Pacific Symposium on Biocomputing 5:165-175 (2000)

RECOGNITION OF PROTEIN STRUCTURE: DETERMINING THE RELATIVE ENERGETIC CONTRIBUTIONS OF β-STRANDS, α-HELICES AND LOOPS

BORIS REVA and SID TOPIOL
Novartis Pharmaceutical Corporation, Core Technologies Area
556 Morris Ave., Summit, NJ 07901, USA
Tel: (908)277 5412; Fax: (908)2774910

We examine the role of residues of secondary structure in recognition of the native structure of proteins. The accuracy of recognition was estimated by computing the Z-score values for fragments of protein chains in threading tests. By testing different combinations of secondary structure fragments of 240 non-homologous proteins we show that the overwhelming majority of proteins can be successfully recognized by the energies of interaction between residues of secondary structure. We also found that β-structures contribute more significantly to fold recognition than α-helices or loops. To validate the Z-score calculations in measuring the accuracy of recognition we evaluated the deviation of the energy distribution from the normal law. The normal law satisfactory approximates the shape of the energy distribution for the majority of proteins and chain fragments; however, deviations are often observed for short fragments and for fragments with relatively high Z-score values. The results of the study justify recognition of remote homologs by threading methods based on a backbone of secondary structure rather than of a whole chain because loops of homologs differ more significantly than strands and helices, and the contribution of loops in structure recognition is relatively small.

1 Introduction

The threading method [1-9] occupies a central position among other approaches to protein structure prediction. In this method a search space for the optimal structure is limited to a relatively small number of folds (usually extracted from the PDB). A query protein chain evaluates each of these structures by sorting different positions (threading) along the backbone of the considered structure. If one of the structures in a folding library has a sufficient number of structural similarities to the native structure (e.g. inter-residue distances) then the energy (free energy) of this structure should be the minimal one, and the protein chain will thereby “recognise” this structure. Hence, a choice of the right structure is determined by an energy balance between two groups of interactions for each structure considered: in the first group, the inter-residue distances and interactions are close to the corresponding native values. In the other group, interactions are different from the native ones. In a typical case [10-12] of structures of the same fold family, the residues of the conservative “core” (mainly residues with secondary structure) contribute to the first group of “recognising” interactions, while the other residues (mainly residues of geometrically variable loops) contribute to an “error” energy.

The idea of miscounting some of the interactions so as to improve for the accuracy of structure recognition has been successfully used by Bryant and co-workers [3-5] who based their fold recognition method on threading over “cores” of protein molecules composed of α-helices and β-strands. However, in the general case, reducing the number of interactions is undesirable for structure recognition because it results in a smaller energy gap separating the native structure from other structures.

There is no accepted recipe for selecting structural “cores” that would be optimal for structure recognition. Indeed, there are many questions to be answered in developing such an approach. In this work we try to answer the simplest ones:

(i) Is it possible to recognise the native structure counting only interactions between residues of secondary structure?

(ii) What are the differences between α-helices, β-strands, and loops in structure recognition?”

2 Method

2.1 Z-score

The accuracy of protein structure recognition is commonly characterised by the value of the Z-score [13]:

$$Z = \frac{E_{Nf} - \langle E \rangle}{D}, \tag{1}$$

where E_{Nf} is the energy of the native fold,

$$\langle E \rangle = \frac{1}{M}\sum_{i=1}^{M} E_i$$

is the average energy,

$$D = \left[\frac{1}{M}\sum_{i=1}^{M}\left(E_i - \langle E \rangle\right)^2\right]^{1/2}$$

is the standard deviation of energies; E_i is the energy of the i-th of M alternative folds $(i = 1,...,M)$. The Z-score gives the average number of standard deviations between the native and the random fold energy. It allows one to estimate the expected number of folds N_Z, among which the native structure can still be selected as the one with the lowest energy. Assuming a normal distribution of energies of competing folds,

$$N_Z = \frac{\sqrt{2\pi}}{\int_{-\infty}^{Z} \exp(-\frac{x^2}{2})dx}. \tag{2}$$

When $Z << -1$, which corresponds to a reasonable accuracy of predicting methods,

$$N_Z = \sqrt{2\pi} Z \exp(\frac{Z^2}{2}) . \qquad (3)$$

The larger N_Z (and hence – Z), the more accurate the protein structure recognition.

