CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS11.html

Microarray Data

Gene	Expression Levels	
	Sample A CONTROL	Sample B TREATMENT
Gene1		
Gene2		
Gene3		
…		

Microarray Analysis

- Is Gene X upregulated? Downregulated? Had no change in expression levels?
  - Genes are represented by probes
  - Experiments may have repeats
- **NULL HYPOTHESIS**
  - *There is no change in gene expression levels for Gene X between Control and Treatment*

Accept/Reject H_0 (Null Hypothesis)?

- P-value thresholds
  - P-value is probability of data assuming H_0 holds
  - P-value threshold of 0.05 means probability of error when H_0 is rejected is 5%
- Fold change
  - If no repeats are done
- t-Test
  - Parametric
  - Non-parametric
    - Wilcoxon rank sum

Multiple Testing & Type I Errors

- Type I Error of 0.05 means that there is a 5% error in prediction of FN by t-Test. IMPLICATIONS?
  - If N=1000 genes & d=40 are differentially expressed (DE), then …
    - 960 X 0.05 = 48 FPs
    - There are more FPs than TPs
    - Type I error and correcting for multiple hypothesis testing are connected

Multiple Test Corrections

- Bonferroni correction
  - Use type I error = α / g = FWER = 0.05/1000
    - Family-wise Error (FWER)
    - Too Conservative! Also reduce true positives!
- Other less conservative corrections possible
  - Sidak correction, Westfall-Young correction, …
- Using False Discovery Rate (FDR) [Benjamini & Hochberg '95, Storey '02 & '03]
  - Earlier: 5% of all tests will result in FPs
  - With FDR adjusted p-value (or q-value): 5% of significant tests will result in false positives.

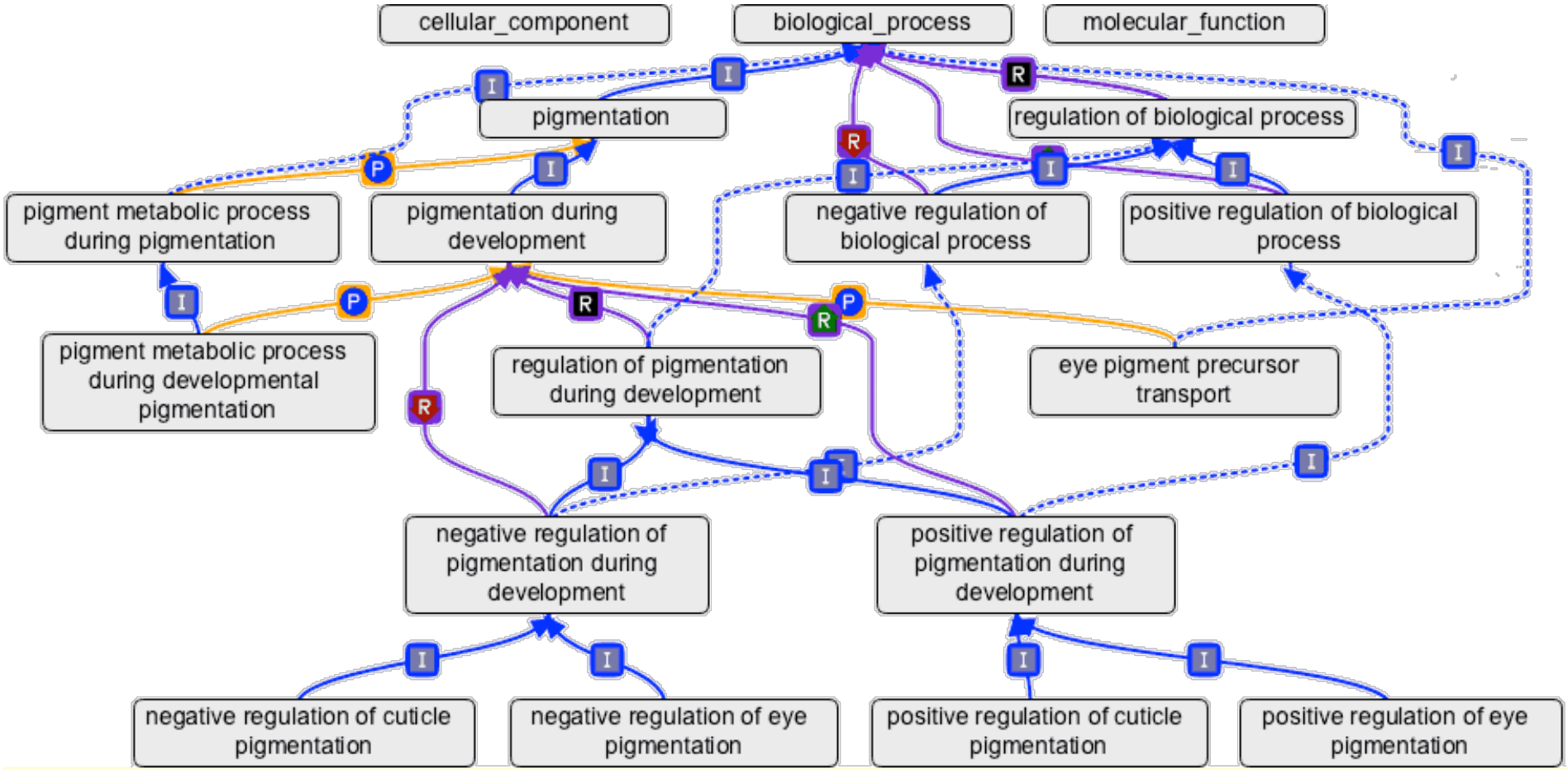
Annotation

- Annotation: association of raw sequence data and useful biological information.
- Integrates:
  - computational analyses,
  - auxiliary biological data and
  - biological expertise.

