МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2011, том 45, № 6, с. 1073–1083

БИОИНФОРМАТИКА

УДК 575.852.112:577.152.321

GHL1–GHL15: НОВЫЕ СЕМЕЙСТВА ГИПОТЕТИЧЕСКИХ ГЛИКОЗИЛГИДРОЛАЗ

© 2011 г. Д. Г. Наумов[1, 2]*

[1]Институт микробиологии им. С.Н. Виноградского Российской академии наук, Москва, 117312

[2]Государственный научный центр "ГосНИИ генетика", Москва, 117545

Поступила в редакцию 15.03.2011 г.
Принята к печати 04.05.2011 г.

Домены пятнадцати недавно обнаруженных семейств гипотетических гликозилгидролаз GHL1–GHL15 использованы для итеративного скрининга базы данных аминокислотных последовательностей. Выявлены их эволюционные связи как между собой, так и с представителями ряда других ранее известных семейств белковых доменов: GH5, GH13, GH13_33, GH17, GH18, GH20, GH27, GH29, GH31, GH35, GH36A, GH36B, GH36C, GH36D, GH36E, GH36F, GH36G, GH36H, GH36J, GH36K, GH39, GH42, GH53, GH66, GH97, GH101, GH107, GH112, GH114, COG1082, COG1306, COG1649, COG2342, DUF3111 и PF00962. Неклассифицированные гомологи сгруппированы в 35 новых семейств гипотетических гликозилгидролаз: GHL16–GHL50. Обсуждается положение семейств GHL1–GHL15 в иерархической классификации гликозилгидролаз и их гомологов, предложены новые суперсемейства белковых доменов.

***Ключевые слова*: гликозилгидролаза, семейства GHL, новые семейства белков, суперсемейство белков, *PSI-BLAST, PSI Protein Classifier, CAZy*, TIM-бочонок, иерархическая классификация белков, эволюция белков, поиск гомологов, аннотация генов.**

GHL1–GHL15: NEW FAMILIES OF HYPOTHETICAL GLYCOSIDE HYDROLASES, by *D. G. Naumoff*[1, 2]* (Winogradsky Institute of Microbiology, Russian Academy of Sciences, Moscow 117312, Russia; State Research Center "GosNII Genetika", Moscow, 117545, Russia; *e-mail: daniil_naumoff@yahoo.com). Domains of fifteen recently found families of hypothetical glycoside hydrolases (GHL1–GHL15) have been used for iterative screening of the protein database. Evolutionary connections between representatives of these families were revealed. Also, their relationship with members of the following known families of protein domains were found: GH5, GH13, GH13_33, GH17, GH18, GH20, GH27, GH29, GH31, GH35, GH36A, GH36B, GH36C, GH36D, GH36E, GH36F, GH36G, GH36H, GH36J, GH36K, GH39, GH42, GH53, GH66, GH97, GH101, GH107, GH112, GH114, COG1082, COG1306, COG1649, COG2342, DUF3111, and PF00962. The unclassified homologues were grouped into 35 new families of hypothetical glycoside hydrolases: GHL16–GHL50. Position of GHL1–GHL15 families in the hierarchical classification of glycoside hydrolases and their homologues is discussed. Several new superfamilies of protein domains are suggested.

***Keywords:* glycoside hydrolase, GHL families, new protein families, protein superfamily, PSI-BLAST, PSI Protein Classifier, CAZy, TIM-barrel fold, hierarchical protein classification, protein evolution, search of homologues, gene annotation.**

> "*The characterization of hypothetical proteins is a major challenge in the post-genomic area*."
>
> E. Rebuffet et al., 2011 [1]

Успешное осуществление тысяч геномных проектов [2] превращает функциональную аннотацию закодированных в геномах белков в одну из важнейших задач современной молекулярной биологии. Домены – строительные модули, из которых построены белки – претерпевают изменения и соединяются в различных комбинациях в процессе эволюции, но обычно они не возникают заново [3]. Объединение структурно сходных (т.е. эволюционно близкородственных) доменов в семейства облегчает их функциональную аннотацию. Обычно белки одного семейства обладают одинаковыми или сходными функциями, например, энзиматическими активностями. Однако многие семейства белковых доменов не содержат функционально охарактеризованных в эксперименте представителей, что затрудняет их аннотацию. Один из возможных подходов в такой ситуации – поиск дальних гомологов, относящихся к другим семействам, для которых

* Эл. почта: daniil_naumoff@yahoo.com

уже существует функциональная аннотация. Однако использование неохарактеризованных белков для поиска предполагаемых гомологов путем итеративного скрининга белковой базы данных не всегда оказывается успешным. При скрининге часто не выявляются какие-либо дополнительные гомологи, особенно в тех случаях, когда семейство представлено лишь небольшим числом близкородственных белков. В результате этого, функционально охарактеризованными (как экспериментально, так и методами биоинформатики) в первую очередь оказываются крупные семейства, хотя число их существенно меньше, чем число мелких семейств. Часто проблема поиска уже охарактеризованных эволюционно родственных семейств может быть снята, если задачу решать в обратном направлении: искать дальние гомологи для представителей многочисленных и хорошо исследованных семейств белковых доменов. Особенно сложна аннотация семейств белков, представленных, главным образом или исключительно, в организмах, принадлежащих к экспериментально малоизученным филогенетическим группам прокариот.

Обширная группа ферментов, катализирующих гликолитическое расщепление углеводов и их производных, – гликозилгидролазы (К.Ф. 3.2.1), на основании сходства их каталитических доменов, была распределена по 120 семействам: GH1–GH125, кроме GH21, GH40, GH41, GH60 и GH69 [4]. При сравнительном анализе первичных и третичных структур соответствующих белков во многих случаях выявляются эволюционные связи между представителями различных семейств, что позволило разработать иерархическую классификацию каталитических доменов гликозилгидролаз и их гомологов [5, 6]. При этом почти все семейства гликозилгидролаз объединены в шесть основных групп, в пределах которых каталитические домены обладают сходной пространственной структурой и имеют, вероятно, общее эволюционное происхождение [6]. На основании наиболее близкого родства 51 семейство было сгруппировано в 14 кланов (GH-A–GH-N) [4, 6].

Использование каталитического домена семейства GH43 (клан GH-F) для итеративного скрининга базы данных позволило проследить его родство с группой функционально неохарактеризованных белков, названной GHLP (впоследствии – COG2152 или DUF377), а также с белком SCO3481 (GenPept, CAB61805.1) из бактерии *Streptomyces coelicolor* [7]. Недавно этот белок вместе с группой его близких гомологов был отнесен к новому семейству GH117 гликозилгидролаз [1, 4]. Показано также близкое родство белков семейства GHLP с еще одной группой функционально неохарактеризованных белков – семейством DUF1861 [6, 8].

