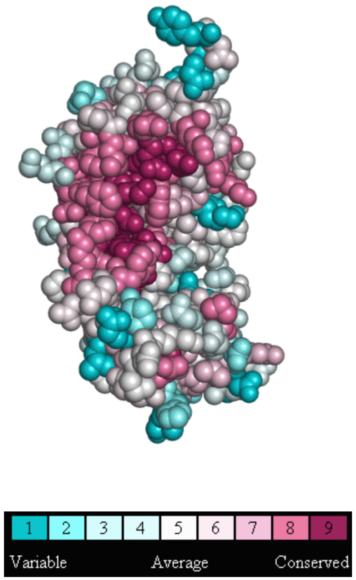
Bioinformatics for molecular biology

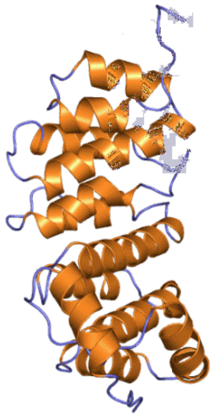
Structural bioinformatics tools, predictors, and 3D modeling – Structural Biology Review

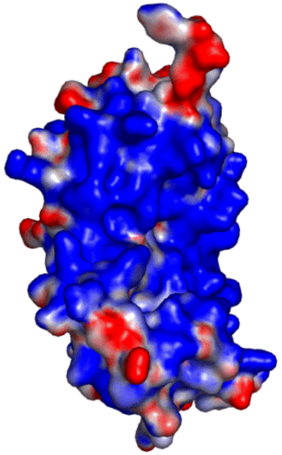
Dr Jon K. Lærdahl, Research Scientist

Department of Microbiology,
Oslo University Hospital - Rikshospitalet &
Bioinformatics Core Facility/CLS initiative,
University of Oslo

E-mail: jonkl@medisin.uio Phone: +47 22844784
Group: Torbjørn Rognes
(http://www.ous-research.no/rognes)
CF: Bioinformatics services
(http://core.rr-research.no/bioinformatics)
CLS: Bioinformatics education
(http://www.mn.uio.no/ifi/english/research/networks/clsi)
Main research area:
Structural and Applied Bioinformatics







Oslo universitetssykehus

Overview

Now:

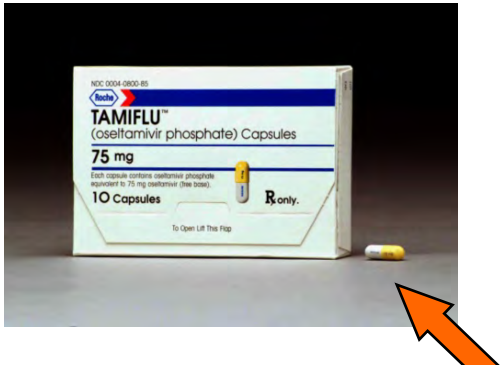
- Protein Structure Review
  - Amino acids, polypeptides, secondary structure elements, visualization, structure determination by X-ray crystallography and NMR methods, PDB

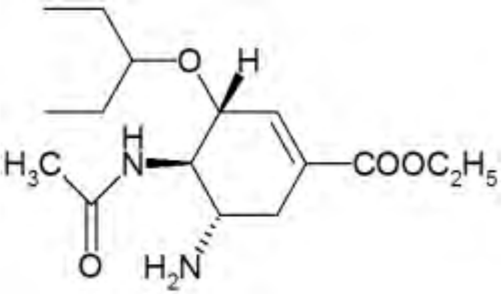
Later…

- Structure comparison and classification (CASP & SCOP)
- Predictors
- 3D structure modeling
  - *Ab initio*
  - Threading/fold recognition
  - Homology modeling
- Practical exercises
  - PyMOL & visualization
- Practical Exercises
  - Homology modeling of influenza neuraminidase (Tamiflu resistance?)
  - Other homology modeling
  - Threading
  - Your own project?

Stop me and ask questions!!

Structural bioinformatics

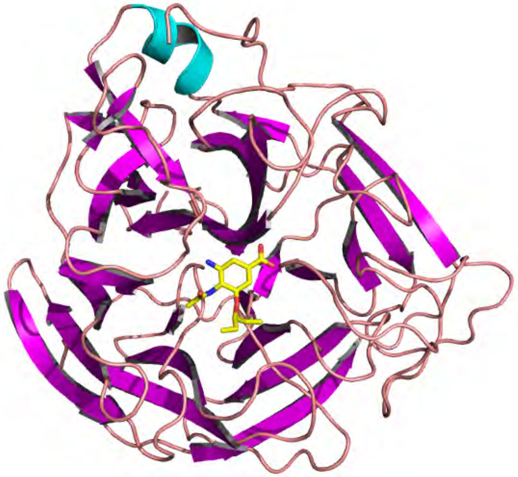

NDC 0004-0800-85
Roche
TAMIFLU™
(oseltamivir phosphate) Capsules
75 mg
Each capsule contains oseltamivir phosphate equivalent to 75 mg oseltamivir (free base).
10 Capsules
℞ only.
To Open Lift This Flap

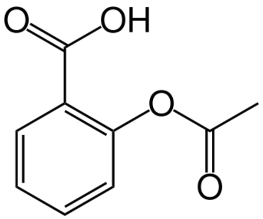


To understand what is really going on in biology you need the 3D structure of the macromolecules, *i.e.* the proteins in particular!

```
ACACACTGGCCTTGGACTCAACCTGATGGGCTTCTGGGCCCAGCCCCAGACAAACCCCCGGCAAACGTC
CCATTCCGAGGAAGCATGAGCAGATGGAGTATGGAAGAAATGCCCAAGACGGCAGGCAGCAGCTGTGGC
GGCCGGCGGACGACAATCCGAGGAGAGGCCTCTGATGTCCTGAGGTCTCAGAGGACGCCTAAAGGCCTT
GAATGGACAAGCTTAGCGGGCGGGCGCAGAAGAGAATAATACTCTGGAGACACTTCCCGAGGGCTCTGG
GGCCGGAGCTGTGTTCGCTCCGGTTCTTGGTGAAGACAGGGTTCGTGGGAGGCGGCCCAAGGAGGGCGAA
CGCCTAAGACTGCAAAGGCTCGGGGGAGAACGGCTCTCGGAGAACGGGCTGGGGAAGGACGTGGCTCTGA
AGACGGACAGCCCTGAGGAACCGCGGGGCGCCCAGATGGAACTCGTTAGCGCCCCGAGTGCAGACAATCC
CGGAGGGGGAAAGGCGAGCAGCTGGCAGAGAGCCCAGTGCCGGAAACCGCGCGAGCGCCTCAGAACGGC
```

Neuraminidase is a glycoside hydrolase enzyme found on the surface of the influenza virus

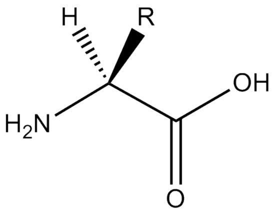


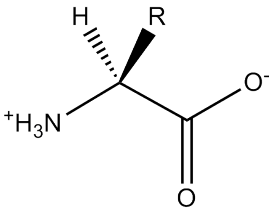


Amino acids – the building blocks of proteins

Proteins are built from 20 naturally occurring amino acids. They have an amino (-NH_2) and acidic (-COOH) functional group

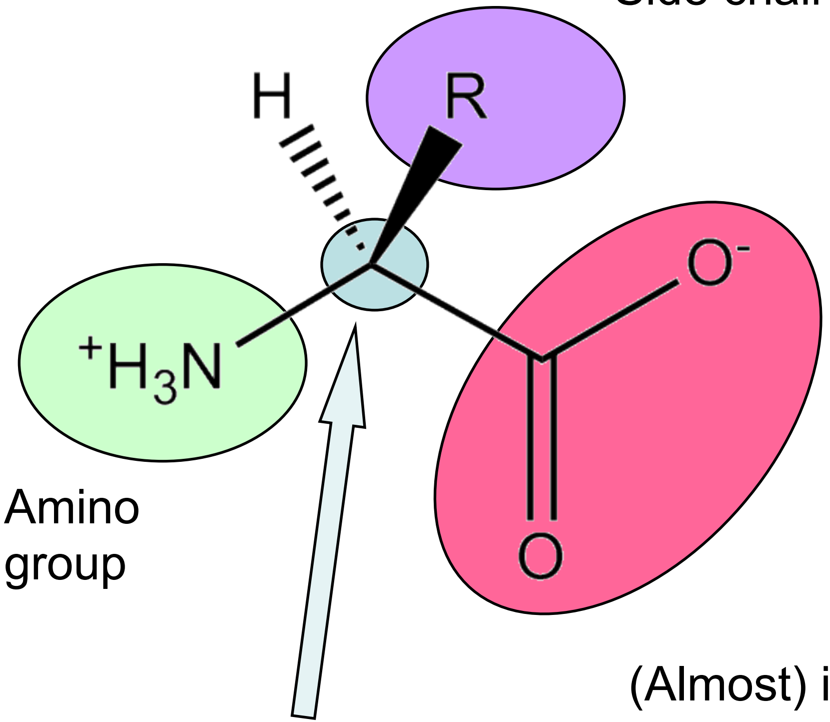
The side chain group (R) determines the properties of the amino acid





Zwitterionic form found at physiological pH

Side chain



Carboxyl group

Amino group

(Almost) invariably enantiomeric L-form (or *S*-form)

alpha carbon (C_α)

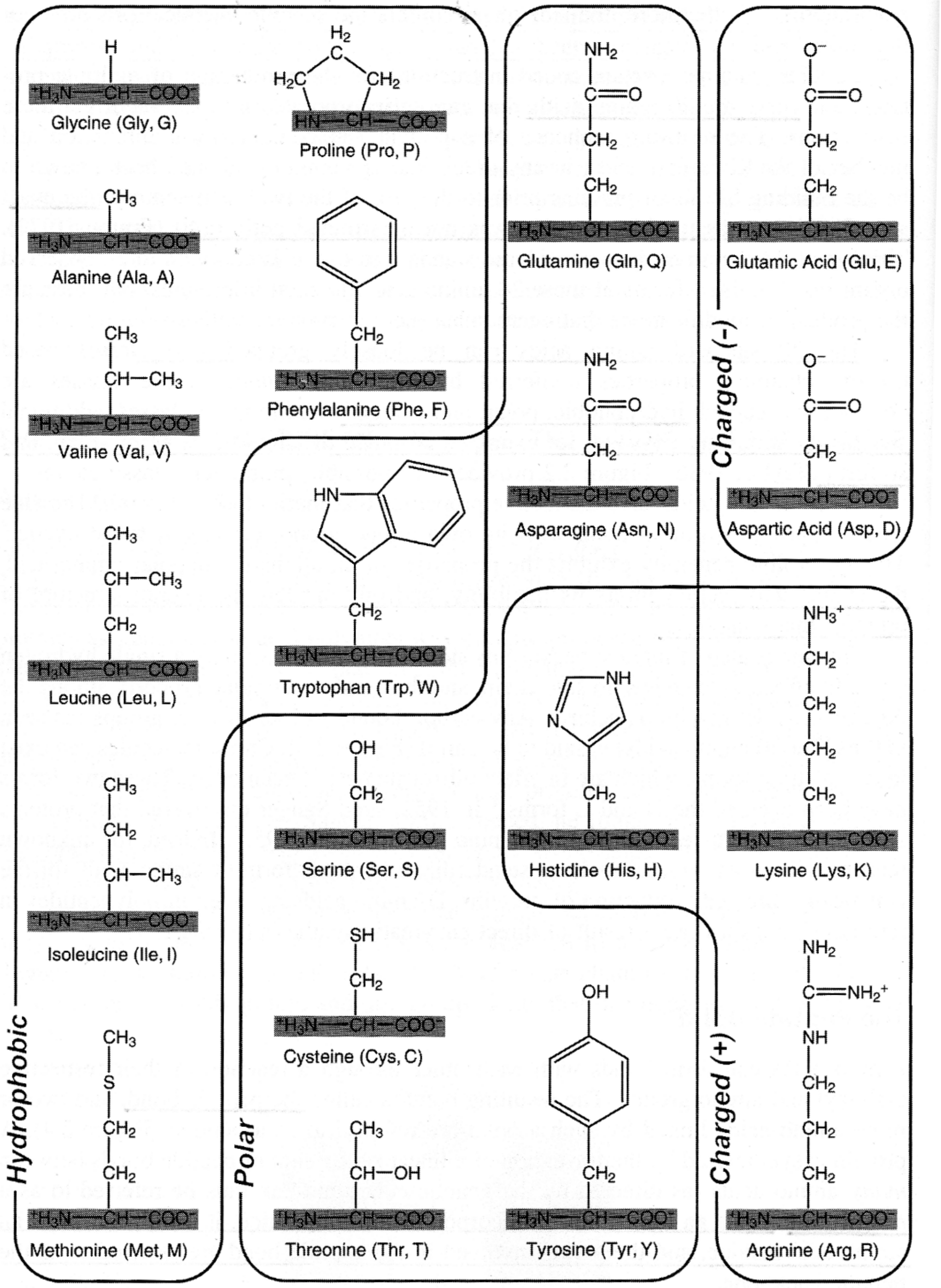
Amino acids

R-group properties:

- Large
- Small

- Hydrophobic
  - Aliphatic
  - Aromatic
- Polar
- Charged
  - Positive/negative charge

Increasing hydrophilicity/higher water (solvent) affinity

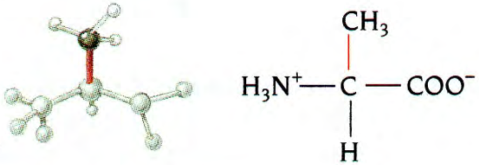


Structural Bioinformatics,
Eds. P.E. Bourne & H. Weissig
(Wiley, Hoboken, NJ, 2003)

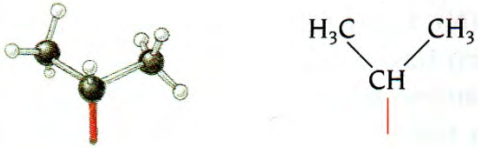
Amino acids

Introduction to Protein Structure, C. Branden & J. Tooze
(Garland, New York, 1998)

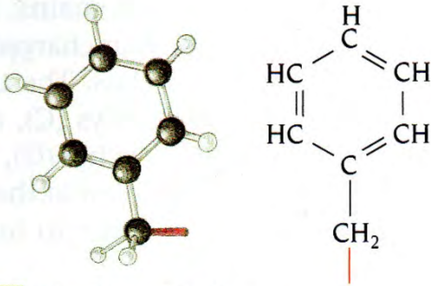
(a) Hydrophobic amino acids



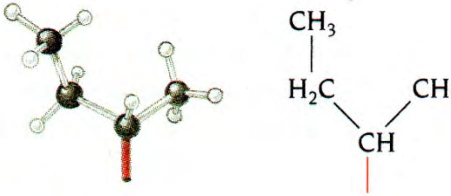
A Ala, Alanine



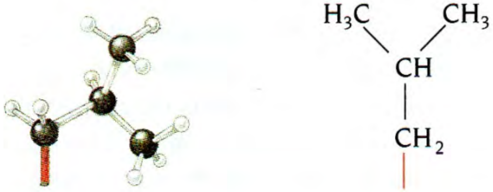
V Val, Valine



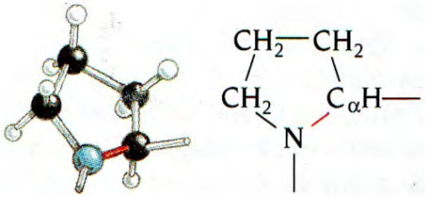
F Phe, Phenylalanine



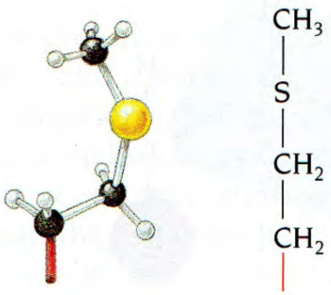
I Ile, Isoleucine



L Leu, Leucine



P Pro, Proline



M Met, Methionine

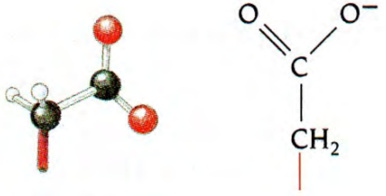
- Hydrophobic
  - Aliphatic
  - Aromatic

- 3-letter code
- 1-letter code

Amino acids

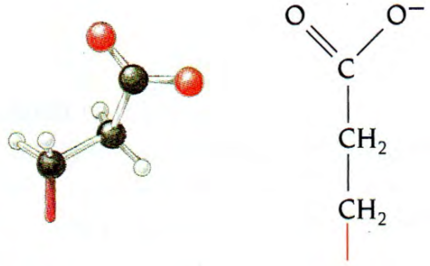
Introduction to Protein Structure, C. Branden & J. Tooze (Garland, New York, 1998)

(b) Charged amino acids



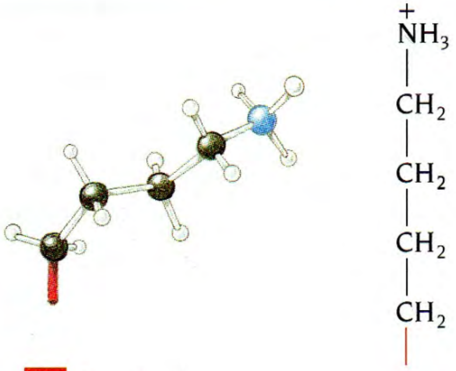
D *Asp, Aspartic acid*

Aspartate



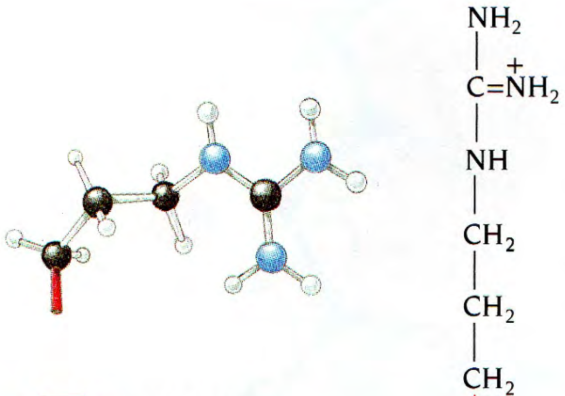
E *Glu, Glutamic acid*

Glutamate



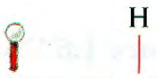
K *Lys, Lysine*

- Hydrophilic
  - Positive charge/basic
  - Negative charge/acidic



R *Arg, Arginine*

(d) Glycine



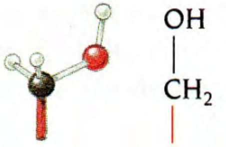
G *Gly, Glycine*

Amino acids

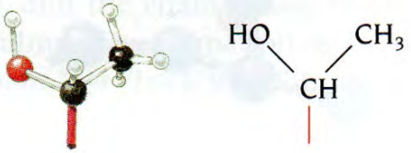
Introduction to Protein Structure, C. Branden & J. Tooze (Garland, New York, 1998)

