# Cotranslational protein folding with L-systems

Gemma B. Danks[1,2], Susan Stepney[2], and Leo S. D. Caves[2]

[1] Computational Biology Unit, Bergen Centre for Computational Science,
University of Bergen, 5008 Bergen, Norway
[2] York Centre for Complex Systems Analysis, University of York,
York, YO10 5YW, United Kingdom

**Abstract.** A protein molecule adopts a specific 3D structure, necessary for its function in the cell, through a process of folding. Modelling the folding process and predicting the final fold from the unique amino acid sequence remain challenging problems. We have previously described the application of L-systems, parallel rewriting rules, to modelling protein folding using two complementary approaches: a physics-based approach, using calculations of interatomic forces, and a knowledge-based approach, using data from fragments of known protein structures. Here we describe a model combining these two approaches creating an *adaptive* stochastic open L-systems model of protein folding. L-systems were originally developed to model growth and development. Here we also describe extensions of our L-systems models to investigate *cotranslational* protein folding, i.e. folding during protein biosynthesis on the ribosome, which is increasingly thought to play an important role. We demonstrate that cotranslational folding fits very naturally into the L-systems framework.

**Key words:** Cotranslational protein folding, L-systems

## 1 Introduction

Proteins are crucial for life. They carry out numerous essential molecular functions in the cell. The function of a protein molecule is determined by its specific 3D shape or *conformation*. A newly formed protein in the cell rapidly *folds* to its functional conformation. The thermodynamic hypothesis [1] states that this final fold, the *native state* of a protein, is its lowest free energy state. The native state of a protein, under physiological conditions, is determined solely by its amino acid sequence. Many features of protein folding are now well understood [2]. However, predicting the native state of a protein from its amino acid sequence alone and modelling the folding process at the atomic level on timescales greater than a microsecond are still not computationally feasible [3].

Proteins may be the simplest example of a biological complex system. They exhibit emergent properties at a range of spatial scales including: the partial double bond characteristic of the peptide bond; patterns of hydrogen bonding; *secondary* structure such as helices and sheets; and the compact and hydrophobic nature of the protein core. Each property is the result of interactions at lower spatial scales. These and other emergent properties of proteins allow them to fold to stable conformations with specific biological functions. Protein folding itself can be viewed as an emergent phenomenon that results from underlying local interactions. We use L-systems [4–6], parallel rewriting rules, to investigate what *global protein-like* characteristics emerge from modelling protein folding at a *local* level using local interactions represented by local rewriting rules. We have previously described the use of L-systems in modelling protein folding using a physics-based approach [7], using a calculation of local interatomic forces to guide rewriting rules, and a knowledge-based approach [8], using data from fragments of known protein structures. Here we describe an L-systems model that combines these two complementary approaches.

Most computational models of protein folding start with a fully formed protein chain. However, in the cell protein folding may occur during protein synthesis (*cotranslational* protein folding [9]). L-systems provide a natural framework for modelling growth. Here we also describe the extension of our L-systems models for *cotranslational* protein folding, i.e. protein folding during the "growth" of the chain.

Firstly, in section 2, we give a brief overview of how we use L-systems to model proteins. In section 3 we summarise the physics-based L-systems model (see [7] for more detail) and the knowledge-based L-systems model (see [8] for more detail). Section 4 describes the integration of these two approaches into a combined model leading to an *adaptive* stochastic L-systems model. Section 5 describes the extension of our L-systems models to incorporate cotranslational protein folding.

## 2 Modelling proteins with L-systems

Lindenmayer systems, or L-systems, were originally developed for the mathematical modelling of plant growth and development [4–6]. An L-system consists of a set of parallel rewriting rules, or *productions*, and an initial string called the *axiom*. The productions replace a symbol, or *module*, called the *predecessor*, with a string called the *successor* (e.g. $a \rightarrow ab$ replaces $a$ with $ab$) repeatedly for a number of specified *derivation steps*.

The axiom in our L-systems models consists of the amino acid sequence of a protein using the single letter amino acid code. We then use context-free productions to rewrite each amino acid letter in the axiom with a string that represents its component atoms and bonds using a bracketed system to capture the 3D structure (Fig. 1) of amino acid specific side chains.

For the folding process, we use context-sensitive productions to rewrite the structural state of each amino acid (captured as parameters in the L-system representation), depending on its local environment. The details depend on the particular model used.

