

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

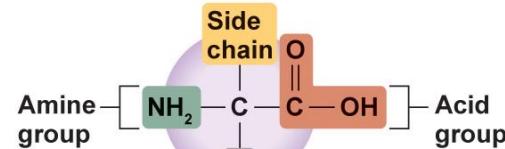
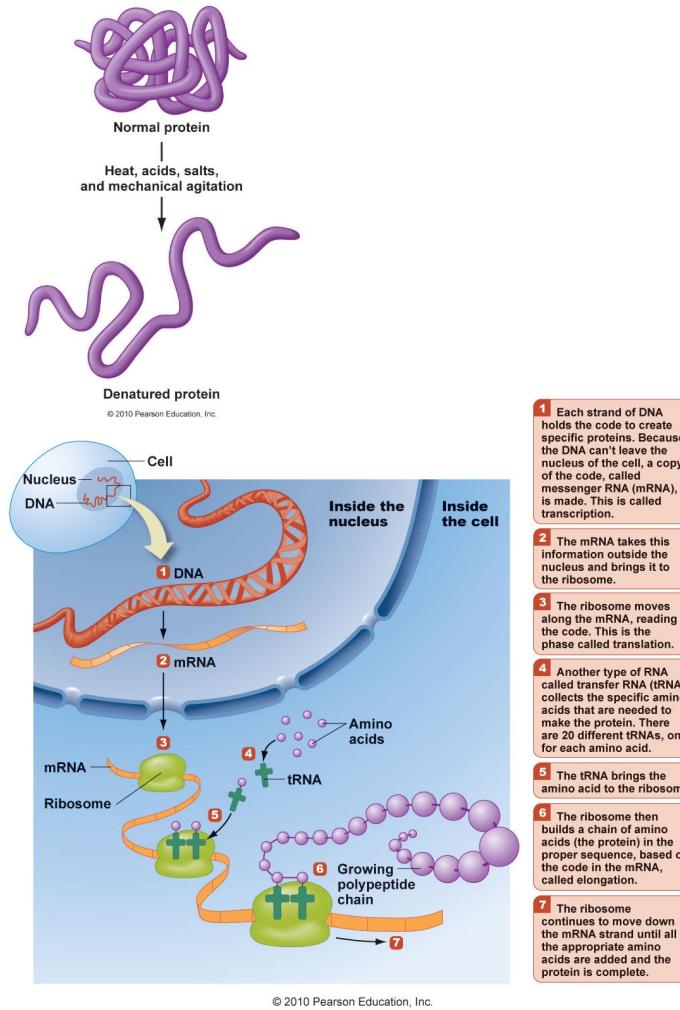
Point 1: This is an interactive lecture based on “project-based learning”:

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

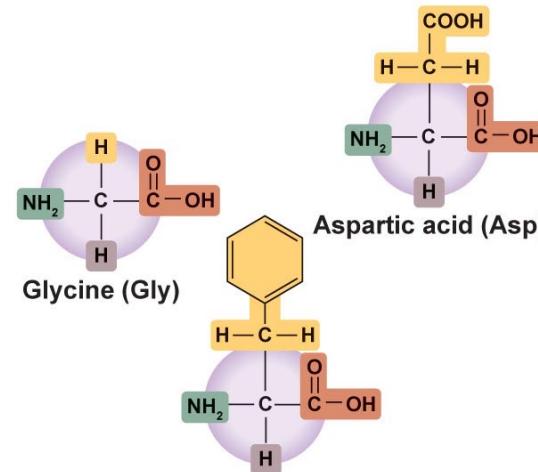
Point 2: All analyses can be done using R/python libraries. We only show a few analyses here.

Part 1: Protein Sequence, Structure, Function

Proteins and Amino Acids



a Amino acid structure

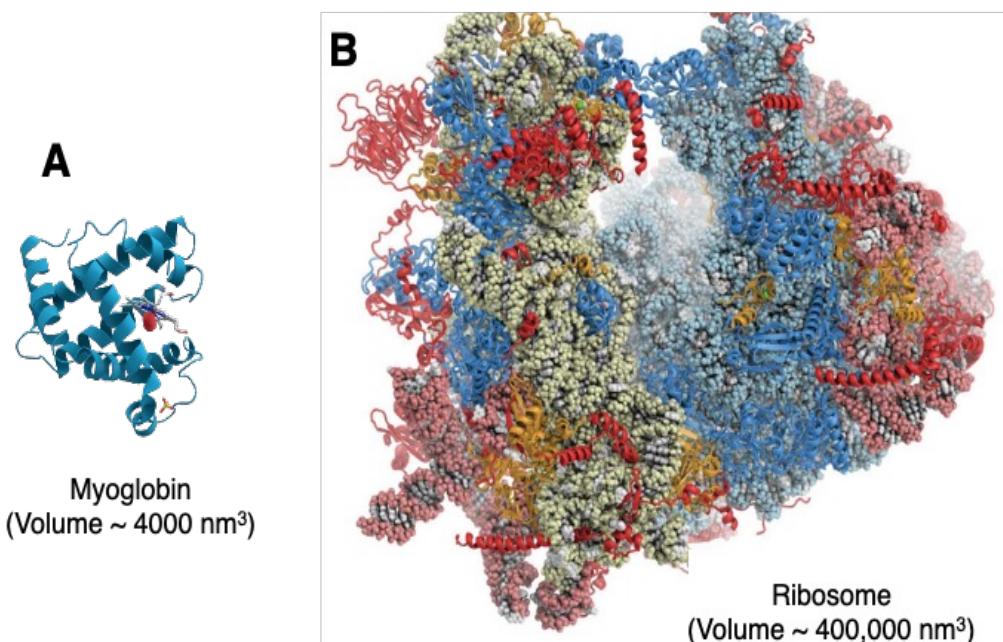


b Different amino acids, showing their unique side chains

© 2010 Pearson Education, Inc.

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



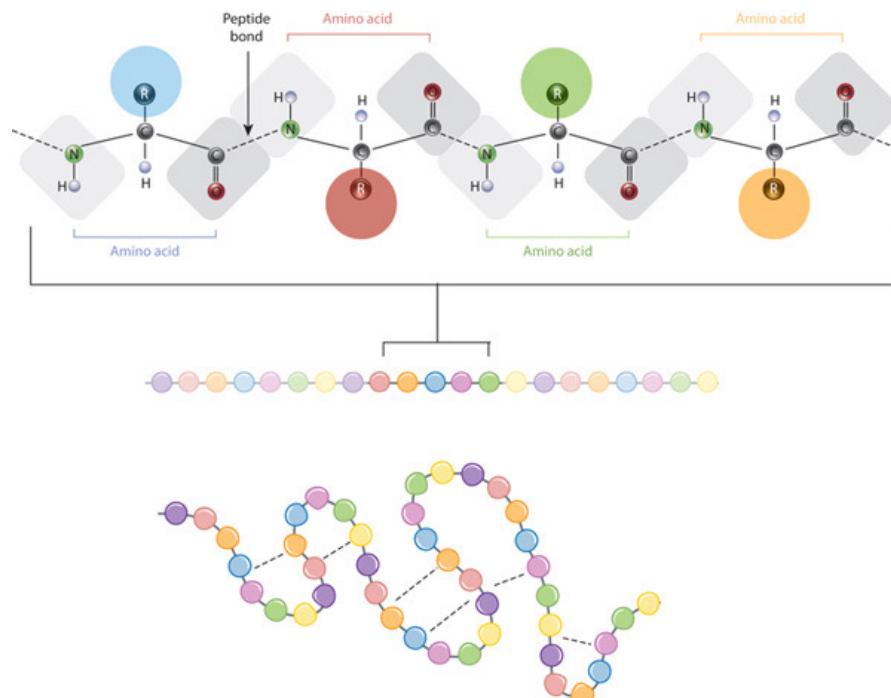
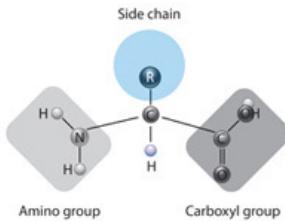
Structure of Proteins

- 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 synthesized 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

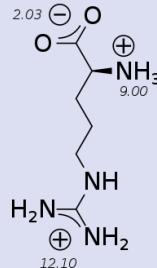
⊕ Positive
⊖ Negative

A. Amino Acids with Electrically Charged Side Chains

Positive

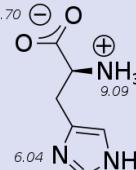
Arginine

Arg R



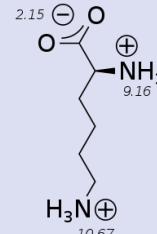
Histidine

His H



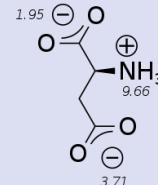
Lysine

Lys K



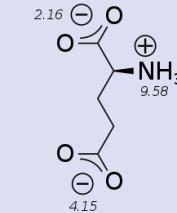
Aspartic Acid

Asp D



Glutamic Acid

Glu E

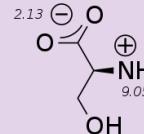


Negative

B. Amino Acids with Polar Uncharged Side Chains

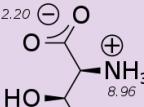
Serine

Ser S



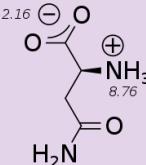
Threonine

Thr T



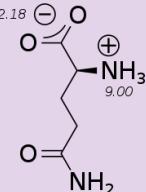
Asparagine

Asn N



Glutamine

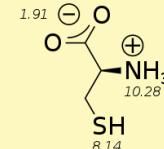
Gln Q



C. Special Cases

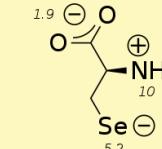
Cysteine

Cys C



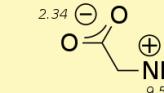
Selenocysteine

Sec U



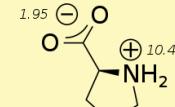
Glycine

Gly G



Proline

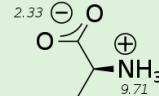
Pro P



D. Amino Acids with Hydrophobic Side Chains

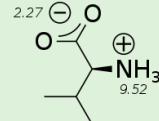
Alanine

Ala A



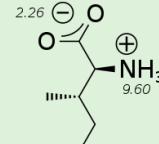
Valine

Val V



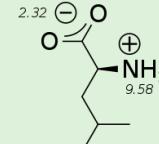
Isoleucine

Ile I



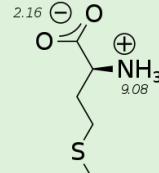
Leucine

Leu L



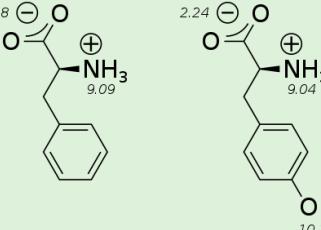
Methionine

Met M



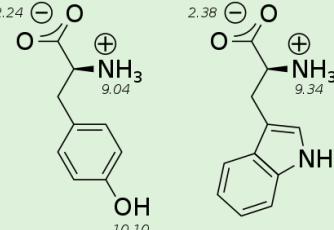
Phenylalanine

Phe F



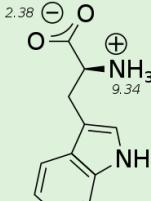
Tyrosine

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

P68871 · HBB_HUMAN

Protein¹ Hemoglobin subunit beta
 Gene¹ HBB
 Status UniProtKB reviewed (Swiss-Prot)
 Organism Homo sapiens (Human)

Amino acids 147
 Protein existence Evidence at protein level
 Annotation score 60

Entry Variant viewer Feature viewer Publications External links History

BLAST Download Add Community curation (3) Add a publication Entry feedback

Fur
 Involved in LVV-h
 Spans in LVV-h
 Functions in RDX/XML
 MISC GFF
 One M

Text
 FASTA (canonical)
 FASTA (canonical & isoform)
 JSON
 XML
 RDF/XML
 GFF
 One M

ang to the various peripheral tissues. [Publication]
 ity of bradykinin, causing a decrease in blood pressure.
 Enkephalin-degrading enzymes such as DPP-3, and as a selective antagonist of the P2RX3 receptor which is involved in pain signaling; these properties implicate it as a regulator of pain
 One M

Download the Fasta file

```
>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens OX=9606 GN=HBB PE=1 SV=2
MVHLTPEEKSATLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLNDNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAHHFG
KEFTPQAAQKVVAGVANALAHKYH
```

