Journal of Biomolecular NMR (2007) 37:31–41
DOI 10.1007/s10858-006-9097-7

Article

A Bayesian-probability-based method for assigning protein backbone dihedral angles based on chemical shifts and local sequences

Jun Wang & Haiyan Liu*
Hefei National Laboratory for Physical Sciences at the Microscale, and Key Laboratory of Structural Biology, School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230027, China

Received 11 April 2006; Accepted 19 September 2006

Key words: backbone dihedral angles, Bayesian probability, chemical shifts

Abstract

Chemical shifts contain substantial information about protein local conformations. We present a method to assign individual protein backbone dihedral angles into specific regions on the Ramachandran map based on the amino acid sequences and the chemical shifts of backbone atoms of tripeptide segments. The method uses a scoring function derived from the Bayesian probability for the central residue of a query tripeptide segment to have a particular conformation. The Ramachandran map is partitioned into representative regions at two levels of resolution. The lower resolution partitioning is equivalent to the conventional definitions of different secondary structure regions on the map. At the higher resolution level, the α and β regions are further divided into subregions. Predictions are attempted at both levels of resolution. We compared our method with TALOS using the original TALOS database, and obtained comparable results. Although TALOS may produce the best results with currently available databases which are much enlarged, the Bayesian-probability-based approach can provide a quantitative measure for the reliability of predictions.

Abbreviations: CSI – chemical shift index; ROC curve – receiver operating characteristic curve.

Introduction

Chemical shifts are among the most important parameters measured by NMR spectroscopy. They are sensitive to local environments and can be used as indicators of local conformations. As an important example, it has been known that the chemical shifts of protein backbone atoms correlate strongly with the backbone dihedral angles or secondary structure types (Spera and Bax, 1991; Wishart et al., 1991; Luginbuhl et al., 1995). This correlation has been exploited in two mutually opposite directions.

In one direction, many efforts have been made to predict chemical shifts from structures, including the developments of statistical approaches deriving empirical chemical shift hypersurfaces from databases of observed chemical shifts (Le and Oldfield, 1994; Wishart and Nip, 1998; Iwadate et al., 1999), and of non-statistical approaches using classical physics or empirical equations to compute chemical shifts (Osapay and Case, 1991, 1994; Neal et al., 2003). First principle or quantum mechanical approaches have also been proposed for the same purpose (Ando et al., 1998; Dedios et al., 1993; Xu and Case, 2001, 2002). Recently, artificial neural networks have been used for this goal as well (Meiler, 2003). Among the various approaches, those based on statistics are

*To whom correspondence should be addressed. E-mail: hyliu@ustc.edu.cn

usually more rapid, while those based on quantum mechanics can deal with various experimental conditions.

In the other direction, several algorithms have been developed to predict structures (mainly protein backbone dihedral angles and secondary structures) from chemical shifts. The chemical shift index (CSI) approach has been widely used to derive information about secondary structures (Wishart et al., 1992). The first CSI scheme made use of the chemical shifts of $^{1}H^{\alpha}$ to assign protein secondary structures. It was soon extended to include chemical shifts of atoms of other types (Le and Oldfield, 1994; Wishart and Sykes, 1994). The information derived from the chemical shifts of atoms of multiple types is in general more accurate and more reliable than that derived from chemical shifts of atoms of a single type. Wang and Jardetzky presented an approach to identify secondary structures by comparing the joint probabilities of a set of known chemical shifts to be associated with different secondary structure types (Wang and Jardetzky, 2002). There is also a neural network method introduced by Hung and Samudrala (Hung and Samudrala, 2003). These approaches make qualitative predictions of secondary structure types without providing further quantitative information. Beger and Bolton described an empirical hypersurface approach to determine protein backbone dihedral angles (Beger and Bolton, 1997). Cornilescu et al. presented the TALOS method, which extracts restraints on backbone dihedral angles from given sequence and secondary chemical shift data (Cornilescu et al., 1999). More specifically, TALOS searches a predefined database for 10 nearest-neighbor tripeptide segments of a query tripeptide segment, using a similarity measure which is a weighted sum of various sequence and chemical shift similarity terms. Predictions are attempted if the nearest neighbors consistently have similar dihedral angles. While the nearest-neighbor approach can provide useful constraints for structural refinements, the weighted sum form of the similarity measure is somehow artificial. One deficiency of the nearest-neighbor method is its sensitivity to the choice of database. With the original TALOS database which is relatively small, we found that a small number of abnormal members in the database may have large influences on the results.

Here we define a Bayesian-probability-based score function to assign the backbone dihedral angle of a central residue based on the same data as used by TALOS, i.e., the sequence and chemical shifts of a tripeptide segment. We estimated the parameters using database statistics. The probability-based approach provides a measure on the reliability of assignments, so that in structure refinement erroneous restraints could be avoided by enforcing only those restraints with higher reliability. In the following sections we will first describe the method. Then we will analyze the performance of the method for prediction goals at two different resolution levels. The lower resolution prediction goal is comparable to the assignment of single residue backbone conformations into different secondary structure regions, and the higher resolution predictions assign residues into subregions resulted from further dividing each of the α and β regions on the Ramachandran map. The results are compared with those obtained using the TALOS method. The differences between the methods will be discussed.

