На правах рукописи

ПЯТНИЦКИЙ Михаил Алексеевич

ВЫЯВЛЕНИЕ ВЗАИМОСВЯЗАННЫХ БЕЛКОВ МЕТОДАМИ АНАЛИЗА ГЕНОМОВ

03.00.28 – биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2009 г.

Работа выполнена в Учреждении Российской академии медицинских наук Научно-исследовательском институте биомедицинской химии имени В.Н.Ореховича РАМН

Научный руководитель: доктор биологических наук
Лисица Андрей Валерьевич

Официальные оппоненты: доктор биологических наук, профессор
Коротков Евгений Вадимович

кандидат биологических наук
Артамонова Ирена Игоревна

Ведущая организация: Учреждение Российской академии наук Институт проблем передачи информации имени А. А. Харкевича РАН

Защита состоится «22» октября 2009 года в 14:30 на заседании Диссертационного совета Д 001.010.01 при Учреждении Российской академии медицинских наук Научно-исследовательском институте биомедицинской химии имени В.Н.Ореховича РАМН по адресу: 119121, г. Москва, Погодинская ул., д.10.

С диссертацией можно ознакомиться в библиотеке Учреждения Российской академии медицинских наук Научно-исследовательского института биомедицинской химии им. В.Н.Ореховича РАМН.

Автореферат разослан «____» сентября 2009 г.

Ученый секретарь Диссертационного совета,
кандидат химических наук Е.А. Карпова

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

1.1. Актуальность проблемы

Одной из важнейших задач современной биологии является выявление белков, которые либо физически взаимодействуют между собой (например, являются субъединицами белкового комплекса), либо взаимосвязаны функционально (участвуют в одних и тех же метаболических или регуляторных путях). Подобного рода анализ позволит идентифицировать белок-белковые взаимодействия, отвечающие за различные клеточные процессы, а также выявить возможные пути развития патологических состояний на молекулярном уровне.

Исследование взаимосвязанных белков в настоящее время особенно актуально благодаря успехам крупномасштабных проектов по секвенированию геномов различных организмов, что революционизировало современную биологию (Mushegian, 2007). В настоящее время данные о первичной структуре большинства белков получают путем трансляции соответствующих генов *in silico*, вместо непосредственного определения последовательности аминокислот.

Уже накоплены сведения о последовательностях геномов более тысячи организмов. При этом широко употребляющийся термин “расшифровка генома” не отражает реальность, поскольку знание последовательности ДНК само по себе не несет в явном виде информации о роли этого гена и кодируемого им белка в функционировании клетки. Для определения функции белка необходимы трудоемкие экспериментальные исследования. Проведение таких работ является отчасти искусством, в то время как секвенирование геномов – это хорошо отработанная технология. Поэтому основной проблемой, стоящей перед постгеномной биоинформатикой, можно считать наиболее рациональное использование имеющегося массива данных о секвенированных геномах для предсказания функций белков и выявления взаимосвязей между белками, обеспечивающих ключевые клеточные процессы.

Для решения указанной проблемы разработаны вычислительные методы, позволяющие предсказывать функции белков и взаимосвязи между ними. В дополнение к традиционному прогнозированию функции белка на основе гомологии, в последнее время добавились методы, опирающиеся на данные сравнительной геномики. Для поиска взаимосвязей между белками используются контекстные свойства генов – распределение гомологов в ряду организмов (Pellegrini et al., 1999), положение и относительный порядок следования генов на хромосоме (Overbeek et al., 1999), частота слияний генов (Marcotte and Marcotte, 2002). Такого рода методы называют контекстно-ориентированными, поскольку используемые свойства генов

имеют смысл лишь при одновременном их сравнительном исследовании в ряду геномов, то есть в определенном контексте. Анализ контекстных свойств гена показывает, что отдельный геном и их совокупность представляют собой особый тип данных, который нельзя сводить к простому сочетанию последовательностей ДНК (Koonin and Galperin, 2003).

Для предсказания групп взаимосвязанных белков в настоящей работе использовали метод филогенетических профилей, согласно которому функционально взаимосвязанные белки также связаны и эволюционно (Kensche et al., 2008). Предполагается, что гены, кодирующие взаимодействующие белки являются ко-эволюционирующими: либо совместно наследуются вновь образованным видом, либо элиминируются естественным отбором. Каждый белок изучаемого организма характеризуется бинарным вектором (филогенетическим профилем, ФП), определяющим присутствие гомолога гена, кодирующего данный белок, в ряду других геномов, называемых референтными. При наличии достаточного количества референтных геномов, каждая пара взаимосвязанных белков в рамках структурного комплекса, регуляторного или метаболического пути будет иметь схожие ФП.

В большинстве работ метод ФП используется для предсказания взаимосвязей между парами белков, то есть кластерный анализ ФП не применяется. В нашей работе использовали более общий подход, ориентированный на поиск групп взаимосвязанных белков. При этом для оптимизации метода ФП использовали данные о белках *E.coli K12* как наиболее изученного модельного микроорганизма, а применение отработанной методики проводили для *M.tuberculosis H37Rv* в связи с большой социальной и медицинской значимостью туберкулезной микобактерии.

Целью работы явилось выявление групп взаимосвязанных белков *E.coli K12* и *M.tuberculosis H37Rv* путем сравнения соответствующих филогенетических профилей, отражающих закономерности наследования генов в ряду прокариотических организмов.

Для достижения поставленной цели решались следующие **задачи**:

1. Определить численные критерии для оценки соответствия состава предсказанных групп взаимосвязанных белков и метаболических путей БД KEGG.
2. Исследовать степень соответствия между кластеризацией ФП белков *E.coli K12* и распределением белков по разделам БД KEGG в зависимости от набора референтных геномов, метода кластерного анализа и способа расчета различий между ФП. Предсказать группы взаимосвязанных белков *E.coli K12*, используя найденные оптимальные параметры метода ФП.

3. Оценить количество групп взаимосвязанных белков *M.tuberculosis H37Rv*, используя данные о ФП белков. Предсказать группы взаимосвязанных белков *M.tuberculosis H37Rv* путем кластеризации ФП на оцененное число групп.

4. Выявить дополнительные группы белков *M.tuberculosis H37Rv*, взаимосвязи внутри которых определяются наличием устойчивых логических отношений между ФП.

1.2. Научная новизна и практическая значимость

Новизна данной работы по сравнению с аналогичными подходами с применением кластеризации ФП (Glazko and Mushegian, 2004; Yamada et al., 2006) заключается в том, что для изучения алгоритмов предсказания групп взаимосвязанных белков использован математический аппарат для сравнения разбиения белков на группы как результата кластерного анализа ФП, и “эталонного” распределения белков по метаболическим путям БД KEGG. Получаемые значения степени соответствия этих двух разбиений, оцененные с помощью т.н. внешних индексов, позволяют проводить сравнительный анализ влияния различных параметров метода ФП на состав выявляемых групп белков. Этими параметрами являются: набор референтных геномов, метод кластеризации и мера расстояния между ФП. Изучая зависимость значений внешних индексов для различных значений этих параметров, в работе определили набор референтных геномов и комбинацию кластеризация/расстояние, которые обеспечивали максимальную точность работы метода ФП, с точки зрения соответствия состава метаболических путей БД KEGG. и предсказанных групп взаимосвязанных белков.

Практическое применение разработанного подхода иллюстрируется на примере предсказания взаимосвязанных групп белков *E.coli K12* и *M.tuberculosis H37 Rv*. Показано, что найденные кластеры белков соответствуют как физическим взаимодействиям между белками (например, субъединицам NADH-дегидрогеназы), так и функционально взаимосвязанным белкам (например, белкам, участвующим в формировании жгутиков). При этом количество групп взаимосвязанных белков *M.tuberculosis* оценивали с помощью т.н. внутренних индексов, опираясь исключительно на данные о ФП белков без привлечения какой-либо дополнительной информации (например, БД KEGG).

