На правах рукописи

ПОНОМАРЕНКО Елена Александровна

АВТОМАТИЧЕСКИЙ АНАЛИЗ НАУЧНЫХ ТЕКСТОВ ДЛЯ СОЗДАНИЯ СЕМАНТИЧЕСКИХ СЕТЕЙ БЕЛКОВ

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 им. В.Н.Ореховича РАМН

Научный руководитель: доктор биологических наук
Лисица Андрей Валерьевич

Официальные оппоненты: доктор биологических наук, профессор
Каменская Марина Александровна

доктор биологических наук
Веселовский Александр Владимирович

Ведущая организация: Учреждение Российской академии наук Институт молекулярной биологии им. В.А.Энгельгардта РАН

Защита состоится «14» мая 2009 года в 13:00 часов на заседании Диссертационного совета Д 001.010.01 при Учреждении Российской академии медицинских наук Научно-исследовательском институте биомедицинской химии им. 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 «____» апреля 2009 г.

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

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

1.1. Актуальность проблемы

Применение современных высокопроизводительных подходов к исследованию живых систем позволяет выдвигать предположения о молекулярных взаимосвязях, лежащих в основе исследуемых биологических процессов. Эти предположения должны подвергаться критической оценке в свете данных, опубликованных в научной литературе. Ознакомление с соответствующими литературными источниками при анализе результатов высокопроизводительных экспериментов занимает длительное время и не всегда обеспечивает полноту анализа. Это обуславливает необходимость создания средств семантического анализа для повышения эффективности обработки результатов высокопроизводительных транскриптомных (Beissbarth T., 2006) и протеомных экспериментов (UniProt Consortium, 2009).

В настоящее время автоматическая интерпретация результатов высокопроизводительных экспериментов проводится в основном с использованием общедоступных баз данных и баз знаний, таких, как UniProt (Burgoon L.D., Zacharewski T.R., 2008), энциклопедия метаболических путей KEGG (Kanehisa M. *et al.*, 2008) или система онтологий генов Gene Ontology (GO, Ashburner M. *et al.*, 2000). В основе онтологии генов GO лежит семантическая сеть – формализованное описание объектов и явлений молекулярной биологии в виде ориентированного графа. Вершинами графа являются объекты предметной области (гены или белки), а ребра задают отношения между ними. В системе GO для обозначения объектов поддерживается контролируемый словарь молекулярно-биологических терминов. С использованием словаря формируются аннотации генов, и, если гены выполняют сходную функцию или участвуют в одном биологическом процессе, то соответствующие им аннотации содержат одинаковые термины (Beissbarth T., 2006).

Повсеместное применение системы GO для интерпретации результатов транскриптомных и протеомных экспериментов привело к осознанию ее недостатков (Zheng B., Lu X., 2007). Во-первых, белкам в составе одного метаболического или регуляторного пути зачастую присваиваются разные аннотации, что затрудняет их использование для автоматической обработки данных. Во-вторых, анализ только аннотаций не всегда позволяет выявить биологический смысл анализируемого явления. В связи с этим, авторы ряда работ предлагают проводить семантический анализ функциональных взаимосвязей генов и белков, напрямую обращаясь к публикациям (Ананько Е.А. с соавт., 2000; Homayouni R., *et al.*, 2005; Bundschus M. *et al.*, 2008).

Так, в работе (Homayouni R., *et al.*, 2005) с применением семантического индексирования рефератов MEDLINE проводили кластерный анализ генов для

аннотирования генома человека. Bundschus и соавторы (Bundschus M. *et al.*, 2008) предложили метод автоматического распознавания наименований заболеваний в текстах статей и определили таким образом ассоциативные связи между 4939 генами и 1745 нозологическими формами. В работе (Raychaudhuri S., Altman R.B., 2003) семантическую метрику применяли для идентификации функциональных кластеров генов, при этом чувствительность предложенной метрики при сравнения с данными системы GO составила 96%. Этот подход получил развитие в работе (Zheng B., Lu X., 2007), где тематическая декомпозиция научных статей позволила получить графы, узлами которых являются не только белки, но и понятия молекулярной биологии – например, апоптоз. Наряду с описанием новых вычислительных подходов в вышеуказанных работах отмечается важность автоматического распознавания в текстах анализируемых документов специальных терминов, в том числе названий белков (Jenssen T.K. *et al.*, 2001).

Постоянное увеличение количества научных статей в области биомедицины все больше усложняет поиск необходимой исследователю информации (Stapley B., Benoit G., 2000). Сложности обработки такого рода данных без использования автоматизации особенно очевидны, если речь идет об анализе информации о функциях белков и генов, идентифицированных в результате высокопроизводительных экспериментов.

В данной работе рассматривается методика сопоставления результатов высокопроизводительных протеомных экспериментов с информацией, представленной в виде множества рефератов научных публикаций в базе MEDLINE. В работе используются как публикации, найденные контекстным поиском по названию белка (релевантные), так и наиболее близкие им по смыслу (родственные). Предлагаемая методика основана на оценке семантической связности между белками, которая рассчитывается как функция от числа одинаковых релевантных или родственных публикаций, найденных для двух белков. Вычисленные значения семантической связности заносили в матрицу семантического сходства, которая отображается в виде неориентированного графа. Полученные в составе семантического графа изолированные подграфы сопоставляли с распределением белков по разделам базы данных KEGG и по категориям системы GO.

Целью работы являлась разработка метода представления информации о взаимосвязях между белками в виде семантической сети, построенной на основе автоматического анализа научных текстов. Для достижения цели решались **задачи**:

1. для каждого из белков выборки, состоящей из 5-ти произвольно отобранных метаболических путей, сформировать специфичный семантический профиль релевантных публикаций;

2. дополнить полученные профили родственными публикациями, найденными в результате автоматической оценки смыслового сходства документов;

3. рассчитать меру семантической связности между белками как функцию пересечения множеств публикаций, входящих в состав релевантных и родственных профилей. На основе рассчитанной меры семантической связности построить семантическую сеть, отражающую белок-белковые взаимосвязи;

4. выделить в полученной семантической сети изолированные подграфы и сравнить их с распределением белков по разделам базы данных метаболических путей KEGG и по категориям онтологии генов GO.

1.2. Научная новизна и практическая значимость

Новизна данной работы по сравнению с аналогичными подходами (Raychaudhuri S. *et al.*, 2002; Plake C. *et al.*, 2006) заключается в том, что мера семантической связности между белками определяется на основе смысловой близости относящихся к белкам документов. Для смыслового сравнения документов применяется алгоритм поиска родственных публикаций, представленный в библиографической системе PubMed [http://www.ncbi.nlm.nih.gov/pubmed/]. Применение поисковых запросов обеспечивает возможность динамической актуализации семантической сети белков по мере увеличения количества публикаций, депонируемых в базе данных MEDLINE. Впервые показано, что рассчитываемую системой PubMed оценку смыслового сходства документов можно использовать для автоматизированного выявления взаимосвязей между белками и конструировать семантические сети, подграфы которых совпадают с разделами базы данных KEGG и с категориями системы GO.

Практическое применение разработанного подхода в научных исследованиях обусловлено интуитивно понятной схемой его работы. Поиск в базе данных MELDINE информации об идентифицированных белках в настоящее время широко используется для интерпретации результатов высокопроизводительных экспериментов в области протеомики. Разработанный подход позволяет автоматизировать поиск релевантных публикаций и существенно ускорить получение обобщенного представления о распределении сотен идентифицируемых в ходе протеомного эксперимента белков по изученным биологическим процессам.

