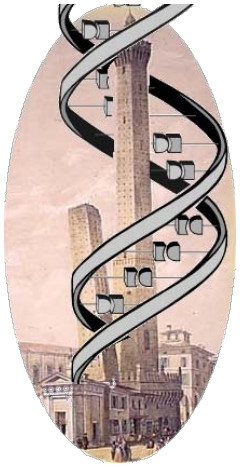
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



Proteins: at the edge of the genomic era

Rita Casadio

BIOCOMPUTING GROUP
University of Bologna, Italy

Bologna Computational Biology Network

WetLab | CompuLab | SME Small Medium Eterprises

AIRBBC

Syllabus:

1) Why proteomics

2) Relevance of proteins

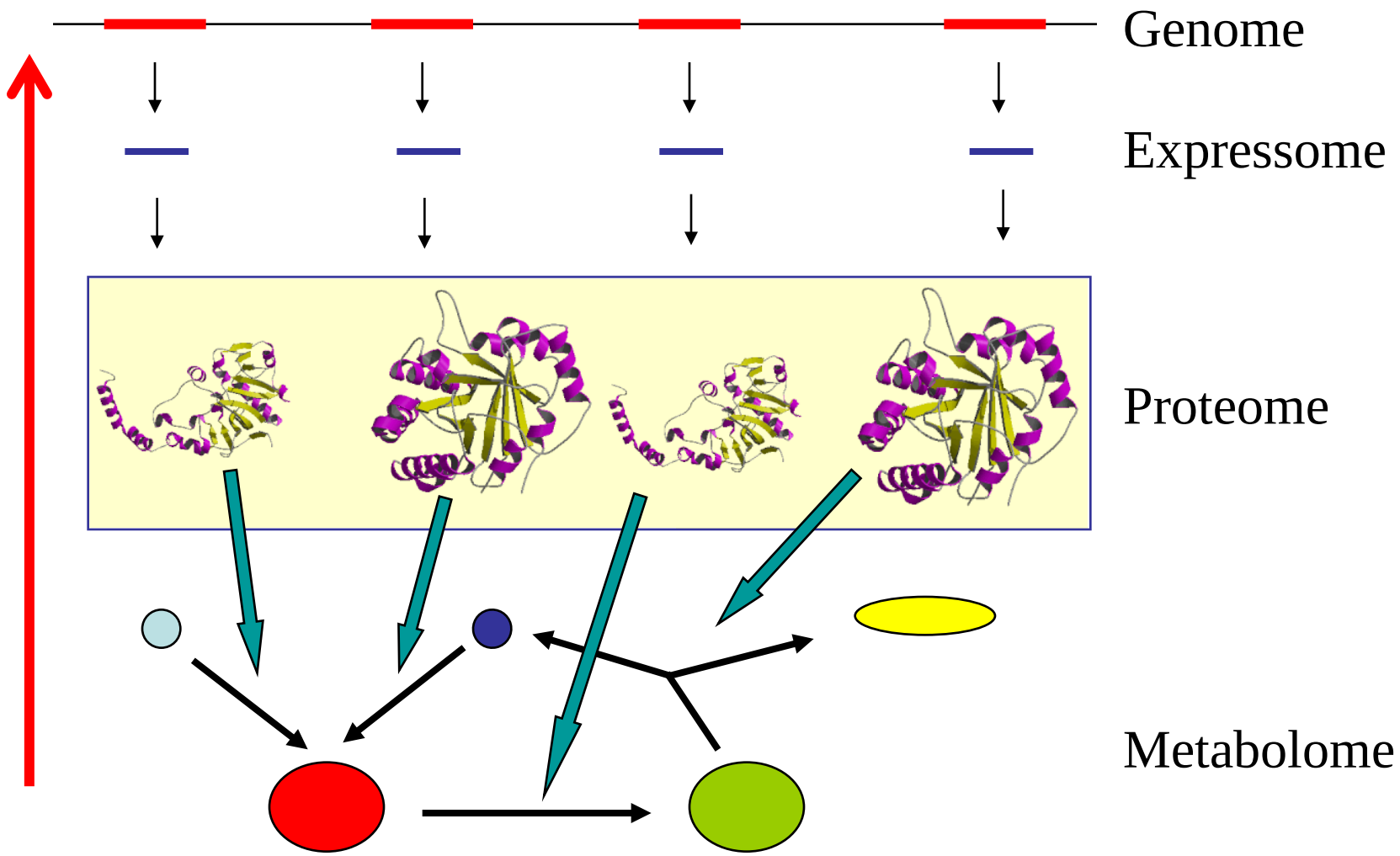
3) Protein structure: the golden standard of our information

4) The protein universe

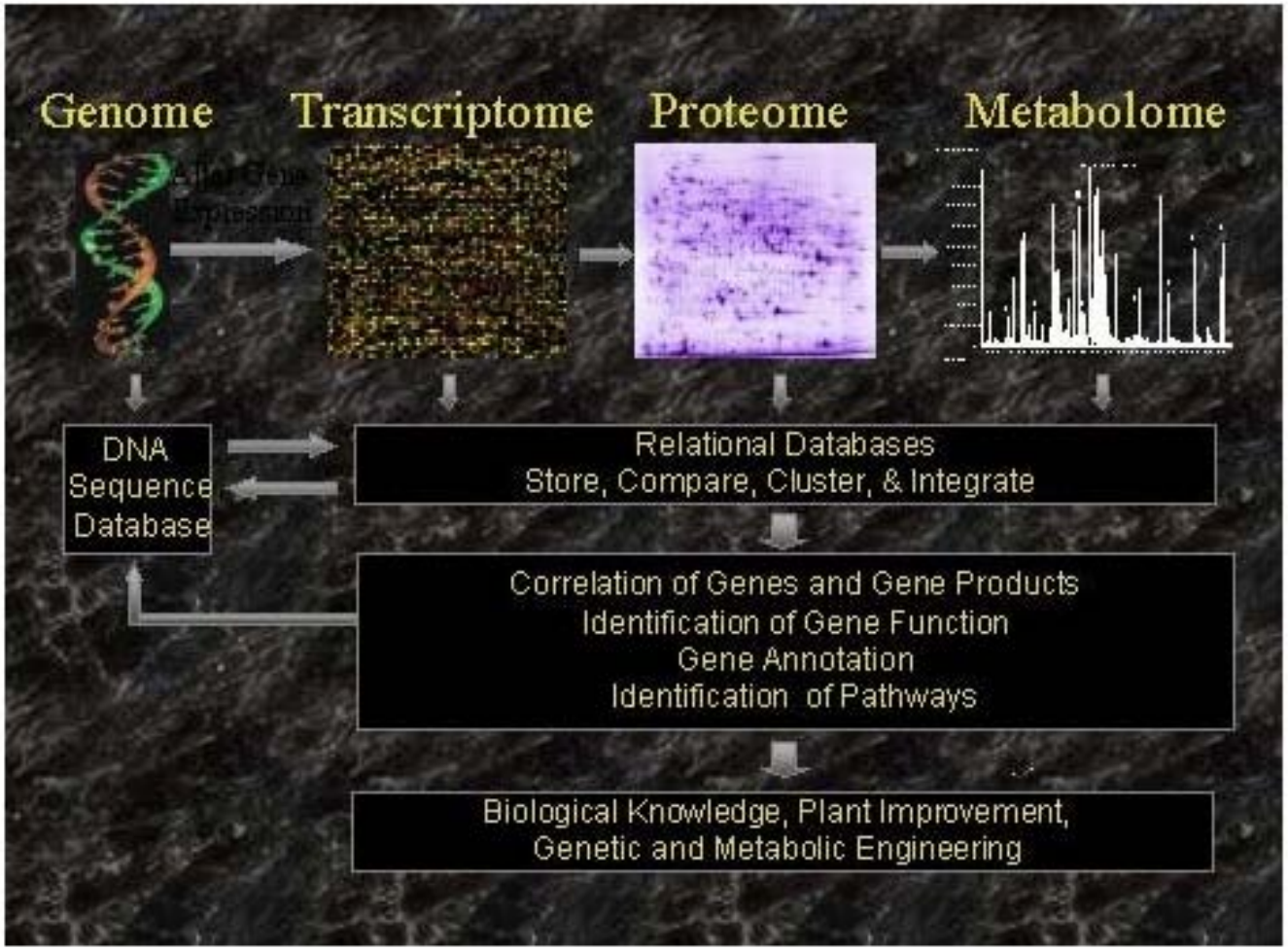
5) Open problems

Functional Genomics

From genes to functions and backward



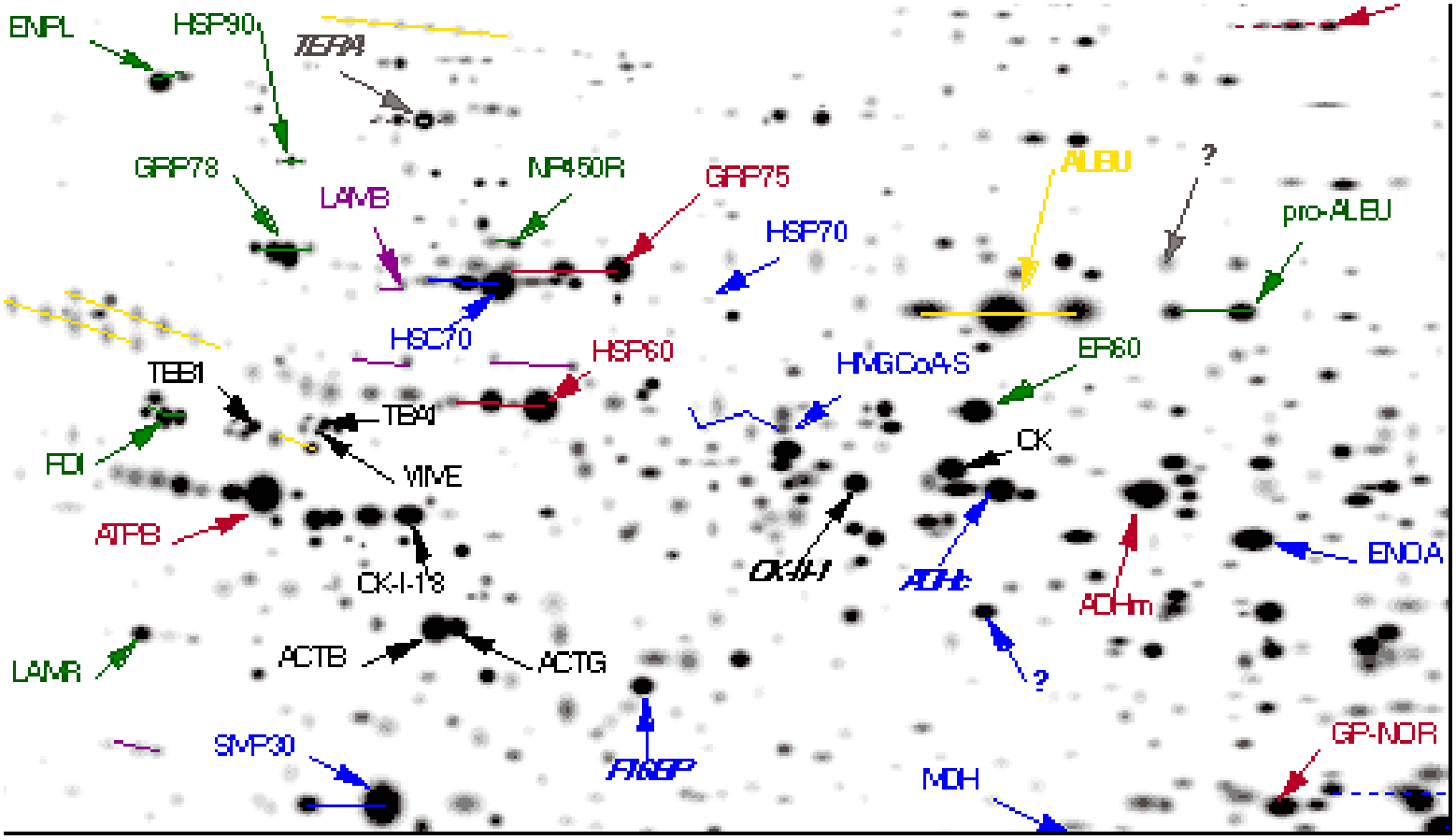
An integrated functional genomic approach (from Genomics to Proteomics)



