[32] GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence

By JEAN GARNIER, JEAN-FRANÇOIS GIBRAT, and BARRY ROBSON

Introduction

How to extract properties from the amino acid sequence of a protein for understanding its function is one of the most timely and competitive areas of the biological sciences. It is related to a fundamental aspect of biology. A linear and ordered sequence of amino acids, coded and conserved by a linear and ordered sequence of nucleotide bases, codes for all that is characteristic of living organisms: specific and organized interactions in space and time between proteins, lipids, nucleic acids, and the cell metabolites. These characteristics depend on how a protein can fold in a unique active three-dimensional structure. This process is spontaneous under given environmental conditions, even though the living cell can add efficiency and control by use of some protein complexes, called chaperones, to catalyze this process.

Much effort has been devoted to the calculation of the spatial structure of a polypeptide chain from its amino acid sequence alone, with only limited but nevertheless encouraging success[1] when a polypeptide is longer than 10–20 amino acids. However, attempts to reduce the problem to simpler features of the protein fold such as α helix, β strands, and aperiodic or coil structure have yielded interesting results (see Refs. 2 and 3). These results have been an aid for designing new proteins, predicting the effect of point mutations, identifying the protein class, for instance, all-α or all-β proteins, predicting epitopes, etc. It is hoped that this information will be increasingly useful to molecular biologists and protein modelers. Usually the computing time is short, and many programs of secondary structure predictions are available on-line to the biologist.

The GOR method is one of the most popular of the secondary structure prediction schemes. This method is theoretically well founded in a series of earlier papers, and it has been the real first prediction of secondary structure implemented as a computer program. The three letters stand for

[1] Protein Structure Prediction Issue. *Proteins* **23,** 3 (1995).

[2] J. Garnier and J. M. Levin, *CABIOS* **7,** 133 (1991).

[3] B. Rost and C. Sander, *Trends Biochem. Sci.* **18,** 120 (1993).

the first letter of the names of the authors of the original publication.[4] This method remains remarkably popular,[5] but users have overlooked the fact that improved versions of the method have since been published, and a full description of them can be found in the book edited by Fasman.[6] The addition of homologous sequence information through multiple alignments has given a significant boost to the accuracy of secondary structure predictions.[7-9] In this chapter, after presenting the major principles used by the GOR method, we give some results obtained with an updated version of this method.

Principles of Method

In a series of articles, Robson *et al.*[10,11] used the formalism of information theory and Baysian statistics to establish the code relating the amino acid sequence and the secondary structures of a protein. This led later to the development of the GOR method.[4]

Information theory was developed in the 1950–1960s[12,13] and the GOR method made use of an information function described by Fano,[14] $I(S; R)$, which is defined as

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

Originally this formulation was concerned mainly with electronic transmission of information. In the present application, S is one of the three conformations, R is one of the 20 amino acid residues, $P(S|R)$ is the conditional probability for observing a conformation S when a residue R is present, and $P(S)$ is the probability of observing S. According to the definition of conditional probabilities, $P(S|R) = P(S, R)/P(R)$ where $P(S, R)$ is the joint probability of observing the events S and R and $P(R)$ is the probability of observing a residue R. It is easy to have an estimation of $I(S; R)$ from

[4] J. Garnier, D. Osguthorpe, and B. Robson, *J. Mol. Biol.* **120,** 97 (1978).

[5] L. B. M. Ellis and R. P. Milius, *CABIOS* **10,** 341 (1994).

[6] J. Garnier and B. Robson, *in* "Prediction of Protein Structure and the Principles of Protein Conformation" (G. D. Fasman, ed.), Chap. 10, p. 417. Plenum Press, New York, 1989.

[7] J. M. Levin, S. Pascarella, P. Argos, and J. Garnier, *Protein Eng.* **6,** 849 (1993).

[8] B. Rost and C. Sander, *J. Mol. Biol.* **232,** 584 (1993).

[9] V. di Francesco, P. J. Munson, and J. Garnier, *28th Annual Hawaii International Conference on System Sciences* (L. Hunter, ed.), p. 285. IEEE Computer Society Press, Los Alamos, 1995.

[10] B. Robson and R. H. Pain, *J. Mol. Biol.* **58,** 237 (1971).

[11] B. Robson, *Biochem. J.* **141,** 853 (1974).

[12] C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication." Univ. of Illinois Press, Urbana, Illinois, 1949.

[13] L. Brillouin, "Science and Information Theory." Academic Press, New York, 1956.

