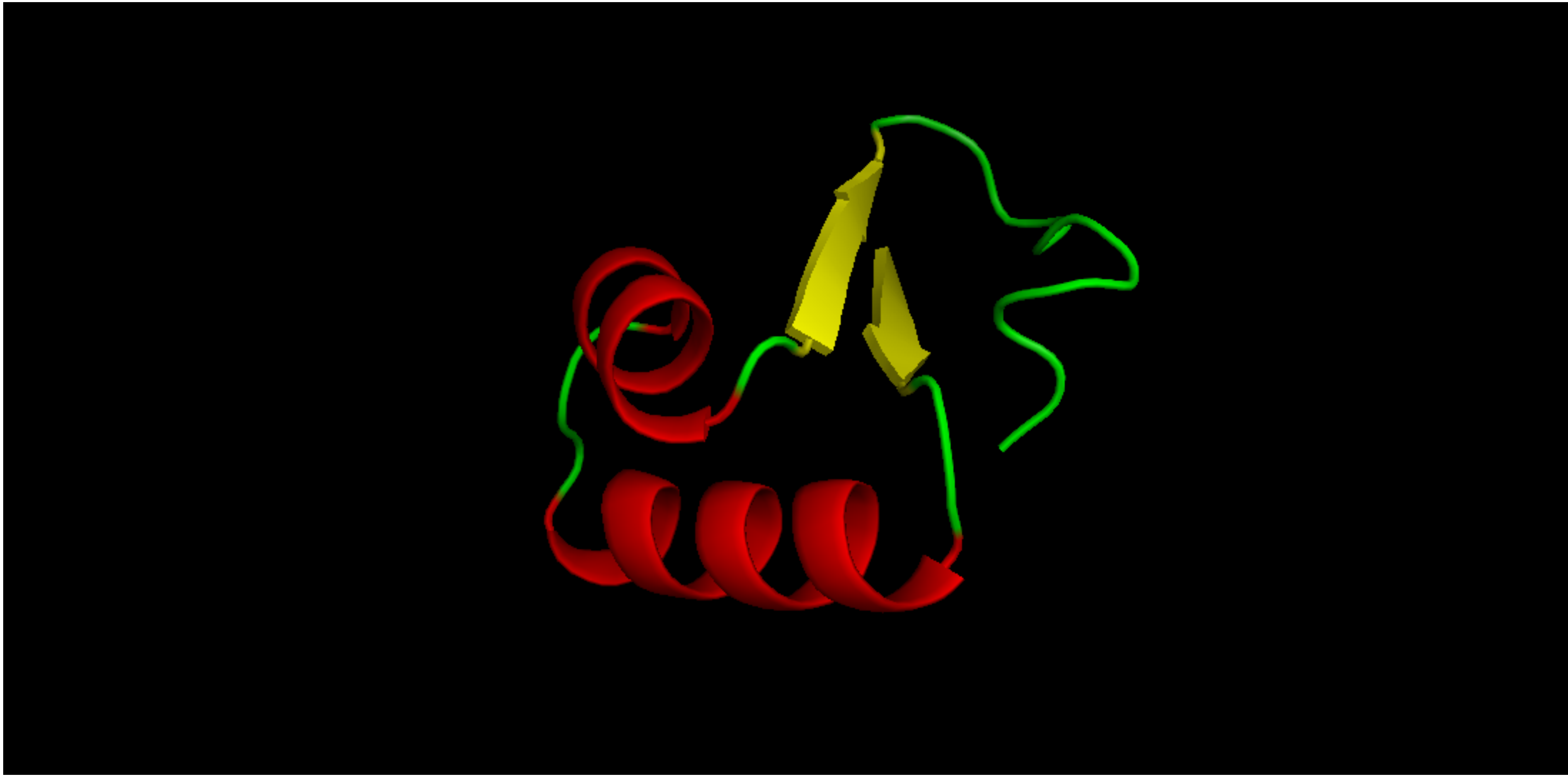
Structural biology

From sequence to structure

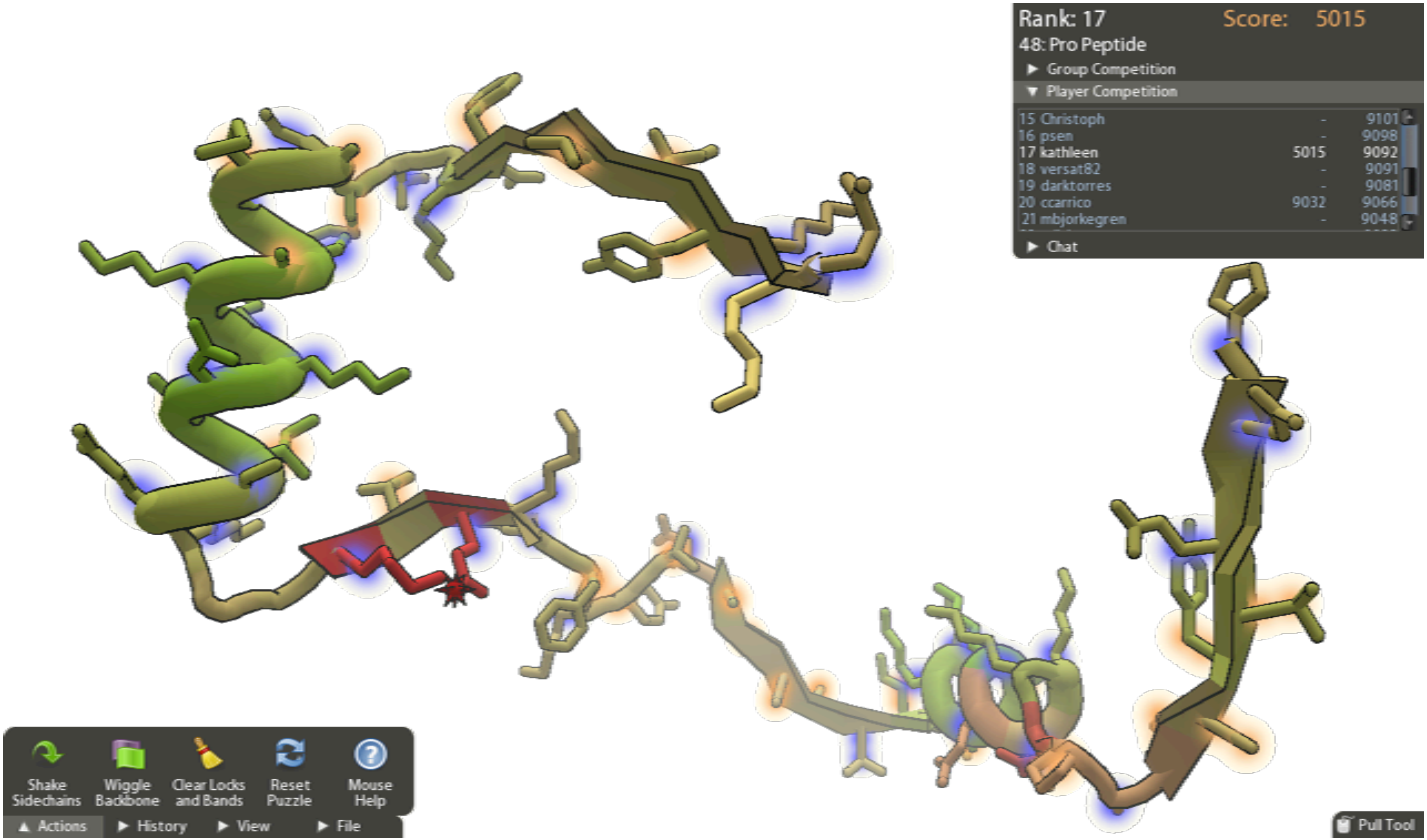
Proteins form into distinct shapes



Aims of structural biology

- Predict the confirmation of a poly-peptide chain

- Predict and analyze the function of a protein



Rank: 17 Score: 5015

48: Pro Peptide

▶ Group Competition

▼ Player Competition

15 Christoph	-	9101
16 psen	-	9098
17 kathleen	5015	9092
18 versat82	-	9091
19 darktorres	-	9081
20 ccarrico	9032	9066
21 mbjorkegren	-	9048

▶ Chat

Protein structures

- Elements of structure
  - The amino acids
  - Levels
  - Databases
  - Folds and families

