Polymer 43 (2002) 441–449

polymer

www.elsevier.com/locate/polymer

Protein secondary structure prediction based on the GOR algorithm incorporating multiple sequence alignment information☆

A. Kloczkowski[a,*], K.-L. Ting[a], R.L. Jernigan[a], J. Garnier[b]

[a]*Laboratory of Experimental and Computational Biology, National Cancer Institute, National Institute of Health, 12 South Drive, Bldg 12B, Rm B116, Bethesda, MD 20892-5677, USA*

[b]*Analytical Biostatistics Section, Mathematical and Statistical Computing Laboratory, CIT, National Institute of Health, 12 South Drive, Bldg 12A, Rm 2039, Bethesda, MD 20892-5626, USA*

Abstract

We have developed a new method for the prediction of the protein secondary structure from the amino acid sequence. The method is based on the most recent version (IV) of the standard GOR (J Mol Biol 120 (1978) 97) algorithm. A significant improvement is obtained by combining multiple sequence alignments with the GOR method. Additional improvement in the predictions is obtained by a simple correction of the results when helices or sheets are too short, or if helices and sheets are direct neighbors along the sequence (we require at least one residue of coil state between them). The imposition of the requirement that the prediction must be strong enough, i.e. that the difference between the probability of the predicted (most probable) state and the probability of the second most probable state must be larger than a certain minimum value also improves significantly secondary structure predictions. We have tested our method on 12 different proteins from the Protein Data Bank with known secondary structures. The average quality of the GOR prediction of the secondary structure for these 12 proteins without multiple sequence alignment was 63.4%. The multiple sequence alignments improve the average prediction to 71.9%. The correction for short helices and sheets and coil states separating sheets and helices improve further the average prediction to 74.4%. Setting the 10% minimum difference between the most probable and the second probable conformation leads to 77.0% accuracy of the prediction, while increasing this limit to 20% increases the average accuracy of the secondary structure prediction to 81.2%. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Protein; GOR algorithm; Secondary structure

1. Introduction

The prediction of the structure of a protein from its sequence is one of the most important and fundamental problems in modern science. With the rapid advancement in the Human Genome Project and the gene sequencing techniques in recent years, an enormous amount of information about gene sequences and translated protein sequences has been obtained. However, the available information about the protein structures (which is more relevant to protein function) is scarce. Only a small fraction of all proteins have known three-dimensional structure published in the Protein Data Bank (PDB). On 29th December 1999, the PDB database contained 11,363 known protein structures. The number of known structures in the PDB will probably exceed 13,000 by the end of 2000, but is still growing much more slower than the number of newly discovered protein sequences.

The prediction of tertiary structure from sequence is an ultimate goal of all protein folding theories. This is however theoretically difficult and will likely remain a computationally challenging problem. Many methods of prediction of the protein tertiary structure use as a starting point some information about the protein secondary structure. The more computationally feasible problem than the prediction of the protein tertiary structure is the prediction of the secondary structure from the protein sequence. The secondary structures of proteins consist of several structural elements such as α-helices, β-sheets, coils and turns. Each residue in the sequence belongs to one of these groups, and by analogy with polymer theory we may define conformational states. The conformational states are α-helices (H), β-strands or extended states (E), coils (C) and turns (T). Frequently (also in the recent version of the GOR method), turns (T) are classified as coil (C), which reduces the number of conformational states to three states: H, E

☆ This paper was originally submitted to *Computational and Theoretical Polymer Science* and received on 14 December 2000; received in revised form on 7 March 2001; accepted on 14 March 2001. Following the incorporation of *Computational and Theoretical Polymer Science* into *Polymer*, this paper was consequently accepted for publication in *Polymer*.

* Corresponding author.
E-mail address: kloczkoa@mail.nih.gov (A. Kloczkowski).

PII: S0032-3861(01)00425-6

and C. The prediction of the protein secondary structure from its amino acid sequence thus corresponds to the prediction of the sequence of the H, E, C conformational states.

There are many diverse methods for the prediction of secondary structures. One of the earliest and most successful methods was the GOR method developed by Garnier and coworkers [1–9]. The GOR method was based on information theory techniques, and the details of this method will be given in the next chapter. Newer methods use a variety of artificial intelligence techniques such as neural networks [11,14,23,25–27] machine learning [12], nearest neighbor algorithms [4,13] and combine approaches [3,15–18]. A detailed review of the problem of the prediction of conformations of proteins from a sequence is given in Ref. [19]. Cuff and Barton developed a set of about 400 protein domains to evaluate the performance of various protein secondary structure prediction algorithms [24]. Several recently developed algorithms have the accuracy of the prediction around 75% [23,25–27,29]. According to Frishman and Argos, the accuracy of the secondary structure prediction may reach 80–85% in near future [28].

2. Method

Our approach is based on combining multiple sequence alignments with the GOR method. The GOR method is one of the most successful schemes for the prediction of the secondary structure from the protein sequence. The method was proposed by Garnier, Osguthorpe and Robson [1] in 1978, and the name of this method is taken from the first letters of the names of the authors. The method has been further developed in a series of publications [2–7]. The basic idea behind the GOR method is the use of the formalism of information theory and Bayesian statistics to relate the amino acid sequence to the protein secondary structure. The basic mathematical object in information theory is the information function I(S,R)

$$I(\mathrm{S};\mathrm{R}) = \log[P(\mathrm{S}|\mathrm{R})/P(\mathrm{S})] \tag{1}$$

defined as a logarithm of the ratio of the conditional probability $P(\mathrm{S}|\mathrm{R})$ of observing conformation S, (where S is one of the three states: helix (H), extended (E) or coil (C)) for residue R (where R is one of the 20 possible amino acids) and the probability $P(\mathrm{S})$ of the occurrence of S. The present version of the GOR program (GOR IV) uses the three conformational states H, E and C. From the protein database of sequences with known secondary structure it is possible to estimate the information function I(S;R).

