

## Семейства белков, кодируемых в пластомах Chlorophyta, Euglenozoa и Rhizaria

О.А. Зверков, А.В. Селиверстов, В.А. Любецкий  
ИППИ РАН  
zverkov@iitp.ru

### Аннотация

Разделение белков по семействам позволяет уточнять их аннотации и выполнять поиск белка по его филогенетическому профилю. Нами выполнено такое разделение (кластеризация) белков, кодируемых в пластомах рано отделившихся ветвей зелёных водорослей и видов с родственными пластидами: *Euglena gracilis*, *E. longa* и *Bigelowiella natans* («хлорофитная» ветвь пластид) на основе того же ранее полученного нами алгоритма. Соответствующая база данных и поиск кластера по филогенетическому профилю белка доступны на нашем сайте по адресу <http://lab6.iitp.ru/ppc/chlorophyta/>. На её основе показано, что отсутствуют белки, специфичные для пластома небольших таксономических групп этих водорослей за исключением отчасти класса *Chlorophyceae*. Эта ситуация расходится с той, которую мы наблюдаем для отдела *Rhodophyta*. Аналогичная база данных была полученная нами для родофитной ветви пластид доступна по адресу <http://lab6.iitp.ru/ppc/redline/>. Мы предполагаем, что донором пластид для *E. longa* и *E. gracilis* служит класс *Trebouxiophyceae*.

### 1. Введение

Рассматриваются пластиды рано отделившихся ветвей зелёных водорослей, объединяемых в таксономическую группу *Chlorophyta* [1, 2], и видов с родственными пластидами, полученными в результате вторичного эндосимбиоза от видов из *Chlorophyta* (по аналогии с термином «родофитная ветвь» будем говорить о «хлорофитной» ветви пластид). Это – *Euglena gracilis*, *E. longa* (из отдела *Euglenozoa*) и *Bigelowiella natans* (из группы *Rhizaria*). Таксономическая группа *Chlorophyta* делится на классы *Chlorophyceae* (роды *Floydiella*, *Schizomeris*, *Stigeoclonium*,

*Chlamydomonas*, *Oedogonium*, *Scenedesmus*), *Mamiellophyceae* (роды *Micromonas* и *Ostreococcus*), *Prasinophyceae* (роды *Monomastix*, *Nephroselmis*, *Pycnococcus*, *Pyramimonas*), *Trebouxiophyceae* (роды *Chlorella*, *Parachlorella*, *Coccomyxa*, *Leptosira*, *Helicosporidium*), *Ulvophyceae* (*Bryopsis* и *Pseudendoclonium*) и род *Oltmannsiellopsis*, который, вероятно, является рано отделившейся ветвью класса *Ulvophyceae*. Более точное деление внутри класса *Chlorophyceae* обсуждается в [3], а внутри класса *Trebouxiophyceae* – в [2]. Заметим, что многие виды из класса *Trebouxiophyceae* (Требуксиевые), входящего в состав *Chlorophyta*, являются обычными симбионтами лишайников и простейших, включая инфузорию *Paramecium bursaria* и амёб *Acanthamoeba borokensis* и *A. amazonas* [4].

Заметим, что изучение родофитной ветви пластид было проведено нами ранее [5, 6].

Два вида простейших *Euglena gracilis* [7] и *E. longa* (другое название *Astasia longa*) [8, 9] являются ближайшими друг к другу представителями отдела *Euglenozoa* [10, 11], хотя они значительно отличаются друг от друга. *E. gracilis* является фотосинтезирующим видом со смешанным типом питания и имеет светочувствительные стигмы, характерные для большинства видов этого отдела. Напротив, *E. longa* не способна к фотосинтезу. Это обусловило значительную редукцию её пластома. Хорошо известно, что пластиды, наряду с митохондриями, являются местом для независимого от света синтеза многих веществ и присутствуют у многих видов, лишённых фотосинтеза. Некоторые ортологичные белки *E. longa* и *E. gracilis* хорошо выравниваются, хотя заметно отличаются друг от друга. В частности, на выравнивании рибосомных белков из пластид *E. longa* и *E. gracilis* доли одинаковых аминокислотных остатков составляют для L2: 68%, для L20: 44%, для L22: 42%, для L23: 49%, для S19: 52%, [8]. Состав светособирающих пигментов *E. gracilis* и

