Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 25, Issue Number 5, (2008)

Hydrolases: The Correlation Between Informational Structure and the Catalytic Sites Organization

Alexei N. Nekrasov*
Alexei A. Zinchenko

Shemyakin-Ovchinnikov
Institute of Bioorganic Chemistry
Russian Academy of Sciences
117997 Moscow
ul. Miklukho-Maklaya 16/10
Russia

http://www.jbsdonline.com

Abstract

The novel method allowing identification of protein structure elements responsible for catalytic activity manifestation is proposed. Structural organization of various hydrolases was studied using the ANIS (**AN**alysis of **I**nformational **S**tructure) method. ANIS allows to reveal a hierarchy of the **EL**ements of **I**nformation **S**tructure (ELIS) using protein amino acid sequence. The ELIS corresponds to the variable length sites with an increased density of structural information. The amino acid residues forming the enzyme catalytic site were shown to belong to the different top-ranking ELIS located in the contact area of the corresponding spatial structure clusters. In the protein spatial structure catalytic sites are located in the area of contact between fragments of polypeptide chain (structural blocs) allocation to the different top-ranking ELIS. According to our results we concluded that structural blocks corresponding to top-ranking ELIS are crucial for protein functioning. Such regions are structurally independent, and their determinate mobility relative to each other is vital for an efficient enzymatic reaction to occur.

Key words: Information structure of proteins; Hydrolases; and Enzyme catalytic site.

Introduction

Virtually all of the chemical transformations that occur in biological systems are catalyzed by enzymes. To this date, hundreds of thousands of native proteins are known to possess an enzymatic activity. Numerous studies of enzyme spatial structures using various theoretical and instrumental approaches still fail to elucidate the exact mechanism of the enzymatic catalysis. This situation calls for the development of novel approaches to the protein research. Herein we propose the ANIS method (**AN**alysis of **I**nformation **S**tructure) which allows to reveal so-called IDIC-sites (**I**ncreased **D**egree of **I**nformational **C**oordination between amino acid residues) within protein sequences. IDIC-sites were previously shown to be concise descriptors of the most characteristic structural features of native proteins (1, 2). The ANIS method also displays the hierarchical relationship between IDIC-sites of various length, which together form the informational structure (IS) of a protein (1). An increased degree of coordination between residues within IDIC-sites indicates a high density of structural information, which determines the certain spatial organization. In this work we used the IS concept to study the distribution of amino acid residues in catalytic sites of several hydrolases between the top-ranking ELIS of corresponding proteins.

The ANIS method is based on the earlier studies of positional informational entropy of non-homologous protein sequences (2), which revealed the following properties:

I. The dependence of positional informational entropy on the inter-residual distance along the primary sequence is S-shaped, which allows

*Email: alexei_nekrasov@mail.ru

to postulate the existence of extended sites with an increased degree of coordination between residues. Such regions may span up to several dozens of residues.

II. The maximum degree of coordination was observed at inter-residual distances of < 6 positions. Accordingly, the interval of five positions was designated as the **H**ighest-**L**evel **C**oordination **D**istance (HLCD), which allowed formalization of protein sequences as series of short, overlapping (by one residue) peptide fragments named information units (IU). This approach provides an ability to study the organization of structural information encoded by the primary sequences.

The existence of IDIC-sites of varying length and the analysis of their distribution along the peptide chain allows to define "the **I**nformational **S**tructure of protein sequence" (IS) as a hierarchy of IDIC-sites (1). IS can be represented in a graphical form (IDIC-diagram), which can be obtained by connecting the centers of closest neighboring IDIC-sites of different ranks.

IDIC-sites of the minimal size (equal to HLCD) correspond to the lowest rank of IS. According to the IS hierarchy, lower-ranking elements are included into higher-ranking elements that correspond to the extensive IDIC-sites of the protein sequence. Such prolonged IDIC-sites represent an Elements of Informational Structure (ELIS) of top ranks in the protein IS. It should be noted that such an hierarchically organized IS can be elucidated for any protein sequence.

When IS of several proteins were compared with their actual spatial structures (1), the boundaries of structural domains in some cases matched with borders of top-ranking ELIS. The present work continues our studies of structural and functional role of protein IS. We used several hydrolases with well-known active sites as objects to study the distribution of amino acid residues in catalytic sites between the top-ranking ELIS of corresponding enzymes.

