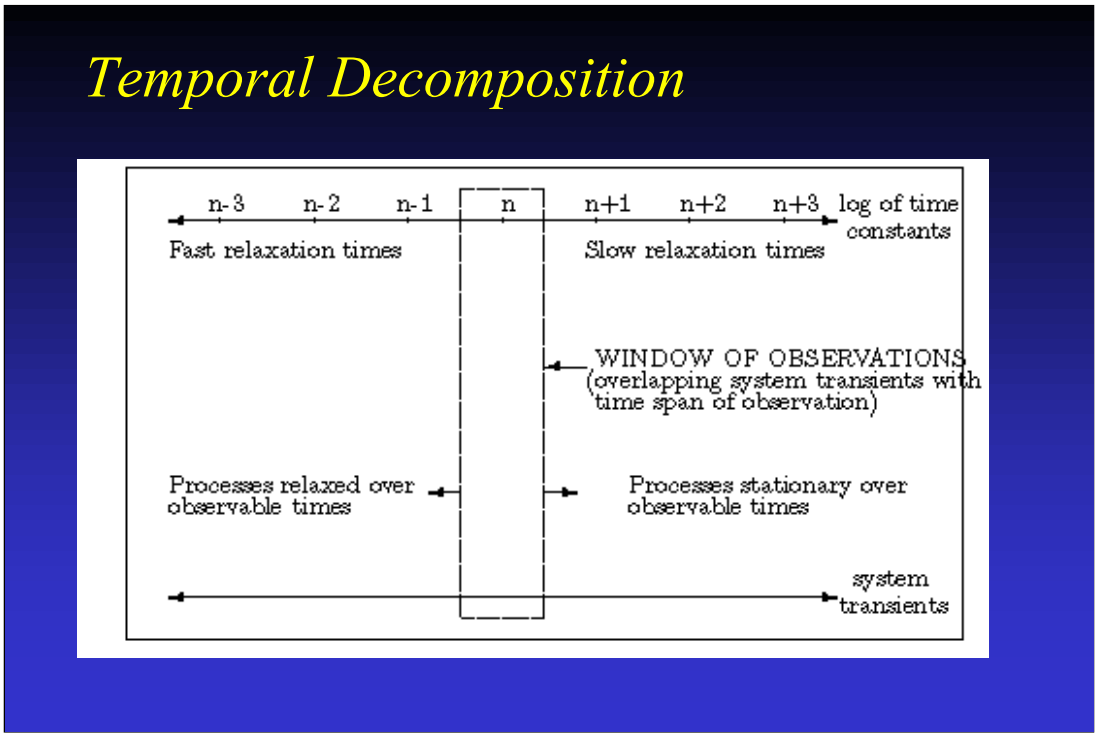


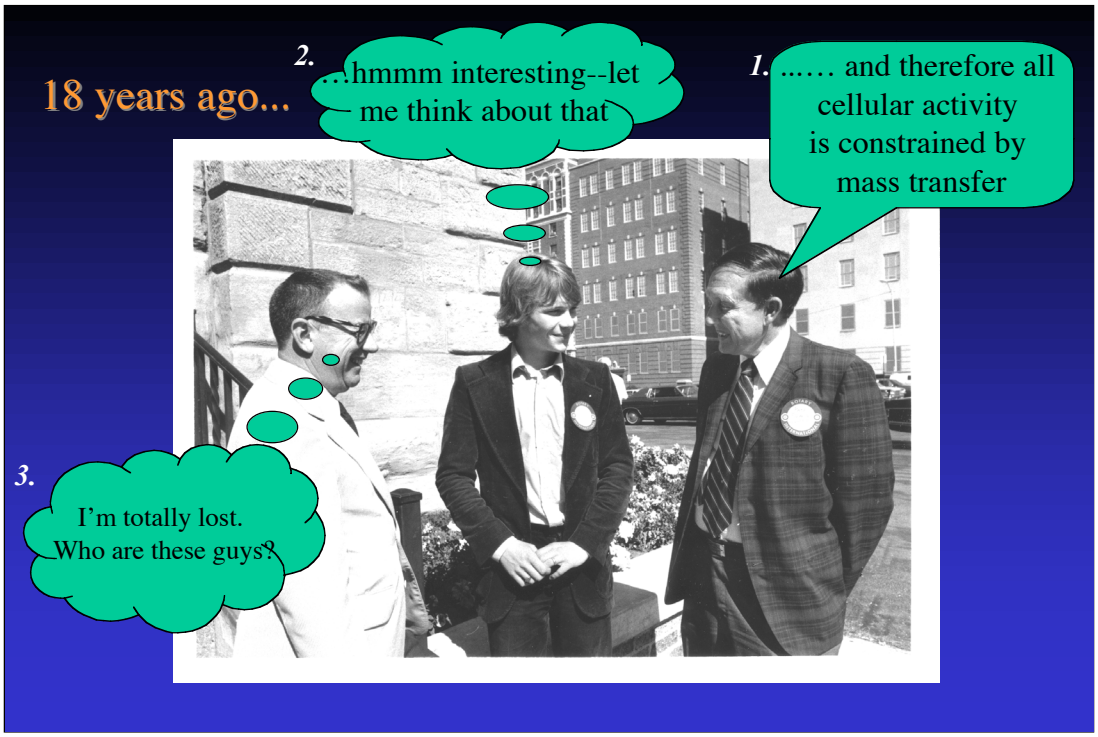
Figure 21: *The steps in finding possible rate laws that can explain observed behavior: (1) estimate the maximum flux rate or capacity of a given pathway by examining experimental data; (2) identify the modes of operation (e.g. we observe from the data points in the figure that the rate of the pathway is either approximately zero or is occurring at its maximum rate); and (3) find the possible regulatory schemes consistent w ith the modes of operation, and for each of these do a least squares fit to find the relative values of the binding constants and a better estimate of the capacity constraints. The figure shows how one of these curves might fit the experimental data. This figure is available in color at* http://gcrg.ucsd.edu/NIH/Figures.pdf.



