Protein folding with stochastic L-systems

Gemma Danks[1], Susan Stepney[1] and Leo Caves[1]

[1]University of York, YO10 5DD, UK
gbd501@york.ac.uk

Abstract

Protein molecules adopt a specific global 3D structure in order to carry out their biological function. To achieve this *native* state a newly formed protein molecule has to fold. The folding process and the final fold are both determined by the sequence of amino acids making up the protein chain. It is not currently possible to predict the conformation of the *native* state from the amino acid sequence alone and the protein folding process is still not fully understood. We are using L-systems, sets of rewriting rules, to model the folding of *protein-like* structures. Models of protein folding vary in complexity and the amount of prior knowledge they contain on existing native protein structures. In a previous paper we presented a method of using open L-systems to model the folding of *protein-like* structures using physics-based rewriting rules. Here we present an L-systems model of protein folding that uses knowledge-based rewriting rules and stochastic L-systems.

Introduction

Protein molecules perform molecular functions in the cell that require a specific 3D structure. This *native* state of a protein is achieved only after a process of folding from an initially unfolded state that it adopts during synthesis on ribosomes in the cell. The thermodynamic hypothesis states that a protein folds to its lowest energy state (Anfinsen, 1973). The folding pathway(s) of a protein are unclear but it is known that the only information necessary to predict the native 3D structure of a protein is contained in its amino acid sequence. A protein cannot find its native state through random sampling as even for a small protein this would take in excess of 10^{27} years (Levinthal, 1969; Zwanzig et al., 1992). The energy landscape theory of protein folding (Onuchic et al., 1997) predicts a rugged funnel-like energy landscape biassed towards the native structure due to the effects of evolution. This theory predicts multiple pathways to the native state that an ensemble of unfolded protein molecules may follow. This is an opposing view to the classical view that there is a single defined pathway for each protein proceeding through a sequence of intermediate states.

Protein molecules are possibly the simplest example of a biological complex system and exhibit many emergent properties that have been selected for during the evolution of life on Earth. Proteins are composed of, and function at, many different levels. The folding of a protein may be viewed as an emergent phenomenon. It is governed by underlying physics involved in the interaction of amino acids that make up the protein chain. These local interactions together give rise to the changing conformation of the whole molecule in a way that leads to the native state. We use L-systems (Prusinkiewicz and Lindenmayer, 1990) to represent these local interactions as a set of rewriting rules. In a previous paper we described an open L-systems model of folding *protein-like* structures using simple physics-based rules (Danks et al., 2007). Here we describe a complementary approach using a stochastic L-systems model of protein folding with knowledge-based rewriting rules. We first give an overview of protein structure, then briefly describe the main aspects of modelling protein folding. We give an overview of L-systems and how we previously used them to model protein folding using physics-based rules. We then describe the development of a knowledge-based L-systems model of protein folding and our initial results.

Protein structure

There are 20 different naturally occurring amino acid monomers that make up proteins. These have the same NH_2-CαH-COOH backbone but differ in their side chain from the central carbon atom (Cα). These different side chains give amino acids different chemical properties. The genetic code specifies a unique linear sequence of amino acids that are covalently linked by peptide bonds during protein synthesis to form polypeptides. This is the *primary* structure of a protein. The length of a single polypeptide varies from around 70 amino acid residues to 1000s of residues. The conformation of the polypeptide chain is defined by the local conformation of each amino acid. Peptide bonds that link amino acids together are fairly rigid. This causes the CO of one amino acid and the NH of the next to lie in the same plane. The two backbone bonds N-Cα and Cα-C allow rotation - these rotations give each amino acid its backbone torsion angles ϕ and ψ respectively. These torsion angles

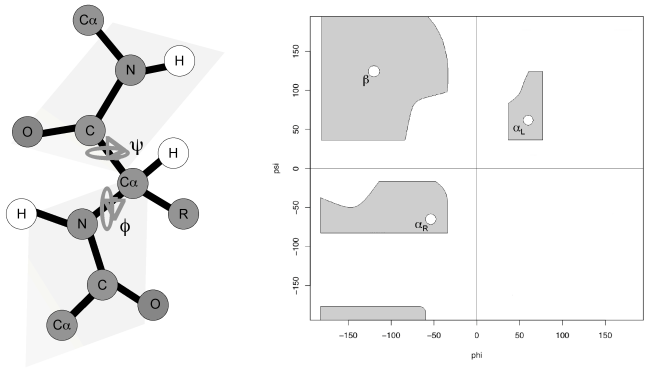


Figure 1: A dipeptide unit showing the structure of an amino acid (NH-CαHR-CO, where R is the amino acid specific side chain). Shaded areas show the atoms lying in the plane of the peptide bonds. The location of the two backbone torsion angles ϕ and ψ are shown and the distribution of sterically allowed values are shown as shaded regions in the ϕ/ψ plot. The two main areas of allowed torsion angles correspond to those that consecutive amino acids adopt to form the two main secondary structure units in folded proteins: the α-helix and β-sheet.

