# «Большие данные» в биологии и медицине

По материалам совместного семинара с представителями международной организации Data-Enabled Life Science Alliance, 4 июля 2013 года, Москва, Россия

О. П. Трифонова[1]*, В. А. Ильин[2,3], Е. В. Колкер[4,5], А. В. Лисица[1]

[1]Научно-исследовательский институт биомедицинской химии им. В.Н. Ореховича РАМН, 119121, Москва, ул. Погодинская, 10, стр. 8

[2]Научно-исследовательский центр «Курчатовский институт», 123182, Москва, пл. Академика Курчатова, 1

[3]Научно-исследовательский институт ядерной физики им. Д.В. Скобельцына Московского государственного университета им. М.В. Ломоносова, 119992, Москва, Ленинские горы, 1, стр. 58

[4] DELSA Global, США

[5] Seattle Children´s Research Institute, 1900 9th Ave Seattle, WA 98101, США

*E-mail: oxana.trifonova@gmail.com

Задача извлечения новых знаний из больших массивов данных обозначается понятием «большие данные» (Big Data). Говоря проще, Big Data – это когда результаты экспериментов не загружаются в таблицу Excel. Например, переписка в Twitter за год на порядки превосходит объем памяти человека, накопленной за всю жизнь. На фоне Twitter все данные о геномах людей представляются пренебрежимо малой величиной [1]. Вопрос о преобразовании массивов данных в знание, обозначенный в 2013 году системой Национальных институтов здоровья США, является приоритетом международного альянса DELSA (Data-Enabled Life Science Alliance, www.delsaglobal.org) [2].

Почему компьютерные вопросы сбора больших данных стимулировали образование сообщества DELSA, куда вошло более 80 ведущих ученых мира, работающих в области медицины, здравоохранения и прикладной информатики? Это новое течение обсуждали участники семинара «Конвергентные технологии: Big Data в биологии и медицине».

В работе семинара приняли участие 35 человек, среди которых были представители научно-исследовательских институтов, занимающиеся анализом больших объемов экспериментально полученных данных, и коммерческие компании, которые осуществляют разработку информационных систем. На семинаре было представлено 16 коротких докладов, целью которых была не столько демонстрация полученных результатов, сколько обсуждение, почему манипулирование большими массивами данных должно иметь отношение к вопросам медицины и здравоохранения.

От имени альянса DELSA Global (Data-Enabled Life Science Alliance, www.delsaglobal.org) открыл семинар проф. Евгений Колкер докладом о миссии данной организации. Альянс выступает за глобализацию биоинформационных подходов в естественных науках и создание научных сообществ в области «омных» наук. Главная идея – ускорить процесс трансляции результатов медико-биологических исследований для удовлетворения потребностей общества.

И без участия биологов во многих науках накапливаются большие массивы данных, которые надо хранить, обрабатывать и анализировать, и в этом нет ничего удивительного. Если говорить о больших объемах данных в области физики высоких энергий, то речь идет о десятках петабайтах, а если о медицине и биологии – то на порядок меньше, но тоже приближается к петабайту. В ходе семинара обсуждался вопрос, за что «зацепиться» российским ученым в мире Big Data – либо за молекулярную биологию в формате «омик», либо за интегративную биологию при моделировании мозга, либо вообще за социальные науки.

Задачи работы с большими объемами данных делятся на две группы: (1) когда данные поступают в режиме реального времени и требуют немедленной обработки и (2) когда есть много накопленных данных, требующих разносторонней интерпретации. Первый тип данных относится к коммерческим


От проблемы к решению: на вызов NIH готовы ответить специалисты в области обработки данных Data-Enabled Life Science Alliance – DELSA

системам, таким, как Google, Twitter и Facebook. Примером второго типа данных являются геномные и протеомные репозитории.

В Институте системного программирования РАН занимаются разработкой систем для работы с большими массивами слабоструктурированных и противоречивых данных, характерных для медико-биологических наук. Для реализации интеллектуальных методов поиска, хранения и анализа данных применяются наборы утилит, библиотек, а также распределенные программные каркасы (фреймворки), работающие на кластерах из сотен и тысяч узлов. Такие проекты, как Hadoop (http://hadoop.apache.org/), Data-Intensive Computing, NoSQL, используются для реализации интеллектуального поиска при работе высоконагруженных веб-сайтов.

