Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 21, Issue Number 5, (2004)

Analysis of the Information Structure of Protein Sequences: A New Method for Analyzing the Domain Organization of Proteins

http://www.jbsdonline.com

Alexei N. Nekrasov

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

Abstract

The amino acid sequences of γ-crystallin, Haloalkane Dehalogenase, Phthalate Dioxygenase, Porphobilinogen Deaminase and Myosin Regulatory Domain c-chain were analyzed for their information content. Sites of increased degree of information coordination between residues (IDIC-sites) were identified, and their organization was studied by means of analyzing the information structure of the protein sequences. Relationships between the structural units forming the spatial and informational structure of proteins were demonstrated. Associations of information-coordinated structural elements (IDIC-associations) were mapped onto compact structural domains found in the spatial structures of globular proteins. The proposed method of analyzing the information structure of protein sequences may find applications in the biotechnology and structural chemistry of proteins.

Introduction

Recent developments in genetic engineering and molecular biology have formed a methodological basis for the research aimed at designing functionally modified protein structures. Functional design usually results in some activities characteristic of native proteins being either expressed or suppressed. The problem of subdividing protein sequences into functionally active areas is most directly approached by identifying those fragments of protein sequences which form separate compact units of the spatial structure (structural domains), as each of these domains can be related to one or another kind of protein activity. At present, the domain organization is studied only in proteins with known spatial structures (1-7). Identification of spatial domain organization by means of analyzing amino acid sequences may be of great importance for structural studies concerning the spatial structures of large protein complexes. Knowing the domain organization from sequences would reduce the problem to identifying the spatial structures of smaller, structurally independent components.

The only theoretical approach currently applied to studying the protein domain organization based on protein sequences is the multi-sequence alignment method (8-11). Unfortunately, the efficiency of this method critically depends on the availability of sequences that are homological to the target protein, which in practice may considerably limit the utility of the multi-alignment method. Developing a method for analyzing protein sequence that would be free from this constraint is therefore an actual problem. This article describes such a method and its application to a number of proteins.

Methods and Archive Data

The theoretical basis underlying the proposed method of analyzing the information structure of protein sequences has been presented in (12) together with data on integral informational properties of unhomological protein sets (UPS). One of the impor-

Email: alexei_nekrasov@mail.ru

tant findings reported in (12) was the detection of an S-shaped component of the positional sensitivity, which is a function of the distance between amino acid residues. This long-range (>30) positional sensitivity, or "mutual sensing", of amino acid residues implies that in native protein sequences there are found sites with increased degree of information coordination between amino acid residues (IDIC-sites). Since the positional sensitivity curve is S-shaped, the phenomenon can be rationalized using Gaussian functions. Another important result from (12) is that the inter-residual positional sensitivity is nearly constant and close to the maximum at distances up to five serial residues. This observation allows short sections of polypeptide chains to be regarded as informationally homogeneous "coding units". Accordingly, it seems more appropriate that detection of IDIC-sites should be based on a novel representation of protein sequences based on fixed-length polypeptide fragments ("information units", or IU) rather than on amino acid residues ("chemical units").

In the proposed representation of protein sequences, each residue together with its immediate serial neighbors are interpreted as "information units", which combine the information content and structural properties of the residues. Note that in this representation, each residue (except the N- and C-terminal residues and their close neighbors) is a IU center. That the positional sensitivity curve is S-shaped (12) indicates co-operative nature of the underlying phenomenon. It is expected therefore that encoding protein sequences in terms of IU should improve the efficiency of IDIC-site detection.

The proposed analysis of the information structure of protein sequences consists of the following steps:

I. Creating a database of IUs found in protein sequences that form a UPS and calculating the frequencies of IU occurrence in this UPS. (Sets of amino acid sequences from (12) were used as UPS).
II. Recoding an amino acid sequence in terms of IU.
III. Profiling the "information content" of the sequence using the frequencies of IU occurrence in the UPS obtained at Step I. The "information profile" is created by summing up the occurrence frequencies of IUs selected using a similarity criterion, which shows the degree of similarity between the IUs of the particular protein sequence and of the whole UPS. Thus, the information content profile indicates the frequency of different IUs in the primary structures of unhomological proteins.
IV. Detecting the centers of IDIC-sites by means of decomposing the information content profile into Gaussian curves and locating the maxima of the decomposition (Gaussian) functions that have the greatest values for the protein sequence in question. The hierarchy of IDIC-sites can be studied by varying the half-width of the Gaussian curves.

The information structure of a protein sequence can be conveniently analyzed using visual representations (IDIC-diagrams) of the decomposition functions that have the greatest values for this protein sequence. Depending on the half-width values of the decomposition functions, their maximums as seen on IDIC-diagrams may disappear, merge, shift along the sequence, and form hierachic structures.

The proposed approach is the first known method for detecting and analyzing the information structure of protein sequences.

Note that the obtained results are stable and insensitive to variations in the IU size (3 to 5 residues) and the IU similarity criterion (80% to 100%) applied during the construction of information content profiles. The sensitivity of the method towards the size and composition of the UPS used for generating the IU database is also insignificant. Variation of any of the above parameters resulted in shifting the centers of the IDIC-sites by ±1 position along the sequence.

