Signatures of Co-translational Folding

Rhodri Saunders [1,*], Martin Mann [2,*] and Charlotte M. Deane [1]

[1] Oxford University, Department of Statistics, 1 South Parks Road, Oxford, OX1 3TG, UK

[2] University of Freiburg, Bioinformatics, Georges-Köhler-Allee 106, 79110 Freiburg, Germany

[*] These authors contributed equally to this work.

Keywords: Co-translational folding, Protein structure, HP model, SCOP domains, Vectorial folding

Correspondence: Charlotte Deane, Oxford University, Department of Statistics, 1 South Parks Road, Oxford, OX1 3TG, UK

Email: deane@stats.ox.ac.uk, Fax: +44 1865 272595, Phone: +44 1865 281301

Abbreviations: CT - Co-Translational; HP - Hydrophobic-Polar; *MCR* - Mean Central Residue; *MoI* - Moment of Inertia; NC_{cen} - relative N- and C-terminal centrality; *PHI* - Possible Hydrophobic Interactions; SSE - Secondary Structure Element; UGEM - Unique Global Energy Minimum

Abstract

Global and co-translational protein folding may both occur *in vivo*, and understanding the relationship between these folding mechanisms is pivotal to our understanding of protein structure formation.

Within this study, over 1.5 million Hydrophobic-Polar sequences were classified as either global or co-translational folders based on their ability to attain a unique minimum energy conformation via co-translational folding. The sequence and structure properties of the sets were then compared to elucidate signatures of co-translational folding.

The strongest signature of co-translational folding is a reduced number of possible favourable contacts in the amino-terminus. There is no evidence of fewer contacts, more local contacts, nor less compact structures. Co-translational folding does produce a more compact amino- than carboxy-terminal region and an amino-terminal biased set of core residues. In real proteins these signatures are also observed and found most strongly in proteins of the SCOP alpha/beta class where 71% have an amino-terminal set of core residues.

The prominence of co-translational features in experimentally determined protein structures suggests that the importance of co-translational folding is currently underestimated.

Supplementary information available at
http://www.stats.ox.ac.uk/proteins/resources.

1 Introduction

Whether protein structure formation occurs concurrently with translation or after release from the ribosome is currently widely debated, see reviews [1, 2]. There is experimental evidence for both folding mechanisms [3–7]. If both occur *in vivo* then understanding the features, interplay and relative abundance of these two folding mechanisms may play a key role in furthering our knowledge of protein structure. Here, we elucidate sequence and structure signatures that are related to the folding mechanism.

A protein has a huge number of possible structures but, in accordance with the Levinthal paradox, it cannot explore every possible conformation in finding its biological structure [8]. It has been suggested that a directed pathway through structure space is used to attain the biological structure [9]. Under *co-translational (CT) folding*, the vectorial emergence of proteins from the ribosome could form the basis for such directed folding pathways.

Experiments have demonstrated that ribosomes can catalyse the folding process [1, 10, 11] and simulations have suggested possible mechanisms for this effect. A simulation on the diamond lattice by Sikorski and Skolnick [12] found that the ribosome accelerated the folding process by preventing formation of off-pathway intermediates. A model by Contreras Martinez et al. [13] found that the ribosome's exit tunnel can facilitate folding of well-designed proteins. Chikenji et al. [14] transformed the normally rugged energy landscape of simplified models into a smooth funnel by preventing certain sub-structures in analogy to the exit tunnel. Recently Jefferys et al. [15] used a CT protein folding algorithm to study the importance of macromolecular crowding on protein folding while Ellis et al. [16] proved that the translation direction has a strong impact on protein folding.

CT folding occurs *in vivo*, however the sequence signatures that drive it and

the resultant structure signatures that define it are both unknown. To identify the signatures of CT folding and to estimate its abundance requires a set of sequences known to fold co-translationally. Unfortunately this set is not available for solved protein structures and so we use the Hydrophobic-Polar (HP) model [17]. Within the HP model we can create a sequence set that folds co-translationally and, for comparison, a sequence set that does not.

Generally, only HP sequences with a unique global energy minimum (UGEM) conformation are considered protein-like. Previously, UGEM sequences/conformations have been shown to exhibit protein-like: hydrophobicity [18, 19]; surface to core ratios [20, 21]; repeating motifs [22–24]; volume exclusion among residues [25]; and hydrophobic cores and polar exteriors [19]. The HP model is particularly useful in studies that require extended coverage of both sequence and conformation space [24]. However, it is essential to partition valid conclusions from system artifacts. Crucially, the focus here is not on the mechanics and energetics of protein folding, but rather on the sequence signatures that direct a protein towards CT folding and the structural signatures that subsequently result.

In the HP model both sequence and conformation space can be fully enumerated [25]. Taking all sequences with a UGEM conformation we produce three sequence sets:

1. *Global-CT* - a sequence reaches its UGEM conformation co-translationally.
2. *Kinetic-CT* - a final unique conformation is formed co-translationally but it is not the UGEM conformation.
3. *Non-CT* - no unique final CT conformation is attained.

The Kinetic-CT set is included because previous research has shown that the UGEM conformation may not always be the most highly populated state under

CT folding [26] whereby CT folding can produce stable but thermodynamically non-optimal structures [27].

Previous studies that explored CT folding were based on hypothetical expectation and/or small data sets. We undertake the first exhaustive sequence classification based on folding ability. Our classification should enable any sequence and structure signatures of CT folding to be identified. Over 1.5 million UGEM sequences are classified; with $17,085$ sequences of length 25 classified as Global-CT. A Markov-chain simulation demonstrates that the three sequence sets differ in their folding behaviour. CT folding increases the likelihood of finding the UGEM conformation by over 20%.

Comparison of our sequence sets demonstrates that many presumed signatures of CT folding are too simplistic. For example, Global-CT conformations are not enriched in local nor previous contacts as postulated by Alexandrov [28] & Deane et al. [29]. They have a more: compact hydrophobic core, amino-terminal core, and centrally orientated amino-terminus. The strongest signature of Global-CT folding is a restriction of the conformational space available to the amino-terminal region: a restriction that has been demonstrated in real protein structures [30].

