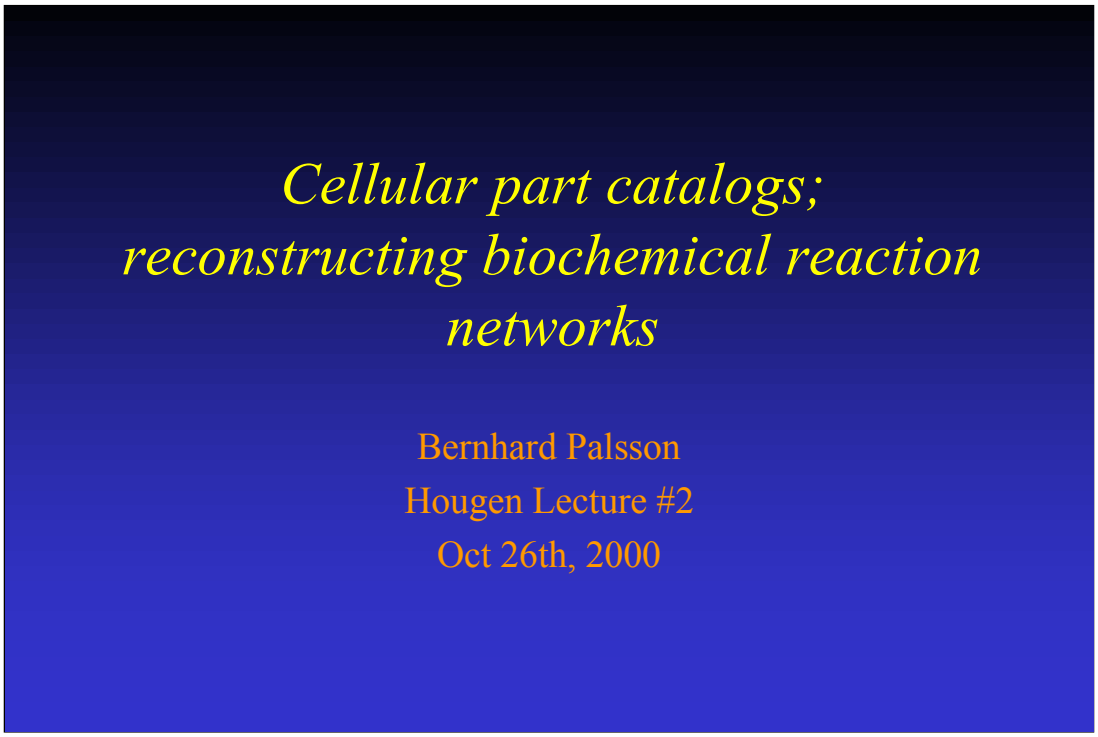


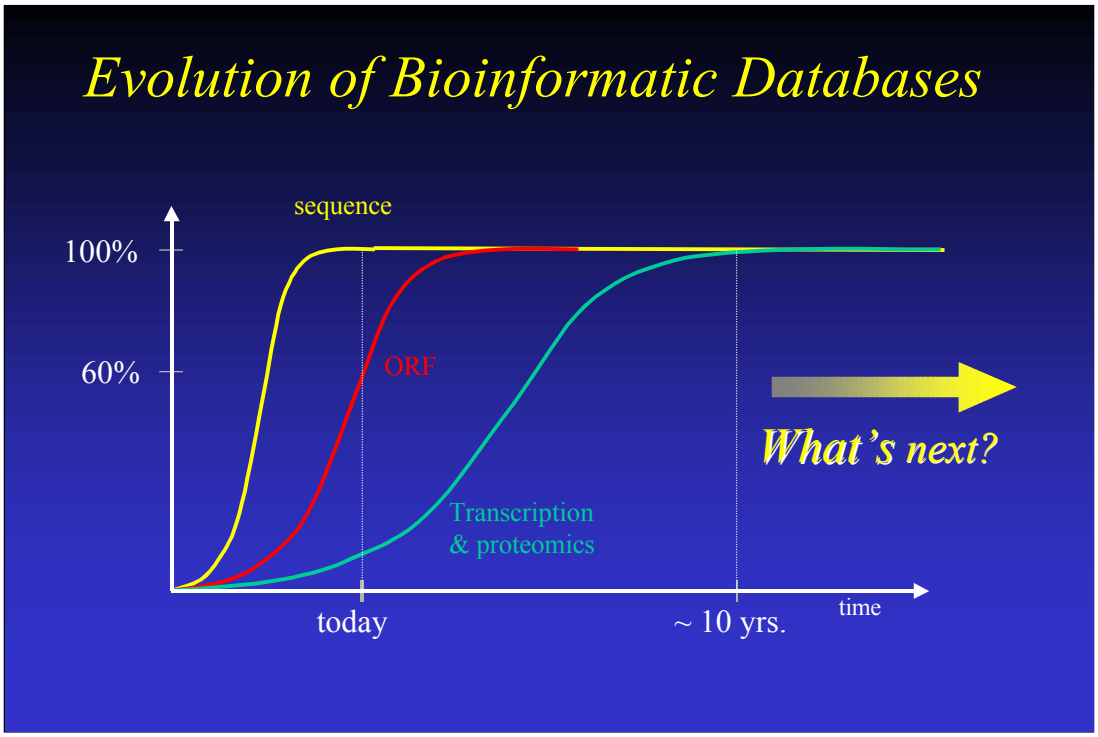
INTRODUCTION

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

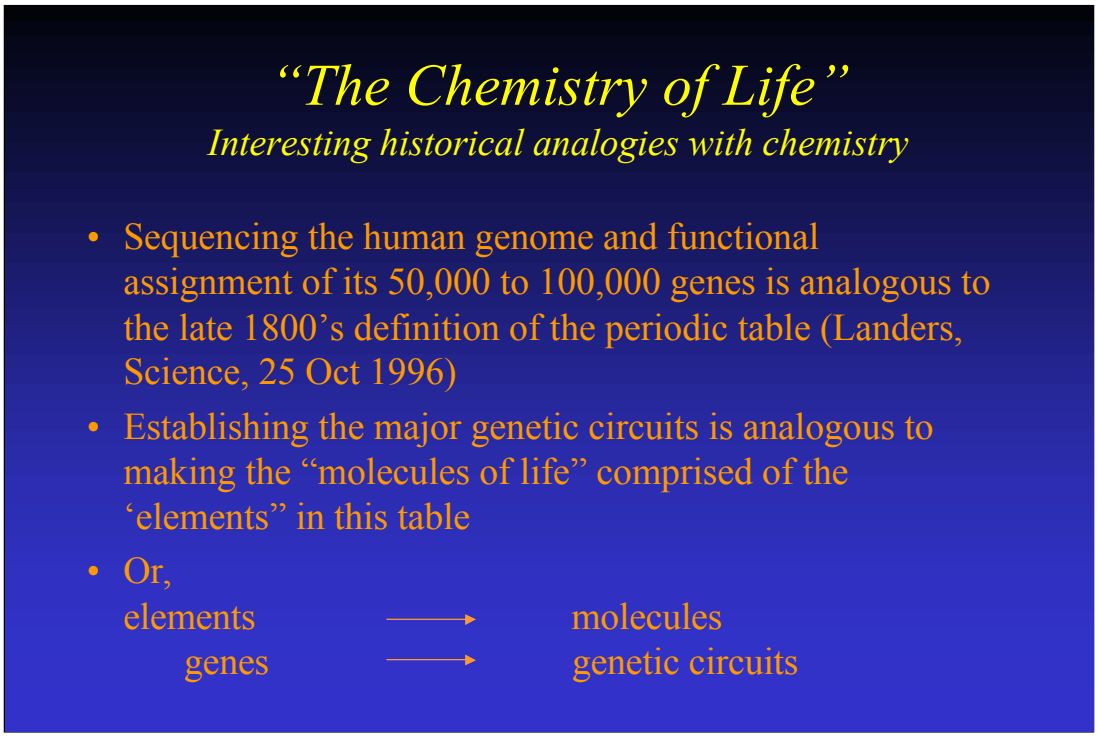
This slide provides just a crude perspective of where we stand today in terms of the evolution of bioinformatic databases and scientific information.

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.

