Understanding biomolecular interactions and simulations

Frederic.Cazals@inria.fr

Biomolecular recognition

PART 1: Introduction to Protein Science
PART 2: Biomolecular recognition

Biomolecular recognition

Computational Structural Biology: what is a protein?

Protein Structure Resolution

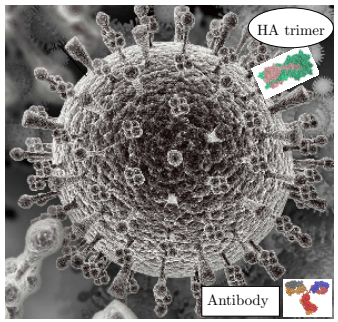
The importance of dynamics

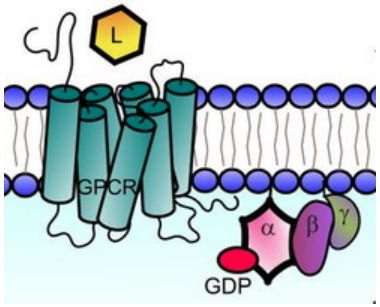
Protein functions: examples

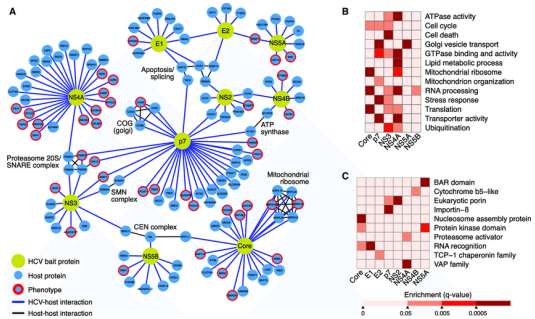
Computational Structural Biology: challenges

Computational structural biology: perspective

towards a new era in biology, medicine, material sciences







- Biology: help unveil all core mechanisms of life at the atomic level
  - metabolism, immune system, genetic information processing, cognition, . . .
- Medicine: foster the design of novel therapeutics
  - optimizing specific molecules e.g. antibodies
  - discovering novel drug targets / transient conformations [1]
  - bridging the gap systems biology – structural biology
- Material sciences at large: atomic level design and engineering

[1] (2017) 35% of FDA approved drugs: 108 GPCRs

Computational Structural Biology

- Goals: unveil the *structure-dynamics-function* conundrum for biomolecules (proteins and nucleic acids)
- Methods: biophysics (crystallography, NMR, cryo-microscopy) + modeling
- Nobel prizes as of 01/2019 [2]: related to molecular/structural biology
  - Chemistry or Physiology-medicine for structures and mechanisms: 64
  - Chemistry or Physics for Methods : 11
  - Chemistry 2013: Levitt, Karplus, Warshel for *the development of multiscale models for complex chemical systems*
- An extraordinary field
  - Technology driven: novel biophysical experiments,
  - Reveals the molecular foundations of biology and medicine,
  - Raises open mathematical / computational questions.

[2] https://pdb101.rcsb.org/learn/flyers-posters-and-other-resources/other-resource/structural-biology-and-nobel-prizes

Methods: molecular simulation

The Nobel Prize in Chemistry 2013
Martin Karplus, Michael Levitt, Arieh Warshel

The Nobel Prize in Chemistry 2013

© Harvard University
Martin Karplus

Photo: © S. Fisch
Michael Levitt

Photo: Wikimedia Commons
Arieh Warshel

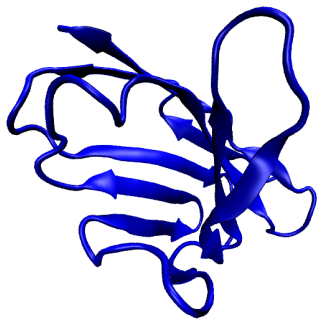
The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel *"for the development of multiscale models for complex chemical systems"*.

What is a protein?

- Primary structure: sequence of amino acids

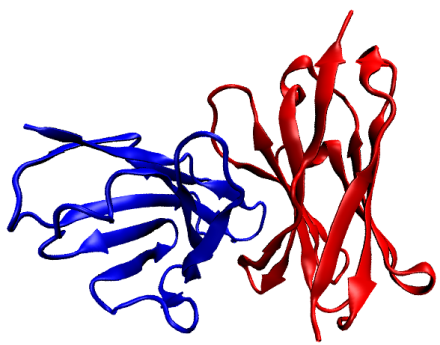
```
P69905 (HBB_HUMAN) MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GS 53
P68871 (HBB_HUMAN) MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN 58
P02144 (MYG_HUMAN) -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKAS 59
                    : *:  :   *   ****         * * *::   * *   *  *  *
```

- Polypeptide chain



- Protein - protein complex

- Heterodimeric protein

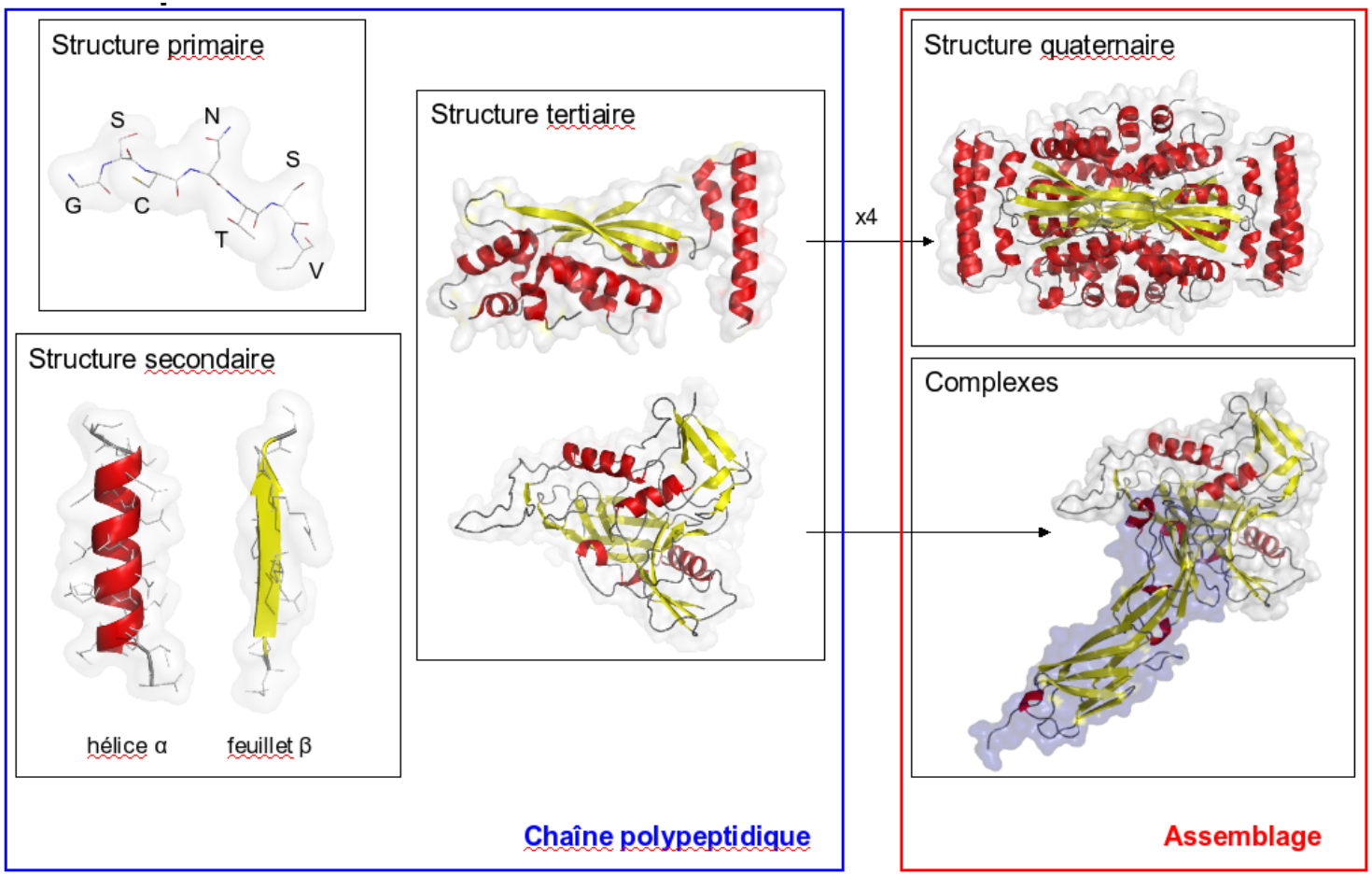


- Nb: median number of a.a. in a chain: ~ 400

What is a protein?

Importance of non-covalent interactions

- Primary to quaternary structure



- Grand Challenges: folding and docking ... related businesses!

The Folding Problem

▷ C. Anfinsen's experiment (Nobel 1972), the sequence determines the structure:

- Identification of the native state —minimum of free energy
- Determination of the folding pathways

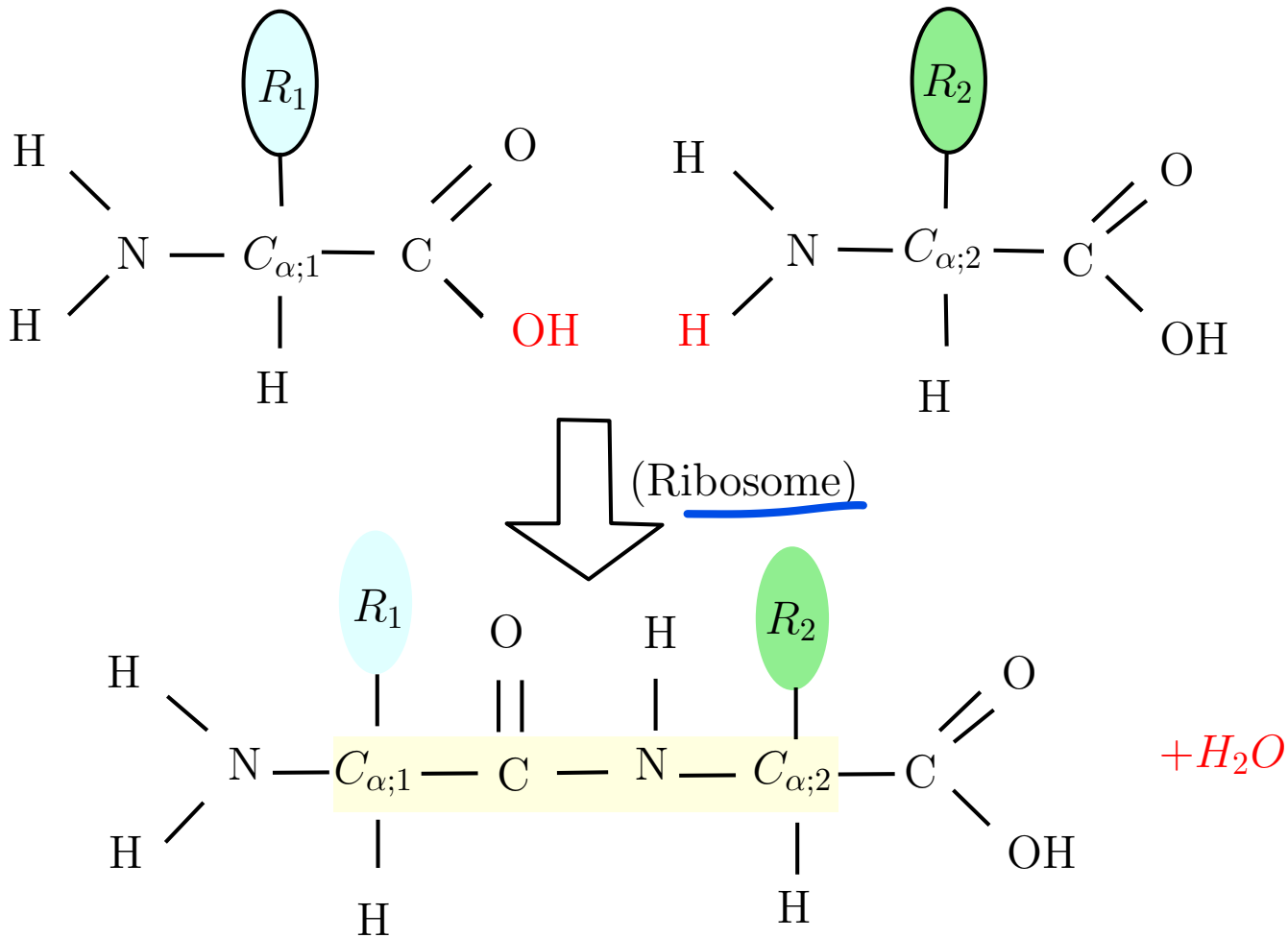
▷ Levinthal's paradox. n amino-acids with r conformations: r^n states.
– Random searches would require astronomical time
– Nature requires from milliseconds (helical prot.) to (tens) seconds (complex geom.)
▷ Other systems: clusters (water molecules, rare gases), crystallization etc

Amino acids and the peptide bond

- ▷ Natural amino acids and their side chains

Nb: 0 to 10 heavy atoms per side chain

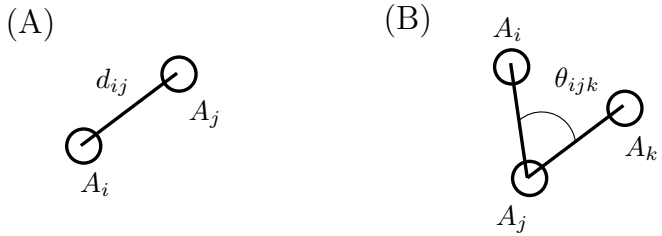
- ▷ Peptide bond synthesis:



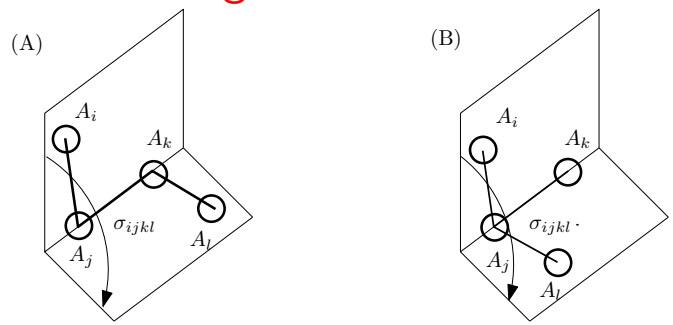
Geometric models: Cartesian and internal coordinates

▷ Cartesian versus internal coordinates: $\{x_i y_i z_i\}_i$ versus $\{d_{ij}, \theta_{ijk}, \sigma_{ijkl}\}$

