МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2011, том 45, № 4, с. 703–714

БИОИНФОРМАТИКА

УДК 575.852.112:577.152.321

ЭНДО-α-1,4-ПОЛИГАЛАКТОЗАМИНИДАЗЫ И ИХ ГОМОЛОГИ: СТРУКТУРА И ЭВОЛЮЦИЯ

© 2011 г. Д. Г. Наумов[1, 2]*, О. О. Степущенко[3]

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

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

[3]Кафедра генетики Казанского (Приволжского) федерального университета, Казань, 420008

Поступила в редакцию 24.09.2010 г.

Принята к печати 02.11.2010 г.

Эндо-α-1,4-полигалактозаминидаза [К.Ф.3.2.1.109] – редкий фермент, каталитический домен которого относится к семейству гликозилгидролаз GH114. Филогенетический анализ белков этого семейства позволил нам показать, что большую роль в эволюции их генов играли дупликации, элиминации и горизонтальный перенос. Рассматриваются доменный состав, вторичная структура и вероятное строение активного центра эндо-α-1,4-полигалактозаминидаз. При итеративном скрининге белковой базы данных выявлены эволюционные связи семейства GH114 с семействами GH13, GH18, GH20, GH27, GH29, GH31, GH35, GH36 и GH66 гликозилгидролаз, а также с семействами COG1306, COG1649, COG2342, GHL3 и GHL4 энзиматически неохарактеризованных белков. Неклассифицированные гомологи сгруппированы в 13 новых семейств гипотетических гликозилгидролаз: GHL5 – GHL15, GH36J и GH36K.

***Ключевые слова*: гликозилгидролаза, эндо-α-1,4-полигалактозаминидаза, семейство GH114, семейство COG2342, семейство GH36, семейства GHL, новые семейства белков, *PSI-BLAST, PSI Protein Classifier, CAZy*, TIM-бочонок, филогенетическое древо белков, иерархическая классификация белков, эволюция белков, поиск гомологов, множественное выравнивание последовательностей, горизонтальный перенос, аннотация генов.**

ENDO-α-1,4-POLYGALACTOSAMINIDASES AND THEIR HOMOLOGUES: STRUCTURE AND EVOLUTION, by *D. G. Naumoff[1, 2]*, O. O. Stepuschenko[3]* ([1]Vinogradsky Institute of Microbiology, Russian Academy of Sciences, Moscow, 117312, Russia; [2]State Research Center "GosNII Genetika", Moscow, 117545 Russia, *e-mail: daniil_naumoff@yahoo.com; [3]Kazan (Volga Region) Federal University, Department of Genetics, Kazan, 420008 Russia).

Endo-α-1,4-polygalactosaminidase is a rare enzyme. Its catalytic domain belongs to the GH114 family of glycoside hydrolases. Phylogenetic analysis of the family proteins allowed us to show an important role of duplications, eliminations, and horizontal transfer in the evolution of their genes. Domain structure, the secondary structure, and proposed structure of the active center of the endo-α-1,4-polygalactosaminidases are discussed. Evolutionary connections of the GH114 family with GH13, GH18, GH20, GH27, GH29, GH31, GH35, GH36, and GH66 families of glycoside hydrolases, as well as, with COG1306, COG1649, COG2342, GHL3, and GHL4 families of enzymatically uncharacterized proteins have been revealed by iterative screening of the protein database. The unclassified homologues have been grouped into 13 new families of hypothetical glycoside hydrolases: GHL5 – GHL15, GH36J, and GH36K.

***Keywords:* glycoside hydrolase, endo-α-1,4-polygalactosaminidase, GH114 family, COG2342 family, GH36 family, GHL families, new protein families, PSI-BLAST, PSI Protein Classifier, CAZy, TIM-barrel fold, protein phylogenetic tree, protein hierarchical classification, protein evolution, search of homologues, multiple sequence alignment, horizontal transfer, gene annotation.**

Эндо-α-1,4-полигалактозаминидаза – очень редкий фермент. Его энзиматическая активность впервые была обнаружена в штамме бактерии *Streptomyces griseus* C-10 [1], однако фермент так и не был выделен из этого организма в гомогенном виде. Тем не менее, в 1984 г. ему был присвоен классификационный номер [К.Ф.3.2.1.109]. Впоследствии аналогичный фермент обнаружили в штамме *Pseudomonas* sp. 881 и исследовали его биохимические свойства [2, 3]. В 1993 г. секвенировали соответствующий ген (GenBank, D14846.1). При анализе аминокислотной последовательности обнаружился короткий участок локального сходства с некоторыми гликозилгидролазами из семейства GH18 (клан GH-K), но близкие гомологи не были найдены [4]. До настоящего времени эндо-α-1,4-полигалактозаминидаз-

* Эл. почта: daniil_naumoff@yahoo.com

ная активность у каких-либо других организмов не выявлена. Однако при осуществлении геномных проектов в целом ряде организмов (в том числе – в штамме *Streptomyces griseus* NBRC 13350) обнаружились последовательности генов гомологичных белков, которые были объединены в семействе COG3868 [5]. Лишь в 2008 г. оно было признано в качестве семейства GH114 гликозилгидролаз [6].

В настоящее время (24 сентября 2010 г.) в семейство GH114 объединяется 67 белков бактерий и низших эукариот [6]. Пространственная структура ни одного из них не установлена. Однако имеются указания [7] на близкое родство семейства COG3868 и семейства COG2342; трехмерная структура одного из его представителей – белка с неизвестной функцией из *Thermotoga maritima* – была установлена (PDB, 2AAM) в 2005 г. и представляет собой $(\beta/\alpha)_8$-бочонок. Этот тип структуры наиболее распространен среди каталитических доменов гликозилгидролаз [6, 8]. Недавно нами показано, что итеративный скрининг базы данных аминокислотных последовательностей с помощью программы PSI-BLAST с использованием представителей семейств GH31 [9] и GH13 [10] гликозилгидролаз позволяет выявить эволюционное родство этих двух семейств с семействами COG3868 (GH114) и COG2342 соответственно. Это свидетельствует в пользу того, что топологии трехмерных структур каталитических доменов всех упомянутых семейств однотипны.

В настоящей работе проведено сравнение последовательностей белков, содержащих домены семейства GH114 гликозилгидролаз, и осуществлен поиск их гомологов. Предварительные результаты этого исследования были представлены на 35-ом конгрессе FEBS [11] и на Седьмой Международной конференции по биоинформатике регуляции и структуры геномов и системной биологии [12].