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

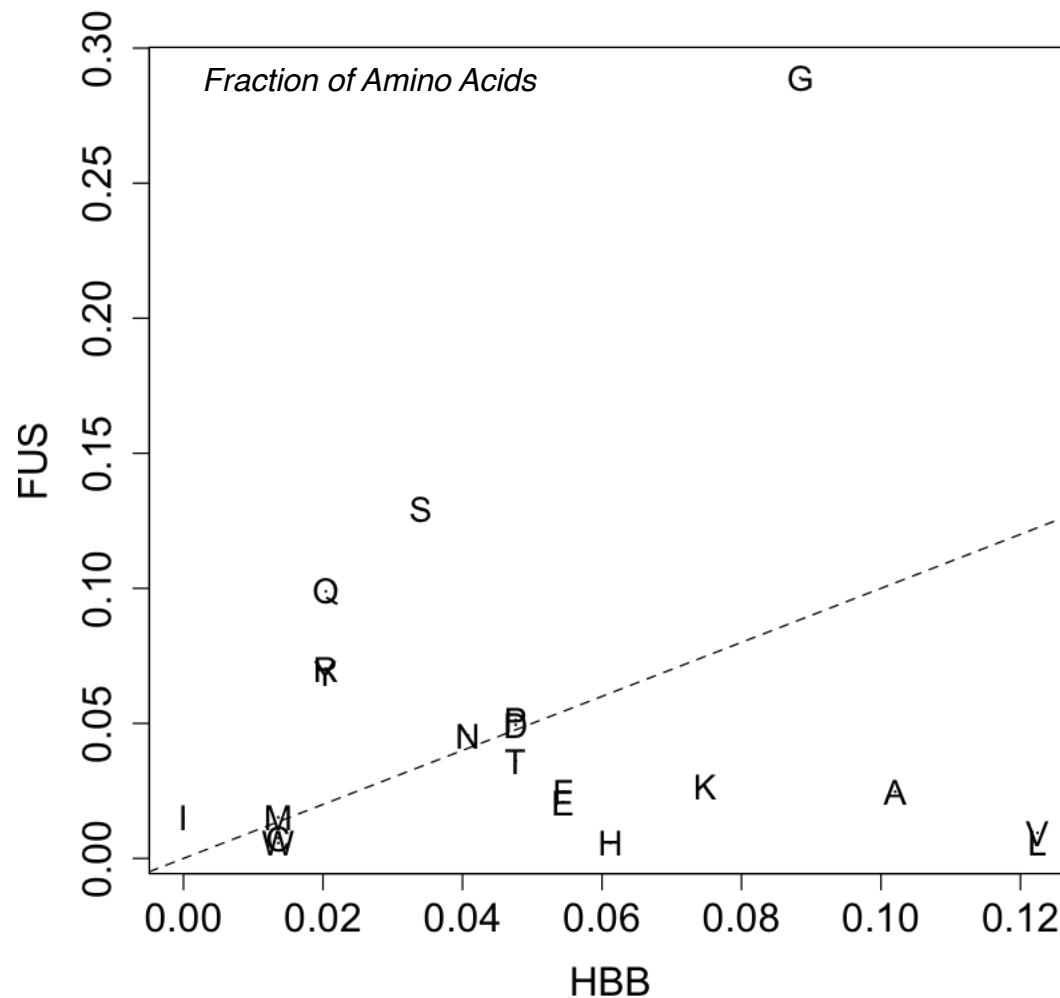
```
# Install and load the protr library
install.packages("protr")
library("protr")

# 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`)

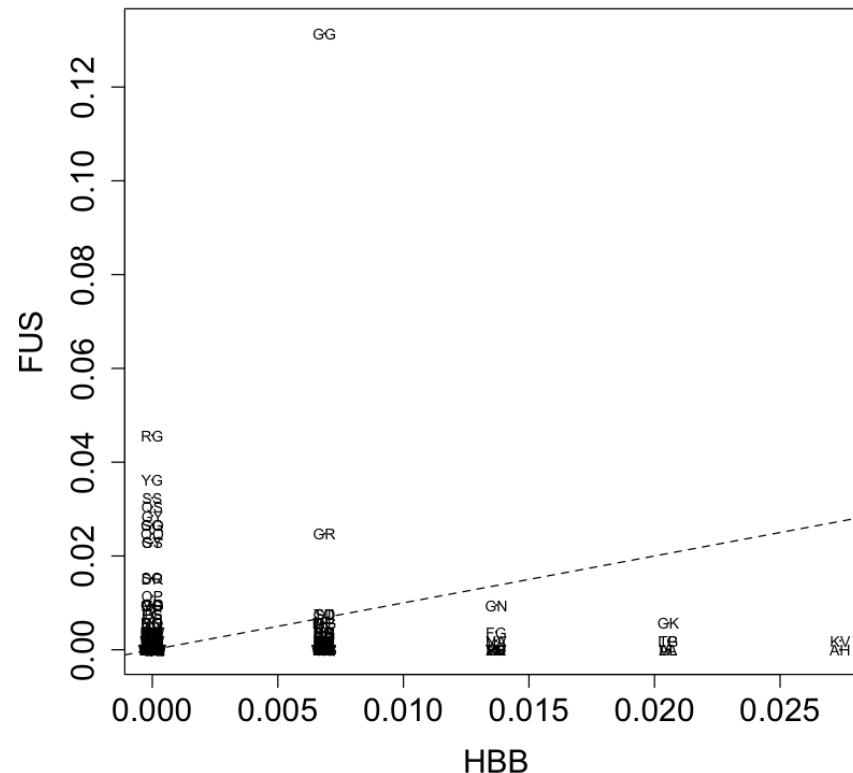
# Plot the composition of two proteins against each other
plot(HBB_frac, FUS_frac, cex=0.1, xlab="HBB", ylab="FUS", cex.axis=1.5, cex.lab=1.5)
text(HBB_frac, FUS_frac, names(HBB_frac), cex=1.3)
abline(a=0, b=1, lty=2)
```

What do you see?



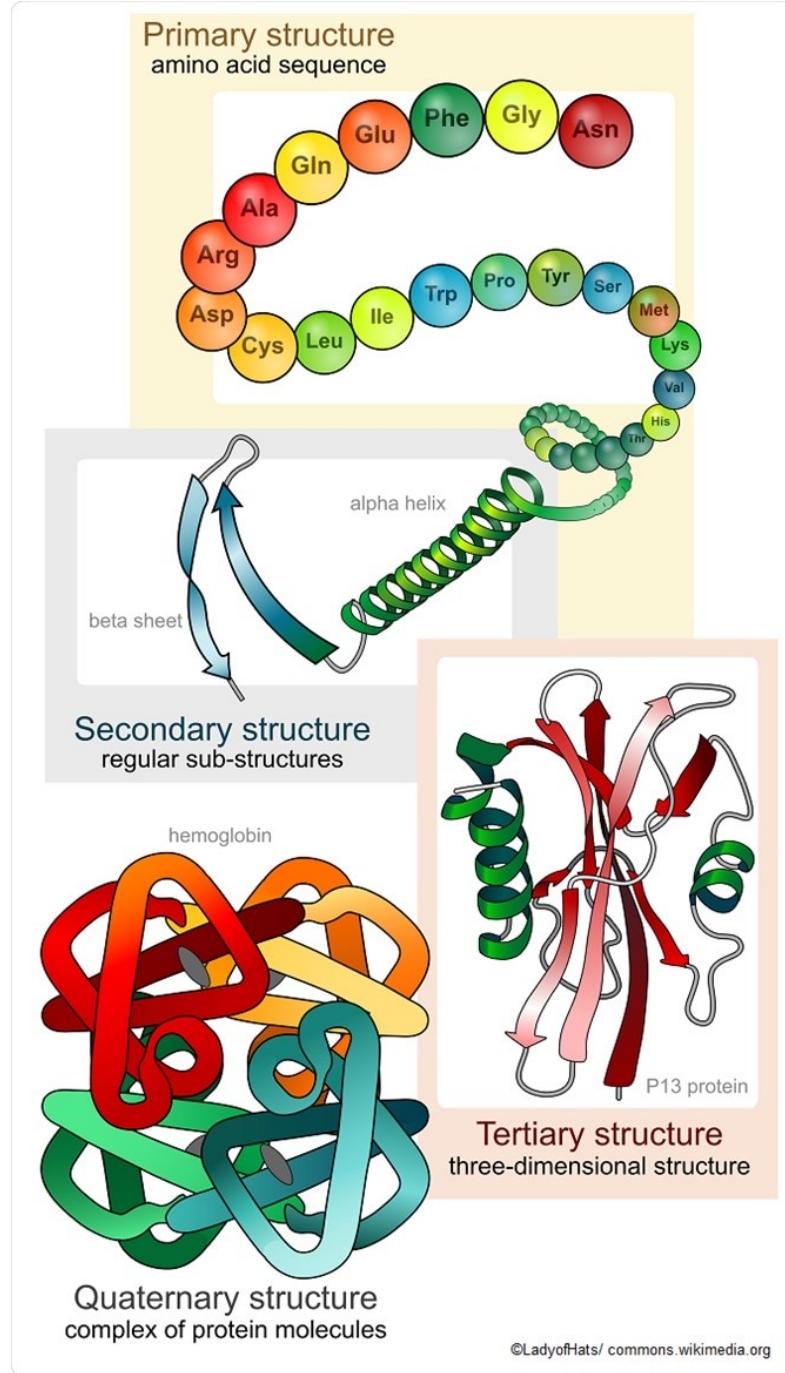
```
# Extract the dipeptide's composition
HBB_frac_dipeptides <- extractDC(HBB$`sp|P68871|HBB_HUMAN`)
FUS_frac_dipeptides <- extractDC(FUS$`sp|P35637|FUS_HUMAN`)

# 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)	Residues (%)*	
	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 Worldwide Protein Data Bank EMD-DB 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.

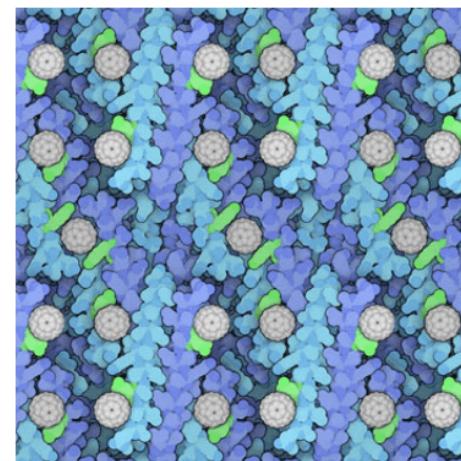
As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

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

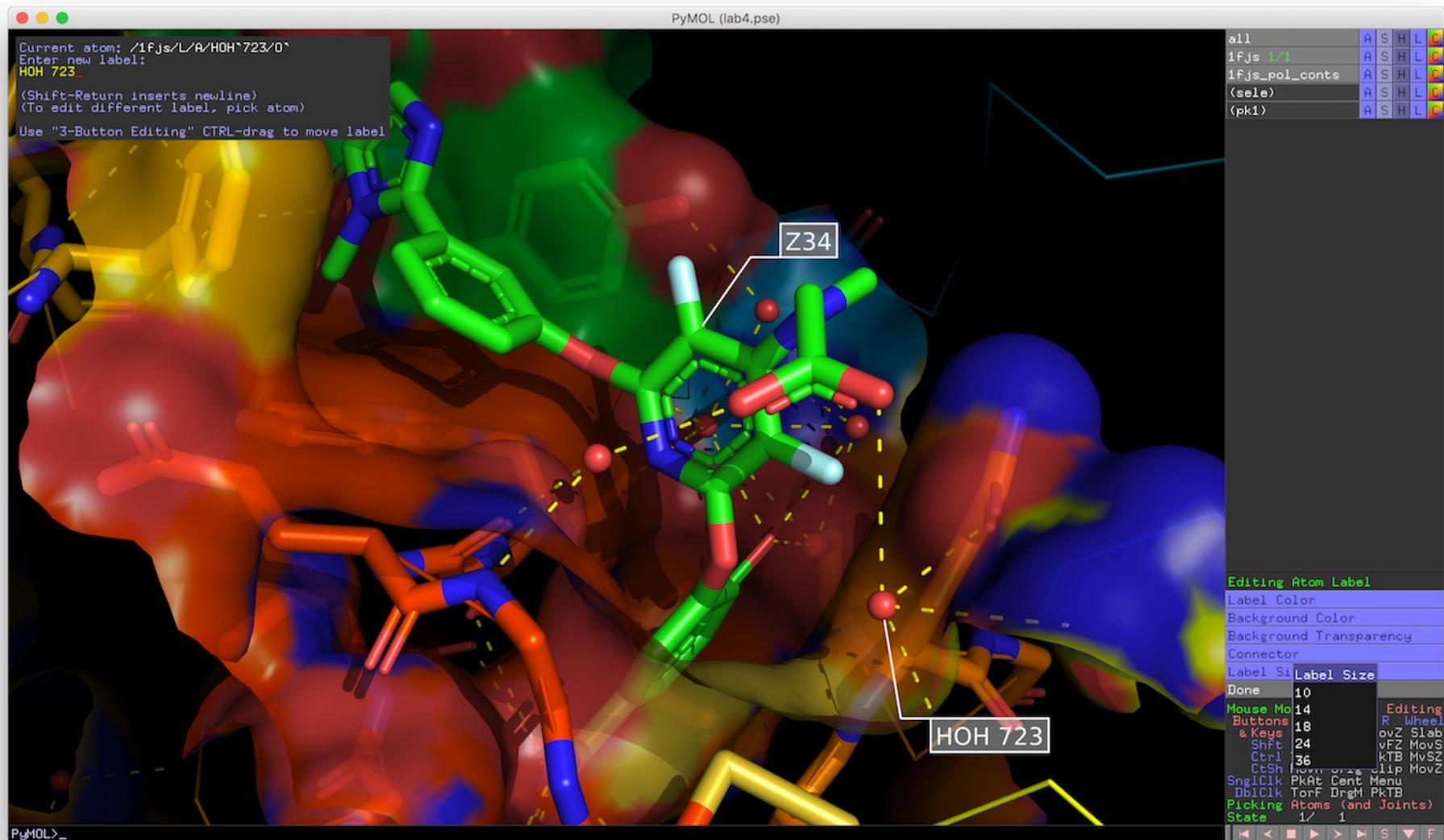
News

Publications ▾

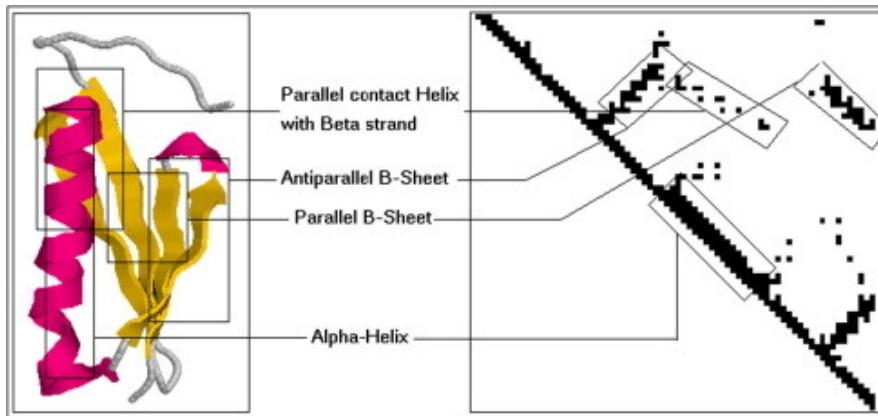


Vote Now for the Viewer's Choice Award

Pymol: Protein and Molecular Visualization



Protein Contact Map

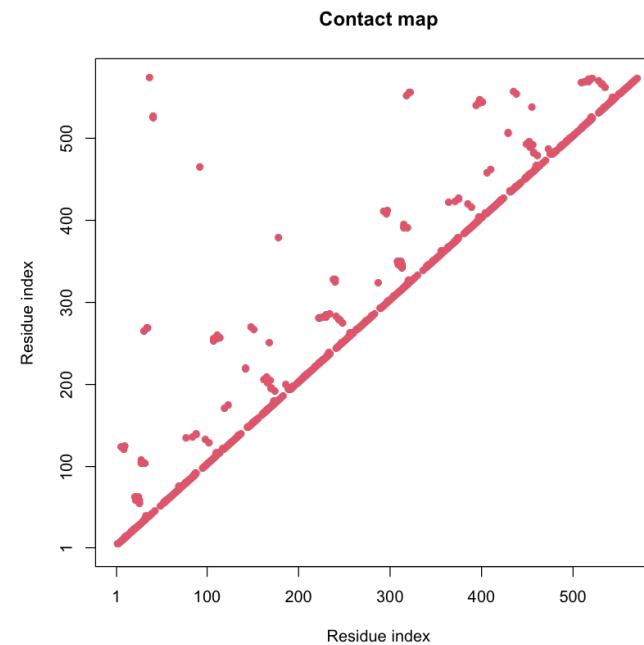


