Protein structure prediction from predicted residue properties utilizing a digital encoding algorithm

R.J. Gilbert

British Bio-technology Limited, Cowley, Oxford OX4 5LY, and Physical Chemistry Laboratory, Oxford Centre for Molecular Science, Oxford OX1 3QZ, UK

Although many disparate methods have been applied to the problem, the accuracy of protein structural prediction still remains disappointingly low, averaging about 65% correct secondary structure assignment. A novel predictive method is presented here, which attempts to address some of the shortfalls inherent in representing a protein as a simple text-like sequence of amino acids, by deriving pattern-matching data from the predicted physical properties of a protein chain rather than from the sequence itself. A unique binary encoding algorithm is used to enable the property profiles to be correlated with known secondary structure, and hence to predict secondary structures for proteins with unknown structures. By treating the sequence in this manner, predictive accuracies averaging over 75% have been achieved.

Keywords: secondary structure prediction, pattern matching, protein structure prediction

PROBLEMS WITH CURRENT METHODS

Current methods of secondary structure prediction from primary sequence give disappointingly low predictive accuracies.[1,2] The two most popular methods, those of Chou and Fasman,[3,4] and Garnier, Osguthorpe and Robson[5] were among the earliest to be described, and very little improvement in performance has been achieved in the succeeding decade and a half, despite numerous novel approaches. The predictive accuracies, which are consistently claimed to be between 65% and 70%, are in most cases not high enough to be a firm basis for the often extremely expensive experimental investigations researchers would like to perform, and so less importance is given to structural predictions compared to structural information derived by other means.

There are several possible reasons for the poor performance of current predictive techniques. In nearly all of the methods, it is assumed that the secondary structure which a given residue actually adopts in the folded protein is a direct result of the local primary sequence, and little consideration is given to the constraints imposed by the particular tertiary fold adopted by the molecule. For example, a section of the peptide chain that must cross from one side of the protein to the other in order to comply with the topology of the tertiary fold, but that contains only a few residues, may be forced into an extended structure, even though its ''inherent'' conformation may be something more compact. In fact, short peptides with identical sequences can be observed in a wide variety of conformations in the protein structure database. Therefore, an improvement might be achieved if some prediction of the likely global position of each residue in the folded protein is incorporated into the secondary structure prediction algorithm.

Many of the existing predictive algorithms are based on a set of empirical rules, which may or may not be determined automatically. The rigid application of a set of rules that do not completely describe the mechanism of protein folding will in all probability lead to incorrectly predicted conformations for many sequences that do not exactly meet the defined criteria. Rule-based predictive algorithms that do not incorporate all of the information provided from available protein structures are therefore liable to misassign conformations when presented with sequences that do not precisely fit the defined rules. Obviously, one way to overcome this limitation is to completely define the rules governing protein folding, and in time this may well be achieved. If, as at the present time, these rules are not completely characterized, an improvement in predictive performance could be generated by incorporating all of the available information from known structures, and not disregarding any possibly relevant data by the application of a limited set of rules.

Predictions that rely on matching patterns in protein sequences may be limited in a similar way to predictions based on folding rules. The likelihood of a short window of residues from the sequence of a protein with an unknown structure matching precisely with one of the identified secondary structure-determining patterns is fairly low, and so secondary structure information can only be predicted for a relatively small fraction of all the possible windows. Also,

Address reprint requests to Dr. Gilbert at British Bio-technology Limited, Watlington Road, Cowley, Oxford OX4 5LY, UK.
Received 13 July 1991; accepted 3 September 1991

in most of these algorithms, only a small fraction of all the possible sequence patterns are recognized as being associated with a particular secondary structure class. In many algorithms, patterns that are only weakly indicative of a particular secondary structure class do not contribute to the final prediction. Thus, as with rule-based predictions, there is likely to be an incomplete set of ''axioms'' used to describe the relationship between sequence and structure.

Predictive methods based on a statistical approach may be limited to a maximum achievable accuracy of less than 100% by the less than ideal probabilistic nature of the structural data. As an example, consider the case of valine, which shows the most marked preference of any residue for any particular secondary conformation, with a propensity to form β-strands P_β of 1.7, when the statistically expected value would be 1.0 (using the Chou and Fasman propensities[3]). About 28% of residues in known structures are in β-strands, so 47.6% (1.7 × 28) of valines, less than half, are actually observed in the ''preferred'' conformation. This means that the secondary structure for valine residues would be incorrectly assigned with a better than even chance for subsequent predictions. Of course, most statistical methods take into account the sequential context of each residue, which should improve the predictive accuracy. However, it is by no means evident that accuracies approaching 100% are achievable with such low correlation levels between residue type and conformational propensity. This leads to the suggestion that there may be alternative characteristics of a peptide chain, other than the simple ''alphabetic'' sequence, that could form the basis of an improved method for secondary structure prediction.

