Folding minimal sequences: the lower bound for sequence complexity of globular proteins

Pedro Romero[a], Zoran Obradovic[a], A.K. Dunker[b,*]

[a]*School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-4660, USA*
[b]*School of Molecular Biosciences, Washington State University, Pullman, WA 99164-4660, USA*

Received 21 September 1999; received in revised form 3 November 1999

Edited by Thomas L. James

Abstract Alphabet size and informational entropy, two formal measures of sequence complexity, are herein applied to two prior studies on the folding of minimal proteins. These measures show a designed four-helix bundle to be unlike its natural counterparts but rather more like a coiled-coil dimer. Segments from a simplified sarc homology 3 domain and more than 2 000 000 segments from globular proteins both have lower bounds for alphabet size of 10 and for entropy near 2.9. These values are therefore suggested to be necessary and sufficient for folding into globular proteins having both rigid side chain packing and biological function.

Key words: Protein folding; Minimal sequence; Complexity; Alphabet size; Entropy

1. Introduction

One of the most interesting approaches for the understanding of protein folding has been the investigation of sequences of reduced amino acid alphabets [1–5]. Although use of formal complexity measures would facilitate meaningful comparisons among simplified-sequence experiments and between these and natural proteins, such measures have yet to be used for these purposes.

Here we apply formal complexity measures to the study of two reduced-alphabet sequences: a simplified sarc homology 3 (SH3) domain [3] and a designed helical protein, DHP$_1$ [5]. Simplified SH3 domains were selected by their biological binding function and characterized by circular dichroism and other methods as folding into a structure similar to that found in their wild-type counterparts [3]. Recent NMR experiments demonstrate an SH3-like fold for one of these simplified proteins (Baker and Yi, personal communication). The DHP$_1$ sequence folded into a four-helix bundle with a well-structured hydrophobic core using just seven amino acids in the entire simplified protein [5].

In this study we used two different formal measures of complexity, alphabet size and informational entropy, to compare the simplified DHP$_1$ and SH3 sequences with selected proteins of known structure and with selected protein databases. The comparisons made possible by these formal measures suggest a complexity lower bound for the folding of globular proteins into domains with biological function.

2. Materials and methods

2.1. Databases

In order to compare simplified sequences to actual proteins, several protein groups and databases were assembled, as shown in Table 1. Swiss-Prot [6] and NRL-3D [7] are sequence databases. The fibrous sequences (silk, collagen, and coiled coils) were compiled from Swiss-Prot. The SH3 sequences were found in Swiss-Prot by key word searches. The four-helix bundles are all from the protein data bank (PDB) [8], again found by key word searches. The amino acid sequences of these four-helix bundles were then acquired from Swiss-Prot.

Globular and fibrous proteins were previously shown to be distinguishable from each other using entropy values over windows of 45 consecutive residues [9,10]. To further test these results, we used NRL-3D, which is an ordered-protein subset of PDB, rather than PDB itself as used by Wootton and co-workers because disordered regions are excluded from NRL-3D.

Disorder is a common element of native protein structure [11–14]. Although disorder and complexity are in the main uncorrelated, a subset of disordered regions do indeed have extremely low complexity. This can be seen by the very low complexities of the most likely to be disordered sequences presented in [12]. Thus, NRL-3D provides a better representation of folded globular proteins than does PDB because the disordered regions are excluded from the former. In addition, all fibrous sequences (coiled coils, collagens and silk fibroins) were removed from the NRL-3D database (241 sequences total), along with nine other low-complexity, non-globular sequences. This modified NRL-3D provides a database of sequences representing well-ordered globular proteins.

2.2. Sequence entropy

From Shannon's information theory [15] the entropy, K, for a window of w consecutive amino acids, is computed as

$$K = -\sum_{i=1}^{N}\frac{n_i}{w}\left(\log_2\frac{n_i}{w}\right) = -\sum_{i=1}^{N} f_i \log_2 f_i \quad (1)$$

where N represents the number of characters in the chosen alphabet and n_i is the number of times the character i appears in the window, while f_i corresponds to the fraction of amino acid i over the window. Base-2 logarithm is used and ($0 \log_2 0$) is defined to be zero.