Materials and methods

A Bayesian-probability-based scoring function

The secondary chemical shift of an atom has been defined as the deviation of its chemical shift from the averaged chemical shift of atoms of the same type in random coil structures (Spera and Bax, 1991). We consider a tripeptide segment with a given amino acid sequence S and a set of backbone atom secondary chemical shifts σ (including chemical shifts of $^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{13}C, ^{15}N and $^{1}H^{\alpha}$). The Bayesian posterior probability for the backbone dihedral angles of the central residue to be in region D, $P(D|S, \sigma)$, can be formulated as

$$P(D|S,\sigma) = \frac{P(\sigma,S|D)P(D)}{\sum_{D'} P(\sigma,S|D')P(D')} = \frac{P(\sigma|D)P(S|D)P(D)}{\sum_{D'} P(\sigma|D')P(S|D')P(D')}. \quad (1)$$

Here we have assumed the conditional independence between the secondary chemical shifts and the sequence given the backbone conformation of

the central residue. This is reasonable because during the conversion of chemical shifts into secondary chemical shifts, the corrections for influence of amino acid types and neighboring residue effects have already been included (in the same way as TALOS (Cornilescu et al., 1999)). This assumption is also necessary as we do not have enough data to obtain a reliable joint distribution of the chemical shifts and the sequences conditioned on D. Now the ratio between the probabilities of two different conformations, D_1 and D_2, given S and σ can be expressed as

$$\begin{aligned}\frac{P(D_1|S,\sigma)}{P(D_2|S,\sigma)} &= \frac{P(\sigma|D_1)P(S|D_1)P(D_1)}{P(\sigma|D_2)P(S|D_2)P(D_2)} \\ &= \frac{P(\sigma|D_1)}{P(\sigma|D_2)}\frac{P(S,D_1)}{P(S,D_2)} \\ &= \frac{P(\sigma|D_1)}{P(\sigma|D_1)}\frac{P(D_1|S)}{P(D_2|S)},\end{aligned} \tag{2}$$

i.e.,

$$P(D|S,\sigma) \propto P(\sigma|D)P(D|S). \tag{3}$$

We assume that $P(\sigma|D)$ follows a multidimensional Gaussian distribution, that is,

$$P(\sigma|D) = \frac{1}{(2\pi)^{15/2}|B|^{1/2}} \exp\left(-\frac{1}{2}(\sigma - \langle\sigma\rangle)'B^{-1}(\sigma - \langle\sigma\rangle)\right), \tag{4}$$

where B is the covariance matrix of secondary chemical shifts, and define the following scoring function which is essentially the negative logarithm of the Bayesian probability defined by Equation 3,

$$\begin{aligned}M(D,S,\sigma) &= -2\ln(P(\sigma|D)P(D|S)) - 15\ln 2\pi \\ &= \ln|B| + (\sigma - \langle\sigma\rangle)'B^{-1}(\sigma - \langle\sigma\rangle) \\ &\quad - 2\ln P(D|S)\end{aligned} \tag{5}$$

We can now assign the central residue to a conformation D corresponding to the highest probability, or equivalently, the lowest M.

As the M value is based on the Bayesian probability, it provides a strict statistical measure for the reliability of predictions. Such a criterion is missing in most other methods.

The partitioning of the Ramachandran map

The partitioning of the Ramachandran map or the two-dimensional backbone dihedral angle space has been based on the distribution of backbone dihedral angles in native protein structures. The distribution was computed using 1296 bins covering the entire map, each bin of size 10° by 10°. We then partitioned the map at two levels of resolution. At the lower level, the bins were clustered into 4 regions as shown in Figure 1. Regions I, II and III correspond to β sheet, α helix and left-handed helix conformations, respectively, and Region IV corresponds to backbone conformations rarely observed in proteins. At the higher resolution level, the larger Regions I and II were further divided into two subregions: Region I into Region Ia and Region Ib representing the extended sheet and the polyproline helix conformations, respectively, and Region II into Region IIa and Region IIb corresponding to one type of turn and the α-helix conformations, respectively.

We note that the subpartitioning corresponds to the clustered distributions of backbone dihedral angles of all residues in native protein structures, not limited to residues contained in certain secondary structure elements. Although residues in β-sheets are more likely to be clustered (but not exclusively) in Regions Ia and Ib and residues in α-helices more likely to cluster in Region IIb, coil residues form densely populated clusters centered in these four sub-regions as well.

Parameter estimation

The parameters in Equations 4 and 5 are all derived statistically and separately from specific datasets. In order to obtain $P(D|S)$, a dataset of tripeptide segments was constructed as follows. High resolution (<2.0 Å) protein crystal structures in Protein Data Bank (PDB) (Berman et al., 2000) were filtered and clustered by CD-HIT (Li et al., 2001, 2002) based on their sequences. The sequence identity threshold used in CD-HIT was at 0.7. Then $P(D|S)$ was calculated as the ratio between the number of occurrences of segments

