BIOINFORMATICS

Vol. 22 no. 14 2006, pages e203–e210
doi:10.1093/bioinformatics/btl248

Modelling sequential protein folding under kinetic control

Fabien P.E. Huard[1,*], Charlotte M. Deane[2] and Graham R. Wood[1]

[1]Department of Statistics, Macquarie University, NSW 2109, Australia and [2]Department of Statistics, 1 South Park Road, Oxford OX1 3TG, UK

ABSTRACT

Motivation: This study presents a novel investigation of the effect of kinetic control on cotranslational protein folding. We demonstrate the effect using simple HP lattice models and show that the cotranslational folding of proteins under kinetic control has a significant impact on the final conformation. Differences arise if nature is not capable of pushing a partially folded protein back over a large energy barrier. For this reason we argue that such constraints should be incorporated into structure prediction techniques. We introduce a finite surmountable energy barrier which allows partially formed chains to partly unfold, and permits us to enumerate exhaustively all energy pathways.

Results: We compare the ground states obtained sequentially with the global ground states of designing sequences (those with a unique global ground state). We find that the sequential ground states become less numerous and more compact as the surmountable energy barrier increases. We also introduce a probabilistic model to describe the distribution of final folds and allow partial settling to the Boltzmann distribution of states at each stage. As a result, conformations with the highest probability of final occurrence are not necessarily the ones of lowest energy.

Availability: Software available on request

Contact: fhuard@efs.mq.edu.au

1 INTRODUCTION

There have been several definitions of cotranslational folding, but it has been elegantly stated that ''co-translational folding has occurred if, following extrusion from the ribosome, the native structure is achieved more quickly than if the full-length, unfolded polypeptide were diluted from chemical denaturant into the same folding milieu as that in which protein biosynthesis occurred'' (Baldwin, 1999). It is recognised that some proteins can fold rapidly and cotranslationally both in eukaryotic and prokaryotic cells (Basharov, 2003; Braakman *et al*., 1991; Fedorov and Baldwin, 1997; Fedorov and Baldwin, 1997; Kolb, 2001; Kolb *et al*., 2000; Netzer and Hartl, 1997) and there is recent evidence that some proteins become *in vivo* biologically active as the polypeptide chain is being translated (Nicola *et al*., 1999). We also know that cotranslational folding can occur spontaneously without additional cellular components (Sanchez *et al*., 2004). Interestingly, nitinol wire, known to remember its annealed shape, has been used to model behaviour of biopolymers and showed that in some cases the native state could only be reached sequentially (Keller, 2003).

Levinthal pointed out that the protein folding process cannot search the entire conformation space due to its vast size. Since proteins are known to fold in the order of milliseconds, we must assume that they follow a restricted set of pathways to reach their native conformation (Levinthal, 1968; Levinthal, 1969). Hence folding is assumed to be under kinetic control, that is, the folding pathway of a protein is unlikely to incorporate folding to a state which would be less thermodynamically stable. It was advanced that protein folding obeys thermodynamical laws and therefore has a native state which is the ground state of lowest free energy (Anfinsen, 1973). It has been theoretically demonstrated (Govindarajan and Goldstein, 1998) that a sequence whose native state has originally a higher energy than the lowest energy state, when submitted to evolution under kinetic control, will most often evolve towards a sequence whose native state is the lowest energy conformation. Thus folding under kinetic control does not necessarily violate the thermodynamical hypothesis.

Surprisingly, state-of-the-art protein folding prediction methods do not incorporate a cotranslational aspect (Bujnicki, 2006); in the latest Critical Assessment of Techniques for Protein Structure Prediction meeting (CASP, 2004) none of the chosen methods exploited the sequential nature of folding. Cotranslation has already been investigated in simulations of biopolymers (Bornberg-Bauer, 1997; Fernandez, 1994; Morrissey *et al*., 2004), but the effect of kinetic control remains unexplored. The method we propose aims at filling this gap; we investigate the effect of energy barriers on cotranslation.

We fold proteins sequentially, mimicking nature as closely as possible. By a ''sequential folding'' we will refer to the path of intermediate and final conformations simulated as the nascent polypeptide chain is elongated. A ''sequential ground state'' is a conformation of lowest energy obtained once all residues are added. We simulate protein fold evolution, as the polypeptide chain length increases, by sequentially elongating the length of protein to be folded, starting from the N-terminus. Amino acids are added one by one at the C-terminus of the chain and each time the chain length increases by one residue, the conformation already simulated is permitted to change. The point here is that the new fold must be a ''restricted evolution'' of the previously predicted fold. By this we mean that the simulation of the newly elongated chain does not start with a random or fully extended conformation, but with the previous model obtained as a base, to which is added the new residue. The latter is added in a fully extended conformation. We also investigate the possibility of adding more than one residue at a time. The final fold of the protein is obtained once all residues are added.

Essential here is the concept of a surmountable energy barrier (Baker, 1998; Guo *et al*., 1997; Sohl *et al*., 1998), the orchestrator of kinetic control. The surmountable energy barrier enables us to partly avoid kinetic traps, and represents the maximum energy gain

*To whom correspondence should be addressed.

possible for the protein at each step of its folding process. It is essentially the unfolding energy available in the system. In the following cases of folding under kinetic control, this surmountable energy barrier is assumed to be finite. Rationale for the imposition of a surmountable energy barrier comes from a number of sources. We know that ~20% of proteins require intervention of chaperones, which play an important role in cotranslation (Frydman, 2001; Hartl and Hayer-Hartl, 2002). It is believed that the primary role of chaperones is to prevent aggregation of nascent polypeptides. The surmountable energy barrier aims at representing the restriction on the folding pathways induced by chaperones. We also know that folding space is restricted by the structure of the ribosome itself (Ban *et al*., 2000; Ramakrishnan, 2002; Wilson *et al*., 2002). In particular, the fold of polypeptides is constrained by the ribosome exit tunnel (Jenni and Bany, 2003; Nakatogawa and Ito, 2002) which favours α-helical secondary structures (Ziv *et al*., 2005).