We use deterministic L-systems in our physics-based model to rewrite the conformation of each amino acid depending on local interatomic forces. The interatomic forces are calculated using open L-systems, which allow an L-system to communicate with an environmental model.

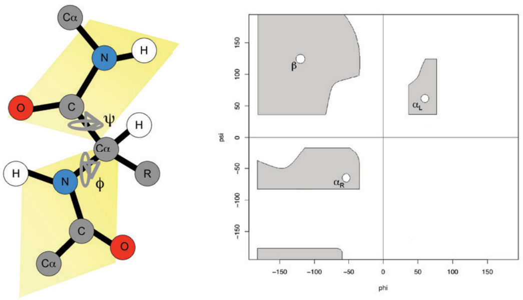


**Fig. 1.** Local conformations of a polypeptide. Left: an amino acid residue within a polypeptide (a chain of amino acids) consists of an -NH-CαHR-CO- backbone structure, where R represents an amino acid specific side chain (there are 20 different amino acids). Amino acids are joined together by semi-rigid peptide bonds (C-N) causing the four surrounding atoms (CαC-N-Cα) to lie in the same plane (shaded regions). The two main variables in protein conformation are therefore the two back bone torsion angles $\phi$ (rotation around the N-Cα bond) and $\psi$ (rotation around the Cα-C bond). Right: sterically allowed regions of $\phi/\psi$ space [10] are shaded in grey on a schematic of a typical Ramachandran plot. These correspond to the most common extended conformations in native structures - the $\alpha$-helix and $\beta$-sheet.

We use stochastic L-systems in our knowledge-based model to rewrite the secondary structure states of each amino acid *residue* with probabilities derived from data on known structures of protein fragments. We use an open L-system environment to store these context-sensitive probabilities.

## 3 Previous results

In our physics-based model [7] we define the conformation of a protein using the two backbone torsion angles, $\phi$ and $\psi$, for each amino acid residue in the chain (see Fig. 1). Productions alter the $\phi$ and $\psi$ values of each residue depending on local interatomic forces calculated in an environmental program. This alters the *local* conformation of each residue in parallel, resulting in a *global* change in conformation at each step (see [7] for more detail). The physics-based approach led to the emergence of *protein-like* compact global conformations.

In our knowledge-based model we use data on known protein structures in stochastic rewriting rules. We calculated the frequency of each amino acid type occurring in each of seven secondary structure states used by the DSSP program [11], given the amino acid type and secondary structure state of one amino acid residue either side. We use these frequencies as probabilities, in stochastic productions, of each residue changing its secondary structure state depending on the states and types of its immediate neighbours (see [8] for more detail).

The knowledge-based approach led to the emergence of bands of secondary structure indicating preferred local conformations for certain residues. The proportion of $\alpha$-helices and $\beta$-sheets emerging in the model for different protein sequences also corresponded well with the structural class of each protein. However, the structures emerging were not necessarily compact, in contrast to the physics-based model, and there was no convergence to a preferred global conformation. This is inevitable when using static probabilities - there is no criterion for choosing one likely state over another.

## 4 Combined model using adaptive stochastic L-systems

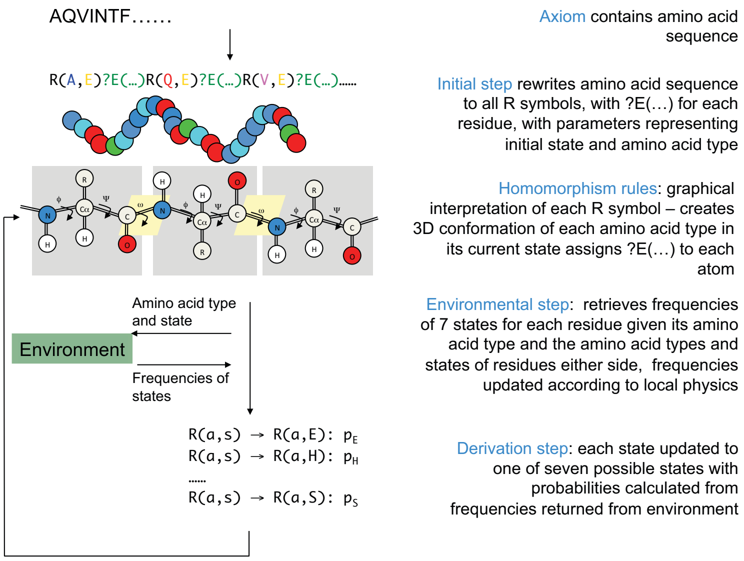
We have developed a model that combines physics-based and knowledge-based information in order to overcome the problems caused by using static probabilities described above. The local physics that informs changes in backbone torsion angles in the physics-based model is used instead to dynamically alter the probabilities of changing to another secondary structure state in the knowledge-based model (see Fig. 2 for an outline of the combined L-systems model).