2.3. Alphabet size

Alphabet size, A, is defined in this study as the number of different characters (residues) in any given sequence window, w_j, and is calculated as

$$A(w_j) = \sum_{i=1}^{N} \delta_i \text{ where } \delta_i = \begin{cases} 1, \text{ if residue } i \text{ is present in window } j \\ 0, \text{ otherwise} \end{cases} \quad (2)$$

where N is the number of characters in the alphabet.

2.4. Data analysis

Complexity comparisons were by means of histograms. These were

*Corresponding author. Fax: (1)-509-335 9688.
E-mail: dunker@mail.wsu.edu

Abbreviations: SH3, sarc homology 3; DHP$_1$, designed helical protein 1

0014-5793/99/$20.00 © 1999 Federation of European Biochemical Societies. All rights reserved.
PII: S0014-5793(99)01557-4

Table 1
Proteins used in this study

Group or protein	Protein descriptions or IDs
Protein sequence databases	
Modified NRL-3D	Whole database minus fibrous and low-complexity, non-globular sequences.
Swiss-Prot	Whole database
Fibrous regions	
Coiled coils	*Swiss-Prot:* mysb_human and pgca_human; *PDB*: 2tma, 1aq5, 1avy, 1svf, 1tn3, 1vdf, 2ara, 2aj3, 1dkg, 1fos, and 1ysa
Collagen	*Swiss-Prot*: ca21_human
Silk fibroins	*Swiss-Prot:* fboh_bommo, fbol_bommo, spd1_arabi, spd1_nepcl and spd2_nepcl.
Study examples	
SH3 domains (wild-types)	*Swiss-Prot:* fyn_xiphe, src_avis2, src_avisr, src_aviss, src_avist, src_chick, src_human, src_rsvh1, src_rsvp, src1_xenla, src2_xenla, srcn_mouse, src_rsvsr, yes_avisy, yes_canfa, yes_chick, yes_human, yes_mouse, yes_xenla and yrk_chick.
SH3 domains (simplified)	Two sequences constructed in [3]
Four-helix bundles (wild-types)	*Swiss-Prot*: hemm_nerdi, hemm_thezo, hem1_phago, rop_ecoli, il4_human, arcb_ecoli, pol_hv2ro, pab_pepma
DHP_1 (designed four-helix bundle)	*PDB*: 4HB1

constructed by calculating the complexity measures A and K over a sliding windows of fixed size, w, with 20 equal bins for A and 50 for K.

For $w \geq 20$, the domain of A is simply 1 to 20, where the lower bound corresponds to a single character (amino acid) in the entire window and the upper bound is simply 20, the total number of different possible characters (amino acids) in the window. For any sized alphabet and window, the domain of K has a lower bound of 0 (one character repeated w times). For a 20 character alphabet with $w \geq 20$, the domain of K has an upper bound of $\sim$4.32 (all 20 characters, randomly distributed).

To compare the complexities of different sequences, the choice of w is important. If w is too small, statistical fluctuations impede comparisons. If w is too large, complexities from different sequences within a particular class of structures tend to converge to the same value. Also, for the analysis of segments within proteins, the resolution of different regions becomes poorer as w increases.

From experimentation on protein sequence databases, Wootton [16] found $w = 45$ to be a good overall compromise. We re-examined this issue (data not shown). For window sizes smaller than about 40, the entropy distributions of NRL-3D showed significant spikes at particular values. The complexity corresponding to the mode of these distributions increased markedly as the window size increased over the 10–40 range, just as expected, but in a non-smooth manner due to the irregularities of the underlying distributions. The distributions became essentially smooth, and length-dependence of the mode of the distributions flattened into a local plateau near $w = 45$. Thus, our analysis agrees the previous studies that $w = 45$ provides a good starting point for the comparison of complexities of different sequences.