We know that some codons are less frequent than others, inducing different translation rates (Andersson and Kurland, 1990; Curran and Yarus, 1989) and that codon substitutions can lead to lower specific activity (Komar *et al*., 1999). Slow codons, usually positioned between domains, can induce a delay required for correct folding of the N-terminus domain (Komar and Jaenicke, 1995). Slow codons can also enhance the formation of secondary structures by preventing domains from interacting with each other (Purvis *et al*., 1987). To model the variation in translation rate imposed by codon selection, we introduce parameter s which represents the number of residues added each time the polypeptide chain is elongated. This creates a primitive ''elongate-pause'' iterative extension process.

We also attach a probability to all partial and fully extended conformations. It has been observed that the biologically active state of some proteins does not correspond to their lowest energy conformation (Sohl *et al*., 1998). We introduce a probabilistic model which captures two factors. The first factor is the number of kinetically controlled energy pathways which can lead to the conformation (relative to the number of possible conformations for the considered sequence). The second factor is the Boltzmann equilibrium distribution for the current set of partial configurations. We balance the two factors using a ''thermodynamic permission factor'' β. This measures the extent to which the Boltzmann distribution is reached. We investigate whether kinetic control together with partial movement to the Boltzmann distribution can result in a sequential ground state whose energy may be a local minimum in the thermodynamic energy path of the protein, as observed experimentally.

HP lattice models have proven a useful tool for modelling protein folding in a simple manner (Chan and Dill, 1993; Chan and Dill, 1994; Dill *et al*., 1995; Pande *et al*., 1997; Shakhnovich, 1998), predicated on the assumption that protein folding is ruled by hydrophobic collapse. Here we use them to assess the impact of sequential folding. Sequences involving only two types of monomer (hydrophobic H and polar P) are considered, with monomer positions restricted to either a two or three-dimensional lattice. Simple models have been used to simulate globular protein folding incorporating cotranslation and restrictions on the folding space, modelling the ribosome as an inert wall (Sikorski and Skolnick, 1990). It was found that α-helical proteins preferred to assemble parallel to the wall, and four member β-barrels slightly preferred assembly perpendicular to the wall. Sikorski and Skolnick ''never observed a successful case of co-translational folding'' and did not consider kinetic control. They used a Monte Carlo algorithm to search the conformation space and pass through local minima, whereas we develop a fully deterministic approach and exhaustively search the conformation space.

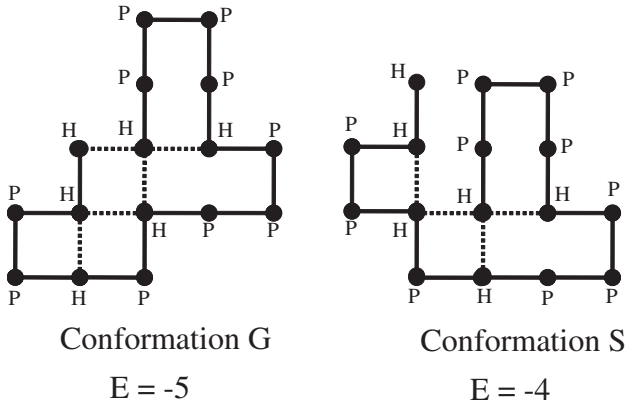

Fig. 1. Conformations obtained for the sequence HPPPPHPPPHPHPPHH. Conformation G represents the global ground state, the unique conformation which has an energy of minus five for this particular sequence. Conformation S is that obtained sequentially, with energy of minus four, when the surmountable energy barrier is zero.

In summary, we explore the consequences of following a sequential route to the final fold. In particular, we study the influence on final conformations of the height of the surmountable energy barrier d and the number of residues s added at each iteration. We find that under kinetic control the sequential ground state of a protein can differ from the global one (Figure 1). The global state of minimum energy can be reached only with a sufficiently high surmountable energy barrier.

We then present the impact of the variation of the main parameters (extrusion length and surmountable energy barrier) on the compactness and multiplicity of the folds. For a given sequence, we observe that final conformations are more compact and less numerous as we increase the surmountable energy barrier.

Finally we enrich our analysis and introduce a probabilistic model based on partial movement to Boltzmann equilibrium at each stage. This enables us to attach a probability to all partial or final conformations obtained for a particular sequence.

2 METHODS

2.1 Principles

Designing sequences We use designing HP sequences in our study. These are sequences with a unique ground state of lowest energy. Irbäck *et al*. (Irbäck and Troein, 2002) present a list of all designing sequences with up to 24 residues. This provides us with reference sequences against which we can test the sequential folding algorithm.

HP Lattice models We use models which fold on a two-dimensional lattice with residues either hydrophobic or polar. They are said to be in contact if they are adjacent in space but not in sequence. The total energy of the chain is determined by the number of contacts in the conformation simulated.

We let n be the number of residues in the full chain. To study the impact of the variation of the chain length, n takes the value 16 or 24. Evidence has been given that such lengths are capable of mimicking relevant protein behaviour (Chan and Dill, 1993; Chan and Dill, 1994; Dill *et al*., 1995; Pande *et al*., 1997; Shakhnovich, 1998).

