ANALYTICAL BIOCHEMISTRY 163, 292–297 (1987)

A Computerized Version of the Chou and Fasman Method for Predicting the Secondary Structure of Proteins

G. DELEAGE,*,[1] B. TINLAND,† AND B. ROUX*

*Laboratoire de Physico-Chimie Biologique, LBTM, CNRS, U.M. 380024, and †Laboratoire de Physico-Chimie, Université Claude Bernard Lyon I, 43, boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France

Received July 8, 1986

A novel computerized program has been developed for predicting the secondary structure of proteins from their amino acid sequences. The scheme of the Chou and Fasman method (1978, *Adv. Enzymol. Relat. Subj. Biochem.* **47**, 45–148) is closely followed. Some of their qualitative rules have been converted to numeric scales to obtain unambiguous predictions. This program has been tested on 21 proteins with known three-dimensional structures constituting a 4457 amino acids data base. The percentage of correctly predicted amino acids is between 41 and 66% for a three-state (helix, sheet, and coil) description of protein secondary structure. © 1987 Academic Press, Inc.

KEY WORDS: computer methods; protein secondary structure.

Over the last 10 years, the prediction of the secondary structure of proteins has been extensively developed and applied to many proteins. The most commonly used methods are statistical (1,2). The Chou and Fasman method (2) uses the known three-dimensional structures of soluble proteins to calculate the frequency of occurrence of amino acid residues in secondary structures such as α helix, β sheet and, β turn. This method, which has been revealed as one of the most accurate, is based on rules which are rather difficult to computerize. This problem was approached by Chou and Fasman, who proposed a computerized predictive method (2) further published by Corrigan and Huang (3). Unfortunately, these computerized predictions do not respect any of the "qualitative" rules and are essentially based on the comparison among P_α, P_β, P_t parameters (2). For example, the helix and sheet nucleation conditions (A1 and B1 from Ref. (2)) and the changes in conformational assignments given to Pro, Asp (near the N terminal) and Arg (near the C terminal) for searching potential helix nuclei are not taken into account. These questionable facts lead to predictions of an α helix shorter than four residues, as the numerous ones recently reported for HTLV III (4), and to a concomitant loss in accuracy (5).

In view of these ascertainments, we report here a computer program which uses Chou and Fasman P_α, P_β, P_t parameters and follows as near as possible their predictive rules and scheme. This program does not predict helices shorter than four residues and presents a good mean accuracy. The results are compared with those obtained by Chou and Fasman (2) and by Corrigan and Huang (3).

MATERIALS AND METHODS

The main features of the program are summarized in the following steps and a typical output is given in Fig. 1 for bovine trypsin inhibitor. The Chou–Fasman $P(N)$ parameters (PA, PB, PT) are averaged on all possible tetrapeptides and listed, as are the α and β

[1] To whom correspondence should be addressed.

