Биомедицинская химия, 2009 том 55, вып. 3, с. 230-246.

БИОИНФОРМАТИКА И МЕТАБОЛОМИКА

УДК 543.51.061:543.54.45:543.8

ПРЕДСКАЗАНИЕ ВЗАИМОСВЯЗАННЫХ БЕЛКОВ МЕТОДАМИ СРАВНИТЕЛЬНОЙ ГЕНОМИКИ *IN SILICO*

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

Обзор посвящен компьютерному поиску взаимосвязанных белков методами сравнительной геномики. Возрастающие возможности биотехнологий по секвенированию новых геномов привели к накоплению данных о последовательностях миллионов генов. Функции большинства этих генов неизвестны, и только для малой части кодируемых ими белков функция может быть определена в эксперименте. Возникает потребность в надежных и достоверных алгоритмах предсказания функций (аннотации) белков методами *in silico*. В обзоре описаны основные подходы сравнительной геномики, использующие как традиционный способ переноса функций на основании гомологии последовательностей, так и с использованием контекстных свойств генов (методы филогенетических профилей и метод генов-соседей). Применение современных методов сравнительной геномики позволяет получить корректные функциональные аннотации для более чем половины всех белков протеома.

Ключевые слова: взаимосвязанные белки, белок-белковые взаимодействия, функциональная аннотация, филогенетические профили, сравнительная геномика

ВВЕДЕНИЕ. Белковые взаимодействия определяют большинство процессов в клетке [1, 2]. Идентификация и изучение сетей взаимосвязанных белков позволяет лучше понять молекулярные механизмы биологических процессов. Функция белка наиболее полно раскрывается в контексте взаимодействия с другими белками: субъединицами, если белок является компонентом молекулярного комплекса, остальными участниками биохимического процесса, если белок участвует в метаболическом пути или внутриклеточной передаче сигнала [3, 4]. Такие взаимосвязи в дальнейшем будут называться структурно-функциональными белок-белковыми взаимодействиями. Изучение этих взаимодействий позволит лучше понять физиологию и патологию клетки, а в конечном итоге и всего организма.

Изучение взаимосвязанных белков сейчас особенно актуально благодаря успехам крупномасштабных проектов по получению полных последовательностей геномов различных организмов. Благодаря прогрессу в области секвенирования ДНК, который революционизировал современную биологию, данные о первичной структуре большинства белков получают путем *in silico* трансляции соответствующих генов вместо прямого определения последовательности аминокислот, например, методом Эдмана.

Однако само по себе знание первичной структуры белка относительно бесполезно, оно приобретает смысл при добавлении биологических фактов в процессе аннотации последовательности. Словарь Webster определяет аннотацию как “заметку, добавленную при комментировании или объяснении”. Базы данных

* - адресат для переписки

по биологическим последовательностям аннотации обычно содержат информацию о клеточной роли и механизмах действия генов и их продуктов. Однако, для определения функции белка необходимы трудоемкие экспериментальные исследования. Проведение таких работ является отчасти искусством, в то время как секвенирование геномов – это хорошо отработанная технология. На момент написания обзора получено 734 полных генома: 659 бактерий, 23 эукариот, 52 архей [http://ncbi.nlm.nih.gov/genbank], и это число постоянно растет. Таким образом, возникает нарастающее отставание между получением биологических последовательностей (генов и белков) и определением функции этих последовательностей. Парадоксальность ситуации состоит в том, что объём получаемых данных намного больше того, который можно осмыслить, проверить и использовать в эксперименте.

Обозначившееся отставание возможно преодолеть путем развития методов, которые позволят проводить функциональную аннотацию всех генов в геноме за допустимое время. При этом возникает альтернатива между медленным и надежным аннотированием последовательностей экспертами-биологами и быстрым, но подверженным ошибкам, аннотированием с помощью полностью автоматизированных программных систем.

На протяжении 80-х и 90-х годов биологическое сообщество полагалось на высоко достоверные аннотации белков, которые разрабатывали относительно небольшие группы экспертов в процессе тщательного анализа опубликованных экспериментальных данных. В настоящее время ситуация в корне изменилась. Анализ количества записей в базах данных по биологическим последовательностям показывает, что для большинства последовательностей их аннотации были получены автоматическими методами. По данным системы RefSeq [http://www.ncbi.nlm.nih.gov/RefSeq/] от 01.05.2008 года, всего 3,2% белков из базы данных было обработано экспертом для проверки качества автоматической функциональной аннотации. Близкие результаты дает и база данных SwissProt [http://www.expasy.org/sprot/] – только для 6,6% функциональных аннотаций белков была проведена проверка. С течением времени доля проверенных специалистами аннотаций неуклонно уменьшается. В связи с экспоненциальным ростом потока данных о последовательностях (удвоение примерно каждые 18 месяцев), экспертам все больше приходится полагаться на предсказание функций методами *in silico*. В сложившейся ситуации совершенствование методов крупномасштабной автоматической аннотации биологических последовательностей приобретает все большую значимость.

1. МОЛЕКУЛЯРНАЯ И КОНТЕКСТНАЯ ФУНКЦИИ БЕЛКА.

Трудность аннотации белковых последовательностей усугубляется отсутствием четкого определения понятия “функция белка”. Для молекулярного биолога функцией белка является участие в сети белок-белковых взаимодействий или локализация белка в определенном компартменте. Для биохимика функция определяется метаболическим процессом, в котором участвует белок, или реакцией, которая катализируется ферментом. Физиолог под функцией белка может понимать временные паттерны экспрессии белка или тканевую специфичность. Для фармаколога важно понять биологические процессы затронутые при ингибировании данного белка и перспективность его в качестве потенциальной мишени для лекарства.

Обычно понятие функции белка разделяют на две составляющие: молекулярную (биохимическую) функцию и контекстную функцию (роль белка в клетке) [3, 5]. При этом под молекулярной функцией понимаются традиционные биохимические характеристики и процессы: связывание, активация, ингибирование, катализ и т.д. Контекстная функция представляет собой систему структурно-функциональных белок-белковых взаимодействий, элементом которой является данный белок. Другие аспекты контекстной функции включают в себя указание внутриклеточной локализации белка и условий его экспрессии.

Роль контекста трудно переоценить – белки практически никогда не функционируют в клетке сами по себе, но часто взаимодействуют с множеством партнеров. По оценкам, в среднем каждый белок физически взаимодействует с 2-10 другими белками [6]. При этом количество белков, с которыми указанный белок является функционально взаимосвязанным (например, участвует в общем метаболическом пути), обычно в 2-3 раза больше. Эта взаимосвязанность является важнейшей особенностью организации и регуляции клетки. По этой причине изучение сетей белок-белковых взаимодействий в последнее время привлекает большое внимание исследователей [7-9].

Существует два основных подхода для предсказания функции белков. Первый подход использует тот факт, что белки, имеющие сходные последовательности аминокислот имеют схожую вторичную и третичную структуру, а следовательно, часто имеют и сходную функцию в клетке. Если для белка с неизвестной функцией удается найти другой белок, схожий на уровне последовательности и уже имеющий функциональную аннотацию, то можно предположить что и первый белок выполняет ту же задачу. Такие методы аннотации (“перенос” функции на основе сходства первичной структуры), в основном предназначены для предсказания молекулярной функции белка.

Другой подход для предсказания функций белков использует данные сравнительной геномики, изучая такие свойства генов как встречаемость и относительное расположение в других геномах. Этот класс методов больше ориентирован на предсказание контекстной функции, поскольку позволяет группировать белки, несущие сходную функциональную нагрузку: например, ферменты, участвующие в одном метаболическом пути, или субъединицы, входящие в состав белкового комплекса.

2. ПРЕДСКАЗАНИЕ МОЛЕКУЛЯРНОЙ ФУНКЦИИ БЕЛКА.