АНАЛИЗ ДАННЫХ

Скрининг базы данных аминокислотных последовательностей GenPept на сайте NCBI (http://www.ncbi.nlm.nih.gov/) проводили 21 июля 2010 г. (кроме случаев, специально отмеченных в тексте) при помощи программы PSI-BLAST. В качестве запроса (query) служили GH114-домены пяти белков: BAG23712.1 (использовали фрагмент белка от 42-го аминокислотного остатка до 282-го), EAA59176.1 (85–336), EDK40790.2 (16–262), EDP52951.1 (71–318) и EDX26248.1 (38–271). Поиск вели по каждому из запросов во всех трех неперекрывающихся частях базы данных GenPept: “non-redundant protein sequences” (далее “nr”), “environmental samples” (“env_nr”) и “patented protein sequences” (“pat”). Пороговое значение величины *E*-value для включения найденного белка в модель, используемую программой PSI-BLAST на каждой следующей итерации, составляло 0.005. Результаты обобщали при помощи разработанной нами программы PSI Protein Classifier [9]. В частности, эта программа позволяет определить принадлежность исследуемых белков к ранее известным семействам на основе уровня сходства аминокислотных последовательностей с известными представителями этих семейств.

Множественное выравнивание аминокислотных последовательностей проводили вручную в программе-редакторе BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html); при этом учитывали результаты попарных выравниваний с помощью PSI-BLAST. Если в процессе построения множественного выравнивания в отдельных аминокислотных последовательностях обнаруживались протяженные делеции или отсутствовало сходство в некоторых участках с остальными последовательностями, то анализировали соответствующие нуклеотидные последовательности в трех рамках считывания и (в случае эукариот) уточняли экзон-интронную структуру генов, чтобы, по возможности, повысить уровень сходства анализируемых аминокислотных последовательностей (идентификаторы отредактированных последовательностей на рисунках даны строчными буквами). С этой целью применяли программу tblastn, позволяющую найти наилучшее попарное выравнивание белка (query) с прочитанными в трех рамках нуклеотидными последовательностями.

По результатам множественного выравнивания (после удаления наиболее вариабельных участков последовательностей) строили филогенетические деревья с использованием программ PROTPARS (метод максимальной экономии, Protein Sequence Parsimony method, MP) и NEIGHBOR (метод ближайших соседей, Neighbor-Joining method, NJ) из пакета PHYLIP (http://evolution.gs.washington.edu/phylip.html). Статистическую надежность узлов оценивали путем бутстреп-анализа (1000 псевдореплик для каждого древа). Программу TreeView Win32 (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) применяли для получения графических изображений деревьев. В работе использовали классификацию подразделения гликозилгидролаз, полисахаридлиаз и карбогидратэстераз на семейства и кланы, приведенную на сайте Carbohydrate-Active Enzymes server [6], и классификацию живых организмов, размещенную на сайте NCBI Taxonomy Homepage (http://www.ncbi.nlm.nih.gov/Taxonomy/).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Скрининг “nr” базы данных аминокислотных последовательностей на сайте NCBI с использованием пяти GH114-доменов в качестве запроса (query) позволил после первой итерации программы PSI-BLAST обнаружить, в общей сложности, 185 неидентичных белков (с *E*-value $\leq$ 0.005). Из них 160 содержат домены семейства GH114, а остальные 25 белков – домены семейства COG2342. Ана-

логичные скрининги баз данных "env_nr" и "pat" добавили (по итогам первых итераций) в общей сложности еще семь и два белка соответственно. Каждый из этих девяти белков имеет домен семейства GH114. Таким образом, всего нами обнаружено 169 белков, содержащих, как минимум, по одному домену семейства GH114. Среди содержащих их организмов – представители ряда таксономических групп бактерий (Actinobacteria, Aquificae, Chloroflexi, Deferribacteres, Deinococci и Proteobacteria), грибы (аскомицеты и базидиомицеты), бурые (*Ectocarpus siliculosus*) и зеленые (*Chlamydomonas reinhardtii* и *Volvox carteri*) водоросли, оомицеты (*Phytophthora infestans*) и инфузории (*Tetrahymena thermophila*). Обращает на себя внимание отсутствие белков этого семейства у бактерий отдела Firmicutes, высших растений, животных и архей. Следует отметить, что из девяти видов дрожжей-аскомицетов, геномы которых представлены в базе данных Génolevures, лишь *Debaryomyces hansenii* имеет белок семейства GH114 [13].

Размеры подавляющего большинства обнаруженных нами белков попадают в диапазон 251–375 а.о., все они содержат лишь один домен. Белок из бактерии *Hahella chejuensis* (GenPept, ABC28688.1) содержит два домена семейства GH114. Два белка из *Ectocarpus siliculosus* (CBJ30803.1 и CBJ30802.1) кодируются соседними генами; первый из белков имеет три GH114-домена, а второй – один. При этом последний домен первого белка и второй белок являются двумя фрагментами одного и того же домена. Суммарно, эти два белка содержат три полноразмерных GH114-домена. Белок из бактерии *Endozoicomonas persephone* (ZP_02533448.1) содержит, наряду с GH114-доменом, еще домен семейства GH9 гликозилгидролаз. У нескольких белков обнаружены дополнительные N-концевые домены. Так, два белка из зеленых водорослей (EDP04954.1 и EFJ52241.1) содержат домены семейства Spherulin4 (или PF12138) [14]. Два белка из бактерий отдела Proteobacteria (ABB75727.1 и CAN94233.1) содержат домен с неизвестной функцией, названный нами GH114_assoc. Его удалось также обнаружить еще у двух энзиматически неохарактеризованных белков (ABF90472.1 и CAN96527.1), которые, наряду с ним, содержат домены COG2342 и полисахаридлиаз PL9 соответственно.

Из дальнейшего анализа были исключены белки, имеющие высокую степень сходства (95% и более идентичных аминокислотных остатков), и короткие фрагменты белков (частичные последовательности). Последовательности остальных 138 GH114-доменов из 136 белков подвергали процедуре множественного выравнивания (его наиболее консервативные фрагменты частично приведены на рис. 1). Полный список 136 белков представлен на рис. 2 и 3. Во множественное выравнивание также включили 41 представителя (список приведен в подписи к рис. 2) семейства COG2342 – эволюционно наиболее близкого к семейству GH114 [7, 11, 12].

Известно, что ключевую роль в активном центре подавляющего большинства гликозилгидролаз играют два карбоксил-содержащих аминокислотных остатка [6, 15–17]. При множественном выравнивании исследуемых белков (рис. 1) выявилось два высококонсервативных в семействах GH114 и COG2342 остатка (Asp и Glu), расположенных на С-конце предсказанных четвертого и шестого β-слоев $(\beta/\alpha)_8$-бочонка. В гомологичных положениях у гликозилгидролаз клана GH-D (семейства GH27, GH31 и GH36) находятся соответственно нуклеофил и донор/акцептор протона [18–20]; можно предположить, что эти два аминокислотных остатка играют аналогичную роль в активном центре эндо-α-1,4-полигалактозаминидаз.

