Term Paper: The Energy Landscape of Biomolecules - Protein Folding

Charlotte Voigt
Albert-Ludwig University Freiburg
Presented: 10.1.2017 Supervisor: Prof. G. Stock

This article describes some aspects of the history and the recent research in protein folding. After general considerations about proteins, Levinthals paradox is described by a simple mathematical model, based on the work of Zwanzig in 1992. In the paper, energy landscapes are introduced and several examples are discussed, mainly led by the work of Dill in 1997. Finally, some of the problems that can occur when producing energy landscapes are observed, following Sittels paper from 2014.

1 Proteins and folding

Proteins are large biomolecules. In the human body, proteins consist of 20 different amino acids and fulfill many important functions. They work as antibodies, enzymes, hemoglobin or many other highly specialised objects in the organisms. If proteins are not in the right structure, they often cannot fulfill their function which can cause hereditary diseases.
When proteins get synthesized by ribosoms, they are just a sequence of amino acids. In this state, they don't fulfill functions. Until they reach their so called native structure, they have to go through a complicated dynamic process of folding. Many proteins even have to bind to each other to finally fulfill their function. This paper is about the process of folding.

2 Levinthals Paradox

Since protein folding is such a complex process with so many degrees of freedom, Levinthal stated in 1968 [5] that it should take an incredible long time for proteins to fold to their native state.
Let's consider an example of a protein with 101 amino acids that are connected via 100 bonds. If each of these bonds can be in 3 different states, there are $3^{100} \approx 5 \cdot 10^{47}$ configurations of the protein. For a protein that randomly changes bonds to search for its native state, even with a sampling rate of $10^{13}\frac{1}{s}$, it would take 10^{27} years to take all possible configurations. This has been seen as a paradox, since biological experiments show that proteins can fold on time scales of seconds or milliseconds.

2.1 A simplified theoretical model

This section describes a highly simplified mathematical model of a protein, presented by Zwanzig [5] to show that Levinthals paradox is easy to solve if interactions between amino acids are taken into account.
This model describes a protein as a sequence of $N+1$ amino acids connected by N bonds. The states of the bonds can either be correct c or incorrect i. Bonds can be incorrect for various reasons which are summarized in one statuts group.
Starting with a random sequence (c, i, i, c, ...), the bonds can change their state with rates of

$$k_0 \text{ for c} \rightarrow \text{i}$$
$$k_1 \text{ for i} \rightarrow \text{c}$$

Defining the number of incorrect bonds as S, the number of correct bonds is $N-S$ and the probability of having S incorrect bonds at a time t, given an initial time t_0 and an initial number of incorrect bonds S_i , can be written as $P(S,t|S_i,t_0) =: P(S,t)$. A master equation describes the probability of gains and losses of a state S. In this case, the master equation can be written as

$$\begin{aligned}\frac{d}{dt}P(S,t) =& \overbrace{(N-S+1)k_0}^{\text{rate}(S-1\rightarrow S)}P(S-1,t) + \overbrace{(S+1)k_1}^{\text{rate}(S+1\rightarrow S)}\,P(S+1,t)\\ &-\underbrace{(N-S)k_0}_{\text{rate}(S\rightarrow S+1)}P(S,t) - \underbrace{Sk_1}_{\text{rate}(S\rightarrow S-1)}\,P(S,t).\end{aligned} \tag{1}$$

Each term of equation 1 consists of the number of bonds being in a neighbouring state, the rate of a single bond to change from the neighbouring to the considered state S and the probability for a bond to be in the neighbouring state.
Equation 1 can easily being put into matrix form by defining the vector $P(t) = (P(S=0,t),...,P(S=N,t))^T$ which will be useful for the following calculations. It is likewise helpful to use absorbing boundary conditions at the final state S_f. This means that the probability for leaving the final state S_f is equal to zero.
Defining the survival probability $s(t,S_f|S_i)$ by

$$s(t,S_f) = \sum_{S\neq S_f} P(S,t) = 1 - \int_0^t F(t';S_f)dt', \tag{2}$$

the mean first-passage time $\tau(S_f|S_i)$ is given by the mean of the probability distribution $F(t;S_f)$.

$$\tau(S_f|S_i) = \int_0^\infty tF(t;S_f)dt \tag{3}$$

This describes the mean time for a protein to reach state S_f from starting state S_i.

Equations 2 and 3 can be inserted into the matrix form of equation 1 to obtain an expression for the mean first-passage time. For large N and not too small k_0, τ can be approximated as

$$\tau(S_f = 0|S_i) \approx \frac{1}{Nk_0}\left(1+\frac{k_0}{k_1}\right)^N. \tag{4}$$

This is actually similar to what Levinthal stated in his paradox. Nk_0 can be associated with a sampling rate and $\left(1+\frac{k_0}{k_1}\right)^N$ with the number of possible states.