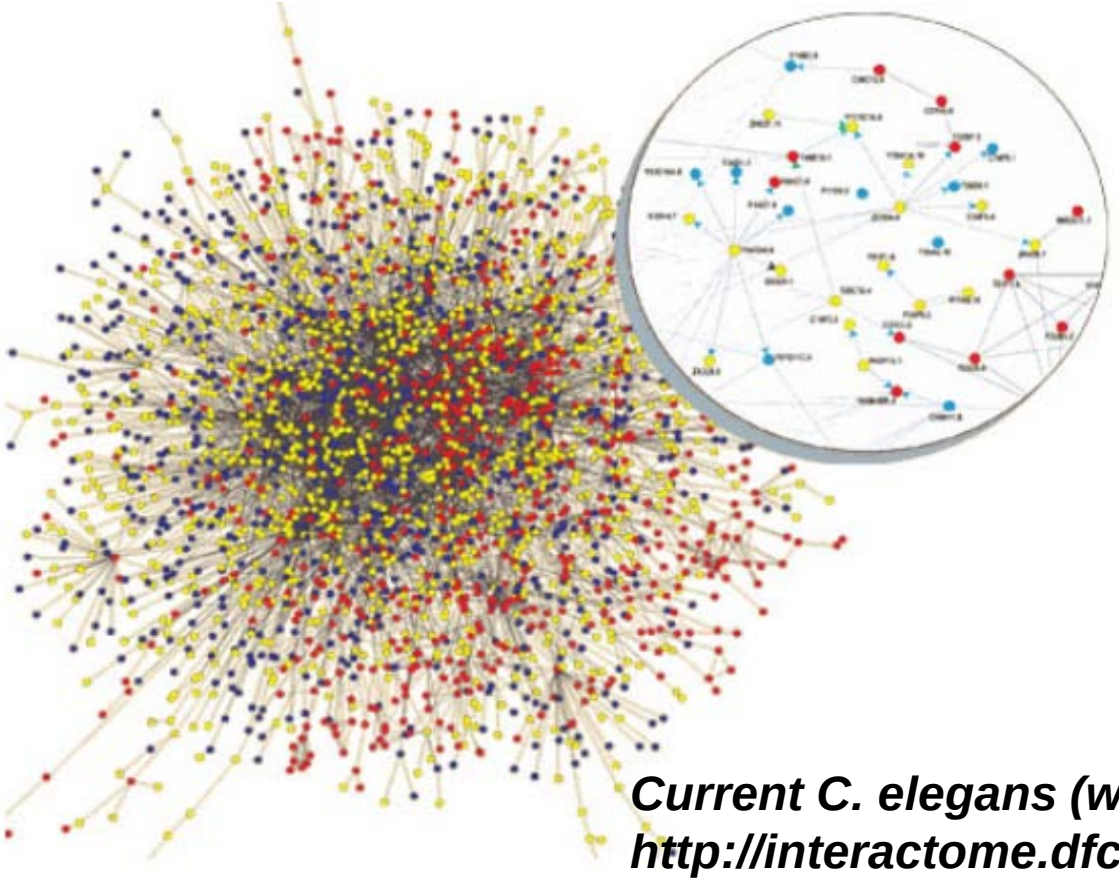
An integrated functional genomic approach monitors quantitative and qualitative differences in the transcriptome, proteome, and metabolome as a means to study gene function and cellular responses to external stimuli.

http://www.noble.org/PlantBio/MS/FG.html

2D gel (made by Large Scale Biology Corp.) of proteins from rat liver cells

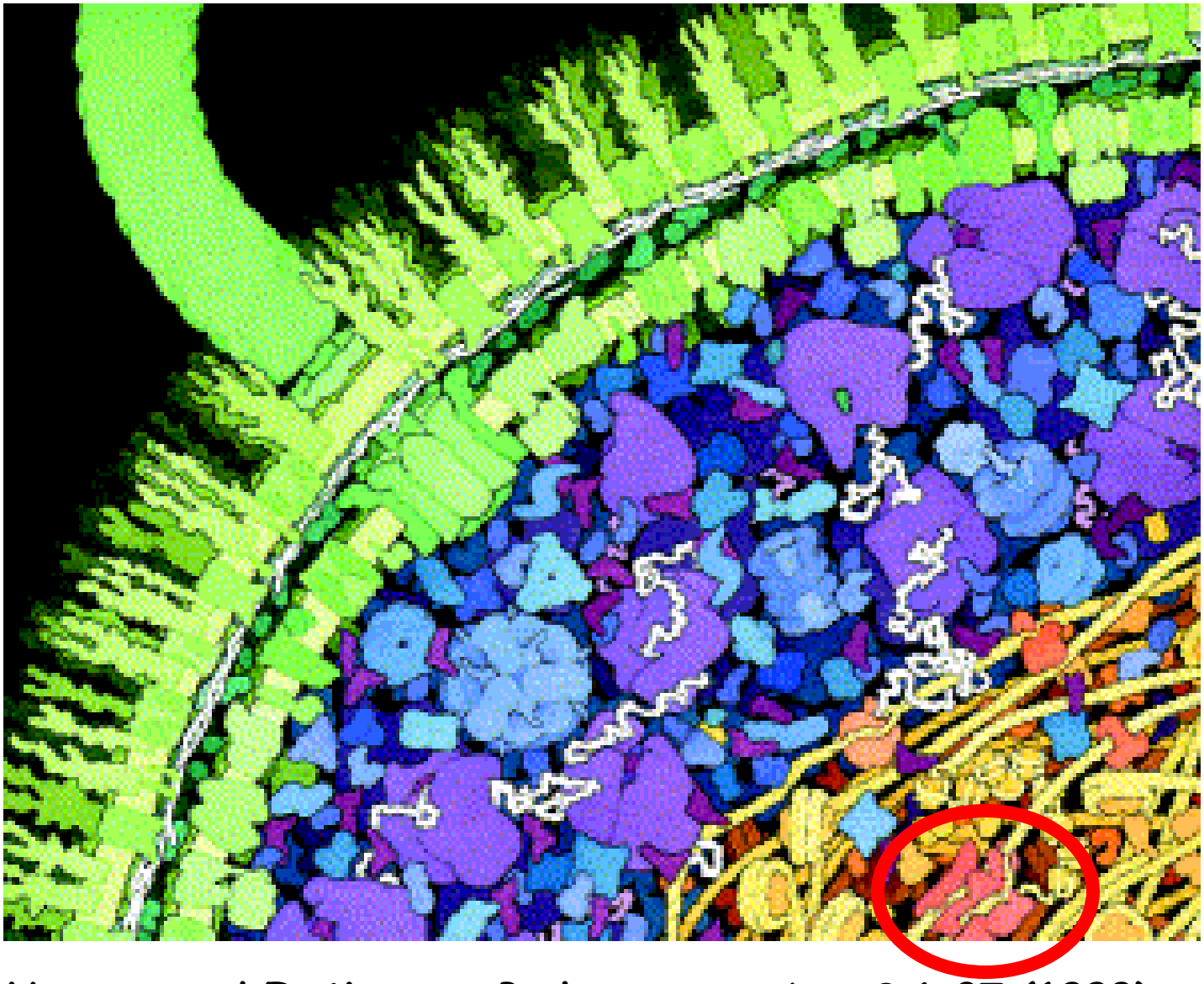


In terms of proteomics, interactomics refers to protein-protein interaction networks



Current C. elegans (worm) interactome
http://interactome.dfci.harvard.edu/C_elegans/

Macromolecular crowding: obvious but underappreciated



M. Hoppert and F. Mayer , Prokaryotes. *Am. Sci.* 87 (1999)

PDB
PROTEIN DATA BANK
Molecule of the Month
by David S. Goodsell
index of installments

RNA Polymerase

Not Just for Messages

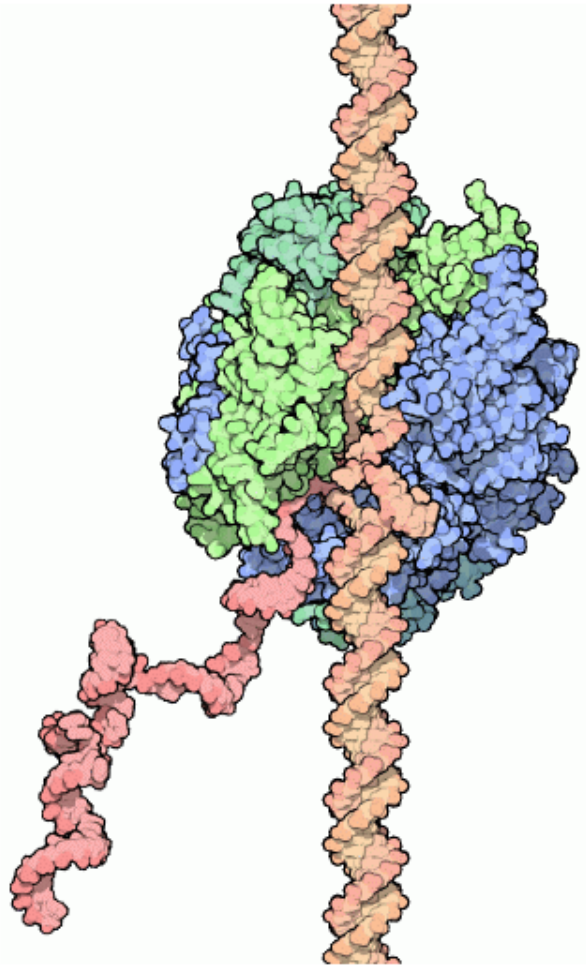
RNA is a versatile molecule. In its most familiar role, RNA acts as an intermediary, carrying genetic information from the DNA to the machinery of protein synthesis. RNA also plays more active roles, performing many of the catalytic and recognition functions normally reserved for proteins. In fact, most of the RNA in cells is found in ribosomes--our protein-synthesizing machines--and the transfer RNA molecules used to add each new amino acid to growing proteins. In addition, countless small RNA molecules are involved in regulating, processing and disposing of the constant traffic of messenger RNA. The enzyme RNA polymerase carries the weighty responsibility of creating all of these different RNA molecules.

The RNA Factory

RNA polymerase is a huge factory with many moving parts. The one shown here, from PDB entry 1i6h, is from yeast cells. It is composed of a dozen different proteins. Together, they form a machine that surrounds DNA strands, unwinds them, and builds an RNA strand based on the information held inside the DNA. Once the enzyme gets started, RNA polymerase marches confidently along the DNA copying RNA strands thousands of nucleotides long.

Accuracy

As you might expect, RNA polymerase needs to be accurate in its copying of genetic information. To improve its accuracy, it performs a simple proofreading step as it builds an RNA strand. The active site is designed to be able to remove nucleotides as well as add them to the growing strand. The enzyme tends to hover around mismatched nucleotides longer than properly added ones, giving the enzyme time to remove them. This process is somewhat wasteful, since proper nucleotides are also occasionally removed, but this is a small price to pay for creating better RNA transcripts. Overall, RNA polymerase makes an error about once in 10,000 nucleotides added, or about once per RNA strand created.



Next: Poisoning Polymerase

RCSB PDB PROTEIN DATA BANK → RCSB PDB-101

An Information
As of Tuesday Dec 06, 2011 at 4 PM PST t

All Categories | Author | Macromolecule | Sequence | Ligand | ?