[14] R. Fano, "Transmission of Information." Wiley, New York, 1961.

a database of known sequences and corresponding observed secondary structures since $P(S, R) = f_{S,R}/N$, $P(R) = f_R/N$ and $P(S) = f_S/N$ with N being the total number of amino acids in the database, $f_{S,R}$ the number of residues R observed in the conformation S in the same database, f_R the total number of residues R, and f_S the total number of residues observed in the conformation S in the same database. Then

$$I(S; R) = \log[(f_{S,R}/f_R)/(f_S/N)] \tag{2}$$

This quantity can be obtained easily from the database provided it is large enough.

A more general treatment requires corrections for levels of data (see Robson[11] and below). Robson[11] introduced the information difference,

$$I(\Delta S; R) = I(S; R) - I(n\text{-}S; R) = \log(f_{S,R}/f_{n\text{-}S,R}) + \log(f_{n\text{-}S}/f_S) \tag{3}$$

where n-S stands for the conformations other than S (non-S); for instance, if S is α helix (H), n-S will be β strand (E) and coil (C) for a three-state prediction. It gives the extra information for S on the two others. It represents a kind of normalization where the total number of amino acids, N, and residues, R, in the database have disappeared from the equation. In effect the positive hypothesis (S; R) and the complementary negative hypothesis (n-S; R) are treated in concert. This quantity also corresponds to one-residue information or single-residue information or self-information. Calculated for the three conformations, the highest value of Eq. (3) for one of the conformations S will be the predicted conformation and will be the propensity for that residue to be in that conformation, usually expressed in centinat units when natural logarithms are used. This underlines one of the differences with the Chou–Fasman propensities which correspond approximately to the mantissa of the log of Eq. (2).

Equations (1) to (3) can be extended to a local sequence along the polypeptide chain of n consecutive residues R:

$$\begin{aligned} I(\Delta S_j; R_1, \ldots, R_n) = {} & \log[P(S_j, R_1, \ldots, R_n)/P(n\text{-}S_j, R_1, \ldots, R_n)] \\ & + \log[P(n\text{-}S)/P(S)] \end{aligned} \tag{4}$$

where $P(S_j, R_1, \ldots, R_n)$ is the joint probability of the conformation S at position j in the sequence and the local sequence $R_1, \ldots, R_n$. One may remark that

$$P(S_j, R_1, \ldots, R_n) + P(n\text{-}S_j, R_1, \ldots, R_n) = 1 \tag{5}$$

and that

$$P(S_j, R_1, \ldots, R_n)/P(n\text{-}S_j, R_1, \ldots, R_n) = P(S)/P(n\text{-}S)e^{I(\Delta S_j; R_1, \ldots, R_n)} \tag{6}$$

In predicting a residue to be in one conformation, one can predict either the one having the highest value of the information with Eq. (4) or the highest probability value taken from Eqs. (5) and (6). Probability values have been used for the prediction of Ramachandran zones.[15] They are more precise than confidence scales developed for other methods[8,16] and underline the fact that the decision to predict the conformation of the highest probability leaves the possibility that the other conformations have a definite probability to occur which can be close to the highest, and thus should not necessarily be ruled out.

One faces a fundamental problem when calculating information values. One needs to estimate terms such as $P(S_j, R_1, \ldots, R_n)$ involving N residues. It is impossible to evaluate such terms directly from the database, so one must resort to various approximations. The different versions of the GOR method correspond to various types of approximations we have tried in an effort to improve the accuracy of the method.

Approximations Involved in GOR Method

The first GOR version,[4] named GOR I, added to the single-residue information the so-called directional information of eight residues on each side of the residue to be predicted in the sequence. This limit of eight was not arbitrary but was based on studies of information content at increasing separations. To obtain the information measure, one starts by calculating from the database the frequency of each of the 20 amino acids residues at different positions, up to eight residues on the N-terminal and C-terminal side, when the central residue is observed in a given conformation but independently of the nature of that residue. In fact, in this approximation one assumes that there is no correlation between residues occurring at different positions in the window of 17 residues so defined. Then

$$I(\Delta S_j;\, R_1, \ldots, R_n) \approx I(\Delta S_j;\, R_j) + \Sigma_{m,m\neq 0}\, I(\Delta S_j;\, R_{j+m}) \qquad (7)$$

where j stands for any position j in the amino acid sequence of which the conformation ought to be predicted and m is between -8 (N-terminal side) and $+8$ (C-terminal side). The same version, GOR II, was updated with a new database in 1989.[6] Both versions predicted four conformations, H, E, C, and T, with T for turns. Subsequent versions predicted only three conformations H, E, and C, although the method has no intrinsic limitation in the number and nature of conformations. The different turn types and the relative difficulty of distinguishing between them using the DSSP program[17] led us to limit the prediction for the time being to three conforma-

[15] J. F. Gibrat, B. Robson, and J. Garnier, *Biochemistry* **30,** 1578 (1991).

