A Lattice Simulation Approach to Protein Folding

Joseph Bullock

January 31, 2018

Abstract

Proteins play a vital role in everyday bodily function, thus understanding their behaviour is paramount to biological, pharmaceutical and medical research. A protein can only perform correctly if it is in a biologically 'active' state, which requires the structure to be precisely folded into the right shape. This is the subject of Levinthal's paradox, which suggests that proteins need to calculate the energies of an enormous number of folded states to determine the native local minimum. In this paper, protein folding is modelled using lattice simulation Monte Carlo techniques, which discretises the folding process. This mechanism presents new insights into the problem of how a protein folds so rapidly, and significantly reduces the computational time compared to that put forward by Levinthal.

Introduction

Proteins are critical to the functioning of biological life. They are complex molecules made up of amino acids (monomers) that fold to form unique 3 dimensional structures which can act as: antibodies, enzymes, messengers etc. [1]. A protein's folded structure defines it's functionality; they are either biologically 'active', when folded into the correct shape for their task (*tertiary state*), or biologically 'dead'. Dead proteins can be not only waste bodily resources, but may also be dangerous if they block pathways and channels for other proteins and molecules crucial for bodily processes [2].

Proteins are constructed out ∼20 to ∼1000 amino acids, which come in 20 unique types, bonded together by covalent bonds in a specific order (*primary structure*), forming a polypeptide chain [1]. Along with forming covalent bonds between neighbouring chain monomers, amino acids feel other attractive and repulsive forces between each other, as well as with the fluid they are in; Van der Waals forces, and Hydrogen bonds are attractive at short distances, and some amino acids are hydrophobic or hydrophilic [2]. For a given primary structure, a protein will fold into a unique state [3], which (at least locally) minimises its energy. Moreover, entropic considerations must be accounted for given the number of possible monomers in a protein chain. The difference between the entropic preference for an unfolded, less chaotic, structure and the attractive interactions favouring a folded molecule, is of the order of 5-10 kcal mol^{-1} [2][1]. Therefore the stability of the protein structure is very fragile.

In 1961 Anfinsen and Haber demonstrated that a protein, once unfolded, will reform back to the same shape every time [4]. Indeed, the refolded protein remains biologically active, meaning the information about the folded shape had not been lost. This discovery posed the question of how the protein 'knows' how to fold to get back to its tertiary structure. This problem was raised by Levinthal in 1969 when he noted that there were an enormous number of possible molecular conformations which a protein could, potentially, fold back into (one estimate came to 10^{143} possible configurations) [5]. Additionally, Levinthal stipulated that the real folding process only takes a matter of milliseconds to complete, which has been confirmed by modern experiments (e.g. [6]).

Levinthal's proposed 'paradox' assumes that a protein tries every possible state in its search for a stable structure. However, this is clearly unrealistic, and ignores many subtleties related to a protein's desire to minimise its energy during folding [7]. To overcome the enormous computational challenge a protein would face in examining the energy of each possible final molecular structure, Levinthal suggested that a protein followed a well defined, unique, pathway to energy minimisation [5]. Such a route, he proposed, would guide the protein to a biologically active structure and minimised the chance for making mistakes along the way. Many theories have been posed in an attempt to resolve Levinthal's paradox, however, it has also, in large part, been down to the use of computer modelling and, in particular, lattice simulations (in which space is discretised) that have led to a better understanding of the folding mechanism [8].

Monte Carlo processes used in these models clearly do not use one unique pathway for folding, due to their random nature, but rather the protein undergoes folding on many different pathways simultaneously with "search biased toward the native state[2] in an essential way by the variation of the effective energy of the polypeptide chain as a function of its conformation" [9]. This has challenged and generalised Levinthal's original solution and likens the process of folding more to that of a chemical reaction with numerous intermediate states.

