Volume 276, number 1,2, 172-174 FEBS 09210 December 1990

A semi-empirical method for prediction of antigenic determinants on protein antigens

A.S. Kolaskar and Prasad C. Tongaonkar

Biotechnology Training Programme, Department of Zoology, University of Poona, Pune-411 007, India

Received 15 August 1990; revised version received 19 October 1990

Analysis of data from experimentally determined antigenic sites on proteins has revealed that the hydrophobic residues *Cys*, *Leu* and *Val*, if they occur on the surface of a protein, are more likely to be a part of antigenic sites. A semi-empirical method which makes use of physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes was developed to predict antigenic determinants on proteins. Application of this method to a large number of proteins has shown that our method can predict antigenic determinants with about 75% accuracy which is better than most of the known methods. This method is based on a single parameter and thus very simple to use.

Antigenic determinant; Protein; Prediction; Semi-empirical method; Sequential antigenic site

1. INTRODUCTION

The delineation of B cell epitopes on protein antigens has attracted the attention of several scientists in recent years [1-5]. Identification of epitopes on proteins would be useful for diagnostic purposes and also in the development of peptide vaccines [6]. To aid experimental workers, Hopp and Woods have developed a method for prediction of antigenic determinants [7]. The approach of Hopp and Woods has been modified to take into account the fact that antigenic sites are on the surface of the protein and most surface residues are antigenic. Recently, Parker et al., have used three parameters - hydrophilicity, accessibility and flexibility - to predict B cell epitopes using a composite plot [8]. This method has improved prediction of antigenic determinants as compared to Hopp and Woods' method, but misses some of the experimentally observed determinants. On the other hand, Welling et al. have calculated the antigenicity value for each amino acid from its frequency of occurrence in epitopes and used these values to predict epitopes [9]. The database used by these workers is very small and consists of only 606 amino acids from 20 proteins. Therefore, frequency values calculated have large errors. We have derived a parameter using experimental antigenic determinant data and physicochemical properties of amino acids. Using this single parameter, a method has been developed to predict antigenic determinants which works with about 75% accuracy.

Correspondence address: A.S. Kolaskar, Biotechnology Training Programme, Department of Zoology, University of Poona, Pune-411 007, India

2. MATERIALS AND METHODS

2.1. *Calculation of antigenic propensity (A_p) values*

169 antigenic determinants are experimentally determined in 34 different proteins (see Table II for the list of proteins) [2-5]. Out of these 169 known antigenic determinants, 156 which have less than 20 amino acid residues per determinant, were used in this study. These 156 experimentally determined antigenic determinants contained 2066 amino acid residues. Using this data, frequency of occurrence of each type of amino acid in antigenic determinants (f_{Ag}) was calculated. Residues on the surface of protein were predicted using the following procedure. Hydrophilicity (P_h), accessibility (P_a) and flexibility (P_f) values given in Table II of Parker et al. for 20 proteinous amino acids were used [8]. In a given protein using these parameter values, averaged, $<P_h>$, $<P_a>$, and $<P_f>$ were calculated for every overlapping heptapeptide from N-terminal to C-terminal. These values were assigned to the middle $(i+3)$ residue in every segment. A residue 'i' was considered to be on the surface if: $<P_{ai}> > \bar{P}_a$ or $<P_{hi}> > \bar{P}_h$ or $<P_{fi}> > \bar{P}_f$, where $\bar{P}_h$, $\bar{P}_a$ and $\bar{P}_f$ are the average values for the protein. Using this data, the frequency of occurrence of amino acids on the surface (f_s) was calculated. The antigenic propensity (A_p) value for each of the amino acids was calculated using the relation -

$$A_p = f_{Ag}/f_s$$

2.2. *Algorithm to predict antigenic determinants*

Step 1: The average antigenic propensity values $<A_p>$ for each of the overlapping heptapeptides from N-terminal to C-terminal of the protein were calculated. These average values were assigned to the fourth, $(i+3)$, residue in the segment.

Step 2: the average antigenic propensity ($\bar{A}_p$) value for the protein was determined.

Step 3: if $\bar{A}_p \geqslant 1.0$ then those residues having $<A_p> \geqslant 1.0$ were termed as potential antigenic residues. If $\bar{A}_p < 1.0$ then those residues having $<A_p> \bar{A}\bar{P}_p$ were termed as potential antigenic residues.

Step 4: to pick up antigenic determinants a condition was set that *six* consecutive residues must satisfy step 3.

This algorithm was used to predict antigenic determinants on 34 proteins for which some experimental results are available. The computer program is available for PC compatible systems.

Published by Elsevier Science Publishers B.V. (Biomedical Division)

Table I

Occurrence of amino acids in epitopes, proteins and on the surface, and their antigenic propensity values

Amino acid	Occurrence of amino acids in			f_{Ag}	A_p
	Epitopes	Surface	Protein		
A	135	328	524	0.065	1.064
C	53	97	186	0.026	1.412
D	118	352	414	0.057	0.866
E	132	401	499	0.064	0.851
F	76	180	365	0.037	1.091
G	116	343	487	0.056	0.874
H	59	138	191	0.029	1.105
I	86	193	437	0.042	1.152
K	158	439	523	0.076	0.930
L	149	308	684	0.072	1.250
M	23	72	152	0.011	0.826
N	94	313	407	0.045	0.776
P	135	328	411	0.065	1.064
Q	99	252	332	0.048	1.015
R	106	314	394	0.051	0.873
S	168	429	553	0.081	1.012
T	141	401	522	0.068	0.909
V	128	239	515	0.062	1.383
W	19	55	103	0.009	0.893
Y	71	158	245	0.034	1.161
Total	2066	5340	7944		