Analysing the SCOP database a general trend towards CT folding is observed. This is seen most strongly in the α/β class with 66% having a more centrally orientated amino- than carboxy-terminus and 71% having a more amino-terminal set of core residues.

2 Material and Methods

Initially we introduce our sequence classification algorithm (`M1`), a deterministic method for high-throughput studies. The Markov-chain simulation schemes for global folding (`M2`) and CT folding (`M3`) follow. Finally, the data sets and mea-

sures are described. Our algorithms are model independent and here applied to the Hydrophobic-Polar model [17], where each amino acid is represented by a single monomer classed as either Hydrophobic or Polar. A conformation is any self-avoiding walk within the chosen lattice. Throughout this study protein sequence P has conformation space S, i.e. the set of all possible conformations. Under global folding all of S is available, whereas CT folding explores only a subset of S. The function $E(s)$ evaluates the energy involved in P forming $s \in S$. The protein length is assigned n.

All our algorithms are freely available in `LatPack` [31] implementewd here as v1.8.1. Programs as follows, `M1`:`latVec`, `M2`:`latFold`, `M3`:`latFoldVec`.

`M1` - Sequence Classification based on CT folding

Our three sequence sets are created using the following classification scheme that evaluates folding via low-energy pathways available within the complete energy landscape. CT folding is modeled by a chain-growth procedure [32], which we apply in accordance with Huard et al. [26]. For each elongation event $l \in [1, n]$, all prior conformations are extended to generate S_l, a set of conformations of length l. Only conformations within ΔE of the current minimum energy are accepted; thus ΔE is the energy in the system available for refolding [26]. At full elongation length n, all conformations reached via CT folding S_n are evaluated to classify the sequence as either Global-CT, Kinetic-CT, or Non-CT.

More precisely, we start with one monomer, i.e. conformation set S_1. For each elongation event l ($2 \leq l \leq n$), all conformations from the last elongation $s \in S_{l-1}$ are extended to produce all possible elongated conformations S'_l of length l. The minimal energy E_m of all $s \in S'_l$ is calculated by $E_m(S'_l) = \min(E(s) \mid s \in S'_l)$. Only conformations $S_l \subseteq S'_l$ within ΔE of E_m are retained for extension at the next elongation event, such that $S_l = \{s \mid s \in S'_l \wedge E(s) \leq (E_m(S'_l) + \Delta E))\}$.

The final set S_n is restricted to the minimal energy conformations reachable, i.e. $S_n = \{s \mid s \in S'_n \wedge E(s) = E_m(S'_n)\}$. Given that P has a UGEM conformation of energy $E(\text{UGEM})$ we can classify P according to S_n:

- Global-CT if $(E(\text{UGEM}) = E(s \in S_n))$; i.e. $|S_n| = 1$,
- Kinetic-CT if $(E(\text{UGEM}) \neq E(s \in S_n) \wedge |S_n| = 1)$,
- else Non-CT.

Further details are provided in supplementary information. Our classification scheme does not examine folding kinetics or thermodynamics; for this we use Markov-chain simulation protocols.

M2 - Global Folding Simulations

Global folding (M2) from a full-length conformation $s_0 \in S$ is simulated using a standard Metropolis 1st-order Markov-chain approach [33]. This produces a time series of conformations $(s_0 \rightarrow s_1 \rightarrow \ldots \rightarrow s_{t_{\max}})$ after $t_{\max}$ simulation steps. Possible transitions $s_{\text{from}} \rightarrow s_{\text{to}}$ are defined by the structural neighbourhood $M(s_i) \subseteq S$ such that $s_{\text{to}} \in M(s_{\text{from}})$, where M has to be symmetric and ergodic. Here, the semi-local *pull-moves* introduced by Lesh et al. [34] are used to define M because they have been demonstrated to simulate folding on a relevant timescale [35]. The probability of accepting a transition $s_{\text{from}} \rightarrow s_{\text{to}}$ is given by the Metropolis criterion $min(1,\ \exp(-[E(s_{\text{to}}) - E(s_{\text{from}})]/kT))$. Symmetric conformations due to rotation or reflection are analysed only once.

M3 - Markov-chain CT Folding Simulations

Translation speed affects protein folding [36] and we hypothesise that folding during elongation can restrict the final conformation set produced. A *CT folding* simu-

lation should incorporate such *interim folding* during chain elongation. `M3` is used to investigate this principle and we term the resulting conformation a *CT-fold*. To model interim folding, a Markov-chain simulation is used after each elongation event. This provides an energy-biased search of local conformation space during elongation. This type of simulation has been informative [37–39] but, unlike our classification method `M1`, cannot easily be applied to high-throughput screens.

To simulate the proposed *CT folding process (`M3`)* an iterative scheme of $l \in [1, n]$ elongation events (`M3.1`) is applied, each followed by interim folding (`M3.2`). The phases in detail are:

`M3.1`: Given the final conformation of the last iteration s_{l-1} of length $(l-1)$, we produce all its possible elongations S_l of length l. From this conformation ensemble we pick a random elongation $s \in S_l$ according to its Boltzmann probability within the ensemble given by

$$Pr(s \in S_l) = \frac{\exp(-E(s)/kT)}{\sum_{s' \in S_l} \exp(-E(s')/kT)} \; . \tag{1}$$

`M3.2`: The chosen elongation s (from `M3.1`) defines the starting conformation for interim folding. Interim folding is simulated via a Markov-chain protocol similar to `M2`; with the local conformation space M defined by *pull-moves* whereby each residue can explore M through a length dependent number of folding events $(c \cdot l)$. The conformation resulting from interim folding s_l either starts another iteration `M3.1` $(l < n)$ or is a *CT-fold* of the sequence.