cannot adopt all possible values due to steric hindrance - some of the atoms branching from the backbone as well as the side chain atoms would collide if certain torsion angles were adopted (Ramachandran et al., 1963). The allowed torsion angles can be plotted to show the regions of ϕ/ψ space that each amino acid can occupy (figure 1). Two main regions of this ϕ/ψ, or Ramachandran, plot of torsion angles at the local amino acid level also correspond to the two main *secondary* structural elements found in native protein structures - the α-helix and the β-sheet. These are formed when a number of consecutive amino acid residues adopt the same torsion angles, and are stabilised by hydrogen bonding between the backbone N-H of one amino acid and the backbone C-O of another. The arrangement of these structural units gives the *tertiary* structure of a protein, i.e. the *native* state of a single polypeptide. The units are generally connected by turns and loops, smaller structural elements.

There are currently over 49,000 known protein structures. These can be divided into classes based on the arrangement and proportion of α-helix and β-sheet units (Murzin et al., 1995; Orengo et al., 1997). Two of the classes are 'all-alpha' and 'all-beta' containing mainly α-helices and β-sheets respectively. Two other major classes in the SCOP database (Murzin et al., 1995) contain a mix of α-helices and β-sheets: the α and β ($\alpha+\beta$) proteins contain segregated alpha and beta regions; the α and β (α/β) proteins contain alternating alpha and beta structures (figure 2). These classes are further subdivided into structurally related proteins. Proteins with similar structures often share a common evolutionary

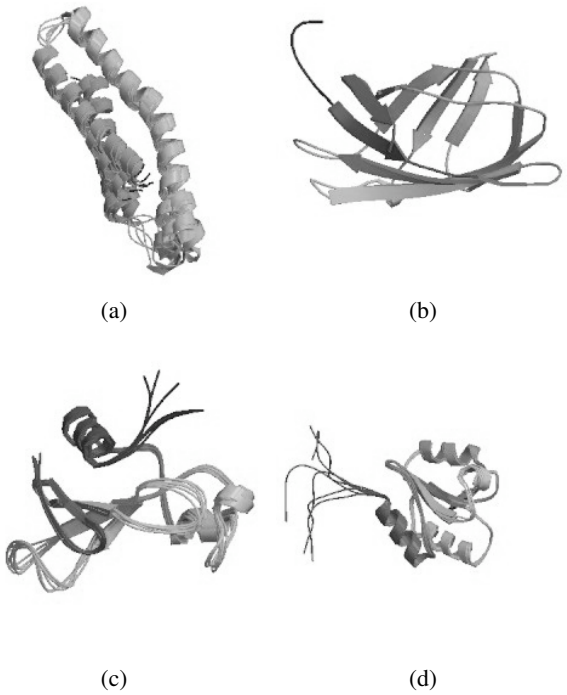


Figure 2: Examples of the four main SCOP classes. All images have been taken from www.rcsb.org. Structures are drawn using ribbons to represent secondary structure (arrows show the direction of a β-strand within a β-sheet). Multiple strands show different experimentally determined structures. (a) An all-alpha protein, PDB ID: 1aj3 (b) An all-beta protein, PDB ID: 1exg (c) Barnase (1bnr) an alpha and beta (a+b) protein - α-helices and β sheets are separated in the protein. (d) 2bjx, an alpha and beta (a/b) protein - α-helices and β-sheets are dispersed throughout the protein.

origin and will have a similar amino acid sequence. Comparative modelling uses related sequences with known structures to predict the fold of a new sequence (Ginalski, 2006). However some very different protein sequences can fold to similar structures and occasionally similar sequences fold to different structures.

Modelling protein folding

There are a wide number of existing models of protein folding (Duan and Kollman, 2001). These range in their representation of space (e.g. lattice or off-lattice, 2D or 3D) as well as the level of detail in the protein molecule itself (from all-atom models to those representing each amino acid as a single bead), which also largely defines the representation of interactions within the protein. Models also differ in their assessment of the *protein-like* nature of the final fold and the method used to sample conformations and find the native state. The simplest models can sample every possible conformation to find the most *native-like* state - usually the

