Prediction of sequential antigenic regions in proteins

Gjalt W. Welling, Wicher J. Weijer, Ruurd van der Zee and Sytske Welling-Wester

Laboratorium voor Medische Microbiologie, Rijksuniversiteit Groningen, Oostersingel 59, 9713 EZ Groningen, The Netherlands

Received 1 July 1985

Prediction of antigenic regions in a protein will be helpful for a rational approach to the synthesis of peptides which may elicit antibodies reactive with the intact protein. Earlier methods are based on the assumption that antigenic regions are primarily hydrophilic regions at the surface of the protein molecule. The method presented here is based on the amino acid composition of known antigenic regions in 20 proteins which is compared with that of 314 proteins [(1978) Atlas of Protein Sequence and Structure, vol. 5, suppl. 3, 363–373]. Antigenicity values were derived from the differences between the two data sets. The method was applied to bovine ribonuclease, the B-subunit of cholera toxin and herpes simplex virus type 1 glycoprotein D. There was a good correlation between the predicted regions and previously determined antigenic regions.

Antigenicity *Prediction* *Antipeptide antibody* *Synthetic peptide*

1. INTRODUCTION

The preparation of antibodies against synthetic protein fragments which are reactive with the intact protein is a rapidly growing field of investigation with many applications, e.g. synthetic vaccines, detection of gene products, isolation of proteins [1].

Methods have been presented to locate hydrophilic regions in a protein [2,3], since it was argued that antigenic determinants are surface-located and often contain charged and polar residues [2]. These methods are very useful to obtain a rough estimate of potentially antigenic regions. However, as shown by Hopp and Woods [2] not all antigenic regions are hydrophilic and not all hydrophilic regions are antigenic. Therefore we developed a predictive method based on the percentage of each amino acid present in known antigenic determinants compared with the percentage of the amino acids in the average composition of a protein.

The procedure was applied to 3 proteins for which partial evidence of antigenic regions was previously present, i.e. bovine pancreatic ribonuclease [4,5], the B-subunit of cholera toxin [6] and the N-terminal part of the herpes simplex virus type 1 glycoprotein D (unpublished; [7]).

2. METHOD

The observation that in addition to hydrophilic amino acids hydrophobic amino acid residues are often present in antigenic determinants was the starting point of our method. Antigenic determinants of a number of proteins were analyzed with respect to their amino acid composition. Of homologous proteins the antigenic determinants of only one of the proteins and the amino acid residues which differed in these regions were included. Antigenic regions and residues from the following 7 proteins [2] were used to determine the average composition of an antigenic determinant: (1) sperm whale myoglobin (residues 15–22, 56–62, 94–99, 113–119, 145–151); (2) hen egg white lysozyme (residues 5, 7, 13, 14, 33, 34, 62, 87, 89, 93, 96, 97, 113, 114, 116, 125); in addition, residues 65–79 of the lysozyme loop were included

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/85/$3.30 © 1985 Federation of European Biochemical Societies

[10]; (3) ferredoxin from *Clostridium pasteurianum* (residues 1–7, 51–55); (4) bovine myelin basic protein (residues 64–73, 74–85, 113–121, 153–166); (5) human IgG heavy chain constant regions (residues of the 214 CH1 domain, 296 and 309 of the CH2 domain and 355–358 of the CH3 domain); (6) bovine α-lactalbumin (residues 10–18, 60–80, 91–94, 105–117); (7) leghemoglobin (residues 15–23, 52–59, 92–98, 107–116, 132–142). In addition, based on other or more recently published evidence, the following residues were included: (8) horse heart cytochrome *c* and cytochromes *c* from different species (residues 11–15, 44, 46–50, 58–62, 88–92, 96, 103) [2,8]; (9) human hemoglobin α (residues 15–23, 49–56, 82–94, 102–107, 121–127) [9]; the antigenic regions from human hemoglobin β are sufficiently different to be included; (10) human hemoglobin β (residues 13–24, 27–38, 72–84, 108–119, 134–146) [9]; (11) serum albumin (bovine and human); the numbering of human serum albumin was used (residues 138–147, 170–181, 309–315, 360–363, 527–536, 560–566, 554–557) [9]; (12) tobacco mosaic virus (vulgare) coat protein (residues 1–10, 34–39, 55–61, 62–68, 80–90, 105–112, 153–158) [11]; (13) foot and mouth disease virus VP1 (residues 200–213, 144–159) [12,13]; (14) influenza virus hemagglutinin of type A (residues 91–108) [14]; also included were antigenic residues in the original and mutant hemagglutinins detected by monoclonal antibodies (residues 53, 133, 143–145, 205) [15]; (15) hepatitis B surface antigen (residues 140–146) [2,16]; (16) small t-antigen of SV40 (residues 169–174); middle T-antigen of polyoma virus (C-terminal hexapeptide and residues 311–319) [18,19]; (18) *Streptococcus pyogenes* M protein (residues 18–29 of the CB7 fragment) [20]; (19) diphtheria toxin (residues 188–201) [21]; (20) poliovirus type 1 (Mahoney) and type 3 (Leon) VP1 (residues 95–103) [22,23].