На основании множественного выравнивания после удаления наиболее вариабельных участков были построены филогенетические деревья семейства GH114 гликозилгидролаз. В качестве внешней группы использовали 41-го представителя семейства COG2342. Деревья, построенные методами ближайших соседей (рис. 2) и максимальной экономии (рисунок не приводится), имеют сходную топологию. На них могут быть выделены три крупных кластера GH114-доменов (рис. 2 и 3). Два из них (I и II), имея очень высокую бутстреп-поддержку (более 92% на обоих деревьях), содержат почти исключительно белки из бактерий отдела Actinobacteria. Исключение составляют два белка из кластера I: один – из базидиомицета *Schizophyllum commune* (GenPept, EFJ00854.1) и второй – из α-протеобактерии *Magnetospirillum magnetotacticum* (ZP_00047953.1). При этом второй белок содержит неполноразмерный GH114-домен, а последовательность его гена, вероятно, является следствием технической ошибки (загрязнения) при секвенировании полного генома соответствующей бактерии (т.к. соответствующий участок ДНК не удалось картировать в геноме). Скорее всего, этот белок принадлежит какой-то бактерии из отдела Actinobacteria.

Третий крупный кластер (III), образованный всеми белками грибов-аскомицетов (рис. 2 и 3), формируется на обоих деревьях, однако имеет очень низкую бутстреп-поддержку (37.9 и 21.9% на NJ- и MP-деревьях соответственно), что свидетельствует о малой достоверности его обособления от обширной группы других белков бактериального и эукариотического происхождения. Еще один кластер с бутстреп-поддержкой 72.4 и 73.0% (на NJ- и MP-деревьях соответственно) образован пятью (из семи) GH114-доменами бурой водоросли *Ectocarpus siliculosus* (рис. 2).

Анализ взаимного расположения ветвей на филогенетических деревьях позволяет констатировать, что в эволюции генов, кодирующих белки семейства GH114 гликозилгидролаз, велика роль эли-

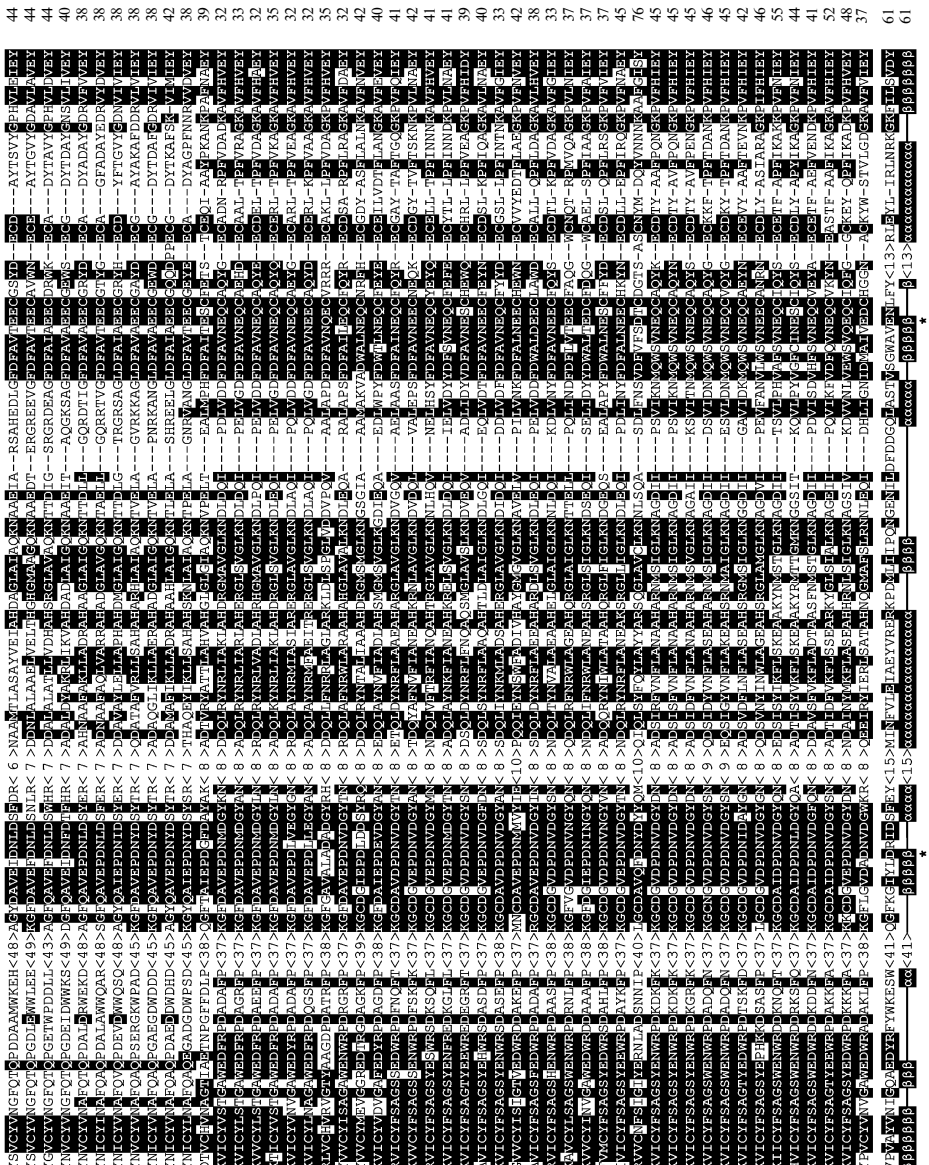

Рис. 1. Фрагменты множественного выравнивания аминокислотных последовательностей гликозилгидролаз семейства GH114. Слева указаны идентификаторы белков согласно базе данных GenPept (NCBI). Пробелы вставлены для улучшения выравнивания. Цифрами указаны расстояния до N- и C-концов белков и длина протяженных инсерций, не указанных на рисунке. Высококонсервативные аминокислотные остатки выделены черным. Звездочками над и под выравниванием отмечены два аминокислотных остатка (Asp и Glu) играющие, вероятно, в активном центре фермента роль нуклеофила и донора/акцептора протона соответственно. Внизу рисунка представлена аминокислотная последовательность (AAD36481.1) белка семейства COG2342 и его вторичная структура согласно базе данных трехмерных структур белков (PDB, 2AAM): α – α-спирали, β – β-слои.