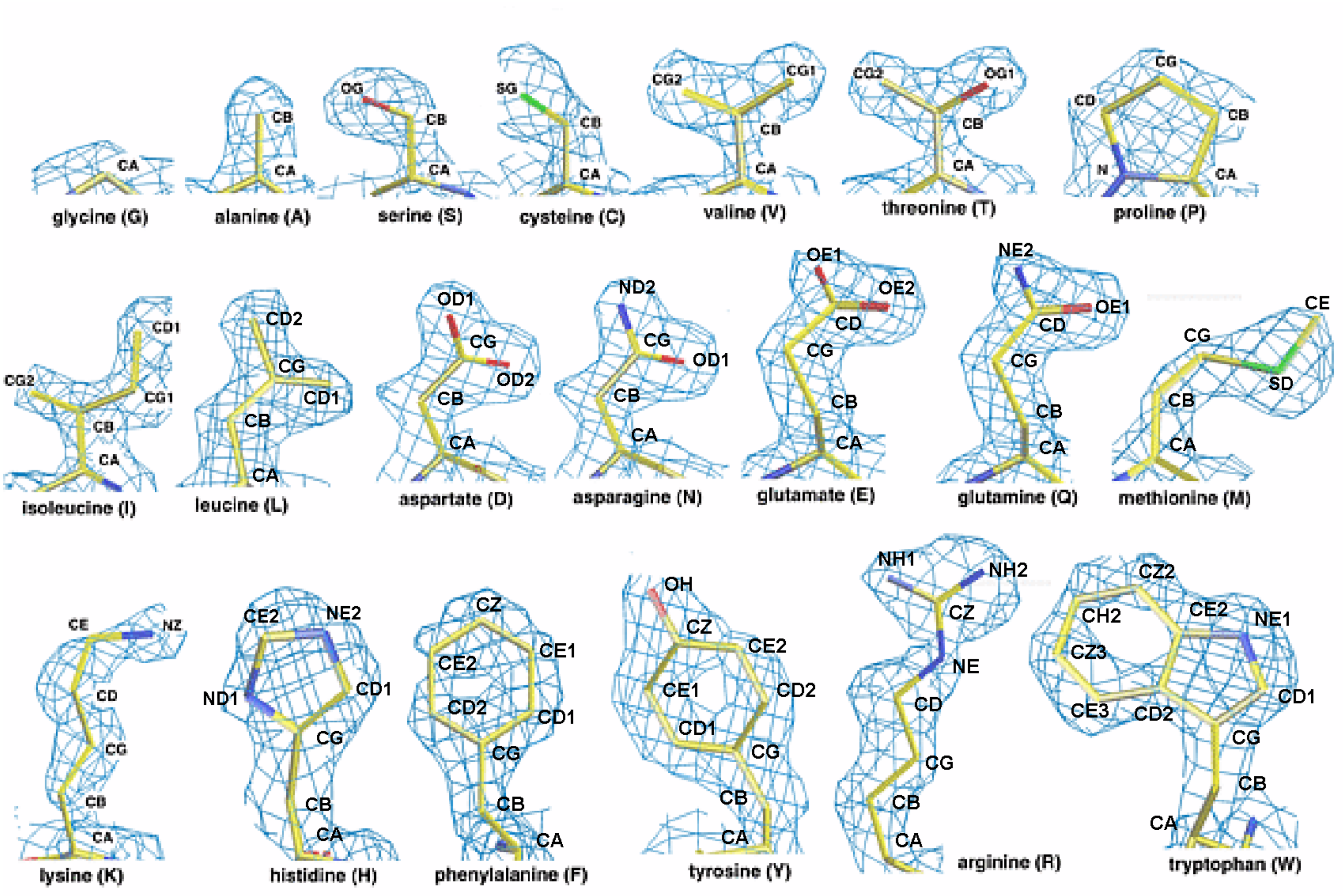
- Prediction

- How to check for correct assignment
  - Ramachandran plot

Structures of biomolecules

- Primary structure
  - Amino acid sequence
- Secondary structure
  - Local elements
    - Helices
    - Sheets
- Tertiary structure (3D)
  - Fold
  - Classification
- Quarternary structure
  - Interactions between chains
  - Protein-protein interaction

AMINO ACIDS

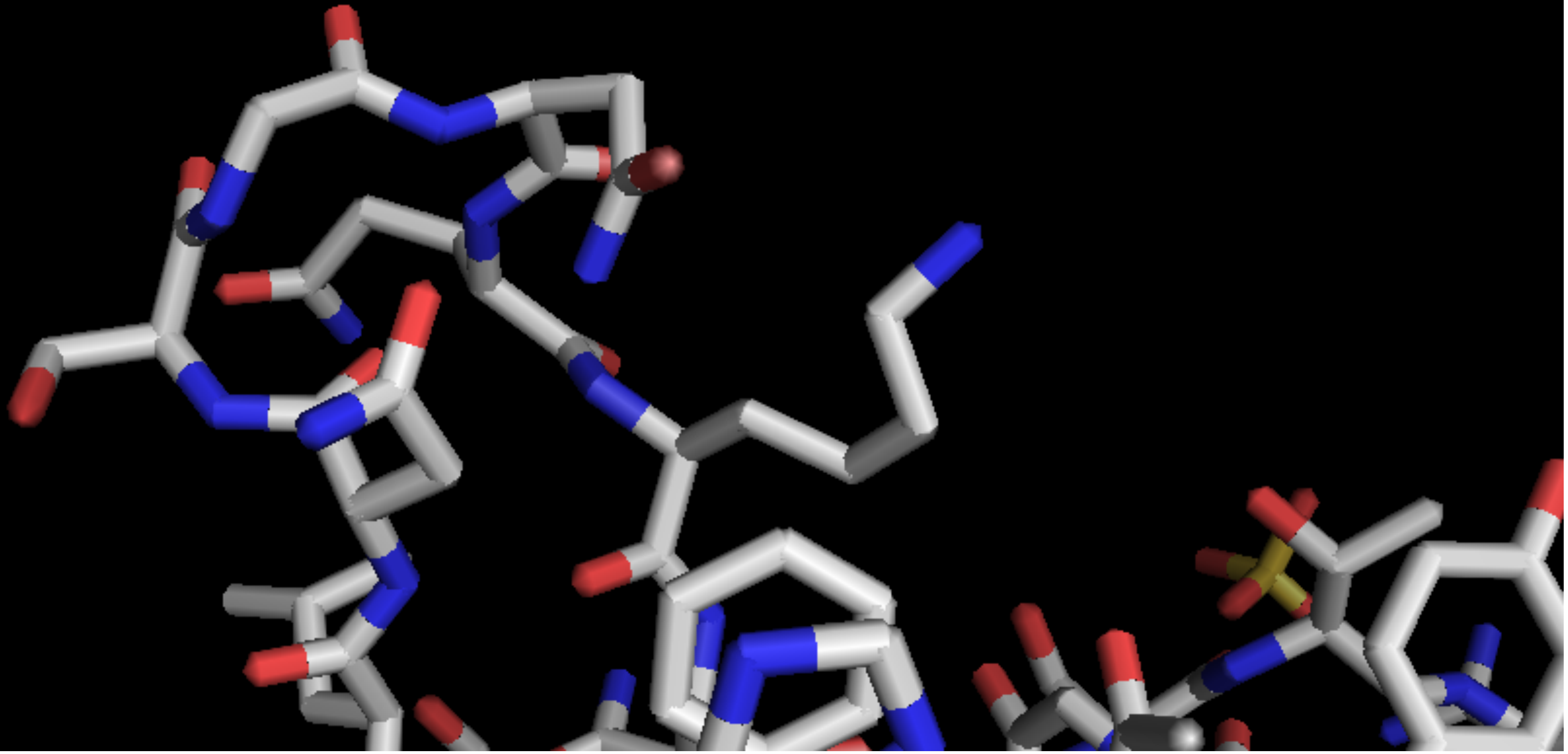

CA
glycine (G)
CB
CA
alanine (A)
OG
CB
CA
serine (S)
SG
CB
CA
cysteine (C)
CG2
CG1
CB
CA
valine (V)
CG2
OG1
CB
CA
threonine (T)
CG
CD
CB
N
CA
proline (P)
CD1
CG2
CG1
CB
CA
isoleucine (I)
CD2
CG
CD1
CB
CA
leucine (L)
OD1
CG
OD2
CB
CA
aspartate (D)
ND2
CG
OD1
CB
CA
asparagine (N)
OE1
OE2
CD
CG
CB
CA
glutamate (E)
NE2
OE1
CD
CG
CB
CA
glutamine (Q)
CE
CG
SD
CB
CA
methionine (M)
CE
NZ
CD
CG
CB
CA
lysine (K)
CE2
NE2
ND1
CD1
CG
CB
CA
histidine (H)
CZ
CE2
CE1
CD2
CD1
CG
CB
CA
phenylalanine (F)
OH
CZ
CE2
CE1
CD2
CD1
CG
CB
CA
tyrosine (Y)
NH1
NH2
CZ
NE
CD
CG
CB
CA
arginine (R)
CZ2
CH2
CE2
NE1
CZ3
CE3
CD2
CD1
CG
CB
CA
tryptophan (W)

Visualizing Proteins

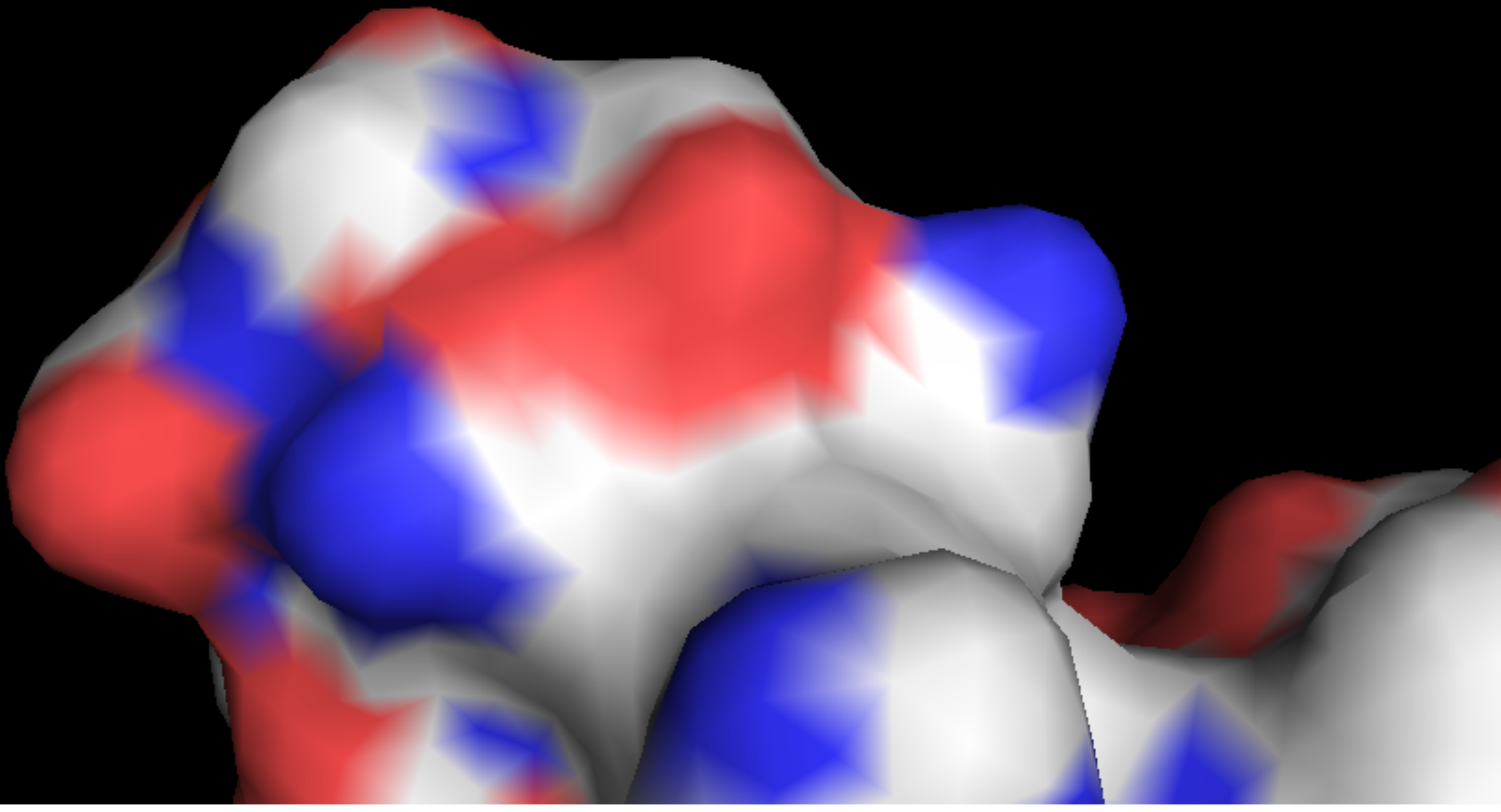
- High complexity
- Multiple levels of structure
- Important properties are “distributed
- throughout the 3D structure

