Published in final edited form as:
Structure. 2009 November 11; 17(11): 1515–1527. doi:10.1016/j.str.2009.09.006.

Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction

Eshel Faraggi[a,†], **Yuedong Yang**[a,†], **Shesheng Zhang**[a], and **Yaoqi Zhou**[a,b,*]

[a] Indiana University School of Informatics, Indiana University-Purdue University and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 719 Indiana Ave., Walker Plaza Building Suite 319, Indianapolis, IN 46202, USA

[b] Kavli Institute for Theoretical Physics China, CAS, Beijing 100190, China

Abstract

Local structures predicted from protein sequences are employed extensively in every aspect of modeling and prediction of protein structure and function. For more than 50 years, they have been predicted at a low-resolution coarse-grained level (e.g. three-state secondary structure). Here, we combine a two-state classifier with real-value predictor to predict local structure in continuous representation by backbone torsion angles. The accuracy of the angles predicted by this approach is close to that derived from NMR chemical shifts. Their substitution for predicted secondary structure as restraints for ab initio structure prediction doubles the success rate. This result demonstrates the potential of predicted local structure for fragment-free tertiary-structure prediction. It further implies potentially significant benefits from employing predicted real-valued torsion angles as a replacement of or supplement to the secondary-structure prediction tools employed almost exclusively in many computational methods ranging from sequence alignment to function prediction.

Prediction of the local structure of proteins is dominated by secondary structure prediction for which the accuracy has stagnated between 77% and 80% for more than a decade (Rost, 2001; Dor and Zhou, 2007a). This stagnation contributes to the slow progress of ab-initio structure prediction that depends on predicted secondary structure to reduce the conformational space (Simons et al., 1997; Ortiz et al., 1998; Eyrich et al., 1999; Hardina et al., 2000; Fain and Levitt, 2003; Nanias et al., 2003) or guide folding pathways (Ozkan et al., 2007; DeBartolo et al., 2009). The accuracy of secondary structure prediction further significantly affects the quality of many applications including multiple sequence alignment (Zhou and Zhou, 2005), fold recognition (Fischer and Eisenberg, 1996) and prediction of structural disorder and flexibility (Schlessinger and Rost, 2005; Young et al.,1999; Radivojac et al., 2007), and function (Godzik et al., 2007). However, the effectiveness of predicted secondary structure is often limited by its coarse-grained representation of local backbone structures. Helices and strands often deviate from their ideal shapes in protein structures while the structures commonly characterized as coils do not have a definite shape. Moreover, defining secondary

[*] To whom correspondence should be addressed: Phone: (317) 278-7674, Fax: (317) 278-9201, yqzhou@iupui.edu.
[†] Equal contribution

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

structure is somewhat arbitrary. The arbitrarily defined boundaries and structural flexibility limit the theoretically possible prediction accuracy to 88–90% (Rost, 2001; Kihara, 2005).

Local structures of proteins can be exactly described by their backbone torsion angles in an unambiguous way. In practice, these continuous angles were discretized because various secondary-structure types are clustered at different regions in the Ramachandran $\phi-\psi$ diagram (Ramachandran and Sasisekharan, 1968). Discretized angle states were employed as a replacement of, or supplement to, secondary structure for refined local-structure classifications. They were also utilized to construct simplified protein models for sampling efficiency. This development led to sequence-based methods for multi-state torsion-angle prediction (Gibrat et al., 1991; Rooman et al., 1991; Kang et al., 1993; Bystroff and Baker, 1998). Unfortunately, multi-state torsion angles are as difficult as secondary structure to predict accurately. For example, a recent study (Zimmermann and Hansmann, 2008) shows that a three-state prediction has an accuracy of 79%, the same level of accuracy of secondary structure prediction (Dor and Zhou, 2007a). Lower accuracy is reported in earlier studies with more states defined (Bystroff et al., 2000; Kuang et al., 2004). No apparent advantage of predicted multi-state torsion angles over predicted secondary structure led to few applications that actually use predicted multi-state torsion angles.

Recently, we have demonstrated that local structures of proteins can be predicted in continuous representation with reasonable accuracy. We introduced Real-SPINE (Real-value prediction of protein Structural Properties by Integrated NEural network) for a real-value prediction of ψ (Dor and Zhou, 2007b) and ϕ torsion angles (Xue et al., 2008) and subsequently improved its accuracy by guided learning and other techniques (Real-SPINE 3) (Faraggi et al., 2009). However, not all ϕ and ψ angles can be sampled by the backbone of proteins because of internal steric constraints between backbone atoms (Ramachandran and Sasisekharan, 1968). These three-dimensional physical constraints are difficult to learn by a machine learning technique with one-dimensional sequence-based information as the only input. By contrast, multi-state prediction avoids this problem by defining discrete $\phi - \psi$ states in sterically allowed regions only.

Here, we propose an approach that combines the advantage of discrete state classifier (avoiding sterically prohibited regions in the Ramachandran $\phi - \psi$ diagram) with that of real-value prediction (removing arbitrary definition of discrete states). We refer to this method as SPINE X with X denoting the combination of the two approaches. The resulting predicted angles are further refined with a conditional random field model (SPINE XI with I denoting refinement). The final predicted angles are not only substantially more accurate than multi-state classifiers for predicting discrete states but also their accuracy is close to the accuracy of the real-value angles derived from NMR chemical shifts. We further found that this new level of accuracy and real-value detail of predicted local structures brings substantial improvement in the success rate of fragment-free tertiary-structure prediction. The result has significant implication for other computational methods that have relied on discrete secondary structure prediction for characterizing unknown local structures of proteins.

Results

Ramachandran and his coworkers (Ramachandran and Sasisekharan, 1968) found that not all torsion angles of protein backbones are sampled by proteins because of internal steric constraints. Fig. 1d shows the statistical distribution of native angles of a database of 2640 proteins (Dor and Zhou, 2007a). The angles are distributed around preferred values, as determined by their secondary structure states, while in the rest of the $\phi-\psi$ plane no angles are found. These forbidden regions caused by three-dimensional physical constraints are difficult to learn by a machine learning technique with one-dimensional sequence-based information as the only input. Indeed, Fig. 1a shows that many angles predicted by a previously-developed real-value predictor (Real-SPINE 3) are in sterically forbidden regions when compared to the same diagram for native values of torsion angles (Fig. 1d). In fact, these predicted angles cover essentially every area within a square boundary.

To solve this problem, we observed that both ϕ and ψ angles have a bimodal distribution as revealed from the four populated regions in Ramachandran diagram of native angles (Fig. 1d). That is, we can make an unambiguous assignment of two separate states for each angle. Following this line of reasoning, we first classified ψ (or ϕ) into two states (peak I or peak II), i.e., made a two-state prediction of the angle. The two-state prediction was followed by a prediction of the real-value deviation of the native angle from the peak. Fig. 1b indicates that this SPINE X method separates predicted torsion angles into four isolated regions as designed and eliminates angles in non-allowed regions. However, predicted angles are distributed too narrowly around predefined centers. Much wider distributions are observed in the native Ramachandran diagram (Fig. 1d). Moreover, the population distribution in the bottom right is approximately parallel to the axis ϕ (Fig. 1b), rather than to the $\phi = -\psi$ axis (Fig. 1d).

To further improve the accuracy of angle prediction given by SPINE X, we applied a conditional random field (CRF) model (Lafferty et al., 2001) that utilizes both ϕ^{NN} and ψ^{NN} produced by the neural networks to predict these errors (i.e., $\phi^{Native} - \phi^{NN}$ and $\psi^{Native} - \psi^{NN}$). We employed the CRF model because a combination of two totally different methods often yields improvement in machine learning. In particular, the CRF model optimizes the conditional probability of the entire sequence, information that is not utilized as input to the neural networks. Furthermore, the CRF model takes into account the possible coupling between ϕ and ψ angles by using both angles together. The predicted errors from the CRF model are used to refine the prediction of torsion angles. As Fig. 1c shows, the $\phi - \psi$ diagram obtained from SPINE XI is much closer to the native angle distribution (Fig. 1d).

The improvement of the angle distribution in the Ramachandran diagram from Fig. 1a to Fig. 1c is accompanied by improvement in prediction accuracy. One can measure accuracy by $Q_{60°}$, the fraction of residues for which both predicted angles are within 60 degree from their corresponding native values (Cavalli et al., 2007). We found that $Q_{60°}$ increases from 74.6% by Real-SPINE 3 (Faraggi et al., 2009) to 81.5±0.4% by SPINE X and to 82.7±0.4% by SPINE XI. That is, there is an 8% absolute (11% relative) improvement from Real-SPINE 3. By comparison, if ideal angles are used for predicted helical and strand residues and an average angle for coil residues (−25.9° for ϕ and 22.3° for ψ), $Q_{60°}$ is only 61% (a separate average-coil angle value for each residue type would increase $Q_{60°}$ to 64%). The large improvement can be confirmed by other measures of the accuracy. For example, $Q_{36°}$, the fraction of residues for which both predicted angles are within 36° from their corresponding native values, increases from 62% in Real-SPINE 3 to 72% in SPINE X and 74% in SPINE XI, a total of 12% absolute improvement. The mean absolute error (MAE) of the angles is defined as the average absolute difference between predicted and measured angles, e.g., for $\psi MAE=\sum_{i=1}^{N}\left|\psi_i^{\text{pred}}-\psi_i^{\text{Native}}\right|/N$. The MAE for ψ reduces from 36.1±0.8° by a “pure” real-value prediction method (Real-SPINE 3) (Faraggi et al., 2009), to 35.2±0.6° by SPINE X, and to 33.4±0.7° by SPINE XI. This is a total of 2.7° (7%) reduction in MAE from Real-SPINE 3. The CRF refinement produces a smaller improvement in $Q_{60°}$ or $Q_{36°}$ than it did for the MAE because the CRF tends to make small (less than 60°) corrections to the angles. For example, $Q_{10°}$, the fraction of residues for which both predicted angles are within 10° from their corresponding native values, increases from 29% in SPINE X to 34% in SPINE XI, a 5% absolute improvement. Significant improvement from Real-SPINE 3 to SPINE XI (11% relative improvement in $Q_{60°}$) confirms the advantage of combing a two-state classifier with a real-value predictor followed by subsequent refinement.