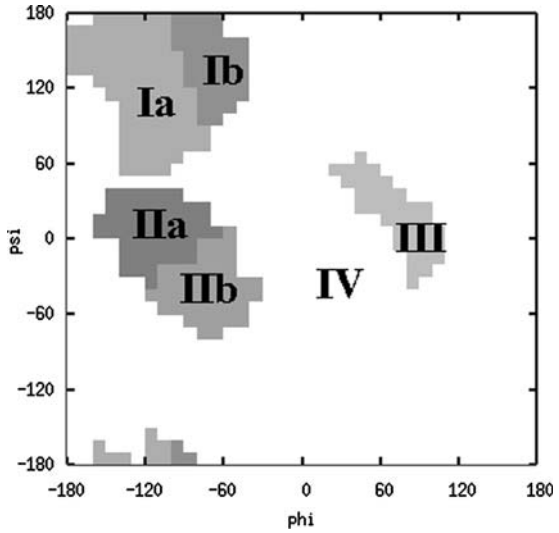

Figure 1. Partitioning of the Ramachandran map. Region I consists of Regions Ia and Ib, and Region II consists of Region IIa and IIb.

with sequence S and a central residue conformation in region D and the total number of occurrences of segments with sequence S in the dataset. We note that in each CD-HIT cluster all segments with the same sequence and falling within the same bin on the Ramachandran map were counted as one occurrence. This avoided over counting of repeatedly occurring segments in conserved domains of different proteins or in repeated domains of the same protein.

The parameters $\langle\sigma\rangle$ and B have been obtained separately using three different datasets. One of the dataset has been derived from BioMagResBank (BMRB, Seavey et al., 1991) and PDB (Berman et al., 2000). We filtered proteins contained in BMRB using the following criteria: (1) all five types of backbone chemical shifts have been assigned; (2) the resolution of the corresponding X-ray structure is below 2.5 Å; (3) the chain contains more than 80 residues; (4) the reference compound was DSS; and (5) the temperature and pH ranges of NMR experiments were 293–313 K and 4.5–8.5, respectively. Criteria 4 and 5 are to ensure that experimental conditions do not have significant effects on the observed chemical shift variations. After filtering 34 proteins remained. The BMRB ID numbers of these proteins as well as the corresponding PDB IDs are listed in Table 1. We then extracted tripeptide segments from these proteins. Any tripeptide segment meeting one of the following criteria has been excluded: (1) its amino acid sequences in the PDB file and in the BMRB file are not the same; (2) there are missing or multiple sets of coordinates for atoms in the tripeptide segments; (3) the

Table 1. Proteins contained in the dataset derived from BioMagResBank (BMRB)

Protein name	BMRB ID	Protein Data Bank (PDB) ID	N_{seg}^{a}
Peptidyl–prolyl *cis–trans* isomerase	bmr5305	1PIN	126
Cyay protein	bmr5792	1EW4	80
Endonuclease V	bmr5244	2END	120
Barnase	bmr4964	1A2P	88
ArcB	bmr4857	2A0B	105
Tetrahymena GCN5	bmr4321	1QST	120
Heme-binding protein A	bmr5081	1DK0	155
Ribosomal protein L25	bmr4395	1DFU	80
Frataxin	bmr4342	1EKG	110
Cellular retinoic-acid-binding protein type II	bmr4186	1CBS	83
Adapter-related protein	bmr5761	1GYU	109
Endo-1,4-β-xylanase	bmr5679	1XYF	109
3C proteinase	bmr4836	1QA7	168
Ribonuclease A	bmr4031	7RSA	116
VP4	bmr5275	1KQR	133
Cytochrome *B5*	bmr4803	1CYO	80
Flavodoxin	bmr5540	5FX2	116
Tetranectin	bmr6008	1TN3	95
Matrix metalloprotease-13	bmr4679	830C	125
Psd Zip45 (homer 1c/vesl 1l)	bmr4766	1I2H	107
Class B β lactamase	bmr4102	2BMI	187
Diphtheria toxin repressor	bmr4183	1BI1	81
Bet v 1	bmr4417	1BV1	141
Interferon regulatory factor 2	bmr4161	2IRF	88
DNA polymerase β	bmr4326	1BPY	71
Bleomycin resistance protein	bmr4786	1BYL	119
γ δ-resolvase	bmr4269	2RSL	82
Replication protein A	bmr5823	1JMC	175
Major urinary protein	bmr4340	1MUP	106
Nitrogen regulatory IIA protein	bmr5789	1A6J	117
Ribonuclease H	bmr5931	1HRH	96
Ba3-type cytochrome *C* oxidase	bmr5819	2CUA	105
Cytosolic phospholipase A2	bmr4188	1CJY	99
Bovine adrenodoxin	bmr4566	1CJE	83

[a]Number of tripeptide segments considered.

B-factor of any backbone atom exceeds 1.5 times the average backbone B-factors of the corresponding protein; (4) less than 3 backbone chemical shifts have been assigned; and (5) the absolute secondary chemical shift of any backbone atom exceeds 4.5 times the corresponding standard deviation. This resulted in 3775 tripeptide segments from which $\langle\sigma\rangle$ and B have been derived.

We also constructed a dataset based on the RefDB database (Zhang et al., 2003) using similar criteria, except that the restrictions on experimental conditions including reference compound, temperatures and pH have been removed. The resulting dataset contains 59 proteins and 6243 tripeptide segments. In cases there are several sets of chemical shift assignments in the RefDB database corresponding to the same set of protein coordinates in the PDB database, one set of chemical shift assignments have been carefully selected.

