Bio390: Introduction to Bioinformatics

Protein Structure and Interactions

12.11.2018

Elif Ozkirimli

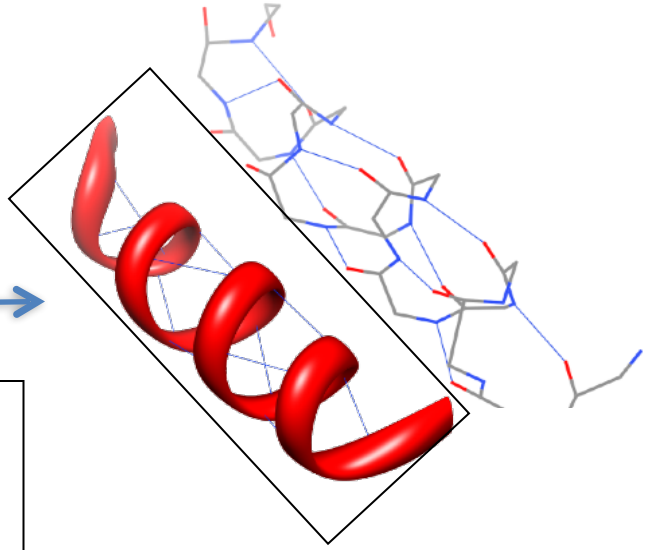
e.ozkirimli@bioc.uzh.ch

Four Levels of Protein Structure

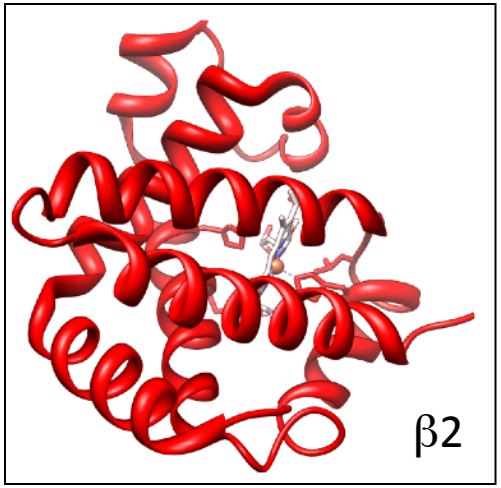
- ***Primary, 1°***

TPEEKSAVTALWGKV

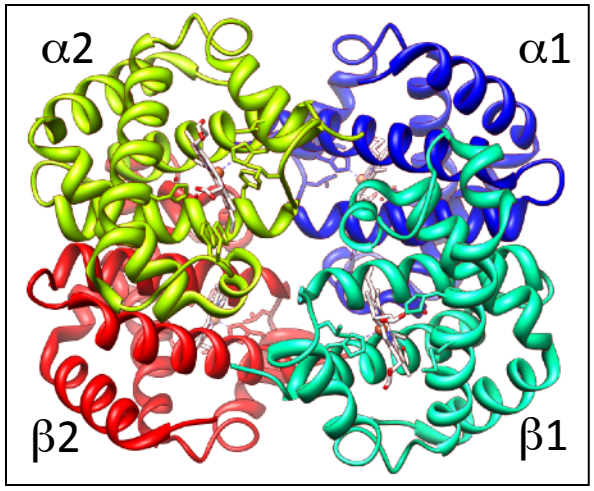
- ***Secondary, 2°***



- ***Tertiary, 3°***

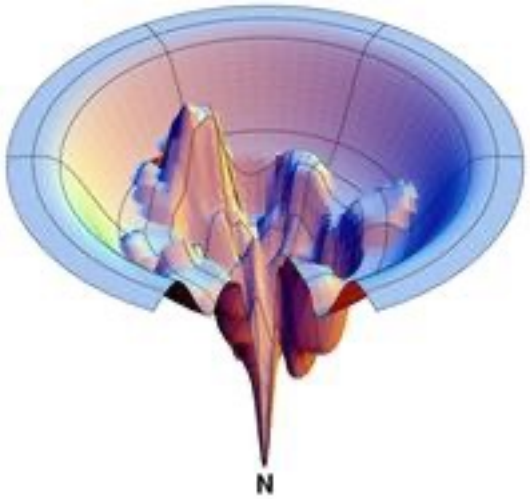


- ***Quaternary, 4°***



Protein folding

- Consider a 100 residue protein. If each residue can take only 3 positions, there are $3^{100} = 5 \times 10^{47}$ possible conformations.
  - If it takes 10^{-13}s to convert from 1 structure to another, exhaustive search would take 1.6×10^{27} years!
- Folding must proceed by progressive stabilization of intermediates.



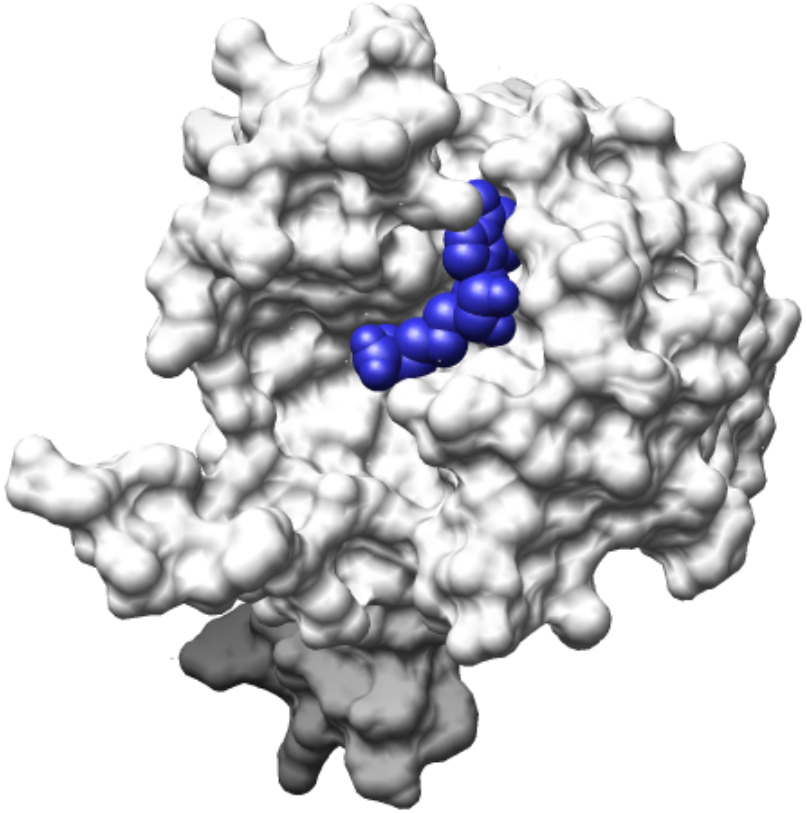
http://dillgroup.stonybrook.edu/#/landscapes

Investigating & visualizing protein structures

Protein 3D Structures

Structure – shape guides function

Binding pockets

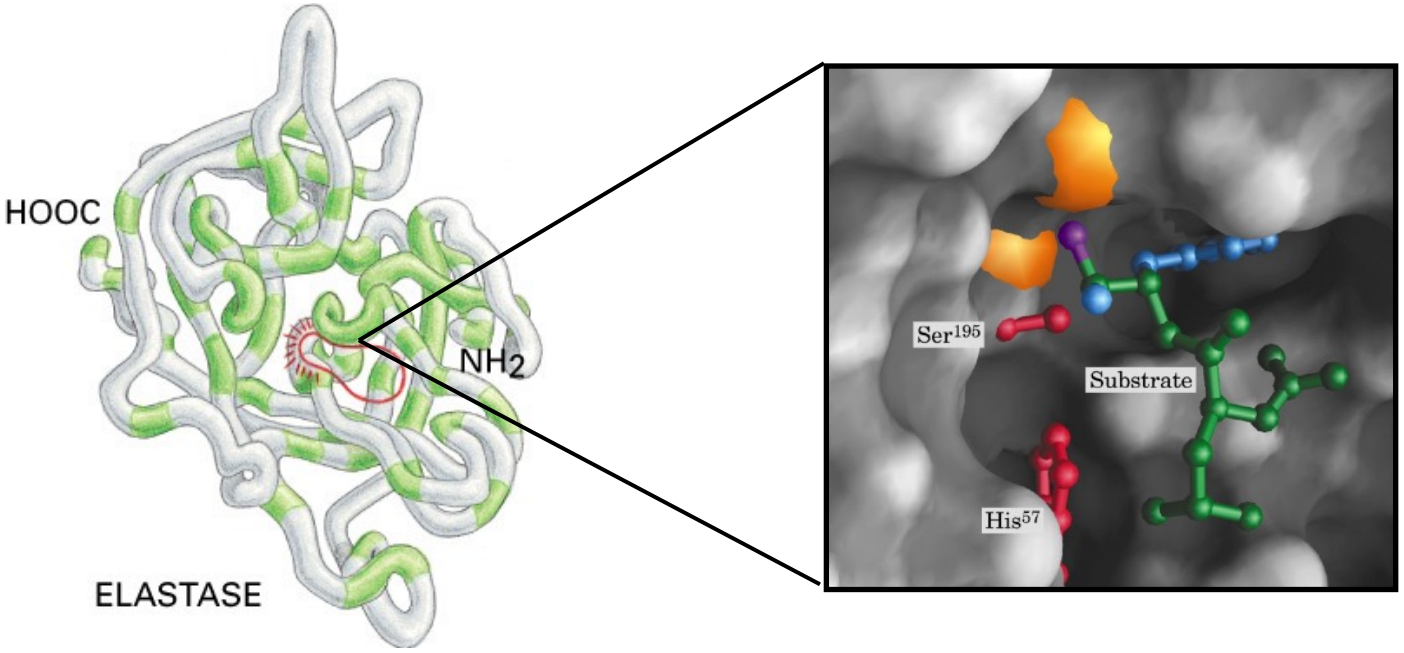


PDB ID 1nw7

Investigating & visualizing protein structures

Protein 3D Structures

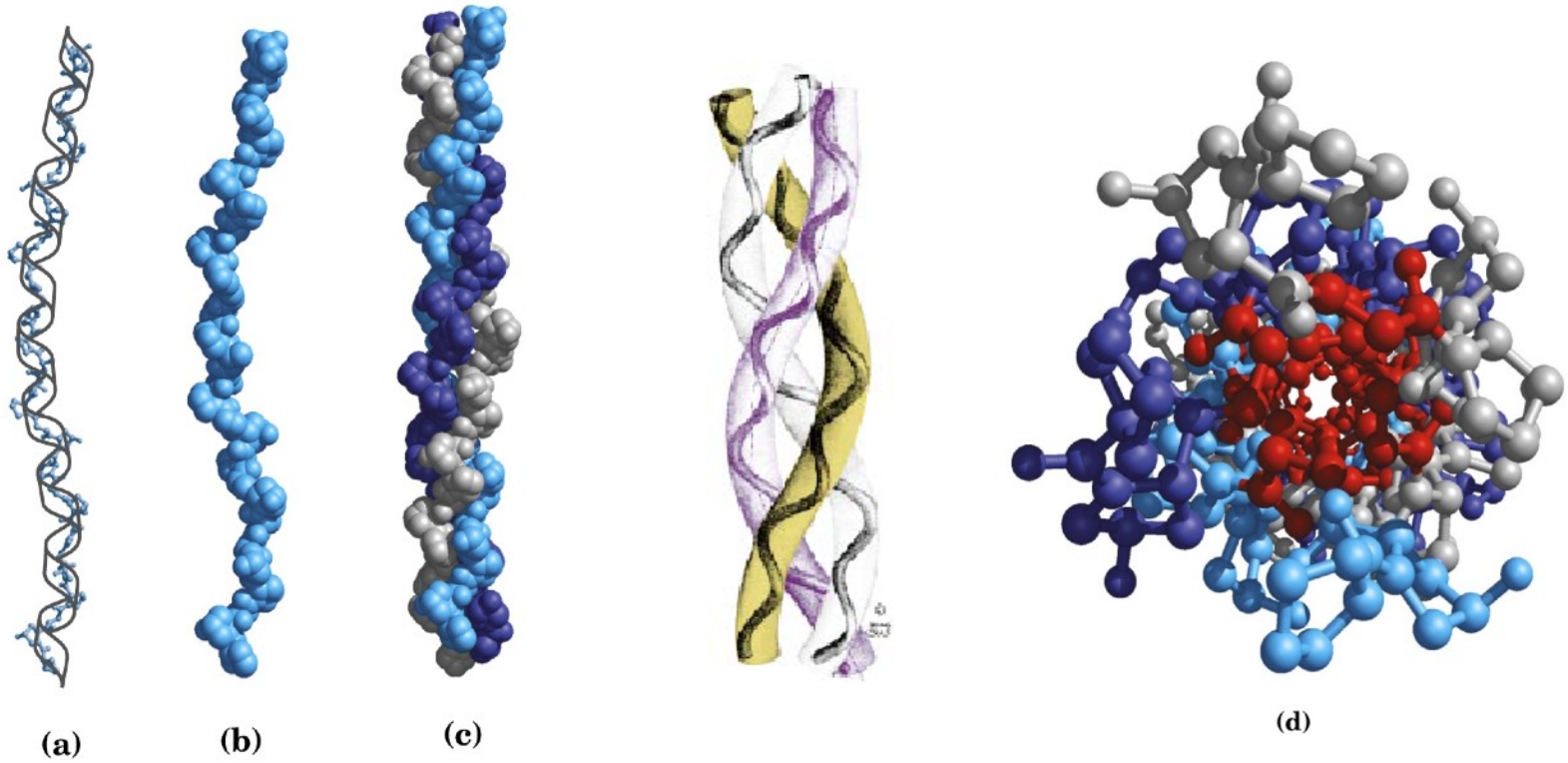
Structure – chemical properties guide function



Investigating & visualizing protein structures

Protein 3D Structures

Structure – global fold guides function

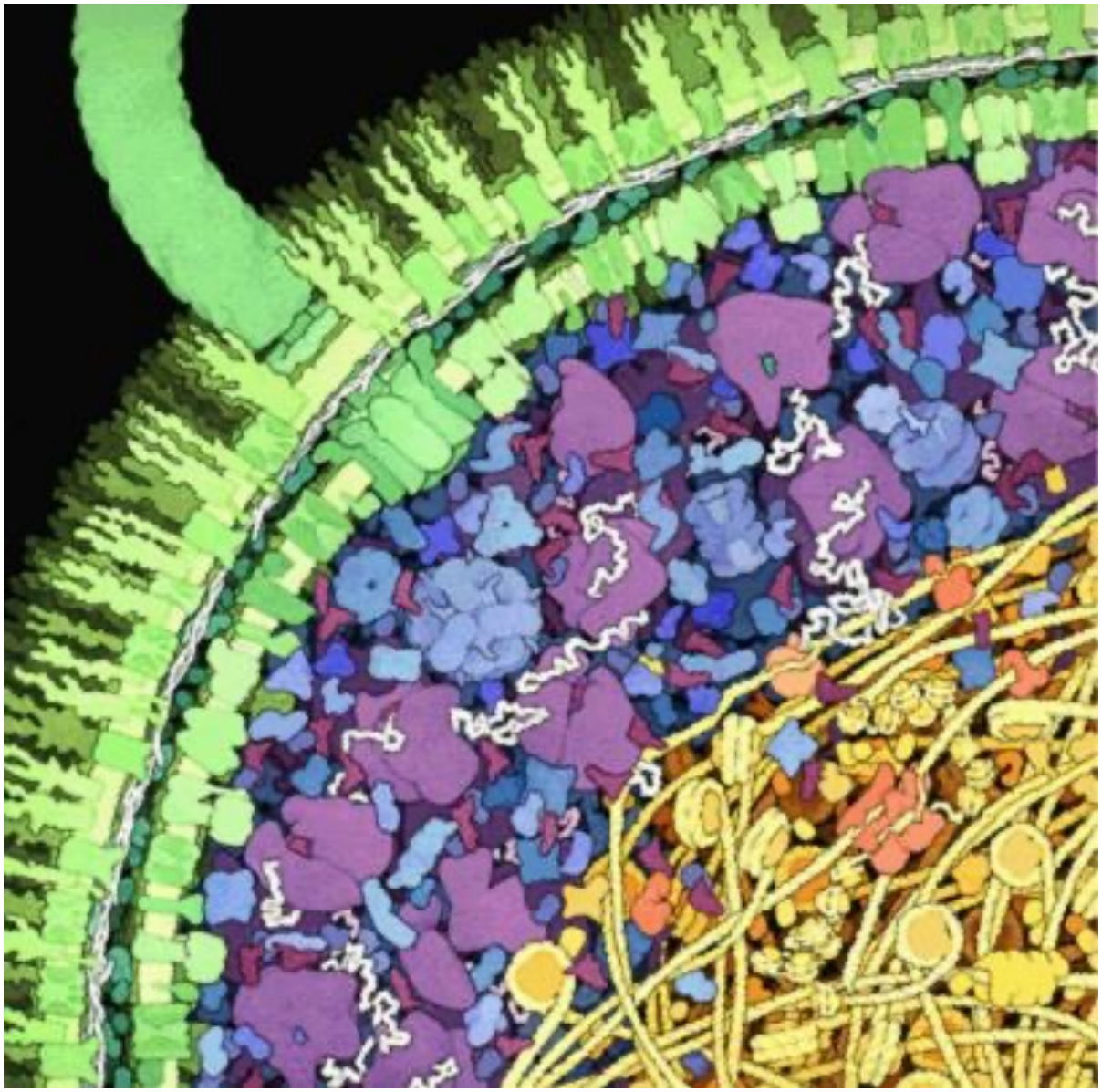


Protein interactions

- Proteins interact with
  - Other proteins (protein complexes)
  - Metabolites (enzymatic activity)
  - DNA (regulatory activity)
  - Signaling molecules (signaling activity)

All carefully regulated

Cells are crowded places!



