

# A priori and a posteriori mixture distributions for using databases in protein structure determination

Stanley L. Sclove<sup>a,\*</sup>, Simon A. Sherman<sup>b</sup>

<sup>a</sup>Information and Decision Sciences Dept. M/C 294, University of Illinois at Chicago, 601 S. Morgan St., Chicago, IL 60607-7124, USA

<sup>b</sup>Eppley Institute for Research in Cancer and Allied Diseases, University of Nebraska Medical Center, 600 South 42nd St., Omaha, NE 68198-6805, USA

Received 1 November 1996; accepted 10 April 1997

---

## Abstract

The twists and turns of protein molecules correspond to the rotational angles of the main and side chains in their constituent amino acid residues. Owing to stereochemical restrictions the joint distribution of these angles falls into several well-separated clusters. Consequently, we fit a statistical finite mixture model to data from the Brookhaven Protein Data Bank (PDB), obtaining a new classification of conformational states of amino acids. The results of this database modeling are important for knowledge-based approaches to protein structure determination and analysis and can be used in conjunction with experimental data in the determination of protein spatial structure. As part of this process, the mixture distributions we have fit can be used as a priori distributions. A posteriori distributions corresponding to these a priori mixture distributions are illustrated. © 1997 Elsevier Science B.V.

**Keywords:** Protein local structure determination; Cluster analysis; Statistical finite mixture model; A priori and a posteriori distributions

---

## 1. Introduction and summary

This paper is directed toward the further use of statistical methods in the study of protein structure. It is a companion to the papers in Refs. [1,2] (earlier work is in Ref. [3]) and discusses the procedures presented there from the viewpoint of mathematical statistics.

First we present our work on statistical description of the dihedral angle values in the Brookhaven Protein Data Bank (PDB) by means of finite mixture model cluster analysis. Then we discuss statistical methods

for applying this description to the assessment and refinement of alternative hypothesized protein structures. Refinement involves the use of the statistical mixture model as a prior distribution for use in Bayesian classification and estimation. The nature of the corresponding posterior distribution is discussed.

## 2. Background

A protein is a directed sequence (“chain”)  
 $R_1, R_2, \dots, R_j, \dots, R_n$

of amino-acid residues. A residue is the combining form of a molecule, i.e. the form it takes as part of a

---

\* Corresponding author.

macromolecule. Each  $R_i$  ( $i = 1, 2, \dots, n$  residues) is one of 20 amino acids, alanine (Ala), arginine (Arg), asparagine (Asn), aspartate (Asp), cysteine (Cys), glutamate (Glu), glutamine (Gln), glycine (Gly), histidine (His), isoleucine (Ile), leucine (Leu), lysine (Lys), methionine (Met), phenylalanine (Phe), proline (Pro), serine (Ser), threonine (Thr), tryptophan (Trp), tyrosine (Tyr), valine (Val).

*Amino acids.* An amino acid typically consists of a main chain and a side chain. The main chain or “backbone” is formed by two successive single bonds, nitrogen to carbon to carbon (N-C-C). The first carbon, the  $\alpha$  carbon, is denoted by  $C^\alpha$ ; the second, the carbonyl carbon, is denoted by  $C'$ . Let the index  $i$  range over the residues in a sequence;  $i = 1, 2, \dots, n$ . One can think of successive residues

$R_{i-1} R_i R_{i+1}$

in terms of

$\{N_{i-1}, C_{i-1}^\alpha, C'_{i-1}\}, \{N_i, C_i^\alpha, C'_i\},$

$\{N_{i+1}, C_{i+1}^\alpha, C'_{i+1}\}$

*Dihedral angles.* In general, a dihedral angle is an angle between planes. Given a system of four successive atoms A-B-C-D, a dihedral angle describes the relationship between the C-D and A-B bonds. This angle is the one between the plane determined by the bonds B-C and C-D (i.e. by the three atoms B, C and D) and the plane determined by the bonds A-B and B-C (i.e. by the atoms A, B, C). (The dihedral angles are also called torsion angles or rotational angles.)

The angle  $\phi$  is the dihedral angle of the system with  $A = C'_{i-1}$ ,  $B = N_i$ ,  $C = C_i^\alpha$ ,  $D = C_i$ .

The angle  $\psi$  is the angle of the system with  $A = N_i$ ,  $B = C_i^\alpha$ ,  $C = C_i$ ,  $D = N_{i+1}$ .

All but two of the amino acids (Ala, Gly, Pro) have a side chain with one or more angles of rotation. Pro has a side chain but its rotational angle is fixed. The individual amino acids differ according to these side chains. The side chain is linked to  $C^\alpha$  and begins with a carbon atom, denoted by  $C^\beta$ , typically followed by a  $\gamma$  atom,  $\delta$  atom, etc. Thus we may extend our description of the  $i$ th residue  $R_i$  to  $\{N_i, C_i^\alpha, C'_i, C_i^\beta\}$ . The side-chain dihedral angle  $\chi_1$  is defined as the angle between the plane determined by N,  $C^\alpha$  and  $C^\beta$  and that determined by  $C^\alpha$ ,  $C^\beta$  and the  $\gamma$  atom. The angle  $\chi_1$  is the first rotational angle in the side chain; successive such angles  $\chi_2, \chi_3$ , etc., are defined. Here we are concerned only with  $\chi_1$ . The three bivariate distributions are described to some extent in the literature, but the trivariate distribution of  $\phi$ ,  $\psi$  and  $\chi_1$  is of interest and is a subject of our investigation.