We further found that the accuracy of angle prediction strongly depends on the secondary structure types. The prediction accuracy is the highest for helical residues. $Q_{60°}$ = 95.2%, 89.4%, and 64.7% for helical, strand, and coil residues, respectively. These values are 0.1%, 1.0%, and 2.6% improvement from the respective values prior to the CRF refinement. Torsion angles of coil residues are most difficult to predict because of their irregular structures. Thus, predicted secondary structure types can be used as a coarse-grained estimation of the accuracy of angle prediction. The proposed SPINE XI method can be compared to multi-state predictors by mapping real-value prediction onto pre-defined states. Kuang et al. (Kuang et al., 2004) and Bystroff et al. (Bystroff et al., 2000) followed a definition that divides the $\phi-\psi$ plane into 81 square grids and predicted grid-defined four states (A, B, G, and E) (Oliva et al., 1997) by neural networks and hidden Markov model, respectively. We achieved 84% correct identification, compared to 74% by Bystroff et al. (Bystroff et al., 2000) and 77% by Kuang et al. (Kuang et al., 2004). Zimmermann and Hansmann (Zimmermann and Hansmann, 2008), on the other hand, predicted 16-state 5-residue building blocks of 8 consecutive torsion angles (de Brevern et al., 2000). They achieved 61% for correctly predicting 16 states and 79% for predicting three secondary-structure states. By comparison, we obtained 66% and 81%, respectively. This 5% or 2% improvement is made despite that both SPINE X and XI were trained to predict the angles of a single residue at a time. Thus, multi-state assignment based on real-value prediction is more accurate than the methods dedicated to predict them whether or not they are defined in single residue, in a fragment of several residues, in a small number of states, or in a large number of states.

How accurate are real value torsion angles predicted by SPINE XI relative to the values obtained from NMR chemical shifts? Recently, Cavalli et al. (Cavalli et al., 2007) developed and used a method called TOPOS, which is similar to TALOS (Cornilescu et al., 1999), to determine the backbone torsion angles most compatible with the experimental chemical shifts in a database of three-residue fragments. They applied TOPOS to 11 proteins and the results are compared to SPINE XI in Fig. 2 (also Table 1). The average accuracy is 88% by TOPOS (Cavalli et al., 2007) and 86.9% by SPINE XI.

To further confirm the result described above, we obtained a set of proteins with both the chemical shift data and PDB structures from the BioMagResBank database (Doreleijers et al., 2003) (http://restraintsgrid.bmrb.wisc.edu/NRG/MRGridServlet). We removed the proteins whose sequences from the chemical shift data mismatch with the corresponding sequences from the PDB structures. This led to a total of 37 proteins. For some proteins with multiple PDB entries, we used the X-ray crystal structure with the highest resolution. This set of proteins is a challenging set because the average fraction of coil residues (43%) is significantly higher than either the above 11-protein set (36%) or the training/test benchmark of 2640 proteins (39%) (Dor and Zhou, 2007a) (See methods). Nevertheless, the accuracy given by TALOS (Cornilescu et al., 1999) based on experimental chemical shifts (79.8%) is only 4% higher than the proposed method (75.5%), a successful result considering that our method is trained entirely on X-ray structures (Fig. 2 and Table 2). A smaller difference (2.6%) in accuracy (89.3% by TALOS and 86.7% by SPINE XI) is observed when X-ray crystal structures (a total of 7) are used for obtaining native torsion angles as shown in Fig. 2 and Table 2.

The above comparison in $Q_{60°}$ does not mean that the accuracy of our predicted local structures is close to the accuracy of local structural information contained in NMR chemical shifts. Both TOPOS and TALOS are only approximate methods that derive torsion angles from NMR chemical shifts. In fact, NMR chemical shifts, rather than the torsion angles derived from NMR chemical shifts, are directly employed as restraints in structure prediction (Cavalli et al., 2007; Gong et al., 2007; Shen et al., 2008; Wishart et al., 2008; Montalvao et al., 2008; Robustelli et al., 2008; Shen et al., 2009). The purpose of the comparison made here is only to illustrate the new level of accuracy achieved by SPINE XI in real-value torsion-angle prediction.

It is of interest to know the relative contribution of various inputs employed in neural networks to the accuracy of torsion angle prediction. There are three key components for neural-network inputs: sequence profiles, representative residue properties, and predicted one-dimensional structure properties (secondary structure and solvent-accessible surface area) [See methods]. Table 3 shows the effect of different input combinations on Q_{36° for ϕ and ψ, respectively. It is clear that sequence profile or predicted one-dimensional structure properties alone achieve similar accuracy while the combination of the three inputs further improves the accuracy by an additional 1–2% for ϕ and 3% for ψ. The improvement, although small, reaffirms the usefulness of combined inputs. In the method described above, peak prediction and the deviation from the peak are predicted separately. We also tested the idea of including predicted peaks as an input for predicting peak deviations. We did not observe significant improvement in either Q_{60° or Q_{36°. This suggests that predicting the deviation from the peak has utilized the peak location implicitly contained in same neural-network inputs employed for peak prediction.

To illustrate the usefulness of the improved real-value prediction relative to that of secondary structure prediction, we predicted protein structures via restraining local structure represented by three predicted secondary-structure states and by continuous representation of predicted torsion angles. A coarse-grained energy function is employed in the absence of any native information (i.e. no protein fragments or templates). Here, we examine whether or not the restraints based on predicted torsion angles or predicted secondary structure can guide a coarse-grained energy function to the right structural folds (Root-mean-squared distance, RMSD, <6Å) (Reva et al., 1998) within top predicted structures – the critically important first step of protein structure prediction (see Methods). We defined a successful prediction of a structural fold if the best structure in top 15 predicted structures has an RMSD of 6Å or less.

Fig. 3 displays one typical example of the effects of different angle restraints on conformational sampling and structure prediction. The energy values of sampled structures for protein 1shf are shown as a function of their RMSD values with various restraints as labeled from top to bottom in Fig. 3. Without any restraints, the RMSD values of all sampled structures are greater than 8Å. Adding the restraints based on ideal secondary structure (only predicted strand residues for this protein) improves the best-sampled structure to near 6Å RMSD. However, the best structure in top 15 predicted structures (Fig. 3 Structure b) continues to have a high RMSD value of 8.5Å. Restraining around real values of predicted torsion angles for predicted strand residues leads to β strands of the best structure in top 15 that are non-ideally shaped but more native-like (Fig. 3 Structure c). The RMSD value of the best sampled structure and that of the best structure in top 15 reduce to 4.5Å and 6.9Å, respectively. It is clear that the predicted deviation from an ideal β strand conformation leads to significantly more effective conformational sampling of near native structures. Adding the predicted restraints of predicted coil residues further decreases the RMSD values to 2.5Å for the best sampled structure and to 3.1Å for the best structure in top 15. This best structure in top 15 has the lowest energy. That is, the structure information contained in coil residues leads to the correct prediction of the structure for 1shf.

Table 4 summarizes the results for a benchmark of 16 proteins (Bradley et al., 2005). Adding predicted secondary-structure restraints to a coarse-grained energy increases the number of successful prediction from 2 to 6 out of 16 proteins and the median RMSD value for the best in top 15 decreases from 9Å to 6Å. This significant change results from highly accurate secondary structure prediction for this set of benchmark proteins (an average Q_3 of 85%). More importantly, this level of success can be further significantly improved by simply replacing ideal angles of predicted helical and strand residues with predicted real values of the same restrained residues. The median value of the best in top 15 structures decreases by an additional 1Å (from 6Å to 5Å) and the number of successful predictions increases by another 67% (from 6 to 10). Thus, predicting torsion angles in real values makes more powerful restraints than predicting secondary structure in ideal representation even before the predicted angles of coil residues are put into use.

Adding restraints on coil residues further improves the overall accuracy of structure prediction. As shown in Table 4, additional restraints on predicted coil residues increase the number of successful predictions by another 20% from 10 to 12 and decrease the median RMSD value by an additional 1Å from 5Å to 4Å. The improvement is observed for the majority of proteins (11 out of 16). Thus, employing the real-value prediction doubles the success rate (from 6 to 12) over employing the predicted secondary structure, by taking into account non-ideal helical and strand conformations and the structures of coil residues. As shown in Fig. 4, the predicted structures reproduce the overall structures of the 12 native structures very well.

There are only four exceptions to the improvement with additional restraints on coil residues. A large increase in the RMSD value of the best structure in top 15 is observed for three proteins: 1b72, 1o2f, and 1mky. While restraining all residues for these three proteins is not successful, restraining only predicted helical and sheet residues successfully predicts their respective structural folds within top 15 structures. The RMSD values of the best structures in top 15 for the three proteins are 6.5Å, 9.7Å, and 8.0Å, respectively, with coil restraints, and 5.1Å, 5.7Å, and 4.5Å, respectively, without coil restraints. This suggests a strategy that combines top candidates predicted with and without coil restraints for further all-atom refinements. Such a simple combination will lead to correct structural folds within top 30 structures for 15 out of 16 proteins. This result also highlights that there is room for further improvement in torsion angle prediction of coil residues.

The above improvement is based on RMSD values. It is of interest to know if other criteria (Sippl, 1982; Siew et al., 2000; Zhang and Skolnick, 2004a) for measuring the accuracy of structures predicted would reveal similar improvement as well. Here we apply the recently developed TM-Score (Template-modeling score) that was designed to be independent of protein sizes. A TM-Score is 1 for perfect match and below 0.2 between two random structures. The median TM-score for the best in top 15 is 0.30, 0.38, 0.45, and 0.54 as we change from no angle restraints, ideal helical and strand restraints, real-value helical and strand restraints to angle restraints on all residues. That is, there is a 42% improvement in structural quality from secondary structure to real-value torsion-angle restraints. This 42% improvement is higher than 32% reduction in RMSD from 6.3Å to 4.3Å. Similar result (38% improvement from secondary structure to real-value torsion-angle restraints) is obtained if the MaxSub score is employed (Siew et al., 2000).

This work predicts ϕ and ψ torsion angles only and all ω torsion angles are fixed to 180° in structure prediction. However, there is a small population of Pro residues that adopt the cis conformation ($\omega = 0°$). For the 16-protein benchmark, this is true for Pro 394 in 1mky, Pro 54 in 1tif, and Pro 20 and Pro 72 in 1dcj. To test the effect of the cis conformation, we fixed ω to 0° for Pro 394 in 1mky and Pro 54 in 1tif. This leads to a significant reduction of the RMSD value of the best structure in top 15 prediction for both 1mky (from 8.0 to 5.2Å) and 1tif (from 4.2 to 3.4Å). Thus, a large improvement of one torsion angle can impact significantly the accuracy of predicted structures. On the other hand, incorrect ω angles do not prohibit the ability to predict correct structural folds for 1tif (4.2Å RMSD, ranked 9) and 1dcj (3.2Å RMSD, ranked 1). That is, the error of a few incorrect angles can be corrected by the adjustment of other torsion angles. This explains the observed success of imperfectly predicted torsion angle restraints for structure prediction. In principle, one could also develop a method to predict ω angles as well. However, the rare events of cis conformations are more difficult to predict because they mostly result from nonlocal interactions. We will incorporate cis conformations during conformational sampling in future studies.

