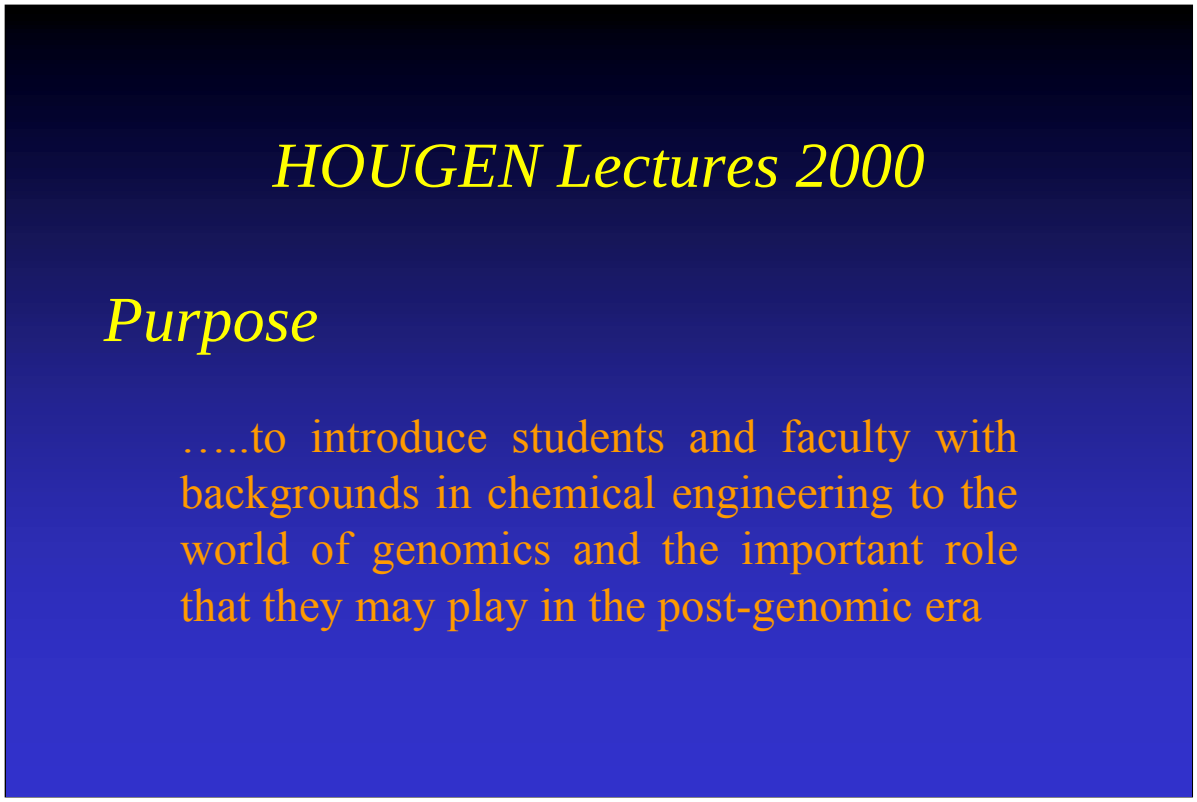
THE HOUGEN 2000 LECTURES

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

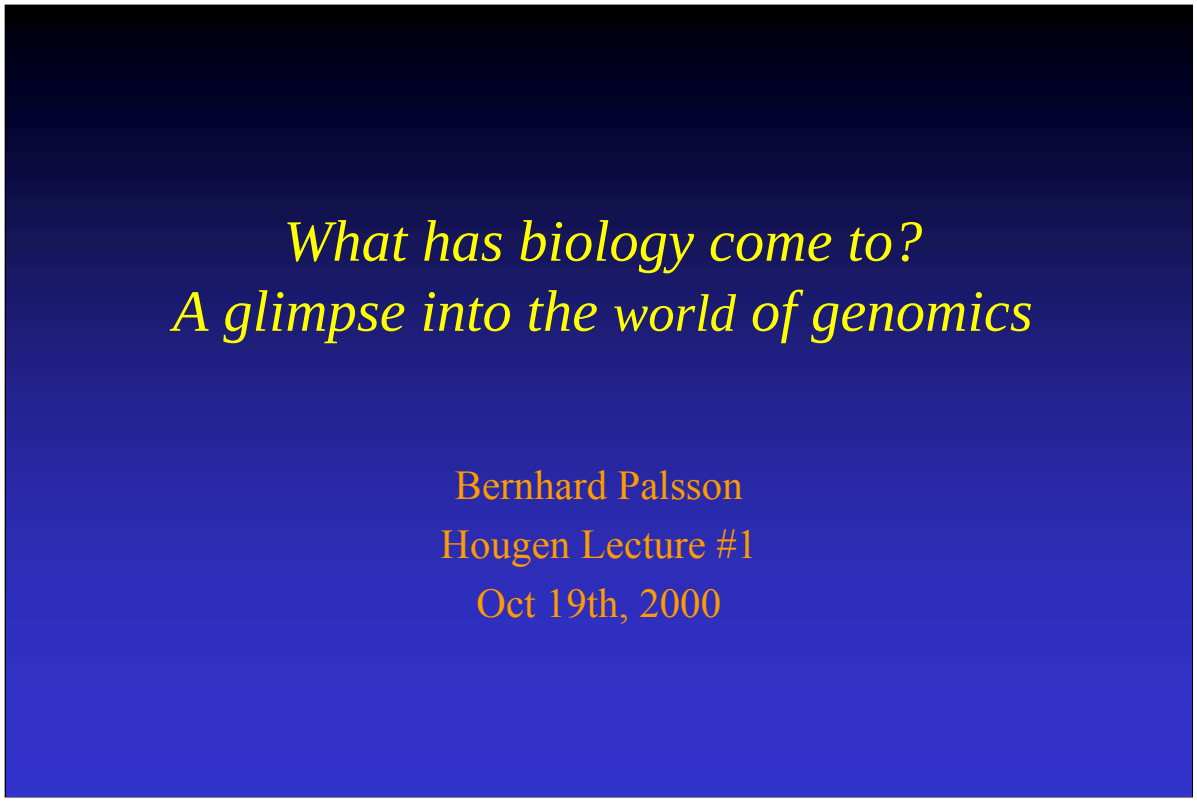
The Hougen visiting professorship was established to enable scholarly and free exchange amongst the visitor, the faculty, and students of chemical engineering. The Wisconsin department has always placed emphasis on fundamental issues and problems that have long-term consequences. It is the opinion of 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