2.1. *Antigenicity values*

The percentage of each amino acid in antigenic regions is listed in table 1 as well as that of each amino acid in the amino acid composition derived from the amino acid sequences of 314 families of sequences [24]. From these data the relative occurrence of each amino acid in an antigenic region was calculated. The antigenicity value of each amino acid was expressed as the $\log_{10}$ relative occurrence (see table 1).

2.2. *Computerization and prediction of regions for peptide synthesis*

To facilitate analysis of proteins with unknown antigenic regions, a program similar to that of Hopp and Woods [2] was developed. With this program, antigenicity values of regions containing a certain number of amino acids can be repetitively averaged down the peptide chain. Since the approximate size of an antigenic determinant is 6–7 amino acids, an average group length of 7 was chosen. This is also in agreement with evidence [5] which shows that in order to produce antipeptide antibodies reactive with the intact protein, the peptide should be at least 7 amino acids long. However, a higher immune response can be obtained with longer peptides [5]. Therefore, regions of 7 amino acids with relatively high antigenicity values ($\geq 5 \times 10^{-2}$) are extended to 11–13 amino

Table 1

	Percentage in antigenic region	Percentage in average protein	Relative occurrence	Antigenicity value[a] ($\log_{10}$ relative occurrence)
Ala	11.2	8.6	1.30	0.115
Gly	5.5	8.4	0.66	−0.184
Leu	8.8	7.4	1.19	0.075
Ser	6.6	7.0	0.94	−0.026
Val	6.4	6.6	0.97	−0.013
Lys	10.6	6.6	1.61	0.206
Thr	5.5	6.1	0.90	−0.045
Glu	5.1	6.0	0.85	−0.071
Asp	6.4	5.5	1.16	0.065
Pro	4.6	5.2	0.89	−0.053
Arg	5.6	4.9	1.14	0.058
Ile	2.3	4.5	0.51	−0.292
Asn	3.6	4.3	0.84	−0.077
Gln	3.8	3.9	0.97	−0.011
Phe	2.6	3.6	0.72	−0.141
Tyr	3.5	3.4	1.03	0.013
Cys	2.2	2.9	0.76	−0.120
His	4.1	2.0	2.05	0.312
Met	0.7	1.7	0.41	−0.385
Trp	1.0	1.3	0.77	−0.114

[a] Antigenicity values were directly derived from the percentages of amino acids

acids, depending on the antigenicity values of the neighboring residues.

3. RESULTS AND DISCUSSION

A total of 606 amino acids from 20 proteins was used to determine the average composition of an antigenic region. This resulted in the approximate composition of an antigenic region, since some of the regions have been defined exactly, e.g. lysozyme and myoglobin, while from others a stretch of 20 amino acids is given which contains an antigenic determinant (e.g. lactalbumin). This composition was compared with that of 314 proteins from the Atlas of Protein Sequence and Structure 1978 [24]. This composition was chosen rather than the most recent data in the Protein Sequence Database [25], because, since 1978, many amino acid sequences have been added with signal peptides and transmembrane segments. It is assumed that these parts of a protein do not play a direct role in the antigenicity. Indeed, the 1984 data set from 2676 proteins shows that 20.3% large hydrophobic residues (Leu, Ile, Val) are present, instead of 18.5% in 1978. His, Lys, Ala, Leu, Asp and Arg, in decreasing order, occur more often in an antigenic region than in the average composition of 314 proteins. Met, Ile, Gly, Phe, Cys and Trp are less frequently found in an antigenic region. The other amino acids show a relative occurrence between 0.84 (Asn) and 1.03 (Tyr).