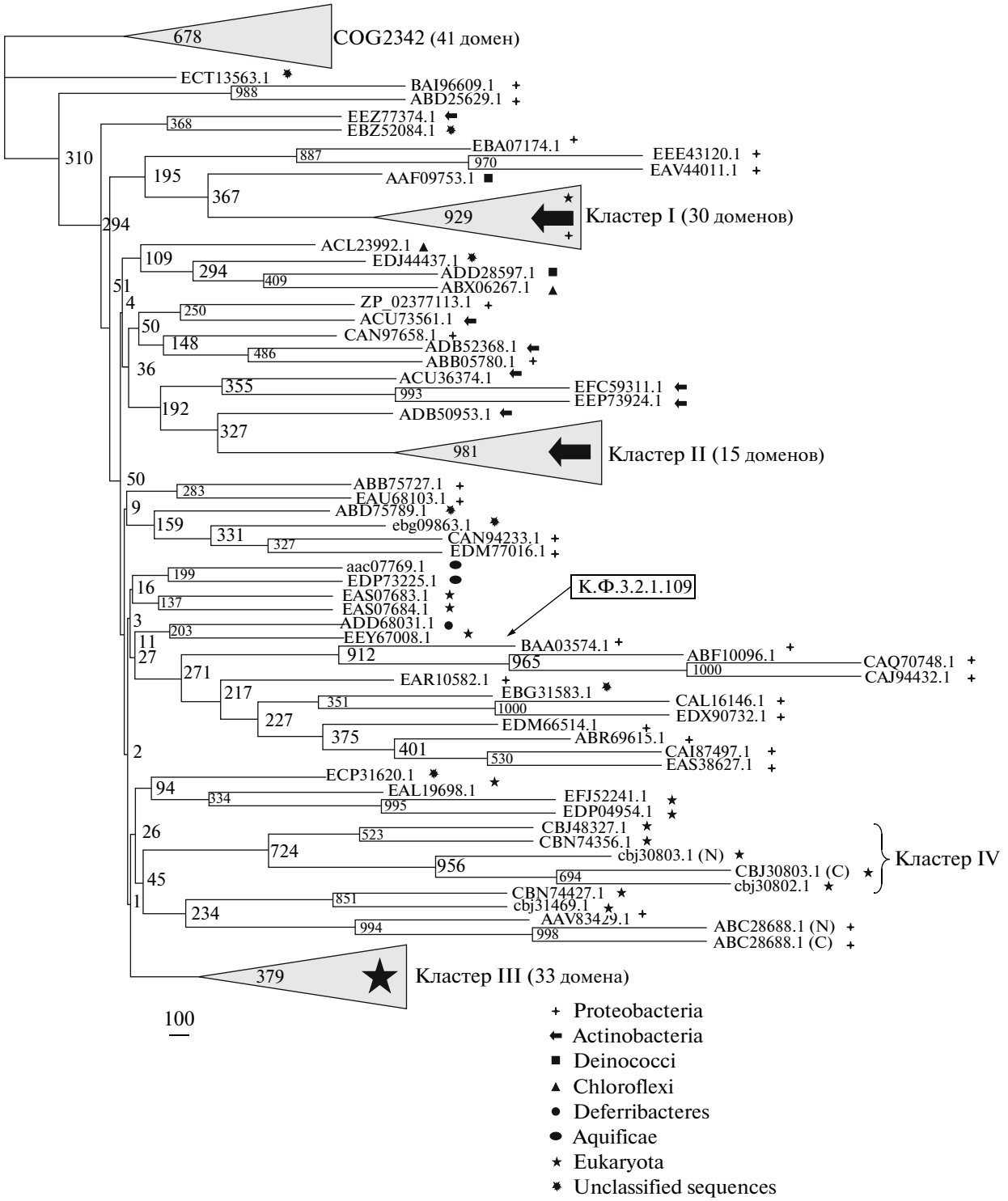

Рис. 2. Схема филогенетического древа семейства GH114 гликозилгидролаз, построенного NJ-методом (ближайших соседей). Статистическую надежность узлов древа оценивали с помощью бутстреп-анализа, около каждого узла указано число подтверждающих псевдореплик из 1000. Группа из 41-го белка семейства COG2342 (GenPept: AAB99489.1, AAD36481.1, AAF10284.1, AAM02338.1, AAM02400.1, AAS44482.1, ABD24977.1, ABD26061.1, ABD26936.1, ABE30011.1, ABF11022.1, ABF90472.1, ABK45762.1, ABP72922.1, ABP72931.1, ABR30852.1, ACB07361.1, ACB09764.1, ACC72301.1, ACD28534.1, ACJ16173.1, ACK41676.1, ACM23260.1, ACN98302.1, ACR79406.1, ACR79895.1, ACS33039.1, ACS63332.1, ACS80746.1, ADD28593.1, ADI31399.1, CAD15948.1, EDP68671.1, EDP76229.1, EDQ03526.1, EEA95822.1, EEB74966.1, EEO40203.1, EEP04836.1, EEP24476.1 и EFG70166.1) взята в качестве внешней. Строение кластеров ветвей I, II и III представлено на рис. 3. Для каждого из белков с помощью условных знаков указано таксономическое положение организма-хозяина. Буквами "N" и "C" обозначены N- и C-концевые GH114-домены у белков ABC28688.1 и CBJ30803.1. Строчными буквами (например, cbj31469.1) отмечены идентификаторы белков, аминокислотные последовательности которых подверглись редактированию при множественном выравнивании. Классификационным номером (К.Ф. 3.2.1.109) отмечен единственный биохимически охарактеризованный белок семейства (BAA03574.1).

