Pacific Symposium on Biocomputing 4:278-289 (1999)

USING INFORMATION THEORY TO DISCOVER SIDE CHAIN ROTAMER CLASSES: ANALYSIS OF THE EFFECTS OF LOCAL BACKBONE STRUCTURE

JACQUELYN S. FETROW
Department of Molecular Biology
The Scripps Research Institute
La Jolla, CA 92037, USA

GEORGE BERG
Department of Computer Science
University at Albany, SUNY
Albany, NY 12222, USA

An understanding of the regularities in the side chain conformations of proteins and how these are related to local backbone structures is important for protein modeling and design. Previous work using regular secondary structures and regular divisions of the backbone dihedral angle data has shown that these rotamers are sensitive to the protein's local backbone conformation. In this preliminary study, we demonstrate a method for combining a more general backbone structure model with an objective clustering algorithm to investigate the effects of backbone structures on side chain rotamer classes and distributions. For the local structure classification, we use the Structural Building Blocks (SBB) categories, which represent all types of secondary structure, including regular structures, capping structures, and loops. For classification of side chain data, we use Minimum Message Length (MML) clustering from information theory. We show an example of how MML clustering on data classified by backbone SBBs can reveal different distributions of rotamer classes among the SBBs. Using these preliminary results, some of the characteristics of a rotamer library created using MML clustering on SBB dependent rotamer data are demonstrated.

1. Introduction

The regularities in the residue side chain conformations in proteins are known as rotamers[1]. Typically, though not always, the side chain rotamers for each amino acid are those conformations that are most energetically favorable[2]. Rotamers simplify the task of modeling, predicting and analyzing protein side chains by reducing the complexity of searching and representing the otherwise continuous conformational space of each side chain[1,3-5].

Because the amino acid side chain interacts with the protein backbone, the distribution of the rotamers for an amino acid will depend not only on the identity of the amino acid, but also on its backbone conformation. Incorporating regular secondary structure and related backbone information into a rotamer library improves the usefulness of rotamers in modeling and design projects[6-8].

In this paper, we introduce techniques to evaluate the effects of general protein backbone structure on rotamer distributions. To effectively describe the backbone conformations of both regular and non-regular secondary structures, we utilize the Structural Building Block (SBB) model of local structure[9,10]. SBB classes, described in more detail below, are a general representation of local structures. The advantages of using the SBB model to develop a backbone-dependent rotamer library include: 1) coverage of all backbone conformations including capping structures, loops, and turns; 2) a manageable number (six) of SBB categories; and 3) no arbitrary cutoffs in conformational space. Since the SBB classes represent all

backbone conformations, using them to build a backbone-dependent rotamer library has the potential to reveal interesting and useful information about the general effect of different backbone structure on the regularities of side chain conformations.

For the rotamer classes, we have used Minimum Message Length (MML)[11,12] from information theory[13] to discover the conformation clusters. MML provides an objective method of finding rotamer categories based on the clustering of the side chain conformation data.

In this preliminary study, we briefly describe MML clustering and SBB structural categories. We then describe the application of MML to developing a backbone-dependent rotamer library based on the SBBs. This preliminary study allows us to explore the application of MML theory to describe side chain groups classified by SBB structural classes. In the results, the analysis of one residue, isoleucine, is described in detail. By finding additional rotamer categories and distributions that are dependent on the backbone environment, the rotamer model is made more specific. The preliminary data are encouraging and we can now investigate our rotamer library's value in the prediction of side chain conformations in protein modeling, and as a tool in protein structure and function analysis.

2. Background and previous work

2.1 Rotamer libraries

Ponder and Richards[1] used rotamers to represent and predict side chain positions in protein cores. They created a rotamer library where the entry for each amino acid contains the side chain conformational information for that amino acid. Rotamer conformations are given as representative side chain dihedral angles, χ_i. For each rotamer, the probability that the amino acid side chain will be found in that conformation is given. In the prediction task, side chains are modeled on backbone protein models using the rotamer library and an algorithm to enforce overall quality criteria for the resulting model. Many different algorithms have been used to incorporate rotamer libraries into protein modeling tools (summarized by Vasquez[14]).

