Биомедицинская химия, 2015 том 61, вып. 2, с. 169-175.

УДК 577.1
©Коллектив авторов

РОССИЯ В МЕЖДУНАРОДНОМ ПРОЕКТЕ “ПРОТЕОМ ЧЕЛОВЕКА”: ПЕРВЫЕ ИТОГИ И ПЕРСПЕКТИВЫ

Е.А. Пономаренко*, В.Г. Згода, А.Т. Копылов, Е.В. Поверенная, Е.В. Ильгисонис, А.В. Лисица, А.И. Арчаков

Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича, 119121, Москва, ул. Погодинская, 10; эл. почта: 2463731@gmail.com

Статья обобщает достижения пилотной фазы (2010-2014 гг.) российской части международного проекта “Протеом человека” и определяет направления дальнейших работ по исследованию протеома, кодируемого генами хромосомы 18 человека (2015-2022 гг.).

Пилотная фаза проекта была сфокусирована на детектировании как минимум одного белка для каждого белок-кодирующего гена хромосомы 18 в трёх типах биологического материала: плазме крови, клетках печени и клеточной линии HepG2. Использование масс-спектрометрической детекции белков методом мониторинга множественных реакций (ММР) и гено-центричного подхода позволило детектировать 95% мастерных (немодифицированных белков, содержащих канонические аминокислотные последовательности) форм белков, из которых для 60% получена оценка количественного содержания белка хотя бы в одном типе биологического материала.

Задачей основной фазы проекта является определение размеров протеома здорового человека с учётом модифицированных форм белков; оно предусматривает как биоинформатическое предсказание количества видов белков, так и выборочное экспериментальное измерение отдельных форм. Работы основной фазы проекта сфокусированы на исследовании здорового (обследованного) человека, поскольку не определены диапазоны концентраций белков, соответствующие нормальному физиологическому состоянию. Отсутствие этих данных существенно осложняет интерпретацию протеомных профилей крови пациентов, что препятствует созданию диагностических тестов.

В долгосрочной перспективе реализация проекта предусматривает создание на основе ММР диагностической тест-системы для количественного измерения форм белков, ассоциированных с развитием заболеваний. Создание подобных тест-систем позволит предсказывать степень риска возникновения тех или иных заболеваний, диагностировать заболевания на ранних стадиях и проводить мониторинг эффективности лечения.

Ключевые слова: протеом, протеомика, транскриптомика, масс-спектрометрия.

DOI: 10.18097/PBMC20156102169

ВВЕДЕНИЕ

Продолжением проекта “Геном человека” [1] стало начало в 2010 году международного проекта “Протеом человека” [2]. Россия не участвовала в выполнении геномного проекта, однако, была одной из стран-инициаторов международного протеомного проекта [3] с еще более амбициозной целью: измерить содержание продуктов экспрессии генов (белков) в организме человека [4].

Проект “Протеом человека” значительно более масштабен, чем завершённый проект “Геном человека”: если геном человека состоит из порядка 20 тыс. генов [5], то количество белков в организме человека за счёт несинонимичных однонуклеотидных замен и процессов альтернативного сплайсинга транскриптов соответствующих генов, а также посттрансляционных модификаций самих белков, превышает 2 млн. [6]. Кроме того, в отличие от генома, протеом меняется в зависимости от влияния различных факторов [7] и представляет собой сиюминутную совокупность белков в конкретном биологическом объекте и в определенной ситуации.

Среди принятых международным сообществом способов исследования протеома, одним из наиболее успешно развивающихся стал хромосомо- или геноцентричный подход [8], при котором результаты анализа белков картируются на соответствующие им белок-кодирующие гены. Сужение научной задачи за счёт геноцентричности делает её более реалистичной по сравнению с альтернативными подходами, позволяя сконцентрировать усилия на белках, кодируемых генами определённой хромосомы. В рамках выполнения российской части проекта хромосомоцентричный

* - адресат для переписки

подход комбинировали с направленным масс-спектрометрическим анализом белков.

Схема проведения масс-спектрометрического анализа (см. рис. 1) включает два основных компонента: регистрацию масс-спектров синтетических протеотипических пептидов для создания эталонной библиотеки и ВЭЖХ анализ с масс-спектрометрической детекцией трипсинолизированных проб биологических образцов. Для формирования перечня детектированных в биоматериале белков, полученные спектры сравнивают с эталонной библиотекой.