Сравнительный анализ аминокислотных последовательностей показал, что имеется структурное сходство ряда лизоцимов с белками двух функционально слабоизученных семейств: белки семейства COG3926 (или DUF847) оказались наиболее близки к лизоцимам семейства GH24 (клан GH-I), а белки из семейства COG5526 – к представителям семейства GH23 [9]. Следует отметить, что семейство COG3926/DUF847 в настоящее время рассматривается как семейство GH108 гликозилгидролаз [4].

Использование каталитических доменов семейства GH27 (клан GH-D) позволило с помощью итеративного скрининга базы данных обнаружить их родство с представителями ряда семейств гликозилгидролаз, а также семейств COG1649 (или DUF187) и GHX [10, 11]. Впоследствии семейство GHX было включено в классификацию гликозилгидролаз под названием GH97 [4, 12]. При использовании доменов семейств GH13 (клан GH-H) и GH31 (GH-D) обнаружены их эволюционные связи соответственно с семействами COG1649 и COG2342 [13] и COG1306, COG1649 и COG3868 [14]. Следует отметить, что некоторая часть представителей семейств COG2342 и COG3868 в настоящее время объединена в новое семейство гликозилгидролаз GH114 [4, 15]. Однако мы в настоящей работе, как и ранее [6, 15], будем называть GH114-доменами только представителей семейства COG3868.

Исследование белков семейства GH36 (клан GH-D) и близких гомологов позволяет их рекласифицировать в 11 новых семейств – GH36A–GH36K, многие из которых не содержат экспериментально охарактеризованных белков [5, 6, 10, 11, 14, 15].

Сравнительный анализ белков, полученных при итеративном скрининге базы данных аминокислотных последовательностей с помощью каталитических доменов семейств GH101 [16] и GH114 (COG3868) [15], позволил идентифицировать 15 новых семейств гипотетических гликозилгидролаз – GHL1–GHL15 – и предсказать вероятное строение активного центра у этих белков. В настоящей работе проведен итеративный скрининг базы данных аминокислотных последовательностей с использованием доменов семейств GHL1–GHL15 для исследования эволюционных связей между семействами белков и поиска новых семейств гипотетических гликозилгидролаз.

АНАЛИЗ ДАННЫХ

Основной скрининг базы данных аминокислотных последовательностей GenPept на сайте NCBI (http://www.ncbi.nlm.nih.gov/) проводили 9 февраля 2011 г. с помощью программы PSI-BLAST. В качестве запроса (query) служили GHL-домены

Таблица 1. Представители семейств GHL1–GHL15, использованные для скрининга базы данных с помощью программы PSI-BLAST

Белок	Семейство	Организм (отдел прокариот)	Размер белка	Область гомологии	Аннотация (NCBI)
AAN24642.1	GHL1	*Bifidobacterium longum* (Actinobacteria)	631	290–532	hypothetical protein
EFB00786.1	GHL2	*Victivallis vadensis* (Lentisphaerae)	960	471–750	»
BAH37732.1	GHL3	*Gemmatimonas aurantiaca* (Gemmatimonadetes)	785	302–575	»
EFB01375.1	GHL4	*Victivallis vadensis* (Lentisphaerae)	1210	544–800	»
ACB52584.1	GHL5	*Cyanothece* sp. ATCC 51142 (Cyanobacteria)	845	198–611	unknown
ADE53551.1	GHL6	*Coraliomargarita akajimensis* (Verrucomicrobia)	356	40–272	hypothetical protein
ABQ90769.1	GHL7	*Roseiflexus* sp. RS-1 (Chloroflexi)	434	36–427	»
ACU02990.1	GHL8	*Pedobacter heparinus* (Bacteroidetes)	392	18–392	»
AAM01554.1	GHL9	*Methanopyrus kandleri* (Euryarchaeota)	389	14–245	predicted membane protein
ABF43386.1	GHL10	*Koribacter versatilis* (Acidobacteria)	494	40–369	conserved hypothetical protein
EFB00125.1	GHL11	*Victivallis vadensis* (Lentisphaerae)	519	145–478	hypothetical protein
ACB73617.1	GHL12	*Opitutus terrae* (Verrucomicrobia)	1221	719–919	α-N-arabinofuranosidase
AAW77166.1	GHL13	*Xanthomonas oryzae* (Proteobacteria)	663	350–635	HmsF protein
ACL40874.1	GHL14	*Arthrobacter chlorophenolicus* (Actinobacteria)	360	2–356	conserved hypothetical protein
BAH54049.1	GHL15	*Rhodococcus opacus* (Actinobacteria)	460	7–458	hypothetical protein

Примечание. В первой колонке указан номер соответствующей аминокислотной последовательности белка в базе данных GenPept (NCBI). В колонке "Размер белка" указано суммарное число аминокислотных остатков в белке-предшественнике. В колонке "Область гомологии" даны номера аминокислотных остатков в последовательности белка, которые ограничивают участок, соответствующий исследуемому домену.

белков семейств GHL1–GHL15 (табл. 1). Границы доменов определяли на основании гомологии с ранее известными доменами, имеющими структуру $(\beta/\alpha)_8$-бочонка. По каждому из запросов поиск вели в разделе "non-redundant protein sequences" базы данных GenPept. В качестве порогового значения величины *E*-value для включения найденного белка в модель, используемую программой PSI-BLAST на каждой следующей итерации, брали величину 0.005. Результаты обобщали с помощью недавно разработанной нами программы PSI Protein Classifier [14]. В работе также анализировали результаты скринингов, проведенных в более ранние сроки (с января 2009 по февраль 2011 г.).

Критерий для отнесения белка к ранее известному семейству – обнаружение этого белка с помощью соответствующего домена какого-либо представителя данного семейства в процессе первой итерации программы PSI-BLAST с $E\text{-value} \le 10^{-7}$. В качестве новых семейств рассматривали существенно изолированные группы белков-гомологов. Соответствующие домены этих белков использовали в качестве запросов для скрининга базы данных с помощью программы PSI-BLAST. Если ни один из этих доменов не позволял обнаружить в процессе первой итерации (с $E\text{-value} \le 10^{-7}$) ни одного фрагмента аминокислотной последовательности, принадлежащего домену из ранее известного семейства, то эти домены рассматривались как представители новых семейств. Их объединяли в семейства на основании того же критерия, что и при отнесении белков к ранее известным семействам (т.е. $E\text{-value} \le 10^{-7}$). Для определения принадлежности белков конкретным семействам (ранее известным или описываемым впервые) применяли программу PSI Protein Classifier.

