Биомедицинская химия, 2009 том 55, вып. 5, с. 534-543.

БИОИНФОРМАТИКА

УДК 543.51.061:543.54.45:543.8

СРАВНЕНИЕ АЛГОРИТМОВ ПРЕДСКАЗАНИЙ ВЗАИМОСВЯЗАННЫХ БЕЛКОВ НА ПРИМЕРЕ МЕТОДА ФИЛОГЕНЕТИЧЕСКИХ ПРОФИЛЕЙ

М.А. Пятницкий*, А.В. Лисица, А.И. Арчаков

НИИ Биомедицинской химии им. В.Н. Ореховича РАМН, 119121, Москва, Погодинская ул., д.10; эл. почта: mpyat@bioinformatics.ru

Задача вычислительной интерактомики заключается в предсказании функциональных взаимосвязей между белками. Одним из подходов к решению этой задачи средствами сравнительной геномики является анализ сходства филогенетических профилей белков. В отличие от большинства предложенных методов, позволяющих выявлять попарные взаимодействия, в работе рассматривается применение кластерного анализа для определения функциональных модулей, состоящих из нескольких белков. Проведен кластерный анализ филогенетических профилей белков *E. coli* с использованием нескольких методов кластерного анализа и различных способов оценки расстояния между профилями. Показано, что при оценке расстояния между профилями по формуле Хэмминга и кластеризации методом Уорда состав кластеров в наибольшей степени отвечает распределению белков по известным метаболическим путям. Предложенную методику оценки точности результатов предсказания взаимосвязанных белков предлагается применять для сопоставления различных алгоритмов вычислительной интерактомики.

Ключевые слова: взаимосвязанные белки, кластерный анализ, филогенетические профили, интерактомика.

ВВЕДЕНИЕ. В постгеномной молекулярной биологии происходит переход от расшифровки последовательностей генов к детальному анализу функций кодируемых ими белков и построению интерактома – сетей взаимосвязей между биомолекулами. Примерами могут служить сети белок-белковых взаимодействий, регуляторные и метаболические сети. Сетевое представление молекулярных процессов позволяет достичь нового уровня понимания функционирования и эволюции генома и клетки в целом.

За последние 10 лет созданы вычислительные методы, предсказывающие функциональные взаимосвязи между белками. Такие методы опираются на данные сравнительной геномики, а степень взаимосвязи между белками определяется свойствами соответствующих генов: характер распределения гомологов в других геномах [1], положение и относительный порядок следования генов на хромосоме [2], частота слияний генов [3]. Подобные методы называются контекстно-ориентированными, поскольку используемые свойства генов имеют смысл лишь при одновременном их исследовании в ряду геномов.

Предложены различные варианты модификации основных алгоритмов для предсказания взаимосвязанных белков. В большинстве работ используется единый методический подход: на первом этапе вводится метрика, определяющая взаимосвязь между парой белков. Затем выбирается определенное значение метрики (порог), и все пары белков для которых метрика больше порога

* - адресат для переписки

считаются взаимосвязанными. Результаты работы алгоритма сравниваются с базой данных по взаимодействующим белкам, оценивается число ложнопозитивных и ложнонегативных предсказаний.

В данной работе используется альтернативный подход, ориентированный на поиск групп взаимосвязанных белков, то есть функциональных модулей [4]. Это достигается применением кластерного анализа к матрице расстояний между белками. Такая постановка задачи является более осмысленной с биологической точки зрения, позволяет раскрыть контекст предсказанных взаимодействий [5, 6], облегчает определение функций для неохарактеризованных белков [7].

В представленной работе мы предлагаем методику сравнения алгоритмов предсказания взаимосвязанных белков. Предлагается оценивать работу таких алгоритмов с точки зрения задачи о сравнении разбиений – определить степень соответствия заданного экспертами истинного группирования взаимосвязанных белков с результатом работы алгоритма. Это позволяет количественно сопоставлять различные подходы, учитывать возможность принадлежности белков к нескольким группам и автоматически находить оптимальное число групп.

Для предсказания групп взаимосвязанных белков нами использовался метод филогенетических профилей (ФП), согласно которому функционально взаимосвязанные белки также связаны и эволюционно [1]. Предполагается, что гены, кодирующие взаимодействующие белки, либо совместно наследуются вновь образованным видом, либо элиминируются естественным отбором. Каждый белок изучаемого организма характеризуется бинарным вектором (профилем), определяющим наличие гомолога данного белка в ряду геномов. При наличии достаточного числа геномов сравнения каждая пара взаимосвязанных белков (в рамках структурного комплекса или метаболического пути) будет иметь схожие ФП.

Целью настоящей работы было сравнение результатов кластеризации ФП с известными данными о распределении белков по метаболическим путям. Сведения о метаболических путях были взяты из базы данных KEGG, ФП для соответствующих белков были получены из базы данных COG. Матрицу попарных расстояний между ФП анализировали с помощью различных методов кластерного анализа. Состав полученных кластеров сопоставляли с метаболическими путями KEGG путем расчета внешнего индекса. С использованием нескольких внутренних индексов получали оптимальное количество кластеров и его также сравнивали с количеством метаболических путей в KEGG. Предложенная методика может применяться для оптимизации параметров алгоритмов предсказания интерактома.