Although this methodology allows for easier computation, the issue of multiple local energy minima is still unresolved. A protein's native state does not necessarily a global energy minimum, however, the protein has to find the 'correct' minimum to be biologically active. Finding this is non-trivial and a simulation can get 'stuck' in a local, but non-native, energy minimum, expending much computational time oscillating around this point in phase space. Sali *et al.* found that this is less problematic in the case of small proteins where the native state energy minimum is well pronounced [3], however, in the case of large proteins, these minima are generally less deep. Further simulation techniques are being developed to help overcome this problem, and others, which will be discussed in more detail.

[1] In comparison, the thermal energy of a mole of gas at 298K (room temperature) is ≈ 0.6kcal mol^{-1}.

[2] Here the native state is taken to imply a biologically active state.

Methodology

To examine the benefits of the lattice simulation approach, a generalised program was designed upon which optional input parameters may be specified, to allow for greater user flexibility [10]. The protein is initially generated on a 3 dimensional lattice space by specifying the chain length and then stating, or randomly generating, a primary structure. An unfolded (straight) structure can then be initialised, however, it can also be beneficial to start with an already folded arrangement to analyse how the formation develops from such initial conditions. In the case of the latter arrangement, the positions of the amino acids are determined by a non-crossing random walk.

Two factors, mentioned above, impact on whether the protein is allowed to fold: the energy of the post-folded structure relative to its pre-folded state, and the entropic random energy fluctuations. The energy of a given structure is determined by:

$$E = J^i_j \, \delta^j_i \, , \tag{1}$$

where J^i_j is the interaction energy between two amino acids of types i and j, and:

$$\delta^j_i = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are nearest neighbours}^3 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The values of J^i_j incorporate the different forces of attraction and repulsion described above and can be represented in a symmetric 20 × 20 matrix. Due to the complexity involved in determining these energies, biologists often take these values to be randomly chosen from a uniform or Gaussian distribution [3]. However, to allow for specific field theoretic results to be incorporated into the programme, the values calculated by Miyazawa and Jernigan using a quasi-chemical approximation have been made available (see [11] Table VI[4]).

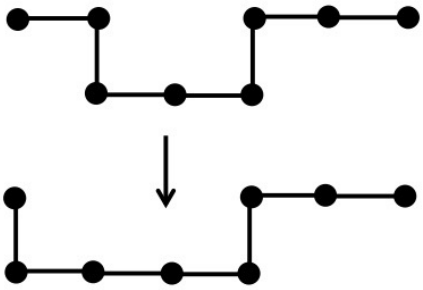

Figure 1: Example of an amino acid moving in 2 dimensions, whilst not breaking the covalent bonds between itself and chain neighbours. Here the second monomer in the chain moves diagonally, all the other amino acids remain fixed, and the Euclidean distance between chain neighbours stays equal to 1.

In each Monte Carlo time step the programme selects, at random, an amino acid to determine if it can be 'moved'. A move constitutes repositioning the selected monomer to one of its neighbouring (unoccupied) lattice positions, whilst not breaking the bonds between the two amino acids adjacent in the primary structure (i.e. keeping $(x_0^2 - x_i^2) + (y_0^2 - y_i^2) + (z_0^2 - z_i^2) = 1$, $i = \{-1, 1\}$, where indices denote the position of a given amino acid in the chain, centred about the randomly selected monomer, denoted by the 0 subscript), as depicted in Fig.1. A move can only be made if one of two conditions are met:

1. The difference in the energy of the protein after the proposed move and its current state, ΔE_{move}, is negative (i.e. the new energy is less than the current).

2. $\exp\left(\frac{-\Delta E_{\text{move}}}{k_B T}\right) > r$, where Boltzmann's constant, k_B, is set to unity, T is the temperature of the solution, and r is a randomly generated number between 0 and 1.

The second criterion corresponds to calculating the probability that the system, represented by a canonical ensemble of energy microstates, makes an energy transition, and then tests to see if the system makes such a transition, given that probability. This condition accounts for the statistical uncertainty in the system due to entropic considerations. If one of the conditions is met then the amino acid is moved and the process repeats again for as many time steps as specified by the user.

