Four Spatial Points That Define Enzyme Families

Gábor Iván[1,2], Zoltán Szabadka[1,2], Vince Grolmusz[1,2] and Gábor Náray-Szabó[3,4]

[1]Protein Information Technology Group, Eötvös University, 1117 Budapest, Hungary

[2]Uratim Ltd. 4400 Nyíregyháza, Hungary

[3]Laboratory of Structural Chemistry and Biology, Eötvös University, 1117, Budapest, Hungary

[4] Protein Modeling Group of the Hungarian Academy of Science, 1117, Budapest, Hungary

ABSTRACT

Motivation: The catalytic properties of enzymes, containing the ASP-HIS-SER triads are deeply investigated for a long time. Serine endopeptidases, cutinases, acetylcholinesterases, cellulases contain these triads. We found that solely the geometric properties *of just four points* in the spatial structure of these enzymes are characteristic to their family.

1 INTRODUCTION

The Protein Data Bank (1) contains more than 54,000 three-dimensional structures of proteins and nucleic acids today. The data deposited in the PDB make possible intricate data mining and knowledge discovery applications that was unthinkable ten years ago. These applications may address the steric properties of the proteins, while genetic-, or just protein sequence databases are not adequate for these studies, since residues lying far apart in the polypeptide sequence may be neighbors in the three-dimensional structure.

One classical example of three residues of large pairwise distances in the sequence but small enough distances in the three-dimensional space to break up peptide bonds are the Asp-His-Ser and Glu-His-Ser triads in serine proteases and some other enzymes (2). The corresponding amino acids may lie far apart in the primary sequence; however, they are spatially related and form a hydrogen-bonded catalytic machinery, which is especially well suited to accelerate the enzymatic reaction. In a number of cases a (- + -) charge pattern is formed, which is stabilized by neighboring hydrogen bonds (3). This triad works also in active sites of enzymes like serine proteases (2), acetylcholinesterases, dipeptidyl peptidases, and lipases (3).

In the present work, the entries the Protein Data Bank (1) were screened for appearances of the Asp-His-Ser and Glu-His-Ser catalytic triads, and it was found that the geometric positions *of just four points* relative to the histidine ring well characterize some enzyme families.

2 METHODS

From the version of July 20, 2007 of the PDB, 44,840 entries were screened. Alternate locations were not allowed, and only the first model was considered in case of multiple models in NMR data. After an initial filtering for obvious atomic distance requirements (By the notation of the PDB atomic entries, only those entries were considered where in the histidines the distance of ND1 and NE1 was less than 3 Å, only those aspartic acids where the distance of OD1 and OD2 was less than 3 Å, and only those glutamines were considered, where the distance of OE1 and OE2 was less than 3 Å).

Next, those Asp-His-Ser and Glu-His-Ser triads were thrown away where any of the following conditions hold:

- the angle of the halving point of the two oxygens of the COO (denoted by *v* on Figure 1), the halving point of the two nitrogens of the ring of the HIS (denoted by *z* on Figure 1) and SER OG (denoted by *x* on Figure 1) was less than 120°;
- or the *v*-*z* distance is greater than 4.5 Å;
- or the *z*-*x* distance is greater than 4.5 Å;
- the pair-wise distance of any two members of the Asp-His-Ser or Glu-His-Ser triads on the polypeptide chain is less than 10 residues.

2,174 triads were the result of this filtering process.

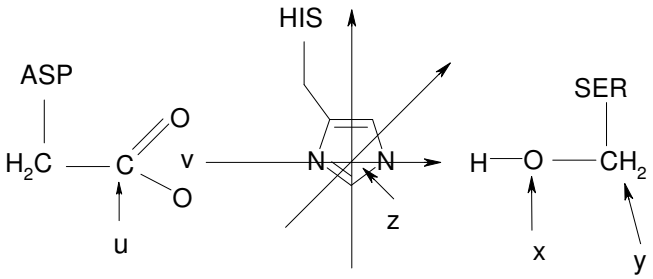


Figure 1: Points *u, v, z, x* and *y* in the triad.

*To whom correspondence should be addressed.