The antigenicity values were used to predict antigenic regions of an arbitrary length of 12 amino acids in bovine ribonuclease, the B-subunit of cholera toxin and herpes simplex virus type 1 glycoprotein D. The antigenicity plots (– – –) are shown in fig.1 and compared with the hydrophilicity values (——) obtained by the method of Hopp and Woods [2]. In ribonuclease 4 antigenic regions were predicted, 1–12, 32–43, 46–57 and 95–106. From the antigenic reactivity of homologous pancreatic ribonuclease with antiserum directed against bovine ribonuclease A, it was deduced that residues 34, 35, 103, 50 and/or 99 and possibly 19 and 37 are part of antigenic reactive regions [4]. In addition, it was determined recently that antibodies against the N-terminal region (residues 1–7, 1–13 and 1–20) reacted with intact ribonuclease [5].

In the B-subunit of cholera toxin regions 7–18, 56–67, 75–86 and 88–99 were predicted. Jacob et al. [6] synthesized 6 peptides: 8–20, 30–42, 50–64, 69–85, 75–85 and 83–97. Immunoprecipitation experiments showed that antibodies against peptides 8–20, 50–64 and 83–97 showed the highest cross-reactivity with the intact toxin. Antibodies against peptide 50–64 could also neutralize the biological activity of the toxin.

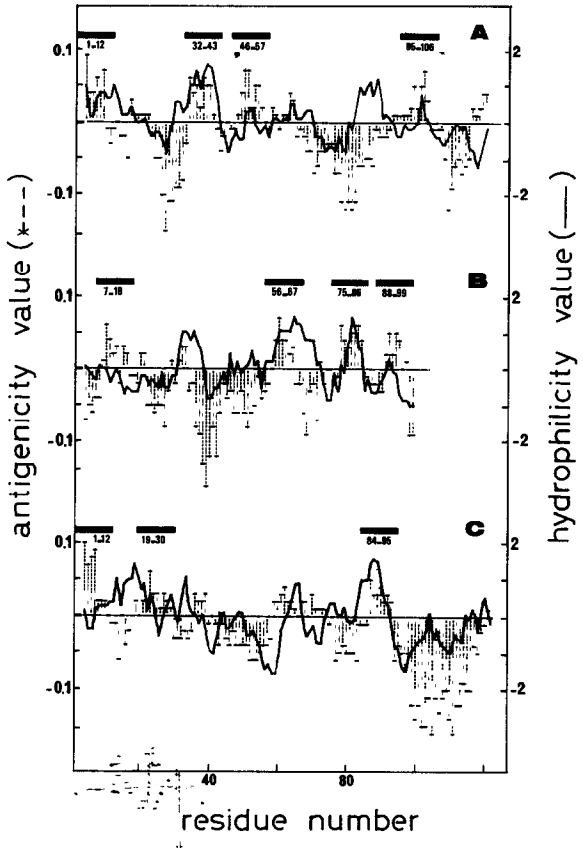


Fig.1. Prediction of antigenic regions in (A) bovine ribonuclease, (B) the B-subunit of cholera toxin and (C) the N-terminal part of herpes simplex virus type 1 glycoprotein D. An average group length of 7 amino acids was chosen. The antigenicity value of each segment of 7 amino acids was plotted at the position of the fourth residue (– – –). Regions of 12 amino acids (indicated by bars) were predicted in the following manner: heptapeptides with high antigenicity values ($\geq 5 \times 10^{-2}$) were, depending on the antigenicity values of neighboring residues, extended to 12 amino acids. For comparison, the hydrophilicity values obtained by the method of Hopp and Woods [2] with an average group length of 7 amino acids are shown as a continuous line.

In the N-terminal part of herpes simplex virus type 1 glycoprotein D, regions 1–12, 19–30 and 84–95 were predicted. Cohen et al. [7] have determined that antibodies against 8–23 could neutralize virus infectivity. We have synthesized peptides 1–13 and 9–21. Antibodies against these peptides could neutralize herpesvirus infectivity in vitro (unpublished). The antigenicity of 84–95 has not yet been investigated. A comparison with the method of Hopp and Woods [2] shows that of the antigenic regions, 46–57 and 95–106 in ribonuclease and 7–18 and 88–99 in cholera toxin were not predicted by the latter method. In the herpes simplex virus glycoprotein, several peptides with amino acid sequences comprising parts of the N-terminal amino acid sequence appeared to be antigenic [7], which was predicted by both methods. All peptides predicted by our method turned out to be antigenic, or did contain antigenic residues, except for cholera toxin peptide 75–86, while the herpes simplex virus peptide 84–95 has not yet been investigated. These results show that with a high success rate peptides are predicted for synthesis which may elicit antibodies reactive with the intact protein and in some cases interfere with the biological activity.

