The antigenic index: a novel algorithm for predicting antigenic determinants

B.A. Jameson and H. Wolf

Abstract

In this paper, we introduce a computer algorithm which can be used to predict the topological features of a protein directly from its primary amino acid sequence. The computer program generates values for surface accessibility parameters and combines these values with those obtained for regional backbone flexibility and predicted secondary structure. The output of this algorithm, the antigenic index, is used to create a linear surface contour profile of the protein. Because most, if not all, antigenic sites are located within surface exposed regions of a protein, the program offers a reliable means of predicting potential antigenic determinants. We have tested the ability of this program to generate accurate surface contour profiles and predict antigenic sites from the linear amino acid sequences of well-characterized proteins and found a strong correlation between the predictions of the antigenic index and known structural and biological data.

Introduction

The availability of protein sequence information has increased the need for accurate methods of predicting both antigenic and immunogenic sites directly from such sequences. Although it is commonly believed that most surface exposed regions of a protein are potential antigenic sites, very little is known about the characteristics of the subset of antigenic sites which are also immunogenic sites (for review, see Benjamin et al., 1984; Berzofsky, 1985). Therefore, until more information is uncovered concerning the nature of immunogenic sites, the general focus of predictive computer algorithms should be directed toward antigenic determinants. The best predictive methods of determining antigenic sites must, in turn, rely on an accurate prediction of the surface contour of a given protein. However, obtaining reliable estimates of a surface exposed region is, at best, a difficult endeavor.

The primordial approach to locating potentially exposed protein sequences involves plotting hydropathic characteristics of amino acids using a 6—10 amino acid window on the protein sequence of interest (Hopp and Woods, 1981; Kyte and Doolittle, 1982). This type of approach has been generally employed because highly charged clusters of amino acids are almost always exposed to an aqueous environment and strongly hydrophobic clusters are usually buried within the interior of the protein. Based on hydropathy alone, however, one cannot accurately predict whether sequences between these two extremes will be exposed on the surface or hidden inside the protein.

In a similar vein, a 'surface probability' plot has been developed (Emini et al., 1985) based on side-chain solvent accessibility values of the individual amino acids (Janin et al., 1978). Although this type of plot gives a set of qualitatively different values than the hydropathy plots, it suffers from the same inherent limitations.

Highly mobile protein segments (flexible regions) seem to correlate well with known antigenic determinants and, in some cases, known immunogenic determinants (Atassi, 1984; Tanier et al., 1984; Westhof et al., 1984). Consequently, flexibility plots have been proposed as indicators of potential antigenic determinants. Although such plots are rarely referred to as surface contour plots, it is reasonable to assume that highly mobile protein segments are located on the surface of a protein due to an entropic energy potential. Most of the flexibility plots are based on calculations of X-ray diffraction-derived B factors (factors which describe the isotropic harmonic oscillation of localized atomic regions). Karplus and Schulz (1985) have obtained average B-factors calculated from 31 different protein structures in order to plot segmental mobility directly from primary sequence information. Although such plots are informative, not all antigenic peaks can be predicted by flexibility parameters alone.

Historically, antigenic prediction plots have utilized a single approach, i.e. hydropathy, surface probability or flexibility. Although several groups have recently superimposed hydropathy values onto secondary structural predictions (Cohen et al., 1984; Dietzschold et al., 1984; Modrow and Wolf, 1986), no group has yet attempted to integrate these various parameters discussed above into a single plot. We report here a novel means of integrating flexibility parameters with hydropathy/solvent accessibility values in a weighted fashion in order to produce a plot of surface contour, referred to as the ‘antigenic index’. This algorithm has been termed the ‘antigenic index’ to reflect that it is an index, i.e. an indication, of potentially exposed surface peaks of a protein. This new plot offers a means of overcoming many of the inherent limitations of the individual surface prediction algorithms.
Systems and methods

The programs for the antigenic index were written in VAX/ FORTRAN (version 4.1). The host computer used for these programs was a VAX 750 (Digital Equipment Corporation). The protein sequences used in this study were all taken from the NBRF protein data bank. The programs were written to function within the University of Wisconsin Computer Group (UWCGC) Sequence Analysis Software Package environment (Devereux et al., 1984). The antigenic index is currently available in VAX-form in the UWCGC Sequence Analysis Software Package and in a portable form, suitable for either IBM PC-compatible or Apple-compatible systems, through International Biotechnologies Incorporated.