The length-dependent folding time $(c \cdot l)$ incorporates the locality of single *pull-moves*. Thus, c is the average number of folding events per residue. It is utilised because the simulation applies a series of local structural changes that, in biology, could occur simultaneously.

HP-Model and Folding Simulation Parameters

The HP model represents each amino acid as a single monomer classified as either Hydrophobic (`H`) or Polar (`P`). In the energy function each non-consecutive HH-contact contributes -1, all other interactions 0. Viable conformations are self-avoiding walks within a given lattice. Many previous protein studies have utilised this model and identified biologically relevant characteristics [18, 21, 40, 41].

Based on previous work [31, 42] we use $kT = 0.3$ as the relative folding temperature within the Metropolis criterion and in Eq. 1. Other values of kT (0.1 to 0.5) produced similar results.

Sequence and Structure Sets

The "designing sequences" of Irback and Sandelin [18] are used in the 2D-square lattice. "Designing sequences" are every HP sequence, $l \leq 25$, with a UGEM conformation. Our classification scheme `M1` grouped them into three folding sets: Global-CT, Kinetic-CT, and Non-CT. Additionally, all remaining HP sequences ($10 \leq n \leq 16$) were tested to see if they are Kinetic-CT sequences; that is whether they have a single conformation after $CT\ folding$ (`M1`). All our 2D studies used a surmountable energy barrier of zero ($\Delta E = 0$) that prevented any chance of global folding.

In the 3D-cubic lattice, two non-exhaustive sets of $\sim 10,000/17,000$ random sequences of length 27/36 with a unique global fold were derived using the CPSP approach of Backofen and Will [43] as implemented in `HPoptdeg` from the `CPSP-tools` package v2.4.2 [44]. In this case, a surmountable energy barrier of 1 ($\Delta E = 1$) was used.

Care must be taken when transfering measures designed on the HP model conformations to real protein structures. Proteins often have flexible, essentially unstructured, termini - the requirement for a single conformation in the HP model

occludes these regions from our analysis. Hence, we analyse protein structures from the start of the most N-terminal secondary structure element (SSE) to the end of the most C-terminal SSE. JOY [45] is used to assign secondary structure, only helix and strand (when in a run of three or more) are considered to be SSEs.

A set of $10,311$ domains from SCOP [46] (release 1.75) with a sequence identity cut-off of 40% was analysed. When comparing between SCOP classes only the 1969 α, 2174 β, 2652 α/β, and 2640 $\alpha+\beta$ domains were considered; other classes had less than 1000 occurences.

Sequence and structure measures

Measures of sequence and structure properties are designed or adapted from the literature [26, 29]. The most informative are detailed here, a full set of tests undertaken can be found in the supplementary material. Throughout, R_i denotes the i-th residue of a sequence of length n. $\delta(R_i, R_j)$ denotes the structural distance between R_i and R_j and $h(R_i) = 1$ if R_i is hydrophobic and $= 0$ otherwise. In solved protein structures, R_1 is the most N-terminal and R_n the most C-terminal residue assigned to a run of at least three helix residues or three strand residues. The sequence record of PDB files is sometimes incomplete if not all residues are resolved in the X-ray structure - in these cases n is based on the actual residue number to incorporate chain breaks.

Sequence measures

- *Hydrophobicity* measures the percentage of residues in the sequence classified as hydrophobic ($= \frac{1}{n}\sum_i h(R_i)$). Hydrophobic residues are Ala, Cys, Ile, Leu, Met, Phe, Pro, Trp and Val [19]. *Hydrophobicity by quartiles (HBQ)*, examines the Hy-

drophobicity of the intervals $[1, \frac{n}{4}], [\frac{n}{4}, \frac{n}{2}], [\frac{n}{2}, \frac{3n}{4}]$, and $[\frac{3n}{4}, n]$.

- *Possible Hydrophobic Interactions* (*PHI*, Eq. 2) describes the relative possibility of each hydrophobic residue to make favourable (hydrophobic) interactions.

$$PHI(i) = \frac{\sum_{1 \leq j \leq n}^{|i-j|>d} I(R_i, R_j)}{n - 2d - 1} \tag{2}$$

where d defines the minimal distance in sequence considered for contacts and is set to 3 throughout. To evaluate interactions we use $I(R_i, R_j) = -1$ if both residues are hydrophobic ($h(R_i) = h(R_j) = 1$); otherwise $I(R_i, R_j) = +1$. Since rectangular lattices suffer the parity problem[1] [44], we halve the normalisation of the *PHI* score in lattice models using $(n - 2d - 1)/2$. For analysis, the *PHI* score is averaged for each position over all sequences per sequence set (Non-CT, Global-CT, Kinetic-CT).

- *Neutral nets* are a network where nodes represent sequences that share a common UGEM conformation and edges connect sequences that differ by a single point mutation. The most connected node, the *hub-node*, has been described as the most protein-like sequence because it is more robust to mutation [24]. Neutral nets were built for the 2D HP sequences of $n = 25$. Only networks for structures that are the UGEM conformation for ≥ 50 sequences (including at least one Global-CT sequence) were considered. It is possible that the sequence space of a UGEM conformation contains more than one distinct neutral net [24]; in which case each net is analysed separately.

[1] Neighbored nodes in rectangular lattices show different parity in coordinate sum resulting in two classes of nodes, i.e. with even or odd coordinate sum. Due to the connectivity along the structure, only monomers with different sequence index parity can form contacts, i.e. even to odd and vice versa.

