Taylor & Francis
Taylor & Francis Group

Journal of Biomolecular Structure and Dynamics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbsd20

Functional identification of ‘hypothetical protein’ structures with unknown function

Y. N. Chirgadze, E. A. Boshkova, A. M. Kargatov & N. Y. Chirgadze

To cite this article: Y. N. Chirgadze, E. A. Boshkova, A. M. Kargatov & N. Y. Chirgadze (2022): Functional identification of ‘hypothetical protein’ structures with unknown function, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2022.2085806

To link to this article: https://doi.org/10.1080/07391102.2022.2085806

Published online: 11 Jun 2022.

JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS
https://doi.org/10.1080/07391102.2022.2085806

Taylor & Francis
Taylor & Francis Group

LETTER TO THE EDITOR

Functional identification of ‘hypothetical protein’ structures with unknown function

Y. N. Chirgadze[a], E. A. Boshkova[a], A. M. Kargatov[a] and N. Y. Chirgadze[b,c]

[a]Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, Russia; [b]Campbell Family Cancer Research Institute, Ontario Cancer Institute, Princess Margaret Hospital, University Health Network, Toronto, Ontario, Canada; [c]X-CHIP Technologies Inc, Toronto, Ontario, Canada

Communicated by Ramaswamy H. Sarma

ARTICLE HISTORY Received 3 February 2022; Accepted 30 May 2022

The work is devoted to 50 years of the Protein Data Bank –basic structural collection of biological macromolecules

1. Introduction

Problem of identification of ‘hypothetical protein’ has arisen about fifteen years ago after deciphering of DNA genomes of the man, animals and bacteria, and appearance of a number of proteins with known three-dimensional structure but unknown function. Since that time the different definitions have been used, such as ‘hypothetical protein’ structure (Mizohata et al., 2005; Hattori et al., 2005), protein with unknown function (Grotthuss et al., 2006), unannotated protein structure, uncharacterized protein, unidentified, and others. All these terms can be found in the Protein Data Bank. We have decided to use here term ‘hypothetical protein’ to attribute *the proteins with known spatial structure but unknown function.*

Suppose the sequence of protein is determined but spatial structure is unknown. Then we have no information about arrangement of its active center; no data about the protein association in solution during its functioning; etc. Thus, general term mentioned as protein with unknown function can be used in different cases provided all things become clear from definite context. As a rule, all proteins with known protein structure are deposited in Protein data Bank (Berman et al., 2002), while GenBank (Benson et al., 2013) also includes proteins for which only sequences are known. As a result, several thousands of such protein structures have been deposited in Protein Data Bank and almost ten times more sequences in GenBank. One of the first protein databases derived from the prediction of their functions was PDB-UF (Grotthuss et al., 2006). In fact, the problem of protein identification appeared at the birth time of molecular biology as a science, almost simultaneously with such important subdivided sciences as structural biology, protein crystallography, more later modern cryo-electron protein microscopy, etc. Protein structure classification is also an important part in this row, and it is also related to the current communication. The latter, however, does not include description and evaluation of efficiency of different genetic, structural, and various mathematical approaches to identification of ‘hypothetical proteins’. Our aim is to summarize the results of ten years of application of original method of identification of a number of unannotated proteins. It should be noted that appearance of this method is based on the known works of classification of protein structures (Chirgadze, 1987; Efimov, 1993, 1997; Murzin et al., 1995; Gordeev et al., 2010; Das & Orengo, 2016). Nearly thirty databases and various corresponding methods of protein annotation have been recently reviewed in Das & Orengo, 2016.

In simple cases the problem is solved by comparing the sequences with those of well-known homologous proteins. However, such direct approach can be applied only to proteins with average sequence identity of more than 30-40%. In the cases of low homology with the known structures, one can use only original methods based on structural homology of some unique part of the protein structures. During last ten years, we have been involved in the identification of several ‘hypothetical protein’ structures. Most of studied here proteins are enzymes, and this has been used in the assignment of the proteins to the corresponding class of the structure. In this communication, we present a short description of the original method of structure identification of such ‘hypothetical protein’. This approach has been applied to four protein structures with unknown function. As a results, 14 new protein subfamilies have been introduced which includes a total of 496 newly identified protein structures and amino acid sequences.