Работа с массивами данных путем интеграции информации, полученной на различных организационных уровнях, составляет сущность принципиально новой дисциплины – коннектомики, о которой участникам семинара рассказал чл.-корр. РАН и РАМН Константин Анохин (НИЦ «Курчатовский институт»). Большие объемы данных будут накапливаться в области нейронаук из-за слияния двух основополагающих факторов. Во-первых, в области нейронаук накоплено огромное количество результатов, полученных с использованием аналитических методов с высоким разрешением. Во-вторых, основной интерес ученых вызывает не работа отдельных синапсов, а как работает целый мозг и как эта работа проецируется на систему – сознание, мысль, действие. Получение информации о работе мозга как системы включает в себя методы визуализации – томографию высокого разрешения, световую микроскопию, электронную микроскопию. Мегапроекты по моделированию мозга уже стартовали (например, Human Brain Project в Европе). Со временем вложения в процесс получения новых экспериментальных данных будут девальвироваться, а вот анализ полученных данных станет приоритетной задачей.

Экстракция и интерпретация информации из существующих баз данных за счет новых алгоритмов анализа играет определяющую роль в будущей науке. Наличие большого количества открытых источников информации, включая различные базы данных и поисковые системы, часто затрудняет поиск необходимой информации. По словам чл.-корр. РАМН Андрея Лисицы (Институт биомедицинской химии РАМН), существующие базы данных по интерактомике совпадают друг с другом максимум на 55% [3]. Задачей при работе с большими массивами становится получение непротиворечивой картины при сведении данных из разных источников воедино.

В Медико-генетическом научном центре РАМН интегрированный подход используют для поиска панелей биомаркеров по существующим геномным, транскриптомным и протеомным данным. Для описания нормальной ткани человека (условно-статистическое понятие нормы) требуется несколько тысяч показателей, которые будут различны для здоровых людей, и понятие усредненного показателя нормы некорректно отражает ситуацию. Проф. Анча Баранова (Медико-

генетический научный центр РАМН) предлагает использовать понятие многомерного пространства нормальной ткани и с помощью коэффициентов корреляции определять расстояние каждого анализируемого образца от условного центра – «идеального состояния». Больные ткани будут находиться дальше от центра, чем здоровые. Предложенный подход позволяет перейти от вероятностной диагностики (например, болен с вероятностью 30%) к измерению расстояния конкретного пациента от нормы.

В докладе д. ф.-м. н. Всеволода Макеева (Институт общей генетики РАН) утверждалось, что в ближайшем будущем мы будем все больше работать с большими массивами постгеномных данных. Эти данные будут двух типов – данные типа индивидуального генома (проект «1000 геномов»), которые получают один раз и хранят в базах данных, чтобы при необходимости скачать. Второй тип – это, например, результаты транскриптомного или протеомного анализа, который проводится регулярно для получения интегрированного «омикс»-профиля индивидуума [4]. В случае геномов существует несколько провайдеров данных; и российские лаборатории, пользуясь этими хранилищами и применяя свои биоинформатические подходы, могут получать новые результаты [5].

По мере внедрения современных методов анализа в клиническую практику будет расти поток динамических данных от индивидуума (результаты мониторинга параметров организма). Встанет необходимость быстрой обработки непрерывно получаемых данных и передачи информации в хранилища для последующего аннотирования и автоматической выработки решений. Это вызовет модификацию технологий хранения и передачи данных для более быстрого обмена информацией. Подобные сервисы для хранения и передачи больших объемов данных активно развиваются, например AmazonS3.

Большую роль в данном процессе играет, конечно, разработка более быстрых методов математического анализа. В докладе д. ф.-м. н. Ивана Оселедца (Институт вычислительной математики РАН) был рассмотрен математический аппарат компактного представления многомерных массивов на основе тензорных поездов (tensor train, TT-формат). В медико-биологических приложениях многомерные задачи возникают постоянно, а TT-формат позволяет выявить основные переменные, достаточные для описания исследуемой системы или процесса.

Медицинские данные требуют обработки в реальном времени, чтобы можно было поставить предварительный диагноз максимум через несколько минут. Компанией «Прогресс» разрабатывается система удаленного мониторинга медицинских показателей с использованием мобильных устройств и сотовой сети передачи данных (Telehealth, доклад Олега Гашникова). Данный метод позволяет круглосуточно наблюдать за пациентом за пределами стационара, что в будущем должно сократить расходы на медицинское обслуживание. На данном этапе требуется разработка методик формирования паттернов тревог на основе накопленных данных и адаптация алгоритмов под конкретного пациента.

