CABIOS

Vol.4. no.1. 1988
Pages 187 – 191

An integrated family of amino acid sequence analysis programs

H.Wolf[1]*, S.Modrow[1], M.Motz[1], B.A.Jameson[1,2], G.Hermann[3] and B.Förtsch[1]

Abstract

During the last years abundant sequence data has become available due to the rapid progress in protein and DNA sequencing techniques. The exact three-dimensional structures, however, are available only for a fraction of proteins with known sequences. For many purposes the primary amino acid sequence of a protein can be directly used to predict important structural parameters. However, mathematical presentation of the calculated values often makes interpretation difficult, especially if many proteins must be analysed and compared. Here we introduce a broad-based, user-defined analysis of amino acid sequence information. The program package is based on published algorithms and is designed to access standard protein data bases, calculate hydropathy, surface probability and flexibility values and perform secondary structure predictions. The data output is in an 'easy-to-read' graphic format and several parameters can be superimposed within a single plot in order to simplify data interpretations. Additionally, this package includes a novel algorithm for the prediction of potential antigenic sites. Thus the software package presented here offers a powerful means of analysing an amino acid sequence for the purpose of structure/function studies as well as antigenic site analyses. These algorithms were written to function in context with the UWGCG (University of Wisconsin Genetics Computer Group) program collection, and are now distributed within that package.

Introduction

Nucleic acid and protein sequences have been evaluated and are available for most laboratories by accessing sequence libraries. Precise information about the structures of these proteins can be reliably obtained by methods such as X-ray diffraction (Liljas and Rossman, 1974) and two-dimensional NMR spectroscopy (Wagner and Wüthrich, 1982); these methods, however, require relatively large amounts of pure crystalline polypeptides and can be applied only to a very small number of proteins with known primary amino acid sequences. A great deal of information, however, may also be obtained for many proteins by a careful examination of the linear amino acid sequence. For these reasons computer algorithms were developed for analysis of such sequences, enabling the prediction of probable secondary structural features (Chou and Fasman, 1978; Garnier *et al.*, 1978), regional backbone flexibility (Karplus and Schulz, 1985) as well as parameters relating to surface accessibility (Janin *et al.*, 1978). The subsequent application of these predictions may facilitate the identification of functional and structural parameters, i.e. transmembrane-spanning regions and signal peptides and may lead to a better understanding of the three-dimensional arrangement of a given amino acid sequence.

Here we present an integrated protein analysis software package which uses the algorithms of Chou and Fasman (1978) and Garnier *et al.* (1978) for prediction of the secondary structures. Since numerical presentation of the calculated values is often rather difficult to interpret, the data may be presented in a two-dimensional or linear graphic output. These structural values may be superimposed with additional parameters such as hydrophilicity (Hopp and Woods, 1981; Kyte and Doolittle, 1983), flexibility (Karplus and Schulz, 1985) and surface probability (modified from Emini *et al.*, 1985). The combination of these parameters with the secondary structure predictions facilitates the identification of continuous antigenic sites in an amino acid sequence. Such structures are most often located in regions with a high content of β-turns and/or high values for hydrophilicity, flexibility and surface probability; the immunodominant antigenic determinants generally exist in loop-like structures at the surface of a protein molecule. As a first step towards automated prediction of antigenic sites, these parameters were combined in a weighted manner resulting in a novel algorithm, the antigenic index (Jameson and Wolf, 1988).

Systems and methods

The programs PROTCALC for calculation of secondary protein structures and additional parameters such as hydrophilicity, surface probability, flexibility, antigenicity and potential N-glycosylation, and PROTPLOT for graphic design were written in VAX/FORTRAN (version 4.1); the VAX 750 (Digital Equipment Corporation) was used as host computer. All drawings were produced on a Hewlett Package Plotter 7475A. The protein sequences were taken from the GENBANK sequence data library, EMBO sequence collection or NBRF protein data

[1] *Max von Pettenkofer Institute, Pettenkoferstr. 9a, D-8000 Munich 2, FRG,* [2] *California Institute of Technology, Division of Biology, Pasadena, CA 91125, USA and* [3] *Gesellschaft für Strahlen- und Umweltforschung, D-8042 Neuherberg, FRG*

**To whom reprint requests should be sent*

bank. The algorithms were written to function within the UWGCG Sequence Analysis Software Package (Devereux *et al.*, 1984); both the analysis program and the graphics output are part of this program collection.

