CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS08.html

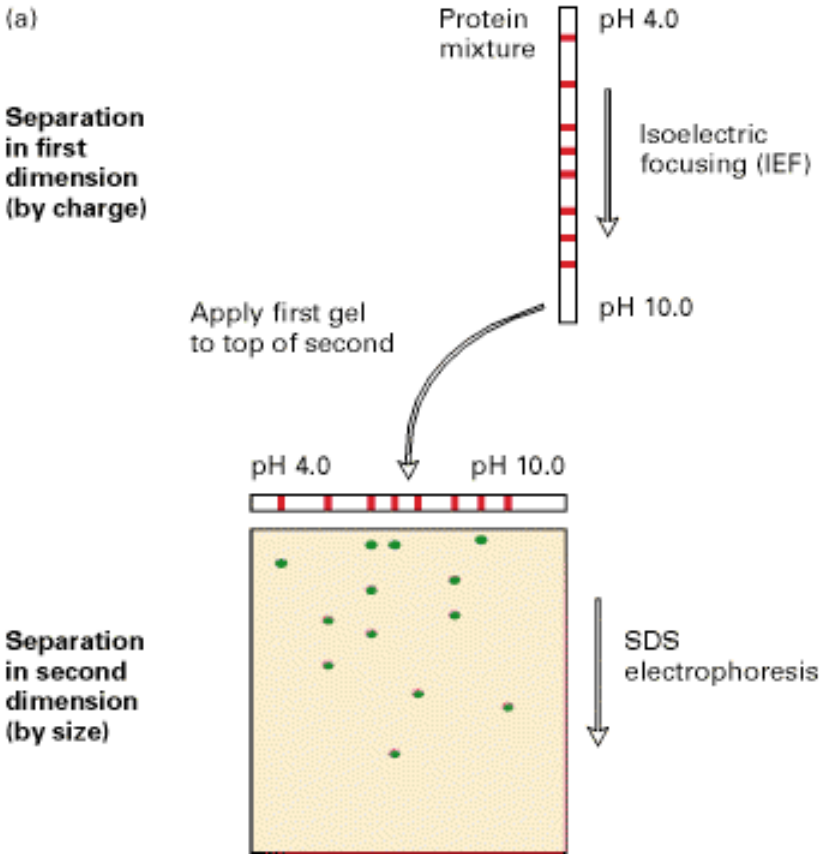
Sequencing (Review by Shendure et al., Nat. Rev. Gen. 2004)

- Goal: ULCS (Ultra low cost sequencing)
  - Personal genome project
- Microelectrophoretic sequencing
  - Electrophoretic separation with single-base resolution
  - Microfabrication, multiplexing, miniaturization technology
- Hybridization sequencing
  - Differential hybridization with short probes
- Cyclic-array sequencing of amplified molecules
  - Multiple cycles of manipulation of spatially separated fragments
  - 2 Types:
    - Pyrosequencing (454 Sequencing)
    - fluorescent in situ sequencing (Solexa and Polony sequencing)
- Non-cyclical, single molecule, real-time methods
  - DNA sent through nanopore & differences in pore conductance measured

