Folding Protein-Like Structures with Open L-systems

Gemma B. Danks, Susan Stepney, and Leo S. D. Caves

York Centre for Complex Systems Analysis,
University of York, York, YO10 5YW, United Kingdom
gbd501@york.ac.uk

Abstract. Proteins, under native conditions, fold to specific 3D structures according to their 1D amino acid sequence, which in turn is defined by the genetic code. The specific shape of a folded protein is a strong indicator of its function in the cell. The mechanisms involved in protein folding are not well understood and predicting the final conformation of a folded protein from its amino acid sequence alone is not yet achievable despite extensive research efforts, both theoretical and experimental. The protein folding process may be viewed as an emergent phenomenon, a result of underlying physics controlling the interaction of amino acids with their local environment, leading to the complex global fold. In this spirit we present a model for investigating protein folding using open L-systems, local rewriting rules with environmental interaction.

Key words: Protein-folding, L-systems, open-L-systems

1 Introduction

In physiological solution, a protein molecule needs only the information contained in its 1D amino acid sequence – a string of typically several hundreds of amino acids of 20 different types in a specific order – to fold to its lowest energy, stable, native state [1]. This specific 3D structure is necessary for the biological function of a protein. In general different sequences fold to different structures and similar sequences fold to similar structures, however there are exceptions where two very different protein sequences share a similar native state. The number of possible conformations of a given sequence is far greater than the number the protein can adopt during folding, indicating that folding is not a random or exhaustive process but follows some pathway(s) [2]. These pathways may be thought of as resulting from the underlying physics of interactions between amino acids in the protein chain. In this sense, protein folding is a paradigm of emergence - the development of well defined global order from a process of self-organised assembly. We present a model for investigating the application of parallel rewriting rules to study protein folding. We use open L-systems with turtle interpretation to model the protein structure and its subsequent folding through application of rewriting rules to local regions of the protein structure over a number of generations leading to global changes in conformation.

L-systems were developed as a mathematical theory of plant development [3, 4] facilitated by an interpretation based on turtle geometry. L-systems are sets of parallel rewriting rules acting repeatedly on symbols in an initial string, the *axiom*, over a number of *derivation* steps. At each step the string may be interpreted graphically leading to visual models of plant growth and development [4]. L-system rules can be quite flexible and several extensions to L-systems have allowed for more complex models of plant development to be created. The following summarises the extensions we used in this work, for further details and formal definitions see [4, 5].

The simplest L-system consists of rules that each rewrite one symbol, called the *predecessor*, with another symbol or string of symbols, called the *successor*, whenever that symbol appears in the string. Context-sensitive L-systems take into account the context of the predecessor - i.e. its neighbouring symbols. Parametric L-systems allow parameters to be assigned to symbols in the string. Conditions on these parameters may then be used in the L-system rules and C-like statements may also be incorporated for further flexibility.

Open L-systems [5] include a separate environmental process, interacting with the L-system via environmental query modules $?E(...)$, in a bi-directional communication process. The environmental program is sent information from the L-system using parameters of the environmental query modules. This information is processed by the environment to determine a response, which is returned to the environmental query modules in the L-system string. The L-system rules can then use this information in the environmental query modules. In this way an open L-system can model a plant interacting with its environment over a number of derivation steps. This has been used to model for example collision avoidance in branching structures and competition for light [5].

2 The Rules of Protein Folding

Up to twenty naturally occurring amino acids can be found in a protein sequence. Each amino acid has a backbone of a central carbon atom, called Cα, attached to an amino group (NH_2), a carboxyl group (COO), a hydrogen and an amino acid specific side chain or R group (see Fig. 1a). Each R group has a distinctive structure and chemical characteristics. The prototypical R group, found in the amino acid alanine, is a methyl group ($–CH_3$), glycine is simpler but a special case as it has a single H atom instead of a side chain. Other side chains vary from long hydrocarbon chains to ring structures or charged groups. Amino acids are linked together, to form *polypeptides*, via a planar peptide bond.