Algorithm

Algorithms for the prediction of secondary protein structure were taken from Garnier *et al.* (1978) and Chou and Fasman (1978). The overlapping regions of an α-helix and β-sheet were resolved by using the 'overall probability' introduced by Nishikawa (1983). The same procedure was also applied to locate turn regions which are inconsistent with other secondary structures. The following modifications of Chou–Fasman rules were used for α-helical regions: the boundary conditions of p(bound) > 1.0 and necessary conditions of $p(\alpha) > p(\beta)$ are removed.

The values of hydrophilicity were determined according to Hopp and Woods (1981) or Kyte and Doolittle (1983). The latter values were multiplied by −1 to allow the same orientation of peak values as with the calculations according to Hopp and Woods and in order to facilitate the usual interpretation, i.e. that in all graphs positive or higher numeric values favor increased potential presence at the surface of a protein structure.

The calculation for the backbone flexibility of the amino acid sequence was performed as described by Karplus and Schulz (1985).

Surface probabilities were based on the individual amino acid data obtained by Janin *et al.* (1978) and calculated using a modification of the algorithm by Emini *et al.* (1985). In the equation below the surface probability at position *n* is defined for sequential hexapeptide sequences as

$$S_n = \left(\prod_{i=1}^{x} \delta_{n+4-i}\right)*(0.37)^{-x}$$

where δ_n is the fractional surface probability and x has the value of 6 for values of *n* further away from the ends than three amino acids and decreased as *n* approaches the ends.

The antigenic index was calculated from an experimentally derived equation based on the data derived from hydrophilicity (H), flexibility (F), surface probability (S), Chou–Fasman secondary structure prediction (CF) and Garnier secondary structure prediction (RG):

$$A_n = 0.3(H_n) + 0.15(S_n) + 0.15(F_n) + 0.2(CF_n) + 0.2(RG_n)$$

(Values for H_n, S_n, F_n, CF_n and RG_n are given in Table I.)

N-glycosylation sites are indicated by the sequences Asn-X-Ser or Asn-X-Thr, and with minor probability, when X is represented by amino acids Asp, Trp or Pro.

Implementation

PROTCALC

The operator must enter the amino acid sequence which is to be analysed in the single letter code through the keyboard or—in the case of an already published sequence—must retrieve it from one of the sequence libraries. PROTCALC calculates the values by running a window of seven (default setting) amino acids from the beginning to the end of the protein sequence; the parameters obtained for every residue in this window are averaged; the size of the window may be altered. A further option is given in order to broaden the peaks of the antigenic index (from *n* − 4 to *n* + 4) by adding 80, 60, 40 or 20% of the peak value to the flanking values in descending order to account for the influence of the additional free energy derived from the mobility of surface regions relative to regions buried inside the protein. The output of PROTCALC is presented as tables with the numerical values for hydrophilicity, surface probability, flexibility and antigenic index. Glycosylation sites of high and low probability are indicated by G and g respectively. Secondary structure parameters are given by H (α-helix), B (β-sheet) or T (β-turn) or by h, b or t for weak secondary structure parameters; random coil regions are not indicated by a special letter. The printout of the calculation for p17 of the human immunodeficiency virus 1 (HIV/HTLV-3, BH10) (Ratner *et al.*, 1985) is shown in Figure 1.