0003-2697/87 $3.00

```
PROTEIN: BOVINE TRYPSIN INHIBITOR
RPDFCLEPPY TGPCKARIIR YFYNAKAGLC QTFVYGGCRA KRNNFKSAED CMRTCGGA
AMINO-ACID NUMBER: 58

***TETRAPEPTIDES***

 AA   PA   IA   PB   IB   PT    FT
 1   .923   i  .85    i  1.132  2.451     *
 2   .853   B  .915   B  1.193  .934   *
 3  1.013   I 1.103   B  .96    .494   *
 4  1.138   b 1.06    b  .78    .072   *
 5   .998   i  .853   b  1.01   .195   *
 6   .965   H  .692   h  1.093  .085   *
 7   .835   H  .735   B  1.23   .714    *
 8   .665   B  .94    B  1.285  2.765       *
 9   .665   B  .99    B  1.295  .655    *
10   .665   b  .99    H  1.295  1.144    *
11   .668   i  .92    h  1.308  .318    *
12   .75    B  .808   b  1.32   3.413       *
13   .963   B  .828   B  1.095  .226   *
14  1.065   i  .923   h  .953   .51    *
15  1.14    h 1.025   b  .772   .232   *
16  1.14    H 1.24    i  .638   .046   *
17  1.03    i 1.265   i  .71    .026   *
18  .958    h 1.4     H  .758   .181   *
19  .97     h 1.345   H  .79    .338   *
20  .873    i 1.313   i  .957   .37    *
21  .795    b 1.303   H  1.11   .349   *
22  .978    h 1.143   h  .99    .425   *
23  .985    b  .982   H  1.093  .226   *
24 1.168    b  .823   i  .973   .511    *
25 1.143    H  .788   i  .973   .367   *
26 1.09     h  .905   b  .955   .556    *
27  .975    H 1.018   i  1      .235   *
28  .898    B 1.085   b  1.08   .292   *
29  .943    H 1.195   h  .93    .095   *
30  .943    i 1.215   h  .933   .617    *
31 1.033    h 1.343   h  .76    .275   *
32  .928    i 1.435   h  .8     .123   *
33  .863    b 1.325   h  .95    .491   *
34  .723    h 1.168   H  1.19   1.164    *
35  .632    b 1.04    H  1.363  1.695      *
36  .705    B  .905   b  1.315  .862    *
37  .918    B  .925   b  1.09   .31    *
38 1.065    i  .923   h  .953   .525   *
39 1.135    i  .858   i  .893   .326   *
40 1.058    H  .848   i  1.045  .622    *
41  .87     h  .843   b  1.27   1.013    *
42  .863    i 1.023   i  1.168  .721    *
43  .907    b  .975   i  1.183  .825    *
44  .933    b  .94    i  1.15   .504    *
45 1.12     h  .925   h  .925   .492   *
46 1.215    h  .673   b  .96    .171   *
47 1.178    i  .623   b  1.073  .569    *
48 1.16     H  .733   i  1.013  .825    *
49 1.168    H  .788   B  .998   .396   *
50 1.035    I  .928   B  1.05   .093   *
51  .99     i 1.09    h  .925   .956    *
52  .99     H 1.09    h  .925   .6      *
53  .77     i 1.015   i  1.165  1.344    *
54  .668    i  .97    h  1.318  1.316    *
55  .815    i  .88    h  1.243  1.396    *
56  .64     B  .583   b  .945   0      *
57  .497    B  .395   b  .555   0      *
58  .355    H  .208   i  .165   0      *

*** POSSIBLE HELIX NUCLEI ***

   15-20
   25-30 Best
   44-49
   45-50
   48-53 Best

*** LINKED HELIX NUCLEI ***

   15-20  PA= 1.12   PB= 1.11
   25-30  PA= 1.08   PB= .94
   44-50  PA= 1.1    PB= .79
   48-53  PA= 1.18   PB= .82

*** POSSIBLE SHEET NUCLEI ***

   17-21
   18-22 Best
   19-23
   20-24
   21-25
   28-32
   29-33
   30-34
   31-35 Best
   32-36
   51-55

*** LINKED SHEET NUCLEI ***

   17-25  PA= .97 PB= 1.23
   28-36  PA= .87 PB= 1.2
   51-55  PA= .93 PB= 1.11

*** FINAL RETAINED STRUCTURE ***

   17-24 SHEET
   28-36 SHEET
   44-53 HELIX

*** TURNS RETAINED ***

   1-4    8-11    10-13    12-15
   41-44  53-56   54-57    55-58

*** MAXIMAL LENGTH OF HELICES ***

   1-7    PA= 1.02 PB= .89
   39-54  PA= 1.05 PB= .89

*** MAXIMAL LENGTH OF SHEETS ***

   17-23  PA= .95  PB= 1.34
   28-35  PA= .91  PB= 1.26
```

FIG. 1. Program output for bovine trypsin inhibitor. Under the IA and IB columns, strong formers are designated by *H*, formers by *h*, weak formers by *I*, indifferent by *i*, breackers by *b*, strong breackers by *B* for, respectively, helix and sheet assignements of numbered amino acids (column AA). The PA, PB, and PT columns correspond to the average values on tetrapeptides of P_α, P_β, and P_{turn} parameters, respectively. The FT column shows the probability of turn beginning at residue *i* based on the product of the frequency of the residues in each position in the tetrapeptide. The position of the star is moved on the right when the FT value increases. In the right side of the listing, the results concerning each section is given. "Best" noticed the nuclei which exhibits the highest potential (see Material and Methods) in case of overlapping nuclei. PA and PB designate the average value over the concerned segment.

assignments of amino acids. The β turn frequency parameters are also given with an optional threshold chosen for discriminating the most probable tetrapeptides presenting a β turn structure. A tetrapeptide is kept as a β turn if $(P_t) > (P_\alpha)$ and $(P_t) > (P_\beta)$ and $F(T) > 1.10^{-4}$ in a coil region (see below). The β turn values are visualized as stars leading to a profile which allows easy detection of high values. If the threshold for $F(T)$ is nil, all the tetrapeptides are printed as in Fig. 1.