```
# Instal and load the library
install.packages("bio3d")
library("bio3d")

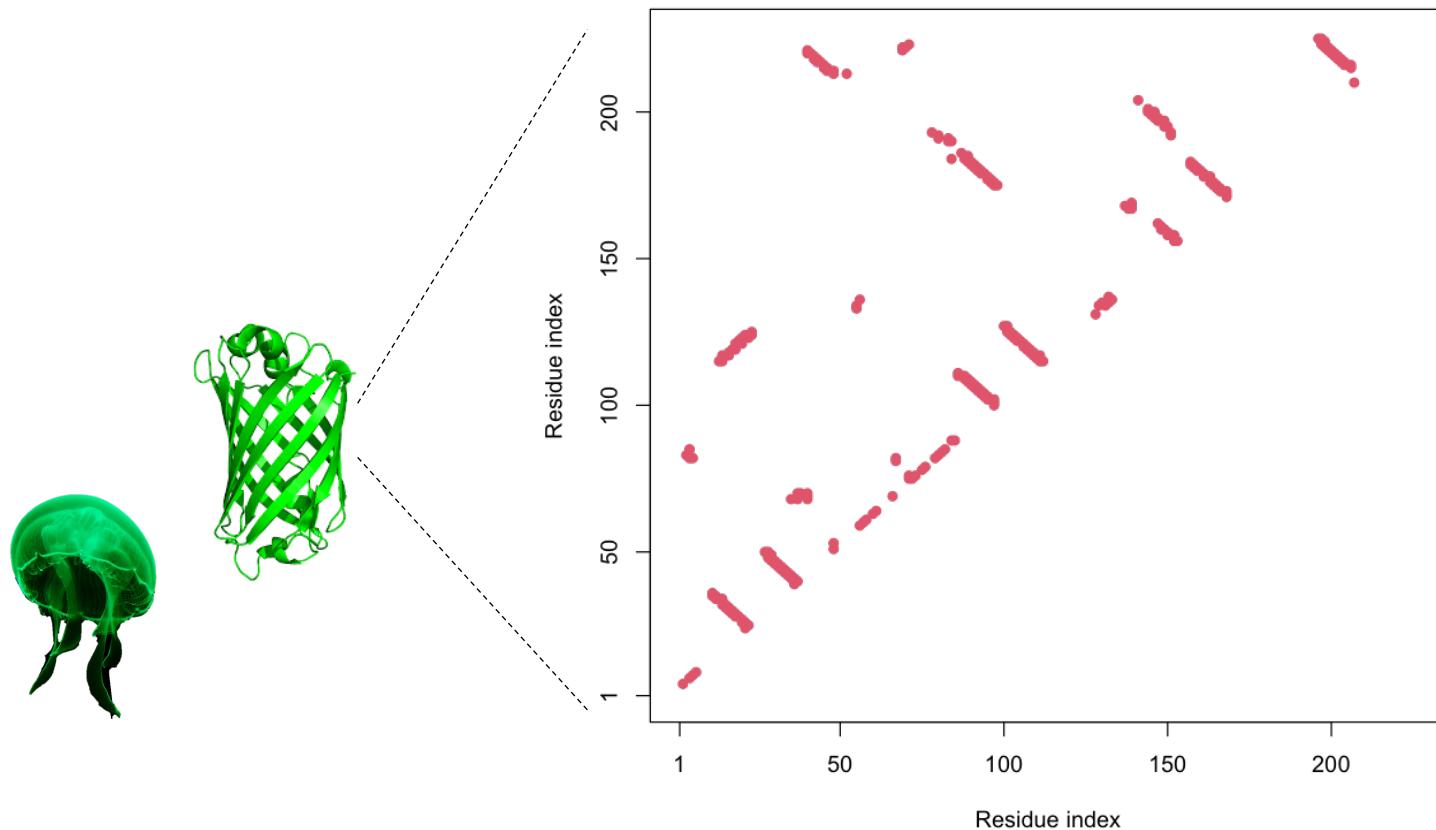
#Get the pdb file and load it
get.pdb("1A01")
pdb <- read.pdb("1A01.pdb")