To compare with TALOS, we also derived the parameters using the dataset used in the original TALOS paper.

Receiver operating characteristic curves

When making a dichotomic prediction based on a continuous value, one needs to choose an appropriate threshold value. After comparisons with the experimental results, predictions can be classified into four groups: true positive, false positive, true negative and false negative. Shifting the threshold changes the sensitivity (the number of true positive predictions over the number of actual positives cases) and specificity (the number of false positive predictions over the number of actual negatives cases) of the predictions in opposite directions. The receiver operating characteristic (ROC) curve can be used to represent the tradeoff between sensitivity and specificity. Conventionally the x-axis of the ROC curve corresponds to one minus the specificity and the y-axis the sensitivity, and the area under the ROC curve measures the discriminative ability of the variable employed to make the predictions. Usually, an area of 0.8 or larger indicates a well-chosen variable, while an area of 0.5 suggests that the prediction is hardly better than random predictions. More details of this analysis technique can be found in the literature (Metz, 1978).

Results and discussions

Statistics of secondary chemical shifts

For the 34 proteins in the BMRB-derived dataset, the distributions of various secondary chemical shifts of residues in each of the four regions on the Ramachandran map (see Figure 1) are shown in Figure 2. The distributions of $^{13}C^{\alpha}$ or $^{13}C^{\beta}$ secondary chemical shifts vary greatly with backbone dihedral angles, whereas the distributions of ^{15}N secondary chemical shifts of residues in different Ramachandran regions do not have very apparent differences. Larger differences between the distributions of chemical shifts of residues in different regions indicate that these chemical shifts contain more specific information about the backbone conformation of the central residue.

To quantify the discriminative capability of secondary chemical shifts of different atom types,

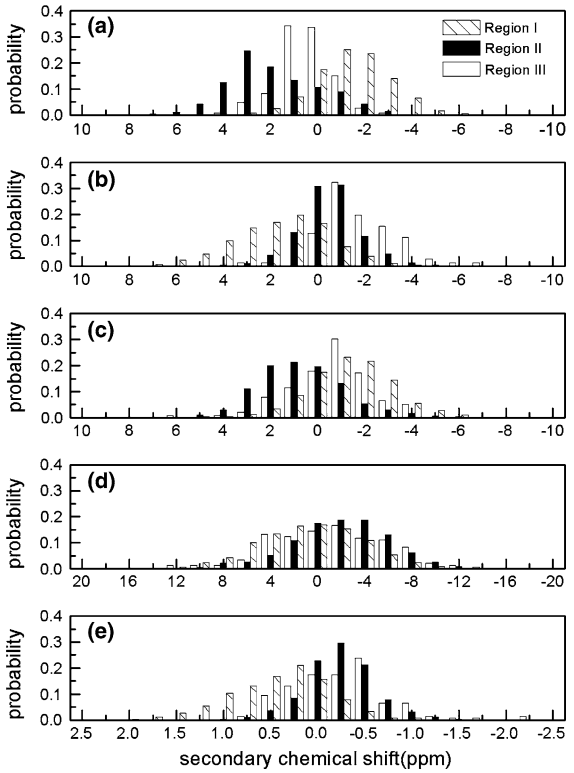

Figure 2. Distributions of secondary chemical shifts of (a) $^{13}C^{\alpha}$ (b) $^{13}C^{\beta}$ (c) $^{13}C'$ (d) ^{15}N (e) $^{1}H^{\alpha}$ for residues with backbone conformations in Regions I, II and III, respectively. Residues contained in the BioMagResBank (BMRB)-derived dataset have been considered.

the following test was performed using the BMRB-derived dataset. For each pair of Ramachandran regions excluding Region IV, each of the tripeptide segments belonging to either of the pair of regions was assigned to one of the two regions based on the secondary chemical shift of only a single atom of the central residue (that is, tripeptide segments with secondary chemical shifts above a varying cutoff were assigned into one region, others were assigned into the other region). The areas under the corresponding ROC curves, which represent the discriminative ability of individual single-residue-single-atom chemical shifts, were calculated and listed in Table 2. The results indicate that the $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{1}H^{\alpha}$ secondary chemical shifts are able to discriminate well the tripeptide segments in Region I from those in Region II or Region III, but any individual chemical shift cannot discriminate between the tripeptide segments in Regions II and III. Using only $^{1}H^{\alpha}$ secondary chemical shifts, the area under the ROC curve for discriminating between tripeptide segments in Regions II and III is only slightly larger than 0.5, as the averaged $^{1}H^{\alpha}$ secondary chemical shifts in the two regions are very close. The weak discriminative ability of ^{15}N secondary chemical shifts corresponds well with that they show similar distributions for different Ramachandran regions. In the following, we retained chemical shifts of backbone atoms whose discriminative abilities are not strong, as using these data can still slightly increase the accuracy of the assignments.

Table 2. Areas under the receiver operating characteristic (ROC) curves for predictions using single-atom secondary chemical shifts to discriminate between two Ramachandran regions

Atom type[a]	Area[b]		
	Region I vs. Region II	Region I vs. Region III	Region II vs. Region III
$^{13}C^{\alpha}$	0.889	0.835	0.733
$^{13}C^{\beta}$	0.808	0.913	0.760
$^{13}C'$	0.789	0.603	0.719
^{15}N	0.678	0.616	0.551
$^{1}H^{\alpha}$	0.856	0.825	0.504

[a]The atom whose secondary chemical shift was used for prediction.
[b]The predictions were performed for residues whose actual backbone dihedral angles fall into one of the two Ramachandran regions considered in each column. See text for details.