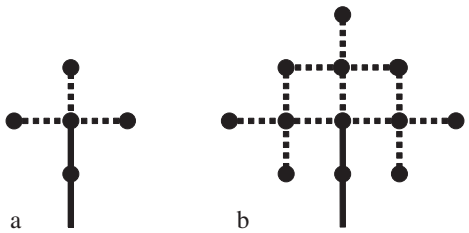

Fig. 2. The different ways to extend a conformation, adding one residue (a) or two (b) at a time. The plain line represents the extremity of the conformation already simulated, and the dashed lines the possible extensions.

Sequential Folding Since we work with relatively short lengths n, the number of monomers s added at each iteration is chosen to be one or two. Sequences of length 16 can have a maximum of nine contacts, so it is reasonable to perform simulations with d, the surmountable energy barrier, equal to zero, one or two.

The first s monomers are laid down, locating them in a conformation with minimum energy, at the same time retaining all configurations within energy d of this minimum. We then have a first set of local conformations of length s and proceed to expand these by adding s monomers to all of these partial configurations, retaining those with minimum energy and all within energy d of this new local minimum. Parameter d remains the surmountable energy barrier, so leading to a new set of local conformations of length $2s$. This procedure is repeated until all monomers are used. A configuration with minimum final energy is termed a ''sequential ground state'', and the one of lowest energy the ''global ground state''.

A conformation C_l, of length l, is extended by s residues using s steps of the three possible single step directions (Figure 2). These three possible directions are—in relative moves—forward, left and right. Only conformations which are self-avoiding and non equivalent are retained. Two conformations are deemed equivalent if one can be obtained either by rotation or reflection on the lattice from the other. At each step we obtain a maximum of three new conformations of length $l+s$. The process is then repeated with each one of these conformations of length $l+s$, and so on until we generate conformations of length n. If n is not a multiple of s, then the algorithm is run for $\lfloor n/s \rfloor$ steps; the last iteration handles the remaining residues.

2.2 Measures of fold compactness

As explained in the introduction, we wish to study the impact of folding sequentially, considering the surmountable energy barrier d, the number of residues added at each iteration of the algorithm s and length of the polypeptide chain n. To assess the final fold we use several measures.

Radius of gyration We calculate the radius of gyration of conformations, as used in real protein structure prediction (Rohl *et al*., 2004; Simons *et al*., 1997; Simons *et al*., 1999), using

$$R_g = \sqrt{\sum_{i=1}^{n} \sum_{j\langle i} [(x_{i1}-x_{j1})^2 + (xi_2-xj_2)^2]}$$

where $x_i = (x_{i1}, x_{i2})$ represents the two coordinates of point i and n is the number of residues in the conformation.

Moment of inertia We use the moment of inertia (MI) as an indicator of the compactness of the structure. It reflects the variance of distances from residues to the centre of mass of the conformation,

$$MI = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_{i1}-\mu_1)^2 + (xi_2-\mu_2)^2]$$

where $\mu = (\mu_1, \mu_2)$ with $\mu_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1}$ and $\mu_2 = \frac{1}{n}\sum_{i=1}^{n} xi_2$

We also use a MI restricted to hydrophobic residues. In this case we term the result the hydrophobic moment of inertia (HMI).

Contact signature We define the contact signature S of a conformation to be the average distance in sequence between two residues in contact. So we have

$$S = \frac{\sum_{i<j} d(i,j)\Delta(i,j)}{N_{contacts}}$$

where $d(i,j)=j-i$ is the distance in sequence between the residues at position i and position j and $\Delta(i,j)$ equals one if residues i and j are in contact and zero otherwise; $N_{contacts}$ is the number of contacts in the chain.

3 RESULTS

We use HP models to investigate the difference between the minimum energy state of a controlled sequential folding and the globally minimum energy state. A difference in these two end states will be found if nature is incapable of pushing a partially formed protein back over a sufficiently high free energy barrier. We explore the influence on this difference of n, d and s.

For a particular sequence, the number of final sequential conformations at the minimum energy level decreases as the surmountable energy barrier increases We focus on 149 randomly selected designing sequences of length 16 whose unique global ground state is known. We extrude one residue at a time, so s is equal to one. We first set the surmountable energy barrier d at zero. We observe that for 48 sequences (32.2%) we obtain a unique sequential ground state, which is not necessarily the global ground state. The number of sequences with a unique sequential ground state increases to 95 (63.8%) as we raise d to one. These results suggest that for a given sequence, the number of final conformations decreases as the surmountable energy barrier increases.

As we increase d, the number of local conformations (as described in methods) retained at each step of the elongation increases. Those which are kept have an energy within d of the lowest. If more conformations are simulated, the probability of retaining the global ground state of energy rises. With a surmountable energy barrier sufficiently high, it is possible to enumerate all conformations and then be sure of obtaining the global ground state. Increasing the number of residues extruded at a time has a similar effect. Adding more than one residue at a time increases the number of intermediate conformations simulated as well as the odds of retaining the global ground state.

For the sequence HPPPPHPPPHPHPPHH, for example, a surmountable energy barrier of one is sufficient to access the global state (Figure 3).

Conformations become tighter as the surmountable energy barrier increases Given a particular sequence there are many final sequential folds (with the same energy) for a given s and d. We measure the compactness of the structure with the radius of gyration R_g. We determine the average R_g over all such conformations sequentially generated for a particular sequence. As d increases, the average R_g decreases. We find that for 88% of the sequences, this average R_g remains the same or registers a decrease when we increase d from zero to one, with s equal to one. An example is given in Figure 4.

