doi:10.1016/S0022-2836(02)00942-7 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 323, 297–307

JMB

Small Libraries of Protein Fragments Model Native Protein Structures Accurately

Rachel Kolodny[1,2]*, Patrice Koehl[1], Leonidas Guibas[2] and Michael Levitt[1]

[1]*Department of Structural Biology, Stanford University Medical School Fairchild Building Stanford, CA 94305, USA*

[2]*Department of Computer Science, Stanford University Gates Building, Stanford CA 94305, USA*

**Corresponding author*

Prediction of protein structure depends on the accuracy and complexity of the models used. Here, we represent the polypeptide chain by a sequence of rigid fragments that are concatenated without any degrees of freedom. Fragments chosen from a library of representative fragments are fit to the native structure using a greedy build-up method. This gives a one-dimensional representation of native protein three-dimensional structure whose quality depends on the nature of the library. We use a novel clustering method to construct libraries that differ in the fragment length (four to seven residues) and number of representative fragments they contain (25–300). Each library is characterized by the quality of fit (accuracy) and the number of allowed states per residue (complexity). We find that the accuracy depends on the complexity and varies from 2.9 Å for a 2.7-state model on the basis of fragments of length 7–0.76 Å for a 15-state model on the basis of fragments of length 5. Our goal is to find representations that are both accurate and economical (low complexity). The models defined here are substantially better in this regard: with ten states per residue we approximate native protein structure to 1 Å compared to over 20 states per residue needed previously.

For the same complexity, we find that longer fragments provide better fits. Unfortunately, libraries of longer fragments must be much larger (for ten states per residue, a seven-residue library is 100 times larger than a five-residue library). As the number of known protein native structures increases, it will be possible to construct larger libraries to better exploit this correlation between neighboring residues. Our fragment libraries, which offer a wide range of optimal fragments suited to different accuracies of fit, may prove to be useful for generating better decoy sets for *ab initio* protein folding and for generating accurate loop conformations in homology modeling.

Keywords: protein representations; discrete models

Introduction

The three-dimensional structure of proteins has been a subject of intense study for several decades. A common way to simplify these complex structures is to consider restrictions on the local main-chain conformation. Almost 50 years ago, Corey & Pauling described the two common types of local secondary structure, the α-helix and the β-sheet.[1] Ten years later, Ramachandran ascribed the limited (ϕ, ψ) torsion angles of each residue due to the interactions of the side-chain with its backbone.[2] In 1986, Jones & Thirup discovered that almost all regions of the protein backbone are comprised of repeating canonical structures.[3] These regions, up to ten residues long, provided an efficient method for interpreting electron density maps. Unger *et al.* followed by classifying peptide backbone units four to ten residues long, into a collection of fragments.[4] These building block units constitute an intermediate level of protein structure representation between single residues and secondary structure. Since then, many studies have investigated the classification of protein fragments and, in particular, the classification of loop structures.[5–8]

Abbreviation used: cRMS, coordinate root mean square.

E-mail address of the corresponding author: rachel.kolodny@stanford.edu

Table 1. PDB identifiers of the proteins used

A. *Test set*[10,a]

d1iiba_	d1gsoa3	D1burs_	d1csh__	d1dfma_	d1pina_	d1c1ka_	d1bsma1	d2pth__	d1lkka_
d1mtyg_	d1kpta_	d3cla__	d1php__	d1ra9__	d3pte__	d1a4ia2	d1bsma2	d1dcs__	d1mfma_
d1pcfa_	d3btoa1	d1ako__	d1aop_3	d1bfd_2	d1krn__	d2cba__	d2end__	d7rsa__	d2erl__
d2gsta1	d1b2pa_	d1tx4a_	d1mrj__	d3ezma_	d1lam_1	d1poa__	d1qhva_	d1ifc__	d1cxqa_
d1bm8__	d1cjca2d1	d1rzl__	d1qqqa_	d1rie__	d1kpf__	d1mla_1	d2eng__	d1mroa1	d1aho__
d1mjha_	ush_1	d1czfa_	d3grs_3	d1ptf__	d1cipa1	d1tc1a_	d3ebx__	d3chbd_	d1a6m__
d1svy__	d1utea_	d1ctf__	d1b6a_1	d1ah7__	d1b3aa_	d1yge_1	d1qaua_	d1qu9a_	d1ixh__
d1tfe__	d1pdo__	d1vns__	d1b6a_2	d8abp__	d1nox__	d1yge_2	d1qh4a1	d1d4oa_	d1cex__
d1thw__	d1vcc__	d7odca1	d2cpl__	d1b4va1	d1dpsa_	d7atja_	d1qh4a2	d1jhga_	d1byi__
d1db1a_	d1pda_2	d1d3va_	d1kapp1	d1b4va2	d1qsaa1	d1utg__	d1rhs__	d1vfya_	d1b0ya_
d1doza_	d1yvei1	d1qgxa_	d1ppn__	d1qh5a_	d1orc__	d1di6a_	d1bi5a1	d1mun__	d1nls__
d1aoha_	d3stda_	d1phc__	d2ilk__	d1b67a_	d1qgwa_	d1sgpi_	d256ba_	d1a7s__	d7a3ha_
d1vhh__	d2ahja_	d1fmk_3	d3cyr__	d1dcia_	d1hfel1	d1qtsa2	d1qksa2	d1swua_	d2fdn__
d1a44__	d1dhn__	d1ay7b_	d1ubpa_	d1ezm_1	d1hfes_	d1aba__	d1msi__	d1nkd__	d1bxoa_
d1fnd_1	d1qhfa_	d1b8za_	d1ubpc1	d1ezm_2	d1c3wa_	d1bgf__	d1dg6a_	d1bkra_	d3lzt__
d1fnd_2	d1ttba_	d1smd_1	d1qq5a_	d1whi__	d1qgua_	d1qfma1	d1qdda_	d1rgea_	d2pvba_
d1atza_	d1qipa_	d1a1ia1	d1moq__	d1dgfa_	d1bx4a_	d3vub__	d1aac__	d5pti__	d1rb9__
d1dmr_2	d2bbkl_	d1qs1a1	d1d7pm_	d1qrea_	d1dpta_	d3euga_	d1cy5a_	d1qj4a_	d3pyp__
d1gsoa1	d2cpga_	d1ajsa_	d1bfg__	d1cyo__	d1aie__	e1pid.1a	d2lisa_	d2igd__	d1cbn__
d1gsoa2	d1kid__	d1t1da_	d1gai__	d1g3p_1	d1byqa_	e1pid.1b	d1amm_1	d3sil__	d1gci__

