FEBS 28436 FEBS Letters 567 (2004) 239–242

Functionally and structurally relevant residues of enzymes: are they segregated or overlapping?

Csaba Magyar, Éva Tüdős, István Simon*

Biological Research Center, Institute of Enzymology, Hungarian Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary

Received 22 January 2004; accepted 23 April 2004

Available online 8 May 2004

Edited by Robert B. Russell

Abstract There is a delicate balance between stability and flexibility needed for enzyme function. To avoid undesirable alteration of the functional properties during the evolutionary optimization of the structural stability under certain circumstances, and vice versa, to avoid unwanted changes of stability during the optimization of the functional properties of proteins, common sense would suggest that parts of the protein structure responsible for stability and parts responsible for function developed and evolved separately. This study shows that nature did not follow this anthropomorphic logic: the set of residues involved in function and those involved in structural stabilization of enzymes are rather overlapping than segregated.

Keywords: Binding sites; Catalytic sites; Stabilization center; Statistical analysis

1. Introduction

Current proteins are the products of multimillion years of evolution. During this long period many proteins have been developed with similar structures but different functions. The tertiary structure of proteins must fulfill more constraints than the primary structure, thus the number of significantly different structures and the number of protein folds are several orders of magnitude smaller than the number of proteins [1–3]. Many proteins from various organisms with different structural stability exhibit the same function [4–6]. These observations suggest that functional regions of the polypeptide chains evolved independently from the regions which are responsible for the structural stability, in order to avoid interference in the course of optimization. In fact, some authors classify conservative residues as structural or functional residues [7]. Furthermore, it is generally accepted that functionally important residues are mainly solvent-accessible residues on the protein surface, while structurally important residues are likely part of the protein core [8].

Our earlier works suggested that residues responsible for the function of major histocompatibility complex proteins play a key role in their structural stabilization [9]. In PD-(D/E)XK (type II) endonucleases, the active site residues and some residues involved in DNA recognition are frequently involved in stabilization centers (SCs; see Section 2) [10,11]. Luque et al. [12] discussed the structural stability of binding sites as characterized by hydrogen/deuterium exchange and found that catalytic residues are mostly located in high stability regions, while binding sites can be found in both high and low stability regions. Likewise, active site residue mutations in AmpC β-lactamase caused decreased activity but increased structural stability [13]. These observations indicate an overlap between functionally and structurally important residues.

The aim of this work was to analyze the relationship between functionally and structurally relevant residues using statistical approaches on a dataset of 417 polypeptide chains, which will be referred as overlapping functional–structural residue (OFSR) dataset. For this purpose subsets of these residue classes were analyzed. As a subset of functionally relevant residues (F subset) of enzymes, we considered residues that had SITE records in the protein data bank (PDB) entries, representing active site, substrate, coenzyme or effector binding site residues. A similar subset of functionally relevant residues with SITE records was also used by Sternberg and co-workers [14]. To probe the dependence of the results on the selection of the dataset or the definition of functional residues, the analysis has been repeated using the CATRES dataset [15]. Due to the smaller size of the CATRES dataset, however, greater statistical uncertainty is expected, a more detailed statistical analysis was done only on the larger OFSR dataset. As a subset of structurally important residues (S subset), we considered SC residues. SC residues are elements of clusters of residues involved in cooperative long-range interaction, i.e., elements of non-covalent crosslinks in proteins. They are involved in 70% more nonlocal interactions than other residues and tend to have lower than average B-factors in X-ray structures and have elevated conservation [10]. A slightly increased number of SC elements was found in proteins from thermophilic sources, compared to their mesophilic counterparts [16]. SCs were shown to modify the helix–helix orientation in four-helix bundles and to contribute to the stabilization free energy [17]. These observations indicate that SC residues are of structural relevance.

2. Materials and methods

2.1. SC definition

SC residues are defined based on the contact map of a protein with known three-dimensional structure [10]. Two residues are in contact if there is at least one pair of heavy atoms with a distance less than the

* Corresponding author. Fax: +36-1-466-5465.
E-mail address: simon@enzim.hu (I. Simon).

Abbreviations: EC, enzyme commission; PDB, protein data bank; SC, stabilization center; OFSR, overlapping functional–structural residue

0014-5793/$22.00
doi:10.1016/j.febslet.2004.04.070

sum of the van der Waals radii of the two atoms plus 1.0 Å. Long-range contacts are defined as contacts between residues, which are separated by at least 10 residues in the amino acid sequence or they are not part of the same polypeptide chain. Two residues are SC elements if they are involved in long-range contacts and at least one supporting residue could be found in each of the flanking tetra-peptides of these residues in such a way that at least seven out of the possible nine interactions were formed between the two triplets. Stabilization centers were identified with the SCide public server (http://www.enzim.hu/scide) [18].