Structure measures

- The *Mean Central Residue (MCR, Eq. 3)* calculates the sequence position closest to the protein's centre of mass (M) (similar to *pMIN* from [29]). Equation 3 utilises $core(k)$ to access the index of the k-th closest residue R according to $\delta(R, M)$ (i.e. $\delta(R_{core(k)}, M) \leq \delta(R_{core(k+1)}, M)$). Through normalisation MCR maps to the interval $[0, 1]$; thus a more N-terminal set of η core residues has a score less than 0.5. We used $\eta = 4$ and $= 8$ in 2D and 3D respectively.

$$MCR = \frac{\sum_{i=1}^{\eta} core(i) \cdot W(i)}{n \cdot \sum_{i=1}^{\eta} W(i)} \quad \text{with } W(i) = \frac{1}{\delta(R_{core(i)}, M)} \tag{3}$$

- NC_{cen} (Eq. 4), the *relative distance of the N and C termini to the protein's center* M, assesses terminal bias of M from a structural rather than sequence perspective. A negative NC_{cen} reveals that the N-terminus R_1 is more centrally orientated then the C-terminus R_n.

$$NC_{cen} = log \frac{\delta(R_1, M)}{\delta(R_n, M)} \tag{4}$$

- The *moment of inertia (MoI, Eq. 5)* measures structural compactness as the average distance of any residue to M [26]. The *hydrophobic MoI* can be calculated by only considering hydrophobic residues; in this case n is the number of hydrophobic residues and M is their average position.

$$MoI = \frac{1}{n} \sum_{i} [\delta(R_i, M)]^2 \tag{5}$$

3 Results and discussion

Our methods and the resulting sequence sets are first validated with comparison to the literature. Subsequently, the sequence and structure properties of our three sequence sets are investigated to identify signatures of CT folding. Finally, we move away from simplified models and identify a subset of these CT folding signatures within a large, non-redundant set of real protein domains.

Co-translational folding via Markov-chain

In general, we envision that CT folding produces a limited set of structures (CT folds) that provide a beneficial start point for reaching the biological conformation. One CT fold may, of course, be the biological conformation. In this study CT folds are produced using the Markov-chain method `M3`.

To validate our method of producing CT folds (`M3`) we investigate the 10-mer sequence `HPHPPHPPHH` that has a UGEM conformation and was shown by Huard et al. [26] to benefit from CT folding. A full enumeration of all possible CT folding paths demonstrated that this sequence folded more efficiently co-translationally than globally. The sequence also benefits from CT folding under our folding procedure `M3`.

Different intermediate folding times (c constant) between elongation events are tested; with four test sets of CT-folds produced ($c = 0, 1, 2, 3$). The success rate of attaining the UGEM conformation from CT-folds is compared to that of attaining the UGEM conformation from a set of random start conformations. In each case, $k = 10^5$ global folding simulations (`M2`) of 100, 200 and 500 simulation steps are undertaken. The protein's folding rate (r_f) is calculated from the number of successful (reaching the UGEM conformation) folding simulations h_{succ} in the k simulations: $r_f = (h_{succ}/k)$.

Figure 1 summarises the results. Our model produces the folding behaviour for HPHPPHPPHH as predicted by full-enumeration [26]. Starting from CT-folds is always superior to random start conformations for this sequence. As the number of folding events per elongation is increased between 0 and 2, the folding rate also increases. Above two intermediate folding events no improvement in the folding rate is observed for this sequence. This may result from the 1st-order Markov-chain implemented because as time increases the protein chain becomes increasingly independent of the starting conformation. These results demonstrate that the model can reproduce full enumeration results and that model proteins can benefit from CT folding.

Furthermore, using M3 we can test whether our folding sets, generated in the following by M1, have different folding properties.

Identification of Co-translational Folding

Using our classification method M1, we have classified over 1.5 million HP sequences into three sets: Global-CT, Kinetic-CT, and Non-CT folders (see Methods). The $765,147$ sequences of length 25 with a UGEM conformation in the 2D-square lattice are separated into $17,085$ Global-CT, $74,502$ Kinetic-CT, and $673,560$ Non-CT sequences. In general, between 10 % and 20 % of our longer sequence sets seem to fold co-translationally. The percentage is higher for shorter sequences (43% at length 13). It suggests that a significant number of proteins may use co-translation to attain their biological conformation. M1 is a very restrictive implementation of CT folding and many more sequences may use CT-folds as a springboard for finding their native conformation.

Our classification procedure (M1) has no kinetic component. It is a purely deterministic measure of whether a sequence can benefit from CT folding under the most restrictive definition of CT folding. To evaluate our classification procedure

`M1` we investigate the folding rates of a random subset of each group (Global-CT, Kinetic-CT, Non-CT) using the Markov-chain CT folding simulation (`M3`) followed by global folding (`M2`). It is expected that Global-CT sequences will exhibit the highest success rate when global folding is initiated from CT folds. Furthermore, Kinetic-CT sequences should have a preference for reaching their kinetic fold over their UGEM conformation. Figure 2 shows the ratio of the folding rates when starting from CT-folds compared to starting from random conformations. Folding rates are averaged over 50 randomly selected sequences per group and 10^4 folding simulations of 200 global folding steps per sequence. As expected, Global-CT sequences benefit the most from CT folding and show a higher folding rate increase than Non-CT sequences. In reaching their UGEM conformation, Kinetic-CT sequences show a similar folding rate increase to Non-CT sequences *but* show a significantly higher rate to adopt their proposed kinetic fold. For all classifications, starting from the CT-folds increases the folding rate and we suggest that CT folding is an efficient way to explore fold space. Increasing the number of intermediate folding events per elongation is beneficial up to a particular point. The optimal number of intermediate folding events increases with sequence length (see supplementary material) but we elucidate no general rule underlying it. The effect probably results from a longer energy driven exploration of CT structure space.

Overall, our extensive sequence classification is supported by the differing folding properties of our three sets under a Markov-chain simulation.

Model Features of Co-translational Folding

A large number of sequences fold co-translationally to a unique final conformation. As described in the HP-model, the percentage of sequences over length 20 that are Global-CT folders is low - their actual number, however, is large. The low percentage

but large number of CT-folders is not counter to biology: protein sequence space is vast, yet only a very small percentage of sequences are observed in nature [47–49].

Global-CT sequences behave differently to the other sets when our measures of sequence space are considered. In our examination of neutral nets we find that Global-CT sequences are often the most robust to mutation. Hub nodes are enriched in Global-CT sequences; at length 25, Global-CT sequences make up just 12.8% of neutral net sequence space and account for 41.9% of hub sequences.

