Computing with evolving proteins

J.L. Fernández-Villacañas, J.M. Fatah and S. Amin

BT Labs, Martlesham Heath, Ipswich IP5 7RE, UK

Abstract. Dynamic Local Search [1] has been applied to the evolution of interactions between protein-like structures. These are composed of a randomly selected sequence of amino acids that are linked together to form linear polymers in three dimensions. The objective function chosen for optimisation is the potential energy given by a Toy protein model. Proteins fold, move and interact with other chains to minimise their objective function at a given rate, F_{rate}, depending on the sum of the rates for re-organisation of their structures. The interaction between different proteins gives a whole range of local attraction/repulsion regimes that result in new structures with new bonds, broken bonds and recursive loops.

1 Introduction

The aim of this paper is to study the interaction between proteins as structures capable of computation. These structures are composed of a string of basic computationally active units (amino acids) that fold to give a functional computational unit based on their shape.

Protein structure can be discussed in terms of three levels of complexity: a primary structure that refers to the linear sequence of amino acids that codes for the protein, a secondary structure that describes the local internal arrangements (alpha helices and beta sheets) and a tertiary structure that studies the spatial 3-d configuration.

Primary structure has been modelled [2] [3] by representing amino acid side chains as spheres connected to their C_α atoms that are themselves linked forming the chain backbone.

Secondary structure has been calculated for a number of RNA molecules [4] with a substantial percentage of correctly predicted helices. Nevertheless these results do not account for the three dimensional structure of the molecules; its prediction still constitutes a great unresolved challenge.

Previous optimisation techniques for deriving the tertiary structure of proteins [5] [6] have also produced limited results. More recent attempts [7] use real protein helices as test cases to derive a folding matrix that is evolved through a genetic algorithm.

This paper is concerned with what lessons the folding and interaction processes between these structures may bring in terms of their use as computational elements. In our work we consider proteins as creatures that interact amongst themselves and with the environment they live in. They behave by folding, making or breaking bonds into low energy configurations minimising the total energy

(intramolecular and intermolecular). The optimisation process thus resembles the dynamics in the system and has been tailored to account for different energy transfer regimes. On one hand, the system may tend towards thermodynamic equilibrium when the strength and frequency of the interactions is low enough for the exchanges in energy to be absorbed by other proteins changing their internal configuration while, on the other hand, the system dynamics could be governed by kinetic exchanges. The latter happens when the proteins become trapped into metastable states if the energy barriers for re-organisation of their structures are too large to be overcome by pure thermal fluctuations.

2 The Protein Model

Each protein is composed of a sequence of amino acids that are linked together by rigid unit-length bonds in three dimensions. No attention is paid to the composition, positions and properties of the individual atoms; only each amino acid as a whole is considered.

The amino acids are the active computational units in our model. In our approach each of the 20 different amino acids is encoded by a normalised variable $\xi_1, \ldots, \xi_n \in [-1, 1]$ that represents the dominant type of interaction. In this case ξ mimics the degree of hydrophobicity or aversion to non-polar residues. Isoleucine is the most hydrophobic of the set, with $\xi = 1$ while arginine is the least hydrophobic with $\xi = -1$. Any polymer with n amino acids is uniquely described by the $n-2$ polar and azimuth angles, θ_i and ϕ_i, between non-terminal amino acids, $i = 2, \ldots, n-1$.

3 Energy calculation

In order to evaluate and rank different protein configurations we have built on the Protein Toy model by Stillinger *et al.* [8]. It assumes two components to the intramolecular potential energy for each molecule: backbone interactions V_1 between bonded residues and interactions V_2 between non-bonded residues; the former will be independent of the amino sequence except for di-sulphide bonds $C-C$, while the latter will depend on the residues and their properties.

For non-bonded residues interaction we can assume (see [7] and references within) that the main force responsible for driving the folding process is the *hydrophobicity* or aversion for water for non-polar molecular residues. This force dominates the Van der Waals interactions between dipoles, hydrogen bond interactions between two electronegative atoms and general electrostatic interactions between charged residues.