B. *Training set*[b,c]

d1gci__ 0.78 1.33	d3lzt__ 0.92 1.15	d1b0ya_ 0.93 1.07	d1aho__ 0.96 1.04	d3sil__ 1.05 0.99
d1cbn__ 0.83 1.23	d1bxoa_ 0.95 1.10	d1byi__ 0.97 1.07	d1cxqa_ 1.02 1.03	d2igd__ 1.10 0.98
d3pyp__ 0.85 1.20	d2fdn__ 0.94 1.10	d1cex__ 1.00 1.07	d2erl__ 1.00 1.02	d1qj4a_ 1.10 0.94
d1rb9__ 0.92 1.17	d7a3ha_ 0.95 1.09	d1ixh__ 0.98 1.06	d1mfma_ 1.02 1.01	d5pti__ 1.00 0.92
d2pvba_ 0.91 1.15	d1nls__ 0.94 1.07	d1a6m__ 1.00 1.05	d1lkka_ 1.00 1.00	d1rgea_ 1.15 0.92
d1bkra_ 1.10 0.92	e1pid1b 1.30 0.70	d1rie__ 1.50 0.63	d1kapp1 1.64 0.59	d1qhfa_ 1.70 0.56
d1nkd__ 1.07 0.92	e1pid1a 1.30 0.70	d3ezma_ 1.50 0.63	d2cpl__ 1.63 0.59	d1dhn__ 1.65 0.56
d1swua_ 1.14 0.91	d3euga_ 1.43 0.69	d1bfd_2 1.60 0.63	d1b6a_2 1.60 0.59	d2ahja_ 1.70 0.56
d1a7s__ 1.12 0.88	d3vub__ 1.40 0.68	d1ra9__ 1.55 0.62	d1b6a_1 1.60 0.59	d3stda_ 1.65 0.56
d1mun__ 1.20 0.88	d1qfma1 1.40 0.67	d1dfma_ 1.50 0.62	d3grs_3 1.54 0.59	d1yvei1 1.65 0.56
d1vfya_ 1.15 0.85	d1bgf__ 1.45 0.67	d1a4ia2 1.50 0.62	d1qqqa_ 1.50 0.59	d1pda_2 1.76 0.56
d1jhga_ 1.30 0.83	d1aba__ 1.45 0.67	d1c1ka_ 1.45 0.62	d1mrj__ 1.60 0.59	d1vcc__ 1.60 0.56
d1d4oa_ 1.21 0.82	d1qtsa2 1.40 0.67	d1byqa_ 1.50 0.62	d1aop_3 1.60 0.59	d1pdo__ 1.70 0.56
d1qu9a_ 1.20 0.82	d1sgpi_ 1.40 0.67	d1aie__ 1.50 0.62	d1php__ 1.65 0.59	d1utea_ 1.55 0.55
d3chbd_ 1.25 0.82	d1di6a_ 1.45 0.67	d1dpta_ 1.54 0.62	d1csh__ 1.60 0.58	d1ush_1 1.73 0.55
d1mroa1 1.16 0.81	d1utg__ 1.34 0.67	d1bx4a_ 1.50 0.62	d1t1da_ 1.51 0.58	d1cjca2 1.70 0.55
d1ifc__ 1.19 0.81	d7atja_ 1.47 0.66	d1qgua_ 1.60 0.62	d1ajsa_ 1.60 0.58	d1b2pa_ 1.70 0.55
d7rsa__ 1.26 0.80	d1yge_2 1.40 0.66	d1c3wa_ 1.55 0.62	d1qs1a1 1.50 0.58	d3btoa1 1.66 0.55
d1dcs__ 1.30 0.79	d1yge_1 1.40 0.66	d1hfes_ 1.60 0.61	d1a1ia1 1.60 0.58	d1kpta_ 1.75 0.55
d2pth__ 1.20 0.79	d1tc1a_ 1.41 0.66	d1hfel1 1.60 0.61	d1smd_1 1.60 0.58	d1gsoa3 1.60 0.55
d1amm_1 1.20 0.78	d1mla_1 1.50 0.66	d1qgwa_ 1.63 0.61	d1b8za_ 1.60 0.58	d1gsoa2 1.60 0.55
d2lisa_ 1.35 0.78	d1poa__ 1.50 0.66	d1orc__ 1.54 0.61	d1ay7b_ 1.70 0.58	d1gsoa1 1.60 0.55
d1cy5a_ 1.30 0.77	d2cba__ 1.54 0.65	d1qsaa1 1.65 0.61	d1fmk_3 1.50 0.58	d1dmr_2 1.82 0.55
d1aac__ 1.31 0.77	d3pte__ 1.60 0.65	d1dpsa_ 1.60 0.61	d1phc__ 1.60 0.58	d1atza_ 1.80 0.54
d1qdda_ 1.30 0.76	d1pina_ 1.35 0.65	d1nox__ 1.59 0.61	d1qgxa_ 1.60 0.57	d1fnd_2 1.70 0.54
d1dg6a_ 1.30 0.76	d1g3p_1 1.46 0.65	d1b3aa_ 1.60 0.61	d1d3va_ 1.70 0.57	d1fnd_1 1.70 0.54
d1msi__ 1.25 0.75	d1cyo__ 1.50 0.64	d1cipa1 1.50 0.61	d7odca1 1.60 0.57	d1a44__ 1.84 0.54
d1qksa2 1.28 0.75	d1qrea_ 1.46 0.64	d1kpf__ 1.50 0.60	d1vns__ 1.66 0.57	d1vhh__ 1.70 0.54
d256ba_ 1.40 0.73	d1dgfa_ 1.50 0.64	d1lam_1 1.60 0.60	d1ctf__ 1.70 0.57	d1aoha_ 1.70 0.54
d1bi5a1 1.56 0.72	d1whi__ 1.50 0.64	d1krn__ 1.67 0.60	d1czfa_ 1.68 0.57	d1doza_ 1.80 0.54
d1rhs__ 1.36 0.72	d1ezm_2 1.50 0.64	d1gai__ 1.70 0.60	d1rzl__ 1.60 0.57	d1db1a_ 1.80 0.54
d1qh4a2 1.41 0.72	d1ezm_1 1.50 0.64	d1bfg__ 1.60 0.60	d1tx4a_ 1.65 0.57	d1thw__ 1.75 0.54
d1qh4a1 1.41 0.72	d1dcia_ 1.50 0.64	d1d7pm_ 1.50 0.60	d1ako__ 1.70 0.57	d1tfe__ 1.70 0.54
d1qaua_ 1.25 0.72	d1b67a_ 1.48 0.64	d1moq__ 1.57 0.60	d3cla__ 1.75 0.57	d1svy__ 1.75 0.54
d3ebx__ 1.40 0.71	d1qh5a_ 1.45 0.63	d1qq5a_ 1.52 0.60	d1burs_ 1.80 0.56	d1mjha_ 1.70 0.54
d2eng__ 1.50 0.71	d1b4va2 1.50 0.63	d1ubpc1 1.65 0.60	d1kid__ 1.70 0.56	d1bm8__ 1.71 0.54
d1qhva_ 1.51 0.70	d1b4va1 1.50 0.63	d1ubpa_ 1.65 0.60	d2cpga_ 1.60 0.56	d2gsta1 1.80 0.53
d2end__ 1.45 0.70	d8abp__ 1.49 0.63	d3cyr__ 1.60 0.59	d2bbkl_ 1.75 0.56	d1pcfa_ 1.74 0.53
d1bsma2 1.35 0.70	d1ah7__ 1.50 0.63	d2ilk__ 1.60 0.59	d1qipa_ 1.72 0.56	d1mtyg_ 1.70 0.53
d1bsma1 1.35 0.70	d1ptf__ 1.60 0.63	d1ppn__ 1.60 0.59	d1ttba_ 1.70 0.56	d1iiba_ 1.80 0.53

[a] Test set described by PDB name.
[b] Training set described by PDB name, structure resolution, SPACI score.
[c] Training set has 200 polypeptides with the highest SPACI scores.