Finally, we are unable to locate an obvious reason for the inability to predict correct structural folds for 1tig within top 15 with or without restraints. Nevertheless, the best structure sampled with restraints (6.2Å RMSD) in all sampled structures is substantially more accurate than the best structure in the absence of any restraints (8.9Å RMSD). It is quite possible that the restraints for this protein are unable to correct or compensate for the error contained in the two-term coarse-grained energy function.

Further Discussion and Concluding Remarks

This paper demonstrates a new level of accuracy achieved for predicted local structures either in exact description by real-value torsion angles or by mapping to multiple predefined states. The application to structure prediction reveals that predicted real-value torsion angles are substantially more effective than predicted secondary structure as local structural restraints for tertiary structure prediction, a technique commonly used in structure prediction methods including ROSETTA (Simons et al., 1997) and TASSER (Zhang and Skolnick, 2004b). These results highlight the power of the newly proposed approach for local structure prediction: defining only states that have clear boundaries from other states and further refining the states by real-value assignments within each one. SPINE XI is available as a server at http://sparks.informatics.iupui.edu.

More importantly, almost all sequence-based prediction methods utilized predicted secondary structures. Examples are protein domain divisions (Cheng et al., 2006; Gewehr and Zimmer, 2006; Tress et al., 2007), dynamical properties of structures (residue fluctuation or temperature B-factor (Schlessinger and Rost, 2005; Yuan et al., 2005), conformationally variable or switching regions (Yoon and Welsh, 2005; Young et al., 1999; Gross, 2000; Boden et al., 2006; Kuznetsov, 2008), intrinsic disorder (Ferron et al., 2006; Dosztanyi et al., 2007; Bourhis et al., 2007; Radivo jac et al., 2007)), and function prediction methods (See recent reviews (Godzik et al., 2007; Yang, 2004; Lopez et al., 2007)). In fact, Libley et al. showed that predicted secondary structures make the largest contribution to function prediction (Lobley et al., 2007). Several initial applications of torsion angles predicted by earlier methods appear to be promising for improving fold recognition (Karchin et al., 2003; Wu and Zhang, 2008; Zhang et al., 2008) and sequence alignment (Huang and Bystroff, 2006). Because our newly predicted torsion angles are more accurate even when mapped onto a few coarse-grained states, a simple update or addition of real-value local structure prediction will likely improve those computational methods relying on the accuracy of secondary structure prediction.

Lastly, it is worth mentioning an alternative approach. We have employed predicted angles as restraints for structure prediction. Another possibility is to predict preference of torsion angles given sequence and secondary structure information during conformational sampling using a probabilistic model (Boomsma et al., 2008; Zhao et al., 2008). Both continuous sampling approaches provide an alternative to discrete fragment-based sampling.

Method

A. Torsion-angle predictions by neural networks (SPINE X)

Torsion angles were first predicted by a neural-network based method with sequence as its only input. The proposed method combines the advantage of multi-state prediction (avoiding sterically prohibited regions) and that of real-value prediction (Xue et al., 2008; Faraggi et al., 2009) (removing arbitrary definition of states). This is done by first designating the two most popular ϕ and ψ angles as the peaks in their distribution. Then, two separate neural networks are employed. The first neural network is used for predicting the peak that ϕ or ψ is closest to. The second neural network is used for predicting the angle deviation from the peak. The results from the two networks are combined to yield the real values of torsion angles. ϕ and ψ angles are predicted separately. The architecture of neural networks can be found in Fig. 5. The dataset (2640 non-redundant high-resolution protein structures with 25% sequence identity cutoff), training and testing (ten-fold cross validation) used in this work are the same as those published previously (Xue et al., 2008; Faraggi et al., 2009). More specifically, this dataset was built on the protein sequence culling server PISCES (Dunbrack, 2006) with sequence identity less than 25% and X-ray resolution better than 3Å. The chains with unknown structure regions were removed. The final dataset contains a total of 591,797 residues.

Input for the neural networks—We used a window size of 21 amino acids with ten residues to either side of the central residue whose angles we wish to predict. Vacant locations in the windows around residues near the terminals of the protein are explicitly excluded from the machine learning by limiting the size of the window for those regions. Each amino acid residue has 31 input features. They include the seven representative amino acid properties (steric parameter, hydrophobicity, volume, polarizability, isoelectric point, helix probability and sheet probability) (Dor and Zhou, 2007a; Meiler et al., 2001), 20 values from the Position Specific Scoring Matrix (PSSM) obtained from PSI-BLAST (Altschul et al., 1997) with three iterations of searching against a non-redundant sequence database, a real-value solvent accessibility prediction (Faraggi et al., 2009), and predicted three-state secondary-structure probabilities (3 values). The latter is obtained from an improved version of SPINE (Dor and Zhou, 2007a) based on guided learning and predicted torsion angles (Faraggi et al., 2009) (details will be published elsewhere).

Network architecture—As shown in Fig. 5, we employ two separate neural networks: one makes a two-state prediction (Peak I and Peak II) and the other predicts the deviation from the peak. The two-state prediction has two output neurons that represent the absolute distance from peak I and II, respectively. A smaller distance indicates the peak location. The network for peak deviation has one output neuron that predicts the difference between the angle and its peak location of native-angle distribution. Both networks consist of two hidden layers with a bipolar activation function [tanh(αx) with α = 0.2] (Faraggi et al., 2009). Each hidden layer contains 51 neurons. For ϕ, peaks I and II are located at −62° and 57°, respectively. For ψ, they are located at −40° and 140°, respectively. The final predicted real-value angle is the location of the predicted peak (one of the four angle pairs above) plus the predicted deviation. ϕ and ψ angles are predicted separately. Each predicted angle is a consensus prediction over five predictions independently trained by different random initial weights. This is done first by a consensus vote on the peak assignment and then by averaging the predicted deviations from the peaks over the five runs. The final predicted angles are the sum of the consensus-peak position and the averaged deviation from the peak. The accuracy of peak prediction is 96.3% for ϕ and 85.9% for ψ.

Algorithm—ψ angles are shifted by 50° such that the two peaks are situated approximately symmetrically about 0°. ϕ angles are not shifted. Neural network weights are trained by the back propagation algorithm (Rumelhart et al., 1986) with guiding factors designed to satisfy the intuitive condition that for most residues, the contribution of a residue to the structural properties of another residue is smaller for greater separation in the protein-sequence distance between the two residues (Faraggi et al., 2009). The learning rate and momentum are set as 0.01 and 0.4, respectively. The initial weights are randomly selected in the range (−0.5, 0.5) and all inputs to the neural networks are normalized in the range of (−1, 1). The weights training is tested with a 10 fold cross validation on the database of 2640 proteins with sequence identity less than 25% between them and X-ray resolution better than 3Å(Zhou and Zhou, 2004). The dataset is divided randomly into 10 groups with 264 proteins each. Each time, nine groups are used for training and one group is used for independent testing. This procedure is repeated 10 times and the prediction accuracies is averaged over the ten folds. In addition, 5% of the training set is excluded from training and used solely as an independent convergence test to avoid possible over-training (overfitting protection). The final predicted ϕ and ψ angles are labeled as ϕ^{NN} and ψ^{NN}, respectively.

B. Torsion-angle refinement by conditional random field models (SPINE XI)

The real values of both ϕ and ψ torsion angles predicted by SPINE X are used as an input to the conditional random field model (Lafferty et al., 2001) to predict the errors of those predicted angles. A CRF model (Lafferty et al., 2001), like a statistical hidden Markov model (HMM) (Baum and Petrie, 1966), is an undirected graphical (random fields) model. Unlike HMM, it directly computes the distribution of a to-be-predicted variable conditioned on known observations. CRF models can contain any number of feature functions and have been recently applied to protein fold recognition (Liu et al., 2006) and conformational sampling (Zhao et al., 2008). In our case, we used it to predict the errors in neural-network predicted angles ($\Delta\tau = \tau - \tau^{NN}$ with τ either ϕ or ψ) conditioned on given observations such as residue type or predicted secondary structure or their combination. These errors are employed to refine the predicted angles.

State-dependent angles—One innovative implementation of the CRF model in this work is to weight observations according to their predictive power for Δτ (see below). We used 60 states to characterize different combinations of 20 amino-acid residue types and predicted three-state secondary structures. The state-dependent angle ($T_m(X)$, m = 1, 60) is approximated as a linear combination of predicted angles and corresponding position specific scoring matrix from PSI-BLAST (P) (Altschul et al., 1997). That is, $T_m(A_i{=}A_m, S_i{=}S_m)=\sum_{j=1}^{20} c_j^m P_{ij}+c_{21}^m\varphi_i^{NN}+c_{22}^m\psi_i^{NN}$ for a given residue *i* whose amino-acid type A_i is A_m and predicted secondary structure S_i is S_m. Here, the coefficients c_j^m are obtained by maximizing the number of residues with $|T_m(A_i{=}A_m, S_i{=}S_m)-\tau_i| < 36°$, where τ_i is the native value of either ϕ_i or ψ_i in training. Note that we optimized the coefficients c_j^m separately for ϕ and ψ.