выравнивание белков, кодируемых в пластидах, показывают родство пластид *Euglena* spp. и пластид зелёных водорослей. Более того, пластиды *Euglena* spp. ближе к пластидам водорослей из классов Chlorophyceae и Trebouxiophyceae, чем к таковым из класса Prasinophyceae [12]. Однако состав пластома, положение интронов и взаимное расположение генов на хромосоме в пластомах *E. gracilis* и других водорослей значительно различаются, что затрудняет определение непосредственного донора её пластид. Происхождения пластид *E. longa* также остаётся не вполне ясным.

Состав пигментов *Bigelowiella natans* соответствует кодовому у зелёных водорослей, и многие белки, кодируемые в пластеде, хорошо выравниваются с белками пластид зелёных водорослей. Напомним также, что вторичное происхождение пластид *B. natans* от зелёных водорослей непосредственно подтверждается наличием нуклеоморфа, остатка от ядра водоросли [13]. Напротив, у *Euglena* spp. нуклеоморф отсутствует.

Разделение белков на семейства (кластеризация белков) позволяет уточнять аннотации белков и выполнять поиск белка по его филогенетическому профилю. Такое разделение позволяет судить, например, о работоспособности белковых комплексов и об эволюции пластома.

Напомним, что филогенетическим профилем белка (гена) называется функция, определённая

на множестве видов и равная +1, если белок (ген) представлен в этом виде, и –1, если не представлен. Филогенетические профили некоторых консервативных пластомных генов из Chlorophyta получены в [1], однако там отсутствуют данные о наличии многих белков (обычно их функция неизвестна); филогенетические профили всех белков родофитной ветви пластид собраны нами в базу данных [5, 6].

Часто возникает обратная задача: поиск белка, который соответствует филогенетическому профилю, уже заданному из каких-то внешних соображений. Такой поиск белка по профилю облегчается, если белки уже разбиты на семейства, организованные в базу данных. Известно несколько таких баз данных [14]. Однако большинство из них ограничивается небольшим числом видов, лишь немногие включают пластиды, и то в очень ограниченном количестве. Нами получена новая база данных, отражающая кластеризацию всех пластомных белков видов, указанных в табл. 1, на основе оригинальных алгоритма и программы, представленных в [5, 6]. Эта база данных и поиск кластеров по филогенетическому профилю белка на её основе доступны на нашем сайте по адресу <http://lab6.iitp.ru/ppc/chlorophyta/>. Там же расположено описание упомянутого алгоритма и инструкции по его применению.

**Таблица 1. Пластома из хлорофитной ветви пластид и количества кластеров.**

Обозначения: в третьем столбце указано общее количество белков в пластома данного вида, (>1) – количество кластеров, которые содержат белки из двух и более видов, представленных в данном пластома из данного вида; (1) – аналогичное количество кластеров, состоящих из белков одного вида.

Пластом	Вид	Белков	(>1)	(1)
NC_001865.1	<i>Chlorella vulgaris</i>	174	94	78
NC_000927.1	<i>Nephroselmis olivacea</i>	155	111	13
NC_002652.1	<i>Euglena longa</i>	46	37	6
NC_001603.2	<i>Euglena gracilis</i>	67	61	5
NC_005353.1	<i>Chlamydomonas reinhardtii</i>	69	65	2
NC_008101.1	<i>Scenedesmus obliquus</i>	77	73	1
NC_008099.1	<i>Oltmannsiellopsis viridis</i>	93	80	3
NC_008100.1	<i>Helicosporidium sp. ex Simulium jonesi</i>	26	25	1
NC_008114.1	<i>Pseudendoclonium akinetum</i>	105	84	17
NC_008289.1	<i>Ostreococcus tauri</i>	61	58	2
NC_008372.1	<i>Stigeoclonium helveticum</i>	79	72	6
NC_008408.1	<i>Bigelowiella natans</i>	61	57	0
NC_009681.1	<i>Leptosira terrestris</i>	88	82	4
NC_011031.1	<i>Oedogonium cardiacum</i>	99	80	1
NC_012099.1	<i>Pyramimonas parkeae</i>	94	83	4
NC_012097.1	<i>Pycnococcus provasolii</i>	68	66	2
NC_012101.1	<i>Monomastix</i> sp. OKE-1	82	70	10
NC_012568.1	<i>Micromonas pusilla</i> CCMP1545	27	26	0
NC_012575.1	<i>Micromonas</i> sp. RCC299	57	56	0
NC_012978.1	<i>Parachlorella kessleri</i>	84	81	0
NC_013359.1	<i>Bryopsis hypnoides</i>	69	68	1
NC_014346.1	<i>Floydiella terrestris</i>	74	69	4
NC_015084.1	<i>Coccomyxa</i> sp. C-169	80	78	2
NC_015359.1	<i>Chlorella variabilis</i>	80	80	0
NC_015645.1	<i>Schizomeris leibleinii</i>	77	72	4