The accuracy and reliability of lower resolution assignments

In this study, the backbone conformation of a central residue of a tripeptide segment was assigned into the backbone dihedral angle region with the lowest value of M (Equation 5).

To test the accuracy of our method, we performed a leave-one-out test on the BMRB-derived dataset, that is, we assigned the backbone dihedral angles of one protein in this dataset using the parameters statistically obtained from the dataset excluding this protein. The accuracy of assignments for each backbone dihedral angle region, which is defined as the ratio of the number of correctly assigned segments over the total number of tripeptide segments, is summarized in Table 3. For the most populated Regions I and II, our assignments gave an accuracy of 89% and 87%, respectively. The accuracy for Region III is only 47%. This may have been caused by the lack of data in the statistics, as there are only 146 tripeptide segments with their central residue conformations belonging to Region III. Region IV corresponds to the backbone dihedral angles which rarely occur, so the accuracy for this region is the lowest, as expected. We also considered a model in which Region IV is strictly disallowed in the prediction results and recalculated the accuracy of our predictions. Then the accuracy for Region

Table 3. The accuracy of the predictions for residues in different Ramachandran regions[a]

	Number of tripeptide segments in dataset	Number of correctly predicted segments[b]	Number of correctly predicted segments[c]
Region I	1874	1660 (89%)	1690 (90%)
Region II	1562	1356 (88%)	1375 (88%)
Region III	146	69 (47%)	82 (56%)
Region IV	193	72 (37%)	–

[a]Predictions were attempted for all residues in the BMRB-derived dataset, disregarding the reliability measure.
[b]Predicted with Region IV allowed in the results. Numbers in parentheses are portions of correctly predicted segments.
[c]The same as the previous column, except that Region IV is strictly disallowed in the predictions.

III increased by 9%, and those for Region I and II also increased slightly.

The above levels of accuracy correspond to including all predictions without rejecting predictions which can be considered as unreliable based on the computed probabilities. Naturally, ΔM, the difference between the two lowest values of M, could be considered as a measure of the reliability of an assignment. In order to check this assumption, we can define a (varying) cutoff value for ΔM, and only consider assignments with ΔM less than the cutoff as reliable. The ROC curve associated with varying the cutoff value is shown in Figure 3a, in which the sensitivity (defined as the ratio of the number of true positive assignments over the total number of actual positive cases) has been plotted versus one minus the specificity (defined as the ratio of the number of true negative assignments over the total number of actual negative cases) at different cutoff values.

As the cutoff for ΔM increases, the sensitivity increases, accompanied by decreases in specificity. This verifies that ΔM provides a quantitative measure for reliability of predictions. In addition, the area under the above ROC curve is 0.83, indicating again that ΔM is suitable to discriminate reliable assignments from unreliable ones. Figure 3b shows the curves of the coverage and accuracy of assignments considered as reliable versus ΔM, where the coverage is defined as the ratio of the number of assignments considered as reliable over the total number of queries, and the accuracy as the ratio of the number of correct reliable assignments over the total number of assignments considered as reliable. In protein structure refinement, imposing a wrong restraint may severely distort the resulting structure. So we set $-\ln 100$, a relatively strict ΔM value, as the cutoff to ensure the accuracy of assignments. This means that for the accepted assignments, the probability for the query tripeptide segment to be in the assigned region is at least 10 times larger than the probabilities for the same tripeptide segment to be in other regions. At this cutoff value, the accuracy and coverage of accepted assignments are 94% and 69%, respectively, in the leave-one-out test using the BMRB-derived dataset.

The same leave-one-out test has been performed using the larger RefDB-derived dataset. The ROC, coverage and accuracy curves are also plotted in Figure 3. With the same cutoff of $-\ln 100$ for ΔM, 4672 assignments (4447 true assignments and 225 false ones) out of 6243 queries were accepted. The accuracy and coverage of accepted assignments are 95% and 75%, respectively. The improvement over the test using the BMRB-derived dataset may be attributed to higher data quality and more amount of data employed to estimate the statistical parameters.

