Protein Science (1995), *4*:2517-2525. Cambridge University Press. Printed in the USA.

A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%

PERDEEP K. MEHTA,[1] JAAP HERINGA, AND PATRICK ARGOS

European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10 22 09, D-69012 Heidelberg, Germany

(RECEIVED July 31, 1995; ACCEPTED September 11, 1995)

Abstract

To improve secondary structure predictions in protein sequences, the information residing in multiple sequence alignments of substituted but structurally related proteins is exploited. A database comprised of 70 protein families and a total of 2,500 sequences, some of which were aligned by tertiary structural superpositions, was used to calculate residue exchange weight matrices within α-helical, β-strand, and coil substructures, respectively. Secondary structure predictions were made based on the observed residue substitutions in local regions of the multiple alignments and the largest possible associated exchange weights in each of the three matrix types. Comparison of the observed and predicted secondary structure on a per-residue basis yielded a mean accuracy of 72.2%. Individual α-helix, β-strand, and coil states were respectively predicted at 66.4, 66.7, and 75.8% correctness, representing a well-balanced three-state prediction. The accuracy level, verified by cross-validation through jack-knife tests on all protein families, dropped, on average, to only 70.9%, indicating the rigor of the prediction procedure. On the basis of robustness, conceptual clarity, accuracy, and executable efficiency, the method has considerable advantage, especially with its sole reliance on amino acid substitutions within structurally related proteins.

Keywords: amino acid sequences; multiple sequence alignment; protein secondary structure; secondary structure prediction

Successful prediction of protein tertiary structure is presently and principally based upon a knowledge-based modeling of side chains in proteins with sequence homologous to one with known structure. With the gap widening between the number of known tertiary and primary structures, and the inability of experimental techniques such as X-ray crystallography or solution NMR to determine the structure of all known proteins, it is necessary to develop theoretical approaches that deduce structure from sequence. To make feasible such efforts, a prediction of secondary structural elements (α-helices, β-strands, coil segments) along the sequence is likely to be of utility. Though early attempts at such predictions appeared some 20 years ago (e.g., Chou & Fasman, 1974; Lim, 1974; Garnier et al., 1978), precise and accurate prediction of secondary structural elements from the amino acid sequence alone has not been achieved (Garnier & Levin, 1991). With accuracies for single sequences generally ranging between 55 and 65%, theoretical determination of the fold of a protein appears unreachable.

Within the last few years, several new methods based on complex analytical procedures have surfaced, generally for single sequence prediction. Muggleton et al. (1992) have applied inductive logic programming. Their machine-learning algorithm allowed a high 81% prediction accuracy, albeit on only four $\alpha\beta$ domain proteins. The use of single three-layered neural networks (cf. Qian & Sejnowski, 1988; Holley & Karplus, 1989; Zhang et al., 1992) have resulted in 63–67% correctness for single sequence prediction, and enhanced neural network procedures (Kneller et al., 1990) have achieved 79% if the query protein is known to contain largely α-helices. The multilayered neural network method of Rost and Sander (1993), applied to multiply aligned protein sequences, yielded a per-residue prediction accuracy near 71%. Leng et al. (1994) developed a complicated two-level case-based reasoning architecture to predict protein secondary structure in single sequences and achieved a mean 68.2% accuracy over several proteins. Salamov and Solovyev (1995) have reported an improvement of the already published technique of Bowie et al. (1991) based on a complex combination of nearest-neighbor algorithms, including secondary structural terminal features and multiple sequence alignments, to yield an accuracy of about 72%. The method presented here, which also relies on knowledge of homologous sequences, is significantly simpler and straightforward. It probes with facility residue exchange statistics to deter-

Reprint requests to: Patrick Argos, European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10 22 09, D-69012 Heidelberg, Germany.

[1] Present address: Biochemisches Institut der Universität Zürich, CH-8057 Zürich, Switzerland.

mine residue variations characteristic of each of the three secondary structural types by examination of substitutions at structurally equivalent positions in evolutionarily related proteins. Given a multiple sequence alignment, mutations in given subsequence regions that correlate best with those preferred by a particular secondary structural state can be utilized as prediction signals. The approach appears significant on the basis of accuracy, robustness, conceptual clarity, and executable efficiency. Donnelly et al. (1994) used protein environment-specific amino acid substitution tables (Lüthy et al., 1991; Overington et al., 1992) to predict and orient α-helices from sequence alignments for protein structural modeling but did not attempt an overall three-state prediction from sequence alone.

Methodology

A computer algorithm has been developed to predict secondary structural elements (α-helix, β-strand, coil) in a query amino acid sequence that can be multiply aligned with those of other structurally related proteins. The technique involved three major calculational steps and procedures. The computer routine PreferCal was first written to determine the preference or avoidance weights for each possible pair of residue exchanges and for each of the three secondary structural states considered here. The PreferPred was used to predict secondary structural elements within a query sequence multiply aligned to related primary structures. Finally, PreferEval allowed evaluation of the accuracy of the secondary structure predictions relative to those known from three-dimensional structural determinations. All routines were written using the standard ANSI C language under a VMS operating system on a VAX station 3100.

Construction of secondary structure-specific residue exchange matrices

Input to PreferCal was taken from the 3D_Ali database of Pascarella and Argos (1992), which contains more than 70 protein families, each defined by proteins with similar main-chain fold. Membrane-spanning proteins were not included in the calculation of the residue exchange statistics. Each 3D_Ali file corresponds to a protein family with a uniquely folded domain (or monomer). Familial sequences with experimentally determined tertiary structures were aligned by structural superposition, albeit in about half of the families, only one homologous three-dimensional structure was known. In each case, the secondary structural state of individual residues was annotated as determined from the tertiary folds. Protein sequences, taken from large molecular biological databases and for which a structure had not yet been determined but were found to be at least 50% identical in residue matches after alignment to a sequence with determined structure, were then multiply aligned to the structural sequence(s). A duplicate set of files was also used, but now the 3D_Ali entries contained known primary structures at the 35% identity level or greater (referred to as 3D_Ali_35 data). The work of Pascarella and Argos (1992) should be consulted for details.