The spontaneous folding of a protein from its *unfolded state* to its lowest energy stable *native state* is driven by physical interactions [6]. The main interactions between atoms thought to drive protein folding consist of the following:

1. Van der Waals forces: attraction and repulsion between atoms, representing general short-range cohesion and excluded volume effects.
2. Electrostatic forces: attraction/repulsion between (partially) charged atoms.

3. Hydrogen bonding: Hydrogen atoms bonded to, and interacting with electronegative atoms (e.g. oxygen/nitrogen) form characteristic spatial interactions. These hydrogen bonds may drive or stabilise the formation of *secondary structure* in a protein – local structural regularities in the protein chain, mainly the α-helix and β-strand. These structures are stable and are the main ordered structural elements occurring in folded proteins.
4. The hydrophobic effect: in aqueous solution, hydrophobic amino-acids tend to pack together at the core of globular proteins, while hydrophilic amino acids tend to be located at the surface.

It is still unclear which of the above interactions are the dominant driving forces in protein folding. Under physiological conditions, all the information a protein needs to fold to its native state is encoded in its amino acid sequence. Different sequences will give rise to different interactions between amino acids in the chain and lead to different native conformations. These sequences have been selected by evolution to fold quickly and spontaneously to stable states [7].

Understanding of the process of protein folding and the accurate prediction of the native state has been the goal of numerous models of protein folding (see [8] for a detailed review). These models range in their level of complexity. The simple 2D lattice HP models [9, 10] assume the hydrophobic effect is the driving force and model short proteins, or peptides, as beads of two types - hydrophobic (H) and hydrophilic or polar (P) - on a string while finding the 2D conformation that maximises hydrophobic contacts. Complex all-atom continuous 3D space models calculate forces between each atom pair[8, 11].

Our approach is to investigate how underlying local rules, governing the interaction of amino acids with their local environment, can be used to model the process of protein folding as an emergent phenomenon leading to a complex global fold. We have developed a three dimensional model using 20 amino acid types and physics-based open L-system rules that drive the folding of an initial protein conformation. Previous work [12, 13] using L-systems to model proteins has focussed on obtaining the native conformation through evolving L-systems rules and an initial axiom that grow the native structure of small (up to 34 residues) proteins under the two-dimensional lattice HP model. Our work focusses on modelling the dynamics of the process of protein folding, rather than on structure prediction. We summarise details of simple L-systems models that have been constructed using different sets of rewriting rules that differ in the level of detail in the representation of physical interactions. For further details see [14]. Both models contain an initial axiom defining a protein sequence in single letter amino acid code. An initial rule set replaces the single letter code of each amino acid with a string of symbols representing the 3D structure of each amino acid type. The string at this stage can be interpreted graphically to give the initial 3D conformation of the protein. Turtle interpretation of the string is used to define the geometrical properties of the system and communicate this to the environment for inter-atomic force calculation. A further rule set is applied over a number of derivation steps to alter the conformation of each amino

acid according to physical interactions in its local environment. The repeated reapplication of these rules leads to global folding of the protein.

The first folding rule set uses a simple environmental model to detect collisions between atoms and leads to local conformational changes that depend on the presence of local collisions. This rule set requires knowledge of the direction of folding, i.e. the successor of the rule has to be specified. The second more sophisticated model uses a more realistic model of the physical interactions between atoms. This rule set uses information from an environment that calculates physical forces to determine the direction of folding. The following sections describe these models in more detail.