CRF model—In the CRF model, the conditional probability of a finite set of Δτ states is defined by $P(\Delta\tau|X)=\exp\left[\sum_{i=1}^{N} F(\Delta\tau,X,i)\right]/Z(X)$ with the normalization factor $Z(X)=\sum_{l=1}^{L}\exp\left[\sum_{i=1}^{N} F(\Delta\tau,X,i)\right]$ for *N* sequentially-linked amino-acid residues and *L* Δτ states. Here, the function *F(Δτ, X, i)* is given by

$$F(\Delta\tau,X,i)=\sum_{j}\lambda_j t_j(\Delta\tau_{i-1},\Delta\tau_i,X_i)+\sum_{k}\mu_k s_k(\Delta\tau_i,X_{i+l(k)})w_i(X) \tag{1}$$

where t_j ($\Delta\tau_{i-1}$,$\Delta\tau_i$, X_i) is a transition feature function of the entire observation sequence and Δτ at positions *i* and *i − 1*, s_k ($\Delta\tau_i$, $X_{i+l(k)}$) is a state feature function of Δτ at position *i* and the observation sequence at position *i* + *l(k)*, $w_i(X)=T_m(A_i{=}A_m, S_i{=}S_m)-\tau_i^{NN}$ is the weight for the observation sequence obtained above, *l(k)* is an index for a sliding window around i (from i − 10 to i + 10), and λ_j and μ_k are parameters to be optimized based on training data. Here, the summation over j is over all 60 observations and 11 Δτ states with regions that are separated by the boundaries at ±2°, ±8°, ±18°, ±32°, ±50° and ±180°. $t_j(\Delta\tau_{i-1},\Delta\tau_i, X_i)$ is a delta function. For example, when j corresponds to $\Delta\tau_1$ at $i - 1$, $\Delta\tau_2$ at position i, and $X = [A_j][S_j]$ at position i, t_j $(\Delta\tau_{i-1},\Delta\tau_i, X_i)$ is nonzero, or 1, only if $\Delta\tau_{i-1} = \Delta\tau_1$, $\Delta\tau_i = \Delta\tau_2$, $A_i = A_j$ and $S_i = S_j$. The state feature function s_k $(\Delta\tau_i, X_{i+l(k)})$ includes not only amino-acid type and secondary structure at position i but also its neighboring residues (a window size of 21 residues is employed). That is, the summation over k is over 60 observations, 11 Δτ states, and a 21-residue window. It should also be noted that $w_i(X)$, depending on observations only, has only 60 values corresponding to 60 observations. When k corresponds to $\Delta\tau_2$ at position i and observation $X = [A_k][S_k]$ at position $i + 2$, s_k $(\Delta\tau_i, X_{i+l(k)})$ is also a delta function which is nonzero, or 1, only if $\Delta\tau_i = \Delta\tau_2$, $A_{i+2} = A_k$ and $S_{i+2} = S_k$.

Solving for λ_j and μ_k parameters—The CRF model trains its parameters by maximizing the conditional log-likelihood L of the data:

$$L(\lambda,\mu)=\ln P(\Delta\tau|X)=\sum_{i=1}^{N}F(\Delta\tau,X,i)-\ln Z(X)-\sum_{j}\lambda_j^2/2\sigma^2-\sum_{k}\mu_k^2/2\sigma^2 \tag{2}$$

where the last two terms are employed to regularize the variation of model parameters, to avoid over-fitting we set $\sigma^2 = 50$ after a few trials. This equation is solved by a slightly modified Powell method for function maximization (Press et al., 1992). The optimization function is convex and guarantees a global minimum.

Predicting Δτ—Once λ_j and μ_k parameters are known, one can efficiently calculate the expected Δτ value via defining “forward” and “backward” vectors and using a simple dynamic programming technique (Liu et al., 2006). The final predicted angles are calculated from τ^{NN}+ Δτ.

Training and Testing—We employed the same 10-fold cross validation technique to estimate the accuracy of the predictions as for ϕ^{NN} and ψ^{NN}.

C. Structure prediction with torsion angle restraints

Proteins are represented by a backbone-only model. In addition to flexible torsion angle restraints, we employed an energy function made of two terms: a statistical energy for C_α and C_β atoms based on the distance-scaled finite ideal gas reference state (DFIRE) (Zhou and Zhou, 2002; Yang and Zhou, 2008b) and a geometric-based hydrogen-bonding term between main-chain amine hydrogen and oxygen atoms. Relative weights of the energetic terms were determined by trial and error. A genetic algorithm with local optimization and clustering techniques (Yang and Zhou, 2008a) was used to search the global energy minimum with (1) the energy function only, (2) the energy function with restraints around ideal angles of predicted secondary structure (helices and strands) by an improved version of SPINE (Dor and Zhou, 2007a; Faraggi et al., 2009), (3) the energy function with 14 restraints around predicted real values of the torsion angles of predicted helical and strand residues, and (4) the energy function with predicted real-value angle restraints for all residues. Ideal angles of helical residues are −57° for ϕ and −45° for ψ and that of strand residues are −129° for ϕ and 124° for ψ (Park and Levitt, 1996). Note that protein structures are predicted without employing fragments or templates of known structures. This fragment free method is described in detail below and its flow chart is shown in Fig. 6.

Protein-specific torsion restraints (E_I^{φ} and E_I^{ψ})—We restrained the torsion angles according to the neural network predictions by adding two terms to the energy function corresponding to ϕ and ψ. These two terms are given by the following equation:

$$E_I^{\tau}=\begin{cases}0, & \Delta\tau\le\Delta\tau^0_{A_I,S_I},\\ K(\Delta\tau/\Delta\tau^0_{A_I,S_I}-1)^2, & \Delta\tau^0_{A_I,S_I}<\Delta\tau\le 2\Delta\tau^0_{A_I,S_I},\\ K, & \Delta\tau>2\Delta\tau^0_{A_I,S_I},\end{cases} \tag{3}$$

where τ can be either the ϕ or ψ angle, $\Delta\tau(=|\tau_I-\tau_I^{pred}|)$ is the absolute deviation of the τ_I angle for a given residue *I* from the corresponding predicted angle (τ_I^{pred}), $\Delta\tau^0_{A_I,S_I}$ is a pre-defined, allowed angle deviation that depends on the residue type (A_I) and predicted secondary structure (S_I) for residue *I*, and *K* is a weight parameter. Δτ is evaluated with the consideration of angle periodicity (e.g., the difference between −180° and 180° is 0°, not 360°). We set *K* to 100 after a few trials. Predefined 60 values of $\Delta\tau^0_{A_I,S_I}$ are obtained from a statistical analysis of the deviations between predicted and actual angles of 2640 protein chains (Zhou and Zhou, 2004). $\Delta\tau^0_{A_I,S_I}$ is chosen so that $\Delta\tau^0_{A_I,S_I}$ is greater than the deviations between predicted and actual angles for 70% of residues of type A_I and secondary structure S_I.

The DFIRE energy function ($E_{ij}^{DFIRE}(r)$)—A statistical energy function based on the distance-scaled finite ideal-gas reference state is used to calculate the interaction energy function between two atoms *i* and *j* (C_α and C_β atoms only) at a distance *r* apart (Zhou and Zhou, 2002; Zhou et al., 2006). A version with finer distance bins (DFIRE 2.0) is used (Yang and Zhou, 2008b).

Hydrogen bonding energy (*E*$^{h\text{-}bond}$)—We adopted a continuous version of an empirical hydrogenbonding energy (Gong et al., 2005; Kortemme et al., 2003) for backbone atoms as follows:

$$E^{h-bond}=w_{hb}\sum_{I,|J-I|>3} w_{O_IH_J}\,f_r(r_{O_IH_J})f_\theta(\theta_1,\theta_1^0)f_\theta(\theta_2,\theta_2^0) \tag{4}$$

with

$$f_r(r)=\begin{cases}0, & r\le 1.75A,\\ 1, & 1.75A<r<2.25A,\\ 4(r-2.75)^2, & 2.25\mathrm{A}<r\le 2.75A\\ 0, & r>2.75A\end{cases} \tag{5}$$

and

$$f_\theta(\theta,\theta^0)=\begin{cases}1, & \cos\theta\le\cos\theta^0-0.25,\\ 4(\cos\theta^0-\cos\theta), & \cos\theta^0-0.25<\cos\theta\le\cos\theta^0,\\ 0, & \cos\theta>\cos\theta^0,\end{cases} \tag{6}$$

where r is the distance between atoms H and O, θ_1 is the angle C-O-H, θ_2 is the angle O-H-N, $\theta_1^0=120°$, $\theta_2^0=126.9°$, $w_{hb} = 10$, and w_{OIHJ} is a proportionality constant to distinguish nonlocal hydrogen bonds ($|I - J| > 4$) from local hydrogen bonds ($|I - J| = 4$, separated by four sequential residues between a hydrogen-bond donor of residue J, H_J, and a hydrogen-bond acceptor of residue I, O_I, no hydrogen bonds were included for $|I - J| < 4$). We set w_{OIHJ}= 1.0 for local hydrogen bonds, 3.6 for nonlocal hydrogen bonds between predicted strand residues, 0.6 for nonlocal hydrogen bonds between predicted helical residues, and 1.2 for other nonlocal hydrogen bonds. These values were obtained by trial and error. Here, we have increased weights for strands because they are entropically more difficult to form. Secondary structures are predicted by an improved version of SPINE (Dor and Zhou, 2007a; Faraggi et al., 2009). The different geometric parameters for the angles C-O-H (θ_1^0) and O-H-N (θ_2^0) are based on statistical analysis of protein structures (Kortemme et al., 2003). We neglect the hydrogen-bond energy for $|I - J| \le 3$ because the DFIRE energy function captures local interactions adequately (Yang and Zhou, 2008a; Yang and Zhou, 2008b).

The energy function—The final energy function is the sum of the three terms described above. That is,

$$E=\sum_I (E_I^{\varphi}+E_I^{\psi})+\sum_{i>j} E_{ij}^{DFIRE}(r)+E^{h-bond}. \tag{7}$$

Initial setup—Proteins are represented by main-chain atoms (N, C_α, C, O, and H) and C_β atoms only and are described by internal coordinates: the bond lengths, covalent-bond angles, and backbone torsion angles of ϕ, ψ, and ω. The bond lengths and covalent angles are fixed to ideal values from the CHARMM force field (Brooks et al., 1983) except for the covalent angle between backbone N, C_α and C atoms. This angle is initially set to the ideal value of 111.6° but is allowed to change within 10° from the ideal value so that a local minimization technique can be employed (as described below). The planar torsion angle ω is fixed to 180° while the initial backbone ϕ/ψ angles are assigned in the following way. When predicted angle restraints are not employed ($E_I^{\varphi}=E_I^{\psi}=0$), ϕ/ψ angles are assigned randomly according to the observed residue-specific probability in the 2640 proteins (Zhou and Zhou, 2004). When predicted angle restraints are employed, backbone angles are randomly assigned with a random value within a range ($\Delta\varphi^0_{A_I,S_I}$ and $\Delta\psi^0_{A_I,S_I}$) around the predicted φ_I^{pred} or ψ_I^{pred} angles.