2.2 Generation of alternative structures by gapless threading

For evaluation of an average energy, $\langle E \rangle$, and a standard deviation, D, one needs to use a representative set of alternative structures. Commonly, such structures are obtained by the method [13] of gapless threading of a query sequence onto all possible *3D* structures provided in the form of backbones of a set of non-homologous proteins. No internal gaps or insertions are allowed; thus, a chain of *N* residues in length can be threaded through a host protein molecule of *M* residues in length in *M-N+1* different ways. Because threaded structures which differ by a small number of register shifts are similar to each other (measured [15] by RMSD) we sample threaded structures with register shifts of 10. Only a tiny fraction of the possible conformations of the query sequences is generated by this procedure. However, this fraction is enough for a crude estimation of the Z-score values. (All the protein structures used in threading were taken from the PDB according to Sander's 25% similarity list [14] of Oct.97. We used [15] only 364 structures from this list as follows: those with no chain breaks, with a resolution better than 2.5A and R factor less than 0.2, and with no structural homologs.)

To determine a particular contribution of a given group of residues to structure recognition we calculated separately the energy of this group in the native structure as well as the corresponding energies in alternative structures. These energies were used for estimating the Z-score for a given group of residues. We considered separately the residues of: (i) α-helices and β-strands; (ii) α-helices, (iii) β-strands, (iv) loops, (v) α-helices and loops, and (vi) β-strands and loops. In our tests we used 240 non-homologous proteins (from the list of 364) of 60 to 350 residues in length. Threading provided us with ~8,000 structurally non-related folds for the shortest chains and ~1,200 for the longest ones. The secondary structure assignment was done according to the information given in the annotations of the PDB files.

2.3 Potentials

In our energy calculations we used pairwise C_α atom based potentials [17-18] as illustrated in Figure 1.

The derivation of these potentials is based on the theory that explains the nature of Boltzmann statistics of protein structure [16]. According to this theory [17], the energy $\varepsilon_{\alpha\beta}(r)$ between remote residues α and β along a chain, at a distance *r=R** (*R** is the maximal distance of interactions, *R*=14.5Å* in this wor*k*), can be

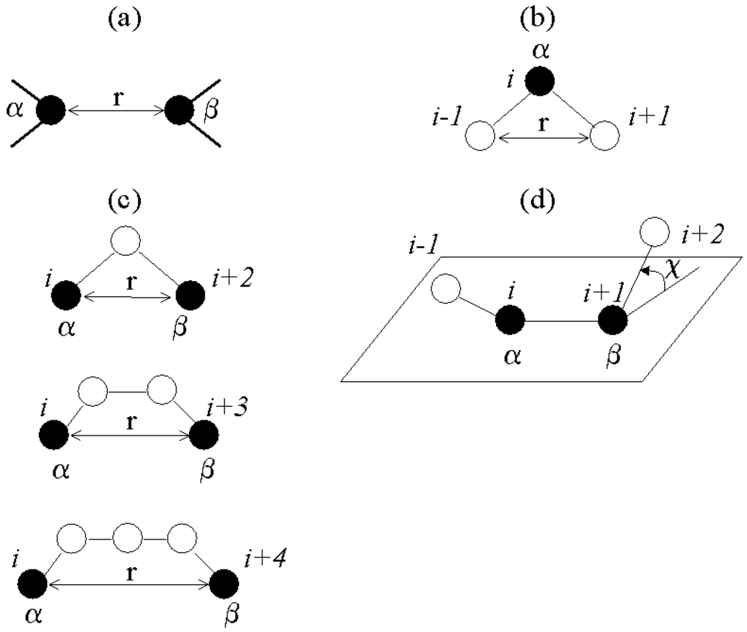

Figure 1. Scheme of interactions taken into account; filled circles show C_α atoms of residues to which a potential is applied. **a.** Interactions depending on the distance **r** between residues α and β ("long-range" potentials [17-18]); **b.** Chain bending potential [17-18]: bending at the intervening residue α affects the distance **r** between terminal residues i-1 and i+1; **c.** Short-range potentials [17-18] depending on the distance between terminal residues α and β. **d.** Chiral potential [18] depending on the dihedral angle χ between two planes of C_α atoms, (i-1,i,i+1) and (i,i+1,i+2), and the residues α and β. Potentials are derived at a resolution of Δ=1A; the angular resolution for chiral potential is Δχ=30 degrees. This crude resolution of energy functions renders insignificant the statistical errors connected with the inclusion (or deletion) of each individual protein in the database used in the derivation of the potentials [17-18].

