Биомедицинская химия, 2009 том 55, вып. 1, с. 5-14.

ПРОТЕОМИКА

УДК 004.912, 615.03, 577.29

ВЫЯВЛЕНИЕ ДИФФЕРЕНЦИАЛЬНО-ЭКСПРЕССИРУЮЩИХСЯ БЕЛКОВ С ИСПОЛЬЗОВАНИЕМ АВТОМАТИЧЕСКОГО МЕТА-АНАЛИЗА ПРОТЕОМНЫХ ПУБЛИКАЦИЙ

Е.А. Пономаренко[1]*, А.В. Лисица[1], И. Петрак[2], С.А. Мошковский[1], А.И. Арчаков[1]

[1]Государственное учреждение Научно-исследовательский институт биомедицинской химии им. В.Н. Ореховича РАМН, 119121, Россия, Москва, ул. Погодинская, 10; тел.: +7(499)246-37-31; эл. почта: 2463731@gmail.com
[2] Карлов Университет в Праге, факультет общей медицины, Институт патофизиологии, ул. Немошници, 5, Прага 2, Чехия

В работе рассматривается автоматический метод поиска дифференциально-экспрессирующихся белков (ДЭБ) в научных статьях по протеомике. Из электронной версии журнала “Proteomics” и открытого ресурса PubMedCentral были отобраны полнотекстовые публикации, посвященные различным протеомным исследованиям. В результате подсчета частот встречаемости названий белков в статьях получен список, на 86% совпадающий с ранее опубликованным в работе (*Petrak et al.,* Proteomics (2008) **8**, 1744) перечнем ДЭБ тканей человека. Независимо от целей экспериментального исследования и применявшегося аналитического метода анализа, различия в уровне экспрессии наблюдаются для белков семейств аннексинов, пероксиредоксинов, а также альфа-енолазы, триозофосфатизомеразы и HSP60. Среди опубликованных ДЭБ также присутствуют сывороточный альбумин, катепсин D и виментин. В статьях наиболее часто встречаются белки, участвующие в воспалении и формировании иммунного ответа, либо входящие в состав системы регуляции и передачи сигнала. Большинство из найденных белков изучаются в контексте исследования патогенеза онкологических заболеваний, болезней сердечно-сосудистой и нервной систем.

Ключевые слова: протеомика, мета-анализ, анализ текстов, 2-DE, LC-MS/MS.

ВВЕДЕНИЕ. Постгеномные технологии объединяют методы геномики, транскриптомики, протеомики и другие высокоэффективные методы исследования живых систем. Одним из аспектов применения постгеномных технологий является клиническая протеомика, и, в частности, поиск дифференциально-экспрессирующихся белков (ДЭБ). Особую важность при этом занимает изучение ДЭБ человека в норме и при различных патологиях. Перспективы развития разработок в этой области связаны с возможностью использования постгеномных технологий в медицинской молекулярной диагностике [1].

В большинстве исследований, направленных на выявление ДЭБ, не учитываются технические ограничения используемых методов. Основными методами исследования экспрессии белков остается метод двумерного электрофореза в полиакриламидном геле (2-DE) в сочетании с масс-спектрометрическими методами идентификации белков [2]. Также для поиска потенциальных биомаркеров используется комбинация технологий высокоэффективной жидкостной хроматографии с масс-спектрометрией (LC-MS/MS) [3]. Концентрационная чувствительность этих широко применяемых методов составляет 10^{-7}-10^{-8} М [4, 5], поэтому определение белков, концентрации которых ниже 10^{-10} М, как правило, невозможно. По мнению ряда авторов [6, 7], ограничение по чувствительности приводит к низкой специфичности выявляемых биомаркеров.

* - адресат для переписки

Систематический подход к анализу опубликованных ДЭБ впервые был предложен в работе [8]. Этой группой исследователей был проведен мета-анализ статей, вышедших в журнале *"Proteomics"* в 2004-2006 годах. Авторы проанализировали работы, в которых были исследованы ткани человека и грызунов с применением метода двумерного электрофореза. На выборке из 99 статей было установлено, что независимо от характера исследуемого процесса, в совершенно разных по своим целям работах исследователи отмечали изменение уровня экспрессии преимущественно одних и тех же белков. На основании этих данных был сформирован "хит-парад" наиболее часто упоминающихся в статьях белков, а определение одинаковых ДЭБ в различных экспериментах получило название протеомного *"déjà vu"* (от французского "уже виденное").

