«Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014 – 2020 годы»

<по этапу № 1 >

Номер Соглашения о предоставлении субсидии: 14.576.21.0073

Тема: «Разработка Интернет-доступного сервиса поисковой протеомной машины для идентификации белков живых организмов».

Приоритетное направление: науки о жизни.

Критическая технология: геномные, протеомные и постгеномные технологии

Период выполнения: 05.11.2014-31.12.2015г.

Плановое финансирование проекта: 7 млн. руб.

Бюджетные средства 3,5 млн. руб.,

Внебюджетные средства 3,5 млн. руб.

Исполнитель: Общество с ограниченной ответственностью «КуБ».

Ключевые слова: протеомика, протеом человека, биоинформатика, хромато-масс-спектрометрия, белки, пептиды.

Цель прикладного научного исследования и экспериментальной разработки

Основной задачей проекта является разработка Интернет-доступного сервиса для хранения и обработки данных хромато-масс-спектрометрических измерений, полученных в рамках высокопроизводительного панорамного протеомного анализа на содержание белков в клинических биопробах, включая экстракты клеток тканей и физиологических жидкостей. В основе сервиса – протеомная поисковая машина открытого кода на языке программирования с динамической типизацией Питон (Python), предназначенная для:

а) автоматизированной обработки массивных данных хромато-масс-спектрометрических измерений, проводимых в рамках глубокого панорамного протеомного анализа биологических проб;

б) идентификации пептидов и белков в анализируемых пробах, для которых были проведены хромато-масс-спектрометрические измерения, с контролируемым пользователем уровнем достоверности;

в) отчета результатов поиска для пользователя в стандартизированных форматах, рекомендованных международным консорциумом глобального проекта Протеом Человека (HPP).

1. Основные результаты проекта

Для решения основной задачи проекта планируется:

а) разработать протеомную поисковую машину открытого кода доступа (далее – «протеомная машина») на высокоуровневом языке с динамической типизацией, позволяющая повысить эффективность (по числу достоверно определяемых компонент анализируемой смеси белков) и чувствительность поиска (динамический диапазон достоверно определяемых компонент анализируемой смеси белков), по сравнению с существующими коммерческими решениями;

б) интегрировать в едином программно-аналитическом комплексе основные составляющие системы обработки и анализа хромато-масс-спектрометрических протеомных данных: (1) первичная обработка экспериментальных данных, представленных в стандартизированных форматах; (2) поиск по геномным базам данных и идентификация белков; (3) подтверждение (валидация) полученных идентификаций; (4) количественный анализ полученных идентификаций белков.

В ходе реализации проекта будут разработаны новые принципы обработки данных масс-спектрометрических измерений, поиск последовательностей белковых кандидатов, ранжирова-

ние полученных хитов и их подтверждение с использованием всей совокупности получаемой экспериментальной информации.

В результате реализации проекта в 2014 г. были получены следующие основные результаты:

1. В рамках работ по созданию программного обеспечения для протеомной поисковой машины была разработана специализированная модульная библиотека парсеров и функций, MSMS_pySearchLib, на языке программирования открытого кода с динамической типизацией Питон (Python), а также локализованная техническая документация к ней. Библиотека подпрограмм и функций MSMS_pySearchLib предназначена для разработчиков прикладного программного обеспечения, используемого в протеомных исследованиях для обработки результатов хромато-масс-спектрометрических измерений, поиска пептидных идентификаций в пробах на основе протеомных баз данных, обработки результатов поиска и их представления в стандартизованных форматах. Целевым прикладным программным обеспечением библиотеки MSMS_pySearchLib являются протеомные машины. Библиотека состоит из 6-ти базовых модулей, разделенных по основным функциональным стадиям работы протеомной машины по обработке первичных данных, идентификации пептидов и белков и представления результатов идентификации в стандартизированных форматах. Библиотека MSMS_pySearchLib является первой библиотекой парсеров и функций, написанной на высокоуровневом языке программирования с динамической типизацией и специализированной для разработки протеомных машин, используемых для обработки результатов панорамного хромато-масс-спектрометрического анализа сложных белковых смесей. Поиск основан на идентификации пептидов и белков в протеомных базах данных пептидных кандидатов на измеряемые прекурсорные ионы в масс-спектрах первого уровня и сравнения соответствующих пептидным кандидатам теоретических масс-спектров второго уровня с экспериментально измеряемым для прекурсорных ионов, а также оценки достоверности полученных идентификаций и составления списка пептидов, присутствовавших в анализируемой пробе, для заданного уровня ложно-положительных идентификаций. Созданная библиотека парсеров и функций содержит все необходимые элементы для разработки протеомной машины на ее основе, а заложенные в библиотеку функциональности полностью соответствуют техническим требованиям разрабатываемой на следующем этапе работ протеомной машины, включая: (1) функции парсинга первичных данных в стандартных форматах; (2) работы со стандартными форматами представления протеомных (белковых) баз данных; (3) расчета изотопных масс, масс фрагментов, масс протеолитических пептидов для любых задаваемых пользователем правил гидролиза; (4) расчет хроматографических времен с использованием различных хроматографических моделей, включая BioLCCC; (5) расчета индексов достоверности пептидных идентификаций; и (6) представления результатов работы ППМ в стандартном формате pepXML.