2. Description of method and data

The proposed method of the identification is based on conservation of the residues related to functionally important regions of protein structure. So, we call it ‘functional identification’. It includes four stages as follows:

- search the amino acid sequence homology of the ‘hypothetical protein’ against annotated proteins;

CONTACT Y. N. Chirgadze chir@vega.protres.ru Russian Academy of Sciences, Institute of Protein Research, Pushchino, Moscow Region 142290, Russia

Table 1. List of annotated 'hypothetical protein' structures considered by the authors.

Identified protein	Source of protein	Sequence identity, %	Number of proteins	References
Alkyl hydroperoxidase D PA0269	*Pseudomonas aeruginosa*	9	5 5 subfamilies	Clarke et al., 2011
CN-hydrolase SA0302	*Staphylococcus aureus*	9	11 1 subfamily	Gordon et al., 2013
Ribosome-associated protein SAV1646	*Staphylococcus aureus*	25-90	13 1 subfamily	Chirgadze et al., 2015
Zn-glyoxalase I SA0856	*Staphylococcus aureus*	27	7 1 subfamily	Chirgadze et al., 2018
Proteins from family glyoxalase I	Bacteria, and *Homo sapiens*	23-47	460 6 subfamilies	Kargatov et al., 2018

Total identified: 496 'hypothetical protein'.
14 new subfamilies.

- select the homologous sequence fragments related to a function of 'hypothetical protein';
- define the structural homology of 'hypothetical protein' with well-identified proteins;
- identify 'hypothetical protein' structure and assign it to a class of protein structure.

Each stage is carried out with the help of well-known procedures. At the first stage it is performed with the help of the protein sequence alignment based on the software package BLAST (Altschul et al., 1997). We have paid here more attention to describe second stage of identification of 'hypothetical protein' structure. At this stage several functional and structural criteria has been used as well. Application of these criteria are important, particularly, in the case of low sequence homology of considered protein with other proteins. When we analyze enzymes, the positions of invariant amino acid residues or short peptide fragments can be found in the active center region. In this case we have used criterion based on the type of enzymatic activity and the corresponding ***sequence signature invariants***. Another criterion can be a ***family specific structural invariants***. For example, a protein molecule exists as a dimer. In this case, there are specific inter-domain residue contacts where these residues are conserved. We have faced with both types of invariants. The third stage of identification is to analyze the structural homology of 'hypothetical protein' in the group of homological proteins in order to define class of protein structure. The last stage defines the correct place of 'hypothetical protein' in the family hierarchy. Selection of the group of homologous proteins in the subfamily faces no problem in the case when at least one protein is well determined. We define this protein as ***representative*** for the subfamily. On this basis we assign other 'hypothetical protein' from GenBank.

In some cases, we have faced with introduction of new subfamilies. Protein subfamily is a level of classification of protein groups based on their close evolutionary relationship, and a number of other features, in particularly, topology of protein structure. At present time there are more than 3,000 families and many more subfamilies. The subfamilies are determined on the basis of variability, such as sequence homology, structural motif related to active site, phylogenetic profiles, etc. There are about thirty such approaches and corresponding data bases described in review by Das & Orengo, 2016. Most reliable of them are suggested to be based on the experimental approaches. Our results are based on the experimental functional data, sequence and structural data as well. Description of our method in details and its application are presented in the references (Chirgadze et al., 2018; Kargatov et al., 2018).

3. Examples of annotated 'hypothetical protein' structures

Thus, we can identify the considered protein. If required, one can introduce a new additional subfamily which allow to correct and expand the classification. To illustrate this, we present the results in Table 1 which have been published in (Clarke et al., 2011; Gordon et al., 2013; Chirgadze et al., 2015; Chirgadze et al., 2018; Kargatov et al., 2018). Homology of residues in the region of enzyme proteins active center of other homologous proteins have been used. The third protein SAV1646 is not an enzyme. In this case we have used the surface conservative residues which are necessary for functioning of protein dimer (Chirgadze et al., 2015). Below we consider examples of functional identification for presented proteins.

