ECE 370
Introduction to Biomedical Engineering

Computational Biology and Bioinformatics

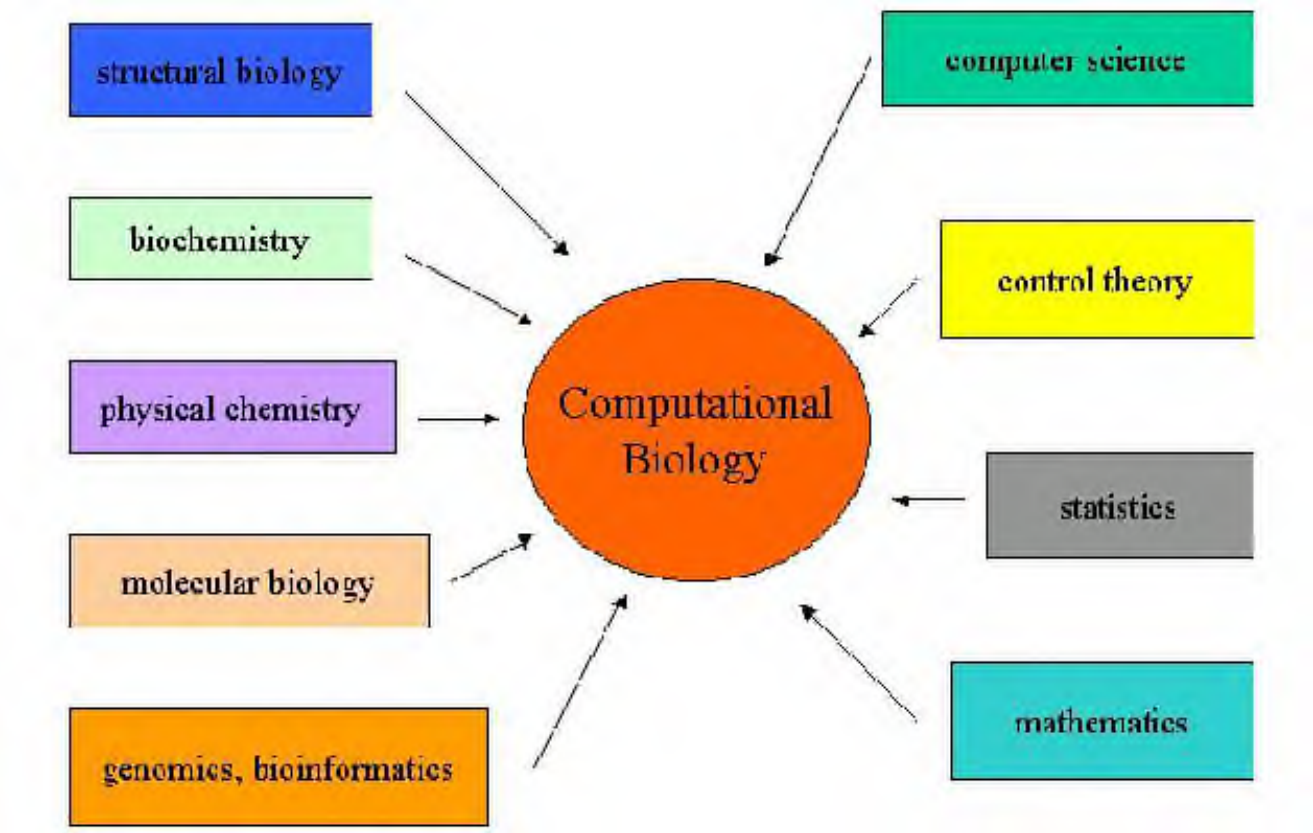
Definitions

- **NIH Definitions**
- **Computational Biology:**
  - The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.
- **Bioinformatics:**
  - Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.
- **The terms computational biology and bioinformatics are often used interchangeably.**
  - However, computational biology sometimes connotes the development of algorithms, mathematical models, and methods for statistical inference, while bioinformatics is more associated with the development of software tools, databases, and visualization methods.

- No definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

Computational Biology

- **What is Computational Biology?**
  - The use of computational techniques to model biological systems at various levels of complexity - atomic, metabolic, cellular and pathologic.



Computational Biology

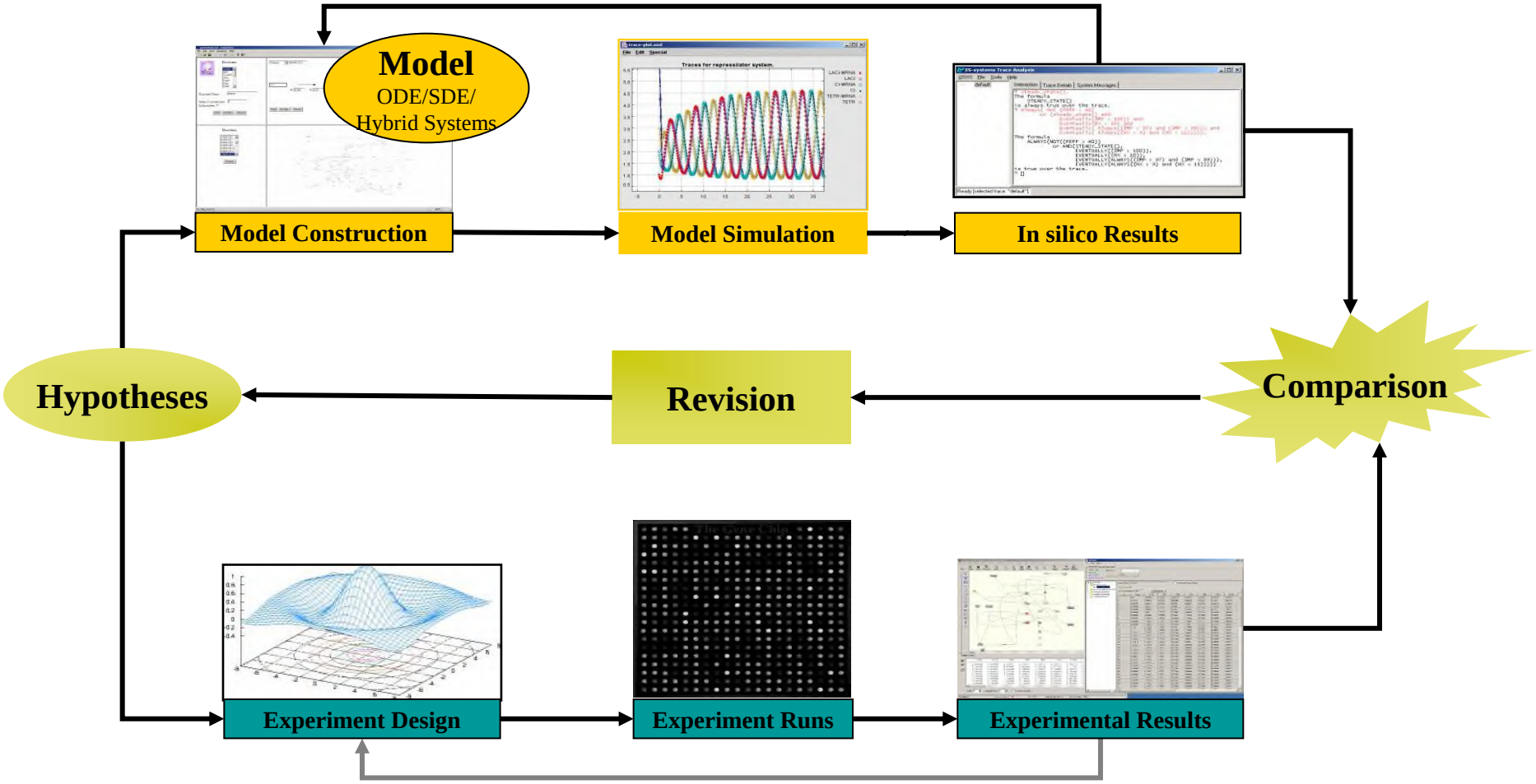
- **Biology of the (not so) past**
  - Isolated
  - Low level (one variable at a time)
  - Slow accumulation of knowledge
- **Biology of the present**
  - Global
  - High level (organismal/theoretical)
  - Rapid accumulation of knowledge
  - Rapid generation of open questions
- **Computers – not just for analysis**
  - Before: Biologists generate data, computers analyze it
  - Now: Computers generate experiments, biologists perform them
  - Biotech has greatest opportunity for science to be done, and computation is crucial!

Computational Biology

- **Drawing upon mathematical approaches developed in the context of**
  - dynamical systems,
  - kinetic analysis,
  - computational theory and
  - logic
- **Possible to create powerful simulation, analysis and reasoning tools for working biologists to be used in**
  - deciphering existing data,
  - devising new experiments and ultimately,
  - understanding functional properties of genomes, proteomes, cells, organs and organisms.