Для иллюстрации возможностей практического применения предлагаемого подхода были установлены взаимосвязи между белками, упомянутыми в журнале Nature. Для 260 таких белков с использованием ресурса PubMed отобрали родственные по смыслу публикации, вышедшие в разных журналах. Содержащуюся в этих публикациях информацию обобщили по принципу: чем больше родственных

публикаций найденно в системе PubMed для двух белков – тем выше степень семантической взаимосвязи между ними. Получили семантическую сеть, отображающую основные белки, которые обсуждались в Nature за последние два года. В составе полученных изолированных подграфов семантической сети были выявлены белки, связанные с развитием онкологических и аутоиммунных заболеваний, а также белки апоптоза. Таким образом, за 3 часа компьютер обработал в сумме более 62 тыс. публикаций из разных журналов и сгенерировал дайджест журнала Nature, выявив приоритеты современной молекулярной биологии.

1.3. Апробация работы

Основные положения диссертационной работы докладывались и обсуждались в ходе следующих конференций: “Информационно-вычислительные технологии в науке, Москва, 2005”, “XIII Российский национальный конгресс «Человек и Лекарство», Москва, 2006”, “XV Российский национальный конгресс «Человек и Лекарство», Москва, 2008”, “Международный конгресс «Протеом человека», Сеул, 2007”; “Международный конгресс «Протеом человека», Амстердам, 2008”, “Международный конгресс «Геномика, протеомика, биоинформатика и нанобиотехнологии для медицины», Москва, 2008”.

1.4. Основные положения, выносимые на защиту

1. Содержащиеся в базе данных UniProt наименования белковых молекул могут использоваться для автоматического поиска релевантных для заданного белка статей в системе PubMed/MEDLINE.

2. Родственные публикации, предоставляемые системой PubMed согласно критерию смыслового сходства документов, содержат дополнительную информацию о взаимосвязях между белками, которая отсутствует в релевантных публикациях.

3. Для выборки, включающей 150 белков из 5-ти различных метаболических путей, построена семантическая сеть, узлы которой являются белками, а ребра отражают меру семантической связности, пропорциональную мощности пересечения множеств ассоциированных с белками публикаций.

4. Подграфы в составе семантической сети согласуются с распределением белков по разделам базы данных метаболических путей KEGG и по категориям онтологии генов GO.

1.5. Публикации

Материалы диссертационной работы отражены в 14 публикациях: в 4 статьях и 10 публикациях в сборниках докладов научных конференций.

1.6. Объем и структура диссертации

Диссертация изложена на 105 страницах машинописного текста, включая 8 таблиц и 16 рисунков. Состоит из глав: «Введение», «Обзор литературы», «Материалы и методы», «Результаты и обсуждение», «Заключение», «Выводы», «Список литературы».

2. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Реферативная база данных. В работе использовали реферативную базу данных MEDLINE. Доступ к рефератам осуществляли через поисковый сервер PubMed.

Поиск контекстной строки *t* проводили путем направления к серверу PubMed запроса Q(*t*) по протоколу HTTP: «entrez?db=pubmed& cmd=search&term='*t*'». Различия между строчными и прописными буквами в поисковой строке не учитывали. Каждой отвечающей запросу Q(*t*) библиографической записи MEDLINE сопоставляли *релевантный* идентификатор PubMed, далее обозначаемый *pmid*. Для релевантного идентификатора *pmid* из поля «Related Links» загружали множество сходных по тематической направленности (*родственных*) публикаций: Q(*pmid*) = *{pmid'}*.

Выборку белков **M** формировали на основе базы данных KEGG (Kanehisa M. *et al.*, 2008). В состав выборки включали белки, относящиеся к метаболическим путям из разных разделов этой базы данных. Выбор разделов осуществляли случайным образом, при этом единственным критерием служило наличие более 15 белков в выбранном метаболическом пути. В итоге отобрали 150 белков, включая 44 белка, относящихся к метаболизму жирных кислот, 32 белка – к метаболизму аргинина и пролина, 18 белков – к биосинтезу гепаран сульфата, 35 белков, участвующих в репликации ДНК и 21 белок метаболизма азота. Коды доступа белков, присвоенные в базе данных UniProt (далее используется обозначение «*un*») использовали в качестве идентификаторов. Для белка с кодом доступа *un* определяли множество $\{pn\}_{un}$ альтернативных названий.

Мера семантического сходства между белками. Каждый белок из выборки **M** характеризовали с помощью семантического профиля, представляющего собой множество идентификаторов релевантных публикаций $\mathbf{P_{un}}$. Релевантные публикации определяли путем направления к серверу PubMed запроса Q(*t*), где в качестве поискового термина *t* указывали обозначения белка из множества $\{pn\}_{un}$:

$$P_{un} = Q(\{pn\}_{un}) = Q('pn_1 \text{ OR } pn_2 ... \text{OR } pn_j') = \{pmid\}_{un} \quad (1)$$

В результате пересечения множества релевантных публикаций, найденных для белка *a* – $\mathbf{P_a}$, и множества релевантных публикаций, найденных для белка *b* – $\mathbf{P_b}$ получали множество $\mathbf{P_{ab}}$ совпадающих публикаций для двух белков *a* и *b*:

$$P_{ab} = P_a \cap P_b \quad (2)$$

Каждому белку из выборки **M** с использованием выражения (1) сопоставляли множество идентификаторов родственных публикаций $\mathbf{R_{un}}$:

$$R_{un} = Q(\{pmid\}_{un}) = \left\{Q(pmid_1), Q(pmid_2), Q(pmid_j)\right\} = \{pmid'\}_{un} \quad (3)$$

При объединении множества $\mathbf{P_{ab}}$, содержащего идентификаторы одинаковых публикаций для двух белков *a* и *b*, с множеством родственных публикаций $\mathbf{R(P_{ab})}$, содержащим близкие по смыслу публикации к публикациям из множества $\mathbf{P_{ab}}$, получали множество **P**:

$$P = P_{ab} \cup R(P_{ab}) \quad (4)$$

Для белка с кодом *un* из выборки **M** семантический профиль $\mathbf{S_{un}}$ для случая родственных публикаций формировали следующим образом: из множества всех родственных публикаций для этого белка $\mathbf{R_{uc}}$ исключали множество публикаций **P**, включающих названия одновременно двух белков из **M**, а также родственные к таким публикациям статьи $\mathbf{R(P_{ab})}$:

$$S_{un} = R_{un} - P \quad (5)$$

Аналогично выражению (2), множеством $\mathbf{S_{ab}}$ обозначали пересечение семантических профилей, построенных для двух белков *a* и *b* с использованием родственных публикаций:

$$S_{ab} = S_a \cap S_b \quad (6)$$

Меру семантического сходства *T(a,b)* между двумя белками *a* и *b* вводили с использованием нормировки Танимото (Rogers D.J., Tanimoto T.T., 1960) исходя из мощностей множеств $\mathbf{P_{ab}}$, $\mathbf{P_a}$ и $\mathbf{P_b}$ (или $\mathbf{S_{ab}}$, $\mathbf{S_a}$ и $\mathbf{S_b}$):

$$T(a,b) = \frac{|P_{ab}|}{|P_a| + |P_b| - |P_{ab}|};\ T'(a,b) = \frac{|S_{ab}|}{|S_a| + |S_b| - |S_{ab}|} \quad (7)$$

Построение и анализ семантической сети. Значения семантического сходства *T*(*i*,*j*) рассчитывали для каждой пары белков (*i*,*j*) из выборки **M** согласно формуле (7). При заданном пороговом значении сходства *l* определяли элементы матрицы смежности *L*, в которой единица обозначала наличие ребра, соединяющего соответствующие белки в графе, ноль – его отсутствие:

$$L_{ij} = \begin{cases} 1, T(i,j) > \ell \\ 0, T(i,j) \le \ell \end{cases} \quad (8)$$

Матрицу смежности визуализировали в виде неориентированного невзвешенного графа с использованием программы Gvedit [http://www.graphviz.org/]. На графе определяли изолированные подграфы, причем подграфы, состоящие из единственной вершины, исключали из рассмотрения. Пороговое значение семантического сходства *l* подбирали так, чтобы среднее число вершин в одном подграфе и количество подграфов было максимально. Для подграфов с количеством вершин $n>5$ вероятность случайного вхождения в состав подграфа x белков, относящихся к одному метаболическому пути, рассчитывали на основе гипергеометрического распределения (Zheng B., Lu X., 2007):

$$p(x \mid M,K,n) = \frac{C_x^K C_{n-x}^{M-K}}{C_n^M}, \quad (9)$$

где M – общее количество белков в выборке, а K – количество белков, относящихся к данному метаболическому пути.