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

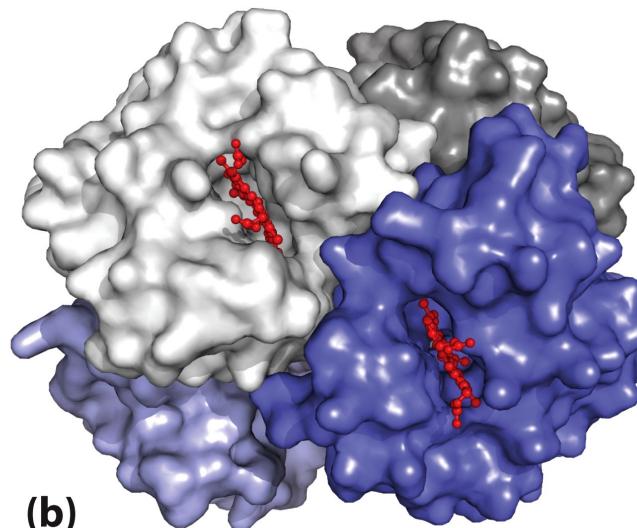
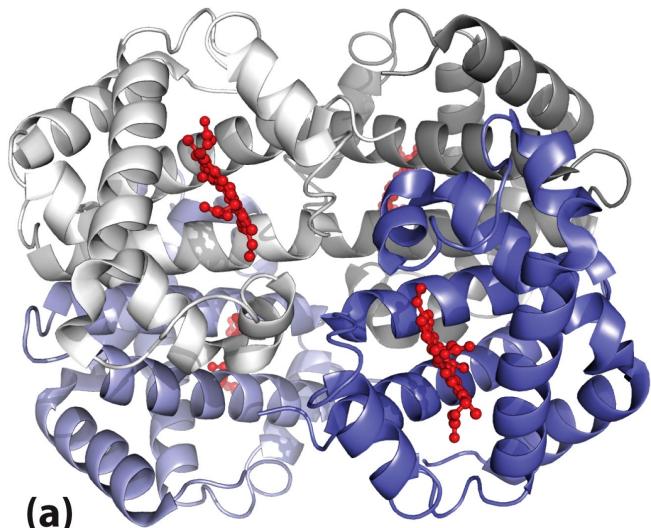
## Reference contact map
ref.cont <- cmap( pdb$xyz[inds$xyz], dcut=6, scut=3 )
plot.cmap(ref.cont)
```



Contact Map for Green Fluorescence Protein

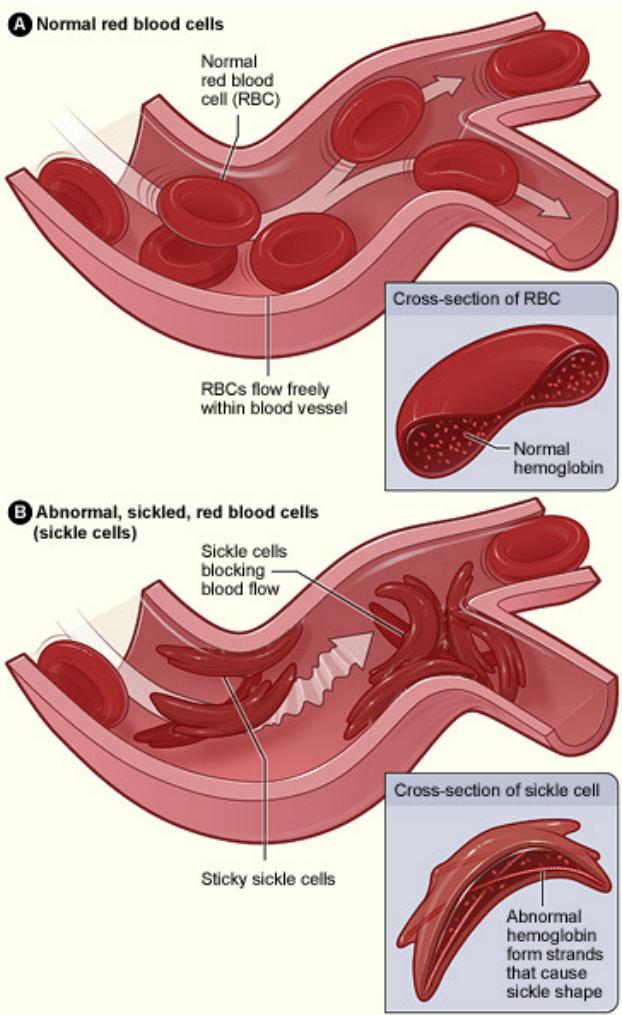


Quaternary Structure of Hemoglobin



- 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 β chains (146 residues each).
- The subunits of hemoglobin are arranged in symmetric pairs, each pair having one α and one β subunit. Hemoglobin can therefore be described either as a tetramer or as a dimer of $\alpha\beta$ protomers.

Sickle Cell Anemia



Sickle Cell Anemia

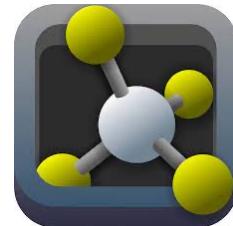
Sickle cell anemia is caused by a specific mutation. The mutation causes a change in a single amino acid in the beta-globin chain of hemoglobin, where the amino acid **glutamic acid** is replaced by **valine** (E6V)

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

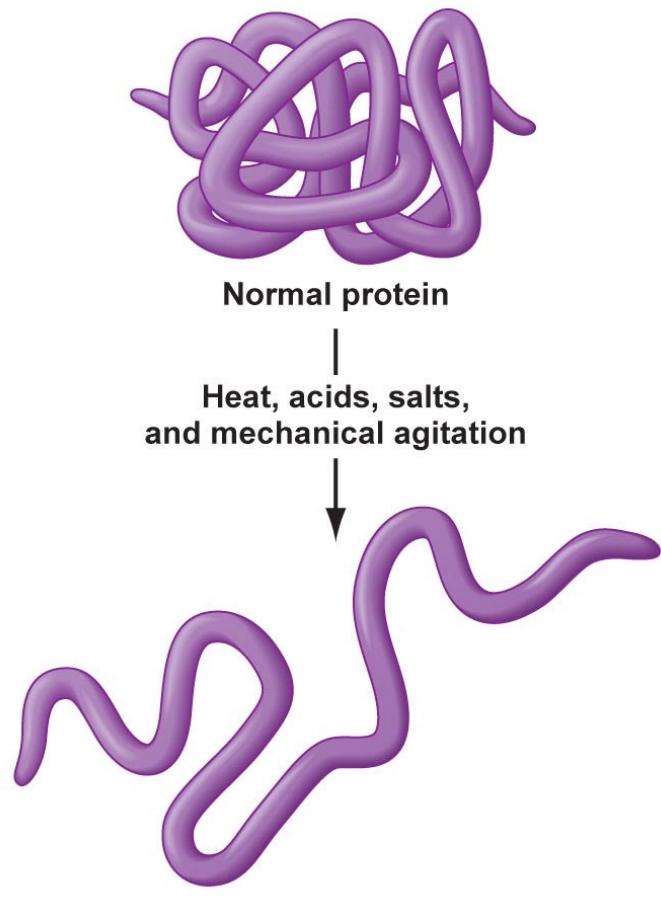
1. Check Hemoglobin B on Uniprot
2. Get the PDB structure and open in Pymol (or *fetch* command)
3. Investigate it
4. How can E6V mutation cause the disease?



Denaturing a Protein

- Desaturating agent/factors:

- Heat
- Acids
- Bases
- Salts
- Mechanical agitation
- Mutations
(interference with folding)

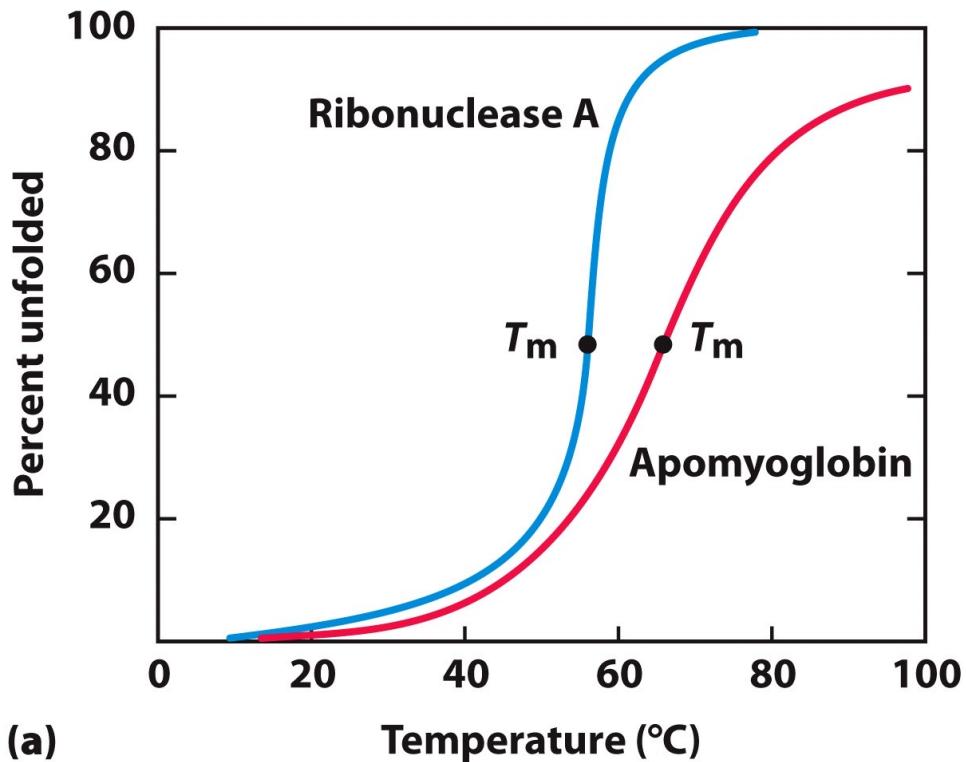


Denatured protein

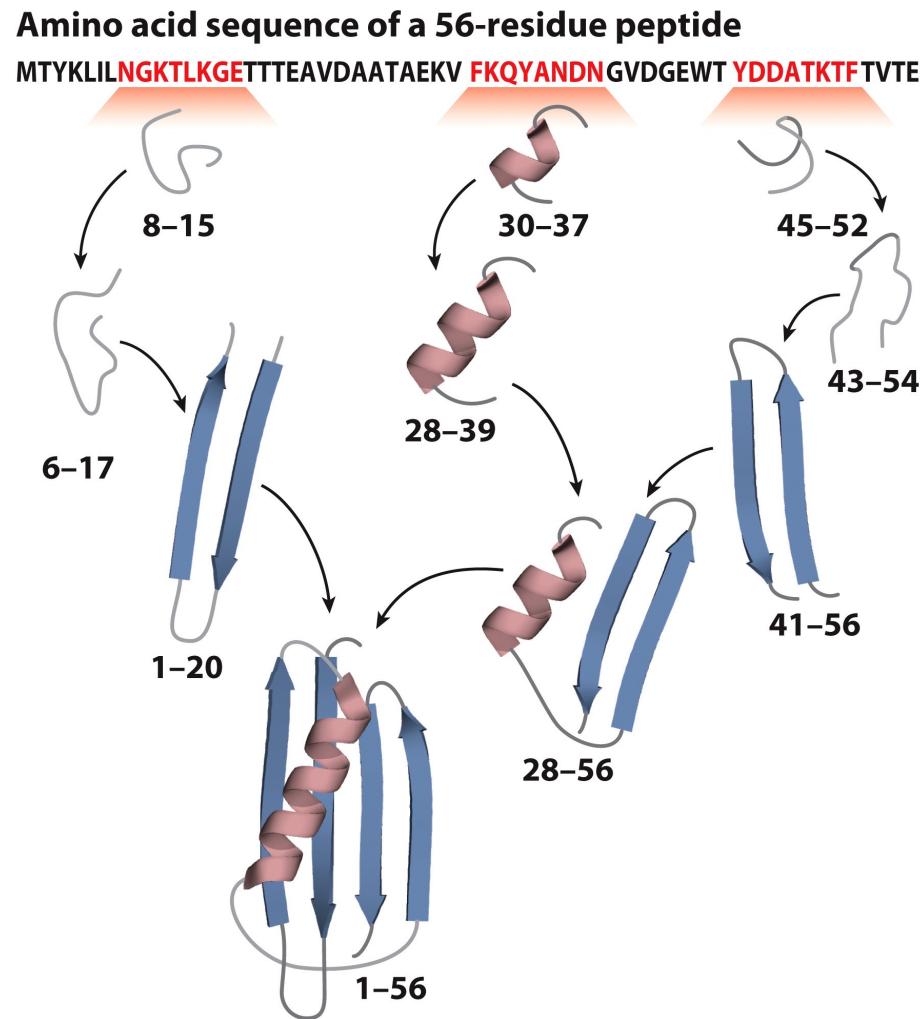
© 2010 Pearson Education, Inc.

Figure
6.5

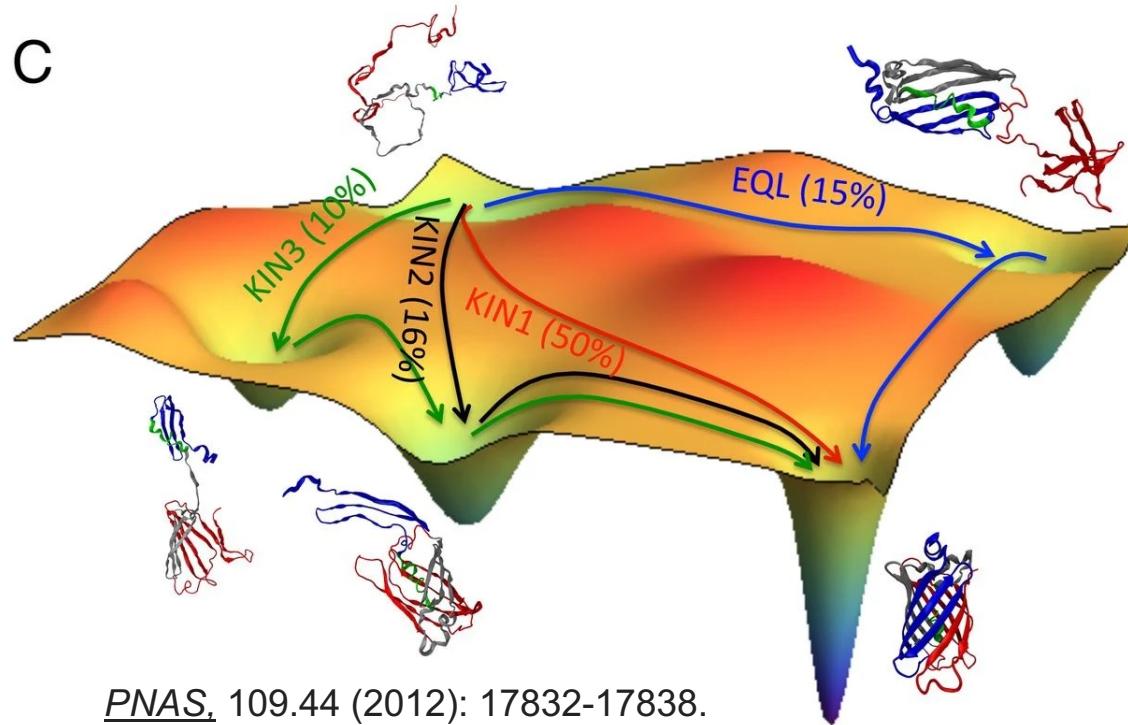
Protein Denaturation



Overview of Protein Folding



Protein Folding Intermediates



Mutational Effects on Protein Stability

(FireProt database)

FIREPROT DB

v1.1 Mutational data for protein stability

Search Enter search phrase... ADVANCED

Home **Browse database** **Datasets** **Use cases** **Help** **Acknowledgement**

ABOUT

FireProtDB is a comprehensive, manually curated database of the protein stability data for single-point mutants.

Proteins find their use in numerous biomedical and biotechnological applications. Naturally occurring proteins usually cannot withstand harsh industrial environments since they have evolved to function under mild conditions. Increasing protein stability is one of the key determinants of protein applicability. The predictive power of the current computational tools is compromised by the limited experimental data that would allow a rigorous training and testing.

This database combines the published datasets from ProTherm and ProtaBank, the data extracted from the recent literature, and the measurements collected in our laboratory. The annotations were obtained from VariBench and HotSpot Wizard. The graphical user interface is designed to facilitate both types of the expected use: (i) the interactive explorations of individual entries on the level of a protein or a mutation and (ii) the construction of highly customized, machine learning-friendly datasets using advanced searching and filtering.

REFERENCES

Stourac, J., Dubrava, J., Musil, M., Horackova, J., Damborsky, J., Mazurenko, S., Bednar, D. 2020: FireProt^{DB}: Database of Manually Curated Protein Stability Data. *Nucleic Acids Research* 49: D319-D324