3 Building Proteins in L-systems

There are two main variables responsible for the conformation of a chain of amino acids, which are the two backbone torsion angles of each amino acid. The torsion angle ϕ is an angle of rotation around the bond between the backbone nitrogen and Cα. The second torsion angle ψ is similarly an angle of rotation around the bond connecting Cα and the following carbon atom (see Fig. 1a). Other torsion angles are present in side chains but do not directly define the conformation of a protein backbone. Rotations around ϕ and ψ cause a polypeptide chain to alter in conformation. For example, all amino acids in a chain adopting both ϕ and ψ torsion angle values of 180° results in an extended chain (as in Fig. 1a). If consecutive amino acids adopt torsion angles $(\phi, \psi) = (-57°, -47°)$ the result is an

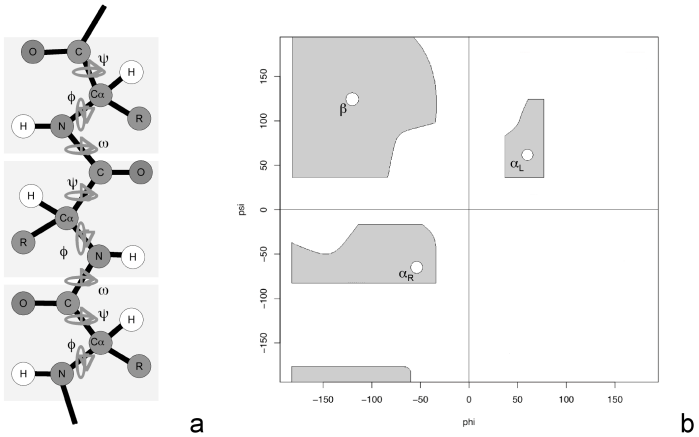

Fig. 1. a. Three amino acids (shaded) linked by peptide bonds with backbone torsion angles shown. The torsion angle ω varies little due to the rigid peptide bond. Variations in torsion angles ϕ and ψ result in different backbone conformations. **b.** A schematic diagram representing a typical Ramachandran plot showing allowed regions of ϕ, ψ space shaded grey. The common secondary structures are shown (β = β-strand, α_R= right-handed α-helix, α_L= left-handed α-helix)

```
Lsystem: alpha-helix

derivation length: 1

Axiom: h    A(-57,-47)A(-57,-47)A(-57,-47)A(-57,-47)A(-57,-47)
            A(-57,-47)A(-57,-47)A(-57,-47)A(-57,-47)A(-57,-47)   F(1.32)o

A(phi,psi) --> F(1.329)n [+(57)Fh]
                  -(58.3)\(phi)
                  F(1.458)c[+(55.6)&(54.73)Fh] [+(55.6)^(54.73)F(1.521)c /(150)\(-60)
                                                                        [/(60)+(70.5)Fh]
                                                                        [/(180)+(70.5)Fh]
                                                                        [\(60)+(70.5)Fh]]
                  -(68.8)\(psi)
                  F(1.525)c[+(59.2)F(1.231)o]
                  -(63.8)\(180)


homomorphism

c --> ;(83)@O(0.8);(99)
h --> ;(99)@O(0.8);(99)
n --> ;(50)@O(0.8);(99)
s --> ;(115)@O(0.8);(99)
o --> ;(67)@O(0.8);(99)
```

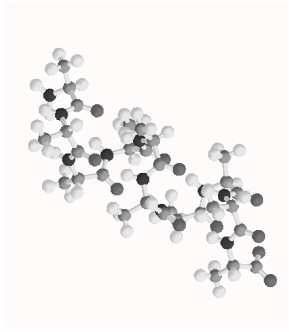

Fig. 2. An L-system to create an α-helix from 10 alanine (A) amino acids with initial torsion angles $(\phi, \psi) = (-57°, -47°)$ defined in the axiom. Each symbol A is replaced by the string in the successor of the rule shown. Graphical interpretation of the string results in an α-helical structure with coloured spheres representing different atom types created for graphical interpretation only (i.e. not rewritten) by *homomorphism rules*

α-helix, a stable secondary structure in proteins due to the presence of hydrogen bonds. Not all combinations of possible ϕ, ψ torsion angles are physically possible due to collisions of neighbouring atoms (*steric hindrance*) at some angles. In 1963 Ramachandran et al. examined all possible conformations of two linked peptide units and plotted the resulting *allowed* ϕ, ψ combinations [15]. This plot, known as a Ramachandran, or ϕ, ψ, plot, shows two main regions of allowed ϕ, ψ space (see Fig. 1b). These regions correspond with the torsion angles defining the α-helix and β-strand, which are the two main secondary structures found in proteins. These occur when consecutive amino acids adopt these angles, and so these extended secondary structures emerge from local amino acid conformations. Further, global structure is also achieved through the organisation of these secondary structures to form the overall 3D *tertiary* structure of a protein.