PREDICTED PROPERTIES AS A BASIS FOR STRUCTURE PREDICTION

Several physical and chemical properties can be predicted for a peptide chain from its primary sequence, with a good correlation to the observed features in proteins having known tertiary structures. The various hydropathy scales that have been developed, based on data gathered by a variety of different experimental techniques, can all be used to answer what is essentially the same question: Is residue x likely to be internalized into the core of the folded protein? The correlation between these properties also agrees well with the generally accepted rules of thumb that have been used to describe features of folded proteins, e.g., ''hydrophobic residues form the core of the protein,'' ''surface regions are often flexible,'' and ''antigenic regions are on the surface.''

By displaying a computer-generated representation of a protein structure with each atom colored according to a scale based on a predicted peptide property, such as backbone flexibility, it is possible to assess visually the correlation between the prediction and the observed location for each residue in the protein (although, of course, the property prediction does not provide a three-dimensional (3D) structural prediction). Not only are surface and core residues easily distinguishable, but residues that show partial accessibility to the solvent have intermediate predicted flexibilities. Similar results are achievable using most of the alternative predicted properties, such as hydropathy, solvent accessibility or antigenicity.

The good correlation between the predicted properties and global location of a residue in the folded protein provides a means to link sequential data to structural features with confidence. The primary sequence gives rise to a predicted property profile that correlates well with the accessibility of the residue in the folded protein. Thus, the property profile can be used to divorce the tertiary structure from the specific details of the primary sequence, and so may allow a greater latitude in identifying sequences that give rise to similar 3D structures.

Because of the topological constraints enforced by a particular tertiary fold, proteins that share a fold have similar predicted property profiles. Although these profiles may differ, for example when the sizes of loops connecting secondary structure elements are variable, they should all have equivalent hydrophobic and hydrophilic regions in the same order and general location within the sequence. Too great a variation from the standard associated with a particular tertiary fold would imply that the protein had a significantly different tertiary structure to the known structures with which it is being compared. Thus, the overall features of the property profile calculated from the primary sequence could well lead to the identification of a particular tertiary fold for a protein of unknown structure, and local profile features may enable the identification of specific 3D structures in the folded protein, including secondary structure elements.

Significant sequence conservation is not necessary for two protein sequences to give rise to similar property profiles. In most cases, the property value depends not only on the nature of each individual residue, but also on the natures of the flanking residues, and so a variety of short sequences may give rise to a similar value of the predicted property at a particular location within the sequence. Thus, in the cases where property profiles (and therefore tertiary folds) are similar between two proteins, individual residues at equivalent locations in the structures may vary, as long as the average properties of the local region are not disturbed too much.

A closer examination of the relationship between predicted properties and residue locations shows that not only does the property prediction discriminate between internal and external regions at a residue level, but also between the buried and accessible sides of many secondary structure elements. For example, the amphipathic nature of many α-helices can easily be discerned visually using color-coded structural models. There is a recognizable pattern in the predicted properties for the residues in this region, with peaks and troughs occurring every three to four residues, reflecting the periodic nature of the α-helix structure (a similar observation for hydrophobicity was made by Eisenberg et al.[6]). Such patterns are, of course, the basis of pattern-matching secondary structure prediction methods, such as that of Lim,[7,8] which group residues according to their hydropathy, and identify patterns by string-matching techniques applied to the sequence characters. However, it is a much more difficult task to identify and match patterns in the aperiodic data of a property prediction, which form an essentially random line in displays like hydropathy plots. For this reason, it was necessary to develop an efficient algorithm to facilitate pattern-matching with predicted property profiles.

PATTERN RECOGNITION IN APERIODIC DATA

There are several statistical methods to determine if two sets of aperiodic data are similar. However for proteins, where it is necessary to quantify the degree of similarity between both global and local regions of the sequences, it is best to split the sequences into short overlapping peptide windows and to calculate the similarity between these, rather than between the entire sequences as a whole.

Figure 1 shows how a sequence can be divided into short overlapping windows that enable comparisons to be made between local sequence regions in isolation. A property profile can be predicted for the entire sequence, and split into windows in a similar fashion. It is then possible to compare the profile segments associated with each window in order to detect local regions with similarly shaped property profiles.