Even when using the (ϕ, ψ) torsion angles as the degrees of freedom, a protein chain has an infinite number of different conformations due to continuous changes in the torsion angles. By restricting the local conformations of individual residues to a handful of states, one can discretize protein conformation so that any chain has a finite number of spatial arrangements. The utility of any discrete

model depends on the accuracy with which it models real protein conformations as well as on its complexity, the number of allowed states per residue. Rooman *et al.*[9] and Park & Levitt[10] showed that discrete models that take into account the uneven (ϕ, ψ) distribution of single residue conformations in proteins are more accurate (for a fixed complexity).

Here, we combine these two previous approaches by finding a finite set of protein fragments that can be used to construct accurate discrete conformations for any protein. We begin by following Unger *et al.*[4] and Micheletti *et al.*,[5] who used the unsupervised learning technique of clustering to identify representative fragments of protein backbone. We use a novel clustering scheme to find better libraries of fragments. These fragment libraries are used to construct discrete approximation to real protein structure. Indeed, as observed by Simon *et al.*,[11] considering only protein models constructed from valid protein fragments yields smaller structural spaces.

We carry out an extensive study with many different-sized libraries of fragments of length 4, 5, 6 and 7. The accuracy with which these discrete representations capture native structure depends on the complexity and varies from 1.9 Å for a four-state model on the basis of fragments of length 7 to 0.76 Å for a 15-state model on the basis of fragments of length 5. With discrete representations, a protein conformation is reduced to a string of symbols that define the local states (alphabets of four and 15 letters, respectively, in the above examples). These strings specify the conformation completely: all possible conformations are generated by all possible strings. Thus, discretization converts a three-dimensional structure into a one-dimensional string akin to the amino acid sequence. We find that longer fragments are more accurate, as they include more correlation than shorter fragments. However, the complexity that can be explored with the longer fragment lengths is limited severely by the relatively small number of known protein structures.

Our clustering method, known as simulated-annealing *k*-means, is likely to be useful for many clustering tasks that involve biological data with an unknown and uneven distribution of objects. The method is relatively efficient when used on very large datasets. Our fragment libraries may prove to be useful for generating better decoy sets for *ab initio* protein folding (as done by Park & Levitt for four-state models[10]), for generating accurate loop conformations in homology modeling, and for analyzing strings of conformational states that define protein structure strings in the same way that is done for strings of amino acid residues that define protein sequences.

Results

Fragments from proteins in the training sets are clustered using the simulated annealing *k*-means technique. The libraries, which are the fragments at the centroid of each cluster, are evaluated by their ability to reconstruct the protein structures in the test set proteins. We consider two criteria: (1) local-fit, which measures the coordinate root mean square (cRMS) deviation of all fragments of the target protein from the library at hand. (2) Global-fit, which measures the cRMS of the reconstructed three-dimensional structure from the entire native structure of the target. We consider fragments of length, f, varying from four to seven residues and library sizes, s (i.e. the number of clusters) varying from 10 to 300; this results in library complexities (calculated as $s^{1/(f-3)}$) that range from 3.16 to 15 states per residue. The libraries found are available on our server† and as Supplementary Material.

The Park & Levitt[10] set of proteins is used as the protein test set in this study. It includes 145 proteins of different structural motifs, varying in length from 36 to 753 residues. The use of the same test set as that used by Park & Levitt allows easier comparison with the results of their study, and offers an extension to their results regarding the complexity and accuracy of discrete approximations of protein structures. For completeness, Table 1 lists the PDB identifiers of the 145 proteins in the test set. As with the training set, we approximate the chain paths of the test set folds by the atomic coordinates of their C^{α} atoms.

Local-fit approximations

Table 2 summarizes the accuracy of the best local-fit approximations for all libraries considered in this study, while Figure 1 plots these data as a function of the complexity. We also calculated the average cRMS deviation of the best local-fit approximations of the test set proteins using the five and six-residue fragment libraries published by Micheletti *et al.*[5] Figure 1 shows that the fragments of the proteins in the test set can be described very well by any of the libraries considered: the average cRMS deviation is below 1 Å in all cases. For libraries of a fixed fragment length, the accuracy of the local-fit approximations is improved when the complexity (or the library size) is increased. This makes intuitive sense: libraries with a greater variety fit the fragments of the test set proteins better. For a library of the same complexity, the accuracy of the local-fit approximations is improved with shorter fragments. This also makes sense: shorter fragments give a better local-fit as there are fewer C^{α} atoms involved in each fragment-to-fragment comparison. Stated differently, there are six additional degrees of freedom for the rigid-body rotation and translation of each fragment. With shorter fragments, there are, therefore, more degrees of

† http://csb.stanford.edu/rachel/fragments

Table 2. Average accuracy of global and local cRMS deviations

Fragment length	Library size	Complexity (states/residue)[c]	Average[a,b] Local cRMS (Å)	Average[a,b] Global cRMS (Å)
4	4	4.00	0.39	2.23
4	6	6.00	0.35	1.64
4	7	7.00	0.33	1.48
4	8	8.00	0.32	1.39
4	10	10.00	0.30	1.12
4	12	12.00	0.28	1.01
4	14	14.00	0.26	0.92
5	10	3.16	0.57	2.57
5	20	4.47	0.47	1.85
5	30	5.48	0.43	1.59
5	40	6.32	0.40	1.41
5	50	7.07	0.39	1.28
5	60	7.75	0.37	1.20
5	80	8.94	0.35	1.06
5	100	10.00	0.34	0.99
5	150	12.25	0.31	0.86
5	225	15.00	0.29	0.76
6	40	3.42	0.65	2.30
6	60	3.91	0.59	2.02
6	70	4.12	0.58	1.92
6	80	4.31	0.56	1.87
6	100	4.64	0.54	1.72
6	200	5.85	0.48	1.41
6	300	6.69	0.45	1.26
7	50	2.66	0.85	2.89
7	100	3.16	0.76	2.41
7	150	3.50	0.72	2.16
7	200	3.76	0.68	2.04
7	250	3.98	0.66	1.91
Micheletti *et al*. fragment libraries[d]				
5	40	6.32	0.48	1.64
6	100	4.64	0.57	1.88

The libraries found are available on our server (http://csb.stanford.edu/rachel/fragments

[a] The average is taken over the approximations of the test set proteins.

[b] The data are plotted in Figures 1 and 2.

[c] The complexity, or the average number of states per residue, of a fragments library is $|L|^{1/(f-3)}$ where $|L|$ is the library size and f is the length of the fragments in the library.

[d] We calculate the Local cRMSD and Global cRMSD of the libraries published by Micheletti *et al*.[5] on the test set of proteins.

freedom to fit a protein structure of given length in the local approximation.

Global-fit approximations

Table 2 summarizes the accuracy of global-fit approximations constructed from the libraries considered in this study, while Figure 2 plots these data as a function of the library complexity. The average cRMS deviation of the global-fit approximations over the test set varies from 2.58 Å for the lowest complexity library to 0.76 Å for the highest complexity library. The inset in Figure 2 plots these data on a log scale along with linear regression lines. We compare our results to those of: (1) Park & Levitt,[10] where the test set and the complexity measure are the same, so their results are merely quoted here; and (2) Micheletti *et al*.[5] In the latter case, we use the libraries of five and six residues published on the World Wide Web,† construct global-fit approximations for the test set and calculate the average cRMS deviation between the test set and its approximations.