The second step is the search for helix and β sheet nuclei. For helices, the averaged values of P_α and P_β parameters, calculated over the whole sequence with a constant run window of six residues, are compared with the critical value of 1.03. The helical assignments are also taken into account to fulfil the following helix nucleation condition: a weak helix former (I_α) counts as half of a former (h_α, H_α) in the segments. The helical assignments for Pro (B_α) near the N terminal and for Arg (i_α) near the C terminal of the putative helix are changed into I_α to find the four formers' residues initiating the helix. All the helix nuclei are listed and a subroutine is activated to select the nucleus with the higher helical potential in case of overlapping nuclei (indicated by "Best" in Fig. 1). Then the helices with five common successive residues are linked into a longer helix.

In order to find the β sheet nuclei the complete sequence is examined again with similar calculations using a constant run window of five residues. In this case, the critical value of (P_β) is 1.05 and the minimal number of former residues is three. All the β sheet nuclei with four common successive residues are linked together into a longer sheet.

After these descriptive steps the overlapping procedure begins. In the case of α/β overlapping regions, the mean value of P_α and P_β is calculated over the ambiguous segment. In addition, as recommended by Chou and Fasman (2), the conformational assignments are also considered. For this purpose, a numeric conversion has been established; i.e., $H = 10$, $h = 5$, $I = 2$, $i = 1$, $b = -2$, $B = -5$ to compare the α and β assignments over the overlapping region. Thus a segment with $(H_2h_2ib)_\alpha$ and $(Hh_3iB)_\beta$ assignments will have a score of 29 for α assignment (A_α) and 21 for β assignment (A_β). Thus, if $(P_\alpha) > P(\beta)$ and $A_\alpha \geqslant A_\beta$, the region is predicted as helical; if $(P_\alpha) \leqslant (P_\beta)$ and $A_\alpha > A_b$, then (P_α) must be superior to 1.05 to predict the region as helical. In all other cases, the segment is predicted as β. A typical example of the utility of such a procedure is shown for the 44–50 and 48–53 possible helices and the 51–55 putative sheet of trypsin inhibitor. The final structure retained was a 53–58 helix, in agreement with X-ray data.

When the overlaps for α/β are solved, the remaining segments are checked to test whether these regions reach the minimal length of 5 for helix and 4 for β; if they do not they are eliminated. The final retained structure is given and is used to estimate the accuracy of the prediction. A search for β turns in coil regions is performed since the overlapping procedure can change or delete some of the secondary structure previously found.

An optional subroutine which elongates all the possible nuclei by tetrapeptides is included for obeying the elongation rules of Chou and Fasman. At this step no choice is made except for the maximal length of a helix (40 residues) or a β sheet (30 residues). All the possible regions, as well as their (P_α) and (P_β) mean parameters, are listed.

The percentage of total residues in the protein identified correctly is

$$\% N = \frac{100(N - N_x)}{N},$$

where N is the total number of residues in the protein and N_x is the total number of incorrectly predicted residues in the protein

$$N_x = \alpha_m + \alpha_o + \beta_m + \beta_o - n_{double},$$

where α_m and β_m are, respectively, the number of helical of β sheet residues missed in the prediction, α_o and β_o are the number of helical or β sheet residues overpredicted, and

n_{double} is the number of incorrectly predicted residues that were counted twice.

All residues predicted in a helix shorter than four amino acids are considered as incorrectly predicted even if they correspond to an X-ray helical structure.

This compiled program was written for an Apple IIe (128K) microcomputer. A program has been included to predict successively (in an automatic mode) the structure of 50 proteins. A listing of this program may be obtained upon request to the authors.

An extended version is being developed on an IBM mainframe at the Centre de Calcul du CNRS Circe (Orsay).

RESULTS

This program has been assessed on 21 proteins belonging to the four classes of proteins: α, β, $\alpha + \beta$, α/β. The analysis concerned 4457 amino acids and the results are summarized in Table 1.