At this point, the ratio $\frac{k_0}{k_1}$ is interpreted as an equilibrium constant by considering the kinetic scheme of a single bond

$$\frac{d}{dt}[c] = -k_0[c] + k_1[i], \qquad [c]+[i] = 1$$

at equilibrium ($\frac{d}{dt}[c] = 0$). Then $\frac{k_0}{k_1} = \frac{[i]_{eq}}{[c]_{eq}} =: K$.

This equilibrium constant can be found using thermodynamics. As described before, there is only one possibility for a bond to be in the correct state c, but many possibilites to be incorrect. It is additionally assumed, that incorrect bonds have higher energy than correct bonds. The energy of a correct bond is defined as ε_c, the energy of an incorrect bond $\varepsilon_c + U$ and the degeneracy of the incorrect state ν. This leads to the equilibrium constant

$$K = \frac{k_0}{k_1} = \frac{\nu e^{-(\varepsilon_c+U)/kT}}{e^{-\varepsilon_c/kT}} = \nu e^{-U/kT}$$

that depends on U, the energy penalty for incorrect bonds. Here, k is the Boltzmann constant and T the temperature. Using these considerations, equation 4 becomes dependent on the energy penalty U.

$$\tau(S_f = 0, S_i) \approx \frac{1}{Nk_0}(1+K)^N$$

The dramatic effect is shown in figure 1, describing that an energy penalty of a few kT changes the mean first-passage time in several orders of magnitude. With this choice of parameters, $\tau(0,S_i)$ becomes biologically significant (of the order of 1 second) when U is about $2kT$. Taking an energy penalty into account is a physically reasonable approach as seen in section 3 and shows what Levinthal missed when he stated his paradox. The folding of a protein is not a random search.

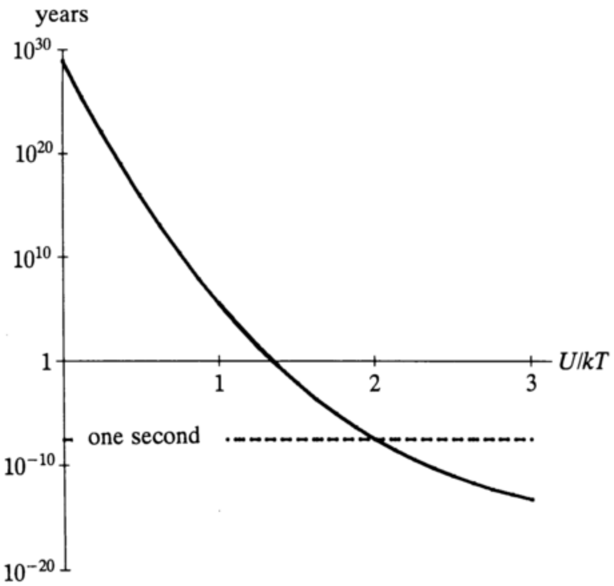

Fig. 1. Mean first-passage time τ in years. With $N = 100$, $\nu = 2$, $S_i = 66$, $k_1 = 10^9 s^{-1}$ and $k_0 = 2\exp(-U/kT)\cdot 10^9 s^{-1}$. [5]

3 Computational Modelling

A more realistic model of biological molecules and processes is given by the so called force fields, used in computational modelling. These force fields U are used in solving Newtons equations of motion

$$m_i \frac{\partial^2 \vec{r}_i}{\partial t^2} = \vec{F}_i \tag{5}$$

$$\vec{F}_i = -\frac{\partial U(\vec{r}_1,...,\vec{r}_N)}{\partial \vec{r}_i} \tag{6}$$

where $i = 1,2,...,N$ denotes the $i'th$ of N atoms (or clusters of atoms) in the system.

A typical force field would be

$$U = \sum_{\text{bonds}} K_r(r-r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}\left(1+cos(n\phi-\delta_n)\right) + \sum_{i<j}\left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_iq_j}{\varepsilon R_{ij}}\right)$$

containing bond and angle interactions quadratically, dihedral and non-bonded interactions such as Lennad-Jones and electrostatic interactions. All constants in this force field depend on the associated atoms.

With computational modelling, it is possible to simulate the folding of a protein step by step and to observe each single atom at each time step. This is different from biological experiments where only macrostates can be observed.

Since it is computationally costly, the modelling approach is limited on smaller proteins and not too long time scales.

Nevertheless for those proteins that can be simulated, new methods of analyzation become possible. One of these methods is described in the following.