Для проведения кластерного анализа использовали программу NEIGHBOR (метод ближайших соседей, Neighbor-Joining method) из пакета PHYLIP (http://evolution.gs.washington.edu/phylip.html). Программу TreeView Win32 (http://taxonomy.zoology.gla.ac.uk/ rod/treeview.html) применяли для визуализации деревьев.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Для итеративного скрининга базы данных аминокислотных последовательностей на сайте NCBI с помощью программы PSI-BLAST было использовано по одному домену семейств GHL1–GHL15 (табл. 1). По итогам вторых итераций с доменами семейств GHL12 и GHL15 были обнаружены те же два белка, что и по итогам их первых итераций (ACB73617.1 и EFL62392.1 для GHL12; BAH54049.1 и CAO81213.1 для GHL15); дальнейший поиск их гомологов на этом прекращали. Скрининг с помощью домена семейства GHL7

останавливали на тринадцатой итерации, так как она не обнаруживала новых белков в сравнении с предыдущей итерацией (с *E*-value ≤ 0.005). С остальными двенадцатью доменами (GHL1–GHL6, GHL8–GHL11, GHL13 и GHL14) проводили от четырех до семи итераций, их число обусловлено техническими ограничениями интернет-версии программы PSI-BLAST. В общей сложности обнаружено 26235 неидентичных белков. Большинство из них в области гомологии содержат домены ранее известных семейств гликозилгидролаз: GH5 (клан GH-A), GH13 (GH-H), GH17 (GH-A), GH18 (GH-K), GH20 (GH-K), GH27 (GH-D), GH29, GH31 (GH-D), GH35 (GH-A), GH36 (GH-D), GH39 (GH-A), GH42 (GH-A), GH53 (GH-A), GH66, GH97, GH101, GH107 и GH114 (табл. 2). Среди белков бывшего семейства GH36 оказались представители всех 11 ранее выделенных нами семейств GH36A–GH36K, кроме семейства GH36I. Помимо этого были обнаружены белки, содержащие домены ряда других семейств, которые до сих пор экспериментально энзиматически не охарактеризованы, но, как показано нами ранее, они имеют эволюционные связи с некоторыми семействами гликозилгидролаз. Помимо семейств GHL1–GHL15, это семейства COG1306 [18], COG1649 [19] и COG2342 [13]. Обнаружены также белки семейства DUF3111, родство которых с гликозилгидролазами до сих пор не было известно. При этом наибольшее число найденных белков – 12976 – принадлежит семейству GH13 (табл. 2). Среди них оказался лишь один белок, относящийся к подсемейству GH13_33, и ни одного белка из подсемейства GH13_25 (или KOG3625). Ранее мы предложили эти два подсемейства рассматривать в качестве самостоятельных семейств клана GH-H [13]. Из найденных белков только 135 не содержат домены ранее известных семейств в области гомологии с последовательностями-запросами. На основании гомологии эти белки удалось объединить в 26 семейств (табл. 2 и 3), названных нами GHL16–GHL41 (от англ. glycoside hydrolase-like).

Все обсуждаемые выше этапы скрининга базы данных аминокислотных последовательностей проводили 9 февраля 2011 г. Однако в ряде случаев скрининги, проведенные в более ранние сроки, позволили выявить целый ряд других связей между семействами белковых доменов. При этом в качестве запросов тоже использовали только белки семейств GHL1–GHL15, однако не всегда те же самые фрагменты и из тех же белков, что указаны в табл. 1. Большинство обнаруженных дополнительных связей соединяли тот же набор семейств белковых доменов, но в других комбинациях (отмечены знаком “плюс” в табл. 2). Вместе с тем, в некоторых случаях удавалось обнаружить связи и с белками, представляющими дополнительные семейства (т.е. не представленные в табл. 2). При использовании в качестве запросов доменов семейств GHL6 и GHL11 обнаружены представители семейств GH112 гликозилгидролаз, COG1082 предсказанных изомераз/эпимераз фосфосахаров и PF00962 аденозиндезаминаз (табл. 4). Некоторые из найденных белков содержат домены, не относящиеся ни к одному из ранее известных семейств. На основании гомологии они объединены нами в шесть новых семейств GHL42–GHL47 (табл. 3 и 4). Следует отметить, что с использованием трех GHL6-запросов (перечислены в табл. 4) выявлено суммарно пять белков семейства GHL44 (GenPept, EDO54080.1, EDV03574.1, EEF88260.1, EEF88266.1 и EFA21664.1). Кроме того, нами обнаружены белки EEN67543.1 и XP_001813879.1 (табл. 4), однако они, вероятно, не являются гомологами гликозилгидролаз, так как первый из них в области сходства с доменами GHL7 и GHL14 представлен множественным повтором короткого фрагмента, а аминокислотная последовательность второго обладает пониженной сложностью (low complexity).

Среди всех семейств экспериментально не охарактеризованных белковых доменов, обнаруженных нами 9 февраля 2011 г., наиболее многочисленны представители семейства COG1649 – 988 белков (табл. 2). В связи с этим следует отметить, что это семейство очень гетерогенно (данные не приводятся). Его более подробное исследование должно привести к выделению вместо него нескольких более мелких семейств близкородственных белков. Семейства GHL6 и GHL9 очень близки к COG1649 (табл. 2). Домены этих семейств выявляют представителей COG1649 уже по итогам первой итерации: GHL6-домен находит белок ABJ82818.1 (*E*-value = 0.0000006), а GHL9-домен – белок ABC57725.1 (*E*-value = 0.0000009). Более того, при скринингах в более ранние даты (30 ноября 2010 и 4 февраля 2011 г.) тот же GHL6-домен (GenPept, ADE53551.1) находил белок ABJ82818.1 с *E*-value = = 0.00000007, что, согласно используемому нами критерию (см. раздел “анализ данных”), позволяло отнести эти белки к одному семейству, т.е. объединить GHL6 и COG1649. При скрининге 4 февраля 2011 г. тот же GHL9-домен (GenPept, AAM01554.1) нашел белок ABC57725.1 с *E*-value = 0.0000007, что также указывает на близкое родство семейств GHL9 и COG1649. Более того, скрининг базы данных с помощью белков ABJ82818.1 и ABC57725.1 показал, что они представляют две небольшие по численности группы белков, достаточно четко обособленных от остальных представителей семейства COG1649 (данные не приводятся). Это позволяет нам рассматривать эти две группы белков как самостоятельные семейства GHL48 и GHL49 (табл. 3).