lowest energy state where the free energy of the model is represented by the sum of interactions. For example the HP-model (Lau and Dill, 1989; Dill et al., 1995) represents the amino acids in a protein as two different kinds of beads on a string - H and P for hydrophobic and hydrophillic - confined to a 2D lattice with each bead on a point in a grid. The interactions between two H beads (i.e. two H beads next to each other on the grid but not in the string) are favourable and are summed for each conformation to give the energy. With proteins of a small number of beads it is possible to calculate the energy for every possible conformation and find the arrangement on the lattice of the native state. With a more detailed representation this is impossible and a sampling method must be adopted. Two main methods are used in these more detailed models. Monte carlo techniques, based on small random changes in conformation combined with an acceptance criterion using a Boltzman distribution, are widely used to find a conformations of progressively lower energy states (Hansmann and Okamoto, 1999). Molecular dynamics is also used extensively to model protein folding using Newton's laws of motion (Scheraga et al., 2007). However, calculating the forces between all atoms in a protein is computationally intensive and so it is not currently possible to model folding on biological time scales for any but the smallest and fastest folding proteins.

Alternatively, a move-set can be biased using knowledge of native protein structures. The most successful methods of protein structure prediction are those based on fragment assembly (Bujnicki, 2006). These model folding by alternating local conformations of the protein chain between different conformations of short fragments of native protein structures. The ϕ/ψ plot gives the conformations that a single dipeptide unit is allowed to adopt, sterically, which differs slightly between amino acid types. This places restrictions on many of the possible local conformations. However, the choice between allowed conformations can not be readily determined from such a local level. Further restrictions on local conformations are governed by the neighbouring residues and their local conformations (Fitzkee et al., 2005).

L-systems

L-systems were developed as a mathematical theory of plant development (Prusinkiewicz and Lindenmayer, 1990; Lindenmayer, 1968). The simplest L-system consists of an *axiom* containing an initial string of symbols together with a set of rewriting rules, or *productions*, one for each symbol. These rules are applied in parallel to each symbol in the string over a number of *derivation steps*. The current symbol, or *predecessor*, is rewritten by another symbol, or string, the *successor* as defined by the rule for that symbol. For example a simple rule might be:

$$a \rightarrow ab$$

This rule would be applied to every a that appears in the string.

Context sensitive L-system rules are applied only if the symbol is preceeded by and/or followed by a specific string. For example the rule:

$$c < a > d \rightarrow ab$$

is only applied to a if it is preceeded by c in its left context and followed by d in its right context.

Parametric L-systems allow each symbol to have one or more parameters associated with it. The rules can then incorporate conditions on these parameters. For example:

$$a(x) : x > 1 \rightarrow a(2)b(1)$$

will be applied only if the parameter associated with a is greater than 1.

Stochastic L-systems allow a number of different rules to match a certain predecessor. Each rule is applied with a given probability. For example using the rules:

$$a \rightarrow ab : 0.75$$
$$a \rightarrow b \ \ : 0.25$$

the predecessor a will be replaced by ab 75% of the time and by b 25% of the time.

Open L-systems (Mech and Prusinkiewicz, 1996) incorporate an interacting model of the environment. The L-system and an environmental program communicate using environmental query modules, $?E(...)$. Information is sent to the environment using the parameters of $?E(...)$. The environment uses this information to determine a response and communicates this information back to the L-system using $?E(...)$ parameters, which can be used in productions.

Using L-systems to model protein folding

The backbone conformation of a protein can be described using only the backbone torsion angles (ϕ, ψ) of each amino acid in the chain. The native state of a protein molecule has specific torsion angles associated with each residue. Secondary structure assignment is largely determined by torsion angles together with hydrogen bonding patterns. Folding of the protein involves the torsion angles within each residue changing to their native conformations. L-systems provide a natural way to model this process. Rewriting rules can be used to alter the ϕ, ψ angles in each residue in parallel across the whole molecule. This leads to the emergence of a global 3D fold as a result of local changes in conformation.

In a previous paper (Danks et al., 2007) we described the development of an open L-systems model of protein folding using physics-based rules. A brief outline is given below.

The axiom contains an amino acid sequence, using the single letter amino acid code, with initial backbone

torsion angles, ϕ and ψ, as parameters. For example, the first 4 amino acids in the protein barnase in a β-strand conformation (where ϕ is approximately -120° and ψ is approximately 120°) gives the following axiom

$$A(-120,120)Q(-120,120)V(-120,120)I(-120,120)$$

An initial derivation step is used to rewrite each symbol representing an amino acid with symbols that represent individual atoms, bonds, bond angles and torsion angles. An initial local conformation of each amino acid is formed by using the initial backbone torsion angles contained in the axiom. Each atom is associated with an environmental query module containing information on the atom that is communicated to an environmental program. At each subsequent folding derivation step an environmental step is performed where the L-system sends this information and the position of each atom in the protein to the environmental program. The environment processes this information and sends a response for each atom back to the environmental query modules in the L-system.