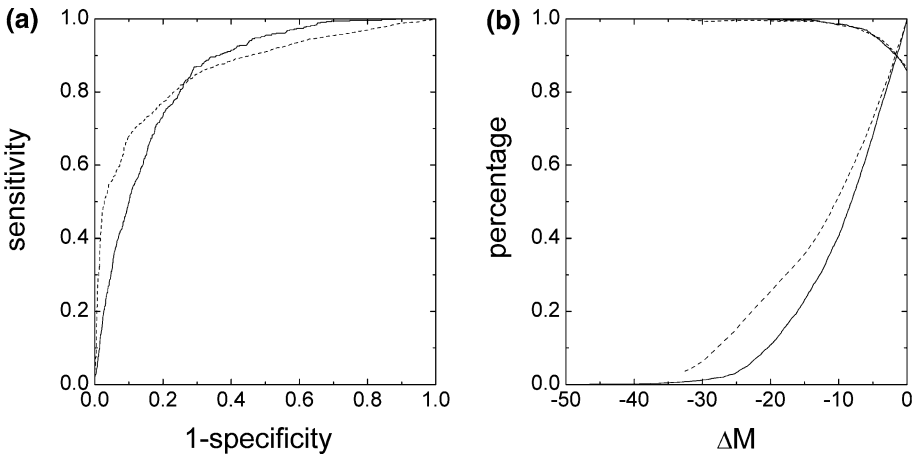

Figure 3. (a) Sensitivity and specificity of predictions accepted at different ΔM values for assignments at the lower resolution level. From left to right ΔM increases monotonically, corresponding to lowering the threshold for a prediction to be accepted. (b) Accuracy (the descent line) and coverage (the ascent line) of accepted assignments versus ΔM for assignments at the lower resolution level. The accuracy is the number of correct accepted assignments divided by the total number of accepted assignments. The coverage is the number of accepted assignments divided by the total number of input targets. Solid lines: results for the BioMagResBank (BMRB)-derived dataset. Dashed lines: results for the RefDB-derived dataset.

In Table 4 the results obtained using the three different datasets are summarized. Results from the include-all tests (in which the Bayesian probability parameters have been estimated using each complete dataset, without excluding the protein containing the query tripeptide segment) are compared with the leave-one-out test results. In principle, to eliminate query information from the training data, we only need to remove each query tripeptide segment instead of each of the entire proteins in the leave-one-out test. This would however result in too tedious repetitions of the statistical process. In addition, even for the smaller datasets, removing a single query peptide from thousands of training peptides would have negligible effects compared to keeping all training data (include-all). So the leave-one protein-out results instead of the leave-one tripeptide-out results are reported here. For the smaller datasets, the leave-one protein-out tests did lead to slightly degenerated performance compared to the include-all tests. This mainly reflects poorer statistics rather than excluding queries from training data. For the larger RefDB-derived dataset results from the include-all tests approaches those from the leave-one-out tests.

The accuracy and reliability of higher resolution assignments

For the reliable assignments in Regions I and II, we can further assign the query tripeptide segments into one of the subregions using the same Bayesian approach. Using the BMRB-derived dataset and setting $-\ln 100$ as the cutoff, we obtained 1352 reliable assignments (1275 true assignments and 77 false ones) for Region I and 1200 assignments (1138 true ones and 62 false ones) for Region II in the leave-one-out tests. Figure 4a, c show the ROC curves for the further assignments of these tripeptide segments into subregions of Region I and Region II, respectively. The areas under the ROC curves are 0.81 and 0.86, respectively. This indicates that ΔM can again be used to discriminate reliable assignments from unreliable ones. Figure 4b, d show the corresponding coverage and accuracy versus ΔM curves for the two regions, respectively. As before, we set $-\ln 100$ as the cutoff. Then for Region I, our method produced 902 reliable assignments into Ia and Ib (838 true assignments and 64 false ones), and for Region II, 1019 reliable assignments into IIa and IIb (962 true assignments and 57 false ones). These results suggest that most tripeptide segments in these two regions can be further reliably assigned into sub-regions. This is of importance in practical structure refinements.

The same tests were carried out using the RefDB-derived dataset. Using a cutoff of $-\ln 100$, we obtained 2060 reliable assignments (1954 true assignments and 106 false ones) for Region I and 2511 reliable assignments (2432 true ones and 79 false ones) for Region II in the leave-one-out tests. The ROC, accuracy and coverage curves for the further assignments into subregions are shown in Figure 4. For Region I, further assignments into Ia and Ib gave 1338 reliable ones (1290 true assignments and 48 false ones), and for Region II, further assignments into IIa and IIb gave 2193

Table 4. Summary of the results by the Bayesian-probability-based methods

Dataset[a]	Parameter estimation[b]	Number of segments	Good[c]	Amb[c]	Bad[c]	Coverage[d]	Accuracy[e]
TALOS	Leave-one-out	2889	1969	806	114	72%	95%
	Include-all	2889	2037	765	87	74%	96%
BioMagResBank (BMRB)-derived	Leave-one-out	3775	2444	1167	164	69%	94%
	Include-all	3775	2539	1098	138	71%	95%
RefDB-derived	Leave-one-out	6243	4447	1571	225	75%	95%
	Include-all	6243	4501	1540	202	75%	96%

[a]See text for definitions of datasets.
[b]In "leave-one-out" tests, the protein containing the query tripeptide segment was excluded from the dataset and parameters re-estimated. In the "include-all" tests, all proteins contained in a given dataset have been used in parameter estimation.
[c]Ambiguous predictions (Amb) refer to those with ΔM above $-\ln 100$. Predictions with ΔM below $-\ln 100$ were marked as either "Good" or "Bad" depending on whether they agree with the actual dihedral angles. Each column lists the corresponding number of predictions.
[d]Number of non-ambiguous predictions divided by the total number of segments in each dataset.
[e]Number of good predictions divided by the total number of non-ambiguous predictions.

