WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

The relationship between N-gram patterns and protein secondary structure

John K. Vries,[1]* Xiong Liu,[1,2] and Ivet Bahar[1]

[1] Department of Computational Biology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15213

[2] Department of Information Science and Telecommunications, School of Information Sciences, University of Pittsburgh, Pennsylvania 15213

ABSTRACT

An n-gram pattern (NP{n,m}) in a protein sequence is a set of n residues and m wildcards in a window of size n+m. Each window of n+m amino acids is associated with a collection of NP{n,m} patterns based on the combinatorics of n+m objects taken m at a time. NP{n,m} patterns that are shared between sequences reflect evolutionary relationships. Recently the authors developed an alignment-independent protein classification algorithm based on shared NP{4,2} patterns that compared favorably to PSI-BLAST. Theoretically, NP{4,2} patterns should also reflect secondary structure propensity since they contain all possible n-grams for $1 \leq n \leq 4$ and a window of 6 residues is wide enough to capture periodicities in the $2 \leq n \leq 5$ range. This sparked interest in differentiating the information content in NP{4,2} patterns related to evolution from the content related to local propensity. The probability of α-, β-, and coil components was determined for every NP{4,2} pattern over all the chains in the Protein Data Bank (PDB). An algorithm exclusively based on the Z-values of these distributions was developed, which accurately predicted 71–76% of α-helical segments and 62–67% of β-sheets in rigorous jackknife tests. This provided evidence for the strong correlation between NP{4,2} patterns and secondary structure. By grouping PDB chains into subsets with increasing levels of sequence identity, it was also possible to separate the evolutionary and local propensity contributions to the classification process. The results showed that information derived from evolutionary relationships was more important for β-sheet prediction than α-helix prediction.

Proteins 2007; 68:830–838.

Key words: n-gram patterns; protein sequence analysis; secondary structure; structural bioinformatics.

INTRODUCTION

The preference or local propensity of individual amino acids for specific secondary structure elements has been known for decades.[1,2] Early secondary structure prediction algorithms based on these preferences achieved prediction accuracies in the 60–65% range.[3,4] As more sequences were deposited in the Protein Data Bank (PDB),[5] it was possible to demonstrate that selected *n*-grams (contiguous runs of *n* amino acids) correlated better with secondary structure than their amino acid components.[6,7] A second generation of secondary structure prediction algorithms was developed that used the statistical properties of *n*-grams in windows of fixed length to train machine learning algorithms such as neural nets (NN), hidden Markov models (HMM), and Support Vector Machines (SVM). Accuracy levels in the 70% range were achieved with these approaches.[8] Early studies also showed that information derived from multiple alignment studies (MSAs) had a significant correlation with secondary structure.[9] Recently, evolutionary information derived from MSAs has been combined with local propensity information to achieve accuracies approaching the 80% range.[10–12] The currently available state-of-the-art services in the public domain include YASPIN,[13] PHDpsi,[14] SSPro2,[15] PSIPRED,[16] SAM-T99,[17] and SAM-T04.[18] Additional approaches include efforts to combine different prediction servers into meta-servers,[19] efforts to relate short sequence segments to secondary structures and motifs in the PDB,[20–22] and efforts to relate secondary structures to other features such as kinetics of protein folding.[23]

N-gram patterns provide another useful means for classifying and characterizing proteins. An *n*-gram pattern (NP{*n*,*m*}) in a protein sequence is a set of *n* residues and *m* wildcards in a window of size *n*+*m*. Each window of *n*+*m* residues in a sequence is associated with a collection of NP{*n*,*m*} patterns based on the combinatorics of *n*+*m* objects taken *m* at a time. Any protein sequence can be parsed into a series of overlapping *n*-gram patterns by advancing a window of size *n*+*m* along the sequence. NP{*n*,*m*} patterns that are shared between

Grant sponsor: NIH; Grant number: 5R01LM007994-02

*Correspondence to: John K. Vries, Department of Computational Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213. E-mail: vries@ccbb.pitt.edu

Received 21 September 2006; Revised 13 December 2006; Accepted 16 February 2007
Published online 21 May 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21480

sequences reflect evolutionary relationships. Recently the authors developed an alignment-independent protein classification algorithm based on shared NP{4,2} patterns that compared favorably to PSI-BLAST.[24] Theoretically, NP{4,2} patterns should also reflect secondary structure propensity since they contain all possible n-grams for $1 \leq n \leq 4$ and a window of 6 residues is wide enough to capture periodicities in the $2 \leq n \leq 5$ range.[25] Since NP{4,2} patterns may derive from both evolutionary constraints and local energy-driven propensities, it is of interest to quantifying the contribution from either effect. NP{4,2} patterns also have other interesting characteristics including: (1) the existence of all theoretically possible NP{4,2} patterns in nature; (2) the low probability of finding redundant NP{4,2} patterns in the same sequence; (3) the high probability of family membership if two nonoverlapping NP{4,2} patterns are shared between sequences; and (4) an implied substitution matrix for matches between sequences based on the variable position of the wildcards in NP{4,2} patterns.[24] NP{4,2} patterns capturing evolutionary as well as local propensity information might be useful as additional inputs to machine learning algorithms for predicting secondary structure if a strong correlation between these patterns and secondary structure could be established.

In the present study, we quantify the statistical relationship between NP{4,2} patterns and secondary structure. We analyze the distribution of NP{4,2} patterns and introduce a new algorithm based on their distribution for correlating these patterns with secondary structure. The methodology applied to a representative set of PDB structures reveals a strong correlation that is based on a combination of evolutionary and local propensity information, the former effect being more pronounced in the case of β-strands.