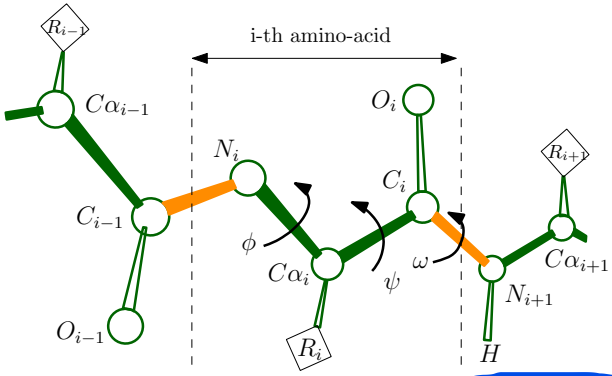
▷ Bond length and valence angle



▷ Dihedral angles



▷ Protein backbone

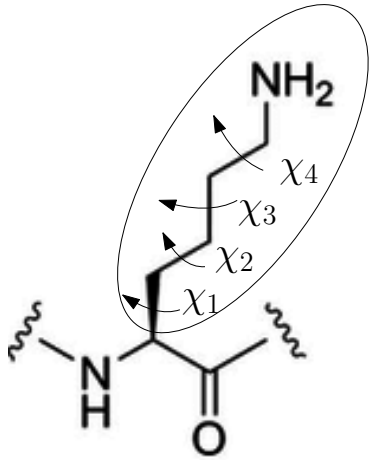


Ramachandran diagram per a.a. type:

- bivariate distribution for (ϕ, ψ)

▷ Side chain: 20 natural amino acids

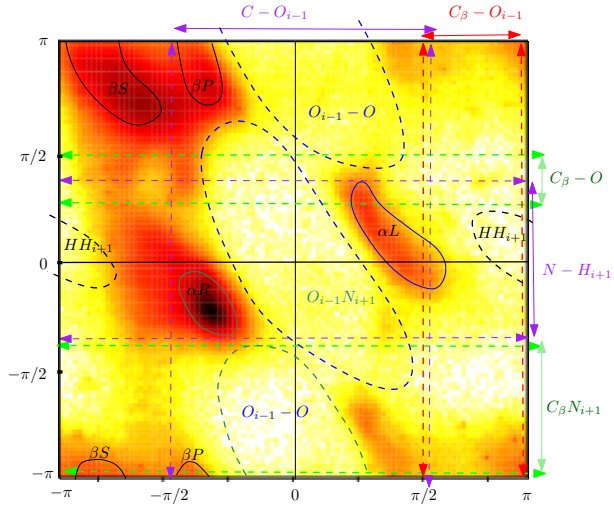
Exple: Lysine, 4 dihedral angles



LYS

The Ramachandran diagrams

▷ Ramachandran diagrams and populated regions

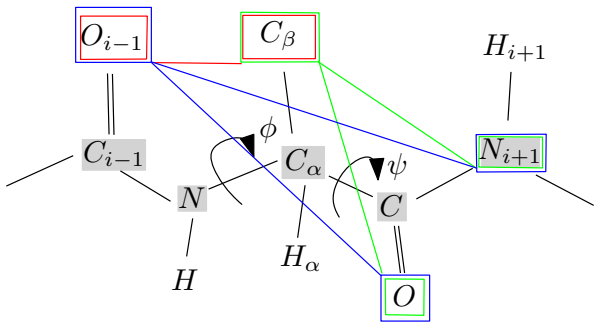


- Main regions: $\alpha L, \alpha R, \beta S, \beta P$
- Three prototypical diagrams
  - Glycine
  - Proline
  - Others – e.g. Aspartic acid

▷ Distance constraints and the Ramachandran tetrahedron

$C1 : C_\beta - O_{i-1}$ $C2 : C_\beta - O + C_\beta N_{i+1}$

$C3 : O_{i-1} - O + O_{i-1} N_{i+1}$



▷Ref: Stereochemistry of polypeptide chain configurations, JMB, 1963; Ramachandran et al

▷Ref: Revisiting the Ramachandran plot, Protein Science, 2003; Ho et al

Biomolecular recognition

Computational Structural Biology: what is a protein?

Protein Structure Resolution

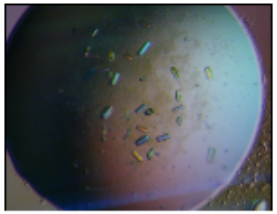
The importance of dynamics

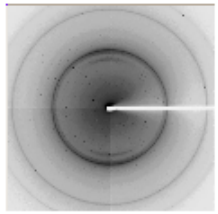
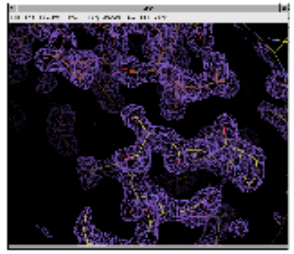
Protein functions: examples

Computational Structural Biology: challenges

Structure resolution:
X ray crystallography, NMR, cryo-electron microscopy

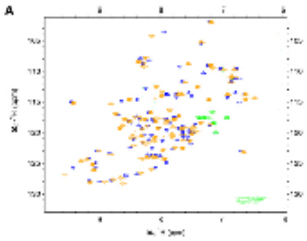
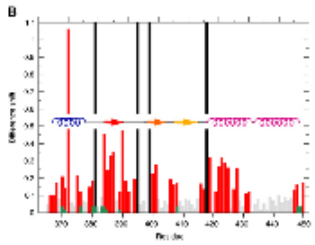
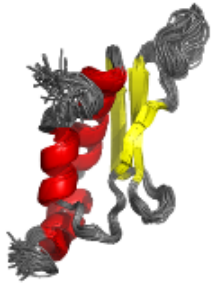
Crystallography



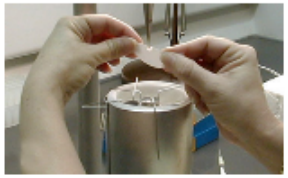



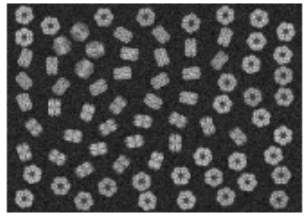
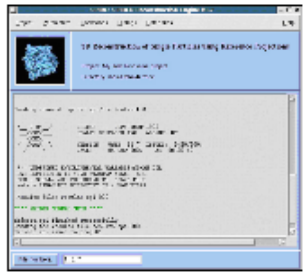
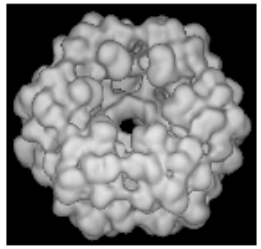

NMR



Cryo electron microscopy



Note: resolutions between 1 and 15 Å

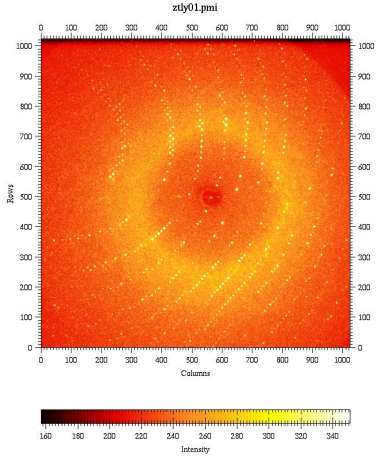
X ray crystallography

- (Selenium) crystals

- Protein crystals

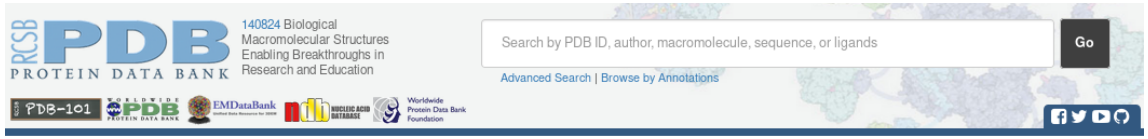
- X ray diffraction

- Diffraction pattern



The Protein Data Bank

- Structures in the PDB: origin and molecular type



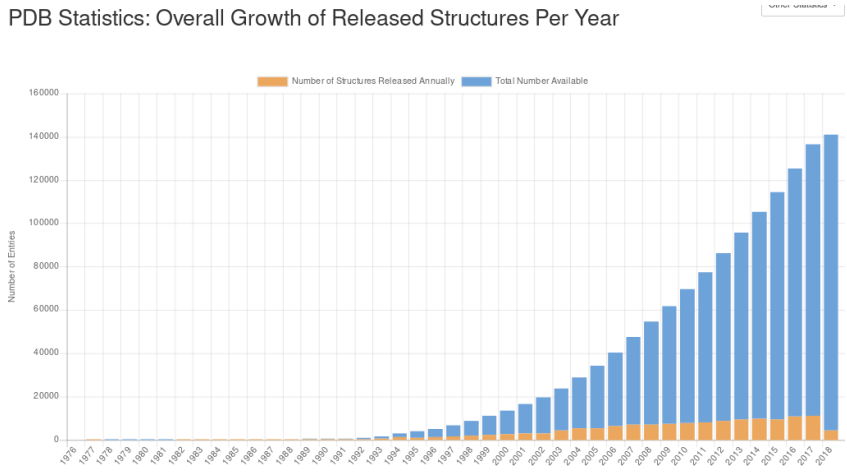
PDB Data Distribution by Experimental Method and Molecular Type

Other Statistics

Copy | CSV

Experimental Method	Proteins	Nucleic Acids	Protein/NA Complex	Other	Total
X-Ray	118103	1931	6027	10	126071
NMR	10723	1243	249	8	12223
Electron Microscopy	1569	31	543	0	2143
Other	241	4	6	13	264
Multi Method	116	4	2	1	123
Total	130752	3213	6827	32	140824

- Growth of the PDB



- To learn more: `https://www.rcsb.org/stats`

A typical PDB file

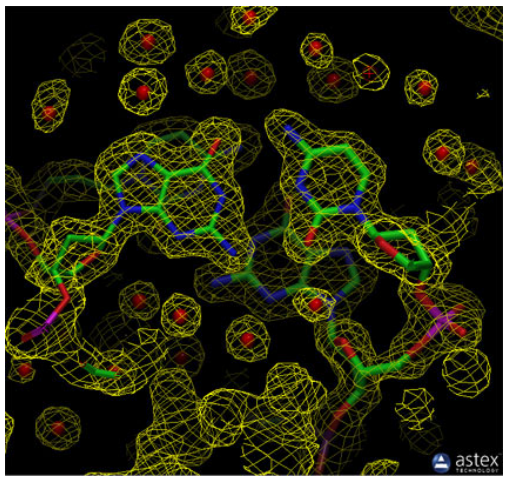
▷ Geometry information: n atoms yield $3n$ Cartesian coordinates . . . and $3n - 6$ degrees of freedom

```
ATOM      1  N   ASP A   1      23.963  -0.947  -1.031  1.00 37.52           N
ATOM      2  CA  ASP A   1      25.119  -0.797  -1.881  1.00 32.56           C
ATOM      3  C   ASP A   1      25.715   0.493  -1.356  1.00 29.72           C
ATOM      4  O   ASP A   1      24.964   1.396  -0.971  1.00 28.87           O
ATOM      5  CB  ASP A   1      24.721  -0.606  -3.341  1.00 34.71           C
ATOM      6  CG  ASP A   1      24.061  -1.777  -4.067  1.00 35.11           C
ATOM      7  OD1 ASP A   1      23.841  -2.849  -3.496  1.00 35.99           O
ATOM      8  OD2 ASP A   1      23.798  -1.612  -5.255  1.00 38.08           O
ATOM      9  H1  ASP A   1      23.429  -0.061  -1.100  1.00 20.00           H
ATOM     10  H2  ASP A   1      23.417  -1.821  -1.194  1.00 20.00           H
ATOM     11  H3  ASP A   1      24.348  -0.968  -0.067  1.00 20.00           H
ATOM     12  N   ILE A   2      27.025   0.577  -1.277  1.00 26.56           N
ATOM     13  CA  ILE A   2      27.669   1.808  -0.873  1.00 25.29           C
ATOM     14  C   ILE A   2      27.740   2.665  -2.147  1.00 26.50           C
ATOM     15  O   ILE A   2      28.123   2.164  -3.216  1.00 26.25           O
```

▷ Other pieces of information: organism, molecules / sequences (and their engineering), crystal resolution and symmetry group, secondary structures, disulfide bonds.

PDB files: pitfalls

- Focus on files from X ray crystallography:
  - Crystal structures: a confined environment
  - Asymetric unit versus biological unit
  - Extra atoms/molecules: water, chemical, co-factors, etc
  - Missing atoms: H systematically, heavy atoms . . . often
  - Alternate locations – if several conformations
  - Atoms retain dynamics encoded in B factors
  - Resolution and precision on coordinates – a complex problem



- To learn more: `https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure`

Biomolecular recognition

Computational Structural Biology: what is a protein?