Structural compactness (*MoI*) did not segregate our sequence sets. However, examining the *hydrophobic MoI* did. In contrast to the theorised properties, Global-CT sequences had, on average, the most compact hydrophobic cores. Kinetic-CT sequences, in accordance with the theory, had the least compact cores; but the distributions do overlap. It would appear that by selecting UGEM conformations that can be found co-translationally we have also selected those conformations with the most compact hydrophobic cores. It may be that a compact core is related to the robustness to mutation exhibited by Global-CT sequences.

Undertaking a closer examination of core properties through the mean central residue (*MCR*) and the relative centrality of the N- and C-termini (NC_{cen}) we found, as expected, that Non-CT sequences show no overall bias. Our other sets do: 73% of Kinetic-CT and 75% of Global-CT conformations have a more centrally orientated N- than C-terminus at length 25. When considering the MCR, we found that 97% of Kinetic-CT and 93% of Global-CT sequences have an N-terminal core (length 25).

It was theorised that CT folding would produce a more compact N- than C-terminal region. We calculate the *MoI* of the extreme eight terminal residues and then compute $\log \frac{MoI(N)}{MoI(C)}$ where a negative result indicates a more compact N- than C-terminal region. Overall a negative score is observed for both Kinetic-CT and Global-CT sets. Non-CT folders have, on average, equally compact N- and C-

termini.

Global-CT sequences also have a significantly lower percentage hydrophobicity (χ^2 test) and exhibit a general decrease in hydrophobicity moving from the N- to C-terminus (see supplementary material).

Of all our tested measures, the possible hydrophobic interactions (*PHI*) score most clearly segregates Global-CT sequences from our other sequence sets. A positive *PHI* score for a position means that there are more unfavourable than favourable interactions possible at that position. Through our *PHI* score we demonstrate that Global-CT sequences are characterised by a low number of possible favourable (hydrophobic) interactions in N-terminal regions (positive *PHI*, see Figure 3). We suggest that specific contacts form in the N-terminal region that restrict and guide the subsequent folding process. The restriction on conformation space is unique to Global-CT sequences. *PHI* is an absolute measure of CT folding potential, and in principle Global-CT sequences could be isolated from the whole of sequence space. To test this we used the *PHI* score and relative terminal hydrophobicity to select possible Global-CT sequences of length 30 and found an 8.5 fold enrichment in expected UGEM sequence identification (see supplementary material).

Co-Translational Folding Features in Real Proteins

Protein structures have a modular design composed of domains, where each domain is assumed to be able to fold independently. Therefore, in order to study the possible signatures of CT folding within real protein structures the SCOP domain database is used.

As mentioned, care must be taken when expanding measures from the HP model to real protein domains and, as such, unstructured termini are occluded from our analysis. Additionally, it is unknown to which of our folding sets (Global-CT,

Kinetic-CT, or Non-CT) each domain would belong. Hence, it is only possible to see if the data set as a whole is biased toward signatures of CT folding. We compare our measures between different SCOP classes as they have been suggested to vary in their propensity for CT folding [29, 50]. We have previously shown that, in contrast to Laio and Micheletti [51], the N-terminal region of a SCOP domain is, on average, more compact than the C-terminal region [30]. In this manuscript we link this observation to CT folding.

SCOP domains in general exhibit a bias towards the signatures we ascribe to CT folding: 56% having an $MCR < 0.5$ and 58% having an $NC_{cen} < 0$. As expected, the bias varies between SCOP domain classes: α ($[MCR < 0.5] = 51\%$, $[NC_{cen} < 0] = 50\%$), β (47%, 60%), α/β (71%, 66%), and $\alpha+\beta$ (54%, 57%) (Fig. 4). The α class alone shows no bias in these tests. All other sets have a significant bias towards CT folding under our NC_{cen} measure.

4 Concluding remarks

Undertaking an exhaustive model study of co-translational (CT) folding, we have classified over 1.5 million HP sequences in to three sequence sets (Global-CT, Kinetic-CT, and Non-CT) based on their CT folding properties. Global-CT sequences are optimised to find their unique global energy minimum (UGEM) conformation via a path of directed growth starting from their N-terminus (CT folding). Kinetic-CT sequences fold co-translationally to a unique final conformation but this conformation is not the unique global energy minimum. Non-CT sequences have a UGEM conformation but cannot attain this nor a unique final conformation co-translationally. A 1st-order Markov-chain simulation demonstrated the different folding behaviour of our sets and suggested that CT folding benefited all sequence sets in attaining their UGEM conformation.

Our three sequence sets allow the first test of the CT properties theorised in the literature [28, 29, 51]. In general, signatures of CT folding are more subtle than theorised. CT conformations are not generally enriched in local nor previous contacts, nor are they less compact. We do find that CT conformations have a more N-terminal core; a more centrally orientated N- than C-terminus; and a more compact N- than C-terminal region. The real protein structures tested are biased towards these signatures of CT folding. As Deane et al. [29], we find that SCOP classes differ in their propensity for CT folding properties and we highlight the α/β class as a strong candidate for CT folding.

Global-CT sequences dominate the hubs of large neutral nets and are thus, on average, more robust to mutation; a result consistent with that of Wang and Klimov [39]. Xia and Levitt [25] demonstrate that, through the evolution of folding rates and protein stability, there is a funnel-like organisation of sequence-space towards these hubs. Govindarajan and Goldstein [52] suggest that CT folding may be the standard and that sequences evolve such that the structure found co-translationally becomes the UGEM conformation. In reference to this we find that some Kinetic-CT conformations are UGEM conformations for other sequences.

