ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

УДК 577.322

СОСТАВЛЕНИЕ И АНАЛИЗ МАТРИЦ АМИНОКИСЛОТНЫХ ЗАМЕН ДЛЯ ОПТИМАЛЬНОГО ВЫРАВНИВАНИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ МИКРОБНЫХ РОДОПСИНОВ

В.Н. Новоселецкий*, Г.А. Армеев, К.В. Шайтан

Кафедра биоинженерии, биологический факультет, Московский государственный университет имени М.В. Ломоносова, Россия, 119234, г. Москва, Ленинские горы, д. 1, стр. 12
**e-mail: valery.novoseletsky@yandex.ru*

Парное выравнивание аминокислотных последовательностей является основным инструментом биоинформатики, используемым как самостоятельно, так и в большом числе более сложных методов. Эффективность этого инструмента критически зависит от используемой оценочной функции, которая состоит из матрицы замен и штрафов за вставку. В настоящей работе выполнены построение и анализ матриц аминокислотных замен для надсемейства микробных родопсинов (RHOD), а также проведено их сопоставление с рядом матриц замен общего назначения (BLOSUM, VTML, PFASUM). Показано, что все матрицы позволяют строить выравнивания последовательностей микробных родопсинов практически одинакового качества, но лишь матрицы семейств BLOSUM и VTML, а также их линейные комбинации с матрицами семейства RHOD позволяют обнаруживать гомологию между микробными родопсинами и гелиородопсином.

Ключевые слова: *микробные родопсины, гелиородопсин, матрица замен, энтропия, выравнивание последовательностей, поиск гомологов*

Микробные родопсины – это обширное надсемейство фотохимически активных белков прокариот и низших эукариот, распространенных по всему земному шару. Со структурной точки зрения они представляют собой трансмембранный пучок из 7 α-спиралей, образующих карман для ретиналя, который ковалентно связывается с остатком лизина в VII спирали. Поглощение кванта света приводит к изомеризации ретиналя, что, в свою очередь, запускает свойственные каждому семейству родопсинов процессы. Что касается биологической функции, то микробные родопсины являются либо приемниками энергии света, которые затем превращают ее в электрохимические потенциал клеточной мембраны (светоуправляемые ионные насосы), либо датчиками, которые используют информацию об окружении для регуляции клеточных процессов [1]. Несмотря на длительную историю изучения и широчайшие исследования микробных родопсинов, они по-прежнему дают почву для новых открытий, ярчайшим примером которых стало недавнее открытие гелиородопсина [2]. Этот ретиналь-содержащий семиспиральный трансмембранный белок имеет чрезвычайно низкое сходство последовательности с последовательностями других микробных родопсинов, а его принципиальной особенностью является инвертированная ориентация относительно мембраны: N-конец располагается внутри клетки, а C-конец – снаружи. В настоящий момент нет полного понимания филогенетических отношений между обширным семейством гелиородопсин-подобных белков и ранее известными микробными родопсинами, но можно ожидать, что семиспиральные трансмембранные белки таят в себе еще немало удивительного.

Математически строгим и вычислительно эффективным методом нахождения оптимальных глобального и локального выравниваний пары последовательностей является динамическое программирование [3, 4], причем результат работы этого алгоритма критически зависит от используемой оценочной функции, которая обычно состоит из матрицы замен аминокислот и штрафов за вставки. Несмотря на появление новых методов выравнивания [5], основанных на статистических подходах, динамическое программирование по-прежнему широко используется, например, для моделирования по гомологии [6] или построения множественных выравниваний такими программами, как CLUSTALW [7] и др. Выбор матрицы замен является наиболее важ-

ным решением перед построением выравнивания. Большинство таких матриц построены путем расчета сходства между аминокислотами как частоты их встречаемости в соответствующих позициях различных последовательностей. Если оказывается, что аминокислоты A и B часто встречаются в эквивалентных позициях, то они обладают сходными свойствами и могут быть заменены друг на друга в процессе эволюции белков. Различие между матрицами обуславливается, главным образом, тем, какая группа белковых последовательностей использована для получения этих матриц и каким образом были определены эквивалентные позиции [8]. Так, широко используемые матрицы замен BLOSUM [9] были получены путем оценки встречаемости аминокислотных замен в выравниваниях эволюционно родственных белков.