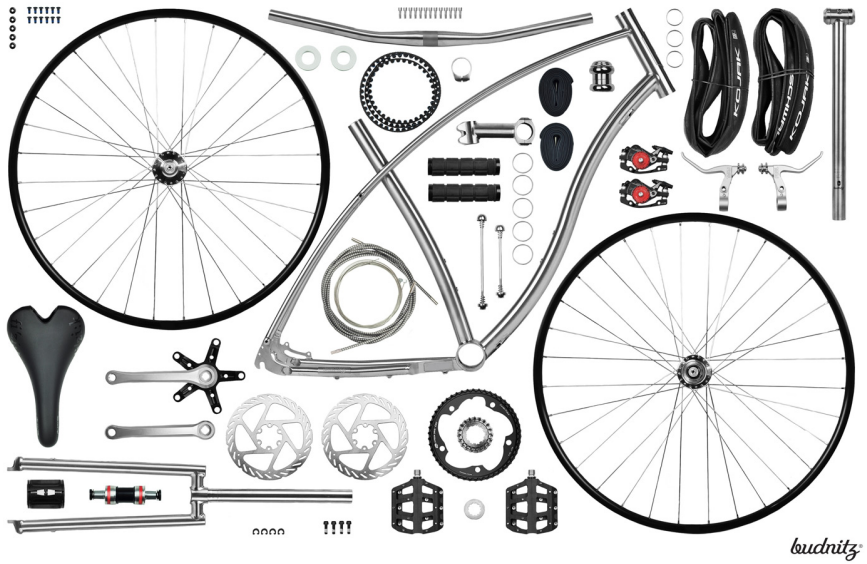
Protein Structure Resolution

The importance of dynamics

Protein functions: examples

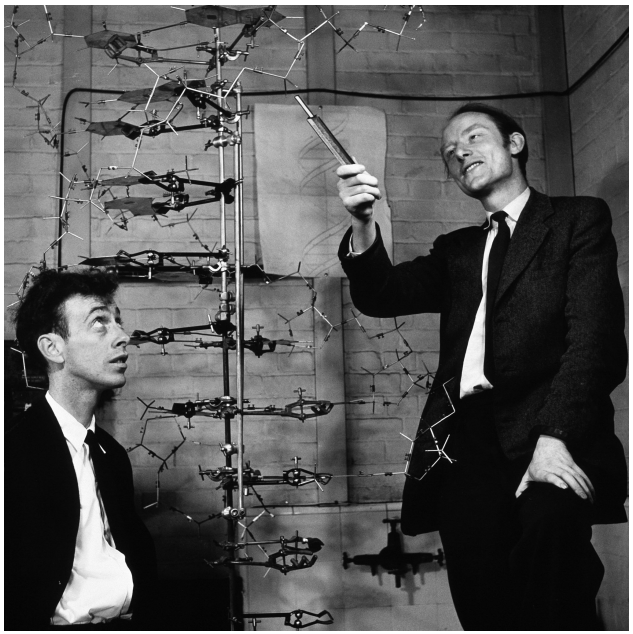
Computational Structural Biology: challenges

Statics vs dynamics



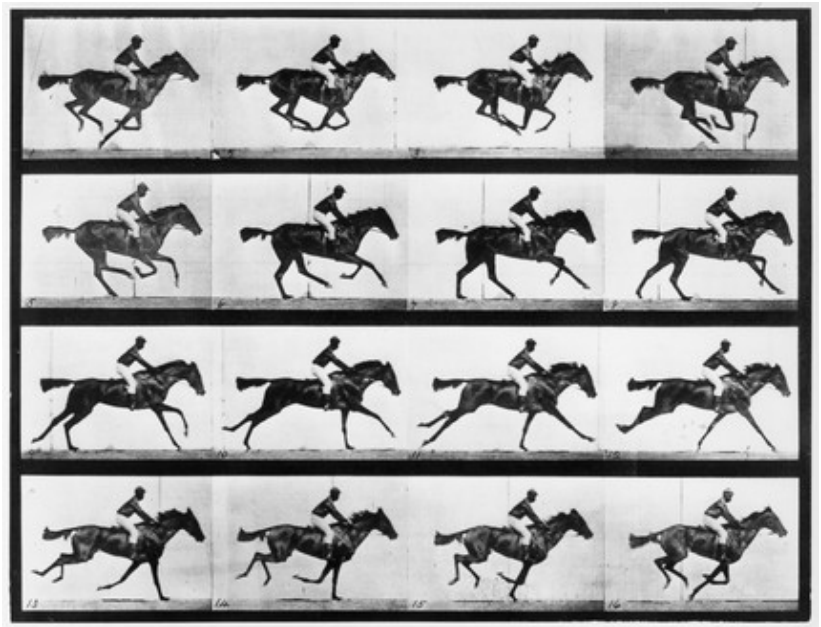
Two schools: static versus dynamic studies

▷ Balls and sticks



(Watson and Crick, DNA model)

▷ The Ballet & time lapse

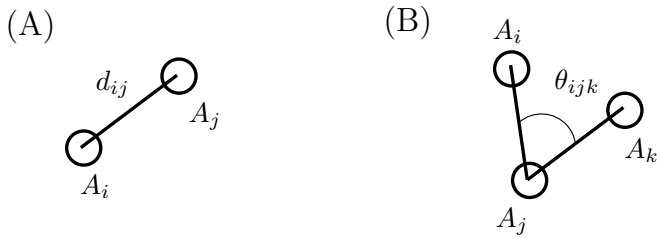


- Static analysis using crystal structure from the Protein Data Bank
  http://rcsb.org
- Dynamical analysis using molecular mechanics

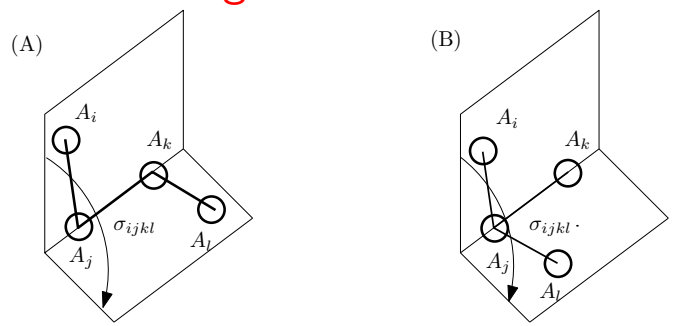
Geometric models: Cartesian and internal coordinates

▷ Cartesian versus internal coordinates: $\{x_i y_i z_i\}_i$ versus $\{d_{ij}, \theta_{ijk}, \sigma_{ijkl}\}$

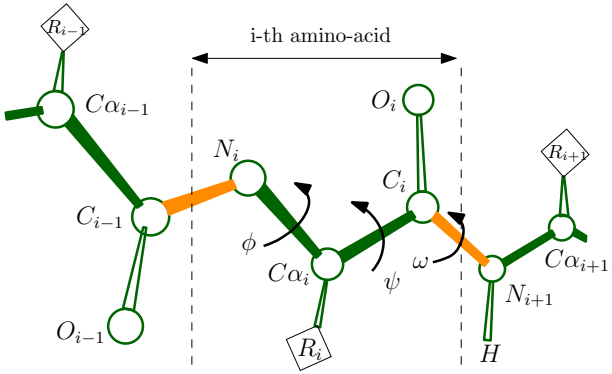
▷ Bond length and valence angle



▷ Dihedral angles



▷ Protein backbone

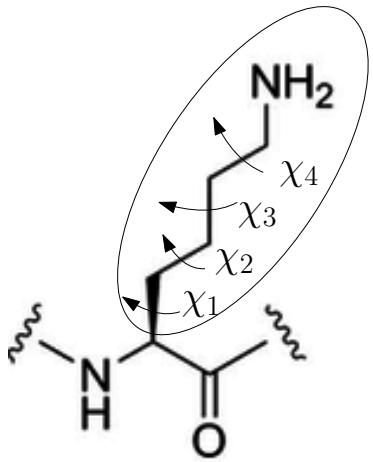


Ramachandran diagram, per a.a. type:

- bivariate distribution for (ϕ, ψ)

▷ Side chain: 20 natural amino acids

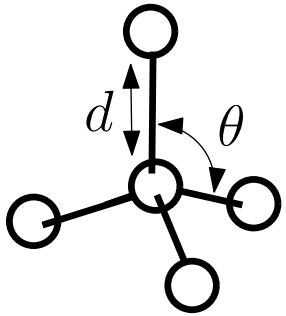
Exple: Lysine, 4 dihedral angles



LYS

The potential energy of (bio-)molecules: force fields

- The $3n-6$ degrees of freedom of a molecule:
  - types for atoms (element, bonds)
  - covalent: bond lengths, angles
  - non covalent: pairwise distances
  - solvent model

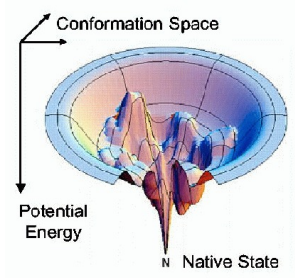


- Potential energy: non linear function

$$V_{\text{total}} = V_{\text{bond}} + V_{\text{angle}} + (V_{\text{proper}} + V_{\text{improper}}) + (V_{\text{vdw}} + V_{\text{electro}}) \qquad (1)$$

V_{bond}: bonds
V_{angle}: covalent angles
V_{proper}: proper dihedrals
V_{improper}: improper dihedrals
V_{vdw}: van der Walls
V_{electro}: electrostatics

- Examples:
  - AMBER: $S_u = (73, 133, 112, 3, 14, 758)$
    1093 unique parameters
  - CHARMM: $S_u = (85, 152, 209, 13, 33, 1)$
    493 unique parameters
  - MARTINI: $S_u = (16, 4, 0, 2, 21, 3)$
    46 unique parameters



(Open problem) Complexity of Potential Energy Landscape

- Consider a force field of the following type:

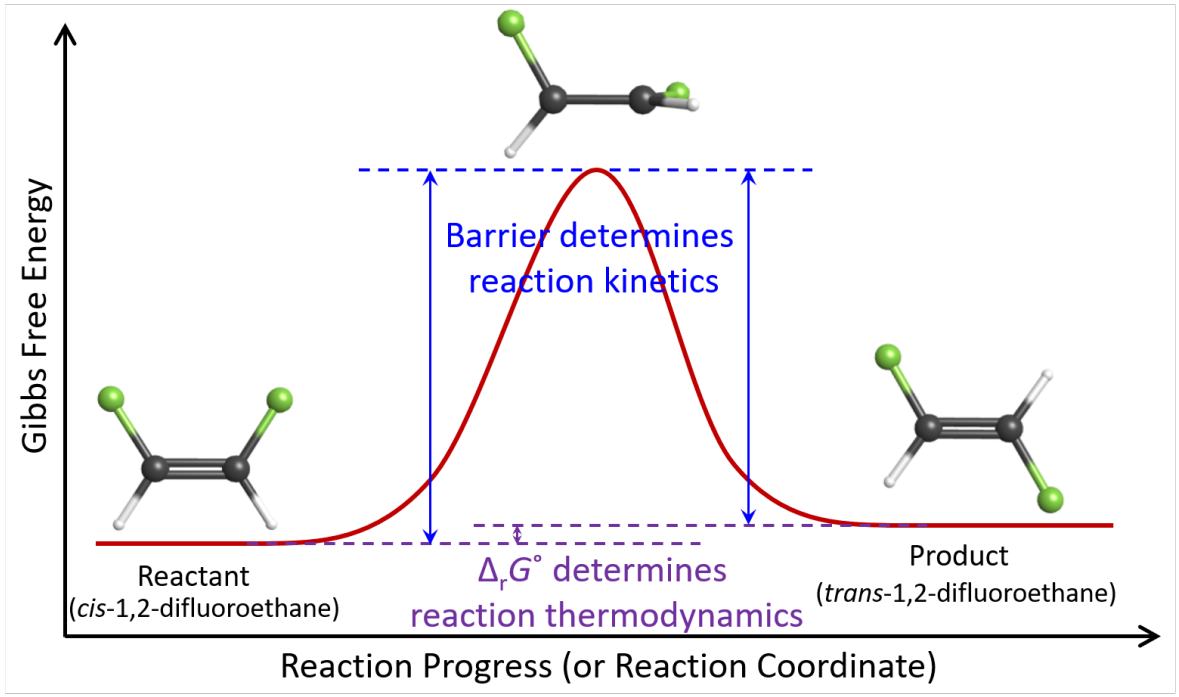
$$V_{BLN} = \frac{1}{2} \cdot K_r \sum_{i=1}^{N-1} (R_{i,i+1} - R_e)^2 + \frac{1}{2} K_0 \sum_{i=1}^{N-2} (\theta_i - \theta_e)^2$$
$$+ \epsilon \cdot \sum_{i=1}^{N-3} [A_i(1 + \cos\phi_i) + B_i(1 + 3\cos\phi_i)]$$
$$+ 4\epsilon \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \cdot C_{ij}[(\frac{\sigma}{R_{i,j}})^{12} - D_{ij}(\frac{\sigma}{R_{i,j}})^6]$$

- Open questions:
  - Number of critical points (local minima, index one saddles)
  - Geometry of the catchment basins (stable manifolds for $-\nabla V$)
  - (Topological) Persistence of local minima
- Rationale:
  - Separation bounds for polynomials
  - Complexity results *à-la* Yomdin-Comte / Tame geometry

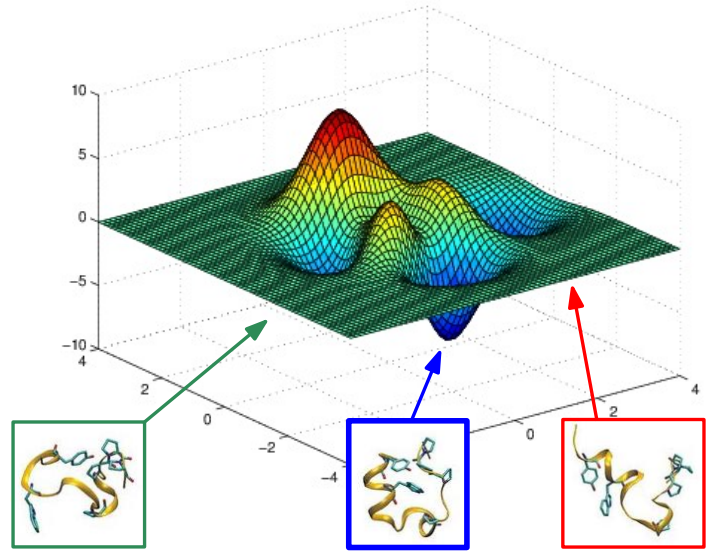
Thermodynamics

- Quantities defined for a conformation x:
  - potential energy: $V(x)$
  - kinetic energy: $K(x)$
  - total energy: $E(x) = V(x) + K(x)$
  - Boltzmann's distribution: $P^{\text{eq}}(x) = e^{-\beta E(x)}/Z, Z = \sum_{\text{Conformation}x} P^{\text{eq}}(x)$
- Quantities defined for ensembles:
  - Average of observable $\mathcal{O}$ wrt an ensemble: $< \mathcal{O} > \equiv \sum_{\text{Conformation}x} \mathcal{O}(x) P^{\text{eq}}(x)$
  - Exple: average total energy $U =< E >$
  - NVT: Helmholtz free energy $A = U - TS = k_B T \ln Z$
  - NPT: Gibbs free energy $G = U + PV - TS = H - TS$

Emergence of macromolecular function(s) from Structure – Thermodynamics – Kinetics



Emergence of macromolecular function(s) from Structure – Thermodynamics – Kinetics

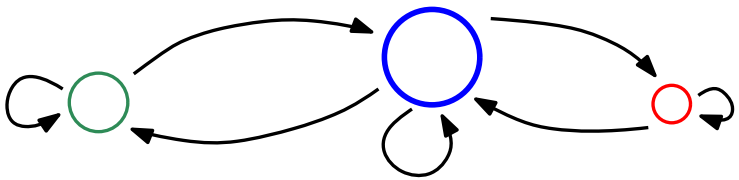


Potential Energy Landscape

- large number of local minima
- enthalpic barriers
- entropic barriers

Structure: stable conformations i.e. local minima of the PEL

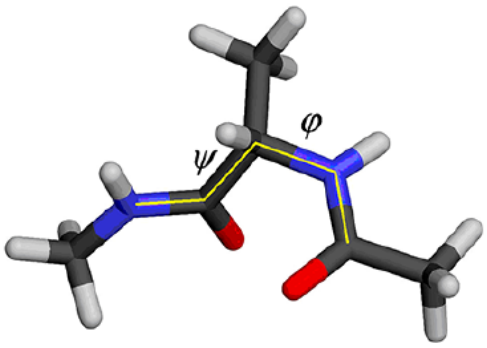
Thermodynamics: meta-stable conformations i.e. ensemble of conformations easily inter-convertible into one - another.