[16] V. Biou, J. F. Gibrat, J. M. Levin, B. Robson, and J. Garnier, *Protein Eng.* **2,** 185 (1988).

[17] W. Kabsch and C. Sander, *Biopolymers* **22,** 2577 (1983).

tions so that helices and strands represent the major architectural structures of the conserved core of homologous proteins.

The next level of approximation, introduced in the GOR III version,[18] considered the correlation between the type of residues in the window and the type of the residue to be predicted. This version uses the so-called pair information:

$$I(\Delta S_j;\, R_1, \ldots, R_n) \approx I(\Delta S_j;\, R_j) + \Sigma_{m,m\neq 0}\, I(\Delta S_j;\, R_{j+m}|R_j) \qquad (8)$$

The second term on the right-hand side of Eq. (8) is a conditional information.[15] It involves the calculation from the database of pair frequencies of residues R_j and R_{j+m} with R_j having the observed conformations S_j and $n\text{-}S_j$ at position j, with a frequency of f_{S_j,R_{j+m},R_j} and $f_{n\text{-}S_j,R_{j+m},R_j}$, respectively, but the conformation of residue R_{j+m} is not taken into consideration. We have

$$I(\Delta S_j;\, R_{j+m}|R_j) = \log(f_{S_j,R_{j+m},R_j}/f_{n\text{-}S_j,R_{j+m},R_j}) + \log(f_{n\text{-}S_j,R_j}/f_{S_j,R_j}) \qquad (9)$$

When this approximation was used, the database (at that time containing roughly 12,000 residues) was barely large enough to allow an easy calculation of terms for Eq. (9). Each of these terms involves two amino acids and a secondary structure conformation, so there are 1200 entries in the table. The average number of observations per entry was therefore 10. However, the amino acids are not all equiprobable; some like Trp or Met are rarer, and the number of observations for entries involving such amino acids were less than 10. As a consequence, the probabilities estimated using the ratio of such sparse frequencies were unreliable and were responsible for a decrease of the prediction accuracy. To circumvent this problem, we introduced so-called dummy frequencies. Readers interested in the precise definition of these dummy frequencies are referred to Refs. 15 and 17.

Dummy frequencies amount to the following considerations. Let us assume that we observe in the database two occurrences of a Met at position $j - 1$ when the residue at j is a Trp in helical conformation. We can calculate easily the expected number we would observe if the two events were uncorrelated, namely, this is the frequency of Met at position $j - 1$ multiplied by the frequency of Trp at position j having a helical conformation divided by the total number of residues. This number is relatively reliable since it is calculated using frequencies that involve only one residue (which thus are greater than frequencies for pairs by a factor of 20, on average). Now we can ask the question, How much do we trust the frequencies for pairs we observed in the database? If we trust them 100%, we just use these frequencies in the calculations of information values. If we mistrust them 100%, we can always use the numbers calculated assuming that

[18] J. F. Gibrat, J. Garnier, and B. Robson, *J. Mol. Biol.* **198,** 425 (1987).

the events are independent, but then we are back to the approximation of Eq. (7). In fact, empirically, to improve the accuracy prediction we need to consider an intermediary stage when we bias the observed frequencies toward the calculated (uncorrelated) ones by a given amount.

The database available (see Table I) now contains about 63,000 residues, so the average number of observations for pairs per entry is 50. This is large enough for us to compute terms involving pairs of residues without the need for introducing dummy frequencies. Note, however, that this new database does not allow the calculation of terms involving triplets of amino acids. We thus decided to include more pairs in our description of the window of 17 residues. Instead of considering only the 16 pairs R_{j+m}, R_j, with m varying from -8 to $+8$ and $m \neq 0$, that is, the pair formed by each residue in the window and the central one, we consider all the possible pairs in the window [there are $(17 \times 16)/2$ such pairs]. We thus have used for the results presented below the following approximation (GOR IV version):

$$\log \frac{P(S_j, LocSeq)}{P(n\text{-}S_j, LocSeq)} = \frac{2}{17} \sum_{\substack{m=-8,\\ n>m}}^{+8} \log \frac{P(S_j, R_{j+m}, R_{j+n})}{P(n\text{-}S_j, R_{j+m}, R_{j+n})} - \frac{15}{17} \sum_{m=-8}^{+8} \log \frac{P(S_j, R_{j+m})}{P(n\text{-}S_j, R_{j+m})} \tag{10}$$

where *LocSeq* stands for the local sequence $R_1, \ldots, R_n$ of 17 residues around the residue to be predicted. The values of $P(S_j, LocSeq)$ from Eq. (10) for the three conformations are then used directly for the predictions instead of using probabilities calculated from Eqs. (5) and (6) with the information value calculated from Eq. (9).