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

3.1. Контекстный поиск названий белков в рефератах статей

Задача идентификации обозначений белков в текстах рефератов научных публикаций была решена с использованием информации из номенклатурного подраздела белковой базы UniProt (UniProt Consortium, 2009). Из базы данных UniProt загружали рекомендуемые кураторами ресурса обозначения белков, собранные в полях «Alternative names» и «Synonyms». Обозначения белков включали в состав поискового запроса с использованием логического оператора «ИЛИ» и сформированный запрос направляли в PubMed. Полученное в результате обработки запроса множество идентификаторов библиографических записей MEDLINE рассматривали как семантический профиль белка, составленный по релевантным публикациям.

Специфичность семантического профиля по отношению к определенному белку определяется тем, насколько указанные в ресурсе UniProt обозначения белков совпадают с наименованиями белков, используемыми авторами статей. В результате автоматической обработки поисковых запросов возможно появление артефактов, поскольку используемая для обозначения белка аббревиатура может совпадать с общеупотребительными или специальными терминами, не являющимися названиями биомакромолекул. Например, одним из сокращенных названий ингибитора транскрипционного фактора NF-kappa-B (код Q9BYH8) является слово «MAIL»

(Molecule possessing Ankyrin repeats Induced by Lipopolysaccharide), которое встречается в большинстве рефератов при указании адреса для переписки.

Достоверность автоматического определения названий белков в текстах рефератов оценивали путем сравнения результатов поиска в системе PubMed с опубликованными в литературе данными о частоте употребления названий белков в статьях по протеомике (Petrak J. *et al.*, 2008). Из базы данных UniProt загружали коды доступа для 11,5 тыс. записей с меткой «evidence at protein level» (метка обозначает, что экспрессия соответствующих белков установлена экспериментальными методами). Практически все отобранные записи содержали несколько названий для одного белка. Только в 6 случаях из 11,5 тыс. записей указано единственное обозначение белка; для 7 белков известно более 20 синонимов, а у одного белка – интерлейкина-8 – существует 42 альтернативных названия. Для большинства (61%) белков в базе данных UniProt указано не более 3-х обозначений. Суммарное количество обозначений белков во всех загруженных из UniProt записях составило величину порядка 30 тыс. терминов, включая аббревиатуры.

В работе (Petrak J. *et al.*, 2008) был опубликован список TOP15, который в данной работе используется для оценки качества автоматического определения названий белков. В список TOP15 вошли белки, встречающиеся с высокой частотой в 99 научных статьях в области протеомики за период 2004-2006 г.г. В текстах этих статей был произведен контекстный поиск обозначений белков, загруженных из базы данных UniProt. Для каждого белка подсчитывали количество публикаций, в которых было найдено как минимум одно текстуальное совпадение с соответствующими данному белку альтернативными или синонимичными обозначениями. Результаты подсчета, выраженные как частота встречаемости белка в выборке из 99 публикаций, приведены в таблице 1 в сравнении со значениями, ранее полученными аналогичным образом для выборки TOP15. Специфичность распознавания названий белков оценивали путем экспертного анализа контекста, в котором упомянут найденный в автоматическом режиме термин.

В таблицу 1 вошли названия белков, упомянутые с частотой встречаемости выше 5% в статьях по протеомике. При их сопоставлении со списком TOP15 видно, что 12 белков (80%) совпадают. Одно из разночтений несущественно: в таблице 1 объединены пероксиредоксины 1 и 2, а в работе (Petrak J. *et al.*, 2008) эти белки учтены раздельно. Два других, не выявленных автоматической процедурой белка, – пируват-киназа и ингибитор диссоциации ГДФ. Разночтения связаны с несовпадением обозначений белков, используемых в записях UniProt и текстах статей. Например, в записи UniProt с кодом P52565 указано единственное обозначение «RHO GDP-DISSOCIATION

INHIBITOR», при этом в статьях этот белок упоминается в форме сокращений: «RHO-GDI», «RHO GDI» или «GDIR».

В таблице 1 есть 3 белка, которые отсутствуют в списке сравнения TOP15: белок теплового шока 60 кДа, глицеральдегид-3-фосфат дегидрогеназа и сывороточный альбумин (отмечены в таблице курсивом). Экспертный анализ контекста показал, что выявление дополнительных белков в результате автоматического поиска связано с тем, что список TOP15 составляли только по приведенным в статьях таблицам, а автоматический поиск проводили по полным текстам статей.

Таблица 1. Список белков, найденных более чем в 5% статей журнала *Proteomics* за 2004-2006 г.г. Колонки «А» и «TOP15» содержат результаты автоматического и экспертного поиска, соответственно. Курсивом отмечены белки, выявленные только в результате автоматического поиска, но отсутствующие в TOP15.

№	Код доступа в UniProt	Название белка	Частота, %	
			А	TOP15
1	P06733	Альфа-енолаза	16	31
2	P60174	Триозофосфатизомераза[a)]	26	22
3	P04083, P07355, P09525, P08758	Аннексины (A1, A2, A4, A5)	36	19
4	Q06830, P32119	Пероксиредоксины[a)] (пероксиредоксин 1, пероксиредоксин 2)	42	21
5	P10809	*Белок теплового шока 60 кДа*[a)]	18	-
6	P04406	*Глицероальдегид-3-фосфат дегидрогеназа*	12	-
7	P62937	Пептидилпролил-цис-транс-изомераза А (циклофилин А)[a)]	10	17
8	P08670	Виментин[a)]	23	20
9	P02768	*Сывороточный альбумин*	13	-
10	P24539	АТФ-синтаза, F-тип, субъединица B[a)]	12	15
11	P07339	Катепсин D[a)]	21	16
12	P11021	GRP78	11	14
13	P11142	HSC70	15	18
14	P04792	Белок теплового шока бета-1[a)] (HSP27)	21	34
15	P05787	Кератин, тип 2[a)]	14	17

[a)] Название белка указано в рефератах публикаций.

Сопоставление с опубликованными данными показало, что сведения базы данных UniProt в целом пригодны для поиска релевантных статей по названиям белков. Контекстным поиском для 150 отобранных из разделов KEGG белков было найдено 65,9

тыс. релевантных рефератов. Идентификаторы рефератов вошли в состав семантических профилей каждого белка. В среднем профиль содержал 445 идентификаторов релевантных статей, причем для >40% белков в семантический профиль входило менее 100 статей.

3.2. Родственные публикации и матрица семантического сходства

Родственные публикации были получены в результате расчета системой PubMed критерия смыслового сходства документов. Для определения смыслового сходства в системе PubMed применяется методика, основанная на сравнении относительных частот встречаемости слов в сопоставляемых текстах. После выполнения поискового запроса по идентификатору *pmid,* сервер PubMed отображает Веб-страницу, где под заголовком «Related Links» выведено пять гипертекстовых ссылок на родственные публикации. Для них извлекали библиографические идентификаторы *pmid'*. Таким образом, для каждого белка было получено два семантических профиля: один по релевантным публикациям и один – по родственным.