Figure 2 offers insight to the relationship between libraries of fixed fragment length and varying complexity, as well as the relationship between libraries of fixed complexity and varying fragment length. For a fixed fragment length, more complex libraries offer better global-fit approximations. This observation makes intuitive sense: the complexity of libraries of fixed length depends on the number of fragments in the library and libraries with greater variety will result in more accurate approximations. More surprisingly, for a fixed complexity, libraries of greater length fragments give better global-fit approximations. All the Global-RMS data from our libraries of different complexity, c, and fragment length, f, can be well fit by a single function:

$$\text{Global-RMS} = e^{(0.094f+1.373)}(\text{Complexity})^{-0.1039f-0.280}$$

or, more simply:

$$(\text{Global-RMS}) \propto (\text{Complexity})^{(-0.1039f-0.280)}$$

and:

$$(\text{Complexity}) \propto (\text{Global-RMS})^{(0.1039f+0.280)}$$

Park & Levitt[10] found that for a model that used non-optimized (ϕ,ψ) torsion angle states:

$$(\text{Global-RMS}) \propto (\text{Complexity})^{-0.5}$$

For a fragment length of 4, which is most like the (ϕ,ψ) states, the corresponding dependence is:

$$(\text{Global-RMS}) \propto (\text{Complexity})^{-0.7}$$

For longer fragments, the power becomes more negative, so that for a length of 7, the dependence is:

$$(\text{Global-RMS}) \propto (\text{Complexity})^{-1.0}$$

This more rapid fall-off of Global-RMS with Complexity for longer fragments means that the models on the basis of longer fragments can model proteins better for a given complexity.

The computer implementation of the global-fit approximation construction procedure uses a heap storing the best approximations found so far; the heap size should be selected to balance between the desire to explore a greater (polynomial) portion of approximation space and the reality of run time and memory constraints. The number of possible global-fit approximations to a target protein is exponential and therefore it is impossible to explore them all, instead only N_{keep} approximations are

† http://www.sissa.it/~michelet/prot/repset/index.html

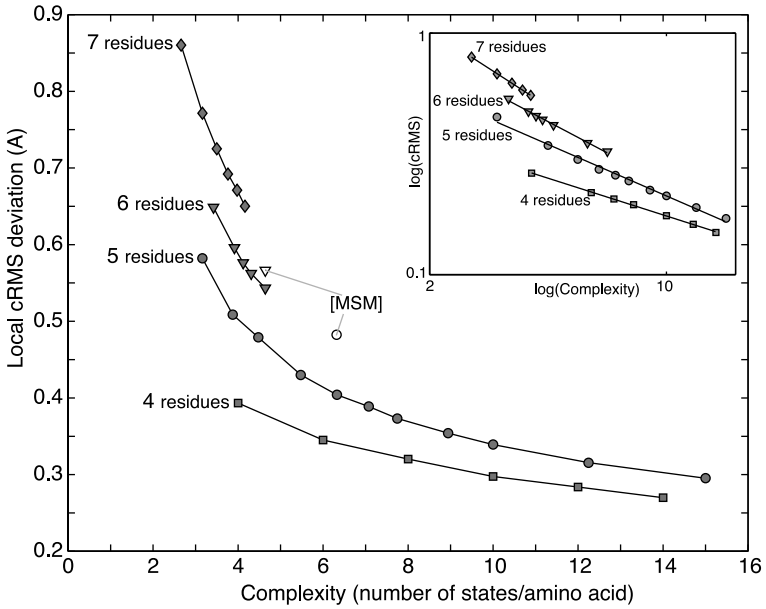

Figure 1. The average local cRMS deviation of test set proteins constructed using various libraries is plotted against the complexity of the library. The libraries vary by size and are of fragments of lengths 4 (squares), 5 (circles), 6 (triangles) and 7 (diamonds), respectively. The complexity is determined by the library size and the fragment length as $s^{1/(f-3)}$. For fixed fragment length, f, more complex libraries with more members, s, give more accurate approximations. The inset shows the same data on a log scale: the linear fit of the data is $y = -0.313x - 0.450$, $y = -0.427x - 0.103$, $y = -0.518x + 0.186$ and $y = -0.633x + 0.459$ for fragment lengths of 4, 5, 6, and 7, respectively. More generally, Local-RMS depends on library complexity and fragment length, f, as:

$$\log(\text{local RMS}) = A \log(\text{Complexity}) + B$$

where $A = -0.1051(f - 1)$ and $B = 0.3016f - 1.6358$.

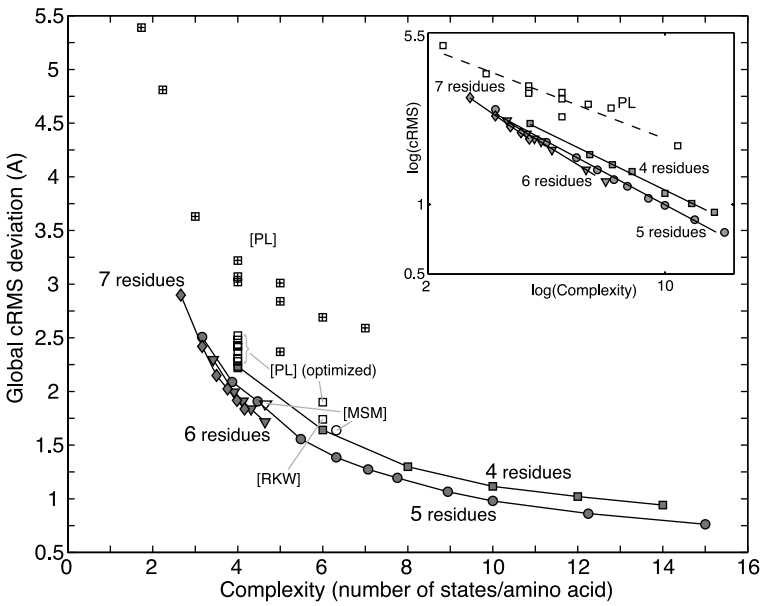

Figure 2. The average Global cRMS deviation of the test set proteins is plotted as a function of the complexity of the library used for constructing the approximations. The libraries vary in size and are of fragments of lengths 4 (squares), 5 (circles), 6 (triangles) and 7 (diamonds), respectively. The libraries compiled in this study are shown in opaque shapes, while the libraries reported by Park & Levitt[10] (libraries with four-residue fragments) and Micheletti *et al*.[5] (libraries with five and six-residue fragments) are shown in hollow shapes. The inset shows the same data in log scale: the linear fit of the data is $y = -0.712x + 1.78$, $y = -0.78x + 1.80$, $y = -0.895x + 1.93$ and $y = -1.016x + 2.05$ for fragment lengths of 4, 5, 6, and 7, respectively. More generally, the Global-RMS depends on library complexity and fragment length, f, as:

$$\log(\text{Global-RMS}) = A \log(\text{Complexity}) + B$$

where $A = -0.1039f - 0.280$ and $B = 0.094f + 1.373$.

