FEBS Letters 356 (1994) 188–190

FEBS 14870

Residues responsible for distinct biological functions are characterized by different statistical features of sequential and spatial neighborhoods: a thermolysin example

A.E. Gabrielian[a,*], S.A. Kostrov[b], M.P. Kirpichnikov[a]

[a]*Institute of Molecular Biology, Russian Academy of Sciences, Vavilov St. 32, 117984 Moscow, Russian Federation*
[b]*Institute of Molecular Genetics, Russian Academy of Sciences, Kurchatov Sq. 46, 123098 Moscow, Russian Federation*

Received 24 October 1994

Abstract An investigation of the functional topography of thermolysin was carried out using frequency analysis of its primary and tertiary structures. The statistical validity of this approach was estimated for the enzyme active site, the substrate-binding pocket, the inter-domain interface and calcium-binding sites' predictions. We showed that frequency analysis of primary structure could be employed to predict the localization of contiguous parts of the inter-domain interface. The same approach appears to be unsuitable to a search for conformation-dependent enzyme active sites and substrate-binding pockets. In contrast, frequency analysis of the spatial neighborhood is not effective for predicting the inter-domain interface as distinct from the active site, substrate-binding pocket and calcium-binding sites. These differences should be taken into account when investigating and understanding protein structure–function relationships.

Key words: Active site; Prediction; Statistical approach; Primary and tertiary structure investigation

1. Introduction

It is usually assumed that amino acid sequences contain all the information determining both structure and function of the protein. Predictions of structural and functional features are mainly limited by the complexity of a straight numerical approach to the folding problem. Functionally important sites (FISs) prediction is in some sense similar to secondary structure prediction, namely, they both are usually inaccurate when non-local interactions are dominant. FISs are arbitrarily divided into two main groups [1]. The first one includes sites formed by amino acids situated near to each other in the primary structure of the protein: they are called contiguous, or linear FISs. Some protein–protein interaction sites, nuclear localization signals and DNA–protein interaction sites were prescribed to this group. Another type of FIS is the discrete, or conformation-dependent one. It consists of amino acid residues that are distant in the primary, but close in the tertiary structure of the protein. A typical example of conformational FIS is the enzyme active site. A necessary, but not absolute, prerequisite for conformational FISs predictions is, therefore, a knowledge of the protein spatial structure. In contrast, to search for linear FISs it is often enough to have the amino acid sequence only. Frequency analysis methods [1–5], that originated from Shannon's information theory, were proposed as effective tools for linear FISs searching. Frequency analysis is based on the assumption that probabilities of elements of a protein's structure could be associated with their functional importance. The hypothesis for the biological basis of this approach was formulated in [5]. We supposed that rare oligomers (in the simplest case, rare amino acids, such as Trp, Cys, His) could have a greater probability to being involved in the biological function of the protein. Their low frequency of occurrence (uniqueness) provided the proof against an accidental similarity or identity with regions in other proteins. The uniqueness of a functional site is important because similar sites could cross-react (mimic each other) and by this way proteins could occasionally interact with 'non-self' receptors and ligands. Such an occasional contact will interfere with the normal biological action of the receptor–ligand system. To prevent this, it is important that the structure of functional sites cannot be mimicked by other sites. Using the classical 'lock-and-key' analogy, the more sophisticated the key, the less is the probability of its imitation. This hypothesis was formulated as a common rule for both linear and conformational FISs. However, the uniqueness of linear FISs is concerned with the frequency characteristics of the contiguous site, which itself is responsible for biological function. This site, as mentioned above, should not be homologous with other sites that do not possess this function. On the other hand, conformational FISs should be unique with regard to the spatial arrangement of residues constituting it, but the unique geometry does not depend on the frequencies of these residues.

We supposed that taking into account the information concerning the spatial structure of the protein should allow us to apply frequency analysis to conformational FISs. The evident limitation of this approach is the necessity of having a well-resolved structure of the protein, but the developments of crystallographic and NMR methods allow us to look forward to fast acquisition of structural data. Moreover, knowledge of structure does not imply a knowledge of functional organization. Besides that, it seems interesting to examine the statistical characteristics of conformational FISs and to compare them with ones of linear FISs.

