

*Original Russian text*

*Igolkina A.A., Andronov E.E., Porozov Yu.B., 2013 published in Matematicheskaya biologiya i bioinformatika, 2013, V. 8. № 2. P. 407-418. URL: [http://www.matbio.org/2013/Igolkina\\_8\\_407.pdf](http://www.matbio.org/2013/Igolkina_8_407.pdf).*

===== BIOINFORMATICS =====

UDC: 577.151.43 + 519.171.2 + 577.151.45

## **Method of Search for Substrate Specificity Regions in Cellulase Class Enzymes Based on their Primary and Tertiary Structures**

**Igolkina A.A.<sup>\*1</sup>, Andronov E.E.<sup>\*\*2</sup>, Porozov Yu.B.<sup>\*\*\*3</sup>**

<sup>1</sup> *St. Petersburg State Polytechnic University, St. Petersburg, 195251, Russia*

<sup>2</sup> *All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, 196608, Russia*

<sup>3</sup> *St. Petersburg National Research University of Information Technologies, Mechanics and Optics, St. Petersburg, 197101, Russia*

**Abstract.** In nature, the degradation of plant biomass, which mostly consists of plant cell walls, is implemented by microorganisms synthesizing cellulase class enzymes (CCEs). Cell walls fibers are composed of complex of polysaccharides, which is split by complicated CCEs. CCEs contain two types of domains. The first type is the catalytic domains that decompose polysaccharides. The second type is the binding domains with substrates (carbohydrate binding module, CBM). The ability of enzymes to decompose polysaccharides is due to the configuration of the catalytic site in the catalytic domain; in particular, the catalytic site should contain the complimentary binding site to the substrate. In this work, we have developed and tested a combined approach to identify CCEs' binding sites which could make the contact zone with plant polysaccharide substrates. This approach was applied to the 90 proteins identified with cellulase activity based on data from M. Hess et al.. As a result, we have found two consensus sequences of CCEs' binding sites which are complimentary with polysaccharide substrates, Carboxymethyl Cellulose (CMC) and Xylan. On the basis of the approach, we have developed a software that implements the basic stages of search and detection of binding sites. The developed method and the software can be used in the analysis of large groups of proteins with diverse substrate specificity to detect functional areas.

**Key words:** *cellulase, graph, site searching algorithm, structure of the protein, binding site, biofuels.*

### **INTRODUCTION**

Nowadays, there are several computational approaches to the assessment of macromolecular interactions; in particular, there are widely used varieties of docking [1]. This method of molecular modeling involves calculations on large arrays of input data, which requires a fair amount of time and resources. Therefore, attempting to narrow the searching area of the intermolecular contacts to performance of the docking protocols is an important step of optimization calculations.

---

\*[igolkinaanna11@gmail.com](mailto:igolkinaanna11@gmail.com)

\*\*[ceandr@gmail.com](mailto:ceandr@gmail.com)

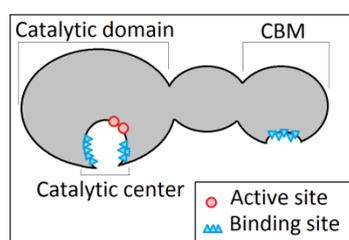
\*\*\*[porozov@ifc.cnr.it](mailto:porozov@ifc.cnr.it)

This paper describes an approach for filtering binding sites in CCEs, which consists of several steps. Its use as a pre-filter in the virtual screening enables to significantly accelerate the process of finding areas of binding.

The objectives of this work are to create methods of finding a consensus sequence of binding sites and search complementary binding sites to various substrates, based on the analysis of the primary and tertiary structures and substrate specificity of 90 proteins [2].

Nowadays, a full set of enzymes needed to degradate energy-intensive cereals and lignocellulosic biomass (cellulose, hemicellulose and lignin), is unknown. The Search of such CCEs is directed to increase efficiency of technologies of biofuel production [3–5]. For this purpose, there are created microbial strains that are transformed by vectors carrying genes of various CCEs [6, 7]. These modifications allow the microorganisms to produce a balanced set of CCEs needed to degrade lignocellulose. However, the currently known strains do not synthesize the optimal complex of CCEs decomposing biomass completely. In particular, this is due to the ability of individual enzymes to interact only with certain components of the biomass. Therefore, searching for genes and amino acid sequences of natural or synthetic CCEs, that includes the structural features of substrate specificity, leads to increased efficiency in production biofuels.

In nature, the splitting of plant polysaccharides is performed by several classes of microorganisms [2, 8]; in particular, the bacteria living in the digestive system of herbivores. M. Hess et al. [2] researched bacteria taken from a cow's rumen which adhered to the fibers of a vegetative substrate (switchgrass) placed in the rumen. After sequencing of metagenomic DNA of bacteria and de novo assembly of genomes, there were selected 90 genes of proteins that belong to the seven glycoside hydrolase family (GH3, GH5, GH8, GH9, GH10, GH26, GH48), part of which had specific binding domains with carbohydrates (CBM\_6, CBM\_4\_9). After testing the products of expression of these genes on the enzymatic activity, the data of the hydrolytic activity of each of the 90 proteins on six substrates (CMC, Xylan, IL-Switchgrass, IL-Miscanthus, IL-Avicel, Lichenan) were obtained. Selected results of the experiment are shown in Table 1.



