Public Databases in Health and Life Sciences

“the potential to translate big data into big discovery”

Academic year 2023-2024

Public Databases in Health and Life Sciences

Topic 3. Protein Sequence Databases: from primary sequence to the tertiary structure.

- The primary sequence databases for proteins.
- Raw databases vs. curated databases.
- Redundancies in databases: RefSeq, UniRef and the Swiss-Prot.
- UniProt: the next generation of protein sequence databases.
- Databases for protein structures. The PDB.

Practical session #3

http://www.ncbi.nlm.nih.gov/

FASTA file

```
>NM_000517.4 Homo sapiens hemoglobin subunit alpha 2
CATAAACCCTGGCGCGCTCGCGGGCCGGCACTCTTCTGGTCCCCACAGACTCAGAGA
GAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGT
AAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTG
TCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCC
CAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCAC
GTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTT
CGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCC
GCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTG
GCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTAGCC
GTTCCTCCTGCCCGCTGGGCCTCCCAACGGGCCCTCCTCCCCTCCTTGCACCGGCCC
TTCCTGGTCTTTGAATAAAGTCTGAGTGGGCAGCAAAAAAAAAAAAAAAAAA
```

https://en.wikipedia.org/wiki/FASTA_format

>NM_000517.6 Homo sapiens hemoglobin subunit alpha 2 (HBA2), mRNA

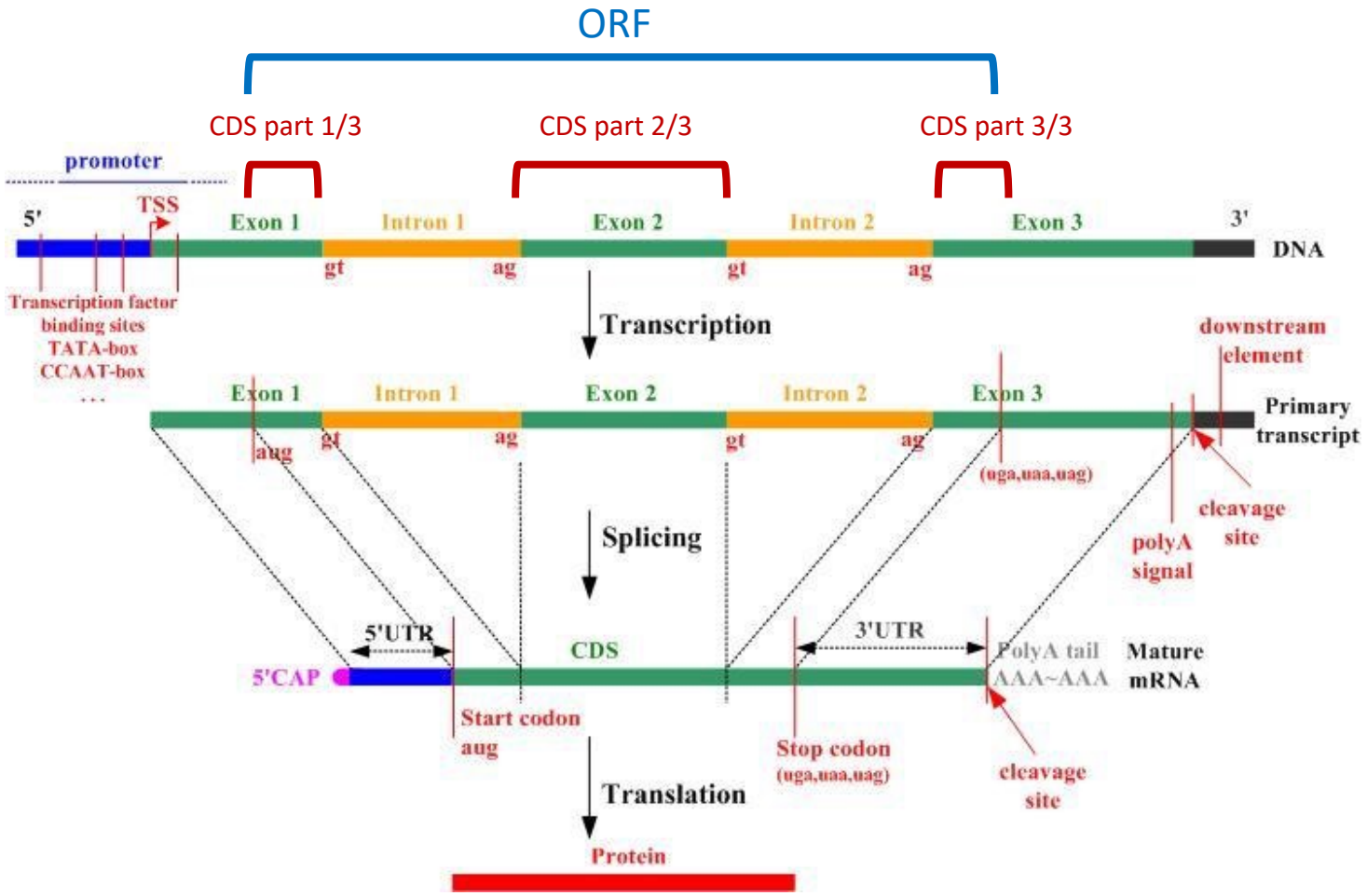
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTC
AAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCC
TGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGG
CCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTG
TCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACT
GCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAA
GTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTAGCCGTTCCT
CCTGCCCGCTGGGCCTCCCAACGGGCCCTCCTCCCCTCCTTGCACCGGCCCTTCCTGGTCTTTGAATAAA
GTCTGAGTGGGCAGCA

>NP_000508.1 hemoglobin subunit alpha [Homo sapiens]

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVK
GHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL
PAEFTPAVHASLDKFLASVSTVLTSKYR

CDS = coding sequence

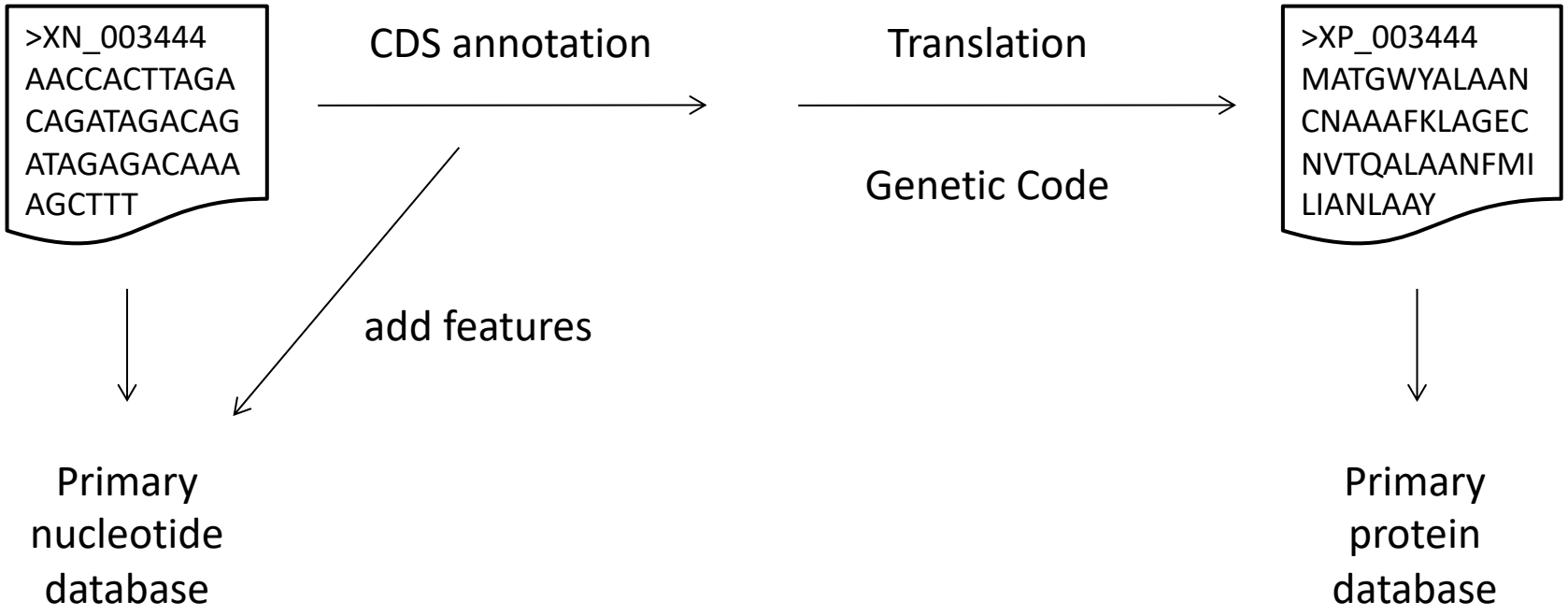
ORF = open reading frame



Example: https://www.ncbi.nlm.nih.gov/nuccore/NG_009238.3
https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP020401.2

Image source: http://i.imgur.com/OjYRouN.jpg

CDS annotation steps



https://www.ncbi.nlm.nih.gov/nuccore/AM743169.1?from=2707467&to=2709688
https://www.ncbi.nlm.nih.gov/nuccore/NC_010943.1?from=2707467&to=2709688
https://www.uniprot.org/uniprot/?query=Smlt2685
https://www.uniprot.org/uniprot/?query=SMLT_RS12800
https://www.uniprot.org/uniprotkb?query=(xref:refseq-WP_005409835.1)

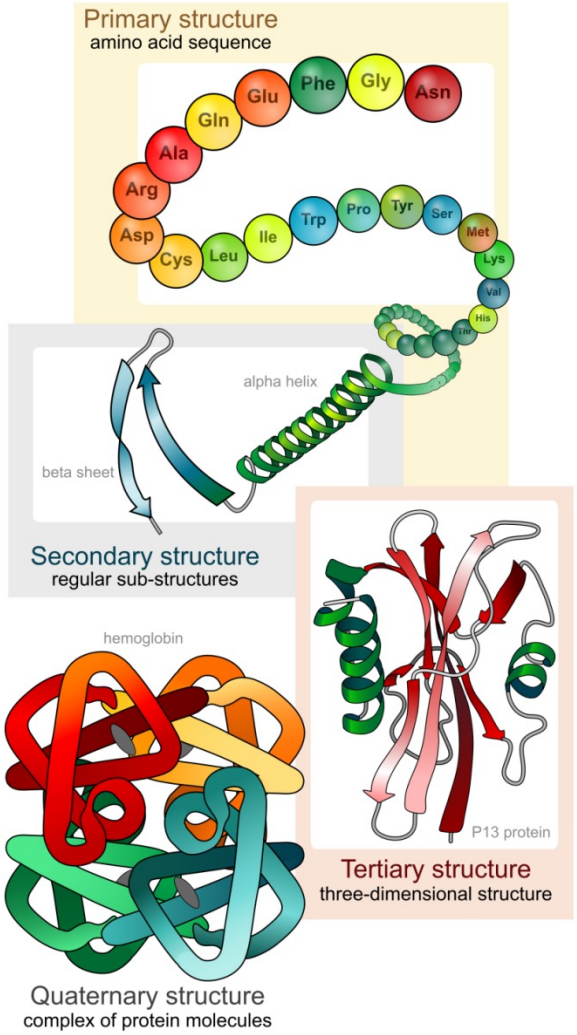
Genetic codes: https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

Alternative genetic codes

UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA Leu	UCA Ser	UAA Stp, Gln[3]	UGA Stp,Trp[4,5],Cys[6],SeC[7]
UUG Leu	UCG Ser	UAG Stp, Gln[3]	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC Leu	CCC Pro	CAC His	CGC Arg
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu, Ser[1]	CCG Pro	CAG Gln	CGG Arg, Usp[8]
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC Asn	AGC Ser
AUA Ile, Usp[2]	ACA Thr	AAA Lys	AGA Arg, Usp[9]
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU Ala	GAU Asp	GGU Gly
GUC Val	GCC Ala	GAC Asp	GGC Gly
GUA Val	GCA Ala Res[10]	GAA Glu	GGA Gly
GUG Val	GCG Ala	GAG Glu	GGG Gly

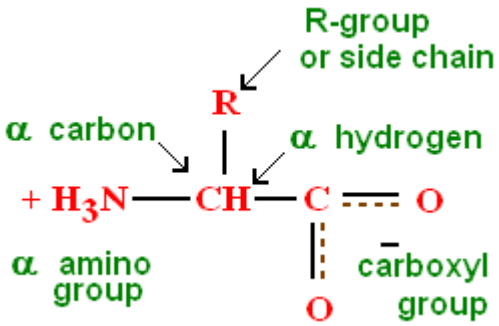
Figure 1.1.1. **The standard genetic code and known variant nuclear codes**. (1) Candida, a unicellular yeast. (2) Micrococcus. (3) ciliated protozoans and green algae. (4) Mycoplasma. (5) suppressor codon in bacteria. (6) Euplotes. (7) the selenocysteine codon (8) Spiroplasma. (9) Micrococcus. (10) resume codon in ssrA RNA (Lehman 2001).