Задачей данной работы было создание алгоритма мета-анализа протеомных публикаций, основанного на оценке частот встречаемости названий белков в выборке текстов. В работе предлагается автоматизировать мета-анализ с целью широкомасштабного поиска в текстах публикаций ДЭБ.

МЕТОДИКА.

Анализ встречаемости названий белков в текстах статей. Названия белков и дополнительная информация об их свойствах (например, данные о локализации, участии в патологическом процессе и функциональной активности) были получены из базы знаний UniProt [9], совмещающей в себе ресурсы SwissProt и TreMBL. Из UniProtKB (UniProt Knowledgebase) были отобраны записи, отвечающие поисковому запросу [*"organism:"Human [9606]""*] и [*"Protein existence: "evidence at protein level"*], т.е. белки человека, экспрессия которых подтверждена на уровне протеома.

Информация о номенклатуре белка, согласно формату ресурса UniProtKB, представлена в виде основного названия, альтернативных названий (синонимов), номера доступа в системе UniProtKB и идентификатора (см. рис. 1б-д). Для определения названий белков в текстах статей формировали подстроку поиска, в которую входили общепринятые названия и синонимы белка, а также идентификаторы и номера доступа к этим белкам в системе UniProtKB.

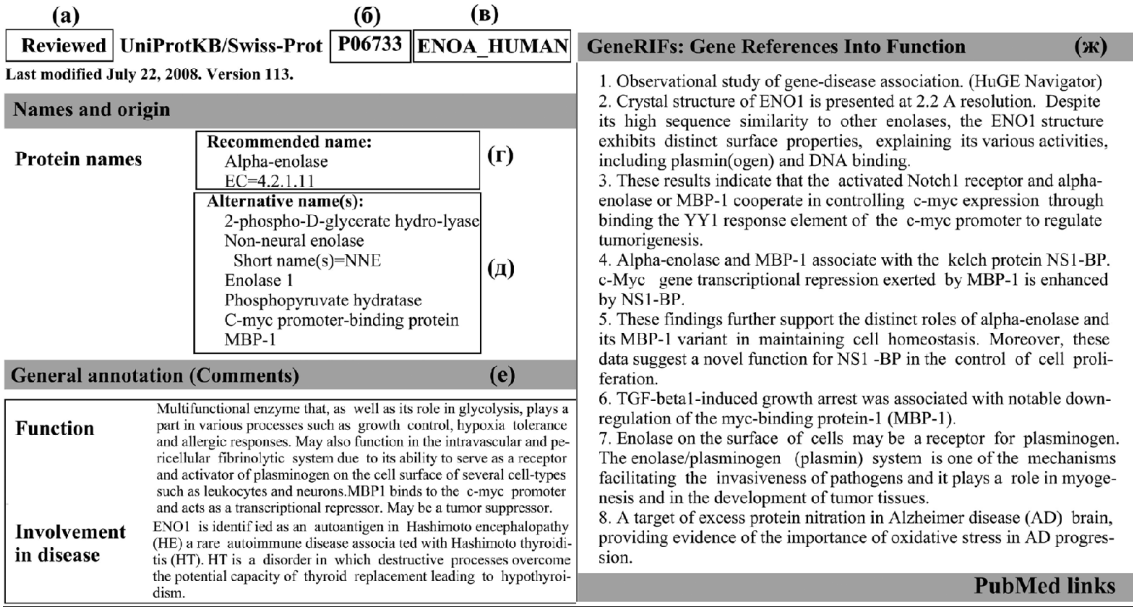

Рисунок.

Фрагменты информационных записей в системах UniProt (а-е) и GenRif (ж) для белка альфа-енолаза: (а) отметка о прохождении экспертизы; (б) номер доступа в системе UniProt; (в) идентификатор белка в системе UniProt; (г) рекомендованное название белка; (д) альтернативные названия белка и синонимы; (е) функциональная аннотация белка и взаимосвязь с развитием заболевания; (ж) выписка GenRif о функциональной роли белка.

Названиями белка считались встречающиеся в тексте слова, по своему написанию идентичные подстроке поиска. Если отсутствовало однозначное соответствие между найденным названием и записью в базе знаний UniProtKB (например, если в нескольких записях было указано одинаковое название для разных белков), то найденный термин не рассматривался как белковый идентификатор. Если название белка совпадало по написанию с широко используемыми терминами, дополнительно проводилась экспертная оценка контекста употребления этого термина.

