Journal of Biomolecular NMR (2005) 32: 71–81
DOI 10.1007/s10858-005-5705-1

Article

Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements

Hamid R. Eghbalnia[a,b,*], Liya Wang[a,c,d], Arash Bahrami[a,c,d], Amir Assadi[b] & John L. Markley[a,c,d]

[a]*Biochemistry Department, National Magnetic Resonance Facility at Madison, 433 Babcock Drive, Madison, WI, 53706, USA;* [b]*Mathematics Department, University of Wisconsin–Madison, 811 Van Vleck Hall, 480 Lincoln Drive, Madison, WI, 53706, USA;* [c]*Center for Eukaryotic Structural Genomics, University of Wisconsin–Madison, Madison, WI, 53706, USA;* [d]*Graduate Program in Biophysics, University of Wisconsin–Madison, Madison, WI, 53706, USA*

Received 28 December 2004; Accepted 08 March 2005

Key words: chemical shifts, protein secondary structure, statistical energy model, statistical decision

Abstract

We present an energy model that combines information from the amino acid sequence of a protein and available NMR chemical shifts for the purposes of identifying low energy conformations and determining elements of secondary structure. The model (''PECAN'', Protein Energetic Conformational Analysis from NMR chemical shifts) optimizes a combination of sequence information and residue-specific statistical energy function to yield energetic descriptions most favorable to predicting secondary structure. Compared to prior methods for secondary structure determination, PECAN provides increased accuracy and range, particularly in regions of extended structure. Moreover, PECAN uses the energetics to identify residues located at the boundaries between regions of predicted secondary structure that may not fit the stringent secondary structure class definitions. The energy model offers insights into the local energetic patterns that underlie conformational preferences. For example, it shows that the information content for defining secondary structure is localized about a residue and reaches a maximum when two residues on either side are considered. The current release of the PECAN software determines the well-defined regions of secondary structure in novel proteins with assigned chemical shifts with an overall accuracy of 90%, which is close to the practical limit of achievable accuracy in classifying the states.

Introduction

Protein secondary structure plays an important role in classifying proteins (Lesk and Rose, 1981) and in analyzing their functional properties (Przytycka et al., 1999). A host of methods have been developed for the prediction of secondary structure from atomic coordinates (determined from X-ray crystallography or NMR spectroscopy), NMR chemical shifts, or simply peptide sequences. The primary forces that govern secondary and tertiary structure are closely related, and it is generally assumed that a detailed characterization of the energetic genesis of secondary structure is a key step toward understanding protein folding.

The accuracy of secondary structure prediction from amino acid sequence alone has been reported to be as high as 78% on selected datasets (Albrecht et al., 2003). The fact that secondary structure can be predicted from sequence with some measure of success indicates that amino acid sequences encode

*To whom correspondence should be addressed. E-mail: eghbalni@nmrfam.wisc.edu.

local information about peptide conformation. On the other hand, the widely demonstrated finding that structure is more highly conserved than sequence in proteins places a limit on the practical accuracy of conformational predictions from sequence alone. On the experimental side, NMR chemical shifts have been considered as indicators of secondary structure since the late 1960s (Markley et al., 1967; Sternlicht and Wilson, 1967), and strong relationships between chemical shifts and secondary structure have been elucidated (Spera and Bax, 1991; Wishart et al., 1991; Le and Oldfield, 1994; Luginbuhl et al., 1995; Iwadate et al., 1999; Sibley et al., 2003). Several algorithms have been developed for identifying secondary structure from NMR chemical shifts alone; these include the $\Delta\delta$ method (Reiley et al., 1992), the chemical shift index method (Wishart and Sykes, 1994), the database approach by TALOS (Cornilescu et al., 1999), and the probability-based method (Wang and Jardetzky, 2002).

Given the separate predictive potential of sequence and chemical shifts, it is reasonable to assume that secondary structure predictions obtained by combining the two would be better than those from either alone. Hung and Samudrala (2003) have applied a supervised machine learning approach to combining sequence and chemical shift data. Our approach described here has been to develop an energetic model that presents a framework for combining the interdependent information from sequence and chemical shifts in a manner that optimizes their joint predictive potential.

The remarkable effectiveness of statistical potentials in specific applications, for example, in molecular mechanics and molecular dynamics simulations (Kuszewski et al., 1996; Moult, 1997), provides reasonable grounds for attempting to devise an energy-based approach for secondary structure prediction. The standard approach is to convert occurrence frequencies to estimates of free energy through application of the Boltzmann hypothesis and to assert that the logarithm of the probability of a specific conformational state is proportional to its energy. In other words, amino acids populate each structural feature with a probability that can be calculated by the familiar Boltzmann–Gibbs weighting factor of statistical thermodynamics.

We used a database containing ~37,000 residues from 310 protein sequences in constructing our statistical potential and in obtaining results from secondary structure determinations. We used a second (non-overlapping) database containing ~12,000 residues from 97 protein sequences in determining that the model is independent of the dataset. Equivalent unbiased criteria were used in selecting the members of each dataset, which consisted of proteins with known structure and assigned chemical shifts. With both datasets we show how the combined use of amino acid sequence and chemical shift information yields a marked increase in accuracy over chemical shift alone in determining secondary structure. We discuss the information content of multidimensional statistical potentials and the basis for our choice of optimal chemical shift dimensions.