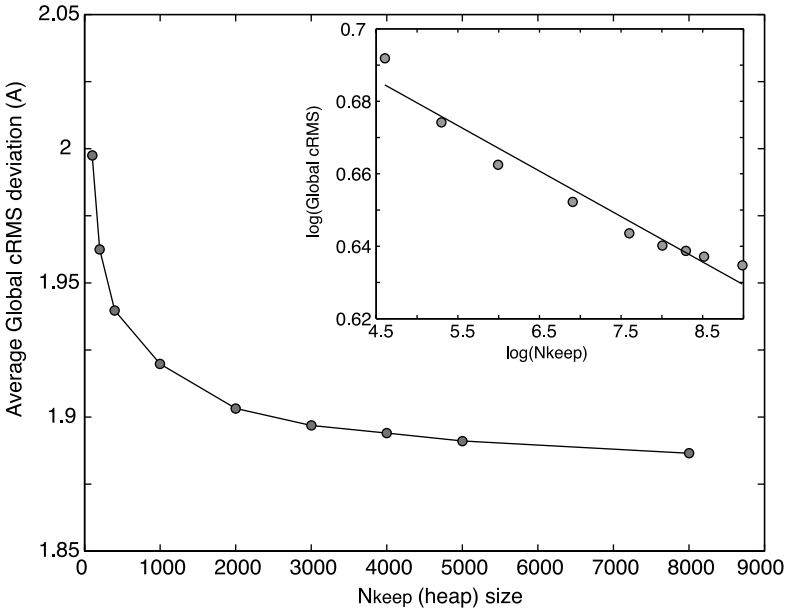

Figure 3. cRMS deviation of best approximation, averaged over the Park & Levitt[10] data set as a function of N_{keep} size. The library used to compile these data is of 20 fragments, each five amino acid residues long. The inset shows the log(Global cRMS) as a function of log(heap size). This same functional behavior is observed in other libraries, making $N_{keep} = 4000$ a reasonable choice for reconstruction from all libraries. From the log–log plot in the inset, we find:

$$\log(\text{Global-RMS}) = -0.013 \log(N_{keep}) + 0.742$$

or:

$$\text{Global-RMS} = 2.1 N_{keep}^{-0.013}.$$

considered, where N_{keep} is the heap size. The running time of the procedure is linear in this size $-O(N_{keep} n|L|)$, and maintaining the heap requires $O(N_{keep})$ memory.

Figure 3 shows a plot of the average accuracy of the best global-fit approximations found for the proteins in the test set, *versus* the heap size used in the construction procedure, for one representative library, L_5^{20} (library of 20 fragments, five residues each). As expected, better approximations are found when using a larger heap. However, the accuracy improves dramatically with increasing heap size for small heaps and remains relatively constant for larger values. Therefore, in this study we used a heap size of 4000 when searching for global-fit approximations, which is an appropriate balance between the quest for accuracy and the limitations on running time. Similar behavior was observed in all the libraries we considered.

Figure 4 shows a plot of the average cRMS deviation of the local-fit approximations from the proteins in the test set, *versus* the same measure of global-fit approximations, for all libraries considered in this study. For any particular library, the local-fit cRMS is always smaller than the corresponding global fit cRMS. This is to be expected, as the local-fit ignores the connection between adjacent fragments along the chain completely. These results show that local-fit approximations can be used to predict the accuracy of the global-fit approximations: libraries that provide accurate local-fit approximations will also provide accurate global-fit approximations. It is clear that for the same level of global fit cRMS deviation, the local cRMS deviation decreases sharply with fragment length.

Dependency on the polypeptide length

The dependency of the accuracy of the approximations on the length of the approximated protein was studied. Specifically, we considered the cRMS deviations of the best local-fit approximation and the cRMS deviation of the best global-fit approximation that we found *versus* the polypeptide length, for all proteins in the test set (data not shown). The accuracy of the local-fit approximations is independent of the chain length, while that of the global-fit approximations is only very slightly dependent on the chain length. As an example of the quality of fit we can obtain, Figure 5 shows three approximations of 1tim of various accuracies.

Discussion

Independence of test set

The training set used to compile the libraries and the test set used to evaluate them are independent. The training set for our procedures is a collection of fragments extracted from proteins with accurate structural data (on the basis of their SPACI[12] scores), while the test set is an accepted set for testing questions of this type. Although lack of overlap was not a criterion used to select the training set, there is only one protein (256b) that is in both sets. This independence of these two sets assures that the results presented do not follow from learning a specific set of proteins, but rather from several properties of protein structure.

Choice of clustering technique

The simulated annealing *k*-means clustering technique is the best we have found for clustering the fragments data set: it performs significantly better than other clustering techniques, including hierarchical (bottom up) clustering, top down clustering and standard *k*-means. Simulated annealing *k*-means surpasses the other clustering technique

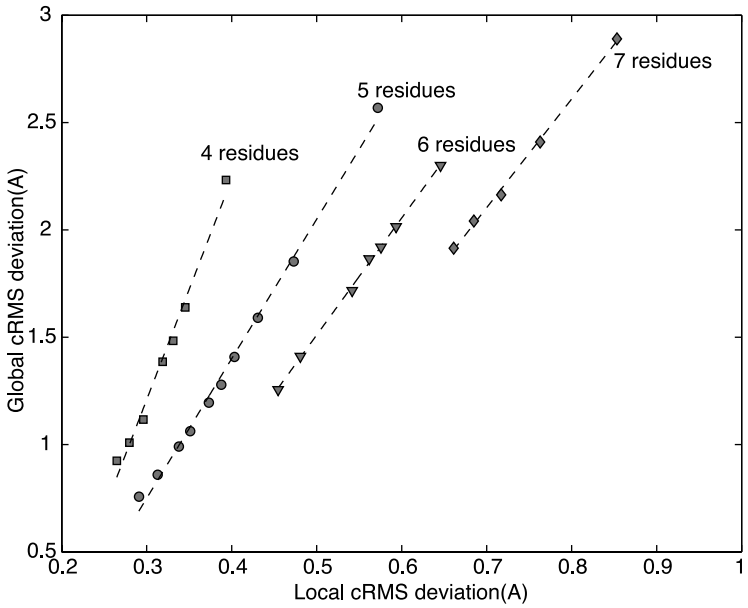

Figure 4. Global-fit accuracy as a function of local-fit accuracy. The average cRMS deviation of the global-fit approximations is plotted as a function of the average cRMS deviation for local-fit approximations for all proteins in the data set. Libraries of fragments of length 4, 5, 6, and 7 of various sizes are plotted here. The data in this Figure are the data of Figures 1 and 2, replotted for illustration. When fitting a line to the data, we have: $y = 10.277x - 1.874$, $y = 6.5x - 1.2$, $y = 5.45x - 1.21$ and $y = 5.066x - 1.443$ for fragment lengths of 4, 5, 6, and 7, respectively.

in two measures: (1) the total variance of the results; and (2) the average cRMS deviation of the local-fit approximations of the proteins in the test set built with the library compiled from the result. The first measure is a general-purpose statistical measure of any clustering result, while the second is specific for this setting. A detailed comparison of the different clustering methods is beyond the scope of this paper will be given elsewhere (our unpublished results). Simulated annealing *k*-means is more robust than the other random clustering techniques: it is less sensitive to the initial randomly picked cluster centers. The hierarchical (bottom up) clustering was the second best method for clustering these data, but it required an order of magnitude more computer time.