Thus, the proposed method is suitable for analyzing the information structure of protein sequences. All the data (see below) for particular proteins were obtained using the following parameters: the IU length of 5 residues; the BASE 2 set from (12) was used as the UPS at Step III; the similarity criterion value of 80%. The computer program for analyzing the information structure of protein sequences was written using C++.

Let us define various information-structural elements of protein sequences. The center of an IDIC-site is the amino acid residue at which the decomposition function having the greatest value is centered when its half-width equals half of the IDIC-site length (this length itself may vary). *IDIC-trees* are independent hierarchies of IDIC-sites (information-structural elements). An *IDIC-branch* is any structural fragment of an IDIC-tree. An *independent IDIC-branch* is any isolated information-structural element which does not merge with other elements in the whole range of variation of the decomposition function half-width. IDIC-trees and independent IDIC-branches may form *IDIC-associations*, or groups of information-structural elements separated from each other by segments of the protein sequence. There are two criteria defining IDIC-associations. The sequence segments separating two IDIC-associations must consist of amino acid residues for which the decomposition functions have significantly lower values (Criterion 1). Information-structural elements are considered as IDIC-association members if the centers of the corresponding IDIC-sites are shifting and converging towards a certain position along the sequence upon increasing the decomposition function half-width (Criterion 2).

An IDIC-branch is assigned a label indicating the amino acid residue on which the decomposition function is centered and the hierarchy rank, which is initially set to 1. Thus, ^{1}E46 denotes a first-order IDIC-branch centered at GLU46 (Fig. 1), ^{1}P27 denotes a first-order IDIC-branch with the decomposition function centered at PRO27 (Fig. 1), etc. If an IDIC-branch undergoes merging, the central residue indicated by its label changes, and the rank is incremented by 1. A similar system applies to IDIC-trees. For example, ^{3}P27 denotes a third-order IDIC-tree, for which the decomposition function is centered at PRO27. This IDIC-tree is a result of two IDIC-branch fusions (Fig. 1). Upon increasing the half-width of the decomposition function, the centers of the IDIC-sites may shift along the sequence, but the labels for IDIC-branches and IDIC-trees are preserved until the next fusion occurs. An IDIC-association is denoted by a bracketed list of its constituent elements (IDIC-trees and independent IDIC-branches), e.g., {^{1}G10, ^{3}P27, ^{1}E46, ^{2}G59} (Fig. 1).

Now applications of the proposed method to proteins with known spatial structures will be demonstrated.

Results and Discussion

As a model protein of the β-sheet structural type, consider γ-crystallin (1GCR.PDB) (13). Figure 1 shows an IDIC-diagram of γ-crystallin obtained by the proposed method. Dotted arrows indicate the direction of shifting of the decomposition functions upon increasing their half-width. Using Criterion 2 (see above), the information-structural elements ^{1}G10, ^{3}P27, ^{1}E46, ^{2}R59 and ^{2}C109, ^{5}W131, ^{1}S166 can be grouped in two IDIC-associations. Using Criterion 1 gives similar results. Thus,

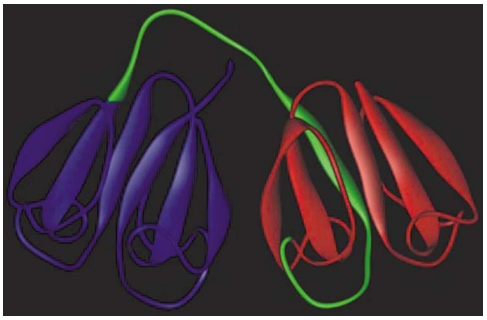

Figure 2: The spatial structure of γ-crystallin (1GCR.PDB). Structural elements that correspond to the IDIC-associations {^{1}G10, ^{3}P27, ^{1}E46, ^{2}G59} and {^{2}C109, ^{4}W131, ^{1}S166} are marked red and blue, respectively. The polypeptide stretch corresponding to the IDIC-tree ^{2}Cys78 is marked green.

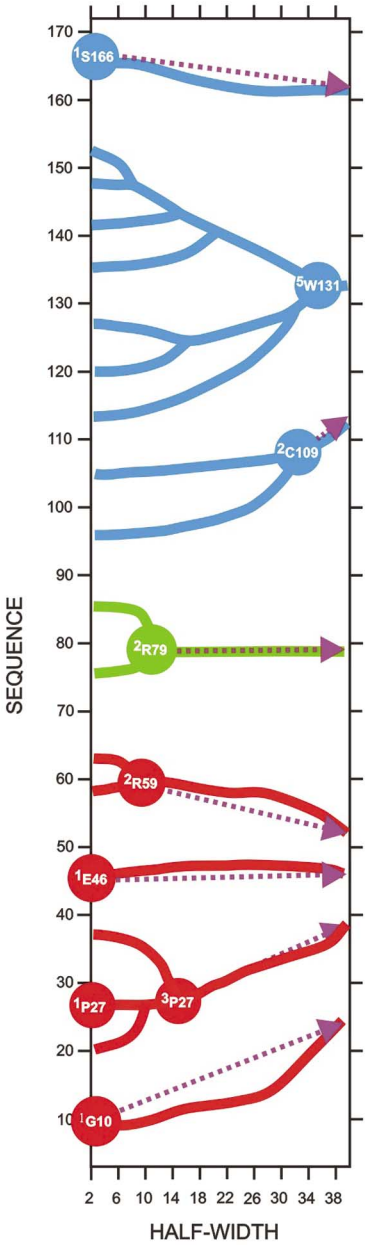

Figure 1: The information structure (an IDIC-diagram) of γ-crystallin. The axes show serial numbers of amino acid resiques in the protein sequence and the half-width values of the decomposition function. IDIC-associations {^{1}G10, ^{3}P27, ^{1}E46, ^{2}G59} and {^{2}C109, ^{4}W131, ^{1}S166} are colored red and blue, respectively. An independent IDIC-tree ^{2}Cys78 is marked green. Dashed arrows indicate the directions in which the centers of the decomposition functions are shifted along the sequence upon increasing their half-width.

