Supplementary material for: Signatures of Co-translational Folding

Rhodri Saunders[1,3], Martin Mann[2,3], and Charlotte Deane[1,4]

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

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

[3]Both authors have contributed equally

[4]Correspondance to: Charlotte M. Deane, University of Oxford, Department of Statistics, 1 South Parks Road, Oxford, England, OX1 3TG. Email: deane@stats.ox.ac.uk, telephone: +44(0)1865 281301, fax: +44(0) 272595

A1 - Co-translation Algorithm for Sequence Classification

Co-translation can be emulated by a chain-growth procedure, which we apply in accordance with [1, 2] for our classification. In the following we discuss the algorithm in detail.

Given:

$P = P_1, \ldots, P_n$	:	monomer sequence from some alphabet A to fold
S	:	structure space of P
$E(P, s \in S)$	:	energy function
N	:	the neighboring vectors of the lattice model to use, e.g. in 2D-square $N = \{\pm(1,0,0), \pm(0,1,0)\}$
ΔE	:	energy interval above the minimal energy for this iteration that are going to be extended in the next, i.e. the surmountable energy barrier [2]
S_i	:	co-tranlational structure space of subsequence $(P_1, \ldots, P_i)$ according to ΔE

Result:

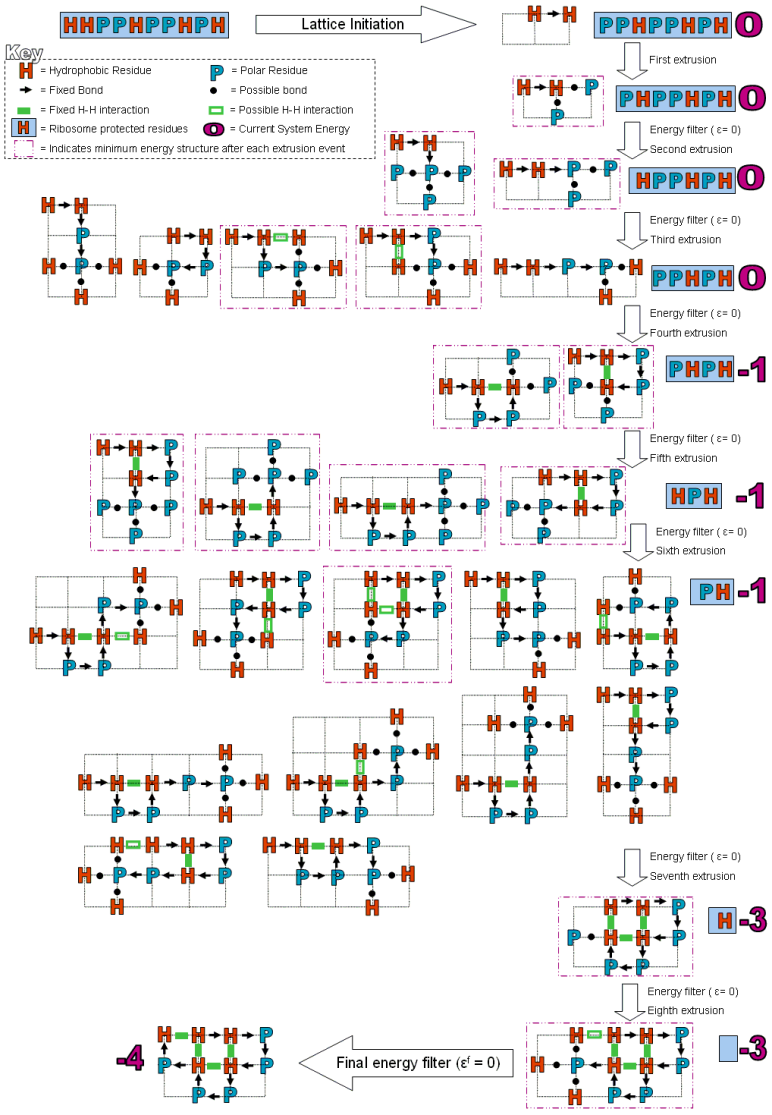
S_n	:	the set of the energetically best structures reachable via co-translational folding pathways with an energy barrier below ΔE

Method:

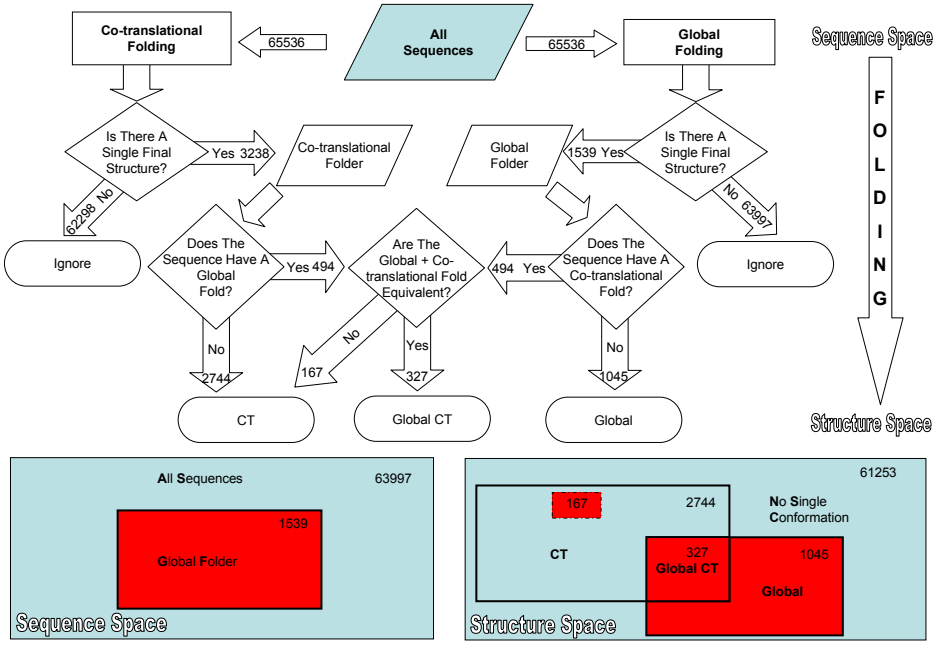
The method follows a greedy structure-elongating chain-growth approach:

$S_1 \leftarrow \{((0,0,0))\}$ ▷ initialized by placing the first monomer to $(0,0,0)$
for $i = 2 \ldots n$ **do**
  $S'_i \leftarrow \emptyset$ ▷ structures generated in current iteration
  for all $s \in S_{i-1}$ **do** ▷ s has length $(i-1)$
    for all $\vec{v} \in N$ **do**
      if $s_{(i-1)} + \vec{v} \notin \{s_1, \ldots, s_{(i-1)}\}$ **then** ▷ check selfavoidingness
        $s' \leftarrow (s_1, \ldots, s_{(i-1)}, (s_{(i-1)} + \vec{v}))$
        $S'_i \leftarrow S'_i \cup \{s'\}$ ▷ store extension
      end if
    end for
  end for
  $minE \leftarrow$ minimal energy of all elements in S'_i
  $S_i \leftarrow \{s \mid s \in S'_i \text{ and } E(P_{1\ldots i}, s) \leq (minE + \Delta E)\}$ ▷ restrict according to ΔE
end for
$S_n \leftarrow \{s \mid s \in S'_n \text{ and } E(P, s) = minE\}$ ▷ store all best

The following figure exemplifies the applied algorithm to fold the HP sequence `HHPPHPPHPH` in the 2D-square lattice. As each structure considered for elongation can have up to three co-translationally accessible lattice nodes the number of structures at each iteration is combinatorial.



Below the relation between sequence and structure sets created by the co-translational folding algorithm above is depicted. Data shown is for HP sequences of length 16 in 2D-square lattice.



A2 - Data Sets

Protein selection criteria for proteins obtained from PISCES webserver are outlined in Table 1. A full list of the proteins used is available at:

`www.stats.ox.ac.uk/bioinfo/resources`

Max % Identity	Resolution	R-value	Chain Length
20	$\leq$ 2Å	$\leq$ 0.3	80 $\leq$ chain $\geq$ 1000 Residues

Table 1: Selection criteria for proteins from PISCES webserver.

The 3D HP sequences used are also available at the above address. The 2D designing sequences of Irbačk and Troein contain over 1.5 million sequences (for a summary see Table 2) can be found here:

`http://cbbp.thep.lu.se/activities/hp/index.html`

HP length	UGEM sequences	Global-CT	Kinetic-CT
4	4	4	0
5	0	0	0
6	7	7	0
7	10	10	0
8	7	7	0
9	6	5	0
10	6	5	0
11	62	33	2
12	87	38	2
13	173	74	1
14	386	133	26
15	857	230	55
16	1539	327	167
17	3404	597	301
18	6349	781	699
19	13454	490	554
20	24900	1864	2691
21	52183	3008	5016
22	97478	4238	10667
23	199290	7121	19230
24	380382	10269	40072
25	765147	17085	74502

Table 2: The designing sequences of Irbačk and Troein broken down by sequence length, column 1. The number of sequences with a unique global energy minimum conformation (UGEM) are shown in column 2; the number of these that we classify as Global-CT and Kinetic-CT are shown in columns 3 and 4 respectively.

A3 - Other Tested Measures of Co-translational Folding

Over the course of our study a number of measures were tested and evaluated. A number of these came, or were adapted from, the relevant literature. Here we outline the measures not discussed in the main paper because they did not adequately define our folding sets. Throughout, R_i denotes the i-th residue with $1 \leq i \leq n$ 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 helix or strand. The sequence record of PDB files is sometimes incomplete, not all residues are resolved in the X-ray structure, thus in these cases n is based on the actual residue number to incorporate chain breaks.

Localness of interaction (LI)

Localness of interaction (LI) measures the proportion of local contacts within a structure (Eq. 1). For the HP model two residues R_i and R_j are in *contact* if they occupy neighbored lattice nodes. A contact is a *local contact* if $|j - i| \leq 5$.
For protein structures a contact is defined as two C_α atoms within 5Å of each other; for glycine the C_β is used. The contact is local if the contacting residues are separated by less than 15 residues. Contacting residues separated by less than 5 residues are discounted because their proximity is likely to be sequence rather than fold dependent.

$$LI = \frac{\sum_{1 \leq i < j \leq n} LocalContact(R_i, R_j)}{\sum_{1 \leq i < j \leq n} Contact(R_i, R_j)} \tag{1}$$

LI with direction (LID)

The measure can be given directionality, LI with direction (LID) is given in Eq. 2 for a sequence of length n. Here a guard region (g) is used because residues within g of the termini cannot form local contacts in both directions. In the HP model $g = 4$ and for proteins $g = 7$. To remove bias we calculate the LI for each residue in turn from R_{1+g} to $\frac{n}{2}$; while simultaneously mirroring the calculation from R_{n-g} to $\frac{n}{2}$. If no contacts are formed for either residue the residues are grouped with the subsequent residue(s) until both numerator and denominator are non-zero. If LID is greater than zero then residues in the N-terminus makes more local contacts than residues in the C-terminus. Thus, for co-translational folding we would expect a LID score > 0.

$$LID = \sum_{i=g}^{n/2} \log \frac{LI(R_{1+i})}{LI(R_{n-i})} \tag{2}$$

Contact Previousness (CPrev)

Identical to the sum of logs ratio developed by Deane *et al.* [3]; we use $CPrev$ in both the HP model and protein structures to determine if there is N- or C- terminal bias towards previous contacts. The question is: does R_{1+i} make more N-terminal contacts (i.e. to R_s with $s \in [1, 1+i]$), than R_{n-i} makes C-terminal contacts (i.e. to R_e with $e \in [n-i, n]$)?
Thus, we use the function $Contact(R_i, R_j)$ to determine if two residues are in contact or not. The function follows the contact definition as given above for LI. The calculation is mirrored about the central point in sequence space, so we calculate the $CPrev$ for each residue in turn from R_{1+g} to R_{n-g}; while simultaneously mirroring the calculation from R_{n-g} to R_{1+g}, see Eq. (3).

$$CPrev = \frac{1}{n-2g} \sum_{i=g}^{n-g} \log \frac{\sum_{s=1}^{1+i} Contact(R_{1+i}, R_s)}{\sum_{e=n}^{n-i} Contact(R_{n-i}, R_e)} \tag{3}$$

If no previous contacts are formed by either residue the residues are grouped with the subsequent residue(s) until both numerator and denominator are non-zero. In these circumstances division is by the number of residue groups formed rather than by the sequence length n.

