Sequence-space selection of cooperative model proteins

2004 J. Phys. A: Math. Gen. 37 L197

(http://iopscience.iop.org/0305-4470/37/18/L02)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **37** (2004) L197–L202 PII: S0305-4470(04)75838-2

LETTER TO THE EDITOR

Sequence-space selection of cooperative model proteins

Ariel Fernández[1,2] **and Kristina Rogale**[2,3]

[1] Indiana University School of Informatics and Center for Computational Biology and Bioinformatics, Indiana University Medical School, 714 N. Senate Ave., Indianapolis, IN 46202, USA
[2] Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL 60637, USA
[3] Department of Mathematics, The University of Chicago, Chicago, IL 60637, USA

E-mail: arfernan@iupui.edu

Received 8 February 2004, in final form 22 March 2004
Published 20 April 2004
Online at stacks.iop.org/JPhysA/37/L197 (DOI: 10.1088/0305-4470/37/18/L02)

Abstract

We introduce a cooperative lattice model to account for specific many-body effects in protein folding. Cooperativity arises as local solvent environments are determined by the large-scale organization of the chain. In contrast to an uncorrelated lattice, a selection pressure in sequence space arises: foldable sequences endowed with sharp equilibrium conformation ensembles are scarce, indicating that a few sequences are able to exclude the solvent where it matters to warrant the integrity of the structures they form in solution.

PACS numbers: 87.15.He, 87.10.+e, 87.15.Da

1. Introduction

Cooperativity remains a poorly understood attribute of the protein folding process [1–5]. The term is often used to imply that a spatial concurrence of distant units along the protein chain is necessary to preserve structural integrity along folding pathways. In general, the basic secondary structure motifs (α-helix or β-sheet) cannot prevail over time when taken in isolation. Thus, cooperativity suggests a many-body problem [1–5], since the removal or alteration of a single residue often has consequences for the overall organization of the chain [1–3]. However, the physical origin of cooperativity remains a subject of debate.

In this regard, a building constraint for soluble proteins has been recently discovered which is likely to provide a physical basis for cooperativity: soluble protein structure can only prevail if most intramolecular hydrogen bonds and other electrostatic interactions are kept 'dry in water' by clustering hydrophobic groups around them [3, 4]. Thus, a thorough examination of the protein data bank (PDB) reveals three basic features [5, 6]: (a) over 92% of the backbone hydrogen bonds in stable folds are thoroughly dehydrated intramolecularly;

(b) the dehydration of such bonds requires the participation of the units paired by the hydrogen bond assisted by the spatial concurrence of other residues required to remove the surrounding water and (c) the identification of packing defects in the form of insufficiently dehydrated hydrogen bonds is robust to variations (up to 2 Å) in the dimensions of the desolvation domain associated with the bond (the extent of desolvation of a hydrogen bond obviously changes with the domain dimensions, but the statistical outliers to complete dehydration remain the same) [3–6]. Similar observations hold for other intramolecular electrostatic interactions which can only prevail if they are kept dry in water, thus enhancing the Coulombic force and preventing water attack and the concurrent solvation of the polar groups involved.

Recent theoretical developments [4] reveal that a considerable intramolecular desolvation of electrostatic pairwise interactions is required for them to prevail. Such interactions require a large-scale context provided by hydrophobic groups brought to proximity to displace surrounding water [4]. This intramolecular 'wrapping' stabilizes and enhances the electrostatics. Not surprisingly, defective intramolecular wrapping has been linked to protein–protein associations [5].

The synergism between interactions and their wrapping required for water exclusion [5, 6] points to the possible origin of cooperativity. In the few cases investigated, this synergism also accounts for the two-state kinetics [4, 7, 8] of small single-domain folders. These observations suggest that the building constraint requiring all intramolecular electrostatic interactions to be kept dry in water must be built into models capturing the essential features of cooperativity. This is precisely the aim of this work.

Here we introduce a minimal model to deal with the evolutionary consequences of cooperativity. We aim at identifying the sequences for which cooperativity becomes advantageous to generate a sharp equilibrium ensemble [9–11]. This feature has a counterpart in energy landscape theory, which proposes that natural sequences are characterized by a large energy gap separating the native state from other conformations [12–14].

Our results help resolve the issue of whether naturally foldable sequences are scarce or simply under-reported within all *a priori* possible sequences.