We also evaluate an average hydrophobic moment of inertia (HMI) of all sequential ground states obtained for a particular

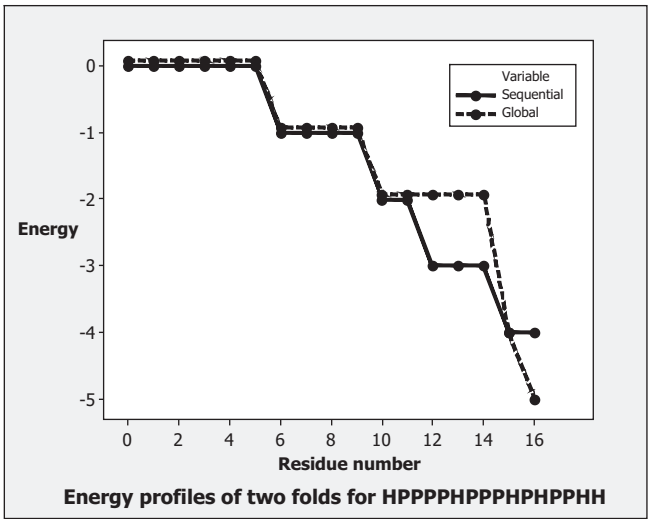

Fig. 3. The energy, in units of $-\varepsilon$, is plotted against the number of residues in the sequence as the chain elongates. The solid line represents the sequential energy path (with d=0 and s=1). The dashed line represents the energy path of the global ground state; note that this path is not influenced by s or d. We observe that the global ground state path is eliminated from the pool of sequential local conformations when the 12th residue is added. At this point the sequential algorithm produces partially extruded conformations with lower energy (one contact). So in this case a surmountable energy barrier of one would be sufficient to retain the path leading to the ground state.

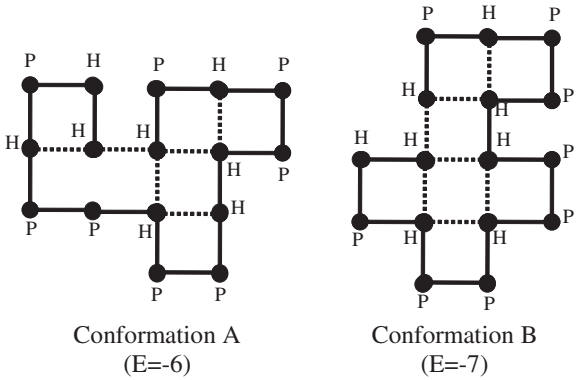

Fig. 4. The graphics show, for the sequence HPHPPHHPPHPPHPHH, the sequential ground state simulated with a surmountable energy barrier equal to zero (left), and equal to one (right). The radius of gyration decreases from 2.405 (left) to 2.377 (right). Both simulations led to a unique sequential ground state.

sequence. We find that the hydrophobic core forms as the surmountable energy barrier increases. We then calculate the difference between this average HMI and the HMI of the global ground state of minimum energy. We observe that as d increases the average HMI of the sequential ground states simulated moves closer to the global HMI. We observe that as d increases, the energy level of final conformations simulated tends to be closer to the energy level of the unique global ground state. The global ground state has the maximum number of contacts possible; hence it generally also has the tightest hydrophobic core. So the closer the conformations are to the

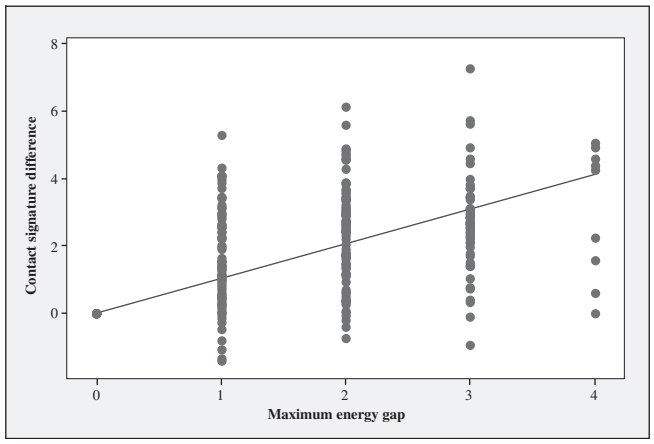

Fig. 5. We evaluate the difference between the average of the sequential contact signatures and the global contact signature. We also determine the maximum energy gap (in the energy paths) between the sequential path of lowest energy and that of the global ground state. The signature difference is plotted against the largest gap in energy. The superimposed points with a null energy gap and a null contact signature difference correspond to the cases where the final sequential conformation is always the global one.

global ground state, the tighter their hydrophobic core is likely to be. In all of the 149 sequences simulated, we observe that 65% (97) have an average HMI which decreases when we increase d from zero to one, and 19.5% (29) have an average HMI which remains the same.

Sequentiality favours short range contacts The further the energy of the sequential path is from that of the global path, the more localized the contacts become. We randomly select 242 sequences of 24 residues and run simulations with d equal to zero and s equal to one. For each sequence we then evaluate the average of the sequential contact signatures, and calculate the difference with the global contact signature. We find that in 89.7% of the cases, the average sequential contact signature is less than the global. We also notice a positive relation when we plot the biggest energy gap for each sequence against the difference in contact signature (Figure 5). These results confirm a previous study which showed that cotranslationality favours local contacts (Morrissey *et al*., 2004).