Computational Biology

- Reasoning and Experimentation



Computational Biology

Functional genomic hypothesis generation and experimentation by a robot scientist

Ross D. King[1], Kenneth E. Whelan[1], Ffion M. Jones[1], Philip G. K. Reiser[1], Christopher H. Bryant[2], Stephen H. Muggleton[3], Douglas B. Kell[4] & Stephen G. Oliver[5]

[1]*Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, UK*
[2]*School of Computing, The Robert Gordon University, Aberdeen AB10 1FR, UK*
[3]*Department of Computing, Imperial College, London SW7 2AZ, UK*
[4]*Department of Chemistry, UMIST, P.O. Box 88, Manchester M60 1QD, UK*
[5]*School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Manchester M13 9PT, UK*

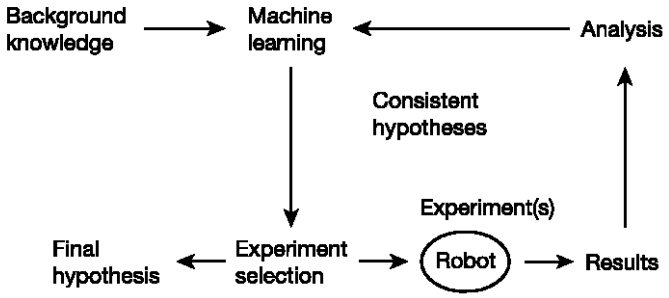


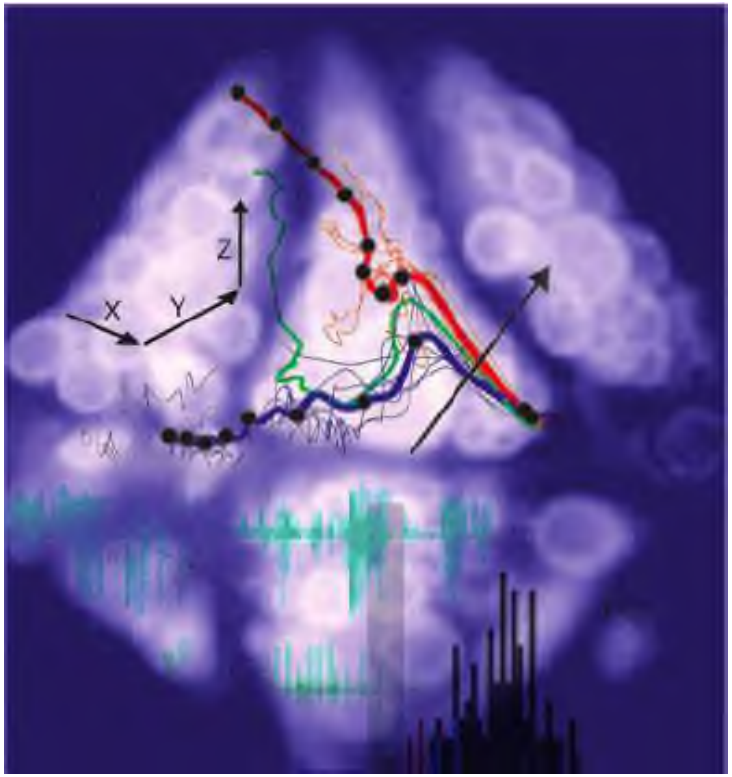
Figure 1 The Robot Scientist hypothesis-generation and experimentation loop.

Future Biology

- **Biology of the future should only involve a biologist and his dog:**
  - the biologist to watch the biological experiments and understand the hypotheses that the data-analysis algorithms produce
  - the dog to bite him if he ever touches the experiments or the computers.

Computational Biology

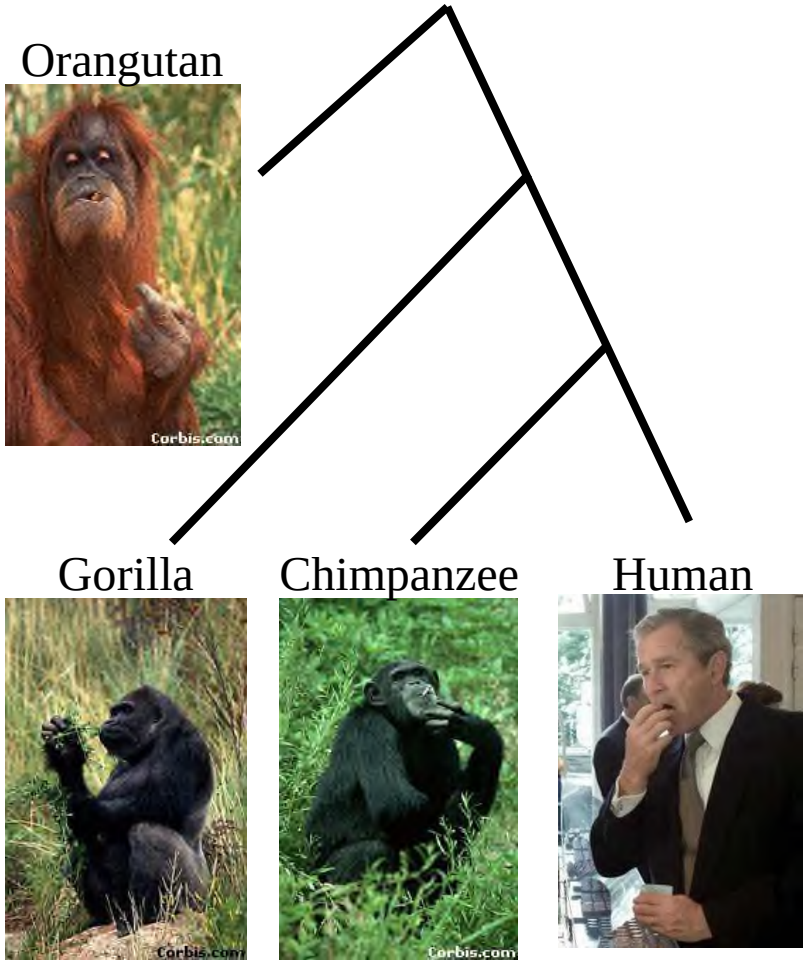
- **Various Sub-fields**
  - Computational biomodeling
  - Computational neuroscience
  - Computational pharmacology
  - Computational evolutionary biology
  - Cancer computational biology
  - Computational genomics (Computational genetics)



Computational Biology

Example Application: Evolution & Phylogeny

- **Phylogeny**
  - A tree representation for the evolutionary history relating the species we are interested in.
  - At each leaf of the tree is a species – we also call it a taxon in phylogenetics (plural form is taxa). They are all distinct.
  - Each internal node corresponds to a speciation event in the past.
  - When reconstructing the phylogeny we compare the characteristics of the taxa, such as their appearance, physiological features, or the composition of the genetic material.



From the Tree of the Life Website, University of Arizona

Computational Biology

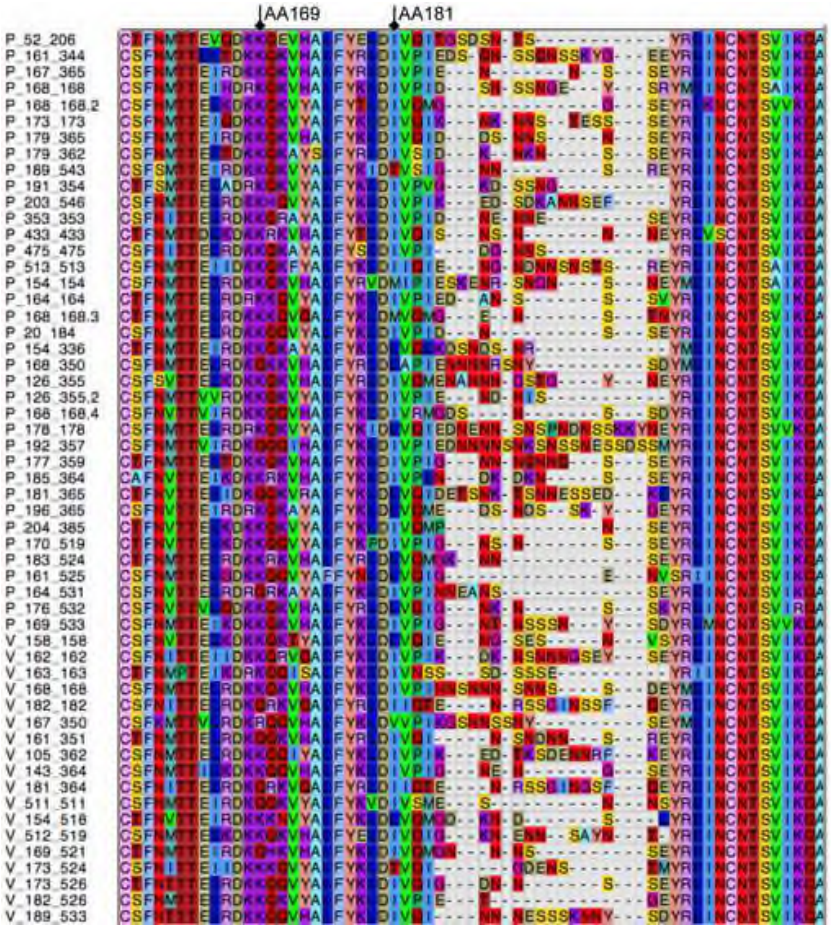
Example Application: Evolution & Phylogeny

- **Phylogenetic Analysis**
  - Step 1: Gather sequence data, and estimate the multiple alignment of the sequences.
  - Step 2: Reconstruct trees on the data. (This can result in *many* trees.)
  - Step 3: Apply consensus methods to the set of trees to figure out what is reliable.