Мера семантической связности отражала количество одинаковых публикаций в семантических профилях двух белков. Попарным пересечением профилей релевантных публикаций было получено множество идентификаторов рефератов, в текстах которых совместно встречаются названия, по крайней мере, двух белков из выборки. Это множество состояло из 9838 элементов и обозначалось $\mathbf{P_{ab}}$ согласно выражению (2). При создании семантического профиля по родственным публикациям учитывали условие (4), в результате чего в профиль включали только те идентификаторы, которые не вошли во множество **P**. При таком условии для 65,9 тыс. релевантных рефератов из системы PubMed было отобрано 196,7 тыс. идентификаторов родственных статей. В среднем семантический профиль по родственным рефератам содержал 1525 идентификаторов *pmid'*, при этом для 40% белков в составе профиля насчитывалось менее 400 статей.

Расчет меры семантической связности был произведен для каждой пары белков с использованием отдельно профилей по релевантным и по родственным публикациям. Полученные две матрицы попарного семантического сходства были симметричными относительно диагонали и содержали (148*147)/2 = 10'878 элементов[1], соответствующих парам белков. Из этого количества значений для случая релевантных публикаций 1147 ячеек матрицы имели ненулевые значения. Таким образом, для ~10% пар белков нашлась как минимум одна публикация, в реферате которой встретились названия обоих белков. Для родственных публикаций доля заполненных ненулевыми значениями ячеек

[1] – матрицу рассчитывали без учета диагонали для уникальных 148 белков (два белка входили одновременно в состав двух метаболических путей).

матрицы семантического сходства оказалась выше более чем в три раза и составила ~34%.

Использование родственных публикаций для расчета семантического сходства позволило установить 72% взаимосвязей, которые в явном виде содержались в релевантных публикациях. Кроме того, обработка родственных публикаций выявила дополнительные сведения о семантической связности между 2829 парами белков (~25% от общего количества пар), для которых информация о взаимосвязях отсутствовала в рефератах релевантных публикаций. Например, белки P49189 (альдегид дегидрогеназа) и A8YXX4 **(**глутамин синтаза) совместно не встречаются ни в одной публикации MEDLINE, однако, в родственных профилях этих белков обнаружили 5 одинаковых статей.

Для пар белков с наиболее высокими значениями семантического сходства был проведен выборочный анализ родственных публикаций. Оказалось, что семантическая связность белков с кодами O15460 и P13674 обусловлена тем, что они функционируют как субъединицы в составе комплекса пролил-4-гидроксилазы (Annunen P. *et al.,* 1997). Два других белка (Q7LGA3 и O14792) связаны, поскольку оба являются ферментами-сульфотрансферазами. Один из них (Q7LGA3) катализирует перенос сульфатной группы с 3`- фосфоаденозин-5`-фосфосульфата в 2-OH позицию гексуроновой кислоты (Xu D. *et al.,* 2007), а другой (O14792) переносит сульфатную группу в 3-OH позицию при биосинтезе гепаран сульфата (Chen J. *et al.,* 2003).

Из приведенных примеров видно, что сравнение профилей родственных рефератов позволяет выявить скрытое семантическое сходство, отражающее взаимодействие белков либо как субъединиц в составе функционального комплекса, либо взаимосвязь белков в цепочке биохимического синтеза.

Доля заполненных ячеек в матрице семантического сходства пропорциональна количеству белковых пар, для которых пересечение семантических профилей содержит хотя бы один элемент. По такому характеристическому показателю было проведено сравнение двух типов выборок белков из базы данных KEGG. В первом случае в выборку случайным образом включали белки, относящиеся к одному метаболическому пути, а во втором – к разным метаболическим путям. Для каждой выборки на основе сравнения профилей родственных и релевантных публикаций рассчитывали число взаимосвязанных белков. Результаты сравнения 50 выборок каждого типа представлены на рисунке 1.

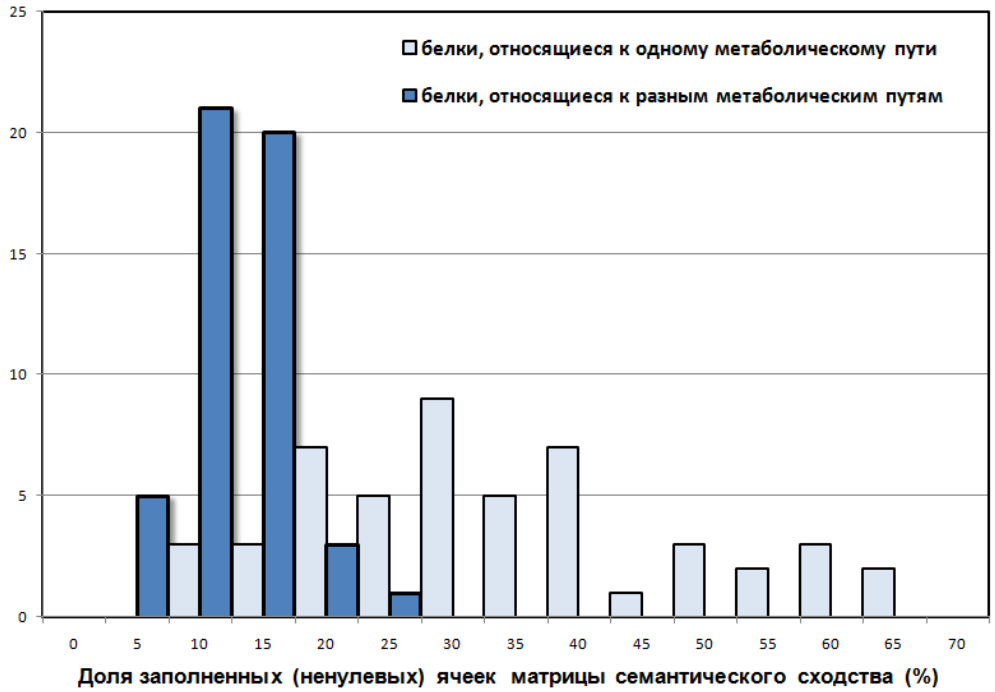

Рисунок 1. Гистограммы уровня заполненности матриц семантического сходства (%), полученные для двух типов выборок белков из базы данных KEGG при пересечении профилей релевантных публикаций.

Из рисунка 1 видно, что для метаболически-связанных белков уровень заполнения матрицы семантического сходства значениями выше, чем для белков, входящих в различные метаболические пути. Пик распределения для последних приходится на 10%, то есть для большинства произвольно сформированных выборок белков заполнено только 10% матрицы сходства. В то же время для белков, отнесенных в базе данных KEGG к одному пути, уровень заполненности матрицы сходства достигает в некоторых случаях 65%. Полученные значения вполне согласуются с результатами, представленными на стр.10: для выборки, состоящей из белков 5-ти различных метаболических путей число заполненных ячеек в матрице семантического сходства также составляет около 10%. Это свидетельствует о том, что семантическая связность между белками внутри одного метаболического пути существенно выше семантической связности между белками, принадлежащими разным метаболическим путям.

3.3. Семантические сети белков

Для графического отображения связей между белками матрицу семантического сходства преобразовали в матрицу смежности, заменяя по формуле (8) все значения ниже порога l на нули, а выше или равные – на единицы. Матрица смежности была визуализирована в виде неориентированного графа с невзвешенными ребрами в программе Gvedit.

На рисунке 2 показана семантическая сеть, построенная для белков пяти метаболических путей человека с использованием релевантных публикаций. В составе сети можно выделить 10 изолированных подграфов, из которых 4 подграфа, обозначенные на рисунке буквами, содержат 8 и более вершин. Такое распределение вершин по крупным изолированным подграфам наблюдалось при пороговом значении семантического сходства l = 0,0045.