Our energy-based method for secondary structure prediction (PECAN, Protein Energetic Conformational Analysis from NMR chemical shifts) is part of a larger effort on the automated analysis of protein NMR data. A single software platform has been developed that uses as input the sequence of the protein and peak lists derived from various experimental multidimensional, multinuclear magnetic resonance datasets and that provides as output chemical shift assignments and secondary structure analysis. We have named this package "PISTACHIO" (Probabilistic Identification of Spin sysTems and their Assignments including Coil–Helix Inference as Output). It is available for general use from a server at http://bija.nmrfam.wisc.edu. The chemical shift assignment algorithm used in PISTACHIO will be presented in a separate publication.

Methods

Overview of the mathematical model

Suppose that w_n is the number of observations of state n and E_n is the energy of state n. Further assume that w_n is approximately proportional to the probability for state n. Then, for a system in equilibrium in one of N possible states, the energy and probability are related as:

$$p(E_n) = \left(\frac{1}{Z}\right)e^{-\beta E_n} \quad Z = \sum_{n=1}^{N} e^{-\beta E_n}$$
$$\beta = 1/kT \tag{1}$$

where k is the Boltzmann constant, T is the temperature and Z is the normalization factor called the partition function, and $p(E_n)$ probability of state E_n. The constant β sets a monotonic scale for units of energy and for our purposes can be set equal to a convenient constant.

For our model we retain the general form of this distribution but modify the energy term and its algebraic interpretation. The form for the energy term E can be very general and for our model has two components: a term representing the statistical (local) bias potential arising from chemical shifts, followed by a second term describing the interaction potential arising from propensities of pairwise sequences (dipeptides) to be in specific classified states (Equation 2). The dependence of the pairwise potential on the local bias potential extends the pairwise potential to an effective three-body interaction potential (tripeptide) that accounts for sequence dependent effects while at the same time keeping the problem computationally tractable. We make no assumptions about the analytic form of these terms; instead each energy term and its proportional influence on the final energy is first estimated and subsequently refined.

$$E_i = -\sum_{j=i-l}^{i+l}\left(\Lambda_{j,j\pm1}(V_j) + c_j V_j\right) \text{ where } \Lambda_{j,j\pm1}(V_j) = \left(H_{j,j+1} + H_{j-1,j}\right)V_j \tag{2}$$

Equation 2 represents a finite model of length $(l+1)$, where the term V_i represents the bias potential vector, which will be determined by constructing optimized residue-specific density estimates that are related to energy through Equation 1. $H_{j,\,j+1}$ is a transition matrix for each state and represents the propensities of residues j and $j+1$ to be in each of the possible state combinations. The values for the term $H_{j,j+1}$ are derived from a database of experimental values. The two products $H_{j,j+1}V_j$ and $H_{j-1,j}V_j$ represent two different stochastic mixings of the initial vector V_j. The averaging operation is performed after transformations that are described in more detail in the Supporting Information.

The parameter c_j, which we can replace with a constant c, represents the proportional influence of the two terms and is optimized by reference to the database of experimental data. Details for computing with the above expression are presented in the Supporting Information.

Each site can be viewed as being in one of three geometrically defined states: helix (H), extended (E), or ''random coil'' (R). R is defined simply as neither H nor E. Each site has an associated energy that is coherent in the finite neighborhood l. When the coherent energy associated with a given site for one state is ''low enough'', then we say that the given residue is in one of the indicated states. When this energy (or equivalently its corresponding probability) does not reach a threshold level, we say that the given residue is in a ''chemical-shift coherent'' state (CSC state). ''CSC state'' is our designation for any site that shows distinguishing chemical shift signatures but cannot be strictly classified in one of the states (H, E, or R).

The ''core structural regions'' are heuristically defined as those for which various methods of secondary structure identification agree on the designation. The secondary structures of these core segments can be determined with a high degree of accuracy. Although the remaining regions may have no strictly definable structure, it is important to take note of their propensities, because their energy patterns may indicate pivotal residues such as turns between extended strands. In our model, the regions are given a numerical value between 1 and -1. A value of 1 indicates H, and a value of -1 indicates E. A value near zero indicates a state equivalent to what is called in the literature a random coil. Absolute values near unity are strong indications of a structured state. Fractional values indicate states with a transition classification; this can be considered as complementary to the DSSPcont approach (Carter et al., 2003), which extends the classification of secondary structure determined from 3D atom coordinates, and has the potential for analyzing inconsistencies between secondary structure designations obtained by use of different schemes for secondary structure classification.

Theoretical factors influencing the model

In theory, global phenomena that obey the Boltzmann–Gibbs distribution can be described in terms of local potentials. Conversely, a set of local potentials could lead to a global description given by the Boltzmann–Gibbs distribution. The key for

such a link is based on a theorem commonly identified as the Hammersly–Clifford theorem (Hammersley and Clifford, 1971, unpublished; Spitzer, 1971; Besag, 1974). An implication of this theorem is that the efficacy of the description given by the Boltzmann–Gibbs distribution is intimately tied to the accuracy of the local potential description. To understand the behavior of our model, its sensitivity to our estimates, and the nature of its solutions, we use an approach that combines elements of statistical mechanics and statistical decision theory (Chentsov, 1982; Janyszek and Mrugal-a, 1989; Ruppeiner, 1995).