The most significant signature of CT folding we identified is a sequence-mediated restriction in N-terminal structure space. The N-terminus of Global-CT sequences can make relatively few favourable contacts and we suggest that the formation of these favourable bonds directs the rest of the folding pathway towards the UGEM structure. In this way a directed folding path is created as suggested by Karplus [9]. There is experimental evidence that local sequence effects on structure are stronger at the N- than C-terminus. Native N-terminal structures have recently been observed in otherwise denatured protein [53]. The restriction in structure space and stronger local sequence signals at the N-terminus should make prediction of structure in

this region more accurate. Indeed increased prediction accuracy at the N-terminus was identified in secondary structure prediction by Holley and Karplus [54] and has recently been revisited by Saunders and Deane [30]. Unfortunately our *PHI* score is currently not directly applicable to real protein data. In a related test, the Miyazawa-Jernigan matrix [55] is used to assess contacts in real protein structures and indicates that there are generally fewer favourable contacts at the N-terminus (see supplementary material).

Overall, our results suggest that the abundance and importance of co-translational folding *in vivo* is currently underestimated.

Conflict of interest statement

The authors have declared no conflict of interest.

5 References

[1] G. Kramer, D. Boehringer, N. Ban, and B. Bukau. The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nat Struct Mol Biol*, 16:589–597, 2009.

[2] L.D. Cabrita, C.M. Dobson, and J. Christodoulou. Protein folding on the ribosome. *Curr Opin Struct Biol.*, 20(1):33–45, 2010.

[3] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci USA*, 47(9):1309–1314, 1961.

[4] I. E. Sanchez, M. Morillas, E. Zobeley, T. Kiefhaber, and R. Glockshuber. Fast folding of the two-domain semliki forest virus capsid protein explains co-translational proteolytic activity. *J Mol Biol*, 338(1):159–67, 2004.

[5] V. A. Kolb, E. V. Makeyev, and A. S. Spirin. Co-translational folding of an eukaryotic multidomain protein in a prokaryotic translation system. *J Biol Chem*, 275(22):16597–601, 2000.

[6] A. V. Nicola, W. Chen, and A. Helenius. Co-translational folding of an alphavirus capsid protein in the cytosol of living cells. *Nat Cell Biol*, 1(6):341–5, 1999.

[7] A. N. Fedorov and T. O. Baldwin. Cotranslational protein folding. *J Biol Chem*, 272(52):32715–8, 1997.

[8] Cyrus Levinthal. Are there pathways for protein folding? *J Chim. Phys.*, 65(1):44, 1968.

[9] M. Karplus. The Levinthal paradox: yesterday and today. *Folding & design*, 2(4), 1997.

[10] B. Hardesty and G. Kramer. Folding of a nascent peptide on the ribosome. *Progress in nucleic acid research and molecular biology*, 66:41–66, 2001.

[11] R. Saunders and C.M. Deane. Synonymous codon usage influences the local protein structure observed. *Nucl. Acids Res.*, 38(19):6719–6728, 2010.

[12] A. Sikorski and J. Skolnick. Dynamic monte carlo simulations of globular protein folding. model studies of in vivo assembly of four helix bundles and four member beta-barrels. *J Mol Biol*, 215(1):183–98, 1990.

[13] L. M. Contreras Martinez, Martinez F. J. Veracoechea, P. Pohkarel, A. D. Stroock, F. A. Escobedo, and M. P. Delisa. Protein translocation through a tunnel induces changes in folding kinetics: a lattice model study. *Biotechnol Bioeng*, 94(1):105–17, 2006.

[14] G. Chikenji, Y. Fujitsuka, and S. Takada. Shaping up the protein folding funnel by local interaction: lesson from a structure prediction study. *Proc Natl Acad Sci USA*, 103(9):3141–6, 2006.

[15] B.R. Jefferys, L.A. Kelley, and Michael J. E. Sternberg. Protein folding requires crowd control in a simulated cell. *Journal of Molecular Biology*, 397(5):1329–1338, 2010.

[16] J.J. Ellis, F.P. Huard, C.M. Deane, S. Srivastava, and G.R. Wood. Directionality in protein fold prediction. *BMC Bioinformatics*, 11(1):172+, 2010.

[17] Kit Fun Lau and Ken A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, 1989.

[18] A. Irback and E. Sandelin. On hydrophobicity correlations in protein chains. *Biophys J*, 79(5):2252–8, 2000.

[19] E. Sandelin. On hydrophobicity and conformational specificity in proteins. *Biophys J*, 86:23–30, 2004.

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

[21] Y. Z. Guo, E. M. Feng, and Y. Wang. Optimal HP configurations of proteins by combining local search with elastic net algorithm. *J Biochem Biophys Methods*, 70(3):335–40, 2007.

[22] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill. A test of lattice protein folding algorithms. *Proc Natl Acad Sci USA*, 92(1):325–9, 1995.

[23] R. Helling, H. Li, R. Melin, J. Miller, N. Wingreen, C. Zeng, and C. Tang. The designability of protein structures. *J Mol Graph Model*, 19(1):157–67, 2001.

[24] R. Wroe, E. Bornberg-Bauer, and H. S. Chan. Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: robustness of the superfunnel paradigm. *Biophys J*, 88(1):118–31, 2005.

[25] Y. Xia and M. Levitt. Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins*, 55(1):107–114, 2004.

[26] Fabien P. E. Huard, Charlotte M. Deane, and Graham R. Wood. Modelling sequential protein folding under kinetic control. *Bioinformatics*, 22(14):e203–210, 2006.

[27] M. P. Morrissey, Z. Ahmed, and E. I. Shakhnovich. The role of cotranslation in protein folding: a lattice model study. *Polymer*, 45(2):557–571, 2004.

[28] N. Alexandrov. Structural argument for n-terminal initiation of protein folding. *Protein Sci*, 2(11):1989–91, 1993.

[29] C. M. Deane, M. Dong, F. P. Huard, B. K. Lance, and G. R. Wood. Cotranslational protein folding - fact or fiction? *Bioinformatics*, 23(13):i142–8, 2007.

[30] Rhodri Saunders and Charlotte M. Deane. Protein structure prediction begins well but ends badly. *Proteins*, 78(5):1282–90, 2009.