Computational Biology

Example Application: Evolution & Phylogeny

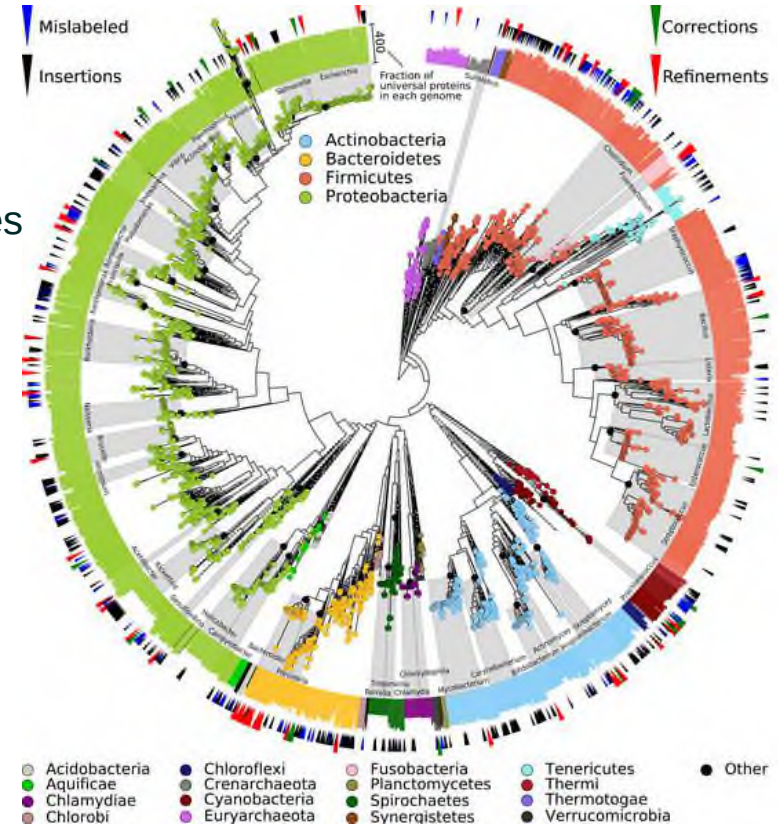
- **Data (Past)**
  - Physiological and morphological features
- **Data (Present)**
  - Biomolecular sequences: DNA, RNA, amino acid, in a multiple alignment
  - Molecular markers (e.g., SNPs, RFLPs, etc.)
  - Morphology
  - Gene order and content
  - These are “character data”: each character is a function mapping the set of taxa to distinct states (equivalence classes), with evolution modelled as a process that changes the state of a character



Computational Biology

Example Application: Evolution & Phylogeny

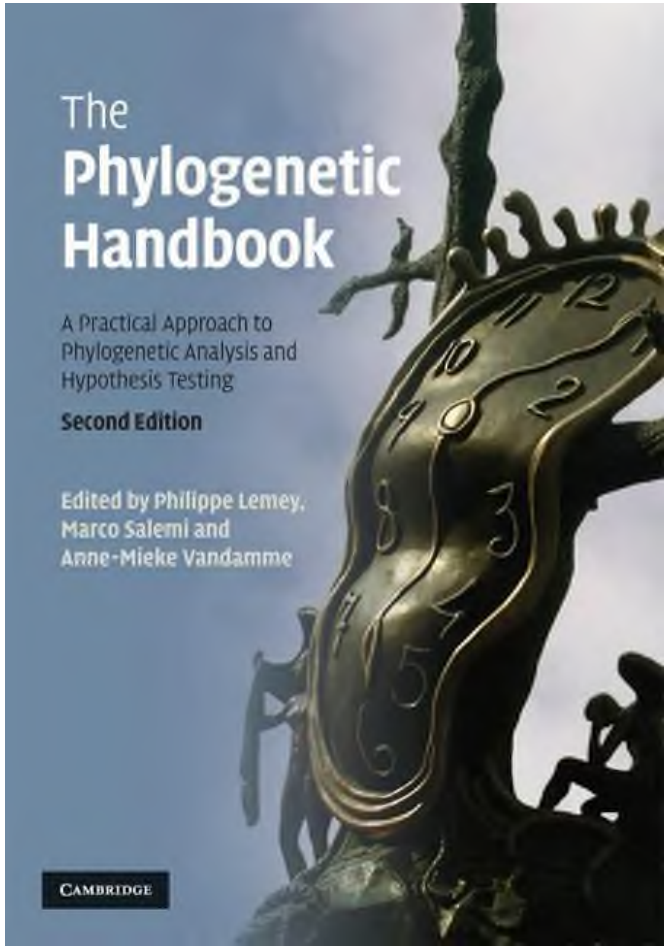
- **Reconstruction methods**
  - Most attempt to solve one of two major optimization criteria:
    - Maximum Parsimony (preferring the simpler of two otherwise equally adequate theorizations)
    - Maximum Likelihood (selecting the set of values of the model parameters that maximizes the "agreement" of the selected model with the observed data)
  - Methods for phylogeny reconstruction are evaluated primarily in simulation studies, based upon stochastic models of evolution.
- **Consensus and agreement methods**
  - Take a set of trees on the same set of taxa, and return a single tree on the full set.
  - Consensus
    - Strict consensus and majority tree.
  - Agreement
    - Maximum agreement subtree.
  - Much new research needs to be done



Computational Biology

Example Application: Evolution & Phylogeny

- **Major challenges**
  - Main challenge: make it possible to obtain good solutions to MP or ML in reasonable time periods on large datasets
  - Speed up searches through tree-space
  - Incorporate new data (e.g., gene order and content)
  - Evaluate novel methodologies
    - Non-tree models
    - Supertree methods



Bioinformatics

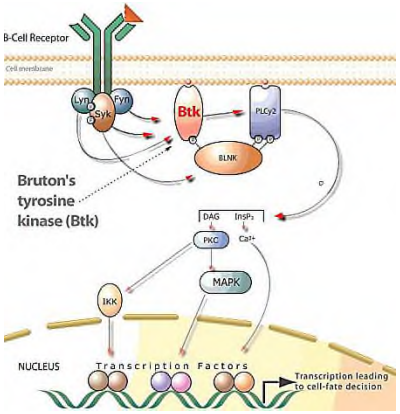
- **What Is Bioinformatics?**
  - Bioinformatics is the unified discipline formed from the combination of biology, computer science, and information technology.
  - "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.“ –Frank Tekaia

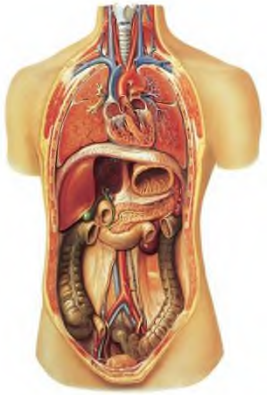
Bioinformatics

- **Central Paradigm of Bioinformatics**

Genetic Information → **Molecular Structure** → **Biochemical Function** → **Symptoms (Phenotype)**

```
TTGGAAAACATTCATG
ATTTATGGGATAGAGC
TTTAGATCAAATTGAA
AAAAAATTAAGCAAAC
CTAGTTTTGAAACCTG
GCTCAAATCGACAAAA
GCTCATGCTTTACAAG
GAGACACGCTCATTAT
TACTGCACCTAATGAT
TTTGCACGGGACTGGT
TAGAATCTAGGTATTC
TAATTTAATTGCTGAA
ACACTTTATGATCTTA
CGGGGGAAGAGTTAGA
TGTAAAATTTATTATT
```





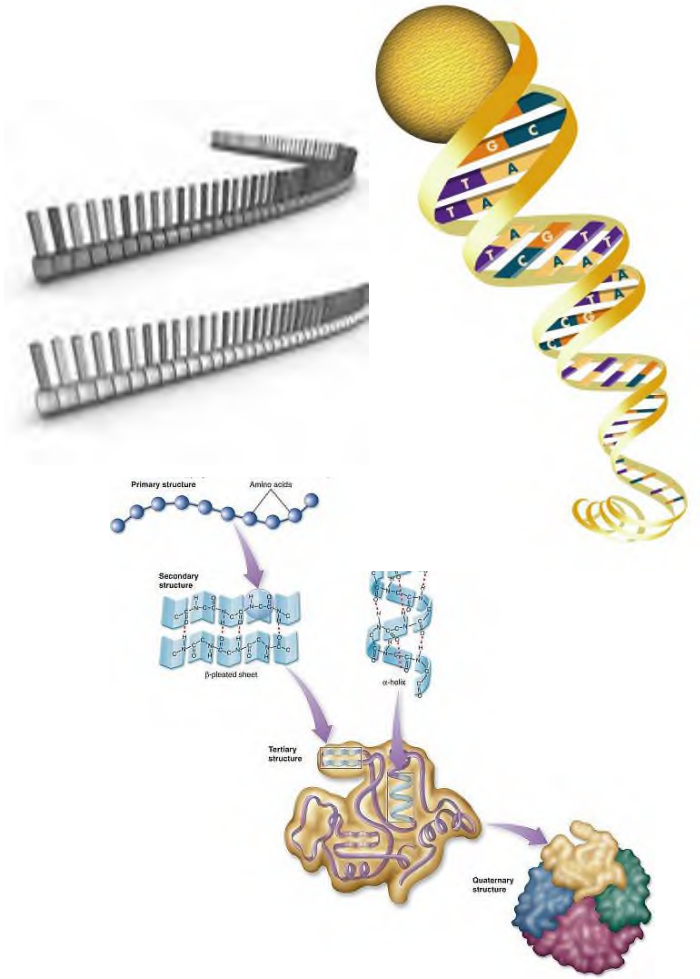
Bioinformatics

Computational Goals of Bioinformatics

- **Learn & Generalize:**
  - Discover conserved patterns (models) of sequences, structures, interactions, metabolism & chemistries from well-studied examples.
- **Predict:**
  - Infer function or structure of newly sequenced genes, genomes, proteins or proteomes from these generalizations.
- **Organize & Integrate:**
  - Develop a systematic and genomic approach to molecular interactions, metabolism, cell signaling, gene expression…
- **Simulate:**
  - Model gene expression, gene regulation, protein folding, protein-protein interaction, protein-ligand binding, catalytic function, metabolism…
- **Engineer:**
  - Construct novel organisms or novel functions or novel regulation of genes and proteins.
- **Treat:**
  - Gene Therapy: Target specific genes, or mutations, RNAi to change a disease phenotype.

Bioinformatics

- **Explosion of "Omes" & "Omics!“**
  - Genome
    - The complete collection of DNA (genes and "non-genes") of an organism
    - 4-letter base code
    - ~ 1,000 base pairs in a small gene
    - ~ 3 X 109 bp in a genome (human)
  - Transcriptome
    - The complete collection of RNAs (mRNAs & others) expressed in an organism
  - Proteome
    - The complete collection of proteins expressed in an organism
    - 20 letter alphabet (amino acids)
      - ACDEFGHIKLMNPQRSTVWY (but not BJOUXZ)
    - ~ 300 aa in an average protein
    - ~ 3 X 10^6 known protein sequences
  - Note:
    - The set of specific RNAs or proteins expressed varies greatly in different cells and tissues -- and critically depends on the age, developmental stage, disease state, etc. of the organism



Genomics

- **Genome**
  - complete set of genetic instructions for making an organism
- **Genomics**
  - any attempt to analyze or compare the entire genetic complement of a species
- **Early genomics was mostly recording genome sequences**
  - The Human Genome sequence is published
  - 3 Gb
  - And the peasants rejoice!