Spatial Previousness - SPrev

$SPrev$ is a measure of how close a residue is in space to residues that are more N-terminal. $SPrev$ is difficult to transfer from the 2D HP model to the 3D-space of proteins; therefore, two different equations are formulated. In the HP model we use $SPrev_{HP}$ (Eq. 4), which calculates the minimum distance between R_i and residues R_1 to R_{i-3}. This is done for each residue in turn from R_1 to R_n; hence, tracking $SPrev_{HP}$ along the folding pathway if co-translational folding is assumed. For global folders it is expected that $SPrev_{HP}$ will be greater than for co-translational folders because there is no presumed requirement that residues are placed close to previously synthesised residues.

$$SPrev_{HP}(R_i) = \min_{1 \le j \le i-3} \delta(R_i, R_j) \tag{4}$$

For proteins, $SPrev_{PROT}$, we utilise the solvent accessibility (EX) of secondary structure elements (SSEs). Assuming co-translational folding, we calculate the mean solvent accessibility for each SSE greater than 3 residues. The co-translational solvent accessibility of each SSE is normalised to the

mean solvent accessibility of the SSE in the full structure (Eq. 5). Solvent accessibility is calculated as described by Lee and Richards [4] utilising the PSA program within JOY [5]. An $SPrev_{PROT}$ score greater than 1 indicates that there is a large disparity between the co-translational and global solvent accessibility of the SSE; this would not be expected for a co-translational folder. However, we found that a SSE completely buried in the global fold would often produce a result greater than 1. For this reason we also take into account the co-translational solvent accessibility of the SSE, its exposure. We investigate exposure at three levels: 70%, 80% and 90%. We expect that co-translational folders will have fewer exposed SSEs with a $SPrev > 1$; and that these SSEs will be shorter. For this calculation the most N-terminal SSE is ignored because it will, by definition, be co-translationally exposed.

$$SPrev_{PROT}(R_i) = \frac{\sum_{i=1}^{n} EX_{CT}(R_i)}{\sum_{i=1}^{n} EX_G(R_i)} \tag{5}$$

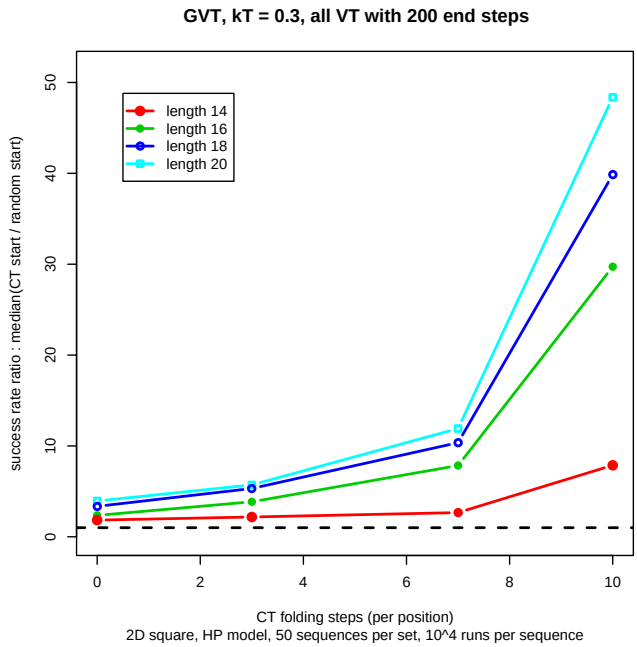
Where EX_G is the exposure of the SSE in the full protein and EX_{CT} is the exposure of the SSE when only the residues N-terminal to it are present.

Terminal distance - NCdis

$NCdis$ measures the distance between the N-terminus and the C-terminus, assumed to be shorter under co-translational folding.