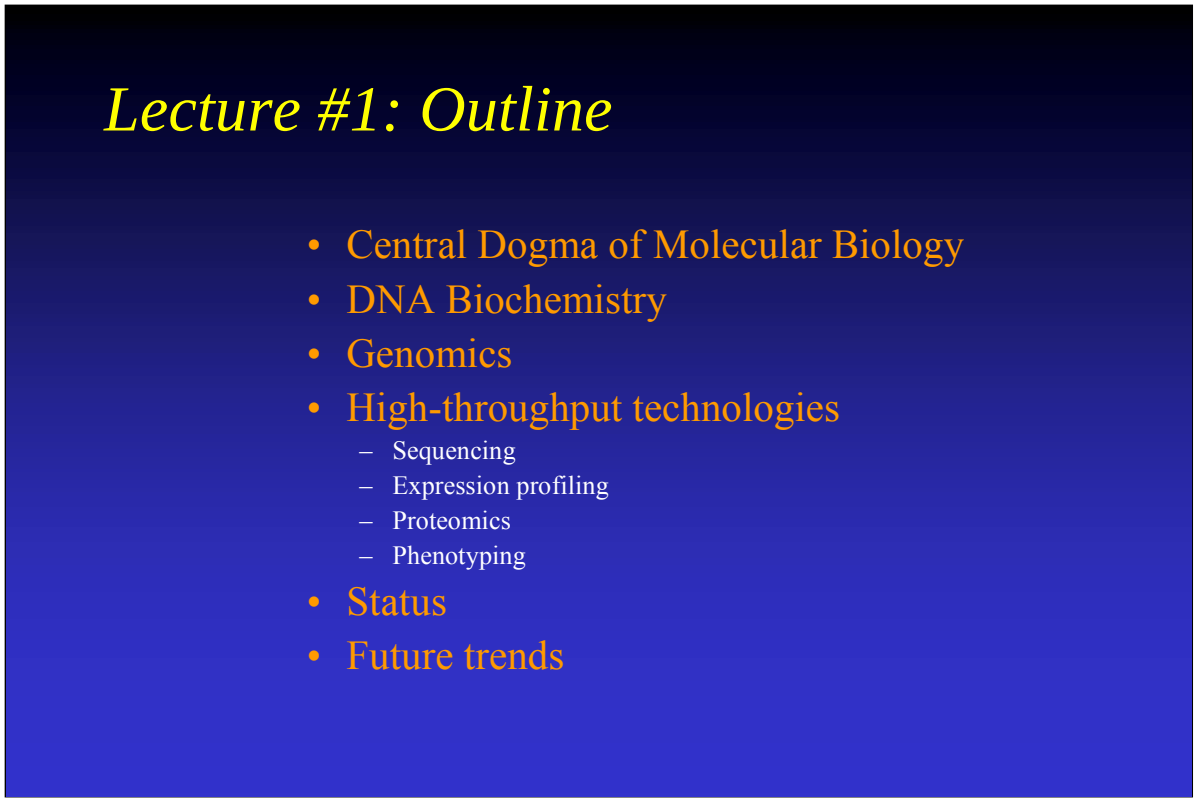
The lectures will be delivered in a very casual setting over lunch on Thursdays. The tentative schedule of topics is given above. We expect that this outline will evolve as the lectures proceed in response to the interest and the expertise of the audience that will attend. An extra time slot on December 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.

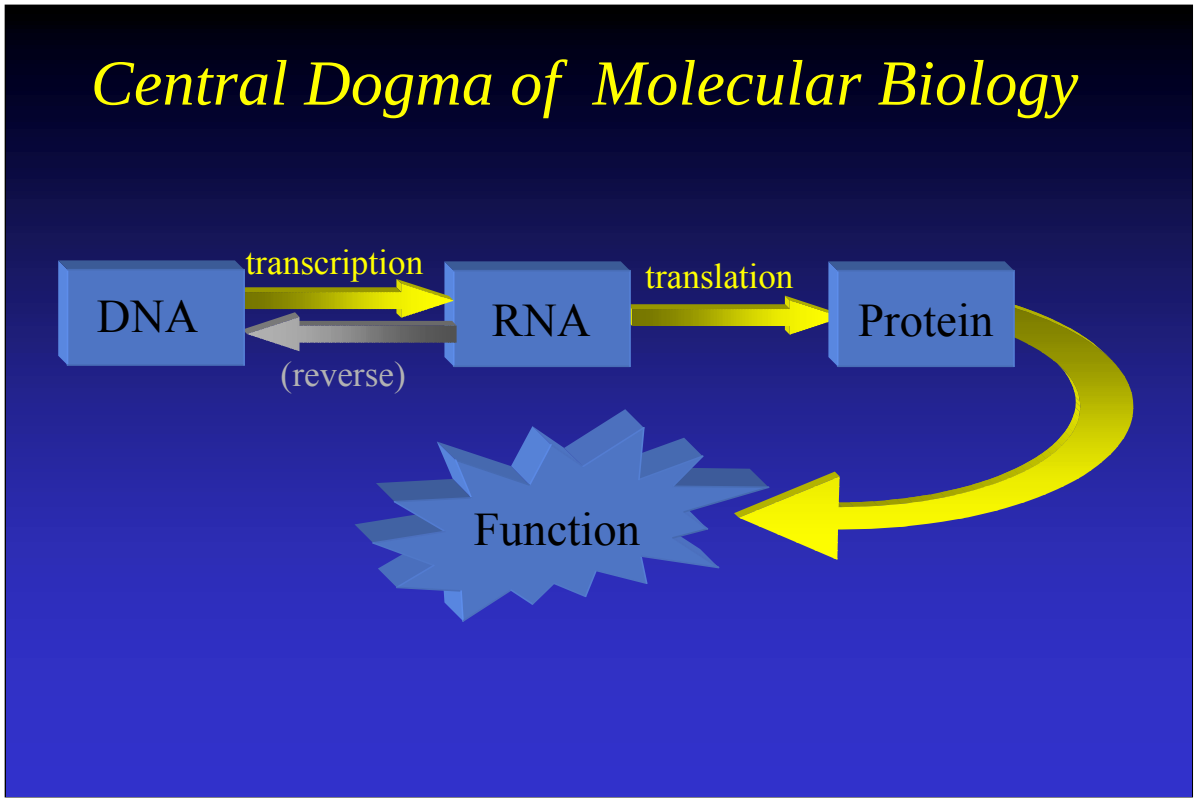


LECTURE #1

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

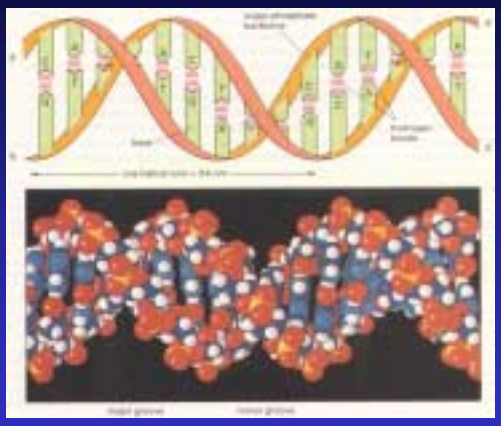
This schema illustrates the central dogma of molecular biology as it was developed about 40 years ago. The DNA, a long thread like molecule of a specific base-pair sequence, carries the inherited information. Short segments of the DNA molecule (called the open reading frames or ORFs) are transcribed 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

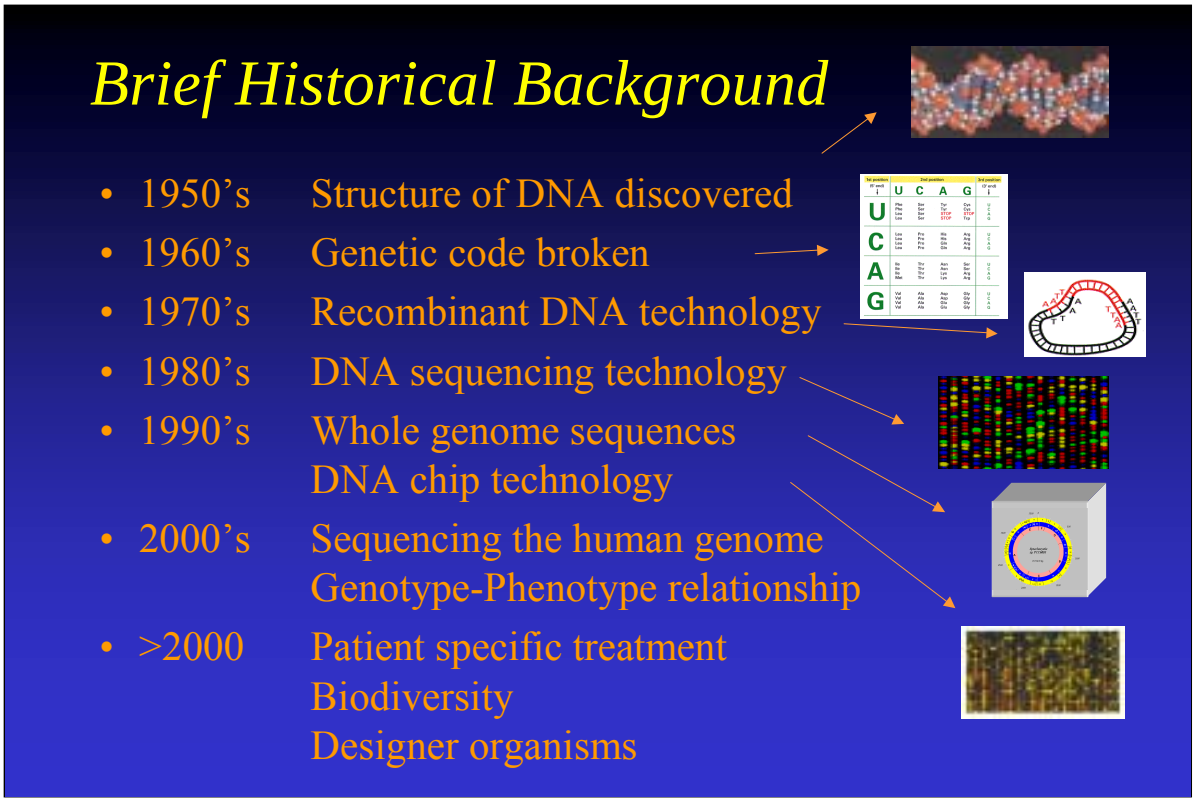
- What is DNA?
  - a linear polymer of nucleotides
  - DNA exists as a molecule of 2 anti-parallel strands that are complementary in their nucleotide sequence.