Algorithm

An algorithm was designed to integrate the predicted influence of hydrophathy/surface probability with flexibility factors. The computational section of this algorithm is comprised of six major subroutines. The hydrophilicity values are determined according to the method of Hopp and Woods (1981). Surface probabilities are based on the individual amino acid data obtained by Janin et al. (1978) and calculated using a modification of the equation described by Emini et al. (1985). In the equation below, the surface probability (S) at position n is defined for sequential hexapeptide sequences as:

\[ S_n = (\prod_{i=1}^{6} \delta_{n+i} + 4) \times (0.37)^6 \]

where \( \delta_n \) is the fractional surface probability. We have modified this equation to compensate for the ends of proteins by allowing the (hexapeptide) factor 6 to vary according to the remaining number of amino acids at the carboxy end of the protein (6, 5, 4, etc.). The analysis of backbone flexibility is performed as described by Karplus and Schulz (1985). Predictions of secondary structure were computed in two different ways. The first method utilizes the individual predictions of helix, sheet and turn according to the rules of the original Chou-Fasman method (15). The overlapping regions of helix and \( \beta \)-sheet are resolved using the 'overall probability' introduced by Nishikawa (1983). The same procedure is also applied here to locate the turn regions which are inconsistent with other secondary structures. We have also used the following modification of the Chou-Fasman rules:

Helix: the boundary condition of \( p(\text{bound}) > 1.0 \) and the necessary condition of \( p(\alpha) > p(\beta) \) are removed.

The other method of secondary structure prediction is carried out as described by Garnier et al. (1978). Although the prediction of secondary structures is calculated as described by either Chou and Fasman (1978) or Garnier et al. (1978), the output has been simplified such that the relative abilities of individual residues, based on windowed averages, to participate in a \( \alpha \)-helix, \( \beta \)-sheet or \( \beta \)-turn are indicated as either 'strong' or 'weak' structure formers.

The last subroutine combines flexibility parameters with hydrophathy/solvent accessibility factors and is used to determine a surface contour value (antigenic index). The antigenic index is based on the information derived from the various subroutines described above and is calculated according to the equation shown in Table I.

After a composite value for the antigenic index has been calculated, a final peak-broadening function is overlayed upon the generated profile. The major surface peaks are broadened (from \( n - 4 \) to \( n + 4 \)) by adding 80, 60, 40 or 20% of the peak value, respectively, to account for the influence of the additional free energy derived from the mobility of surface peaks, relative to regions buried in the interior of the protein. This calculation is performed for all major peaks except for those peaks which are predicted to occur within a region of a helix because such regions tend to be less flexible.

Implementation

A computer program was written for predicting a protein's surface contour (potential antigenic determinants) directly from its primary amino acid sequence. This program, the antigenic index, was designed to integrate flexibility components such as averaged

### Table I. Computation of the antigenic index

<table>
<thead>
<tr>
<th>Value used for calculation of ( A_1 )</th>
<th>Values calculated according to references listed above</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_1 = 2 )</td>
<td>( H &gt; 0.05 )</td>
</tr>
<tr>
<td>( H_1 = 1 )</td>
<td>( 0.5 &gt; H &gt; 0 )</td>
</tr>
<tr>
<td>( H_1 = -1 )</td>
<td>( 0 &gt; H &gt; -0.4 )</td>
</tr>
<tr>
<td>( H_1 = -2 )</td>
<td>( -0.4 &gt; H )</td>
</tr>
<tr>
<td>( S_1 = 1 )</td>
<td>( 1.0 &gt; S )</td>
</tr>
<tr>
<td>( S_1 = 0 )</td>
<td>otherwise</td>
</tr>
<tr>
<td>( F_1 = 1 )</td>
<td>( 1.0 &gt; F )</td>
</tr>
<tr>
<td>( F_1 = 0 )</td>
<td>otherwise</td>
</tr>
<tr>
<td>( CF_1 = 2 )</td>
<td>( CF = \text{strong turn} )</td>
</tr>
<tr>
<td>( CF_1 = 1 )</td>
<td>( CF = \text{weak turn or random coil} )</td>
</tr>
<tr>
<td>( CF_1 = 0 )</td>
<td>otherwise</td>
</tr>
<tr>
<td>( RG_1 = 2 )</td>
<td>( RG = \text{strong turn} )</td>
</tr>
<tr>
<td>( RG_1 = 1 )</td>
<td>( RG = \text{weak turn or random coil} )</td>
</tr>
<tr>
<td>( RG_1 = 0 )</td>
<td>otherwise</td>
</tr>
</tbody>
</table>

\[
A_1 = \sum_{i=1}^{N} 0.3(H_i) + 0.15(S_i) + 0.15(F_i) + 0.2(CF_i) + 0.2(RG_i)
\]
The antigenic index