Some sequences are not foldable sequentially with a low surmountable energy barrier The method explores exhaustively all possible conformations accessible sequentially. Some particular sets of intermediate conformations may result in non-extendable conformations. These are conformations which have folded into a state that cannot be extended to reach the full length conformation. It is possible to avoid these dead-end conformations by increasing the surmountable energy barrier. An increase in d permits a higher number of intermediate conformations to be retained at each iteration of the elongation, and thus reduces the chance that an iteration results only in conformations which cannot be extended. We assume that these conformations which cannot be modelled sequentially with a low surmountable energy barrier cannot represent proteins which have mutated through evolution. We conclude that biological sequences must evolve to avoid sequences which can fall into such traps.

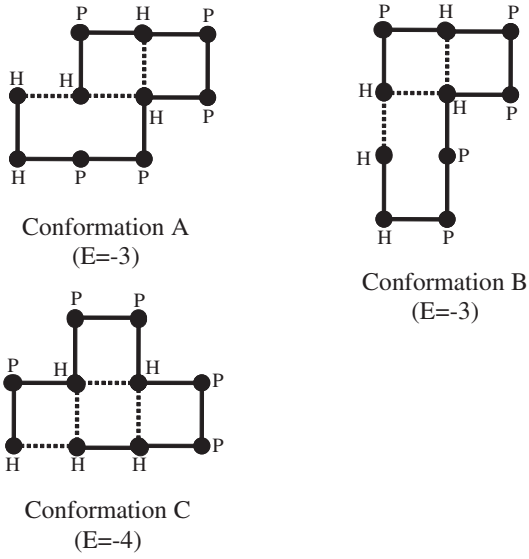

Fig. 6. Conformations A and B represent respectively two (out of eight) sequential ground states with three contacts each, and conformation C shows the unique global ground state for the sequence.

Analysis of energetic pathways of sequentially folded proteins We focus on the 10-mer HPHPPHPPHH. We simulate the ground states obtained with a surmountable energy barrier d=0, adding one residue at a time, so s=1. This sequence corresponds to the shortest designing sequence available for which the unique global ground state fold differs from the global ground state under the preceding conditions. Figure 6 shows two of the conformations obtained sequentially and that of the global ground state. The global ground state can only be reached with a surmountable energy barrier of one.

Figure 7 shows the energy paths of the sequential ground states and the global ground state. When the sixth residue is added, the best fold modelled sequentially has one more contact than the path towards the global ground state of energy. Since the surmountable energy barrier is zero, the path to the global ground state is not retained. Having a null probability of occurrence, the ground state is eliminated from the pool of potential final folds (Figure 8).

Definition of a probabilistic 2D simple lattice model The surmountable energy barrier allows a set of conformations to be retained at each elongation of the chain, and these may have different energies. As a consequence, there may also be a set of final conformations for a given sequence. We want to be able to attach a probability to each of these conformations, partially or fully elongated.

We know that some proteins in their native state are not in their lowest Gibbs free energy state, and fold to a state more stable than the native one (Baker, 1998; Sohl *et al*., 1998). Baskakov *et al*. showed for instance that the folding of mouse prion protein was under kinetic control when folding to its α-helical native conformation, separated by a large energy barrier from a more thermodynamically stable β-sheet-rich isoform (Baskakov *et al*., 2001). Therefore we accept that the intermediate conformations accessed by the polypeptide, as it is elongated, may also not be in a lowest free energy state. In order to model this we do not permit the

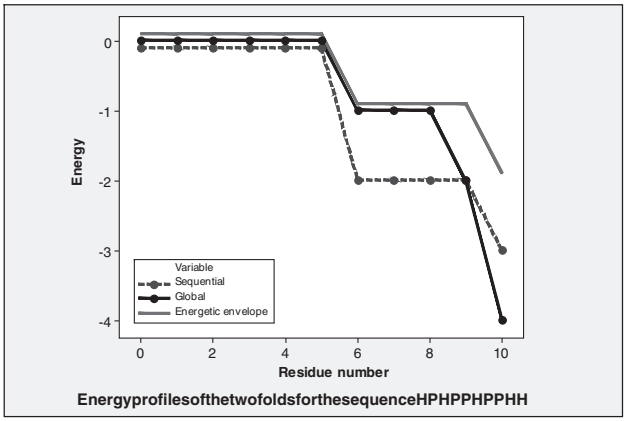

Fig. 7. A plot of the common energy path of the minimum energy sequential folds (with d=0 and s=1) and the global fold for the sequence HPHPPHPPHH. Also shown is the upper energy envelope; all energy paths lying below this envelope are considered in the analysis.

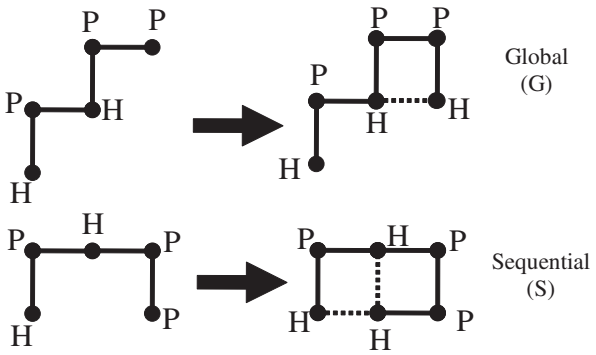

Fig. 8. Conformation G with one contact leads to the global state of energy (for the full length). Conformation G is one of the seven possible conformations of length six with one contact. Conformation S shows the only possible conformation of length six with two contacts. Conformation S is the intermediate conformation of length six which has the lowest energy.

distribution of conformations to reach the Boltzmann energy distribution completely and we introduce a ''thermodynamic permission factor'' β ($0 \leq \beta \leq 1$). This factor is a coefficient permitting movement to the Boltzmann equilibrium probability of every conformation, partially or fully extended.