Methods

Informational Structure Analysis

IDIC-diagrams can be composed using following algorithm:
- the amino acid sequence of a given protein is encoded as a set of informational units (IU);
- the population profile of the target protein structure by IU is determined;
- IDIC-sites is localized within the protein sequence.
- the graphic representation of IS is constructed.

Encoding Protein Sequences as IU Sets

Let $\Re$ be the set of all known amino acid residues forming primary structure of native protein sequences, *i.e.*, $B_i = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. Then the protein sequence written in a traditional form is a function $B : \{1,\ldots, N\} \rightarrow \Re$, where N is the number of residues in the sequence. When encoded as a set of informational units (IU) of $\varepsilon = 2\delta + 1$ size, each amino acid residue i is represented not by the elements of $\Re$, but by the set $U_i = B_{i-\delta},\ldots,B_{i+\delta})$, so the protein sequence is determined by the function $U : \{1,\ldots,N\} \rightarrow \Re^{\varepsilon}$. All the protein sequences included in the PIR database rel. 8.0 (3) were formalized in this manner, and the occurrence frequencies P_{U_k} of information units of type k (U_k) were computed for each sequence. The IU size was selected not no exceed HLCD value ($\varepsilon \leq R_{HLCD}$) then the positional informational entropy is minimal, *i.e.*, $\varepsilon \leq 5$ (2). The occurrence frequencies of IUs were used for the study of information structure of protein sequences.

The next step in the study of protein information structure is the composition of the profile $F = \{F_j\}$ of the protein sequence populated by information units U_k:

$$F(j) = \begin{cases} \sum\limits_{i=j\pm\varepsilon/2} P_{U_i} & \text{if} \quad U_j \in \Im \\ 0 & \text{if} \quad U_j \notin \Im \end{cases} \qquad [1]$$

where $j = 1 + \delta,\ldots N - \delta$, $\varepsilon = 2\delta + 1$ (IU size). $\Im$ is the multitude of all possible information units that occur in the database obtained by the analysis of sets of non-homologous protein sequences. Eighty percent of the information units from $\Im$ is equal in terms of their physical and chemical characteristics.

Localization of IDIC-sites

In order to identify the location of IDIC-sites we introduce a Gaussian function $f(j')$ for each $j = 1,2,\ldots N$

$$f(j') = De^{-\frac{(j'-j)^2}{\rho^2}} \qquad [2]$$

which is determined for all values $j = 1,2,\ldots N$ and which complies with the condition

$$f(j') - F(j) \geq 0. \qquad [3]$$

The later condition is the limitation of $f(j')$ function D parameter value.

For each $j = 1,\ldots N$ the $D = D(j)$ parameter is chosen to be the maximal of D complying with the condition [3].

During the information structure studies the ρ values matching interval $\delta < \rho < N - \delta$ were used, where N is the length of target sequence. At the same time,

$$\omega(j,\rho) = \sum_{j'} D(j)e^{-\frac{(j'-j)^2}{\rho^2}} \qquad [4]$$

can be calculated for each $f_j(j')$ function.

Graphic Representation of a Protein IS (The IDIC-diagram)

The IS of a protein can be graphically represented as a surface Γ_ω, which is the graph of $\omega = \omega(j,\rho)$ function determined for all possible ρ and j values. A simpler and more convenient IDIC-diagram can be constructed by using a restricted set of Γ_ω surface values, which are local maximums complying with the condition:

$$\omega(j-1,\rho) \leq w(j,\rho) \geq \omega(j+1,\rho). \qquad [5]$$

As the result of such substitution the structurally complicated Γ_ω surface is reduced to a limited set of points $K = \{\omega(j,\rho) : \omega(j-1,\rho) \leq \omega(j,\rho) \geq \omega(j+1,\rho)\}$ $1 + \delta \leq j \leq N - \delta$, $\delta \leq \rho \leq N - \delta$. By connecting the closest points that correspond to positions of IDIC-sites of different length ρ, hierarchical graphs-ELIS are produced. Hereinafter, in the discussion of protein information structure we will operate with the different rank ELIS. The ELIS at the given node point is characterized by the rank that is equal to the number of node points along the most distant path from lowest-level element to target node point. The ELIS located at the lowest level of hierarchy are assigned a rank equal to 1.