It is known that the distribution of  $\chi_1$  is tri-modal, with modes at about  $-60$ ,  $+60$  and  $180^\circ$ . (These modes are sometimes denoted by  $g^-$ ,  $g^+$  and  $t$ , but the notation is not consistent across all amino acids.) This trimodality is illustrated in Fig. 1, a dot plot of  $\chi_1$  for Val.

We analyze the conformations of 58 inhomologous proteins of high resolution ( $2 \text{ \AA}$  or better) from the Brookhaven PDB, which includes X-ray measurements of proteins in the crystalline state. Our data set consisted of several thousand observations from this data bank. Each case in the data set corresponds to a particular amino acid in a particular position in the sequence of a particular protein. The aim of the phase of our research reported here is the statistical modeling and description of this database of angular

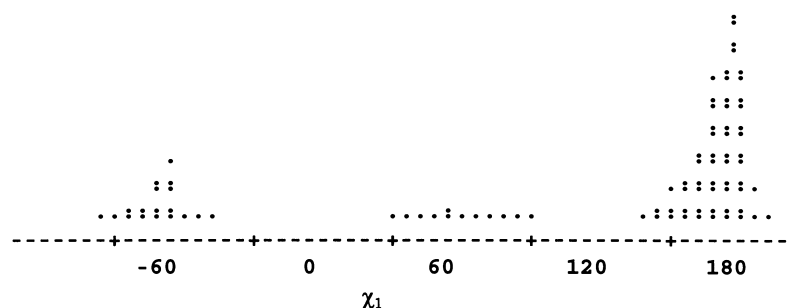


Fig. 1. Dot plot of  $\chi_1$  for Val. (Each dot represents ten points.)

values, to see how they cluster and how the clusters vary across amino acids. The result is a characterization, in terms of the statistical finite mixture model, of the joint distributions, for data in the PDB, of  $\phi$ ,  $\psi$  and  $\chi_1$  for the amino acids with side chains, and the joint distribution of  $\phi$  and  $\psi$  for the others. Dunbrack [4] and Dunbrack and Karplus [5,6] have also done extensive work to describe these distributions. In contrast to their work, which produces descriptive statistics for these data, we perform probabilistic modeling of the data, using the statistical finite mixture model to fit the clustered trivariate distribution of angular values.

The description of the database is interesting and useful in its own right. Also, the resulting parameter estimates are used in the determination of protein structure, as follows. For proteins in solution, NMR is used to obtain estimates of inter-atomic distances; from these, estimates of the dihedral angles are obtained. These estimates are imprecise. We supplement them with the additional information from the modeling of the data bank. The finite-mixture distribution fit to the X-ray data from the PDB is used as an a priori (“prior”) distribution. A given NMR measurement is classified into the one or two most likely states (clusters). The cluster mean is combined with the NMR measurements to provide improved a posteriori (“posterior”) estimates. A mathematical description of this procedure is given later in the paper. A description of the procedure and its computer implementation are given in the companion papers [1,2].

The fitting of distributions to the angular data in the PDB by finite mixture model cluster analysis is described next.

### 3. Finite mixture cluster analysis: the model and the method

The finite mixture model [7,8] was used to fit the clusters of points in angular space. The probability density function (p.d.f.) of a finite mixture distribution is of the form

$$f(\mathbf{x}) = p_1 f_1(\mathbf{x}) + p_2 f_2(\mathbf{x}) + \dots + p_c f_c(\mathbf{x})$$

where the mixing probabilities  $p_c$ ,  $c = 1, 2, \dots, C$ , are positive and sum to unity and the p.d.f.s  $f_c(\mathbf{x})$  are the components of the mixture. Often the components are assumed to belong to a specified parametric family,

e.g. Gaussian. The integer  $C$  is the number of components. The number of components  $C$  can be greater than the number of clusters (in the sense of modes) because two component densities close together might result in only one mode.

Fitting the model involves determining  $C$  and estimating the mixing probabilities and the distributional parameters.

In our case the variable  $\mathbf{x}$  is the vector of the three angles  $\phi$ ,  $\psi$ ,  $\chi_1$ . (Later we shall refer to this vector as  $\mathbf{A}$ , for “angles”.) In our work to date, the component distributions are taken to be trivariate normal with different covariance matrices as well as different means. Different covariance matrices are used because it is clear from steric contour plots (Ramachandran diagrams) and data plots that the clusters have different orientations. In two dimensions, a cluster running from southwest to northeast is indicative of positive correlation; from northwest to southeast, negative correlation. As an example, a plot of  $\psi$  vs.  $\phi$  for Val is shown in Fig. 2.

The E-M algorithm [7–10] was used for estimation. It is an iterative, maximum likelihood estimation procedure. In the finite-mixture model situation, it starts with preliminary estimates of the distributional parameters, computes posterior probabilities based on these, then uses the posterior probabilities to update the estimates of the distributional parameters, etc.

When a mixture of  $C$  distributions is to be fit, the parameters to be estimated are the  $C$  mixing probabilities and the distributional parameters for the  $C$  component distributions. When the components are multivariate Gaussian with  $p$  variables, the distributional parameters for each component are  $p$  means,  $p$  standard deviations and  $p(p-1)/2$  correlations. In our case there are three variables  $\phi$ ,  $\psi$  and  $\chi_1$ , so there are three means, three standard deviations and three correlations:  $\phi$  with  $\psi$ ,  $\phi$  with  $\chi_1$ , and  $\psi$  with  $\chi_1$ .