The comparison between the profile segments can be performed in a variety of ways, depending on how the information is to be used. For example, a simple average root-mean-squared (RMS) difference between the property values of equivalent residues in the windows can be used as a homology score in automatic alignment procedures. However, for secondary structure prediction purposes, it is more useful to be able to classify the profile segment in some way, thus allowing observed secondary structure propensities to be assigned to particular profile segment classes. Such a classification needs to be able to assign similar, but not necessarily identical, profile segments to the same class, as well as having the capacity to distinguish all of the possible profile shapes that could be encountered in proteins.

One way of classifying profile segments which meets the above criteria is to use a form of digital encoding, or *hashing*, to generate a single index number based on the shape of the profile segment. If the hashing technique allows enough variability in the specific shape of the profile segment that gives rise to the same index number, it provides a simple way by which similar, but not identical, sequences can be correlated using predicted structural information, as opposed to using the primary sequences directly.

The binary encoding technique outlined in Figure 2 calculates a unique index number derived from the shape of the predicted profile segment. Segments that give rise to the same index number fall into the same class, and are said to *match*. The number of zones determines the *granularity* of the hashing algorithm, and so defines the amount of variability allowed in the shape of the profile segments in order to assign them to the same indexed class. A greater number of zones puts a tighter constraint on the shape of the profiles that would generate each index number, and results in fewer profile segments in each class. However, too few zones would not give the specificity required to enable sufficient profile segment classes to be useful for secondary structure prediction purposes. For an eight-residue window with a granularity of 2 bits per zone (i.e., four zones), it is possible to distinguish 65536 profile segment classes. These segment classes are equivalent to the sequence patterns recognized by current widely used pattern-matching secondary structure prediction algorithms, and so it is possible to develop a new predictive algorithm using pattern-matching in the predicted structural properties of a sequence, rather than using the linear sequence directly.

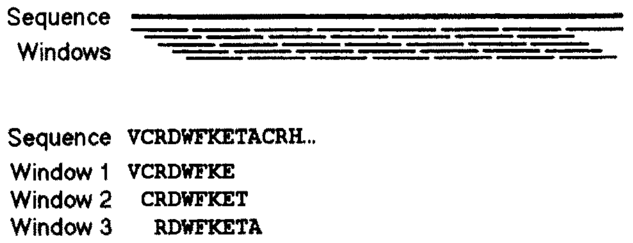


Figure 1. To enable the detection of local similarities, the sequence is split into many overlapping windows. The window length should be about the same size as the structural feature of interest. A window of eight residues is about the same size as most β-strands and the minimum stable α-helix (two turns).

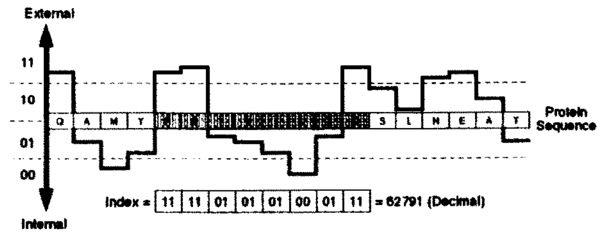


Figure 2. The predicted property profile for a given protein sequence (in the central bar) is shown as a bold line. The range of possible values for the predicted property can be split into zones, each of which is assigned a unique binary number (shown on the vertical axis). The unique index number for the profile segment relating to the sequence window (shown as shaded residues) is then calculated as a binary number derived from each of the zones in which the property profile falls. The values for the cut-offs between the zones are calculated to ensure an equal number of residues in each zone, and depend on the profiles predicted for a suitable "training set" of proteins.

THE NEW PREDICTION ALGORITHM

The development of a suitable algorithm for matching patterns in the shapes of short segments of predicted structural property profiles enables a new method of secondary structure prediction to be proposed. This algorithm aims to address some of the problems inherent in the alternative algorithms that are currently most popular. It incorporates predictions of the global location of each residue within the folded protein, does not disregard any information provided by the training set of protein structures, does not rely on possibly incomplete axiomatic folding rules, and attempts to reduce the problems associated with a low statistical correlation between residue type and secondary structure propensity by allowing a greater variability in the sequences that give rise to the patterns used to generate the predicted conformational probabilities.

Calculation of the Property Index Profile

Several predicted structural properties are suitable for this algorithm. Hydropathy,[9] solvent accessibility,[10] backbone flexibility,[11] antigenicity[12] and even side-chain mutability[13] all give acceptable results when used with this method (Table 1 lists these scales). To avoid possible aberrations, however, a consensus property index (PI) is calculated as an average of these five predicted properties.