Objects of Study

The following hydrolases with well-defined catalytic sites were studied: pepsin

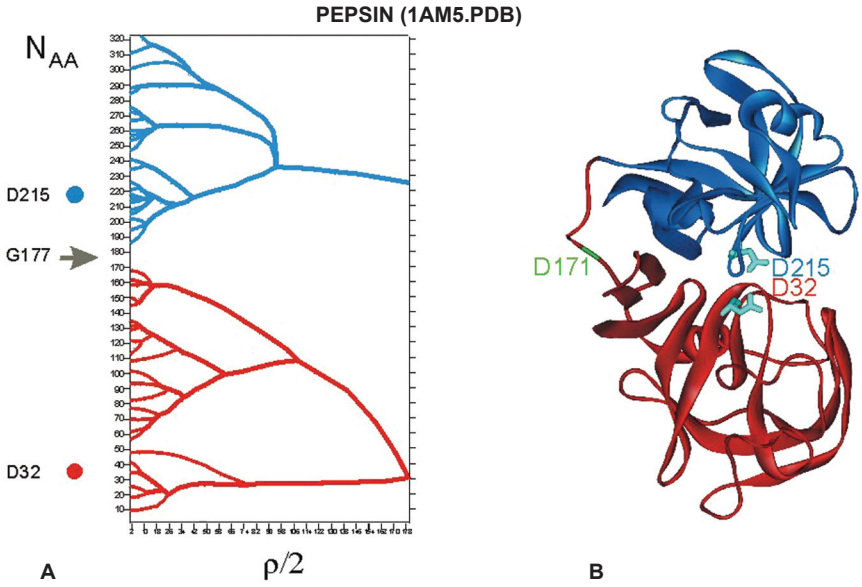

Figure 1: Pepsin. (**A**) The IDIC-diagram of pepsin IS. Here and further on Figures 2A-6A, the half-width values (ρ/2) of the decomposition function (x-values) are plotted against the serial numbers (N_{AA}) of amino acid residues (y-values) in the protein sequence. The primary structure of pepsin is represented by two fragments (1-177 and 178-326) corresponding to two top-ranking ELIS. ELIS borders are marked by an arrow. The active site residues D32 and D215 are highlighted. (**B**) Top-ranking ELIS in the structure of pepsin are marked with unique colors. The active site residues D32 and D215 are cyan-colored. The residue D171, which, according to the 3Dee database, lies between the structural domains of the enzyme, is colored in green.

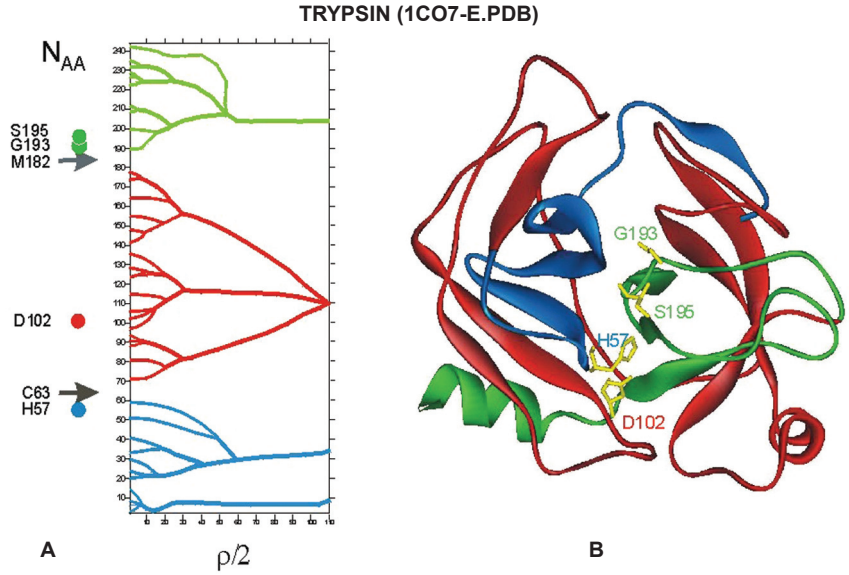

Figure 2: Trypsin. (**A**) The IDIC-diagram of trypsin IS. The three top-ranking ELIS are colored blue, red, and green. They correspond to sequence fragments R1-C63, Y64-M182, and V183-N245. The catalytic triad residues H57, D102, S195, and G193 and ELIS borders are marked. (**B**) The spatial structure of trypsin (8) with top-ranking ELIS marked with unique colors. The H57, D102, S195, and G193 residues are highlighted.