**Figure 1.** Structure of CCEs.

Used in the testing substrates have a linear polymeric structure, and each of them is a component of plant biomass [3]. In order to retain and degrade a molecule of the polysaccharide, the structure of CCEs should have the characteristic pattern of the catalytic domain: a tunnel (groove, "pocket") which contains the active site and the binding site with the substrate (Fig. 1), [4, 9, 10]. Developed in this paper method of search binding sites considers both specified features of the tertiary structure of proteins and different versions of site models: strict and lax

Substrate Specificity (specificity to one or a group of substrates).

## ANALYSIS AND METHODS OF PROCESSING OF INITIAL DATA

### Analysis of initial data

The results of Matthias Hess et al. (Table 1) showed that CCEs, in the same domain hydrolase family, have unequal cellulase activity on a set of substrates. As a consequence, the activity of CCEs depends on the configuration of their catalytic sites, and not from the accessories to hydrolase family. Thus, one of the possible factors that determine the behavior of CCEs on the set of substrates belonged to one family can be a deletion of active sites in the catalytic center. However, the results of search with the help of Pfam [11] show that almost all (71 of 90) proteins have active sites relevant to their domains. Moreover, some proteins have not predicted active sites but could decompose substrates. Another possible factor which can explain the ambiguous behavior of CCEs is the inaccessibility of the active site on the molecular surface in the catalytic center. In this case, the two known active sites of GH5

family could give no more than four different specific bindings, which were not confirmed by an experiment (Table 1). Thus, it was concluded that the activity of CCEs is determined by the presence or absence of the areas essential for a complementary binding to the substrate.

Area of contact CCEs with the substrate consists of a set of binding sites (one or more separate short segments of a protein chain) [9]. The analysis of presented data [2] led to the conclusion that the binding region for all six substrates (CMC agar, Xylan, IL-Switchgrass, IL-Miscanthus, IL-Avicel, Lichenan) possess unique to their properties. Therefore, the main purpose of the search of binding areas was to find a set of amino acid areas length  $k$  ( $k$ -mers) for each of substrates, where each set has no common  $k$ -mers with other sets.

**Table 1.** Properties of the family GH5 by the results of work by Matthias Hess et al. [2]

Gene ID	CAZy family	Hydrolytic Activity <sup>3)</sup>						Domain organization <sup>4)</sup>
		C	X	S	M	A	L	
458803_07710	GH5							
0_06533	GH5					✓		
3271578_13460	GH5						✓	
2698429_129360	GH5	✓						
1696514_56150	GH5	✓				✓	✓	
2395619_81340	GH5	✓			✓		✓	
3671981_90060	GH5	✓		✓	✓	✓		
3932955_213080	GH5	✓			✓	✓	✓	
3953955_160820	GH5	✓		✓	✓	✓	✓	
558318_19410	GH5	✓	✓	✓	✓	✓		

### The method of processing initial data

Conjectural binding site to any of the six substrates was  $k$ -mer found in all protein sequences that are active on this substrate, and was not detected in any of the other proteins' sequences. Therefore, a simple search for areas of binding sites might look like that. If the products of genes G1 and G2 are active on two substrates (C and X), and the product of the gene G3 is active on one the substrate (X), then the common  $k$ -mers for sequences of genes G1 and G2 (and not found in the sequence of the gene G3) can be considered candidates for the substrate binding site of C. The more sequences G1 and G2 differ from G3, the more number of found candidates for binding sites will be. Thus, in order to reduce the "noise", which complicates search for binding sites of CCEs, it was better to separate CCEs genes into groups, representatives of which were slightly different in the amino acid sequence, but had unequal activity on substrates. Then search for  $k$ -mers was made in each group.

Thus, the search was divided into four stages:

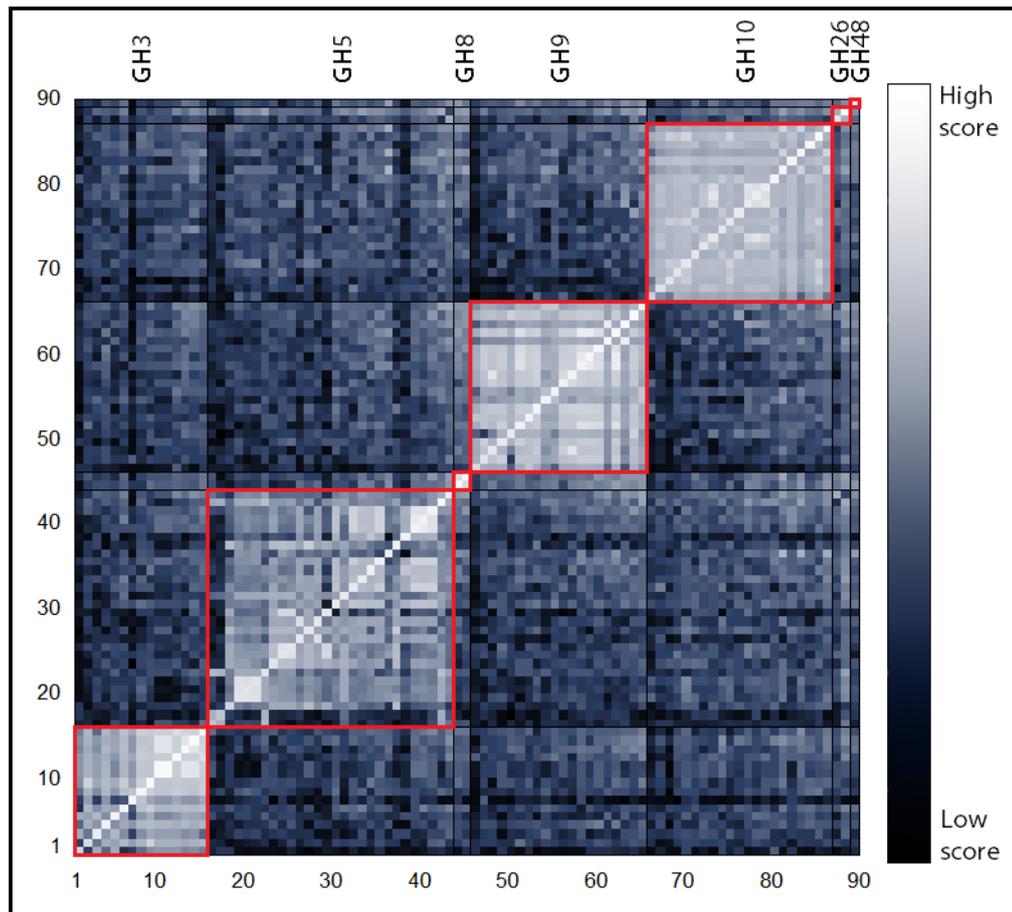
1. Grouping hydrolases based on simple and multiple global alignments.
2. Grouping hydrolases, based on the results of the construction of phylogenetic trees.
3. The algorithm of search of binding domains in the group.
4. Filtering using modeling of protein structure.

### Grouping hydrolases based on simple and multiple global alignments

Among the 90 participants from studied hydrolase families, there was conducted global pairwise alignment between all proteins using Needleall [12]. The results of alignments

confirm the presence of specific relations between proteins within the same family (Fig. 2). At the same time, the similarities between the hydrolase families (GH3, GH5, GH8, GH9, GH10, GH26, GH48) were not found. This absence of similarity between the sequences of proteins from different families means that the binding region for the substrates should be found in each family separately.

Further analysis of each protein family was performed with a series of multiple sequence alignments (ClustalW2 [13]) with varying amounts of protein in sets for MSA and alignment parameters. The result of alignment did not give enough reason to separate proteins in the family into meaningful groups. This may be due to different domain structure, intra-domain variability and the need to use "floating" parameters MSA.

