Macromolecular Docking - project seminar

Peter Mühlbacher - a1253030

March 23, 2015

Contents

Abstract

Weekly reports on advances, encountered difficulties and implementations for the seminar on artificial intelligence will be gathered in this document.

Goals

In this seminar I aim to examine the problem of (flexible) protein-protein docking.

1 Week 1

Originally I wanted to approach the problem of protein folding by investigating the asymptotic ($t \to \infty$) behaviour of the probability density function of a given protein. As the changes over time of a given protein[1] can be modelled

[1] understood as a single point in high dimensional space

by a Langevin equation $\dot{x} = -\nabla U(x) + \beta \dot{w}$, the corresponding probability density function $p(t, x)$ is described by the forward Fokker-Planck equation $\frac{\partial p}{\partial t} = \tilde{\beta} \Delta p + \operatorname{div}(p \nabla U)$ whose asymptotic behaviour has been investigated in Nadler et al. (2008).

However, there are (at least) two reasons why this does not work:

- In order to get a meaningful low dimensional representation of the system, a considerable margin between two adjacent eigenvalues $\lambda_{k+1} \gg \lambda_k$ of the Fokker-Planck operator is needed and it can be shown that this roughly corresponds to having a potential function U with k local minima[2]. Taking these properties into consideration it seems futile to employ this method of dimensionality reduction as the potential function in the area of protein folding has thousands of local minima.

- Even if solutions could be found explicitly in a reasonable amount of time the problem of finding and exact potential function U still remains unsolved[3]. As we are working with approximations it does not seem to make any sense trying to study its asymptotic long-term behaviour.

As a result of above considerations my topic of this seminar was restricted to docking. In the special case of rigid docking both molecules are assumed to be in a metastable state which drastically cuts the the complexity of the state space.

2 Week 2

2.1 Benchmarks

Visit `http://zlab.umassmed.edu/benchmark/` for a benchmark on various docking problems that are ordered by difficulty. There are test cases for rigid docking as well as flexible docking, again ordered by amount of typically changing parameters like torsion angles.

2.2 PDB

The *Protein Data Bank Format* is a plain text data file, storing (amongst other information) the coordinates of every single atom in a protein. As most

[2] c.f. Nadler et al. (2008), p.9

[3] c.f. Neumaier (2006), p.14ff

other atomic coordinate files PDB does not specify covalent bonds between atoms unless the two atoms are not members of the *Standard Residues*[4] in protein or nucleic acid chains. Typically, any two non-hydrogen atoms within 1.9 Ångstroms of each other are deemed to be covalently bonded. The distance for a bond involving a hydrogen atom is less.

For a list of PDB files for various proteins visit `http://www.rcsb.org/`.

2.3 Modelling the Potential

Given two metastable molecules $M_i = (\mathbf{x_i}, \alpha_i, \beta_i, \gamma_i, \{\mathbf{a_{ij}}\}_{j=1}^{n_i})$, $i \in \{1, 2\}$. In the following $\mathbf{x_i} \in \mathbb{R}^3$ will be called the molecule's (or the system's) point of reference, $\alpha_i, \beta_i, \gamma_i \in [0, 2\pi)$ the system's rotation (where α_i corresponds to the rotation around the x-axis relative to its point of reference; analogously for β_i, γ_i) and $\mathbf{a_{ij}} \in \mathbb{R}^3$ the system's atoms with (absolute) coordinates $\mathbf{x_i}+\mathbf{a_{ij}}$, where n_i is the number of atoms.

The Lennard-Jones potential of a system consisting of two points $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$ with distance $r^2 = ||a - b||$ is defined as

$$V_{LJ}(r) = \varepsilon \left(\left(\frac{\sigma}{r} \right)^{12} - 2 \left(\frac{\sigma}{r} \right)^{6} \right) \tag{1}$$

where ε is the size of the potential well and σ the distance at which the potential reaches its minimum. Both are usually given by experiments.

In order to investigate the energy landscape for a docking problem it is natural to consider the Lennard-Jones potential for a system consisting of two proteins M_1, M_2 given by

$$\tilde{V}_{LJ}(M_1, M_2) = \sum_{i=1}^{n_1+n_2} \sum_{j>i}^{n_1+n_2} V_{LJ}(r_{ij}) \tag{2}$$

if we set r_{ij} to be the distance between the ith and jth atom[5].