Search | All Categories: e.g., PDB ID, molecule name, author

Customize This Page

MyPDB Hide

Login to your Account
Register a New Account

Home Hide

News & Publications
Usage/Reference Policies
Deposition Policies
Website FAQ
Deposition FAQ
Contact Us
About Us
Careers
External Links
Sitemap
New Website Features

Deposition Hide

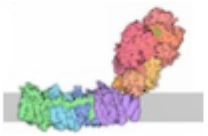
All Deposit Services

Biological Macromolecular Resource

Full Description

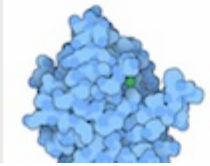
Featured Molecules

Structural View of Biology

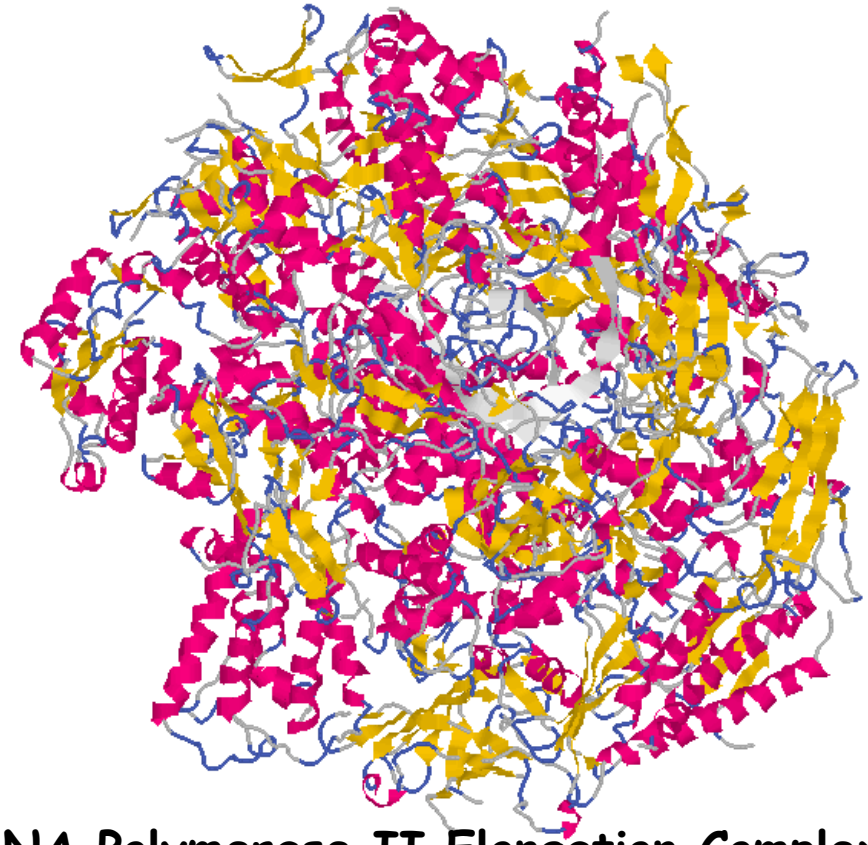


Molecule of
Complex I
Complex
transport

Full Artic



Protein Stru
Superbugs
Antibiotics
Natural a
might exp



By: T

Biolo
Ener

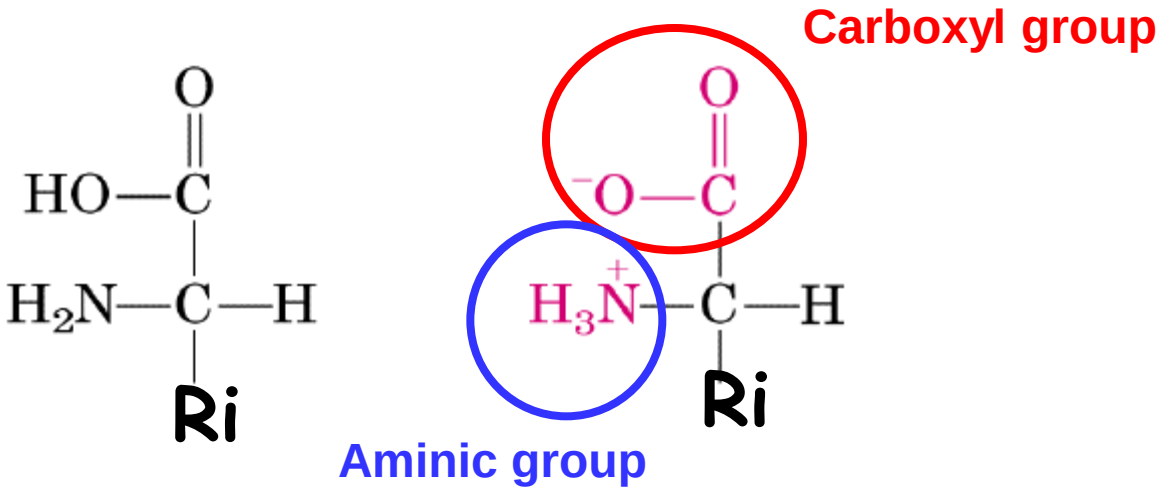
n res

disc
them

E.G:: RNA Polymerase II Elongation Complex

Some elements of previous knowledge

To start with…..you have the amino acids :



Ri are molecular residues with different structure (**20**) and different physico-chemical characteristics.

Example: http://webhost.bridgew.edu/fgorga/proteins/default.htm

Name (Residue)	3-letter code	Single code	Relative abundance (%) E.C.	MW	pK	VdW volume ($Å^3$)	Charged, Polar, Hydrophobic
Alanine	ALA	A	13.0	71		67	H
Arginine	ARG	R	5.3	157	12.5	148	C+
Asparagine	ASN	N	9.9	114		96	P
Aspartate	ASP	D	9.9	114	3.9	91	C-
Cysteine	CYS	C	1.8	103		86	P
Glutamate	GLU	E	10.8	128	4.3	109	C-
Glutamine	GLN	Q	10.8	128		114	P
Glycine	GLY	G	7.8	57		48	-
Histidine	HIS	H	0.7	137	6.0	118	P,C+
Isoleucine	ILE	I	4.4	113		124	H
Leucine	LEU	L	7.8	113		124	H
Lysine	LYS	K	7.0	129	10.5	135	C+
Methionine	MET	M	3.8	131		124	H
Phenylalanine	PHE	F	3.3	147		135	H
Proline	PRO	P	4.6	97		90	H
Serine	SER	S	6.0	87		73	P
Threonine	THR	T	4.6	101		93	P
Tryptophan	TRP	W	1.0	186		163	P
Tyrosine	TYR	Y	2.2	163	10.1	141	P
Valine	VAL	V	6.0	99		105	H

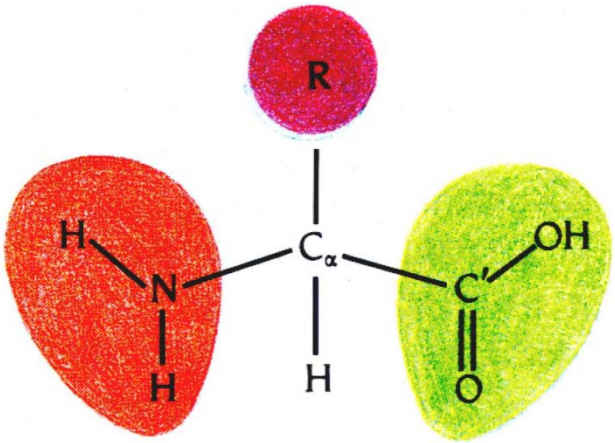
Also aminocids have a name .....

Principles of Protein Structure

Basic elements of the protein covalent structure

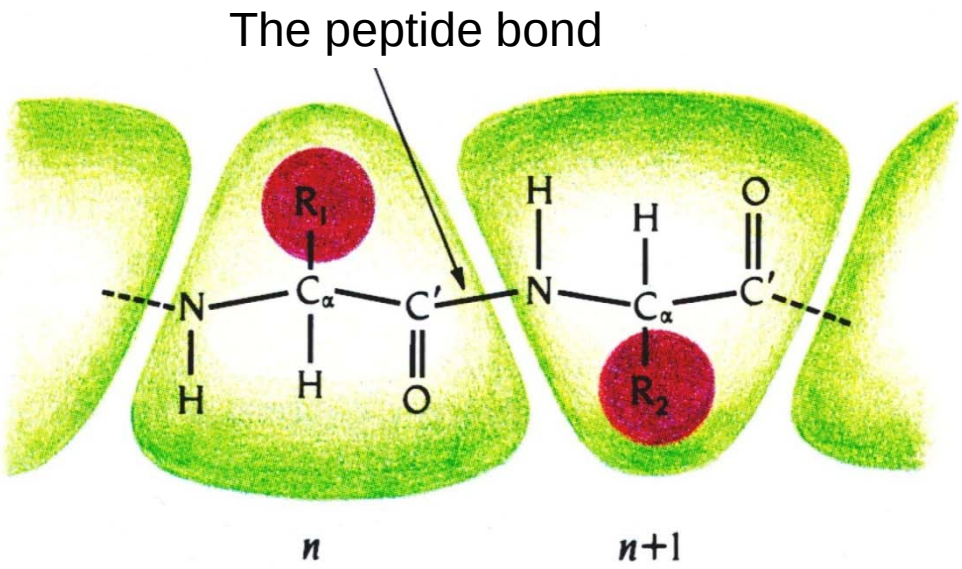
20 Different Monomers:
with Ri polar and apolar