The scales used to calculate the PI values are given in Table 1. The hydropathy scale is an aggregate of a vapor–aqueous phase partitioning scale and two statistical scales, and is widely thought to be the best general-purpose scale for this purpose. The antigenicity scale is based on a polar–apolar solvent partitioning scale, and has been found to be reliable in predicting residues likely to form antigenic sites. The accessibility scale is based on the fraction of residues of a particular type that are at least 95% buried in a set of 12 observed protein structures. The mutability scale is based on the observed frequency with which each residue type changes in a set of homologous proteins from different organisms. Residues on the surface of proteins are more likely to mutate than those in the core, as any changes are less likely to disrupt the correct folding of the protein. Finally, the flexibility scale is based on observed temperature factors (B values), which indicate atom motion in crystal structures.

In general, the predicted property value P_n for residue R_n is calculated as an average of a window of residues centered around R_n, in order to take account of the sequence context of R_n. Thus, P_n is not necessarily the same as V_n, the scale value for a residue of type R_n. Conversely, a particular value of P_n can be produced by a variety of different sequences.

In all cases except backbone flexibility, which was calculated using the method suggested by the authors, a slight improvement was introduced into the averaging method used to derive the property values. Where the original algorithm would assign the property value of residue n by averaging across a window of residues, say from residues $(n - 3)$ to $(n + 3)$, it has been found preferable to use a weighted average, with the weighting inversely proportional to the sequential distance from residue n (Figure 3). In this way, residues that are more likely to be close together spatially have more of an influence on the property values assigned to each other. For a seven-residue window, if the weighting of residue $(n + m)$ is $W_{(n+m)}$, and the property scale value for a residue of type $(n + m)$ is $V_{(n+m)}$, then the property value for residue n, P_n is calculated as

$$P_n = \sum_{m=-3}^{m=3} (W_{(n+m)} \times V_{(n+m)})$$

The calculation of backbone flexibility incorporates a form of smoothing by having three different values for each residue type, selected according to whether residue n has 0, 1 or 2 "rigid" neighbors (defined as residues of type ALA, LEU, HIS, VAL, TYR, ILE, PHE, CYS, TRP or MET). It was considered inappropriate to use an additional averaging filter to smooth the flexibility profile.

In addition to the averaging window refinement, all the property scales are recalibrated to the same numerical range (0–1000), simply by assigning 0 to the residue type with

Table 1. Predicted properties used to calculate the property index values are *a* hydropathy,[9] *b* antigenicity,[12] *c* solvent accessibility,[10] *d* side chain mutability[13] and *e* flexibility.[11] Properties *a–d* are calculated by averaging over a window of 7 residues. Property *e* is assigned according to whether the residue has 0, 1 or 2 "rigid" neighbors (indicated by the subscripts, and defined as ALA, LEU, HIS, VAL, TYR, ILE, PHE, CYS, TRP or MET), and so has three value scales

	a	*b*	*c*	*d*	e_0	e_1	e_2
ALA	1.8	−0.5	0.38	75	1.041	0.946	0.892
CYS	2.5	−1.0	0.47	15	0.960	0.878	0.925
ASP	−3.5	3.0	0.15	79	1.033	1.089	0.932
GLU	−3.5	3.0	0.18	76	1.094	1.036	0.933
PHE	2.8	−2.5	0.50	31	0.930	0.912	0.914
GLY	−0.4	0.0	0.36	37	1.142	1.042	0.923
HIS	−3.2	−0.5	0.17	49	0.982	0.952	0.894
ILE	4.5	−1.8	0.60	72	1.002	0.892	0.872
LYS	−3.9	3.0	0.03	42	1.093	1.082	1.057
LEU	3.8	−1.8	0.45	30	0.967	0.961	0.921
MET	1.9	−1.3	0.40	70	0.947	0.862	0.804
ANS	−3.5	0.2	0.12	100	1.117	1.006	0.930
PRO	−1.6	0.0	0.18	42	1.055	1.085	0.932
GLN	−3.5	0.2	0.07	69	1.165	1.028	0.885
ARG	−4.5	3.0	0.01	49	1.038	1.028	0.901
SER	−0.8	0.3	0.22	90	1.169	1.048	0.923
THR	−0.7	−0.4	0.23	72	1.073	1.051	0.934
VAL	4.2	−1.5	0.54	55	0.982	0.927	0.913
TRP	−0.9	−3.4	0.27	13	0.925	0.917	0.803
TYR	−1.3	−2.3	0.15	31	0.961	0.930	0.837