This is a personal *preprint version* of the paper, that is published in the Biochemical and Biophysical Research Communications in a more polished, peer-reviewed form. The DOI: doi:10.1016/j.bbrc.2009.04.022

PDB ID	Content description	Chain ID	Triad positions	excl/incl	verified	Ec number	EC numb. modified
1E66	STRUCTURE OF ACETYLCHOLINESTERASE COMPLEXED WITH (-)- HUPRINE X AT 2.1A RESOLUTION	A	327|440|200	E	OK	3.1.1.7	3.1.1.7
1H22	STRUCTURE OF ACETYLCHOLINESTERASE (E.C. 3.1.1.7) COMPLEXED WITH (S,S)-(-)-BIS(10)-HUPYRIDONE AT 2.15A RESOLUTION	A	327|440|200	E	OK	3.1.1.7	3.1.1.7
1H23	STRUCTURE OF ACETYLCHOLINESTERASE (E.C. 3.1.1.7) COMPLEXED WITH (S,S)-(-)-BIS(12)-HUPYRIDONE AT 2.15A RESOLUTION	A	327|440|200	E	OK	3.1.1.7	3.1.1.7
1U65	ACHE W. CPT-11	A	327|440|200	E	OK	3.1.1.7	3.1.1.7
1W4L	COMPLEX OF TCACHE WITH BIS-ACTING GALANTHAMINE DERIVATIVE	A	327|440|200	E	OK	3.1.1.7	3.1.1.7
1W6R	COMPLEX OF TCACHE WITH GALANTHAMINE DERIVATIVE	A	327|440|200	E	OK	3.1.1.7	3.1.1.7
1W76	ORTHORHOMBIC FORM OF TORPEDO CALIFORNICA ACETYLCHOLINESTERASE (ACHE) COMPLEXED WITH BIS-ACTING GALANTHAMINE DERIVATIVE	B	327|440|200	E	OK	3.1.1.7	3.1.1.7
1ZGB	CRYSTAL STRUCTURE OF TORPEDO CALIFORNICA ACETYLCHOLINESTERASE IN COMPLEX WITH AN (R)-TACRINE(10)-HUPYRIDONE INHIBITOR.	A	327|440|200	E	OK	3.1.1.7	3.1.1.7
2BAG	3D STRUCTURE OF TORPEDO CALIFORNICA ACETYLCHOLINESTERASE COMPLEXED WITH GANSTIGMINE	A	327|440|200	E	OK	3.1.1.7	3.1.1.7

Table 1: A small portion of the TRIAD table, available at the http://pitgroup.org/triads used for hand-curating the redundant hits found in the Protein Data Bank. Each row contained the PDB ID, the name of the entry, the chain where the triad was found, positions of the triad residues in the chain, a checkmark, the EC code found by automatic parsing from the SwissProt database, and the hand-inserted EC code, if the automatic parsing did not yield one.

Some PDB entries, mostly oligomers, contained more than one such triads (e.g., the octamer 1P8J contains eight copies of the triad). Clearly, the symmetric copies of the same triad in oligomers carry no new structural information, and need to be filtered out.

Similarly, there are numerous redundancies in the PDB itself: different entries exist in the PDB with minor differences: different entries solely with point-mutations or the same protein crystallized with different ligands (e.g., PDB entries 1XRP, 1XRQ, 1XRM, 1XRN contain the same triad-containing *proline iminopeptidase* protein, crystallized with different peptides, or PDB entries 1H22 and 1H23 contain different co-crystallized ligands in Table 1).

In order to getting rid of redundancies, we prepared the full version of Table 1, available at http://pitgroup.org/triads. In this table, the PDB ID, the name and the residue sequence are associated with the chain ID, the positions of the triad residues and with the EC numbers of the proteins, containing the triads.

The EC numbers of the PDB ID's were identified and inserted into the "E.C. number" column of the TRIAD table by fully automatic queries using the Swiss-Prot database (7). Some PDB ID's clearly have related EC numbers (e.g., crystallized parts of a larger enzyme), but they were not found in the Swiss-Prot database. In this case we inserted the code by hand in column "E.C. num. modified" in the table.

We were left with 350 non-redundant PDB entries, containing the triad.

Instead of selecting some arbitrary topologic or structural property of the triad, we constructed 12-dimensional vectors from the Cartesian coordinates of points *u*, *v, x* and *y* (Figure 1), relative to the origin *z*, simply as follows:

- the first three coordinates of the vector are the coordinates of vector *u*,
- the second three coordinates are those of *v*
- the third three coordinates are those of *x*,
- the fourth three coordinates are those of *y*.

Therefore, our vectors are generated in a much more straightforward way than those of (5). Every protein of the PDB corresponds to a point in the 12-dimensional space:

$$(u_1, u_2, u_3, v_1, v_2, v_3, x_1, x_2, x_3, y_1, y_2, y_3)$$

This is a personal *preprint version* of the paper, that is published in the Biochemical and Biophysical Research Communications in a more polished, peer-reviewed form. The DOI: doi:10.1016/j.bbrc.2009.04.022

Next, we applied a widely used, density-based clustering algorithm, the OPTICS algorithm (6) for these 350 points in the 12-dimensional space.

