CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

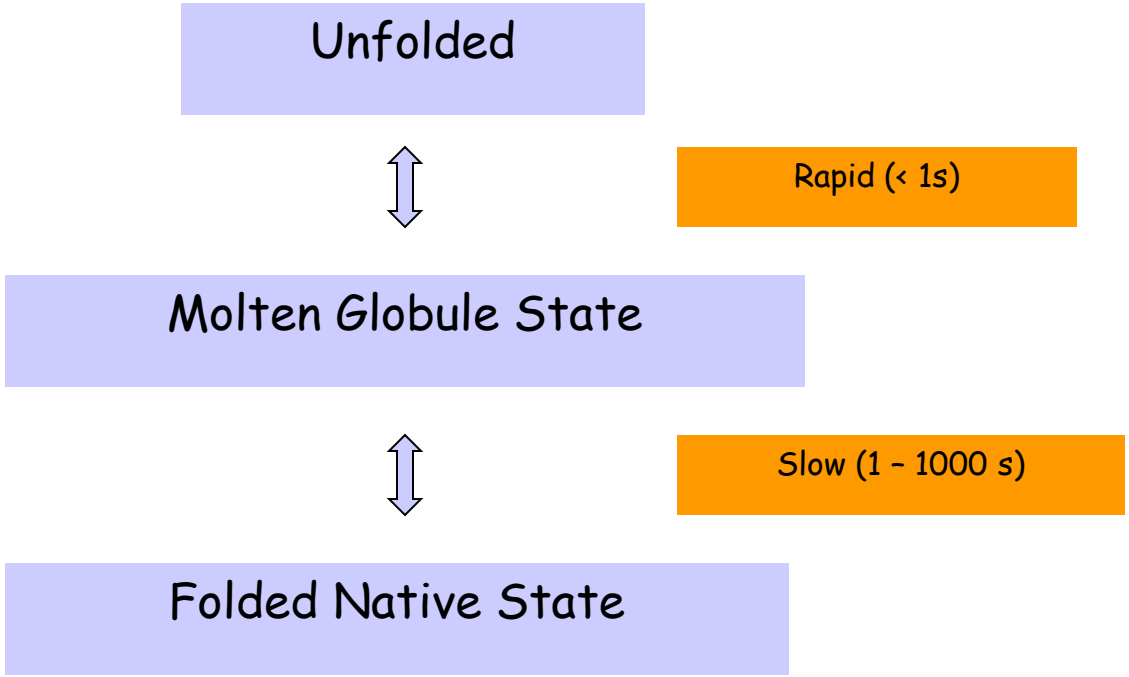
Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

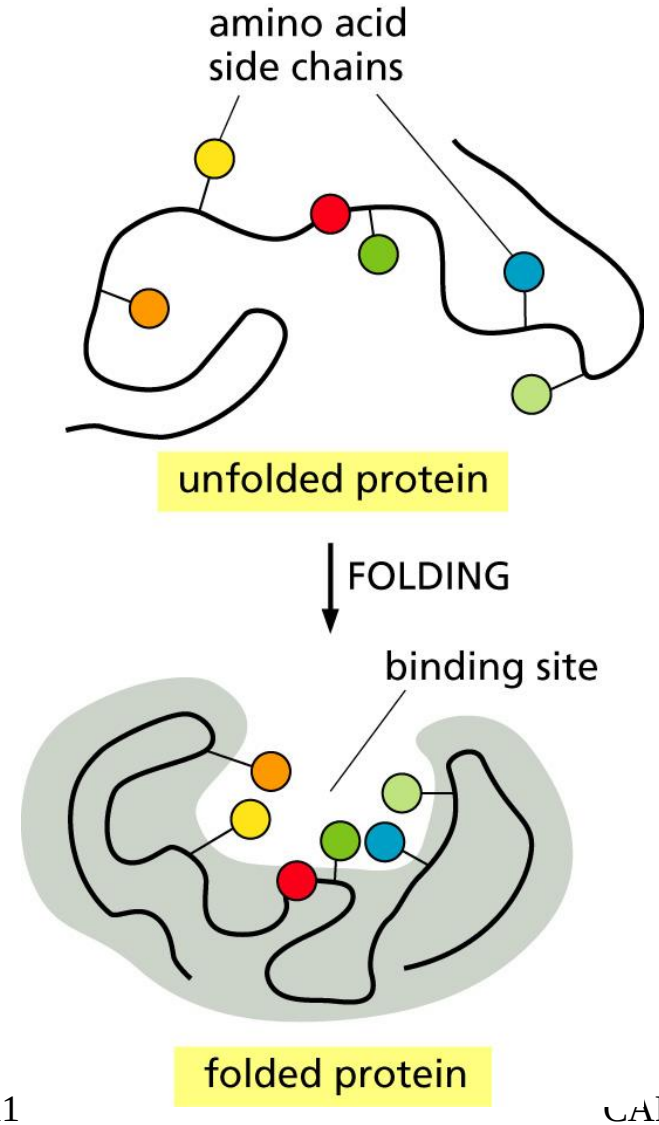
www.cis.fiu.edu/~giri/teach/BioinfS13.html

Protein Folding

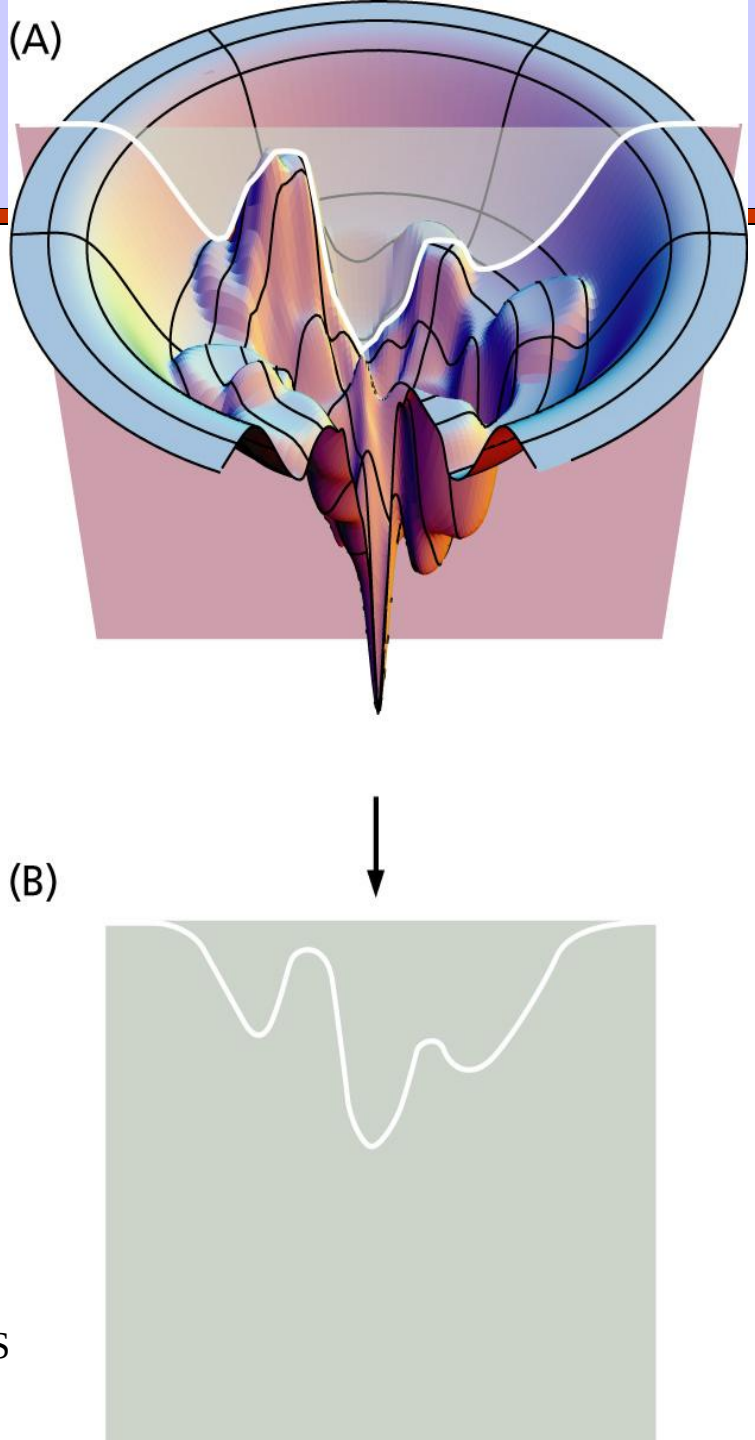


- How to find minimum energy configuration?