estimated as

$$\varepsilon_{\alpha\beta}(r) = -RT\ln\{N_{\alpha\beta}(r,\Delta)\big/[N(r,\Delta)\,f_{\alpha\beta}(R^*)]\}\,, \qquad (4)$$

where $N_{\alpha\beta}(r,\Delta)$ is the number of αβ pairs observed in proteins at distances [*r*-*Δ/2*,*r*+*Δ/2*]; $N(r,\Delta) = \sum_{\alpha\geq\beta}\sum_{\beta} N_{\alpha\beta}(r,\Delta)$ is the total number of residue pairs at a distance *r*; $f_{\alpha\beta}(R^*) = \tilde{N}_{\alpha\beta}(r > R^*)\Big/\sum_{\alpha\geq\beta}\sum_{\beta}\tilde{N}_{\alpha\beta}(r > R^*)$ is the fraction of αβ pairs at "no-interaction" distances *r*>*R** ($\tilde{N}_{\alpha\beta}(r > R^*)$ is the total number of αβ pairs at *r*>*R**).

Similarly, short-range (u^s), bending ($b^{(2)}$), and chiral (*h*) potentials are estimated as:

$$u^s{}_{\alpha\beta}(r) = -RT\ln\{[N^s{}_\alpha(r,\Delta)\big/(N^s(r,\Delta)\, f^s{}_\alpha)][N^s{}_\beta(r,\Delta)\big/(N^s(r,\Delta)\, f^s{}_\beta)]\}\,, \tag{5}$$

$$b^{(2)}{}_\alpha(r) = -RT\ln\{M^{(2)}{}_\alpha(r,\Delta)\big/[M^{(2)}(r,\Delta)\, f^{(2)}{}_\alpha]\}\,, \tag{6}$$

$$h_{\alpha\beta}(\chi) = -RT\ln\{Q_{\alpha\beta}(\chi,\Delta\chi)\big/[Q(\chi,\Delta\chi)\, f_{\alpha\beta}]\}\,, \tag{7}$$

where $N^s{}_\alpha(r,\Delta) = \sum_\beta N^s{}_{\alpha\beta}(r,\Delta)$ is the number of *i,i+s* pairs at a distance *r*, at a resolution Δ, and with a residue α in a position *i*; $f^s{}_\alpha = \sum_r N^s{}_\alpha(r,\Delta) / \sum_r N^s(r,\Delta)$ is a fraction of residues α in a position *i*; and $N^s(r,\Delta) = \sum_\alpha N^s{}_\alpha(r,\Delta)$. The bending and the chiral energies are computed in a similar way using the corresponding statistics $M^{(2)}{}_\alpha(r,\Delta)$ and $Q_{\alpha\beta}(\chi,\Delta\chi)$.

2.4 Normal distribution for alternative fold energies

In estimating the accuracy of protein structure recognition by a Z-score value we assume the normal distribution for alternative fold energies. This type of energy distribution is typical for systems approximated by the random energy model, REM [19]. The total energy in REM is a sum of many independent random interactions. The energy of a protein globule is a sum of thousands of inter-residue interactions. Alternative random structures allow for a huge variety of inter-residue contacts. Therefore, it is generally believed that the energy spectrum of a protein molecule has the normal form. However, because we compare the energies of interactions for relatively small and specially chosen groups of residues it is especially important to validate the applicability of the REM for separate chain fragments. A deviation between a theoretical and an observed distribution is usually measured by the χ^2 value [20]. To compute this quantity we divide an energy distribution into bins and calculate the observed and the expected bin populations. The χ^2 value for a protein *p* is computed as:

$$\chi_p{}^2 = \frac{1}{Q}\sum_{i=1}^{K}\frac{\left(n_{p,i}^{(o)} - n_{p,i}^{(e)}\right)^2}{\left(n_{p,i}^{(e)}\right)^2}\,, \tag{8}$$