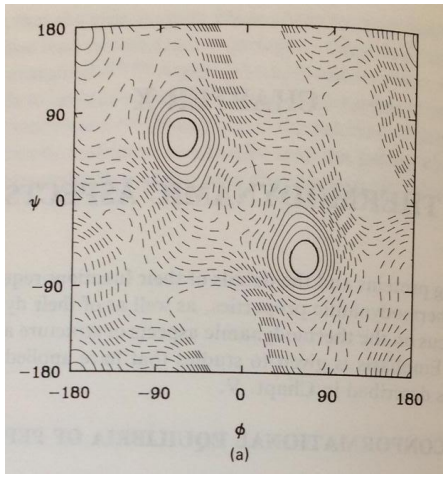


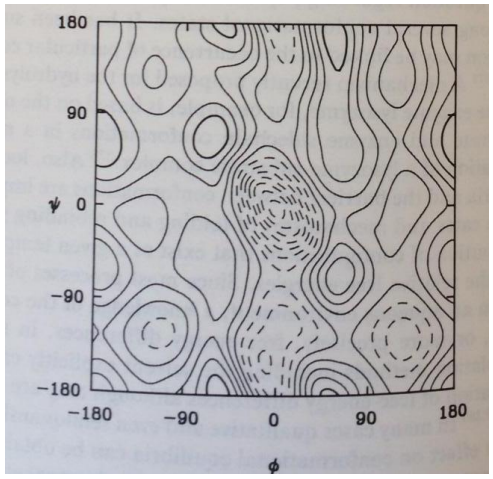
Kinetics: transitions between meta-stable conformations e.g. Markov state model

Potential energy landscapes: illustration

- Potential energy map: vacuum (PE) versus solvated (PMF):

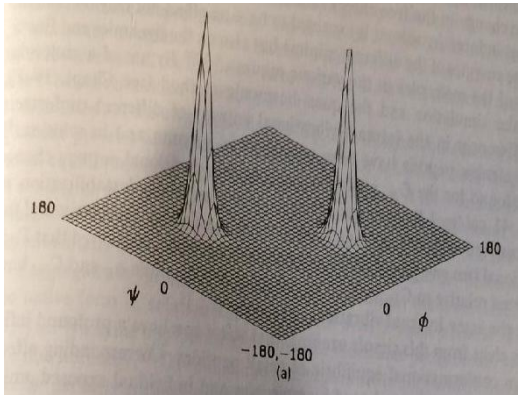


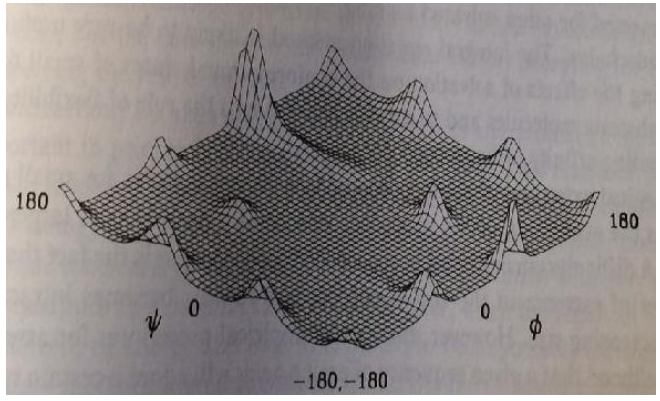




- Corresponding Boltzmann-weighted probability maps:

Solvent stabilizes many more conformers–hydrogen bonding.





- Ref: Petitt, Karplus, Chem. Phys. Lett., 121, 1985

Dynamics of biomolecules: first molecular simulation of a protein

About the simulation duration, quoting M. Levitt "*Cannot remember, but likely less than 100 picoseconds.*
Nb: from the late eighties"

Challenge *Dynamics of proteins*: specification

Youtube

- Input: structure(s) of biomolecules + potential energy model
- Output
  - Thermodynamics: meta-stable states and observables
  - Kinetics: transition rates, Markov state models
- Time-scales
  - Biological time-scale $>$ millisecond
  - Integration time step in molecular dynamics: $\Delta t \sim 10^{-15} s$

- 162 amino acids, $>$ 2000 atoms
- 5.058ms of simulation time
- $\sim$ 230 GPU years on NVIDIA GeForce GTX 980 processor

Ref: Chodera et al, eLife, 2019

Protein motions: time scales

Table 1. Characteristic Time Scales for Protein Motions

event	spatial extent (nm)	amplitude (nm)	time (s)	appropriate simulations
bond-length vibration	0.2−0.5	0.001−0.01	10^{-14}−10^{-13}	QM methods
elastic vibration of globular domain	1.0−2.0	0.005−0.05	10^{-12}−10^{-11}	conventional MD
rotation of solvent-exposed side chains	0.5−1.0	0.5−1.0	10^{-11}−10^{-10}	conventional MD
torsional libration of buried groups	0.5−1.0	0.05	10^{-11}−10^{-9}	conventional MD
hinge bending (relative motion of globular domains)	1.0−2.0	0.1−0.5	10^{-11}−10^{-7}	Langevin dynamics, enhanced sampling MD methods?
rotation of buried side chains	0.5	0.5	10^{-4}−1	enhanced sampling MD methods?
allosteric transitions	0.5−4.0	0.1−0.5	10^{-5}−1	enhanced sampling MD methods?
local denaturation	0.5−1.0	0.5−1.0	10^{-5}−10^{1}	enhanced sampling MD methods?
loop motions	1.0−5.0	1.0−5.0	10^{-9}−10^{-5}	Brownian dynamics?
rigid-body (helix) motions		1.0−5.0	10^{-9}−10^{-6}	enhanced sampling MD methods?
helix−coil transitions		>5.0	10^{-7}−10^{4}	enhanced sampling MD methods?
protein association	≫1.0			Brownian dynamics

▷Ref: Adcock and McCammon, Chem. Rev., 2006

Potentials of Mean Force and free energies

▷ Rationale: decouple the slow and fast dof of a system. Example: solvated protein:

- slow dof: protein
- fast dof: solvent molecules

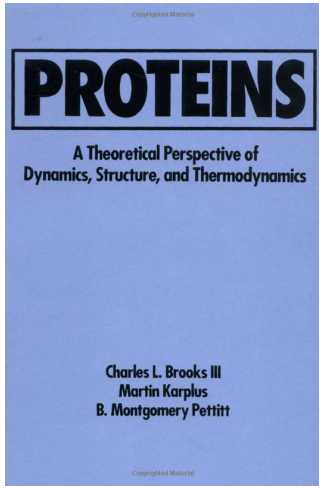
▷ How to: replace the overall potential energy by an average, computed over the fast dof

▷ PMF definition:

$$\exp(-\beta PMF(x_1,\ldots,x_n)) \propto \frac{\int \exp(-\beta V(x_1,\ldots,x_d))dx^{n+1,\ldots,d}}{\rho_{\text{unif.}(x_1,\ldots,x_n)}} \qquad (2)$$

Nb: in this equation, $\rho_{\text{unif.}}$ stand for the uniform distribution on the slow dof, which naturally depends on the nature of these parameters – cartesian or internal coordinates.

Dynamics: *alea jacta est* in the mid eighties



CONTENTS

I. INTRODUCTION 1

II. PROTEIN STRUCTURE AND DYNAMICS—AN OVERVIEW 7
- A. The Structure of Proteins 7
- B. Overview of Protein Motions 14

III. POTENTIAL FUNCTIONS 23
- A. Theoretical Basis 23
- B. Form of Potential Functions 25
- C. Parameter Determination 30

IV. DYNAMICAL SIMULATION METHODS 33
- A. General Features of Molecular Dynamics Methods 33
- B. Molecular Dynamics with Conventional Periodic Boundary Conditions 36
- C. Molecular Dynamics with Stochastic Boundary Conditions 38
- D. Stochastic Dynamics with a Potential of Mean Force 44
- E. Activated Dynamics 46
- F. Harmonic and Quasi-Harmonic Dynamics 49
- G. Algorithms for Molecular and Stochastic Dynamics 51
- H. Minimization Algorithms 54

V. THERMODYNAMIC METHODS 59
- A. Vacuum Calculations 59
- B. Free Energies in the Condensed Phase 62
- C. Thermodynamic Perturbation Theory 66

VI. ATOM AND SIDECHAIN MOTIONS 75
- A. Atom Motions 75
  - 1. Amplitudes and Distributions 76
  - 2. Time Dependence: Local and Collective Effects 84
  - 3. Harmonic Dynamics 87
  - 4. Biological Role of Atom Fluctuations 94
- B. Sidechain Motions 95
  - 1. Aromatic Sidechains 95
  - 2. Ligand-Protein Interaction in Myoglobin and Hemoglobin 111

VII. RIGID-BODY MOTIONS 117
- A. Helix Motions 117
- B. Domain Motions 119
- C. Subunit Motions 125

VIII. LARGER-SCALE MOTIONS 127
- A. Helix-Coil Transition 128
- B. Protein Folding 129
- C. Disorder-to-Order Transitions 132
  - 1. Trypsinogen-Trypsin Transition 133
  - 2. Triosephosphate Isomerase 135

IX. SOLVENT INFLUENCE ON PROTEIN DYNAMICS 137
- A. Global Influences on the Structure and Motional Amplitudes 137
- B. Influence on Dynamics 142
  - 1. Alanine Dipeptide Results 143
  - 2. Protein Results 146
  - 3. Stochastic Dynamics Simulations of Barrier Crossing in Solution 153
- C. Solvent Dynamics and Structure 154
- D. Role of Water in Enzyme Active Sites 161
- E. Solvent Role in Ligand-Binding Reactions 169

X. THERMODYNAMIC ASPECTS 175
- A. Conformational Equilibria of Peptides 175
- B. Configurational Entropy of Proteins 180
- C. Ligand Binding, Mutagenesis, and Drug Design 183

XI. EXPERIMENTAL COMPARISONS AND ANALYSIS 191
- A. X-Ray Diffraction 191
- B. Nuclear Magnetic Resonance 199
- C. Fluorescence Depolarization 211
- D. Vibrational Spectroscopy 216
- E. Electron Spin Relaxation 218
- F. Hydrogen Exchange 219
- G. Mössbauer Spectroscopy 221
- H. Photodissociation and Rebinding Kinetics 223

XII. CONCLUDING DISCUSSION 225

REFERENCES 233

INDEX 251

▷Ref: Brooks, Karplus, Montgomery Pettitt; Advances in Chemical Physics, Proteins; Wiley, 1988

Biomolecular recognition

Computational Structural Biology: what is a protein?

Protein Structure Resolution

The importance of dynamics

Protein functions: examples

Computational Structural Biology: challenges

Molecular dynamics and protein functions: movies

▷ Selected (great) movies:

- Protein synthesis by the ribosome:
  `https://www.youtube.com/watch?v=TfYf_rPWUdY`
- Membrane fusion and infection by SARS-Cov-2:
  `https://youtu.be/e2Qi-hAXdJo`
- Molecular motors: `https://www.youtube.com/watch?v=X_tYrnv_o6A`

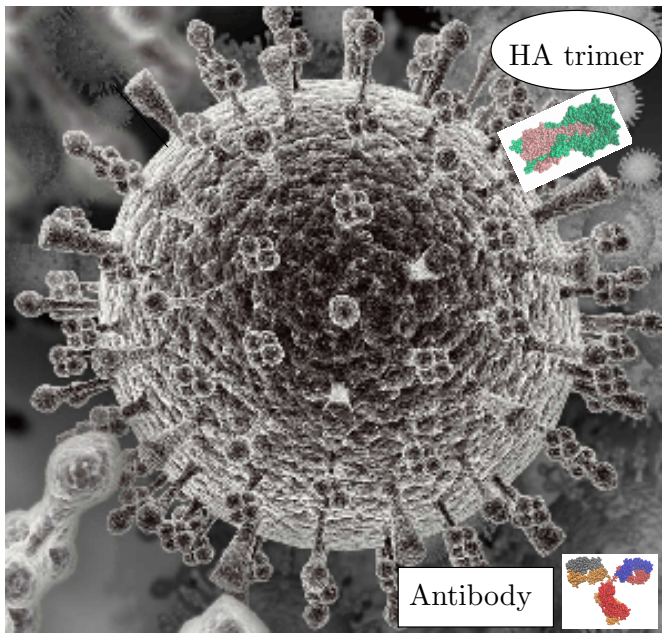
▷ Other videos of interest:

- Various phenomena in this movie:
  `https://www.youtube.com/watch?v=wJyUtbn005Y`
- More XVivo movies at
  `https://www.youtube.com/channel/UCAUL7Wl_lydKXI8q0oi4CUw`

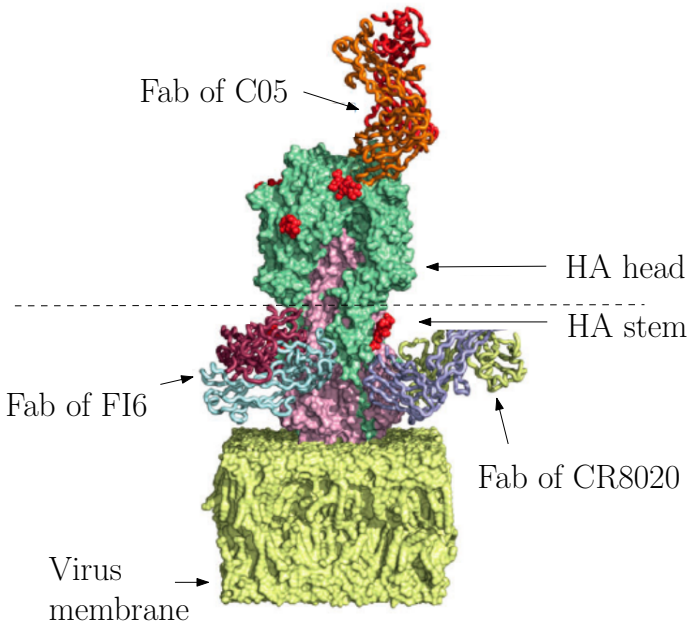
▷ Rmk. Remarkable illustration of the aforementioned mechanisms can be found in the book [?]; see also the gallery on the PDB portal, at `https://pdb101.rcsb.org/sci-art/goodsell-gallery`.

