Biophysical Chemistry

Applying molecular simulation to biomolecules

Jonathan Doye

jonathan.doye@chem.ox.ac.uk

Introduction

The polymer theory we looked at the last lecture is about the predicting the properties of disordered polymers, where the theory is about "counting" the number of possible states. It's just about entropy.

However, many biopolymers (e.g. proteins, RNA) adopt ordered, well-defined structures that result from the strong and specific interactions between different parts of the polymer, and which are crucial to their function.

Theory is more limited in its ability to describe such structures and the ordering processes that lead to such structures.

An alternative to theory or experiment: Molecular simulation.

Outline:

1. Protein and RNA structure and folding.

2. Molecular simulation: The basics

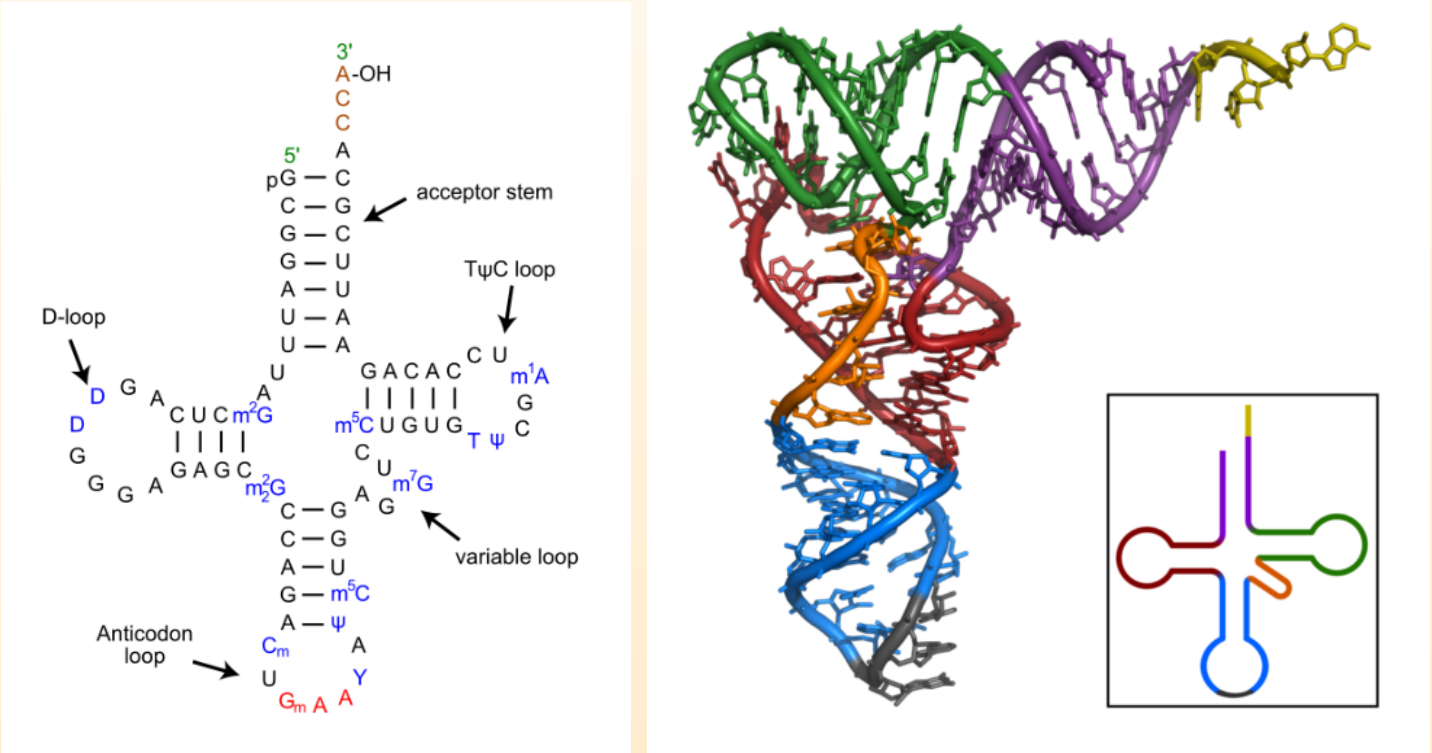
3. Example applications

RNA folding

Primary structure: sequences of bases (A, U, G, C)

Secondary structure: Which bases are base-paired to each other.