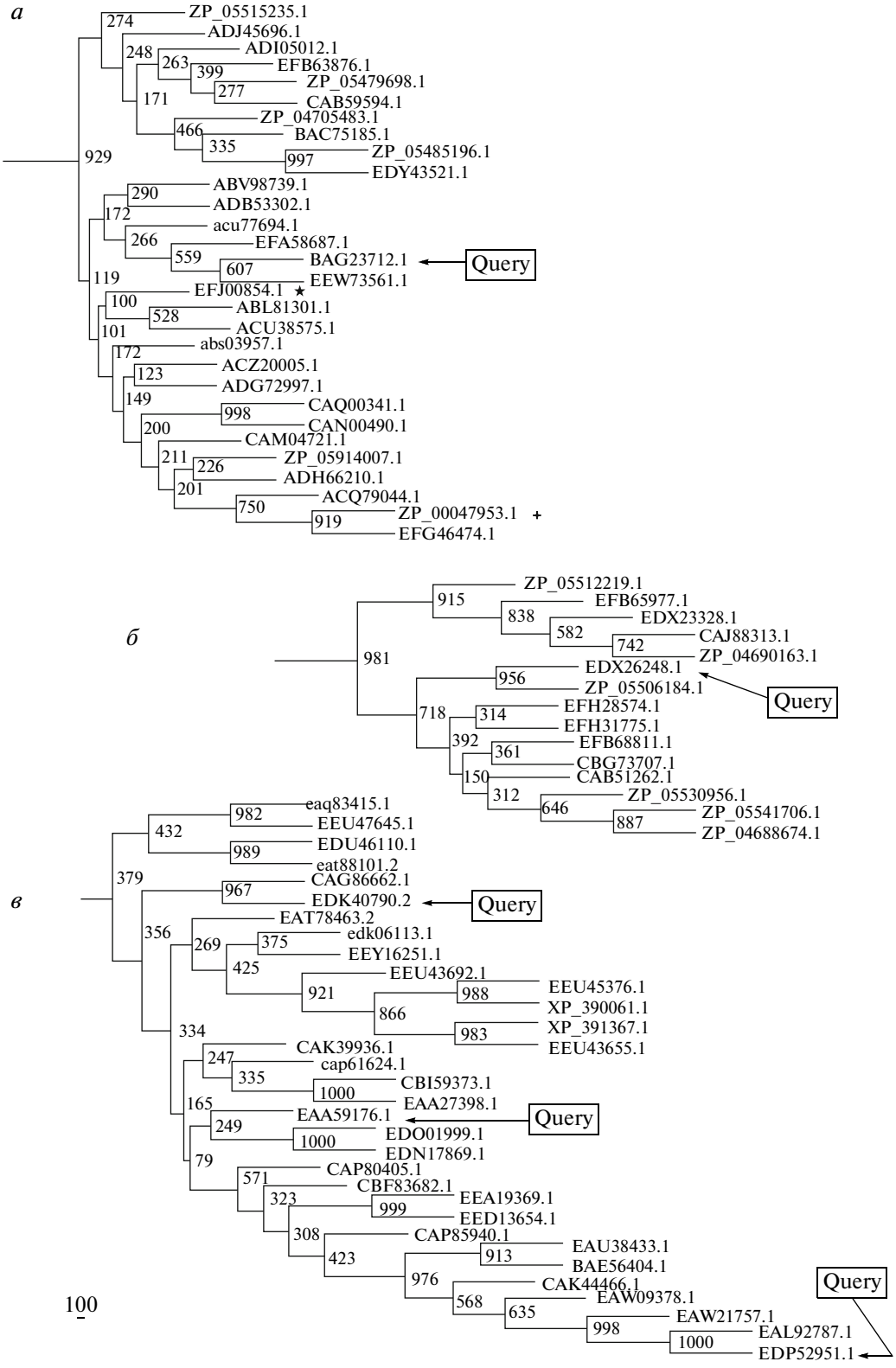

Рис. 3. Фрагменты филогенетического древа семейства GH114 гликозилгидролаз, построенного NJ-методом (ближайших соседей). Фрагменты *а*, *б* и *в* соответствуют кластерам ветвей I, II и III, изображенным на рис. 2. Обозначением "Query" отмечено пять белков, использованных в качестве запросов при скрининге базы данных с помощью программы PSI-BLAST (см. табл. 1). Остальные обозначения – см. рис. 2.

минаций, дупликаций и горизонтального переноса. В частности, многочисленные белки протеобактерий не формируют крупных стабильных кластеров на деревьях (рис. 2).

Для исследования эволюционных связей между белками семейства GH114 и белками других семейств проводили итеративный скрининг базы данных аминокислотных последовательностей с помощью программы PSI-BLAST с использованием пяти GH114-доменов в качестве запроса. Было проведено не менее чем по 8 итераций скрининга "nr"-базы данных с каждым из пяти доменов: 20 итераций с GH114-доменом белка BAG23712.1, 10 – с EDK40790.2, 9 – с EDP52951.1 и по 8 – с EAA59176.1 и EDX26248.1. Число итераций определялось техническими ограничениями интернет-версии программы PSI-BLAST. Скрининг баз данных "env_nr" и "pat" проводили до прекращения появления новых белков (в зависимости от запроса – от трех до семи и две итерации соответственно). В двух последних базах данных удалось найти суммарно лишь 23 белка (с E-value ≤ 0.005), все они принадлежат к семействам GH114 и COG2342.

Скрининг базы данных "nr" позволил обнаружить, в общей сложности, 14732 неидентичных белка. Большинство из них в области гомологии содержит домены ранее известных семейств гликозилгидролаз: GH13 (клан GH-H), GH18 (GH-K), GH20 (GH-K), GH27 (GH-D), GH29, GH31 (GH-D), GH35 (GH-A), GH36 (GH-D), GH66 и GH114 (табл. 1). Среди белков бывшего семейства GH36 оказались представители ранее выделенных нами семейств GH36A, GH36B, GH36E, GH36F, GH36G и GH36H [8, 9], а также двух новых – GH36J (GenPept, ADB29787.1) и GH36K (CBL19648.1). Кроме того, были обнаружены белки, содержащие домены ряда других семейств, которые до сих пор энзиматически не охарактеризованы в эксперименте, но для которых нами были показаны ранее эволюционные связи с некоторыми семействами гликозилгидролаз (семейства COG1306 [21], COG1649 [22], COG2342 [10], GHL3 [23] и GHL4 [23]). При этом наибольшее число найденных белков – 10186 – принадлежит к семейству GH13 (табл. 1). Среди них нет ни одного белка, относящегося к подсемействам GH13_25 (или COG3625) и GH13_33, которые ранее мы предложили рассматривать в качестве самостоятельных семейств клана GH-H [10].

Из найденных только 22 белка не содержат домены ранее известных семейств в области гомологии с последовательностями-запросами. На основании гомологии эти белки удалось объединить в девять семейств (табл. 1 и 2), названных нами GHL5 – GHL13 (от англ. Glycoside Hydrolase-Like). Один из этих белков (GenPept, ACB73617.1), отнесенный нами к семейству GHL12, также содержит домен семейства GH62 гликозилгидролаз, а четыре белка (BAE70447.1, AAW77166.1, ZP_06485502.1 и ZP_06488656.1), отнесенные к семейству GHL13, содержат домен CE4 карбогидратэстераз.