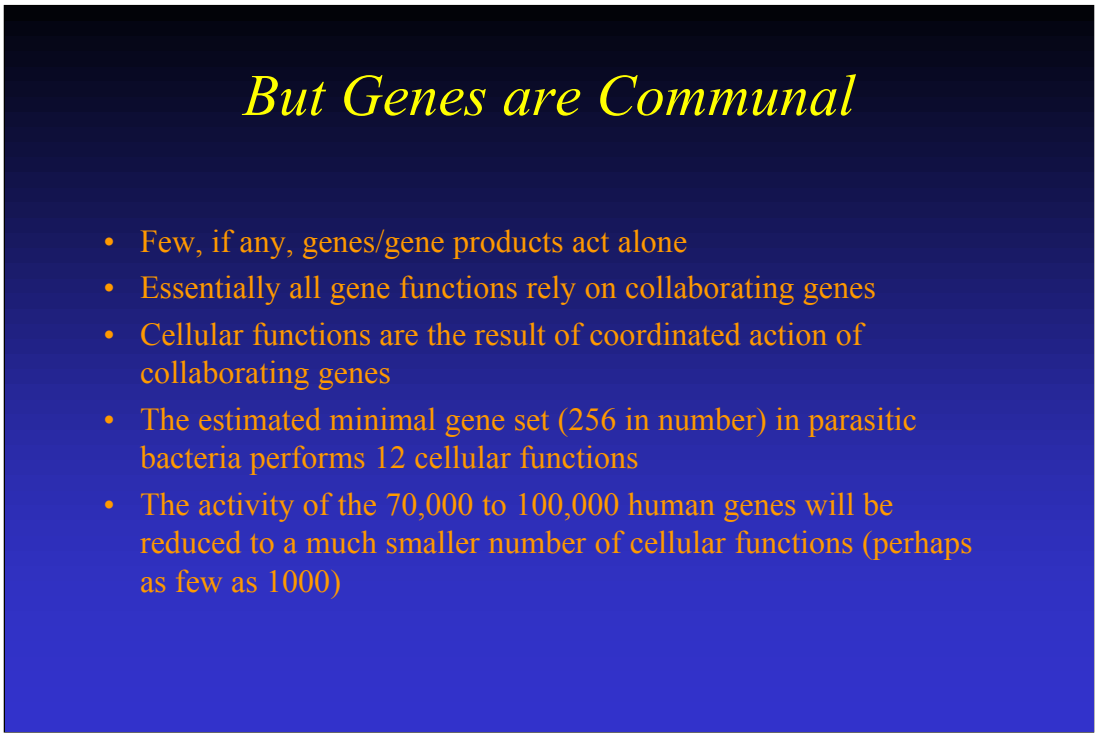
If we look ahead into the not too distant future we can expect to have enormous amounts of information pertaining to the content, structure, and expression of the genotype.

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?



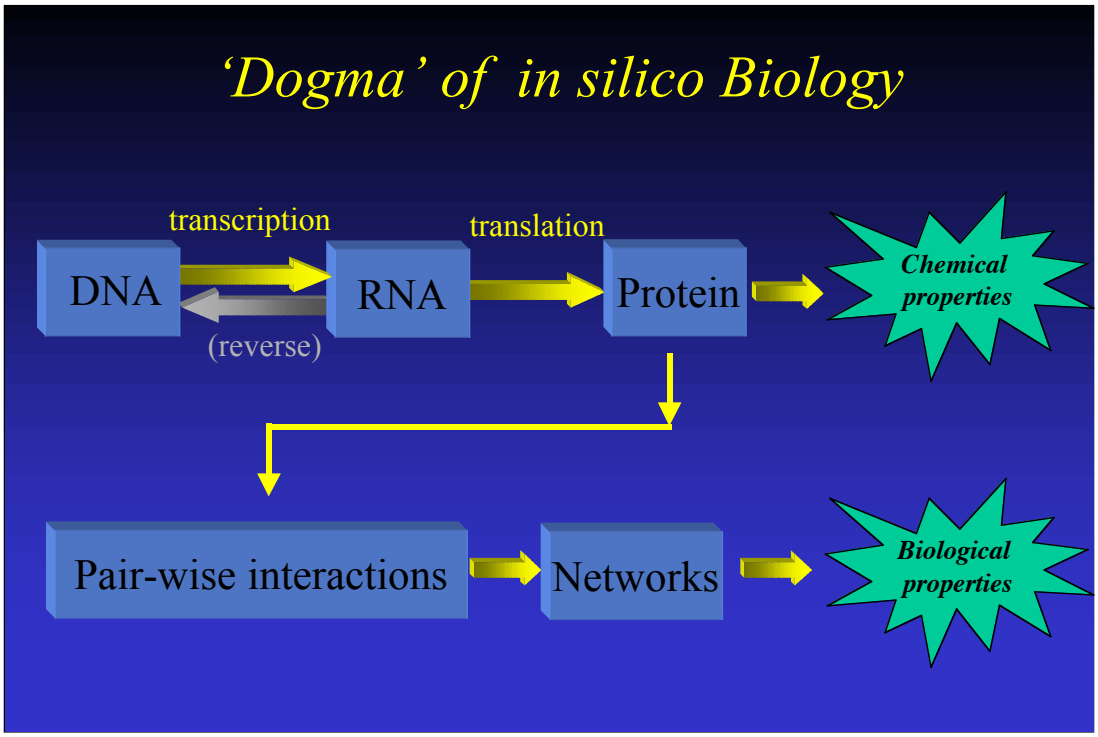
THE LANDER ANALOGY

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.