Gene Ontology

- Ontology: entities and their relationships
- Ontology: representation of knowledge as a set of concepts within a domain
  - Provides a shared vocabulary
- Gene Ontology (GO): project to
  - Standardize representation of gene & gene product attributes across species and DBs
  - Provide controlled vocabulary for data and features
  - Provide tools to access and process knowledgebase
  - Recent: Renal and Cardiovascular GO

GO Hierarchy is a Graph



- 3 Hierarchies: Cellular Component, Biological Process, Molecular Function

GO Terms

- Every term has a name
  - E.g., ribosome, glucose transport, amino acid binding
- Every term has a unique accession number or ID
  - E.g., GO:0005125, GO:0060092
- Terms may be related by relationships:
  - **is-a**: E.g., GO:0015758 glucose transport is a GO:0015749 monosaccharide transport
  - **part-of**: E.g., GO:0031966 mitochondrial membrane is part of GO:0005740 mitochondrial envelope
  - **regulates**: E.g., GO:0006916 anti-apoptosis regulates GO:0012501 programmed cell death

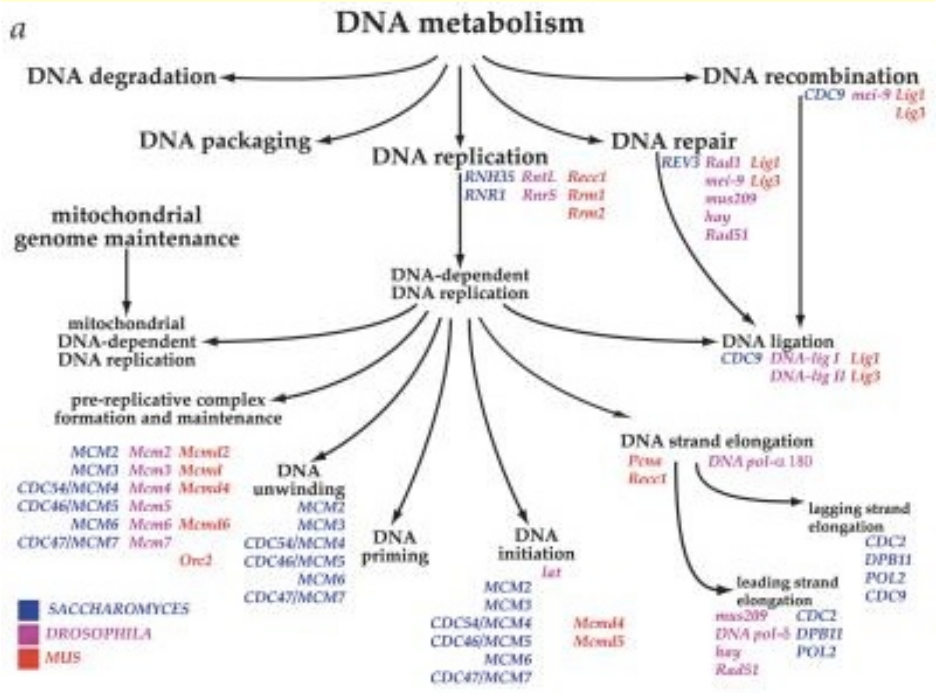
Sample GO Term

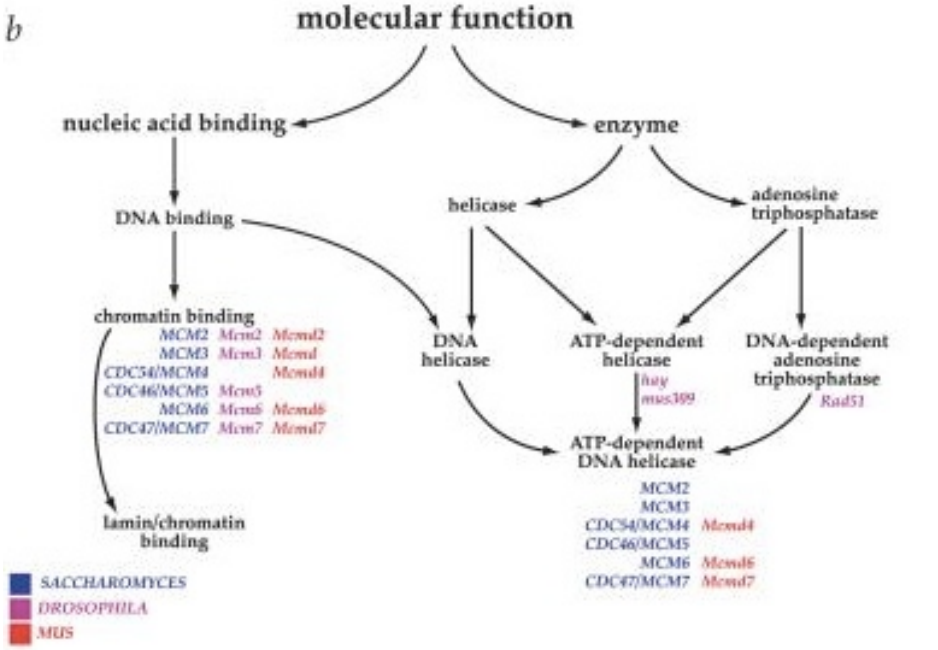
id: GO:0016049
name: **cell growth**
namespace: biological_process
def: "The process in which a cell irreversibly increases in size over time by accretion and biosynthetic production of matter similar to that already present." [GOC:ai]
subset: goslim_generic
subset: goslim_plant
subset: gosubset_prok
synonym: "cell expansion" RELATED []
synonym: "cellular growth" EXACT []
synonym: "growth of cell" EXACT []
is_a: GO:0009987 ! cellular process
is_a: GO:0040007 ! growth
relationship: part_of GO:0008361 ! regulation of cell size

