CABIOS

Vol.4. no.3. 1988
Pages 357 – 365

A simple method for predicting the secondary structure of globular proteins: implications and accuracy

O.Gascuel[1] and J.L.Golmard

Abstract

A method is presented for predicting the secondary structure of globular proteins from their amino acid sequence. It is based on a rigorous statistical exploitation of the well-known biological fact that the amino acid compositions of each secondary structure are different. We also propose an evaluation process that allows us to estimate the capacity of a method to predict the secondary structure of a new protein which does not have any homologous proteins whose structure is already known. This evaluation process shows that our method has a prediction accuracy of 58.7% over three states for the 62 proteins of the Kabsch and Sander (1983a) data bank. This result is better than that obtained by the most widely used methods—Lim (1974), Chou and Fasman (1978) and Garnier et al. (1978)—and also than that obtained by a recent method based on local homologies (Levin et al., 1986). Our prediction method is very simple and may be implemented on any microcomputer and even on programmable pocket calculators. A simple Pascal implementation of the method prediction algorithm is given. The interpretation of our results in terms of protein folding and directions for further work are discussed.

Introduction

The enormous increase in our knowledge of DNA sequences has led to the increased use of protein secondary structure prediction methods from the amino acid sequence. These methods have been dealt with in numerous researches during the last 15 years. In 1983 some important standardization work was done by Kabsch and Sander (1983a,b): the standardization of the data by proposing objective secondary structure assignments from crystallographic data; choice of a data bank containing 62 proteins < 50% homologous; the definition of a benchmark to compare methods; and the comparison of the three most widely used methods: Lim (1974), Chou and Fasman (1978) and Garnier *et al.* (1978).

In 1986 three new methods were published (Sweet, 1986; Nishikawa and Ooi, 1986; Levin *et al*, 1986). These three methods are based on the same hypothesis: short homologous sequences of amino acids have a similar secondary structure, even if they come from non-homologous proteins. In these three methods, the homology has an imprecise meaning, which is between 'close from an evolutionary point of view' and 'having similar amino acid composition'. The best reported result, of these three methods, is that of Levin *et al.* with a prediction accuracy of ~62% over three states for the 62 proteins of the Kabsch and Sander data bank.

Unfortunately, this important improvement over previous methods is mainly due to the presence, in the Kabsch and Sander data bank, of homologous and functionally related proteins, such as chymotrypsin (2GCH) and trypsin (2PTN) which present a homology of 42%. Thus, 62% overestimates the accuracy with which these methods predict the secondary structure of a new protein which does not have any homologous proteins whose structure is already known. However, this problem is particularly important since in the case where the protein to be predicted has an homologue whose structure is already known, its secondary (and tertiary) structure may be derived directly and accurately by aligning.

In this paper we propose a new secondary structure prediction method, which we call GGBSM (Gascuel and Golmard Basic Statistical Method). GGBSM is a purely statistical method which rigorously exploits the well-known biological fact that the amino acid compositions of each type of secondary structure are different. The term 'basic' both expresses the fact that this method is very simple, and also that it may be considered as a starting point for further and more complete approaches to the protein sequences.

We also propose a new way to evaluate the prediction methods which allows us to estimate their ability to predict the secondary structure of a protein which does not have a homologue whose structure is already known. This evaluation process shows that GGBSM has an accuracy over three states of 58.7%. This result is somewhat better than that obtained in the most widely used methods (~50% for Chou and Fasman; ~56% for Lim and Garnier *et al.*) and it is also better than the result obtained by Levin *et al.* (57.5%).

In addition, GGBSM is well balanced: for every state S, the number of residues predicted in the state S is equal to the number of residues observed in the state S. Our method does not have the undesirable feature shown by other methods (Kabsch and Sander, 1983b), which over- or under-predict certain states.

System and methods

All programs described below were developed on a VAX 785.

Unité 194 de l'INSERM, 91 Bd. de l'Hôpital, 75013 Paris, France
[1] *Present address: Centre de Recherche en Informatique de Montpellier, 860 rue de Saint Priest, 34100 Montpellier, France*

GGBSM itself was written in FORTRAN 77. The aligning program used to extract homologies from the Kabsch and Sander data bank and the Levin *et al.* (1976) prediction method were written in VAX LISP (2.1). A Pascal—Microsoft 3.3—version of the GGBSM prediction algorithm (but not the programs for calculating the parameters values) was implemented on an IBM AT and was integrated in the biostation environment (Nanard and Nanard, 1985). A simplified version of this latter program is given later.

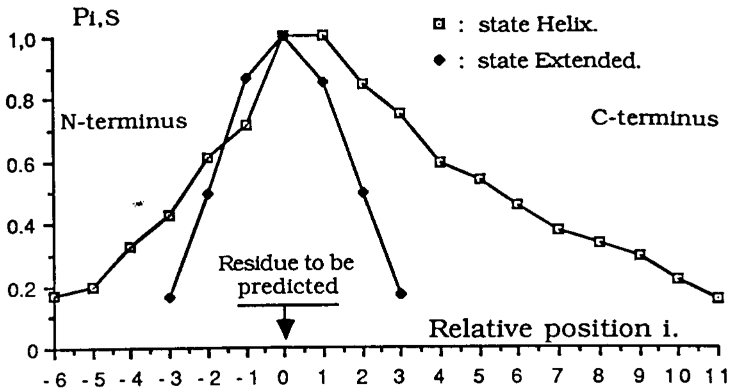

Fig. 1. Importance of the relative positions: □ state *Helix*, ◆ state *Extended*.