В работе впервые применен математический аппарат логической регрессии (Ruczinski et al., 2003) для анализа данных о ФП. Суть этого подхода состоит в поиске логических закономерностей среди ФП, когда присутствие в геноме одного гена может быть предсказано, используя наличие или отсутствие в геноме некоторого

набора других генов (предикторов). Тем самым, могут быть выявлены дополнительные взаимосвязи между белками, кодируемыми соответствующими генами. Использование аппарата логической регрессии для поиска взаимосвязей между белками, является обобщением и развитием метода логического анализа ФП, предложенного в работе (Bowers et al., 2004).

Метод логической регрессии применялся для анализа данных о ФП белков *M.tuberculosis*. Показано, что получаемые таким образом группы логически ассоциированных между собой белков имеют биологический смысл и позволяют выдвигать новые гипотезы о взаимосвязях между белками в клетке. При этом предсказанные взаимосвязи принципиально отличаются от тех, которые могли быть получены при кластерном анализе ФП.

1.3. Апробация работы

Основные положения диссертационной работы докладывались и обсуждались в ходе следующих конференций: “Международный конгресс «Протеом человека», Лонг Бич, США, 2006”, “Международный конгресс «Протеом человека», Сеул, Корея, 2007”, “XV Российский национальный конгресс «Человек и Лекарство», Москва, 2008”, “IV Международная конференция «Геномика, протеомика, биоинформатика и нанобиотехнологии для медицины», Москва, 2008”, “Международная конференция по биоинформатике регуляции и структуры генома, Новосибирск, 2008”, “Московская конференция по вычислительной молекулярной биологии, Москва, 2009”.

1.4. Основные положения, выносимые на защиту

1. Расчет внешних индексов позволяет оптимизировать параметры метода ФП, а также сопоставлять метаболические пути БД KEGG и найденные кластеры взаимосвязанных белков.

2. Наилучшее соответствие состава найденных кластеров взаимосвязанных белков *E.coli K12* и метаболических путей KEGG достигается при использовании набора из 150 референтных геномов, кластеризации методом PAM и использовании взаимной информации в качестве меры расстояния между ФП.

3. Внутренние индексы можно использовать для предсказания количества групп взаимосвязанных белков. Кластеризация ФП на определенное таким образом количество групп, позволяет предсказать как физические взаимодействия, так и функциональные взаимосвязи между белками.

4. Применение логической регрессии для данных о ФП белков позволяет предсказывать взаимосвязи между белками, отличающиеся от тех которые обнаруживаются кластерным анализом.

1.5. Публикации

Материалы диссертационной работы отражены в 12 публикациях: в 3 статьях в журналах, входящих в список ВАК, и 9 материалах научных конференций.

1.6. Объем и структура диссертации

Диссертация изложена на 119 страницах машинописного текста, включая 26 рисунков и 3 таблицы. Состоит из глав: «Введение», «Обзор литературы», «Материалы и методы», «Результаты и обсуждение», «Заключение», «Выводы», «Список литературы», «Приложение».

2. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Филогенетические профили для белков *E.coli K12* и *M.tuberculosis H37Rv* загрузили из БД KEGG, раздел Orthology [http://www.genome.jp/kegg/ko.html]. Матрица ФП для *E.coli* состояла из 1184 строк (соответствовавших ФП каждого белка) и 588 столбцов (референтных геномов). Для каждого белка присутствие/отсутствие ортолога в каждом референтном организме кодировалось единицей или нулем соответственно. Аналогичная матрица ФП для *M.tuberculosis* состояла из 1770 строк и 588 столбцов.

Информацию о метаболических путях для *E.coli K12* загружали из БД KEGG. Для описания принадлежности *i*-го белка к *j*-ому метаболическому пути использовали представление в виде матрицы принадлежности **M**, $m_{ij} = n_j / N_i$, где n_j – количество белков в *j*-ом метаболическом пути, а N_i – суммарное количество белков во всех остальных путях, к которым принадлежит *i*-ый белок. Считали, что каждый из 124 метаболических путей *E.coli K12* представляет собой группу взаимосвязанных белков.

Предсказание групп взаимосвязанных белков осуществляли посредством кластерного анализа матриц ФП, полученных для белков *E.coli* и *M.tuberculosis*. Использовали несколько различных мер расстояний и методов кластерного анализа. Среди мер расстояния использовали: расстояния Хэмминга, Жаккара, Кульчинского; вероятность случайного совпадения ненулевых элементов ФП; взаимную информацию. Использовали иерархические и итеративные методы кластерного анализа. Среди первой группы использовали агломеративные методы (методы ближней, средней и полной связей, метод Уорда) и дивизивный метод (метод DIANA). Результатом работы иерархических методов являлась последовательность объединения белков в кластеры для различных уровней сходства соответствующих ФП (дендрограмма). В качестве итеративного метода кластерного анализа

использовали метод PAM, результатом применения которого являлся непосредственный состав кластеров взаимосвязанных белков.

Сравнение кластеризаций белков. После кластерного анализа данных о ФП белков проводили оценку, насколько адекватным являлось полученное разбиение белков на группы. В соответствии с особенностями задачи рассматривали две ситуации. Если *a priori* было известно "корректное" распределение белков по разделам (метаболическим путям) БД KEGG, то для оценки степени соответствия полученной кластеризации этому "корректному" распределению применяли внешние индексы оценки кластеризации: расстояние между матрицами принадлежности и индекс Рэнда. Если же "истинное" распределение белков на группы считали неизвестным, то использовали внутренние индексы, оценивающие только структуру полученной кластеризации белков.

Расстояние между матрицами принадлежности использовали в качестве меры соответствия составов полученной кластеризации белков и метаболических путей KEGG. Информацию об отнесении белков к метаболическим путям KEGG описывали матрицей принадлежности $\mathbf{M}_1$, а группирование белков, полученное в результате кластерного анализа ФП, описывали матрицей принадлежности $\mathbf{M}_2$. Перемешивание идентификаторов кластеров не изменяет разбиение белков, что эквивалентно перестановке столбцов матрицы принадлежности $\mathbf{M}_2$. Расстояние между матрицами принадлежности $\mathbf{M}_1$ и $\mathbf{M}_2$ определяли как минимальное расстояние с учетом всех возможных перестановок столбцов $\mathbf{M}_2$, т.е. выполнялось:

$$d\left(\mathbf{M}_1, \mathbf{M}_2\right) = \min_{\mathbf{P}} \left\| \mathbf{M}_1 - \mathbf{M}_2\mathbf{P} \right\| \qquad (1)$$

Минимизацию проводили по всем матрицам-перестановкам[1] **P**. Согласно (Hornik, 2005) получили, что минимизация (1) эквивалентна максимизации $tr\left(\mathbf{M}_1'\mathbf{M}_2\mathbf{P}\right) = \sum_{i,k} m_{1ik} m_{2i\pi(k)}$. Нахождение оптимальной матрицы-перестановки **P,** обеспечивающей оптимальное соответствие друг другу метаболических путей KEGG и кластеризации матрицы ФП**,** рассматривали как вариант задачи о назначениях, которую точно решали методом линейного программирования. Меньшие значения расстояния между указанными матрицами принадлежности означали лучшее соответствие состава групп белков, получаемых при кластеризации ФП, и метаболических путей KEGG.

---

[1] Матрица перестановки – квадратная бинарная матрица, в каждой строке и столбце которой находится лишь один единичный элемент. Матрица перестановки размера *n* x *n* является матричным представлением перестановки порядка *n*, т.е. может быть использована для описания перестановки столбцов

Индекс Рэнда согласно (Gan et al., 2007) использовали для сравнения двух разбиений G_1 и G_2, где G_1 – группирование белков в соответствии с метаболическими путями KEGG, а G_2 – группирование белков, полученное при кластеризации матрицы ФП. Индекс Рэнда отличается от расстояния между матрицами принадлежности, поскольку использует фиксированные разбиения (белок принадлежит только к одной группе), тогда как матрица принадлежности может описывать нечеткие разбиения (белок принадлежит к нескольким метаболическим путям).