The 3 hierarchies

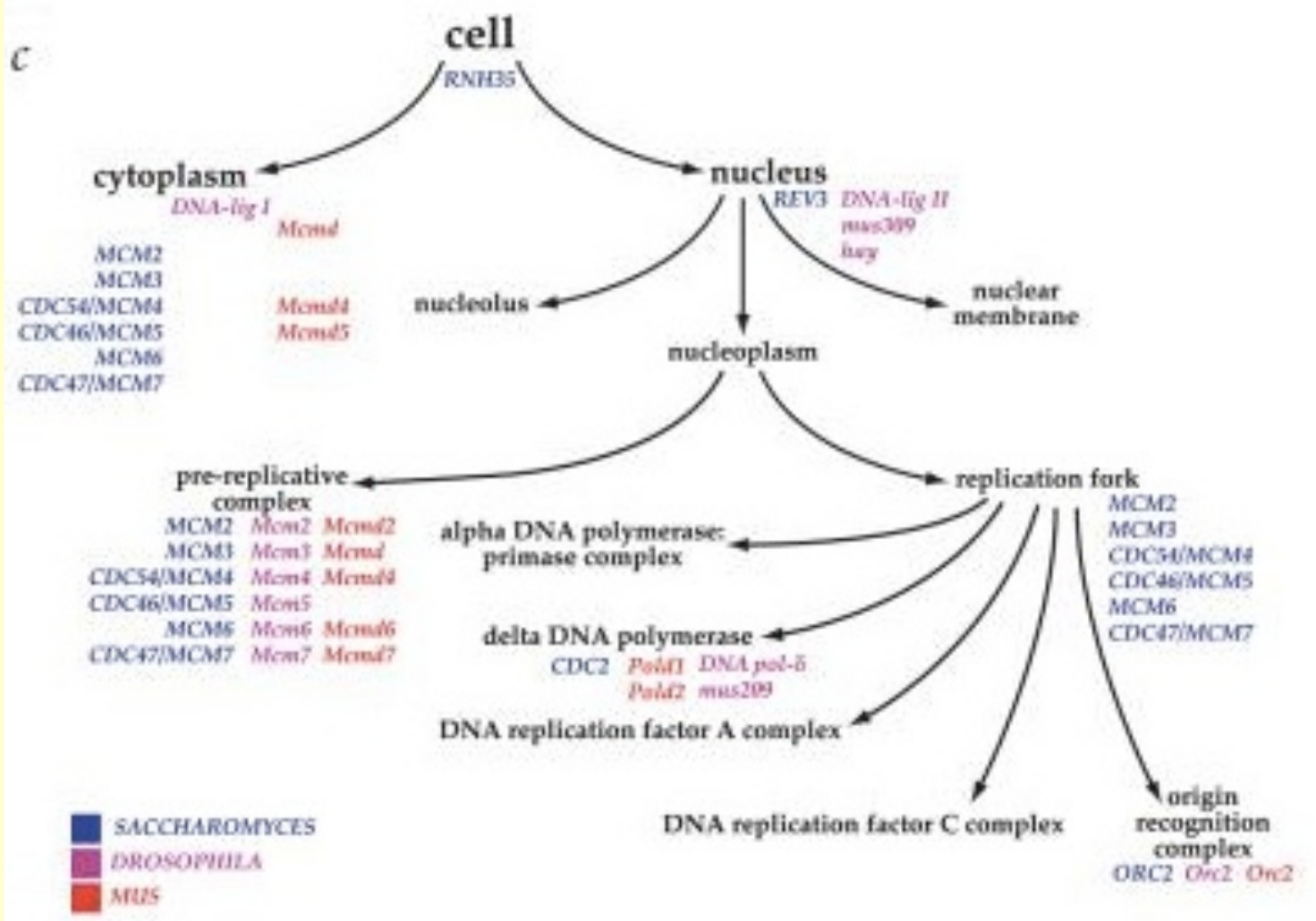
- Cellular Component: A component of the cell, i.e., location
  - E.g., rough endoplasmic reticulum, nucleus, ribosome, proteasome
- Biological Process: A biological process is series of events accomplished by one or more ordered assemblies of molecular functions.
  - E.g., cellular physiological process, signal transduction, pyrimidine metabolic process, alpha-glucoside transport
- Molecular Function: Activities, such as catalytic or binding activities, that occur at the molecular level
  - E.g., catalytic activity, transporter activity, binding; adenylate cyclase activity, Toll receptor binding