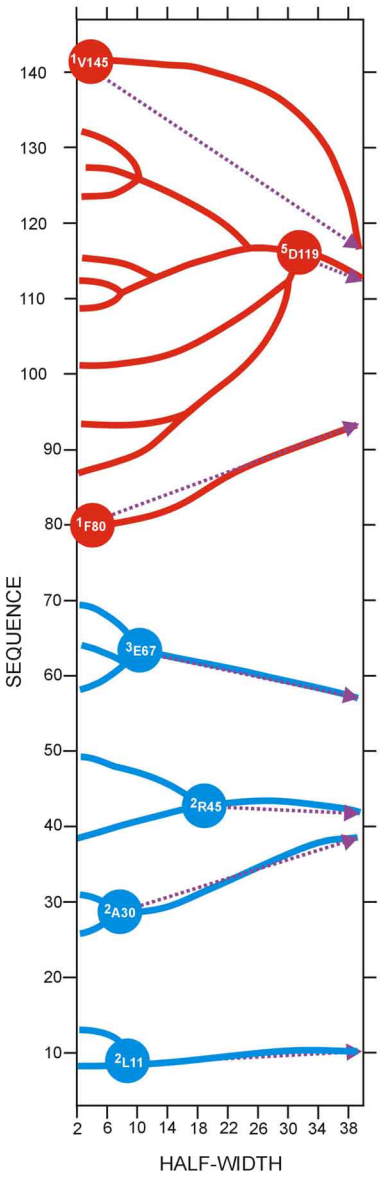

Figure 3: IDIC-diagram of Myosin Regulatory Domain c-chain. The IDIC-associations {^{2}L11, ^{2}A30, ^{2}R45, ^{3}E67} and {^{1}F83, ^{5}D119, ^{1}V145} identified according to Criterions 1 and 2 are marked blue and red. Dashed arrows indicate the directions in which the centers of the decomposition functions are shifted along the sequence upon increasing their half-width.

the information structure of γ-crystallin includes two IDIC-associations, {^{1}G10, ^{3}P27, ^{1}E46, ^{2}G59} and {^{2}C109, ^{5}W131, ^{1}S166}, and an independent IDIC-tree, ^{2}C78 (Fig. 1). The IDIC-associations {^{1}G10, ^{3}P27, ^{1}E46, ^{2}G59} and {^{2}C109, ^{5}W131, ^{1}S166} correspond to the polypeptide fragments GLY1 - TRP68 and MET90 - TYR174, respectively. The independent IDIC-tree ^{2}C78, which is marked green on the IDIC-diagram (Fig. 1) and on the spatial drawing (Fig. 2), covers the residues MET69 - ARG89. There is a correspondence between the elements of the spatial structure (Fig. 2) and the information-structural elements of γ-crystallin. It is seen from Figure 2 that IDIC-associations correspond to the structural domains of γ-crystallin, and the IDIC-tree ^{2}Cys78 corresponds to the elongated polypeptide chain connecting the two domains. At SER77 – ILE81, this polypeptide chain forms a β-strand included in the N-terminal domain. Presumably, being the third β-strand in the β-sheets of the N-terminal domain, it is incorporated in the β-structure at the later stages of its formation during the mutual orientation of the domains.

The N-terminal domain is represented in the information structure by an IDIC-association in which three components out of four are of lower order. Two of the elements, ^{1}G10 and ^{1}E46, are independent, first-order IDIC-branches. Conversely, the IDIC-association corresponding to the C-terminal domain is based on a high-order IDIC-tree, ^{5}W131 (CYS109 - TRP157). However, the x-ray data indicate that both domains have comparable compactness characteristics.

A model protein of the α-helical structural type is Myosin Regulatory Domain (1SCM-C.PDB) (14). Its IDIC-diagram is shown in Figure 3. Arrows indicate the direction of shifting along the sequence of the decomposition functions upon increasing their half-width. Based on Criterion 2 (the use of Criterion 1 gives identical results), the information-structural elements can be combined in two IDIC-associations, {^{2}L11, ^{2}A30, ^{2}R45, ^{3}E67} and {^{1}F83, ^{5}D119, ^{1}V145}, which correspond to the polypeptide chain fragments SER4 – CYS75 and MET76 – PRO152, respectively. The correspondence between these IDIC-associations and the spatial structure is seen from Figure 4. The polypeptide fragments corresponding to the IDIC-associations are marked in red and blue colors. The IDIC-associations {^{2}L11, ^{2}A30, ^{2}R45, ^{3}E67} and {^{1}F80, ^{5}D119, ^{1}V145} are well matched with the visually identifiable structural domains. The structural domains are equally compact regardless of whether the corresponding IDIC-association consists of several lower-order IDIC-trees (viz., the N-terminal domain) or of a single, higher-order IDIC-tree (^{5}D119 for the C-terminal domain). Note that the independent IDIC-branch ^{1}F80 (marked by green color in Fig. 5), which connects the two domains, participates in the IDIC-association {^{1}F80, ^{5}D119, ^{1}V145}. This differs from the case of γ-crystallin, where the IDIC-tree ^{2}R79 connecting the two domains does not participate in IDIC-associations.