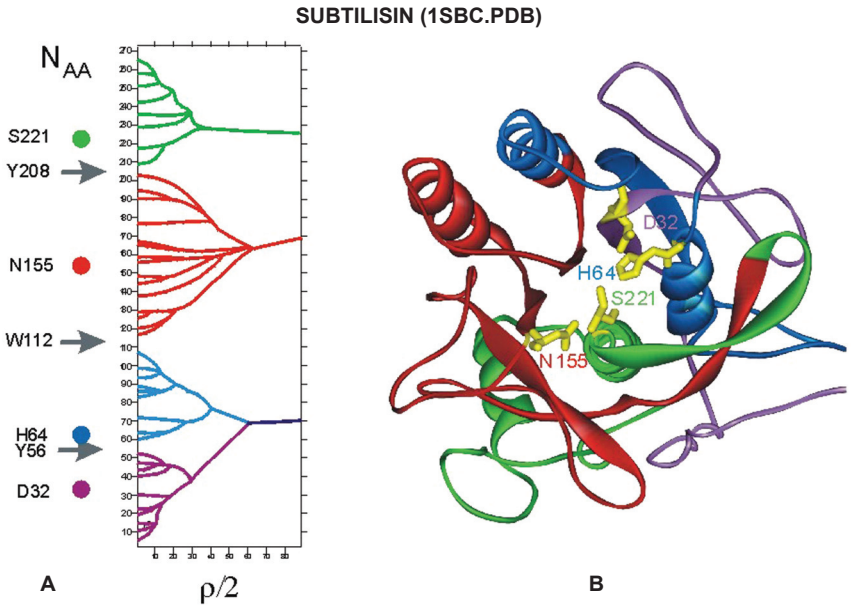

Figure 3: Subtilisin. (**A**) The IDIC-diagram of subtilisin IS. The three top-ranking ELIS are colored violet/blue, red, and green. They correspond to sequence fragments A1-W112, A113-Y208, and P209-Q274. At a lower level of IS hierarchy (dotted line), an additional division of the N-terminal ELIS occurs at Y56. The catalytic site residues D32, H64, S221, and N155 are marked. (**B**) The spatial structure of subtilisin (10) with top-ranking ELIS marked with unique colors. The D32, H64, S221, and N155 residues are highlighted.

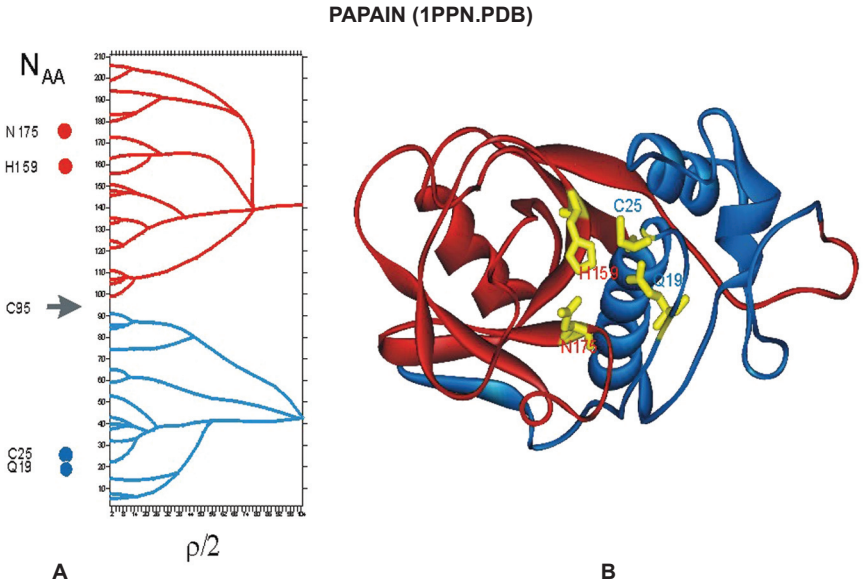

Figure 4: Papain. (**A**) The IDIC-diagram of papain IS. The two top-ranking ELIS (colored red and blue) correspond to the sequence fragments I1-C95 and R96-N217. The border of ELIS is marked by an arrow. The catalytic dyad residues C25 and H159, as well as the essential residues Q19 and N175, are marked. (**B**) The spatial structure of papain (12) with top-ranking ELIS marked with unique colors. The C25, H159, and Q19, N175 residues are highlighted.

