Prediction of Chain Flexibility in Proteins

A Tool for the Selection of Peptide Antigens

P.A. Karplus and G.E. Schulz
Institut für Organische Chemie und Biochemie der Universität,
D-7800 Freiburg i.Br.

Today, a multitude of genes have been sequenced of which the respective proteins remain unknown. These proteins can be identified, localized, and purified by using antibodies raised against oligopeptides that correspond to segments of the hypothetical protein sequence [1, 2]. In special cases, such oligopeptides may even give rise to synthetic vaccines [3, 4]. Currently, the selection of oligopeptide stretches from a protein sequence is based on schemes designed to predict segments of high antigenicity [5], hydrophilicity [6, 7], or reverse-turn potential [8, 9]. However, it has recently been demonstrated that segmental flexibility is more indicative of an antigenic determinant than the selection criteria mentioned above [10], and that it is also better suited for selecting crossreacting peptides [11]. Accordingly, we have analyzed 31 refined protein structures to develop a method for predicting flexible segments from a given amino acid sequence.

The data base used for the prediction of chain flexibility consisted of 31 proteins (given in Fig. 2) of known three-dimensional structure, as deposited in the Protein Data Bank, Brookhaven, USA. The protein structures selected had been refined with individual atomic temperature factors (i.e. B-values); they had more than 30 residues, their resolution was better than or equal to 0.3 nm, and they were at least 50% different in sequence from all other included proteins.

As a measure for chain flexibility we chose the temperature factors, i.e. B-values, of the C_α atoms. An inspection of experimental data showed that the averages and the spreads of B-values varied greatly from protein to protein, which presumably reflects differences in structure refinement methods and stages more than natural variances. In order to avoid bias towards proteins with extreme averages or spreads, the B-value of each C_α atom was normalized following the equation

$$B_{\text{norm}} = (B + D_p)/(<B>_p + D_p),$$

in which $<B>_p$ is the average B-value of all C_α atoms of protein p, omitting the 3-N- and the 3-C-terminal residues. The average B_{norm} of a protein is always 1.0. The value of D_p for a given protein p was chosen in such a way that the root mean square deviation of the B_{norm}-values was 0.3. Before adjustment using D_p the majority of the 31 proteins showed root mean square deviations between 0.2 and 0.4.

The chain ends of proteins are known to have an above average antigenicity [12]. To determine whether this is correlated with flexibility, statistics were compiled on the first (1, 2, ... 20) and the last ($m-19$, $m-18$, ... m) residues of all chains. The average B_{norm}-values of residues 1, 2, 3, $m-1$, m were 1.60, 1.28, 1.17, 1.27, 1.53, respectively, whereas the remaining 35 positions had values fluctuating between 0.87 and 1.07. Thus, the chain termini are exceptionally flexible.

Next, we established the average relationship between B_{norm}-value and amino acid type (Fig. 1a). For flexibility prediction these single-residue statistics were refined by a nearest-neighbor analysis. First, the 20 amino acid types were divided into 2 groups, "rigid" and "flexible". Rigid residue types are those with average B_{norm}-values less than 1.0 (i.e. A, L, H, V, Y, I, F, C, W, M). Then, separate average B_{norm}-values were determined for residues with no rigid neighbors, residues with one rigid neighbor, and residues for which both neighbors are rigid. The resulting neighbor-correlated B_{norm}-values are given in the DATA statements for BNORMØ, BNORM1, and BNORM2 in Fig. 2. There is a striking nearest-neighbor effect, for instance, B_{norm} of S changes from 16.9% above average for no rigid neighbor to 7.7% below average for 2 rigid neighbors.