INTRO to PROTEIN SCIENCE



From Wikimedia Commons

Amino acids: the building blocks



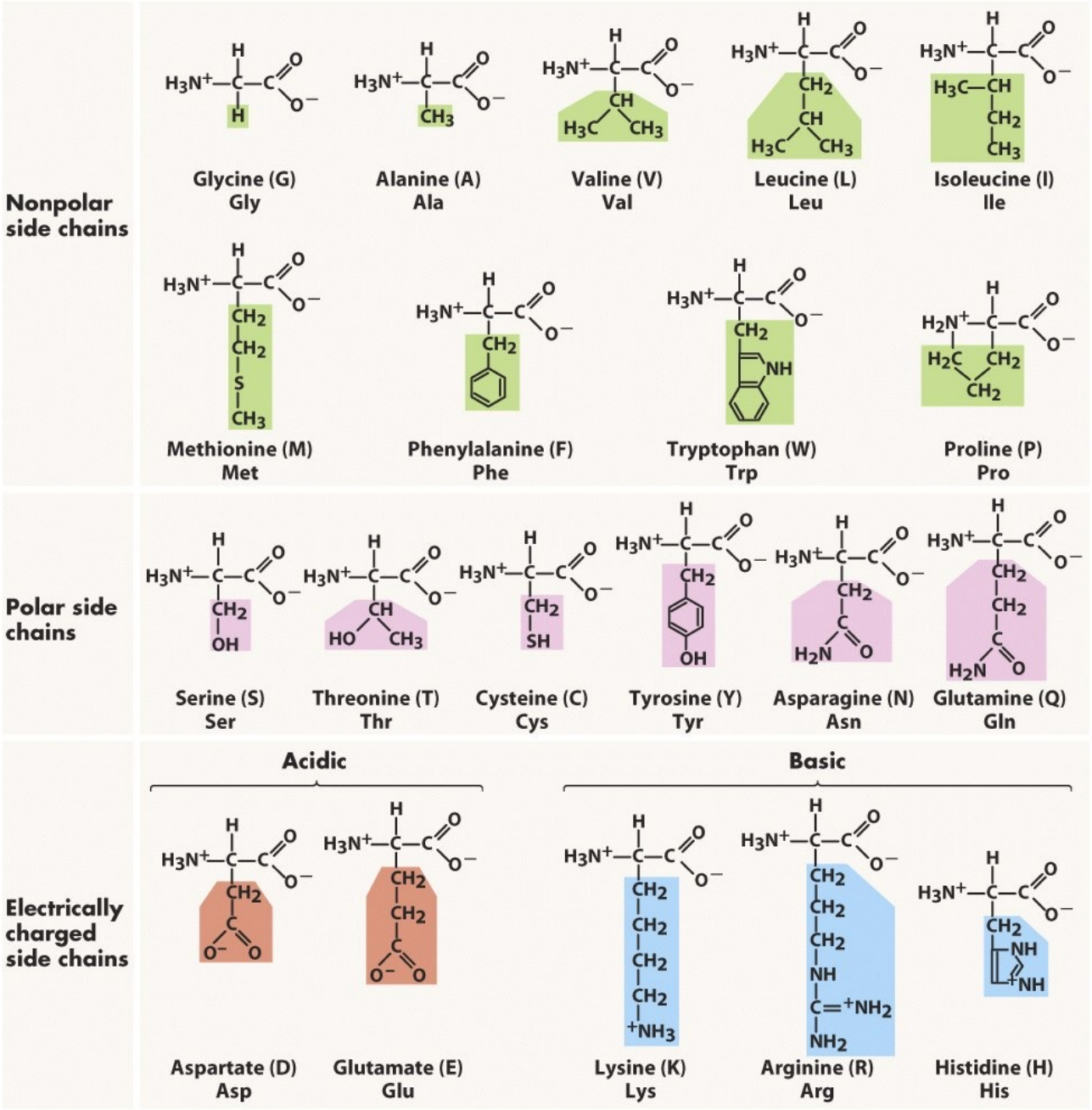
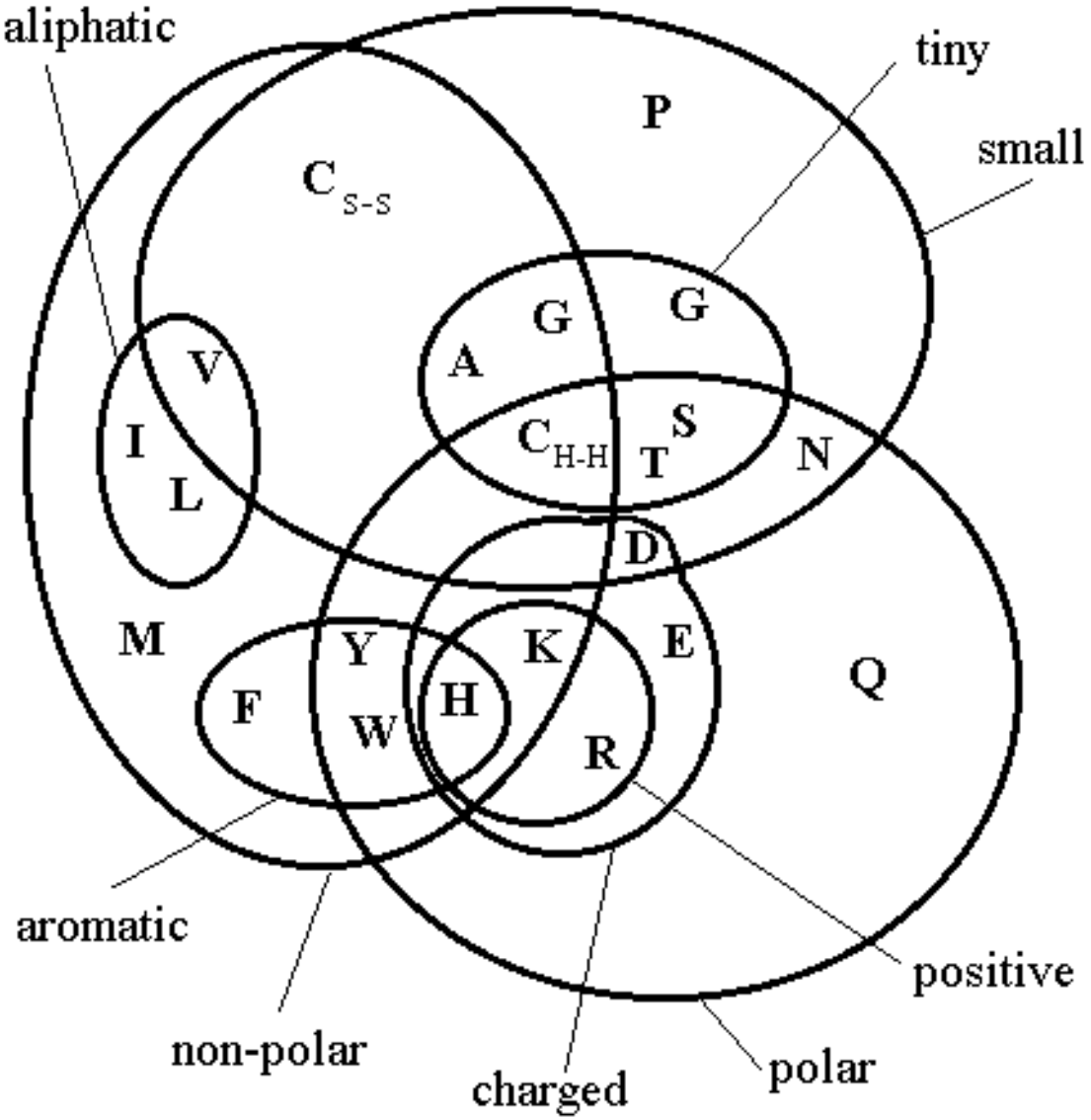
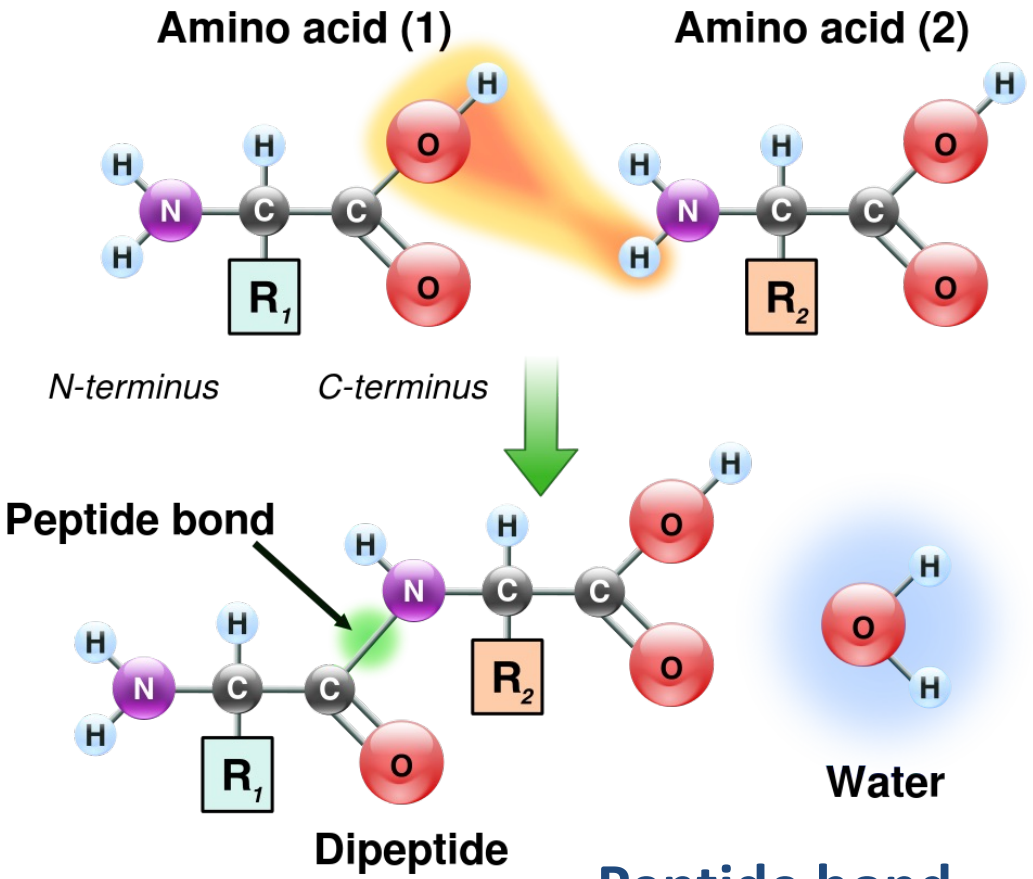


Figure 3-5 Biological Science, 2/e

Amino acids

physical-chemical properties

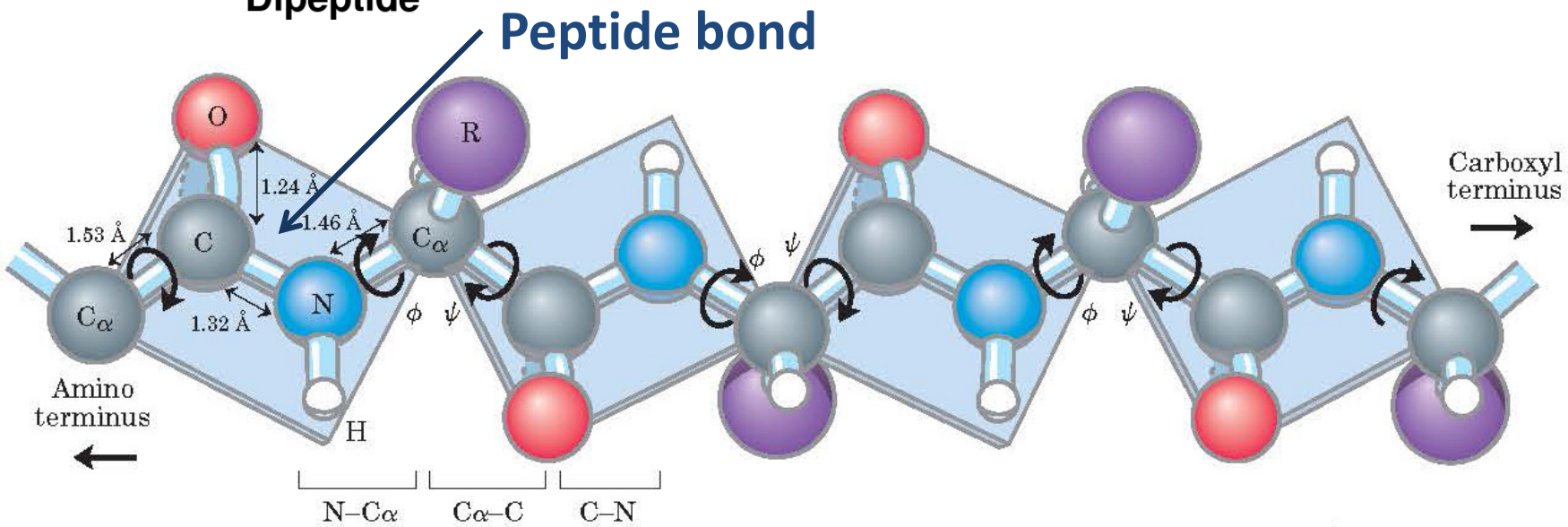


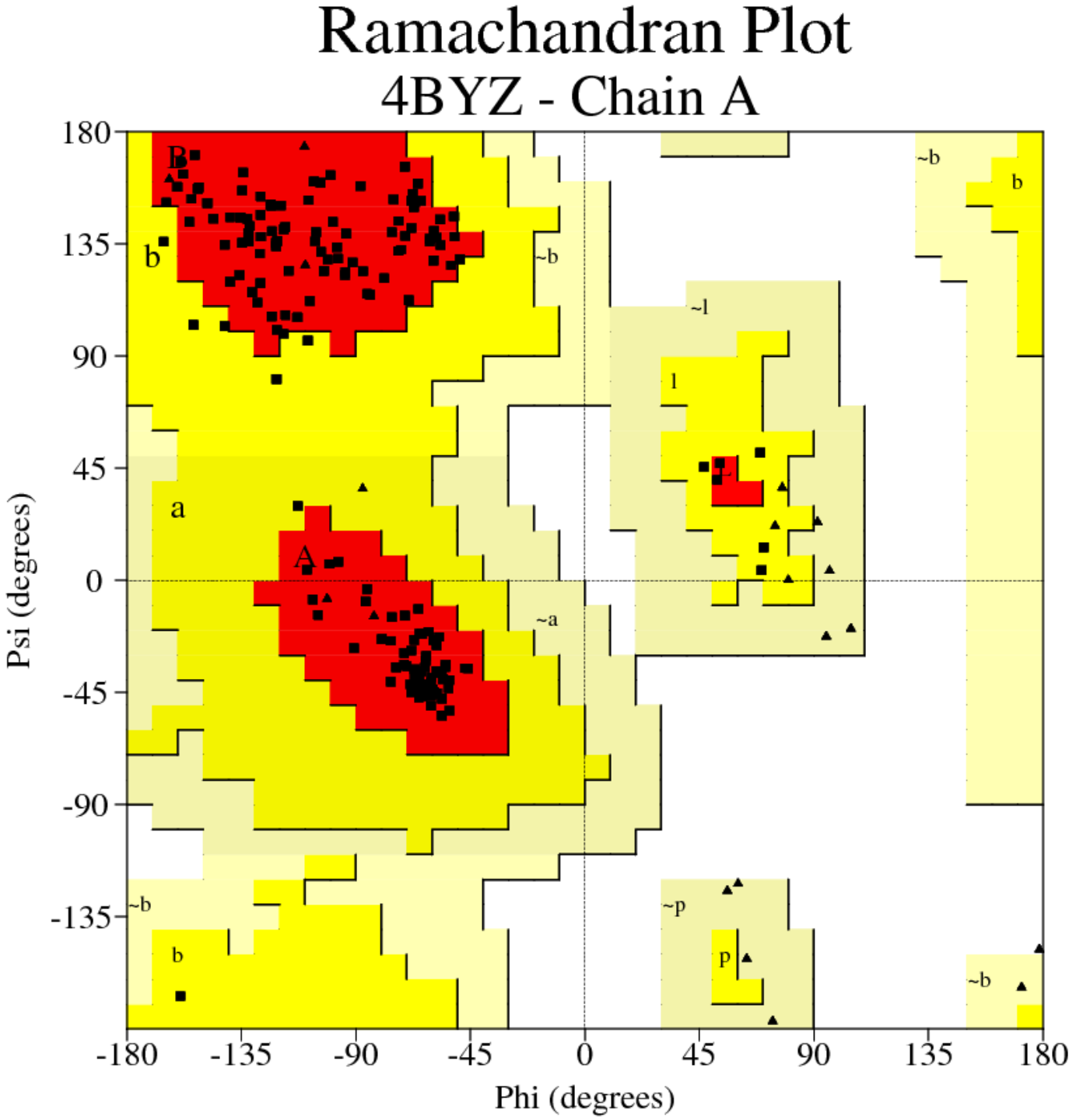


N,CA,C,O: Main chain or backbone.