Thioredoxin

Myohemerythrin

Fig. 1. The plots shown above were generated from the amino acid sequence of thioredoxin S2 (Escherichia coli) (NBRF code TXEC). The hatched lines above and below the plots refer to the amino acid numbers of the protein sequence. The hydrophilicity, surface probability, flexibility and antigenic index plots were calculated as described in Methods. Secondary structure predictions were performed according to the rules of Chou−Fasman (CF) (1978) or Robson−Gamier (RG) (1978), also as described in Methods. The closed boxes (-•-) show regions of the protein's central core; the open boxes (-•-) represent the four major protruding loops of the protein; and the short bars (—) refer to regions of the protein in which reverse turns are present (see text for details).

Fig. 2. Computer plots generated from the amino acid sequence of myohemerythrin (Themiste zostericola) (NBRF code HRTHM). See legend to Figure 1 for a description of the individual plots. The open boxes (-□-) refer to regions from which 'highly reactive' synthetic peptides were derived (amino acids 3−16, 7−16, 37−46 and 57−66; the hatched boxes (-■■-) refer to regions from which 'intermediately reactive' synthetic peptides were derived (amino acids 63−72, 69−82 and 73−82); and the closed boxes (-■■■-) refer to regions from which 'non-reactive (cold)' peptides were derived (amino acids 26−35, 42−51, 96−109 and 100−109) (see text for details).

well with exposed protein regions, but the absolute magnitude of these peaks is not always meaningful.

Another obvious parameter in determining a protein's surface contour is the thermodynamic influence of hydrophobicity. An arbitrary cut-off point was selected to distinguish between major and secondary hydrophobic peaks. The positive contribution of hydrophilic amino acids in a surface peak is acknowledged here as well as the negative influence of hydrophobic amino acids. Thus, the 'antigenic index' reflects the influence of several different physical parameters. The equation was biased such that hydrophilicity/solvent accessibility contributions represent 45% of the calculated value and flexibility components represent 55% of the value. The actual weightings of the individual components of the antigenic index equation shown in Table I were derived empirically through trial and error.

Results

Although a variety of proteins have been tested using the surface contour predictions of the antigenic index, we have selected...
Fig. 3. Computer plots generated from the amino acid sequence of the pre-S region of the large envelope protein of hepatitis B virus (NBRF code SAVLVD). See legend to Figure 1 for a description of the individual plots. The open boxes (---) refer to regions from which immunogenic synthetic peptides were derived; the short bars (—) represent the protein regions to which the human antibody binding sites have been fine mapped; and the closed boxes (-----) refer to regions from which the synthetic peptides were derived which were bound by rabbit antibodies raised against intact hepatitis B virions (see text for details).

three representative samples for presentation in this paper. Thioredoxin-S2 (Escherichia coli) was used as a model protein for testing the predictive ability of the antigenic index against known structural information (Figure 1). Thioredoxin is a small globular protein for which the three-dimensional crystal structure has been solved (Soderberg et al., 1974; Holmgren et al., 1975). The protein has a central core consisting of three parallel and two anti-parallel β-sheets (residues 2–8, 22–29, 53–58, 77–81 and 88–91). The regions which comprise the central core are clearly indicated on the plot produced by the antigenic index algorithm. None of the other predictive routines shown in Figure 1 show the same degree of predictive accuracy. The X-ray diffraction data for this protein indicate that there are four reverse turns (residues 9–11, 32–35, 49–52 and 68–71) which on the antigenic index profile appear as sharp peaks. Finally, the crystal structure shows four external protruding loops (residues 12–19, 29–37, 60–70 and 81–88) which is in agreement with the surface contour prediction of the antigenic index.