Обнаруженный методами масс-спектрометрии белок обозначают как мастерный, то есть соответствующий канонической аминокислотной последовательности без пострансляционных модификаций, сплайс-вариантов и мутаций. Совокупность полученных результатов масс-спектрометрической детекции и измерения концентрации каждого мастерного белка составляют сведения о мастерном протеоме выбранного типа биологического материала.

Цель российской части проекта – измерение протеома, кодируемого генами хромосомы 18 человека, в плазме крови, клетках печени и линии HepG2 на уровне чувствительности 10^{-18} М, соответствующий 1 копии белка (пептида) на 1 мкл плазмы крови или на 10^6-10^7 клеток печени или клеток HepG2 [9].

Для выбора хромосомы проводили сравнительный анализ всех хромосом человека по ряду параметров, таких, как количество белок-кодирующих генов, их связь с развитием заболеваний, степень изученности и др. Результаты показали примерно одинаковое распределение диагностически значимых генов в геноме [10], поэтому было решено сфокусироваться на исследовании протеома, кодируемого генами хромосомы 18, содержащей около трёхсот белок-кодирующих генов (http://www.uniprot.org/docs/humchr18). Основными заболеваниями, ассоциированными с генами выбранной хромосомы, являются колоректальный рак, лимфома, диабет и амилоидная нейропатия [11, 12].

На момент старта проекта в 2010 году, информационные ресурсы содержали сведения о локализации 285 белок-кодирующих генов на хромосоме 18 человека ([9]). Поскольку данные постоянно уточняются, то сейчас с выбранной хромосомой ассоциировано всего 276 белок-кодирующих генов [13] (см. рис. 2), из которых, согласно NextProt/UniProt, статус “protein level” (экспрессия подтверждена на протеомном уровне) имеют лишь 237 (в начале проекта – 197), экспрессия подтверждена на транскриптомном уровне для 92% генов (согласно БД RNAseqAtlas (http://medicalgenomics.org/rna_seq_atlas)), на протеомном при детекции антителами – для 67% генов (согласно БД HumanProteinAtlas (http://www.proteinatlas.org/)), а по результатам масс-спектрометрических исследований – для 77% генов (согласно БД GPMdb (http://gpmdb.thegpm.org/) и PeptideAtlas (http://www.peptideatlas.org/)). В протеомном репозитории PRIDE (http://www.ebi.ac.uk/pride/archive/) на сегодняшний день содержится информация о детекции 275 из 276 белок-кодирующих генов,