Biological Process & Molecular Function

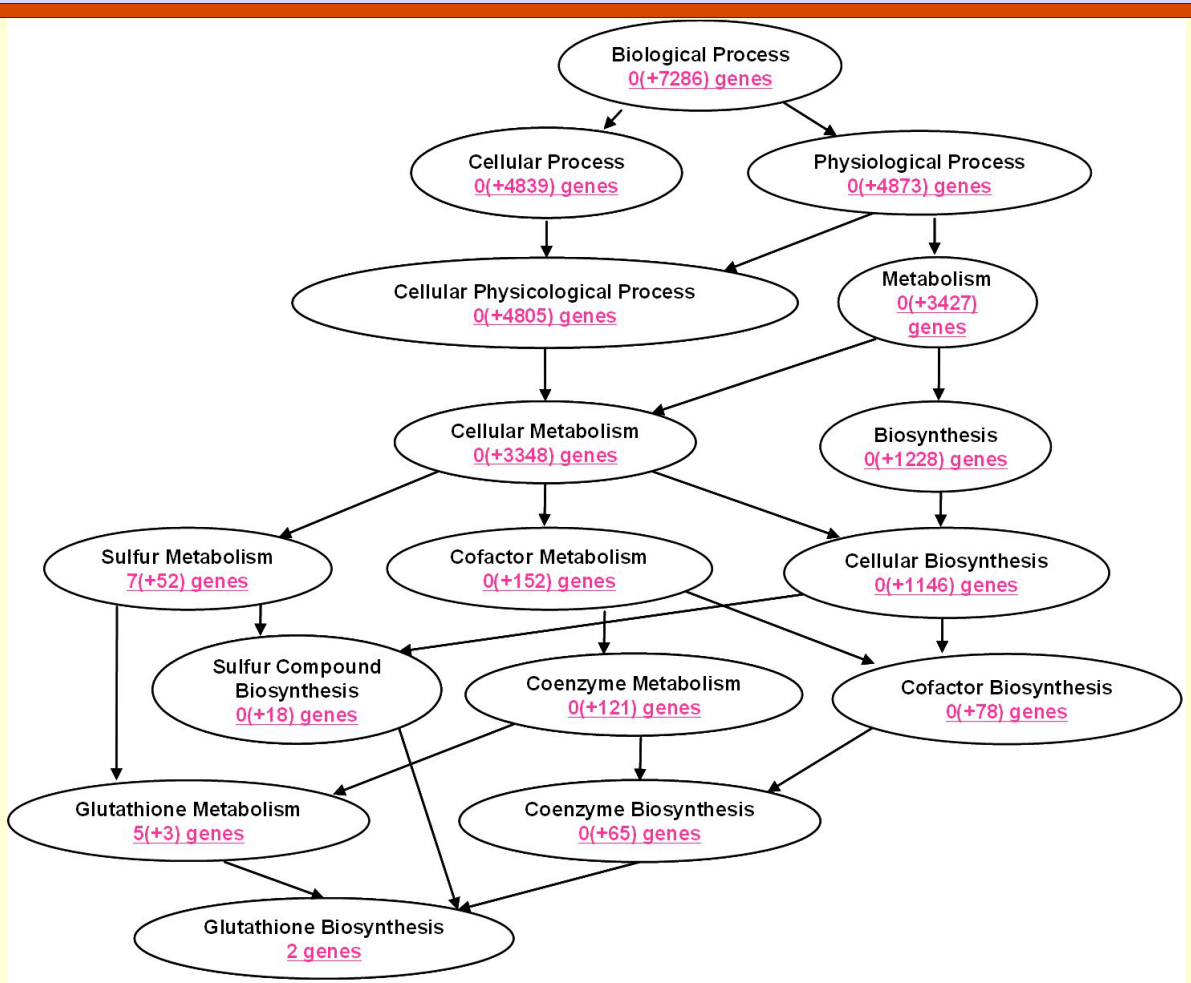




Cellular Component



Go Hierarchy is a Graph: Yeast



Utility of GO Annotations

- Assign annotations to new genes based on their similarities or proximities to annotated genes
- **Enrichment Analysis**: Overrepresentation or underrepresentation in sets of genes
  - 'Developmental Process' was most significantly overrepresented GO term (P = 0.0006), involving 26% of all regulated genes.
  - P-value = $\sum_{i=q}^{m} \frac{\binom{m}{i}\binom{t-m}{k-i}}{\binom{t}{k}}$

Example [Zheng et al., BMC Gen 2010]

Results: Zebrafish were treated with zinc-depleted and zinc-adequate conditions for 2 weeks. Gill samples were collected at 5 time points and transcriptome changes analysed in quintuplicate using a microarray. A total of 333 genes showed differential regulation by zinc depletion (fold-change > 1.8; adjusted P-value < 0.1; 10% FDR). Down-regulation was dominant at most time points and distinct sets of genes were regulated at different stages. GO enrichment analysis showed 'Developmental Process' as the most significantly overrepresented GO term (P = 0.0006), involving 26% of all regulated genes. Other significant terms related to development, cell cycle, cell differentiation, gene regulation, butanoate metabolism, lysine degradation, protein tyrosin phosphatases, nucleobase, nucleoside and nucleotide metabolism, and cellular metabolic processes. Network analysis of the temporal expression profile indicated that transcription factors *foxl1, wt1, nr5a1, nr6a1*, and especially, *hnf4a* may be key coordinators of the homeostatic response to zinc depletion.

Networks

Genes & Proteins form complex network of dependencies

- Regulatory Networks
  - Edge from TFs to genes they regulate
- Protein-protein interaction (PPI) Networks
- Other Networks: KEGG
  - Metabolic Pathways
  - Genetic Info Processing
  - Environmental Info Processing
  - Cellular Processes
  - Organismal Systems
  - Human Disease, …

http://www.genome.jp/kegg/

KEGG Metabolic Pathways

- Carbohydrate Metabolism
  - Glycolysis, citrate, pyruvate, starch, sucrose, ascorbate, …
- Energy Metabolism
  - Photosynthesis, carbon & nitrogen fixation, sulfur & methane metabolism, …
- Lipid Metabolism
  - Biosynthesis of fatty acid, steroid, ketone, bile acid, …
- Nucleotide Metabolism
- Amino Acid Metabolism
- Metabolism of other amino acids
- Glycan Biosynthesis & Metabolism
- Metabolism of Cofactors and Vitamins
- Metabolism of Terpenoids & Polyketides
- Biosynthesis of Other Secondary Metabolites
- Xenobiotics Biodegradation and Metabolism

KEGG: Info Processing

Genetic Info Processing

- Transcription
- Translation
- Folding, Sorting & Degradation
- Replication & Repair

Environmental Info Processing

- Membrane Transport
- Signal Transduction
- Signaling Molecules & Interaction

KEGG: Misc Networks

Organismal Systems

- Immune Systems
- Endocrine, Circulatory
- Digestive, Excretory
- Nervous, Sensory
- Development
- Environmental Adaptation

Cellular Processes

- Transport & Catabolism
- Cell Motility
- Cell Growth & Death
- Cell Communication