Практически все существующие матрицы замен являются матрицами общего назначения, поскольку были получены усреднением частот встречаемости в различных белковых семействах и отражают типичные черты всех белков. Матрицы такого рода востребованы для поиска схожих последовательностей в обширных базах данных, где последовательность запроса сравнивается с миллионами разнообразных последовательностей. Но часто приходится выполнять глобальное выравнивание последовательностей из одного заранее известного семейства и в этом случае даже лучшие матрицы общего назначения могут оказаться не самыми подходящими, поскольку совершенно не учитывают особенностей конкретного семейства [8].

В данной работе получено семейство матриц замен RHOD, специфичных для микробных родопсинов, и проведено исследование их точности и селективности. Показано, что полученные матрицы по качеству выравнивания последовательностей микробных родопсинов соответствуют широко используемым матрицам семейств BLOSUM [9] и VTML [10], лишь немного превосходя их. В то же время при выявлении сходства микробных родопсинов и гелиородопсина при поиске в базах данных SwissProt [11] и PDB [12] матрицы RHOD оказались недостаточно эффективными по сравнению с матрицами BLOSUM и VTML. Тем не менее, матрицы, полученные линейной комбинацией матриц RHOD и VTML, продемонстрировали высокую эффективность в обоих тестах.

Материалы и методы

Расчет матриц замен выполняли на основе множественного выравнивания последовательностей ряда микробных родопсинов и их гомологов из баз данных SwissProt и TrEMBL [11] по стандартному алгоритму [9]. Поиск гомологов выполняли с помощью программы FASTA [14]. Для построения множественного выравнивания использовали веб-сервисы Clustal Omega [15] и Muscle [16]. Матрицы BLOSUM были взяты на ресурсе NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/matrices). Структурное выравнивание выполняли в программе Maestro (Schrqdinger, LLC, США). Для расчета парных выравниваний с различными матрицами замен использовали программу AlignMe [17].

Результаты и обсуждение

Для построения аккуратного множественного выравнивания последовательностей различных семейств микробных родопсинов желательно использовать большое число их гомологов, поскольку сходство между последовательностями весьма невелико. Так, например, при структурном выравнивании галородопсина *Halobacterium salinarium* (pdb-код 1e12) [18] и галородопсина из *Natromonas pharaonis* (pdb-код 3a7k) [19] идентичность составляет 54%, в то время как при его выравнивании с сенсорным родопсином из *Natromonas pharaonis* (pdb-код 3qap) [20] идентичность составляет 23%, а с натриевым светочувствительным насосом из *Dokdonia eikasta* (pdb-код 3x3b) [21] – всего 11%. Однако для структурного выравнивания в базе данных PDB доступны всего лишь около 20 уникальных структур микробных родопсинов, включая четыре указанных. База данных белковых последовательностей SwissProt также содержит небольшое количество уникальных последовательностей микробных родопсинов, поэтому основная часть последовательностей для построения выравниваний была взята из базы данных TrEMBL. После исключения последовательностей с идентичностью более 99% итоговый набор стал содержать 260 последовательностей (235 из TrEMBL и 25 из SwissProt).

Множественное выравнивание последовательностей выполняли с помощью ряда соответствующих программ, включая Clustal Omega [15] и Muscle [16], но на эффективности итоговых матриц замен это практически не отразилось, поэтому далее приведены лишь результаты, полученные на основе выравниваний с помощью программы Clustal Omega. Попарная идентичность последовательностей составила от 11% до 98%, средняя идентичность – 30±14%. Для расчета матриц замен была использована только часть выравнивания, относящаяся к трасмембранному домену рассматриваемых белков (позиции с 130 по 470). Большой (до 350 остатков) цитоплазма-

тический домен ряда последовательностей (каналородопсины) не учитывался.

Альтернативный вариант множественного выравнивания последовательностей был получен после попарного структурного выравнивания невырожденного набора из 22 структур микробных родопсинов, имеющихся в базе данных PDB, и их последующего объединения. Попарная идентичность последовательностей составила от 11% до 64%, средняя идентичность – 25±11%. Однако полученная на основе такого выравнивания матрица замен по своим свойствам оказалась близка к матрице RHOD0020 (см. ниже) и далее отдельно не обсуждается.