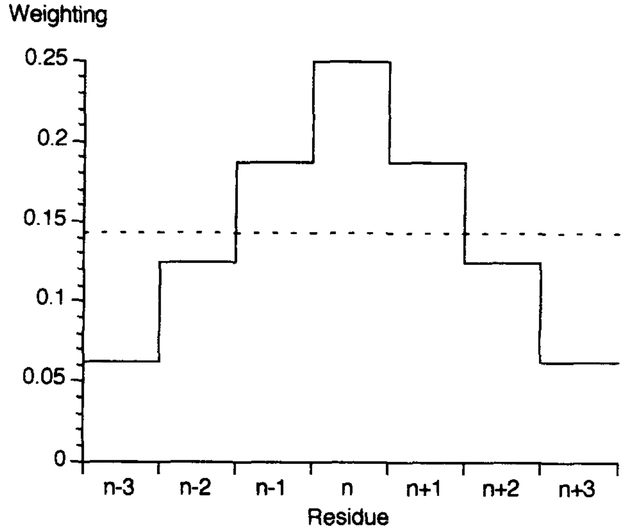


Figure 3. When calculating an average predicted property value from a window of residues, a triangular averaging profile (solid line) is used in preference to the rectangular profile (dotted line) suggested by the authors of the original algorithms. It was considered unrealistic that the residues some distance from the one to which the value is assigned have the same influence as those neighboring it. The profile shown is that for a seven-residue window.

the lowest property value, 1000 to that with the highest, and scaling the others appropriately. In some cases, this also involves an inversion of the scale (e.g., in the case of hydropathy, which is inversely related to properties like flexibility and antigenicity). This recalibration is necessary to enable property values derived by different experimental methods, and with different physical units, to be compared directly. The property index profile is then calculated as a simple numerical average of the five predicted property values of each residue in turn. Residues with PI values less than 500 are more likely to be in the core of the protein, and those with values greater than 500 are more likely to be on the surface.

The Pattern Database

In order to be able to predict secondary structure preferences based on pattern-matching using calculated PI profiles, it is necessary to construct a database relating observed secondary structure propensities[14] to each profile segment class, using a training set of proteins with known structures. The secondary structure of a protein with an unknown tertiary structure can then be predicted by calculating its PI profile, splitting this into segments, and matching the segments to entries in the database, noting the frequencies with which particular secondary structure classes are associated with each segment class.

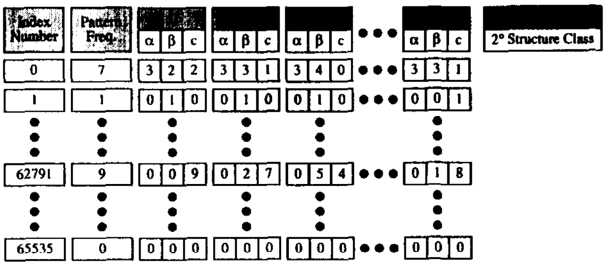
An index into the database is provided by the digital hashing algorithm outlined in the previous section. This encoding method allows similar, but not necessarily identical profile segments to share the same record in the database, and so the stored secondary structure information is common to a set of related profile segments, rather than to just one segment.

Each indexed record in the pattern database (Figure 4) contains a field that stores the number of times its associated pattern was found in the protein data set, and fields that store the secondary structure frequencies observed for each residue in the window. For a window of eight residues, and an index number calculated using 2 bits per zone, there are 65536 distinguishable patterns, and hence 65536 records in the database. If this window size is used to generate a 3-class prediction, the pattern database array requires just over 3 MB (megabytes) of contiguous storage space, and so the method is unsuitable for most personal computers, yet can easily be accommodated on most workstations and minicomputers.

Calculation of Secondary Structure Propensities

The probability of each residue in a given pattern having a particular secondary structure is found by simply dividing the secondary structure frequency by the pattern frequency, summed for every window containing the residue in question. For example, using the data in Figure 4, the probability of residue 2 in pattern 62791 forming a β-strand is 0.222. Secondary structure propensities, roughly equivalent to the P_{ij} values of Chou and Fasman[3] and Levitt[15] can be calculated by dividing the secondary structure probability by the observed proportion of residues adopting that secondary structure in the training set of proteins.

The propensities calculated in this way differ from those of the earlier methods, however, because they depend not only on the residue type (or, more accurately, on the property value at that location in the sequence), but also on the sequence context. Most residues in the sequence are matched against the database in several contexts, as they occur in more than one of the short sequence windows used to generate the profile segments. If the residue in the previous example occurred in just two patterns, with numbers 0 and 62791, then the secondary structure probabilities would be calculated from the summed frequencies of the two database



Index Number	Pattern Freq.	α	β	c	α	β	c	α	β	c		α	β	c	2° Structure Class
0	7	3	2	2	3	3	1	3	4	0	•••	3	3	1	
1	1	0	1	0	0	1	0	0	1	0	•••	0	0	1	
⋮	⋮	⋮			⋮			⋮					⋮		
62791	9	0	0	9	0	2	7	0	5	4	•••	0	1	8	
⋮	⋮	⋮			⋮			⋮					⋮		
65535	0	0	0	0	0	0	0	0	0	0	•••	0	0	0	

Figure 4. The profile pattern database may be stored as a single array, accessed by the index number derived from the shape of the profile segment. The database stores the observed frequency of each segment class, as well as the frequencies of the secondary structures observed for each residue within the sequence window used to generate the profile segment. In the example above, the data is for an eight-residue window, 3-class (α-helix, β-strand, random-coil) prediction.

entries for patterns 0 and 62791. The residue would then have a β-strand probability of $(2 + 3)/(9 + 7) = 0.3125$.