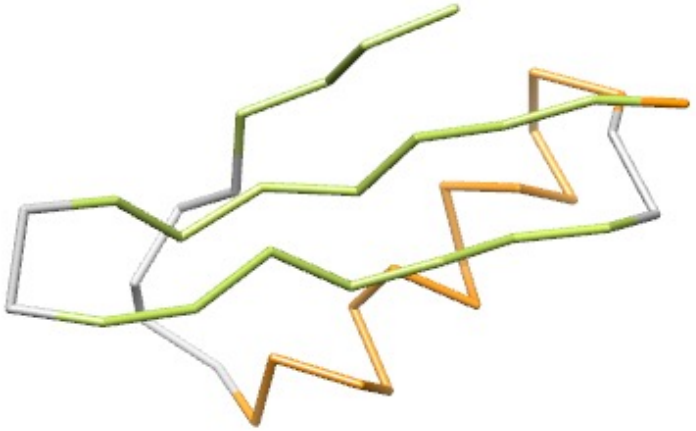
Peptide bond cannot rotate.

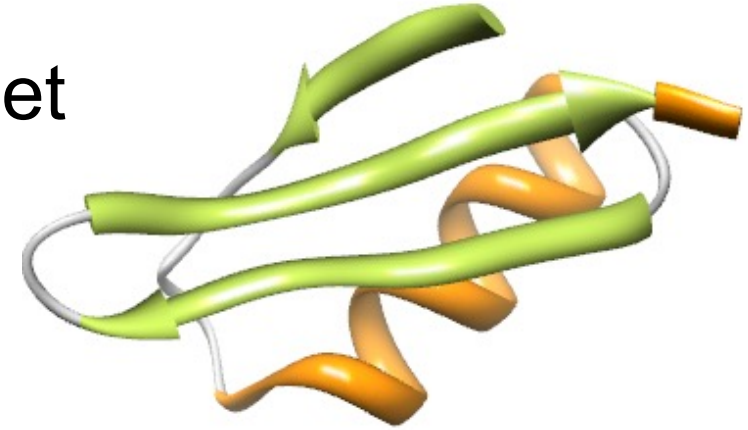
N-CA: phi free rotation dihedral angle
Ca-C: psi free rotation dihedral angle

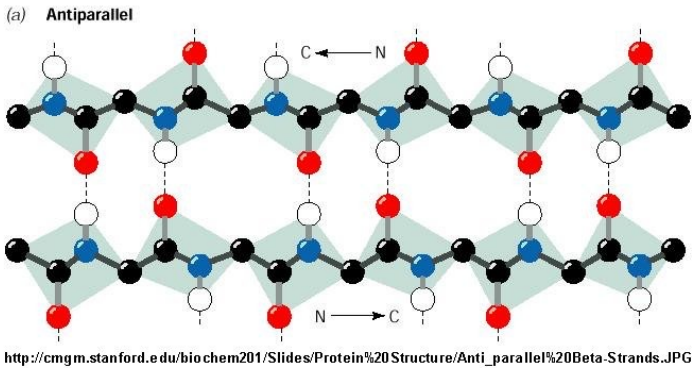



Ramachandran Plot
4BYZ - Chain A
Psi (degrees)
Phi (degrees)
180
135
90
45
0
-45
-90
-135
-180
-135
-90
-45
0
45
90
135
180
B
b
a
A
l
p
~b
~l
~a
~p

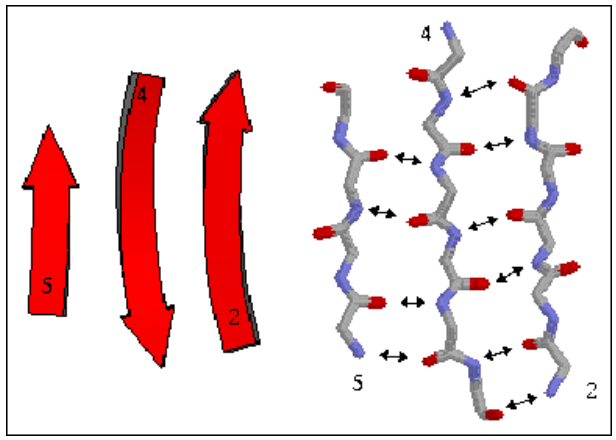
Beta - sheet

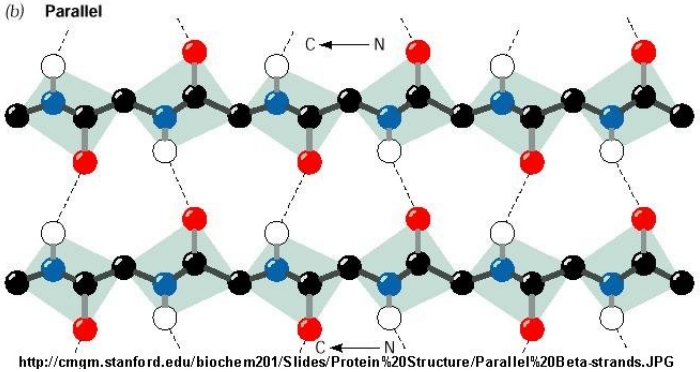




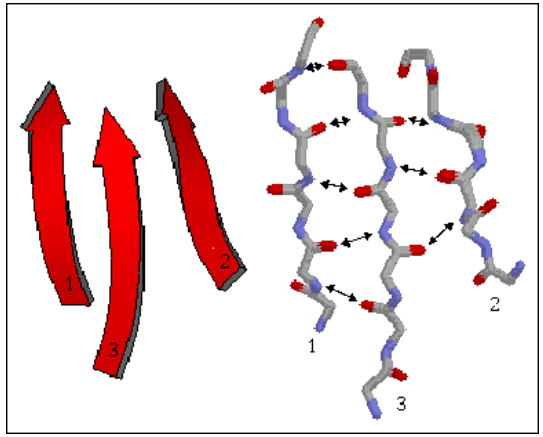


http://cmgm.stanford.edu/biochem201/Slides/Protein%20Structure/Anti_parallel%20Beta-Strands.JPG

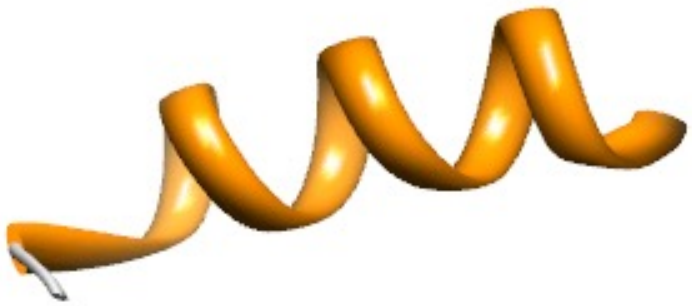


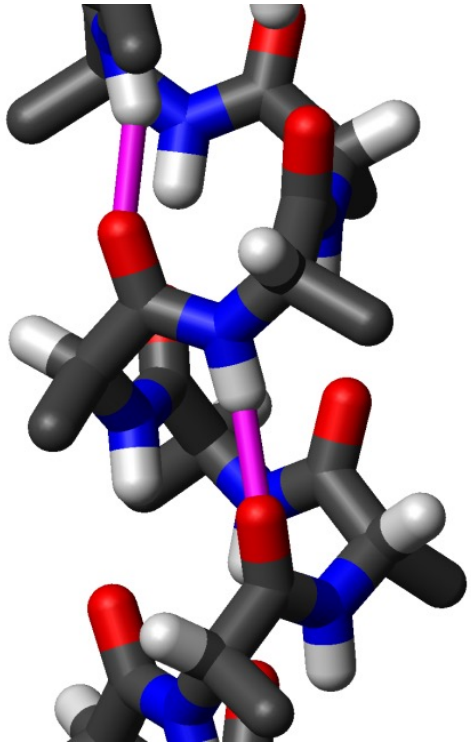


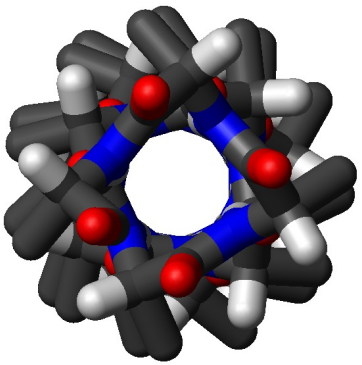
http://cmgm.stanford.edu/biochem201/Slides/Protein%20Structure/Parallel%20Beta-strands.JPG

