http://dx.doi.org/10.17202/JUSO.2015.2.1 *Journal of Universal Science Vol* ***2(1):*** *1-9, 2015*

Journal of Universal Science Online | WWW.JUSO.HU | ORIGINAL PAPER ISSN 2416-0008

# Classification of glycosyl hydrolases based on structural homology

**László FÜLÖP[1,*] and Tamás PONYI[1]**

[1]Department of Chemistry and Biochemistry, Szent István University, Páter K.u.1., Gödöllő, H-2103 Hungary

[*]Corresponding Author; e-mail: lesley.fulop@gmail.com

*Received: 2 August 2015 / Revised: 18 August 2015 / Accepted: 1 September 2015*

Keywords: protein structure, sequence, glycosyl hydrolases, cluster, similarity

Abbreviations**:**

BASH – Bourne-Again Shell
CAZY - Carbohydrate-Active enZymes
E.C. – Enzyme Commission
FSSP – Dali Fold Classification
GH – Glycosyl Hydrolase
ICGEB - International Centre for Genetic Engineering and Biotechnology
PDB ID – Protein Data Bank ID
PRIDE – PRobability of IDEntity

## ABSTRACT

Glycosyl hydrolases are a well-known group of enzymes, which hydrolyze the glycosidic bond between carbohydrates, or between a carbohydrate and different molecules. Glycosyl hydrolases play a vital role in the human body, and are widely used in industrial applications. Glycosyl hydrolases classification is based on substrate specificity and amino acid or nucleotide sequence similarity which reflects their evolutionary relationship.

Our aim, in this study, was to carry out the classification of glycosyl hydrolases, based solely on structural similarity which was made possible by the several structures available in the databases and the availability of computing power to conduct such a computationally intensive task, in a reasonable time-frame. It was also aimed that the structural similarity based classification be compared to the present classification system.

Based on an all-against-all comparison, we conducted a structural comparison of glycosyl hydrolases. The results are presented graphically. The graphical representation defined 24 structurally homologous classes. The classification was validated using $C_\alpha$ - $C_\alpha$ distance analysis and amino acid sequence cluster analysis.

Advantages of this method are that – being an automated method – it is fast, simple and reproducible. Glycosyl hydrolases could be classified into 24 separate classes. N-glycosyl and O-glycosyl hydrolases (both forming binding and catalytic domain classes as well) were clearly different, the former consisting of 8 classes, and the latter consisting of 16 classes. Structural classes simplified the previous classification system. This classification represents the current glycosyl hydrolase family system, but also extends it especially concerning the clan system.

Journal of Universal Science Online
WWW.JUSO.HU
ORIGINAL PAPER
ISSN: 2416-0008

## Introduction

Glycosyl hydrolases (GH) are a well-known and intensively researched group of enzymes, which hydrolyze the glycosidic bond between carbohydrates, or between a carbohydrate and different molecules. Carbohydrates are diverse molecules and participate in many biochemical pathways, and so are glycosyl hydrolases. Therefore their deficiency can cause serious, and often inheritable disorders, like, for example, lactose intolerance [1] or Fabry disease [2]. Apart from the human medical research purposes, these enzymes are utilized widely in biomass degradation, [3] bioethanol synthesis [4], waste processing [5] and pest control [6], to name but a few applications.

Glycosyl hydrolases belong to the E.C. 3.2. group, and are currently categorized into about 160 entries, based primarily on their substrate specificity [7]. Another classification system is based on sequence similarities, hydrophobic cluster analysis and catalytic mechanisms [8, 9], and assigns glycosyl hydrolases into GH families [10]. However, enzymes with different substrate specificities are sometimes found in the same family and enzymes that hydrolase the same substrate are sometimes found in different families. [11] The number of families in this classification system has already grown beyond 100. Based on folding similarities [12], about every other family is assigned to either of the clans.

When the above classification was created, 3-dimensional structures [13] and structural comparison methods barely existed. Shortly after this, to complement the existing systems, a classification, based on folding and structure similarities, was suggested [14]. However, neither the computational performance nor the numbers of known structures nor structural comparison methods were sufficient to develop such a system. In recent years, however, conditions became appropriate to classify macromolecules according to their structure [15]. One such system is Dali Fold classification (FSSP), which is based on an "all-against-all" comparison of the PDB (Protein Data Bank) database [16]. The major drawback of this system is that its classification relies on PDB90, a representative subset of PDB excluding structures that share more than 90 % of sequence similarity, and only the top 20 results are shown.