Tertiary structure: The full three-dimensional structure of the RNA.



Secondary and tertiary structures of transfer RNA

RNA secondary structure prediction

The interactions stabilizing RNA secondary structures (base-pairing and stacking) are well-defined and thermodynamically characterized.

Together with estimates of the entropy of single-stranded loops and tails, one can estimate the free energy of a particular secondary structure.

By searching through all possible secondary structures one can find the lowest free-energy structure, and the relative probabilities of different structures. e.g. RNAfold.

The predicted structures have reasonable agreement with experiment.

Predicting RNA tertiary structures however is notoriously difficult.

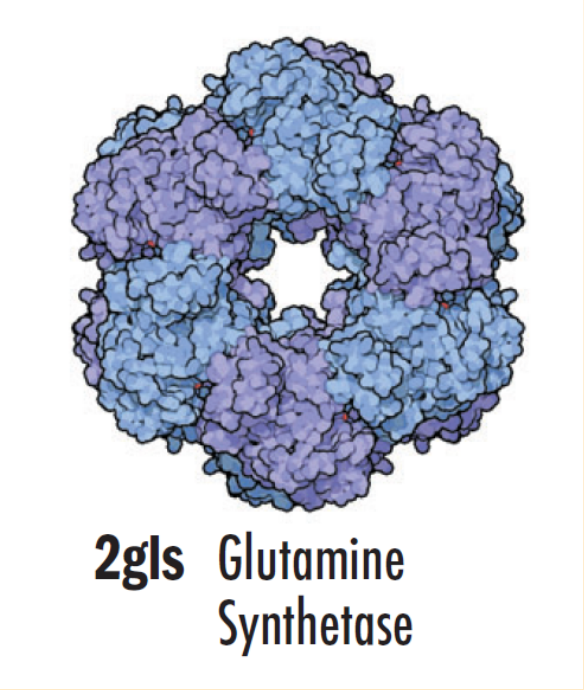
Protein folding

Primary structure: sequence of amino acids

Secondary structure: Motifs such as α helices and β sheets.

Tertiary structure: The full three-dimensional structure of the protein.

Quaternary structure: The assembly of individual proteins into larger complexes.



Many proteins have a compact tertiary structure. For these $R_g \sim N^{1/3}$, rather than $R_g \sim N^{1/2}$.

Both predicting tertiary structure, and understanding how proteins can fold to their native state is both difficult and has been much studied by theory and simulation.

Protein folding: Levinthal paradox

Cyrus Levinthal (1969) estimated that the number of possible conformations for a protein with 150 residues was of the order of 10^{300}. Even if these states were searched at an incredibly fast rate, it would take longer than the age of the universe to find the native state by chance. How then do proteins manage to fold to their native state?

Problems with the paradox:

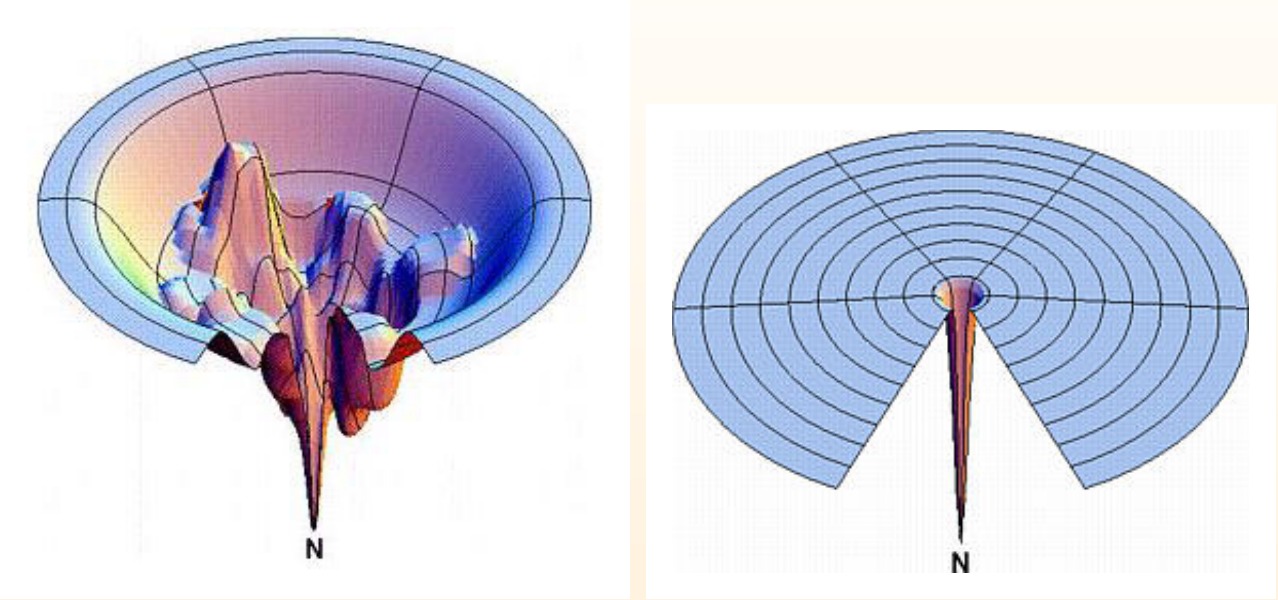
1) Not all states are thermodynamically equally likely. Boltzmann factor $\exp(-\beta E)$ favours lower energy states.