3.1. Identification of protein PA0269 from Pseudomonas aeruginosa

For this protein several other homologous proteins from Protein Data Bank have been found with rather low homologies corresponding to amino acid sequence identity of nearly 9%. The protein structure consists of ten helical fragments. However, the homology was concentrated inside of a limited part which contains two α-helices and includes 27 residues. Among the homologs there was only one annotated as alkyl hydroperoxidase D from *Mycobacterium tuberculosis* which involved in antioxidant defense mechanism. The structures of both proteins are shown in Figure 1. It explains the extremely low homology of 'hypothetical protein' PA0269 from *Pseudomonas aeruginosa* because of transposition of catalytically related α-helical hairpin in structure of identified protein PA0269. This protein and five other newly identified proteins belong to new different subfamilies (Clarke et al., 2011).

3.2. Identification of protein SA0856 from Staphylococcus aureus

Initially, protein SA0856 has been considered as a hypothetical gene product from *Staphylococcus aureus,* and it has been later identified as Zn-glyoxalase I (Chirgadze et al.,

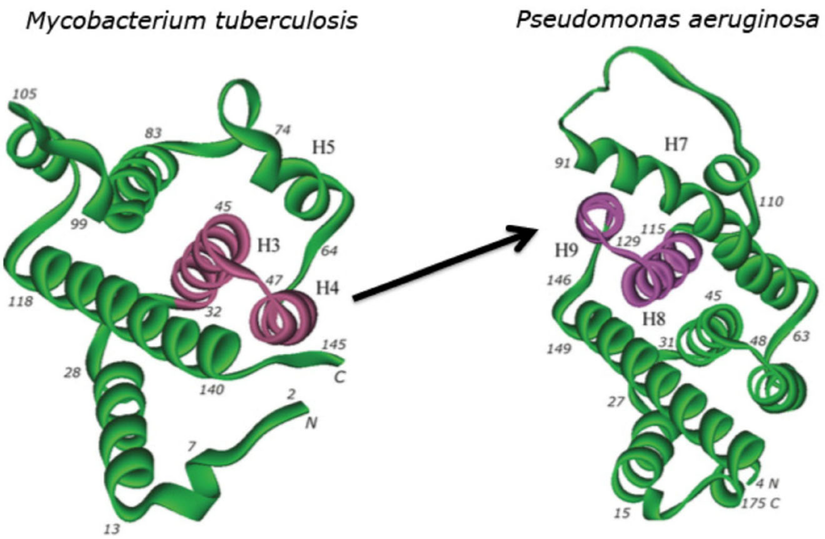

Figure 1. Transposition of catalytically related α-helical hairpin in the structure of newly identified protein PA0269 from *Pseudomonas aeruginosa*. Other details are given in Clarke et al., 2011.

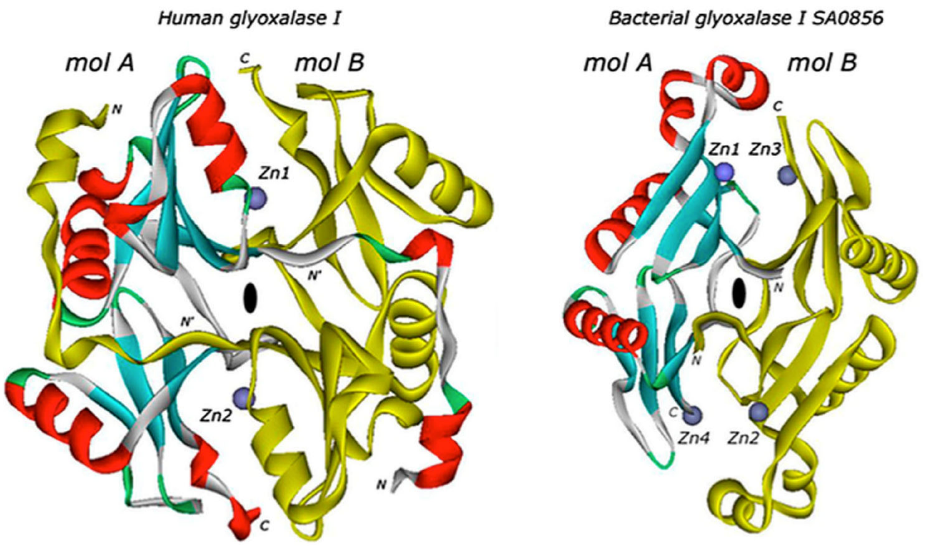

Figure 2. Representative crystal structures of subfamily A (left) and new subfamily B (right) for Zn-glyoxalase I. Larger size of human glyoxalase is explained by extra N-terminal peptide NN′.