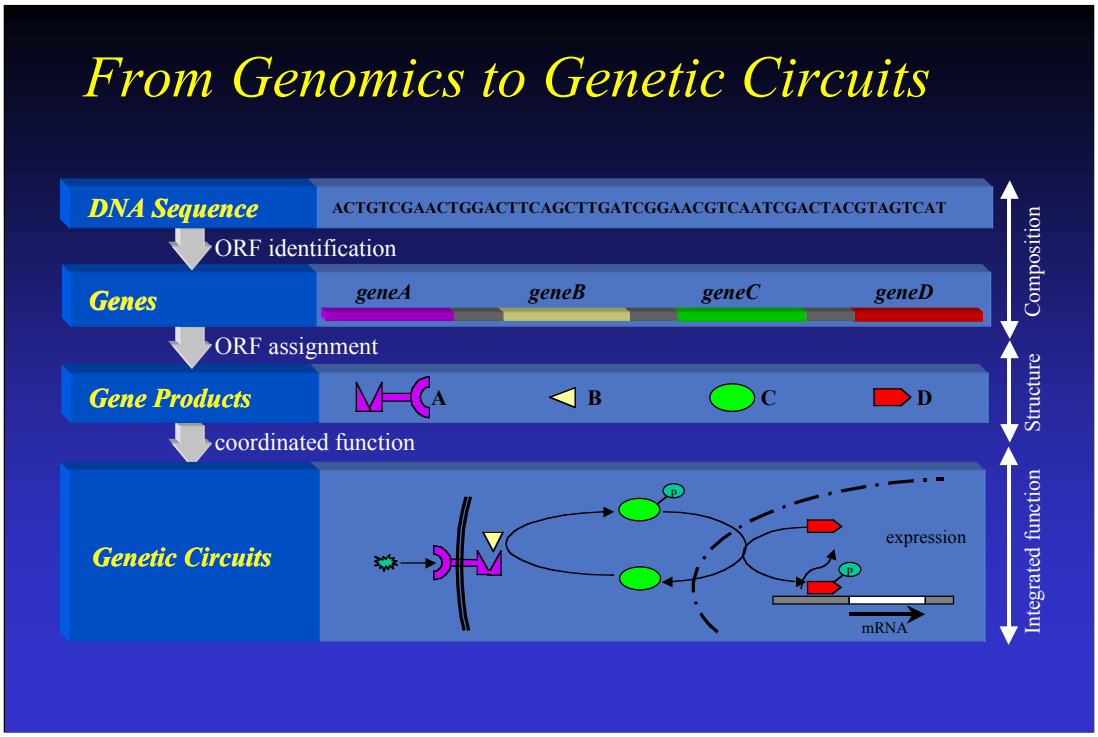


THE DOGMA OF IN SILICO BIOLOGY

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

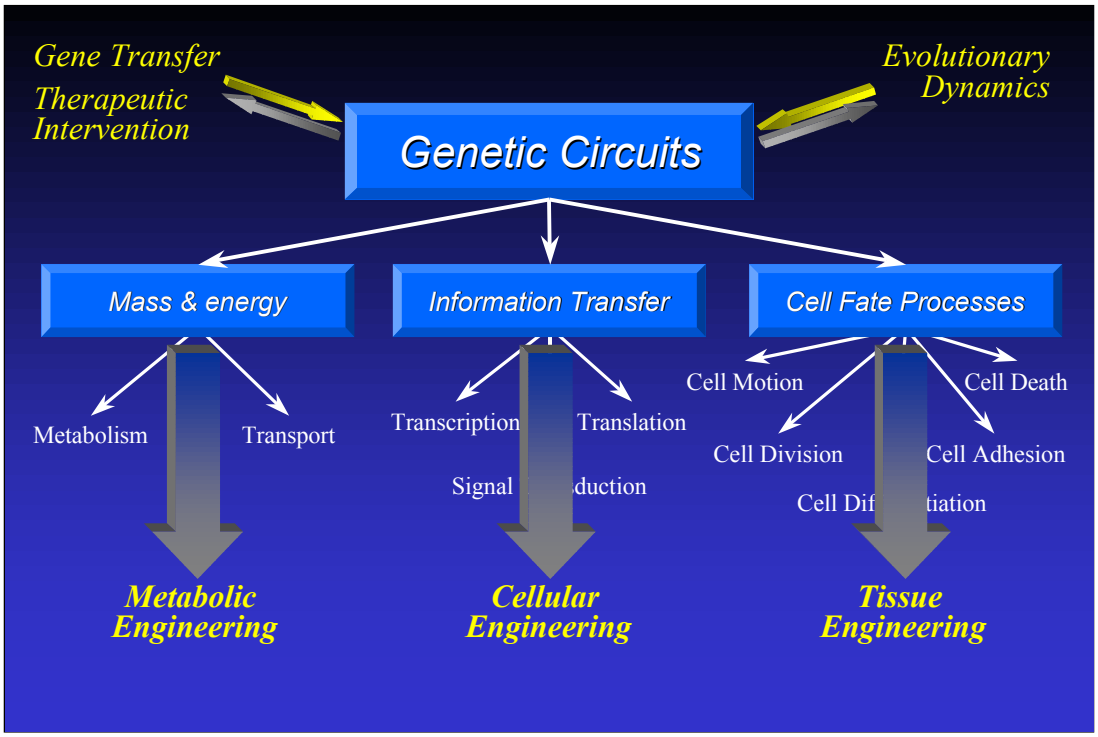
The relationship between the genotype and the phenotype is complex, highly non-linear and cannot be predicted from simply cataloging and assigning gene functions to genes found in a genome.

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

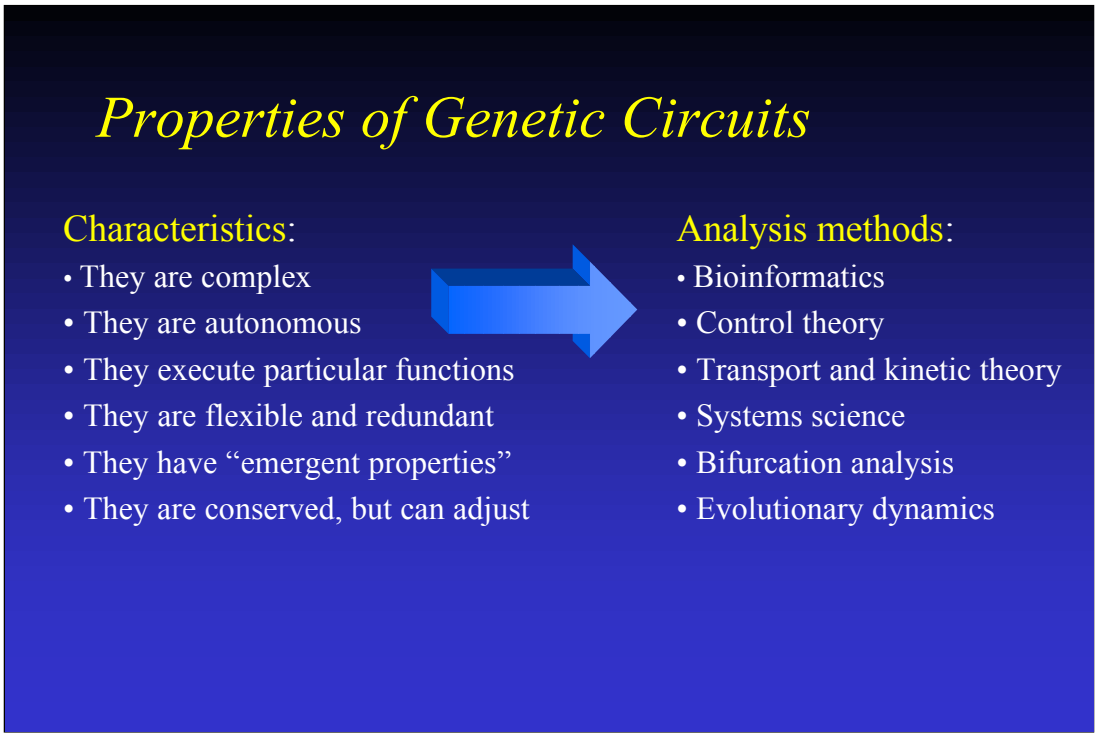
Although we do not know all the genetic circuits found on a genome we can still begin to classify them. A coarse grained classification is illustrated in this slide:

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.

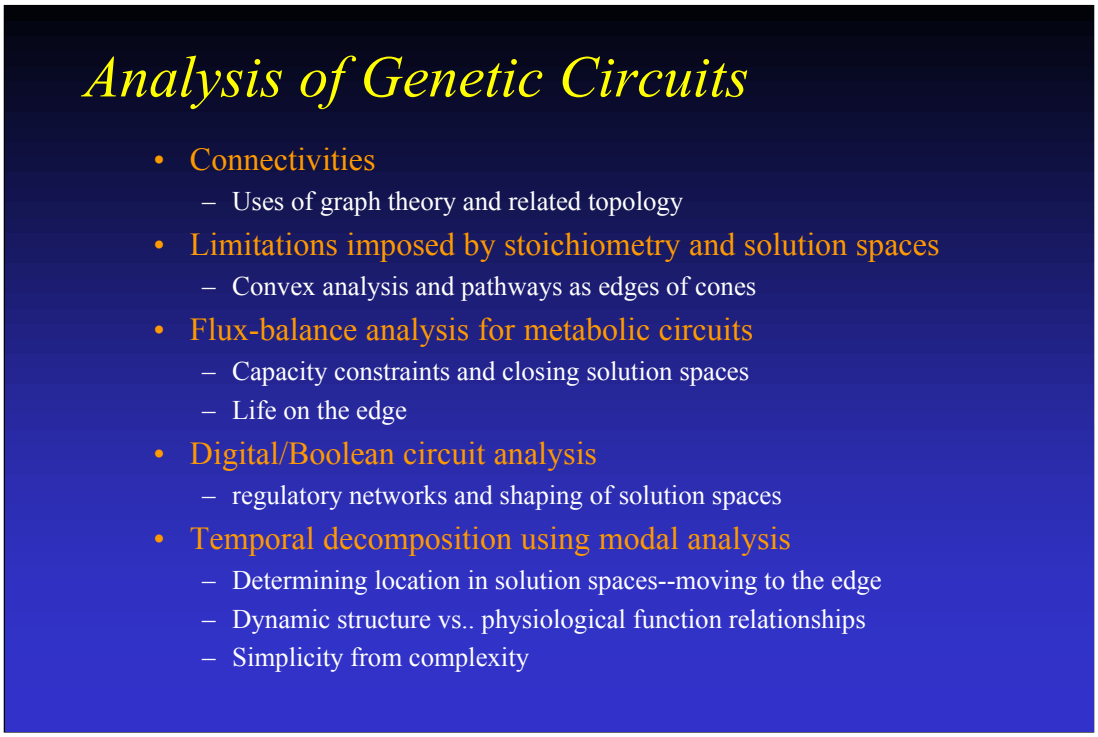
In general they are complex with many components which offer a degree of flexibility in functioning and in evolving. Once genes are expressed, the coordinated function of the gene products is autonomous, and embedded within these built in controls are the capabilities to perform creative functions.

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.