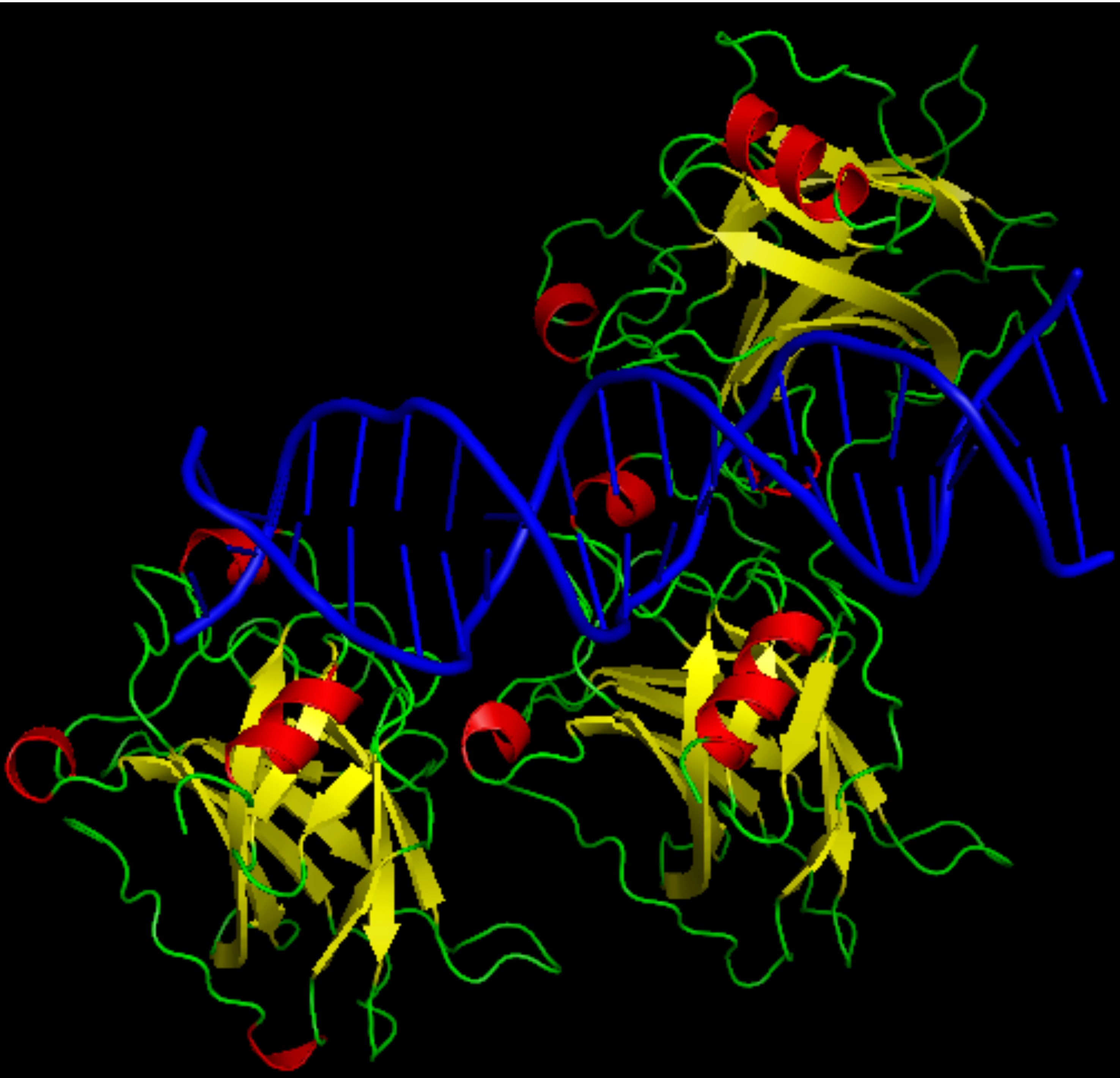
- No single/simple “point” at which to look