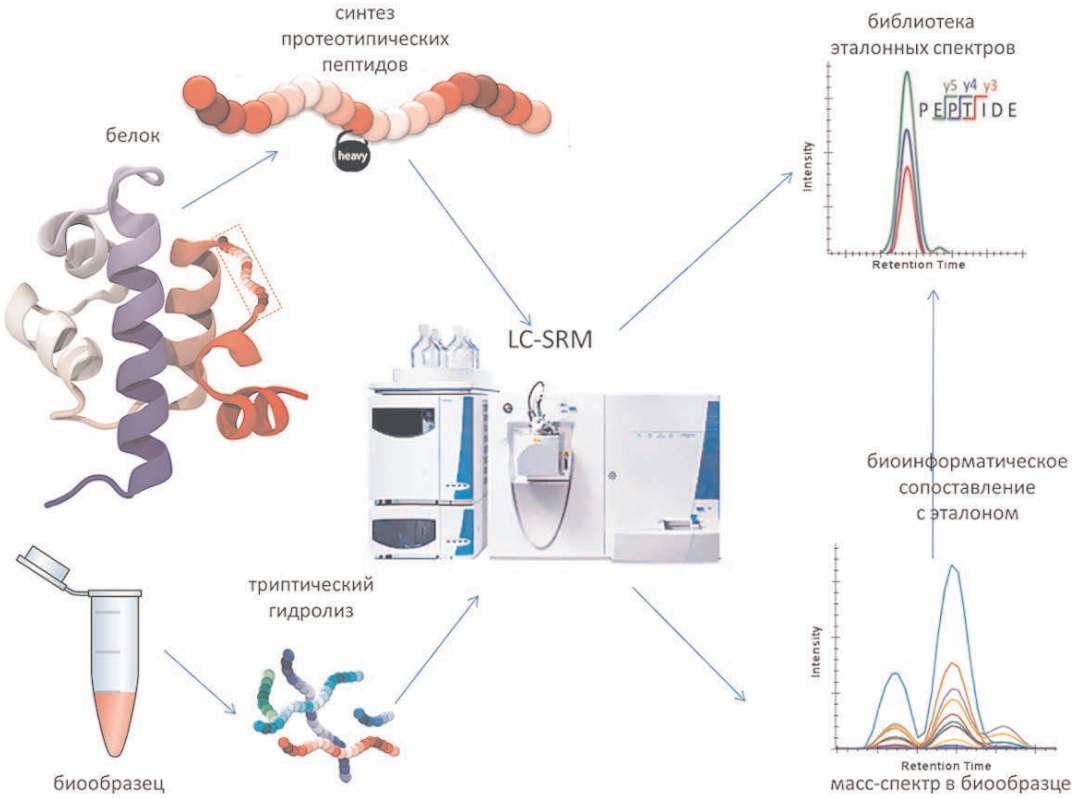

Рисунок 1. Схема проведения масс-спектрометрического анализа.

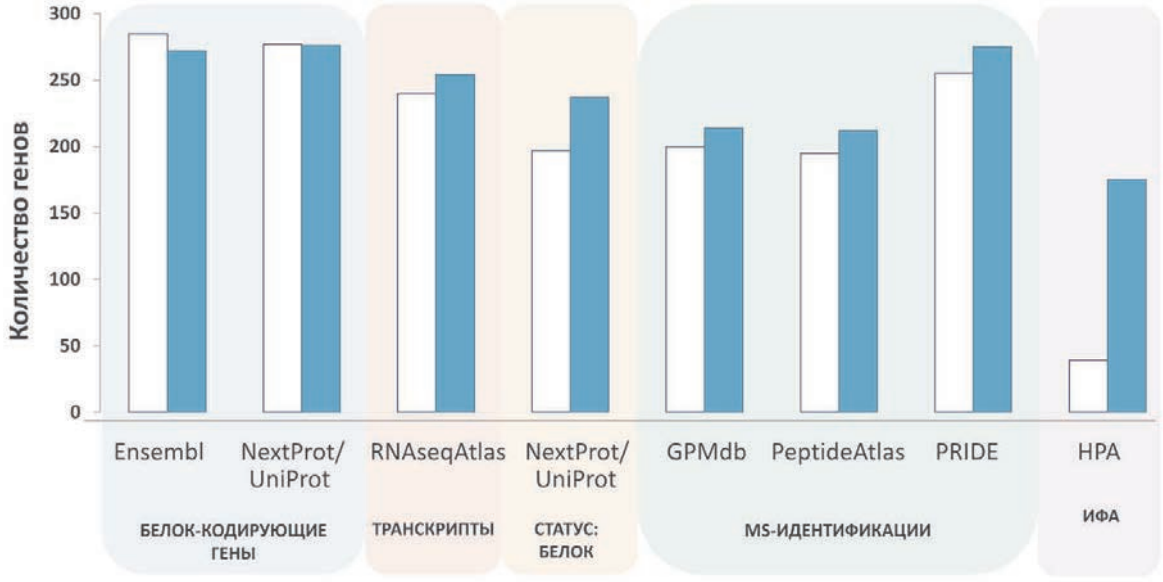

Рисунок 2. Сводная информация о хромосоме 18 (данные на начало проекта - 2010 год - отмечены белым, на 2015 год - цветом).

локализованных на хромосоме 18 человека. Однако для данных сведений отсутствуют критерии надёжности идентификации. Так, для панорамной масс-спектрометрии рекомендовано использование для идентификации белка минимум двух пептидов (FDR<1%), при этом среди перечня пептидов по результатам идентификации отбирают только наиболее вероятные результаты. Например, при использовании программы MASCOT рекомендованный уровень отсечения равен $p<0,05$. Требования к результатам, полученным с использованием направленной масс-спектрометрии, включают проведение надёжной идентификация белка минимум по двум пептидам, желательно с использованием изотопно-меченных пептидов), соответствующие принятым мировым стандартам [14].

К настоящему моменту среди белок-кодирующих генов, локализованных на хромосоме 18 человека, для ~20% генов нет сведений в информационных ресурсах об их идентификации на протеомном уровне; при этом для половины из них вообще отсутствуют сведения об экспрессии. По данным ресурса HumanProteinAtlas (http://www.proteinatlas.org/humanproteome/proteinevidence), аналогичная ситуация наблюдается и для других хромосом.