МЕТОДИКА.

Метаболические пути. Информацию о метаболических путях *E. coli* загружали из базы данных KEGG [8]. Общее число метаболических путей KEGG составило 125, а общее число белков – 1133. Каждый метаболический путь рассматривали как группу взаимосвязанных белков [9,10]. Были исключены 194 белка, принадлежащие только к двухкомпонентным сигнальным системам (код 02020) и белкам-транспортерам семейства ATP-связывающих кассет (код 02010).

Распределение белков по метаболическим путям представляли в виде матрицы принадлежности $\boldsymbol{M}$, где строки соответствовали белкам, а столбцы соответствовали метаболическим путям. Каждый элемент матрицы m_{ij} характеризовал принадлежность i-го белка к j-ому метаболическому пути. Поскольку белок может участвовать в нескольких метаболических путях, то задавали значение $m_{ij} = n_j/N_i$, где n_j - число белков в j-ом метаболическом пути, а N_i –суммарное число белков в остальных метаболических путях, к которым принадлежит i-ый белок. Элементы матрицы нормировали для выполнения ограничения $\Sigma_j m_{ij} = 1$. Матрица принадлежности содержала 939 строк и 123 столбца.

Филогенетические профили. Для загруженных из KEGG белков из базы данных COG [11] получили информацию о том, в геномах каких организмов есть ортологи для исследуемых белков, т.е. построили ФП. Для каждого ФП

единицей отмечали наличие ортолога в определенном геноме сравнения, нулём - его отсутствие. В результате были получены результаты сопоставления 939 белков метаболических путей *E. coli* с 65-ю геномами сравнения, в том числе 7 геномами эукариот и 58 геномами прокариот. После удаления 50 белков, ортологи которых присутствовали во всех геномах сравнения, осталось 889 ФП.

Использовали несколько вариантов определения расстояния между двумя ФП. Для двух профилей X и Y обозначим $S_{ij}(i,j\in\{0,1\})$ число появлений i в X и j в Y на соответствующих позициях ФП, N – длина ФП. Использовали расстояние Хэмминга $D_H = S_{10}+S_{01}/S_{00}+S_{01}+S_{10}+S_{11}$, расстояние Жаккара $D_J = (S_{10}+S_{01})/(S_{11}+S_{10}+S_{01})$, расстояние Кульчинского $D_K = (S_{10}+S_{01} - S_{11}+N)/(S_{10}+S_{01}+N)$.

Дополнительно в качестве расстояния использовали вероятность случайного совпадения двух ФП. Пусть x и y – число единиц в профилях X и Y, соответственно. Обозначим $P(z|x,y,N)$ вероятность наблюдать z случайных совпадений единиц в обоих ФП. Определим w_z как число способов распределить z совпадений между N геномами и $\overline{w}_z$ как число способов распределить оставшиеся $x - z$ *и* $y - z$ белков среди оставшихся $N–z$ геномов. Общее число способов, которое может дать z совпадений при заданных N, x, y равно $\overline{w}_z \times w_z$, где

$$w_z = \binom{N}{z} \quad \text{и} \quad \overline{w}_z = \binom{N-z}{x-z}\binom{N-x}{y-z} .$$

Для получения искомой вероятности, разделив на число способов распределения x и y белков среди N геномов

$$W = \binom{N}{x}\binom{N}{y} \quad \text{имеем} \quad P(z \mid N,x,y) = \frac{\overline{w}_z w_z}{W} \text{ [12].}$$

Для каждого варианта определения расстояния строили матрицу попарных расстояний между ФП для всех белков.

Кластерный анализ. Определение групп взаимосвязанных белков осуществляли посредством кластерного анализа матриц расстояний между ФП, построенным для различных мер сходства. Для построения иерархии кластеров использовали как агломеративный подход (постепенное объединение объектов в кластеры), так и дивизивный подход (последовательное разделение групп). Для агломеративной кластеризации применяли методы ближних, средних, дальних связей и метод Уорда. Для дивизивной кластеризации применяли метод DIANA [13]. Помимо иерархической кластеризации применяли итеративное разбиение, напрямую возвращающее состав кластеров. Для этого использовали метод PAM, устойчивую к выбросам версию известного алгоритма k-средних [13].

Оценка результатов кластерного анализа. Для оценки соответствия результатов кластерного анализа метаболическим путям базы данных KEGG использовали внешний индекс, который рассчитывали как минимальное расстояние между **M** (матрицей принадлежности для KEGG) и **M′** (матрицей принадлежности для результатов кластерного анализа): $d(\mathbf{M},\mathbf{M}') = \min_{\mathbf{P}}||\mathbf{M} - \mathbf{M}'\mathbf{P}||$, минимизацию индекса проводили по всем матрицам-перестановкам **P**.