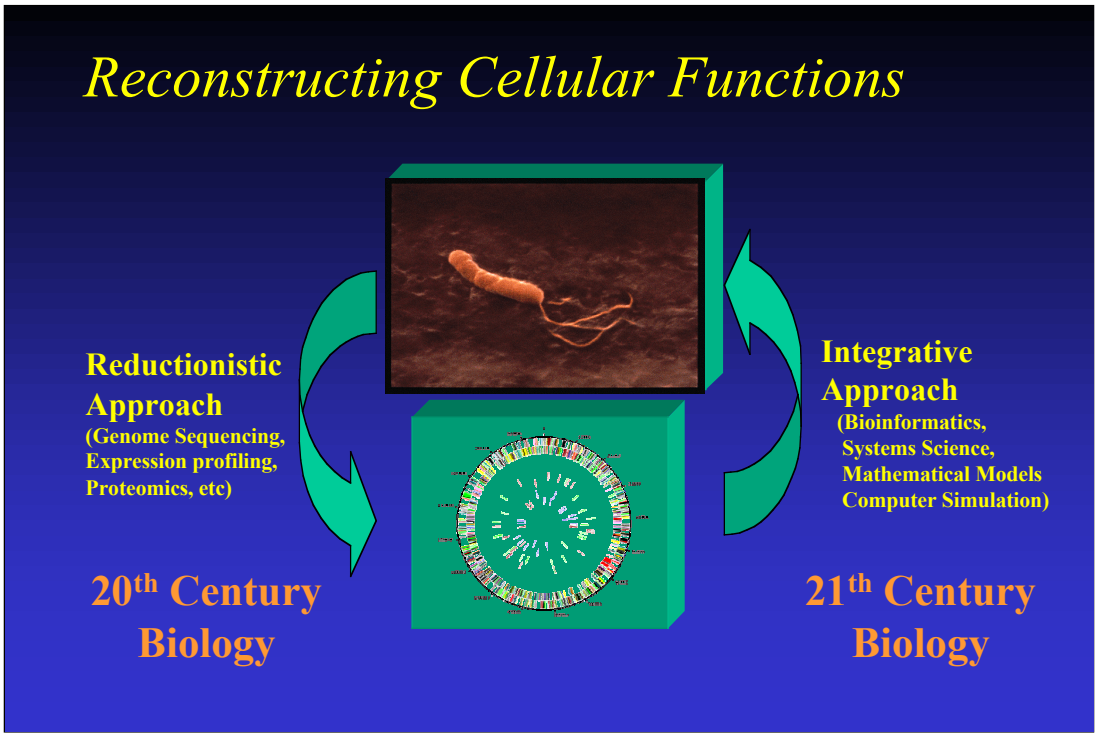
Genetic Circuits; a different point of view

- 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

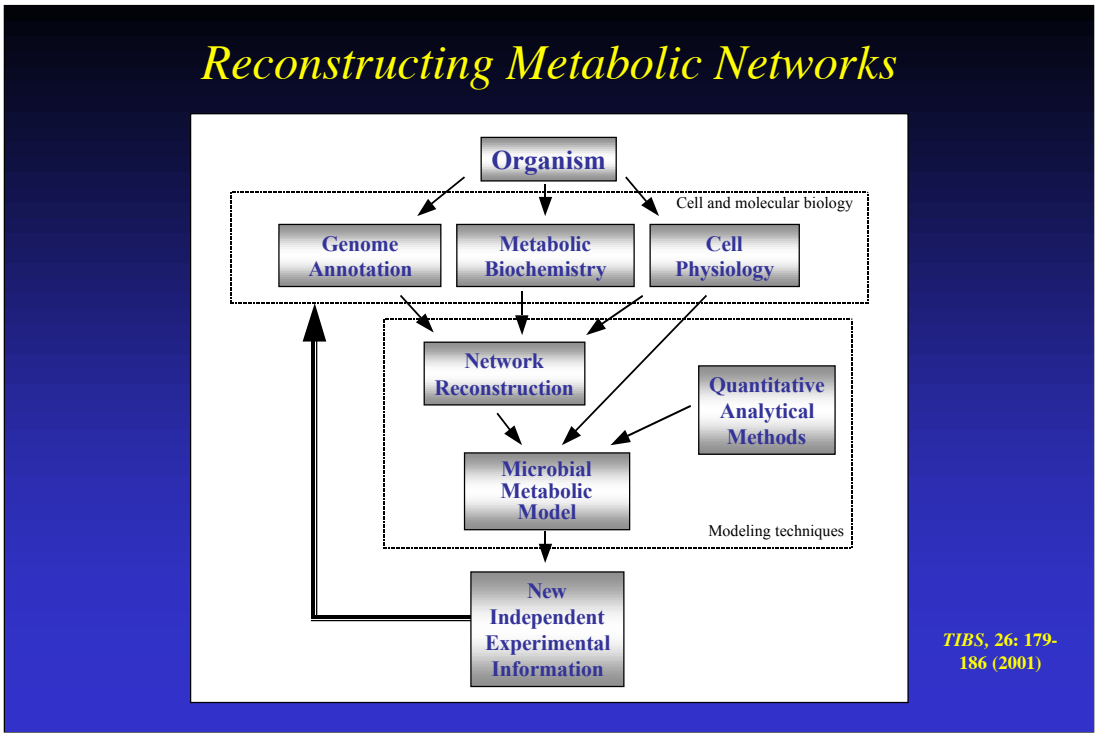
NEW CURRICULA

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.

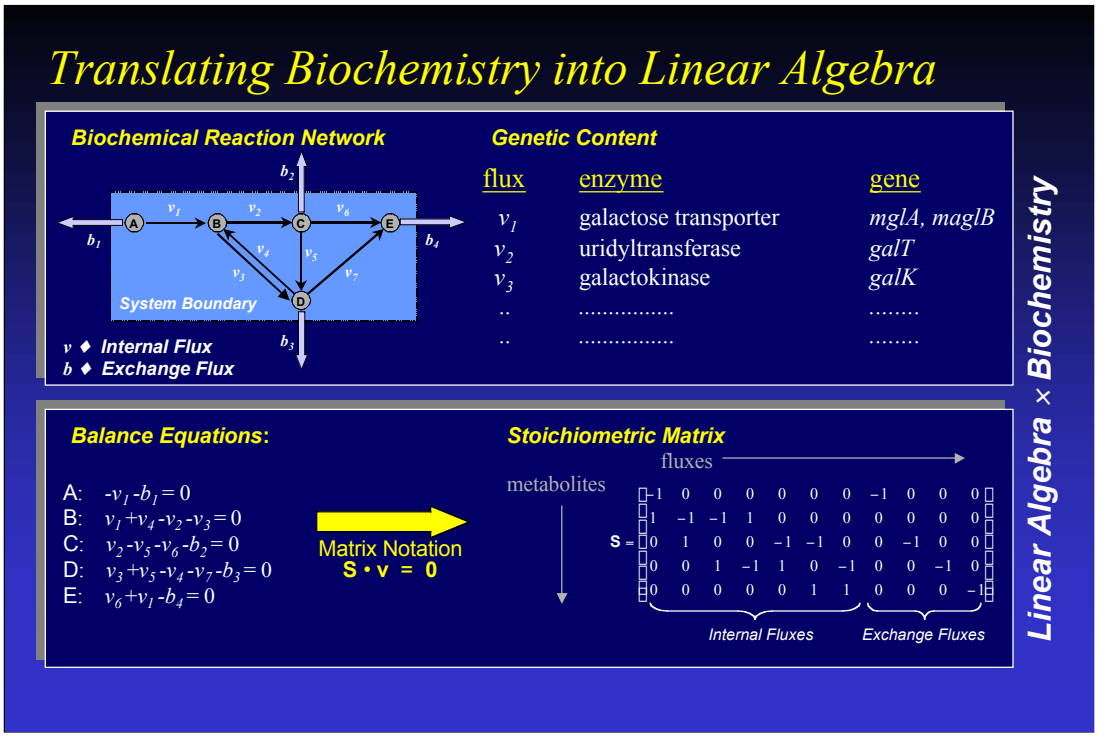


THE RECONSTRUCTION PROCESS

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.

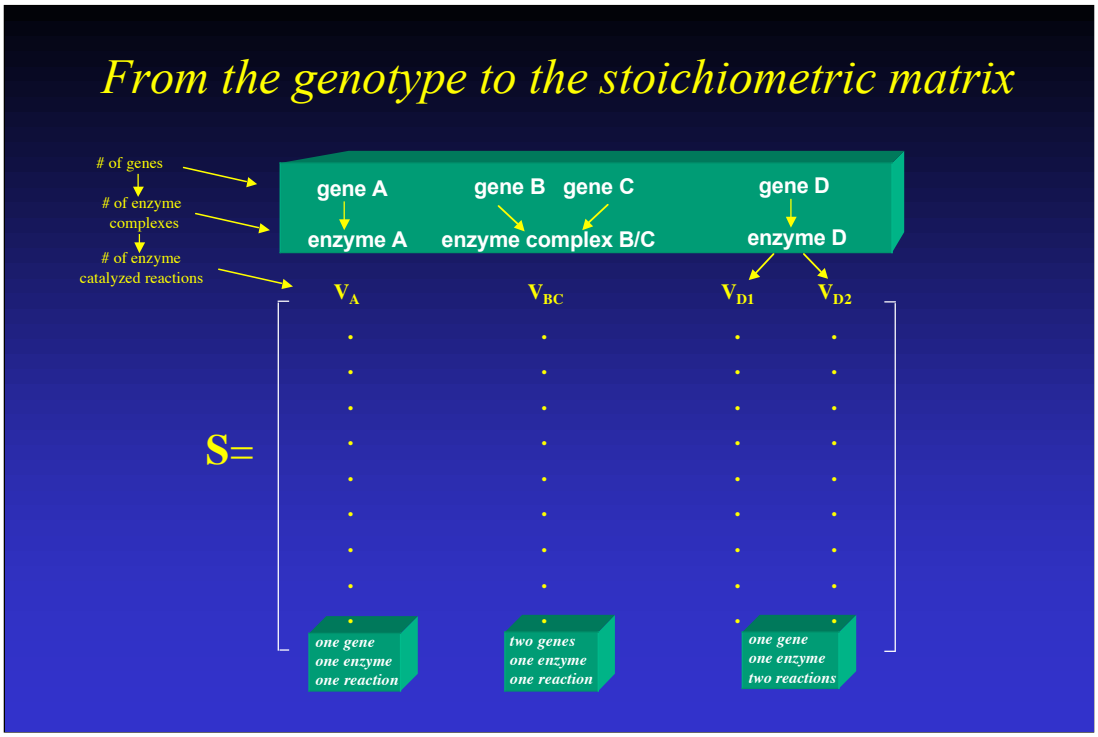
At present this process cannot be automated, and in particular much human input and interpretation is required in reading all the pertinent literature on known biochemical activity reported for the organism in question and to interpret its physiological functions.