Database and Results

We used a database of 267 protein structures having a resolution better than 2.5 Å with an R factor less than 25% and whose length is greater than 50 residues (see Table I). There is no pair of proteins with an identity above 30%. The prediction is carried out using a jackknife: the protein to be predicted is removed from the database, the parameters are estimated using the 266 remaining proteins, and the prediction is done using these parameters. As mentioned above, this database is large enough that we do not need to use dummy frequencies anymore. Moreover, we do not use decision constants to adjust the predicted number of secondary structures to the observed number in the database. In fact, there is no optimization of any sort; we just estimate the probabilities according to Eq. (10) from the frequencies observed in the database.

TABLE I

Database Proteins[a]

1aaj.x	1aak.x	1aap.a	1aba.x	1abk.x	1abm.a	1add.x
1ads.x	1alk.a	1aoz.a	1apa.x	1apm.e	1arb.x	1atr.x
1avh.a	1ayh.x	1bab.a	1bbh.a	1bbp.a	1bet.x	1bge.a
1bll.e	1bmd.a	1bov.a	1bpb.x	1brs.d	1btc.x	1c2r.a
1caj.x	1cau.a	1cau.b	1cde.x	1cdt.a	1cew.i	1cgt.x
1chm.a	1cmb.a	1cob.a	1col.a	1cpc.a	1cpc.b	1cpt.x
1crl.x	1cse.i	1ctf.x	1ctm.x	1cus.x	1ddt.x	1dhr.x
1dog.x	1dsb.a	1eaf.x	1eco.x	1ede.x	1end.x	1epa.a
1fba.a	1fdd.x	1fha.x	1fia.a	1fkb.x	1fna.x	1fnr.x
1fxi.a	1gal.x	1gd1.o	1gdh.a	1gky.x	1glt.x	1gmf.a
1gof.x	1gox.x	1gp1.a	1gpb.x	1gpr.x	1gsr.a	1hbq.x
1hdx.a	1hiv.a	1hlb.x	1hle.a	1hmy.x	1hoe.x	1hpl.a
1hrh.a	1hsl.a	1huw.x	1ifc.x	1ipd.x	1isu.a	1ith.a
1l29.x	1le4.x	1len.a	1lga.a	1lis.x	1lla.x	1lmb.3
1lts.a	1lts.d	1mdc.x	1mgn.x	1min.a	1min.b	1mjc.x
1mpp.x	1mup.x	1nar.x	1nba.a	1ndk.x	1noa.x	1nsb.a
1nxb.x	1ofv.x	1olb.a	1omf.x	1omp.x	1onc.x	1osa.x
1pda.x	1pfk.a	1pgb.x	1pgd.x	1phh.x	1php.x	1pii.x
1plf.a	1poc.x	1poh.x	1pox.a	1ppa.x	1ppf.e	1ppf.i
1ppn.x	1prc.c	1prc.h	1prc.l	1prc.m	1pts.a	1pya.a
1pya.b	1pyd.a	1rcb.x	1rec.x	1rib.a	1rnd.x	1rop.a
1rve.a	1s01.x	1sac.a	1sbp.x	1ses.a	1sgt.x	1sha.a
1shf.a	1sim.x	1slt.b	1snc.x	1spa.x	1stf.i	1tbe.a
1tca.x	1tie.x	1tml.x	1tnd.a	1tpl.a	1trb.x	1trk.a
1tro.a	1ttb.a	1utg.x	1vaa.a	1vaa.b	1vmo.a	1wht.a
1wht.b	1wsy.a	1wsy.b	1yhb.x	1zaa.c	256b.a	2aai.b
2aza.a	2bop.a	2ccy.a	2cdv.x	2chs.a	2cmd.x	2cp4.x
2cpl.x	2cro.x	2ctc.x	2cts.x	2cyp.x	2dnj.a	2er7.e
2hbg.x	2hhm.a	2hip.a	2hpd.a	2ihl.x	2lh2.x	2liv.x
2mhr.x	2mnr.x	2msb.a	2mta.c	2mta.h	2mta.l	2pf1.x
2pia.x	2pol.a	2por.x	2reb.x	2rn2.x	2rsl.a	2sar.a
2sas.x	2scp.a	2sga.x	2sn3.x	2spc.a	2tgi.x	2tmd.a
2tpr.a	2tsc.a	3aah.a	3aah.b	3adk.x	3b5c.x	3cd4.x
3chy.x	3cla.x	3cox.x	3dfr.x	3eca.a	3gap.a	3gbp.x
3ink.c	3rub.l	3rub.s	3sdh.a	3tgl.x	451c.x	4blm.a
4enl.x	4fgf.x	4gcr.x	4ts1.a	4xis.x	5fbp.a	5p21.x
5tim.a	6fab.h	6fab.l	6taa.x	8abp.x	8acn.x	8atc.a
8atc.b	8cat.a	8i1b.x	8rxn.a	8tln.e	9ldt.a	9rnt.x
9wga.a						