3D_Ali families containing only one (or more) sequence(s) with known tertiary structure are referred to here as single (or multiple) structure families. There are 38 3D_Ali sequence sets, each containing two or more sequences with experimentally solved fold. However, three of these families were not utilized in this work because either their structure depended on binding another protein (inhibitors) or they contained metal clusters relying on disulfide bridges (two ferrodoxin families). There are an additional 25 families that contain only one sequence of known fold and more than one primary structure (with unknown fold) taken from large protein sequence databases (excluding crambin, which is a peculiarly hydrophobic protein not amenable to prediction).

For the calculation of the residue exchange statistics for each secondary structural type, entries with at least two sequences with experimentally determined three-dimensional structures were used. The secondary structure of each residue as determined from the known fold was used to assign a structural state to each alignment position in the multiple sequence blocks. Unique residue exchange frequencies were counted for each possible pair of amino acid types occurring at a given match site. Possible exchanges were only counted once; for example, if AAACCF occurred at an alignment position, the exchanges A→C, C→F, and A→F constituted the only counts and then each with a count of one. The frequencies were then summed over all alignment sites and output into an appropriate cell in each of the corresponding secondary structure-specific substitution matrices (H, E, and C for helical, strand, and coil/loop regions, respectively). For the sake of simplicity, the different types of helices, strands, turns, or bends were lumped together: H (α-helix), G (3_{10}-helix), and I (π-helix) states were called H (see Kabsch & Sander [1983] for detailed structural definitions); E (extended strand) and B (residue in an isolated β-bulge) as E; and T (H-bonded turn), S (bend), and blank (coil) as C.

To extract a consensus secondary structure from the superposed structures in a given 3D_Ali file over each of the multiple alignment positions, three different procedures were attempted; namely, highly stringent, semi-stringent, and flexible (or majority) rules. The stringent process considers only those families with at least three or more experimentally determined tertiary structures. A secondary structure type was not assigned unless 80% or more of the proteins with known structure had displayed a secondary structural state of the same type (H, E, or C). The semi-stringent rules, which assigned a consensus secondary structure, changed according to the number of known structures in a particular family. For instance, if four or more structures were superposed, more than 50% of them must be similar in secondary structure for assignment. If fewer than 50% of the structures contributed an amino acid to a given alignment site due to an insertion or deletion, then more than 80% of those appearing must agree for a consensus declaration. If two or three tertiary structures appeared in the file, all must show the same structural type. The flexible or majority rule simply assigned the secondary structural type that appeared most often at a given alignment position regardless of the number of structures contributing. If two substructural types were observed most often and equally in frequency, then a hierarchy of E > H > C was used to decide the consensus type. For all the rules, when an assignment criterion was not met, the position was declared of unknown type and not used in gathering statistics and not considered as a predictable site.

Certain thresholds were applied to the length of consensus secondary structural elements for their validation; namely, five or more contiguous positions for helix, three or more for strands, and five or more for coil. For purposes of calculating exchange matrices, alignment positions in the too-short regions were ignored, while, for prediction evaluation, the segments were as-

signed as coil. No limit was placed on the maximum allowable lengths for these elements.

The following motivations prompted the secondary structural length limits and consensus assignment rules. Colloc'h et al. (1993) analyzed three different methodologies for assigning three-state secondary structure types from known tertiary protein structures, one of which was DSSP (Kabsch & Sander, 1983) used in the Pascarella and Argos (1992) database. They found, for 154 proteins, that only 63% of sequence sites were designated the same by all techniques. The DSSP approach was unique in delineating a plethora of four-residue helices and two-residue strands. Thus, to be more assured of proper secondary structural designation, limits of five and three were used for helices and strands, respectively. For consensus assignment of structure, considerable noise can be introduced in the residue exchange statistics unless the structural type is adopted by several proteins at a given alignment position. Here several rules were attempted with varying strictness (see above) and those selected resulted in optimal predictions. Russell and Barton (1993) and Levin et al. (1993) have also examined these issues.

Once unique exchanges and frequencies of amino acids in multiple alignment positions assigned to secondary structural types were counted over all protein families, the amino acid substitution weight matrices were constructed for each secondary structural type. The matrix values were symmetric about the diagonal because no direction could be assigned to the residue substitutions. For example, helix-specific matrix cells $M(i,j)_{\text{helix}}$ were calculated as follows:

$$M(i,j)_{\text{helix}} = \frac{S_{i,j,\text{helix}} \Big/ \sum_{i=1}^{20} \sum_{j=1}^{20} S_{i,j,\text{helix}}}{S_{i,j,\text{all}} \Big/ \sum_{i=1}^{20} \sum_{j=1}^{20} S_{i,j,\text{all}}}$$

$$i \neq j$$

and

$$M(i,i)_{\text{helix}} = \frac{A_{i,\text{helix}} \Big/ \sum_{i=1}^{20} A_{i,\text{helix}}}{A_{i,\text{all}} \Big/ \sum_{i=1}^{20} A_{i,\text{all}}},$$

where A_i refers to the ith amino acid type (amino acid frequency for given secondary structural state), $S_{i,j}$ to substitutions of the ith and jth types, "helix" to only those residues that occur in helical secondary structural regions, and "all" to residues over the entire tertiary structure regardless of secondary structural state. Similar determinations were made for the strand and coil conformations. A matrix cell value greater (less) than 1.00 would indicate a preference (avoidance) for the residue substitution or residue type within the given secondary structural state, whereas values at 1.00 suggest neutrality.