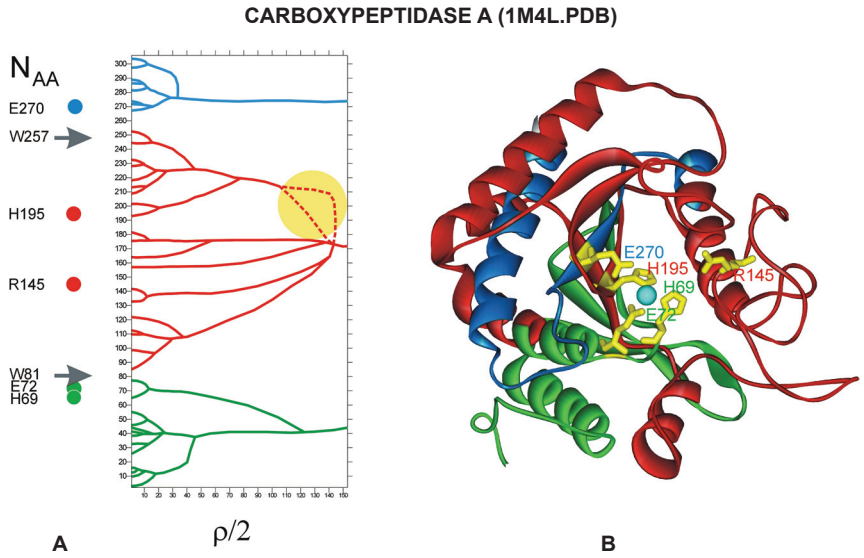

Figure 5: Carboxypeptidase A. (**A**) The three highest-ranking ELIS corresponding to sequence fragments A1-W81, F82-W257, and S258-N307 are colored green, red, and blue. The borders are marked by arrows. The area of IDIC-branch splitting is colored yellow. (**B**) The spatial structure of carboxypeptidase A (14) with top-ranking ELIS marked with unique colors. The catalytic site residues E270 and H69, E72, H196 are highlighted.

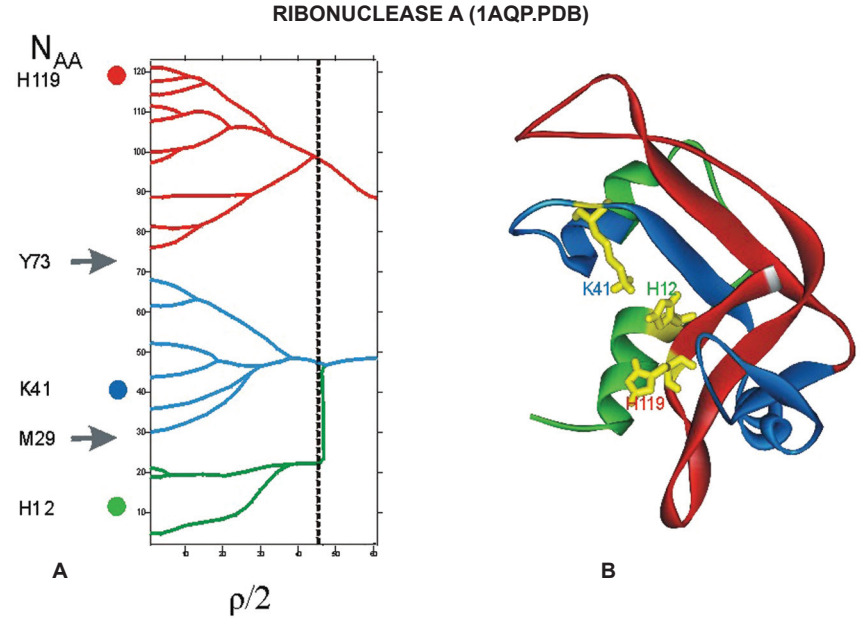

Figure 6: RNase A. (**A**) The IDIC-diagram of RNase A IS. The two top-ranking ELIS (colored green/blue and red) correspond to sequence fragments K1-Y73 and Q74-V124. At a lower level of IS hierarchy (dotted line), an additional division of the N-terminal ELIS occurs, resulting in two ELIS 1-29 and 30-73 of the lower rank. The ELIS boundaries are marked by arrows. (**B**) The spatial structure of RNase A (15) with top-ranking ELIS marked with unique colors. The active site residues H12, H119 and K41 are highlighted.

