*МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2006, том 40, № 3, с. 541–545*

## МАТЕМАТИЧЕСКАЯ И СИСТЕМНАЯ БИОЛОГИЯ

*УДК 577.1*

# ВЕРОЯТНОСТНЫЙ МЕТОД ПРЕДСКАЗАНИЯ ТРАНСМЕМБРАННЫХ УЧАСТКОВ ПО МНОЖЕСТВЕННОМУ ВЫРАВНИВАНИЮ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

**© 2006 г. Р. А. Сутормин[1]*, А. А. Миронов[1, 2, 3]**

*[1] Государственный научный центр "ГосНИИгенетика", Москва, 117545*

*[2]Институт проблем передачи информации Российской академии наук, Москва, 127994*

*[3]Факультет биоинформатики и биоинженерии Московского государственного университета им. М.В. Ломоносова, Москва, 119992*

Поступила в редакцию 25.01.2006 г.

Предсказание положений мембранных участков на последовательностях мембранных белков – давно известная и, безусловно, важная задача, точность решения которой методами, не использующими гомологичного поиска по дополнительному банку данных, может быть улучшена. В данной области ощущается нехватка тестовых данных из-за малого объема реальных структурных данных по мембранным белкам. В данной работе сформирована тестовая выборка структурных выравниваний мембранных белков, разметка которых является согласованием информации об известных трехмерных структурах, входящих в выравнивания аминокислотных последовательностей белков. Предлагается метод предсказания мембранной разметки выравнивания, использующий алгоритм Forward-backward из теории скрытых марковских моделей. Метод позволяет не только предсказывать положения мембранных участков, но и формировать вероятностный мембранный профиль, который может быть использован в дальнейшем в методах множественного выравнивания, учитывающих информацию о вторичной структуре последовательностей. Метод реализован в компьютерной программе, доступной в Internet по адресу http://bioinf.fbb.msu.ru/fwdbck/. Предложенный метод дает результаты, лучшие, чем метод MEMSAT, являющийся чуть ли не единственным методом предсказания мембранной разметки множественного выравнивания без использования гомологичного поиска.

*Ключевые слова*: мембранный белок, предсказание вторичной структуры, скрытые марковские модели, алгоритм forward-backward, вероятностный мембранный профиль, тестовая выборка.

MEMBRANE PROBABILITY PROFILE CONSTRUCTION BASED ON AMINO ACIDS SEQUENCES MULTIPLE ALIGNMENT, by *R. A. Sutormin[1]*, A. A. Mironov[1, 2, 3]* ([1]State Scientific Center GosNIIGenetika, Moscow, 113545 Russia, *e-mail: sutor_ra@mail.ru; [2]Institute for Information Transmission Problems, Moscow, 127994 Russia; [3]Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, 119992 Russia). Prediction of membrane segments in sequences of membrane proteins is well known and important problem. Accuracy of the solution of this problem by methods that don't use homology search in additional data bank can be improved. There is a lack of testing data in this area because of small amount of real structures of membrane proteins. In this work, we create a testing set of structural alignments of membrane proteins, in which positioning of the membrane segments reflects agreement of known 3D-structures of proteins in the alignment. We propose a method for predicting position of membrane segments in multiple alignment based on forward-backward algorithm from HMM theory. This method not only allows to predict positions of membrane segments but also forms probability membrane profile, which can be used in multiple alignment methods that take into account secondary structure information about sequences. Method is implemented in computer program available on the World-Wide Web site http://bioinf.fbb.msu.ru/fwdbck/. Proposed method provides results better than MEMSAT method, which is nearly only tool for prediction of membrane segments in multiple alignments without additional homology search.

*Key words*: membrane protein, secondary structure prediction, hidden markov models, forward-backward algorithm, probability membrane profile, testing data set.