(c) Polar amino acids

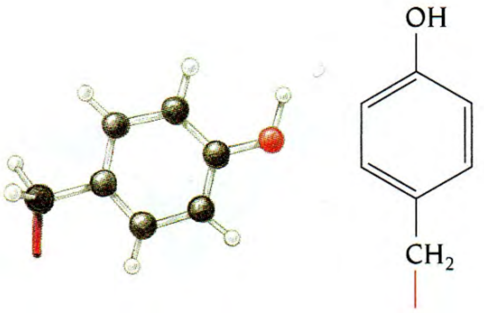
- Hydrophilic



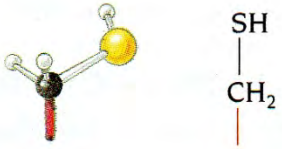
S Ser, Serine



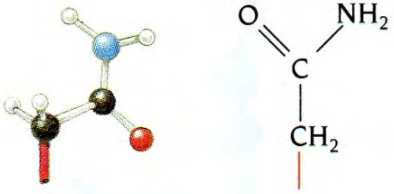
T Thr, Threonine



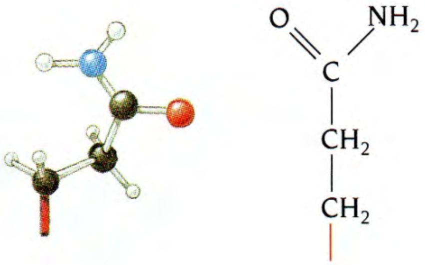
Y Tyr, Tyrosine



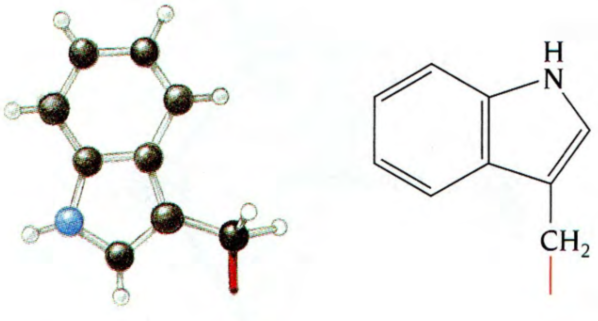
C Cys, Cysteine



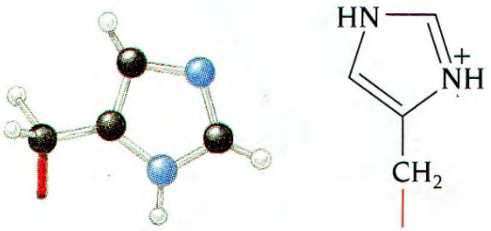
N Asn, Asparagine



Q Gln, Glutamine



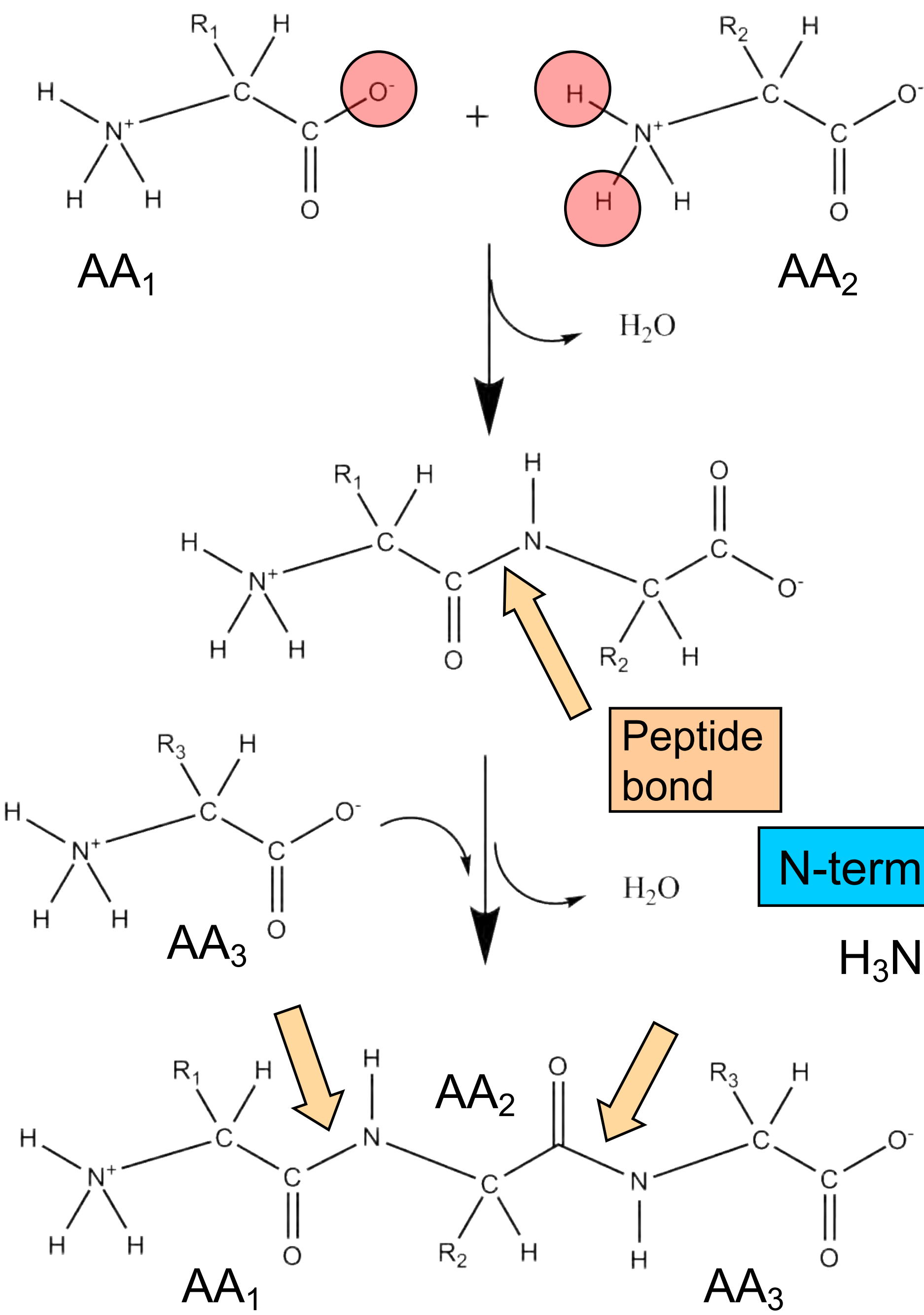
W Trp, Tryptophan



H His, Histidine

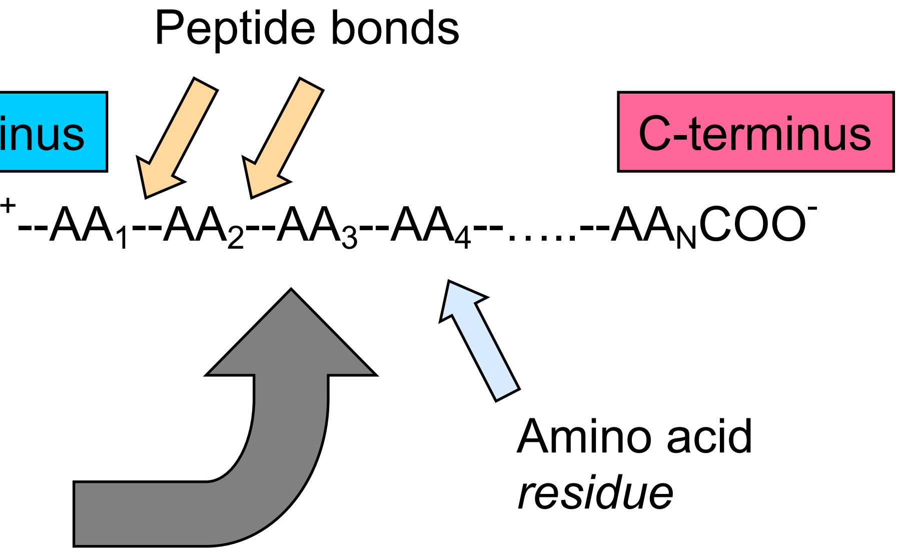
Polypeptides

Jon K. Lærdahl,
Structural Bioinformatics



Proteins are polypeptides, *i.e.* many amino acids connected by peptide bonds

N-terminus



Peptide bonds

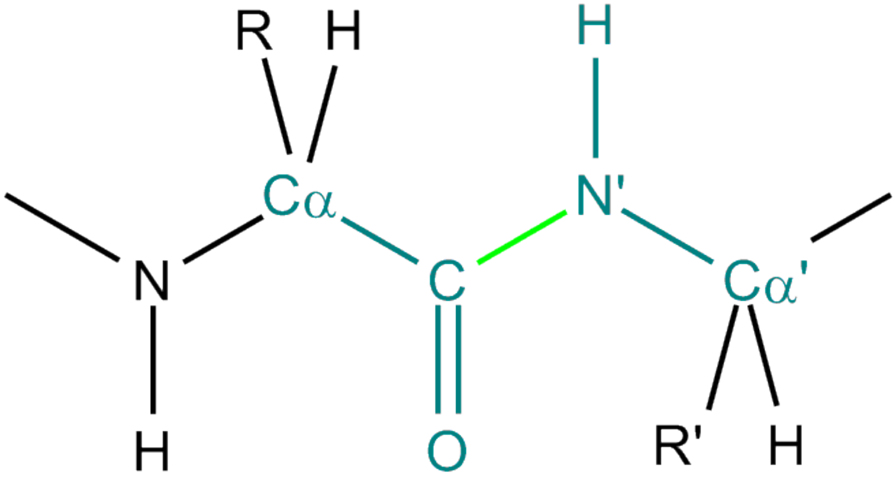
C-terminus

H_3N^{+}--AA_1--AA_2--AA_3--AA_4--…..--AA_NCOO^-

Amino acid *residue*

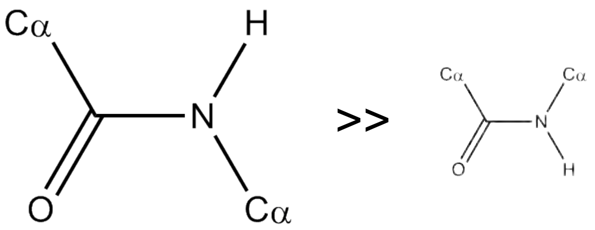
Dihedral angles

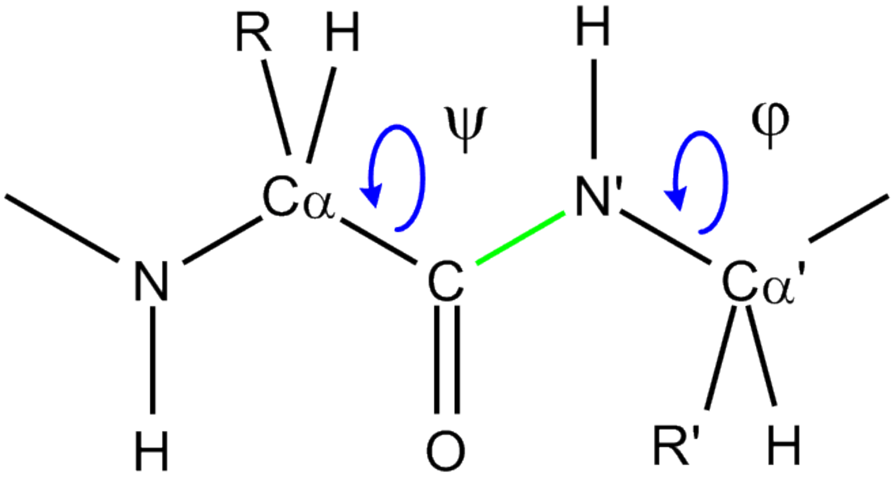
Proteins are polypeptides, *i.e.* many amino acids connected by peptide bonds



The peptide bond (light green) is a partial double bond and is fixed at ~180º, *i.e.* the green part is flat

Cis-form for peptide bond is extremely rare except for prolines (~25%).

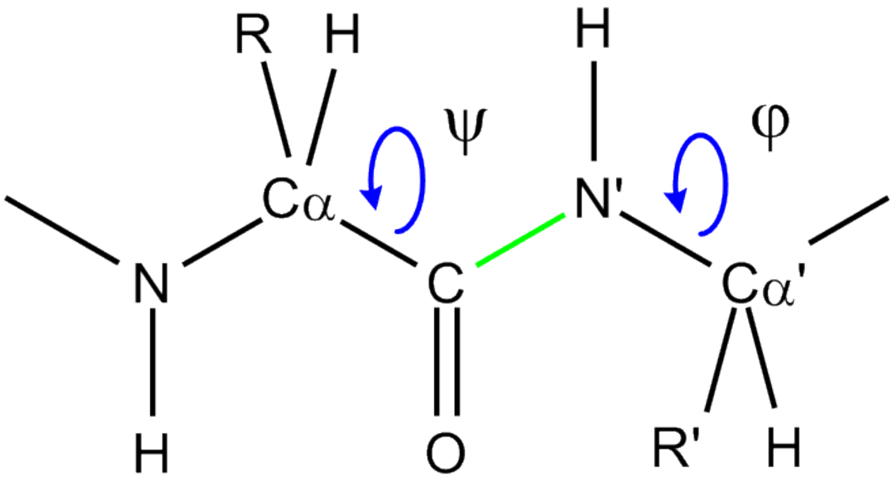




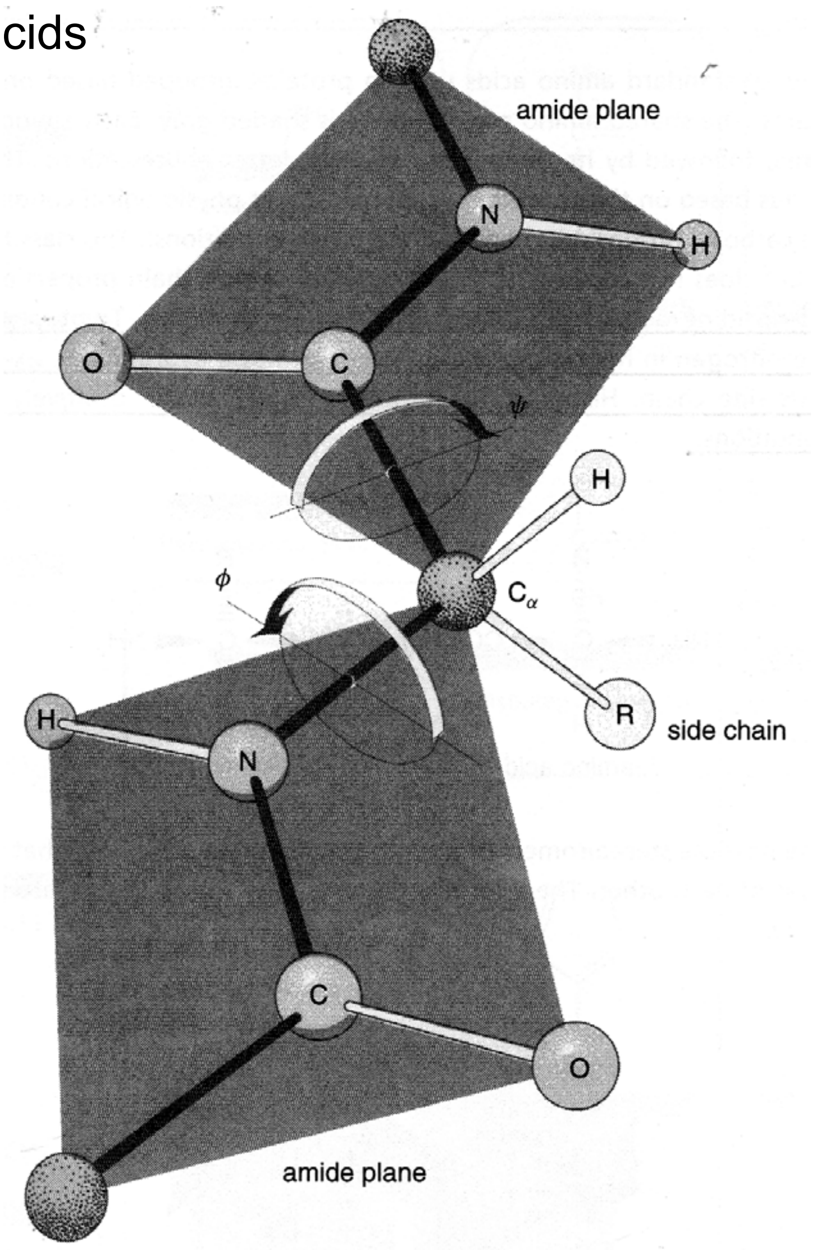
The dihedral angles phi (φ) and psi (ψ) determines the conformation of the peptide backbone

Dihedral angles

Proteins are polypeptides, *i.e.* many amino acids connected by peptide bonds



One (φ,ψ) pair for each residue in a protein

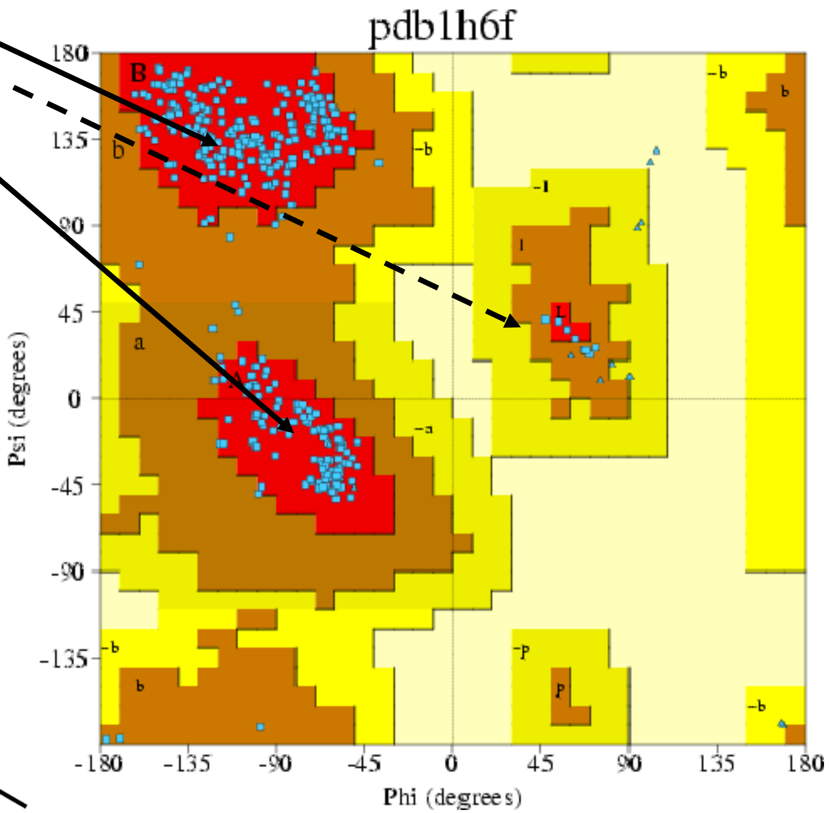


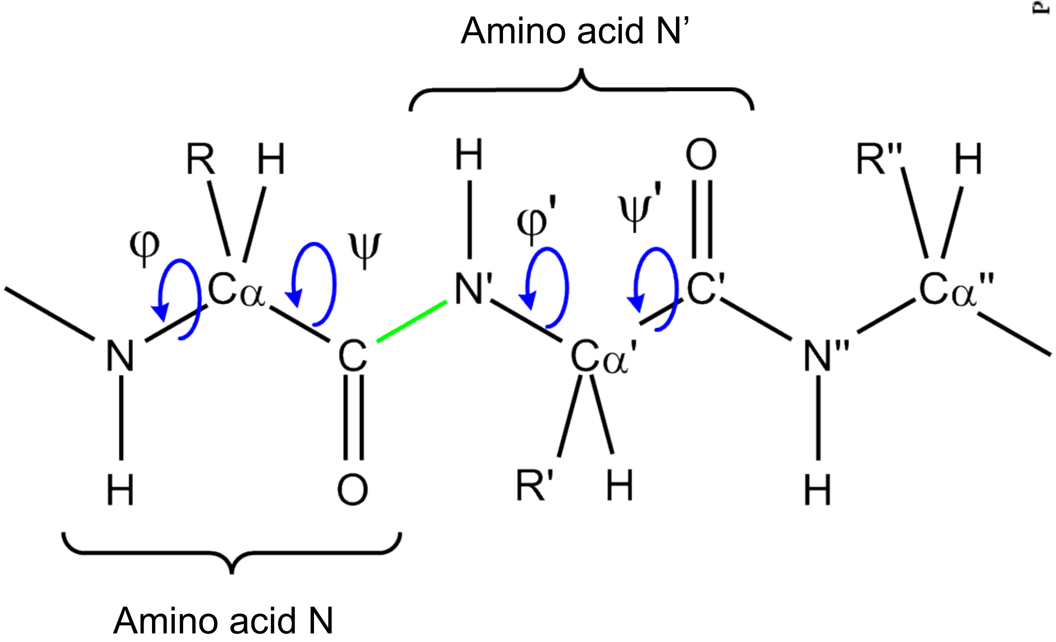
Structural Bioinformatics,
Eds. P.E. Bourne & H. Weissig
(Wiley, Hoboken, NJ, 2003)

Ramachandran plot

- Dihedral angles
  - Phi (φ)
  - Psi (ψ)
- Plot of (φ,ψ) angle pairs for each residue in a protein: *Ramachandran plot*

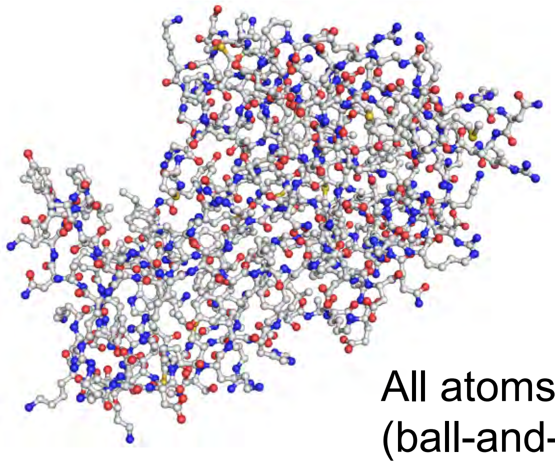
Most (φ,ψ) pairs in two (three) regions



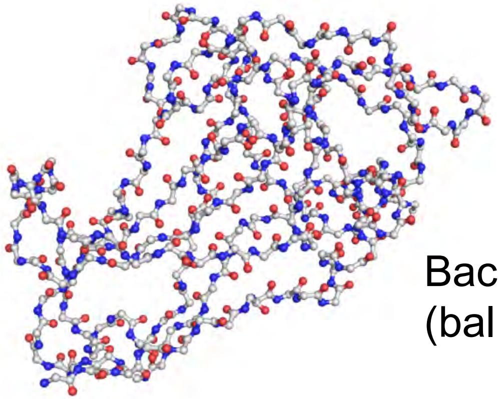


One point (blue spot) for each of the 184 residues in this protein (1H6F) (a human a transcription factor)

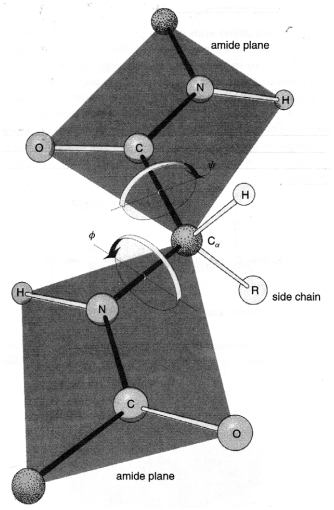
Ramachandran plot



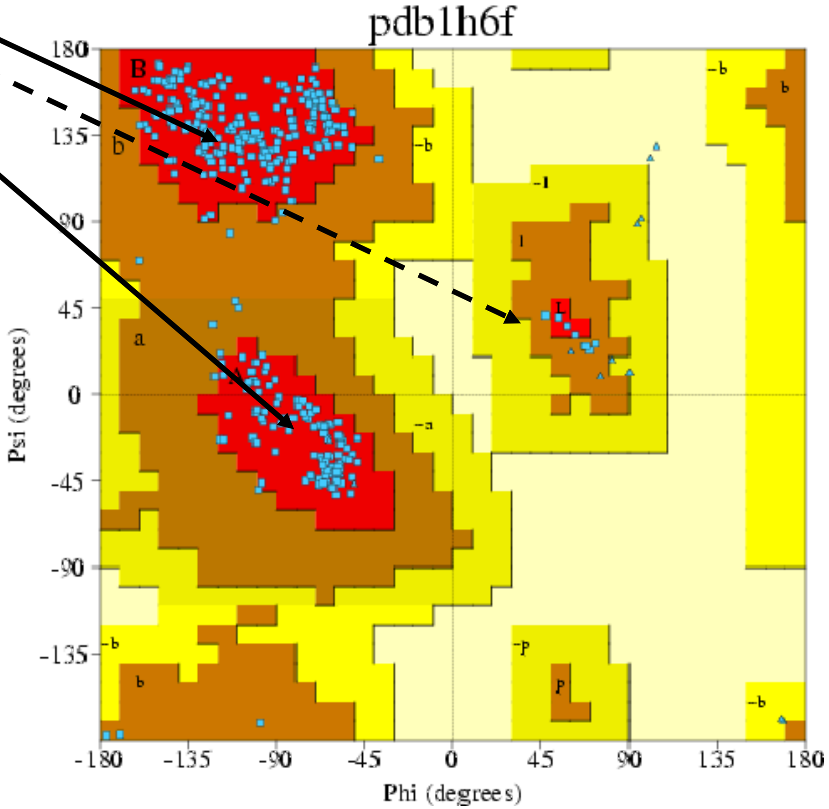
All atoms
(ball-and-stick)



Backbone atoms only
(ball-and-stick)



Most (φ,ψ) pairs in
two (three) regions

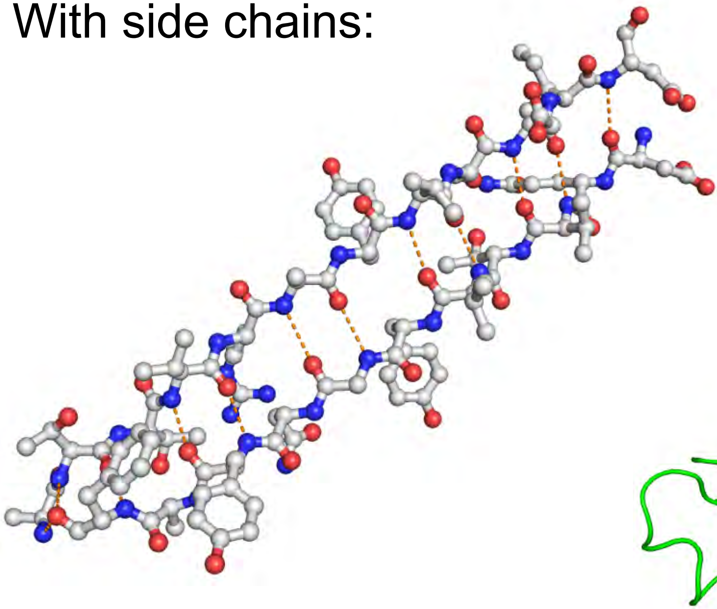


One point (blue spot) for each of the 184 residues in this protein (1H6F) (a human a transcription factor)

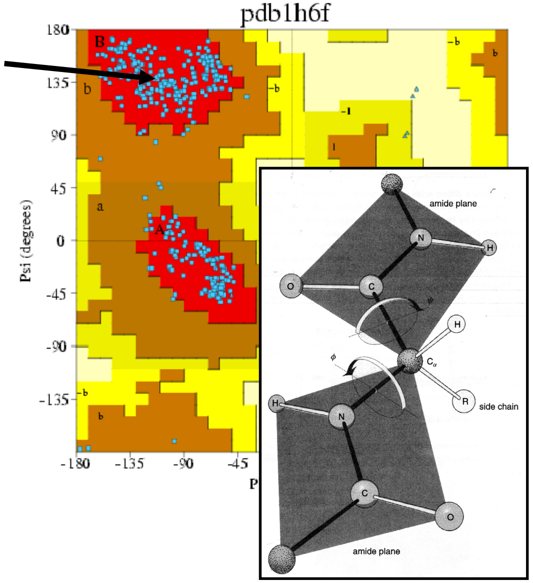
Jon K. Lærdahl,
Structural Bioinformatics

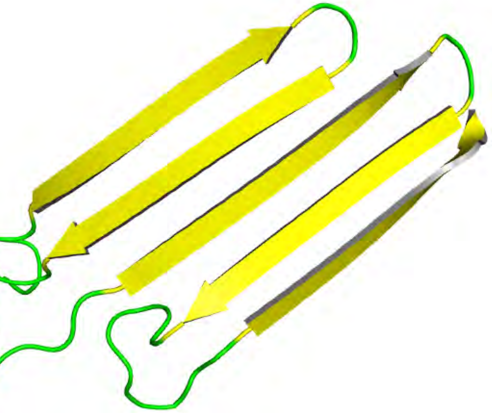
Secondary structure – β-sheets

With side chains:



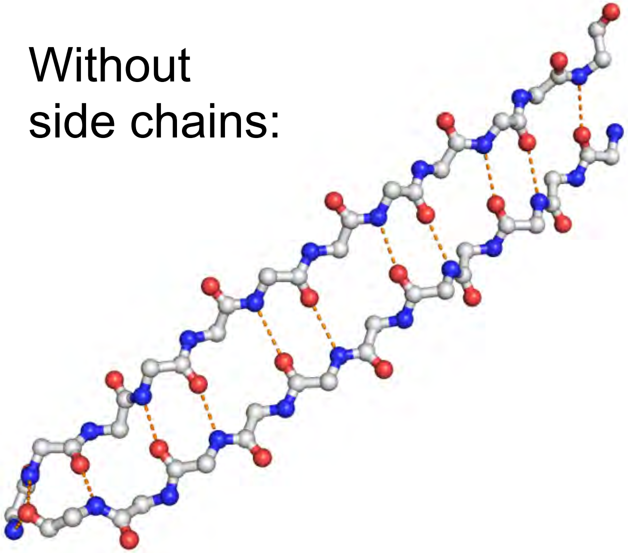
β-strands & β-sheets

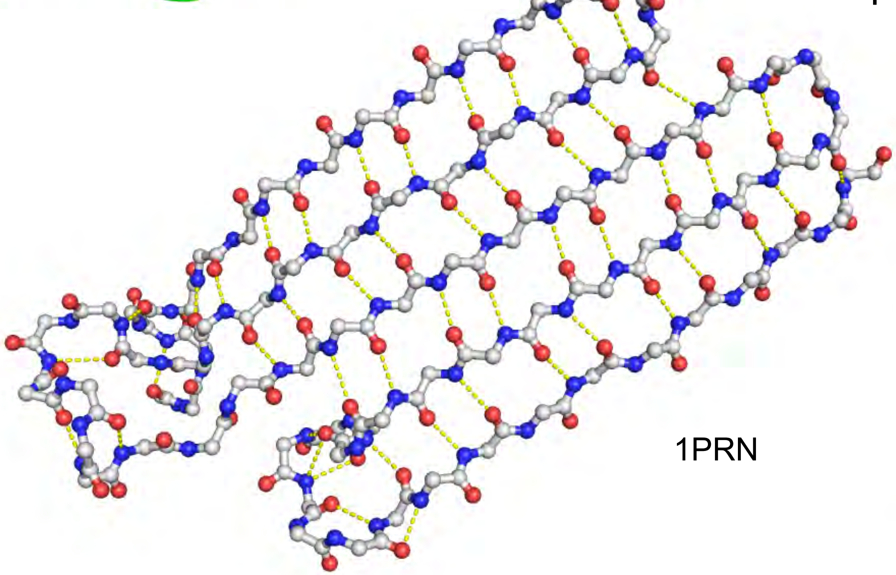




Psi ~ 135º
Phi ~ -100º

Without
side chains:

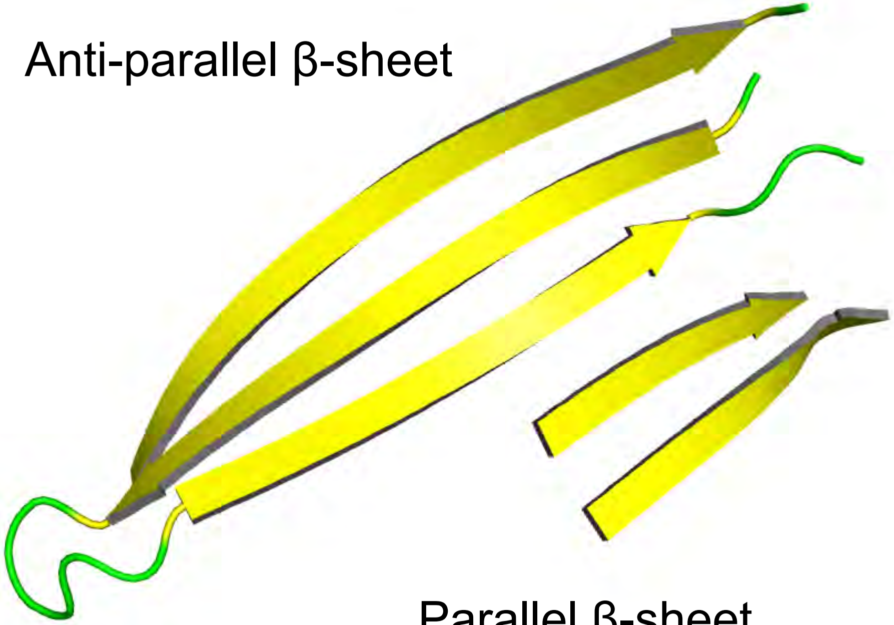




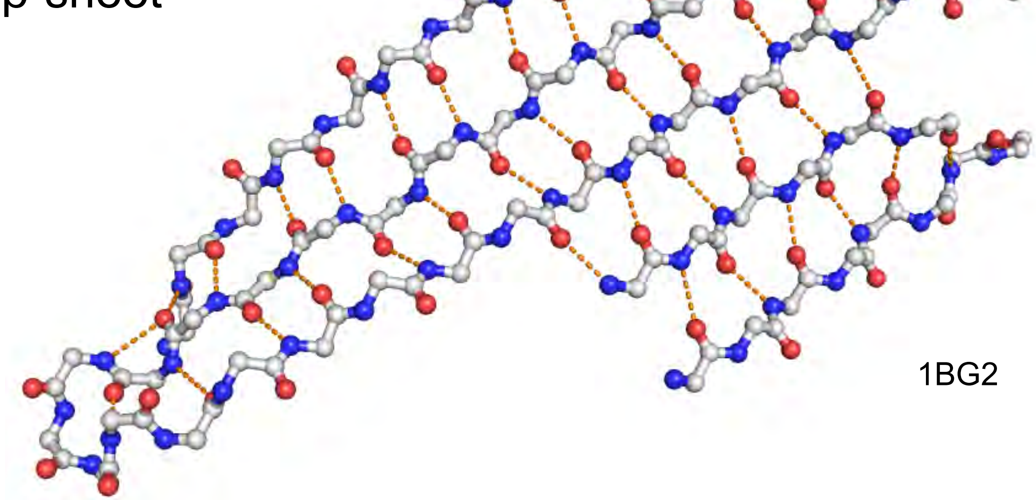
1PRN

Secondary structure – β-sheets

Anti-parallel β-sheet



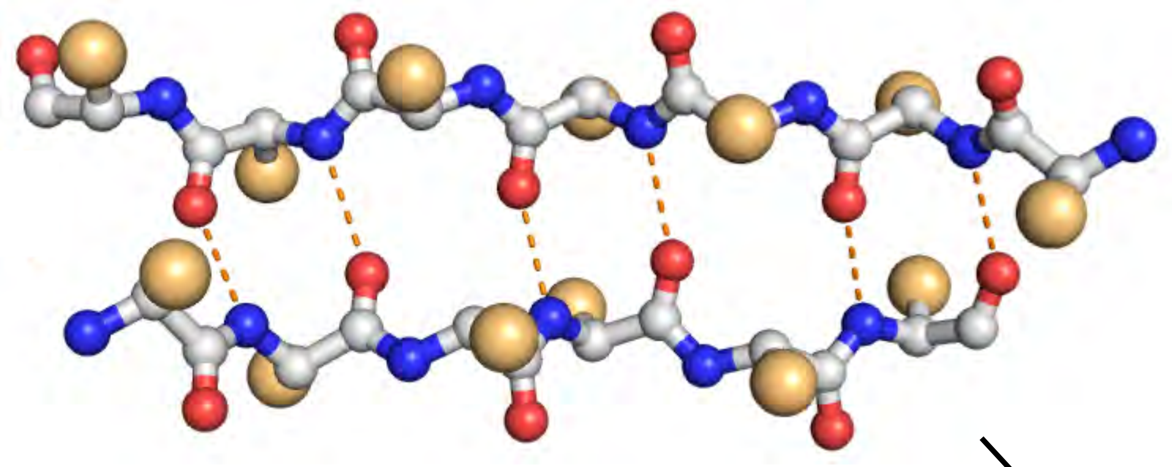
Parallel β-sheet



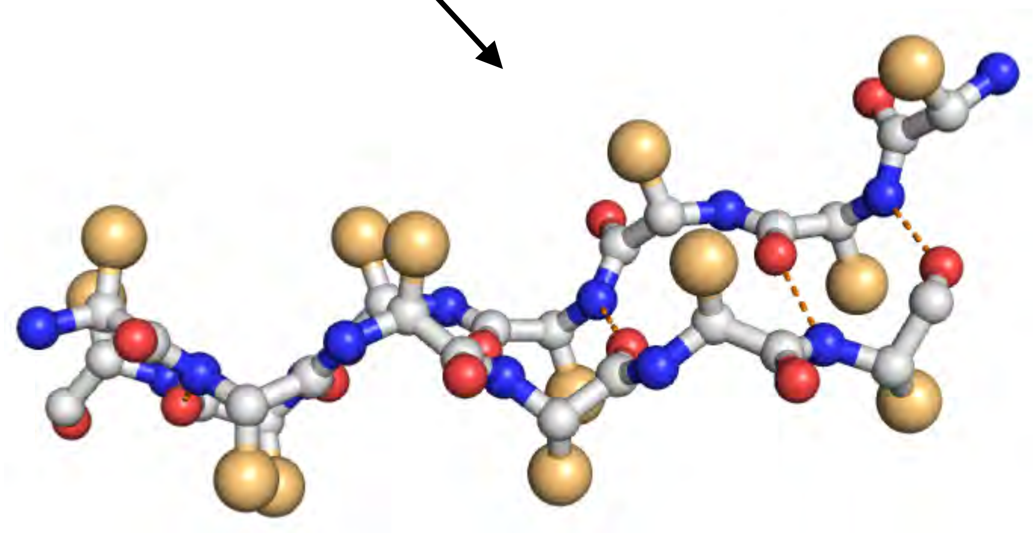
1BG2

β-sheets can be parallel, anti-parallel or mixed

Secondary structure – β-sheets



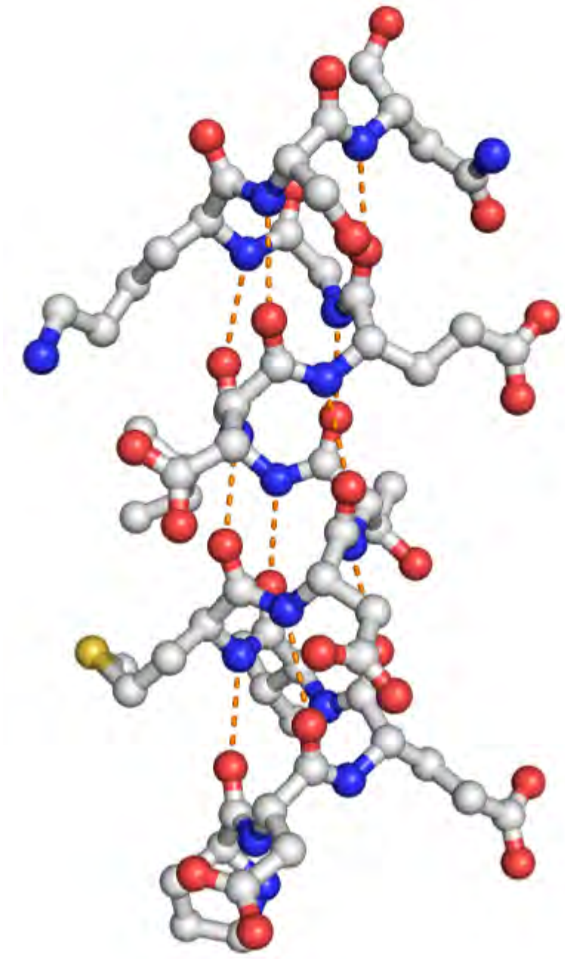
90º rotation



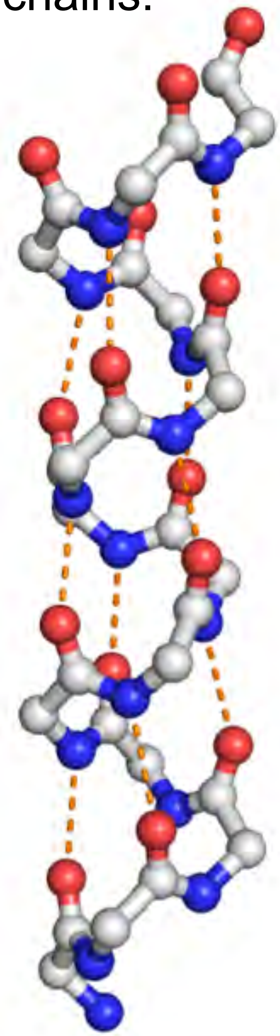
In β-sheets each side chain R-group is alternately on opposite sides of the plane of the sheet

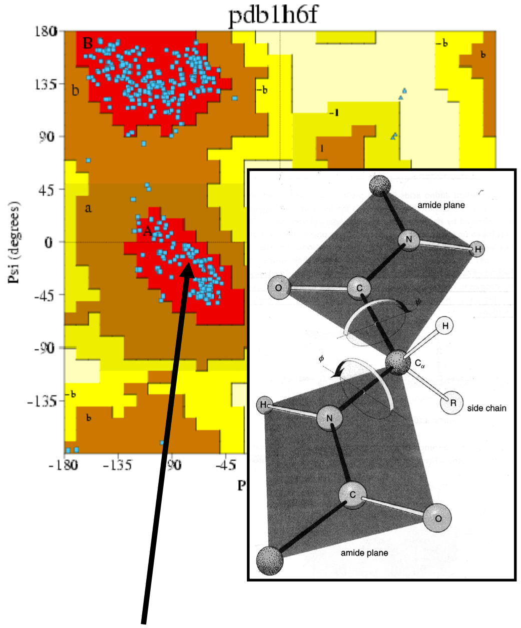
Secondary structure – α-helices

With side chains:



Without side chains:





α-helix

3.6 amino acids/turn

H-bonds between amino acids n & n+4

Partial positive charge at N-terminus and negative charge at C-terminus, *i.e.* it is a *dipole*

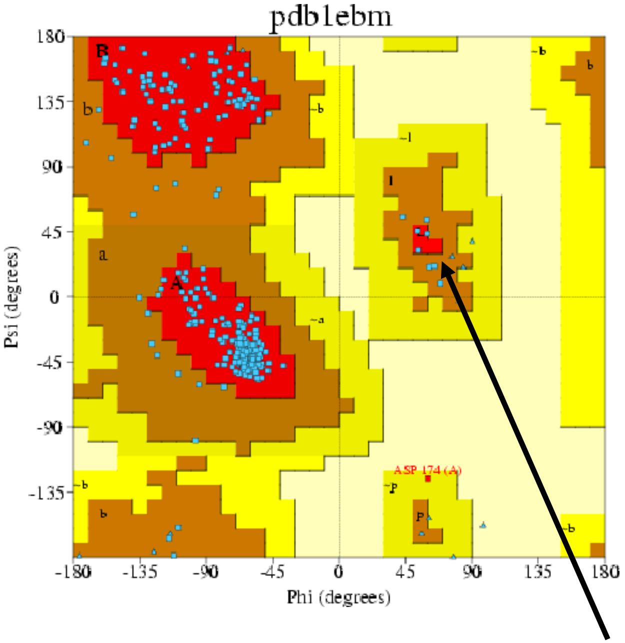
Jon K. Lærdahl,
Structural Bioinformatics

Secondary structure – 3 states

Three ”states”:
α-helices (H)
β-sheets (E)
Loops/coils (C)

Loops/coils:

- Loops may be hairpins or sharp turns
- Random coils/irregular loops
- Often “allowed” with insertions/deletions, *i.e.* evolutionary variable regions



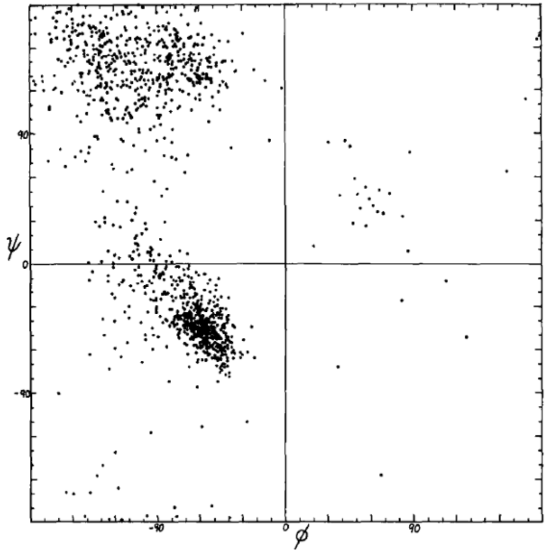
Left-handed helices

Coil here: “Everything that is not helix or sheet”

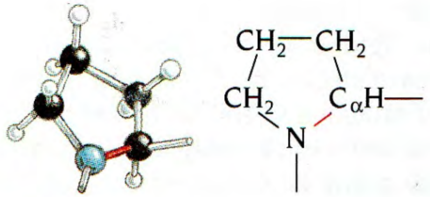
Coil often means: “Everything that is not helix or sheet or some characteristic loops”

Often contains Gly (to give flexibility) or Pro (to “break up” secondary structure elements)

Secondary structure – Gly & Pro

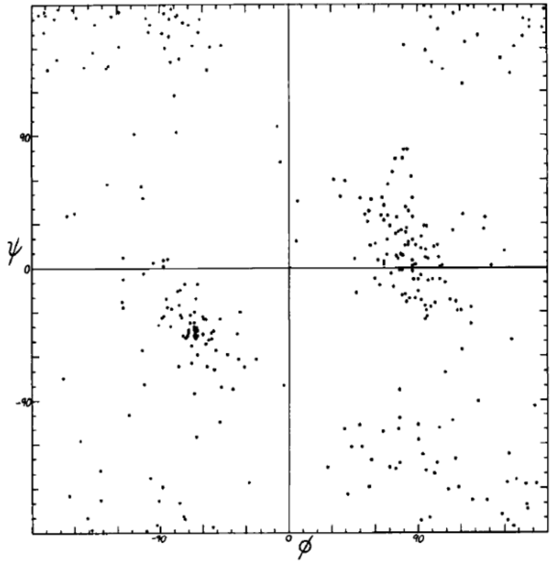


Non-glycine residues are mainly in α-helices and β-sheets



P Pro, Proline

Proline has very little flexibility in the backbone (disruptive to normal secondary structure)



Glycine has no side chain and a more flexible backbone

(d) Glycine

G Gly, Glycine

J. Richardson, *Adv. Prot. Chem.* **34**, 167 (1981)

Protein structure

- Primary structure: Linear amino acid sequence

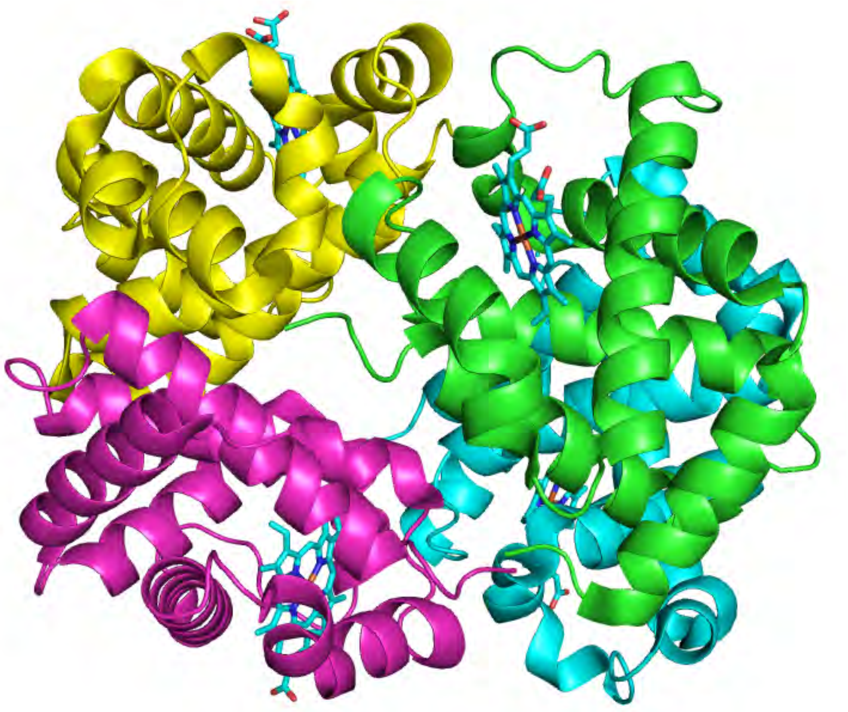
Met-Ala-Leu-Asp-Asp-…

- Secondary structure: Local conformation of the peptide chain:
    - α-helix
    - β-sheet

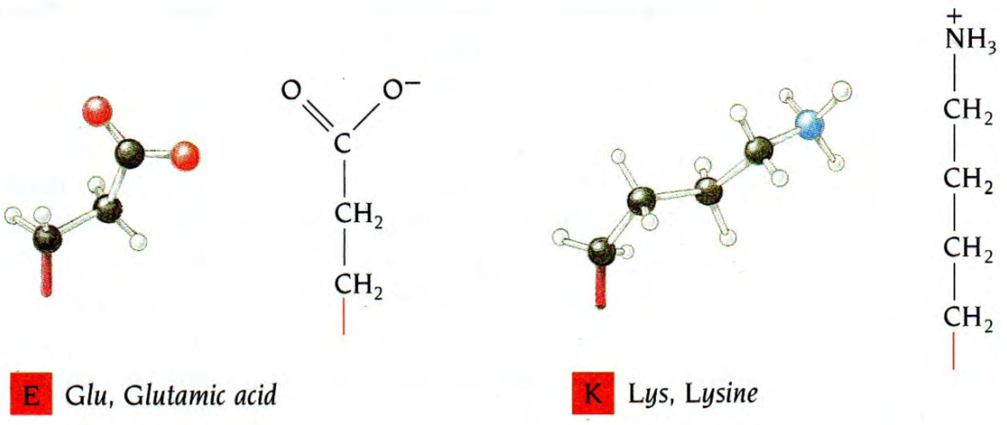
- Tertiary structure: The full 3D structure

- Quaternary structure: Association of several proteins/peptide chains into protein complexes

Hemoglobin, 1GZX



Residue properties



pK$_a$ depends on local environment

e.g. Glu close to negatively charged moiety: higher pK$_a$
Glu close to Lys is more willing to give off H$^+$, *i.e.* lower pK$_a$

pK$_a$ = 4.25

pK$_a$ = 10.53

Free amino acid

N-terminal amino group, pK$_a$ ~ 7.4
C-terminal acidic group, pK$_a$ ~ 3.9

In a protein

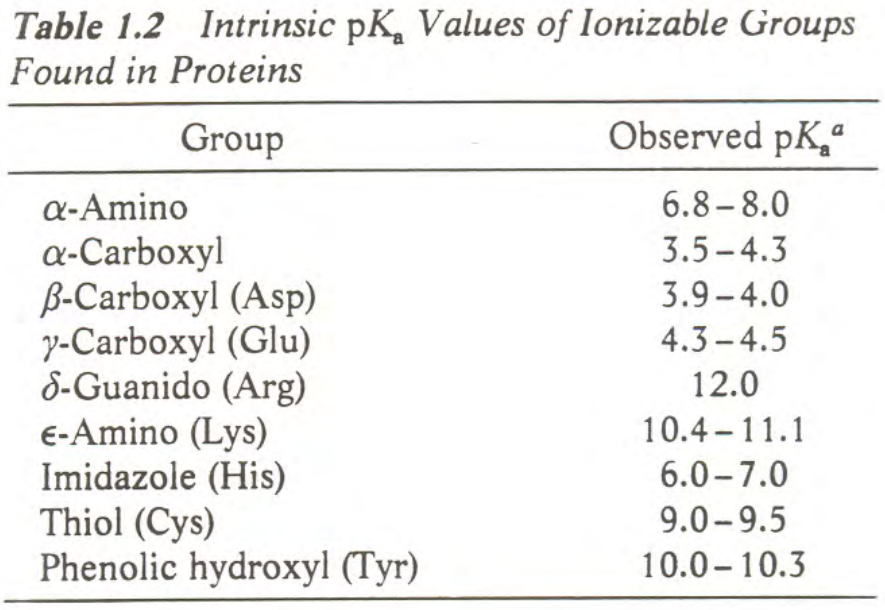
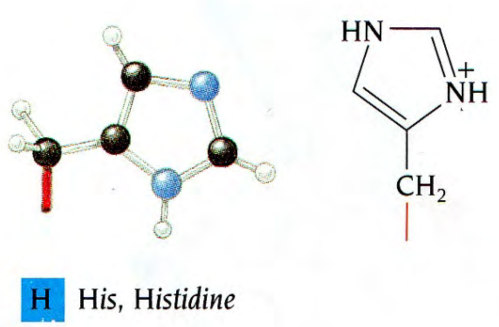


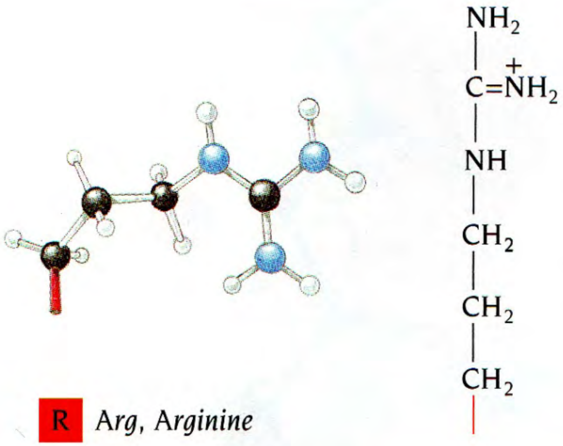
Table 1.2 *Intrinsic* pK_a *Values of Ionizable Groups Found in Proteins*

Group	Observed pK_a[a]
α-Amino	6.8 – 8.0
α-Carboxyl	3.5 – 4.3
β-Carboxyl (Asp)	3.9 – 4.0
γ-Carboxyl (Glu)	4.3 – 4.5
δ-Guanido (Arg)	12.0
ϵ-Amino (Lys)	10.4 – 11.1
Imidazole (His)	6.0 – 7.0
Thiol (Cys)	9.0 – 9.5
Phenolic hydroxyl (Tyr)	10.0 – 10.3

Proteins, T.E. Creighton (Freeman, New York, 1997)

Residue properties





Arg is "always" positively charged with pKa close to 12

His has pK_a close to 7 and the local environment is often tuned to to give correct acid/base chemistry. Strong base at neutral pH/Strong nucleophile. Often a catalytic residue.

