Molekulare Grundlagen genetisch bedingter Erkrankungen (2023)

Pouria Dasmeh

Center for Human Genetics, Marburg University Hospital

Course agenda and organizational notes

Part 1: Protein Sequence, Structure, Function
Part 2: Protein Interactions
Part 3: Protein Networks

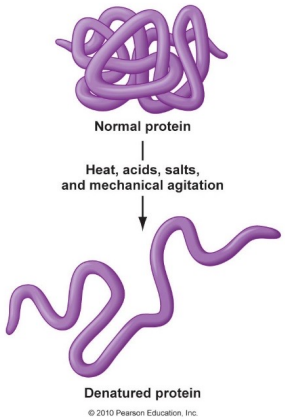
Point 1: This is an interactive lecture based on “project-based learning”:

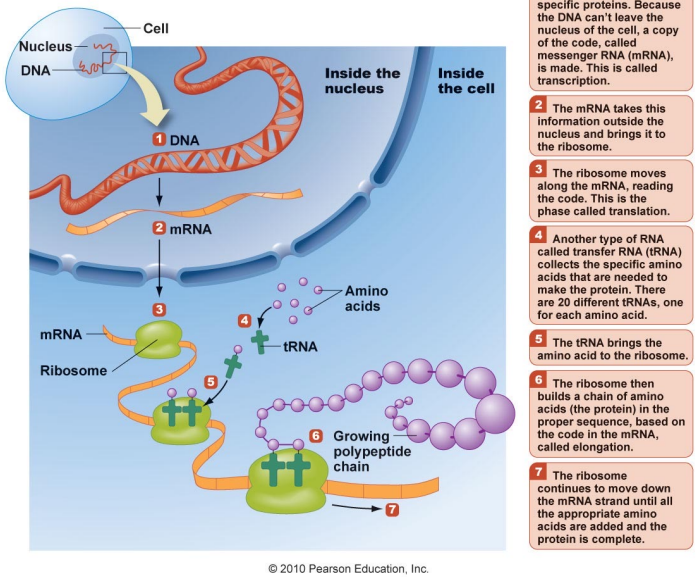
- You have to talk!
- We have class Activities
- We have three mini projects (One collaborative, Two competitive)

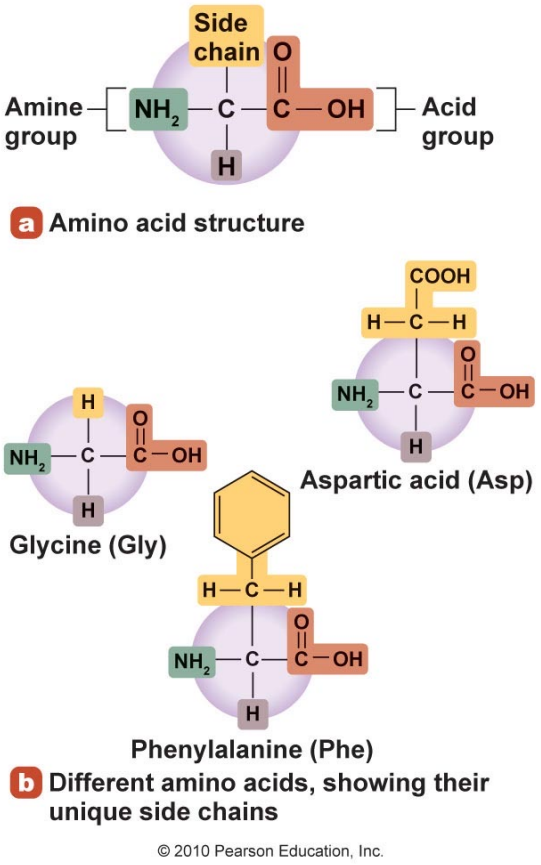
Point 2: All analyses can be done using R/python libraries. We only show a few analyses here.

Part 1:
Protein Sequence, Structure, Function

Proteins and Amino Acids

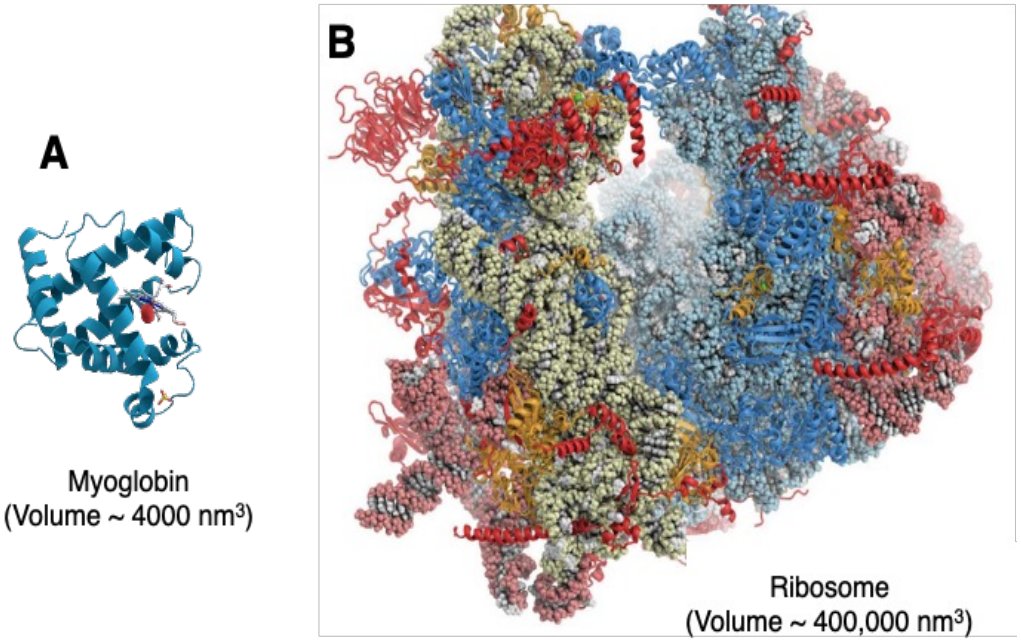






Proteins:

- Made up of chains of amino acids
- Are involved in most of the body's functions and life processes
- The sequence of amino acids is determined by DNA



Structure of Proteins

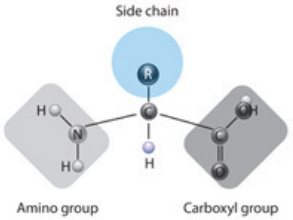
- Made up of chains of amino acids; classified by number of amino acids in a chain
  - Peptides: fewer than 50 amino acids
    - Dipeptides: 2 amino acids
    - Tripeptides: 3 amino acids
    - Polypeptides: more than 10 amino acids
  - Proteins: more than 50 amino acids
    - Typically 100 to 10,000 amino acids linked together
- Chains are synthesizes based on specific bodily DNA.
- Amino acids are composed of carbon, hydrogen, oxygen, and nitrogen.

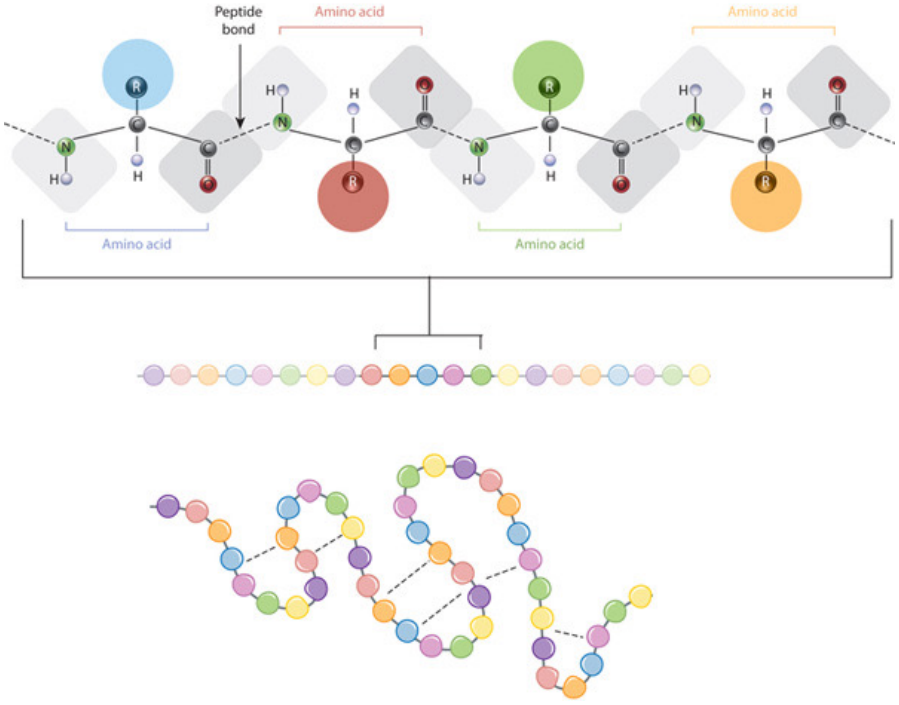
Class Activity:

Find (one of the) shortest and (one of the longest) proteins

in the UniProt database and profile them.

An amino acid





An amino acid

TWENTY-ONE PROTEINOGENIC α-AMINO ACIDS

Side chain charge at physiological pH 7.4

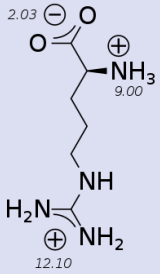
pK_a values shown italicized

⊕ Positive
⊖ Negative

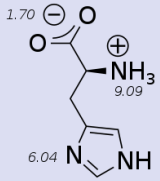
A. Amino Acids with Electrically Charged Side Chains

Positive

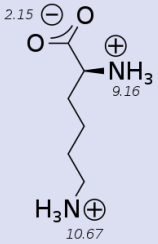
Arginine Arg R



Histidine His H

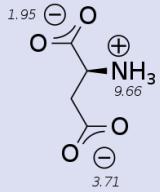


Lysine Lys K

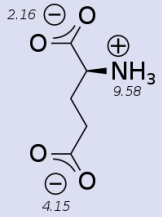


Negative

Aspartic Acid Asp D

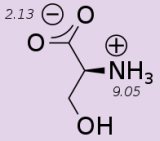


Glutamic Acid Glu E

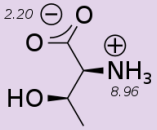


B. Amino Acids with Polar Uncharged Side Chains

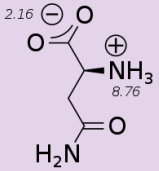
Serine Ser S



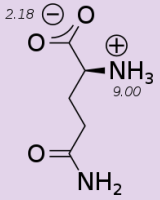
Threonine Thr T



Asparagine Asn N

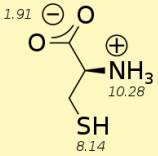


Glutamine Gln Q



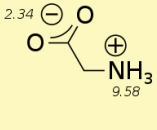
C. Special Cases

Cysteine Cys C

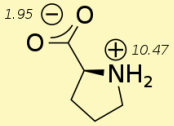


Selenocysteine Sec U

Glycine Gly G

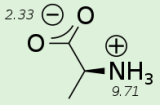


Proline Pro P

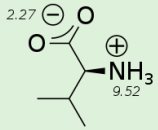


D. Amino Acids with Hydrophobic Side Chains

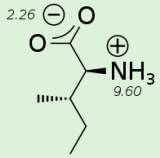
Alanine Ala A



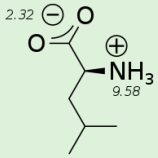
Valine Val V



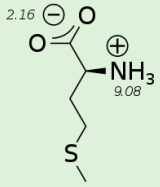
Isoleucine Ile I



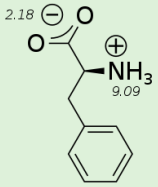
Leucine Leu L



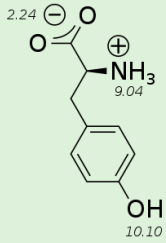
Methionine Met M



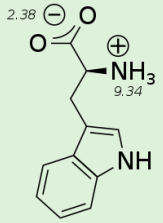
Phenylalanine Phe F



Tyrosine Tyr Y



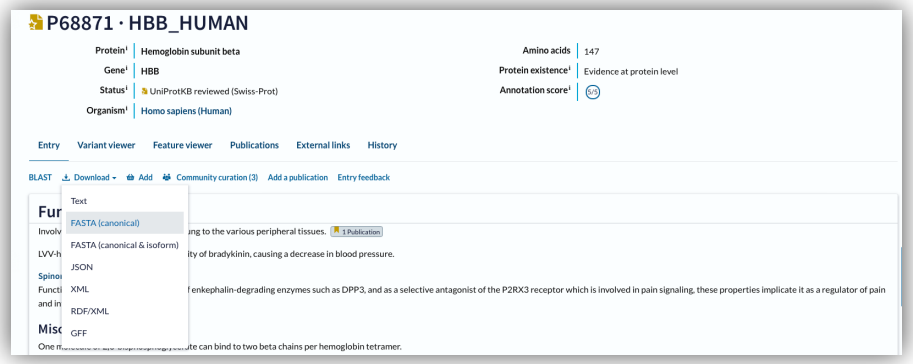
Tryptophan Trp W



Composition of peptides, and dipeptides in two proteins:

Hemoglobin beta (HBB)
RNA binding protein FUS

Search HBB gene on Uniprot



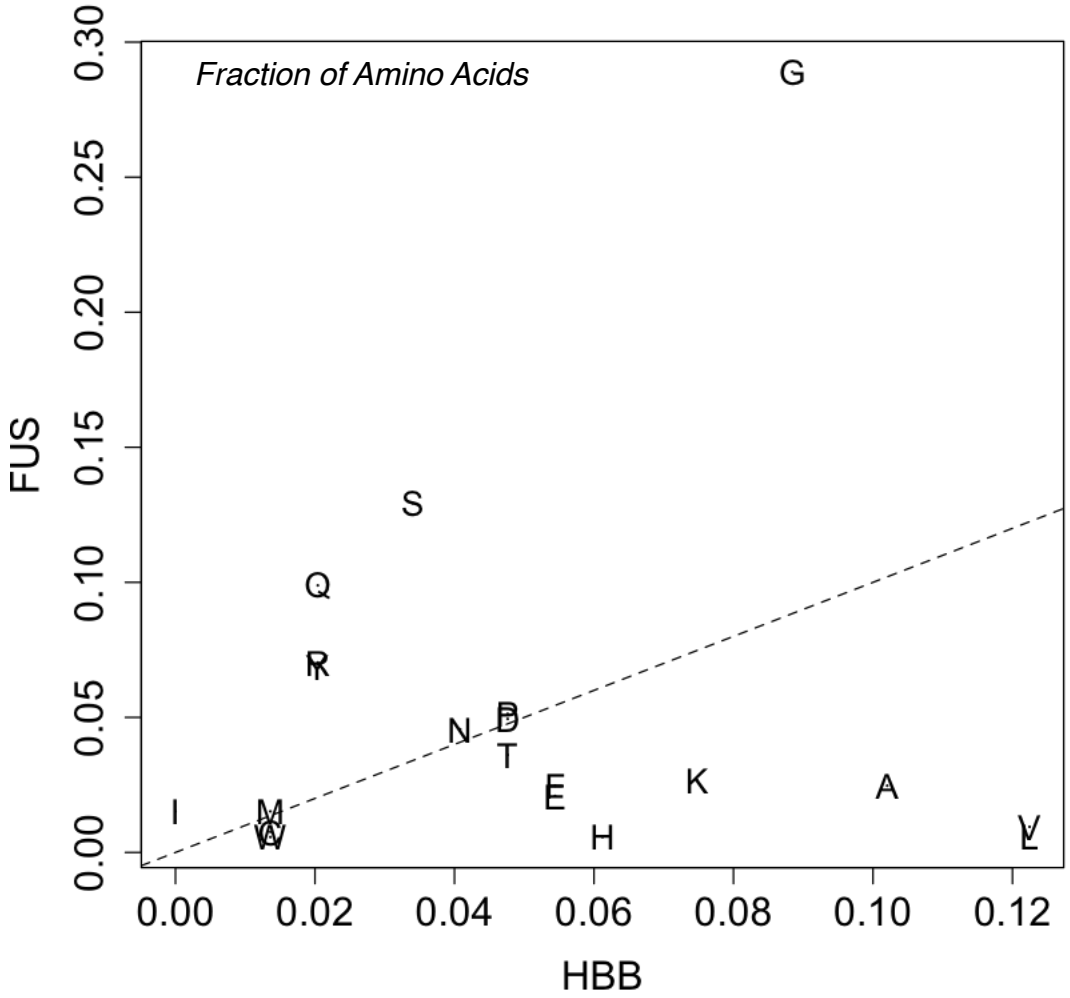
Download the Fasta file

```
>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens OX=9606 GN=HBB PE=1 SV=2
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

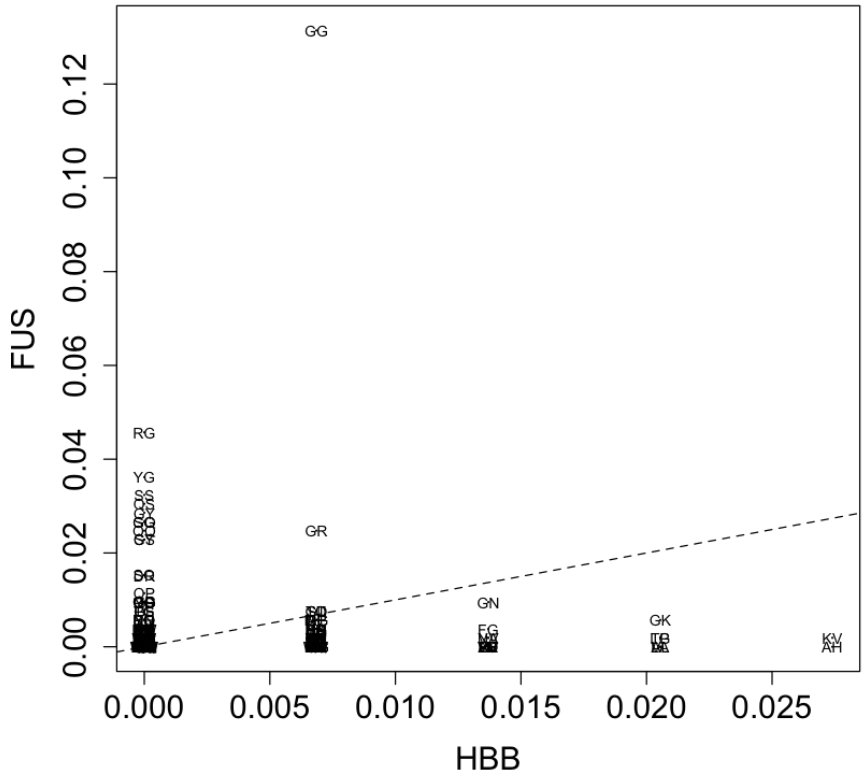
Run the following code to extract and compare the amino acid compositions