METHODS

Database selection and secondary structure determination

The studies reported in this paper were based on the 3D structures extracted from the Protein Data Bank (PDB) (http://www.rcsb.org).[5] A total of 32,434 files were downloaded on 05-Sep-2005. A total of 58,831 chains were selected from these files after applying the following criteria: (1) crystallographic data only; (2) chain length between 75 and 1500 residues; (3) no nucleotide chains; and (4) no chains with more than 10 missing side chains. The secondary structure was determined for each selected chain using the STRIDE program.[26] Secondary structure assignment by STRIDE is based on hydrogen bond formation patterns derived from the distances and angles between hydrogen bond forming pairs. This is similar to the approach employed in the popular DSSP program.[27] In STRIDE, additional constraints related to backbone dihedral angles also contribute to the determination. A recent comparison of DSSP with STRIDE showed no significant differences in prediction with respect to the principal secondary structure components.[28] The structural categories generated by STRIDE include the 3_{10}-helix (G), the α-helix (H), the π-helix (I), extended β-strands (E), isolated β-bridges (B), hydrogen-bonded turns (T), and nonhydrogen bonded turns (S). For the analysis in this paper, these seven states were consolidated to A (helix) = {G, H, I}, B (sheet) = {B, E}, and C (coil) = {T, S}.

Clustering of chains based on sequence identity level

Studies involving the statistical properties of sequences contained in the PDB must take into account the skewed distribution of the sequences in this database. Membrane proteins are underrepresented because they are difficult to crystallize and they are not soluble in water.[29] Site mutagenesis studies have resulted in large clusters of near identities for selected target proteins.[30] Finally, the attention of the scientific community has often been focused on small subsets of proteins with known roles in physiology or disease.[31] To quantify database skew, the 58,831 chains in the PDB database were assigned to 100 identity cutoff (IC) levels based on the maximum allowable similarity between sequences at the same level (i.e. IC90 would contain no sequences with similarity greater than 90%). Similarity was determined by measuring the percentage of unique 4-grams in a query sequence that were also present in a target sequence of equal or greater length. For this purpose, a 4-gram was defined as any instance of 4 consecutive residues in a sequence. Unique 4-grams represent the collection of all possible 4-grams in a sequence after purging for redundancy. The specific algorithm involved the following steps: (1) sort the 58,831 chains in the PDB into ascending order based on chain length; (2) select query sequences starting at the top of the list and moving toward the bottom; (3) for each query sequence determine the percentage of unique 4-grams contained in each sequence below its level; (4) populate a histogram whose bins are levels of identity from 0.00 to 0.99; and (5) store the chains in separate databases based on 100 IC levels. The histogram of the IC levels is shown in Figure 1(a). The sharp drop seen below the IC10 level is related to exclusion of the 4-gram patterns that occur in sequence pairs by random chance. The steep rise beyond the IC95 level is related to overrepresentation and the near-identities stemming from site mutagenesis studies.

The features in the histogram are accentuated when the relative entropy of the 4-gram distributions is measured over this interval. Relative entropy is defined as $S(P\|Q) = \sum_{i} \log(p_i/q_i)$, where p_i is the probability of

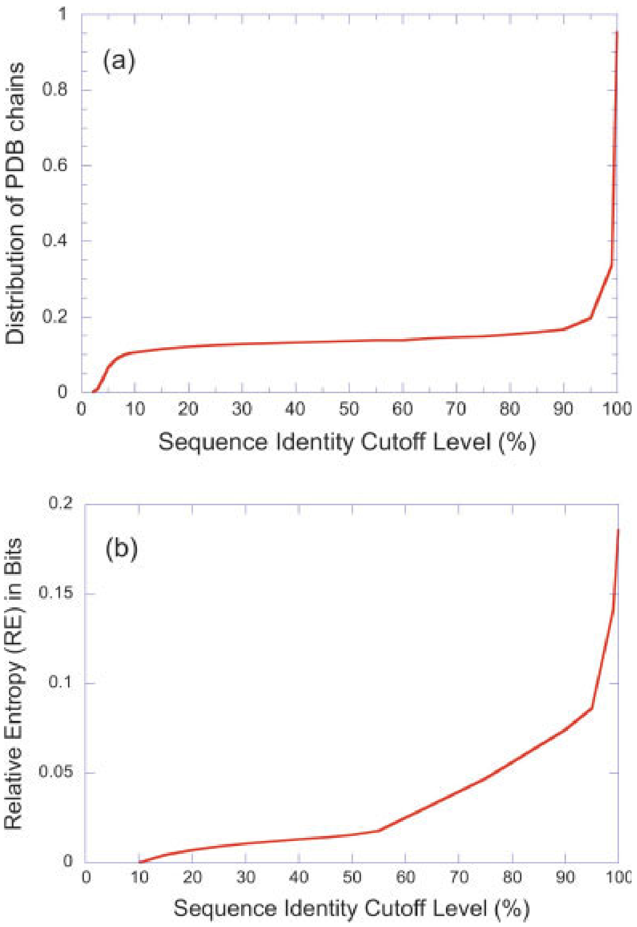
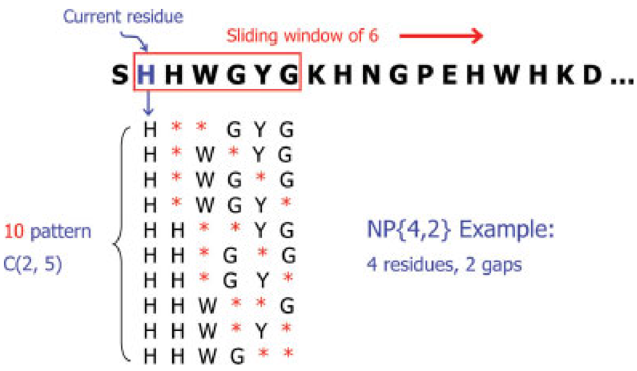