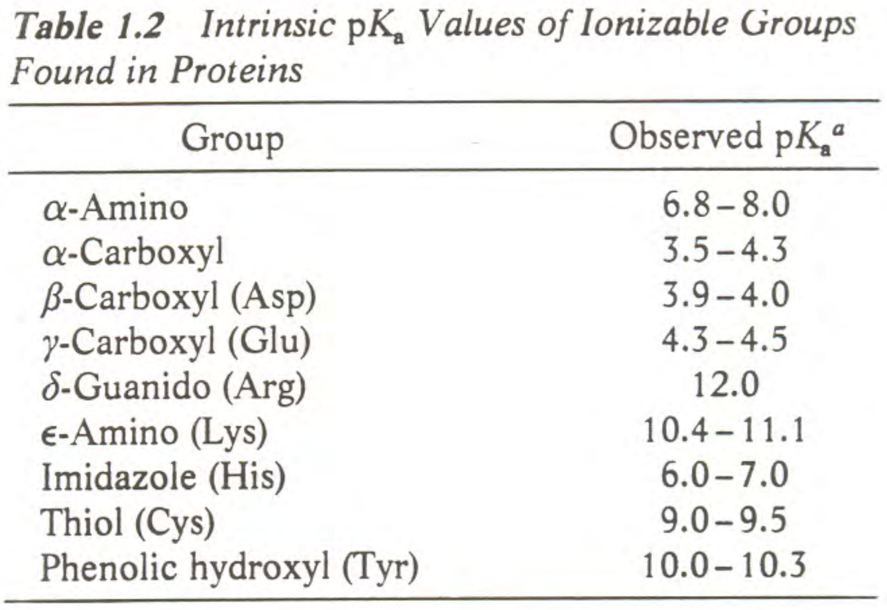
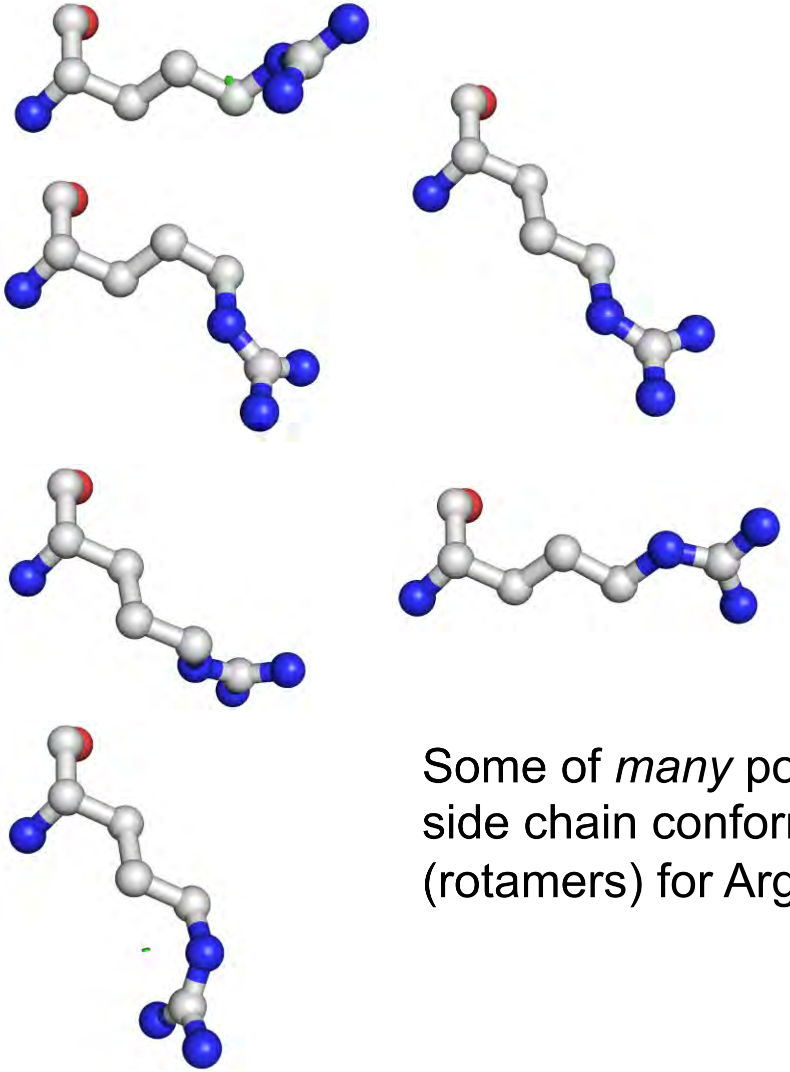


Table 1.2 *Intrinsic pK_a Values of Ionizable Groups Found in Proteins*

Group	Observed pK_a[a]
α-Amino	6.8 – 8.0
α-Carboxyl	3.5 – 4.3
β-Carboxyl (Asp)	3.9 – 4.0
γ-Carboxyl (Glu)	4.3 – 4.5
δ-Guanido (Arg)	12.0
ϵ-Amino (Lys)	10.4 – 11.1
Imidazole (His)	6.0 – 7.0
Thiol (Cys)	9.0 – 9.5
Phenolic hydroxyl (Tyr)	10.0 – 10.3

Proteins, T.E. Creighton (Freeman, New York, 1997)

Side chain conformations (Rotamers)



Some of *many* possibly side chain conformations (rotamers) for Arg

Analysis of many structures have shown that residues prefer one or a few conformations. These are called *rotamers* and are collected and distributed in *rotamer libraries*

These libraries are used in computational modeling of protein 3D structure.

Very simply put:

1. Determine overall 3D structure of backbone
2. Add side chains
3. Optimize side chains using conformations from rotamer libraries

Stabilizing forces

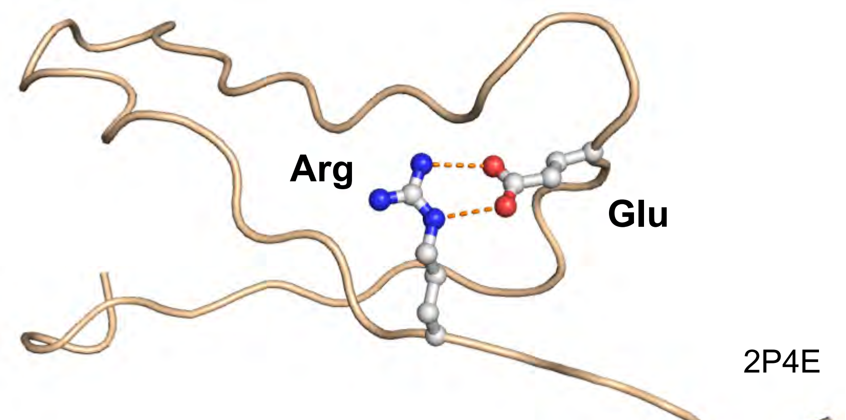
What is making proteins fold and associate into a well-defined 3D structure?

- Electrostatic interactions (salt bridges)

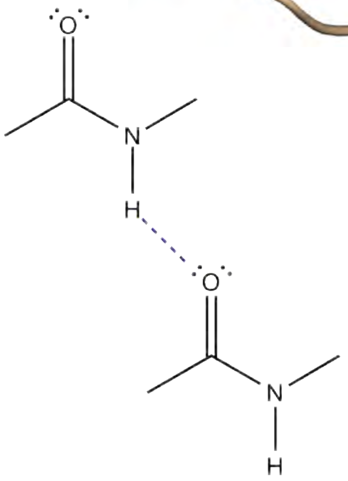
- Hydrogen bonds (H-bonds)

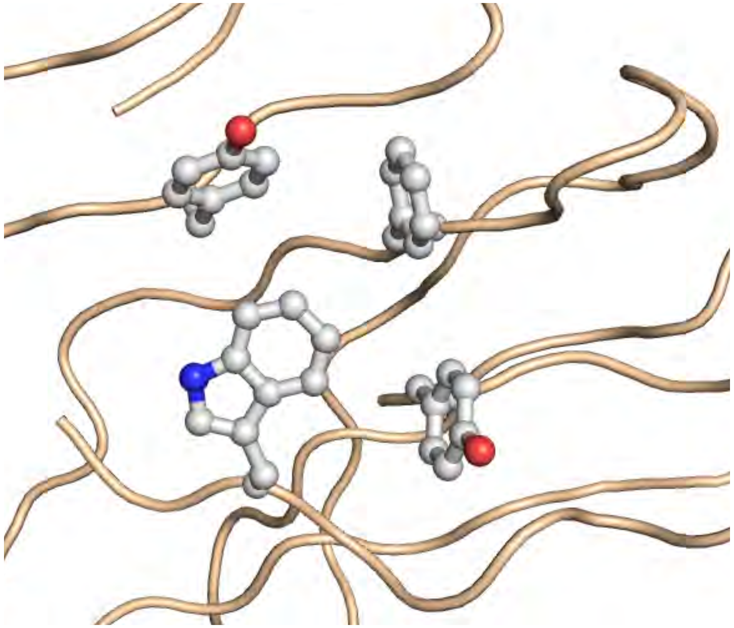
- van der Waals forces (weak)

- IMPORTANT: Hydrophobic interaction forces (minimizing the surface area of hydrophobic side chains exposed to solvent)



2P4E

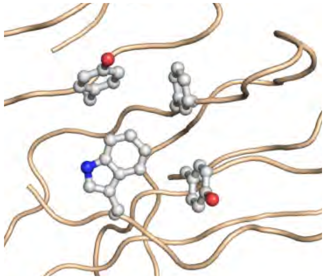


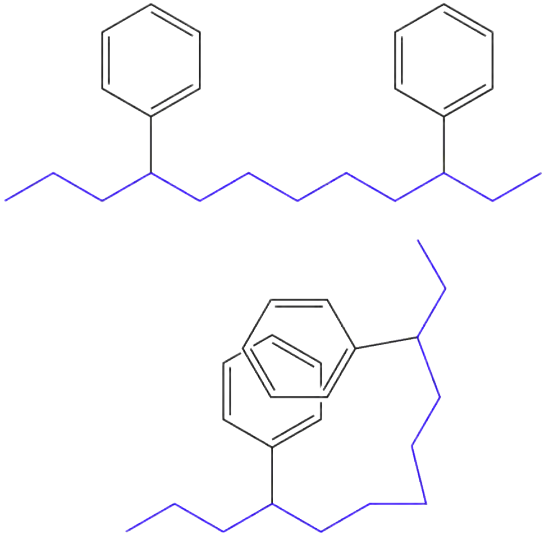


2P4E

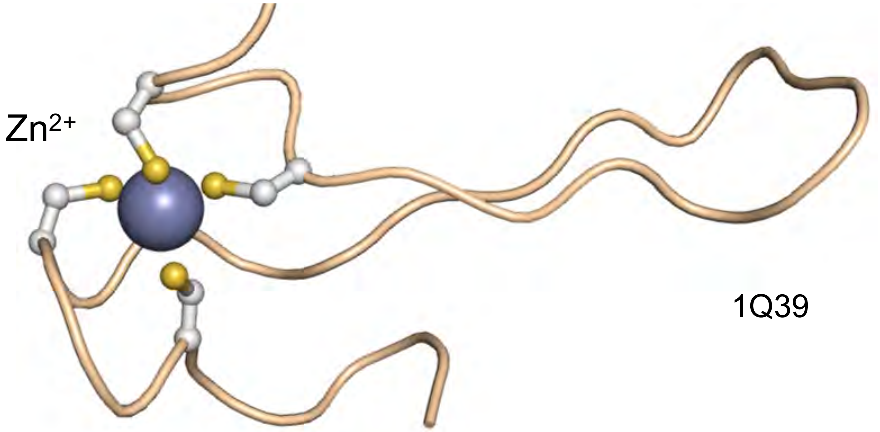
Stabilizing forces

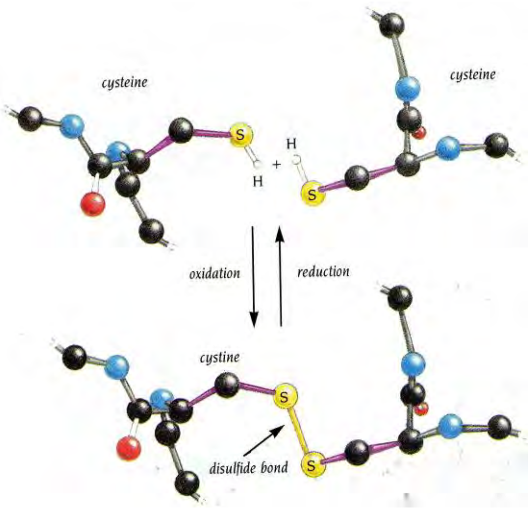
IMPORTANT: Hydrophobic interaction forces (minimizing the surface area of hydrophobic side chains exposed to solvent)





Reduced surface area exposed to solvent (water) for the hydrophobic side chains





Covalent Cys-Cys disulfide bonds

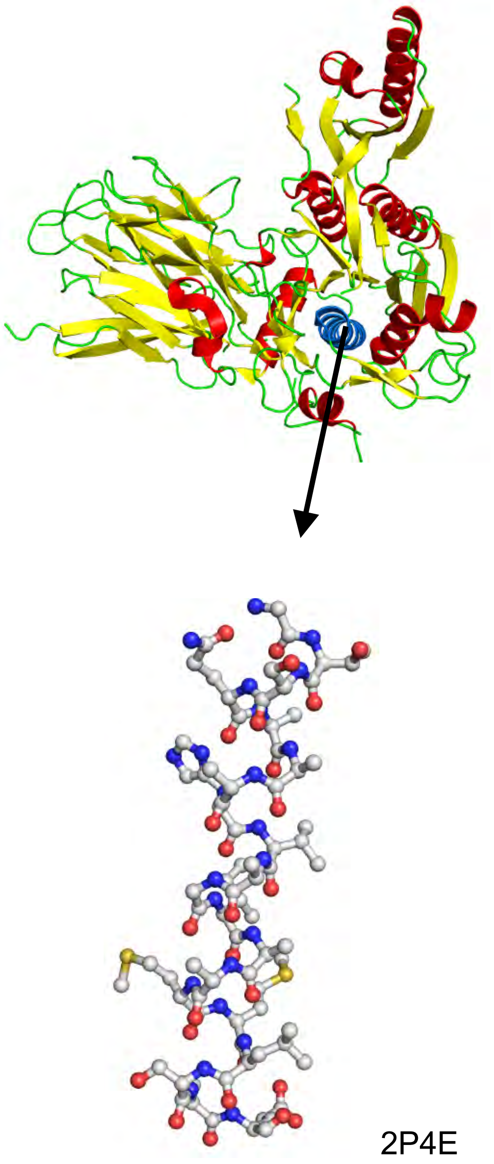
Introduction to Protein Structure,
C. Branden & J. Tooze
(Garland, New York, 1998)

Metal ions may stabilize the protein structure (e.g. in zinc fingers)

Protein folding

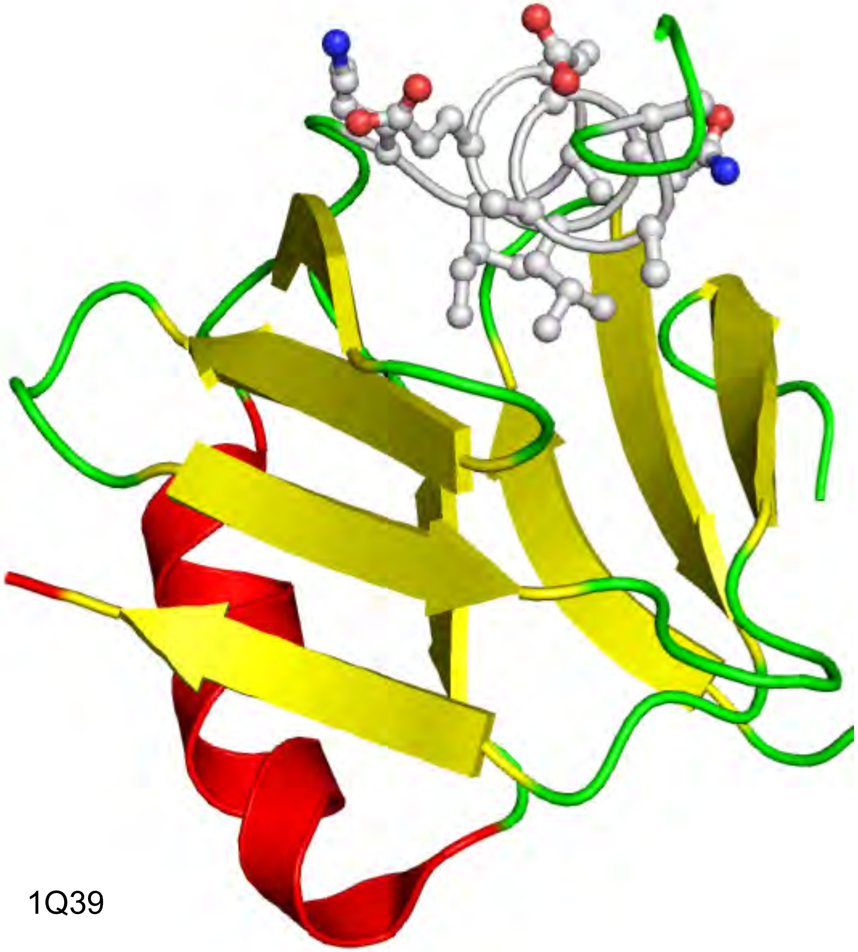
What is making proteins fold and associate into a well-defined 3D structure?

- Proteins are often found in water and both protein-protein and protein-water interactions must be taken into account (*i.e.* interactions in folded vs. denatured state)
- *Dominant* forces responsible for tertiary structure are (believed to be) the hydrophobic interaction forces
    - Residues with hydrophobic side chains are packed in the interior of the protein
    - Charged and polar residues tend to be on the protein surface
    - Polar backbone in the protein interior is "hidden" by building secondary structure elements
- Polar residue side chains in the core must be "neutralized" by interacting with other residues, e.g. in H-bond donor-acceptor pairs
- Charged residue side chains in the core must be "neutralized" by interacting with other residues through salt bridges



Protein folding

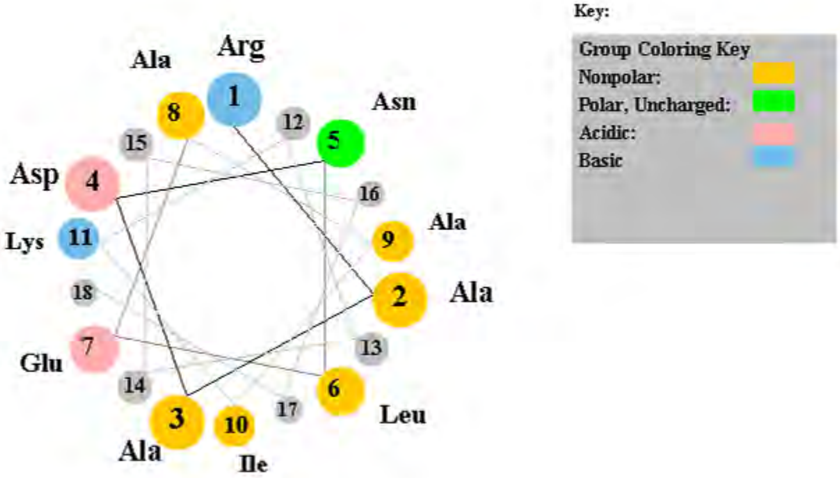
Secondary structure elements (α-helices & β-sheets) on the surfaces of proteins are often amphipathic (one hydrophilic and one hydrophobic side)



1Q39

“Pattern” of every 3-4 residues hydrophobic

Patterns can be used for predictions by computational methods, *e.g.* predict secondary structure from primary sequence

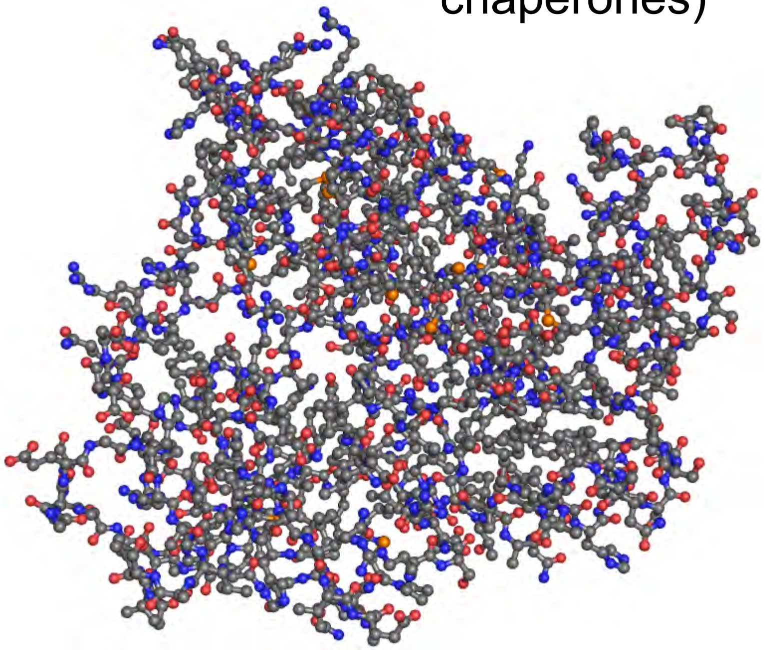


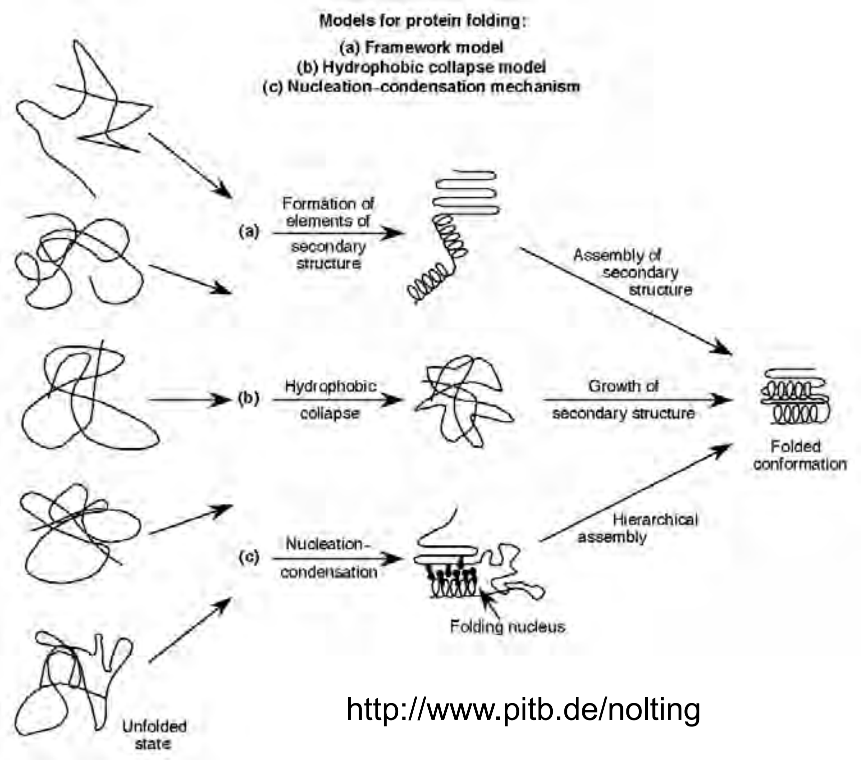
http://cti.itc.virginia.edu/~cmg/Demo/wheel/wheelApp.html

Protein folding

**TLASTPALWASIPCPRSELRLDLV
LPSGQS**

Folding is spontaneous in the cell (but often with helper molecules, chaperones)




Models for protein folding:
(a) Framework model
(b) Hydrophobic collapse model
(c) Nucleation-condensation mechanism
(a)
Formation of elements of secondary structure
Assembly of secondary structure
(b)
Hydrophobic collapse
Growth of secondary structure
Folded conformation
(c)
Nucleation-condensation
Hierarchical assembly
Folding nucleus
Unfolded state

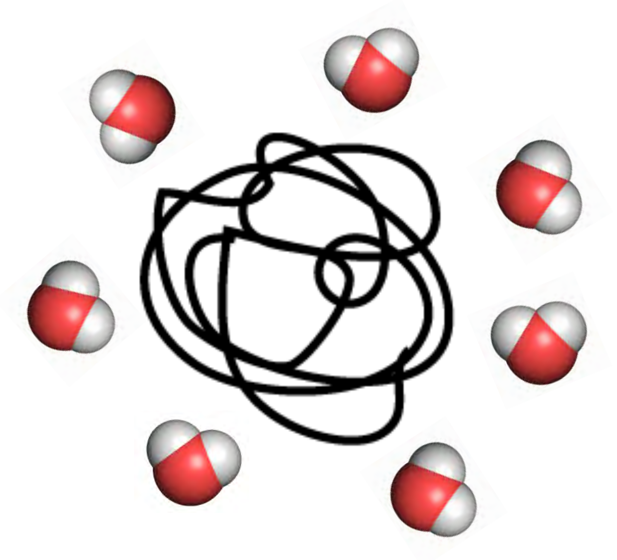
http://www.pitb.de/nolting

Put *very* simply:

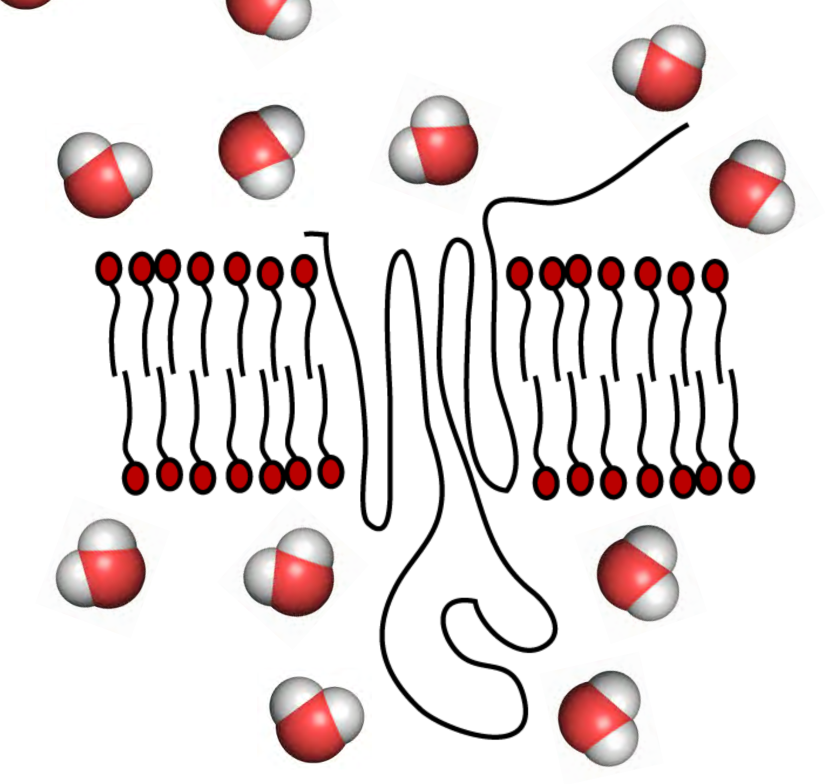
1. Secondary structure forms transiently
2. Hydrophobic collapse, formation of stable secondary structure
3. Folding completes, formation of tertiary interactions