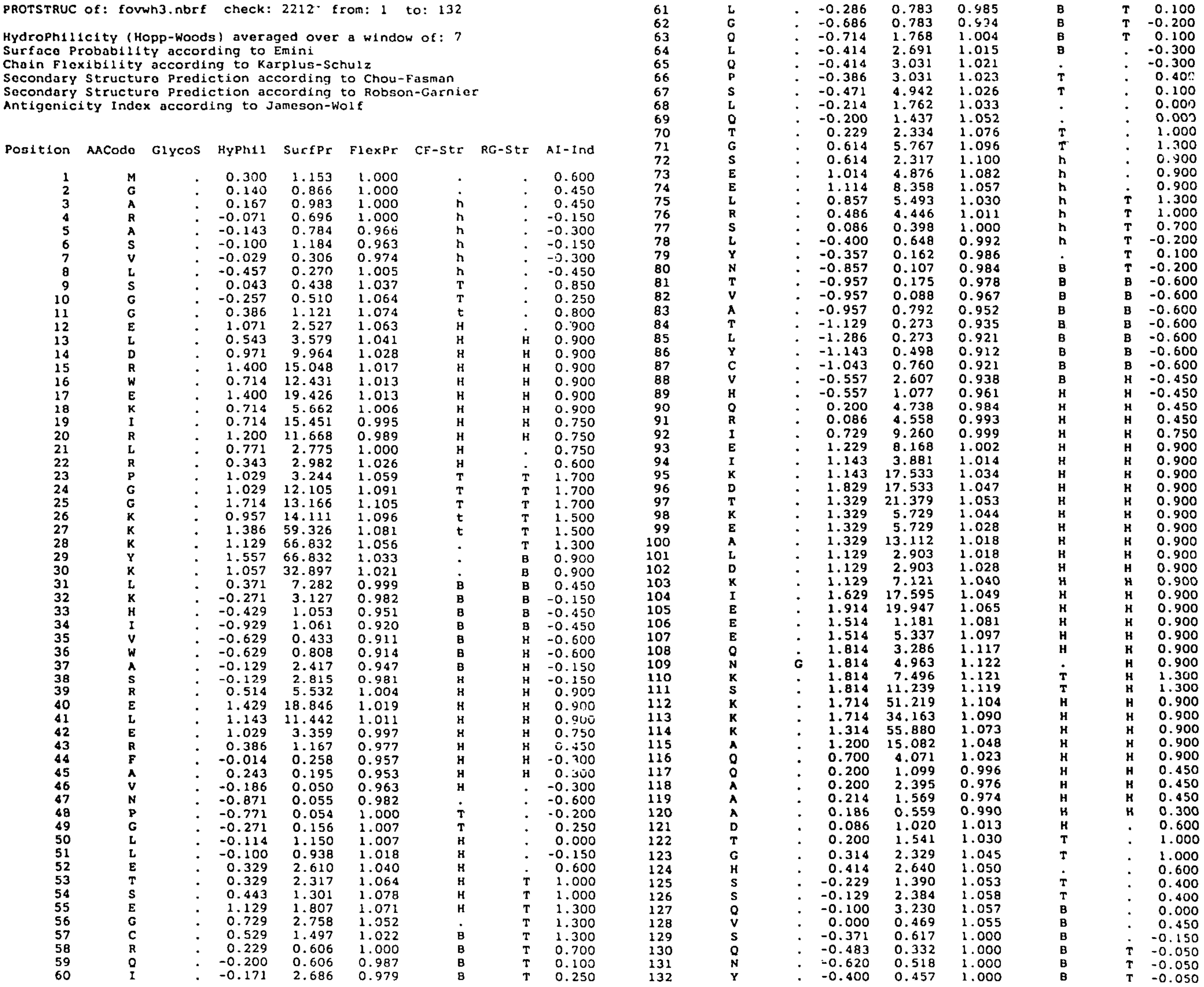
Table I. Computation of the antigenic index

Values used for calculation of AI for calculated values of column 2	Values calculated according to references listed above
$H_i = 2$	H > 0.5
$H_i = 1$	0.5 > H > O
$H_i = -1$	O > H > −0.4
$H_i = -2$	−0.4 > H
$S_i = 1$	1.0 > S
$S_i = 0$	Otherwise
$F_i = 1$	1.0 > F
$F_i = 0$	Otherwise
$CF_i = 2$	CF = strong turn
$CF_i = 1$	CF = weak turn or random coil
$CF_i = 0$	Otherwise
$RG_i = 2$	RG = strong turn
$RG_i = 1$	RG = weak turn or random coil
$RG_i = 0$	Otherwise

H, hydrophilicity; S, surface probability; F, backbone flexibility; CF, secondary structure prediction; RG, secondary structure prediction; AI, antigenic index.