In this paper we describe the structural comparison of glycosyl hydrolases utilizing an all-against-all approach of domain 3D structure comparison. The results of the analysis were represented graphically. Using hierarchical cluster analysis [17, 18] and graphical representation, structurally homologous classes were identified. This classification was validated by amino acid sequence cluster [19, 20] (ClustalW-Protdist) and $C_\alpha$ - $C_\alpha$ distance analysis [21] (PRIDE Cluster), and contrasted to the existing classifications.

## Material and methods

*Structures*

The glycosyl hydrolase protein structures were downloaded from the Brookhaven Protein Data Bank [16, 22]. Every structure that contained the words MUTANT or COMPLEX anywhere in the PDB-file (exceptions to REMARK lines) were automatically removed from the comparison by a custom made BASH script.

*Structural comparison by DaliLite*

A total of 789 domain structures from 498 PDB entries were compared. DaliLite [17, 23] v2.1 was used for the extraction and the structural comparison of domains in list mode (one-against-all). Every single structure was compared to all the other available structures. Output files were transformed to HTML by DaliLite, and processed by different custom made BASH scripts. A 2-dimensional matrix,

Journal of Universal Science Online
WWW.JUSO.HU
ORIGINAL PAPER
ISSN: 2416-0008

with the compared domain structures on both axes and the normalized similarity indices in the intersections, was constructed from the output data. Normalization was required because the value of similarity indices depend not only on the level of homology, but also on the size of the compared structures. Normalization was applied by transforming the value of similarity indices (Z-scores) to percentage, the index value of self-comparison (comparing the structure to itself, total identity) being 100 % in each line (1 line = 1 structure) of the matrix.

Together with the construction of the matrix, a supporting chart was also created automatically, which contained the domain ID (derived from the PDB name), sequence length, E.C. number, CATH [24, 25] ID, CAZY [26, 27] family number and clan ID, name (e.g. endo-1,4-beta-xylanase), source organism and secondary structure composition of the structure. Occasional corrections were made manually.

On the matrix, a hierarchical cluster analysis was conducted using the Cluster program [17] utilizing the Kendall's Tau algorithm [28], then the cluster was refined and corrected manually. Only full lines or columns were moved at once, to preserve the relative position of structures. The classes were already revealed by the cluster analysis; only less homologous parts needed some manual intervention. The order of the classes was also changed according to the nomenclature. Then the supporting chart was sorted according to the matrix.

*Structural comparison by PRIDE Cluster*

The PRIDE algorithm calculates structure similarity based on $C_\alpha$ - $C_\alpha$ distances. Therefore the results are based only on the atomic coordinates of $C_\alpha$ atoms; amino-acids sequence, secondary structure content, or the topology of secondary structural elements are disregarded. PRIDE Cluster makes a classification of several structures.

The same structure database was used as for the comparison with DaliLite, in the order of the sorted DaliLite matrix. The coordinates of the $C_\alpha$ atoms were extracted and stored in a separate document by a custom made BASH script, and run on the PRIDE Cluster. The results were presented in the same way as with using DaliLite.

*Amino acid sequence comparison by ClustalW and ProtDist*

In order to identify amino acid sequence clusters, multiple alignment of domain sequences were conducted, using ClustalW v1.83. The same structure database was used as for the DaliLite comprison, in the order of the sorted DaliLite matrix. The sequences of the proteins were extracted and stored in a separate document by a custom made BASH script, then were multiply aligned using ClustalW. From the multiple alignment, a protein distance matrix was produced by ProtDist, utilizing the Jones-Taylor-Thornton model. The results were presented in the same way as of DaliLite.