[a] The database was prepared by J. M. Levin and checked for homologous sequences with the help of V. Di Francesco. This database has been modified to restore the total length of the sequences as defined in the SEQRES field of the Protein Data Bank (PDB) file (the DSSP program omits residues whose coordinates are missing in the PDB file, and thus if this occurs in the middle of the polypeptide chain it is split into two or more chains). Residues having no coordinates were assigned the conformation X and were not taken into account for the prediction accuracy although the prediction was done with the whole sequence length. The PDB code is followed by the chain name a, b, c, d, h (heavy), l (light), x (one chain only), e (enzyme), or i (inhibitor).

TABLE II
GLOBAL RESULTS FOR DATABASE PREDICTION

	Observed			
	H	E	C	Total
Predicted				
H	14,460	3094	4790	22,344
E	1124	4965	2089	8178
C	6002	5546	21,496	33,044
Total	21,586	13,605	28,375	63,566
Q_{prd} [a]	64.7	60.7	65.1	
Q_{obs} [b]	67.0	36.5	75.8	
Q_3 [c] = 64.4%				

[a] Number of correctly predicted residues/number of predicted residues.

[b] Number of correctly predicted residues/number of observed residues

[c] Total number of correctly predicted residues/total number of residues.

However, this sometimes leads to predictions that are not physically meaningful, for example, helices having only two residues, or mixtures of strand (E) and helix (H) residues. Several attempts have been made to solve that problem (see Rost and Sander[8] and Zimmermann[19]). Here we added a simple filter after the prediction which requires helices to be at least four residues and strands to be at least two residues. For instance, let us assume that we predict two isolated H residues. We then look for all the possibilities of extension of the two H residues (in this case, there are three possibilities: adding two H's before, adding one H before and one H after, and adding two H's after the predicted H's). We then calculate the product of the probabilities of the different secondary structures for the three segments so defined. The segment that is the most probable is selected leading either to an extension of the helix to four residues or to the suppression of the two isolated residues. Although this filter affects the prediction of particular proteins, on average for the whole database it has no effect on the prediction accuracy; it neither improves nor decreases the percentage of correctly predicted residues.

The global results for the database are shown in Table II. The percentage of correctly predicted residues is 64.4%, and an individual prediction output of the program is given in Table III with an extra column to compare with

[19] K. Zimmermann, *Protein Eng.* **7**, 1197 (1994).

TABLE III
PREDICTION OF EGLIN[a]