Structure - dynamics - function: illustration on antibody - antigen complexes

▷ Influenza



▷ (Broadly) neutralizing antibodies

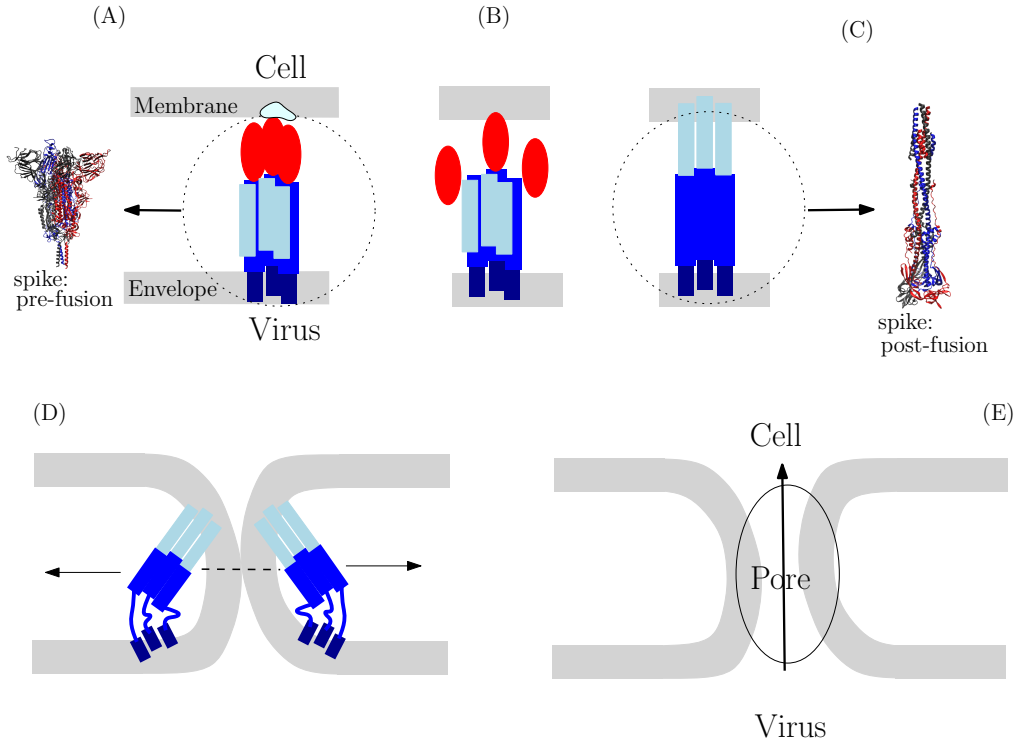


▷ Core questions – illustrated on on IG-Ag complexes

– Binding affinity: geometry (cf lock and key) + dynamics (entropy / free energy)
– Interaction specificity
– Multivalent binding: affinity - avidity - virus entry inhibition

SARS-Cov-2: cell entry mechanism

▷ SARS-Cov-2: cell entry mechanism via virus envelope - cell membrane fusion:



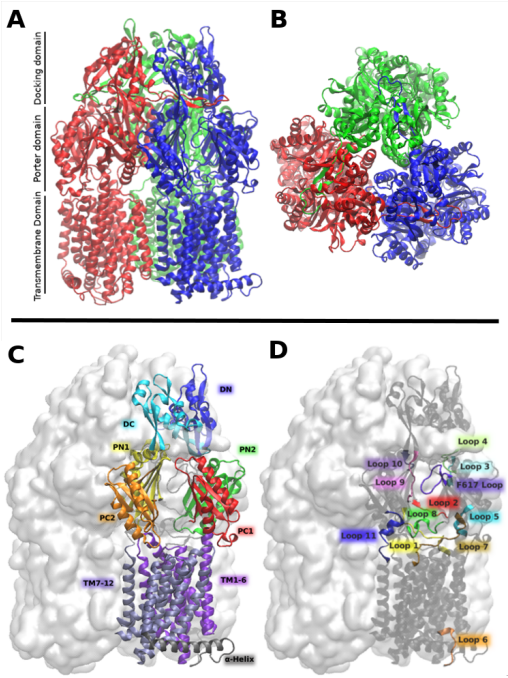
▷ Spike, the S1 and S2 domains: S1: the receptor binding domain (RBD, red ellipsis); S2: the fusion machinery (blue rectangles)

- **(A)** Attachment of the RBD to its receptor ACE2
- **(B)** Cleavage step removing the S1 subunit
- **(C)** Fusion machinery: refolding + membrane anchoring
- **(D,E)** Formation of the hemi-pore and pore

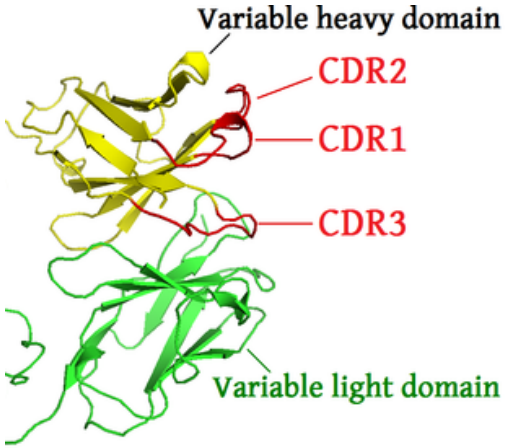
▷ Biophysics and biology of SARS-Cov-2/Omicron – Marc Gozlan:
`https://www.lemonde.fr/blog/realitesbiomedicales/2022/02/09/omicron-une-biologie-et-une-dynamique-virale-differentes-de-celles-observees-c`

Loops: biological relevance and dynamics

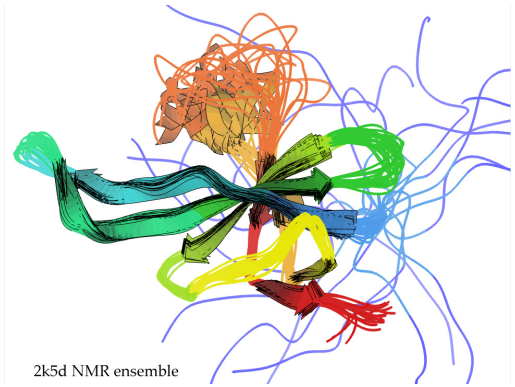
- Loops in biological processes



AcrB/RND



CDRs



IDP (2k5d/NMR)

- Action modes
  - (Structure) Global dynamics: global motions of domains
  - (Thermodynamics) Localized dynamics of CDR in antibodies (binding affinity)
  - (Mix) IDP and more generally highly flexible regions

- Open problems: accurate predictions for structure / thermodynamics / kinetics

Keywords: AI, deep learning, `AlphaFold2`, Covid19, protein design, flexibility, thermodynamics

Biomolecular recognition

Computational Structural Biology: what is a protein?

Protein Structure Resolution

The importance of dynamics

Protein functions: examples

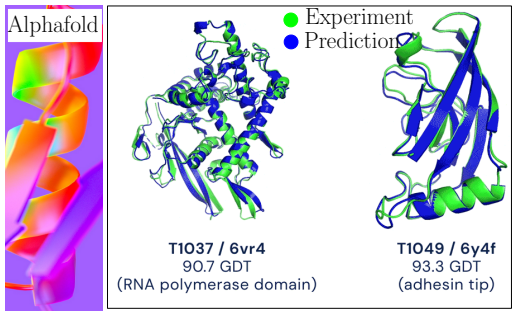
Computational Structural Biology: challenges

Challenge *Structure of proteins*: specification

- Input: sequences from genome sequencing projects

```
P69905 (HBB_HUMAN) MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GS 53
P68871 (HBB_HUMAN) MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN 58
P02144 (MYG_HUMAN) -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKAS 59
                    : *:  :   *   ****         * * *::   * *   *  *  *
```

- Output: plausible structures i.e. atomic coordinates $\{(x_i, y_i, z_i)\}$

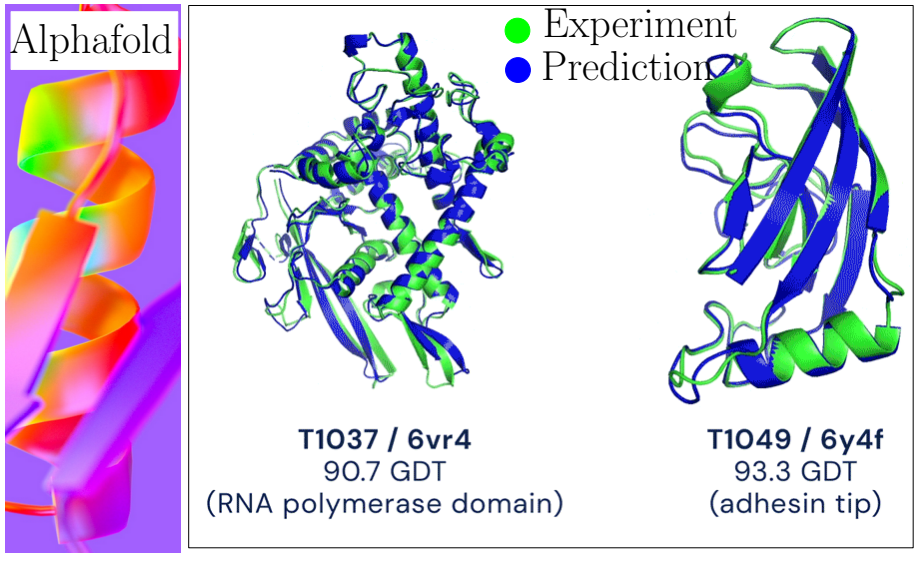


- Protein sequences versus structures: numbers
  - Num. sequences in `UniProtKB`: `TrEMBL` ($\sim 10^8$), `Swiss-Prot` ($\sim 10^5$)
  - Num. structures in the Protein Data Bank: $\sim 150,000$ structures
- Recent & notable: the Deepmind combined approach (DL, optimization)
  - Bias towards well folded structure – no disorder (IDP)
  - Structure only – neither thermodynamics nor kinetics
  - Predicting is not explaining

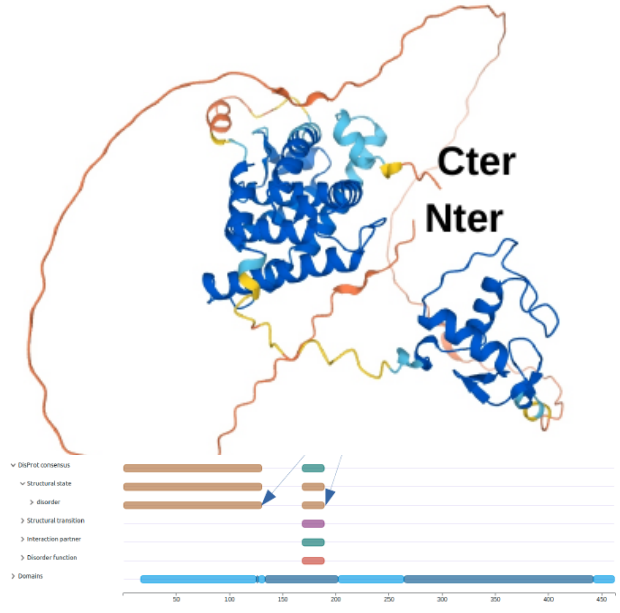
Alphafold by Deepmind

AI: what is being learned?

▷ Successes



▷ . . . and failures



▷ Recent & notable: the Deepmind combined approach (DL, optimization)

- Structure only – neither thermodynamics nor kinetics
- Bias towards well folded structure – no disorder (IDP)
- Predicting is not explaining
- Heavy engineering (team: 34 scientists/engineers)

▷Ref: Jumper et al, Nature, 2021

Challenge *Dynamics of proteins*: specification

Youtube

- Input: structure(s) of biomolecules + potential energy model
- Output
  - Thermodynamics: meta-stable states and observables
  - Kinetics: transition rates, Markov state models
- Time-scales
  - Biological time-scale $>$ millisecond
  - Integration time step in molecular dynamics: $\Delta t \sim 10^{-15} s$

- 162 amino acids, $>$ 2000 atoms
- 5.058ms of simulation time
- $\sim$ 230 GPU years on NVIDIA GeForce GTX 980 processor

Ref: Chodera et al, eLife, 2019

Modeling dynamics: shear difficulties

- Three sources of difficulties
  - System size: $3n - 6$ degrees of freedom: typically $> 10^4$
  - Time scales: 15 orders of magnitude
  - Spatial scales: 3 orders of magnitude

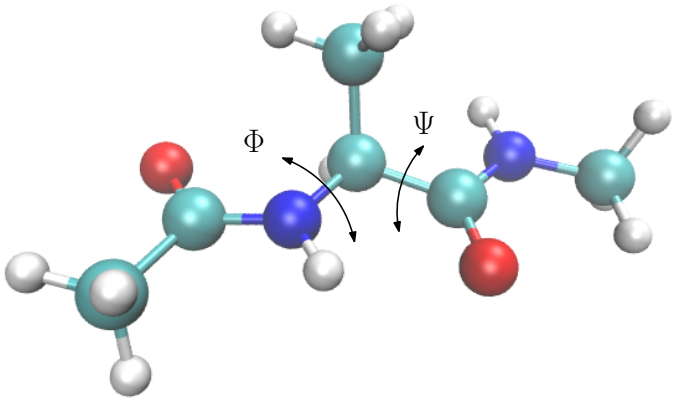
Table 1. Characteristic Time Scales for Protein Motions

event	spatial extent (nm)	amplitude (nm)	time (s)	appropriate simulations
bond-length vibration	0.2−0.5	0.001−0.01	10^{-14}−10^{-13}	QM methods
elastic vibration of globular domain	1.0−2.0	0.005−0.05	10^{-12}−10^{-11}	conventional MD
rotation of solvent-exposed side chains	0.5−1.0	0.5−1.0	10^{-11}−10^{-10}	conventional MD
torsional libration of buried groups	0.5−1.0	0.05	10^{-11}−10^{-9}	conventional MD
hinge bending (relative motion of globular domains)	1.0−2.0	0.1−0.5	10^{-11}−10^{-7}	Langevin dynamics, enhanced sampling MD methods?
rotation of buried side chains	0.5	0.5	10^{-4}−1	enhanced sampling MD methods?
allosteric transitions	0.5−4.0	0.1−0.5	10^{-5}−1	enhanced sampling MD methods?
local denaturation	0.5−1.0	0.5−1.0	10^{-5}−10^{1}	enhanced sampling MD methods?
loop motions	1.0−5.0	1.0−5.0	10^{-9}−10^{-5}	Brownian dynamics?
rigid-body (helix) motions		1.0−5.0	10^{-9}−10^{-6}	enhanced sampling MD methods?
helix−coil transitions		>5.0	10^{-7}−10^{4}	enhanced sampling MD methods?
protein association	≫1.0			Brownian dynamics