If we want to explicitly calculate its gradient with respect to the free parameters $\mathbf{x_1}, \alpha_1, \beta_1, \gamma_1$ it is instructive to write down the entire formula as a function of those variables. First of all, to faciliate computation, we split up the summation into three parts $\tilde{V}_{LJ}(M_1, M_2) = \tilde{V}_{LJ}^1(M_1) + \tilde{V}_{LJ}^2(M_2) +$

[4] For further information consult `http://www.wwpdb.org/documentation/file-format-content/format33/sect10.html` and `http://www.wwpdb.org/documentation/file-format-content/format33/sect4.html`.

[5] $i, j \in \{1, \ldots, n_1 + n_2\}$

$\tilde{V}^3_{LJ}(M_1, M_2)$, where $\tilde{V}^1_{LJ}$ refers to all combinations of atoms of M_1, $\tilde{V}^2_{LJ}$ to those of M_2 and $\tilde{V}^3_{LJ}$ to the summation over all pairs $(a_{1i}, a_{2j})_{i\in\{1,\ldots,n_1\}, j\in\{1,\ldots,n_2\}}$. It is apparent that $\tilde{V}^1_{LJ}$ and $\tilde{V}^2_{LJ}$ do not change as a function of the arguments chosen above (however, they would in case we also introduced variable torsion angles and such). To calculate the gradient it thus suffices to inspect $\tilde{V}^3_{LJ}$ as a function of the free parameters $x, y, z, \alpha, \beta, \gamma$ describing the position of M_1:

$$\tilde{V}^3_{LJ}(\mathbf{x}_1, \alpha_1, \beta_1, \gamma_1) = \varepsilon \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\frac{\sigma}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma}{r_{ij}}\right)^{6} \tag{3}$$

where $r_{ij}^2 = \|\tilde{a}_{1i} - a_{2j}\|$ and $\tilde{a}_{1i}$ are the first system's ith particle's absolute coordinates given by

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_1 & -\sin\alpha_1 \\ 0 & \sin\alpha_1 & \cos\alpha_1 \end{pmatrix} \begin{pmatrix} \cos\gamma_1 & 0 & -\sin\gamma_1 \\ 0 & 1 & 0 \\ \sin\gamma_1 & 0 & \cos\gamma_1 \end{pmatrix} \begin{pmatrix} \cos\gamma_1 & -\sin\gamma_1 & 0 \\ \sin\gamma_1 & \cos\gamma_1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{a}_{1i} + \mathbf{x}_1$$

.

Now one can explicitly compute the partial derivatives $\frac{\partial \tilde{V}_{LJ}}{\partial p} = \frac{\partial \tilde{V}^3_{LJ}}{\partial p}$, for p being one of the parameters.

As of now we can compute the potential $\tilde{V}_{LJ}(\mathbf{X})$ for a given point $\mathbf{X} = (\mathbf{x}_1, \alpha_1, \beta_1, \gamma_1)$ in the state space and its gradient $\nabla \tilde{V}_{LJ}(\mathbf{X})$ and that is all we need for algorithms like steepest descent (which would still require rescaling of the parameters as to avoid extensive zig-zagging) or diffusion maps.

Furthermore, note that the computation of the potential and the gradient can be combined into one double loop, making it $O\left((n_1 + n_2)^2\right)$.[6]

2.4 Implications for Flexible Docking

Given the file formats' property of not storing information of covalent bonds I expect the problem of flexible docking (i.e. some (e.g. torsion) angles are considered to be free variables as well) to be much harder as the gradient can hardly be computed explicitly anymore and numerical methods of differentiation may be necessary. In order to make them work reasonably fast it is not only important to consider the potential function, but also the function

[6] although the computation of the gradient is likely to result in a big constant factor of operations

computing the coordinates of the entire system, given a (small) change of some (torsion) angle. However this is just (part of) the problem of protein folding and as such very likely to be very hard.

Another problem is that $\tilde{V}_{LJ}^1(M_1)$ and $\tilde{V}_{LJ}^2(M_2)$ would not be constants anymore and thus a considerable amount of additional computation is necessary.

Finite element methods are also less likely to be applicable as the number of the state space's dimensions increases when adding more free variables.

References

Nadler, Boaz, Ronald R. Coifman, Ioannis G. Kevrekidis, Stéphane Lafon, and Mauro Maggioni (2008), "Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems." *Society for Industrial and Applied Mathematics*, 842864, URL `http://www.wisdom.weizmann.ac.il/~nadler/Publications/diffusion_map_MMS.pdf`. Multiscale Model. Simul.

Neumaier, Arnold (2006), "Molecular modeling of proteins and mathematical prediction of protein structure." URL `http://www.mat.univie.ac.at/~neum/ms/protein.pdf`.