PROTPLOT

This program uses the protein calculation file from PROTPLOT and draws a colored two-dimensional or linear graph of p17 of HIV-1 (Figure 2a and b). In the linear plot all derived parameters derived from PROTCALC may be presented in a combined graphic output. Alternatively, the operator may select single parameters to be plotted. The *y* axis may be enlarged by a factor of two in both cases. The operator can also choose



```
PROTSTRUC of: fovwh3.nbrf  check: 2212˙ from: 1  to: 132

HydroPhilicity (Hopp-Woods) averaged over a window of: 7
Surface Probability according to Emini
Chain Flexibility according to Karplus-Schulz
Secondary Structure Prediction according to Chou-Fasman
Secondary Structure Prediction according to Robson-Garnier
Antigenicity Index according to Jameson-Wolf

Position  AACode  GlycoS  HyPhil  SurfPr  FlexPr  CF-Str  RG-Str  AI-Ind

    1       M       .     0.300   1.153   1.000      .       .     0.600
    2       G       .     0.140   0.866   1.000      .       .     0.450
    3       A       .     0.167   0.983   1.000      h       .     0.450
    4       R       .    -0.071   0.696   1.000      h       .    -0.150
    5       A       .    -0.143   0.784   0.966      h       .    -0.300
    6       S       .    -0.100   1.184   0.963      h       .    -0.150
    7       V       .    -0.029   0.306   0.974      h       .    -0.300
    8       L       .    -0.457   0.270   1.005      h       .    -0.450
    9       S       .     0.043   0.438   1.037      T       .     0.850
   10       G       .    -0.257   0.510   1.064      T       .     0.250
   11       G       .     0.386   1.121   1.074      t       .     0.800
   12       E       .     1.071   2.527   1.063      H       .     0.900
   13       L       .     0.543   3.579   1.041      H       H     0.900
   14       D       .     0.971   9.964   1.028      H       H     0.900
   15       R       .     1.400  15.048   1.017      H       H     0.900
   16       W       .     0.714  12.431   1.013      H       H     0.900
   17       E       .     1.400  19.426   1.013      H       H     0.900
   18       K       .     0.714   5.662   1.006      H       H     0.900
   19       I       .     0.714  15.451   0.995      H       H     0.750
   20       R       .     1.200  11.668   0.989      H       H     0.750
   21       L       .     0.771   2.775   1.000      H       .     0.750
   22       R       .     0.343   2.982   1.026      H       .     0.600
   23       P       .     1.029   3.244   1.059      T       T     1.700
   24       G       .     1.029  12.105   1.091      T       T     1.700
   25       G       .     1.714  13.166   1.105      T       T     1.700
   26       K       .     0.957  14.111   1.096      t       T     1.500
   27       K       .     1.386  59.326   1.081      t       T     1.500
   28       K       .     1.129  66.832   1.056      .       T     1.300
   29       Y       .     1.557  66.832   1.033      .       B     0.900
   30       K       .     1.057  32.897   1.021      .       B     0.900
   31       L       .     0.371   7.282   0.999      B       B     0.450
   32       K       .    -0.271   3.127   0.982      B       B    -0.150
   33       H       .    -0.429   1.053   0.951      B       B    -0.450
   34       I       .    -0.929   1.061   0.920      B       B    -0.450
   35       V       .    -0.629   0.433   0.911      B       H    -0.600
   36       W       .    -0.629   0.808   0.914      B       H    -0.600
   37       A       .    -0.129   2.417   0.947      B       H    -0.150
   38       S       .    -0.129   2.815   0.981      H       H    -0.150
   39       R       .     0.514   5.532   1.004      H       H     0.900
   40       E       .     1.429  18.846   1.019      H       H     0.900
   41       L       .     1.143  11.442   1.011      H       H     0.900
   42       E       .     1.029   3.359   0.997      H       H     0.750
   43       R       .     0.386   1.167   0.977      H       H     0.450
   44       F       .    -0.014   0.258   0.957      H       H    -0.300
   45       A       .     0.243   0.195   0.953      H       H     0.300
   46       V       .    -0.186   0.050   0.963      H       .    -0.300
   47       N       .    -0.871   0.055   0.982      .       .    -0.600
   48       P       .    -0.771   0.054   1.000      T       .    -0.200
   49       G       .    -0.271   0.156   1.007      T       .     0.250
   50       L       .    -0.114   1.150   1.007      H       .     0.000
   51       L       .    -0.100   0.938   1.018      H       .    -0.150
   52       E       .     0.329   2.610   1.040      H       .     0.600
   53       T       .     0.329   2.317   1.064      H       T     1.000
   54       S       .     0.443   1.301   1.078      H       T     1.000
   55       E       .     1.129   1.807   1.071      H       T     1.300
   56       G       .     0.729   2.758   1.052      .       T     1.300
   57       C       .     0.529   1.497   1.022      B       T     1.300
   58       R       .     0.229   0.606   1.000      B       T     0.700
   59       Q       .    -0.200   0.606   0.987      B       T     0.100
   60       I       .    -0.171   2.686   0.979      B       T     0.250