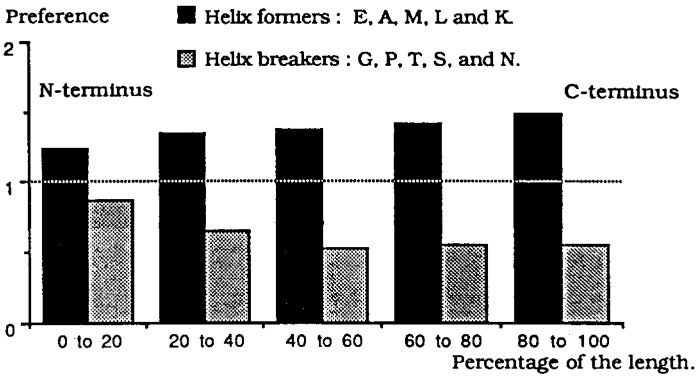

Fig. 2. Histogram of the preferences along helices of the five stronger helix formers and of the five stronger helix breakers, calculated over the 62 proteins of the Kabsch and Sander data bank.

Algorithms

Mathematical model

GGBSM is a 'local' and 'residue-by-residue' method. It determines which state is the most probable for each residue of the protein, on the basis of local considerations on the sequence. In this study, as in most recent studies, only three states were considered: *Helix* (helix 3 or 4), *Extended* and *Coil* (remaining residues or random coil).

The biological concept which underlies GGBSM is that the amino acid composition of each type of secondary structure is different. The problem is reversing this knowledge, in order to express the secondary structure as a function of the local amino acid composition. Numerous mathematical transcriptions are possible, and these form the basis of different methods (e.g. Chou and Fasman, 1978; or Garnier *et al.*, 1978). GGBSM proposes a new solution to this problem. It is based on three sets of parameters whose values, calculated over the 62 proteins of the Kabsch and Sander data bank, are given in the Implementation section.

The parameters $P_{i,S}$. Given R as the residue to be predicted, the parameters $P_{i,S}$ indicate, for the state S, the importance of the position i relative to R. The $P_{i,S}$ are independent of residue type. The intuitive idea is that the state of R is primarily determined by R's own type, slightly less by the types of its two neighbors, again less by the types of the two following residues on both sides, and so on. The $P_{i,S}$ measure this effect. Given a state S, the larger i is (in absolute value), the smaller is $P_{i,S}$. The $P_{0,S}$ which indicate the importance of the central position, taken by R, are arbitrarily fixed to 1. Figure 1 represents the values of these parameters for the state *Helix* and *Extended*, plotted against the relative position i.

The curve of the parameters attached to the state *Helix* is clearly asymmetric. The state *Helix* (versus *not Helix*) of a residue is mostly determined by the type of amino acids located after it, towards the C terminus. This may be explained by the irregular distribution, along helices, of helix formers and breakers (see Figure 2). At the start of the helix, towards the N terminus, the preferences (i.e. the ratio between the frequency in the given region and the average frequency) of helix breakers and formers are close to 1. The gap between these two preferences, which is a measure of propensity to helix, increases when moving from beginning to end and reaches its maximum near the helix end, towards the C terminus. Briefly, propensity to helix is higher at the helix end than at the beginning. When considering amino acids one at a time, distribution histograms are not so simple as those in Figure 2: negatively charged residues, prolines and glycines are mainly located at the start of the helix; positively charged and hydrophobic residues are located at the end (Chou and Fasman, 1974). Finally, all these facts suggest that helix folding preferably starts near the helix end and mostly propagates towards the N terminus.

On the other hand, the curve of the parameters attached to the state *Extended* is symmetrical. The state *Extended* (versus *not Extended*) of a residue R is determined with the same strength by residues located before and after R. This is the result of the amino acid distribution in extended regions which is quite regular and symmetrical. Another difference between *Helix* and *Extended* curves is that the *Extended* curve decreases more quickly when moving away from R. This derives from the average length of extended regions (~4.5 residues) which is smaller than the average length of helices (~9 residues).

The curve of the parameters attached to the state *Coil* appears between the *Helix* curve and the *Extended* curve (to simplify matters, this curve has not been represented on Figure 1, but values of the parameters are given later). The state *Coil* (versus *not Coil*) of a residue is determined more by the type of residues located after it, towards the C terminus, than by the type of residues located before it, towards the N terminus. However, this effect is weaker than that observed in the case of helices. It is due to the amino acid distribution in the random coil which is complex and non-regular. The breadth of the *Coil* curve is in-between the breadth of the *Helix* curve and of the *Extended* curve. This is because the average length of random coil regions ($\simeq$6.5 residues) is intermediate between the other two (see above).

The parameters $I_{t,S}$. These measure the preference for the state S of the amino acid of type t. We define the type of an amino acid either as being a given amino acid, e.g. a Phe or a Trp, or belonging to a given category of amino acids, e.g. a negatively charged amino acid, that will be expressed by being an Asp-Glu. Intuitively, $I_{Pro,H}$ measuring the preference of prolines for helices is low. On the other hand, $I_{Ala,H}$ measuring the preference of alanines for helices is high.

In the GGBSM version presented here, 12 types of amino acids have been empirically retained: Ala, Cys, Asp-Glu, Phe-Trp-Tyr, Gly, His, Ile-Val, Lys-Arg, Leu-Met, Asn-Gln, Pro, Ser-Thr. Amino acids grouped in the same category have common physico-chemical properties and/or have similar preferences for each of the three states.

From a qualitative point of view, these parameters have a meaning close to those of the conformational parameters of Chou and Fasman (1974). Nevertheless certain differences exist: e.g. $I_{His,H}$ which expresses (for GGBSM) the histidine's preference for the state *Helix*, is relatively high, while the corresponding conformational parameter is neutral (1.079 when calculated on the 62 proteins of the Kabsch and Sander data bank). These slight differences arise from the fact that these two types of parameters are very different from a mathematical point of view. This aspect will clearly appear below.

The parameters N_S. Using the parameters N_{Coil}, N_{Helix} and $N_{Extended}$, the method may be adjusted so that when applied to a given set of proteins, it exactly predicts a given percentage of residues in each of the three states. These parameters play a similar role to the decision constants in the Garnier *et al.* (1978) model (GOR model). Only the relative values of these parameters are important. Thus, we have fixed $N_{Coil}=1$. These parameters may be used in two different ways.