Local optimization—We used two types of moves for local energy optimization: pivot move and local "fixed-end" moves (Betancourt, 2005). In a pivot move, a random value between −10° and 10° is added to the ϕ and ψ angles for a randomly selected residue I. In "fixed-end" moves, two residues, I and J, are chosen randomly with their sequence positions 2 to 10 residues apart ($2 \le |I - J| \le 10$). The atoms in between the C_α atom of residue I and that of residue J are then rotated along the axis of the two C_α atoms by an angle randomly chosen with a constraint that the covalent angle between N, C_α and C atoms does not deviate more than 10° from its ideal value (111.6°) after the rotation. After each move, the Metropolis criterion is employed to accept or reject the move with a Boltzmann factor: $0.01\Delta E^{l-1}$, with ΔE^{l-1} the root-mean squared energy value of all conformations in the last generation, $l-1$, and the effective temperature is set to a low value of 0.01. Because a pivot move often leads to a large change in a protein's conformation and reduces the acceptance rate of moves, we gradually reduced the probability to make a "pivot move" from 1.0 to 0.0 in the first 30% of moves from a total of 30N moves of local optimization for a protein of chain length N.

Genetic Algorithm (GA)—The genetic algorithm described below is modified slightly from what was used for ab initio refolding of terminal fragments (Yang and Zhou, 2008a). We used N_p (=160) parents for the genetic algorithm. Initial N_p conformations are generated (initial setup) and locally optimized (local optimization). These optimized conformations serve as the first-generation parent conformations, and are used to generate another N_p conformations by crossover, mutations and local optimization. This leads to $2N_p$ conformations that are clustered as follows: The lowest energy conformation is chosen as the representative conformation of the first cluster which contains all the conformations with a drms of 5Å or less from the conformation. drms is the root-mean-squared difference between the pairwise distances of the two conformations. Then, the lowest energy conformation from the remaining conformations is chosen as the representative conformation of the second cluster. This process continues until all conformations are clustered. We employed a fitness function: $f_m^l = \exp(-E_m/\Delta E^{l-1})/(\rho_m o_m)$ where E_m is the energy of conformation m, ρ_m is the number of conformations in the cluster that conformation m belongs to, $o_m(=1+0.3n_m^g)$ mimics the age of the conformation m that is related to number of generations the conformation m existed, n_m^g. To ensure structural diversity, the first 30 conformations for the next generation are chosen from the first 30 clusters, respectively, according to the above fitness function. The remaining 130 conformations are chosen according to the fitness function from the remaining 290 conformations. A maximum of 400 generations is used in protein conformational search.

D. Effect of homologs

One question is if the existence of homologs between training and test proteins will increase the accuracy of predicted angles when we compared our angle-prediction results with TOPOS (11 protein set (Cavalli et al., 2007)) and TALOS (37 protein set) and employed them in structure prediction (16 protein set (Bradley et al., 2005)). To answer this question, we made a specific prediction server that was trained without sequence homologs to the 16 proteins used for structure prediction. We found that the average accuracy of 16 proteins changes only 0.6% ($Q_{60°}$ =88.8% with homologues versus 88.2% without). This highlights our effective use of overfitting protection in training. Here, we have performed structure prediction using a server trained without sequence homologues.

Acknowledgments

This work was supported by the NIH (R01 GM 966049 and R01 GM 068530). This research was also supported in part by the Project of Knowledge Innovation Program (PKIP) of Chinese Academy of Sciences, Grant No. KJCX2.YW.W10. We would like to thank Martin Karplus and Song Liu for their valuable comments and thank Frank Delaglio and Ad Bax for making the program NMRPipe (TALOS) available for download. Authors declare no competing interests.

References

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Aci Res 1997;25:3389–3402.

Baum LE, Petrie T. Statistical inference for probablistic functions of finite state Markov chains. Annals of Mathematical Statistics 1966;37:1554–1563.

Betancourt MR. Efficient Monte Carlo trial moves for polypeptide simulations. J Chem Phys 2005;123:174905. [PubMed: 16375567]

Boden M, Yuan Z, Bailey TL. Prediction of protein continuum secondary structure with probabilistic models based on NMR solved structures. BMC Bioinformatics 2006;7:68. [PubMed: 16478545]

Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T. A generative, probabilistic model of local protein structure. Proc Natl Acad Sci USA 2008;105:8932–8937. [PubMed: 18579771]

Bourhis JM, Canard B, Longhi S. Predicting protein disorder and induced folding: from theoretical principles to practical applications. Current Protein & Peptide Science 2007;8:135–149. [PubMed: 17430195]

Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871. [PubMed: 16166519]

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.

Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577. [PubMed: 9698570]

Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–190. [PubMed: 10926500]

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 2007;104:9615–9620. [PubMed: 17535901]

Cheng J, Sweredoski MJ, Baldi P. DOMpro: Protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. Data Mining and Knowledge Discovery 2006;13:1–10.

Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 1999;13:289–302. [PubMed: 10212987]

de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 2000;41:271–287. [PubMed: 11025540]

DeBartolo J, Colubri A, Jha AK, Fitzgerald JE, Freed KF, Sosnick TR. Mimicking the folding pathway to improve homology-free protein structure prediction. Proc Natl Acad Sci 2009;106:3734–3739. [PubMed: 19237560]

Dor O, Zhou Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins 2007a;66:838–845. [PubMed: 17177203]

Dor O, Zhou Y. Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins 2007b;68:76–81. [PubMed: 17397056]

Doreleijers JF, Mading S, Maziuk D, So journer K, Yin L, Zhu J, Markley JL, Ulrich EL. BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. J Biomol NMR 2003;26:139–146. [PubMed: 12766409]

Dosztanyi Z, Sandor M, Tompa P, Simon I. Prediction of protein disorder at the domain level. Current Protein & Peptide Science 2007;8:161–171. [PubMed: 17430197]

Dunbrack, R. A protein sequence culling server (cullpdb pc25 res3.0). 2006. http://dunbrack.fccc.edu/Guoli/piscesdownload.php

Eyrich VA, Standley DM, Friesner RA. Prediction of protein tertiary to low resolution: performance for a large and structurally diverse test set. J Mol Biol 1999;288:725–742. [PubMed: 10329175]

Fain B, Levitt M. Funnel sculpting for in silico assembly of secondary structure elements of proteins. Proc Natl Acad Sci USA 2003;100:10700–10705. [PubMed: 12925740]

Faraggi E, Xue B, Zhou Y. Improving the accuracy of predicting real-value backbone torsion angles and residue solvent accessibility by guided learning through two-layer neural networks. Proteins 2009;74:857–871. [PubMed: 18704938]

Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. Proteins 2006;65:1–14. [PubMed: 16856179]

Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955. [PubMed: 8732766]

Gewehr JE, Zimmer R. SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles. Bioinformatics 2006;22:181–187. [PubMed: 16267083]

Gibrat JF, Robson B, Garnier J. Influence of the local amino acid sequence upon the zones of the torsional angles phi and psi adopted by residues in proteins. Biochemistry 1991;30:1578–1586. [PubMed: 1993174]

Godzik A, Jambon M, Friedberg I. Computational protein function prediction: are we making progress? Cellular & Molecular Life Sciences 2007;64:2505–2511. [PubMed: 17611711]

Gong H, Fleming PJ, Rose GD. Building native protein conformation from highly approximate backbone torsion angles. Proc Natl Acad Sci USA 2005;102:16227–16232. [PubMed: 16251268]

Gong HP, Shen Y, Rose GD. Building native protein conformation from NMR backbone chemical shifts using monte carlo fragment assembly. Protein Sci 2007;16:1515–1521. [PubMed: 17656574]

Gross M. Proteins that convert from alpha helix to beta sheet: implications for folding and disease. Current Protein & Peptide Science 2000;1:339–347. [PubMed: 12369904]

Hardina C, Eastwood M, Luthey-Schulten Z, Wolynes PG. Associative memory hamiltonians for structure prediction without homology: alpha-helical proteins. Proc Natl Acad Sci USA 2000;97:14235–14240. [PubMed: 11114172]

Huang YM, Bystroff C. Improved pairwise alignments of proteins in the twilight zone using local structure predictions. Bioinformatics 2006;22:413–422. [PubMed: 16352653]

Kang HS, Kurochkina NA, Lee B. Estimation and use of protein backbone angle probabilities. J Mol Biol 1993;229:448–460. [PubMed: 8429556]

Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 2003;51:504–514. [PubMed: 12784210]

Kihara D. The effect of long-range interactions on the secondary structure formation of proteins. Protein Sciencd 2005;14:1955–1963.

Kortemme T, Morozov A, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Molec Biol 2003;326:1239–1259. [PubMed: 12589766]

Kuang R, Lesliei CS, Yang AS. Protein backbone angle prediction with machine learning approaches. Bioinformatics 2004;20:1612–1621. [PubMed: 14988121]

Kuznetsov IB. Ordered conformational change in the protein backbone: prediction of conformationally variable positions from sequence and low-resolution structural data. Proteins 2008;72:74–87. [PubMed: 18186479]

Lafferty, J.; Mccallum, A.; Pereira, F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proc. 18th International Conf. on Machine Learning; San Francisco, CA: Morgan Kaufmann; 2001. p. 282-289.

Liu Y, Carbonell J, Weigele P, Gopalakrishnan V. Protein fold recognition using segmentation conditional random fields (SCRFs). J Comp Bio 2006;13 (2):394–406.

Lobley A, Swindells MB, Orengo CA, Jones DT. Inferring function using patterns of native disorder in proteins. PLoS Comput Biol 2007;3:e162. [PubMed: 17722973]

Lopez G, Ro jas A, Tress M, Valencia A. Assessment of predictions submitted for the casp7 function prediction category. Proteins 2007;69(Supplement 8):165–174. [PubMed: 17654548]

Meiler J, Muller M, Zeidler A, Schmaschke F. Generation and evaluation of dimension reduced amino acid parameter representations by artificial neural networks. J Mol Model 2001;7:360–369.

Montalvao RW, Cavalli A, Salvatella X, Blundell TL, Vendruscolo M. Structure Determination of ProteinProtein Complexes Using NMR Chemical Shifts: Case of an Endonuclease ColicinImmunity Protein Complex. J Amer Chem Soc 2008;130:1599015996.

Nanias M, Chinchio M, Pillardy J, Ripoll D, Scheraga H. Packing helices in proteins by global optimization of a potential energy function. Proc Nat Acad Sci USA 2003;100:1706–1710. [PubMed: 12571353]

Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ. An automated classification of the structure of protein loops. J Mol Biol 1997;266:814–830. [PubMed: 9102471]

Ortiz AR, Kolinski A, Skolnick J. Native-like topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. Proc Natl Acad Sci USA 1998;95:1020–1025. [PubMed: 9448278]

Ozkan SB, Wu GA, Chodera JD, Dill KA. Protein folding by zipping and assembly. Proc Natl Acad Sci 2007;104:11987–11992. [PubMed: 17620603]

Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996;258:367–392. [PubMed: 8627632]

Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. Numerical Recipes in C. Vol. 2. Cambridge, UK: Cambridge University Press; 1992.