Figure 1

*(**a**) The histogram of 58,831 PDB chains as a function of IC level. The sharp drop below the IC10 identity is related to exclusion of the 4-gram patterns that occur in pairs of sequences by random chance. The steep rise beyond IC95 is related to the overrepresentation of certain proteins and the near identities stemming from site mutagenesis studies. The gradual rise between IC10 and IC95 represents the smooth accumulation of homologous protein family members. (**b**) The relative entropy of the 4-gram distribution as a function of identity cutoff. There is little effect from skewing between IC10 and IC55. The break at ~IC55 appears to be related to overrepresentation of members of immunoglobulin fold. After IC95 database distortion from overrepresentation is marked.*



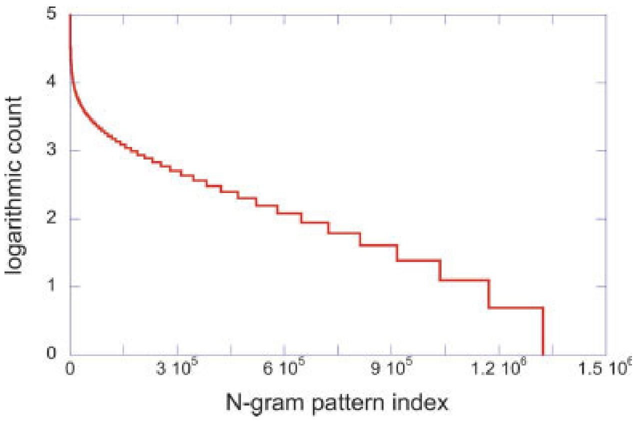


Figure 3

Histogram of NP{4,2} patterns for a typical 90–10 training set from the IC95 level. There are a small number of patterns toward index 0 that are overrepresented. The majority of patterns, however, have a relatively flat distribution over the range of the index.

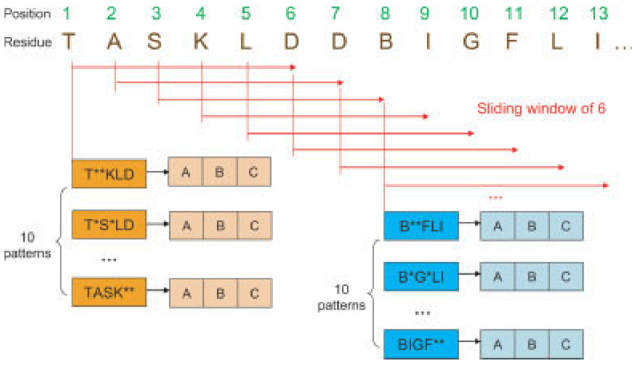


Figure 4

Creation of classification tables from training sets. A window of width 6 is advanced one residue at a time from left to right. At each position the 10 patterns associated with that window are generated. The average secondary structure content for each type (A–C) is recorded for each pattern. These percentages are averaged over all sequences in the training set. The data triplet in the final array represents the Z-value for each structure type (A–C) with respect to its NP{4,2} pattern.

the ith member in distribution P and q_i is the probability of the ith member in distribution Q[32]. Family members are characterized by different 4-gram distributions.[24] If family members were added evenly over the identity range from IC10 to IC100, the relative entropy would show a gradual rise reflecting the relative distribution of family sizes. It can be seen from Figure 1(b) that there is a

Figure 2

An example of extracting NP{4,2} patterns. At each position (e.g., H), there are 10 NP{4,2} patterns.

break at ~IC55 and another break at IC95. Direct inspection of the family membership at IC55 showed that this break was strongly associated with overrepresentation of the 4-grams associated with immunoglobulin folds. The break at IC95 picked up the overrepresentation from selected study targets and site mutagenesis studies. The curves in Figure 1 suggest that studies carried out between IC10 and IC55 would not be affected by skewed distribution and that studies carried out below the IC95 level would have only modest effects.

Distribution of N-gram patterns

The set of 58,831 chains from the PDB were divided into eight subsets consisting of IC = 10, 25, 50, 75, 90, 95, 99, and 100. Sample size in these levels ranged from 6600 chains in IC10 to 58,831 chains in IC100. A total of 160 training and test sets were generated from these eight IC levels. Half of the sets reflected a 90–10% split between training and test sets. The remainder reflected a 50–50% split. At each IC level and split, training and test set membership were determined from a random number generator.

n-Gram patterns (NP{*n*,*m*}) are sets of *n* residues and *m* wildcards in windows of size *n*+*m*. In this study NP{4,2} patterns were chosen for study based on optimization studies conducted in pervious research.[24] The constraint was also added that NP{4,2} patterns must start with a residue. Figure 2 shows an example of NP{4,2} patterns containing 4 residues and 2 wildcards in a 6 residue window. For each residue position, there are 10 patterns based on combinatorics. There are 1.6 million ($20^4 \times 10$) theoretically possible patterns considering all combinations of 20 amino acids at occupied sites. The number of patterns actually observed over the 160 training sets ranged from a high of 1.52 million to a low of 1.48 million. Therefore, in all training sets, most of the theoretically possible patterns were observed. A typical training set histogram (the distribution of particular NP{4,2} patterns) from the IC95 level is shown in Figure 3. The patterns have been sorted in descending order by count. It can be seen that the majority of NP{4,2} patterns have a smooth distribution over a broad index range. The overrepresentation on the left side of the graph affects less than 3% of the NP{4,2} patterns.