(i) They may be used in order to make the method well balanced. Given *T-SET*, the set of proteins of known structure taken into account by the method (in technical terms, *T-SET* is called the training set), the values of the parameters N_S may be adjusted so that for each state S, the number of residues of *T-SET* predicted in the state S is exactly equal to the number of residues of *T-SET* observed in the state S.

(ii) In practice, when trying to determine the secondary structure of a new protein, one often has access to complementary information about the relative abundances of helix, extended and random coil, in this protein. This information may come from physical measurements such as circular dichroism or Raman spectroscopy. It can be more hypothetical and derived from the use of methods for classifying proteins into structural classes, such as Klein and Delisi (1986) or Nakashima *et al.* (1986). Then, the values of the parameters N_S may be adjusted so that when applied to the protein under study, the GGBSM method exactly predicts the desired percentage of residues in each of the three states. Practical outcomes of this interesting feature of the method will be studied in further work.

Let us briefly explain the way in which these parameters are used (more details are given below). Given the sequence $\cdots$ Ala-Asp-Gly-Asp $\cdots$ where the residue to predict R is the first Asp (the Ala is in position−1, the first Asp is in position 0, the Gly is in position 1, the second Asp is in position 2), the prediction is based on the comparison of three numbers, attached to each of the three states: *CBLF(R, Coil), CBLF(R, Helix), CBLF(R, Extended).* Using these three numbers, the probability of each state is calculated. The chosen state is the one with the highest *CBLF* and consequently the highest probability. The *CBLF* are calculated in the following way:

$$CBLF(R, Coil) = N_{Coil}{*}\,(\ldots I_{Ala,C}{*}P_{-1,C} + I_{Asp\text{-}Glu,C}{*}P_{0,C} + I_{Gly,C}{*}P_{1,C} + I_{Asp\text{-}Glu,C}{*}P_{2,C}\ldots)$$

$$CBLF(R, Helix) = N_{Helix}{*}\,(\ldots I_{Ala,H}{*}P_{-1,H} + I_{Asp\text{-}Glu,H}{*}P_{0,H} + I_{Gly,H}{*}P_{1,H} + I_{Asp\text{-}Glu,H}{*}P_{2,H}\ldots)$$

and similarly for *CBLF* (*R*, *Extended*).

GGBSM takes into account, with the use of the parameters $P_{i,S}$, that when predicting R, the R's type is most important and that when moving away from R, the type of residues becomes less and less important. The importance of the relative position is considered as independent of the type of the residue which takes this position. This is intermediate between the two methods commonly used to extract information from sequences.

(i) The first of these methods relies on the calculation of an averaged amino acid composition on a window which slides along the sequence. The method of Hopp and Woods (1981) for locating protein antigenic determinant uses this principle. The method of Chou and Fasman (1978) is also partly based on this principle. Given the sequence above ($\cdot\cdot\cdot$ Ala-Asp-Gly-Asp $\cdot\cdot\cdot$), it is considered that in the window there is one Ala, one Gly and two Asp; the calculation is based on sums such as: $\ldots + W_{Ala,S} + W_{Gly,S} + 2{*}W_{Asp,S}\cdots$ where $W_{t,S}$ expresses the amino acid t's preference for the state S. Relative positions of amino acids included in the window are not taken into account. This model is less complete than the one used by GGBSM and will give most often less good results.

(ii) In the second of these two methods, the importance of the position, for each type of amino acid, is taken into account. The calculation is based on sums such as: . . . + $W_{Ala,-1,S}$ + $W_{Asp,0,S}$ + $W_{Gly,1,S}$ + $W_{Asp,2,S}$ + . . ., where $W_{t,i,S}$ expresses the importance for the state S of an amino acid type t in the relative position i. The GOR method is based on this principle. The method of Nishikawa and Ooi (1980) for predicting the surface-interior diagram of globular proteins is also based on this principle. This implies a more complete model than that used by GGBSM and should theoretically give better results, under certain conditions which we will discuss below.

First, there exists an important difference between GGBSM and both previous methods. They are based on coefficients (called $W_{t,i,S}$ above), which are assumed to be independent and independently evaluated. On the other hand, in GGBSM no such assumption is made and an appropriate statistical method (see below) is used in order to learn globally the optimal values of the whole set of parameters. Thus, roughly speaking, GGBSM's parameters are better estimated than GOR's or Nishikawa and Ooi's coefficients.

We could question whether it would be possible to learn globally the values of the coefficients $W_{t,i,S}$. As observed by Nishikawa and Ooi (1980), the answer would probably be negative because there are too many coefficients: 1020 (three states × 20 amino acids × a window of length 17) for a relatively small amount of data (< 11 000 residues).

Another difference, at a practical level, is that our method requires few parameters: 70 ($32P_{i,S}$ + $36I_{t,S}$ + $2N_S$, details are given below), while a method such as GOR requires many more coefficients: 1022 ($1020W_{t,i,S}$ + two decision constants).

Calculating the parameter values

Let us give some notations, definitions and precisions

(i) The secondary structure is predicted independently for each monomer of the protein.

(ii) *SOS* is the set of states. Here *SOS* = {*Coil, Helix, Extended*}.

(iii) *SOT* is the set of types (see above).

(iv) $F_{R,S}$ is the window in which R is in relative position 0, and which is taken into account when determining whether R is or is not in state S. Here, for the state *Coil* the window is (−3, + 6) in size, for *Helix* (−6, + 11) and for *Extended* (−3, + 3). These sizes have been empirically determined so that for each relative position i inside the window attached to state S, $P_{i,S}$ has sufficient value: $\sim \geq 0.1$ (see above Figure 1 and Implementation). These sizes have also been chosen so that they optimize the method's accuracy.

(v) C is a Boolean function receiving three parameters: a window $F_{R,S}$; a relative position i; and an amino acid type t (see above). $C(F_{R,S}, i,t)$ returns 1 when the amino acid in the relative position i in the window $F_{R,S}$ has type t and otherwise 0.

(vi) $P\ (R,S)$ is the R's probability of being in state S.