Many rotamer modeling algorithms use a single table entry of rotamer categories and distributions for each amino acid. However, Janin et al.[2] observed that the rotamer distribution changes with the backbone conformation of the residue. Elaborating on this observation, several researchers have created rotamer libraries and rotamer modeling algorithms that are backbone-dependent[6-8]. In these approaches, there is a separate rotamer library entry for each different backbone conformation.

The question then becomes how to adequately describe a protein's backbone conformation. One early approach[6] examined only α helices and β strands, as defined by the DSSP program[15]. Another group used generalized regions of regular secondary structure from a backbone φ and ψ dihedral angle (Ramachandran) map[8]. A third method divided the Ramachandran map into 20° by 20° sections[7]. These approaches either ignore non-regular backbone structure regions, or represent them in a discontinuous fashion. Thus, none of these descriptions is fully adequate for describing the continuous range of non-regular secondary structures in proteins[9].

There is an additional difficulty in constructing rotamer libraries. While the fundamental definition of a rotamer is an energetically favorable side chain conformation[2], in practice some amino acids have significant clusters of side chain conformations that are not energetically favorable[3,8]. If a rotamer library were restricted to energetically favorable rotamer classes, the energy minima would provide a principled definition for the set of classes. However, when defining rotamers based on actual conformational data, there is no such clear-cut criterion, which has led to great variation in the definition of categories in various rotamer libraries[3,8].

2.2 The SBB classification of backbone structures

In contrast to previous work, the backbone environment used in this study is the SBB classification of local protein structure[9]. The SBB classes are based on the regularity of C_α atom distances and angles in a seven-residue segment of protein. The classes themselves are discovered automatically. The distance and angle data for each segment are encoded as a length 43 vector. These vectors are then used to train an autoassociative artificial neural network[16]. The hidden layer representations of the network, which are of length eight, are used as a reduced representation of the data and these vectors are then clustered using k-means. Details of the process can be found in Zhang et al.[10]. The most general results are found when the vectors are clustered into six clusters. These clusters are used as the SBB local backbone structure categories, α, β, ζ, η, τ, ι[9].

After the automated clustering procedure, the backbone segments corresponding to these six clusters were analyzed and it was discovered that they corresponded to both known, and in some cases previously unknown, protein backbone classifications[9]. Two SBB classes were found to represent the regular secondary structures – α-helices (SBB α) and β-strands (SBB β). In addition, the other four SBB categories correspond almost uniquely to the capping structures for the helices (N-cap SBB ζ; C-cap SBB η) and strands (N-cap SBB τ; C-cap SBB ι). All six SBB classes are also found in the non-regular structures or "loop" regions of proteins. In the capping structures and in the loops, the SBBs are often found in recurring patterns, which describe distinct, local structures that can be analyzed for

specific hydrogen bonding patterns and unique interactions[9]. The library of SBB classifications for the database of proteins described here can be found at http://www.cs.albany.edu/compbio.

Use of the SBB classification of local backbone structure for a backbone-dependent rotamer library provides advantages over previous methods. First, it avoids the discontinuities inherent in simply representing backbone geometry by dividing the φ-ψ map into arbitrary 20° by 20° blocks[7]. Second, it provides a more realistic representation of the non-regular secondary structures than division of the backbone conformational space based on the regular secondary structures[8]. The six SBB classifications provide a relatively smaller number of categories compared to the work of Dunbrack[7] or some other methods for automatic recognition of backbone conformations[17,18], thus reducing the number of low population, statistically problematic categories.

2.3 Minimum Message Length (MML) Classification

MML classification[11,12] is a concept from information theory. Simplifying somewhat, it can be viewed as inference based on the Minimum Description Length (MDL) principle[13]. In MDL, an encoding is defined for a set of data so that the expected size of the message needed to transmit a data item selected at random is minimized. In MML, the minimal encoding is produced using knowledge of the structure and regularities in the data. Thus, the MML encoding can be used as an automatically generated, objective clustering of the data.