▷Ref: Adcock and McCammon, Chem. Rev., 2006

Density of states and partition functions

Dialanine

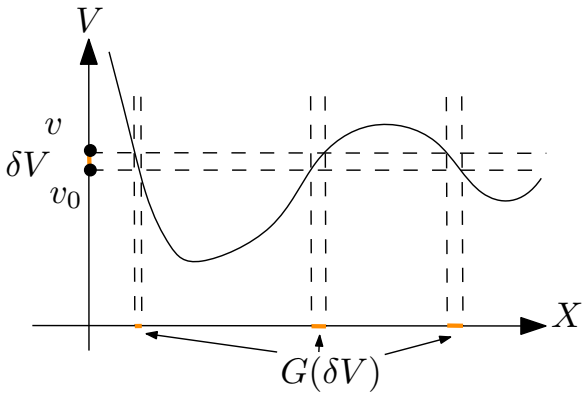


Molecule in water at temperature T

- q: vector of positions of atoms
- Potential energy:

$$V(q)$$

▷ Potential energy landscape:



▷ Density of states (DoS):

- Push forward of the Lebesgue measure by the potential energy V:
- For any $v_0 < v_1$:

$$g([v_0, v_1]) = \int_X 1_{[v_0, v_1]}(V(q))dq$$

▷ Partition function for $A \subset X$: integrate Boltzmann's factor

$$Z_A(T) = \int_{\mathbb{A}} e^{-\beta u} dg(u)$$

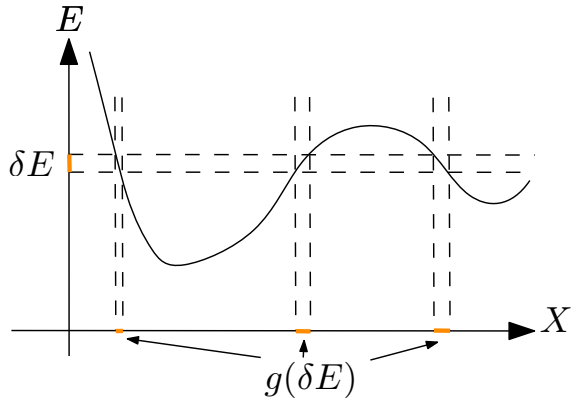
▷ NB: n atom: $d = 3n$ Cartesian coordinates. Exple: antibody: $d \sim 42,000$

Free energy, Density of States, Volumes

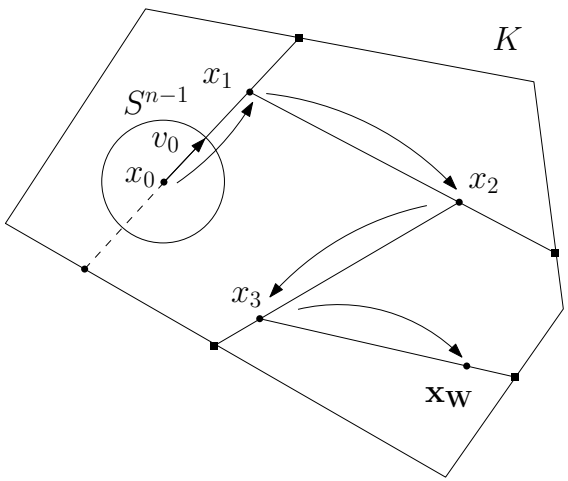
▷ Partition function and density of states:

$$Z = \sum_{x_i:\text{state}} e^{-\beta E(x_i)}$$

$$= \sum_{j:\text{energy level}} g(E_j) e^{-\beta E_j}$$



▷ Learning from simpler cases: polytopes in $\mathbb{R}^d$, $d \in [100 \ldots 1000]$



▷ Unless P=NP: no polynomial time algorithm with approx factor $(cd/\log d)^d$

▷ But: probabilistic algorithms running in $O^\star(d^{3.5})$

▷Ref: Cousins and Vempala, Math. Prog. Comp., 2016

▷Ref: Chalkis, Emiris, Fisikopoulos, arXiv:1905.05494, 2019

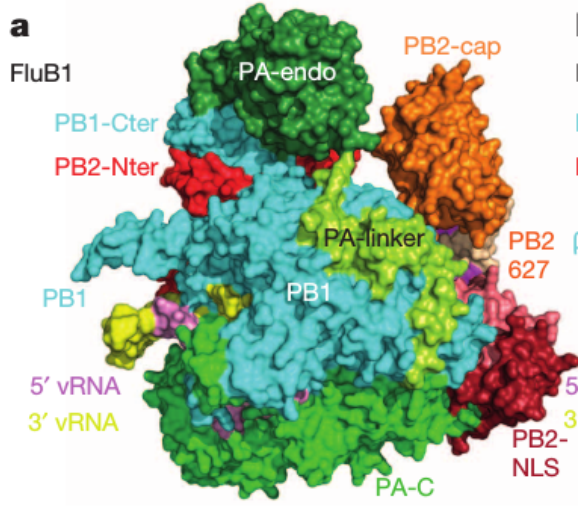
▷Ref: Chevallier et al, AISTATS, 2022

Challenge *Molecular machines–structure and dynamics*: specification

- ▷ Molecular machines: assemblies with tens / hundreds of subunits
- ▷ Input
  - cryo-electron microscopy (cryo-EM) maps of whole assemblies
  - crystal structures of subunits
  - other data: native mass spectrometry data, . . .
- ▷ Output: structure(s) + mechanism(s)

▷ Polymerase of E. coli: structure+dynamics

▷ Polymerase of influenza: structure



▷Ref: Scheres et al, Elife, 2015; ▷Ref: Cusak et al, Nature, 2015

Modeling dynamics: shear difficulties

- Three sources of difficulties
  - System size: $3n - 6$ degrees of freedom: typically $> 10^4$
  - Time scales: 15 orders of magnitude
  - Spatial scales: 3 orders of magnitude

Table 1. Characteristic Time Scales for Protein Motions

event	spatial extent (nm)	amplitude (nm)	time (s)	appropriate simulations
bond-length vibration	0.2−0.5	0.001−0.01	10^{-14}−10^{-13}	QM methods
elastic vibration of globular domain	1.0−2.0	0.005−0.05	10^{-12}−10^{-11}	conventional MD
rotation of solvent-exposed side chains	0.5−1.0	0.5−1.0	10^{-11}−10^{-10}	conventional MD
torsional libration of buried groups	0.5−1.0	0.05	10^{-11}−10^{-9}	conventional MD
hinge bending (relative motion of globular domains)	1.0−2.0	0.1−0.5	10^{-11}−10^{-7}	Langevin dynamics, enhanced sampling MD methods?
rotation of buried side chains	0.5	0.5	10^{-4}−1	enhanced sampling MD methods?
allosteric transitions	0.5−4.0	0.1−0.5	10^{-5}−1	enhanced sampling MD methods?
local denaturation	0.5−1.0	0.5−1.0	10^{-5}−10^{1}	enhanced sampling MD methods?
loop motions	1.0−5.0	1.0−5.0	10^{-9}−10^{-5}	Brownian dynamics?
rigid-body (helix) motions		1.0−5.0	10^{-9}−10^{-6}	enhanced sampling MD methods?
helix−coil transitions		>5.0	10^{-7}−10^{4}	enhanced sampling MD methods?
protein association	≫1.0			Brownian dynamics

▷Ref: Adcock and McCammon, Chem. Rev., 2006

Biomolecular recognition

PART 1: Introduction to Protein Science
PART 2: Biomolecular recognition

Main points

Main points:

- Proteins and binding affinity
- Enthalpy - entropy compensation
- The time dimension $1/K_{\text{off}}$
- Application: antibodies binding viruses

Biomolecular recognition

Biomolecular recognition: proteins and binding affinity

Association and dissociation constants K_a, K_d

Enthalpy - entropy compensation

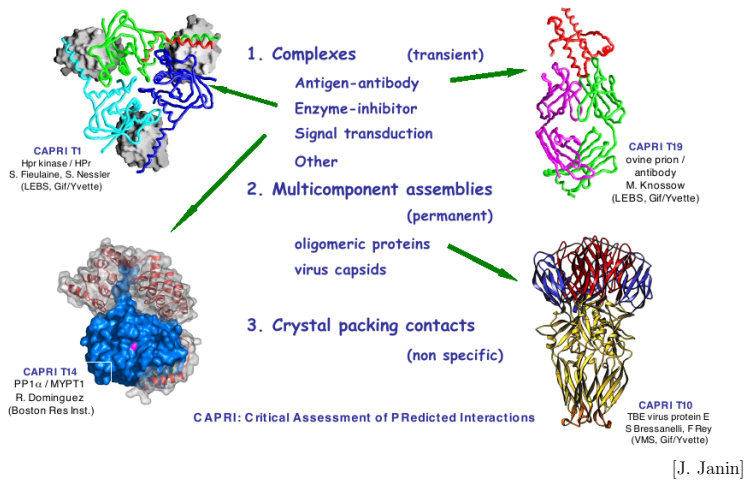
The time dimension: $K_d, K_{\text{on}}, K_{\text{off}}$ and $1/K_{\text{off}}$

Application: influenza - antibody complexes

Biological complexes: structural diversity

▷ Biology rests on interactions biomolecules make with one another. A remarkable variety of such complexes exist, both in size and time scales spanned [?].

▷ Size-wise, complexes span a range from $O(100\ kDa)$ up to 120 MDa (mammalian NPC). Note that the nuclear pore complex is the largest assembly known (to date) in eukaryotic cells, as it involves circa 500 polypeptide chains.



[J. Janin]

Figure: **Biological complexes: diversity.** From [?].

Biological complexes: time-wise

- Time-wise, biological complexes also span several orders of magnitude, say from the millisecond to years for permanent ones (Fig. 2).

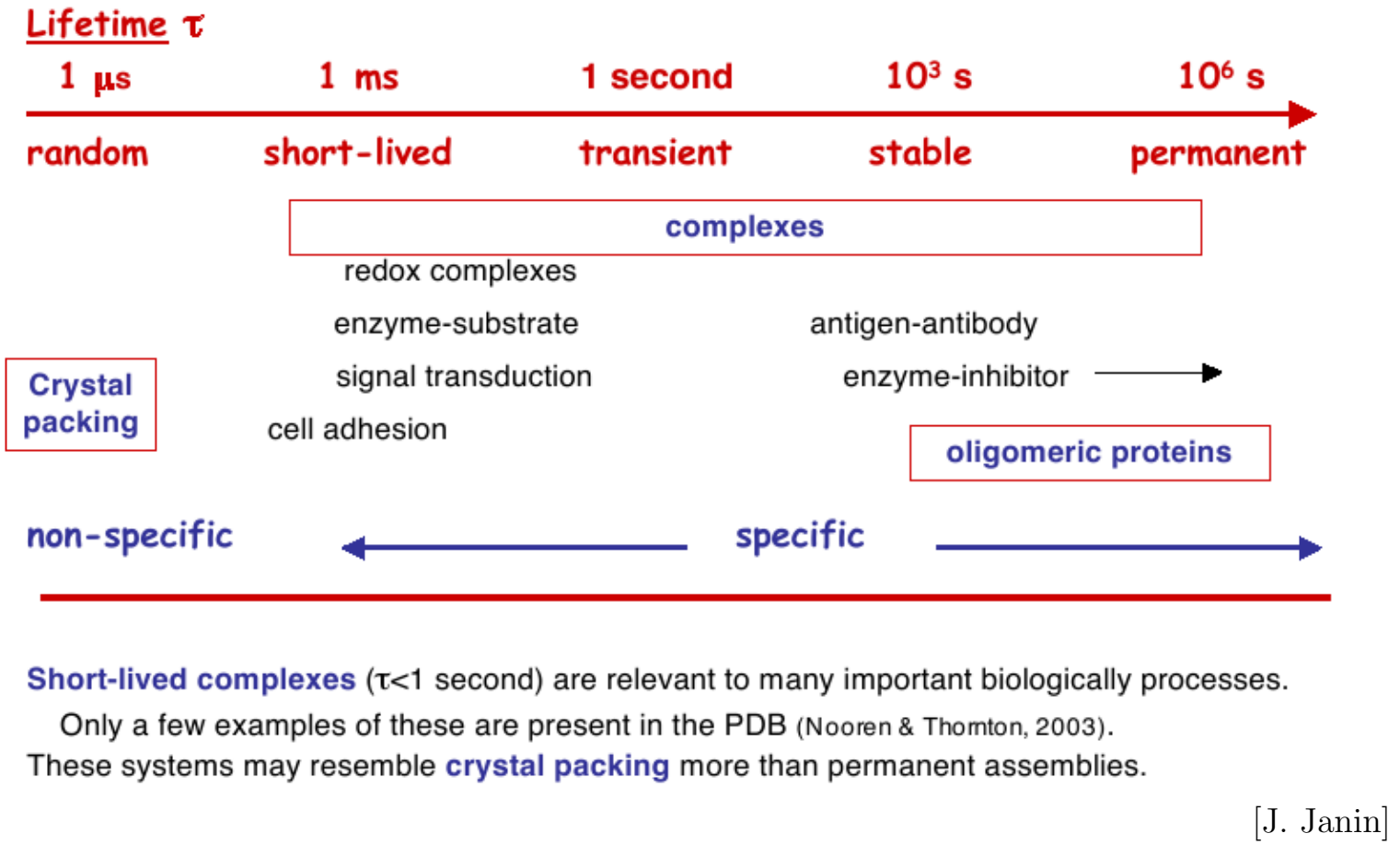


Figure: **Biological complexes: time scales.** From [?].

Docking models

- Over the years, several docking models have been proposed (Fig. 4):
  - Lock-and-key Fisher, 1894. In this model, the two partners associate as rigid bodies.
  - Induced fit: Koshland, 1958. While getting close, the partners *shape* one-another, resulting in the conformations found in the complex.
  - Conformer selection, Monod-Wyman-Changeux, 1965. In solution or in the cell, each molecule exists in a variety of conformations. In the course of their diffusion, *compatible* conformations stumble onto one-another, and the complex gets formed.

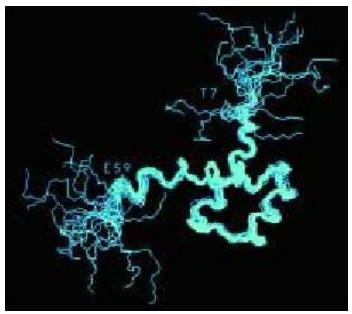


Figure: **Flexibility of biomolecules: illustration.** From the Nobel lecture of K. Wütricht.