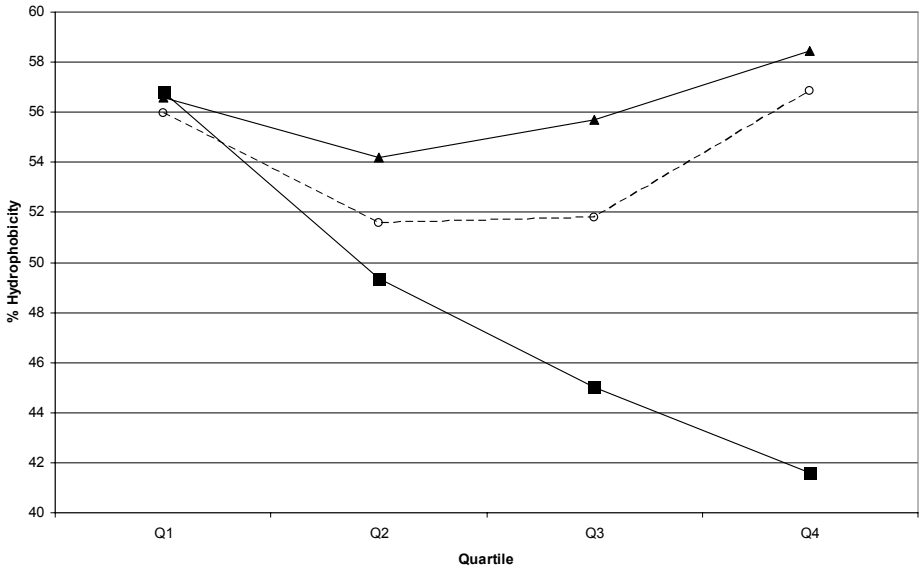
$$NCdis = \delta(R_1, R_n) \tag{6}$$

A4 - Intermediate Folding Events



The influence of the number of intermediate folding events. Data within the given figure is shown for Global-CT sequences folded using the Markov chain approach from the Methods section followed by a standard global folding simulation.
At each length 50 sequences have been randomly selected from the overall sets of Global-CT folders. For each sequence 10^4 global folding simulations of 200 global folding steps have been performed, either starting from random compact structures (see Materials) or from CT-folds derived with different numbers of intermediate folding steps (X-axis). The figure visualises the relative success rate when starting from CT-folds to find the unique global energy minimum (UGEM) fold relative to the success rate when starting from random compact structures.
An increasing number of intermediate folding events (X-axis) significantly increases the success rate of finding the UGEM fold (Y-axis) relative to success rates in pure global folding from a randomly compact structure. This effect is clearly length dependent, i.e. the longer a sequence is the more it is influenced by CT folding. This results from the exponentially growing structure space where CT folding enables a potent restriction of the folding pathways towards the UGEM structure.

A5 - Hydrophobicity by Quartiles



Hydrophobicity By Quartiles at length 25 in the 2D-square lattice clearly shows that Global-CT sequences (squares) behave differently than both Kinetic-CT (triangles) and Global (circles, broken line) sequences.

A6 - Putative Global-CT Sequences at length 30

In order to assess the effectiveness of our HIP score in identifying Global-CT sequences from all other HP sequences we attempted to identify Global-CT sequences from the 2^{30} HP sequences of length 30. To our knowledge no UGEM data exists at this length. For each sequence we calculate the mean HIP score for residue 1 to 8 - where our fold sets diverge the most - and the relative terminal hydrophobicity (tH). tH is given by the log of N-terminal hydrophobicity over C-terminal hydrophobicity. Under co-translation we expect a positive tH score. Sequences with a HIP > 0.3 and tH > 0.5 were selected as possible Global-CT sequences. In total, 981,143 sequences were selected, 0.09% of the sequence set. Of those, 1087 sequences were found to have a final single conformation following vectorial folding at $\Delta E \in \{0, 1, 2, 3\}$ (not all sequences could be tested at all energy intervals). Extrapolating from data at length 25 we expect to identify 0.01% of sequences as Global-CT. Our simple selection on HIP and tH thus provides a 8.5 fold enrichment.

The 92 sequences we identified using HIP and HBQ as being strong candidates for Global-CT folders are given in Table 3. The protein structure is described in terms of absolute moves: **U**p (Y+1), **D**own (Y-1), **L**eft (X-1) and **R**ight (X+1); with all conformations normalised to start with an Up.

Table 3: Putative Global-CT sequences where E denotes the energy of the given sequence-structure pair.

