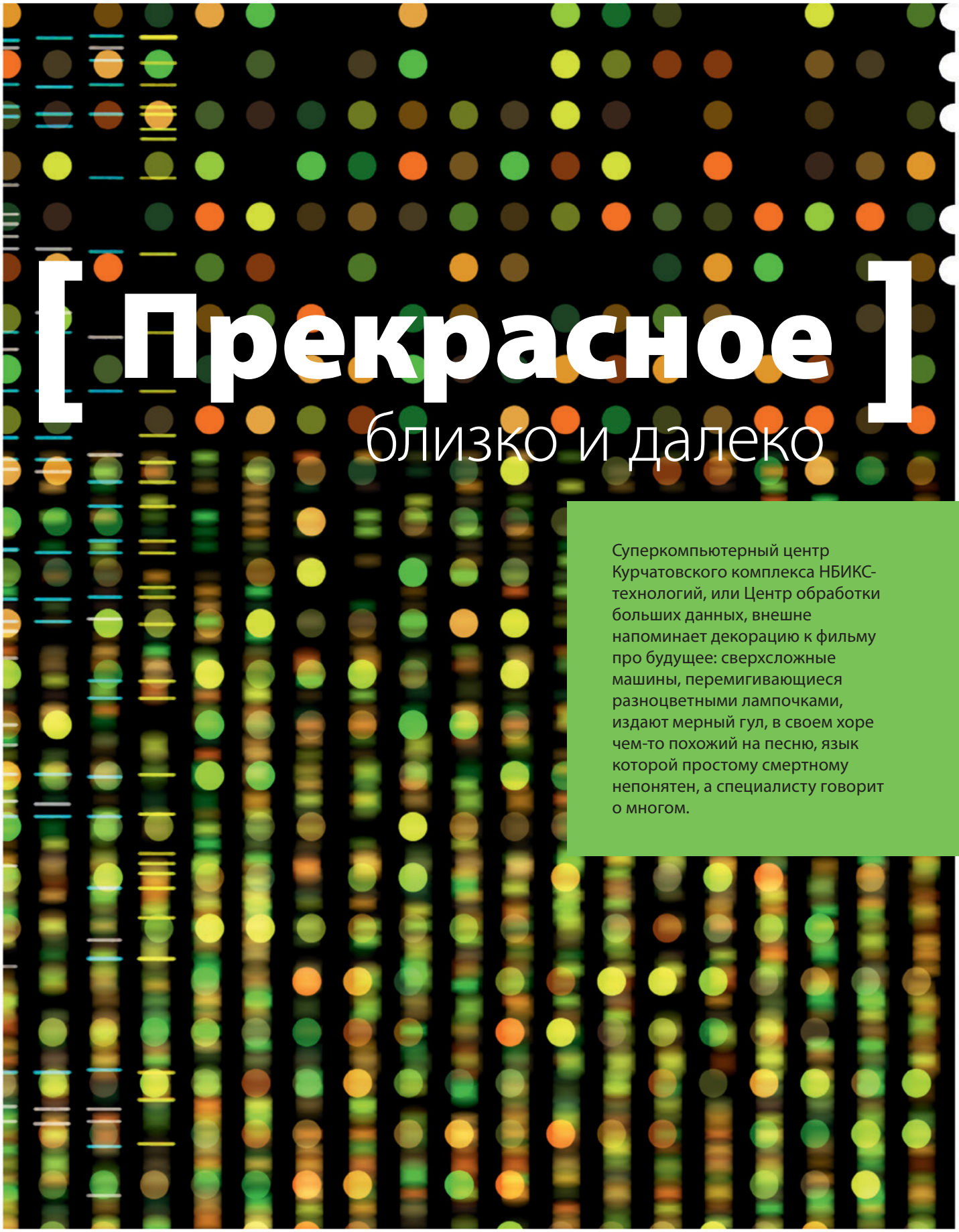

[Прекрасное]

близко и далеко

Суперкомпьютерный центр Курчатовского комплекса НБИКС-технологий, или Центр обработки больших данных, внешне напоминает декорацию к фильму про будущее: сверхсложные машины, перемигивающиеся разноцветными лампочками, издают мерный гул, в своем хоре чем-то похожий на песню, язык которой простому смертному непонятен, а специалисту говорит о многом.