TEMPORAL DECOMPOSITION

The hierarchy of intrinsic times can be represented by the time axis. Fast transients are characterized by the processes at the extreme left and slow transients at the extreme right. The process time scale, i.e. the time scale of interest, can be represented by a *window of observation* on this time axis. One can conceptualize this readily by looking at a three-dimensional system where one time constant represents the fast motion; the second, the time scale of interest; and the third, a slow motion.

The terms which have time constants faster than the observed window can be eliminated from the dynamic description as these terms are small. However, the mechanisms which have transients slower than the observed time exhibit high "inertia"' and hardly move from their initial state and can be considered constants. One can thus remove slow or fast terms by the appropriate use of the eigenrows and eigenvectors.



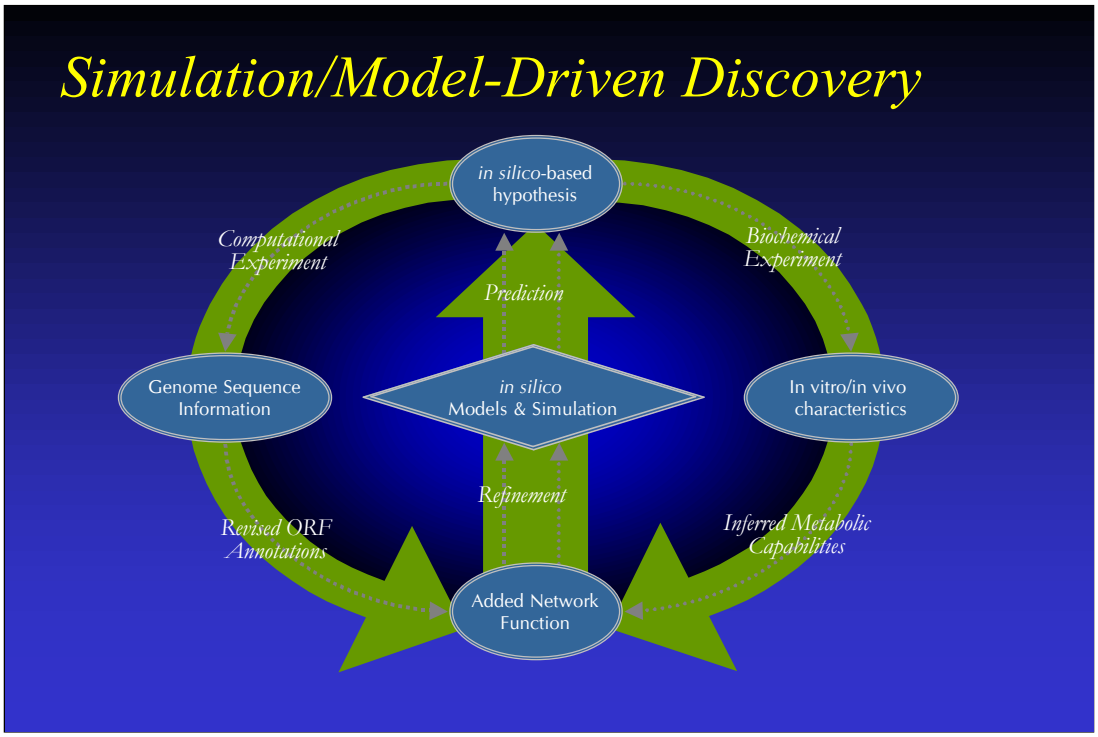
A Personal Reflection

Some Lessons: towards principles

- Importance of Constraints
  - Cells are constrained in their behavior and seem to push close to these constraints ('life on the edge')
  - Extension of the concept of Mass Transfer limitations
    - (E.N. Lightfoot)
- A large number of components (complex genotypes) display relatively few overall types of behaviors (phenotypes)

Simplicity from complexity: the evidence mounts

- Singular value decomposition of genome-scale expression data is in uncovering simple underlying patterns
- Modal analysis of dynamic models of metabolism shows simple dynamic structures
- Robustness analysis of kinetic models of biochemical systems models reveals insensitivity to individual kinetic constants



The model building process is an iterative one. We must learn to embrace failure.

Summary

- Metabolic genotypes can be formulated based on annotated sequence data
- Using the biochemical properties of the gene products and other information, a genome-scale metabolic network can be formulated
- Flux distributions through this network cannot be uniquely calculated, but optimal phenotypes can
- Testable experimental hypotheses can be generated in this way and have been put forth for *E. coli* growth on acetate and succinate
- Further testing is needed to assess the generality of the approach
- It forms the basis for iterative model building within the framework of applying successive constraints

--The End--

Hougen 2000 Lectures