На первый взгляд в стороне от темы семинара прозвучал доклад о проблеме сбора и обработки данных геопозиционирования, которые накапливаются операторами сотовой связи, а также собираются агрегаторами типа Google, Facebook и AlterGeo с использованием современных приложений для смартфонов. Докладчик Артем Вольфтруб (ведущий разработчик ООО «Грамант») сообщил, что с 2009 года появилась серия публикаций группы Alex Pentland и David Laser из Массачусетского технологического института, в которых обосновывается, что анализ геоданных может быть ничуть не менее информативен для прогноза социально значимых заболеваний, чем геном. В патогенезе мультигенного заболевания значительную роль играют факторы внешней среды, так называемый экспозом. Получить информацию об экспозоме с достаточной степенью детализации можно анализируя перемещения человека, сравнивая общие закономерности перемещений для популяций и выявляя паттерны, коррелирующие с рисками для здоровья, например с развитием сердечно-сосудистых заболеваний или ожирения [6].

В дискуссиях участников семинара в разных контекстах упоминался суперкомпьютер «Ватсон», созданный компанией IBM для ответа на вопросы (теоретически на любые!), сформулированные на естественном языке. Это один из первых примеров экспертных систем, функционирующих по принципу Big Data. В 2011 году было объявлено о применении этого суперкомпьютера для обработки слабоструктурированных массивов данных для решения задач медицины и здравоохранения [7].

Анализируя проблему больших данных в области биологии и медицины, следует отметить, что для этих наук еще со времен натурфилософии характерно накопление больших массивов информации, фиксирующей результаты наблюдений. В геномную эру накопление информации осуществляли, как казалось, с понятной целью. Однако, когда технический аспект был решен и геном расшифровали, оказалось, что собственно к вопросам сохранения

здоровья эти данные имеют слабое отношение [8].

В постгеномный период мировая биомедицинская наука вернулась на уровень феноменологического описания, которое направлено на сбор данных без однозначной перспективы их дальнейшей интерпретации. Таков проект «Протеом человека» – описательный атлас, где количественные данные собираются по каждому белку, однако неясно, как применять эти результаты к прикладным задачам лабораторной диагностики. Аналогичен и проект «Коннектом», целью которого является накопление данных о передаче сигналов между нейронами в ожидании, что скопившись до какого-то критического уровня эти данные позволят сымитировать в компьютере работу человеческого мозга.

Подводя итоги семинара, участники отметили, что феномен Big Data связан с открывшейся возможностью современной техногенной сферы генерировать и хранить данные, при этом четкое понимание для чего собирать и хранить эти данные отсутствует. Российским ученым следует в первую очередь сосредоточиться на анализе Big Data, чтобы превратить массив данных в гипотезы, пригодные для проверки точечным биохимическим экспериментом. В качестве основного направления развития российского отделения DELSA следует наметить задачу ознакомления с данными проекта «Коннектом». ●

СПИСОК ЛИТЕРАТУРЫ

1. Hesla L. Particle physics tames big data // Symmetry. August 01, 2012. (http://www.symmetrymagazine.org/article/august-2012/particle-physics-tames-big-data)
2. Kolker E., Stewart E., Ozdemir V. // OMICS. 2012. V. 3. № 16. P. 138–147.
3. Lehne B., Schlitt T. // Human Genomics. 2009. № 3. P. 291–297.
4. Li-Pook-Than J., Snyder M. // Chemistry & Biology. 2013. № 20. P. 660–666.
5. Tsoy O.V., Pyatnitskiy M.A., Kazanov M.D., Gelfand M.S. // BMC Evolutionary Biology. 2012. № 12. (doi: 10.1186/1471-2148-12-200)
6. Pentland A., Lazer D., Brewer D., Heibeck T. // Studies in Health Technology and Informatics. 2009. № 149. P. 93–102.
7. Wakeman N. IBM's Watson heads to medical school. Washington Technology. February 17, 2011. (http://washingtontechnology.com/articles/2011/02/17/ibm-watson-next-steps.aspx)
8. Bentley D.R. // Nature. 2004. V. 429. № 6990. P. 440–445.