Radivo jac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. Biophysical J 2007;92:1439–1456.

Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem 1968;23:283–437. [PubMed: 4882249]

Reva B, Finkelstein AV, Skolnick J. What is the probability of a chance prediction of a protein structure with an RMSD of 6Å? Folding & Design 1998;3:141–147. [PubMed: 9565758]

Robustelli P, Cavalli A, Vendruscolo M. Determination of Protein Structures in the Solid State from NMR Chemical Shifts. STRUCTURE 2008;16 (12):1764–1769. [PubMed: 19081052]

Rooman MJ, Kocher JP, Wodak SJ. Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions. J Mol Biol 1991;221:961–979. [PubMed: 1942039]

Rost B. Review: protein secondary structure prediction continues to rise. J Structural Biology 2001;134:204–218.

Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back propagating errors. Nature 1986;323:533–536.

Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins 2005;61:115–126. [PubMed: 16080156]

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A. Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 2008;105:4685–4690. [PubMed: 18326625]

Shen Y, Vernon R, Baker D, Bax A. De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 2009;43:63–78. [PubMed: 19034676]

Siew N, Elofsson A, Rychlewski L, Fischer D. Maxsub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 2000;16:776–785. [PubMed: 11108700]

Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulate anealing and bayesian scoring functions. J Mol Biol 1997;268:209–225. [PubMed: 9149153]

Sippl MJ. On the problem of comparing protein structures development and applications of a new method for the assessment of structural similarities of polypeptide conformations. J Molec Biol 1982;156:359–388. [PubMed: 7086905]

Tress M, Cheng J, Baldi P, Joo K, Lee J, Seo JH, Lee J, Baker D, Chivian D, Kim D, Ezkurdia I. Assessment of predictions submitted for the casp7 domain prediction category. Proteins 2007;69 (Supplement 8):137–151. [PubMed: 17680686]

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G. CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 2008;36 (Suppl S):W496–W502. [PubMed: 18515350]

Wu S, Zhang Y. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 2008;72:547–556. [PubMed: 18247410]

Xue B, Dor O, Faraggi E, Zhou Y. Real-value prediction of backbone torsion angles. Proteins 2008;72:427–433. [PubMed: 18214956]

Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 2008a;72:793–803. [PubMed: 18260109]

Yang Y, Zhou Y. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely-related all-atom statistical energy functions. Protein Sci 2008b; 17:1212–1219. [PubMed: 18469178]

Yang ZR. Biological applications of support vector machines. Briefings in Bioinformatics 2004;5:328–338. [PubMed: 15606969]

Yoon S, Welsh WJ. Rapid assessment of contact-dependent secondary structure propensity: Relevance to amyloidogenic sequences. Proteins 2005;60:110–117. [PubMed: 15849755]

Young M, Kirshenbaum K, Dill KA, Highsmith S. Predicting conformational switches in proteins. Protein Science 1999;8:1752–1764. [PubMed: 10493576]

Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. Proteins 2005;58:905–912. [PubMed: 15645415]

Zhang W, Liu S, Zhou Y. SP5: improving protein fold recognition by using predicted torsion angles and profile-based gap penalty. PLoS ONE 2008;6:e2325. [PubMed: 18523556]

Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004a;57:702–710. [PubMed: 15476259]

Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004b;101:7594–7599. [PubMed: 15126668]

Zhao F, Li S, Sterner BW, Xu J. Discriminative learning for protein conformation sampling. Proteins 2008;73:228–240. [PubMed: 18412258]

Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure derived potentials of mean force for structure selection and stability prediction. Protein Science 2002;11:2714–2726. [PubMed: 12381853]

Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 2004;55:1005–1013. [PubMed: 15146497]

Zhou H, Zhou Y. SPEM: improving multiple-sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics 2005;21:3615–3621. [PubMed: 16020471]

Zhou Y, Zhou H, Zhang C, Liu S. What is a desirable statistical energy function for proteins and how can it be obtained? Cell Biochem Biophys 2006;46:165–174. [PubMed: 17012757]

Zimmermann O, Hansmann UHE. LOCUSTRA: Accurate prediction of local protein structure using a two-layer support vector machine approach. J Chem Info Modeling 2008;48:1903–1908.

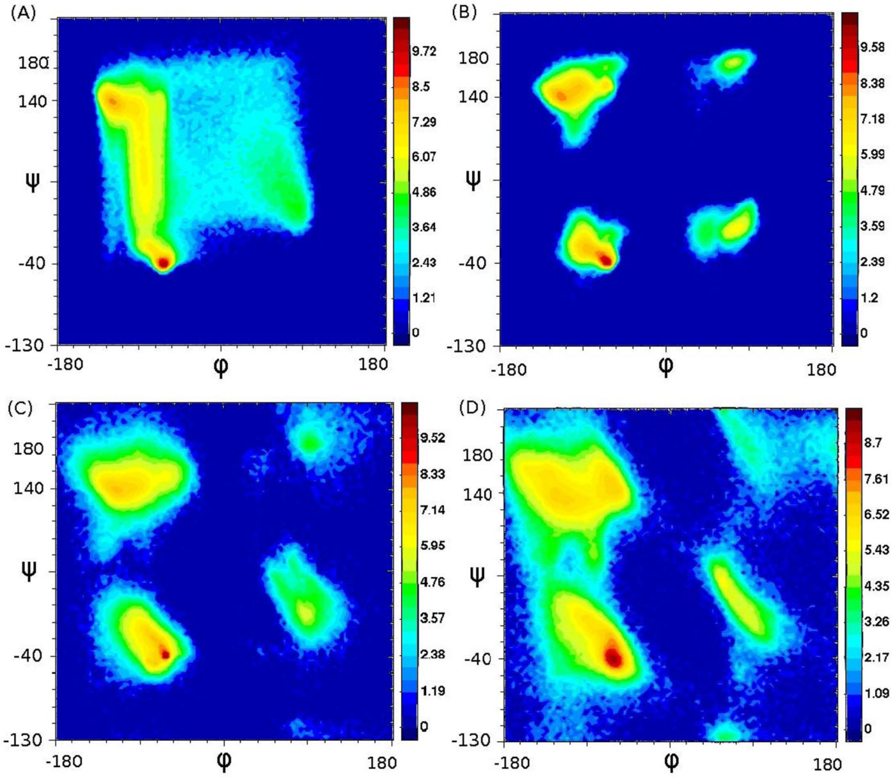

Figure 1.

The Ramachandran plot of 10-fold cross-validated predictions of 2640 proteins by a previous real-value prediction method called Real-SPINE 3 (A), by combining a two-state classifier with real value prediction, SPINE X (B), post-refinement, SPINE XI (C), and native dihedral angles (D). The ψ axis is shifted for a clearer view of the population separation. No shift was done on the ϕ angles.

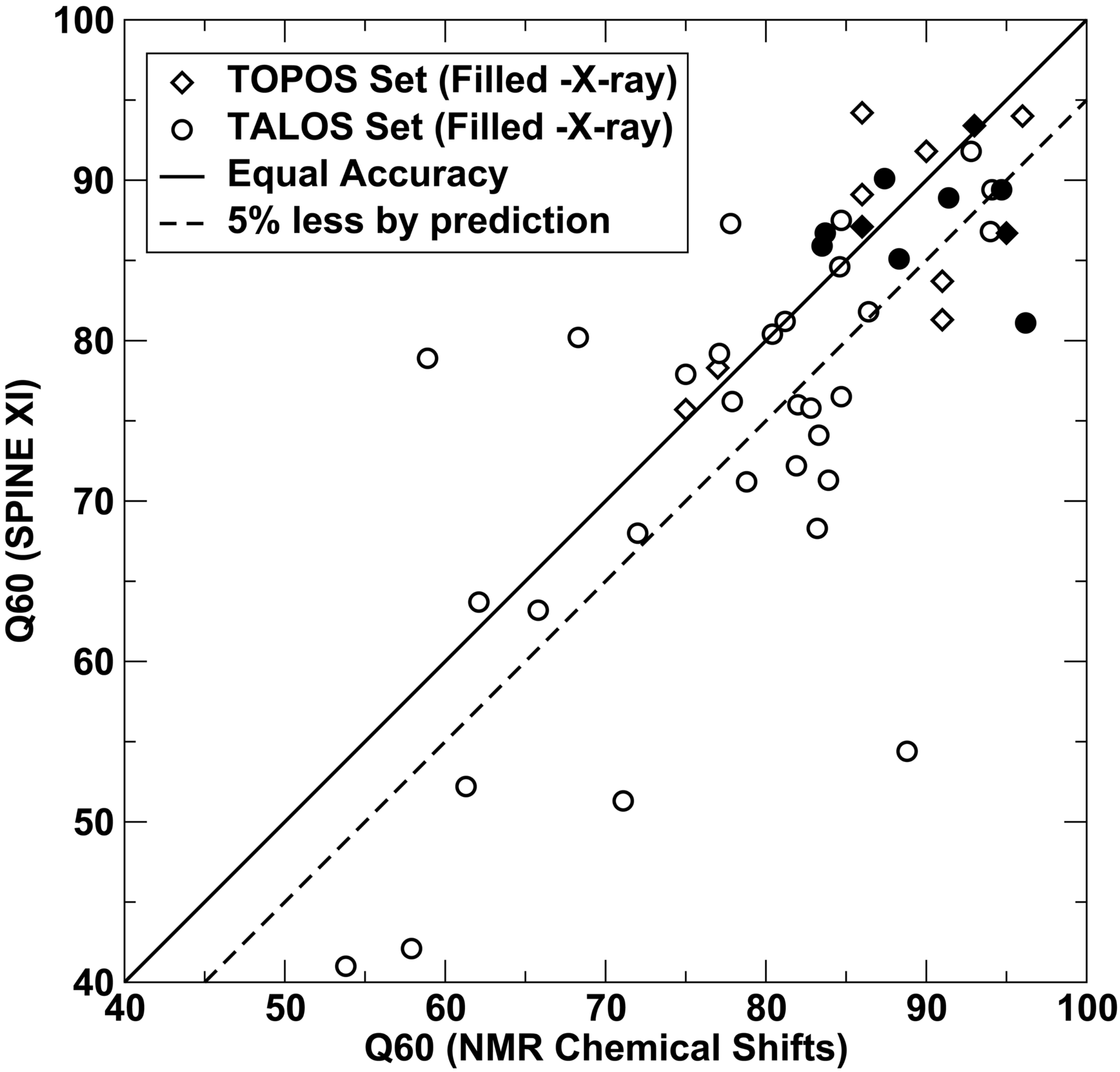

Figure 2.