## Results and discussions

*Structural comparison by DaliLite*

A total of 789 domain structures from 498 PDB entries were processed. Six hundred and eighty-one O-glycosyl (E.C. 3.2.1.x), 93 N-glycosyl (3.2.2.x) and 15 S-glycosyl hydrolase (E.C. 3.2.3.x, now E.C. 3.2.1.147, see also the Enzyme Nomenclature, 1992) structures were compared. The result of the all-against all comparison is a 2-dimensional matrix, converted to percentage and represented graphically (1 value = 1 pixel). The clustered matrix showed distinct nodes (classes) on the graph. The clustered matrix showed distinct structural classes of different size (black rectangles), as well as similarities

Journal of Universal Science Online
WWW.JUSO.HU
ORIGINAL PAPER
ISSN: 2416-0008

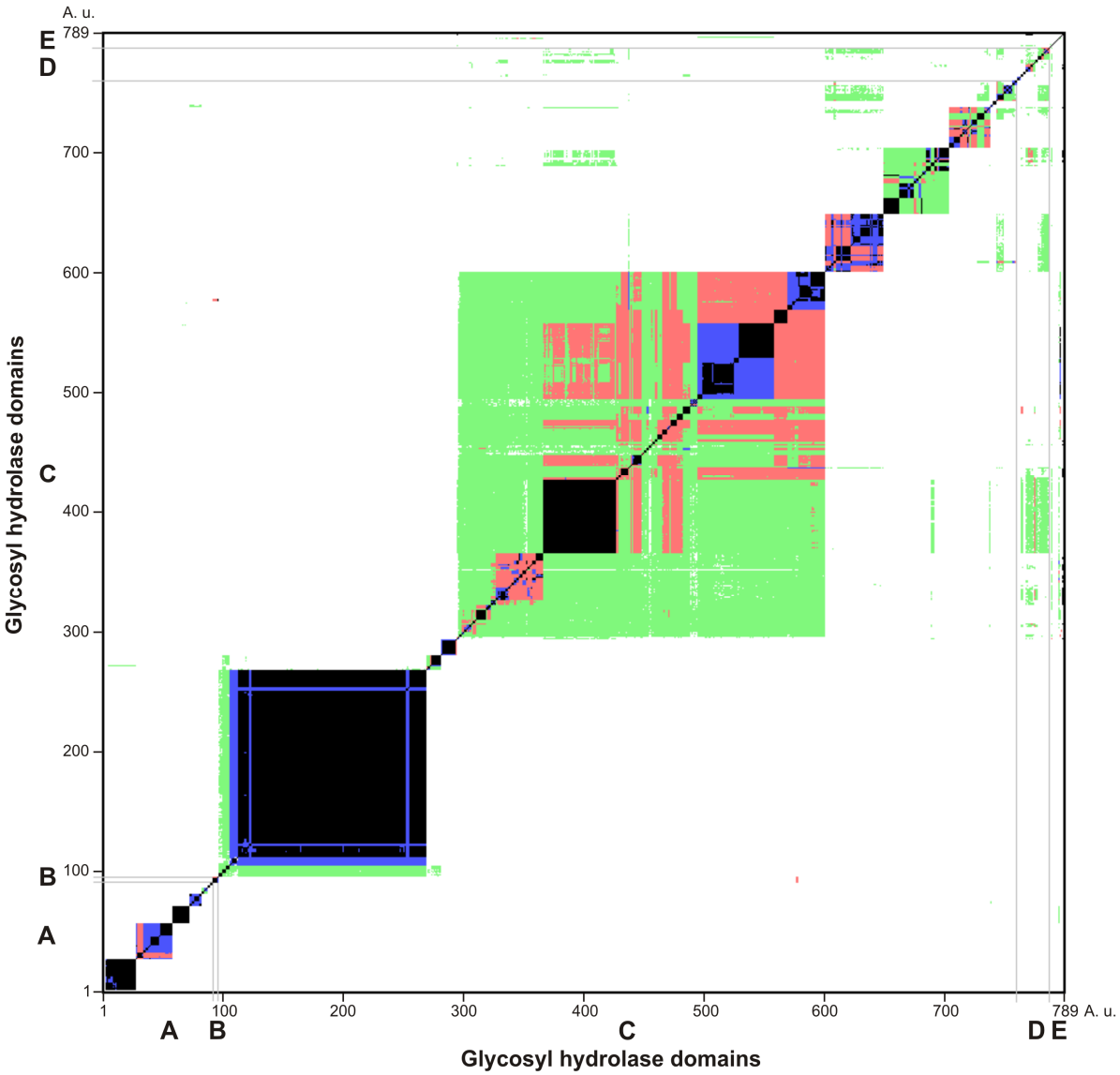


**Fig. 1.** Image of the structural similarity matrix of glycosyl hydrolases, according to DaliLite. Each pixel represents one pair wise comparison.
*Similarity*: white: 0 - 8 %, green: 9 - 25 %, orange: 26 - 50 %, blue: 51 - 75 %, black: 76 - 100 %. *Regions*: A: N-Glycosyl hydrolase catalytic domains, B: N-Glycosyl hydrolase binding domains, C: O-(S)-Glycosyl hydrolase catalytic domains, D: O-(S)-Glycosyl hydrolase binding domains, E: Unclassified domains.