Белки называют гомологичными, если кодирующие их гены с большой вероятностью имеют общее эволюционное происхождение [10]. Гомологичные белки у разных видов, которые эволюционировали от общего предка, называют ортологами. Например, геном человека и геном мыши содержат в себе ген, кодирующий α-субъединицу гемоглобина. Такие гены называют ортологами, поскольку оба гена эволюционировали от α-гемоглобинового гена у общего предка человека и мыши. Гомологичные гены в одном геноме, произошедшие путем дупликации одного гена в эволюционном прошлом, называют паралогами. Например, в геноме человека такими генами являются ген кодирующий α-субъединицу гемоглобина и ген кодирующий β-субъединицу. Эти гены образовались в результате дупликации предкового гена и получившиеся копии, дивергируя в процессе эволюции, дали начало генам α- и β-субъединицам гемоглобина.

Гомологичные белки обычно сходны между собой на всех уровнях белковой структуры и, следовательно, с большой вероятностью выполняют близкие функции, по крайней мере, на молекулярном уровне [11]. Предсказание функции белка по гомологии основано на экстраполяции на этот белок экспериментально установленных знаний о функциях его гомологов. Для поиска гомологов в базах данных о генах и белках различных организмов обычно используют программу локального выравнивания последовательностей BLAST [12]. Результатом выравнивания является информация об уровне сходства соответствующих белковых последовательностей.

Например, пусть в секвенированном геноме обнаружен ген, который кодирует белок с неизвестной функцией. Поиск предполагаемых гомологов этого белка в базе данных может найти последовательности с уже установленной функцией. Предполагается, что интересующий исследователя белок с большой вероятностью будет выполнять ту же или близкую роль в клетке, что и его гомологи.

Рисунок 1 иллюстрирует метод предсказания функциональной аннотации белка по гомологии с уже известным белком. В основе метода лежит

представление о том, что первичная структура белка определяет его структуру, а структура в свою очередь определяет функцию белка. Это формулируется как “первый факт анализа биологических последовательностей” и является фундаментом сравнительной геномики: биомолекулы, имеющие значительное сходство на уровне последовательности, обычно имеют и схожие функции и/или структуры [13]. Стоит отметить, что обратное утверждение несправедливо – белки, имеющие одинаковую функцию (изофункциональные), могут быть абсолютно не схожи на уровне первичной последовательности.

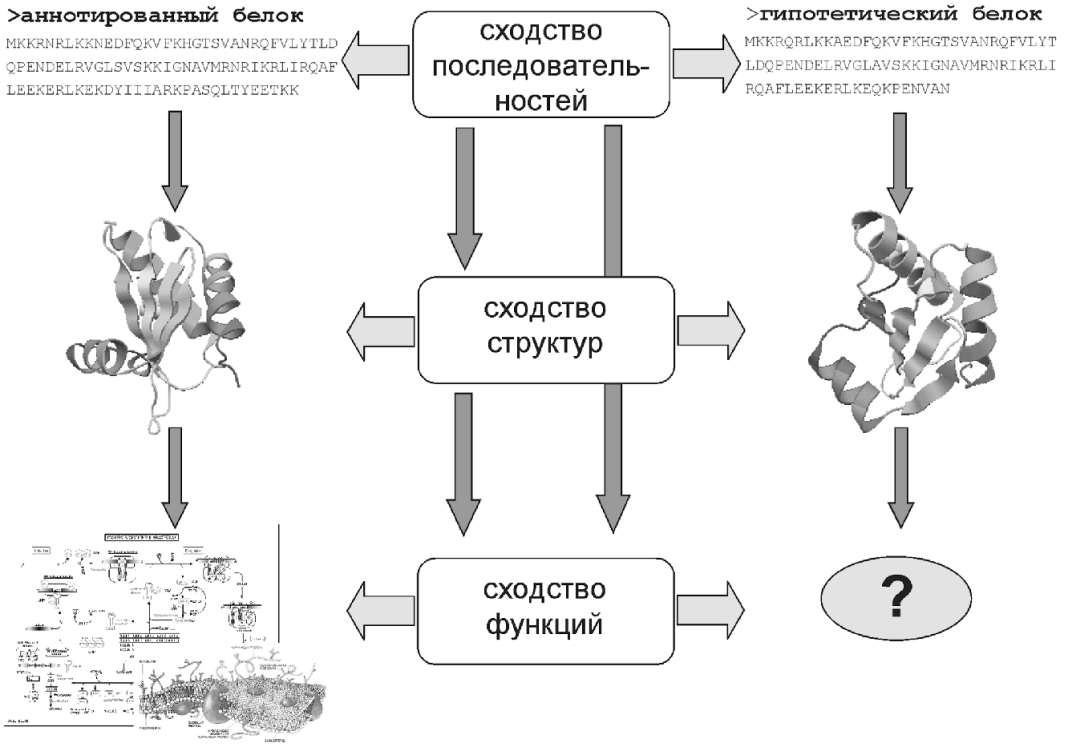

Рисунок 1.
Предсказание функциональной аннотации белка на основании гомологии с известным белком.

Высокая степень сходства первичных структур обычно указывает на существование эволюционной взаимосвязи. В теории всегда требуется сначала установить общность происхождения двух белков, а уже на основе этого делать выводы о сходстве структуры и функции. Однако наличие гомологии между белками практически никогда не может быть доказано абсолютно достоверно, поскольку для этого нужно экспериментально воссоздать ход эволюции и проследить за изменениями соответствующих генов. Поэтому на практике поступают наоборот: об общности происхождения судят на основании сходства последовательностей и структур.

Наличие гомологии между двумя белками является свойством типа “все или ничего”. Часто этот термин используется неправильно: говорят о “60% гомологии”, когда хотят выразить “последовательности схожи на 60%”. Выражение “уровень гомологии” допустимо использовать лишь в том случае, когда требуется выразить степень уверенности в заключении о наличии гомологии. То есть, можно говорить о байесовском подходе к установлению гомологии [14]. При этом отсутствует линейная зависимость между уровнем сходства последовательностей и степенью уверенности исследователя в существовании гомологии. Например, если две последовательности схожи на 80%, то можно быть на 100% уверенным в общности их происхождения, но известны примеры, когда удается достоверно установить гомологию для белков с уровнем сходства близким к случайному шуму.

Так в работе [15] удалось установить гомологию белков, участвующих в регулировании метаболизма азота у прокариот, с уровнем сходства последовательностей достигающим 1-2%. Повышению чувствительности поиска гомологов способствует применение новых методов, основанных на построении профилей белковых семейств: итеративный поиск дальних гомологов (PSI-BLAST [12]) и алгоритмов, использующих скрытые марковские модели (HMMer [16]). Тем не менее, в случае низкого уровня сходства последовательностей процесс установления наличия гомологии является нетривиальным и почти никогда не может быть автоматизирован.

Специфичность взаимодействия между белками определяется структурными и физико-химическими свойствами аминокислотных остатков, формирующих интерфейс межмолекулярного контакта. Поэтому наблюдается определенная степень эволюционной консервативности последовательностей взаимодействующих белков (или доменов). Большинство белков содержат, по крайней мере, один участок, который обладает значительным сходством с последовательностями из филогенетически далеких видов, разошедшихся в процессе эволюции сотни миллионов лет назад. Белки, которые имеют гомологов в филогенетически удаленных видах, составляют 60-80% всего протеома. Исключение из этого общего правила составляют протеомы вирусов и бактериофагов. Для них доля белков, имеющих гомологов в удаленных видах, обычно равна 20-30% [17].

При сравнительном анализе функций белка по первичной структуре важно определить, является ли сходство между последовательностями локальным или глобальным. Локальное сходство между белками наблюдается и в том случае, если они имеют общий домен, однако функция белков может быть различной. Наличие общей функции часто проявляется в виде коротких консервативных участков (мотивов), разделенных неконсервативными областями переменной длины.