In this paper, the evolution of a typical protein-like system will be treated as a process of energy minimisation (see section 6 for details). Our objective function is the total intramolecular and intermolecular potential energy. The intramolecular relative energy to be minimised can be expressed as,

$$\Phi_{intra}(k) = \sum_{i=2}^{n_k-1} V_1(\theta_i, \phi_i) + \sum_{i=1}^{n_k-2} \sum_{j=i+2}^{n_k} V_2(r_{ij}, \xi_i, \xi_j) + V_3 \tag{1}$$

where n_k is the number of amino acids for protein k and $k = 1, \ldots, N$ is the protein index, N being the total number of proteins. The distances between amino acids, r_{ij} are calculated from the relative spherical co-ordinates of the preceding residues in the chain. V_3 is the change in intramolecular potential energy due to the presence of internal bonds. $V_3 = 0$ when an individual protein has no loops and,

$$V_3 = V_1(\theta_{ij}, \phi_{ij}) - V_2(r_{ij}, \xi_i, \xi_j)$$

when there is an internal bond between amino acids i and j. The intermolecular energy is,

$$\Phi_{inter} = \sum_{k=1}^{N} \sum_{l=k+1}^{N} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} V_2(r_{ij}, \xi_i, \xi_j) \tag{2}$$

where again r_{ij} represents the distance between two amino acids in different chains. The total potential energy will be,

$$\Phi = \sum_{k=1}^{N} \Phi_{intra}(k) + \Phi_{inter} \tag{3}$$

The expressions for the bonded backbone and non-bonded potential terms are (see ref. [8]),

$$V_1(\theta_i, \phi_i) = \frac{1}{4}(1 - cos\theta_i \; cos\phi_i) \tag{4}$$

$$V_2(r_{ij}, \xi_i, \xi_j) = 4\left(\frac{1}{r_{ij}^{12}} - \frac{C(\xi_i, \xi_j)}{r_{ij}^{6}}\right) \tag{5}$$

$$C(\xi_i, \xi_j) = \frac{1}{8}(1 + \xi_i + \xi_j + 5\,\xi_i\xi_j) \tag{6}$$

The potential term in (5) is the one responsible for the attraction and repulsion between amino acids and consequently between proteins if the energy conditions allow it. Expression 6 controls the strength of these interactions per unit distance; it varies from $C(\xi_i, \xi_j) = 1$ for a II pair to $C(\xi_i, \xi_j) = -\frac{1}{2}$ for a IR pair.

4 The Basics of Computing

In a previous paper [9] we demonstrated the capability of protein-like structures to move in a solvent medium and interact with other similar structures in various regimes of attraction and repulsion. In this section we intend to introduce the basic ingredients for these structures to be able to compute sums, subtractions and form loops. This would imbue the proteins with computing capabilities based on their shape. The set of operands responsible for the computing capabilities are just two: breaking internal amino acid bonds and making new bonds, both between different proteins and within the same structure (i.e. loops). Let's consider them separately:

- *Breaking bonds:* We will assume that at every stage in the evolution of a typical protein-like system, each structure has got the ability of breaking every single amino to amino bond. If a breaking transition proves to be energy efficient, that is the resulting system with an extra protein has got a lower total potential energy given by Eq. 3, the breakage will be given a chance to happen. If we start with a system of N proteins, the net result will be an increased number of proteins $N + 1$.

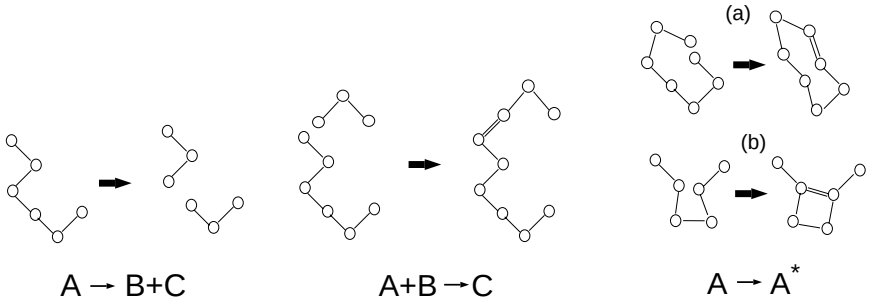

Fig. 1. Breaking of a bond (left), Making a bond (double line) (center) and Loop formation (right): End-to-end loop (a) and internal di-sulphide loop (b).

- *Making bonds:* This process allows for the formation of complex structures by the addition of two ingredient chains $A + B \rightarrow C$ at a time. We will assume that bonds between two different chains can only be made from and to the terminal amino acids of each protein.
- *Loops:* A special case of making bonds, loop bonds can happen between terminal amino acids or two non-consecutive special amino acids all in the same protein. Their net effect is the closure of all or part of the sequential amino acid chain. As we mentioned in section 2, this type of bond resembles the di-sulphide bond between Cysteine amino acids in real proteins. The net effect of this transition is an alteration in the internal structure, thus changing the potential energy value of ingredient structure A as $A \rightarrow A^*$.