As has been reported (Braun et al., 1994; Lukin et al., 1997; Schwarzinger et al., 2000; Wishart and Case, 2001; Wang and Jardetzky, 2002), mean random coil chemical shifts can vary by more than 2 ppm, and chemical shifts are often a single value representing a distribution (Lukin et al., 1997). Therefore, the pursuit of accuracy beyond existing methods calls for a detailed statistical analysis of chemical shift data under a "minimal" set of assumptions.

Databases

Our database of chemical shifts was derived from two sets of BMRB entries with verified matched PDB entries. The first set (derived from data downloaded in October 2003) was used to derive statistical potential estimates and subsequently to test the accuracy of secondary structure identification. The second set (from more recent entries downloaded in June 2004) was used for the purpose of testing the ability of our model to deal with "unseen" data. All entries were included that had sequence lengths ≥50, at least one reported $^{13}C^{\alpha}$ chemical shift, and reported experimental errors within a reasonable window ($\Delta\delta^{13}C < 0.4$ ppm; $\Delta\delta^{15}N < 0.4$ ppm;$\Delta\delta^{1}H < 0.04$ ppm). To avoid an artifactual increase in claimed accuracy, only one chain of each homo-oligomeric protein was included in the dataset. In total, our first database comprised ~37,000 residues from 310 different protein chains, and the second database comprised ~12,000 residues from 98 different proteins. For the more abundant amino acids (>3.5%), the average frequency difference between the two databases was ~3.2%, while for the less abundant amino acids an average frequency difference of 15% was observed. No bias in the selection of amino acids in the two sets was detected (a detailed table is presented in the Supporting Information). The secondary structure designations for the set of 310 proteins were obtained by applying the DSSP algorithm (Kabsch and Sander, 1983) to the PDB entries corresponding to the BMRB entries. We obtained the dipeptide frequencies for the corresponding conformations from the PDBselect database, dated March 2004 (Hobohm and Sander, 1994), which contains 1621 structures with specified secondary structure defined by DSSP results.

Optimized residue-specific potentials

The relationship between probabilities and energy allowed us to build accurate local energy models from the corresponding densities. A good way of approximating densities is by regularizing sample data applying a smoothing kernel at the appropriate bandwidth (Silverman, 1986). To obtain a robust estimate of densities with the "appropriate" bandwidth for our entire dataset, we exploited the difference in the qualitative behavior of two distinct kernels and their bias-variance characterization.

The *hybrid* nearest neighbor estimator can be written as

$$\hat{f}(x) = \frac{1}{n d_k(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{d_k(x)}\right) \tag{3}$$

$$K(x) = \begin{cases} 1/2 & \text{if } |x| < 1 \\ 0 & \text{otherwise} \end{cases}.$$

$$d_k(x) = \max(h, m_k(x))$$

$m_k(x)$ is the largest distance between x and its k nearest neighbors and h is a fixed value larger than expected experimental error. The sample size is n, and $\hat{f}(x)$ is the kernel estimate evaluated at x with window width $d_k(x)$. The Ep kernel is defined as

$$K_e(x) = \frac{3}{4\sqrt{5}}\left(1 - x^2/5\right)_+ \tag{4}$$

The "+" indicates that only the positive part is considered.

The hybrid nearest neighbor estimator (Equation 3) can lead to more bias in high-dimensional

settings but is more sensitive to concentrated distributions, whereas the Ep kernel (Equation 4) is smooth and provides a true density. The basic nearest neighbor method has been used for secondary structure prediction from protein sequences (Salzberg and Cost, 1992; Zhang et al., 1992; Yi and Lander, 1993; Salamov and Solovyev, 1995; Levin, 1997; Jiang, 2003).

A comparison of the hybrid nearest neighbor kernel estimates with the Ep kernel, at a given bandwidth, can then be used to test the adequacy of the number of data points and the reliability of the estimate. Let $f_e(x)$ be the density estimate using the Ep kernel and $f_e(h)$ be the estimate obtained using the hybrid nearest neighbor approach. Then,

$$(A(f > \varepsilon))^{-1} \int_{f>\delta} (f_e - f_h)_\varepsilon dx \leq \delta \quad \delta \ll 1$$
$$(f_e - f_h)_\varepsilon = \begin{cases} |f_e - f_h| & \text{if } (f_e - f_h) > \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

requires that the integrated error between two kernels, over a region where there is a non-negligible positive probability in either kernel, be small. The factor A is used to normalize for the area of the region over which integration is performed. The ε-insensitive criterion is robust and more suitable to our purposes than approaches that optimize the mean square integrated error of a single kernel.

Optimality of multidimensional chemical shifts

When multidimensional data are available, the dimension of the space in which densities should be estimated is an important consideration in determining the local potential. The optimal dimension for this space depends on the residue type and the conformational state and must be determined for each residue. The Kullback–Liebler (KL) information divergence, a measure of the distance between two distributions p and q, is defined as

$$D(p, q) = \sum p \log \frac{p}{q} \tag{6}$$

Let $p = p(x_1, \ldots, x_2, \ldots, x_n)$ and $q = p(x_1, x_2, \ldots, x_{i\text{-}1}, x_{i+1}, \ldots, x_n)\ p(x_i)$. The first term in q is obtained by considering the joint density after removing the i_{th} variable. With the addition of each dimension some information may be gained. A higher statistical dependence for additional dimensions yields more negative values for D. The additional information content must be balanced against changes in the accuracy of the local potential estimates, as measured by the accuracy of secondary structure identification. We then used this measure to find the smallest subset of chemical shift combinations, or chemical shift space (CSS) (Labudde et al., 2003), that is the most accurate.