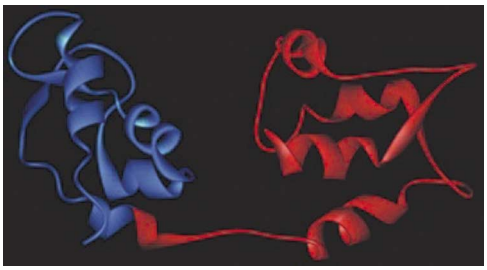

Figure 4: The spatial structure of Myosin Regulatory Domain C-chain. The polypeptide chain fragments corresponding to the IDIC-associations {^{2}L11, ^{2}A30, ^{2}R45, ^{3}E67} and {^{1}F80, ^{5}D119, ^{1}V145} are marked blue and red.

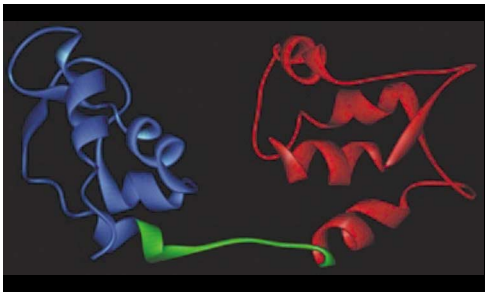

Figure 5: The spatial structure of Myosin Regulatory Domain C-chain. The polypeptide chain fragments corresponding to the IDIC-trees ^{2}L11, ^{2}A30, ^{2}R45 and ^{3}E67 are marked blue; the IDIC-trees ^{5}D119 and ^{1}V145, red. The polypeptide chain fragment (MET76 – ALA84) corresponding to the IDIC- branch ^{1}F80 is marked green. In the spatial structure this fragment connects two domains.

The information-analyzed structures of γ-crystallin and Myosin Regulatory Domain suggest that structural domains tend to be clearly separated in cases when the IDIC-association corresponding to the second domain is based on high-order IDIC-trees. Apparently, an IDIC-association consisting of several lower-order IDIC-trees tends to be incorporated in spatial structures corresponding to the IDIC-associations that form the preceding domain. Some relevant examples will be discussed below.

At a glance, the spatial structure of Haloalkane Dehalogenase (1EDE.PDB) (14, 15) looks like a tightly packed globule devoid of domain structures. Our method, however, may reveal more detailed structural features than a visual analysis permits. An IDIC-diagram of this protein is shown in Figure 6. The IDIC-associations {^{1}N14, ^{4}P32, ^{3}V77, ^{2}R112, ^{2}R140}, {^{2}S162, ^{2}S183, ^{4}D212}, and {^{3}I256, ^{3}H289} identified by Criterions 1 and 2 are marked in red, blue, and green colors. The spatial structure elements that correspond to these IDIC-associations are marked by the respective colors in Figure 7. It is seen that the N-terminal domain (MET1 - CYS150, marked by red color) is formed into a β-sheet consisting of six β-strands. In terms of information structure, this domain corresponds to an IDIC-association {^{1}N14, ^{4}P32, ^{3}V77, ^{2}R112, ^{2}R140}. The following, central domain (colored blue) is consists of five α-helices forming a nearly flat structure that partially shields the globule. This central domain, which is located in a cavity between the N- and C-terminal domains and interacts with segments of their polypeptide chains, seems to stabilize the whole spatial structure. The central domain corresponds to an IDIC-association {^{2}S162, ^{2}S183, ^{4}D212}. Finally, the C-terminal domain (ASN246 – GLU310), marked by green color in Figure 7, is described by an IDIC-association {^{3}I256, ^{3}H289}. It is the smallest in size and consists of β-α-β-α structures forming a β-sheet, which is an extension of the N-terminal β-sheet. This double-stranded β-sheet is covered from two sides by the α-helices. Thus, information analysis is capable of identifying separate substructures even in tightly packed globular proteins. The above example demonstrates that even apparently homogeneous structural units, such as the eight-stranded β-sheet, are in fact composite structures.

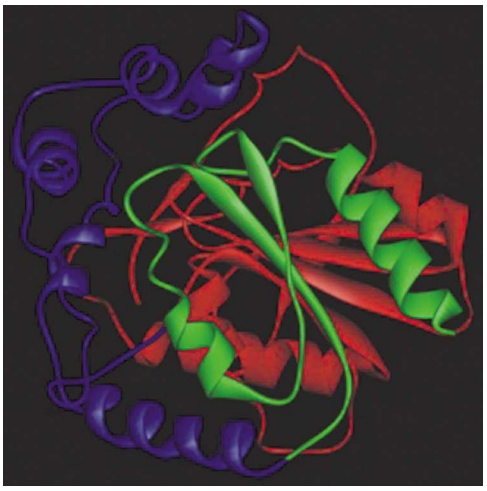

Figure 7: The spatial structure of Haloalkane Dehalogenase. The polypeptide chain fragments corresponding to the IDIC-associations {^{1}N14, ^{4}P32, ^{3}V77, ^{2}R112, ^{2}R140}, {^{2}S162, ^{2}S183, ^{4}D212} and {^{3}I256, ^{3}H289} are colored red, blue, and green, respectively.

