PROTEINS: Structure, Function, and Genetics 27:329–335 (1997)

Seventy-Five Percent Accuracy in Protein Secondary Structure Prediction

Dmitrij Frishman* and Patrick Argos
European Molecular Biology Laboratory, Postfach 102209, 69012 Heidelberg, Germany

ABSTRACT **In this study we present an accurate secondary structure prediction procedure by using a query and related sequences. The most novel aspect of our approach is its reliance on local pairwise alignment of the sequence to be predicted with each related sequence rather than utilization of a multiple alignment. The residue-by-residue accuracy of the method is 75% in three structural states after jack-knife tests. The gain in prediction accuracy compared with the existing techniques, which are at best 72%, is achieved by secondary structure propensities based on both local and long-range effects, utilization of similar sequence information in the from of carefully selected pairwise alignment fragments, and reliance on a large collection of known protein primary structures. The method is especially appropriate for large-scale sequence analysis efforts such as genome characterization, where precise and significant multiple sequence alignments are not available or achievable. Proteins 27:329–335, 1997.**

Key words: protein structure; protein sequence analysis; hydrogen bonds; sequence alignment

INTRODUCTION

A major improvement in protein secondary structure prediction accuracy from sequence alone resulted from the exploration of additional information contained in often numerous sequences homologous to that predicted. This was made possible by the unprecedented speedup in nucleotide sequencing capabilities, resulting in a near 15-fold increase in the number of sequences over the last decade. For 85% of the known protein sequences, at least one homologous sequence is known (as inferred from the ProDom database[1]), making secondary structure prediction from multiple sequences realizable.

Although the benefits of multiple sequences for secondary structure prediction were noted long ago,[2] most of the consistent methodological work on this subject was made over the last decade[3-5] with the best available programs surpassing the 70% accuracy level[6-9]; reviewed recently in ref. 10. Many recent secondary structure predictions are based on sequence families.[11-14] It is generally accepted that the utilization of multiply aligned sequences brings about a gain in prediction accuracy of 6–8%, relative to the single sequence case.[6,15,16]

The framework of current approaches includes automatic multiple alignment of related sequences and derivation of amino acid residue variation patterns at individual alignment positions or within fixed-length sequence spans of the multiple alignment. To generate the secondary structure prediction for the query sequence, an entire range of mathematical formalisms has been used from simple statistical rules to sophisticated machine learning algorithms.

Multiple sequence alignment remains a difficult task in molecular bioinformatics. Rigorous algorithms based on dynamic programming have the computational complexity of at least L^n (where L is the sequence length and n is the number of sequences) and can be impractical if many or long sequences are involved. Although several shortcuts based on incorporation of biologically relevant information to limit the search space have been suggested,[17-20] the currently used approaches almost always rely on hierarchical clustering of the sequences by pairwise alignment beginning with the most closely related pairs, so that the overall alignment quality depends largely on the pairwise similarity scores of different sequences along the evolutionary tree.[21-27] Once aligned, two sequences preserve their register and gaps introduced at earlier stages of the alignment procedure are never reconsidered, following the dictate "once a gap, always a gap."[23] Such procedures represent a compromise between pairwise and overall alignment quality.

Most of the gaps introduced in the alignment can be irrelevant for secondary structure prediction, which focuses on the relationship of the sequence to be predicted with all the other sequences and not on all pairwise relationships. Very distantly related proteins often share only short functional and structural sequence patterns, making attempts to multiply, align, and utilize the entire sequences futile. Important structural elements present in some fam-

*Correspondence to: Dr. Dmitrij Frishman, Max-Plank Institute for Biochemistry, Martinsried Institute for Protein Sequences, 82152 Martinsried, Germany.
Received 20 June 1996; accepted 30 September 1996.

ily members can be matched against gaps in other sequences,[28] which will either mislead the recognition procedure or leave certain alignment regions unassigned.[29] Sequences completely foreign to the given family can also be recruited by database searching techniques with inappropriate or ambivalent parametric setting, further reducing the information content of the multiple alignment. Recent evidence shows that misaligned sequences can reduce prediction accuracy to a level lower than that achieved with mere single sequence information.[15,16]

In this study we propose an alternative way to use the additional information contained in a set of related sequences. A careful pairwise alignment of the query sequence with all related sequences is performed. Only significant alignment fragments are subsequently considered. The secondary structure propensities of the auxiliary-related sequences are combined with (projected onto) those of the base sequence and weighted according to their degree of similarity.