```r
# Install and load the protr library
install.packages("protr")
library("protr")

# Read the protein sequences in FASTA format
HBB <- protr::readFASTA("FM_proteins/P68871.fasta")
FUS <- protr::readFASTA("FM_proteins/P35637.fasta")

# Extract the amino acid composition
HBB_frac <- extractAAC(HBB$`sp|P68871|HBB_HUMAN`)
FUS_frac <- extractAAC(FUS$`sp|P35637|FUS_HUMAN`)

# Plot the composition of two proteins against each other
plot(HBB_frac, FUS_frac, cex=0.1, xlab="HBB", ylab="FUS", cex.axis=1.5, cex.lab=1.5)
text(HBB_frac, FUS_frac, names(HBB_frac), cex=1.3)
abline(a=0, b=1, lty=2)
```

What do you see?

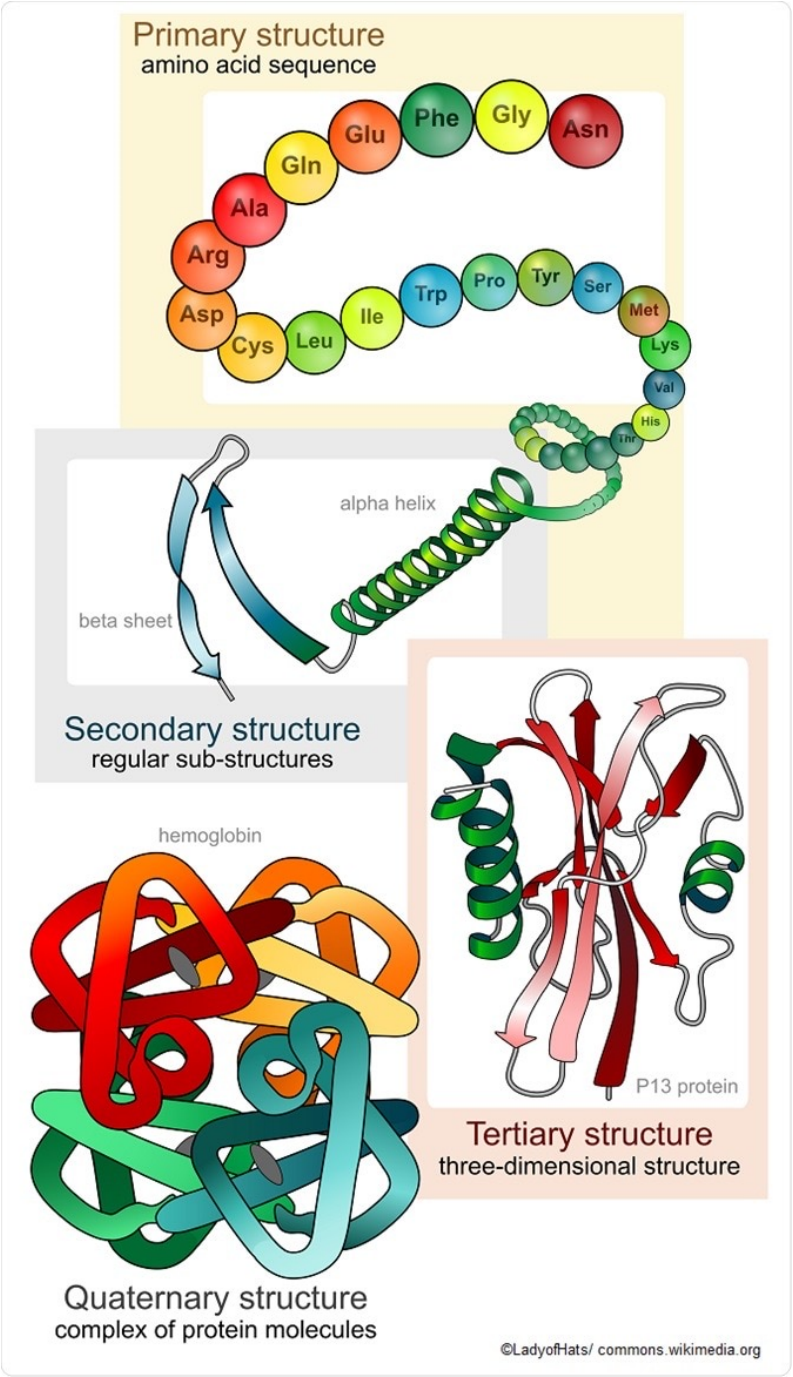


```r
# Extract the dipeptide's composition
HBB_frac_dipeptides <- extractDC(HBB$`sp|P68871|HBB_HUMAN`)
FUS_frac_dipeptides <- extractDC(FUS$`sp|P35637|FUS_HUMAN`)

# Plot the fraction of different dipeptides
plot(HBB_frac_dipeptides, FUS_frac_dipeptides, cex=0.1, xlab="HBB", ylab="FUS", cex.axis=1.5, cex.lab=1.5)
text(HBB_frac_dipeptides, FUS_frac_dipeptides, names(HBB_frac_dipeptides), cex=0.7)
abline(a=0, b=1, lty=2)
```



Q: How to determine that certain dipeptides are more frequent than by random chance?

Structure of the Protein



Different proteins have different percentages of secondary structure elements

TABLE 4–4 **Approximate Proportion of α Helix and β Conformation in Some Single-Chain Proteins**

Protein (total residues)	Residuevs (%)*	
	α Helix	β Conformation
Chymotrypsin (247)	14	45
Ribonuclease (124)	26	35
Carboxypeptidase (307)	38	17
Cytochrome *c* (104)	39	0
Lysozyme (129)	40	12
Myoglobin (153)	78	0

Source: Data from Cantor, C.R. & Schimmel, P.R. (1980) *Biophysical Chemistry,* Part I: *The Conformation of Biological Macromolecules*, p. 100, W. H. Freeman and Company, New York.

*Portions of the polypeptide chains not accounted for by α helix or β conformation consist of bends and irregularly coiled or extended stretches. Segments of α helix and β conformation sometimes deviate slightly from their normal dimensions and geometry.

Class Activity: Check the secondary structure elements in Hemoglobin (beta subunit) and the RNA binding protein FUS.

PDB database

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

RCSB PDB PROTEIN DATA BANK

140824 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands Go

Advanced Search | Browse by Annotations

PDB-101 wwPDB EMDataBank Nucleic Acid Database Worldwide Protein Data Bank Foundation

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

A Structural View of Biology

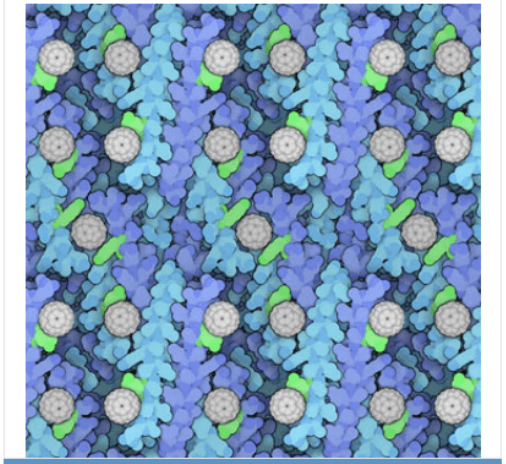
This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

RCSB PDB Services and Impact

June **Molecule of the Month**



Proteins and Nanoparticles

Latest Entries *As of Tuesday May 29 2018*

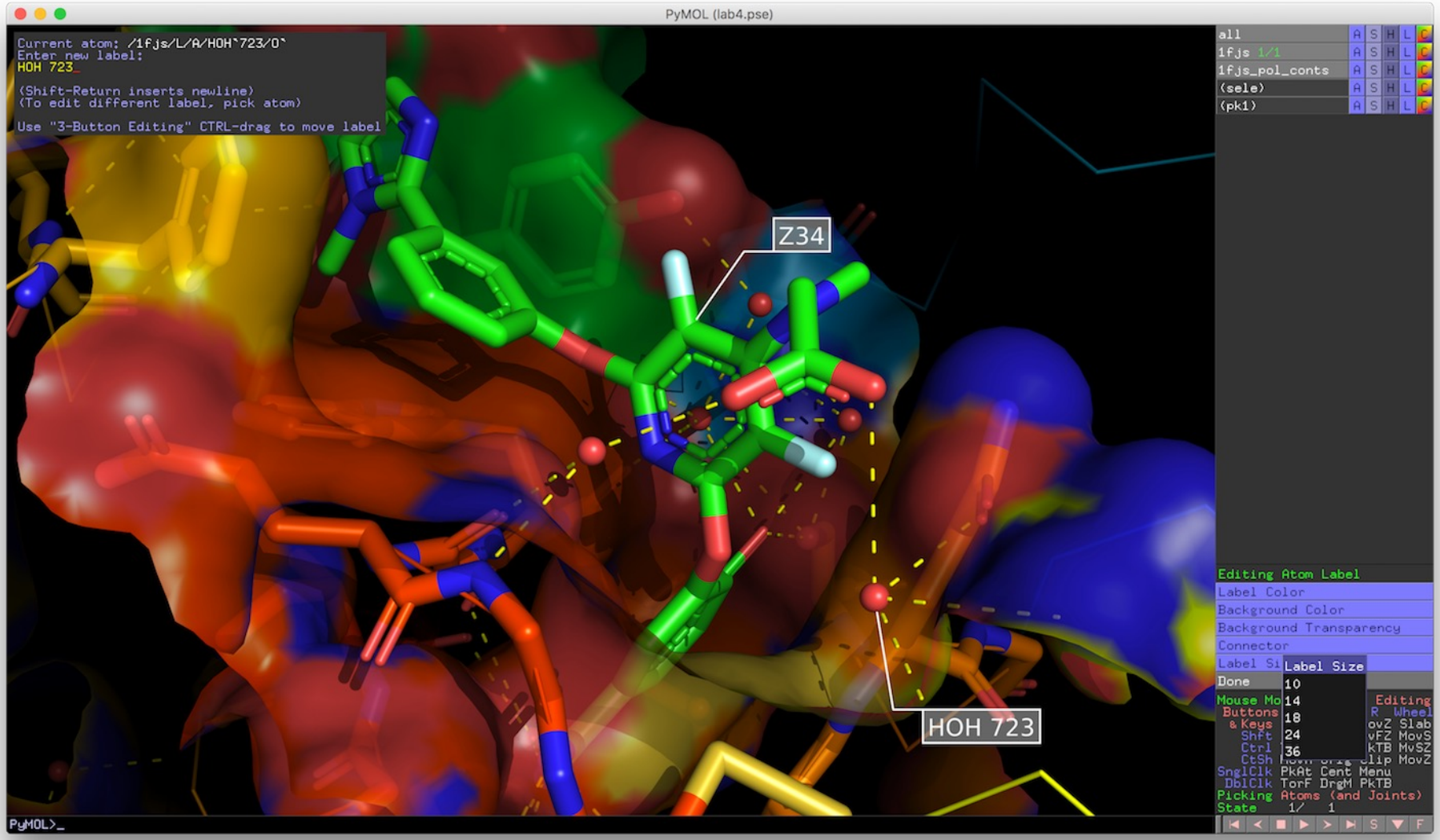
Features & Highlights

New Architecture and Services Enable Faster Access to More Information

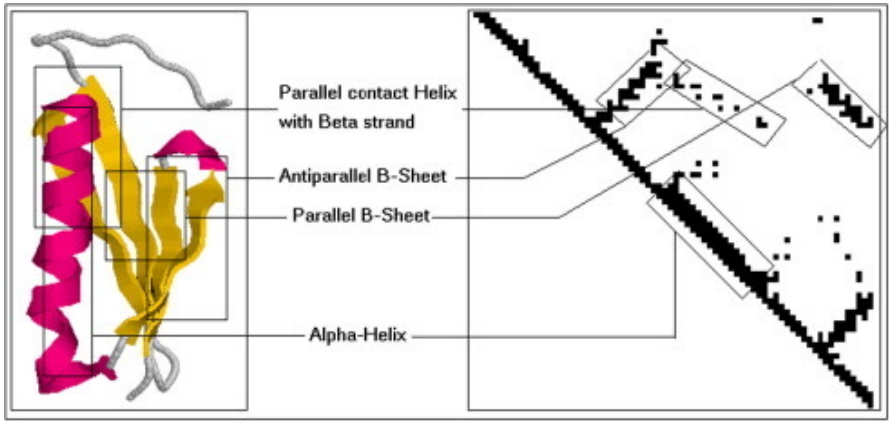
News Publications

Vote Now for the Viewer's Choice Award

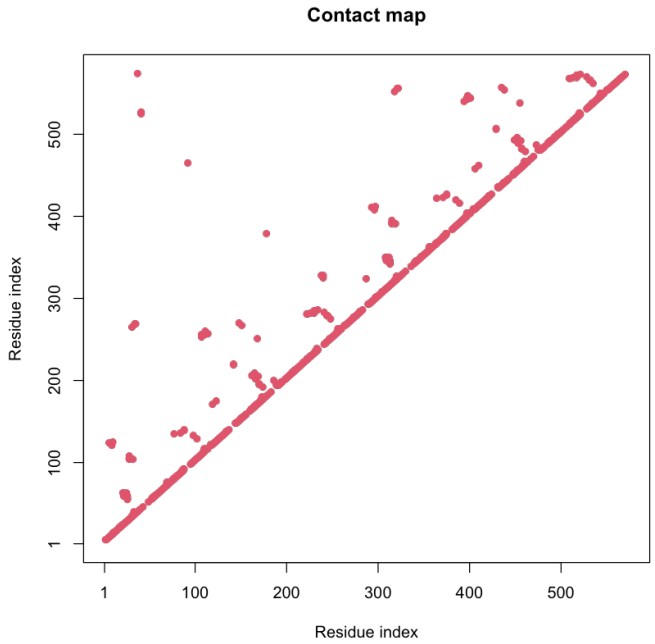
Pymol: Protein and Molecular Visualization



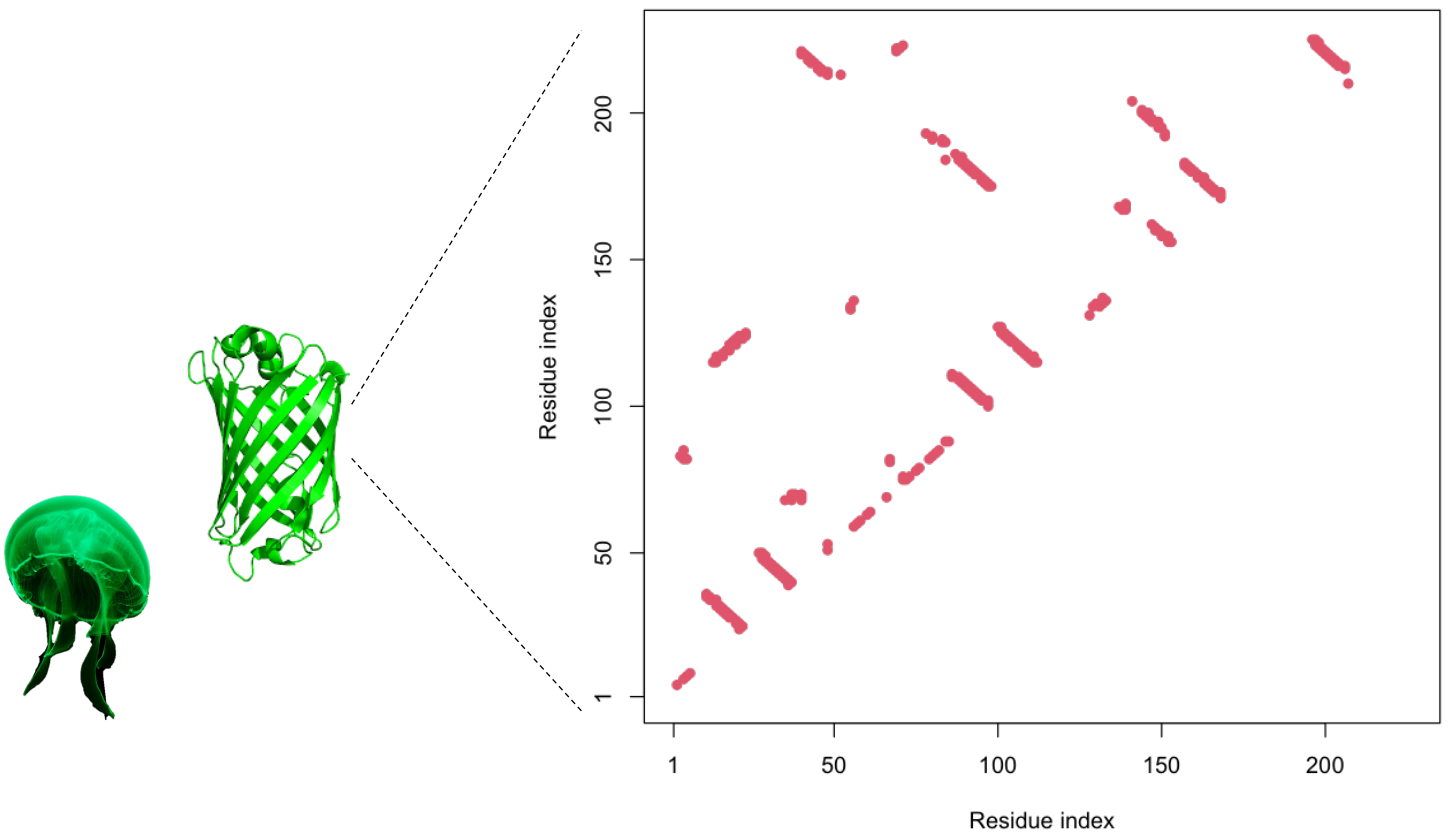
Protein Contact Map



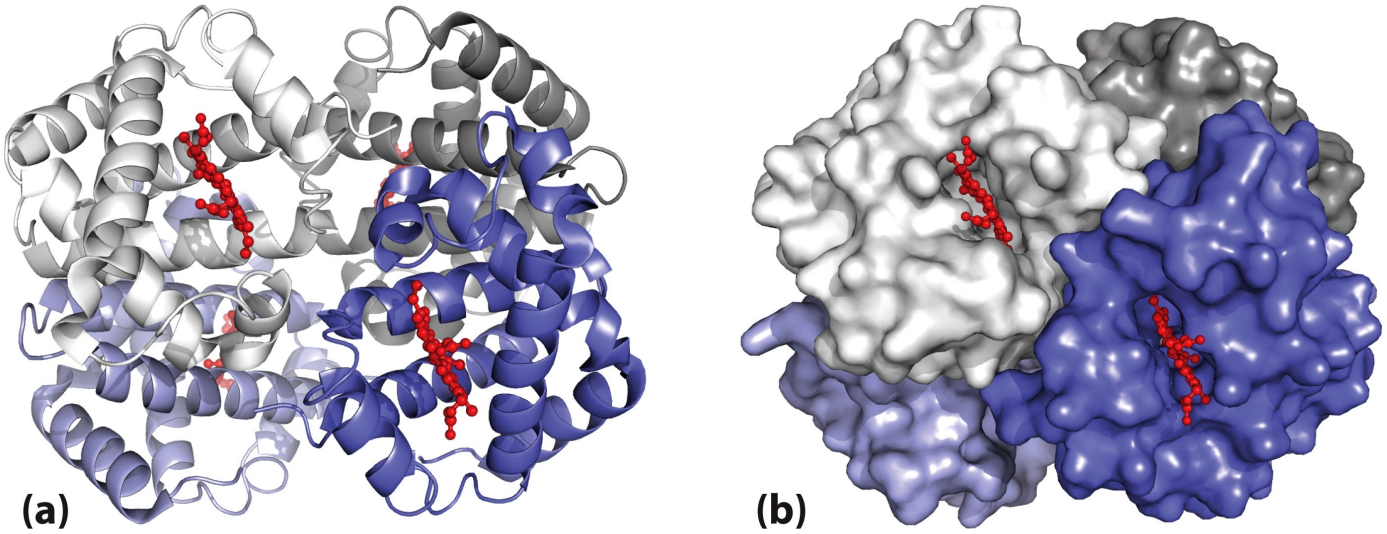