1. ТРАНСКРИПТОМНОЕ ПРОФИЛИРОВАНИЕ

Транскриптомное профилирование ткани печени и клеточной линии HepG2 проводили с использованием трёх методов: секвенирование на платформах Illumina и SOLiD, а также количественный анализ содержания транскриптов в клетке методом qRT-PCR [13, 15].

В клетках ткани печени было найдено 259 транскриптов, соответствующих генам хромосомы 18, при этом, наибольшее число транскриптов детектировано с использованием метода количественной полимеразной цепной реакции в реальном времени qRT-PCR (рис. 3а). Объединение результатов, полученных на клеточной линии HepG2, при помощи методов Illumina, SOLiD и qRT-PCR позволило получить информацию о содержании 252 транскриптов, из которых более 70% зарегистрированы независимо каждым из используемых аналитических методов (рис. 3б).

Применение комбинированного транскриптомного анализа с использованием двух платформ транскриптомного секвенирования и метода qRT-PCR позволило идентифицировать более 97% транскриптов для белок-кодирующих генов хромосомы 18: суммарно в обоих исследуемых типах биологического материала было найдено 260 транскриптов (рис. 3в).

2. ПРОТЕОМНОЕ ПРОФИЛИРОВАНИЕ

В трёх типах биологического материала (плазме крови, клетках печени и клеточной линии HepG2) было детектировано 268 белков и измерено их содержание [13, 15], что составляет 95% от всего мастерного протеома хромосомы 18 [16]. Около 60% всех детектированных белков были измерены по двум протеотипическим пептидам (то есть пептидам, соответствующим только одной аминокислотной последовательности белка для данного организма и пригодными для масс-спектрометрии), с разницей количественного содержания в биологическом материале не более чем в 2 раза. В каждом из трёх выбранных типов биологического материала детектировано сопоставимое количество мастерных белков, а пересечение между ними составляет около 96% (258 белков из 268 найденных (http://kb18.ru/protein/matrix).