The overall percentage of correctly predicted residues (based on a three-state description of secondary structure, α, β, coil) obtained by our version is 51.4. By comparing the programs based on the Chou and Fasman method (2), one can see that only 4 proteins (representing a total number of 859 amino acids) are predicted more correctly by Corrigan and Huang (3) than by the present program. A maximal difference of 5.4% is observed for serine protease B. For the 17 other proteins (representing a total number of 3598 amino acids) our program gives better results, the maximal difference being as high as 23.8% for myohemerythrin.

The prediction of the secondary structure

TABLE 1

NUMBER OF CORRECTLY PREDICTED AMINO ACIDS (THREE-STATE MODEL)

			Corrigan and Huang		This work	
Proteins	N	Class[a]	N_c[b]	%	N_c	%
Myohemerythrin	118	α	41	(34.7)	69	(58.5)
Myoglobin	153	α	72	(47)	100	(65.4)
Calcium-binding protein B	108	α	49	(45.4)	71	(65.7)
Superoxide dismutase	151	β	102	(67.5)	94	(62.3)
Bence Jones dimer MEG	215	β	101	(46.9)	109	(50.7)
Serine protease B	185	β	86	(46.5)	76	(41.1)
Lysozyme phage T4	164	$\alpha + \beta$	63	(38.4)	80	(48.8)
Ribonuclease S	124	$\alpha + \beta$	58	(46.8)	76	(61.3)
Papaine	212	$\alpha + \beta$	104	(49.1)	104	(49.1)
Nuclease (*Staphylococcus Aureus*)	149	$\alpha + \beta$	91	(61.1)	87	(58.4)
Thermolysin	316	$\alpha + \beta$	117	(37)	156	(49.4)
Carbonic anhydrase B	256	$\alpha + \beta$	128	(50)	132	(51.6)
Thioredoxin of *Escherichia Coli*	108	α/β	54	(50)	60	(55.6)
Flavodoxin	138	α/β	70	(50.7)	75	(54.3)
Adenylate kinase	194	α/β	93	(47.9)	110	(56.7)
Triose phosphate isomerase	248	α/β	117	(47.2)	158	(63.7)
Alcohol dehydrogenase	374	α/β	167	(44.7)	160	(42.8)
Glyceraldehyde dehydrogenase	333	α/β	136	(40.8)	153	(45.9)
Lactate dehydrogenase–NAD	329	α/β	127	(38.6)	144	(43.8)
Subtilisin NOVO	275	α/β	116	(42.2)	132	(48)
Carboxypeptidase A	307	α/β	137	(44.6)	143	(46.6)
TOTAL	4457		2029	(45.8)	2289	(51.4)

[a] Classification of Richardson (6).

[b] N_c is the number of correctly predicted amino acids.

for three proteins (triose phosphate isomerase, super oxide dismutase, myohemerythrin) is shown in Fig. 2. These proteins have been chosen for their representativity of the structure of many proteins. The comparisons have been made among X-ray structure (a), Chou and Fasman predictions (b), our program (c), and Corrigan and Huang (d). The first result is that our program yields a plausible structure; it does not predict helices shorter than 5 residues and it gives a good idea of the sequence of alternating structures. This is particularly illustrated for triose phosphate isomerase (α/β protein) in which most of the helical regions have been correctly assigned. Moreover, the succession of alternating α and β structures is clearly elucidated.

Although the Corrigan and Huang algorithm yields better results for superoxide dismutase (see Table 1), the 72–79 helix which is not found in X-ray data is predicted in agreement with the Chou and Fasman prediction (2). For myohemerythrin (Fig. 2C) the improvement brought by the implementation of the Chou and Fasman rules is obvious.

To check if the prediction accuracy of a given protein is reliable to the class to which it belongs (Table 1), the mean accuracy attainable by different methods has been compared as a function of the class of proteins (Table 2). For α proteins, our program (63.3%) appears largely better than that of Corrigan and Huang (3) (42.7%). For β proteins, the later (52.5%) is barely better than our version (50.6%). For the last two classes ($\alpha + \beta$, α/β) our program is significantly better than the other one (see Table 2). It should be noticed that for a three-state model a random prediction would give an accuracy of 33%.