Local-fit approximations

Local-fit approximations are interesting, even though the resulting structures consist of disjointed fragments that can be found only by fitting a known protein backbone. In building these local approximations we seek, as studies before us,[4,5] a short list of fragments that is representative of all fragments of known proteins. Local-fit approximations capture this notion of similarity, and offer an efficient, linear time method of evaluating libraries of fragments. Comparison between the accuracy of local-fit approximations using libraries found in this study and those constructed by Micheletti *et al*.[5] indicates that the elaborate clustering scheme used here leads to better results. In addition, local-fit approximations serve as

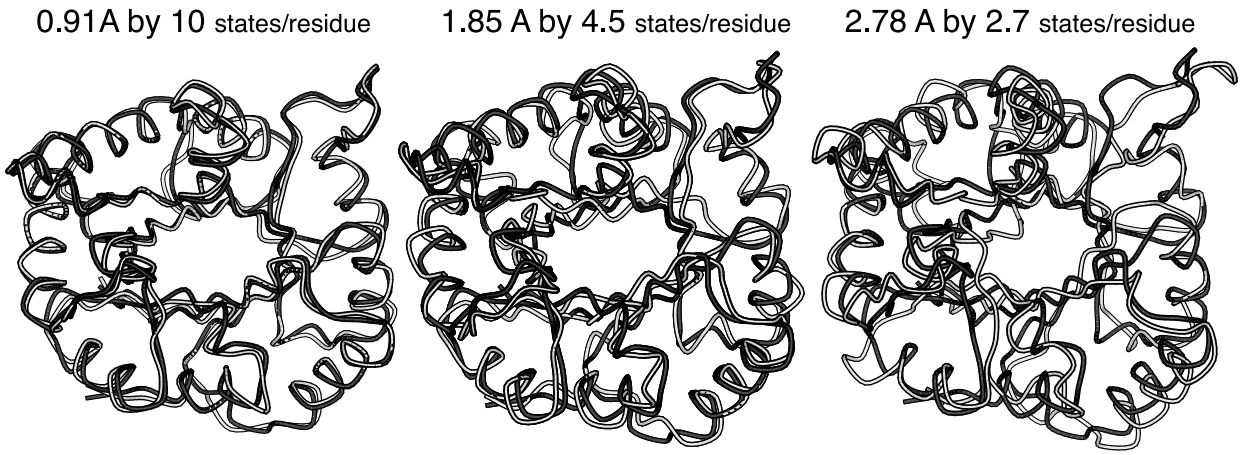

Figure 5. Three global-fit approximations to the alpha–beta barrel protein with PDB identifier 1tim. The protein is drawn[20] in black and the approximations in gray. The libraries used when modeling the protein in: (a) has 100 fragments of five residues each and achieves an overall cRMS distance of 0.9146 Å (ten states per residue); (b) has 20 fragments of five residues each and achieves an overall cRMS distance of 1.8454 Å (4.47 states per residue), and (c) has 50 fragments of seven residues each and achieves an overall cRMS distance of 2.7805 Å (2.66 states per residue). The clear improvement in global fit with increasing library complexity is apparent.

predictors to the accuracy of the computationally more expensive global-fit approximations.

Building better approximation models

Park & Levitt[10] and Rooman *et al.*[9] showed that discrete approximation models for protein structure that take the uneven distribution of residue conformations in real proteins into account are more accurate than models of comparable complexity that do not. The discrete models they constructed treat all residues along the chain equally, in the sense that each residue can have any one of c conformations (where c is the complexity of the model). These conformations are described by the pair of angles (ϕ, ψ) that defines the positioning of the residue with respect to the previous residue along the chain.

The discrete approximations we construct with libraries of four residue fragments are equivalent to the optimized models considered in earlier studies.[9,10] Indeed, the complexity of the library L_4^s of size s is $s^{1/(4-3)} = s = |L^s|_4$, which is just the number of its elements. In effect, the library fragments encode the (ϕ, ψ) angles of their last residue with respect to the previous parts of the approximation, while the first three residues position the fragment. The results we achieve with libraries of fragments of four residues are similar to those obtained previously by others. Park & Levitt's[10] best four-state model has an accuracy of 2.22 Å when approximating the test set of proteins, while the best four-state model we found achieves an accuracy of 2.23 Å. Rooman *et al.*[9] found a six-state model with an average accuracy of 1.74 Å over the proteins test set, compared with an average accuracy of 1.64 Å in our six-state model.

The conformations of consecutive residues along the backbone of proteins are correlated to one another. Under the reasonable assumption that the libraries we find for a specific size and fragment length are optimal, our results show this correlation between conformations of neighboring residues. For illustration, consider the correlation of conformations of pairs of neighboring residues, which is reflected in the relative accuracy of models from libraries of four and five-residue fragments, respectively. Imagine that the conformation of two consecutive residues is independent and without any correlation. If L_4^s is an optimal s-element library of four-residue fragments, it can be used to construct equivalent library $L_5^{s'}$ of five-residue fragments (concatenate all pairs of four residues with a three-residue overlap to give a new library containing $s' = s^2$ five-residue fragments). Clearly, L_4^s and $L_5^{s'}$ span exactly the same space of approximating structures and have the same complexity, s. If the conformation of two consecutive residues along the chain was independent, and L_4^s is optimal, then $L_5^{s'}$ is optimal too. This would mean that global-fit models would have the same accuracy for libraries of four and five-residue fragments. Here, we find that for the same complexity, models with five-residue fragments are significantly more accurate than those with four-residue fragments (in Table 2 for a complexity of 10, the five-residue fragments fit to 0.99 Å, whereas four-residue fragments fit to 1.12 Å) indicating very significant correlations.

Our use of fragment libraries in construction of proteins approximation space allows exploitation of the correlations of conformations along the backbone to achieve better low-complexity models. This effect is particularly important if one wants to reproduce native structures to better than 1 Å. Here, we can achieve such accuracy with a complexity of 12 for a four-residue fragment or 10 for a five-residue fragment. By comparison, the Park & Levitt[10] model would require a complexity of over 50 states-per residue to achieve an accuracy of 1 Å. As the number of possible conformation for a chain of length n residues, depends on $(\text{Complexity})^n$, these differences have a huge impact on the size of the particular protein's conformation space. We expect to be able to get even better results with libraries of six or seven-residue fragments. Unfortunately, very large datasets of refined protein coordinates are needed to make reliable libraries for the longer fragments. Here, we have a 250-fragment library of length 7 that has a complexity of 4 and attains a Global-RMS of 1.91 Å. To obtain a Global-RMS value of 1 Å would require a complexity of about 8 and a library of $8^4 = 4096$ fragments. With the rapid pace of protein structure determination, we believe that such a library may soon be possible.

Conclusions

The fragment libraries that we have derived can approximate native structures with Global-fit cRMS deviations that vary from 2.9 Å to 0.76 Å for libraries whose complexities vary from 2.66 to 15 states per residue, respectively. When employed together with the buildup method of protein structure approximation, this gives a wide range of different-resolution models that are expected to be useful for a wide range of applications, including: protein structure prediction, loop fitting, exhaustive enumeration of peptide conformations, and low-resolution structure determination by NMR or X-ray crystallography.