Construction of secondary structure classification tables

Secondary structures were grouped into three classes (A, B, and C) as outlined above. The relationship between NP{4,2} patterns and secondary structure was determined for each of the 160 training sets using the following steps: (1) pass a window of size 6 over each chain advancing it one residue at a time until the last position that will accommodate the full window is reached; (2) generate the 10 possible NP{4,2} patterns at each position (the combinatorics of 4 residues taken 2 at a time when the first position is always occupied by a residue); (3) for each NP{4,2} pattern add up the number of A's, B's and C's in the 6 positions in the window expressing the result as a percentage of each secondary structure type; (4) when all chains have been processed, determine the mean and standard deviation for each category and express the result as a *Z* score using the formula $Z = (X_i - X)/S$ where X_i is the observed value, X is the mean, and S is the standard deviation; (5) for each of the 160 training sets create a hashtable with NP{4,2} patterns as keys and arrays containing the *Z* scores for A, B, and C as values. Figure 4 depicts the process for creating classification tables for the 160 training sets.

Jackknife testing algorithm

The *Z*-scores for the A, B, and C levels in the hashtables generated for each of the 160 training sets were used to predict the secondary structure in the corresponding test sets. The members of the training and test sets were mutually exclusive. Because the training and test sets spanned IC levels from 10 to 100, the effects of database skew, secondary structure propensity and evolutionary content on accuracy were also measured. The following steps were employed in jackknife testing: (1) slide a window of size 6 over each test sequence; (2) look up each NP{4,2} pattern in the corresponding classification table and retrieve the *Z* values for its A, B, and C components; (3) average all 10 patterns; (4) average over all six positions as the window advances down the sequence (adjust appropriately for end effects); (5) for each completed sequence assign the secondary structure type associated with the greatest positive *Z*-value; (6) record the actual secondary structure type determined by STRIDE; and (7) determine the accuracy as the percentage of correct predictions for each sequence. Figure 5 depicts the jackknife protocol for obtaining the *Z*-values used in prediction.

For each jackknife run, the secondary structure prediction, the true secondary structure, the most positive *Z*-value and the difference between this *Z*-value and its nearest neighbor (ΔZ) were recorded for each position. This information was used to analyze the prediction accuracy separately for α, β, and coil regions and to generate confidence levels for each predicted position.

RESULTS

Prediction accuracy

The secondary structure prediction profiles for two example chains are illustrated in Figure 6. The first case was drawn from the IC50 level with a 50–50% training/

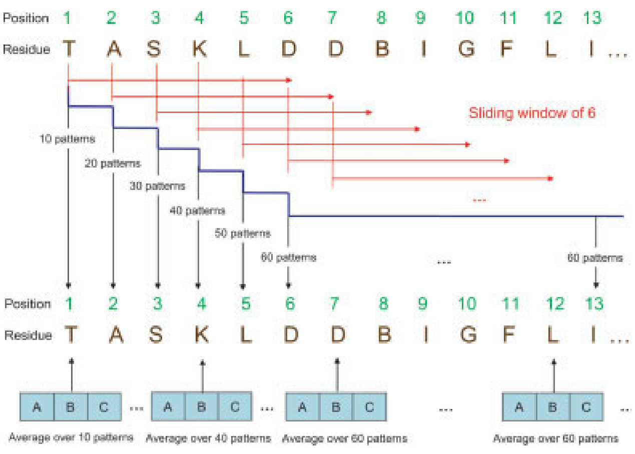


Figure 5

*Jackknife testing protocol. A window of size 6 is advanced one residue at a time from left to right. For each pattern, the Z-values associated with the secondary structure (A–C) components are looked up in the classification table. The values for all 10 patterns are then averaged over all six positions associated with each window (after correcting for end effects). Secondary structure assignment is based on the secondary structure component with the largest positive Z-*value.

test set split. The second case was drawn from the IC95 level with a 90–10% training/test set split. Figure 6(a) shows the profile for an aspartate receptor, which is an example of an all-alpha structure (PDB code 1lih, CATH code 1.20.120.30). The *y*-axis represents the average *Z*-values for α, β, and coil from the corresponding training set. The two color bars show the actual (experimental) secondary structure and the predicted (theoretical) secondary structure, respectively. It can be seen that the largest *Z* score is significantly higher than its nearest neighbor at most positions. For this sequence, the secondary structure prediction accuracy was 83%. Figure 6(b) shows the profile for carbonic anhydrase, an example of an α+β structure (PDB code 1ca2, CATH code 3.10.200.10). For this sequence, the secondary structure prediction accuracy was 84%. The ribbon diagrams of the two proteins color-coded according to the actual secondary structure (left) and the predicted secondary structure (right) are illustrated in Figure 7 for the two examples.

