Свободный дух БИОИНФОРМАТИКИ

Ю.Ю. ВАСЬКИН, Ю.Э. ДАНИЛОВА

Темп развития, набранный молекулярной биологией, не может не поражать. Рутинным делом стала расшифровка ДНК: современное оборудование для высокопроизводительного секвенирования позволяет полностью определить нуклеотидные последовательности генома человека всего за один день и 1 тыс. долларов! И сейчас внимание ученых сконцентрировано на исследовании функциональной роли тех или иных участков генома и последствий, к которым приводят изменения в них. В распоряжении исследователей уже находятся большие базы данных, и каждый день генерируются все новые и новые гигабайты биологической информации. Адекватно оперировать столь гигантскими объемами данных невозможно без автоматизированных средств анализа. Именно биоинформатика становится в наши дни связующим звеном между компьютерными и молекулярно-генетическими технологиями, обеспечивая прогресс в новых динамично развивающихся фундаментальных и прикладных областях биологии

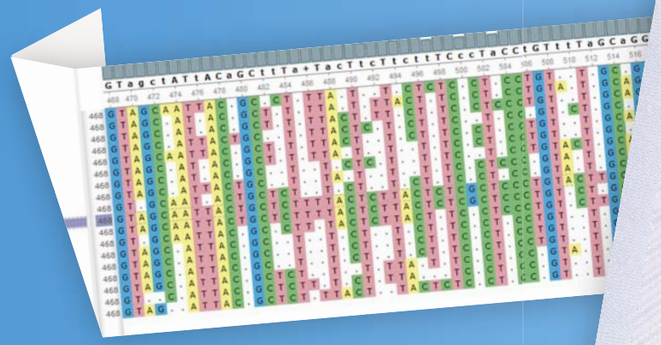
В биоинформатике можно выделить три основных направления: анализ и статистическая обработка биологических данных, разработка алгоритмов анализа и создание специализированных программных средств. Первые два направления по большей части относятся к исследовательской деятельности и получению значимых научных результатов. Последнее же очень близко по духу к традиционной разработке программного обеспечения, с которой имеют дело IT-компании.

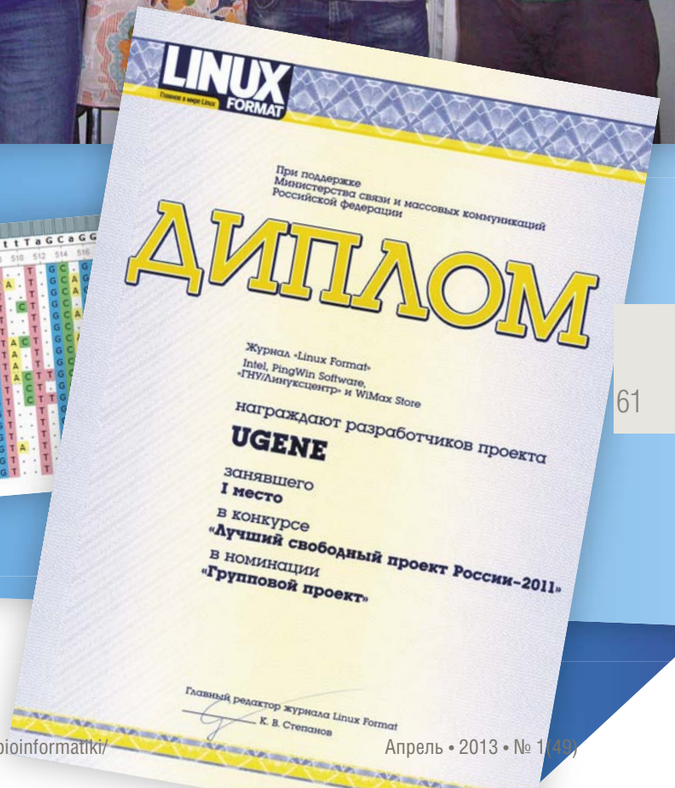
В зарубежной практике принято, что научные лаборатории нанимают специалистов или даже содержат целые отделы по обработке биологических данных и созданию программно-аппаратного обеспечения,