Using the L-systems software 'L-studio' [16] we have developed a set of rules that when applied to an initial amino acid sequence in the axiom leads to a string, which when interpreted graphically represents an all-atom 3D structure of a protein. The conformation of this initial structure is defined in the axiom as parameters on each amino acid defining initial (ϕ, ψ) torsion angles. Through the use of different values for initial torsion angles any conformation of a structure can be specified. For example all $(\phi, \psi) = (-57°, -47°)$ would produce an all α-helix conformation (see Fig. 2). Inserting the torsion angles representing the native structure of a protein will create the native backbone conformation.

4 Folding Proteins in L-systems

Once an initial 3D conformation of a protein sequence has been created a further rule set in the L-system rewrites the initial torsion angles in each amino acid repeatedly over a number of *derivation* steps. Altering the torsion angles in

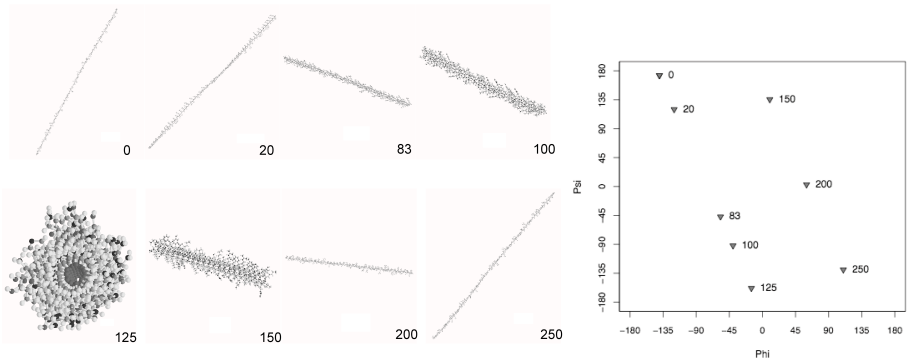

Fig. 3. Derivation steps emerging from a set of L-system rules: $\phi \rightarrow \phi + 1, \psi \rightarrow \psi - 2.65$ applied to an initial conformation in an extended state of the amino acid sequence of the protein barnase (110 residues). A Ramachandran plot shows the ϕ, ψ angles of every amino acid at corresponding step numbers. These rules cause the folding of the structure to a β-strand conformation at derivation step 20, and an α-helix at step 83. Continuing to apply these rules leads to physically impossible structures e.g. step 125.

parallel across the whole chain results in global changes in the protein fold as a consequence of local conformational changes in each individual amino acid.

4.1 Simple Geometric Model

The rewriting rules in this simplest case are of the format: $\phi \rightarrow \phi \pm \Delta\phi$ where $\Delta\phi$ is a constant value of increment for ϕ and similarly for ψ. This results in a uniform change in local conformation across the whole protein chain leading to ordered changes in global conformation. However, with no restrictions in place on the torsion angles allowed, the structure is free to adopt physically impossible conformations (see Fig. 3) both globally (the entire protein chain may occupy a flattened disk shape) and locally (ϕ, ψ combinations causing overlapping neighbouring atoms within an amino acid). Imposing restrictions to local conformational changes is possible in the L-system but with information limited to being local in the sequence (i.e. individual amino acid torsion angles) regions of the chain that are brought close together spatially but distant in the sequence are not having an effect on the folding. As folding is in 3 dimensions it is important that the local rules are governed by spatially local regions not just regions local in sequence. This requires the use of open-L-systems in the model to communicate with the L-system rules in order to include local spatial information. Two sets of rules were developed with different levels of simplification of the physics involved and are described below.