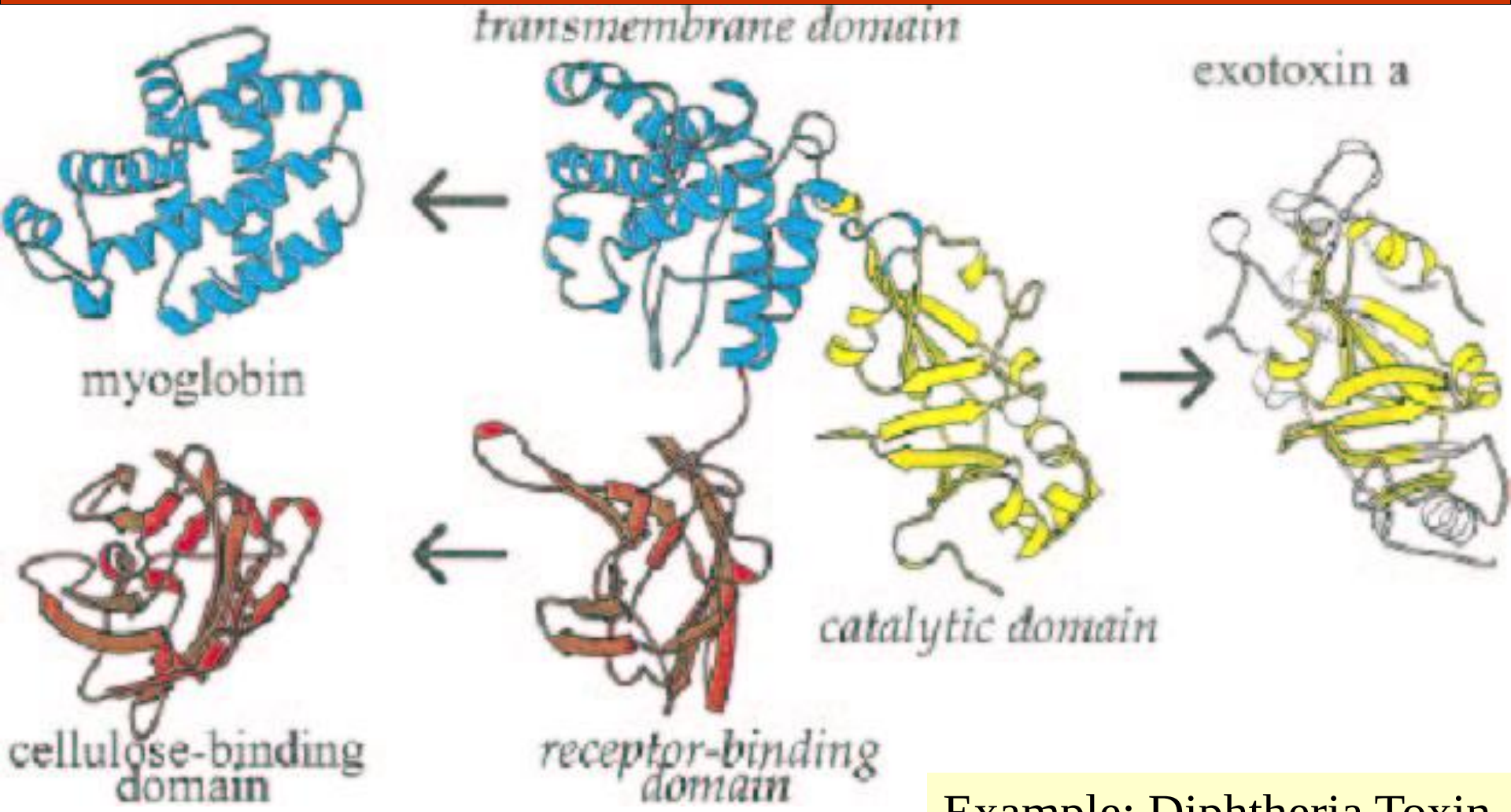
Protein Folding



Energy Landscape



Modular Nature of Protein Structures



Example: Diphtheria Toxin

Protein Structures

- Most proteins have a hydrophobic core.
- Within the core, specific interactions take place between amino acid side chains.
- Can an amino acid be replaced by some other amino acid?
  - Limited by space and available contacts with nearby amino acids
- Outside the core, proteins are composed of loops and structural elements in contact with water, solvent, other proteins and other structures.

Active Sites

Active sites in proteins are usually hydrophobic pockets/crevices/troughs that involve sidechain atoms.

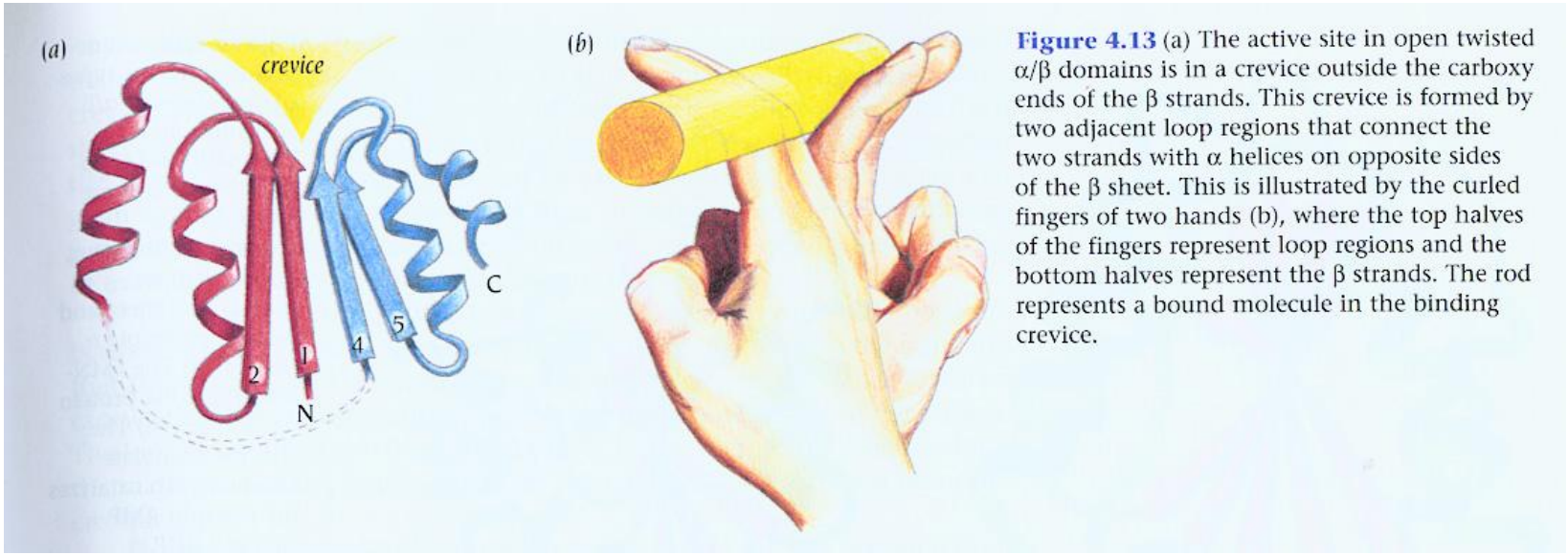
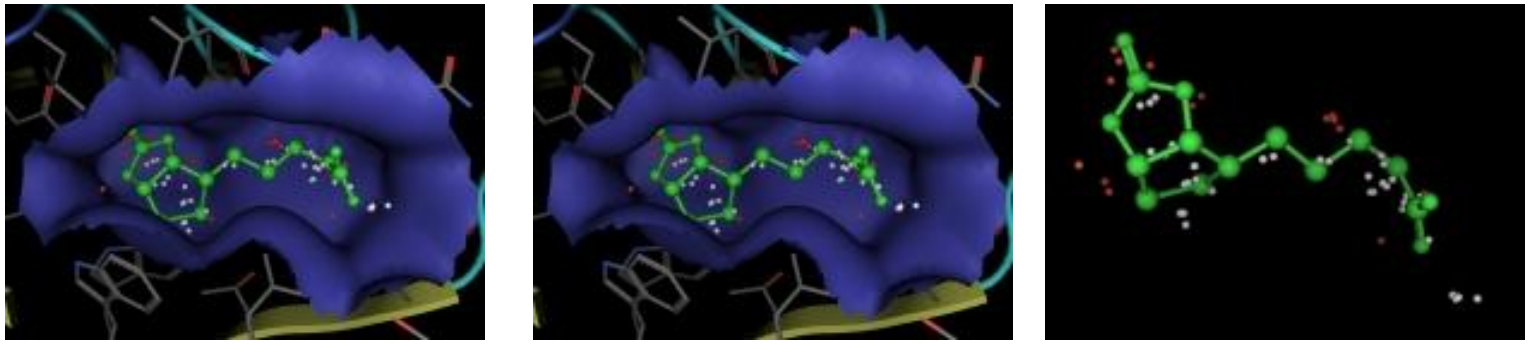


Figure 4.13 (a) The active site in open twisted α/β domains is in a crevice outside the carboxy ends of the β strands. This crevice is formed by two adjacent loop regions that connect the two strands with α helices on opposite sides of the β sheet. This is illustrated by the curled fingers of two hands (b), where the top halves of the fingers represent loop regions and the bottom halves represent the β strands. The rod represents a bound molecule in the binding crevice.

Active Sites



Left PDB 3RTD (streptavidin) and the first site located by the MOE Site Finder. **Middle** 3RTD with complexed ligand (biotin). **Right** Biotin ligand overlaid with calculated alpha spheres of the first site.

Viewing Protein Structures

- SPDBV
- RASMOL
- CHIME

Structural Classification of Proteins

- Over 1000 protein families known
  - Sequence alignment, motif finding, block finding, similarity search
- SCOP (Structural Classification of Proteins)
  - Based on structural & evolutionary relationships.
  - Contains ~ 40,000 domains
  - Classes (groups of folds), Folds (proteins sharing folds), Families (proteins related by function/evolution), Superfamilies (distantly related proteins)

SCOP Family View

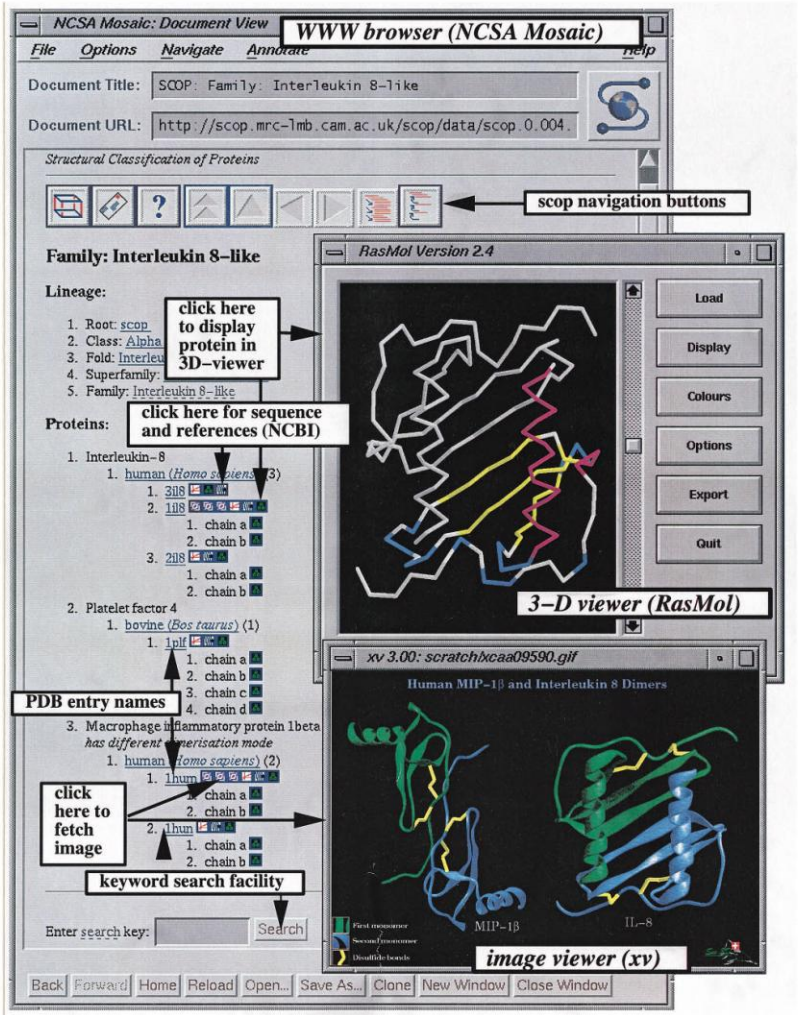


Figure 2. A typical scop session is shown on a unix workstation. A scop page, of the Interleukin 8-like family, is displayed by the *WWW browser program (NCSA Mosaic)* (Schatz & Hardin, 1994). Navigating through the tree structure is accomplished by selecting any underlined entry, by clicking on buttons (at the top of each page) and by keyword searching (at the bottom of each page). The static image comparing two proteins in this family was downloaded by clicking on the icon indicated and is displayed by image-viewer program *xv*. By clicking on one of the green icons, commands were sent to a molecular viewer program (*RasMol*) written by Roger Sayle (Sayle, 1994), instructing it to automatically display the relevant PDB file and colour the domain in question by secondary structure. Since sending large PDB files over the network can be slow, this feature of scop can be configured to use local copies of PDB files if they are available. Equivalent WWW browsers, image-display programs and molecular viewers are also available free for Windows-PC and Macintosh platforms.