Ключевые слова: биоинформатика, молекулярно-генетическая информация, высокопроизводительное секвенирование, вычислительные конвейеры, свободное ПО
Key words: bioinformatics, molecular-genetic information,high throughput sequencing, computational chains, open source

ВАСЬКИН Юрий Юрьевич – инженер-программист НЦИТ УНИПРО (Новосибирск), магистрант факультета Новосибирского национального исследовательского государственного университета. Победитель конкурса УМНИК Фонда содействия развитию малых форм предприятий в научно-технической сфере (2012 г.).
Автор и соавтор 5 научных работ

ДАНИЛОВА Юлия Эдуардовна – кандидат физико-математических наук, заместитель директора НЦИТ УНИПРО (Новосибирск).
Автор и соавтор 10 научных работ





Молодежная биоинформационная команда, работающая над проектом UGENE

однако в нашей стране ученые часто самостоятельно осваивают или даже разрабатывают нужные им программные средства. И такой переход от традиционных «пробирок» к высокопроизводительным вычислениям не дается легко. У исследователей могут возникать сложности при использовании разнородных комплектов программного обеспечения, которые могут работать с несовместимыми форматами данных и только на непривычных операционных системах. Возникают проблемы и с редактированием огромных файлов, с поиском и установкой программ, часто представляющих собой файлы с исходным кодом, который нужно сначала скомпилировать и настроить, а также с решением других компьютерных задач, которыми не должен заниматься биолог.

В этой связи в 2003 г. Новосибирский центр информационных технологий «Унипро» решил приложить свой опыт программных разработок в бурно развивающейся области оцифрованной биологии. Тогда же были налажены первые научные контакты, результатом которых стало первое биоинформационное приложение, разработанное новичком компании, выпускником физфака НГУ М. Фурсовым.

В течение последующего десятилетия компания разрабатывала специализированные биоинформационные приложения и расширяла круг потребителей и разработчиков своих программных продуктов.

Основным наукоемким проектом «Унипро» стал UGENE, появившийся в июне 2008 г. И хотя на старте этого проекта участвовало всего три инженера-программиста, задачу сразу поставили амбициозную – перегнать коммерческие аналоги.