An aminoacid



The protein hetero polymer

The peptide bond

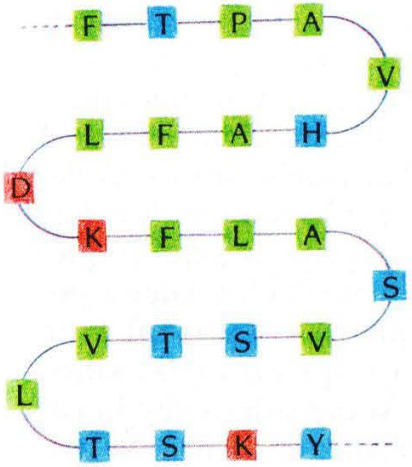


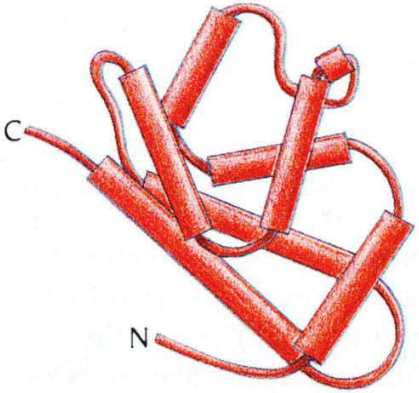
The protein backbone

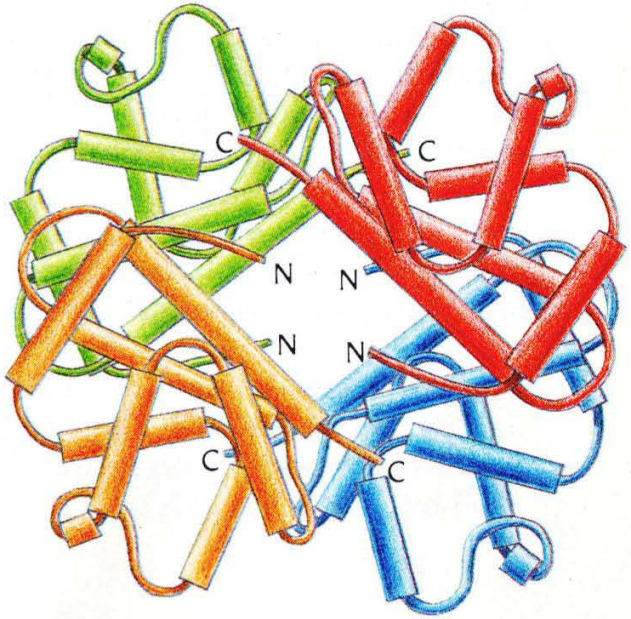
Principles of Protein Structure

Hierarchical organization

Covalent *Secondary* *Tertiary* *Quaternary*

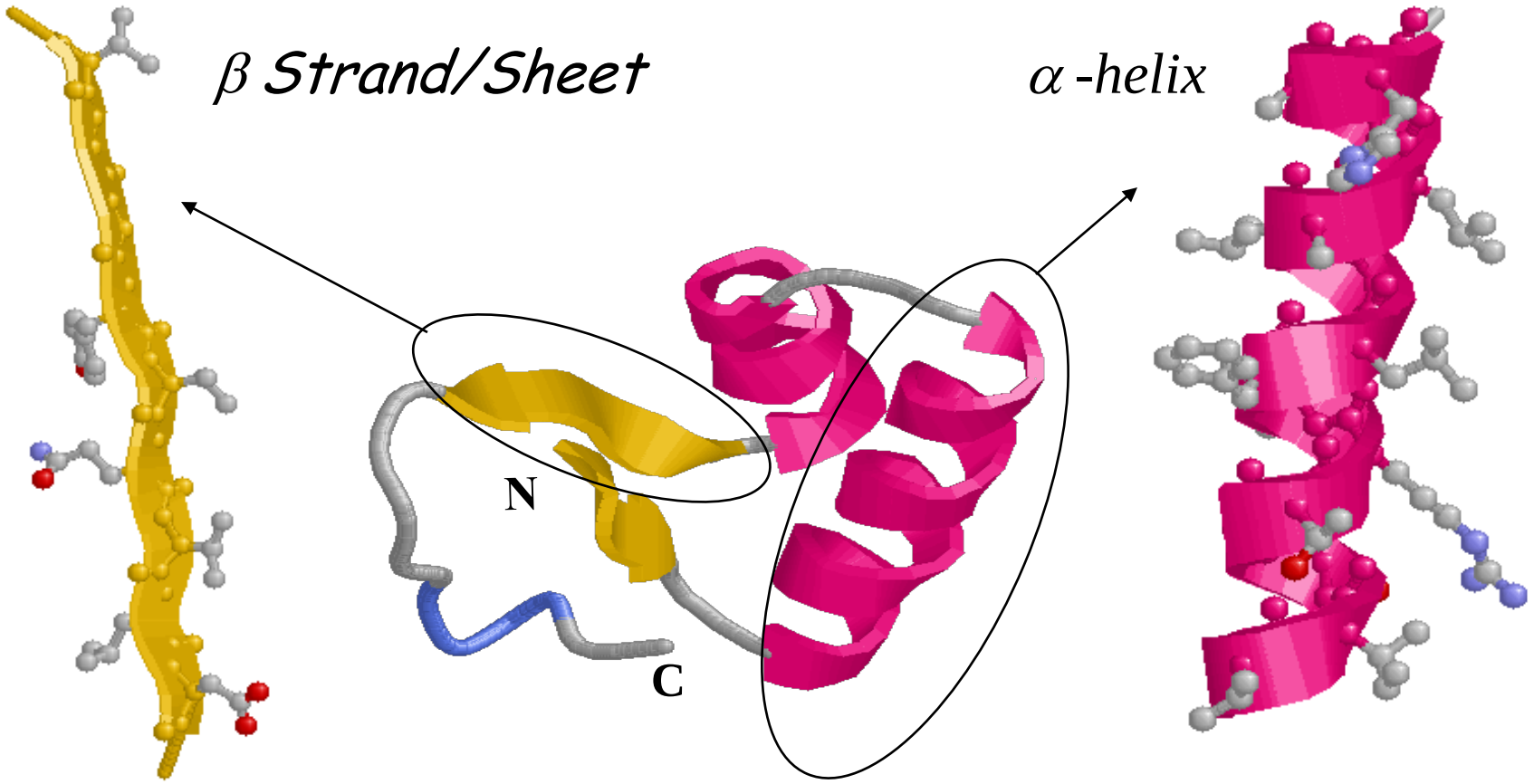






Principles of Protein Structure

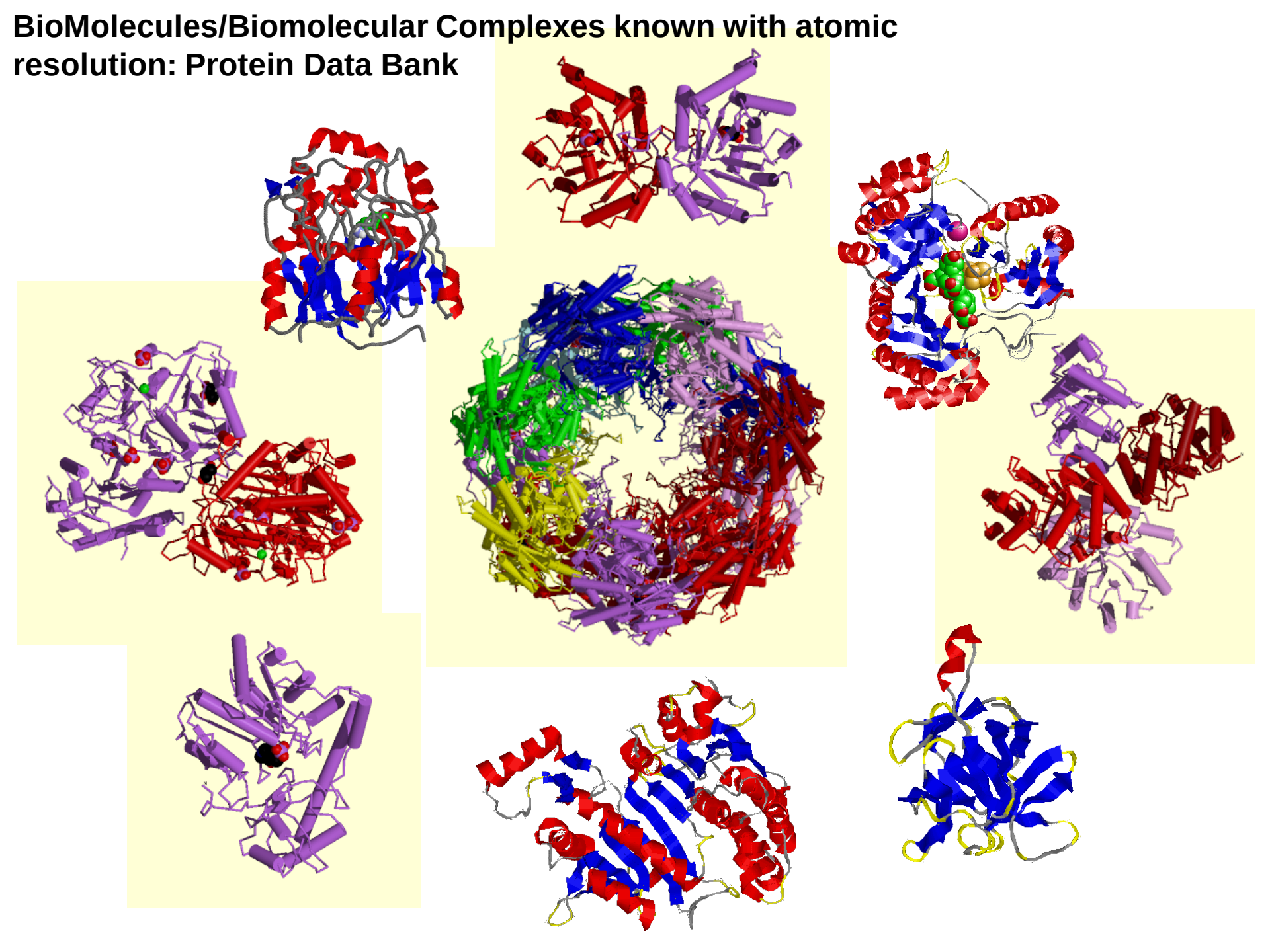
Elements of secondary structure

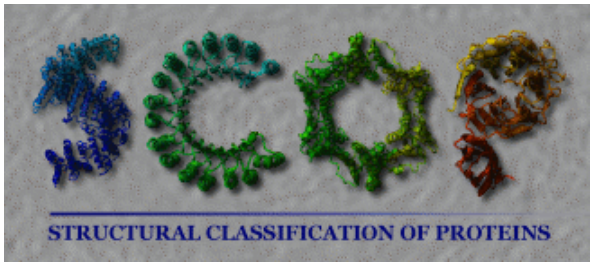


More structural details……

http://webhost.bridgew.edu/fgorga/proteins/default.htm

BioMolecules/Biomolecular Complexes known with atomic resolution: Protein Data Bank





SCOP: Structural Classification of Proteins

Classes:

1. All alpha proteins [46456] (284)
2. All beta proteins [48724] (174)
3. Alpha and beta proteins (a/b) [51349] (147)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) [53931] (376)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) [56572] (66)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [56835] (58)
   Does not include proteins in the immune system
7. Small proteins [56992] (90)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins [57942] (7)
   Not a true class
9. Low resolution protein structures [58117] (26)
   Not a true class
10. Peptides [58231] (121)
    Peptides and fragments. Not a true class
11. Designed proteins [58788] (44)
    Experimental structures of proteins with essentially non-natural sequences.

Domains are hierarchically classified:

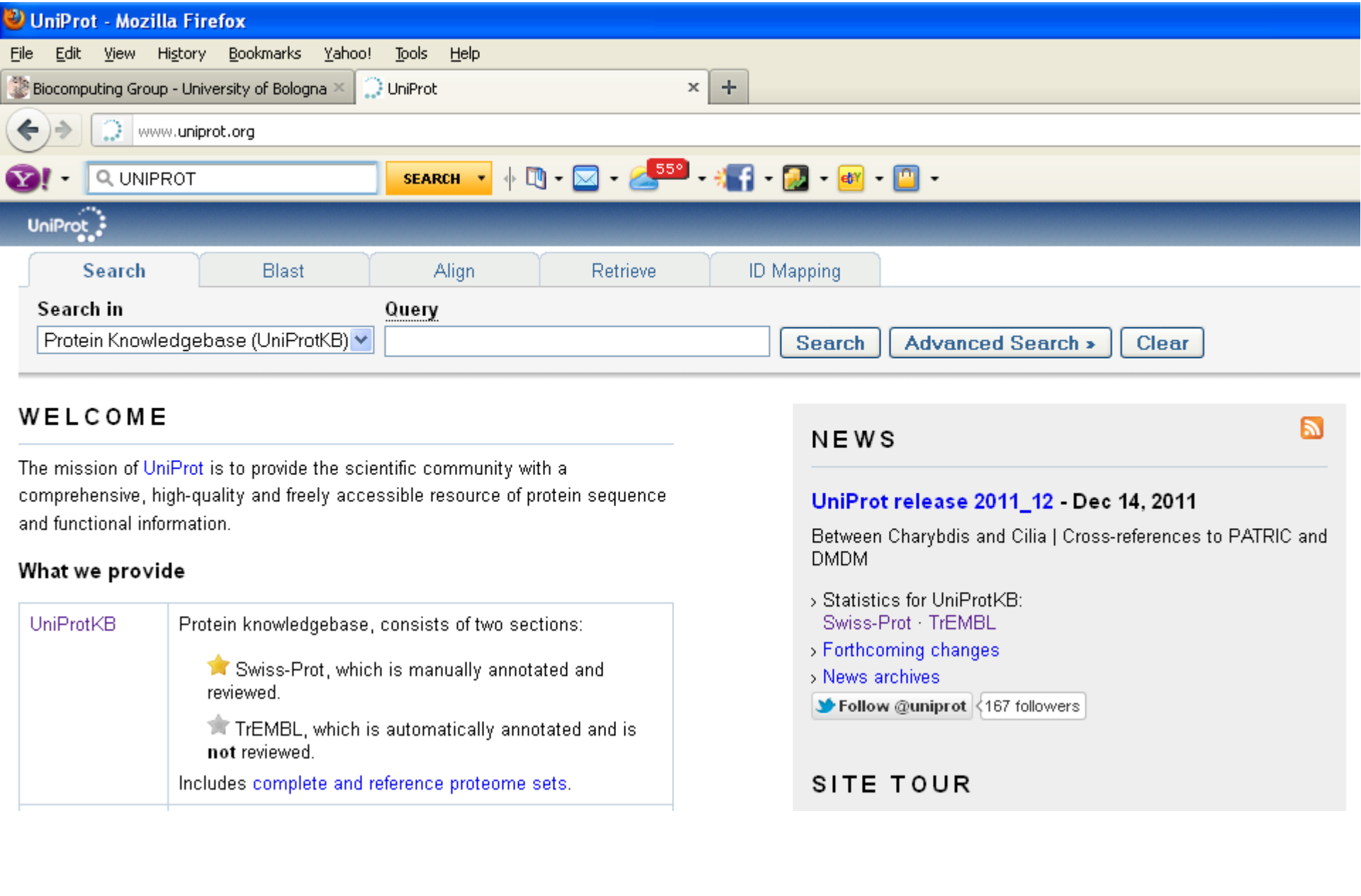
- class

- fold: proteins with secondary structures in same arrangement with the same topological connections