It is quite conceivable that a new sequence may differ substantially from the proteins used to generate the pattern database, and that it may produce profile segments with shapes that have not been previously encountered. In such cases, no specific secondary structure information is available for the prediction, and so the structure probabilities are set to the observed structure frequencies in the training set of proteins. Secondary structure predictions that are based on a relatively large number of unique profile segments should be treated with even more caution than is usual, and some indication that this is the case should be provided to the user of the algorithm.

Presentation of the Results

Secondary structure predictions for proteins of unknown structure are usually given as unambiguous single-class assignments for each residue. Often, however, two or more secondary structure classes are predicted for the same residue with approximately equal probabilities, and so a single unambiguous prediction may give a misleading impression of the confidence in the prediction for each residue. In addition, the human brain is much more adept than automated algorithms at interpreting subtle differences in the predicted structure propensities. For this reason, it may be preferable for the output from the predictive algorithm to be presented as a series of shaded bars (Figure 5), which allows a rapid visual determination of the relative likelihood of each secondary structure class. This form of representation is believed to be more digestible than the two-dimensional graphical representation that is more commonly used.

For the less-experienced user of the algorithm, who may be uncertain as to how to interpret the shaded-bar representation of the secondary structure propensities, an acceptable automatic single-class assignment can be derived from the secondary structure probabilities by applying the following rules (which have been parametized for a 3-class prediction):

(1) For residue n, if any of the 3 secondary structure propensities is appreciably greater than the others (i.e., more than 0.2 greater than the next highest propensity) or the propensity for a structural class P_n is 0.2 greater than either $P_{(n-1)}$ or $P_{(n+1)}$, the residue is assigned to that structure class; otherwise rule 2 is applied.

(2) If two or more secondary structure probabilities are within 0.2 of each other, then the propensities are recalculated using a 3-residue triangular averaging window (i.e., the new $P_n = 0.25\ (P_{(n-1)}) + 0.5\ (P_n) + 0.25\ (P_{(n+1)})$, where P is the structural class propensity). Rule 1 is then reapplied. If the structure still cannot be assigned, the propensities are recalculated from the original ones, using a window 2 residues larger. Rule 2 is repeated until the window size reaches the sequence length, at which point it is decided that the secondary structure class cannot be determined, so the residue is assigned as "coil."

Figure 5 also shows a unique assignment calculated by this method, allowing a comparison of the two forms of prediction display. Notice, for example, that the shaded bars clearly show that the second and third α-helices are predicted to occur with a much higher probability than the first, but that there is no indication of this in the automatically derived structure assignment.

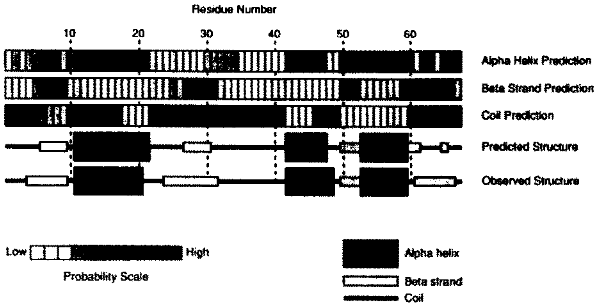


Figure 5. The secondary structure prediction for human insulin-like growth factor 2 [IGF2]. The sequence is represented as a horizontal bar, with the residues intensity-coded according to the predicted probability of each secondary structure class. The structure observed in a 3D model of IGF2 (from the Brookhaven entry 1gf2) is shown for comparison, as is an automatically derived unique class prediction. The Brookhaven entries for the proteins used to generate the pattern database were as follows: 1acx, 1cac, 1ccr, 1crn, 1eca, 1etu, 1fdx, 1fxl, 1hip, 1hmq, 1hoe, 1gcr, 1lzl, 1mbd, 1mlt, 1nxb, 1ppt, 1rn3, 1tim, 1tnf, 1ubq, 2act, 2alp, 2app, 2aza, 2b5c, 2cab, 2lbp, 2lhb, 2mev, 2ovo, 2pab, 2sod, 2utg, 3bp2, 3cln, 3cna, 3cpv, 3fab, 3grs, 3wrp, 4ilb, 5cpa, 5rxn, 5tnc, 6pcy, 6pti, 7tln, 8dfr and 8ldh. None of these proteins contain the insulin-like fold of IGF2.