Interatomic forces (from an empirical potential) are calculated between each atom attached to the backbone and any other nearby atoms. Changes in both $\phi$ and $\psi$ are calculated for each residue according to these local forces. The environment also calculates, using typical torsion angle values for each residue in each secondary structure state, the change in $\phi$ and change in $\psi$ that would be required for each residue to move from its current secondary structure state to each of the other possible states. These are compared to the changes in $\phi$ and $\psi$ that were calculated from the local forces. The frequencies are then updated in proportion to these differences. This is repeated at each derivation step - frequency values are updated by scaling values from the previous step according to the forces that result from the new conformation at the current step. This allows the gradual accumulation of a physics-bias into the frequencies, some of which may decrease to zero.

This model leads to a better *protein-like* convergence to a preferred global conformation than the knowledge-based model, while retaining bands of local secondary structure preferences with proportions of $\alpha$-helix and $\beta$-sheet that fit well with the structural class of each protein sequence. Convergence to a preferred global conformation is better in the all-$\alpha$ and all-$\beta$ structural classes. However, these preferred conformations are not necessarily compact. The final conformation is sensitive to the choice of initial states (particularly in $\alpha/\beta$ or $\alpha$+$\beta$ structural classes). An all-extended initial conformation leads to a greater number of residues adopting extended states. Similarly an all-$\alpha$ initial conformation leads to a greater number of residues adopting $\alpha$-helix states. This may be important in the context of cotranslational folding.

## 5 Modelling cotranslational protein folding with L-systems

The specific amino acid sequence of a protein molecule is formed during its biosynthesis. Protein-coding genes are *transcribed* into messenger RNA (mRNA)



Axiom contains amino acid sequence

Initial step rewrites amino acid sequence to all R symbols, with ?E(…) for each residue, with parameters representing initial state and amino acid type

Homomorphism rules: graphical interpretation of each R symbol – creates 3D conformation of each amino acid type in its current state assigns ?E(…) to each atom

Environmental step: retrieves frequencies of 7 states for each residue given its amino acid type and the amino acid types and states of residues either side, frequencies updated according to local physics

Derivation step: each state updated to one of seven possible states with probabilities calculated from frequencies returned from environment

**Fig. 2.** An outline of the stages in the combined L-systems model. Information at both residue-level (specified in the string) and atomic-level (via homomorphism rules) is sent to the environment via communication modules. The environment retrieves the frequencies for the probabilistic rewriting rules using the information in the residue-level communication modules. The frequencies are altered using physics-based information calculated using the information in the atomic-level communication modules.

molecules that are *translated* by ribosomes, the macromolecular protein factories of the cell. During translation a ribosome concatenates amino acid building blocks in the order specified by the mRNA being read. Amino acids are concatenated one at a time, by peptide bonds, to form a growing polypeptide which is gradually extruded through a tunnel in the ribosome [12]. Upon exiting the ribosome the polypeptide is free to begin the folding process. Since formation of secondary structure and compact states is faster than protein synthesis [13, 14], this simultaneous growth and folding may be important in finding the native state. Furthermore the ribosome itself imposes physical constraints to the initial conformation [12]. L-systems were originally developed to model plant growth and development [4–6]. We have extended our L-systems models described above to model the growth of a polypeptide chain and its simultaneous folding, i.e. co-translational protein folding.

Three main features of cotranslational folding have been simplified and incorporated into our L-systems model: protein synthesis; passage through the ribosome; and extrusion from the ribosome.

**Modelling protein synthesis.** The full protein sequence is contained in the axiom. A parameter is added to each amino acid module in the axiom to represent its position in the sequence (in the N-terminal to C-terminal direction). A condition is added to the rewriting rules that generate the structural representation of each amino acid. This condition allows one amino acid module to be rewritten at the C-terminal end of a partially formed chain of length, $r$ (representing the number of residues that are in the polypeptide exit tunnel but are unable to fold), every $e$ derivation steps (representing the rate of protein synthesis): $N > ((n * e) - r)$, where $N$ is the current derivation step number and $n$ is the amino acid number in the sequence.