REFERENCES

[1] Lerner, R.A. (1982) Nature 299, 592–596.
[2] Hopp, T.P. and Woods, K.R. (1981) Proc. Natl. Acad. Sci. USA 78, 3824–3828.
[3] Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105–132.
[4] Welling, G.W. and Groen, G. (1976) Biochim. Biophys. Acta 446, 331–335.
[5] Welling, G.W. and Fries, H. (1985) FEBS Lett. 182, 81–84.
[6] Jacob, C.O., Sela, M. and Arnon, R. (1983) Proc. Natl. Acad. Sci. USA 50, 7611–7615.
[7] Cohen, G.H., Dietzschold, B., Ponce De Leon, M., Long, D., Golub, E., Varrichio, A., Pereira, L. and Eisenberg, R.J. (1984) J. Virol. 49, 102–108.
[8] Urbanski, G.J. and Margoliash, E. (1977) J. Immunol. 118, 1170–1180.
[9] Atassi, M.Z. (1984) Eur. J. Biochem. 145, 1–20.
[10] Arnon, R., Maron, E., Sela, M. and Anfinsen, C.B. (1971) Proc. Natl. Acad. Sci. USA 68, 1450–1455.
[11] Westhof, E., Altschuh, D., Moras, D., Bloomer, A.C., Mondragon, A., Klug, A. and Van Regenmortel, M.H.V. (1984) Nature 311, 123–126.
[12] Bittle, J.L., Houghten, R.A., Alexander, H., Shinnick, T.M., Sutcliffe, J.G., Lerner, R.A., Rowlands, D.J. and Brown, F. (1982) Nature 298, 30–33.
[13] Pfaff, E., Mussgay, M., Böhm, H.O., Schulz, G.E. and Schaller, H. (1982) EMBO J. 1, 869–874.
[14] Müller, G.M., Shapira, M. and Arnon, R. (1982) Proc. Natl. Acad. Sci. USA 79, 569–573.
[15] Webster, R.G., Laver, W.G., Air, G.M. and Schild, G.C. (1982) Nature 296, 117–121.
[16] Bhatnagar, P., Papas, E., Blum, H.E., Milich, D.R., Nitechi, D., Karels, M.J. and Vyas, G.N. (1982) Proc. Natl. Acad. Sci. USA 79, 4400–4404.
[17] Harvey, R., Faulkes, R., Gillett, P., Lindsay, N., Paucha, E., Bradbury, A. and Smith, A.E. (1982) EMBO J. 1, 473–477.
[18] Walter, G., Hutchinson, M.A., Hunter, T. and Eckart, W. (1981) Proc. Natl. Acad. Sci. USA 78, 4882–4886.
[19] Schafhausen, B., Benjamin, T.L., Pike, L., Casnellie, J. and Krebs, E. (1982) J. Biol. Chem. 257, 12467–12470.
[20] Beachy, E.H., Seyer, J.M., Dale, J.B., Simpson, W.A. and Kang, A.H. (1981) Nature 292, 457–459.
[21] Audibert, F., Jolivet, M., Chedid, L., Alouf, J.E., Boquet, F., Rivaille, P. and Siffert, O. (1981) Nature 289, 593–594.
[22] Evans, D.M.A., Minor, P.D., Schild, G.S. and Almond, J.W. (1983) Nature 304, 459–462.
[23] Emini, E.A., Jameson, B.A. and Wimmer, E. (1983) Nature 304, 699–703.
[24] Dayhoff, M.O., Hunt, L.T. and Hurst-Calderone, S. (1978) in: Atlas of Protein Sequence and Structure, vol.5, suppl.3, pp.363–373, Natl. Biomed. Res. Foundn, Washington DC.
[25] Barker, W.C., Hunt, L.T., Orcutt, B.C., George, D.G., Yeh, L.S., Chen, H.R., Blomquist, M.C., Johnson, G.C., Seibel-Ross, E.I. and Ledley, R.S. (1984) Atlas of Protein Sequence and Structure – Protein Sequence Database, version 1; Natl. Biomed. Res. Foundn, Washington DC.