Hopper & Mayer, 1999, Prokaryotes. Am.Sci. 87:518

Overview

- Collect PPI data
  - Experimental methods to identify PPI
  - Text mining
  - Computational tools to predict PPI
- Curate PPI data - databases
- Analyze PPI data
  - Molecular level
  - Network level
- Predict novel PPIs

Protein-protein Interactions - Methods

Y2H [47,48]	High throughput
Affinity purification–MS [61]	
DNA microarrays/Gene coexpression [113]	
Protein microarrays [114–116]	
Synthetic lethality [85,86]	
Phage display [117]	
X-ray crystallography, NMR spectroscopy [84]	Low throughput
Fluorescence resonance energy transfer [89]	
Surface plasmon resonance [91]	
Atomic force microscopy [93]	
Electron microscopy [118]	

Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e43

Protein-protein Interactions - Methods

Low throughput (molecular level)

- Small scale
- One or a few proteins at a time
- Small number of data points
- Expensive and slow

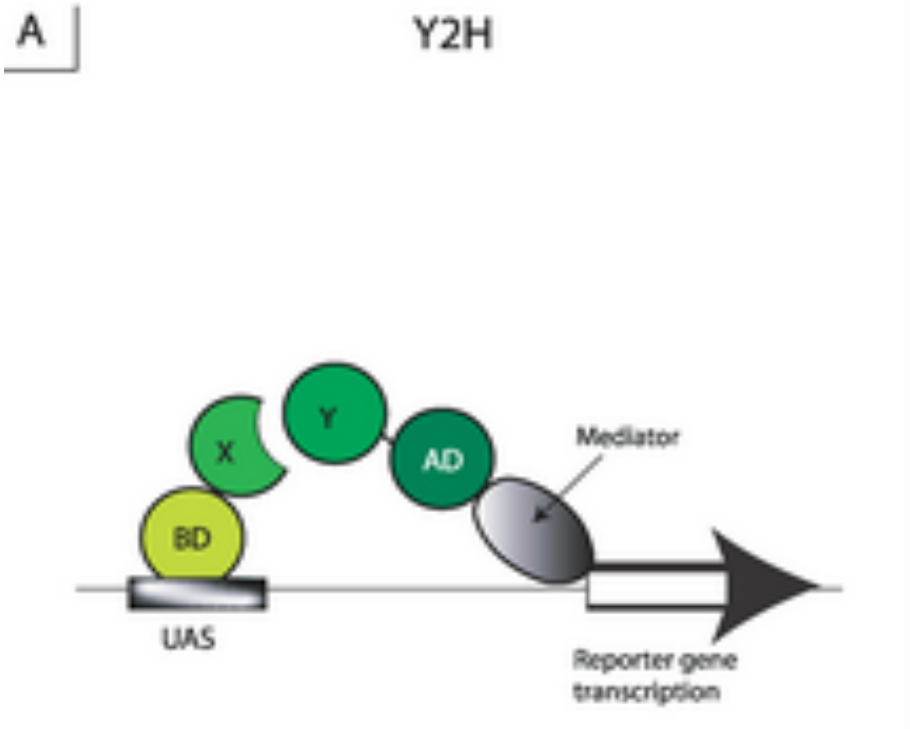
High throughput

- Hundreds/thousands of proteins
- Noisy and incomplete data
- Little overlap across data sets

Protein-protein Interactions - Methods

- Various techniques are used to investigate protein-protein interactions, including:
- Biochemical/biophysical methods (*in vitro*)
  - ➢Isothermal calorimetry
  - ➢Surface plasmon resonance (e.g. BIACore)
  - ➢Mass spectrometry e.g. from protein complexes
  - ➢"Pull-down" assays – one protein can be bound by an antibody (immunoprecipitation) or via a "tag"
- Molecular/cellular biological methods (*in vivo*)
  - ➢Two-hybrid experiments
  - ➢Fluorescent proteins

Experimental Techniques for High-Throughput PPI analysis



Most transcription activators have at least two distinct domains, one that directs binding to a promoter DNA sequence (BD) and another that activates transcription (AD)

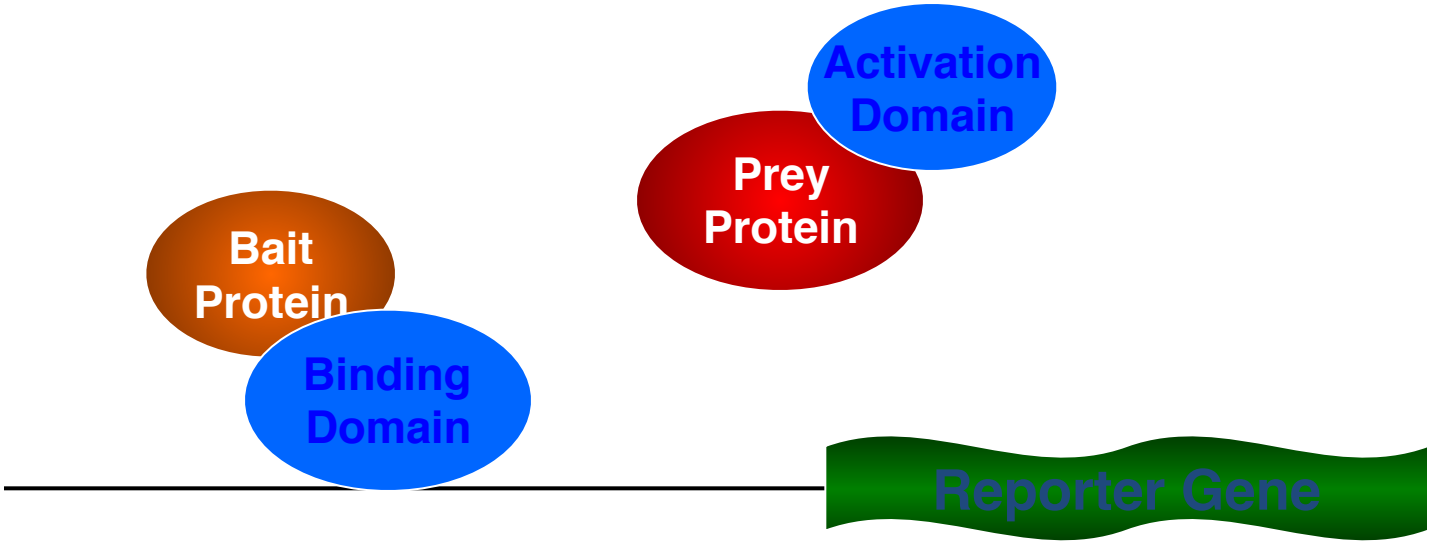
In Y2H, X is fused to BD (bait). Y is fused to AD (prey).

If X and Y interact, AD and BD are brought together and the gene is activated.

Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e42

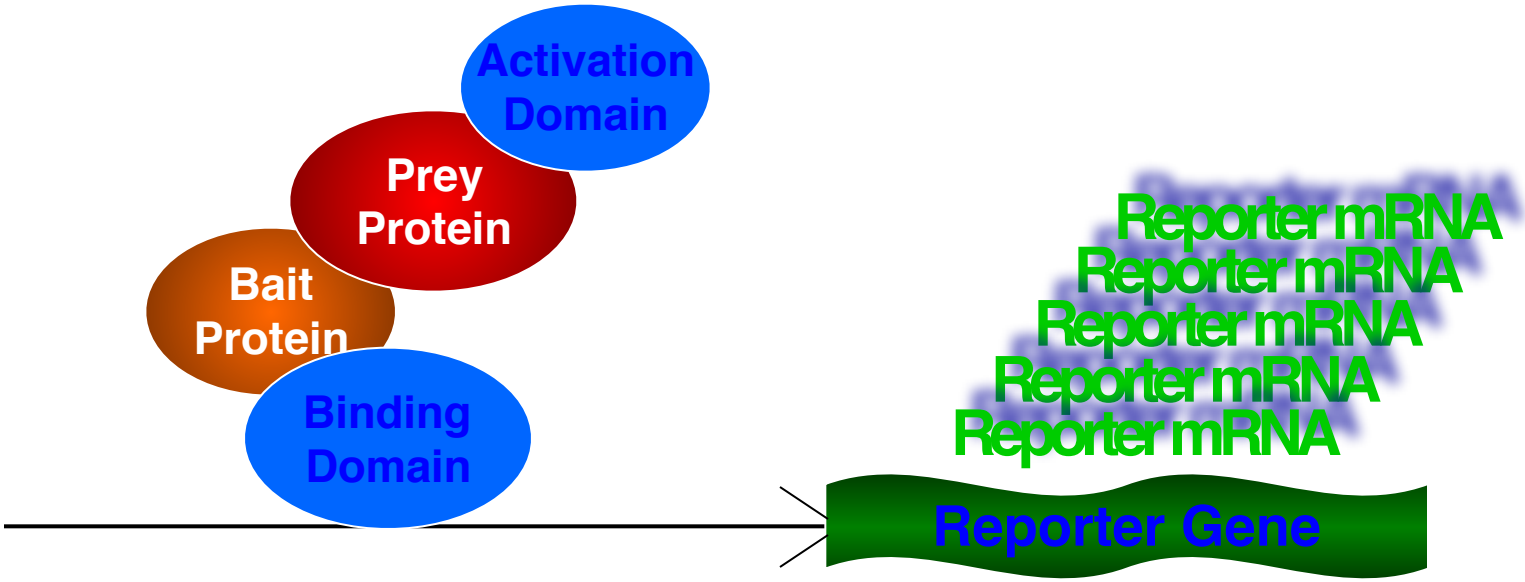
The Two-Hybrid System

- Two hybrid proteins are generated with transcription factor domains
- Both fusions are expressed in a yeast cell that carries a reporter gene whose expression is under the control of binding sites for the DNA-binding domain



The Two-Hybrid System

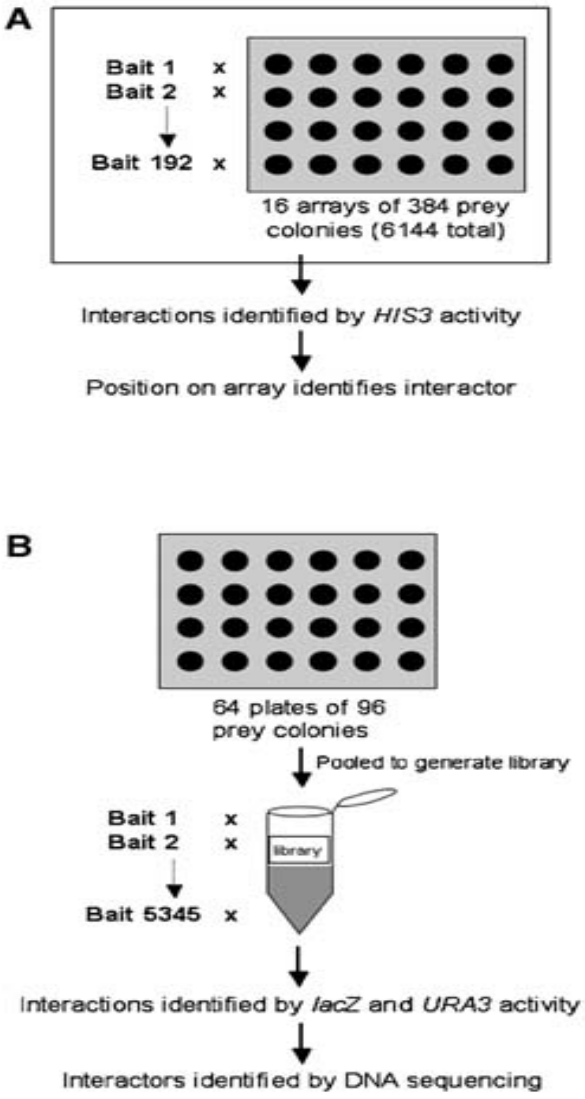
- Interaction of bait and prey proteins localizes the activation domain to the reporter gene, thus activating transcription.
- Since the reporter gene typically codes for a survival factor, yeast colonies will grow only when an interaction occurs.



Genome-wide analysis by Y2H

- Matrix approach: a matrix of prey clones is added to the matrix of bait clones. Diploids where X and Y interact are selected based on the expression of a reporter gene.

- Library approach: one bait X is screened against an entire library. Positives are selected based on their ability to grow on specific substrates.