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.



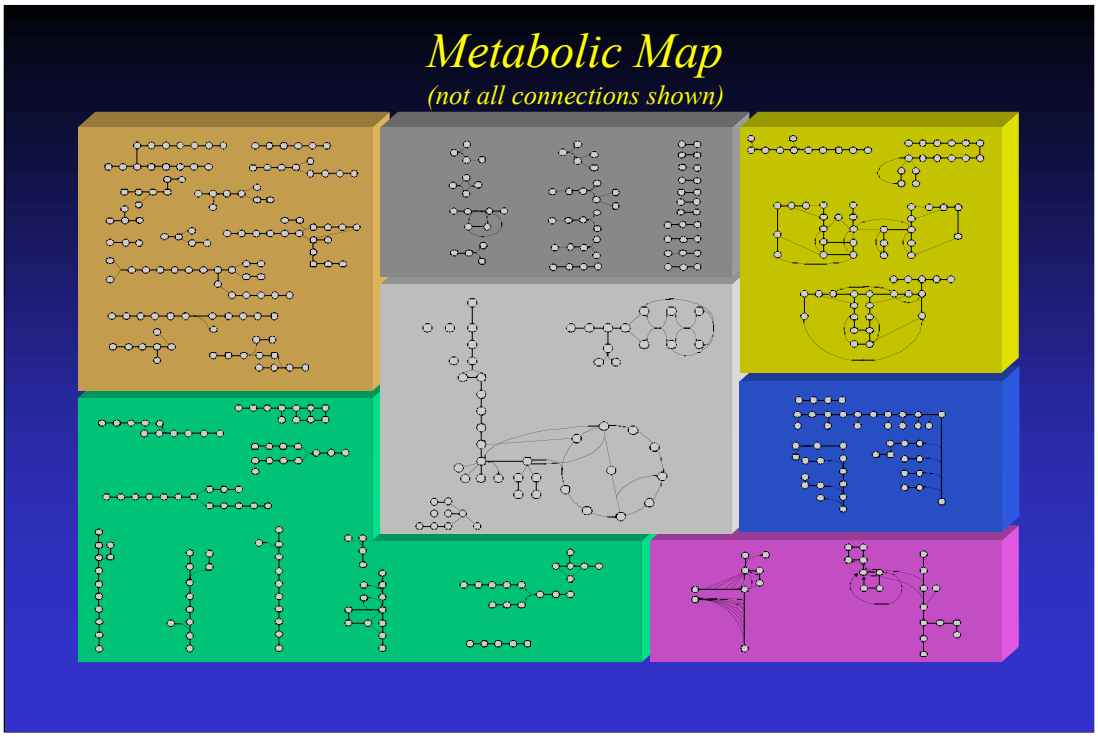
THE NUMBER OF REACTIONS IN A METABOLIC GENOTYPE IS NOT THE SAME AS THE NUMBER OF GENES IN THE GENOTYPE

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.

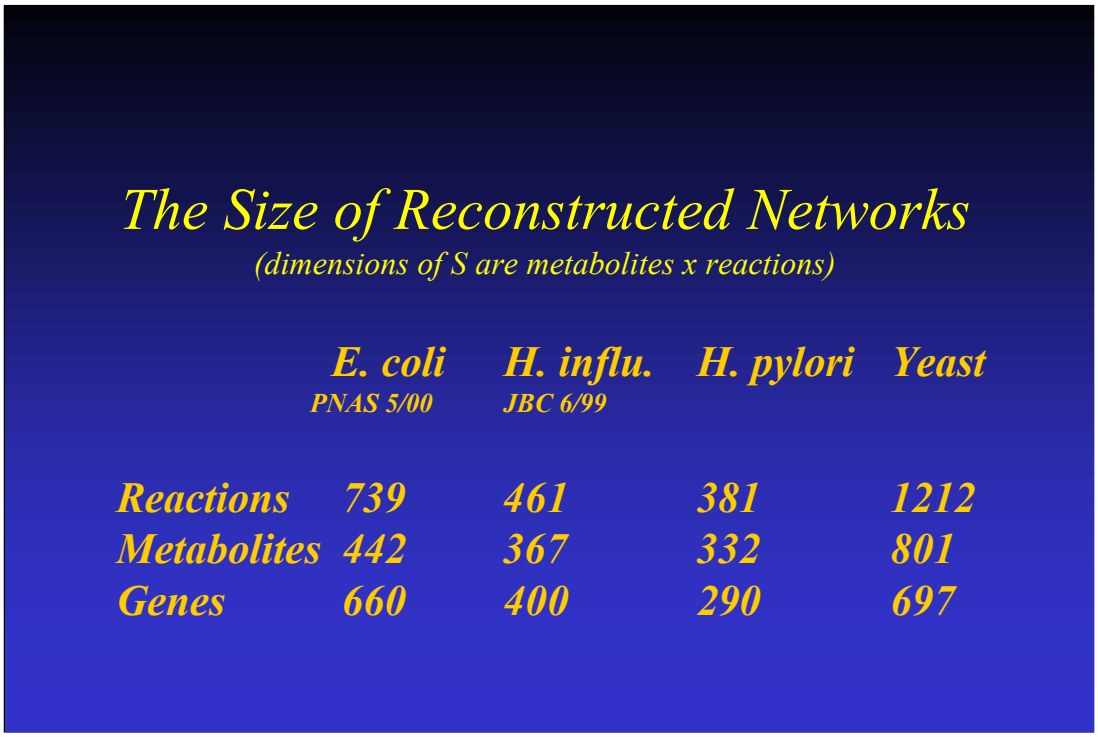


THE METABOLIC MAP REPRESENTATION OF THE *ESCHERICHIA COLI* K-12 METABOLIC GENOTYPE

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.

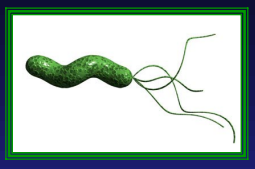


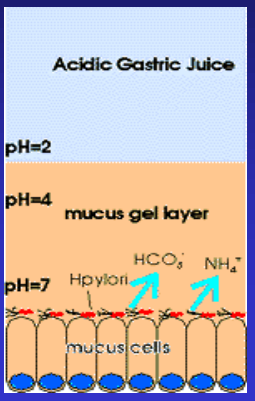
DIMENSIONS OF S

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.

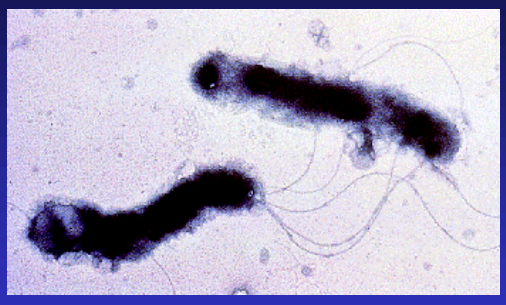
The stomach is protected from its own gastric juice by a thick layer of mucus that covers the stomach lining. *Helicobacter pylori* takes advantage of this protection by living in the mucus lining.