Для оценки результатов кластеризации без привлечения данных о метаболических путях использовали следующие внутренние индексы: Γ-статистику Хьюберта (корреляция между матрицей расстояний и матрицей принадлежности); индекс Дана (отношение минимального межкластерного расстояния к максимальному внутрикластерному диаметру); индекс Дэйвиса (среднее расстояние между каждым кластером и ближайшим к нему кластером); отношение среднего внутрикластерного расстояния к среднему межкластерному расстоянию; индекс “ширина силуэта“,

$$sw = \frac{1}{n}\sum_{i=1}^{n}\frac{\min\left(d(i,C)\right) - a_i}{\max\left(a_i, \min d(i,C)\right)}$$, где a_i – среднее расстояние между i-ым и другими

объектами одного кластера, *d(i,C)* – среднее расстояние от объекта *i* до других объектов кластера *C* [13].

РЕЗУЛЬТАТЫ. Для получения эталонных групп взаимосвязанных белков использовали базу данных KEGG. Филогенетические профили для 3764 белков *E. coli K12* были загружены из базы данных COG. Белки, для которых отсутствовала аннотация в базе данных KEGG, были исключены. В результате построенная матрица филогенетических профилей состояла из 889 строк и 65 столбцов.

В работе использовали различные способы определения расстояний между ФП белков. Был проведен предварительный анализ того, насколько различные меры расстояния между ФП определяют взаимосвязи соответствующих белков в KEGG. Для всех пар белков были вычислены различные меры расстояний между ФП с шагом 0,02 для расстояний Хэмминга, Жаккара и Кульчинского и с шагом *1/e* для вероятности случайного совпадения. Для каждого интервала расстояний было вычислено отношение числа пар белков, участвующих в общем пути KEGG к общему числу пар в данном диапазоне расстояний. Для малых расстояний между ФП эта доля должна быть максимальной, а для больших расстояний – минимальной.

Результаты представлены на рисунке 1. Для расстояния Хэмминга (рис. 1а) доля пар белков участвующих в общем метаболическом пути снижается до 16% при расстоянии между соответствующими ФП, превышающем 0,08, т.е. при 5 различающихся элементах. Для большого диапазона расстояний доля взаимосвязанных согласно KEGG пар белков не превышает 3%. В то же время для белков на расстоянии 0,9 (58 различающихся элементов) доля таких пар достигает 5%, что связано с феноменом неортологичного замещения генов (гены с различной последовательностью выполняют близкие функции) [14]. В этом случае белки имеют комплементарные ФП, и расстояние Хэмминга для таких пар велико.

Для расстояния, определяемого как вероятность случайного совпадения двух профилей (рис. 1б), зависимость имеет другой характер. Для вероятности меньшей чем 1E-12 доля взаимосвязанных в KEGG пар белков превышает 50%, однако число таких пар составляет всего 0,3% от общего количества пар. При вероятности случайных совпадений, превышающей 1E-6 доля взаимосвязанных пар выходит на плато и составляет 2-3%. С поправкой Бонферрони на число произведенных попарных вычислений (394716) такая вероятность не дает оснований отклонить гипотезу о случайном совпадении ФП.

Для расстояний Кульчинского и Жаккара характер зависимости сходен с таковым для расстояния Хэмминга. При этом соответствующие значения доли связанных в KEGG белков для одинаковых интервалов расстояний меньше по сравнению с расстоянием Хэмминга. Следовательно, расстояния Жаккара и Кульчинского хуже отражают взаимосвязи KEGG.

Согласно методу филогенетических профилей, белки, имеющие схожее распределение гомологов в ряду организмов (а значит и схожие ФП), считаются взаимосвязанными. Для выявления кластеров взаимосвязанных белков *E. coli K12* провели кластерный анализ ФП используя ряд методов кластеризации и несколько способов определения расстояний между ФП (см. раздел “Методика”).

Был проведен перебор 24 комбинаций методов кластеризации и способов определения расстояний. Для каждой комбинации рассчитывали внешний и внутренние индексы, оценивающие результаты проведенного кластерного анализа, что позволило количественно сравнивать различные варианты. Показаны только некоторые характерные кривые в виде зависимости значений индексов от числа кластеров, на которые проводилось разбиение. Дополнительные материалы в виде графиков для всех индексов и комбинаций представлены на веб-сайте www.bioinformatics.ru.

На рисунке 2 представлена зависимость минимального расстояния между матрицей принадлежности KEGG и матрицей принадлежности полученной кластеризации от числа кластеров. Меньшие значения индекса обозначают лучшее соответствие полученных кластеров белков с метаболическими путями KEGG.