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**

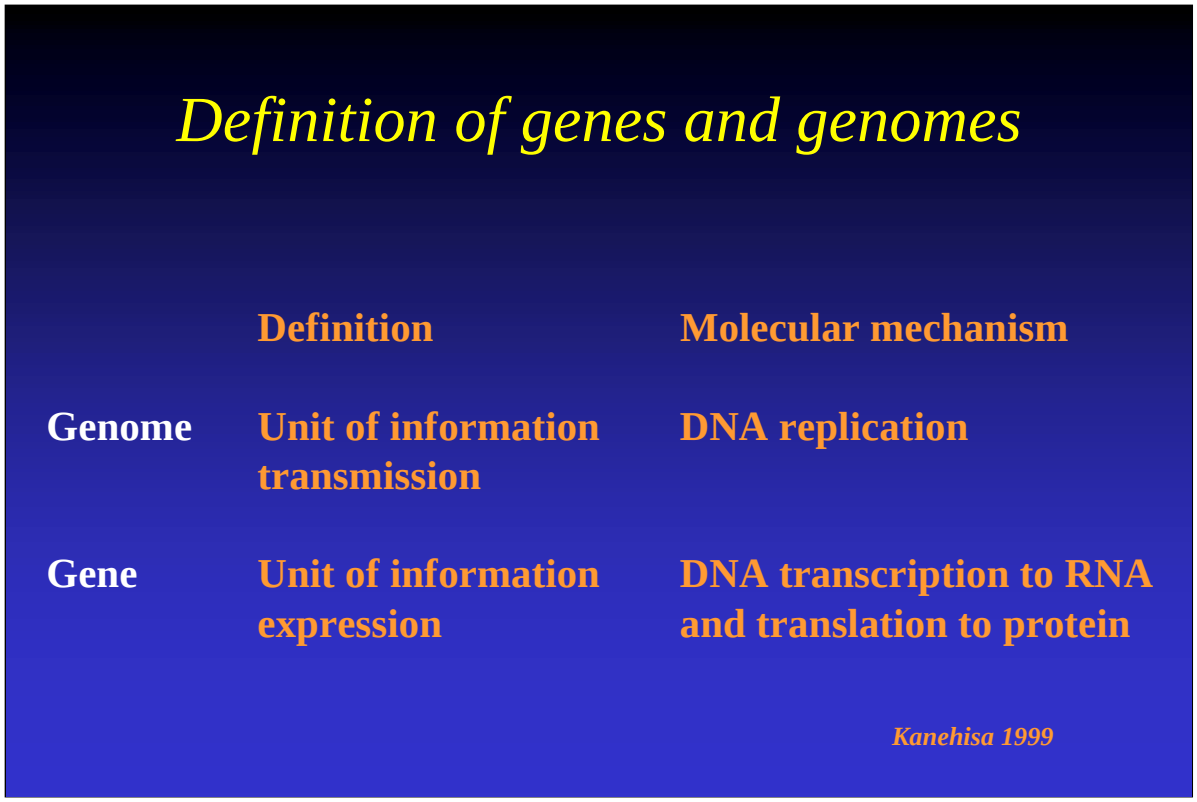
General Genomics Information:

- 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.

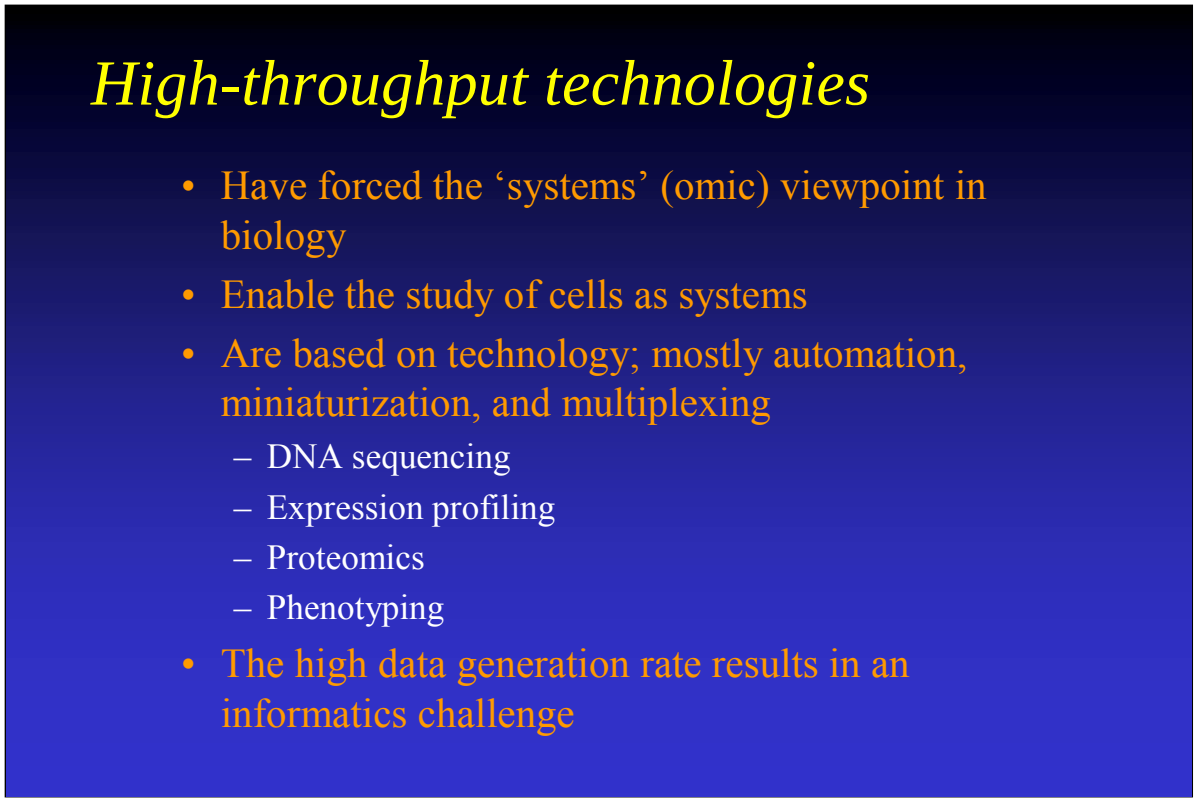
Other 'omics' words are proteome, transcriptome, metabolome, physiome, and phenome, with their obvious meanings.



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.