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*

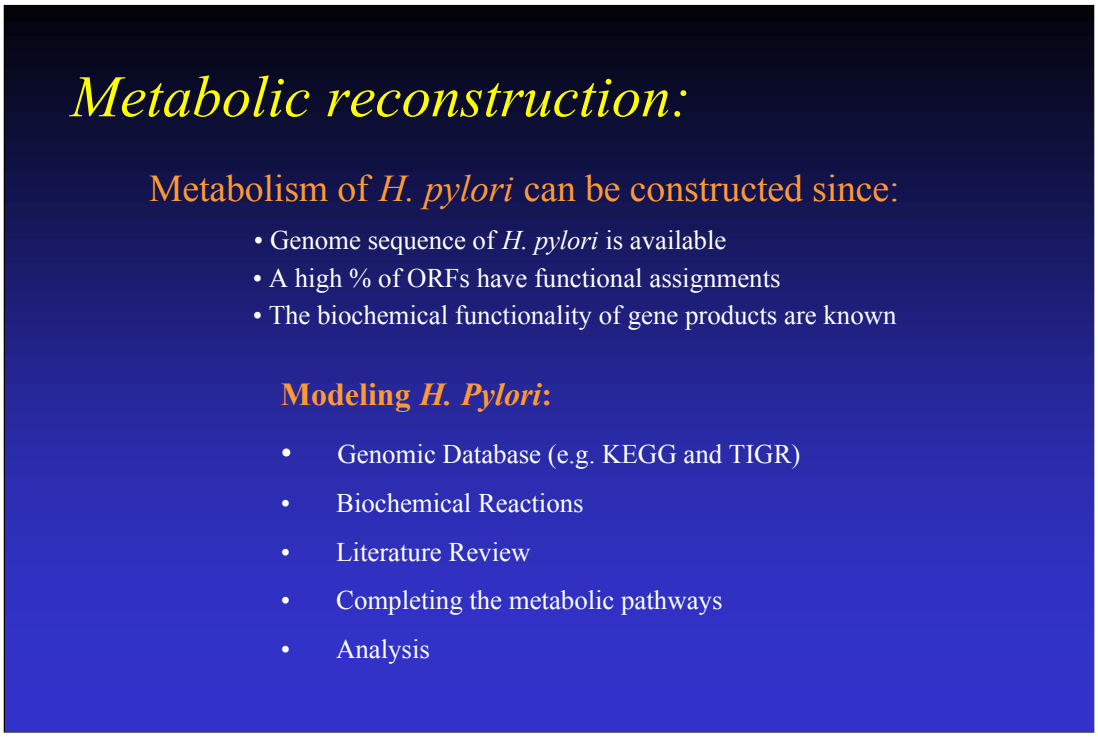
- Immune response cannot reach the infection through stomach lining
- Immune response buildup degrades stomach lining cells (superoxide radicals) – gastritis or peptic ulcers can result within days
- *H. pylori* feeds on nutrients sent to reinforce the white cells
- Carried by >50% of world's population, 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.

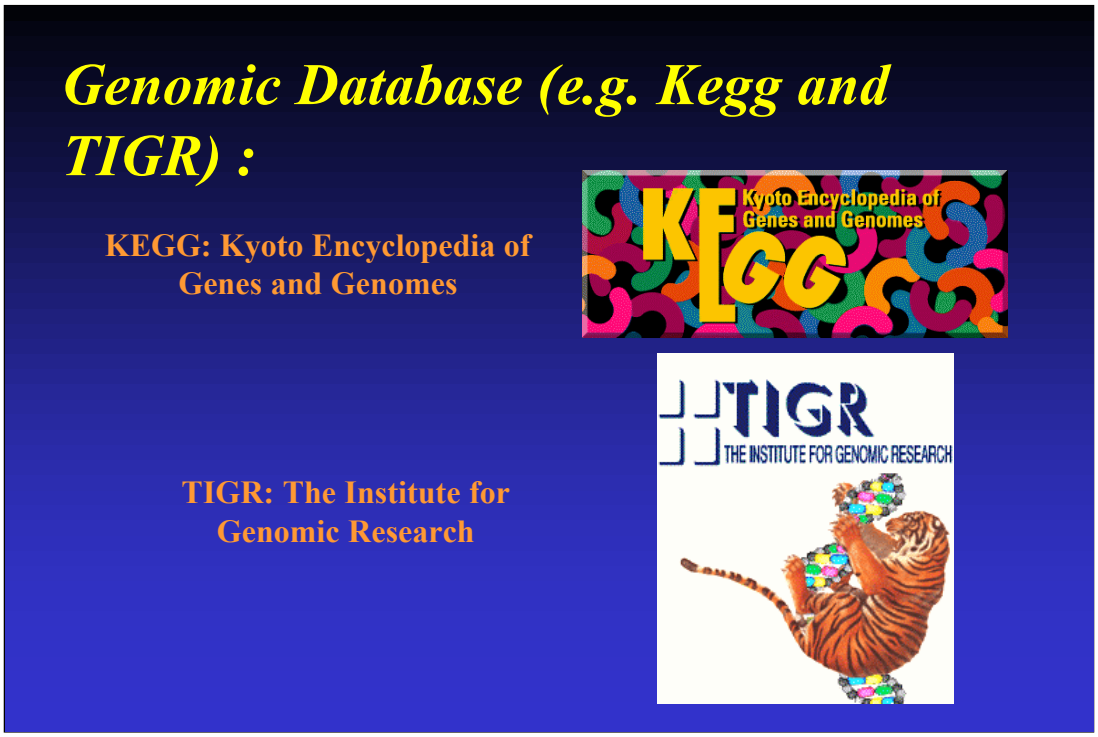
In developing countries most adults are infected. Acquisition occurs in about 10% of children per annum between the ages of 2 and 8 years so that most are infected by their teens. It is evident from careful surveys that the majority of persons in the world are infected with *H.. pylori*. (from www.hpylori.com)



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:

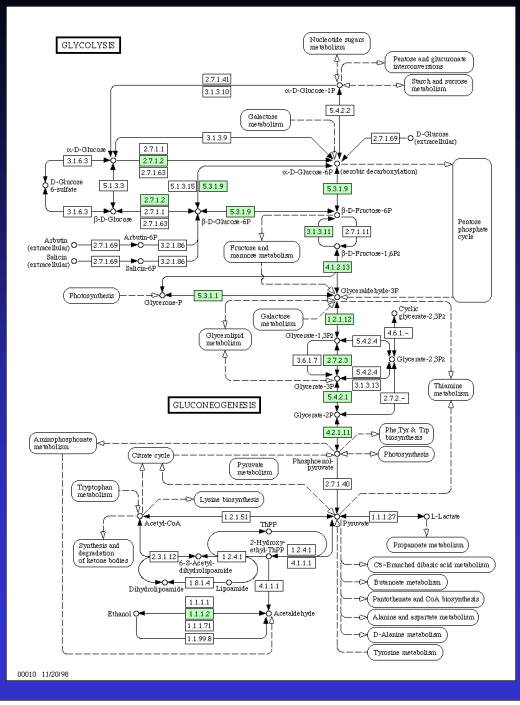
KEGG: www. kegg.com

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:

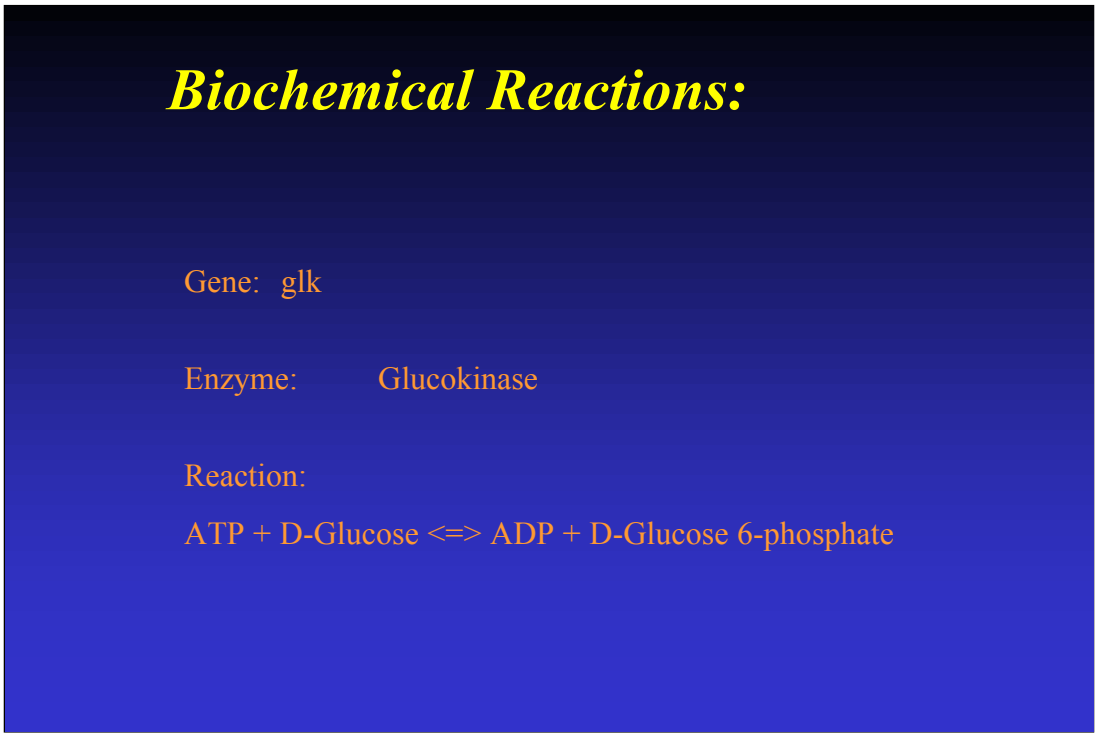
- Genes
- Gene Products
- Metabolic Pathways



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.