Наихудшие результаты были получены при использовании метода ближней связи. Отметим, что этот метод оказывается наихудшим для всех использованных индексов вне зависимости от используемого расстояния. Это является следствием эффекта “образования цепочки”, когда все объекты постепенно включаются в один доминирующий кластер. Наилучшие результаты дает кластеризация методом Уорда и расстояние Хэмминга; достигаемый минимум индекса равен 33,45 при 128 кластерах (отмечено стрелкой). Поскольку общее число метаболических путей KEGG для *E. coli* равно 123, то можно отметить хорошее согласие в оценке числа кластеров. Остальные варианты комбинирования расстояний и кластеризации дают постепенное уменьшение индекса с выходом на плато в районе 130-140 кластеров; это означает, что достигнуто оптимальное соответствие метаболическим путям KEGG.

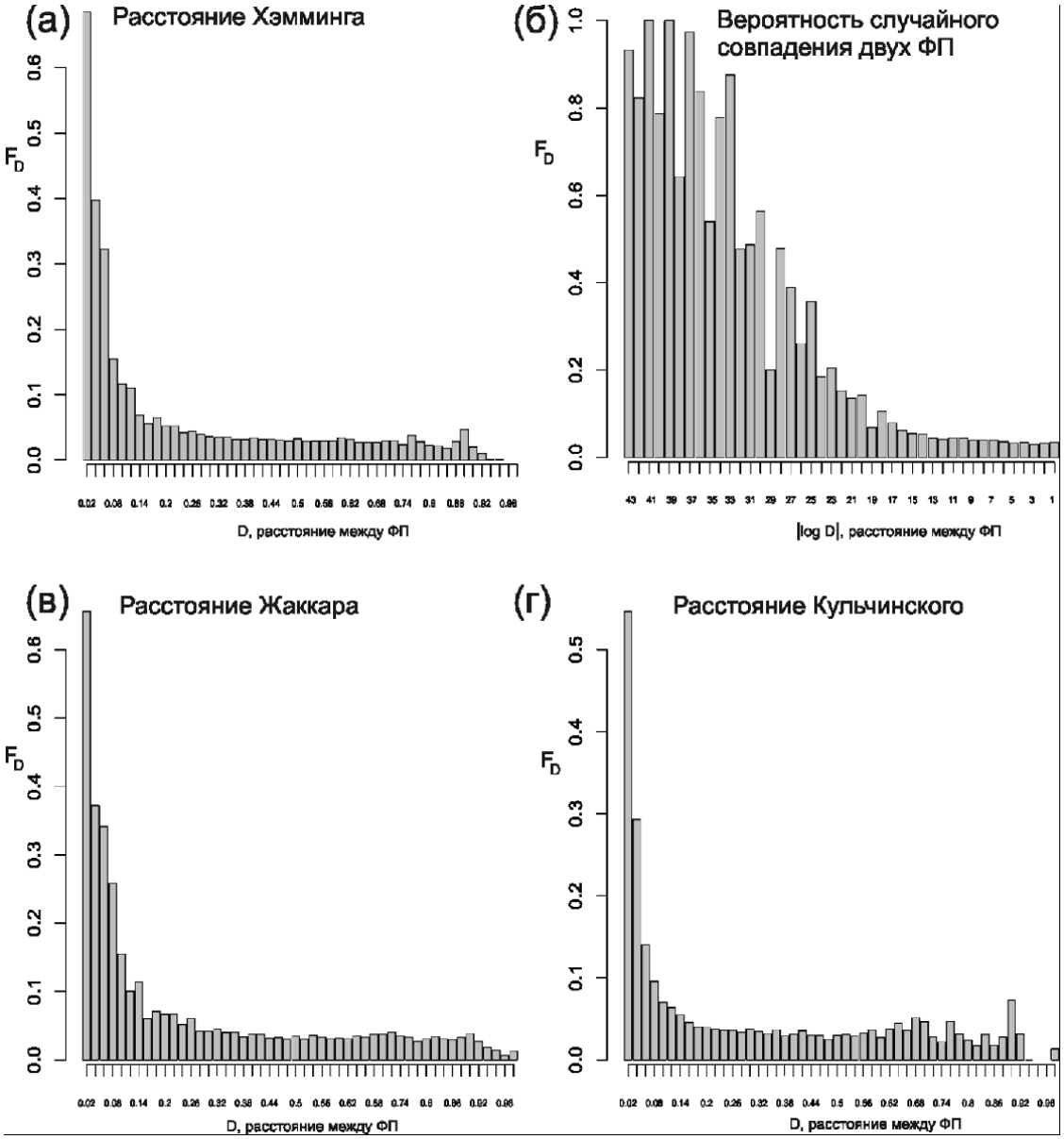

Рисунок 1.

Сравнение различных расстояний между ФП как отражение данных о взаимосвязанных белках KEGG. По оси абсцисс – расстояние D между ФП для всех пар белков. По оси ординат – доля пар белков F_D, участвующих в общем метаболическом пути KEGG для заданного диапазона расстояний. (а) – расстояние Хэмминга; (б) – модуль логарифма вероятности случайного совпадения ФП; (в) – расстояние Жаккара; (г) расстояние Кульчинского.