The observation of the given residue in state S depends not only on the type of the amino acid R but on other amino acids in the sequence, especially those which are relatively close sequential neighbors.

It is convenient to use in the analysis the information difference

$$I(\Delta\mathrm{S};\mathrm{R}_1,\mathrm{R}_2,\ldots,\mathrm{R}_n) = I(\mathrm{S};\mathrm{R}_1,\mathrm{R}_2,\ldots,\mathrm{R}_n) - I(n - \mathrm{S};\mathrm{R}_1,\mathrm{R}_2,\ldots,\mathrm{R}_n) \tag{2}$$

where $n - \mathrm{S}$ denotes the conformations different than S, i.e. if S is H then $n - \mathrm{S}$ is E and C.

Information theory in general enables decomposing the information brought to a complex event into the sum of information of simpler events, generally

$$\begin{aligned} I(\Delta\mathrm{S};\mathrm{R}_1,\mathrm{R}_2,\ldots,\mathrm{R}_n) &= I(\Delta\mathrm{S};\mathrm{R}_1) + I(\Delta\mathrm{S};\mathrm{R}_2|\mathrm{R}_1) \\ &\quad + I(\Delta\mathrm{S};\mathrm{R}_3|\mathrm{R}_1,\mathrm{R}_2) + \cdots \\ &\quad + I(\Delta\mathrm{S},\mathrm{R}_n|\mathrm{R}_1,\mathrm{R}_2,\ldots,\mathrm{R}_{n-1}) \end{aligned} \tag{3}$$

The GOR method uses a window of 17 residues, i.e. for a given residue j only the first eight sequentially neighboring residues on each side of j are included. The version IV of GOR method uses the pair-wise approximation for the information function so that

$$\begin{aligned} \log\frac{P(\mathrm{S}_j,\mathrm{LocSeq})}{P(n-\mathrm{S}_j,\mathrm{LocSeq})} &= \frac{2}{17}\sum_{\substack{n,m=-8\\ n>m}}^{+8}\log\frac{P(\mathrm{S}_j,\mathrm{R}_{j+m},\mathrm{R}_{j+n})}{P(n-\mathrm{S}_j,\mathrm{R}_{j+m},\mathrm{R}_{j+n})} \\ &= \frac{15}{17}\sum_{m=-8}^{+8}\log\frac{P(\mathrm{S}_j,\mathrm{R}_{j+m})}{P(n-\mathrm{S}_j,\mathrm{R}_{j+m})} \end{aligned} \tag{4}$$

Here $\sum_{m=-8}^{+8}$ denotes the summation over the window m between −8 and +8. The pair frequencies of residues R_{j+m} and R_{j+n} with R_j occurring in conformations S_j and $n - \mathrm{S}_j$ are calculated from the database. The GOR IV program uses a database of 267 proteins with known structure. All proteins in the database have their structure determined with resolution better that 2.5 Å. Based on these experimentally determined frequencies (from the GOR database) and the approximations mentioned above the GOR program can calculate probabilities of conformational states for any sequence. The GOR IV program can be accessed through the WWW server located at NIH through the link http://abs.cit.nih.gov/gor/.