We note that as a routine part of kernel density estimation, cross-validation is a highly recommended practice, which we follow. The methods described in this section were used to obtain estimates for combinations of N = 1,2,3 chemical shifts of different nuclei to obtain an optimal predictor for each residue type for the variety of chemical shift information available. N is the dimensionality of the CSS. To make full use of all data and to improve robustness with respect to all available data, we have employed a kernel extension approach that makes use of data available in lower dimensional CSS to obtain estimates in a higher dimensional CSS. Figure 1a shows the estimates for the nearest neighbor and Ep kernels in the $^{13}C^{\alpha}$–$^{1}H^{\alpha}$ CSS for alanine.

Computational procedure

Our computational realization of the above model involved a number of steps that are detailed in the Supporting Information. We defined the accuracy of our secondary structure identification for each of the states s (H, E, or R) in each protein chain by a parameter $Q'_s = N_s/(T_s + F_s)$, where N_s is the number of residues identified by our method to be in the particular state s and confirmed by DSSP to be in the given state, T_s is the total number of residues in the s state from DSSP, and F_s is the number of residues predicted in the s state by PECAN but not confirmed by DSSP. (Table 2 s in the Supporting Information lists the results for different CSS choices.)

We optimized the accuracy and robustness of our energetic model by varying the neighborhood size of each amino acid (l), the parameter (c), and by retaining a running tally of mean predication accuracy and variance across the dataset for each of the core secondary structure regions (Riis and Krogh, 1996). The size of the neighborhood was minimized to be the smallest size that attained the minimum error rate. Each amino acid was assigned

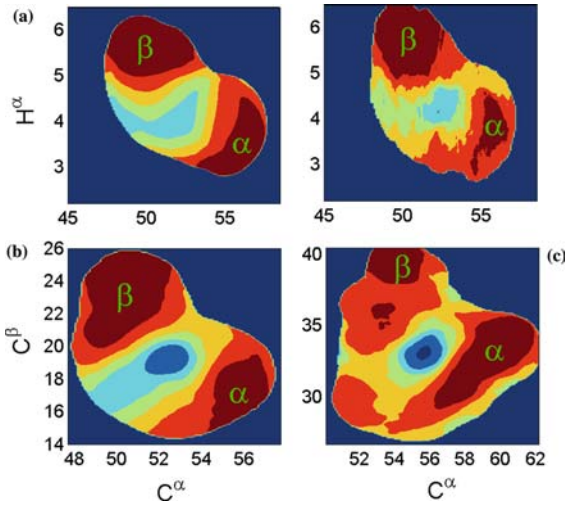

Figure 1. Estimated relative densities in two-dimensional $^{13}C^{\alpha}$–$^{1}H^{\alpha}$ and $^{13}C^{\alpha}$–$^{13}C^{\beta}$ chemical shift space (2D CSS). Alanine $^{13}C^{\alpha}$–$^{1}H^{\alpha}$ densities derived from the Epanechnikov kernel (Ep kernel) (a left) for different conformational states and from the locally adaptive nearest neighbor estimator (a right) showing qualitative agreement. Estimated densities in two-dimensional $^{13}C^{\alpha}$–$^{13}C^{\beta}$ chemical shift space (CSS) derived from the Ep kernel for (b) alanine and (c) methionine showing well separated states in alanine vs. the mixed states in methionine.

a probability vector on the basis of a density estimate designed to have the smallest error rate.

Results and discussion

In determining the accuracy of our method, we used as ''correct'' the elements of secondary structure derived from DSSP analysis (Kabsch and Sander, 1983) of three-dimensional structural models from the PDB entry identified in the BMRB entry as corresponding to the chemical shift values. In cases where multiple structures (from X-ray or NMR) were available, we used the one we considered to be best resolved. Two different Q3 accuracy results were computed for each dataset. In the first Q3 calculation, one residue at the border of two different identifiable regions was removed from consideration prior to the comparison against DSSP. These results are shown in the first row of Table 1 and demonstrate an overall Q3 accuracy of approximately 90%. In the second Q3 calculation, predictions against DSSP were made only for those residues not considered in the first calculation. Use of a uniform decision threshold across all residues in all proteins resulted in an accuracy for these regions of 60%. In other words, in regions designated by our method to be in a CSC state, our secondary structure identification differed from the DSSP designation for 40% of residues tested. The pattern of energies indicates that our approach has sufficient information content to distinguish between core structural regions and transition regions along their boundaries. For the full set of residues (core and CSC state), our agreement with DSSP was 83%.