Two models were developed that use a different level of representation of interactions between atoms. One model calculates whether any of the atoms are colliding and another more detailed model calculates the forces between nearby atoms. This information is returned to the L-system. A rule set uses the collision or force information returned to each atom. The rules alter the backbone torsion angles of each amino acid depending on the interactions of atoms within that amino acid with any other atom in the protein that is spatially local. This is repeated over a number of derivation steps leading to a physics-based folding at the global level of the whole protein molecule. The resulting structures at each step were assessed for *protein-like* qualities that included a measure of compactness, which is characteristic of folded protein structures. We found that using local rules in this way to model folding, while not giving *native-like* folds, did lead to compact structures.

Developing a knowledge-based model of protein folding using stochastic L-systems

The physics-based L-systems models allow a protein to sample conformations by moving through time: forces between atoms determine the next conformation. We have used a different approach in developing a knowledge-based L-systems model. A protein alters in conformation over a number of derivation steps, but this is not representative of time. Instead of local moves based on physical forces, local conformations sample those that are most often found in native, i.e. fully folded, structures.

The backbone torsion angles that describe the conformation of a protein are used to assign secondary structure. Taking into account hydrogen bonding and the state of neighbouring residues each residue can be assigned one of seven different secondary structure states (Kabsch and Sander, 1983). These are: α-helix (H), extended strand (E), residue in isolated β-bridge (B), 3/10 helix (G), π-helix (I), hydrogen bonded turn (T) and bend (S). Residues not taking part in secondary structure units are not assigned a state. We have developed an L-systems model that uses these secondary structure states instead of individual torsion angles. We use stochastic rules with probabilities based on data obtained from the DSSP database (ftp://ftp.cmbi.kun.nl/pub/molbio/data/dssp) instead of using physics-based deterministic rules.

Obtaining frequencies of context dependent states

Data obtained from the DSSP database include backbone torsion angles, amino acid type and secondary structure state of each amino acid residue in 35,492 proteins from the protein data bank (www.rcsb.org). Each protein sequence was split into fragments using a window of 3 residues long. Fragments where secondary structure was not assigned were removed leaving 10,954,172 fragments.

Frequencies of each of the 20 residue types in each of the 7 secondary structure states were calculated in all possible contexts of one residue either side. Where R represents an individual amino acid residue, A is the amino acid type and S is its state, a 3-residue fragment contains the following information:

$$R_{i-1}(A_{i-1},S_{i-1})R_i(A_i,S_i)R_{i+1}(A_{i+1},S_{i+1})$$

For each unique combination of $A_{i-1}, S_{i-1}, A_i, A_{i+1}, S_{i+1}$ the frequency of each possible S_i is calculated. There are 20^3 possible 3 residue sequences and 7^3 possible state contexts. All possible 3 residue sequences (8000) appear in the data used here. However, of a possible 2,744,000 unique 3 residue sequence and state combinations only 230,250 appear in the data. Where there is no data for an amino acid in a particular 3 residue fragment in a specific conformation, that state is allocated a low frequency of 10^{-2}, rather than zero, to allow these states to be sampled with a low probability in the L-systems model.

Developing stochastic L-systems rules

An L-systems model has been developed to use the frequencies calculated from the data in stochastic rewriting rules. The axiom contains an amino acid sequence using the single letter amino acid code. An initial derivation step rewrites this code to replace each amino acid by the symbol R with parameters defining the amino acid type and its initial state. For example the first five amino acids in barnase, $AQVIN$, in an initial extended (E) conformation are replaced by:

$$R(A,E)R(Q,E)R(V,E)R(I,E)R(N,E)$$

where the first parameter represents the amino acid type and

the second parameter represents the initial conformation (numbers are used in the model).

Each R is also accompanied by an environmental query module containing the same information. At each subsequent derivation step the information on each residue is first sent to an environmental program. This stores all the residue amino acid types and states. Open L-systems are used here only to store and return specific frequencies from a matrix of values. For each residue, excluding the first and last, given the amino acid type of that residue and the amino acid type and state of one residue either side the frequency of that residue in each of the 7 secondary structure states is found. The first and last residues are given equal probabilities for each state. These 7 frequencies are returned, for each residue, to the environmental query modules in the L-system. A set of 7 stochastic rewriting rules, one for each state, use the corresponding frequency from the environment as its probability of being applied. These rules then rewrite the secondary structure state of each residue depending on the 3 residue sequence that it is within (constant) and the secondary structure state of the residues either side (variable). The form of the rewriting rules are as follows:

$$R(a,s) >?E(p0,p1,p2,p3,p4,p5,p6) \rightarrow R(a,E) : p0$$
$$R(a,s) >?E(p0,p1,p2,p3,p4,p5,p6) \rightarrow R(a,H) : p1$$
$$:$$
$$R(a,s) >?E(p0,p1,p2,p3,p4,p5,p6) \rightarrow R(a,S) : p6$$