4.2 Simple Collision Avoidance Model

The first model incorporating open L-systems uses an existing environmental program 'Ecosystem' included in the L-studio software package [17]. At each

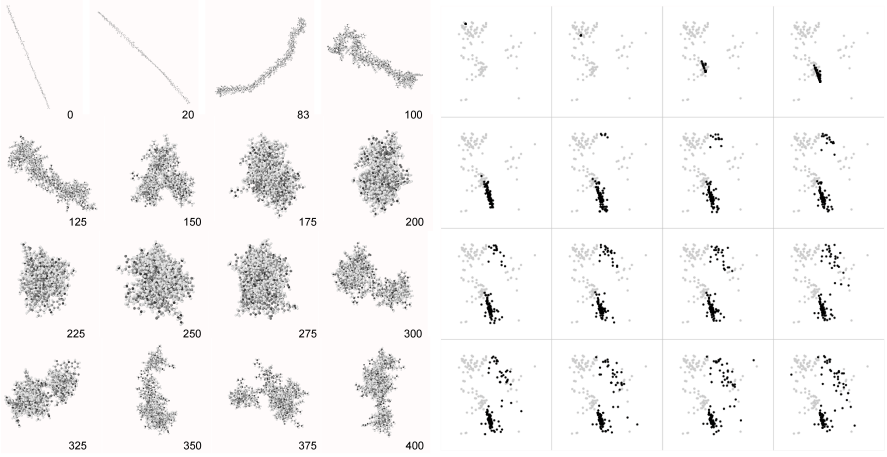

Fig. 4. General features emerging from the L-system using rules as in Fig. 3 but modified such that the sign of the angle increment is reversed with each consecutive local collision. Images show the global changes in conformation obtained using the amino acid sequence of barnase, Ramachandran plots show the ϕ, ψ angles for each amino acid at corresponding derivation steps with the native state angles shown in grey for reference.

derivation step this environment is sent the radii and positions of spheres, using environmental query modules ('?E(r)' where r is the radius of a sphere), and detects if any sphere is overlapping any other sphere. This information is returned to the L-system which then incorporates it into context-sensitive, parametric rules with conditions on the parameters of all communication modules in the string local to the torsion angle being rewritten. Therefore the increment of each torsion angle may depend on whether there is a collision between any atom close to the torsion angle and any other atom in the protein. The information on local collisions is used to alter torsion angles in one way if no collisions occur and another if there is a collision. For example, simply reversing the sign of the angle increment on detection of a collision causes the local conformation to back out of its previous move. The effects on the global and local conformational changes in such a model can be seen in Fig. 4. Local conformational changes vary across the chain resulting in complex sequence-dependent 3D global folds.

Due to restrictions imposed in the environmental program and to keep the model as simple as possible all atomic radii were kept equal and much smaller than their actual radii. This leads to problems since local conformations that would produce collisions, were the radii realistic sizes, are then allowed. This causes the L-system structures to adopt less *protein-like* conformations. The environmental program was modified to allow realistic radii, which improved the protein-like nature of the resulting local conformations but it became clear that defining constant angular increments needed to be replaced by a model which incorporates more physics.