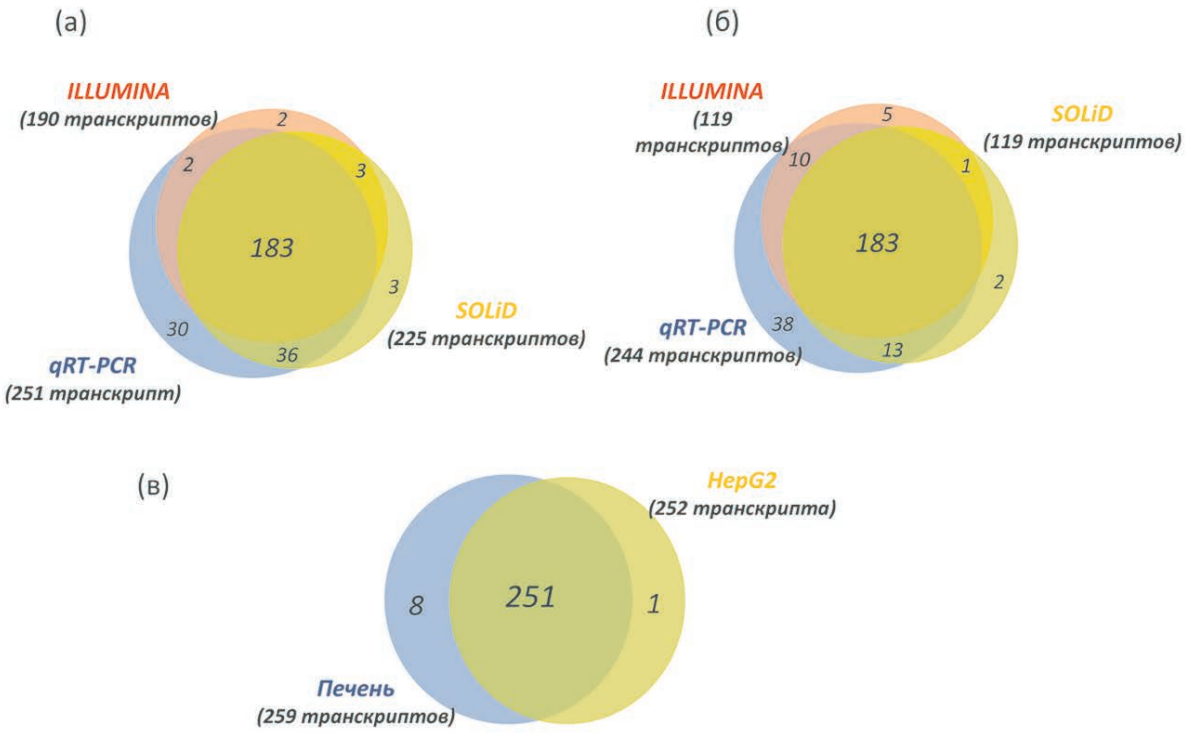

Рисунок 3. Результаты транскриптомного профилирования продуктов экспрессии генов хромосомы 18 в клетках ткани печени и клеточной линии HepG2 с использованием платформ SOLiD, Illumina и метода qRT-PCR. (а) - Транскрипты, детектированные в клетках ткани печени; (б) - транскрипты, детектированные в клетках линии HepG2; (в) - пересечение транскриптов, детектированных (всеми методами) в клетках ткани печени и клетках линии HepG2.

Из общего числа мастерных белков хромосомы 18 человека не было обнаружено ни в одном из типов биологического материала только восемь белков. Пять из них, по данным ресурсов PlasmaProteome Database и PRIDE, присутствуют в плазме крови; один из этих белков (SERPINB10) упомянут в обзорных работах Farrah и соавт. [17, 18]. Кроме того, экспрессия генов, кодирующих три белка, подтверждена в результате транскриптомного профилирования клеток ткани печени и клеточной линии HepG2, выполненного в рамках данной работы. Это свидетельствует об экспрессии кодирующих данные белки генов в клетках ткани печени и клеточной линии HepG2. Вероятно, отсутствие детекции на протеомном уровне может быть объяснено недостаточной чувствительностью аналитического метода.

На рисунке 4 приведены гистограммы распределения полученных концентраций белков, кодируемых генами хромосомы 18, выраженных в копиях молекул белков в микролитре плазмы или в одной клетке. Как видно из рисунка 4, распределение концентрации белков носит колоколообразный характер, с максимумом на уровне 10^8 копий на 1 мкл для плазмы крови человека и 10^5 копий на клетку печени или клеточной линии HepG2. При этом, в плазме крови более представлены высоко- и средне-копийные белки по сравнению с клетками ткани печени и клеточной линии HepG2. В рамках работы был разработан метод для детекции и измерения белков в области низких (10^{-12}-10^{-14} М) и ультранизких (10^{-15}-10^{-18} М) концентраций путём их необратимого связывания на биогранулах (bio-beads) [19]. Интересно отметить, что количество детектированных в клетках печени и клеточной линии HepG2 с применение усовершенствованного метода низко-копийных белков (отмечены заштрихованными областями) превышает число найденных низко-копийных белков в плазме крови.

Для хранения и обработки полученных результатов транскриптомного и протеомного профилирования продуктов экспрессии генов хромосомы 18 была создана геноцентричная база знаний (www.kb18.ru). Ресурс позволяет визуализировать сведения в формате тепловой карты [20], при котором характеристики каждого экспрессируемого с гена продукта закодированы цветом. Помимо представления данных в виде тепловой карты, позволяющего получать информацию о степени изученности белка, в системе реализована возможность поиска, сортировки и настройки цветокодировки данных, а также работы с выборками белков. Гибкая структура системы позволяет хранить данные не только в геноцентричном формате, но и создавать новые выборки с иным структурированием объектов: например, возможно хранение информации о пептидах или модифицированных вариантах белков.