Genomics

DNA Sequencing

- **Bioinformatics is based on the fact that DNA sequencing is cheap, and becoming easier and cheaper very quickly.**
  - the Human Genome Project cost roughly $3 billion and took 12 years (1991-2003).
  - Sequencing James Watson's genome in 2007 cost $2 million and took 2 months
  - Today, you could get your genome sequenced for about $100,000 and it would take a month.
  - The Archon X prize: you win $10 million if you can sequence 100 human genomes in 10 days, at a cost of $10,000 per genome.
  - It is realistic to envision $100 per genome within 10 years: everyone's genome could be sequenced if they wanted or needed it.

Genomics

- **Why it's useful?**
  - All of the information needed to build an organism is contained in its DNA. If we could understand it, we would know how life works.
    - Preventing and curing diseases like cancer (which is caused by mutations in DNA) and inherited diseases.
    - Curing infectious diseases (everything from AIDS and malaria to the common cold). If we understand how a microorganism works, we can figure out how to block it.
    - Understanding genetic and evolutionary relationships between species
    - Understanding genetic relationships between humans. Projects exist to understand human genetic diversity. Also, sequencing the Neanderthal genome.
      - Ancient DNA: currently it is thought that under ideal conditions (continuously kept frozen), there is a limit of about 1 million years for DNA survival. So, Jurassic Park will probably remain fiction.

Genomics

- **From DNA to Gene**
  - But: extracting that information is difficult. How to convert a string of ACGT's into knowledge of how the organism works is hard.
  - Most of the work is on the computer, with key confirming experiments done in the "wet lab".
  - The sequence below contains a gene critical for life: the gene that initiates replication of the DNA. Can you spot it?
  - Data mining techniques are used for this purpose
    - Data mining: the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns for business advantage.

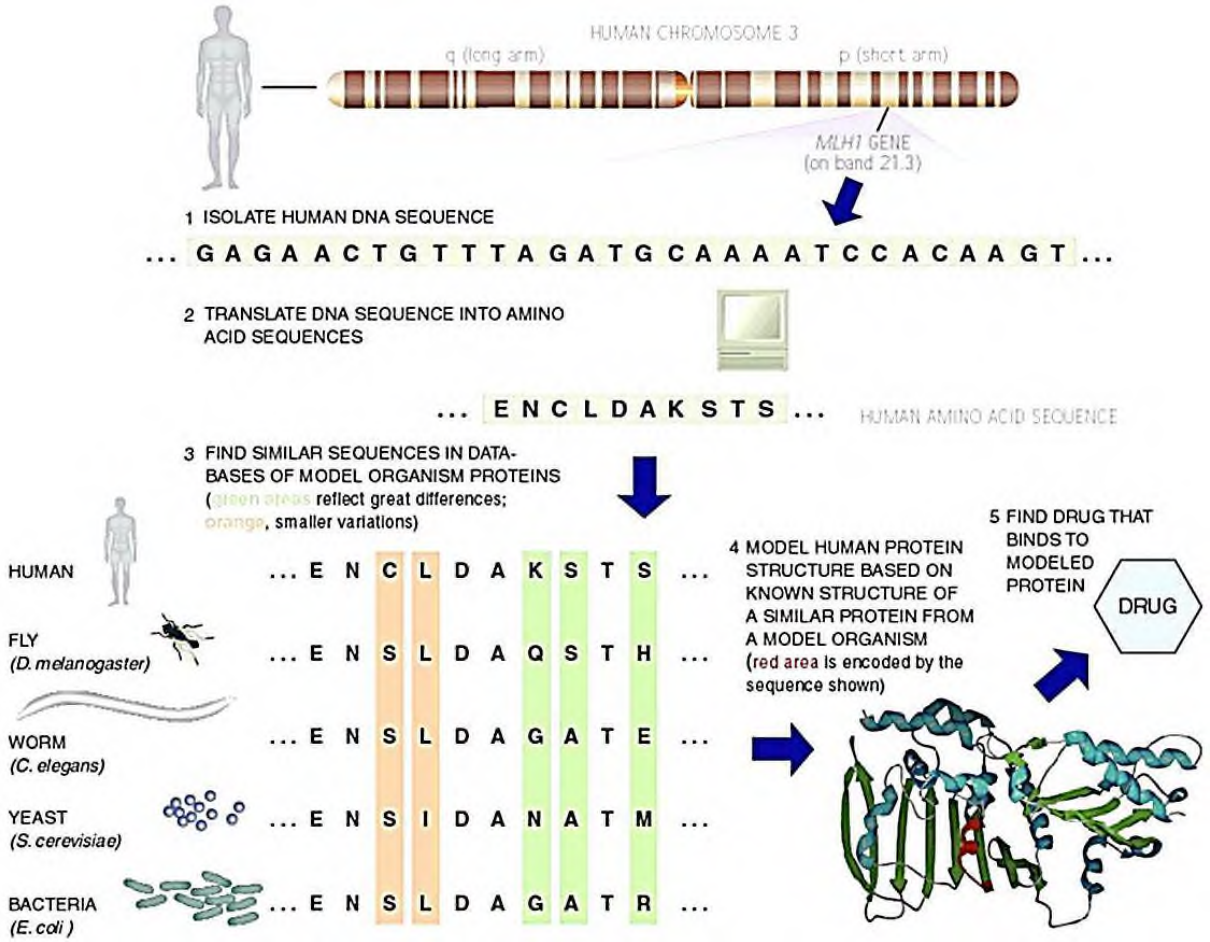
```
TTGGAAAACATTCATGATTTATGGGATAGAGCTTTAGATCAAATTGAAAAAAAATTAAGCAAACCTAGTTTTGAAACCTG
GCTCAAATCGACAAAAGCTCATGCTTTACAAGGAGACACGCTCATTATTACTGCACCTAATGATTTTGCACGGGACTGGT
TAGAATCTAGGTATTCTAATTTAATTGCTGAAACACTTTATGATCTTACGGGGGAAGAGTTAGATGTAAAATTTATTATT
CCTCCTAACCAGGCCGAGGAAGAATTCGATATTCAAACTCCTAAAAAGAAAGTCAATAAAGACGAAGGAGCAGAATTTCC
TCAAAGCATGCTAAATTCGAAGTATACCTTTGATACATTTGTTATCGGATCTGGAAATCGGTTTGCGCATGCAGCTTCTT
TAGCAGTAGCAGAAGCGCCGGCTAAAGCGTATAATCCGCTTTTTATTTACGGGGGAGTAGGATTAGGCAAAACACACTTA
ATGCACGCCATAGGCCACTATGTGTTAGATCATAATCCTGCCGCGAAAGTCGTGTACTTATCATCTGAAAAATTCACAAA
CGAGTTTATTAACTCTATTCGTGACAATAAAGCAGTAGAATTCCGCAACAAATACCGTAATGTAGATGTTTTACTGATTG
ATGATATTCAATTCTTAGCAGGTAAAGAGCAGACACAAGAAGAATTTTTCCATACGTTTAATACGCTTCACGAAGAAAGC
AAGCAGATTGTCATCTCAAGTGATCGACCGCCGAAAGAAATTCCTACACTTGAAGATCGACTTCGCTCTCGCTTTGAATG
GGGCCTTATTACAGACATCACACCACCAGATTTGGAAACACGAATTGCTATTTTGCGTAAAAAAGCCAAAGCGGACGGCT
TAGTTATTCCAAATGAAGTTATGCTTTATATCGCCAATCAGATTGATTCAAATATTAGAGAATTAGAAGGCGCACTTATT
```

Genomics

- **What's next? → Post Sequencing Era**
  - Comparative Genomics
    - the management and analysis of the millions of data points that result from Genomics
      - Sorting out the mess
  - Functional Genomics
    - Other, more direct, large-scale ways of identifying gene functions and associations
      - for example yeast two-hybrid methods
  - Structural Genomics
    - emphasizes high-throughput, whole-genome analysis.
      - outlines the current state
      - future plans of structural genomics efforts around the world and describes the possible benefits of this research