Состав показанных на рисунке 2 подграфов «А», «Б» и «Г» согласуется с предлагаемым в базе данных KEGG распределением белков по метаболическим путям. В подграфе «А» – 10 вершин и все они соответствуют карбоангидразам, отнесенным согласно KEGG к метаболизму азот-содержащих соединений. Другая часть этого метаболического пути, представленная 3 белками, вошла в состав подграфа «В». Эти три белка принимают участие в метаболизме аммиака, причем два из них (коды P00367 и P31327 на рис. 2) относятся не только к метаболизму азот-содержащих соединений, но и к биосинтезу аргинина и пролина. Белок с кодом P31327 является карбамоил-фосфатсинтетазой (КФ 6.3.4.16) и представляет собой связующее звено между метаболизмом азот-содержащих соединений и циклом мочевины, который согласно номенклатуре KEGG, отнесен к разделу «метаболизм аминокислот».

Подграф «Б» содержит 29 белков, принимающих участие в репликации ДНК. В данном случае белки образуют функциональные комплексы за счет физического взаимодействия друг с другом, а не являются сопряженными звеньями метаболической цепочки. Механизм репликации ДНК эукариот опубликован в деталях, поэтому на семантической сети нашли отражения все его элементы: структурные компоненты комплекса ДНК-полимеразы (коды P09884 и Q14181) и ДНК-праймазы (коды P49643 и P49642), белки репликационных комплексов (коды A4D2J4 и P56282, соответственно), белки в составе пререпликационного геликазного комплекса MCM (см. рис. 2).

Основная часть подграфа «Б» представлена белковыми факторами репликации (коды Q13156, A8KY9, A4D105), в том числе факторами репликации C («clamp loader», коды P35250, P40938, P35251). ДНК лигаза и ядерный антиген пролиферирующих клеток (коды B2RAI8 и B2R897, соответственно) находятся на периферии подграфа.

Подграф «В» смешанный: в него вошли белки из разных разделов базы данных KEGG. Всего в составе подграфа 55 белков, часть которых относится к метаболизму жирных кислот, а другая часть – к метаболизму аргинина и пролина. Белки из раздела 01103 «метаболизм липидов» базы данных KEGG распределились между двумя областями, обозначенными «В2» и «В3» на рисунке 2. В области «В2» находится 14 белков: 6 алкогольдегидрогеназ и 8 альдегиддегидрогеназ. Область «В3» представлена 26 белками, в числе которых ферменты, участвующие в деградации жирных кислот, –

ацетил-коА-ацетилтрансферазы (B2R6H1 и P42765), еноил-коА-гидратазы (Q58EZ5, P30084), ацил-коА дегидрогеназы и гидроксиацил-коА-дегидрогеназы.

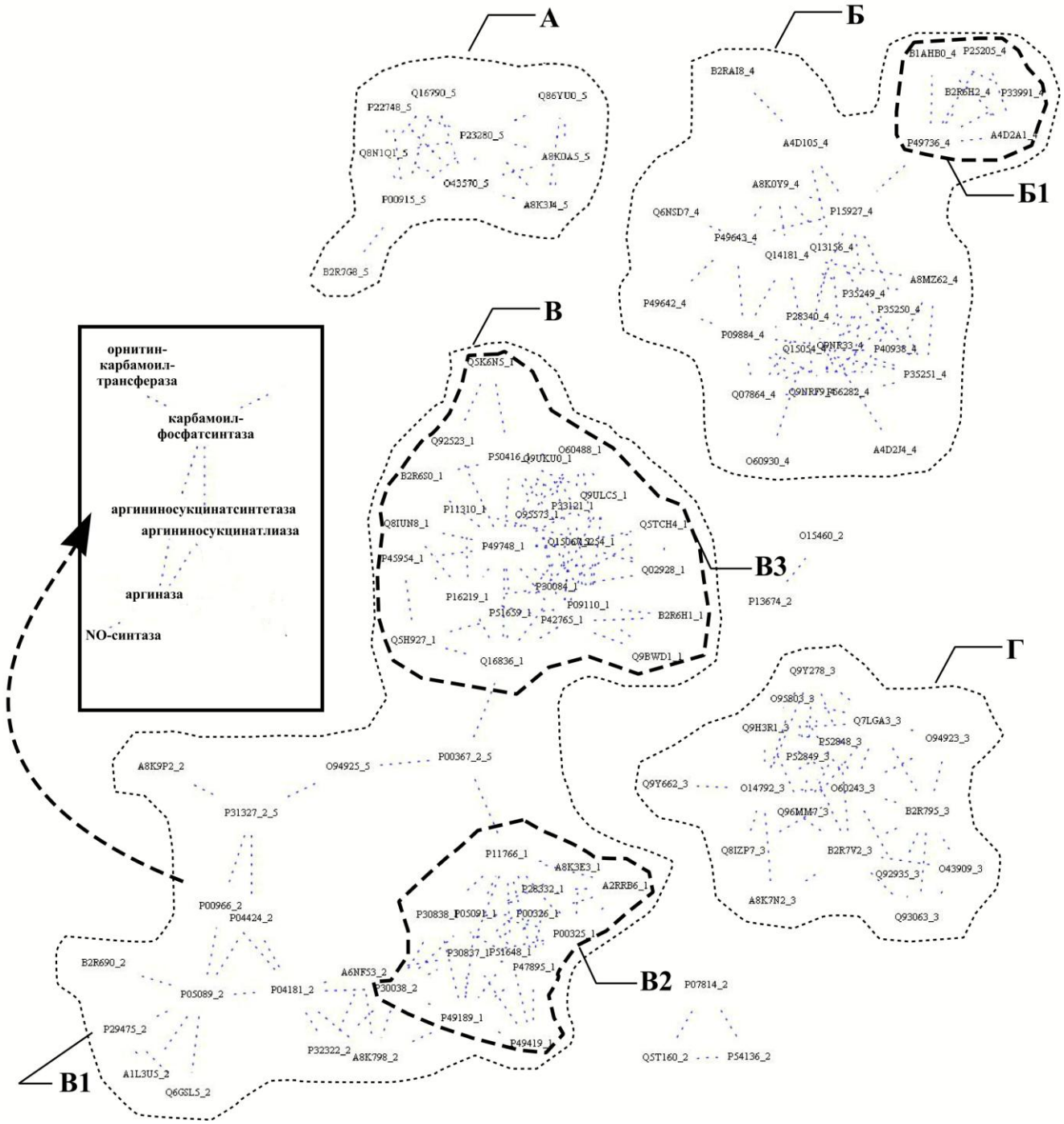

Рисунок 2. Семантическая сеть белков пяти метаболических путей, полученная с использованием оценки семантического сходства по релевантным публикациям. Вершины обозначены «xxxxx_y», где «xxxxx» – код доступа белка в базе данных UniProt, «y» – номер метаболического пути: 1 – метаболизм жирных кислот, 2 – метаболизм аргинина и пролина, 3 – биосинтез гепарансульфата, 4 – репликация ДНК и 5 – метаболизм азот-содержащих соединений. «Б1» – геликазный комплекс MCM; «В1» – цикл мочевины (см. врезку); «В2» и «В3» – белки метаболизма липидов.

В подграф «В» вошли все ферменты цикла мочевины, включая аргининосукцинатсинтетазу (код P00966), аргининосукцинатлиазу (P04424), аргиназу (P05089), орнитинтранскарбамоилазу (A8K9P2). Также в состав этого подграфа вошла NO-синтаза (P29475), катализирующая в цикле мочевины превращение аргинина в цитруллин с образованием NO (Husson A. *et al.*, 2003).

Подграф «Г» представлен 18 ферментами биосинтеза гепаран сульфата – гликозаминогликана, по структуре сходного с гепарином. Биосинтез гепарансульфата осуществляют различные виды гликозилтрансфераз, сульфотрансфераз и эпимераз (Nadanaka S., Kitagawa H., 2008). Представители соответствующих подклассов ферментов наблюдаются в составе подграфа «Г»: коды O43909 и Q93063 – N-ацетилглюкозаминил трансферазы (КФ 2.4.1), коды O95803 и Q9H3R1 – N-деацетилаза и N-сульфотрансфераза, код Q9Y278 – глюкозамин сульфотрансфераза (КФ 2.8.2.29).