Для каждого белка подсчитывали количество статей, в которых встречается его название, при этом контекст упоминания белка не анализировали. Отбирались те белки, названия которых встретились как минимум в 5% публикаций в соответствующей выборке (см. табл. 1).

Таблица 1. Выборки научных публикаций.

№	*Обозначение выборки*	*Источник публикаций*	*Количество статей*	*Поисковый запрос*
1	**REVIEWED99**	**Proteomics**	**99**	–
2	2-DE	PMC	100	«(2-DE **OR** 2-D electrophoresis) **AND** human»
3	**LC-MS/MS**	**Proteomics**	**94**	**«LC-MS/MS AND human»**
4	Random	PMC	100	«proteomics»

Загрузку публикаций и поиск названий белков проводили в автоматическом режиме с использованием вычислительных мощностей кластера SysBiont [10] под управлением скриптов, написанных на языке Perl [11].

Выборки текстов научных публикаций. В работе использовали 4 выборки, каждая из которых содержала около 100 текстов научных статей (см. табл. 1). В первую выборку, обозначенную как REVIEWED99, вошли публикации, ранее проанализированные в работе [8]. Во вторую (2-DE) и третью (LC-MS/MS) выборки вошли публикации, найденные по ключевым словам (см. табл. 1) в открытой системе PubMedCentral (PMC) [12] и в электронной версии журнала *"Proteomics*", соответственно. Четвертая выборка (Random) включала в себя отобранные случайным образом публикации из списка научных работ, предоставленных по поисковому запросу "proteomics" ресурсом PMC. Дополнительных ограничений, касающихся характера исследуемого биоматериала и экспериментальных условий, при формировании выборок не вводилось.

Отобранные публикации, включая таблицы, названия таблиц и подписи к рисункам, преобразовывали в текстовый формат. Дополнительные материалы к статьям не анализировали.

Аннотация белков. Найденные с наибольшей частотой встречаемости белки были аннотированы с использованием информации, предоставляемой ресурсами GenRif [13] и UniProt. Система GenRif является одним из подразделов информационного портала EntrezGene и связывает название гена с краткой фразой, опубликованной в библиотеке PubMed [14], описывающей функции этого гена. В данной работе в системе GenRif анализировали информацию о связи гена с развитием патологического процесса и о функциях соответствующего белка (см. рис. 1ж). Данные о функциях и связи белков с развитием конкретного заболевания также были получены из соответствующих полей системы UniProt (см. рис. 1е).

РЕЗУЛЬТАТЫ. Версия базы знаний UniProt 13.1, датированная 18 марта 2008 г., содержала сведения по более чем 70 тыс. белкам человека. Из них было отобрано 11 тысяч белков, существование которых доказано на протеомном уровне, т.е. подтверждено масс-спектрометрической идентификацией, методами рентгеноструктурного анализа или с использованием антител. Для 95% отобранных белков найдено более одного названия; с учетом альтернативных и синонимичных вариантов в словарь названий белков вошло около 30 тысяч терминов.

Для каждой из четырех выборок (REVIEWED99, 2-DE, LC-MS/MS и Random) в автоматическом режиме был сформирован список названий белков, которые встречаются как минимум в 5% статей внутри выборки. Выборке REVIEWED99 при этом соответствовали два перечня названий белков: один был сформирован в автоматическом режиме с использованием мета-анализа, а второй создан экспертами (TOP15, контрольный) и опубликован в работе [8] (см. табл. 2).

В результате обработки четырёх выборок было получено 48 уникальных названий белков, из них 19 присутствовало только в одном из пяти сформированных списков. Такие белки не рассматривались в данной работе.

Во всех исследуемых выборках публикаций с высокой частотой были найдены следующие белки: альфа-енолаза, пероксиредоксины (пероксиредоксин 1 или пероксиредоксин 2), триозофосфатизомераза, HSP60 и белки из семейства аннексинов (аннексин A1, A2, A4 или A5). Одновременно в трёх исследуемых выборках статей встретились следующие белки: эпидермальный фактор роста, глицеральдегид-3-фосфатдегидрогеназа и интерлейкин-8. Остальные белки присутствовали только в двух выборках, при этом ATP-синтаза, катепсин D, GRP78, HSP70, HSP27 и кератин с высокой частотой наблюдались только в выборке публикаций REVIEWED99.