```
# Instal and load the library
install.packages("bio3d")
library("bio3d")

#Get the pdb file and load it
get.pdb("1A01")
pdb <- read.pdb("1A01.pdb")

## Atom Selection indices
inds <- atom.select(pdb, "calpha")

## Reference contact map
ref.cont <- cmap( pdb$xyz[inds$xyz], dcut=6, scut=3 )
plot.cmap(ref.cont)
```



Contact Map for Green Fluorescence Protein

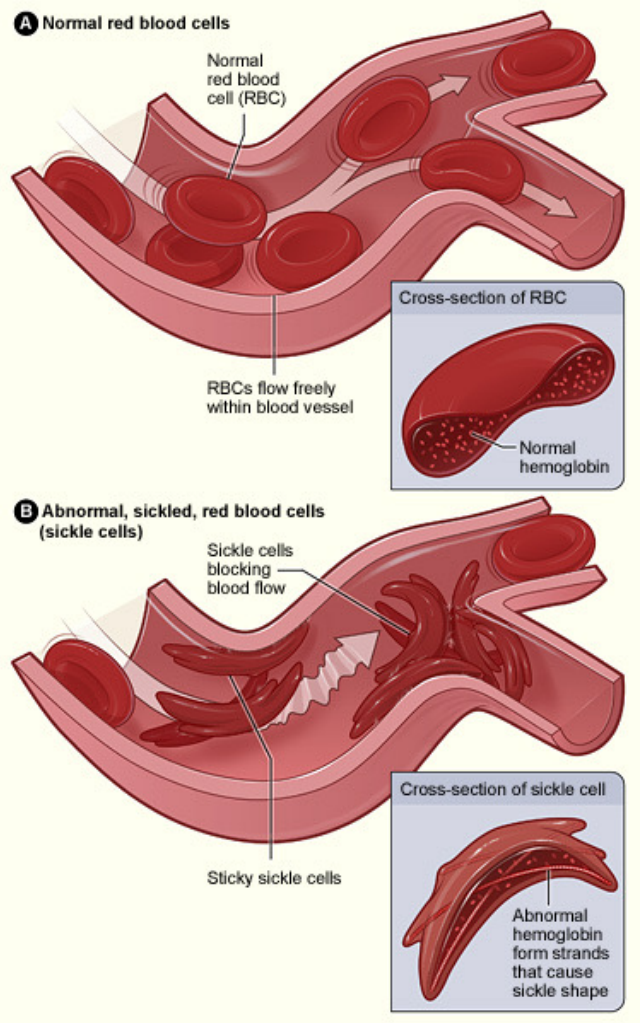


Quaternary Structure of Hemoglobin



- Hemoglobin (M_r 64,500) contains four polypeptide chains with one heme group each.
- The protein portion, globin, consists of two α chains (141 residues each) and two ß chains (146 residues each).
- The subunits of hemoglobin are arranged in symmetric pairs, each pair having one α and one ß subunit. Hemoglobin can therefore be described either as a tetramer or as a dimer of αß protomers.

Sickle Cell Anemia



Sickle Cell Anemia

Sickle cell anemia is caused by a specific mutation. The mutation causes a change in a single amino acid in the beta-globin chain of hemoglobin, where the amino acid **glutamic acid** is replaced by **valine** (E6V)

This single substitution changes the physical properties of hemoglobin, causing it to form stiff, sticky, and sickle-shaped red blood cells that can't easily flow through small blood vessels.

Class Activity:

Investigate the causes of E6V mutation leading to sickle cell anemia

1. Check Hemoglobin B on Uniprot
2. Get the PDB structure and open in Pymol (or *fetch* command)
3. Investigate it
4. How can E6V mutation cause the disease?

Denaturing a Protein

- Desaturating agent/factors:
  - Heat
  - Acids
  - Bases
  - Salts
  - Mechanical agitation
  - Mutations

  (interference with folding)

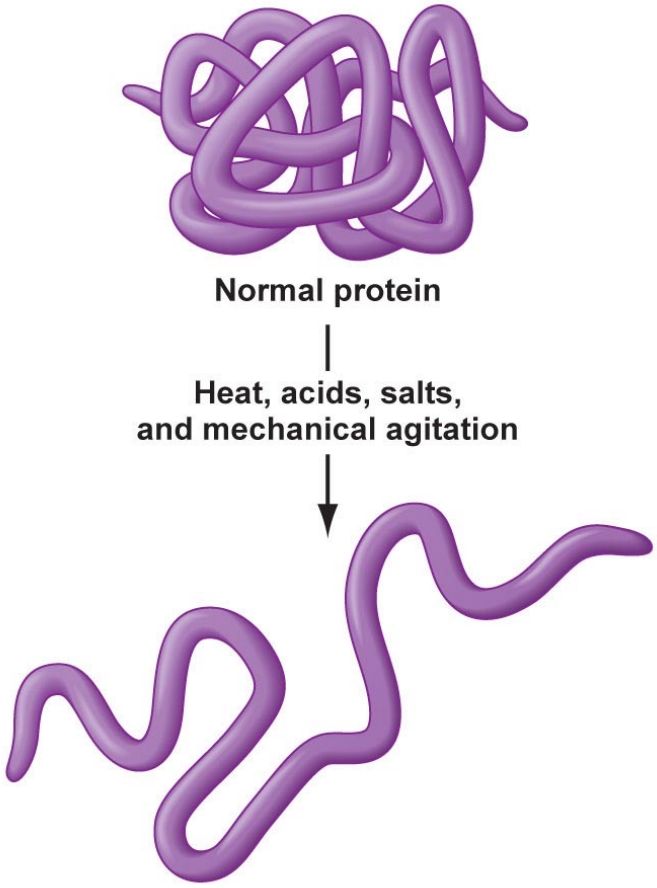
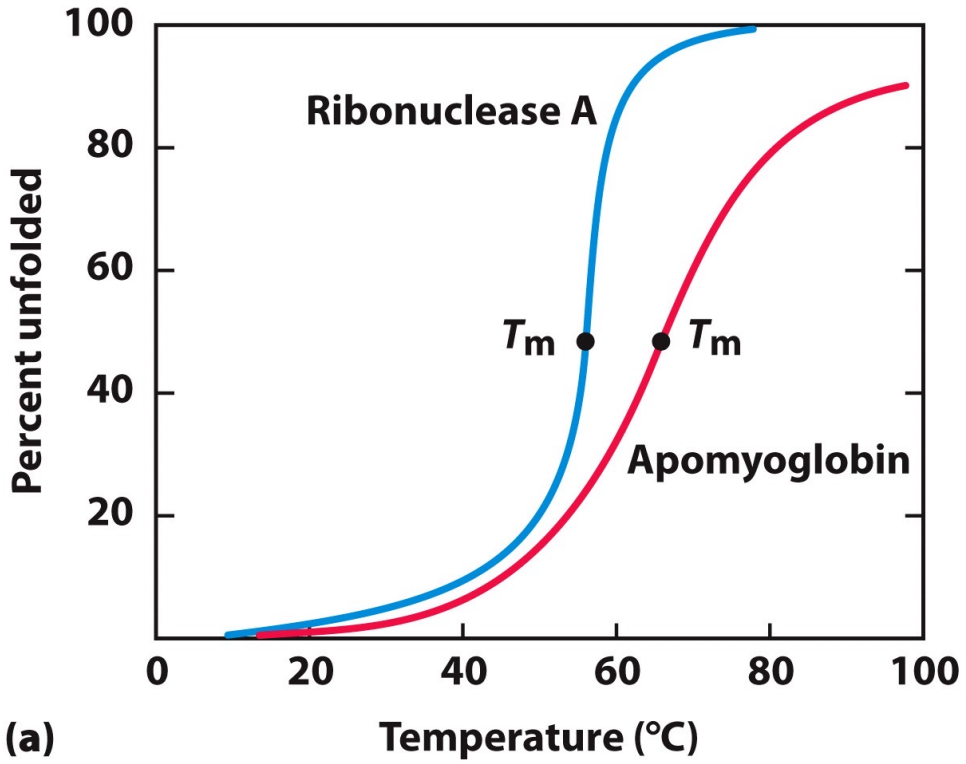
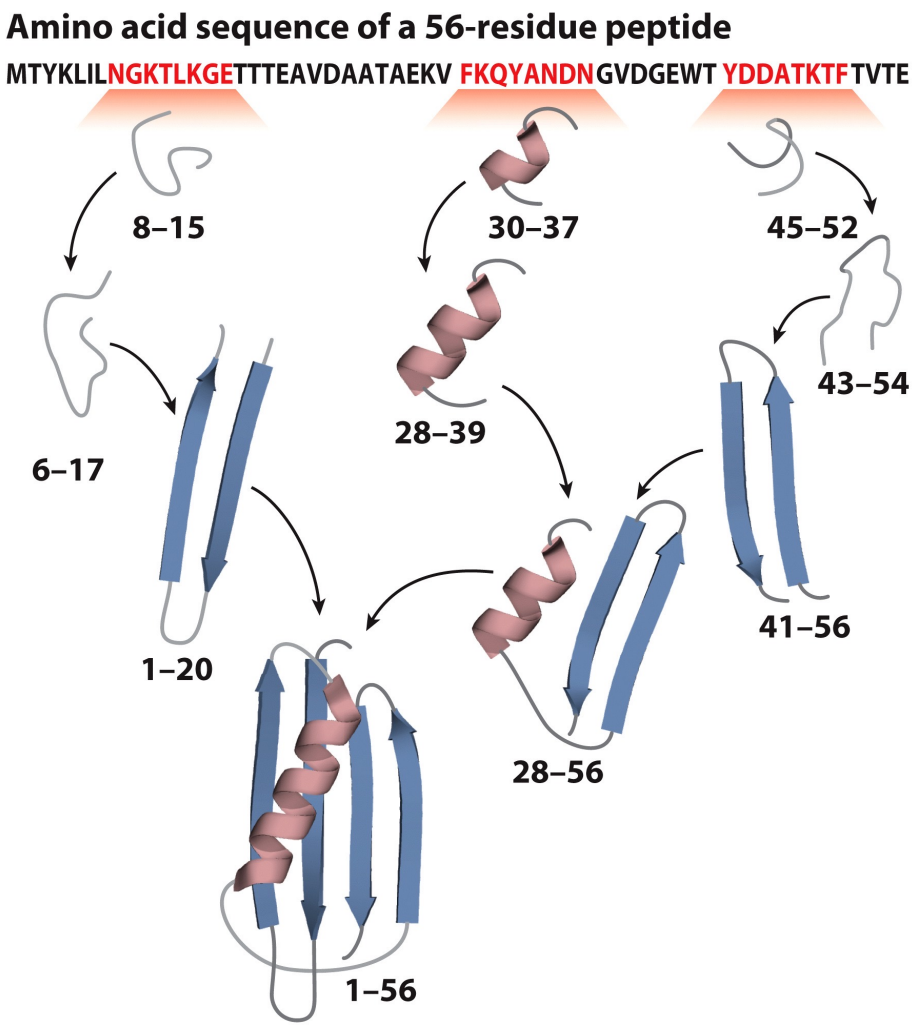


Figure 6.5

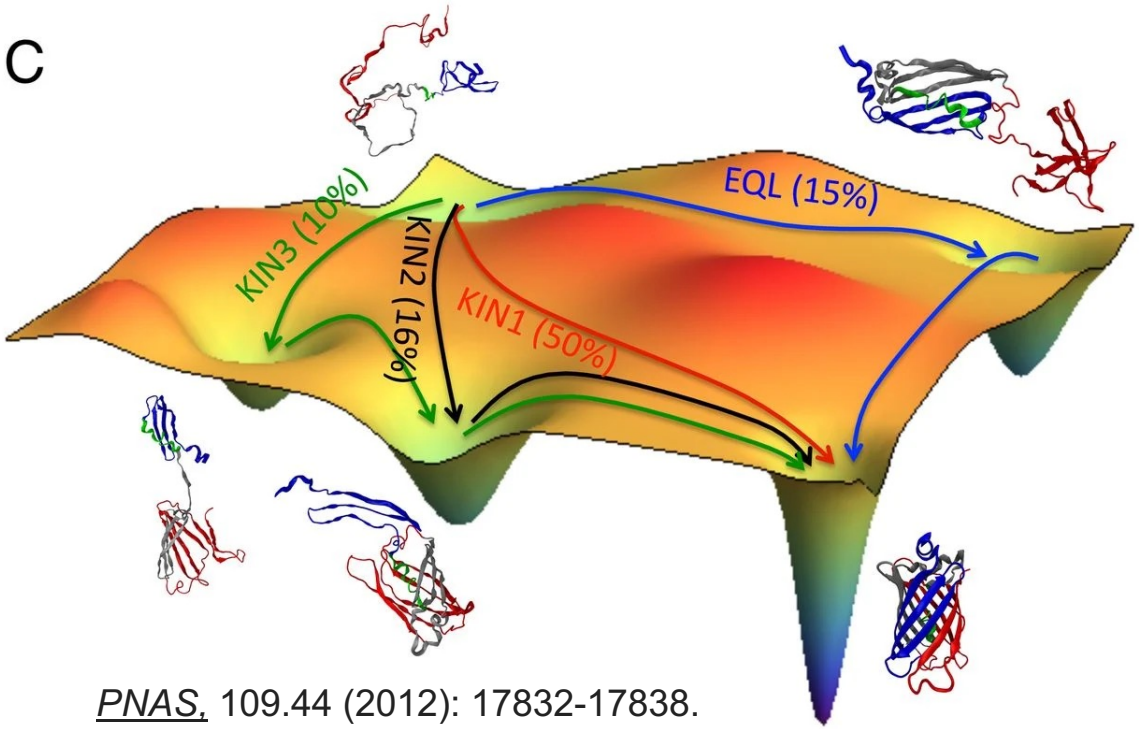
Protein Denaturation



Overview of Protein Folding



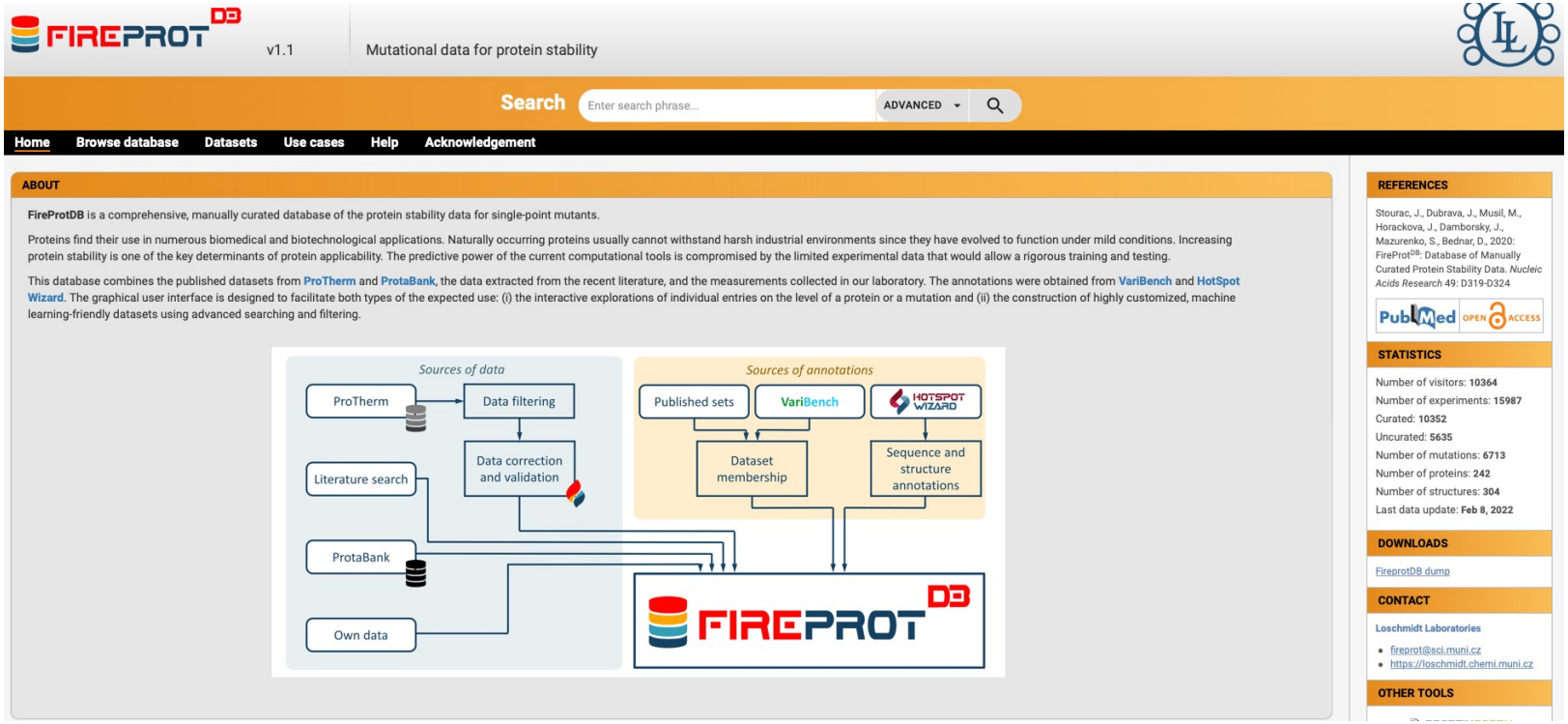
Protein Folding Intermediates

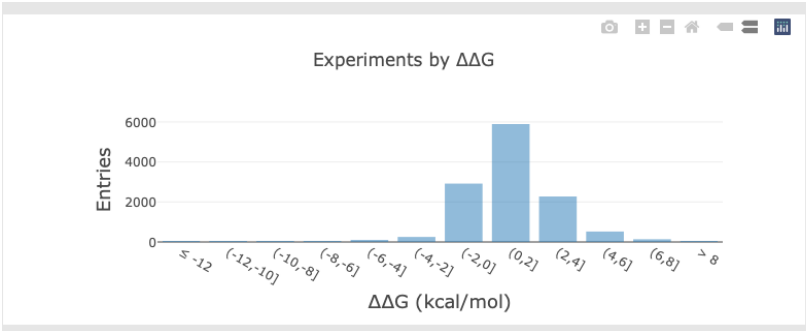


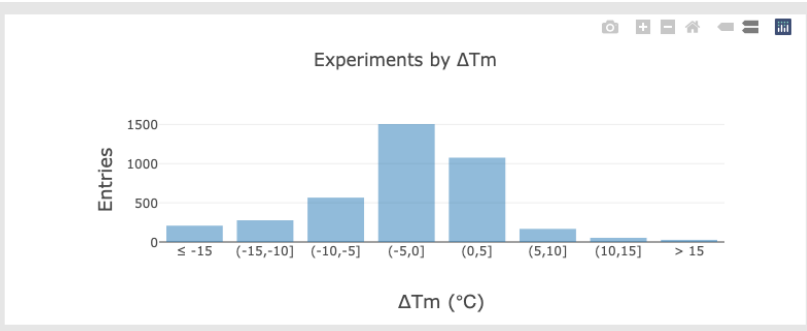
PNAS, 109.44 (2012): 17832-17838.