Seq	Obs	Prd	pH	pE	pC
T	X	C	0.00	0.00	1.00
E	X	C	0.00	0.02	0.98
F	X	C	0.00	0.08	0.92
G	X	C	0.01	0.13	0.87
S	X	C	0.02	0.14	0.84
E	X	C	0.04	0.24	0.72
L	X	C	0.09	0.33	0.59
K	C	C	0.15	0.24	0.61
S	C	C	0.19	0.16	0.65
F	C	C	0.12	0.12	0.77
P	C	C	0.29	0.12	0.59
E	C	C	0.35	0.26	0.39
V	C	C	0.37	0.30	0.34
V	C	C	0.35	0.24	0.41
G	C	C	0.25	0.20	0.55
K	C	C	0.26	0.27	0.47
T	C	C	0.24	0.34	0.42
V	H	C	0.37	0.19	0.44
D	H	H	0.54	0.08	0.38
Q	H	H	0.58	0.11	0.30
A	H	H	0.61	0.11	0.28
R	H	H	0.56	0.19	0.25
E	H	H	0.50	0.27	0.24
Y	H	H	0.46	0.35	0.19
F	H	H	0.34	0.44	0.22
T	H	H	0.29	0.38	0.32
L	H	C	0.20	0.35	0.44
H	H	C	0.09	0.22	0.69
Y	C	C	0.03	0.11	0.86
P	C	C	0.05	0.06	0.89
Q	C	C	0.09	0.15	0.76
Y	C	C	0.08	0.29	0.63
N	E	C	0.07	0.31	0.61
V	E	E	0.06	0.65	0.30
Y	E	E	0.04	0.75	0.21
F	E	E	0.02	0.76	0.22
L	E	C	0.01	0.30	0.69
P	E	C	0.02	0.09	0.88
E	C	C	0.02	0.03	0.95
G	C	C	0.01	0.01	0.98
S	C	C	0.01	0.02	0.97
P	C	C	0.09	0.12	0.79
V	E	C	0.18	0.33	0.49
T	E	H	0.23	0.51	0.26

TABLE III (*continued*)

Seq	Obs	Prd	pH	pE	pC
L	C	H	0.36	0.46	0.18
D	C	H	0.38	0.33	0.29
L	C	H	0.51	0.17	0.32
R	C	C	0.39	0.16	0.46
Y	C	C	0.33	0.22	0.46
N	C	C	0.24	0.18	0.57
R	E	C	0.20	0.31	0.49
V	E	E	0.17	0.57	0.26
R	E	E	0.12	0.71	0.17
V	E	E	0.07	0.80	0.13
F	E	E	0.05	0.71	0.25
Y	E	E	0.03	0.49	0.48
N	E	C	0.01	0.12	0.87
P	C	C	0.01	0.03	0.96
G	C	C	0.01	0.03	0.96
T	C	C	0.02	0.10	0.88
N	C	C	0.03	0.29	0.68
V	E	E	0.04	0.54	0.41
V	E	E	0.05	0.68	0.27
N	C	E	0.02	0.71	0.27
H	C	E	0.01	0.58	0.41
V	C	C	0.01	0.22	0.78
P	C	C	0.01	0.09	0.91
H	E	C	0.00	0.02	0.98
V	E	C	0.00	0.00	1.00
G	C	C	0.00	0.00	1.00

[a] Amino acid sequence (Seq) of eglin, a subtilisin inhibitor (1cse), with observed conformations (Obs), predicted conformations with filter (Prd), and the probability values pH, pE, and pC for the predicted α helix (H), β strands (E), and coil (C), respectively. The conformation X corresponds to residues for which the crystallographer gave no coordinates. For some residues, for instance, V-13, although the probability for H is higher, the filter assigned a coil (see text). The accuracy of prediction for the three conformations (Q_3) is 73%.

the observed conformations. The result when considering the prediction of individual proteins is 64.7% with a standard deviation of 9.3%. Figure 1a shows the number of proteins as a function of the percentage of correctly predicted residues. Figure 1b is just a check that this distribution does not depart significantly from a Gaussian distribution.

As suggested by Levin,[20] Fig. 2 shows a scatter plot of the percentage of correctly predicted residues as a function of the size of the protein

[20] J. M. Levin, to be published.