Prediction of secondary structure

The secondary structure prediction method discussed here requires as input residue exchange weight matrices for each of the three substructural states and a single query sequence to be predicted or a multiple sequence alignment that contains the query sequence and from which predictions are made. The program can accept a single sequence or a multiple sequence alignment in the database format of 3D_Ali (Pascarella & Argos, 1992) and HSSP (Sander & Schneider, 1991) or it can recognize MSF files generated by the multiple alignment routine PileUp of the GCG package (Genetics Computer Group, 1991). A single sequence input will prompt a search of large protein sequence databases for related members, and then a subsequent multiple alignment (if possible), followed by a secondary structure prediction for the query sequence. The substitution matrices, acting as input to PreferPred, are precalculated with PreferCal. The PreferPred prediction is for three states (α-helix, β-strand, or coil) shown along the length of the query amino acid sequence as a single sequence or within a multiple alignment.

The prediction technique is composed of the following procedures. The occurrence of each amino acid type is noted only once at each alignment position (or query sequence site for a single sequence prediction). All possible exchanges are collated for each position, independent of substitution direction. Nondiagonal and diagonal values, taken from the exchange weight matrices for each substitution and residue type respectively, are then summed for each of the three structural types and for each alignment (single sequence) position. The resulting sums are then added and averaged for each structural type over a sliding window that encompasses a given number of contiguous alignment positions and is centered on the position to be predicted. This process is repeated for all possible windows along the alignment or single sequence, i.e., it slides along the multiple alignment in steps of one. At each alignment position or window center, the highest window average over the three states predicts the site to be in α-helix (H), β-strand (E), or coil (C). Tests were made with various window lengths in the range 3–13 in steps of 2 to achieve optimal predictions.

Before the final prediction is made, certain cleaning or filtering rules are introduced to remove the predicted secondary structure elements of unacceptable length or unacceptably interrupted by other substructural-type predictions. These cleaning rules are completed in three successive cycles such that the results from application of a previous filter are taken before effecting the next filter.

Round I cleans single position interruptions such that, if two flanking alignment (single) sequence sites are predicted in one structural state with the middle in another state, the middle position is assigned according to the consistent flanks. For example, three successive predictions of (H, C/E, H) becomes (H, H, H) where C/E indicates C or E. Round II cleans double position interruptions such that in five positions, three flanks are of one structural state and two middle sites are of another. For instance, (H, H, C/E, C/E, H) or (H, C/E, C/E, H, H) becomes (H, H, H, H, H). Finally, in Round III, all helices less than or equal to 4 in length and all strands less than or equal to 2 in length are altered to coil predictions; for instance, (H, H, H) becomes (C, C, C). Various length thresholds in each of the cleaning steps were attempted; those reported were found optimal.

Evaluation of predicted secondary structure and cross-validation

The PreferEval routine was used to evaluate the predicted secondary structure of a protein. The assessment criterion used was

the fraction of residues predicted correctly in a given protein or alignment relative to the observed or consensus secondary structure for the protein or alignment sites. Alignment sites that could not be assigned a consensus secondary structure, as previously described, were ignored in the evaluation. The total number of amino acids appearing at all sequence alignment sites, regardless of type or frequency of appearance and assigned structural state, was used to normalize the number of residues predicted correctly, which were counted in similar fashion at each of the alignment sites. The final, overall evaluation for a set of protein families was taken as their mean accuracy.

For cross-validation, a comprehensive jack-knife test was used where secondary structure-specific substitution matrices were calculated using all but one protein family in a data set and then prediction and evaluation performed on the deleted family. Hence, no information on the predicted family was included in the generation of substitution matrices. As the calculation time for the substitution matrices and predictions were minimal, a jack-knife test could be repeated for each family. An average over all such tests was calculated to yield a stringent estimate of the overall prediction performance.

Results and discussion

The first task involved the generation of secondary structure-specific residue exchange weights, which are represented as 20×20 symmetric matrices with each cell containing a preference (or avoidance) value according to the acceptability of a particular residue substitution (e.g., Gly↔Ala) or residue type (for diagonal elements) in a given secondary structural state including α-helix, β-strand, and coil. The matrices were calculated from aligned sequences of protein families as they appear in the 3D_Ali database of Pascarella and Argos (1992). The 70 families of aligned sequences consist of at least one protein sequence with known three-dimensional structure and then one or more related sequences (taken from protein sequence databases) with at least 50% residue identity after alignment to any one sequence with known fold. A second set of files over the same 70 families with the sequence identity at the 35% identity level was also used. For both sets, the sequences associated with experimentally determined structures were aligned by structural superposition of main-chain C_α atoms; secondary structural assignments were ascertained for these primary structures from the main-chain dihedral angles and hydrogen bonding states (Kabsch & Sander, 1983).

Several conditions involving the use of these files were attempted to achieve the optimal secondary structure predictions. The structure-specific residue exchange weights were calculated utilizing all the 3D_Ali files or only those with two or more known structures to allow better delineation of conserved secondary structures through consensus. Files based on 50% residue identity or 35% provided further distinction. The rules for consensus secondary structure assignment were varied according to three levels of stringency for agreement amongst the known tertiary architectures as described in the Methodology section. Further, various minimal lengths were attempted respectively for assignment and prediction of α-helices, β-strands, and coil segments. Finally, secondary structure predictions were optimized according to several odd-numbered window lengths encompassing contiguous alignment positions over which summed residue exchange weights at each site were added and averaged for each structural state and then assigned to the central window position. The largest of the three state values provided the secondary structural type predicted. For all the various tested conditions mentioned, the prediction evaluation was performed over all families used and all alignment sites with assigned secondary structure.

It was found that substitution matrices used here for prediction are best calculated with families consisting of two or more known structures and database sequences added at the 50% residue identity level or greater. Figure 1 shows the matrices for each of the three substructural states. The use of both single- and multiple-structure families in combination reduced the prediction accuracy by as much as 5%. The presence of several experimentally determined folds in a given family (multiple structure) as well as highly related sequences (50% identity) allowed better recognition of secondary structurally conserved regions, which in turn yielded exchange statistics more characteristic of the structural type. The 35 multiple-structure families employed are listed in Table 2.