Mutational Effects on Protein Stability

(FireProt database)







What does an average mutation do to a protein ?!

Class Mini Project 1 (collaborative)

Compare the sequences of HBB in Human with other mammalian orthologs (using UniProt database)

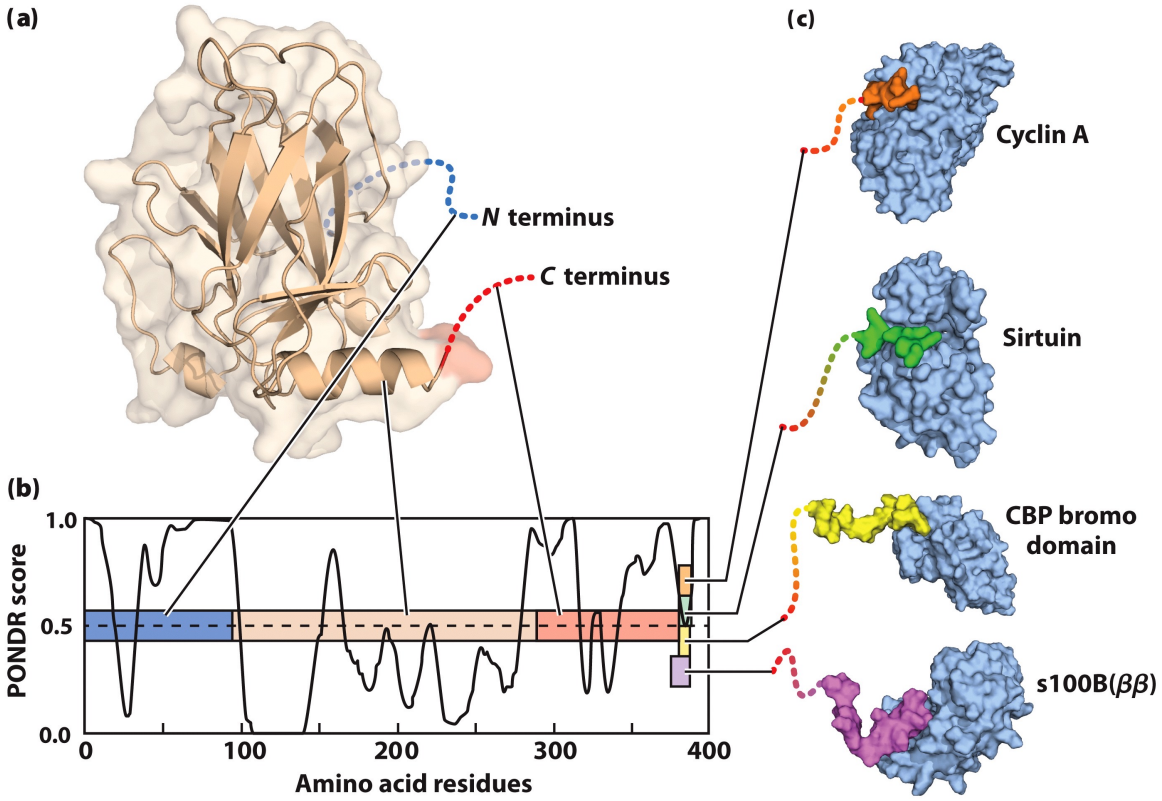
UniProtKB 7,462 results or search "HBB" as a Gene Name, Protein Name, or Strain

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share 6 rows selected out of 25

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
P68871	HBB_HUMAN	Hemoglobin subunit beta[...]	HBB	Homo sapiens (Human)	147 AA
P02070	HBB_BOVIN	Hemoglobin subunit beta[...]	HBB	Bos taurus (Bovine)	145 AA
P68873	HBB_PANTR	Hemoglobin subunit beta[...]	HBB	Pan troglodytes (Chimpanzee)	147 AA
P02075	HBB_SHEEP	Hemoglobin subunit beta[...]	HBB	Ovis aries (Sheep)	145 AA
P02112	HBB_CHICK	Hemoglobin subunit beta[...]	HBB	Gallus gallus (Chicken)	147 AA
P68872	HBB_PANPA	Hemoglobin subunit beta[...]	HBB	Pan paniscus (Pygmy chimpanzee) (Bonobo)	147 AA

1) Work in groups
2) Discuss which mutations to choose (and divide)
3) Discuss which species to include
4) Report the conservation of “disease mutations”

Intrinsically Disordered Proteins



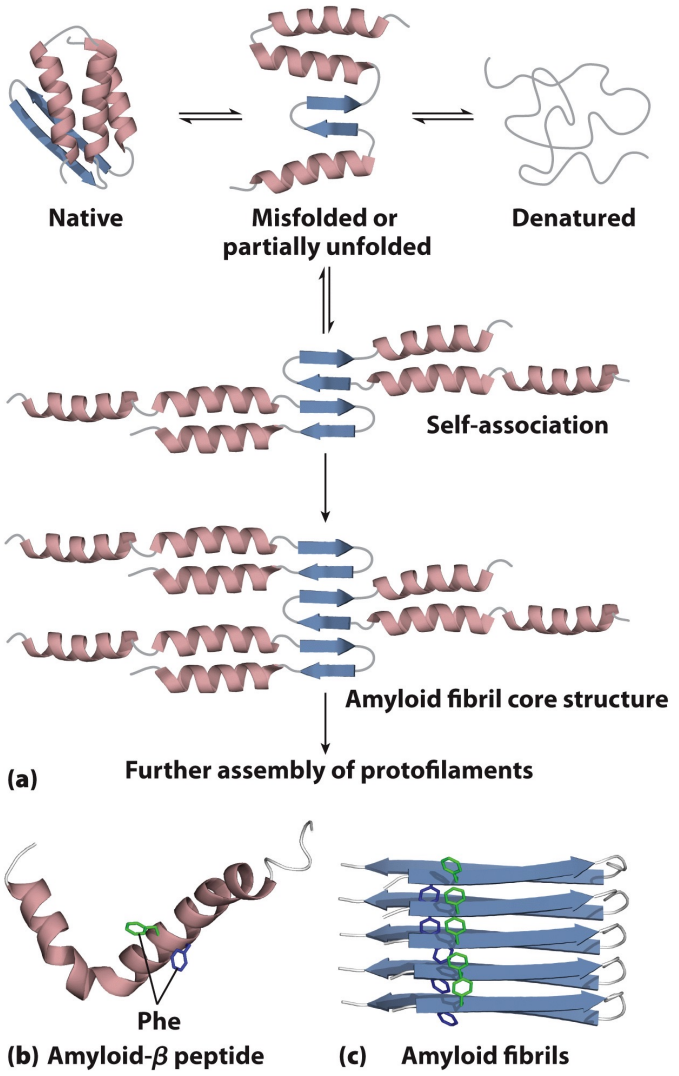
TP53 plays a crucial role in the control of cell division. It features both structured and unstructured segments. An unstructured region at the carboxyl terminus interacts with at least four different binding partners and assumes a different structure in each of the complexes.

Protein Folding Disorders: Amyloidosis

A soluble protein that is normally produced by a cell adopts a misfolded state and converts into an insoluble extracellular amyloid fiber.

Fiber formation is promoted by the aggregation of regions of proteins that have a ß conformation. In Alzheimer disease, proteolytic cleavage of a neuronal membrane protein (amyloid-ß precursor protein, APP) produces an α-helical membrane-spanning peptide (amyloid-ß peptide) that converts to the ß conformation and aggregates into amyloid fibrils.