We now model the probabilities of intermediate conformations along the different energy pathways. The probabilistic model defines a distribution for each intermediate and final model which is the sum of two components, an initial probability weighted by *1-*β and the Boltzmann probability weighted by β. The initial probability is the parent conformation probability divided by the number of offspring of this parent conformation, so is determined by the different elongation paths. If several conformations, after elongation, result in the same offspring conformation, the latter has a chance of occurrence which is the sum of the probabilities of the common offspring. As we assume that the pool of intermediate conformations may not reach the Boltzmann equilibrium, the Boltzmann equilibrium distribution is weighted by β.

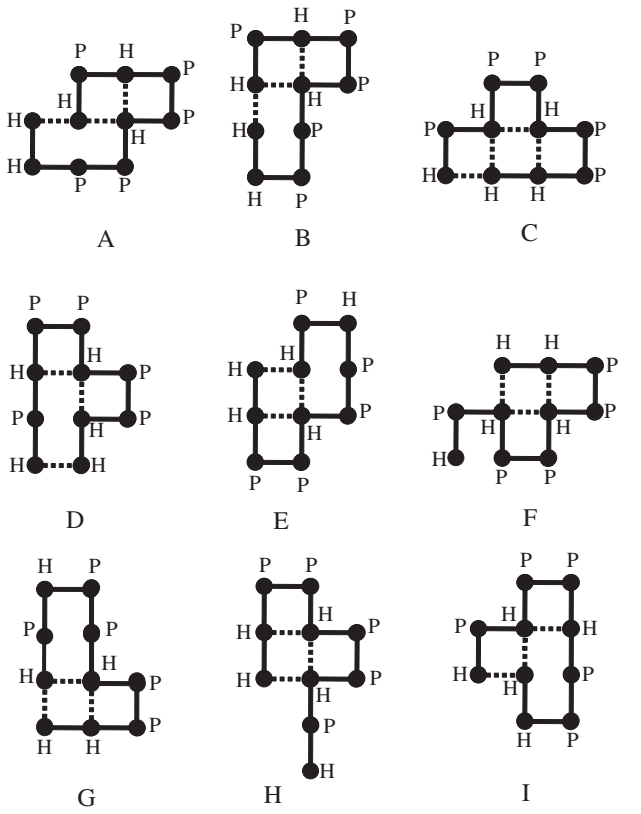

Fig. 9. The nine final conformations obtained sequentially for the sequence HPHPPHPPHH using *s=1* and *d=1*.

Given a surmountable energy barrier d we have, for a chain of l residues, a known distribution of n^l intermediate conformations C_i^l, $i=1,\ldots,n^l$ with known probabilities p_i^l. We elongate all intermediate conformations of length l by s residues. There arises a new set of n^{l+s} intermediate conformations of length $l+s$.

We assume that all newly modeled C_i^{l+s}conformations of length $l+s$ have an immediate probability I_i^{l+s} which is followed in time by a final probability F_i^{l+s}. We know that a given conformation C_i^l can give birth to a number b_i^l of kinetically permissible different conformations of length $l+s$, and that a given conformation C_i^{l+s} can have a_i^{l+s}different ancestors of length l.

We define the initial probability of C_i^{l+s} which has $a_i^{l+s}=a$ ancestors $C_{i1}^l, C_{i2}^l, \ldots, C_{ia}^l$ by

$$I_i^{l+s} = \sum_{j=1}^{a} \frac{F_{ij}^l}{b_{ij}^l}$$

We define the final probability of C_i^{l+s} by

$$F_i^{l+s} = (1-\beta) \times I_i^{l+s} + \beta \times \frac{e^{E_i^{l+s}/kT}}{Q^{l+s}}$$

where E_i^{l+s} is the number of contacts of C_i^{l+s} and

$$Q^{l+s} = \sum_{h=0}^{cl+s} g_{l+s}(h) e^{-h\varepsilon/kT}$$

Table 1. The probability of the nine folds obtained for HPHPPHPPHH

Configuration	Energy	Prob. T=0.2, β=0.75	Prob. T=0.2, β=0.25	Prob. T=0.8, β=0.25
A	−3	0.058	0.274	0.2
B	−3	0.048	0.196	0.152
C	−4	0.737	0.281	0.139
D	−3	0.03	0.05	0.093
E	−3	0.03	0.048	0.089
F	−3	0.03	0.048	0.089
G	−3	0.03	0.048	0.087
H	−3	0.03	0.048	0.087
I	−3	0.005	0.008	0.06

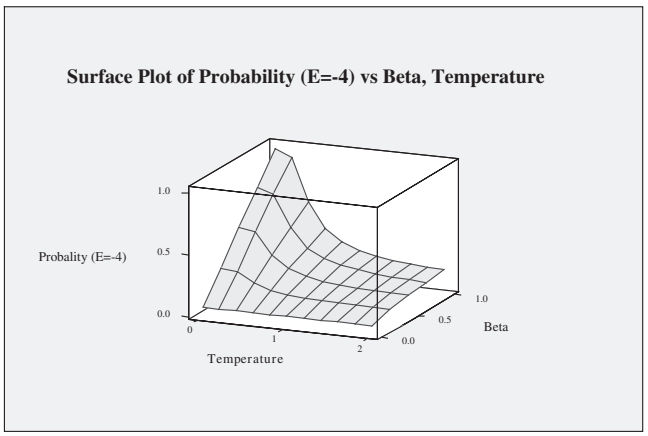

Fig. 10. A graphic showing the probability that the 10-mer HPHPPHPPHH is in the lowest state of energy (−4) as temperature and thermodynamic permission factor change. We observe that the probability decreases as the temperature rises and as the thermodynamic permission factor β drops. When we increase the temperature we allow more energy for unfolding, favouring states which have a higher energy than the ground state. As β decreases to zero, we allow the distribution at each elongation stage less freedom to settle to the Boltzmann distribution, favouring higher energy states. Note that a β of zero results in a model which is independent of temperature, whence the non-zero probability of a final conformation is solely determined by the initial probabilities at each stage.

where Q^{l+s} is the partition function and $g_{l+s}(h)$ is the density of states, which is the number of all sequential conformations of length $l+s$ with h contacts, c_{l+s} is the maximum number of contacts among all conformations of length $l+s$, T is the temperature and k is the Boltzmann constant.