The predicted relative flexibility at residue position n of a given amino acid

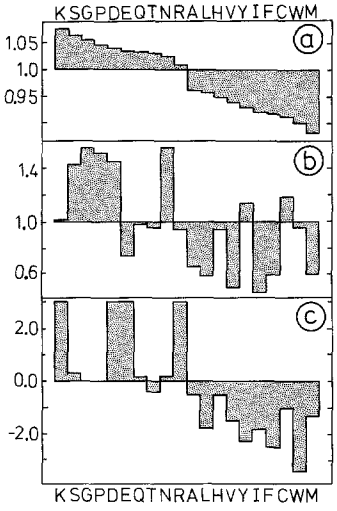

Fig. 1. Single-residue coefficients. a) Average B_{norm}-values derived from all 31 proteins of the data base. The first 3 and the last 3 residues of each chain were excluded. b) Reverse-turn conformational potentials of [8]; c) antigenicity according to [5]

```
      PROGRAM FLEXPLOT
c*************************************************************************
c     The input to this program is an amino acid sequence in
c     the one letter code which is entered from the terminal.
c     Based on amino acid flexibility coefficients the
c     predicted flexibility profile is calculated.
c*************************************************************************
c     The Brookhaven Protein Data Bank abbreviations for the
c     31 proteins selected are:      2ACT, 2CPVB, 1ICB, 5CPA,
c     2GCH, 1CRN, 4CYT, 1CZCB,  351C, 3DFR, 1EDD, 2FD1, 1HMQ,
c     1HDS, 1FB4, 1INS,  1LH2,  1LYM, 1MBD, 1NXB, 1SN3, 10V0,
c     1BP2, 1PCY, 1SGA,  4RSA, 2RXNA, 2SOD, 3TLN, 2PTN, 4PTI
c*************************************************************************
      CHARACTER*1 SEQ,AA,AAIN,PLOT,ILINE,IF,IBLANK
      INTEGER ISEQ(999)
      DIMENSION SEQ(999),NAYB(999),FPRED(999),AA(21),WT(7),
     $BSCAN(7),BNORM0(20),BNORM1(20),BNORM2(20),PLOT(31)
      DATA AA /'K','S','G','P','D','E','Q','T','N',
     $ 'R','A','L','H','V','Y','I','F','C','W','M','^'/
c*****flexibility parameters for no 'rigid' neighbors
      DATA BNORM0/1.093,1.169,1.142,1.055,1.033,1.094,1.165,
     $1.073,1.117,1.038,1.041,0.967,0.982,0.982,0.961,1.002,
     $0.930,0.960,0.925,0.947/
c*****flexibility parameters for one 'rigid' neighbor
      DATA BNORM1/1.082,1.048,1.042,1.085,1.089,1.036,1.028,
     $1.051,1.006,1.028,0.946,0.961,0.952,0.927,0.930,0.892,
     $0.912,0.878,0.917,0.862/
c*****flexibility parameters for two 'rigid' neighbors
      DATA BNORM2/1.057,0.923,0.923,0.932,0.932,0.933,0.885,
     $0.934,0.930,0.901,0.892,0.921,0.894,0.913,0.837,0.872,
     $0.914,0.925,0.803,0.804/
      DATA WT /0.25,0.50,0.75,1.00,0.75,0.50,0.25/
      DATA IBLANK,ILINE,IF /' ','|','F'/
c*****type the sequence in one letter code; finish with "^".
      TYPE 101
      DO 2 I=1,999
      READ (5,100)AAIN
      IF (AAIN.EQ.AA(21)) GOTO 3
      SEQ(I)=AAIN
      DO 1 J=1,20
 1    IF (AAIN.EQ.AA(J)) GOTO 2
      STOP 'Illegal amino acid type'
 2    ISEQ(I)=J
c*****the complete sequence is in
 3    LENGTH=I-1
c*****do nearest neighbor analysis
      DO 4 I=2,LENGTH-1
      NAYB(I)=0
      IF ((ISEQ(I-1).GE.11).OR .(ISEQ(I+1).GE.11))NAYB(I)=1
 4    IF ((ISEQ(I-1).GE.11).AND.(ISEQ(I+1).GE.11))NAYB(I)=2
c*****calculate predicted relative flexibility
      DO 5 I=5,LENGTH-4
      DO 5 J=1,7
      NJ=ISEQ(I-4+J)
      IF (NAYB(I-4+J).EQ.0) BSCAN(J)=BNORM0(NJ)
      IF (NAYB(I-4+J).EQ.1) BSCAN(J)=BNORM1(NJ)
      IF (NAYB(I-4+J).EQ.2) BSCAN(J)=BNORM2(NJ)
 5    FPRED(I)=FPRED(I)+BSCAN(J)*WT(J)/4.0
c*****plot the flexibility profile with mean set to 16
      DO 7 I=1,LENGTH
      DO 6 J=1,31          ! initialize the plot array
 6    PLOT(J)=IBLANK
      PLOT(16)=ILINE       ! mark the mean
      NPOS=NINT((FPRED(I)-1.0)*100.0+16.0)
      IF (NPOS.GE.1.AND.NPOS.LE.31)PLOT(NPOS)=IF
      WRITE (6,102)I,SEQ(I),FPRED(I),PLOT
 7    CONTINUE
      STOP
100   FORMAT(A1)
101   FORMAT(' Enter amino acid sequence in single letter',
     $       ' code'/' one amino acid per line. Signal end',
     $       ' by entering "^".')
102   FORMAT(I5,A5,F8.3,5X,31A1)
      END
```