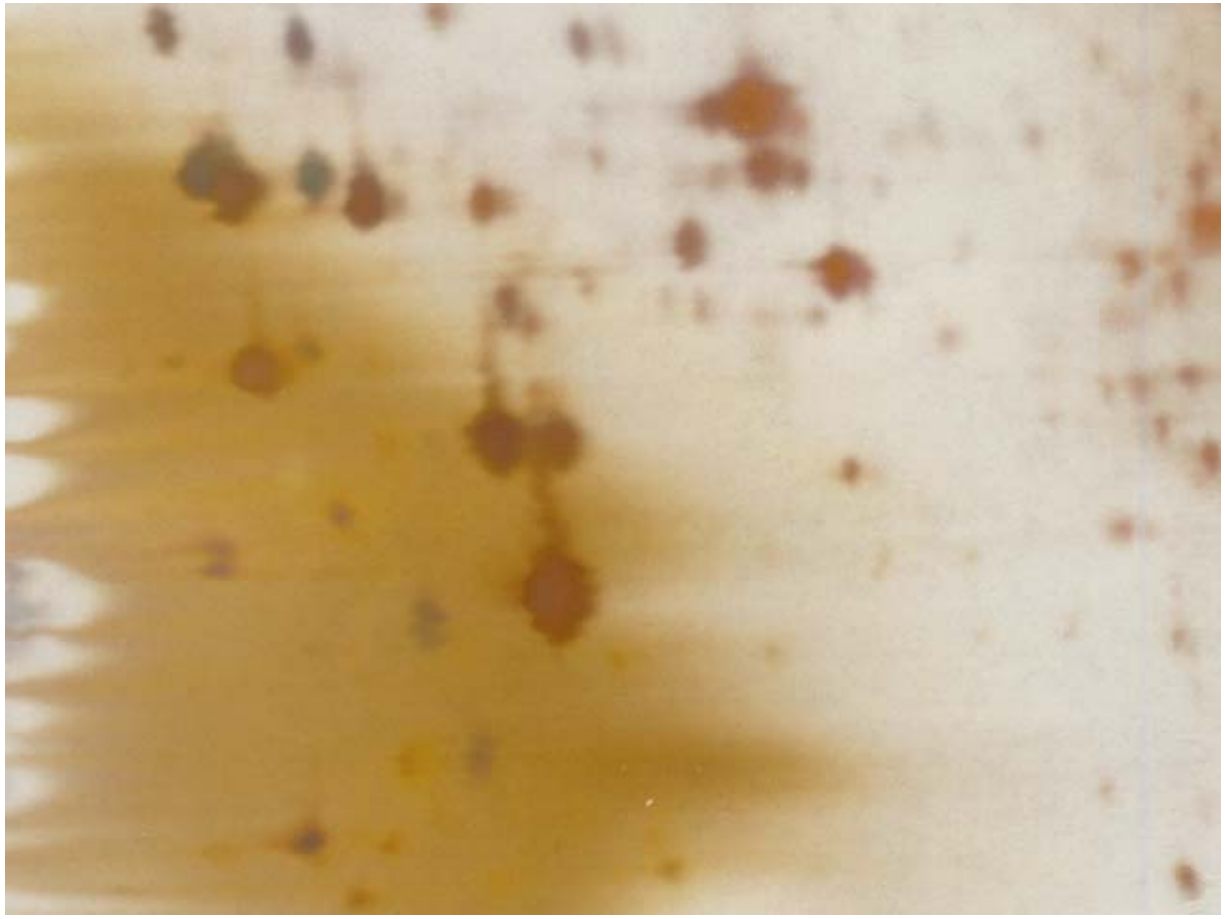
Assembly Software

- Parallel EST alignment engine (http://corba.ebi.ac.uk/EST") with a CORBA interface to alignment database. Can perform ad hoc assemblies. Can act as foundation for CORBA-based EST assembly and editing package. [Parsons, EBI]
- Software using multiple alternative sequence assembly "engines" writing to a common format file [Staden, Cambridge] (http://www.mrc-lmb.cam.ac.uk/pubseq/index.html).
- Phrap (http://bozeman.genome.washington.edu/phrap.docs/phrap.html)
- Assembler (TIGR) for EST and Microbial whole-genome assembly (http://www.tigr.org/softlab/)
- FAK2 and FAKtory (http://www.cs.arizona.edu/people/gene/) [Myers]
- GCG (http://www.gcg.com)
- Falcon [Gryan, Harvard] fast (rascal.med.harvard.edu/gryan/falcon/)
- SPACE, SPASS [Lawrence Berkeley Labs] (http://www-hgc.lbl.gov/inf/space.html)
- CAP 2 [Huang] (http://www.tigem.it/ASSEMBLY/capdoc.html)

2D-Gels


(a)
Protein mixture
pH 4.0
Separation in first dimension (by charge)
Isoelectric focusing (IEF)
pH 10.0
Apply first gel to top of second
pH 4.0
pH 10.0
Separation in second dimension (by size)
SDS electrophoresis

2D Gel Electrophoresis



2D-Gels

First Dimension Methodology of a 2D Gel:

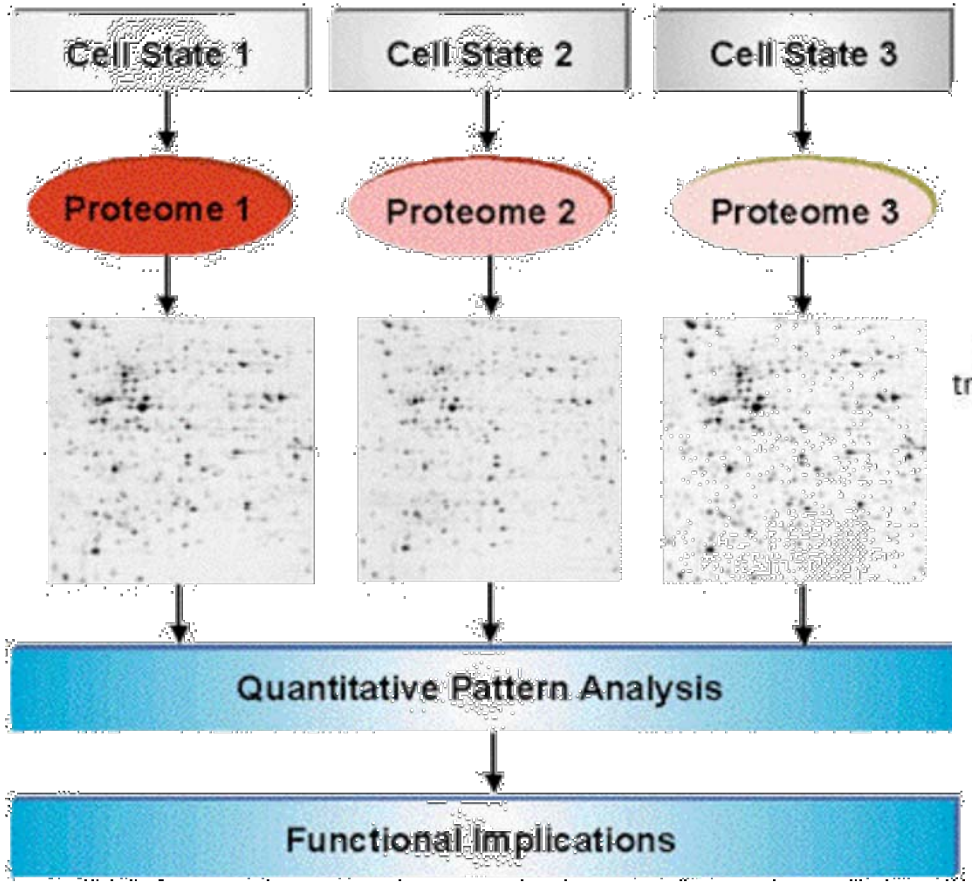
Denatured cell extract layered on a glass tube filled with polyacrylamide saturated with solution of ampholytes, a mixture of polyanionic[(-) charged] and polycationic [(+) charged] molecules. When placed in an electric field, the ampholytes separate and form continuous gradient based on net charge. Highly polyanionic ampholytes will collect at one end of tube, highly polycationic ampholytes will collect at other end. Gradient of ampholytes establishes pH gradient. Charged proteins migrate through gradient until they reach their pI, or isoelectric point, the pH at which the net charge of the protein is zero. This resolves proteins that differ by only one charge.

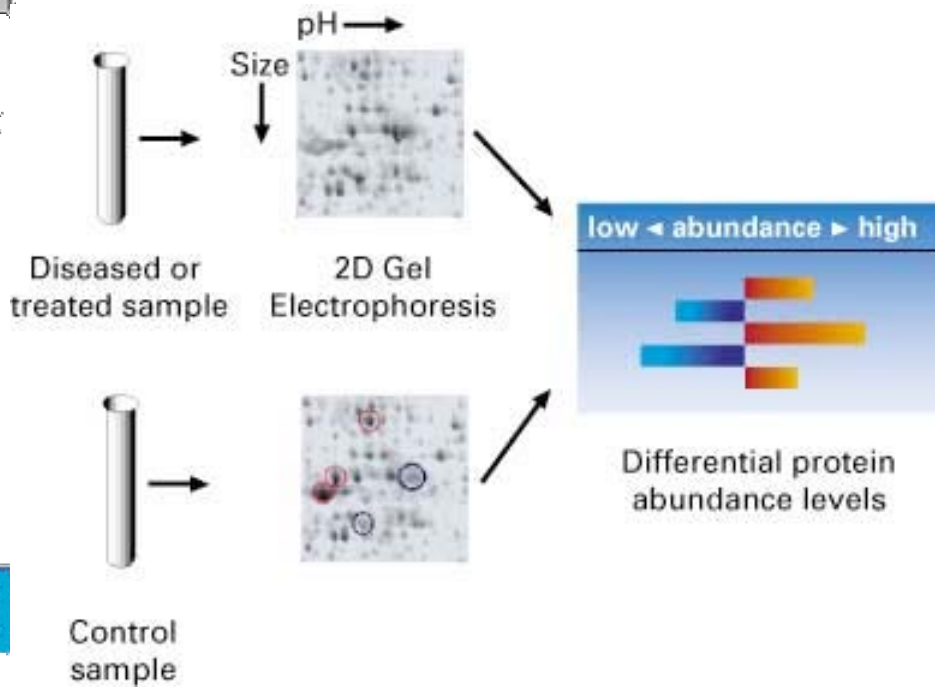
Entering the Second Dimension:

Proteins that were separated on IEF gel are next separated in the second dimension based on their molecular weights. The IEF gel is extruded from tube and placed lengthwise in alignment with second polyacrylamide gel slab saturated with SDS. When an electric field is imposed, the proteins migrate from IEF gel into SDS slab gel and then separate according to mass. Sequential resolution of proteins by their charge and mass can give excellent separation of cellular proteins. As many as 1000 proteins can be resolved simultaneously.

*Some information was taken from Lodish *et al.* Molecular Cell Biology.

2D-gels

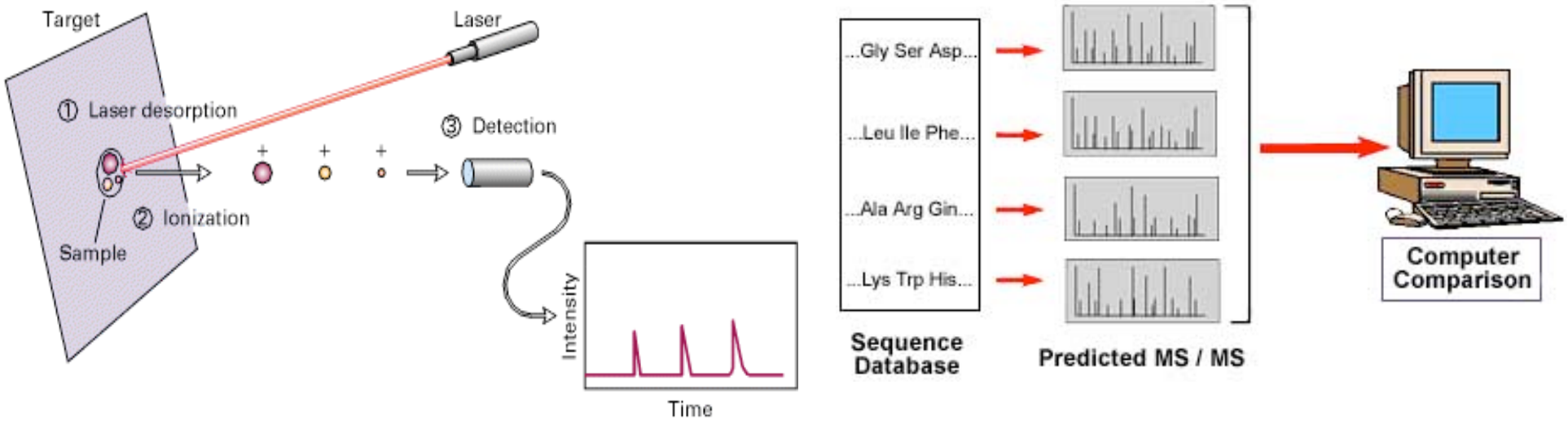




Comparing Proteomes For Differences in Protein Expression

Comparing Different Sample Types For Changes in Protein Levels

Mass Spectrometry



Mass Spectrometry

- **Mass measurements By Time-of-Flight**
  Pulses of light from laser ionizes protein that is absorbed on metal target. Electric field accelerates molecules in sample towards detector. The time to the detector is inversely proportional to the mass of the molecule. Simple conversion to mass gives the molecular weights of proteins and peptides.
- **Using Peptide Masses to Identify Proteins:**
  One powerful use of mass spectrometers is to identify a protein from its peptide mass fingerprint. A peptide mass fingerprint is a compilation of the molecular weights of peptides generated by a specific protease. The molecular weights of the parent protein prior to protease treatment and the subsequent proteolytic fragments are used to search genome databases for any similarly sized protein with identical or similar peptide mass maps. The increasing availability of genome sequences combined with this approach has almost eliminated the need to chemically sequence a protein to determine its amino acid sequence.