2. В рамках работ по выбору наиболее оптимальных алгоритмических решений для обработки результатов протеомного анализа и получения идентификаций на основе сопоставления масс-спектров второго уровня для пептидов протеомной базы данных с экспериментальными был проведен сравнительный анализ работы трех наиболее широко используемых протеомных поисковых машин, Mascot, X!Tandem и OMSSA. Для проведения сравнительного анализа были получены экспериментальные данные для смесей стандартов рекомбинантных белков и клинических проб. Основное сравнение существующих поисковых алгоритмов осуществлялось на данных панорамного хромато-масс-спектрометрического анализа протеома клеточной линии рака, в котором было получено 45 000 масс-спектров второго уровня МС/МС для идентифика-

ции экспрессированных белков. В результатах измерений в каждом техническом повторе идентифицировалось более 6000 пептидов. Проведенное исследование выявило несколько критических входных параметров, задаваемых пользователем при настройке протеомной машины, которые оказывают наиболее существенное влияние на количество идентифицируемых пептидов в одних и тех же экспериментальных данных. К таким критическим параметрам относятся, в первую очередь, диапазон разрешенных к поиску масс в масс-спектрах. В частности, было показано существование ограниченного диапазона разрешенных к поиску масс для масс-спектров второго уровня, вне рамок которого количество идентификаций многократно уменьшается вплоть до нулевых значений. Также интересным наблюдением является высокий уровень толерантности количества идентификаций пептидов к диапазону допустимых к поиску масс в масс-спектрах первого уровня. Превышение в несколько порядков этим диапазоном уровня точности измерения масс в масс-спектрах первого уровня практически не оказывает влияния на количество идентифицируемых пептидов. Было показано различие сравниваемых протеомных поисковых машин в соответствующих зависимостях количества идентификаций от этих критических параметров. Кроме того, было обнаружено, что алгоритмы поиска, заложенные в работу сравниваемых протеомных машин, дают существенно отличающиеся результаты при работе с масс-спектрами второго уровня, измеренными с использованием неэргодичных методов диссоциации пептидов (в частности, был рассмотрен метод диссоциативного захвата электрона ECD). Еще одним принципиальным результатом сравнения явилось то, что коммерческое и наиболее широко используемое в настоящее время решение, Mascot, по эффективности работы не имеет существенных конкурентных преимуществ по сравнению с некоммерческим ПО, такими как X!Tandem. Следующим принципиальным выводом из результатов сравнения является вывод о необходимости автоматизированной настройки критических входных параметров работы протеомной машины. Следует отметить, что реализация такой возможности заложена в требования к разрабатываемой в рамках проекта протеомной машине (режим автооптимизации входных параметров). Наконец, еще одним принципиальным выводом из результатов сравнения является вывод о необходимости совмещения в рамках работы протеомной машины функции валидации результатов поиска (в настоящее время эта стадия протеомного анализа проводится отдельно) с целью повышения количества идентификаций за счет дополнительного поиска по спектрам фрагментации второго уровня МС/МС, которые по индексам достоверности совпадения с теоретическими спектрами пептидов базы данных попали ниже порогового, определяемого алгоритмом сравнения спектров протеомной машиной. В разрабатываемой в рамках проекта протеомной машины функция валидации идентификаций будет интегрирована в работу протеомной машины, для реализации чего в библиотеке парсеров и функций MSMS_pySearchLib был разработан специальный модуль.