Table II

Results of application of our method to predict antigenic determinants

Protein	N	C	M
(1) IL-3 (mouse)	9	6	3 (2)
(2) Interferon-1 (human)	3	3	0 (0)
(3) Interferon (human)	1	0	1 (1)
(4) Coat protein (TMV)	8	6	2 (4)
(5) M5 protein (*S. pyogenes*)	3	2	1 (3)
(6) Myohemerythrin	8	5	3 (3)
(7) HCG/B	10	7	3 (4)
(8) Rotavirus s11/VP6	3	2	1 (2)
(9) Erythropoietin	6	6	0 (2)
(10) MBP (human)	8	7	1 (2)
(11) MBP (bovine)	8	7	1 (0)
(12) Myoglobin (sperm whale)	5	4	1 (1)
(13) Myoglobin (bovine)	3	2	1 (0)
(14) Poliovirus 1/VP2	2	1	1 (1)
(15) Poliovirus 1/VP1	15	9	6 (8)
(16) Cholera toxin/B	11	8	3 (6)
(17) Labile toxin/B (*E. coli*)	2	2	0 (1)
(18) α-Lactalbumin (bovine)	3	2	1 (0)
(19) Serum albumin (human)	3	3	0 (1)
(20) Thyroglobulin (human)	1	1	0 (1)
(21) Apamine (honey bee)	1	1	0 (0)
(22) Renin (human)	7	5	2 (4)
(23) MS II protein	7	6	1 (0)
(24) Ferredoxin (*Clostridium*)	2	2	0 (1)
(25) Lysozyme (hen egg)	4	3	1 (2)
(26) Delta antigen (HDV)	10	5	5 (3)
(27) Rhinovirus/14/VP3	1	1	0 (0)
(28) Rhinovirus/14/VP1	1	1	0 (0)
(29) Cytochrome *c* (horse)	1	0	1 (0)
(30) Cytochrome *c* (bovine)	1	0	1 (0)
(31) AChR (*Torpedo*) α	12	8	4 (5)
(32) AChR (*Torpedo*) δ	3	2	1 (1)
(33) AChR (human) α	3	2	1 (1)
(34) Interferon γ (mouse)	1	0	1 (1)
Total	169	122	47 (60)

N = total number of sites determined experimentally; C = number of sites correctly predicted using our method; M = number of sites missed (number of sites missed by the method of Parker et al. given in brackets).

3. RESULTS AND DISCUSSION

As can be seen from Table I, Ser, Lys, Thr, Glu and Ala occur with relatively high frequency in antigenic determinants. However, these frequency values can be misleading since it is known that several amino acids, among these, also occur with high frequency in total protein (this can be seen from Table I, which gives the occurrence of amino acids in 34 proteins considered). We are aware that complete antigenic structure of very few proteins is known today and thus f_{Ag} can change. Due to such changes, A_p values given here are tentative. However, it is interesting to note that A_p values are very large for Cys, Val and Leu which are hydrophobic amino acids. Thus, whenever these residues occur on the surface, they are likely to be a part of antigenic determinants. This we consider a very important observation from our present analysis. Results of application of the algorithm to 34 proteins are given in Table II. Number of sites missed are fewer by our method than by the method of Parker et al. (see Table II). As seen from Table II, our of the 169 experimentally known antigenic determinants, our method has correctly picked up 122 antigenic determinants with average accuracy of about 75%. Thus our method can pick up antigenic determinants with good accuracy. Number of additional sites predicted by our method are small in well-studied proteins such as IL-3, Coat protein/TMV, Erythropoietin and Myohemerythrin, in which the number of sites missed are 0, 2, 2 and 1, respectively. In the light of these results we feel that most of the additional sites predicted by our method are likely to be antigenic determinants. Thus, in short, a single-parameter based semi-empirical method which judiciously makes use of physicochemical properties of amino acid residues and experimental data, is developed to predict antigenic determinants, and its accuracy has been tested by application to a large number of proteins.

Acknowledgements: We would like to thank Mr Stephen Samuel for doing the computer programming and Ms Vijaya Avhad for typing the manuscript. Facilities used at Bioinformatics, DIC, Pune are acknowledged.

REFERENCES

[1] Getzoff, E.D., Tainer, J.A. and Lerner, R.A. (1988) in: The Chemistry and Mechanism of Antibody Binding to Protein Antigens, Adv. Immunol., vol. 43, (Dixon, F.J. ed) pp. 1-98, Academic Press, London.

[2] Milich, D.R. (1989) in: Synthetic T and B Cell Recognition Sites: Implications for Vaccine Development, Adv. Immunol. vol. 45) (Dixon, F.J. ed.) pp. 195-282, Academic Press, London.

[3] Dyall-Smith, M.L., Lazdins, I., Tregear, G.W. and Holmes, S.M. (1990) Proc. Natl. Acad. Sci. USA 83, 3465-3468.

[4] Wang, J., Jansen, R.W., Brown, E.A. and Lemon, S.M. (1990) J. Virol. 64, 1108-1116.

[5] Atassi, M.Z. (1984) Eur. J. Biochem. 145, 1-20.

[6] Schmidt, A.M. (1989) Biotect. Adv. 7, 187-213.

[7] Hopp, T.P. and Woods, K.R. (1981) Proc. Natl. Acad. Sci. USA 78, 3824-3828.

[8] Parker, J.M.R., Guo, D. and Hodges, R.S. (1986) Biochemistry 25, 5425-5432.

[9] Welling, G.W., Weijer, W.J., van der Zee, R. and Welling-Wester, S. (1985) FEBS Lett. 188, 215-218.