(EC 3.4.23.1), a peptide hydrolase from MEROPS family A1 (aspartyl peptidases), clan AA (4-7); trypsin (EC 3.4.21.4), a peptide hydrolase from subfamily S1A (serine peptidases), clan PA(S) (4, 7-9); subtilisin (EC 3.4.21.62), a peptide hydrolase from subfamily S8A (serine peptidases), clan SB (4, 7, 9, 10); papain (EC 3.4.22.2), a peptide hydrolase from family C1 (cysteine peptidases), clan CA (4, 7, 11, 12); carboxypeptidase A (EC 3.4.17.1) a peptide hydrolase from family M14 (metallopeptidases), clan MC (4, 7, 13, 14); and a pancreatic endonuclease-ribonuclease A (EC 3.1.27.5) (7, 15, 16).

Results and Discussion

The characteristic feature of the domain organization of enzymes is that the amino acid residues of the catalytic site are located in the area of inter-domain contact. For example, D32 and D215 residues of the catalytic site of pepsin (Fig. 1A) belong to different domains. As protein domains show an increased mobility relative to each other, the distribution of catalytic site residues between different domains is directly related to the mechanism of enzyme action (17-19).

Whereas the domain organization is not necessarily present in all enzymes, the sequence-derived IS is a universal attribute of all native proteins. We postulated that top-ranking ELIS correspond to independent elements that posses stable spatial structure and determinate mobility typical for actual structural domains. In this case, one should expect the residues forming the catalytic sites to be distributed over different elements of the spatial structure that correspond to top-ranking ELIS. Such a distribution would provide a mobility of structural elements that is required for an efficient enzymatic reaction to occur.

Let us discuss in detail the results obtained with our objects of study. The domain organization of proteins with known spatial structure was taken from the Database of Protein Domain Definitions – 3Dee (20). In the pepsin molecule there are two domains formed by the polypeptide fragments 1-170 and 171-326. The IS of this protein (*cf*. the IDIC-diagram on Fig. 1A) consists of two top-ranking ELIS separated by the G177 residue. Thus, there is a good agreement between the ANIS results and the 3Dee data. The spatial structure can be conveniently compared to IS using color-coding. Figure 1B shows the spatial structure of pepsin (8) with the polypeptide chain colored in correspondence to the top-ranking ELIS.

The catalytic site of pepsin is formed by the side chains of D32 and D215 aspartate residues. It can be seen from the IDIC-diagram (Fig. 1A) that these residues are located in different top-ranking ELIS.

Thus, analysis of pepsin IS demonstrates there is a good agreement between the domain organization and the localization of top-ranking ELIS, which may lead to conclusion that top-ranking ELIS can in fact be treated as actual domains The localization of the catalytic site residues in different ELIS seems quite logical in this case.

The catalytic sites of trypsin and subtilisin are represented by the catalytic triad typical for serine proteases, although these proteins are not homologues. They belong to different structural classes (trypsin is a β-protein, and subtilisin is an α/β-protein) and differ in the domain organization (trypsin consists of two structural domains, subtilisin possesses a sole structural domain).

The IS of trypsin (Fig. 2A) comprises three top-ranking ELIS, which do not match the domain organization of protein. However, each of the residues forming the catalytic triad (H57, D102, and S195) (7) is located in its own different top-ranking ELIS. The G193 residue, which participates in the formation of the “oxyanion hole” and stabilizes the negative charge in the tetrahedral transition complex, is

sometimes also regarded as part of the catalytic site of trypsin. This residue is located in the same ELIS with the S195 residue, which is directly involved in the formation of the acyl-enzyme intermediate. Thus, the IS of trypsin reflects the structure of its catalytic site. The spatial structure of trypsin, top-ranking ELIS and the catalytic site residues are shown in Figure 2B. It can be from Figure 2B that the catalytic site is located in the area of contact between all three ELIS.