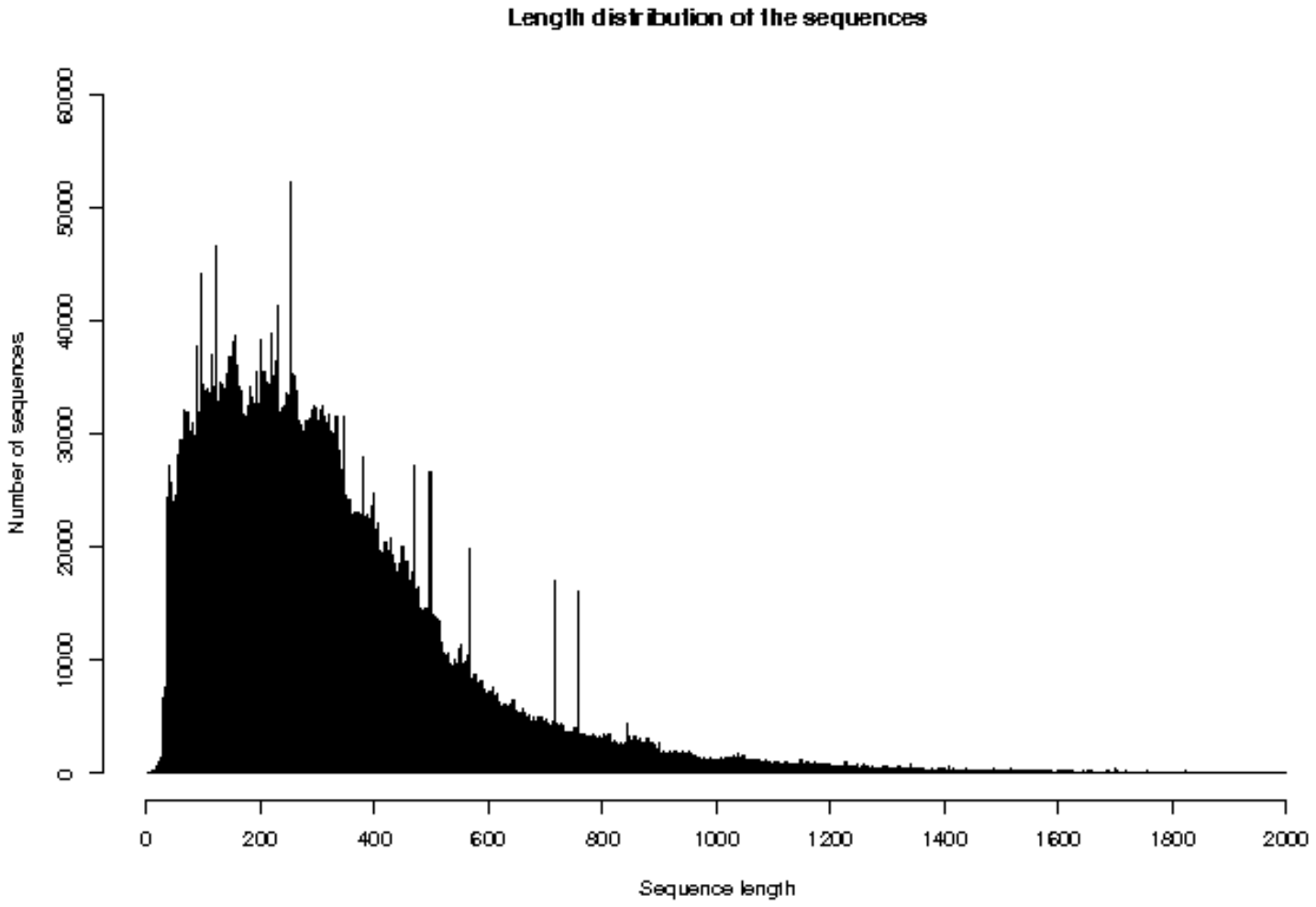
- superfamily: structures and functional features suggest a common evolutionary origin

- family: proteins with identities ≥30%; with identities <30% but with similar structures and functions

The UniProt Universe

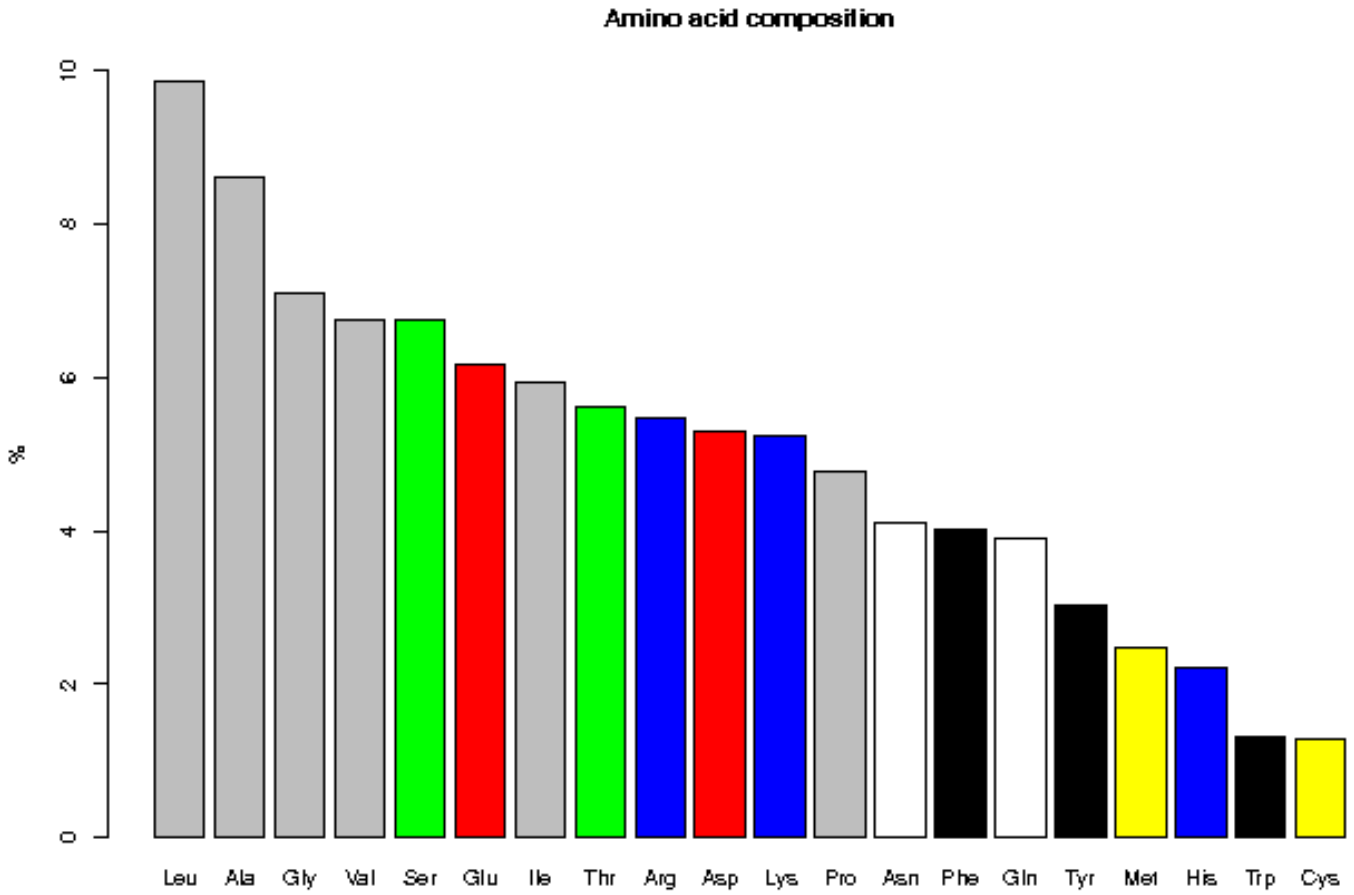


UniProtKB/TrEMBL - Current Release Statistics



The average sequence length in UniProtKB/TrEMBL is 327 amino acids.

The shortest sequence is G0XMK1_9MYRT: 1 amino acids.
The longest sequence is Q3ASY8_CHLCH: 36805 amino acids.



Legend: gray = aliphatic, red = acidic, green = small hydroxy,
blue = basic, black = aromatic, white = amide, yellow = sulfur

5.2 Classification of the amino acids by their frequency

Leu, Ala, Gly, Val, Ser, Glu, Ile, Thr, Arg, Asp, Lys, Pro, Asn, Phe,
Gln, Tyr, Met, His, Trp, Cys

Some chemico-physical properties..

Ri lateral side chains are classified as:

1) Hydrophobic (escape the interaction with the polar solvent):
Alanine, Valine, Leucine, Isoleucine, Glycine, Proline, Cisteine, Methionine (A, V, L, I, G, P, C, M)

2) Aromatic (with an aromatic ring)
Phenilalanin, Tyrosin, Triptophane (F, Y, W)

3) Polar (are stabilised by interaction with polar solvent)
Histidine, Arginine, Glutamine, Serine, Treonine (H, N, Q, S, T)

4) Charged (characterised by local positive or negative charged groups at physiological pH)
Lysine, Asparagine, Aspartic Acid, Glutammic Acid(K, R, D, E)

The protein folding problem……

Proteins are frustrated systems…..

KVFGRCELAA AMKRHGLDNY RGYSLGNWVC AAKFESNFNT
QATNRNTDGS TDYGILQINS RWWCNDGRTP GSRNLCNIPC
SALLSSDITA SVNCAKKIVS DGNGMNAWVA WRNRCKGTDV
QAWIRGCRL

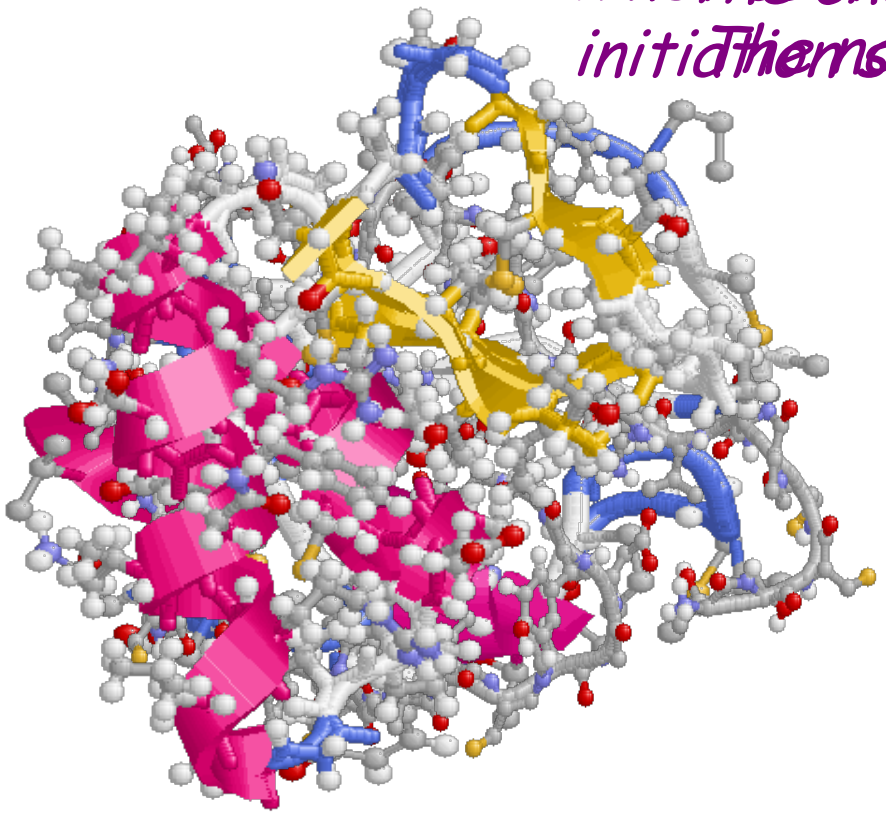
Too many tendencies when they in contact with the polar solvent ….the hydrophobic effects dominates….