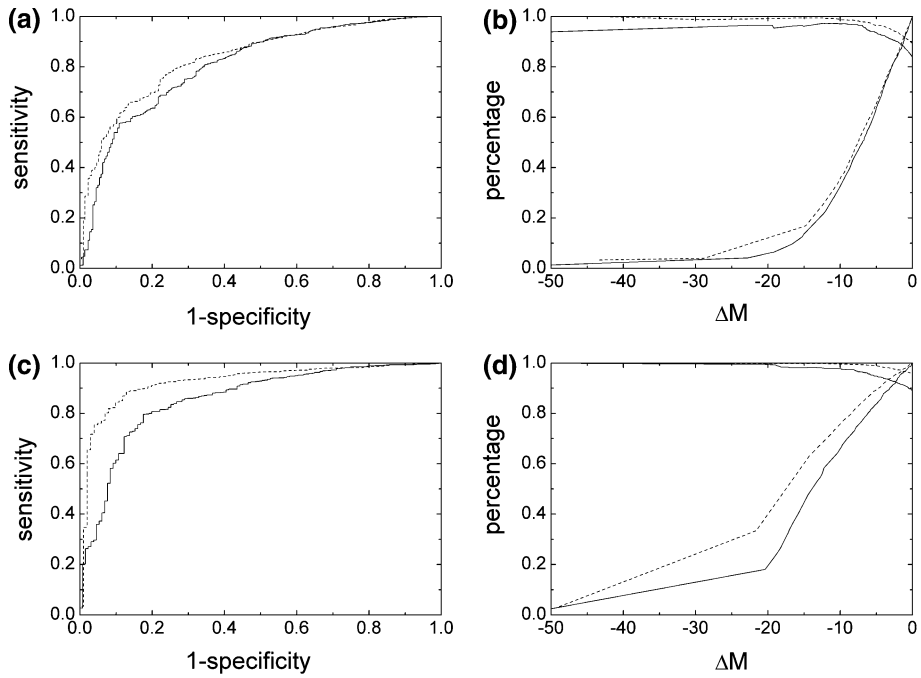

Figure 4. (a) Sensitivity and specificity of predictions accepted at different ΔM values for further assignments of residues in Region I at the higher resolution level. (b) Accuracy and coverage of accepted assignments versus ΔM for further assignments of residues in Region I at the higher resolution level. (c) Similar to (a), but for Region II. (d) Similar to (b), but for Region II. See caption of Figure 3 for other details.

reliable ones (2166 true assignment and 27 false ones). Compared with using the BMRB-derived dataset, the results are again improved.

Contributions of neighboring residues

Most previous methods assigned backbone dihedral angles based on only the chemical shifts of atoms of a single residue. Following TALOS, the Bayesian probability approach presented here utilizes extra information contained in chemical shifts of immediately adjacent residues in sequence. To verify the contributions of adjacent residues in our assignments, a test using the BMRB-derived dataset has been performed in which the backbone dihedral angles were assigned without using the data of neighboring residues. At the ΔM value of $-\ln 100$, the accuracy and coverage of this test assignment are 95% and 57%, respectively. And in Figure 5, the curve without considering these chemical shifts is always below the curves obtained with them taken into account. Clearly, considering adjacent residues is helpful to the prediction of backbone dihedral angles.

Comparisons of performance with TALOS

TALOS is the only previous method which uses information from adjacent residues, so in the following paragraph, we compare the performance of our method only with that of TALOS. Before presenting the results, we emphasize that the comparisons here are made based on applying both the Bayesian method and the TALOS method to the original TALOS dataset. It has been indicated that with current much larger databases, TALOS can achieve 71% coverage and $>98\%$ accuracy.

We first compare our method with TALOS for predictions at the lower resolution level. We applied both the TALOS procedure and the Bayesian-based approach to the original TALOS dataset, which contains 21 proteins and 2899 tripeptide segments. We did not include tripeptide segments with missing backbone chemical shift data for the first or the last residue in this comparison. The following criteria for "good" (true positive) and "bad" (false positive) predictions were applied to mark the results: an attempted prediction is marked as "good" if the actual

backbone dihedral angles are in the predicted region (with the Bayesian probability approach), or are in (or just outside) the same region as the predicted dihedral angle values (with the TALOS approach), otherwise it is marked as ''bad''. In principle, the target dihedral angles and its nearest neighbors can all be in a less populated region and still cluster well. This would still be considered as a good prediction. With the small dataset such cases did not occur. The accuracy versus coverage curve using the Bayesian approach is plotted in Figure 5. The point corresponding to result obtained using the TALOS method is indicated in the same figure. This point is slightly below the Bayesian curve, indicating that both methods have similar overall performance at this level of prediction resolution. When the accuracy of our method is equal to that of TALOS, the coverage of our method is 69% and when the coverage of our method is equal to that of TALOS, the accuracy of our method is 97%.

TALOS also gives the predictions of specific angles by using the averaged backbone dihedral angles of contributing nearest-neighbor tripeptide segments. It has been reported that the root mean square differences between the good predictions by TALOS and the dihedral angles observed in crystal structures were about 15° for φ, and 14° for ψ (Cornilescu et al., 1999). If the backbone dihedral angles of all query tripeptide segments reliably assigned into Region Ia using the Bayesian method were predicted as (−122°, 136°), and the backbone dihedral angels of all query tripeptide segments reliably assigned into Region I but not reliably assigned into either Region Ia or Region Ib were predicted as (−106°, 138°), and so on, the root mean square differences between the predictions and the actual values would be 17° for φ and 18° for ψ. These results would be slightly inferior to but on the same order of magnitude as that given by TALOS.