Jackknife test results

Table I shows the jackknife test results for the complete set of eight IC levels (rows) and the 10 different random seeds (columns), for the training/test set split of 90–10%. Table II shows the same results for a training/test set split of 50–50%. Training and test sequences are mutually exclusive in all cases. Inspection of Table I reveals that the classification accuracy does not vary significantly over the different random seeds. The average classification accuracy (listed in the last column) rises from a low of 61% to a high of 75% as the IC level rises from IC10 to IC100. The results in Table II show no significant differences compared to the results in Table I. All subsequence analysis will therefore be confined to results obtained with a 90–10% split.

The improvement in the accuracy of prediction in Table I with advancement from the IC10 level to IC100 level is partly due to the incorporation of evolutionary information from related family members. There is also an artifact related to skewing of the database from overrepresentation of selected proteins. Accuracy as a function of IC level is shown by the red curve in Figure 8(a) which attempts to separate these factors. It can be seen that accuracy rises gradually in a linear fashion (approximated by the blue dotted line) from IC10 to IC90. Above the IC95 level there is a sharp break. Comparing the results in this plot with the plots in Figure 1(a,b) suggests that significant skewing of the results does not come into play until the IC95 level is exceeded.

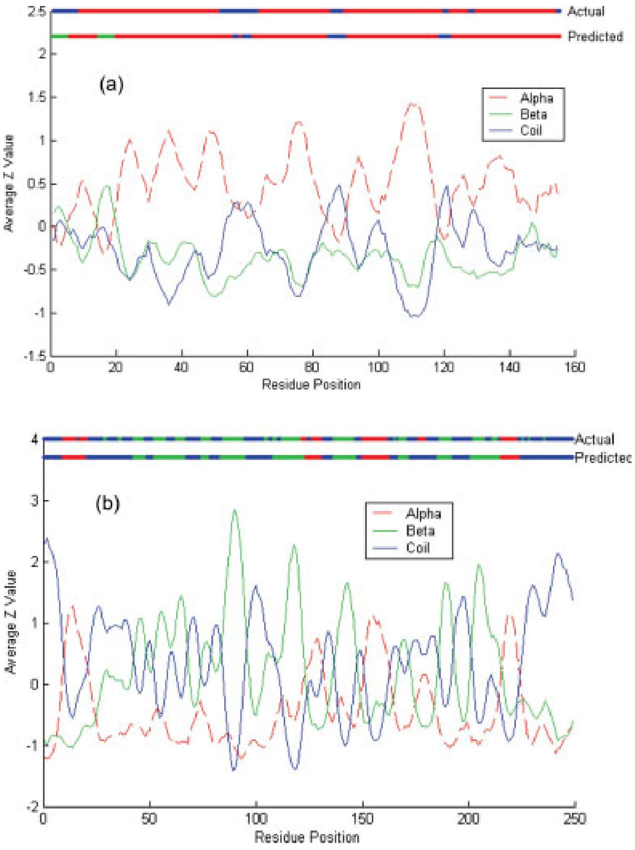


Figure 6

*(**a**) Secondary structure Z-score profiles for the aspartate receptor 1lih. At most positions, the difference between the largest Z-score and its nearest neighbor is significant. Predictions based on the largest Z-score in this case were 83% accurate. (**b**) Secondary structure Z-score profiles for carbonic anhydrase (1ca2). Predictions based on the most positive Z-score in this case were 84% accurate.*