Некоторые из исследованных нами новых семейств гипотетических гликозилгидролаз содержат белки из мало изученных отделов бактерий. Пять GHL10-доменов (ABF43386.1, ABJ85023.1, ADV82350.1, ADV82352.1 и ADW71029.1) и один

Таблица 2. Эволюционные связи семейств GHL1–GHL15 с другими семействами белковых доменов по данным скрининга базы данных GenPept программой PSI-BLAST 9 февраля 2011 года

Семейство	GHL1	GHL2	GHL3	GHL4	GHL5	GHL6	GHL7	GHL8	GHL9	GHL10	GHL11	GHL12	GHL13	GHL14	GHL15	Число белков
GH5	–	–	–	–	–	–	–	–	–	–	4	–	+	–	–	1
GH13	3	4	3	2	2	3	+	6	–	3	2	–	4	6	+	12975
GH13_33	5	+	–	–	–	–	–	–	–	–	–	–	–	–	–	1
GH17	–	–	–	–	–	–	–	–	4	–	–	–	–	–	–	1
GH18	–	–	–	–	–	4	–	–	2	–	+	–	4	4	–	5195
GH20	–	–	+	4	–	+	–	–	–	4	3	–	5	–	+	35
GH27	–	6	4	+	–	+	–	+	–	5	–	–	–	+	–	54
GH29	+	6	–	+	–	2	–	+	–	3	4	–	5	–	+	1112
GH31	+	6	2	3	+	4	+	4	–	3	3	–	6	5	–	2777
GH35	–	–	–	+	–	+	–	–	–	4	+	–	6	–	–	35
GH36A	+	4	3	3	–	3	+	6	–	3	2	–	4	7	–	804
GH36B	+	6	3	3	–	+	+	5	–	+	4	–	+	6	–	110
GH36C	–	6	+	–	–	–	–	–	–	–	–	–	–	–	–	1
GH36D	–	6	4	–	–	+	–	–	–	–	–	–	–	–	–	10
GH36E	–	6	4	–	–	+	–	–	–	–	–	–	–	–	–	13
GH36F	–	6	4	4	–	–	–	+	–	+	–	–	–	+	–	2
GH36G	4	5	2	3	–	4	+	6	–	4	4	–	5	6	–	9
GH36H	–	6	3	4	4	3	+	+	–	3	3	–	4	4	–	16
GH36J	–	–	4	–	–	–	–	–	–	–	–	–	–	–	–	1
GH36K	–	–	4	–	–	+	–	–	–	–	–	–	–	+	–	3
GH39	–	–	–	+	–	–	–	–	–	4	–	–	–	–	–	3
GH42	–	–	–	–	–	3	–	–	–	2	4	–	5	–	–	640
GH53	–	–	–	–	–	–	–	–	–	–	–	–	6	–	–	1
GH66	–	–	3	3	–	+	+	5	–	+	4	–	+	4	–	56
GH97	–	–	–	–	–	4	–	–	–	–	–	–	–	–	–	3
GH101	1	3	3	–	–	–	–	–	–	–	–	–	–	–	–	194
GH107	–	–	–	4	–	4	–	–	–	–	+	–	–	–	–	1
GH114	–	–	–	–	–	–	6	–	–	–	–	–	–	6	+	179
COG1306	+	–	4	3	–	3	11	6	2	3	2	–	4	6	–	182
COG1649	5	4	2	1	–	1	9	4	1	2	1	–	2	4	+	988
COG2342	–	6	4	+	–	–	2	+	–	–	–	–	+	3	+	225
DUF3111	2	4	+	–	–	–	–	–	–	–	–	–	–	–	–	133
GHL1	1	2	2	–	–	–	–	–	–	–	–	–	–	–	–	22
GHL2	2	1	2	–	–	–	–	–	–	–	–	–	–	–	–	2
GHL3	2	2	1	3	–	+	–	–	–	+	+	–	5	6	–	11
GHL4	3	4	3	1	–	–	–	–	–	–	4	–	–	–	–	3
GHL5	–	–	–	–	1	–	–	–	–	–	–	–	–	–	–	17
GHL6	–	–	–	–	–	1	–	–	–	4	3	–	5	–	–	3
GHL7	–	–	4	–	–	–	1	2	–	–	–	–	–	2	+	2
GHL8	–	–	–	–	–	–	1	1	–	–	–	–	–	2	–	4

Таблица 2. Окончание

Семейство	GHL1	GHL2	GHL3	GHL4	GHL5	GHL6	GHL7	GHL8	GHL9	GHL10	GHL11	GHL12	GHL13	GHL14	GHL15	Число белков
GHL9	–	–	–	–	–	–	–	–	1	–	–	–	2	7	–	1
GHL10	–	–	–	3	–	3	–	+	–	1	2	–	4	–	–	81
GHL11	–	–	–	3	–	3	–	–	–	2	1	–	5	–	–	3
GHL12	–	–	–	–	–	–	–	–	–	–	+	1	–	–	–	2
GHL13	–	–	–	3	–	3	+	6	–	3	2	–	1	–	–	175
GHL14	–	–	–	–	–	–	2	3	–	–	–	–	–	1	+	12
GHL15	–	–	–	–	–	–	5	–	–	–	–	–	–	4	1	2
GHL16	3	2	–	–	–	–	–	–	–	–	–	–	–	–	–	43
GHL17	5	+	–	4	–	–	–	–	–	–	4	–	–	–	–	3
GHL18	–	5	2	–	–	–	–	–	–	–	–	–	–	–	–	1
GHL19	–	6	–	–	–	+	–	–	–	+	–	–	–	–	–	1
GHL20	–	–	2	–	–	–	–	–	–	–	–	–	–	–	–	1
GHL21	–	–	–	–	–	3	–	–	–	–	–	–	–	–	–	2
GHL22	–	–	–	–	–	–	3	4	–	–	–	–	–	4	–	1
GHL23	–	–	–	–	–	–	4	4	–	–	–	–	–	3	–	4
GHL24	–	–	–	–	–	–	4	3	–	–	–	–	–	3	+	1
GHL25	–	–	–	–	–	–	–	–	4	–	–	–	+	–	–	18
GHL26	–	–	–	–	–	–	–	–	4	–	–	–	–	–	–	2
GHL27	–	–	–	–	–	–	–	–	4	3	3	–	5	–	–	15
GHL28	–	–	–	–	–	–	–	–	4	–	–	–	5	–	–	4
GHL29	–	–	–	–	–	–	–	–	4	–	–	–	–	–	–	1
GHL30	–	–	–	–	–	+	–	–	–	2	3	–	5	–	–	12
GHL31	–	–	–	–	–	+	–	–	–	2	3	–	–	–	–	1
GHL32	–	–	–	–	–	+	–	–	–	3	3	–	6	–	–	5
GHL33	–	–	–	–	–	–	–	–	–	4	–	–	–	–	–	7
GHL34	–	–	–	–	–	–	–	–	–	4	–	–	–	–	–	1
GHL35	–	–	–	–	–	+	–	–	–	4	+	–	–	–	–	1
GHL36	–	–	–	–	–	–	–	–	–	4	–	–	–	–	–	2
GHL37	–	–	–	–	–	+	–	–	–	–	4	–	–	–	–	3
GHL38	–	–	–	–	–	–	–	–	–	–	4	–	–	+	–	1
GHL39	–	–	–	–	–	–	–	–	–	–	4	–	+	–	–	1
GHL40	–	–	–	–	–	–	–	–	–	–	–	–	4	–	–	2
GHL41	–	–	–	–	–	–	–	–	–	–	–	–	–	5	–	2