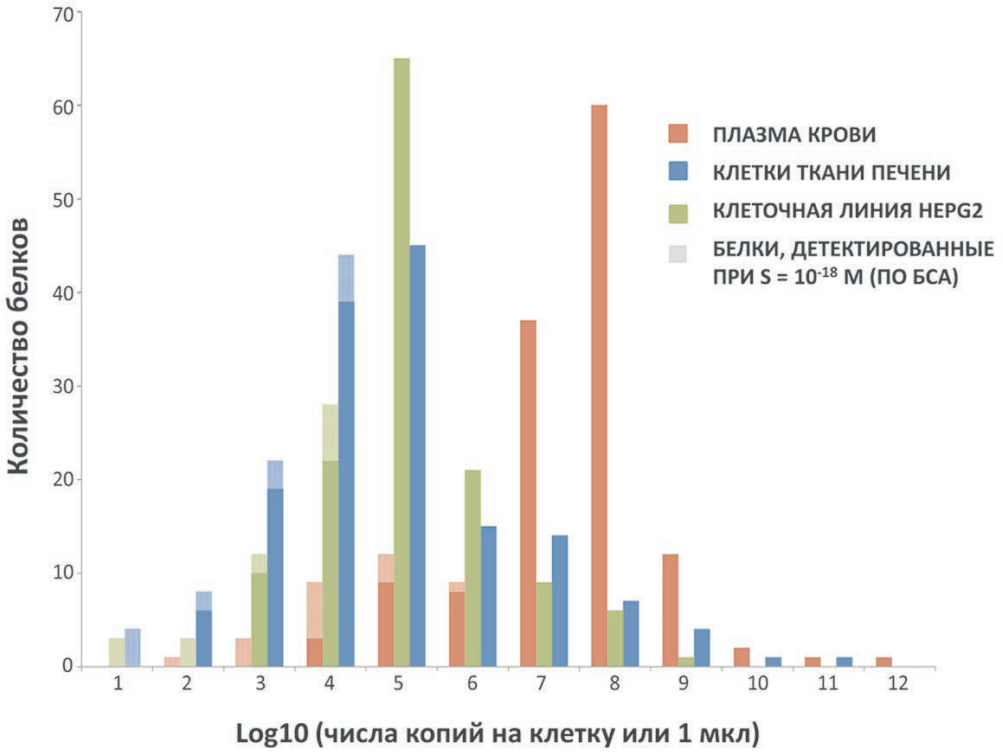

Рисунок 4. Распределение копийностей белков, кодируемых генами хромосомы 18 человека, плазме крови, клетках печени и гепатоцеллюлярной клеточной линии HepG2. В качестве примера распределение построено для белков, измеренных по 2-м пептидам с разницей между копийностью менее 1 порядка.

В системе депонированы данные из трёх основных источников: (1) внешние информационные ресурсы, такие как NextProt или UniProt; (2) собственные экспериментальные данные (например, результаты протеомного и транскриптомного профилирования (http://kb18.ru/protein/matrix/212020)); (3) результаты предсказаний, например, наличия модифицированых вариантов белков, локализации белка или сведения об интерактоме.

3. ЗАДАЧИ ОСНОВНОЙ ФАЗЫ ПРОЕКТА

Целью основной фазы проекта “Протеом человека” является дополнение мастерного протеома сведениями о формах белков, возникающих в результате различных модификаций – наличия несинонимичных одно-нуклеотидных полиморфизмов в геноме, процессов альтернативного сплайсинга и постранслящионных модификаций. Это означает, что вместо полученного на пилотной фазе мастерного протеома, когда одному гену соответствовал только один (мастерный) белок, в основной фазе проекта будет проанализировано всё многообразие возможных протеоформ, кодируемых одним геном. Принимая во внимание, что один ген может кодировать до 100 белков [21] (то есть несколько миллионов протеоформ в масштабах организма [22]) или несколько десятков тысяч для хромосомы 18 [15], существующие экспериментальные методы мало пригодны для решения этой задачи: так, на детектирование 300 белков в ходе пилотной фазы проекта ушло 3 года. Очевидно, что проверить фактическое существование каждой формы таргетным (направленным) масс-спектрометрическим методом нереально. Потому первый этап осуществления основной фазы проекта связан с развитием биоинформатических методов предсказания протеоформ. К этим методам относится, например, предсказание вариантов нуклеотидной последовательности на основе данных расшифровки геномов. Сборка экзомных (то есть всей совокупности транскрибируемых на мРНК экзонов после удаления интронов) последовательностей даёт информацию о возможных изменениях в аминокислотной последовательности; однако, метода предсказывающего вероятность экспрессии белкового продукта с изменённой первичной структурой, сегодня не существует. Основная фаза проекта “Протеом человека” позволит осуществить критическую оценку существующих биоинформатических предсказательных алгоритмов и, возможно, разработать новые алгоритмы, опирающиеся на экспериментальные данные по детектированию неканонических протеоформ.