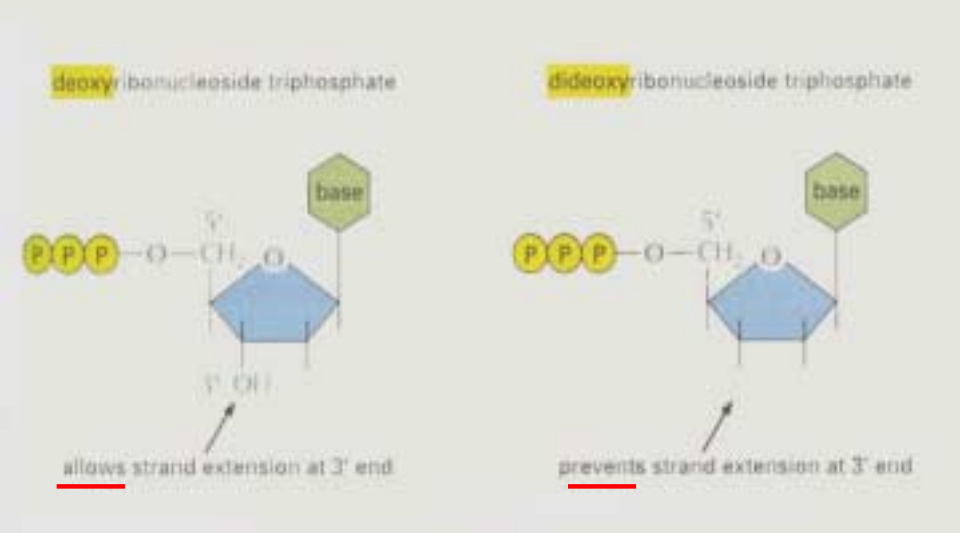
DNA Sequencing

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.

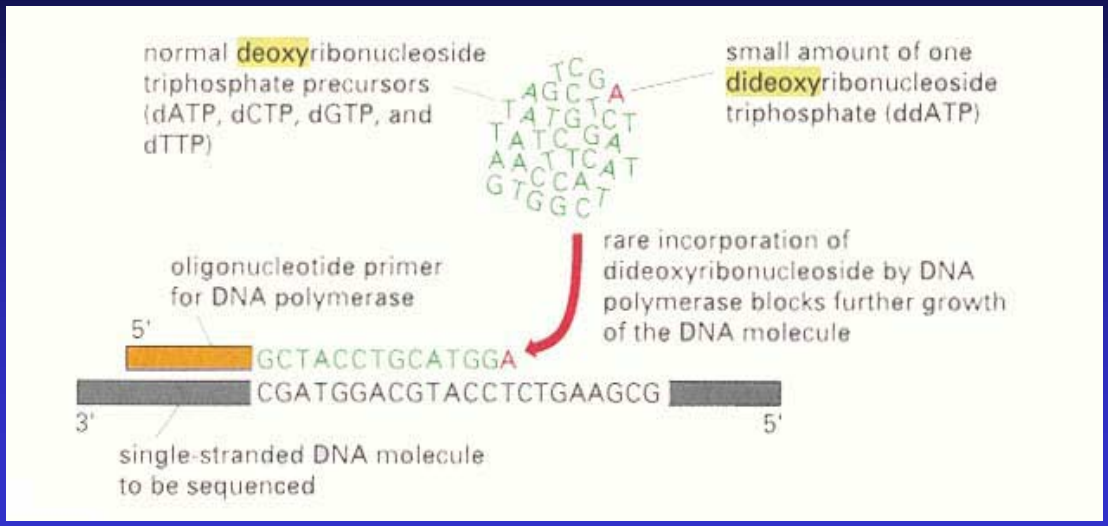
Dideoxy-nucleotides will stop DNA polymerization: they are terminators of polymerization



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



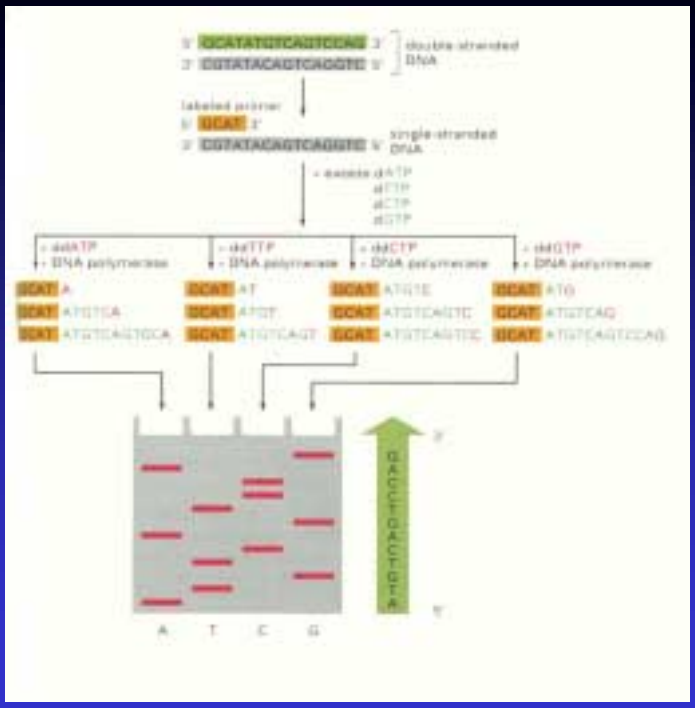
CHAIN TERMINATION

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.

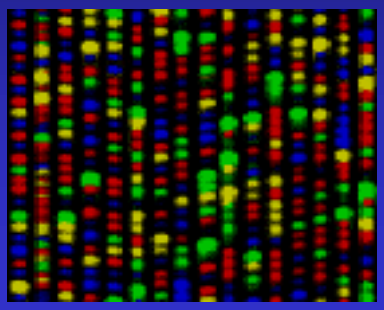
For comparison the alpha and beta chains of hemoglobin are about 150 aa.

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.



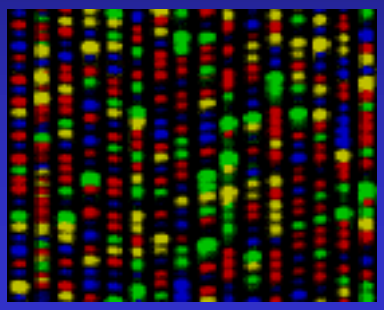
AUTOMATED SEQUENCING

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

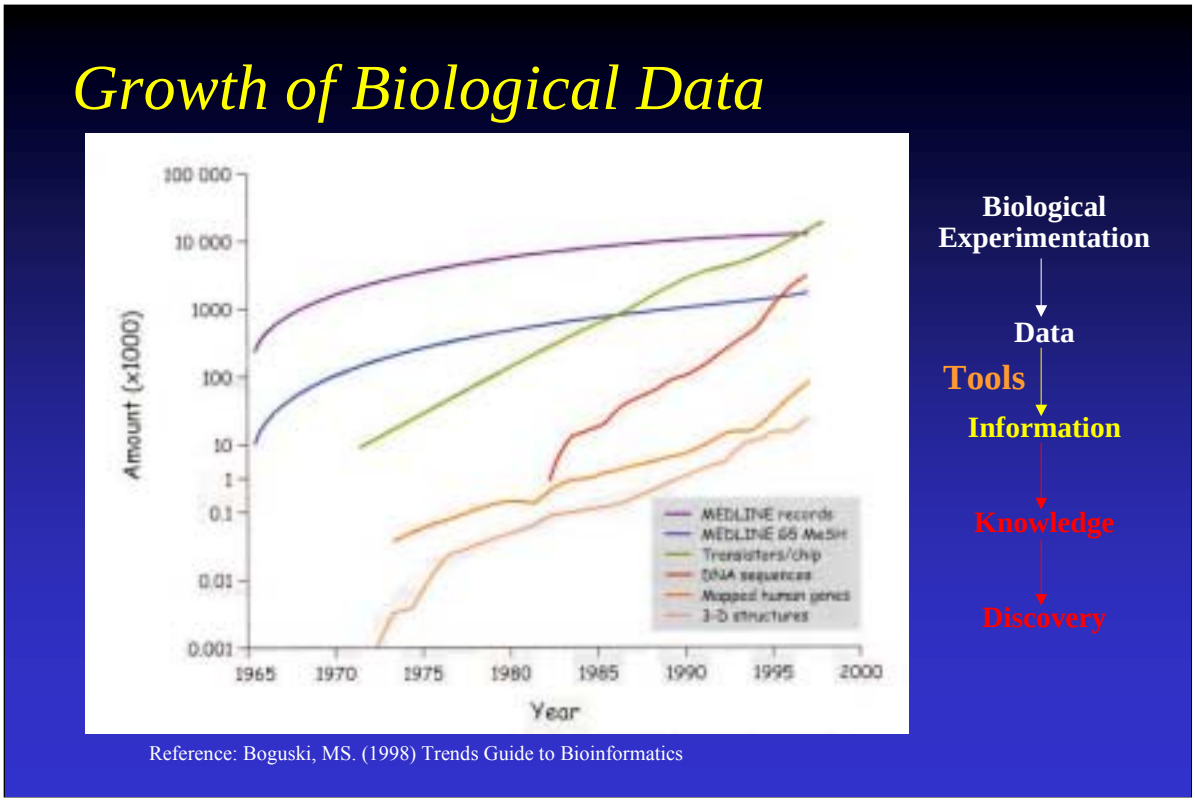
DNASYSTEM (www-biology.ucsd.edu/others/dsmith/dnasys.html) Doug Smith, UCSD Biology Dept.
Provides brief descriptions and links to most bioinformatic databases and web sites