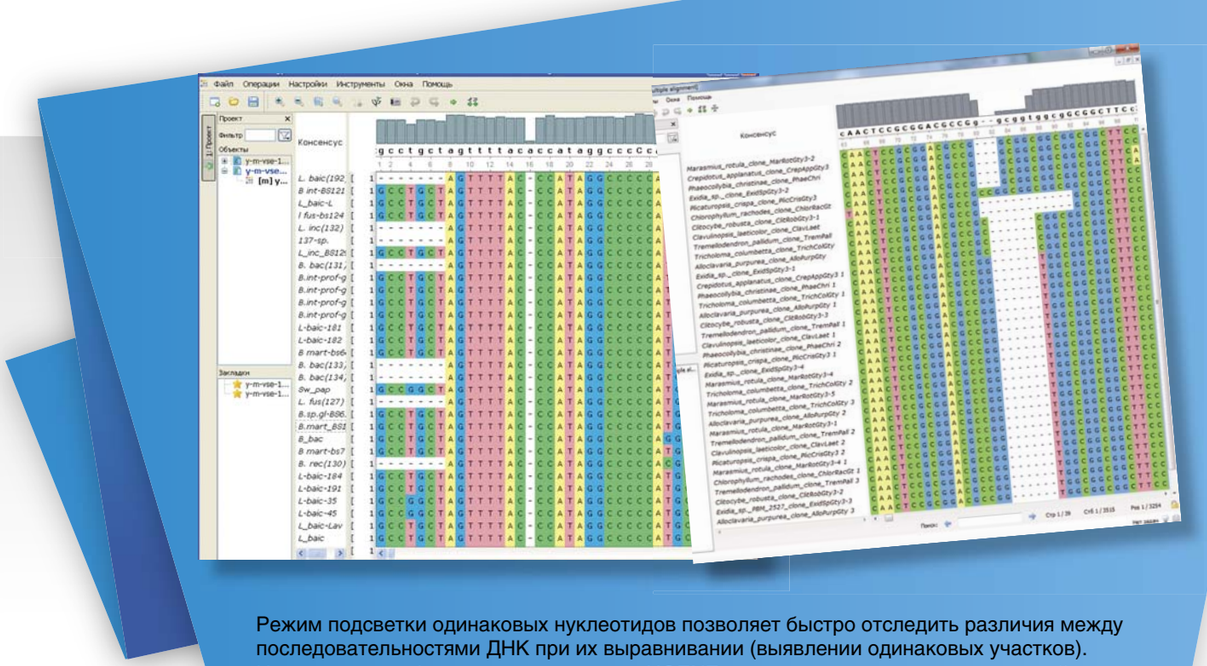
Что умеет UGENE

UGENE объединяет «под одной крышей» множество популярных инструментов для работы с молекулярно-генетической информацией. В то же время изрядную долю пакета составляют результаты реализации оригинальных идей самих разработчиков, а также предложений, которые постоянно вносят пользователи. Немаловажным фактором привлекательности UGENE является и удобный графический интерфейс: пользователь может запускать интересующие его алгоритмы и сразу отслеживать результат в одном из окон визуализации.

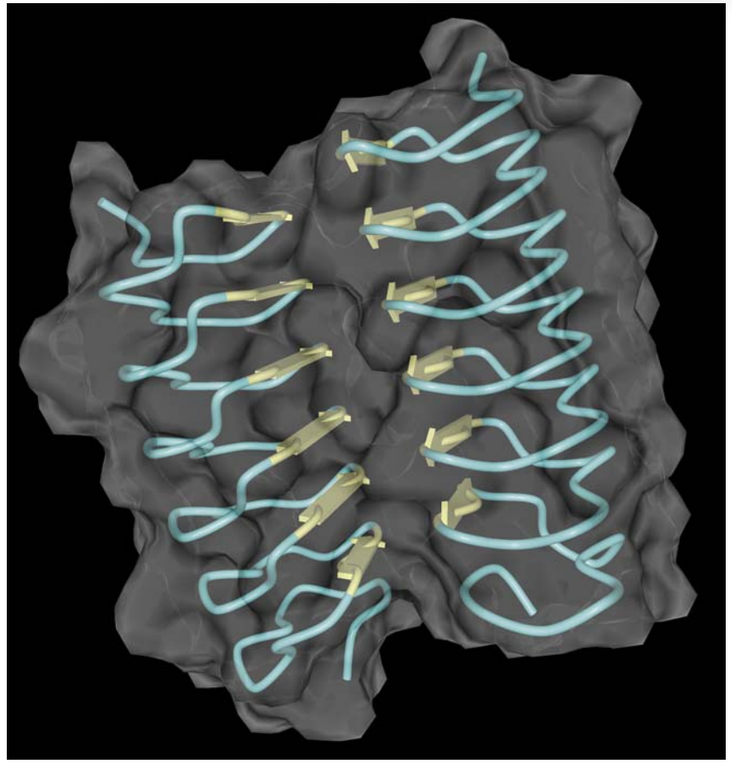
Программные инструменты, с помощью которых реализован UGENE, позволяют использовать его на большинстве операционных систем. К тому же пакет распространяется бесплатно и с открытым исходным кодом, т. е. любой пользователь может видеть, как устроена программа изнутри, и при желании адаптировать ее под собственные нужды – лицензионное соглашение GPL позволяет встраивать в пакет другие программные разработки с совместимыми лицензиями.

Как известно, высокое сходство макромолекул обычно указывает на их значительное структурное, функциональное и эволюционное родство. Поэтому одной из главных задач биоинформатики является *выравнивание* – поиск сходных участков молекулярных последовательностей.