CATH: Protein Structure Classification

- Semi-automatic classification; ~36K domains
- 4 levels of classification:
  - Class (C), depends on sec. Str. Content
    - α class, β class, α/β class, $\alpha+\beta$ class
  - Architecture (A), orientation of sec. Str.
  - Topolgy (T), topological connections &
  - Homologous Superfamily (H), similar str and functions.

DALI/FSSP Database

- Completely automated; 3724 domains
- Criteria of compactness & recurrence
- Each domain is assigned a Domain Classification number DC_l_m_n_p representing fold space attractor region (l), globular folding topology (m), functional family (n) and sequence family (p).

Structural Alignment

- What is structural alignment of proteins?
  - 3-d superimposition of the atoms as "best as possible", i.e., to minimize RMSD (root mean square deviation).
  - Can be done using VAST and SARF
- Structural similarity is common, even among proteins that do not share sequence similarity or evolutionary relationship.

Other databases & tools

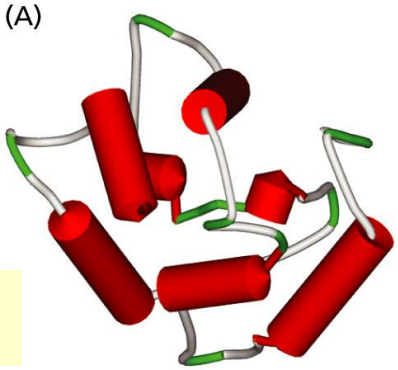
- MMDB contains groups of structurally related proteins
- SARF structurally similar proteins using secondary structure elements
- VAST Structure Neighbors
- SSAP uses double dynamic programming to structurally align proteins

5 Fold Space classes

Attractor 1 can be characterized as alpha/beta,
attractor 2 as all-beta, attractor 3 as all-alpha,
attractor 5 as alpha-beta meander (1mli), and
attractor 4 contains antiparallel beta-barrels e.g. OB-fold (1prtF).

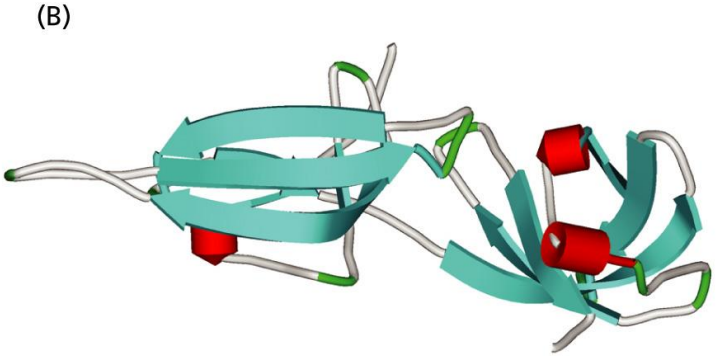
Examples of protein classes

Parvalbumin



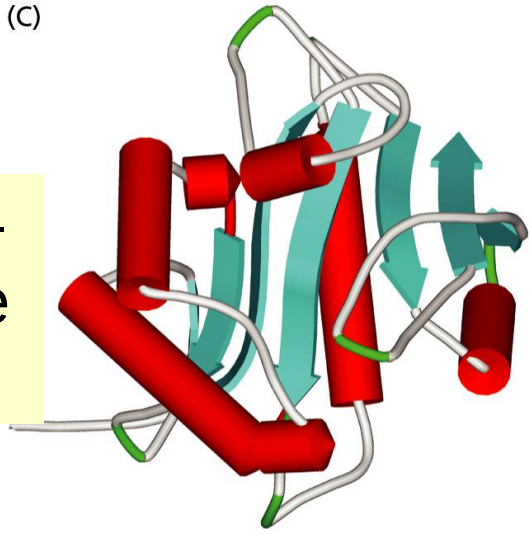
1B8C

Translation Initiation Factor 5A

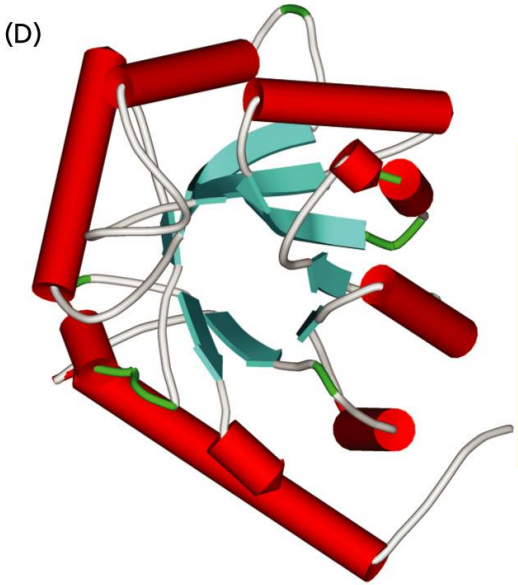


1BKB

Serotonin N-acetyltransfe rase



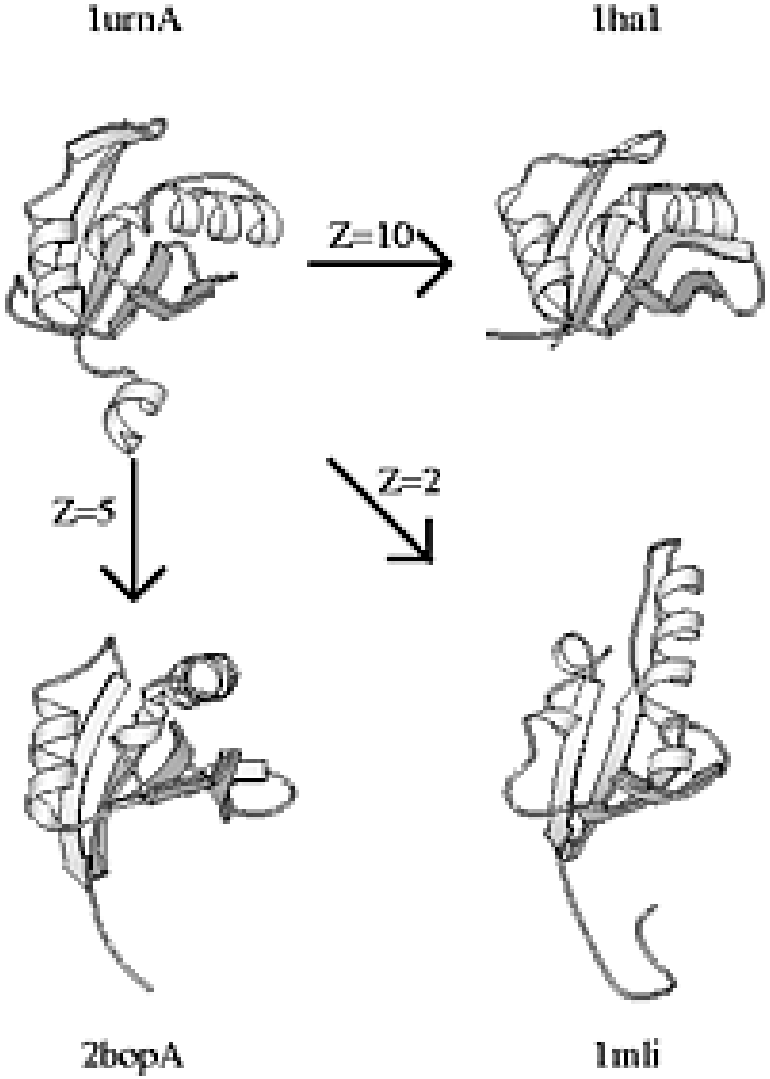
1CJW



1CT5

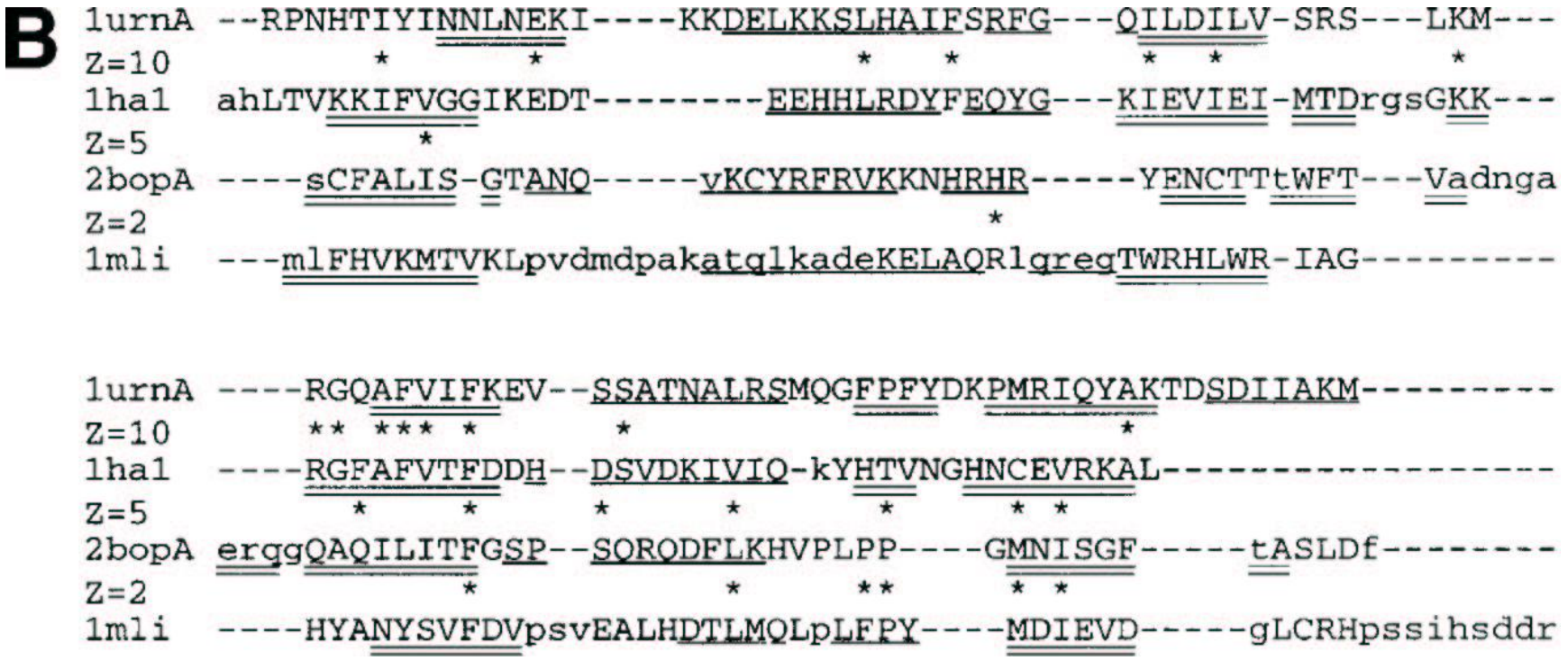
Hypothetical protein from yeast (α/β alternating fold)