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)

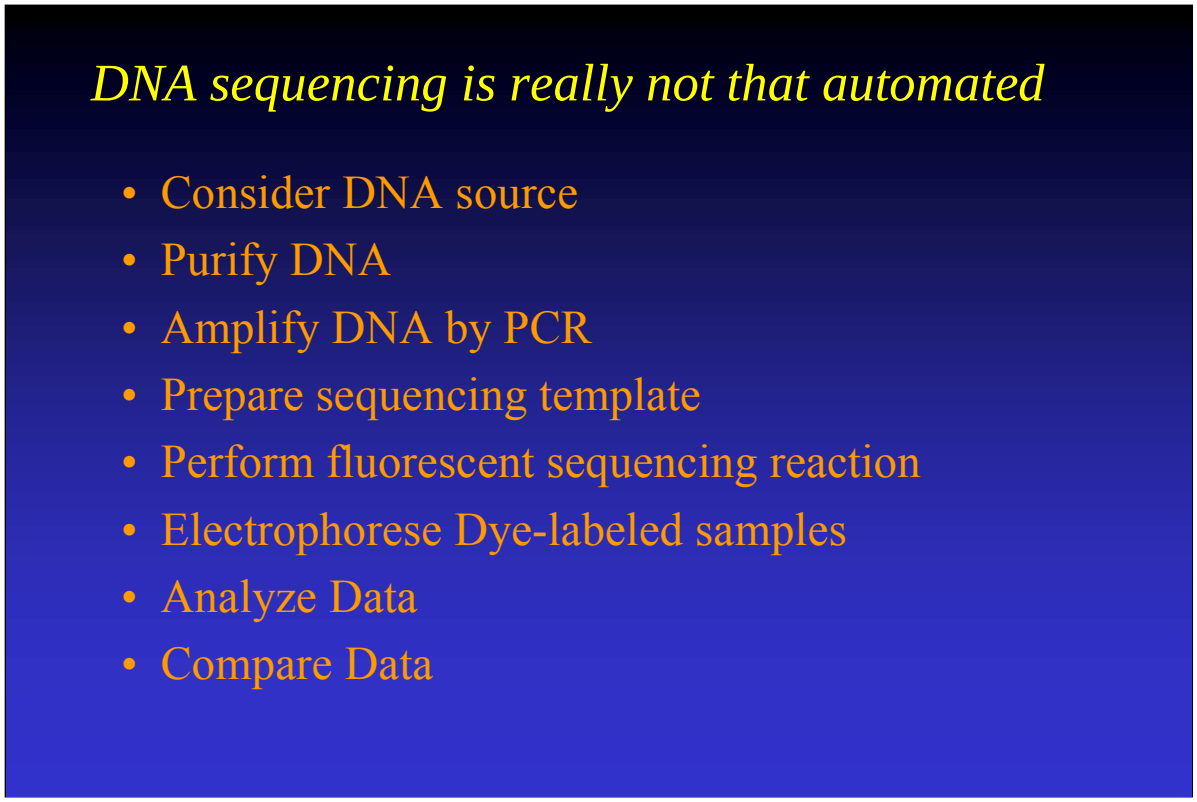


GROWTH OF BIOLOGICAL DATA

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.