Fig. 2. Fortran program example for the prediction of chain flexibility. Abbreviations for the 31 proteins used as data base are given. The array AA reflects the ordering of amino acids according to flexibility as given in Fig. 1a. The arrays BNORMØ, BNORM1, and BNORM2 contain the neighbor-correlated B_{norm}-values in the order of AA

sequence is taken as the weighted sum of the neighbor-correlated B_{norm}-values for the amino acids at positions $n-3$, $n-2$, $n-1$, n, $n+1$, $n+2$, and $n+3$ using the weights 0.25, 0.50, 0.75, 1.00, 0.75, 0.50 and 0.25, respectively. A Fortran program for this calculation is given in Fig. 2. The quality of the method can be visualized in Fig. 3, which shows evident correspondence between the predicted and the observed flexibility of lysozyme. Residues 115 through 119 are predicted to be flexible, but in the native structure they are held tight by the disulphide bond between residues 30 and 115. It should be noted that the two highest peaks of the predicted profile correspond to the known continuous epitopes of lysozyme [14].

A comparison of the single-residue coefficients of our scheme with those used for reverse-turn [8] and antigenicity prediction [5] reveals significant differences (Fig. 1). In consequence, the proposed flexibility prediction deviates appreciably from the presently used methods [5–9] and thus provides novel information. The presented scheme used in conjunction with other methods should aid future selection of chain segments suitable for inducing antibodies which crossreact with native proteins.

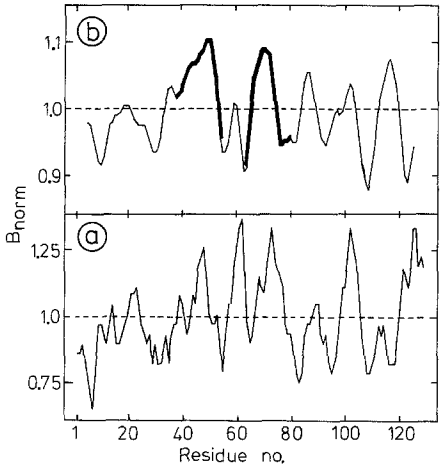

Fig. 3. Profiles for hen egg white lysozyme (entry 1LYM of the Brookhaven Protein Data Bank). a) Experimental B_{norm}-values derived from the X-ray structure of lysozyme [13]. b) Predicted flexibility using the proposed scheme. Coefficients used for calculating this prediction were derived from the data base after eliminating lysozyme (data not shown). Continuous antigenic determinants [14] are indicated in bold