*Эл. почта: sutor_ra@mail.ru

Многие задачи биоинформатики включают стадию выравнивания аминокислотных последовательностей белков [1]. Поэтому качество построения выравниваний часто является критическим фактором при анализе последовательностей [2]. В случае мембранных белков процесс установления трехмерных структур встречает достаточно много трудностей [3], и кристаллографических данных по ним известно немного. Вследствие этого трудно строить структурные выравнивания для проверки работы автоматических методов построения множественного выравнивания [4]. С другой стороны, есть основания полагать, что общепринятые методы выравнивания не гарантируют хорошего качества вследствие необычного и неравномерного аминокислотного состава мембранных белков [5]. Неоднократно высказывалось мнение, что вовлечение в процесс построения выравнивания информации о вторичной структуре белков должно приводить к улучшению качества выравнивания [6]. В случае мембранных белков в качестве вторичной структуры логичнее рассматривать области белковой последовательности, находящиеся в мембране [7], так как участки, не лежащие в мембране, по свойствам похожи на глобулярные белки, которые, в свою очередь, выравниваются довольно хорошо [8]. Таким образом, прежде чем разрабатывать алгоритм выравнивания, учитывающий мембранную разметку последовательностей, требуется научиться правильно строить эту разметку. К сожалению, качество методов, предсказывающих положения мембранных участков по аминокислотной последовательности, далеко от идеального [9]. Есть ряд методов, такие как PHDpsihtm (часть сервера PredictProtein) [10] или MEMSAT [11] (в режиме online), которые строят предсказания для последовательности, опираясь на результаты поиска гомологичных последовательностей по некоторому белковому банку. В этой работе мы сфокусировали внимание на решении задачи, не прибегая к дополнительному гомологичному поиску. Нам не удалось найти другого метода, который предсказывает положения мембранных участков на основе множественного выравнивания и не требует для работы дополнительных гомологичных данных, кроме метода MEMSAT (в режиме offline). Метод основан на выборе подходящей трансмембранной модели с использованием динамического программирования в соответствии с тем, предпочитают ли аминокислотные остатки располагаться на краях мембраны, в середине или вне ее. Метод дает в качестве результата некоторую последовательность–маску, каждый символ которой "говорит" о том, где находятся аминокислотные остатки из соответствующего столбца выравнивания – в мембране или снаружи. Так как нельзя быть в точности уверенным, что положение мембранного участка жестко фиксировано (белковая молекула "дышит", т.е. происходят слабые колебания звеньев цепи), то более адекватным было бы предсказание того, какова вероятность для данной аминокислоты находиться в мембране (назовем это "мембранный вероятностный профиль"). Для аминокислот, лежащих внутри мембраны, эта вероятность должна быть высока, а на краях она должна плавно опускаться до нуля. Такие вероятности можно получать на основе построения скрытой марковской модели (HMM) мембранных и внешних областей. Используя алгоритм forward-backward [12], мы можем вычислять вероятность того, насколько та или иная аминокислота укладывается в мембранную часть модели.

Целью работы было построение набора кластеров мембранных белков с известной трехмерной структурой, для которых можно построить адекватное структурное выравнивание, а также разработка метода построения мембранного профиля для столбцов множественного выравнивания.

## МЕТОДЫ

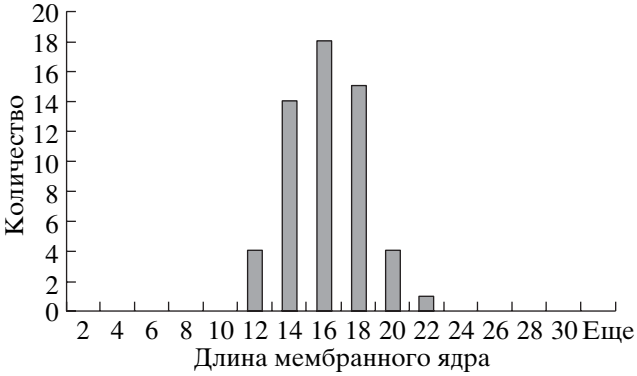
Метод формирования мембранного профиля, где каждому столбцу множественного выравнивания приписывается вероятность того, что аминокислотные остатки этого столбца лежат в мембране, состоит в следующем. На основе множественного выравнивания строится частотный аминокислотный профиль (частотная матрица). Для этого строится матрица попарных эволюционных расстояний между последовательностями на основе величин попарного сходства (identity) методом пуассоновской коррекции [13]: $d = -\log((20\max\{1.1/20, id\} - 1)/19)$, где $d$ – эволюционное расстояние, а $id$ – доля столбцов выравнивания с совпавшими аминокислотами. Далее строится филогенетическое дерево методом ближайшего соседа [14] и каждой последовательности приписывается вес простым, но эффективным методом, предложенным Герштейном и соавт. [15]. Веса обладают следующим свойством. Если у нас имеется $k$ одинаковых последовательностей, то они получают веса $1/k$, а последовательность, не похожая ни на одну другую, получает вес 1. Частотный профиль формируется путем усреднения всех единичных профилей последовательностей с учетом их веса. Для формирования результата используется алгоритм forward-backward на основе скрытой марковской модели (HMM), аналогичной той, что использована в сервере TMHMM [16]. В этой модели различаются состояния для аминокислот, находящихся в цитоплазме, для аминокислот, смотрящих наружу клетки, и двух последовательностей состояний, соответствующих белковой цепи, пересекающей мембрану изнутри наружу и наоборот. Выделяются две группы состояний модели, приходящиеся на границы мембраны. Параметры модели обучены на выборке размеченных