Fold Types & Neighbors

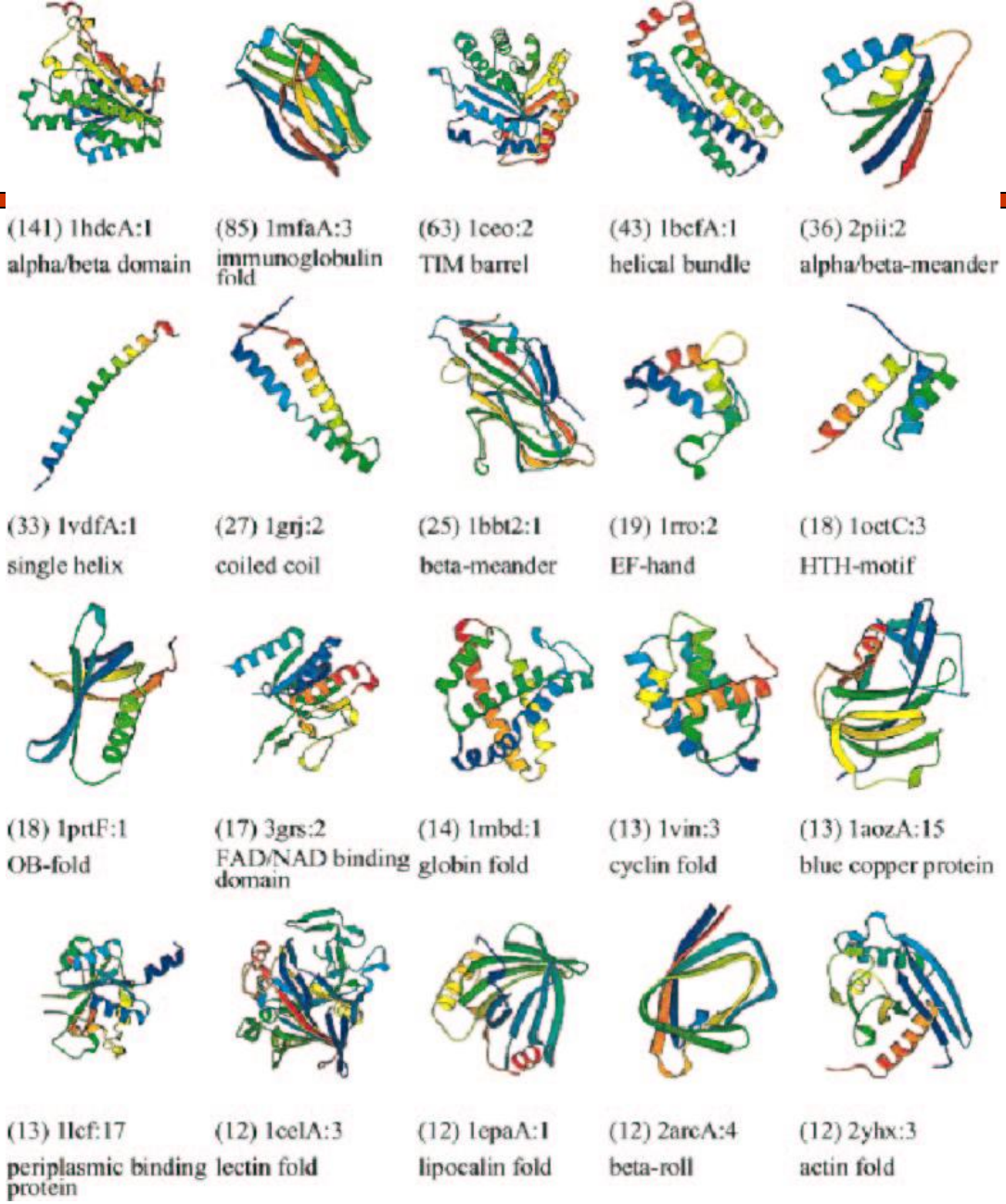


Structural neighbours of 1urnA (top left). 1mli (bottom right) has the same topology even though there are shifts in the relative orientation of secondary structure elements.

Sequence Alignment of Fold Neighbors



Frequent Fold Types



Protein Structure Prediction

- Holy Grail of bioinformatics
- Protein Structure Initiative to determine a set of protein structures that span protein structure space sufficiently well. WHY?
  - Number of folds in natural proteins is limited. Thus a newly discovered proteins should be within modeling distance of some protein in set.
- CASP: Critical Assessment of techniques for structure prediction
  - To stimulate work in this difficult field

PSP Methods

- homology-based modeling
- methods based on fold recognition
  - Threading methods
- *ab initio* methods
  - From first principles
  - With the help of databases

ROSETTA

- Best method for PSP
- As proteins fold, a large number of partially folded, low-energy conformations are formed, and that local structures combine to form more global structures with minimum energy.
- Build a database of known structures (I-sites) of short sequences (3-15 residues).
- Monte Carlo simulation assembling possible substructures and computing energy

Threading Methods

- See p471, Mount
  - **http://www.bioinformaticsonline.org/links/ch_10_t_7.html**

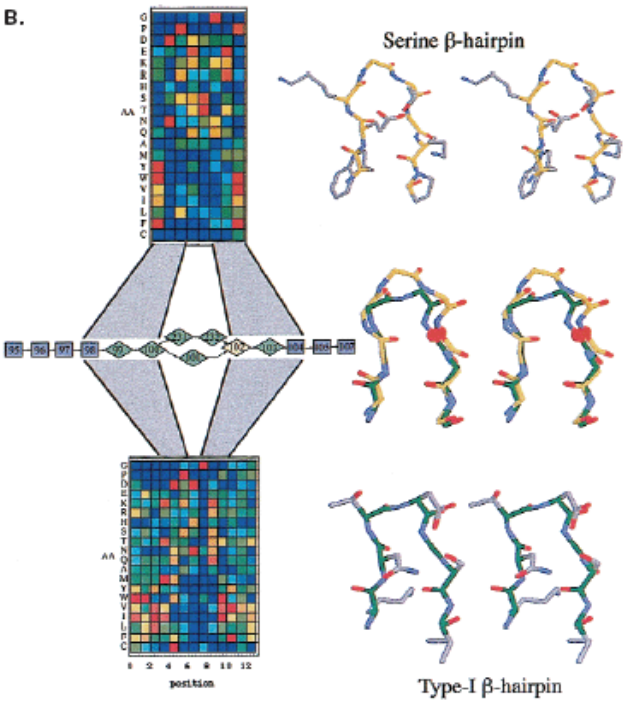


FIGURE 10.30. A hidden Markov model (discrete state-space model) of protein three-dimensional structure. (*B*) HMM called HMMSTR based on I-sites, 3- to 15-amino-acid patterns that are associated with three-dimensional structural features. The two matrices with colored squares represent alignment of sets of patterns that are found to be associated with a structure, in this case the hairpin turns shown on the right. Each column in the table corresponds to the amino acid variation found for one structural position in one of the turns. (*Blue* side chains) Conserved nonpolar residues; (*green*) conserved polar residues; (*red*) conserved proline; and (*orange*) conserved glycine. The two hairpins are aligned structurally in the middle structure on the right and the observed variation in the corresponding amino acid positions is represented by the HMM between the matrices on the left. The HMM represents an alignment of the two hairpin structural motifs in three-dimensional space and an alignment of the sequences. A short mismatch in the turn is represented by splitting the model into two branches. The shaped icons represent states, each of which represents a structure and a sequence position. Each state contains probability distributions about the sequence and structural attributes of a single position in the motif, including the probability of observing a particular amino acid, secondary structure, Φ-Ψ backbone angles, and structural context, e.g., location of β strand in a β sheet. Rectangles are predominantly β-strand states, and diamonds are predominantly turns. The color of the icon indicates a sequence preference as follows: (*blue*) hydrophobic; (*green*) polar; and (*yellow*) glycine. Numbers in icons are arbitrary identification numbers for the HMM states. There is a transition probability of moving from each state in the model to the next, as in HMMs that represent msa's. This model is a small component of the main HMMSTR model that represents a merging of the entire I-sites library. Three different models, designated λ^D, λ^C, and λ^R, are included in HMMSTR, which differ in details as to how the alignment of the I-sites was obtained to design the branching patterns (topology) of the model and which structural data were used to train the model. HMMSTR may be used for a variety of different predictions, including secondary structure prediction, structural context prediction, and Φ-Ψ dihedral angle prediction. Predictions are made by aligning the model with a sequence, finding if there is a high-scoring alignment, and deciphering the highest-scoring path through the model. The HMMSTR program may be downloaded or used on a server that can be readily located by a Web search. (*B*, reprinted, with permission, from Bystroff et al. 2000 [©2000 Elsevier].)

Modeling Servers

- SwissMODEL
- 3DJigsaw
- CPHModel
- ESyPred3D
- Geno3D
- SDSC1
- Rosetta
- MolIDE
- SCWRL
- PSIPred
- MODELLER
- LOOPY

Hypothesis Testing

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

Hypothesis Testing Logic

		Hypothesis Choice	
		H0	H1
Decision	H0	Correctly Accept (TN)	Type II Error (FN) β
	H1	Type I Error (FP) α	Correctly Reject (TP)

- Typical Values:
  - Type I error of 0.05
  - Type II error of 0.2

Problem with Hypothesis Testing

- Not testing just one gene
- If multiple genes are tested, then t-Test assumes each test is independent
- Are the tests independent?
  - No!
- Need Correction
  - P-values need to be adjusted
  - Bonferroni or other correction methods needed
  - Achieved by controlling Type I error

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.