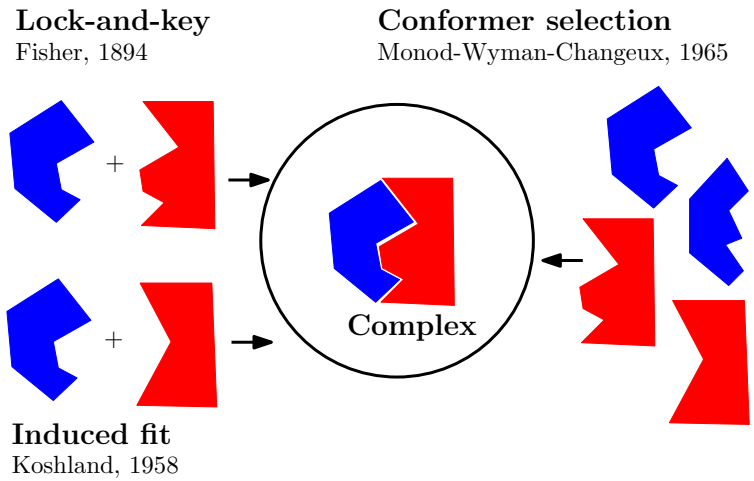


Figure: **Docking models.**

Chemical equilibrium

▷ Setup: we consider a protein P and a ligand L which interact in a non-covalent fashion. This means that no chemical bonds get created or removed. We further assume that these two species for a chemical equilibrium:

$$P + L \rightleftharpoons PL, K_{eq} = \frac{[PL]_{\text{Eq.}}}{[P]_{\text{Eq.}}[L]_{\text{Eq.}}} \tag{3}$$

▷ The notion of *equilibrium* is central here, and owes to competing effects:

- Due to attraction forces, P and L get closer to one another.
- Due in particular to thermal fluctuations, they get away.

▷ In the medium considered (test tube, cell): three chemical species: P, L, and the complex PL.

▷ In the sequel, we consider the standard setup:

- We start from std concentrations of the individual species, say 1 Molar
- We consider the equilibrium concentrations

Binding affinity: spectrum

- Typical binding affinity values are presented in Table 5.

Type of Interaction	K_D (molar)	ΔG^0_{bind} (at 300K) kJ mol^{-1}
Enzyme:ATP	~1×10^{-3} to ~1×10^{-6} (millimolar to micromolar)	–17 to –35
signaling protein binding to a target	~1×10^{-6} (micromolar)	-35
Sequence-specific recognition of DNA by a transcription factor	~1×10^{-9} (nanomolar)	-52
small molecule inhibitors of proteins (drugs)	~1×10^{-9} to ~1×10^{-12} (nanomolar to picomolar)	–52 to –69
biotin binding to avidin protein (strongest known non-covalent interaction)	~1×10^{-15} (femtomolar)	-86

Figure: **binding affinity: typical examples.** Table from [**?**, Chapter 12].

Binding affinity and specificity

- The two critical notions for protein interactions are
  - Binding affinity: the *strength* of the interactions.
  - Binding specificity: the variety of partners a molecules binds sufficiently strongly with.

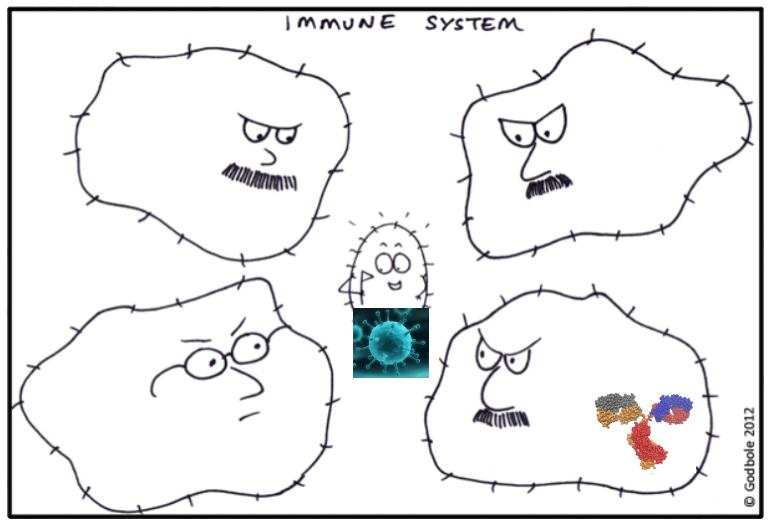


Figure: **Binding affinity and specificity: how to for the immune system.** The molecules secreted should bind strongly enough the pathogens; but they should also be quite specific.

Biomolecular recognition

Biomolecular recognition: proteins and binding affinity

Association and dissociation constants K_a, K_d

Enthalpy - entropy compensation

The time dimension: $K_d, K_{\text{on}}, K_{\text{off}}$ and $1/K_{\text{off}}$

Application: influenza - antibody complexes

Equilibrium constants K_a, K_d

- Consider the non-covalent interaction $P + L \rightleftharpoons PL$
- The law of mass action yields the association and dissociation constants:

$$\begin{cases} \text{Association constant} : K_a = \dfrac{[PL]_{\mathbf{Eq.}}}{[P]_{\mathbf{Eq.}}.[L]_{\mathbf{Eq.}}} \\ \text{Dissociation constant} : K_d = \dfrac{[P]_{\mathbf{Eq.}}.[L]_{\mathbf{Eq.}}}{[PL]_{\mathbf{Eq.}}} \end{cases} \tag{4}$$

Using std units, K_a is expressed in moles^{-1}, and K_d is in moles.

- Determine the concentration of the molecular species, here P, L, and PL, when the binding reaction reaches an equilibrium.
- The relationship between K_a and the variation of free energy satisfies:

$$\Delta G_a^0 = -RT \log c^0 K_a = RT \log \frac{K_d}{c^0}. \tag{5}$$

- Rmk. In Eq. 5, c^0 is meant to obtain a unit-less number: if K_a is expressed in moles^{-1} ,then C^0 is equal to 1 molar.

Fractional saturation

- The fraction of proteins with bound ligand satisfies:

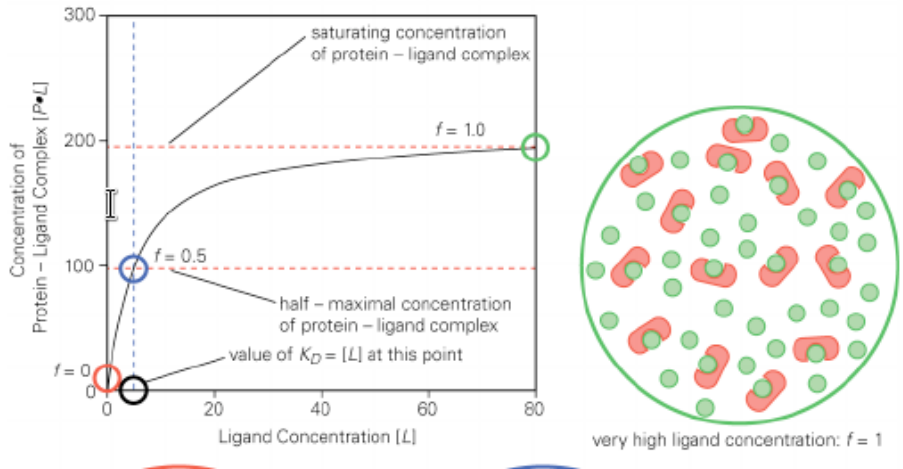
$$f = \frac{\#\text{num proteins with bound ligand}}{\text{total } \#\text{ proteins}} \tag{6}$$

$$= \frac{[PL]_{\text{Eq.}}}{[P]_{\text{Eq.}} + [PL]_{\text{Eq.}}} \tag{7}$$

$$= \frac{[P]_{\text{Eq.}}[L]_{\text{Eq.}}}{K_d([P]_{\text{Eq.}} + \frac{[P]_{\mathbf{Eq.}}[L]_{\mathbf{Eq.}}}{K_d})} \tag{8}$$

$$= \frac{1}{K_d(\frac{1}{K_d} + \frac{1}{[L]_{\mathbf{Eq.}}})} = \frac{[L]_{\text{Eq.}}}{[L]_{\text{Eq.}} + K_d} \tag{9}$$

- Varying the concentration of the ligand, one gets from Eq. 6:
- Observation: K_d is the concentration of the ligand such that the fraction of bound equals 1/2.



Biomolecular recognition

Biomolecular recognition: proteins and binding affinity

Association and dissociation constants K_a, K_d

Enthalpy - entropy compensation

The time dimension: $K_d, K_{\text{on}}, K_{\text{off}}$ and $1/K_{\text{off}}$

Application: influenza - antibody complexes

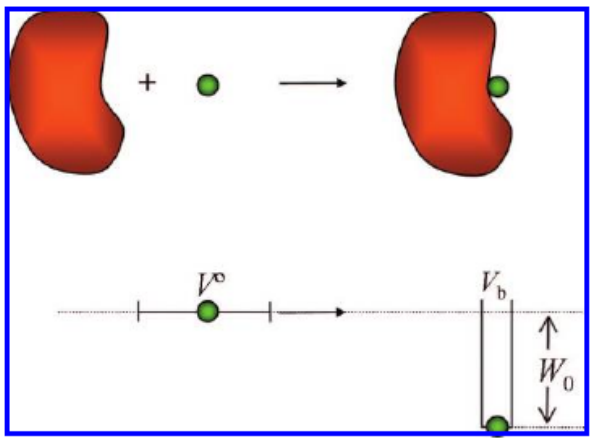
Enthalpy - entropy compensation - I

▷ To understand the components of binding, let us recall:

$$\Delta G_a^0 = -RT \log c^0 K_a = RT \log \frac{K_d}{c^0} = \Delta H - T\Delta S. \tag{10}$$

▷ To understand the relative variations of ΔH and $T\Delta S$, we need to discuss several components in turn:

- (1) System protein + ligand, enthalpy
- (2) Mixing: Two versus three species
- (3) Ligand and its translational / rotational entropy
- (4) System protein + ligand, conformational + vibrational entropy
- (5) Solvent and its entropy



Binding affinity: enthalpy-entropy competition illustrated along the binding process. The volume accessible to the ligand decreases, whence $T\Delta S < 0$ and $-T\Delta S > 0$. On the other hand, the interaction energy (enthalpy) decreases by W_0. From [**?**].

Enthalpy - entropy compensation - II

1. **System protein + ligand, enthalpy:**
   - Energy minimization when P and L get closer. (Exple: strong electrostatic interactions.)
2. **Mixing: Two versus three species:**
   - Three species (P, L, PL) have more entropy than two.
3. **Ligand and its translational / rotational entropy:**
   - Assuming P fixed: 6 dof of the ligand get constrained. Translation/rotational entropy decreases.
4. **System protein + ligand, conformational + vibrational entropy:**
   - In PL, conformational changes hindered + coupled harmonic oscillators: conformational and vibrational entropy decrease.
5. **Solvent and its entropy:**
   - Buried surface area at the interface $\Rightarrow$ the solvent S increases.

Summary:

- During association, grossly speaking: ΔH is negative, and $-T\Delta S$ is positive.
- Variation of enthalpy and entropy are very subtle, and the balance depends in general on the temperature.
- For biological systems: this subtlety is key to **regulation**. By slightly changing the conditions (temperature, pH, ionic strength which alter the electrostatic interactions), the behavior changes.

Enthalpy-entropy competition: illustration on protein unfolding

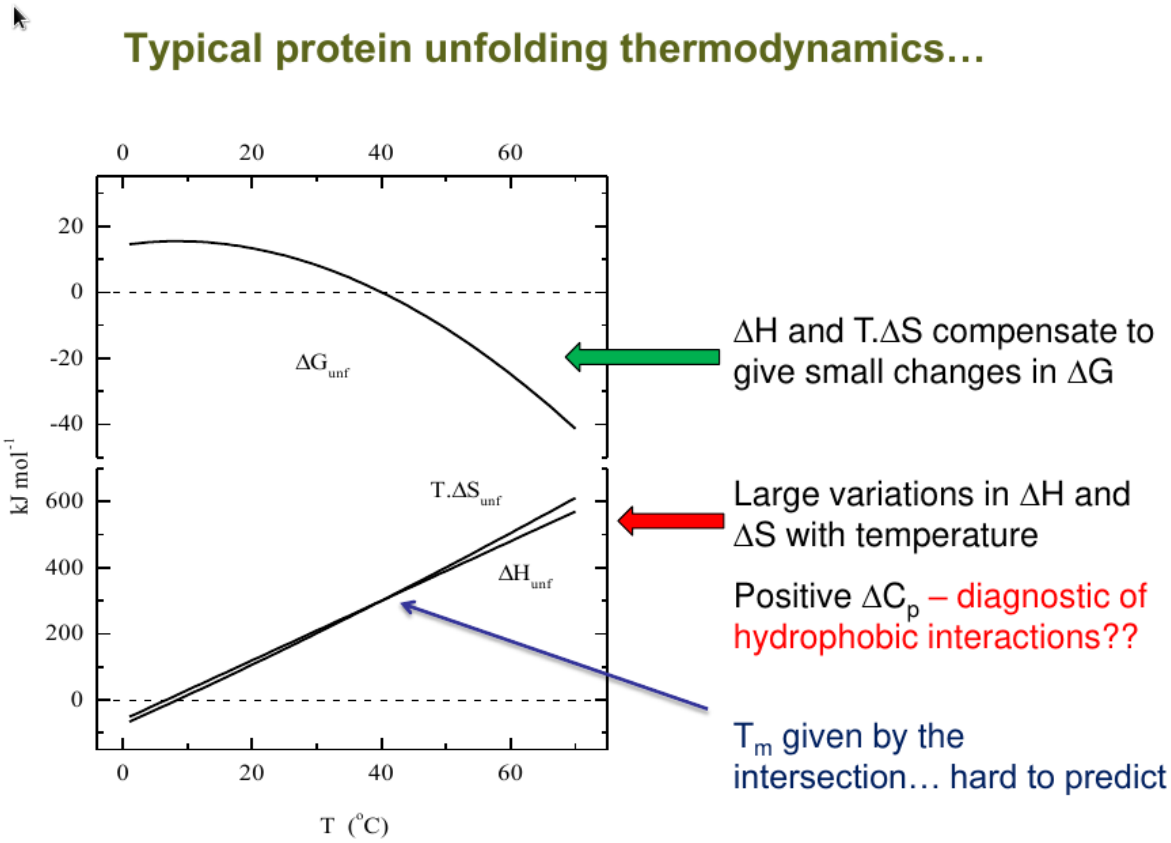


Figure: **Protein unfolding: illustration of the enthalpy-entropy competition.** Courtesy of Alan Cooper.