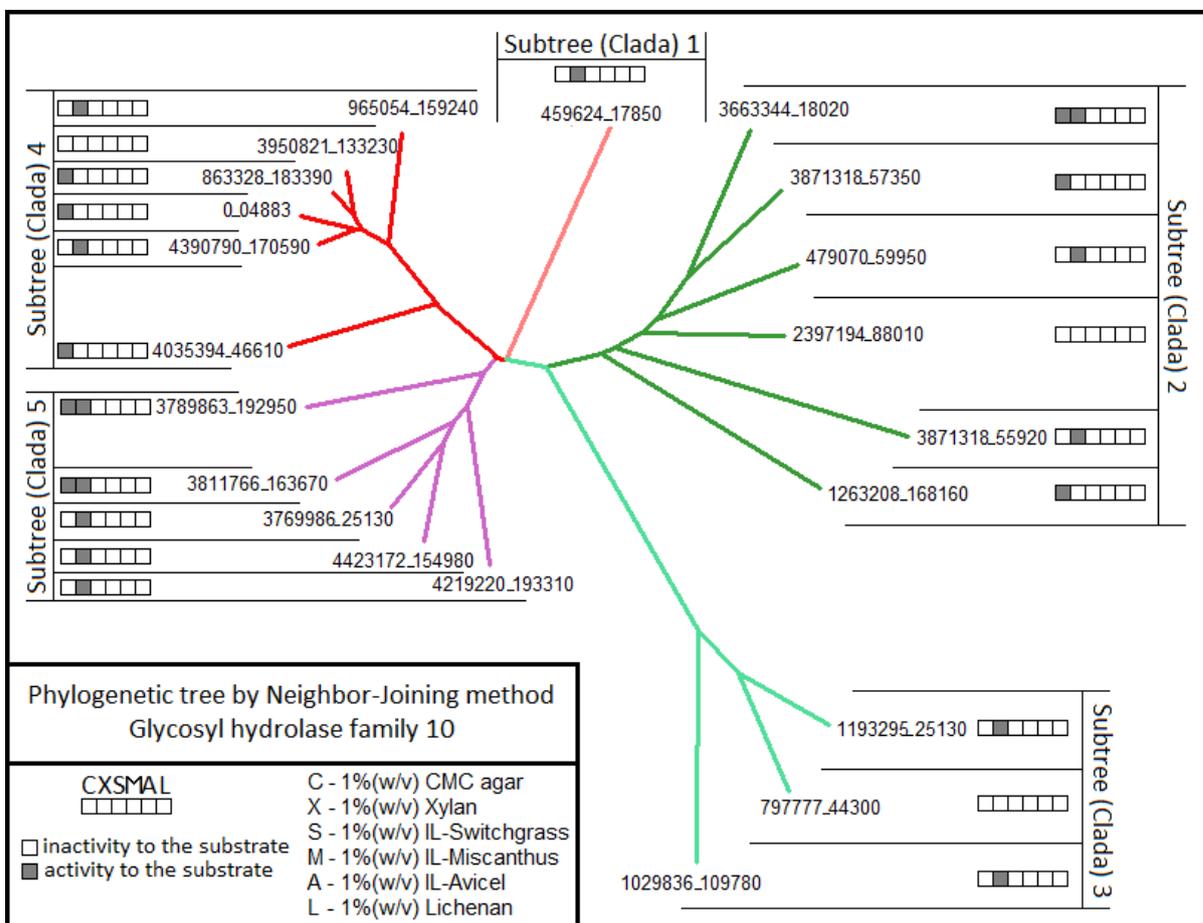


**Figure 2.** Results of pairwise comparison (alignment) of 90 proteins. The areas for a couple of proteins from one a family (in red squares) are the areas of great similarity.

### ***Grouping hydrolases, based on the results of the construction of phylogenetic trees***

The results of the multiple global alignments have shown that the protein sequences have many differences that prevent identifying meaningful groups (the groups of proteins that differ in activity to the substrate and have minimal differences in the amino acid sequence). Therefore, there was made an analysis of proteins domains responsible for the functionality CCEs. From each protein in [2] its functional site (domain name or part of it) was isolated and there was made the pair alignment of the site with the consensus sequence of the corresponding domain of the family (using Pfam [11]). After that the all obtained pairwise alignments of the each family were joined for the multiple alignments. Composite multiple alignments were obtained with the help of the supporting sequence of the domains for each of the seven hydrolase families in Pfam.

The results of the MSA were functional regions of proteins of each family. These results were a starting point for analysis with phylogenetic trees. The phylogenetic trees were constructed with using the algorithm Neighbor-Joining (package MEGA [14]). For each tree there were identified several large clades, in which the search for binding domains was made (Fig. 3). While dividing the tree into clades we took into account that within the same clade, on the one hand, there must be proteins different in substrate specificity (if possible), but on the other hand, these proteins have a relation in sequences.



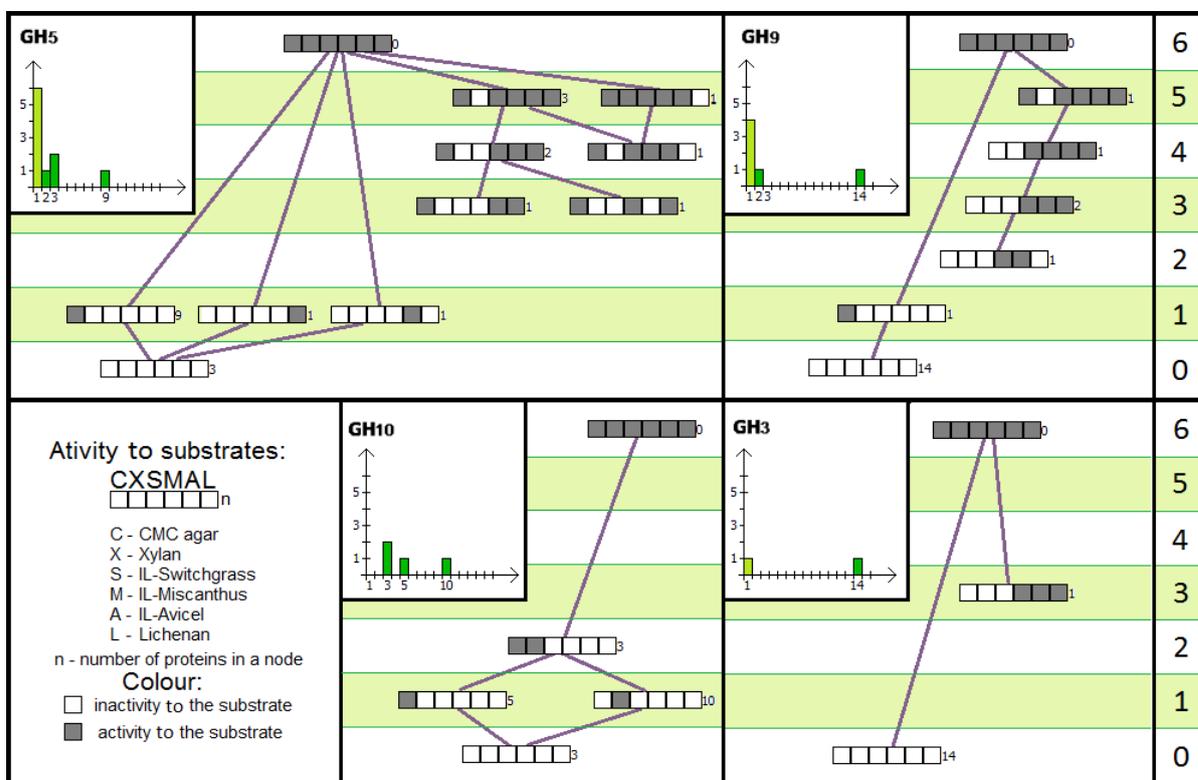
**Figure 3.** Phylogenetic tree constructed for members of the family GH10. The substrate specificity for each representative of family. The searching algorithm of binding sites was done in selected subtrees (clades). Gene ID of the corresponding genes takes place on the leaves.