The following examples further demonstrate the utility of the information analysis method.

Let us consider Phthalate Dioxygenase (2PIA.PDB) (16). Unlike the previously considered cases, its IDIC-diagram (Fig. 8) shows that here Criterions 1 and 2 produce different results when detecting IDIC-associations. If Criterion 1 is used, the IDIC-associations are separated by a polypeptide chain segment TRP164 - HIS214, whereas according to Criterion 2, this segment contains IDIC-trees ^{3}P177 and ^{2}T207 that belong to different IDIC-associations. The directions in which the centers of the decomposition functions are shifted upon increasing the half-width (Criterion 2) are indicated by arrows in Figure 8. Thus, according to Criterion 2, the information structure of this protein includes three IDIC-associations: {^{2}K12,

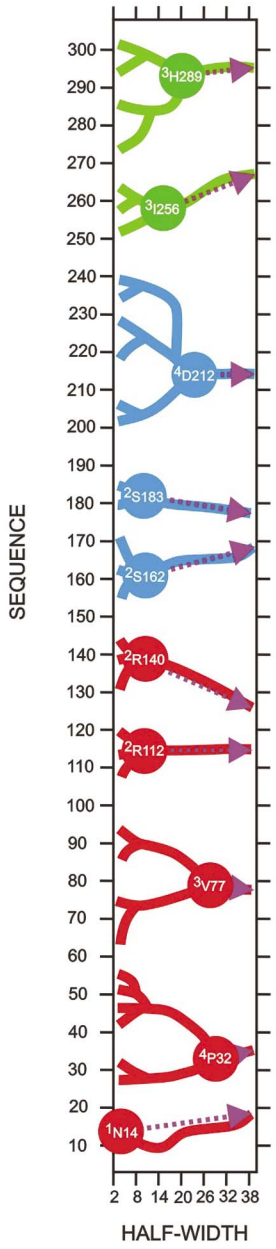

Figure 6: The IDIC-diagram of Haloalkane Dehalogenase. The IDIC-associations {^{1}N14, ^{4}P32, ^{3}V77, ^{2}R112, ^{2}R140}, {^{2}S162, ^{2}S183, ^{4}D212} and {^{3}I256, ^{3}H289} are colored red, blue, and green, respectively. Dashed arrows indicate the directions in which the centers of the decomposition functions are shifted along the sequence upon increasing their half-width.

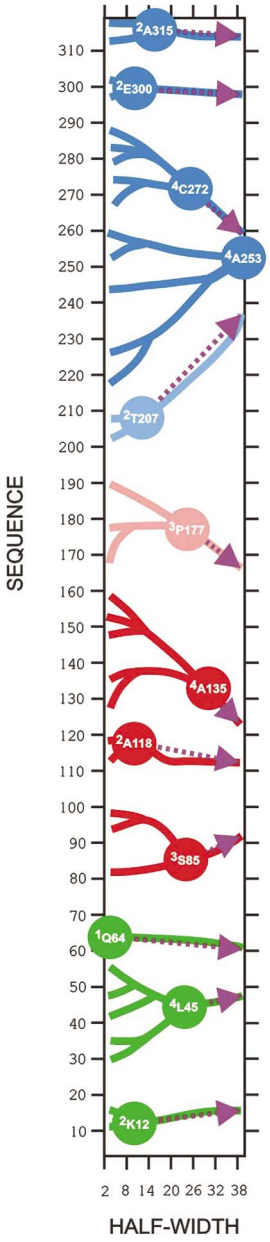

Figure 8: The IDIC-diagram of Phthalate Dioxygenase. The information structure consists of three IDIC-associations {^{2}K12, ^{2}R45, ^{1}Q64}, {^{3}S85, ^{2}A118, ^{4}A135, ^{3}P177} and {^{2}T207, ^{4}A253, ^{4}C272, ^{2}E300, ^{2}A315}, which are marked by green, red, and blue color, respectively. The IDIC-trees ^{3}P177 and ^{2}T207, for which the decomposition functions have anomalously low values, are marked by pink and light blue, respectively. Dashed arrows indicate the directions in which the centers of the decomposition functions are shifted along the sequence upon increasing their half-width.

^{2}R45, ^{1}Q64}, {^{3}S85, ^{2}A118, ^{4}A135, ^{3}P177}, and {^{2}T207, ^{4}A253, ^{4}C272, ^{2}E300, ^{2}A315} (marked by green, red, and deep blue colors, respectively). Although they belong to the respective IDIC-associations, the IDIC-trees ^{3}P177 and ^{2}T207 are marked in Figure 8 by special colors (pink and light blue), because the corresponding decomposition functions have anomalously low values. The color scheme of Figure 9 corresponds to that of Figure 8. It is seen that the C-terminal domain has a compact structure some what apart from the α-helix that corresponds to the IDIC-tree ^{2}T207. As we have just observed, this IDIC-tree ^{2}T207 is notable for the lower value of its decomposition function. Thus, peculiarities of the information structure are reflected in the spatial structure of this protein. Although the N-terminal part (THR1 - HIS214) is much more structurally complex, its information analysis reveals a number of interesting features.