The procedure used is iterative, and hence requires initial values. Initial values for the means were suggested by the Ramachandran diagrams and from earlier work [11–13]. Ramachandran diagrams are plots in the  $(\phi, \psi)$ -plane showing which  $(\phi, \psi)$ -pairs are sterically allowed. A  $(\phi, \psi)$ -pair is disallowed if it would cause some pair of atoms to become too close to one another (e.g. see Ref. [14], pp. 258–260).

The initial values of the standard deviations were set at  $10^\circ$  for all three angles; the correlations were

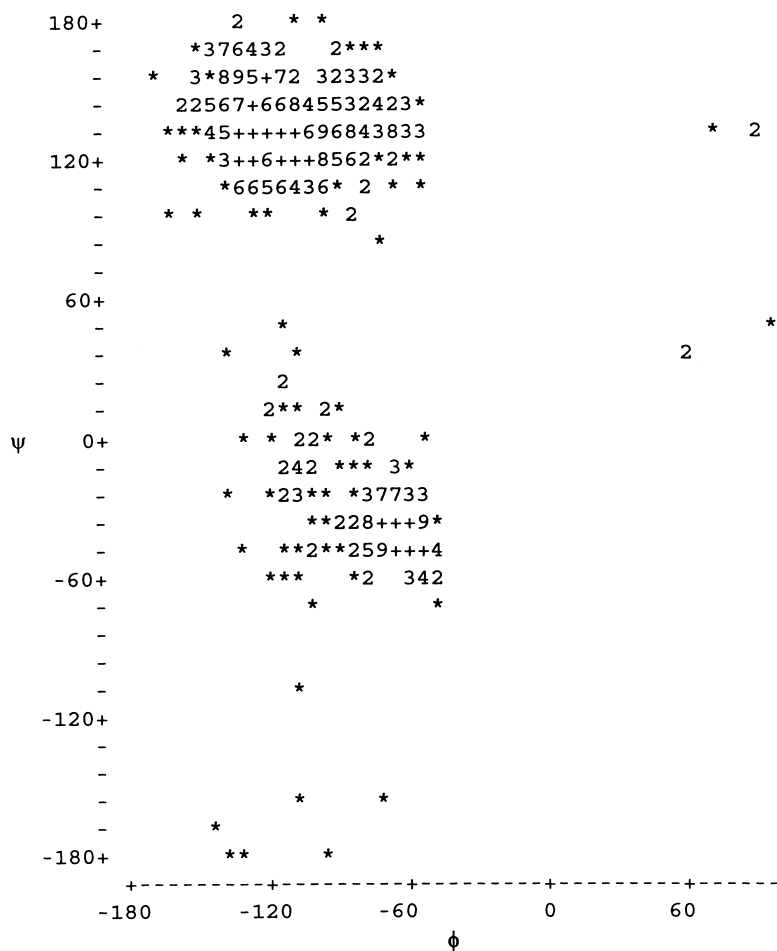


Fig. 2. Scatterplot of  $\psi$  vs.  $\phi$  for Val. Numbers are counts at each point (\* means 1, + means >9).

initialized at zero. To start, the values of the mixing probabilities were taken as equal, i.e. all equal to  $1/C$ .

## 4. Data pre-processing

### 4.1. Data coding

The data are angular. A recorded value of  $-179$  is the same as  $181$  even though their numerical values are far apart. Consequently, some folding was necessary.

It is known that  $\chi_1$  has three modes, near  $-60$ ,  $+60$ , and  $+180$  ( $= -180$ ). Values from  $-360$  to about  $-120$  were folded over to the corresponding positive values.

### 4.2. Further data coding

An interesting approach to the problem of angular data, not used here because it doubles the number of variables, is to use trigonometric coordinates, representing an angle  $\theta$  by the pair  $(\cos \theta, \sin \theta)$ . This would have resulted in six, rather than three, variables, nonlinearly related in pairs, so we did not opt for this procedure. Rather, since clustering was to be done anyway, the approach used was first to fit the clusters and then combine coinciding clusters by suitably recoding the data in one of them. For example, if there were a cluster C1 centered at  $(\phi, \psi, \chi_1) = (170, 180, 60)$ , say, and another C2 centered at  $(-190, -180, 60)$ , i.e. the same centers coded differently, then