Создание матриц замен

Создание семейства матриц замен RHOD, специфичных для семейства микробных родопсинов, и расчет их относительной энтропии были выполнены по стандартному алгоритму, использованному ранее при создании семейства матриц BLOSUM [9] (табл. 1). Нумерация матриц этого семейства означает идентичность последовательностей в выравнивании, использованном для построения матриц. Аналогичная нумерация используется и в семействе PFASUM [13], которое было построено с использованием структурных выравниваний, представленных в базе данных PFAM [22]. По аналогии с этими семействами матриц, нумерация полученных нами матриц также опирается на идентичность последовательностей и содержит два числа, первое из которых означает минимальную, а второе – максимальную идентичность последовательностей в использованном выравнивании. Так, название RHOD0020 указывает на то, что при расчете этой матрицы было использовано выравнивание последовательностей, идентичность которых лежит в пределах от 0 до 20%. Напротив, числа в названиях матриц семейства VTML обозначают степень эволюционного расхождения последовательностей, использованных для построения соответствующих матриц [10]. Свойства матриц указанных семейств приведены в табл. 1 (данные по матрицам VTML взяты из файла upam.h пакета fasta [14]). Матрицы расположены по возрастанию относительной энтропии и средней идентичности последовательностей. Интересно отметить большую относительную энтропию матриц семейства RHOD по сравнению с матрицами семейства VTML, построенными по выравниваниям сравнимой идентичности, однако выяснение причин этого требует глубокого анализа аминокислотного состава соответствующих последовательностей и выходит за пределы данной публикации.

Таблица 1

Относительная энтропия (H) и средняя идентичность последовательностей в исходных выравниваниях (I) ряда матриц замен

Матрица	H	I,%	Матрица	H	I,%	Матрица	H	Матрица	H
						BLOSUM35	0,21	PFASUM31	0,23
RHOD0020	0,53	17				BLOSUM40	0,29	PFASUM43	0,34
RHOD0030	0,59	22	VTML200	0,41	23			PFASUM51	0,41
						BLOSUM50	0,48	PFASUM60	0,49
RHOD0040	0,65	25	VTML160	0,56	25			PFASUM67	0,56
RHOD0050	0,68	26				BLOSUM62	0,70	PFASUM78	0,69
RHOD0100	0,79	30				BLOSUM70	0,84		
RHOD20100	0,95	33	VTML120	0,94	37	BLOSUM80	0,99		
RHOD30100	1,51	44	VTML80	1,39	50	BLOSUM100	1,45		
RHOD40100	2,72	60							
RHOD50100	3,18	65	VTML40	2,27	69				
RHOD60100	3,99	74							
RHOD70100	4,80	81	VTML20	2,92	83				
RHOD80100	5,76	89	VTML10	3,46	91				
RHOD90100	6,56	94							

Тестирование матриц замен

При расчете оценки конкретного парного выравнивания помимо матрицы замены важную роль играют штрафы за открытие вставки

и за ее удлинение. Подбор оптимальных значений штрафов является необходимым шагом для дальнейшего использования матрицы замен. Для каждой матрицы поиск оптимальных значений штрафов выполняли в диапазоне от 1 до 20 для штрафа за открытие вставки и от 1 до 5 для штрафа за удлинение. Критерием оптимальности параметров стало качество выравнивания *Q(i, j)* [8] для последовательностей *i* и *j*:

$$Q(i,j) = \frac{1}{2}\left(\frac{n_I(i,j)}{L_R(i,j)} + \frac{n_I(i,j)}{L_T(i,j)}\right)$$

где $n_I(i,j)$ – число одинаково выровненных пар остатков в референсном и тестовом выравниваниях, $L_R(i,j)$ – длина референсного выравнивания, $L_T(i,j)$ – длина тестового выравнивания.

Тестирование матриц замен выполняли на последовательностях микробных родопсинов, структура которых была определена экспериментально, и соответствующие данные размещены в базе данных PDB. В качестве референсных выступали выравнивания последовательностей, полученные на основе парных структурных выравниваний соответствующих родопсинов (231 выравнивание), а тестовые выравнивания были построены для этих же последовательностей, но с использованием тестируемых матриц замен и штрафов за вставки. Оптимальные значения штрафов и среднее значение качества выравниваний с использованием различных матриц приведены в табл. 2. Видно, что все исследуемые матрицы замен демонстрируют близкие средние значения качества выравнивания. Это не позволяет рекомендовать какую-либо из них в качестве лучшей для выполнения выравнивания последовательностей микробных родопсинов.