between some of the classes (black dots/lines/boxes outside the rectangles) (Fig. 1). Rectangles are domains that can be considered as structurally homologous groups, because of their similarity in 3D structure. In some of the classes, a number of subclasses could be obtained. Domains in the same subclass usually show at least 50 %, or higher structural similarity value. The size of the rectangles, only represent the number of domains available for analysis. The symmetry anomaly seen in the upper and lower part of Fig. 1 is due to the nature of structural comparison of DaliLite, because the Z-score (similarity index) value is different if one compares structure 'A' to structure 'B' or structure 'B' to structure 'A'. Classes were numbered separately for N- and O-glycosyl hydrolases. The classes were set against CATH, CAZY, and E.C. Both O-glycosyl and N-glycosyl hydrolases formed separate classes, while S-glycosyl hydrolases were indistinguishable in one of the O-glycosyl hydrolase classes. Less than 2 percent of the

Journal of Universal Science Online
WWW.JUSO.HU
ORIGINAL PAPER
ISSN: 2416-0008

domains did not fit into any class. Apart from one sole domain, O-glycosyl and N-glycosyl hydrolases were not mixed. As a matter of fact, O- and N-glycosyl hydrolase classes barely show any, even distant structural similarity to each other, according to DaliLite. Altogether, from the 622521 (789*789) pair wise comparisons, we had 152253 results, which means 24.45 % of the compared pairs showed some level of structural homology.

*Structural classes and subclasses*

The summary of the structurally homologous classes is presented in Table 1. In the E.C., CATH, CAZY FAMILY and CAZY CLAN columns of Table 1, each and every entry present in the class is shown, regardless of whether there was only a single representative, or multiple instances. The classes are shown in the same order as in Fig. 1.

N-glycosyl hydrolases formed 8 classes altogether, 7 of them contained catalytic domains (88 domains, designated region **A**), and 1 contained binding domains (5 domains, designated region **B**). N-glycosyl catalytic domains seem to be quite uniform within their class, only 2 of the 7 classes contained multiple E.C. numbers (i.e. were polyspecific). Concerning CATH, 2 classes contained multiple CATH entries. No CAZY entries were present in the catalytic domain classes. The sole binding domain class contained multiple E.C. and CATH entries, as well as a CAZY entry.