Все обсуждаемые выше результаты скрининга баз данных аминокислотных последовательностей проводили 21 июля 2010 г. Однако в ряде случаев скрининги, проведенные в более ранние сроки, позволили выявить в базе данных "nr" некоторые дополнительные белки. Среди них оказались белки, содержащие домены, не относящиеся ни к одному из известных семейств. Так, при использовании GH114-домена белка EAA59176.1 9 мая и 21 февраля 2010 г. по итогам четвертой итерации обнаружен белок ACL40874.1 (E-value = 0.001 и 0.004 соответственно), а по итогам пятой итерации – белок ABM10768.1 (E-value = 0.00007 и 0.00008 соответственно). На основании гомологии эти два белка объединены нами в новое семейство GHL14 (табл. 2). Использование в качестве запроса GH114-домена белка EDK40790.2 позволило, по итогам шестой итерации 4 февраля 2010 г., обнаружить белок BAH54049.1 (E-value = 0.002), который мы рассматриваем в качестве одного из представителей нового семейства GHL15 (табл. 2).

Следует отметить, что некоторые белки семейства GHL5 нам удавалось ранее выявлять с помощью программы PSI-BLAST при использовании в качестве запроса каталитических доменов семейств GH13 [10] и GH101 [23] гликозилгидролаз.

Множественное выравнивание аминокислотных последовательностей доменов семейств GH114 и COG2342 с представителями новых семейств гипотетических гликозилгидролаз (GHL3 – GHL15, GH36J и GH36K) показало, что большинство из них имеют консервативные остатки Asp и Glu на С-конце предсказанных четвертого и шестого β-слоев $(\beta/\alpha)_8$-бочонка (рис. 4). Эти данные позволяют, во-первых, предположить общность механизма катализа (с вероятным сохранением оптической конфигурации гликозидной связи у продукта реакции) и, во-вторых, сходство расположения каталитически важных остатков (на С-конце предсказанных четвертого и шестого β-слоев) в большинстве новых семейств. Следует отметить, что у белков семейств GHL3 и GHL4 предполагаемые нуклеофил и донор/акцептор протона находятся на С-конце четвертого и пятого β-слоев $(\beta/\alpha)_8$-бочонка, подобно гликозилгидролазам семейства GH101 [23].

Полученные в работе данные поддерживают ранее высказанную нами гипотезу [8], что все каталитические домены гликозилгидролаз с трехмерной структурой в виде $(\beta/\alpha)_8$-бочонка имеют общее эволюционное происхождение.

Уже после прохождения стадии рецензирования данной статьи, 21 октября 2010 г. мы повторили скрининг "nr"-базы данных аминокислотных последовательностей на сайте NCBI с использованием пяти GH114-доменов в качестве запроса. По итогам первых двух итераций программы PSI-

Таблица 1. Эволюционные связи GH114 с другими семействами белковых доменов по данным скрининга “nr” базы данных программой PSI-BLAST

Семейство	BAG23712.1	EAA59176.1	EDK40790.2	EDP52951.1	EDX26248.1	Число белков
GH13	17	5	6	6	5	10186
GH18			3	3	3	6
GH20	20	7		9		2
GH27	19	7	10	8	6	27
GH29	19	8		9	7	8
GH31	17	4	5	5	3	2286
GH35		7				1
GH36A	17	6	10	8	6	693
GH36B	18	5	8	4	5	98
GH36E					8	1
GH36F	20	6	9	8	6	6
GH36G	18	7	10	7	6	7
GH36H	17	6	9	6	5	13
GH36J		8			7	1
GH36K		8		9		1
GH66	17	4	10	6	5	40
GH114	1	1	1	1	1	160
COG1306	17	2	4	3	7	152
COG1649	6	4	8	5	5	817
COG2342	2	1	1	1	1	203
GHL3	18	7		8	7	1
GHL4	19					1
GHL5	20			9	8	5
GHL6	19					1
GHL7		6		6		2
GHL8		3	8	3	3	1
GHL9				7		1
GHL10	18				7	6
GHL11	20					1
GHL12			10	6		1
GHL13					7	4

Примечание. Каждая колонка соответствует одному GH114-домену, использованному в качестве запроса (query). Указано минимальное число итераций, необходимое для выявления белка соответствующего семейства с *E*-value ≤ 0.005. Бывшее семейство GH36 рассматривается как группа из 11 самостоятельных семейств GH36A – GH36K. COG1306, COG1649, COG2342, а также GHL3 – GHL13 – семейства гипотетических гликозилгидролаз. В последней колонке указано общее число неидентичных представителей данного семейства, обнаруженных при скрининге хотя бы с одним из пяти запросов.

Таблица 2. Белки новых семейств гипотетических гликозилгидролаз, гомологичных эндо-α-1,4-полигалактозаминидазам

Белок	Семейство	Организм (отдел прокариот)	Размер белка	Область гомологии	Аннотация (NCBI)
BAH37732.1	GHL3	*Gemmatimonas aurantiaca* (Gemmatimonadetes)	785	358–575	hypothetical protein
EFB01375.1	GHL4	*Victivallis vadensis* (Lentisphaerae)	1210	546–792	hypothetical protein
ACB52584.1	GHL5	*Cyanothece* sp. (Cyanobacteria)	845	298–533	unknown
EDX77985.1	GHL5	*Microcoleus chthonoplastes* (Cyanobacteria)	892	346–581	hypothetical protein
EDY37787.1	GHL5	*Cyanobium* sp. (Cyanobacteria)	843	318–553	conserved hypothetical protein
EAQ75221.1	GHL5	*Synechococcus* sp. (Cyanobacteria)	850	312–547	hypothetical protein
EAZ89764.1	GHL5	*Cyanothece* sp. (Cyanobacteria)	845	298–533	hypothetical protein
ADE53551.1	GHL6	*Coraliomargarita akajimensis* (Verrucomicrobia)	356	85–249	hypothetical protein
ADE53552.1*	GHL6	*Coraliomargarita akajimensis* (Verrucomicrobia)	389	95–262	hypothetical protein
EEF76716.1*	GHL6	*Bacteroides coprophilus* (Bacteroidetes)	363	84–246	hypothetical protein
ABQ90769.1	GHL7	*Roseiflexus* sp. (Chloroflexi)	434	73–363	hypothetical protein
ABU59141.1	GHL7	*Roseiflexus castenholzii* (Chloroflexi)	412	51–291	conserved hypothetical protein
ACU02990.1	GHL8	*Pedobacter heparinus* (Bacteroidetes)	392	18–295	hypothetical protein
ZP_02383235.1*	GHL8	*Burkholderia ubonensis* (Proteobacteria)	400	46–300	hypothetical protein
AAM01554.1	GHL9	*Methanopyrus kandleri* (Euryarchaeota)	389	65–216	predicted membane protein
ABY39946.1	GHL10	*Brucella suis* (Proteobacteria)	393	83–252	hypothetical protein, conserved
ABJ85023.1	GHL10	*Candidatus Solibacter usitatus* (Acidobacteria)	493	130–283	conserved hypothetical protein
ACS32889.1	GHL10	*Thermococcus gammatolerans* (Euryarchaeota)	403	85–253	conserved hypothetical protein
ABF43386.1	GHL10	*Candidatus Koribacter versatilis* (Acidobacteria)	494	124–286	conserved hypothetical protein
EEY02255.1	GHL10	*Brucella neotomae* (Proteobacteria)	389	79–248	conserved hypothetical protein
ZP_05449056.1	GHL10	*Brucella neotomae* (Proteobacteria)	393	83–252	hypothetical protein