METHODS

Protein Structure Data Sets

For training, testing, and comparing our algorithm, we used the same nonredundant set of 125 globular protein tertiary structures, as listed by Rost and Sander[6] (set RS). The atomic coordinates were taken from the Protein Structure Bank (PDB).[30] For the final training we created our own set with the automated procedure of Heringa et al.[31] (set FA). The latter contained 556 protein chains determined by X-ray analysis and NMR with no more than 30% pairwise sequence identity, no sequence with length less than 50 residues, and crystallographic resolution >2.5 Å.

Generating Related Sequence Sets

For each protein with a known three-dimensional structure as used in this analysis, related protein sequences were extracted from the largest protein sequence data bank (TREMBL), which was created by T. Etzold and G. Schaefer at the European Molecular Biology Laboratory (EMBL)[32] and contains translations of all coding frames without internal stop codons in the EMBL nucleotide sequence database.[33] Searching for similar sequences was based on the improved FASTA technique[34] (version 2.0), which provides an estimate of statistical significance of the hits found based on the extreme value distribution.[35] Because the evaluation of alignment quality is incorporated in our technique at a later stage (see below), a very generous cutoff for extreme values (0.1) was used to ensure that a full sequence set is generated. Every set of sequences similar to a given sequence with known topology was made nonredundant with the procedure of Heringa et al.[31] such that no two sequences of the set had more than 95% residue identity. This step was necessary because the TREMBL database often contains identical or nearly identical sequences resulting from different sequencing projects, as well as fragments included in other database entries.

Secondary Structure Propensities

The principal step in our procedure involves generating seven secondary structural propensities (P_i, i = 1,7) for the query sequence and each sequence in the related set as described earlier for the algorithm PREDATOR, which relies only on single sequence information for secondary structure prediction.[36] Three propensities are based on long-range interactions involving potential hydrogen bonding residues in antiparallel (P_1) and parallel (P_2) β-strands as well as α-helices (P_3); three further propensities for helix (P_4), strand (P_5), and coil (P_6) rely on the similarity of the sequence segment to be predicted with those of known conformation (nearest neighbor approach[37]), and finally a statistically based turn propensity (P_7) used over a four-residue window as described by Hutchinson and Thornton.[38] These propensities rely on different concepts (hydrogen bonded pairing, sequence fragment similarity, and knowledge-based statistics) that complement each other with appropriate weighting and allow a high prediction accuracy (68%) by using single sequence information only.

Combination of the Secondary Structure Propensities of the Base Sequence With Those of Related Sequences

This section describes the primary novel element of our method. Instead of relying on protein sequences multiply aligned over their entire length, PREDATOR uses pairwise alignments of the base sequence with each sequence from the related set identified by the SIM technique of Huang and Miller.[39] SIM produces Q best nonintersecting local alignments between a pair of sequences by dynamic programming.

Let $P_i^0(l)$ be the secondary structure propensities of the sequence being predicted, where i refers to a given propensity (i = 1,7) and l is the residue site (l = 1,L^0) in a sequence of length L^0. Let $P_i^m(l)$ (i = 1,7; l = 1,L^m; m = 1,M) represent secondary structure propensities for M-related sequences with respective lengths L^m. After aligning the base sequence with the m-th similar sequence, we obtain in general Q best local nonintersecting alignments with residue percentage identity of the aligned fragments $\Omega_q^{0,m}$ (q = 1,Q) and length $S_q^{0,m}$ (Fig. 1). The percentage of identity is relative to the number of matched residue pairs where gaps are not considered, although they may appear in the SIM local alignments. The quality of every pairwise alignment with the base sequence was characterized by the pseudo-

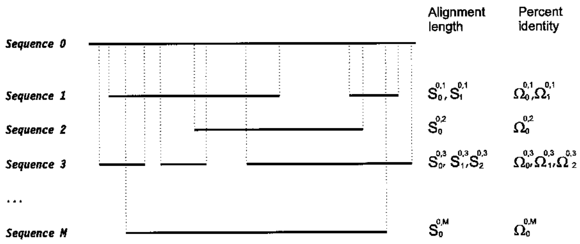

Fig. 1. Pairwise local alignments of the query sequence 0 with the related sequences m = 1,2 . . . M. Every alignment is characterized by its length $S_q^{0,m}$ and residue percentage identity $\Omega_q^{0,m}$ (q = 1,Q), where Q is the total number of local alignments between the sequences 0 and m.

information

$$I_q^{0,m} = -\Omega_q^{0,m} \ln \Omega_q^{0,m} \tag{1}$$

and the measure of Sander and Schneider[40]

$$\tilde{\Omega}_q^{0,m} = 290.15 * S_q^{0,m^{-0.562}} \tag{2}$$

which gives the minimum threshold of percentage of identical residues for a given length of residue matches necessary for true structural homology (Fig. 2).