Binding affinity: ab initio calculations

- Model from molecular mechanics: potential energy / force field. With $X \in \{P, L, PL\}$:
  - r_X internal coordinates of molecular species X
  - $V(r_X)$ the potential energy, and $W(X)$ be the solvation energy.
- Dissociation free energy reads as – std concentration $c^\circ (= 1M)$:

$$\Delta G_d^\circ = -RT \ln\Big(\frac{c^\circ}{8\pi^2} \frac{(\int e^{-(V(r_P)+W(r_P))/RT} dr_P)(\int e^{-(V(r_L)+W(r_L))/RT} dr_L)}{\int e^{-(V(r_{PL})+W(r_{PL}))/RT} dr_{PL}}\Big). \quad (11)$$

- Major difficulties:
  - Very high dimensionality
  - Complex energy functions

▷Ref: Gilson, Zhou; Ann. Rev. Biomol. Struct., 2007

Biomolecular recognition

Biomolecular recognition: proteins and binding affinity

Association and dissociation constants K_a, K_d

Enthalpy - entropy compensation

The time dimension: $K_d, K_{\text{on}}, K_{\text{off}}$ and $1/K_{\text{off}}$

Application: influenza - antibody complexes

Equilibrium constants and reaction rates

- To account for kinetics, one resorts to the reaction rates

$$P + L \underset{K_{\mathbf{off}}}{\overset{K_{\mathbf{on}}}{\rightleftharpoons}} PL. \tag{12}$$

Note that $K_{\mathbf{on}}$ is expressed $mol^{-1}s^{-1}$ while $K_{\mathbf{off}}$ is expressed in s^{-1}. These rates account for the fact that in order to assemble, the molecules must first meet/collide.

- The relationship with dissociation is as follows:

$$K_a = \frac{K_{\mathbf{on}}}{K_{\mathbf{off}}}. \tag{13}$$

Residence times

Binding affinity is a thermodynamic quantity. On the other hand, time is clearly involved in biomolecular interactions – Chapter **??**.

- Mean life of the complex, $1/K_{\text{off}}$: average life span of the PL complex.
- Half-time of the complex, $\log 2/K_{\text{off}}$: the time required for half of a population of complexes to unbind.

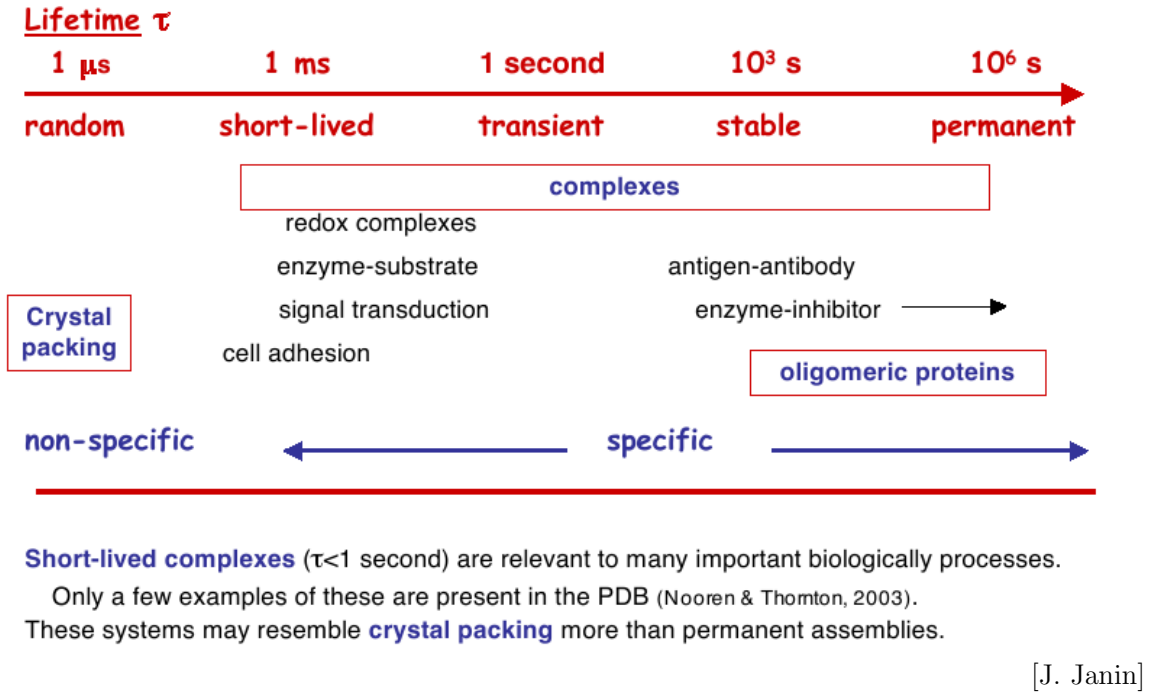


Figure: **Biological complexes: time scales.** From [**?**].

Example: ligand binding for GPCR

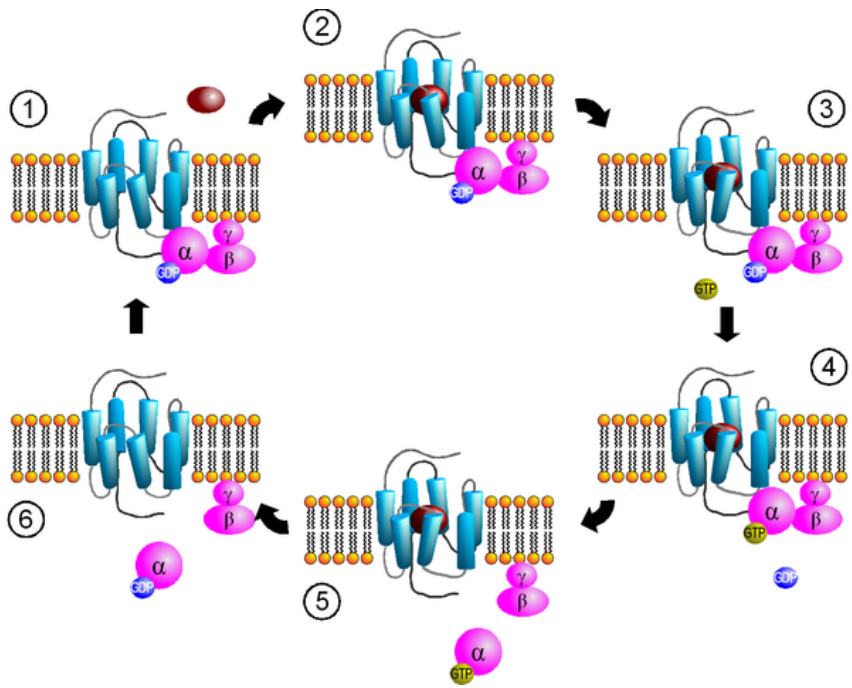
▷ GPCR - protein G complexes: involved in signal transduction inside the cell. The structure of such complexes is as follows (Fig. 10):

- GPCR are receptors involving 7 trans-membrane helices.
- Heterotrimeric G proteins, made of three subunits denoted α, β, γ.

▷ Ligand binding on the extra-cellular side: N-ter region, within the helices.

▷ Triggers (Fig. 10): (1) conformational changes in the cytoplasmic side of the receptor (2) dissociation of the subunits G_α (+GDP) on the one hand, and the dimer $G_{\beta\gamma}$ on the other hand (3) . These trigger signaling cascades

▷ Time constraint: ligand must stay long enough for the conformational change to occur; if not, abortive complex.



Biomolecular recognition

Biomolecular recognition: proteins and binding affinity

Association and dissociation constants K_a, K_d

Enthalpy - entropy compensation

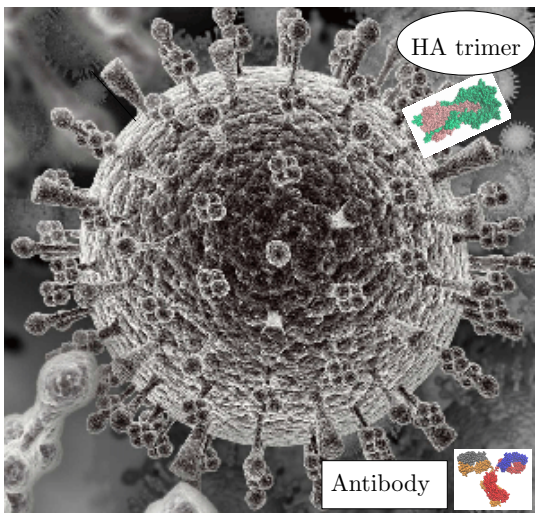
The time dimension: $K_d, K_{\text{on}}, K_{\text{off}}$ and $1/K_{\text{off}}$

Application: influenza - antibody complexes

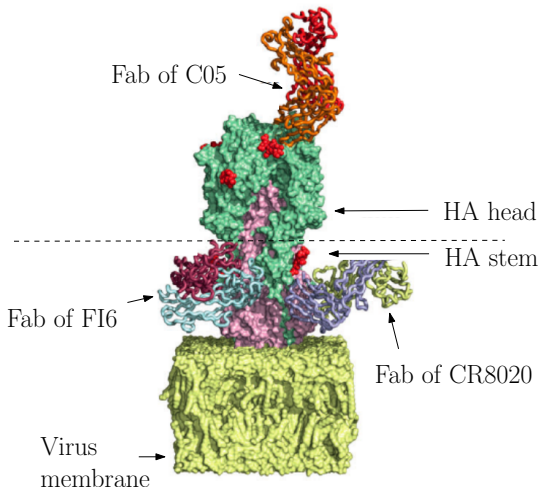
Virus neutralization by antibodies: the problem

- Enveloped viruses: the case of influenza
- Broadly neutralizing antibodies targeting the fusion protein of influenza:
  - Ig on top: prevent the virus attachment
  - Ig on stem: preventing the conformational changes required for envelope-membrane fusion

▷ The influenza virus. Drawn to scale a trimer of the fusion protein (HA)

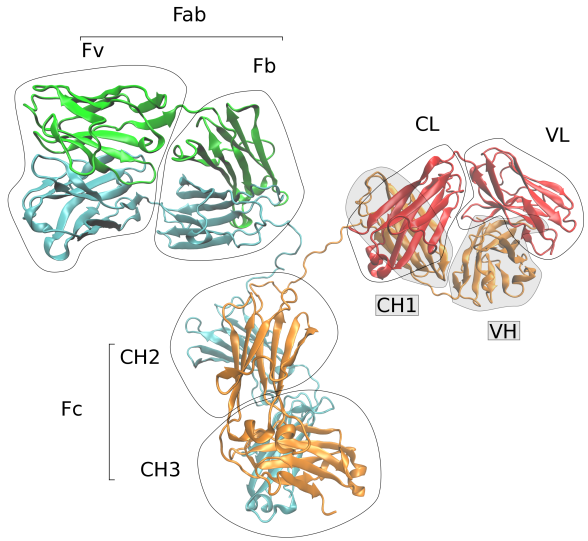


▷ Broadly neutralizing antibodies : hemaglutinin (HA) of influenza is depicted in green

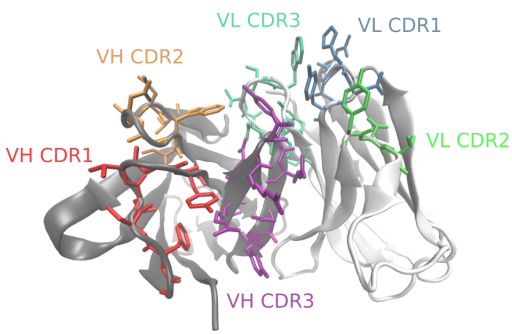


The structure of antibodies – IgG immunoglobulins

- Overall structure



- FABs and CDRs



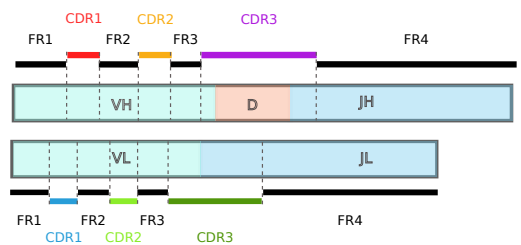


Figure: **(A) Antigen-binding fragment (FAB) and Complementarity Determining Regions (CDRs) (B) Encoding of CDRs and Frs by the V, D and J genes**

Affinity maturation: process

- Affinity maturation: secretion of more potent antibodies
- IgG lineage

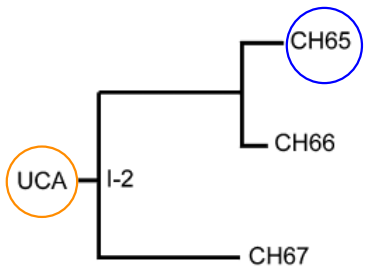


Figure: **Lineage of IgG observed during an immune response against influenza**.

- Evolution of the affinity

Fab	$K_d(\mu M)$
UCA	118 ± 14
I-2	142 ± 15
CH65	$0.49 \pm .10$
CH67	0.36 ± 0.04

Table: **Binding affinities: K_d analysis by SPR** NB: CH65 $\sim$ CH67; wrt UCA: $\Rightarrow \sim$ 200-fold improvement

Affinity enhancement: origin

▷ Ancestor and matured IgG have similar binding modes

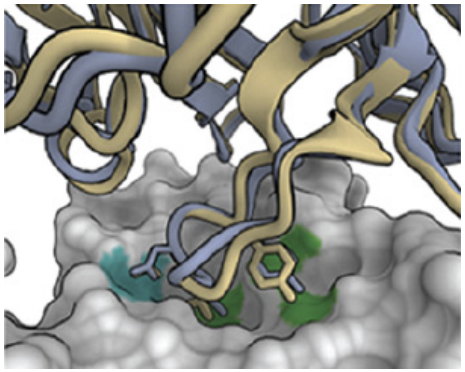


Figure: **But UCA and CH65 have similar binding modes.** Displayed: backbone traces of the CDR3. From [**?**].

▷ But matured IgG have a pre-formed binding site:

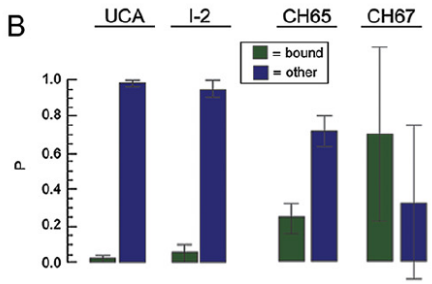


Figure: **CDR3: time spent in bound and unbound conformations.** Maturated IG (CH65, CH67): more time in the bound conformation. From [**?**].

▷ Origin of the affinity enhancement: lesser entropic penalty.
"*In both branches (CH65, CH67)), increased conformational restriction of CDR H3 has been the principle consequence of affinity maturation.* "

Bibliography