2.2. Dataset

The Families of Structurally Similar Proteins [19] database released on 16th June 2002 was downloaded from "ftp://ftp.ebi.ac.uk/pub/databases/fssp". The initial dataset contained 27 180 protein chains from 14 437 proteins, belonging to 2859 representative families. All representative polypeptide chains were checked against SITE records in the PDB [20] entries, enzyme commission (EC) numbers [21] in the COMPND records, and whether the chain contains SC elements. Only 280 polypeptide chains fulfilled all these conditions. For the rest of the representative enzymes, homologous structures were selected with the highest possible similarity, utmost SITE records and lowest RMSD values. This search resulted in 140 new entries. After checking the whole dataset for redundancy using the BLASTclust program [22] (with a score density threshold value of 0.8), 417 entries remained containing 3618 functionally important residues. The complete dataset is available under: "http://www.enzim.hu/~magyarcs/func_stab.html". The dataset has been divided into several secondary structural subclasses based on the SCOP classification [23]. Protein chains with differently classified domains were considered as multi-domain (class-E) proteins. Different biochemical functions were also taken into account by dividing the dataset according to the EC numbering [21]. The CATRES dataset [15], containing 614 catalytic residues from 177 entries in 181 polypeptide chains was also analyzed.

2.3. Data analysis

The expected number of residues occurring in both F and S subsets of the ith polypeptide chain was calculated as

$$N^i_{\mathrm{exp}} = \frac{N^i_{\mathrm{F}}}{N^i_{\mathrm{total}}}\frac{N^i_{\mathrm{S}}}{N^i_{\mathrm{total}}}N^i_{\mathrm{total}} = \frac{N^i_{\mathrm{F}}N^i_{\mathrm{S}}}{N^i_{\mathrm{total}}}$$

where N^i_{F} is the number of residues in the F subset; N^i_{S} is the number of residues in the S subset and N^i_{total} is the total number of residues. The measure of the overlap between functionally and structurally relevant residues for a single protein was calculated as

$$O^i = \frac{N^i_{\mathrm{obs}}}{N^i_{\mathrm{exp}}} = \frac{N^i_{\mathrm{obs}}}{N^i_{\mathrm{F}}N^i_{\mathrm{S}}/N^i_{\mathrm{total}}} = \frac{N^i_{\mathrm{obs}}N^i_{\mathrm{total}}}{N^i_{\mathrm{F}}N^i_{\mathrm{S}}}$$

where N^i_{obs} is the observed number of residues belonging to both F and S subsets. For the whole dataset three different values were calculated:

$$O_1 = \frac{\sum_{i=1}^{N_{\mathrm{ch}}} O^i}{N_{\mathrm{ch}}}, \quad O_2 = \frac{\sum_i N^i_{\mathrm{obs}}}{\sum_i N^i_{\mathrm{exp}}} = \frac{\sum_i N^i_{\mathrm{obs}}}{\sum_i \frac{N^i_{\mathrm{F}}N^i_{\mathrm{S}}}{N^i_{\mathrm{total}}}} \text{ and}$$

$$O_3 = \frac{\sum_i N^i_{\mathrm{obs}}}{\sum_i N^i_{\mathrm{F}}\sum_i N^i_{\mathrm{S}}/\sum_i N^i_{\mathrm{total}}} = \frac{\sum_i N^i_{\mathrm{obs}}\sum_i N^i_{\mathrm{total}}}{\sum_i N^i_{\mathrm{F}}\sum_i N^i_{\mathrm{S}}}$$

where N_{ch} is the number of polypeptide chains in the dataset (417 and 181 for the OFSR and the CATRES dataset, respectively).

The mean value (O_1) is calculated as an average of the O^i values. The O_2 value is calculated as the ratio of the sums of expected and observed values of overlapping residues over all protein chains. While O_1 gives equal weights to the proteins of various sizes, O_2 implicitly weighs each protein with its total number of residues. We created a hypothetical protein by merging all sequences with SC and SITE residues already identified into a single sequence and calculated the O value (O_3). The median of the O^i values was derived from the whole dataset and its functional (EC) and structural (SCOP) subclasses in order to show that the observed overlap is not due to a few proteins with high O^i values. To check the significance of the observed high overlap between the S and F subsets, 100 randomized control datasets were generated. For each protein of the control datasets, the same number of residues as in the F subsets (N^i_{F}) was selected randomly and the statistical analysis was repeated with these residues instead of the SITE residues. The same O_1, O_2 and O_3 overlap values and their standard deviations were calculated for the 100 randomized datasets.