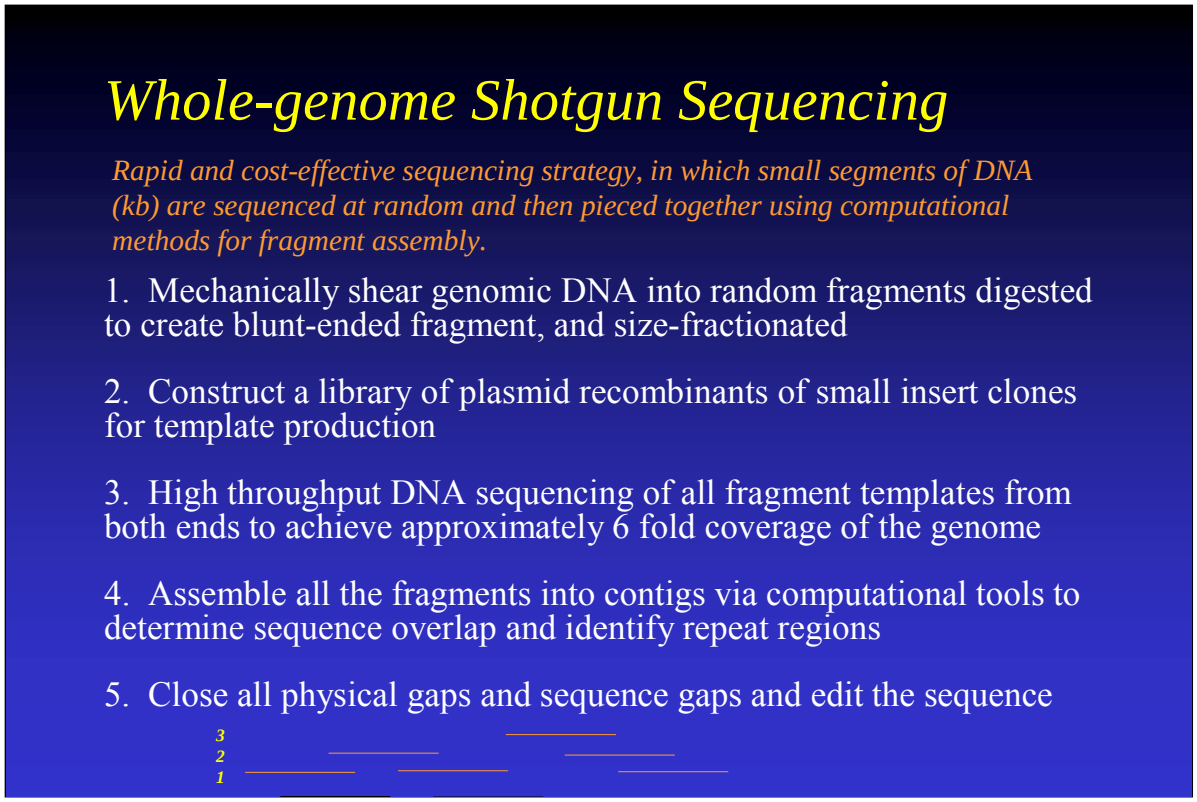


AUTOMATION

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

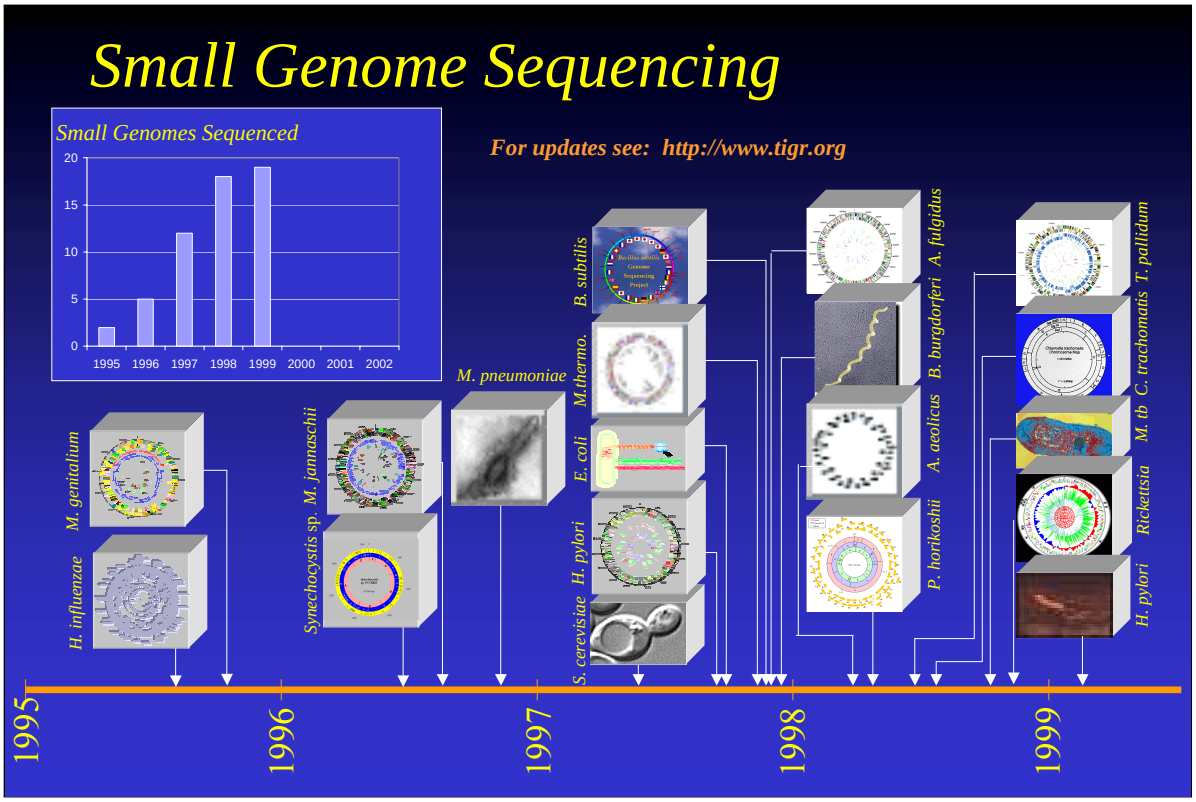
Whole genome sequencing is used to establish a full sequence. The basic idea is to randomly (mechanically) break the DNA into fragments size fractionate these fragments, capture these fragments and sequence them as described. Islands of the whole genome sequence are obtained as shown schematically at 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.

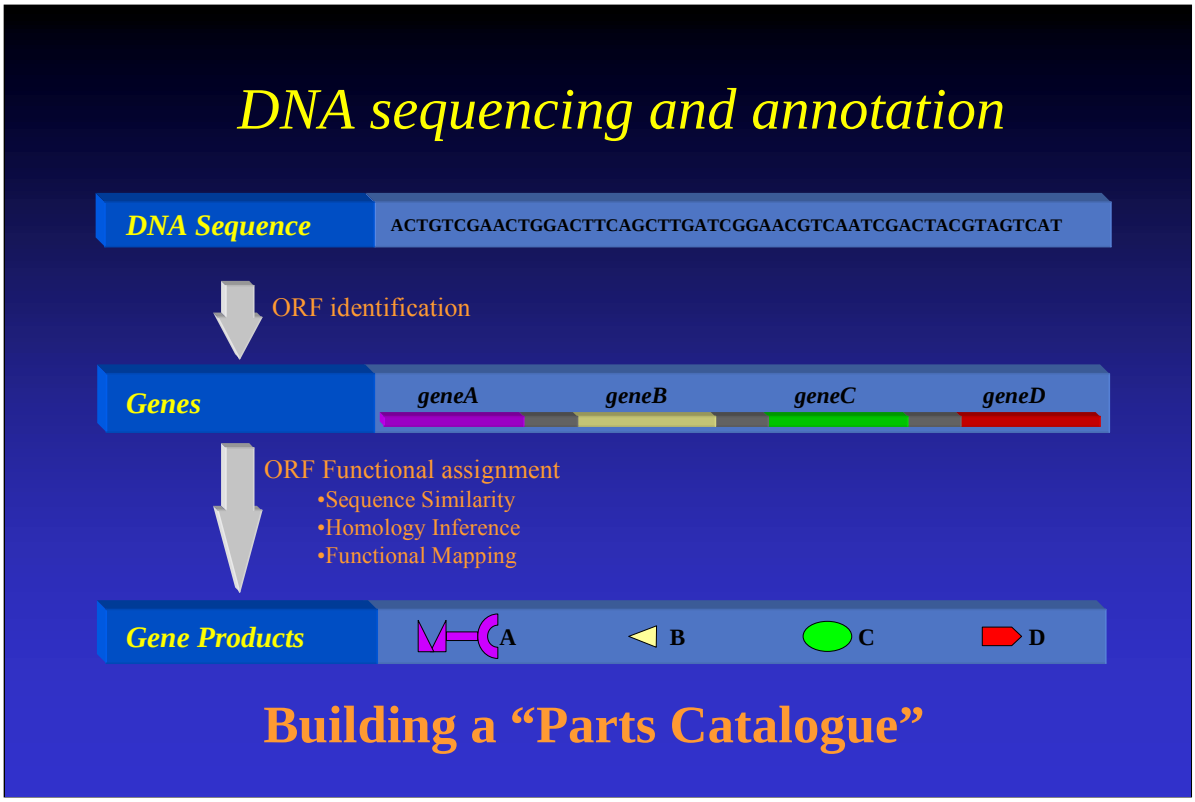
Currently, the complete genome sequence has been determined for hundreds of microorganisms (>30 in public domain), and a handful of multicellular organisms, including human, fruit fly and the nematode C. elegans. The number of these sequences is expected to grow rapidly.