In assigning a consensus secondary structural state to the aligned tertiary structures, the most flexible (majority) rule proved to be the best. Though this result could seem contrary to expectation, greater residue exchange statistics were provided by the expanded number of assigned positions. Also in *calculating* the residue exchange matrices, only helical, extended, and coil segments with minimal lengths of 5, 3, and 5, respectively, were utilized, once again providing assurance of structural type with appropriate residue substitutions. It must be emphasized that in *evaluating* the predictions, all helices and strands less than five and three residues in length, respectively, were not ignored and were assigned as coil regions; all coil segments were accepted for evaluation regardless of length. The percentage of residues predictable was found to be 75.1% when alignment positions where no secondary structural assignment could be elicited are excluded. Our method relies on multiple structures and, if specific structurally equivalenced positions are not in agreement regarding secondary structural type, it is impossible to check for prediction accuracy as there is no standard-of-truth. It should be noted, however, that the mean prediction accuracy for single-structure families, where all sequence sites are predicted, is 68%, which represents the lower limit should all residues be included. It should also be noted that the single-structure families contain fewer multiply aligned sequences and therefore yield less accurate predictions.

It was observed that sliding windows over the alignment positions to predict the central window residue should be of different length for each structural class; namely, 9, 7, and 5 for helical, strand, and coil exchange statistics, respectively. The prediction accuracy was generally improved by about 5% when using different lengths in contrast to the same window for all types. To validate properly the optimization of the basic parameters, jack-knife tests were performed while varying the specific structure length thresholds. For helices (H), 3–12 was used in steps of two with the strand and coil lengths fixed at 7 and 5, respectively. In each case, the mean percent accuracy was calculated with the appropriate substitution matrices and then jack-knife tests performed for each protein family. The mean percent accuracy over all families ranged between 69.9% ($H = 3$) and 72.1% ($H = 9$), and the corresponding jack-knife results encompassed 68.7% and 70.9%. The prediction accuracy changed only about 1.2% in eliminating each family from the exchange

A

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	1.23																			
Cys	1.41	1.04																		
Asp	1.31	1.44	1.11																	
Glu	1.34	1.43	1.31	1.19																
Phe	0.94	1.16	0.79	0.91	0.82															
Gly	1.11	1.36	1.02	1.06	0.68	0.86														
His	1.17	1.50	1.05	1.23	0.94	1.10	0.99													
Ile	1.08	1.24	0.93	0.96	0.74	0.79	0.86	0.90												
Lys	1.29	1.31	1.18	1.22	0.78	1.05	1.10	0.94	1.13											
Leu	1.19	1.41	1.00	1.01	0.90	0.88	1.08	0.99	1.07	1.04										
Met	1.20	1.36	0.94	1.02	0.91	0.81	1.08	1.01	0.99	1.08	1.04									
Asn	1.28	1.36	1.08	1.21	0.83	0.96	1.13	0.89	1.14	0.98	1.01	1.02								
Pro	0.94	1.30	0.84	0.96	0.69	0.92	0.80	0.67	0.67	0.75	0.59	0.74	0.75							
Gln	1.31	1.49	1.25	1.26	0.88	1.11	1.22	1.02	1.21	1.01	1.10	1.20	0.91	1.16						
Arg	1.17	1.21	1.18	1.15	0.74	0.99	1.06	0.91	1.11	1.02	0.97	1.03	0.69	1.10	1.06					
Ser	1.17	1.33	0.97	1.02	0.78	0.94	1.03	0.89	0.98	1.01	0.59	1.01	0.79	1.09	0.90	0.93				
Thr	1.06	1.18	1.01	0.94	0.72	0.86	0.95		0.88	0.90	0.96	0.93	0.80	1.02	0.88	0.87	0.85			
Val	1.09	1.33	1.01	1.04	0.76	0.95	1.01	0.83	0.99	0.88	1.02	0.96	0.77	1.06	0.98	0.93	0.84	0.83		
Trp	0.90	0.78	0.87	0.86	0.66	0.58	0.92	0.53	0.74	0.72	0.75	0.88	0.40	0.95	0.63	0.69	0.74	0.60	0.91	
Tyr	0.86	1.00	0.73	0.76	0.71	0.58	0.81	0.68	0.79	0.79	0.86	0.83	0.54	0.85	0.75	0.63	0.65	0.67	0.66	0.83

B

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	0.75																			
Cys	0.75	1.07																		
Asp	0.56	0.67	0.66																	
Glu	0.63	0.69	0.58	0.74																
Phe	1.16	1.04	1.17	1.17	1.33															
Gly	0.83	0.80	0.61	0.87	1.30	0.97														
His	0.79	0.69	0.68	0.78	1.05	0.71	0.99													
Ile	1.11	0.96	1.06	1.15	1.54	1.25	1.24	1.33												
Lys	0.65	0.79	0.64	0.78	1.11	0.76	0.82	1.11	0.78											
Leu	0.97	0.84	0.96	1.12	1.27	1.09	1.07	1.21	0.99	1.12										
Met	0.97	0.97	0.93	0.96	1.21	1.18	0.96	1.16	1.03	1.06	1.05									
Asn	0.71	0.72	0.69	0.74	1.11	0.71	0.78	1.12	0.72	0.97	0.92	0.80								
Pro	0.77	0.67	0.75	0.75	1.07	0.75	0.83	1.30	0.92	1.21	1.34	0.91	0.82							
Gln	0.68	0.65	0.64	0.73	1.07	0.78	0.79	1.10	0.75	1.07	0.97	0.71	0.83	0.80						
Arg	0.80	0.84	0.63	0.82	1.26	0.84	0.92	1.23	0.84	1.05	1.14	0.88	1.03	0.83	0.90					
Ser	0.80	0.81	0.77	0.92	1.28	0.91	0.85	1.24	0.89	1.09	1.15	0.84	0.87	0.85	0.99	0.96				
Thr	0.97	0.93	0.88	1.11	1.32	1.03	0.91	1.27	1.09	1.18	1.11	0.99	0.93	1.02	1.08	1.10	1.16			
Val	1.05	0.91	0.99	1.08	1.51	1.09	1.13	1.42	1.06	1.35	1.22	1.05	1.13	1.03	1.08	1.21	1.27	1.41		
Trp	1.18	1.47	1.04	1.33	1.31	1.28	1.18	1.71	1.11	1.37	1.35	0.95	1.09	1.17	1.33	1.34	1.42	1.63	1.22	
Tyr	1.23	1.22	1.12	1.31	1.34	1.27	1.23	1.51	1.18	1.41	1.23	1.08	1.29	1.11	1.25	1.41	1.46	1.60	1.29	1.31