3. Results and discussion

The overlap between functionally and structurally relevant residues has been investigated by calculating the correlation between these two residue classes in our OFSR dataset of 417 polypeptide chains containing 136 810 residues. O values for all studied proteins are available at "http://www.enzim.hu/~magyarcs/func_stab.html", while the result of the statistical analysis is summarized in Fig. 1 and Table 1. Fig. 1 shows the

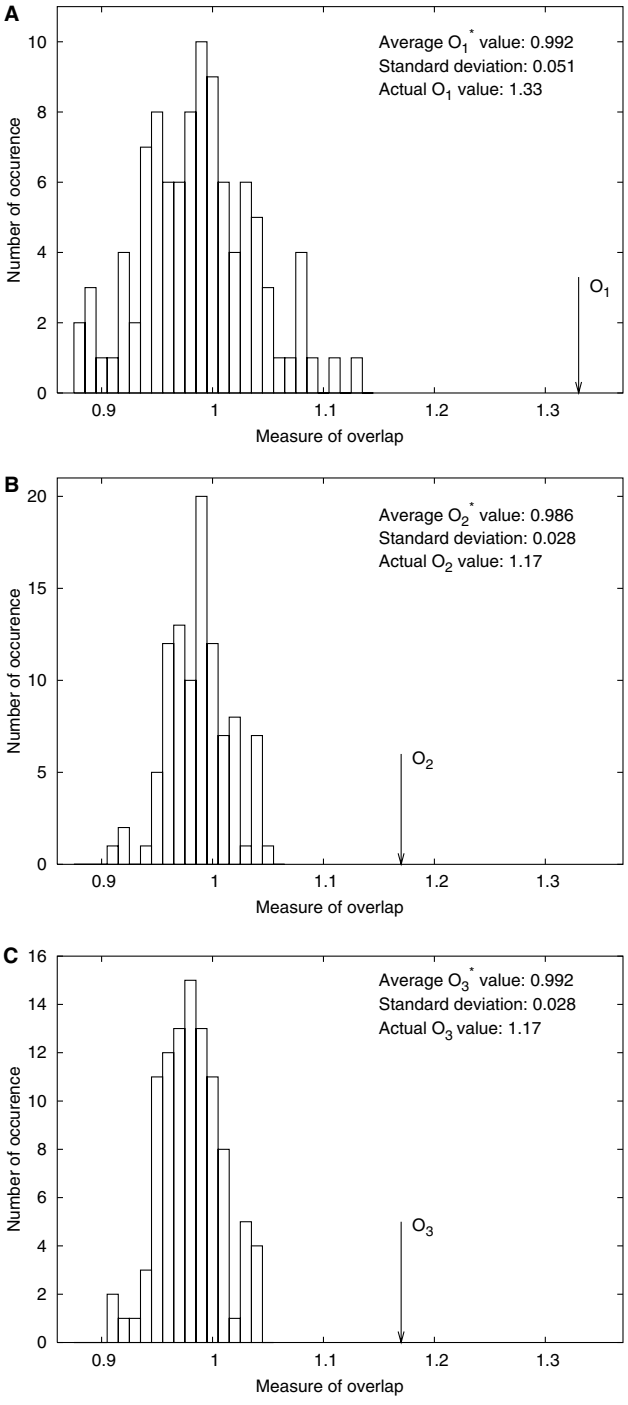

Fig. 1. Distribution of O^*_1 (A), O^*_2 (B) and O^*_3 (C) over 100 randomized datasets and the actual O_1, O_2 and O_3 values of the OFSR dataset.

Table 1
Main properties and O values for the OFSR dataset with its various subsets and for the CATRES dataset [15]

Subset	N_{ch}	N_{total}	O_1	O_2	O_3	M
EC 1	95	35 625	1.31	1.12	1.13	1.14
EC 2	81	26 196	1.19	1.19	1.20	1.09
EC 3	166	49 043	1.37	1.24	1.21	1.37
EC 4	46	16 131	1.38	1.07	1.06	1.03
EC 5	18	5612	1.65	1.35	1.38	1.48
EC 6	11	4203	1.13	0.94	0.88	0.98
Total	417	136 810	1.33	1.17	1.17	1.20
SCOP A	26	8879	1.30	1.27	1.22	1.35
SCOP B	53	14 650	1.14	1.07	1.04	1.11
SCOP C	135	44 189	1.29	1.11	1.14	1.15
SCOP D	76	18 285	1.28	1.34	1.26	1.25
SCOP E	113	47 870	1.46	1.22	1.19	1.23
SCOP other	14	2937	1.76	1.36	1.26	2.05
CATRES	181	62 075	1.13	1.19	1.20	1.05