The server takes as input the query sequence in FASTA format and predicts its secondary structure. For each residue i along the sequence the program calculates the probabilities p_{H}, p_{E} and p_{C} and the secondary structure conformational prediction (H, E or C). The probabilities are normalized

```
!         JX0184
-VKVPEPFAWNESFATSYKNIDLEHRTLFNGLFALS-EFNTRDQLLACKEVFVMHFRDEQGQMEK-ANYE-HFEEHRGIHEGFLEKMGHWKAPVAQKDIKFGMEWLVNHIPTEDFKYKGKL
!         S50177
-VKVPAPFAWNEDFATSYKFIDLEHRTLFNGLFALS-EFNTRDQLLACKEVFVMHFRDEQGQMEK-ANYE-HFEEHKGIHEGFLEKMGHWKAPVAQKDIRFGMEWLVNHIPAEDFKYKGKL
!         JT0560
-MKIPVPYAWTPDFKTTYENIDSEHRTLFNGLFALS-EFNTQHQLNAAIEVFTLHFHDEQGQMIR-SNYV-NTKEHTDIHNGFMDTMRGWQSPVPQKALKDGMEWLANHIPTEDFKYKGKL
!         S50178
-MKVPAPYAWNSDFATTYENIDSEHRTLFNGLFALA-EFNTLTQLNAAIEVFTLHFHDEQGQMIR-SNYV-NTKEHTDIHNGFMDVMRGWRSPVPQQDLLAGMAWLANHIPTEDFKYKGKL
!         1HMO_A
GFPIPDPYCWDISFRTFYTIIDDEHKTLFNGILLLSQA-DNADHLNELRRCTGKHFLNEQQLMQS-SQYA-GYAEHKKAHDDFIHKLDTWDG-----DVTYAKNWLVNHIKTIDFKYRGKI
!         HRTHBD
GFPIPDPYCWDISFRTFYTIVDDEHKTLFNGILLLSQA-DNADHLNELRRCTGKHFLNEQQLMQA-SQYA-GYAEHKKAHDDFIHKLDTWDG-----DVTYAKNWLVNHIKTIDFKYRGKI
!         HRGG
GFPIPDPYVWDPSFRTFYSIIDDEHKTLFNGIFHLAID-DNADNLGELRRCTGKHFLNEQVLMQA-SQYQ-FYDEHKKEHEGFIHALDNWKG-----DVKWAKSWLVNHIKTIDFKYKGKI
!         HRTH
GFPIPDPYGWDPSFRTFYSIIDDEHKTLFNGIFHLAID-DNADNLGELRRCTGKHFLNEQVLMQA-SQYQ-FYDEHKKAHEEFIRALDNWKG-----DVKWAKSWLVNHIKTIDFKYKGKI
!         S38261
GFEIPEPYKWDESFQVFYEKLDEEHKQIFNAIFALCGG-NNAGNLKSLVDVTANHFADEEAMLKASASYG-DFDSHKKKHEDFLAVIRGLGAPVPQDKINYAKEWLVNHIKGTDFGYKGKL
!         S16190
GFEVPEPFKWDESFQVFYDKLDEEHKQIFNAIFALGGG-NNADNLKKMIDVTANHFADEEAMMLASAAYKSEHPGHKKKHEDFLAVIRGLSAPVPNDKLLYAKDWLVNHIKGTDFTYKGKL
!         S29264
-YDIPEPFRWDESFKVFYE------------------------------------------------------------------------------------------------------
!         S23922
PFDIPEPYVWDESFRVFYDNLDDEHKGLFKGVFNCAADMSSAGNLKHLIDVTTTHFRNEEAMMDA-AKYE-NVVPHKQMHKDFLAKLGGLKAPLDQGTIDYAKDWLVQHIKTTDFKYKGKL
!         PS0350
GFDIPEPYVWDESFRVFYDLLDDEHKGLFQG------------------------------------------------------------------------------------------
!         HRTHM
GWEIPEPYVWDESFRVFYEQLDEEHKKIFKGIFDCIRD-NSAPNLATLVKVTTNHFTHEEAMMDA-AKYS-EVVPHKKMHKDFLEKIGGLSAPVDAKNVDYCKEWLVNHIKGTDFKYKGKL
!         HRIN
GFPVPDPFIWDASFKTFYDDLDNQHKQLFQAILTQG-NVGGATAGDNAYACLVAHFLFEEAAMQV-AKYG-GYGAHKAAHEEFLGKVKGGSA-----DAAYCKDWLTQHIKTIDFKYKGKL
!         A29667
-FPIPIPYCW--LLRTLIKKIQ---AVIPKGVLAMT--------VAQVCHVVP--------LLVG----------------GIIQQL----------VIEYSVIL-TD-------------
!         G69605
-MIFMKTLIEG--ETHMAKKVDAEYYRQLEQIQAAD------FVLVELSLYLNTHPHDEDALKQFN-QYSGYSRHLKRQFESSYGPLLQFG------NSPAGKDWDWGKGP---WPWQV--
```

Fig. 1. The multiple alignment of chain A of protein hemerythrin (oxy) 1HMO.

between 0 and 1 with

$$p_H + p_E + p_C = 1 \tag{5}$$

Usually the predicted conformational state corresponds to that with the highest probability, but sometimes there are exceptions. The GOR IV program has been previously tested on various sequences with known structure. Usually the predictions of the GOR IV program are accurate in the 60–70% range, with the average accuracy (tested on the relatively large sample of proteins) being 64.4% [7]. The accuracy of the prediction is measured by calculating the percentage of the sequence residues with the correctly predicted conformational states. Other methods used for the secondary structure predictions such as neural network methods or nearest-neighbor methods have similar or lower success rates when based on single sequence [8]. A big advantage of the GOR method over other methods is that it clearly identifies all factors that are included in the analysis and calculates probabilities of all three conformational states.