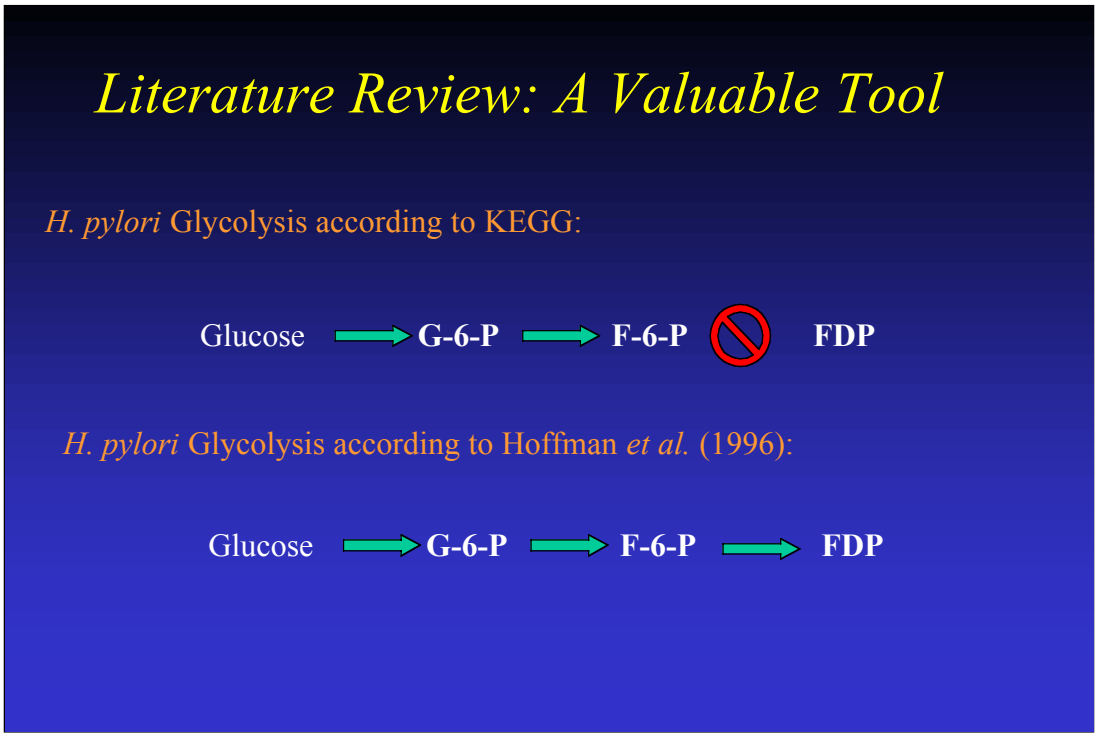
KEGG uses the same maps for many organisms, so not all of the pathways shown in this map are actually available to *H. pylori*. Some are for *E. coli*, for example. The genes actually found in *H. pylori*, according to this map, are the ones which are highlighted in green.



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.

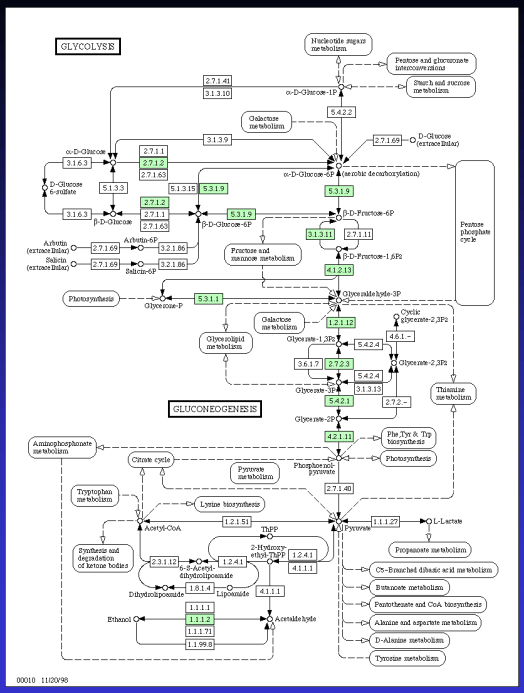


THE NEED FOR USING
PHYSIOLOGY AND INFERRING REACTIONS

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



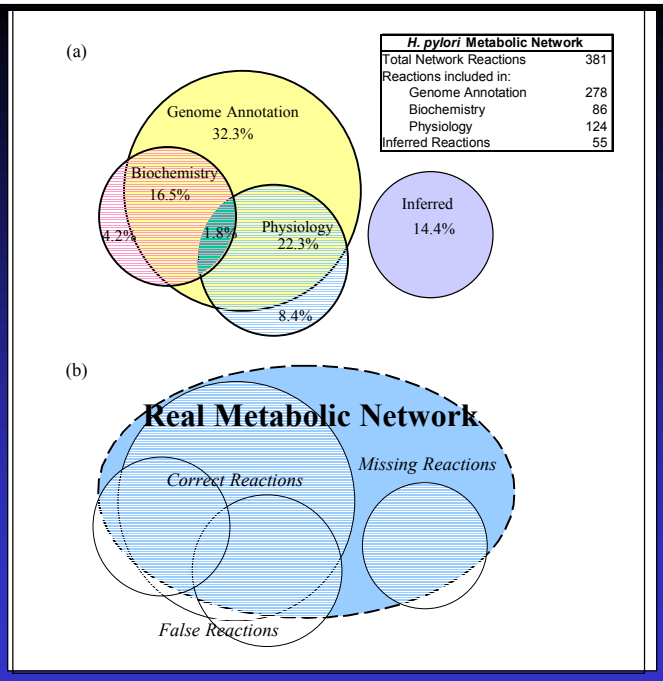
THE NEED FOR USING PHYSIOLOGY AND INFERRING REACTIONS, CONT'D

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^+

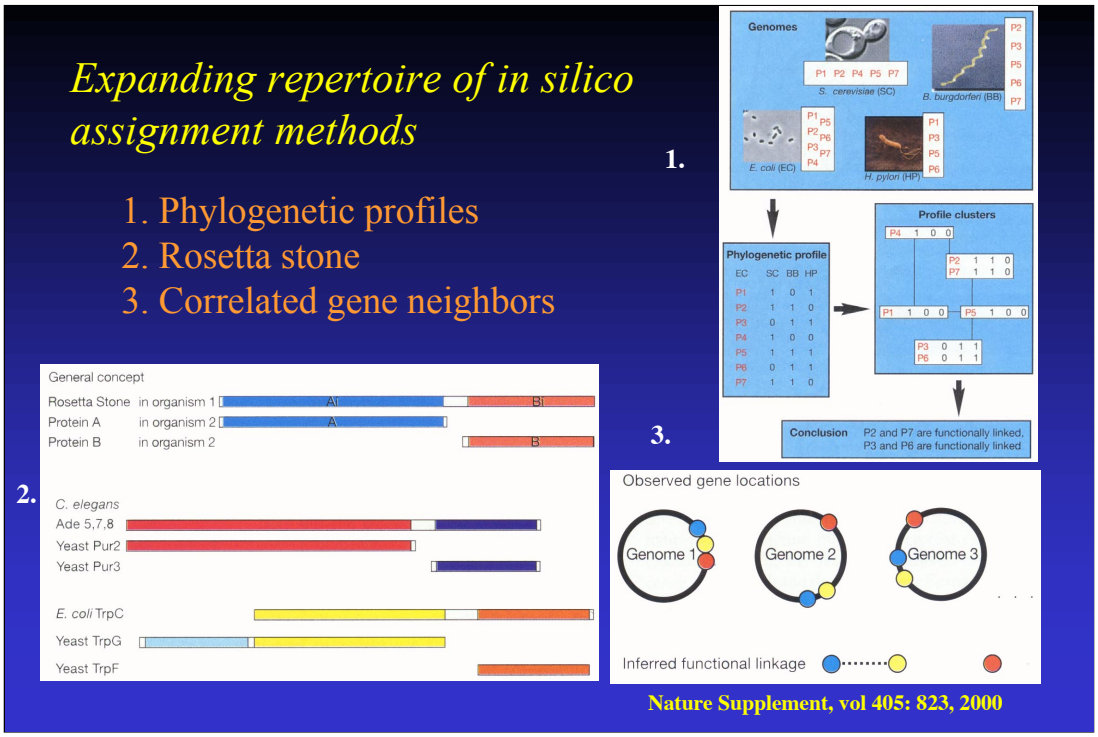
Computational Verification:

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.