Genomics

- Study of all genes in a genome, or comparison of whole genomes.
  - Whole genome sequencing
  - Whole genome annotation & Functional genomics
  - Whole genome comparison
    - PipMaker: uses BLASTZ to compare very long sequences (> 2Mb); http://www.cse.psu.edu/pipmaker/
    - Mummer: used for comparing long microbial sequences (uses Suffix trees!)

Genomics

- Study of all genes in a genome
  - Gene Expression
    - Microarray experiments & analysis
      - Probe design (CODEHOP)
      - Array image analysis (CrazyQuant)
      - Identifying genes with significant changes (SAM)
      - Clustering

Comparative Genomics

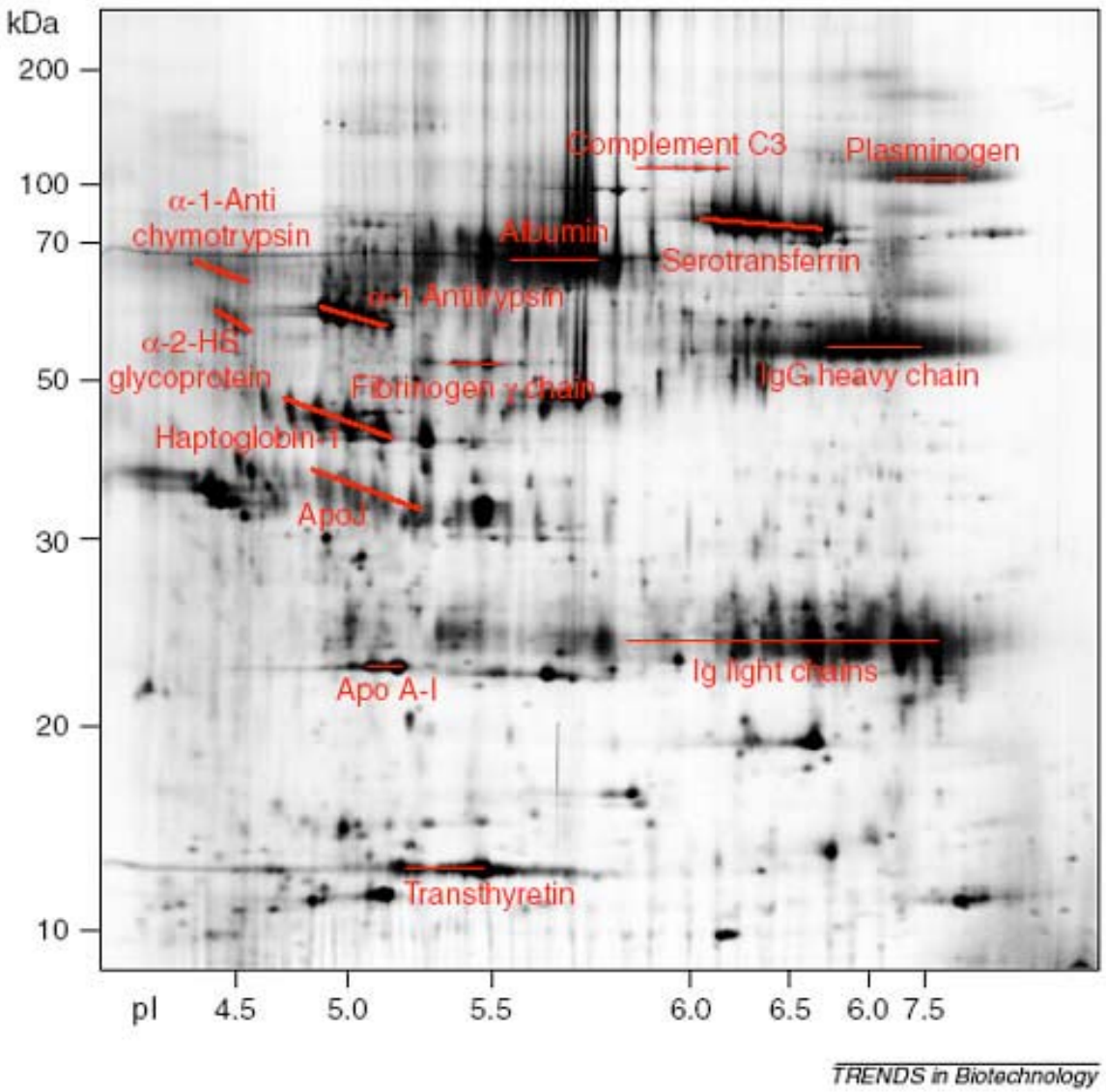
- Comparison of whole genomes.
  - Whole genome sequencing
  - Whole genome annotation & Functional genomics
  - Whole genome comparison
    - PipMaker, MultiPipMaker, EnteriX: PipMaker uses BLASTZ to compare very long sequences (> 2Mb); http://www.cse.psu.edu/pipmaker/
    - Mummer: used for comparing long microbial sequences (uses Suffix trees!)
    - Many more!