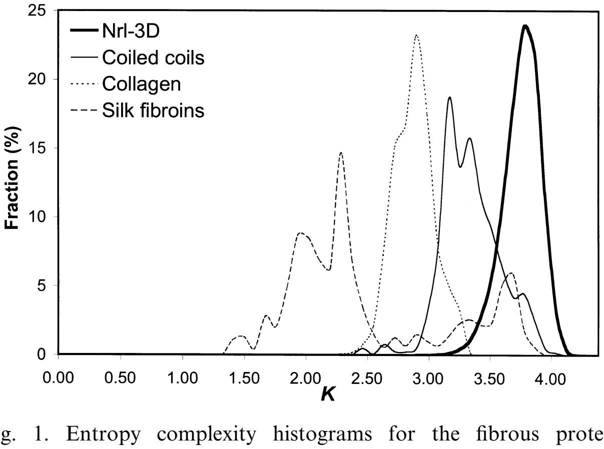

Fig. 1. Entropy complexity histograms for the fibrous protein groups from Table 1 and the NRL-3D protein database. The entropy values were calculated as described in Section 2.2 and binned as indicated in Section 2.4 for several of the databases listed in Table 1.

3. Results and discussion

3.1. Distinguishing globular and non-globular proteins

Determination of entropy (K) values for the various databases show that this measure shifts to lower values in the order globular > coiled coils > collagen > silk (Fig. 1). By considering different types of fibrous proteins, these results extend the previous findings of Wootton and co-workers [9] and at the same time further demonstrate the utility of entropy as a method for comparing sequence types.

The reason for the roughly bimodal complexity distribution of the silk sequences is unclear. Perhaps silk fibroins contain globular domains that were not properly excised in these studies. Further investigations are needed on this point.

3.2. Comparing entropy and alphabet size

Experimentalists have used alphabet size rather than Shannon's entropy to characterize their reduced-alphabet folding experiments [3–5]. Alphabet size has the advantage of being more intuitive.

Using a window size of 45 as discussed above and applying this measure to silk, collagen, coiled coils and the modified NRL-3D, we find the alphabet size (A) to yield results similar

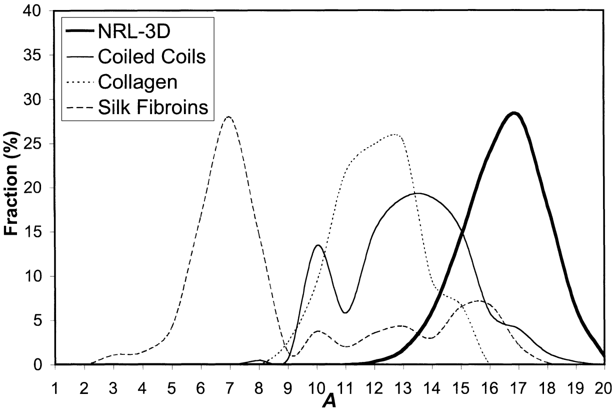

Fig. 2. Alphabet size complexity histograms for the fibrous protein groups from Table 1 and the NRL-3D protein database. The alphabet size values were calculated as described in Section 2.3 and binned as indicated in Section 2.4 for several of the databases indicated in Table 1.

Table 2
Complexity distributions for the proteins and groups used in this study

Protein, group or database	Average complexity measure					
	Alphabet size (A)	A Domain		Entropy (K)	K Domain	
		min	max		min	max
Swiss-Prot	16.1 ± 1.8	1	20	3.76 ± 0.24	0.00	4.32
NRL-3D	16.5 ± 1.4	10	20	3.81 ± 0.16	2.90	4.24
Coiled coils	13.2 ± 2.1	8	19	3.38 ± 0.24	2.45	4.06
Collagen	12.2 ± 1.4	9	15	2.92 ± 0.16	2.36	3.35
Silk fibroins	9.1 ± 3.7	3	18	2.54 ± 0.65	1.44	3.98
Wild-type SH3 domains	17.1 ± 0.8	15	18	3.93 ± 0.06	3.74	4.07
Simplified SH3 domains	12.3 ± 0.9	10	13	3.28 ± 0.09	3.05	3.40
Wild-type helix bundles	16.6 ± 1.2	13	19	3.80 ± 0.12	3.38	4.08
DHP_1	6.0 ± 0.2	6	7	2.45 ± 0.08	2.33	2.56

to those of the entropy, K, although collagen and coiled coils are very poorly separated (Fig. 2). Thus, K is clearly superior to A in distinguishing the fibrous structural classes, which probably results from K's sensitivity to the distribution of residues within a window.

