Structural bioinformatics

Computational methods for analyzing and predicting structures of biomacromolecules

Guillaume Postic, PhD

Assistant Professor of bioinformatics
Université Paris-Saclay, IBISC laboratory at Evry

Nov. 15, 2021

Structural bioinformatics: table of contents

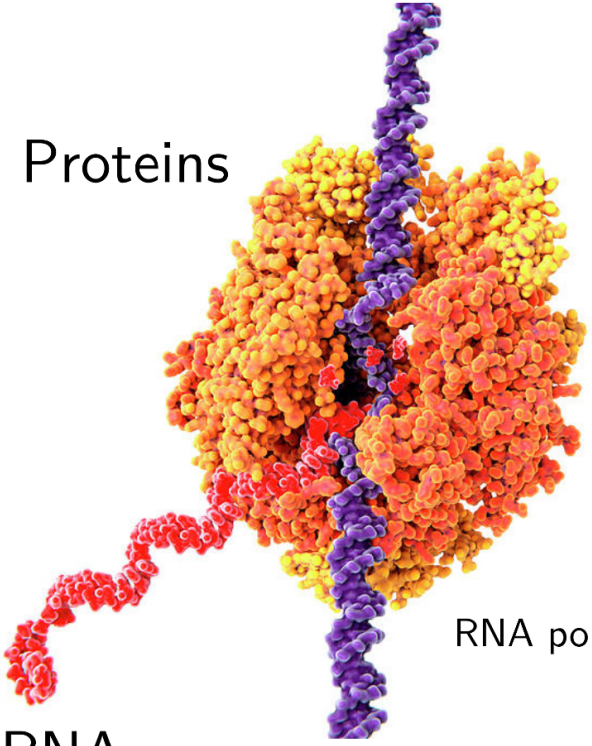
1. Introduction
   - Macromolecules
   - Structure and function
   - Molecular design
2. Experimental determination
   - X-ray, NMR, cryo-EM
   - X-ray, NMR, cryo-EM
   - The Protein Data Bank
3. Protein structure: analysis
   - Primary: sequence
   - Secondary: local conformations
   - Tertiary: global conformation
   - Intermediary
   - Quaternary: oligomery
   - Representation
4. Protein folding
   - Thermodynamics
   - Kinetics
5. Structure prediction
   - Preamble
   - Sampling
   - Scoring
   - Performance
   - Applications
6. Molecular flexibility
   - Molecular dynamics simulations
   - Advanced techniques
   - Normal mode analysis
7. Nucleic acids structures
   - DNA
   - RNA

Structural bioinformatics: table of contents

1. Introduction
   - Macromolecules
   - Structure and function
   - Molecular design
2. Experimental determination
   - X-ray, NMR, cryo-EM
   - X-ray, NMR, cryo-EM
   - The Protein Data Bank
3. Protein structure: analysis
   - Primary: sequence
   - Secondary: local conformations
   - Tertiary: global conformation
   - Intermediary
   - Quaternary: oligomery
   - Representation
4. Protein folding
   - Thermodynamics
   - Kinetics
5. Structure prediction
   - Preamble
   - Sampling
   - Scoring
   - Performance
   - Applications
6. Molecular flexibility
   - Molecular dynamics simulations
   - Advanced techniques
   - Normal mode analysis
7. Nucleic acids structures
   - DNA
   - RNA

Macromolecules

Proteins

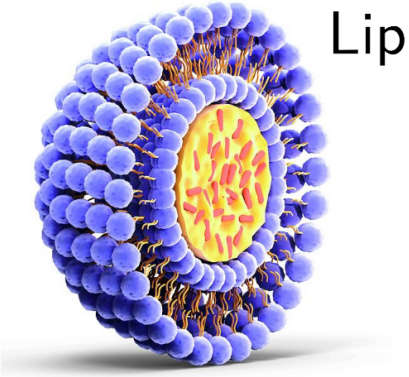


RNA polymerase II

RNA

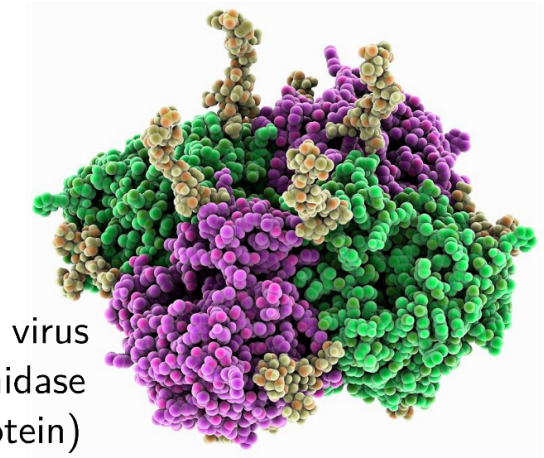
DNA

Lipids



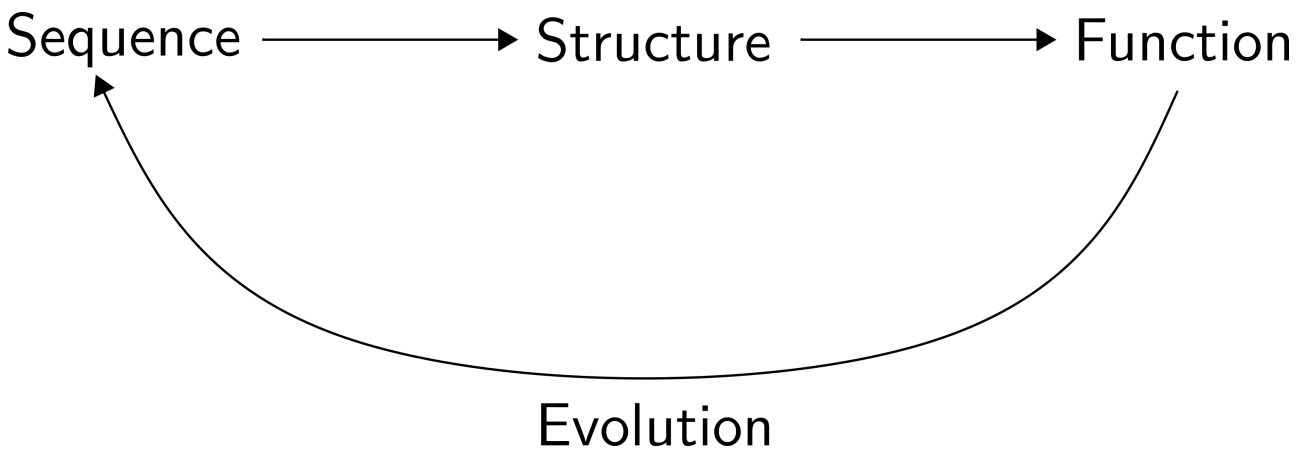
Liposome

Polysaccharides



Influenza virus neuraminidase (glycoprotein)

The sequence–structure–function paradigm



Protein structure is more conserved than sequence throughout evolution, thus maintaining function

Function ⇔ structure (and interaction)

Evolution:
different structures
but similar functions

A single object
with multiple functions

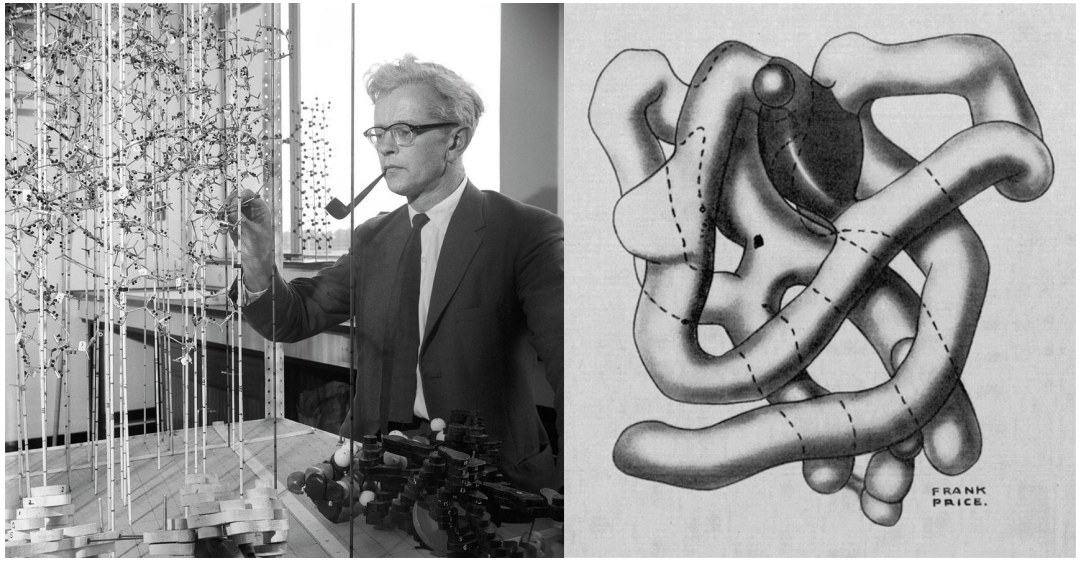
Understanding molecular mechanisms: applications

- In silico drug design requires structural information
  - Receptor: transmembrane or cytosolic protein
  - Drug ligand: peptide
- Protein engineering: rational design
  - Artificial proteins with new properties
  - Production in vitro (e.g. drugs)
  - Genetically modified organisms

Structural bioinformatics: table of contents

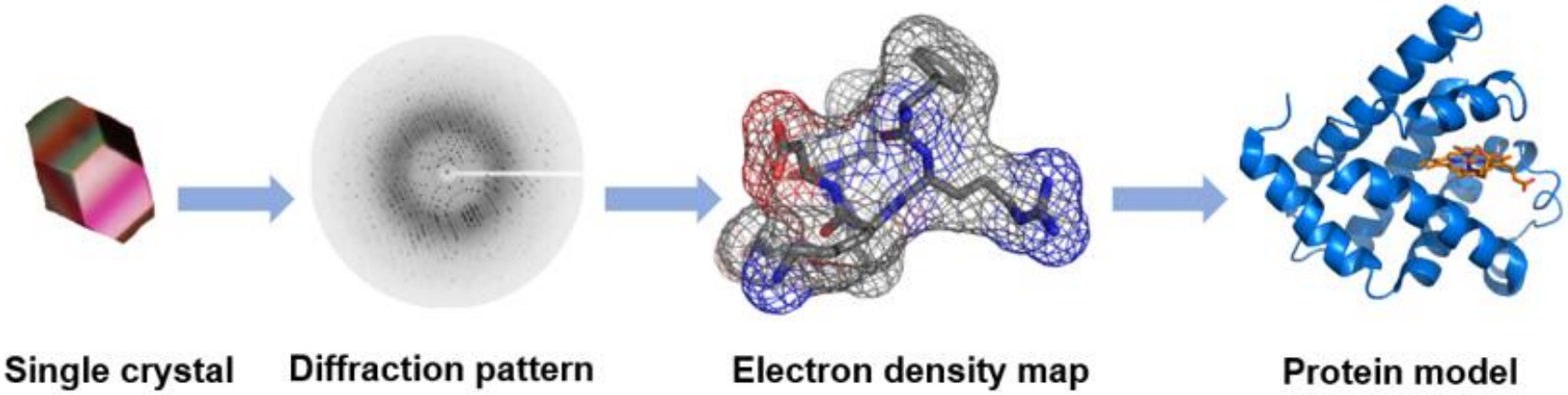
1. Introduction
   - Macromolecules
   - Structure and function
   - Molecular design
2. Experimental determination
   - X-ray, NMR, cryo-EM
   - X-ray, NMR, cryo-EM
   - The Protein Data Bank
3. Protein structure: analysis
   - Primary: sequence
   - Secondary: local conformations
   - Tertiary: global conformation
   - Intermediary
   - Quaternary: oligomery
   - Representation
4. Protein folding
   - Thermodynamics
   - Kinetics
5. Structure prediction
   - Preamble
   - Sampling
   - Scoring
   - Performance
   - Applications
6. Molecular flexibility
   - Molecular dynamics simulations
   - Advanced techniques
   - Normal mode analysis
7. Nucleic acids structures
   - DNA
   - RNA

Determination of the myoglobin structure



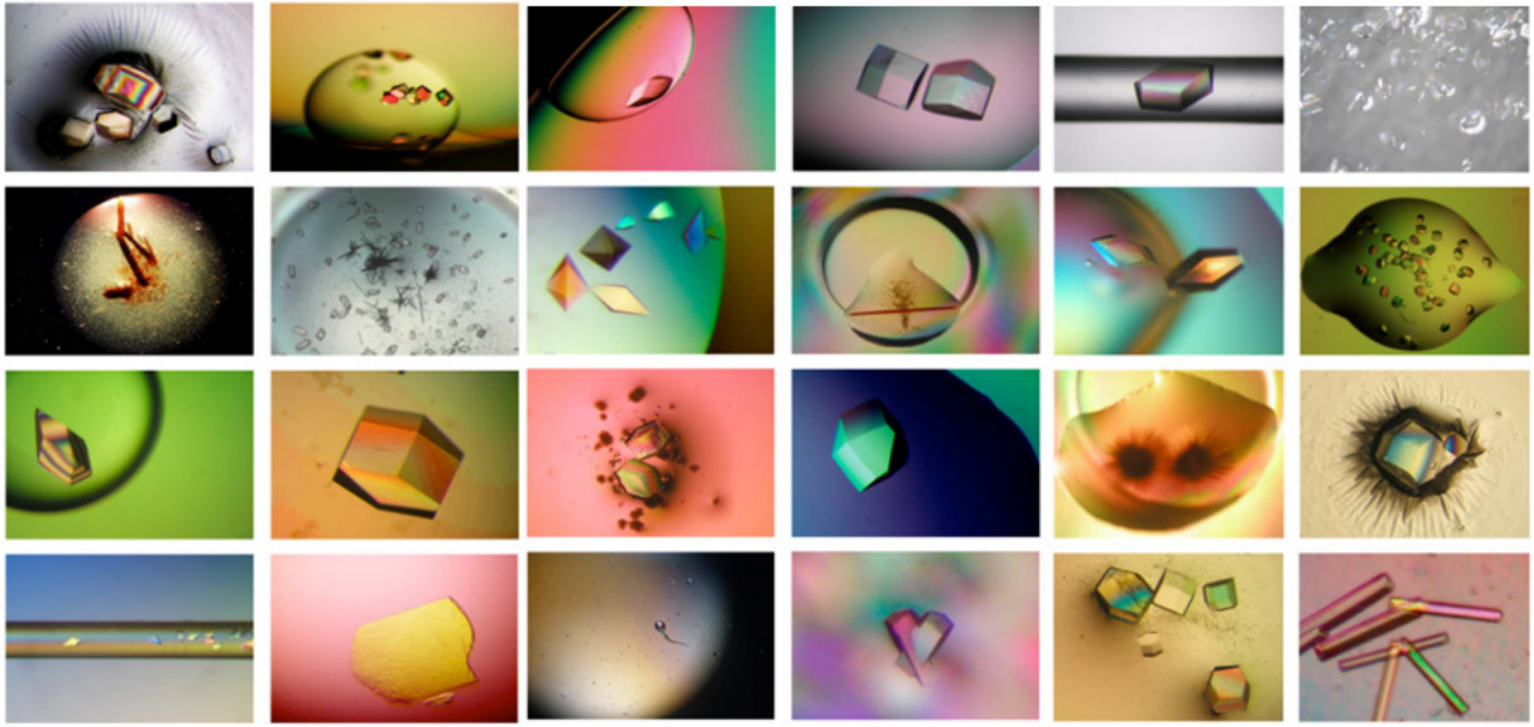
John Kendrew, Nobel Prize for Chemistry (1962)

X-ray crystallography



90% of the published structures

X-ray crystallography



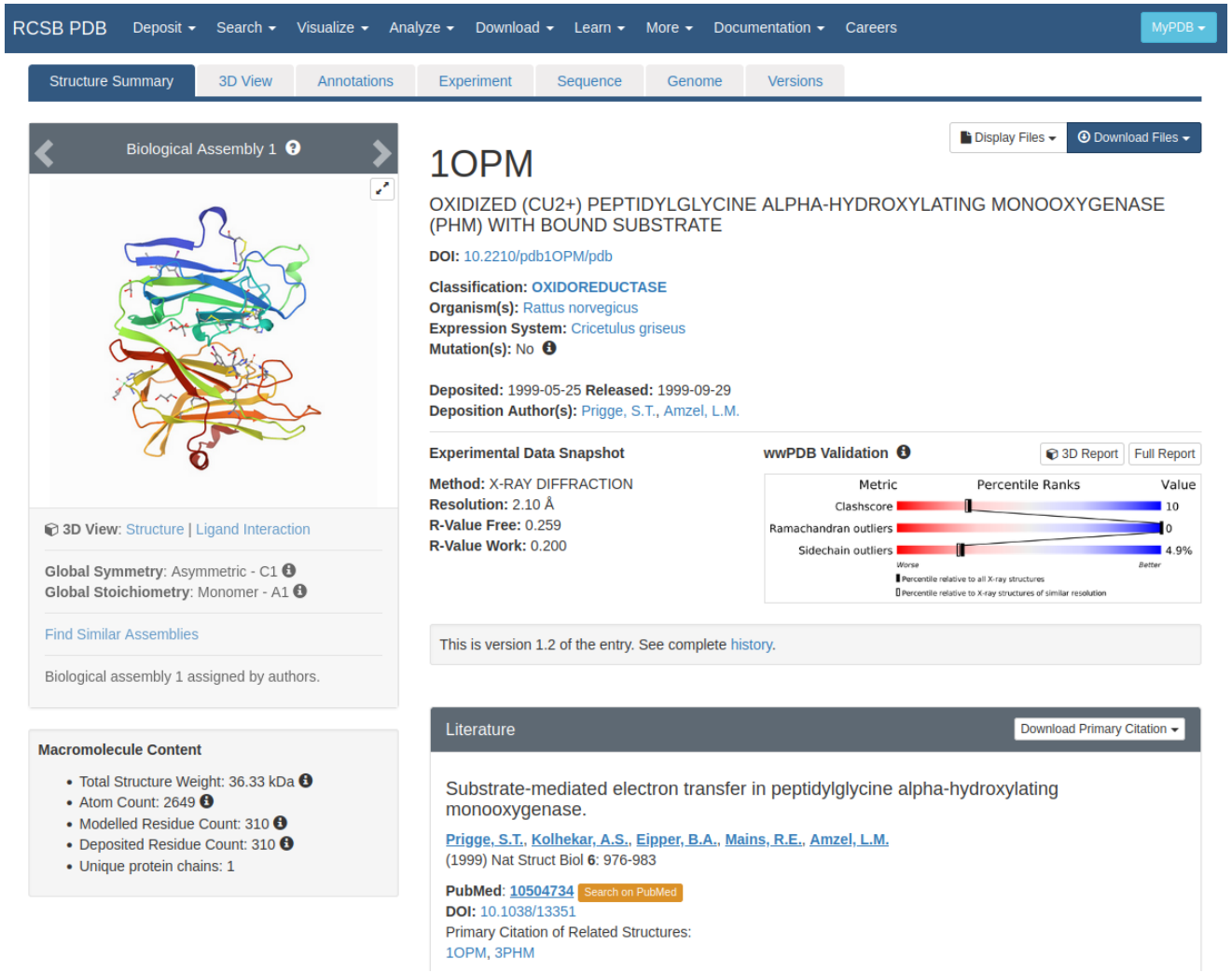
Solution nuclear magnetic resonance (NMR) spectroscopy