(vii) The calculation of the probabilities $P\ (R,S)$ is done using the following formulas:

$$BLF(R,S) = \Sigma\ P_{i,S} * I_{t,S} * C\ (F_{R,S},\ i,\ t)$$

$$\text{(for } i \text{ inside } F_{R,S} \text{ and } t \text{ in } SOT)$$

$$CBLF\ (R,S) = N_S * BLF\ (R,S)$$

where *BLF* means the bilinear form; where *CBLF* means the corrected bilinear form and $P_{i,S}$, $I_{t,S}$ and N_S are as defined above; and

$$P(R,\ S) = \frac{e^{CBLF(R,\ S)}}{\Sigma\ e^{CBLF(R,\ X)}} \quad \text{for every } X \text{ in } SOS$$

which is a very classical way to turn functions into a probability-like form.

At both termini, it is difficult to distinguish a local amino acid composition and to initiate the windows. In addition, the sequence termini are most often flexible and unstructured. Therefore, when R is one of the three first or of the three last residues of the monomer under study, R is predicted as *Coil*. In all other cases the state with the largest probability is chosen.

Calculating the parameters $P_{i,S}$ and $I_{t,S}$. The parameters $P_{i,S}$ and $I_{t,S}$ are calculated independently for each state S. The principle is as follows.

(i) Given a state S, the criterion to be minimized is:

$$\Sigma\ [BLF\ (R,S) - TRUE\text{-}STATE\ (R,S)]^2 \quad \text{for every residue } R \text{ in the training set.}$$

$$TRUE\text{-}STATE\ (R,S) = 1 \text{ when the state of } R \text{ is } S \text{ and otherwise } 0.$$

(ii) The functions called *BLF* are bilinear in relation to the two sets of parameters $P_{i,S}$ and $I_{t,S}$.

(iii) If the parameter values of one of these two sets are fixed, e.g. the $P_{i,S}$ values, one may calculate by the least-squares method (Kendall and Stuart, 1976) the optimal parameter values of the other set, e.g. the $I_{t,S}$ values.

(iv) The calculation of the parameters relies on an iterative use of the above property. The algorithm is as follows: fixing arbitrarily the $P_{i,S}$ values and calculating the $I_{t,S}$; fixing the $I_{t,S}$ to the values calculated previously and calculating the $P_{i,S}$ fixing the $P_{i,S}$, to the values calculated previously and calculating the $I_{t,S}$, etc. until convergence (or similarly: fixing arbitrarily the $I_{t,S}$ values and calculating the $P_{i,S}$; fixing the $P_{i,S}$ to the values calculated previously, etc.).

(v) This algorithm converges rapidly: <10 iterations are necessary. Without additional constraints, it does not give a unique solution. This follows from the fact that *BLF(R,S)* remains equal when multiplying all $P_{i,S}$ by given constant ($\neq 0$) and dividing all $I_{t,S}$ by the same constant. But, when fixing $P_{0,S} = 1$ (this constraint is coherent with the model, see above), the algorithm converges towards a unique minimum. We have not formally demonstrated this, but in practice we have verified that, after shifting the starting point, the same minimum is reached.

Calculating the parameters N_{State}. Once the $P_{i,S}$ and $I_{t,S}$ values have been determinded, the N_{Helix} and $N_{Extended}$ values (N_{Coil} is fixed to 1, see above) are calculated by minimizing the criterion:

$$\Sigma\ [\textit{Number of predicted (S)} - \textit{Number of observed (S)}]^2$$
on the three states: *Coil, Helix* and *Extended.*

The minimization algorithm is the non-linear simplex one (Nelder and Mead, 1965). This algorithm is robust and appeared perfectly suited to our problem.

The evaluation process

The evaluation process we propose estimates the capacity of a method to predict the secondary structure of a new protein which does not have any homologous proteins whose structure is already known. The principle is as follows: for all proteins *P* of the Kabsch and Sander data bank, the true secondary structure of *P* is compared to the secondary structure predicted by the method when removing *P* itself from the data bank, as usual, but removing the proteins which are homologous to *P*. Various other solutions, based on the same idea, are conceivable. In order to facilitate comparisons, the solution we chose is as close as possible to the benchmark of Kabsch and Sander (1983b). In particular, we have used the same data bank and we have not added to it the few new proteins which have been elucidated since 1983 and which are weakly homologous to the proteins already in the data bank. The only small difference, at the level of data, is that for all proteins, the most recent crystallographic data—and secondary structure assignment—were used.

In the present study, we considered that two proteins are homologous when they have an homology > 30%, or when they contain two segments of length > 60 amino acids, with an homology > 30%. Homologies were computed by an algorithm similar to Lipman and Pearson (1985). In order to avoid artefacts, possible with such methods, only functionally related proteins—in a broad sense, very close to that given in Kabsch and Sander (1983a)—were aligned. Nineteen pairs of homologous proteins were found. These are (using the Brookhaven identifiers):

(155C, 3C2C), (155C, 3CYT), (351C, 3C2C), (3C2C, 3CYT),
(1APR, 2AAP), (1CTX, 1NXB), (1ECD, 1LHB), (1LHB, 1MBN),
(1LHB, 2MHB), (1MBN, 2MHB), (1REI, 3FAB), (2ACT, 8PAP),
(1EST, 2ALP), (1EST, 2GCH), (1EST, 2PTN), (2ALP, 2GCH),
(2ALP, 2SGA), (2GCH, 2PTN), (2SGA, 2PTN).

Homologies between monomers of a protein, e.g. between 2MHB α-chain and 2MHB β-chain, are not indicated above because they are not useful for the evaluation process we propose. Indeed, when predicting a given monomor of a protein, the whole protein (all monomers) is removed from the training set.