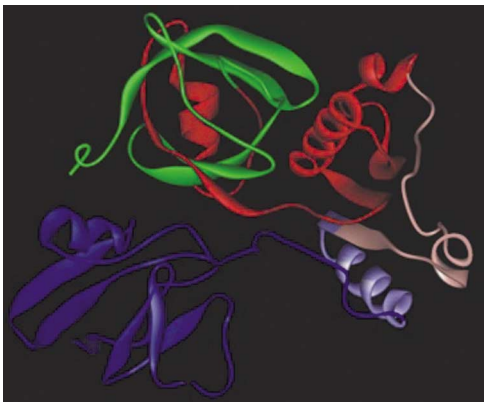

Figure 9: The spatial structure of Phthalate Dioxygenase. The polypeptide chain fragments corresponding to the IDIC-associations {^{2}K12, ^{2}R45, ^{1}Q64}, {^{3}S85, ^{2}A118, ^{4}A135, ^{3}P177} and {^{2}T207, ^{4}A253, ^{4}C272, ^{2}E300, ^{2}A315} are marked by green, red, and blue color, respectively. Elements of the spatial structure corresponding to the IDIC-trees ^{3}P177 and ^{2}T207 are marked by pink and light blue, respectively.

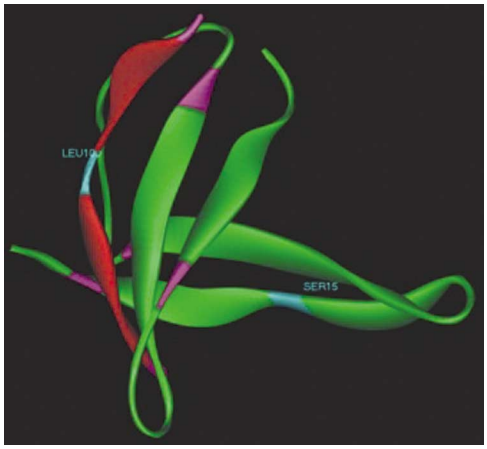

Figure 10: A fragment of the spatial structure of Phthalate Dioxygenase consisting of two β-sheets. Polypeptide chains corresponding to the IDIC-association {^{2}K12, ^{2}R45, ^{1}Q64} and IDIC-tree ^{3}S85 are marked in green and red colors, respectively. The residues SER15 and LEU100 (cyan color) cause deformations of the elongated β-structures (LEU9 – ILE19 and ALA95 – GLU104), breaking them into separate β-strands (LEU9 – ALA14, LYS16 – ILE19, ALA95 – SER99 and ARG102 – GLU104). The first-rank IDIC-branch centers (magenta color) ^{1}D30, ^{1}A43, ^{1}R55, ^{1}A95 _ ^{1}F105 correlate well with the limits of the β-strands.

Figure 11: A β-barrel like structure formed by the structural elements belonging to the IDIC-tree ^{3}S85 (red color) and the IDIC-association {^{2}K12, ^{2}R45, ^{1}Q64} (green color).

First, the IDIC-association {^{2}K12, ^{2}R45, ^{1}Q64}, which is labeled green in Figures 8 and 9, corresponds to five β-strands, which form two β-sheets. Note the deformation of the elongated polypeptide fragment LEU9 - ILE19 at the SER15 residue, which breaks this fragments in two β-strands. These β-strands (LEU9 – ALA14 and LYS16 - ILE19) belong to different β-sheets. These two β-sheets each include, apart from the β-strands that belong to the IDIC-association {^{2}K12, ^{2}R45, ^{1}Q64}, a β-strand from {^{3}S85, ^{2}A118, ^{4}A135, ^{3}P177}. The latter β-strands (ALA95 – SER99 and ARG102 – GLU104) are parts of an elongated β-structure ALA95 – GLU104 with a deformation at LEU100 (Fig. 10), similarly to the LEU9 – ILE19 fragment, in which the deformation at SER15 divides the β-structure into two β-strands. Note that both deformation sites (SER15 and LEU100, marked by cyan color in Fig. 10) correspond to IDIC-branches ^{1}SER15 and ^{1}LEU100 (Fig. 8). Location of the first-rank IDIC-branches ^{1}D30, ^{1}A43, ^{1}R55, ^{1}A95 and ^{1}F105 correlates well with the β-strand borders (marked by magenta color in Fig. 10). Thus, all the structural elements (the two β-strands ALA95 – SER99 and ARG102 – GLU104 and the α-helix GLY82 – ASP88) corresponding to the IDIC-tree ^{3}S85 are integrated into the spatial structure corresponding to the IDIC-association {^{2}K12, ^{2}R45, ^{1}Q64} (Fig. 11). The β-strands ALA95 – SER99 and ARG102 – GLU104 are incorporated in the β-sheet formed by the β-strands from the IDIC-association {^{2}K12, ^{2}R45, ^{1}Q64} and form a stable spatial structure resembling a β-barrel (Fig. 11), which is stabilized by the α-helix (GLY82 – ASP88). As β-barrels are among the most stable spatial structures, one may assume that inclusion of the IDIC-tree ^{3}S85 into spatial structures corresponding to the IDIC-association {^{2}K12, ^{2}R45, ^{1}Q64} stabilizes the spatial structures formed by these polypeptide fragments.