2018). The protein structure is a dimer which forms two cavities for the active sites. Sequence analysis of other found homologous proteins showed that the sequences can be divided in two separate subfamilies. Subfamily A is based on the representative structure Zn-glyoxalase I from *Homo sapiens* (Cameron et al.,1997) and includes enzymes from various organisms from bacteria to plants and mammals. Among other sequences we can select new subfamily B on the bases of annotated protein SA0856 from *Staphylococcus aureus*. This subfamily includes proteins from bacteria only. The structures of representative proteins for two subfamilies are shown in Figure 2. Active sites of both enzymes are formed with two single protein units associated in dimer. In subfamily A active center includes one zinc ion but in subfamily B two ions. The key difference of the sequences, the so-called ‘signature’, is directly related to the conservative catalytic residues around metal ion of the active sites, and also with the contacting residues of monomers in the dimer.

3.3. Identification of ‘hypothetical proteins’ in glyoxalase I family

Glyoxalase I (S-D-lactoyl glutathione lyase) is one of two enzymes of the glyoxalase detoxification system acting against methyl glyoxal and other aldehydes, which are the metabolites derived from glycolysis. The glyoxalase system is very common and available almost in all living organisms: bacteria, protozoa, plants, and animals, including humans. It is related to the class of ‘life essential proteins’. The enzyme glyoxalase I belongs to the expanded *Glyoxalase/Bleomycin resistance protein/Dioxygenase* superfamily. To date the

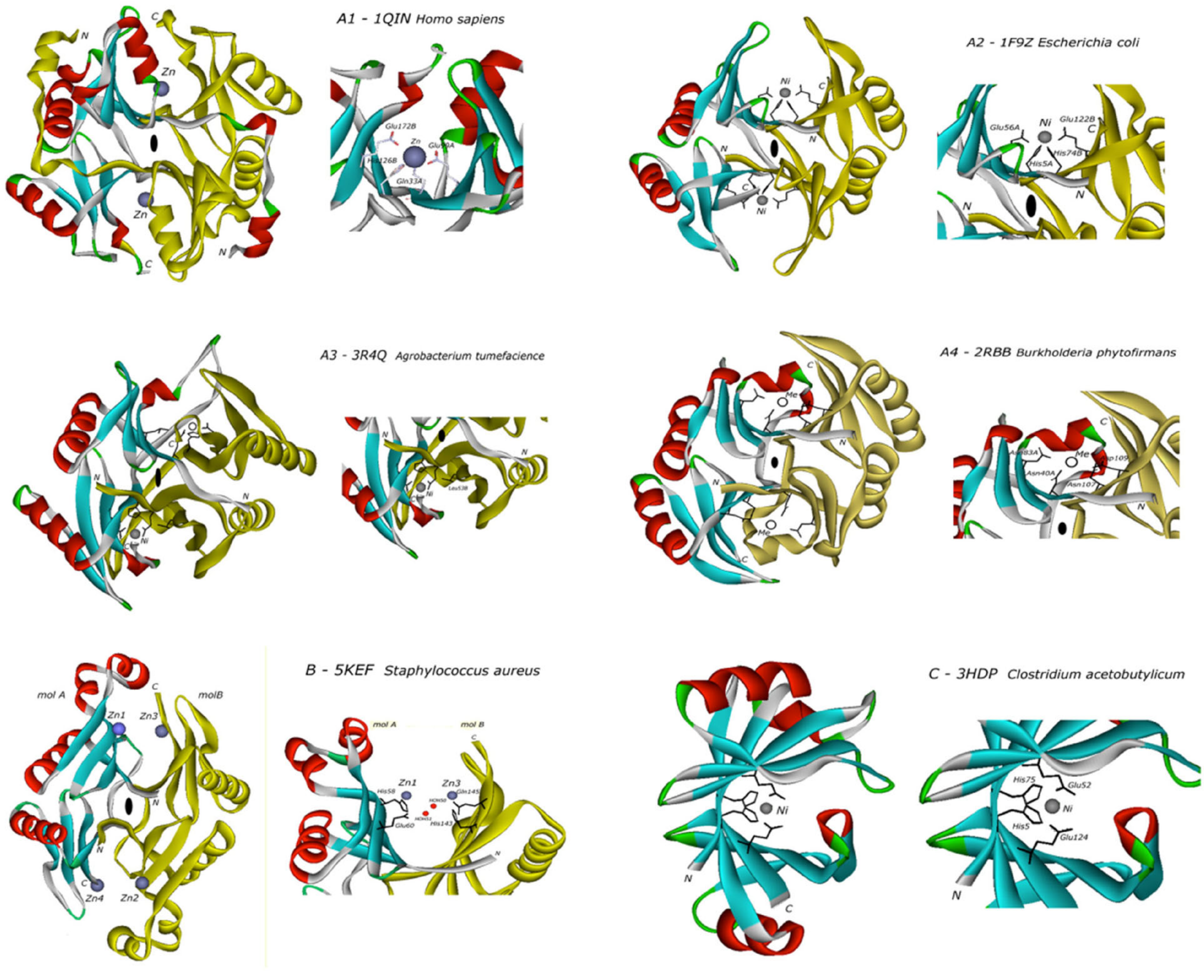
A1 - 1QIN Homo sapiens
A2 - 1F9Z Escherichia coli
A3 - 3R4Q Agrobacterium tumefacience
A4 - 2RBB Burkholderia phytofirmans
B - 5KEF Staphylococcus aureus
C - 3HDP Clostridium acetobutylicum
Zn
Ni
Me
mol A
mol B
Zn1
Zn3
Zn4
Zn2
N
C
Glu172B
Gln33A
Glu56A
Glu122B
His74B
His5A
Asn40A
Asn107
His58
Glu60
Gln145
His143
His75
Glu52
His5
Glu124