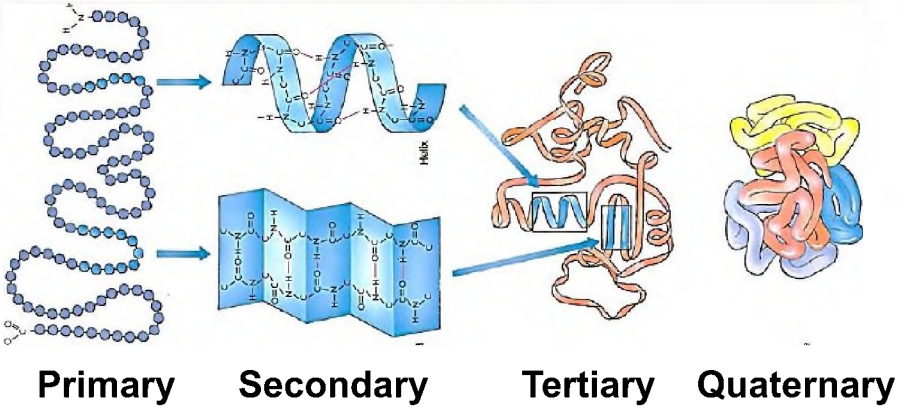
Genomics

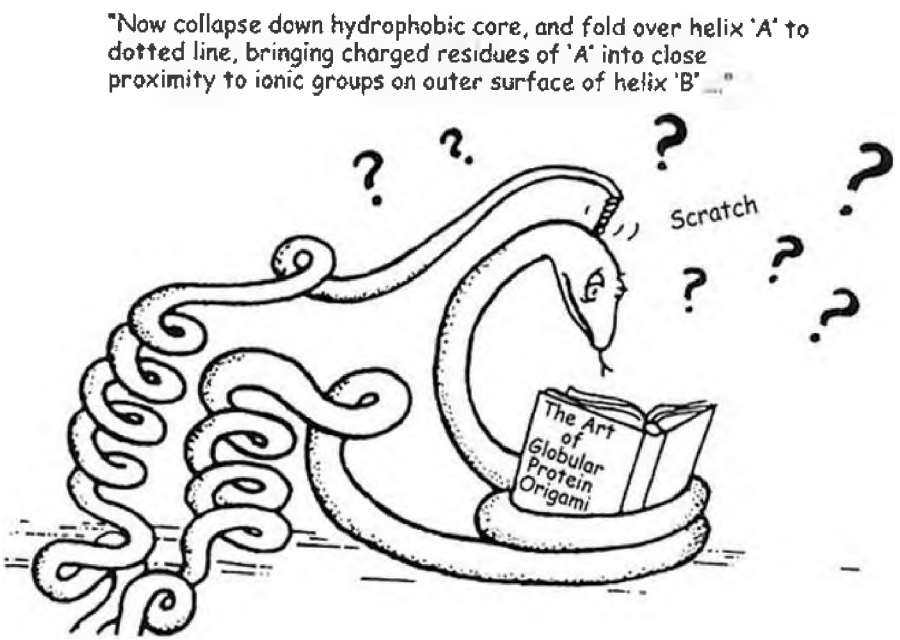
- **Finding Homologs (Comparative Genomics)**
  - Homologs - "same genes" in different organisms
  - Search large portions of entire genome to find a gene which is similar to human
    - Human vs. Mouse vs. Yeast
    - Much easier to do experiments on yeast!



Genomics

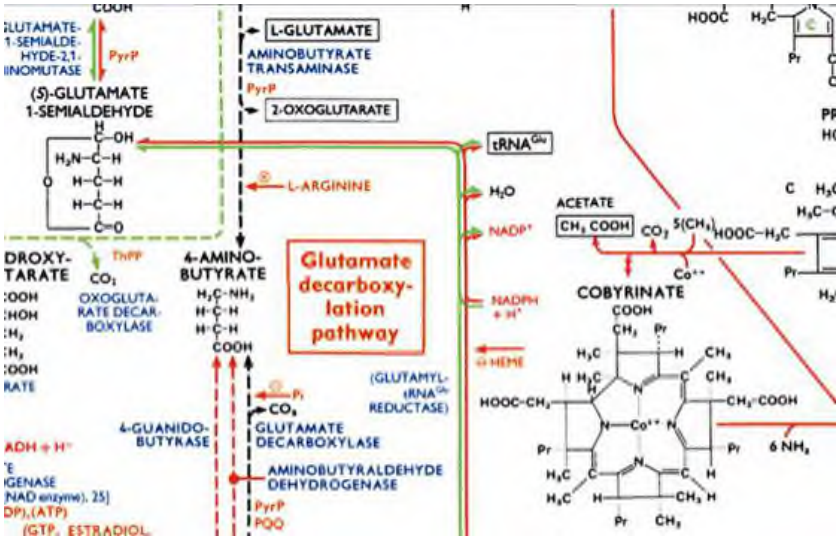
- **From Gene to Protein (Structural Genomics)**
  - Macromolecular Structures
    - Proteins acquire a 3-d shape and may bind with other molecules
  - How does a protein (or RNA) sequence fold into an active 3-dimensional structure?
  - Can we predict structure from sequence?
  - Can we predict function from structure (or perhaps, from sequence alone?)
  - We don't yet understand the protein folding code - but we try to engineer proteins anyway!
    - Extensive modeling and simulations

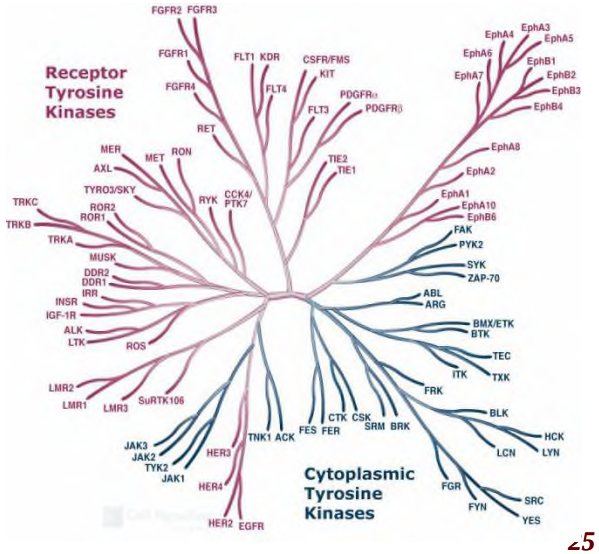




Genomics

- **From Gene to Function (Functional Genomics)**
  - How do patterns of gene expression determine phenotype?
  - How do proteins interact in biological networks?
  - Which genes and proteins are required for differentiation during development?
  - Which genes and pathways have been most highly conserved during evolution?





Transcriptomics

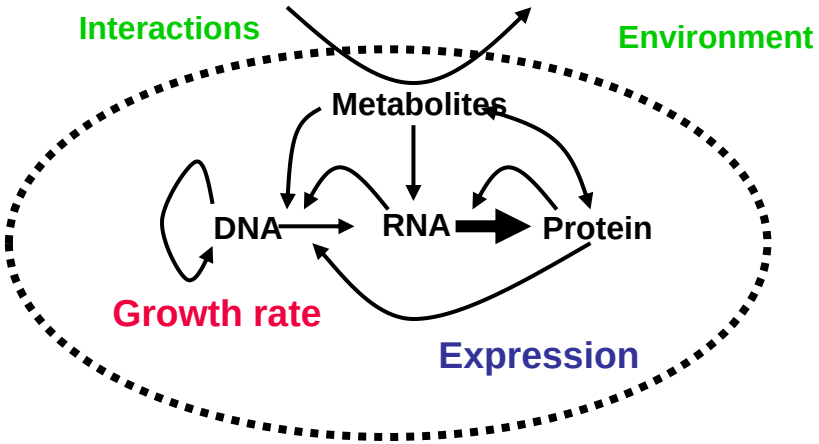
- **Transcriptome**
  - The RNA (of all kinds) produced in a cell
- **What makes transcriptomics important?**
  - A cell's DNA (genome) is a blueprint for the cell's potential but not the cell's actual, current form
    - All cells in an organism contain the same DNA.
    - This DNA encodes every possible cell type in that organism—muscle, bone, nerve, skin, etc.
    - There are more than 160,000 genes in each cell, only a handful of which actually determine that cell's structure.
- **Transcriptomics in Disease Treatment**
  - Nearly all major diseases—more than 98% of all hospital admissions—are caused by an particular pattern in a group of genes.
  - Isolating this group by comparing the hundreds of thousands of genes in each of many genomes would be very impractical.
  - Looking at the trascriptomes of the cells associated with the disease is much more efficient.

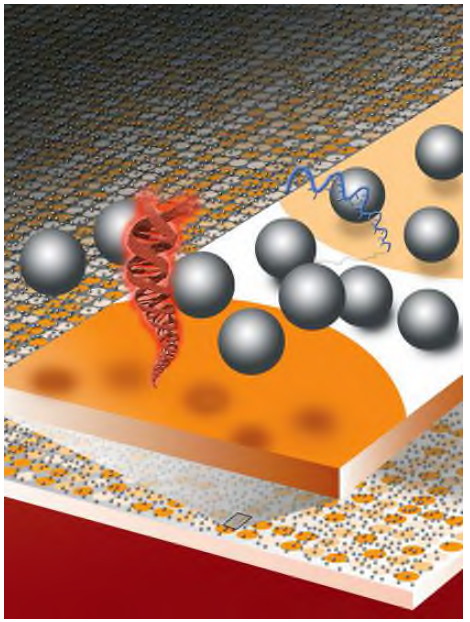
Proteomics

- **Proteome**
  - The proteins expressed in a cell
- **What Makes Proteomics Important?**
  - Many of the interesting things about a given cell's current state can be deduced from the type and structure of the proteins it expresses.
  - Changes in, for example, tissue types, carbon sources, temperature, and stage in life of the cell can be observed in its proteins.
- **Proteomics In Disease Treatment**
  - Many human diseases are caused by a normal protein being modified improperly. This also can only be detected in the proteome, not the genome.
  - The targets of almost all medical drugs are proteins. By identifying these proteins, proteomics aids the progress of pharmacogenetics.