http://webhost.bridgew.edu/fgorga/proteins/default.htm

The Protein Folding Problem

The folding process

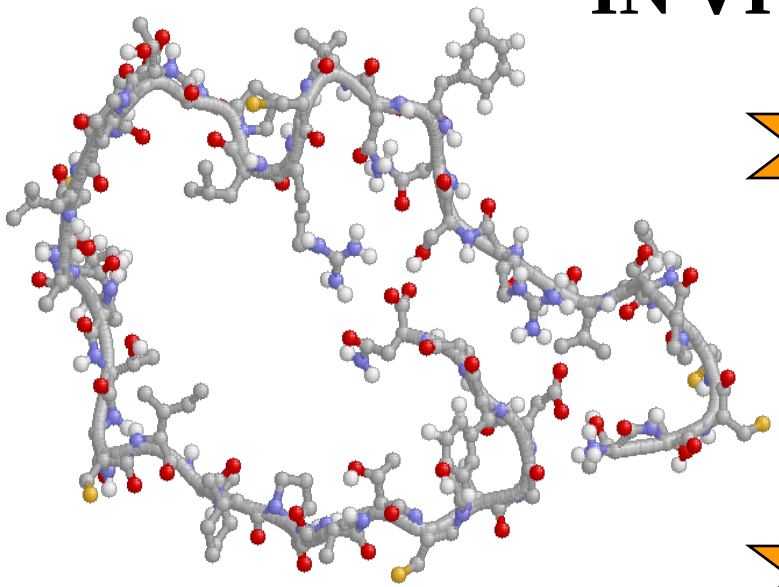
The folding kine The chain initid The native protein



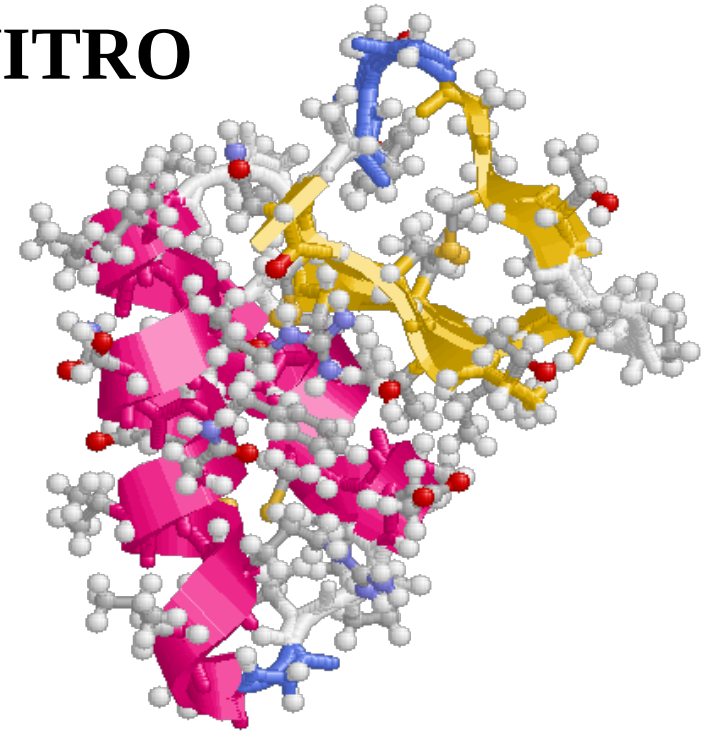
The Protein Folding problem

The polypeptide chain *The native structure*

IN VIVO & IN VITRO



?



IN SILICO

Little exercise……

Observation: the Residue composition of protein chains is rather conserved

Question: How many theoretical proteins can be predicted on the basis of a protein length of 100 residues?

How many of the predicted ones do you think to find in UniProtKB?

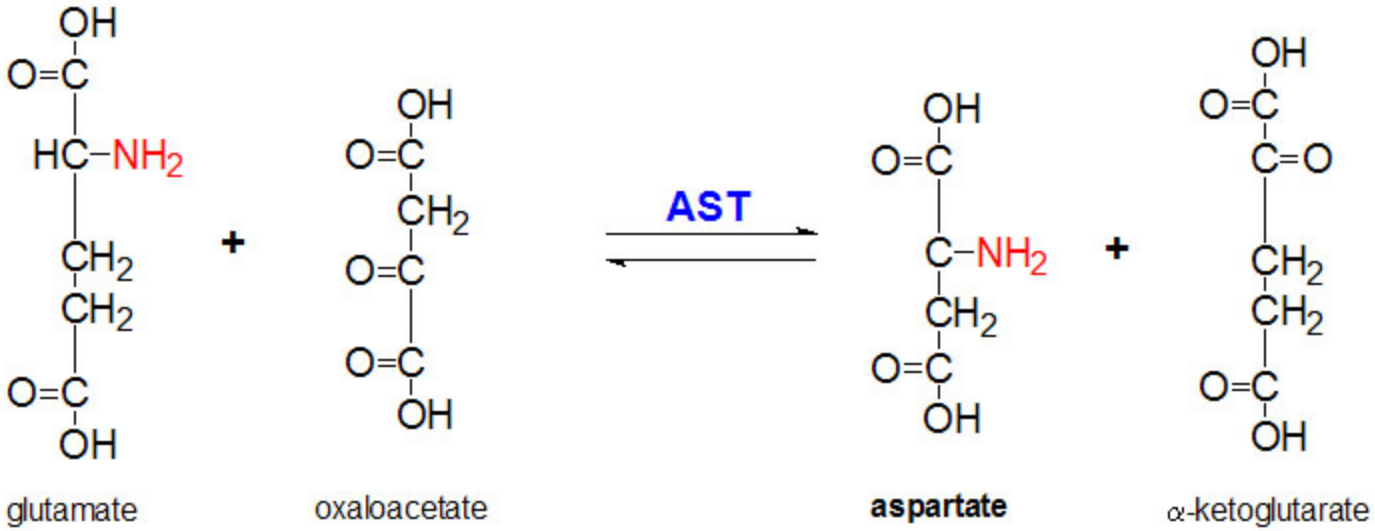
From the Protein Sequence to the Structure and Function space

What is protein function?

What is a function?

For enzymes: function can be defined on the basis of the catalysed molecular reaction.

e.g. aspartic aminotransferase (AST)



In biochemistry, a **transaminase** or an **aminotransferase** is an enzyme that catalyzes a type of reaction between an amino acid and an α-keto acid.

Specifically, this reaction (transamination) involves removing the amino group from the amino acid, leaving behind an α-keto acid, and transferring it to the reactant α-keto acid and converting it into an amino acid. The enzymes are important in the production of various amino acids, and measuring the concentrations of various transaminases in the blood is important in the diagnosing and tracking many diseases. Transaminases require the coenzyme *pyridoxal-phosphate*, which is converted into *pyridoxamine* in the first phase of the reaction, when an amino acid is converted into a keto acid.

Enzyme-bound pyridoxamine in turn reacts with pyruvate, oxaloacetate, or alpha-ketoglutarate, giving alanine, aspartic acid, or glutamic acid, respectively.

The presence of elevated transaminases can be an indicator of liver damage.

Enzyme Commission (E.C.) classification

A hierarchical classification for enzymes

1. -. -.- Oxidoreductases.

2. -. -.- Transferases.

3. -. -.- Hydrolases.

4. -. -.- Lyases.

5. -. -.- Isomerases.

6. -. -.- Ligases.

2. -. -.- Transferases.
2. 1. -.- Transferring one-carbon groups.
2. 1. 1.- Methyltransferases.
2. 1. 2.- Hydroxymethyl-, formyl- and related transferases.
2. 1. 3.- Carboxyl- and carbamoyltransferases.
2. 1. 4.- Amidinotransferases.
2. 2. -.- Transferring aldehyde or ketone residues.
2. 2. 1.- Transketolases and transaldolases.
2. 3. -.- Acyltransferases.
2. 3. 1.- Transferring groups other than amino-acyl groups.
2. 3. 2.- Aminoacyltransferases.
2. 3. 3.- Acyl groups converted into alkyl on transfer.
2. 4. -.- Glycosyltransferases.
2. 4. 1.- Hexosyltransferases.
2. 4. 2.- Pentosyltransferases.
2. 4.99.- Transferring other glycosyl groups.
2. 5. -.- Transferring alkyl or aryl groups, other than methyl groups.
2. 5. 1.- Transferring alkyl or aryl groups, other than methyl groups.
2. 6. -.- Transferring nitrogenous groups.
2. 6. 1.- Transaminases (aminotransferases).
2. 6. 3.- Oximinotransferases.
2. 6.99.- Transferring other nitrogenous groups.

EC 2.6 Transferring nitrogenous groups
EC 2.6.1Transaminases
EC 2.6.1.1 Aspartate transaminase

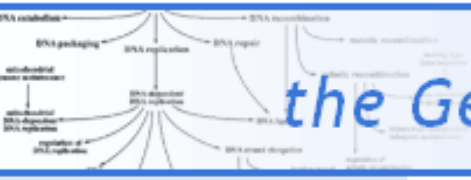
Other name(s): glutamic-oxaloacetic transaminase; glutamic-aspartic transaminase; transaminase A; AAT; AspT; 2-oxoglutarate-glutamate aminotransferase; aspartate α-ketoglutarate transaminase; aspartate aminotransferase; aspartate-2-oxoglutarate transaminase; aspartic acid aminotransferase; aspartic aminotransferase; aspartyl aminotransferase; AST; glutamate-oxalacetate aminotransferase; glutamate-oxalate transaminase; glutamic-aspartic aminotransferase; glutamic-oxalacetic transaminase; glutamic oxalic transaminase; GOT (enzyme); L-aspartate transaminase; L-aspartate-α-ketoglutarate transaminase; L-aspartate-2-ketoglutarate aminotransferase; L-aspartate-2-oxoglutarate aminotransferase; L-aspartate-2-oxoglutarate-transaminase; L-aspartic aminotransferase; oxaloacetate-aspartate aminotransferase; oxaloacetate transferase; aspartate:2-oxoglutarate aminotransferase; glutamate oxaloacetate transaminase

Systematic name: L-aspartate:2-oxoglutarate aminotransferase

Problems:

Isoforms
e.g How to differentiate the function of the cytoplasmic aspartate aminotransferase from that of mitochondrial isoform?

Non enzymatic proteins



the Gene Ontology

Open menus

GO function vocabulary:

http://www.geneontology.org/

The Ontologies

- Cellular component
- Biological process
- Molecular function

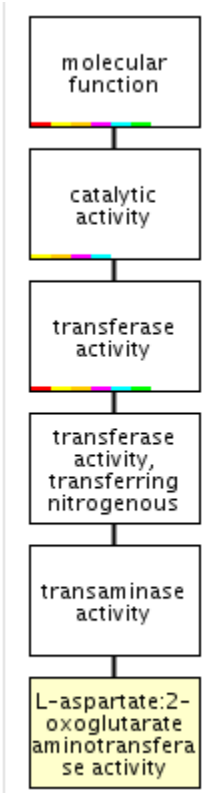
Ontology Structure

The Gene Ontology is a **controlled vocabulary**, a set of standard terms—words and phrases—used for indexing and retrieving information. In addition to defining terms, GO also defines the **relationships** between the terms, making it a **structured** vocabulary.

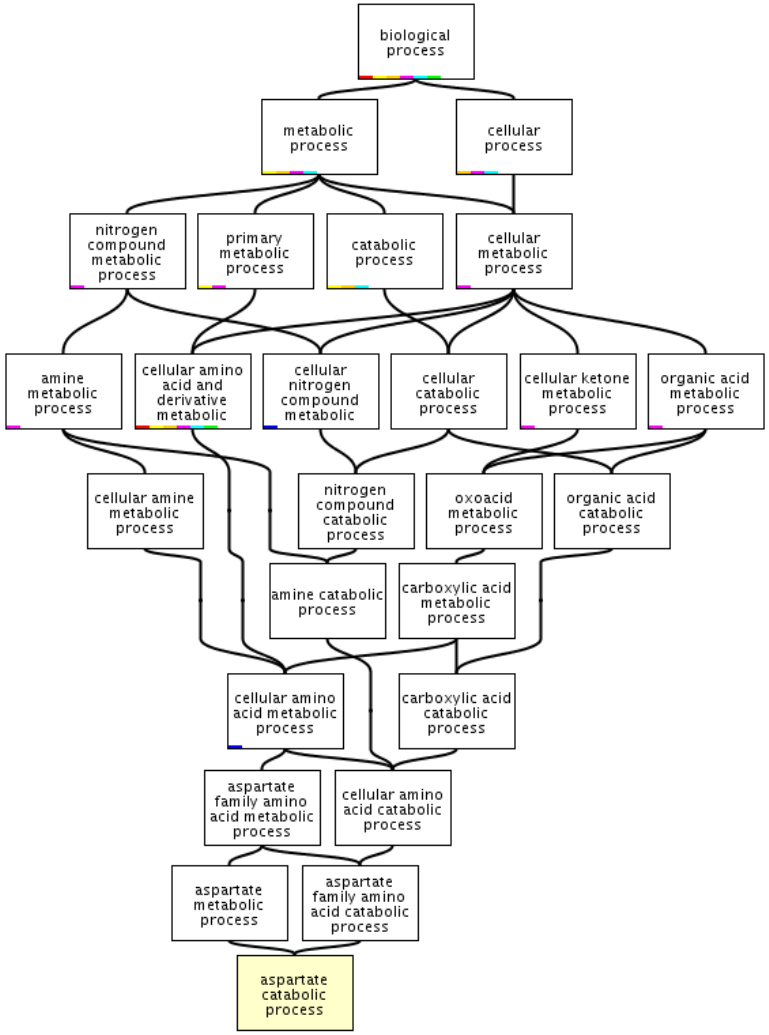
GO as a Graph

The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are arcs between the nodes. The relationships used in GO are **directed**—for example, a mitochondrion *is an* organelle, but an organelle is not a mitochondrion—and the graph is **acyclic**, meaning that cycles are not allowed in the graph. The ontologies resemble a hierarchy, as child terms are more specialized and parent terms are less specialized, but unlike a hierarchy, a term may have more than one parent term. For example, the biological process term hexose biosynthetic process has two parents, hexose metabolic process and monosaccharide biosynthetic process. This is because biosynthetic process is a type of metabolic process and a hexose is a type of monosaccharide.