4 Energy Landscapes

Energy landscapes are a tool to analyze protein folding. They plot the free energy of a protein over its degrees of freedom. For example, the cartesian r_1-coordinates of two atoms of the protein might be taken as the axes of the energy landscape. The free energy of the coordinate x of a system is given by

$$\Delta G(x) = -kT \ln P(x).$$

where P(x) denotes the probability distribution of the coordinate x. To get the free energy correctly, the potential of mean force would have to be calculated, which is an integral over all degrees of freedom. This is hard to manage, so usually a histogram of the coordinate of interest is associated with a probability distribution.

4.1 Levinthals Paradox in Energy Landscape terms

As a first example of an energy landscape, consider figure 2. This shows a random search of a protein for its native state N described in Levinthals paradox. Since there is a large flat space of configurations, the protein never 'knows' whether a taken step was a step into the right direction. In this flat area, the protein wanders around slowly and aimlessly.

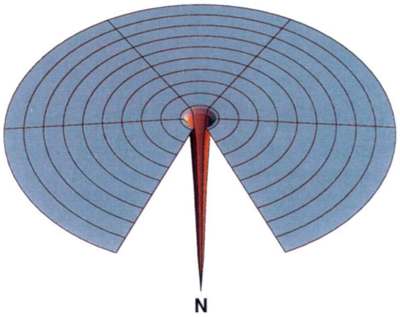

Fig. 2. Energy landscape of a random search of a protein as in Levinthals paradox. [2]

Since Levinthal also noticed that his random search was unbiological, he thought about solutions and suggested the concept of pathways as presented in figure 3. This pathway solution is oriented at the process of chemical reactions. Simple chemical reactions always follow the same path, so the steps of the process has a defined order. Levinthal and others tried to impute this to protein folding. In this view, the protein starts at an initial state and changes its bonds, dihedrals and atom positions in a fixed order of steps which finally leads to the native state. Every step out of the tunnel is seen as an off-pathway step which is useless and has to be inverted to come back onto the right pathway.

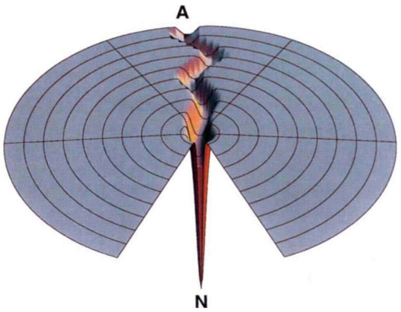

Fig. 3. Energy landscape of the pathway solution of the random search problem [2]

This model is not according to experience from experiments. Proteins will barely fold in exactly the same way twice. One reason for this is the low change of energy in a single step in protein folding. In chemical reactions, each change of energy is much larger than kT, so brownian motion doesn't affect the process very much. Since in protein folding, the changes of energy are not much larger than kT, brownian motion has a large influence on the variability of folding processes.

4.2 Energy landscapes are funnel-like

In reality, energy landscapes of protein folding can be described as funnel-like. Concidering the simple mathematical model described in section 2.1, the corresponding energy landscape looks like the one in figure 4. Each step towards the native state (so changing a bond from incorrect to correct) leads to a slightly smaller energy. Therefore, the native state can be reached in biologically significant times. But as the model in 2.1 is highly simplified, so is the energy landscape in figure 4. A more realistic landscape is shown in figure 5. This energy landscape contains a global energy minimum which is the native state. But is also consists of local minima, energy barriers and flat areas. This complexity is due to the complexity of the folding process with its many degrees of freedom.

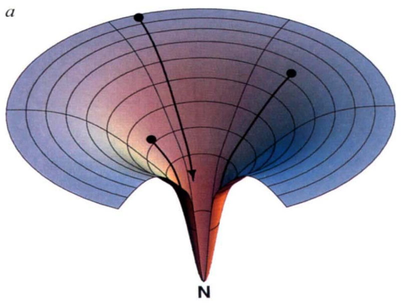

Fig. 4. A funnel-like energy landscape [2]

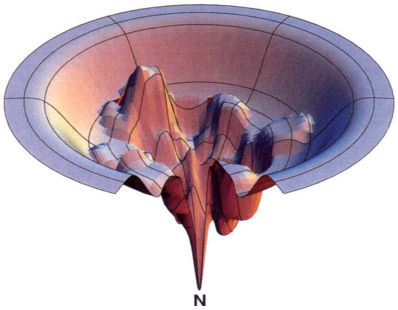

Fig. 5. A more realistic, bumpy funnel-like energy landscape [2]

4.3 Dimensionality reduction

An important step in producing energy landscapes is the reduction of dimensions. For a protein of N atoms, there are $3N$ cartesian coordinates, $N-1$ bond stretch vibrations, $N-2$ bend vibrations or $N-3$ dihedral angles. The data obtained by computational modelling can be expressed in these coordinates. It is difficult to interpret data in these high-dimensional spaces which makes a reduction necessary. One possibility is to define a single reaction coordinate such as the number of native contacts, the radius of gyration or the root mean square distance from the native structure. For the model in section 2.1, the number of correct bonds would be an obvious choice. Some of these natural coordinates can also be obtained from experiments. Nevertheless, in using just a single coordinate, a good part of information is not used.