2) The structure of the energy landscape (potential energy or free energy as a function of the coordinates) may make certain states more kinetically accessible than others.

Two extreme landscapes:

1. A funnel. I.e. Generally the system becomes more like the native state as one goes down in energy. The funnel biases the system towards finding the native state.

2. A "Levinthal-like" flat putting green. Just chance if one happens to find the hole.



Extreme landscapes

Proteins have evolved to have landscapes sufficiently funnel-like to fold on reasonable time scales.

Molecular simulations

They are "numerical experiments".

Molecular simulations have two main types

- Molecular Dynamics: $\{\mathbf{r}^N(t), \mathbf{p}^N(t)\}$. I.e. a movie
  n.b. $\mathbf{r}^N = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$
- Monte Carlo: $\{\mathbf{r}_i^N\} \quad \Rightarrow \quad \langle B \rangle$
  I.e. ensembles of configurations, hence give thermodynamics and equilibrium properties

Molecular simulations are a "new" form of science. First reported Monte Carlo simulation (1953), molecular dynamics simulation (1956).

Growth of this science due to growth in computational power (cf Moore's law), advances in algorithms and availability of user-friendly codes.

Complementary to experiment. Although limited in system size (max $O(10^8 - 10^9)$) and time scales (max $O(\mu s - ms)$), the direct visualization of molecular configurations is exactly what is often hardest to achieve in experiment.

Molecular Dynamics

Molecular dynamics simulations involve numerically solving Newton's equations of motion for all the atoms in a system.

It assumes the motion of the nuclei is classical, which is a good approximation for many materials. Quantum effects are most relevant for lighter atoms and lower temperature.

It assumes the nuclei move on a potential energy surface, i.e. Born-Oppenheimer approximation and the separability of nuclear and electron motion.

Trajectories are generated by iteratively integrating Newton's equations forward in time.

Molecular Dynamics: Basic scheme

1. Initialize the system: choose $\mathbf{r}^N(t=0)$ and $\mathbf{p}^N(t=0)$

2. Compute forces

3. Integrate equations of motion: from $t \to t + \delta t$.

4. Repeat 3 and 4 until trajectory of desired length.

Note: a Newtonian system not acted on by any external forces will conserve the total energy. NVE (microcanonical) ensemble not NVT. Same in bulk limit. Temperature can be calculated through the equipartition theorem

$$\left\langle \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} \right\rangle = \frac{3}{2} NkT \tag{1}$$

Need: (i) Initial coordinates and initial velocities (e.g. choose from Maxwell-Boltzmann distribution), (ii) an integration algorithm (iii) An interatomic potential $\mathcal{V}(\mathbf{r}^N)$. ($F_i = -\partial \mathcal{V}/\partial \mathbf{r}_i$).

Integration Scheme

All based on Taylor expansions of position in time.

A simple example. Expand position to third order in time.