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	

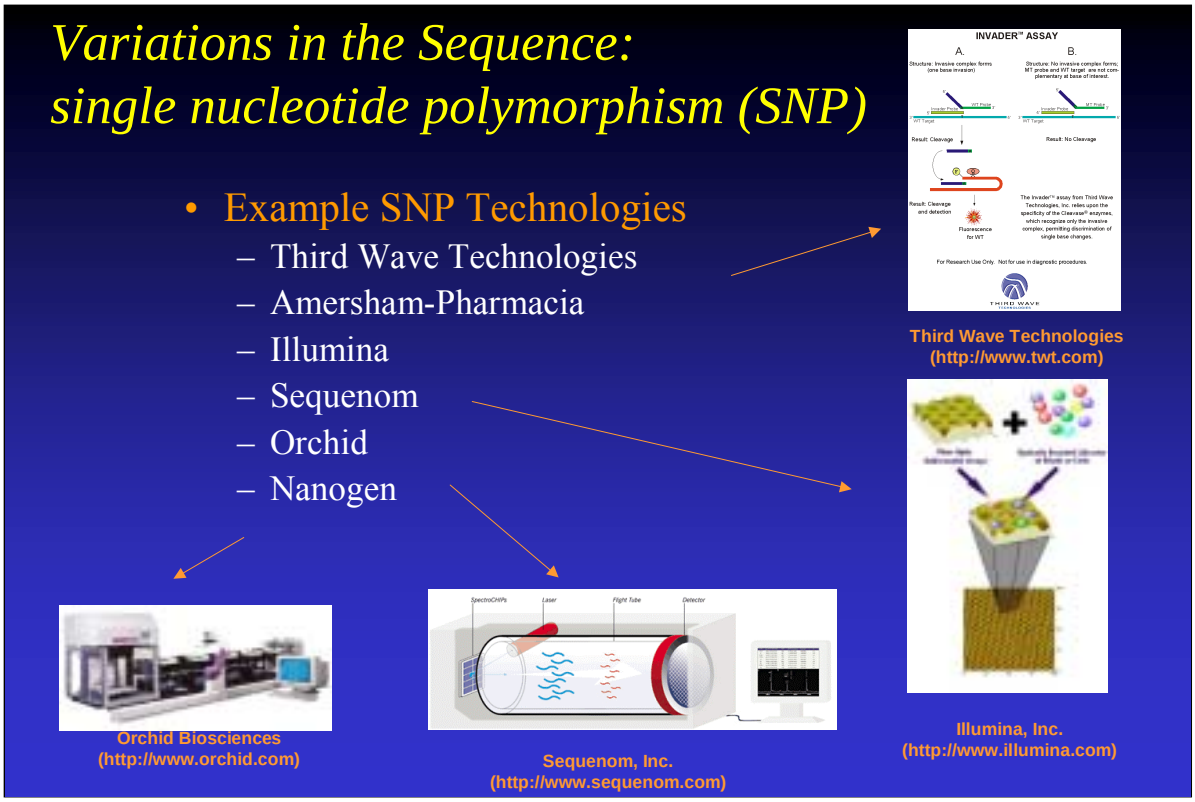
Comparing genomes and sequences

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

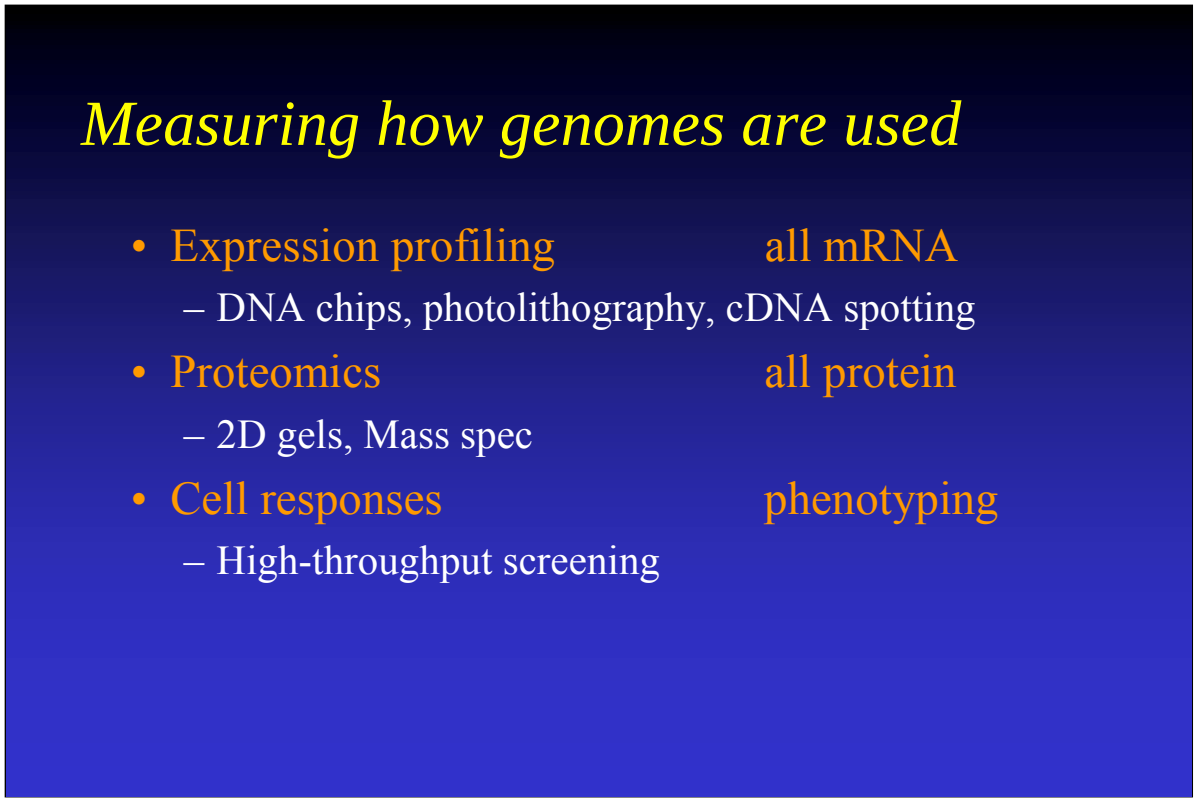


SNPs

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

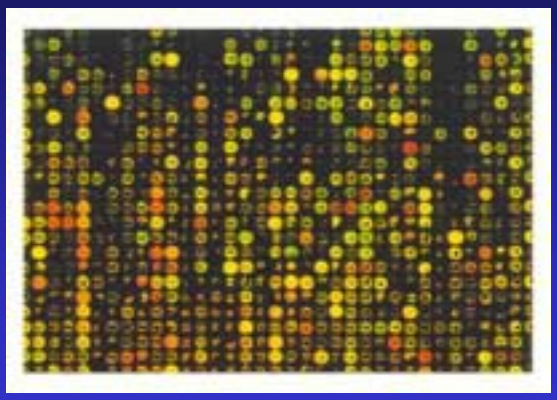
Now that we have full DNA sequences and gene complements, there are a number of approaches emerging that allow us to measure on a genome-scale how these genes are deployed by an organism and what the resulting phenotypic behavior is.

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.

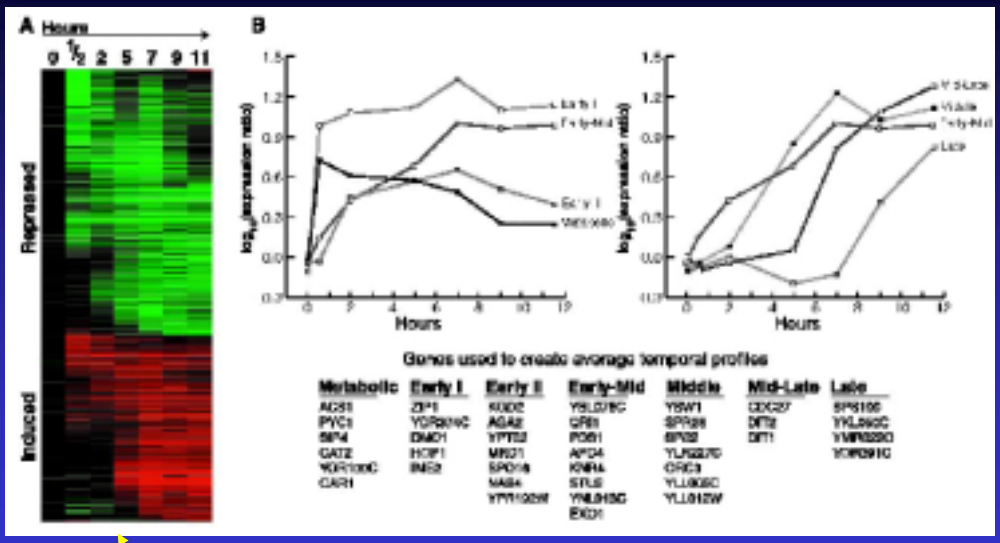
In addition to the slides to follow; here are a couple of studies of great interest:

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



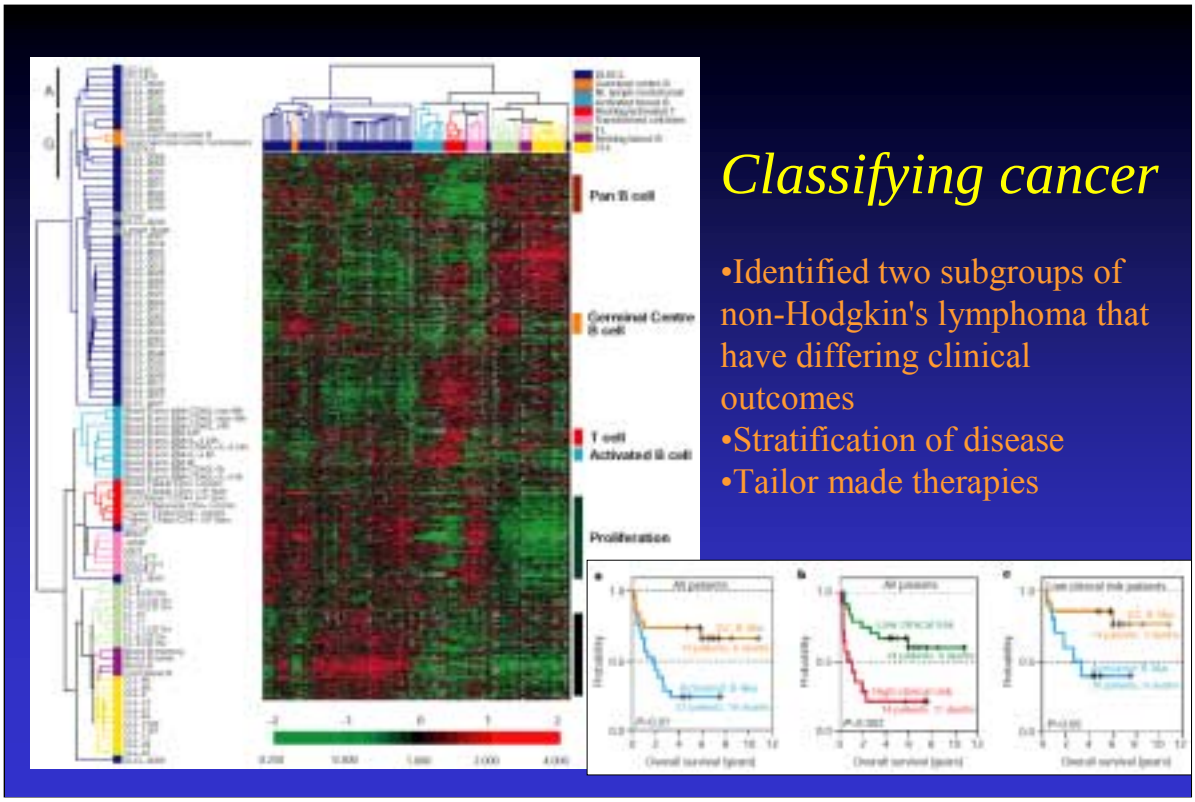
Each array is converted into a column, or a vector.
This vector is a snapshot of the state variables of the cell

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

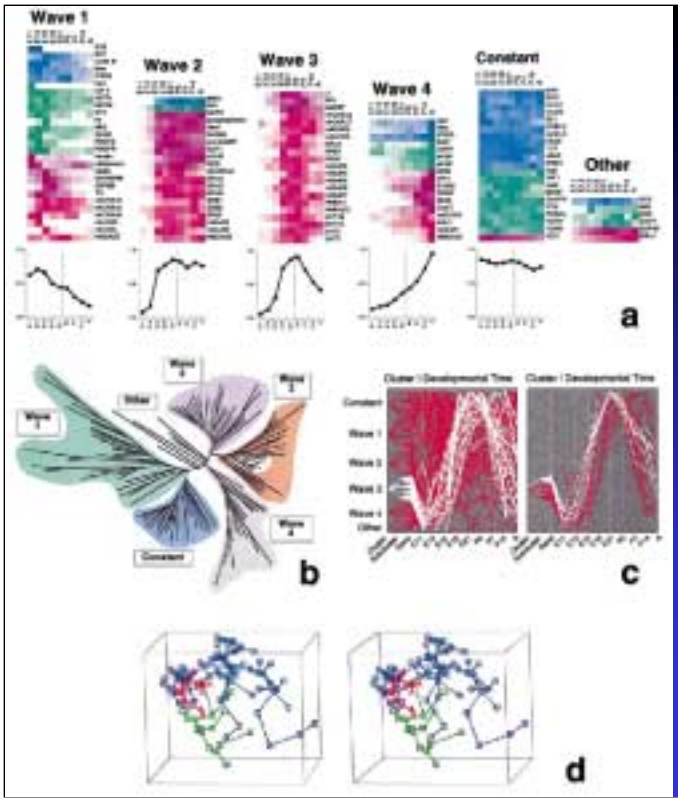
ABSTRACT: Diploid cells of budding yeast produce haploid cells through the developmental program of sporulation, which consists of meiosis and spore morphogenesis. DNA microarrays containing nearly every yeast gene were used to assay changes in gene expression during sporulation. At least seven distinct 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

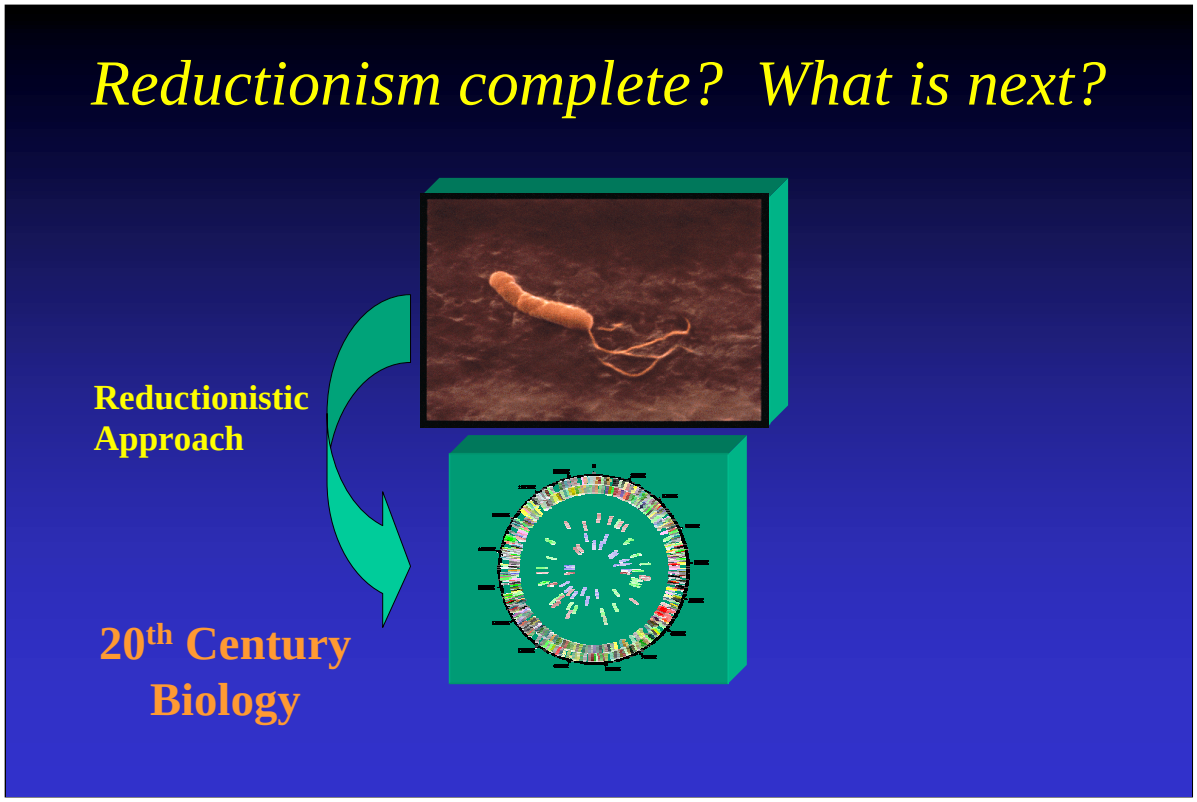
XILING WEN*, STEFANIE FUHRMAN*, GEORGE S. MICHAELS † ,DANIEL B. CARR † , SUSAN SMITH*, JEFFERY L. BARKER*, AND ROLAND SOMOGYI* ‡

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