Однако, даже если для гена известны его гомологи в других организмах, это не всегда помогает установить его функцию. В каждом геноме достаточна велика доля генов, кодирующих белки, принадлежащие к высоко консервативным семействам, для которых биологическая роль не установлена ни для одного члена семейства. Такие белки аннотируются как консервативные гипотетические белки. Например, в геноме одного из наиболее изученных модельных микроорганизмов *E. coli* такие гены составляют 22,1% (на момент написания обзора, [www.tigr.org]). В настоящее время в базе данных NCBI Protein содержится более 850000 последовательностей с аннотацией “консервативный гипотетический белок” и более 10500000 последовательностей с аннотацией “гипотетический белок”.

Установление гомологии и собственно предсказание функции белка может осложняться различными факторами, в частности, наличием функциональной или структурной конвергенции, при которой белки обладают схожей структурой/функцией, но не имеют общего предка. Явление, когда два изофункциональных белка не являются ортологами, называется неортологичным замещением гена [18]. По меньшей мере 10% всех групп в системе Enzyme Classification (EC) содержат изофункциональные, но не ортологичные белки [19].

Другим фактором надежной функциональной аннотации на основании гомологии является правильное разделение гомологов белка на ортологов и паралогов. Для ортологов сохранение первоначальной функции более вероятно по сравнению с паралогами, поскольку благодаря дупликации генов последние имеют тенденцию приобретать дополнительные функции. Тем самым использование только ортологов для функциональной аннотации будет давать более надежные предсказания. Разработаны алгоритмы и базы данных, хранящие информацию о наборах ортологов (COG http://ncbi.nih.gov/COG, InParanoid http://inparanoid.sbc.su.se). Наиболее надежным способом разделения гомологов на ортологи и паралоги является построение филогенетических деревьев, однако это требует существенных вычислительных затрат и не всегда соответствующий анализ может быть надежно автоматизирован.

Наконец, различные характерные особенности анализируемых последовательностей могут негативно влиять на качество работы алгоритмов выравнивания. Например, наличие мультидоменной структуры белков и обусловленное этим локальное сходство последовательностей осложняет принятие решения о наличии гомологии, поскольку существование общего домена между белками еще не гарантирует их гомологию. Кроме того, могут возникнуть затруднения при выравнивании участков последовательностей с необычным аминокислотным составом или повторами. Такие области должны отфильтровываться, поскольку их наличие также приводит к снижению достоверности результатов работы алгоритма установления гомологии.

Несмотря на указанные недостатки, метод переноса функциональных аннотаций на основании гомологии белков является основным средством для предсказания функциональной роли белков в клетке. В настоящее время использование технологий сравнительной геномики *in silico* в сочетании с данными о структуре и организации генома дает возможность предсказывать достаточно надежно функции примерно половины белков закодированных во вновь секвенированном геноме [20]. Несомненно, эта доля будет расти в будущем по мере накопления новых экспериментальных данных.

3. ПРЕДСКАЗАНИЕ КОНТЕКСТНОЙ ФУНКЦИИ БЕЛКА.

Другим аспектом функции белка является участие в сложной сети белок-белковых взаимодействий. Тем самым изучение партнеров по взаимодействию является важным шагом к пониманию роли белка в клетке. Предсказание структурно-функциональных взаимодействий между белками методами *in silico* стало возможным благодаря применению алгоритмов, которые опираются не только на гомологию последовательностей.

В то время как присваивание функции генов на основе гомологии использует перенос аннотации гена с известной функцией на его гомолог, методы предсказания контекстной функции белка основаны на принципе “виновен по ассоциации” (guilt by association) [21]. Согласно этому принципу, если ген А участвует в функции Х и есть данные в пользу того, что ген В ассоциирован с А, то ген В тоже участвует в функции Х. Для оценки ассоциации между генами используются такие свойства как характер распределения гомологов в других геномах [22], положение и относительный порядок следования генов на хромосоме [23], частота слияний генов [24], паттерны экспрессии гена [25]. Если неохарактеризованный белок может быть включен в кластер ассоциированных между собой белков, где установлена функция для одного или нескольких белков, то тем самым может быть установлена функциональная связь и сделано предположение о функции этого белка.

Важным косвенным фактором, лежащим в основе таких ассоциативных подходов, является использование предположения о давлении естественного отбора как движущей силы эволюции генома и протеома. Взаимосвязанные белки с большой вероятностью будут иметь близкие эволюционные истории, поскольку сохранение белок-белковых взаимодействий и биохимических функций требует координации эволюционных изменений (метод филогенетических профилей). Также в случае если порядок следования генов или слияние генов будут давать селективное преимущество в виде улучшенного функционального взаимодействия между белками, то такие геномные перестройки будут сохраняться в ходе эволюции (метод генных кластеров и метод розеттского камня). Общий обзор методов предсказания контекстной функции белка приведен в работах [26-28].

Методы изучения коэволюции или кластеризации генов часто называют негомологичными методами для предсказания функций белков. Этот термин трудно назвать удачным, поскольку указанные методы все равно опираются на определение набора гомологов (в идеале – ортологов). Только после надежного определения всех гомологов можно ожидать, что анализ их совместной встречаемости или относительного расположения в геномах даст полезную

информацию для предсказания функций белков. Возможно, более правильным будет называть такие методы “пост-гомологичными” [14].

3.1. Анализ коэволюционирующих белков: филогенетические профили.

Тот факт, что два гена встречаются совместно в одном геноме, дает мало информации об их возможной функциональной взаимосвязи. Если же эти гены совместно встречаются и отсутствуют в большом числе геномов, то подобное трудно счесть случайным совпадением. Сравнительный анализ совместных распределений генов в ряде геномов называется методом филогенетических профилей [22, 29]. Метод основан на предположении, что функционально связанные гены также связаны и эволюционно, то есть в ходе эволюции такие гены будут либо совместно унаследованы вновь образованным видом, либо будут элиминированы. Таким образом, организм находится под давлением естественного отбора: одновременное присутствие или отсутствие обоих взаимосвязанных генов будет способствовать приспособленности, а наличие в геноме лишь одного из генов - наоборот, ухудшать приспособленность. Например, гены, кодирующие белки жгутика, есть только в геноме бактерий имеющих жгутик, но не в геномах других безжгутиковых организмов [30].

Исходными данными является большое число последовательностей геномов организмов (т.н. референтные геномы, рис. 2а). Каждый белок изучаемого организма (в данном случае – *E. coli*) характеризуется своим филогенетическим профилем. Филогенетический профиль (ФП) – это бинарный вектор, компоненты которого показывают, присутствует ли гомолог гена, кодирующий данный белок, в каждом из референтных геномов (рис. 2б). Если такой гомолог присутствует, то в позиции соответствующей данному референтному организму ставится 1, в противном случае – 0. Каждая пара взаимодействующих белков (составляющих структурный комплекс или участвующих в одном метаболическом пути) будет совместно присутствовать в одних геномах, и отсутствовать в других, т.е. иметь схожие профили. Проведение кластерного анализа ФП (рис. 2в) даёт группы функционально взаимосвязанных белков. В рассмотренном примере белки P2 и P7, а также белки P3 и P6 предсказаны как функционально взаимосвязанные, т.к. имеют идентичные ФП. Так для белков P2 и P7 из *E. coli* есть гомологи у *S. cerevisiase* и *B. burgdorferi*, но при этом нет гомологов у *H. pylori* ни для одного, ни для другого белка.