where $p0, p1, p2, p3, p4, p5, p6$ are the probabilities of being in states E, H, G, I, B, T, S respectively. Each $R(a,s)$ is followed by an associated environmental query module $?E(p0,p1,p2,p3,p4,p5,p6)$ in its right context. This contains frequencies of each of the 7 secondary structure states, returned from the environment, for that residue while its neighbours are in their current states. Each $R(a,s)$ is then rewritten to change its state, s, to one of the secondary structure states with probabilities calculated from the frequencies in $?E(...)$. The environmental modules are also rewritten to again store the amino acid residue type and the updated state of the preceding $R(a,s)$ to send to the environment at the next derivation step. The states of the neighbours also change at each derivation step as it is a parallel rewriting process.

The aim of this model is to detect the emergence of any locally encoded secondary structure preference and to assess its ability to produce *protein-like* global features. The 3D protein structure is obtained by using *homomorphism* rules. These are applied after each derivation step but are used only for graphical interpretation and do not rewrite any symbols in the string. A rule for each amino acid type draws out the structure of that amino acid with amino acid specific ϕ, ψ angles for each of the seven secondary structure states. These angles were obtained from the data used to calculate the probabilities. Each secondary structure occupies a specific region(s) of the ϕ/ψ plot. As an approximation we took the most common ϕ, ψ angles for each residue type in each state.

Folding proteins using stochastic L-systems

The folding behaviour of four example amino acid sequences using the knowledge-based stochastic L-systems rules are shown in figure 3. Each sequence represents a protein from one of the four major SCOP classes: all-α, all-β, $\alpha + \beta$ and α/β. Each plot shows the change in state of each residue in the protein over 5000 derivation steps. Each protein starts in the same all-extended state. There is a marked difference in patterns of secondary structure, across all derivation steps, between different protein sequences. However, comparison to the native secondary structure states for each protein shows that the structures emerging are not necessarily *native-like*. The horizontal bands that are visible for some residues show that some local secondary structure preference is emerging using these local rules.

Secondary structure is one characteristic of protein structures. Another key feature of globular proteins is their compactness. The 3D structures of each protein at each derivation step was obtained by mapping secondary structure states for each residue type to typical ϕ, ψ torsion angles taken from the data. The radius of gyration (Rg) is a measure of compactness and this was calculated for each structure resulting from each derivation step. Figure 4 shows the change in Rg of one protein, barnase (1bnr), over 2000 derivation steps. This gives an indication of how *protein-like* the global structures are at each step in the L-systems model. The results of the physics-based L-systems model for barnase as well as the value of the native state are also shown. It is clear that the knowledge-based rules are not folding the protein to a very compact structure, and there seems to be little convergence to one structure over time. At most steps the radius of gyration is above the native state and consecutive steps may allow the protein to fold and unfold rapidly. This can also be seen by looking at the global conformations at a number of derivation steps (figure 5).

The physics-based rules seem to be forming more compact structures. There is no constraint on which states a residue may take at the next step in the knowledge-based model other than the probability of being in that state in the context of its neighbours. Torsion angles at subsequent steps could jump dramatically across ϕ/ψ space and this is causing the global structure to also change dramatically. There is also little convergence to a preferred global structure, although this appears to vary between protein sequences - those with more β-sheet conformations seem to maintain a more consistent pattern in the state images (figure 3). This problem is largely due to the fixed probabilities that drive the rules. For convergence to a preferred structure the prob-