Our average of 90% accuracy with the first dataset (310 proteins; 37,000 residues) demonstrates the ability of the model to compute accurate energetics for NMR data collected under

Table 1. Results from the application of our energetic model for chemical shift space and sequence (PECAN) to proteins of known structure and with assigned chemical shifts[a]

Residues considered	Correctly assigned residues			Q_3
	α-helix	β-strand	random coil	
Dataset A				
Core	94.0%	83.8%	89.6%	90.1%
All	86.1%	70.6%	85.7%	83.0%
Dataset B				
Core	92.9%	83.1%	89.0%	89.7%
All	86.4%	70.6%	83.8%	82.4%

[a]Restriction of potential candidates to those of length >50 amino acid residues with $^{13}C^{\alpha}$ chemical shift data, yielded 407 proteins (48,638 residues), which were divided (not selected) into two sets. Dataset A, which consisted of 310 proteins (36,491 residues), was used for model construction and testing. Dataset B, which consisted of 97 proteins (12,147 residues), was not used in any part of model construction and served primarily as a control dataset to measure robustness of our model.

varying conditions of pH and temperature (Table 1). The similar rate of success with the second dataset (97 proteins; 12,000 residues) shows the robustness of the PECAN approach in handling data not previously considered. For 99% of the proteins, our identification accuracy exceeded 75%, and for 96% of the proteins the accuracy exceeded 80%.

Our information content criteria led to the choice of residue-specific combinations of chemical shifts. For example, in the case of glutamine, a weak sheet former, the optimal set of chemical shifts for sheet prediction was provided by the 3D $^{1}H^{\alpha}$–^{13}C–^{15}N CSS. In no case was the performance improved by using more than three chemical shift values for a given residue type. Thus, although higher dimensional information has been used in the past (Wishart and Sykes, 1994; Wang and Jardetzky, 2002), our results show no justification for this.

Figure 1b provides an example of how the 2D $^{13}C^{\alpha}$–$^{13}C^{\beta}$ CSS can yield a better identification of secondary structure for a particular amino acid than the 2D $^{1}H^{\alpha}$–$^{13}C^{\alpha}$ CSS (Figure 1a) and suggests how the two can be combined to yield better overall secondary structure identifications. The estimated density for alanine based on the 2D $^{13}C^{\alpha}$–$^{13}C^{\beta}$ CSS (Figure 1b) shows a distinctly clearer separation of densities for helix, strand, and coil than the 2D $^{13}C^{\alpha}$–$^{1}H^{\alpha}$ CSS for the same residue (Figure 1a). Our results indicate that, among the 2D CSS measures, $^{13}C^{\alpha}$–$^{13}C^{\beta}$ is the best general secondary structure predictor for alanine but not for methionine (Figure 1c). Depending on the chemical shift values available, a distinct CSS can be chosen as optimal for identifying the secondary structure of each residue. The dependence of CSS on residue type holds for all residues, even when all chemical shift data are available.

Figure 2 illustrates secondary structure identification results for four proteins with different folds. The transition regions, or regions in CSC state, are shown in yellow in the probability plots (Figure 2a–d). Figure 2e shows the energetic estimates used in determining the probabilities for the protein shown in Figure 2d. Figure 3 illustrates that the accuracy of secondary structure identification increased when sequence information was included.

The optimization curves in Figure 4a indicate that the addition of sequence information tends to create a bias toward more structure in the sequence. For example, the E to R transition propensities (Supporting Information) favor extended strand over random coil. The H to R transition is more balanced, but has a slight preference for H. Note that as the accuracy for the E and H regions increases, the accuracy for R first increases and then decreases (Figure 4a). When the information mixture parameter, c is approximately 1, Q_3 reaches its maximal value. The accuracy of the predictions increased significantly when one neighbor on each side was included (Figure 4b). A smaller increase in accuracy was observed when a second neighbor on each side was included (Figure 4b), but the inclusion of additional neighbors did not improve the energy estimate or accuracy curve (Figure 4b, inset). Thus, our optimization procedure selected a five-residue stretch as the optimal length for the detection of secondary structure. This result is consistent with the idea that secondary structure formation is primarily a local phenomenon that extends over one or two neighboring residues. Including tripeptide propensities did not alter this result. Thus, given the current available data, it is unlikely that the neighborhood size for the optimal detection of secondary structure is strongly influenced by the scale of sequence interaction data (pairwise vs. three-long).

As a means of evaluating the performance of PECAN relative to the current state-of-the-art, we submitted our second database of proteins with chemical shifts to analysis by the PSSI software (Wang and Jardetzky, 2002). By use of the PSSI algorithm, Wang and Jardetzky (2002) reported a Q3 accuracy of 88% for predictions of ''highly structured'' secondary structures for 36 proteins with ~6100 residues; they compared this to 81% accuracy obtained by the CSI approach. In carrying out the analysis, we used PSSI version 2, as available from the website (http://www.pronmr.com/) on August 24, 2004. As output from PSSI, we chose the raw secondary structure identifications and the identifications achieved following application of PSSI heuristics. We then used our protocol to compare the predicted results from PSSI with those from DSSP for each structure. According to our core scoring protocol, PSSI obtained averages of 74% correct for the raw identifications and 82% correct following heuristic corrections (Figure 5). With the same dataset and