Wireframe

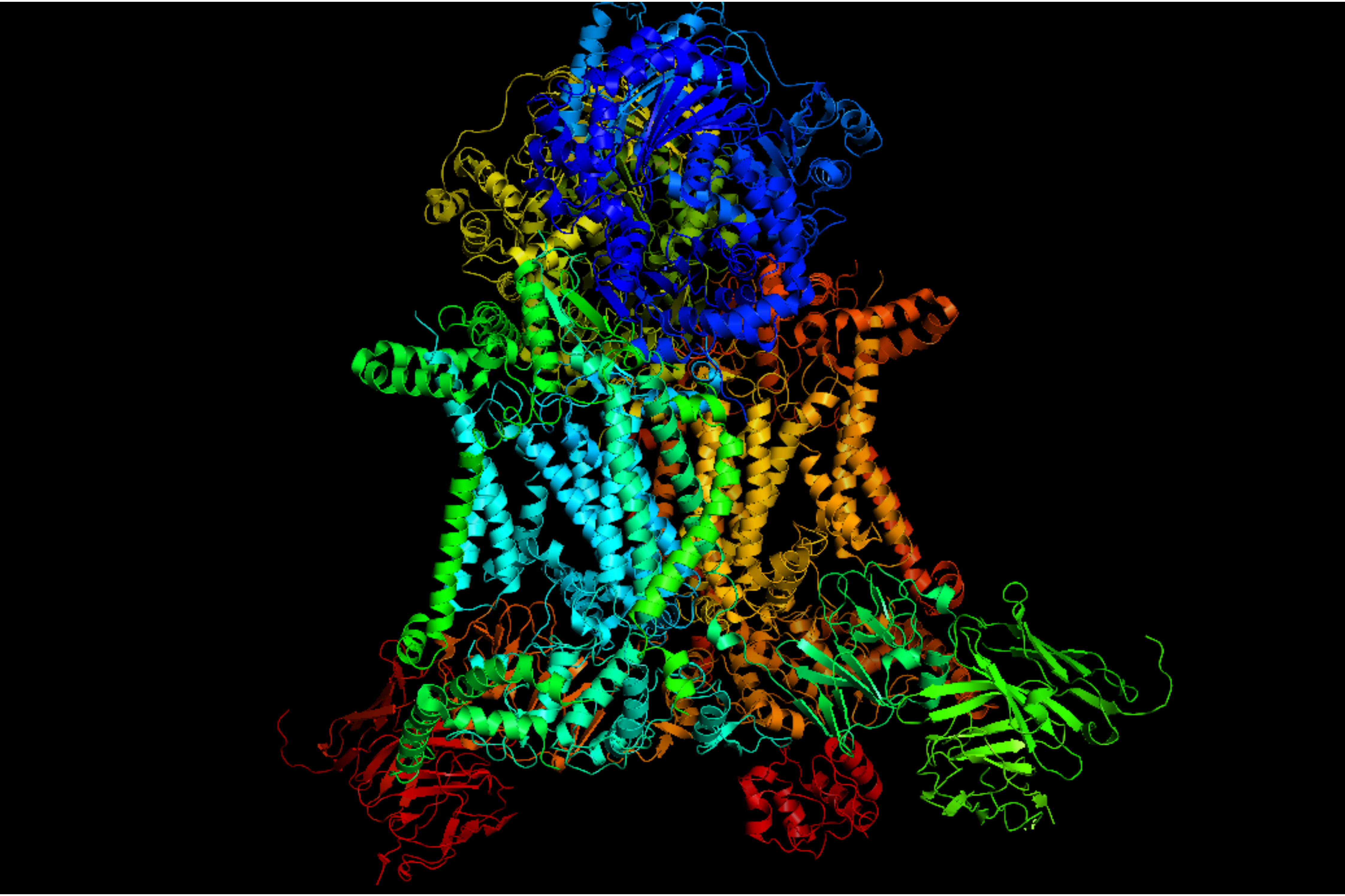


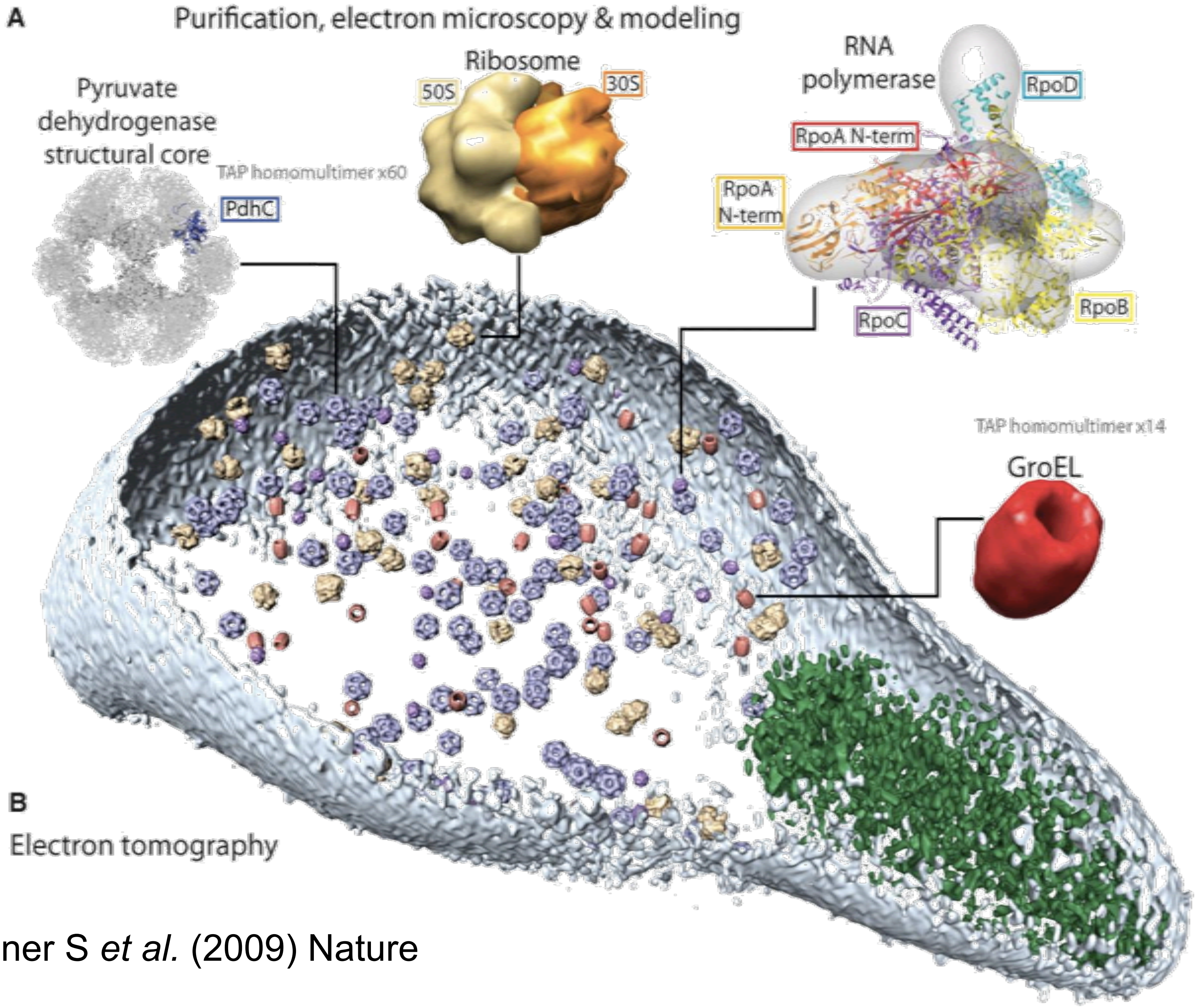
Surface





Cartoon





Kühner S *et al.* (2009) Nature

PDB/RCSB database

- Protein Data Bank – One of the oldest databases on molecular biology
- Repository of all known structures
  - All published structures must be deposited
- Four-character identifier

Classification of protein structures

CATH database

- Fold
- Superfamily – Secondary structure contacts
- Sequence families
- **Domains**

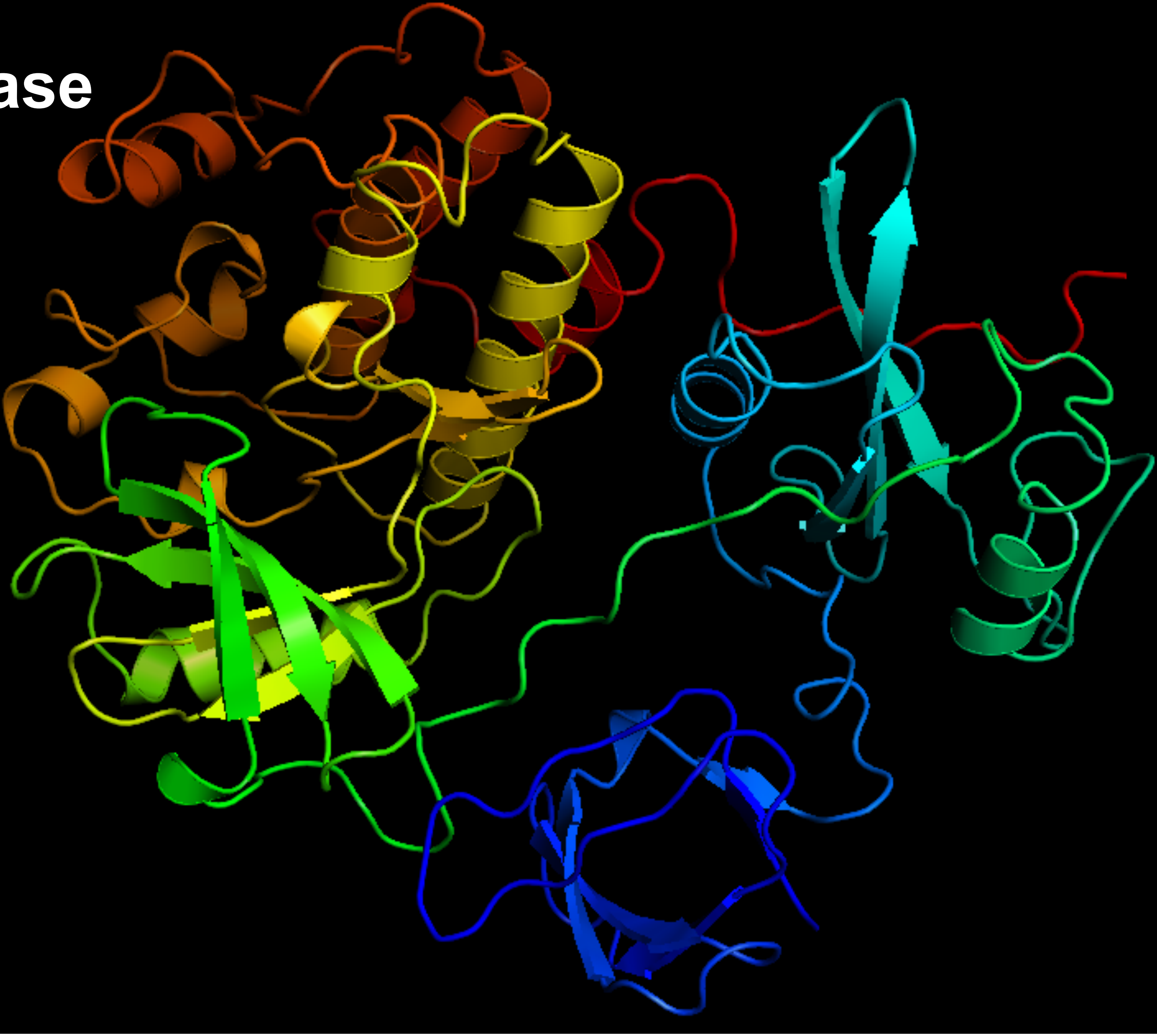
- Rule based on secondary structure content, contacts and domain boundaries

SCOP database

- Class
  - All α, all β
  - α / β – Parallel sheets
  - α + β – Antiparallel sheets
  - **Multi-domain proteins**
  - Membrane
  - Unstructured proteins