Результаты метода ФП сильно зависят от набора используемых референтных геномов. В пионерской работе [22] использовалось 16 геномов, причём прогнозировалось, что с увеличением числа референтных геномов и появлением данных по эукариотам, информативность и точность метода возрастут. Это оправдалось лишь частично. В работе [31] показано, что при использовании более 86 референтных геномов точность метода практически перестаёт повышаться и достигает насыщения при 145 организмах. Включение в набор геномов нескольких штаммов одного вида ухудшает точность предсказаний. Также результаты ухудшаются при большом числе включенных геномов эукариот. В работе [32], где этот вопрос исследован наиболее подробно, рекомендуется использовать неизбыточные наборы референтных геномов представителей всех 3 надцарств (археи, бактерии и эукариоты), исключая паразитические организмы (поскольку для них характерна интенсивная редукция геномов) и близкородственные штаммы. В последнем случае наличие гомолога в одном виде не является независимым событием относительно его присутствия в другом близкородственном виде, поскольку геномы не успели сильно измениться в процессе эволюции. Тем самым, схожесть ФП будет указывать не на имеющуюся функциональную взаимосвязь между белками, а отражать факт эволюционной близости двух геномов. В простейшем варианте метода филогения референтных геномов никак не учитывается, однако в последнее время разработаны алгоритмы, позволяющие использовать имеющееся филогенетическое дерево при предсказании функции белка [33]. К сожалению, такие методы требуют значительных вычислительных ресурсов [34].

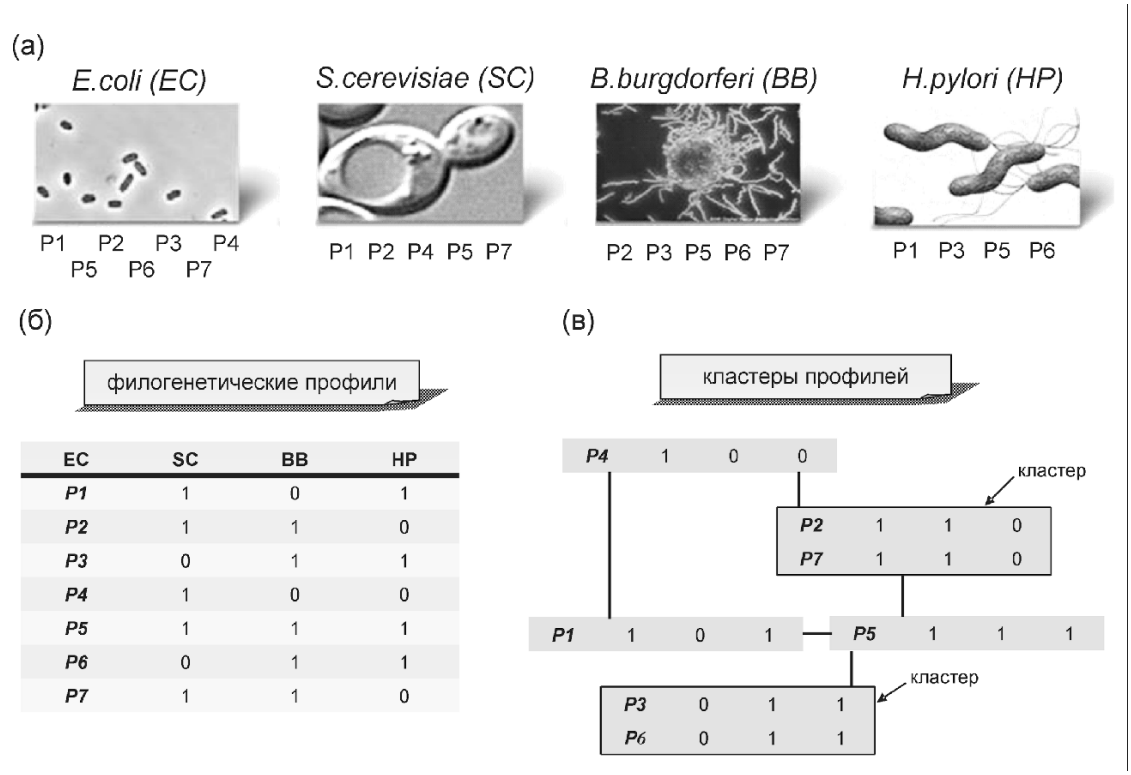

Рисунок 2.
Метод филогенетических профилей. (а) Набор референтных организмов. В геноме каждого референтного организма может быть закодирован гомолог соответствующего белка P1-P7 из изучаемого организма (*E. coli*). (б) Филогенетические профили белков *E. coli*. Для каждого белка строится филогенетический профиль, представляющий собой двоичный вектор, где наличие/отсутствие гомологов данного белка закодированных в каждом референтном геноме обозначается 1 или 0 соответственно. (в) Кластеризация филогенетических профилей позволяет выявить группы функционально взаимосвязанных белков (белки P2 и P7, а также белки P3 и P6).

В исходном варианте метода ФП гипотеза о функциональной связи между двумя белками выдвигалась на основе сходства ФП, соответствующим этим белкам. Столь простых соотношений можно ожидать только в тех случаях, когда рассматривается структурный комплекс, в который входят оба белка, или если оба белка интегрированы в состав неветвящегося метаболического пути [6]. Если же учитывать ветвление, наличие параллельных и альтернативных путей, а также функциональную конвергенцию и горизонтальный перенос генов, то можно ожидать более сложных логических отношений, чем простое попарное сходство между ФП. В работе [35] рассматриваются всевозможные логические отношения в триадах ФП, основанные на конъюнкции (оператор “И”) и дизъюнкции (оператор “ИЛИ”). Такой подход представляется весьма перспективным, поскольку он позволяет обнаруживать связи между белками, которые принципиально не выявляются попарным сравнением ФП.

Типичным примером успешного применения метода ФП является исследование гена *FRDA*, кодирующего белок фратаксин. Хотя и было известно, что мутация в этом гене вызывает нейродегенеративное заболевание – атаксию Фридрейха, однако молекулярная функция белка оставалась неизвестной [36]. В работе [37] было показано, что ФП фратаксина схож с профилями для нескольких белков, участвующих в сборке железосерного кластера ферредоксина.

Это наблюдение позволило предположить, что и фратаксин участвует в этом процессе. Экспериментальное подтверждение предсказания было получено годом позднее в работе [38].

В работе [34] приведена таблица экспериментальных подтверждений предсказаний взаимосвязанных белков, сделанных на основе метода ФП. В целом, метод позволяет делать высокоспецифичные предсказания при относительно малой чувствительности, т.е. преобладают ложноположительные результаты. Несмотря на то, что для прокариот наблюдается сильная функциональная взаимосвязь для генов, имеющих схожие ФП [39], применение метода для геномов эукариот вызывает определенные трудности [40]. С одной стороны, это можно объяснить малым числом секвенированных геномов эукариот по сравнению с таковыми прокариот. С другой стороны, возможной причиной является несостоятельность для описания сложной организации геномов эукариот простых допущений, использующихся в методе ФП.

В то же время подходы, отработанные на методе ФП, могут быть применены и в других областях, поскольку сама идея изучения совместной встречаемости взаимосвязанных сущностей является достаточно общей. Любые признаки, которые могут быть представлены в виде последовательности двоичных векторов, могут быть изучены с помощью подобных алгоритмов. Такими признаками могут быть различные фенотипические проявления. Так, например, в работе [41] рассмотрены взаимодействия между белковыми доменами, где в качестве элементов ФП использовалось наличие/отсутствие определенного домена в белке. В статье [42] применен аналогичный подход для анализа регуляции генов рибонуклеотидредуктазы, с использованием распределения консервативного сигнала NrdR-box в бактериальных геномах. Корреляция распределения сигнальных последовательностей в белке с ФП белков различных семейств успешно применялась для определения специфичности транспортных систем белков в цитоплазме [43].