## 2. Материалы и методы

Пластомы 25-ти видов получены из базы данных NCBI и перечислены в табл. 1. Для контроля результатов и построения филогенетических деревьев использовались пакет программ MEGA 5 [15] и база данных Pfam [16].

Сам алгоритм выделения белковых семейств (кластеризации) был представлен на 53-й научной конференции МФТИ [5] и описан в статье [6]. Веб-страницы <http://lab6.iitp.ru/ppc/chlorophyta/> и <http://lab6.iitp.ru/ppc/redline/> обеспечивают, по крайней мере, следующие две функции. Поиск всех кодируемых в пластидах белков: 1) по заданному филогенетическому профилю, 2) по фрагменту аминокислотной последовательности, при этом вместе с каждой найденной последовательностью указывается, кластер, которому она принадлежит.

Некоторые алгоритмически полученные кластеры приходилось затем объединять из биологических соображений. Приведём примеры такого дополнительного анализа кластеров.

(1) Три кластера, состоящие из рибосомных белков S3, объединены в один кластер. Эти белки имеют консервативные участки на N- и C-концах, но в середине у некоторых видов произошли длинные неконсервативные вставки; вероятно, это – не имеющие описания интроны с длинами, кратными 3 п.н.

(2) Два кластера, состоящие из рибосомных белков S4, объединены в один кластер.

(3) К большому кластеру, состоящему из  $\alpha$ -субъединиц РНК-полимераз бактериального типа (RpoA), добавлены два маленьких кластера из гомологичных белков: один состоял из белков двух видов *Euglena* spp., другой – из белков вида *Rupnococcus provasolii*. Наибольшее значение величины  $e$ -value при сравнении с доменами из базы данных Pfam, характерными для RpoA, наблюдаются у белка из *P. provasolii*, однако ближайший к нему гомолог у *Pyramimonas parkae*

несомненно является  $\alpha$ -субъединицей РНК-полимеразы.

(4) Объединены в один кластер белки АТФ-связывающей субъединицы протеазы ClpP. На выравнивании видно, что на N-конце белков имеется консервативный участок длиной около 50 аминокислотных остатков. У видов из класса Chlorophyceae присутствует вставка длиной около 300 аминокислотных остатков, и вблизи C-конца имеется ещё один консервативный участок длиной около 130 аминокислотных остатков. При проверке по базе данных Pfam у всех белков обнаружены домены, характерные для Clp протеазы.

(5) Объединены в один кластер некоторые белки, гомологичные белку Ycf1, для которых удалось построить хорошее множественное выравнивание.

(6) Объединены в один кластер некоторые белки, гомологичные белку FtsH и связанные с делением пластид, для которых удалось построить хорошее множественное выравнивание и которые имеют домен, характерный для АТФаз.

(7) Объединены в один кластер белки, являющиеся фрагментами  $\beta$ -субъединицы РНК-полимеразы бактериального типа (RpoB). Так полученный кластер подробнее описан в конце следующего раздела 3.