However, in order to use MML for dihedral angle values, a representation is needed that can model the circular nature of the data. The von Mises distribution is a generalization of the Normal distribution suitable for circular data. It models the continuous and circular properties found in data such as dihedral angles[19]. Thus, MML theory combined with the von Mises distribution can be used automatically to cluster side chain conformations based on their SBB backbone classifications.

2.4 Related Work

Dowe et al. used MML and the von Mises distribution to find distinct protein backbone structures, using φ and ψ angle data[18]. The MML principle provided an objective method of finding the clusters in the data. The von Mises distribution effectively modeled the circular nature of the dihedral angle backbone data. Using these techniques they found 27 distinct backbone classes. In subsequent work[20], they augmented their analysis with a Hidden Markov Model, and refined their number of classes to 17.

In related work, Thompson and Goldstein used mutual information to find substitution classes for amino acids, where the classes were optimized for

substitutions that did not change a protein's local structure[21]. They used these classes in an algorithm to predict solvent accessibility in proteins[22].

3. Application of MML to side chain rotamers classified by backbone geometry

3.1 The database

In order to do MML classification of rotamers it was necessary to first construct a high quality database of protein structures. The OBSTRUCT Internet server[23] was used to obtain a list of protein structures from the Protein Databank (PDB)[24]. Protein structures were limited to those with a resolution of 2.0 Å or better, 30% sequence identity or less, no NMR structures, and no structures using only the C_α backbone.

The list of PDB structures was subsequently examined. Structures without refinement R-values in their PDB files were removed. Non-globular proteins or domains (e.g. fibrous or membrane structures) were removed. Also any entries with abnormalities that compromised the side chain data (e.g. large regions that were modeled rather than built from experimental data) were removed from the list. A final list constructed according to these criteria contained 339 protein structures. This list can be found at http://www.cs.albany.edu/compbio/sbb/rotamers.

To minimize the effect of solvent on side chains and to further limit the number of modeled side chains in the data, the database was filtered to remove any residues with more than 5.0 Å of solvent accessible surface area. The solvent accessible surface area was measured by the DSSP algorithm[15].

The rotamer analysis was performed on 17 amino acids. The side chains for Ala, Gly and Pro have no rotatable side chain dihedral angles and thus cannot be modeled by rotamers. For this preliminary study, cystine and cysteine residues are combined. This conflates two possibly distinct conformational distributions. The resulting high-quality database for the 17 amino acid types contained 13,939 residues.

For most of the amino acids, the side chain dihedral data were used directly. However, to account for symmetry in their side chains, χ_2 values for several amino acids were adjusted [7]. For Asn and Asp, 180° was added to all χ_2 values less than –90°, and 180° was subtracted from values greater than +90°. This mapped the χ_2 angles into the range [-90°, +90°]. For His, Phe and Tyr, all χ_2 values less than 0° have 180° added to them to bring all values into the range [0°, 180°].

3.2 Application of MML clustering to discover side chain rotamers

The Snob program[18] is an implementation of MML classification that supports the von Mises distribution[19]. The FORTRAN software was downloaded from the ftp site given by Dowe et al. and compiled for use in this research. Separate Snob clusterings were made for each amino acid and SBB class combination. Each resulting clustering was a backbone-dependent rotamer library entry. For comparison, Snob was also run on all dihedral angle data for each amino acid, thus producing a backbone-independent rotamer library.

Clustering was done on χ_1 and χ_2 side chain dihedral angle data for all amino acids, except for Cys, Ser, Thr and Val, which have only χ_1 data. Since the χ_2 angles for Asn, Asp, His, Phe and Tyr are not in the expected range [-180°, +180°], these χ_2 angle values were linearly transformed into this range for use by Snob. Following clustering, the results were mapped back into the amino acids' specified χ_2 angle ranges.

An error term must be supplied to the Snob program since the precision of the data affects its information content, which affects the MML clustering[18]. At the 2.0 Å resolution of the database, the precision of the dihedral angles should be within 10°[8]. For this reason, the error term was set to 10° in the Snob runs.