PubMed OPEN ACCESS

STATISTICS

Number of visitors: 10364
Number of experiments: 15987
Curated: 10352
Uncurated: 5635
Number of mutations: 6713
Number of proteins: 242
Number of structures: 304
Last data update: Feb 8, 2022

DOWNLOADS

FireprotDB dump

CONTACT

Loschmidt Laboratories
• fireprot@sci.muni.cz
• <https://loschmidt.chemi.muni.cz>

OTHER TOOLS

Experiments by $\Delta\Delta G$

ΔΔG (kcal/mol)	Entries
-12	~10
-12,-10	~10
-10,-8	~10
-8,-6	~10
-6,-4	~10
-4,-2	~3500
-2,0	~5800
0,2	~2200
2,4	~500
4,6	~100
6,8	~10
> 8	~10

Experiments by ΔT_m

ΔTm (°C)	Entries
≤ -15	~150
(-15,-10]	~250
(-10,-5]	~550
(-5,0]	~1450
(0,5]	~1050
(5,10]	~100
> 15	~10

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

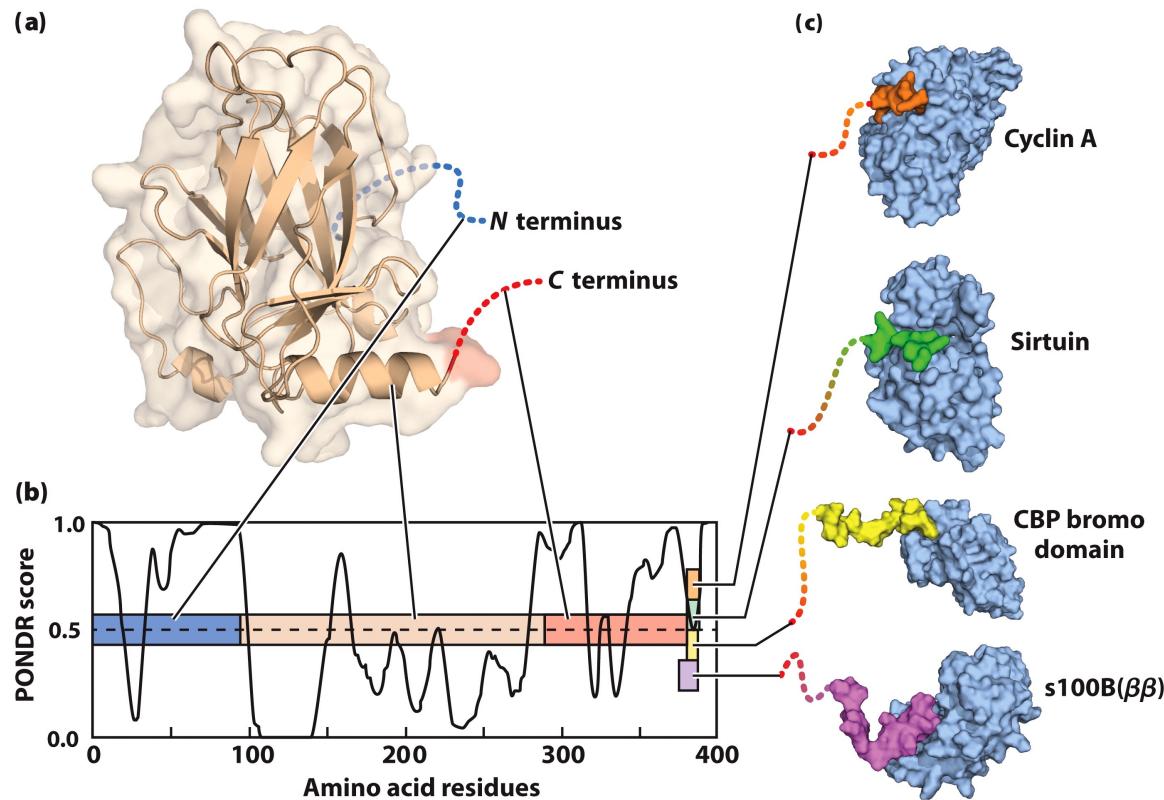
BLAST Align Map IDs ↴ Download ⌂ Add View: Cards ○ Table ⚡ Customize columns ☰ Share ▾ 6 rows selected out of 25

Entry ▾	Entry Name ▾	Protein Names ▾	Gene Names ▾	Organism ▾	Length ▾
<input checked="" type="checkbox"/> P68871	¤ HBB_HUMAN	Hemoglobin subunit beta[...]	HBB	Homo sapiens (Human)	147 AA
<input checked="" type="checkbox"/> P02070	¤ HBB_BOVIN	Hemoglobin subunit beta[...]	HBB	Bos taurus (Bovine)	145 AA
<input checked="" type="checkbox"/> P68873	¤ HBB_PANTR	Hemoglobin subunit beta[...]	HBB	Pan troglodytes (Chimpanzee)	147 AA
<input checked="" type="checkbox"/> P02075	¤ HBB_SHEEP	Hemoglobin subunit beta[...]	HBB	Ovis aries (Sheep)	145 AA
<input checked="" type="checkbox"/> P02112	¤ HBB_CHICK	Hemoglobin subunit beta[...]	HBB	Gallus gallus (Chicken)	147 AA
<input checked="" type="checkbox"/> 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



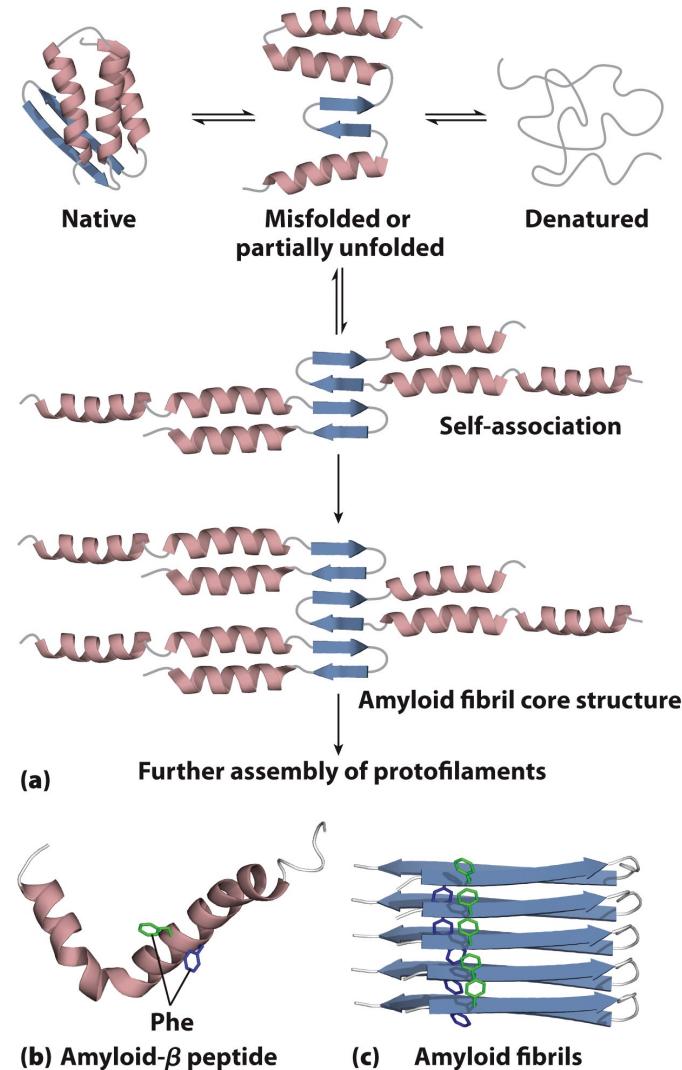
TP53 plays a crucial role in the control of cell division. It features both structured and unstructured segments. An unstructured region at the carboxyl terminus interacts with at least four different binding partners and assumes a different structure in each of the complexes.

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 β conformation. In Alzheimer disease, proteolytic cleavage of a neuronal membrane protein (amyloid- β precursor protein, APP) produces an α -helical membrane-spanning peptide (amyloid- β peptide) that converts to the β conformation and aggregates into amyloid fibrils.

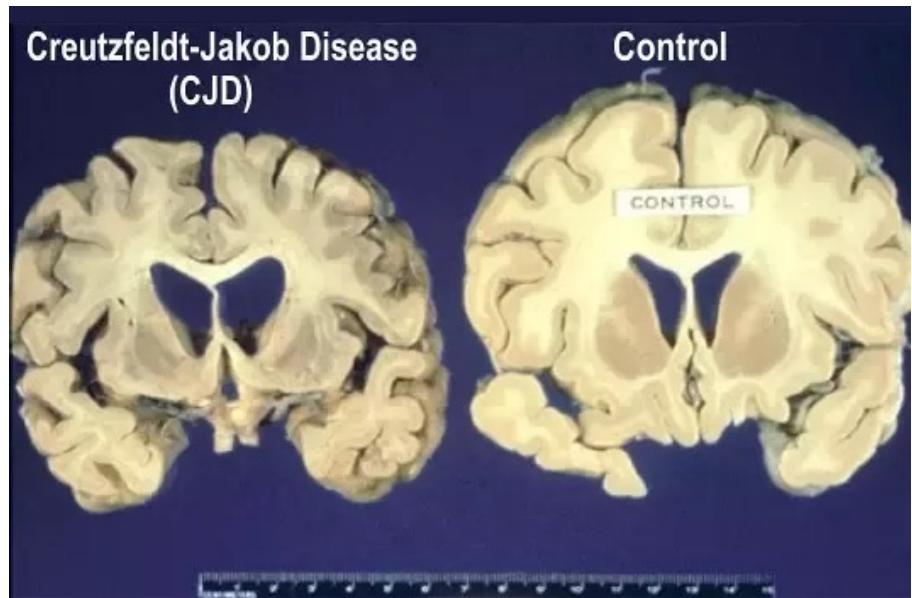
The extracellular deposition of amyloid fibrils is associated with plaque formation and ultimately death of the nearby neurons.



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.

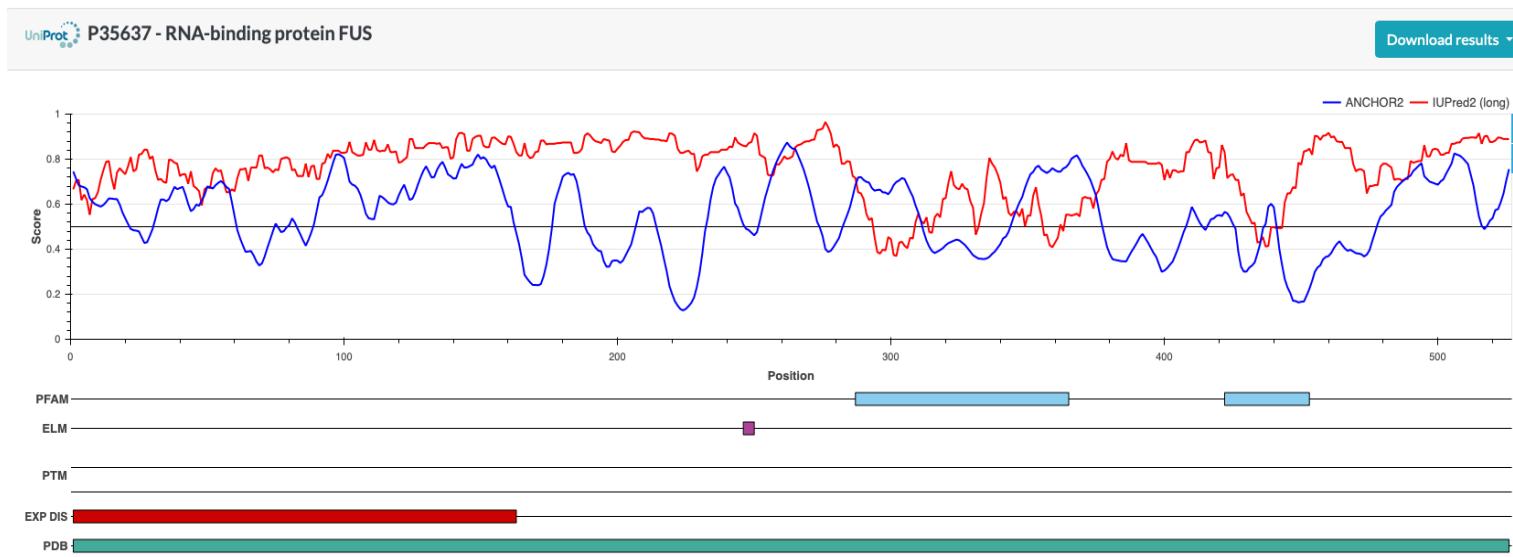
Susan Lindquist (1949-2016)

Class Mini Project 2 (competitive)

PART 1

Start from the protein FUS (Uniprot ID: [P35637](#))

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



Class Mini Project 2 (competitive)

PART 1

Q2: Mutations in this protein can cause what diseases?