На рисунке 3 приведен семантический граф, полученный на основе родственных публикаций. Как следует из выражения (5), семантический профиль родственных публикаций не содержал рефератов, в которых контекстным поиском одновременно были определены названия двух белков из выборки. Поэтому, появление одинаковых идентификаторов *pmid'* в семантических профилях двух разных белков указывает, что соответствующие публикации содержат семантическую связность в неявном виде.

Из рисунка 3 видно, что определение скрытой семантической связности между белками возможно с использованием встроенных функций системы PubMed, поскольку при использовании родственных публикаций структура семантической сети сохранилась. Изолированные подграфы, обозначенные на рисунке 3, совпали по своему составу с распределением белков по метаболическим путям.

В отличие от сети, полученной для релевантных публикаций (ср. рис. 2 и рис. 3), построенный по родственным семантическим отношениям граф содержит большее число подграфов – 15. Среди этих подграфов можно выделить 6 крупных, содержащих более 5 вершин (эти подграфы обозначены буквами на рис. 3). Белки, участвующие в метаболизме жирных кислот, представлены в подграфах «А» (15 вершин) и подграфе «Г» (21 вершина). В подграфе «Д» присутствуют карбоангидразы, а в состав подграфа «Е» вошли 6 белков метаболизма аргинина и пролина: пирролин-5-карбоксилат редуктазы (A8K798 и P32322), альдегиддегидрогеназы (P49189 и P30038), орнитин-аминотрансфераза (P04181) и пролиндегидрогеназа (A6NF53). Белки, принимающие участие в биосинтезе гепарансульфата, сгруппировались в подграфе «Б», содержащем 16 вершин. Состав подграфа совпадает с аналогичным подграфом «Г» на рисунке 2.

При выбранном уровне отсечения (l=0,009) на графе для родственных публикаций появилось 11 взаимосвязей, которых не было на графе для релевантных

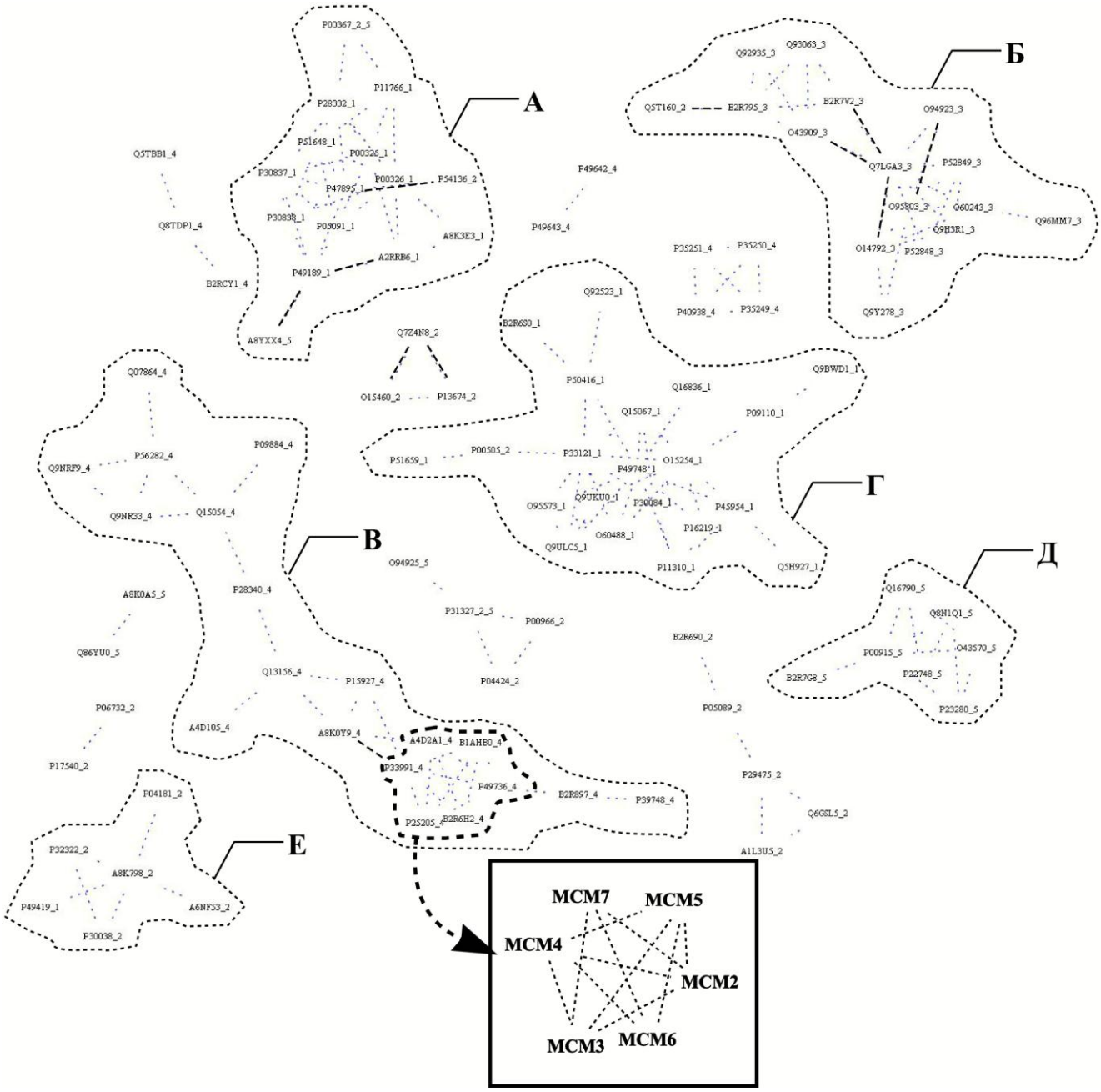

Рисунок 3. Семантическая сеть белков пяти метаболических путей, полученная с использованием родственных публикаций. Обозначения вершин – в соответствии с рисунком 2. Жирным пунктиром показаны ребра, отсутствующие на рисунке 2. На врезке указаны белки геликазного комплекса MCM.

публикаций. Из них 5 новых взаимосвязей выявлены между ферментами биосинтеза гепарансульфата. В частности, появились ребра, соединяющие белок EXT1, гепаран-сульфат сульфотрансферазу и экзостозин-подобный белок. Эти новые взаимосвязи обусловлены ведущей ролью нарушений синтеза гепарансульфата в хондроцитах и

остеобластах при развитии тяжелого наследственного заболевания костной ткани – экзостоза (Nadanaka S., Kitagawa H., 2008).

В составе подграфа «А» установилась ранее на выявленная в ходе анализа релевантных публикаций взаимосвязь между ферментами алкоголь- и альдегиддегидрогеназой, участвующими в метаболизме алкоголей. Две другие новые взаимосвязи в составе подграфа «А» соединяют альдегиддегидрогеназу, входящую в состав пути метаболизма липидов, с ферментами биосинтеза белков. Альдегиддегидрогеназа класса 6 и аргинил-тРНК синтетазой связаны в 42 родственных публикациях. Анализ этих публикаций показывает, что связующим звеном является витамин А (ретинол). Его равновесие с ретиналем регулируется альдегиддегидрогеназой, и одновременно ретинол оказывает стимулирующее воздействие на уровень экспрессии аминоацил-тРНК синтаз.

Одно новое ребро наблюдается в подграфе «В», который объединяет белки репликации ДНК. Взаимосвязь появилась между фактором регуляции репликации MCM4 и наиболее хорошо охарактеризованным белком этой группы – репликационным белком А, без которого невозможно протекание большинства процессов, в которых расплетается двойная спираль ДНК.