Figure 3. Structures of representative proteins of six new subfamilies of the glyoxalase I family. The active site cavity can contain one or two different metal ions. The data of the protein for subfamily A4 is available only as apo-enzyme, and possible locations of metal ions are shown here as white circles. The data taken from Kargatov et al., Chirgadze et al., 2018.

GenBank contains about seven hundreds amino acid sequences of this enzyme, and the Protein Data Bank includes nearly thirty spatial structures. Carrying out functional identification of Zn-glyoxalase I (Chirgadze et al., 2018) we have observed a significant diversity of homology. In particulary, it was connected with the structure of active site cavity. For example, sometime active site includes one ion metal atom, while in most subfamilies the active site contains two ion atoms. It impels us to expand the classification and to introduce a few novel subfamilies (Kargatov et al., 2018). Thus, it was applied for analysis of all set of 'hypothetical proteins' of glyoxalase I available at the moment in GenBank and Protein Data Bank. As a result, six new subfamilies A1, A2, A3, A4, B, and C of glyoxalase I have been disclosed, and a total of 460 'hypothetical proteins' have been identified and classified. The pair sequence identities in these subfamilies were ranged from 20 to 47% (Table 1). The differences of the representative structures for these subfamilies are shown in Figure 3, data taken from Protein Data Bank. Other details are given in (Kargatov et al., 2018).

4. Conclusion

In this communication we have presented the results for several 'hypothetical proteins' using an approach of functional identification. The results show how the novel annotated data can be obtained for the 'hypothetical proteins' already available in Protein Data Bank and GenBank. Generally, two main principles are the keys for use of the method. First is the sequence homology of given protein with other annotated proteins along the whole or only part of the protein chain. And the second key is a specific conservative sequence 'signature' connected with the structure of the active site. Finally, the important steps should be carried out: an analysis of the residues' arrangement in catalytic center and experimental test of functional identity. Those have been done for considered proteins as seen from the original referenced papers. Thus, in the case of expanded family glyoxalase I, as an example, we have identified nearly 460 'hypothetical proteins', from about seven hundreds in GenBank, and introduce six new unknown earlier subfamilies.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Res*, *41*, 36–42.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., & Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallographica. Section D, Biological Crystallography*, *58* (Pt 6 No 1), 899–907. https://doi.org/10.1107/s0907444902003451

Cameron, A. D., Olin, B., Ridderstrom, M., Mannervik, B., & Jones, T. A. (1997). Crystal structure of human glyoxalase I – Evidence for gene duplication and 3D domain swapping. *The EMBO Journal*, *16*(12), 3386–3395.

Chirgadze, Y. (1987). Deduction and systematic classification of spatial motifs of the antiparallel β-structure in globular proteins. *Acta Cryst*, *A43*, 405–417.