The screenshot shows the ClinVar homepage. At the top, there's a banner with the NIH logo and text: "An official website of the United States government [Here's how you know](#)". Below the banner, the National Library of Medicine logo and "National Center for Biotechnology Information" text are visible, along with a "Log in" button. A "ClinVar home" link is also present. The main navigation bar includes "ClinVar", "Search" (with dropdown options), "Advanced", "Help", and a menu bar with "Home", "About", "Access", "Help", "Submit", "Statistics", and "FTP". On the left, a sequence of DNA is displayed: ACTGATGGTATGGGCCAAGAGATATATCT CAGGTACGGCTGTCACTCACCTAGACCTCAC CAGGGCTGGGCATAAAAGTCAGGGCAGAGC CCATGGTGCATCTGACTCCTGAGGAGAAGT GCAGGTTGGTATCAAGGTTACAAGACAGGT GGCACTGACTCTCTGCCTATTGGTCTAT. To the right of the sequence, the word "ClinVar" is prominently displayed in a large white font. Below it, a dark text box contains the text: "ClinVar aggregates information about genomic variation and its relationship to human health." At the bottom, three columns provide links to "Using ClinVar", "Tools", and "Related Sites".

Using ClinVar	Tools	Related Sites
About ClinVar	ACMG Recommendations for Reporting of Incidental Findings	ClinGen
Data Dictionary	ClinVar Submission Portal	GeneReviews ®
Downloads/FTP site	Submissions	GTR ®
FAQ	Variation Viewer	MedGen
Contact Us	Clinical Remapping - Between assemblies and RefSeqGenes	OMIM ®
Factsheet	RefSeqGene/LRG	Variation

Q3: Where in the protein?

Q4: Can we learn anything from the PDB database?

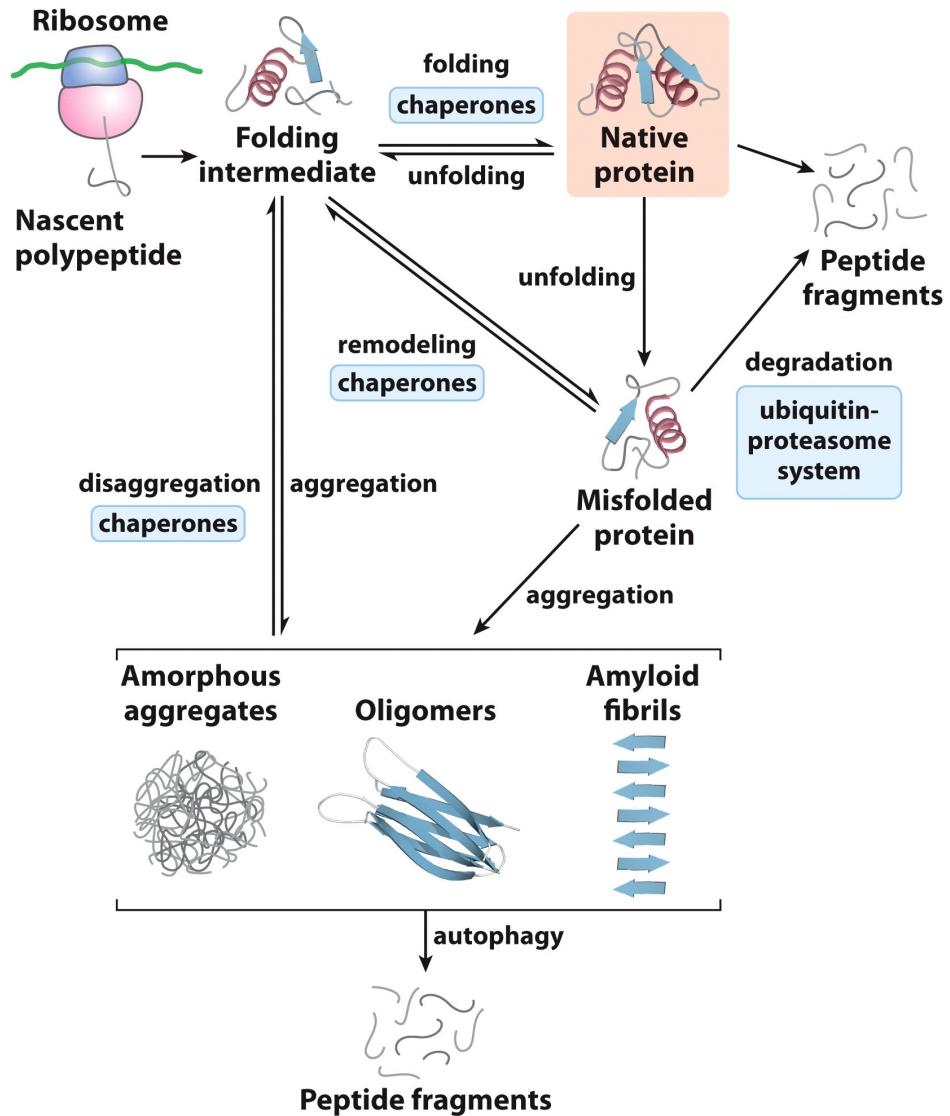
Proteostasis

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.

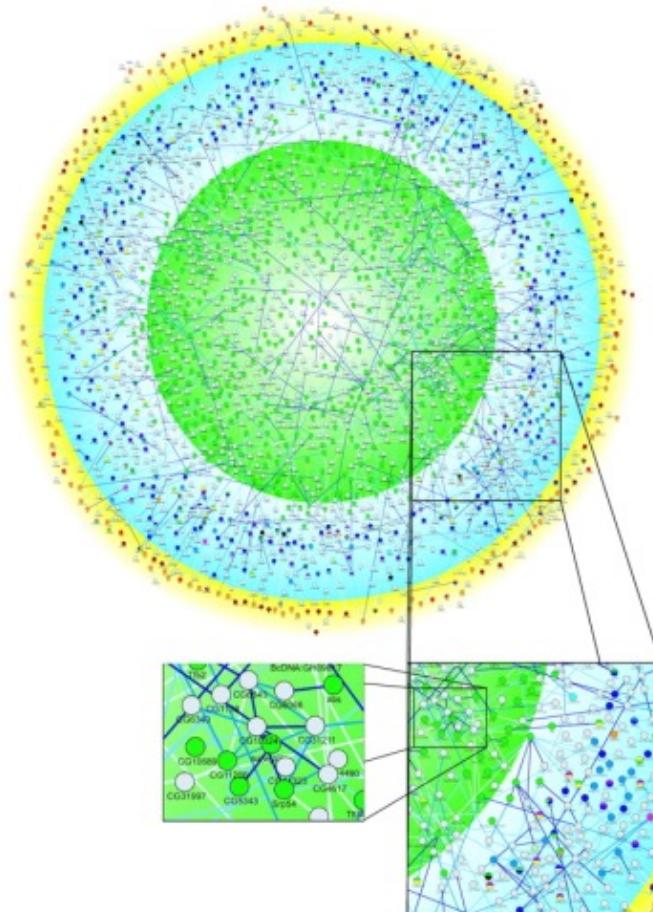
Second, multiple pathways contribute to protein folding, many of which involve the activity of complexes called chaperones. Chaperones (including chaperonins) also contribute to the refolding of proteins that are partially and transiently unfolded.

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*
Gint, Rothberg et al. Science 302, 1727-1736 (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

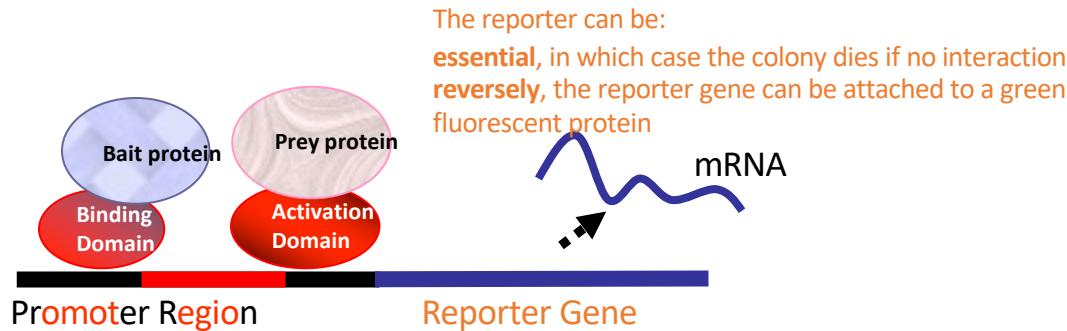
Yeast Two-Hybrid System

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

Protein structure determines interactions with other molecules

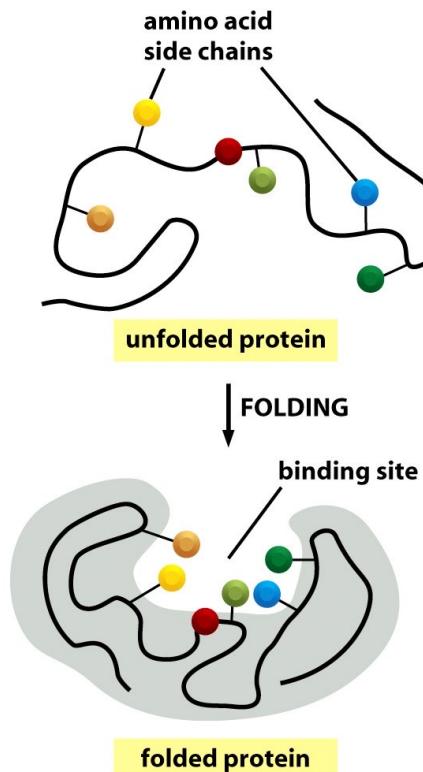
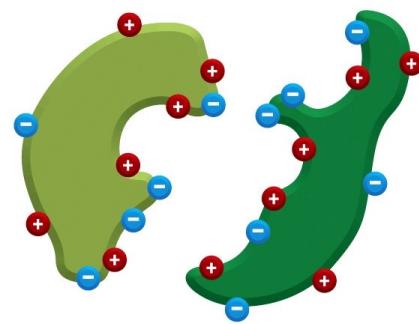


Figure 3-37a Molecular Biology of the Cell 5/e (© Garland Science 2008)

Protein structure affects protein-protein interactions



Protein structure affects protein-protein interactions

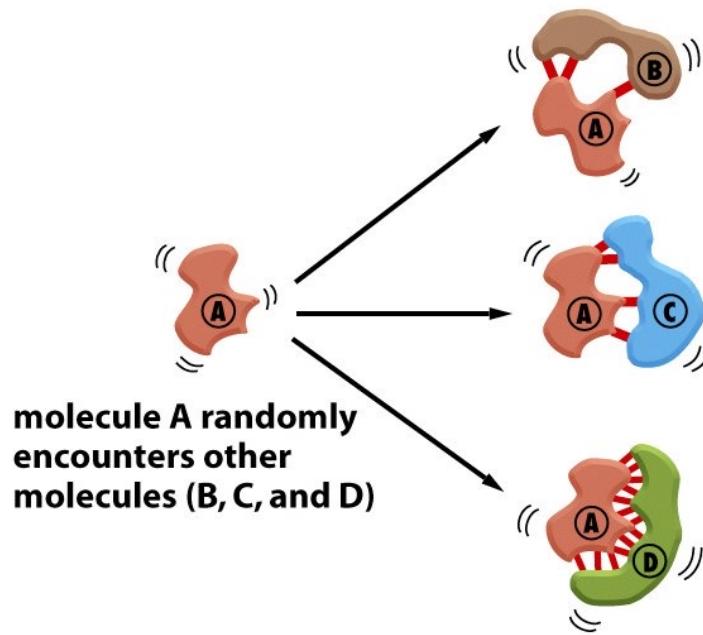
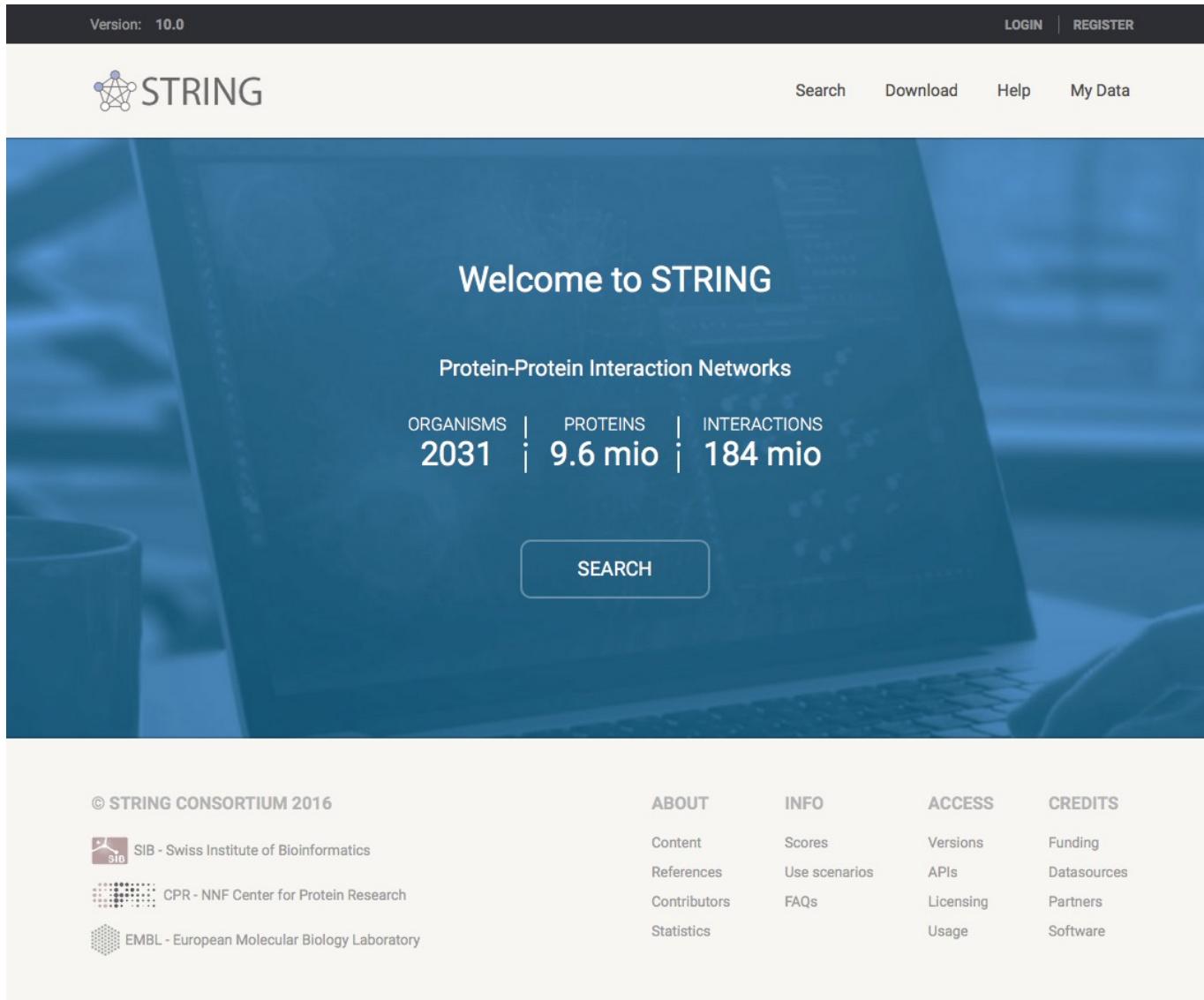


Figure 3-42 Molecular Biology of the Cell 5/e (© Garland Science 2008)

STRING database



The screenshot shows the STRING database homepage. At the top, a dark header bar displays "Version: 10.0" on the left and "LOGIN | REGISTER" on the right. Below the header is a navigation bar with the STRING logo (a network icon), "Search", "Download", "Help", and "My Data" links. The main content area has a blue-toned background image of a computer monitor displaying code. The title "Welcome to STRING" is centered at the top of this area. Below it, the subtitle "Protein-Protein Interaction Networks" is displayed. A summary of data statistics follows: "ORGANISMS 2031", "PROTEINS 9.6 mio", and "INTERACTIONS 184 mio". A large "SEARCH" button is positioned below these statistics. At the bottom of the page, there is a footer with copyright information and links to various partners and resources.