Implementation of the Algorithm

The algorithm for secondary structure prediction outlined above has been coded as a computer program using the C programming language, although it should be reasonably easy to implement in other languages. The only limitation to the portability of the program is that the computer hardware should be able to support large contiguous arrays of the size needed to contain the pattern database, typically around 3 MB. Unfortunately, most microcomputers are unable to support arrays of this size, and so the program is, at present, limited to mainframes, minicomputers and workstations. The program has successfully run, without modification, on DEC VAX minicomputers and Silicon Graphics IRIS workstations. A cut-down version of the program, which uses a window length of 5 residues rather than the optimum of 8 (and so drastically reduces the size of the pattern database, to 32 KB), has been implemented on the Apple Macintosh range of microcomputers, although this version shows much poorer predictive accuracies than the full implementation.

PERFORMANCE OF THE ALGORITHM

The performance of the prediction method outlined above was tested using a set of 50 proteins with known structures (listed in Table 2). The proteins selected for this training

Table 2. **The 50 protein structures used to test the performance of the secondary structure prediction algorithm. The proteins are all dissimilar to each other, with different specific tertiary folds**

Brookhaven code	Protein name	Brookhaven code	Protein name
1acx	Actinoxanthin	2b5c	Cytochrome B5
1ccr	Cytochrome C	2cab	Carbonic anhydrase B
1crn	Crambin	2hla	Histocompatability antigen
1eca	Erythrocruorin	2lbp	Leucine binding protein
1etu	Elongation factor tu	2mev	Mengo virus
1fdx	Ferredoxin	2ovo	Ovomucoid third domain
1fx1	Flavodoxin	2pab	Prealbumin
1hip	High potential iron protein	2sod	Superoxide dismutase
1hmq	Hemerythrin	2utg	Uteroglobin
1hoe	Alpha-amylase inhibitor	3bp2	Phospholipase A2
1gcn	Glucagon	3cln	Calmodulin
1gcr	Gamma-II crystallin	3cna	Concanavalin A
1lz1	Lysozyme	3cpv	Ca-binding parvalbumin
1mbd	Myoglobin	3fab	Immunoglobin fab
1mlt	Melittin	3grs	Glutathione reductase
1nxb	Neurotoxin B	3wrp	Trp repressor
1ppt	Pancreatic polypeptide	4i1b	Interleukin 1B
1rn3	Ribonuclease A	5cpa	Carboxypeptidase A
1tim	Triose phosphate isomerase	5rxn	Rubredoxin
1tnf	Tumour necrosis factor	5tnc	Troponin C
1ubq	Ubiquitin	6pcy	Plastocyanin
2act	Actinidin	6pti	Trypsin inhibitor
2alp	Alpha-lytic protease	7tin	Thermolysin
2app	Penicillopepsin	8dfr	Dihydrofolate reductase
2aza	Azurin	8ldh	Lactate dehydrogenase

set cover a wide range of sizes and tertiary folds, and are not significantly homologous to each other.

To assess the predictive accuracy of the algorithm, the secondary structure of each protein was predicted using a pattern database generated from the other 49 proteins in the test set, and the predicted secondary structure was then compared with the observed secondary structure in the crystallographic model. As the proteins are nonhomologous, this procedure is analogous to predicting the structure of 50 proteins with novel tertiary folds.

The predictive performance Q is given by the equation

$$Q = \frac{N_{\text{Correct}}}{N_{\text{Total}}} \times 100\%$$

where N_{Correct} is the number of residues whose secondary structure conformation is correctly predicted, and N_{Total} is the total number of residues in the protein. Most of the current widely used prediction methods are claimed to achieve average Q values of between 65% and 70%.[2] The Q values for the predictions on the 50 proteins listed in Table 2 are given in Table 3, and average 77% for this set of proteins.

For most of the test proteins, this algorithm gives similar or more accurate results than other predictive algorithms. However, the accuracy does appear to depend on the characteristics of the secondary structure elements in the protein. This method is much more accurate at predicting the class of amphipathic or hydrophilic secondary structure elements than of purely hydrophobic or hydrophilic ones (see Table 4). Since the secondary structure elements of smaller proteins are almost entirely amphipathic in nature, predictive accuracy is higher for shorter proteins. The most likely explanation for this bias towards amphipathic elements is that these are the ones which are associated with profile segment shapes particularly specific to the secondary structure class. The amphipathic secondary structure elements, having easily recognized patterns of alternating hydrophobic and hydrophilic residues, form the basis of several other pattern-matching prediction methods.