Для каждой пары белков x_i и x_j был возможен один из четырех вариантов:

а) x_i и x_j находились в одной группе в G_1 и в G_2.

б) x_i и x_j находились в разных группах в G_1 и в G_2.

в) x_i и x_j находились в одной группе в G_1, но в разных группах в G_2.

г) x_i и x_j находились в одной группе в G_2, но в разных группах в G_1.

Количество случаев, соответствующих приведенным пунктам, обозначили как a, b, c, d соответственно. Рассчитывали модифицированный индекс Рэнда:

$$R' = \frac{2(ab - cd)}{\left((a+d)(d+b)+(a+c)(c+b)\right)} \qquad (2)$$

Значения индекса Рэнда находились в диапазоне от 0 до 1, причем большие значения означали лучшее соответствие состава кластеров, получаемых при кластеризации ФП, метаболическим путям KEGG.

Внутренние индексы: индекс Дэвиса-Болдуина, индекс “ширина силуэта”, индекс Дана, Г-статистику Хьюберта, отношение среднего внутрикластерного расстояния к среднему междукластерному, вычисляли согласно работе (Gan et al., 2007).

Оценку оптимального количества кластеров проводили с помощью L-метода. Искали точку перегиба зависимости внутреннего индекса от количества кластеров, на которое проводилось разбиение. Производили подбор двух линейных регрессий для оптимальной аппроксимации этой зависимости в смысле минимизации суммы квадратов отклонений. Поиск наилучшей точки разделения двух линейных регрессий проводили методом динамического программирования. Оптимальное количество кластеров определяли как абсциссу точки пересечения указанных регрессий. Наилучшая аппроксимация зависимости внутреннего индекса от числа кластеров двумя прямыми с нулевой суммарной ошибкой показана на рисунке 1-в. Пример определения оптимального числа кластеров показан на рисунке 1-д.

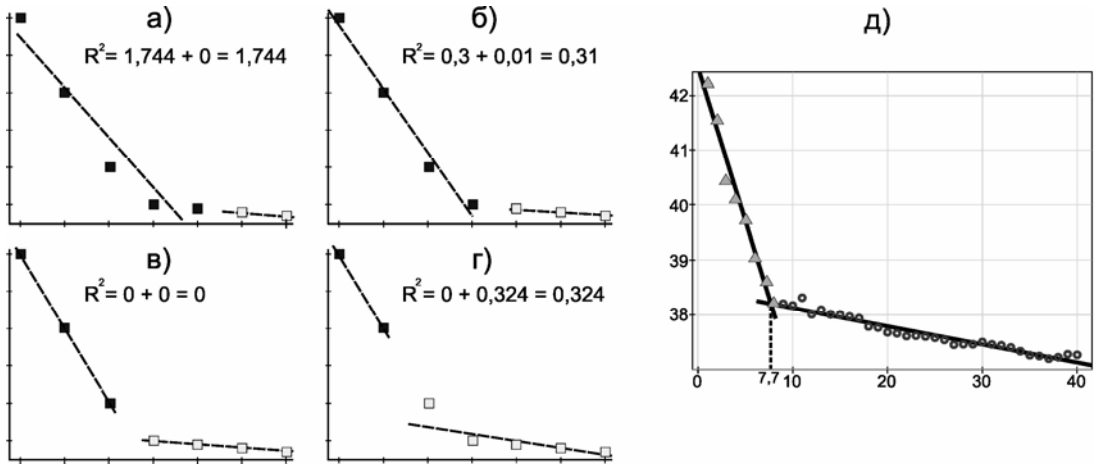

Рисунок 1. а-г) Принцип работы L-метода. Поиск оптимальной аппроксимации зависимости внутреннего индекса от количества кластеров двумя прямыми с наименьшей возможной ошибкой. Абсцисса – количество кластеров, ордината – значения любого внутреннего индекса для разбиения на соответствующее количество кластеров. д) Пример определения оптимального количества кластеров с помощью L-метода – абсцисса точки пересечения двух найденных линейных регрессий

Логическую регрессию (Ruczinski et al., 2003) применяли для поиска логических закономерностей в ФП. Логическая регрессия решает задачу предсказания значений целевого ФП (зависимой переменной) как логическую комбинацию других ФП-предикторов.

Для поиска оптимальной логической комбинации предикторов искомое логическое выражение представляли в виде двоичного дерева. Узлами дерева являются логические операции AND (И) и OR (ИЛИ), листьями дерева являются бинарные предикторы или их логические отрицания (NOT, НЕ). Оценочную функцию, определяющую, насколько модель соответствует исходным данным, вводили как количество элементов целевого ФП, для которых не выполнялось найденное логическое правило.

Оптимизацию оценочной функции проводили пошаговым методом. Определили набор обратимых шагов, позволяющих переходить от одного логического дерева к другому, производя, таким образом, поиск наилучшего логического выражения, предсказывающего целевой ФП. Использовали шаги: изменить лист (ФП-предиктор), изменить оператор (узел дерева), добавить новую ветвь, удалить ветвь, разделить лист, удалить лист. В новой точке пространства логических деревьев вычисляли оценочную функцию, чтобы определить, произошло ли улучшение модели (уменьшение значения оценочной функции) по сравнению с предыдущим шагом. Для минимизации значений оценочной функции использовали метод симулированного отжига (Kirkpatrick et al., 1983).

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

3.1 Общая методология работы

Согласно основному допущению метода ФП белки, обладающие сходными ФП, являются взаимосвязанными. Однако, в большинстве работ подобные предсказания проводятся лишь для пары белков, при этом используется единый подход. Прежде всего, вводится метрика, определяющая взаимосвязь между парой белков. Обычно с этой целью используют расстояние Хэмминга, расстояние Жаккара или взаимную информацию между ФП. Затем выбирается определенное значение метрики (порог), и все пары белков, для которых метрика оказывается больше порога, считаются взаимосвязанными. Результаты работы указанного алгоритма обычно сравниваются с базой данных по взаимодействующим белкам, оценивается число ложноположительных и ложноотрицательных предсказаний.

В нашей работе развит подход, ориентированный на поиск групп взаимосвязанных белков (функциональных белковых модулей). Группы взаимосвязанных белков формируются в результате применения методов кластерного анализа к матрице расстояний между ФП белков. Постановка задачи в форме поиска функциональных модулей по сравнению с предсказанием пар взаимосвязанных белков отражает системный подход, позволяющий полнее раскрыть контекст найденных взаимосвязей между белками. Например, изучение аннотаций белков, входящих в одну группу с неохарактеризованным белком, может способствовать формированию гипотез о возможной функции этого белка.

Для применения сформулированного подхода к поиску функциональных белковых модулей требуется указать ряд параметров, влияющих на результаты работы алгоритма: метод кластеризации, способ определения расстояния между ФП и набор референтных геномов. Соответственно для максимизации точности предсказания функциональных белковых модулей в смысле соответствия данным KEGG было необходимо провести поиск оптимальных значений указанных параметров.

Для количественной оценки точности предсказаний функциональных белковых модулей в работе использовали математический аппарат сравнения двух разбиений. Под разбиениями понимали распределение ряда белков на группы. Например, разбиением являлся результат кластерного анализа ФП – распределение белков на группы согласно мере схожести соответствующих ФП. Распределение белков по метаболическим путям, известным из БД KEGG, также рассматривали как разбиение, с той лишь разницей, что один белок мог быть аннотирован как относящийся одновременно к нескольким метаболическим путям. Математически разбиение

описывается с помощью матрицы принадлежности. Для анализа точности предсказаний функциональных белковых модулей сравнивали два разбиения – матрицу принадлежности $\mathbf{M_1}$, описывавшей данные о метаболических путях, и матрицу принадлежности $\mathbf{M_2}$, описывавшей результаты кластерного анализа ФП.