## 3. Результаты и обсуждение

Кластеризация охватывает 1992 белка, из которых сформированы 166 одноэлементных кластеров и 156 кластеров, включающих более одного белка; среди них 87 содержат не более одного белка из каждого вида, 68 содержат пары белков из одного и того же вида, один кластер содержит три белка из одного и того же вида. Детали показаны в табл. 1. Распределение кластеров по числу представленных в них видов показано на рис. 1. Из них 13 кластеров имеют представителей в каждом виде группы Chlorophyta (из табл. 1). Более

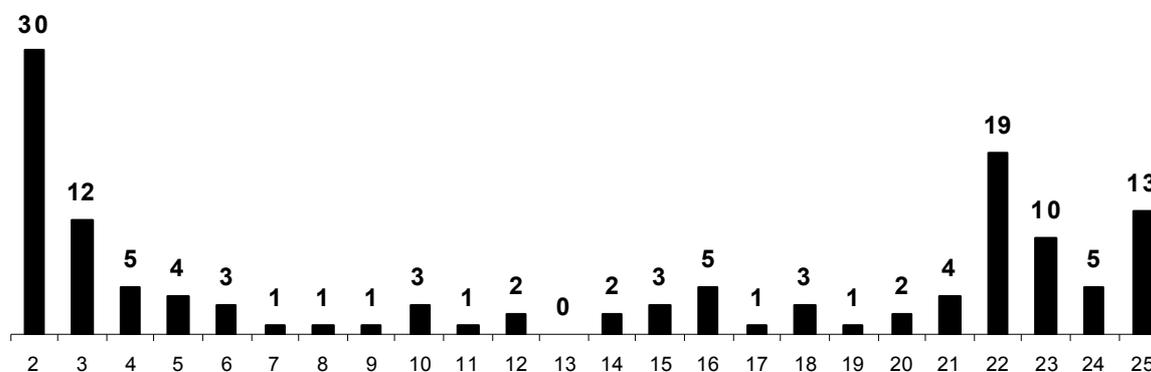


Рисунок 1. Распределение кластеров по числу видов.

Приведено количество кластеров пластоных белков из хлорофитной ветви пластид в зависимости от точного количества видов, представленных в кластере

того, представители этих 13-ти кластеров найдены и у вторичных эндосимбионтов *Euglena gracilis*, *E. longa* и *Bigelowiella natans*. Для каждого из классов Mamiellophyceae, Prasinophyceae, Trebouxiophyceae и Ulvophyceae не найдено ни одного специфичного кластера белков, т.е. такого, что его белки присутствуют в каждом виде данного класса, но отсутствуют в других классах группы Chlorophyta. Среди Chlorophyta только класс Chlorophyceae имеет, и при том ровно один, специфичный для него кластер белков, вероятно, связанных с делением пластид. Это показывает значительную близость пластомов рассмотренных видов из группы Chlorophyta. Здесь ситуация принципиально отличается от ситуации в отделе Rhodophyta, [5, 6].

Всего 29 кластеров имеют представителей в обоих видах *Euglena longa* и *E. gracilis*. Из них только один кластер представлен ровно в двух видах *E. longa* и *E. gracilis*; он состоит из белков NP\_074988.1 и NP\_041917.1 с неизвестной функцией; 14 кластеров содержат белок из *E. longa*, но не содержат соответствующего белка из *E. gracilis*; 37 кластеров содержат белок из *E. gracilis*, но не содержат белка из *E. longa*. Наиболее часто белки из пластома *E. longa*, отсутствующие в *E. gracilis*, имеют гомологи у водорослей *Chlorella vulgaris* и *Leptosira terrestris* из класса Trebouxiophyceae (Требуксиевые). Наша гипотеза: донором пластид для *E. longa* и *E. gracilis* служит класс Trebouxiophyceae. Это хорошо согласуется с высокой частотой вхождения водорослей именно этого класса в состав симбиотических сообществ, включая лишайники и некоторые простейшие [4]. Асимметрия (14<37) между их пластомами связана с заметно меньшим размером пластома *E. longa* по сравнению с *E. gracilis*.