Отдел перспективных компьютерных технологий и систем занимается целым рядом важнейших задач — от расшифровки молекул и вирусов до исследования дальних галактик. Именно об этом мы беседуем с молодым ученым, руководителем группы вычислительных технологий **Антоном Борисовичем Теслюком**.

— Антон, какова ваша основная исследовательская задача?

— Наша задача связана с анализом больших данных и применением суперкомпьютерных технологий для прикладных исследований, в которых вся мощь этих технологий может проявиться в полную силу.

Одно из главных направлений в нашей работе — поддержка больших экспериментов, того, что называется установками мегакласса: это масштабный международный проект, в котором мы участвуем, когда десятки стран совместно строят крупную экспериментальную установку. Типичный пример — Большой адронный коллайдер. Характерная особенность этого проекта — то, что во время его работы генерируются гигантские объемы данных, петабайты. Работа с этими данными очень сложна с точки зрения как информационных технологий, так и алгоритмов для анализа данных. Для поддержки такого проекта были созданы ресурсные центры более чем в ста лабораториях и исследовательских центрах. Курчатовский институт начал участвовать в развитии грид-центра в рамках *CERN* более десяти лет назад, а последние два года Курчатовский центр — это центр первого уровня (*Tier 1*) для экспериментов *ATLAS* и *CMS*. Мы поддерживаем для *CERN* необходимые сервисы и программное обеспечение, развиваем нашу вычислительную инфраструктуру и осуществляем передачу, хранение и распределение гигантских массивов информации, которые генерируются в *CERN*.

Видеостена для визуализации результатов расчетов

Еще один международный проект, относящийся к установкам мегакласса, — лазер на свободных электронах *XFEL*, строящийся сейчас в Гамбурге. Россия стала крупнейшим после Германии партнером этого проекта по своему финансовому, технологическому, интеллектуальному вкладу. Мы в Курчатовском институте разрабатываем новые алгоритмы для анализа данных, которые будут генерироваться на лазере свободных электронов. Эта установка уникальна, но обработка информации, повторюсь, — большая алгоритмическая и математическая проблема. По сей день стоит ряд вопросов. Как извлекать смысл, как восстанавливать структуру молекул или вирусов, которые будут исследоваться на этом лазере? Важно создать новые алгоритмы, позволяющие получить те знания, ради которых эта установка строится.

— Для какой цели она строится?

— Лазеры на свободных электронах известны уже более полувека. Но лишь совсем недавно люди научились делать их настолько мощными, яркими и излучающими такой короткий импульс (порядка нескольких фемтосекунд), что можно наблюдать на детекторе рассеяние фотонов на отдельных молекулах. За столь короткое время молекула не успевает деформироваться, и ее можно считать неподвижной. Благодаря такому эффекту можно, анализируя большое количество дифракционных изображений, получить структуру отдельной молекулы или даже атома. Но это пока дело будущего. То, о чем можно говорить сейчас, — возможность исследования крупных макромолекулярных комплексов или отдельных вирусов, которые невозможно изучать традиционными методами типа рентгеноструктурного анализа. Они просто не замораживаются в кристаллы. Когда эта установка заработает, появятся новые инструменты, при помощи которых можно заглянуть внутрь удивительного наномира.

— Как можно туда заглянуть? Как в микроскоп?

— Нет, совсем иначе. В микроскоп можно посмотреть и увидеть объект, который мы хотим изучить. Здесь же мы видим косвенную информацию о том, как отдельные фотоны дифрагировали,

В Курчатовском суперкомпьютерном центре

рассеивались на объекте. Вычислительная сложность состоит в том, что необходимо из косвенной информации собрать точную информацию о том, как эта молекула устроена. Сложность в самом построении эксперимента. После каждого импульса молекула уничтожается, и каждый раз мы видим новую молекулу, которая, как мы считаем, должна быть такой же, как предыдущая, но на самом деле это может быть не так.