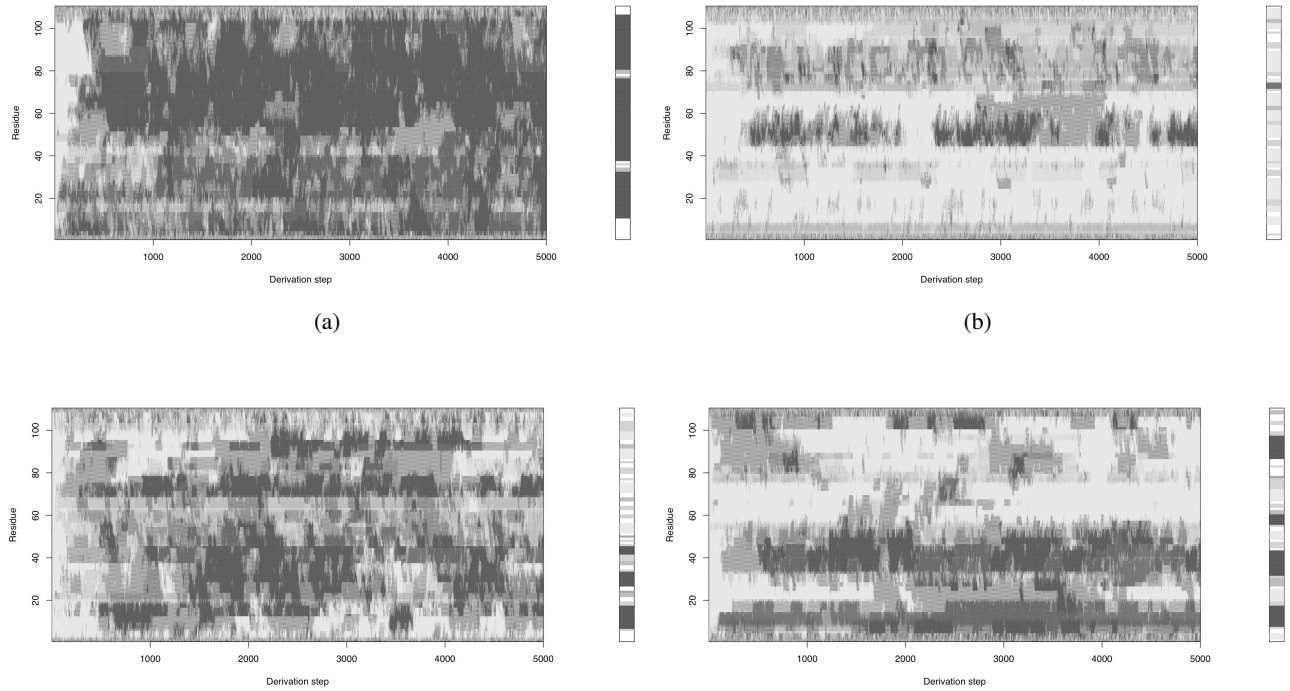


Figure 3: Results from four protein sequences, each from a different SCOP class, using the knowledge-based stochastic L-system rules for 5000 derivation steps. Each image shows the states of individual residues (y-axis) at each derivation step (x-axis). Lightest grey represents the extended state, black represents the α-helix. The 3/10 helix, π-helix, isolated beta bridge, turn and bend are shown in shades of grey from dark to light. Horizontal bands show the emergence of preferred local secondary structure. A bar to the right of each plot shows the native secondary structure for each protein (white represents unassigned states). Native global structures are shown in figure 2. (a) 1aj3 (all-alpha) (b) 1exg (all-beta) (c) 1bnr (α + β) (d) 2bjx (α/β).

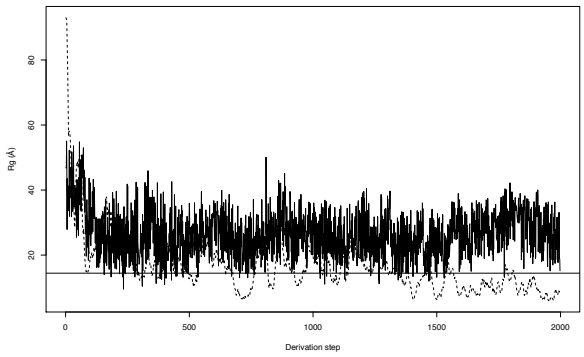


Figure 4: The radius of gyration, Rg - a measure of compactness, for each structure at each derivation step. The solid line shows the change in Rg in the knowledge-based model while the dashed line corresponds to the physics-based model for the the amino acid sequence of barnase, 1bnr. The horizontal line shows the Rg value of the native state.

abilities must be altered during folding to give each residue a final probability of being in only one state. Although not converging to one preferred structure each protein sequence seems to maintain a consistent cycling through similar states. Horizontal bands emerge for certain residues in the states images (figure 3) indicating that there is some secondary structure preference locally in the sequence. Each protein sequence also tends to adopt its particular pattern of states with different initial conformations (figure 6).

A difficulty with assessing global conformations in the knowledge-based model is the inaccuracies in mapping from individual residue secondary structure states to backbone torsion angles. This is particularly difficult when dealing with turns and bends where more than one region of torsion angle space appears in the data. The local conformations of residues that form a turn are dependent on their positions in that turn structure. This issue may be resolved by incorporating context dependence in the homomorphism rules.