Alignments are discarded as insignificant by applying the following selection criteria: 1) $S_q^{0,m} < 10$; 2) $I_q^{0,m} < \tilde{I}$ where $\tilde{I}$ is an empirically chosen threshold; or 3) $\Omega_q^{0,m} < \tilde{\Omega}_q^{0,m}$ for a given alignment length $S_q^{0,m}$ (Fig. 2).

The final propensity values for each residue l of the base sequence are calculated as a weighted sum of the native propensities $P_i^0(l)$ and all propensities of the residues from homologous sequences projected onto the residue l from the local pairwise alignment procedure such that

$$P_i^{0,Final}(l) = \frac{P_i^0(l) + \sum_{m=1}^{M} \begin{cases} I_q^{0,m} P_i^m(h), & \text{if a residue } h \text{ of sequence } m \text{ is projected onto residue } l \text{ of sequence 0 through local alignment q} \\ 0, & \text{otherwise} \end{cases}}{1 + \sum_{m=1}^{M} \begin{cases} I_q^{0,m}, & \text{if a residue of sequence } m \text{ is projected onto residue } l \text{ of sequence 0 through local alignment q} \\ 0, & \text{otherwise} \end{cases}}$$

Generating and Evaluating the Prediction

The rules for assigning the secondary structural type at each reside site l from the final propensities $P_i^{0,Final}(l)$ were the same as for the single sequence PREDATOR (see ref. 36 for details). If ($P_1(l) > \tau_1$ *or* $P_2(l) > \tau_2$) and $P_3(l) < \tau_3$, then predict sheet; otherwise if $P_3(l) > \tau_3$, then predict helix; otherwise predict coil. If $P_6(l) > \tau_6$, then predict coil. If $P_5(l) > \tau_5$, then predict sheet. If $P_4(l) > \tau_4$, then predict helix. If $P_7(l) > \tau_7$, then predict coil. The threshold values τ_i (i = 1,7) were determined to achieve the best possible prediction accuracy by a global optimization procedure involving multiple steps of random generation of starting threshold values in reasonable ranges with a subsequent Nedler-Mead simplex function minimization.[41] Postprocessing of the prediction consisted of eliminating α-helices of four residues and fewer in length and β-strands of two or fewer residues in length.

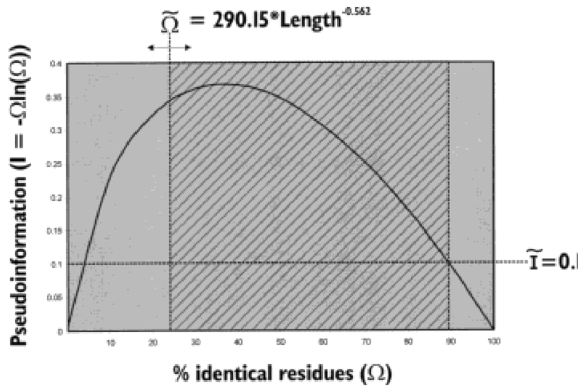

Fig. 2. Alignment quality (pseudoinformation) as a function of residue identity fraction (see Methods). The value of the pseudoinformation is used as a weight to combine the secondary structure propensities of matched residues in the related sequences with each of those in the query sequence. Note that the curve has its maximum at ~35% identity. The accepted range of identity is shown over the hatched area and is bracketed by two thresholds, $\tilde{\Omega}$ and $\tilde{I}$, where $\tilde{\Omega}$ depends on the alignment length according to the Sander and Schneider[40] formula (illustrated in the figure for an 80-residue alignment span) while constant $\tilde{I}$ was empirically found. Note that sequence segments with lower identity to the query sequence contribute more to the secondary structure prediction because of their greater information content.

To ensure the absence of a relationship between sequences in the training set used to optimize propensity thresholds and the protein sequence under prediction, we implemented a simple one-at-a-time jack-knife procedure. Each of the protein sequences with known tertiary structure was iteratively removed from the training set, all propensities recalculated, optimal thresholds found, and the resulting secondary structure prediction procedure applied to the removed sequence. Prediction accuracy was defined as the fraction of residues whose secondary structural conformation was correctly predicted in three states (helix, sheet, and coil). DSSP secondary structure assignments[42] were used in this study to compare with past efforts that were always reliant on DSSP. However, the final version of PREDATOR has an option where the user can specify one of the two target secondary structure assignment methods, DSSP or STRIDE.[43] The average accuracy over all such jack-knife tests was taken to indicate the overall prediction accuracy.