   61       L       .    -0.286   0.783   0.985      B       T     0.100
   62       G       .    -0.686   0.783   0.994      B       T    -0.200
   63       Q       .    -0.714   1.768   1.004      B       T     0.100
   64       L       .    -0.414   2.691   1.015      B       .    -0.300
   65       Q       .    -0.414   3.031   1.021      .       .    -0.300
   66       P       .    -0.386   3.031   1.023      T       .     0.400
   67       S       .    -0.471   4.942   1.026      T       .     0.100
   68       L       .    -0.214   1.762   1.033      .       .     0.000
   69       Q       .    -0.200   1.437   1.052      .       .     0.000
   70       T       .     0.229   2.334   1.076      T       .     1.000
   71       G       .     0.614   5.767   1.096      T       .     1.300
   72       S       .     0.614   2.317   1.100      h       .     0.900
   73       E       .     1.014   4.876   1.082      h       .     0.900
   74       E       .     1.114   8.358   1.057      h       .     0.900
   75       L       .     0.857   5.493   1.030      h       T     1.300
   76       R       .     0.486   4.446   1.011      h       T     1.000
   77       S       .     0.086   0.398   1.000      h       T     0.700
   78       L       .    -0.400   0.648   0.992      h       T    -0.200
   79       Y       .    -0.357   0.162   0.986      .       T     0.100
   80       N       .    -0.857   0.107   0.984      B       T    -0.200
   81       T       .    -0.957   0.175   0.978      B       B    -0.600
   82       V       .    -0.957   0.088   0.967      B       B    -0.600
   83       A       .    -0.957   0.792   0.952      B       B    -0.600
   84       T       .    -1.129   0.273   0.935      B       B    -0.600
   85       L       .    -1.286   0.273   0.921      B       B    -0.600
   86       Y       .    -1.143   0.498   0.912      B       B    -0.600
   87       C       .    -1.043   0.760   0.921      B       B    -0.600
   88       V       .    -0.557   2.607   0.938      B       H    -0.450
   89       H       .    -0.557   1.077   0.961      H       H    -0.450
   90       Q       .     0.200   4.738   0.984      H       H     0.450
   91       R       .     0.086   4.558   0.993      H       H     0.450
   92       I       .     0.729   9.260   0.999      H       H     0.750
   93       E       .     1.229   8.168   1.002      H       H     0.900
   94       I       .     1.143   3.881   1.014      H       H     0.900
   95       K       .     1.143  17.533   1.034      H       H     0.900
   96       D       .     1.829  17.533   1.047      H       H     0.900
   97       T       .     1.329  21.379   1.053      H       H     0.900
   98       K       .     1.329   5.729   1.044      H       H     0.900
   99       E       .     1.329   5.729   1.028      H       H     0.900
  100       A       .     1.329  13.112   1.018      H       H     0.900
  101       L       .     1.129   2.903   1.018      H       H     0.900
  102       D       .     1.129   2.903   1.028      H       H     0.900
  103       K       .     1.129   7.121   1.040      H       H     0.900
  104       I       .     1.629  17.595   1.049      H       H     0.900
  105       E       .     1.914  19.947   1.065      H       H     0.900
  106       E       .     1.514   1.181   1.081      H       H     0.900
  107       E       .     1.514   5.337   1.097      H       H     0.900
  108       Q       .     1.814   3.286   1.117      H       H     0.900
  109       N       G     1.814   4.963   1.122      .       H     0.900
  110       K       .     1.814   7.496   1.121      T       H     1.300
  111       S       .     1.814  11.239   1.119      T       H     1.300
  112       K       .     1.714  51.219   1.104      H       H     0.900
  113       K       .     1.714  34.163   1.090      H       H     0.900
  114       K       .     1.314  55.880   1.073      H       H     0.900
  115       A       .     1.200  15.082   1.048      H       H     0.900
  116       Q       .     0.700   4.071   1.023      H       H     0.900
  117       Q       .     0.200   1.099   0.996      H       H     0.450
  118       A       .     0.200   2.395   0.976      H       H     0.450
  119       A       .     0.214   1.569   0.974      H       H     0.450
  120       A       .     0.186   0.559   0.990      H       H     0.300
  121       D       .     0.086   1.020   1.013      H       .     0.600
  122       T       .     0.200   1.541   1.030      T       .     1.000
  123       G       .     0.314   2.329   1.045      T       .     1.000
  124       H       .     0.414   2.640   1.050      .       .     0.600
  125       S       .    -0.229   1.390   1.053      T       .     0.400
  126       S       .    -0.129   2.384   1.058      T       .     0.400
  127       Q       .    -0.100   3.230   1.057      B       .     0.000
  128       V       .     0.000   0.469   1.055      B       .     0.450
  129       S       .    -0.371   0.617   1.000      B       .    -0.150
  130       Q       .    -0.483   0.332   1.000      B       T    -0.050
  131       N       .    -0.620   0.518   1.000      B       T    -0.050
  132       Y       .    -0.400   0.457   1.000      B       T    -0.050
```