3.3. Comparing complexities

Table 2 gives the average complexity values, the standard deviations, and the ranges for all of proteins and groups in this study. Fig. 3 reproduces this data, in reverse order, to facilitate comparisons.

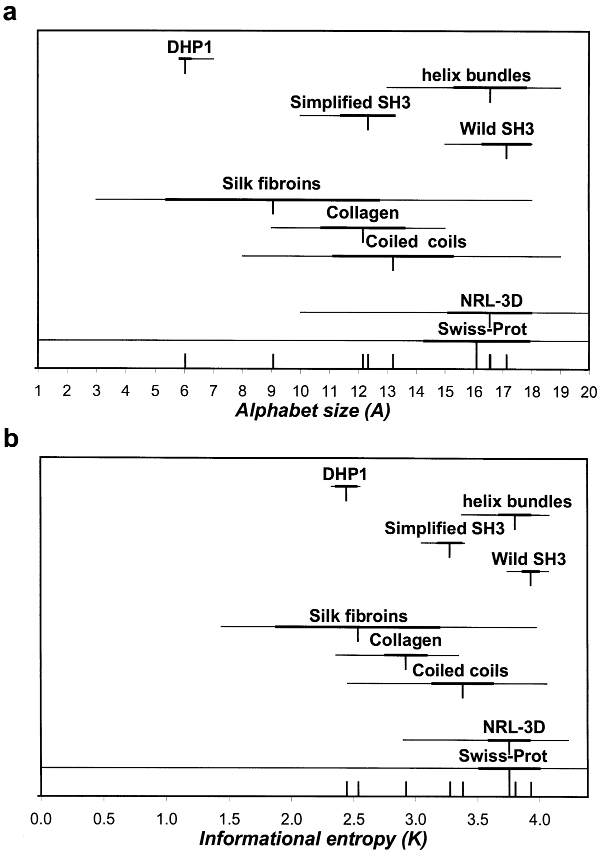

Fig. 3. Alphabet size (a) and entropy (b) distribution diagrams for all proteins, groups and databases used in this study. The average values are indicated by vertical lines, which are reproduced on the x-axis to show small differences. The thick horizontal lines indicate the standard deviations. The entire complexity domains for the various sets of proteins are represented as thin horizontal lines.

Swiss-Prot and NRL-3D (Table 2 and bottom lines in Fig. 3) have similar average entropy values and alphabet sizes, but very different complexity value domains. For windows of 45, the sequences in Swiss-Prot span the entire domains for both measures of complexity. That is, Swiss-Prot contains sequences with 45 amino acid segments having just one amino acid, corresponding to an alphabet size of one and a complexity of zero. Swiss-Prot also has 45 amino acid segments with random arrangements of all 20 amino acids, corresponding to an alphabet size of 20 and an entropy of 4.32. On the other hand, the NRL-3D distributions exhibit limits. For alphabet size, NRL-3D has a lower bound of 10 and an upper bound of 20, the entire alphabet. For entropy, the lower limit for NRL-3D is 2.90 and the upper limit is 4.24, close to, but not quite reaching a random distribution of amino acids within the window.

3.4. Comparing DHP_1 with natural proteins

Our collection of native four-helix bundles, which may or may not be representative of all four-helix bundles, have complexities with averages and standard deviations that are very similar to those of Swiss-Prot and NRL-3D, both with respect to alphabet size and entropy. Compared to the folded sequences in NRL-3D, the four-helix bundle distributions have narrower domains by both measures. However, given the similar sizes of the standard deviations, the narrower domains for the A and K values are very likely due to the smaller sample size.