Version: 10.0

LOGIN | REGISTER

STRING

Search Download Help My Data

Welcome to STRING

Protein-Protein Interaction Networks

ORGANISMS | PROTEINS | INTERACTIONS
2031 | 9.6 mio | 184 mio

SEARCH

© STRING CONSORTIUM 2016

SIB - Swiss Institute of Bioinformatics

CPR - NNF Center for Protein Research

EMBL - European Molecular Biology Laboratory

ABOUT

Content

References

Contributors

Statistics

INFO

Scores

Use scenarios

FAQs

Usage

ACCESS

Versions

APIs

Licensing

Usage

CREDITS

Funding

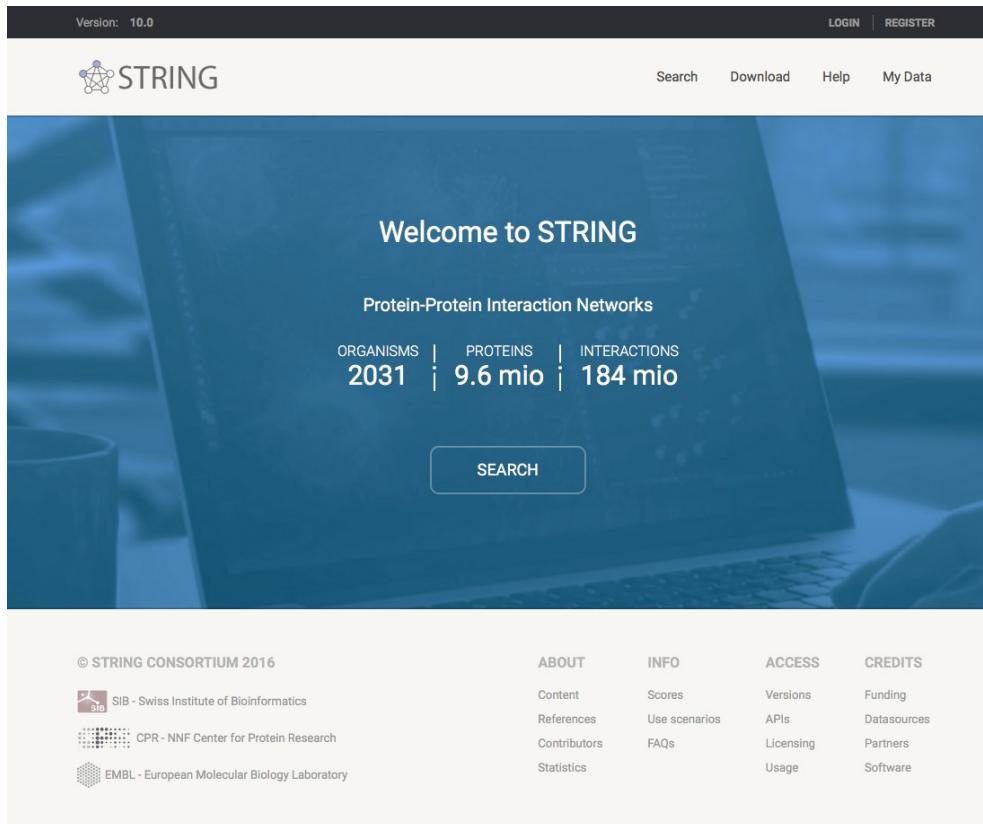
Datasources

Partners

Software

Class Mini Project 2 (competitive)

PART 2



The screenshot shows the STRING 10.0 homepage. At the top, it displays "Version: 10.0" and navigation links for "LOGIN" and "REGISTER". The main header "STRING" is accompanied by a logo of a molecular structure. Below the header, there are links for "Search", "Download", "Help", and "My Data". The central feature is a large banner with the text "Welcome to STRING" and "Protein-Protein Interaction Networks". It highlights statistics: "ORGANISMS | 2031", "PROTEINS | 9.6 mio", and "INTERACTIONS | 184 mio". A prominent "SEARCH" button is located in the center of the banner. The footer contains copyright information ("© STRING CONSORTIUM 2016") and logos for SIB, CPR-NNF, and EMBL, along with links to "ABOUT", "INFO", "ACCESS", and "CREDITS" sections.

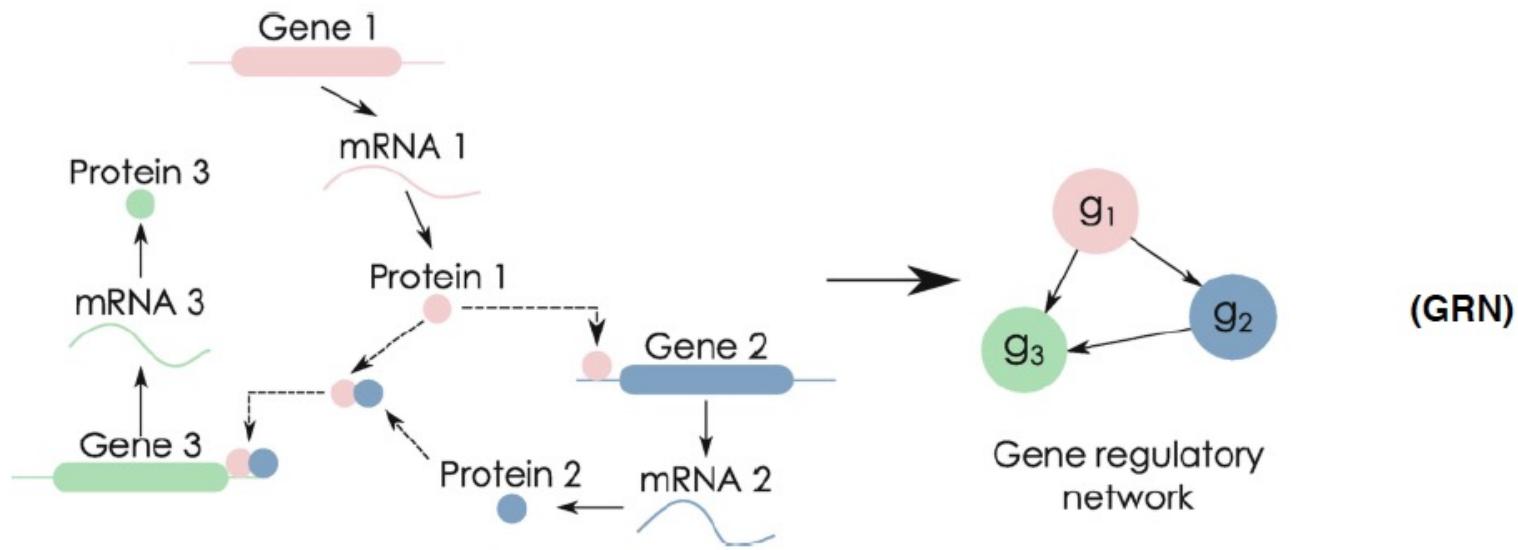
Q5: How protein-protein interactions can help us understand the role of FUS in neurological diseases?

Part 3: Protein Networks

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

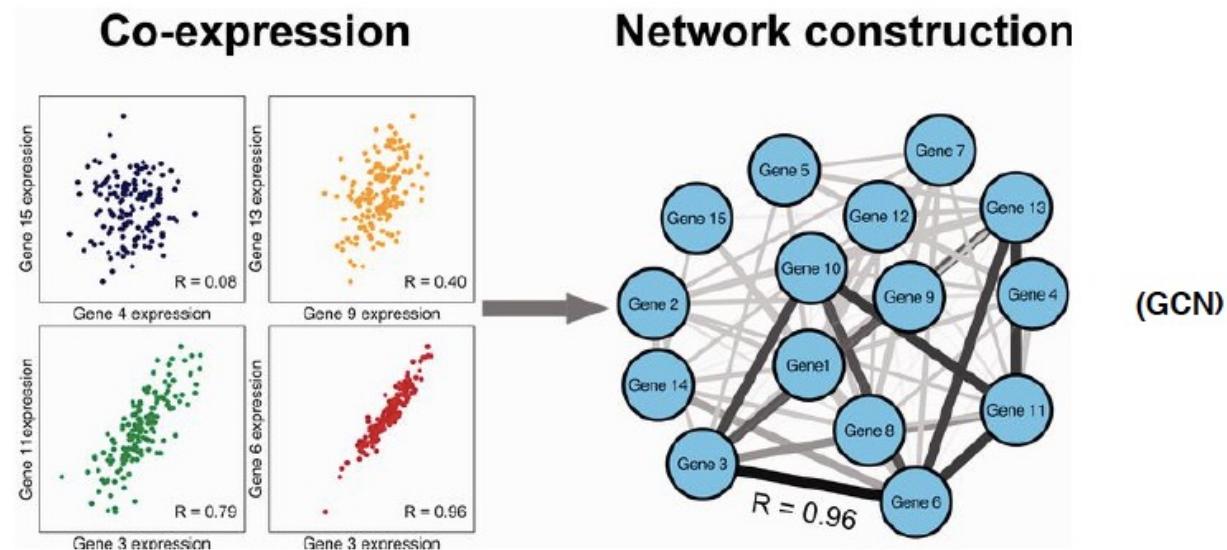


Gene Regulatory Networks



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



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

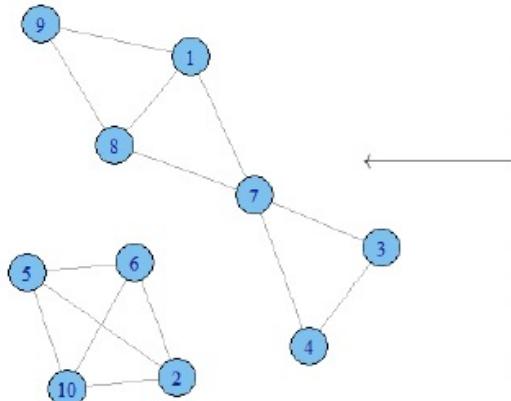


	S_1	S_2	S_3	
G_1	43.26	40.89	5.05	
G_2	166.6	41.87	136.65	
G_3	12.53	39.55	42.09	
G_4	28.77	191.92	236.56	
G_5	114.7	79.7	99.76	$ r(G_i, G_j) $
G_6	119.1	80.57	114.59	Pearson correlation
G_7	118.9	156.69	186.95	
G_8	3.76	2.48	136.78	
G_9	32.73	11.99	118.8	
G_{10}	17.46	56.11	21.41	

Gene expression values

	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

$|r(G_i, G_j)| \geq -0.8$
Significance threshold

Network adjacency matrix

Metabolic Networks

A set of chemical reactions that produces:

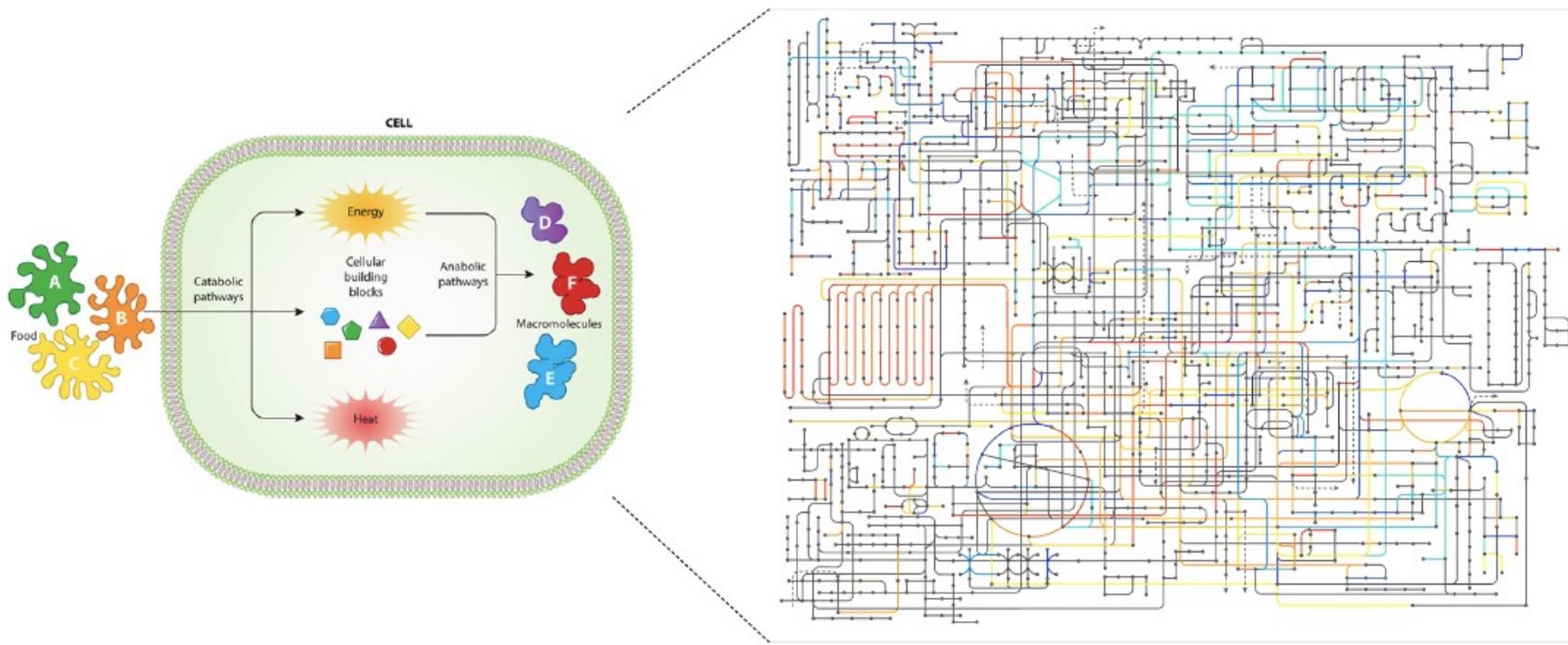
1) energy

(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



The screenshot shows the KEGG PATHWAY Database homepage. At the top left is the KEGG logo. Below it, the title "KEGG PATHWAY Database" and the subtitle "Wiring diagrams of molecular interactions, reactions and relations". A navigation bar below the title includes links for KEGG2, PATHWAY, BRITE, MODULE, KO, GENES, COMPOUND, DISEASE, and DRUG. Below the navigation bar is a search interface with fields for "Select prefix" (map or Organism), "Enter keywords", and buttons for "Go" and "Help". A link "[New pathway maps | Update history]" is also present. The main content area is titled "Pathway Maps" and contains a list of pathway categories: 1. Metabolism, 2. Genetic Information Processing, 3. Environmental Information Processing, 4. Cellular Processes, 5. Organismal Systems, 6. Human Diseases, and 7. Drug Development. A note at the bottom states: "KEGG PATHWAY is the reference database for pathway mapping in KEGG Mapper."