Еще одним применением парного сравнения последовательностей и, соответственно, матриц замен является поиск схожих последовательностей в базах данных. В этой связи особенный интерес представляет возможность нахождения таких последовательностей, которые не использовались в параметризации матриц, но имеют определенное сходство с последовательностями из обучающего набора. В нашем случае такой последовательностью стала последовательность недавно открытого гелиородопсина [2], практически не имеющая идентичных остатков с остальными микробными родопсинами. С другой стороны, при наличии в базах данных последовательностей, гарантированно гомологичных последовательности запроса, итоговый список схожих последовательностей должен содержать все гомологичные последовательности и никаких других. Для проверки такой селективности матриц замен была использована последовательность дельта-родопсина, структура которого известна (pdb-код 4fbz) [23]. Результаты проверок также приведены в табл. 2 и представляют особенный интерес. При использовании последовательности дельта-родопсина в качестве последовательности запроса в базах данных SwissProt и PDB все матрицы продемонстрировали практически одинаковую способность выявить все или почти все гомологичные последовательности в обоих ресурсах. Большее число последовательностей в базе данных PDB обусловлено тем, что некоторые белки имеют по несколько структур, размещенных в ней. Отметим, что в случае дельта-родопсина схожая результативность поиска наблюдалась в широком диапазоне штрафов. Напротив, результаты поиска гомологов гелиородопсина сильно зависели от использованных штрафов и в целом совершенно иные: в то время как матрицы семейств BLOSUM, VTML и отчасти PFASUM оказались способны выявить сходство последовательностей гелиородопсина и микробных родопсинов, матрицы семейства RHOD оказались к нему нечувствительны.

Для создания матриц замен, имеющих как высокое качество выравнивания, так и достаточную селективность, были построены матрицы Mix, являющиеся линейными комбинациями матриц RHOD0030 и VTML200:

где X и Y – весовые коэффициенты, при-

$$\text{Mix\_X-Y}[a, b] = X\cdot\text{RHOD0030}[a, b] + Y\cdot\text{VTML200}[a, b],$$

чем $X + Y = 100\%$, а a и b – матричные индексы.

Выбор матриц RHOD0030 и VTML200 обусловлен низкой относительной энтропией обеих при общей схожестью их свойств (табл. 1). Были рассмотрены следующие комбинации X и Y: 10% и 90%, 20% и 80%, 30% и 70%. Оказалось, что все три смешанные матрицы унаследовали лучшие черты исходных матриц, сочетая в себе способность давать высокое качество парных выравниваний и высокую селективность при поиске в базах данных (табл. 2).

Проведенное в данной работе исследование показывает, что матрицы аминокислотных замен RHOD, построенные на основе множественного выравнивания последовательностей микробных родопсинов, сравнимы по эффективности с матрицами замен общего назначения семейств BLOSUM, VTML и PFASUM при выравнивании последовательностей микробных родопсинов, но недостаточно селективны для выявления сходства по-

Таблица 2

Оптимальные штрафы (ОШ), среднее качество <Q> парных выравниваний микробных родопсинов и результативность матриц в поиске дельта- и гелиородопсина в базах данных SwissProt и PDB. В дробной записи: слева – число истинных гомологов среди результатов поиска, справа – общее число результатов поиска

Матрица	Парные выравнивания		Поиск гомологов дельтародопсина				Поиск гомологов гелиородопсина			
	ОШ	<Q>	ОШ	в SP	ОШ	в PDB	ОШ	в SP	ОШ	в PDB
BLOSUM35	16,2	0,70	1,2	44/45	6,1	317/323	20,1	3/13	2, 2	39/39
BLOSUM40	20,2	0,70	1,3	44/45	6,1	317/332	19,3	5/11	19, 1	6/7
BLOSUM80	20,5	0,69	1,3	44/46	4,2	317/323	13,1	9/28	2, 2	39/39
BLOSUM100	20,7	0,67	6,1	44/45	2,3	317/329	9,2	3/13	2, 3	39/45
VTML200	14,3	0,67	1,3	44/45	1,3	317/323	13,3	6/15	8, 3	6/7
VTML120	15,2	0,70	1,2	44/46	1,2	317/323	10,1	8/21	4, 3	6/6
VTML80	18,5	0,70	1,2	44/46	2,2	317/342	1,3	5/20	4,2	6/7
PFASUM31	14,2	0,69	1,3	44/45	3,3	317/323	–	0	9,2	27/29
PFASUM43	14,1	0,70	1,3	44/45	2,3	317/323	13,1	5/24	12,2	6/6
RHOD0020	12,1	0,70	6,3	44/48	3,3	317/324	–	0	–	0
RHOD0030	15,1	0,72	6,3	44/47	7,3	317/324	–	0	–	0
RHOD20100	15,1	0,70	12,3	44/46	2,3	317/323	–	0	–	0
RHOD30100	15,3	0,70	3,3	44/45	1,3	317/332	–	0	–	0
Mix_10-90	17,3	0,70	5,1	44/45	5,1	317/323	14,2	6/13	12,2	6/6
Mix_20-80	18,3	0,71	4,1	44/46	5,1	317/323	12,1	9/25	12,1	6/6
Mix_30-70	18,3	0,72	4,1	44/45	4,1	317/323	13,1	8/23	13,1	6/9