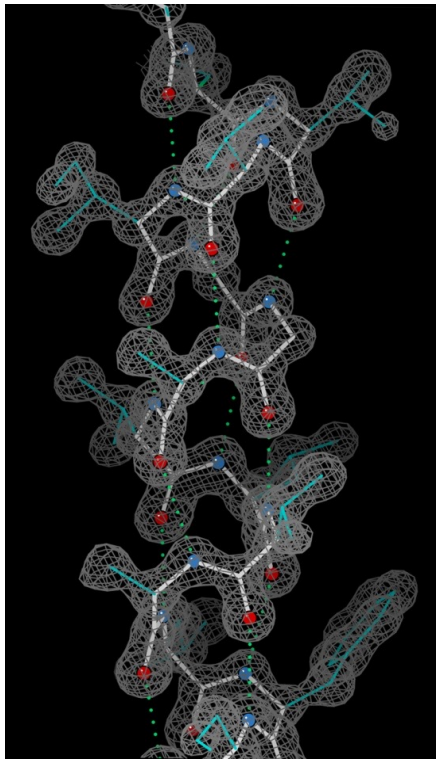


Alpha - helix

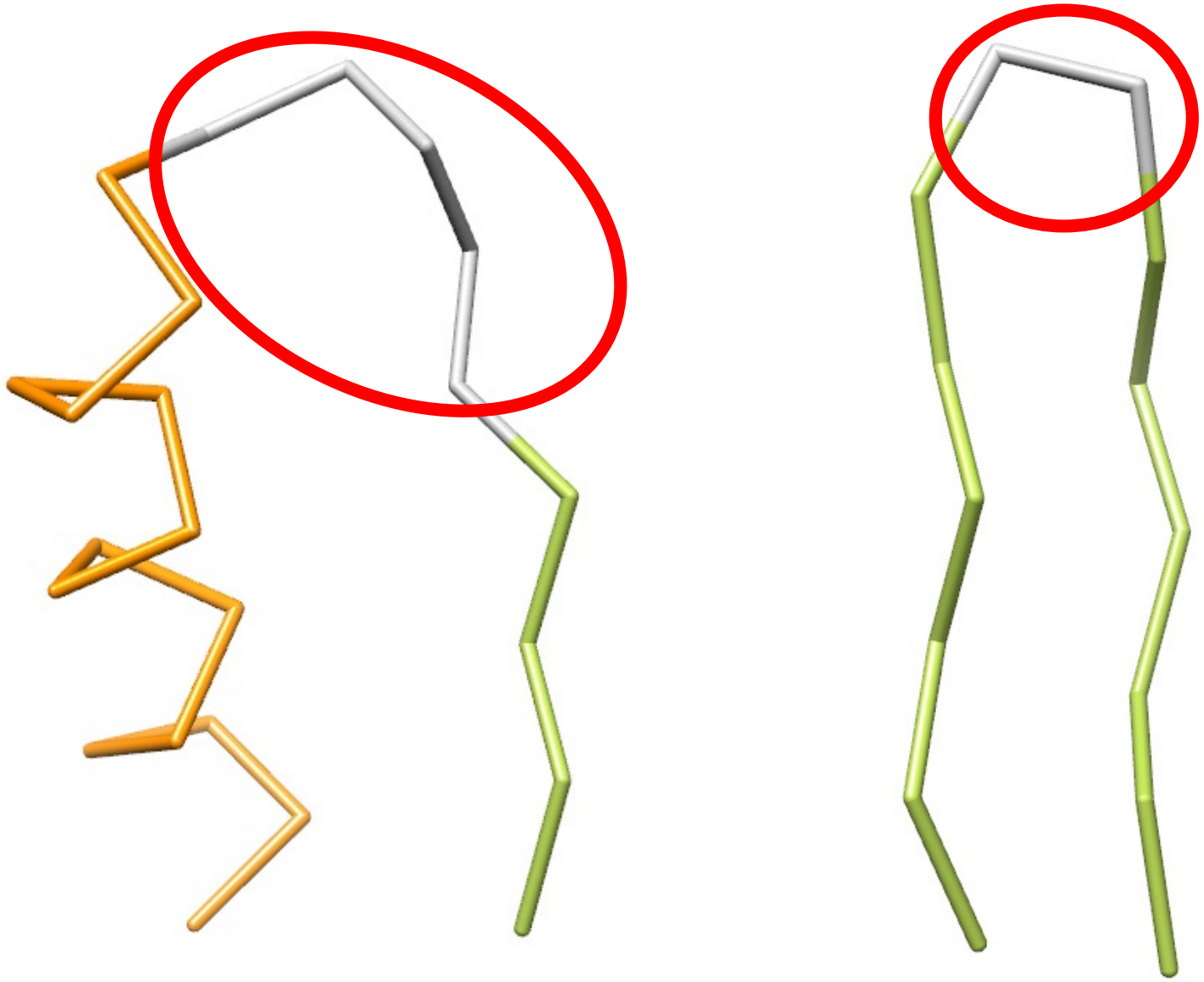








Loop: Residues between secondary structures



Large variety of protein structures

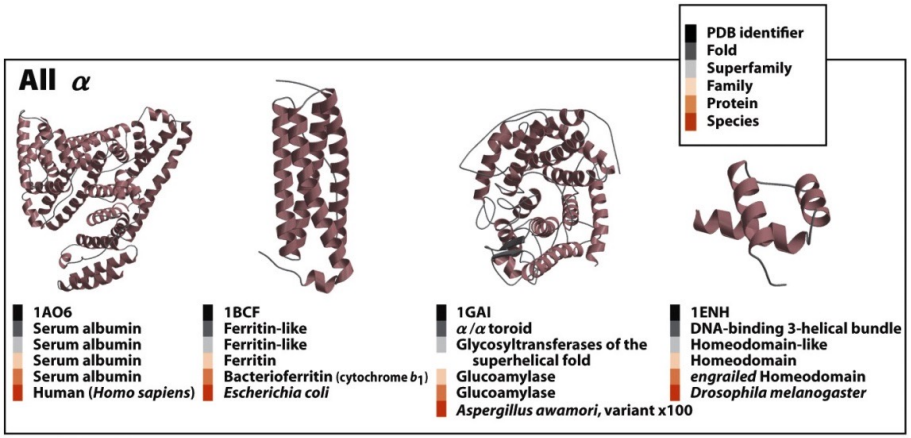


Figure 4-21 part 1
Lehninger Principles of Biochemistry, Fifth Edition

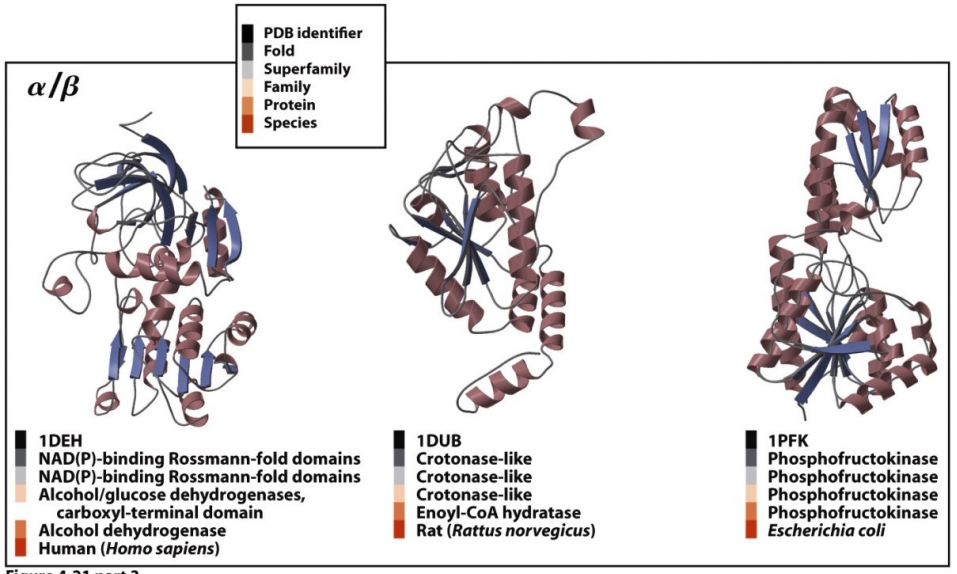


Figure 4-21 part 3
Lehninger Principles of Biochemistry, Fifth Edition

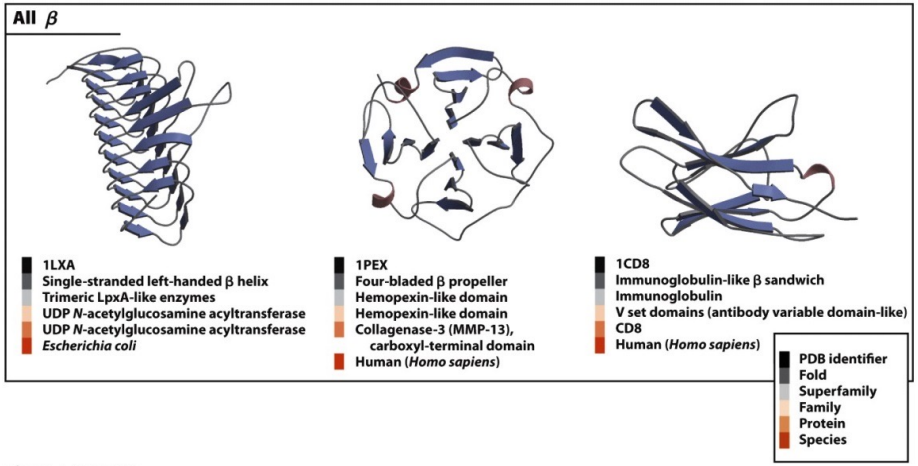


Figure 4-21 part 2
Lehninger Principles of Biochemistry, Fifth Edition

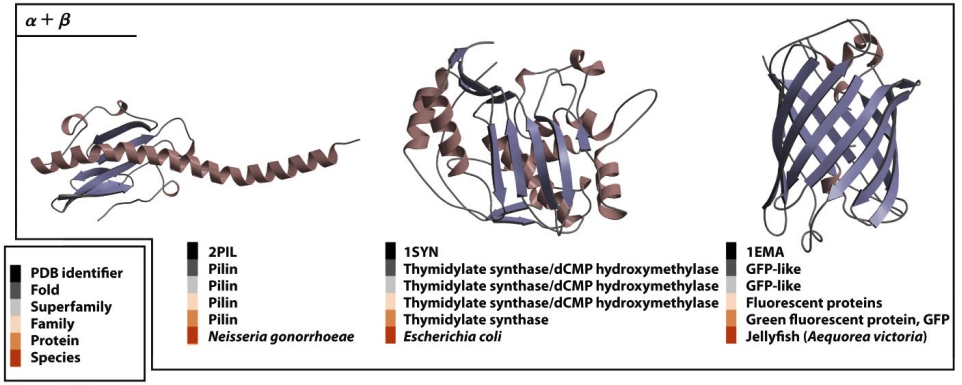
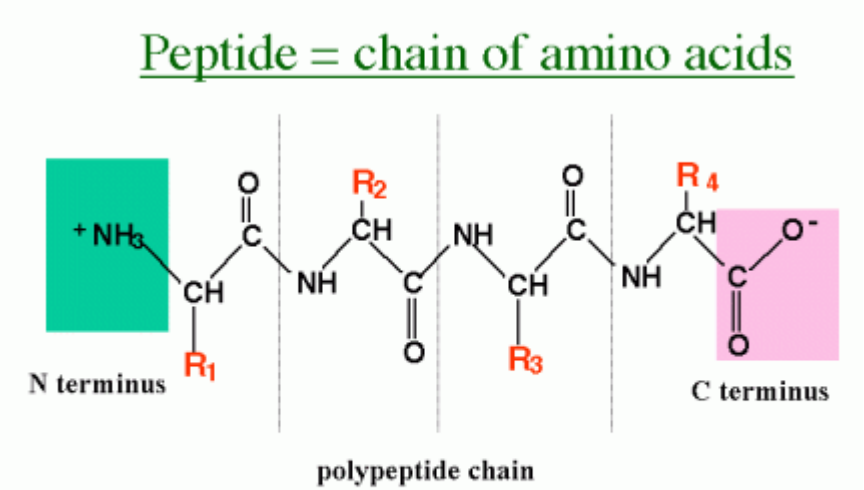


Figure 4-21 part 4
Lehninger Principles of Biochemistry, Fifth Edition

The structure levels in proteins

The primary structure is the "one-dimensional" sequence of amino acids in the protein or peptide.




```
>NP_000508.1
MVLSPADKTNVKAAWGKVGAHAGEYGAEALER
MFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVA
DALTNAVAHVDDMPNALSALSDLHAHKLRVDP
VNFKLLSHCLLVTLAAHLPAEFTPAVHASLDK
FLASVSTVLTSKYR
```

The structure levels in proteins

Secundary structure: Regular structural unit formed by sequence regions in a protein (from a sequence perspective: two-dimensional). Basically, these are of two types: α-helices and β-sheets.

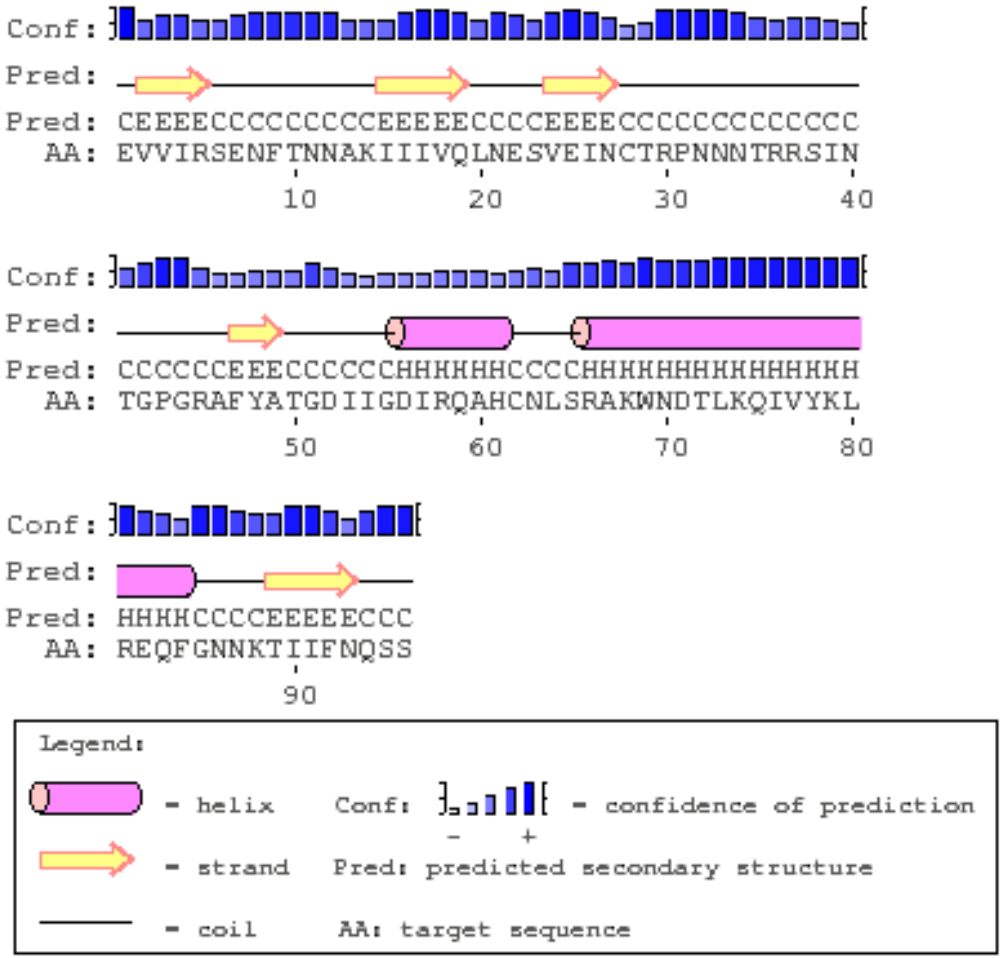
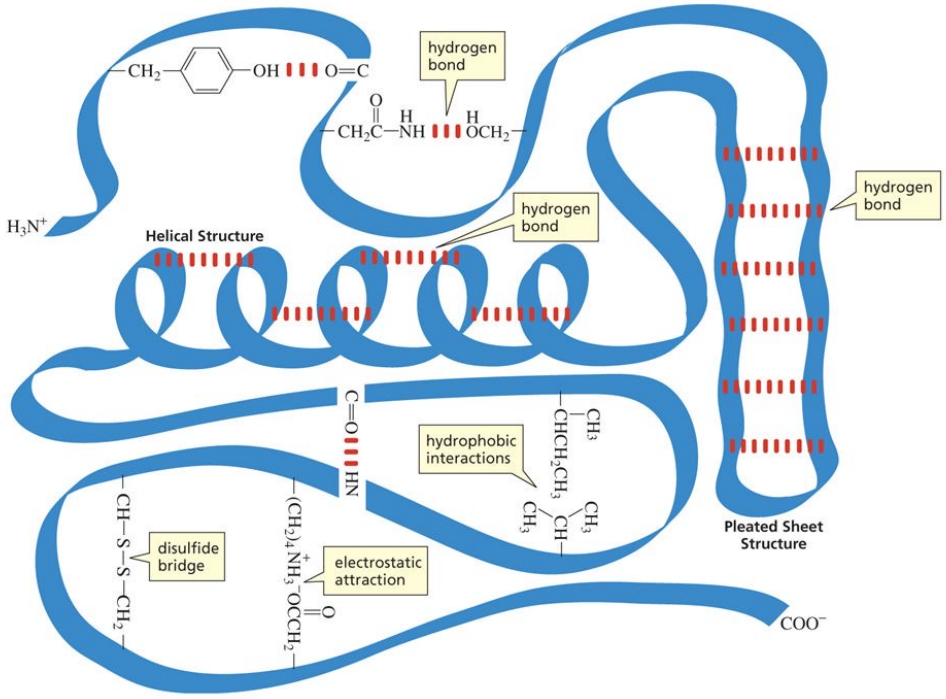
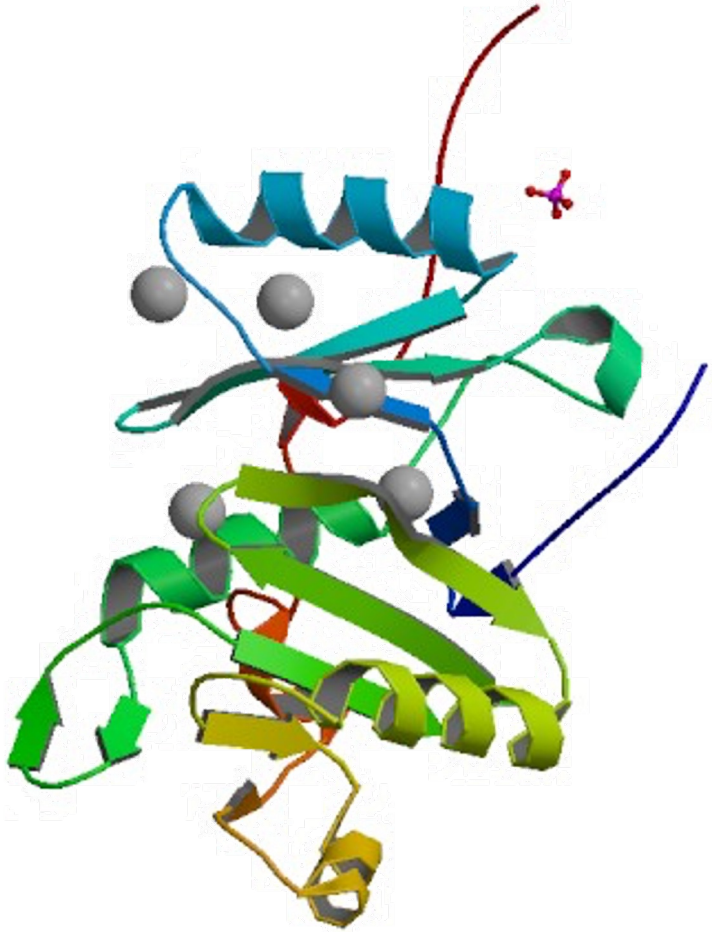