- Less than 8% of the published structures
- May provide unique information about dynamics and intermolecular interactions
- Not adapted to large molecules
- Generate ensemble of models, rather than a single one

Cryogenic electron microscopy (cryo-EM)

- 2% of the published structures (newest approach)
- Resolution of very large multimeric structures
- e.g. Heteromultimeric structure (360-mer) of the Zika virus

The Protein Data Bank



Web portals: RCSB PDB (USA); PDBe (Europe); PDBj (Japan);

The PDB format

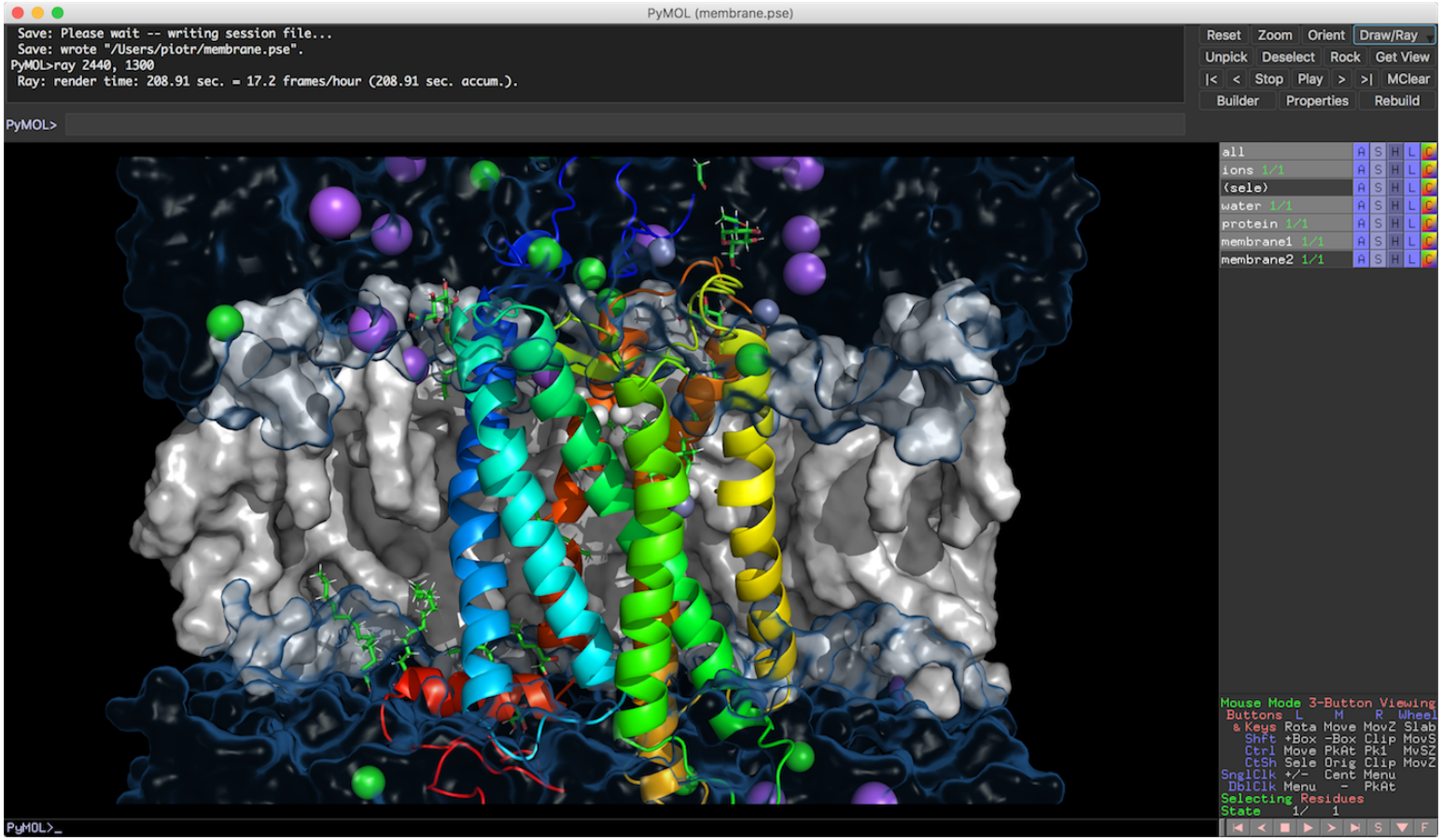
```
column          1         2         3         4         5         6         7         8
number   12345678901234567890123456789012345678901234567890123456789012345678901234567890

         ATOM    1  N   GLY A   3      17.119   0.186  36.320  1.00 64.10           N
         ATOM    2  CA  GLY A   3      16.944  -0.800  35.208  1.00 63.46           C
         ATOM    3  C   GLY A   3      16.818  -0.087  33.851  1.00 61.22           C
         ATOM    4  O   GLY A   3      15.721   0.337  33.463  1.00 62.81           O
         ATOM    5  N   PRO A   4      17.944   0.077  33.129  1.00 57.39           N
         ATOM    6  CA  PRO A   4      17.950   0.742  31.815  1.00 53.27           C
         ATOM    7  C   PRO A   4      18.005  -0.247  30.629  1.00 49.78           C
         ATOM    8  O   PRO A   4      19.086  -0.678  30.218  1.00 48.17           O
         ATOM    9  CB  PRO A   4      19.191   1.613  31.898  1.00 54.33           C
         ATOM   10  CG  PRO A   4      20.161   0.686  32.625  1.00 55.45           C
         ATOM   11  CD  PRO A   4      19.305   0.019  33.701  1.00 55.83           C

field id  1   2   3 4 5 6 7 8   9    10    11    12   13          14 15
```

field id	definition	length	format	range	string slicing (Python)
1	"ATOM " or "HETATM"	6	{:6s}	01-06	[0:6]
2	atom serial number	5	{:5d}	07-11	[6:11]
3	atom name	4	{:^4s}	13-16	[12:16]
4	alternate location indicator	1	{:1s}	17	[16:17]
5	residue name	3	{:3s}	18-20	[17:20]
6	chain identifier	1	{:1s}	22	[21:22]
7	residue sequence number	4	{:4d}	23-26	[22:26]
8	code for insertion of residues	1	{:1s}	27	[26:27]
9	orthogonal coordinates for X (in Angstroms)	8	{:8.3f}	31-38	[30:38]
10	orthogonal coordinates for Y (in Angstroms)	8	{:8.3f}	39-46	[38:46]
11	orthogonal coordinates for Z (in Angstroms)	8	{:8.3f}	47-54	[46:54]
12	occupancy	6	{:6.2f}	55-60	[54:60]
13	temperature factor	6	{:6.2f}	61-66	[60:66]
14	element symbol	2	{:>2s}	77-78	[76:78]
15	charge on the atom	2	{:2s}	79-80	[78:80]

Visualizing PDB files

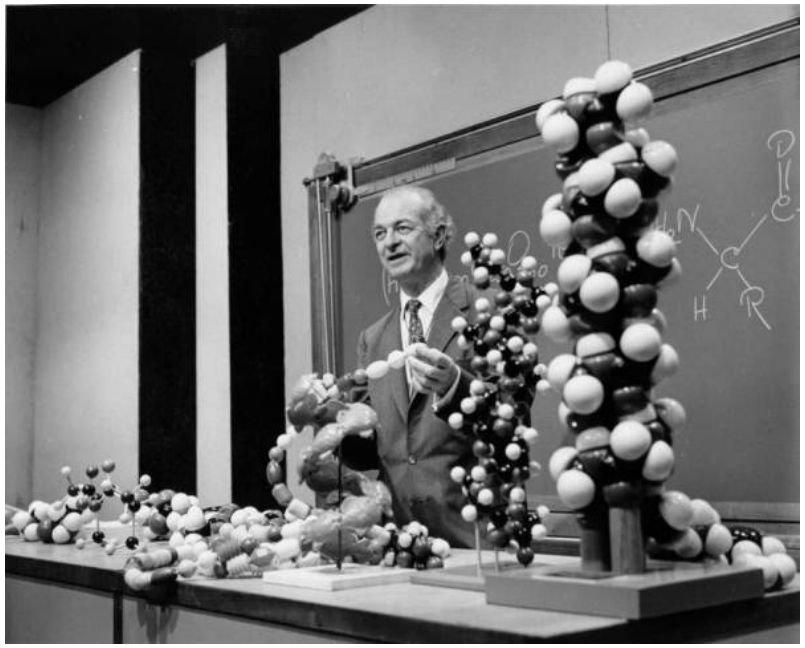


Structural bioinformatics: table of contents

1. Introduction
   - Macromolecules
   - Structure and function
   - Molecular design
2. Experimental determination
   - X-ray, NMR, cryo-EM
   - X-ray, NMR, cryo-EM
   - The Protein Data Bank
3. Protein structure: analysis
   - Primary: sequence
   - Secondary: local conformations
   - Tertiary: global conformation
   - Intermediary
   - Quaternary: oligomery
   - Representation
4. Protein folding
   - Thermodynamics
   - Kinetics
5. Structure prediction
   - Preamble
   - Sampling
   - Scoring
   - Performance
   - Applications
6. Molecular flexibility
   - Molecular dynamics simulations
   - Advanced techniques
   - Normal mode analysis
7. Nucleic acids structures
   - DNA
   - RNA

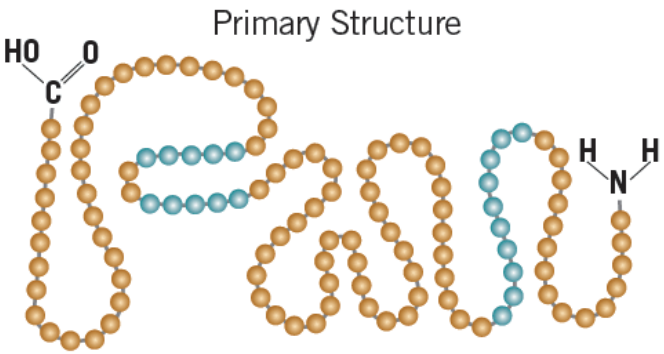
Definition

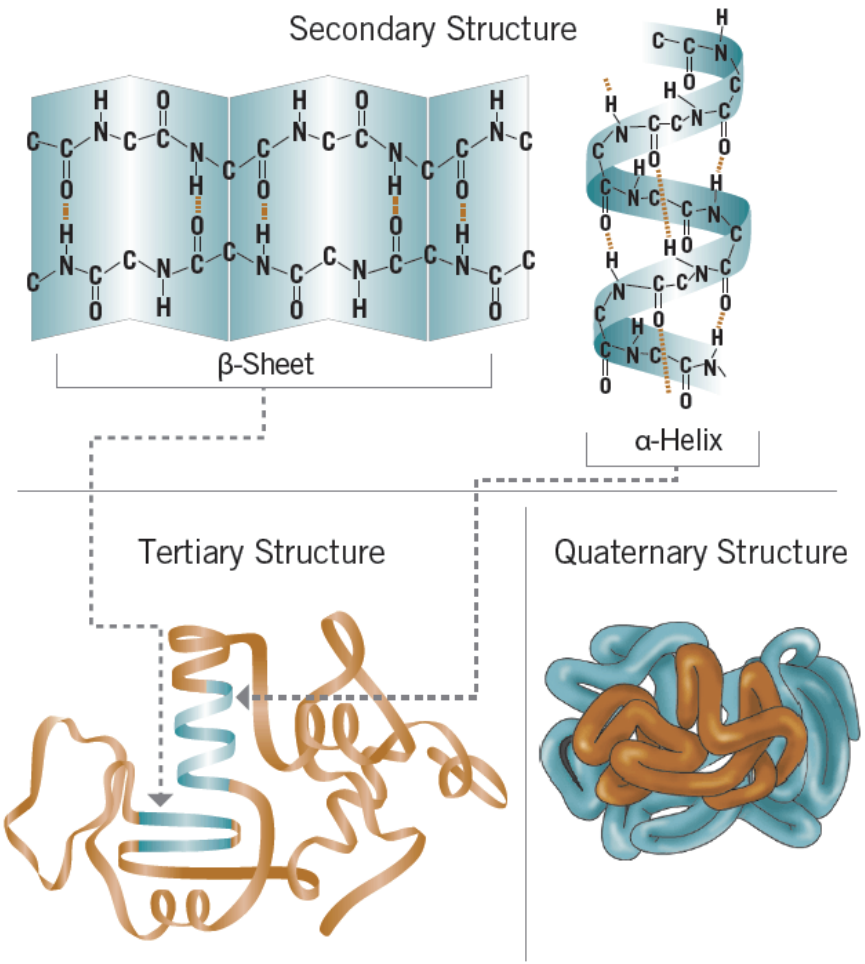
Analysis is the process of separating a whole into its constituent parts to gain a better understanding of it.