where $n_{p,i}^{(o)}$ and $n_{p,i}^{(e)}$ are the observed and the expected population, respectively, of energies in bin i for protein *p*; *K* is the number of bins taken into account and *Q* is the number of degrees of freedom. For the normal distribution *Q = K - 3* because the total population, mean value and dispersion of the expected statistics are extracted from the observed statistics. When χ^2 ~ 1, the experimental data can be treated as confirming the expected statistics. We used a threshold of χ^2 = 3 (corresponding to 5% significance at *Q*=2) to distinguish the energy distributions that differ from the normal one. Because the region of low energies ($E < \langle E \rangle$) is of major interest, the χ^2 values were estimated for the left-hand sides of the energy distributions. The interval $\left(-\infty,\langle E\rangle\right)$ was divided into K = 5 bins: $\left(-\infty,E_{40}\right)$, $\left(E_{40},E_{40}+\delta\right),\ldots,$ $\left(E_{40}+3\delta,\langle E\rangle\right)$, where E_{40} is the 40-th lowest observed energy, $\delta=\left(\langle E\rangle-E_{40}\right)/4$.

We also tried to estimate changes in the Z-scores caused by simple changes in the number of interactions. It is easy to see that under the assumption of REM (i.e., inter-residue interactions are random and non-correlated) both the energy difference, $\Delta E=E_{Nf}-\langle E\rangle$, and the energy dispersion, D^2 , must be proportional to the total number of interactions, N . Using the proportionality $Z=\Delta E/D\propto\sqrt{N}$ one can easily estimate the expected value of the Z-score

$$Z_{predict_\nu}\approx Z_0\sqrt{\frac{N_\nu}{N_0}}\,, \tag{9}$$

where the index 0 corresponds to the full length of the protein in question and ν corresponds to a specific subset, e.g., the subset of β+α fragments.

3 Results and Discussion

The results of the study are summarised in Tables I-IV. In these tables we compare the averaged Z-scores, energies and numbers of inter-residue interactions obtained for the whole chains and for the selected fragments of protein chains. The percentages of β-, α- and loop structures indicate how many residues actually contribute to structure recognition. For a more accurate estimate of the quality of recognition we introduce a threshold value of Z^*= -4.5 (which corresponds to N_z~300,000, Eq.(3)) and give the percentage of proteins with Z < Z^*. The $Z_{predict}$ estimates according to Eq.(9) and χ^2 values of Eq.(8) are used to compare the deviations between the observed energy distributions and the REM based estimates.
Table IA presents the results of averaging over all proteins used in the study. It shows that the residues of secondary structure (i.e. β+α) occupying an average ~56% (24%β+32%α) of the total sequence length and contributing ~35% of the total number of interactions play a decisive role in structure recognition. (The average $<Z_1>$-score is -7.4; 92% of proteins had a Z_1-score less than -4.5.) The data of Table 1A allows comparison of the relative contributions of β-structures, α-helices and loop fragments. Surprisingly, β-structures occupying only ~1/4th of the total sequence length and contributing ~1/10th of the total interactions provide significantly more accurate structure recognition than α-helices or loops (based on lower $<Z_1>$ values and higher numbers of proteins with $Z_1<Z^*$).

A critical question for this study is the legitimacy of the approach used for estimating the accuracy of fold recognition. The Z-score comparative analysis is based on the idea that distributions of alternative fold energies are always well approximated by the normal law. This assumption as well as the REM based dependency on a number of interactions given by Eq.(9) is also examined in Table IA.

Table IA. Averaged characteristics of energy distributions obtained in threading tests for different combinations of structural fragments*. Averaging is done over all considered proteins.

Combination **of fragments**	**Number** **of proteins**	**$\langle Z_0 \rangle$**	**$\langle Z_1 \rangle$**	**$\langle Z_{predict} \rangle$**	**% of** **Z<-4.5**	**% of** **interactions**	**% β**	**% α**	**% l**	**% of proteins** **with χ2<3.0**
β+α+l	240	-8.8	-8.8	-8.8	99	100	**24**	**32**	**44**	74
β+α	240	-8.8	-7.4	-5.2	92	35	**24**	**32**	44	65
β	220	-8.9	-6.0	-2.7	71	9	**26**	30	44	34
α	232	-8.9	-4.0	-3.8	32	19	23	**34**	43	54
l	240	-8.8	-3.8	-4.0	24	21	24	32	**44**	50
β+l	220	-8.9	-6.8	-6.4	91	52	**26**	30	**44**	71
α+l	232	-8.9	-6.2	-7.1	75	64	23	**34**	**43**	69

Table IB. Averaged characteristics of energy distributions obtained in threading tests for different combinations of structural fragments*. Averaging is done over proteins with normal energy distributions (χ^2<3.0).