Globular vs. membrane proteins

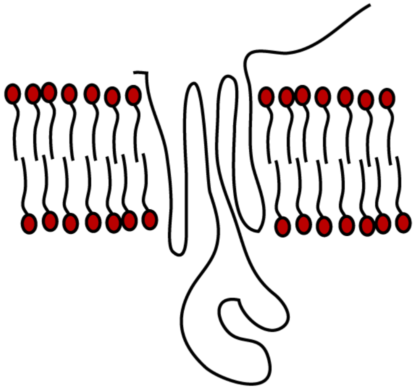
Globular proteins
- Soluble
- Surrounded by water

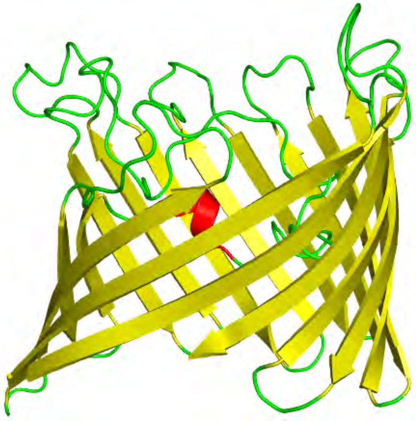


Membrane proteins
- In lipid bilayers
- Hydrophobic surface facing membrane interior

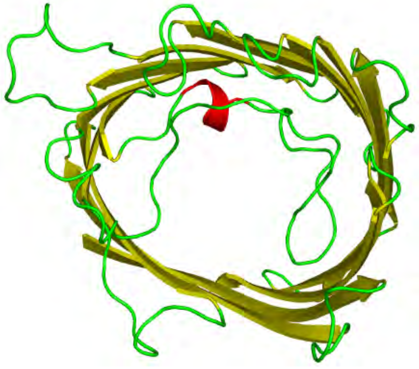


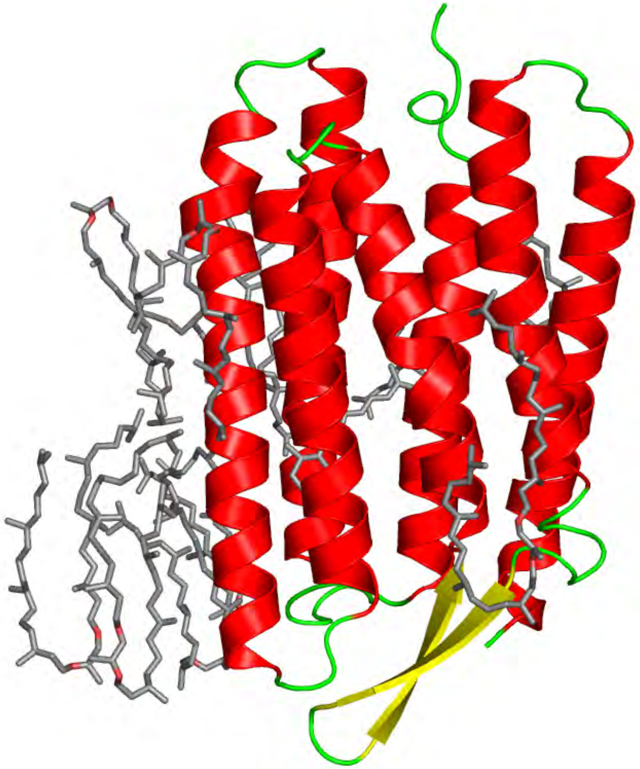
Membrane proteins





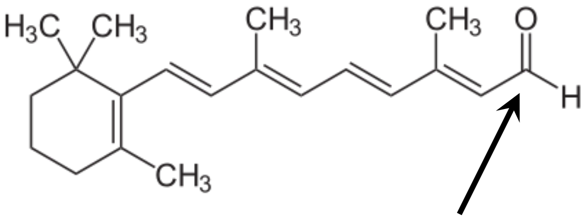
Beta-barrel porin (1PRN)





Rhodopsin (1QHJ)

Co-factor/prosthetic group retinal:



Covalent (Schiff bond) linkage to protein Lys residue

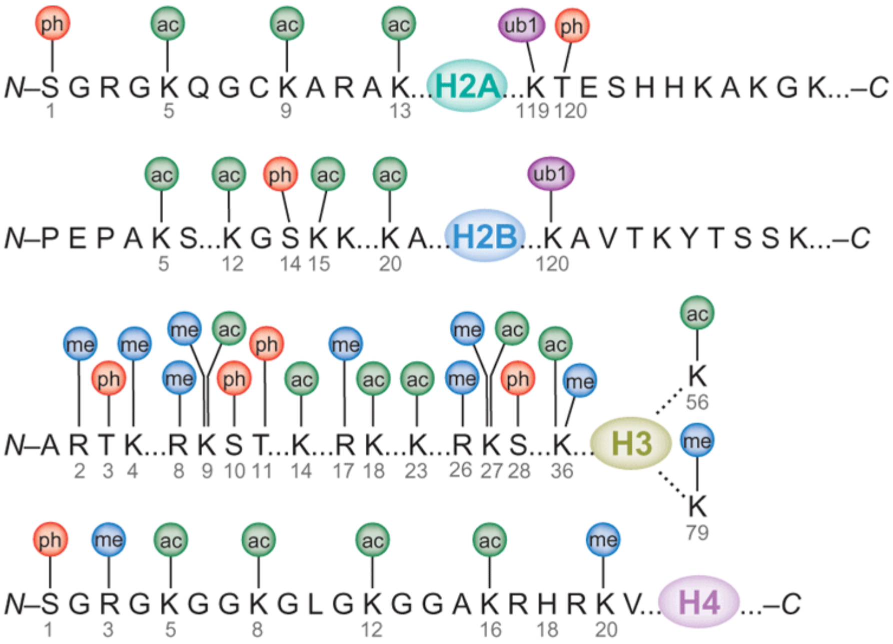
Many apo-proteins need co-factors/prosthetic groups to become functional

PTMs

Post-translational modifications (PTMs), *i.e.* chemical modification after translation, *e.g.*

- Glycosylation (addition of sugar groups to *e.g.* Asn, Ser, or Thr)
- Phosphorylation of Ser/Thr by kinases
- Methylation of Lys in histones
- Ubiquitination (addition of the protein ubiquitin to Lys)
- Methionine aminopeptidases may remove N-terminal Met
- *Many, many more!!*

Bhaumik *et al.*, *Nat. Struct. Mol. Biol.* **14**, 1008 (2007)


ph ac ac ac ub1 ph
N–S G R G K Q G C K A R A K...H2A...K T E S H H K A K G K...–C
1 5 9 13 119 120
ac ac ph ac ac ub1
N–P E P A K S...K G S K K...K A...H2B...K A V T K Y T S S K...–C
5 12 14 15 20 120
me me ph me ac me ph ph ac me ac ac me ac me ph ac me ac
N–A R T K...R K S T...K...R K...K...R K S...K...H3
2 3 4 8 9 10 11 14 17 18 23 26 27 28 36
K 56
me
K 79
ph me ac ac ac ac me
N–S G R G K G G K G L G K G G A K R H R K V...H4...–C
1 3 5 8 12 16 18 20

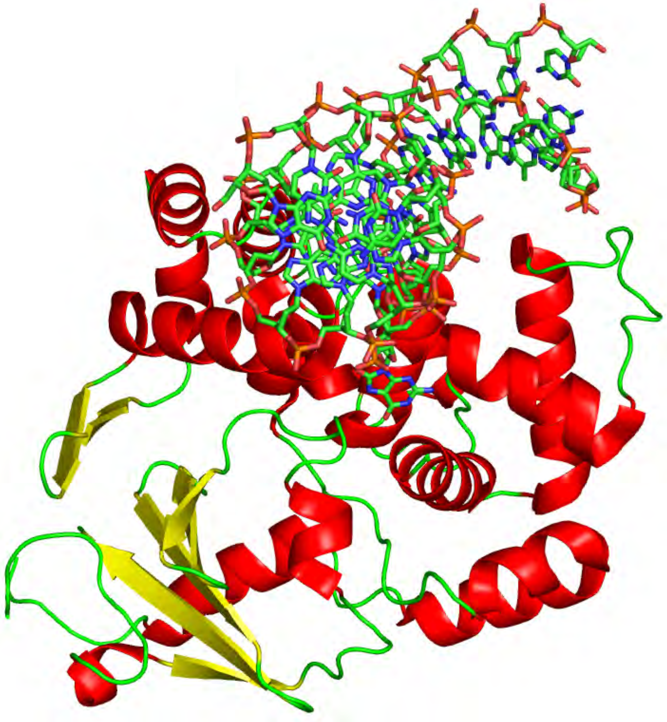
PTMs of human histones include acetylation (ac), methylation (me), phosphorylation (ph) and ubiquitination (ub1)

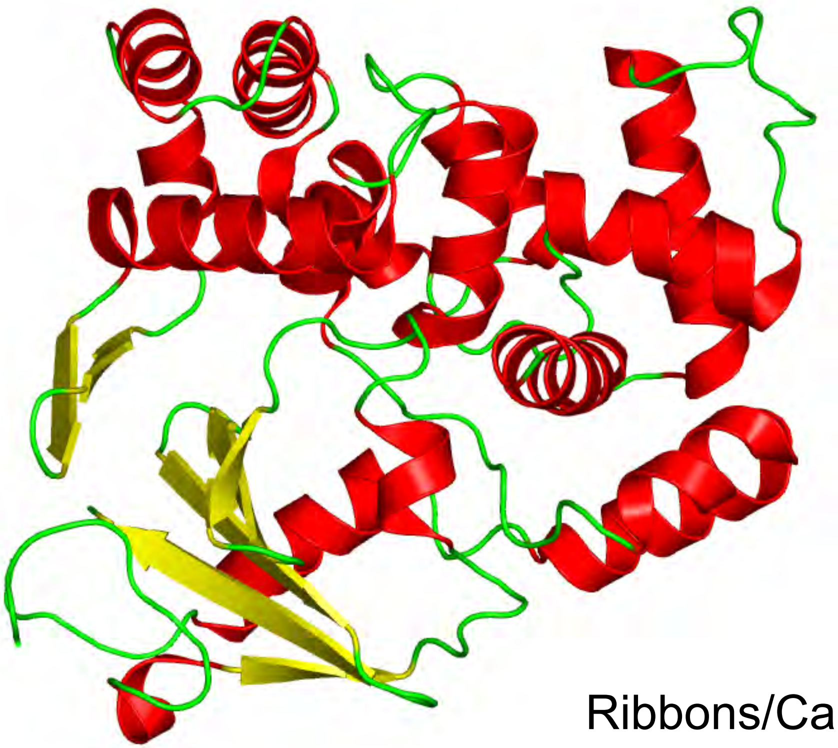
Even if you know the complete 3D structure of the apo-protein you may be unable to understand the function of the protein if you have no information about the PTMs!

Visualization of protein structure

Jon K. Lærdahl,
Structural Bioinformatics

Human OGG1, a DNA repair enzyme that recognizes and excises oxidized DNA bases





Ribbons/Cartoon

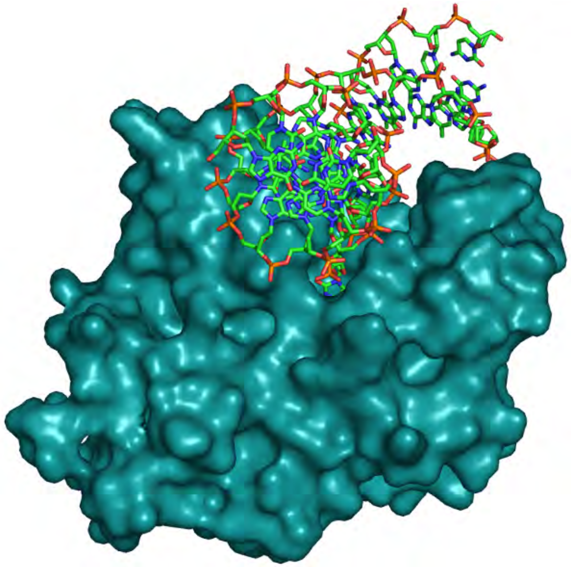
Software (advanced graphics rendering):

- RasMol
- Swiss-PDBViewer (freeware; also homology modeling)
- Molscript (command-line-based)
- Jmol (open-source Java viewer)
- PyMOL (open-source, user-sponsored)
- Many more both free and very expensive

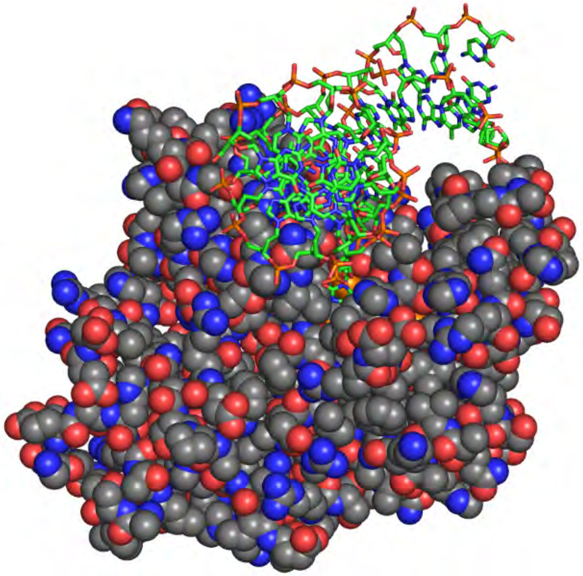
We will use some of these at the Exercises!

Visualization of protein structure

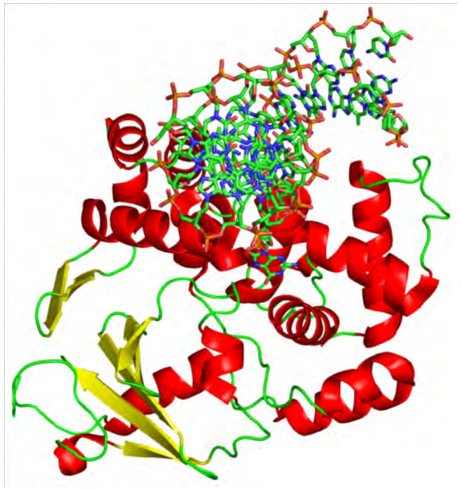
Human OGG1, a DNA repair enzyme that recognizes and excises oxidized DNA bases

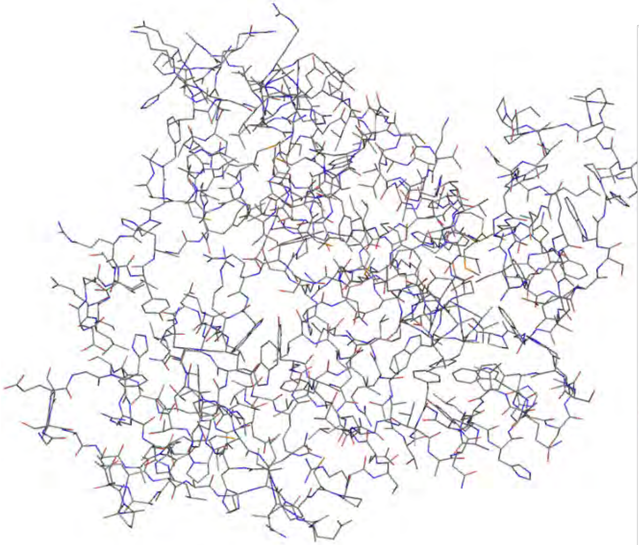


Surface

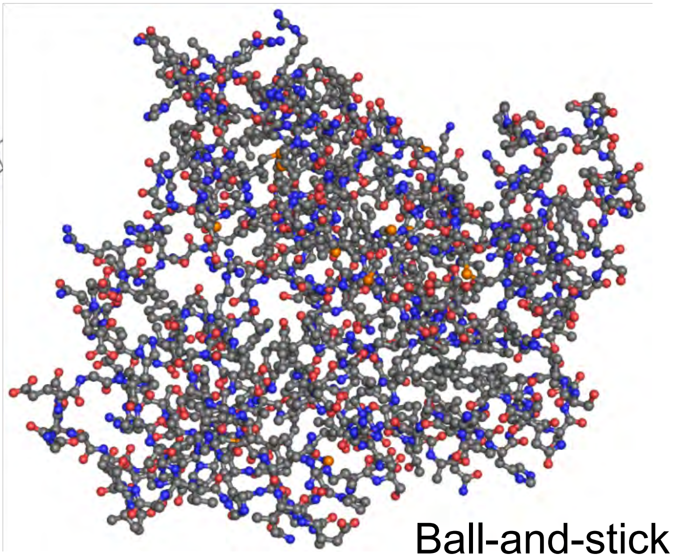


Space-filling spheres (CPK)





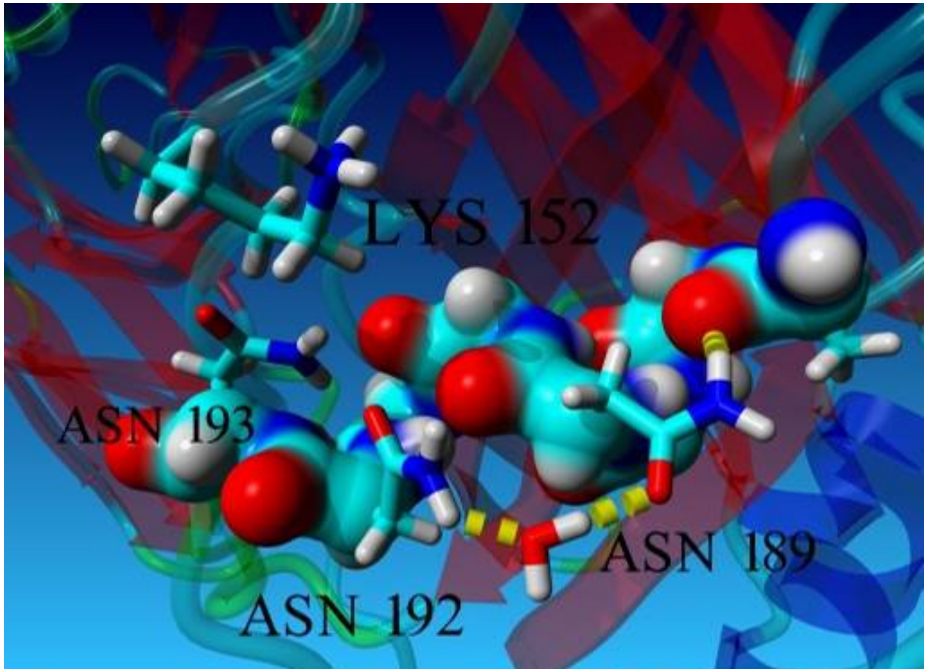
Wireframes

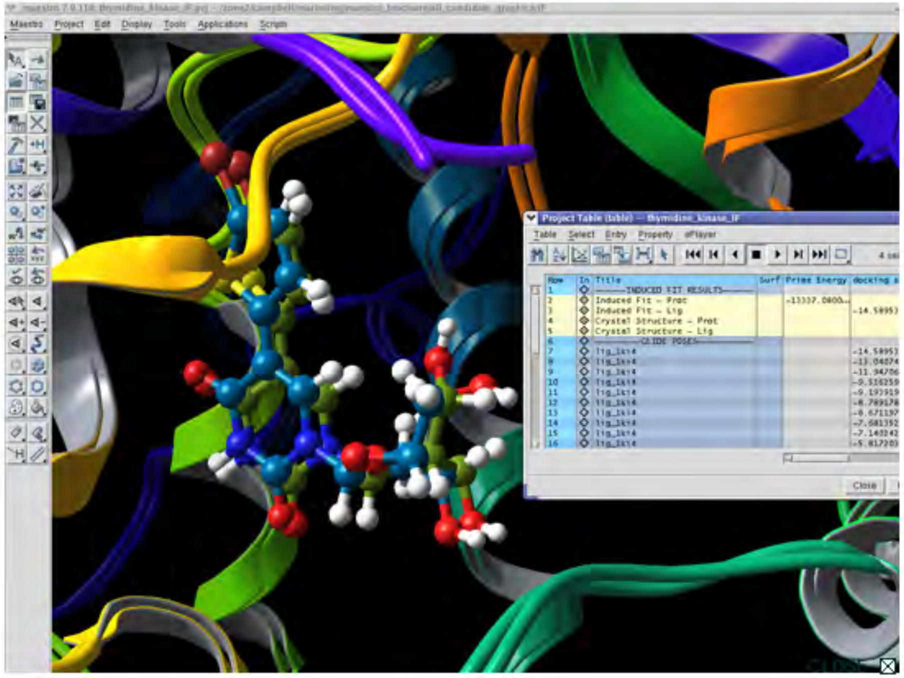


Ball-and-stick

Visualization of protein structure

YASARA





Project Table (table) — thymidine_kinase_IF

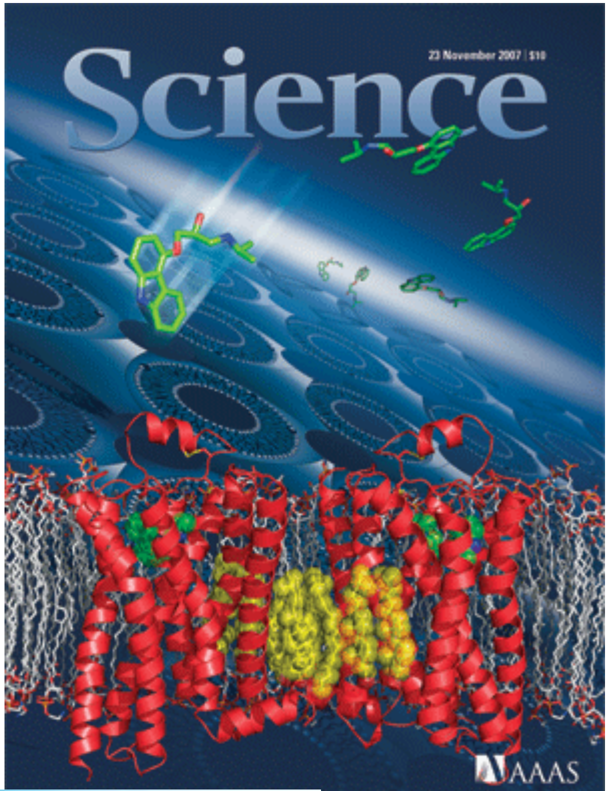
Row	In	Title	Surf	Prime Energy	docking s
1	◇	———INDUCED FIT RESULTS———			
2	◇	Induced Fit – Prot		-13337.0800...	
3	◇	Induced Fit – Lig			-14.58953
4	◇	Crystal Structure – Prot			
5	◇	Crystal Structure – Lig			
6	◇	———GLIDE POSES———			
7	◇	lig_1ki4			-14.58953
8	◇	lig_1ki4			-13.04074
9	◇	lig_1ki4			-11.94706
10	◇	lig_1ki4			-9.516259
11	◇	lig_1ki4			-9.193919
12	◇	lig_1ki4			-8.789178
13	◇	lig_1ki4			-8.671197
14	◇	lig_1ki4			-7.681352
15	◇	lig_1ki4			-7.140242
16	◇	lig_1ki4			-5.817203

Maestro

PyMOL

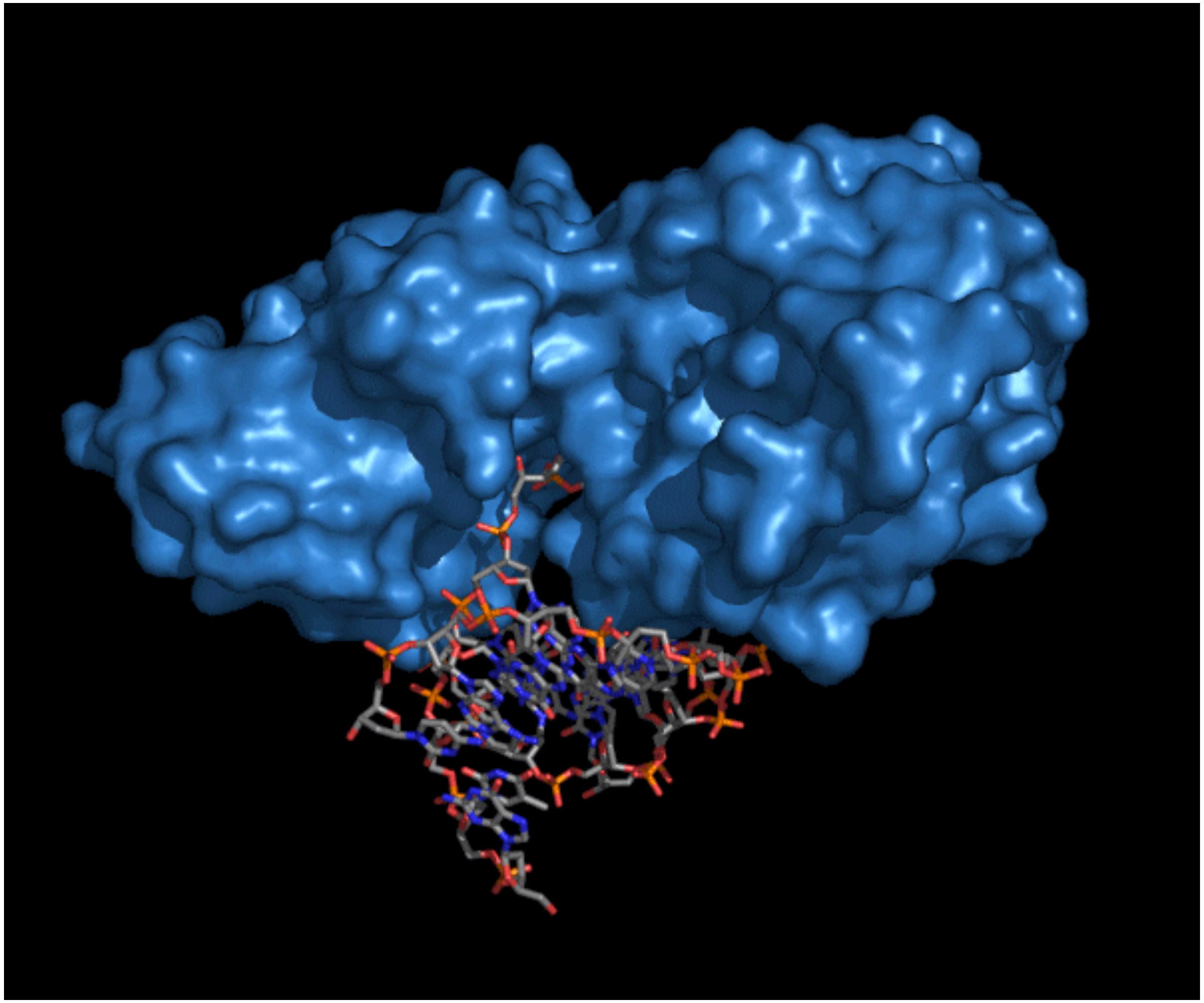
Visualization of protein structure

Jon K. Lærdahl,
Structural Bioinformatics



Publication quality graphics from PyMOL

Movies, interactivity etc.