Примечание. Каждая колонка соответствует одному GHL-домену, использованному в качестве запроса (query), список соответствующих белков приведен в табл. 1. Указано минимальное число итераций, необходимое для выявления белка соответствующего семейства (название строки) с *E*-value ≤ 0.005. Знаки “минус” и “плюс” означают, что белки этого семейства с соответствующим запросом не были обнаружены. В последней колонке указано общее число неидентичных представителей данного семейства, обнаруженных при скрининге хотя бы с одним из 15 запросов. Белок ACU60063.1, наряду с доменом GH18, также содержит домен GH5, а белок EAA58873.1, наряду с доменом GH31, еще содержит домен GH18, однако эти два белка были посчитаны только по одному разу – в составе того семейства, к которому принадлежит домен, соответствующий найденной области гомологии. Целый ряд белков содержит по два и более GH31-домена [17], при подсчете числа белков они считались по одному разу. Знак “плюс” стоит в тех случаях, когда белки соответствующего семейства были найдены при поиске в более ранние сроки (с января 2009 по февраль 2011 г.), при этом могли использоваться не те же самые фрагменты или фрагменты не из тех же самых белков, что 9 февраля 2011 г. (см. текст).

Таблица 3. Белки новых семейств GHL16–GHL50 гипотетических гликозилгидролаз

Белок	Семейство	Организм (название отдела)	Размер белка	Область гомологии	Аннотация (NCBI)
ABX42632.1	GHL16	*Clostridium phytofermentans* (Firmicutes)	869	495–742	hypothetical protein
EAS06878.2	GHL17	*Tetrahymena thermophila* (Alveolata)	1849	464–728	glycosyl transferase, group 1 family protein
EEF89911.1	GHL18	*Bacteroides cellulosilyticus* (Bacteroidetes)	711	270–561	hypothetical protein
EDM27996.1	GHL19	*Lentisphaera araneosa* (Lentisphaerae)	874	400–585	hypothetical protein
EDQ88515.1	GHL20	*Monosiga brevicollis* (Choanoflagellida)	1365	918–1167	predicted protein
ADW74608.1	GHL21	*Rahnella* sp. Y9602 (Proteobacteria)	1617	895–1084	hypothetical protein
EDY96574.1	GHL22	*Bacteroides plebeius* (Bacteroidetes)	414	133–410	»
EDM25438.1	GHL23	*Lentisphaera araneosa* (Lentisphaerae)	409	6–402	»
EDZ68011.1	GHL24	*Nitrosococcus oceani* (Proteobacteria)	590	2–411	»
BAC70894.1	GHL25	*Streptomyces avermitilis* (Actinobacteria)	412	190–369	putative membrane protein
EFC35845.1	GHL26	*Naegleria gruberi* (Heterolobosea)	276	75–249	predicted protein
ACM56874.1	GHL27	*Halorubrum lacusprofundi* (Euryarchaeota)	286	34–252	conserved hypothetical protein
ACX65432.1	GHL28	*Paenibacillus* sp. Y412MC10 (Firmicutes)	408	150–372	copper amine oxidase domain protein
EAS01825.3	GHL29	*Tetrahymena thermophila* (Alveolata)	326	107–219	hypothetical protein
ACZ10290.1	GHL30	*Sebaldella termitidis* (Fusobacteria)	392	4–274	»
EDL56436.1	GHL31	*Planctomyces maris* (Planctomycetes)	477	11–354	»
ADL26670.1	GHL32	*Fibrobacter succinogenes* (Fibrobacteres)	324	12–270	conserved domain protein
ADU31410.1	GHL33	*Bacillus cellulosilyticus* (Firmicutes)	1041	386–644	glycoside hydrolase family 42 domain protein
EDY21786.1	GHL34	*Chthoniobacter flavus* (Verrucomicrobia)	873	274–466	hypothetical protein
ACU74079.1	GHL35	*Catenulispora acidiphila* (Actinobacteria)	723	75–283	»
BAH74220.1	GHL36	*Desulfovibrio magneticus* (Proteobacteria)	789	52–344	»
EFX45617.1	GHL37	*Paenibacillus larvae* (Firmicutes)	477	35–290	»
EEF59093.1	GHL38	*Bacterium Ellin514* (Verrucomicrobia)	610	121–300	»
EFD00658.1	GHL39	*Clostridium hathewayi* (Firmicutes)	592	1–190	conserved hypothetical protein
ADH61837.1	GHL40	*Thermoanaerobacter mathranii* (Firmicutes)	656	426–640	copper amine oxidase domain protein
BAJ49185.1	GHL41	*Candidatus Caldiarchaeum subterraneum*	832	375–783	hypothetical protein
EDM26940.1	GHL42	*Lentisphaera araneosa* (Lentisphaerae)	275	2–87	hypothetical protein
EFB02667.1	GHL43	*Victivallis vadensis* (Lentisphaerae)	1337	927–1113	glycosyl hydrolase BNR repeat-containing protein
EFA21664.1	GHL44	*Bacteroides* sp. D20 (Bacteroidetes)	518	52–191	conserved hypothetical protein
EDP96618.1	GHL45	*Kordia algicida* (Bacteroidetes)	1840	1554–1684	putative cell wall-associated protein precursor
ACU02565.1	GHL46	*Pedobacter heparinus* (Bacteroidetes)	714	100–317	hypothetical protein
CBT77424.1	GHL47	*Arthrobacter arilaitensis* (Actinobacteria)	435	200–405	»
ABJ82818.1	GHL48	*Solibacter usitatus* (Acidobacteria)	674	21–235	»
ABC57725.1	GHL49	*Methanosphaera stadtmanae* (Euryarchaeota)	834	586–819	member of asn/thr-rich large protein family
ADB52725.1	GHL50	*Conexibacter woesei* (Actinobacteria)	1405	547–680	hypothetical protein

Примечание. В первой колонке указан номер аминокислотной последовательности соответствующего белка в базе данных GenPept (NCBI). В колонке “Размер белка” указано суммарное число аминокислотных остатков в белке-предшественнике. В колонке “Область гомологии” даны номера аминокислотных остатков в последовательности белка, которые ограничивают участок, соответствующий исследуемому домену.