The OPTICS (*Ordering Points To Identify the Clustering Structure*) (6) is capable to visualize high-dimensional clusters in 2 dimensions by ordering the points and creating the *reachability plot* of the data.
The reachability plot is generated by assigning a value called *reachability distance* to all the objects of the database, while going through the database points in a specific order. The smaller is the reachability distance the closer is the point in question to the already visited points. The algorithm traverses the data points (the 12-dimensional vectors in our case), and a new, still unused point is visited going through densely populated regions (6), taking the smallest possible steps. For each scanned data point (corresponding to a point on the X axis on Figures 2 and 3) a reachability distance (given on the Y axis) is assigned that roughly speaking describes the local density in its neighborhood. Therefore the boundary of a cluster is depicted in the reachability plot of Figure 2 as a peak, while clusters are depicted as concave regions.
The OPTICS reachability plot contains a lot of the information about the clustering structure of the database, although it does not assign the objects to clusters. After the creation of the reachability plot, cluster membership assignments can be created by cutting the reachability plot with a horizontal line.

Figure 3: The concave regions show different clusters of proteins; the clustering were done
inset. The color code changes continuously from red (EC number = 1) to violet (EC numbe
by a minpts parameter of 3 of the OPTICS algorithm (6). The horizontal line suggests a po
cussed in the article

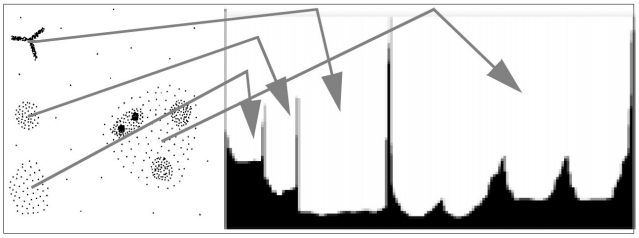
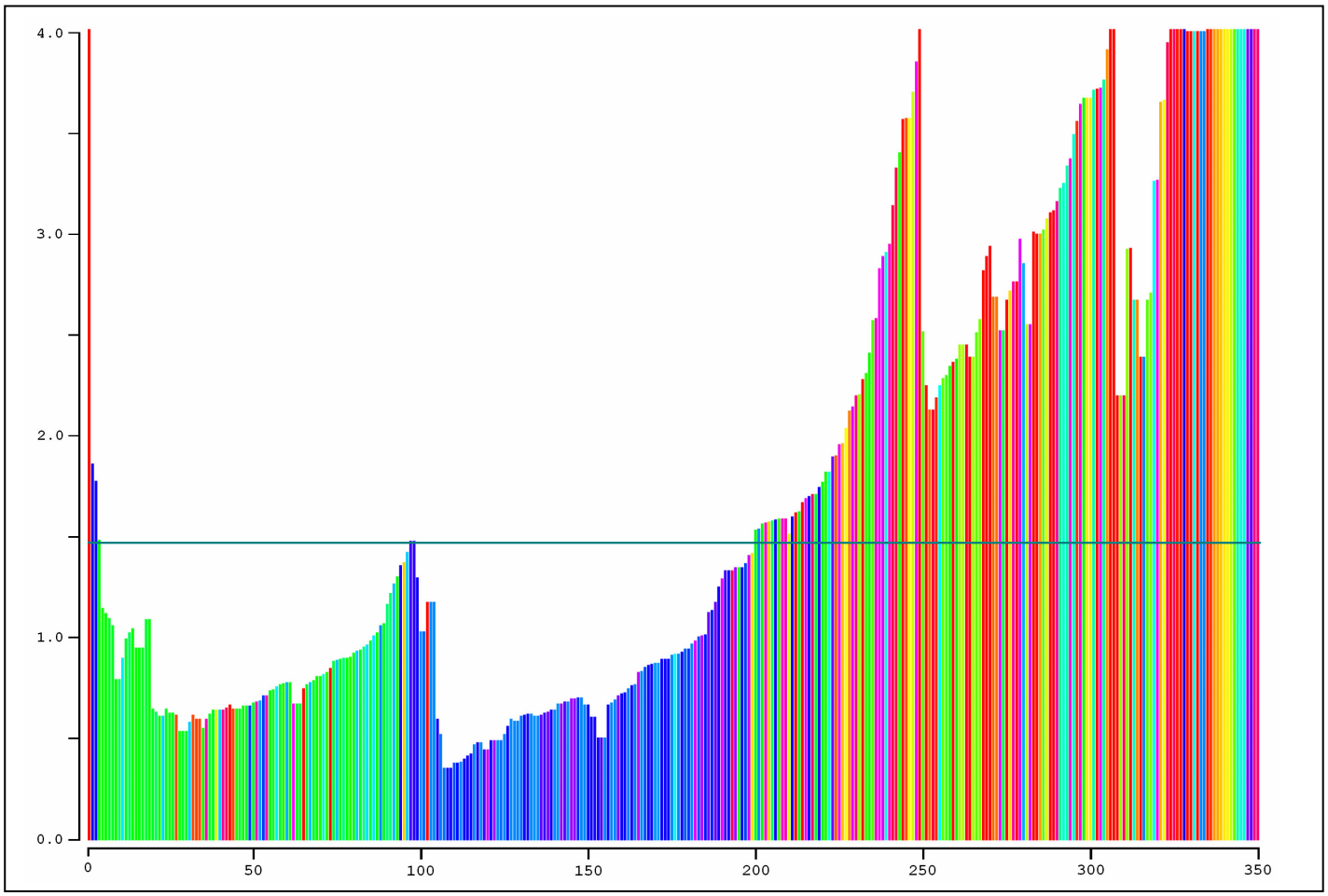