### The searching algorithm of binding sites in the group

One of the conditions of the group formation was the presence in the one clade of proteins with a variety (different and same) of substrate specificity. This requirement made it possible to form in a clade several subgroups containing CCEs with the same activity. On the basis of such grouping for each clade there was constructed a graph of activity which nodes were matched to subgroup (activity graph) (Fig. 4). Each node of the graph corresponds to a set of objects (properties, attributes attributes): a set of protein sequences of members of the subgroup, activity on the substrates (the same for all members of the subgroup) and some lists of  $k$ -mers. Additionally, one extra node (zero vertex) was added to the graph; it did not contain a list of sequences of proteins and  $k$ -mers, but possessed a hypothetical activity on all substrates. The data of activity of 90 CCEs gave the opportunity to separate nodes in the graph into seven levels. The node's level number is the number of substrates; all of the node's

proteins are active on this number of substrates. The edge of the graph connects V1 with V2, if all the following conditions are met:

- Activity corresponding node V2 includes activity corresponding node V1.
- Node V2 is on neighboring level with V1. If the neighboring level with the V1 was not a node that satisfies the above condition, then V2 assumes zero vertex.



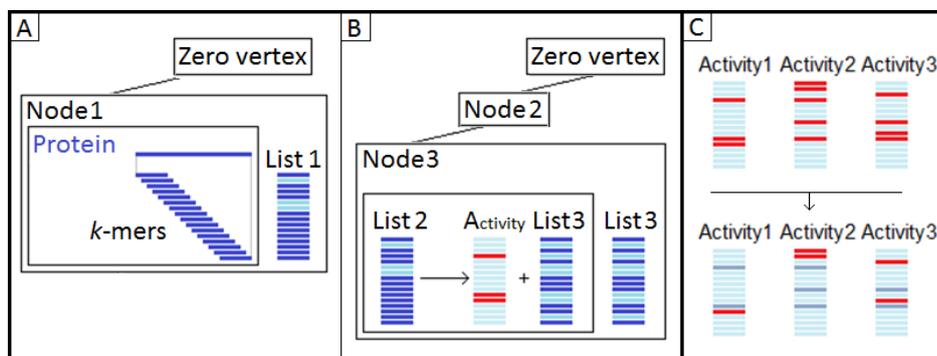
**Figure 4.** Activ graph made up of representatives of the studied hydrolase families. Histograms of the number of sequences in the nodes.

Algorithm of searching for candidates for binding sites consisted of two traversing of this graph. Traversings were realized consistently by levels and started at 6 (from the zero vertex), ended at 0.

At the first graph traversing the list of  $k$ -mers was formed in each node. Each list contains  $k$ -mers which are common for all sequences in the node. If on the new step of the traversing a node appears connected by an edge to the zero vertex (Fig. 5,A), then the list of  $k$ -mers was formed from the all  $k$ -mers from the random sequence.  $k$ -mer was added to the list of this node, if it is found in all of node's sequences. If on the new step of the traversing a node appears not connected with the zero vertex, then  $k$ -mers were chosen from lists of all nodes (with greater level) which were connected with this node. Similarly, if a  $k$ -mer was found in all sequences of the node, it was added to the node's list. If the  $k$ -mer was not found in any sequence of the node, it was suspected as a binding site with some substrate. The substrate was defined as a difference between the current node and the node with greater level, which the  $k$ -mer was belonged to (Fig. 5,B). Thus, after the first graph traversal each node's list of  $k$ -mers was created and the candidates of binding sites have been identified.

After the first graph traversing found  $k$ -mers could be associated with different substrates. Such a result was due to the branching of the graph and the independence of a list in the nodes. At the same time for a fair division of  $k$ -mers by substrates it required to detect that each  $k$ -mers corresponded to only one substrate. Therefore, after the first graph traversal the same  $k$ -mers corresponding to different substrates, were excluded from further consideration (Fig. 5,C).

As it has already been noted, selected  $k$ -mers could be obtained from different (independent) paths from level 6 to level 0, so the second graph traversal was followed by the first. The second traversal checked for the presence of each found candidate for the binding site in all nodes that was not involved in its formation.



**Figure 5.** The steps of the searching algorithm for binding sites.

### Filtering with modeling the structure of proteins

Additional testing of found sites was performed on the three-dimensional structure of proteins. By using ModWeb [15], for each of the 90 proteins a 3D-model was constructed. Then found  $k$ -mers were marked on the constructed structures and their locations were identified as one of the list: inside a protein, on the surface of the protein not in a indentation, on the protein surface in the indentation. Sites, which formed the surface of indentations (pockets or grooves) were taken by us as reliable binding sites. Specific active sites (of each family) from Pfam were also marked on protein structures (Fig. 6).