Snob was run ten times for each amino acid/SBB classification data set, to find a minimal or near minimal clustering. Each run was started with a different randomly chosen set of nine clusters, and allowed to continue until convergence, or a maximum of 100 iterations. No manual intervention was done on the runs. The low cost run, as measured by the number of nits ($log_2\ e$ times the number of bits)[18] required to represent the data, was chosen as the clustering for the entry.

In general, Snob returns the mean χ_1 and χ_2 values for the clusters it finds, as well as measures of the cluster's concentration, κ_1 and κ_2, along both dimensions, and the number of residues in the cluster.

4. Results: Preliminary analysis of the rotamer classes

Clustering side chain conformations into rotamers based on an SBB classification of the backbone uncovers many interesting details. While a complete statistical and conformational analysis of the results is beyond the scope of this paper, we show some overall results, as well as detailed results for isoleucine. These data illustrate the advantages and limitations of the approach described here.

The number of rotamer classes found for each entry in the SBB backbone-dependent rotamer library is given in Table 1. Also given is the number of rotamer classes found for the backbone-independent library ("All SBBs"). As suggested by Janin[2] and confirmed by others[6-8], the overall rotamer distribution for a given

Table 1. The number of MML-discovered rotamer classes for each amino acid and amino acid/SBB combination.

Amino	Number of rotamer classes						
Acid	All SBBs	SBB α	SBB β	SBB η	SBB τ	SBB ζ	SBB ι
Arg	6	4	3	1	3	1	2
Asn	3	3	3	3	3	3	3
Asp	4	3	4	2	3	3	3
Cys*	4	3	3	3	3	2	3
Gln	7	4	6	1	3	2	1
Glu	8	5	4	3	2	3	3
His	3	2	3	1	3	1	3
Ile	9	6	7	3	5	5	4
Leu	13	8	6	4	4	3	4
Lys	5	2	3	1	1	1	1
Met	8	6	6	3	4	3	4
Phe	6	4	4	2	3	3	3
Ser	4	4	4	3	3	3	3
Thr	4	3	4	2	3	3	3
Trp	6	5	5	1	3	1	2
Tyr	4	4	4	2	3	3	3
Val	4	3	4	3	4	3	3

* Cysteine and cystine residues were combined into one category.

amino acid is actually the combination of different rotamer distributions for different backbone conformations.

In many cases MML clustering finds the standard, canonical rotamer classes. For the backbone-independent clustering of Ile, MML finds five of the nine canonical, energetically favorable rotamers: (g+,g+), (g+,t), (g-,t), (t,g-), and (t,t) (where the first symbol refers to χ_1 and the second symbol refers to χ_2, and g+ is *gauche*+, corresponding to angles around –60°, g- is *gauche*-, corresponding to angles around +60°, and t is *trans*, corresponding to angles around 180°[2]) (Figure 1). The other four canonical rotamer classes are not populated by Ile residues. The missing classes are indicated by O's in Figure 1. The g- conformation of χ_1 is generally avoided in Ile because of the β-branching of the isoleucine residue and this agrees with data described by others[1,3,7,8].

MML also finds clusters in the side chain conformation data that do not correspond to the canonical rotamers. For instance, in Ile clusters are found around (-95°,g-), (g-,93°) and (g+,150°) (Figure 1). In general, these rotamer classes are not found in other rotamer libraries, although a few correspond to non-canonical

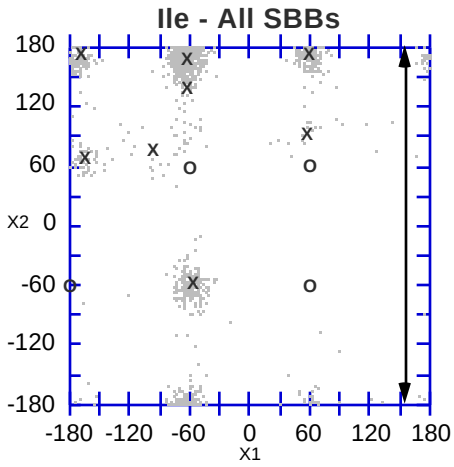

Figure 1. The distribution of χ_1 and χ_2 data and backbone independent MML rotamers for the amino acid isoleucine. The χ_1 and χ_2 values for individual residues in the database are indicated by dots. Rotamer classes discovered by MML clustering are indicated by X's and classes described by MML along only one dimension are indicated by arrows. Canonical rotamer classes that are not populated by isoleucine residues are indicated by O's.