Another important observation is that IDIC-trees ^{2}A118, ^{4}A135, and ^{3}P177 of the IDIC-association {^{3}S85, ^{2}A118, ^{4}A135, ^{3}P177} form, together with the IDIC-tree ^{2}T207 of the IDIC-association {^{2}T207, ^{4}A253, ^{4}C272, ^{2}E300, ^{2}A315}, a compact structure consisting of a β-sheet formed by three β-strands (SER13 – GLY119, SER140 – THR147 and GLN194 – CYS199) and covered by four α-helices (ILE123 – GLU136, PHE156 – THR160, PHE183 – PHE187 and GLN202 – MET211). Note that the decomposition function for the IDIC-tree ^{2}T207 is anomalously low, this peculiarity of the information structure is reflected in the spatial structure of the protein.

It is important that in both cases the "folded-in" polypeptide fragments exactly correspond to elements of the information structure; namely to IDIC-trees ^{3}S85 and ^{2}T207. Perhaps IDIC-trees are the most stable elements of the information structure.

Another interesting example is Porphobilinogen deaminase (1PDA.PDB) (17), in which one of IDIC-trees corresponds to elements of secondary structure that belong to two different domains. Its information structure is shown in Figure 12. IDIC-associations {^{2}I8, ^{5}K51, ^{2}E78}, {^{3}G112, ^{3}V142, ^{4}R174}, and {^{1}G230, ^{2}E241, ^{2}VE265,

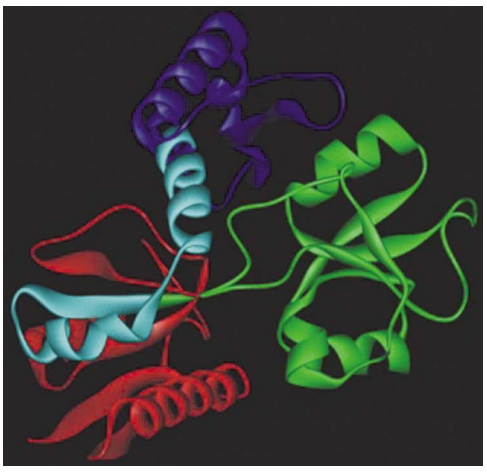

Figure 13: The spatial structure of Porphobilinogen deaminase. The polypeptide chain fragments corresponding to the IDIC-associations {^{2}I8, ^{5}K51, ^{2}E78}, {^{3}G112, ^{3}V142, ^{4}R174} and {^{1}G230, ^{2}E241, ^{2}VE265, ^{1}E283} are colored red, green, and blue, respectively. The structural fragment corresponding to IDIC-tree ^{3}L209 is marked by cyan color.

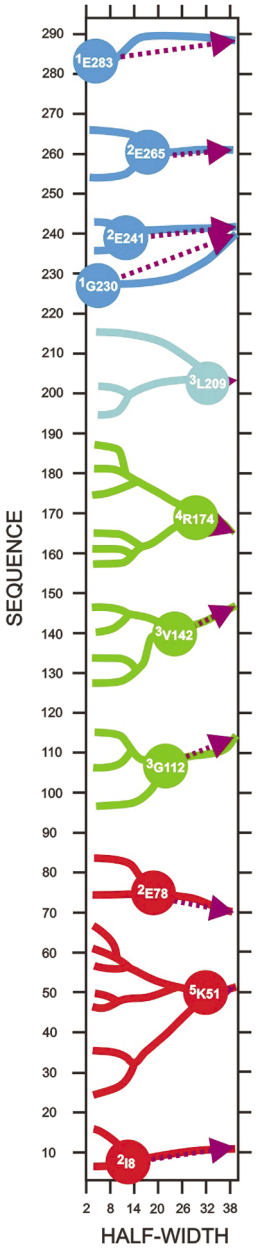

Figure 12: The IDIC-diagram of Porphobilinogen deaminase. The IDIC-associations {^{2}I8, ^{5}K51, ^{2}E78}, {^{3}G112, ^{3}V142, ^{4}R174}, and {^{1}G230, ^{2}E241, ^{2}VE265, ^{1}E283} are colored red, green, and blue, respectively. An independent IDIC-tree ^{3}L209 is marked by cyan color. Dashed arrows indicate the directions in which the centers of the decomposition functions are shifted along the sequence upon increasing their half-width.