Ask another question

Rank	Anova (p)	q Value	Power	Cluster
30	0.00436	0.0119	0.993	
77	0.00536	0.0119	0.987	
97	0.00631	0.0119	0.98	
29	0.00655	0.0119	0.979	
43	0.00605	0.0119	0.982	
23	0.0067	0.0119	0.977	
36	0.00632	0.0119	0.98	
28	0.00698	0.0119	0.975	
76	0.00685	0.0119	0.976	
60	0.0067	0.0119	0.977	
10	0.00479	0.0119	0.991	
13	0.00467	0.0119	0.991	
51	0.00432	0.0119	0.993	
91	0.0062	0.0119	0.981	
21	0.00611	0.0119	0.982	
46	0.00414	0.0119	0.994	
45	0.00739	0.0127	0.972	
25	0.00822	0.0137	0.964	
53	0.00903	0.0137	0.956	
6	0.00919	0.0138	0.955	
52	0.01	0.0141	0.946	
2	0.00976	0.0141	0.949	
87	0.0101	0.0141	0.946	
19	0.0109	0.0141	0.938	
96	0.0102	0.0141	0.944	
55	0.011	0.0141	0.937	
50	0.00949	0.0141	0.952	
49	0.0115	0.0144	0.931	
32	0.0127	0.0144	0.918	

P-value vs Q-value

Consider example shown. Let N = 839.

Marked item has p-value 0.01 and q-value 0.0141. P-value threshold of 0.01 implies a 1% chance of false positives. Thus, we expect 839*0.01 = 8.39 FPs. Since item has rank 52, we expect to have 8 or 9 of these to be FPs.

Q-value threshold of 0.0141 implies a 1.41% of all spots with q-value less than this to be FPs. Thus, we expect 52*0.0141 = 0.7332 FPs, i.e., less than one FP.

Annotations and Gene Ontology

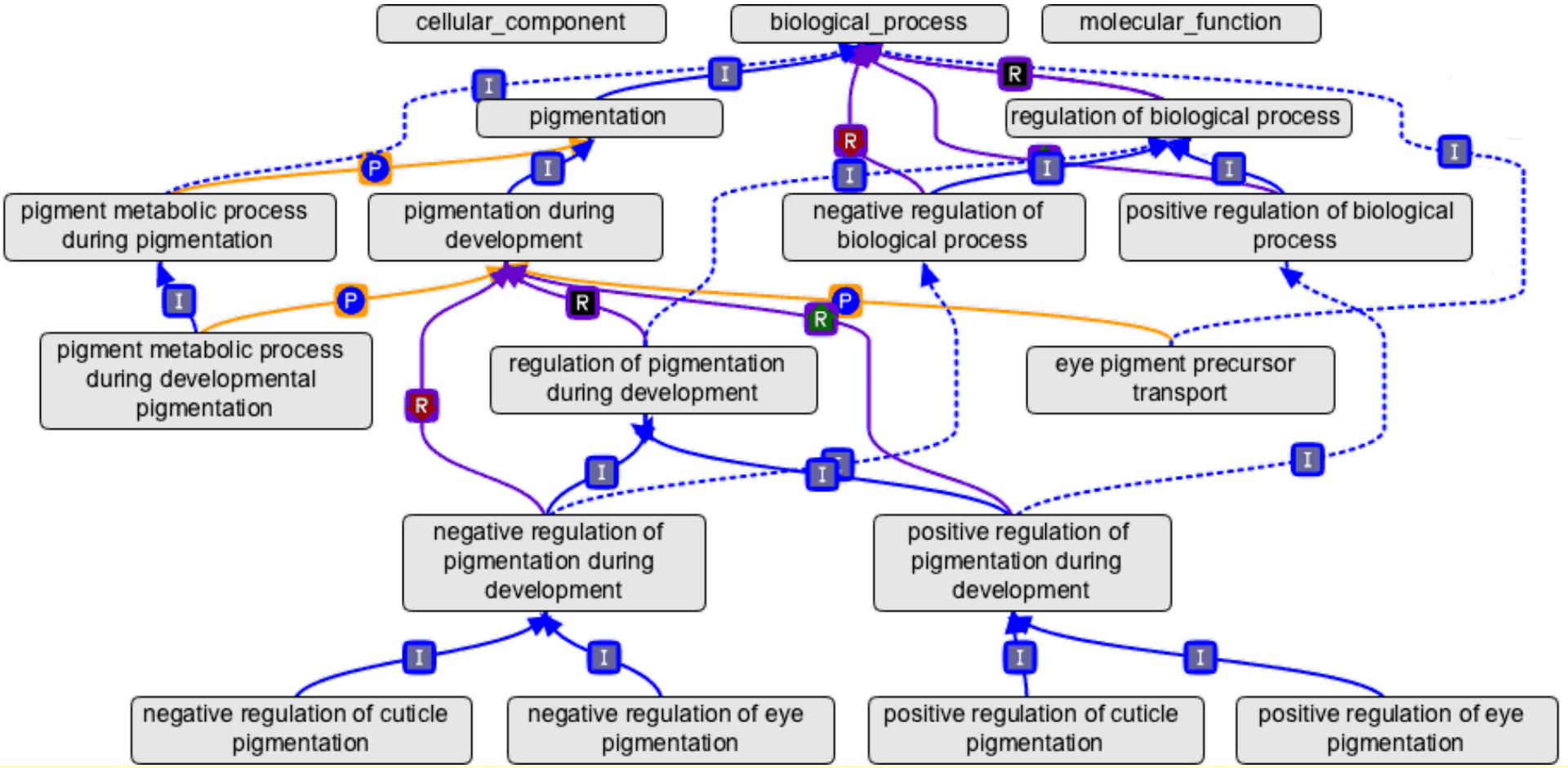
Annotation

- 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

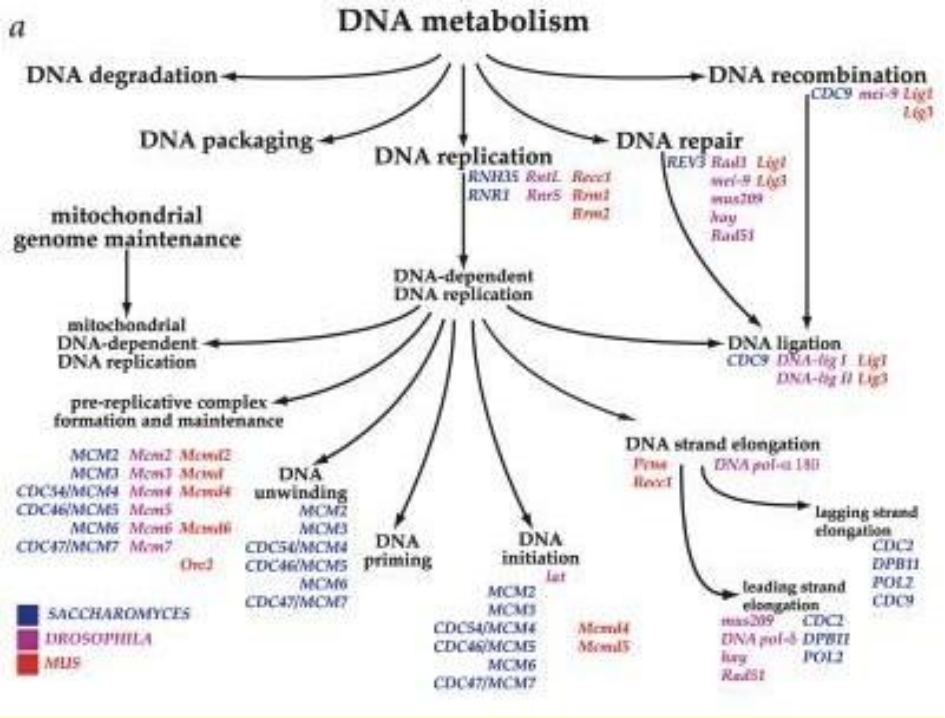
Sample GO Term

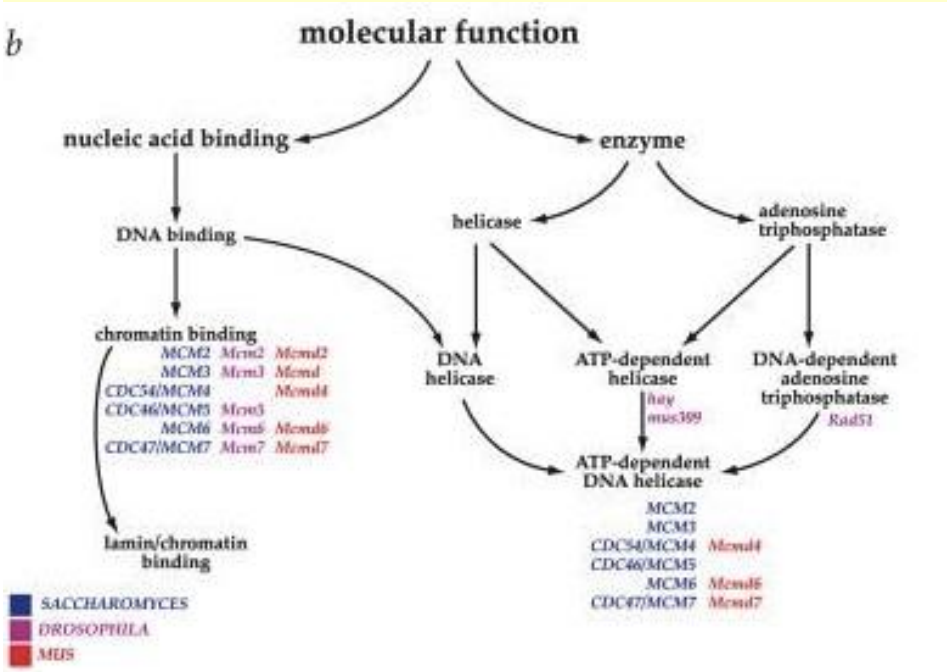
id: GO:0016049
name: **cell growth**
namespace: biological_process
def: "The process in which a cell irreversibly increases in size over time by accretion and biosynthetic production of matter similar to that already present." [GOC:ai]
subset: goslim_generic
subset: goslim_plant
subset: gosubset_prok
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