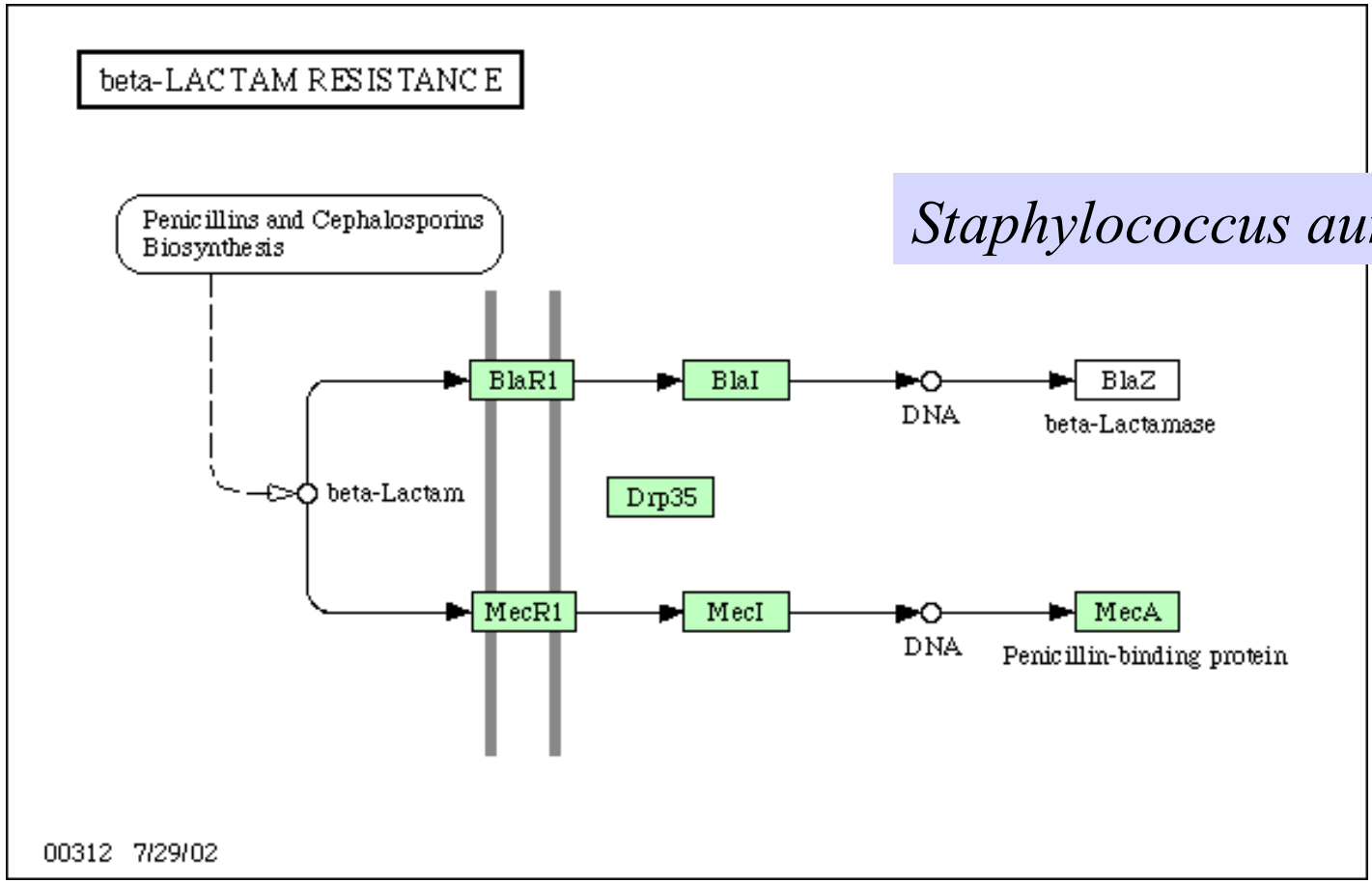
KEGG: More Networks …

Disease

- Cancers
- Immune System Diseases
- Neurodegenerative
- Cardiovascular
- Metabolic
- Infectious

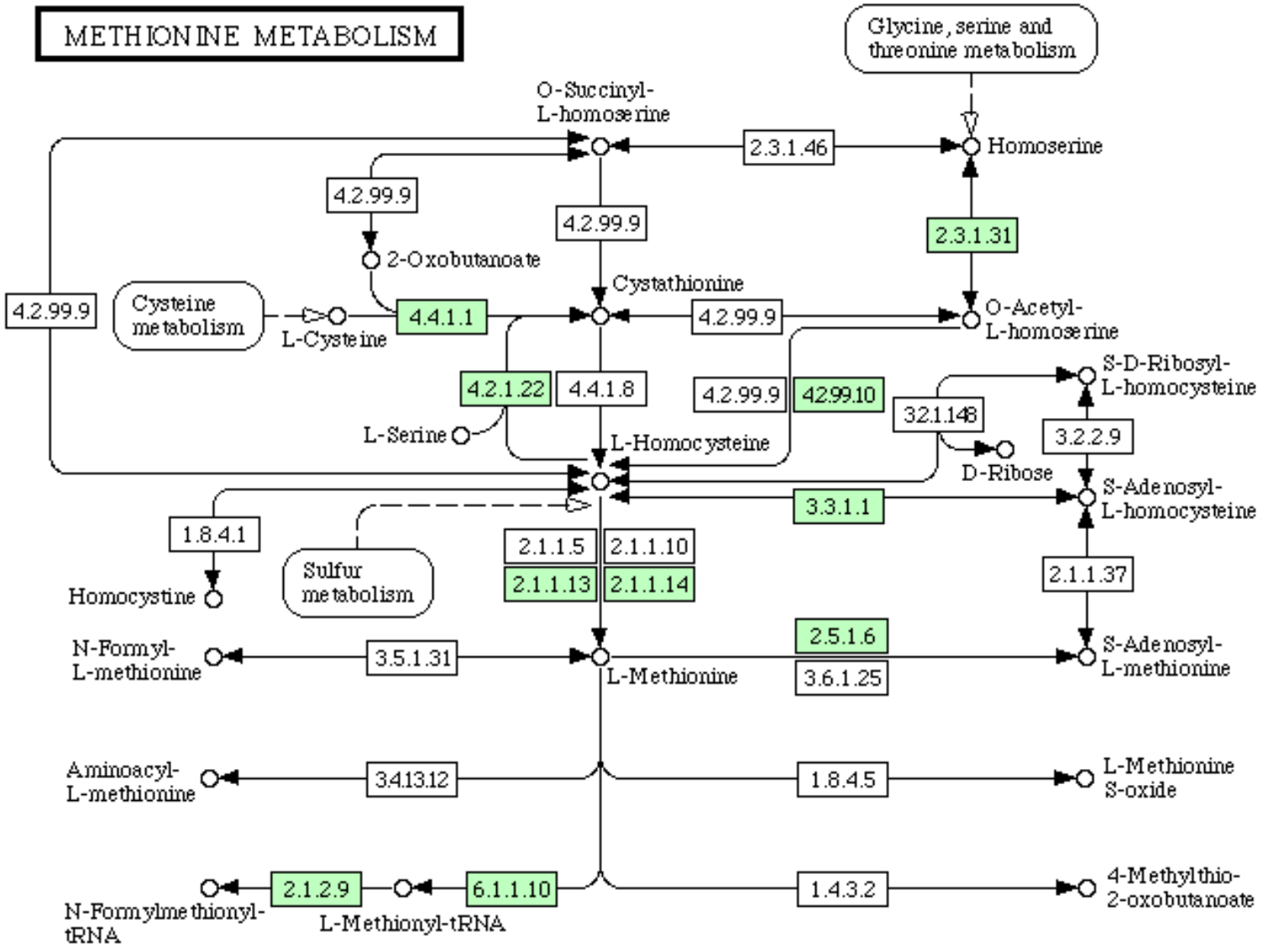
- Drugs
- Antibiotics
- Chronology: Antineoplastics, nervous system agents, misc., …
- Target-based: GPCRs, Nuclear, Ion Channels, Enzymes
- Structure-based
- Skeleton-based