В основной фазе предстоит также уточнить количественные данные о содержании мастерных белков в биологическом материале. Это будет осуществляться с использованием изотопно-меченных стандартов, представляющих собой пептиды с “тяжёлым” аминокислотным остатком аргинина, лизина и лейцина (эти аминокислотные остатки чаще всего присутствуют в пептидах, полученных

в ходе трипсинолиза белков анализируемого биологического материала). При постановке эксперимента исходные концентрации изотопно-меченных стандартов известны заранее, а поскольку площадь хроматографического пика пропорционально концентрации пептида, то по соотношению площади хроматографических пиков нативного и изотопно-меченного протеотипического пептида возможно оценить концентрацию пептида, а соответственно, и белка в биоматериале. Данные эксперименты будут выполнены на образцах плазмы крови здоровых, обследованных врачебной комиссией Института медико-биологических проблем, добровольцев. Будут получены данные о естественной изменчивости содержания белков в крови здорового человека, и определены диапазоны концентраций, соответствующих нормальному физиологическому состоянию. Сравнение полученных данных о концентрации белков в группе здоровых добровольцев (кандидатов в космонавты) и пациентов с диагностированными заболеваниями позволит выделить две основные группы белков плазмы крови: (1) группу консервативных белков, содержание которых не варьирует или меняется незначительно как в норме, так и при патологии; (2) группу белков, концентрация которых сильно варьирует у здоровых людей; (3) группу белков, содержание которых в норме и патологии достоверно отличается.

В перспективе, сравнение полученных данных о концентрации белков позволит установить границы содержания белков в плазме крови для здорового человека и определить форму кривой, описывающей распределение числа видов белков и их копийности в плазме крови человека, что, в свою очередь, необходимо для развития диагностических методов.

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

За время выполнения Российской части проекта “Протеом человека” на примере продуктов экспрессии генов хромосомы 18 была показана возможность использования комбинирования геноцентричного подхода и масс-спектрометрических методов (МС и ММР) для идентификации и измерения содержания белков в биологических образцах. Предложенная стратегия открывает перспективы для исследования протеомов отдельных органов и тканей, что имеет принципиальное значение для создания новых диагностических систем.

Основным результатом проекта является определение размеров протеома плазмы крови здорового человека, а именно детекция всех видов белков (оценка ширины протеома) и измерение содержания каждого вида белка (измерение глубины протеома) в плазме крови здоровых добровольцев.

Результаты реализации проекта могут быть использованы при создании диагностической тест-системы для количественного измерения с использованием ММР форм белков, ассоциированных с развитием заболеваний. Создание подобных тест-систем позволит в будущем предсказывать степень риска возникновения тех или иных заболеваний, диагностировать ранние стадии развития заболеваний и проводить мониторинг эффективности лечения.