одиночных последовательностей, доступной на сайте сервера TMHMM. Этот сервер предсказывает положение мембранных участков на одиночной последовательности и не умеет работать с аминокислотным частотным профилем или с множественным выравниванием. В его основе лежит алгоритм Витерби (см. [12]), который применяется для нахождения оптимального пути, но, в отличие от алгоритма forward-backward, не может быть использован для построения вероятностного профиля.

*Построение тестовой выборки*

Для проверки метода сформирована выборка эталонных множественных выравниваний. Для этого взяты все последовательности мембранных белков с известной пространственной структурой (442 белка) с сайта сервера PDBTM [17]. Далее были построены все попарные выравнивания с использованием программы CLUSTALW [1]. Если встречали пары белков со сходством (identity) не менее 95%, то из них оставляли один. Далее проводили кластеризацию по попарному сходству методом ближайшего соседа [14] с нижним порогом 20%. Если кластер оказывался размером более 20 белков, то нижний порог для него поднимали до тех пор, пока он не разделялся на меньшие кластеры. После этого рассматривали только кластеры, содержащие не менее 3 белков. Для каждого кластера проводили множественное структурное выравнивание трехмерных структур белков с использованием сервера MAMMOTH [18]. Если качество выравнивания было очень низким (мало столбцов выравнивания, достоверных с точки зрения метода), то отбрасывали самый дальний представитель кластера и кластер выравнивали вновь.

В результате использования данной процедуры получили 11 кластеров из 55 белков. Доля структурно надежных столбцов выравниваний находится в диапазоне от 24 до 86%, в среднем составляя 63%. Размер кластеров характеризуется диапазоном от 3 до 8 белков, средний размер – 5 белков. Далее проверяли принадлежность белков кластеров к структурным семействам по классификации SCOP [19] и CATH [20]. В одном кластере обнаружили двухдоменную структуру, причем есть белки, в которых присутствует только один из двух доменов. В трех кластерах имеются белки, структурные семейства которых в обеих классификациях не обозначены. В двух кластерах присутствуют белки из разных семейств, в одном случае смешаны семейства бактериородопсинов (f.13.1.1 по SCOP) и сукцинатдегидрогеназ/фумаратредуктаз (f.21.2.2 по SCOP), во втором – семейства белков реакционного центра фотосистемы I (f.31.1.1 по SCOP) и (c.37.1.12 по SCOP), ATPазного домена ABC-транспортера.




Гистограмма распределения длин мембранных ядер.

*Построение достоверной мембранной разметки*

В каждом белке каждого кластера разметили участки белковой последовательности, лежащие в мембране, на основе алгоритма TMDET [21], определяющего наиболее вероятное положение мембраны в трехмерной структуре. Для того чтобы избежать ошибочной классификации участка белковой цепи как мембранного из-за неточного предсказания положения мембраны алгоритмом TMDET, были введены "серые" области по краям мембраны толщиной в 5 ангстрем. Если какой-то участок белковой цепи лежит только в "серой" области, то он не считается мембранным. Разметки наносили на структурные выравнивания, и на этой основе сформировали общую мембранную разметку (мембранные ядра). В ядра вошли те колонки структурного выравнивания, в которых все безделеционные позиции помечены как мембранные. Руководствуясь выходной информацией сервера MAMMOTH о достоверности структурного выравнивания в тех или иных столбцах, ядра разделяли на два класса – заслуживающие доверия и не заслуживающие. В первый класс попадали ядра, где две трети столбцов имеют выравнивание, достоверное с точки зрения MAMMOTH, а также длина которых не меньше пяти столбцов. Ядра второго класса были изъяты из рассмотрения.

Всего в результате работы процедуры получили 56 мембранных ядер; в среднем, на выравнивание приходится 5 ядер; количество ядер в выравнивании колеблется между 1 и 12. Кроме этого, в трех выравниваниях изъяли из рассмотрения пять сомнительных ядер, в которых менее 60% столбцов являются достоверными с точки зрения структурного выравнивания. Распределение длин ядер представлено на рисунке.