Databases for Comparative Genomics

- PEDANT useful resource for standard questions in comparative genomics. For e.g., *how many known proteins in XXX have known 3-d structures, how many proteins from family YYY are in ZZZ, etc.*
- COGs Clusters of orthologous groups of proteins.
- MBGD Microbial genome database searches for homologs in all microbial genomes

Proteomics

- Study of all proteins in a genome, or comparison of whole genomes.
  - Whole genome annotation & Functional proteomics
  - Whole genome comparison
  - Protein Expression: 2D Gel Electrophoresis


kDa
200
100
70
50
30
20
10
pI
4.5
5.0
5.5
6.0
6.5
6.0
7.5
Complement C3
Plasminogen
α-1-Anti
chymotrypsin
Albumin
Serotransferrin
α-1 Antitrypsin
α-2-HS
glycoprotein
Fibrinogen γ chain
IgG heavy chain
Haptoglobin-1
ApoJ
Ig light chains
Apo A-I
Transthyretin
TRENDS in Biotechnology

Other Proteomics Tools

From ExPASy/SWISS-PROT:

- **AACompIdent** identify proteins from aa composition

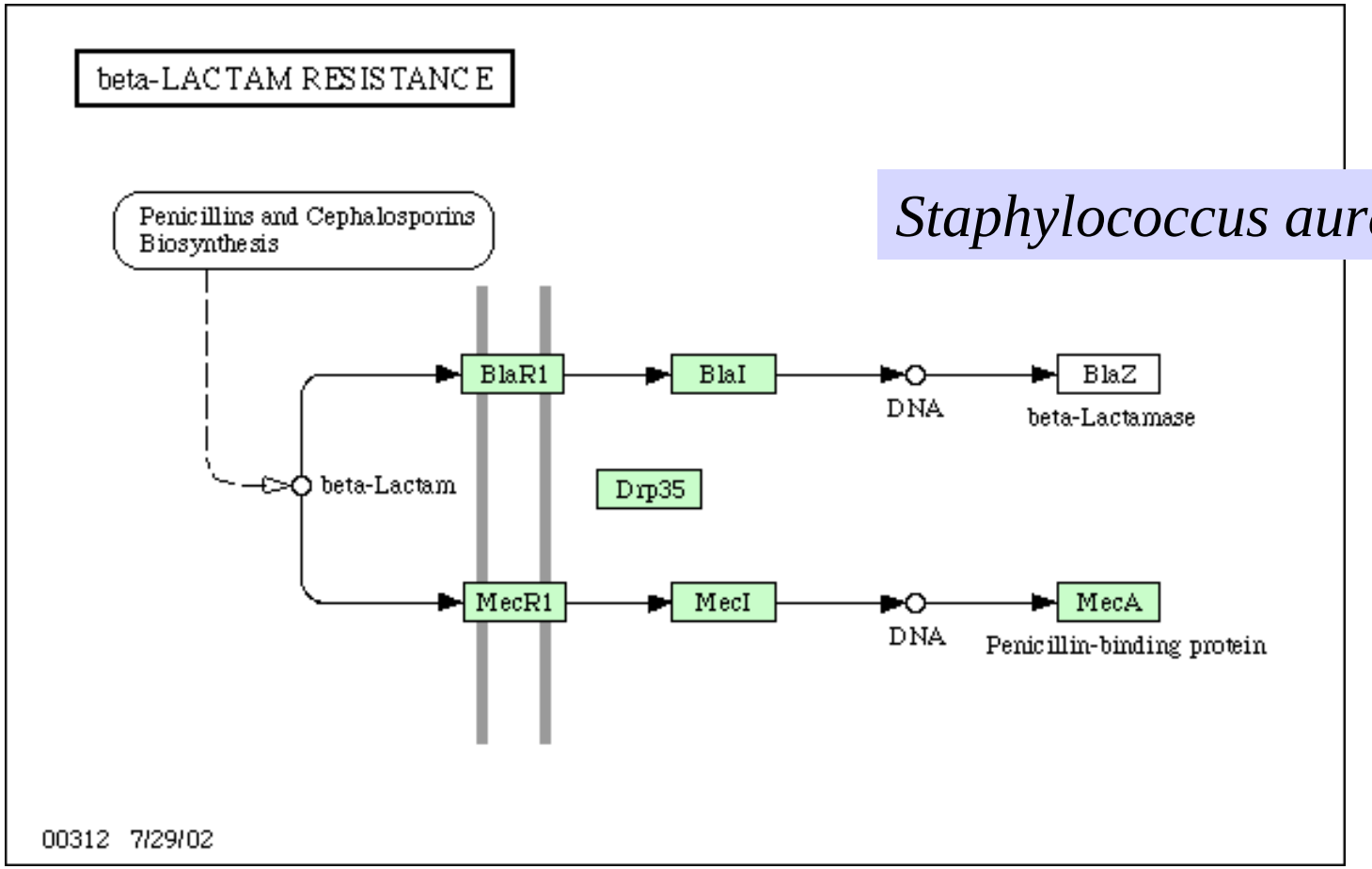
[Input: aa composition, isoelectric point, mol wt., etc. Output: proteins from DB]

- **AACompSim** compares proteins aa composition with other proteins
- **MultIdent** uses mol wt., mass fingerprints, etc. to identify proteins
- **PeptIdent** compares experimentally determined mass fingerprints with theoretically determined ones for all proteins
- **FindMod** predicts post-translational modifications based on mass difference between experimental and theoretical mass fingerprints.
- **PeptideMass** theoretical mass fingerprint for a given protein.
- **GlycoMod** predicts oligosaccharide modifications from mass difference
- **TGREASE** calculates hydrophobicity of protein along its length