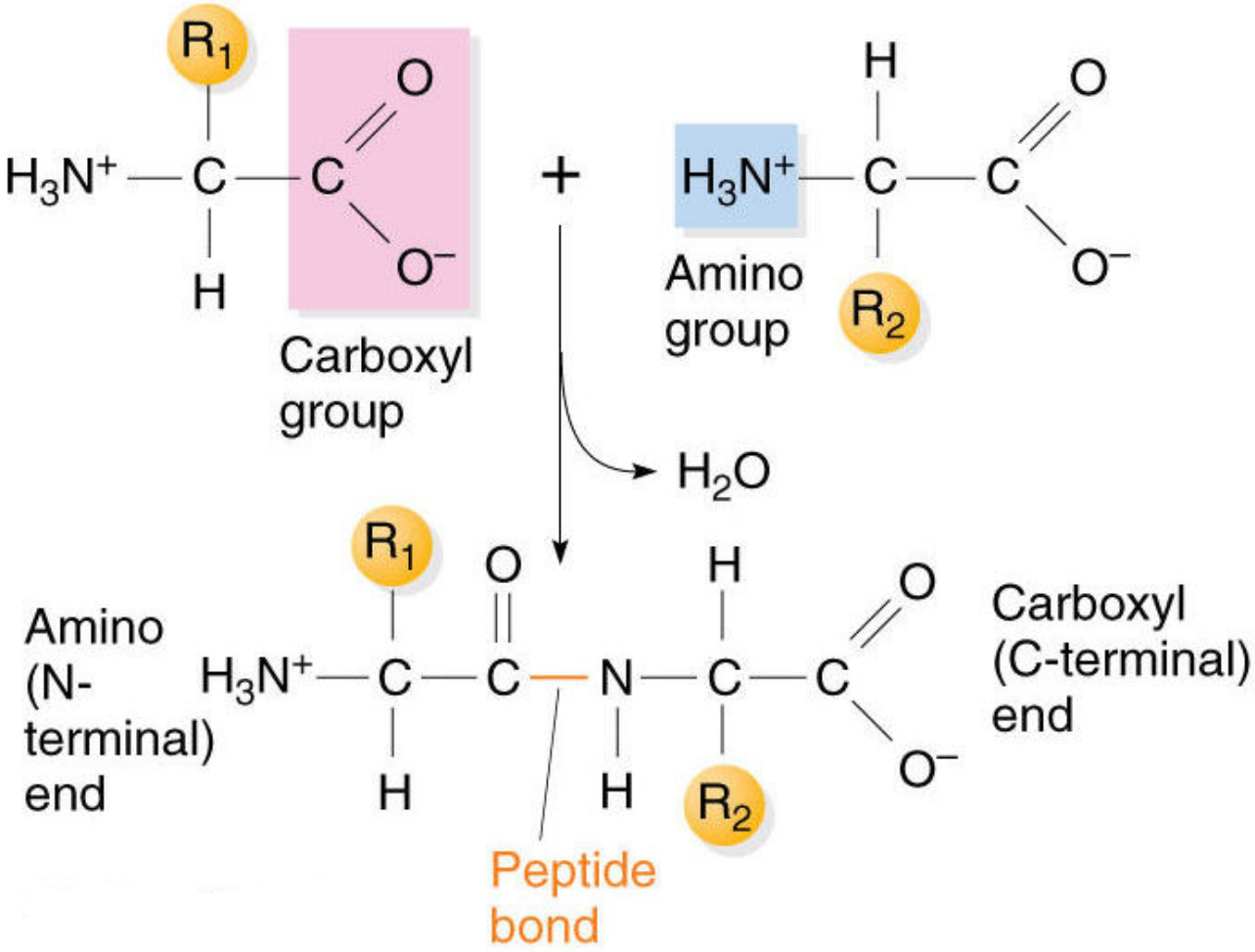
Linus Pauling

Overview

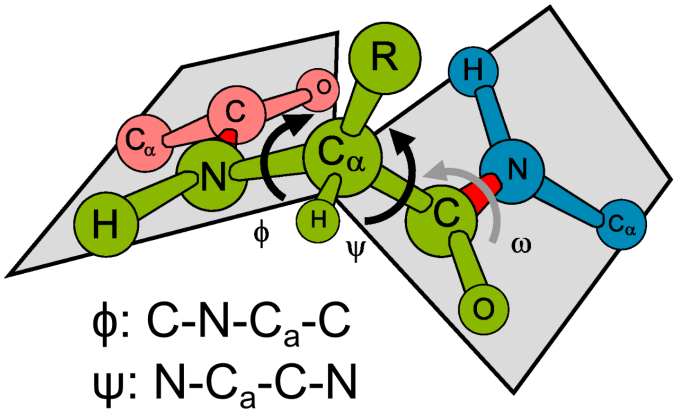

Primary Structure
HO
O
C
H
N
H


Secondary Structure
β-Sheet
α-Helix
Tertiary Structure
Quaternary Structure

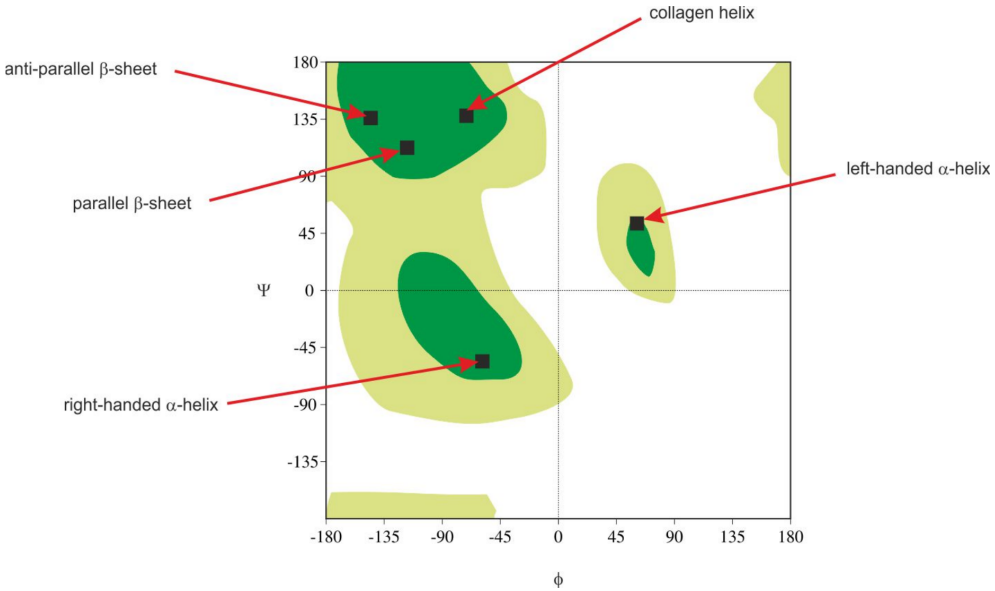
- 20 AA + non-standard
- Peptide bond + backbone + side chain
- `https://www.uniprot.org/`



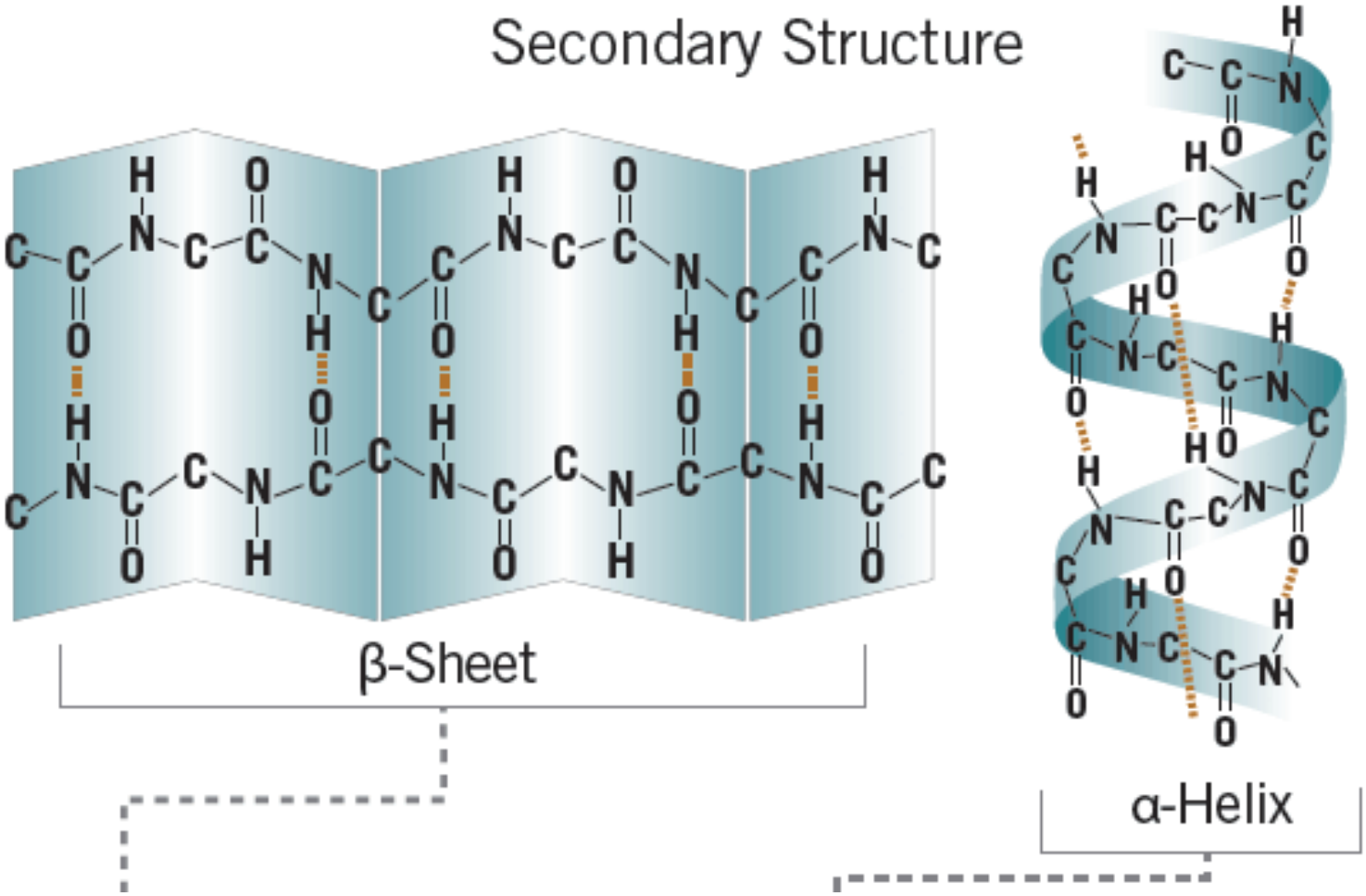
- Defined by backbone atoms
- Dihedral (torsion) angles
- Ramachandran phi/psi plot: clusters of values



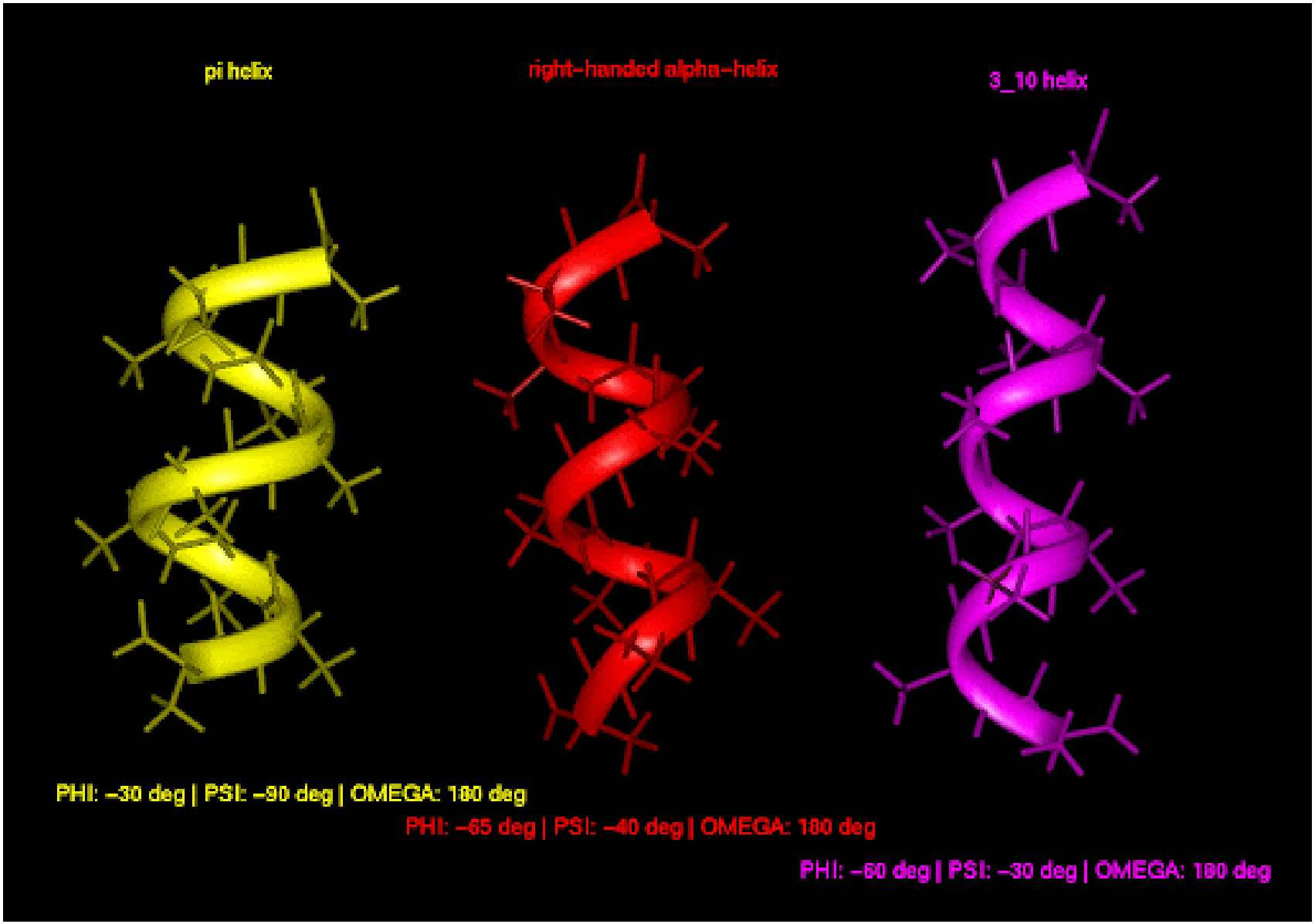
ϕ: C-N-C_a-C
ψ: N-C_a-C-N
ω: C_a-C-N-C_a



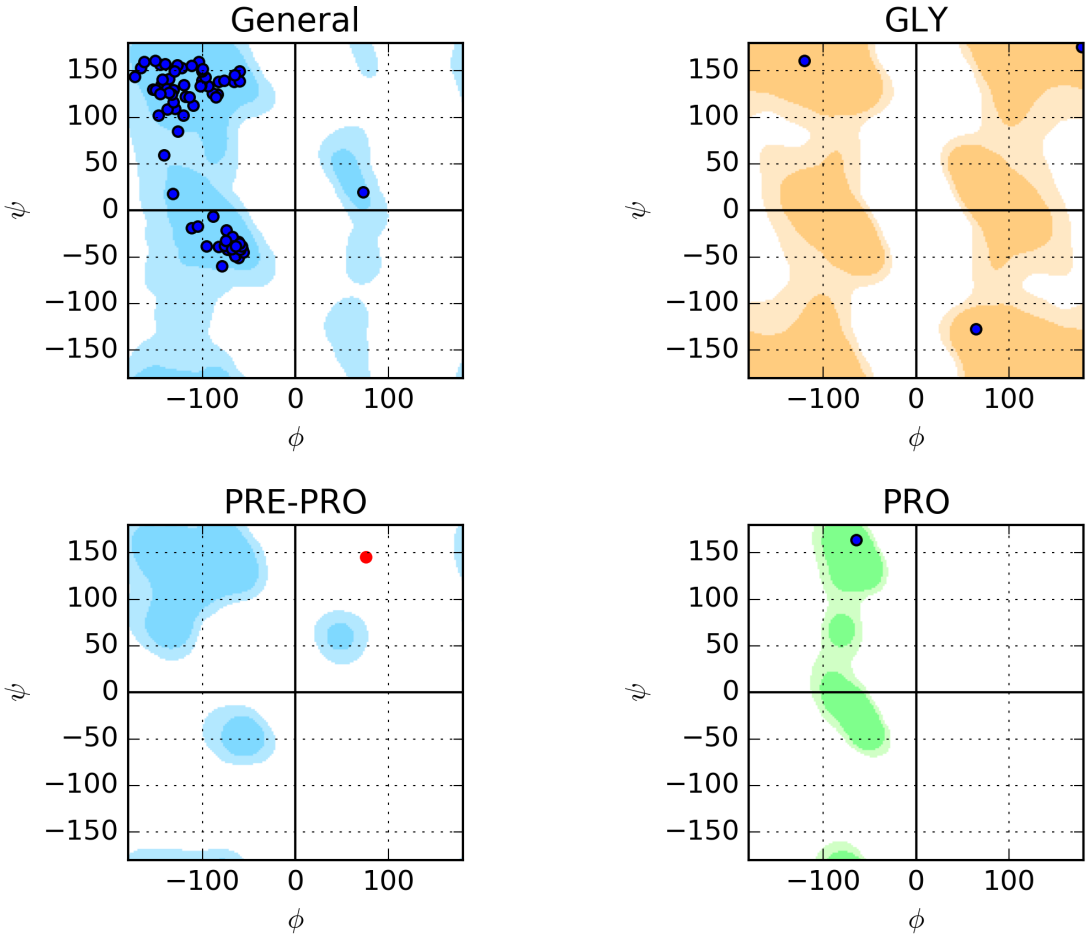
Secondary: local conformations



Secondary: local conformations



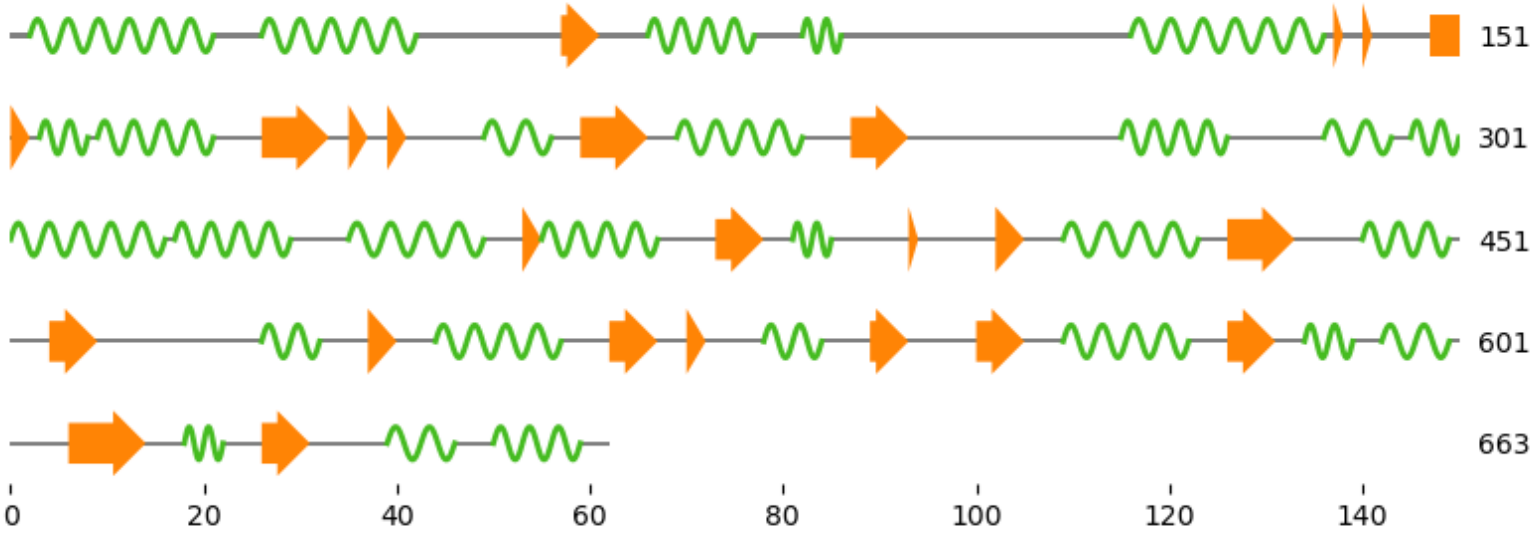
Differences in helicogenicity



Secondary structure assignment

Based on experimentally determined coordinates

- DSSP (Define Secondary Structure of Proteins), Kabsch and Sander, 1983
- STRIDE (STRuctural IDEntification), Frishman and Argos, 1995
- 95% agreement due to flexibility (helix vs strand) on some positions
- Many other methods using other criteria → higher discrepancies