*Методы предсказания мембранной разметки по выравниванию*

Проверяли следующие методы предсказания мембранной разметки: MEMSAT, FWDBCK, ос-

Качество предсказания мембранной разметки разными методами

| Метод | quality_five[а] | quality_half[б] |
|---|---|---|
| MEMSAT | 0.964 | 0.964 |
| FWDBCK | 0.977 | 0.966 |
| HMMTOP (усp.) | 0.934 | 0.934 |
| HMMTOP (ориг.) | 0.916 | 0.914 |

[а] Качество предсказания, которое есть число ядер, покрытых любым предсказанным мембранным участком хотя бы на пять столбцов, деленное на максимум из числа ядер и числа предсказанных участков.
[б] Качество предсказания, которое есть число ядер, в каждом из которых хотя бы 50% столбцов покрыто любым предсказанным мембранным участком, деленное на максимум из числа ядер и числа предсказанных участков.

нованный на описанном выше методе формирования трансмембранного вероятностного профиля, и метод усреднения результатов сервера HMMTOP [22] по белкам в выравнивании (далее усреднение HMMTOP). На вход серверу MEMSAT подавали аминокислотные частотные профили выравниваний с учетом весов последовательностей, но без учета делеций. Разметку на мембранные участки FWDBCK формировали так, что столбцы, вероятность нахождения которых в мембране была не меньше 0.8, объявляли мембранными. Если встречалось менее пяти мембранных столбцов, стоящих вместе, то их не считали мембранными.

Метод усреднения HMMTOP устроен так. При выравнивании на каждую последовательность наносится мембранная разметка, предсказываемая сервером HMMTOP. Столбцы, в которых, как минимум, две трети безделеционных позиций помечены как мембранные, объявляли мембранными. Если встречалось менее пяти мембранных столбцов, стоящих вместе, то их не считали мембранными.

### *Оценка качества работы метода HMMTOP*

Для того чтобы убедиться в том, что методы предсказания мембранной разметки, опирающиеся на выравнивание, работают лучше, чем методы, имеющие дело только с одной последовательностью, проверяли качество работы метода HMMTOP для каждой белковой последовательности каждого кластера. С этой целью для каждой последовательности формировали "сужение" информации о достоверности столбцов в структурном выравнивании соответствующего кластера путем выбрасывания столбцов, в которых рассматриваемая последовательность имеет делецию. Аналогично строили разметку последовательности на ядра, которая есть сужение разметки на ядра всего выравнивания. Далее, к разметке последовательности на ядра и к разметке, предсказанной методом HMMTOP, применяли аналогичный описанному выше фильтр, позволяющий игнорировать мембранные участки и ядра с малой длиной и с малой степенью пересечения с "маской достоверности". Результат приведен в таблице напротив пункта "HMMTOP (ориг.)".

### *Оценка качества предсказания*

Прежде чем оценивать качество предсказанной тем или иным методом разметки, из нее выкидывают те мембранные участки, которым нельзя доверять. Полагали, что участку можно доверять, если две трети покрываемых им столбцов имеют структурное выравнивание, достоверное с точки зрения MAMMOTH, а также если длина участка не меньше пяти столбцов. Для каждого метода предсказания и для каждого кластера были посчитаны две величины оценки качества. Первая – под названием quality_five, которая есть число ядер, покрытых любым предсказанным мембранным участком хотя бы на пять столбцов, деленное на максимум из числа ядер и числа предсказанных участков. Вторая – quality_half, которая есть число ядер, в каждом из которых хотя бы 50% столбцов покрыты любым предсказанным участком, деленное на максимум из числа ядер и числа предсказанных участков. Как видно из таблицы, лучшие результаты дает метод FWDBCK.

## РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

На данный момент развития биоинформатики ощущается нехватка данных по мембранным белкам, на которых можно проверять качество работы методов автоматического предсказания положений мембранных участков и методов, строящих множественные выравнивания аминокислотных последовательностей. В секции базы данных Balibase [4], посвященной мембранным белкам, для большей части выравниваний не представлена мембранная разметка, которая могла бы быть получена на основе анализа известных трехмерных структур, а также не выделяются столбцы, выравниванию в которых можно доверять с точки зрения метода, строящего структурные выравнивания.