As a model to validate our hypothesis about the possibility of predicting conformational FISs, we used thermolysin, a secretory metalloproteinase with unusually high thermostability. The structure of thermolysin consists of two domains [6–8]. Sites of contacts between these domains (inter-domain interface) can be regarded as intramolecular sites of protein–protein interaction, possibly responsible for the thermostability of the

*Corresponding author. Fax: (7) (095) 135-1405.
E-mail: andy@imb.mb.free.net

SSDI 0014-5793(94)01257-1

Table 1
Functional residues of thermolysin

Function	Residues responsible for function
Enzyme active site	142, 146, 166, 143, 157, 170, 203, 226, 231
Substrate binding	130, 133, 139, 188, 189, 192, 202
Ca^{2+} binding	57, 59, 61, 138, 187, 174, 177, 185, 190, 183, 191, 182
Inter-domain interface	14, 79, 80, 82, 83, 84, 86, 87, 88, 90, 127, 128, 129, 130, 131, 132, 133, 135, 136; 172, 175, 176, 179, 180, 182, 191, 192, 193, 194, 195, 202, 264, 265, 268, 272

molecule. The key advantage of thermolysin is its well-defined 3D structure and presence of data on functional organization. Known functional sites of thermolysin are: enzyme active site, substrate-binding site, Ca^{2+}-binding sites and inter-domain interface (see Table 1). Data from the first of these three groups were taken from the literature [6–8]. The search for residues constituting the inter-domain interface was performed according to the following procedure: (i) the structure of thermolysin was divided into three parts: 1–136 (domain 1), 137–157 (inter-domain linker) and 158–316 (domain 2); (ii) an amino acid from domain 1 or 2 was considered to form an inter-domain interface if it was in contact with any amino acid from another domain.

It could be concluded (see Table 1) that the inter-domain interface is formed mostly by three contiguous sites. Thus, the inter-domain interface is the only linear FIS in the thermolysin structure while the active site, substrate- and calcium-binding sites are conformational.

2. Materials and methods

To calculate linear uniqueness U^L, we used frequencies of tripeptides in non-homologous proteins from the PIR database [10]. Each residue, A_k, was characterized by the frequency of the tripeptide $A_{k-1}A_kA_{k+1}$, where A_{k-1} and A_{k+1} are preceding and subsequent residues in the primary structure of thermolysin.

$$U^L_{A_k} = -fr(A_{k-1}A_kA_{k+1}) \quad (1)$$

The linear uniqueness profile was smoothed using a 5-residue window that was previously shown as optimal [5].

To characterize the frequency of occurrence of the spatial neighborhood we introduced a new parameter called 'spatial uniqueness'. Simply stated, the more unusual a residue's neighborhood, the more is its spatial uniqueness. Spatial uniqueness, U^S, was calculated using data on amino acid contact frequencies in protein structures [11]. Amino acid residues were considered as contacting if the distance between any two atoms of these residues did not exceed 4.5 Å. Each residue A_k was characterized by the sum of the frequencies of its contacts A_kA_x, where A_x is the amino acid contacting A_k in the structure of thermolysin.

$$U^S_{A_k} = -\sum_x fr(A_kA_x) \quad (2)$$

We have also calculated the averaged spatial uniqueness U^A that is normalized to the total number of contacts for this residue. Because of this normalization, average spatial uniqueness does not depend on shielding of the residue.

$$U^A_{A_k} = -(\sum_x fr(A_kA_x))/i \quad (3)$$

where i stands for total number of residues that are contacting with residue A_k in the protein (e.g. thermolysin) structure.

Results of the prediction were analyzed using χ^2 statistics as described in [12]. This type of analysis characterizes the probability of obtaining the given numbers of correct and wrong predictions by chance. According to [12], we interpreted the prediction as correct if the uniqueness value for a functional amino acid exceeded the average plus 0.7 × S.D., and/or the value for the non-functional amino acid was less than the average minus 0.7 × S.D. Here S.D. is the standard deviation from the average value for all residues. Analogously, the prediction was considered as wrong both if the uniqueness value for functional residue was less than the average minus S.D., and/or if the value for the non-functional residue was greater than the average plus 0.7 × S.D. The 2 × 2 contingency table included the number of correct and wrong predictions for functional and non-functional amino acids. The values of χ^2 coefficients, calculated for a 2 × 2 table, not exceeding 3.84 are not statistically significant at the level of 0.05. It should be noted that for every type of biological activity we separately considered non-functional amino acids as ones not involved in particular biological function. For example, if we are analyzing calcium-binding amino acids, all others are treated as non-functional, including active site, substrate-binding and interface residues.