В работе использовали два внешних индекса для оценки степени соответствия двух разбиений: евклидово расстояние между матрицами принадлежности $\mathbf{M_1}$ и $\mathbf{M_2}$ и индекс Рэнда. В результате, задачу оптимизации метода ФП для предсказания групп взаимосвязанных белков свели к сравнению внешних индексов разбиений, получаемых при определенных значениях параметров. Нужно отметить, что подобная методика является универсальной и может быть применена для изучения любого алгоритма предсказания групп взаимосвязанных белков.

3.2 Определение оптимального набора референтных геномов

Точность предсказаний взаимосвязанных белков методом ФП существенно зависит от набора используемых референтных геномов. Поэтому мы определили оптимальный набор референтных геномов, максимизирующий соответствие состава полученных кластеров белков и метаболических путей KEGG. Для этого из референтных геномов исключили как чрезмерно филогенетически близкие, так и чрезмерно филогенетически удаленные организмы.

ФП для белков *E.coli K12* загрузили из БД KEGG, раздел Orthology. Первоначально набор референтных геномов включал в себя данные о 1067 организмах (827 геномов бактерий, 62 генома архей и 178 геномов эукариот). Далее из матрицы ФП исключили все геномы эукариот, поскольку использование таких чрезмерно филогенетически удаленных видов ухудшает точность работы метода. Затем были выявлены все случаи, когда определенный вид микроорганизмов был представлен более чем одним штаммом, и из матрицы ФП были исключены все геномы таких штаммов за исключением одного, выбранного случайным образом.

На этой стадии отобранный набор включал 588 геномов, в том числе 54 генома архей и 534 генома бактерий. Для определения относительных расстояний между референтными геномами, построили неукорененное филогенетическое дерево, используя объединенные последовательности генов, присутствующих одновременно во всех геномах (*rpsC*, *rpsD*, *rpsG*). Для множественного выравнивания использовали программу MUSCLE 3.6 с параметрами по умолчанию. Филогенетическое дерево построили методом максимального правдоподобия с помощью программы protml из пакета PHYLIP 3.6 с параметрами по умолчанию. Используя полученную матрицу филогенетических расстояний между геномами, определяли группы

близкородственных микроорганизмов. Из каждой группы случайным образом выбирали один геном.

Сравнительный анализ результатов для матриц ФП с различным количеством столбцов позволил изучить влияние набора референтных геномов на точность предсказаний взаимосвязанных белков. Кластеризацию матрицы ФП проводили для различных комбинаций метода кластеризации и способа определения расстояния между ФП. Количество кластеров задавали равным 124, согласно количеству известных метаболических путей для *E.coli K12*.

На рисунке 2 представлена зависимость индекса Рэнда от количества референтных организмов. Индекс Рэнда рассчитывали на основе сопоставления состава полученных кластеров взаимосвязанных белков и метаболических путей KEGG, причем большие значения индекса означали лучшее соответствие. На рисунке показаны только комбинации кластеризация/расстояние, давшие наилучшие результаты. Количество референтных организмов находилось в диапазоне от 11 до 588, при этом большее значение соответствовало присутствию групп близкородственных организмов.

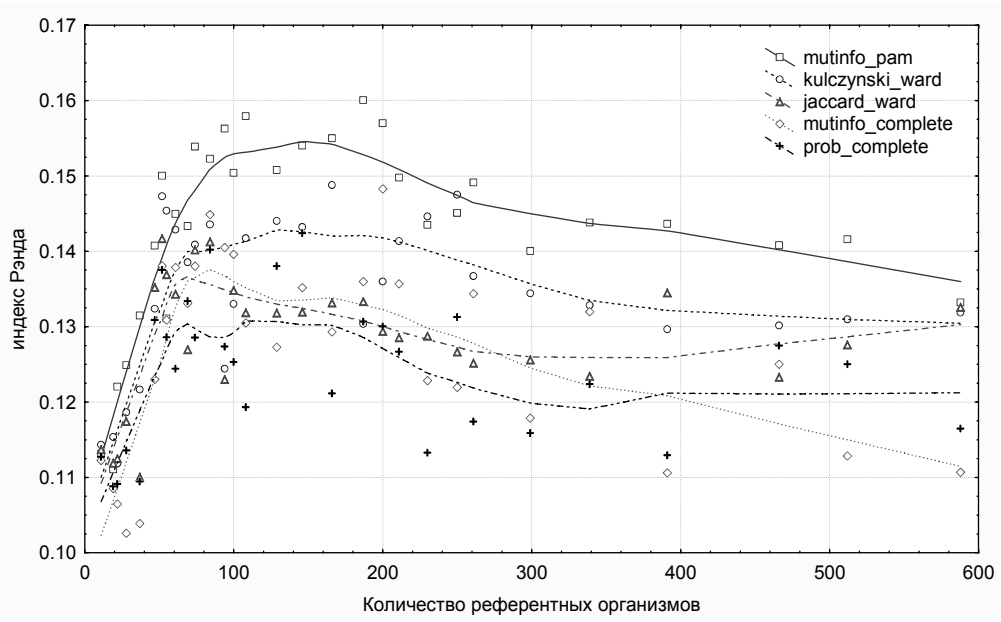

Рисунок 2. Зависимость индекса Рэнда от количества референтных организмов, используемых в матрице ФП. Меры расстояния между ФП: по Кульчинскому (kulczynski), по Жаккару (jaccard), взаимная информация (mutinfo), вероятность случайного совпадения ненулевых элементов ФП (prob). Методы кластеризации: Уорд (ward), PAM, метод дальней связи (complete). Сглаживание произведено методом Loess с параметром 0,4.

Как следует из рисунка 2, для всех комбинаций расстояние/кластеризация зависимость индекса Рэнда от количества используемых референтных организмов имеет приблизительно одинаковый характер. Индекс Рэнда быстро возрастает при

увеличении количества референтных геномов от 11 до 100, выходит на плато в диапазоне от 100 до 200 и далее постепенно снижается. Наилучшие результаты дает кластеризация методом PAM по взаимной информации, среднее значение индекса Рэнда в диапазоне 100-200 референтных организмов составило 0,155. Применение метода Уорда и расстояния Кульчинского дало несколько худшие результаты: среднее значение индекса Рэнда в том же диапазоне оказалось равным 0,140. Максимальная точность остальных комбинаций кластеризация/расстояние была в среднем ниже и составила 0,13.

Исходя из полученных данных можно заключить, что использование 100-200 референтных организмов в ФП является оптимальным с точки зрения соответствия состава полученных кластеров белков и метаболических путей KEGG. Применение же большего количества референтных геномов ухудшает результаты метода ФП. В дальнейшем использовали набор из 150 референтных организмов, названия которых приведены в Приложении 1 диссертационной работы. Полученный характер зависимости точности метода ФП от количества референтных геномов соответствует литературным данным, опубликованным ранее (Sun et al., 2005). В этой работе было показано, что при использовании более 86 референтных геномов точность работы метода ФП для *E.coli* перестает существенно повышаться, и достигает максимума при 145 референтных организмах.

3.3 Определение оптимальных параметров кластеризации ФП

Провели перебор 25 комбинаций методов кластеризации и способов определения расстояний для ФП белков *E.coli K12*, используя определенный ранее набор из 150 референтных геномов. Для всех использованных методов кластерного анализа было необходимо указывать количество кластеров, на которые происходит группирование белков. Количество кластеров варьировали в диапазоне от 5 до 250, для каждого значения количества кластеров и комбинации кластеризация/расстояние рассчитывали внешние индексы: расстояние между матрицами принадлежности KEGG и полученной кластеризации, а также индекс Рэнда.