5 The Local Optimiser

Dynamic Local Search (DLS) [1] provides a suitable approach for solving multivariable constrained functions. The basic idea is to explore local and global space simultaneously. One exploration allows us to search the local space in detail and the other to escape from the local space if the system becomes trapped in a local minimum. Along each dimension the search is performed independently of other dimensions; each variable is randomly perturbed in opposite directions and the magnitude of the amplitude along those two directions are different from each other and from one iteration to the next. The size of the perturbed amplitude is designed to explore local space for a fixed percentage of the time and global space for the remaining time.

The application of the DLS method to the problem of the optimisation of the potential energy given by Eq. (3) for a system containing N proteins is as follows;

In the simplest case, when $N = 1$, only $2 \times (n-2)$ variables need to be perturbed to achieve energy minimisation. These correspond to the polar and azimuth angles (θ_i, ϕ_i) for each amino acid $i = 2, \ldots, n-1$. The situation becomes more complicated when $N > 1$. In this case the number of degrees of freedom is,

$$N_{var} = 2 \times \sum_{i=1}^{N} (n_i - 2) + 3 \times (N-1) \tag{7}$$

that is, the sum of the free angles for each of the proteins in the set plus the (x_i, y_i, z_i) relative co-ordinates of the first amino acid for each protein $i = 2, \ldots, N$ with respect to the first protein $i = 1$. In this way both the relative distances among proteins and the folding which results from angle perturbation are responsible for the energy minimisation. Each iteration of the local optimiser consists of checking the returned objective function along opposite directions for every degree of freedom (as in Eq. (7)). The number of function evaluations per iteration is thus, $3 \times N_{var}$.

6 How does evolution proceed?

The main working hypothesis is that we are cooling the system down slowly enough so that we can assume that the solvent and the proteins are in local thermodynamic equilibrium; that is, individual proteins will exchange energy with the solvent and with other proteins of the order of KT_{eq}, where T_{eq} is the average temperature of the system. As evolution proceeds, the energy of the system is minimised progressively; this means that T_{eq} is decreased, or that the system cools down. When the returned objective function is below a desired fixed temperature, we stop the optimisation process and keep the folded structures as the valid configurations.

On the other hand, protein interactions may be governed by kinetic dynamics outside thermal equilibrium. This is the case when the reorganisation for the

structure of the protein cannot be achieved by thermal fluctuations in a reasonable time. Obviously this approach does lead to the protein being trapped in local minima, making the final potential energy of the whole chain always higher than its thermodynamic equilibrium value especially for very long chains.

In order to avoid the latter scenario, we will assume that we can slow down the rate of evolution so that the protein can re-organise itself and still be in thermodynamic equilibrium; this will imply a slower convergence, longer evolving times to the desired final configuration. The working assumption is that the rate of evolution for a fixed system of proteins is proportional to the sum of the transition rates, both for making and breaking bonds. This means that if, at a given stage of its evolutionary path, a system is very likely to undergo transitions with a high probability that may affect its configuration, the rate of evolution is slowed down; on the other hand if the sum of the rates at a given time is low, then the system can continue being optimised and cooled at a quicker rate.

Let's define what we understand by rates of transition for re-organisation of the proteins structure: if a system composed of N proteins makes or breaks a bond and as a result of this transition there is a change in the total potential energy of the system ΔE, then the associated rate for that transition is,

$$k = e^{-(\frac{\Delta E}{KT_{eq}})}$$

where we have normalised the energy changes by a factor that accounts for the equilibrium temperature of the system at that evolution stage. This value is calculated before any transition rates are computed by evaluating the total energy of the system and equating it to KT_{eq}.

It follows that transitions which lead to lower energy values $\Delta E < 0$ have rates $k > 1$ and those which increase the energy of the system have $k < 1$. The rate of evolution is made proportional to,

$$N_{iter} = A + B\frac{1}{\sum_{i=1}^{N_r} k_i} \tag{8}$$

where A and B are constants and N_r is the number of possible make and break transitions of the whole system at a given time; this quantity depends on the number of proteins, their chain sizes and the distances between their amino acids. N_{iter} is the number of iterations that we subject the whole system to at a given time; thus when the sum of the rates is different, N_{iter} changes and consequently the speed of evolution is altered.

7 Results and Discussion

In this section we will present one example for each of the different protein reactions quoted in section 4; that is, breaking of a bond, making of a bond between two proteins and making of a bond to form a loop in one protein.