Methods

Datasets of protein fragments

A set of proteins from the Protein Data Bank[13] (PDB) with the most reliable structural data served as our initial data set for the clustering analysis. Specifically, we use the 200 unique protein domains as defined by SCOP version 1.57[14] with the highest-ranking SPACI scores[12] (see Table 1). The 200 domains, all with a SPACI score greater than 0.534, have a total of 36,397 residues. In our study, we approximate the chain path describing

the fold of each of these proteins by the atomic coordinates of its C^{α} atoms.

Four training sets of protein backbone fragments were extracted from the above set of proteins. These training sets differ in fragment length, and include sets of four, five, six and seven-residue fragments. Each of these sets is comprised of all consecutive non-overlapping fragments of the appropriate length, starting at a random initial position. It is not advantageous to include overlapping fragments in these sets, as any two neighboring fragments are very close to each other solely because they have a large overlapping part. This structural overlap introduces noise into the training set and makes the clustering task significantly harder. The numbers of fragments in the data sets we compiled are 8949, 7123, 5910 and 5029 for the four, five, six and seven-residue data sets, respectively.

A distance measure in structure space

We use the coordinate root-mean-square (denoted cRMS) deviation of the C^{α} atom to measure the structural similarity of any two fragments. This measure satisfies the triangle inequality:

$$\mathrm{cRMS}(AC) \leq \mathrm{cRMS}(AB) + \mathrm{cRMS}(BC)$$

for three fragments A, B, C of the same size, making it acceptable for use in clustering.[15] The cRMS deviation is measured between pairs of atoms in the two fragments after optimal least-squares superposition.[16] We considered using other measures such as: (1) the RMS value of the f residue (ϕ, ψ) torsion angles, where f is the fragment length; (2) the RMS deviation of the $f-3$ chain α angles (the torsion angle defined by four consecutive C^{α} atoms[17]); (3) the RMS value of $f(f-1)/2$ inter-C^{α} distances in each fragment. None of these alternatives was satisfactory, as the (ϕ, ψ) torsion angles were too noisy and reflective of local change, the α angles were too coarse a description of the fragment shape, and the inter-C^{α} distances did not discriminate a right-handed from a left-handed structure.

Pruning and clustering the fragments datasets

Two special characteristics of our fragment data-sets that need to be considered before clustering are the outlier fragments and the very high concentration of α-helical fragments. Outliers are fragments with a relatively large cRMS deviation from all other fragments, and therefore cannot be considered representative of common structural protein motifs. We facilitate the clustering task by weeding out these outliers according to a threshold. In all cases, approximately 10% of the fragments are discarded using the threshold values 0.074 Å, 0.307 Å, 0.487 Å and 0.755 Å for the data sets of fragments of length four, five, six and seven residues, respectively; we do this by eliminating any fragment whose cRMS deviation from the closest other fragment in the dataset is greater than the threshold value. An additional unique characteristic of our training set is a highly populated region of fragments from α-helices, which complicate the clustering procedure.

We cluster each of the different length fragment datasets using *k*-means simulated annealing, a novel clustering technique that varies *k*-means clustering by using simulated annealing to improve the cluster centroids. The *k*-means simulated annealing repeatedly runs *k*-means clustering and then merges two clusters and splits another, in a Monte Carlo fashion. The clusters to be merged are selected at random, with nearby clusters more likely to be chosen; the cluster to be split is selected at random, with disperse clusters more likely to be selected. We tried a number of different scoring functions and the one that performed best was total variance of the clustering (the sum over all clusters of the square of the distance of any fragment to its cluster centroid). The desired number of clusters is given to the clustering procedure as input and the improvement step described maintains the number of clusters. This scheme surpasses normal *k*-means by its improved handling of the wide range of fragment concentrations (there are many more α-helical fragments), and by its insensitivity to the initial choice of cluster centers. It even works a little better than the much more time-consuming hierarchical clustering method that merges clusters on the basis of the maximum distance between any members of the different clusters. A more detailed description of the simulated annealing *k*-means clustering technique as well as a comparison to other clustering methods will be given elsewhere (our unpublished results).

Libraries of common protein structural motifs

The clustering result is used to compile a library, a small representative set of protein fragments. The libraries are succinct representations of specific clustering runs and they consist of the cluster centroids – the fragment that has the minimum sum of cRMS deviations from all the other cluster fragments. Our study explores many clustering runs (that vary by the number of clusters, the length of the fragments in the dataset or the specific cluster labels they assign to each fragment), each one giving rise to a different library of motifs. In each case, we use *k*-means simulated annealing clustering starting with 50 different random seeds and choose the best library as that with the minimal total variance score. The representative fragments in each library are, therefore, determined by the clustering procedure used to generate it. The size of the library is the required number of clusters; the length of the fragments in the library is that of the fragments in the dataset clustered.

Evaluating the quality of a library

Next, we turn to evaluating the quality of a library: a library is considered "good" if it approximates real protein structure accurately. The clustered fragments are used to construct a library that is representative of all fragments in the training set data. We aspire to a set of fragments that also represents well all protein motifs (of this length) found in real proteins. The quality of a library is evaluated using a test set of protein structures that is independent of the training set. For comparison with earlier work, we use the set of 145 proteins used by Park & Levitt.[10] We use two criteria to evaluate a library: (1) local-fit; and (2) global-fit.

The local-fit is a measure of how well the library fits the local conformation of all the proteins in the test set. Each protein structure is broken into the set of all its overlapping fragments of the specified length, f. The best local-fit approximate structure for a protein is constructed, in linear time, by finding for each of the protein fragments the most similar fragment in the library (in terms of cRMS). Averaging this cRMS value over all the fragments in all the proteins in the test set gives the local-fit score for the particular library.

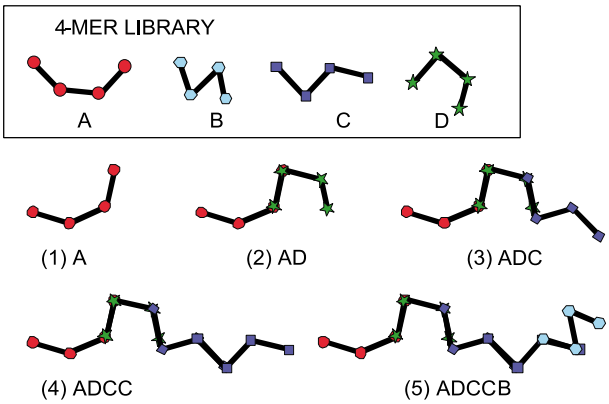

Figure 6. A simple two-dimensional example showing the construction of a chain formed by superimposed fragments. The four-state library contains four fragments, each four residues long and is enclosed in the box. The fragments have library serial numbers of A–D. The first two residues of any fragments can be superimposed on the last two residues of the preceding fragment, thus determining the position of the fragment. The four stages of the construction process of the chain ADCCB are shown.