Sequence comparison is one of the most important methods in computational biology. Different sequences are compared to each other to find out which parts of the sequences are alike and which are dissimilar. The similarity of sequences is usually related to their evolutionary dependence. In the alignment of two sequences the sequences are treated as text strings where each letter in the string corresponds to the one letter code for a given residue at the specific position along the protein sequence. The alignment of two sequences can be visualized as a process of sliding one character string over another string trying to find the best possible match, where most characters of two strings are similar or belong to the same classes of residues (such as residues that are charged positively or negatively, aromatic, aliphatic, etc.). The quality of alignment of the two sequences is measured by a properly defined score. Additionally, in the process of the alignment the cutting of the text strings and the creation of gaps is allowed (but subject to some penalties in the calculation of the score). There are standard programs developed for the alignment and comparison of two as well as multiple sequences. One of the most popular programs is FASTA, one of the family of FAST programs for sequence database search [10]. Because the problem of sequence alignments is relatively unknown to the polymer community we illustrate it below. We show the alignment of two protein sequences. The first (upper) sequence corresponds to chain A of the protein hemerythrin (oxy) which has 113 residues and the known structure (1HMO) deposited in the PDB, the second (lower) sequence corresponds to S16190 myohemerythrin-polychaete (*Nereis diversicolor*) a protein of known sequence but unknown structure. The alignment of these two sequences was obtained with the FASTA program. The vertical bars | between these two sequences correspond to the exact matching of residues (identical characters in two strings), while the colons: illustrate the partial matching when two residues belong only to the same category of aminoacid (like I and V at the fourth position in two strings, because Isoleucine (I) and Valine (V) are both bulky aliphatic). In order to get the best sequence alignment two single gaps (after Met62 and Gly69) marked by — and a longer gap of the length of five residues after Thr92 were created in the sequence of 1hmo_a. The total length 120 of the alignment consists therefore of 113 residues of 1HMO_A and seven gaps.

```
                      10        20        30        40        50        60

1HMO_A        GFPIPDPYCWDISFRTFYTIIDDEHKTLFNGILLLSQADNADHLNELRRCTGKHFLNEQQ

              || :|:|: || ||::||  :|:||| :||:|: |: ::|||:|:::   |::|| :|: 

S16190        GFEVPEPFKWDESFQVFYDKLDEEHKQIFNAIFALGGGNNADNLKKMIDVTANHFADEEA
                      10        20        30        40        50        60


                        70        80        90             100       110

1HMO_A        LM-QSSQYAG-YAEHKKAHDDFIHKLDTWDGDVT-----YAKNWLVNHIKTIDFKYRGKI

              :|  |: | : :  ||| |:||:  :   :: |      |||:|||||||  || |:||:

S16190        MMLASAAYKSEHPGHKKKHEDFLAVIRGLSAPVPNDKLLYAKDWLVNHIKGTDFTYKGKL
                      70        80        90       100       110       120
```

Multiple sequence alignment is a similar problem, but instead of aligning two sequences we have to align together larger numbers of sequences, and because of this the problem is computationally more complicated, especially if both insertions and deletions in sequence are allowed. Several standard programs have been developed in computational biology for this purpose, and one of the most popular is the family of CLUSTAL programs. Fig. 1 shows the multiple sequence alignments. The same chain A of the protein 1HMO is now aligned with 16 other sequences by using CLUSTALX — one of the programs for multiple alignments from the CLUSTAL family. The lines starting with ! contain the names of the proteins whose sequences are displayed in the line below. The alignment of 1HMO_A and S16190 obtained in the multiple alignment is now slightly different than the alignment of these two sequences as obtained with FASTA.

The multiple sequence alignments have been proposed and used earlier to improve the secondary structure prediction. The inclusion of evolutionarily related sequences can improve significantly the prediction of secondary structure [6,15–16,20–24,29].

Such improvement in the accuracy of the prediction is due to the fact that the sequence alignment of homologous proteins coincides with their structural alignment, and therefore aligned residues, especially those inside the protein core, have mostly similar secondary structures. Multiple alignments, besides the information about the identity of the aligned residues, provide information about the location of gaps and the distribution of mutations in the multiple aligned sequences of homologous proteins. However, the detailed basis for the improvement of the secondary structure prediction by using multiple sequence alignments has not been yet fully developed.

Our method consists of several steps. First we have chosen a set of proteins with known structures published in the Protein Data Bank:

www.rcsb.org/pdb. All these proteins have well determined structures (the resolution of the structure is better than 2.5 Å, mostly better than 2.0 Å). The proteins were chosen to represent various classes of proteins and various folds. The list of the structures we have chosen is given below:

1. 1A58 (Cyclophilin)
2. 1HMO (Hemerythrin (oxy)) chain A
3. 12GS (Glutathione S-Transferase) chain A
4. 1HGE (Hemagglutinin) chain A
5. 1TF4 (T. Fusca Endo/Exo-Cellulase) chain A
6. 1HXN (Hemopexin)
7. 1A8L (Protein Disulfide Oxidoreductase)
8. 1AVA (Barley α-Amylase 2(Cv Menuet) chain A
9. 1AVA (Barley α-Amylase/Subtilisin Inhibitor) chain C
10. 1ARL (Apo-Carboxypeptidase)
11. 1DQI (Superoxide Reductase) chain A
12. 1AVM (Superoxide Dismutase) chain A

None of these proteins belongs to the databank of the GOR program, nor has a strong identity to any of 267 proteins in the GOR databank. We checked for such possible similarities by using the BLAST program available on the National Center of Biotechnology Information of NIH webpage http://www.ncbi.nlm.nih.gov/BLAST/.

BLAST calculations detected only four similarities between the set of the proteins listed above and the databank of GOR proteins: 28% identity between 1HMO_A and 3SDH_A, 29% identity between 1AVA_C and 1TIE, the 23% identity between 1AVA_A and 1CGT, and the 49% identity between 1AVM_A and 1ABM_A. All these identities are less than 50% and therefore do not introduce a strong bias for our predictions of the secondary structure.

For each of these 12 target proteins we make the sequence alignments against remote homologous sequences available in the PIR database by using the FASTA program (Wisconsin Package, version 8 of the Genetic Computer Group). PIR is the protein sequence database of PIR-International. The database is a collaborative project between the Protein Information Resource (PIR) at the National Biomedical Research Foundation (NBRF) affiliated with the Georgetown University Medical Center, the International Protein Information Database in Japan (JIPID) and the Munich Institute for Protein Sequences (MIPS). The PIR database is available on the webpage: http://pir.georgetown.edu/.