RNA and Protein Expression

- **Cells are different because of differential gene expression.**
  - About 40% of human genes are expressed at one time.
- **Gene expression**
  - Gene is expressed by transcribing DNA into single-stranded mRNA
  - mRNA is later translated into a protein
  - Microarrays measure the level of mRNA or protein expression
- **The cellular environment is dynamic**
  - Expression alone is not enough to complete the picture
  - Interactions within the cell
    - Pathways involving proteins, enzymes, and metabolites
  - Interactions with the environment
    - Milieu, other cells, etc
- **mRNA or protein expression represents dynamic aspects of cell**
  - Can be measured with microarray technology



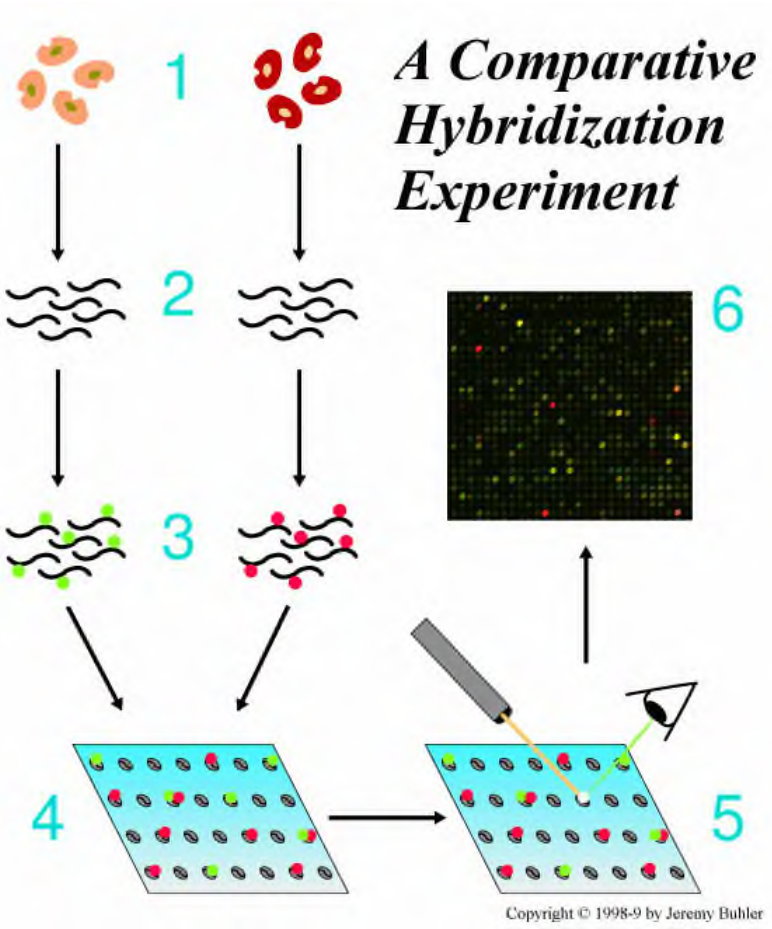


Microarrays

- **What are Microarrays?**
  - Small glass or silicon slides upon the surface of which are arrayed thousands of features (probes)
    - usually between 500 up to a million
    - Can test several probes from several cells at the same time
  - Probes can be
    - DNA or RNA
    - Protein microarrays (Proteomics)
    - Antibody Arrays
    - Tissue Arrays

Microarrays

- **Steps of a Microarray Experiment**
  1. Prepare microarray(s) by choosing probes and attaching them to substrate. Note location and properties of each probe.
  2. From 2 cell samples (say one normal and another with cancer) collect mRNA (make more stable cDNA from them) or proteins and add fluorescent labels (green to normal and red to cancer).
  3. Generate a hybridization solution containing a mixture of fluorescently labelled targets.
  4. Incubate hybridization mixture.
  5. Detect probe hybridization using laser technology
  6. Scan the arrays and store output as images
  7. Quantify each spot, Subtract background and Normalize
  8. Export a table of fluorescent intensities for each gene in the array and build a database
  9. Analyze data using computational (bioinformatics) methods : Statistical analysis, data mining, pathway analysis



Microarrays

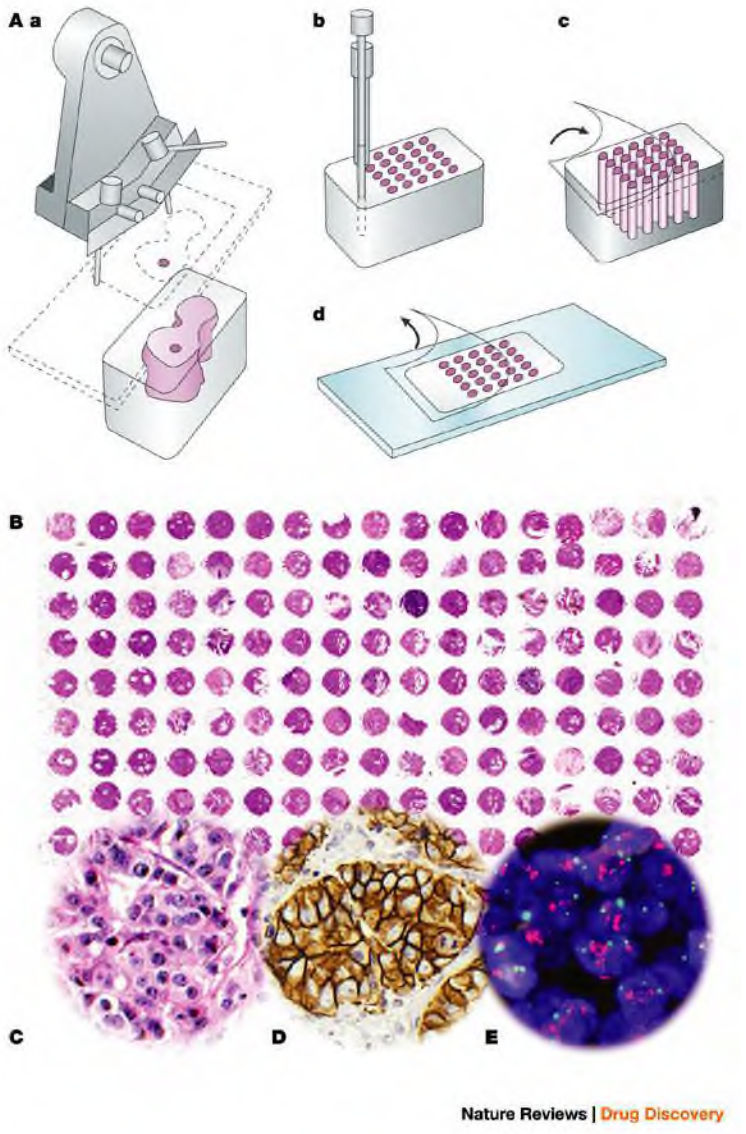
- **Why use Microarrays?**
  - Determine what genes are active in a cell and at what levels
  - Compare the gene expression profiles of a control vs treated
  - Determine what genes have increased or decreased in during an experimental condition
  - Determine which genes have biological significance in a system
  - Discovery of new genes, pathways, and cellular trafficking
- **Why analyze so many genes?**
  - Just because we sequenced a genome doesn't mean we know anything about the genes. Thousands of genes remain without an assigned function.
  - Patterns or clusters of genes are more informative regarding total cellular function than looking at one or two genes – can figure out new pathways

Microarrays

- **Microarray Formats**
  - Cartridge-based
    - Miniaturized, high density arrays of DNA oligos within a plastic housing
    - 1,300,000 DNA oligos 1-cm by 1-cm
    - One sample=One chip
    - Affymetrix, Agilent, Applied Biosystems…
    - Generally used with expression and DNA arrays
    - More expensive
    - Spotted or “photolithographic”
  - Spotted Glass Slide
    - Uses cDNA, Oligonucleotide, protein, antibody
    - Robotically spotted cDNAs or Oligonucleotides
    - Printed on Nylon, Plastic, or Glass microscope slide
    - Involves two dyes on the same slide