Image source: https://openwetware.org/wiki/BIOL368/F14:Isabel_Gonzaga_Week_9

The structure levels in proteins

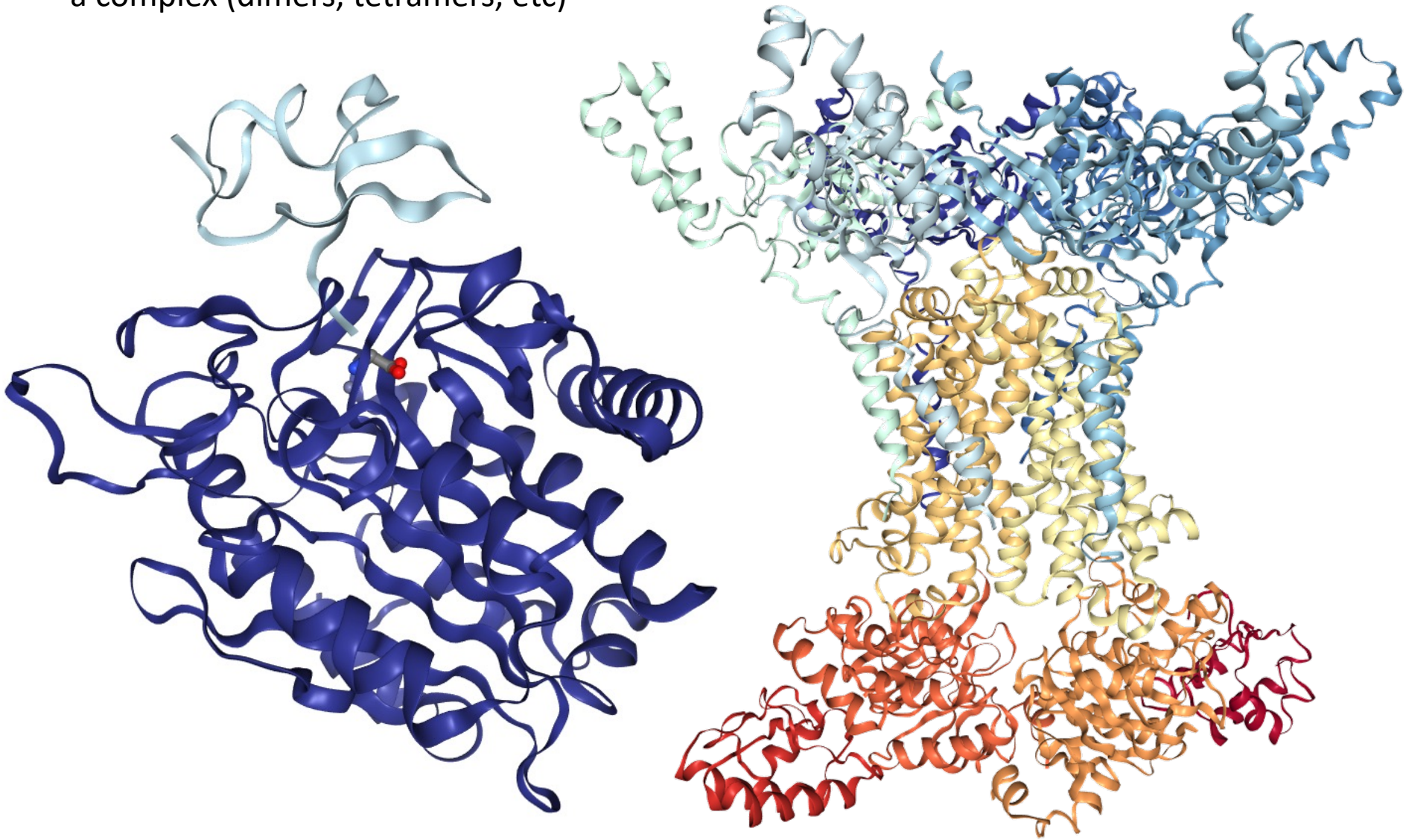
Tertiary structure is the three-dimensional fold of the protein, or how the secondary structure elements are packed against each other.



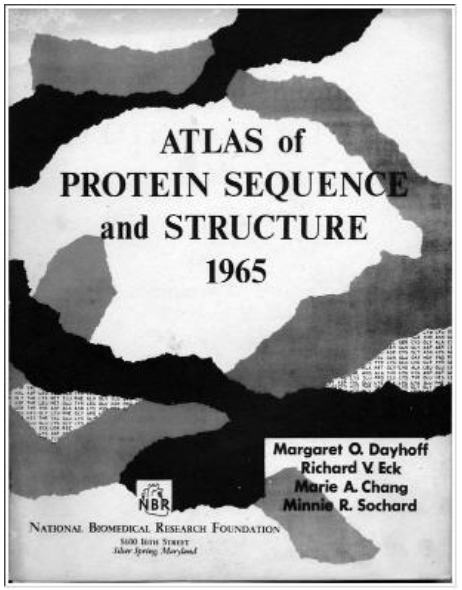


The structure levels in proteins

Quaternary structure is the arrangement in the space of multiple peptide chains to form a complex (dimers, tetramers, etc)



Historical aspects of major protein databases



Atlas of Protein Sequence and Structure, published from 1965-1978 under the editorship of Margaret O. Dayhoff

The Protein Information Resource (PIR) created in 1984

http://pir.georgetown.edu/pirwww/index.shtml

In 2002 PIR, along with its international partners, EBI (European Bioinformatics Institute) and SIB (Swiss Institute of Bioinformatics), were awarded a grant from NIH to create

UniProt

http://www.uniprot.org/

Major protein databases:

Sequence databases:

Primary (direct CDS translations from nucleotide databases)

GenBank → **GenPept** (Protein database at NCBI)

ENA → **TrEMBL**

Swiss-Prot (https://www.uniprot.org/uniprot/?query=keyword:KW-0903)

https://www.uniprot.org/keywords/KW-0903

Curated

RefSeq (NCBI)

Swiss-Prot (EMBL-EBI)

PIR protein information resource Georgetown University

Integrated and knowledge

UniProt = Swiss-Prot + TrEMBL + PIR

Structure database (not only proteins!!!)

PDB Protein Databank sequences from structures

NCBI protein

https://www.ncbi.nlm.nih.gov/protein/

NCBI Resources How To Sign in to NCBI

Protein Protein Search

Advanced Help

Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and

Using Protein

- Quick Start Guide
- FAQ
- Help

Protein Tools

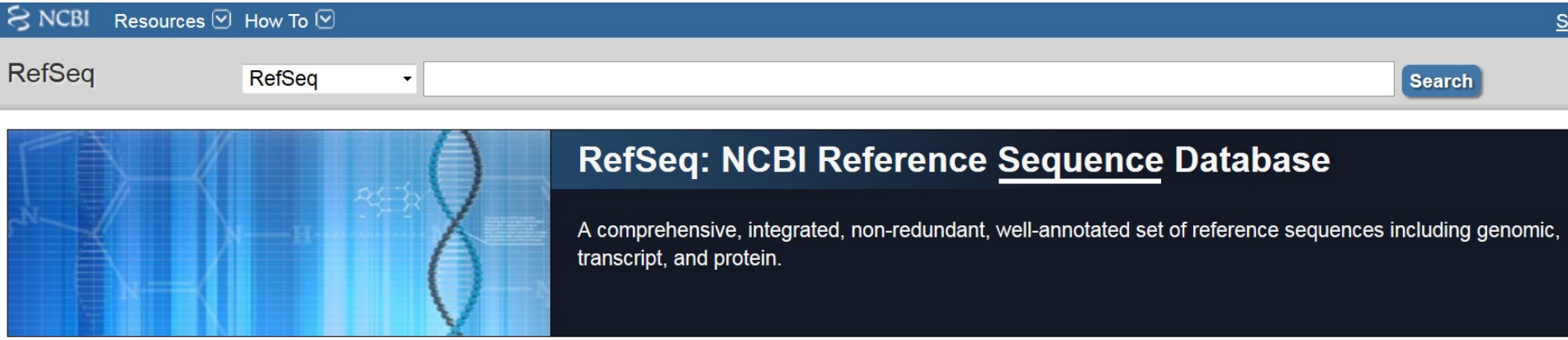
- BLAST
- LinkOut
- E-Utilities

Other Resources

- GenBank Home
- RefSeq Home
- CDD

NCBI RefSeq

https://www.ncbi.nlm.nih.gov/refseq/

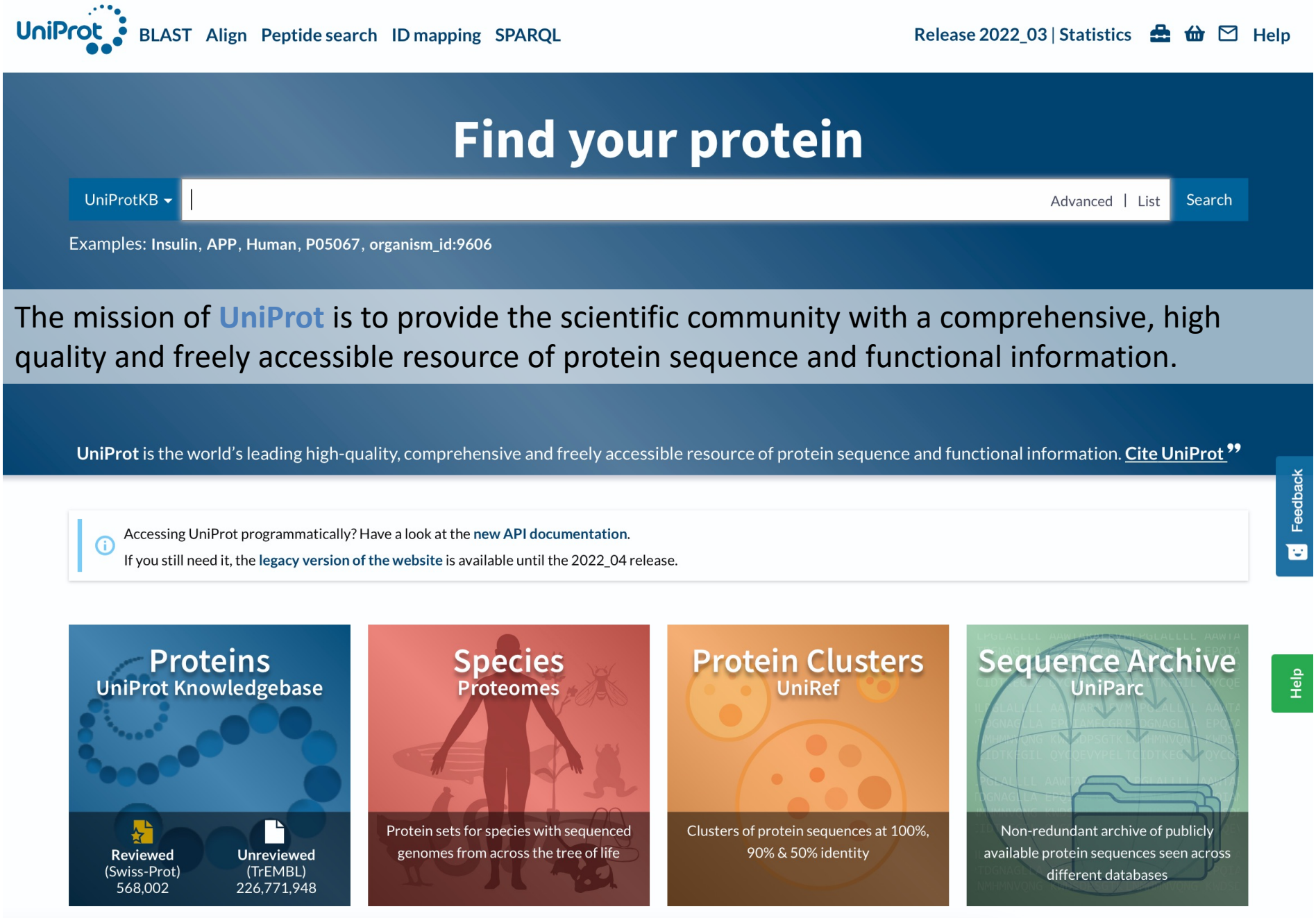


genomes ⇢ transcripts ⇢ proteins

- non-redundant; best representative
- updates to reflect current sequence data and biology
- distinct, stable accession series

WARNING:
Non redundancy removes species origin

https://www.ncbi.nlm.nih.gov/protein/WP_005414279.1
https://www.ncbi.nlm.nih.gov/ipg/WP_005414279.1



http://www.uniprot.org/

UniProtKB http://www.uniprot.org/

• The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on **proteins**, with accurate, consistent and rich annotation.

- amino acid sequence,
- protein name or description,
- taxonomic data
- citation information
- functional annotations
- as much annotation information as possible is added.

• "**UniProtKB/Swiss-Prot**" is a high quality manually annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions.

• "**UniProtKB/TrEMBL**" computationally analyzed records derived from the translation of the coding sequences (CDS) that await full manual annotation.

- Once in UniProtKB/Swiss-Prot, a protein entry is removed from UniProtKB/TrEMBL.

- About 98 % of the protein sequences provided by UniProtKB are derived from the translation of the CDS which have been submitted to the public nucleic acid databases, the EMBL-Bank/GenBank/DDBJ databases (INSDC).