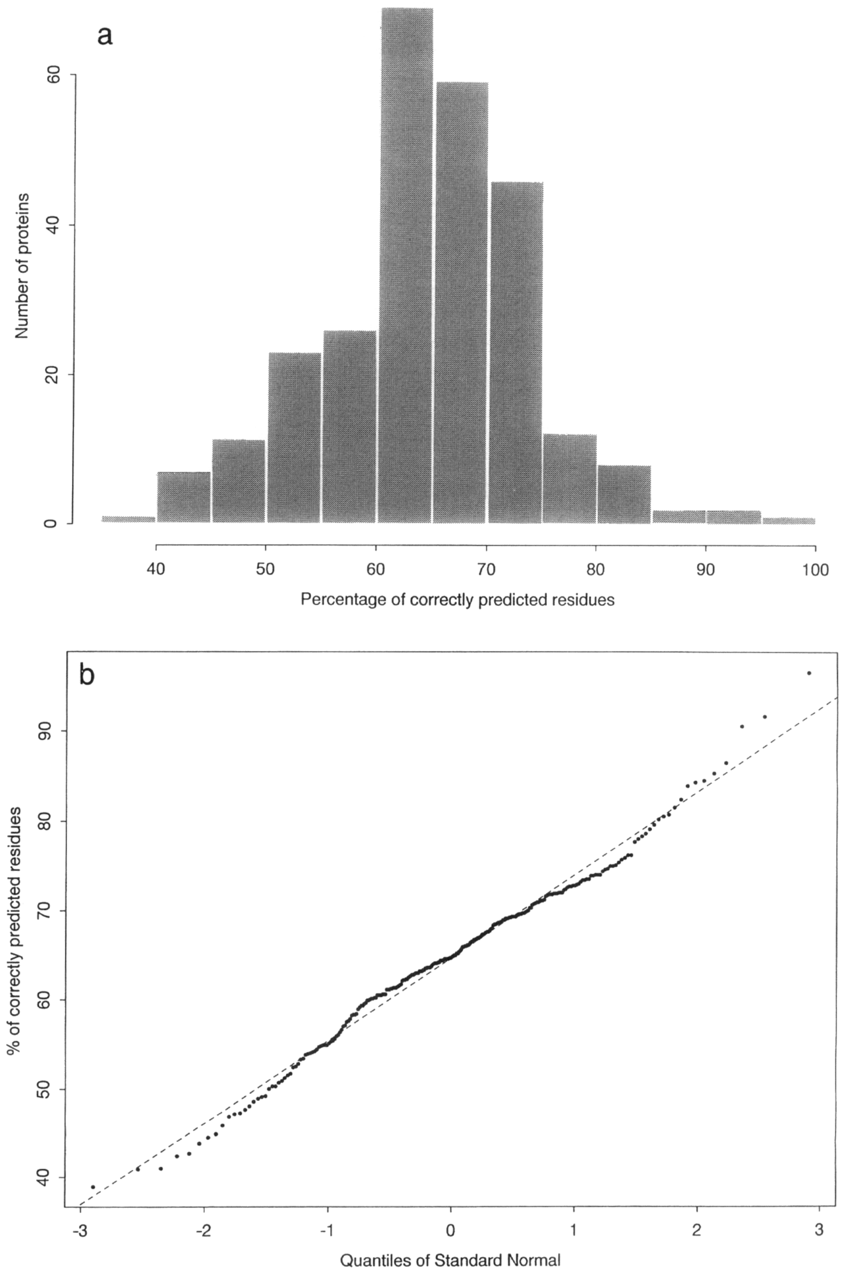

FIG. 1. (a) Histogram of secondary structure prediction accuracies (Q_3) for all the proteins of the database. (b) Normal quantile–quantile plot of the results showing the agreement of the distribution with a normal distribution. The individual values of Q_3 for each protein of the database are sorted according to the quantiles of a normal distribution with a mean of 64.4% and a standard deviation of 9.3%. The dashed line is the least-squares fit of the points.

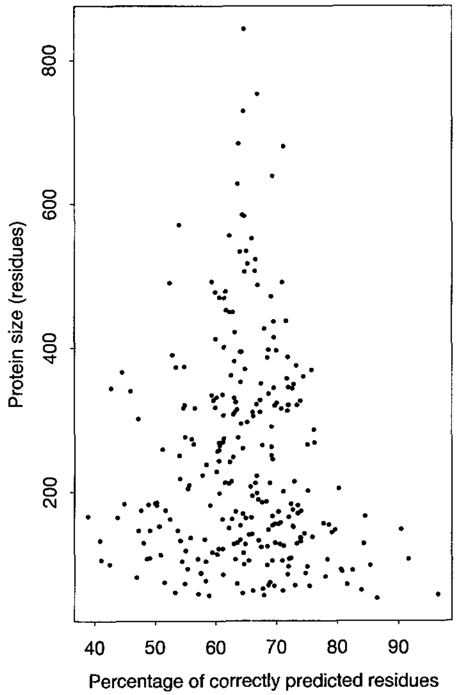

FIG. 2. Distribution of the sequence length (number of amino acid) of the database proteins as a function of the number of correctly predicted residues.

(number of residues). There seems to be no apparent effect of the length of the protein on the accuracy of prediction, except that the longer the protein, the closer the accuracy comes to the average value. For proteins of less than 200 residues the accuracy can lie anywhere between 40 and 90%. A consequence of this observation is that no protein is easier to predict than any other, but rather there are segments of the sequence easier to predict; thus the shorter is the protein, the more likely its accuracy of prediction will be different from the mean, and inversely for the longest ones.