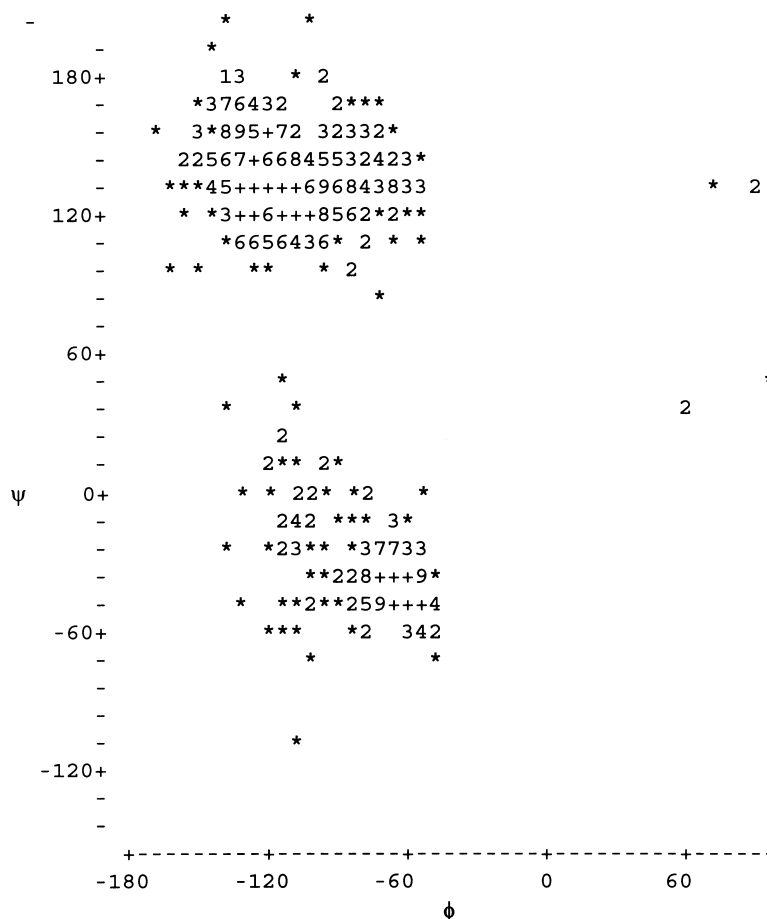


Fig. 3. Scatterplot of  $\psi$  vs.  $\phi$  for Val. Numbers are counts at each point (\* means 1, + means >9). Fig. 2 with portion near  $-180$  moved to  $+180$  to put together a cluster.

the points in C2 would be recoded according to  $\phi' = \phi + 360$ ,  $\psi' = \psi + 360$ ,  $\chi'_1 = \chi_1$ . Comparison of Fig. 3 with Fig. 2 illustrates this; the cluster at the upper left in Fig. 3 is formed from clusters at the upper left and lower left in Fig. 2. There is a number of clusters present; in fact, we fit a model with 16 clusters for the combined data set for all 17 amino acids with variable side-chain angles. There are different conventions in recording the data for the various amino acids. For example,  $120^\circ$  must be added to  $\chi_1$  measurements for Ile and Thr to make them comparable with those for the other amino acids.

## 5. Results of the cluster analysis

Mixture distributions were fit to the data for the 20 amino acids.

The computer programs used, implementing the E-M algorithm for multivariate Gaussian component distributions, were programmed by Sclove [15]. Similar programs are available elsewhere also (e.g. see Ref. [7]). The estimate of the mean vector for a given component distribution is a weighted average of all  $n$  observations, the weights being proportional to the estimate at that stage of the probability that the observation arose from that component, i.e. the conditional probability of component  $c$ , given observation  $i$ . The estimate of the covariance matrix for a given component can be interpreted similarly.

The clusters obtained admit interpretations in terms of six  $(\phi, \psi)$  configurations constituting structures (called B, P, R, T, L and E), taken in combination with the three modes for  $\chi_1$ . Our B denotes

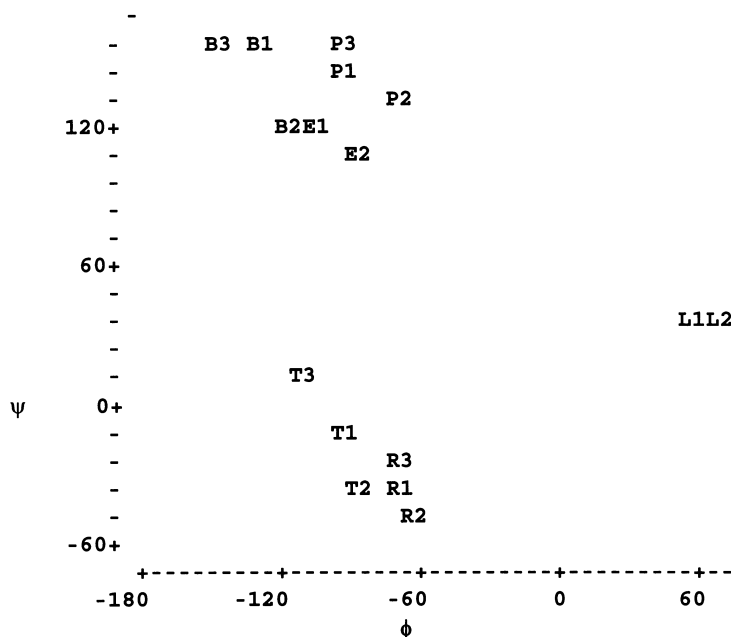


Fig. 4. The 16 cluster mean values. See columns 3 and 4 of Table 1 for the coordinates plotted (cluster mean values of  $f$  and  $y$ ). Code for the corresponding regular structures: B,  $\beta$ -structure; P, polyproline II helix; E, extended transition state; R, right-hand  $\alpha$ -helix; T, twisted transition state; L, left-hand  $\alpha$ -helix. Code for  $\chi_1$ : 1,  $\chi_1$  approx.  $-60^\circ$ ; 2,  $\chi_1$  approx.  $+180^\circ$ ; 3,  $\chi_1$  approx.  $+60^\circ$ . Example: ‘L2’ indicates that the mean of that cluster has values of  $\phi$  and  $\psi$  approximating those for a left-hand  $\alpha$ -helix and  $\chi_1$  is approximately  $180^\circ$ .

$\beta$ -structure; P, polyproline II helix; R, right-hand  $\alpha$ -helix; L, left-hand  $\alpha$ -helix; T, twisted transition state; E, extended transition state. These six, taken in combination with the three rotamers of  $\chi_1$ , would give 18 clusters, except two of these (L and E) cannot (or at least are very unlikely to) exist with  $\chi_1$  at  $+60^\circ$ . Within the six configurations there is a consistent pattern of  $(\phi, \psi)$  locations corresponding to the three  $\chi_1$  states, with  $+60$  to the upper left,  $180$  to the lower right and  $-60$  between these. See Fig. 4.