rotamers found by others. For example, (g-, 93°) is near a rotamer identified by Schrauber at (61°,100°)[8].

Snob also finds a cluster at $\chi_1 = 153°$. While in some instances, described later, such one dimensional clusters can represent uniform distributions of data along the second dimension, more often they are symptomatic of the clustering algorithm trying to handle sparse or scattered residue data. This is the case for the one-dimensional clustering in the backbone independent library entry for Ile (Figure 1).

Figures 2a-f show the distributions of χ_1 and χ_2 data found by MML clusterings of Ile residues in each of the SBB classes. The results for Ile are representative of other well-populated amino acids in our database, but complete analysis of all residues is beyond the scope of this paper. A summary of all data is presented in Table 1.

The MML clusterings find distinctly different clusters for Ile, both among the individual SBBs (compare Figures 2a-2f), and when each SBB distribution is compared to the backbone-independent distribution (compare Figure 1 to Figures 2a-2f). While in a few cases, this can be attributed to sparse data (discussed below), most of the SBB classes in Ile are well enough populated that the clustering shows distinct differences in the side chain conformations exhibited by Ile in different

backbone conformations. For example, while the (t,t) conformation is well populated in general, it is almost completely absent in SBB η and SBB ι. The locations of these unpopulated rotamer classes are indicated by O's in Figures 2d and 2f. In general, other work on backbone dependent rotamer libraries[6-8] detects only differences corresponding to our results for SBB classes α and β (data not shown).

As in the backbone-independent case, the MML clustering finds non-canonical rotamers for the SBB-classified data. For instance, in SBB α, a cluster is found around (-109°,84°) (Figure 2a). This cluster would appear to be an amalgamation of the canonical rotamer (g+,g-) and the (–95°,g-) conformation described above for the backbone-independent data. Neither the canonical (g+,g-) nor the non-canonical (–95°,g+) cluster is found in any of the other SBB categories. While statistical analysis of this clustering is beyond the scope of this paper, 32% of all the data points in the backbone-independent dataset are found in the SBB α category, suggesting that either the (g+,g-) or the (–95°,g-) conformations are only found when the Ile backbone is in the α−helical conformation in proteins.

The MML clustering results are clearly sensitive to sparse data. For SBB ζ MML finds a single "rotamer" at (-165°,115°). However, there is no residue data

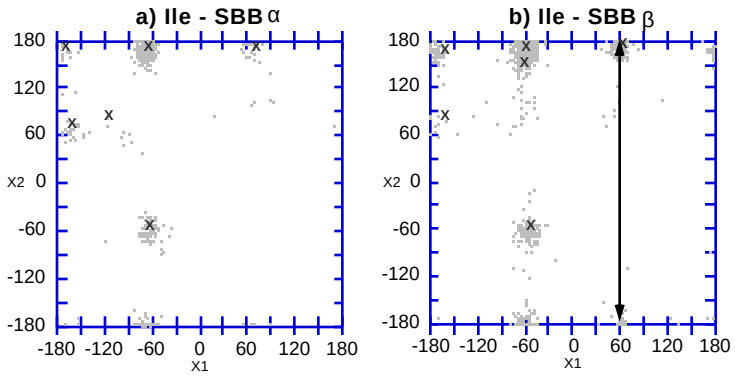

Figure 2. The distribution of χ_1 and χ_2 data and backbone dependent MML rotamers for the amino acid isoleucine. The χ_1 and χ_2 values for individual residues in the database are indicated by dots. Rotamer classes discovered by MML clustering are indicated by X's and classes described by MML along only one dimension are indicated by arrows. Unpopulated instances of the rotamer class (*trans,trans*) are indicated by O's. (a) rotamer classes and data for residues classified as SBB α (helix) (b) SBB β (strand) (c) SBB ζ (d) SBB η (e) SBBτ (f) SBB ι.