For each of the 12 target proteins we have chosen through sequence alignments homologous sequences from the PIR database having the pairwise sequence identity to the target protein greater than 20–30%. The identity percentage limit was dependent on the number of sequences found. If, for a given target protein, FASTA found large numbers of homologous sequences the identity limit was set to 30%, whereas for proteins for which FASTA produces a relatively small number of homologous sequences the identity limit was lowered to 20%. This range of sequence identity limits was suggested in recent work of Geetha et al. [9]. All sequences satisfying the identity criterion (for a given target protein) obtained from the PIR database through the pairwise FASTA alignments are subsequently used in sequence multiple alignments with CLUSTALX. Fig. 1 shows the case of the multiple alignment for the chain A of 1HMO. The FASTA alignments of sequences from the PIR database led to the selection of 16 homologous sequences satisfying the similarity criterion. Those 16 sequences and the sequence of the target protein (1HMO_A) were used in the multiple sequence alignments of the 17 sequences shown in Fig. 1.

The number of sequences in the multiple alignments is shown in the last column of Table 1. Table 1 shows also the number of residues (chain length) of each target protein. The multiple alignments obtained by the CLUSTALX program were then reformatted to the FASTA format (as shown in Fig. 1) and used as input for the GOR IV program. For each (jth) residue in the (ith) sequence GOR IV calculates probabilities of helix (H), extended (E) and coil (C) state, and the conformational state prediction, as discussed in detail in the earlier part of this paper. The GOR program output neglects information about gaps in the sequence, and therefore the results of the calculations were reformatted by adding information about gaps. As the result of calculations we obtained three matrices $P_H(i,j)$, $P_E(i,j)$, $P_C(i,j)$ with elements showing the GOR IV program probabilities of conformational states H, E and C (normalized according to Eq. (5)) for the jth residue in the ith multiple alignment sequence. The size of these matrices is m_n, where m is the number of aligned sequences and n is the total length of the alignment. For the multiple sequence alignment shown in Fig. 1 the number of aligned sequences is $m = 17$ and the length of the alignment $n = 121$. The length of the alignment is longer then the length (113 residues) of the 1HMO_A chain, because of the extra eight gaps in the sequence 1HMO_A in Fig. 1.

In the final step, we calculate the averages $\langle P_H(j)\rangle$, $\langle P_E j)\rangle$, $\langle P_C(j)\rangle$ by summing up the elements of matrices $P_H(i,j)$, $P_E(i,j)$, $P_C(i,j)$ at the jth position $(1 \leq j \leq n)$ in the sequence in the multiple alignment over all m sequences (rows) and dividing the results of the summation by the total number of entries (excluding gaps) in the jth colum of multiple sequence alignment file. The vectors $\langle P_H(j)\rangle$, $\langle P_E(j)\rangle$ and $\langle P_C(j)\rangle$ were then contracted in size by skipping elements corresponding to gaps in the sequence of the target protein. For example Fig. 1 shows the size of vectors $\langle P_H(j)\rangle$, $\langle P_E(j)\rangle$ and $\langle P_C(j)\rangle$ was reduced from 121 to 113 — the original length of the 1HMO_A chain. After this contraction the index j in $\langle P_H(j)\rangle$, $\langle P_E(j)\rangle$ and $\langle P_C(j)\rangle$ corresponds directly to the jth residue in the original sequence of the target protein.

The probabilities $\langle P_H(j)\rangle$, $\langle P_E(j)\rangle$ and $\langle P_C(j)\rangle$ allow us to predict the conformational state of the jth residue in the sequence of the target protein. The conformational state ν ($\nu = \mathrm{H, E}$ or C) of the jth residue which has the highest probability $P_\nu(j) = \max(\langle P_H(j)\rangle, \langle P_E(j)\rangle, \langle P_C(j)\rangle)$ is the natural choice for the prediction. We call this method the average multiple alignment prediction and use the notation P_{multi} for the calculated accuracy of this prediction.

We have tried also to use other schemes for predicting the

Table 1
Protein structures studied

Protein	Chain	Number of residues	Number of protein in the multiple alignment file
1A58		177	112
1HMO	A	113	17
12GS	A	210	62
1HGE	A	328	132
1TF4	A	605	63
1HXN		210	9
1A8L		226	51
1AVA	A	403	104
1AVA	C	181	73
1ARL		307	50
1DQI	A	124	14
1AVM	A	201	145

Table 2
Proteins from the multiple alignment file for 1A58 with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage with the target protein
1	S48567	79.1	57.8
2	T27882	76.3	60.2
3	CSTO	73.5	60.9
4	A46579	71.2	57.4
5	T27371	70.6	60.2
25	1A58	65.5	100.0

Table 4
Proteins from the multiple alignment file for 12GS with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	A46048	71.9	30.0
2	JC6529	69.5	97.1
3	S24330	68.6	31.8
4	S43431	68.1	32.1
5	T21898	68.1	31.3
7	12GS_A	67.6	100.0

secondary structure. A different method for making a prediction was based on matrices $P_H(i,j)$, $P_E(i,j)$ and $P_C(i,j)$ and the assumption that for a given jth position in the sequence the ith sequence among m sequences which has the largest probability max ($P_H(i,j)$, $P_E(i,j)$, $P_C(i,j)$) (where now we seek the maximum not only over three conformational states, but also over m sequences) at this position is the proper conformational choice. We call this method the strongest prediction scheme and use the notation P_{strong} for the accuracy of this prediction.