Application of the probabilistic model We apply the probabilistic model to the 10-mer HPHPPHPPHH. We study the impact of β and the temperature T on the distribution of conformations at each step of the elongation process, using $d=1$ and $s=1$. Figure 9 (A-I) shows the nine final conformations obtained; Table 1 shows the final probabilities of these nine conformations. We see that the probability of being in the lowest state of energy (conformation C) decreases as we raise T and lower β. With $T=0.8$ and $\beta=0.25$

Table 2. Summary of biophysical principles modelled

Biophysical mechanisms	Examples	Corresponding parameters in computational experiments	Results in computational experiments	Qualitative prediction to be tested on computational experiments
Cotranslational folding occurs	Semliki Forest virus capsid protein becomes biologically active before the full length polyprotein is produced (Baldwin, 1999)	*s* symbolizes the number of residue(s) added each time the chain is elongated	Models become more compact and less numerous as *s* increases	Results are kinetically controlled folds. Evidence of real protein models to be in kinetic traps is expected (if simulated sequentially under kinetic control)
Folding is under kinetic control	Mouse prion protein native conformation is not the most thermodynamically stable conformation (Baskakov *et al*., 2001)	*d* symbolizes the finite surmountable energy barrier β symbolizes the thermodynamic permission factor which releases the Boltzmann energy distribution	Models become more compact and less numerous as *d* increases Models with highest probability of occurrence are not always the ones of lowest energy	

we have conformations A and B more likely to occur than the lowest energy conformation C. Figure 10 shows the probability that the 10-mer HPHPPHPPHH is in the lowest energy state as T and β vary.

Consequences of cotranslational folding of real proteins Should cotranslational folding prove to be the norm, then we can make predictions about the effect on protein structure:

(i) The N-terminus may be more likely to be buried; the C-terminus, being ''held'' by the ribosome, may be more likely to be peripheral in the final structure.

(ii) Protein structure may favour local contacts.

(iii) The active state of a protein may not be the lowest energy state.

(iv) Designed sequences may often fail to produce the desired structure because cotranslational folding is not taken into account. Therefore designing artificial proteins with local interactions vectorised from the N- to the C- terminus may be advantageous.

(v) New folds of lower energy may be found if we relax kinetic control, increasing the surmountable energy barrier.

CONCLUSION

We have modelled the folding of proteins cotranslationally and under kinetic control, with the help of simple lattice models. We selected intermediate conformations, within the surmountable energy barrier, as the polypeptide chain elongated. We saw that the globally minimum energy, that with the maximum number of contacts, was not always accessible with a low surmountable energy barrier. As we increased this barrier, we obtained final sequential conformations which were more compact and less numerous. A sufficiently high barrier enabled us to reach a final conformation which had the maximum number of contacts.

We attached a probability to each of the intermediate and final folds obtained. We introduced a thermodynamic permission factor, capturing the property that intermediate and final conformations under constraints may not always reach the Boltzmann equilibrium. We found that folds with lowest energy were not always the ones with highest probability. We summarized our results in Table 2.

The study is restricted to short, two-dimensional designing sequences. Modelling could be improved through use of longer sequences, folding three-dimensionally. The thermodynamic permission factor modelled various *in vivo* constraints on the folds, summarizing these constraints in a single parameter. Future developments could include use of a length-dependent thermodynamic permission factor. Finally, we know that the ribosome imposes spatial restrictions on the fold; these should also be taken into account.

REFERENCES

Andersson,S.G.E. and Kurland,C.G. (1990) Codon preferences in free-living microorganisms. *Microbiological Reviews*, **54**, 198–210.

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

Baker,D. (1998) Metastable states and folding free energy barriers. *Nature Structural Biology*, **5**, 1021–1024.

Baldwin,T.O. (1999) Protein folding in vivo: the importance of ribosomes. *Nature Cell Biology*, **1**, 154–155.

Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, **289**, 905–920.

Basharov,M.A. (2003) Protein folding. *Journal of Cellular and Molecular Medicine*, **7**, 223–237.

Baskakov,I.V., Legname,G., Prusiner,S.B. and Cohen,F.E. (2001) Folding of prion protein to its native alpha-helical conformation is under kinetic control. *Journal of Biological Chemistry*, **276**, 19687–19690.

Bornberg-Bauer,E. (1997) Chain growth algorithms for HP-type lattice proteins. *RECOMB 97*, 47–55.

Braakman,I., Hoover-Litty,H., Wagner,K.R. and Helenius,A. (1991) Folding of influenza hemagglutinin in the endoplasmic reticulum. *Journal of Cell Biology*, **114**, 401–411.

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

Chan,H.S. and Dill,K.A. (1993) Energy landscapes and the collapse dynamics of homopolymers. *Journal of Chemical Physics*, **99**, 2116–2127.

Chan,H.S. and Dill,K.A. (1994) Transition states and folding dynamics of proteins and heteropolymers. *Journal of Chemical Physics*, **100**, 9238–9257.