Мы занимаемся тем, что разрабатываем новые алгоритмы, позволяющие классифицировать изображения, которые мы получаем. Здесь мы используем методы машинного обучения — кластеризацию, регрессию, классификацию, метод многомерных сложных данных. Кроме того, мы занимаемся методами восстановления структуры, исходного объекта по тем изображениям, которые прошли предобработку нашими системами фильтрации и сортировки. В результате мы учимся восстанавливать изображение и структуру той молекулы, которую исследуем.

— Какие еще задачи вы решаете?

— Другая область знаний, где наши технологии могут эффективно применяться, — это наука биоинформатика, работа с генетической информацией. Применение наших суперкомпьютеров эффективно тогда, когда мы имеем дело с большими данными, которые нельзя обработать на каком-нибудь отдельном персональном компьютере или отдельном сервере. Биоинформатика — пример такой области знания, где также востребована мощь наших технологий. Эксперименты по секвенированию генома производят гигантское количество данных. Человеческий геном, например, — это примерно 3 млрд нуклеотидов, которые при секвенировании прочитываются десятки сотни раз. То есть в работе с одним геномом мы имеем дело с сотнями гигабайт информации, и поиск новых смыслов из большого массива данных, сопоставление множества геномов, — область, куда мы можем привнести что-то новое.

В Курчатовском комплексе НБИКС-технологий уже несколько лет работает отдел, который занимается геномным секвенированием. Одна из наших нынешних задач — организовать информационно-технологическую поддержку таких экспериментов, используя наши наработки, уже полученные для работы с *CERN*. Для поддержки работы Большого адронного коллайдера было разработано программное обеспечение, которое умеет распределять задачи по сотням центров. Это такая большая грид-система. Сейчас актуальная задача состоит в том, чтобы научить этот грид применяться не только для задач БАК и задач физики, но и для других областей знаний.

— Для каких?

— В частности, мы сейчас делаем систему сервисов, которые могут решать биоинформатические задачи. У нас в Курчатовском институте есть лаборатория, которая занимается большими данными, и совместно с этой лабораторией мы привносим биоинформатические алгоритмы в систему, которая называется «Панда». Это новое поколение программного обеспечения, которое может масштабироваться на очень большое количество узлов и потоков данных.

— Есть ли что-то принципиально новое в расшифровке генома человека?

— Есть, и эта работа еще не опубликована. В Курчатовском институте был проект, когда мы генотипировали представителей более 40 этносов и исследовали их генотип, пытаясь понять, чем они отличаются друг от друга. Интересная задача — попробовать найти по генетическим данным возможные исторические взаимосвязи между различными этносами, понять, насколько изменился за века генотип, например, русского человека и, в частности, есть ли в нем заметные следы монголо-татарского ига. Мы могли сравнивать их, например, с современными бурятами и монголами, а в качестве усредненных русских, живущих в центральных регионах, взяли популяцию староверов из Новосибирской области, которые, как известно, не практикуют браки с представителями других этносов и национальностей и живут замкнуто. Иначе говоря, староверы должны быть чисто русскими.

— И что же выяснилось?

— Оказалось, у староверов азиатские следы видны чуть лучше, чем у современных русских. Поэтому можно предположить, что если такой след и был, то сейчас в популяции он уже практически растворился, а вот, например, у староверов его

еще немного видно. Любопытный результат: мы-то предполагали, что будет как раз наоборот.

— Какие прикладные возможности дают эти методы?

— Анализ геномов широко применяется в медицине, в прогнозировании риска всевозможных заболеваний начиная от разных форм рака и заканчивая шизофренией. Считается, что склонность к шизофрении опять же заложена генетически.

Коллеги из нашего геномного отдела сейчас занимаются также исследованием генетического материала мамонта, который был найден в очень хорошей сохранности на территории Якутии.

— У нас скоро появится живой мамонт?