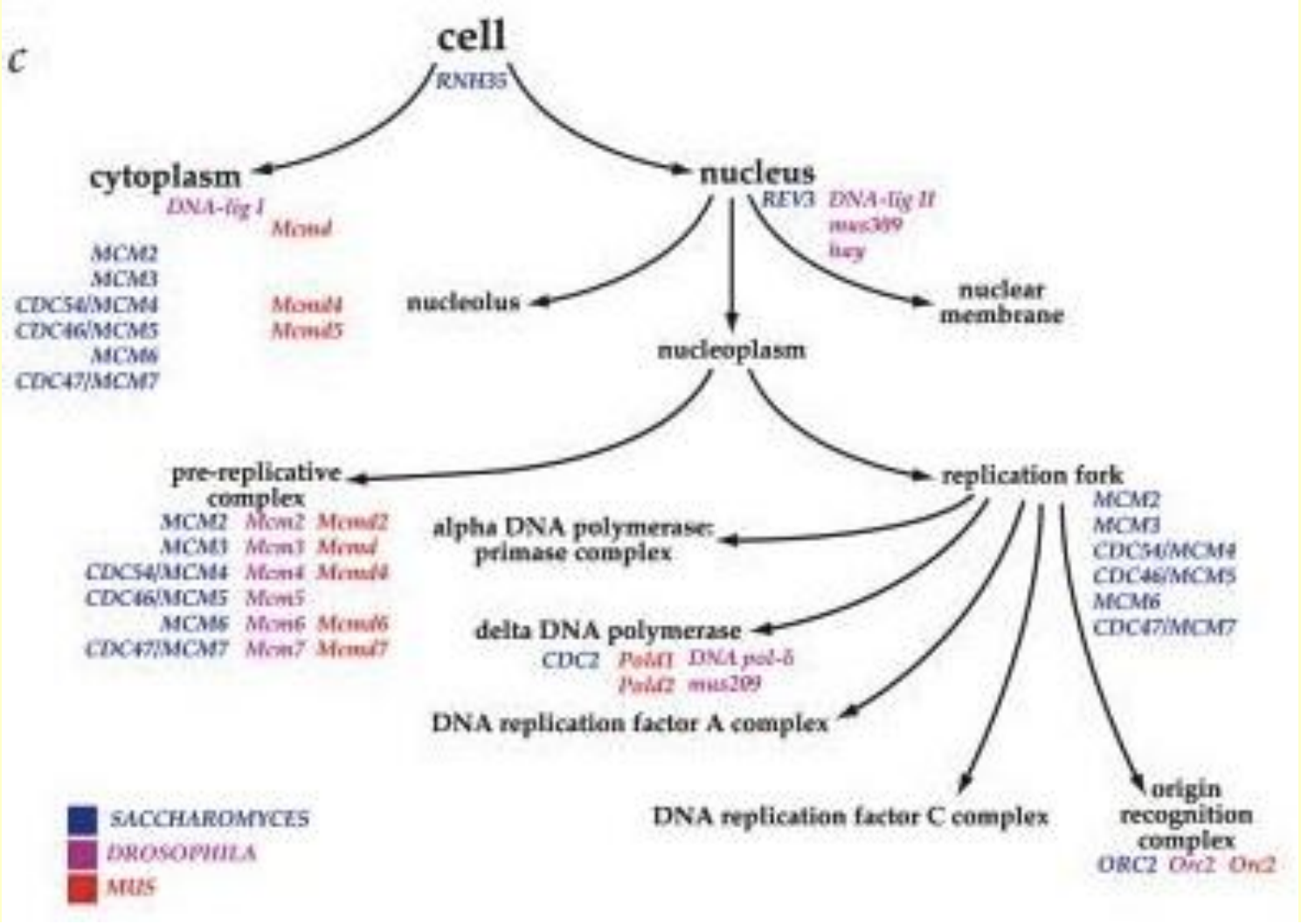
- Cellular Component: A component of the cell, i.e., location
  - E.g., rough endoplasmic reticulum, nucleus, ribosome, proteasome
- Biological Process: A biological process is series of events accomplished by one or more ordered assemblies of molecular functions.
  - E.g., cellular physiological process, 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