Because conformational predictions of the GOR method do not always coincide with states with the largest probabilities, we also use a prediction scheme based not on probabilities but on the counting of actual GOR conformational predictions for each residue in the multiple aligned sequences. The criterion of the predicted state was the maximum number of such predictions counted for a given residue in the sequence. In the case when two states had similar numbers of predictions the state with the larger average probability was chosen. We call this method the counting scheme and use the notation P_{count} for the accuracy of this prediction.

We have also tried the combination of the GOR single sequence prediction with the multiple sequence alignments. In the case when the GOR single sequence prediction of the conformational state for a given jth residue of the target protein is strong enough to have the probability $p(j)$ exceeding a specified minimum value p_{min}, we used this single sequence prediction. If the single sequence prediction of the conformational state of the target protein was weak ($p(j) < p_{\mathrm{min}}$) we use the multiple sequence alignments (similarly as for P_{multi} predictions) to improve the prediction. We have tried several values of the parameter p_{min} measuring the required minimal strength of the single sequence prediction; the best results were obtained for $p_{\mathrm{min}} = 0.7$ for probabilities $0 \leq p(j) \leq 1$ satisfying Eq. (5). We call this method which is the mixing of GOR single sequence predictions with the multiple sequence alignment prediction a mixed method and use the notation P_{mix} for the accuracy of this prediction.

3. Results and discussion

The results of all calculations are summarized in Tables 2–13 and in Table 14. Tables 2–13 show the results of the GOR calculations for each of the 12 target proteins with known structure. Sequences with the unknown structures in the multiple sequence alignments are assumed to have the structure of the target protein. Based on this assumption the GOR (single sequence) prediction of the secondary structure for each sequence in the multiple alignments is performed and the accuracy of the prediction for each sequence is calculated. We used the information about secondary structure of target proteins directly from the Protein Data Bank. We did not use the DSSP algorithm to assign 'correct' secondary structures. The sequences from the multiple alignments are then ranked according to their prediction accuracy. Tables 2–13 show for each target

Table 3
Proteins from the multiple alignment file for 1HMO with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	S50177	78.8	39.1
2	HRIN	73.5	46.9
3	S16190	72.6	42.5
4	JX0184	70.8	40.0
5	S38261	69.9	43.7
13	1HMO_A	52.5	100.0

Table 5
Proteins from the multiple alignment file for 1HGE with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	HMIVHM	74.1	95.1
2	HMIVDU	72.3	95.1
3	HMIV77	70.4	96.0
4	HMIVBH	69.8	23.4
5	S01882	69.8	23.7
18	1HGE_A	68.6	100.0

Table 6
Proteins from the multiple alignment file for 1TF4 with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	JC5874	61.3	53.4
2	A39199	58.4	73.1
3	1TF4_A	57.7	100.0
4	I40807	55.9	56.0
5	S12021	55.5	57.1

Table 8
Proteins from the multiple alignment file for 1A8L with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	H71239	74.8	88.5
2–3	1A8L	72.1	100.0
2–3	S54843	72.1	100.0
4	F75204	71.7	88.9
5	A72669	65.5	38.1

protein the first five sequences with the highest prediction agreement and the position of the target protein in the ranking. For each sequence the percentage identity with the target protein is displayed. Tables 2–13 show very interesting features, the GOR prediction of the secondary structure for the sequence of the target protein is always worse than the prediction incorporating some other homologous sequences. For 1A58 (Table 2) there are 24 out of 112 homologous sequences that have better predictions of the secondary structure than 1A58 itself, and sequences with the best prediction have about 60% identity to 1A58. The best prediction 79.1% is obtained for a sequence (S48567), which has 57.8% identity with 1A58. The GOR prediction for 1A58 based on its sequence is 65.5% correct. For 1HMO_A (Table 3) there are 12 out of 19 sequences, which provide better prediction of the secondary structure than 1A58_A. Sequences with the best ranking of the prediction have identity to 1A58_A of the order of 40%. The best prediction (with 78.8% accuracy) for S50177 is much above the GOR prediction for 1HMO_A, which is a mere 52.5%. For the case of 12GS_A shown in Table 4, the target protein takes 7th place (out of 62 sequences) in the ranking of the accuracy of the prediction, and the gap between 12GS_A (67.6% accuracy) and the best prediction (71.9%) is much lower than in the previous case (1HMO_A). The sequence that has the best prediction (A46048) has only 30% identity with the target protein. This shows that homologous sequences with relatively low identity are also important for improving secondary structure prediction. In some cases (such as 12GS_A, or 1AVA_C) the best prediction is given by sequences with low identity to the target protein, in some other cases (such as 1HGE_A) homologous sequences very similar to the target protein give the best prediction.

Table 14 summarizes the results of our calculations. For each target protein the accuracy of the single sequence GOR prediction P_{single} is compared with various multiple sequence alignment prediction schemes discussed in the previous chapter. The average multiple alignment prediction P_{multi} is based on probabilities $\langle P_{\mathrm{H}}(j)\rangle$, $\langle P_{\mathrm{E}}(j)\rangle$ and $\langle P_{\mathrm{C}}(j)\rangle$ calculated from the GOR multiple sequence alignment predictions. Results for the strongest prediction method with accuracy P_{strong} are based on the largest elements of the matrices $P_{\mathrm{H}}(i,j)$, $P_{\mathrm{E}}(i,j)$ and $P_{\mathrm{C}}(i,j)$. The accuracy of the prediction P_{count} is based on the counting of the GOR conformational predictions, for all sequences in the multiple alignment. The mixed method with accuracy P_{mixed} is the prediction scheme combining the GOR single sequence prediction with the multiple sequence alignment if the single sequence prediction is not strong enough (less than $p_{\mathrm{min}} = 0.7$).