Received January 14, 1985

1. Walter, G., et al.: Proc. Nat. Acad. Sci. USA *77*, 5197 (1980)
2. Sutcliffe, J.G., et al.: Nature *287*, 801 (1980)
3. Bittle, J.L., et al.: ibid. *298*, 30 (1982)
4. Pfaff, E., et al.: EMBO J. *1*, 869 (1982)
5. Hopp, T.P., Woods, K.R.: Proc. Nat. Acad. Sci. USA *78*, 3824 (1981)
6. Rose, G.D.: Nature *272*, 586 (1978)
7. Kyte, J., Doolittle, R.F.: J. Mol. Biol. *157*, 105 (1982)
8. Chou, P.Y., Fasman, G.D.: Biophys. J. *26*, 367 (1979)
9. Garnier, J., Osguthorpe, D.J., Robson, B.: J. Mol. Biol. *120*, 97 (1978)
10. Westhof, E., et al.: Nature *311*, 123 (1984)
11. Tainer, J.A., et al.: ibid. *312*, 127 (1984)
12. Walter, G., Doolittle, R.F., in: Genetic Engineering: Principles and Methods, Vol. 5, p. 61 (Setlow, J.K., Hollaender, A., eds.). New York: Plenum 1983
13. Rao, S.T., Hogle, J., Sundaralingam, M.: Acta Crystallogr. *C39*, 237 (1983)
14. Ibrahimi, I.M., et al.: Mol. Immunol. *17*, 37 (1980); Takagaki, Y., et al.: Biochemistry *19*, 2498 (1980)

Synthesis of Acetylcholine Receptors in *Xenopus* Oocytes Induced by Poly(A)$^+$-mRNA from Locust Nervous Tissue

H. Breer and D. Benke

Abteilung für Zoophysiologie der Universität, D-4500 Osnabrück

This paper demonstrates the synthesis of acetylcholine receptors, as defined by specific α-bungarotoxin (α-BGTX) binding, in *Xenopus* oocytes after microinjection of poly(A)$^+$-mRNA from the nervous system of insects.

The central nervous tissue of insects has been shown to contain high concentrations of α-toxin binding sites [1], which obviously represent functional recpetors for acetylcholine with a distinct nicotinic pharmacology [2, 3] quite similar to the receptors in neuromuscular junctions and in electroplaques. Thus insects offer the possibility to analyse nicotinic acetylcholine receptors which are produced and which operate in nerve cells, not in muscle cells or electrocytes. Recent biochemical analyses have shown that the nicotinic binding site from insect ganglia represents a macromolecule with a sedimentation coefficient of 9–10 S and that it is obviously an oligomer of several identical subunits [4]. The insect receptor thus comprises a molecular structure quite different from the well characterized *Torpedo* receptor.

The very powerful techniques of molecular genetics have greatly facilitated the analysis of the *Torpedo* receptor [6]; if mRNA encoding the receptor polypeptides in insects can be obtained, a molecular approach will give insight into the detailed structure of the insect receptor protein. Such information could elucidate the molecular functioning of the receptor and thus enable an interesting comparison with the molecular structure of the *Torpedo* receptor.

In a first approach, the nervous tissue of locust was probed for receptor-specific mRNA. As an assaying system, the *Xenopus* oocytes have proved very useful in identifying the mRNA and genes coding for the nicotinic acetylcholine receptor from *Torpedo* [7]. Therefore, RNA extracts from locusts were injected into oocytes and analysed with regard to their capability of directing the synthesis of the receptor protein.

RNA was isolated from the head and thoracic ganglia of locust (*Locusta migratoria*) using the guanidinium-isothiocyanate/CsCl-gradient procedure [8] and was subsequently fractionated by oligo(dT)-cellulose chromatography. The poly(A)$^+$-mRNA fraction thus isolated was microinjected into fully grown oocytes which were incubated for 48 h at 20° C. Proteins were then extracted from injected and non-injected oocytes by homogenizing in phosphate buffer, pH 7.4 containing 1% detergent. BGTX binding was de-