Adapted from B. Causier, Mass Spectroscopy Reviews, 2004

Limitations of Y2H

- Fusion of a protein into chimeras can change the structure of a target
- Protein interactions can be different in yeast and the organisms where the genes came from
- It is difficult to target extracellular proteins
- It is hard to detect interactions between proteins active only in a *complex*
- Proteins which can interact in two-hybrid experiments, may never interact *in vivo*

Identifying Protein–Protein Interactions

Protein complex isolation

- Epitope tag one protein in the complex
- Isolate tagged protein to isolate stably interacting proteins
- Separate and identify all isolated proteins

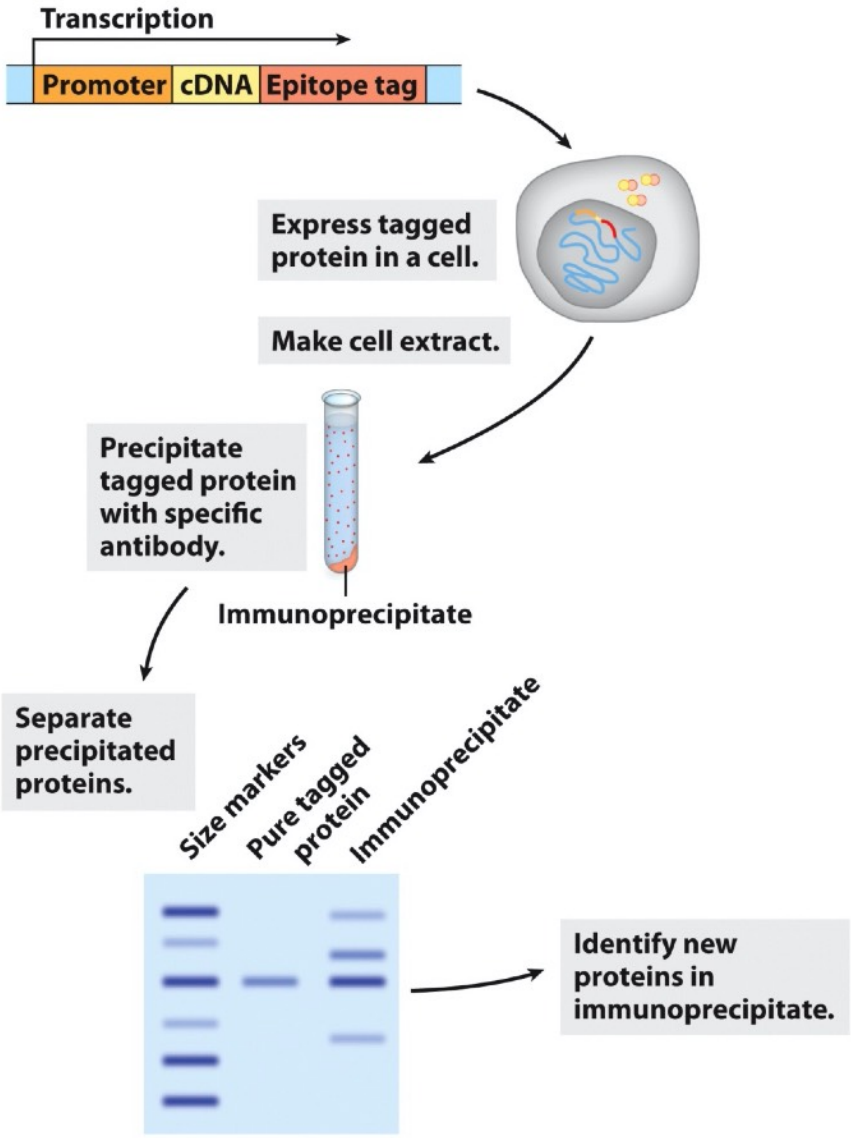


Figure 9-19
Lehninger Principles of Biochemistry, Sixth Edition

Tandem Affinity Purification

A target gene is fused with the TAP tag (protein A + calmodulin binding peptide) and is expressed in yeast.

Protein forms native complexes with other proteins.

1st step of purification: protein A interacts with IgG matrix.

Eluate incubated with calmodulin coated beads.

After washing: protein complexes harvested and identified by MS.

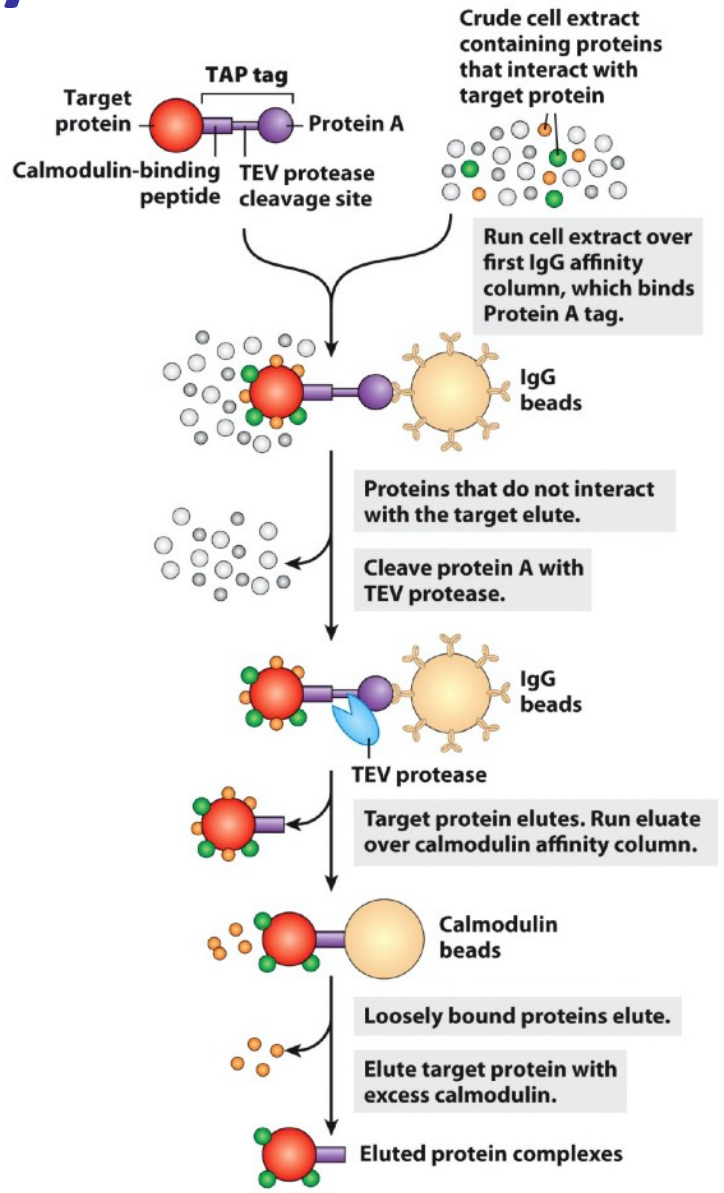
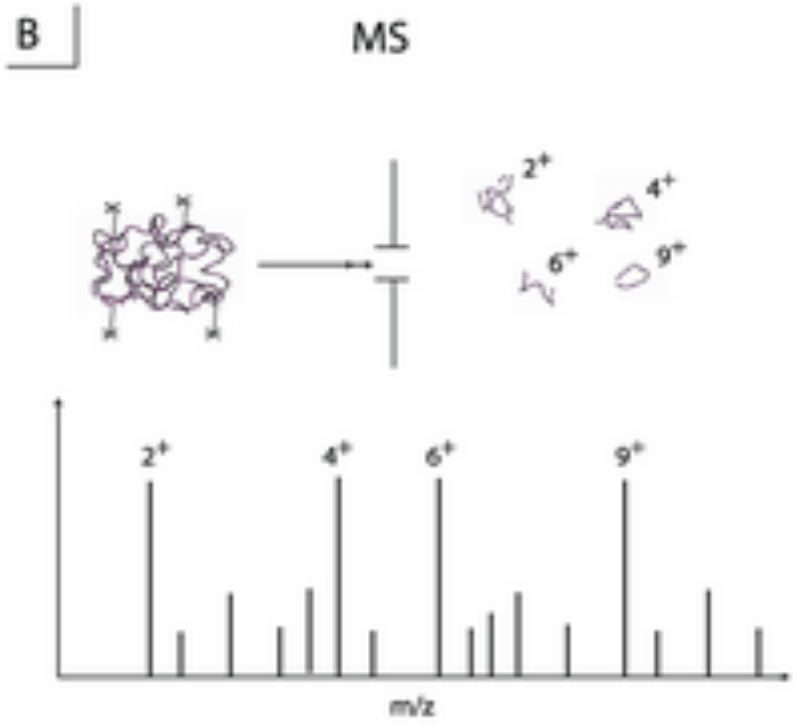


Figure 9-20
Lehninger Principles of Biochemistry, Sixth Edition

Experimental Techniques for High-Throughput PPI analysis

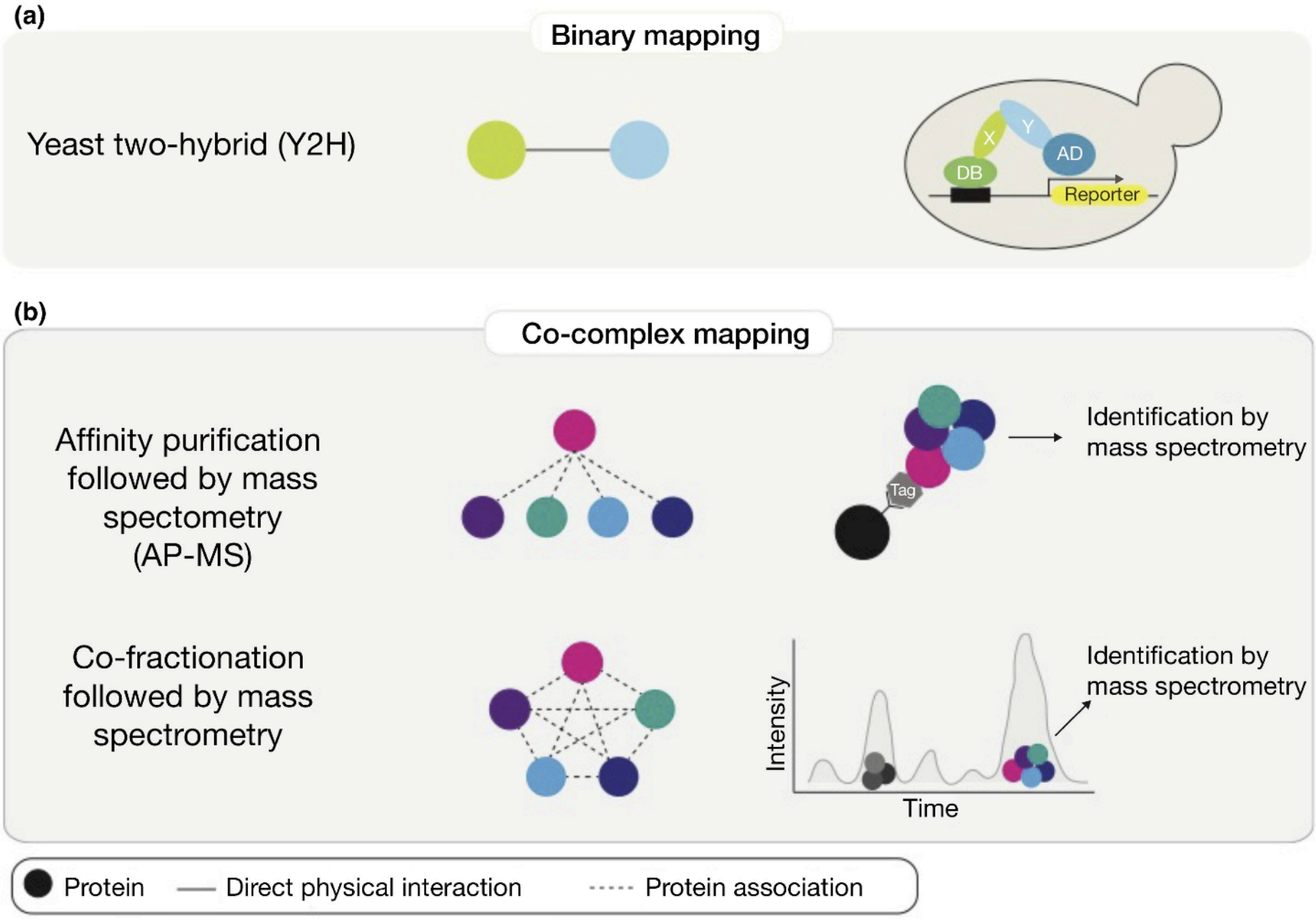


Mass spectroscopy

Protein complexes are purified.

MS identifies polypeptide sequence.

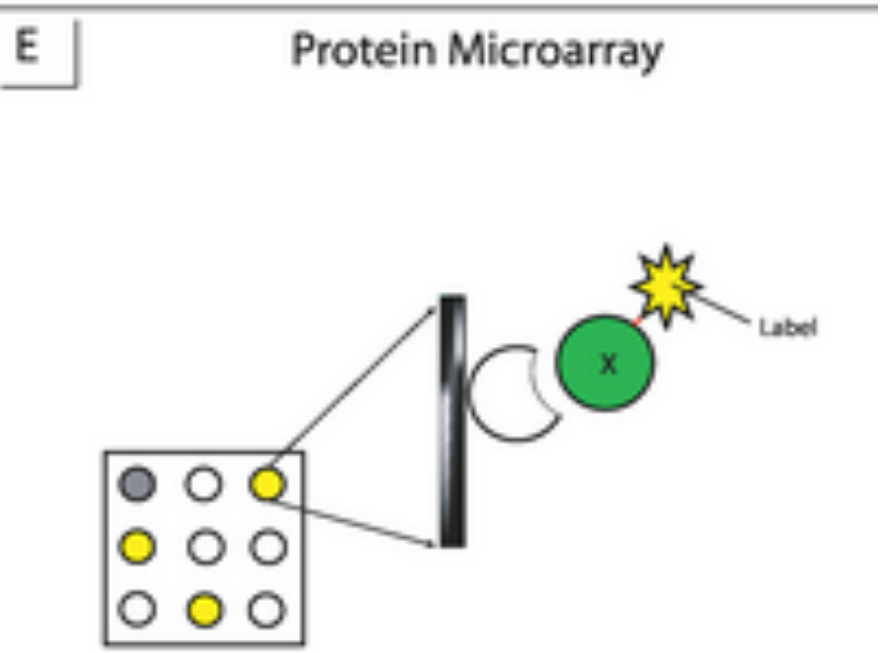
Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e42


(a)
Binary mapping
Yeast two-hybrid (Y2H)
X
Y
AD
DB
Reporter
(b)
Co-complex mapping
Affinity purification followed by mass spectometry (AP-MS)
Tag
Identification by mass spectrometry
Co-fractionation followed by mass spectrometry
Intensity
Time
Identification by mass spectrometry
Protein
Direct physical interaction
Protein association

Differences and similarities between Y2H and MS-TAP

- TAP permits protein *complexes* to be isolated, but cannot detect weak/transient PPIs
- Both methods generate a lot of false positives, only ~50% interactions are biologically significant
- Y2H is an *in vivo* technique
- MS can detect large stable complexes and networks of interactions

Experimental Techniques for High-Throughput PPI analysis



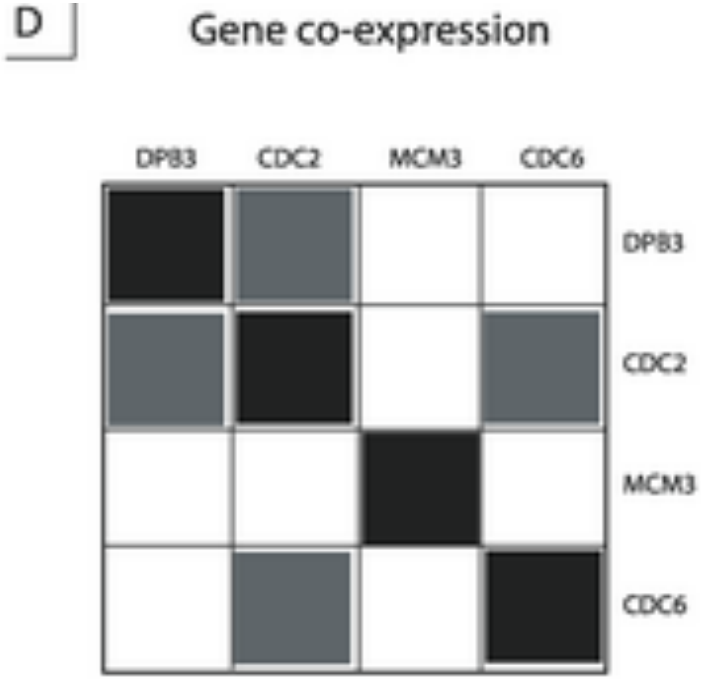
Protein microarrays (protein chips) can detect interactions between actual proteins rather than genes: target proteins immobilized on the solid support are probed with a fluorescently labeled protein.

Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e42

Protein-protein Interactions

- Direct methods
  - Yeast 2 hybrid
  - TAP tagging
  - Mass spectrometry
- Indirect
  - Correlated mRNA expression profiles
  - Genetic interaction data
  - Interolog analysis
  - Co-localization, co-expression
  - Correlated mutations

Experimental Techniques for High-Throughput PPI analysis



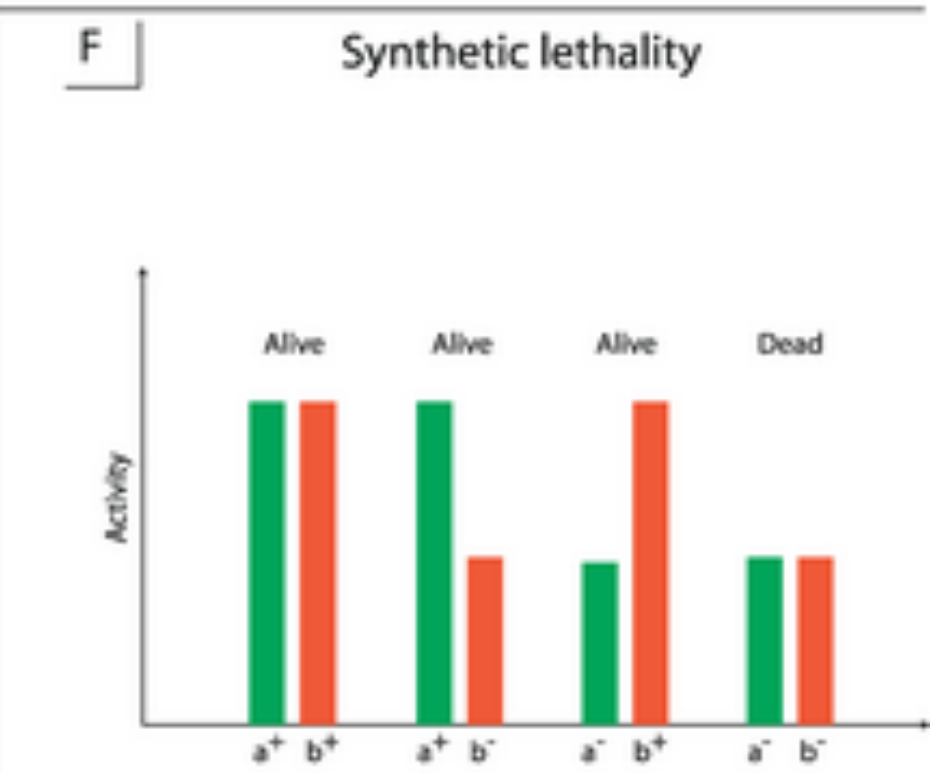
Hypothesis: Functional subunits of a protein complex should be present in stoichiometric amounts in cell.

Gene coexpression analysis produces a correlation matrix where the dark areas show high correlation between expression levels of corresponding genes.

Especially high correlation for permanent complexes (ribosome, proteasome).

Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e42

Experimental Techniques for High-Throughput PPI analysis



Synthetic lethality method describes the genetic interaction when two individual, nonlethal mutations result in lethality when combined together (a– b–).

Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e42

Study of Protein-protein Interactions *In Vivo*

- Fluorescent techniques:
  - ➢FRET – fluorescence resonance energy transfer; reports on distance between 2 fluorophores
  - ➢Fluorescent reporters – expressed proteins emit fluorescence at specific wavelength
  - ➢FRAP (FLIP) – fluorescence recovery after photobleaching (fluorescence loss in photobleaching); allows movement of reporters to be monitored

Bioinformatics - tasks

- Collect PPI information
- Analyze PPIs
  - Molecular level

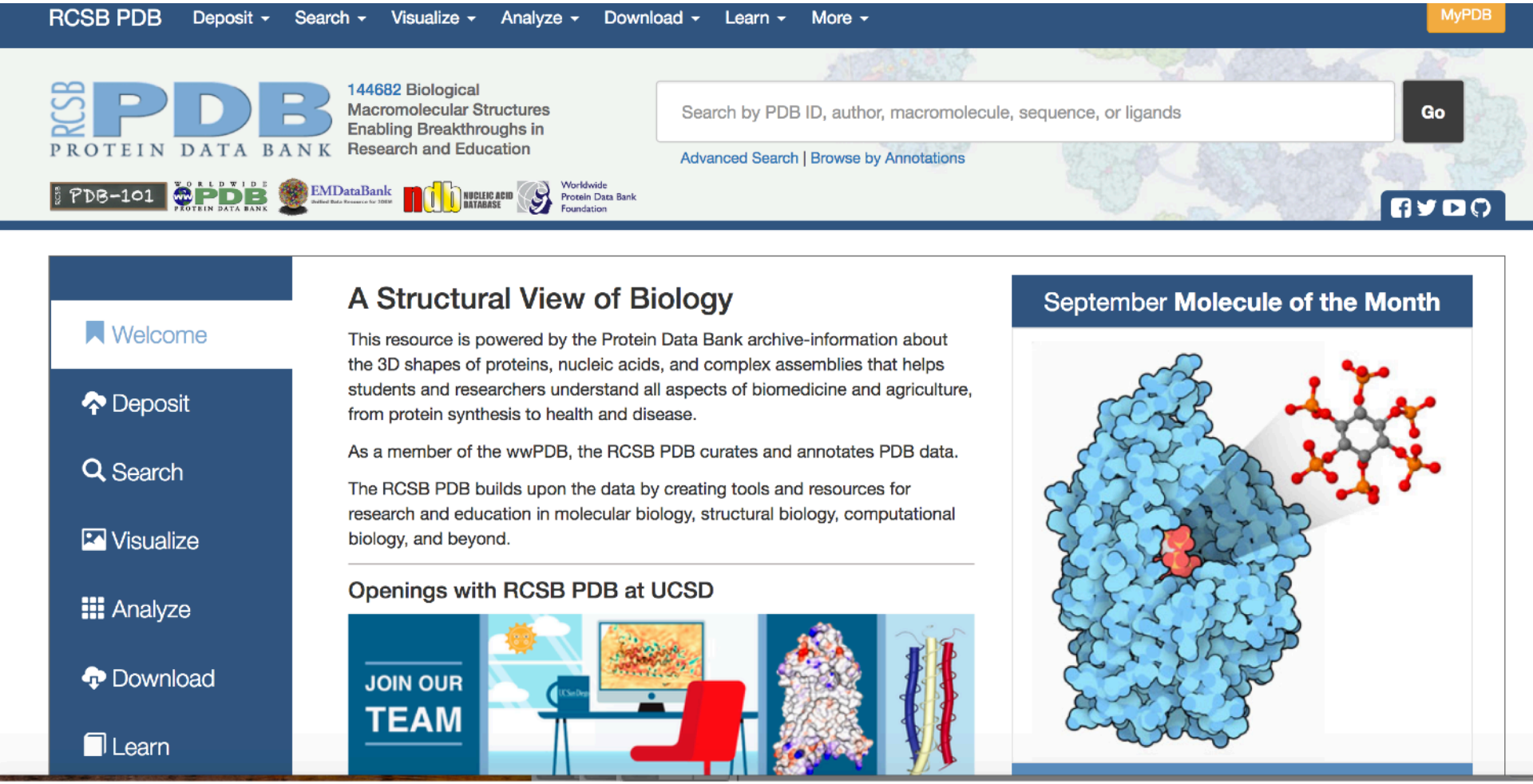
Identifying Regions Involved in Protein-protein Interactions

- Once protein-protein interactions have been identified, it is important to establish how the interactions occur i.e. what regions or specific amino acids are important for the interaction?
- Determine effect of mutations on the protein-protein interaction to identify critical residues for binding
- Results usually confirmed by more than one experimental technique

Structural information on protein complexes

- Protein Data Bank: maintained by the Research Collaboratory of Structural Bioinformatics(RCSB)
  - http://www.rcsb.org/pdb/
  - As of Sunday Sep 30, 2018 there are 144,682 Structures
  - including structures of Protein/Nucleic Acid Complexes, Nucleic Acids, Carbohydrates
  - The number of protein complexes is around 26,000.
- Each structure has a PDB ID: a 4 character unique identifier
- Most structures are determined by X-ray crystallography (about 130000 entries). Some other methods are NMR (about 12000 entries) and electron microscopy (EM) (about 2500 entries).

Structural information



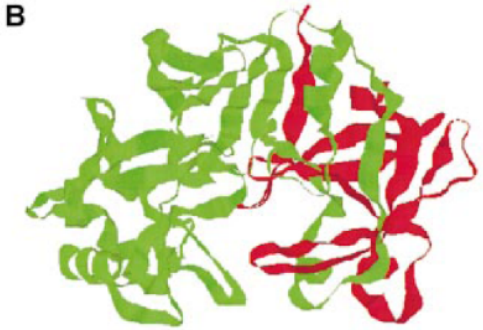
Types of protein-protein interactions (PPI)

Obligate PPI

the protomers are **not** found as stable structures on their own *in vivo*

Obligate homodimer

P22 Arc repressor DNA-binding



Obligate heterodimer

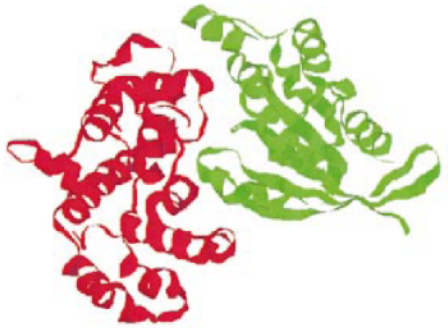
Human cathepsin D

1LYB

Non-obligate PPI

Non-obligate homodimer

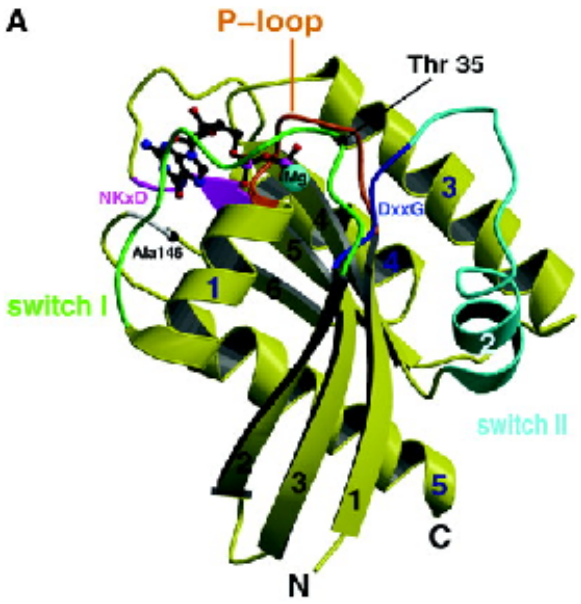
Sperm lysin

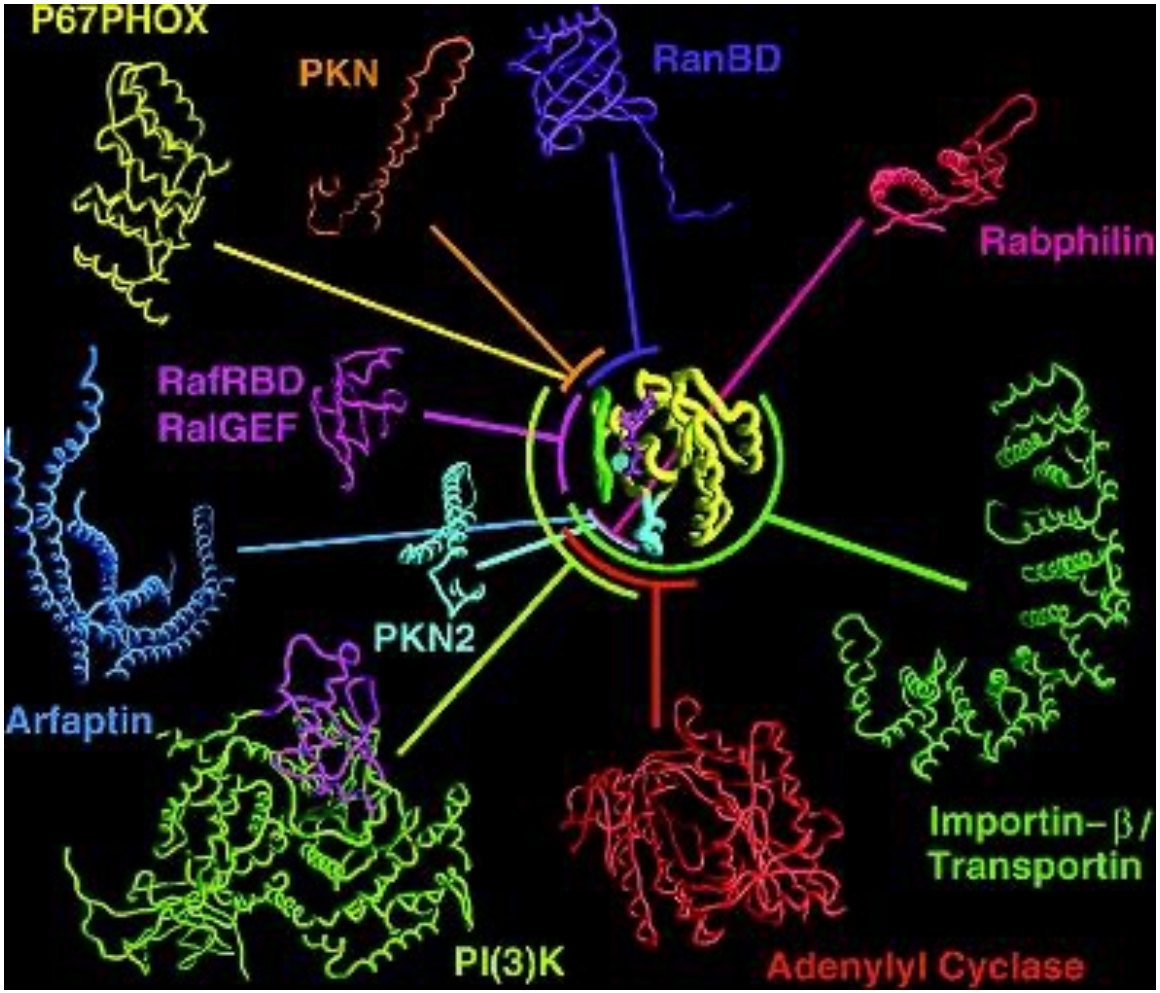


Non-obligate heterodimer

RhoA and RhoGAP signaling complex

Multiple interactions: Guanine-nucleotide binding protein

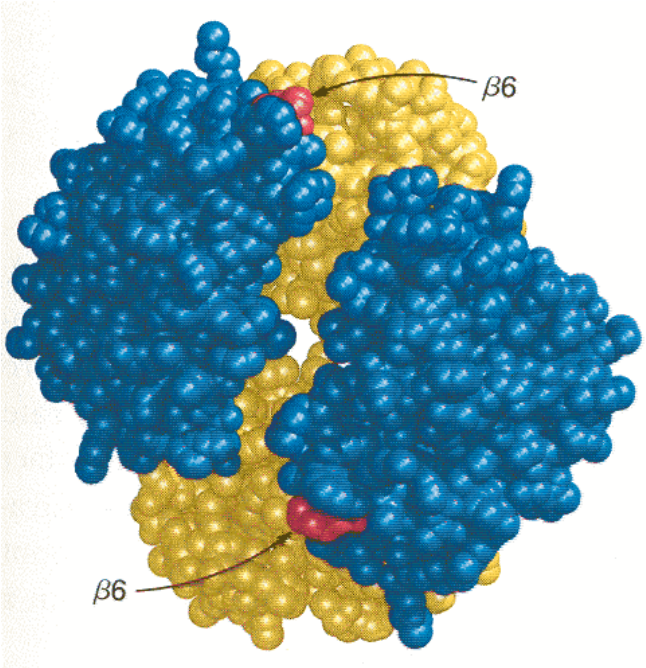




Adapted from Vetter & Wittinghofer, *Science* 2001

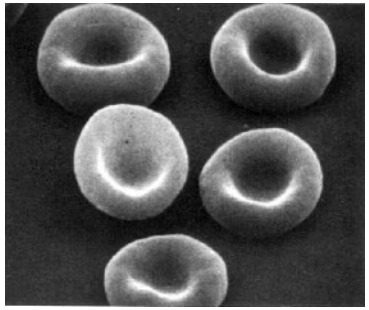
Effect of a single mutation
Sickle Cell Anemia