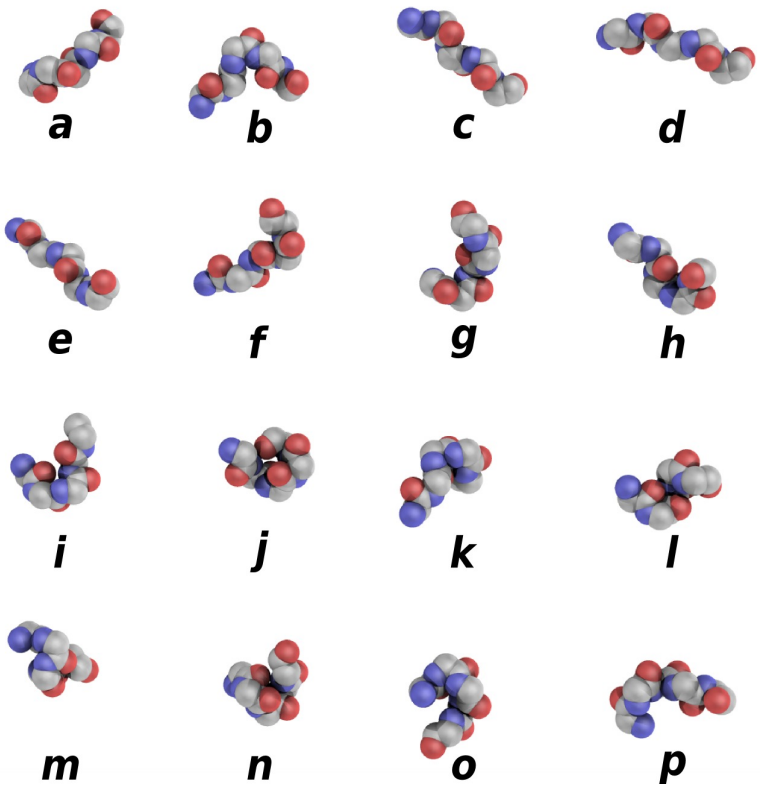
Beyond the 3 regular states

The 8 types of secondary structure that DSSP defines

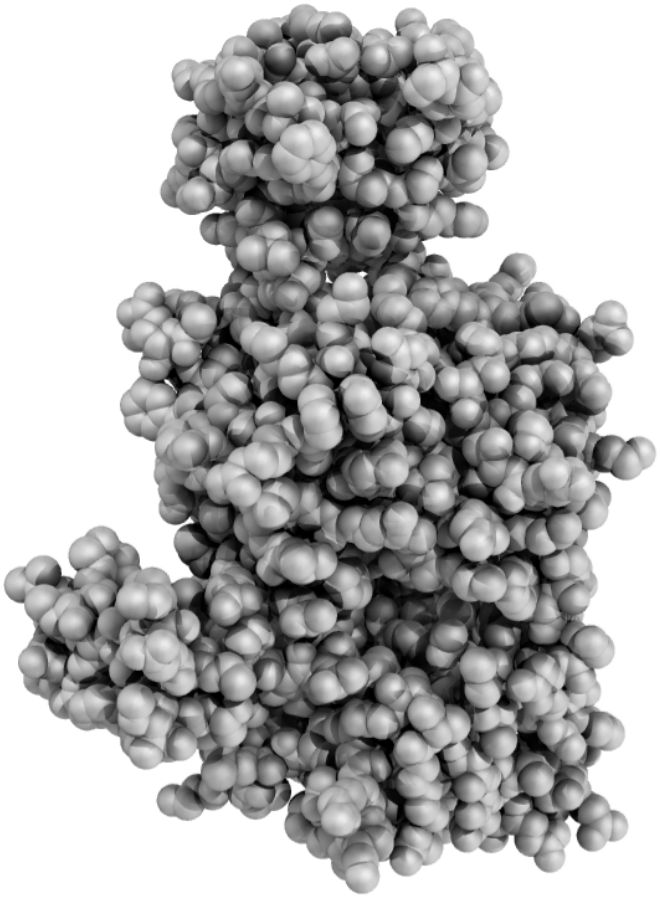
- G = 3-turn helix (3_{10} helix). Min length 3 residues.
- H = 4-turn helix (α helix). Minimum length 4 residues.
- I = 5-turn helix (π helix). Minimum length 5 residues.
- T = hydrogen bonded turn (3, 4 or 5 turn)
- E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues.
- B = residue in isolated β-bridge (single pair β-sheet hydrogen bond formation)
- S = bend (the only non-hydrogen-bond based assignment).
- C = coil (residues which are not in any of the above conformations).

Beyond the 3 regular states

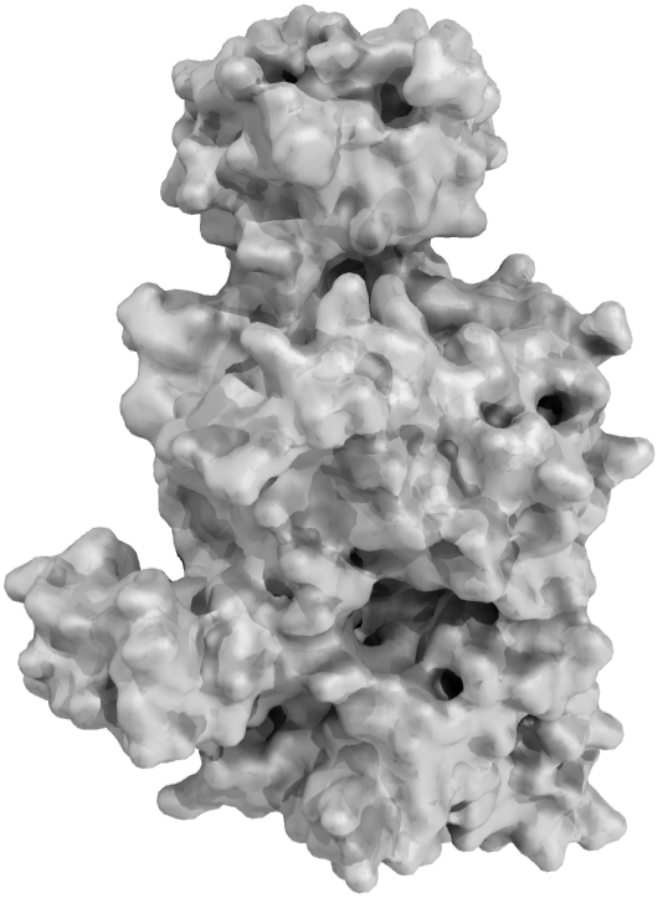
Example of a "structural alphabet" with 16 states
→ 3D structure as a **sequence (1D) of local conformations**
→ Powerful approach



The three dimensional shape of a protein (PDB entry 1PKN)

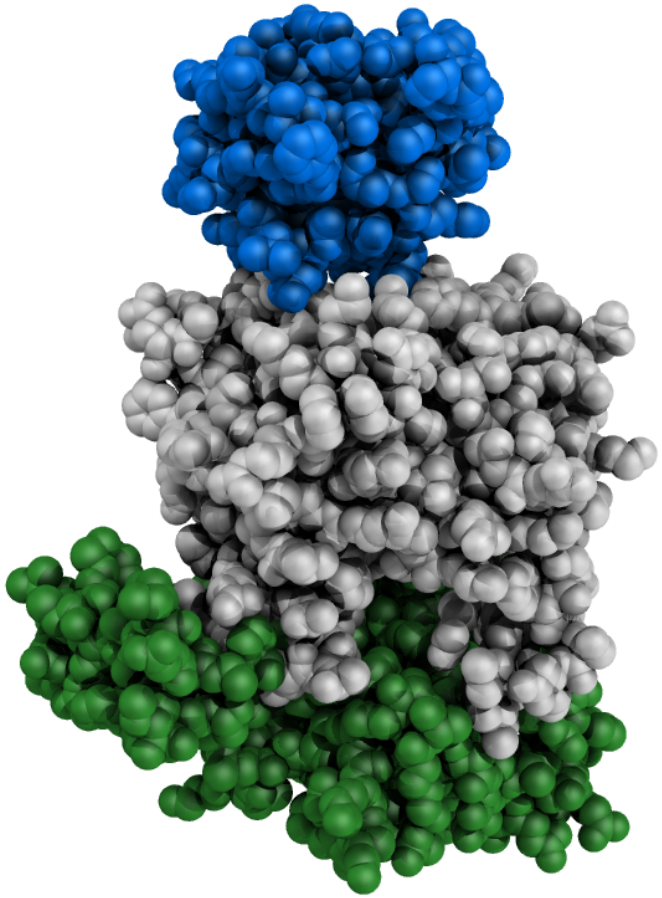


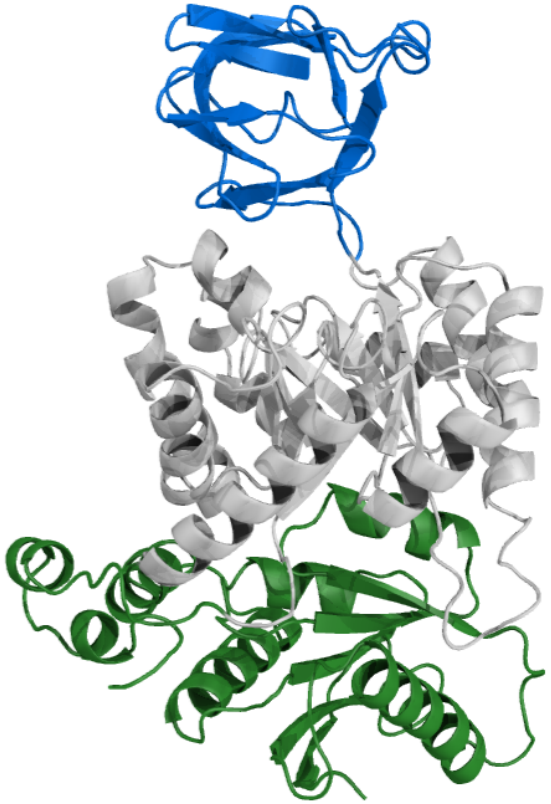
Sphere representation
(van der Waals radius)



Molecular surface
Solvent accessibility is calculated with NACCESS
(Hubbard and Thornton, 1992)

Structural domains





Minimum size = 30 residues (flexibility)

Note: peptides are proteins below 50 amino acids

Structural domains

Wetlaufer (1973), *PNAS*

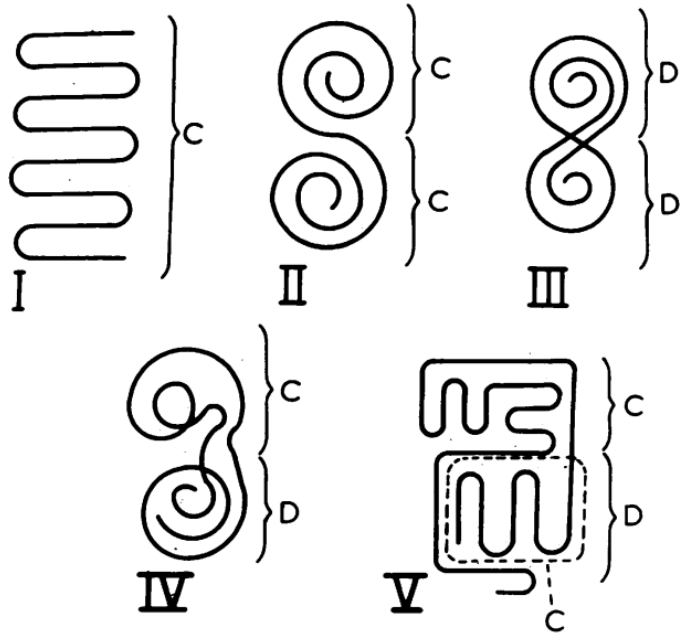


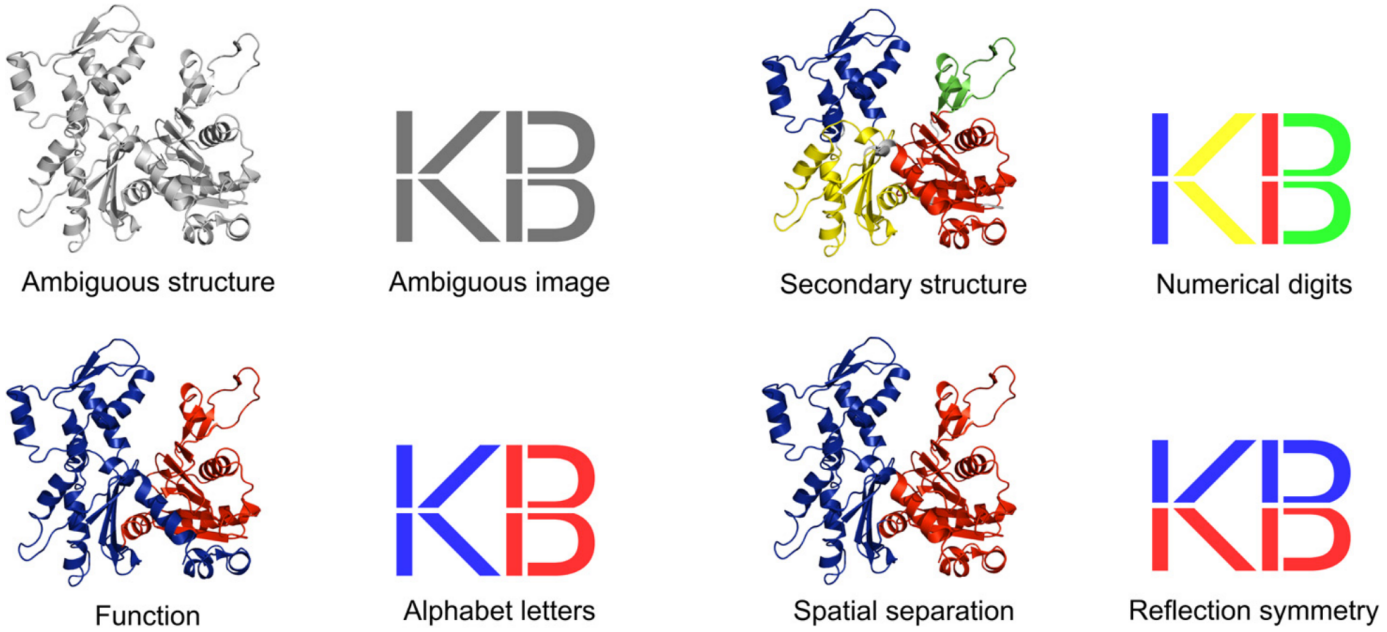
FIG. 1. Schematic peptide chain arrangements to illustrate continuous (*C*) and discontinuous (*D*) structural regions. Scheme I represents a one-region protein, which is by definition continuous. II represents a two-region protein, both regions of which are continuous. In III the two regions represented are both discontinuous. In IV the upper region is continuous, the lower discontinuous. In V the upper region is continuous and, taken as a whole, the lower region is discontinuous. A continuous region in the lower part of Scheme V is enclosed by a *dashed line*.

Structural domains

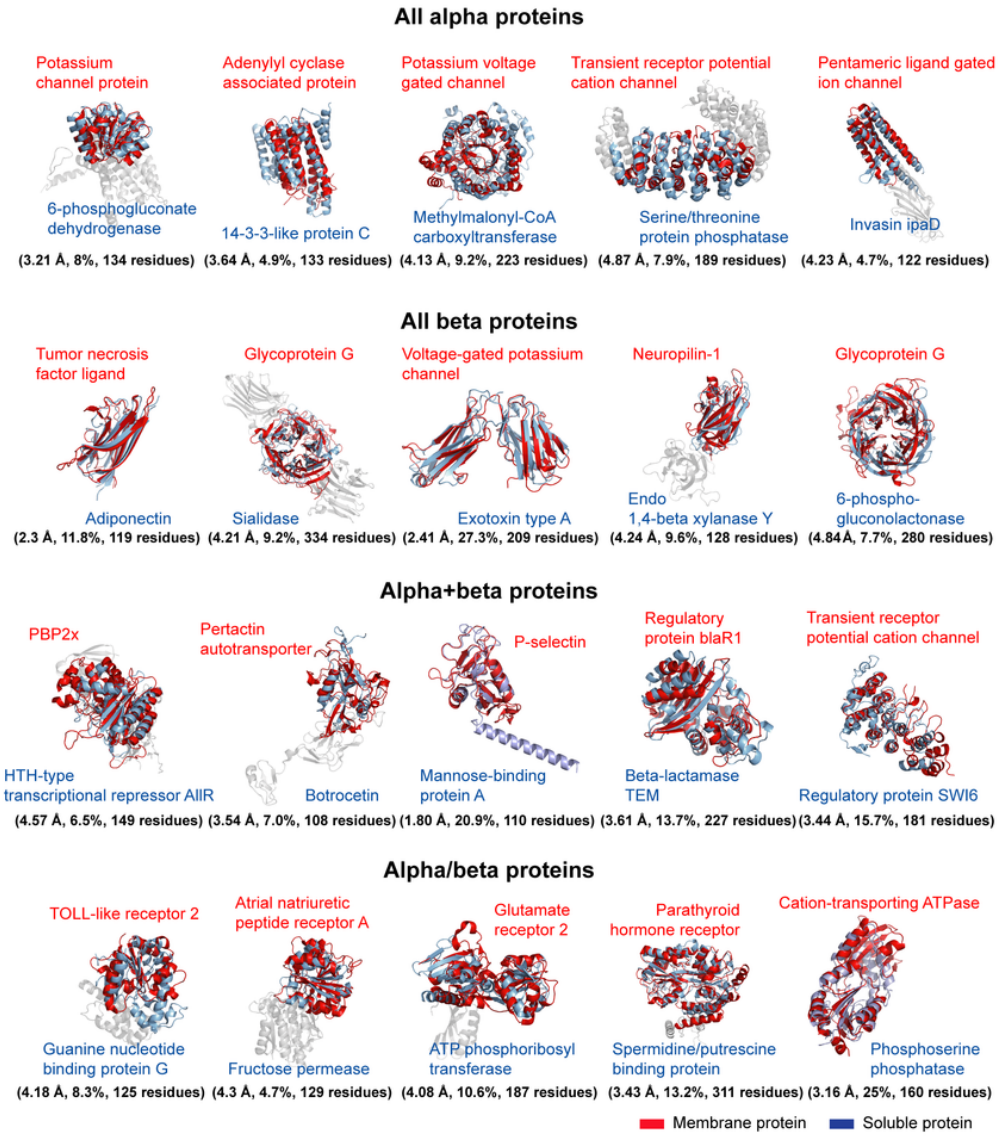
Definition based on multiple criteria: e.g. geometry, function, evolution, independent folding, motion, secondary structure content

→ Convergence shows importance of the concept

→ Discrepancies between the CATH and SCOP databases (automated domain partitioning)