The IS of the single-domain protein, subtilisin Carlsberg (Fig. 3A) comprises three top-ranking ELIS separated by the W112 and Y208 residues. The catalytic site of subtilisin (7) is formed by the serine protease triad of H64, D32, and S221. Similar to trypsin, the N155 residue, which stabilizes the tetrahedral transition state of a substrate, may be regarded as part of the catalytic site. Of these four residues, D32 and H64 belong to the top-ranking ELIS (A1-W112), whereas N155 and S221 belong to top-ranking ELIS (W112-T207) and (Y208-Q274), respectively. At the same time a lower level of the IS hierarchy, the N-terminal ELIS (A1-W112) becomes divided into two ELIS of lower rank separated by the Y56 residue, so that D32 and H64 become distributed over different lower-ranking ELIS. Thus, each of the active site residue of subtilisin falls into a different ELIS when two high levels of the IS hierarchy are examined instead of one. The spatial structure of subtilisin marked with colors of ELIS of both ranks and the catalytic site residues are shown in Figure 3B. It can be seen that the catalytic triad is again situated in the area of ELIS contact.

The active site of papain, a cysteine peptidase, is formed by the C25 and H159 residues. This classic catalytic dyad is usually augmented by two important residues, N175 and Q19. The N175 residue is believed to form a hydrogen bond with the side-chain imidazole ring of H159 and fixes it, and the Q19 residue stabilizes the tetrahedral intermediate (11). It can be seen from the IDIC-diagram shown in Figure 4A that the top-ranking ELIS divide the papain sequence into two fragments: I1-C95 and R96-N217. Each fragment includes two of the catalytic site residues. The residues of the catalytic dyad-C25 and H159, belong to different top-ranking ELIS. It is important to note that, similar to trypsin, the intermediate-stabilizing Q19 residue and the acyl-enzyme-forming C25 residue are located in the same ELIS. The spatial structure, top-ranking ELIS, and the catalytic site residues of papain are shown in Figure 4B.

The catalytic sites of metallopeptidases, including carboxypeptidase A, contain zinc ions. The active site of carboxypeptidase A comprises H69, E72, and H196 residues (which form three out of four coordination bonds with the metal), a catalytic glutamyl residue E270, and R145, which binds to the α-carboxyl group of the substrate (13). In this protein structure (Fig. 5A) there are three top-ranking ELIS corresponding to sequence fragments of unequal length, A1-W81, F82-W257, and S258-N307. The (258-307) fragment contains the E270 residue, and each of the other two fragments contains two of the remaining residues of the active site (H69, E72 and R145, H196). Note that each of the residues coordinated to the zinc ion is located in a different ELIS. Such distribution pattern of zinc-associated residues may reflect the peculiarities mechanism of cofactor-assisted enzymatic catalysis. Figure 5A demonstrates a peculiarity of central ELIS that manifests as a split of IDIC-branch upon increasing ρ value. This feature can also be connected with a presence of cofactor in the enzyme structure. Figure 5B shows the spatial structure, the top-ranking ELIS, and the catalytic residues of carboxypeptidase A.

The above-described distribution pattern of the catalytic site residues over top-ranking ELIS was observed not only for the large proteins, but also for smaller hydrolases. For example, the IDIC-diagram for RNase A (Fig. 6A) shows two top-ranking ELIS, K1-Y73, and Q74-V124. The catalytic site of RNase A is formed by H12, K41, and H119 (18). Two essential catalytic residues, H12 and H119, are located in different ELIS. At the next level of the IS hierarchy, ELIS K1-Y73 becomes divided in two fragments, K1-M29 and M30-Y73 corresponding to ELIS of one lower ranks.

At this lower level, each of the catalytic residues (H12, K41, and H119) is associated with its own IS element, and is located in the area of ELIS contacts (Fig. 6B).

Our data shows that in hydrolases the catalytic site residues are localized in different top-ranking ELIS in the area of contact between the corresponding parts of protein spatial structures. The further investigations are required to learn whether this observation is common for other classes of enzymes as well.

The obtained results allow us to postulate that the efficiency of an enzymatic reaction is determined not only by the spatial proximity and productive orientation of the substrate and catalytic residues of the active site (20-21), but also by the non-stochastic mobility of the structural elements, which determines their spatial configuration of catalytic residues at various stages of the catalytic act. We assume that the ability towards the coordinated determined rearrangements of the protein structure is a fundamental property of enzymes, which is enabled by the complexity of their structural organization. The non-stochastic mobility is determined by the elements of spatial structure corresponding to top-ranking ELIS. The high degree of coordination between the residues located in top-ranking ELIS is reflected in the structural stability of the corresponding fragments of the spatial structure. As a result, the conformational transitions occurring within the protein globule during enzymatic catalysis are mainly caused by this non-stochastic movement of structural elements that correspond to top-ranking ELIS. This movement pre-determines the mutual orientation of the enzyme active site residues and the substrate at various stages of the catalytic act and thus ensures high efficiency of the process.