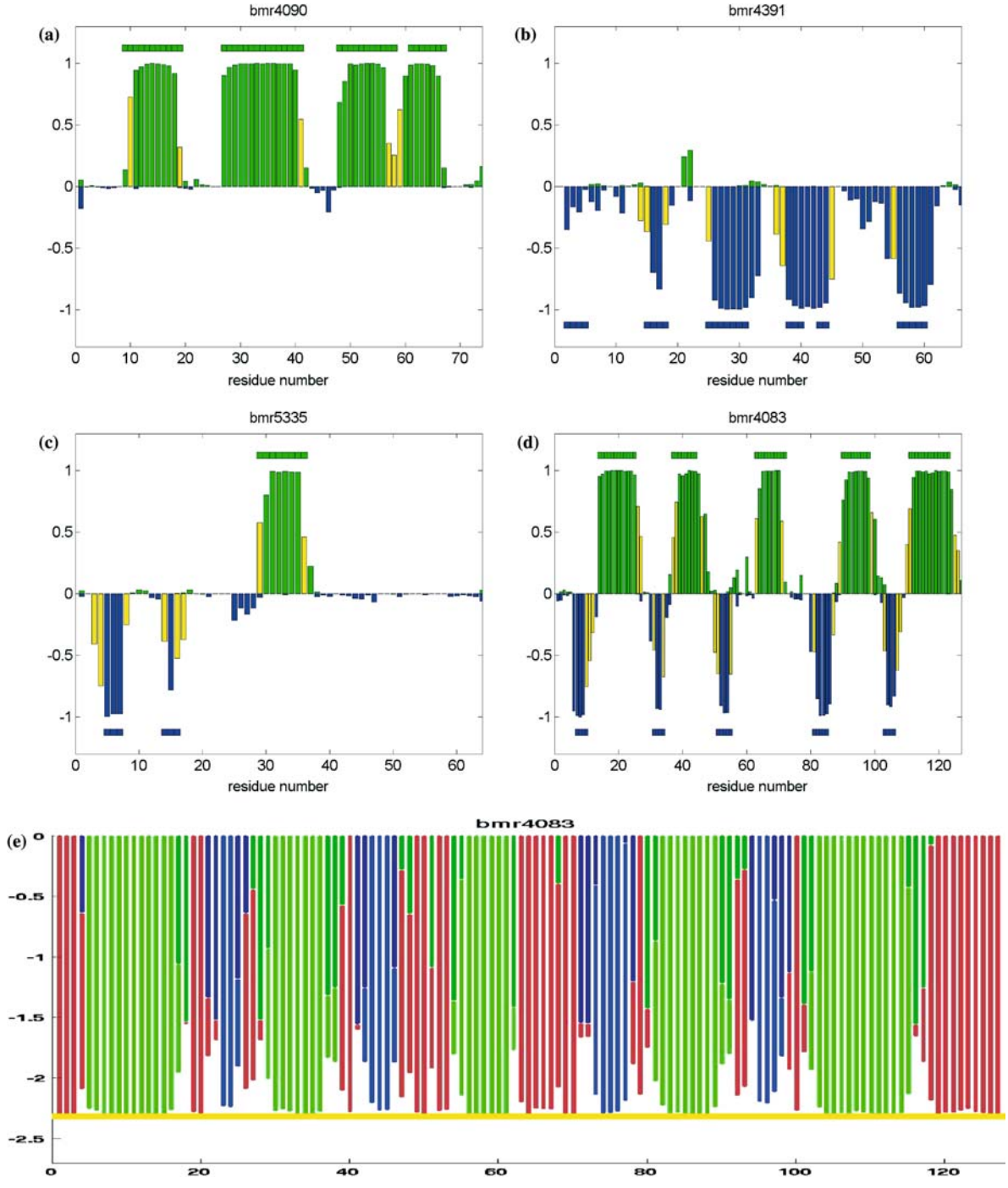
bmr4090
(a)
residue number
bmr4391
(b)
residue number
bmr5335
(c)
residue number
bmr4083
(d)
residue number
bmr4083
(e)

Figure 2. Use of the algorithm to identify the secondary structural elements in four proteins from chemical shifts and the peptide sequence. The horizontal axis represents the sequence number. The positive vertical axis represents the probability of helix (green); and the negative vertical axis represents the probability of extended structure (blue). Values near zero represent random coil residues. The yellow bars indicate identification of a region without a distinct structural designation. The horizontal green bars (above) and blue bars (below) indicate the designation, respectively, as helix and sheet by DSSP. The four proteins correspond to BMRB entries (a) 4090, (b) 4391, (c) 5335, and (d) 4083. The first residue for entries 4090 and 5335 are not shown because of missing chemical shifts. (e) The energy pattern used for the identification of secondary structure for BMRB entry 4083 (d). The horizontal axis represents the sequence number, and the negative vertical axis represents the energy level for each residue compared to the ideal stable energy level for the assigned secondary structure (shown as the yellow line across the bottom of the figure). Helix energy is shown in green, extended structure energy in blue, and red represents structural forms not classified as helix or extended. Bars with a mixture of two or three colors indicate regions where the chemical shift data point toward a structure that cannot be classified strictly into one of the three structural categories. The ideal energy scale is normalized to a common level and was selected for the best exposition.

scoring protocol, PECAN achieved 90% accuracy without heuristics, which is close to the practical limit of achievable accuracy in classifying the states.