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}_i(t) + \dot{\mathbf{r}}_i(t)\delta t + \frac{\ddot{\mathbf{r}}_i(t)}{2}\delta t^2 + \frac{\dddot{\mathbf{r}}_i(t)}{6}\delta t^3 + \mathcal{O}(\delta t^4) \tag{2}$$

As $\dot{\mathbf{r}}_i = \mathbf{v}_i$ and using Newton's 2nd law ($\mathbf{f}_i = m_i\ddot{\mathbf{r}}_i$), rewrite as

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\delta t + \frac{\mathbf{f}_i(t)}{2m_i}\delta t^2 + \frac{\dddot{\mathbf{r}}_i(t)}{6}\delta t^3 + \mathcal{O}(\delta t^4) \tag{3}$$

Similarly,

$$\mathbf{r}_i(t-\delta t) = \mathbf{r}_i(t) - \mathbf{v}_i(t)\delta t + \frac{\mathbf{f}_i(t)}{2m_i}\delta t^2 - \frac{\dddot{\mathbf{r}}_i(t)}{6}\delta t^3 + \mathcal{O}(\delta t^4) \tag{4}$$

Adding these two equations gives

$$\mathbf{r}_i(t+\delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\delta t) + \frac{\mathbf{f}_i(t)}{m_i}\delta t^2 + \mathcal{O}(\delta t^4) \tag{5}$$

Note that due to the cancellation of odd time derivatives of $\mathbf{r}_i$:

- this expression is accurate to third order in time, even though only forces are used.
- velocities do not appear in the expression.

Get velocity, by subtracting Eq. 4 from Eq. 3.

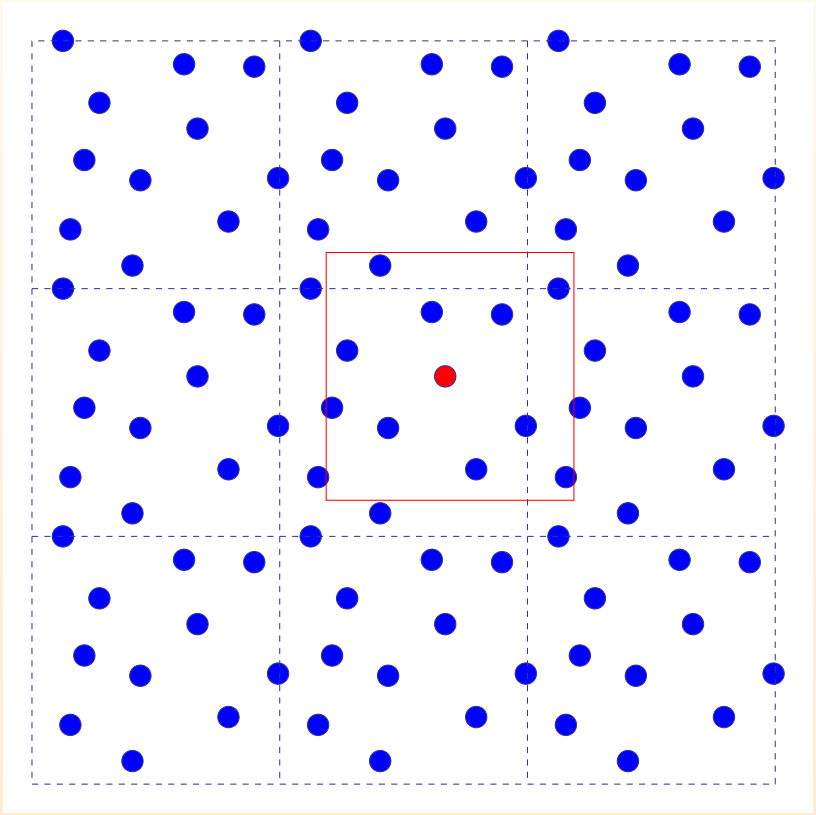
$$\mathbf{v}_i(t) = \frac{1}{2\delta t}\left[\mathbf{r}_i(t+\delta t) - \mathbf{r}_i(t-\delta t)\right] + \mathcal{O}(\delta t^2) \tag{6}$$

This is the Verlet scheme.

Time steps need to be sufficiently small, so that conservation of energy is maintained. Typically O(fs).

Periodic Boundary conditions

Normally, one wants to simulate a bulk system using a simulation of a finite number of particles. But one does not want to introduce a surface (for small systems, the fraction of particles at any surface is high)



One approach is to use periodic boundary conditions.

Particles are usually restricted to only be able to interact with the nearest image of any other particle.