The structure of *Bacillus stearothermophilus* Fpg protein borohydride-trapped with DNA oligo as determined by Fromme and Verdine, *Nat. Struct. Biol.* **9**, 544 (2002), PDB: 1L1Z.

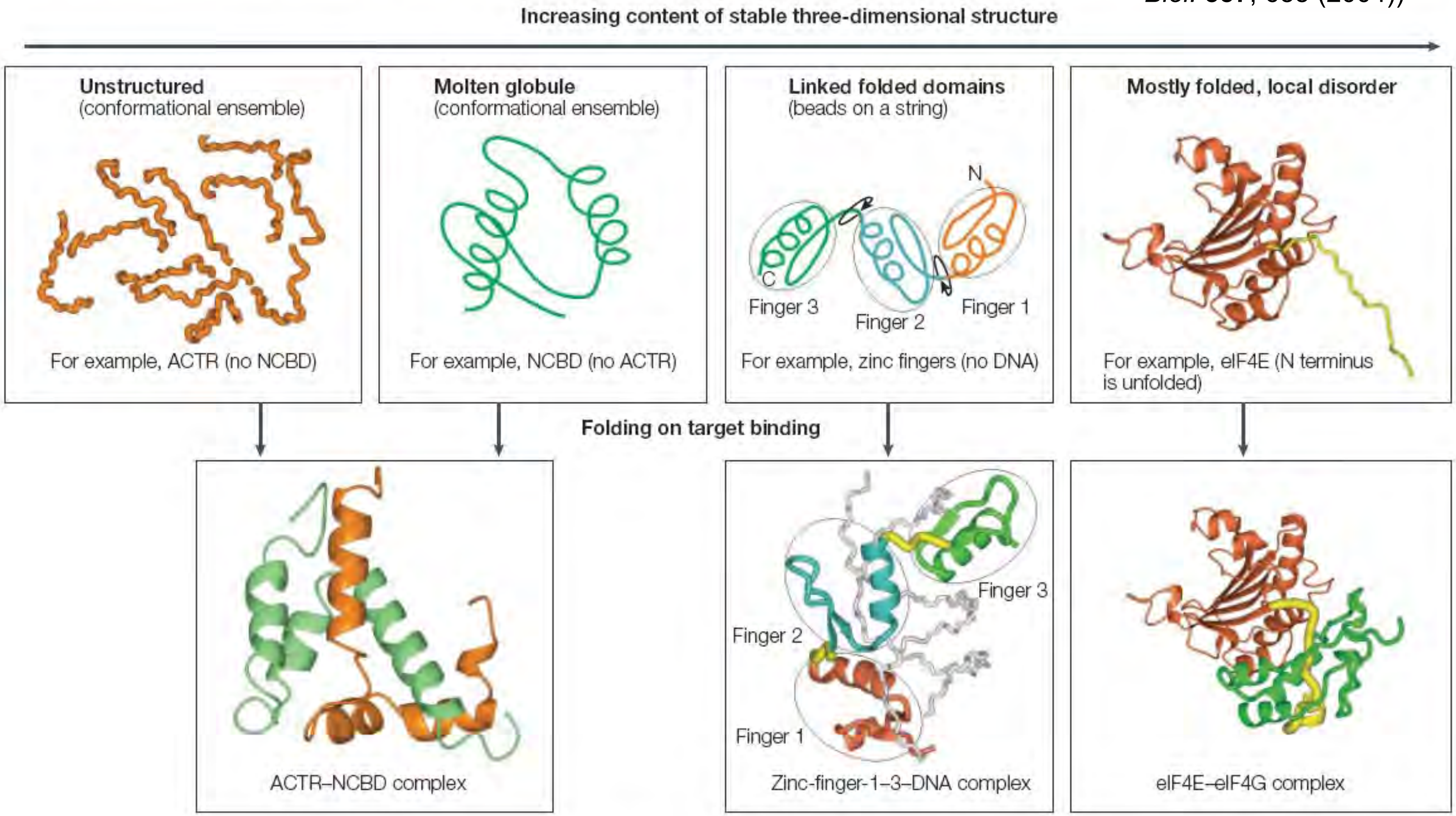
The graphics was generated with PyMOL

Structural disorder in proteins

Jon K. Lærdahl,
Structural Bioinformatics

- Not all proteins have a regular 3D structure for the full sequence
- The full protein, segments or small parts may be structurally disordered/intrinsically unstructured

Predicted 20% of human proteins have disordered segments of length >50 residues (1% in *E. coli*) (J.J. Ward *et al.*, *J. Mol. Biol.* **337**, 635 (2004))

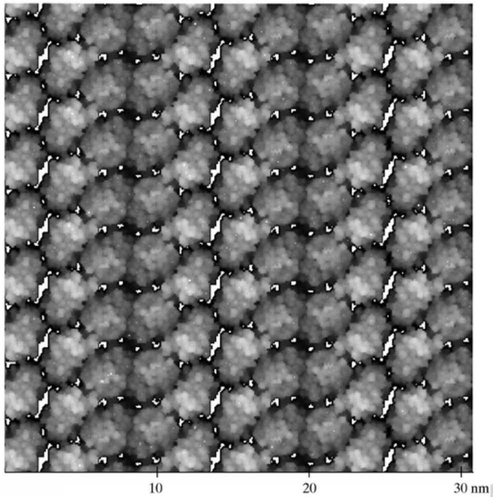

Increasing content of stable three-dimensional structure
Unstructured
(conformational ensemble)
For example, ACTR (no NCBD)
Molten globule
(conformational ensemble)
For example, NCBD (no ACTR)
Linked folded domains
(beads on a string)
N
C
Finger 3
Finger 2
Finger 1
For example, zinc fingers (no DNA)
Mostly folded, local disorder
For example, eIF4E (N terminus is unfolded)
Folding on target binding
ACTR–NCBD complex
Finger 3
Finger 2
Finger 1
Zinc-finger-1–3–DNA complex
eIF4E–eIF4G complex

H.J. Dyson & P.E. Wright, *Nat. Rev. Mol. Cell Biol*. **6**, 197 (2005)

Experimental determination of protein structure – X-ray Crystallography

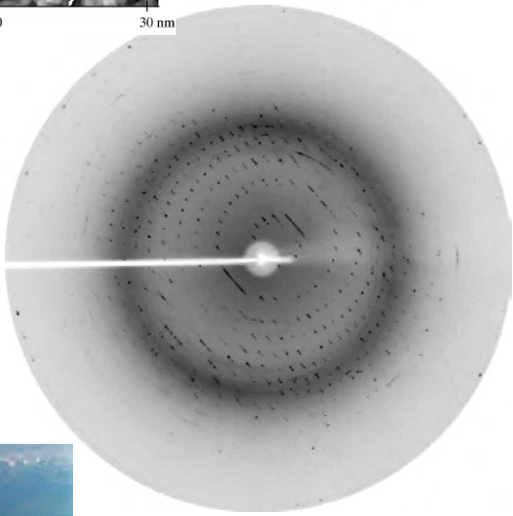
Jon K. Lærdahl,
Structural Bioinformatics

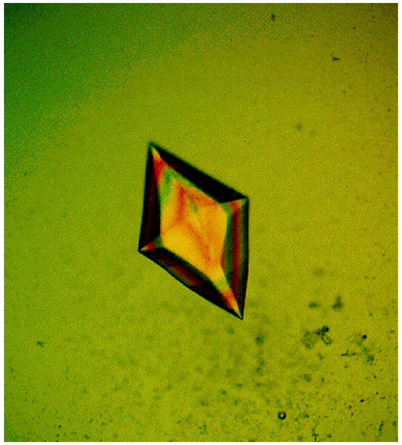
- Necessary to grow protein crystals
    - Often (extremely) difficult
- Diffraction in X-ray beam
- Must solve “phase problem” (due to unknown timing of diffraction waves hitting the detector):
    - Molecular replacement (use the known structure of similar protein)
    - Multiple isomorphus replacement (generate crystals with heavy atoms, *e.g.* by soaking)
- Strong X-ray source needed to get high accuracy (Synchrotron)

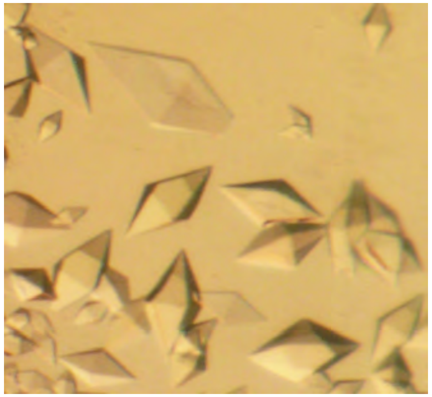


Li *et al*., *Acta Cryst.* **D55**, 1023 (1999)

Proteins are located in a lattice, in a repeated and oriented fashion





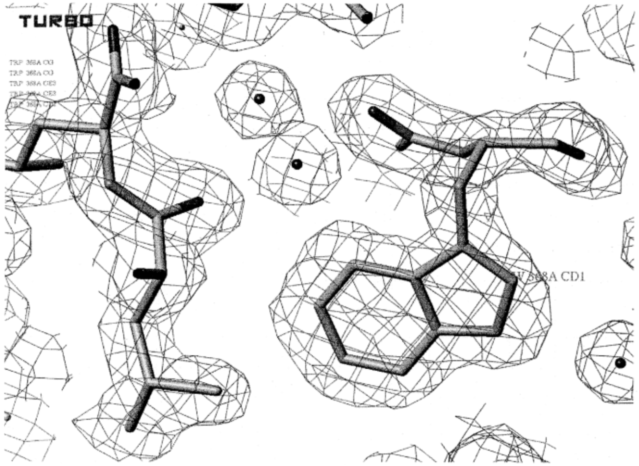


Experimental determination of protein structure – X-ray Crystallography

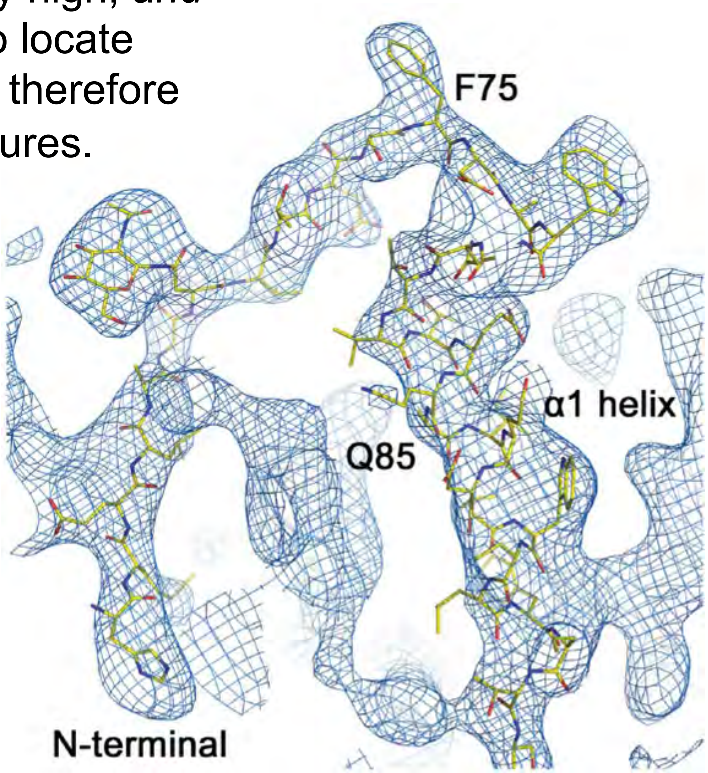
Diffraction pattern & solved phases: Electron density map ("electron cloud"):

- Model protein primary sequence into electron density map
- Resolution:
    - Low ~5.0 Å
    - Intermediate ~2.0-2.5 Å
    - High ~1.2 Å (Only at this very high, *and rare*, resolution it is possible to locate hydrogen atoms. H-atoms are therefore usually not visible in the structures.

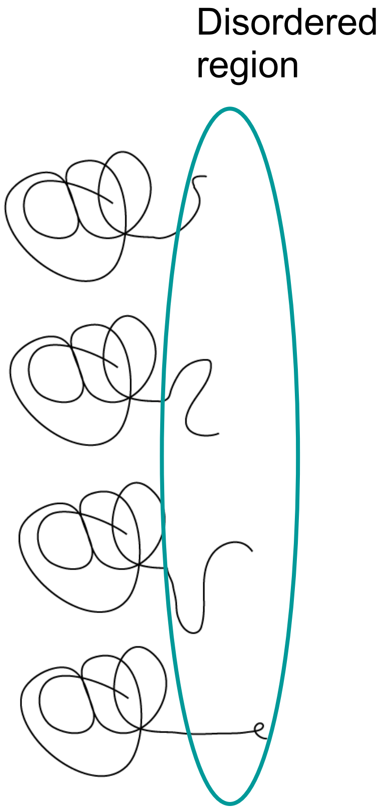
- Gives a *static* picture of the protein in the crystal which might not correspond closely to situation in solution
- Bottleneck: Crystallization (and phase problem)
- No electron density for structurally disordered regions



A.R. Slabas *et al.*, *Biochem. Soc. Trans.* **28**, 677 (2000) (1.9 Å resolution)



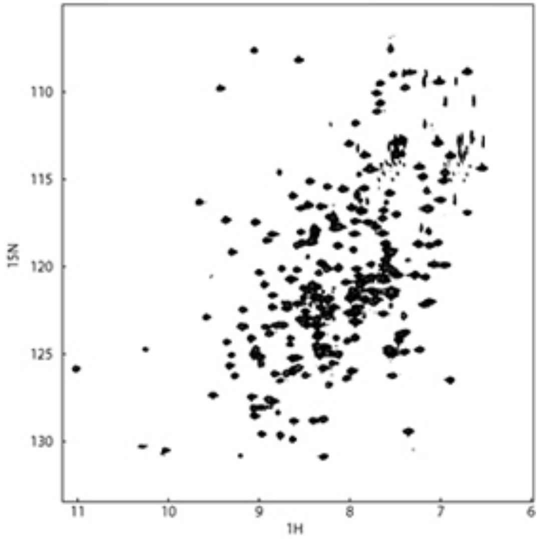
X. Chen *et al.*, *Acta Cryst.* **D65**, 339 (2009) (~3.5 Å resolution)

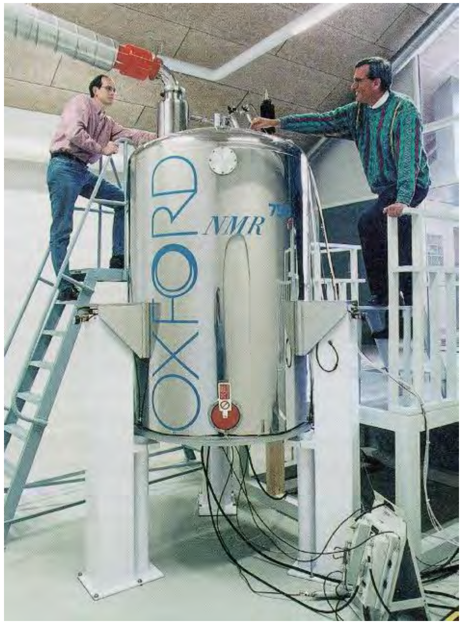


Experimental determination of protein structure – NMR Spectroscopy

Nuclear Magnetic Resonance (NMR) Spectroscopy:

- Based in energy levels of magnetic nuclei *(e.g.* ^{13}C and ^{15}N*)* in a *very* strong external magnetic field probed my a radio frequency signal
- Determines distances between all labeled atoms in a protein
- Structure model built from distances
- Structure solved in solution
    - No need to grow crystals
- Can be used to study proteins dynamics & behavior in solution
- Can currently only be employed for proteins of limited size (a few hundred residues)



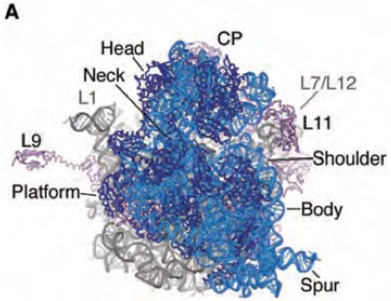


Experimental determination of protein structure

X-ray Crystallography:

Pros:

- Can be used for huge protein complexes
  - 10.000s of atoms in *e.g.* complete ribosomes



B.S. Schuwirth, *Science* **310**, 827 (2005)

- Can in fortunate cases give very high resolution (Atom position uncertainty ~0.2 Å or less)

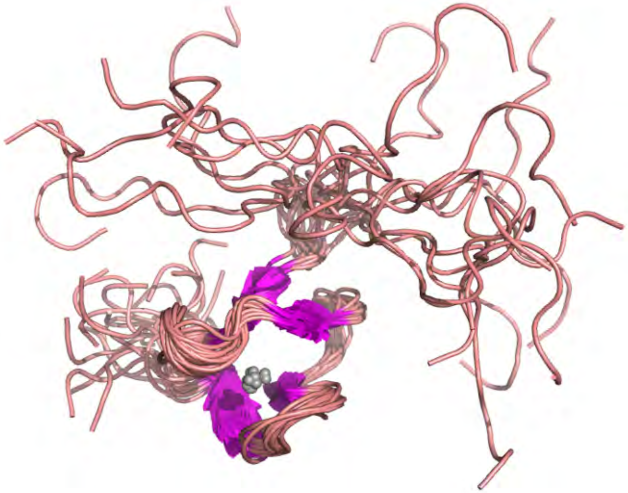
Cons:

- Usually (extremely!) tricky to grow crystals
  - Membrane proteins are particularly difficult
  - Proteins with disordered segments are difficult
- Need to solve phase problem
- Does not give insight into dynamics and protein disorder
- Large amounts of protein needed
- Usually missing H-atoms
- Disordered loops/regions are not visible

NMR Spectroscopy:

Pros:

- Can be used directly on proteins in solution
- No need for crystallization
- Dynamics studies
- Both ordered and disordered proteins (usually an ensemble of 20-40 models)



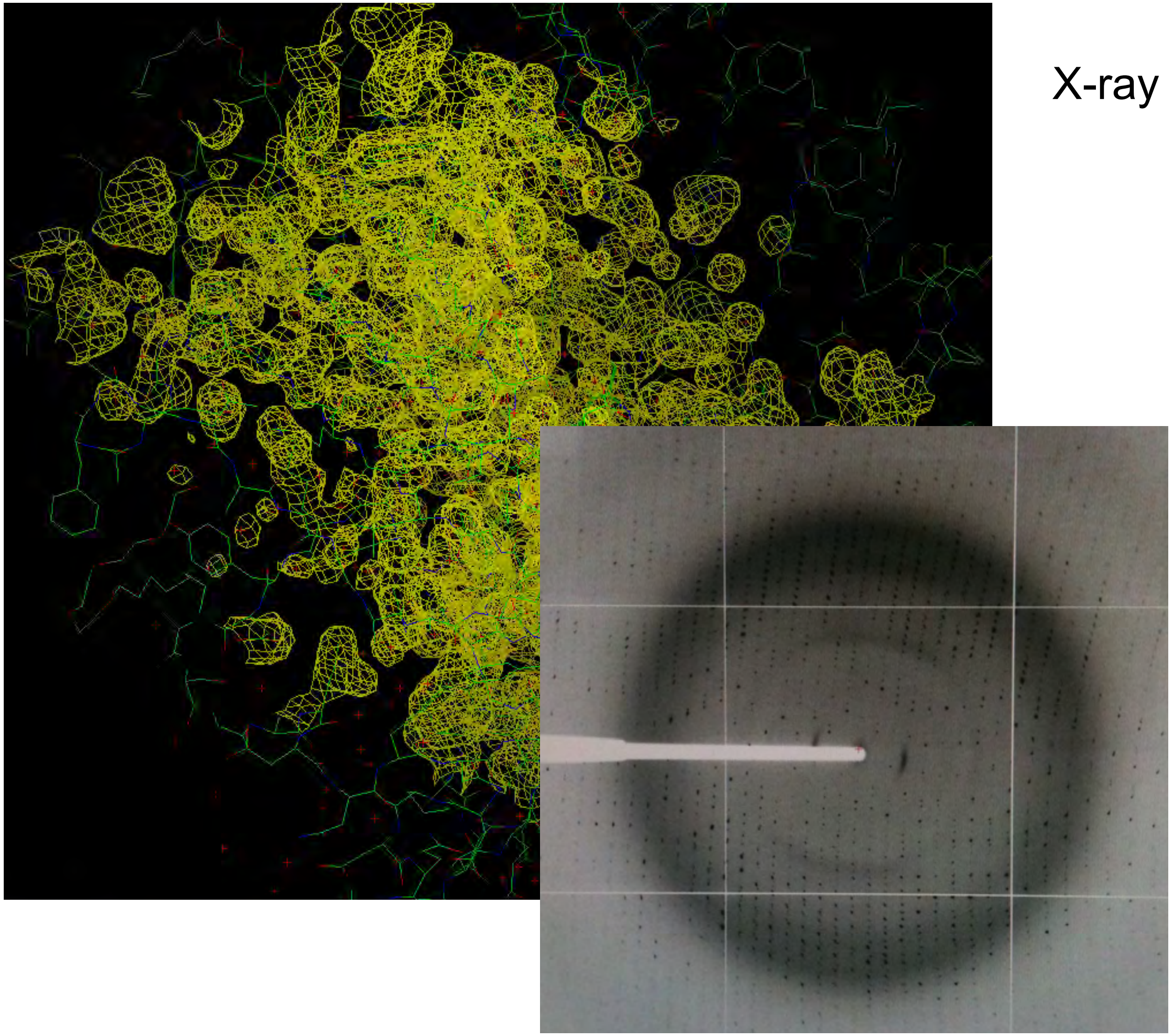
1N0Z

Cons:

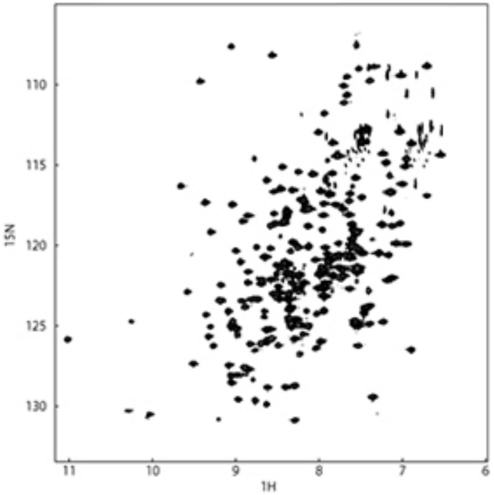
- Only applicable for small proteins (<200 residues?)
- Huge amounts of protein needed

All experimental methods: Labor intensive and requiring (very) expensive instruments
Membrane proteins *extremely tricky*
The experimental structures are also models!

Modeling of atoms into electron density

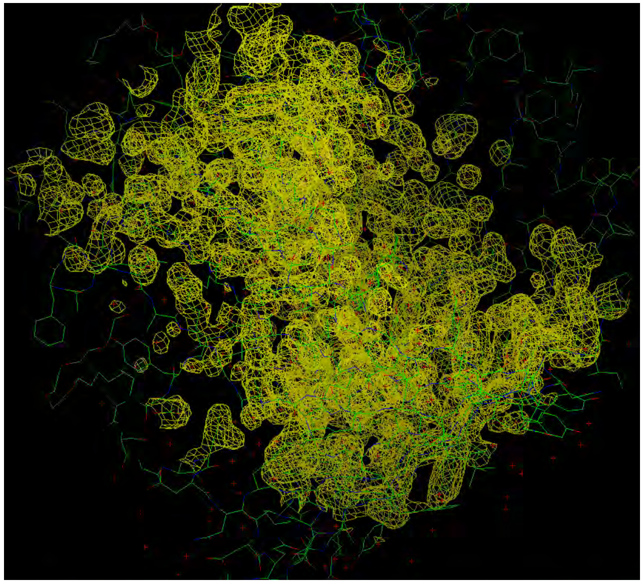


X-ray crystallography



NMR

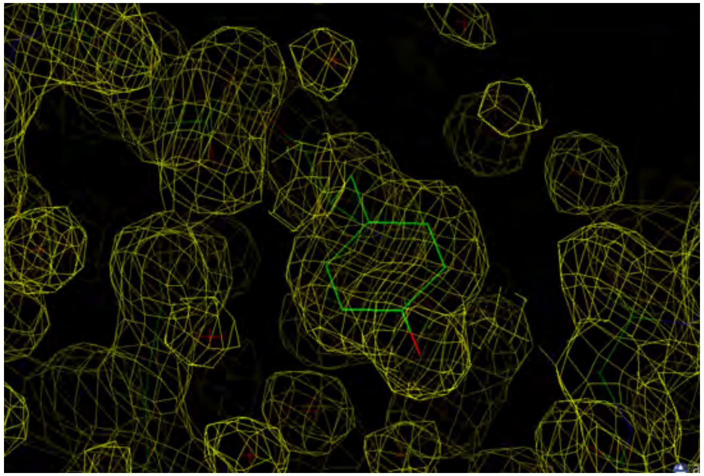
Modeling of atoms into electron density



The experimental structures are also "models"!