Such an evaluation process applies to all methods in which there is clearly a training set. Thus, it applies to statistical methods and to pattern-recognition-like methods such as Levin *et al.* (1986). However, it does not apply to expert-system-like methods, such as Lim (1974), Chou and Fasman (1978) or Cohen *et al.* (1983). These methods are based on sets of empirical rules, given by the authors of the methods and theoretically universal (i.e. not only valid for proteins whose structure is known but also for all other proteins). This universality is clearly theoretical because the authors of the methods were inevitably inspired by what they knew. This is well demonstrated for Lim's method, which obtains much better results on proteins whose structure was known before its conception (~ 64%), than on proteins whose structure has been elucidated afterwards (~ 56%) (see Kabsch and Sander 1983b).

What is the difference between this evaluation process and the more classical one called 'jack knife' (Efron, 1982), which is often used? In this test, when predicting the protein *P*, *P* itself is removed from the training set, but not its homologues. Because proteins homologous to *P* are similar to *P* at several levels, an overestimation of the method's capacity to predict *P* usually follows. From a statistical point of view and given a method, results will be better when using the 'jack knife' than when using our evaluation process. However, the gap will very likely be larger for methods based on local homology, such as Levin *et al.* (1986), than for statistical methods such as GOR or GGBSM.

Implementation

A simple version of the GGBSM prediction algorithm is given in Figure 3. It is written in Pascal Microsoft (version 3.3) on an IBM AT. Let us give some additional comments.

MAX-LENGTH is the maximum length of the sequence to be predicted. Here, it is fixed to 5000 amino acids. The program deals with sequences of such length without any problem. Longer sequences may be dealt with by dividing them into subsequences of length ≤5000. The program's speed is ~ 1.5 s CPU for a sequence of 100 amino acids.

CBLF-P is a record which contains all parameters necessary to compute what is called *CBLF(R,S)* above. FIRST is the beginning of the window and LAST the end. Parameter values of $P_{i,S}$, $I_{t,S}$ and N_S are respectively stored in PIS, ITS and NS. CBLF-P-COIL is associated to the state *Coil*, CBLF-P-HELIX to the state *Helix*, etc.