Close inspection of the results indicates the classification accuracy of PECAN could be

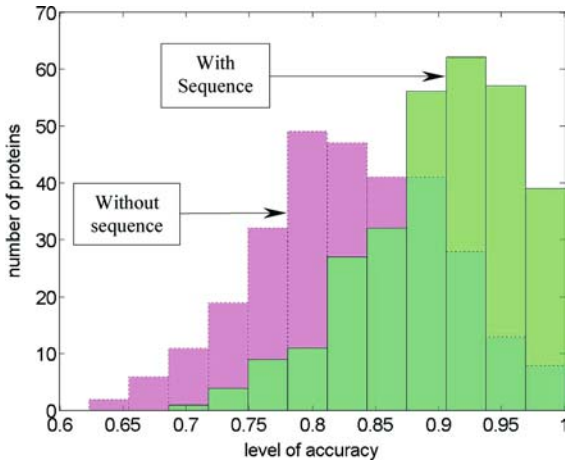

Figure 3. Distribution of Q3 for dataset 1. Light red bars represent the distribution of accuracy (1 = 100%) without the incorporation of sequence information. Green bars represent the distribution of accuracy after incorporating sequence information.

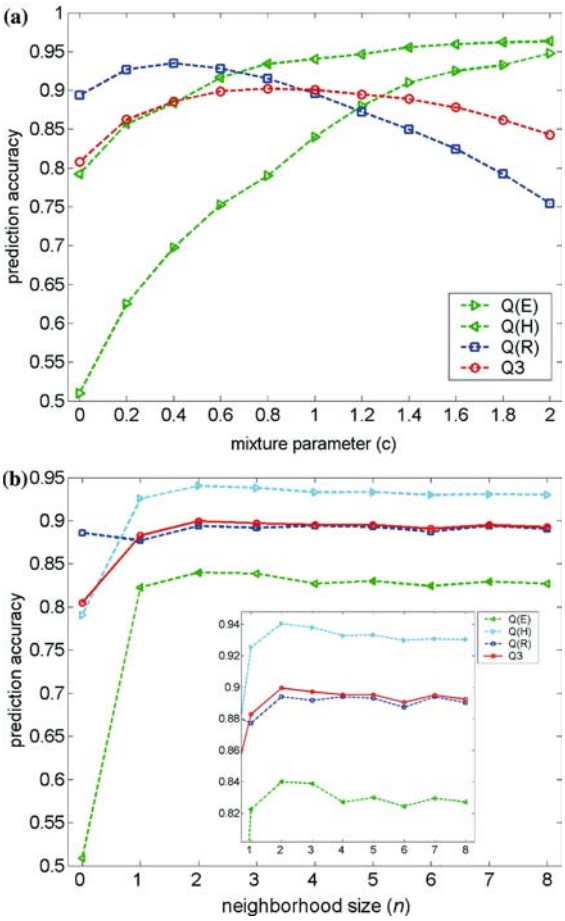

Figure 4. Illustration of the optimization step. Q3 is optimized against the parameter c that determines the relative strengths of sequence vs. bias potential. Q3 is also optimized on same set for lattice size. (a) Prediction accuracy vs. the mixture parameter c. (b) Prediction accuracy vs. $n = l+1$, the lattice neighborhood size, where $2n + 1$ is the total lattice size. In the inset, the $n = 0$ point (no sequence information) has been removed to better illustrate the impact of lattice size.

improved by the addition of heuristic rules. For example, regions identified as helix or strand with length shorter than three are less prominent in existing databases, and the reassignment of these regions as coil would improve the scores (Wishart and Sykes, 1994). Such rules also could improve the identification 3/10-helix structures (2.5% of the total residues) as an additional post-processing step. However, we have elected not to do this, because our goal has been to devise a purely energetic model that can provide insight and be refined as the database of structures with assigned chemical shifts grows.

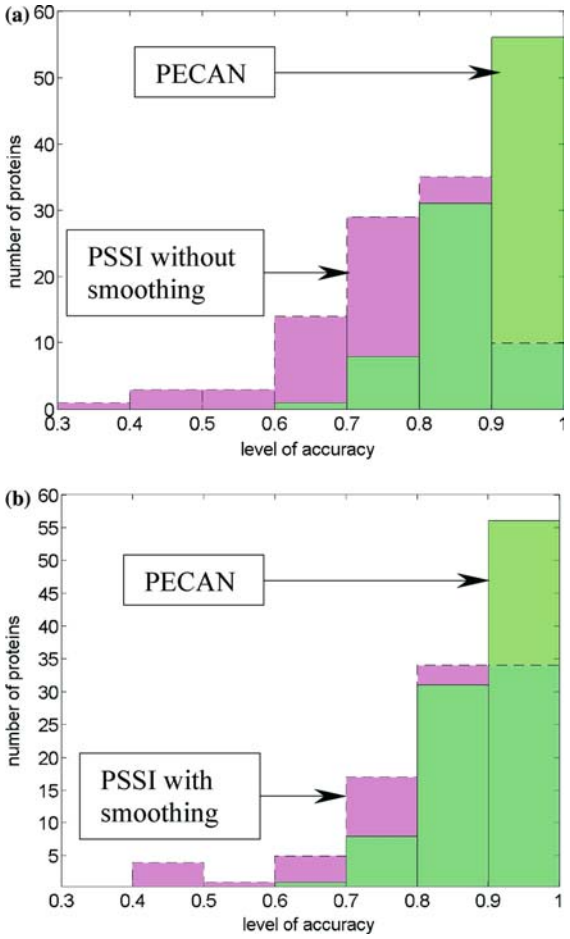

Figure 5. Comparison of results from the analysis of data set 2 by PECAN (light green) and PSSI (magenta): (a) with raw PSSI output, (b) with smoothed PSSI output. The results show that PECAN obtains very accurate secondary structure information without heuristic modifications.