Domain and secondary structure contents: classes



Supersecondary structure

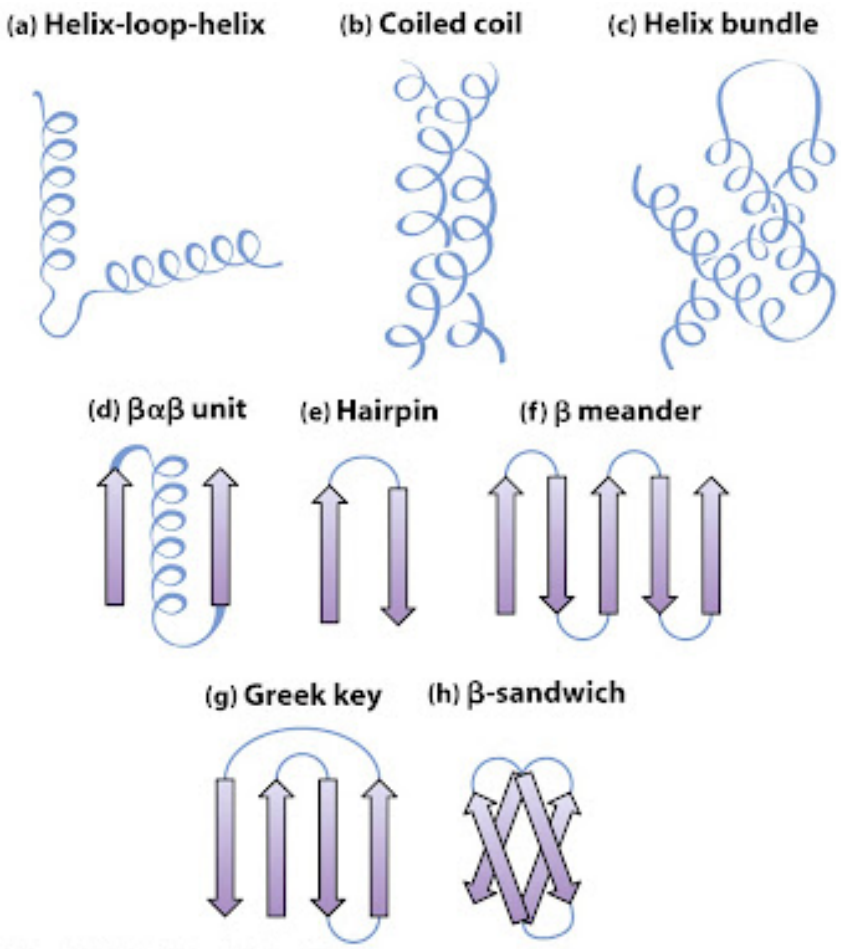
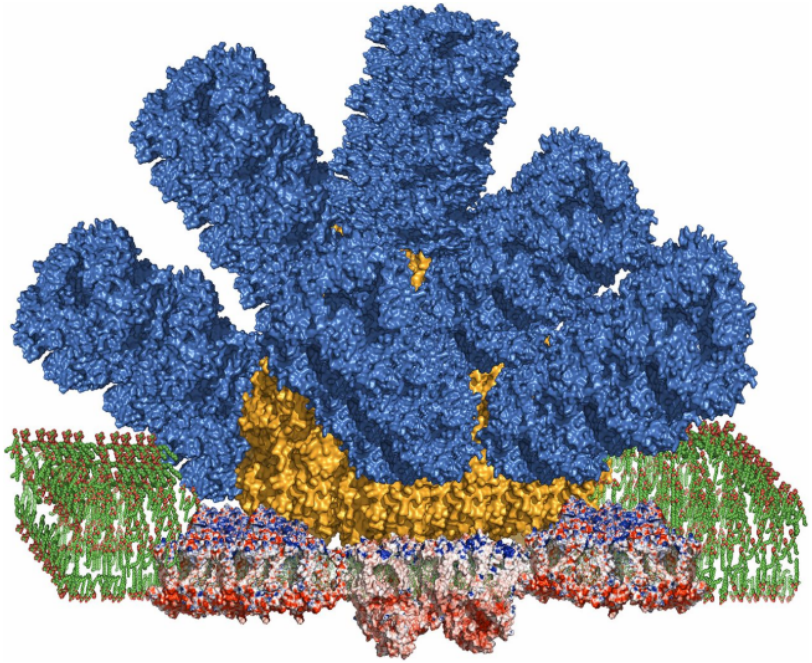


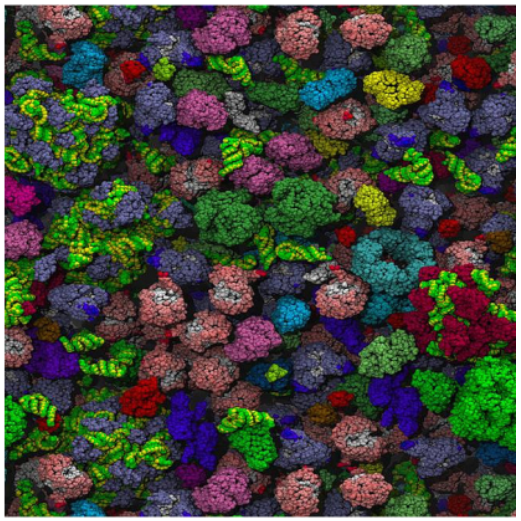
Figure 4-19 Principles of Biochemistry, 4/e
© 2006 Pearson Prentice Hall, Inc.

Numerous computational methods for subdomain structures
e.g. Protein Peeling, TEF2

Structural complexes

- Assembly of multiple subunits (protein and/or RNA) into homo- or hetero-oligomeric structures
- Not to be confused with non-specific interactions





Macromolecular docking

- Predict interaction between two macromolecules
- For two known partners: predict molecular details
- Criteria: shape complementarity, calculated binding free energy, co-evolution
- e.g. InterEvDock3, HADDOCK

Note

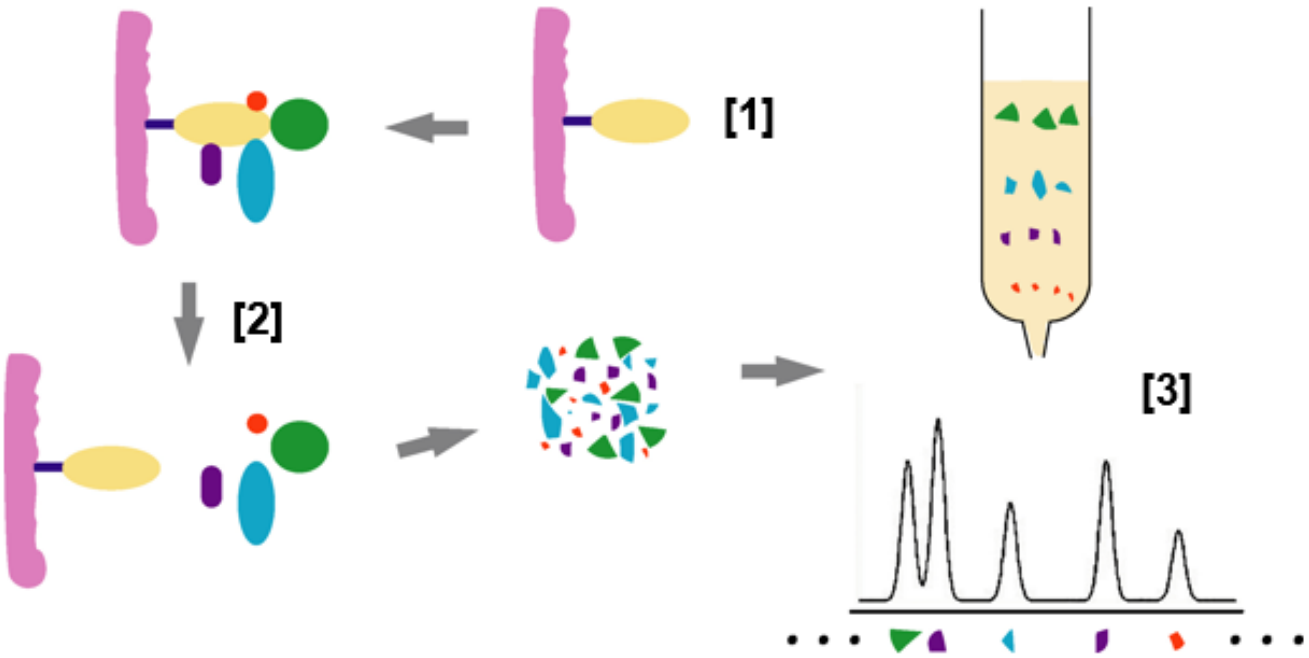
For protein-peptide interactions, drug design docking methods can be used: namely, AutoDock, Vina, and Smina

Experimental techniques: biochemical methods

Affinity Purification Mass Spectrometry (AP-MS)

Capture of the "prey" proteins with a tagged "bait" protein

→ Immunoprecipitation or affinity column, then elution and MS

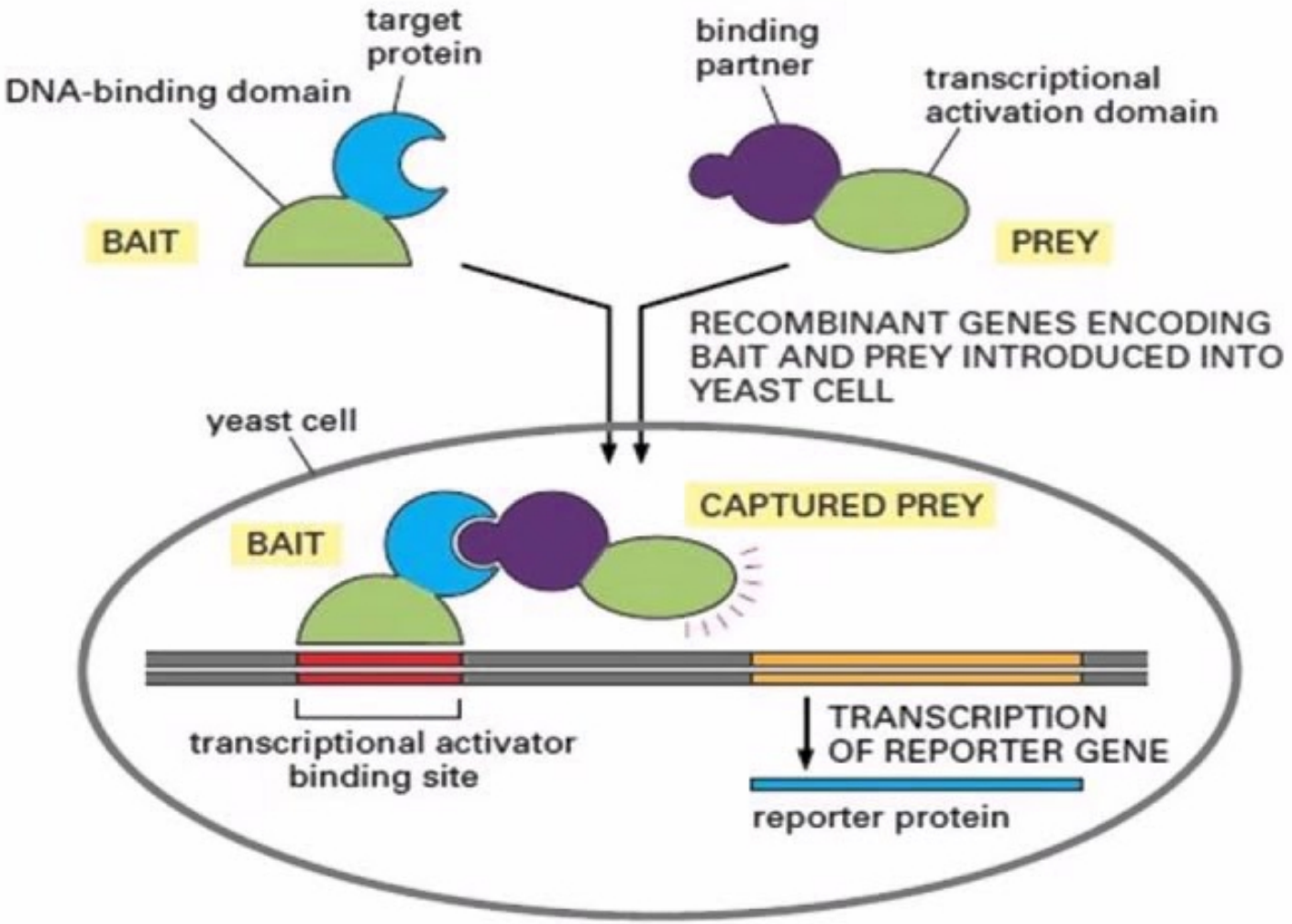


Can detect transient interactions (e.g. signaling pathways)

May include false positives

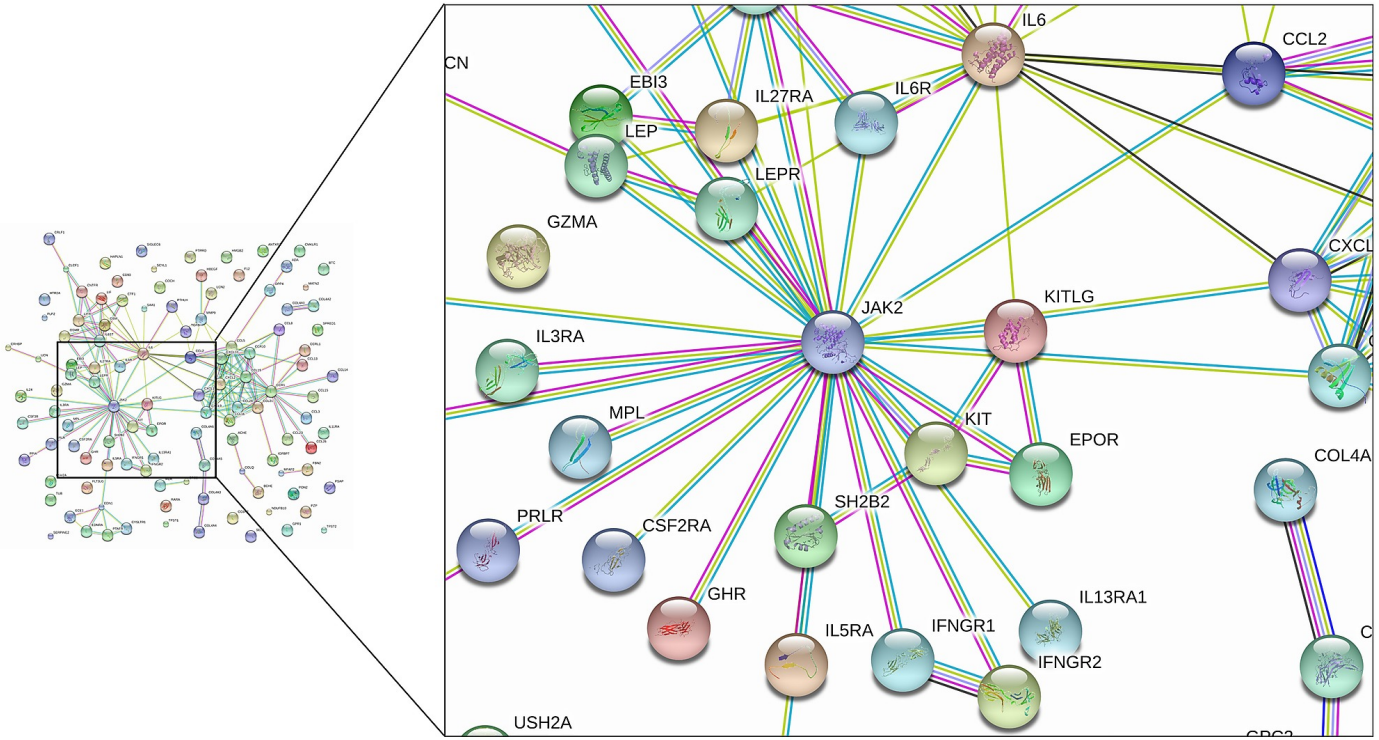
Experimental techniques: genetic approaches

Two-hybrid system (→ pairwise interactions)



Network biology

- "Interactomics" data stored in databases: e.g. BioGRID, IntAct, PDB (complexes)
- Data integration and visualization servers: e.g. STRING, GeneMANIA