The results include estimates of the nine distributional parameters (means and variances of the three angles and covariances between the three pairs of angles) for each component distribution.

The probabilities across the 16 clusters for the combined data for the 17 amino acids with variable side-chain angles, as well as means, standard deviations and correlations, are given in Table 1. A negative correlation means that the corresponding elliptical cluster runs from northwest to southeast; a positive correlation, that the cluster runs from southwest to northeast.

The estimates of the mixing probabilities in the finite-mixture model differ across amino-acid residues;

i.e. different amino-acid residues have different propensities for the 16 alternative states. This can be seen from a cross-tabulation (not shown here) of amino acid by cluster, the row-percentages of which are the mixing probabilities for the corresponding amino acid across the final 16 component distributions.

For the three amino acids not having a variable side-chain angle (Ala, Gly, Pro) we fit the bivariate mixture distribution of  $(\phi, \psi)$ . See Table 2 (Pro), Table 3 (Ala) and Table 4 (Gly). Note that various correlational patterns were observed within clusters; e.g. sometimes the correlation of  $\phi$  with  $\psi$  is positive, sometimes negative, sometimes essentially zero, and similarly for the correlations of  $\phi$  with  $\chi_1$  and of  $\psi$  with  $\chi_1$ . There were earlier claims (e.g. see Ref. [16]) that the side-chain angle varies only slightly with the main-chain angles. In fact, within clusters, these angles are highly correlated for some amino acids.

## 6. Applications of the results of the cluster analysis

The procedures discussed here for determination

Table 1

Probabilities, means, standard deviations and correlations for the combined data of the 17 amino acids with variable side-chain angles ( $n = 6730$ )

Cluster	Prob.	Mean (deg)			Standard deviation (deg)			Correlations		
		$\phi$	$\psi$	$\chi_1$	$\phi$	$\psi$	$\chi_1$	$\phi, \psi$	$\phi, \chi_1$	$\psi, \chi_1$
B1	0.062	-129	153	-61	10	14	9	-0.2	-0.1	0.1
B2	0.160	-114	125	180	21	14	9	-0.2	0.0	-0.1
B3	0.051	-146	160	64	13	14	10	-0.3	0.1	0.0
P1	0.105	-89	148	-63	20	15	11	-0.1	-0.2	0.2
P2	0.038	-67	133	180	11	8	9	0.0	0.1	-0.1
P3	0.025	-91	157	66	21	19	10	-0.2	0.0	0.1
E1	0.029	-111	115	-61	16	26	10	-0.5	-0.1	0.2
E2	0.010	-84	103	185	8	19	7	0.1	0.3	-0.4
R1	0.124	-64	-37	-69	6	9	9	-0.4	0.0	0.0
R2	0.142	-62	-44	178	7	6	10	-0.3	0.0	0.0
R3	0.047	-69	-22	67	11	13	11	-0.7	0.1	0.0
T1	0.139	-92	-12	-63	19	22	10	-0.5	-0.1	0.0
T2	0.029	-84	-39	179	23	21	11	0.1	0.1	-0.1
T3	0.022	-109	8	63	21	30	11	-0.2	0.2	-0.1
L1	0.017	59	36	-62	8	13	11	-0.6	0.0	0.0
L2	0.002	65	39	190	18	11	6	-0.4	0.0	-0.2

and refinement of protein structures are bottom-up procedures, working at the atomic level within successive residues. The work can be done in terms of atomic coordinates on the one hand or the dihedral angles on the other hand. Our work is in terms of the angles, though the mathematical formulation here could be applied similarly to coordinates.

Procedures for protein structure determination involve combining experimental data with the existing knowledge base, the database. Thus the mixture-distribution fit to the database has applications in the

determination and refinement of protein structures. One way to do this is simply to use the results of the cluster analysis to suggest starting values for optimization of a target function. (The “target function” includes an energy function and penalty terms for violation of constraints provided by other experimental data.) Another way to use the results is to use the Bayesian statistical paradigm to combine the information in the database with experimental information. The mixture-distribution is taken as a prior distribution and combined with the experimental

Table 2

Probabilities, means, standard deviations and correlations for Pro ( $n = 446$ )

Cluster	Prob.	Mean (deg)		Standard deviation (deg)		Correlation between $\phi$ and $\psi$
		$\phi$	$\psi$	$\phi$	$\psi$	
1	0.399	-63	-28	11	14	-0.6
2	0.022	-76	58	9	12	-0.8
3	0.115	-73	138	13	20	-0.3
4	0.438	-64	148	8	12	-0.5
5	0.026	-75	180	10	21	0.0

Table 3  
Probabilities, means, standard deviations and correlations for Ala ( $n = 923$ )

Cluster	Prob.	Mean (deg)		Standard deviation (deg)		Correlation between $\phi$ and $\psi$
		$\phi$	$\psi$	$\phi$	$\psi$	
1	0.200	-131	142	25	20	-0.5
2	0.148	-71	143	14	14	-0.5
3	0.180	-83	-13	25	30	-0.7
4	0.427	-63	-38	6	9	-0.4
5	0.028	57	19	22	72	-0.1
6	0.013	-128	-135	29	41	0.3
7	0.003	162	162	12	15	-0.1

data. Note that in this application of the Bayesian paradigm, there is no subjective element: the prior distribution is obtained from empirical data, i.e. the existing database.