следовательностей микробных родопсинов с последовательностью гелиородопсина. В то же время, создание смешанных матриц замен позволяет устранить этот недостаток при сохранении высокого качества парных выравниваний.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект №17-00-00166 КОМФИ – создание матриц замен и №18-02-40010 мега – тестирование матриц замен). Вычисления выполнены с использованием суперкомпьютера «Ломоносов» [24].

Исследования выполнены без использования животных и без привлечения людей в качестве испытуемых.

Авторы заявляют, что у них нет конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Govorunova E.G., Sineshchekov O.A., Li H., Spudich J.L.* Microbial rhodopsins: diversity, mechanisms, and optogenetic applications // Annu. Rev. Biochem. 2017. Vol. 86. N 1. P. 845–872.

2. *Pushkarev A., Inoue K., Larom S. et al.* A distinct abundant group of microbial rhodopsins discovered using functional metagenomics // Nature. 2018. Vol. 558. N 7711. P. 595–599.

3. *Needleman S.B., Wunsch C.D.* A general method applicable to the search for similarities in the amino acid sequence of two proteins // J. Mol. Biol. 1970. Vol. 48. N 3. P. 443–453.

4. *Smith T.F., Waterman M.S.* Identification of common molecular subsequences // J. Mol. Biol. 1981. Vol. 147. N 1. P. 195–197.

5. *Lqytynoja A.* Alignment methods: strategies, challenges, benchmarking, and comparative overview // Evolutionary genomics. Methods in molecular biology (methods and protocols). Vol 855 / Ed. M. Anisimova. New Jersey: Humana Press, 2012. P. 203–235.

6. *Khan F.I., Wei D.Q., Gu K.R., Hassan M.I., Tabrez S.* Current updates on computer aided protein modeling and designing // Int. J. Biol. Macromol. 2016. Vol. 85. P. 48–62.

7. *Thompson J.D., Higgins D.G., Gibson T.J.* ClustalW: improving the sensitivity of progressive multiple sequence aligment through sequence weighting, position specific gap penalties and weight matrix choice // Nucleic Acids Res. 1994. Vol. 22. N 22. P. 4673–4680.

8. *Kuznetsov I.B.* Protein sequence alignment with family-specific amino acid similarity matrices // BMC Res. Notes. 2011. Vol. 4:296.

9. *Henikoff S., Henikoff J.G.* Amino acid substitution matrices from protein blocks // Proc. Natl. Acad. Sci. USA. 1992. Vol. 89. N 22. P. 10915–10919.

10. *Müller T., Spang R., Vingron M.* Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method // Mol. Biol. Evol. 2002. Vol. 19. N 1. P. 8–13.

11. The UniProt Consortium. UniProt: the universal protein knowledgebase // Nucleic Acids Res. 2018. Vol. 46. N 5. P. 2699–2699.

12. *Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E.* The Protein Data Bank // Nucleic Acids Res. 2000. Vol. 28. N 1. P. 235–242.

13. *Keul F., Hess M., Goesele M., Hamacher K.* PFASUM: A substitution matrix from Pfam structural alignments // BMC Bioinformatics. 2017. Vol. 18:293.

14. *Pearson W.R.* Rapid and sensitive sequence comparison with FASTP and FASTA // Methods Enzymol. 1990. Vol. 183. P. 63–98.

15. *Sievers F., Higgins D.G.* Clustal Omega for making accurate alignments of many protein sequences // Protein Sci. 2018. Vol. 27. N 1. P. 135–145.

16. *Edgar R.C.* MUSCLE: Multiple sequence alignment with high accuracy and high throughput // Nucleic Acids Res. 2004. Vol. 32. N 5. P. 1792–1797.