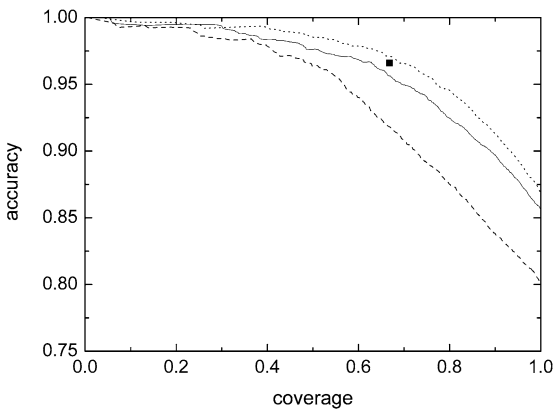

Figure 5. Accuracy (the number of correct predictions divided by the number of accepted predictions) versus coverage (the number of accepted predictions divided by the total number of queries) of the predictions. Solid line: predictions made using the BioMagResBank (BMRB)-derived dataset and using information from flanking residues; dashed line: predictions made using the BMRB-derived dataset without using information from flanking residues; dotted line: predictions made on the dataset used by TALOS using the Bayesian probability method; square: predictions made on the dataset used by TALOS using the TALOS method.

One advantage of the nearest-neighbor-based TALOS method is that it can capture the correlation between unusual local structures and sequence and chemical shifts, providing that such unusual local structures occurred sufficient number of times in the dataset. Current dataset size usually does not allow for such unusual local structures to be characterized by the Bayesian-based approach

The nearest-neighbor method may, however, have some undesirable effects. With the original TALOS dataset, we observed that some of the ambiguous predictions of different peptide segments made by the TALOS method are actually caused by that some tripeptides in the dataset are repeatedly but wrongly selected as nearest neighbors. For example, Asp^{50} of β-hydroxydecanoyl thiol ester dehydrase was identified as the nearest-neighbor of 16 query segments which have different central dihedral angles, resulting in ambiguous predictions for these queries. On the contrary, the Bayesian-based method assigned these queries correctly. General experiences have shown that enlarging the dataset in the TALOS method will reduce such effects (the test for the latest version of TALOS containing 78 proteins shows the coverage of assignment reaches 71%). It, however, seems that the problem does not disappear completely as there are good, unambiguous predictions made using the smaller dataset turned into ambiguous ones using the larger dataset.

Program availability

The method reported here was implemented in Fortran. It reads protein sequences and chemical

shifts from a text file, and its output contains the assigned backbone dihedral region and the values of M and ΔM for each residue. The program can be downloaded at http://www.sg.ustc.edu.cn/download, as well as the datasets and statistical results.

Acknowledgments

This work has been supported by Chinese Department of Science and Technology (Grant No. 2004AA235110) and National Natural Science Foundation of China (Grant Nos. 90403120 and 30121001).

References

Ando, I., Kameda, T., Asakawa, N., Kuroki, S. and Kurosu, H. (1998) *J. Mol. Struct.*, **441**, 213–230.
Beger, R.D. and Bolton, P.H. (1997) *J. Biomol. NMR*, **10**, 129–142.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.
Dedios, A.C., Pearson, J.G. and Oldfield, E. (1993) *Science*, **260**, 1491–1496.
Hung, L. and Samudrala, R. (2003) *Protein Sci.*, **12**, 288–295.
Iwadate, M., Asakura, T. and Williamson, M.P. (1999) *J. Biomol. NMR*, **13**, 199–211.
Le, H. and Oldfield, E. (1994) *J. Biomol. NMR*, **4**, 341–348.
Li, W., Jaroszewski, L. and Godzik, A. (2001) *Bioinformatics*, **17**, 282–283.
Li, W., Jaroszewski, L. and Godzik, A. (2002) *Bioinformatics*, **18**, 77–82.
Luginbuhl, P., Szyperski, T. and Wuthrich, K. (1995) *J. Magn. Reson. B*, **109**, 229–233.
Meiler, J. (2003) *J. Biomol. NMR*, **26**, 25–37.
Metz, C.E. (1978) *Semin. Nulc. Med.*, **8**, 283–298.
Neal, S., Nip, A.M., Zhang, H. and Wishart, D.S. (2003) *J. Biomol. NMR*, **26**, 215–240.
Osapay, K. and Case, D.A. (1991) *J. Am. Chem. Soc.*, **113**, 9436–9444.
Osapay, K. and Case, D.A. (1994) *J. Biomol. NMR*, **4**, 215–230.
Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.
Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.
Wang, Y. and Jardetzky, O. (2002) *Protein Sci.*, **11**, 852–861.
Wishart, D.S. and Nip, A.M. (1998) *Biochem. Cell Biol.*, **76**, 153–163.
Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.
Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311–333.
Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) *Biochemistry*, **31**, 1647–1651.
Xu, X. and Case, D.A. (2001) *J. Biomol. NMR*, **21**, 321–333.
Xu, X. and Case, D.A. (2002) *Biopolymers*, **65**, 408–423.
Zhang, H., Neal, S. and Wishart, D.S. (2003) *J. Biomol. NMR*, **25**, 173–195.