An integrated family of amino acid sequence analysis programs

H. Wolf¹*, S. Modrow¹, M. Motz¹, B.A. Jameson¹,², G. Hermann³ and B. Förtsch¹

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 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 Würthrich, 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 Packard Plotter 7475A. The protein sequences were taken from the GENBANK sequence data library, EMBO sequence collection or NBRF protein data

¹Max von Pettenkofer Institute, Pettenkoferstr. 9a, D-8000 Munich 2, FRG
²California Institute of Technology, Division of Biology, Pasadena, CA 91125, USA and ³Gesellschaft für Strahlen- und Umweltforschung, D-8042 Neuherberg, FRG
*To whom reprint requests should be sent
bank. The algorithms were written to function within the UWCGG 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(α) > p(β) 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

\[ S_n = \frac{1}{2} \sum_{i=1}^{x} (I_i \delta_i + 4, \ldots + 4) \times (0.37)^{-i} \]

where \( \delta_i \) 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):

\[ A_n = 0.3(H_n) + 0.15(F_n) + 0.15(F_n) + 0.2(CF_n) + 0.2(RG_n) \]

(Values for \( H_n, F_n, CF_n \) and \( RG_n \) are given in Table 1.)

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 pl7 of the human immunodeficiency virus 1 (HIV/HTLV-3, BH10) (Ratner et al., 1985) is shown in Figure 1.

PROTPLOT

This program uses the protein calculation file from PROTPLOT and draws a colored two-dimensional or linear graph of pl7 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

### Table 1. Computation of the antigenic index

<table>
<thead>
<tr>
<th>Values used for calculation of column 2</th>
<th>Values calculated according to references listed above</th>
</tr>
</thead>
<tbody>
<tr>
<td>H, hydrophilicity; S, surface probability; F, backbone flexibility; CF, secondary structure prediction; RG, secondary structure prediction; AI, antigenic index.</td>
<td></td>
</tr>
<tr>
<td>H = 2</td>
<td>H &gt; 0.5</td>
</tr>
<tr>
<td>H = 1</td>
<td>0.5 &gt; H &gt; 0</td>
</tr>
<tr>
<td>H = -1</td>
<td>O &gt; H &gt; -0.4</td>
</tr>
<tr>
<td>H = -2</td>
<td>-0.4 &gt; H</td>
</tr>
<tr>
<td>S = 1</td>
<td>1.0 &gt; S</td>
</tr>
<tr>
<td>S = 0</td>
<td>Otherwise</td>
</tr>
<tr>
<td>F = 1</td>
<td>1.0 &gt; F</td>
</tr>
<tr>
<td>F = 0</td>
<td>Otherwise</td>
</tr>
<tr>
<td>CF = 2</td>
<td>CF = strong turn</td>
</tr>
<tr>
<td>CF = 1</td>
<td>CF = weak turn or random coil</td>
</tr>
<tr>
<td>CF = 0</td>
<td>Otherwise</td>
</tr>
<tr>
<td>RG = 2</td>
<td>RG = strong turn</td>
</tr>
<tr>
<td>RG = 1</td>
<td>RG = weak turn or random coil</td>
</tr>
<tr>
<td>RG = 0</td>
<td>Otherwise</td>
</tr>
</tbody>
</table>
Family of amino acid sequence analysis programs

PROSTTRUC analysis of: low1h.vurf check: 2121; from 1 to: 132

Hydrophillibity (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 Janson-Weid

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/wave 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. \( \alpha \)-helical structures: \( \beta \)-sheet: \( \beta \)-turn: \( \alpha \)-random coil: \( \delta \): glycosylation site; O: hydrophilic regions; O: 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