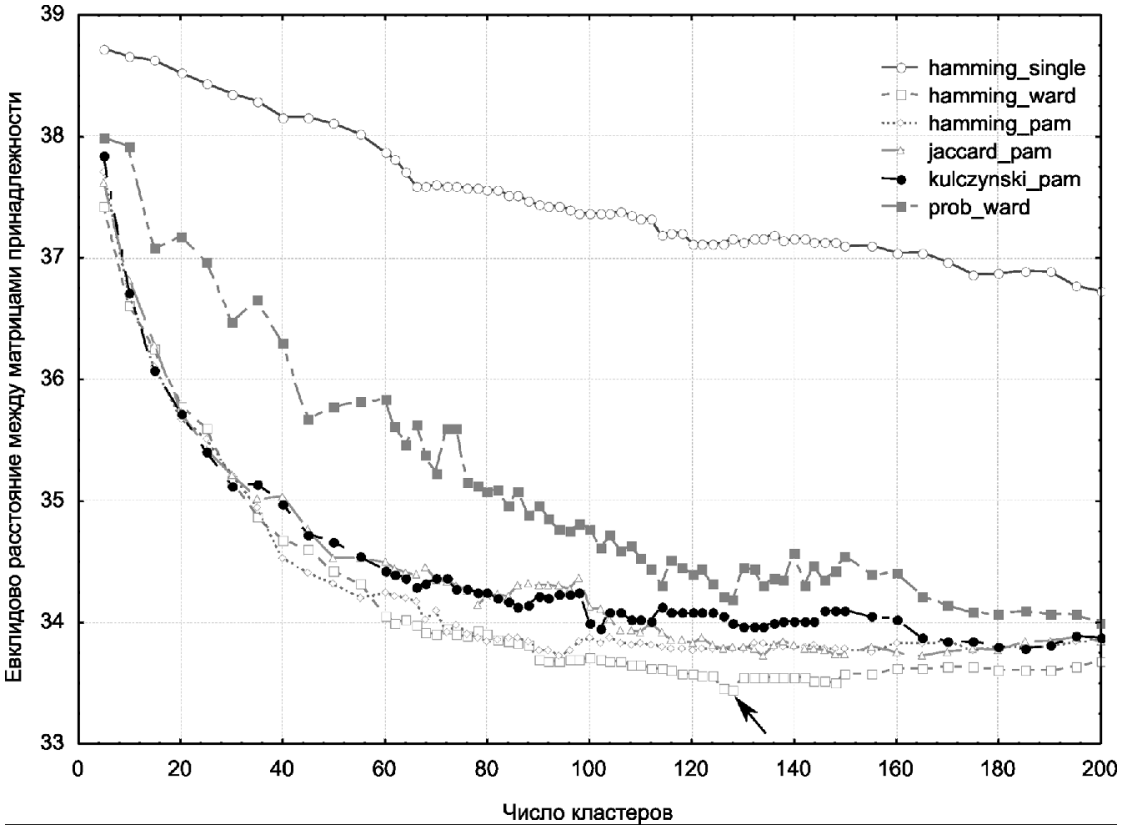

Рисунок 2.
Значения внешнего индекса в зависимости от количества кластеров. Использованы расстояние Хэмминга (hamming), Жаккара (jaccard), Кульчинского (kulczynski), вероятность случайного совпадения ФП (prob). В качестве метода кластеризации использован метод Уорда (ward), метод PAM и метод ближней связи (single). Стрелкой отмечен минимум индекса (128 кластеров), обеспечивающий наилучшее совпадение с KEGG (123 группы).

В результате минимизации данного индекса для каждой комбинации метода кластеризации и меры расстояния получили матрицу перестановки **P**, оптимально сопоставляющую каждую группу полученного разбиения соответствующему метаболическому пути KEGG. Для кластеризации методом Уорда с расстоянием Хэмминга при 128 кластерах 38,9% белков, входящих в один из метаболических путей KEGG, оказались в одном кластере.

На рисунке 3 показаны значения индекса “ширина силуэта“ в зависимости от числа кластеров. Результаты принципиально отличаются от представленных на рисунке 2, поскольку индекс “ширина силуэта” является внутренним, то есть при его расчете не использовали данные о метаболических путях KEGG. Большие значения индекса соответствуют лучшему качеству разбиений. Локальный максимум достигается при 120 кластерах методом Уорда с расстоянием равным вероятности случайного совпадения ФП, что хорошо согласуется с истинным числом кластеров (123). При этом 34,8% белков входящих в один из метаболических путей KEGG оказались в одном кластере. Использование метода полной связи дает скачок при 132 кластерах. Тем самым, анализ динамики данного индекса позволяет оценить оптимальное количество кластеров, не прибегая к использованию обучающих данных (метаболических путей KEGG). В то же время, для некоторых методов (в особенности для кривых, полученных на основе

расстояния Хэмминга) затруднительно определить выраженную точку перелома, в которой происходит резкое замедление агломерации кластеров. Для большинства методов отмечается почти монотонный рост значений индекса с увеличением числа кластеров, что характерно для данного индекса, при этом значения выходят на плато при числе кластеров превышающим 140.