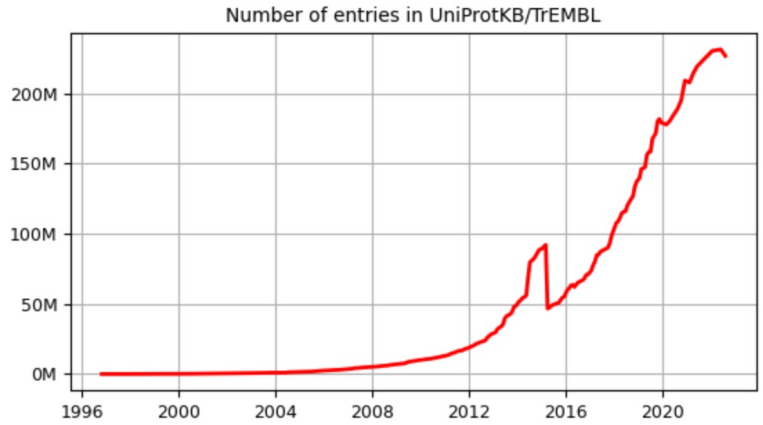
Search for Field Descriptions and Tags in your Database

NCBI	UNIPROT
NC_0000*[Accession]	accession:P62988
Human[Organism]	active:false
horse[taxonomy]	lit_author:ashburner
neoplasms[MeSHTerms]	protein_name:CD233
prolactin[Protein Name]	chebi:18420
APOE[gene]	xref_count_pdb:[20 TO *]
srcdb_refseq[Properties]	date_created:[2012-10-01 TO *]
2010/06[Publication Date]	database:pfam xref:pdb-1aut
110:500[Sequence Length]	ec:3.2.1.23
gene_symbol[sym]	existence:3
1.1.1.53[ecno]	family:serpin
gbdiv_est[PROP]	gene:HPSE
: : :	: : :
etc	etc

http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers
https://www.ncbi.nlm.nih.gov/books/NBK49540/
http://www.uniprot.org/help/query-fields <- chek here "active" current and legacy
https://www.uniprot.org/help/api_queries
https://www.uniprot.org/help/programmatic_access

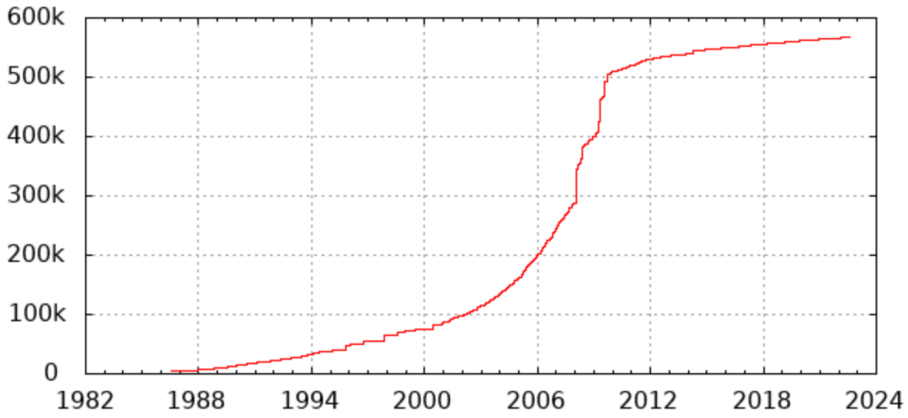
Growth of the UniProt database

Release 2022_03 of 03-Aug-2022 of UniProtKB/TrEMBL contains 226771949 sequence entries.



http://www.ebi.ac.uk/uniprot/TrEMBLstats

Release 2022_03 of 03-Aug-2022 of UniProtKB/Swiss-Prot contains 568002 sequence entries.



https://web.expasy.org/docs/relnotes/relstat.html

How redundant are sequences in UniProtKB?

Non-redundancy means in:

UniProtKB/TrEMBL: one record for 100% identical full-length sequences in one species;

UniProtKB/Swiss-Prot: one record per gene in one species;

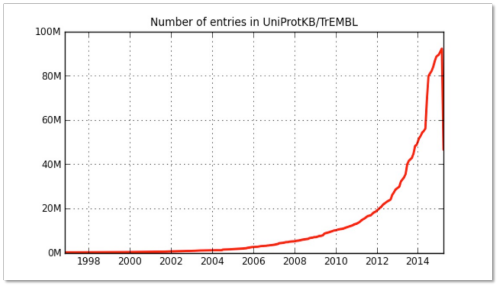
UniParc: one record for 100% identical sequences over the entire length, regardless of the species.

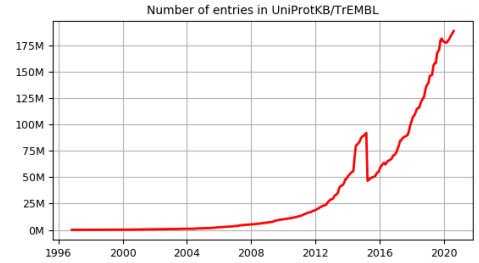
UniRef100: one record for 100% identical sequences, including fragments, regardless of the species.

'Non-redundancy' in UniProtKB/Swiss-Prot implies that identical sequences are represented in a single record. However, if identical protein sequences are produced in different species, they are stored in different records (e.g. ubiquitin).

Proteome redundancy removal

https://www.uniprot.org/help/proteome_redundancy
https://insideuniprot.blogspot.com/2015/05/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5199198/



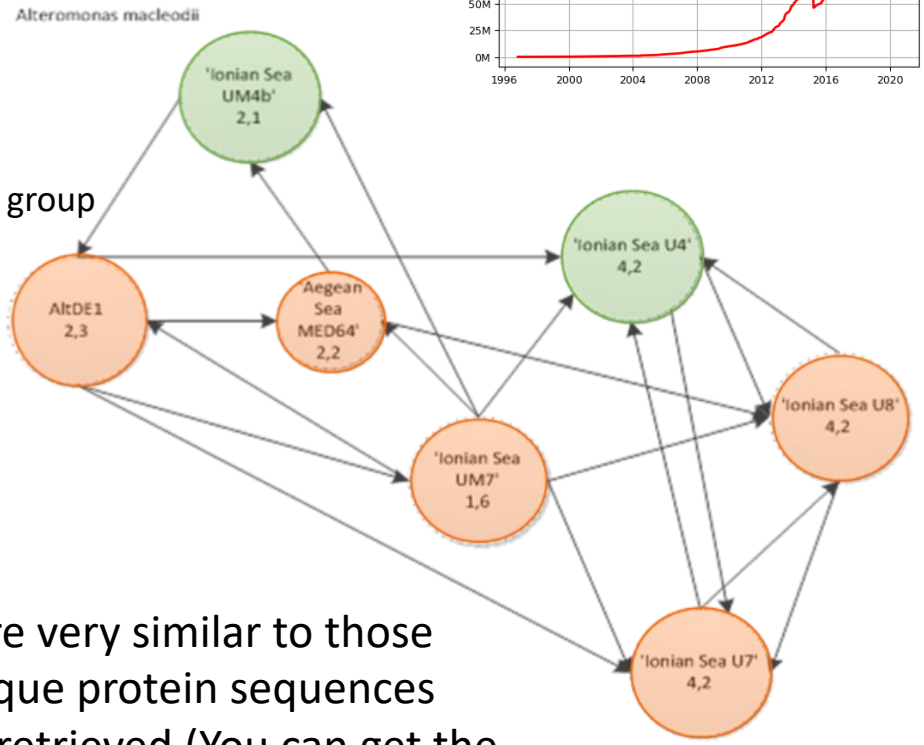


Step 1: Group proteomes by taxonomy level

Step 2: Pairwise comparison of proteomes within each group

Step 3: Graph analysis

Step 4: Elimination of redundant proteomes

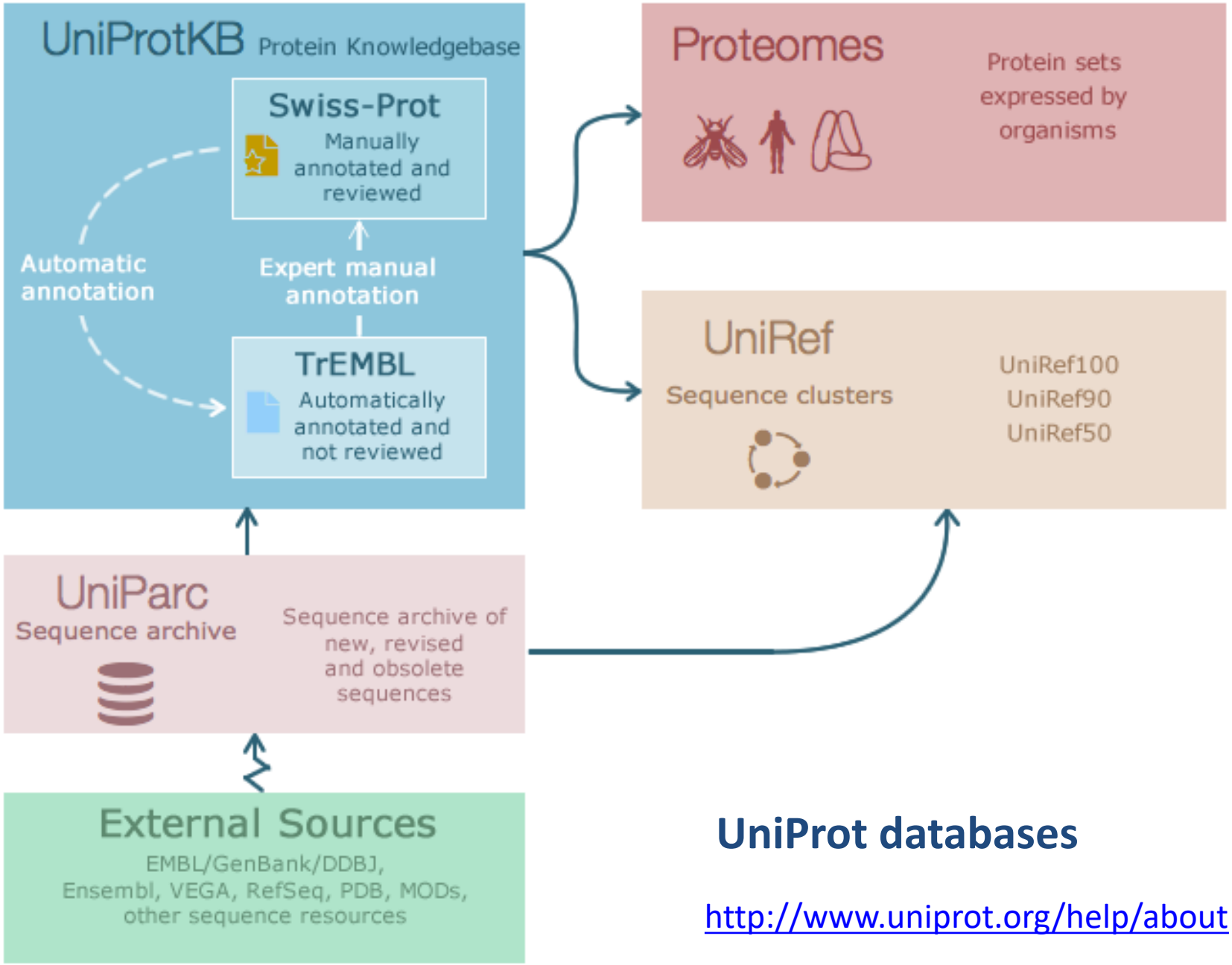


Although removed proteomes are very similar to those that are not, they all contain unique protein sequences which UNIPROT data can not be retrieved (You can get the sequence from UniParc)

What can we get from Uniprot

https://www.uniprot.org/uniprot/P0DTC2

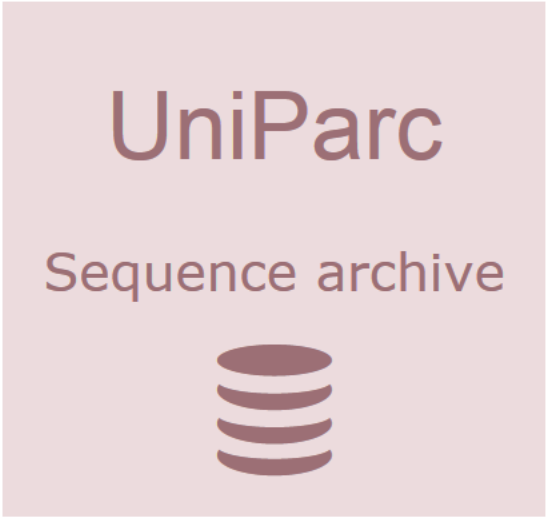
- Publications
- Graphical view
- Function
- Names (synonyms)
- Origin Organism, genes …
- Location
- Topology
- Structure
- Known mutations
- PTM
- Sites
- Interaction
- Motifs, Domains and Regions
- Expression
- Etc …
- Cross-references to other databases (including source of previous data)



UniProt databases

http://www.uniprot.org/help/about

UniParc archive



It contains only protein sequences (without annotations)

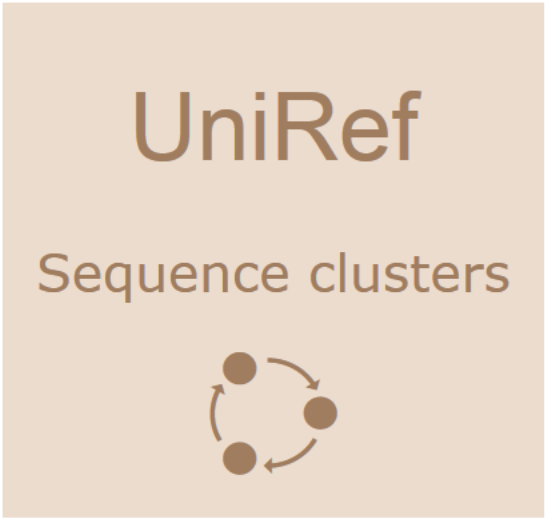
It provides cross-references to the source databases

Non-redundant archive of protein sequences from the public databases

- Non-redundancy means each unique protein sequence is stored only once and is assigned a unique stable UniParc identifier (UPI)

See http://www.uniprot.org/help/uniparc

UniRef – UniProt redundancy reducing system for protein sequences



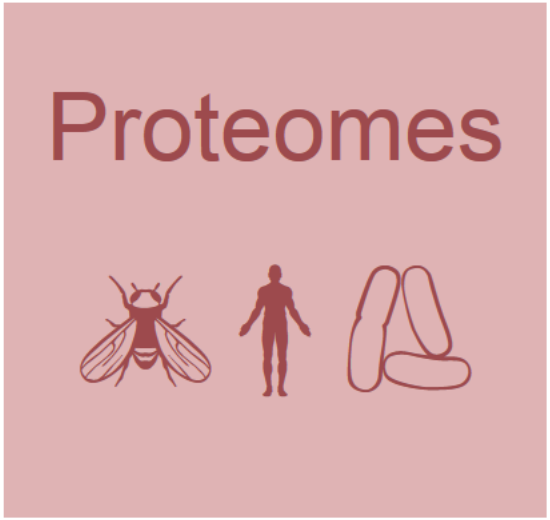
It provides non-redundant protein sequences from UniProt allowing faster and more informative sequence similarity searches

Hiding redundant sequences by clustering them

- UniRef100 = Identical sequences and sub-fragments
- UniRef90 = 90% identical UniRef100 clusters
- UniRef50 = 50% identical UniRef90 clusters

See http://www.uniprot.org/help/uniref

UniProt proteomes



It provides proteome sets of proteins whose genomes have been completely sequenced.