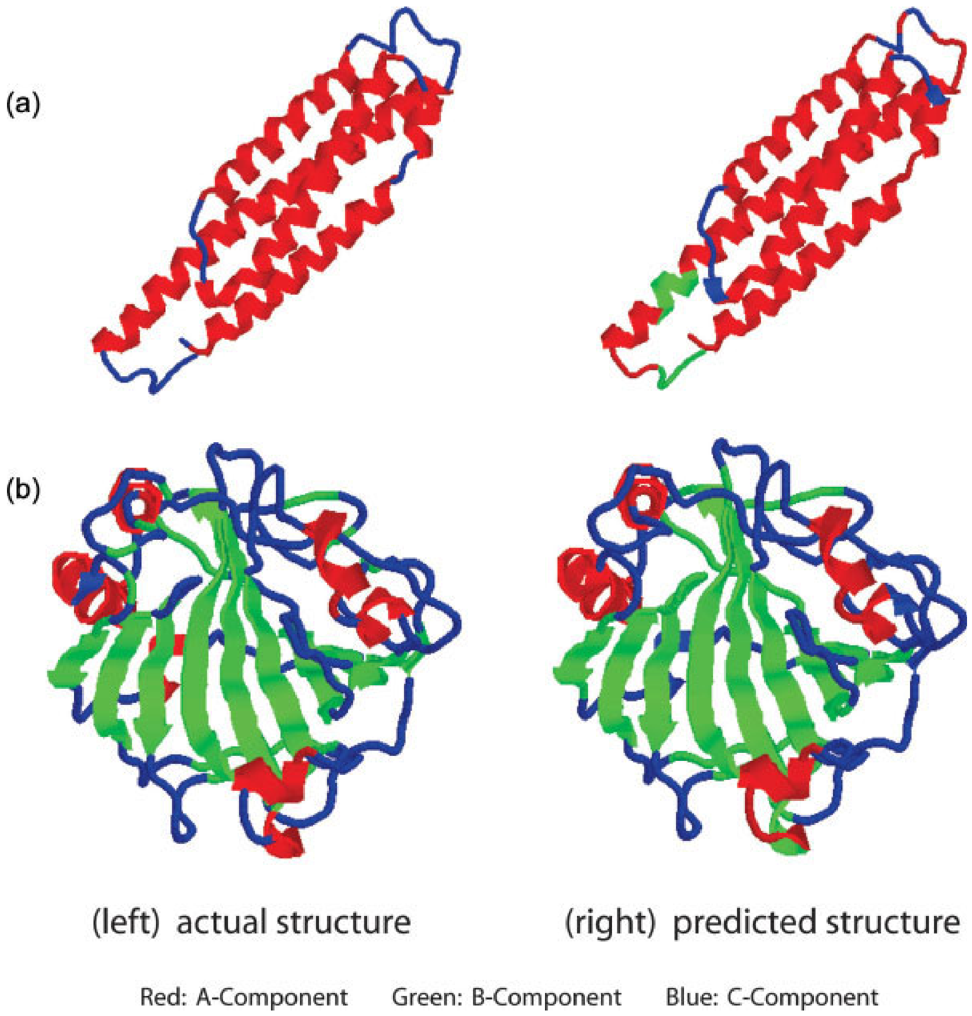


Figure 7

*Ribbon diagrams of two protein chains color-coded according to the actual secondary structure (left) and the predicted secondary structure (right), where the red color represents the A component, the green color represents the B component, and the blue color represents the C component. (**a**) Aspartate receptor (1lih). (**b**) Carbonic anhydrase (1ca2).*

Table I

Jackknife Results as a Function of Identity Level for a 90–10 Training/Test Split

	Seed										
IC%	3	5	7	11	13	17	19	23	29	31	Ave
10	0.606	0.605	0.609	0.610	0.611	0.609	0.615	0.612	0.607	0.611	0.609
25	0.613	0.612	0.615	0.611	0.618	0.616	0.614	0.619	0.622	0.610	0.615
50	0.623	0.620	0.621	0.617	0.621	0.622	0.623	0.621	0.621	0.617	0.621
75	0.624	0.628	0.627	0.629	0.630	0.632	0.629	0.633	0.628	0.631	0.629
90	0.639	0.640	0.641	0.646	0.644	0.638	0.639	0.641	0.641	0.645	0.641
95	0.654	0.656	0.660	0.656	0.649	0.656	0.653	0.660	0.658	0.659	0.656
99	0.707	0.706	0.707	0.707	0.704	0.706	0.705	0.704	0.703	0.706	0.705
100	0.745	0.746	0.745	0.745	0.745	0.745	0.746	0.744	0.747	0.746	0.745

Table II
Jackknife Results as a Function of Identity Level for a 50–50 Training/Test Split

	Seed										
IC%	3	5	7	11	13	17	19	23	29	31	Ave
10	0.604	0.603	0.603	0.603	0.602	0.605	0.604	0.604	0.604	0.603	0.604
25	0.610	0.609	0.610	0.609	0.608	0.610	0.610	0.610	0.612	0.609	0.610
50	0.617	0.615	0.615	0.615	0.616	0.616	0.617	0.614	0.616	0.613	0.615
75	0.622	0.623	0.624	0.625	0.623	0.623	0.626	0.625	0.624	0.625	0.624
90	0.635	0.635	0.633	0.636	0.634	0.636	0.634	0.635	0.637	0.635	0.635
95	0.649	0.651	0.650	0.649	0.651	0.651	0.650	0.650	0.650	0.649	0.650
99	0.701	0.699	0.700	0.700	0.702	0.701	0.701	0.700	0.702	0.701	0.701
100	0.744	0.743	0.743	0.743	0.743	0.743	0.744	0.743	0.744	0.744	0.743

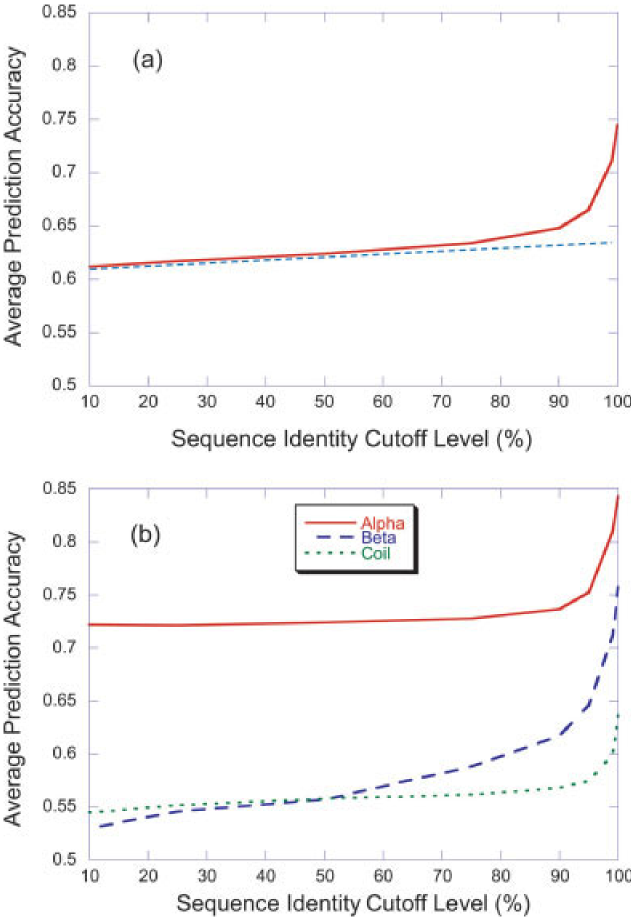


Figure 8
(***a***) *Average accuracy of secondary structure prediction as a function of IC level if the examined set of structures (red curve). A slow gradual rise exists between IC10 and IC90. Above IC95 a sharp rise occurs, as indicated by the departure from the best fitting line (blue, dashed). This suggests that significant database skewing does not enter the picture until the IC90-IC95 level.* (***b***) *Prediction accuracy as a function of structure type. Alpha helix prediction is much better than beta sheet or coil prediction. Beta sheet prediction is more strongly influenced by inclusion of additional protein family members that alpha helix or coil prediction.*

