Volume 223, number 2, 402–404 FEB 05265 November 1987

Prediction of exposure degree diagram and sites of limited proteolysis in globular proteins as an approach to computer-aided design of protein bioregulators with prolonged action

M.A. Rodionov, S.G. Galaktionov and A.A. Akhrem

Institute of Bioorganic Chemistry, Byelorussian SSR Academy of Sciences and *All-Union Institute of Microbiological Industry, Minsk, USSR*

Received 30 July 1987; revised version received 8 September 1987

In order to prolong the lifetime of protein bioregulators in blood it is possible to engineer analogs with protected sites of limited proteolysis. To determine the sites, primarily accessible to trypsin-like proteases, a computer procedure has been developed, including a prediction algorithm, to produce the residue diagram of a globular protein and a discriminant algorithm to determine the sites most liable to proteolysis. The accuracy of prediction of amino acid residue exposure is characterised by correlation coefficients between experimental and theoretical exposure values, the coefficients being about 0.7 as calculated for 10 globular proteins. The classification of Arg and Lys residues into two groups, susceptible or insusceptible to protease, has an error percentage of about 25.

Protein structure; Limited proteolysis; Protein bioregulator

1. INTRODUCTION

Many protein bioregulators circulating in blood have relatively short half-lives – usually a few minutes. This is quite adequate for the physiological function of native bioregulators, but if they are used as drugs it is often necessary to protect them from breakdown, caused by numerous proteases, especially by the most active one – the trypsin-like protease plasmin. An obvious way to solve this problem is to determine the sites most susceptible to attack by proteases, and to modify selectively the appropriate residues. Of course, there also exists the risk that such residues belong to the active site of the bioregulator and therefore that the modified protein will lose its biological activity; if the active center remains intact, the action time of the analog will be significantly greater.

Enzymatic cleavage of a peptide bond depends on different characteristics of interacting molecular surfaces in the substrate protein and enzyme [1]. Here, we discuss a particular aspect of this problem, considering the degree of exposure of the residues attacked by trypsin-like proteases (Lys or Arg) and their nearest neighbors in the polypeptide chain as factors of their susceptibility to proteolysis.

2. PREDICTION OF THE DEGREE OF AMINO ACID RESIDUE EXPOSURE

The degree of exposure individual residue in globular protein is determined here as the share of free volume (i.e. not filled with atoms) in an 8 Å

Correspondence address: M.A. Rodionov, Institute of Bioorganic Chemistry, Byelorussian SSR Academy of Sciences, Minsk 220600, USSR

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/87/$3.50 © 1987 Federation of European Biochemical Societies

radius sphere with the center placed in the geometrical center of the residue.

The degree of exposure of q_i of the i-th residue in an amino acid sequence is supposed to be linearly dependent on q_{i+j} values of its nearest neighbors, on the share of the l-th type residues in an amino acid composition Z_l and on the number of residues in a given protein, N.

$$q_i^{ks} = \sum_{\substack{j=-4 \\ j\neq 0}}^{4} a_j^{ks} q_{i+j} + g^{ks}, \quad (1)$$

where

$$g^{ks} = \sum_{l=1}^{20} b_l^{ks} Z_l + c^{ks} N + d^{ks}$$

The multiple regression equation coefficients a_j^{ks}, b_l^{ks}, c^{ks} and d^{ks} for each k-th type of amino acid residue and s-th type of secondary structure were obtained using the data on 60 non-homologous proteins [2]. The system of eqns 1 for all residues in a protein may be rewritten in matrix form

$$\vec{Q} = A\vec{Q} + \vec{G} \text{ or } (E - A)\vec{Q} = \vec{G} \quad (2)$$

where E is the unity matrix and the evaluation of vector $\vec{Q}$ is the solution of the linear equation system, eqn 2.

The procedure described above was tested by calculation of the exposure diagrams for 10 proteins of known tertiary structure not included in the above set of 60 proteins; the location of secondary structure elements was determined using the algorithm of Ptitsyn and Finkelstein [3]. Predicted values q_i were compared with experimental ones by calculating the correlation coefficients between two of these sets; these correlation coefficients range from 0.60 to 0.80.

3. EVALUATION OF THE SUSCEPTIBILITY OF BASIC RESIDUES TO THE ATTACK BY TRYPSIN-LIKE PROTEASES

Having found the reliability of the prediction algorithm to be moderately satisfactory, we used it in a procedure that permitted us to evaluate the susceptibility of a given basic residue containing sites to the attack of trypsin-like proteases. This susceptibility is supposed to depend on the steric conditions in the vicinity of the i-th basic residue, which are determined by q_{i+k} values of the residue in question and its neighboring residues; in our calculations the number of neighboring residues k ranged from -3 to 3.