Figure 2: (source: (6)): On the left side a planar point set is given, on the right side the reachability plot is visualized. Deeper concave regions correspond to the denser clusters.

Our main result, Figure 3, is the reachability plot of the 350 data points, gained from the non-redundant triad-containing proteins from the Protein Data Bank. For visualizing the different EC numbers of the proteins, corresponding to the data points, we colored data bars continuously.



3 RESULTS AND DISCUSSION

By analyzing Figure 3, it is easy to identify the green cluster (between the items 5 and 95) and the blue cluster (from item no. 95 through 190).

Table 2 in the on-line supplementary material contains the PDB ID's , the EC numbers, the triad positions and also the colored data bars of Figure 3, in a six-page-long complete list. In order to facilitate the analysis of the enzyme-families of the clusters, we also included the SCOP ID's in Table 2 (8).

Using the Structural Classification of Proteins (SCOP), the two clusters are again strongly characteristic to protein superfamilies. The first cluster contains 61 structures with an alpha/beta hydrolase fold (cf. Table 2 in the on-line supporting material), while only *one* structure falls outside. The second cluster has 66 structures with trypsin-like and three ones with subtilisin-like fold, all but two of these (PDB codes: 1WXR and 1DXP) refer to serine proteases (EC 3.4.21). 15 trypsin-like structures are outside the cluster, for most of these the resolution is relatively low, indicating a lower precision of the structure determination. 29 and 27 structures fall within the alpha/beta hydrolase and trypsin-like fold clusters, respectively, for which no SCOP classification is available. We may suppose that these structures possess the same SCOP fold as all other ones within their cluster.

We conclude that after a thorough redundancy filtering, we recovered 350 entries from the PDB, containing the ASP-HIS-SER or GLU-HIS-SER triads. Next, using only four spatial points, denoted by *u, v, x* and *y* on Figure 1 and the highly capable OPTICS clustering algorithm (6), we clearly identified two clusters of triad-containing enzymes. Both clusters were validated by EC-number analysis though the color codes (Figure 3) and SCOP codes (Table 2 in the on-line supporting material).

ACKNOWLEDGEMENTS

Funding: The following grants are acknowledged: Hungarian National Research Fund (OTKA): NK67800, NI68466, NK67867, NKTH: TB-INTER, OMFB-01295/2006.

REFERENCES

(1) Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P.E.; *Nucleic Acids Research* **28**, 235-242 (2000).
(2) B. W. Matthews, P. Sigler, K. Henderson, and D. Blow, *Nature* **214,** 652–656 (1967).
(3) Dodson, G., and A. Wlodawer. 1998. Catalytic triads and their relatives. *Trends Biochem. Sci*., **23**:347-352.
(4) Gérczei, T., Asbóth, B., Náray-Szabó, G.: *J. Chem. Inf. Comput. Sci*. **39,** 310-315 (1999).
(5) Hamelryck, T.: *Proteins: Structure, Function, and Bioinformatics* Volume 51 Issue 1, Pages 96 - 108, 2003.
(6) M. Ankerst, M. M.Breunig, H. Kriegel, J. Sander, OPTICS: Ordering Points To identify the Clustering Structure, *Proc. ACM SIGMOD '99 Int. Conf. on Management of Data, Philadelphia PA,* (1999).
(7) B. Boeckmann et al.:The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. 31:365-370(2003).
(8) Murzin AG, Brenner SE, Hubbard T, Chothia C (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures". *J. Mol. Biol.* **247** (4): 536–40