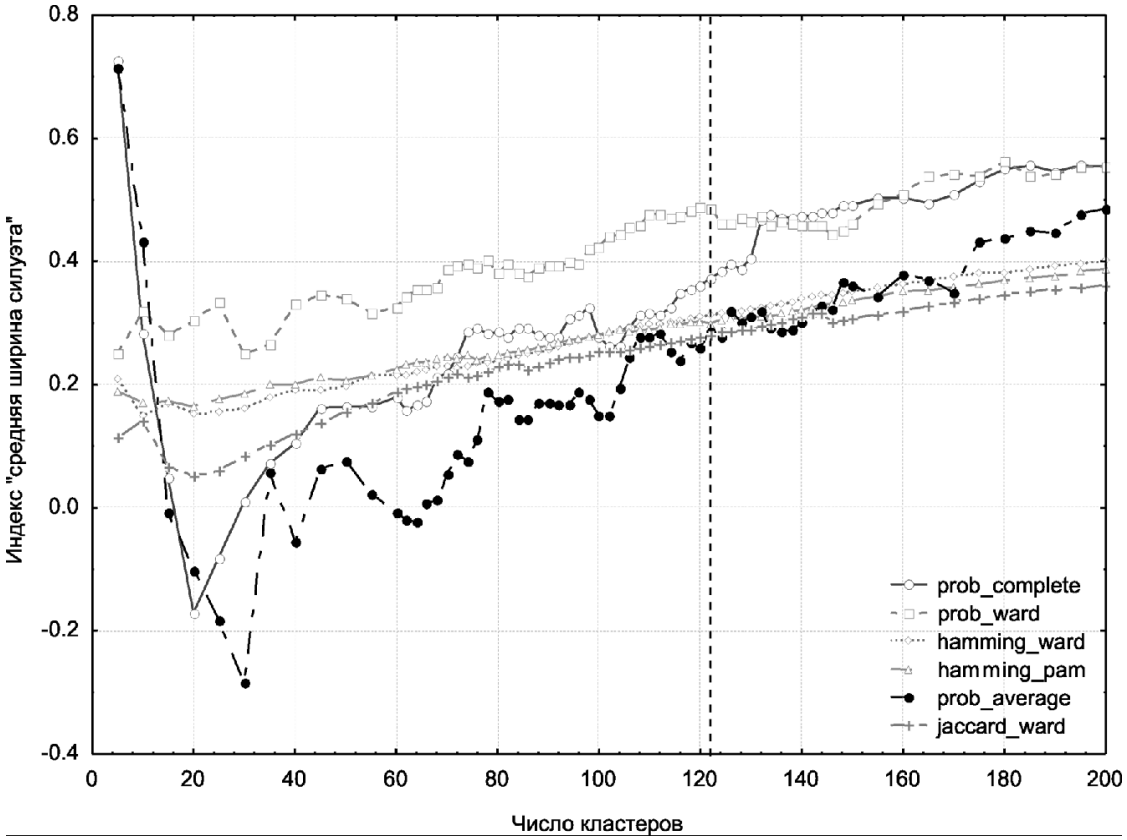

Рисунок 3.
Значения внутреннего индекса “ширина силуэта” в зависимости от количества кластеров. Использованы расстояние Хэмминга (hamming), Жаккара (jaccard), вероятность случайного совпадения ФП (prob). В качестве метода кластеризации использован метод Уорда (ward), метод PAM и метод дальней связи (complete). Пунктиром отмечено количество метаболических путей KEGG (123).

ОБСУЖДЕНИЕ. Метод ФП широко используется для предсказания функциональных взаимосвязей между белками [15-17], их внутриклеточной локализации [18] и аннотации геномов[19]. В работе [1] предполагается, что белки являются взаимосвязанными, если их ФП удалены менее чем на 3 по расстоянию Хэмминга.

Наибольший интерес представляет выявление групп взаимосвязанных белков (функциональных модулей) соответствующих метаболическим путям и белковым комплексам, что дает возможность изучать сети внутриклеточных белок-белковых взаимодействий, разрабатывать новые лекарственные средства [20], осуществлять поиск неохарактеризованных клеточных систем [21].

В большинстве работ по применению метода ФП рассматриваются взаимосвязи исключительно между парами белков [9, 10, 17, 21]. Два белка считаются взаимосвязанными, если расстояние между их ФП меньше выбранного порогового значения. При этом изменение величины порога, очевидно, влияет на результат предсказания. Подобный подход не позволяет предсказывать функциональные белковые модули, он также уязвим с точки зрения произвольности в задании порога.

Для поиска групп взаимосвязанных белков в данной работе мы применили кластерный анализ к набору ФП белков протеома *E. coli K12*. Поскольку существует ряд причин, искажающих данные о ФП белков, и, следовательно, ограничивающих число выявляемых взаимосвязей между белками (некорректное установление ортологичных отношений и видоспецифичные утраты генов [22]), то для максимизации предсказательной силы метода были использованы различные варианты кластерного анализа и меры расстояния между ФП.