Representation

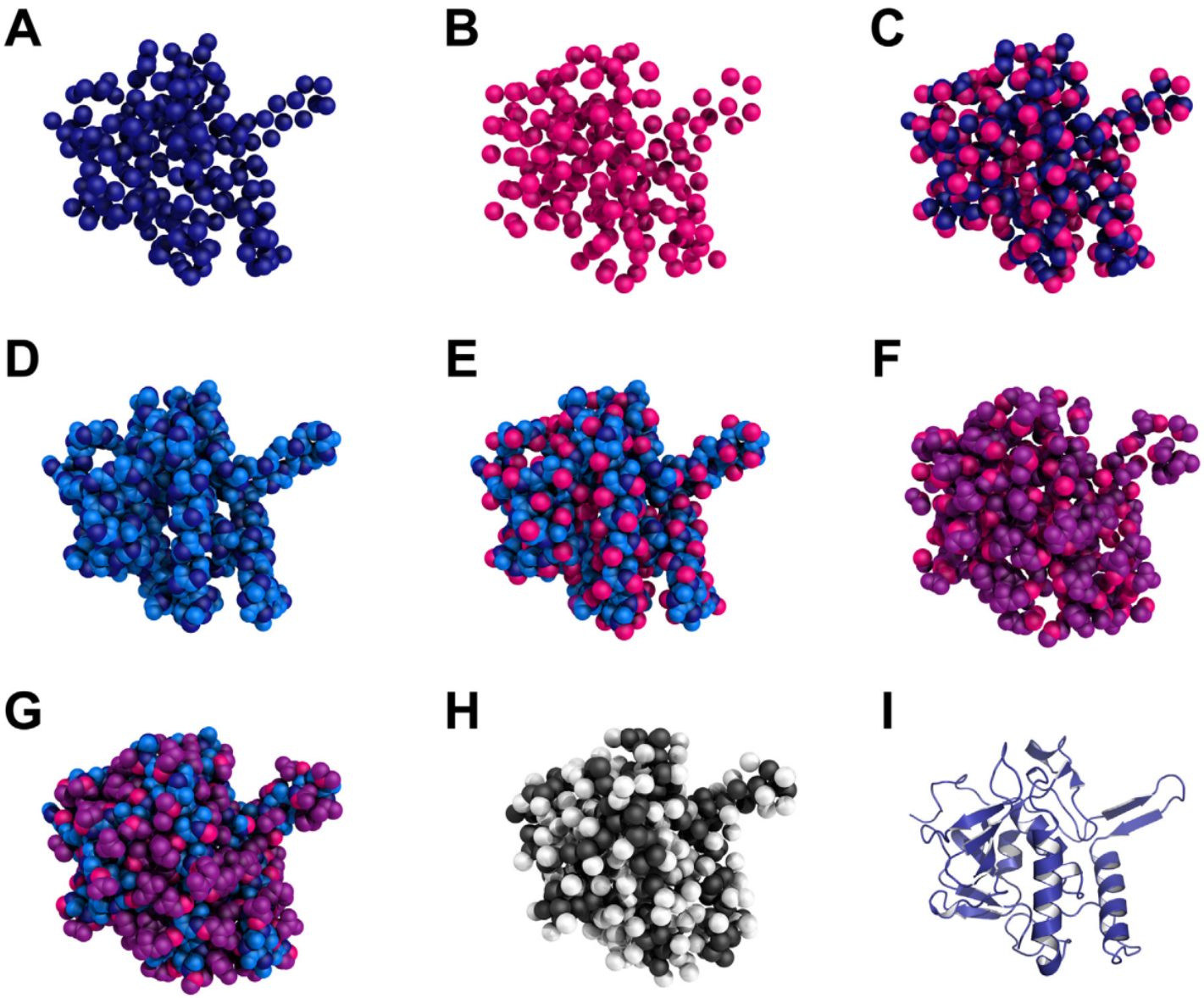
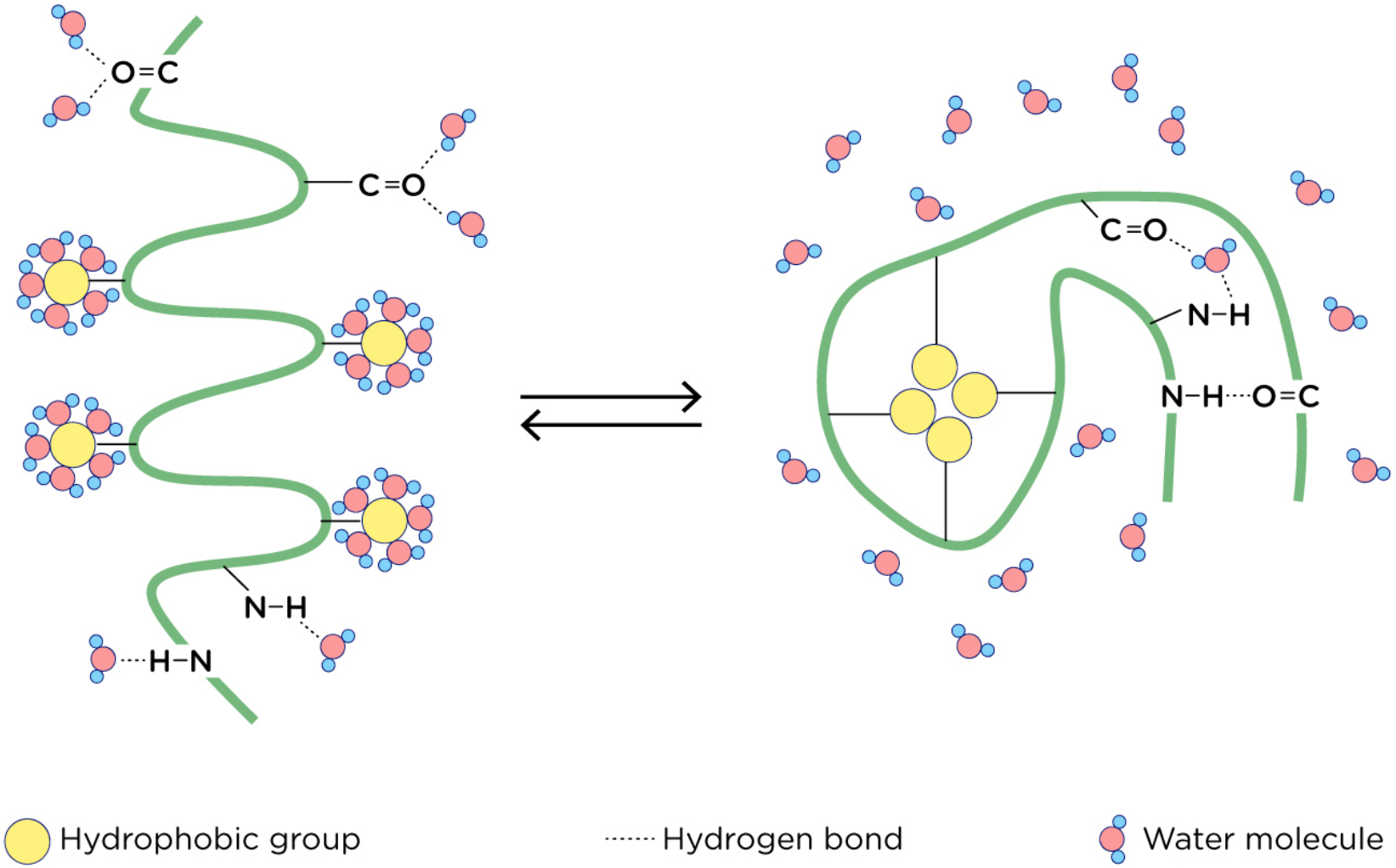


Fig. 1. Structure of the peptidoglycan hydrolase RipA, from *Mycobacterium tuberculosis* H37Rv (PDB code: 3ne0), in increasing-resolution order: (A) Cα, (B) Cβ, (C) Cα + Cβ, (D) backbone, (E) backbone + Cβ, (F) side-chains, (G) all-atom. The MARTINI coarse-grained model (H) is represented with backbone beads in black and side-chain beads in white. (I) Cartoon representation of the protein.

Structural bioinformatics: table of contents

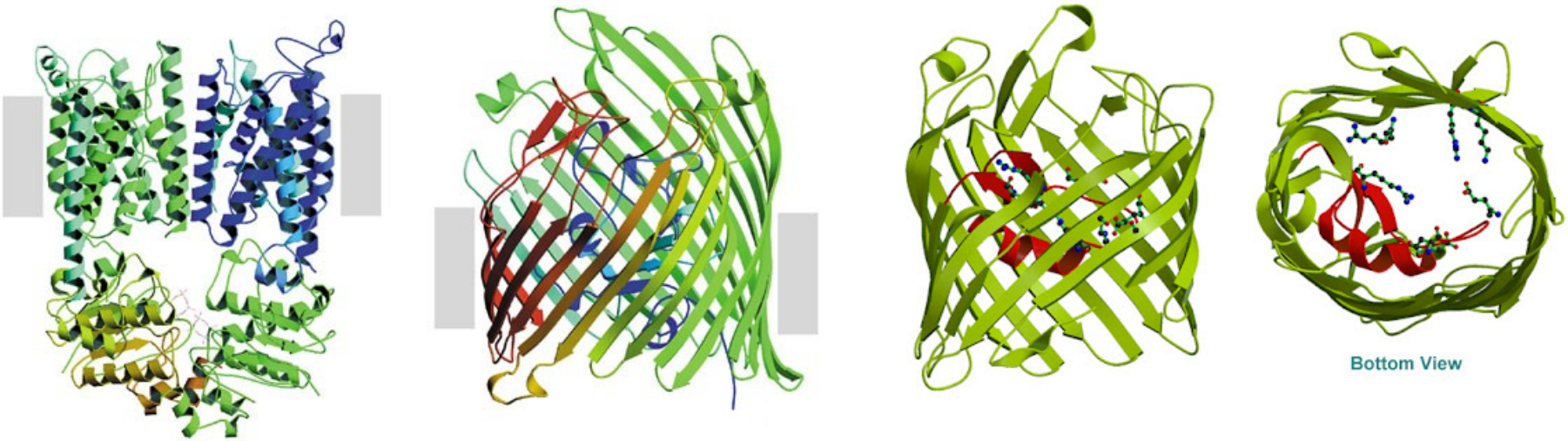
1. Introduction
   - Macromolecules
   - Structure and function
   - Molecular design
2. Experimental determination
   - X-ray, NMR, cryo-EM
   - X-ray, NMR, cryo-EM
   - The Protein Data Bank
3. Protein structure: analysis
   - Primary: sequence
   - Secondary: local conformations
   - Tertiary: global conformation
   - Intermediary
   - Quaternary: oligomery
   - Representation
4. Protein folding
   - Thermodynamics
   - Kinetics
5. Structure prediction
   - Preamble
   - Sampling
   - Scoring
   - Performance
   - Applications
6. Molecular flexibility
   - Molecular dynamics simulations
   - Advanced techniques
   - Normal mode analysis
7. Nucleic acids structures
   - DNA
   - RNA

Entropic force



Entropic force

- Water-soluble proteins → globular shape
- Transmembrane proteins: hydrophobic environment of the lipid bilayer
- ↪ Orthogonal to the membrane
- ↪ Receptors, channels



Besides hydrophobic interactions: van der Waals force, hydrogen bonds, ionic bonding (charged residues), covalent bonding (cysteines)

Anfinsen's dogma (Nobel Prize in Chemistry,1972)

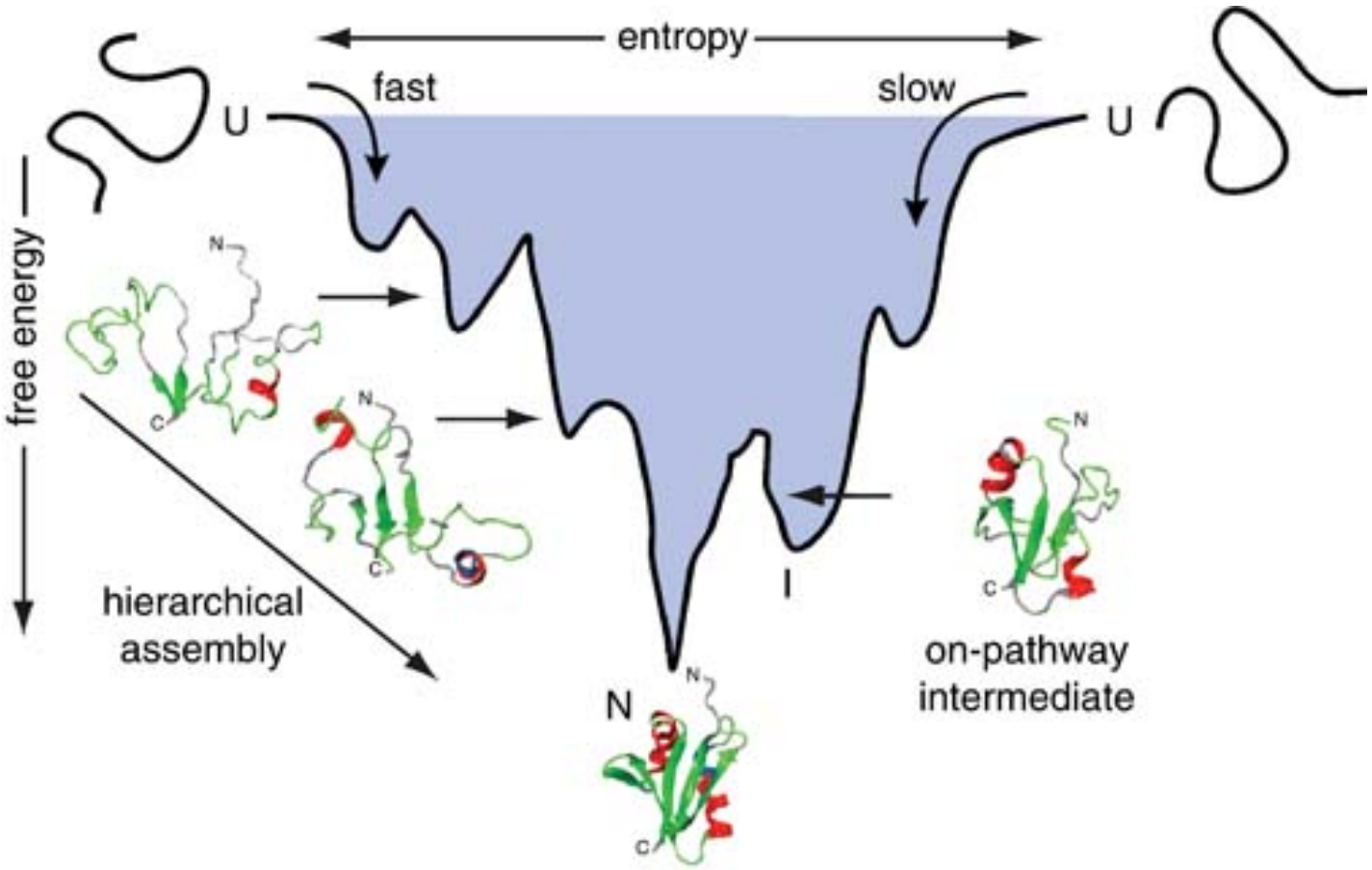
In the native conditions, the polypeptide chain folds into its native conformation **spontaneously**. Thus, of all possible conformations, the native protein structure has the lowest Gibbs free energy.
Found by tracking enzyme activity in either native or denaturing conditions

Limitations

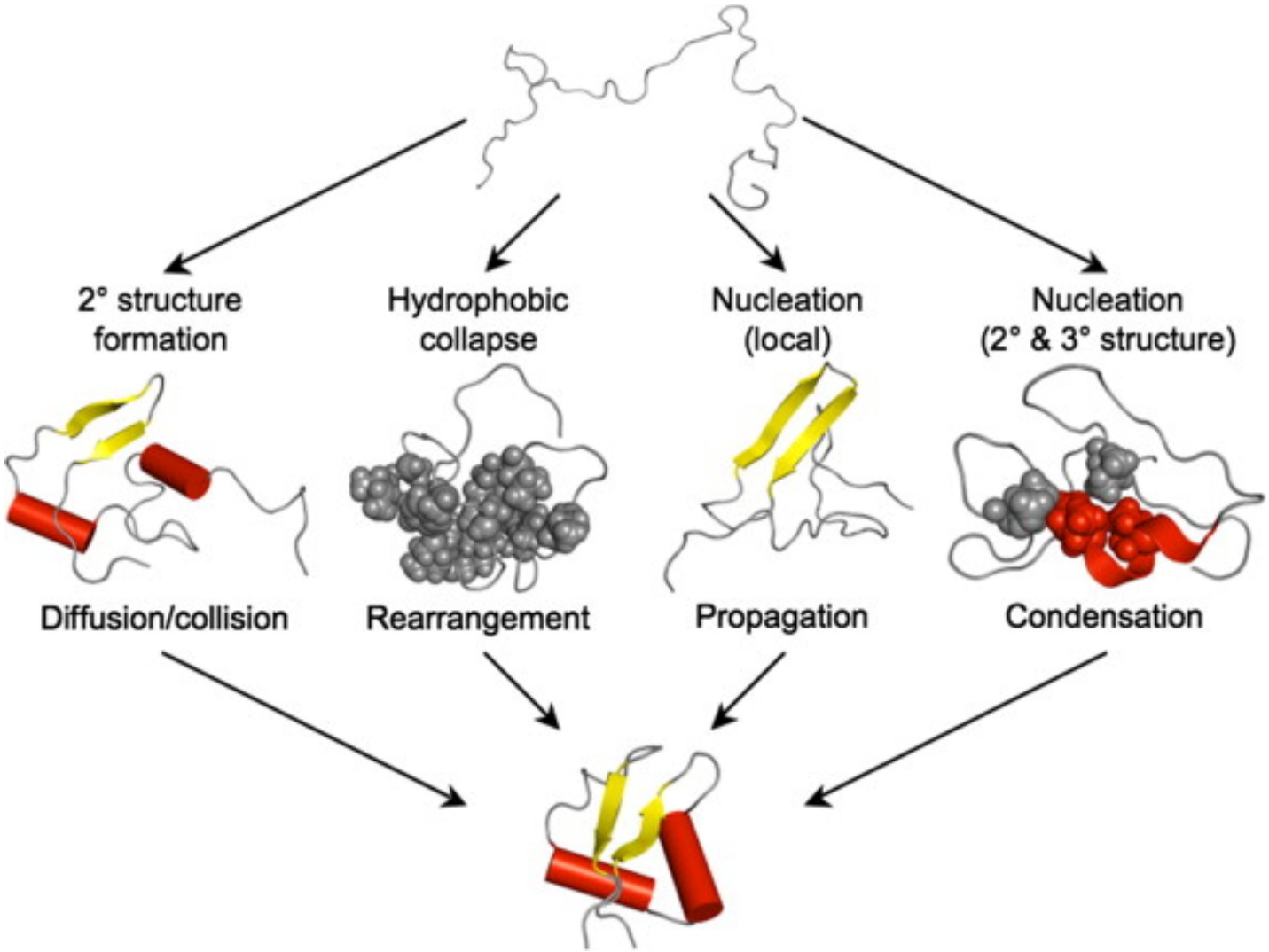
- Protein structures are flexible objects
- Chaperone are sometimes required
- Intrinsically disordered proteins (IDP)

Levinthal's paradox and folding funnels

An averaged size protein of 100 residues can have 10^{30} conformations. By a random exploration of this conformational space, it would take 10^{77} s to find the native structure.