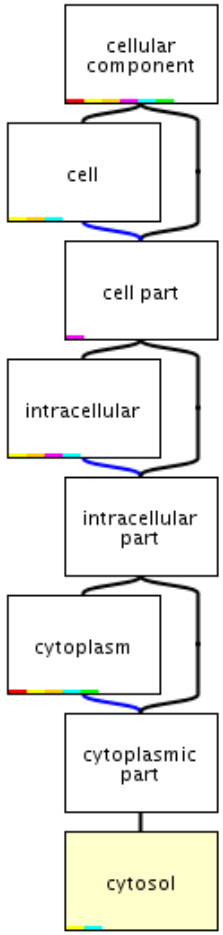
Gene Ontology classification:
The human cytoplasmic aspartate transaminase



GO:0004069



GO:0006533



GO:0005829

Genomic data and the problem of protein validation

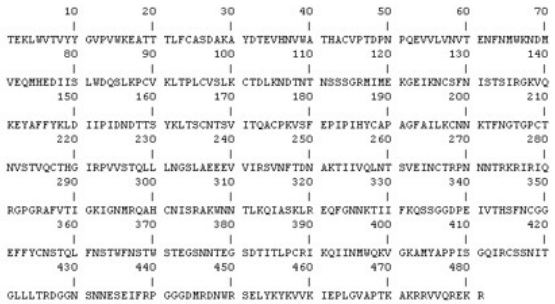
Data production→Data analysis

DNA sequencing →gene recognition → protein translation

Experiments to validate protein structure and function produce data in a time >> than that required to deposit putative protein sequences into data bases

A "BIG" problem of the "omic era" after genome sequencing:

code for



that are endowed with

Protein sequences (~17 millions)

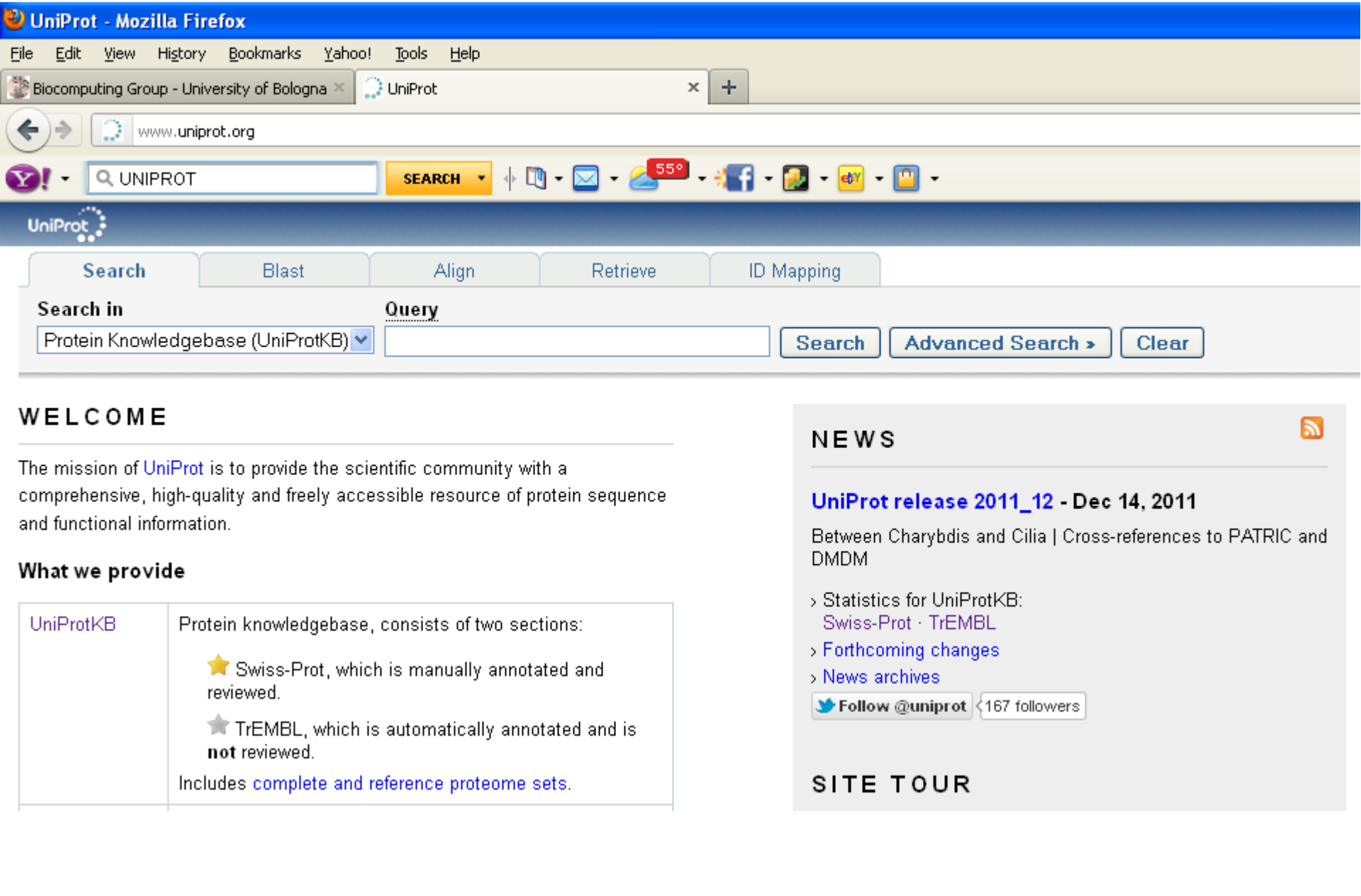
Genomes (~3000)

Protein structures and functions

Protein sequence Annotation:

to endow with structural and functional features protein sequences after gene translation

The UniProt Universe



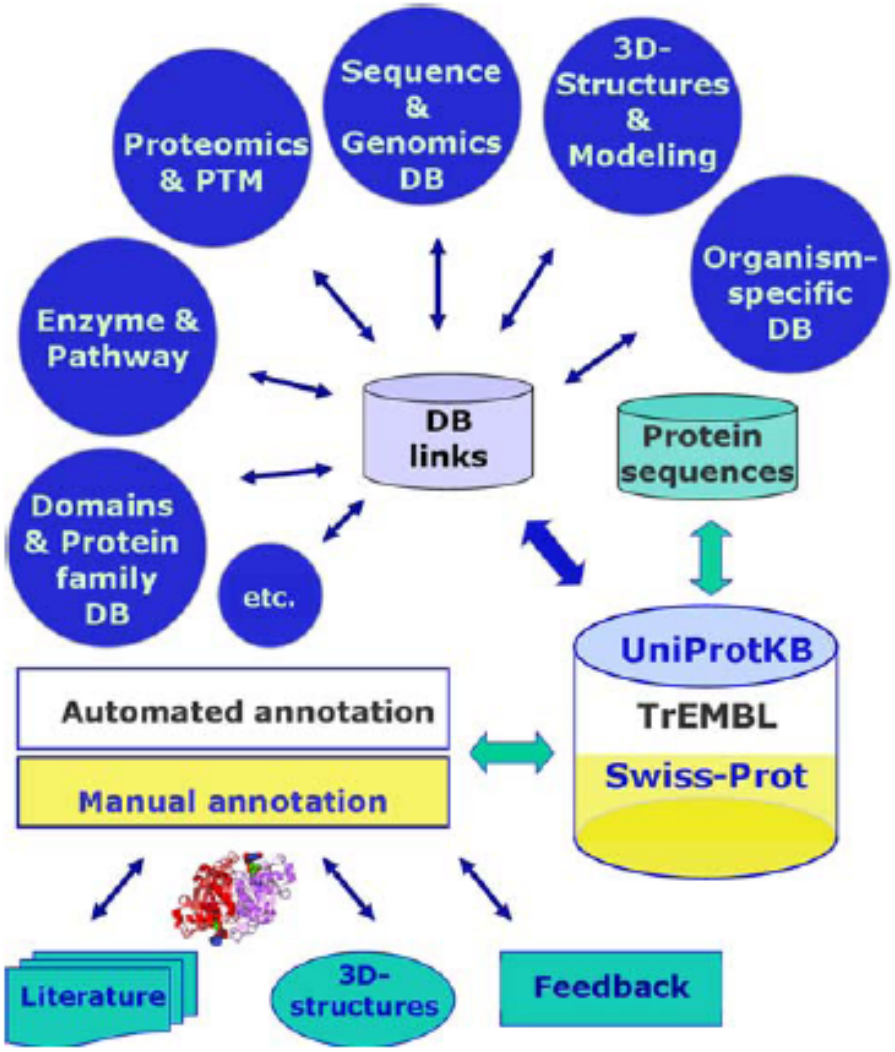
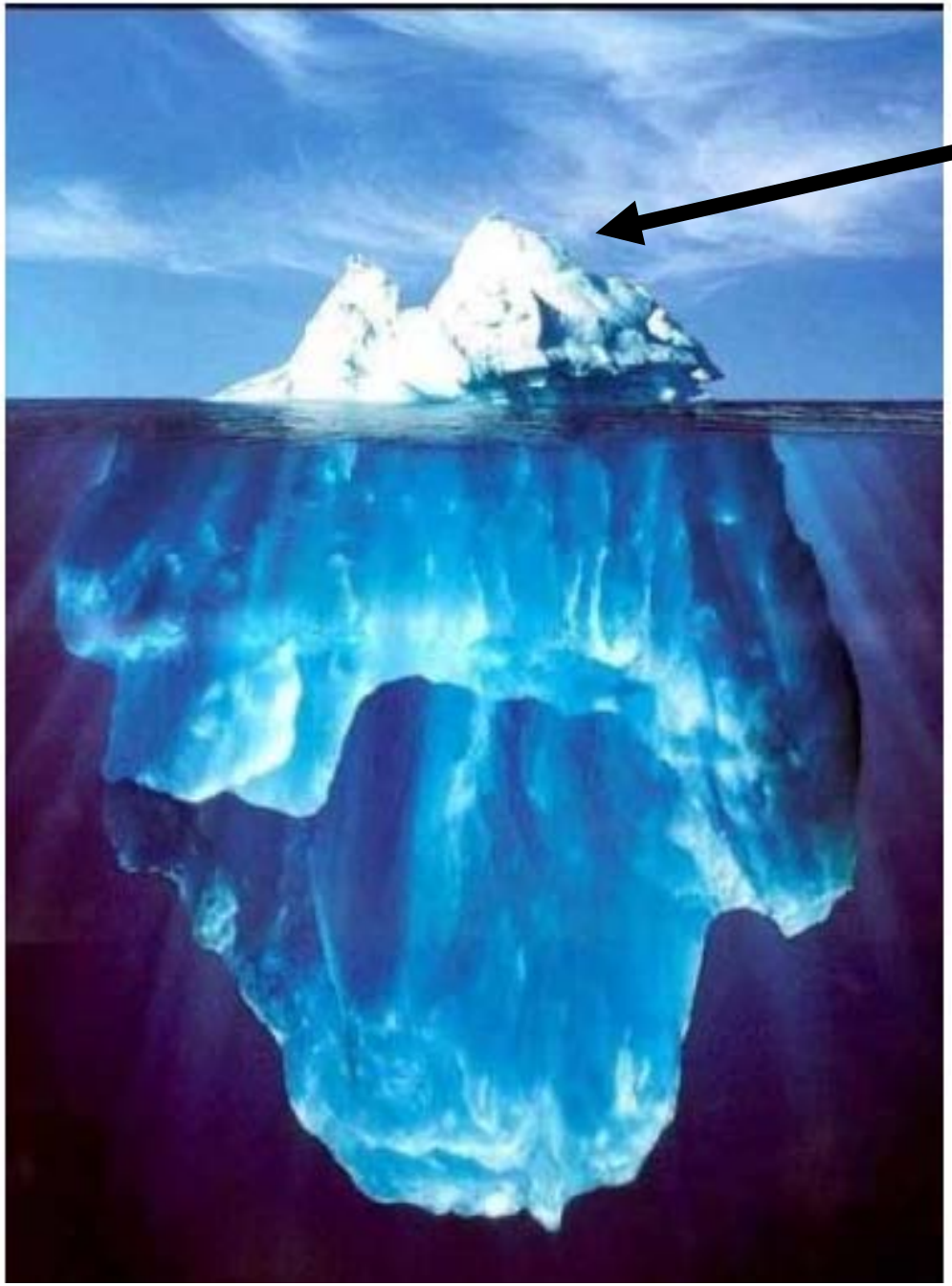