Для количественной оценки точности предсказаний функциональных модулей в работе предложен единый подход к сравнению алгоритмов предсказания взаимосвязанных белков. Полученное в результате работы алгоритма объединение белков в группы может быть охарактеризовано с помощью ряда индексов. Если известно, “истинное” разбиение белков (в работе использовали метаболические пути KEGG), то применение внешних индексов позволяет оценить соответствие этого разбиения и результатов работы алгоритма. Другим возможным источником экспертных данных о взаимосвязанных белках является система Gene Ontology [23]. Если же экспертные данные о группах белков недоступны, то для оценки полученных кластеров применяются внутренние индексы. Анализ поведения внутренних индексов может дать указания об “истинном” количестве кластеров без каких-либо априорных сведений о распределении белков по группам.

Поскольку каждый индекс учитывает различные аспекты построенной кластеризации (компактность кластеров, расстояние между кластерами, соответствие матрице расстояний и т.д.), то возможны ситуации, при которых однозначный выбор наилучшего группирования белков невозможен. В таких случаях следует выбирать наиболее часто встречающийся вариант или учитывать специфику индексов. Из-за принципиальной разницы между внутренними и внешними индексами их результаты также могут не совпадать. Например, согласно внешнему индексу, равному расстоянию между матрицами принадлежности, оптимальным является сочетание расстояния Хэмминга и метода Уорда, в то время как согласно внутреннему индексу “ширина силуэта”, таковым является использование метода полной связи и вероятности случайного совпадения ФП. Поскольку внешние индексы используют “золотой стандарт”, то их значения точнее оценивают соответствие кластеров “правильному” разбиению. Использование же внутренних индексов соответствует часто встречающемуся на практике случаю, когда “истинное” распределение белков на группы неизвестно.

Нами проведен сравнительный анализ влияния различных параметров метода ФП на точность предсказаний модулей взаимосвязанных белков. Показано, что варьирование способа кластеризации и определения меры схожести между двумя профилями существенным образом влияет на согласие результатов с известными метаболическими путями. Согласно полученным результатам, можно заключить, что в качестве расстояния между ФП предпочтительно использовать либо расстояние Хэмминга, либо вероятность случайного совпадения ФП. В качестве способа кластеризации оптимальным является использование метода Уорда или метода полной связи. При этом соответственно 38,9% и 34,8% белков из одного кластера имеют общий метаболический путь KEGG.

В работе [24] изучались белковые модули, полученные при наложении результатов кластеризации ФП и метаболических путей KEGG. К сожалению, в работе не приведены численные данные о соответствии кластеризации ФП

и данных KEGG, за исключением иллюстрации для белков участвующих в биосинтезе лизина. Авторы использовали метод полной связи и расстояние Жаккара. Интересно отметить, что согласно полученным нами результатам, способ кластеризации в работе [24] оптимален, однако расстояние таковым не является.

Метод ФП может быть применен к большому количеству белков, функция которых не охарактеризована в эксперименте [25]. Например, из 4430 белков *E. coli* только 1416 белков аннотированы в базе данных KEGG, в то время как согласно базе данных TIGR, неизвестна функция 674 неконсервативных и 949 консервативных белков. Предложенная в работе кластеризация ФП может применяться для аннотирования белков с неизвестными функциями: включение неохарактеризованного белка в найденный кластер может быть использовано для предсказания его функциональной роли в клетке. Использование найденных в работе оптимальных параметров метода позволит проводить такие предсказания с максимальной точностью.

Принципиальной трудностью при предсказании взаимосвязанных белков является то, что большинство алгоритмов кластеризации относят белок только к одному кластеру, в то время как согласно базе данных KEGG, из 889 использовавшихся в работе белков *E. coli* 342 белка принадлежат двум и более метаболическим путям. Для расчета всех индексов кроме расстояния между матрицами принадлежности, нечеткое разбиение KEGG было преобразовано в фиксированное. Каждый *i*-ый белок включали только в кластер *j*, для которого значение m_{ij} было максимальным. Избежать эту проблему возможно, если применить алгоритмы нечеткой кластеризации [13], что является предметом работы, проводящейся в настоящее время.

Другим перспективным направлением является разработка методов статистического анализа для автоматического поиска особенностей кривых, подобных приведенным на рисунках 2 и 3. Так, например, для автоматической проверки статистической значимости полученных разбиений необходимо построить распределение интересующего индекса при случайном делении белков на группы, что будет соответствовать нулевой гипотезе об отсутствии структурированности исходных данных.

ЗАКЛЮЧЕНИЕ. В работе был проведен поиск функциональных белковых модулей путем применения кластерного анализа к набору ФП белков протеома *E. coli K12*. Согласно результатам сравнительного анализа различных сочетаний способов кластеризации и расстояний между ФП, наилучшее согласие предсказаний с метаболическими путями из базы данных KEGG достигается при использовании расстояния Хэмминга и кластеризации методом Уорда. Предложенная методика сравнения групп взаимосвязанных белков может применяться для оптимизации параметров алгоритмов предсказания взаимосвязанных белков.