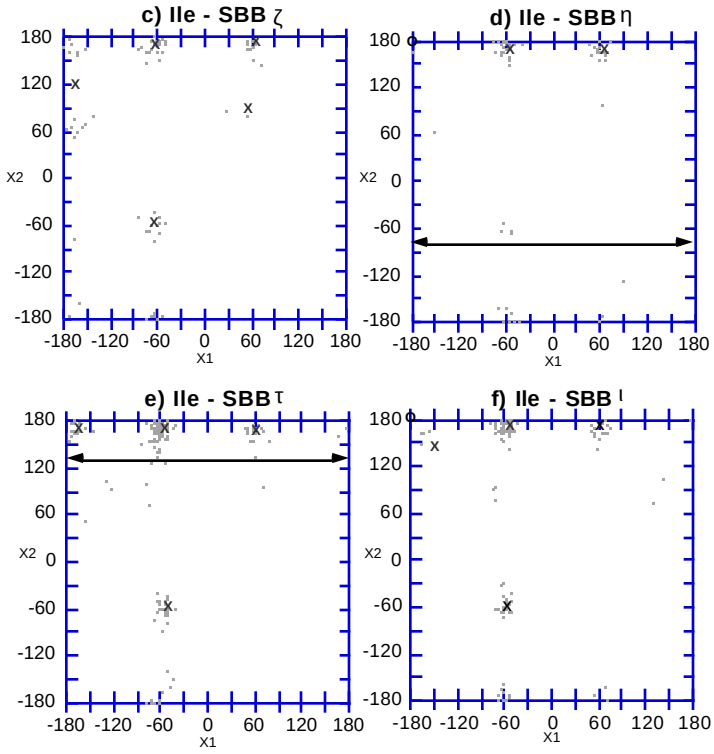

Figure 2. (Continued)

near this position. Rather, this purported rotamer conflates two sparsely populated clusters of residues at (t,t) and (t,g-) (Figure 2c).

Also, MML occasionally finds a cluster using only one dimension (Figures 1, 2b, 2d and 2e). For some amino acids this can reflect continuous distributions along one dimension. Examples of this are one-dimensional clusters for Asn (occurring in the clusterings of - All SBBs, SBB η, SBB ι, and SBB β) and Asp (All SBBs, SBB ζ, SBB ι, and SBB β) (data not shown) that correspond to continuous χ_2 distributions reported in the literature[8]. But, more often one-dimensional clusters occur when there are sparse, widely dispersed data. In these cases, the clusters do not reflect the distribution of the data, but are rather an artifact of the clustering algorithm. The one-dimensional clusterings in Figures 1, 2b, 2d and 2e are examples of this.

5. Conclusions and future work

The results in this paper show that MML clustering and SBB local structure classifications can be combined to create a backbone-dependent rotamer library. In many cases these rotamer classes show differences in the distribution of rotamers for different SBB classes. It has been known that regular secondary structures have an effect on rotamer class distributions, but the results here show this effect also generalizes to non-regular backbone structures. In addition, in some cases there are rotamer classes that are unique to specific amino acid/SBB combinations, and completely absent in the other entries for the same amino acid. These results show that using a more general backbone structure model reveals additional dependence of side chain conformations on a residue's backbone environment.

To extend this preliminary work, we will first relax the restrictions on allowed sequence identity, greatly increasing the database size. This will alleviate the scarcity of data in some of the lightly populated rotamer library entries. In the cases where there is still too little data in library entries, statistical back-off techniques can be used to build approximate entries[25]. The two Cys populations will be separated and rotamer libraries created for each of them. Once this is done a new library can be created, and we can then analyze the side chain distributions for each amino acid in detail. A variety of factors, such as variable solvent accessibility of residues can be analyzed to assess their effects on the rotamer libraries.