Sequence	E	Structure
HHHPHHPPPHHPPPPHPPPHPPPHPPPPHP	-10	UULDDDRRUURRULLUULDLLDDLLDRRD
HHHPHHPPPPHPPHPPHHPPPHPPHPPPPHP	-10	UULDDDRRURULUULUULDDLDDLLDRRD
HHHPHHPPPPHPPPHHPPPPHPPPPPHPHP	-10	UULDDDDRURRULURRULLULLLLDDLDRD
HHHPHPHPPPHHPPPPHPPPPPHHPPHPPHP	-10	ULURRDRRDLLDLLULUUURRRRULULDL
HHHPHPHPPPHHPPPPHPHHPPPPHPPPPHP	-10	ULURRDRRDLLDLLULUUUULDDLLDRRD
HHHPHPPPHPHHPPPPHPHHPPPPHPPPPHP	-10	ULURURDDRDLDLLULUUUULDDLLDRRD
HHHPHPPPHPPPPHPPHPPPPPHPPPHPHP	-9	UULLDLDRDRRURULURRULLLLURURDR
HHHPHPPHPPPPHPPHPPPPHPPPHPPPHP	-9	UULDLDRDRRURULURULLLLURUURDDR
HHHPPPHHHPPHPPHPPHPHPPHPPPPPHP	-10	UULLDRDDDRURUURULULULDLLDDDRD
HHHPPPHPHPHHPPPPHPHHPPPPHPPPPHP	-10	ULUURDRDRDLDLLULUUUULDDLLDRRD
HHHPHHPPPPHHPPHPPPPHPPPHPPPPHP	-10	UULDDDRRUUURULUULDDLLDDLLDRRD
HHHPHHPPPPPHPPPHPPHPPPPHPPPPHP	-9	UULDDLUUURRURDDRDLDRDLLDDLUUL
HHHPHHPPPPPHPHPPPPHPPPPPHPPHP	-9	UULDDLUUURRURDDRRDLLDRDLLDLUL
HHHPHHPPPPPHPPPHPPPPHPPPPPHP	-9	UULDDLUUURRURDDRRDLLDDLDDLUUL
HHHPHHPPPPPHPHHPPPHPPPPPHPPHP	-10	UULDDDDDRUURUURRULLULLLLDDLDRD
HHHPHHPPPPPHPHHPPPHPPPPPHPPPP	-10	UULDDDDDRUURUURULULLLLDDLLDRRD
HHHPHHPPPPPHPHPPPPHPPPPPHPPPP	-9	UULDDLUUURRRDRRDLLDRDLLDDLUUL
HHHPHHPPPPHHPHPPPPHPPPPPHPPPP	-10	UULDDDRRUURDRUULLULLLLDDLLDRRD
HHHPHHPPPPHHPHPPPPHPPPPPHPPPP	-10	UULDDDRRUURDRUULLULLLLDDLLDRRD
HHHPHHPPPPHPHPPPPHPPPPPHPPPHP	-10	UULDDDRRURULUULULDLLDRDLLDRRD
HHHPHHPPPPHPHPPHPPPPHPPPPHPHP	-10	UULDDDRRURULUULUULDDLLDRDLDRD
HHHPHHPPPPHPPHPPHPPPPHPPHPPPHP	-10	UULDDDRRURULUULUULDDLDDLLDRRD
HHHPHHPPPPHPPPPPHPPHPPPPHPPPHP	-9	UULDDLLURUURRRDRDLDRDLLLDDLUUL
HHHPHHPPPPHPPPPPHPPPPPHPPPPHP	-9	UULDDLLURUURRRDRRDLLDRDLLDLUUL
HHHPHHPPPPHPPPPPHPPPPHPPHPPPHP	-9	UULDDLLURUURRRDRRDLLDDLDDLUUL
HHHPHHPPPPHPPPHHPPHPPPPPHPPPHP	-10	UULDDDDRRURULUULUULDDLLDDLLDRD
HHHPHHPPPPHPHHPPPHPPPPPHPPPPP	-10	UULDDDDRRURULUULUULDDLLDRRDLDRD
HHHPHHPPPPHPHPPPPHPPPPPHPPPHP	-10	UULDDDRRURULUULUULDDLLDDLLDRRD
HHHPHHPPPPHPHPPPPHPPPPPHPPPPP	-9	UULDDLLURUURRRDRDLDRDLLLDDLUUL
HHHPHHPPPPHPPPPPHPPPPPHPPPHP	-9	UULDDLLURUURRRDRRDLLDRDLLDLUUL
HHHPHHPPPPHPPPPPHPPPPPHPPPPP	-9	UULDDLLURUURRRDRRDLLDDLDDLUUL
HHHPHHPPPPHPPPHHPPHPPPPPHPPPP	-10	UULDDDDRRURULUULUULDDLLDRRDLD
HHHPHHPPPPHPPPHHPPHPPPPPHPPPH	-10	UULDDDRRURULUULUULDDLLDRRDLDL
HHHPHHPPPPHPPPHHPPHPPPPPHPPPP	-9	UULDDDDRRURULUULUULDDLLDRRDLD
HHHPHPHPHHPPPPHPPPPPHHPPPPHPHP	-10	ULURRDRDLDLLULUUURRRRRULLULDL
HHHPHPHPHHPPPPHPPPPPHHPPPHPPHP	-10	ULURRDRDLDLLULUUURRRRULUULDDL
HHHPHPPPHHPPPPHPPPPPHPPHPPPPHP	-8	ULURRRDLDDLLULUUURRURDRRDDDLD
HHHPHPPPPHPHHHPPPHHPHPPPPPPPHP	-10	UULDLLDRRDRRURDDLLDRRRUUUULLU
HHHPPPHHHHPPHPPPPHPPPPHPPPPHPH	-11	ULURRDDRUURDRRDLLDDLULDDLUULU
HHHPPPHHHHPPPPHPPPPHPPPPHPPHPH	-11	ULURRDDRUUURDDRRDLLDDLULDLULU
HHHPPPHHHHPPPPHPPHPPPPHPPPPHPH	-11	ULURRDDRUUURDDRRDLDDLULDDLUULU
HHHPPPHHHHPHPPPPHPPPPPPPPPHPH	-11	ULURRDDRUUUURDDRRDLLDDLLLULU
HHHPPPHHHPHPPPPHPPPPPPPHHPPHPH	-11	UUULDDDDLULLURRUUURRRDDDRDLDL
HHHPPPHHHPHPPPPHPPPPPPPHPPPHPH	-10	ULURRDDDRURRULLUULLLLDDLDRRDR
HHHPPPHHHPPHPPPHPPPPPPPPPHPPH	-11	ULURRDDDDRUUUURDRDLDDDLLLUULU
HHHPPPHHHPPHHHPPPHPPPPPHPPPHP	-10	ULURRDDDDLULULDDRDRRRUURRULLU
HHHPPPHHHPPHPPHPPHPPPPPPPHPPH	-10	UUULDDDDDRURUURULUULLLDDDLDRD
HHHPPPHHPPHHPHPPPPPHPPPHPPPPH	-9	ULURRDDRDLLLULDDRRRRRDLDDLUUL
HHHPPPHHPPHPPPPHPPPPPPPPPHPPH	-9	ULURRDDDRURRULLUULLLLDDLDRRDR
HHHPPPHHPPPPPHHPHPPPPPHPPPHPH	-9	ULURRDDRRDLLLLULDDRRRRRDLDLUL
HHHPPPHHPPPPPHPPHPPPPPPPPHPPH	-9	ULURRDDDDRUURULUULLLLDDLDRRDR
HHHPPPHHPPHPPHPPHPPPPPPPPHPPH	-9	ULURRDDDLLULULLURRURRRDDRRDLLD
HHHPPPHHPPPPPHPHPPHPPPPPPPHPH	-9	ULURRDDDLLULULLURRURRRRDDRDLD
HHHPPPHHPPHPPHPPHPPPPPPPPHPPH	-10	ULURRDDDLDLUULLURUURRRRDDRDLD
HHHPPPHHPHPPHHPHPPPPPPPHPPPHP	-10	ULURRDDDLDLUUUUURRRRDDRRDLLD
HHHPPPHHPHPPHPPHPPPPPPPHPPPHP	-10	UULLDRDDRDRUUUURULULDLLLDDDRD
HHHPPPHHPHPPHPPPPHPPPPPPPHPPH	-9	UUULDDDDRRUURRULLUULLLDDDLDRD
HHHPPPHHPHPPPHPPPHPPPPPPPHPPH	-9	ULURRDDDLDLUULLURUURRRRDDRDLD
HHHPPPHHPHPPPHPHPPHPPPPPPPHPPH	-10	ULURRDDDLDLUULULURURRRRDDRDLD
HHHPPPHHPHPPPHPPPPPPPPPHPPPHP	-9	ULURRDDDLDLUUULUUURRRRDDRRDLLD