While native four-helix bundles have sequence complexities like those of other globular proteins, DHP_1 has complexities unlike those of globular proteins but more like the lower extreme of coiled coils (compare entropy values in Table 2 and Fig. 3). This kinship extends to the structural level, for the helix crossing angles in DHP_1 are essentially the same as those in the leucine zipper coiled coil [5]. Since the initial design of simplified helical bundles [1] was derived from stereochemical considerations [17] extended from coiled coils [18], it is not surprising that these highly simplified helical bundles are apparently more akin to a coiled-coil dimer than to native four-helix proteins.

Although coiled coils often have sequences that are simpler than those of globular proteins, it is also clear from Table 2 and Fig. 3 that there is a great deal of overlap between their complexity distributions. In addition there is the complication that several of the highly simplified bundles constructed to date have flexible rather than rigid side chain packing [3]. Thus, there may be no clear division between globular helical bundles on one hand, and coiled coil-like bundles on the other, but rigid versus non-rigid side chain packing is a very useful experimental distinction.

3.5. Comparing the simplified SH3 domains with natural proteins

Wild-type SH3 domains appear to have a higher complexity, a narrower standard deviation, and a narrower complexity domain compared to the folded proteins in NRL-3D, by both the alphabet and entropy measures (Table 2 and Fig. 3). The significance of these findings is unclear, but it is perhaps worth noting that the SH3 domain not only folds into a compact structure but also has a specific binding function to a sequence of amino acids.

The selection protocols by Riddle et al. [3] led to a very substantial reduction in complexity for this domain, from 17.1 to 12.3 (alphabet size) or from 3.93 to 3.28 (entropy), while preserving function. These reduced values are substantially outside the complexity value domains established by the standard deviations for the NRL-3D and Swiss-Prot distributions. Thus, the Riddle et al. experiments accomplished a very significant reduction in sequence complexity, from values higher than typical K_s and A_s, to values below those at the standard deviations from the average.

Although the Riddle et al. experiments accomplished substantial complexity reductions, the entire domain of the complexity values lies within the complexity domain of NRL-3D. Overall, the laboratory selection experiments failed to produce a functional, globular protein that was simpler than proteins that have evolved by natural selection.

3.6. Proposed lower bound for sequence complexity

Using a multi-step phage display process, Riddle et al. [3] selected the simplified SH3 sequences from a nominal library of 3×10^{11} different molecules (calculated from tables 1 and 2 in [3]). The version of the NRL-3D database used herein contained about 2×10^{6} 45-residue segments from a collection of a few thousand proteins representing about 800 different protein families.

PDB, the parent of NRL-3D, is clearly non-representative of the proteins in nature. The 'average protein' from genomic studies is significantly different in length, composition, and secondary structure from those in PDB [19]. Also, fibrous proteins are clearly under-represented in PDB [10] as are membrane proteins. Likewise, sequences that fail to fold into ordered three-dimensional structures, which have been called natively unfolded [20], natively disordered [21] or intrinsically unstructured [14], are also under-represented [11–13]. For both fibrous and intrinsically disordered proteins, the under-representation is likely to be due in large measure to the filter imposed by the need for protein crystals.

Given that PDB is non-representative of all proteins in nature, a much more restricted question is whether the ordered parts of the proteins in PDB, that is the sequences in NRL-3D, are representative of the structured, or the structured parts, of globular proteins in nature. The roughly 800 families providing the segments in NRL-3D represent a significant fraction of the total number of intrinsically ordered protein families that are likely to exist [22]. Furthermore, a more recent attempt to estimate the number of protein folds and superfamilies suggests specifically that the proteins in PDB are representative. This conclusion was based on the observation that, if the protein structure data are re-evaluated over time using a rigorous statistical analysis, the same number of folds is estimated repeatedly. In contrast, if PDB were non-representative, the estimated number would be expected to increase over time as unrepresented members of structure space are added to the set [23]. Thus, with the usual caveats due to the uncertainties regarding the assumptions that underlie such types of studies, the 2×10^{6} 45-residue segments in NRL-3D are likely to be representative of all such ordered segments of globular proteins that have evolved by natural selection.