RESULTS AND DISCUSSION

This study concentrates on a new and optimal way to use and extract similar sequence information for

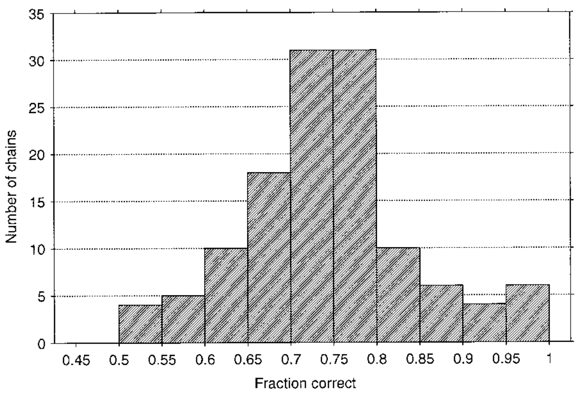

Fig. 3. Distribution of the PREDATOR prediction accuracy from related multiple sequences in three states over the 125 protein chains in the RS set (see Methods).

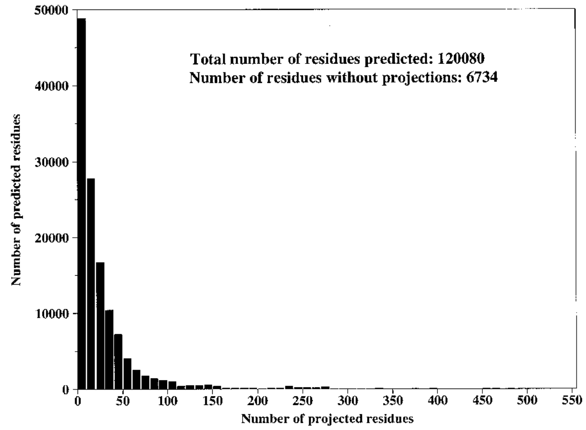

Fig. 4. Number of residues projected onto each of the predicted residues through the search and alignment procedure.

secondary structure prediction. The general approach is applicable to any propensity-based prediction technique. Rigorous dynamic programming pairwise alignment of the base sequence with each and all related sequence fragments or entire sequences results in more precise relationships than those arising from multiple alignment procedures that demand significant global sequence matching and can miss distantly related sequence spans. Any relaxation in significance can yield mistaken alignments, which in turn reduce prediction performance. Our method also projects secondary structure propensities of individual residues from sequences in the related set onto the correspondingly matched residues in the query sequence through the use of weights proportional to the similarity of the aligned fragments.

The mean residue-by-residue prediction accuracy of the technique described here is 74.8% resulting from a one-at-a-time protein jack-knife procedure applied to a carefully selected set (RS) of 125 nonhomologous protein sequences with known tertiary structure as originally listed by Rost and Sander,[6] albeit one chain of an inappropriate membrane-buried protein was excluded. Matthews[54] correlation coefficients for α-helix, β-sheet, and coil were 0.61, 0.45, and 0.44, respectively. The distribution of the accuracy values for different chains of the RS set is shown in Figure 3. This set has become a comparative standard to assess the quality of prediction schemes.[7,44] The accuracy without the jack-knife procedure was 77.5%, only 2.5 percentage points higher than with jackknifing. However, because each protein structure in the 125-protein set represents on average ~0.8% of the total information, statistics gathered from the training set may be insufficient for this method. The data bank of known protein structures[30] is ever increasing, and currently there are 556 protein chains (see Methods) whose sequences are maximally related at the 30% residue identity level. The accuracy of PREDATOR on this larger set (FA) was 74.6% without jack-knife calculations, which would have involved prohibitive computational requirements. However, because each protein in the latter set contains on average less than 0.2% of the total information, the prediction accuracy, based on the experience with the 125-protein set, would be expected to drop marginally by ~0.6% due to the jack-knife procedure, leading to a real expected accuracy near 74%. The accuracy with respect to the STRIDE[43] secondary structure assignments was ~0.5% lower than that achieved with the DSSP[42] method.