Continued on next page

Table 3 – continued from previous page

Sequence	E	Structure
HHHPPPHHPHPPPHPHPPPPPPPHPPPHHP	-9	ULURRDDDLDLUULUUURRRRDDRRDLLD
HHHPPPHHPHPPPHPHPPPPPHPHPPPPHP	-9	ULURRDDDLDLUULUUURRRRDDRRDLLD
HHHPPPHHPHPPPHPHPPPHPPPHPPPPHP	-9	ULURRDDDLDLUULUUURRRRDDRRDLLD
HHHPPPHHPHPHPPPHPPPPPPPHPPPHP	-9	ULURRDDDLLULLLURUURRRRDDRRDLLD
HHHPPPHHPHPHPPPHPPPPPPPHPPPHHP	-9	ULURRDDDLLULLLURUURRRRDDRRDLLD
HHHPPPHHPHPHPHPPHPPPPPPHPPPPHP	-10	ULURRDDDLLULULURURRRRDDRRDLLD
HHHPPPHHPHPHPHPPPPPHPPPPPPHPPHP	-10	ULURRDDDLLULULLURRURRRRDDRDLD
HHHPPPHPHHPPPHPPPPPPPHPPPHPPHP	-9	ULLURRRDDDLLULLUUURRRRRDDRDLD
HHHPHHPPPPPPPHPPPPHPPPPHPPPPHP	-8	UULDDLUUURRRDRRDLLDRDLLDDLUUL
HHHPHPPPHPPPPPHHHPPPPHPPPHPPHPH	-11	ULURURDDRRDLLDLDDDLUULLUUURDDR
HHHPPPHHPPPHPPPHPPPPPPPHPPPPHP	-8	ULURRDDDLLULLURUURRRRDDRRDLLD
HHHPPPHHPPPHPPPHPPPPPPPHPPPHHP	-8	ULURRDDDLLULLURUURRRRDDRRDLLD
HHHPPPHHPPPHPPPHPPPPPPHPHPPPPHP	-8	ULURRDDDLLULLURUURRRRDDRRDLLD
HHHPPPHHPPPHPPPHPPPHPPPHPPPPHP	-8	ULURRDDDLLULLURUURRRRDDRRDLLD
HHHPPPHHPHHHPPPHPPPPPPPHPPPPHP	-9	ULURRDDDLLULLURUURRRRDDRRDLLD
HHHPPPHHPHPPHPPHPPPPPPPHPPPPHP	-9	UUULDDDDRRUURULUULLLDDDLLDRRD
HHHPPPHHPHPHPPPHPPPPPPHPHPPPPHP	-9	ULURRDDDLLULLURUURRRRDDRRDLLD
HHHPPPHHPHPHPPPHPPPHPPPHPPPPHP	-9	ULURRDDDLLULLURUURRRRDDRRDLLD
HHHPPPHPHPHHPPPHPPPPPPHPPPHPHH	-9	ULLURRRDRDLDDLLULLUUURRUURDRDL
HHHPHHPPPHHPPHPPPPPPPHPPPPHPHH	-10	UULDDLDRRDRURRULLURUUULLLDRRD
HHHPHHPPHPHPPPHPPPPPPHPHPPPPHH	-9	UULDDLDRRDRUUURDDDDLDLLLUURDR
HHHPHHPPPHPHPPPHPPPPPPPHPHPPHH	-9	UULDDLDRRDRUUURDDDDLLLDLLURUR
HHHPHHPPPPHPHPPHPPPPHPPPPPHHH	-9	UULDDLLDRRDDRUURRULURRRULLULD
HHHPHHPPPPHPPHPHPPPPPPPPHPHH	-9	UULDDDRRURULUULLLDLLDDDRRULUR
HHHPHPPPHPHPHPHHPPPPPPHPPPHHH	-9	ULURURDDRDLDDRURDDLLLULLUURDR
HHHPHPPPHPHPHPHHPPPPPPHPPPPHH	-9	ULURURDDRDLDDRURDDLLLULLUURDR
HHHPHPPPHPHPHPHPPPPPPPHPPPHHH	-9	ULURURDDRDLDDRURDDLLLULLUURDR
HHHPHPPPHPHPHPHPPPPPPPHPPPPHH	-9	ULURURDDRDLDDRURDDLLLULLUURDR
HHHPHPPPPHHPHPHPPPPPPPHPPHPHHP	-8	UULDLLDRRDRRUURRUUULLLLDRRDRU
HHHPHPPPPHPHHPHPPPPPPPHPPPPHH	-9	UULDLLDRRDRDRURRULLURRRULLULD
HHHPPPHHHPHPPPHPPPPPPPHPPPPHH	-10	UUULDDDDLUULURUURRRDDDRRDLLDL
HHHPPPHHHPHPPPHPPPPPPPHPPHPHHP	-8	UULLDRDLDRRRUURRUUULLLLDRRDRU
HHHPPPHHHPHPPPHPPPPPPPHPPHPHPH	-8	UULLDRDLDRRRUURRUUULLLLDRRDRU
HHHPPPHHHPPPHPHPPPPPPPHPPHPHPH	-9	UUULDDDDLLURUUUURRRDDDRDLDLDL
HHHPPPHHHPPHHPPPPPPPHPPPPHPHPH	-10	ULURRDDRRULURRDDDLLLDDLUULULU
HHHPPPHHPPHPHPHPPPPPPPHPPHPHHP	-8	UULLDRDLDRRRUURRUUULLLLDRRDRU

A7 - Local Contacts

Local contacts (LC) are those between residue i and residues j with $|j-i| \leq 5$. The tables below show, for each sequence/structure set, the sequence length considered (column 1), the number of sequences in the set (column 2), the percentage of local contacts (columns 3 and 5) and the mean number of local contacts per structure (columns 4 and 6). Data in columns 5 and 6 is only calculated from the first six residues of each structure. The first six residues were analysed because HIP best distinguishes between the sequence/structure sets in this window.