Results

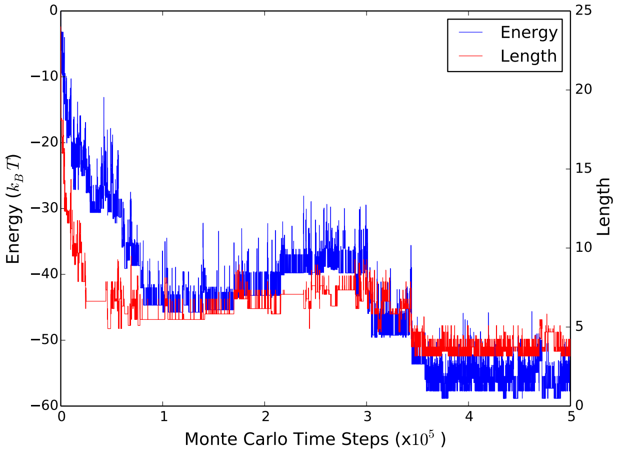

Figure 2: Energy and length, as a function of Monte Carlo time, of a protein with chain length 25, at a temperature of 1, embedded within 3D lattice space. The interaction energy matrix, J^i_j, has been filled with randomly chosen values from a uniform distribution between −4 and −2, consistent with the literature.

At each time step the new energy of the protein and its 'length' (defined as the Euclidean distance in lattice space between the first and final amino acid in the chain) was calculated and recorded. This length was used as a measure of how 'folded' the protein was at any given time.

[3] An amino acid is a monomer's nearest neighbour if it is not covalently bonded to it, but is still one unit of distance away in lattice space.

[4] Values differ to those in the table as a correction of the gas constant, $R \approx 8.314$ J mol^{-1}K^{-1}, is made to convert from energy units of RT to that of $k_B T$ used in this paper, and commonly in the literature.

Fig.2 shows that starting with an unfolded protein and allowing it to undergo the natural folding process, its energy swiftly decreases and then gradually levels off, as expected. At around 2.5×10^5 time steps the energy can be seen to rise to a local maximum value, suggesting that the protein had been in a false local energy minimum, shallow enough to be able to escape. The energy is finally relatively stable around a value of $-67\, k_B T$, indicating a significantly deeper energy well, and thus suggesting that the protein may be in a 'native' molecular state [3]. The length follows a very similar trajectory to the energy, suggesting that a more compact protein makes for a more stable structure.

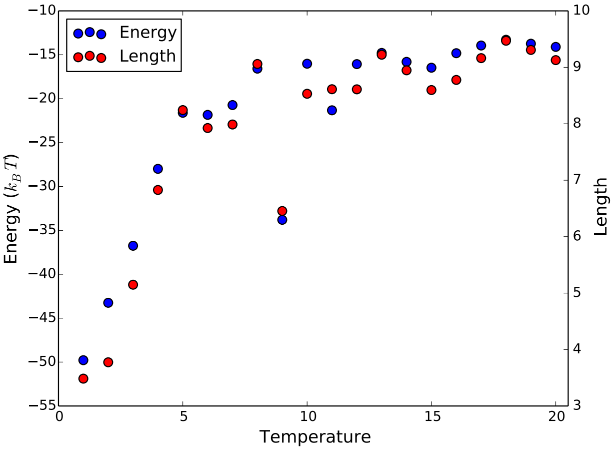

Figure 3: Average energy and length over 10^5 Monte Carlo time steps, as a function of temperature, of a protein, with chain length 30, embedded within a 3D lattice space. The interaction energy matrix, J^i_j, has been filled with randomly chosen values from a uniform distribution between -4 and -2, consistent with the literature. Error bars have been omitted to avoid confusion, the standard deviation for all points is in the range 8.1-16. The result at $T = 9$ can be justified as a random anomaly, since, after running several simulations, such large fluctuations frequently occur for one temperature value in 20.