SUMMARY

The prediction of secondary structure by matching patterns taken from structural properties predicted from the primary sequence, rather than from the sequence itself, is a logical extension to many of the current widely used secondary structure prediction algorithms, and does appear to give an improvement in the predictive accuracy for small, soluble, globular proteins. The unique binary encoding algorithm for indexing entries into the parent database is both fast and accurate, and may be applicable to other problems where a similar method for matching patterns in nonperiodic or unsmooth data is required.

This algorithm, using pattern-matching with predicted structural properties rather than the amino acid sequence

Table 3. The % prediction accuracy (or Q value) was calculated for the 50 structures listed in Table 2. Using this set of proteins, the average Q value is approximately 77%. Note, however, that the accuracies range from 57% to 98%, and so there is still a chance that the secondary structure prediction could be up to 40% incorrect. Many of the best accuracies are achieved with the smaller proteins

Brookhaven code	% Accuracy	Brookhaven code	% Accuracy	Brookhaven code	% Accuracy	Brookhaven code	% Accuracy	Brookhaven code	% Accuracy
1acx	87	1gcn	92	1ubq	69	2ovo	60	3wrp	80
1ccr	73	1gcr	79	2act	69	2pab	85	4i1b	69
1cm	80	1lz1	73	2alp	80	2sod	91	5cpa	68
1eca	64	1mbd	69	2app	85	2utg	76	5rxn	90
1etu	70	1mlt	57	2aza	91	3bp2	81	5tnc	73
1fdx	77	1nxb	98	2b5c	72	3cln	73	6pcy	90
1fx1	74	1ppt	63	2cab	82	3cna	81	6pti	94
1hip	83	1rn3	92	2hla	61	3cpv	76	7tln	74
1hmq	62	1tim	64	2lbp	70	3fab	89	8dfr	76
1hoe	88	1tnf	90	2mev	61	3grs	74	8ldh	74

Table 4. Both α-helix and β-strand secondary structure elements in the training set of 50 proteins listed in Table 2 were classified visually as being mostly hydrophobic, amphipathic or hydrophilic, and the prediction accuracies were calculated for each class. As can be seen from the results, the highest accuracies are achieved with amphipathic secondary structure elements, presumably because these are associated with particularly specific profile segment shapes

2° Structure type	% Accuracy
Hydrophobic α-helix	49
Amphipathic α-helix	87
Hydrophilic α-helix	63
Hydrophobic β-strand	68
Amphipathic β-strand	85
Hydrophilic β-strand	54

itself, suggests that there is a stronger correlation between the property profile shape and secondary structure class than there is between sequence and secondary structure class, at least for amphipathic structures. The prediction accuracies for the other types of secondary structure are about the same as with the alternative prediction methods. The relative abundance of amphipathic structural elements in water-soluble globular proteins means that for this type of protein, the property-based algorithm has a better overall performance than sequence-based predictions.

REFERENCES

1 Nishikawa, K. and Ooi, T. Amino acid sequence homology applied to the prediction of protein structures, and joint prediction with existing methods. *Biochem. Biophys. Acta* 1986, **871**, 45–54

2 Kabsch, W. and Sander, C. How good are predictions of protein secondary structure? *FEBS Lett.* 1983, **155**, 179–182

3 Chou, P.Y. and Fasman, G.D. Empirical predictions of protein conformations. *Annu. Rev. Biochem.* 1978, **47**, 251–276

4 Chou, P.Y. and Fasman, G.D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* 1978, **47**, 45–148

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

6 Eisenberg, D. et al. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. U.S.A.* 1984, **81**, 140–144

7 Lim, V.I. Structural principles of the globular organisation of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.* 1974, **88**, 857–872

8 Lim, V.I. Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.* 1974, **88**, 873–894

9 Kyte, J. and Doolittle, R.F. A simple method for displaying the hydropathic nature of a protein. *J. Mol. Biol.* 1982, **157**, 105–132

10 Chothia, C. The nature of accessible and buried surfaces in proteins. *J. Mol. Biol.* 1976, **105**, 1–14

11 Karplus, P.A. and Schultz, G.E. Prediction of chain flexibility in proteins. *Naturwissenschaften* 1985, **72**, 212–213

12 Hopp, T.P. and Woods, K.R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* 1981, **78**, 3824–3828

13 *Atlas of Protein Sequence and Structure* (M.O. Dayhoff, Ed.) National Biomedical Research Foundation (1978) vol. 5, suppl. 3.

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

15 Levitt, M. Conformational preferences of amino acids in globular proteins. *Biochemistry* 1978, **17**, 4277–4285