C

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	0.86																			
Cys	0.49	0.71																		
Asp	1.23	0.60	1.43																	
Glu	0.96	0.57	1.17	1.01																
Phe	0.79	0.46	1.17	0.85	0.83															
Gly	1.11	0.51	1.88	1.15	1.16	1.50														
His	1.04	0.39	1.63	0.91	1.05	1.40	1.05													
Ile	0.51	0.43	1.04	0.74	0.42	0.98	0.81	0.58												
Lys	1.06	0.68	1.38	0.93	1.33	1.43	1.15	0.91	1.10											
Leu	0.57	0.28	1.09	0.69	0.63	1.10	0.62		0.84	0.61										
Met	0.53	0.10	1.34	1.02	0.74	1.07	0.88	0.59	0.97	0.63	0.77									
Asn	0.95	0.70	1.52	1.04	1.19	1.80	1.18	1.00	1.30	1.13	1.17	1.36								
Pro	1.69	0.96	2.04	1.71	1.66	1.80	1.96	1.18	2.08	1.19	1.28	1.91	2.13							
Gln	0.92	0.51	1.17	0.96	1.16	1.22	0.92	0.70	1.02	0.80	0.79	1.13	1.63	0.96						
Arg	1.02	0.82	1.41	1.02	1.09	1.41	1.02	0.69	1.08	0.82	0.74	1.19	1.76	1.13	1.06					
Ser	1.04	0.55	1.63	1.13	0.94	1.37	1.29	0.74	1.32	0.74	0.79	1.35	1.87	1.12	1.29	1.29				
Thr	0.92	0.68	1.27	0.89	0.97	1.30	1.33	0.79	1.12	0.84	0.83	1.23	1.71	0.89	1.13	1.11	1.08			
Val	0.65	0.32	0.99	0.71	0.44	0.90	0.66	0.45	0.89	0.49	0.44	0.99	1.32	0.77	0.86	0.70	0.78	0.63		
Trp	0.86	0.47	1.24	0.61	1.19	1.46	0.78	0.56	1.44	0.88	0.83	1.46	2.39	0.73	1.22	1.01	0.69	0.57	0.78	
Tyr	0.84	0.48	1.46	0.91	0.99	1.51	0.97	0.65	1.13	0.57	0.84	1.27	1.54	1.14	1.10	1.03	0.85	0.48	1.22	0.83

Fig. 1. Secondary structure-specific residue exchange weight matrices determined from 35 structural families in the 3D_Ali_50 database (Pascarella & Argos, 1992). **A:** α-Helix. **B:** β-Strand. **C:** Coil.

matrix calculations. Given the smallness of the change and similar results in the final jack-knife control (see above), no further controls were performed regarding length threshold combinations. Various filtering or cleaning procedures were also tested and revolved about the number of residues to use in flanks and sandwiched positions; the optimal filters are detailed in the Methodology section.

For the 35 3D_Ali_50 families, where at least two sequences have known fold and the other sequences (if any) are at least at the 50% residue identity level in alignment with at least one of the sequences of known topology, a prediction accuracy of 70.6% was achieved under optimal prediction parameters given in the previous paragraph (Table 1). However, if the same structure-specific residue exchange matrices based on the 35 families at the 50% identity level were applied to the same 35 families but now with sequences from databases added that had at least 35% residue identity with at least one of the sequences with known tertiary architecture, the prediction correctness increased to 72.2%. Obviously the greater substitution information improved the result. The standard deviation of the prediction accuracy among the families was 9–10% for both cases (Fig. 2). The percentages of correctly predicted residues in helical, extended, and coil structural states were 66.4%, 66.7%, and 74.8%, respectively, a balanced prediction (Table 2).

The computer time required to predict a given family on a VAX 9600 mainframe ranged between a few seconds to a few minutes. The largest family, with 437 immunoglobulin domain sequences, each with length about 120 residues, required only 4 min. The calculation of the exchange matrices also needed little computer effort, which was never more than 30 s for a given parametric and database setting.

The robustness of the exchange matrix prediction technique was tested in several ways. An exhaustive cross-validation was performed where structure-specific residue exchange weights were derived under optimal parametric settings from the 50% identity 35-family database (3D_Ali_50) and where predictions utilized the same 35-entry data but with sequences at the 35% identity level (3D_Ali_35). For each jack-knife test, one

Table 1. *Secondary structure prediction accuracy for 35 multiple-structure protein families*