Pathway Example from KEGG



Staphylococcus aureus

Pseudomonas aeruginosa



Omics

- Genomics
- Proteomics
- Transcriptomics
- Metabolomics
- Glycomics
- Cytomics
- Lipidomics
- …

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
  - All aspects of total gene content
  - Gene Expression
    - Microarray experiments & analysis
    - RNA-Seq

Comparative Genomics

- Comparison of whole genomes.
  - Sequence comparison
  - Content comparison
  - Functional annotation comparison
  - …

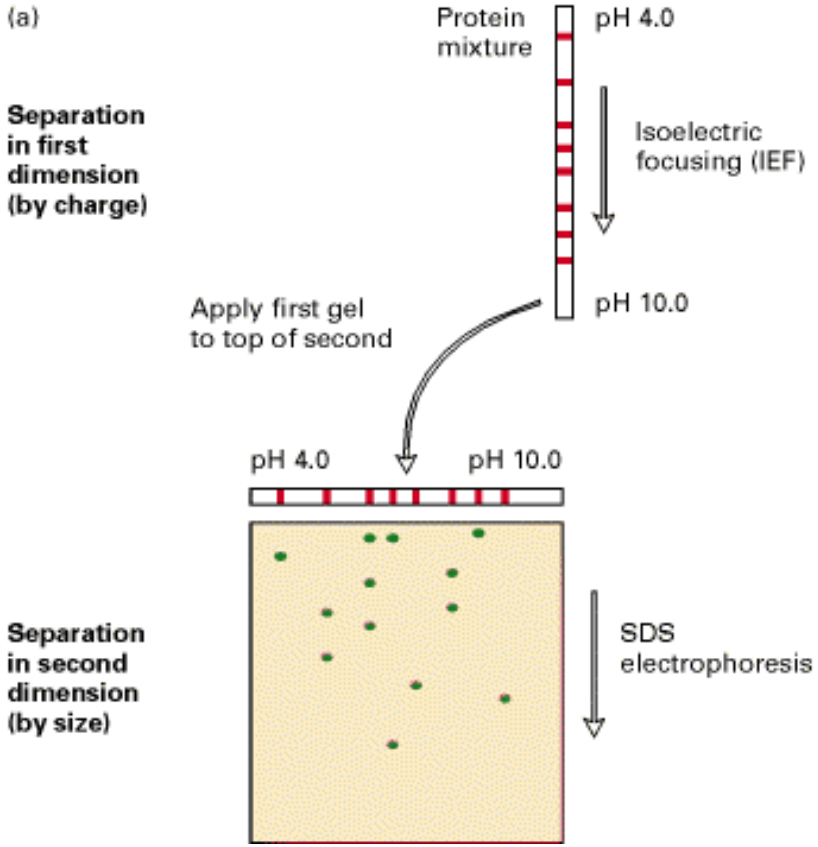