It includes both manually reviewed (Swiss-Prot) and unreviewed (TrEMBL) entries.

It detects highly redundant proteomes within species groups which are deleted from UniProtKB

See http://www.uniprot.org/help/proteome

Proteome = set of proteins thought to be expressed by an organism

Evidences that supports the existence of a protein

- Evidence at protein level
- Evidence at transcript level
- Inferred from homology
- Predicted
- Uncertain

Putative, uncharacterized and hypothetical, are also used for a protein with predicted or unknown function at any level of evidences.
But take good care with it, names usually do not get updated. Ex: https://www.uniprot.org/uniprotkb/Q9HVW4

http://www.ebi.ac.uk/uniprot/TrEMBLstats

http://www.uniprot.org/help/protein_existence

Determining protein function

- Experimentally determining function (time consuming and expensive)
    - Functional assays
    - **3D structure**
    - Co-localization
    - Protein-protein interaction (experimental approaches)
    - Microarray experiments

- Predictive methods
    - Transfer of function based on homology
        - **BLAST**
    - Based on the amino acid sequence (predicted protein secondary structure, transmembrane helices, subcellular localization, and posttranslational modifications)
    - 3D modeling
    - Motif and Domains

- Combination of all above

Determining 3D structure of proteins

Experimentally

- X-ray Crystallography
- NMR Spectroscopy
- Electron Microscopy

Predictive

- Homology modeling
  - It uses a known experimental structure of a homologous protein (the template)
- Others (eg: Threading)
- THE NEW ONE: AlphaFold

https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction

Protein structure databases

- Protein Data Bank PDB https://www.rcsb.org/pdb

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization http://www.wwpdb.org/

  - PDB-101 is an online portal for education http://pdb101.rcsb.org/
  - Protein Data Bank in Europe PDBe http://www.ebi.ac.uk/pdbe/

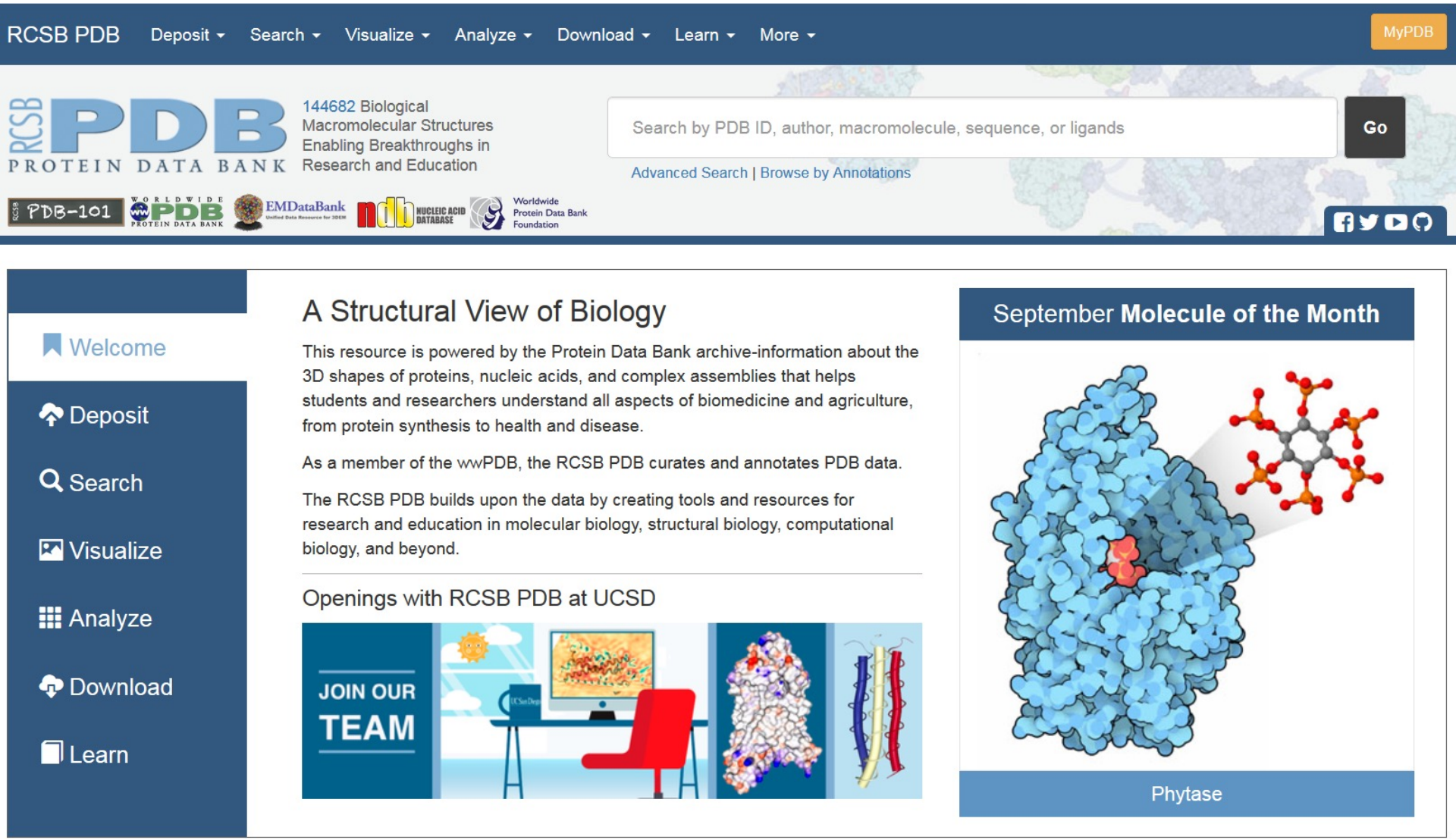
- NCBI protein structures https://www.ncbi.nlm.nih.gov/structure/
- Protein Model Portal (PMP) http://www.proteinmodelportal.org/
- The SWISS-MODEL Repository https://swissmodel.expasy.org/repository/

Annotated 3D protein structure models generated by the SWISS-MODEL homology-modelling pipeline.

- Others (https://en.wikipedia.org/wiki/Protein_structure_database)
- AlphaFoldDB: https://alphafold.ebi.ac.uk

The Protein Data Bank (PDB)

https://www.rcsb.org/pdb



PDB history at
https://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/index.html

The Protein Data Bank (PDB)

https://www.rcsb.org/pdb

IT IS NOT A PROTEIN DATABASE!

PDB history at
https://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/index.html

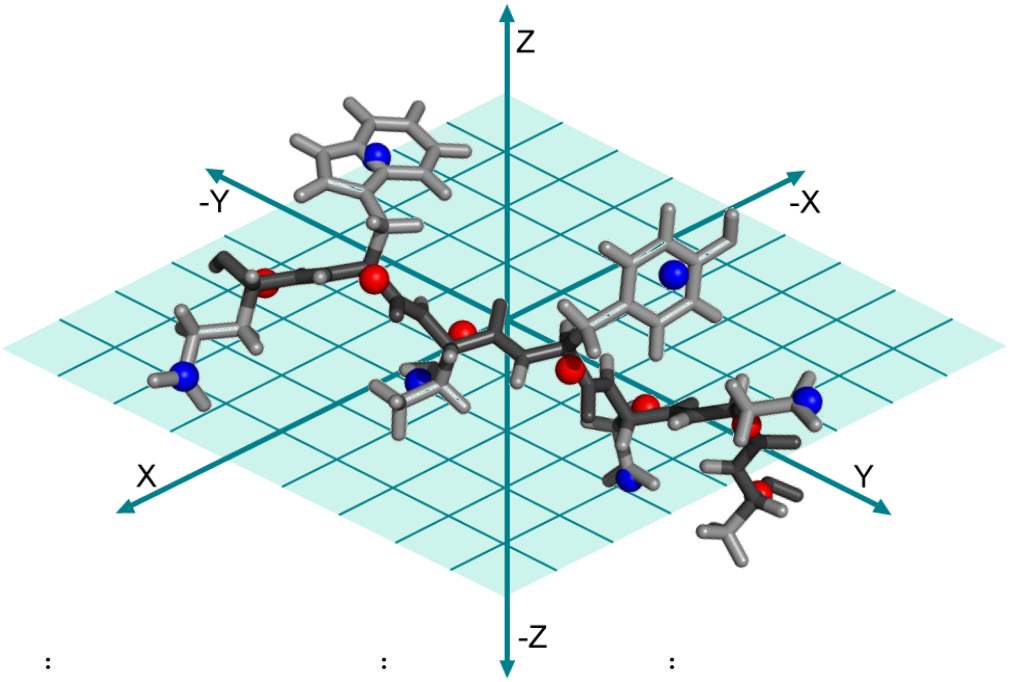
Utility of protein 3D models

- Annotate protein function
- Functional site of proteins
- Templates for homology modeling (make better predictions)
- Protein-protein interactions
- Protein complexes and macromolecular structures
- Discover new drugs or new targets
- ……..

A typical PDB format file

http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html

The primary information stored in the PDB archive consists of coordinate files for biological molecules. These files list the atoms in each protein, and their 3D location in space.




```
   :              :                    :                 :
   :              :                    :                 :
ATOM      1  N    VAL A   1      -0.686  14.852  16.090  1.00 42.31           N
ATOM      2  CA   VAL A   1       0.621  15.543  16.081  1.00 53.86           C
ATOM      3  C    VAL A   1       1.823  14.549  16.200  1.00 27.97           C
ATOM      4  O    VAL A   1       1.628  13.369  15.953  1.00 23.24           O
ATOM      5  CB   VAL A   1       0.521  16.582  17.208  1.00 36.55           C
ATOM      6  CG1  VAL A   1      -0.444  17.695  16.811  1.00 35.23           C
ATOM      7  CG2  VAL A   1       0.119  15.972  18.559  1.00 53.38           C
ATOM      8  N    LEU A   2       3.112  14.944  16.272  1.00 13.35           N
ATOM      9  CA   LEU A   2       4.140  13.989  16.663  1.00 16.70           C
ATOM     10  C    LEU A   2       3.978  13.813  18.182  1.00 18.20           C
ATOM     11  O    LEU A   2       3.770  14.833  18.864  1.00 17.87           O
   :              :                    :                 :
   :              :                    :                 :
```

list of the atoms and their coordinates

http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html#ATOM

A typical PDB format file

```
HEADER    OXYGEN STORAGE                          27-AUG-81   1MBO
TITLE     STRUCTURE AND REFINEMENT OF OXYMYOGLOBIN AT 1.6 ANGSTROMS RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MYOGLOBIN;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: PHYSETER CATODON;
SOURCE   3 ORGANISM_COMMON: SPERM WHALE;
SOURCE   4 ORGANISM_TAXID: 9755
KEYWDS    OXYGEN STORAGE
EXPDTA    X-RAY DIFFRACTION
AUTHOR    S.E.V.PHILLIPS
REMARK   2 RESOLUTION.    1.60 ANGSTROMS.
   :             :
   :             :
HELIX    6   F LEU A   86  THR A   95  1                                   10
HELIX    7   G PRO A  100  ARG A  118  1                                   19
HELIX    8   H GLY A  124  LEU A  149  1                                   26
LINK         NE2 HIS A  93                FE   HEM A 155     1555   1555  2.07
LINK        FE   HEM A 155                 O1  OXY A 555     1555   1555  1.83
SITE     1 AC1  4 SER A  58  GLU A  59  ASP A  60  HOH A 227
SITE     1 AC2 22 THR A  39  LYS A  42  PHE A  43  ARG A  45
SITE     2 AC2 22 HIS A  64  VAL A  68  LEU A  89  SER A  92
   :             :                   :                :
   :             :                   :                :
ATOM      1  N   VAL A   1      -0.686  14.852  16.090  1.00 42.31           N
ATOM      2  CA  VAL A   1       0.621  15.543  16.081  1.00 53.86           C
ATOM      3  C   VAL A   1       1.823  14.549  16.200  1.00 27.97           C
ATOM      4  O   VAL A   1       1.628  13.369  15.953  1.00 23.24           O
ATOM      5  CB  VAL A   1       0.521  16.582  17.208  1.00 36.55           C
ATOM      6  CG1 VAL A   1      -0.444  17.695  16.811  1.00 35.23           C
ATOM      7  CG2 VAL A   1       0.119  15.972  18.559  1.00 53.38           C
ATOM      8  N   LEU A   2       3.112  14.944  16.272  1.00 13.35           N
ATOM      9  CA  LEU A   2       4.140  13.989  16.663  1.00 16.70           C
ATOM     10  C   LEU A   2       3.978  13.813  18.182  1.00 18.20           C
ATOM     11  O   LEU A   2       3.770  14.833  18.864  1.00 17.87           O
   :             :                   :                :
   :             :                   :                :
```

"header" section (protein, citations, details of the structure solution, and remarks.

keywords

Sequence information, secondary structure and binding sites

list of the atoms and their coordinates

https://www.rcsb.org/structure/1MBO
https://files.rcsb.org/view/1MBO.pdb

A typical PDB format file

```
HEADER    OXYGEN STORAGE                          27-AUG-81   1MBO
TITLE     STRUCTURE AND REFINEMENT OF OXYMYOGLOBIN AT 1.6 ANGSTROMS RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MYOGLOBIN;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: PHYSETER CATODON;
SOURCE   3 ORGANISM_COMMON: SPERM WHALE;
SOURCE   4 ORGANISM_TAXID: 9755
KEYWDS    OXYGEN STORAGE
EXPDTA    X-RAY DIFFRACTION
AUTHOR    S.E.V.PHILLIPS
REMARK   2 RESOLUTION.    1.60 ANGSTROMS.
   :             :
   :             :
HELIX    6   F LEU A   86  THR A   95  1                                  10
HELIX    7   G PRO A  100  ARG A  118  1                                  19
HELIX    8   H GLY A  124  LEU A  149  1                                  26
LINK         NE2 HIS A  93                FE   HEM A 155     1555   1555  2.07
LINK        FE   HEM A 155                 O1  OXY A 555     1555   1555  1.83
SITE     1 AC1  4 SER A  58  GLU A  59  ASP A  60  HOH A 227
SITE     1 AC2 22 THR A  39  LYS A  42  PHE A  43  ARG A  45
SITE     2 AC2 22 HIS A  64  VAL A  68  LEU A  89  SER A  92
   :             :                   :               :
   :             :                   :               :
ATOM      1  N   VAL A   1      -0.686  14.852  16.090  1.00 42.31           N
ATOM      2  CA  VAL A   1       0.621  15.543  16.081  1.00 53.86           C
ATOM      3  C   VAL A   1       1.823  14.549  16.200  1.00 27.97           C
ATOM      4  O   VAL A   1       1.628  13.369  15.953  1.00 23.24           O
ATOM      5  CB  VAL A   1       0.521  16.582  17.208  1.00 36.55           C
ATOM      6  CG1 VAL A   1      -0.444  17.695  16.811  1.00 35.23           C
ATOM      7  CG2 VAL A   1       0.119  15.972  18.559  1.00 53.38           C
ATOM      8  N   LEU A   2       3.112  14.944  16.272  1.00 13.35           N
ATOM      9  CA  LEU A   2       4.140  13.989  16.663  1.00 16.70           C
ATOM     10  C   LEU A   2       3.978  13.813  18.182  1.00 18.20           C
ATOM     11  O   LEU A   2       3.770  14.833  18.864  1.00 17.87           O
   :             :                   :               :
   :             :                   :               :
```