## RESULTS AND DISCUSSION

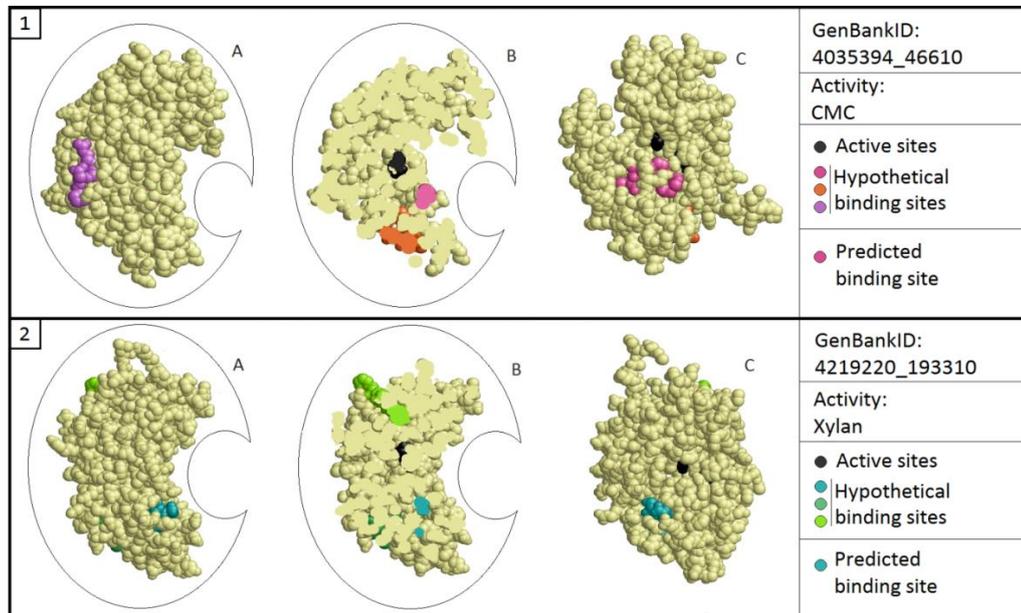
The developed method for searching binding sites was applied to the data on the activity of 90 protein from seven hydrolase families (GH3, GH5, GH8, GH9, GH10, GH26, GH48). At the stage grouping (grouping hydrolases, based on simple and multiple global alignment) proteins into families, it appeared that in the data there is one protein family GH48 in which was represented by only one protein, and both families GH8 and GH26 were represented by two proteins. Such a number of representatives in the family was not enough to reveal the characteristic of binding sites to substrates. Furthermore, an activity of only one of these five CCEs was detected. Therefore, these genes were excluded from CCEs consideration.

At the stage grouping (grouping hydrolases, based on the results of the construction of phylogenetic trees) by phylogenetic trees (Fig. 3), it appears that the separation of protein families by clades produced very small groups. This was due to the insufficient number of primary data. That is why, the searching algorithm for binding sites was applied graphs which was constructed for families GH3, GH5, GH9, GH10 on the whole(Fig. 4).

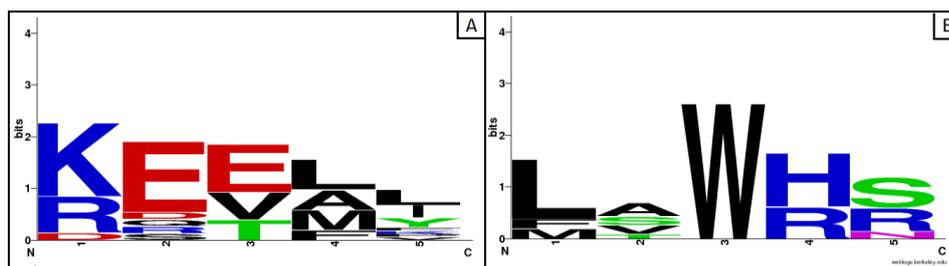
A large number of nodes in the graph which had only one element seriously hampered the correct searching for binding sites ( $k$ -mers). The algorithm involves the forming of lists of  $k$ -mers which were common to characteristics sequences in a node. Thus, the greater number of sequences contained in the node, the more specific  $k$ -mers, forming a list, are. On the contrary, if the node has only one sequence, its list of  $k$ -mers is found uninformatively, leading very noisy output. Therefore, besides the activity graphs of families there were constructed histograms of the number of sequences in the nodes (Fig. 4). It appeared) that only one graph (graph of the family GH10) could be considered quite complete, so the further work was done only with it.

GH10 family contains 21 CCEs with known activity on the substrates. Also if a member of the family has been active on a substrate, this substrate was the CMC and / or Xylan. In

this family the algorithm had worked correctly with with the parameter  $k$  equaled 5. As a result there were found three binding sites to the substrate CMC and three sites to the substrate Xylan. Determination of their location on the modeled protein structures showed that all known binding sites were located on the surface of the protein, but only one in each triplet (group of three predicted binding sites) was observed in the indentation of the catalytic site, which contains the active site (Fig. 6). Using the data of identified binding sites for each members of the family GH10 (Table 2), conserved sequences for binding sites of the CCEs to substrates CMC and Xylan (Fig. 7) were obtained.