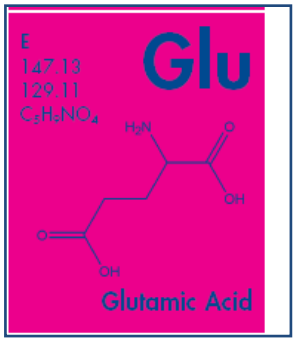
- Hemoglobin is the protein in red blood cells (erythrocytes) responsible for binding oxygen

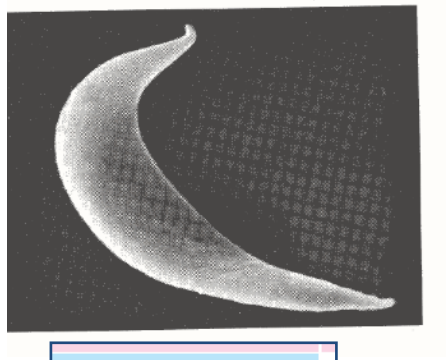


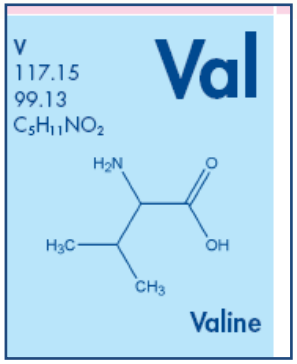
Sickle Cell Anemia

- The mutation E→V in the β chain replaces a charged Glu by a hydrophobic Val on the surface of hemoglobin
- The resulting "sticky patch" causes hemoglobin to agglutinate (stick together) and form fibers which deform the red blood cell and do not carry oxygen efficiently
- Sickle cell anemia was the first identified molecular disease





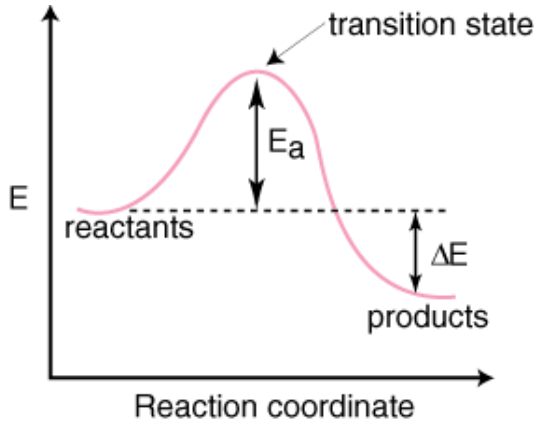




Protein protein complex formation driven by encounter, concentration of components and free energy of complex relative to unbound form

PPIs can be controlled by

- Altering local concentration of the protein components
- Influencing the binding affinity, determined by the physicochemical and geometrical interface properties



General Properties of Protein-Protein Binding Sites

- More **hydrophobic** than the rest of the protein surface.
- Relatively **flat** (except for enzyme-substrate binding sites) with a core (buried) and rim (solvent accessible)
- The binding free energy is not distributed equally across the protein interfaces: a small subset of residues at the interfaces forms energy **hot spots**, enriched in tyrosine, tryptophan, and arginine.
- **Structurally conserved residues**, especially polar resides, correspond to the energy hot spots.
- Most sites are **specific** but some have **promiscuous** binding characteristics

Bioinformatics - tasks

- Collect PPI information
- Analyze PPIs
  - Molecular level
- Curate PPIs in a database

PPI Databases

- DIP
  - http://dip.doe-mbi.ucla.edu/
- MIPS (small scale)
  - http://mips.gsf.de/proj/ppi/
- BIND (PPI, Prot-DNA, Prot-SM)
  - http://www.bind.ca (now owned by Unleashed)
- OPHID (predicted interactions)
  - http://ophid.utoronto.ca/ophid/
- MINT - Molecular Interactions Database
  - http://mint.bio.uniroma2.it/mint/Welcome.do
- IntAct (EBI)
  - http://www.ebi.ac.uk/intact/site/
- InterDom (domain interactions)
  - http://interdom.lit.org.sg/
- STRING (EMBL)
  - http://string.embl.de/

PPI databases

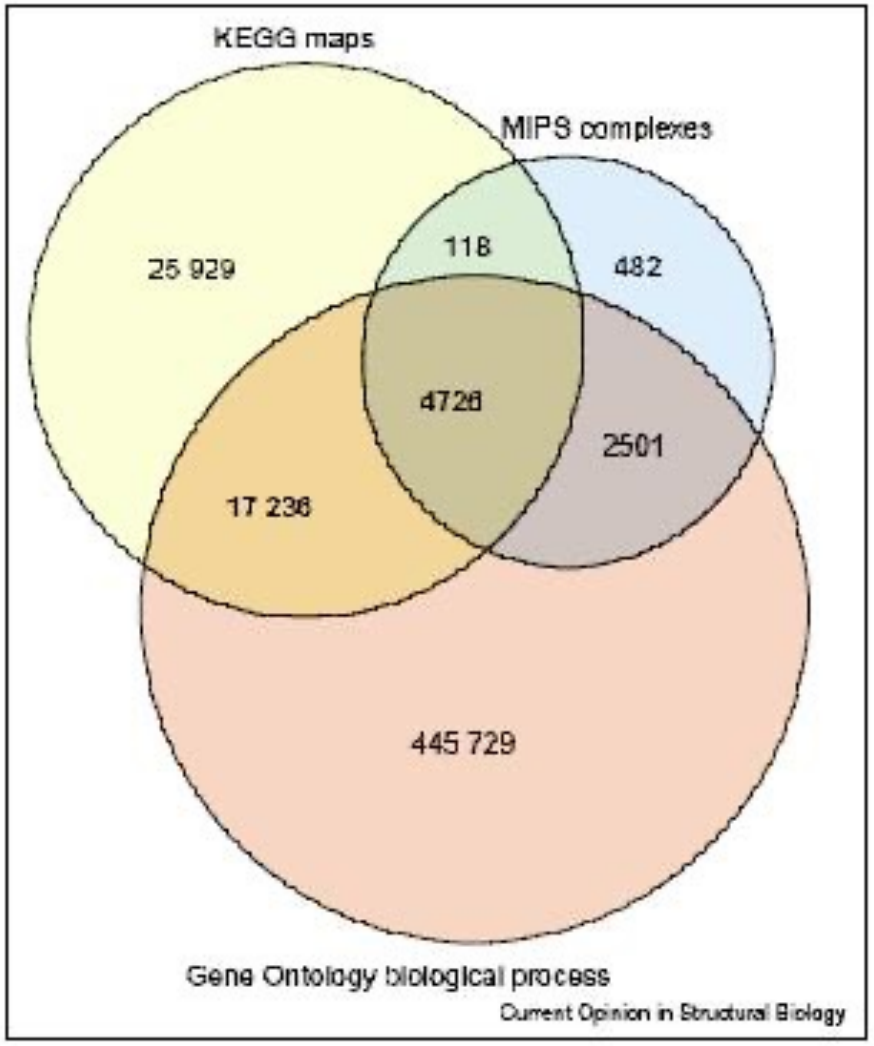
Types

- Experiment (E)
- Structure detail (S)
- Predicted
  - Physical (P)
  - Functional (F)
- Curated (C)
- Homology modeling (H)
- *International Molecular Exchange (IMEx) consortium

Database	Proteins/Domains	Type
DIP*, LiveDIP	P	E,S
BIND*	P	E,C,S
MPact/MIPS*	P	E,C,F
STRING	P	E,P,F
MINT*	P	E,C
IntAct*	P	E,C
BioGRID*	P	E,C
HPRD	P	E,C
3did, Interprets	D	S,H
Pibase, ModBase	D	S,H
CBM	D	S
iPfam	D	S
InterDom	D	P
DIMA	D	F,S
Prolinks	P	F

Comparing the DBs

- High FP rate in high-throughput experiments
- Disagreement between benchmark sets
- Experimental PPI data is sparse relative to all PPIs, so dataset overlap is small and hard to confirm with multiple sources



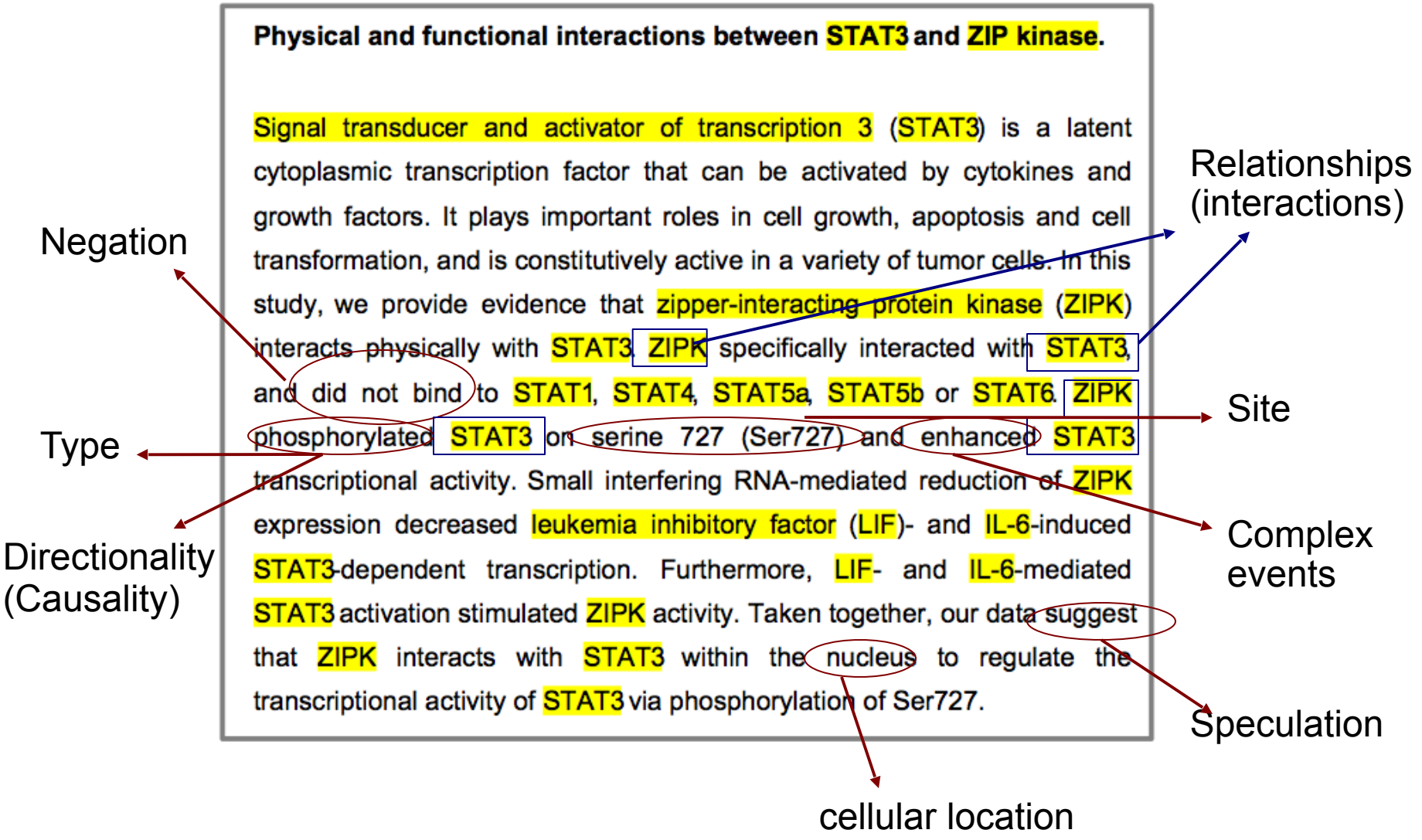
Bioinformatics - tasks

- Collect PPI information
- Analyze PPIs
  - Molecular level
- Curate PPIs in a database
- Extract information about PPI from texts

Text Mining

- Searching Medline or PubMed for words or word combinations
- Co-occurrence of terms is the simplest metric, yet lends to a higher FP rate
- Natural Language Processing (NLP) methods are more specific (e.g., “X binds to Y”; “X interacts with Y”; “X associates with Y” etc.) yet are difficult to detect so it has a higher FN rate
- Normally requires a list of known gene names or protein names for a given organism

What can be extracted



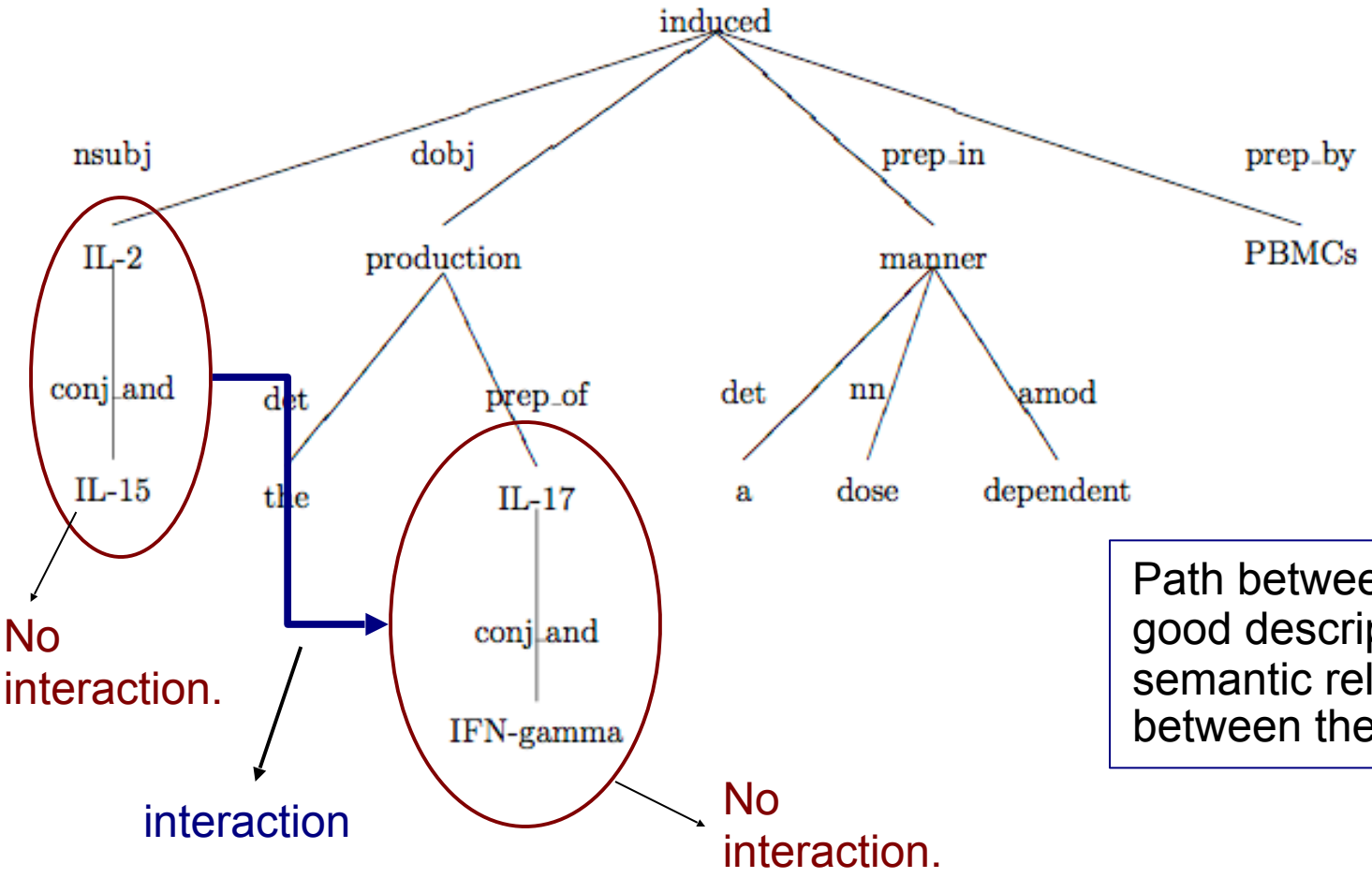
NLP pipeline

Tokenization	• What are the words? • Break sentence into words
Morphology	• Named entity recognition (NER) • Part of speech tagging (classification)
Syntax	• Dependency parsing • Using grammar rules
Semantics	• Codependency resolution • Use context info
Applications	• Machine translation • Language generation + dialog

Interaction Extraction

IL-2 and IL-15 induced the production of IL-17 and IFN-gamma in a dose dependent manner by PBMCs.