The procedure INITIALIZATION assigns the previous variables to their value. Firstly, using the procedure FF, each amino acid is associated to a numerical type, e.g. alanine (A in single-letter code) is associated to type 1, isoleucine and valine (I and V in single-letter code) which have the same type (see above) are associated to 4. Letters which do not correspond to any amino (X, O, a, etc.) have type 0. The correspondence between the ASCII code of a character and its type is stored in AA-TYPES. Secondly, the CBLF-P's are initialized, e.g.

```
PROGRAM GGBSM(INPUT,OUTPUT);
 CONST MAX_LENGTH = 5000;

 TYPE CBLF_P = RECORD
                FIRST,LAST : INTEGER;
                PIS : ARRAY[-6..+11] OF REAL;
                ITS : ARRAY[0..12] OF REAL;
                NS : REAL;
               END;

 VAR AA_TYPES : ARRAY[33..255] OF 0..12;
     CBLF_P_COIL,
     CBLF_P_EXTENDED,
     CBLF_P_HELIX : CBLF_P;
     SEQUENCE : ARRAY[1..MAX_LENGTH] OF 0..255;
     LENGTH : INTEGER;
     FILE_O : TEXT;

 PROCEDURE INITIALIZATION;
  VAR CHARACTER : INTEGER;
  PROCEDURE FF(C:CHAR;I:INTEGER);
   BEGIN AA_TYPES[ORD(C)]:=I END;
  BEGIN
  FOR CHARACTER:=33 TO 255 DO
   AA_TYPES[CHARACTER]:=0;
  FF('A',1);
  FF('C',2);
  FF('H',3);
  FF('I',4);FF('V',4);
  FF('L',5);FF('M',5);
  FF('G',6);
  FF('P',7);
  FF('F',8);FF('W',8);FF('Y',8);
  FF('S',9);FF('T',9);
  FF('D',10);FF('E',10);
  FF('N',11);FF('Q',11);
  FF('K',12);FF('R',12);
  WITH CBLF_P_COIL DO
   BEGIN
   FIRST:=-3;LAST:=+6;NS:=1.0;ITS[0]:=0.0;
   ITS[ 1]:= 0.018074;ITS[ 2]:= 0.073735;
   ITS[ 3]:= 0.071903;ITS[ 4]:=-0.03926 ;
   ITS[ 5]:=-0.018393;ITS[ 6]:= 0.294359;
   ITS[ 7]:= 0.445666;ITS[ 8]:= 0.012034;
   ITS[ 9]:= 0.16152 ;ITS[10]:= 0.17015 ;
   ITS[11]:= 0.187471;ITS[12]:= 0.109817;
   PIS[-3]:= 0.154212;PIS[-2]:= 0.434159;
   PIS[-1]:= 0.706102;PIS[ 0]:= 1.0      ;
   PIS[ 1]:= 0.915659;PIS[ 2]:= 0.559442;
   PIS[ 3]:= 0.336078;PIS[ 4]:= 0.109662;
   PIS[ 5]:= 0.098532;PIS[ 6]:= 0.098014;
   END;
  WITH CBLF_P_HELIX DO
   BEGIN
   FIRST:=-6;LAST:=+11;NS:=1.2311;ITS[0]:=0.0;
   ITS[ 1]:= 0.147418;ITS[ 2]:= 0.024205;
   ITS[ 3]:= 0.114178;ITS[ 4]:= 0.037925;
   ITS[ 5]:= 0.111157;ITS[ 6]:=-0.081108;
   ITS[ 7]:=-0.144076;ITS[ 8]:= 0.063689;
   ITS[ 9]:=-0.031711;ITS[10]:= 0.050114;
   ITS[11]:= 0.01434 ;ITS[12]:= 0.093082;
   PIS[-6]:= 0.169079;PIS[-5]:= 0.195383;
   PIS[-4]:= 0.326344;PIS[-3]:= 0.430292;
   PIS[-2]:= 0.610814;PIS[-1]:= 0.711503;
   PIS[ 0]:= 1.0     ;PIS[ 1]:= 0.996892;
   PIS[ 2]:= 0.845693;PIS[ 3]:= 0.74947 ;
   PIS[ 4]:= 0.589796;PIS[ 5]:= 0.538197;
   PIS[ 6]:= 0.458475;PIS[ 7]:= 0.374935;
   PIS[ 8]:= 0.332311;PIS[ 9]:= 0.291095;
   PIS[10]:= 0.221007;PIS[11]:= 0.159564;
   END;
  WITH CBLF_P_EXTENDED DO
   BEGIN
   FIRST:=-3;LAST :=+3;NS:=1.5451;ITS[0]:= 0.0;
   ITS[ 1]:=-0.005853;ITS[ 2]:= 0.096215;
   ITS[ 3]:= 0.016752;ITS[ 4]:= 0.214617;
   ITS[ 5]:= 0.098072;ITS[ 6]:= 0.01415 ;
   ITS[ 7]:=-0.078651;ITS[ 8]:= 0.124579;
   ITS[ 9]:= 0.079342;ITS[10]:=-0.062711;
   ITS[11]:=-0.001842;ITS[12]:=-0.014743;
   PIS[-3]:= 0.16148 ;PIS[-2]:= 0.498551;
   PIS[-1]:= 0.864631;PIS[ 0]:= 1.0      ;
   PIS[ 1]:= 0.850783;PIS[ 2]:= 0.49398 ;
   PIS[ 3]:= 0.169746;
   END;
  END;

 PROCEDURE READ_SEQUENCE;
  VAR FILE_I : TEXT;
      NAME : STRING(80);AA : CHAR;
  BEGIN
  LENGTH:=0;
  WRITE('Name of the sequence');READLN(NAME);
  ASSIGN(FILE_I,NAME);RESET(FILE_I);
  REPEAT
   READ(FILE_I,AA);
   IF AA>' ' THEN
    BEGIN
    LENGTH:=LENGTH+1;SEQUENCE[LENGTH]:=ORD(AA)
    END;
  UNTIL EOF(FILE_I) OR (LENGTH=MAX_LENGTH);
  CLOSE(FILE_I);
  END;

 PROCEDURE ASK_OUTPUT;
  VAR NAME : STRING(80);
  BEGIN
  WRITELN;
  WRITE('Name of the output file');
  READLN(NAME);
  ASSIGN(FILE_O,NAME);REWRITE(FILE_O);
  END;

 FUNCTION C_SUM(POS:INTEGER;CBLF:CBLF_P):REAL;
  VAR TOTAL : REAL;
      INDICE,FIRST_P,LAST_P : INTEGER;
  BEGIN
  WITH CBLF DO
   BEGIN
   TOTAL:=0.0;
   FIRST_P:=FIRST+POS;LAST_P:=LAST+POS;
   IF FIRST_P<1 THEN FIRST_P:=1;
   IF LAST_P>LENGTH THEN LAST_P:=LENGTH;
   FOR INDICE:=FIRST_P TO LAST_P DO
    TOTAL:=TOTAL+PIS[INDICE-POS]*
            ITS[AA_TYPES[SEQUENCE[INDICE]]];
   C_SUM:=TOTAL*NS;
   END;
  END;

 PROCEDURE SECONDARY_STRUCTURE_PREDICTION;
  VAR ECS_COIL,ECS_HELIX,
      ECS_EXTENDED,TECS : REAL;
      P : INTEGER;
  BEGIN
  FOR P:=1 TO LENGTH DO
   BEGIN
   ECS_COIL:=EXP(C_SUM(P,CBLF_P_COIL));
   ECS_HELIX:=EXP(C_SUM(P,CBLF_P_HELIX));
   ECS_EXTENDED:=EXP(C_SUM(P,CBLF_P_EXTENDED));
   TECS:=ECS_COIL+ECS_HELIX+ECS_EXTENDED;
   WRITE(FILE_O,P:5,CHR(SEQUENCE[P]):2);
   IF ((ECS_COIL>=ECS_HELIX) AND
       (ECS_COIL>=ECS_EXTENDED))
      OR (P<=3) OR (P>LENGTH-3)
   THEN WRITE(FILE_O,'  C  ')
   ELSE IF ECS_HELIX>=ECS_EXTENDED
        THEN WRITE(FILE_O,'  H  ')
        ELSE WRITE(FILE_O,'  E  ');
   WRITELN(FILE_O,
      ECS_COIL/TECS:3:4,'  ',
      ECS_HELIX/TECS:3:4,'  ',
      ECS_EXTENDED/TECS:3:4);
   END;
  END;

 BEGIN
 INITIALIZATION;
 READ_SEQUENCE;
 ASK_OUTPUT;
 SECONDARY_STRUCTURE_PREDICTION;
 CLOSE(FILE_O);
 END.
```

Fig. 3. A simple Pascal version of the GGBSM secondary structure prediction algorithm.

(i) the preference of alanine for the state *Coil*, called $I_{Ala,\ C}$ above, corresponds to ITS[1] (alanine has the numerical type 1) inside CBLF-P-COIL and has a value of 0.018074; (ii) the importance of position 5 for the state *Helix*, called $P_{5,H}$ above, corresponds to PIS[5] inside CBLF-P-HELIX and has a value of 0.538197; (iii) the decision constant of the state *Extended*, called $N_{Extended}$ above, corresponds to NS inside CBLF-P-EXTENDED and has a value of 1.5451.

The sequence (in single-letter code) to be predicted is read by the procedure READ-SEQUENCE and stored in SEQUENCE which is an array of characters, given in ASCII code. Special characters (ASCII code <33) such as carriage return or space are not read. LENGTH is the length of the sequence to be predicted.

The function C-SUM has two parameters: a residue number, POS, and a CBLF-P, CBLF. It returns what is called *CBLF(R,S)* above, *R* being the residue of number POS and *S* the state corresponding to CBLF.

Predictions are computed by the procedure SECONDARY-STRUCTURE-PREDICTION exactly as defined above. They are stored in a file whose name is given by the user during the execution of the procedure ASK-OUTPUT. Each line of this file corresponds to one residue and contains the residue number, the amino acid, the state which is predicted and probabilities of the three states *Coil, Helix* and *Extended* respectively.