Averaging over the energies and lengths of different temperature states, starting with a pre-folded formation, demonstrates the dependence of these properties on the solution temperature. Fig.3 shows that, initially, the protein manages to fold relatively quickly to a low energy state, however, despite being initialised in the same conformation, beyond a transition temperature of $\sim T = 5$, the protein quickly struggles to locate a suitable energy minimum[5]. A similar transition has been found in experiments and is known as the 'glass transition'; it occurs when the "dynamic behaviour is dominated by large-scale collective motions of bonded and non-bonded groups of atoms" rather than "simple harmonic vibrations" [12]. It is due to these large scale motions, that the effects of the entropic motion become dominant, thus the protein struggles to fold. In fact, at higher temperatures the system will become even more chaotic as the protein starts to denature [13]. This chaotic behaviour is more explicitly demonstrated in Fig.4; here, an increase in temperature from $T = 1$, where the total energy change is around $18\, k_B T$, to $T = 5$ visibly creates a greater degree of chaos in the energy profile and within 10^5 time steps the energy fails to significantly decrease at all. A similar effect is also seen at $T = 10$, where the protein, again, fails to reduce its energy. Above the transition temperature the energy profile is so volatile that the protein struggles to lower its energy at all, whereas for the case of $T = 1$ the energy decreases gradually over time.

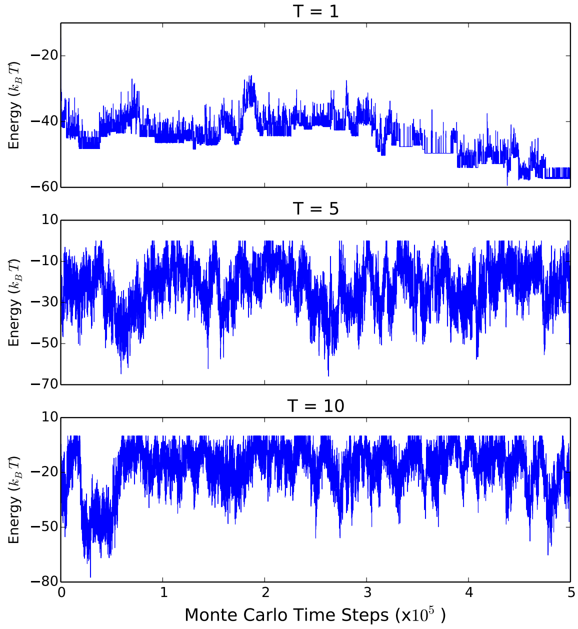

Figure 4: Energy, as a function of Monte Carlo time, of a protein a three different temperatures, with a chain length of 30, embedded within a 3D lattice space, and the same interaction energy matrix as previously used. The protein began in a randomly folded state and the programme was seeded such as to initialise it in the exact same state each time, with only the temperature changed.

Discussion

Conclusion

This work presents a method of calculating certain variables concerned with the mechanism of protein folding. By using lattice simulations, computational techniques are much more readily applied than in cases where space is not discretised, whilst still retaining a high degree of accuracy in calculations. The success of this technique is apparent in so much as it has been a driving force for theoretical understanding in recent decades [8].

The particular algorithm and programme discussed [10] has been verified using various indicators mentioned above. Fig.2 demonstrates the overall ability of the programme to produce the desired, global, objective of modelling a protein folding from an unfolded state to some deep-welled 'native' state [3], whilst Fig.3 depicts the dependence on temperature for both energy and length. In this case, these results are backed by experiment [12] and highlight important features of protein dynamics. These two examples are complemented by Fig.4 which draws attention to the large effect temperature has on an individual system's progression; how it can impede protein folding speed, along with the necessity of having deep energy minima for stability.

[5] From repeated runs the transition temperature is $T = 5 \pm 1$.

Along with simulating the mechanics of the folding process, one of the aims of this paper is to reduce the computational time of the folding process down to that of a real protein. As mentioned, proteins regularly fold in a time on the order of milliseconds, yet simulations carried out here take approximately 10^5 times longer. However, despite this shortfall, the programme presented in this paper demonstrates that even by just applying the relatively simple pathway condition that the protein aims to minimise it's energy, as well as move under statistical thermodynamic conditions, the folding time it dramatically reduced compared to that suggested by Levinthal. Clearly, however, there is still a critical deficit in theoretical understanding about this process, which, if better understood, could lead to further reductions in processing time.