Chirgadze, Y. N., Boshkova, E. A., Battaile, K. P., Mendes, V. G., Lam, R., Chan, T. S. Y., Romanov, V., Pai, E. F., & Chirgadze, N. Y. (2018). Crystal structure of *Staphylococcus aureus* Zn-glyoxalase I: New subfamily of glyoxalase I family. *Journal of Biomolecular Structure and Dynamics*, *36*(2), 376–386. https://doi.org/10.1080/07391102.2016.1278038

Chirgadze, Y. N., Clarke, T. E., Romanov, V., Kisselman, G., Wu-Brown, J., Soloveychik, M., Chan, T. S. Y., Gordon, R. D., Battaile, K. P., Pai, E. F., & Chirgadze, N. Y. (2015). The structure of SAV1646 from *Staphylococcus aureus* belonging to a new 'ribosome-associated' subfamily of bacterial proteins. *Acta Crystallographica Section D: Biological Crystallography*, *D71*, 332–337.

Clarke, T. E., Romanov, V., Chirgadze, Y. N., Klomsiri, C., Kisselman, G., Wu-Brown, J., Poole, L. B., Pai, E. F., & Chirgadze, N. Y. (2011). Crystal structure of alkyl hydroperoxidase D like protein PA0269 from *Pseudomonas aeruginosa*: Homology of the AhpD-like structural family. *BMC Structural Biology*, *11*, 27. https://doi.org/10.1186/1472-6807-11-27

Das, S., & Orengo, C. A. (2016). Protein function annotation using protein domain family resources. *Methods (San Diego, California)*, *93*, 24–34. https://doi.org/10.1016/j.ymeth.2015.09.029

Efimov, A. V. (1993). Standard structures in proteins. *Progress in Biophysics and Molecular Biology*, *60*(3), 201–239. https://doi.org/10.1016/0079-6107(93)90015-C

Efimov, A. V. (1997). Structural trees for protein superfamilies. *Proteins: Structure, Function, and Genetics*, *28*(2), 241–261. https://doi.org/10.1002/(SICI)1097-0134(199706)28:2<241::AID-PROT12>3.0.CO;2-I

Gordeev, A. B., Kargatov, A. M., & Efimov, A. V. (2010). PCBOST: Protein classification based on structural trees. *Biochemical and Biophysical Research Communications*, *397*(3), 470–471. https://doi.org/10.1016/j.bbrc.2010.05.136

Gordon, R. D., Qiu, W., Romanov, V., Lam, K., Soloveychik, M., Benetteraj, D., Battaile, K., Chirgadze Yu, N., Pai, E., & Chirgadze, N. Y. (2013). Crystal structure of the CN-hydrolase SA0302 from the pathogenic bacterium *Staphylococcus aureus* belonging to the Nit and NitFhit Branch of the nitrilase superfamily. *Journal of Biomolecular Structure & Dynamics*, *31*(10), 1057–1065. https://doi.org/10.1080/07391102.2012.719111

Grotthuss, M., Plewczynski, D., Ginalski, K., Rychlewski, L., & Shakhnovich, E. I. (2006). PDB-UF: Database of predicted enzymatic functions for unannotated protein structures from structural genomics. *BMC Bioinformatics*, *7*, 53.

Hattori, M., Mizohata, E., Manzoku, M., Bessho, Y., Murayama, K., Terada, K., Kuramitsu, S., Shirouzu, M., & Yokoyama, S. (2005). Crystal structure of the hypothetical protein TTHA1013 from *Thermus thermophilus* HB8. *Proteins*, *61*(4), 1117–1120. https://doi.org/10.1002/prot.20692

Kargatov, A. M., Boshkova, E. A., & Chirgadze, Y. N. (2018). Novel approach for structural identification of protein family: Glyoxalase I. *Journal of Biomolecular Structure & Dynamics*, *36*(10), 2699–2712. https://doi.org/10.1080/07391102.2017.1367330

Mizohata, E., Hattori, M., Kuramitsu, S., Shirouzu, M., & Yokoyama, S. (2005). Crystal structure of hypothetical protein TTHA1013 from an extremely thermophilic bacterium Thermus thermophilus HB8. *The Protein Data Bank,* PDB ID 1wv8. https://doi.org/10.2210/pdb1wv8/pdb

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, *247*(4), 536–540.