^{1}E283} are marked by red, green, and blue colors, respectively. An independent IDIC-tree ^{3}L209 is marked by cyan color. This color scheme is preserved in Figure 13 which shows the spatial corresponding structure elements. Visual analysis easily identifies the domains corresponding to IDIC-associations. The spatial structure elements corresponding to the IDIC-tree ^{3}L209 are of special interest. In the primary structure, this IDIC-tree ^{3}L209 is located between the central and C-terminal domains. The corresponding spatial structural elements (a β-strand and two α-helices) belong to the N-terminal and C-terminal spatial domains. The residues (VAL192 – LEU198) from IDIC-tree ^{3}L209 form a β-strand which is located in the hydrophobic nucleus of the N-terminal domain, being the fourth of the five strands that form the N-terminal β-sheet. Note that the fifth β-strand (LEU84 – VAL87) is surrounded by long, irregular polypeptide fragments. This may be the mechanism by which the more distant β-strand (VAL192 – LEU198) is also incorporated in this β-sheet. Apart from the β-strand (LEU84 – VAL87), the residues covered by the IDIC-tree ^{3}L209 also form two α-helices (ARG202 – ASN211 and GLU214 – MET225) in the, respectively. Thus, the residues from the IDIC-tree ^{3}L209 play crucial role in the mutual location of the N- and C-terminal domains. Changing either the mutual orientation of the α-helices or their degree of compactness will affect the spatial location of the C-terminal domain with respect to the N-terminal and central domains. Note that the mutual orientation of the central and N-terminal domains is less variable, because these domains are interconnected by polypeptide stretches GLU91 – ARG96 and VAL187 – VAL192. Thus, the IDIC-tree ^{3}L209 is very important in the overall spatial organization. Conformational changes in the α-helices belonging to this independent IDIC-tree ^{3}L209 may have important consequences regarding both the folding and biological functions of this protein. It is of primary importance that this peculiar feature of the spatial organization is reflected in the information structure of the primary sequence analyzed by the proposed method.

Conclusions

I. In protein sequences there are regions of increased "information coupling" between amino acid residues (IDIC-sites). As a rule, these regions are organized in hierarchic structures. A network of IDIC-sites determines the information structure of a protein sequence.
II. A method is proposed, which allows the information structure of protein sequences to be analyzed and described in terms of "information units" and higher-order information-structure elements (IDIC-branches and IDIC-trees). Criteria for combining IDIC-branches and IDIC-trees into IDIC-associations are proposed and tested using proteins with known spatial structures.
III. Spatial structures and information structures of proteins are shown to correlate. It is demonstrated that protein folding may depend on structural elements of polypeptide chains which have their counterparts in the information structure.
IV. IDIC-associations were shown to correspond well to the structural domains identified in the spatial structures of selected proteins.

Acknowledgment

The author wishes to thank Dr. Vladimir Ovcharenko for useful discussions and help in preparation of the manuscript.

References and Footnotes

1. R. Sowdhamini and T. L. Blundell. *Protein Science 4*, 506-520 (1995)
2. A. S. Siddiqui and G. J. Barton. *Protein Science 4*, 872-884 (1995)
3. R. Sowdhamini, S. D. Rufina and T. L. Blundell. *Fold. Des. 1*, 209-220 (1996)
4. Chung-Jung Tsai and R. Nussinov. *Protein Science 6*, 24-42 (1997)
5. R. Sowdhamini, D. F. Burke, C. Deane, J. F. Huang, K. Mizuguchi, H. A. Nagarajaram, J. P. Overington, N. Srinivasan, R. E. Steward and T. L. Blundell. *Acta Crystallogr. D Biol. Crystallogr. 54*, 1168-1177 (1998)

6. C. A. Orengo, A. M. Martin, G. Hutchinson, S. Jones, D. T. Jones, A. D. Michie, M. B. Swindells and J. M. Thornton. *Acta Crystallogr. D Biol. Crystallogr. 54*, 1155-1167 (1998)
7. U. Dengler, A. S. Siddiqui and G. J. Barton. *Proteins 42*, 332-344 (2001)
8. X. Guan and L. Du. *Bioinformatics 14*, 783-788 (1998)
9. A. Marchler-Bauer, A. R. Panchenko, N. Ariel and S. H. Bryant. *Proteins 48*, 439-446 (2002)
10. R. A. George and J. Heringa. *Proteins. 48*, 672-681 (2002)
11. D. J. Rigden. *Protein Eng. 15*, 65-77 (2002)
12. A. N. Nekrasov. *J. Biomol. Struct. Dynam. 20*, 87-92 (2002)
13. G. Wistow, B. Turnell, L. Summers, C. Slingsby, D. Moss, L. Miller, P. Lindley and T. Blundell. *J Mol. Biol. 170*, 175-202 (1983)
14. X. Xie, D. H. Harrison, I. Schlichting, R. M. Sweet, V. N. Kalabokis, A. G. Szent-Gyorgyi and C. Cohen. *Nature 368*, 306-312 (1994)
15. K. H. G. Verschueren, S. M. Franken, H. J. Rozeboom, K. H. Kalk, B. W. Dijkstra. *J Mol. Biol. 232*, 856-872 (1993)
16. S. M. Franken, H. J. Rozeboom, K. H. Kalk and B. W. Dijkstra. *EMBO J. 10*, 1297-1302 (1991)
17. C. C. Correll, C. J. Batie, D. P. Ballou and M. L. Ludwig. *Science 258*, 1604-1610 (1992)
18. G. V. Louie, P. D. Brownlie, R. Lambert, J. B. Cooper, T. L. Blundell, S. P. Wood, M. J. Warren, S. C. Woodcock and P. M. Jordan. *Nature 359*, 33-39 (1992).

Date Received: October 27, 2003

Communicated by the Editor Valery Ivanov