Curran,J.F. and Yarus,M. (1989) Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *Journal of Molecular Biology*, **209**, 65–77.

Dill,K.A., Bromberg,S., Yue,K., 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**, 561–602.

Fedorov,A.N. and Baldwin,T.O. (1997) Cotranslational protein folding. *Journal of Biological Chemistry*, **272**, 32715–32718.

Fedorov,A.N. and Baldwin,T.O. (1997) GroE modulates kinetic partitioning of folding intermediates between alternative states to maximize the yield of biologically active protein. *Journal of Molecular Biology*, **268**, 712–723.

Fernandez,A. (1994) Ascribing weights to folding histories: explaining the expediency of biopolymer folding. *Journal of Physics A (Mathematical and General)*, **27**, 6039–6052.

Frydman,J. (2001) Folding of newly translated proteins in vivo: the role of molecular chaperones. *Annual Review of Biochemistry*, **70**, 603–647.

Govindarajan,S. and Goldstein,R.A. (1998) On the thermodynamic hypothesis of protein folding. *Proceedings of the National Academy of Sciences*, **95**, 5545–5549.

Guo,Z., Brooks,C.L. and Boczko,E.M. (1997) Exploring the folding free energy surface of a three-helix bundle protein. *Proceedings of the National Academy of Sciences*, **94**, 10161–10166.

Hartl,F.U. and Hayer-Hartl,M. (2002) Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, **295**, 1852–1858.

Irbäck,A. and Troein,C. (2002) Enumerating designing sequences in the HP model. *Journal of Biological Physics*, **28**, 1–15.

Jenni,S. and Bany,N. (2003) The chemistry of protein synthesis and voyage through the ribosomal tunnel. *Current Opinion in Structural Biology*, **13**, 212–219.

Keller,S.L. (2003) Sequential folding of a rigid wire into three-dimensional structures. *American Journal of Physics*, **72**, 599–604.

Kolb,V.A. (2001) Cotranslational protein folding. *Molecular Biology*, **35**, 584–590.

Kolb,V.A., Makeyev,E.V. and Spirin,A.S. (2000) Co-translational folding of an eukaryotic multidomain protein in a prokaryotic translation system. *Journal of Biological Chemistry*, **275**, 16597–16601.

Komar,A.A. and Jaenicke,R. (1995) Kinetics of translation of gamma B crystallin and its circularly permutated variant in an in vitro cell-free system: possible relations to codon distribution and protein folding. *FEBS Letters*, **376**, 195–198.

Komar,A.A., Lesnik,T. and Reiss,C. (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Letters*, **462**, 387–391.

Levinthal,C. (1968) Are there pathways for protein folding, *Journal of Chemical Physics*, **65**, 44–45.

Levinthal,C. (1969) Mossbauer spectroscopy in biological systems. *University of Illinois Press, Urbana*, 22–24.

Morrissey,M.P., Ahmed,Z. and Shakhnovich,E.I. (2004) The role of cotranslation in protein folding: a lattice model study. *Polymer*, **45**, 557–571.

Nakatogawa,H. and Ito,K. (2002) The ribosomal exit tunnel functions as a discriminating gate. *Cell*, **106**, 629–636.

Netzer,W.J. and Hartl,F.U. (1997) Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature*, **388**, 343–349.

Nicola,A.V., Chen,W. and Helenius,A. (1999) Co-translational folding of an alphavirus capsid protein in the cytosol of living cells. *Nature Cell Biology*, **1**, 341–345.

Pande,V.S., Grosbergy,A.Y. and Tanaka,T. (1997) Statistical mechanics of simple models of protein folding and design. *Biophysical Journal*, **73**, 3192–3210.

Purvis,I.J., Bettany,A.J.E., Santiago,T.C., Coggins,J.R., Duncan,K., Eason,R. and Brown,A.J.P. (1987) The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis, *Journal of Molecular Biology*, **193**, 413–417.

Ramakrishnan,V. (2002) Ribosome structure and the mechanism of translation. *Cell*, **108**, 557–572.

Rohl,C.A., Strauss,C.E.M., Misura,K.M.S. and Baker,D. (2004) Protein structure prediction using Rosetta. *Methods Enzymol*, **383**, 66–93.

Sanchez,I.E., Morillas,M., Zobeley,E., Kiefhaber,T. and Glockshuber,R. (2004) Fast folding of the two-domain semliki forest virus capsid protein explains co-translational proteolytic activity. *Journal of Molecular Biology*, **338**, 159–167.

Shakhnovich,E.I. (1998) Protein design: A perspective from simple tractable models. *ArXiv Condensed Matter e-prints/9804199 (web publication)*.

Sikorski,A. and Skolnick,J. (1990) Dynamic monte carlo simulations of globular protein folding. *Journal of Molecular Biology*, **215**, 183–198.

Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, **268**, 209–225.

Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B.A., Bystroff,C. and Baker,D. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, **34**, 82–95.

Sohl,J.L., Jaswal,S.S. and Agard,D.A. (1998) Unfolded conformations of alpha-lytic protease are more stable than its native state. *Nature*, **392**, 817–819.

Wilson,D.N., Blaha,G., Connell,S.R., Ivanov,P.V., Jenke,H., Stelzl,U., Teraoka,Y. and Nierhaus,K.H. (2002) Protein synthesis at atomic resolution: mechanistics of translation in the light of highly resolved structures for the ribosome. *Current Protein and Peptide Science*, **3**, 1–53.

Ziv,G., Haran,G. and Thirumalai,D. (2005) Ribosome exit tunnel can entropically stabilize alpha-helices. *Proceedings of the National Academy of Sciences*, **102**, 18956–18961.