2. Назначение и область применения результатов проекта

Предлагаемый к созданию Интернет-доступный сервис на основе протеомной поисковой машины, предназначен для определения белкового состава биообразцов и интерпретации результатов масс-спектрометрических измерений в следующих областях:

(1) лабораторной диагностики: количественное масс-спектрометрическое выявление белковых маркеров социально значимых заболеваний, которые могут быть использованы для ранней диагностики и мониторинга патологии, а также анализа эффективности фармакологического воздействия;

(2) биомедицинских исследований в области протеомики и постгеномных технологий: разработка методов глубокого панорамного (в широком диапазоне молекулярных масс и кон-

центраций) анализа белкового состав билогических проб и количественного измерения содержания целевых белков.

(3) развитие методов персонализированной медицины: выявление протеомных маркеров, включая как индивидуальные белки, так и белковые «сигнатуры», заболеваний человека, а также проявлений экспрессии генов (и/или мутаций генов) на протеомном уровне.

Успешная реализация проекта позволит уменьшить зависимость от зарубежных коммерческих биоинформационных продуктов отечественных иссследований в области протеомики, а также диагностических центров, осуществляющих анализ содержания белков в пробах в практике клинического анализа.

3. Эффекты от внедрения результатов проекта

Выявление и разработка новой протеомной поисковой машины для идентификации пептидов на основе имеющихся экспериментальных данных является серьезной и актуальной задачей в современных протеомных исследованиях. Новая поисковая машина должна не просто «дополнять» стандартные масс-спектрометрические решения, а предоставлять независимую оценку достоверности пептидных идентификаций. В рамках предлагаемого проекта планируется разработать универсальную протеомную поисковую машину, в которой будут интегрированы методы первичной обработки профильных хромато-масс-спектрометрических спектров, включая спектры фрагментации, методы извлечения комплементарной информации о последовательностях и аминокислотном составе идентифицируемых пептидов, включая хроматографические времена и изоэлектрические точки, и методы валидации полученных идентификаций на основе вероятностных алгоритмов.

В свою очередь, повышение достоверности и чувствительности (глубины) протеомного анализа позволит осуществлять направленный поиск белковых панелей, количественно характеризующих стадии развития социально-значимых заболеваний живых организмов, включая человека, а также раннего развития патологий на клеточном уровне. Так, например, сопоставление данных геномного секвенирования с данными глубокого протеомного анализа позволит решить проблему подтверждения обнаруженных экспрессий тех, или иных генов, связанных с развитием патологий, на белковом уровне. В свою очередь, выявление экспрессии белков на клеточном уровне позволит существенно продвинуть вперед понимание биологической функции белков и определение активных функциональных групп вдоль аминокислотной последовательности с целью разработки и оценки новых лекарств и фармацевтических продуктов более эффективно и с использованием систематического подхода. В этой связи, довольно трудно переоценить экономическую значимость развития протеомных технологий (в том числе, соответствующих биоинформационных ресурсов). Например, уже существующие результаты протеомных исследований позволяют в ряде случаев сократить на 30% расходы на разработку новых лекарственных препаратов. В среднем, эти расходы составляют, например, только в США порядка 500-700 млн долларов на один лекарственный препарат, предназначенный для лечения того, или иного, серьезного заболевания (данные на 2009 г.). Если учесть, что Агентство США по продуктам питания и лекарственным препаратам утверждает к производству и продаже около 100 новых лекарств в год, то эффективность результатов работ в области протеомики может составлять десятки миллиардов долларов в год. Одновременно, сокращаются не только расходы, но и время, требуемое на разработку лекарства от лабораторных исследований до использования в клинической практике.