The global-fit is a measure of how well the library fits the global three-dimensional conformation of all the proteins in the test set. One possible way to construct contiguous three-dimensional structures from the library fragments is by concatenating the best local-fit library fragments found above. If the first C^{α} atom of each fragment is superimposed on the last C^{α} atom of the preceding fragment, one would need to specify the relative orientation of the two fragments. This could be done using two polar coordinates but is, in any case, unsatisfactory, as the reconstruction would not be discrete, in that one would need to specify the list of fragments as well as the values of the continuous polar angles. More succinctly, there would not be the desired one-to-one correspondence between the string of library fragment codes and the global three-dimensional structure.

We therefore use a different scheme, denoted as global-fit approximation, for constructing chains from protein fragments: the position of each added fragment is determined by best superimposing its first three C^{α} atoms onto the last three C^{α} atoms of the preceding fragment already constructed. Even if the two C^{α} triplets do not match perfectly, this will define the relative orientation of the fragments uniquely, provided that the atoms of each C^{α} triplet do not lie along a line. In polypeptide chains, the distance between consecutive C^{α} atoms is fixed at 3.8 Å and the angle formed by three consecutive C^{α} atoms is between 90° and 130°.[17]

Figure 6 demonstrates a two-dimensional analog of this scheme for constructing two structures from fragments of a four-element library. Notice that in two dimensions, any two (rather than three) consecutive amino acid residues can be superimposed on any two consecutive amino acid residues in another fragment. We emphasize that the library fragments are used as mere templates; any fragment can be used repeatedly along the constructed chain. In this global-fit scheme, a structure is described completely by the list of library fragments that construct it. There is a one-to-one correspondence between the space of all approximations and the space of all strings of library serial numbers. Therefore, the space of all approximations constructed using a library is discrete, and when the length of the target structure is fixed it is also finite.

Computer implementation of global-fit approximations

While the best local-fit approximation is found easily by finding the library fragment that best fits each local fragment, the sequence of library fragments needed for the global-fit is much harder to find. The optimal sequence of library fragments must define the three-dimensional structure with the minimal cRMS deviation from the real structure of the target protein. The number of possible sequences of fragments is, unfortunately, exponential in the protein's length, so that it is impossible to consider all sequences in search for the best global-fit approximation. We, therefore, follow Park & Levitt[10] and use a greedy algorithm for finding a good rather than the best global-fit approximation. Let f denote the length of the fragments in the library. Starting at the N terminus, we construct approximations for increasingly larger segments of the protein. Given a partial approximation, we extend it using the best library fragment, i.e. the one whose concatenation yields a structure of minimal cRMS deviation from the corresponding segment in the protein. This concatenation is achieved by superimposing the first three residues of the added fragment on the last three residues of the already constructed segment so that $f - 3$ residues are added each time. We repeat this process until the C terminus of the target protein is reached. This process is deterministic and takes linear time.

An important property of this model-building method is that each step is local, while our criterion for evaluating the goodness of it is global. A fragment used in the construction of the approximating structure influences the overall accuracy of the approximation *via* the accuracy of the local protein segment it describes, and through the positioning it determines for its following fragments. Consequently, it may be beneficial to make a less greedy choice: a less-well fitting library fragment may allow better positioning of subsequent fragments, improving the overall quality of the approximation.

Our algorithm is therefore improved by keeping a set of candidate structures for extension, rather than a single one as described above. Specifically, we allow the algorithm to be slightly less greedy and keep a set (or heap) of the best N_{keep} greedily constructed approximations for the segment of the protein approximated thus far. At each step, we extend each of the N_{keep} approximations with all possible library fragments, and then greedily keep the best N_{keep} approximations. This algorithm is still greedy, and therefore does not guarantee the globally optimal solution, yet it explores a slightly bigger part of the approximations space. Greedy algorithms like this were first used in computational biology by Vasquez & Scheraga to build-up low energy conformations of polypeptide chains.[18,19]

Complexity of the approximation models

Here, the complexity of the space of global-fit approximations is the average number of states per residue or equivalently the nth root of the size of the total number of states for a chain of length n, following the convention

set by Park & Levitt.[10] Equivalently, the complexity of a library measures the size of the space of structures defined by it, normalized so that it is independent of the lengths of the approximated proteins. Let L_f^s be a library of s fragments, each f residues long. Adding one library fragment to a growing chain extends the chain by $f - 3$ residues. The number of fragments, m, needed for n residues is the first integer larger than or equal to $n/(f - 3)$. Thus, a string of m fragment serial numbers from the library L defines an approximating structure of n residues. The size of the approximation space is equal to the number N of such strings where:

$$N = s^m = s^{n/(f-3)} = (s^{1/(f-3)})^n$$

Thus, the complexity is $s^{1/(f-3)}$ states per residue: it is a property of the fragment library varying with both the library size and the fragment length. Although the local-fit approximations do not have a complexity, as they cannot be used to construct a chain, we sometimes refer to the global-fit complexity measures of the fragment libraries used.

References

1. Corey, R. B. & Pauling, L. (1953). Fundamental dimensions of polypeptide chains. *Proc. R. Soc. London*, **141**, 10.
2. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Conformation of polypeptides and proteins. *J. Mol. Biol.* **7**, 95–99.
3. Jones, A. T. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819–822.
4. Unger, R., Harel, D., Wherland, S. & Sussman, J. L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins: Struct. Funct. Genet.* **5**, 355–373.
5. Micheletti, C., Seno, F. & Maritan, A. (2000). Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins: Struct. Funct. Genet.* **40**, 662–674.
6. de Braven, A. G., Etchebest, C. & Hazout, S. (2000). Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks. *Proteins: Struct. Funct. Genet.* **41**, 271–287.
7. Wintjens, R. T., Rooman, M. J. & Wodak, S. J. (1996). Automatic classification and analysis of alpha–alpha-turn motifs in proteins. *J. Mol. Biol.* **255**, 235–253.
8. Oliva, B. & Bates, P. A. (1997). An automated classification of the structure of protein loops. *J. Mol. Biol.* **266**, 814–830.
9. Rooman, M. J., Kocher, J. I. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry*, **31**, 10226–10238.
10. Park, B. & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **249**, 493–507.
11. Simon, I., Glasser, L. & Scheraga, H. A. (1991). Calculation of protein conformation as an assembly of stale overlapping segments: application to bovine pancreatic typsin inhibitor. *Proc. Natl Acad. Sci. USA*, **88**, 3661–3665.
12. Brenner, S. E., Koehl, P. & Levitt, M. (2000). The ASTRAL compendium for protein structure and sequence analysis. *Nucl. Acids Res.* **28**, 254–256.
13. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
14. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
15. Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*, Wiley, New York.
16. Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallog. sect. A*, **34**, 827–828.
17. Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.
18. Vasquez, M. & Scheraga, H. A. (1985). Use of buildup and energy-minimization procedures to compute low energy structures of the backbone of Enkaphilin. *Biopolymers*, **24**, 1437–1447.
19. Vazquez, M. & Scheraga, H. A. (1988). Calculation of protein conformation by the build-up procedure. Application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data. *J. Biomol. Struct. Dyn.* **5**, 705–755.
20. Kraulis, J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.

Edited by B. Honig

(Received 12 July 2002; received in revised form 29 August 2002; accepted 29 August 2002)

http://www.academicpress.com/jmb

Supplementary Material is available on IDEAL