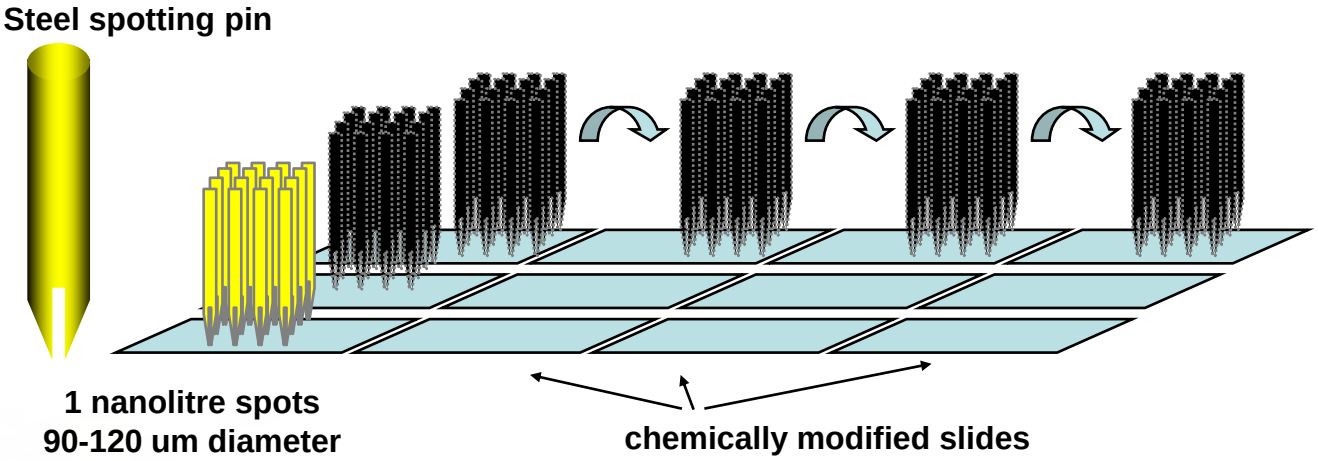
Microarrays

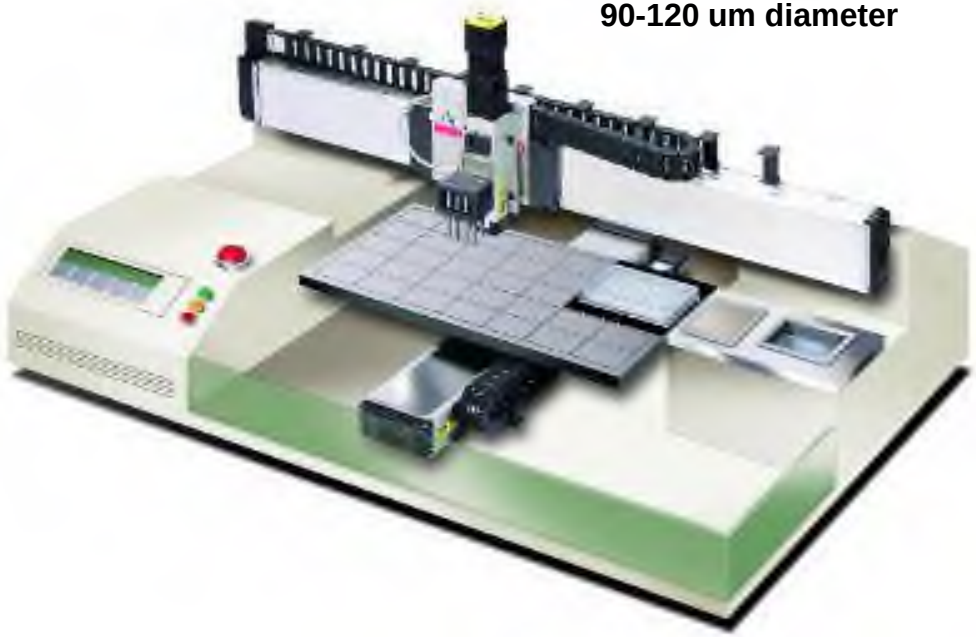
- **Microarray Formats**
  - Tissue Section Slide
    - Slide based “spotted” tissues
    - Coring of embedded paraffin tissues and plugging or inserting into new paraffin block
    - Sectioning and deposition onto a slide



Microarrays

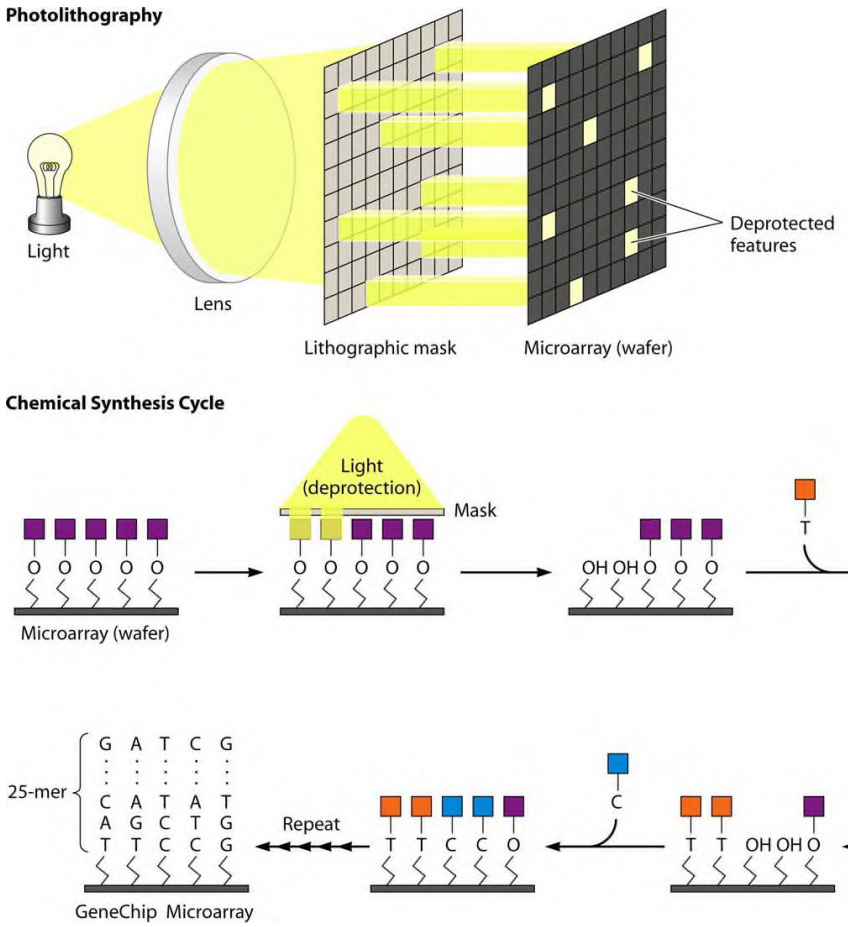
- **Spotted Arrays**





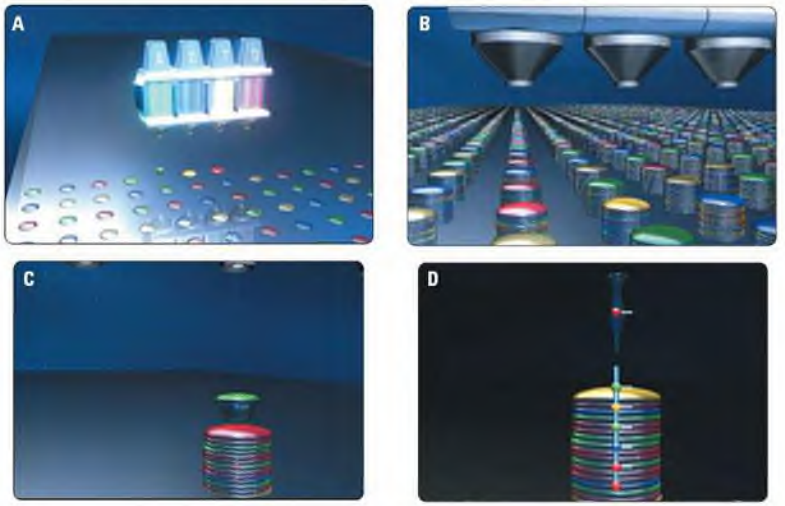
Microarrays

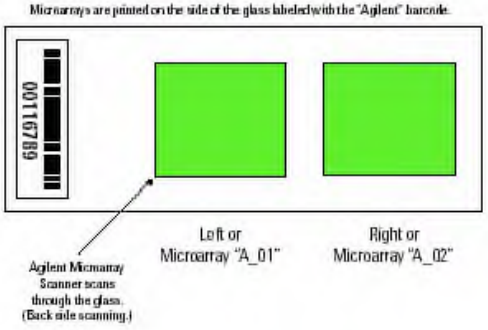
- **Array Fabrication Photolithography**
  - Light activated synthesis
    - synthesize oligonucleotides on glass slides
    - 10^7 copies per oligo in 24 x 24 um square
  - Use 20 pairs of different 25-mers per gene
    - Perfect match and mismatch



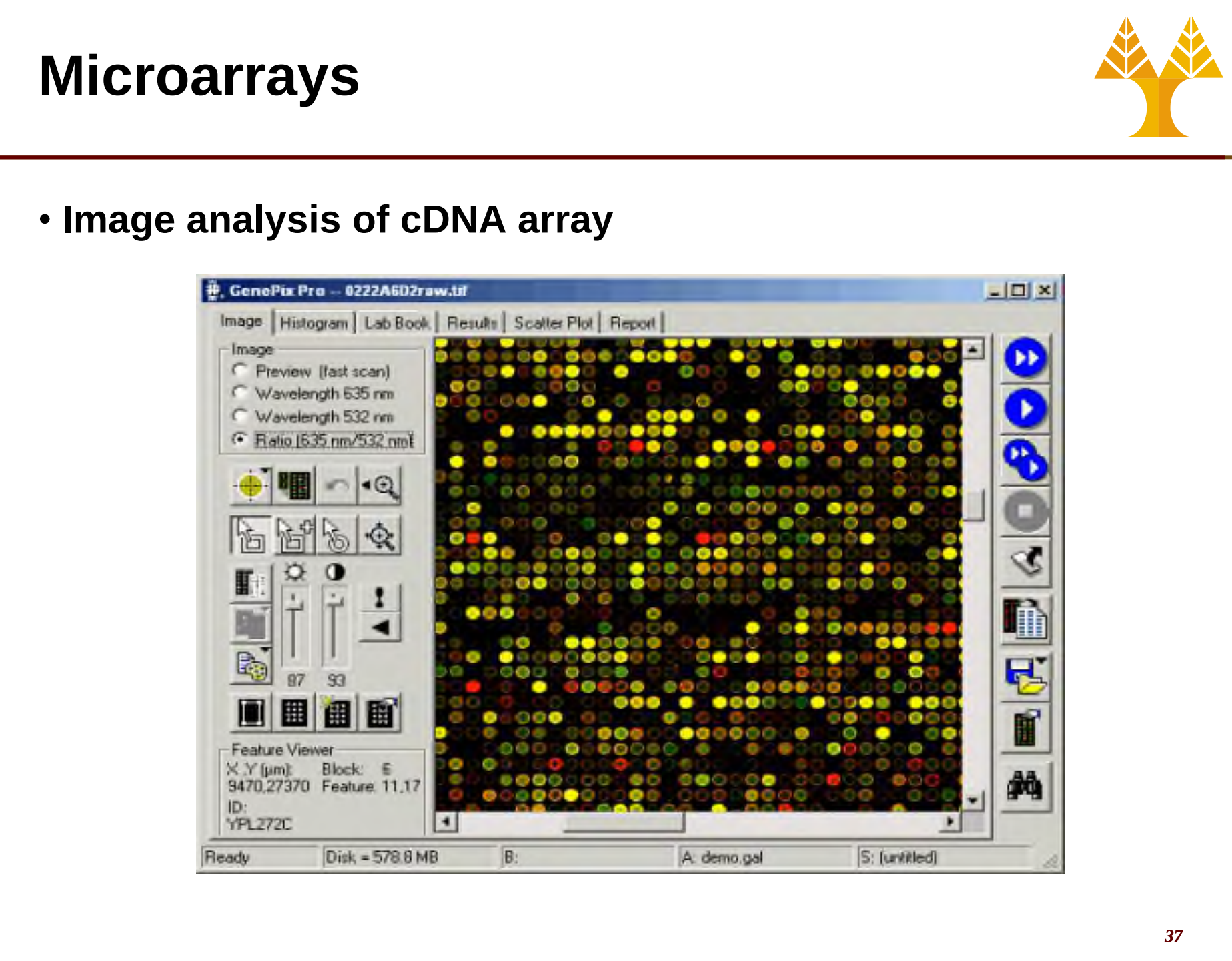
Microarrays

- **Printed Microarrays**
  - Agilent delivering printed 60-mer microarrays in addition to 25-mer formats.
  - The inkjet process uses standard phosphoramidite chemistry to deliver extremely small volumes (picoliters) of the chemicals to be spotted.