Fig. 1. Printout of protein analysis of p17, a core protein of the human immunodeficiency virus (HIV-1).

between a one- or a four-color plot. In the two-dimensional display the secondary structure predictions by either Chou–Fasman (1978) or Garnier *et al.* (1978) are plotted; these drawings may be superimposed on the values for hydrophilicity and hydrophobicity, flexibility, surface probability or antigenic index. Those additional values are indicated by special symbols (circles or diamonds) which have been superimposed on the secondary structures (for explanation of the symbols see Figure 2a). Display of potential N-glycosylation sites may be suppressed; the default numbering is set with an interval of 50 residues but may be user-defined. There is an extra option for plotting weak Chou–Fasman parameters. Particular amino acids can be marked by their single letter code.

The two-dimensional plot is first drawn on an imaginary plane and then adjusted to give maximum filling of the given sheet format. Adjustment is done proportionally for the *x* and *y* axes.

Discussion

The two-dimensional plot allows for a fast screening of large numbers of proteins for structural parameters such as transmembrane regions or potential segments of genes likely to be recognized by the immune system. The latter feature is of particular importance when subclones, e.g. for use as an antigen source for diagnosis (Motz *et al.*, 1986), or the respective synthetic peptide are prepared and used similarly. It should be noted that the given structure in the graphic output is simply a consequence of predicted turns with all other structural elements being represented in a modified sine/zig-zag/waved line and does not resemble in any way the natural structure. However, the graphic representation, especially when superimposed values are displayed, facilitates and accelerates analysis of the gross properties of a protein.