Combination **of fragments**	**Number** **of proteins**	**$\langle Z_0 \rangle$**	**$\langle Z_1 \rangle$**	**$\langle Z_{predict} \rangle$**	**% of** **Z<-4.5**	**% of** **interactions**	**% β**	**% α**	**% l**	<χ2>
β+α+l	177	-9.2	-9.2	-9.2	100	100	**24**	**32**	**44**	1.38
β+α	156	-9.4	-7.8	-5.6	97	36	**22**	**37**	42	1.47
β	74	-9.5	-6.1	-3.1	77	12	**30**	29	42	1.56
α	126	-9.6	-4.7	-4.4	50	21	19	**39**	42	1.43
l	120	-9.6	-4.1	-4.5	35	22	23	32	**45**	1.49
β+l	156	-9.2	-7.0	-6.6	95	53	**26**	29	**45**	1.37
α+l	161	-9.4	-6.6	-7.5	83	65	**21**	36	**43**	1.43

***β, α** and **l** stand, respectively, for α-helices, β-structure and loops; only 220 and 232 proteins of the total set of 240 proteins, have respectively, β– and α– structure; **$\langle Z_0 \rangle$, $\langle Z_1 \rangle$** are the averaged Z-scores for, respectively, the whole chains and for the chain fragments calculated by Eq.(1); **$\langle Z_{predict} \rangle$** is the averaged Z-score calculated by Eq.(9); **% of Z<-4.5** gives a percent of proteins with a Z-score less than -4.5; **% of interactions** is the averaged percentage of inter-residue interactions; **%β, %α, %l** are correspondingly the average percentages of β, α and loop structure, entries corresponding to fragments tested in each row are shown in bold; 2 degrees of freedom are used in χ^2 calculation according to Eq.(8).

One can see that for the majority of the tested chain fragments the energy distributions can be reasonably treated as the normal ones (the average of the last column is 60%). The longer chain fragments show less deviation from the normal law than the shorter ones (not shown). One can also see that the estimates of $Z_{predict}$ (based on the idea that accuracy of recognition is directly dependent on the number of interactions, see Eq.(9)) differ significantly from the observed values ($\langle Z_1 \rangle$) for all types of chain fragments, especially for the fragments of secondary structure and β-structure. For instance, for the β+α fragments $\langle Z_1 \rangle$ value of –7.4 as compared to $\langle Z_{predict} \rangle$ of –5.2 indicates that the contribution of these fragments to structural recognition is significantly higher than expected from their proportional representation in the structures. Loops, on the contrary, show higher Z-scores than predicted from the corresponding numbers of interactions.

Because the number of chain fragments with energy distributions poorly approximated by the normal law is significant and can not be easily ignored we performed two additional tests to compare the data of Table IA. In the first test (Table IB), we extracted the data using only chain fragments with the normal distributions of energies. In the second test, we used a wittingly biased set of alternative folds obtained by threading with a one-residue register shift along the backbones of the host proteins (Table II). The trends of the data of Tables IB and IA are similar in all the feature details. However, the Z-scores and the energies of Table IB are lower and the correlation coefficients are higher than the corresponding ones of Table IA. One can also see that the chain fragments of pure β-structure and α-helices presented in Table IB occupy more of the sequence length than the corresponding fragments of Table IA.

Table II. Averaged characteristics of energy distributions obtained in threading tests for different shifts along a host molecule backbone*.