3.4. Сопоставление сетевых подграфов с разделами KEGG и с аннотациями GO

Для оценки степени соответствия состава изолированных подграфов семантической сети разделам KEGG или аннотациям онтологии генов GO, по формуле (9) рассчитывали вероятность p случайного объединения белков в один подграф. Значения $p<0,05$ указывали на неслучайный характер распределения белков по подграфам. В таблице 2 приведены данные о количестве вершин в каждом подграфе, совпадающих с белками одного метаболического пути KEGG. Видно, что совпадения характеризуются значениями p значительно ниже порогового уровня 0,05. Например, из 29 белков, участвующих в репликации ДНК, в состав одного подграфа вошел 21 белок при анализе релевантных публикаций и 19 белков при анализе родственных публикаций.

Значения p для указанных подграфов составили $1,6*10^{-25}$ и $6,4*10^{-15}$, соответственно. Для других подграфов значения p находятся в диапазоне от 10^{-5} до 10^{-23}, то есть подграфы хорошо совпадают с разделами базы данных KEGG. Это означает, что полученные семантические сети отражают закономерности молекулярных процессов, положенные в основу структуры данных в системе KEGG.

Таблица 2. Сопоставление состава подграфов (ПГ) с распределением белков по разделам базы данных KEGG. КВ – количество вершин в подграфе. *p* – вероятность случайного объединения белков одного метаболического пути в подграф.

		Распределение белков по метаболическим путям KEGG					
ПГ	**КВ**	**Метаболизм жирных кислот**	**Метаболизм аргинина и пролина**	**Биосинтез гепаран сульфата**	**Репликация ДНК**	**Метаболизм азота**	***p***
		44	33	18	35	21	
«А»[а)]	10	0	0	0	0	10	$2{,}8 \times 10^{-10}$
«Б»[а)]	29	0	0	0	21	0	$1{,}6 \times 10^{-25}$
«В»[а)]	55	39	15	0	0	1	$5{,}8 \times 10^{-18}$
«Г»[а)]	18	0	0	18	0	0	$1{,}1 \times 10^{-23}$
«А»[б)]	15	12	2	0	0	1	$2{,}2 \times 10^{-5}$
«Б»[б)]	16	0	1	15	0	0	$8{,}8 \times 10^{-20}$
«В»[б)]	21	20	1	0	0	0	$7{,}5 \times 10^{-12}$
«Г»[б)]	19	0	0	0	19	0	$6{,}4 \times 10^{-15}$
«Д»[б)]	7	0	0	0	0	7	10^{-7}
«Е»[б)]	6	1	5	0	0	0	10^{-4}

[а)] Для релевантных публикаций, в соответствии с рисунком 2.
[б)] Для родственных публикаций, в соответствии с рисунком 3.

Данные таблицы 2 позволяют указать на соответствие между подграфами, полученными в результате расчета меры семантической связности по релевантным и по родственным публикациям. Раздел «репликация ДНК» базы данных KEGG представлен 35 белками, из которых, как уже указывалось, 21 белок оказался в составе подграфа «Б» (по релевантным) и 19 белков оказались в составе подграфа «Г» (по родственным). Белки раздела «биосинтез гепаран сульфата» образовали подграфы из 18 и 15 белков для случаев релевантных и родственных публикаций, соответственно. Белки, участвующие в метаболизме жирных кислот, образовали один смешанный подграф «Б» из 29 вершин по результатам анализа релевантных публикаций, а для родственных этот состоящий из 44 белков метаболический путь разделился на два подграфа «А» и «Г», суммарное количество вершин в которых – 34.

Для отобранных по базе данных KEGG белков были загружены аннотации согласно онтологии генов GO. Аннотации были найдены только для 128 из 150 белков исходной выборки. В GO аннотации белков подразделяются на три раздела: «клеточная локализация», «биологический процесс» и «молекулярная функция». Из анализируемой выборки в указанные категории вошло 100, 102 и 110 аннотированных белков, соответственно. Для 68% белков присутствовали аннотации в терминах всех трех категорий GO. Для полученной выборки из 128 белков была сконструирована семантическая сеть, объединяющая информацию о релевантных публикациях.

Таблица 3 содержит данные, показывающие, что в составе подграфов семантической сети обнаруживаются белки с одинаковыми аннотациями GO. Таблица отсортирована по последнему столбцу, в котором указана вероятность случайного объединения в составе подграфа белков с одинаковой аннотацией GO. Видно, что наибольшая вероятность p=0,0002 получилась для белков, участвующих в транспорте электронов и локализованных в мембране, однако даже это самое высокое значение на два порядка ниже выбранного порога достоверности p<0,05.

В составе одного подграфа наблюдались белки, имеющие одинаковые аннотации по разным разделам онтологии. Например, 22 белка в составе подграфа «Б» (см. рис. 2) участвуют согласно разделу «Биологический процесс» в репликации ДНК, при этом 16 из них локализованы в нуклеоплазме по аннотации в разделе «Клеточная локализация». Аналогично, в подграфе «Г», 9 белков в разделе «Биологический процесс» получили аннотацию «биосинтез гепарансульфата», которая совпадает с классификацией этих белков в базе данных KEGG.

Все 9 белков, обладающих карбонат-дегидратазной активностью согласно разделу «Молекулярные функции», оказались локализованы в подграфе «А». Согласно базе данных KEGG, все белки этого подграфа участвуют в метаболизме азот-содержащих соединений. В целом можно отметить, что состав подграфов семантической сети не только хорошо согласуется с метаболическими путями, но и обладает статистически значимой специфичностью в отношении определенных аннотаций GO.

Таблица 3. Распределение белков (M=128), входящих в состав изолированных подграфов, в соответствии с аннотациями GO. *K* – количество белков с заданной аннотацией в выборке, *x* – количество вершин с заданной аннотацией в подграфе, *p* – вероятность, рассчитанная по формуле (9).

ПГ[a)]	Раздел GO	Категория GO	*K*	*x*	*P*
«Б»	БП	Репликация ДНК	22	22	$1,2*10^{-22}$
«А»	МФ	Карбонат-дегидратазная активность	9	9	$1,6*10^{-21}$
«Г»	БП	Биосинтез гепарансульфата	9	9	$5,2*10^{-14}$
«Б»	Л	Нуклеоплазма	16	16	$3*10^{-13}$
«Б»	МФ	Связывание с ДНК	16	16	$3*10^{-13}$
	БП	Окислительно-восстановительные процессы	30	30	$7,3*10^{-13}$
	Л	Мембрана	21	13	$3,6*10^{-9}$
	МФ	Связывание ионов цинка	21	9	$1,5*10^{-8}$
	Л	Аппарат Гольджи	10	9	$4,8*10^{-8}$
	МФ	Связывание АТФ	9	9	$3,5*10^{-6}$
	МФ	Образование белковых комплексов	23	14	$1,7*10^{-5}$
	Л	Митохондрии	13	13	$2*10^{-4}$
	МФ	Транспорт электронов	15	15	$2*10^{-4}$

[a)]Обозначения подграфов согласно рисунку 2.

4. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Алгоритмический анализ семантической взаимосвязи между парой белков, названия которых встретились в одном реферате, обычно проводится для выявления белок-белковых взаимодействий (Ding J. *et al.*, 2002). Из полученных нами результатов следует, что аналогичный подход может эффективно применяться не только для определения попарных взаимодействий, но и для отображения охватывающих несколько белков сетевых взаимосвязей.

Анализ качества распознавания обозначений белков в текстах научных публикаций показал, что результаты автоматического контекстного поиска на 80% совпадают с опубликованными данными о том, насколько часто тот или иной белок встречается в протеомных статьях. Следовательно, информация из базы данных UniProt может использоваться в поисковых системах для формирования семантических профилей, специфичных для каждого белка. В перспективе можно ожидать увеличения соответствия между используемыми авторами научных статей обозначениями белков и названиями, рекомендуемыми кураторами базы данных UniProt.