Таблица 2. Окончание

Белок	Семейство	Организм (отдел прокариот)	Размер белка	Область гомологии	Аннотация (NCBI)
EFB00125.1	GHL11	*Victivallis vadensis* (Lentisphaerae)	519	223–391	hypothetical protein
ABZ09704.1*	GHL11	uncultured marine crenarchaeote (Thaumarchaeota)	393	79–260	hypothetical protein
EDY17439.1*	GHL11	*Chthoniobacter flavus* (Verrucomicrobia)	715	104–281	PBS lyase HEAT domain protein repeat-containing protein
ACB73617.1	GHL12	*Opitutus terrae* (Verrucomicrobia)	1221	734–919	α-N-arabinofuranosidase
EFL62392.1*	GHL12	*Acetivibrio cellulolyticus* (Firmicutes)	929	446–631	carbohydrate binding family 11
AAW77166.1	GHL13	*Xanthomonas oryzae* (Proteobacteria)	663	402–589	HmsF protein
BAE70447.1	GHL13	*Xanthomonas oryzae* (Proteobacteria)	632	371–558	HmsF protein
ZP_06485502.1	GHL13	*Xanthomonas campestris* (Proteobacteria)	629	368–555	HmsF
ZP_06488656.1	GHL13	*Xanthomonas campestris* (Proteobacteria)	629	368–555	HmsF
ACL40874.1*	GHL14	*Arthrobacter chlorophenolicus* (Actinobacteria)	360	6–263	conserved hypothetical protein
ABM10768.1*	GHL14	*Arthrobacter aurescens* (Actinobacteria)	360	6–263	hypothetical protein
BAI63890.1*	GHL14	*Rothia mucilaginosa* (Actinobacteria)	398	16–305	ABC-type sugar transport system, ATPase component
ZP_07359708.1*	GHL14	*Actinomyces viscosus* (Actinobacteria)	369	10–271	hypothetical protein
BAH54049.1*	GHL15	*Rhodococcus opacus* (Actinobacteria)	460	44–283	hypothetical protein
CAO81213.1*	GHL15	*Candidatus Cloacamonas acidaminovorans* (candidate division WWE1)	635	21–248	hypothetical protein
ADB29787.1	GH36J	*Kribbella flavida* (Actinobacteria)	569	256–521	hypothetical protein
CBL19648.1	GH36K	*Ruminococcus* sp. (Firmicutes)	561	201–500	α-galactosidase

Примечание. В первой колонке указан номер соответствующей аминокислотной последовательности белка в базе данных GenPept (NCBI). Звездочкой отмечены номера белков, не обнаруженных программой PSI-BLAST ни с одним из запросов при скрининге 21 июля 2010 г. 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, гомологичный доменам семейства GH114.

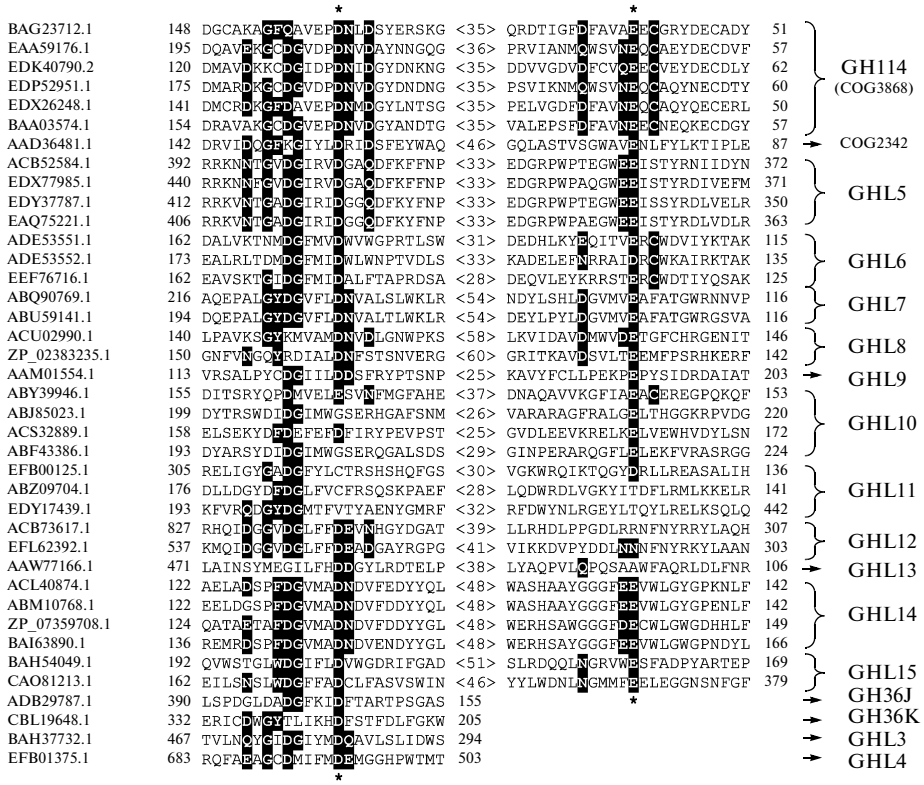

Рис. 4. Два участка аминокислотных последовательностей представителей новых семейств гипотетических гликозилгидролаз (GHL3 – GHL15, GH36J и GH36K), гомологичных участкам последовательностей домена GH114, содержащим остатки Asp и Glu, которые рассматриваются в качестве возможных компонентов активного центра – нуклеофила и донора/акцептора протона соответственно (отмечены звездочками сверху и снизу рисунка). Высококонсервативные аминокислотные остатки выделены черным. Слева указаны идентификаторы белков в базе данных GenPept (NCBI). Цифрами указаны расстояния до N- и С-концов белков и длина протяженных инсерций, не приведенных на рисунке. Справа указана принадлежность белков к определенным семействам. В случае белков семейств GHL3, GHL4, GH36J и GH36K указан только один фрагмент, т.к. для второго не удалось получить надежного выравнивания. Белки семейств GHL3 и GHL4, повидимому, содержат донор/акцептор протона в негомологичном положении аминокислотной последовательности [23].