Protein family[a]	Percent correctly predicted residues		
	3D_Ali_50[b]	3D_Ali_35	Jackknife_35
1. Acid proteases (AC_PROT)	65.6	73.5	67.3
2. Plastocyanins (PLASTO)	78.7	78.8	78.8
3. Tim barrell proteins (BARREL)	78.7	83.1	83.1
4. Sugar binders (BINDING)	62.3	62.3	61.5
5. Carbonic anhydrases (CARBONIC)	74.0	76.8	75.8
6. Calcium binders (CA_BIND)	81.1	78.1	74.9
7. Crystallins (GCR)	55.8	59.8	53.7
8. Cytochrome *b*5s (CYTB)	72.5	69.3	69.3
9. Cytochrome *c*s (CYTC)	70.2	71.4	71.4
10. Cytochrome *c*3's (CYTC3)	70.8	73.9	68.3
11. Cytochrome *b*562+c' (256B)	83.6	86.9	83.3
12. Dihydrofolate reductases (DFR)	69.0	68.7	68.0
13. Trypsin inhibitors (EGLIN)	68.5	76.6	76.6
14. Hydrolases/reductases (FAD_NADH)	52.0	53.2	53.2
15. Globins (GLOBIN)	63.5	63.3	70.6
16. Immunoglobulin domains (IGB)	80.4	80.1	78.0
17. Interleukins (IL)	56.9	56.9	56.9
18. Lectins (LTN)	72.8	73.7	69.9
19. Lysozymes (LZM)	61.0	57.8	58.9
20. Dehydrogenases (NBD)	80.2	90.8	87.7
21. Papains (PAP)	60.6	70.5	70.5
22. Phospholipases (PLIPASE)	77.9	76.5	76.5
23. Phosphofructokinases (KINASE)	63.7	59.8	59.8
24. Rhubredoxins (RDX)	73.9	74.8	74.8
25. DNA repressors (REPRESSOR)	73.2	75.4	75.4
26. Rhodaneses (RHD)	58.7	58.7	58.7
27. Subtilisins (SBT)	68.7	72.9	71.5
28. Serine proteases (S_PROT)	73.3	71.0	68.7
29. Toxins (TOX)	75.2	83.8	82.3
30. Viral capsid proteins (VIRUS)	84.4	84.3	84.3
31. Wheat germ agglutinin (WGA)	99.9	99.9	99.9
32. Hemerythrins (HMR)	67.3	67.3	66.3
33. Viral proteases (VIRUS_PROT)	59.3	52.8	48.0
34. Histocompatibility antigens (HLA_A2)	67.9	70.3	64.4
35. D-Xylose isomerases (XIA)	71.8	71.8	72.4
Mean correctly predicted per residue	70.6%	72.2%	70.9%

[a] The name of the protein family as well as the identifier (in parentheses) within the 3D_Ali database of Pascarella and Argos (1992) are given.
[b] 3D_Ali_50 refers to the prediction accuracy for families where related sequences from large protein sequence databases are added to the multiple alignments such that each is at least 50% identical in matched residues to one of the sequences with known tertiary structure. 3D-Ali_35 is similar except that the 35% identity level is acceptable, resulting in a greater number of sequences in the multiple alignment. The jackknife prediction accuracy for a given family is shown based on the 3D_Ali_35 entries.

of the 35 families was removed from the substitution matrix calculations and then subsequently predicted. Each family was iteratively deleted and the resulting prediction accuracies over the 35 cross-validations averaged. The mean was 70.9%, which compares favorably with the 72.2% where all families were used for the exchange statistics and prediction evaluation. The average prediction accuracy over the 25 single-structure families, including all sequences to the 35% identity level, was determined to be 68.0% where the exchange weight matrices were calculated from the 35 multiple-structure families (3D_Ali_50). The somewhat lower prediction accuracy would be expected because the multiple structural families each contained, on average, sixfold more sequences with known topology than the lone sequence in single-structure families, thus providing considerably more substitution information for prediction. The increase in correctness from 68.0% to 70.9% would also indicate that the greater the number and diversity of sequences in the multiple alignments, the better the general prediction accuracy. Finally, predictions were run on the single sequences with known structure in the 25 single-structure families; here only the diagonal elements in the structure-specific substitution matrices were utilized for prediction. The mean accuracy was a lower 64%, once again pointing to the importance of the multiple sequence information. Nonetheless, this accuracy compares favorably with the best of the single-sequence prediction methods (Levin et al., 1993).

Figure 3 shows the observed and predicted (before and after cleaning) secondary structure for various protein examples. The

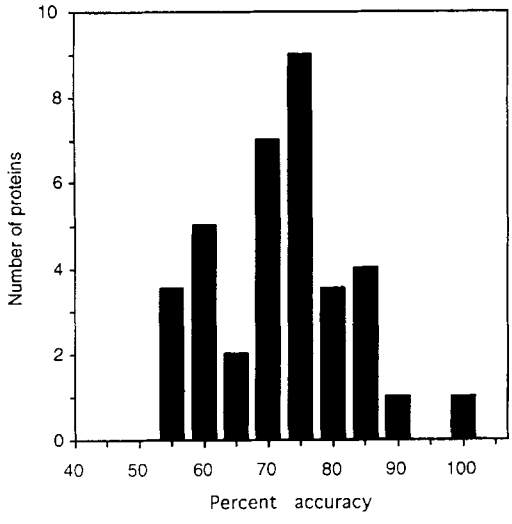

Fig. 2. Distribution of secondary structure prediction correctness for 35 multiple-structure proteins in the 3D_Ali_35 database.

results are shown relative to the first-listed sequence in the appropriate entry of the 3D_Ali_35 database. The percentage accuracy of the predictions shown ranges from 63 to 80% and the proteins selected for illustration display various topologies, including all-helix, all-strand, and mixed strand/helix.

Four methods, according to the knowledge of the authors, have been published to make use of the extra information in multiple protein sequence alignments in an automated fashion for secondary structural prediction. The approaches of Crawford et al. (1987), Bazan (1990), Barton et al. (1991), Benner and Gerloff (1991), and Russell et al. (1992) have not been automated nor have they been applied to a sufficient number of protein families for comprehensive assessment. These latter techniques will thus not be discussed further here.

Zvelebil et al. (1987) examined 11 protein families that were composed of clearly related sequences aligned by automated techniques. They averaged the secondary structure predictions based on the GOR method (Garnier et al., 1978; Gibrat et al., 1987) over the familial members and found a mean 4% prediction improvement over the 11 protein groups. Addition of further rules such as examining conservation patterns characteristic of specific secondary structural types yielded an extra 5% increase in mean prediction accuracy. Levin et al. (1993), who used the most frequent predictions from two methods over the fam-

Table 2. *Prediction accuracy for each structural type over 35 multiple-structure families with sequences at the 35% identity level*

Residues observed in	Residues predicted as Helix	Strand	Coil	Observed (%) in structure	Correctly predicted (%)
Helix	885	171	276	30.1	66.4
Strand	71	732	294	24.8	66.7
Coil	216	269	1,516	45.2	75.8