На рисунке 3 представлена зависимость расстояния между матрицами принадлежности KEGG и полученной кластеризации от количества кластеров для наилучших комбинаций кластеризация/расстояние. Меньшие значения индекса соответствуют лучшему совпадению полученных кластеров белков с метаболическими путями KEGG. В результате минимизации данного индекса для каждой комбинации кластеризация/расстояние получили матрицу перестановки **P**, согласно которой каждый кластер был наилучшим образом сопоставлен

соответствующему метаболическому пути KEGG. Так, для кластеризации на 124 кластера методом PAM с использованием взаимной информации как расстояния было получено, что 38,3% белков, входящих в один из метаболических путей KEGG, оказались в одном кластере. Для всех комбинаций кластеризация/расстояние характерен одинаковый тип зависимости: постепенное уменьшение индекса с увеличением количества кластеров, выход на плато достигается в районе 130-140 кластеров.

Наилучшее совпадение кластеризации ФП с данными KEGG было получено при кластерном анализе методом PAM и использовании взаимной информации в качестве меры расстояния. Для 110 кластеров был достигнут глобальный минимум индекса равный 37,2 (отмечено стрелкой, рис. 2). Поскольку в БД KEGG число метаболических путей для *E.coli K12* равно 124 (показано пунктирной линией), то можно отметить достаточно хорошее согласие в оценке количества кластеров.

Следующее по степени соответствия метаболическим путям KEGG группирование белков было получено при кластеризации по методу Уорда и использовании взаимной информации в качестве меры расстояния. Полученный при этом глобальный минимум индекса равен 37,5 при 122 кластерах, чем почти достигается соответствие числа кластеров и метаболических путей KEGG. В районе 120-125 кластеров отмечается приблизительное равенство значений индекса с методом PAM.

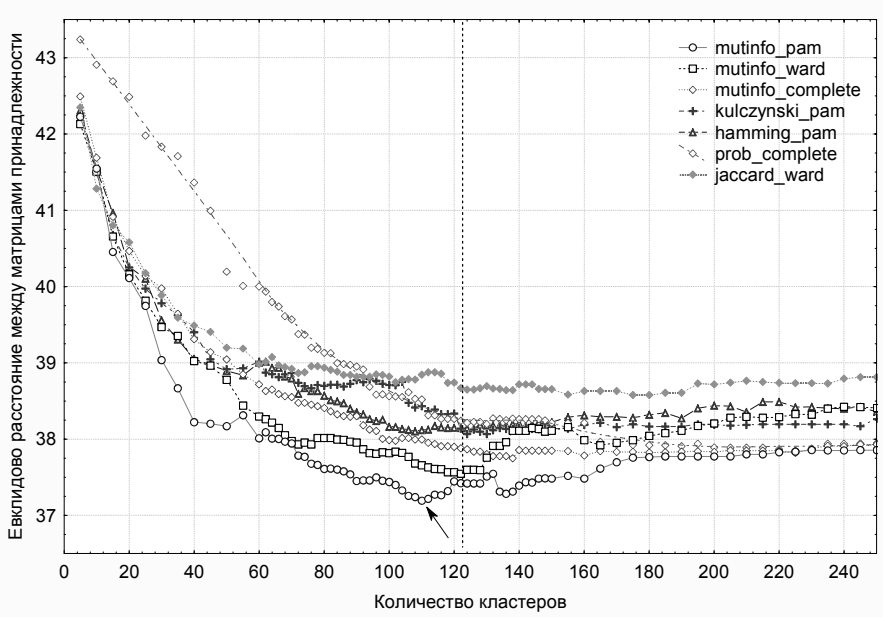

Рисунок 3. Зависимость расстояния между матрицами принадлежности от количества кластеров, по которым происходит распределение ФП. Меры расстояния: взаимная информация между ФП (mutinfo), расстояние Хэмминга (hamming), Жаккара (jaccard), Кульчинского (kulczynski), вероятность случайного совпадения ненулевых элементов ФП (prob). Методы кластеризации: Уорда (ward), PAM и полной связи (complete).

Следует отметить, что для трех наиболее соответствующих данным KEGG комбинаций кластеризация/расстояние используется взаимная информация, в то время как метод кластерного анализа различается. На основании этого можно предположить, что используемая мера расстояния между ФП в большей степени влияет на результаты предсказания групп взаимосвязанных белков, чем метод кластеризации. Среди способов кластеризации лидируют методы PAM, Уорда и полной связи. Остальные методы кластеризации показывают худшие результаты (не приведены на рисунке).

Аналогичные результаты по определению наилучшей комбинации кластеризация/расстояние, полученные для индекса Рэнда, согласовывались с результатами для расстояния между матрицами принадлежности. В обоих случаях списки наилучших комбинаций метода кластеризации и расстояния между ФП во многом совпадают как по составу, так и по ранжированию. При этом среди способов кластеризации также встречаются только методы Уорда, полной связи и PAM.

Таким образом, можно заключить, что наилучшее согласие в составах получаемых кластеров белков и метаболических путей KEGG достигается при использовании взаимной информации как меры расстояния между ФП и метода PAM для кластеризации ФП. Выбор меры расстояния между ФП оказывает большее влияние на состав полученных групп белков по сравнению с использованным способом кластеризации.

3.4 Предсказание состава известных метаболических путей методом ФП

Для выявления взаимосвязанных белков *E.coli K12* использовали взаимную информацию в качестве расстояния между ФП и метод PAM как вариант кластерного анализа согласно ранее полученным результатам. Провели кластеризацию ФП на 124 кластера, в соответствии с количеством известных метаболических путей для *E.coli K12* согласно БД KEGG.

При вычислении расстояния между матрицами принадлежности кластеризации ФП и метаболических путей KEGG, определили матрицу перестановки, задающую наилучшее соответствие полученных кластеров взаимосвязанных белков и метаболических путей. Для каждого метаболического пути определили количество белков, которые входят в кластер, наилучшим образом соответствующий данному пути.

На рисунке 4 приведена диаграмма, отражающая степень соответствия состава кластеров ФП разделам БД KEGG. Как следует из рисунка 4, пять метаболических путей KEGG воспроизводятся с точностью, превышающей 75%, в то время как 80

метаболических путей KEGG воспроизводятся с точностью менее 25%. Полная таблица результатов приведена в Приложении 2 диссертационной работы.

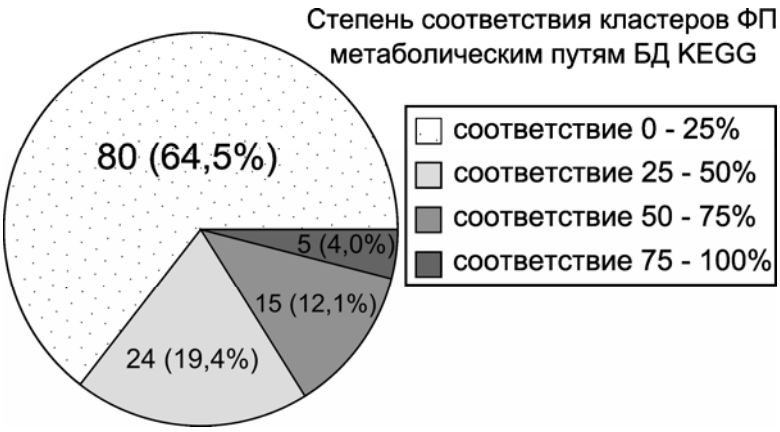

Рисунок 4. Степень соответствия состава кластеров ФП метаболическим путям БД KEGG для белков *E.coli K12*. Для каждого метаболического пути определили кластер, наилучшим образом ему соответствующий. Цветом показана доля белков в метаболическом пути, входящих в указанный кластер.