The extracellular deposition of amyloid fibrils is associated with plaque formation and ultimately death of the nearby neurons.

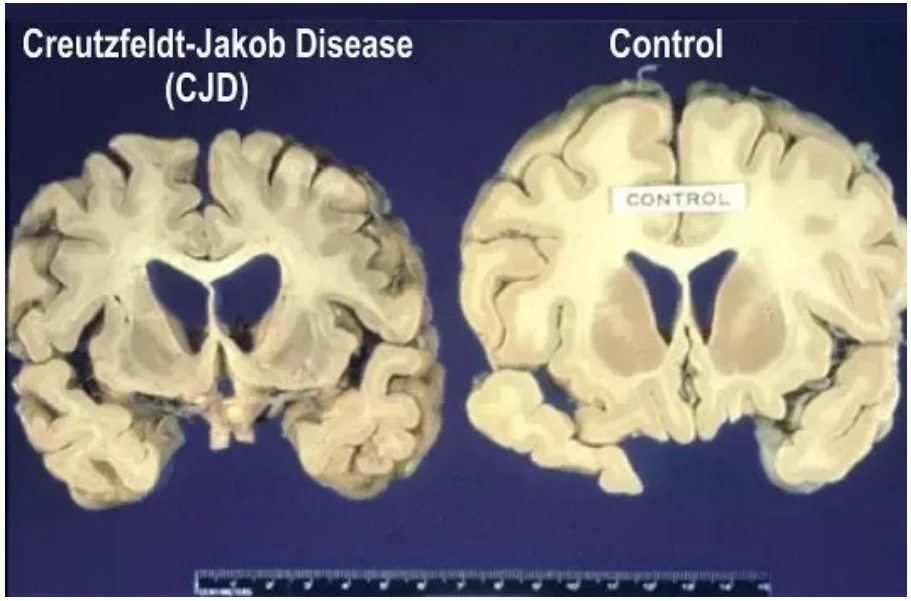


Protein Folding Disorders: Prion Diseases

In the neurodegenerative diseases known as spongiform encephalopathies, a misfolded form of a normal neuronal protein PrP is responsible for disease.

Spongiform encephalopathies occur in many species of animals. In humans, the disorders are known as kuru and Creutzfeld-Jacob disease. In cows, the disorder is known as mad cow disease. In sheep it is called scrapie, and in deer and elk, it is called chronic wasting disease. The diseased brain becomes riddled with holes.

Progressive deterioration of the brain leads to a spectrum of neurological symptoms, and is always fatal.

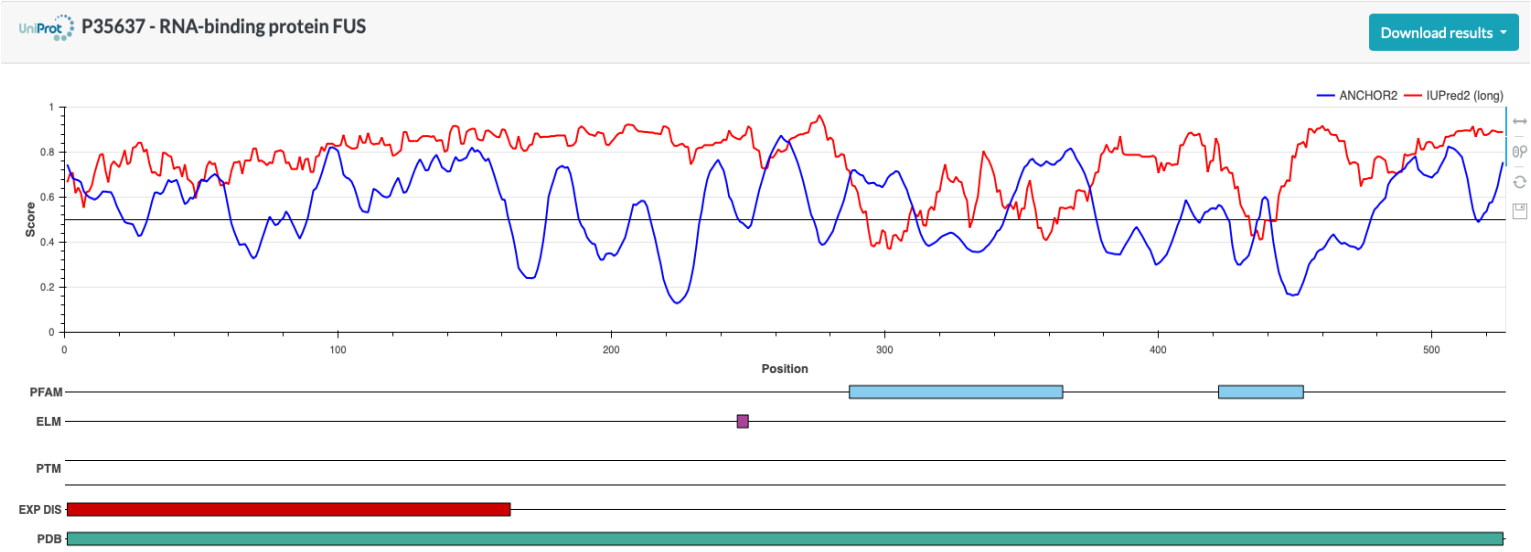


Class Mini Project 2 (competitive)

PART 1

Start from the protein FUS (Uniprot ID: P35637)

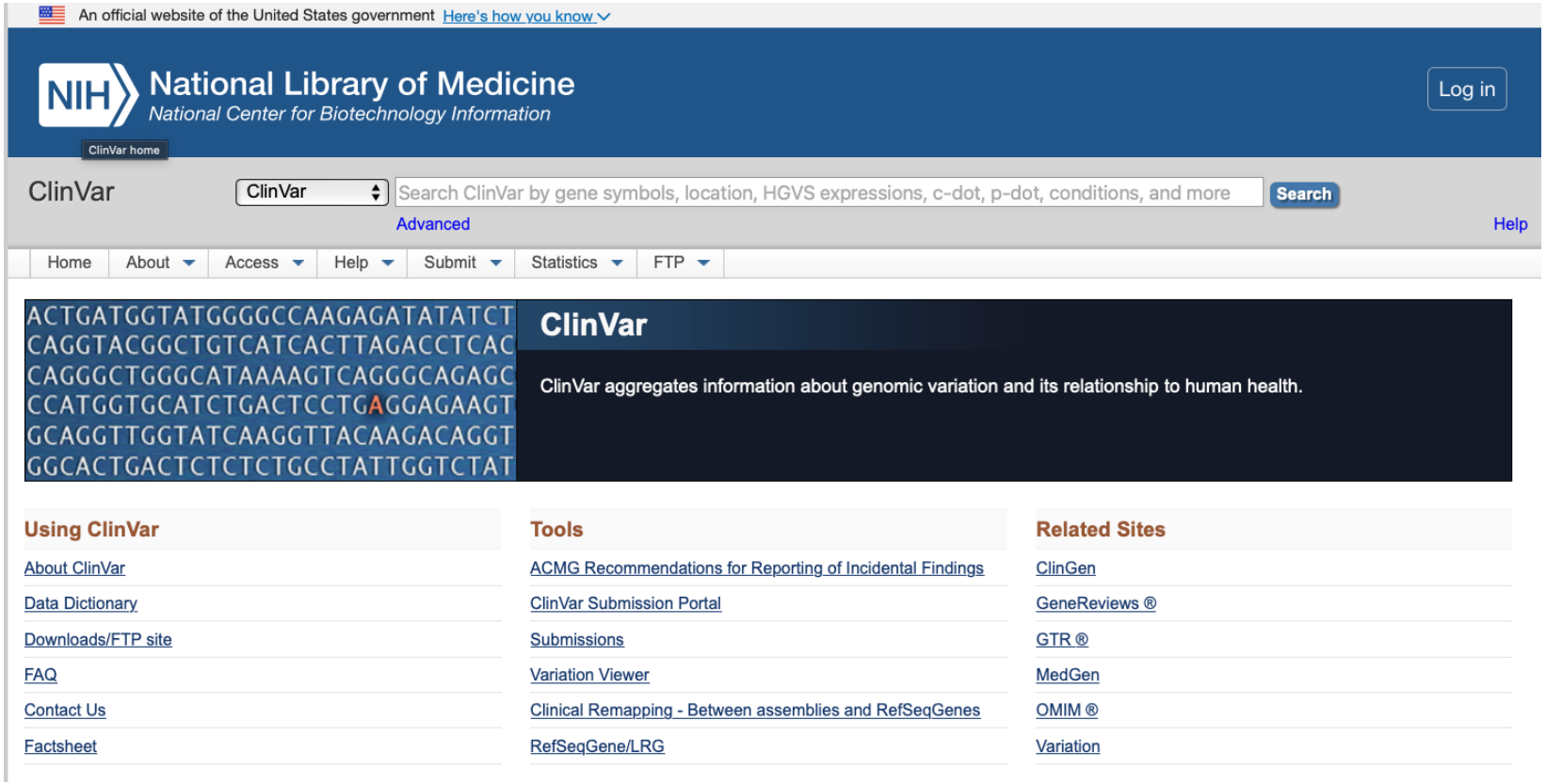
Q1: Is this protein structured or disordered? (Iupred database)



Class Mini Project 2 (competitive)

PART 1

Q2: Mutations in this protein can cause what diseases?



Q3: Where in the protein?

Q4: Can we learn anything from the PDB database?

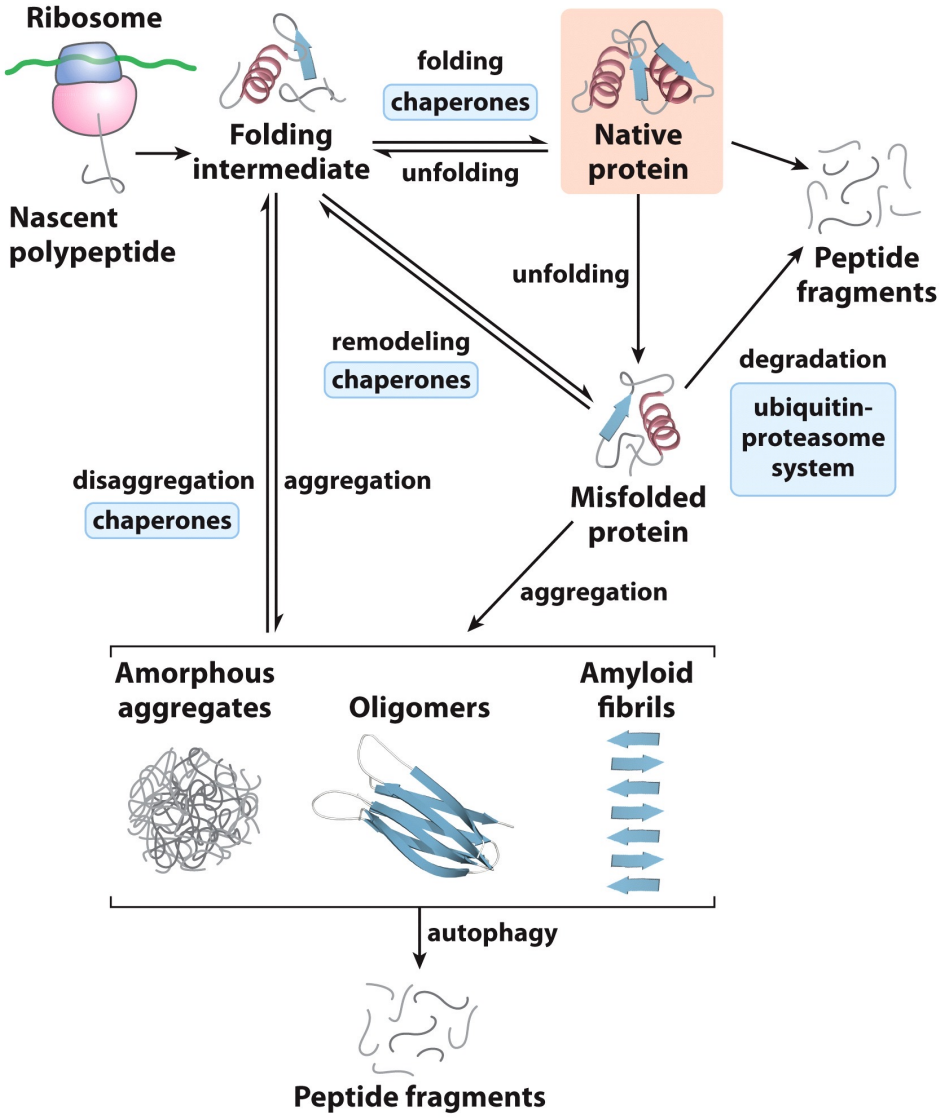
Proteostasis

The maintenance of an active set of cellular proteins required under a given set of conditions is called proteostasis. Three kinds of processes contribute to proteostasis:

First, proteins are synthesized on a ribosome.

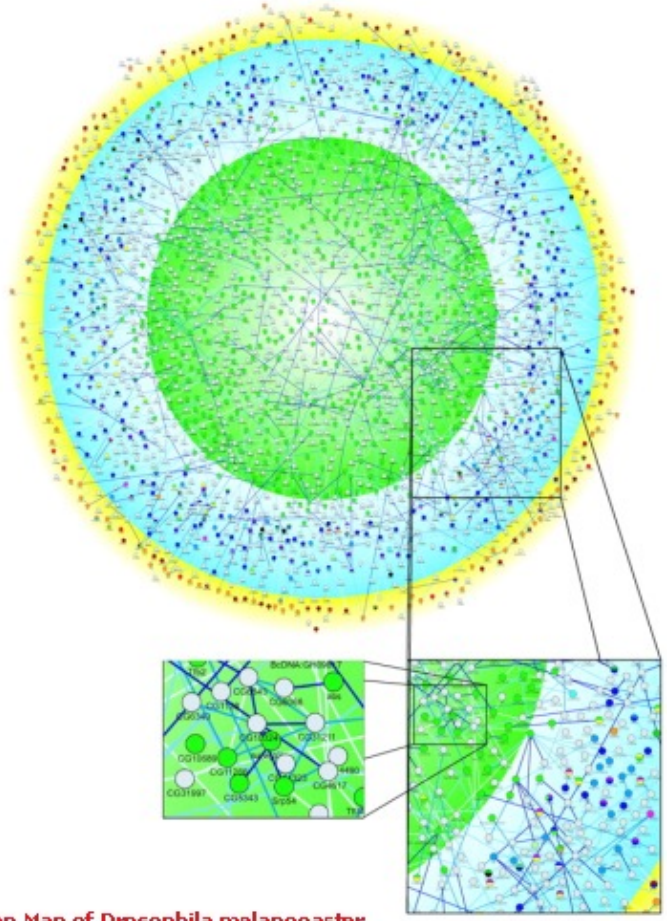
Second, multiple pathways contribute to protein folding, many of which involve the activity of complexes called chaperones. Chaperones (including chaperonins) also contribute to the refolding of proteins that are partially and transiently unfolded.

Third, proteins that are irreversibly unfolded are subject to sequestration and degradation by several additional pathways



Part 2:
Protein Interactions

Proteins work together!



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

Proteins “work together” forming multi complexes to carry out the specific functions

Identification of interactions

Experimental