(Genia Tagger: 71% F-measure)



Path between proteins: good description of semantic relation between them.

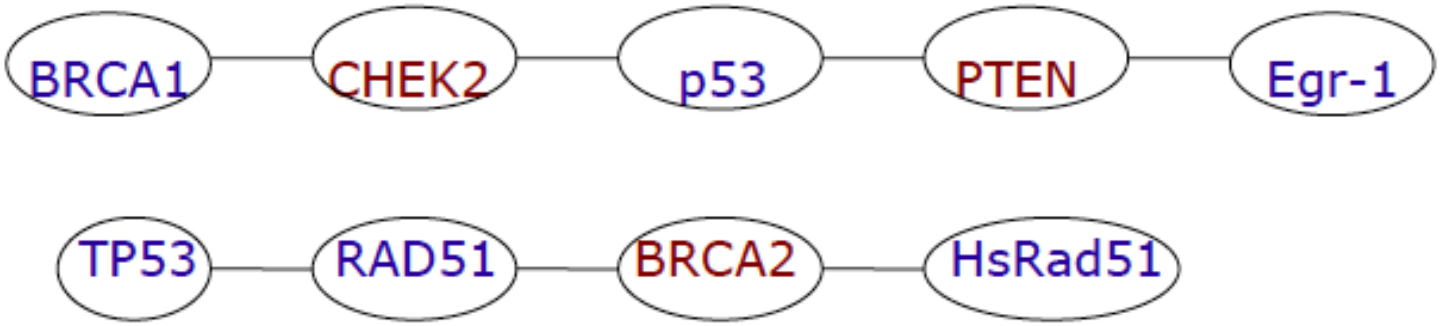
Stanford Parser is used to generate the dependency parse trees (de Marneffe et al., 2006).

Constructing the Interaction Network

Sample extracted interaction sentences:

- PTEN is transcriptionally regulated by transcription factors such as p53 and Egr-1.
- In response to DNA damage, the cell-cycle checkpoint kinase CHEK2 can be activated by ATM kinase to phosphorylate p53 and BRCA1, which are involved in cell-cycle control and apoptosis.
- The interactions of RAD51 with TP53, RPA and the BRC repeats of BRCA2 are relatively well understood (see Discussion).
- The interaction of BRCA2 with HsRad51 is significantly more different to both RadA and RecA (Figure 2c).

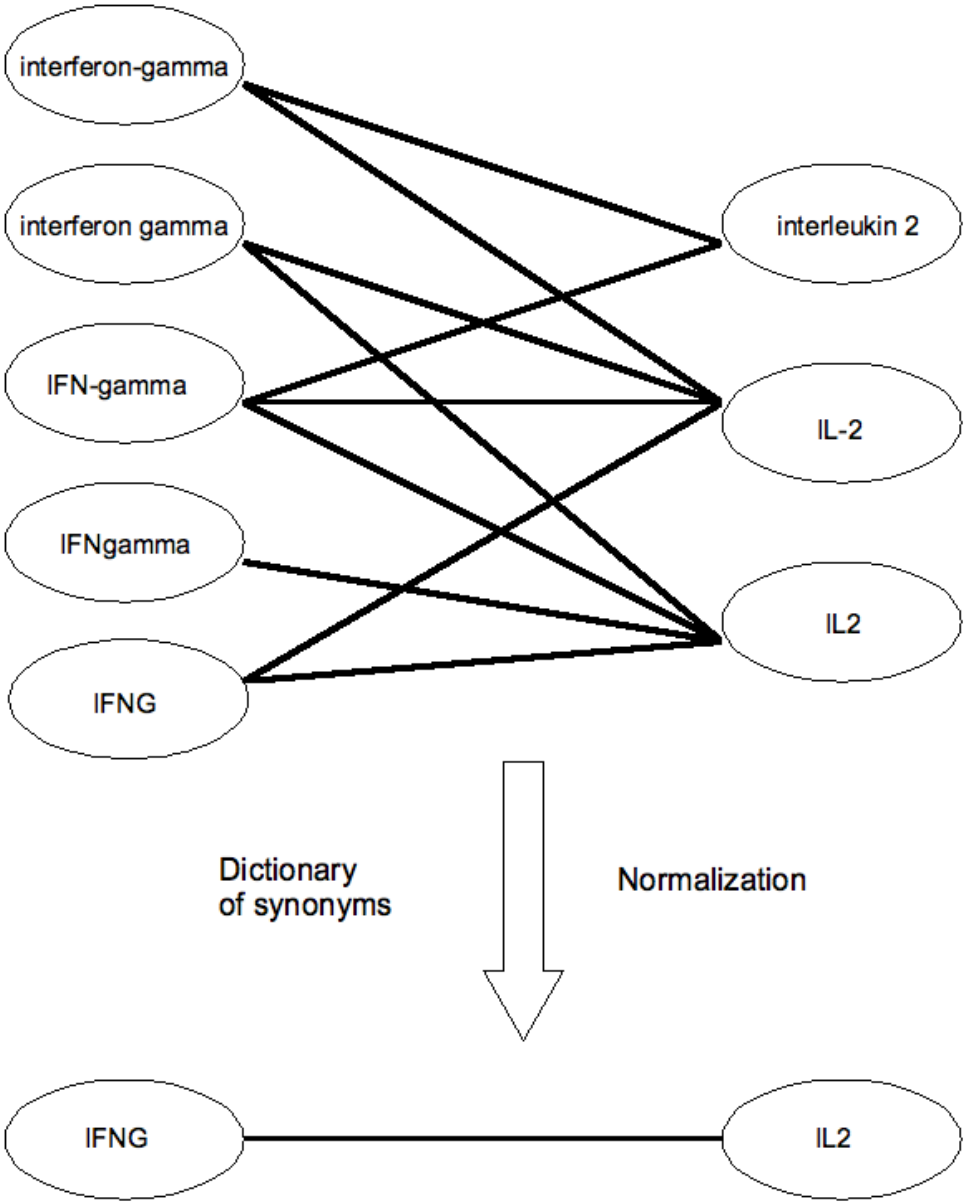
- The constructed graph:



seed genes:

CHEK2
PTEN
BRCA2

Gene Name Normalization

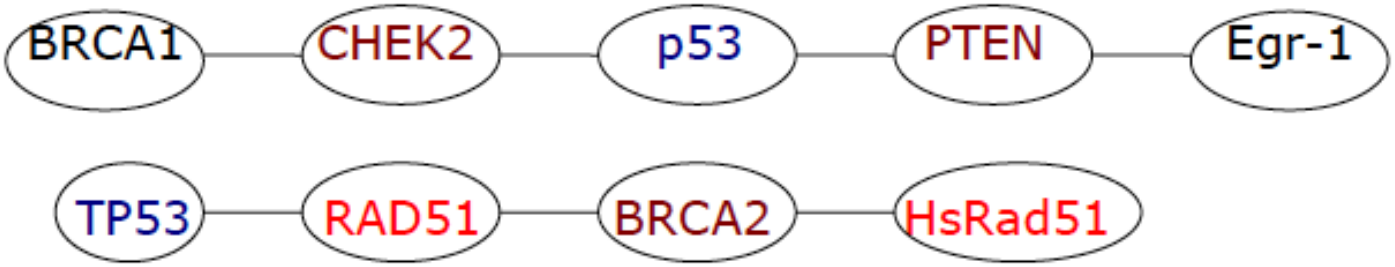


HUGO Gene Nomenclature Committee (HGNC) database used as the dictionary for gene names and their synonyms. ~28,000 gene records

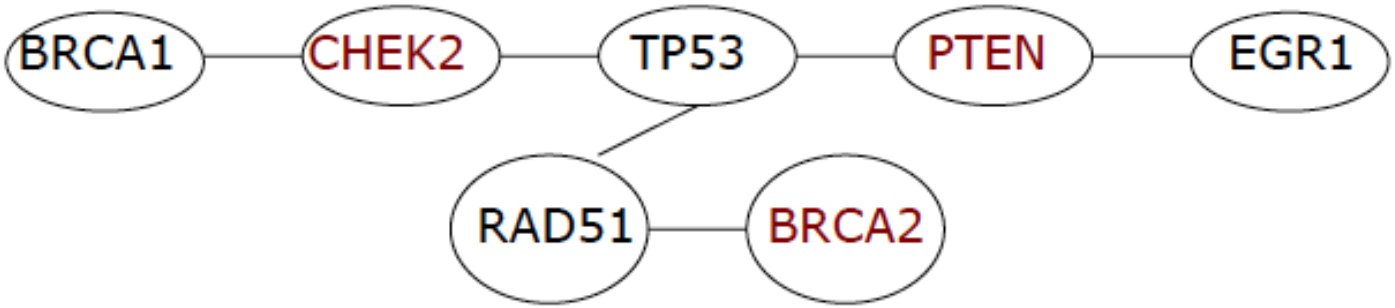
(http://www.genenames.org/)

Constructing the Interaction Network

- The graph before gene name normalization:



- The graph after gene name normalization:



Pre-BIND

- Used Support Vector Machine (SVM) to scan literature for PPIs
- Precision, accuracy and recall of 92% for correctly classifying PPI abstracts
- Estimated to capture 60% of all abstracted protein interactions for a given organism

Donaldson *et al*. BMC Bioinformatics 2003 4:11

Bioinformatics - tasks

- Collect PPI information
- Analyze PPIs
  - Molecular level
- Curate PPIs in a database
- Extract information about PPI from texts
- Analyze PPIs
  - Systems level

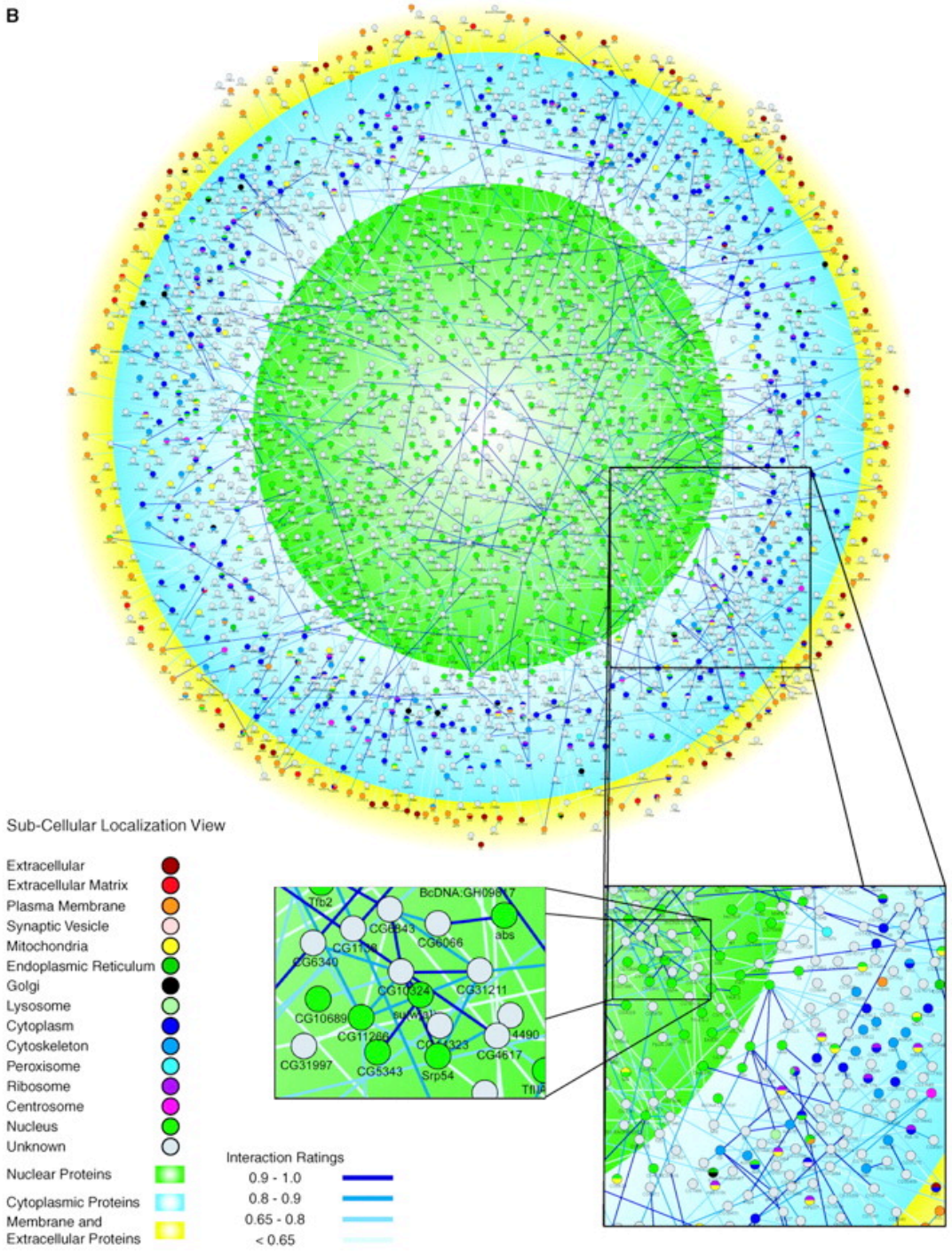
Assessment of Protein-protein Interaction Data

- Currently believed that yeast has >30,000 different interactions (for ~6,000 proteins)
- Overall conclusion is: different techniques identify different complexes!
- Results from protein-protein interaction studies should be confirmed by more than one experimental technique
- Especially important for considering if *in vitro* observations are relevant for *in vivo* situations