[31] Martin Mann, Daniel Maticzka, Rhodri Saunders, and Rolf Backofen. Classifying protein-like sequences in arbitrary lattice protein models using latpack. *HFSP Journal*, 2(6):396, 2008.

[32] Erich Bornberg-Bauer. Chain growth algorithms for HP-type lattice proteins. In *Proc of RECOMB'97*, pages 47–55, 1997.

[33] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[34] Neal Lesh, Michael Mitzenmacher, and Sue Whitesides. A complete and effective move set for simplified protein folding. In *Proc of RECOMB'03*, pages 188–195, New York, NY, USA, 2003. ACM.

[35] K. Steinhöfel, A. Skaliotis, and A.A. Albrecht. Stochastic protein folding simulation in the d-dimensional HP-model. In *Proc of BIRD'07*, pages 381–394, 2007.

[36] Gong Zhang, Magdalena Hubalewska, and Zoya Ignatova. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol*, 16:274–280, 2009.

[37] J. Skolnick and A. Kolinski. Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J Mol Biol*, 221(2):499–531, 1991.

[38] S. Bromberg and K. A. Dill. Side-chain entropy and packing in proteins. *Protein Sci*, 3(7):997–1009, 1994.

[39] P. Wang and D. K. Klimov. Lattice simulations of cotranslational folding of single domain proteins. *Proteins*, 70(3):925–937, 2007.

[40] H. S. Chan, S. Bromberg, and K. A. Dill. Models of cooperativity in protein folding. *Philos Trans R Soc Lond B Biol Sci.*, 348(1323):61–70, 1995.

[41] E. I. Shakhnovich. Modeling protein folding: the beauty and power of simplicity. *Folding & design*, 1(3), 1996.

[42] E. Jacob and R. Unger. A tale of two tails: Why are terminal residues of proteins exposed? *Bioinformatics*, 23(2):e225–30, 2007.

[43] Rolf Backofen and Sebastian Will. A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. *Constraints*, 11(1):5–30, 2006.

[44] Martin Mann, Sebastian Will, and Rolf Backofen. CPSP-tools - exact and complete algorithms for high-throughput 3D lattice protein studies. *BMC Bioinformatics*, 9:230, 2008.

[45] K. Mizuguchi, C. M. Deane, T. L. Blundell, M. S. Johnson, and J. P. Overington. Joy: protein sequence-structure representation and analysis. *Bioinformatics*, 14(7):617–23, 1998.

[46] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, 1995.

[47] John Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225(5232):563–564, 1970.

[48] V. S. Pande, A. Y. Grosberg, and T. Tanaka. Nonrandomness in protein sequences: evidence for a physically driven stage of evolution? *Proc Natl Acad Sci USA*, 91(26):12972–5, 1994.

[49] Y. Xia and M. Levitt. Simulating protein evolution in sequence and structure space. *Curr Opin Struct Biol*, 14(2):202–207, 2004.

[50] William R. Taylor. Topological accessibility shows a distinct asymmetry in the folds of $\beta\alpha$ proteins. *FEBS Letters*, 580(22):5263–5267, 2006.

[51] A. Laio and C. Micheletti. Are structural biases at protein termini a signature of vectorial folding? *Proteins*, 62(1):17–23, 2006.

[52] S. Govindarajan and R. A. Goldstein. On the thermodynamic hypothesis of protein folding. *Proc Natl Acad Sci USA*, 95(10):5545–9, 1998.

[53] Preeti Chugha and Terrence G. Oas. Backbone dynamics of the monomeric lambda repressor denatured state ensemble under nondenaturing conditions. *Biochemistry*, 46(5):1141–1151, 2007.

[54] L. H. Holley and M. Karplus. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA*, 86(1):152–156, 1989.

[55] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256(3):623–44, 1996.

Figure legends

Figure 1: The effect of start conformation on attaining the UGEM conformation of `HPHPPHPPHH`. The proportion of simulations resulting in the UGEM conformation (Y-axis) is plotted against the number of folding steps (X-axis). Starting from randomly compact structures (R, solid line) is never better than starting from a CT-fold. Data is shown for CT-folds created using 0, 1, 2 and 3 (dashed lines) intermediate folding events per elongation. The inset chart shows the relative success rate (Y-axis) of using CT-folds over randomly compact structures as a start point for CT-folds created using a varying number of intermediate folding events (X-axis).

Figure 2: The propensity of our different sets to fold successfully. The ratio of the average folding rates (Y-axis) is the success rate when starting from CT-folds compared to the success rate starting from random structures. Data is shown for different numbers of intermediate folding events (X-axis) using length 25 sequences in the 2D-square lattice. Data for our Kinetic-CT set is split into folding to the UGEM conformation and to the unique final conformation found via pure CT folding under our classification system. NCT = Non-CT folders, GCT = Global-CT and KCT = Kinetic-CT.

Figure 3: Possible Hydrophobic Interaction (*PHI*) score data for length 25 sequences. Global-CT sequences (G) have positive scores in the N-terminal region indicating that only a few hydrophobic interactions are possible, i.e. the majority of an H residue's potential contacts are with P residues and thus not favoured. For Kinetic-CT (K) and Non-CT (N) sequences scores are always negative. Lines are a guide to the eye.

Figure 4: Distribution of the Mean Central Residue (*MCR*) and the relative distance of the N- and C-termini to the centre of mass (NC_{cen}) for each SCOP class. Only the α class shows no bias according to the measures. All other classes, and at most the α/β class, show a trend towards CT folding behavior, i.e. $MCR < 0.5$ and $NC_{cen} < 0$.