- X-ray crystallography
- NMR spectroscopy
- Mass spectrometry (Tandem affinity purification)
- Immunoprecipitation
- Yeast two-hybrid
- Microarrays

Computational

Genomic data

- Phylogenetic profiling
- Gene context
- Gene fusion
- Symmetric evolution

Structural data

- Sequence profile
- 3D structural distance matrix
- Surface patches
- Binding interactions

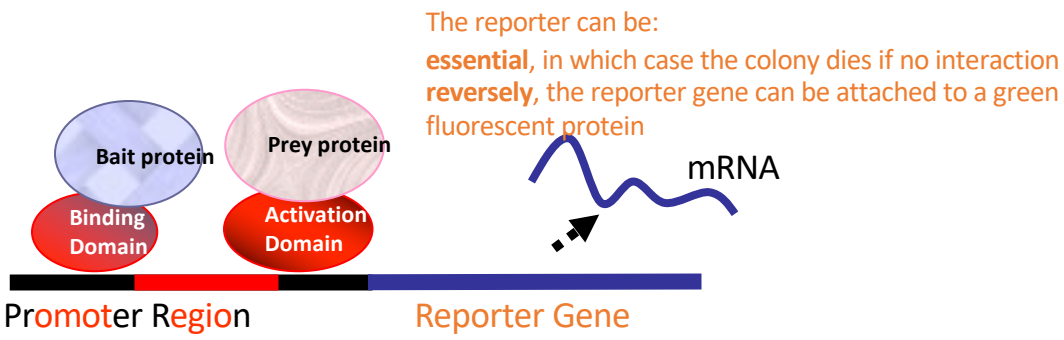
Yeast Two-Hybrid System

A transcription factor is split into 2 domains and two hybrid proteins are designed.

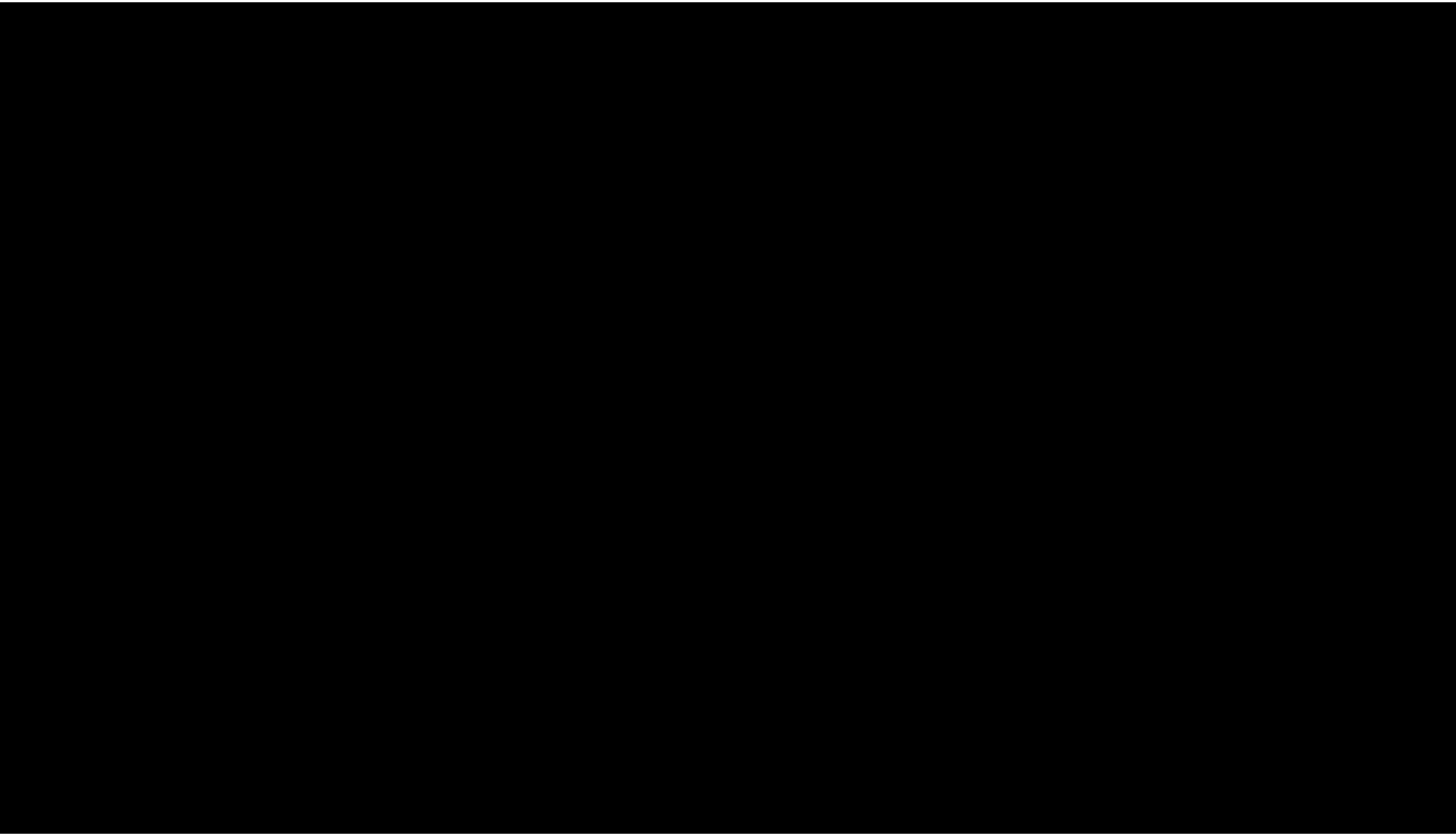
One protein of interest (bait) is typically fused to a DNA-binding domain.

The proteins being screened for interactions with the bait (preys) are fused to a transcription-activating domain.

An interaction between the bait and a prey will bring these 2 domains close together which in turn results in the transcription of a reporter gene.



Clinical Proteomics: Proximity Labeling



Credit: Harvard Medical School

Protein structure determines interactions with other molecules

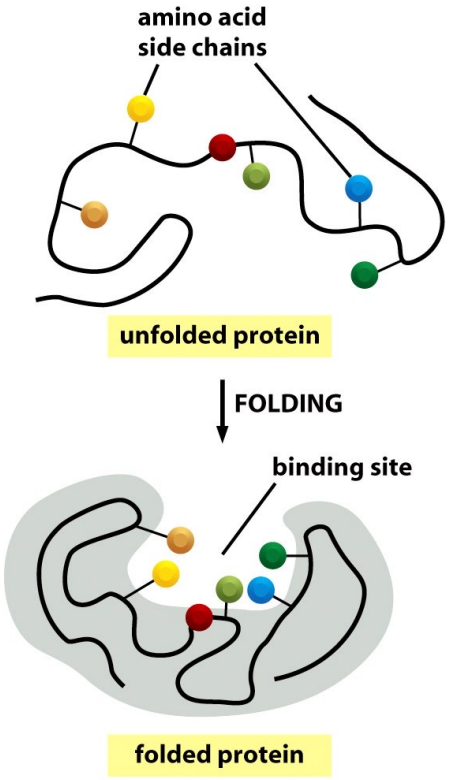
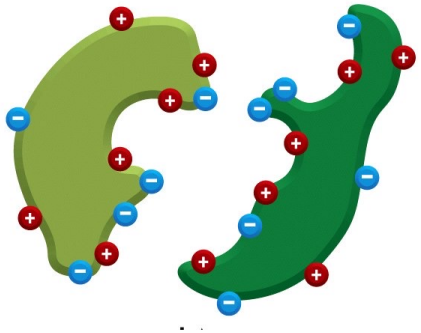


Figure 3-37a Molecular Biology of the Cell 5/e (© Garland Science 2008)

Protein structure affects protein-protein interactions



Protein structure affects protein-protein interactions

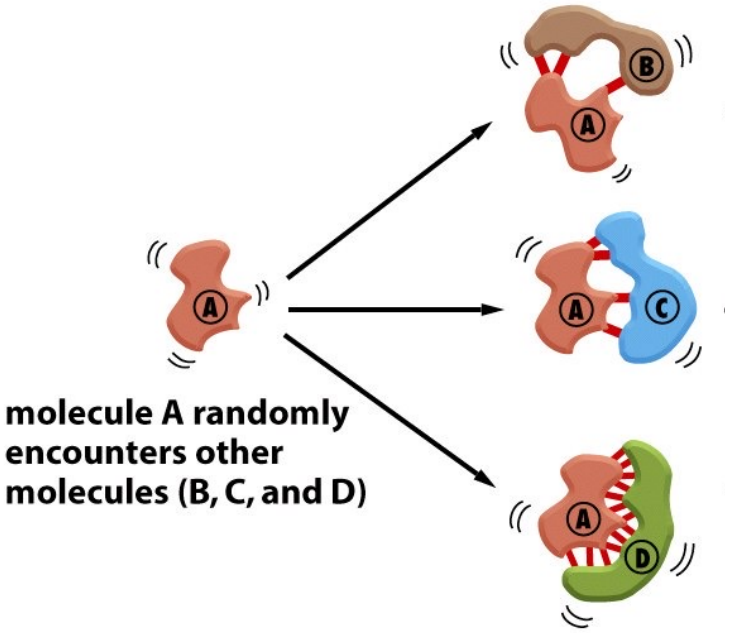
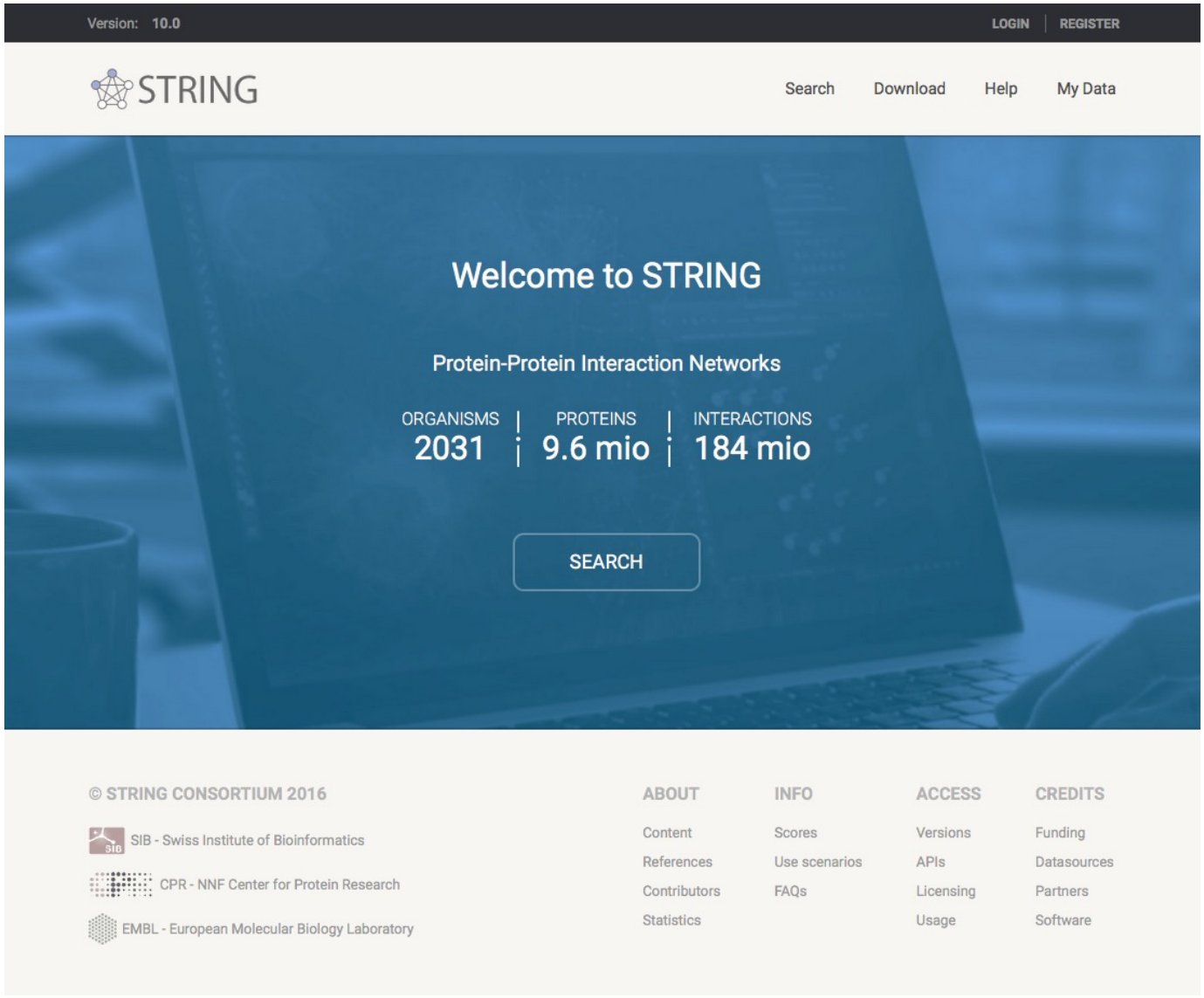


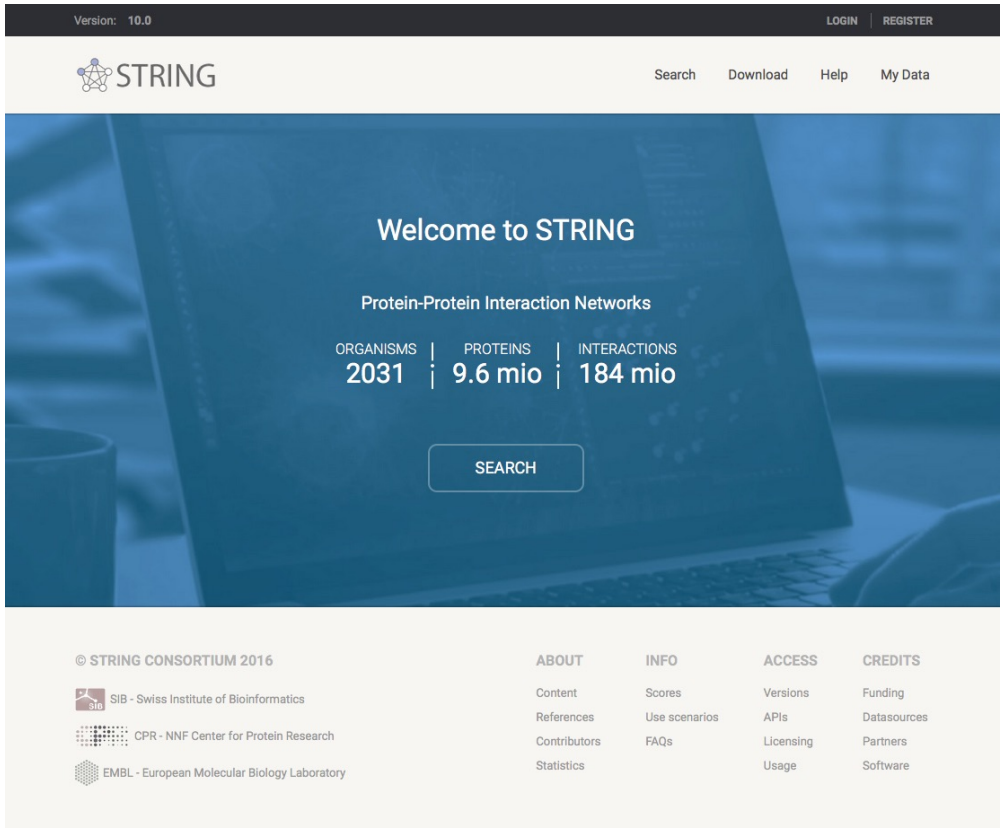
Figure 3-42 Molecular Biology of the Cell 5/e (© Garland Science 2008)

STRING database



Class Mini Project 2 (competitive)

PART 2

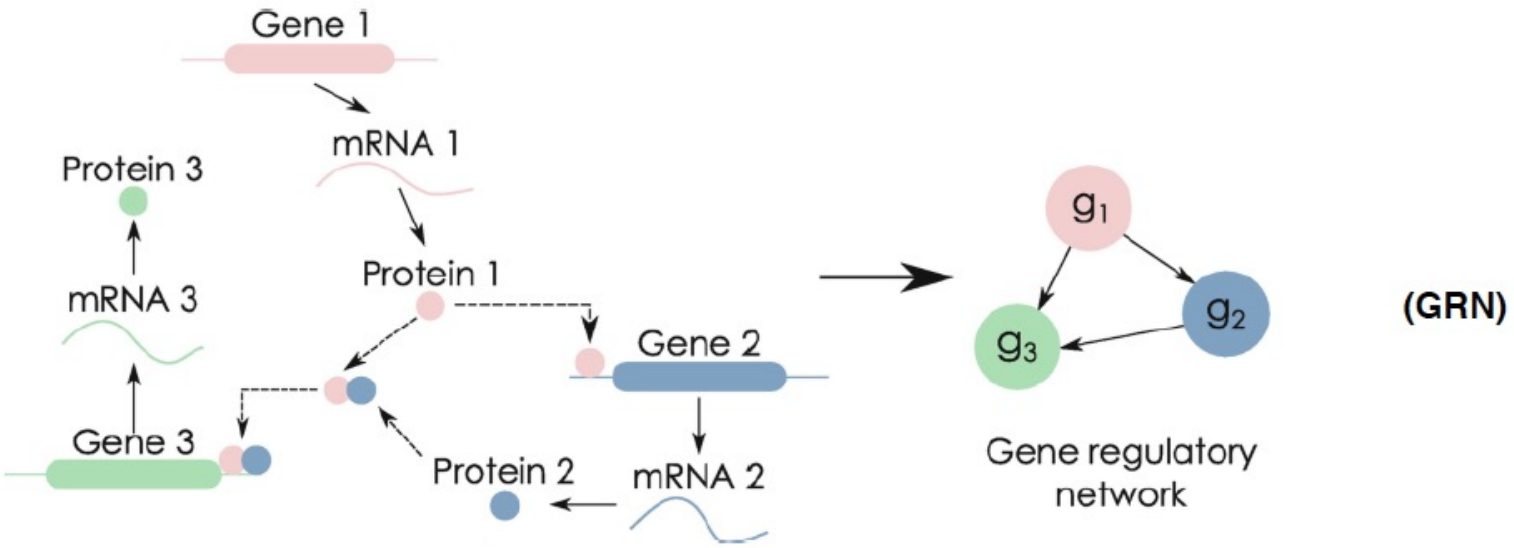


Q5: How protein-protein interactions can help us understand the role of FUS in neurological diseases?

Part 3:
Protein Networks

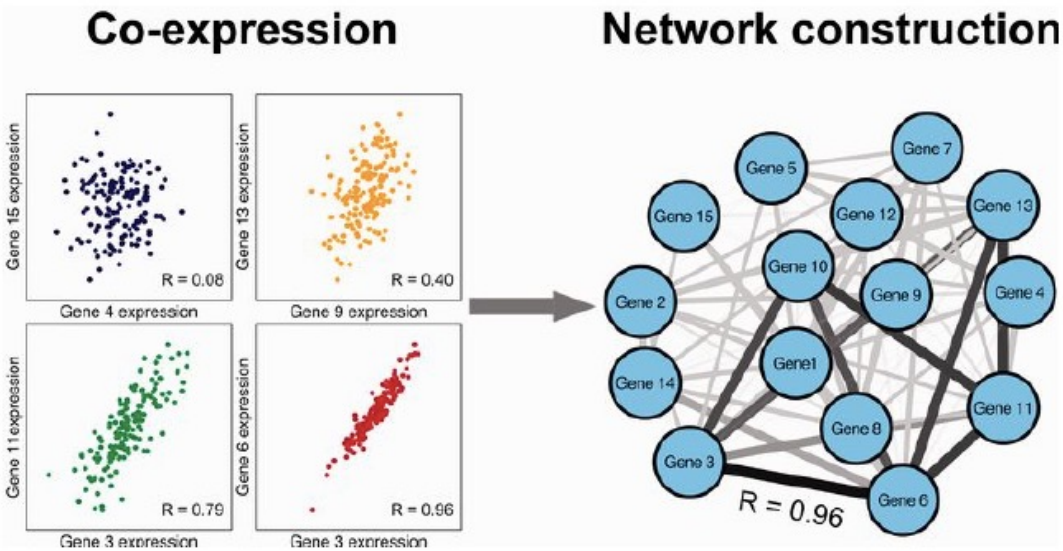
Why do we have different tissues and cell types if we have only one DNA molecule?

Gene Regulatory Networks



A set of molecular **regulators** that interact with each other to govern the **gene expression** levels of mRNA and proteins >> the **function** of the **cell**.

Gene Co-expression Networks



(GCN)