Gene Networks & Pathways

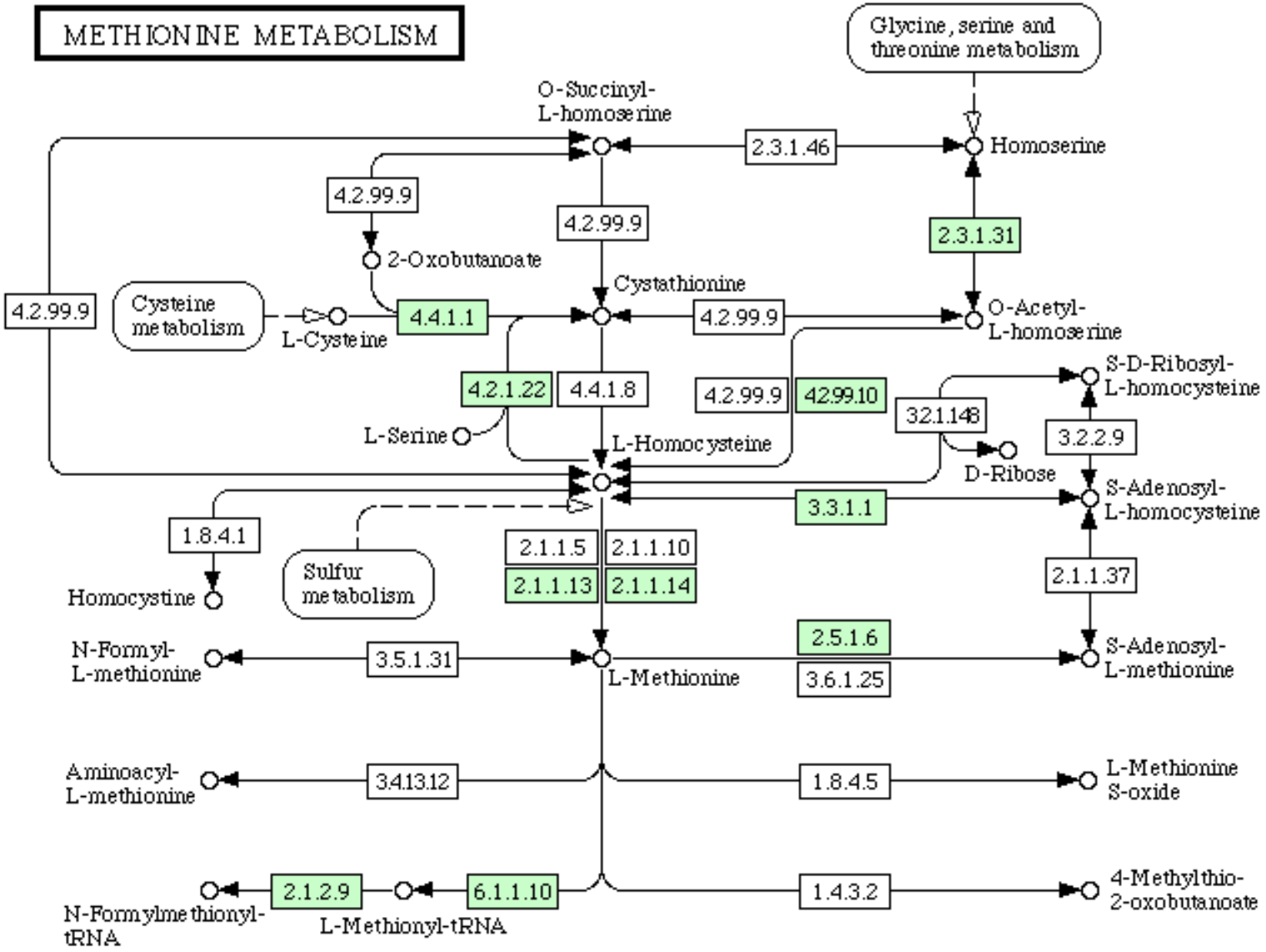
- Genes & Proteins act in concert and therefore form a complex network of dependencies.

Pathway Example from KEGG



Staphylococcus aureus

Pseudomonas aeruginosa



00271 8/6/02

STSs and ESTs

- Sequence-Tagged Site: short, unique sequence
- Expressed Sequence Tag: short, unique sequence from a coding region
  - 1991: 609 ESTs [Adams et al.]
  - June 2000: 4.6 million in dbEST
  - Genome sequencing center at St. Louis produce 20,000 ESTs per week.

What Are ESTs and How Are They Made?

- Small pieces of DNA sequence (usually 200 - 500 nucleotides) of low quality.
- Extract mRNA from cells, tissues, or organs and sequence either end. Reverse transcribe to get cDNA (5' EST and 3'EST) and deposit in EST library.
- Used as "**tags**" or markers for that gene.
- Can be used to identify similar genes from other organisms (Complications: variations among organisms, variations in genome size, presence or absence of **introns**).
- 5' ESTs tend to be more useful (cross-species conservation), 3' EST often in UTR.

DNA Markers

- Uniquely identifiable DNA segments.
- Short, <500 nucleotides.
- Layout of these markers give a map of genome.
- Markers may be polymorphic (variations among individuals). Polymorphism gives rise to alleles.
- Found by PCR assays.

Polymorphisms

- Length polymorphisms
  - Variable # of tandem repeats (VNTR)
  - Microsatellites or short tandem repeats
  - Restriction fragment length polymorphism (RFLP) caused by changes in restriction sites.
- Single nucleotide polymorphism (SNP)
  - Average once every ~100 bases in humans
  - Usually biallelic
  - dbSNP database of SNPs (over 100,000 SNPs)
  - ESTs are a good source of SNPs

SNPs

- SNPs often act as "disease markers", and provide "genetic predisposition".
- SNPs may explain differences in drug response of individuals.
- Association study: study SNP patterns in diseased individuals and compare against SNP patterns in normal individuals.
- Many diseases associated with SNP profile.