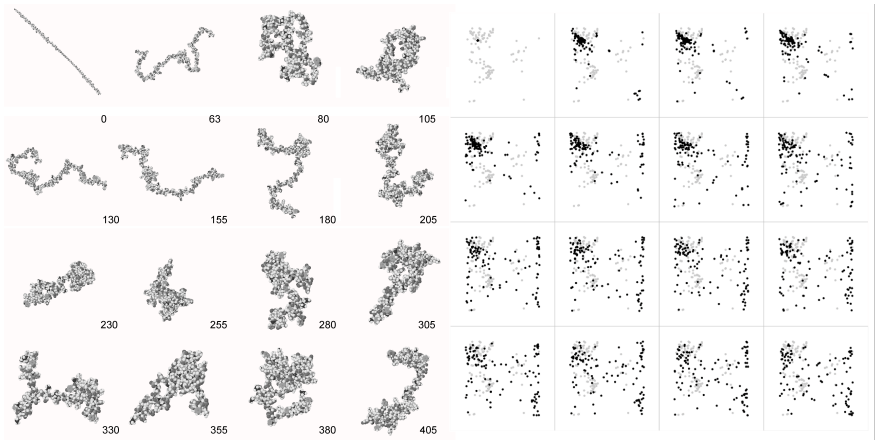

Fig. 5. General features emerging from the L-system using the sum of torque on neighbouring atoms, from the local environment, to increment the angle. The initial state corresponds to a β-strand conformation (as in step 20 in Fig. 4). Images show the global changes in conformation obtained using the amino acid sequence of barnase, Ramachandran 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.

4.3 Physical Forces Model

In a more physical rule set, information on the forces exerted on each atom are returned to the L-system, replacing the simple collision detection. These forces were calculated using a Lennard-Jones potential, to model van der Waals interactions i.e weak attractive forces between distant atoms and strong repulsive forces between very close atoms, and simple Coulombic electrostatics. The following formulae show the forms of the Lennard-Jones potential (left) and the electrostatic potential (right):

$$V(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \qquad V(r) = \frac{q_1 q_2}{r}$$

Where r is the separation distance between two atoms, σ represents the separation distance where the potential is zero and ϵ the energy well-depth, and q_1 and q_2 are the partial charges of two atoms.

In this model, side chain torsion angles are also rewritten. The information (the forces calculated) remains local to the torsion angle. The torque each atom exerts on its nearest rotatable bond is used in the rules to alter each torsion angle by summing the values of torque from nearby atoms. This avoids defining a fixed angular increment and allows the physics to drive the rules. The increments will change at each derivation step due to the application of the rewriting rules altering the locations of atoms. The feedback between the L-system and the environment results in conformational changes following the physics of the model. This necessarily depends on parameters that must be defined for each atom for

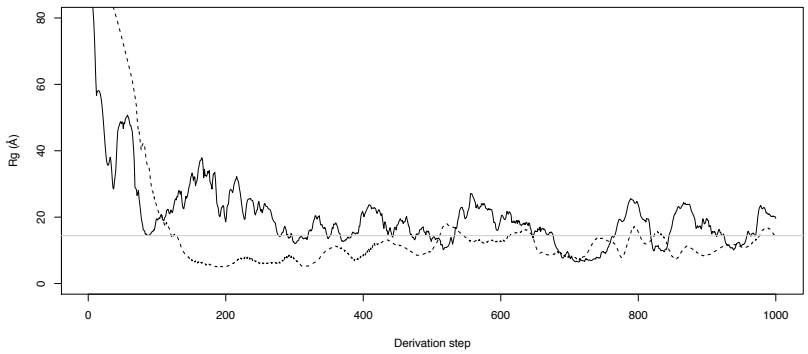

Fig. 6. Comparing the radius of gyration, Rg (a measure of compactness) at each derivation step of the L-system in the simple collision detection model (dashed line) and the forces model (solid line). The grey line shows the value of the native conformation of barnase. Both simulations lead to compact global structures from local rules.

use in calculations of forces. The parameters we use here are taken from the OPLS force field [11] used in atomistic condensed phase molecular simulations. The effects of this more physical set of rules are shown in Fig. 5.

Although neither model produces native-like conformations (plots in Figs. 4 and 5) both produce protein-like conformations as measured by the compactness of global conformations (globular proteins generally adopt compact native conformations) (Fig. 6). The use of a more physical rule set results in folding that is not forced by defining rules that drive the simulation to fold the protein in a predefined direction (as in the simple collision detection model), instead the successor of each rule, i.e. the increment of each torsion angle, is dependent on the local physical forces that change at each derivation step. The local conformations in this model are also more protein-like as seen by the angles adopted by each amino acid (Fig. 5) when compared with the allowed regions of a typical Ramachandran plot (Fig. 1).