The last row in Table 14 shows averages of these predictions over all 12 target proteins studied in this paper. The comparison of these various prediction schemes show that the strongest prediction method P_{strong} is the worst one, and on average (64.1% accuracy) it is only slightly better than the average GOR prediction for the single sequence P_{single} (63.4%). All other three prediction schemes give almost similar predictions averaged over the 12 proteins P_{multi} (71.9%), P_{count} (71.9%), P_{mixed} (71.5%).

Table 7
Proteins from the multiple alignment file for 1HXN with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	OQRT	63.5	73.0
2	A55486	63.5	73.7
3	OQHU	59.8	83.1
4	A40774	58.9	31.1
5–6	OQRB	57.8	100.0
5–6	1HXN	57.8	100.0

Table 9
Proteins from the multiple alignment file for 1AVA (chain A) with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	S14956	65.5	71.6
2	S14957	65.3	72.1
3	JT0946	65.0	70.6
4	JC7138	65.0	70.6
5	S05667	64.5	30.0
20	1AVA_A	55.1	100.0

Table 10
Proteins from the multiple alignment file for 1AVA (chain C) with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	JX0310	69.6	27.8
2	JX0311	69.1	27.8
3	TIWDKB	69.1	31.5
4	TIWDK	69.1	30.9
5	A24082	68.0	29.0
25	1AVA_C	61.9	100.0

Table 12
Proteins from the multiple alignment file for 1DQI with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	H71102	77.4	74.2
2	F75136	69.4	70.2
3	H69292	64.5	67.8
4	G72348	64.5	57.0
5–6	T44571	62.9	100.0
5–6	1DQI_A	62.9	100.0

This shows that multiple sequence alignment improves significantly the prediction of the secondary structure. On average the improvement is of the order of 8.5%, but for some individual proteins, such as 1HMO_A the improvement is as high as 24%.

We have tried several methods to further improve the prediction of the secondary structure. A simple improvement can be made by a critical analysis of the predicted conformational sequence. The GOR algorithm sometimes gives predictions that are unphysical, such as a helix or a sheet having a length of one or two residues, or H and E states, which are nearest neighbors along the sequence. We have employed a correction algorithm to the results of average multiple alignment prediction P_{multi}. We impose the requirement that helices (H) and strands (E) must have at least a length of three residues each, and that these two states cannot be direct neighbors along the sequence, and that E and H conformations must be separated by at least one coil (C) state. If the conformational sequence obtained by using the P_{multi} methods violate at the jth position in the sequence any of these rules, then the second most probable state (instead of the state with the highest probability) was taken as the choice from the $\langle P_{\mathrm{H}}(j)\rangle$, $\langle P_{\mathrm{E}}(j)\rangle$ and $\langle P_{\mathrm{C}}(j)\rangle$ vectors. Such corrected predictions are shown in the fourth column ($P_{\mathrm{multi,cor}}$) of the Table 14. The average improvement (averaged over 12 proteins) over P_{multi} prediction by using this correction scheme is 2.5%.

For the prediction of tertiary structure from secondary structure it is often important to have some secondary structure elements predicted with very high degree of certainty to use as starting points. This means, for example, that instead of having the whole secondary structure predicted with 70% accuracy, it may be better to have only 85% of the secondary structure predicted, but with the greater confidence of 82%. We have tried to find out for each target protein the conformational states, which could be predicted with a higher level of confidence. We impose the requirement that the prediction of the secondary structure by using the average multiple alignment prediction method P_{multi} (based on probabilities $\langle P_{\mathrm{H}}(j)\rangle$, $\langle P_{\mathrm{E}}(j)\rangle$ and $\langle P_{\mathrm{C}}(j)\rangle$) must be strong enough, i.e. that the difference between the probability of the predicted (most probable) state and the probability of the second most probable state at the jth position in the sequence must be larger than a certain minimum value Δ. If the difference is less than Δ the conformation of secondary structure for this position is indeterminate (marked by ?) and the whole conformational sequence is composed of four elements: H, E, C and ? (indeterminate states).

We have performed the calculations using two different values of this parameter

$\Delta = 0.1$ and $\Delta = 0.2$ (with probabilities normalized to 1 according to Eq. (5), so Δ is 10% or 20%). The last columns show results of the secondary structure prediction for these two cases, and the percent of indeterminate states for the two values of Δ.