Выравнивание с помощью MUSCLE белков каждого из 13-ти кластеров, имеющих представителей в каждом из рассмотренных видов (табл. 1), и последующее построение деревьев белков методом Maximum Likelihood показало, что в 11-ти из 13-ти кластеров ближайшим к белку из *Euglena longa* является белок из *E. gracilis*, а в двух других случаях – из *Helicosporidium* sp. ex *Simulium jonesii*. При этом в одном из этих двух случаев три вида *Euglena* spp. и *Helicosporidium* sp. образуют отдельную кладу. Всего в 4-х из 13-ти кластеров три вида *Euglena* spp. и *Helicosporidium* sp. образуют отдельную кладу. В остальных случаях положение белков из *Euglena* spp. относительно белков других родов меняется хаотически, что не позволяет сделать уверенное предположение о доноре пластид для *Euglena* spp., однако принадлежность *Helicosporidium* к классу Trebouxiophyceae согласуется с нашей гипотезой о происхождении пластид *Euglena* spp. от вида из этого класса.

Наконец, опишем содержание кластера  $\beta$ -субъединиц РНК-полимераз бактериального типа (RpoB). Здесь оказалось полезным учесть архитектуру белка RpoB у разных видов. У этих белков с помощью базы данных Pfam было определено взаимное расположение доменов. Ниже используется следующая нумерация доменов из базы данных Pfam: 1-й – RNA\_pol\_Rpb2\_1, 2-й – RNA\_pol\_Rpb2\_2, 3-й – RNA\_pol\_Rpb2\_3, домены 4 и 5 объединены в рассматриваемых видах в один домен RNA\_pol\_Rpb2\_45, 6-й – RNA\_pol\_Rpb2\_6, и 7-й – RNA\_pol\_Rpb2\_7. У вида *Coccomyxa* sp. C-169 в белке RpoB (YP\_004222037.1) домен 1 имеет две копии. В двух видах *Leptosira terrestris* (YP\_001382217.1) и *Oedogonium cardiacum* (YP\_002000391.1) домен 6 имеет по две копии, расположенных в указанных белках рядом. В пяти видах – *Chlamydomonas reinhardtii* (NP\_958398.1), *Oltmannsiellopsis viridis* (YP\_635874.1), *Pseudendoclonium akinetum* (YP\_636174.1), *Scenedesmus obliquus* (YP\_635950.1) и *Stigeoclonium helveticum* (YP\_764419.1) – домен 2 имеет по две копии в указанных белках. В 13-ти видах – *Pyramimonas parkeae* (YP\_002600950.1), *Nephroselmis olivacea* (NP\_050839.1), *Euglena gracilis* (NP\_041951.1), *Oltmannsiellopsis viridis* (YP\_635874.1), *Chlorella vulgaris* (NP\_045893.1), *Ostreococcus tauri* (YP\_717229.1), *Micromonas* sp. RCC299 (YP\_002808641.1), *Pseudendoclonium akinetum* (YP\_636174.1), *Micromonas pusilla* CCMP1545 (YP\_002808499.1), *Chlorella variabilis* (YP\_004347774.1), *Bigelowiella natans* (YP\_778610.1), *Parachlorella kessleri* (YP\_003058290.1), *Coccomyxa* sp. C-169 (YP\_004222037.1) – все домены представлены во всех белках из кластера. К этим 13-ти видам примыкают *Euglena longa* (NP\_074962.1), у которой потеряны только домен 2, и *Monomastix* sp. ОКЕ-1 (YP\_002601004.1), у которой потеряны только домены 4 и 5.

Ещё у семи видов  $\beta$ -субъединица РНК-полимеразы (RpoB) разделилась на два белка, один из которых включает только домены 6 и 7, а другой – только остальные домены с 1 по 5. Это – *Chlamydomonas reinhardtii* (NP\_958398.1, NP\_958397.1), *Scenedesmus obliquus* (YP\_635950.1, YP\_635949.1), *Schizomeris leibleinii* (YP\_004581337.1, YP\_004581350.1), *Stigeoclonium helveticum* (YP\_764419.1, YP\_764412.1), *Floydiella terrestris* (YP\_003795481.1, YP\_003795537.1), *Leptosira terrestris* (YP\_001382216.1, YP\_001382217.1) и *Oedogonium cardiacum* (YP\_002000410.1, YP\_002000391.1). Особняком стоят ещё два вида: *Helicosporidium* sp. ex *Simulium jonesii* (YP\_635922.1) с доменами 3, 6, 7 и *Bryopsis hypnoides* (YP\_003227091.1) с доменами 1, 2, 3 у белка RpoB. У *Pyrococcus provasolii* белок RpoB не был определён. Отметим, что у *P.*

*provasolii* размеры белков RpoC1 и RpoC2, кодирующих  $\beta'$ - и  $\beta''$ -субъединицы РНК-полимеразы, значительно больше, чем у близкого вида *Pyramimonas parkeae*.