Изучение ассоциации генов и фенотипов позволяет предсказать участие продуктов этих генов в различных биологических процессах. Это не требует предварительных знаний о других генах, участвующих в этом процессе. Связь между генотипом и фенотипом хорошо иллюстрирует работа [44], в которой изучались белки, ответственные за биогенез и функционирование базального тела – модифицированной центриоли, участвующей в работе жгутиков и ресничек. Для анализа были выбраны геномы *Arabidopsis thaliana*, жгутиковой водоросли *Chlamydomonas reinhardtii* и человека. Были отобраны 688 генов, которые присутствуют только в организмах имеющих жгутик, т.е. у хламидомонады и у человека. Обнаружилось, что для 12% отобранных генов их причастность к формированию жгутика ранее была неизвестна и в этом списке присутствует ген человека *BBS5*, ранее ассоциированный с тяжелым наследственным заболеванием, синдромом Барда-Бидля. Далее в эксперименте было подтверждено, что продукт гена-гомолога *BBS5* действительно локализован в базальном теле клеток *Caenorhabditis elegans*. Тем самым была установлена взаимосвязь между нарушениями формирования жгутика и многочисленными проявлениями синдрома Барда-Бидля - дистрофией сетчатки, ожирением, полидактилией и т.д. Таким образом, анализ взаимосвязи между фенотипом и генотипом может способствовать установлению функций генов и пониманию молекулярных механизмов болезней. В других случаях ассоциативный анализ был применен для определения генов связанных с патогенностью [45] и гипертермофилией [46].

Иногда в ходе эволюции ген может быть замещен другим геном с такой же функцией, причем первичная последовательность нового гена может сильно отличаться от первоначального варианта. Такой феномен называется неортологичным замещением генов [18]. Поскольку давление естественного отбора обычно элиминирует копии генов с одинаковой функцией, то ФП нового гена будет комплементарен ФП исходного гена. Направленный поиск генов с взаимно

комплементарными ФП может использоваться для функциональной аннотации, т.к. функции генов остаются неизменными. Так, в работе [47] для гипотетического белка была предсказано участие в биосинтеза тиамина на основании комплементарности его ФП и профилей других белков с известной биохимической функцией. Предсказание удалось подтвердить в эксперименте.

Важным преимуществом метода является относительная простота реализации и возможность масштабного применения. Так, в работе [29] ФП были использованы для предсказания всех ко-локализованных белков в клетке. Ещё одним использованием метода является исправление ошибок, возникающих при разметке вновь секвенированных геномов. В процессе автоматического поиска генов иногда возможна ошибка алгоритма, при которой границы гена устанавливаются неверно. В таких случаях аннотация такого "гена" будет противоречить уже известным установленным функциональным или эволюционным соотношениям между группами других генов. В работе [48] ФП были успешно использованы для идентификации 22 белков прокариот при аннотации которых были допущены ошибки в результате неверного определения границ соответствующих генов.

3.2. Анализ генных кластеров и метод "розеттского камня".

Знание последовательностей геномов микроорганизмов позволяет проводить сравнительный анализ положения генов на хромосоме. Информацию о расстоянии между генами и относительном порядке следования генов можно использовать для предсказания функционально взаимосвязанных белков.

Результаты работ по сравнительной геномике микроорганизмов [49] показывают, что в масштабе всего генома порядок следования генов не консервативен и может сильно различаться даже для относительно близкородственных видов, таких как *E. coli* и *H. influenzae* [50, 51]. Однако, для близко расположенных генов вероятность сохранения их относительного расположения в геноме в процессе эволюции значительно повышается [52]. Это позволяет предположить, что причиной является воздействие стабилизирующего отбора, формы естественного отбора. Давление стабилизирующего отбора происходит на уровне групп соседних генов – генных кластеров, не нарушая относительную последовательность генов внутри кластера. При этом положение самих кластеров в геноме можно рассматривать как случайное. Выясняется, что воздействие естественного отбора сохраняет порядок следования для генов, имеющих общую функцию. Такой отбор на относительное расположение в геноме, вероятно, связан с необходимостью синхронизировать экспрессию этих генов для осуществления совместной функции. Тем самым, изучение группирования и порядка следования генов может дать информацию о функциональных взаимосвязях генов.

Возможно несколько типов группирования генов на хромосоме. В крайнем варианте это – трансляционное слияние генов (рис. 3а). В процессе эволюции могут возникать ситуации, когда два гена присутствующие раздельно в одном организме могут сливаться в другом организме, т.е. экспрессироваться как один мультидоменный белок. Такое слияние, вероятно, происходит для оптимизации коэкспрессии генов, кодирующих взаимодействующие белки. Поэтому поиск слившихся генов может способствовать установлению функциональных взаимодействий между белками.

Другим типом слияния генов, характерным для прокариот, является транскрипционное слияние (рис. 3б), когда одна молекула РНК включает в себя несколько открытых рамок считывания, которые являются ко-регулируемыми (имеют один общий промотор) и кодируют несколько различных белков. Такая единица называется опероном. Гены в составе оперона транскрибируются в одном направлении и кодируют ферменты, которые относятся к последовательным этапам какого-либо биохимического процесса, т.е. являются функционально взаимосвязанными.

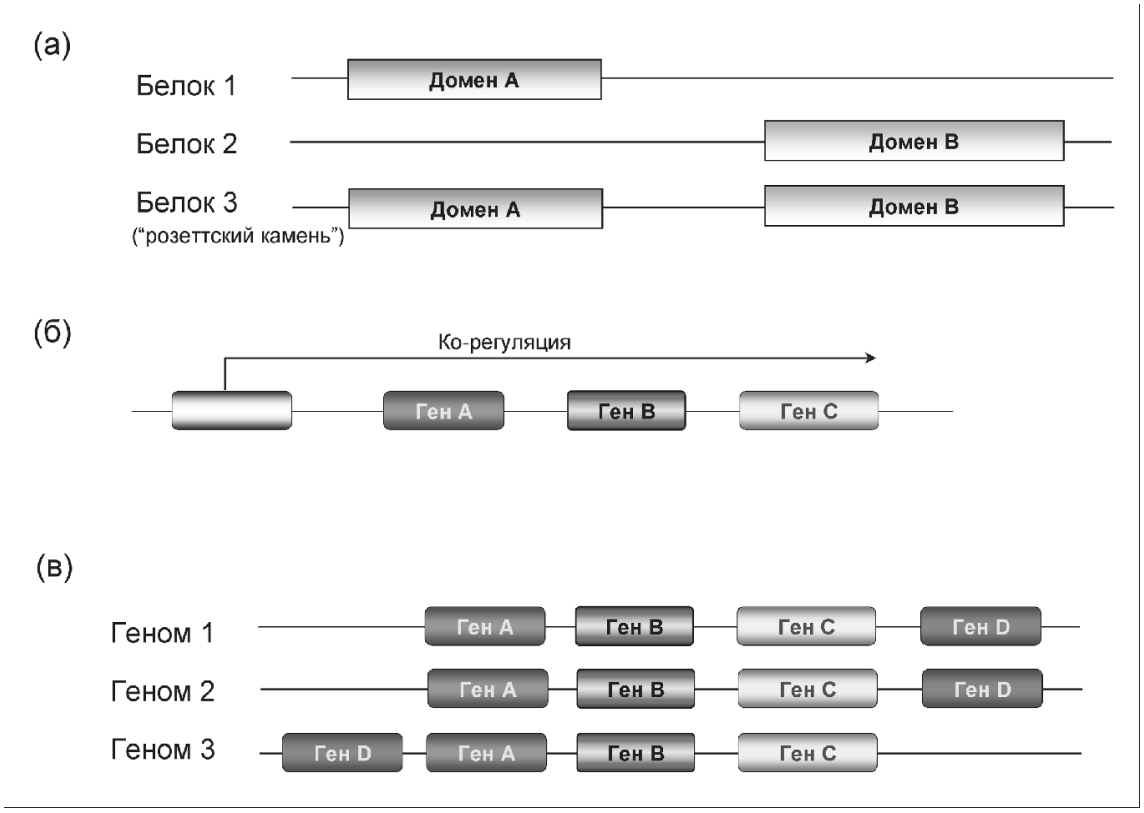

Рисунок 3.
Различные типы слияния генов и их использование для предсказания функциональной взаимосвязи между соответствующими белками. (а) Трансляционное слияние генов - функциональная взаимосвязь между двумя белками устанавливается на основе третьего "розеттского" белка, в котором встречаются участки (домены) схожие с каждым из белков в отдельности. (б) Транскрипционное слияние – оперонная организация генома (для прокариот). Гены входящие в состав оперона считаются взаимосвязанными. (в) Гены-соседи – имеют тенденцию близко располагаться друг относительно друга в нескольких геномах.