Discussion

Table I contains comparative results of five methods: (1) Lim (1974), (2) Garnier *et al* (1978), (3) Chou and Fasman (1978), (4) Levin *et al* (1986) and (5) GGBSM.

Results of Lim's method are those published by Kabsch and Sander (1983b). As indicated above, the result averaged over all proteins, 59.3%, is a very optimistic estimation of the method's accuracy. A better estimation is the result it obtains on the post-1974 proteins: 56.3%.

GOR's and Chou's results are also those published by Kabsch and Sander (1983b). Although their evaluation process is different to ours, one may consider that in the case of these two methods the difference is low (see above) and that the Kabsch and Sander results give a good estimation of the accuracy.

Results of the two last methods were obtained using the evaluation process described above. Method 4 has been rewritten by ourselves as described in Levin *et al.* (1986). Protein-by-protein results of these last two methods are available on request.

GGBSM is the only method to be well balanced. The number of residues predicted in each of the states is very close to the number of residues observed in this state. These numbers are not exactly equal because results given here are obtained through a true evaluation process in which the training set is modified for each predicted protein. The interest of a well balanced method appears at two levels.

(i) At the practical level. Interpreting predictions of a well-balanced method is easier. When using a method that is not well balanced, one has to correct mentally (and roughly) the defects of the method. When the method predicts too many *Coil* and not enough *Extended*, one has to correct this and one has to suppose that some of the predicted *Coil* are actually *Extended*.

(ii) At the performance level. For the same percentage of well-predicted residues, a well-balanced method is usually better than one which is not. Indeed, the size of the classes are very unequal: *Coil* ≈ 50%, *Helix* ≈ 30% and *Extended* ≈ 20%. When a method, e.g. Lim's predicts too many *Coil* (59.4% for Lim's) and not enough *Extended* (15.7% for Lim's), it takes fewer risks and its accuracy is overestimated.

The results of every method fluctuate widely among proteins, e.g. for GGBSM, 27% of well predicted residues in the case of melittin (1MLT) and 81% in the case of avian pancreatic polypeptide (1PPT). Thus, the result averaged over all

Table I. Predictive success of five methods: details of *Coil (C)*, *Helix (H)* and *Extended (E)* predictions averaged over all proteins of the Kabsch and Sander data bank

	(1) Lim			(2) GOR			(3) Chou			(4) Levin			(5) GGBSM		
Predicted/observed matrix (percentage of the total number of residues)															
	Observed			Observed			Observed			Observed			Observed		
	C	*H*	*E*	*C*	*H*	*E*	*C*	*H*	*E*	*C*	*H*	*E*	*C*	*H*	*E*
Predicted *C*	37.4	11.3	10.7	28.4	7.0	6.4	26.9	8.2	6.3	33.3	8.8	7.8	34.9	8.8	7.5
Predicted *H*	7.6	14.3	3.0	11.2	15.9	3.4	10.8	11.9	3.9	11.6	16.8	4.9	9.6	15.3	4.1
Predicted *E*	5.3	2.7	7.6	10.8	5.3	11.6	12.6	8.3	11.1	5.9	3.5	7.4	6.3	5.0	8.5
Success		59.3%			55.9%			49.9%			57.5%			58.7%	
Predicted/observed total (%)															
	C	*H*	*E*	*C*	*H*	*E*	*C*	*H*	*E*	*C*	*H*	*E*	*C*	*H*	*E*
Predicted	59.4	25.0	15.7	41.8	30.5	27.6	41.4	26.5	32.0	49.9	33.2	16.8	51.1	29.0	19.9
Observed	50.4	28.3	21.3	50.4	28.3	21.3	50.4	28.3	21.3	50.9	29.0	20.1	50.9	29.0	20.1

proteins—58.7% of well-predicted residues for GGBSM—does not have any real meaning. It is only of interest in the framework of a comparison with other methods, tested on the same proteins, through the same evaluation process.

Our method, like other comparable methods, will fair badly when applied to membrane-associated proteins, because there is no membrane-associated protein in the Kabsch and Sander data bank. Thus, the method's parameters are calculated under the assumption that proteins fold in water and they are not applicable outside this assumption. In the same way, our method is not reliable when applied to proteins which depart from the secondary structure paradigm, e.g. the disulphide-bridge-rich proteins.

There has long been a general consensus which considers it impossible to obtain a 100% secondary structure prediction. A global folding theory considering tertiary interactions would be needed. However, the different attempts to solve the final problem of conformation prediction for globular proteins from sequence alone have as yet been disappointing. A line of approach seems to try to arrange secondary structures into the form of super-secondary structures and domains (Cohen *et al.*, 1983; Taylor and Thornton, 1983). But such approaches are based on secondary structure prediction methods. Thus, we go back to the necessity of developing and of improving these methods. More generally, in our opinion, the problem of conformation prediction has to be tackled hierarchically, step by step and by successive approximations. Such an approach has been successfully applied (Lathrop *et al.*, 1987) on the related problem of matching proteins which are close at functional level but distant from an evolutionary point of view. This way, accuracy of the initial steps—secondary structure prediction would obviously be one of them—is of prime importance since final prediction would strongly depend on them.

Outside these prospective considerations, secondary structure prediction methods bring to light a large amount of information concerning protein sequences which may be used in various ways such as antigenic site prediction or active site prediction. Specially interesting are statistical methods like GGBSM because they allow spectra to be plotted where the abscissa forms the sequence and the ordinate supports the probabilities of each of the states (see in particular Ptitsyn and Finkelstein, 1983). Such spectra contain much more information than the simple result of the decision process which consists of choosing the most probable state. They not only contain the decision itself but also the confidence one may have in this decision. They may also be used to compare sequences when the aim is detecting structural or functional similarities.