Combination	Number	Shift: 1 residue				Shift: 10 residues			
of	of	$\langle Z_0\rangle$	$\langle Z_1\rangle$	% of	% of proteins	$\langle Z_0\rangle$	$\langle Z_1\rangle$	% of	% of proteins
fragments	proteins			$Z<-4.5$	with $\chi 2<3.0$			$Z<-4.5$	with $\chi 2<3.0$
β+α+l	240	-8.9	-8.9	99	34	-8.8	-8.8	99	74
β+α	240	-8.9	-7.4	91	27	-8.8	-7.4	92	65
β	220	-8.9	-6.1	69	6	-8.9	-6.0	71	34
α	232	-8.9	-4.0	30	19	-8.9	-4.0	32	54
l	240	-8.9	-3.8	22	3	-8.8	-3.8	24	50
β+l	220	-8.9	-6.8	90	30	-8.9	-6.8	91	71
α+l	232	-8.9	-6.2	75	16	-8.9	-6.2	75	69

*See legend for Table IA

Table II shows that the relative length of the shift between two subsequent structures (1 vs.10) obtained by threading causes a drastic change in the normal law energy distributions. Nevertheless, one can see that the Z-score values computed for these different sets of alternative structures are practically the same.

We do not know the exact reasons that could cause a deviation from the normal distribution in threading energies for given chain fragments. (Because of errors in the energy functions and errors in filtering of alternative structures it is impossible to distinguish between a “true” deviation caused by correlation in sequence and an “erroneous” one.) Generally however, the consistent results of Tables 1A, IB and II support the use of the Z-score computation in evaluation of the role of given chain fragments in structure recognition. (Chain fragments should occupy ~1/3 of the total sequence length or more as indicated by the data for β-structural and α-helical fragments of Table IB.)

To better understand the differences between the chain fragments that recognise the native structure and those that do not, we formed two extreme groups of proteins. In the first group we included the fragments that recognise their native structure better than the whole chain (Table III); the fragments with poor structure recognition (Z-score >-4.5) formed the second group (Table IV). One can see immediately from Table III that in all those cases when the Z-score value of a chain fragment is lower than the Z-score value of the whole chain, the energy of this fragment is also lower than the energy of the whole chain. The fragments of Table IV follow the reverse pattern of Z-score values and energies. The data of both tables confirm the special role of β-structure. As it follows from the structural composition, the fragments with significant contribution of β-structure and of pure β-structure

Table III. Averaged characteristics of energy distributuins for proteins with $Z_1<Z_0$*

Combination of fragments	Number of proteins**	$<Z_0>$	$<Z_1>$	$<E_1-E_0>$	% of interactions	% β	% α	% l	% N_{tot}
β+α	34	-7.6	-8.2	-31.0	32	**32**	**23**	45	14.2
	17	-8.2	-8.9	-35.4	35	**26**	**34**	41	7.1
β	16	-7.1	-7.7	-11.1	20	**42**	10	48	7.3
	4	-6.8	-7.4	-28.4	24	**47**	6	47	1.8
β+l	7	-7.2	-7.5	-14.0	88	**48**	7	**46**	3.2
	6	-7.2	-7.6	-15.2	88	**46**	7	**47**	2.7
α+l	1	-7.7	-8.1	-29.1	91	5	**53**	**42**	0.0
α	1	-7.7	-8.0	-117.8	31	5	**53**	42	0.0
l	0								0

***β, α, l ,$<Z_0>$, $<Z_1>$, % of interactions, %β,%α,%l** - see the legend to Table I; **$<E_1-E_0>$** is the averaged energy in RT units[15-16] between the chain fragments and the whole chain; **%N_{tot}** is a percentage of the corresponding proteins;**Averaged data for all the observed proteins and those with the normal distribution of energy ($\chi 2<3.0$) are given in the first and second raws, respectively.

Table IV. Averaged characteristics of energy distributuins for proteins with Z>-4.5*

Combination of fragments	Number of proteins	$<Z_0>$	$<Z_1>$	$<E_1-E_0>$	% of interactioins	% β	% α	% l	% N_{tot}
β+α	22	-6.0	-3.7	23.4	29	**18**	**31**	52	9
β+l	25	-6.5	-3.9	46.8	35	**18**	41	**40**	11
a+l	62	-7.1	-3.7	110.2	41	37	**19**	**44**	27
β	69	-7.7	-3.4	70.9	9	**23**	32	46	31
α	165	-8.4	-3.1	122.8	14	27	**28**	45	71
l	189	-8.6	-3.3	167.3	19	25	34	**42**	79

***β, α, l ,$<Z_0>$, $<Z_1>$, $<E_1-E_0>$, % of contacts, %β,%α,%l,%N_{tot}** - see the legend to Tables I and III.

fragments form the majority of proteins of Table III (only one pure α-helical protein is presented in Table III). Consistently with this, the portion of fragments with β structure is the smallest in the high Z-score group of Table IV.