В третьем случае группировка генов представляет собой только тенденцию некоторых генов (т.н. генов-соседей) располагаться рядом на хромосоме в ряду геномов. Вообще провести четкую эволюционную границу между различными вариантами слияний генов достаточно трудно. Рассмотрим каждый из типов расположения генов более подробно.

При анализе трансляционного слияния генов основное предположение состоит в том, что слияние будет сохраняться, если улучшение функционального взаимодействия между слившимися белками дает организму эволюционное преимущество. Тем самым, найденные слившиеся белки или домены в одном организме могут указать на их взаимосвязь в других видах, где соответствующие белки экспрессируются раздельно. Трансляционное слияние генов является сильным свидетельством в пользу того, что соответствующие белки взаимодействуют физически или синтезируются в эквимолярных концентрациях.

Изучение слияния/разделения генов называется методом "розеттского камня" [53]. Можно сказать, что в этом подходе "язык" белков слияния переводится на "язык" функциональных взаимодействий. Напомним, что розеттский камень представляет собой плиту с идентичной надписью на 3 языках: греческом, египетском демотическом письме и древнеегипетскими иероглифами. Французский исследователь Ж.-Ф. Шампольон, сопоставляя три текста, успешно использовал розеттский камень для расшифровки египетских иероглифов.

В работе [6] для протеома *E. coli* методом "розеттского камня" было обнаружено 6809 предположительно взаимодействующих пар негомологичных белков. Проверка результатов работы метода показала, что для большей части таких пар оба белка действительно являлись функционально связанными. Сравнение с результатами из базы данных DIP показало, что примерно 6,4% экспериментально подтвержденных белок-белковых взаимодействий возможно объяснить с помощью метода "розеттского камня". Иллюстрацией предсказаний взаимосвязанных белков может служить пример с ДНК-гиразами А и В, которые присутствуют раздельно в *E.coli*, но при этом гомологичны различным участкам топоизомеразы II из *S.cerevisiae*. В этом случае топоизомераза II является белком слияния (или т.н. "розеттским" белком). Сами же ДНК-гиразы А и В действительно являются взаимодействующими в *E. coli*.

Не все события слияния имеют одинаковую ценность для установления функциональных взаимосвязей между генами. Например, некоторые домены, такие как АТР-связывающие кассеты и SH3-домены, участвуют в слияниях более чем с сотней других доменов [6]. Поэтому было бы некорректным говорить о функциональной взаимосвязи каждого белка с SH3-доменом с каждым белком с киназным доменом только на основании наличия белков слияния с обоими доменами.

В случае транскрипционного слияния продукты оперонов (генных кластеров) с большой вероятностью взаимодействуют между собой, т.е. вовлечены в один и тот же метаболический путь или являются компонентами одной функциональной системы. Следовательно, поиск и определение оперонной структуры геномов микроорганизмов являются также и способом определения структурно-функционально взаимосвязанных белков.

Разработаны различные методы для идентификации оперонной структуры в геномах микроорганизмов. Для предсказания границ оперонов используется разница в расстояниях между генами принадлежащими одному оперону (несколько нуклеотидов) и генами на границах соседних оперонов (в среднем 190 нуклеотидов) [54]. В работе [55] при анализе оперонной организации генома *E. coli* были корректно предсказаны 85% от всех известных оперонов. В другой статье [56] по результатам работы метода на 34 геномах были определены более 7600 пар генов, для которых вероятность их вхождения в один оперон составляла не менее 98%. Важным достоинством метода оперонов является отсутствие необходимости определения гомологии между генами, что позволяет предсказать функциональную взаимосвязь белков, для которых определение факта гомологии может быть нетривиальным.

Наконец, для предсказания функциональных взаимосвязей могут быть использованы данные об относительном расположении на хромосоме гомологичных генов. Сколько информации о функциональных взаимосвязях между генами содержится в определенном порядке следования генов? Теоретически N генов можно упорядочить $N!$ способами, что является астрономическим числом, учитывая среднее число генов в геноме (несколько тысяч). Если же в ряде геномов сохраняется расположение нескольких генов относительно друг друга, то такое наблюдение явно является неслучайным.

При изучении одного генома, такие гены могут быть выявлены при поиске областей, где межгенные расстояния меньше чем средняя величина по геному. При наличии последовательностей многих родственных геномов возможно выявление групп генов, которые расположены близко друг к другу в нескольких геномах одновременно (рис. 3в). В данном случае методическая сложность заключается в определении критерия близости генов. Например, в работе [23] два гена считаются соседними в том случае, если они оба расположены на одной цепи ДНК и межгенное расстояние составляет не более 600 нуклеотидов.

Гены внутри такого кластера могут быть функционально связанными (хотя обычно это менее вероятно чем в первых двух случаях слияний генов), например, иметь сходные уровни экспрессии [57] или схожую тканевую специфичность [58].

Анализ паттерна сохранения порядка генов среди трех геномов бактерий и архей показал, что 63-75% ко-регулируемых генов являются физически взаимодействующими [59]. В случае эукариот функционально взаимосвязанные гены иногда также могут сохранять положение друг относительно друга. Примером являются гены *hox*, которые играют важную роль в эмбриональном развитии [60].

Принципиальная трудность всех методов основанных на изучении относительного порядка генов такая же, что и в методе ФП: генные кластеры в двух близкородственных видах могут отражать всего лишь расположение генов у их общего предка, а вовсе не результат действия сил отбора для сохранения функциональной взаимосвязи между генами. Также кластеры могут образовываться и в результате хромосомных транслокаций. Уже отмечалось, что на больших эволюционных расстояниях сохранение порядка генов наблюдается редко, что затрудняет применение метода.

4. Методы верификации результатов.

Как и в любом вычислительном эксперименте, при предсказании взаимосвязанных генов/белков *in silico* важную роль играют методы проверки и верификации полученных прогнозов. Однако здесь возникает много проблем, связанных с отсутствием общепринятой унифицированной методики сравнительного анализа качества предсказаний. Это не позволяет напрямую сравнивать между собой работы, в которых предлагаются различные модификации методов поиска взаимосвязанных белков. В первую очередь, отсутствует общепризнанный “золотой стандарт”, т.е. набор последовательностей с известными верными аннотациями, на котором должен тестироваться каждый новый алгоритм. В литературе можно встретить три основных способа проверки методов предсказаний функций белков.

Первым подходом является сравнение с экспериментальными данными, полученных с применением методик, направленных на выявление физических белок-белковых взаимодействий: дрожжевой дигибридный скрининг, белковые микрочипы, иммунопреципитация, масс-спектрометрия и т.д. Результаты таких исследований заносятся в специализированные базы данных [61]. Из наиболее известных следует упомянуть базы MINT http://cbm.bio.uniroma2.it/mint/, MIPS http://mips.gsf.de/proj/yeast/CYGD/interaction/, IntAct http://www.ebi.ac.uk/intact, DIP http://dip.doe-mbi.ucla.edu/. Так, например, в работе [62] предлагаемое улучшение метода ФП путем учета филогении референтных геномов проверяется на данных из MIPS. Для набора из 10551 пар подтвержденных экспериментально взаимодействий между белками удалось на 35% улучшить результаты предсказания методом ФП.

К сожалению, в большинстве экспериментов используются лишь несколько модельных организмов, что затрудняет проверку предсказаний для протеомов других организмов. Например, база данных DIP содержит информацию о белок-белковых взаимодействиях для 110 организмов, но 96% взаимодействий получено всего на восьми организмах. Важным принципиальным недостатком такого способа проверки результатов предсказаний является высокий процент ложноположительных взаимодействий, получаемых в эксперименте [63] и плохое согласование между собой данных, полученных с помощью различных экспериментальных методик [7, 64].