Сравнение результатов анализа тестовой выборки, полученных в автоматическом и экспертном режиме, показало, что 12 белков из сформированного автоматическим путем списка присутствуют в экспертном перечне TOP15 (т.е., совпадают 86% белков). Только 2 белка – пируваткиназа M1/M2 и ингибитор диссоциации GDP-Rho – из TOP15 не вошли в автоматически сформированный список.

Помимо белков, присутствующих в TOP15 (т.е., белков с изменённым уровнем экспрессии), в результате автоматической обработки публикаций также были выявлены другие белки, упоминающиеся в выборке REVIEWED99 с высокой частотой. Это - убиквитин (P62988), HSP60 (P10809), глицеральдегид-3-фосфатдегидрогеназа (P04406) и сывороточный альбумин (P02768).

Одной из причин, по которой в список автоматически выявленных белков вошли дополнительные белки, является различие в объеме анализируемого материала. Экспертный анализ проводился только на основании материала таблиц, в то время как с использованием автоматического мета-анализа обрабатывался весь текст статьи. В результате показано, что в 66% случаев названия белков встречаются в разделе “Результаты”, а в 24% - в разделе “Обсуждение”. В названиях статей, в разделах “Реферат”, “Введение”, “Материалы и методы” название белков встречаются редко – в сумме около 9%. В разделах “Заключение” и “Список литературных источников” названия белков встречаются в единичных случаях. Это наблюдение указывает на целесообразность использования для мета-анализа полнотекстовых публикаций.

Найденные в результате мета-анализа белки в основном локализованы в цитоплазме (пероксиредоксины, триозофосфатизомераза, циклофилин A, виментин, кальмодулин, HSC70) или являются секретируемыми белками (интерлейкин-8, сывороточный альбумин, ангиотензиноген, металлопротеиназа-9, серотрансферрин); в меньшей степени представлены мембранные белки (эпидермальный фактор роста, ErbB2) .

Таблица 2. Список 15 наиболее часто встречающихся белков, найденных более чем в 5% статей исследуемых групп. Результаты поиска для белков семейств аннексинов и пероксиредоксинов объединены в одну графу. Колонки 1 и 2 содержат результаты автоматического и экспертного анализа выборки REVIEWED99; колонки 3, 4 и 5 содержат результаты автоматического анализа выборок 2-DE, LC-MS/MS и Random, соответственно. В колонках 1-5 указана относительная частота упоминания названия белка в соответствующей выборке (в % от количества публикаций внутри выборки). Полный список белков, выявленных в данной работе, доступен в качестве дополнительного материала по адресу http://ibmc.msk.ru/departments/bt/papers/.

№	Название белка (локализация) номер доступа в UniProt	Номер выборки частота (%)					Краткое описание[1]	Связь с развитием патологических процессов[1]
		1	2	3	4	5		
1	**Альфа-енолаза (Ц, М, Я)[2] P06733**	**16**	**31**	**11**	**22**	**12**	**Фермент, участвующий в предпоследнем этапе гликолиза. Катализирует переход 2-фосфо-D-глицерата в фосфоенолпируват.**	**Энцефалопатия Хашимото, тиреоидит Хашимото, астма, кератоконус, болезнь Альцгеймера**
2	Триозофосфатизомераза[3] (Ц) P60174	26	22	7	13	5	Катализирует превращение D-глицеральдегид-3-фосфата в дигидроксиацетонфосфат в ходе пентозофосфатного цикла.	Триозофосфат-изомеразный дефицит, хроническая гемолитическая анемия, нейро-мышечные дисфункции, кардиомиопатия, неонатальная желтуха, нейродегенеративные заболевания, склероз тканей, спленомегалия, холелитиаз
3	Аннексины (A1, A2, A4, A5) (С) P04083, P07355, P09525, P08758	36	19	13	32	9	Относятся к семейству Ca^{2+}- и фосфолипид-связывающих белков (липокортинов), блокируют фосфолипазу А2. Принимают участие в экзоцитозе.	Лейкоз
4	Пероксиредоксины[3] (пероксиредоксин 1, пероксиредоксин 2) (Ц) Q06830, P32119	42	21	6	18	12	Антиоксиданты, участвуют во внутриклеточной системе передачи сигнала за счет контроля индуцируемого цитокинами уровня	Колоректальный рак, кардиомиопатии, сахарный диабет, рак легких, облитерирующий бронхиолит, болезнь Паркинсона, синдром Дауна, болезнь Альцгеймера, сидром Пика
5	**Белок теплового шока 60 кДа[3] (Мт) P10809**	**18**	**-**	**8**	**9**	**5**	**Митохондриальный шаперонин. Участвует в формировании иммунного ответа, в рефолдинге и транспортировке белков из цитоплазмы в матрикс митохондрий, а также в формировании третичной структуры белков.**	**Спастическая параплегия 13 типа, диабет, аутоиммунные заболевания, карциномы желчного пузыря и кишечника, дерматомиозит, сердечная патология, сепсис, прионные заболевания**
6	Эпидермальный фактор роста (М) P01133	-	-	12	16	6	Участвует в регуляции клеточного роста, пролиферации и дифференцировке клеток эпидермальных и эпителиальных тканей.	Гипомагниемия 4 типа (почечная гипомагниемия), нефропатия, аутизм, шизофрения, рак легких, рак шеи, рак простаты, нейробластома, крипторхизм, эндометриоз, глиома, меланома