N_{ch} is the number of polypeptide chains, N_{total} is the sum of total number of residues in the various subsets. O_1, O_2 and O_3 values are measures of overlap between functionally and structurally relevant residues. M values are the medians derived from the O^i values of the individual proteins.

distributions of the O_1^*, O_2^* and O_3^* values for 100 random datasets and the actual data for the OFSR dataset. Table 1 shows the O_1, O_2 and O_3 values, and the median of the O^i values for the CATRES dataset, the OFSR dataset and its various subclasses. The mean values and the standard deviations on the 100 random samples are 0.992 ± 0.051, 0.986 ± 0.028 and 0.992 ± 0.028 for O_1, O_2 and O_3, respectively. It is worth noting that the measured overlap as well as its standard deviation is about twice as high in the case of the O_1 value, than for the other two O values. Therefore, the deviations from the averages of the random cases are the same, 6.6 in standard deviation units in all three cases. The probability that this deviation happens by chance on a database of this size is less than 10^{-9}. Our observations that the overlap between F and S subsets is higher, than in a randomized sample, is valid for the OFSR dataset and not for all individual proteins. In fact, it is not valid for more than one-third of the proteins studied. O_2 and O_3 values are almost the same for the random distribution and also for the OFSR dataset, indicating that the hypothetical protein (created by merging all sequences together) gives the same result as the size weighted average of the O^i values of the individual proteins. To confirm whether this considerable overlap is generally valid for all proteins in the dataset, proteins with different secondary structure or different biochemical function have been analyzed separately. Table 1 shows the results for the six SCOP and six EC subclasses.

To investigate further if our results are generally valid or the consequence of our special definition of functionally important residues, the same statistical analysis was carried out on the CATRES dataset, which contains only catalytic residues in contrast to the SITE residues used in all other statistics. This dataset contained 614 catalytic site residues in 181 polypeptide chains, the obtained results are listed in Table 1. The O_2 and O_3 values are similar to the values obtained on the OFSR dataset. The lower O_1 and median values can be explained with a relatively large number of proteins with zero O^i values with no common residues in the F and S subsets due to the low number of catalytic residues per protein in the CATRES dataset. Therefore, only the calculation of O values, which are based on sums over the whole dataset (particularly O_3), is meaningful.

The expected number and measure of overlaps were also calculated for the 20 different amino acids using the same statistical approach as for the entire dataset (available at "http://www.enzim.hu/~magyarcs/func_stab.html"). Since the majority of the 20 types of residues are missing from the F subset of many individual proteins, only the O_3 values can be used. In a few cases like for Glu, Asp and Gln, the measure of overlap has a statistically significant value (2.09, 1.78 and 1.70). As a rule of thumb, charged and some polar residues appear rather often among the observed overlaps. Except Cys, none of the polar residues have smaller observed than expected value for overlap and there are only a few apolar residues (like Ile and Val) where the observed value falls below the expected one. It indicates that the higher overlap between the functionally and structurally important residues is due to the fact that many functionally important hydrophilic residues do appear among the structurally relevant residues, and not because structurally important hydrophobic residues are over represented in the F subset.

The results obtained on the whole dataset and practically on all of its subsets indicate that the overlap between the functionally and structurally relevant residues is higher than expected on statistical basis. This is just the opposite of what one would expect with an anthropomorphic logic; the division of labor among structurally and functionally relevant residues. For allosteric enzymes, the overlap of functional and structural residues might contribute to the propagation of the conformational changes [12]. However, the high level of overlap as a general rule looks peculiar, especially in the light of the rather apolar character of the majority of structurally relevant residues and the rather polar and charged character of most of the residues involved in various enzyme functions, like substrate binding. Bartlett et al. [15] have recently showed a surprising result about the localization of catalytic site residues. Although they are mostly hydrophilic, most of them are in buried positions even in the apo-enzyme. The mostly hydrophobic SC residues also tend to appear in buried parts of the protein. Our results indicate that from the viewpoint of overlap, the

tendency of being buried for both the structurally and the functionally relevant residues is significantly stronger than the difference in polar–apolar character of the two kinds of biologically important residues.