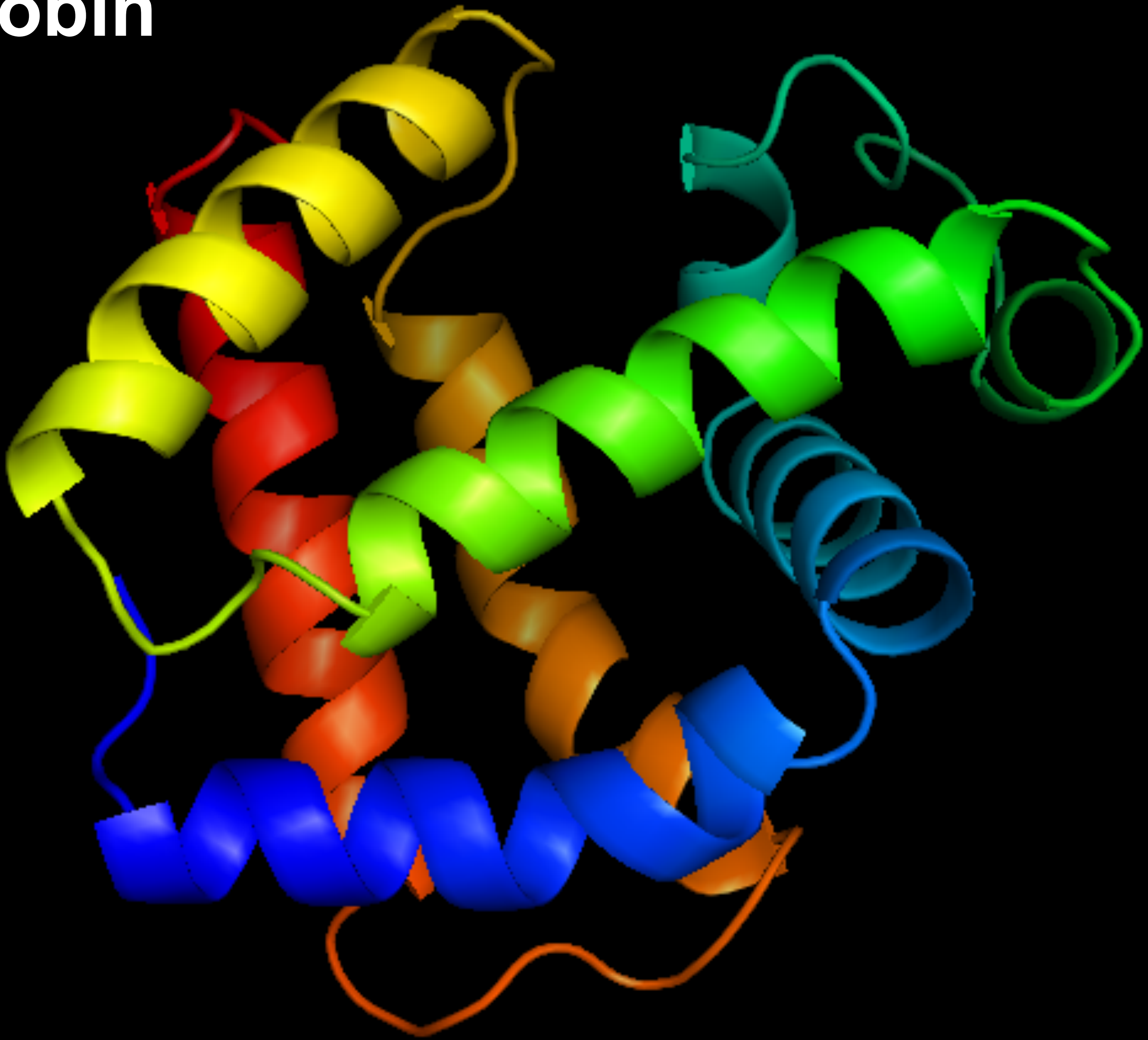
- Folds
- Superfamilies
- Families

SRC kinase



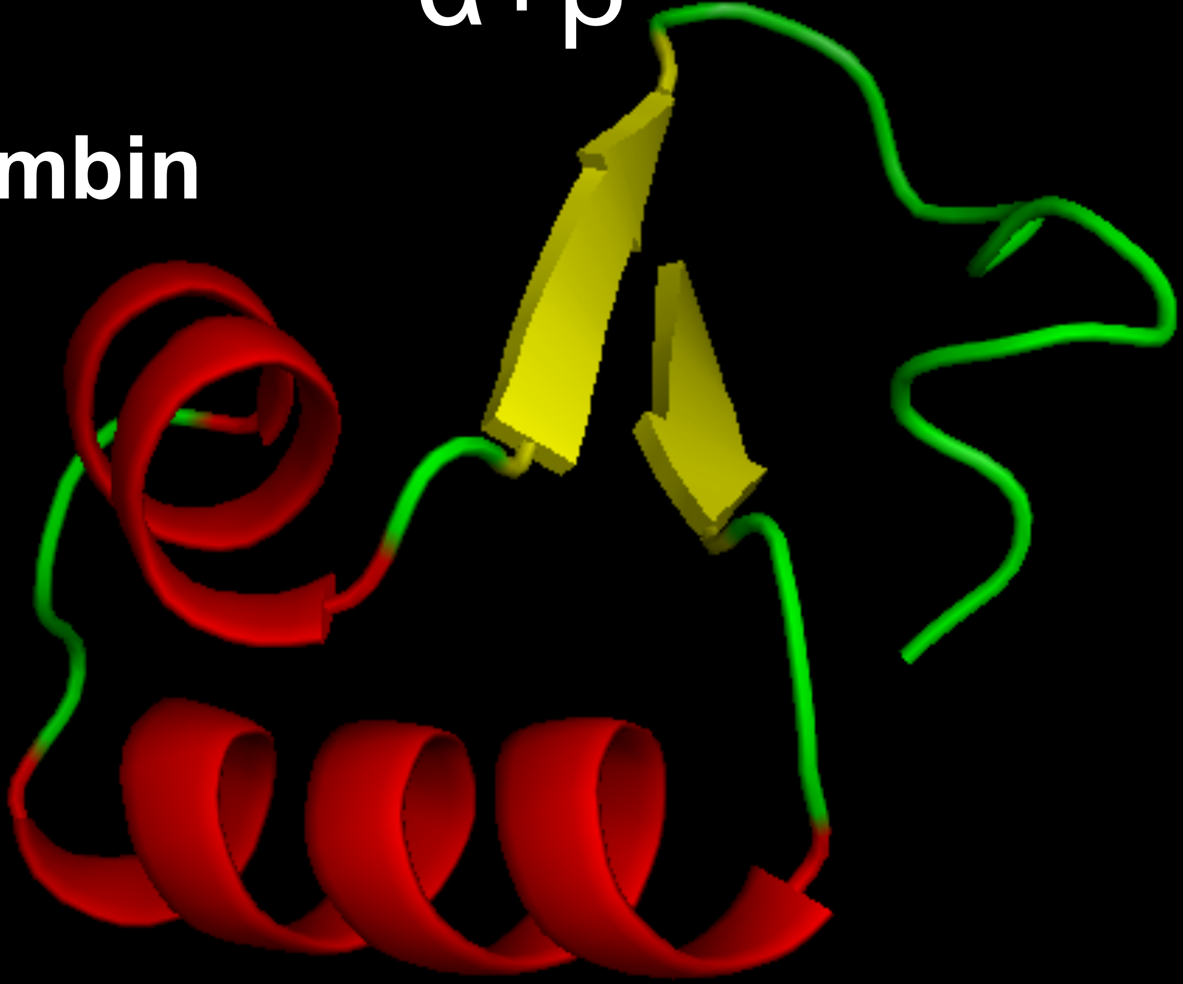
FOLDS

Hemoglobin

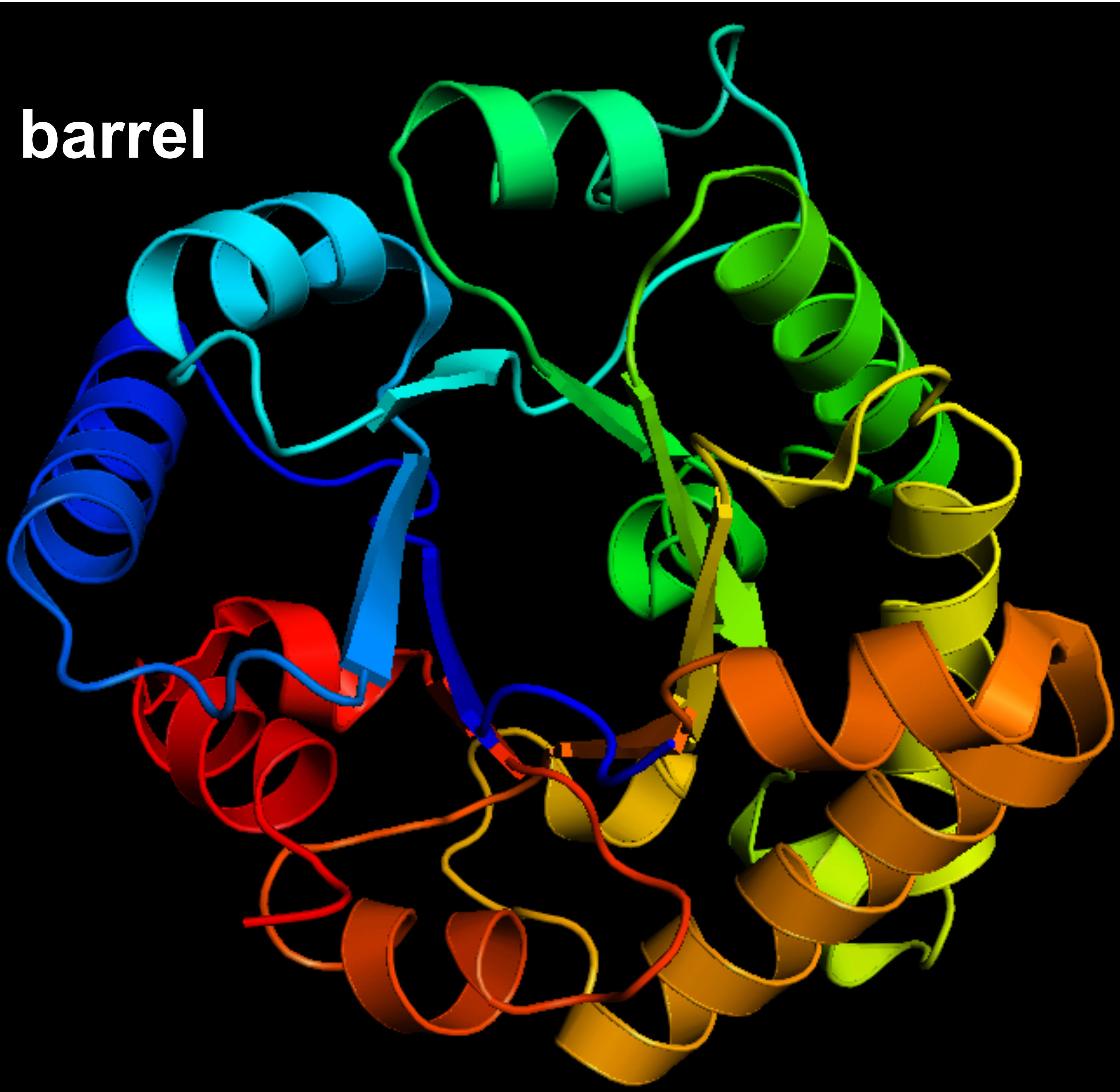


α+β

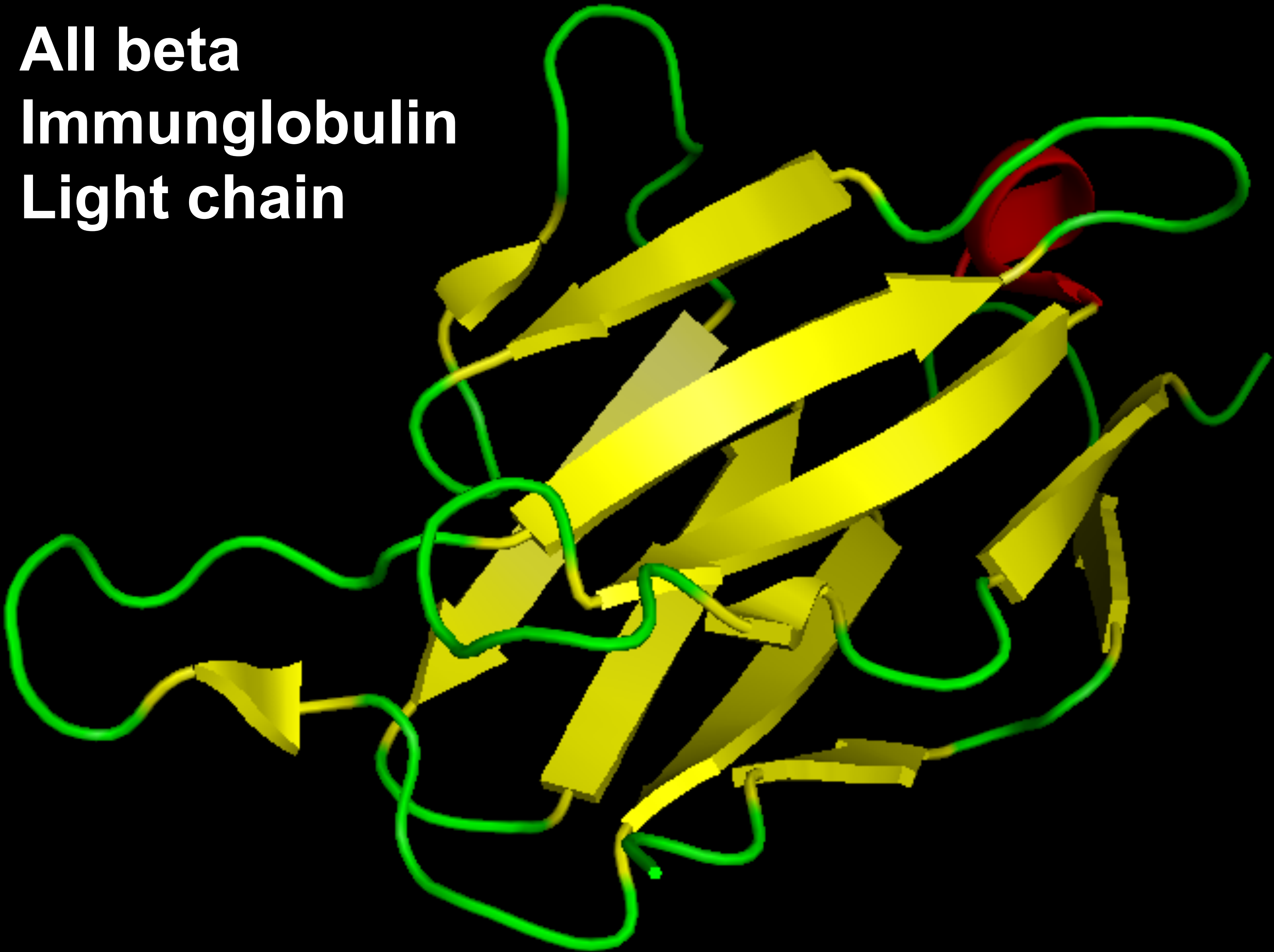
Crambin



α/β TIM barrel



All beta
Immunglobulin
Light chain



PREDICTION OF PROTEIN STRUCTURES

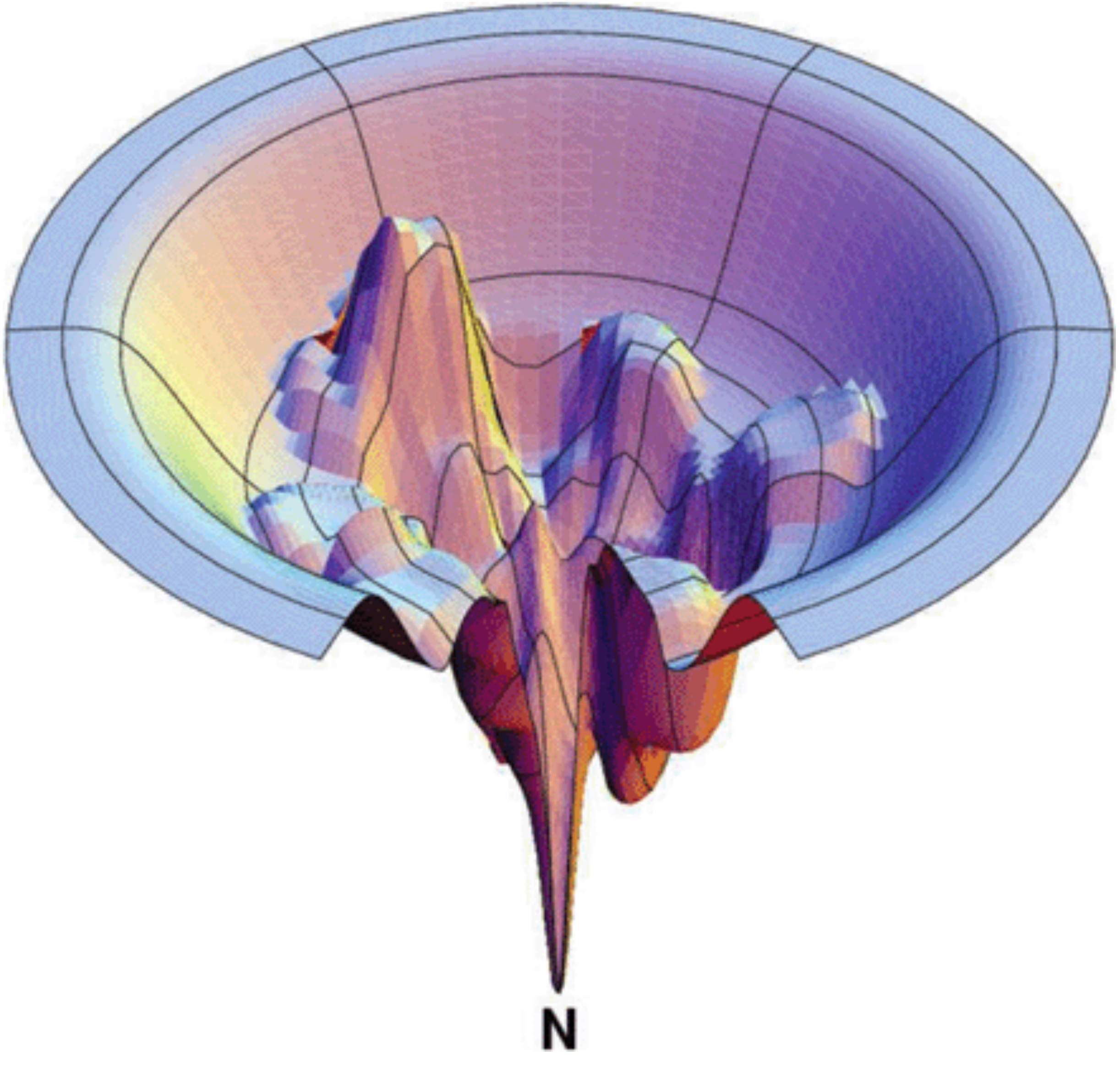
Anfinsen’s dogma (1961)

- Denatured proteins can refold *in vitro*
- No folding machinery required
- All information about the structure resides in the sequence
- Native structure: minimum free energy
  - Unique
  - Stable
  - Kinetically accessible

Levinthal’s Paradox

- Consider a protein with 101 residues
  - 100 Ψ and 100φ angels
  - If we assume only three stable positions and none for ω
  - 3^{200} or 10^{95} confirmations
  - Sampling all confirmations exceeds the life time of the universe
- Proteins fold in milliseconds

Folding landscape



Secondary structure

- Single sequence methods
  - Chou-Fasman
  - GOR
- Neural networks
  - PHD
- HMMs

Chou-Fasman

Nameetc	P(a)	P(b)	P(t)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanin	142	83	66	0.06	0.076	0.035	0.058
Threonie	83	119	96	0.086	0.108	0.065	0.065
…							

Calulate if P(a) > 100 for 4 out 6 AA, assign helix
Calculate if P(b) >100 for 3 out 5 AA assign sheet
Calculate p(t) = f(i) … assign turn
Further rules to resolve clashes