Таблица 4. Обнаружение белков, принадлежащих семействам, не выявленным при скрининге 9 февраля 2011 года

Белок-запрос для PSI-BLAST			Результат PSI-BLAST				
семейство	номер белка	фрагмент	дата	итерация	*E*-value	номер белка	семейство
GHL6	ADE53552.1	42–262	1 декабря 2010	5	0.001	EAQ48079.1	COG1082
GHL11	ABZ09704.1	47–283	3 декабря 2010	3	0.004	ADH97994.1	GH112
GHL11	EFB00125.1	169–391	2 декабря 2010	4	0.0005	ABJ82784.1	PF00962
GHL11	EFB00125.1	169–391	2 декабря 2010	4	0.00005	BAH38121.1	PF00962
GHL11	EDY17439.1	57–281	3 декабря 2010	4	0.0003	BAH38121.1	PF00962
GHL3	AAO78700.1	309–546	19 января 2009	2	0.0000003	EDM26940.1	GHL42
GHL3	AAO78700.1	309–546	5 ноября 2009	2	0.000002	EDM26940.1	GHL42
GHL3	AAO78700.1	309–546	19 января 2009	3	0.00007	EDM96603.1	GHL43
GHL3	AAO78700.1	309–546	5 ноября 2009	2	0.002	EDM96603.1	GHL43
GHL6	ADE53552.1	42–262	1 декабря 2010	3	0.005	EFA21664.1	GHL44
GHL6	EEF76716.1	37–246	2 декабря 2010	3	0.00004	EFA21664.1	GHL44
GHL6	ADE53551.1	40–252	4 февраля 2011	5	0.004	EFA21664.1	GHL44
GHL11	EFB00125.1	169–391	2 декабря 2010	4	0.003	EDP96618.1	GHL45
GHL11	EDY17439.1	57–281	3 декабря 2010	4	0.004	ACU02565.1	GHL46
GHL13	AAW77166.1	352–635	4 февраля 2011	5	0.0003	CBT77424.1	GHL47
GHL6	ADE53551.1	1–356	5 октября 2010	4	0.001	XP_001813879.1	–
GHL7	ABU59141.1	1–412	6 октября 2010	20	0.0003	EEN67543.1	–
GHL14	ZP_07359708.1	10–368	3 декабря 2010	9	0.004	EEN67543.1	–

GHL48-домен (ABJ82818.1) обнаружены в белках представителей Acidobacteria. Единственный известный GHL31-домен (EDL56436.1) принадлежит планктомицетам, а у бактерий отдела Verrucomicrobia обнаружены домены семейств GHL3 (EDY17983.1), GHL6 (ADE53551.1 и ADE53552.1), GHL11 (EDY17439.1), GHL12 (ACB73617.1), GHL34 (EDY21786.1), GHL38 (EEF59093.1) и GHL48 (ADI16743.1, EEG18981.1 и EEG19451.1).

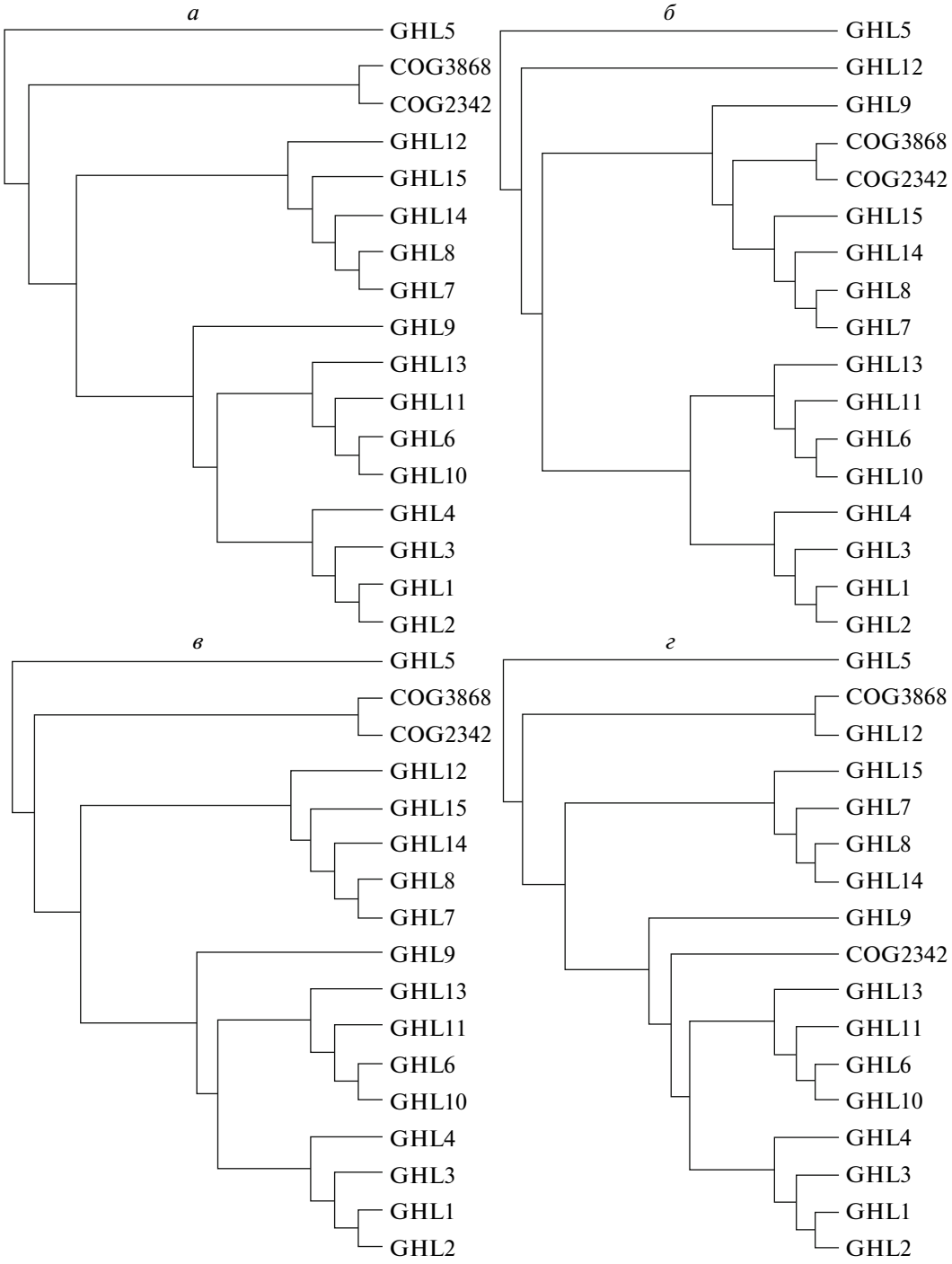
Следует отметить, что в рамках исследования эволюционных связей белков семейства GH114 [15] 9 мая 2010 г. мы провели скрининг базы данных аминокислотных последовательностей с помощью GH114-домена (аминокислотные остатки 38-274) энзиматически не охарактеризованного белка из *Streptomyces* sp. Mg1 (GenPept, EDX26248.1). По итогам девятой итерации программы PSI-BLAST с *E*-value = 0.0003 обнаружен гипотетический белок из актиномицета *Conexibacter woesei* DSM 14684 (ADB52725.1). Вместе с двумя закодированными в том же геноме близкими гомологами (ADB51172.1 и ADB52726.1) этот белок образует новое семейство GHL50 (табл. 3), эволюционно наиболее близкими к которому являются семейства гликозилгидролаз GH36A и GH36B (данные не приводятся).