The next stage in this work is to incorporate some physics into the knowledge-based model. A global driving force, for example to a compact global conformation, may be needed

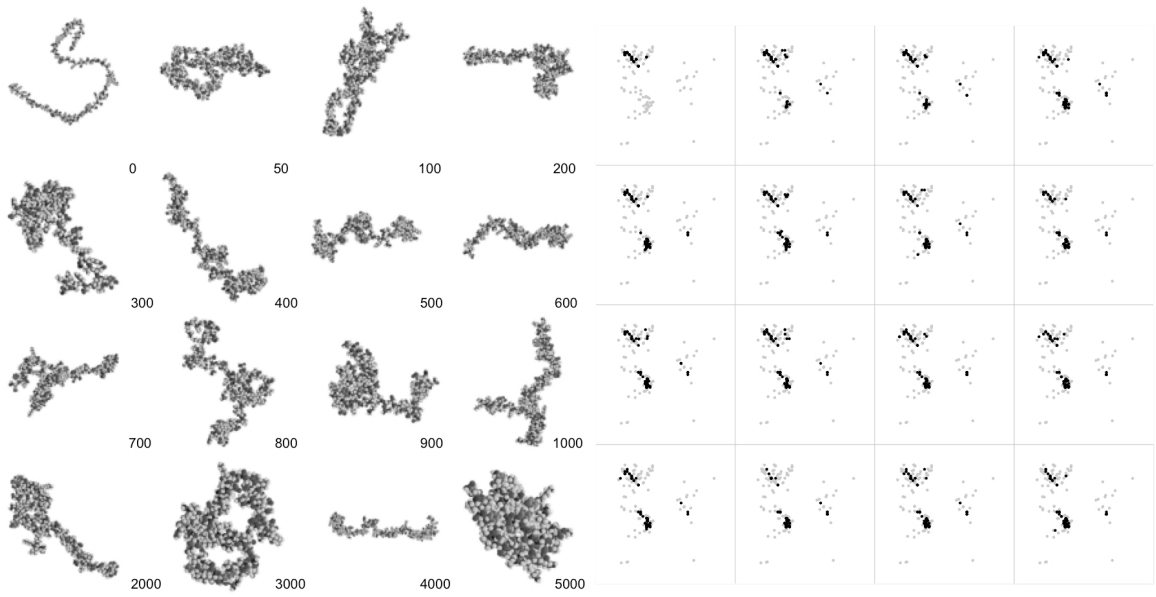
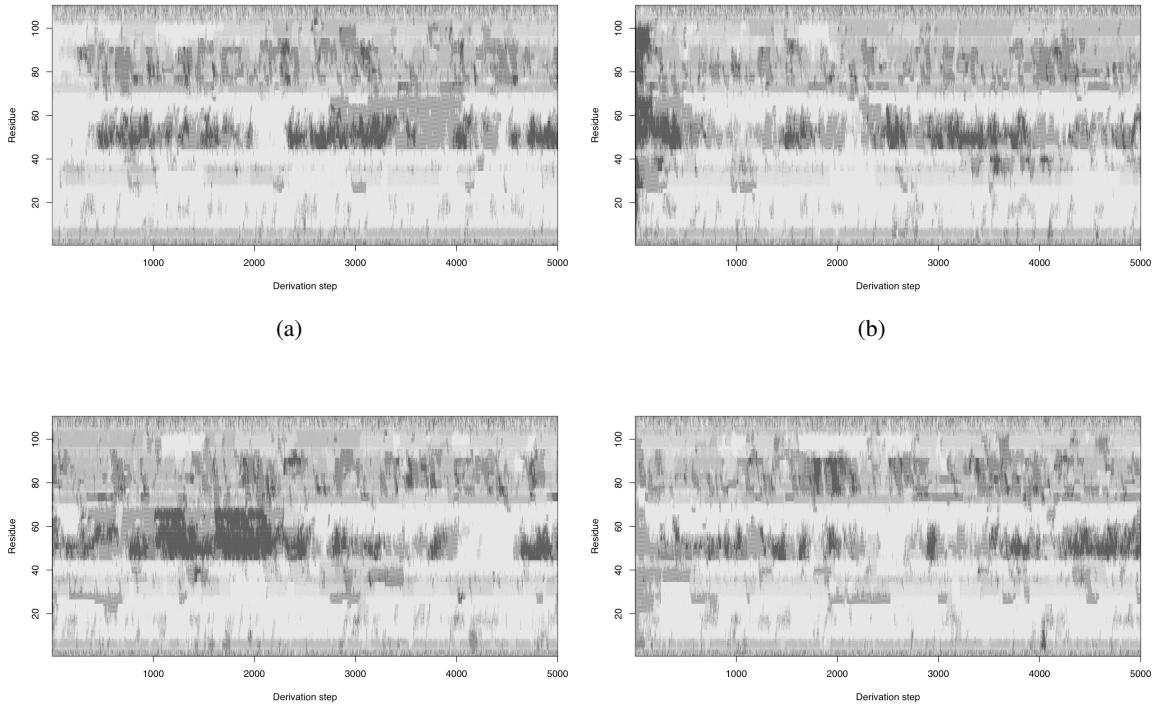


Figure 5: General features emerging from the L-system using the protein sequence of barnase, 1bnr. The initial state corresponds to an all extended conformation (image of state changes across all derivation steps shown in figure 3c). Images show the global changes in conformation, ϕ/ψ plots show the ϕ, ψ angles (black) for each amino acid at corresponding derivation steps with the native state angles shown in grey for reference.