Примечание. [1] - На основании данных ресурсов UniProt и GenRIF; [2] - Ц - цитоплазма, М - мембрана, Мт-митохондрии, Я - ядро, С - секретируемый белок, Л - лизосомы; [3] - Название белка было указано в реферате публикации.

Таблица 2. Продолжение.

№	Название белка (локализация) номер доступа в UniProt	Номер выборки частота (%)					Краткое описание[1]	Связь с развитием патологических процессов[1]
		1	2	3	4	5		
7	**Глицеральдегид-3-фосфатдегидрогеназа (Ц, М) P04406**	**12**	**-**	**-**	**18**	**9**	**Гликолитическая активность, активация транскрипции, инициализация апоптоза, участие в формировании эндосом.**	**Рак простаты, нейродегенеративные заболевания, опухоль мозга.**
8	Интерлейкин-8 (С) P10145	-	-	7	9	5	Действует на нейтрофилы, базофилы, лимфоциты. Хемотаксис нейтрофилов. Противовоспалительная реакция.	Шизофрения, псориаз, сепсис.
9	**Пептидилпролил-*цис-транс*-изомераза А (циклофилин А)[3] (Ц) P62937**	**10**	**17**	**-**	**11**	**-**	**Обладает шаперонной активностью и катализирует *цис-транс*-изомеризацию пептидной связи, предшествующей остатку пролина, придавая активную конформацию пептидам. Ускоряет фолдинг белков.**	**Рак легких, атеросклероз, инфаркт.**
10	Виментин[3] (Ц) P08670	23	20	-	9	-	В виде гомополимера образует интермедиальные филаменты - компоненты цитоскелета	Андроген-независимый рак простаты, рак груди, колоректальная карцинома.
11	**Сывороточный альбумин (С) P02768**	**13**	**-**	**-**	**10**	**-**	**Регуляция коллоидного осмотического давления крови, транспорт стероидных и тиреоидных гормонов, жирных кислот.**	**Эутиреоидная гипертироксинемия.**
12	Ангиотензиноген (С) P01019	-	-	8	11	-	Предшественник ангиотензина 1. Компонент ренин-ангиотензиновой системы (регуляция кровяного давления)	Эссенциальная гипертензия.
13	АТР-синтаза, F-тип, субъединица В (Ц, М)[2,3] P24539	12	15	-	-	-	Синтез ATP из ADP.	Нет данных
14	Кальмодулин (Ц) P62158	-	-	7	11	-	Ca^{2+}-связывающий белок. Является интегральной субъединицей целого ряда ферментов (протеинкиназы, протеинфосфатазы, фосфодиэстеразы, ферменты мышечной подвижности).	Нет данных
15	**Катепсин D[3] (Л) P07339**	**21**	**16**	**-**	**-**	**-**	**Внутриклеточная протеиназа.**	**Нейродегенеративные заболевания, рак молочной железы.**

Примечание. [1] - На основании данных ресурсов UniProt и GenRIF; [2] - Ц - цитоплазма, М - мембрана, Мт-митохондрии, Я - ядро, С - секретируемый белок, Л - лизосомы; [3] - Название белка было указано в реферате публикации.