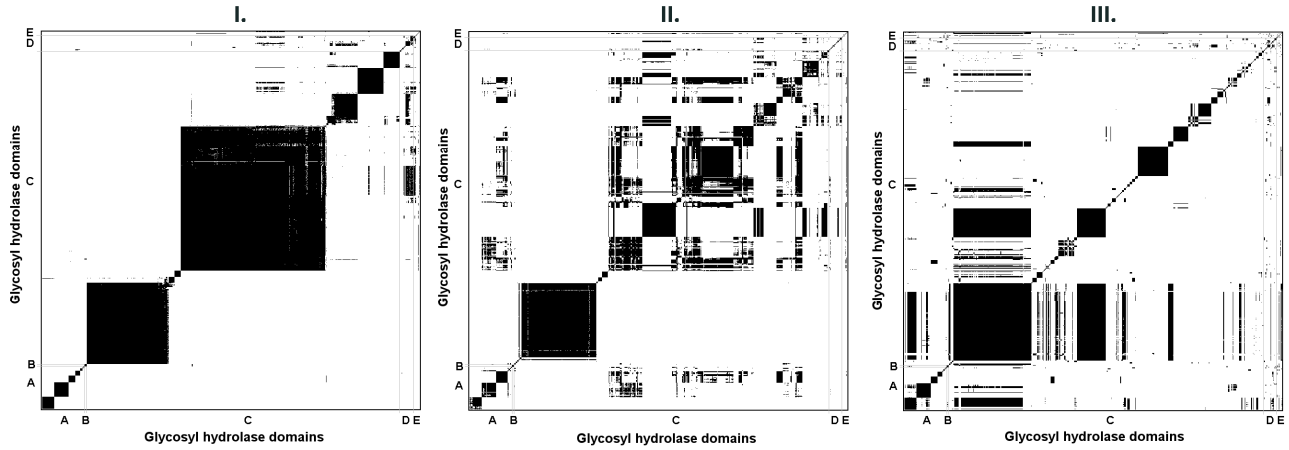


**Fig. 2.** Similarity matrices for I. DaliLite, II. Pride Cluster and III. ClustalW/ProtDist.
Each pixel represents one pair wise comparison.
*I. Similarity*: white: 0 - 8 %, black: 9 - 100 %. *II. Similarity*: white: 0 - 80 %, black: 81- 100 %. *III. Difference*: white: > 4 , black: 0 – 4.

O-glycosyl hydrolases formed 16 classes altogether, 10 of them contained catalytic domains (657 domains, designated region **C**), and 6 contained binding domains (27 domains, designated region **D**). S-glycosyl hydrolases (E.C. 3.2.3.1, now E.C. 3.2.1.147) were only present in class C4. O-glycosyl catalytic domains were more diverse within a class than their N-glycosyl counterparts. Apart from the C3 class (which contained a sole entry from E.C., CATH, CAZY and CAZY CLAN, respectively), only 2 of the 10 classes were monospecific, and 7 others contained multiple entries from both CATH and CAZY, 4 of them even containing multiple CAZY CLANs. Binding domains were more uniform (probably because of the low number of domains), 3 of the 6 classes contained a sole entry from E.C. and CAZY, 3 classes contained a sole entry from CATH.

The remaining domains (12 domains, designated region **E**) could not be assigned to any classes above. These domains were highly diverse concerning E.C., CATH and CAZY.

Journal of Universal Science Online WWW.JUSO.HU ORIGINAL PAPER ISSN: 2416-0008

Despite the low number of N-glycosyl hydrolases, almost the same number of catalytic domain classes could be obtained as O-glycosyl hydrolases. Unlike catalytic domains, N-glycosyl binding domains only formed 1 class, probably because of the extremely small number of binding domains. Domains from the same E.C. entry often appeared in different classes, sometimes even from the same PDB file.

**Table 1.** Summary of the structurally homologous classes of glycosyl hydrolases.
The domain, E.C., CATH, CAZY FAM. and CAZY CLAN cells show how many different entries could be assigned to the given classes.

| Domains | Class | Domains | E.C. | CATH | CAZY FAM. | CAZY CLAN | Representative enzyme |
|---|---|---|---|---|---|---|---|
| N-glycosyl catalytic (***A***) | 1 | 25 | 1 | 1 | - | - | Shiga Toxin A Subunit |
| | 2 | 30 | 1 | 4 | - | - | Ricin A Chain |
| | 3 | 14 | 1 | 1 | - | - | ADP-Ribosyl Cyclase |
| | 4 | 10 | 2 | 1 | - | - | Pyrimidine Nucleoside Hydrolase |
| | 5 | 5 | 2 | 2 | - | - | 3-Methyladenine DNA Glycosylase |
| | 6 | 2 | 1 | 1 | - | - | Uracil-DNA Glycosylase |
| | 7 | 2 | 1 | 1 | - | - | MTA/SAH Nucleosidase |
| N-glycosyl binding (***B***) | 8 | 5 | 2 | 2 | 1 | - | Endo-1,4-beta-Xylanase A |
| O-(S)-glycosyl catalytic (***C***) | 1 | 171 | 2 | 2 | 3 | - | Lysozyme |
| | 2 | 12 | 2 | 4 | 2 | 1 | Chitosanase |
| | 3 | 13 | 1 | 1 | 1 | 1 | Polygalacturonidase |
| | 4 | 303 | 33 | 35 | 28 | 4 | Beta-Galactosidase |
| | 5 | 48 | 2 | 2 | 3 | 2 | Endo-1,4-beta-Xylanase |
| | 6 | 54 | 5 | 7 | 6 | 2 | Cellobiohydrolase |
| | 7 | 34 | 2 | 6 | 4 | 2 | Neuraminidase |
| | 8 | 2 | 2 | - | 1 | - | 6-Phospho-beta-Glucosidase |
| | 9 | 3 | 1 | 1 | 1 | - | Endoglucanase |
| | 10 | 17 | 4 | 2 | 2 | 1 | Beta-Agarase B |
| O-(S)-glycosyl binding (***D***) | 11 | 3 | 1 | 1 | 1 | - | Endo-1,4-beta-Xylanase Y |
| | 12 | 2 | 2 | 1 | 1 | - | Glucoamylase |
| | 13 | 2 | 1 | - | 1 | - | Cellobiohydrolase |
| | 14 | 8 | 3 | 2 | 1 | - | Endo-1,4-beta-Xylanase D |
| | 15 | 2 | 1 | 1 | 1 | - | Endo-1,4-beta-Xylanase A |
| | 16 | 10 | 2 | 3 | 4 | - | Endoglucanase C |
| Unspecified (***E***) | - | 12 | 7 | 9 | 4 | 1 | Lysosomal alpha-Mannosidase |