— Нет, до живого мамонта пока далеко. Суть технологии тут в том, чтобы прочитать его геном с большой степенью покрытия. Это сложно, потому что ДНК деформируется со временем и найти сохранившийся образец не так-то просто. Восстановить мамонта из этого образца — технология послезавтрашнего дня.

— Расскажите о вашей вычислительной инфраструктуре. Ведь это огромное и очень беспокойное хозяйство?

— Да, это серверы и несколько больших вычислительных комплексов, которые сами по себе производят гигантское количество данных. Анализ этих данных — отдельная математическая и алгоритмическая проблема. Важно понять, как оптимизировать работу этих структур и предсказать намечающиеся проблемы.

— Каким образом можно предвидеть будущее, спрогнозировать его?

— Например, у нас работает вычислительный комплекс, решает разные задачи. И вот к нему выстраивается электронная очередь, чтобы получить доступ к ресурсам, которыми этот комплекс располагает. Как оптимизировать этот процесс? Ведь срочность доступа к этим ресурсам у всех разная. Мы должны знать всю историю работы и историю задач тех, кому это требуется, ведь в противном случае мы кого-то можем обслужить быстрее, а другие от этого пострадают. Значит, надо выяснить, какого пользователя мы должны поставить вне очереди, а какой может подождать, и ничего страшного из-за этого не произойдет. Это и есть система управления ресурсами и предсказание будущей нагрузки, развитием которой мы тоже занимаемся.

— Правильно ли я понимаю, что такие математические прогнозы можно делать в отношении больших человеческих общностей, а также конкретного человеческого здоровья? Скажем, прогноз рисков возникновения тех или иных заболеваний, которых пока нет.

— Совершенно верно. Одно из основных практических приложений для задач, связанных с прочтением и анализом ДНК, — предсказание тех рисков, которые несет человек, прогноз возможного заболевания, которое можно предотвратить, если заранее о нем знать. Правда, тут есть свои сложности. Часто с точки зрения биологии не всегда понятна связь между особенностями ДНК и будущей болезнью. Иногда эта связь может быть косвенной. Однако ясно, что в этом направлении необходимо работать.

— Можно ли по геному определить склонность к каким-либо патологиям или особенностям? Например, сказать, что представители той или иной этнической группы в чем-то больше рискуют, чем другие?

— Да. Мы, например, исследовали, какими мутациями один этнос отличается от другого, и обнаружили, что африканские пигмеи, для которых характерен очень маленький рост, имеют частую характерную мутацию *rs484959*, связанную с дефектом костей. Эту особенность они передают генетически. Другой хорошо известный пример — способность у разных этносов вырабатывать фермент лактазу, который отвечает за расщепление молока. Известно, что молоко лучше усваивается жителями Европы, чем жителями Африки, Азии или американскими индейцами, которые исторически не усваивают молоко. Считается, что способность вырабатывать лактазу во взрослом возрасте и перерабатывать молоко появилась у северных людей — датчан, шведов, финнов. Но она часто напрочь отсутствует как у южных этносов, так и у восточных. Например, китайцы практически не усваивают молоко. Кстати, похожим образом себя ведут и коренные северные народы, например эскимосы.

— А алкоголь?

— С алкоголем наоборот. У юго-восточных народов, например у тех же китайцев, есть характерная генетическая особенность, благодаря которой они расщепляют алкоголь в токсичный ацетальдегид гораздо быстрее европейцев, но при этом выводят его из организма медленнее. Ввиду этого употребление алкоголя дает им больше неприятных ощущений и проходит тяжелей, чем у большинства европейцев.

— У вас на стене — необыкновенной красоты космические картинки. Имеют ли они отношение к вашей работе?

— Да, это изображения *CERN*, карта ресурсных центров, которые обслуживают этот большой эксперимент, и мы — один из них. Например, картинка из области астрофизики показывает зарождение двойной звезды. Так что мы участвуем в решении самых разных задач — как удаленных на сверхдальние космические расстояния, так и находящихся вплотную к нам, таких как исследование микробов или вирусов. Думаю, для человечества одинаково важно и то и другое. ■

Беседовала Наталия Лескова