Method to solve the structure

Resolution is a measure of the quality of the data
>3Å is low resolution

https://www.rcsb.org/structure/1MBO
https://files.rcsb.org/view/1MBO.pdb

A typical PDB format file

```
HEADER    OXYGEN STORAGE                          27-AUG-81   1MBO
TITLE     STRUCTURE AND REFINEMENT OF OXYMYOGLOBIN AT 1.6 ANGSTROMS RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MYOGLOBIN;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
   :              :
   :              :
REMARK   2 RESOLUTION.    1.60 ANGSTROMS.
   :              :
   :              :
REMARK 800
REMARK 800 SITE
REMARK 800 SITE_IDENTIFIER: AC1
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE SO4 A 154
REMARK 800
REMARK 800 SITE_IDENTIFIER: AC2
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE HEM A 155
REMARK 800
REMARK 800 SITE_IDENTIFIER: AC3
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE OXY A 555
   :              :
   :              :

SITE     1 AC1  4 SER A  58  GLU A  59  ASP A  60  HOH A 227
SITE     1 AC2 22 THR A  39  LYS A  42  PHE A  43  ARG A  45
SITE     2 AC2 22 HIS A  64  VAL A  68  LEU A  89  SER A  92
SITE     3 AC2 22 HIS A  93  HIS A  97  ILE A  99  TYR A 103
SITE     4 AC2 22 LEU A 104  HOH A 163  HOH A 217  HOH A 229
SITE     5 AC2 22 HOH A 234  HOH A 269  HOH A 337  HOH A 469
SITE     6 AC2 22 HOH A 472  OXY A 555
SITE     1 AC3  5 PHE A  43  HIS A  64  VAL A  68  HEM A 155
SITE     2 AC3  5 HOH A 341
   :              :                   :                :
   :              :                   :                :
```

REMARK records present experimental details, annotations, comments, and information not included in other records
Eg:
REMARK 2 provides the resolution
REMARK 800 informs about sites (SITE record must be present)

https://www.rcsb.org/structure/1MBO
https://files.rcsb.org/view/1MBO.pdb

A typical PDB format file

```
HEADER    OXYGEN STORAGE                          27-AUG-81   1MBO
TITLE     STRUCTURE AND REFINEMENT OF OXYMYOGLOBIN AT 1.6 ANGSTROMS RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MYOGLOBIN;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
   :             :
   :             :
DBREF  1MBO A    1   153  UNP    P02185   MYG_PHYCA        1    153
SEQRES   1 A  153  VAL LEU SER GLU GLY GLU TRP GLN LEU VAL LEU HIS VAL
SEQRES   2 A  153  TRP ALA LYS VAL GLU ALA ASP VAL ALA GLY HIS GLY GLN
SEQRES   3 A  153  ASP ILE LEU ILE ARG LEU PHE LYS SER HIS PRO GLU THR
SEQRES   4 A  153  LEU GLU LYS PHE ASP ARG PHE LYS HIS LEU LYS THR GLU
SEQRES   5 A  153  ALA GLU MET LYS ALA SER GLU ASP LEU LYS LYS HIS GLY
SEQRES   6 A  153  VAL THR VAL LEU THR ALA LEU GLY ALA ILE LEU LYS LYS
SEQRES   7 A  153  LYS GLY HIS HIS GLU ALA GLU LEU LYS PRO LEU ALA GLN
SEQRES   8 A  153  SER HIS ALA THR LYS HIS LYS ILE PRO ILE LYS TYR LEU
SEQRES   9 A  153  GLU PHE ILE SER GLU ALA ILE ILE HIS VAL LEU HIS SER
SEQRES  10 A  153  ARG HIS PRO GLY ASP PHE GLY ALA ASP ALA GLN GLY ALA
SEQRES  11 A  153  MET ASN LYS ALA LEU GLU LEU PHE ARG LYS ASP ILE ALA
SEQRES  12 A  153  ALA LYS TYR LYS GLU LEU GLY TYR GLN GLY
   :             :
   :             :
HELIX    1   A SER A    3  GLU A   18  1                                  16
HELIX    2   B ASP A   20  SER A   35  1                                  16
HELIX    3   C HIS A   36  LYS A   42  1                                   7
HELIX    4   D THR A   51  ALA A   57  1                                   7
HELIX    5   E SER A   58  LYS A   77  1                                  20
HELIX    6   F LEU A   86  THR A   95  1                                  10
HELIX    7   G PRO A  100  ARG A  118  1                                  19
HELIX    8   H GLY A  124  LEU A  149  1                                  26
   :             :                   :                 :
   :             :                   :                 :
```

DBREF: cross-reference with sequence DB (could not exist take a look at https://www.rcsb.org/structure/4BWR)

SEQRES: Sequence used in the experiment, not necessarily all its residues have coordinates (display/download-> fasta sequence)

HELIX /SHEET –> secondary structure information

https://www.rcsb.org/structure/1MBO
https://files.rcsb.org/view/1MBO.pdb

A typical PDB format file

```
HEADER    OXYGEN STORAGE                          27-AUG-81   1MBO
TITLE     STRUCTURE AND REFINEMENT OF OXYMYOGLOBIN AT 1.6 ANGSTROMS RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MYOGLOBIN;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: PHYSETER CATODON;
SOURCE   3 ORGANISM_COMMON: SPERM WHALE;
SOURCE   4 ORGANISM_TAXID: 9755
KEYWDS    OXYGEN STORAGE
EXPDTA    X-RAY DIFFRACTION
AUTHOR    S.E.V.PHILLIPS
REMARK   2 RESOLUTION.    1.60 ANGSTROMS.
   :             :
   :             :
HELIX    6   F LEU A   86  THR A   95  1                                   10
HELIX    7   G PRO A  100  ARG A  118  1                                   19
HELIX    8   H GLY A  124  LEU A  149  1                                   26
LINK         NE2 HIS A  93                FE   HEM A 155     1555   1555  2.07
```

atomic coordinates

Atom # and name, residue name, one letter to specify the chain (in oligomeric proteins), residue number, its x, y, and z coordinates, occupancy and temperature factor

```
    :              :                    :                :
ATOM      1  N   VAL A   1      -0.686  14.852  16.090  1.00 42.31           N
ATOM      2  CA  VAL A   1       0.621  15.543  16.081  1.00 53.86           C
                                                        00 27.97           C
                                                        00 23.24           O
                                                        00 36.55           C
ATOM      6  CG1 VAL A   1      -0.444  17.695  16.811  1.00 35.23           C
ATOM      7  CG2 VAL A   1       0.119  15.972  18.559  1.00 53.38           C
ATOM      8  N   LEU A   2       3.112  14.944  16.272  1.00 13.35           N
ATOM      9  CA  LEU A   2       4.140  13.989  16.663  1.00 16.70           C
ATOM     10  C   LEU A   2       3.978  13.813  18.182  1.00 18.20           C
ATOM     11  O   LEU A   2       3.770  14.833  18.864  1.00 17.87           O
                                                        :
                                                        :
```

ATOM atoms in proteins or nucleic acid atoms

HETATM atoms in non standard residues.

https://www.rcsb.org/structure/1MBO
https://files.rcsb.org/view/1MBO.pdb

PDB: coordinates format

```
#--------------------------------------------------------------------------------
#Field |    Column    | FORTRAN |
#  No. |     range    | format  | Description
#--------------------------------------------------------------------------------
#   1. |    1 -  6    |   A6    | Record ID (eg ATOM, HETATM)
#   2. |    7 - 11    |   I5    | Atom serial number
#   -  |   12 - 12    |   1X    | Blank
#   3. |   13 - 16    |   A4    | Atom name (eg " CA " , " ND1")
#   4. |   17 - 17    |   A1    | Alternative location code (if any)
#   5. |   18 - 20    |   A3    | Standard 3-letter amino acid code for residue
#   -  |   21 - 21    |   1X    | Blank
#   6. |   22 - 22    |   A1    | Chain identifier code
#   7. |   23 - 26    |   I4    | Residue sequence number
#   8. |   27 - 27    |   A1    | Insertion code (if any)
#   -  |   28 - 30    |   3X    | Blank
#   9. |   31 - 38    |  F8.3   | Atom's x-coordinate
#  10. |   39 - 46    |  F8.3   | Atom's y-coordinate
#  11. |   47 - 54    |  F8.3   | Atom's z-coordinate
#  12. |   55 - 60    |  F6.2   | Occupancy value for atom
#  13. |   61 - 66    |  F6.2   | B-value (thermal factor)
#   -  |   67 - 67    |   1X    | Blank
#  14. |   68 - 70    |   I3    | Footnote number
#--------------------------------------------------------------------------------
#
#         1         2         3         4         5         6
#12345678901234567890123456789012345678901234567890123456789012345678
#--------------------------------------------------------------------
#@<<<<<@>>>> @<<<@@<< @@>>>@   @>>>>>>>@>>>>>>>@>>>>>>>
#ATOM   1751  N   GLY C 250      32.286   1.882  43.206  1.00 22.00
#ATOM     31  P   POM     3       8.010   4.280   1.939
#ATOM   3000 SE   MSE A 401     -1.600 -14.228  65.136  0.70 15.94          SE
```

Not in the pdb file. Just here to easy visualize and understand the format.

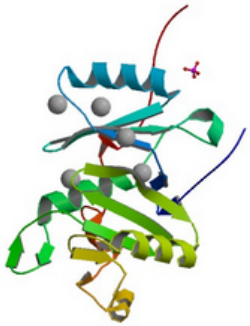
PDB: Web interface

https://www.rcsb.org/structure/4BYZ

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | MyPDB

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

Biological Assembly 1



3D View: Structure | Electron Density | Ligand Interacti...

Standalone Viewers
Protein Workshop | Ligand Explorer

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A

Biological assembly 1 assigned by authors and generated by PISA (software)

Macromolecule Content

- Total Structure Weight: 22489.71
- Atom Count: 1630
- Residue Count: 194
- Unique protein chains: 1

Display Files | Download Files

4BYZ

Structural characterization using Sulfur-SAD of the cytoplasmic domain of Burkholderia pseudomallei PilO2Bp, an actin-like protein component of a Type IVb R64-derivative pilus machinery.

DOI: 10.2210/pdb4BYZ/pdb

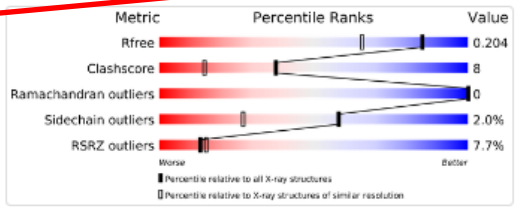
Classification: MOTOR PROTEIN
Organism(s): Burkholderia pseudomallei (strain K96243)
Expression System: Escherichia coli BL21(DE3)

Deposited: 2013-07-22 **Released:** 2014-04-23
Deposition Author(s): Lassaux, P., Manjasetty, B.A., Conchillo-Sole, O., Yero, D., Gourlay, L., Perletti, L., Daura, X., Belrhali, H., Bolognesi, M.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 1.55 Å
R-Value Free: 0.198
R-Value Work: 0.150

wwPDB Validation | 3D Report | Full Report



This is version 1.0 of the entry. See complete history

Literature | Download Primary Citation

Redefining the Pf06864 Pfam Family Based on Burkholderia Pseudomallei Pilo2BP S-Sad Crystal Structure.

Lassaux, P., Conchillo-Sole, O., Manjasetty, B.A., Yero, D., Perletti, L., Belrhali, H., Daura, X., Gourlay, L.J., Bolognesi, M.
(2014) Plos One 9: 94981

PubMed: 24728008 Search on PubMed | Search on PubMed Central
DOI: 10.1371/journal.pone.0094981
Primary Citation of Related Structures:
4BZ0

PubMed Abstract:
Type IV pili are surface-exposed filaments and bacterial virulence factors, represented by the Tfpa and Tfpb types, which assemble via specific machineries. The Tfpb group is further divided into seven variants, linked to heterogeneity in the assembl ...

Expression system

Method used to solve the structure and, if by X-Ray diffraction, its **resolution** (any value below 3A is considered good resolution)

View the structure interactively (rotate, zoom, cartoons, surface, etc) and its electron density

Publications:
- View this pubmed entry
- Search PDB by this pubmed reference

Structure: PDB

Web services

PDB Web services:
https://www.rcsb.org/pages/webservices
https://data.rcsb.org/redoc/index.html

Legacy web services (Not working anymore. Question: what happens with the software that uses them?)
https://www.rcsb.org/pages/webservices/rest-fetch
http://www.rcsb.org/pdb/software/rest.do

The following URLs return data in json format. Firefox formats this results in a human readable way. Safari present them as they are.

Example: **get ligand info for 4HHB**:

- With new API:

#Get data for 4HHB entry
GET https://data.rcsb.org/rest/v1/core/entry/4hhb
#Extract ".rcsb_entry_container_identifiers.non_polymer_entity_ids" (result: 3,4)
Get data for non polymer id 3
GET "https://data.rcsb.org/rest/v1/core/nonpolymer_entity/4hhb/3"
Extract ". pdbx_entity_nonpoly. comp_id" (result: "HEM")
GET https://data.rcsb.org/rest/v1/core/chemcomp/HEM
Repeat the process for non polymer id 4
GET "https://data.rcsb.org/rest/v1/core/nonpolymer_entity/4hhb/4"
…

3D structure of proteins