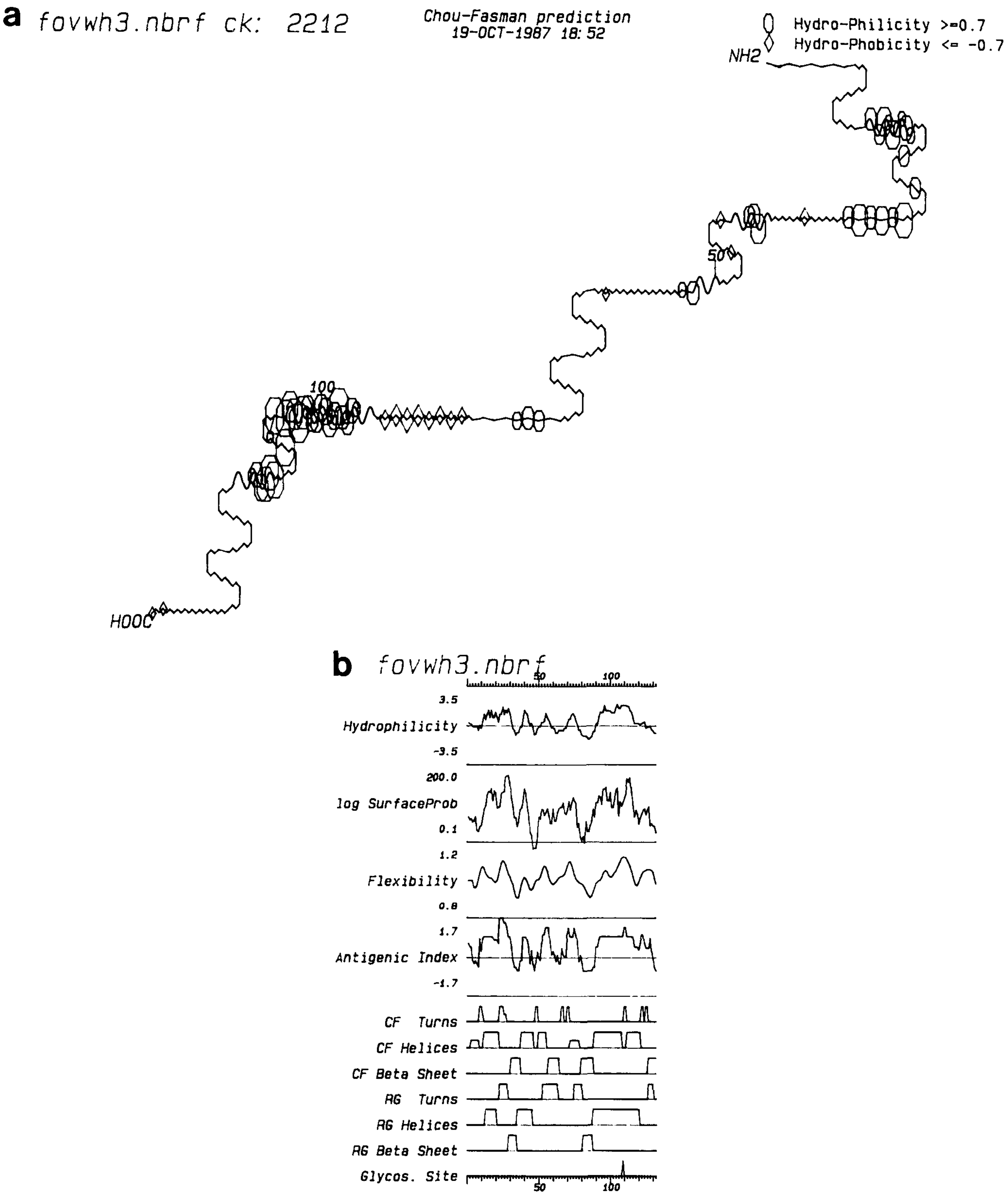


Fig. 2. (a) Two-dimensional graph of p17 of HIV-1 using strong Chou–Fasman parameters. ∿∿ : α-helical structures; ∿∿∿ : β-sheet; ⌒ : β-turn; ∿∿ : random coil; 6 : glycosylation site; ○: hydrophilic regions; ◇ : hydrophobic regions. (b) Linear graph of p17 of HIV-1.