Acid proteases, AC-PROT, mainly β protein, 73.5%

```
Sequence | GEVASVPLTNYLDSQYFGKIYLGTPPQEFTVLFDTGSSDFWVPSIYCKSN
Obs SS   |    EEEEEEE     EEEEEEEE     EEEEEEE      EEE
Pred SS  |    EEEE  EE EEEEEEEEEEE    HHEEEEE    EEEEE   E HH
Clean SS |    EEEEEEEE EEEEEEEEEEE      EEEEE    EEEEE     HH

Sequence | ACKNHQRFDPRKSSTFQNLGKPLSIHYGTGSMQGILGYDTVTVSNIVDIQ
Obs SS   |                EEEEEEEEEE      EEEEEEEEEEEE  EEEEE
Pred SS  | HHHH E       H       EEE      EEEE  EEEEEEE  EEE
Clean SS | HHHH                 EEE      EEEE  EEEEEEEEEEEE

Sequence | QTVGLSTQEPGDVFTYAEFDGILGMAYPSLASEYSIPVFDNMMNRHLVAQ
Obs SS   | EEEEEEEEE            EEEEE          HHHHHHH
Pred SS  | EEEEEEH            EEEEEE       E  EEEEEEE HHEHHHEE
Clean SS | EEEEEE             EEEEEE          EEEEEEE HHHHHHEE

Sequence | DLFSVYMDRNGQESMLTLGAIDPSY
Obs SS   |  EEEEE        EEEEEE
Pred SS  | EEEEEEE       EEEEE
Clean SS | EEEEEEE       EEEEE
```

Globins, GLOBIN, all-α protein, 63.3%

```
Sequence | VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH
Obs SS   |   HHHHHHHHHHHHHH   HHHHHHHHHHHHHH
Pred SS  | HHHH  HHHHEEEEEEHH  HH HHHHHHEEEEE   EE
Clean SS | HHHHHHHHHHHEEEEEEHHHHHHHHHHHHHEEEEE

Sequence | GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKL
Obs SS   |  HHHHHHHHHHHHHHHHH    HHHHHHHHHHHH      HHHHH
Pred SS  |    HH  HHHHHHHHHHHHH    H HHHHHHHHHHH     EEH
Clean SS |    HHHHHHHHHHHHHHHHH    HHHHHHHHHHHHHH

Sequence | LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
Obs SS   | HHHHHHHHHHH      HHHHHHHHHHHHHHHHHH
Pred SS  | HHEEEEEEEEHH        HHHHHHHEEHHHHHHHH
Clean SS |  EEEEEEE             HHHHHHHHHHHHHHHHH
```

Immunoglobulin domains, IGB, all-β protein, 80.1%

```
Sequence | QSVLTQPPSASGTPGQRVTISCSGTSSNIGSSTVNWYQQLPGMAPKLLIY
Obs SS   |   EEEE          EEEEEEEE        EEEEEE        EEEE
Pred SS  | EEEEE           EEEEEEEH EE     EEEEEE        EEEE
Clean SS | EEEEE           EEEEEEE         EEEEEE        EEEE

Sequence | RDAMRPSGVPDRFSGSKSGASASLAIGGLQSEDETDYYCAAWDVSLNAYV
Obs SS   |              EEEEEE   EEEEE          EEEEEEE
Pred SS  | E                  EEEEEE   HHH     EEEEEE
Clean SS | E                  EEEEEE           EEEEEE

Sequence | FGTGTKVTVLG
Obs SS   |      EEEE
Pred SS  |      EEEEEE
Clean SS |      EEEEEE
```

Phospholipases, PLIPASE, mixed α/β structure, 76.5%

```
Sequence | SLVQFETLIMKIAGRSGLLWYSAYGCYCGWGGHGLPQDATDRCCFVHDCC
Obs SS   |  HHHHHHHHHHH                                HHHHHHHHHHH
Pred SS  | HHHHEEHHHHHHH       EEE EEEEE                HHHHHHHHHH
Clean SS | HHHHHHHHHHHHH       EEEEEEEEE                HHHHHHHHHH

Sequence | YGKATDCNPKTVSYTYSEENGEIICGGDDPCGTQICECDKAAAICFRDNI
Obs SS   | HHHH            EEE     EEE     HHHHHHHHHHHHHHHHHH
Pred SS  | HHH HH      EEEEEE                HHHHHHHHHHHHHHHH
Clean SS | HHHHHH      EEEEEE                HHHHHHHHHHHHHHHHH

Sequence | PSYDNKYWLFPPKDCREEPEPC
Obs SS   |
Pred SS  |                HHHH    H
Clean SS |
```

Fig. 3. Comparison of the three-state observed and predicted secondary structures in four different protein folds. The protein name, 3D_Ali database identifier, structural class, and prediction accuracy are first given. Helical and strand residues are indicated for both observed and predicted structures by H and E, and coil is left blank. The sequence shown is the first noted in the 3D_Ali database (Pascarella & Argos, 1992). Obs SS, Pred SS, and Clean SS are, respectively, the X-ray determined, predicted, and cleaned predicted secondary structures.

ily members, achieved a mean 8% improvement for 7 different protein families, with an overall mean prediction accuracy of 69%. They also noted that there must be a minimum 25% sequence identity among all familial sequence pairs, otherwise the prediction improvement could be diminished or even turn negative, a result emphasizing the importance of proper alignment. The sequences added here at the 35% level are clearly above the Levin et al. minimum. Rost and Sander (1993) used a trained and complex multilayer neural network as well as several filtering and sequence weighting rules to achieve an overall accuracy near 71% with standard deviation about 9% over 130 families. Due to the large computational costs in training nets, a full jackknife test could not be performed and the 71% figure is based on only sevenfold cross-validation. A 6–8% prediction improvement was observed when the multiple sequence data was employed over single-sequence prediction. In a review of neural network procedures to predict several protein structural and functional features, Hirst and Sternberg (1992) conclude that purely statistical approaches work equally well, as observed here. The Salamov and Solovyev (1995) procedure is also complex, incorporating many different secondary structural properties and again relying on neural nets.