Global-CT

length	no.Seqs	% LC	Mean LC per structure	% LC 1to6	Mean LC 1to6 per structure
20	1864	51.96	8.34	50.62	3.50
21	3008	51.11	8.55	49.96	3.53
22	4238	51.88	9.24	49.90	3.57
23	7121	51.55	9.62	49.47	3.57
24	10269	51.55	10.15	49.54	3.67
25	17085	51.17	10.49	49.12	3.64

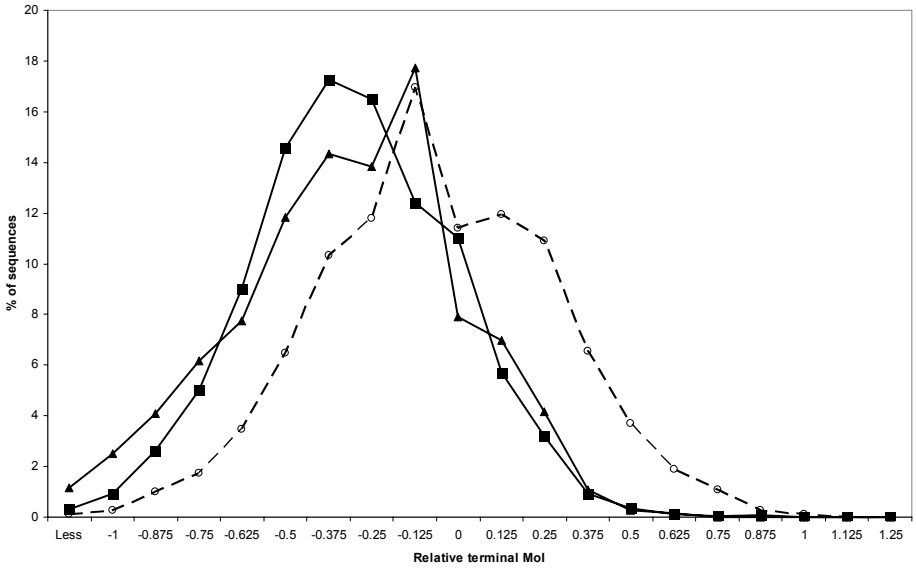
Kinetic-CT

length	no.Seqs	% LC	Mean LC per structure	% LC 1to6	Mean LC 1to6 per structure
20	2691	58.07	9.31	58.32	3.89
21	5016	60.39	10.06	60.82	3.93
22	10667	58.28	10.34	58.05	3.91
23	19230	59.27	10.85	58.88	3.92
24	40072	57.77	11.15	56.94	3.89
25	74502	58.71	11.73	57.66	3.91

Global

length	no.Seqs	% LC	Mean LC per structure	% LC 1to6	Mean LC 1to6 per structure
20	20345	47.13	8.35	43.26	2.81
21	44159	46.71	8.63	42.44	2.72
22	82573	46.87	9.21	41.97	2.80
23	172939	46.48	9.48	41.46	2.72
24	212447	46.12	9.91	40.27	2.75
25	673560	46.13	10.31	40.48	2.71

A8 - Terminal Moment of Inertia



Terminal Moment of Inertia for sequences of length 25 in 2D-square lattice. Global-CT (squares) and Kinetic-CT (triangles) structures have the expected negative score. Global structures (circles, broken line) have an average score of zero.

A9 - SCOP Domains Showing Evidence of CT folding

Exploring the fold level of the SCOP hierarchy, we identify SCOP folds that exhibit strong characteristics of vectorial folding by analysis of results from 6 measures: $NCcen, NCdis, CPrev, MCR$, relative terminal hydrophobicity, and relative terminal MoI (see methods). Of the 835 folds present in our data set we observe 26 folds that occur at least twice in the top fifty of each measure. Interestingly, although the α/β domains show the most vectorial character on analysis at the class level, we find that α domains dominate at multiple occurrences at the fold level. Overall, at two or more occurrences there are: 10 α, 4 β, 3 α/β, 7 $\alpha+\beta$, and 2 other folds. This may suggest that vectorial folding can be characterised by many weak or a few strong indicators. These, suspected, co-translational domains are listed in the Table 4.

SCOP fold	a.229	a.68	d.235	d.9	a.166	a.28	a.49	b.4
Occurences	4	4	4	4	3	3	3	3
SCOP fold	c.130	c.28	d.204	f.39	g.37	g.39	g.51	g.61
Occurences	3	3	3	3	3	3	3	3

Table 4: SCOP domains showing multiple co-translational characteristics

A10 - Contact Potential in SCOP Domains

We investigate the interaction potential of SCOP domains (Figure 1). The profiles of secondary structure elements differ, e.g. helix is more locally compact than strand [6]; for this reason only the all alpha and all beta class are considered as they contain single elements. Using the side-chain centre of mass, we assess all side chain interactions within a 5Å cut-off and assign an energy score via the Miyazawa-Jernigan matrix [7]. Contacts between residues adjacent in sequence are ignored. As with HIP a favourable interaction is scored -1 and an unfavourable interaction +1. Due to different domain sizes we analyse the data in segments (3 to 20). In general the most N-terminal segment has a lower proportion of favourable contacts than subsequent segments. However, the overall energy of the segments does not differ significantly suggesting that these few favourable interactions are stronger than average.