A disadvantage of this system is the variable scale which makes it difficult to compare proteins of different sizes from the sequence of structural elements where one might reflect a subunit of the other. In addition, only a limited number of parameters can be given and judged with any one drawing. These problems can be alleviated by a linear output with a fixed scale and almost unlimited space for presentation of parameters.

Those predictions and calculations have been successfully used for immunological questions, especially in the selection of peptides and protein subunits for production by chemical or gene technological methods (Motz *et al.*, 1986; Modrow and Wolf, 1986; Modrow *et al.*, 1987; Jameson *et al.*, 1987). In particular, when larger genes or genomes need to be analysed, random synthesis of peptides or shotgun expression are both expensive and time consuming. Therefore selection of specific sites is a useful approach.

Correlation with the natural conformation can only be expected to the extent which has been reported for CF and RG predictions, but in comparison with known related structures such approaches should become more reliable.

References

Chou,P.Y. and Fasman,G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.*, **47**, 45–148.

Devereux,J., Haeberli,P. and Smithies,O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.

Emini,E.A., Hughes,J.V., Perlow,D.S. and Bager,I. (1985) Induction of Hepatitis A virus–neutralizing antibody by a virus-specific synthetic peptide. *J. Virol.*, **55**, 836–839.

Garnier,J., Osguthorpe,O.J. and Robson,B. (1978) Analysis of the accuracy and implication 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 antigen determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA*, **78**, 3824–3828.

Jameson,B.A. and Wolf,H. (1988) Predicting antigenicity from protein primary structure: a new algorithm for the prediction of antigenic sites. *CABIOS*, **4**, 181–186.

Jameson,B., Guertler,L. and Wolf,H. (1987) Priming of anti-HIV neutralizing antibodies with an ENV-derived synthetic peptide. *Cold Spring Harbor*, in press.

Janin,J., Wodak,S., Levitt,M. and Maigret,B. (1987) Conformation of amino acid side-chains in proteins. *J. Mol. Biol.*, **125**, 357–386.

Karplus,P.A. and Schulz,G.E. (1985) Prediction of chain flexibility in proteins. *Naturwissenschaften*, **72**, 212–213.

Kyte,J. and Doolittle,R.F. (1983) A simple method for displaying the hydrophathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Liljas,A. and Rossman,M.G. (1974) X-ray studies of protein interactions. *Annu. Rev. Biochem.*, **43**, 475–507.

Modrow,S. and Wolf,H. (1986) Characterization of two related Epstein–Barr virus-encoded proteins by synthetic oligopeptides, which are differentially expressed in Burkitt's lymphoma and *in vitro* transformed cell lines. *Proc. Natl. Acad. Sci. USA*, **83**, 5703–5707.

Modrow,S., Han,B., Shaw,G.M., Gallo,R.C., Wong-Staal,F. and Wolf,H. (1987) Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions. *J. Virol.*, **61**, in press.

Motz,M., Fan,J., Seibl,R., Jilg,W. and Wolf,H. (1986) Expression of the Epstein–Barr virus 138-kDa early protein in *Escherichia coli* for the use as antigen in diagnostic tests. *Gene*, **42**, 303–312.

Nishikawa,K. (1983) Assessment of secondary structure prediction of proteins comparison of computerized Chou–Fasman method with others. *Biochim. Biophys. Acta*, **748**, 285–299.

Ratner,L., Haseltine,W., Patarca,R., Livak,K.J., Starcich,B., Josephs,S.F., Doran,E.R., Rafalske,J.A., Whitehorn,E.A., Baumeister,K., Ivanoffm,L., Petteway,S.R., Pearson,M.L., Lautenberger,J.A., Papas,T.S., Ghrayeb,J., Chang,N.T., Gallo,R.C. and Wong-Staal,F. (1985) Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature*, **313**, 227–233.

Wagner,G. and Wüthrich,K. (1982) Sequential resonance assignments in Protein ^{1}H nuclear magnetic resonance spectra. Basic pancreatic trypsin inhibitor. *J. Mol. Biol.*, **155**, 347–366.

Received on March 20, 1987; accepted September 4, 1987