Другим подходом является сравнение результатов предсказания с ранее накопленными биохимическими знаниями и иерархиями функциональных категорий. В работах [31, 32, 65] результаты предсказаний сравниваются с базой данных KEGG [66], которая содержит данные о метаболических путях. Два белка считаются взаимосвязанными в том случае, если они совместно встречаются хотя бы в одном метаболическом пути. Этот критерий был использован, например, для оценки качества работы метода “розеттского камня” в работе [24]. Согласно использованной в этой работе схеме оценки результатов, предсказания

взаимосвязанных белков *E. coli* могут быть произведены с точностью достигающей 70%. Другим источником для сравнений является иерархия терминов Gene Ontology [67], описывающая структуру и системы клетки. Здесь два белка считаются взаимосвязанными, если их аннотации в этой системе совпадают до некоторого уровня. В работе [68] было предсказано, что примерно для 30% белков дрожжей можно найти в протеоме человека функционально близкие белки. В статье [40] после проведения кластерного анализа ФП белков человека было рассчитано что ~25% кластеров сильно перекрываются по составу с соответствующими группами в Gene Ontology.

Наконец, в работах [69, 70] в качестве меры сходства функций было предложено сравнивать наборы ключевых слов описывающих белок в базе данных SwissProt. Например, при изменении порога расстояний между генами в методе оперона от 0 до 100 нуклеотидов процент общих ключевых слов в аннотации уменьшается с 51% до 45% [70].

Обсуждая возможные ошибки методов для автоматической функциональной аннотации генов и белков, необходимо подчеркнуть неравнозначность двух возможных типов ошибок. Ложнонегативные результаты, то есть пропуски алгоритмом функциональных связей там, где они присутствуют, намного менее опасны, чем ложноположительные результаты (перепредсказание). В последнем случае возможно дальнейшее распространение ошибочной аннотации по базе данных, особенно при переносе функций на основании гомологии. В последнее время, по аналогии с общедоступной энциклопедией Wikipedia, куда каждый может вносить исправления, предлагается использовать подобный подход для того, чтобы привлечь все мировое сообщество биологов-экспертов к проверке существующих и генерации новых функциональных аннотаций.

ЗАКЛЮЧЕНИЕ. Задача определения функций белка или гена на основе первичной последовательности или структуры – одна из важнейших задач постгеномной биоинформатики. В дополнение к традиционному подходу переноса функции на основании гомологии в течение последних лет добавились методы, использующие контекстные свойства генов – распределение гомологов в ряду организмов, относительное расположение генов на хромосоме и т.д. Эти методы активно опираются на достижения постгеномной эры современной биологии, используя информацию о последовательностях большого числа полностью секвенированных геномах. Анализ контекстных свойств гена показывает, что геном и набор геномов являются особым типом данных, которые нельзя сводить к простой совокупности последовательностей генов [20].

Говоря о различных подходах к предсказанию функций белков и генов важно упомянуть интерактивные веб-ориентированные базы данных, которые интегрируют как результаты вычислительной сравнительной геномики, так и результаты крупномасштабных экспериментов по обнаружению физически взаимодействующих белков. Такие системы предоставляют удобный унифицированный интерфейс для доступа ко всему арсеналу доступных методов. Примерами могут служить Prolinks [71], PREDICTOME [72], PLEX [73]. Наиболее известна система STRING [74], которая позволяет предсказывать функции белка методом ФП, методами “розеттского камня” и генов-соседей, проводить анализ транскриптомных данных по ко-экспрессии и осуществлять поиск белковых взаимосвязей, описанных в текстах публикаций. Также доступны данные по экспериментально установленным белок-белковым взаимодействиям. Результаты применения различных методов наглядно визуализируются в виде сети, узлами которой являются белки, а рёбра указывают предсказанные взаимосвязи.

За последние 10 лет область вычислительного предсказания функции генов и белков активно развивается [26, 33, 75]. Важнейшим стимулом для развития является выполнение высокопроизводительных экспериментов, в первую очередь проектов по секвенированию полных геномов. Если первые геномы (*Haemophilus influenzae, Mycoplasma pneumonia*) аннотировались в основном экспертами

[76, 77], то в настоящее время первичная разметка вновь секвенированных геномов и предсказание функций генов осуществляется автоматически. При этом неизбежно возникает необходимость в автоматической же проверке и уточнении сделанных аннотаций [78]. Поэтому можно утверждать, что работы в области определения взаимосвязанных генов и белков методами *in silico* будут крайне востребованы в обозримом будущем.

ЛИТЕРАТУРА

1. *Gavin A.C., Aloy P., Grandi P., Krause R., Boesche M., Marzioch M., Rau C., Jensen L.J., Bastuck S. et al.* (2006) Nature, **440**, 631-636.
2. *Li S., Armstrong C.M., Bertin N., Ge H., Milstein S., Boxem M., Vidalain P.O., Han J.D., Chesneau A. et al.* (2004) Science, **303**, 540-543.
3. *Gabaldon T., Huynen M.A.* (2004) Cell. Mol. Life Sci., **61**, 930-944.
4. *Huynen M.A., Snel B., von Mering C., Bork P.* (2003) Curr. Opin. Cell. Biol., **15**, 191-198.
5. *Eisenberg D., Marcotte E.M., Xenarios I., Yeates T.O.* (2000) Nature, **405**, 823-826.
6. *Marcotte E.M., Pellegrini M., Ng H.L., Rice D.W., Yeates T.O., Eisenberg D.* (1999) Science, **285**, 751-753.
7. *Ravasz E., Somera A.L., Mongru D.A., Oltvai Z.N., Barabasi A.L.* (2002) Science, **297**, 1551-1555.
8. *Wu J., Mellor J.C., DeLisi C.* (2005) Genome Inform., **16**, 142-149.
9. *Yanai I., DeLisi C.* (2002) Genome Biol., **3**, research0064.
10. *Fitch W.M.* (1970) Syst. Zool., **19**, 99-113.
11. *Teichmann S.A.* (2002) J. Mol. Biol., **324**, 399-407.
12. *Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J.* (1997) Nucleic Acids Res., **25**, 3389-3402.
13. *Gusfield D.* (1997) Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, Cambridge, UK.
14. *Mushegian A.R.* (2007) Foundations of Comparative Genomics. Academic Press.
15. *Kinch L.N., Grishin N.V.* (2002) Proteins, **48**, 75-84.
16. *Eddy S.R.* (1998) Bioinformatics, **14**, 755-763.
17. *Liu J., Glazko G., Mushegian A.* (2006) Virus Res., **117**, 68-80.
18. *Koonin E.V., Mushegian A.R., Bork P.* (1996) Trends Genet., **12**, 334-336.
19. *Galperin M.Y., Koonin E.V.* (1999) Genetica, **106**, 159-170.
20. *Koonin E.V., Galperin M.Y.* (2003) Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Kluwer Academic, Boston.
21. *Aravind L.* (2000) Genome Res., **10**, 1074-1077.
22. *Pellegrini M., Marcotte E.M., Thompson M.J., Eisenberg D., Yeates T.O.* (1999) Proc. Natl. Acad. Sci. USA, **96**, 4285-4288.
23. *Overbeek R., Fonstein M., D'Souza M., Pusch G.D., Maltsev N.* (1999) Proc. Natl. Acad. Sci. USA, **96**, 2896-2901.
24. *Marcotte C.J., Marcotte E.M.* (2002) Appl. Bioinformatics, **1**, 93-100.
25. *van Noort V., Snel B., Huynen M.A.* (2003) Trends Genet., **19**, 238-242.
26. *Salwinski L., Eisenberg D.* (2003) Curr. Opin. Struct. Biol., **13**, 377-382.
27. *Shoemaker B.A., Panchenko A.R.* (2007) PLoS Comput. Biol., **3**, e43.
28. *Valencia A., Pazos F.* (2002) Curr. Opin. Struct. Biol., **12**, 368-373.
29. *Marcotte E.M., Xenarios I., van Der Bliek A.M., Eisenberg D.* (2000) Proc. Natl. Acad. Sci. USA, **97**, 12115-12120.
30. *Jim K., Parmar K., Singh M., Tavazoie S.* (2004) Genome Res., **14**, 109-115.
31. *Sun J., Xu J., Liu Z., Liu Q., Zhao A., Shi T., Li Y.* (2005) Bioinformatics, **21**, 3409-3415.
32. *Jothi R., Przytycka T.M., Aravind L.* (2007) BMC Bioinformatics, **8**, 173.