The influence of each aligned fragment on the secondary structure prediction of the query sequence was dependent on the identity level within the aligned region, according to equation 1 (see Methods). Thus, sequences or sequence spans very similar to the query sequence and those insignificantly related to it have little or no influence on the prediction, whereas sequences with similarity in the most informationally rich range of 35–70% make the greatest contribution. Very similar sequences provide little new information with respect to the query sequence. Highly diverged sequences can have considerably different secondary structure even when the overall topological equivalence is preserved and must therefore be downweighted. This local downweighting of very distant sequences and closely related sequences is unique to our approach and certainly different from other sequence weighting schemes.[18,45-47] Furthermore, Vogt et al.[48] have shown that protein sequence alignments, when compared with those derived from three-dimensional structural superposition, display a mean correctness of match near 90% at 35% residue identity, whereas at 30%, 25%, and 20% the respective average accuracies drop quickly to 85%, 75%, and 55%. The sequence identity is calculated locally for the aligned fragments considered significant by the SIM routine. This allows the use of even more related sequence

SWISS-PROT identifier	Protein name	Organism	E-value
FIS_ECOLI	Factor-for-inversion stimulation protein	*E. coli*	1.3e-31
FIS_HAEIN	Factor-for-inversion stimulation protein	*H. influenzae*	4.2e-28
NTRC_AZOBR	Nitrogen assimilation regulatory protein	*A. brasilense*	7.4e-10
NTRC_RHIME	Nitrogen assimilation regulatory protein	*R. meliloti*	2e-08
NTRC_BRASR	Nitrogen assimilation regulatory protein	*Bradyrhizobium Sp*	3.2e-07
NTRC_RHOCA	Nitrogen regulation protein NTRC	*R. capsulatus*	6.4e-07
ATOC_ECOLI	Acetoacetate metabolism regulatory protein	*E. coli*	0.0077
NTRC_KLEPN	Nitrogen assimilation regulatory protein	*K. pneumoniae*	0.0078
FLBD_CAUCR	Transcriptional regulatory protein FLBD	*C. crescentus*	0.019
NTRC_ECOLI	Nitrogen regulation protein NR(i)	*E. coli*	0.024
NTRC_SALTY	Nitrogen regulation protein NR(i)	*S. typhimurium*	0.024
NTRC_PROVU	Nitrogen regulation protein NR(i)	*P. vulgaris*	0.05

a

```
                                                                                                   Percent   Pseudo-
                                                                                                   identity  information
3fis.brk    1    PLRDSVKQALKNYFAQLNGQ-DVNDLYELVLAEVEQPLLDMVMQYTRGNQTRAALMMGINRGTLRKKLKKYGMN      -         -
FIS_ECOLI   26   PLRDSVKQALKNYFAQLNGQ-DVNDLYELVLAEVEQPLLDMVMQYTRGNQTRAALMMGINRGTLRKKLKKYGMN      100.0%    0.00
FIS_HAEIN   27   PLRDSVKQALRNYLAQEDGQ-DVNDLYELVLAEVEHPMLDMIMQYTRGNQTRAANMLGINRGTLRKKLKKYGMG      88.9%     0.11
NTRC_AZOBR  402  GESAAVERHEKDYFAAHKDGMPSNGLYDRVLREVERPLIISLSLSATRGNQIKAAQLLGLNRNTLRKKIRDLDIQ     48.5%     0.35
NTRC_RHIME  403  SHSQAVEENMRQYFASFGDAEPPSGLYDRVLAEMEYPLILAALTATRGNQIKAADLLGLNRNTLRKKIRELGVS      39.7%     0.37
NTRC_BRASR  401  NEGGAVEAYLSSHESGFPNGVPPPGLYHRTEKEIEIPLLTAALAATRGNQIRAADLLGLNRNTLRKKIRDLDIQ      44.1%     0.36
NTRC_RHOCA  376  KESASIARHERRYEDLHGGNEPPPGLYDRILAEMEAPLIELALDATGGNQAKCADLLGINRNTLRKKITDLDIQ      42.7%     0.36
ATOC_ECOLI  389  QIRQPVCNAGEVKTAPVGER-----NLKEEIKRVEKRIIMEVLEQQEGNRTRTALMLGISRRALMYKLQEYGID      40.1%     0.37
NTRC_KLEPN  396  QMLPDSWATLLGQWADRALRSGHQNLLSEAQPEMERTLLTTALRHTQGHKQEAARLLGWGRNTLTRKLKELGME      41.5%     0.37
FLBD_CAUCR  368  -PMAPAPNVAVARGAQMAADAASRAFVGSTVAEVEQQLIHDTLEHCLGNRTHAANILGISIRTLRNKLKEYSDA      48.8%     0.35
NTRC_ECOLI  397  QMQPDSWATLLAQWADRALRSGHQNLLSEAQPELERTLLTTALRHTQGHKQEAARLLGWGRNTLTRKLKELGME      41.5%     0.37
{NTRC_PROVU 398  RLTSQHWSQHLSLWADEALGEGKENILNDALPQFERTLLLSALAYTQGHKQDAARLLGWGRNTLTRKLKELGIE
{NTRC_PROVU 407                      GE-GKENILNDALPQFERTLLLSALAYTQGHKQDAARLLGWGRNTLTRKLKELGI       33.3%     0.37

PREDICTION       CCCHHHHHHHHHHHHHCCCC-CCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHCCCCHHHHHHHHHHCCC
STRIDE           CHHHHHHHHHHHHHHCCCCC-CCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHCCCHHHHHHHHHHHHCCC
```