2. Theoretical results

Due to computational limitations to generate conformations exhaustively, we introduce a model protein in the form of a self-avoiding random walker in a cubic lattice with beads of the H (hydrophobic) or P (polar) type. The solvent is treated implicitly by modelling local environments as dependent on the chain conformation [3, 4]. Thus, the internal energy pair contributions are rescaled according to a computational assessment of conformation-dependent environments. The effect of clusters of non-nearest-neighbour residues responsible for water exclusion from interactive pairs defines the correlated nature of the lattice and makes pairwise electrostatic interactions context dependent. The context is thus provided by the distribution of units surrounding or wrapping the interaction, and accounts for the extent of dryness of the interaction itself.

A zeroth-order or 'in-bulk' potential—itself insensitive to the evolving contexts—is defined: if the distance $d(i, j) = 1 =$ length of lattice side, the (i, j)-zeroth-order pairwise energy contribution, $U^{o}{}_{ij}$, involving noncovalently bonded units i and j, will be assumed to take one of three possible values: $U^{o}{}_{ij} = -1.0$ if both units are of the H-type, $U^{o}{}_{ij} = +0.25$ if they belong to different classes and $U^{o}{}_{ij} = -0.60$ if both are of the P-type. The choice of parameters is justified *a posteriori*, but stems from the assumption that fully solvent-screened electrostatics are commensurate with thermal fluctuations ($\sim RT$, the physical dimensions

have been consistently dropped in the model) [4, 5], while the hydrophobic effect is initially dominant but decreases as the hydrophobes involved are less exposed to the solvent [1–4].

In higher order approximations, the P–P contribution should be enhanced according to the decrease in the local dielectric coefficient brought about by water removal, itself resulting from the proximity of H-type units [3–6]. The H–H hydrophobic attraction becomes weakened due to H-type clustering around it because of the reduction in hydrophobe–water interface [1]. Finally, the H–P repulsion is enhanced since water removal due to H-type clustering prevents hydration of the P-unit, raising its self-energy.

Conformations are generated by moves of the crankshaft, corner-flipping and tail-turn types and the thermodynamic limit is investigated by a 50 000 sample enumeration of self-avoiding walks of chains of length $N = 32$.

A scaling of energetic pair contributions according to the large-scale context generated by the chain, requires introducing three-body (spatial) correlation factors ${C_k}^{i,j}$. Each such factor represents a contribution to the context where the (i, j)-interaction occurs, resulting from the spatial proximity of the hydrophobic unit k, a unit not covalently bonded to either i or j. These correlation factors are essential to properly model the cooperative effects in line with the recent observations described in the introduction and suggesting that only well-wrapped pairwise interactions prevail along the folding process [4]. Generically, the index k denotes a wrapping unit, while i and j denote interactive units. The context of the (i, j)-interaction is then framed by the k-units.

In consonance with the building constraint described in the introduction, the cumulative effect of spatial correlations and its bearing on the local environment of the (i, j)-interaction is introduced in the following way:

$$U_{ij} = {U^{\mathrm{o}}}_{ij} \times \left[\Pi_{k\in\Lambda(i,j)} {C_k}^{i,j}\right] \tag{1}$$

where U_{ij} represents the context-dependent contribution to the overall energy of the chain conformation and is associated with the (i, j)-interaction, and the k, framing the context of the (i, j)-interaction, are units belonging to the family $\Lambda(i, j)$ of hydrophobic residues neighbouring the (i, j)-contact:

$$\Lambda(i, j) = \{k \text{ of H-type: } 0 < d(i, k) < r^* \text{ and } 0 < d(j, k) < r^*;\ 3^{1/2} < r^* < 2\}. \tag{2}$$

For the cubic lattice, each such family is made up of at most 16 units. The results are sensitive to the proximity parameter r^*. A more restrictive condition of proximity ($r^* < 3^{1/2}$) removes the evolutionary edge of cooperative folders, as described below.

The factor ${C_k}^{i,j}$ represents a three-body correlation where a Coulombic (i, j)-interaction is strengthened due to the proximity of the hydrophobic residue k, while a hydrophobic (i, j)-interaction is weakened. Thus, the correlation represents a realistic feature: as a hydrophobe approaches an electrostatically interactive pair, it removes surrounding water molecules, reducing the electrostatic screening, thereby enhancing the interaction [3–6]. On the other hand, a hydrophobic interaction is weakened as the hydrophobe–water interface is reduced by the presence of a third hydrophobe [1–4].