DISCUSSION

This paper describes a new computer program that was developed for use on an Apple IIe which is capable of predicting the secondary structure of proteins according to the Chou–Fasman predictive scheme with good accuracy. Sequences up to 1000 amino acids can be analyzed through the compiled form of this program. As pointed out by Garnier *et al.* (1), the necessity for a predictive method to be computerized is obvious since its accuracy would not depend on the predictor; otherwise, a comparison between different methods would be invaluable. The accuracy attainable with our program is better than that previously described using the same

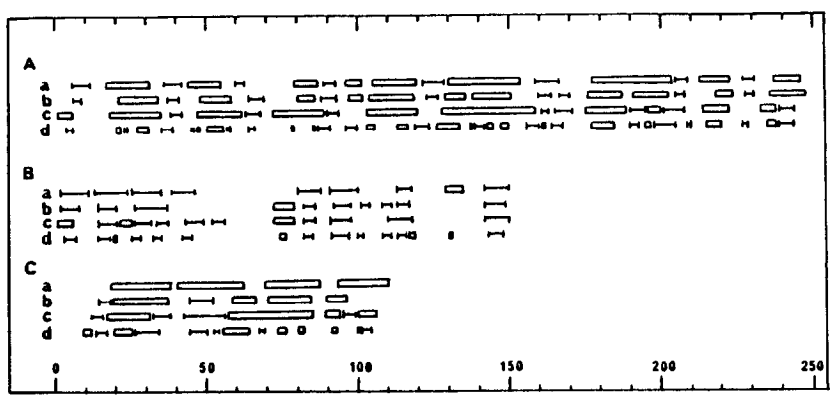

FIG. 2. Comparison between X-ray structure (a), Chou-Fasman prediction (b), our program (c), and the algorithm of Corrigan and Huang (d) for triose phosphate isomerase (A), superoxide dismutase (B), and myohemerythrin (C). Helices are designated by open rectangles and sheets by lines. The blanks in the sequences represents coil regions. The numbered scale indicates the position of amino acids in the three sequences.

TABLE 2

PERCENTAGE OF CORRECTLY PREDICTED AMINO ACIDS

Method	Class of proteins			
	α	β	$\alpha + \beta$	α/β
Corrigan and Huang (3)	42.7	52.5	45.9	44.1
This work	63.3	50.6	52	49.2

method and is of the same order of magnitude as that obtained by the algorithm of Garnier *et al.* (1) with all decision constants equal to zero. An improvement of about 6 residues over 100 is obtained which constitutes a 12% improvement as compared with the Corrigan and Huang program (3).

On the other hand, a program for protein secondary structure analysis has been recently developed (7) using several kinds of structural information. Unfortunately, the results are presented only by plots of the α and β potentials leading to a purely descriptive drawing since no structure is proposed.

Although this type of information is useful, the conclusion drawn by a user may be biased by its rather qualitative interpretation. That is why the combination of at least two deciding computerized methods will be the most powerful method to avoid such kinds of misinterpreting. The main advantages of our program are that it takes into account most of the qualitative rules of Chou and Fasman, it may be used without a precise knowledge of the Chou–Fasman method, and it should be useful for any laboratory equipped with an Apple IIe computer. In addition, this program can help in predicting the super secondary structure such as $\beta\alpha\beta$ units (8) or $\alpha\alpha$ corners (9) since it gives particularly good results for α/β and all α proteins.

REFERENCES

1. Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) *J. Mol. Biol.* **120,** 97–120.
2. Chou, P. Y., and Fasman, G. D. (1978) *Adv. Enzymol. Relat. Subj. Biochem.* **47,** 45–148.
3. Corrigan, A. J., and Huang, P. C. (1982) *Comput. Programs Biomed.* **15,** 163–168.
4. Pauletti, D., Simmonds, R., Dreesman, G. R., and Kennedy, R. C. (1985) *Anal. Biochem.* **151,** 540–546.
5. Kabsch, W., and Sander, C. (1983) *FEBS Lett.* **155,** 179–182.
6. Richardson, J. S. (1981) *Adv. Protein Chem.* **34,** 167–339.
7. Gribskov, M., Burgess, R. R., and Devereux, J. (1986) *Nucleic Acids Res.* **14,** 327–334.
8. Taylor, W. R., and Thornton, J. M. (1983) *Nature (London)* **301,** 540–542.
9. Efimov, A. V. (1984) *FEBS Lett.* **166,** 33–38.