4 Conclusions

In this work we examined the role of secondary structure fragments in structure recognition. The results show that the overwhelming majority of protein structures can be successfully recognised by energies of interactions between residues of secondary structure. Thus, the secondary structure determines protein fold not only geometrically, but also energetically. We also found that the interactions between residues of β-structure are more specific in structure recognition than the interactions between residues of α-helices or loops. Generally, the accuracy of recognition correlates with the energy of a fragment. The chain fragments with lower energy appear to be more specific in recognition. The accuracy of recognition does not depend directly on the number of interactions (as one could expect under assumptions of the Random Energy Model) indicating a significant correlation between specific interactions within protein globule.

In estimating the accuracy of fold recognition we examined the shapes of energy distributions and found that for the majority of proteins and chain fragments their energy distributions are well approximated by the normal law. The deviations from the normal law are observed more often for short fragments and for the fragments with relatively high Z-score values.

It is worth noting that loops virtually do not contribute to structure recognition for ~90% of proteins considered. We believe that this shows an original "division of labour' in proteins: β-strands and α-helices are mainly responsible for structure while loops are mainly responsible for function. (However for ~10% of considered proteins, loops cannot be ignored in structure recognition.)

These results also justify the usage of the secondary structure backbone [3-5] rather than of the whole chain in recognition of remote homologs by threading methods, because loops of homologs differ more than strands and helices and their contribution to structure recognition is relatively small.

Thus, this study shows that the secondary structure (α-helices and β-strands) forms a *core* sufficient for recognition of the native structure for the majority of proteins. Is such a core the *optimal* one for recognition of (i) the native structure; (ii) remote homologs? These questions remain to be addressed.

Acknowledgments

We are grateful to Drs. A.V.Finkelstein, I.M.Gelfand and A.E.Kister for useful discussions.

References

1. H. Flockner, M.Braxenthaler, P. Lackner, M. Jaritz, M. Ortner, M. Sippl, *Proteins* **23**:376 (1995)
2. C.M.R.Lemer, M.J.Rooman,. S.J.Wodak, *Proteins* **23**:337 (1995)
3. S.Bryant, C.Lawrence, *Proteins* **16**:92 (1993)
4. T.Madei, J-F.Gilbrat, S.Bryant, *Proteins* **23**:356 (1995)
5. S. Bryant, *Proteins* **26**:172(1996)
6. D. Jones, J. Thornton, *Curr. Opin. Struct.Biol.* **6**:210 (1996)
7. D.Fisher, D.Rice, J. Bowie, D. Eisenberg, *FASB J.*, **10**:126 (1996).
8. Jaroszewski, L. Rychlewski, B. Zhang and A. Godzik, *Protein Science*, **7:**1431 (1998)
9. A.Finkelstein, *Curr. Opin. Struct.Biol.*, **7:**60 (1997)
10. J.Greer, *J.Mol.Biol.* **153**:1027 (1981)
11. J.Greer, *Proteins* **7**:317 (1990)
12. R.B.Russel, G.J.Barton, *J. Mol. Biol.* **244**: 332 (1994)
13. M. Hendlich, P.Lackner, S. Weitckus, H. Flokner, S. Froschauer,S. K. Gottsbacher, G. Casari, M. Sippl, *J.Mol.Biol.*, **216**:167 (1990)
14. U. Hobohm, M. Scharf, R. Schneider, C. Sander, *Protein Sci*; **1**:409 (1992)
15. B. Reva, A. Finkelstein, J. Skolnick, *Folding & Design*, 1998; **3**:141 (1998)
16. A.Finkelstein, A.Badretdinov, A.Gutin, *Proteins* **23**:142 (1995)
17. B. Reva, A. Finkelstein, M. Sanner, A. Olson, *Protein Eng*., **10**:865 (1997)
18. B. Reva, A. Finkelstein, J. Skolnick, *Protein structure prediction methods and protocols.* Totowa, NJ: Humana Press Inc. in press (1999)
19. B. Derrida, *Phys.Rev.B* **24**:2613 (1981)
20. J. Mathews and B. Walker, *Mathematical methods in physics*. W.A.Benjamin Inc.New York, (1964