It may appear surprising that better secondary structure determination comes from choosing optimal sets of limited chemical shifts for each residue type than from applying all available data. Other approaches, for example, PSSI and TALOS (Cornilescu et al., 1999) attempt to make simultaneous use of all available chemical shifts for each residue type. However, on theoretical grounds it may be expected that the chemical shifts from some atom types are more responsive to side-chain dihedral angles or tertiary structure than others. Chemical shift information not used by PECAN to identify secondary structure can be employed in developing further restraints for the determination of three-dimensional structure.

Supporting Information available: Mathematical details, tables of intermediate results, and computational information (4 tables and 7 figures) at http://dx.doi.org/10.1007/s10858-005-5705-1.

Acknowledgements

This research was supported by Biomedical Research Technology Program, National Center for Research Resources, through NIH Grant P41 RR02301, which supports the National Magnetic Resonance Facility at Madison, and by the National Institute of General Medical Science's Protein Structure Initiative through NIH grant 1 P50 GM64598, which supports the Center for Eukaryotic Structural Genomics. During part of this work H.E. was supported as a postdoctoral trainee by the National Library of Medicine under grant 5T15LM005359. We thank Eldon L. Ulrich and William M. Westler for advice and encouragement. This work made extensive use of the BioMagResBank and the Protein Data Bank.

References

Albrecht, M., Tosatto, S.C., Lengauer, T. and Valle, G. (2003) *Protein Eng*., **16**, 459–462.
Besag, J. (1974) *J. R. Stat. Soc*., **36**, 192–236.
Braun, D., Wider, G. and Wuthrich, K. (1994) *J. Am. Chem. Soc*., **116**, 8466–8469.
Carter, P., Andersen C.A. and Rost, B. (2003) *Nucleic Acids Res*., **31**, 3293–3295.
Chentsov, N.N. (1982) *Statistical Decision Rules and Optimal Inference* American Mathematical SocietyProvidence, RI.
Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.
Hobohm, U. and Sander, C. (1994) *Protein Sci*., **3**, 522–524.
Hung,, L.H. and Samudrala, R. (2003) *Protein Science*, **12**, 288–295.
Iwadate, M., Asakura, T. and Williamson, M.P. (1999) *J. Biomol. NMR*, **13**, 199–211.
Janyszek, H. and Mrugal-a, R. (1989) *Phys. Rev. A*, **39**, 6515.
Jiang, F. (2003) *Protein Eng*. **16**, 651–657.
Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577–2637.
Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1996) *Protein Sci*., **5**, 1067–1080.
Labudde, D., Leitner, D., Kruger, M. and Oschkinat, H. (2003) *J. Biomol. NMR*, **25**, 41–53.
Le, H. and Oldfield, E. (1994) *J. Biomol. NMR*, **4**, 341–348.
Lesk, A.M. and Rose, G.D. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 4304–4308.
Levin, J.M. (1997) *Protein Eng*., **10**, 771–776.
Luginbuhl, P., Szyperski, T. and Wuthrich, K. (1995) *J. Magn. Reson. B*, **109**, 229–233.
Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.

Markley, J.L., Meadows, D.H. and Jardetzky, O. (1967) *J. Mol. Biol.*, **27**, 25–35.
Moult, J. (1997) *Curr. Opin. Struct. Biol.*, **7**, 194–199.
Przytycka, T., Aurora, R. and Rose, G.D. (1999) *Nat. Struct. Biol.*, **6**, 672–682.
Reiley, M.D., Thanabal, V. and Omecinsky, D.O. (1992) *J. Am. Chem. Soc.*, **114**, 6251–6252.
Riis, S.K. and Krogh, A. (1996) *J. Comput. Biol.*, **3**, 163–183.
Ruppeiner, G. (1995) *Rev. Mod. Phys.*, **67**, 605 .
Salamov, A.A. and Solovyev, V.V. (1995) *J. Mol. Biol.*, **247**, 11–15.
Salzberg, S. and Cost, S. (1992) *J. Mol. Biol.*, **227**, 371 374.
Schwarzinger, S., Kroon, G.J., Foss, T.R., Wright, P.E. and Dyson, H.J. (2000) *J. Biomol. NMR*, **18**, 43–48.
Sibley, A.B., Cosman, M. and Krishnan, V.V. (2003) *Biophys. J.*, **84**, 1223–1227.
Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis* Chapman and Hall, New York, NY.
Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.
Spitzer, F. (1971) *Am. Math. Mon.*, **78**, 142–154.
Sternlicht, H. and Wilson, D. (1967) *Biochemistry*, **6**, 2881–2892.
Wang, Y. and Jardetzky, O. (2002) *Protein Sci.*, **11**, 852–861.
Wishart, D.S. and Case, D.A. (2001) *Methods Enzymol.*, **338**, 3–34.
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.
Yi, T.M. and Lander, E.S. (1993) *J. Mol. Biol.*, **232**, 1117 1129.
Zhang, X., Mesirov, J.P. and Waltz, D.L. (1992) *J. Mol. Biol.*, **225**, 1049–1063.