b

Fig. 5. Secondary structure prediction of the factor for inversion stimulation (FIS) protein[53] (chain A, PDB code 3FIS) containing a helix-turn-helix motif for DNA binding.[52] The first 19 residues of the protein were not resolved by X-ray crystallographic techniques and are not included in the sequence. **A:** Results of the FASTA[34] search against the SWISS-PROT[32] database with the sequence of the FIS protein. The SWISS-PROT data bank was only used for demonstration because it is better documented than TREMBL. Sequences NTRC_ECOLI and NTRC_SALTY are nearly identical and only one of them was used for further calculations. The last column provides the estimate of statistical significance for a given database hit based on the extreme value distribution.[35] Only top-scoring hits of the database search with E-values less than 0.1 are shown. **B:** CLUSTAL W[46] multiple sequence alignment of the FIS protein and related sequences. Only a fragment of the alignment corresponding to the full sequence of the PDB entry 3FIS is shown. Local alignments of the related proteins with the query sequence produced by PREDATOR (through the SIM technique) by using the same set of proteins are shadowed and percentage of identity and pseudoinformation value are indicated in the last two columns. Numbered residue sites taken from the SWISS-PROT and corresponding to the NH_2-terminal residue in the spans shown are given in the column after database identifiers. The sequence FIS_ECOLI represents the full version of the FIS protein (with the 19-residue NH_2-terminal portion) and is discarded by PREDATOR because it is 100% identical to the query sequence. Note that the sequence NTRC_PROVU was aligned differently by CLUSTAL W (upper line) and by PREDATOR (lower line). The helix-turn-helix motif participating in DNA binding is in boxes. Secondary structure of the FIS protein predicted by PREDATOR and assigned by the STRIDE algorithm[43] from the known tertiary structure is shown below.

information than that in global multiple alignments because in many cases only relatively short domains or motifs in two proteins are really similar with global identity below the significance threshold.

Experts in certain protein families often work with sequence alignments resulting from careful manual refinement based on both visual criteria and available experimental data. PREDATOR has an option to allow to take into account such cases and preserve the register of multiply aligned sequences for secondary structure prediction. Pairwise alignments of the query sequence with other homologues are directly borrowed from the multiple alignment. Significant fragments of the pairwise alignments are then detected in the usual manner and their secondary structure propensities projected onto the base sequence as previously described. The rest of the procedure remains unchanged.

This method also maximizes the amount of sequence information used for prediction by searching for related sequences in the TREMBL database, the largest collection of protein sequences currently available. The total number of individual residues contained in the FA set of 556 proteins is 120,080, and only 6,734 or less than 1% had no match from related TREMBL sequences whose propensities improved the prediction (Fig 4). Nearly half of the predicted residues had between 1 and 10 related residues, whereas the remaining half had more than 10.

It must be emphasized that the prediction arising from our method is not a consensus over a set of sequences, especially given the use of weights and related fragments, but rather a prediction for just one sequence in the presence of others. Significant variation can be expected between the consensus structure of aligned sequences and the structure of each sequence in the set[49,50] such that if considerably diverged protein families are involved, the consensus approach can result in significant error. Our method is also particularly helpful in molecular modeling where the prediction is centered on the modeled sequence.

Secondary structure predictions generated separately for each sequence in a set could be used for a consensus prediction. This approach, however, has been justifiably criticized[51] because the amplitudes of individual propensities are not considered and decision making is unreliable when secondary structural states of equivalent residues display significant spread.

To demonstrate the performance of PREDATOR, we selected a well-documented example of the helix-turn-helix structural motif found in many protein families involved in DNA binding.[52] A SWISS-PROT database search using one sequence of such proteins, the *Escherichia coli* FIS (factor for inversion stimulation),[53] yielded 12 related sequences (Fig. 5A). Their global multiple alignment and local pairwise alignments used by PREDATOR are compared in Figure 5B. PREDATOR selectively used only significantly related fragments with weights dependent on the identity level. The sequence of the FIS protein from *Haemophilis influenzae*, nearly 90% identical to the query sequence within the local alignment used by PREDATOR, made little contribution to the prediction (pseudoinformation value 0.11), whereas the sequence fragments of the nitrogen assimilation regulatory proteins from different organisms, acetate metabolism regulatory protein, and transcription regulatory protein FLBD made considerable contributions where pseudoinformation values ranged from 0.35 to 0.37, close to the maximum possible (Fig. 2). The resulting prediction (Figure 5B) correctly reproduces both helices of the helix-turn-helix motif, as well as the two helices flanking the motif.