Prediction accuracy as function of secondary structure type

The separate breakdown of prediction accuracy for α-, β-, and coil-regions as a function of IC level is shown in Figure 8(b). This figure demonstrates that the results for

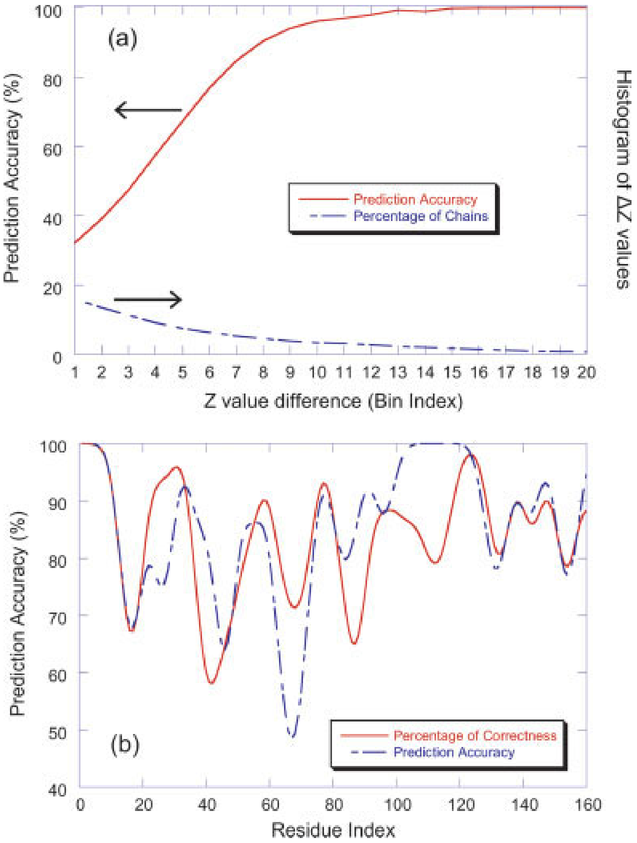


Figure 9
(***a***) *Z difference distribution and associated prediction accuracy. This relation can be used to construct a lookup table to supply percentages of correctness for individual positions in sequences.* (***b***) *Percentage of correctness versus prediction accuracy using the structure of Tnf receptor associated factor 2 (Traf2, PDB ID 1ca4).*

DOI 10.1002/prot

α-helices are significantly better than those for β-sheets or coils. The accuracy for α-helix prediction at IC10 is 71%. This rises to 76% at IC95. At an IC level of 10, most related family members have been purged from the database. The prediction power of the NP{4,2} patterns at this level is based predominantly on local propensity. The gradual increase in accuracy as the IC level rises reflects the evolutionary information gleaned from adding homologous family members to the mix. The final rise probably reflects skewing of the database. The accuracy profile for β-sheets shows significantly more improvement with rising IC levels than α-helices or coils. This suggests that evolutionary information is more important for β-sheet stabilization than for the other secondary structure types. The difference between the accuracy rates for α-helices and β-sheets at the IC10 level suggests that local propensities are more important for α-helix accuracy, consistent with the hydrogen bond pattern between near neighbors (i, $i + 4$) along the sequence.

Prediction confidence analysis

We examined the distributions of the Z values for the α-, β-, and coil-regions as a function of prediction accuracy over all 160 training sets. The distribution of the Z-values for each type of structure was close to the normal distribution. The accuracy values tended to improve the further the Z-value deviated from the mean. We also examined the prediction accuracy as a function of Z-value differences or ΔZ values. The resulting histogram of ΔZ values and associated prediction accuracies is displayed in Figure 9(a).

A lookup table can be constructed from these relationships relating ΔZ value to percentage of correctness (probability of being a correct prediction) for individual positions in sequences. Figure 9(b) shows a sample plot of secondary structure prediction accuracy (dashed line) versus the probability of being correct (solid line) for the TNF receptor (1ca4). It can be seen that the prediction accuracy from jackknife testing closely follows the probability of correctness derived from the lookup table.

DISCUSSION AND CONCLUSIONS

The jackknife test results based on the largest Z-score show that there is a strong correlation between NP{4,2} n-gram patterns and secondary structure type. To quantify the relationship, we had to first deal with the skewed distribution of sequences contained in the PDB and differentiate the contribution of local structural propensities from the contribution of family membership. We designed 100 IC levels based on unique 4-grams to quantify these effects and the effect of database skew. The 58,831 PDB chains were assigned to 100 IC levels based on the maximum allowable similarity between sequences at the same level. The gradual rise between IC10 and IC95 represented the gradual accumulation of homologous protein family members. Overall, the prediction accuracy for 50–50% training/test split ranged from 60.4% at the IC10 level to 65.0% at the IC95 level. We further broke down prediction into secondary structure type. For α-helices, the accuracy ranged from 72.2 to 75.2%. For β-strands, the accuracy ranged from 53 to 64.6%, and for coils from 54.5 to 57.4%. Prediction results improved as the IC level increased from 10 to 95% and improvement for β-strands was significantly higher, in general [see Fig. 8(b)]. This may be explained by the fact that β-structures are more heavily influenced by remote contacts than by local propensities. While this is not a surprising result, the relative contribution of evolutionary information to α-helix and β-sheet prediction has not been quantified in previous studies.