A gene co-expression network can be built by finding pairs of genes which show a similar expression pattern across samples.

Gene Co-expression Networks

RNA-seq or Microarray experiments

	S_1	S_2	S_3
G_1	43.26	40.89	5.05
G_2	166.6	41.87	136.65
G_3	12.53	39.55	42.09
G_4	28.77	191.92	236.56
G_5	114.7	79.7	99.76
G_6	119.1	80.57	114.59
G_7	118.9	156.69	186.95
G_8	3.76	2.48	136.78
G_9	32.73	11.99	118.8
G_{10}	17.46	56.11	21.41

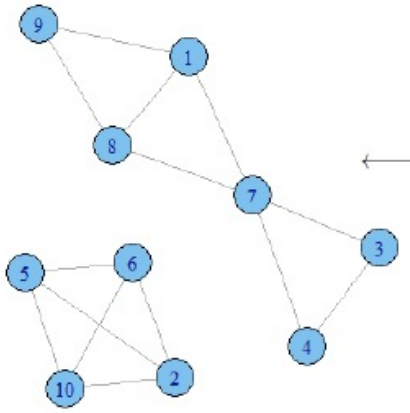
Gene expression values

$|r(G_i, G_j)|$

Pearson correlation

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	**1.00**	0.23	0.61	0.71	0.03	0.35	**0.86**	**1.00**	**0.97**	0.37
G_2	0.23	**1.00**	0.63	0.52	**0.98**	**0.99**	0.29	0.30	0.46	**0.99**
G_3	0.61	0.63	**1.00**	**0.99**	0.77	0.53	**0.93**	0.56	0.41	0.51
G_4	0.71	0.52	**0.99**	**1.00**	0.69	0.41	**0.97**	0.66	0.52	0.40
G_5	0.03	**0.98**	0.77	0.69	**1.00**	**0.95**	0.48	0.09	0.27	**0.94**
G_6	0.35	**0.99**	0.53	0.41	**0.95**	**1.00**	0.17	0.41	0.57	**1.00**
G_7	0.86	0.29	**0.93**	**0.97**	0.48	0.17	**1.00**	**0.83**	0.72	0.16
G_8	**1.00**	0.30	0.56	0.66	0.09	0.41	0.83	**1.00**	**0.98**	0.42
G_9	**0.97**	0.46	0.41	0.52	0.27	0.57	0.72	**0.98**	**1.00**	0.58
G_{10}	0.37	**0.99**	0.51	0.40	**0.94**	**1.00**	0.16	0.42	0.58	**1.00**

Similarity (Co-expression) score



	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	0	0	0	0	0	0	1	1	1	0
G_2	0	0	0	0	1	1	0	0	0	1
G_3	0	0	0	1	0	0	1	0	0	0
G_4	0	0	1	0	0	0	1	0	0	0
G_5	0	1	0	0	0	1	0	0	0	1
G_6	0	1	0	0	1	0	0	0	0	1
G_7	1	0	1	1	0	0	0	1	0	0
G_8	1	0	0	0	0	0	1	0	1	0
G_9	1	0	0	0	0	0	0	1	0	0
G_{10}	0	1	0	0	1	1	0	0	0	0

Network adjacency matrix

$|r(G_i, G_j)| >= 0.8$

Significance threshold

14

Metabolic Networks

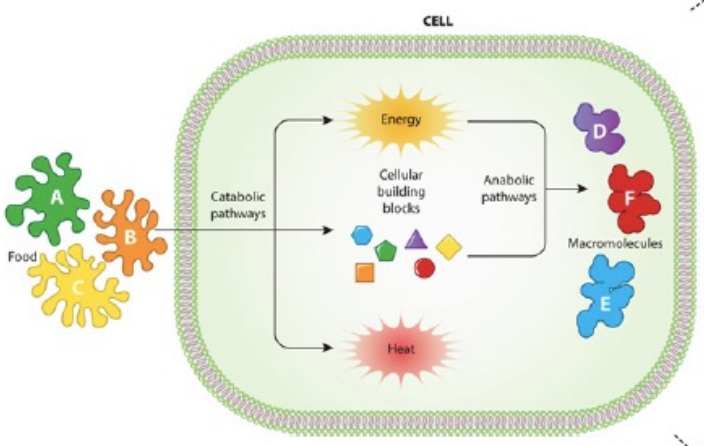
A a set of chemical reactions that produces:

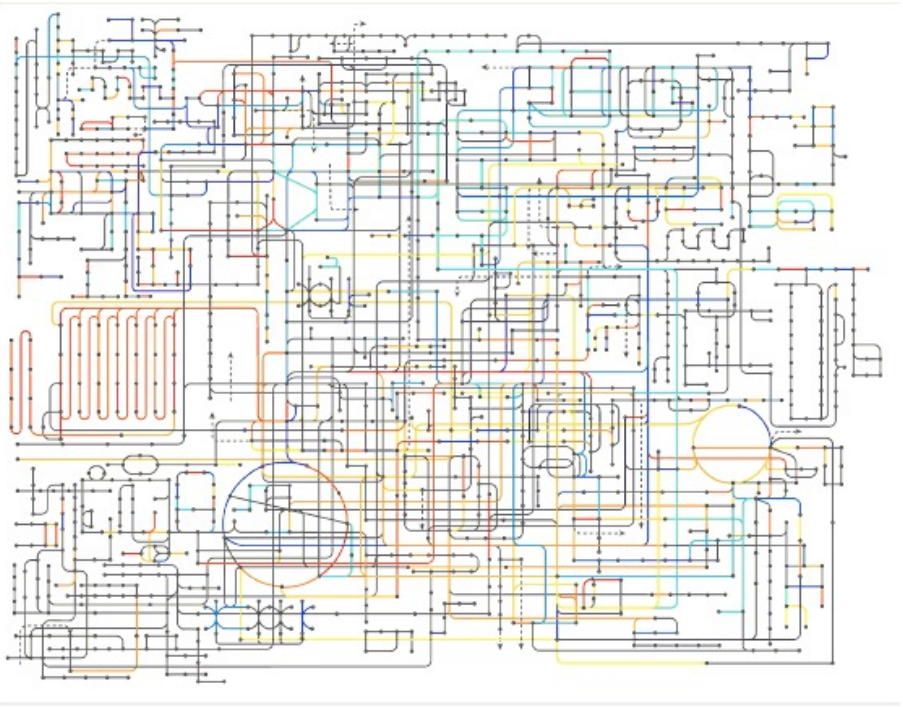
1) energy
(for maintenance of cell functions and for biosyntheses)

2) molecular building blocks for biosyntheses

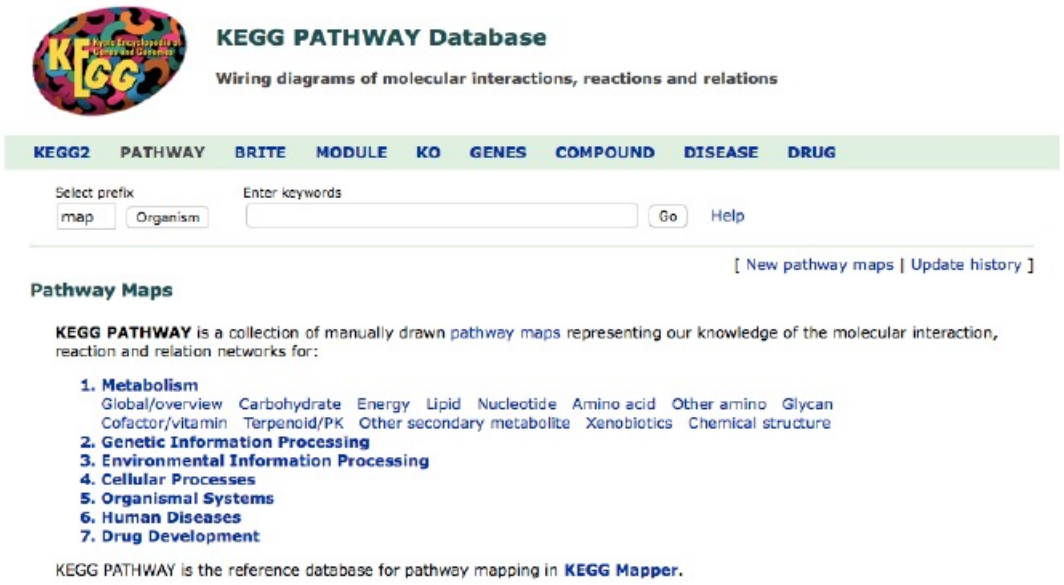
These reactions are catalyzed by enzymes that are encoded by genes.

Metabolic Networks





Metabolic networks



KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

KEGG2 PATHWAY BRITE MODULE KO GENES COMPOUND DISEASE DRUG

Select prefix: map | Organism | Enter keywords | Go | Help

[New pathway maps | Update history]

Pathway Maps

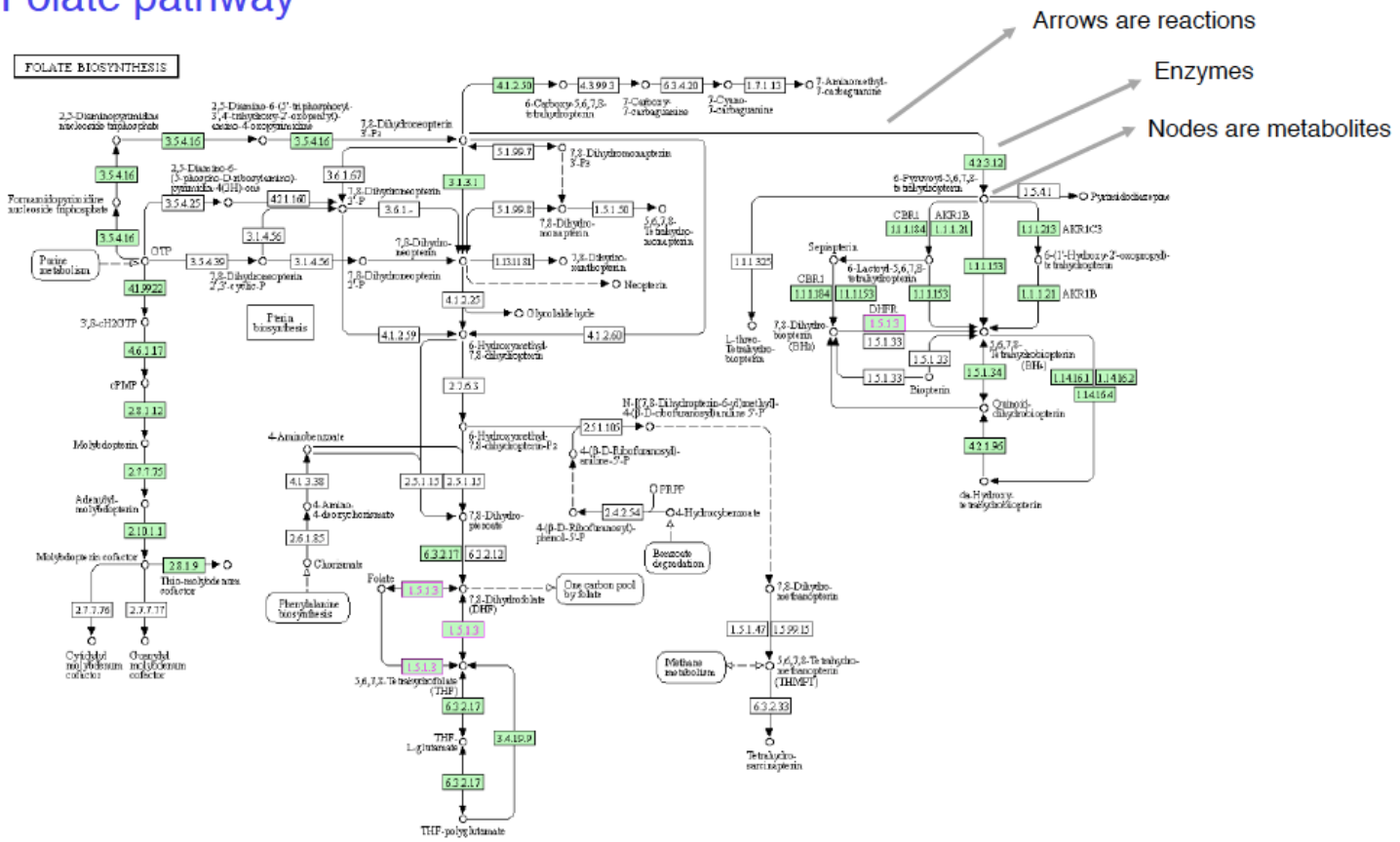
KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:

1. **Metabolism**
   Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan
   Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human Diseases**
7. **Drug Development**

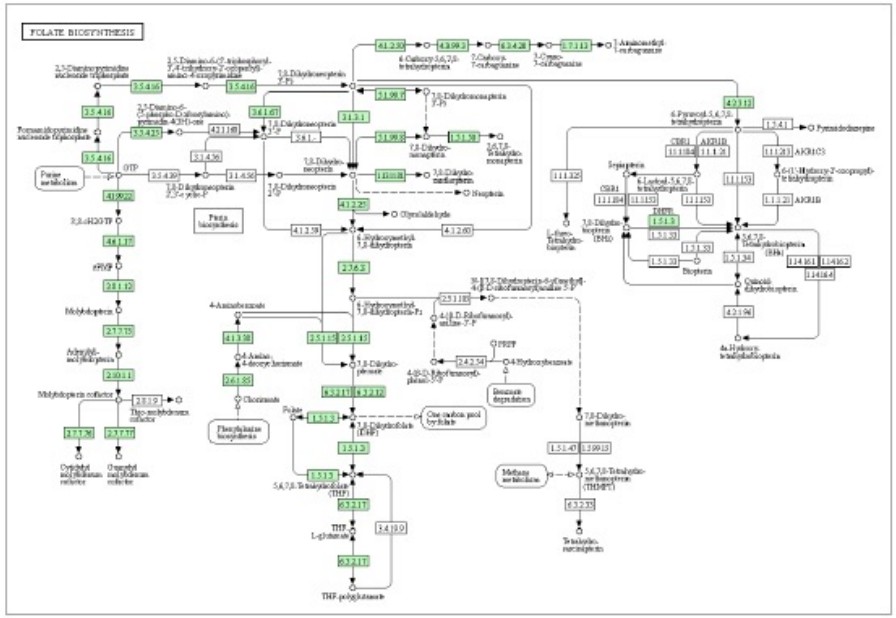
KEGG PATHWAY is the reference database for pathway mapping in **KEGG Mapper**.

Find the folate biosynthetic pathway in Human and *E.coli*

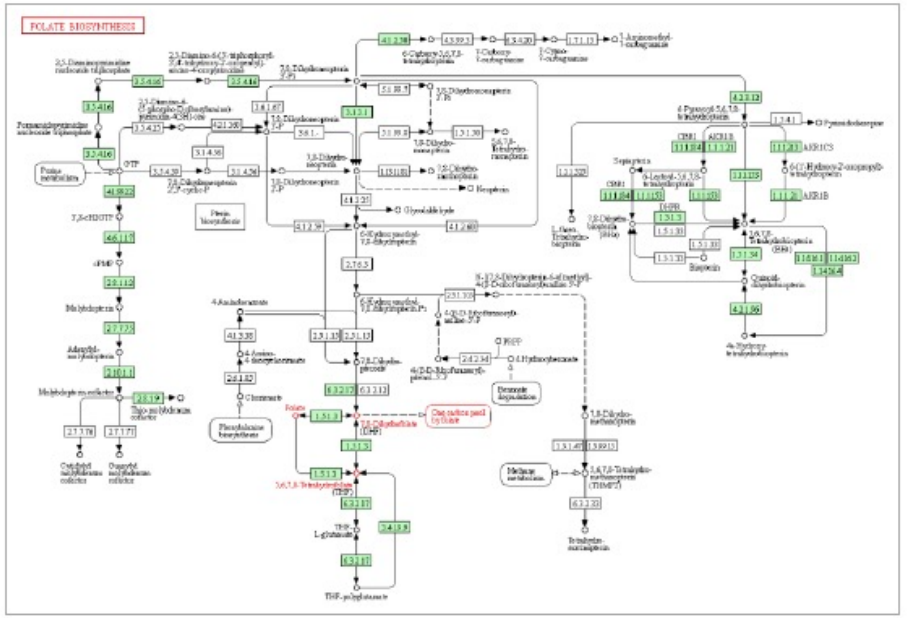
Folate pathway



Folate Pathway in Human and E.coli

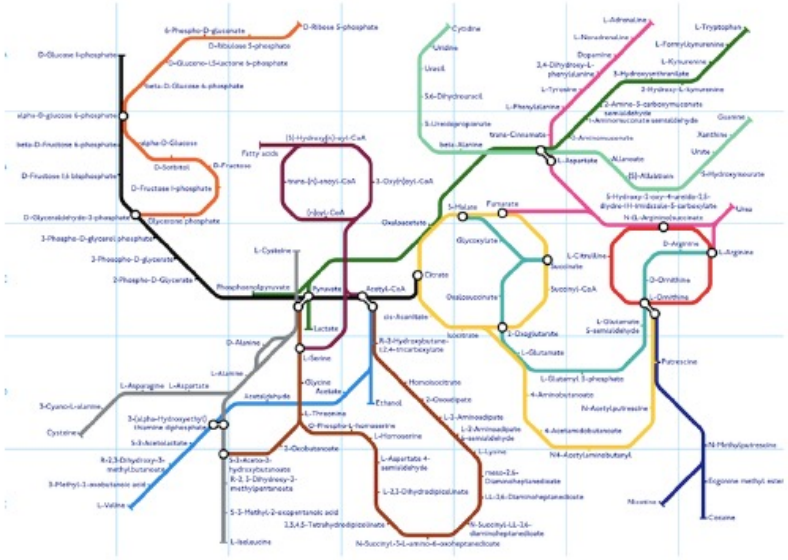


E.coli



Human

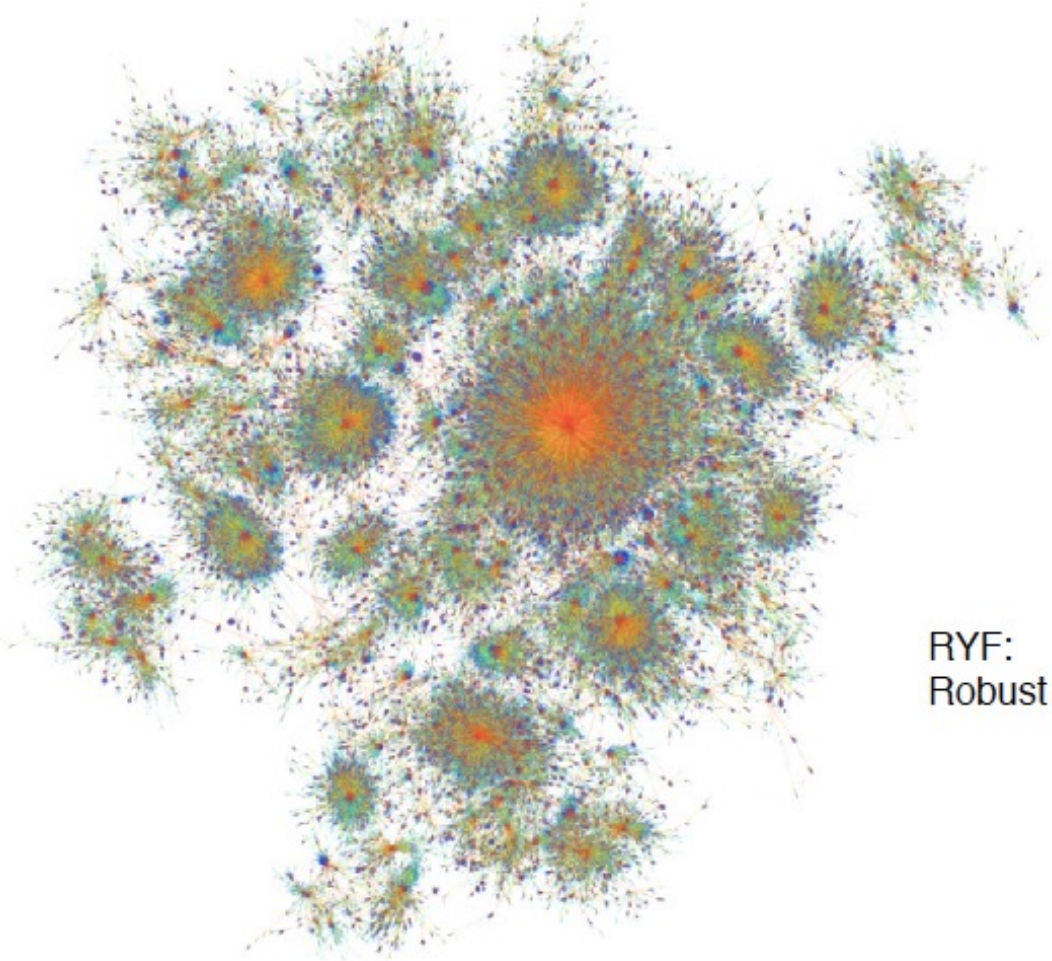
Biological Networks and Complex Networks have similarities



Metabolic map

New York subway map

Network Topology Matters!

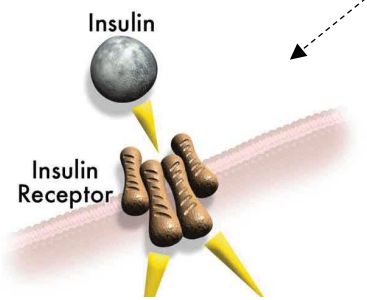


RYF:
Robust yet Fragile!

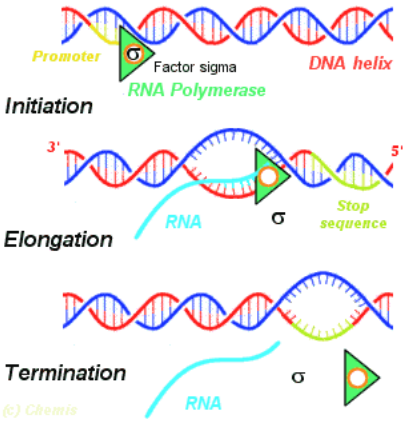
Gene Ontology Terms

Experiments (Differential Expression, Interactions, …)

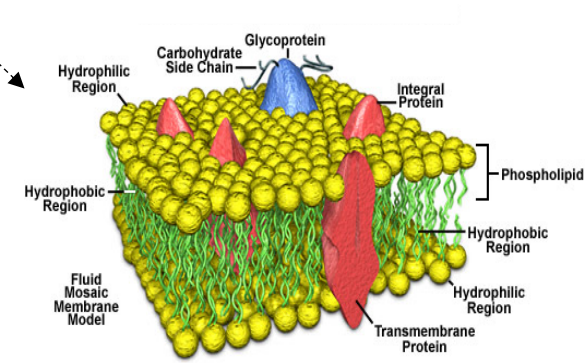
{Gene 1, Gene 2, Gene 3, ..}



Molecular Function



Biological Process

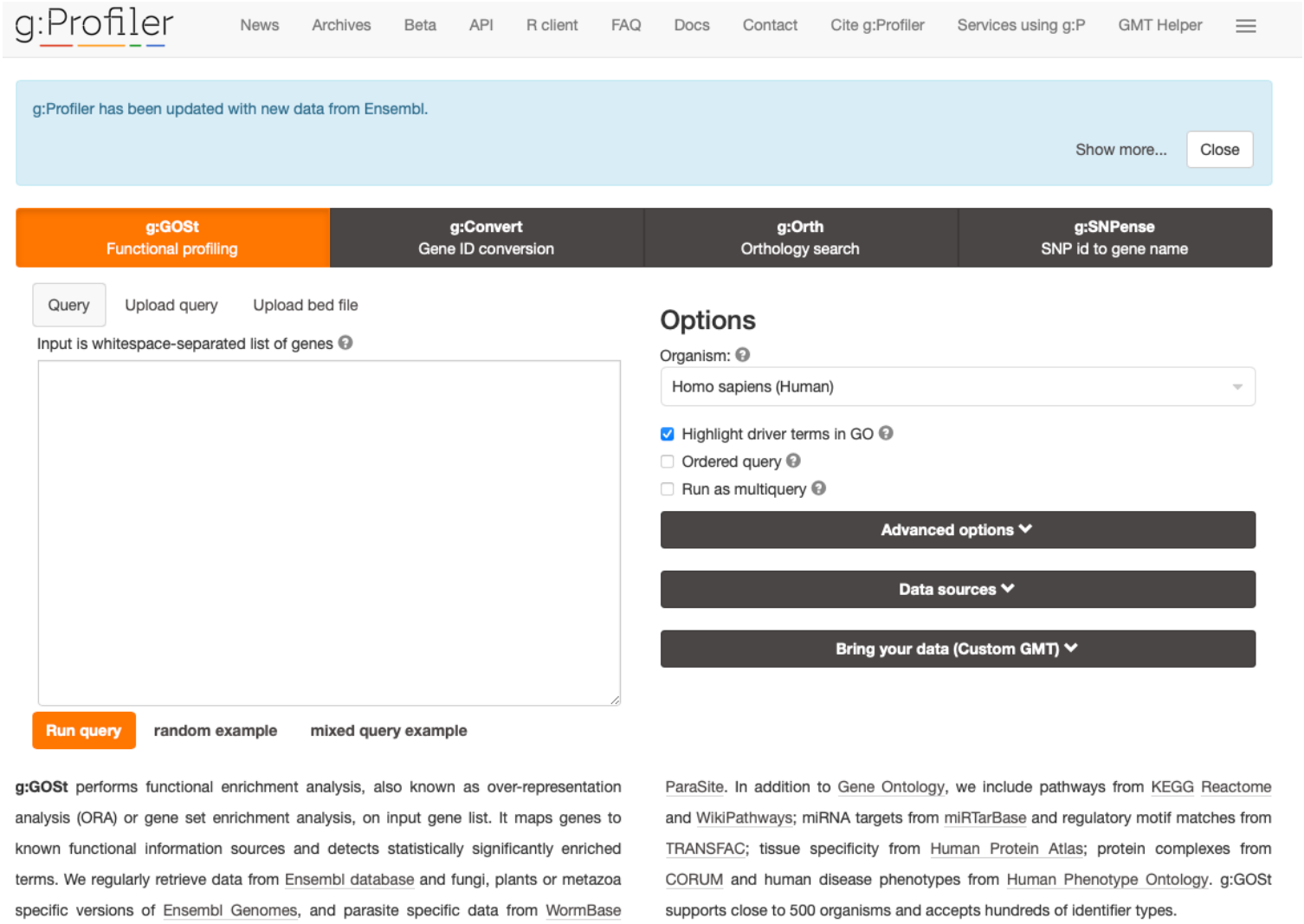


Cellular Component

\+

(Transcription Factors, Tissue Specificity, …)

GO-term enrichment analysis



Class Mini Project 3 (competetive)

Construct the protein interaction networks of the Alzheimer's Disease

STRING (disease module) +

GO term enrichment (processes)

Q: What biological processes are involved in the Alzheimer's disease?

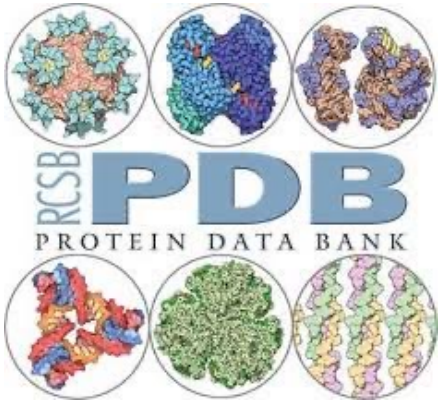
Extra Activities

Check the metabolic networks of the following diseases and find the genes within the metabolic networks whose mutations contribute to the disease.

Group 1:
Phenylketonuria (PKU)

Group 2:
Tai-sachs Disease

An overview of databases and tools reviewed



On exam questions

Question 1: Protein aggregation-based diseases

Question 2: Databases of Proteins and their interactions