A unique feature of the graphical representation of the similarity matrix is the appearance of lines or dots (boxes) outside the rectangles (homologous classes). These domains are the ones which show similarity to domains from different classes. This reveals structural entities within domains that share structural similarities, independent of classes. Some of these 'remote' homologies can be explained. For example structural subsets of

Journal of Universal Science Online WWW.JUSO.HU ORIGINAL PAPER ISSN: 2416-0008

the binding modules can be found in some of the catalytic domains where the catalytic core of the enzyme resembles to that of a binding module (Fig. 1 and Fig. 2/I region C, class 6). These regions share the same – substrate binding – function. For other distant similarities there is no obvious explanation as yet, and the appearance of these is under investigation. These can be caused by the bias found in the structural database (structure determination errors) or the limitations of the comparison method.

*Structural comparison by PRIDE Cluster*

In order to support the structural classification introduced previously, another structural comparison was conducted using PRIDE Cluster. The result matrix is sorted and presented in the same manner as earlier. Similar subclasses were identified as with using DaliLite. The outlines of the classes are recognizable; however the identification of some classes is not obvious. The PRIDE algorithm is more sensitive to distant structural similarities, so the results contain more "noise", as seen on the graph. For example, see class C4, where its outline is visible, but it does not appear as a uniform cluster. (Fig. 2/II).

*Classification using amino acid sequence similarity*

In order to set the structural classification against the CAZY system, we conducted an amino acid sequence comparison. The amino acid sequences of the domains were multiply aligned using ClustalW, and then a protein distance matrix was calculated using the Jones-Taylor-Thornton model of ProtDist. The result matrix is sorted and presented in the same manner as earlier. Whereas the subclasses are perfectly visible and match the results of DaliLite and PRIDE, between some of the subclasses sequence similarities occurred that did not correspond to the structural comparison results. (Fig. 2/III).

## Conclusions

Based on structural similarity, glycosyl hydrolases could be classified into 24 separate classes. N-glycosyl and O-glycosyl hydrolases were clearly different, the former consisting of 8 classes, and the latter consisting of 16 classes. Binding and catalytic domains formed separate classes, respectively. This structural classification clearly differs from both CAZY and CATH databases, despite the expected uniformity between these databases. The data extracted from the graphical representation represents the fold similarity better between domains, but interestingly, structurally homologous classes in glycosyl hydrolases did not concur to a single CAZY CLAN, which is supposed to represent fold similarities. PRIDE analysis created a similar node clustering, although it is more sensitive to distant structural similarities.

Clearly, understanding glycosyl hydrolases (and enzymes as such) the construction of databases based on different approaches is essential. This work hopes to add to the better understanding of this – functionally and structurally diverse – group of enzymes.

Journal of Universal Science Online WWW.JUSO.HU ORIGINAL PAPER ISSN: 2416-0008

## Acknowledgements

Special thank goes to Péter Róka and Péter Czanik PhD students and the Informatics Office of the Faculty of Agricultural and Enviromental Sciences, Szent István University. The authors would like to thank Sándor Pongor, International Centre for Genetic Engineering and Biotechnology (ICGEB), for making possible to use PRIDE on a large dataset.