This discussion builds upon many pioneering examples and further demonstrates how important Monte Carlo lattice simulations can be to furthering understanding of complex systems. Their wide use in other fields, particularly in the physical sciences, has meant that techniques have become highly sophisticated, however, the field of biological research is still catching up with this methodology. Moreover, it has already been demonstrated that, among other things, existing results have been confirmed by this programme, thus giving a good foundation for further research and experimentation.

Further Work

This area of research has presented particularly challenging for several decades, as it combines a lack of understanding in the protein folding process due to the difficulties of experimental measurement, with being a computationally intensive task, particularly due to the lack of precise theoretical hypotheses.

The ultimate aim is to develop a programme which can simulate the mechanics of the folding process, in a time comparable to that of a real protein. One method to increase computational efficiency would be dividing the structure into more manageable pieces and parallelising the computation. Splitting up the protein would speed up the computations by an order comparable to the number of splits made, however, there must be enough interactions between amino acids on a global scale such that the protein still folds into its tertiary structure. To minimise the trade off between computational efficiency from independently calculating the folding of segments in parallel, and the cumbersome task of accounting for more global interactions, research is being conducted into what amino acid sub-primary chains act in an autonomous fashion in nature [14]. Some progress has been made in investigating the questions of when a protein behaves as one large domain or like a multi-domain system [15], however, this work has mainly been confined to cubic lattice space, and ignores the potential for a protein to expand into all space. In the future, the programme discussed here could be adapted to incorporate domain folding theorems, with relative ease, to test whether such a procedure would significantly reduce computational time, whilst also allowing the protein to investigate all points in 3 dimensional lattice space.

References

[1] Lister Hill National Centre for Biomedical Communications, "How Genes Work," U.S. National Library of Medicine, 2018.

[2] R. Callender, R. Gilminshin, B. Dyer, and W. Woodruff, "Protein physics," *Physics World*, vol. 7, no. 8, p. 41, 1994.

[3] A. Sali, E. Shakhnovich, and M. Karplus, "How does a protein fold?," *Nature*, vol. 369, pp. 248–251, 05 1994.

[4] C. B. Anfinsen and E. Haber, "Studies on the Reduction and Re-formation of Protein Disulfide Bonds," *The Journal of Biological Chemistry*, vol. 236, no. 5, pp. 1361–1363, 1961.

[5] C. Levinthal, "How to Fold Graciously," *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois*, pp. 22–24, 1969.

[6] U. Mayor, C. M. Johnson, V. Daggett, and A. R. Fersht, "Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 25, pp. 13518–13522, 2000.

[7] K. A. Dill, "Theory for the folding and stability of globular proteins," *Biochemistry*, vol. 24, no. 6, pp. 1501–1509, 1985.

[8] R. L. Baldwin, "Matching speed and stability," *Nature*, vol. 396, pp. 183–184, 1994.

[9] M. Karplus, "The levinthal paradox: yesterday and today," *Folding and Design*, vol. 2, pp. S69 – S75, 1997.

[10] J. P. Bullock, "Protein." `https://github.com/JosephPB/Protein/`, 2018.

[11] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation," *Macromolecules*, vol. 18, no. 3, pp. 534–552, 1985.

[12] D. Ringe and G. A. Petsko, "The glass transition in protein dynamics: what it is, why it occurs, and how to exploit it," *Biophysical Chemistry*, vol. 105, no. 2, pp. 667 – 680, 2003.

[13] D. C. Rees and A. D. Robertson, "Some thermodynamic implications for the thermostability of proteins," *Protein Science : A Publication of the Protein Society*, vol. 10, no. 6, pp. 1187–1194, 2001.

[14] A. R. Panchenko, Z. Luthey-Schulten, and P. G. Wolynes, "Foldons, protein structural modules, and exons," *Proceedings of the National Academy of Sciences*, vol. 93, no. 5, pp. 2008–2013, 1996.

[15] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, "Domains in folding of model proteins," *Protein Science : A Publication of the Protein Society*, vol. 4, no. 6, pp. 1167–1177, 1995.