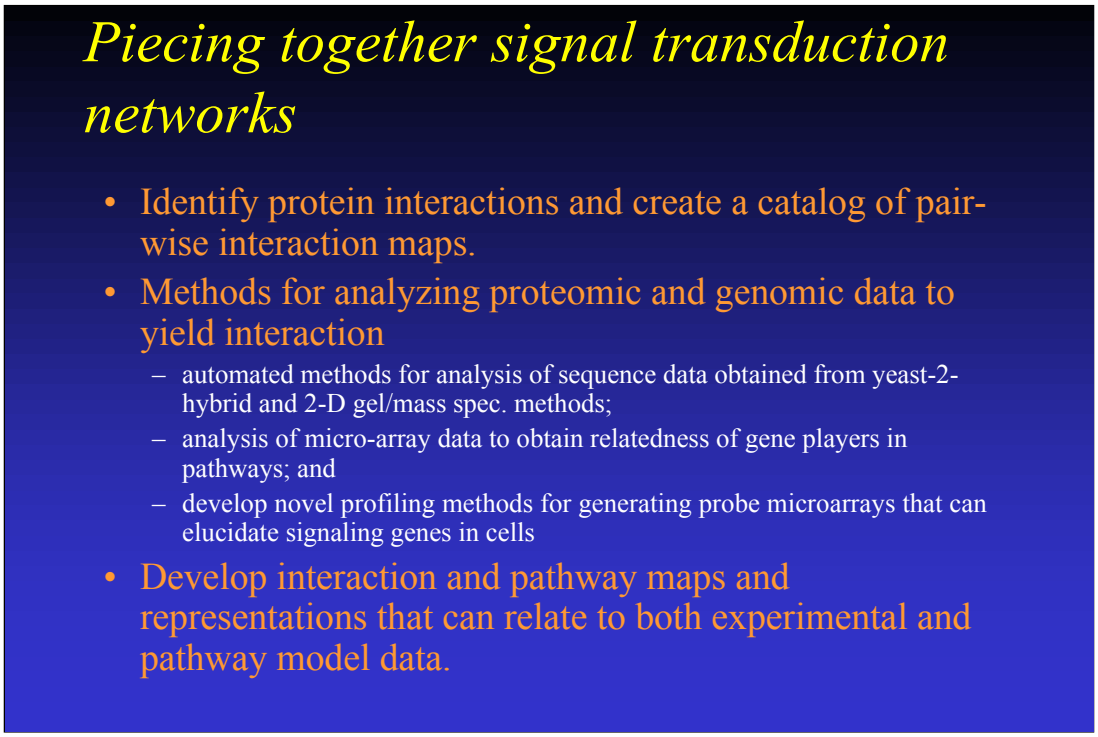


NEW METHODS

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

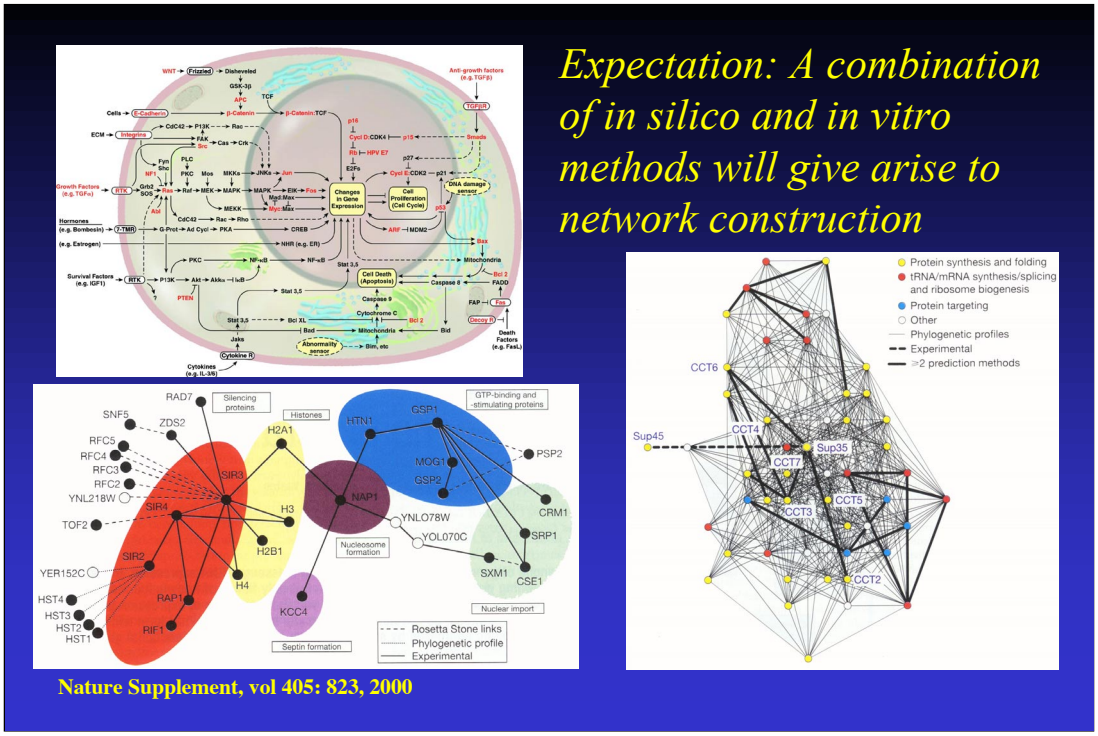
- Make mutants and experimentally determine phenotype
- Expression arrays and cluster analysis
- Computational approach based on co-evolution of protein and analysis of fusion protein (Rosetta Stone)
- Protein-protein interaction maps



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.

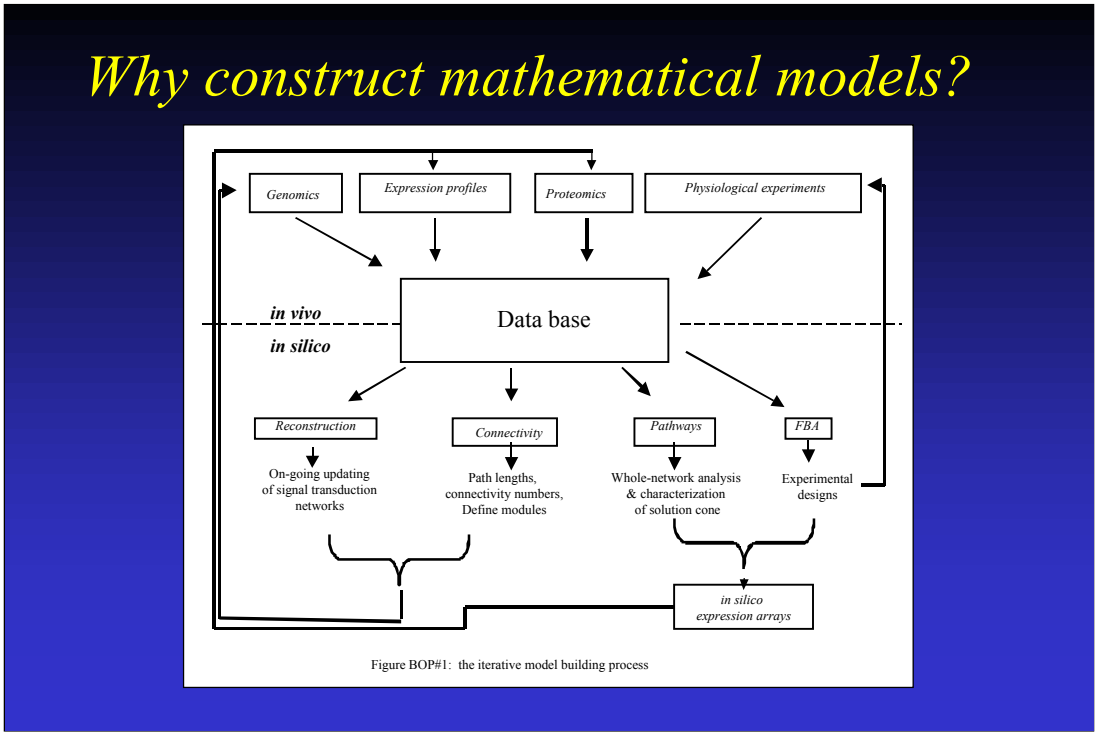
Algorithmic methods include: a) extensive schemes to analyze genomic and proteomic data, b) a high throughput pipeline for sequence comparisons across species and c) validation methods to compare diverse sources of data pertaining to specific molecular interactions. Finally, pair-wise interaction data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



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).