Table IV shows results for the prediction of secondary structure segments, namely, helices and strands. The percentage of correctly predicted segments is given according to the minimum percentage of overlap that is allowed between the predicted and observed segments. For instance, in Table IV, for the row 75% a segment is considered as being correctly predicted if the predicted segment overlaps with at least 75% of the observed segment (counted as the number of residues).

Conclusion

Through the successive incorporation of observed frequencies of single, then pairs of residues on a local sequence of 17 residues, the accuracy of

TABLE IV

Results for Segments: Helices and Strands

	Number of segments			Average length	
	H	E	Total	H	E
Observed	1989	2587	4576	10.9	5.9
Predicted	2148	2043	4191	10.6	4.1

Overlap	H segments	E segments
75%	51.1	23.7
50%	70.0	42.0
25%	75.7	50.2

the GOR method has been improved from about 55% (GOR I using a jackknife[21]) up to 64.4%. The increase of the database size from 67 proteins to the present database of 267 proteins and the use of a more detailed description of the local sequence resulted in an improvement of about 1% (GOR III, Q_3 = 63.3%; GOR IV, Q_3 = 64.4%; the corresponding standard deviations of Q_3 are 0.8% and 0.6%, respectively). However, the result of 63.3% for GOR III was reached using dummy frequencies and adding decision constants that we now believe resulted in a slight bias toward the database we were then using. This causes the overall accuracy of the method to be overestimated by a percent or so, as became apparent when parameters derived from the original database were used to predict new sets of proteins. This is the reason why, here, we avoided the use of decision constants (e.g., to adjust the number of predicted secondary structures to what is observed in the database) in order to obtain a more robust estimation of the accuracy of the method. This small increase in the prediction accuracy is consistent with a previous assumption we made.[15] We estimated that we were able to extract more or less all the information available in the local sequence. Other published methods including neural net methods, using only the protein sequence, are of similar or lower accuracy (see Refs. 2 and 3).

We attributed the limitation in the accuracy of the prediction[15] to the lack of long-distance effects. This appears to be confirmed here by the poor quality of the β-strand prediction. Because β sheets require the pairing of residues that may be distant along the sequence, this secondary structure is presumably more dependent on long-range interactions than are α helices or coils. Clearly the method does not fare well with the prediction of β strands. Although Q_{prd} for β strands is only slightly lower than Q_{prd} for the

[21] W. Kabsch and C. Sander, *FEBS Lett.* **155,** 179 (1983).

two other secondary structures, there is a chronic underprediction of this structure, Q_{obs} for β strands is thus significantly lower compared to Q_{obs} for helices and coils (even taking into account the overestimation of the corresponding Q_{obs} values, which is a consequence of the overprediction of helices and coils). This is also noticeable in the fact that, whereas the average length of predicted and observed helices corresponds closely, the average length of predicted strands is shorter by about one-third compared to the average length of the observed ones.

It is thus very important to consider the possibilities of including long-range interactions in our method (or other methods using only short-range information, that is, local sequence only, for that matter). One way to introduce long-distance effects is to use specific nonlocal pairs to improve β-strand prediction.[22] In other words, the prediction of β strands could be done using the local window as usual to first select putative β-strand segments and then one could slide a window along the sequence to look whether complementary segments to these putative β strand segments can be found. Another possibility is to use multiple alignments. This is based on the assumption that corresponding residues in the alignment, provided the alignment is correct, will have the same secondary structure, being at the same location in the fold. The use of multiple alignments has recently been the source of an improvement of the accuracy ranging from 5, for the GOR[7] and the quadratic logistic[9] methods up to 10 percentage points for a neural network method.[8]

The GOR method has the advantage over neural network-based methods or nearest-neighbor methods in that it clearly identifies what is taken into account for the prediction and what is neglected. Moreover, the method provides estimates of probabilities for the three secondary structures at each residue position, which can be useful for further application of the method.[15] It relies only on observed frequencies in the database; thus, the calculation of the parameters is straightforward and easy to update.

Availability

The corresponding program has been written in C language and currently runs on a platform with a UNIX operating system (but it will run equally well on other operating systems). It can be obtained by anonymous ftp at NCBI (National Center for Biotechnology Information) using the following procedure: ftp ncbi.nlm.nih.gov, move to the directory gibrat/GOR. It is also available at INRA-Jouy-en-Josas: ftp locus.jouy.inra.fr, move to directory/pub/protein/GOR.

[22] T. J. P. Hubbard, *27th Annual Hawaii International Conference on System Sciences,* (L. Hunter, ed.) IEEE Computer Society Press, Los Alamos, 336 (1994).