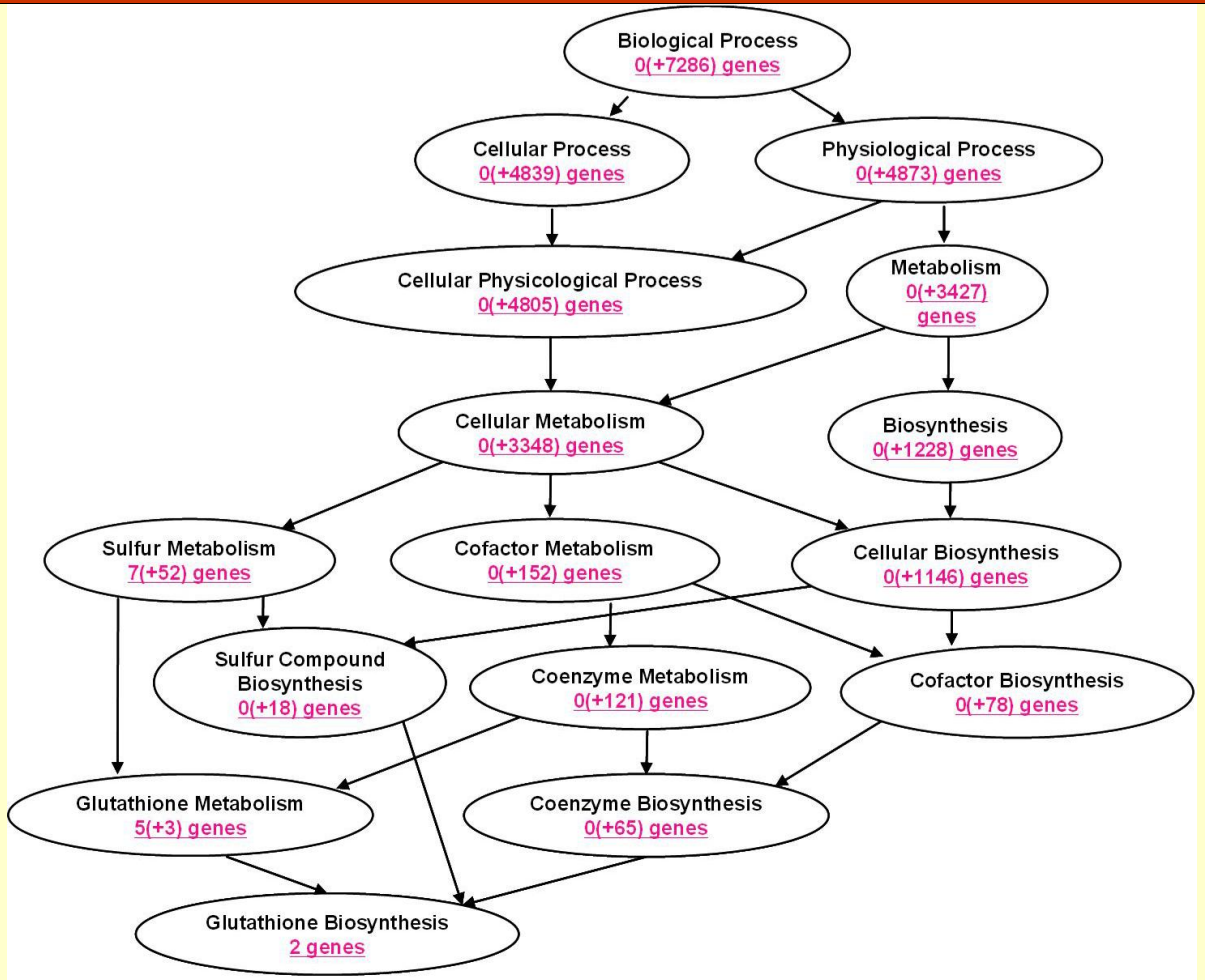




Cellular Component



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

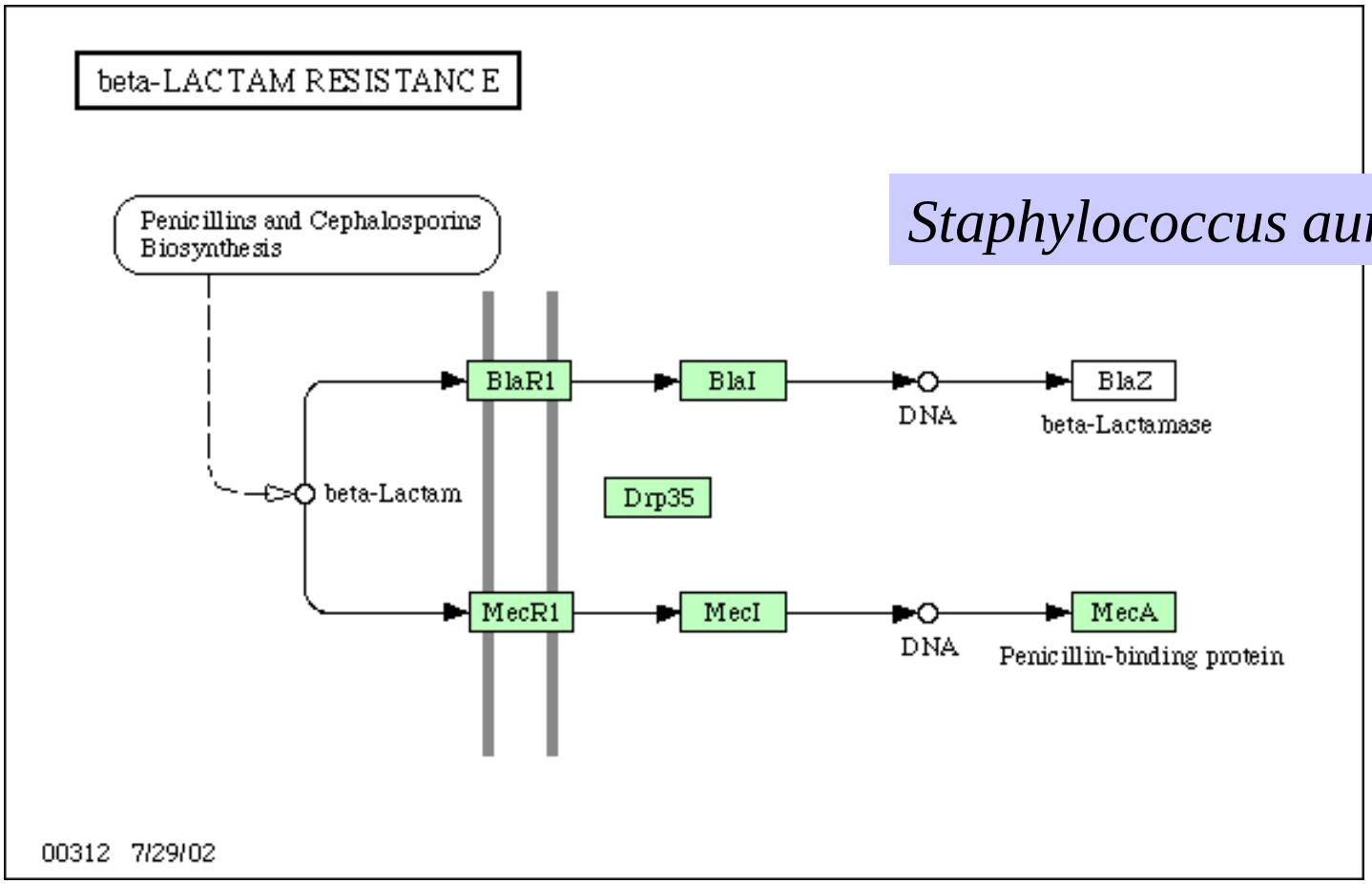
KEGG: More Networks …

Disease

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

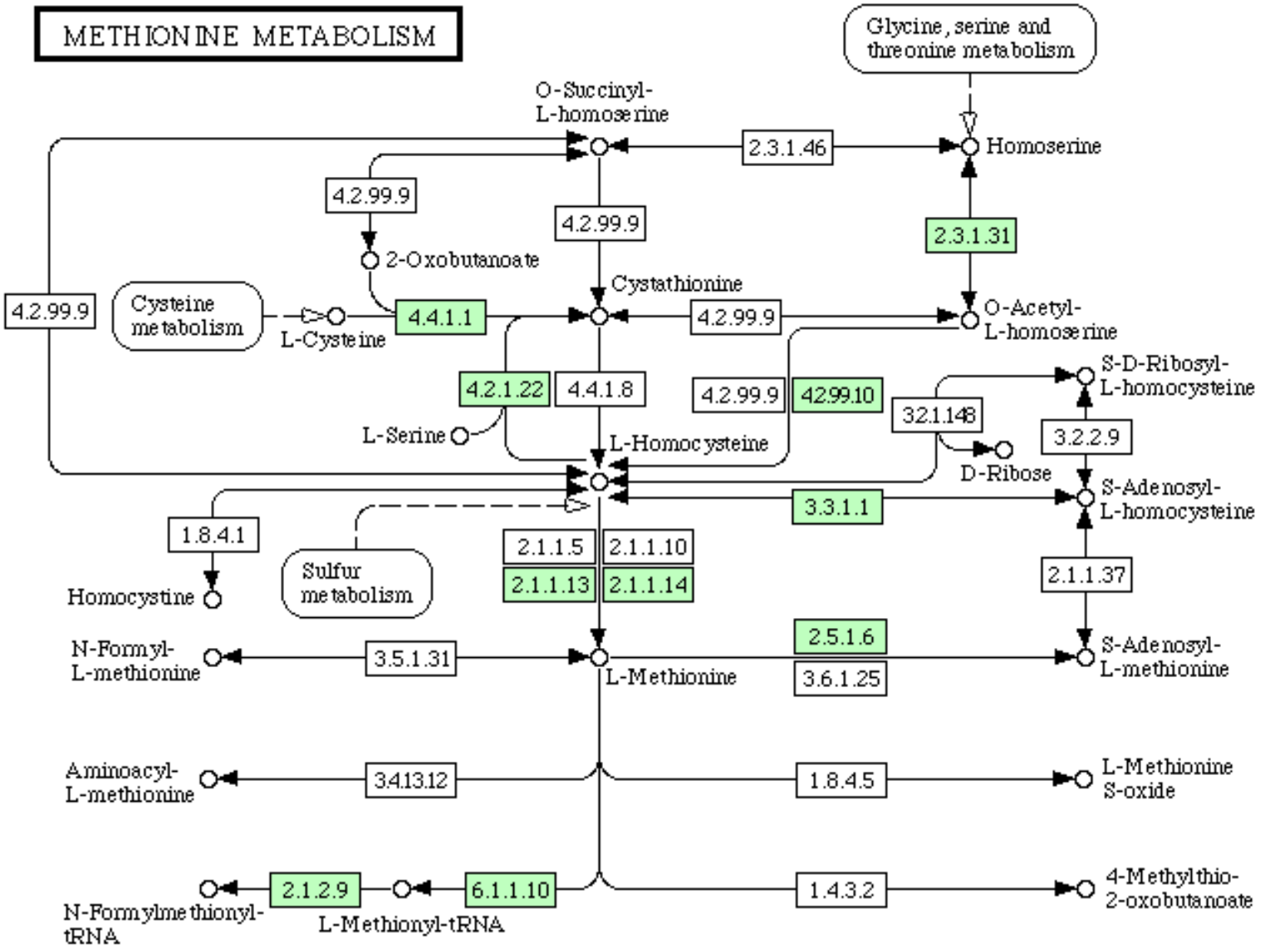
- Drugs
- Antibiotics
- Chronology: Antineoplastics, nervous system agents, misc., …
- Target-based: GPCRs, Nuclear, Ion Channels, Enzymes
- Structure-based
- Skeleton-based