Из полученных данных следует, что чаще всего в протеомных исследованиях идентифицируются белки, участвующие в воспалении, формировании иммунного ответа и передаче сигнала, наряду с белками системы регуляции. Реже встречаются ферменты энергетического обмена (например, глицеральдегид-3-фосфатдегидрогеназа), антиоксиданты (пероксиредоксины), а также белки, участвующие в трансляции и фолдинге. Структурные белки клетки и цитоскелета, такие, как кератин и тропомиозин, а также транспортные белки и ферменты катаболизма упоминаются с частотой 25% .

На основании информации ресурсов UniProt и GenRif (см. табл. 2), можно утверждать, что большинство найденных белков встречаются в литературе в контексте исследования таких патологических состояний, как злокачественные опухоли (рак простаты, кишечника, молочной железы, лёгких, шейки матки, колоректальный рак, рак желчного пузыря), болезни сердечно-сосудистой (кардиомиопатия, эссенциальная гипертензия, атеросклероз) и нервной системы (болезнь Альцгеймера, болезнь Паркинсона, эпилепсия, нейробластома, наследственная нейропатия, шизофрения) .

ОБСУЖДЕНИЕ. Автоматическая идентификация наиболее часто встречающихся белков в группе текстов может быть использована в случаях, если требуется в сжатые сроки получить общее представление о результатах высокопроизводительных исследований по определенному направлению. Предложенный алгоритм позволяет сэкономить время исследователя: обработка четырех выборок в нашей работе заняла 5 часов, в то время как мета-анализ одной выборки, выполненный с участием эксперта в работе [8], потребовал 9 рабочих дней. Автоматический анализ публикаций, хотя и уступает по качеству экспертной обработке (подсчет частоты встречаемости слов не подразумевает проведение анализа контекста статьи и критическую оценку результатов эксперимента), представляется более перспективным в связи с постоянно возрастающим количеством публикуемых данных.

Эффективность предлагаемого подхода в большой степени зависит от качества автоматического распознавания названий белков в тексте публикации. Несмотря на кажущуюся простоту, в действительности автоматическое распознавание названий белков в текстах сопряжено с рядом сложностей [15]. Это обусловлено несколькими факторами, в том числе отсутствием четких правил использования названий белка в статьях и встречающимися иногда синтаксическими ошибками [16]. Нередко название гена или белка совпадает по написанию с другими словами и смысловое наполнение в этом случае зависит от контекста употребления [17]: например, название “big brain” может означать как полное название гена *Drosophila melanogaster* bib (P23645), так и анатомическое описание.

Расхождения в названиях белков между указанными в базе знаний UniProt и обычно используемыми авторами научных статей могут быть проиллюстрированы на следующем примере. Белок катепсин D, часто упоминающийся в научных публикациях, *ранее* обозначался согласно номенклатуре UniProt версии 13.1 как “Катепсин D-прекурсор”, т.е. как предшественник активной формы катепсина D (см. историю редактирования записи P07339 [18]). При этом отдельной записи для катепсина D база знаний не содержала. В современной версии ресурса (14.0) это несоответствие устранено, и запись под кодом P07339 представляет информацию о катепсине D, однако теперь ни одна запись системы не отвечает поисковому запросу “Cathepsin D precursor”.

Запись P01019 (ангиотензиноген) иллюстрирует пример объединения в UniProt нескольких разных белков под одним кодом доступа. Ангиотензиноген является предшественником ангиотензина-1, из которого последовательно образуются ангиотензин-2 и ангиотензин-3 [19]. Каждый из этих белков обладает уникальными функциями и строением, однако названия всех этих белков присутствуют в базе знаний UniProt как варианты названий ангиотензиногена.

Возможно, это целесообразно с точки зрения молекулярной генетики, но является существенной проблемой при обработке результатов протеомных исследований.

При проведении автоматического поиска названий белков также следует учитывать и возможность появления ложноположительных результатов. Например, термин “тиоредоксин” в ходе компьютерного мета-анализа автоматически будет ассоциирован с белком тиоредоксином (запись в базе знаний UniProt P10599), в то время, как в большинстве статей термин “тиоредоксин” упомянут в контексте “тиоредоксинпероксидаза” и указывает на другой белок - пероксиредоксин.