**Modelling restrictions of the ribosome.** The polypeptide exit tunnel is approximately 80Å in length [15] and experimental evidence shows that it contains 30-40 amino acid residues [16, 17]. We took the lower of these estimates for the number of residues held in the ribosome, $r$, in our model.

**Modelling folding on extrusion from the ribosome.** A condition is incorporated into the rewriting rules that alter the secondary structure states of residues (or the backbone torsion angles in the physics-based model) so that only residues at the N-terminal end that are outside of the polypeptide exit tunnel can start folding: $N > (n * e)$. This allows one residue at the N-terminal end to start folding every $e$ derivation steps, as one amino acid structure is added to the C-terminal end. Once the protein is fully formed and all residues are out of the ribosome, all residues can fold in parallel until a specified derivation length.

**Results.** Protein folding in the cell may be cotranslational if the partially formed polypeptide can adopt a stable conformation. In our physics-based cotranslational model the partially formed polypeptide may rapidly adopt a compact conformation once outside the ribosome (Fig. 3).

The emergence of the final fold through global conformational changes may be dependent on the history of local conformational changes. The rate of protein synthesis may affect this history. In the cell, pauses in translation and the use of rare mRNA codons cause the rate of protein synthesis to vary. Using the combined model (the integrated physics-based and knowledge-based model) we find that cotranslational folding alters the local secondary structure preferences in certain residues and that this is dependent on the growth rate, $e$, of the polypeptide chain (Fig. 4).

Our cotranslational L-systems models put protein folding into a more biological context. Folding on the ribosome during protein synthesis may be important to finding the native state [14, 9]. We have shown that L-systems provide a natural modelling framework for investigating cotranslational protein folding. The L-systems framework facilitates the integration of the growth process of protein synthesis and the developmental process of protein folding, through local conformational changes, into a single model.

## 6 Summary

Our previous work describes the development of L-systems models of protein folding using two complementary approaches: a knowledge-based approach and a physics-based approach. Here we describe how these models were integrated to produce a combined *adaptive* stochastic open L-systems model, which gives

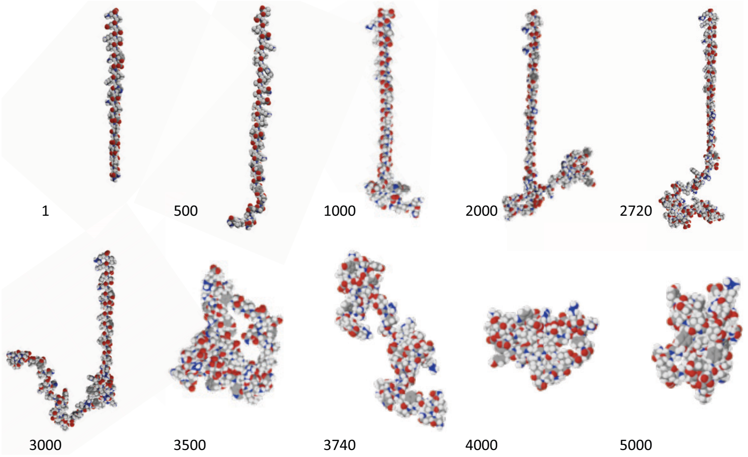
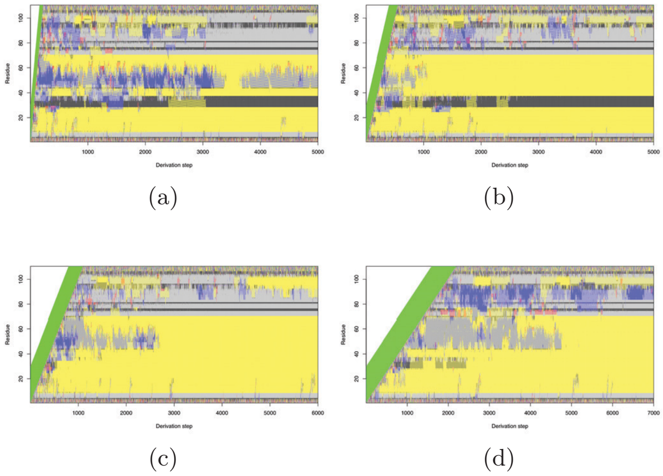


**Fig. 3.** Images showing protein folding in the cotranslational physics-based model, at derivation step numbers as shown, using the protein sequence barnase (PDB ID: 1bnr). Step 1 shows the initial $r$ residues (here 30) in an initial conformation (here a $\beta$-strand) modelling the partially formed polypeptide that is unable to fold inside the ribosome. One residue is added to the C-terminal end every $e$ steps (here 34), while one residue at a time is allowed to start folding at the N-terminal end.