The complexity lower bounds for 45-residue segments obtained by laboratory selection and by the afore discussed sampling of natural selection are compared in Table 2 and Fig. 3. An exact coincidence, 10, is observed for alphabet size. A near coincidence, 2.90 (nature) compared to 3.05 (laboratory), is observed for the entropy lower bounds. Very short-term, but highly biased and intensive laboratory selection experiments evidently achieved almost the same lower bound for complexity as did nature, operating over about 3.5 billion years of evolution by natural selection.

The near coincidence of the alphabet and entropy lower bounds presented here argue that these values characterize the sequence complexity that is both necessary and sufficient for the formation of a globular protein having both rigid side chain packing and biological function. There is a local complexity requirement for binding pocket formation, a global complexity requirement for rigid packing by filling nooks and crannies, and a global complexity requirement for surface characteristics leading to solubility. We speculate that the need to simultaneously meet the ligand binding requirements, the side chain packing requirements and the surface solubility requirements combine to determine the observed lower bounds for K and A.

Acknowledgements: Jose Onuchic shared insights that led to this work. Celeste Brown, Xiohong Li and Ethan Garner are thanked for helpful discussions. Support from NSF research Grant CSE-IIS-9711532 to Z.O. and A.K.D. is gratefully acknowledged.

References

[1] DeGrado, W.F., Wasserman, Z.R. and Lear, J.D. (1989) Science 243, 622–628.

[2] Hecht, M.H., Richardson, J.S., Richardson, D.C. and Ogden, R.C. (1990) Science 249, 884–891.

[3] Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q. and Baker, D. (1997) Nat. Struct. Biol. 4, 805–809.

[4] Plaxco, K.W., Riddle, D.S., Grantcharova, V. and Baker, D. (1998) Curr. Opin. Struct. Biol. 8, 80–85.

[5] Schafmeister, C.E., LaPorte, S.L., Miercke, L.J. and Stroud, R.M. (1997) Nat. Struct. Biol. 4, 1039–1046.

[6] Bairoch, A. and Apweiler, R. (1999) Nucleic Acids Res. 27, 49–54.

[7] Pattabiraman, N., Namboodiri, K., Lowrey, A. and Gaber, B.P. (1990) Protein Seq. Data Anal. 3, 387–405.

[8] Abola, E.E., Berstein, F.C., Bryant, S.H., Koetzle, T.F. and Weng, J. (1987) Crystallographic Database-Information Content, Software Systems. Scientific Application, 107–132.

[9] Wootton, J.C. (1994) Curr. Opin. Struct. Biol. 4, 413–421.

[10] Wootton, J.C. and Federhen, S. (1996) Methods Enzymol. 266, 554–571.

[11] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Dunker, A.K. (1997) Proc. I.E.E.E. International Conference on Neural Networks 1, 90–95.

[12] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E.,

Guilliot, S., Garner, E. and Dunker, A.K. (1998) Pac. Symp. Biocomput. 3, 435–446.

[13] Romero, P.Z., Obradovic, C. and Dunker, A.K. (1998) Artif. Intell. Rev., in press.

[14] Wright, P.E. and Dyson, H.J. (1999) J. Mol. Biol. 293, 321–331.

[15] Shannon, C.E. (1948) Bell Syst. Tech. J., pp. 379–423, 623–656.

[16] Wootton, J.C. (1994) Comput. Chem. 18, 269–285.

[17] Dunker, A.K. and Zaleske, D.J. (1977) Biochem. J. 163, 45–57.

[18] Crick, F.H. (1953) Acta Cryst. 6, 689–697.

[19] Gerstein, M. (1998) Fold. Design 3, 497–512.

[20] Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A. and Lansbury Jr., P.T. (1996) Biochemistry 35, 13709–13715.

[21] Garner, E., Cannon, P., Romero, P., Obradovic, Z. and Dunker, A. (1998) Genome Inform. 9, 201–214.

[22] Chothia, C. (1992) Nature 357, 543–544.

[23] Wang, Z.X. (1998) Protein Eng. 11, 621–626.