(a) (b)

(c) (d)

Figure 6: Results of protein 1exg in different initial conformations. Each image shows the states of individual residues (y-axis) at each derivation step (x-axis). Horizontal bands show the emergence of preferred local secondary structure. (a) initial state as all extended (b) initial state as all alpha (c) initial state all 3/10 helix (d) initial state in alternating alpha-beta.

to alter the probability table during folding. The combination of our simple physics-based L-systems model with the knowledge-based rules would also allow selection between states. This would allow local structural preference to work together with spatially local interactions and may lead to more *protein-like* structures that converge to a final folded state. Incorporating physics into the knowledge-based model may also help to prevent large global changes in conformations caused by unrestricted changes in local residue conformations.

Summary

We have presented an L-systems model that uses data-driven stochastic rewriting rules to fold protein sequences by altering the secondary structure state of individual amino acid residues. The state of each residue is rewritten in parallel across the whole protein. The state that an individual residue changes to depends on the amino acid type of that residue and the amino acid types and the current states of the neighbouring residues on either side. Seven secondary structure states are used based on those used in the DSSP database. The probabilities of adopting each of seven states were obtained from the frequencies of each state, given the states of residues either side, found in 10,954,172 3-residue fragments from 35,492 native protein structures in the DSSP database. Typical backbone ϕ, ψ torsion angles were also obtained for each amino acid type in each of the seven states from the data and used to reconstruct the 3D structure of a protein at each derivation step. This was used to assess the *protein-like* nature of global conformations.

Results are shown for four protein sequences from each major structural class. Local structure preference can be seen to emerge for some residues in a sequence. Overall differences in the proportion of local α-helix and extended conformations can also be seen between protein sequences using these rules. However, the resulting structures do not converge to a preferred global compact conformation. Further work will be to incorporate some physics-based bias into the probability table to allow a preferred global conformation to emerge.

Acknowledgements

This work is supported by the BBSRC. We thank Karim Elsawy for providing the fragment data.

References

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(96):223–230.

Bujnicki, J. M. (2006). Protein-structure prediction by recombination of fragments. *Chembiochem*, 7(1):19–27.

Danks, G. B., Stepney, S., and Caves, L. S. D. (2007). Folding protein-like structures with open L-systems. *ECAL 2007, LNCS*, 4648:1100–1109.

Dill, K. A., Bromberg, S., Yue, K. Z., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S. (1995). Principles of protein-folding - a perspective from simple exact models. *Protein Science*, 4(4):561–602.

Duan, Y. and Kollman, P. A. (2001). Computational protein folding: From lattice to all-atom. *IBM Systems Journal*, 40(2):297–309.

Fitzkee, N. C., Fleming, P. J., Gong, H., Panasik, N., J., Street, T. O., and Rose, G. D. (2005). Are proteins made from a limited parts list? *TRENDS in Biochemical Sciences*, 30(2):73–80.

Ginalski, K. (2006). Comparative modeling for protein structure prediction. *Current Opinion in Structural Biology*, 16(2):172–177.

Hansmann, U. H. E. and Okamoto, Y. (1999). New monte carlo algorithms for protein folding. *Current Opinion in Structural Biology*, 9(2):177–183.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637.

Lau, K. F. and Dill, K. A. (1989). A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules*, 22(10):3986–3997.

Levinthal, C. (1969). How to fold graciously. *Mössbaun Spectroscopy in Biological Systems Proceedings, Univ. of Illinois Bulletin*, 67(41):22–24.

Lindenmayer, A. (1968). Mathematical models for cellular interactions in development. Parts I and II. *Journal of Theoretical Biology*, 18:280–315.

Mech, R. and Prusinkiewicz, P. (1996). Visual models of plants interacting with their environment. *SIGGRAPH 96, Computer Graphics*, pages 397–410.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP - a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540.

Onuchic, J. N., LutheySchulten, Z., and Wolynes, P. G. (1997). Theory of protein folding: The energy landscape perspective. *Annual Review of Physical Chemistry*, 48:545–600.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH - a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108.

Prusinkiewicz, P. and Lindenmayer, A. (1990). *The Algorithmic Beauty of Plants*. Springer.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99.

Scheraga, H. A., Khalili, M., and Liwo, A. (2007). Protein-folding dynamics: Overview of molecular simulation techniques. *Annual Review of Physical Chemistry*, 58:57–83.

Zwanzig, R., Szabo, A., and Bagchi, B. (1992). Levinthal's paradox. *PNAS*, 89(1):20–22.