ЛИТЕРАТУРА

1. *Collins F.S., Lander E.S., Rogers J., Waterston R.H.* (2005) Nature, **50**, 162–168.
2. *Legrain P., Aebersold R., Archakov A., Bairoch A., Bala K., Beretta L., Bergeron J., Borchers C.H., Corthals G.L., Costello C.E. et al.* (2011) Mol. Cell. Proteomics, **10**, M111.009993.
3. *Archakov A., Bergeron J.J.M., Khlunov A., Lisitsa A., Paik Y.* (2009) Mol. Cell. Proteomics., **8**, 2199–2200.
4. *Omenn G.S.* (2014) J. Proteomics, **100**, 3–7.
5. *Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A. et al.* (2001) Science (New York), **291**, 1304-1351.
6. *Archakov A., Ivanov Y., Lisitsa A., Zgoda V.* (2009) Proteomics, **9**,1326-1343.
7. *Archakov A., Zgoda V., Kopylov A., Naryzhny S., Chernobrovkin A., Ponomarenko E., Lisitsa A.* (2012) Expert Rev. Proteomics, **9**, 667–676.
8. *Paik Y.-K., Omenn G.S., Uhlen M., Hanash S., Marko-Varga G., Aebersold R., Bairoch A., Yamamoto T., Legrain P., Lee H.-J. et al.* (2012) J. Proteome Res., **11**, 2005-2013. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22443261
9. *Archakov A., Aseev A., Bykov V., Grigoriev A., Govorun V., Ivanov V., Khlunov A., Lisitsa A., Mazurenko S., Makarov A.A., Ponomarenko E., Sagdeev R., Skryabin K.* (2011) Proteomics, **11**, 1853–1856.
10. *Ponomarenko E., Poverennaya E., Pyatnitskiy M., Lisitsa A., Moshkovskii S., Ilgisonis E., Chernobrovkin A., Archakov A.* (2012) Omics: Journal of Integrative Biology, **16**, 604–611.
11. *Zarzour P., Boelen L., Luciani F., Beck D., Sakthianandeswaren A., Mouradov D., Sieber O.M., Hawkins N.J., Hesson L.B., Ward L.R., Wong H.J.* (2015) Genes, Chromosomes, Cancer.
12. *McDonough C.W., Bostrom M.A., Lu L., Hicks P.J., Langefeld C.D., Divers J., Mychaleckyj J.C., Freedman B.I., Bowden D.W.* (2009) Human Genetics, **126**, 805–817.
13. *Ponomarenko E.A., Kopylov A.T., Lisitsa A.V., Radko S.P., Kiseleva Y.Y., Kurbatov L.K., Ptitsyn K.G., Tikhonova O.V., Moisa A.A., Novikova S.E. et al.* (2014) J. Proteome Res., **13**, 183–190.
14. *Lane L., Bairoch A., Beavis R.C., Deutsch E.W., Gaudet P., Lundberg E., Omenn G.S.* (2015) **13**, 15–20.
15. *Zgoda V.G., Kopylov A.T., Tikhonova O.V., Moisa A.A., Pyndyk N.V., Farafonova T.E., Novikova S.E., Lisitsa A.V., Ponomarenko E.A., Poverennaya E.V. et al.* (2013) J. Proteome Res., **12**, 123–134.
16. *Archakov A., Lisitsa A., Ponomarenko E., Zgoda V.* (2015) Expert Rev. Proteomics, **12**, 111-113.
17. *Farrah T., Deutsch E.W., Omenn G.S., Sun Z., Watts J.D., Yamamoto T., Shteynberg D., Harris M.M., Moritz R.L.* (2014) J. Proteome Res., **13**, 60–75.
18. *Liu X., Valentine S.J., Plasencia M.D., Trimpin S., Naylor S., Clemmer D.E.* (2007) J. Amer. Soc. Mass Spectrom., **18**, 1249–1264.
19. *Kopylov A.T., Zgoda V.G., Lisitsa A.V., Archakov A.I.* (2013) Proteomics, **13**, 727–742.

20. *Poverennaya E.V., Bogolubova N.A., Bylko N.N., Ponomarenko E.A., Lisitsa A.V., Archakov A.I.* (2014) Biochim. Biophys Acta: Proteins and Proteomics, **1844**, 77–81.
21 *Naryzhny S.N., Lisitsa A.V., Zgoda V.G., Ponomarenko E.A., Archakov A.I.* (2013) Electrophoresis, **35**, 1–20.
22. *Kelleher N.L.* (2012) J. Amer. Soc. Mass Spectrom., **23**, 1617–1624.

Поступила: 16. 02. 2015.

THE RUSSIAN PART OF THE HUMAN PROTEOME PROJECT: FIRST RESULTS AND PROSPECTS

E.A. Ponomarenko, V.G. Zgoda, A.T. Kopylov, E.V. Poverennaya, E.V. Ilgisonis, A.V. Lisitsa, A.I. Archakov

Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; e-mail: 2463731@gmail.com

The article summarizes the achievements of the pilot phase (2010-2014) of the Russian part of the international project “Human Proteome” and identifies the directions for further work on the study of the human chromosome 18 proteome in the framework of the project main phase (2015-2022).

The pilot phase of the project was focused on the detection of at least one protein for each chromosome 18 protein-coding gene in three types of the biological material. The application of mass spectrometric detection of proteins by the methods of multiple reactions monitoring (MRM) and gene-centric approach made it possible to detect 95% of master forms of proteins, for 60% of which the quantitative assessment of the protein content was obtained in at least one type of the biological material.

The task of the main phase of the project is to measure the proteome size of healthy individuals, taking into account the modified protein forms, providing for both the bioinformatics prediction of the quantity of proteins types and the selective experimental measurement of single proteoforms.

Since the ranges of protein concentrations corresponding to the normal physiological state have not been identified, the work of the main phase of the project is focused on the study of clinically healthy individuals. The absence of these data complicates significantly the interpretation of the patients’ blood proteomic profiles and prevents creating diagnostic tests.

In the long term prospect, implementation of the project envisages development of a diagnostic test system based on multiple reactions monitoring (MRM) for quantitative measurement of the protein forms associated with some diseases.

Development of such test systems will allow predicting the extent of risk of different diseases, diagnosing a disease at its early stage and monitoring the effectiveness of the treatment.

Key words: proteome, proteomics, transcriptomics, mass-spectrometry.