PPI network properties

Nodes & connections

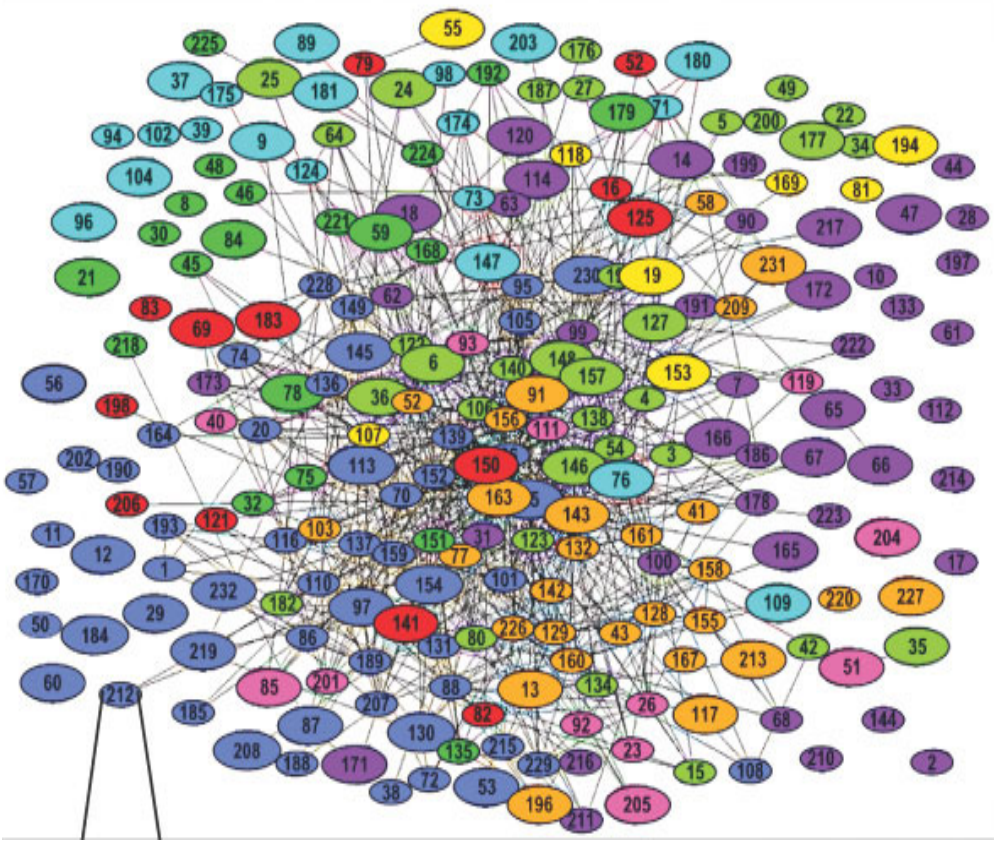
Drosophila interaction map



From:
A Protein Interaction Map of Drosophila
Giot et al. Science **302**, *1727-1136 (2003)*

Functional organization of yeast proteome: network of protein complexes

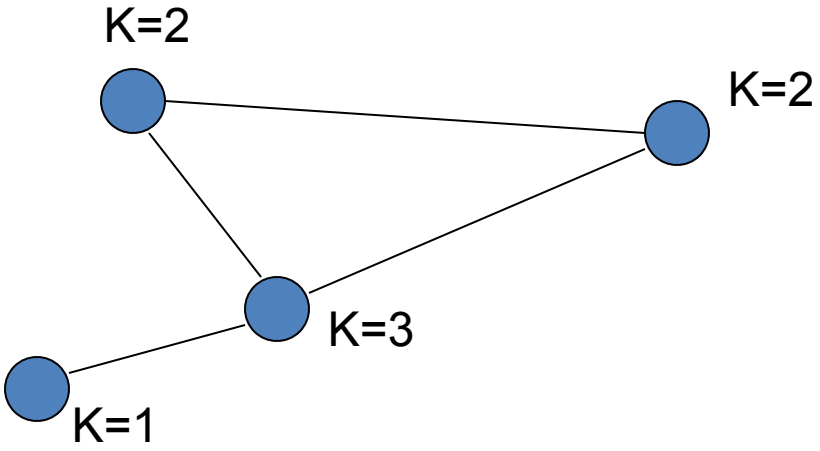
- Essential gene products are more likely to interact with essential rather than nonessential proteins
- Orthologous proteins interact with complexes enriched with orthologs



Gavin *et al*, Nature, 2002

Characteristics of networks

***n* = *nodes*, *k* = connections or "edges"**



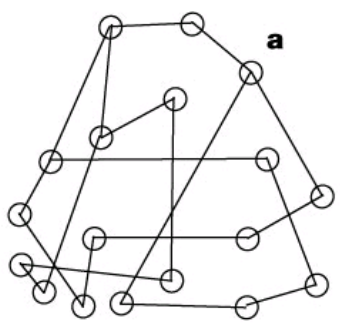
• In biology, n refers to genes/proteins (and/or metabolites) while k refers to interactions

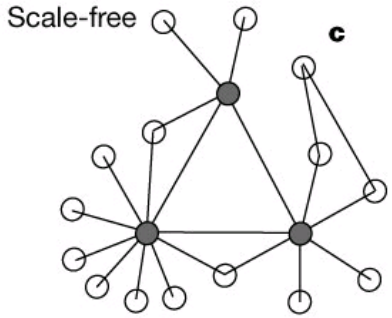
Difference between scale-free and random graph models.

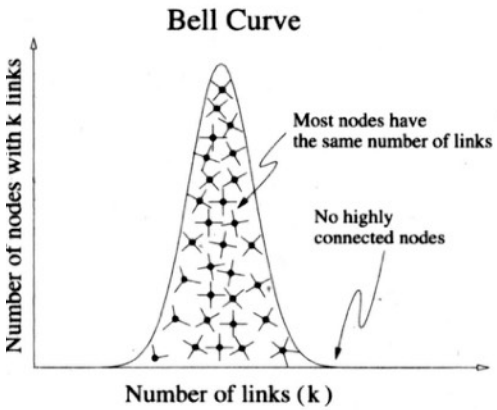
Random networks are homogeneous, most nodes have the same number of links.

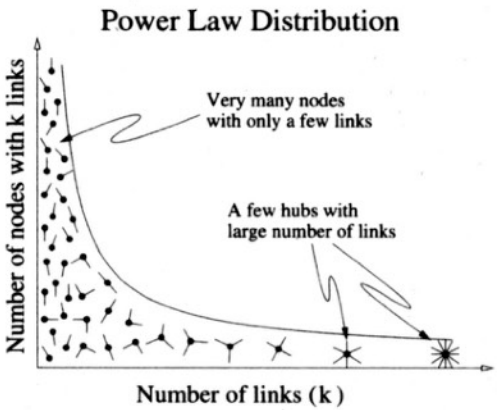
Scale-free networks have a few highly connected hubs and many nodes with few connections.

The presence of the hubs ensures that the nodes with few connections (low degree) are still connected to the network. Thus ensuring efficiency of communication.





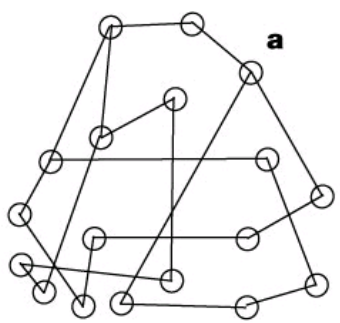


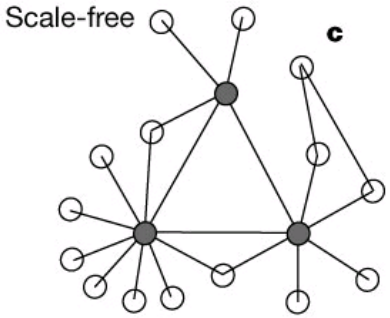


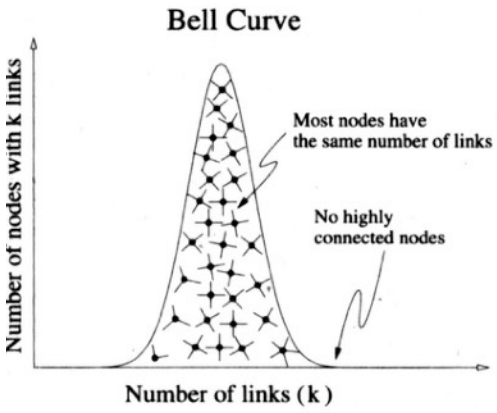
Difference between scale-free and random graph models.

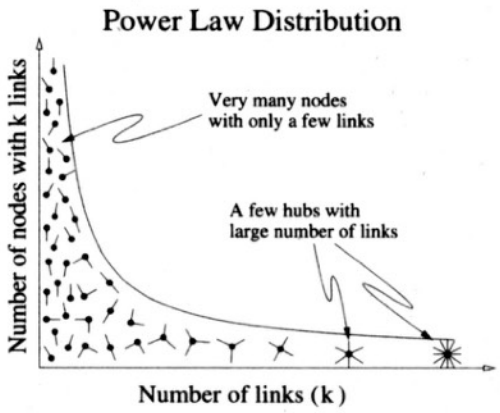
As a result, scale free networks are attack tolerant. A **random** attack does not break network down.

Attacks targeting hubs (highly connected nodes) are more dangerous.





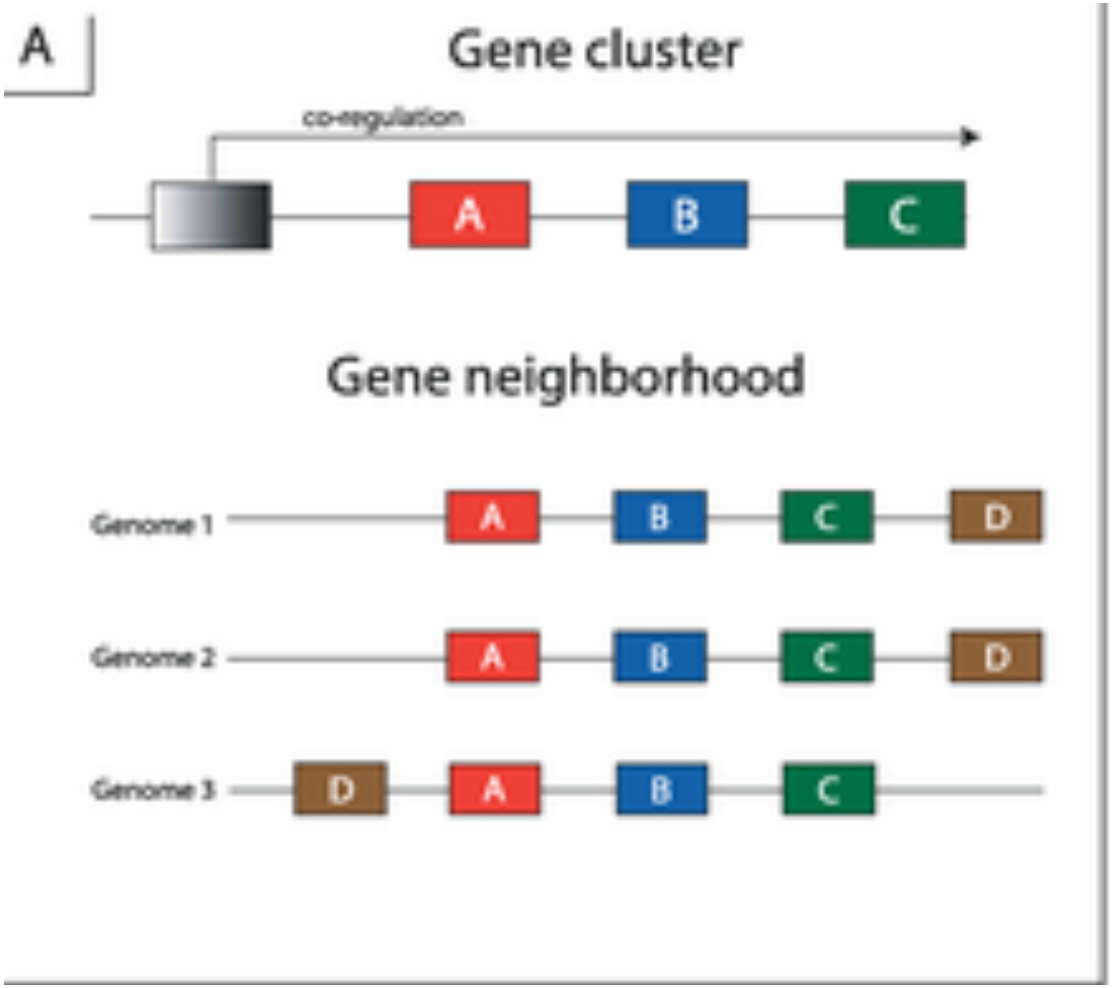




Bioinformatics - tasks

- Collect PPI information
- Analyze PPIs
  - Molecular level
- Curate PPIs in a database
- Extract information about PPI from texts
- Analyze PPIs
  - Systems level
- Prediction
  - Identify PPIs (systems level)

Computational Methods for Protein Interaction Prediction



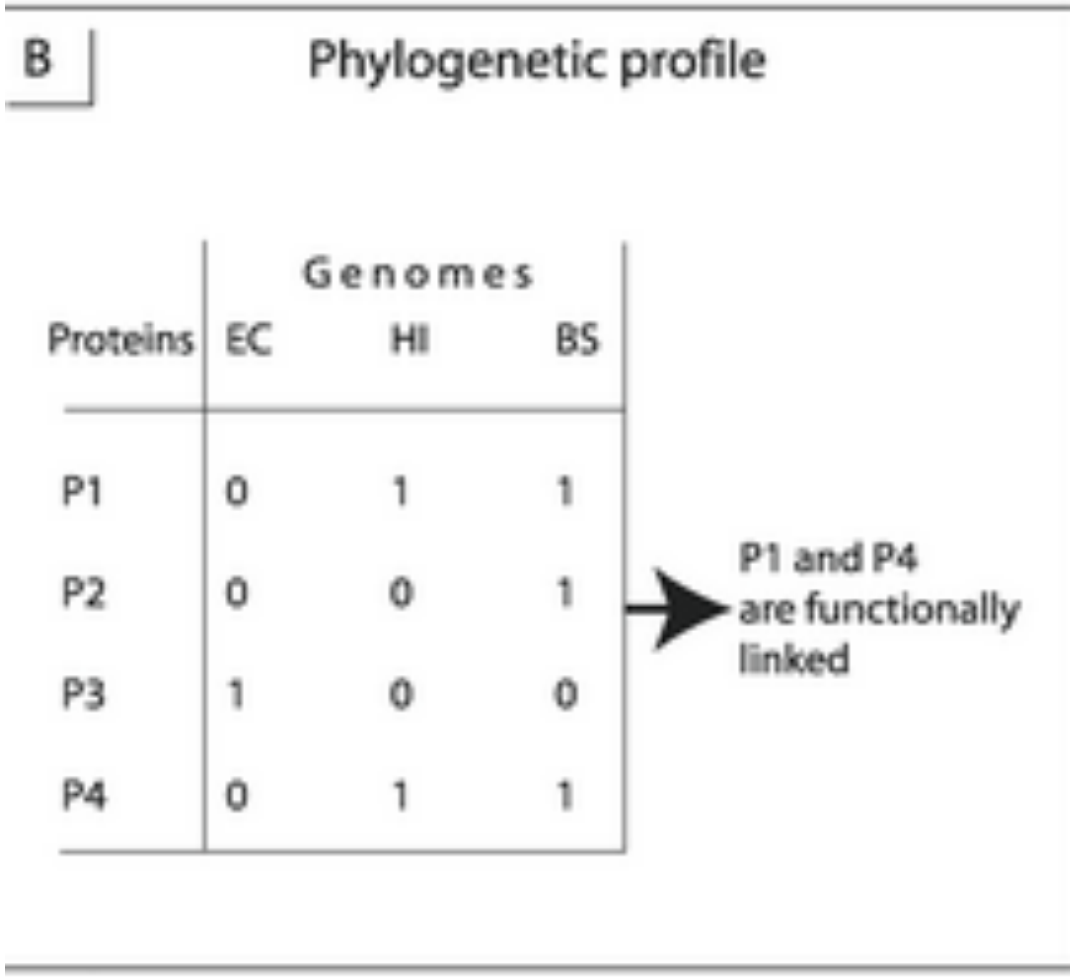
Gene cluster and gene neighborhood methods

Genes with closely related functions encoding potentially interacting proteins are often co-regulated in eukaryotes.