A11 - Terminal Compactness

Our results (Figure 2) suggest that N-terminal regions may be more compact and closer to their global energy minimum structure than C-terminal regions. Similarly, the C-terminal may show more subtle structural variance due to increased flexibility over the N-terminus. These suggestions run counter to the work of Laio and Micheletti [8] whose investigation of 458 proteins showed that the C-terminus is generally more compact than the N-terminus. We reran a number of their tests on a much larger data set, 2618 non-redundant proteins selected via the PISCES web server [9]. Compactness is calculated using both the Radius of Gyration (RoG - Eq.7) and the Moment of Inertia (MoI - Eq.8) following [2]. The radius of gyration is computed by

$$RoG = \sqrt{\sum_{1 \leq i < j \leq n} [\delta(R_i, R_j)]^2} \; . \tag{7}$$

MoI is the average distance of all residues (R_i) in the range $1 \leq i \leq n$ to the their centre of mass (M). M is calculated as the average co-ordinates X,

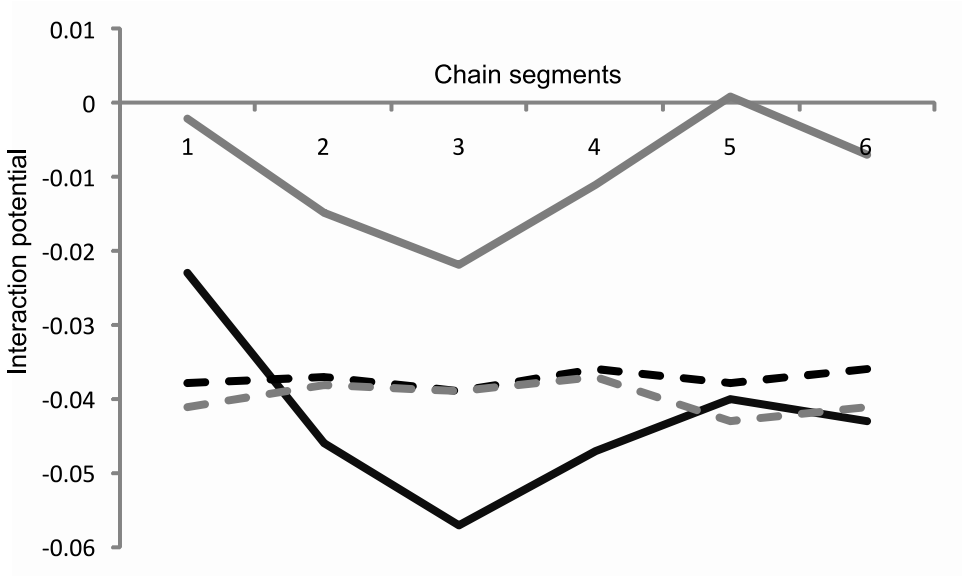


Figure 1: The N-terminus has a low contact potential. In both alpha domains (black) and beta domains (grey) the interaction potential (solid line) is high at the N-terminus (segment 1) and decreases in the subsequent segments. Through out, the overall contact energy of the segments (dotted line) stays relatively constant. This suggests that though fewer favourable contacts form at the N-terminus the favourable contacts are stronger. Contacts are defined as being two side-chain centres of mass within 5Å of each other, the energy of interaction is assigned from the Miyazawa-Jernigan matrix.

Y and Z of all residues. Where n is the number of residues and δ measures the structural distance between two coordinates.

$$MoI = \frac{1}{n}\sum_{i=1}^{n}[\delta(R_i, M)]^2 \tag{8}$$

For each protein in our set we take a structural fragment of length l (where l varies from 6 to 40) from both the N-terminus and the C-terminus. The MoI and RoG of each structural fragment is then calculated and the relative terminal compactness calculated: $log(N{-}compactness)/(C{-}compactness)$. For both MoI and RoG, if the N-terminal fragment is more compact we expect a negative value to be returned. When considering our whole data set our findings support those of Laio and Micheletti. However, as discussed in the main paper β-strand is more prevalent at the N-terminus and is known

to be a less compact secondary structure than α-helix. When we consider only proteins with equivalent termini, that is the most N-terminal secondary structure is the same as the most C-terminal secondary structure then we find that, in general, the N-terminus is more compact than the C-terminus. Thus, measures of relative structural compactness are not independent of secondary structure types and it is only a fair test to compare structures with equivalent termini.

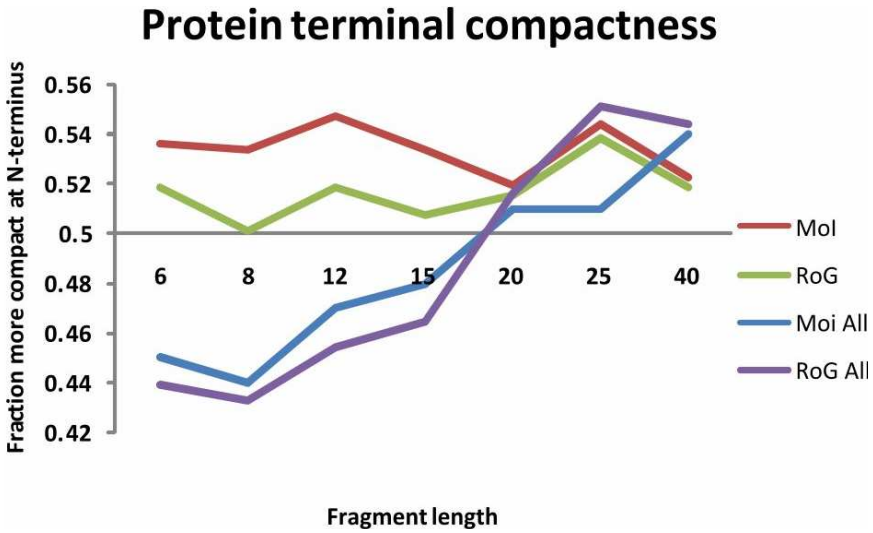


Figure 2: When comparing equivalent termini, we show that the N-terminus is generally more compact than the C-terminus (red and green lines). This is true for two different measure of compactness, Moment of Inertia (MoI) and Radius of Gyration (RoG). Equivalent termini are where both the N- and C-termini have the same type of secondary structure - e.g. either both helix or both strand. If non-equivalent termini are also included (blue and purple lines) the more compact terminus varies with fragment length.

References

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

[2] Huard, F. P. E., C. M. Deane, and G. R. Wood. 2006. Modelling sequential protein folding under kinetic control. *Bioinformatics*. 22:e203–210.

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

[4] Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*. 55:379–400.

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

[6] Saunders, R., and C. M. Deane. 2009. Protein structure prediction begins well but ends badly. *Proteins*. Accepted, DOI 10.1002/prot.22646.

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

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

[9] Wang, G., and R. L. Dunbrack. 2005. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res*. 33:W94–8.