5 Summary

The L-systems models we have presented show that incorporating even very simple collision detection produce complex global conformations that are also sequence dependent. This comes from inclusion of the environment. Protein folding is a problem of translating a 1D code to a 3D structure where the process is driven by physical rules. Replacing simple collision detection with physical forces prevents restrictions imposed by defining the rule successors. This allows folding to be governed by local physics of the environment and leads to more protein-like features. These features include characteristic local conformations shown by more realistic trajectories through ϕ, ψ space and compact global conformations.

The approach of using local rewriting rules has so far given interesting results as proof of concept. Our next step is to analyse the behaviour of the models both

in terms of the characteristics of their trajectories and resulting structures. The models may also be developed further to allow the other driving forces in protein folding - hydrophobic interactions and hydrogen bonding - to be additionally incorporated into the rule sets. The goal of this study is to try to discover to what extent protein folding may be modelled in terms of physical locally-determined conformational changes.

Acknowledgments This work is supported by the BBSRC.

References

1. Anfinsen, C.B.: Principles that govern the folding of protein chains. Science. **181** (1973) 223–239
2. Zwanzig, R., Szabo, A. and Bagchi, B.: Levinthal's paradox. Proc. Natl. Acad. Sci. USA. **89** (1992) 20–22
3. Lindenmayer, A.: Mathematical models for cellular interactions in development. Parts I and II. J. Theor. Biol. **18** (1968) 280–315
4. Prusinkiewicz, P. and Lindenmayer, A.: The Algorithmic Beauty of Plants. Springer, New York. (1990)
5. Mech, R. and Prusinkiewicz, P.: Visual Models of Plants Interacting with Their Environment. Proceedings of SIGGRAPH 96. (1996) 397–410
6. Dill, K.A.: Dominant forces in protein folding. Biochemistry. **29** (1990) 7133–7155
7. Onuchic, J.N. and Wolynes, P.G.: Theory of protein folding. Curr. Opin. Struct. Biol. **14** (2004) 70–75
8. Mirny, L. and Shakhnovich, E.: Protein folding theory: From lattice to all-atom models. Annu. Rev. Biophys. Biomol. Struct. **30** (2001) 361–396
9. Lau, K.F. and Dill, K.A.: A Lattice Statistical-Mechanics Model of the Conformational and Sequence-Spaces of Proteins. Macromolecules. **22** (1989) 3986–3997
10. Dill, K.A., Bromberg, S., Yue, K.Z., Fiebig, K.M., Yee, D.P., Thomas, P.D., and Chan, H.S.: Principles of protein-folding - a perspective from simple exact models. Protein Sci. **4** (1995) 561–602
11. Jorgensen, W.L, Maxwell, D.S. and Tirado-Rives, J.: Development and testing of the OPLS All-Atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. **118** (1996) 11225–11236
12. Escuela, G., Ochoa, G. and Krasnogor, N.: Evolving L-systems to capture protein structure native conformations. Proceedings of EuroGP2005. LNCS. **3447** (2005) 73–83.
13. Ochoa, G., Escuela, G. and Krasnogor, N.: Incorporating knowledge of secondary structures in a L-system-based encoding for protein folding. Proceedings of EA 2005. LNCS. **3871** (2006) 247–258
14. Danks, G.B., Stepney, S. and Caves, L.S.D.: Protein folding with L-systems: encoding the problem. (In prep.)
15. Ramachandran G.N., Ramakrishnan C., and Sasisekharan V.: Stereochemistry of polypeptide chain configurations. J. Mol. Biol. **7** (1963) 95–99
16. Prusinkiewicz, P., Karwowski, R., Mech, R., and Hanan, J.: L-Studio/cpfg: A Software System for Modeling Plants. Proceedings of AGTIVE99. LNCS. **1779** (2000) 457–464.
17. Mech, R. and Prusinkiewicz, P.: Users Manual for Environmental programs. (1998)