Максимально точно был воспроизведен состав метаболического пути с идентификатором 02040, куда входит 38 белков участвующих в сборке жгутиков. Наилучшим образом этому пути соответствует кластер №42, включающий в себя, кроме 35 белков для сборки жгутика, еще 3 белка (csrA, nrfA, cheZ), которые аннотированы как принадлежащие к другим разделам KEGG. Таким образом, для указанного раздела KEGG точность воспроизведения составила 35/38 = 91,2%.

Проведенный анализ показал, хотя в БД KEGG белки csrA и cheZ не отнесены к разделу 02040, тем не менее, оба этих белка связаны с функционированием жгутика. Согласно БД UniProt, белок csrA необходим для пост-трансляционной активации экспрессии генов *FlhC* и *FlhD*. Белок cheZ является фосфатазой, участвующей в генерации регулирующего сигнала для вращения жгутика. Для белка nrfA (периплазматической нитрит-редуктазы) установить прямую взаимосвязь с функционированием жгутика не удалось. Однако, косвенным свидетельством в пользу такой взаимосвязи может являться тот факт, что у мутантных *E.coli* по гену *hmp*, отвечающим за метаболизм оксида азота, по неустановленной причине наблюдались нарушения в формировании жгутиков (Stevanin et al., 2007). Тем самым, практически для всех белков кластера №42 удалось подтвердить их участие в функционировании жгутика, несмотря на то, что согласно БД KEGG, вышеуказанные три белка с этим процессом не связаны.

3.5 Предсказание групп взаимосвязанных белков *de novo*

При предсказании взаимосвязанных белков *E.coli K12* мы опирались на информацию о их распределении по разделам базы KEGG. Эти данные использовали в качестве эталона для определения параметров метода ФП, максимизирующих соответствие состава найденных кластеров взаимосвязанных белков и метаболических путей БД KEGG. При этом количество таких групп полагали заранее известным и равным числу метаболических путей *E.coli*.

На практике возможна другая постановка задачи: как можно использовать информацию о ФП для предсказания групп взаимосвязанных белков *de novo*, то есть в ситуации, когда отсутствуют данные о белковых взаимодействиях и априорные биохимические сведения. Подобная формулировка проблемы возникает при изучении организма, для которого известна только последовательность генома, но нет экспериментальных сведений о функциях белков.

Важнейшим параметром, определяющим состав получаемых групп взаимосвязанных белков, является количество кластеров, на которое производится разбиение. При наличии аннотаций белков из KEGG, количество кластеров должно быть примерно равным количеству молекулярных процессов, обеспечивающих существование изучаемого организма. Поскольку в отношении малоизученных организмов функциональные аннотации для белков отсутствуют или ненадежны, то необходимо установить оптимальное количество кластеров без привлечения какого-либо дополнительного источника информации, помимо матрицы ФП.

Несмотря на то, что в БД KEGG содержится информация о 118 метаболических путях *M.tuberculosis H37Rv*, мы не использовали эти сведения в изложенных далее вычислительных экспериментах. Группы взаимосвязанных белков *M.tuberculosis* выявляли исключительно путем кластерного анализа данных о 1770 соответствующих ФП. Для каждого количества кластеров в диапазоне от 5 до 150 рассчитывали набор внутренних индексов для оценки “качества” полученного разбиения. Использовали 25 комбинаций метода кластеризации и меры расстояния между ФП. Для наилучшей в смысле каждого внутреннего индекса комбинации кластеризация/расстояние анализировали зависимость указанного индекса от количества кластеров. Согласно L-методу определяли точку, в которой характер зависимости внутреннего индекса от числа кластеров изменялся. Полученное значение количества кластеров рассматривали в качестве оценки предполагаемого количества молекулярных процессов.

В таблице 1 приведены оценки оптимального количества групп взаимосвязанных белков *M.tuberculosis* для различных внутренних индексов. В

последней колонке приведены комбинации кластеризации/расстояния, для которых было достигнуто наилучшее “качество” получаемого разбиения белков по кластерам согласно каждому индексу.

Таблица 1. Оценки оптимального количества групп взаимосвязанных белков *M.tuberculosis* для различных внутренних индексов. Показаны четыре лучшие комбинации метода кластеризации и расстояния между ФП.

Внутренний индекс	**Оценка кол-ва кластеров**	**Лучшие сочетания расстояние/кластеризация**
Индекс Дана	99	mutinfo_complete mutinfo_pam mutinfo_ward kulczynski_complete
Индекс Дэвиса-Болдуина	93	prob_average prob_complete prob_ward prob_diana
Г-статистика Хьюберта	85	mutinfo_complete kulczynski_complete kulczynski_average mutinfo_diana
Отношение внутрикластерного расстояния к междукластерному	98	prob_ward prob_complete prob_diana prob_average
“Ширина силуэта”	106	hamming_ward, hamming_pam jaccard_ward prob_ward

Наименьшую оценку количества кластеров (85) дает использование Г-статистики Хьюберта. Максимальную оценку количества кластеров (106) получили с использованием индекса “ширина силуэта”, причем эта оценка примерно соответствует данным KEGG (118 метаболических путей).

Анализируя лучшие комбинации метода кластеризации и расстояния между ФП для каждого внутреннего индекса можно отметить, что мера расстояния оказывает большее влияние на получаемую кластеризацию белков по сравнению с применяемым вариантом кластерного анализа. Это согласуется с данными, полученными ранее для внешних индексов при анализе ФП белков *E.coli*. Так, для индекса Дана в трех из четырех комбинациях кластеризация/расстояние используется взаимная информация (mutinfo). Для индекса Дэвиса-Болдуина и для отношения внутрикластерного расстояния к междукластерному в качестве “наилучшей” меры расстояния используется только вероятность случайного совпадения ненулевых

элементов ФП (prob), в то время как преобладающий способ кластеризации не выявлен.

Использование внутренних индексов не позволяет однозначно определить оптимальную комбинацию кластеризация/расстояние. Поэтому для поиска групп взаимосвязанных белков *M.tuberculosis H37Rv* использовали комбинацию кластеризация/расстояние, установленную ранее на данных о ФП белков *E.coli*, а именно: кластерный анализ методом PAM и взаимную информацию (mutinfo_pam).

Определение количества кластеров взаимосвязанных белков мы проводили на основе данных о поведении внутренних индексов. Усреднив количество кластеров, определенное для каждого внутреннего индекса, получили, что оптимальное количество функциональных модулей для *M.tuberculosis* равно 96. Провели кластерный анализ ФП белков *M.tuberculosis* для разбиения на 96 кластеров. При этом мы не использовали никакой априорной информации, кроме матрицы ФП.

В качестве примера полученных результатов на рисунке 5 представлены ФП из кластера №33. В состав этого кластера полностью вошли все 20 цитохромов P450, закодированные в геноме *M.tuberculosis*. В кластер также вошел белок choD, являющийся оксидоредуктазой, участвующей в метаболизме холестерина. Можно отметить идеальное совпадение ФП для всех цитохромов P450. Присутствие в геноме столь большого количества цитохромов P450 сильно выделяет туберкулезную микобактерию среди остальных прокариот. Высказываются предложения использовать цитохромы P450 в качестве мишени для создания противотуберкулезного лекарственного препарата (McLean et al., 2007), например, на основе веществ из группы азолов

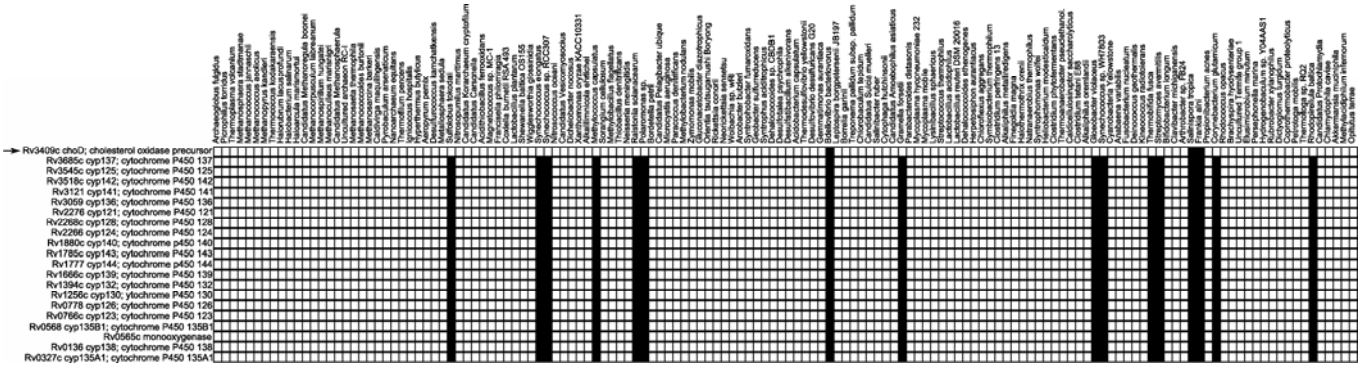

Рисунок 5. Филогенетические профили кластера №33. Большинство белков в кластере являются цитохромами P450. Стрелкой отмечен белок choD, являющийся оксидоредуктазой.