**Figure 6.** Location of predicted binding sites on the three dimensional structure of two proteins of the family GH10. Sites correspond to substrates CMC and Xylan (1, 2). Location of Hypothetical binding sites with substrates CMC and Xylan is found by the searching algorithm. A: lateral view of the protein. B: lateral view of the protein in a section: active sites and predicted binding sites are located in the catalytic site on the section. C: a view from the catalytic site.

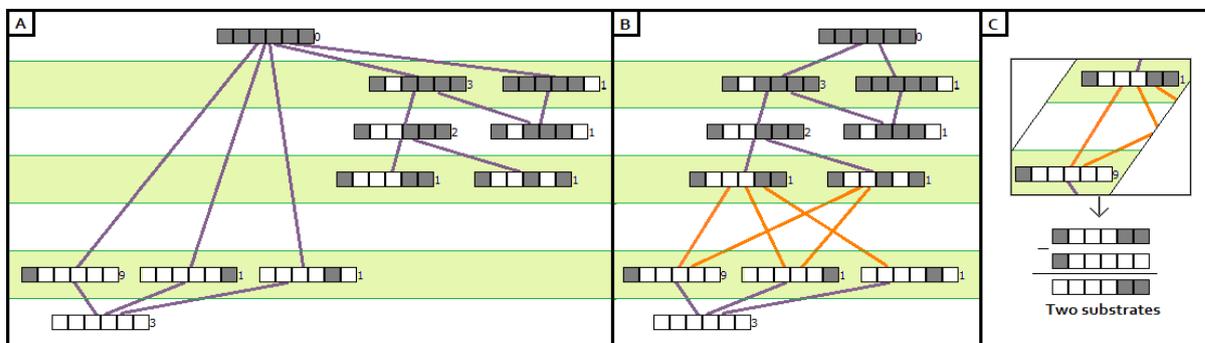


**Figure 7.** A graphical representation of the consensus sequence of the binding site (Sequence Logo [16]). A – binding site with CMC. B – binding site with Xylan.

However, excepting for sites that can complementarily bind only with one particular substrate, in the catalytic site of the enzyme other types of sites could be located. There are sites which are capable of forming contact with a set of molecules. Moreover, one enzyme may have got several binding sites (on the surface) which all are specific to a one substrate (Fig. 2).

Thus, the complete set of specific CCEs binding sites with substrates CMC and Xylan may not confine of found two. Exclusion from consideration other substrate specificity binding sites (if such existed) could be due to the method of construction activity graph. The characteristic detail of the construction is that an edge which connects two nodes placed only in adjacent levels (if any of the nodes is not the zero vertex). If we assume that an edge can

connect nodes that are not only on the adjacent levels (if none of them is zero vertex), this condition adds additional edges to the activity graph. Because of the increasing number of edges the number of steps and the number of  $k$ -mers (at the output of the algorithm) in the searching algorithm also increases (Fig. 8,B). Detected by this graph additional sites probably extend the set of found substrate specificity sites. However, such additional  $k$ -mers will have to pass a procedure of concretization of specificity, since they will be determined by the specificity for one of several substrates (Fig. 8,C).



**Figure 8.** Modification of the activity graph for the family GH10 (see description in the text). *A*: The activity graph contains compounds of nodes that take place only on adjacent levels. *B*: The activity graph; the way to construct allows the connection of nodes, which take place not only on adjacent levels. The difference between numbers of levels could be equal to one or two. *C*: The part of the graph of activity. While the passage of the algorithm on the edge (selected), some new selected  $k$ -mers will be candidates for the binding sites not with a specific substrate, but with one of the two.

**Table 2.** Binding sites of the family GH10 with substrates CMC and Xylan

Gene ID	Predicted binding sites to substrates		Hydrolytic Activity					
	CMC	Xylan	C	X	S	M	A	L
1263208_168160	Leu 102 - Asn 106		V					
4035394_46610	Leu 30 - Ser 34		V					
4423172_154980		Lys 27 - Tyr 31		V				
3789863_192950	Leu 89 - Ser 93	Lys 111 - Val 115	V	V				
4219220_193310		Lys 112 - Tyr 116		V				
3663344_18020	Leu 343 - Arg 347	Arg 148 - Leu 152	V	V				
965054_159240		Arg 172 - Ile 176		V				
0_04883	Leu 121 - Ser 125		V					
479070_59950		Arg 179 - Leu 183		V				
3769986_25130		Lys 160 - Tyr 164		V				
3871318_57350	Met 378 - Arg 382		V					
863328_183390	Leu 104 - Ser 108		V					
4390790_170590		Lys 129 - Leu 133		V				
3950821_133230								
2397194_88010								
1193295_25130		Lys 359 - Leu 363		V				
459624_17850		Arg 183 - Ile 187		V				
1029836_109780		Pro 121 - Ile 125		V				
3871318_55920		Arg 143 - Leu 147		V				
3811766_163670	Leu 305 - Arg 309	Lys 152 - Pro 156	V	V				
797777_44300								

The developed method of search is directed to the detection of sites that match to only one specific substrate, but with it you can also find other types of binding sites. In the proposed search algorithm of binding sites, at some point (Fig 5,C)  $k$ -measures corresponding to several substrates were excluded from consideration. Since these  $k$ -mers are common for some

sequences of CCEs, we can assume that some of them were binding region that is not selective to specific substrates. As shown in Fig. 7,C, adding new edges to the activity graph may lead to the discovery of sites which substrate specificity cannot not be defined explicitly. In that way, the search area of binding sites that do not have substrate specificity could be narrowed by combining two sets of  $k$ -mers: the first – the activity of which was shown on several substrates and found on the modified graph, the second – the activity of which was uncertain.