GGBSM's qualities derive from the simplicity of the mathematical model we used. Because this model is based on few parameters (70 free parameters), we might employ an efficient learning method allowing the optimal values of the parameters to be globally estimated. This reasonable parametrization also causes the method to be robust, i.e. results are very close when directly predicting the training set (59.9% on the Kabsch and Sander data bank) and when testing the method using the process we propose (58.7% on the same data). This highlights the almost complete independence between the method and the data bank used as the training set: predictions on new proteins will not be much modified by modifications of the data bank. Finally, this small parameterization and the simplicity of the calculations present a practical advantage: GGBSM may be implemented on any microcomputer and even on programmable pocket calculators.

We do not think that GGBSM itself may be greatly improved. Very surprizingly (at least for us), when shifting amino acid regroupings (or types), very small changes in the results are observed. For example, when shifting from the 12 amino acid types mentioned above, to 13 types where glutamic acid and aspartic acid have been separated, almost no change is observed. However, glutamic acid's and aspartic acid's preferences for each state are quite different (Chou and Fasman, 1974). In the same way, a 20-type solution (one type for each amino acid) does not lead to better results. Our interpretation is that local amino acid compositions are highly dependent on one another [this somewhat contradicts the Garnier *et al.*'s (1978) assumption of independence of the coefficients, see above]. Information carried by the sequence is in part redundant. The consequences are: (i) it is possible, as we do, to group certain amino acids two by two or three by three; (ii) an imperfect choice of amino acid regroupings does not cause too serious consequences. Similarly, performance does not seem to be very sensitive to the choice of the sizes of the windows attached to each state (see above). Finally, we estimate that no more than 1−2% of accuracy improvement may be obtained by optimizing the empirical choices presented in this paper.

The counterpart of the mathematical model's simplicity is an excessive simplification of the biological model. Certain determinants of the secondary structure are not taken into account. We think in particular of: (i) the hydrophobic moment of the whole protein sequence, which has a predictive power on the relative abundances of helix and extended chain (Eisenberg *et al.*, 1984); (ii) the very particular and asymmetric amino acid distribution in helices (Chou and Fasman, 1974), which currently is only globally taken into account by the method. In addition, the final prediction does not have a global consistency, since it may comprise unobserved state chainings such as: ··· *Extended − Helix − Extended − Helix* ··· or anything else.

The good results obtained by GGBSM seem to indicate that the above-mentioned determinants are somewhat less important than those on which this method is based. Nevertheless, taking these determinants into account is, in our opinion, the best way to improve our method. Two complementary approaches may be considered.

(i) Using GGBSM as a starting point and building on it. As suggested above, coupling GGBSM with a method for classifying proteins into structural classes such as Klein and Delisi (1986), using among other things the hydrophobic moment, would allow the decision constants N_S to be particularized and therefore the method's accuracy to be improved. A method for smoothing and interpretting the probability spectra, conceivably based on certain ideas contained in the works of Chou and Fasman (1978), Cohen *et al.* (1983) and Taylor and Thornton (1983), would result in consistent predictions and better results.

(ii) Searching for new statistical methods which have GGBSM's qualities (i.e. small parametrization and global estimation of the parameters) but taking a more complete structural model into account.

Acknowledgements

With sincere thanks to Nicky, Bruno Bachimont, Olivier Bacquet, François Colonna Cesari, Antoine Danchin. Philippe Marlière, Chris Sander, William Saurin and Pierre Zweigenbaum for help in starting this work, technical assistance, advice, encouragement and plenty of meaningful discussions.

References

Chou,P.Y. and Fasman,G.D. (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**, 211−222.

Chou,P.Y. and Fasman,G.D. (1978) Empirical prediction of protein conformation. *Annu. Rev. Biochem.*, **47**, 251−276.

Cohen,F.E., Abarbanel,R.M., Kuntz,I.D. and Fletterick,R.J. (1983) Secondary structure assignment for alpha/beta proteins by a combinatorial approach. *Biochemistry*, **22**, 4894−4904.

Efron,B. (1982) *The Jack Knife, the Bootstrap and other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia.

Eisenberg,D., Weiss,R.M. and Terwilliger,T.C. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA*, **81**, 140−144.

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

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA*, **78**, 3824−3828.

Kabsch,W. and Sander,C. (1983a) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577−2637.

Kabsch,W. and Sander,C. (1983b) How good are predictions of protein secondary structure? *FEBS Lett.*, **155**, 179−182.

Kendall,M. and Stuart,A. (1976) *The Advanced Theory of Statistics*. Vol. **2**. Charles Griffin, London.

Klein,P. and Delisi,C. (1986) Classifying proteins into structural groups. *Biopolymers*, **25**, 1659−1672.

Lathrop R.H., Webster T.A. and Smith T. (1987) ARIADNE: Pattern-directed inference and hierarchical abstractions in protein structure recognition. *ACM Commun.*, **30**, 909−921.

Levin,J., Robson,B. and Garnier,J. (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.*, **205**, 303−308.

Lim,V.I. (1974) Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J. Mol. Biol*, **88**, 873−894.

Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435−1441.

Nakashima,H., Nishikawa,K. and Ooi,T. (1986) The folding type of a protein is relevant to the amino acid composition. *J. Biochem.*, **99**, 153−162.

Nanard,M. and Nanard,J. (1985) A user-friendly biological workstation. *Biochimie*, **67-5**, 429−432.

Nelder,J.A. and Mead,R. (1965) A simplex method for function minimization. *Computer J.*, **7**, 308−313.

Nishikawa,K. and Ooi,T. (1980) Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int. J. Peptide Protein Res.*, **16**, 19−32.

Nishikawa,K. and Ooi,T. (1986) Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochim. Biophys. Acta*, **871**, 45−54.

Ptitsyn,O.B. and Finkelstein,A.V. (1983) Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*, **22**, 15−25.

Sweet,R.M. (1986) Evolutionary similarity among peptide segments is a basis for prediction of protein folding. *Biopolymers*, **25**, 1565−1577.

Taylor,W.R. and Thornton,J.M. (1983) Prediction of super secondary structure. *Nature*, **301**, 540−542.

Received on December 15, 1987; accepted on April 12, 1988