And heavily depends on computers/software

1PRN

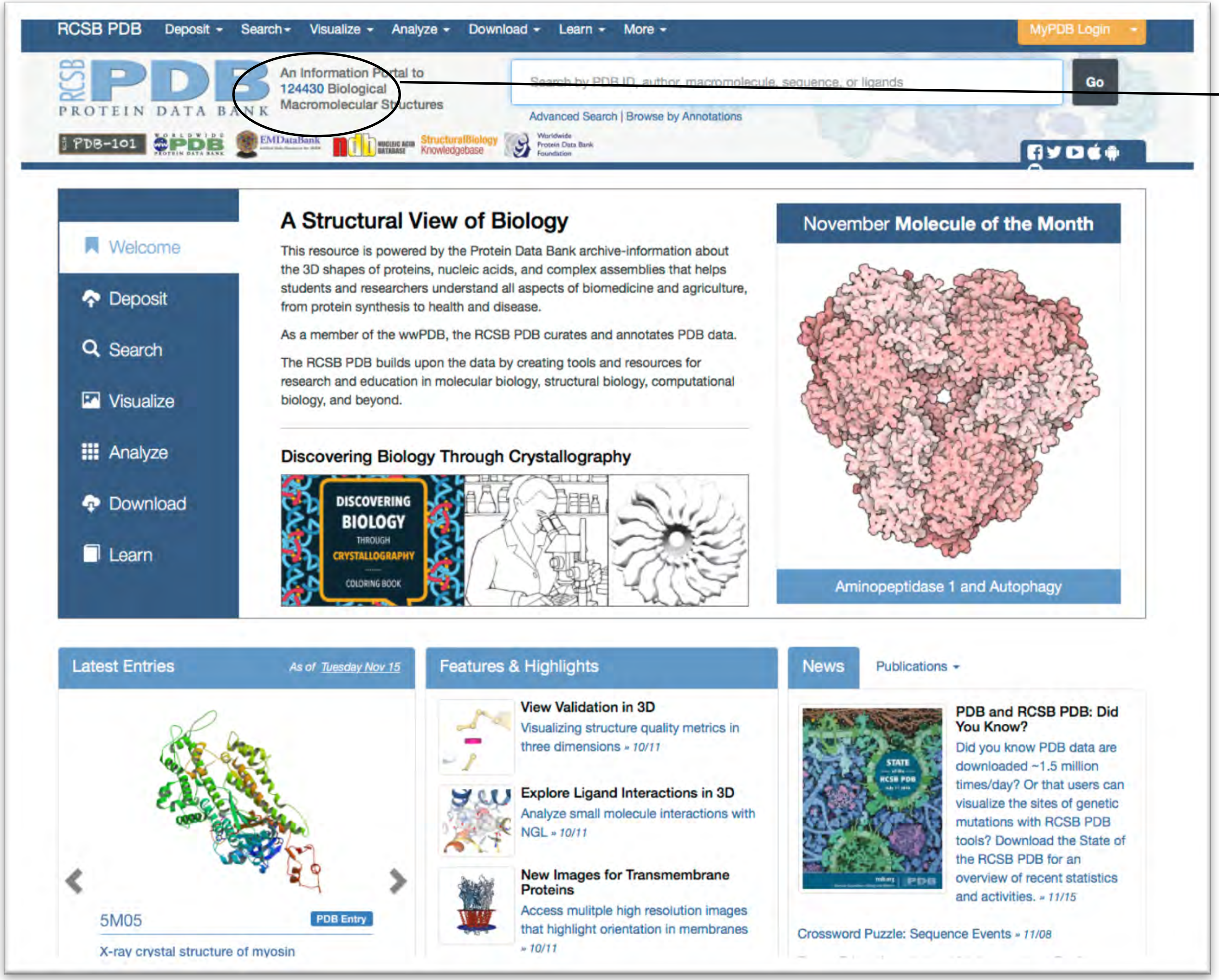


Remember, when looking at an ***experimental structure*** (X-ray):

- Resolution and R-factor gives you an idea about the quality of the experimental model
  - Resolution ~ 3 Å: side chains may be wrong rotamer or missing, main chain normally ok
  - Resolution ~ 2 Å: most side chains should be ok
  - Resolution < 1.5 Å: high accuracy structure
  - Resolution < 1.2 Å: may even be possible to determine positions for hydrogen atoms
- Due to structural flexibility or "problems" in crystals, some regions, typically loops or N-/C-terminus may have little visible electron density.
  - In some cases this gives gaps in the sequences or missing side chains
  - In other cases people put in residues/atoms anyway, in reasonable positions
  - The Uppsala Electron Density Server can be useful

Protein Structure Database

Protein Data Bank (PDB) www.rcsb.org:
The home of all experimental proteins structures


RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB Login
RCSB PDB
PROTEIN DATA BANK
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations
PDB-101
WORLDWIDE PDB PROTEIN DATA BANK
EMDataBank
ndb NUCLEIC ACID DATABASE
StructuralBiology Knowledgebase
Worldwide Protein Data Bank Foundation
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
DISCOVERING BIOLOGY THROUGH CRYSTALLOGRAPHY COLORING BOOK
STATE of the RCSB PDB
PDB Entry
Publications

An Information Portal to 124430 Biological Macromolecular Structures

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.

Discovering Biology Through Crystallography

November Molecule of the Month

Aminopeptidase 1 and Autophagy

Latest Entries

As of Tuesday Nov 15

5M05

X-ray crystal structure of myosin

Features & Highlights

View Validation in 3D
Visualizing structure quality metrics in three dimensions » 10/11

Explore Ligand Interactions in 3D
Analyze small molecule interactions with NGL » 10/11

New Images for Transmembrane Proteins
Access mulitple high resolution images that highlight orientation in membranes » 10/11

News

PDB and RCSB PDB: Did You Know?
Did you know PDB data are downloaded ~1.5 million times/day? Or that users can visualize the sites of genetic mutations with RCSB PDB tools? Download the State of the RCSB PDB for an overview of recent statistics and activities. » 11/15

Crossword Puzzle: Sequence Events » 11/08

>124,000 structures
Not all are unique

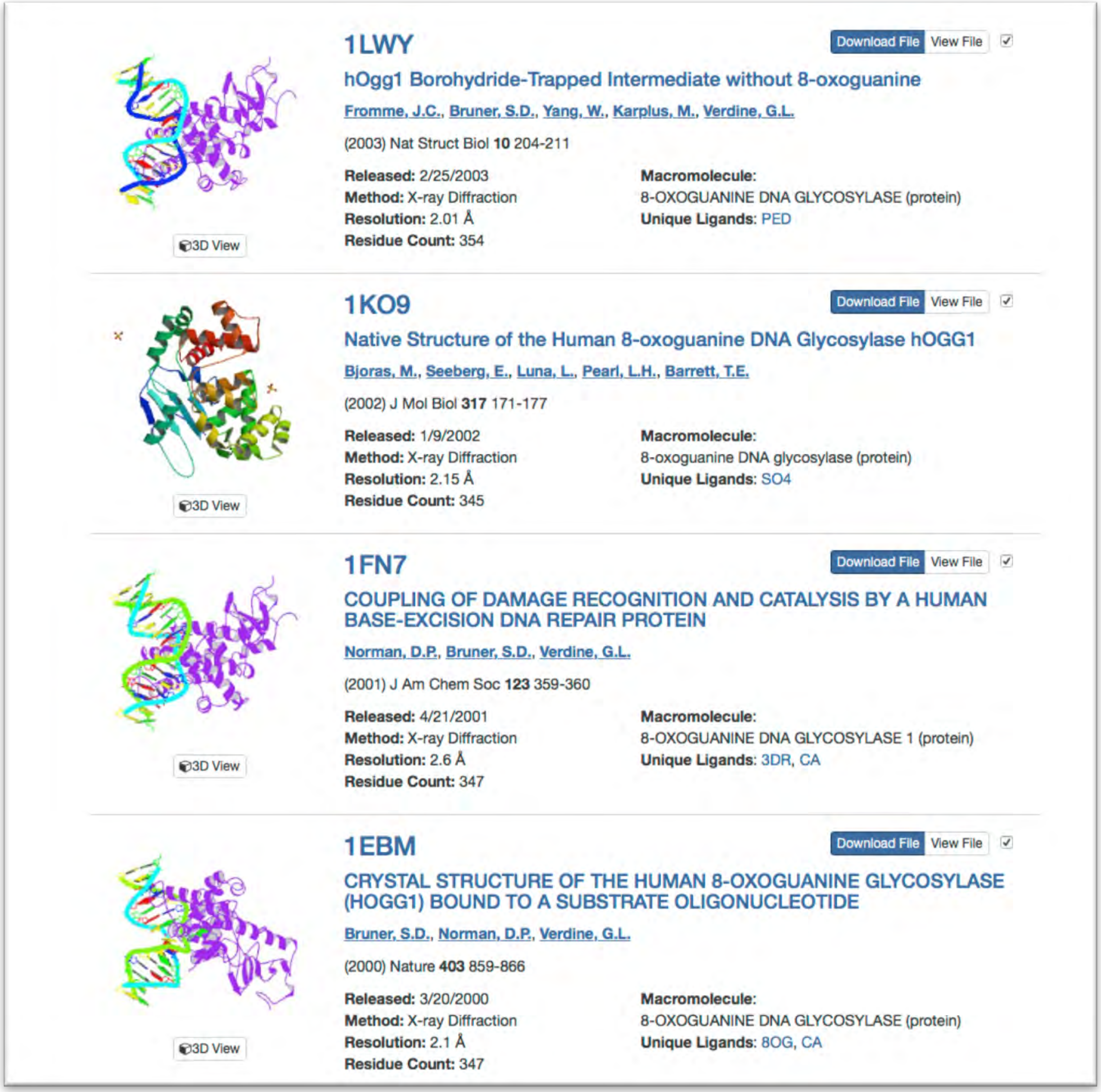
Some few 1000 unique protein folds

126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions as of April 2011

PDB identifiers are on the form 1LYZ, 2B6C, 1T06 (and does not "mean" anything)

Protein Structure Database

Search for "OGG1"



3D View

1LWY

Download File View File ☑

hOgg1 Borohydride-Trapped Intermediate without 8-oxoguanine

Fromme, J.C., Bruner, S.D., Yang, W., Karplus, M., Verdine, G.L.

(2003) Nat Struct Biol **10** 204-211

Released: 2/25/2003	**Macromolecule:**
Method: X-ray Diffraction	8-OXOGUANINE DNA GLYCOSYLASE (protein)
Resolution: 2.01 Å	**Unique Ligands:** PED
Residue Count: 354	

3D View

1KO9

Download File View File ☑

Native Structure of the Human 8-oxoguanine DNA Glycosylase hOGG1

Bjoras, M., Seeberg, E., Luna, L., Pearl, L.H., Barrett, T.E.

(2002) J Mol Biol **317** 171-177

Released: 1/9/2002	**Macromolecule:**
Method: X-ray Diffraction	8-oxoguanine DNA glycosylase (protein)
Resolution: 2.15 Å	**Unique Ligands:** SO4
Residue Count: 345	

3D View

1FN7

Download File View File ☑

COUPLING OF DAMAGE RECOGNITION AND CATALYSIS BY A HUMAN BASE-EXCISION DNA REPAIR PROTEIN

Norman, D.P., Bruner, S.D., Verdine, G.L.

(2001) J Am Chem Soc **123** 359-360

Released: 4/21/2001	**Macromolecule:**
Method: X-ray Diffraction	8-OXOGUANINE DNA GLYCOSYLASE 1 (protein)
Resolution: 2.6 Å	**Unique Ligands:** 3DR, CA
Residue Count: 347	

3D View

1EBM

Download File View File ☑

CRYSTAL STRUCTURE OF THE HUMAN 8-OXOGUANINE GLYCOSYLASE (HOGG1) BOUND TO A SUBSTRATE OLIGONUCLEOTIDE

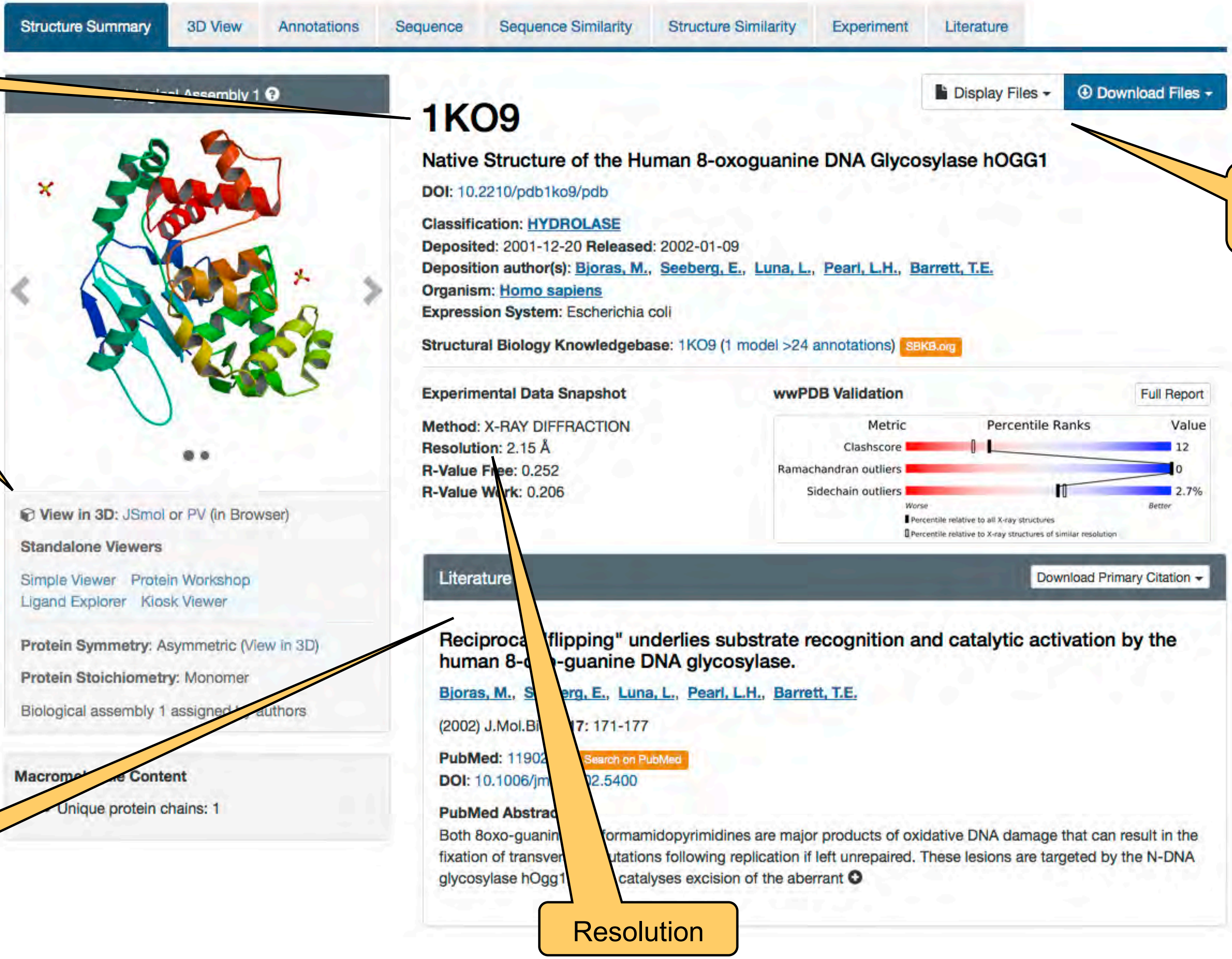
Bruner, S.D., Norman, D.P., Verdine, G.L.

(2000) Nature **403** 859-866

Released: 3/20/2000	**Macromolecule:**
Method: X-ray Diffraction	8-OXOGUANINE DNA GLYCOSYLASE (protein)
Resolution: 2.1 Å	**Unique Ligands:** 8OG, CA
Residue Count: 347	

Protein Structure Database

First hit for "OGG1"



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

ssembly 1

Display Files ▾ | Download Files ▾

1KO9

Native Structure of the Human 8-oxoguanine DNA Glycosylase hOGG1

DOI: 10.2210/pdb1ko9/pdb

Classification: HYDROLASE

Deposited: 2001-12-20 **Released:** 2002-01-09

Deposition author(s): Bjoras, M., Seeberg, E., Luna, L., Pearl, L.H., Barrett, T.E.

Organism: Homo sapiens

Expression System: Escherichia coli

Structural Biology Knowledgebase: 1KO9 (1 model >24 annotations) SBKB.org

View in 3D: JSmol or PV (in Browser)

Standalone Viewers

Simple Viewer Protein Workshop
Ligand Explorer Kiosk Viewer

Protein Symmetry: Asymmetric (View in 3D)

Protein Stoichiometry: Monomer

Biological assembly 1 assigned by authors

Macromo e Content

Unique protein chains: 1

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.15 Å
R-Value Free: 0.252
R-Value Work: 0.206

wwPDB Validation Full Report

Metric	Percentile Ranks	Value
Clashscore		12
Ramachandran outliers		0
Sidechain outliers		2.7%

Worse
Better
▮ Percentile relative to all X-ray structures
▯ Percentile relative to X-ray structures of similar resolution

Literature Download Primary Citation ▾

Reciproca "flipping" underlies substrate recognition and catalytic activation by the human 8-o -guanine DNA glycosylase.

Bjoras, M., S erg, E., Luna, L., Pearl, L.H., Barrett, T.E.

(2002) J.Mol.Bi 7: 171-177

PubMed: 11902 Search on PubMed
DOI: 10.1006/jm 02.5400

PubMed Abstrac

Both 8oxo-guanin formamidopyrimidines are major products of oxidative DNA damage that can result in the fixation of transver utations following replication if left unrepaired. These lesions are targeted by the N-DNA glycosylase hOgg1 catalyses excision of the aberrant

PDB id

PDB file (data file)

View structure in *e.g.* Jmol

Publication

Resolution

PDB entry – an example in PDB format

```
HEADER    LYASE/DNA                               24-JAN-00   1EBM
TITLE     CRYSTAL STRUCTURE OF THE HUMAN 8-OXOGUANINE GLYCOSYLASE
TITLE    2 (HOGG1) BOUND TO A SUBSTRATE OLIGONUCLEOTIDE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: 8-OXOGUANINE DNA GLYCOSYLASE;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: CORE FRAGMENT (RESIDUES 12 TO 325);
COMPND   5 SYNONYM: AP LYASE;
COMPND   6 ENGINEERED: YES;
COMPND   7 MUTATION: YES;
COMPND   8 MOL_ID: 2;
COMPND   9 MOLECULE: DNA (5'-D(*GP*CP*GP*TP*CP*CP*AP*(OXO)
COMPND  10 GP*GP*TP*CP*TP*AP*CP*C)-3');
COMPND  11 CHAIN: C;
COMPND  12 ENGINEERED: YES;
COMPND  13 MOL_ID: 3;
COMPND  14 MOLECULE: DNA (5'-
COMPND  15 D(*GP*GP*TP*AP*GP*AP*CP*CP*TP*GP*GP*AP*CP*GP*C)-3');
COMPND  16 CHAIN: D;
COMPND  17 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   4 EXPRESSION_SYSTEM_COMMON: BACTERIA;
SOURCE   5 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE   6 EXPRESSION_SYSTEM_PLASMID: PET30A-HOGG1;
SOURCE   7 MOL_ID: 2;
SOURCE   8 SYNTHETIC: YES;
SOURCE   9 MOL_ID: 3;
SOURCE  10 SYNTHETIC: YES
KEYWDS    DNA REPAIR, DNA GLYCOSYLASE, PROTEIN/DNA
EXPDTA    X-RAY DIFFRACTION
AUTHOR    S.D.BRUNER,D.P.NORMAN,G.L.VERDINE
REVDAT   1   20-MAR-00 1EBM    0
JRNL        AUTH   S.D.BRUNER,D.P.NORMAN,G.L.VERDINE
JRNL        TITL   STRUCTURAL BASIS FOR RECOGNITION AND REPAIR OF THE
JRNL        TITL 2 ENDOGENOUS MUTAGEN 8-OXOGUANINE IN DNA
JRNL        REF    NATURE                          V. 403   859 2000
JRNL        REFN   ASTM NATUAS  UK ISSN 0028-0836
REMARK   1
REMARK   2 RESOLUTION. 2.10 ANGSTROMS.
REMARK   3
......
```

- Standard since early 1970s
- FORTRAN compatible format
- Some limitations
  - Number of atoms
  - Number of chains
  - Length of fields
- Not good for parsing by computers

PDB entry – an example in PDB format

Jon K. Lærdahl,
Structural Bioinformatics

Atom name

Residue name

Chain

B-factor

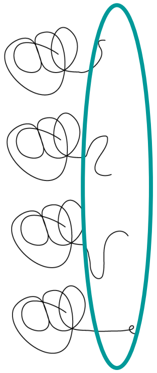
Amino acid field

```
....
ATOM      1  N   GLY A   9      29.382 -12.935  38.434  1.00 39.96           N
ATOM      2  CA  GLY A   9      28.983 -13.096  36.994  1.00 40.83           C
ATOM      3  C   GLY A   9      27.548 -12.643  36.792  1.00 41.51           C
ATOM      4  O   GLY A   9      27.265 -11.724  36.007  1.00 41.29           O
ATOM      5  N   SER A  10      26.631 -13.287  37.505  1.00 41.40           N
ATOM      6  CA  SER A  10      25.222 -12.936  37.418  1.00 41.42           C
ATOM      7  C   SER A  10      24.900 -11.903  38.494  1.00 39.54           C
ATOM      8  O   SER A  10      23.732 -11.620  38.763  1.00 40.12           O
ATOM      9  CB  SER A  10      24.357 -14.176  37.639  1.00 43.12           C
ATOM     10  OG  SER A  10      24.599 -14.728  38.920  1.00 43.93           O
ATOM     11  N   GLU A  11      25.940 -11.343  39.102  1.00 37.35           N
ATOM     12  CA  GLU A  11      25.764 -10.360  40.166  1.00 36.30           C
ATOM     13  C   GLU A  11      26.373  -9.013  39.755  1.00 34.00           C
ATOM     14  O   GLU A  11      27.302  -8.968  38.951  1.00 32.56           O
ATOM     15  CB  GLU A  11      26.451 -10.849  41.454  1.00 38.36           C
ATOM     16  CG  GLU A  11      26.387 -12.365  41.740  1.00 39.94           C
ATOM     17  CD  GLU A  11      25.069 -12.823  42.343  1.00 41.33           C
ATOM     18  OE1 GLU A  11      24.963 -14.021  42.693  1.00 40.98           O
ATOM     19  OE2 GLU A  11      24.139 -11.999  42.468  1.00 41.16           O
ATOM     20  N   GLY A  12      25.853  -7.925  40.320  1.00 31.94           N
ATOM     21  CA  GLY A  12      26.368  -6.602  40.009  1.00 30.07           C
ATOM     22  C   GLY A  12      25.925  -6.027  38.674  1.00 29.09           C
ATOM     23  O   GLY A  12      25.174  -6.652  37.919  1.00 28.15           O
ATOM     24  N   HIS A  13      26.392  -4.820  38.379  1.00 29.23           N
ATOM     25  CA  HIS A  13      26.043  -4.159  37.124  1.00 29.36           C
ATOM     26  C   HIS A  13      26.651  -4.913  35.941  1.00 30.04           C
ATOM     27  O   HIS A  13      27.838  -5.247  35.948  1.00 30.64           O
ATOM     28  CB  HIS A  13      26.545  -2.716  37.121  1.00 28.62           C
ATOM     29  CG  HIS A  13      25.874  -1.831  38.127  1.00 27.87           C
ATOM     30  ND1 HIS A  13      26.285  -1.746  39.441  1.00 26.37           N
....
HETATM 3056  O   HOH     5      23.168  15.174  34.624  1.00 18.07           O
HETATM 3057  O   HOH     6      21.609  14.592  31.635  1.00 13.68           O
HETATM 3058  O   HOH     7      14.739  30.965  30.601  1.00 26.62           O
HETATM 3059  O   HOH     9      29.320   3.836  25.672  1.00 27.62           O
....
```

Cofactor field

Atom coordinates

The B-factor (temperature factor) is an indicator of thermal motion. Actually a mixture of real thermal motion and structural disorder (multiple conformations)



PDB entry – an example in mmCIF format

```
data_1EBM
#
_entry.id   1EBM
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    1.044
_audit_conform.dict_location   http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.
_database_2.database_code
PDB  1EBM
NDB  PD0117
RCSB RCSB010437
#
_database_PDB_rev.num             1
_database_PDB_rev.date            2000-03-20
_database_PDB_rev.date_original   2000-01-24
_database_PDB_rev.status          ?
_database_PDB_rev.replaces        1EBM
_database_PDB_rev.mod_type        0
#
_pdbx_database_status.status_code    REL
_pdbx_database_status.entry_id       1EBM
_pdbx_database_status.deposit_site   RCSB
_pdbx_database_status.process_site   RCSB
_pdbx_database_status.SG_entry       .
#
loop_
_audit_author.name
'Bruner, S.D.'
'Norman, D.P.'
'Verdine, G.L.'
#
_citation.id                        primary
_citation.title                     'Structural basis for recognition
_citation.journal_abbrev            Nature
_citation.journal_volume            403
_citation.page_first                859
_citation.page_last                 866
```

Newer data format and alternative to “PDB format”

- No limitations in number of atoms, chains, fields etc.
- Better suited for automatic parsing/processing

Structural bioinformatics

Experimental structure is hard to get

The 3D structure on a protein is determined by the amino acid sequence (primary structure)

There are many orders of magnitude more sequences available than there are structures

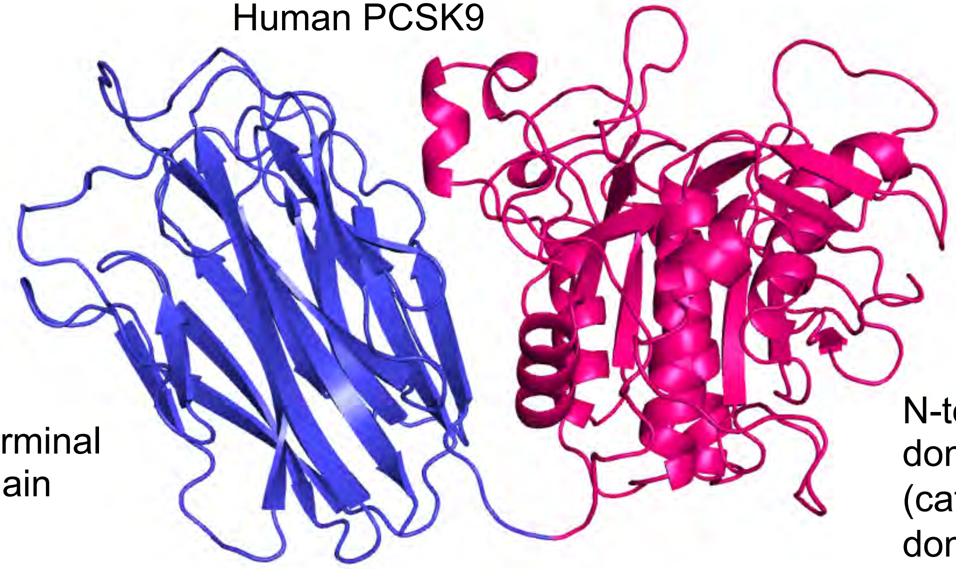
How do we get information about structure from sequence?