3. Results and discussion

According to the χ^2 coefficients (see Table 2), linear uniqueness was significantly higher for calcium-binding residues ($P < 0.05$) and marginally for interface residues ($P < 0.1$), than for the rest of the protein. In contrast, active site and substrate-binding residues do not stand out by their linear uniqueness values. The situation is completely different for spatial uniqueness. There is the distinct increase in χ^2 coefficients for Ca^{2+}-binding residues and, especially, for active site residues, for which the difference between linear and spatial uniqueness is the largest. Very interesting are results obtained for the group of substrate-binding residues. They are *less unique* (have more trivial environments) than other amino acid residues in the thermolysin structure ($P < 0.05$). This could be prescribed to the broad specificity of thermolysin as an enzyme, dictating the necessity for the binding site to adapt to various substrate structures. The results for averaged spatial uniqueness are quite similar to ones for spatial uniqueness. The group of calcium-binding residues has the largest χ^2 coefficient, followed by the active site. Their values are increased even more compared to substrate-binding and interface residues.

According to the results of statistical testing, the best characteristic to search for active center and calcium-binding residues is averaged spatial uniqueness; for substrate-binding site, spatial uniqueness; and for inter-domain interface, linear uniqueness. Spatial uniqueness could be estimated as a significant criterion of whether the given residue forms a conformational FIS ($P < 0.05$). The values of spatial uniqueness for linear FIS are very low. By contrast, linear uniqueness could be used to predict linear intramolecular interaction sites (such as domain interfaces) but is of no use for conformational FIS. We suppose that the conformational active center of the enzyme became

Table 2
Statistical characteristics (χ^2 coefficients) of functional residues' predictions

Group of functional residues	Linear uniqueness	Spatial uniqueness	Average spatial uniqueness
All functional residues	4.19	4.57	10.44
Ca^{2+}-binding	3.91	7.55	10.08
Active site	−0.97	5.8	7.45
Substrate binding	−0.97	−3.83	−1.57
Inter-domain interface	3.19	−0.2	0.49

Predictions significant at the <0.05 level are underlined.

unique only after correct folding of the protein structure. One could speculate that parts of the conformational active site, in general, should not be sequentially unique so as to prevent possible use as a protein–protein interaction site. The only exception we found was the structure of Ca^{2+}-binding sites, that are both sequentially and spatially unique. Even in this case statistical parameters of prediction were better for spatial uniqueness compared with linear uniqueness. Thus we showed that distinct FISs are characterized by different frequency patterns both in primary and in tertiary structures. We concluded that the hypothesis concerning the biological meaning of uniqueness provides a rational basis for prediction of a wide range of functional sites.

References

[1] Eroshkin, A.M., Zhilkin, P.A., Popkov, I.K. and Kulichkov, V.A. (1987) Biofizika (Russ.; Engl. res.) 32, 972–981.
[2] Saroff, H.A. and Pretorius, H.T (1983) Bull. Math. Biol. 45, 117–138.
[3] Claverie, J.-M. and Bougueleret, L. (1986) Nucleic Acids Res. 14, 179–196.
[4] Gabrielian, A.E., Ivanov, V.S. and Kozhich, A.T. (1990) Comp. Appl. Biosci. (CABIOS) 6, 1–2.
[5] Gabrielian, A.E., Nekrasov, A.N. and Kirpichnikov, M.P. (1991) Biomed. Sci. 2, 481–484.
[6] Matthews, B., Jansonius, J., Collman, P., Schoenborn, B. and Dupourque, D. (1972) Nature 238, 37–40.
[7] Matthews, B., Collman, P., Jansonius, J., Titani, K., Walsch, K. and Neurath, H. (1974) Nature 238, 41–43.
[8] Matthews, B., Weaver, L. and Kester, W. (1974) J. Biol. Chem. 249, 8030–8044.
[9] Strongin, A., Kostrov, S. and Kaydalova, N. (1991) Protein Seq. Data Anal. 4, 355–361.
[10] McCaldon, P. and Argos, P. (1988) Proteins: Struct., Funct. Genet. 4, 99–122.
[11] Manavalan, P. and Ponnuswamy, P.K. (1977) Arch. Biochem. Biophys. 184, 476–487.
[12] Pellequer, J.L., Westhof, E. and Van Regenmortel, M.H.V. (1991) Methods Enzymol. 203, 176–201.