It is worth noting that in the CATRES dataset the subset of functional residues, i.e., exclusively the catalytic residues, represents only 1% of the total residues, while in the OFSR dataset the F subset of functional residues, i.e., all residues in the SITE records, came out at 2.7% of all residues. Therefore, the majority of these SITE residues are not catalytic ones, but rather binding residues and other functionally important residues, which more often appear in surface loops of proteins. Despite the different distribution of residues in the interior and the exterior of proteins considered in the two datasets, the measured overlap is almost the same. This suggests that the preference of the location of the structurally important residues and that of the functionally important ones is not the only reason of the observed high level of overlap.

Acknowledgements: We gratefully acknowledge the support of OTKA Grants D38487 and T34131, and we thank Drs. Zsuzsanna Dosztányi and Mónika Fuxreiter for their helpful comments on improving the manuscript.

References

[1] Chothia, C. and Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins. EMBO J. 5, 823–826.
[2] Chothia, C. and Lesk, A.M. (1987) Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol. 196, 901–917.
[3] Liu, X., Fan, K. and Wang, W. (2004) The number of protein folds and their distribution over families in nature. Proteins 54, 491–499.
[4] Jaenicke, R. (1991) Glyceraldehyde-3-phosphate dehydrogenase from *Thermotoga maritima*: strategies of protein stabilization. FEMS Microbiol. Rev. 18, 215–224.
[5] Wallon, G., Lovett, S.T., Magyar, C., Svingor, A., Szilagyi, A., Zavodszky, P., Ringer, D. and Petsko, G.A. (1997) Sequence and homology model of 3-isopropylmalate dehydrogenase from the psychrotrophic bacterium *Vibrio* sp. I5 suggest reasons for thermal instability. Protein Eng. 10, 665–672.
[6] Szilagyi, A. and Zavodszky, P. (2000) Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Struct. Fold. Des. 15, 493–504.
[7] Golovanov, A.P., Efremov, R.G., Jaravine, V.A., Vergoten, G. and Arsieniev, A.S. (1995) Amino acid residue: is it structural or functional? FEBS Lett. 375, 162–166.
[8] Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Farisell, P., Casadio, R. and Ben-Tal, N. (2004) Conseq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics DOI: 10.1093/bioinformatics/bth070.
[9] Simon, A., Dosztanyi, Z., Rajnavolgyi, E. and Simon, I. (2000) Function-related regulation of the stability of MHC proteins. Biophys. J. 79, 2305–2313.
[10] Dosztanyi, Z., Fiser, A. and Simon, I. (1997) Stabilization centers in proteins: identification, characterization and predictions. J. Mol. Biol. 272, 597–612.
[11] Fuxreiter, M. and Simon, I. (2002) Protein stability indicates divergent evolution of PD-(D/E)XK type II restriction endonucleases. Protein Sci. 11, 1978–1983.
[12] Luque, I., Leavitt, S.A. and Freire, E. (2002) The linkage between protein folding and functional cooperativity: two sides of the same coin? Annu. Rev. Biophys. Biomol. Struct. 31, 235–256.
[13] Beadle, B.M. and Shoichet, B.K. (2002) Structural bases of stability-function tradeoffs in enzymes. J. Mol. Biol. 321, 285–296.
[14] Aloy, P., Querol, E., Aviles, F.X. and Sternberg, M.J. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J. Mol. Biol. 311, 395–408.
[15] Bartlett, G.J., Porter, C.T., Borkakoti, N. and Thornton, J.M. (2002) Analysis of catalytic residues in enzyme active sites. J. Mol. Biol. 324, 105–121.
[16] Simon, A., Dosztanyi, Z., Magyar, C., Szirtes, G., Rajnavolgyi, E. and Simon, I. (2001) Stabilization centers and protein stability. Theor. Chem. Acc. 106, 121–127.
[17] Fuxreiter, M. and Simon, I. (2002) Role of stabilization centers in 4 helix bundle proteins. Proteins 48, 320–326.
[18] Dosztanyi, Z., Magyar, C., Tusnady, E.G. and Simon, I. (2003) SCide: identification of stabilization centers in proteins. Bioinformatics 19, 899–900.
[19] Holm, L. and Sander, C. (1996) Mapping the protein universe. Science 273, 595–602.
[20] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The protein data bank. Nucleic Acids Res. 28, 235–242.
[21] Bairoch, A. (2000) The ENZYME database in 2000. Nucleic Acids Res. 28, 304–305.
[22] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
[23] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.