## References

1. McBean LD, Miller GD (1998) Allaying Fears and Fallacies about Lactose Intolerance. **Journal of the American Dietetic Association 6**: *pp.* 671-676

2. Masson C, Cissé I, Simon V, Insalaco P, Audran M. (2004) Fabry disease: a review. **Joint Bone Spine 71**: *pp.* 381-383

3. Coughlan MP, Hazlewood GP (1993) Hemicellulose and hemicellulases. Portland Press, London, UK.

4. Palmarola-Adrados B, Chotěborská P, Galbe M, Zacchi G (2004) Ethanol production from non-starch carbohydrates of wheat bran. **Bioresource Technology 96**: *pp.* 843-850

5. Raunkjaer K, Hvitved-Jacobsen T, Nielsen, PH (1995) Transformation of organic matter in gravity sewer. **Water Environment Research 67**: *pp.*181-188

6. Kramer KJ, Muthukrishnan S (1997) Insect Chitinases: Molecular Biology and Potential Use as Biopesticides. **Insect Biochemistry and Molecular Biology 27**: *pp.* 887-900

7. Enzyme nomenclature (1992) Academic Press, San Diego, California.

8. Adelene Y. L. Sim and Chandra Verma (2015) How does a hydrocarbon staple affect peptide hydrophobicity? **Journal of Computational Chemistry 36**: *pp.* 773–784.

9. Henrissat B (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. **Biochemical Journal 280**: *pp.* 309–316

10. Naumoff, D. (2010) GH101 family of glycoside hydrolases. **Journal of Bioinformatics and Computational Biology 8**: *pp.* 437-451

11. Davies G, Henrissat B (1995) Structures and mechanisms of glycosyl hydrolases. **Structure 3**: *pp.* 853-859

12. Štefan Janeček, Karol Blesák (2011) Sequence-Structural Features and Evolutionary Relationships of Family GH57 α-Amylases and Their Putative α-Amylase-Like Homologues. **The Protein Journal 30**: *pp.* 429-435

13. J. Ecker, L. Fülöp, (2014) Molecular modeling of DDT's and it's major metabolites adsorption in the interlaminar space of montmorillonite. **Journal of Universal Science 1**: *pp.* 12-19

14. Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. **Biochemical Journal 293**: *pp.* 781-788

15. D. G. Naumoff (2011) Hierarchical Classification of Glycoside Hydrolases **Biochemistry (Moscow), 76**: *pp.* 622-635

16. Holm L, Sander C (1996) Mapping the protein universe. **Science 273**: *pp.* 595-602.

17. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. **Proceedings of the National Academy of Sciences 95**: *pp.* 14863-14868

18. Daniel J Rigden, Ruth Y Eberhardt, Harry J Gilbert, Qingping Xu, Yuanyuan Chang, Adam Godzik. (2014) Structure- and context-based analysis of the GxGYxYP family reveals a new putative class of Glycoside Hydrolase. **BMC Bioinformatics 15**: *pp.* 196-208

19. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. **Computer Applications in the Biosciences 8**: *pp.* 275-282

Journal of Universal Science Online
WWW.JUSO.HU
ORIGINAL PAPER
ISSN: 2416-0008

20. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Research 22**: *pp.* 4673-80

21. Carugo O, Pongor S (2002) Protein fold similarity estimated by a probabilistic approach based on C(*alpha*) - C(*alpha*) distance comparison. **Journal of Molecular Biology 315**: *pp.* 887-98

22. Henrik Aspeborg, Pedro M Coutinho, Yang Wang, Harry Brumer, Bernard Henrissat (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). **BMC Evolutionary Biology 12**: *pp.* 186-201

23. Holm L, Park J (2000) DaliLite workbench for protein structure comparison. **Bioinformatics 16**: *pp.* 566-567

24. Pearl FMG, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J,

25. Annalisa Bordogna, Alessandro Pandini and Laura Bonati (2011) Predicting the accuracy of protein–ligand docking on homology models. **Journal of Computational Chemistry 32**: *pp.* 81–98.

26. Coutinho PM, Henrissat B (1999) Carbohydrate-active enzymes: an integrated database approach. In "Recent Advances in Carbohydrate Bioengineering", Gilbert HJ, Davies G, Henrissat B and Svensson B eds., The Royal Society of Chemistry, Cambridge, 3-12

27. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V and Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. **Nucleic Acids Research 37**: *pp.* D233-D238

28. Shieh G.S (1998) A weighted Kendall's tau statistic. **Statistics & Probability Letters 39**: *pp.* 17-24