ЛИТЕРАТУРА

1. *Pellegrini M., Marcotte E.M., Thompson M.J., Eisenberg D., Yeates T.O.* (1999) Proc. Natl. Acad. Sci. USA, **96**, 4285-4288.
2. *Overbeek R., Fonstein M., D'Souza M., Pusch G.D., Maltsev N.* (1999) Proc. Natl. Acad. Sci. USA, **96**, 2896-2901.
3. *Marcotte C.J., Marcotte E.M.* (2002) Appl. Bioinformatics, **1**, 93-100.
4. *Snel B., Huynen M.A.* (2004) Genome Res., **14**, 391-397.
5. *Chen J., Yuan B.* (2006) Bioinformatics, **22**, 2283-2290.
6. *Vinogradov A.E.* (2008) Bioinformatics, **24**, 2814-2817.
7. *Dittrich M.T., Klau G.W., Rosenwald A., Dandekar T., Muller T.* (2008) Bioinformatics, **24**, 223-231.
8. *Ogata H., Goto S., Sato K., Fujibuchi W., Bono H., Kanehisa M.* (1999) Nucleic Acids Res., **27**, 29-34.

9. *Jothi R., Przytycka T.M., Aravind L.* (2007) BMC Bioinformatics, **8**, 173.
10. *Sun J., Xu J., Liu Z., Liu Q., Zhao A., Shi T., Li Y.* (2005) Bioinformatics, **21**, 3409-3415.
11. *Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., Koonin E.V., Krylov D.M., Mazumder R., Mekhedov S.L. et al.* (2003) BMC Bioinformatics, **4**, 41.
12. *Wu J., Kasif S., DeLisi C.* (2003) Bioinformatics, **19**, 1524-1530.
13. *Kaufman L., Rousseeuw P.J.* (2005) Finding Groups in Data. Wiley-Interscience.
14. *Koonin E.V., Mushegian A.R., Bork P.* (1996) Trends Genet., **12**, 334-336.
15. *Karimpour-Fard A., Leach S.M., Gill R.T., Hunter L.E.* (2008) BMC Bioinformatics, **9**, 397.
16. *Strong M., Graeber T.G., Beeby M., Pellegrini M., Thompson M.J., Yeates T.O., Eisenberg D.* (2003) Nucleic Acids Res., **31**, 7099-7109.
17. *Wu J., Mellor J.C., DeLisi C.* (2005) Genome Inform., **16**, 142-149.
18. *Marcotte E.M., Xenarios I., van Der Bliek A.M., Eisenberg D.* (2000) Proc. Natl. Acad. Sci. USA, **97**, 12115-12120.
19. *Barker D., Pagel M.* (2005) PLoS Comput. Biol., **1**, e3.
20. *Archakov A.I., Govorun V.M., Dubanov A.V., Ivanov Y.D., Veselovsky A.V., Lewi P., Janssen P.* (2003) Proteomics, **3**, 380-391.
21. *Date S.V., Marcotte E.M.* (2003) Nat. Biotechnol., **21**, 1055-1062.
22. *Mirkin B.G, Fenner T.I., Galperin M.Y., Koonin E.V.* (2003) BMC Evol Biol, **3**, 2.
23. *Loganantharaj R., Atwi M.* (2007) BMC Bioinformatics, **8** Suppl 7, S25.
24. *Yamada T., Kanehisa M., Goto S.* (2006) BMC Bioinformatics, **7**, 130.
25. *Kensche P.R., van Noort V., Dutilh B.E., Huynen M.A.* (2008) J. R. Soc. Interface, **5**, 151-170.

Поступила: 09. 02. 2009.

COMPARISON OF ALGORITHMS FOR PREDICTION OF RELATED PROTEINS USING THE METHOD OF PHYLOGENETIC PROFILES

M.A. Pyatnitskiy, A.V. Lisitsa, A.I. Archakov

Institute of Biomedical Chemistry RAMS, Pogodinskaya st.10, Moscow, 119121 Russia;
e-mail: mpyat@bioinformatics.ru

Computational interactomics deals with prediction of functionally related proteins. One approach for solving this problem using comparative genomics consists in analysis of similarities between phylogenetic profiles of proteins. In contrast to most methods, which predict only pairwise interactions between proteins, in the present work we have applied cluster analysis techniques in order to find modules of functionally related proteins. We have performed cluster analysis of phylogenetic profiles of *E. coli* proteins using several clustering techniques and distances between profiles. We report here, that the best correspondence in the composition of resultant clusters to known metabolic pathways is achieved using Ward's clustering together with Hamming's distance. The proposed technique of assessing predictions of the modules of functionally related proteins can be used for comparative analysis of different algorithms for computational interactomics.

Key words: related proteins, cluster analysis, phylogenetic profiles, computational interactomics.