The results of the current studies show that NP{4,2} patterns capture a combination of evolutionary and local propensity information that correlates strongly with secondary structure. This suggests that NP{4,2} patterns might be a useful input for prediction algorithms that require both local and global information. A recent study by Birzele and Kramer indeed confirms that the frequency of patterns of conserved amino acids provides information that is complementary to established methods.[33] Although the utility of secondary structure prediction algorithms is limited for proteins with known structural homologs, there is still a small niche where these may be useful. Future studies are planned, which use NP{4,2} patterns as inputs to neural nets using training sets from the PDB, which reflect both secondary structure and evolutionary relationships. The NP{4,2} pattern also appears to be well suited for correlation with other parameters derived from 3D coordinate data such as packing density, surface accessibility, and protein dynamics. Future studies are also planned which will correlate NP{4,2} patterns with the collective modes predicted by elastic network models[34] and the inter-residue contact topologies reflected in Kirchhoff connectivity matrices.[35] The basic idea is to correlate sites with unusual combinations of features with functional sites determined from experimental data.

REFERENCES

1. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 1951;37:205–211.
2. Kendrew JC, Watson HC, Strandberg BE, Dickerson RE, Phillips DC, Shore VC. The amino-acid sequence X-ray methods, and its correlation with chemical data. Nature 1961;190:666–670.
3. Ptitsyn OB, Finkelstein AV. Theory of protein secondary structure and algorithm of its prediction. Biopolymers 1983;22:15–25.
4. Rost B. Rising accuracy of protein secondary structure prediction. In: Chasman D, editor. Protein structure determination, analysis, and modeling for drug discovery. New York: Dekker; 2005.

5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
6. Liu Y, Carbonell J, Klein-Seetharaman J, Gopalakrishnan V.Context sensitive vocabulary and its application in protein secondary structure prediction. ACM International Conference on Research and Development in Information Retrieval; 2004.
7. Wu CH, Zhao S, Chen HL, Lo CJ, McLarty J. Motif identification neural design for rapid and sensitive protein family search. Comput Appl Biosci 1996;12:109–118.
8. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
9. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 1987;195:957–961.
10. Pollastri G, McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 2005;21:1719–1720.
11. Jones DT, Swindells MB. Getting the most from PSI-BLAST. Trends Biochem Sci 2002;27:161–164.
12. Kloczkowski A, Ting KL, Jernigan RL, Garnier J. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins 2002;49:154–166.
13. Lin K, Simossis VA, Taylor WR, Heringa J. A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 2005;21:152–159.
14. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. Proteins 2002;46:197–205.
15. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002;47:228–235.
16. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
17. Karplus K, Hu B. Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. Bioinformatics 2001;17:713–720.
18. Karplus K, Katzman S, Shackleford G, Koeva M, Draper J, Barnes B, Soriano M, Hughey R. SAM-T04: what is new in protein-structure prediction for CASP6. Proteins 2005;61(Suppl 7):135–142.
19. Koh IY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Grana O, Pazos F, Valencia A, Sali A, Rost B. EVA: evaluation of protein structure prediction servers. Nucleic Acids Res 2003;31:3311–3315.
20. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
21. Bystroff C, Shao Y. Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. Bioinformatics 2002;18(Suppl 1):S54–S61.
22. Kuznetsov IB, Rackovsky S. On the properties and sequence context of structurally ambivalent fragments in proteins. Protein Sci 2003;12:2420–2433.
23. Gong H, Isom DG, Srinivasan R, Rose GD. Local secondary structure content predicts folding rates for simple, two-state proteins. J Mol Biol 2003;327:1149–1154.
24. Vries JK, Munshi R, Tobi D, Klein-Seetharaman J, Benos PV, Bahar I. A sequence alignment-independent method for protein classification. Appl Bioinformatics 2004;3:137–148.
25. Penel S, Morrison RG, Mortishire-Smith RJ, Doig AJ. Periodicity in alpha-helix lengths and C-capping preferences. J Mol Biol 1999;293:1211–1219.
26. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
27. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
28. Bornot A, de Brevern AG. Protein beta-turn assignments. Bioinformation 2006;1:153–155.
29. Lodish H, Berk A, Zipursky L, Matsudaira P, Baltimore D, Darnell J. Molecular cell biology. New York: W.H. Freeman; 2000.
30. Rost S, Fregin A, Hunerberg M, Bevans CG, Muller CR, Oldenburg J. Site-directed mutagenesis of coumarin-type anticoagulant-sensitive VKORC1: evidence that highly conserved amino acids define structural requirements for enzymatic activity and inhibition by warfarin. Thromb Haemost 2005;94:780–786.
31. Kabani M, Beckerich JM, Gaillardin C. Sls1p stimulates Sec63p-mediated activation of Kar2p in a conformation-dependent manner in the yeast endoplasmic reticulum. Mol Cell Biol 2000;20:6923–6934.
32. Shannon CE. A mathematical theory of communications. Bell Syst Tech J 1948;27:379–423.
33. Birzele F, Kramer S. A new representation for protein secondary structure prediction based on frequent patterns. Bioinformatics 2006;22:2628–2634.
34. Bahar I, Rader AJ. Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol 2005;15:586–592.
35. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 1997;2:173–181.