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Таким образом, итеративный скрининг базы данных аминокислотных последовательностей с помощью программы PSI-BLAST позволил обнаружить эволюционные связи между доменами семейств GHL1–GHL15 или напрямую, или опосредованно через другие семейства (табл. 2). При этом удалось выявить три группы наиболее близкородственных семейств, способных находить друг друга в течение одной-двух итераций: одна из них объединяет семейства GHL1–GHL3, вторая – семейства GHL7, GHL8 и GHL14, а наименее компактная третья – семейства GHL9–GHL11 и GHL13 (табл. 2).

Применение кластерного анализа к данным, приведенным в табл. 2, показывает, что 12 из 15 семейств формируют три стабильных кластера (рисунок). Один из них объединяет семейства GHL1–GHL4, второй – GHL6, GHL10, GHL11 и GHL13, а третий – GHL7, GHL8, GHL14 и GHL15. Однако следует иметь в виду, что применение кластерного анализа к таким данным не имеет строгого обоснования, так как число итераций между двумя семействами отражает, скорее, наличие “переходных форм” в базе данных, чем истинное эволюционное расстояние.

Обнаружены эволюционные связи доменов семейств GHL1–GHL15 с представителями большого числа семейств гликозилгидролаз, имею-



Иерархическая кластеризация семейств GHL1–GHL15 гипотетических гликозилгидролаз. Деревья построены с использованием метода ближайших соседей (NJ) на основании данных, приведенных в табл. 2. Использованы также результаты скрининга базы данных аминокислотных последовательностей с помощью доменов семейств COG2342 и COG3868. Домены семейства COG2342 при скринингах, проведенных с 21 по 31 января 2011 г., находили представителей семейств GHL3–GHL5, GHL7, GHL9–GHL11, GHL13 и GHL14 соответственно за 5, 8, 7, 2, 4, 7, 8, 5 и 2 итерации (неопубликованные данные). Домены семейства COG3868 (GH114) при скринингах, проведенных 21 июля 2010 г., находили представителей семейств GHL3–GHL13 соответственно за 7, 19, 8, 19, 6, 3, 7, 7, 20, 6 и 7 итераций [15]. Домены семейств COG2342 и COG3868 находили друг друга за одну итерацию. В качестве внешней группы условно выбрано семейство GHL5, относящееся к α-глюкозидазному суперсемейству [6]. При проведении кластерного анализа знаки "плюс" и "минус" в табл. 2 заменялись на условные численные значения: "25" и "30" (рис. *а*), "30" и "50" (рис. *б*), "50" и "50" (рис. *в*) и "50" и "100" (рис. *г*) соответственно.

щих пространственную структуру в виде $(\beta/\alpha)_8$-бочонка (табл. 2). Выявлены также связи с доменами COG1306, COG1649 и COG2342 гипотетических гликозилгидролаз, имеющими такую же структуру. Следует отметить, что ранее нами с помощью итеративного скрининга установлено родство следующих доменов: семейства GH13 гликозилгидролаз с белками семейств GHL4 и GHL5 [6], семейства GH101 – с представителями семейств GHL1–GHL5 [16] и семейства GH114 – с членами семейства GHL3–GHL15 [15]. Анализ аминокислотных последовательностей доменов семейств GHL1–GHL15 на сайте базы данных Pfam [20] показал (данные не приводятся), что семейство GHL1 близко к семейству DUF3111, семейство GHL5 – к GH13, семейства GHL6 и GHL10 – к DUF187 (COG1649), а семейства GHL7–GHL9 и GHL14 – к DUF297 (объединяет GH114 и COG2342). Множественное выравнивание аминокислотных последовательностей показало, что у многих представителей семейств GHL1–GHL15 консервативные аминокислотные остатки (Asp и/или Glu) сохраняются в положениях, гомологичных компонентам активного центра у гликозилгидролаз семейств GH101 или GH114 [15, 16]. Вместе взятые, эти результаты позволяют предполагать, что домены семейств GHL1–GHL15 имеют структуру $(\beta/\alpha)_8$-бочонка, кодирующие их гены возникли из генов гликозилгидролаз, а многие белки, содержащие эти домены, обладают какими-то из гликозилгидролазных активностей. Это подтверждает обоснованность включения семейств GHL1–GHL15 в иерархическую классификацию гликозилгидролаз и их гомологов [6].

Результаты, полученные в настоящей работе, впервые позволили проследить эволюционные связи у гликозилгидролаз семейств GH107 и GH112, аденозиндезаминаз семейства PF00962 и белков семейств COG1082 и DUF3111 с обширной группой семейств гликозилгидролаз, имеющих пространственную структуру каталитического домена в виде $(\beta/\alpha)_8$-бочонка (табл. 2 и 4). Следует отметить, что семейство COG1082 проаннотировано как “Sugar phosphate isomerases/epimerases” [21], а в базе данных Pfam соответствующие белки отнесены к семейству PF01261, названному “Xylose isomerase-like TIM barrel” [20]. Однако у единственного, по всей видимости, экспериментально охарактеризованного представителя семейства COG1082 (GenPept, CAB16008.1) показана *мио*-инозоза-2-дегидратазная активность (К.Ф. 4.2.1.44) [22].

Одним из основных результатов настоящей работы является обнаружение 35 новых семейств доменов гипотетических гликозилгидролаз GHL16–GHL50 (табл. 3), имеющих, по-видимому, структуру $(\beta/\alpha)_8$-бочонка. Следует упомянуть, что представителей семейств GHL17 и GHL18 мы находили ранее в процессе поиска гомологов гликозилгидролаз семейства GH13 [13], а представителей семейств GHL30 и GHL39 – при поиске доменов, гомологичных COG2342 (неопубликованные данные).

Полученные нами результаты подтверждают правомочность объединения семейства GH101 с семействами GHL1–GHL3 в одно суперсемейство [6]. На основании близкого родства (табл. 2) в состав этого суперсемейства также могут быть включены семейства DUF3111, GHL16, GHL18 и GHL20. Для этого суперсемейства мы предлагаем название эндо-α-*N*-ацетилгалактозаминидазное суперсемейство – по единственно известной для его представителей энзиматической активности [К.Ф. 3.2.1.97].

Аналогичным образом, на основании данных, полученных ранее [15] и в настоящей работе (табл. 2), семейства GH114, COG2342, GHL7, GHL8 и GHL14 могут быть объединены в составе эндо-α-1,4-полигалактозаминидазного суперсемейства – по названию активности [К.Ф. 3.2.1.109].