IMPLEMENTATION AND AVAILABILITY

The algorithm described here is implemented as a stand-alone portable C program called PREDATOR. The source code, documentation, and executables for many computer platforms are available for academic users via anonymous ftp from ftp.ebi.ac.uk (directories /pub/software/unix/predator, /pub/software/dos/predator) or from Dmitrij Frishman (frishman@mips.biochem.mpg.de). Protein sequences can be submitted for secondary structure prediction either to WWW URL http://www.embl-heidelberg.de/predator/predator_info.html or through electronic mail to predator@embl-heildelberg.de. A mail message containing HELP in the first line will be appropriately answered.

REFERENCES

1. Sonnhammer, E.L.L., Kahn, D. The modular arrangement of proteins as inferred from analysis of homology. Protein Sci. 3:482–492, 1994.
2. Argos, P. Prediction of the secondary structure of mouse nerve growth factor and its comparison with insulin. Biochem. Biophys. Res. Commun. 3:805–11, 1976.
3. Zvelebil, M.J., Barton, G.J., Taylor, W.R., Sternberg, M.J. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J. Mol. Biol. 195:957–961, 1987.
4. Crawford, I.P., Niermann, T., Kirschner, K. Prediction of secondary structure by evolutionary comparison: Application to the alpha subunit of thryptophan synthase. Proteins Struct. Funct. Genet. 2:118–129, 1987.
5. Benner, S.A. Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. Adv. Enzyme Regul. 28:219–236, 1989.
6. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
7. Salamov, A.A., Solovyev, V.V. Prediction of protein secondary structure by combining nearest-meighbour akgorithms amd multiple sequence alignments. J. Mol. Biol. 247:11–15, 1995.
8. Mehta, P.K., Heringa, J., Argos, P. A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. Protein Sci. 4:2517–2525, 1995.
9. Geourjon, C., Deléage, G. SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple sequences. Comput. Appl. Biosci. 11:681–684, 1995.
10. Barton, G.J. Protein secondary structure prediction. Curr. Opin. Struct. Biol. 5:372–376, 1995.
11. Benner, S.A., Gerloff, D.L., Jenny, T.F. Predicting protein crystal structures. Science 265:1642–1644, 1994.
12. Lupas, A., Koster, A.J., Walz, J., Baumeister, W. Predicted secondary structure of the 20 S proteasome and model structure of the putative peptide channel. FEBS Lett. 354:45–49, 1994.
13. Barton, G.J., Cohen, P.T.W., Barford, D. Conservation analysis and structure prediction of the protein serine/threonine phosphatases. Eur. J. Biochem. 220:225–237, 1994.
14. Bork, P., Holm, L., Koonin, E.V., Sander, C. The cytidyltransferase superfamily: Identification of the nucleotide-binding site and fold prediction. Proteins Struct. Funct. Genet. 22:259–266, 1995.
15. Levin, J., Pascarella, S., Argos, P., Garnier, J. Quantification of secondary structure prediction improvement using multiple alignments. Protein Eng. 6:849–854, 1993.
16. Di Francesco, V., Garnier, J., Munson, P.J. Improving protein secondary structure prediction with aligned homologous sequences. Protein Sci. 5:106–113, 1996.
17. Spouge, J.L. Speeding up dynamic programming algorithms for finding optimal lattice model. SIAM J. Appl. Math. 49:1552–1556, 1989.
18. Altschul, S.F., Carroll, R.J., Lipman, D.J. Weights for data related by a tree. J. Mol. Biol. 207:647–653, 1989.
19. Carrillo, H., Lipman, D.J. The multiple sequence alignment problem. SIAM J. Appl. Math. 48:1073–1082, 1988.
20. Gotoh, O. Optimal alignment between groups of sequences and its application to multiple sequence alignment. Comput. Appl. Biosci. 9:361–370, 1993.
21. Hogeweg, P., Hesper, B. The alignment of sets of sequences and the construction of phyletic trees: An integrated method. J. Mol. Evol. 20:175–186, 1984.
22. Johnson, M.S., Doolittle, R.F. A method for the simultaneous alignment of three of more amino acid sequences. J. Mol. Evol. 23:267–278, 1986.
23. Feng, D.-F., Doolittle, R.F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25:351–360, 1987.
24. Barton, G.J., Sternberg, M.J.E. A strategy for the rapid multiple alignment of protein sequences: Confidence levels from tertiary structure comparisons. J. Mol. Biol. 198:327–337, 1987.
25. Higgins, D.G., Sharp, P.M. CLUSTAL: A package for prerforming multiple sequence alignment on a microcomputer. Gene 73:237–244, 1988.
26. Vingron, M., Argos, P. A fast and sensitive multiple sequence alignment algorithm. Comput. Appl. Biosci. 5:115–121, 1989.
27. Hirosawa, M., Totoki, Y., Hoshida M., Ishikawa M. Comprehensive study on iterative algorithms of multiple sequence alignment. Comput. Appl. Biosci. 11:13–18, 1995.
28. Benner, S.A. Predicting the conformation of proteins from sequences: Progress and future progress. J. Mol. Recogn. 8:9–28, 1995.
29. Gerloff, D.L., Jenny, T.F., Knecht, L.J., Gonnet, G.H., Benner, S.A. The nitrogenase MoFe protein: A secondary structure prediction. FEBS Lett. 318:118–124, 1993.
30. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F.,

Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.

31. Heringa, J., Sommerfeldt, H., Higgins, D., Argos, P. OBSTRUCT: A program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. Comput. Appl. Biosci. 8:599–600, 1992.
32. Bairoch, A., Apweiler, R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res. 24:21–25, 1996.
33. Rice, C.M., Fuchs, R., Higgins, D.G., Stoehr, P.J., Cameron, G.N. The EMBL data library. Nucleic Acids Res. 21:2967–2971, 1993.
34. Pearson, W.R., Lipman, D.J. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85:2444–2448, 1988.
35. Altschul, S.F., Boguski, M.S., Gish, W., Wootton, J.C. Issues in searching molecular sequence databases. Nat. Genet. 6:119–129, 1994.
36. Frishman, D., Argos, P. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Eng. 9:133–142, 1996.
37. Zhang, X., Mesirov, J.P., Waltz, D.L. Hybrid system for protein secondary structure prediction. J. Mol. Biol. 225: 1049–1063, 1992.
38. Hutchinson, E.G., Thornton, J.M. A revised set of poteitials for beta-turn formation in proteins. Protein Sci. 3:2207–2216, 1994.
39. Huang, X., Miller, W. A time-efficient, linear-space local similarity algorithm. Adv. Appl. Math. 12:337–357, 1991.
40. Sander, C., Schneider, R. Database of homology derived protein structures and the structural meaning of sequence alignment. Proteins Struct. Funct. Genet. 9:56–68, 1991.
41. Nash, J.C. 'Compact Numerical Methods for Computers: Linear Algebra and Function Minimization.' New York: John Wiley, 1979.
42. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
43. Frishman, D., Argos, P. Knowledge-base protein secondary structure assignment. Proteins Struct. Funct. Genet. 23: 566–579, 1995.
44. Geourjon, C., Deléage, G. SOPM: A self-optimized method for protein secondary structure prediction. Protein Eng. 7:157–164, 1994.
45. Sibbald, P.R., Argos, P. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. J. Mol. Biol. 216:813–818, 1990.
46. Thompson, J.D., Higgins, D.G., Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680, 1994.
47. Henikoff, S., Henikoff, J.G. Position-based sequence weights. J. Mol. Biol. 243:574–578, 1994.
48. Vogt, G., Etzold, T., Argos, P. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. J. Mol. Biol. 249:816–831, 1995.
49. Russell, R.B., Barton, G.J. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J. Mol. Biol. 234:951–957, 1993.
50. Pascarella, S., Argos, P. A data bank merging related protein structures and sequences. Protein Eng. 5:121–137, 1992.
51. Niermann, T., Kirschner, K. Use of homologous sequence to improve protein secondary structure prediction. Methods Enzymol. 202:54–59, 1991.
52. Suzuki, M., Yagi, N., Gernstein, M. DNA recognition and superstructure formation by helix-turn-helix proteins. Protein Eng. 8:329–338, 1995.
53. Yuan, H.S., Finkel, S.E., Feng, J.-A., Johnson, R.S., Dickerson, R.E. The molecular structure of wild-type and a mutant FIS protein: Relationship between mutational changes and recombinational enhancer function or DNA binding. Proc. Natl. Acad. Sci. USA 88:9558–9562, 1991.
54. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochem. Biophys. Acta 405:422–451, 1975.