Представленные выше примеры иллюстрируют, что на этапе автоматической обработки текста практически неизбежны частичные потери информации, а также получение ложноположительных результатов. Используемый в настоящей работе алгоритм позволяет выявлять до 80% названий белков, то есть, из 100 случайным образом отобранных терминов, которые алгоритм определил как названия белков, 80 при проверке действительно оказались таковыми. Достаточно высокий уровень распознавания указывает на благоприятную тенденцию к стандартизации терминологии обозначения белков. Вероятно, такая тенденция наметилась благодаря накоплению полностью расшифрованных геномов, на основе которых можно определить конечное разнообразие возможных белковых продуктов.

Стандартизация обозначений генов и белков является также следствием применения высокоэффективных методов исследования. Идентификация протеома осуществляется путем сравнения результатов масс-спектрометрии с базами данных известных белков, поэтому результаты поиска, т.е. названия белков, выводятся согласно номенклатуре информационного ресурса. В дальнейшем обозначения идентифицированных белков используются авторами статей в неизмененном виде.

В данной работе на выборке REVIEWED99 показано, что большинство наиболее часто встречающихся белков практически совпадает со списком ДЭБ из списка TOP15, полученного на основе экспертного анализа тех же публикаций. В отличие от эксперта (молекулярного биолога), компьютерный алгоритм не проводит смыслового анализа контекста статьи, где описано увеличение или уменьшение уровня экспрессии, а подсчитывает частоту упоминания названия белка среди статей. Поскольку результаты автоматического мета-анализа согласуются с данными экспертной оценки, то, по-видимому, наиболее часто встречающиеся в статьях белки являются ДЭБ. В том случае, если ДЭБ отсутствуют в списке наиболее часто встречающихся белков, то, вероятно, их можно рассматривать в качестве потенциальных биомаркеров, поскольку они одновременно обладают специфичностью к состоянию и различаются по уровням экспрессии.

В данной работе алгоритм автоматического поиска наиболее часто встречающихся белков применяли для статей, описывающих исследования с использованием методов 2-DE и LC-MS/MS. Более половины белков, наиболее часто встречающихся в выборках публикаций 2-DE и LC-MS/MS, одинаковые. Это можно объяснить, если принять во внимание примерно одинаковый диапазон концентрационной чувствительности используемых методов [6]. Возможно, феномен протеомного *déjà vu* зависит от ограничений экспериментальной платформы, поэтому использование доступных в настоящее время протеомных технологий не позволит в ближайшем будущем существенно расширить список ДЭБ.

Анализ случайно отобранных статей в области протеомики подтверждает эту гипотезу. Наблюдается значительное совпадение списков наиболее часто встречаемых белков, полученных при анализе выборок 2-DE, LC-MS/MS и Random: одновременно во всех трех выборках встречаются 7 белков (см. табл. 2). Это означает, что существует регулярно выявляемая в протеомных экспериментах группа белков, многие из которых, по-видимому, относятся к группе белков “домашнего хозяйства”. К этой группе относятся белки и гены, функционирующие на всех стадиях жизненного цикла организма и обеспечивающие выполнение основных функций клетки, так называемые белки/гены “домашнего хозяйства”

(housekeeping proteins/genes) [20]. Сравнение результатов мета-анализа с генами "домашнего хозяйства", список которых опубликован в работе [21], показало, что 9 белков (альфа-енолаза, аннексин A2, глицеральдегид-3-фосфатдегидрогеназа, ATP-синтаза, кальмодулин, катепсин D, GRP78, HSC71 и убиквитин – см. табл. 2) действительно являются белками "домашнего хозяйства".

ВЫВОДЫ. Применение автоматического мета-анализа становится важной задачей в связи с постоянно возрастающим количеством публикуемых данных в области протеомики. Эффективность автоматического мета-анализа публикаций напрямую зависит от достоверности распознавания названий белков в текстах публикаций. На данный момент наиболее пригодными для автоматического анализа являются публикации, в которых для каждого белка указан код доступа в глобальных протеомных базах знаний (например, в базе знаний UniProt) а также приведены названия исследуемых заболеваний в соответствии с международной классификацией.

Автоматизированный мета-анализ полнотекстовых протеомных публикаций позволяет эффективно выявлять ДЭБ. Неспецифичность белков, экспрессия которых изменяется вне зависимости от характера воздействия, по-видимому, является одной из причин выявления одинаковых ДЭБ при различных состояниях. Другой возможной причиной выявления в протеомных экспериментах одних и тех же ДЭБ могут быть ограничения по концентрационной чувствительности существующих протеомных методов. Эти ограничения не позволяют анализировать различия в уровне экспрессии белков, присутствующих в биоматериале в низких концентрациях.