This is why it is usually tried to reduce the large number of dimensions to just as many dimensions as necessary to describe the main features of the folding. There are many problems that can occur during the reduction process. For example, connectivities between energy minima can easily get lost. In the following, the focus is on a problem described in [4], that occurs when cartesian coordinates are used in large molecules.

Generally, there are many ways to chose the 'important' dimensions to be kept. Here, the dimensions are reduced via principal component analysis (PCA), where a basis transformation via diagonalization of the covariance matrix σ_{ij} is performed. This projects the data into the direction of the greatest variance while the first principal components carry the most variance.

In a cartesian coordinate PCA, first it is necessary to separate overall and internal motion. Therefore, the center of coordinate system is set to the proteins center of mass $\sum_i m_i \vec{r}_i = 0$. The atomic velocities are decomposed via $\vec{v}_i = \vec{u}_i + \vec{\omega} \times \vec{r}_i$ into a vibrational and an angular part. This leads to the kinetic energy

$$\begin{aligned} T &= \frac{1}{2}\sum_{i=1}^{N} m_i \left(\vec{u}_i^2 + (\vec{\omega}\times\vec{r}_i)^2 + \vec{\omega}(\vec{r}_i \times \vec{u}_i)\right) \\ &= T_V + T_R + T_{RV}\,. \end{aligned}$$

To get rid of the rotational part of T, the overall motion gets removed via $\vec{r'}_i = R\vec{r}_i$, by minimizing $\sum_{i=1}^{N} m_i(\vec{r'}_i - \vec{\bar{r}}_i)^2$ with respect to a reference structure $\vec{\bar{r}}$, which can be any state of the protein but is usually chosen as the native state obtained by simulation or experiment.

This procedure works well if observing small rigid structures. For larger molecules as proteins, the overall motion is barely seperated as shown later.

Alternatively, [4] suggests the use of internal coordinates such as dihedral angles ϕ_n. Since these are circular coordinates, they first get transformed into linear ones via

$$\begin{aligned} q_{2n-1} &= cos\phi_n \\ q_{2n} &= sin\phi_n \end{aligned}$$

which doubles the number of coordinates.

On the transformed cartesian or internal coordinates, a PCA can now be performed. The result of both methods when performed on the folding of the villin headpiece is shown in figure 6. It can be seen that the PCA on internal coordinates show much more minima and energy barriers that are not resolved in the PCA on cartesian coordinates. This is due to the fact, that the internal motion is not properly removed for the cartesian coordinate PCA.

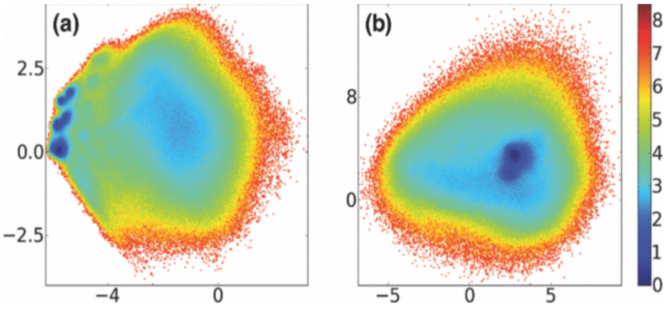

Fig. 6. Free Energy landscape of the folding of the villin headpiece HP35 obtained from a PCA using a) internal coordinates and b) cartesian coordinates [4]

5 Summary

Protein folding is a very complex process which plays an important role in biology. It has been shown that the problem of Levinthals paradox can be resolved by considering an energy penalty for incorrect bonds. This turns an unguided random search into a directed search, which can be described by an energy landscape. Energy landscapes describe the folding process in an illustrative way which allows interpretation. Different kinds of energy landscapes have been discussed. It has been shown that it is important not to lose information during the production of energy landscapes.

References

[1] K. A. Dill. Polymer principles and protein folding. *Protein Science*, 8:1166–1180, 1999.

[2] K. A. Dill and H.S. Chan. From levinthal to pathways to funnels: The ”new view” of protein folding kinetics. *Nature Structural Biology*, 4:10–19, 1999.

[3] K. A. Dill and J. L. MacCallum. The protein folding problem, 50 years on. *Science*, 338:1042–1046, 2012.

[4] A. Jain F. Sittel and G. Stock. Principal component analysis of molecular dynamics: On the use of cartesian vs. internal coordinates. *J. Chem. Phys.*, 141:014111, 2014.

[5] A. Szabo R. Zwanzig and B. Bagchi. Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 89:20–22, 1992.