Databases for Comparative Genomics

- GreenGenes
- 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

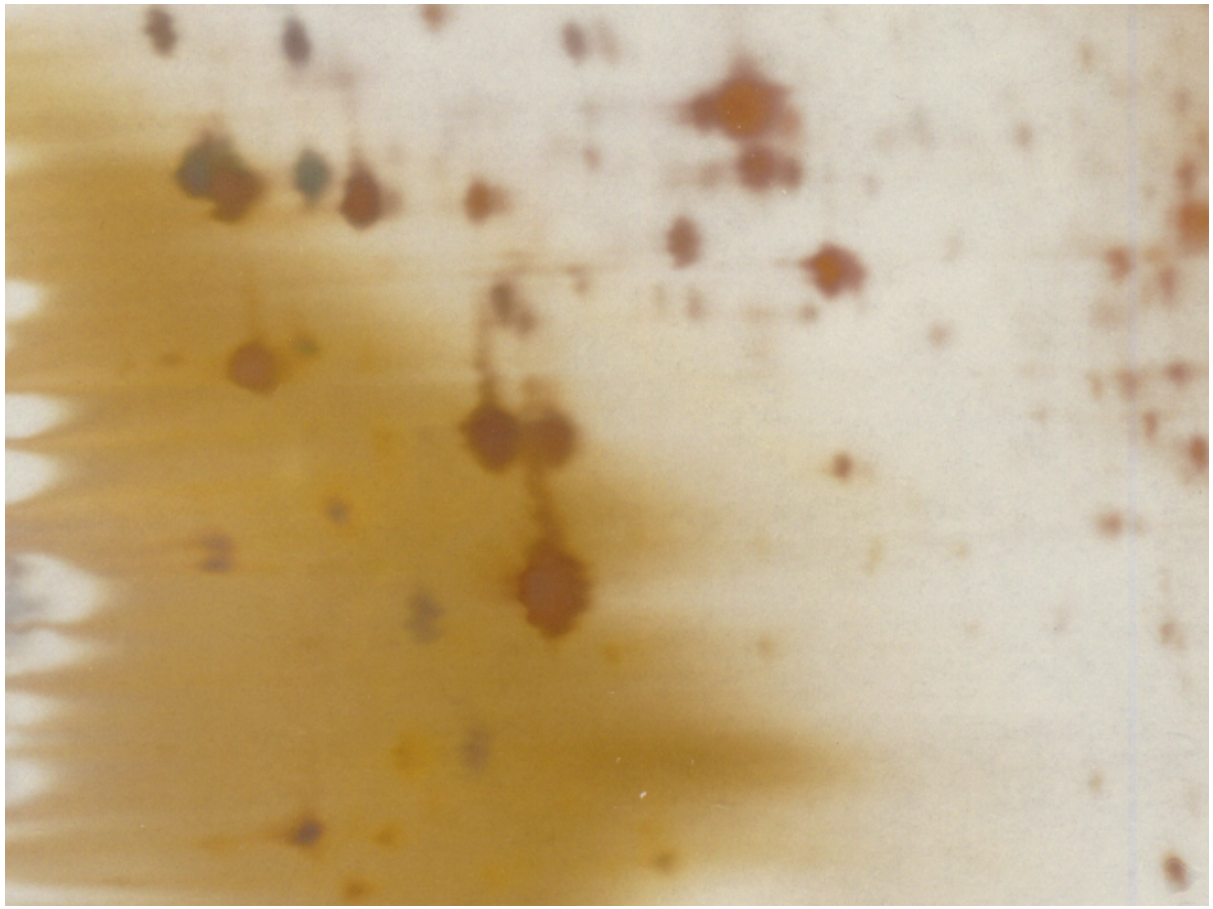
Proteomics

- 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

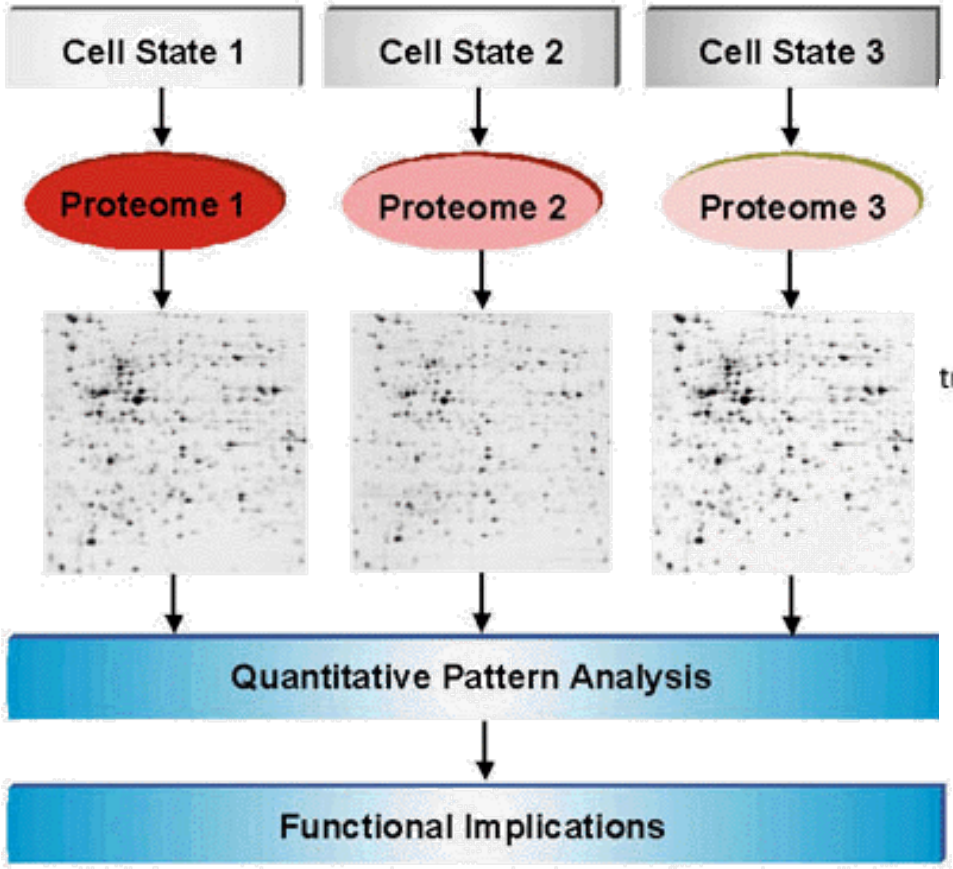
2D-Gels

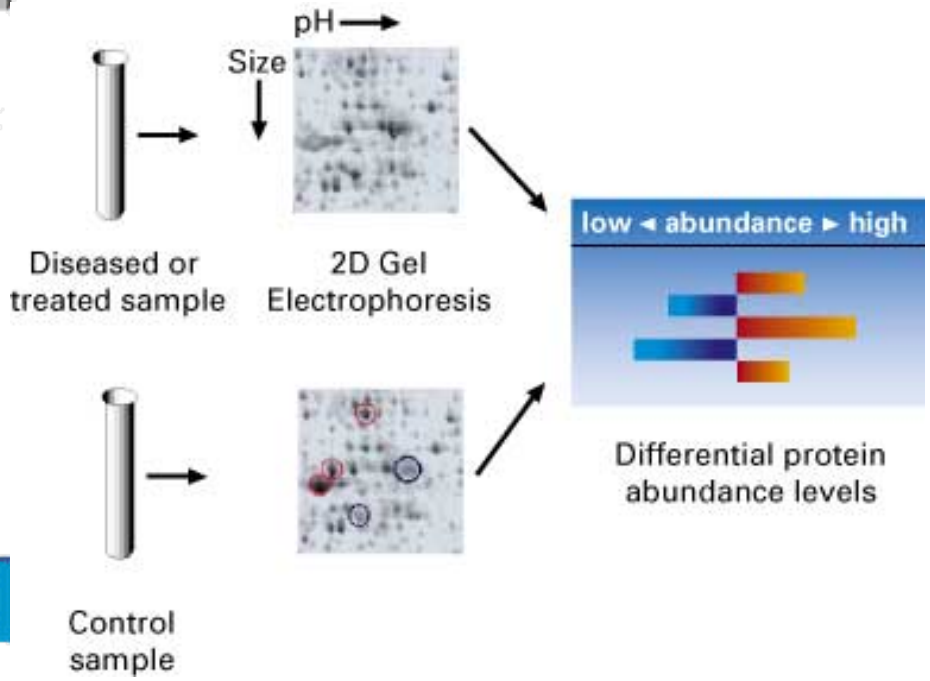

(a)
Protein mixture
pH 4.0
Separation in first dimension (by charge)
Isoelectric focusing (IEF)
pH 10.0
Apply first gel to top of second
pH 4.0
pH 10.0
Separation in second dimension (by size)
SDS electrophoresis