### 6.1. The Bayesian paradigm

Let the vector  $A$  denote the true, unknown value of the dihedral angles at a given residue. The vector  $A$  is an unknown parameter, to be estimated. Experimental data  $x$ , bearing upon the value  $A$ , will be obtained. These experimental data will be combined with the prior knowledge about  $A$ . The p.d.f. of the prior distribution will be denoted by  $f(A)$ . It is this density which has been modeled according to the finite-mixture model. The prior distribution is combined with the experimental data according to the Bayes'

formula to obtain posterior probabilities. The experimental data  $x$  can be either discrete or continuous.

### 6.2. Discrete experimental data

When the experimental data vector  $x$  is discrete, it means  $x$  takes on, say,  $m$  distinct values,  $v_1, v_2, \dots, v_m$ .

Much of the work reported in the companion papers [1,2] is in terms of categorical variables, the “ $d$ -connectivities”. Each such  $d$  is a function of the distance between two protons; e.g.  $d_{N\alpha}$  relates to the protons in the nitrogen and  $\alpha$ -carbon atoms. The  $d$  values are categorical variables indicating the absence or presence and strength of a “cross-peak”, a peak in the bispectrum of the two protons.

We consider first an oversimplified example. Suppose there were three  $d$  values (corresponding to

Table 4  
Probabilities, means, standard deviations and correlations for Gly ( $n = 949$ )

Cluster	Prob.	Mean (deg)		Standard deviation (deg)		Correlation between $\phi$ and $\psi$
		$\phi$	$\psi$	$\phi$	$\psi$	
1	0.027	-162	-164	12	11	0.2
2	0.026	154	172	16	6	0.3
3	0.061	-158	165	12	12	-0.2
4	0.040	161	-163	17	13	-0.2
5	0.044	-100	-158	20	20	-0.4
6	0.061	100	150	20	23	-0.4
7	0.094	-85	-7	25	35	-0.5
8	0.107	95	0	20	25	-0.2
9	0.087	-82	154	19	17	0.0
10	0.101	73	-145	15	20	-0.5
11	0.096	-62	-41	6	8	-0.3
12	0.258	80	9	13	19	-0.9

three pairs of protons),  $d_1, d_2, d_3$ . Then the data vector  $\mathbf{x} = (d_1 d_2 d_3)$ . Suppose further that each  $d$  is either 1 or 0, corresponding to the presence or absence of a cross-peak. Then the vector  $\mathbf{x}$  can have  $m = 8$  possible values (patterns), (1 1 1), (1 1 0), (1 0 1), (0 1 1), (1 0 0), (0 1 0), (0 0 1), and (0 0 0). Note that to discriminate among the 16 possible states, these eight values would have to be supplemented with other information.

In the actual application by program FiSiNOE reported earlier in Ref. [3], the  $d$  values are not coded simply as 1 or 0 but rather are semi-quantitatively coded on an ordinal scale as absent, weak, medium, or strong. Further, there are not just three  $d$  values but nine, five intra-residue  $d$  values and four sequential (inter-residue)  $d$  values. Also, there are three coupling constants (in hertz), coded as strong, medium, weak, and absent/very weak. Now,  $\mathbf{x}$  represents such data at a given site. Given  $\mathbf{x}$ , it is possible to say which components  $c$  are compatible with it. Typically, there will be two or three such components, but sometimes only one. Suppose there are three matching components, and, according to the fitted mixture model, these have prior probabilities ( $p_c$  values) equal to 0.20, 0.12, and 0.08. Then the posterior probabilities of these components are obtained by normalization as  $0.20/(0.20 + 0.12 + 0.08) = 0.5$ ,  $0.12/(0.20 + 0.12 + 0.08) = 0.3$ , and  $0.08/(0.20 + 0.12 + 0.08) = 0.2$ .

A more formal procedure for obtaining these posterior probabilities would be the following. The conditional probability of component  $c$ , given that  $\mathbf{x}$  equals pattern  $\mathbf{v}_j$ ,  $j = 1, 2, \dots, m$ , is, using Bayes' formula and letting  $p(\cdot)$  denote the requisite probability mass functions

$$p(c|\mathbf{x}=\mathbf{v}_j) = p_c p(\mathbf{v}_j|c) / p(\mathbf{x}=\mathbf{v}_j)$$

Note that this formula is analogous to the conditional probability formula

$$Pr(A|B) = Pr(A \text{ and } B) / Pr(B) = Pr(A)Pr(B|A) / Pr(B)$$

Expanding the numerator in terms of the C states, we have

$$p(c|\mathbf{x}=\mathbf{v}_j) = p_c p(\mathbf{v}_j|c) / [p_1 p(\mathbf{v}_j|1) + p_2 p(\mathbf{v}_j|2) + \dots + p_C p(\mathbf{v}_j|C)].$$

The conditional probability  $p(\mathbf{v}_j|c)$  of the particular experimental outcome  $\mathbf{v}_j$ , given state (component)  $c$ ,

is a measure of compatibility, on a zero-to-one scale, between that state and that experimental result. The above procedure, then, corresponds to estimating this as 1 or 0, according to if there is or is not a match between  $c$  and  $\mathbf{v}_j$ . Alternatively, the  $p(\mathbf{v}_j|c)$  could presumably be estimated experimentally. This would mean obtaining experimental data  $\mathbf{x}$  a number of times from a residue or residues whose angles are in configuration  $c$  and obtaining the distribution across the  $\mathbf{v}_j$  values. This would need to be done for each state  $c$ .