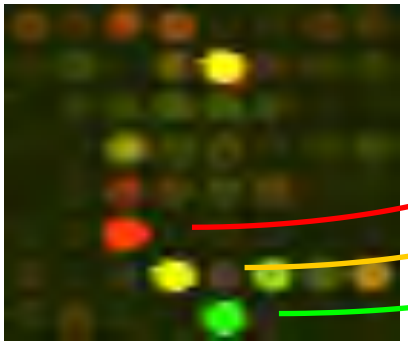
Microarrays



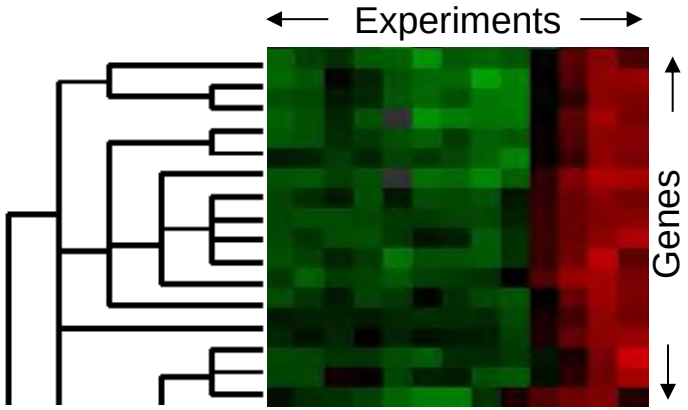
- **Image analysis of cDNA array**

Microarrays

- **Image analysis of cDNA array**



Cy3	Cy5	$\frac{Cy5}{Cy3}$	$\log_2\left(\frac{Cy5}{Cy3}\right)$
~~200~~	10000	50.00	5.64
~~4800~~	4800	1.00	0.00
~~9000~~	300	0.03	-4.91



Underexpressed

Overexpressed

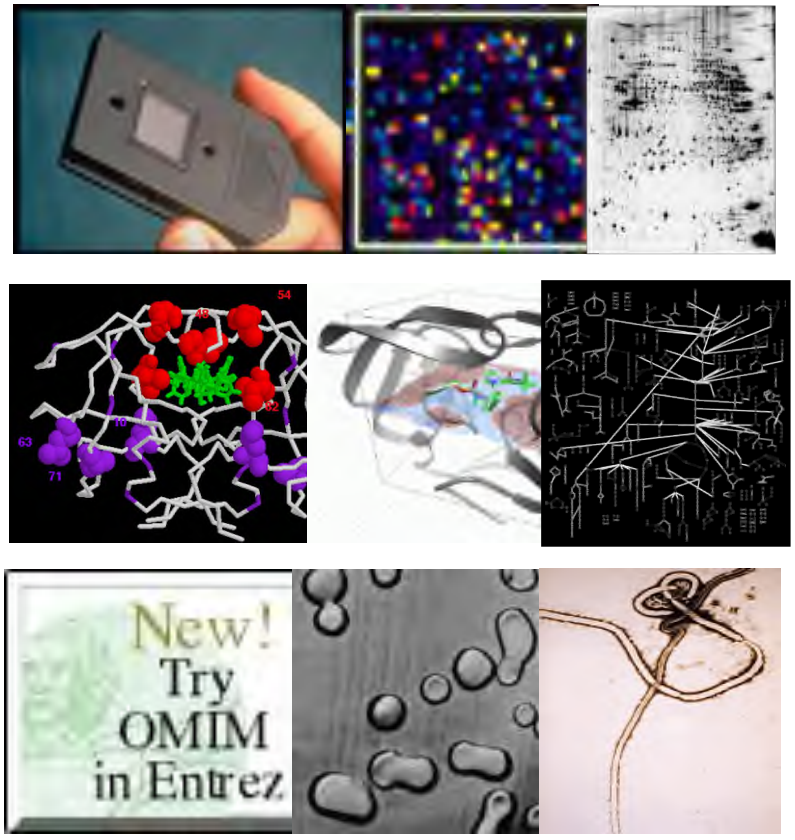
8 4 2 fold 2 4 8

Microarrays

- **Technology is evolving rapidly**
- **Blending of biology, automation, and informatics**
- **New applications are being pursued**
  - Beyond gene discovery into screening, validation, clinical genotyping, etc
- **Microarrays are becoming more broadly available and accepted**
  - Protein Arrays, tissue arrays, etc
  - Diagnostic Applications
- **Diagnostics**
  - Disease detection
  - Tumor classification
  - Patient stratification
  - Intervention therapeutics
- **Treatment and Customized Medicine**

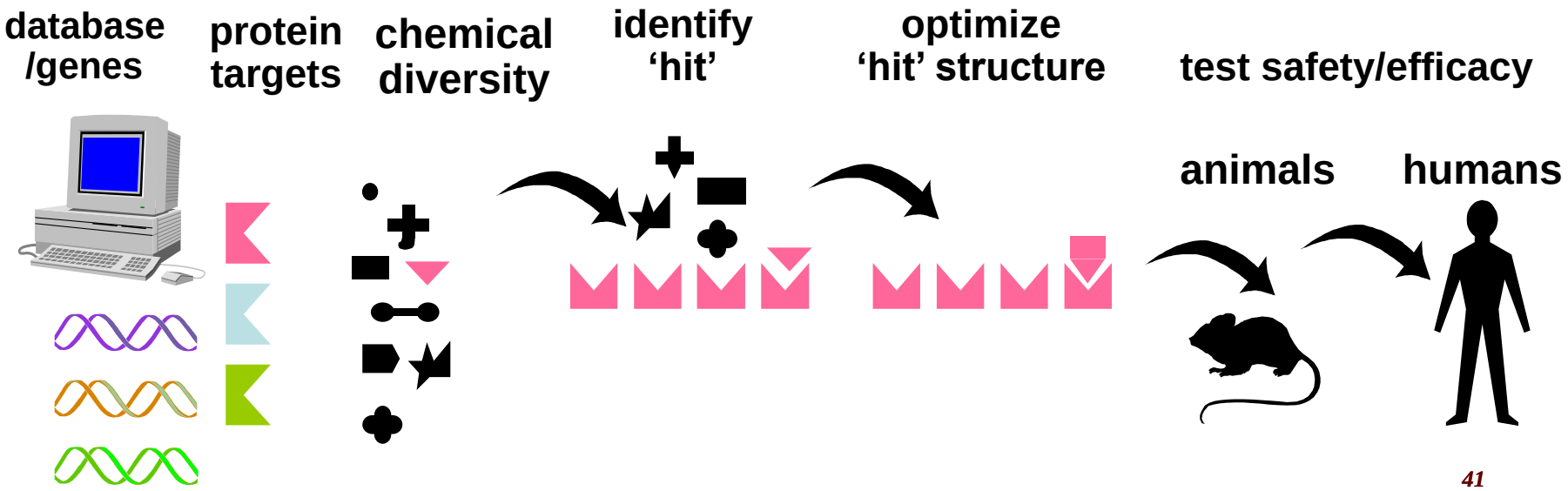
Leveraging Genomic Information in Medicine

- **Novel Diagnostics**
  - Microchips & Microarrays - DNA
  - Gene Expression - RNA
  - Proteomics – Protein
- **Novel Therapeutics**
  - Drug Target Discovery
  - Rational Drug Design
  - Molecular Docking
  - Gene Therapy
  - Stem Cell Therapy
- **Understanding Metabolism**
- **Understanding Disease**
  - Inherited Diseases - OMIM
  - Infectious Diseases
  - Pathogenic Bacteria
  - Viruses



Example Application I: Designing Drugs

- **The aim is to translate new information into new therapies**
- **The Drug Discovery Process**
  - Find genes responsible for disease
  - Determine the protein structure of the target
  - Understanding how proteins bind other molecules
  - Docking & structure modeling
  - Designing inhibitors



Example Application I: Designing Drugs

- **Complexity of Drug Discovery**
  - Finding a Molecule that Satisfies Multiple Criteria
  - “Testing” in silico can reduce the number of candidates significantly (down to 10-100)

Related Fields

- **Medical Informatics**
  - The study and application of computing methods to improve communication, understanding, and management of medical data
  - Generally concerned with how the data is manipulated rather than the data itself
- **Cheminformatics**
  - The study and application of computing methods, along with chemical and biological technology, for drug design and development
- **Pharmacogenomics**
  - The application of genomic methods to identify drug targets
  - For example, searching entire genomes for potential drug receptors, or by studying gene expression patterns in tumors
- **Pharmacogenetics**
  - The use of genomic methods to determine what causes variations in individual response to drug treatments
  - The goal is to identify drugs that may be only be effective for subsets of patients, or to tailor drugs for specific individuals or groups

Bio-inspired Design and Technology

- **Develop new approaches to pressing technological challenges in various fields by exploiting engineering solutions found in nature**
  - Nature had billions of years to optimize its solutions
- **Examples**
  - Computing: genetic algorithms, neural networks, artificial immune networks, etc
  - Materials: smart adaptable materials, unique properties (spider silk, lobster shell, cartilage)
  - Aerodynamics: flight of birds and insects
  - Architecture: mimicking the cooling properties of termite mounds
  - Fabrication: biomorphic mineralization
  - Nanotechnology: self assembling and adapting nanostructures, unique properties