BLAST нам удалось обнаружить, в общей сложности, 317 неидентичных белков (с E-value ≤ 0.005). Из них 167 содержат домены семейства GH114, 148 – домены семейства COG2342, а оставшиеся два белка принадлежат к семействам COG0441 (GenPept, CAH04860.1) и COG1306 (BAD40253.1). Таким образом, число белков, содержащих домены семейства GH114, в "nr"-базе данных за три месяца увеличилось с 160 до 167.

Согласно версии базы данных CAZy [6] от 13 октября 2010 г., семейство GH114 содержит лишь 67 белков. Анализ показал, что три из них (AAF10284.1, CBJ50540.1 и EAR02555.1) на самом деле принадлежат к семейству COG2342. Еще девять представителей семейства COG2342 (ABM41017.1, ABX38175.1, ADI31399.1, ADK67932.1, CBJ37308.1, CBJ37339.1, CBJ42422.1, CBJ42450.1 и EAR02552.1) оказались в списке неклассифицированных гликозилгидролаз (известных также как условное семейство GH0). Согласно последней версии базы данных CAZy [6] от 10 марта 2011 г., семейство GH114 содержит уже 81 белок.

СПИСОК ЛИТЕРАТУРЫ

1. Reissig J.L., Lai W.H., Glasgow J.E. 1975. An endogalactosaminidase from *Streptomyces griseus*. *Can. J. Biochem*. **53**, 1237–1249.
2. Tamura J.-I., Takagi H., Kadowaki K. 1988. Purification and some properties of the endo α-1,4 polygalactosaminidase from *Pseudomonas* sp. *Agric. Biol. Chem*. **52**, 2475–2484.
3. Tamura J.-I., Abe T., Hasegawa K., Kadowaki K. 1992. The mode of action of endo α-1,4 polygalactosaminidase from *Pseudomonas* sp. 881 on galactosaminooligosaccharides. *Biosci. Biotech. Biochem*. **56**, 380–383.
4. Tamura J.-I., Hasegawa K., Kadowaki K., Igarashi Y., Kodama T. 1995. Molecular cloning and sequence analysis of the gene encoding an endo α-1,4 polygalactosaminidase of *Pseudomonas* sp. 881. *J. Ferment. Bioeng*. **80**, 305–310.

5. Clusters of Orthologous Groups of proteins (COGs). 2010. Phylogenetic classification of proteins encoded in complete genomes. (http://www.ncbi.nlm.nih.gov/COG/).
6. Carbohydrate-Active Enzymes server (CAZy). 2011. (http://www.cazy.org/).
7. Iyer L.M., Aravind L., Bork P., Hofmann K., Mushegian A.R., Zhulin I.B., Koonin E.V. 2001. *Quoderat demonstrandum?* The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biol*. **2**, research0051.
8. Naumoff D.G. 2006. Development of a hierarchical classification of the TIM-barrel type glycoside hydrolases. *Proc. Fifth Internat. Conf. Bioinformat. Genome Regulat. Struct*. July 16–22, 2006. Novosibirsk. Russia. **1**, 294–298. (http://www.bionet.nsc.ru/meeting/bgrs2006/BGRS_2006_V1.pdf).
9. Наумов Д.Г., Каррерас М. 2009. Новая программа PSI Protein Classifier автоматизирует анализ результатов программы PSI-BLAST. *Молекуляр. биология*. **43**, 709–721.
10. Gizatullina D.I., Naumoff D.G. 2009. Reclassification of GH13 family of glycoside hydrolases. *Proc. Internat. Moscow Conf. Computat. Mol. Biol.* July 20–23, 2009. Moscow. Russia. P. 249–250. (http://mccmb.belozersky.msu.ru/2009/MCCMB09_Proceedings.pdf).
11. Naumoff D.G., Stepuschenko O.O. 2010. Endo-α-1,4-polygalactosaminidase structure and evolution. *FEBS J*. **277**(S1), 233–234.
12. Stepuschenko O.O., Naumoff D.G. 2010. Sequence analysis of COG3868 and COG2342 families. *Abstr. Seventh Internat. Conf. Bioinformat. Genome Regulat. Struct. and Systems Biol*. June 20–27, 2010. Novosibirsk. Russia. P.193.
13. Naumoff D.G. 2010. Sequence analysis of yeast glycoside hydrolases. *Abstr. Seventh Internat. Conf. Bioinformat. Genome Regulat. Struct. and Systems Biol*. June 20–27, 2010. Novosibirsk. Russia. P. 194. (http://conf.nsc.ru/files/conferences/BGRSSB2010/abstracts/11493/11766/Naumoff_BGRS'2010.doc).
14. The Pfam database (Pfam 24.0). 2009. (http://pfam.sanger.ac.uk).
15. Henrissat B., Davies G. 1997. Structural and sequence-based classification of glycoside hydrolases. *Curr. Opin. Struct. Biol*. **7**, 637–644.
16. Rye C.S., Withers S.G. 2000. Glycosidase mechanisms. *Curr. Opin. Chem. Biol*. **4**, 573–580.
17. Zechel D.L., Withers S.G. 2000. Glycosidase mechanisms: anatomy of a finely tuned catalyst. *Acc. Chem. Res*. **33**, 11–18.
18. Garman S.C., Hannick L., Zhu A., Garboczi D.N. 2002. The 1.9 Å structure of α-N-acetylgalactosaminidase: molecular basis of glycosidase deficiency diseases. *Structure*. **10**, 425–434.
19. Наумов Д.Г. 2004. Филогенетический анализ α-галактозидаз семейства GH27. *Молекуляр. биология*. **38**, 463–467.
20. Lovering A.L., Lee S.S., Kim Y.W., Withers S.G., Strynadka N.C. 2005. Mechanistic and structural analysis of a family 31 α-glycosidase and its glycosyl-enzyme intermediate. *J. Biol. Chem*. **280**, 2105–2115.
21. Naumoff D.G. 2008. The GH31 family of glycoside hydrolases: subfamily structure and evolutionary connections. *Abstr. Sixth Internat. Conf. Bioinformat. Genome Regulat. Struct*. June 22-28, 2008. Novosibirsk. Russia. P. 169. (http://www.bionet.nsc.ru/meeting/bgrs2008/BGRS2008_Proceedings.pdf).
22. Kuznetsova A.Y., Naumoff D.G. 2006. Phylogenetic analysis of COG1649, a new family of predicted glycosyl hydrolases. *Proc. Fifth Internat. Conf. Bioinformat. Genome Regulat. Struct*. July 16-22, 2006. Novosibirsk. Russia. **3**, 179–182. (http://www.bionet.nsc.ru/meeting/bgrs2006/BGRS_2006_V3.pdf).
23. Naumoff D.G. 2010. GH101 family of glycoside hydrolases: subfamily structure and evolutionary connections with other families. *J. Bioinform. Comput. Biol*. **8**, 437–451.