Thus, correlations can be quantified by

$${C_k}^{i,j} = [1 + \Delta \times h(r^* - d(i, k)) \times h(r^* - d(j, k))]^{g(i,j)}. \tag{3}$$

Here $g(i, j) = -1$ if $U_{\mathrm{o},ij} = -1.00$, and $g(i, j) = +1$ if $U_{\mathrm{o}}(i, j) = -0.60$ or $+0.25$; and $h(x)$ denotes the Heaviside function ($h(x) = 0$ if $x < 0$, and $h(x) = 1$ if $x > 0$). Thus, a hydrophobic residue k contributes to the correlation if and only if it is neighbouring both units i and j ($d(i, k) < r^*, d(j, k) < r^*$). The parameter $\Delta > 0$ measures the sensitivity of the pairwise interaction to the local solvent environment defined by the extent of hydrophobic burial.

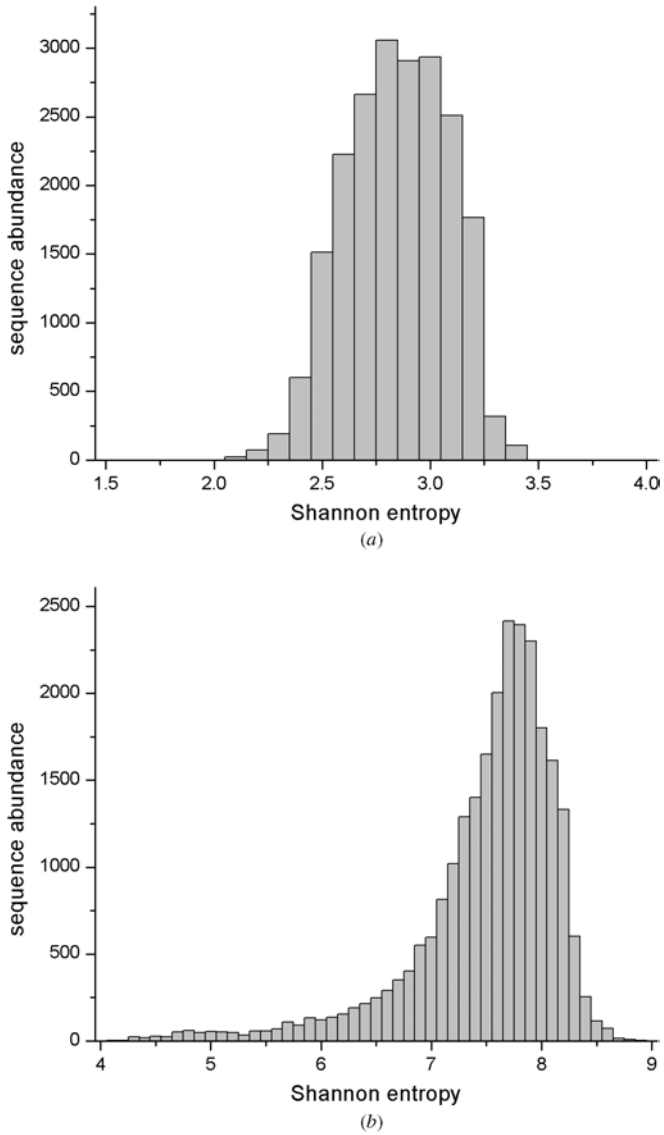

Figure 1. Abundance distribution from a sample of 22 000 H–P sequences of length $N = 32$ grouped according to the Shannon entropy of the equilibrium conformation ensemble for the uncorrelated model with (*a*) $\Delta = 0$ and (*b*) $\Delta = 0.5$ at $T = 303$ K.

The sharpness of the equilibrium conformation ensemble can be measured by the Shannon entropy $S_\Delta = -\sum_E P_\Delta(E) \ln P_\Delta(E) = -\langle \ln P_\Delta(E) \rangle_E$, where $\langle \cdot \rangle_E$ denotes the expected value with respect to the Boltzmann probability distribution, $P_\Delta(E)$, of energy levels. Thus, $P_\Delta(E) = \omega_\Delta(E) \exp(-E/RT)/Z_\Delta(T)$, where $\omega_\Delta(E)$ is the E-level degeneracy, and $Z_\Delta(T) = \sum_{E'} \omega_\Delta(E') \exp(-E'/RT)$ is the partition function at temperature T. The energy of a conformation is computed as $E = \sum_{i<j-1} {U^{\circ}}_{ij} \times \left[\Pi_{k \in \Lambda(i,j)} {C_k}^{i,j}\right]$.