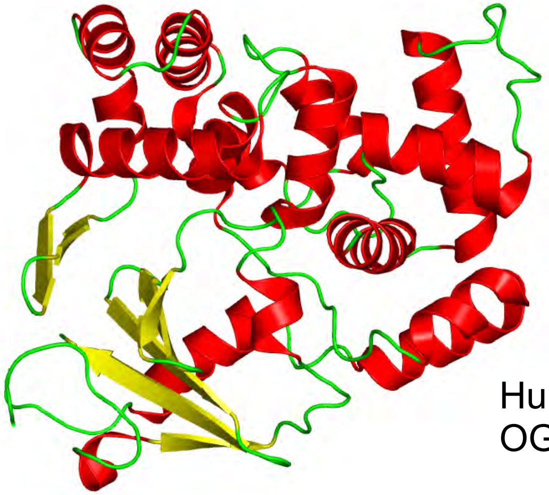
Protein domains

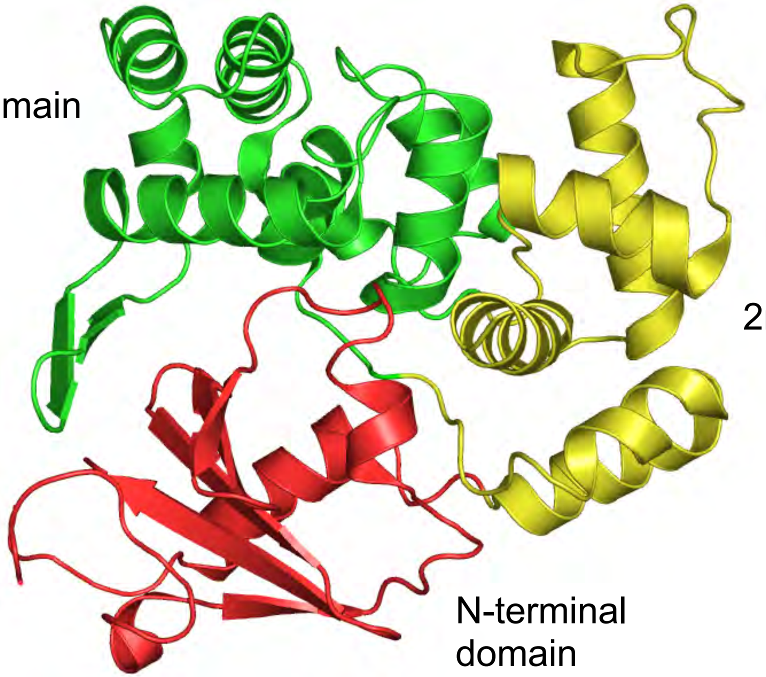
Domain: Compact part of a protein that represents a structurally independent region

Domains are often separate functional units that may be studied separately

Domains fold independently? Not always…

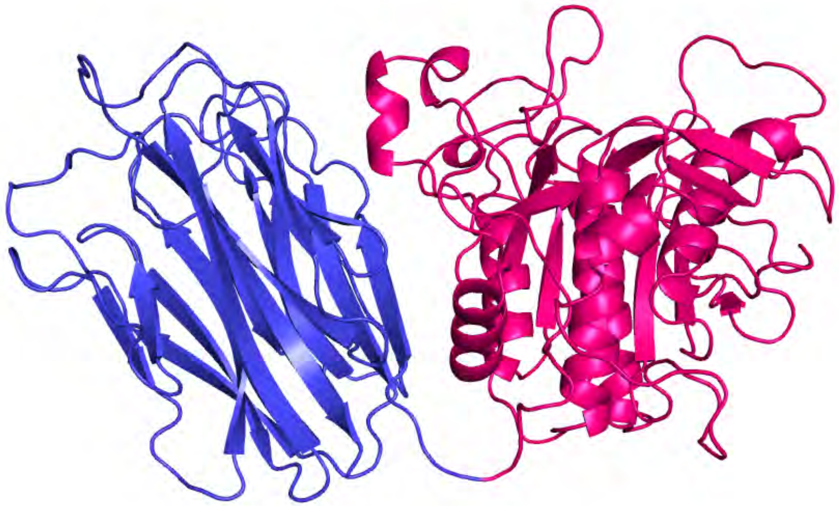




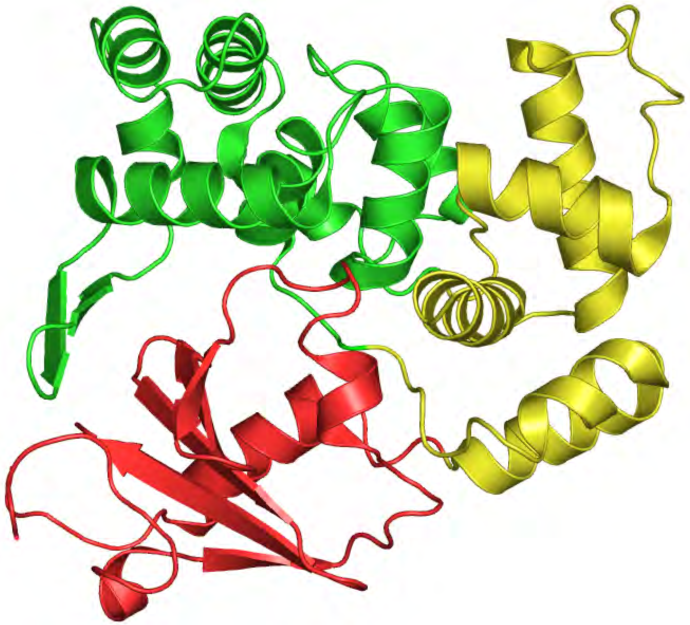


Protein domains

Dividing a protein structure into domains: no "right way to do it" or "correct algorithm", *i.e.* ***a lot of subjectivity involved***



Most people would agree there are two domains here



Three domains?
One domain?
Two?

SCOP vs.
CATH?

Very often we model, compare, classify *domains* – not full-length proteins

Protein domains

Instead of working with full length proteins that may be

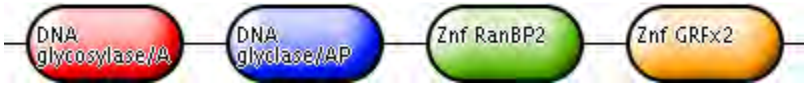
- very large
- contain one or many separate modules (*i.e.* domains)
- have both structured and unstructured parts

We often instead work with protein domains that are

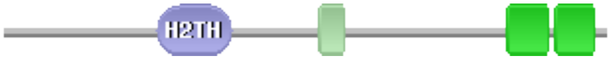
- more compact
- can be studied separately
  - function
  - structure by X-ray crystallography/NMR
  - bioinformatics modeling
- may be viewed as the "spare parts" building up full-length proteins

Many proteins are structured domains, "spare parts", connected by short loops or long disordered regions

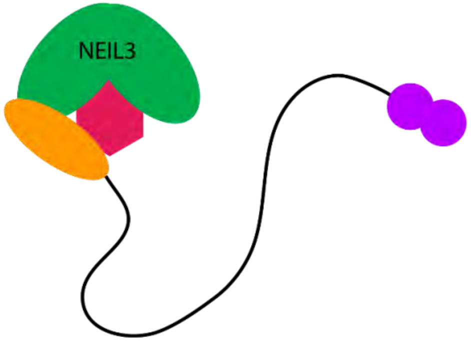
Far from trivial to detect boundaries between domains from sequence only:



InterPro



Pfam



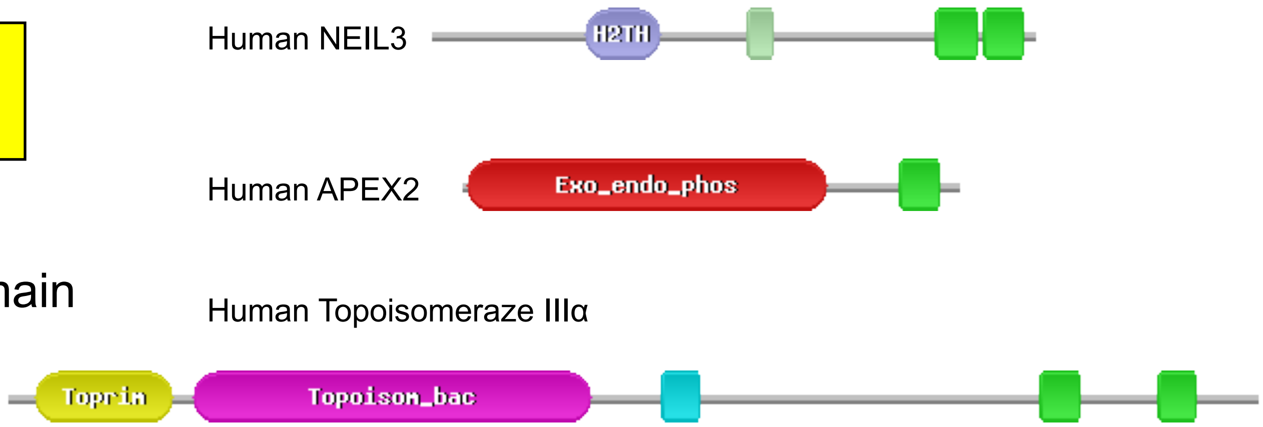
Protein domains

Domains have a “signature sequence” that can be described as a HMM Logo

Important to think in terms of domains!!

🟩 GRF zinc finger domain

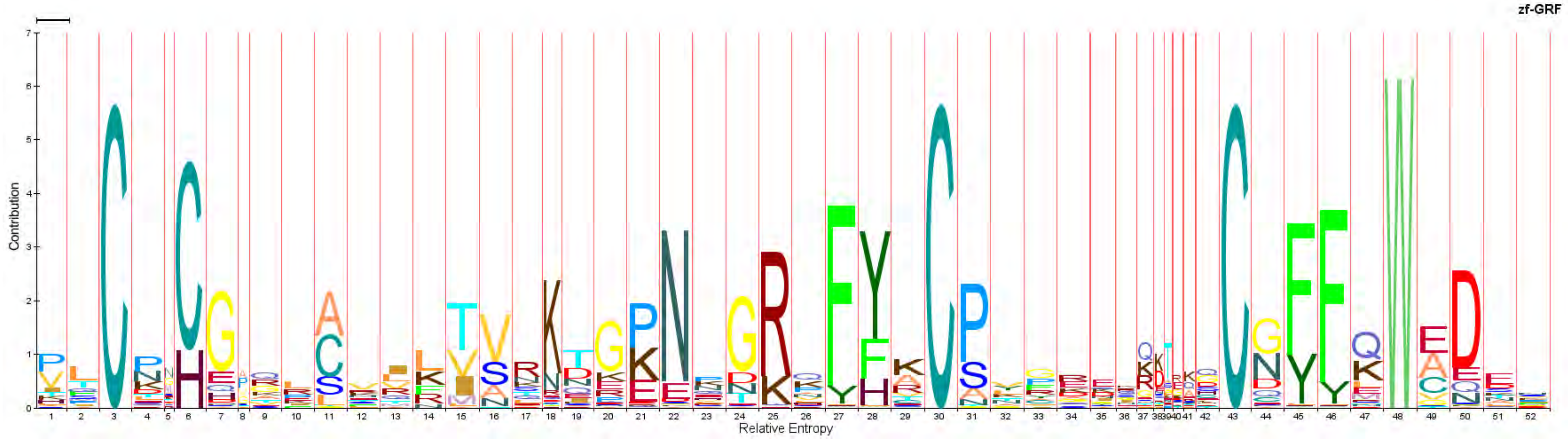
Domains can be “switched”. They can be viewed as “spare parts” that can be used to build new proteins through evolution



Human NEIL3

Human APEX2

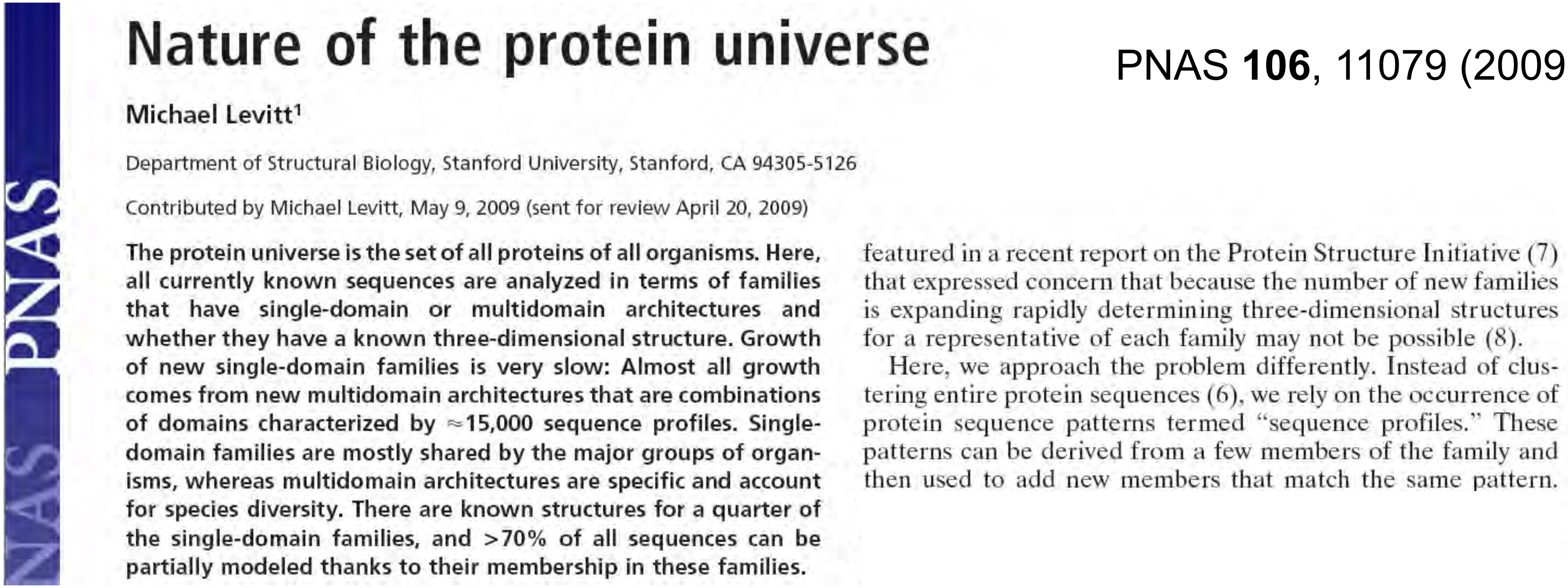
Human Topoisomeraze IIIα



Pfam HMM-logo for the GRF zinc finger domain

Protein domains

PNAS **106**, 11079 (2009)



Nature of the protein universe

Michael Levitt[1]

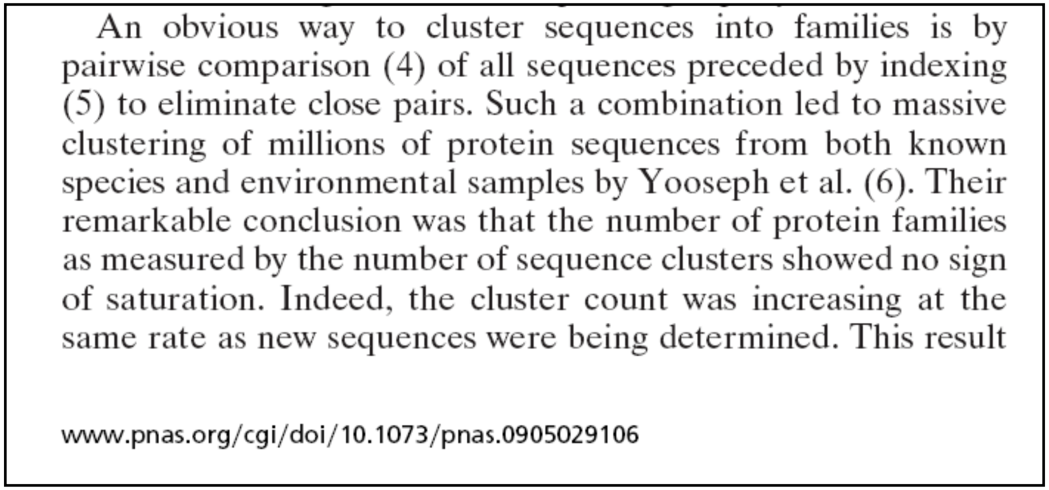
Department of Structural Biology, Stanford University, Stanford, CA 94305-5126

Contributed by Michael Levitt, May 9, 2009 (sent for review April 20, 2009)

The protein universe is the set of all proteins of all organisms. Here, all currently known sequences are analyzed in terms of families that have single-domain or multidomain architectures and whether they have a known three-dimensional structure. Growth of new single-domain families is very slow: Almost all growth comes from new multidomain architectures that are combinations of domains characterized by ≈15,000 sequence profiles. Single-domain families are mostly shared by the major groups of organisms, whereas multidomain architectures are specific and account for species diversity. There are known structures for a quarter of the single-domain families, and >70% of all sequences can be partially modeled thanks to their membership in these families.

featured in a recent report on the Protein Structure Initiative (7) that expressed concern that because the number of new families is expanding rapidly determining three-dimensional structures for a representative of each family may not be possible (8).

Here, we approach the problem differently. Instead of clustering entire protein sequences (6), we rely on the occurrence of protein sequence patterns termed "sequence profiles." These patterns can be derived from a few members of the family and then used to add new members that match the same pattern.



An obvious way to cluster sequences into families is by pairwise comparison (4) of all sequences preceded by indexing (5) to eliminate close pairs. Such a combination led to massive clustering of millions of protein sequences from both known species and environmental samples by Yooseph et al. (6). Their remarkable conclusion was that the number of protein families as measured by the number of sequence clusters showed no sign of saturation. Indeed, the cluster count was increasing at the same rate as new sequences were being determined. This result

www.pnas.org/cgi/doi/10.1073/pnas.0905029106

(6) Yooseph D, *et al.* (2007) The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. PLoS Biol **5**:e16.

Protein domains

Jon K. Lærdahl,
Structural Bioinformatics

PNAS **106**, 11079 (2009)

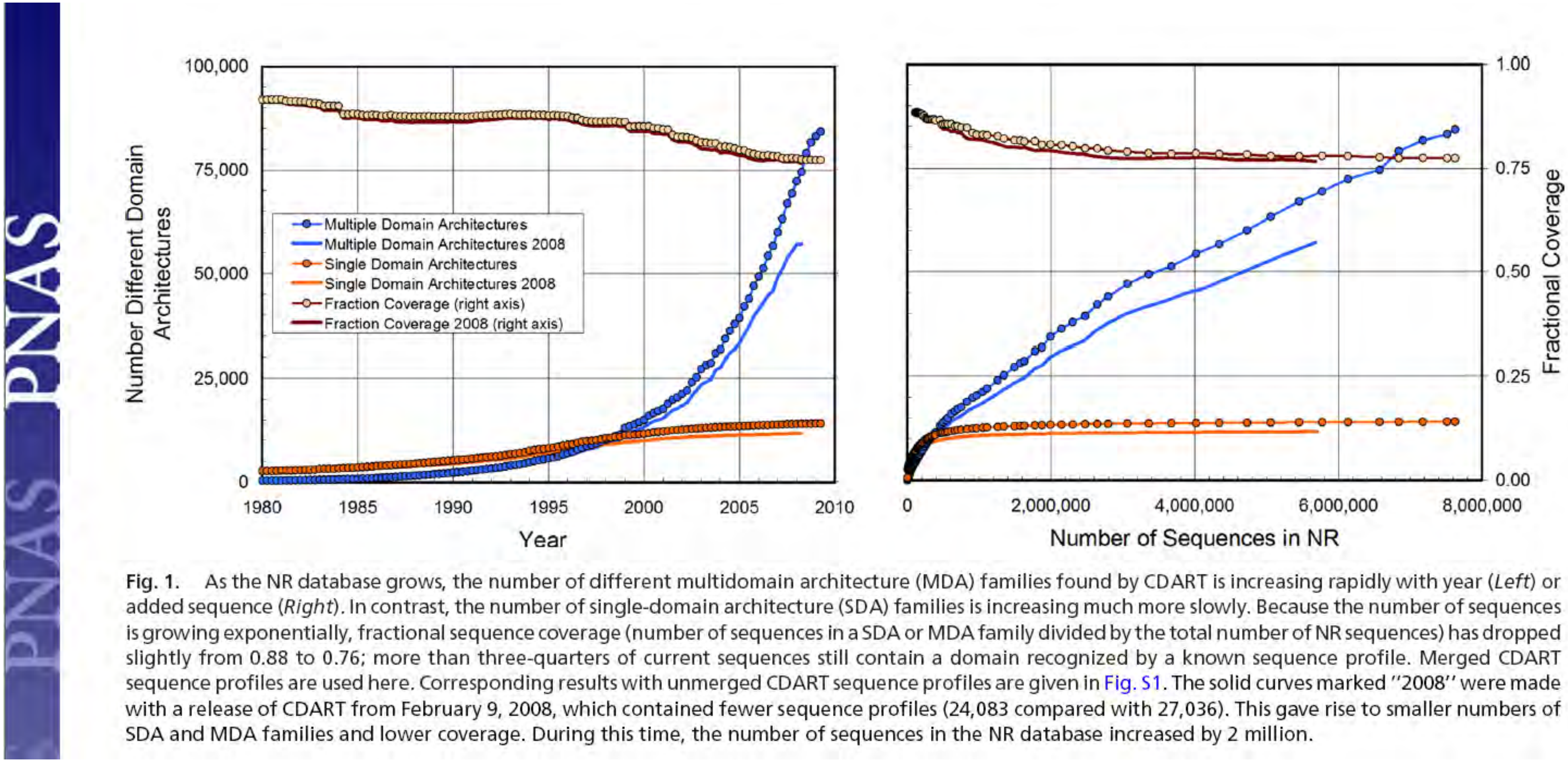


Fig. 1. As the NR database grows, the number of different multidomain architecture (MDA) families found by CDART is increasing rapidly with year (*Left*) or added sequence (*Right*). In contrast, the number of single-domain architecture (SDA) families is increasing much more slowly. Because the number of sequences is growing exponentially, fractional sequence coverage (number of sequences in a SDA or MDA family divided by the total number of NR sequences) has dropped slightly from 0.88 to 0.76; more than three-quarters of current sequences still contain a domain recognized by a known sequence profile. Merged CDART sequence profiles are used here. Corresponding results with unmerged CDART sequence profiles are given in Fig. S1. The solid curves marked ''2008'' were made with a release of CDART from February 9, 2008, which contained fewer sequence profiles (24,083 compared with 27,036). This gave rise to smaller numbers of SDA and MDA families and lower coverage. During this time, the number of sequences in the NR database increased by 2 million.

There are known structures for a quarter of
the single-domain families, and >70% of all sequences can be
partially modeled thanks to their membership in these families.

End