В данной работе построена выборка кластеров мембранных белков, где для каждого кластера построено структурное множественное выравнивание и нанесены мембранные ядра, т.е. группы столбцов, "мембранность" которых подтверждена структурой каждого белка кластера. Хотя ядра имеют среднюю длину 15.5, которая немного меньше, чем 21 (общепринятая средняя длина мембранного участка белковой цепи), но при этом они не содержат сомнительные столбцы. Также выделены столбцы, достоверные с точки

зрения метода структурного выравнивания. Таким образом, данная выборка (несмотря на малый размер) может с уверенностью быть использована для проверки качества методов, предсказывающих мембранную разметку или строящих множественные выравнивания.

С другой стороны, разработан метод формирования мембранного вероятностного профиля. Адекватность метода проверена на основе предсказания по нему мембранной разметки (см. FWD-BCK в таблице). Качество этого предсказания оказалось несколько лучше, чем у наиболее точных методов, в которых не прибегают к гомологичному поиску в дополнительном банке данных.

Также такой профиль может быть использован при построении множественных выравниваний последовательностей мембранных белков. Если метод выравнивания "прогрессивный", то на каждом шаге соединения профилей двух подвыравниваний в один можно улучшать результирующее выравнивание, варьируя для каждого столбца такие параметры, как матрица замен, штрафы за открытие и продолжение делеций, в зависимости от того, какова вероятность для аминокислот данного столбца лежать в мембране.

Кроме того, разработан интернет-сервер, где пользователь может для своего выравнивания получить мембранный вероятностный профиль.

Сервер и тестовая выборка доступны по адресу http://bioinf.fbb.msu.ru/fwdbck/.

Работа получила финансовую поддержку Российской академии наук (программы "Молекулярная и клеточная биология" и "Происхождение и эволюция биосферы"), фонда Howard Hughes Medical Institute (грант 55000309) а также Российского фонда фундаментальных исследований (05-04-48759).

## СПИСОК ЛИТЕРАТУРЫ

1. Thompson J.D., Higgins D.G., Gibson T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**, 4673–4680.
2. Jaroszewski L., Li W., Godzik A. 2002. In search for more accurate alignments in the twilight zone. *Protein Sci*. **11**, 1702–1713.
3. Zhang H., Cramer W.A. 2005. Problems in Obtaining Diffraction-quality Crystals of Heterooligomeric Integral Membrane Proteins. *J. Struct. Funct. Genomics*. **6**, 219–223.
4. Bahr A., Thompson J.D., Thierry J.C., Poch O. 2001. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res*. **29**, 323–326.
5. Sutormin R.A., Rakhmaninova A.B., Gelfand M.S. 2003. BATMAS30: amino acid substitution matrix for alignment of bacterial transporters. *Proteins*. **51**, 85–95.
6. Heringa J. 1999. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Comput. Chem*. **23**, 341–364.
7. Ng P.C., Henikoff J.G., Henikoff S. 2000. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics*. **16**, 760–766.
8. Do C.B., Mahabhashyam M.S., Brudno M., Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*. **15**, 330–340.
9. Chen C.P., Kernytsky A., Rost B. 2002. Transmembrane helix predictions revisited. *Protein Sci*. **11**, 2774–2791.
10. Rost B., Liu J. 2003. The PredictProtein server. *Nucleic Acids Res*. **31**, 3300–3304.
11. Jones D.T. 1998. Do transmembrane protein superfolds exist? *FEBS Lett*. **423**, 281–285.
12. Krogh A., Mian I.S., Haussler D. 1994. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res*. **22**, 4768–4778.
13. Zuckerkandl E., Pauling L. 1965. Molecules as documents of evolutionary history. *J. Theor. Biol*. **8**, 357–366.
14. Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. **4**, 406–425.
15. Gerstein M., Sonnhammer E.L., Chothia C. 1994. Volume changes in protein evolution. *J. Mol. Biol*. **236**, 1067–1078.
16. Sonnhammer E.L., von Heijne G., Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. **6**, 175–182.
17. Tusnady G.E., Dosztanyi Z., Simon I. 2005. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*. **33**, 275–278.
18. Lupyan D., Leo-Macias A., Ortiz A.R. 2005. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*. **21**, 3255–3263.
19. Murzin A.G., Brenner S.E., Hubbard T., Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. **247**, 536–540.
20. Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., Thornton J.M. 1997. CATH – a hierarchic classification of protein domain structures. *Structure*. **5**, 1093–1108.
21. Tusnady G.E., Dosztanyi Z., Simon I. 2005. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*. **21**, 1276–1277.
22. Tusnady G.E., Simon I. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol*. **283**, 489–506.