More accurate measures of foldability [12–14] should probably encompass the kinetic accessibility of the states represented in equilibrium ensembles. Unfortunately, issues such as whether a cluster of conformations constitutes a set of kinetically related states, interconverting

Table 1. The eighteen H–P sequences with no periodicity or symmetry satisfying: $S_\Delta < [S_\Delta] - 2\sigma_\Delta$ for $\Delta = 0.5$ at $T = 303$ K. These sequences are the comparatively good folders representing 0.081% of the sampled sequences.

HPHHPHPPPHPHPHPPHPHHPPPHHPPHPHPH
PHPHPHPHPPHPPPPHPPPHPPHPPPPHHHP
PHPHPPPHPHPHPHPHPHPHPPPHHPPHHHPH
PPPPHHHPPHPPPHPHPHPPPHPHPPPHPPPHP
PPHPPPPHHHHPPPHHHHPPHPHPPHPHPHPP
HPHPHPHPHPHPHPHHPHPHPPPHHPHPPPH
HPHHPPPHHHPPHHPPHPHPPHPPHPHPPHPH
PPHPPPHHPHPPHPHPPHPPHPPHPHPHHPPP
HHHHPHPHPHPPHHPHPPPPPPHHHPHPHPPP
PHHHPPPHHPHHPPPPHPHHPHPPHPHPPHPH
PPPHHPPPHPHHHPHPHHPPHPHHPPPPHHHP
HPPHPPPHPPPHPPHHHPHHHPHHPPHPPPHP
PHPPHPHPHPHPHPPPHPPHPHPHPPHPHPHP
PPPHPHPHPHPPHPHPHPHPHHPPPHPHPH
PHPHPHPHPHPHPPHPHPHPHPHPHPHPPH
PHPHHPHPPPPHPHPPHPHPPHPHPHPPHPHP
PHPPPHPPHPHPPHPHPPHHPHHPPHPHPPPH
PHPPHPHPPHPHPPHPPHPHPPHPPHHHHPPP

through allowed refolding pathways, are difficult to address. Such problems require an accurate *ab initio* simulator able to perform a massive exploration in pathway space. Thus, we have chosen to focus on the thermodynamic ensemble, aware as we are that this ensemble might be sometimes diffuse while the kinetic ensemble might be sharp, as some conformations contributing to equilibrium might simply belong to disjoint ergodic components.

The abundance of $N = 32$ (H, P)-sequences distributed according to their S-value was estimated from a statistical sample of 22 000 sequences and is displayed in figures 1(a) and (b) for uncorrelated ($\Delta = 0$) and cooperative ($\Delta = 0.5$) walkers, respectively. The $\Delta = 0.5$ choice produces the largest spread in Shannon entropies for $0 < \Delta < 1$.

Because of the three-body correlations, the number of populated energy levels for the cooperative walker is vastly larger than in the uncorrelated case. This is reflected in the larger quenched average, $[S_\Delta]$ ($[\cdot]$ = average over sequences): $[S_{0.5}] = 7.7$, versus $[S_0] = 2.8$ in the uncorrelated case. The spread in the S-distribution is also much larger in the cooperative case: 4.8 versus 1.4 for $\Delta = 0$. The standard deviations in the correlated and uncorrelated case are respectively $\sigma_\Delta = 1.4$ and $\sigma_0 = 0.4$.

The larger spread for cooperative walkers imposes a severe selection pressure on sequences required to fold into a functionally competent ensemble: biological function requires a sharp ensemble of native states and a few ($\sim$0.081%) of the sampled cooperative sequences are comparatively focused ($4 < S_\Delta < 5$) in the sense that $S_\Delta < [S_\Delta] - 2\sigma_\Delta$. The nontrivial $N = 32$ sequences satisfying this inequality are given in table 1. On the other hand, the non-cooperative S-distribution is strikingly narrow and uniform: no sequence realizes the inequality $S_0 =< [S_0] - 2\sigma_0$, or $S_0 < 2$, signalling an impossibility of singling out functionally relevant sequences. This essential qualitative difference between both models is strongly dependent on the definition of proximity in the lattice. Thus, regardless of the parameter choice, for $r^* < 3^{1/2}$, we invariably get {sequences: $S_\Delta < [S_\Delta] - 2\sigma_\Delta\} = \emptyset$.