Intermolecular Forces (Reminder)

The main non-bonding interactions are:

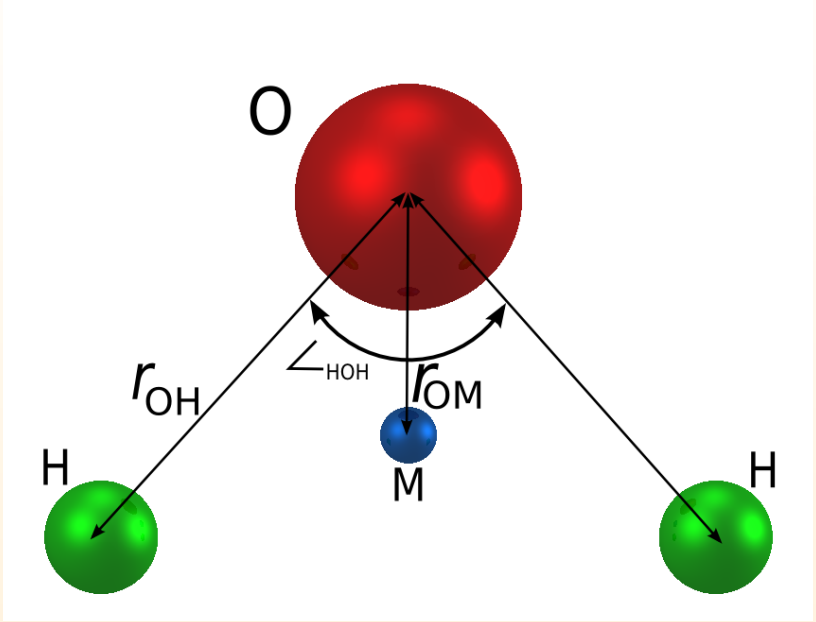
1) Electrostatic:
Interaction between the charge distributions of two molecules.
Can be attractive or repulsive.
Pairwise additive.

2) Induction:
Due to the distortion of the electron density of a molecule by the electric field due to a nearby molecule.
Attractive and non-additive.

3) Dispersion:
Due to correlated fluctuations of the electrons in interacting molecules.
Attractive.
Approximately pairwise

4) Exchange-repulsion:
Due to overlap of charge distributions at short range.
Repulsive because of Pauli principle.

Water Potential: A 4-site model (TIP4P)



Charges $+q$ on two hydrogens.

Charge $-2q$ associated with oxygen at M.

Lennard-Jones interaction ($V = 4\epsilon[(\sigma/r)12 - (\sigma/r)^6]$) centred on oxygen.

Note: Molecule is rigid and not polarizable.

Some of the effects of induction are folded into the charge parameters in an average way. E.g. the static dipole moment in the model is greater than the gas phase dipole moment of water, because induction increases the effective water dipole moment in the liquid.

For historical reasons, the worse TIP3P ($r_{OM} = 0$) is used in most biomolecular force fields.

Biomolecular force-fields

Commonly used force-fields: AMBER, CHARMM.

Forms for potential very simple. E.g.

$$V\left(\mathbf{r}^N\right) = \sum_{bonds} \frac{1}{2} k_b (l - l_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsions} V_n \left[1 + \cos(n\phi - \gamma)\right]$$
$$+ \sum_i \sum_{j>i} \left\{ 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \quad (7)$$

Note: no polarizability; charge distribution simply represented by charges on atoms; non-bonded interactions independent of geometry of bonded framework.

Parameters fitted to a very wide range of data.

Although the force-fields shoud be used with caution, they work surprisingly well.

Appendix

Proof that $\langle z \rangle = -\partial A / \partial F$.

$$Z = \sum_i \exp\left(-\frac{E_i}{k_B T}\right) \tag{8}$$

$$\langle z \rangle = \frac{1}{Z} \sum_i z_i \exp\left(-\frac{E_i}{k_B T}\right) \tag{9}$$

If $E = -Fz +$rest (independent of z)

$$\frac{\partial Z}{\partial F} = \sum_i \frac{z_i}{k_B T} \exp\left(-\frac{E_i}{k_B T}\right) = \frac{\langle z \rangle Z}{k_B T} \tag{10}$$

Therefore

$$\langle z \rangle = k_B T \frac{1}{Z} \frac{\partial Z}{\partial F} = -\frac{\partial A}{\partial F} \tag{11}$$