Figure 1

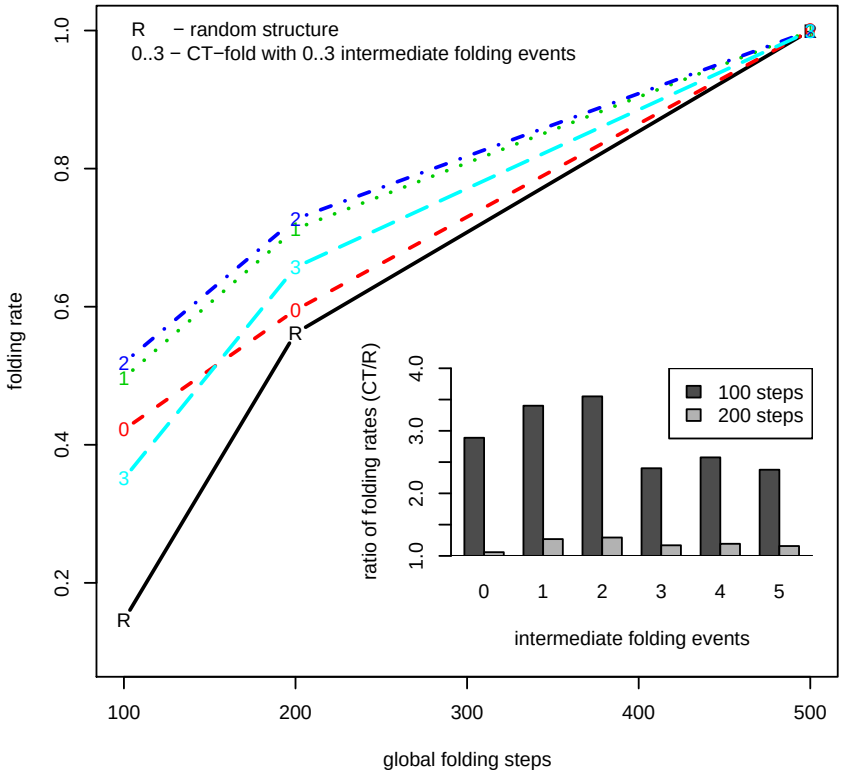
R − random structure
0..3 − CT−fold with 0..3 intermediate folding events
folding rate
global folding steps
ratio of folding rates (CT/R)
intermediate folding events
100 steps
200 steps

Figure 2

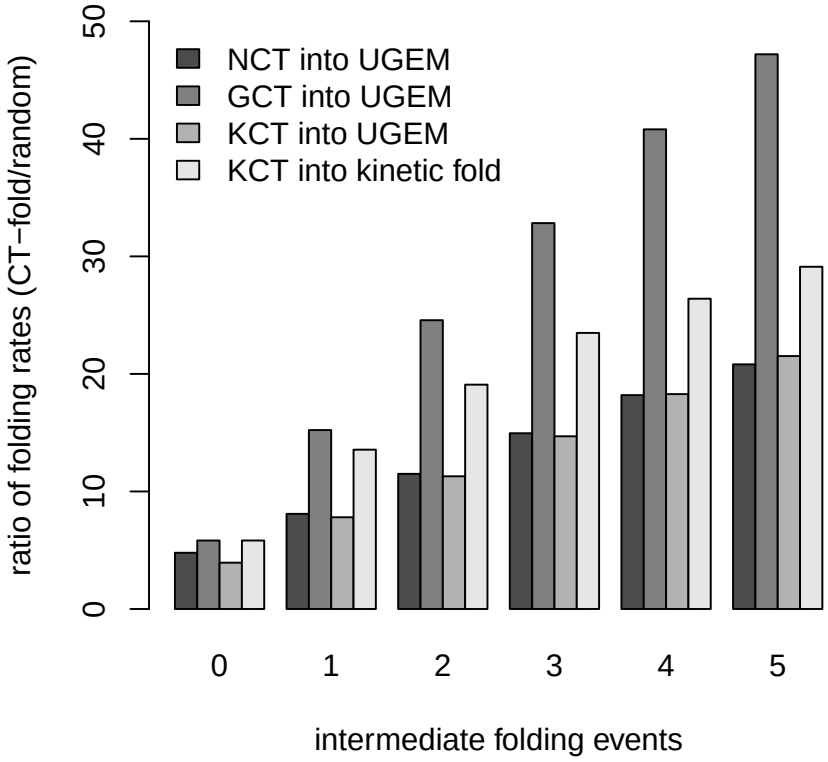
NCT into UGEM
GCT into UGEM
KCT into UGEM
KCT into kinetic fold
ratio of folding rates (CT−fold/random)
0
10
20
30
40
50
0
1
2
3
4
5
intermediate folding events

Figure 3

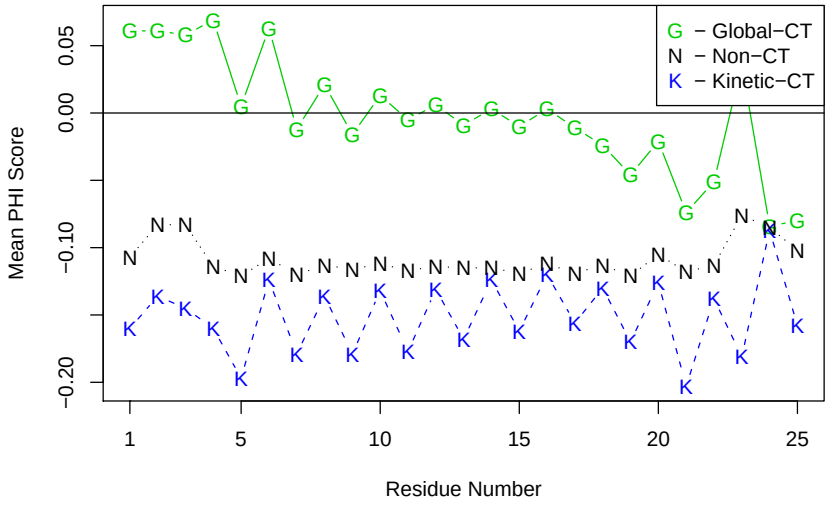
G − Global−CT
N − Non−CT
K − Kinetic−CT
Mean PHI Score
0.05
0.00
−0.10
−0.20
1
5
10
15
20
25
Residue Number

Figure 4