This fundamental distinction between cooperative and uncorrelated models extends to the $\Delta \to 0^+$ limit, as shown by the discontinuities:

$$7.33 \sim \lim_{\Delta \to 0^+} [S_\Delta] \neq [S_0] = 2.8 \tag{4}$$

$$1.18 \sim \lim_{\Delta \to 0^+} \sigma_\Delta \neq \sigma_0 = 0.4 \tag{5}$$

$$4.25 \sim \lim_{\Delta \to 0^+} L(S_\Delta) \neq L(S_0) = 1.40 \tag{6}$$

where $L(\cdot)$ denotes the range in S-values. The limits were estimated by computing $[S]$, σ and $L(S)$ for $\Delta = 0.1, 0.001$ and $\Delta = 0.0001$. Correlations, even at the perturbative infinitesimal level ($\Delta \to 0^+$), determine the collapse of discrete energy-level populations into quasi-continuum spreads. Thus, *the uncorrelated model is not the limit case of very low cooperativity*. Strikingly, even in the $\Delta \to 0^+$ limit, we get {sequences: $S_\Delta < [S_\Delta] - 2\sigma_\Delta\} \neq \emptyset$, while for the uncorrelated model we have {sequences: $S_0 < [S_0] - 2\sigma_0\} = \emptyset$.

3. Conclusions

As is known, the naturally occurring protein sequences of a given length invariably constitute a small fraction of the *a priori* sequence space. The probable reasons for this dearth are subsumed in past choices along evolutionary trends. This fact was rigorously addressed here from a statistical mechanics perspective: we examined the statistics on the sharpness of equilibrium folding ensembles across sequences. We concluded that cooperativity is essential to account for the selection pressure that singles out a small portion (0.081%) of all *a priori* sequences. Foldable sequences might be under-reported as yet, but they are also scarce as this work reveals. This is so because only a few sequences are able to effectively exclude the solvent from the pairwise interactions to warrant the integrity of the structure in solution.

This work emphasizes the need to come to grips with the properties of protein structure that determine cooperativity, and specifically addresses the issue: what sort of building constraint makes folding cooperative? Recent research [3–6] has revealed that intramolecular electrostatics must remain 'dry in water' to guarantee the prevalence of soluble protein structure. This constraint is shown to provide a structural basis for cooperativity and shown to exert a stringent selection pressure.

Acknowledgments

Ariel Fernández gratefully thanks the Indiana Genomics Initiative (INGEN) for financial support and the Eli Lilly Corporation for an unrestricted grant.

References

[1] Baldwin R L and Rose G D 1999 *Trends Biochem. Sci.* **24** 77
[2] Vendruscolo M and Domany E 1998 *J. Chem. Phys.* **109** 11101
[3] Fernández A 2001 *J. Chem. Phys.* **114** 2489
[4] Fernández A, Sosnick T R and Colubri A 2002 *J. Mol. Biol.* **321** 659
[5] Fernández A and Scheraga H A 2003 *Proc. Natl Acad. Sci. USA* **100** 113
[6] Fernández A and Berry R S 2003 *Proc. Natl Acad. Sci. USA* **100** 2391
[7] Krantz B A, Mayne L, Rumbley J, Englander S W and Sosnick T R 2002 *J. Mol. Biol.* **324** 359
[8] Baldwin R L 2002 *Science* **295** 1657
[9] Chan H S and Dill K A 1993 *J. Chem. Phys.* **99** 2116
[10] Pokarowski P, Kolinski A and Skolnick J 2003 *Biophys. J.* **84** 1518
[11] Dima R I and Thirumalai D 2002 *Protein Sci.* **11** 1036
[12] Onuchic J N, Luthey-Schulten Z and Wolynes P G 1997 *Annu. Rev. Phys. Chem.* **48** 545
[13] Jewet A, Pande V S and Plaxco K 2003 *J. Mol. Biol.* **326** 247
[14] Klimov D K and Thirumalai D 1996 *Phys. Rev. Lett.* **76** 4070