The accuracy of torsion angles obtained from NMR chemical shifts by programs TOPOS (11 proteins) and TALOS (37 proteins) is compared to that given by SPINEXI. Accuracy is measured by $Q_{60°}$, the fraction of residues for which both ϕ and ψ angles are less than 60° away from their respective native values. SPINE XI makes equally or more accurate prediction for 20 proteins (42%). NMR-derived angles are only slightly more accurate than SPINE XI for an additional 10 proteins (less than 5% difference) (dashed line).

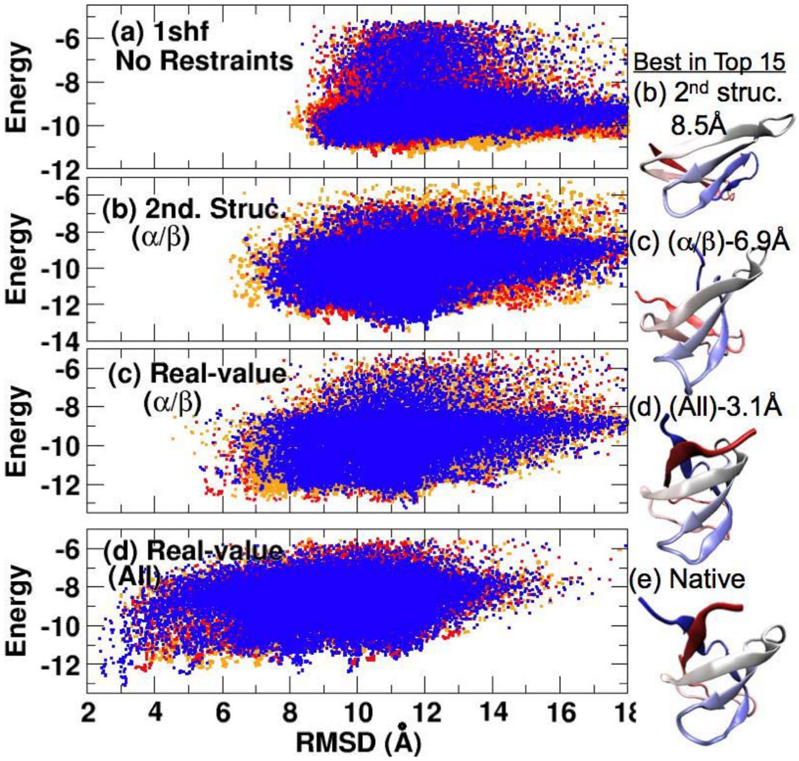

Figure 3.

The effects of various restraints on conformational sampling of 1shf. The energy of all sampled conformations as a function of their RMSD values from the native structure. From top to bottom, conformations are sampled without any restraints (a), with predicted secondary-structure restraints (i.e. restraining around the ideal angles of predicted helical and/or strand residues) (b), with restraints around predicted real-value torsion angles of predicted helical and/or strand residues only (c), and with restraints around predicted torsion angles for all residues (d). The corresponding best structures in top 15 for (b) to (d) are shown in the right panel and the native structure is shown in (e).

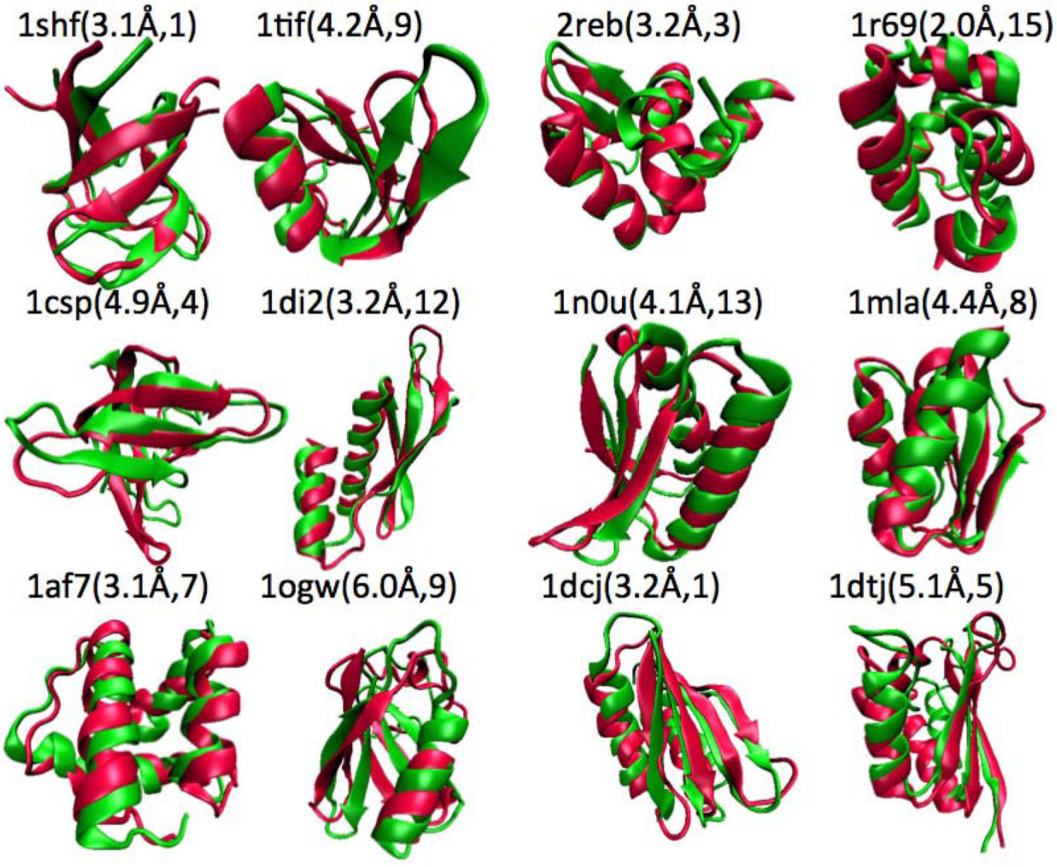

Figure 4.

Comparison between predicted structure (best in top 15, in Green) and native structure (in Red) for 12 proteins as labeled, along with their ranks.

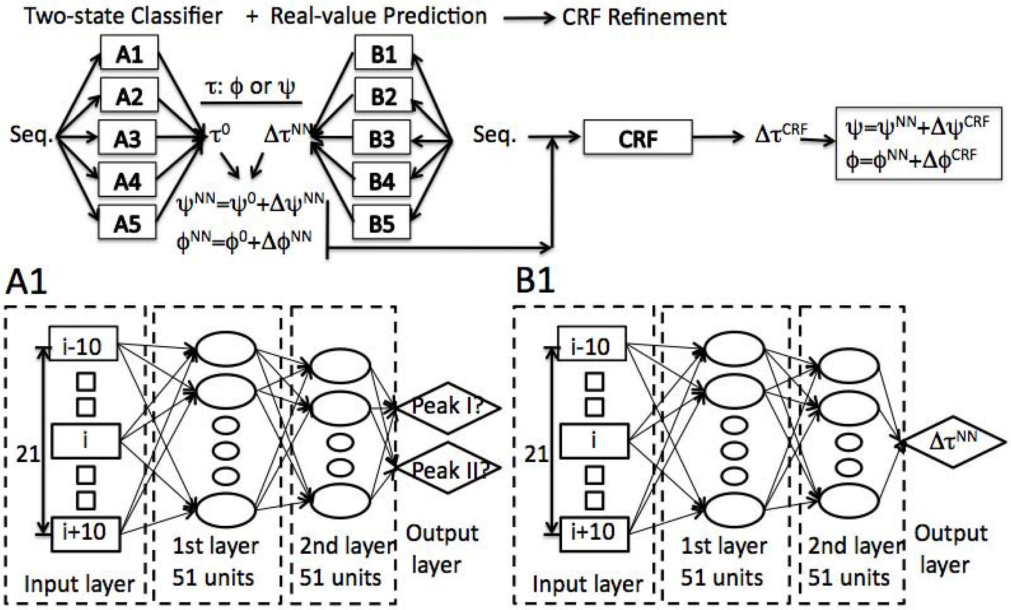

Figure 5.

Network architecture for real-value prediction of backbone torsion angles (SPINE XI). Two sets of neural networks are constructed for predicting angle peaks (Peak I or Peak II) by a two-state classifier (A1–A5) and the deviation from the peak by a real-value predictor (B1–B5). Both predictions result from a consensus of five independent predictions. Each predictor (A1,..A5,B1,…B5) has two hidden layers. Angles predicted by neural networks are further refined by the conditional random field (CRF) model.