Близкое родство семейств COG1649, GHL4, GHL6, GHL9–GHL11, GHL13, GHL30 и GHL31 (табл. 2), а также GHL48 и GHL49 позволяет их рассматривать как одно суперсемейство – “суперсемейство COG1649”.

Как и предлагалось ранее [6], семейство GHL5 целесообразно рассматривать в составе α-глюкозидазного суперсемейства наряду с семействами GH13, GH13_25, GH13_33, GH70 и GH77 гликозилгидролаз.

Семейство GHL50 может быть включено в состав α-галактозидазного суперсемейства, куда, в частности, относятся семейства GH36A и GH36B [6].

Следует отметить, что все пять упомянутые выше суперсемейства – α-галактозидазное, α-глюкозидазное, эндо-α-*N*-ацетилгалактозаминидазное, эндо-α-1,4-полигалактозаминидазное и “суперсемейство COG1649” – находятся между собой в близком родстве и должны быть отнесены к типу II классических $(\beta/\alpha)_8$-гликозилгидролаз в иерархической классификации гликозилгидролаз и их гомологов [6]. Для определения положения в этой классификации семейств GHL12 и GHL15 гипотетических гликозилгидролаз требуются более детальные исследования.

Результаты этого исследования были представлены на 21-ом Международном симпозиуме по гликоконъюгатам [23].

СПИСОК ЛИТЕРАТУРЫ

1. Rebuffet E., Groisillier A., Thompson A., Jeudy A., Barbeyron T., Czjzek M., Michel G. 2011. Discovery and structural characterization of a novel glycosidase family of marine origin. *Environ. Microbiol.* **13**, 1253–1270.

2. Liolios K., Chen I.M., Mavromatis K., Tavernarakis N., Hugenholtz P., Markowitz V.M., Kyrpides N.C. 2010. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucl. Acids Res.* **38**, D346–D354.
3. Heger A., Holm L. 2003. Exhaustive enumeration of protein domain families. *J. Mol. Biol.* **328**, 749–767.
4. Carbohydrate-Active Enzymes server. 2011. (http://www.cazy.org/).
5. Naumoff D.G. 2006. Development of a hierarchical classification of the TIM-barrel type glycoside hydrolases. *Proc. Fifth Int. Conf. Bioinform. Genome Regul. Struct.* July 16–22, 2006. Novosibirsk. Russia. **1**, 294–298. (http://www.bionet.nsc.ru/meeting/bgrs_proceedings/papers/2006/BGRS_2006_V1_067.pdf).
6. Наумов Д.Г. 2011. Иерархическая классификация гликозил-гидролаз. *Биохимия*. **76,** 764–780.
7. Naumoff D.G. 2001. β-Fructosidase superfamily: homology with some α-L-arabinases and β-D-xylosidases. *Prot. Struct. Funct. Genet.* **42**, 66–76.
8. Murzin A.G., Chandonia J.-M., Andreeva A., Howorth D., Lo Conte L., Ailey B.G., Brenner S.E., Hubbard T.J.P., Chothia C. 2011. Structural classification of proteins. (http://scop.mrc-lmb.cam.ac.uk/scop/).
9. Pei J., Grishin N.V. 2005. COG3926 and COG5526: a tale of two new lysozyme-like protein families. *Protein Sci.* **14**, 2574–2581.
10. Наумов Д.Г. 2004. Филогенетический анализ α-галактозидаз семейства GH27. *Молекуляр. биология.* **38**, 463–467.
11. Naumoff D.G. 2004. The α-galactosidase superfamily: sequence based classification of α-galactosidases and related glycosidases. *Proc. Fourth Int. Conf. Bioinform. Genome Regul. Struct.* July 25–30, 2004. Novosibirsk. Russia. **1**, 315–318. (http://www.bionet.nsc.ru/meeting/bgrs_proceedings/papers/2004/BGRS_2004_V1_079.pdf).
12. Naumoff D.G. 2005. GH97 is a new family of glycoside hydrolases, which is related to the α-galactosidase superfamily. *BMC Genomics*. **6**, Art. 112.
13. Gizatullina D.I., Naumoff D.G. 2009. Reclassification of GH13 family of glycoside hydrolases. *Proc. Intern. Moscow Conf. Comput. Mol. Biol.* July 20–23, 2009. Moscow. Russia. P.249–250. (http://mccmb.belozersky.msu.ru/2009/MCCMB09_Proceedings.pdf).
14. Наумов Д.Г., Каррерас М. 2009. Новая программа PSI Protein Classifier автоматизирует анализ результатов программы PSI-BLAST. *Молекуляр. биология*. **43**, 709–721.
15. Наумов Д.Г., Степущенко О.О. 2011. Эндо-α-1,4-полигалактозаминидазы и их гомологи: структура и эволюция. *Молекуляр. биология*. **45**, 703–714.
16. Naumoff D.G. 2010. GH101 family of glycoside hydrolases: subfamily structure and evolutionary connections with other families. *J. Bioinform. Comput. Biol.* **8**, 437–451.
17. Наумов Д.Г. 2007. Структура и эволюция генов мальтазы-глюкоамилазы и сахаразы-изомальтазы млекопитающих. *Молекуляр. биология*. **41**, 1056–1068.
18. Naumoff D.G. 2008. The GH31 family of glycoside hydrolases: subfamily structure and evolutionary connections. *Abstr. Sixth Int. Conf. Bioinform. Genome Regul. Struct.* June 22–28, 2008. Novosibirsk. Russia. P. 169. (http://www.bionet.nsc.ru/meeting/bgrs2008/BGRS2008_Proceedings.pdf#page=169).
19. Kuznetsova A.Y., Naumoff D.G. 2006. Phylogenetic analysis of COG1649, a new family of predicted glycosyl hydrolases. *Proc. Fifth Int. Conf. Bioinform. Genome Regul. Struct.* July 16–22, 2006. Novosibirsk. Russia. **3**, 179–182. (http://www.bionet.nsc.ru/meeting/bgrs_proceedings/papers/2006/BGRS_2006_V3_038.pdf).
20. The Pfam database. 2011. Pfam 25.0. (http://pfam.sanger.ac.uk).
21. Clusters of Orthologous Groups of proteins (COGs). Phylogenetic classification of proteins encoded in complete genomes. 2011. (http://www.ncbi.nlm.nih.gov/COG/).
22. Yoshida K., Yamaguchi M., Ikeda H., Omae K., Tsurusaki K., Fujita Y. 2004. The fifth gene of the *iol* operon of *Bacillus subtilis*, *iolE*, encodes 2-keto-*myo*-inositol dehydratase. *Microbiology*. **150**, 571–580.
23. Naumoff D.G. 2011. New families of TIM-barrel type hypothetical glycoside hydrolases. *Glycoconjugate J.* **28**, 315.