Работа выполнена в рамках Программы "Протеомика в медицине и биотехнологии" и грантов LC06044 и 0021620806 (Ministerstvo školství, mládeže a tělovýchovy, Czech Republic).

ЛИТЕРАТУРА

1. *Hanash S.M., Pitteri S.J., Faca V.M.* (2008) Nature, **452**, 572-579.
2. *Hanash S.M.* (2003) Nature, **422**, 226-232.
3. *Qian W.-J., Jacobs J.M., Liu T., Camp II D.G., Smith R.D.* (2006) Mol. Cell Proteomics, **5**, 1727-1744.
4. *Kuster B., Mann M.* (1998) Curr. Opin. Struct Biol., **8**, 393-400.
5. *Archakov A.I., Ivanov Yu.D., Lisitsa A.V., Zgoda V.G.* (2007) Proteomics, **7**, 4-9.
6. *Diamandis, E.P.* (2003) Clin. Chem., **49**, 1272–1275.
7. *Archakov A.I., Ivanov Yu.D.* (2007) Molecular BioSystems, **3**, 336-342.
8. *Petrak J., Ivanek R., Toman O., Cmejla R., Cmejlova J., Vyoral D., Zivny J., Vulpe C.D.* (2008) Proteomics, **8**, 1744-1749.
9. www.uniprot.org
10. www.supercomputers.ru/?page=rating
11. www.perl.org
12. www.pubmedcentral.nih.gov
13. www.ncbi.nlm.nih.gov/sites/entrez
14. www.pubmed.gov
15. *Krallinger M., Valencia A.* (2005) Genome Biol., **6**, 224.
16. *Zeeberg B., Riss J., Kane D., Bussey K., Uchio E., Linehan W., Barrett J., Weinstein J.* (2004) BMC Bioinformatics, **5**, 80.
17. *Chen L., Liu H., Friedman C.* (2005) Bioinformatics, **21**, 248-256.
18. www.uniprot.org/uniprot/P07339?version=*
19. *Paul M., Mehr A.P., Kreutz R.* (2006) Physiol. Rev., **86**, 747-803.

20. *Butte A.J., Dzau V.J., Glueck S.B.* (2001) Physiol. Genomics, **7**, 95–96.
21. *Eisenberg E., Levanon E.Y.* (2003) Trends in Genetics, **19**, 362-365.

Поступила: 11. 09. 2008.

IDENTIFICATION OF DIFFERENTIALLY EXPRESSED PROTEINS USING AUTOMATIC META-ANALYSIS OF PROTEOMICS-RELATED ARTICLES

E.A. Ponomarenko[1], A.V. Lisitsa[1], I. Petrak[2], S.A. Moshkovskii[1], A.I. Archakov[1]

[1]Institute of Biomedical Chemistry RAMS, Pogodinskaya ul., 10, Moscow, 119121 Russia,
tel.: +7(499)246-37-31; e-mail: 2463731@gmail.com
[2]Charles University in Prague, First Medical Faculty, Institute of Pathological Physiology,
U Nemocnice 5, Praha 2, Czech Republic

We present here a new method for automatic meta-analysis of proteomic articles using assessment of frequency of individual protein names in the text. The list of all possible human protein names including synonyms was retrieved from UniProt knowledgebase. The retrieved names were searched in full-texts of peer-reviewed publications from electronic version of "Proteomics" journal and from PubMedCentral. In the automatic mode we have confirmed the earlier list of proteins [*Petrak et al.,* Proteomics (2008) **8**, 1744] most frequently reported as differentially expressed (DEPs) in human tissues. We have also verified, that the most recurrent proteins were reported in proteomic papers regardless of tissue, experimental goals or, to some extent, experimental methods employed. Frequently reported DEPs were: annexins, peroxiredoxins, alpha-enolase, triosephosphate isomerase, and HSP60. Besides, serum albumin, cathepsin D and vimentin were observed with relatively high frequency. The DEPs were reported in papers related to oncological, cardiovascular and neuronal diseases, and were involved in such biological processes as inflammation, cell regulation, immune responce and signal transduction. We conclude that automatic meta-analysis of proteomic papers enabled extraction of frequently reported proteins that are most likely the differentially expressed ones.

Key words: proteomics, meta-analysis, text mining, 2-DE, LC-MS/MS.