Другой пример приведен на рисунке 6, где представлены ФП из кластера №11. В состав этого кластера входят исключительно субъединицы NADH-дегидрогеназы типа I. Из 14 субъединиц, участвующих в формировании этого белкового комплекса,

в данный кластер вошли 11. Этот пример иллюстрирует применнимость кластеризации ФП для выявления физически взаимодействующих белков, таких как субъединицы в составе одного белкового комплекса. Согласно работе (Velmurugan et al., 2007), одна из субъединиц указанной NADH-дегидрогеназы, а именно *nuoG*, участвует в ингибировании апоптоза макрофагов, что является необходимым фактором вирулентности патогенных микобактерий.

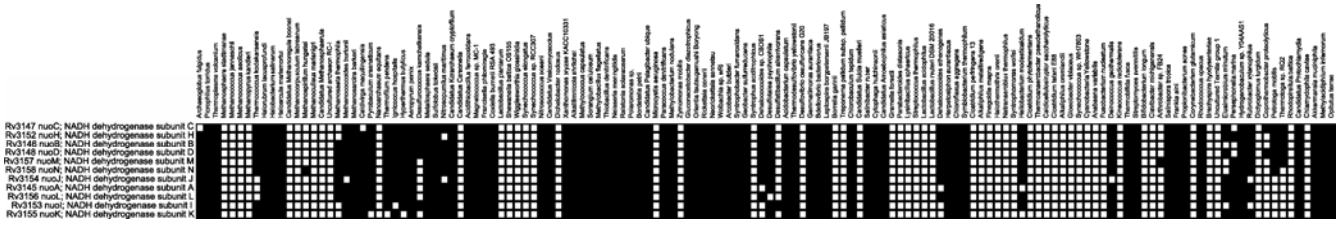

Рисунок 6. ФП для кластера №11. Все белки в кластере являются субъединицами NADH-дегидрогеназы.

Таким образом, кластеризация ФП белков *M.tuberculosis* на оцененное по внутренним индексам число групп может использоваться для предсказания как физических взаимодействий между белками, так и функциональных взаимосвязей.

3.6 Поиск логических взаимосвязей между ФП белков

В дополнение к поиску групп белков на основании сходства ФП, в работе использовали другой подход к выявлению взаимосвязанных белков. Впервые данный метод, названный логическим анализом филогенетических профилей, был предложен в работе (Bowers et al., 2004). Подход состоит в поиске логических конструкций, связывающих между собой более двух ФП. Например, в случае последовательной ферментативной реакции A→B→C, фермент C скорее будет закодирован в геноме только в том случае, если в геноме одновременно также присутствуют гены, кодирующие оба других фермента. В этом случае ФП соответствующих ферментов можно записать как логическое пересечение: *C = A AND B*. В данной работе впервые использовали аппарат логической регрессии для поиска логических правил связывающих заданный ФП с произвольным числом других ФП (предикторов), что обобщает работу Bowers с соавторами, где использовались только тройки ФП.

Выявление логических закономерностей проводили для 1770 ФП белков *M.tuberculosis H37Rv*. Отбор логических правил был ограничен условиями, чтобы в результирующее правило входило не более четырех ФП-предикторов и не более трех логических операций типа AND и OR. При этом каждый ФП-предиктор мог быть заменен своим логическим отрицанием (операция NOT). Среди найденных

логических закономерностей, 56% составили правила, которые выполнялись для всех элементов ФП без исключений.

Некоторые найденные логические правила были проанализированы с целью определения таких взаимосвязей между белками, которые могут представлять интерес для понимания особенностей метаболизма *M.tuberculosis*. Одним из примеров такого логического анализа ФП, является следующее правило:

```
Rv1819c = (Rv2399c AND Rv1970) AND (Rv1607 OR Rv3137)
```
(3)

Выражение (3) означает, что ген Rv1819c присутствует в геномах референтных микроорганизмов только в том случае, если также одновременно присутствуют гены Rv2399c, Rv1970, и один из генов Rv1607 или Rv3137, или оба этих гена вместе. Эта закономерность выполнялась для 147 референтных геномов из 150.

В БД UniProt Rv1819c охарактеризован как трансмембранный белок, принадлежащий к суперсемейству ABC-транспортеров. Было показано что, несмотря на отсутствие глобальных нарушений в целостности мембраны, инактивирование гена Rv1819c приводит к увеличению резистентности мутантных по этому гену микобактерий к блеомицину и способствует развитию хронической инфекции (Domenech et al., 2009). По мнению авторов указанной работы, эти данные свидетельствуют об участии продукта гена Rv1819c в транспорте молекул, обуславливающих взаимодействие патоген-хозяин.

Белок Rv2399c (ген *cysT*) является сульфат-пермеазой и также принадлежит к суперсемейству ABC-транспортеров. В работе (Sassetti et al., 2003) путем высокопроизводительного мутагенеза с помощью транспозонов показано, что *cysT* необходим для роста и выживания *M.tuberculosis*.

Белок Rv1970 (ген *lprM* или *mce3E*) принадлежит к семейству липопротеинов MCE. Точная роль этого белка пока остается неизвестной, но есть свидетельства в пользу того, что Rv1970 обеспечивает механизм проникновения патогена в клетку-хозяина (El-Shazly et al., 2007).

Белок Rv1607 (ген *chaA*) является Ca^{2+}/H^{+} антипортером.

Белок Rv3137 является инозитол-монофосфатазой – ферментом, участвующим в синтезе фосфатидилинозитола из миоинозитола. Этот процесс играет весьма важную роль, поскольку клеточная стенка *M.tuberculosis* содержит различные гликолипиды, содержащие фосфатидилинозитол. Эти гликолипиды выполняют структурную функцию, а также участвуют во взаимодействиях патоген-хозяин.

Таким образом, анализ функциональных аннотаций ФП-предикторов найденного логического правила показал взаимосвязь трансмембранных транспортеров и биосинтеза фосфатидилинозитола. При этом для трех белков из пяти

показано их участие во взаимодействиях патоген-хозяин, что позволяет предложить их как перспективные мишени для противотуберкулезных препаратов. Найденную взаимосвязь между кальциевым антипортером Rv1607 и инозитол-монофосфатазой Rv3137 объясняет работа (Berggard et al., 2002), в которой показано, что связанный с Ca^{2+} калбиндин обуславливает повышение активности инозитол-монофосфатазы до 250-кратного уровня. Тем самым автоматически полученному логическому правилу может быть дана разумная биологическая интерпретация.

4. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В работе использованы методы сравнения разбиений для оптимизации параметров метода ФП. Одним из таких параметров является набор референтных геномов. Удаление из матрицы ФП групп близкородственных организмов повысило точность работы метода. Это объясняется тем, что наличие гена в одном организме увеличивает вероятность нахождения его гомологов в филогенетически близких организмах. События не являются независимыми, поскольку геномы не сильно изменились в процессе эволюции. Следовательно, совпадение элементов в одинаковых позициях ФП, соответствующих филогенетически близким организмам, отражает не функциональную взаимосвязь между белками, а всего лишь эволюционную близость геномов. Поэтому использование близкородственных организмов, увеличивая размерность матрицы ФП, не вносит дополнительной информации, пригодной для предсказания взаимосвязанных белков, а только повышает уровень шума, ухудшая точность работы метода.

Среди “наилучших” способов кластеризации встречаются методы Уорда, полной связи и PAM. Это указывает на наличие шума в данных об ФП, поскольку общим для всех этих методов является относительная устойчивость (робастность) к сильно разнящимся ФП. Так, для определения расстояния между кластерами метод полной связи использует максимальное расстояние между ФП – наиболее консервативную оценку. Метод Уорда минимизирует дисперсию кластеров и также является робастным методом. Метод PAM в качестве центра кластера использует медоиды – обобщение понятия медианы, что является более робастной оценкой по сравнению с математическим ожиданием (центром тяжести в многомерном случае). Соответственно методы средней и ближней связей, а также метод DIANA не являются оптимальными для кластеризации ФП, будучи относительно неробастными.

Одной из причин появления шума в данных о ФП может являться некорректное определение ортологов генов. Также, возможно, как шум проявляют себя отклонения от относительно простого предположения метода ФП о взаимосвязи ко-

эволюционирующих белков. Такая ситуация вероятна, например, при использовании в качестве референтных организмов, ведущих паразитический образ жизни, что часто сопровождается значительной редукцией геномов, и, следовательно, зашумлением данных об ФП.

Относительно невысокие результаты воспроизведения состава метаболических путей *E.coli K12* (только 20 метаболических путей из 124 восстановлены с точностью превышающей 50%) могут быть объяснены рядом причин. Во-первых, следует учитывать, что каждый белок KEGG может быть аннотирован как принадлежащий одновременно к нескольким метаболическим путям (нечеткое разбиение), в то время как в результате применения использованных в работе методов кластеризации каждый белок относится только к одному кластеру (фиксированное разбиение). Во-вторых, как показано в нашей работе, белки внутри каждого кластера могут быть корректно предсказаны как функционально взаимосвязанные, в то время как они относятся к разным разделам KEGG. Поэтому возможные "нарушения" в соответствии состава кластеров метаболическим путям KEGG могут таковыми и не являться при последующем анализе, а иметь биологическое значение. Также, многие пути KEGG, состав которых был воспроизведен в автоматическом режиме менее чем на 25%, содержат относительно малое количество белков. В 24 разделах KEGG, состав которых плохо воспроизводился при кластеризации ФП, в среднем содержались всего 3,8 белка.

Анализируя результаты, полученные с помощью внутренних индексов, можно отметить удовлетворительное согласие в оценке числа взаимосвязанных групп (96 для средней оценки согласно всем индексам и 118 согласно KEGG). В то же время внутренние индексы не позволяют однозначно указать "наилучшую" комбинацию кластеризация/расстояние (в отличие от внешних индексов). Это можно объяснить тем, что каждый внутренний индекс учитывает различные аспекты построенной кластеризации (компактность кластеров, расстояние между кластерами, соответствие матрице расстояний и т.д.).

Логическая регрессия ФП позволяет находить неочевидные функциональные взаимосвязи между белками, которые не могут быть выявлены при простом попарном сопоставлении ФП (что происходит при кластеризации). Используя этот подход возможно выдвигать гипотезы о функциональной роли белков, ФП которых входят в состав логических правил. В то же время необходимо контролировать параметры метода (максимальный размер модели), а предсказанные взаимосвязи между белками требуют экспериментальной проверки.

5. ВЫВОДЫ

1. Показано, что для оценки соответствия состава предсказанных кластеров взаимосвязанных белков и метаболических путей БД KEGG можно использовать внешние индексы для сравнения разбиений: расстояние между матрицами принадлежности и индекс Рэнда.

2. Наилучшее соответствие состава кластеров белков *E.coli K12* данным БД KEGG достигается при использовании 150 референтных геномов, кластеризации методом PAM и взаимной информации как меры расстояния между ФП. Степень соответствия в большей степени зависит от меры расстояния, чем от способа кластеризации.

3. Внутренние индексы для оценки кластеризаций можно использовать для предсказания количества групп взаимосвязанных белков, опираясь исключительно на данные о ФП. Кластеризация ФП на оцененное количество групп, позволяет предсказывать физические и функциональные взаимосвязи между белками *M.tuberculosis*.

4. Метод логической регрессии позволяет выявлять взаимосвязи между белками, которые отличаются от результатов кластеризации ФП. Для белков *M.tuberculosis* выявили 991 логическую взаимосвязь, выполнявшуюся для всех референтных геномов.

6. СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Pyatnitskiy M.A., Lisitsa A.V., Archakov A.I. Comparative Analysis of Methods for Clustering Phylogenetic Profiles // Сборник трудов конференции HUPO – 2006 – С.581
2. Пятницкий М.А., Курганский А.Д., Лисица А.В., Арчаков А.И. Приложение методов распознавания образов для поиска антимикробных препаратов // Материалы международной школы-конференции молодых ученых «Системная биология и биоинженерия», МАКС Пресс – Москва – 2006 – С.84.
3. Pyatnitskiy M.A., Lisitsa A.V., Archakov A.I. Prediction of Protein-Protein Interactions: Phylogenetic Profiles and Cluster Analysis // Материалы конференции CMTPI – 2007 – С.137.
4. Pyatnitskiy M.A., Lisitsa A.V., Archakov A.I. Clustering Functionally Related Proteins In Prokaryotes Using Graph Representation of High Throughput Data // Сборник трудов конференции HUPO – 2007 – С.118.
5. Moshkovskii S.A., Vlasova M.A., Pyatnitskiy M.A., Tikhonova O.V., Safarova M.R., Makarov O.V., Archakov A. I. Acute phase serum amyloid A in ovarian cancer as an important component of proteome diagnostic profiling // PROTEOMICS - Clinical Applications – 2007 – № 1(1) – С.107-117.
6. Pyatnitskiy M.A., Lisitsa A.V., Archakov A.I. Prediction of Functionally Related Proteins: Phylogenetic Profiles and Cluster Analysis // Сборник материалов международной конференции GPBNM – 2008 – С.38.
7. Пятницкий М.А., Лисица А.В., Арчаков А.И. Предсказание взаимосвязанных белков: филогенетические профили и кластерный анализ // Сборник трудов конференции "Человек и лекарство" – 2008 – С.412.
8. Pyatnitskiy M.A., Lisitsa A.V., Archakov A.I. Prediction of Functionally Related Proteins: Phylogenetic Profiles and Cluster Analysis // Сборник материалов международной конференции BGRS – 2008 – С.199.
9. Пятницкий М.А., Лисица А.В., Арчаков А.И. Предсказание взаимосвязанных белков методами сравнительной геномики in silico //Биомедицинская химия – 2009 – т.55(3) – С.230-246.
10. Пятницкий М.А., Лисица А.В., Арчаков А.И. Сравнение алгоритмов предсказаний взаимосвязанных белков на примере метода филогенетических профилей // Биомедицинская химия – 2009 – т.55(5) – С.534-538.
11. Пятницкий М.А., Лисица А.В., Арчаков А.И. Оптимизация метода филогенетических профилей для предсказания взаимосвязанных белков // Сборник трудов конференции "Математика. Компьютер. Образование" – 2009 – С.65.
12. Pyatnitskiy M.A., Lisitsa A.V., Archakov A.I. Cluster Analysis of Phylogenetic Profiles // Сборник материалов международной конференции MCCMB – 2009 – С.300-301.