Мера семантической связности между белками была введена путем сопоставления ассоциированных с белками профилей публикаций. Явная связность отражала количество релевантных публикаций, в которых названия двух белков встречаются одновременно. Скрытая связность вводилась пропорционально количеству одинаковых для двух белков родственных публикаций, при этом все релевантные публикации, совпадающие для двух белков, исключались из анализа. Родственные публикации не содержат совместного упоминания названий белков, но лежащее в основе оценки родственности публикаций сравнение частот употребления терминов в научных статьях позволяет установить значительную часть семантических отношений между объектами.

Проведенный анализ семантического графа белков, построенного с использованием релевантных публикаций (см. рис. 2), показывает, что предложенная информационная модель отражает существующие в клетке метаболические взаимосвязи между белками. Из таблицы 2 видно, что, в зависимости от метаболического пути, на графе было отображено от 45% (метаболизм аргинина и пролина) до 100% (метаболизм гепарансульфата) входящих в этот путь белков.

Информативность семантической сети существенно не снижается при переходе от релевантных к родственным публикациям (ср. рис. 2 и рис. 3). Группировка белков в изолированные подграфы соотносится со структурой данных в ресурсах KEGG и GO, и носит неслучайный характер, что следует из низких значений вероятности p (см. табл. 2 и 3).

Наибольший интерес с точки зрения анализа белок-белковых взаимосвязей представляет возможность выявления скрытой (латентной) связности. Обычно для этого применяются специализированные алгоритмы; например, метод скрытого семантического индексирования, основанный на сравнении частот встречаемости терминов в текстах рефератов MEDLINE (Homayouni R. *et al.*, 2005). В данной работе мы показали, что семантическая сеть (см. рис. 3) может быть получена в результате сравнения идентификаторов релевантных публикаций, выводимых в поле «Related Links» системы PubMed. Следует обратить внимание, что множество родственных публикаций было получено таким образом, что ни в одном реферате одновременно не встречались названия двух любых белков из выборки. Кроме того, из выборки были исключены даже публикации, родственные по отношению к рефератам, где совместно встречаются названия двух белков (см. выражение (5). Несмотря на намеренное обеднение множества родственных публикаций, были получены подграфы, хорошо совпадающие с метаболическими путями (см. табл. 2). Более того, на семантической сети, построенной по родственным публикациям, появились дополнительные ребра. Они отражают такие отношения между белками, которые невозможно выявить контекстным поиском названий белков в системе PubMed.

Неоспоримым преимуществом предложенного в работе подхода является возможность сопоставления результатов широкомасштабных транскриптомных или протеомных экспериментов с текущим уровнем знаний, отраженном в рефератах рецензируемых научных статей. В целом проанализированные в работе семантические отношения между белками не выходят за рамки современного курса лекций по биохимии и молекулярной биологии. В тоже время, расширение предметной области за рамки классических представлений возможно за счет выявления скрытой семантической связности с использованием родственных публикаций.

5. ВЫВОДЫ

1. С использованием номенклатурных обозначений из базы данных UniProt в автоматическом режиме получены релевантные семантические профили, специфичные для каждого белка. Релевантный семантический профиль представлял собой множество идентификаторов публикаций из ресурса MEDLINE, найденных контекстным поиском в текстах рефератов по наименованиям белков.

2. Множество входящих в состав семантических профилей рефератов расширяется в три раза при включении в него родственных публикаций, предлагаемых автоматическими средствами оценки смыслового сходства документов. За счет родственных публикаций было установлено дополнительно около 3-х тыс. неявных семантических связей между белками из пяти различных метаболических путей базы данных KEGG.

3. Мера семантического сходства между двумя белками определена как множество рефератов публикаций, совпадающих при сравнении семантических профилей белков. С использованием этой меры получены семантические сети, отображающие взаимосвязи между белками в составе хорошо изученных биохимических и молекулярно-биологических процессов. Семантические сети, построенные на основе сопоставления релевантных и родственных профилей, сходны между собой, причем сравнение родственных профилей позволяет выявить дополнительные семантические взаимосвязи между белками.

4. В составе каждой семантической сети выявлено несколько изолированных подграфов. Показано, что в состав подграфов входят белки, относящиеся к одному метаболическому пути и имеющие одинаковые аннотации в системе онтологии GO, с вероятностью случайного объединения $p<10^{-3}$.

6. СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Пономаренко Е.А., Лисица А.В., Карузина И.И., Мирошниченко Ю.В. Автоматизированное аннотирование функциональных свойств белков надсемейства цитохромов Р450 // Аллергия, астма и клиническая иммунология – 2003 – № 7(8) – 95-99.
2. Lisitsa A.V, Ponomarenko E.A., Karuzina I.I., Ivanov A.S., Archakov A.I. Balance Sheet for Cytochrome P450 Knowledgebase // In: Proceedings 13-th International Conference on Cytochromes P450 – Prague – 2003 – 67-73.
3. Иванов Н.А., Лисица А.В., Пономаренко Е.А., Арчаков А.И. Тематический анализ резюме научных публикаций в области цитохромов Р450 //Сборник материалов Сессии ИВТН – Москва – 2003 – 28-29.
4. Лисица А.В., Мирошниченко Ю.В., Пономаренко Е.А. База знаний по цитохромам Р450 // Сборник научных трудов X Российского национального конгресса «Человек и лекарство» – 2003 – 730.
5. Lisitsa A.V., Ponomarenko E.A., Gusev S.A., Kuznetsova G.P., Karuzina I.I., Lewi P., Archakov A.I. Cytochrome P450 knowledgebase: structure and functionality //In: Proceedings 14th International conference on cytochromes P450: biophysics and bioinformatics – Dallas, USA – 2005 – 29-34.
6. Пономаренко Е.А., Лисица А.В., Гусев С.А. База знаний по цитохромам Р450 // Материалы международной школы-конференции молодых ученых «Системная биология и биоинженерия», МАКС Пресс – Москва – 2005 – 50.
7. Пономаренко Е.А., Лисица А.В., Карузина И.И., Гусев С.А. База знаний по цитохромам Р450 // Сборник материалов Сессии ИВТН – Москва – 2006 –32.
8. Ponomarenko E.A., Lisitsa A.V., Archakov A.I. Text Mining Tools in Analysis of High-Throughput Data // Материалы конференции CMTPI – 2007 – 135.
9. Ponomarenko E.A., Lisitsa A.V., Archakov A.I. Searching for Related Proteins Using Textomic Approach // Сборник трудов конференции HUPO – 2007 – 103.
10. Ponomarenko E.A., Lisitsa A.V., Petrak J., Moshkovskii S.A., Archakov A.I. Textomics Tools for Automatically update the Hit-parade of repeatedly identified proteins // Сборник материалов международной конференции GPBNM – 2008 – 36.
11. Пономаренко Е.А., Лисица А.В., Арчаков А.И. Лингвистические методы поиска взаимосвязанных белков // Сборник трудов конференции "Человек и лекарство" – 2008 – 523.
12. Ponomarenko E.A., Lisitsa A.V., Petrak J., Moshkovskii S.A., Archakov A.I. Automated meta-analysis confirms the Hit-parade of repeatedly identified proteins // Сборник материалов международной конференции HUPO – 2008 – 1669.

13. Пономаренко Е.А., Лисица А.В., Петрак И., Мошковский С.А., Арчаков А.И. Выявление дифференциально-экспрессирующихся белков с использованием автоматического мета-анализа протеомных публикаций.//Биомедицинская химия, 2009 - №55(1) – 5-14.

13a. Ponomarenko E.A., Lisitsa A.V., Petrak J., Moshkovskii S.A., Archakov A.I. Identification of Differentially Expressed Proteins Using Automated Meta-Analysis of Proteomics-Related Articles //Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry – 2009 – № 3(1) – 10-16.

14. Пономаренко Е.А., Лисица А.В., Арчаков А.И. Лингвистические методы поиска взаимосвязанных белков // Сборник трудов конференции "Математика. Компьютер. Образование" – 2009 – 68.