Pathway Example from KEGG



Staphylococcus aureus

Pseudomonas aeruginosa



00271 8/6/02

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
  - …

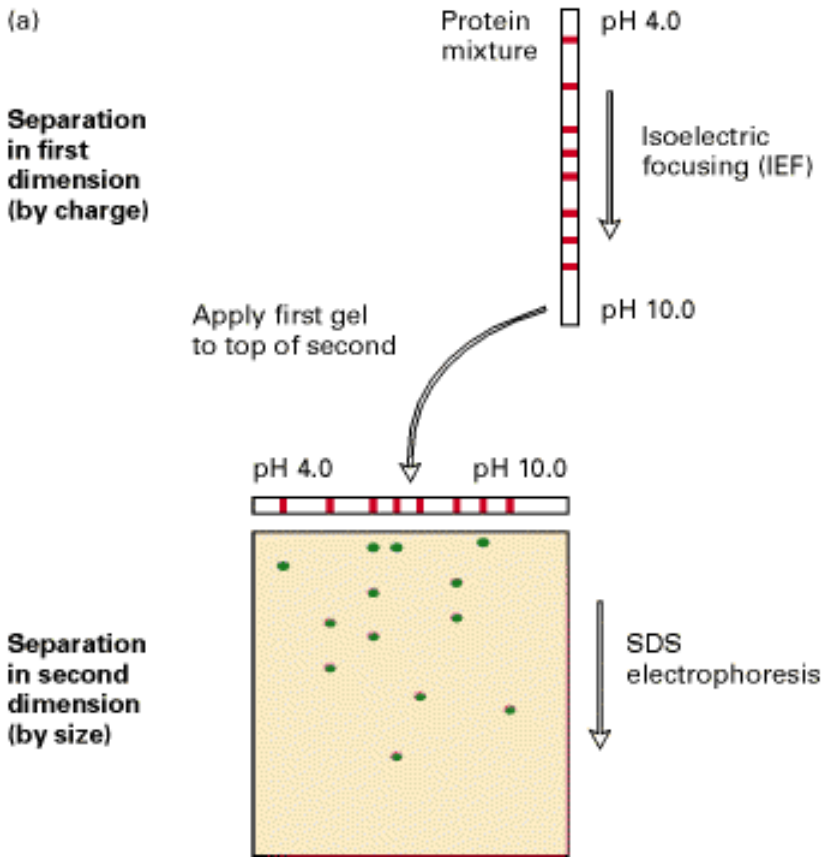
Databases for Comparative Genomics

- 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

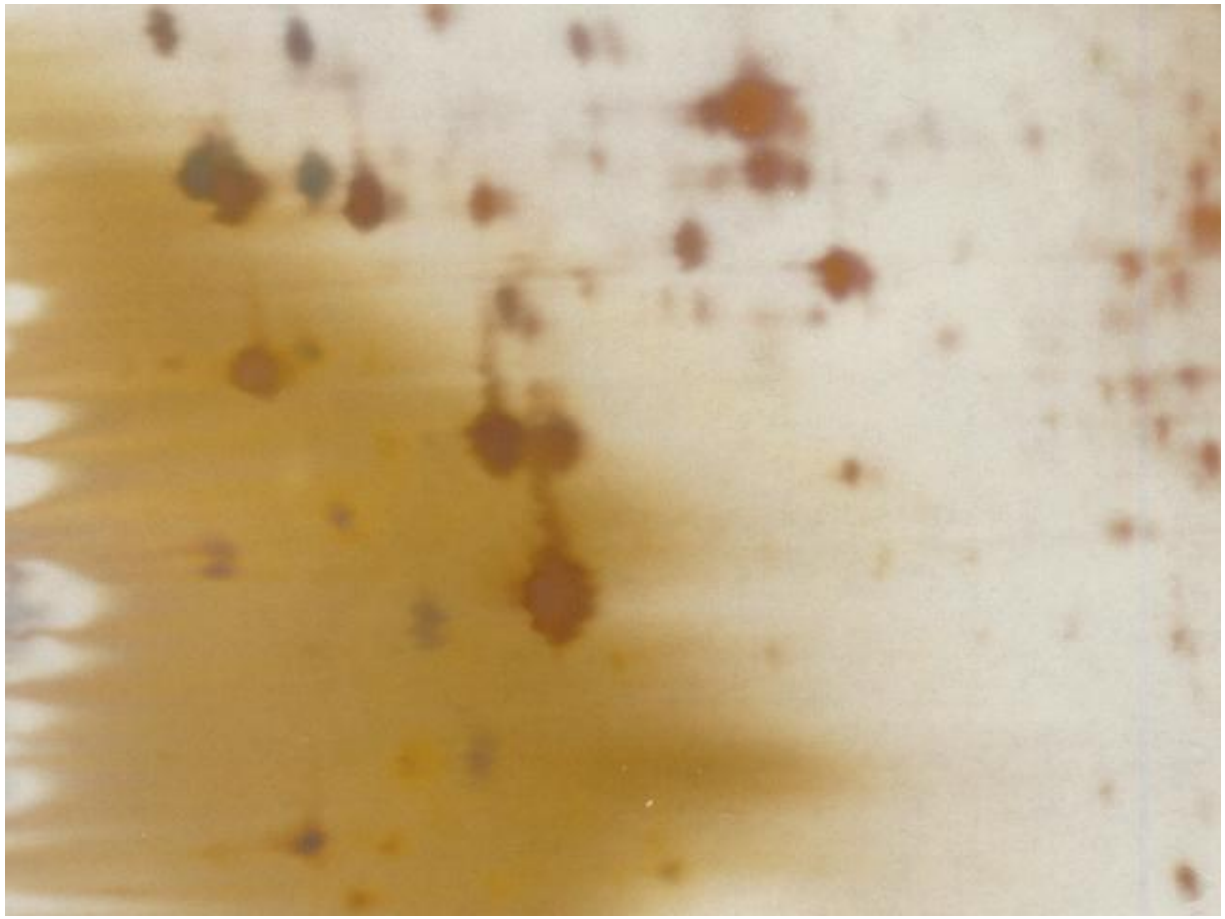
Proteomics

- 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

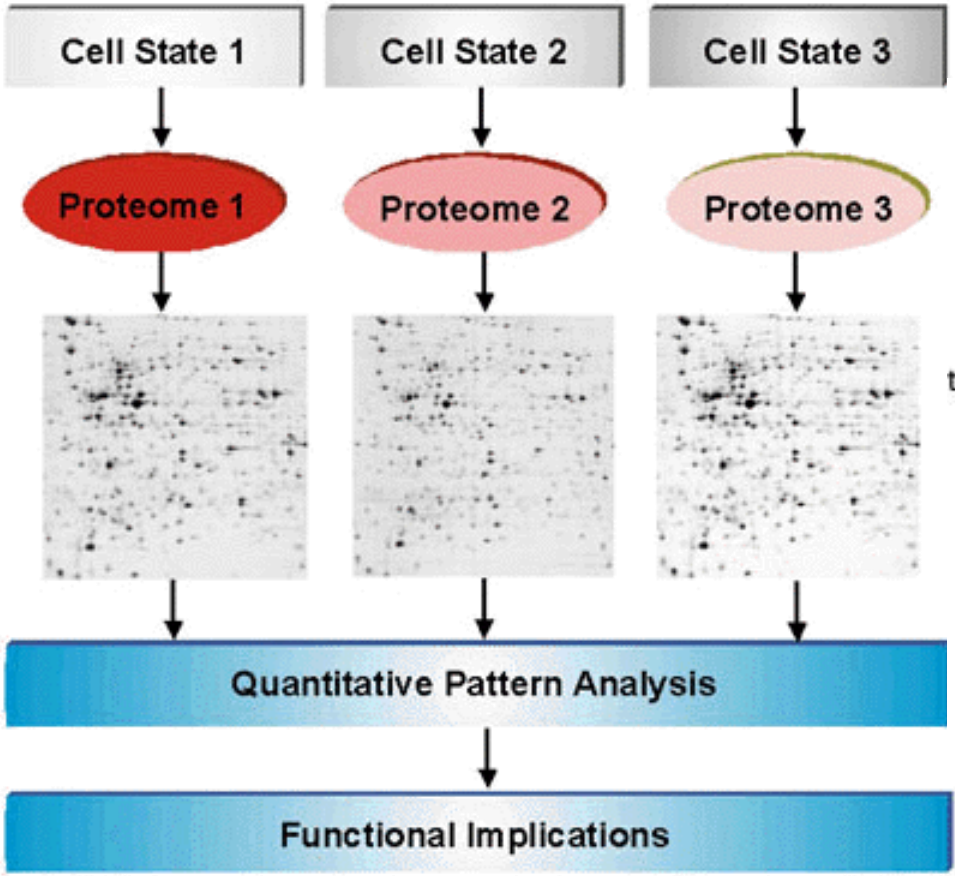
2D-Gels

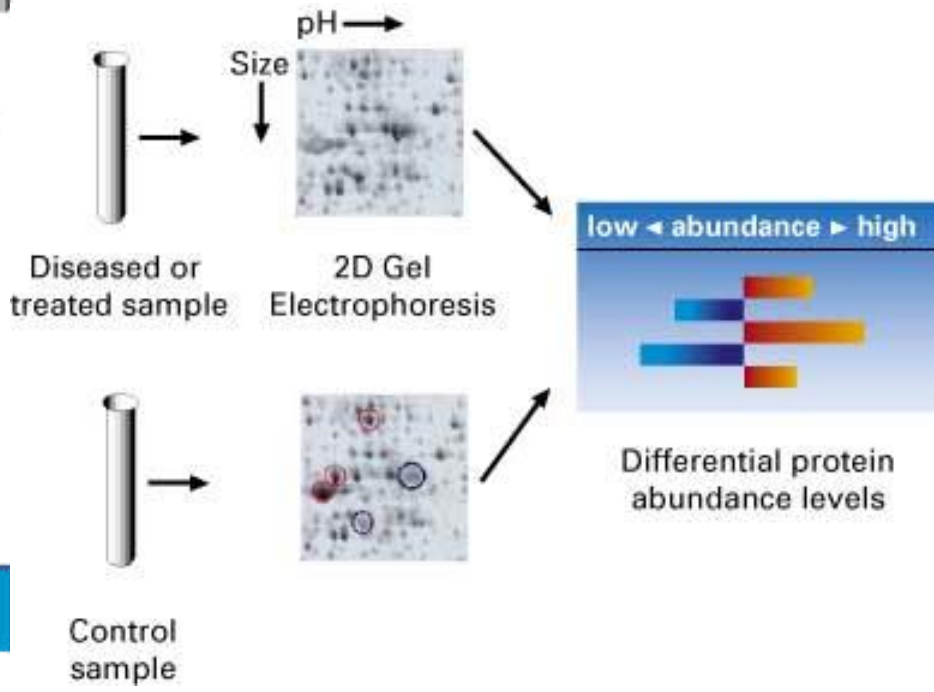

(a)
Protein mixture
pH 4.0
Separation in first dimension (by charge)
Isoelectric focusing (IEF)
pH 10.0
Apply first gel to top of second
pH 4.0
pH 10.0
Separation in second dimension (by size)
SDS electrophoresis

2D Gel Electrophoresis



2D-gels

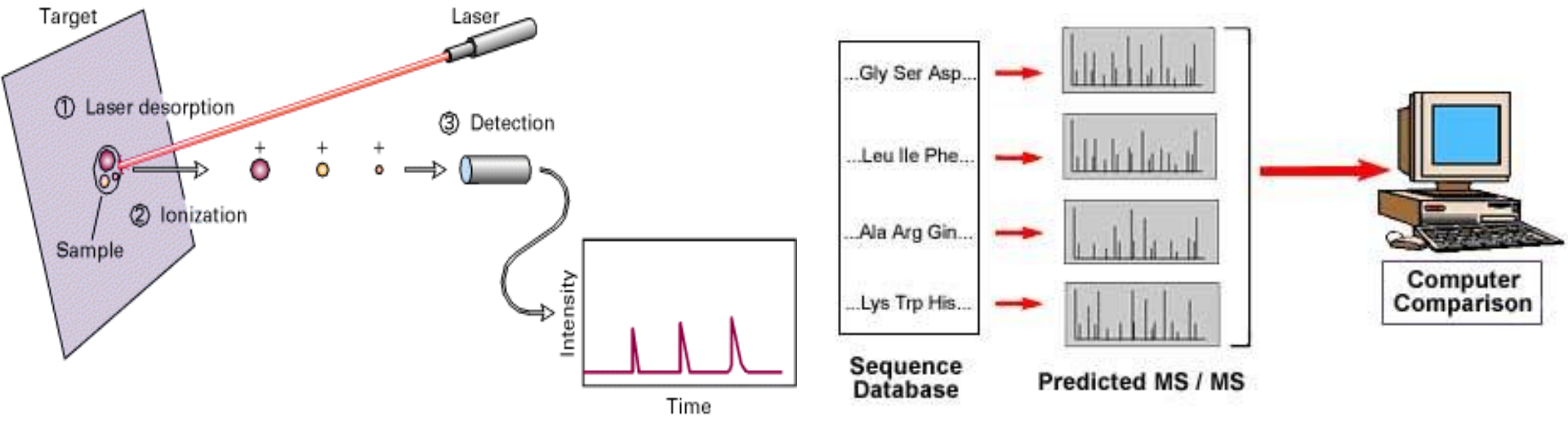




Comparing Proteomes For Differences in Protein Expression

Comparing Different Sample Types For Changes in Protein Levels

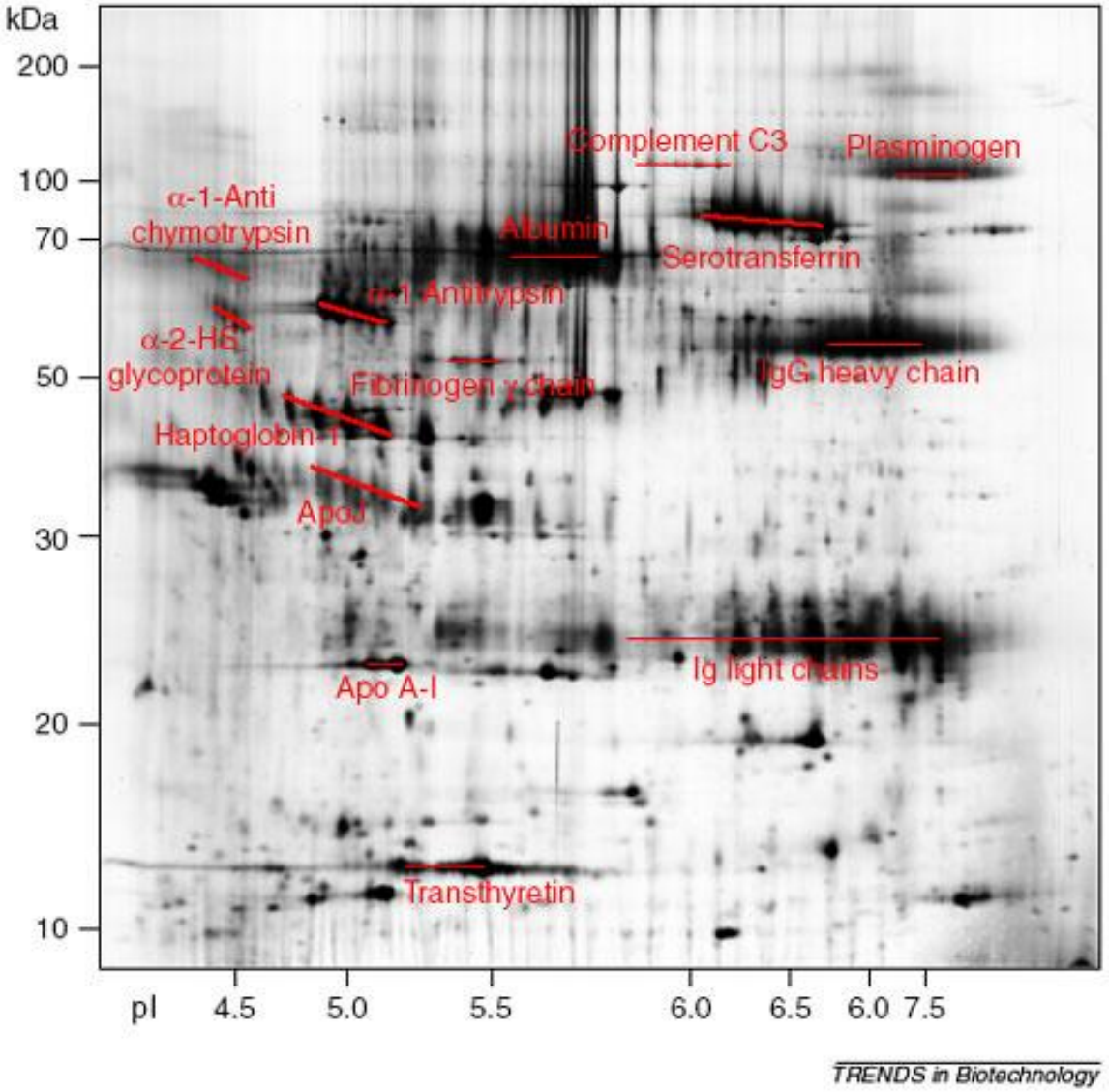
Mass Spectrometry



- **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