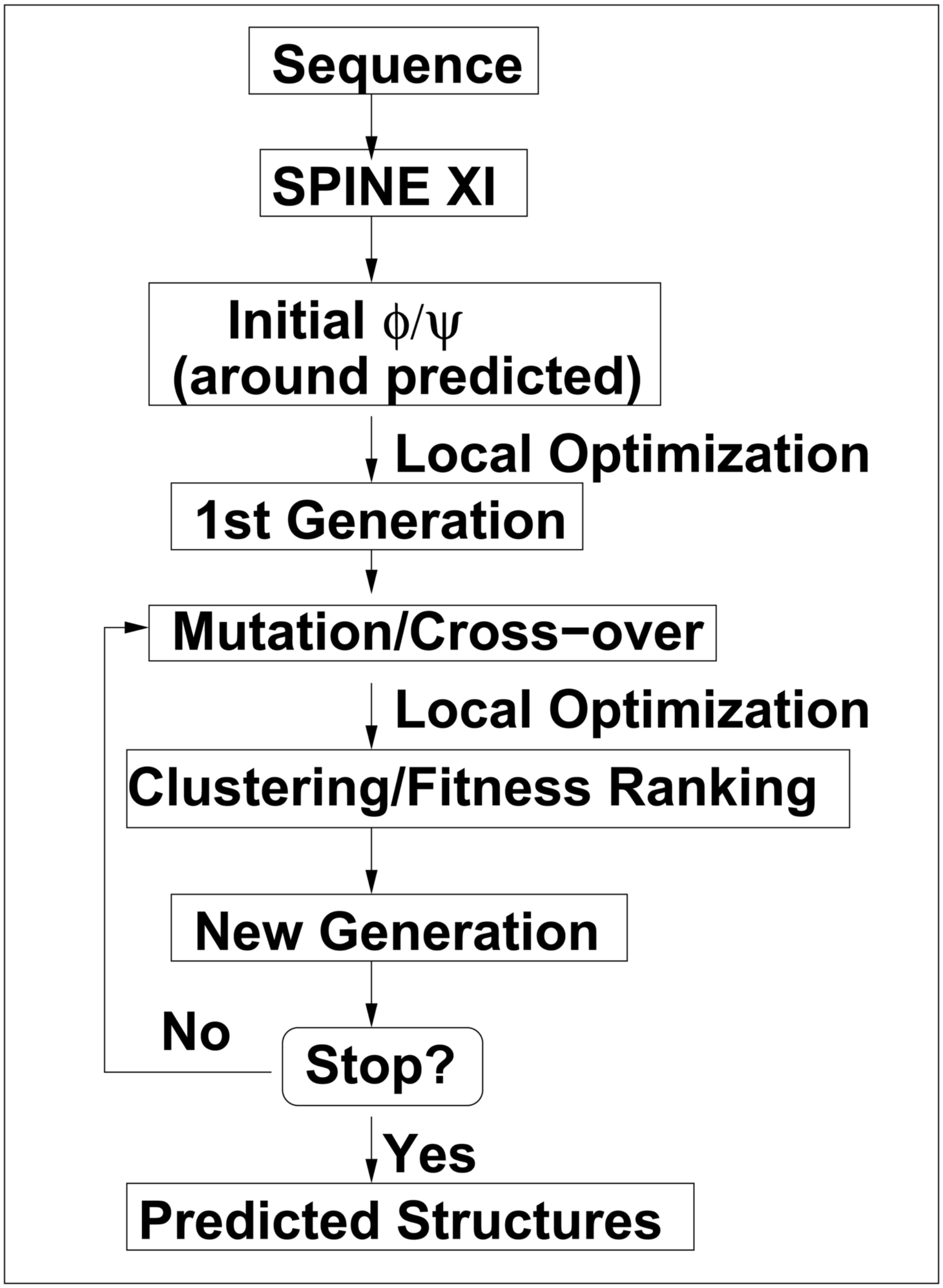

Figure 6.

The flow chart for the method used for fragment-free protein-tertiary-structure prediction. Details for the first part of the method, obtaining SPINE XI, are given in Fig. 5.

Table 1

The accuracy measured by Q_{60° for torsion angles obtained by TOPOS based on NMR chemical shifts (Cavalli et al., 2007) and predicted by SPINE XI for 11 proteins.

Protein	PDB ID[a]	L[b]	%α/β/coil[c]	TOPOS(%)[d]	SPINE XI(%)
Bet v 4	1h4b	84	64/4/32	96	94.0
Calbindin	3icb(X)	74	60/0/40	95	86.7
FF domain	1uzc	54	77/0/23	86	94.2
HPr	1poh(X)	85	37/29/34	86	87.1
Sda	1pv0	46	60/0/40	86	89.1
Δ27-GG	1sa8	106	0/65/35	77	78.3
TM1442	1sbo	110	44/20/36	90	91.8
Ubiquitin	1ubq(X)	76	25/32/43	93	93.4
MrR5	1yvc	70	0/51/49	75	75.7
PhS018	2glw	92	21/50/29	91	83.7
Sen15	2gw6	123	32/29/39	91	81.3
Ave.			38/26/36	88	86.9

[a] Protein Data Bank Identification Number. The method used to solve the structure is NMR unless it is specifically indicated as (X) for X-ray.

[b] Number of residues.

[c] Fraction of helical, strands, and coil residues (Cavalli et al., 2007).

[d] From Ref. (Cavalli et al., 2007).

Table 2

The accuracy measured by $Q_{60°}$ for torsion angles obtained by TALOS based on NMR chemical shifts (Cornilescu et al., 1999) and predicted by SPINE XI for 37 proteins.

#bmr[a]	PDB[b]	L[c]	%α/β/coil[d]	TALOS(%)[e]	SPINE XI(%)
10118	2exd	72	0/40/60	81.9	72.2
11022	2rng	76	28/15/57	71.1	51.3
11036	2ro1	186	36/2/62	61.3	52.2
11055	1jyr(X)	94	17/25/58	88.3	85.1
15125	2jns	85	46/0/54	94.1	89.4
15177	1r1v(X)	94	65/10/25	94.7	89.4
15247	2v9h	151	32/35/33	94.0	86.8
15283	1igd(X)	53	25/42/33	96.2	81.1
15393	2jss	190	57/0/43	62.1	63.7
15439	2juc	52	64/0/36	78.8	71.2
15444	2ckx(X	81	70/0/30	91.4	88.9
15469	2jzd	125	44/15/41	88.8	54.4
15528	2jwn	122	23/26/51	77.9	76.2
15534	2juj	54	61/0/39	83.3	74.1
15579	2jy0	25	56/0/44	72.0	68.0
15622	2job	90	43/30/27	58.9	78.9
15677	2k1e	101	68/0/32	68.3	80.2
15953	3e2b(X)	85	34/43/23	83.5	85.9
287	1ner	63	53/0/47	77.8	87.3
4090	2ezi	72	61/0/39	84.7	87.5
4224	2myo	114	47/0/54	65.8	63.2
4333	2gf5	97	74/0/26	92.8	91.8
4540	1nla	50	62/0/38	82.0	76.0
5141	1t4z	87	35/23/42	83.9	71.3
5211	1d4t(X)	111	17/29/54	87.4	90.1
5596	1n91	101	22/30/48	83.2	68.3
5824	1v2z(X)	98	81/0/19	83.7	86.7
6085	1s6w	19	0/29/71	57.9	42.1
6127	1se7	80	58/0/42	81.2	81.2
6128	1se9	99	15/29/56	82.8	75.8
6147	1y6d	112	79/0/21	80.4	80.4
6324	1x9a	85	39/17/44	84.7	76.5
6392	1o7c	96	18/24/58	77.1	79.2
6432	1wvk	78	17/7/76	53.8	41.0
6553	1yza	104	63/0/37	75.0	77.9
6560	1yyx	104	76/0/24	84.6	84.6
6571	2ae9	66	65/0/35	86.4	81.8
Ave (X ray)			44/21/35	89.3	86.7
Ave (All)			45/13/43	79.8	75.5

[a] bmr identification number.

[b] Protein Data Bank Identification Number. The method used to solve the structure is NMR unless it is specifically indicated as (X) for X-ray.

[c] Number of residues predicted by TALOS.

[d] Fraction of helical, strands, and coil residues.

[e] TALOS program was downloaded from http://spin.niddk.nih.gov/NMRPipe/talos/.

Table 3

The accuracy ($Q_{36°}$) of SPINE X given by different combinations of inputs (Sequence profiles, representative residue properties, and predicted structural properties including secondary structure and solvent accessible surface area) employed in neural networks.[a]

Seq. Profile	**Properties**	**Predicted SP**	**ϕ**	**ψ**
X			83%	76%
	X		79%	66%
		X	84%	76%
X	X		84%	78%
X		X	84.5%	78.5%
	X	X	84.5%	78%
X	X	X	85%	79%

[a] This result is from testing on one fold.

Table 4

The root-mean-squared distance (RMSD) from the native structure of the best structure in top five clusters of three independent runs (top 15) by an energy function only, with restraints around ideal angles of predicted helical and strand residues (secondary structure restraints), around predicted real-value angles of predicted helical and strand residues, and around predicted real-value angles of all residues.

					Best RMSD in Top 15 in Å			
PDB ID[a]	L[b]	α/β/coil (%)[c]	$Q_{60°}$ (%)[d]	Q_3 (%)[e]	No Res.[f]	Ideal α&β[g]	Real α&β[h]	All[i]
1b72	49	71/0/29	66.6	67.3	4.9	5.3	5.1	6.5
1shf	59	0/42/58	86.0	79.7	9.1	8.5	6.9	3.1
1tif	59	24/37/39	89.5	84.7	8.5	6.2	4.5	4.2 (3.4[j])
2reb	60	62/22/16	91.4	85.0	6.6	3.2	5.6	3.2
1r69	61	71/0/29	93.2	96.7	6.7	9.0	6.2	2.0
1csp	67	0/55/45	86.2	77.6	10.4	7.3	6.1	4.9
1di2	69	42/33/25	92.5	95.7	11.6	3.8	6.5	3.2
1n0u	69	45/28/27	98.5	87.0	9.4	4.1	4.2	4.1
1mla	70	37/37/26	91.2	90.0	8.0	3.0	3.3	4.4
1af7	72	71/0/29	95.7	97.2	5.8	6.1	4.9	3.1
1ogw	72	26/32/42	90.0	79.2	8.0	6.8	6.5	6.0
1dcj	73	36/32/32	90.1	91.8	8.9	4.2	4.1	3.2
1dtj	74	41/27/68	83.3	83.8	10.2	6.3	5.2	5.1
1o2f	77	48/29/23	84.0	81.8	9.5	6.6	5.7	9.7
1mky	81	35/25/40	91.1	86.4	11.9	7.0	4.5	8.0 (5.2[j])
1tig	88	39/35/26	82.6	75.0	10.7	11.7	10.5	10.6
Median:		40/27/33[k]	88.2[k]	84.9[k]	9.0	6.3	5.4	4.3
Success Rate (RMSD<6Å)[l]:					2/16	6/16	10/16	12/16

[a] Protein Data Bank Identification Number (a dataset of Ref. (Bradley et al., 2005)).

[b] Chain length (Bradley et al., 2005).

[c] Fractions of native helical, strand and coil residues.

[d] Fraction of residues for which predicted angles are within 60° from their native values for both ϕ and ψ angles.

[e] The accuracy of secondary structure prediction by an improved version of SPINE (Dor and Zhou, 2007a; Faraggi et al., 2009). The best structure (in RMSD) in top five clusters of three independent runs (top 15)

[f] without torsion angle restraints (DFIRE plus hydrogen bonding only).

[g] The energy function plus restraints around ideal angles of predicted helices and strands.

[h] The energy function plus restraints on predicted real values of torsion angles of predicted helical and strand residues only.

[i] The energy function plus restraints on all residues.

[j] A native cis-conformation is assigned to Pro 54 in 1tif and Pro 394 ($\omega = 0°$) in 1mky.

[k] Average.

[l] The success rate: the number of proteins having a correct structural topology (RMSD<6Å) ranked within top 15.