### 6.2.1. Using posterior probabilities in protein spatial structure determination

In this way, posterior probabilities of the states are obtained for each of the  $n$  sites (residues). Random structures are then built up as follows, using a Metropolis-type algorithm. Consider a case where three components, say C1, C2 and C3, have posterior probabilities of 0.5, 0.3, and 0.2. A random number  $u$  between 0 and 1 is obtained. If  $u$  is between 0 and 0.5, the site is classified as C1; if  $u$  is between 0.5 and 0.8, as C2; if  $u$  is between 0.8 and 1, as C3. The mean values for these components can then be used as the values of the dihedral angles at that site. This is done for all  $n$  sites. This gives one configuration for the chain. The procedure is repeated for all  $n$  sites, a number, say  $M$ , times. ( $M$  might be 100 for example.) This yields  $M$  suggested configurations for the given chain. These  $M$  can all be tried as starting configurations in an energy minimization. This Metropolis-type algorithm constitutes a distinct improvement over pure Monte Carlo simulation. A Metropolis-type algorithm has been coded in the program FiSiNOE-3; see Ref. [2]. The maximum likelihood estimate of the structure is the one where, at each site, the estimate of the dihedral angles is taken to be the mean of the component with the highest probability. In addition to this highest-probability structure, on the order of 100 000 alternative structures can be generated; on the order of 1000 of these are chosen as starting configurations for the target-function optimization. These 1000 can be chosen to be those with the highest probabilities.

### 6.3. Continuous experimental data

Numerical values for the dihedral angles are continuous data. Such values are available both from the

build-up process and from final suggested structures. Bayesian procedures of classification and estimation may be applied to such values.

### 6.3.1. Classification

Now let  $\mathbf{x}$  denote the vector of three angles. The state from which the vector  $\mathbf{x}$  arose can be estimated by Bayesian classification as follows. Maximum a posteriori (MAP) classification can be used. This consists of simply assigning  $\mathbf{x}$  to that state for which the posterior probability is highest. That is

$$\text{assign } \mathbf{x} \text{ to } c^* \text{ iff. } c^* = \arg \max_{c=1,2,\dots,C} \{p(c|\mathbf{x})\}$$

Here

$$p(c|\mathbf{x}) = p_c f_c(\mathbf{x}) / f(\mathbf{x}), \quad c = 1, 2, \dots, C$$

The multivariate Gaussian p.d.f. with mean vector  $\mathbf{m}_c$  and covariance matrix  $\mathbf{S}_c$  is

$$f_c(\mathbf{x}) = (2\pi)^{-p/2} (\det \mathbf{S}_c)^{-1/2} \exp[-(\frac{1}{2})D^2(\mathbf{x}, \mathbf{m}_c; \mathbf{S}_c)]$$

where  $p$  is the number of variables (three in our case), ‘‘det’’ denotes determinant, and, given any  $p$ -dimensional vectors  $\mathbf{u}$ ,  $\mathbf{v}$  and any non-singular  $p \times p$  matrix  $\mathbf{M}$ , the function  $D^2$  is the squared statistical distance (Mahalanobis distance), the quadratic form

$$D^2(\mathbf{u}, \mathbf{v}; \mathbf{M}) = (\mathbf{u} - \mathbf{v})' \mathbf{M}^{-1} (\mathbf{u} - \mathbf{v})$$

where  $\mathbf{v}'$  denotes the transpose of the (column) vector  $\mathbf{v}$ . (If the covariance matrix was the identity, this would be the squared Euclidean distance.)

Consequently, we have

$$\begin{aligned} \ln p(c|\mathbf{x}) &= \ln p_c - (\frac{1}{2}p) \ln(2\pi) - (\frac{1}{2}) \\ &\quad \times \ln \det \mathbf{S}_c - (\frac{1}{2})D^2(\mathbf{x}, \mathbf{m}_c; \mathbf{S}_c) - \ln f(\mathbf{x}) \end{aligned}$$

Thus, multiplying by  $-2$  and dropping terms which do not vary with  $c$ , it follows that

$$\begin{aligned} c^* &= \arg \min_{c=1,2,\dots,C} \{D^2(\mathbf{x}, \mathbf{m}_c; \mathbf{S}_c) \\ &\quad + p \ln \det \mathbf{S}_c - 2 \ln p_c\} \end{aligned}$$

(If all the covariance matrices  $\mathbf{S}_c$  were equal and all the prior probabilities were  $p_c$  equal, this would simply be the minimum distance classification, where the distance is the appropriate statistical (Mahalanobis) distance.)

Having classified  $\mathbf{x}$  as having arisen from state  $c^*$ , one might take the cluster mean  $\mathbf{m}_{c^*}$  as the estimate of

the vector  $\mathbf{A}$  of dihedral angles at that site. Alternatively, Bayesian estimation could be used.

### 6.3.2. Estimation

If one wishes to combine  $\mathbf{x}$  with the information in the database to obtain an improved estimate of the vector  $\mathbf{A}$  of dihedral angles at a site, one can proceed as follows.

Let us just illustrate what happens in the estimate of a single angle, say  $\phi$ . The posterior estimate (mean of the posterior distribution) is

$$E[\phi|\mathbf{x}] = w_1 x + w_2 E(\phi)$$

where  $E(\phi)$  is the mean of the mixture prior distribution and the weights  $w_1$  and  $w_2$  are proportional to the precisions (reciprocal variances):

$$w_1 = \text{Prec}(x|\phi) / \text{Const.}, \quad w_2 = \text{Prec}(\phi) / \text{Const.}$$

where  $\text{Prec}(\phi) = 1/\text{Var}(\phi)$ ,  $\text{Prec}(x|\phi) = 1/\text{Var}(x|\phi)$ , and  $\text{Const.} = \text{Prec}(\phi) + \text{Prec}(x|\phi)$ . That is, the estimator is a weighted average of the prior mean and the experimental result, where the weights depend upon the relative precisions.