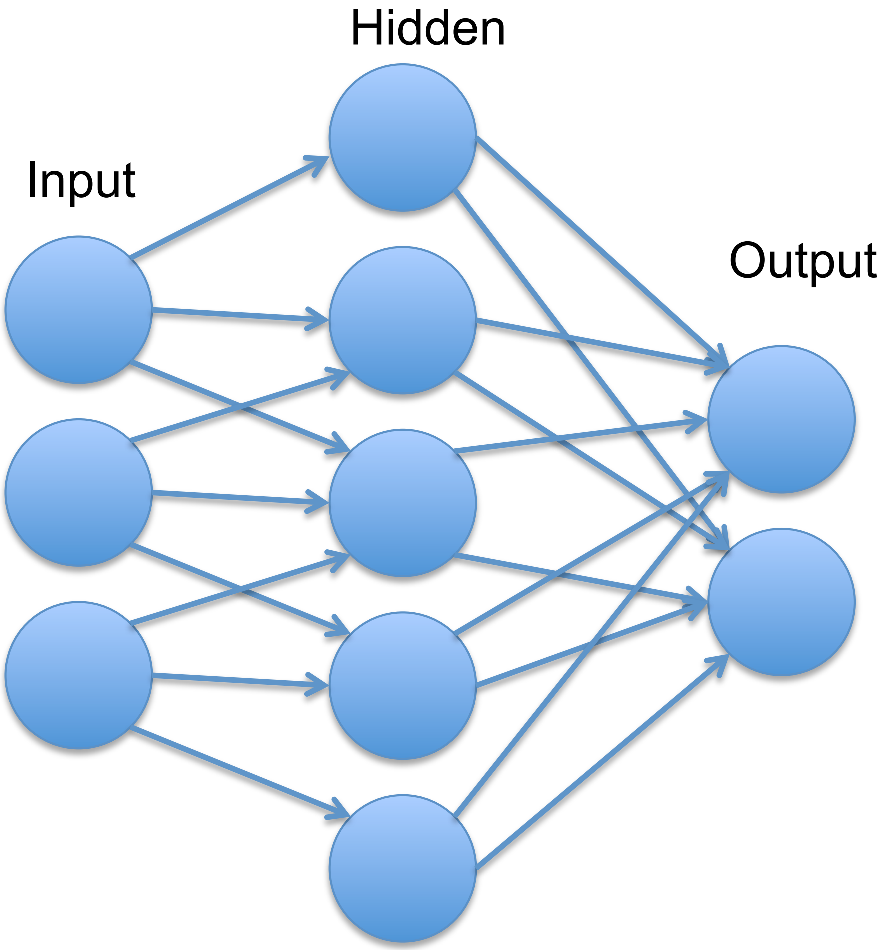
Chou and Fasman (1974) Biochemistry

Single sequence methods

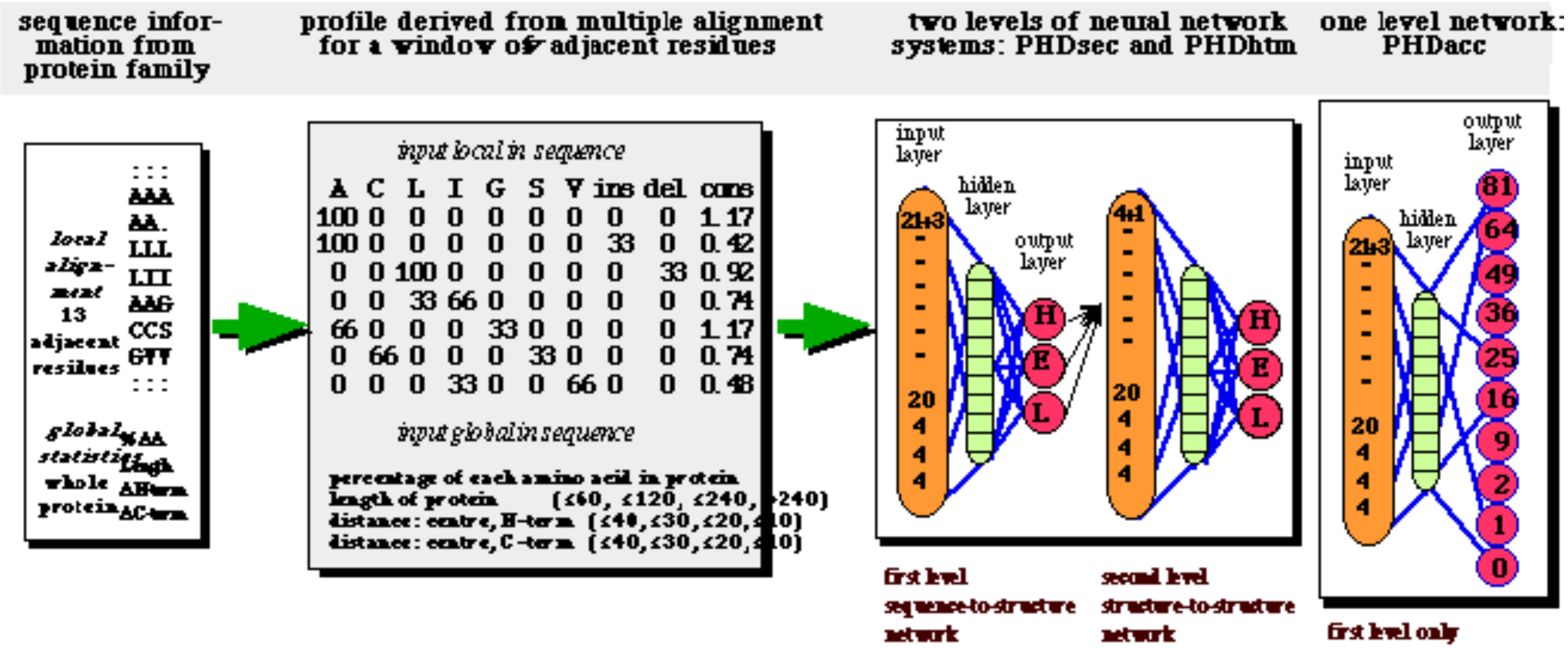
- Prediction based on propensity of an AA to occur in helix, sheet or turn
- Chou-Fasman
  - Empirical, rule based
- GOR
  - Log-odds score, Bayesian statistics

Neural network

Machine learning technique
inspired by neuronal structures



PHD



sequence information from protein family

profile derived from multiple alignment for a window of adjacent residues

two levels of neural network systems: PHDsec and PHDhtm

one level network: PHDacc

input local in sequence

A	C	L	I	G	S	V	ins	del	cons
100	0	0	0	0	0	0	0	0	1.17
100	0	0	0	0	0	0	33	0	0.42
0	0	100	0	0	0	0	0	33	0.92
0	0	33	66	0	0	0	0	0	0.74
66	0	0	0	33	0	0	0	0	1.17
0	66	0	0	0	33	0	0	0	0.74
0	0	0	33	0	0	66	0	0	0.48

input global in sequence

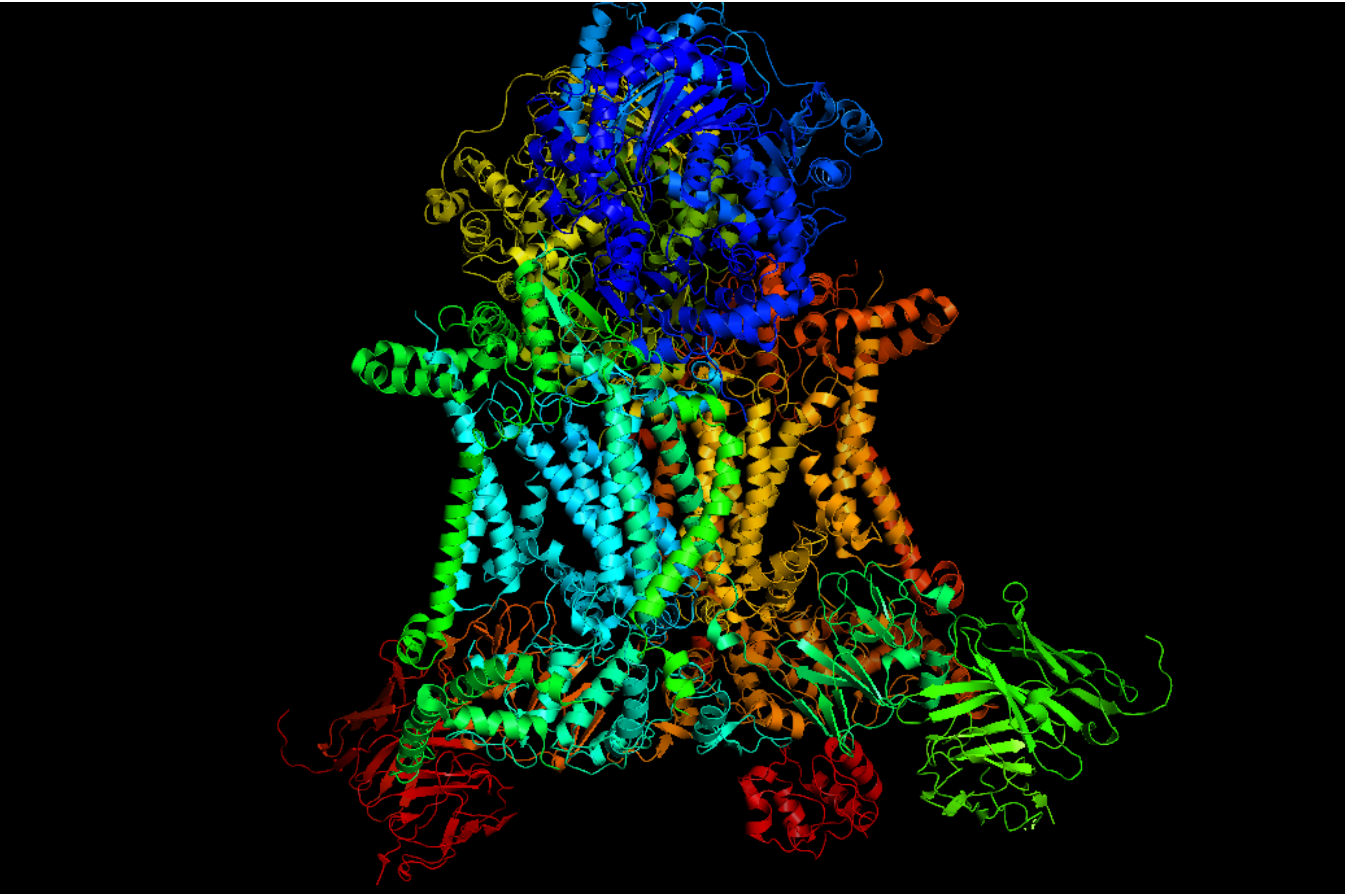
percentage of each amino acid in protein
length of protein (≤60, ≤120, ≤240, >240)
distance: centre, N-term (≤40,≤30,≤20,≤10)
distance: centre, C-term (≤40,≤30,≤20,≤10)

first level sequence-to-structure network

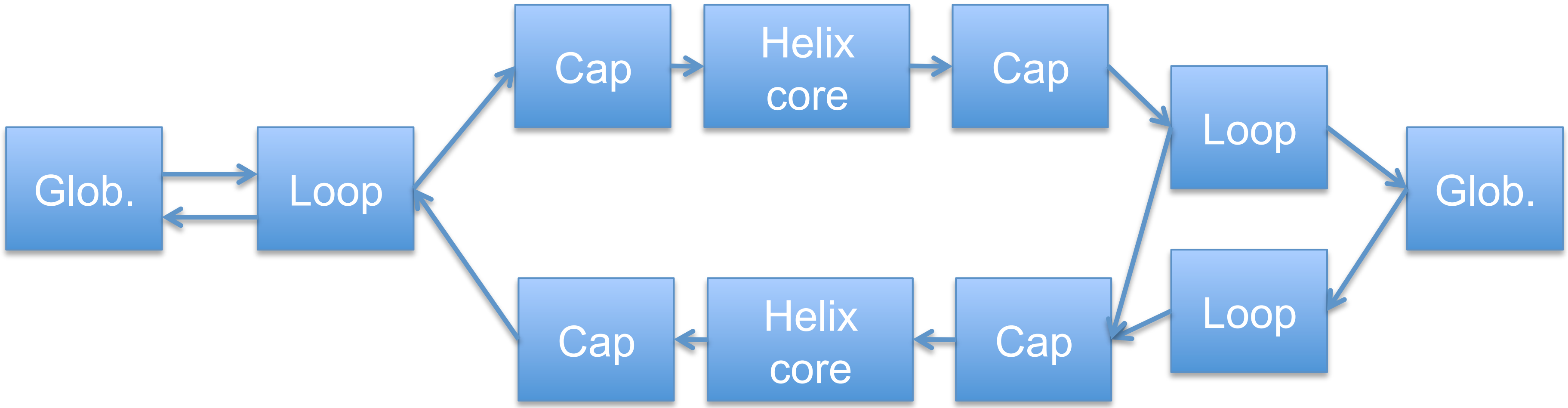
second level structure-to-structure network