Работа выполнена при частичной финансовой поддержке госконтрактами 14.740.11.0624, 14.740.11.1053, 14.740.12.0830 Министерства образования и науки РФ.

## Список литературы

- [1] F. Lü, W. Xü, C. Tian, G. Wang, J. Niu, G. Pan, S. Hu, *The Bryopsis hypnoides plastid genome: multimeric forms and complete nucleotide sequence*, PLoS ONE 6(2), 2011, e14663.
- [2] M. Turmel, C. Otis, C. Lemieux *The chloroplast genomes of the green algae Pedinomonas minor, Parachlorella kessleri, and Oocystis solitaria reveal a shared ancestry between the Pedinomonadales and Chlorellales*, Mol. Biol. Evol. 26(10), 2009, 2317–2331.
- [3] J.S. Brouard, C. Otis, C. Lemieux, M. Turmel, *The exceptionally large chloroplast genome of the green alga Floydiella terrestris illuminates the evolutionary history of the Chlorophyceae*, Genome Biol. Evol. 2, 2010, 240–256.
- [4] А.С. Карлов, *Взаимодействие зоохлорелл с новым потенциальным хозяином – крупными свободно живущими амёбами*, Цитология. 34(4), 1992, 73.
- [5] О.А. Зверков, А.В. Селиверстов, В.А. Любецкий, *Об одном алгоритме кластеризации белков*, Труды 53-й научной конференции МФТИ. М.: МФТИ, 2010, часть 1, Т. 1, С. 118–119.
- [6] О.А. Зверков, А.В. Селиверстов, В.А. Любецкий, *Белковые семейства, специфичные для пластомов небольших таксономических групп водорослей и простейших*, Молекулярная биология, 2012, (в печати).
- [7] R.B. Hallick, L. Hong, R.G. Drager, M.R. Favreau, A. Monfort, B. Orsat, A. Spielmann, E. Stutz, *Complete sequence of Euglena gracilis chloroplast DNA*, Nucleic Acids Res. 21(15), 1993, 3537–3544.
- [8] G. Gockel, S. Baier, W. Hachtel, *Plastid ribosomal protein genes from the nonphotosynthetic flagellate Astasia longa*, Plant Physiol. 105, 1994, 1443–1444.
- [9] G. Gockel, W. Hachtel, *Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate Astasia longa*, Protist. 151(4), 2000, 347–351.
- [10] E.W. Linton, A. Karnkowska-Ishikawa, J.I. Kim, W. Shin, M.S. Bennett, J. Kwiatowski, B. Zakryś, R.E. Triemer, *Reconstructing euglenoid evolutionary relationships using three genes: nuclear SSU and LSU, and chloroplast SSU rDNA sequences and the description of Euglenaria gen. nov. (Euglenophyta)*, Protist. 161(4), 2010, 603–19.
- [11] S. Brosnan, W. Shin, K.M. Kjer, R.E. Triemer, *Phylogeny of the photosynthetic euglenophytes inferred from the nuclear SSU and partial LSU rDNA*, Int. J. Syst. Evol. Microbiol. 53(4), 2003, 1175–86.
- [12] M. Turmel, C. Otis, C. Lemieux, *The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: Insights into the architecture of ancestral chloroplast genomes*, Proc. Natl. Acad. Sci. USA. 96, 1999, 10248–10253.
- [13] P.R. Gilson, V. Su, C.H. Slamovits, M.E. Reith, P.J. Keeling, G.I. McFadden, *Complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus*, Proc. Natl. Acad. Sci. U.S.A. 103(25), 2006, 9566–9571.
- [14] A.M. Altenhoff, C. Dessimoz, *Phylogenetic and functional assessment of orthologs inference projects and methods*, PLoS Comput. Biol. 5(1), 2009, e1000262.
- [15] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods*, Molecular Biology and Evolution. 28, 2011, 2731–2739.
- [16] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman, *The Pfam protein families database*, Nucleic acids research, 38, Database issue, 2010, D211–D222.