Two sets of exposure degree values for residues Lys or Arg and its neighbors in sites susceptible (group A) and insusceptible (group B) to limited proteolysis were formed using experimental data on the location of such sites in 9 proteins of known three-dimensional structure (chymotrypsinogen A [4], concanavalin [5], cytochrome c and myoglobin [6], elastase [7], insulin [8], ribonuclease S [9], staphylococcal nuclease [10] and trypsin [11]); the cases when Arg or Lys residues preceded the Pro residues are not considered. These sets were used to obtain the parameters of a discriminant analysis algorithm in the form described in [12]. This algorithm provides an accuracy of classification into groups A or B of about 75%.

To test the procedure as a whole, we have predicted the basic residues accessible to trypsin-like enzymes in two proteins, where the sites of limited proteolysis have been determined experimentally: proinsulin and somatotropin. Such sites in proinsulin are residues Arg_{31} and Arg_{65} [13]; in somatotropin – Arg_{134} and Lys_{145} [14]. In both cases the calculation correctly revealed the

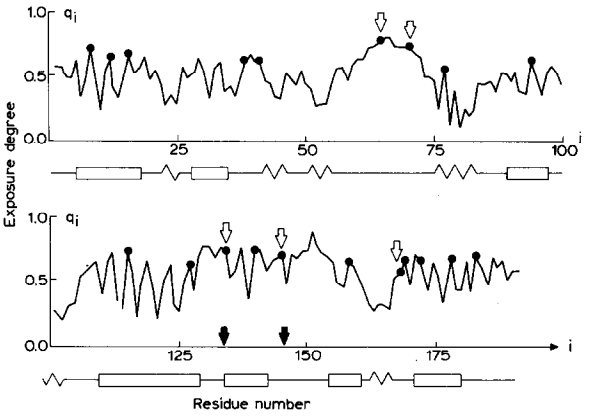

Fig.1. Predicted exposure degree diagram of somatotropin. Residues Arg and Lys are indicated by circles; the unfilled and filled arrows show the predicted and experimentally found [13] sites of trypsin attack. Symbols at the bottom indicate the conformational states predicted by the method [3]: (⟨⟩) α-helix, (⋀⋁⋀) β-sheet, (——) turns or undefined structure.

above residues, but in addition, some other residues were also qualified as possible sites of the primary enzyme attack. In proinsulin the errors concern the nearest neighbors of the attacked residue Arg_{31}, i.e. Lys_{29} and Arg_{32}; in somatotropin the erroneously determined residues are Arg_{64}, Lys_{70} and Lys_{168}. For the last protein the calculated exposure diagram and localization of sites of limited proteolysis – both calculated and experimentally found – are given in fig.1.

Many proteins, being the subject of intensive attention in modern biotechnology, are not available in quantities sufficient for the experimental determination of limited proteolysis sites; therefore, we hope that the approach we have developed, though not quite perfect as yet, can be useful in such cases.

ACKNOWLEDGEMENTS

The authors express their gratitude to Professor O.B. Ptitsyn and Dr A.V. Finkelstein for the computer program for the prediction of protein secondary structure.

REFERENCES

[1] Antonov, V.K. (1983) Chemistry of Proteolysis, Nauka, Moscow.
[2] Bernstein, C., Koetzle, F., Williams, J.B., Meyer, E.F., Brice, D., Rodjers, J.R., Kennard, O., Shimanoucki, T. and Tasumi, M. (1977) Eur. J. Biochem. 80, 319–324.
[3] Ptitsyn, O.B. and Finkelstein, A.V. (1983) Biopolymers 22, 15–25.
[4] Bettelneim, F.R. and Neirath, H. (1955) J. Biol. Chem. 212, 241–253.
[5] Wang, J.L. and Edelman, G.M. (1978) J. Biol. Chem. 253, 3008–3015.
[6] Fong, S.L. and Hargrave, P.A. (1977) Int. J. Peptide Protein Res. 10, 139–145.
[7] Ghelis, C., Tempete-Gaillourdet, M. and Yon, J.M. (1979) Biochem. Biophys. Res. Commun. 84, 31–36.
[8] Grestfield, A.M., Moore, S. and Stein, W.H. (1963) J. Biol. Chem. 238, 622–627.
[9] Allende, J.E. and Richards, F.M. (1962) Biochemistry 1, 295–304.
[10] Taniuchi, H. and Anfinsen, C.B. (1968) J. Biol. Chem. 243, 4778–4786.
[11] Schroeder, D.D. and Shaw, E. (1968) J. Biol. Chem. 243, 2943–2949.
[12] Afifi, A.A. and Azen, S.P. (1979) Statistical Analysis. A Computer Oriented Approach, Academic Press, New York.
[13] Oyer, P.E., Cho, S., Peterson, J.D. and Steiner, D.F. (1971) J. Biol. Chem. 246, 1375–1386.
[14] Graf, L., Choh, H.L. and Jibson, M.D. (1982) J. Biol. Chem. 257, 2365–2369.