Additional work can be done to evaluate the quality of the MML rotamers. The κ concentration parameters can be evaluated to distinguish those purported rotamers that truly represent residue data clusters. Appropriate statistical techniques can be used to determine the statistical significance of the difference in rotamer distributions among SBB categories.

Traditionally, rotamer analysis has been restricted to single residues. However, since SBB patterns along the protein backbone have been found to correlate with specific structures[9], the SBB model is ideally suited for finding rotamer distributions for entire local structures. Patterns of SBBs along the protein backbone will be analyzed in this fashion to discover rotamer populations that occur in specific protein structures.

All of this information will be used to improve protein modeling, design, and prediction tools. Rotamer classes for specific backbone conformations, as identified by SBB patterns, may also facilitate the use of rotamers as a tool for analyzing structure and function in proteins.

References

1. Ponder, J.W. & Richards, F.M. *J Mol Biol* **193**, 775-791 (1987).
2. Janin, J. & Wodak, S. *J Mol Biol* **125**, 357-386 (1978).

3. Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. *J Biomol Struct Dyn* **8**, 1267-1289 (1991).
4. Shenkin, P.S., Farid, H. & Fetrow, J.S. *Proteins* **26**, 323-352 (1996).
5. Doig, A.J., MacArthur, M.W., Stapley, B.J. & Thornton, J.M. *Protein Sci* **6**, 147-155 (1997).
6. McGregor, M.J., Islam, S.A. & Sternberg, M.J. *J Mol Biol* **198**, 295-310 (1987).
7. Dunbrack, R.L., Jr. & Karplus, M. *J Mol Biol* **230**, 543-574 (1993).
8. Schrauber, H., Eisenhaber, F. & Argos, P. *J Mol Biol* **230**, 592-612 (1993).
9. Fetrow, J.S., Palumbo, M.J. & Berg, G. *Proteins* **27**, 249-271 (1997).
10. Zhang, X., Fetrow, J. & Berg, G. in *Techniques in Protein Chemistry V* (ed. Crabb, J.) 397-404 (Academic Press, 1994).
11. Wallace, C.S. & Boulton, D.M. *Computer Journal* **11**, 185-194 (1968).
12. Wallace, C.S. & Dowe, D.L. in *6th International Workshop on Artificial Intelligence and Statistics* 529-536 (, 1997).
13. Cover, T.M. & Thomas, J.A. *Elements of Information Theory*, (John Wiley and Sons, New York, 1991).
14. Vasquez, M. *Curr Opin Struct Biol* **6**, 217-221 (1996).
15. Kabsch, W. & Sander, C. *Biopolymers* **22**, 2577-2637 (1983).
16. Rumelhart, D.E., Hinton, G. & Williams, R.J. in *Parallel Distributed Processing*, Vol. I (eds. Rumelhart, D.E. & McClelland, J.L.) 318-362 (MIT Press, Cambridge, MA, 1986).
17. Unger, R., Harel, D., Wherland, S. & Sussman, J. *Proteins* **5**, 355-373 (1989).
18. Dowe, D.L. *et al. Pac Symp Biocomput* **1**, 242-255 (1996).
19. Fisher, N.I. *Statistical Analysis of Circular Data*, (Cambridge University Press, Cambridge, 1993).
20. Edgoose, T., Allison, L. & Dowe, D.L. *Pac Symp Biocomput* **3**, 583-594 (1998).
21. Thompson, M.J. & Goldstein, R.A. *Proteins* **25**, 28-37 (1996).
22. Thompson, M.J. & Goldstein, R.A. *Proteins* **25**, 38-47 (1996).
23. Heringa, J., Sommerfeldt, H., Higgins, D. & Argos, P. *CABIOS* **8**, 599-600 (1992).
24. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. & Weng, J. *Protein Data Bank in crystallographic databases - Information content, software systems, scientific application*, 107-132 (Data Commision of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987).
25. Dunbrack, R.L., Jr. & Cohen, F.E. *Protein Sci* **6**, 1661-1681 (1997).