33. *Barker D., Meade A., Pagel M.* (2007) Bioinformatics, **23**, 14-20.
34. *Kensche P.R., van Noort V., Dutilh B.E., Huynen M.A.* (2008) J. R. Soc. Interface, **5**, 151-170.
35. *Bowers P.M., Cokus S.J., Eisenberg D., Yeates T.O.* (2004) Science, **306**, 2246-2249.
36. *Campuzano V., Montermini L., Molto M.D., Pianese L., Cossee M., Cavalcanti F., Monros E., Rodius F., Duclos F., et al.* (1996) Science, **271**, 1423-1427.
37. *Huynen M.A., Snel B., Bork P., Gibson T.J.* (2001) Hum. Mol. Genet., **10**, 2463-2468.
38. *Chen O.S., Hemenway S., Kaplan J.* (2002) Proc. Natl. Acad. Sci. USA, **99**, 12321-12326.
39. *Wu J., Kasif S., DeLisi C.* (2003) Bioinformatics, **19**, 1524-1530.
40. *Loganantharaj R., Atwi M.* (2007) BMC Bioinformatics, **8,** Suppl. 7, S25.
41. *Pagel P., Wong P., Frishman D.* (2004) J. Mol. Biol, **344**, 1331-1346.
42. *Rodionov D.A., Gelfand M.S.* (2005) Trends Genet., **21**, 385-389.
43. *Haft D.H., Paulsen I.T., Ward N., Selengut J.D.* (2006) BMC Biol., **4**, 29.
44. *Li J.B., Gerdes J.M., Haycraft C.J., Fan Y., Teslovich T.M., May-Simera H., Li H., Blacque O.E., Li L. et al.* (2004) Cell, **117**, 541-552.
45. *Huynen M.A., Diaz-Lazcoz Y., Bork P.* (1997) Trends Genet., **13**, 389-390.
46. *Makarova K.S., Wolf Y.I., Koonin E.V.* (2003) Trends Genet., **19**, 172-176.
47. *Morett E., Korbel J.O., Rajan E., Saab-Rincon G, Olvera L., Olvera M., Schmidt S., Snel B., Bork P.* (2003) Nat. Biotechnol., **21**, 790-795.
48. *Mikkelsen T.S., Galagan J.E., Mesirov J.P.* (2005) Bioinformatics, **21**, 464-470.
49. *Watanabe H., Mori H., Itoh T., Gojobori T.* (1997) J. Mol. Evol., **44,** Suppl. 1, S57-S64.
50. *Mushegian A.R., Koonin E.V.* (1996) Trends Genet., **12**, 289-290.
51. *Tatusov R.L., Mushegian A.R., Bork P., Brown N.P., Hayes W.S., Borodovsky M., Rudd K.E., Koonin E.V.* (1996) Curr. Biol., **6**, 279-291.
52. *Yanai I., Mellor J.C., DeLisi C.* (2002) Trends Genet., **18**, 176-179.
53. *Enright A.J., Ouzounis C.A.* (2001) Genome Biol., **2**, RESEARCH0034.
54. *Moreno-Hagelsieb G., Collado-Vides J.* (2002) Bioinformatics, **18,** Suppl. 1, S329-S336.
55. *Price M.N., Huang K.H., Alm E.J., Arkin A.P.* (2005) Nucleic Acids Res., **33**, 880-892.
56. *Ermolaeva M.D., White O., Salzberg S.L.* (2001) Nucleic Acids Res., **29**, 1216-1221.
57. *Rogozin I.B., Makarova K.S., Murvai J., Czabarka E., Wolf Y.I., Tatusov R.L., Szekely L.A., Koonin E.V.* (2002) Nucleic Acids Res., **30**, 2212-2223.
58. *Li Q., Lee B.T., Zhang L.* (2005) BMC Genomics, **6**, 7.
59. *Huynen M.A., Snel B.* (2000) Adv. Protein Chem., **54**, 345-379.
60. *Negre B., Casillas S., Suzanne M., Sanchez-Herrero E., Akam M., Nefedov M., Barbadilla A., de Jong P., Ruiz A.* (2005) Genome Res., **15**, 692-700.
61. *Shoemaker B.A., Panchenko A.R.* (2007) PLoS Comput. Biol., **3**, e42.
62. *Barker D., Pagel M.* (2005) PLoS Comput. Biol., **1**, e3.
63. *Han J.D., Dupuy D., Bertin N., Cusick M.E., Vidal M.* (2005) Nat. Biotechnol., **23**, 839-844.
64. *Bader J.S., Chaudhuri A., Rothberg J.M., Chant J.* (2004) Nat. Biotechnol., **22**, 78-85.
65. *Date S.V., Marcotte E.M.* (2003) Nat. Biotechnol., **21**, 1055-1062.
66. *Ogata H., Goto S., Sato K., Fujibuchi W., Bono H., Kanehisa M.* (1999) Nucleic Acids Res., **27**, 29-34.
67. *Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S. et al.* (2000) Nat. Genet., **25**, 25-29.
68. *Schlicker A., Domingues F.S., Rahnenfuhrer J., Lengauer T.* (2006) BMC Bioinformatics, **7**, 302.
69. *Strong M., Graeber T.G., Beeby M., Pellegrini M., Thompson M.J., Yeates T.O., Eisenberg D.* (2003) Nucleic Acids Res., **31**, 7099-7109.

70. *Strong M., Mallick P., Pellegrini M., Thompson M.J., Eisenberg D.* (2003) Genome Biol., **4**, R59.
71. *Bowers P.M., Pellegrini M., Thompson M.J., Fierro J., Yeates T.O., Eisenberg D.* (2004) Genome Biol., **5**, R35.
72. *Mellor J.C., Yanai I., Clodfelter K.H., Mintseris J., DeLisi C.* (2002) Nucleic Acids Res., **30**, 306-309.
73. *Date S.V., Marcotte E.M.* (2005) Bioinformatics, **21**, 2558-2559.
74. *von Mering C., Jensen L.J., Kuhn M., Chaffron S., Doerks T., Kruger B., Snel B., Bork P.* (2007) Nucleic Acids Res., **35**, D358-D362.
75. *Wu J., Hu Z., DeLisi C.* (2006) BMC Bioinformatics, **7**, 80.
76. *Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A. et al.* (1995) Science, **269**, 496-512.
77. *Himmelreich R., Hilbert H., Plagens H., Pirkl E., Li B.C., Herrmann R.* (1996) Nucleic Acids Res., **24**, 4420-4449.
78. *Artamonova I.I., Frishman G., Frishman D.* (2007) BMC Bioinformatics, **8**, 261.

Поступила: 01. 10. 2008.

PREDICTION OF FUNCTIONALLY RELATED PROTEINS BY COMPARATIVE GENOMICS *IN SILICO*

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

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

Review is devoted to computational prediction of functionally related proteins by comparative genomics. Growing possibilities of biotechnology for genome sequencing lead to generation of sequences for millions of genes. However, function of majority of these genes is unknown, and can be determined experimentally only for a few of them. Therefore, accurate and robust methods for *in silico* prediction (annotation) of gene functions are highly required. We describe here the main techniques of comparative genomics, including the standard method based on transferring functions between homologous sequences and also context-based methods, including phylogenetic profiles and gene-neighbor approaches. Modern methods of comparative genomics allow obtaining correct functional annotations for more than a half of all organism proteins.

Key words: related proteins, protein-protein interactions, functional annotation, phylogenetic profiles, comparative genomics.