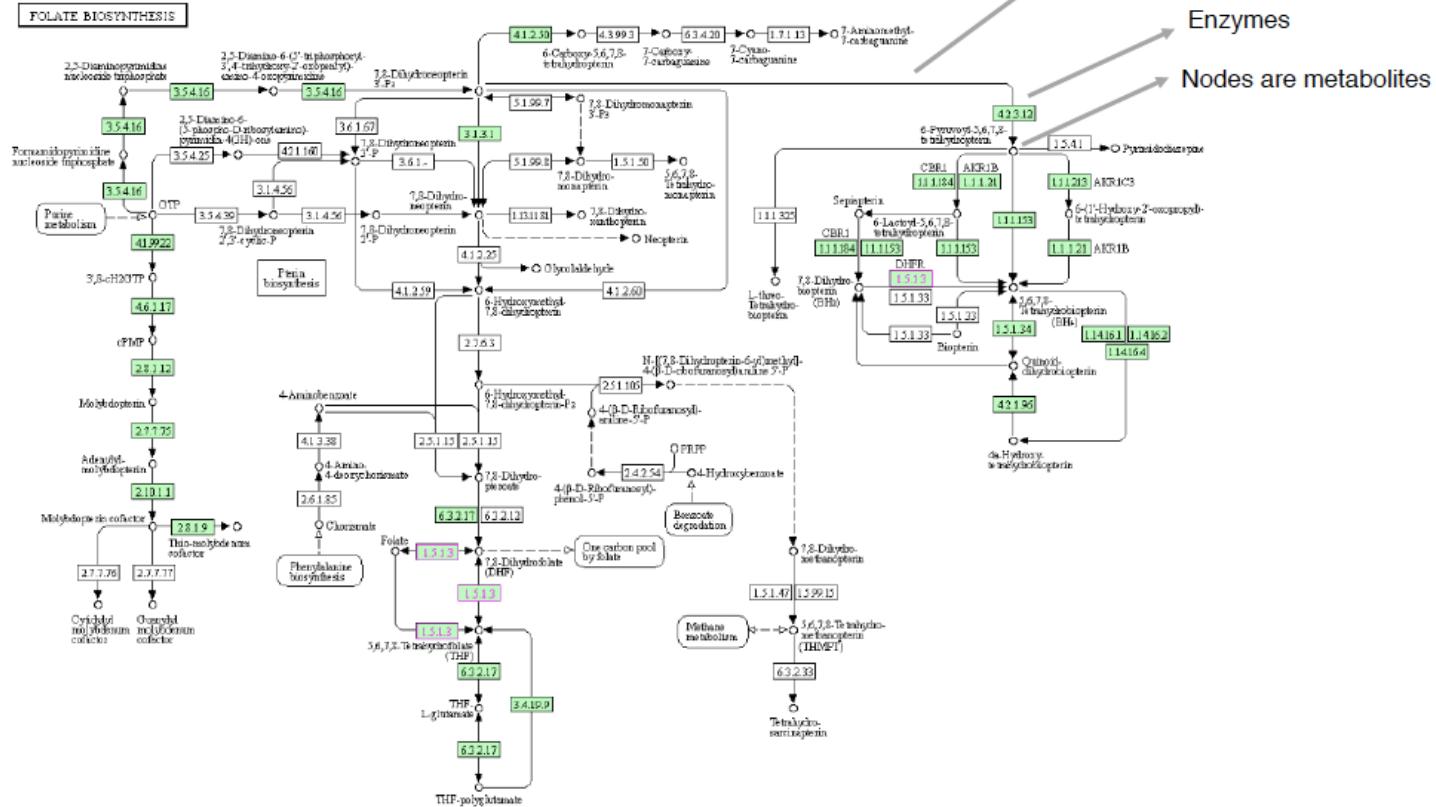
Find the folate biosynthetic pathway
in Human and *E.coli*

Folate pathway

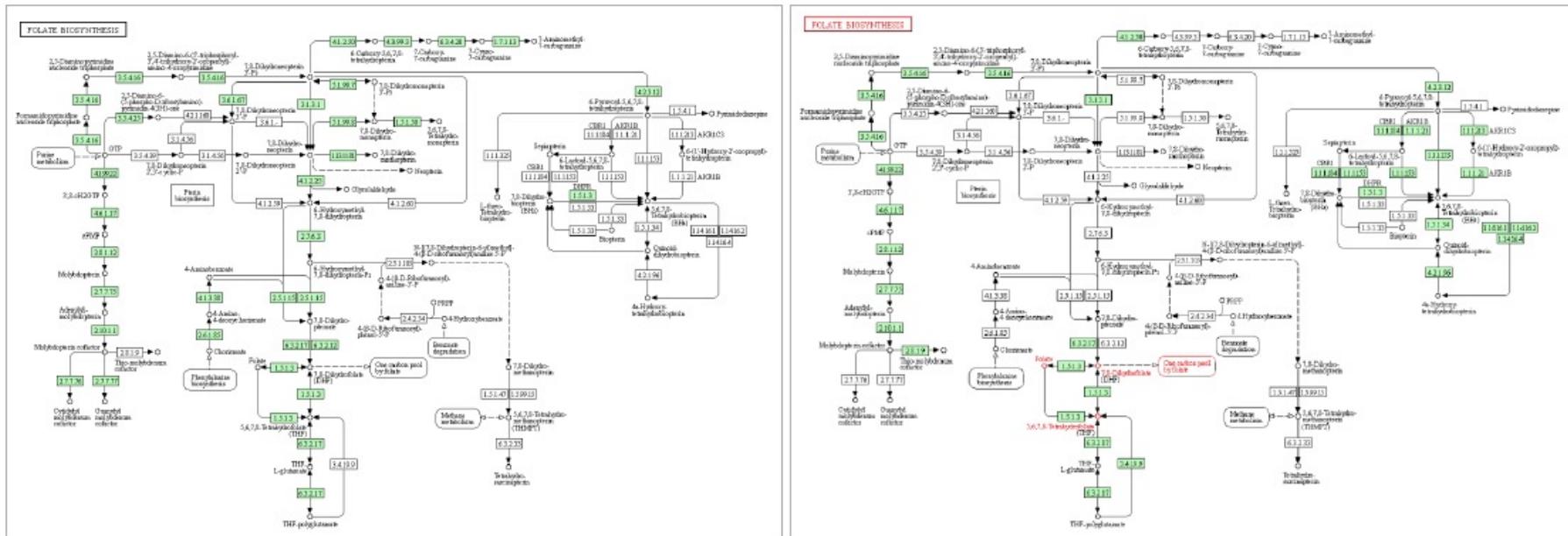
Arrows are reactions

Enzymes

Nodes are metabolites



Folate Pathway in Human and E.coli



Biological Networks and Complex Networks have similarities

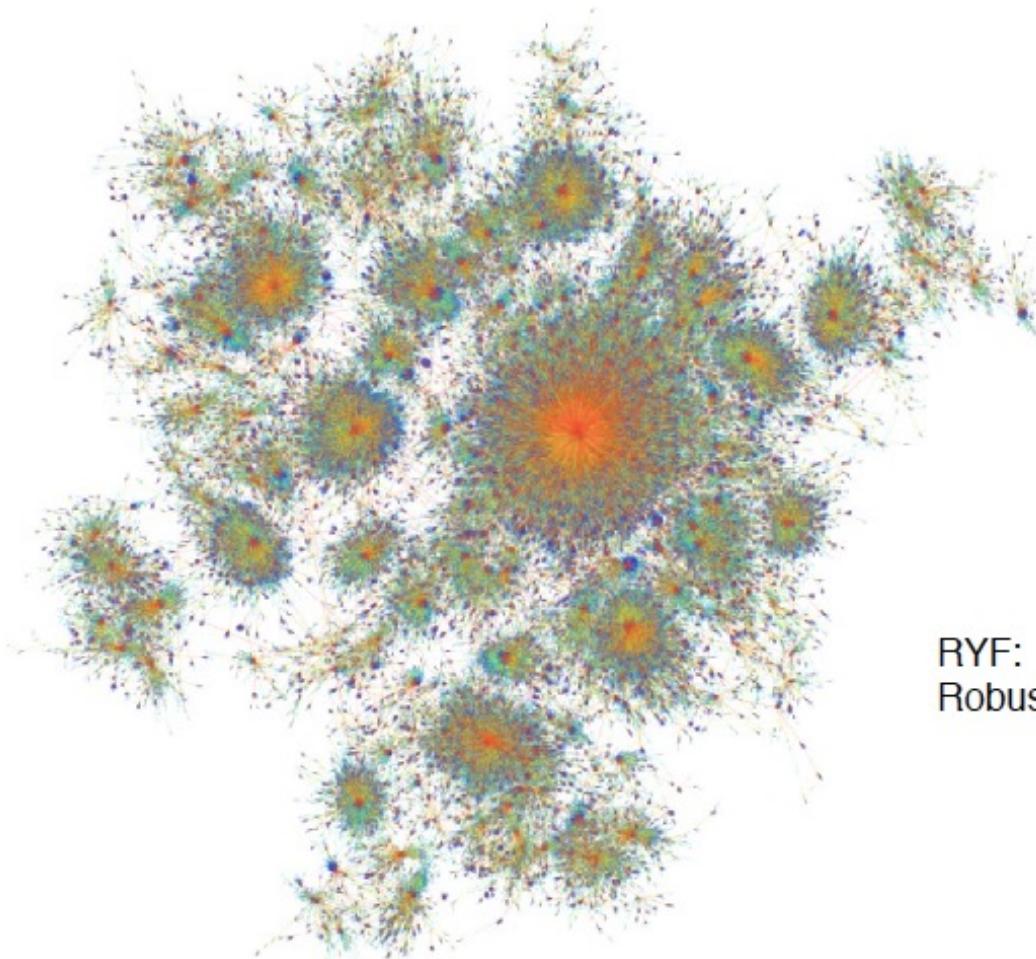


Metabolic map



New York subway map

Network Topology Matters!



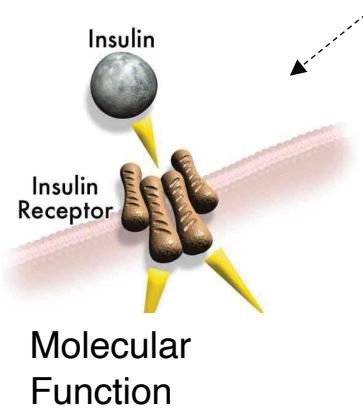
RYF:
Robust yet Fragile!

Gene Ontology Terms

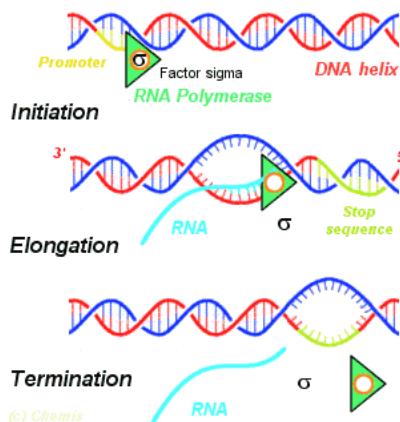
Experiments (Differential Expression, Interactions, ...)



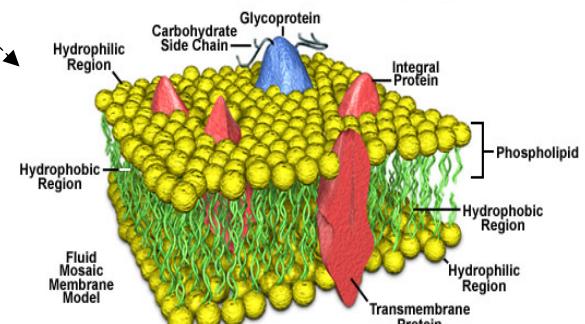
{Gene 1, Gene 2, Gene 3, ...}



Molecular Function



Biological Process



Cellular Component

+
(Transcription Factors,
Tissue Specificity, ...)

GO-term enrichment analysis

g:Profiler has been updated with new data from Ensembl.

Show more... Close

g:GOST
Functional profiling

g:Convert
Gene ID conversion

g:Orth
Orthology search

g:SNPense
SNP id to gene name

Query Upload query Upload bed file

Input is whitespace-separated list of genes ?

Run query random example mixed query example

Options

Organism: ?

Homo sapiens (Human)

Highlight driver terms in GO ?

Ordered query ?

Run as multiquery ?

Advanced options ▾

Data sources ▾

Bring your data (Custom GMT) ▾

g:GOST performs functional enrichment analysis, also known as over-representation analysis (ORA) or gene set enrichment analysis, on input gene list. It maps genes to known functional information sources and detects statistically significantly enriched terms. We regularly retrieve data from [Ensembl database](#) and fungi, plants or metazoa specific versions of [Ensembl Genomes](#), and parasite specific data from [WormBase](#)

ParaSite. In addition to Gene Ontology, we include pathways from KEGG Reactome and WikiPathways; miRNA targets from miRTarBase and regulatory motif matches from TRANSFAC; tissue specificity from Human Protein Atlas; protein complexes from CORUM and human disease phenotypes from Human Phenotype Ontology. g:GOST supports close to 500 organisms and accepts hundreds of identifier types.

Class Mini Project 3 (competitive)

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)

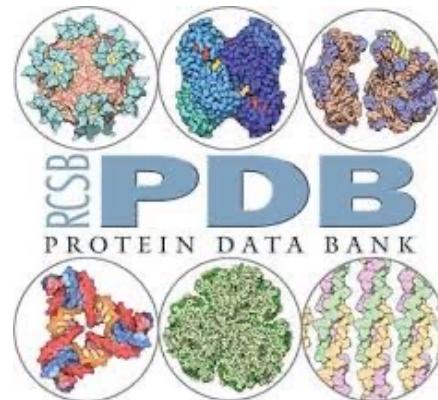
Group 2:
Tai-sachs Disease

An overview of databases and tools reviewed



IUPred2A

 FIREPROT^{DB}

The logo for FIREPROT consists of three colored cylinders (red, orange, and blue) stacked vertically to the left of the text "FIREPROT". A small "DB" superscript is positioned above the letter "T".

g:Profiler



On exam questions

Question 1: Protein aggregation-based diseases

Question 2: Databases of Proteins and their interactions