Protein folding models: not mutually exclusive



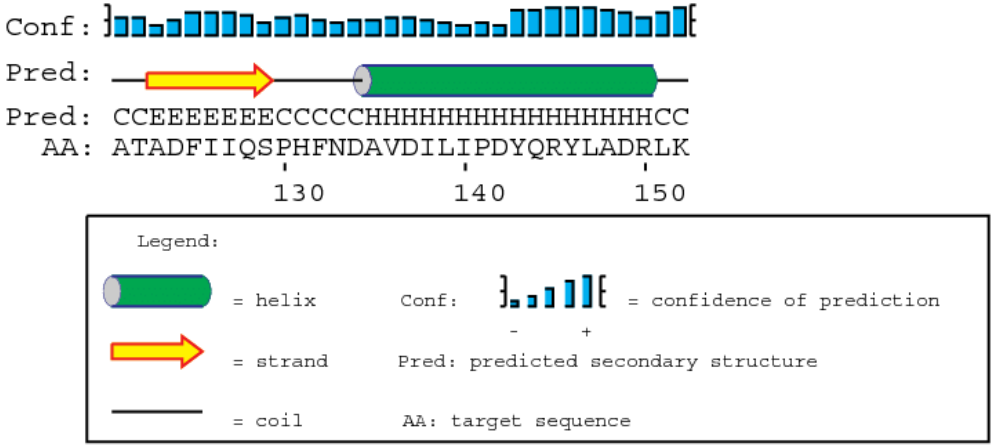
Structural bioinformatics: table of contents

1. Introduction
   - Macromolecules
   - Structure and function
   - Molecular design
2. Experimental determination
   - X-ray, NMR, cryo-EM
   - X-ray, NMR, cryo-EM
   - The Protein Data Bank
3. Protein structure: analysis
   - Primary: sequence
   - Secondary: local conformations
   - Tertiary: global conformation
   - Intermediary
   - Quaternary: oligomery
   - Representation
4. Protein folding
   - Thermodynamics
   - Kinetics
5. Structure prediction
   - Preamble
   - Sampling
   - Scoring
   - Performance
   - Applications
6. Molecular flexibility
   - Molecular dynamics simulations
   - Advanced techniques
   - Normal mode analysis
7. Nucleic acids structures
   - DNA
   - RNA

Secondary structure prediction

PSIPRED (Jones, 1999)

- Q3 = 90%, i.e. rate of correct 3-state prediction (random Q3 = 33.3%)
- Based on evolution (i.e. multiple sequence alignments)
- Performance has increased along with the number of sequences in the databases
- Note: Q3's theoretical upper limit due to uncertainty on some residues (local flexibility)



Tertiary structure: the protein folding problem

- A major scientific question[1]
- **NP-hardness**[2]
- For an average size protein of 100 residues,
  find the structure with the lowest Gibbs free energy,
  on a total of $\sim 10^{30}$ conformations

Heuristic: sampling and scoring

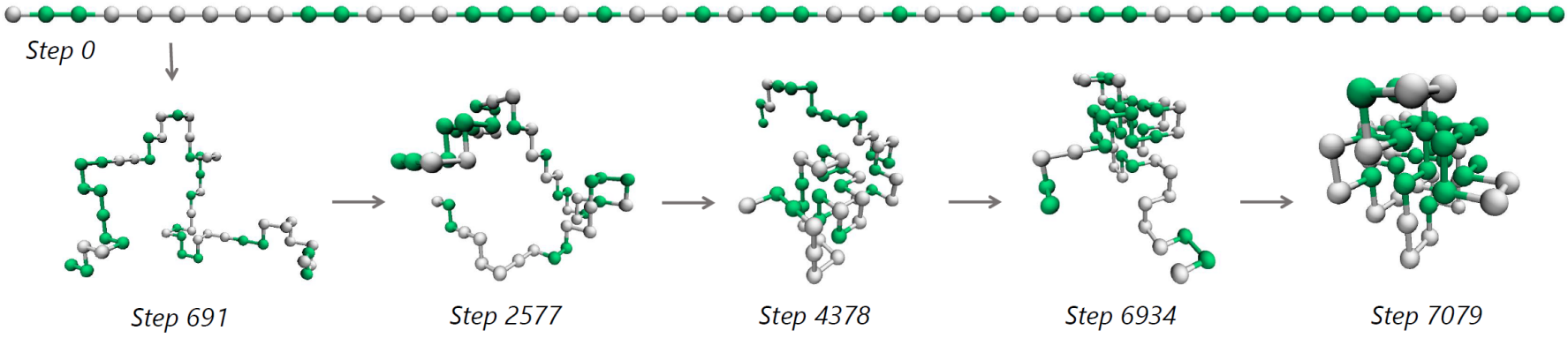
- Non-exhaustive algorithmic search of the conformational space
- Optimization of an objective function

[1] "What Don't We Know?", *Science*, 2005
[2] Unger and Moult, 1993

Ab initio modeling

Exploration of a **simplified conformational space**

- 3D → 1D representation, by using a structural alphabet
- Lattice protein, e.g. hydrophobic-polar (HP) protein folding model



Comparative modeling

Detection of experimental structures used as **templates** compatible with the sequence, by using alignment methods

- BLAST to detect homolog proteins (seq. id. >30%)
- If no homolog: detection of distant homologies
  → **fold recognition** methods
  - Alignments of PSSMs (position-specific scoring matrices, also called profiles); e.g. PSI-BLAST, ORION
  - Alignments of HMM-based (hidden markov model) profiles; e.g. HHsearch
  - Sequence-to-structure alignments (1D–3D methods)

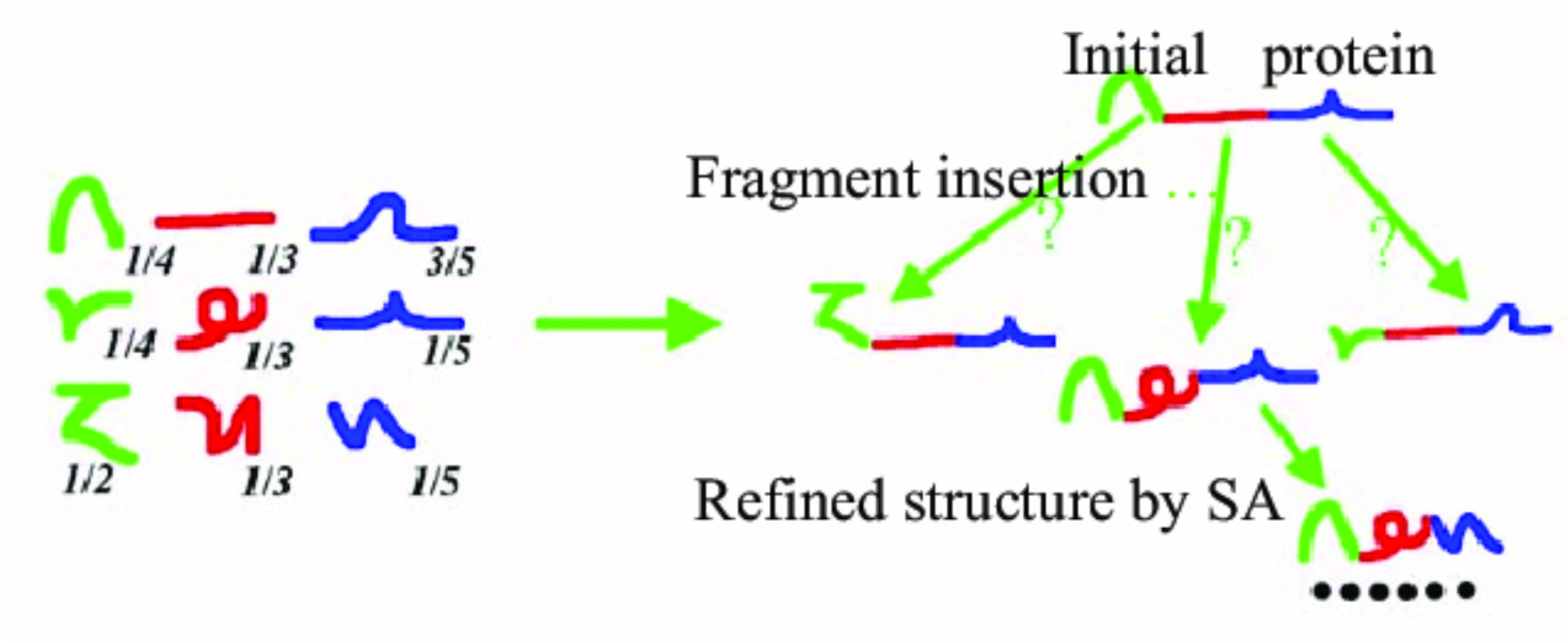
Note

In genomics, fold recognition methods can be used to annotate genes of unknown function.

Once a template found: generation of the model by satisfaction of spatial restraints; e.g. MODELLER (most cited method)

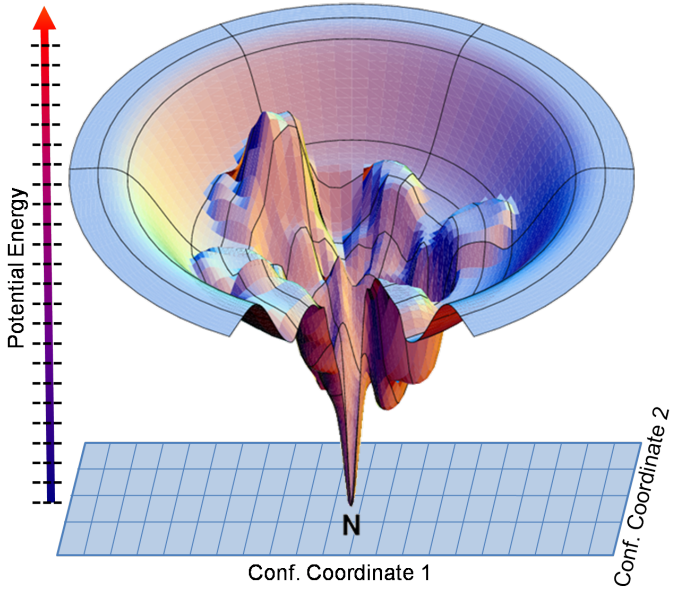
De novo modeling

- In between ab initio and comparative modeling
- Based on protein **fragment assembly**
- Most cited method: Rosetta (Baker, 2005)



Objective function

Must approximate Gibbs free energy → minimization



Assess the quality of the predicted structures

- At each step of the exploration, to guide the sampling algorithm
- At the end of the sampling, if several solutions are produced

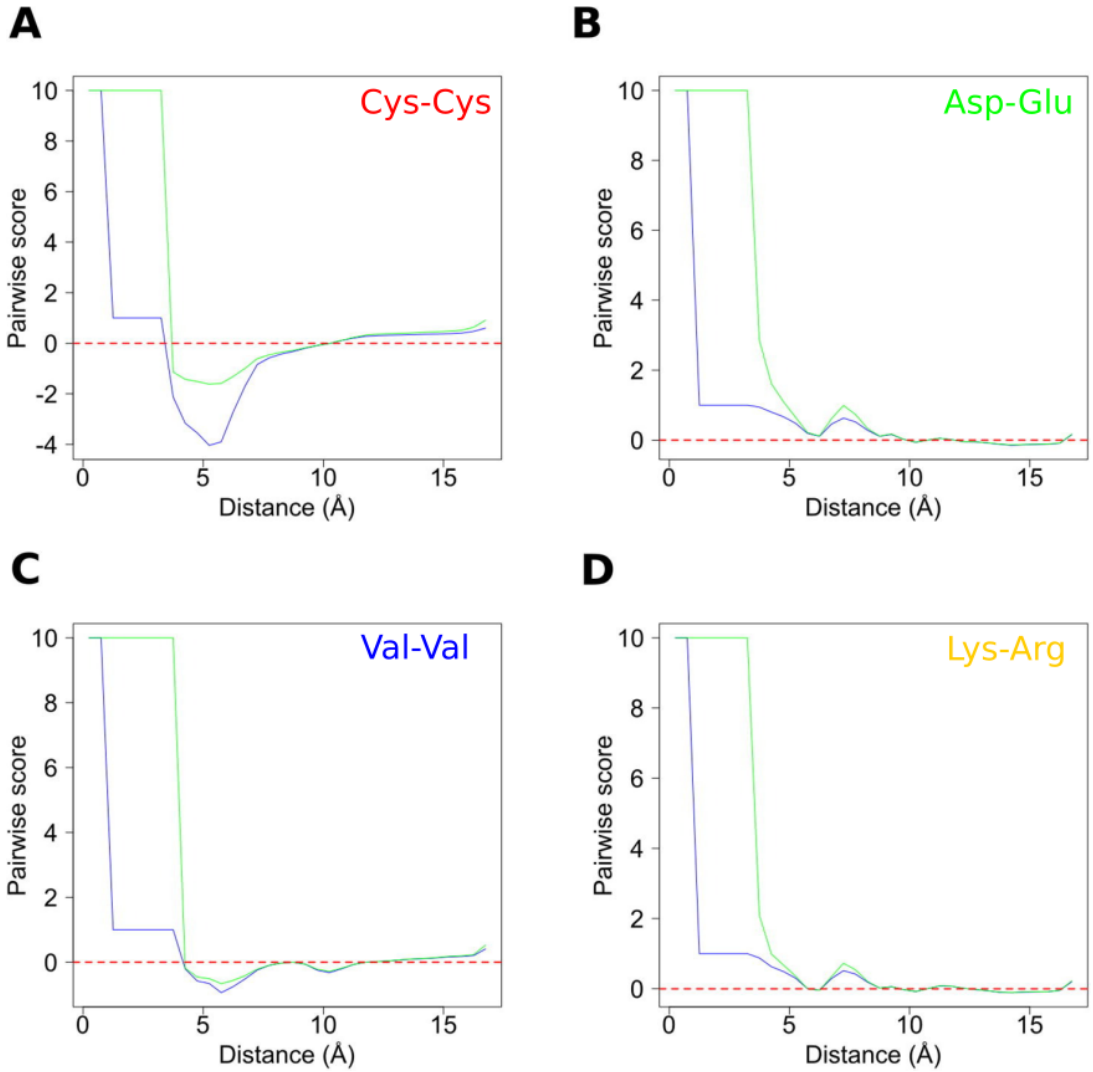
Physical functions

- Based on physical potentials (e.g. Lennard-Jones, Coulomb, Morse)
- Account for the entropic contribution to the protein folding in a way that is either too approximate or computationally inefficient
- The same goes for quantum effects
- $\rightarrow$ Physical energy functions rarely meet success :'(

Statistical (or "knowledge-based") functions

- Built upon statistics computed on a set of protein structures that have been experimentally determined
- Any structural feature can be used: e.g. interatomic distances, torsion angle values, solvent accessibility
- Can be composite, combining several features or including physical potentials
- $\hookrightarrow$ Structural descriptors combined by using machine learning techniques
- $\hookrightarrow$ Used in Google DeepMind's AlphaFold :-)

Statistical potentials: examples



Statistical potentials: Sippl's formalism (1990)

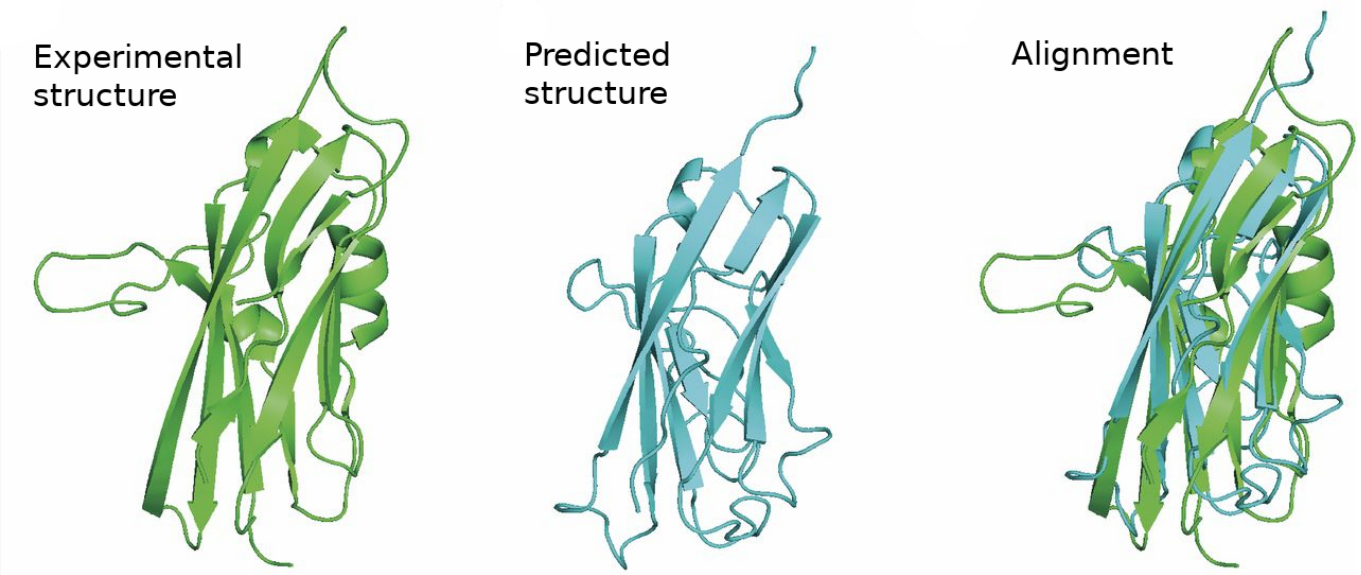
two distance distributions were derived from a set of native structures: one is conditioned on the residue types of the atom pair (conditional model, M_1), whereas the other is not (background model, M_2, the so-called "reference state"). The score of a protein conformation is then calculated as the negative log-likelihood ratio of the distances under the statistical models M_1 and M_2, summed over all pairs of atoms i, j:

$$score = -\sum_{i,j,i\neq j} \log\left(\frac{P\left(d_{ij}|M_1\right)}{P\left(d_{ij}|M_2\right)}\right) \tag{1}$$

where $P(d_{ij}|M_1)$ and $P(d_{ij}|M_2)$ are the observed probabilities for two atoms i and j to be separated by a distance d, with and without consideration of the atom types, respectively. Often improperly labeled "energy", the score actually measures the relative support of the statistical models M_1 and M_2, by the distances observed in the protein conformation evaluated. Thus, the more negative the score, the more the assessed structure supports a native distribution of interatomic distances (M_1), over a random one (M_2).