For $\Delta = 10\%$ average prediction of the secondary structure has 77.0% accuracy with 22.0% of unpredicted states. By increasing Δ to 20% the confidence of the prediction of

Table 11
Proteins from the multiple alignment file for 1ARL with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	A56171	73.6	67.5
2	S29127	72.6	81.1
3	CPBOA	71.0	99.3
4	1ARL	69.7	100.0
5	A32128	69.1	65.7

Table 13
Proteins from the multiple alignment file for 1AVM with the best GOR predictions

Ranking position	Protein name	GOR prediction agreement (in %)	Percentage identity with the target protein
1	T42080	73.6	57.9
2	S15205	71.1	63.1
3–4	1AVM_A	70.7	100.0
3–4	JC4396	70.7	100.0
5	S04423	70.2	40.9

Table 14
Comparison of the results of the GOR secondary structure prediction for a target protein with various methods using multiple alignments. All predictions are measured by the percentage of correct residue sequence conformations

Target protein	P_{single}	P_{multi}	$P_{\text{multi\_cor}}$	P_{strong}	P_{count}	P_{mixed}	P_{multi} $\Delta = 10\%$	Fraction of indeterminate states $\Delta = 10\%$	P_{multi} $\Delta = 20\%$	Fraction of indeterminate states $\Delta = 20\%$
1A58	65.5	76.8	79.1	74.6	78.0	76.8	80.3	14.1	82.3	36.2
1HMO_A	52.2	76.1	79.6	65.5	75.2	75.2	80.3	28.3	84.6	54.0
12GS_A	67.6	71.0	72.4	60.0	71.4	71.9	76.9	19.5	82.4	37.6
1HGE_A	68.6	73.8	77.4	69.5	73.8	72.9	77.8	16.2	82.9	35.7
1TF4_A	57.7	64.0	65.6	62.0	63.3	64.5	69.2	24.8	72.5	41.7
1HXN	57.1	66.2	70.8	58.9	63.0	65.8	71.2	22.4	81.9	42.0
1A8L	72.1	76.1	76.5	61.1	77.9	76.1	80.6	22.6	80.8	46.9
1AVA_A	55.1	66.3	67.0	58.6	63.8	66.3	71.0	29.8	77.2	49.9
1AVA_C	61.9	78.5	82.3	70.7	76.2	75.7	85.0	18.8	88.6	37.0
1ARL	69.7	73.6	76.9	69.7	74.9	73.6	80.9	26.7	82.0	44.0
1DQI_A	62.9	70.2	71.8	63.7	73.4	69.4	76.8	23.4	82.1	46.0
1AVM_A	70.7	70.1	73.1	55.2	72.1	69.7	73.9	17.9	76.6	36.3
Averages	63.4	71.9	74.4	64.1	71.9	71.5	77.0	22.0	81.2	42.3

secondary structure was increased to 81.2% but the fraction of indeterminate states became significantly higher at 42.3%.

We have shown that by incorporating the multiple sequence alignment information into the GOR algorithm, we substantially improve the prediction of the secondary structure. The improvement of 8.5% to the 71.9% prediction accuracy is close to the improvements obtained by neural network programs (PHD and PsiPred). The correction for short helices and strands and exclusion of HE neighbors further improves the prediction to 74.4%. The new method also enables us to predict the secondary structure of a substantial part of the sequence with a confidence level greater than 80%. This shows that this new method is promising, and may successfully compete with artificial intelligence techniques. The method will be tested in the near future for more protein structures. The inclusion of additional information such as hydrophobicity of residues in the multiple alignments may help in further improving the method.

References

[1] Garnier J, Osguthorpe D, Robson B. J Mol Biol 1978;120:97.
[2] Gibrat JF, Garnier J, Robson B. J Mol Biol 1987;198:425.
[3] Biou V, Gibrat JF, Levin JM, Robson B, Garnier J. Protein Engng 1988;2:185.
[4] Levin JM, Robson B, Garnier J. FESB Lett 1986;205:303.
[5] Garnier J, Robson B. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York: Plenum Press, 1989. p. 417.
[6] Levin JM, Pascarella S, Argos P, Garnier J. Protein Engng 1993;6:849.
[7] Garnier J, Gibrat JF, Robson B. Meth Enzym 1996;266:540.
[8] Garnier J, Levin JM. CABIOS 1991;7:133.
[9] Geetha V, Di Francesco V, Garnier J, Munson PJ. Protein Engng 1999;12:527.
[10] Pearson WR, Lipman DJ. Proc Natl Acad Sci 1988;85:2444.
[11] Qian N, Sejnowski TJ. J Mol Biol 1989;202:865.
[12] King RD, Sternberg MJ. J Mol Biol 1990;216:441.
[13] Salzberg S, Cost S. J Mol Biol 1992;227:371.
[14] Holley LH, Karplus M. Proc Natl Acad Sci 1989;86:152.
[15] Rost B, Sander C. Proteins, Struct Funct Genet 1994;19:55.
[16] Salamov AA, Solovyev VV. J Mol Biol 1995;247:11.
[17] Stolorz P, Lapedes A, Xia Y. J Mol Biol 1992;225:363.
[18] Zhang X, Mesirov JP, Waltz DL. J Mol Biol 1992;225:1049.
[19] Benner SA, Jenny TF, Cohen MA, Gennet GH. Adv Enzym Regul 1994;34:269.
[20] Zvelebil MJ, Barton GJ, Taylor WR, Sterling MJE. J Mol Biol 1987;195:957.
[21] Rost B, Sander C. J Mol Biol 1993;232:584.
[22] Di Francesco V, Garnier J, Munson PJ. Protein Sci 1996;5:106.
[23] Cuff JA, Barton GJ. Proteins, Struct Funct Genet 2000;40:502.
[24] Cuff JA, Barton GJ. Proteins, Struct Funct Genet 1999;34:508.
[25] Ouali M, King RD. Protein Sci 2000;9:1162.
[26] Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G. Bioinformatics 1999;15:937.
[27] Jones DT. J Mol Biol 1999;292:195.
[28] Frishman D, Argos P. Folding Design 1997;2:159.
[29] Frishman D, Argos P. Proteins, Struct Funct Genet 1997;27:329.