The first of the examples uses one copy of the Crambin protein (Abyssinian Cabbage seed, 46 amino acids). Its sequence has been obtained from [10]. At the

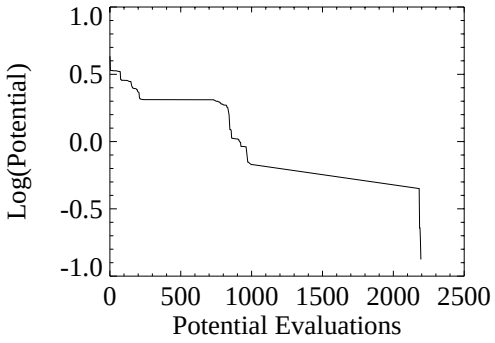

Fig. 2. Total energy vs. number of function evaluations for the Crambin system.

beginning of the simulation the protein is placed in random spatial location and the total potential energy is measured. After the re-organisation transition rates are calculated a breakage is chosen ($k = 1.023$), and the total energy is reduced by a factor of 0.02 splitting the protein in two of sizes 32 and 14 by breaking the bond between amino acids 32 and 33.

The optimisation is then carried out for 5 full iterations when the rates are re-computed. No transition is selected and again the optimiser reduces progressively the total energy of the system below a value of $KT_{eq} = 0.1$ after a total 9 iterations. Figure 2 describes the evolution of the total energy in this system; note that each iteration corresponds to 267 function evaluations for the broken protein as can be deduced from Eq. (7). Three different stages of the system spatial evolution can also be found in Fig. 3; (a) shows the initial configuration (iteration=0), (b) corresponds to the stage right after the break has been made and (c) is the final energy optimised configuration.

We have also studied the evolution of random sequence distributions of various lengths. For instance, the $n5$ system (five randomly chosen sequences of lengths between 3 and 10 amino acids) produces an initial configuration of five protein of sizes 8, 5, 7, 5 and 5 amino acids. The total potential energy versus the number of times the potential function is evaluated has been shown in Fig. 4a. There are two transitions; the first one after 12 iterations when proteins 3 and 4 (sizes 7 and 5 respectively) make an end-to-end bond between the first amino acids of each protein. A transition rate $k = 2.48$ reduces the energy of the system by a factor of 0.9 and forms a new chain of length 12. The transition is clearly visible in Fig. 4a at 2233 function evaluations.

The system, now with only four proteins, is optimised for a further 16 iterations, when another transition is chosen, $k = 1.01$, and an internal bond is made between the terminal amino acids of protein 4 (initially protein 5). The system decreases its energy by a factor of 0.01 and protein 4 forms a closed loop.

Evolution proceeds for a further 11 iterations until the total energy of the system falls below $KT_{eq} = 0.1$ at 7160 function evaluations. Figure 5 displays three stages in the evolution of this system; (a) is the initial configuration, (b) corresponds to 2233 function evaluations, when the first extra bond is made while (c) shows the state of the system after protein 4 forms a closed loop at

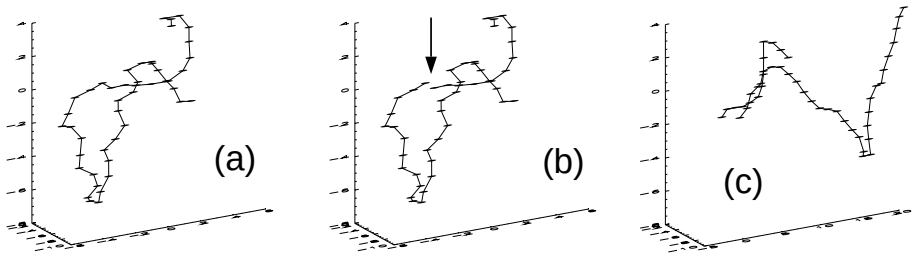

Fig. 3. Three stages for the evolution of the Crambin protein system (see text for explanation).

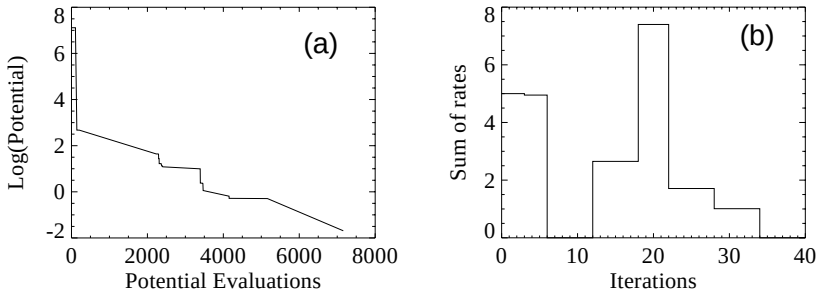

Fig. 4. Total energy vs. number of function evaluations (a) and change in the rate evolution (b) for the $n5$ system.