Thus, we believe that an application of the ANIS method to enzymes led to the discovery of novel, stable structural elements, which play an important role in enzyme functioning, and allowed to propose several hypotheses concerning the course of enzymatic catalytic reactions.

Acknowledgments

We thank Dr. Alexey Perkovsky for helpful discussions and critical reading of the manuscript.

References and Footnotes

1. Nekrasov A. N. *J Biomol Struct Dyn 21*, 615-623 (2004).
2. Nekrasov A. N. *J Biomol Struct Dyn 20*, 87-92 (2002).
3. Wu C. H., Yeh L. S., Huang H., Arminski L., Castro-Alvear J., Chen Y., Hu Z., Kourtesis P., Ledley R. S., Suzek B. E., Vinayaka C. R., Zhang J., and Barker W. C. *Nucleic Acids Res 1*, 345-347 (2003).
4. MEROPS the peptidase database Release 7.50 http://merops.sanger.ac.uk/
5. James M. N. G. Catalytic pathway of aspartic peptidases, in *Handbook of Proteolytic Enzimes, 2 Ed.,* pp. 12-19. Eds., Barrett A. J., Rawlings N. D., and Woessner J. F. Elsevier, London (2004).
6. Sielecki, A. R., Fedorov, A. A., Boodhoo, A., Andreeva, N. S., and James, M. N. *J Mol Biol 214*, 143-170 (1990).
7. The Comprehensive Enzyme Information System - BRENDA http://www.brenda.uni-koeln.de/
8. Transue, T. R., Gabel, S. A., and London, R. E. *Bioconjug Chem 17*, 300-308 (2006).
9. Rawlings, N. D. and Barret, A. J. Introduction: Serine peptidase and their clans, in *Handbook of Proteolytic Enzimes, 2 Ed.*, pp. 1417-1439. Eds., Barrett, A. J., Rawlings, N. D., and Woessner, J. F. Elsevier, London (2004).
10. Neidhart, D. J. and Petsko, G. A. *Protein Eng 2*, 271-276 (1988).
11. Polgar, L. Catalytic Mechanisms of Cysteine Peptidases, in *Handbook of Proteolytic Enzimes, 2 Ed.*, pp. 1072-1079. Eds., Barrett, A. J., Rawlings, N. D., and Woessner, J.F. Elsevier, London (2004).
12. Yamamoto, D., Matsumoto, K., Ohishi, H., Ishida, T., Inoue, M., Kitamura, K., and Mizuno, H. *J Biol Chem 266*, 14771-14777 (1991).
13. Auld, D. S. Catalytic Mechanisms of Metallopeptidases, in *Handbook of Proteolytic Enzimes, 2 Ed.*, pp. 268-289. Eds., Barrett, A. J., Rawlings, N. D., and Woessner, J.F. Elsevier, London (2004).
14. Kilshtain-Vardi, A., Glick, M., Greenblatt, H.M., Goldblum, A., and Shoham, G. *Acta Crystallogr, Sect D 59*, 323-333 (2003).

15. Balakrishnan, R., Ramasubbu, N., Varughese, K. I., and Parthasararhy, R. *Proc Nat Acad Sci. USA 94*, 9620-9625 (1997).
16. Blackburn, P. and Moore, S. Pancreatic Ribonuclease, in *The Enzymes* 15, pp. 317-433. Ed., P. D. Boyer. Acad. press, N.Y. (1982).
17. Fersht, A. *Structure and Mechanism in Protein Science*. W. H. Freeman and Company, New York (1999).
18. Finkelstein, A. V. and Ptitsyn, O. B. *Protein Physics. A Course of Lectures*. Academic Press (2002).
19. Antonov, V. K. *Chemistry of Proteolysis*. Springer-Verlag, Berlin – Heidelberg (1993).
20. 3Dee - Database of Protein Domain Definitions http://www.compbio.dundee.ac.uk/3Dee/search/domains_server.html
21. Jencks, W. P. *Annu Rev Biochem 66*, 1-18 (1997).
22. Blow, D. *Structure 8*, R77-R81 (2000).

Date Received: December 13, 2007

Communicated by the Editor Valery Ivanov