Consider now the expression for the posterior p.d.f. in terms of the prior p.d.f.:

$$h(\mathbf{A}|\mathbf{x}) = f(\mathbf{A})g(\mathbf{x}|\mathbf{A})/k(\mathbf{x})$$

where  $g(\mathbf{x}|\mathbf{A})$  is the p.d.f. of  $\mathbf{x}$ , for fixed  $\mathbf{A}$  ( $g$  is sometimes called the ‘‘likelihood’’), and  $k(\mathbf{x})$  is the marginal p.d.f. of  $\mathbf{x}$ , i.e. the integral of the numerator with respect to  $\mathbf{A}$ . When the prior has the finite-mixture model form, it can be shown that the posterior can be expressed in terms of posteriors corresponding to the component p.d.f. values. In particular, the mean of the posterior comes out in terms of the means of the posteriors corresponding to the component p.d.f. values. In turn, the means of the component posteriors involve the means of the components of the mixture prior. When the experimental data  $\mathbf{x}$  consists of numerical estimates of the dihedral angles, this provides a way of combining experimental data  $\mathbf{x}$  with the means of the clusters corresponding to the mixture prior. Details will be given in future work.

### 6.3.3. Scoring hypothesized alternative structures

Suppose there are  $T$  competing suggested structures for a protein of  $n$  residues. Denote the dihedral angles

in the  $t$ th suggested structure by  $\{a_{ti}, i = 1, 2, \dots, n\}$ . The statistical likelihood function provides an appropriate way of scoring these suggested structures, i.e. the likelihood is a figure-of-merit for alternative suggested structures. It provides a measure of consistency of the suggested structure with the information in the database. The likelihood for structure  $t$  is  $f(a_{t1}) \times f(a_{t2}) \times \dots \times f(a_{tn})$ , where  $f(a)$  is the fitted mixture p.d.f., evaluated at  $a$ . This product is computed for each suggested structure  $t$ ; the highest value is the best suggested structure. When the  $a$  values have been obtained from the Bayesian posterior probability formula, the resulting product  $h(a_{t1}|x) \times h(a_{t2}|x) \times \dots \times h(a_{tn}|x)$  is the posterior probability of the  $t$ th suggested structure,  $t = 1, 2, \dots, T$ .

## 7. Conclusions

By finite mixture model cluster analysis we have fit distributions to dihedral angles from the PDB. The trivariate distribution of  $(\phi, \psi, \chi_1)$  is fit for amino acid residues with variable side-chain angles and the bivariate distribution of  $(\phi, \psi)$  is fit for those without variable side-chain angles.

The resulting 16 clusters admit interpretation as combinations of six states (B, P, R, E, T and L) with three modes for  $\chi_1$ .

Different amino-acid residues have different propensities for the 16 alternative states.

The fitted mixture model is useful in obtaining starting values for energy minimization and in Bayesian refinement of estimates of dihedral angles.

## References

- [1] L. Kurnarsky, O. Shats, S. Sherman, Improving the efficiency of protein structure determination from NMR, ECC3, Paper #52.
- [2] O. Shats, S.A. Sherman, The FiSiNOE-3 program for determination of protein and peptide conformations from NMR data, ECC3, Paper #53.
- [3] S. Sherman, S. Sclove, L. Kurnarsky, I. Tomchin, O. Shats, J. Mol. Struct. (Theochem) 368 (1996) 153–161.
- [4] R.L. Dunbrack Jr., The backbone-dependent rotamer library Webpage, 1996. World Wide Web URL <http://www.cmpchem.ucsf.edu/~dunbrack>.
- [5] R.L. Dunbrack Jr., M. Karplus, J. Mol. Biol. 230 (1993) 543–574.
- [6] R.L. Dunbrack Jr., M. Karplus, Nature Struct. Biol. 1 (1994) 334–340.
- [7] G.J. McLachlan, K.E. Basford, Mixture Models: Inference and Applications to Clustering, Dekker, New York, 1987.
- [8] D.M. Titterton, A.F.M. Smith, U.E. Makov, Statistical Analysis of Finite Mixture Distributions, Wiley, New York, 1985.
- [9] A.P. Dempster, N.M. Laird, D.B. Rubin, J. R. Stat. Soc. B 39 (1977) 1–38.
- [10] J.H. Wolfe, Multivariate Behav. Res. 5 (1970) 329–350.
- [11] J. Moulton, M.N.G. James, Proteins: Struct. Function Genetics 1 (1986) 146–163.
- [12] S.A. Sherman, A.M. Andrianov, A.A. Akhrem, J. Biomol. Struct. Dyn. 4 (1987) 869–884.
- [13] S.A. Sherman, A.M. Andrianov, A.A. Akhrem, J. Biomol. Struct. Dyn. 5 (1988) 785–801.
- [14] C.R. Cantor, P.R. Schimmel, Biophysical Chemistry, Freeman, New York, 1980.
- [15] S.L. Sclove, CLUSPAC: computer programs for mixture-model cluster analysis, CRIM Working Paper No. 92-2, 1992, Center for Research in Information Management, University of Illinois at Chicago, Chicago, IL.
- [16] J. Janin, S. Wodak, M. Levitt, B. Maigret, J. Mol. Biol. 125 (1978) 357–386.