Analysis of protein sequences using Pfam confirmed that the CCEs have a complex domain organization, and, except a domain that defines the family, the enzyme can contain multiple domains CBM. For example, three members of the family GH10 have a CBM\_6 domain in their structure. It was noticed that these three CCEs formed the clade (subtree) number three in the phylogenetic tree (Fig. 3). Thus, the presence of a larger number of entries data, not only allows to search sites in clades of phylogenetic tree, but also makes it possible to find the area of CBM domains (in this case - of domain CBM\_6).

The developed method found the substrate specificity binding sites even when the original sample of CCEs is small. This approach can be modified by offered options and applied to representative samples of objects, which represent a sequence of an alphabet, to highlight specific areas.

This work was supported by the Ministry of Education and Science of Russian Federation (Contract № 16.552.11.7085), by Ministry of Education and Science of Russian Federation in the context of Federal Program "Scientific and pedagogical personnel of innovative Russia", agreement № 14.B37.21.0562.

## REFERENCES

1. Woodcock S., Henrissat B., Sugiyama J. Docking of Congo Red to the Surface of Crystalline Cellulose Using Molecular Mechanics. *Biopolymers*. 1995. V. 36. P. 201–210.
2. Hess M., Sczyrba A., Egan R., Kim T.W., Chokhawala H., Schroth G., Luo S., Clark D.S., Chen F., Zhang T. et al. Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science*. 2011. V. 331. P. 463–467.
3. Rubin E.M. Genomics of cellulosic biofuels. *Nature*. 2008. V. 454. P. 841–845.
4. Himmel M.E., Ding S.Y., Johnson D.K., Adney W.S., Nimlos M.R., Brady J.W., Foust T.D. Biomass Recalcitrance: Engineering Plants and enzymes for Biofuels Production. *Science*. 2007. V. 315. P. 804–807.
5. Samuel R., Pu Y., Foston M., Ragauskas A.J. Solid-state NMR characterization of switchgrass cellulose after dilute acid pretreatment. *Biofuels*. 2010. V. 1. P. 85–90.
6. Gilkes N.R., Kilburn D.G., Langsford M.L., Miller Jr R.C., Wakarchuk W.W., Warren R.A.J., Whittle D.J., Wong W.K.R. Isolation and Characterization of *Escherichia coli* Clones Expressing Cellulase Genes from Cellulomonas Jimi. *Journal of General Microbiology*. 1984. V. 130. P. 1377–1384.
7. Gilkes N.R., Langsford M.L., Kilburn D.G., Miller R.C.Jr., Warren R.A. Mode of Action and Substrate Specificities of Cellulases from Cloned Bacterial Genes. *The Journal of Biological Chemistry*. 1984. V. 259. № 16. P. 10455–10459.
8. Abu Bakar N.K., Abd-Aziz S., Hassan M.A., Ghazali F.M. Isolation and Selection of Appropriate Cellulolytic Mixed Microbial Cultures for Cellulases Production from Oil Palm Empty Fruit Bunch. *Biotechnology*. 2010. V. 9. P. 73–78.
9. Koivula A., Reinikainen T., Ruohonen L., Valkeajärvi A., Claeysens M., Teleman O., Kleywegt G.J., Szardenings M., Rouvinen J., Jones T.A., Teeri T.T. The active site of *Trichoderma reesei* cellobiohydrolase II: the role of tyrosine 169. *Protein Engineering*. 1996. V. 9. № 8. P. 691–699.

10. Knowless J., Lehtovaara P., Teeri T. Cellulase families and their genes. *Trends in biotech.* 1987. V. 5. P. 255–261.
11. Bateman A., Coin L., Durbin R., Finn R.D., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E.L.L. et al. The Pfam protein families database. *Nucleic Acids Research.* 2004. V. 32. P. 138–141.
12. Rice P., Longden I., Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends in Genetics.* 2000. V. 16. № 6. P. 276–277.
13. Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R. et al. Clustal W and Clustal X version 2.0. *Bioinformatics.* 2007. V. 23. № 21. P. 2947–2948.
14. Tamura K., Dudley J., Nei M., Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution.* 2007. V. 24. P. 1596–1599.
15. Pieper U., Eswar N., Braberg H., Madhusudhan M.S., Davis F.P., Stuart A.C., Mirkovic N., Rossi A., Marti-Renom M.A., Fiser A et al. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Research.* 2004. V. 32. P. D217–D222.
16. Crooks G.E., Hon G., Chandonia J.M., Brenner S.E. WebLogo: A Sequence Logo Generator. *Genome Research.* 2004. V. 14. P. 1188–1190.

Received Apr 24, 2013.

Published Jul 16, 2013.