(a) (b) (c) (d)

**Fig. 4.** Changing the rate of growth, $e$, of the polypeptide chain, using the protein sequence '1exg' in an all-$\pi$-helix initial state, in the cotranslational combined model. Plots show the secondary structure states of each residue (y-axis) at each derivation step (x-axis). Yellow = extended, blue = $\alpha$-helix, red = 3/10 helix, green = $\pi$-helix, purple = isolated beta bridge, grey = turn and black = bend. (a) $e$ = 2 steps per residue (b) $e$ = 5, (c) $e$ = 10 and (d) $e$ = 20.

a greater degree of *protein-like* convergence to a preferred global conformation. We also present a framework for investigating *cotranslational* protein folding using L-systems. This puts protein folding into a more biological context. We demonstrate that L-systems provide a natural framework for modelling the simultaneous growth and folding of a polypeptide chain. Initial results show that the rate of protein synthesis influences the preferred secondary structure preference of some residues in our combined model. Further work will include a more explicit model of the ribosome and will investigate the effects of the rate of protein synthesis across a wide range of protein sequences.

**Acknowledgments.** This work was funded by the BBSRC.

## References

1. Anfinsen, C.B.: Principles that govern the folding of protein chains. Science **181**(96) (1973) 223–230
2. Dill, K.A., Ozkan, S.B., Shell, M.S., Weikl, T.R.: The protein folding problem. Annu. Rev. Biophys. **37** (2008) 289–316
3. Daggett, V.: Protein folding-simulation. Chem. Rev. **106**(5) (2006) 1898–1916
4. Lindenmayer, A.: Mathematical models for cellular interactions in development. I. Filaments with one-sided inputs. J. Theor. Biol. **18**(3) (1968) 280–299
5. Lindenmayer, A.: Mathematical models for cellular interactions in development. II. Simple and branching filaments with two-sided inputs. J. Theor. Biol. **18**(3) (1968) 300–315
6. Prusinkiewicz, P., Lindenmayer, A.: The Algorithmic Beauty of Plants. Springer (1990)
7. Danks, G.B., Stepney, S., Caves, L.S.D.: Folding protein-like structures with open L-systems. In: ECAL2007. Volume 4648 of LNAI., Springer (2007) 1100–1109
8. Danks, G.B., Stepney, S., Caves, L.S.D.: Protein folding with stochastic L-systems. In: Artificial Life XI, MIT Press (2008) 150–157
9. Kolb, V.A.: Cotranslational protein folding. Mol. Biol. **35**(4) (2001) 584–590
10. Ramachandran, G.N., Sasisekharan, V.: Conformation of polypeptides and proteins. Adv. Protein Chem. **23** (1968) 283–438
11. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**(12) (1983) 2577–2637
12. Nissen, P., Hansen, J., Ban, N., Moore, P.B., Steitz, T.A.: The structural basis of ribosome activity in peptide bond synthesis. Science **289**(5481) (2000) 920–930
13. Kubelka, J., Hofrichter, J., Eaton, W.A.: The protein folding 'speed limit'. Curr. Opin. Struc. Biol. **14**(1) (2004) 76–88
14. Basharov, M.A.: Protein folding. J. Cell. Mol. Med. **7**(3) (2003) 223–237
15. Voss, N.R., Gerstein, M., Steitz, T.A., Moore, P.B.: The geometry of the ribosomal polypeptide exit tunnel. J. Mol. Biol. **360**(4) (2006) 893–906
16. Malkin, L.I., Rich, A.: Partial resistance of nascent polypeptide chains to proteolytic digestion due to ribosomal shielding. J. Mol. Biol. **26**(2) (1967) 329–346
17. Blobel, G., Sabatini, D.D.: Controlled proteolysis of nascent polypeptides in rat liver cell fractions. I. Locations of polypeptides within ribosomes. J. Cell Biol. **45**(1) (1970) 130–145