Fig. 1 UniProtKB serves as a knowledge repository and as a central hub that provides links to numerous other databases. New protein sequences are integrated in UniProtKB/TrEMBL and annotated by an automated procedure. UniProtKB/Swiss-Prot entries are manually annotated, combining carefully checked protein sequences with information from the scientific literature, protein 3D-structures, and specialised databases, together with feedback from the scientific community

UniProt KB: The largest annotation resource

Ursula Hinz • *The UniProt Consortium*
Cell. Mol. Life Sci. (2010) 67:1049–1064

DATA -INTEGRATION



The "omic" era

Complexity

- Genomics
- Transcriptomics
- Proteomics
- Metabolomics
- Regulomics

Systems Biology

Release 2011_11 of 16-Nov-2011 of UniProtKB/TrEMBL contains 18,215,214 sequence entries

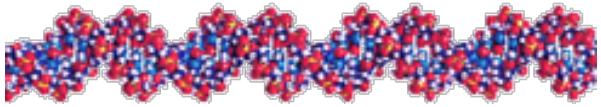
Protein existence (PE):	entries	
1: Evidence at protein level	13085	0.07%
2: Evidence at transcript level	547306	3.00%
3: Inferred from homology	3857630	21.18%
4: Predicted	13797193	75.75%

Release 2011_11 of 16-Nov-11 of UniProtKB/Swiss-Prot contains 533,049 sequence entries

Protein existence (PE):	entries	
1: Evidence at protein level	73298	13.8%
2: Evidence at transcript level	69925	13.1%
3: Inferred from homology	373485	70.1%
4: Predicted	14452	2.7%
5: Uncertain	1889	0.4%

Only 3.4 % sequences has evidence at the protein and trascript level and only 0.4 % proteins have structures in the Protein Data Bank.

The Data Bases of Biological Sequences and Structures

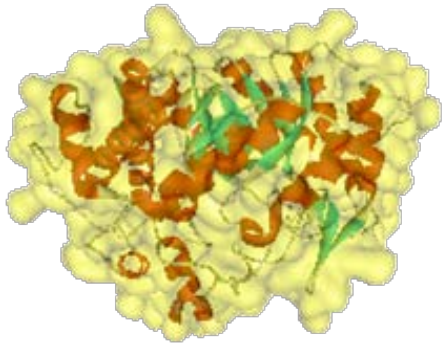


GenBank: 135,440,990 sequences
126,551,501,141 nucleotides

```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus.
MYSFPNSFRFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI
KSVSKKPVGIIYANSSFQPLTDKDMEAVEMAENDNRWWFFDAIIRGEITR
GNEKIVRDDLKGRLDWIGVNYYTRTVVKRTEKGYVSLGGYGHGCERNSVS
LAGLPTSDFGWEFFPEGLYDVLTKYWNRYHLYMYVTENGIADDADYQRPY
YLVSHVYQVHRAINSGADVRGYLHWSLADNYEWASGFSMRFGLLKVDYNT
KRLYWRPSALVYREIATNGAITDEIEHLNSVPPVKPLRH
```

UniProt/Tremble:
18,215,214 sequences
5,957,253,786 residues

UniProt/SwissProt:
533,049 sequences
189,064,225 residues



PDB: 75,4633 structures
membrane proteins <2%

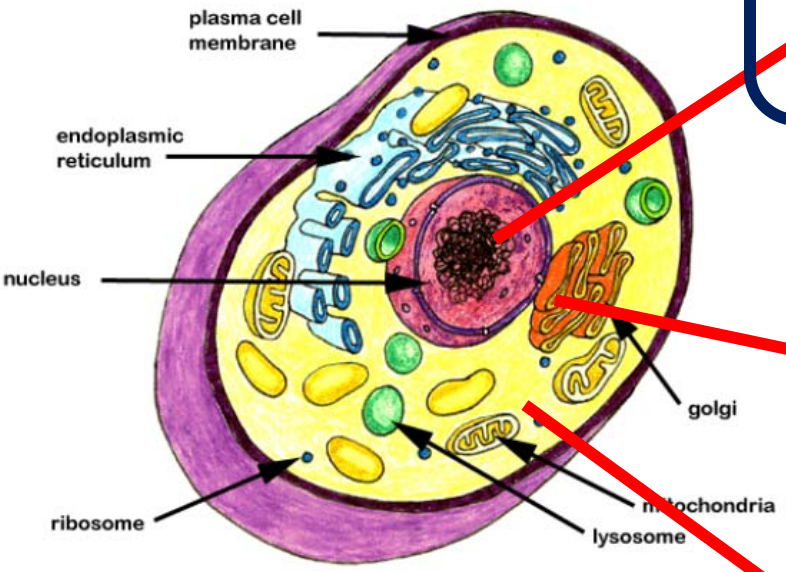
≅43 HGE!

Update:
December 2011

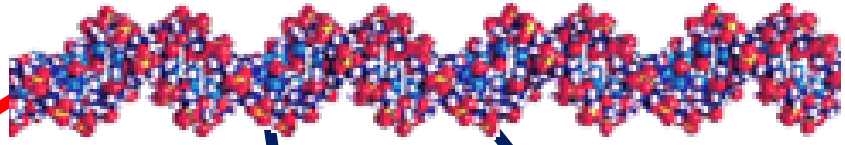
The ingredients of biological complexity at the cell level

From genes to proteins and their interaction

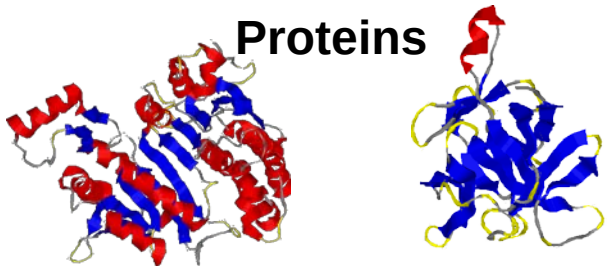
The cell

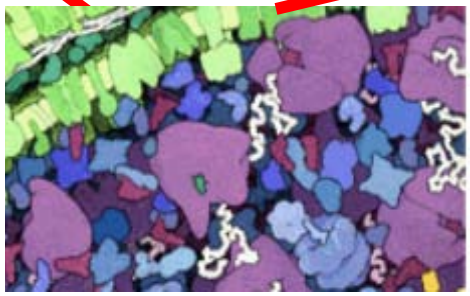


DNA & Genes



Proteins

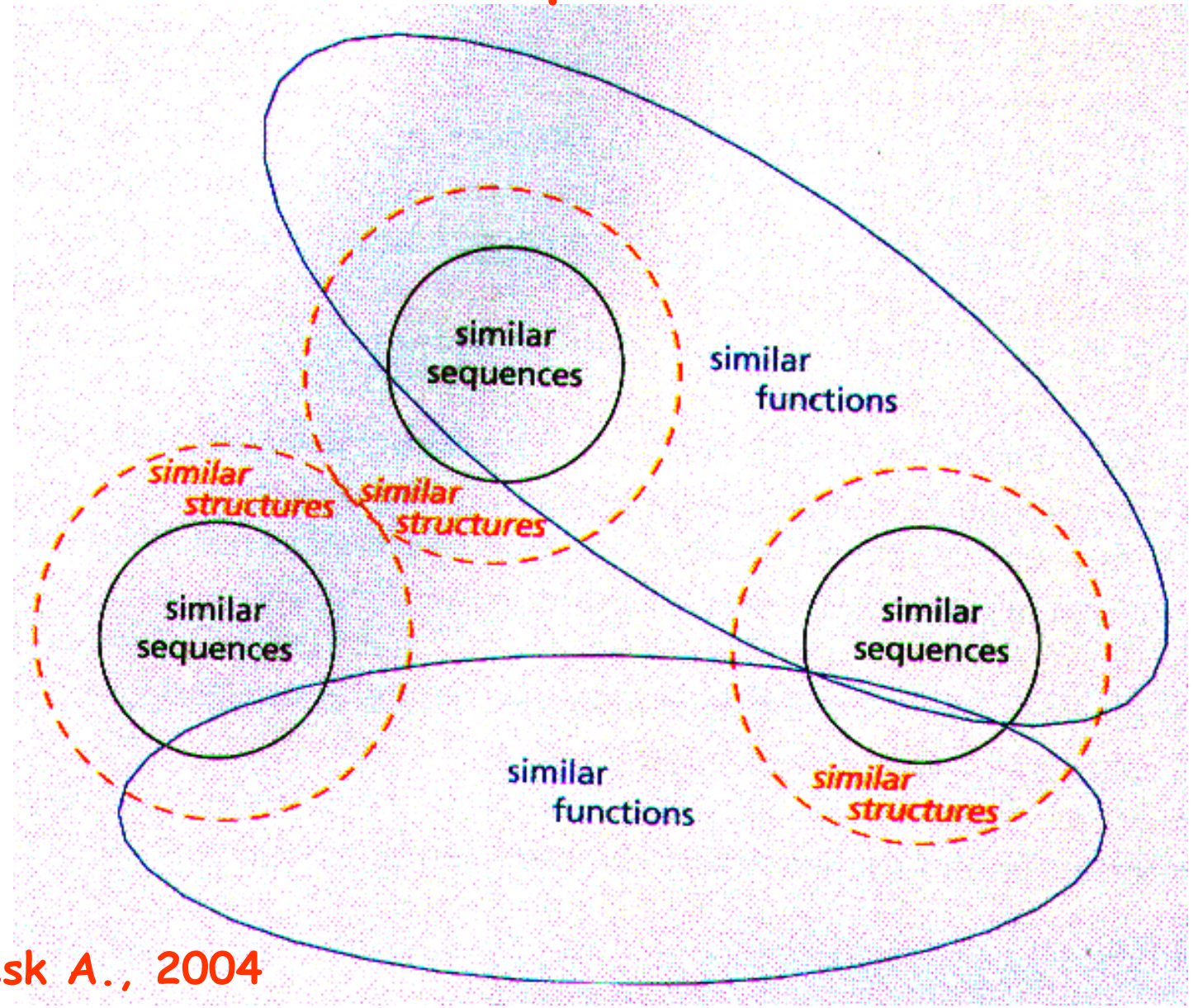




Macromolecular crawding

Molecular Interactions and Functions

How can we infer function and structure from sequence?



Lesk A., 2004

Summing up…..

Open problems:

1)Protein structural and functional annotation

BIOINFORMATICS

Data Bases

(Biosequences, Structures, Genomes, DNA Chips, Proteomes, Interatomics, Literature)

- Implementation
- Data Mining
- Links

Computational Biology

Tools for:

- Sequence analysis
- Functional genomics
- Proteomics

Systems Biology

Models for:
Interatomics, Methabolomics, Evolving complex biosystems (Cell, Organism,..)