5162 potential evaluations.

The rate at which evolution takes place for the $n5$ system is shown in Fig. 4b. Once the sum of the rates for the re-organisation is known, the system is optimised for a number of iterations given by Eq. (8). Constants A and B have been chosen so that when the sum of the rates is $\sim 2.7N_r$, the system evolves only for one iteration (slow regime), while when the sum of the rates is $\sim N_r$ the number of iterations is $N_{itermax}/2$ (normal regime). N_r is the total number of possible transitions (see section 6) and $N_{itermax}$ is the maximum number of iterations allowed in the fast regime when the sum of the rates is close to zero.

8 Conclusions

- The Dynamic Local Search method used proves effective in minimising the potential energy for different evolving regimes and system configurations.
- The dynamically evolving system approaches the conditions of local thermodynamic equilibrium by making the rate of evolution dependent on the total sum of the rates for re-organisation of the protein structures.
- The minimised evolved structures are useful as cases for a folding and bond based computational output. The dynamics show examples of additions, $A + B \rightarrow C$, subtractions, $A \rightarrow B + C$, and recursive loops, $A \rightarrow A^*$.

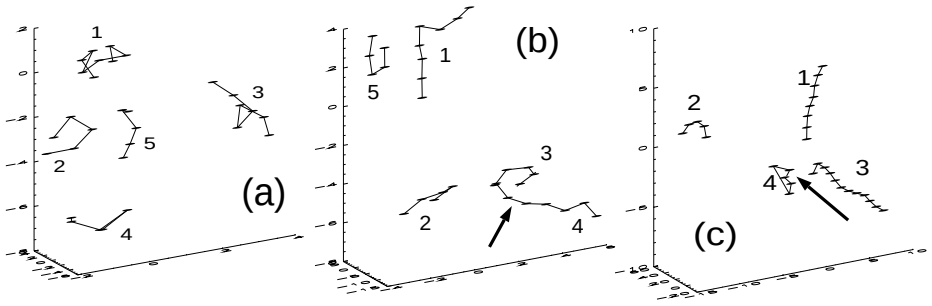

Fig. 5. Three stages for the evolution of the $n5$ protein system (see text for explanation)

Our main interest is not to reproduce exactly the shape of real protein distributions but to strip away the complexity introduced by pseudo-energy functions that could describe the atomic force fields in detail in order to get an insight into the types of operations a protein-like system can achieve.

References

1. Amin, S., Fernández-Villacañas, J.L.: Dynamic Local Search. Proc. Galesia97 conference, Glasgow, September 1997
2. Lau, K.F., Dill, A.: Theory for protein mutability and biogenesis. Proc. Nat. Acad. Sci. USA **87** 1990 638–642
3. Unger, R., Moult, J.: Genetic Algorithms for Protein Folding Simulations. J. Mol. Biol. **231** 1993 75–81
4. Jaeger, J.A., Turner, D.H., Zucker, M.: Improved predictions of secondary structures for RNA. Proc. Nat. Acad. Sci. USA **86** 1989 7706–10
5. Quian, N., Sejnowski, T.J.: Predicting the Secondary structure of Globular proteins using Neural Network models. J. Mol. Biol. **202** 1988 865–884
6. Fariselli, P., Compiani, M., Casadio, R.: Predicting Secondary structures of Membrane proteins with Neural Networks. Euro. Biophys. J. **22** 1993 41–51
7. Calabretta, R., Nolfi, S., Parisi, D.: An Artificial Life model for predicting the Tertiary Structure of unknown proteins that emulates the folding process. Proc. of Third European Conference of Artificial Life, Granada 1995
8. Stillinger, F.H., Head-Gordon, T., Hirsfield, C.L.: Toy model for protein folding. Physical Review E **48** 2 1993 1469
9. Fernández-Villacañas, J.L., Fatah, J., Amin, S.: Evolution of Protein Interactions. Proc. BCEC'97 Bio-Computing and Emergent Behaviour, Skövde, Sweden, September 1997
10. Berstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M.: The protein data bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. **112** 1977 535–542

This article was processed using the LaTeX macro package with LLNCS style