Myohemerythrin was used as a second example to test the predictive ability of the antigenic index. Not only has the crystal structure for this protein been solved but its antigenic determinants have been mapped by peptide analysis (Tanier et al., 1984, and references therein). The correlation between the plot of the antigenic index and the known antigenic determinants is very strong (Figure 2). The anti-peptide antibodies directed toward the amino terminus of this protein have the highest avidity of the 'reactive' peptide antibodies. In general, the amino termini of proteins seem to have potent antigenic determinants (Walter and Doolittle, 1983). This unusual reactivity may be due to an enhanced flexibility of the amino termini relative to internal protein segments. At present, the antigenic index algorithm does not take this phenomenon into account and, consequently, may underestimate such regions. However, the myohemerythrin-specific peptide antibodies displaying the second highest reactivity correspond to the region 37–42. This region appears on the antigenic index as the highest peak of the profile. It should also be noted that the non-reactive peptides correspond to local minima on the antigenic index plot. Recently, Geyson et al. (1987) have studied the chemistry of antibody binding to a protein (myohemerythrin) using X-ray crystallographic data. In this study, they have constructed a surface contour profile for the myohemerythrin. Their surface exposure/shape accessibility plot is essentially superimposable with the predicted contour produced by the plot of the antigenic index [the shape accessibility plot is shown in Geyson et al. (1987)]. Furthermore, the results of Geyson et al. (1987) fully support the notion that all surface exposed regions of a protein are poten-
The antibody index

The amino acid sequence of the pre-S region of the large envelope protein of hepatitis B virus (HBV) has been plotted in Figure 3. Antigenic and immunogenic epitopes of this protein have been mapped and the putative host receptor attachment site identified (Neurath et al., 1984, 1985a,b, 1986a,b). The correlation between the available biological information and the antigenic index profile is striking (see Figure 3). The pre-S protein provides an example of the difficulty in predicting naturally immunogenic determinants. Results using synthetic peptides indicate that antibodies developed during the course of a natural hepatitis infection recognize only two primary regions of the protein, amino acids 12—32 and 120—145 (Neurath, 1984, 1985a,b). The B-cell epitopes of these determinants have been fine mapped to residues 26—32 and residues 132—137 (Neurath et al., 1986). On the other hand, antibodies raised against HBV in rabbits recognize five primary determinants, amino acids 1—12, 12—32, 32—53, 94—117 and 120—145 (Neurath et al., 1986a). Thus, the immune system’s ability to recognize viral determinants is species dependent, a phenomenon also observed with poliovirus (Jameson et al., 1985). Based on analyses of several immunogenically defined proteins, such as the VP1 protein of poliovirus, human immunogenic sites appear as “isolated” peaks in the antigenic index profile, i.e. a peak which is not clustered among other major peaks. All of the surface exposed regions predicted by the antigenic index plot apparently correspond to actual antigenic sites. A priori, it is impossible to predict which of these antigenic determinants are immunogenic in the context of the native virus.

Discussion

Major antigenic/immunogenic determinants (structures which, in the context of the native protein, are naturally recognized by the immune system) seem to correlate well with regions of extraordinary surface exposure and flexibility. We have combined these parameters into a novel algorithm designed to predict potential antigenic sites directly from protein sequence information.

The antigenic index yields a surface profile with a strong overall correlation between the predicted contour of a protein and that obtained by physical and biological methods (in most examples tested agreement was >90%). Because most, if not all, prominently exposed surface regions of a protein represent antigenic peaks, this technique offers a means of evaluating potential regions of antigenicity directly from the primary sequence of a protein. Although the antigenic index is a composite function of several other subroutines, occasionally the individual subroutines, e.g. flexibility or surface probability, will generate plots which are extraordinarily similar to the plot produced by the antigenic index. However, the antigenic index plots offer the greatest consistency and overall accuracy in predicting a protein’s surface contour profile directly from its linear amino acid sequence. The use of such an algorithm will always have inherent limitations. The user must supply biological considerations, e.g. that cytoplasmically exposed regions of transmembrane proteins will not be recognized by the immune system, to the interpretation of generated computer plots. We feel that the antigenic index provides an integrated approach that is well suited for structural and immunological protein analyses.

References

Amsterdam, pp. 113–132.

Received on August 17, 1987; accepted on December 31, 1987.

Circle No. 32 on Reader Enquiry Card