2D Gel Electrophoresis



2D-gels

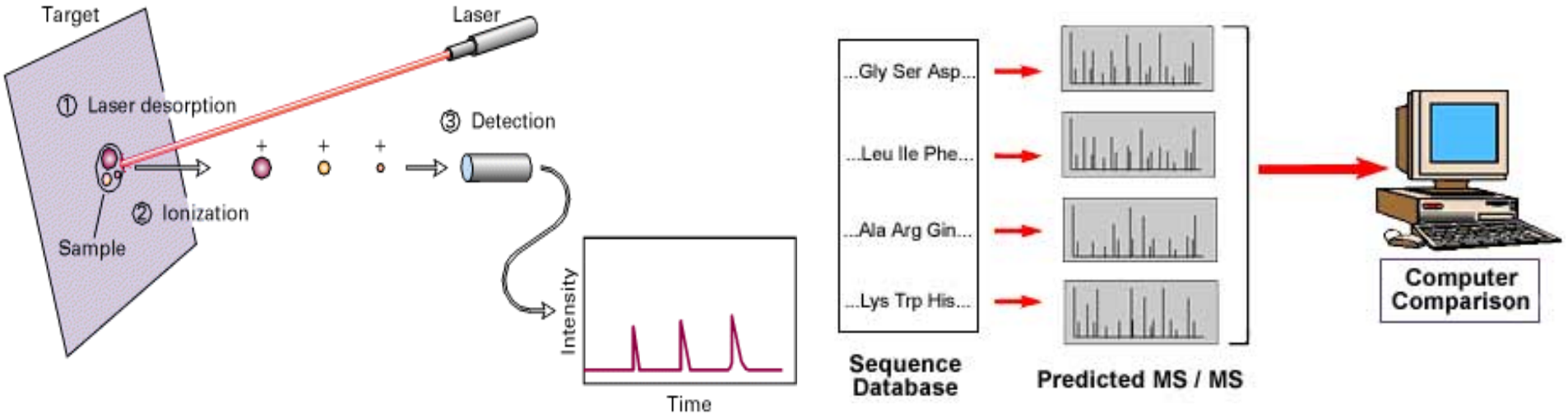




Comparing Proteomes For Differences in Protein Expression

Comparing Different Sample Types For Changes in Protein Levels

Mass Spectrometry

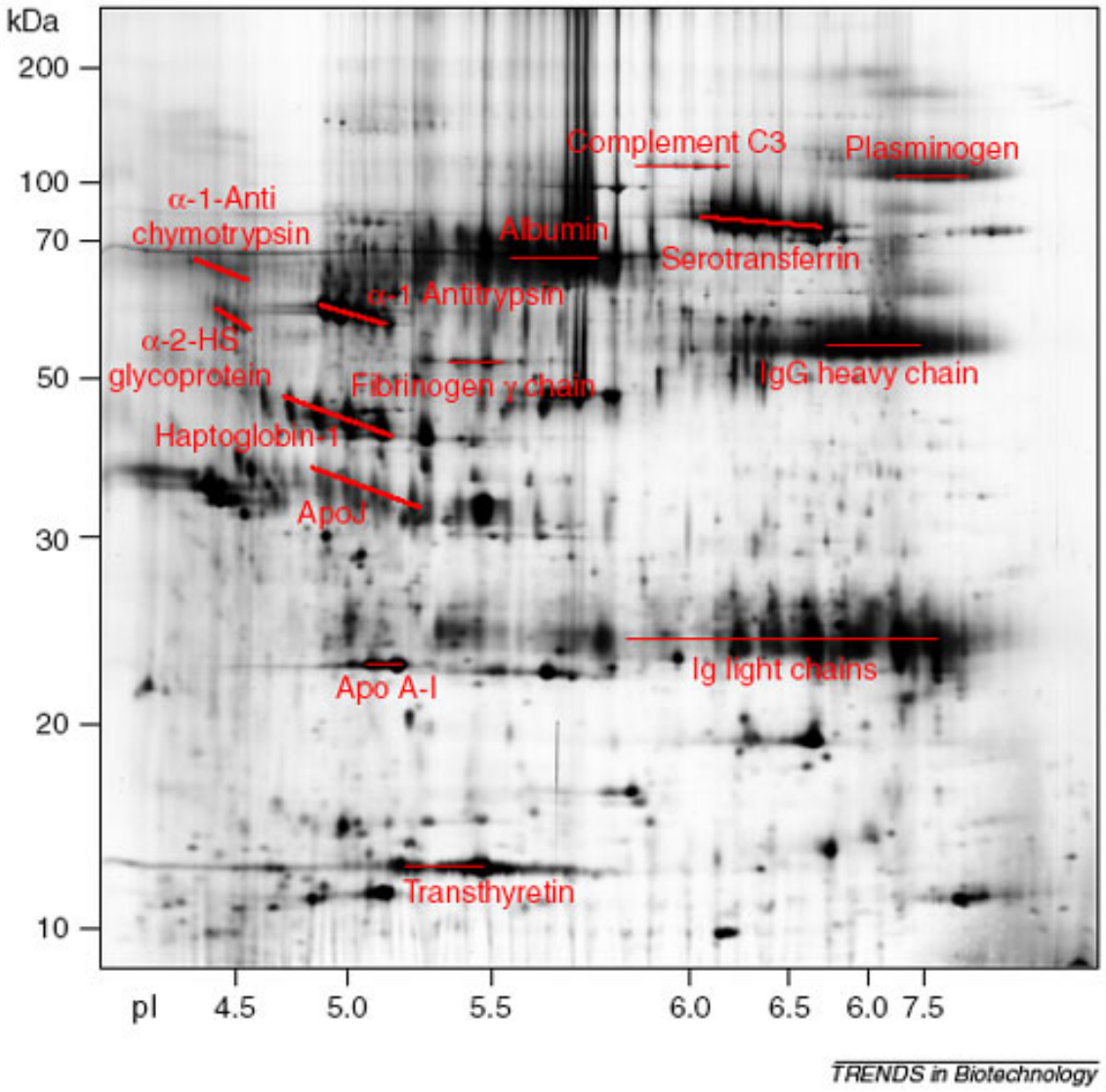


- **Mass measurements By Time-of-Flight**
  - Laser ionizes protein
  - Electric field accelerates molecules in sample toward detector
  - Time to detector is inversely proportional to mass of molecule
  - Infer molecular weights of proteins and peptides

Mass Spectrometry (MS)

Using Peptide Masses to Identify Proteins

- Peptide mass fingerprint is a compilation of molecular weights of peptides
- Use molecular weight of native protein and MS signature to search database for similarly-sized proteins with similar MS maps
- Fairly easy to sequence proteins using MS


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

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.

Comparative Interactomics