Gene clusters or operons encoding for co-regulated genes are usually conserved

Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e43

Computational Methods for Protein Interaction Prediction



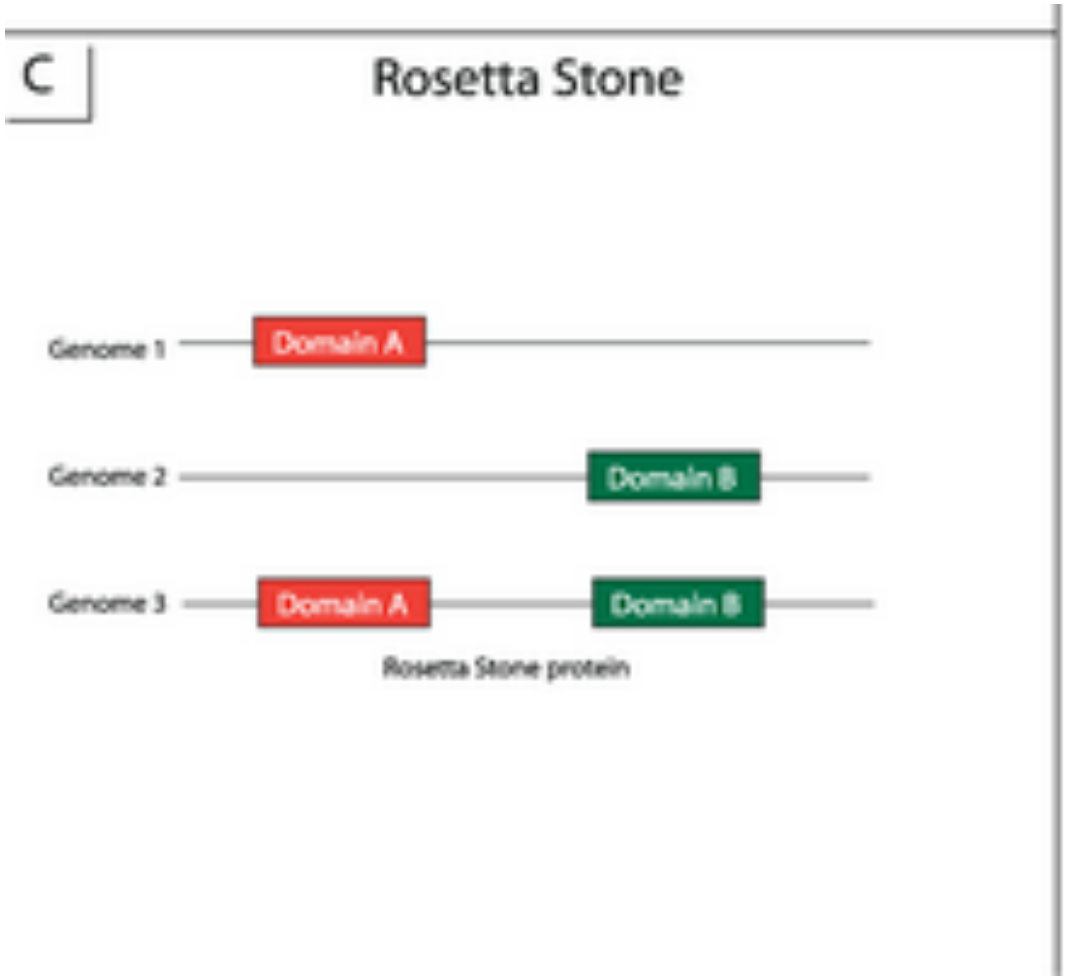
Phylogenetic profile method, showing the presence/absence of four proteins in three genomes.

Hypothesis: functionally linked and potentially interacting nonhomologous proteins co-evolve

A phylogenetic profile is constructed for each protein, as a vector of N (number of genomes). Proteins that form clusters are functionally related.

Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e43

Computational Methods for Protein Interaction Prediction

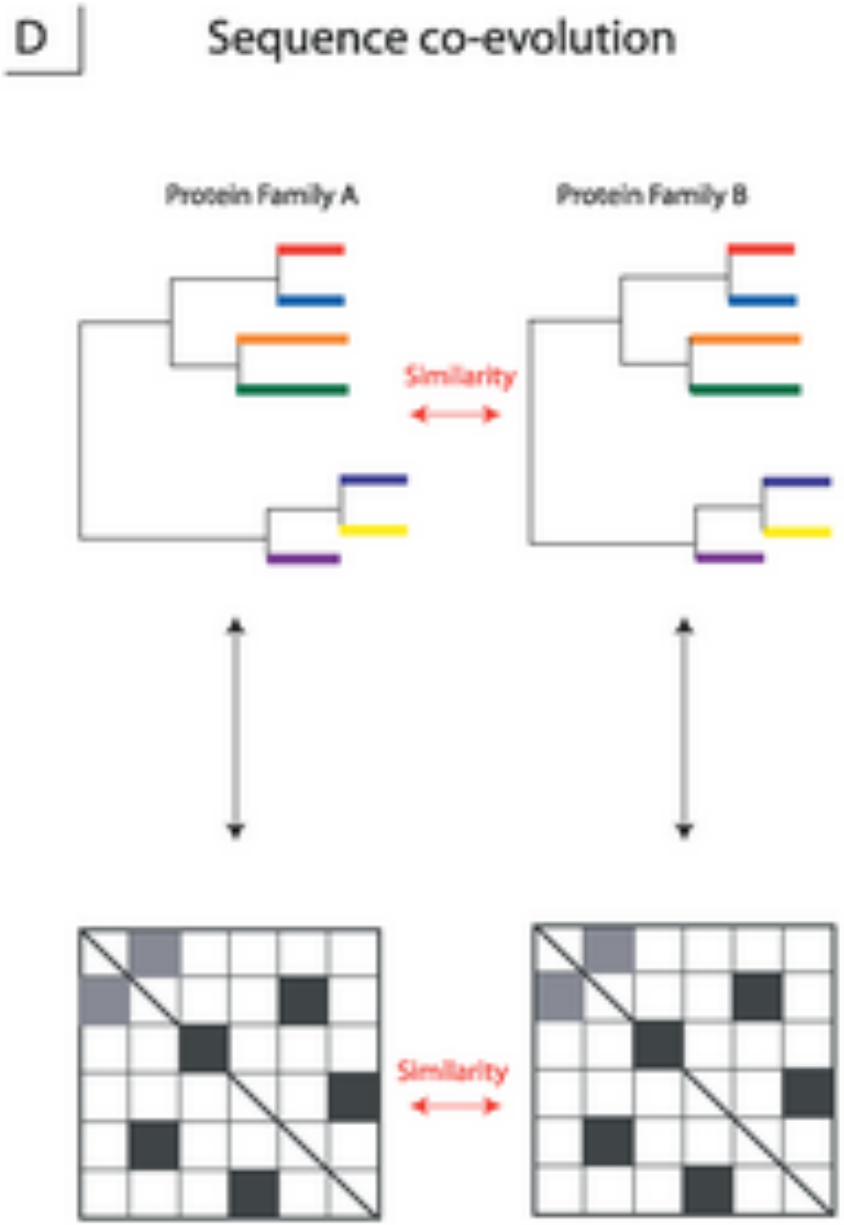


Rosetta Stone method

Based on the observation that some interacting proteins/domains have homologs in other genomes that are fused into one protein chain, a so-called Rosetta Stone protein

Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e43

Computational Methods for Protein Interaction Prediction



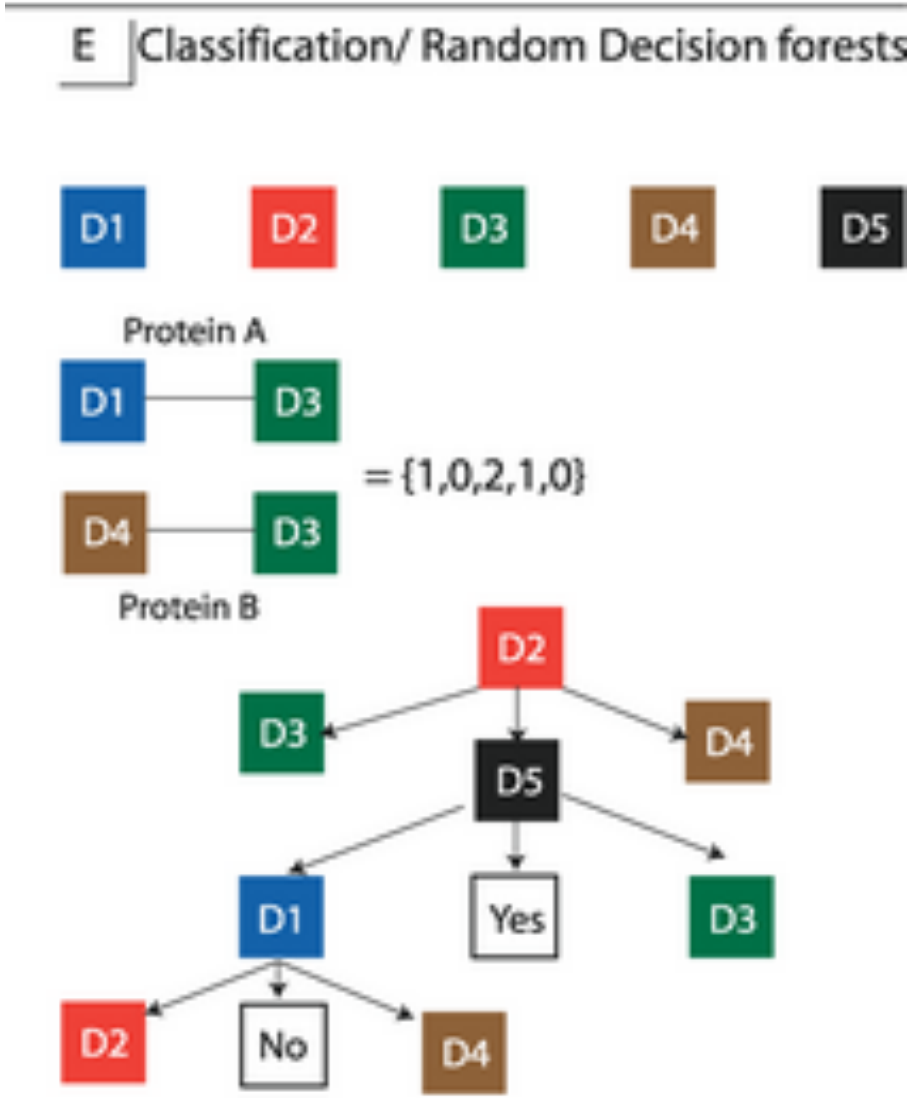
Sequence co-evolution method looking for the similarity between two phylogenetic trees/distance matrices

Proteins co-evolve so that changes in one protein leading to the loss of function or interaction should be compensated by the correlated changes in another protein.

Calculated based on similarity between phylogenetic trees.

Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e43

Computational Methods for Protein Interaction Prediction



Classification methods shown with the example of Random Decision Forest (RDF) method. Here, 5 different features/domains are used and each pair is encoded as a string of 0, 1, and 2.

The decision trees are constructed based on the training set of interacting protein pairs and decisions are made if proteins under the question interact or not.

Shoemaker BA, Panchenko AR (2007) PLOS Computational Biology 3(4): e43

Bioinformatics - tasks

- Collect PPI information
- Analyze PPIs
  - Molecular level
- Curate PPIs in a database
- Extract information about PPI from texts
- Analyze PPIs
  - Systems level
- Prediction
  - Identify PPIs (systems level)
  - Identify PPI interfaces (molecular level)

Computational Methods to Identify Protein-Protein Binding Sites

Use evolutionary information from sequence alignments, and/or residue properties, and/or the geometric information from structures.

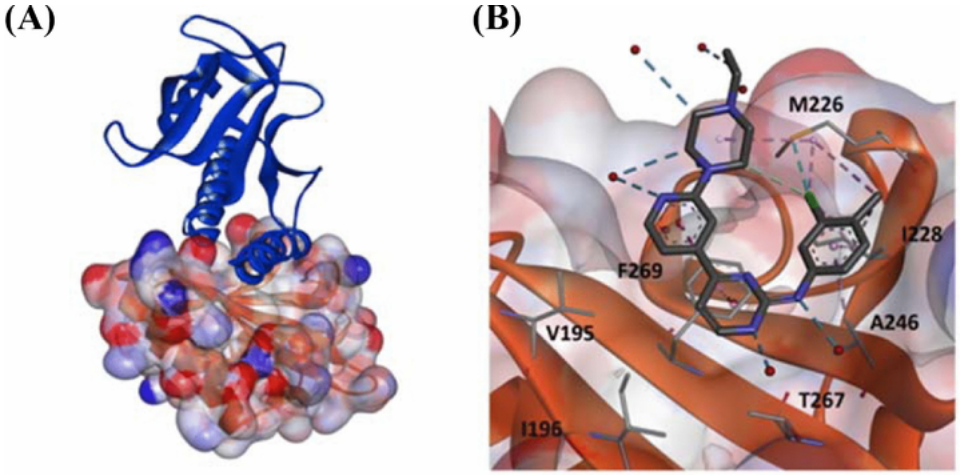
- Docking
- Threading and homology modeling
- Evolutionary tracing
- Correlated mutations
- Properties of patches
- Hydrophobicity
- Neural networks and support vector machines

Perspective – PPI sites as drug binding sites

PPIs can be modulated to **target** cancer, viruses, autoimmune disorders.

Direct binding occurs when small molecule modulators bind to the protein-protein interface, resulting in disruption or stabilization of the PPI.

Or allosteric binding occurs when a small molecule binds to a site away from the active site, resulting in a conformational change to the PPI binding site which then inhibits or enhances the PPI.



(A) Crystal structure of FtsZ (Blue) binding ZipA whole protein (Orange) (PDB: 1F46) (B) Compound 26 binding to the interface (PDB:1Y2F)

References

- Shoemaker BA, Panchenko AR (2007) Deciphering Protein–Protein Interactions. Part I. Experimental Techniques and Databases. *PLoS Comput Biol* **3**(3): e42. https://doi.org/10.1371/journal.pcbi.0030042
- Shoemaker BA, Panchenko AR (2007) Deciphering Protein–Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. *PLoS Comput Biol* **3**(4): e43. https://doi.org/10.1371/journal.pcbi.0030043
- von Mering C et al. (2002) Comparative assessment of large-scale data sets of protein–protein interactions *Nature* **417**:399–403
- Sudha, G., Nussinov, R., & Srinivasan, N. (2014). An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles. *Progress in Biophysics and Molecular Biology*, ***116***(2-3), 141–150. http://doi.org/10.1016/j.pbiomolbio.2014.07.004
- Cossar PJ, Lewis PJ, McCluskey A (2018) Protein-protein interactions as antibiotic targets: A medicinal chemistry perspective *Medicial Research Reviews* https://doi.org/10.1002/med.21519