first level only

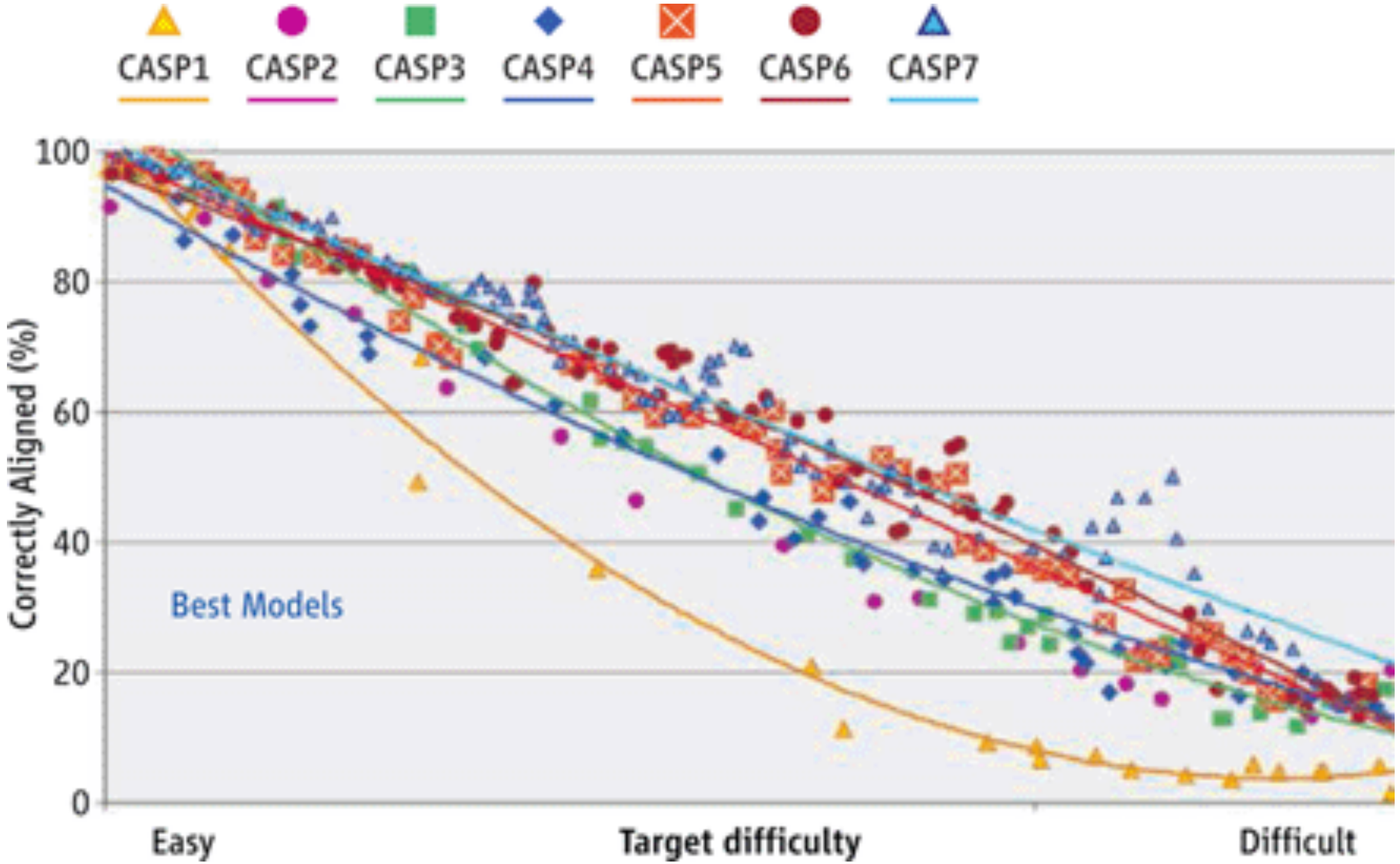
Rost (1996) Methods in Enzymology



TMHMM


Glob.
Loop
Cap
Helix core
Cap
Loop
Glob.
Loop
Cap
Helix core
Cap

CASP



A. KRYSHTAFOVYCH ET AL., PROTEINS: STRUCTURE, FUNCTION, AND BIOINFORMATICS (5 OCTOBER 2007)

Tertiary structure

- Homology modeling
- Threading
  - Fold recognition
- Ab initio modeling

RMSD

- Root-mean-square deviation

- Distance of backbone atoms
  - Usually cα

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

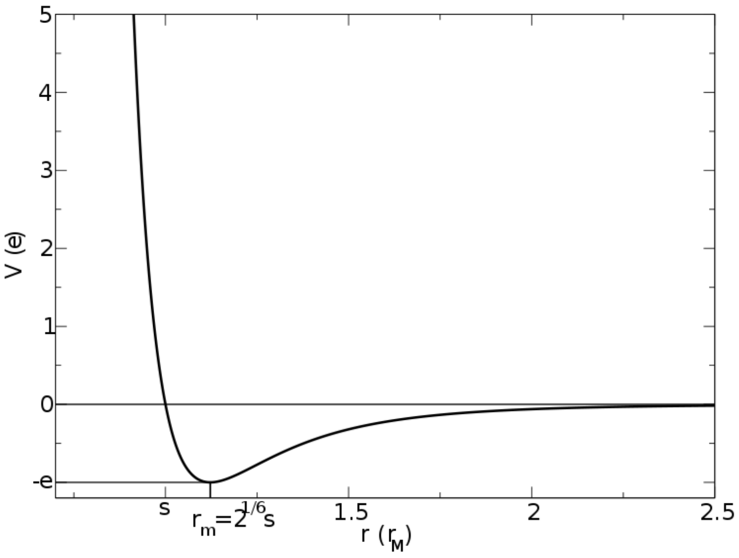
Some chemistry

- Intramolecular forces
  - Covalent bonds (400 kcal)
  - Strong but only relevant for cystin
- Intermolecular forces
  - Hydrogen bonds (12 – 16 kcal)
  - Van der Waals forces
    - Dipole-dipole (0.5 -2 kcal)
    - London (<1 kcal)
  - Buried hydrophobic faces

Lennard-Jones potential

- Summarizes the repulsion of atoms and attraction by van der Waals forces

$$V_{LJ} = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$
$$= \varepsilon \left[\left(\frac{r_m}{r}\right)^{12} - 2\left(\frac{r_m}{r}\right)^{6}\right]$$



Souce: Wikipedia. Lennard-Jones potential

Structure prediction

1. Find backbone structure
   1. Homology modeling
   2. Threading
   3. Ab initio prediction
2. Loop modeling
3. Sidechain packing
4. Refinement

Homology modeling

- Find homologous sequence (BLAST etc)
- Multiple alignment (Muscle etc)
- Replace backbone in defined, conserved parts
- Check core model and re-align
- Model side chain
- Model loop regions
- Energy minimization

Homology modeling

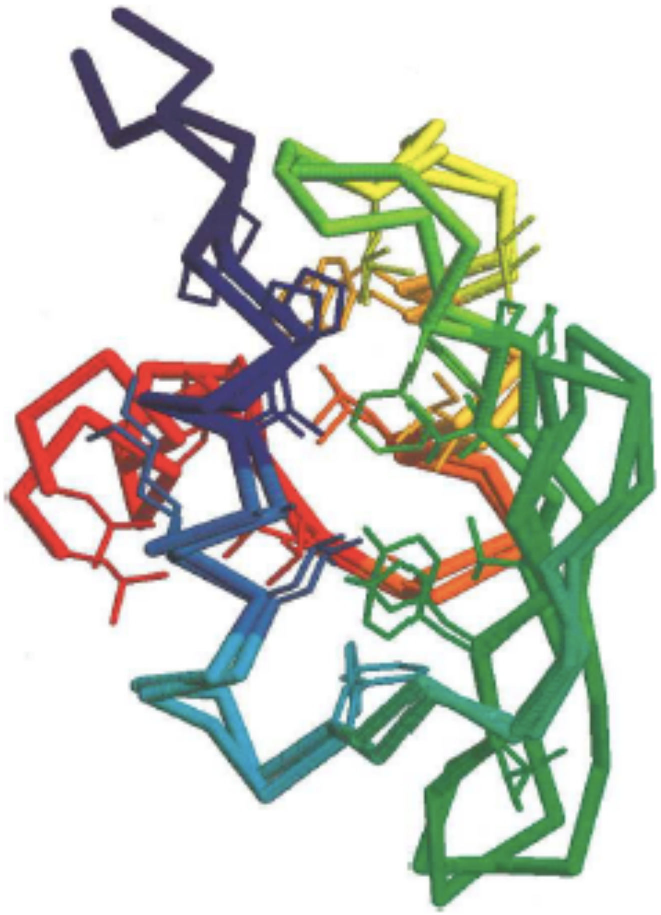
- Simple procedure for ID>40% over 50 AA (typical values, check for plausbility)
- Difficult if ID <25% over reasonable range
- Automated, SWISSMODEL available for all suitable targets
- If no template can be found:
  - Search template with sensitive methods: threading
  - Build from scratch: *ab initio*

Threading

- Naïve approach: Perform Homology Modeling for many/all templates, score the best
- Alignments at low %ID become problematic
- Fold recognition occasionally works, models often fail

Ab initio prediction

- Library of k-mers from known structures
- Build „random“ structures of k-mers
- Optimize in cycles, using a custom scoring function
- Analyze the top structures according to protein-like appearance and/or expectations from the literature.
- ROSETTA (Baker et al. (1998) outperformed contestants in CASP3.



Problem solved?

- Great improvements for globular proteins

- Open issues
  - Membrane proteins
  - Unstructured regions
  - Large assemblies