The results presented here are consistent with those previously reported. A near 71% accuracy is achieved for 35 families after full cross-validation; the prediction standard deviation is about 9%. If all 60 families are utilized (single- and multiple-structural), close to 70% correctness ensues after cross-validation. The improvement over single sequences is 7%. It also appears that, if more and varied sequences can be added to the multiple alignments, the prediction accuracy will improve. The prediction is balanced amongst the three substructural types (Table 2).

The novelty of the residue substitution approach to secondary structure prediction encompasses various aspects. The method is fast and simple and its operational process easily comprehended and directly relatable to protein structural and evolutionary characteristics. The filtering procedures are straightforward. These properties are in contrast to those of other techniques that use more arcane and complex processes, often requiring long computations such as neural networks or inductive logic algorithms. The present method is applicable to both single sequences and multiple alignments and assures usage of all available information. As the number of known tertiary structures increases, the substitution matrices can be easily updated and optimal parameterization and verification tests quickly determined.

The exchange matrices should be useful for protein design, engineering, and modeling, where knowledge of preferred amino acid substitutions in given structural environments is typically required. Addition of more rules to the present technique should result in its improvement. Exemplary strategies are annotation of regions with several insertions and deletions, signals for likely coil segments, as well as searches in multiple alignments for hydrophobic conservation patterns characteristic of given secondary structural types.

Electronic submission

A single sequence in GCG or PIR format (Devereux et al., 1984) or a multiple alignment in 3D_Ali (Pascarella & Argos, 1992) or MSF format (see the PileUp routine in the GCG package) can be submitted via electronic mail to SSPRED@EMBL-HEIDELBERG.DE or using the World Wide Web facilities (http://www.embl-heidelberg.de/sspred/ssppred-info.html). The prediction response will always be provided through an electronic mail message. HELP information is also provided by constituting the first and only line of an electronic message with the single word HELP at the line start.

Acknowledgments

We thank our colleagues Philip Lijnzaad and Thure Etzold for computer programming succor and critical comments, as well as Frank Eisenhaber, André Juffer, and Martin Vingron for helpful discussions. Our thanks also go to the EMBL for providing space and facilities and to the EMBL Computer Group for system support. The Swiss National Science Foundation provided financial support to Perdeep Mehta for the accomplishment of this work.

References

Barton GJ, Newman RH, Freemont PF, Crumpton MJ. 1991. Amino acid sequence analysis of the annexin super-gene family of proteins. *Eur J Biochem 198*:749–760.

Bazan JF. 1990. Structural design and molecular evolution of a cytokine receptor superfamily. *Proc Natl Acad Sci USA 87*:6934–6938.

Benner S, Gerloff D. 1991. Patterns of divergence in homologous proteins and tertiary structure. A prediction of the structure of the catalytic domain of protein kinases. *Adv Enyzme Regul 31*:121–181.

Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry 13*:211–215.

Colloc'h N, Etchebest C, Thoreau E, Henrisaat B, Mornon JP. 1993. Comparison of three algorithms for the assignment of secondary structure in proteins: The advantage of a consensus assignment. *Protein Eng 6*(4):377–382.

Crawford IP, Niermann T, Kirchner K. 1987. Prediction of secondary structure by evolutionary comparison: Application to the alpha subunit of tryptophan synthetase. *Proteins Struct Funct Genet 2*:118–129.

Devereux J, Haberli P, Smithies O. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucl Acids Res 12*:387–395.

Donnelly D, Overington JP, Blundell TL. 1994. The prediction and orientation of alpha-helices from sequence alignments: The combined use of environment-dependent substitution tables, fourier transform methods and helix capping rules. *Protein Eng 7*:645–653.

Garnier J, Levin JM. 1991. The protein structure code: What is its present status. *CABIOS 7*:133–142.

Garnier J, Osguthorpe DJ, Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol 120*:97–120.

Genetics Computer Group. 1991. *Program manual for the GCG package, version 7*. Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA 53711.

Gibrat JF, Garnier J, Robson B. 1987. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J Mol Biol 198*:425–443.

Hirst JD, Sternberg MJE. 1992. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry 31*:7211–7218.

Holley HL, Karplus M. 1989. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA 86*:152–156.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Kneller DG, Cohen FE, Langridge R. 1990. Improvements in protein secondary structure prediction by an enhanced neural network. *J Mol Biol 214*:171–182.

Leng B, Buchanan BG, Nicholas HB. 1994. Protein secondary structure prediction using two-level case-based reasoning. *J Computational Biol 1*: 25–38.

Levin JM, Pascarella S, Argos P, Garnier J. 1993. Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng 6*:849–854.

Lim VI. 1974. Structural principles of the globular organization of protein

chains. A stereochemical theory of globular protein secondary structure. *J Mol Biol 88*:857–872.

Lüthy R, McLachlan AD, Eisenberg D. 1991. Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins Struct Funct Genet 10*:229–239.

Muggleton S, King RD, Sternberg MJE. 1992. Protein secondary structure prediction using logic-based machine learning. *Protein Eng 7*(5):647–657.

Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. 1992. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci 1*:216–226.

Pascarella S, Argos P. 1992. A databank merging related protein structures and sequences. *Protein Eng 5*(2):121–137.

Qian N, Sejnowski TJ. 1988. Predicting the secondary structure of globular proteins using neural networks. *J Mol Biol 202*:865–884.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Russell RB, Barton GJ. 1993. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J Mol Biol 234*: 951–957.

Russell RB, Breed J, Barton GJ. 1992. Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Lett 304*:15–20.

Salamov AA, Solovyev VV. 1995. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J Mol Biol 247*:11–15.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct Funct Genet 9*:56–68.

Zhang X, Mesirov JP, Waltz DL. 1992. Hybrid system for protein secondary structure prediction. *J Mol Biol 225*:1049–1063.

Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJE. 1987. Prediction of protein secondary structure and active site using the alignment of homologous sequences. *J Mol Biol 195*:957–961.