Prediction: distance from the ground truth

- Similarity between predicted and experimental structures
- Requires (i) structural alignment algorithm and (ii) metrics
- ↪ TM-align and TM-score (Zhang lab, 2005) are the most popular
- RMSD (Root-Mean-Square Deviation) is the historical measure; however, it overestimates the structural difference due to the loop (flexible) regions



CASP (critical assessment of structure prediction)

- Competition in modeling protein structure
- Evry two years, since 1994
- Global distance test (GDT): measure of accuracy (close to TM-score)
- Latest edition (CASP14) won by AlphaFold2, with a median score of 92.4 (out of 100)
- $\hookrightarrow$ Protein folding problem solved? (Folding pathways)

Statistical potentials have continuously proven useful in numerous applications, which all boil down to predicting the lowest-energy conformations

- Molecular docking, which also consists of a sampling and a scoring procedures
- Prediction of protein stability upon point mutations; e.g. FoldX
- Prediction of protein solubility and aggregation

Structural bioinformatics: table of contents

1. Introduction
   - Macromolecules
   - Structure and function
   - Molecular design
2. Experimental determination
   - X-ray, NMR, cryo-EM
   - X-ray, NMR, cryo-EM
   - The Protein Data Bank
3. Protein structure: analysis
   - Primary: sequence
   - Secondary: local conformations
   - Tertiary: global conformation
   - Intermediary
   - Quaternary: oligomery
   - Representation
4. Protein folding
   - Thermodynamics
   - Kinetics
5. Structure prediction
   - Preamble
   - Sampling
   - Scoring
   - Performance
   - Applications
6. Molecular flexibility
   - Molecular dynamics simulations
   - Advanced techniques
   - Normal mode analysis
7. Nucleic acids structures
   - DNA
   - RNA

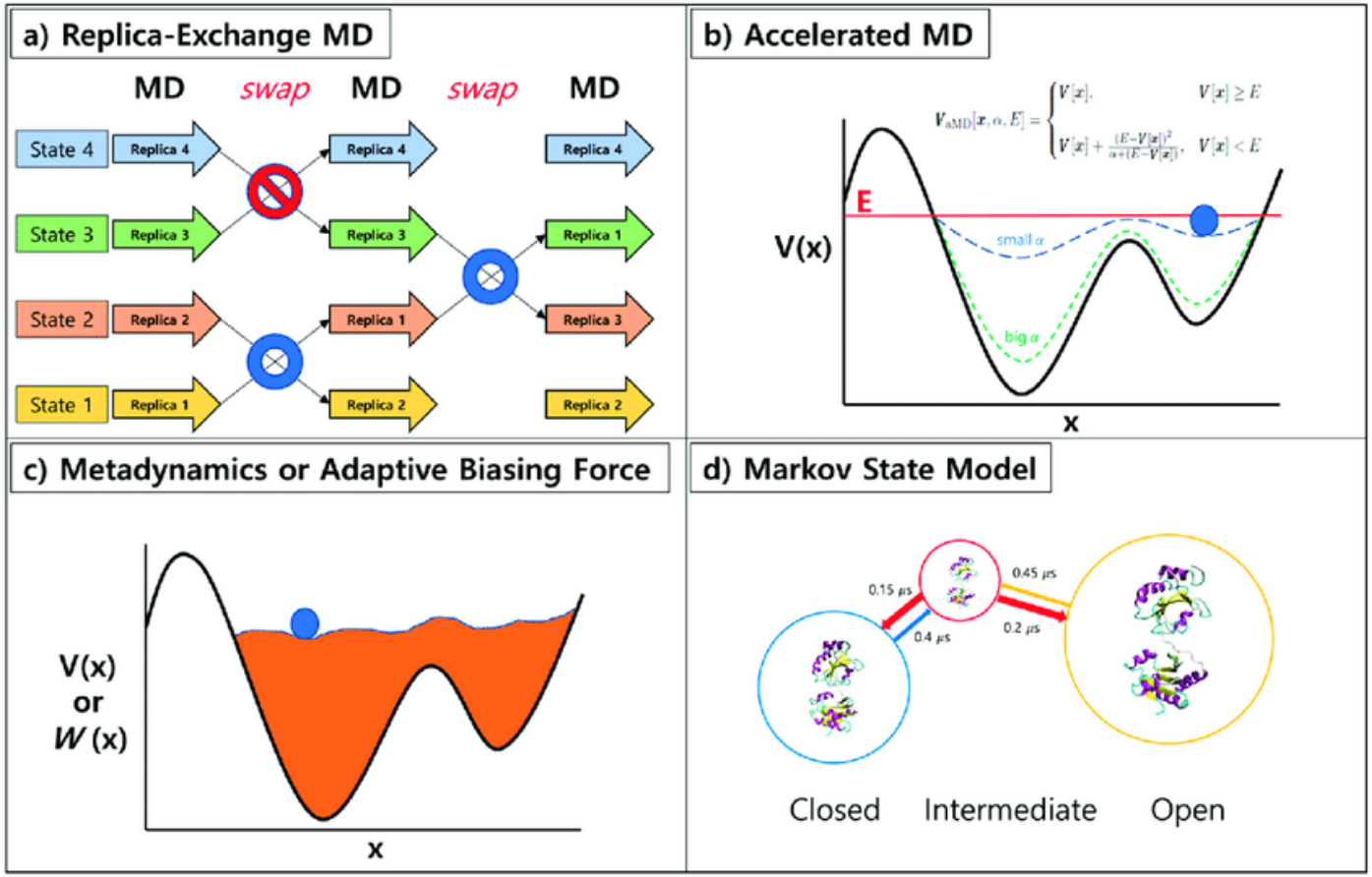
- Trajectories of atoms and molecules are determined by numerically solving Newton's equations of motion
- Numerous variations of the two main force fields, CHARMM (Chemistry at HARvard Macromolecular Mechanics) and AMBER (Assisted Model Building with Energy Refinement)
- GROMACS (GROningen MAchine for Chemical Simulations) is the most popular package; it is free and open-source
- Implicit vs explicit solvent

Actual use

The approximations made (due to the computational cost) are too numerous to make MD simulations a predictive approach. However, when knowing the initial and final states of the system (from experimental results), MD simulations may sometimes **provide insights into the molecular mechanisms**.

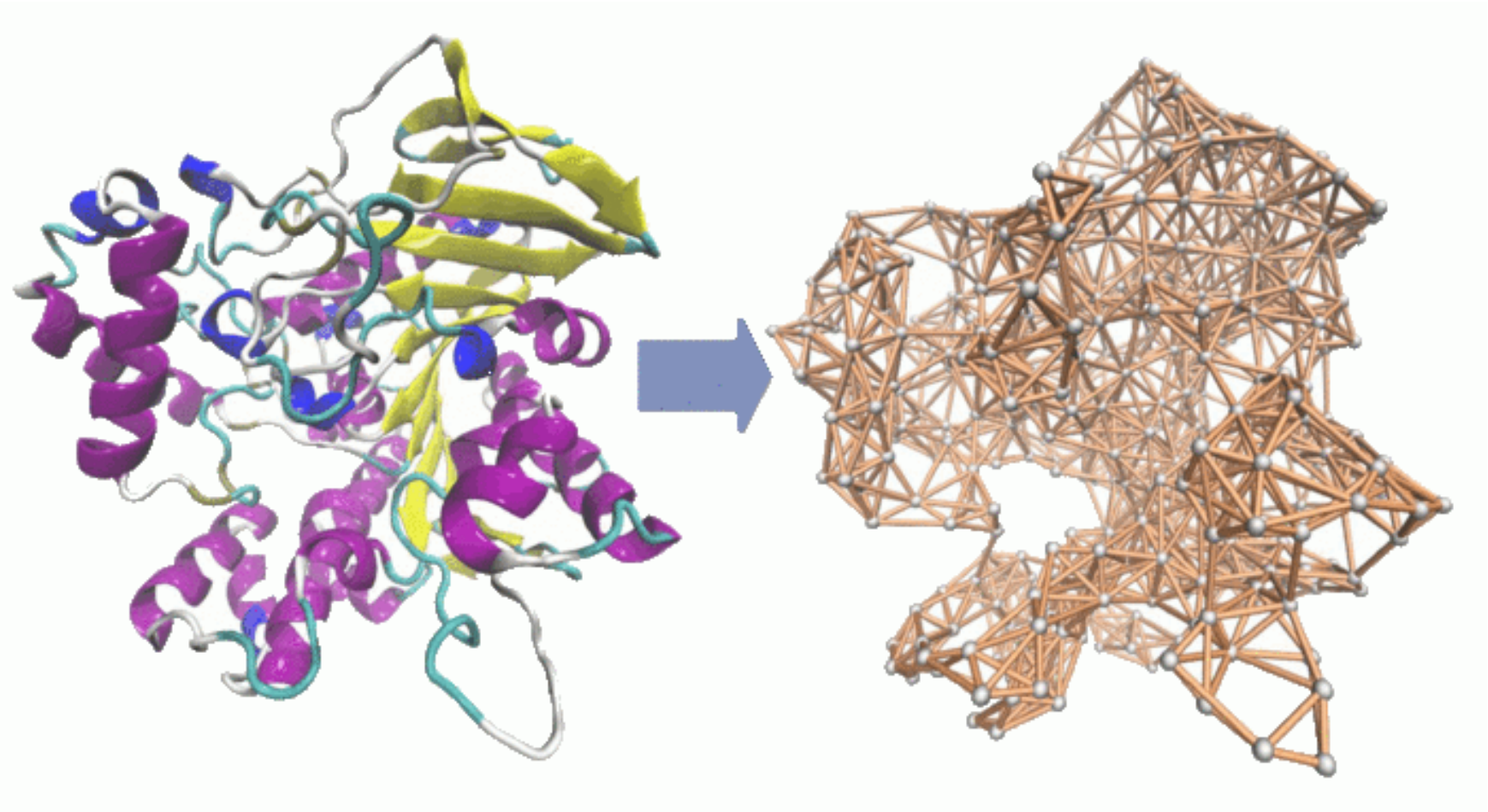
Estimation of the free energy by enhanced sampling

- Cross energy barriers to explore more configurations
- Computationally expensive: small systems or fast molecular motions



Normal mode analysis

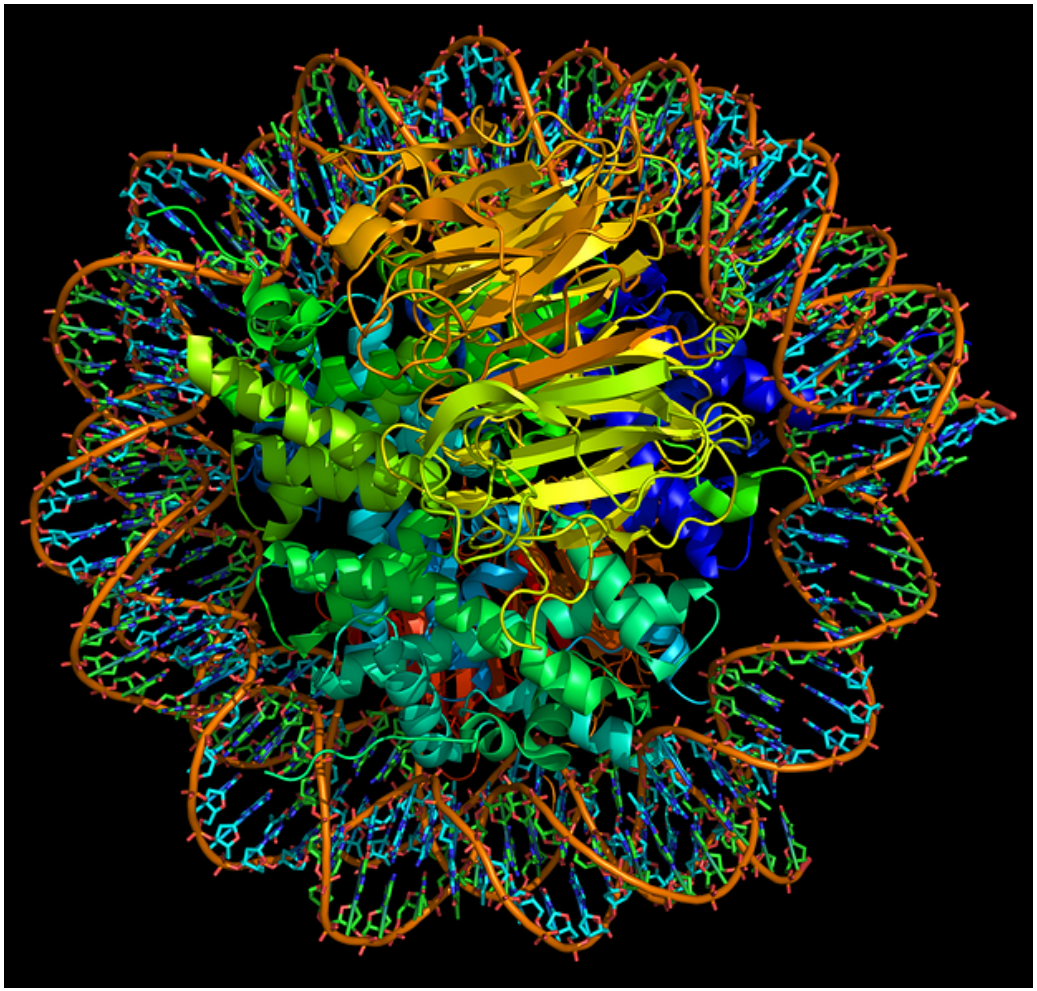
- Structure represented as an elastic network model (ENM)
- Computationally efficient: large systems or slow movements



Structural bioinformatics: table of contents

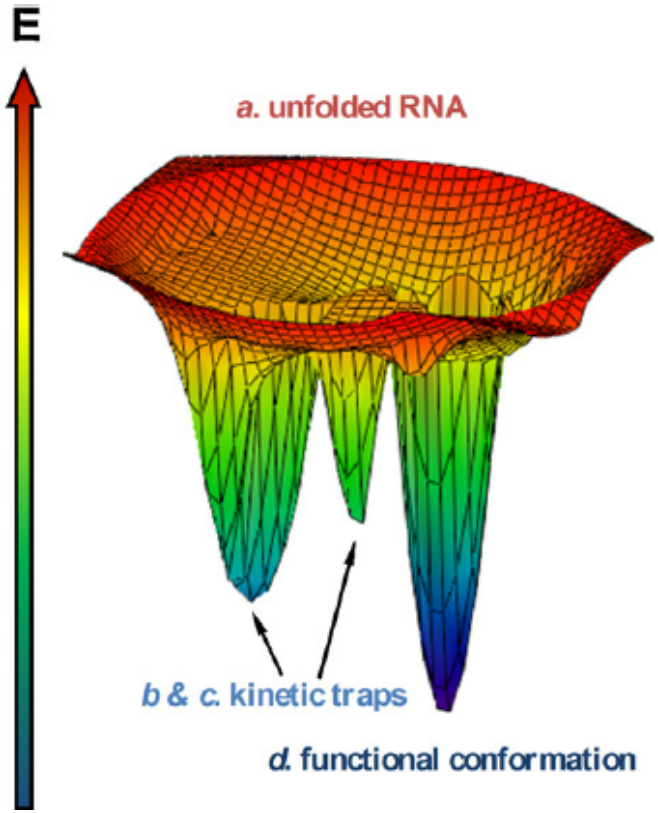
1. Introduction
   - Macromolecules
   - Structure and function
   - Molecular design
2. Experimental determination
   - X-ray, NMR, cryo-EM
   - X-ray, NMR, cryo-EM
   - The Protein Data Bank
3. Protein structure: analysis
   - Primary: sequence
   - Secondary: local conformations
   - Tertiary: global conformation
   - Intermediary
   - Quaternary: oligomery
   - Representation
4. Protein folding
   - Thermodynamics
   - Kinetics
5. Structure prediction
   - Preamble
   - Sampling
   - Scoring
   - Performance
   - Applications
6. Molecular flexibility
   - Molecular dynamics simulations
   - Advanced techniques
   - Normal mode analysis
7. Nucleic acids structures
   - DNA
   - RNA

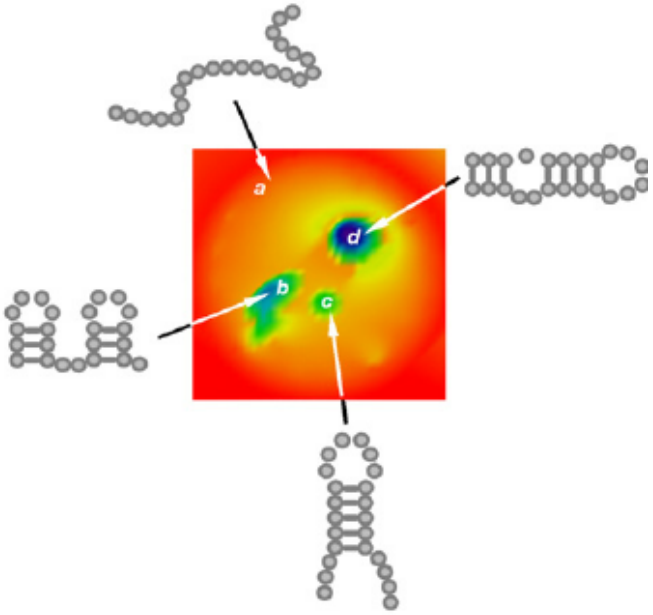
Most structural bioinformatics researches focus on complex formation and dynamics with proteins, e.g. transcription factors, regulators, histones



RNA

- Secondary structure: numerous prediction tools
  `https://evryrna.ibisc.univ-evry.fr/`
- Tertiary structure: still challenging
  - Experimental data limited in both quantity and quality
  - Base interactions rather than entropic force
  - Rugged energy landscape





The end

Thank you for your attention