17. *Stamm M., Staritzbichler R., Khafizov K., Forrest L.R.* AlignMe—a membrane protein sequence alignment web server // Nucleic Acids Res. 2014. Vol. 42. N W1. P. W246–W251.

18. *Kolbe M., Besir H., Essen L.O., Oesterhelt D.* Structure of the light-driven chloride pump halorhodopsin at 1.8 E Resolution // Science. 2000. Vol. 288. N 5470. P. 1390–1396.

19. *Kouyama T., Kanada S., Takeguchi Y., Narusawa A., Murakami M., Ihara K.* Crystal structure of the light-driven chloride pump halorhodopsin from Natronomonas pharaonis // J. Mol. Biol. 2010. Vol. 396. N 3. P. 564–579.

20. *Gushchin I., Reshetnyak A., Borshchevskiy V., Ishchenko A., Round E., Grudinin S., Engelhard M., Bьldt G., Gordeliy V.* Active state of sensory rhodopsin II: Structural determinants for signal transfer and proton pumping // J. Mol. Biol. 2011. Vol. 412. N 4. P. 591–600.

21. *Kato H.E., Inoue K., Abe-Yoshizumi R. et al.* Structural basis for Na^+ transport mechanism by a light-driven Na^+ pump // Nature. 2015. Vol. 521. N 7550. P. 48–53.

22. *Finn R.D., Bateman A., Clements J., Coggill P., Eberhardt R.Y., Eddy S.R., Heger A., Hetherington K., Holm L., Mistry J., Sonnhammer E.L.L., Tate J., Punta M.* Pfam: The protein families database // Nucleic Acids Res. 2014. Vol. 42. N D1. P. D222–D230.

23. *Zhang J., Mizuno K., Murata Y., Koide H., Murakami M., Ihara K., Kouyama T.* Crystal structure of deltarhodopsin-3 from Haloterrigena thermotolerans // Proteins: Struct., Funct., Bioinf. 2013. Vol. 81. N 9. P. 1585–1592.

24. *Sadovnichy V., Tikhonravov A., Voevodin V., Opanasenko V.I.* "Lomonosov": Supercomputing at Moscow State University // Contemporary High Performance Computing: From Petascale toward Exascale. Boca Raton: CRC Press, 2013. P. 283–307.

Поступила в редакцию
03.09.2018
Поступила после доработки
28.12.2018
Принята в печать
11.01.2019

RESEARCH ARTICLE

CONSTRUCTION AND EVALUATION OF AMINO ACID SUBSTITUTION MATRICES FOR OPTIMAL ALIGNMENT OF SEQUENCES OF MICROBIAL RHODOPSINS

V.N. Novoseletsky*, G.A. Armeev, K.V. Shaitan

Department of Bioengineering, School of Biology, Lomonosov Moscow State University, Leninskiye gory 1–12, Moscow, 119234, Russia
**e-mail: valery.novoseletsky@yandex.ru*

Pairwise alignment of amino acid sequences is the basic tool of bioinformatics and it is widely used itself and in numerous approaches. Performance of this tool is critically depends on scoring function, which consisted of substitution matrix and gap penalties. In this work we constructed and evaluated a set of family-specific substitution matrices for microbial rhodopsins (RHOD) and compared them with general-purpose matrices (BLOSUM, VTML, PFASUM). We showed that all matrices demonstrated similar quality of pairwise alignment of microbial rhodopsins and only BLOSUM and VTML matrices, and linear combinations of them with RHOD matrices, are able to detect distant homology between microbial rhodopsins and heliorhodopsin.

Keywords: *microbial rhodopsins, heliorhodopsin, substitution matrix, entropy, sequence alignment, homology search*

Сведения об авторах

Новоселецкий Валерий Николаевич – канд. физ-мат. наук, доц. кафедры биоинженерии биологического факультета МГУ.
Тел.: 8-495-939-57-38; e-mail: valery.novoseletsky@yandex.ru

Армеев Григорий Алексеевич – мл. науч. сотр. кафедры биоинженерии биологического факультета МГУ. Тел.: 8-495-939-57-38;
e-mail: satary@yandex.ru

Шайтан Константин Вольдемарович – докт. физ-мат. наук, проф., зам. зав. кафедры биоинженерии биологического факультета МГУ.
Тел.: 8-495-939-57-38; e-mail: shaytan49@yandex.ru