Сегодня имеются сотни различных видов и инструментов выравнивания. В зависимости от задачи ученый может работать с одной или с миллионами последовательностей, искать похожие участки в удаленной базе данных или проводить выравнивание нескольких десятков



Режим подсветки одинаковых нуклеотидов позволяет быстро отследить различия между последовательностями ДНК при их выравнивании (выявлении одинаковых участков). Использование функции выравнивания UGENE: *слева* – при поиске некодирующей митохондриальной ДНК, пригодной для видовой идентификации восьми видов байкальских губок, *справа* – для сравнения последовательностей одного из ферментов грибов, обратной транскриптазы Tcn1. *По данным ЛИН СО РАН (Иркутск) и ИЦиГ СО РАН (Новосибирск)*



Так выглядит построенная в программе UGENE 3D-структура антифризного белка, который, встраиваясь в кристаллическую решетку льда, препятствует росту ледяных кристаллов в клетках арктических животных. *Реконструкция по данным: (Howard et al., 2011)*

«УНИПРО» – 20 ЛЕТ!

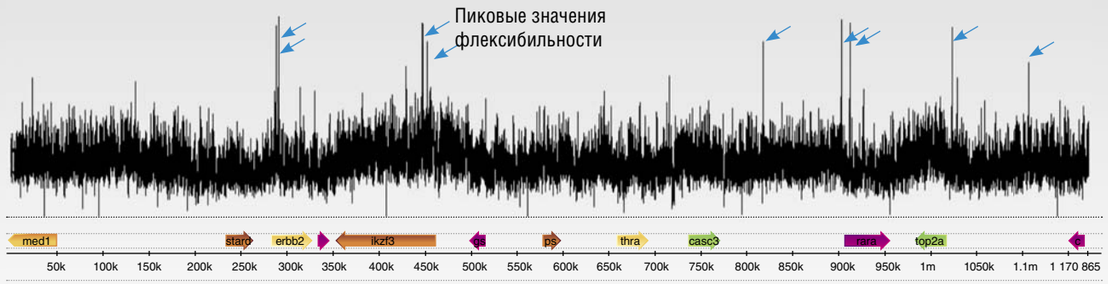
НЦИТ «Унипро» был создан в 1993 г. в новосибирском Академгородке командой профессионалов системного программирования, работавших в российском суперкомпьютерном проекте «Эльбрус». В течение 12 лет центр в сотрудничестве с американской компанией Sun Microsystems, известным производителем программного и аппаратного обеспечения, тестировал и улучшал язык Java – ныне самый востребованный язык программирования. Сейчас в компании работают свыше 70 человек, в большинстве своем выпускники НГУ и НГТУ.

Программисты Унипро много лет успешно разрабатывают сложные тестовые и моделирующие системы, встроенное программное обеспечение для электроэнергетики, оптимизирующие компиляторы, корпоративные системы управления. Большинство проектов компании – долгосрочные, а среди заказчиков есть такие громкие имена, как Google.

Биоинформатикой в Унипро стали заниматься с начала 2000-х гг. Почему же компания, столь успешно работающая в традиционных компьютерных областях, вступила в такую сложную, рискованную и «чуждую» научную область? Наверное, не последнюю роль сыграло обыкновенное человеческое любопытство, ведь речь идет о близкой всем науке о жизни, только рассмотренной под другим углом.

Кроме того, находясь в научном центре, трудно удержаться от сотрудничества с учеными: у компании к этому времени уже имелся успешный опыт сотрудничества с геофизиками и математиками.

Биоинформатику в Унипро сегодня можно назвать научной отдушиной среди заказных коммерческих проектов, которая предполагает полную свободу творчества. К тому же она служит и хорошим полигоном для тренировки начинающих программистов: большинство биоинформационной команды составляют студенты, магистранты и вчерашние выпускники университетов, работающие под руководством опытных инженеров-программистов



Программа UGENE была использована для оценки степени флексибильности ДНК и обнаружения в ней потенциальных мест цепочечных разрывов ДНК, которые связаны с образованием рака молочной железы.
Анализ выполнен на участке нуклеотидной последовательности 18-й хромосомы человека , включающем семь генов (база данных GenBank).
Данные Н.Ю. Маценко (НИИМББ СО РАМН, Новосибирск)

последовательностей, полученных в эксперименте. Соответственно и подходы к обработке и визуализации данных должны быть совершенно разными. К тому же нужно помнить об удобстве пользователя и предоставить ему возможность выполнять рутинные операции самым простым способом.

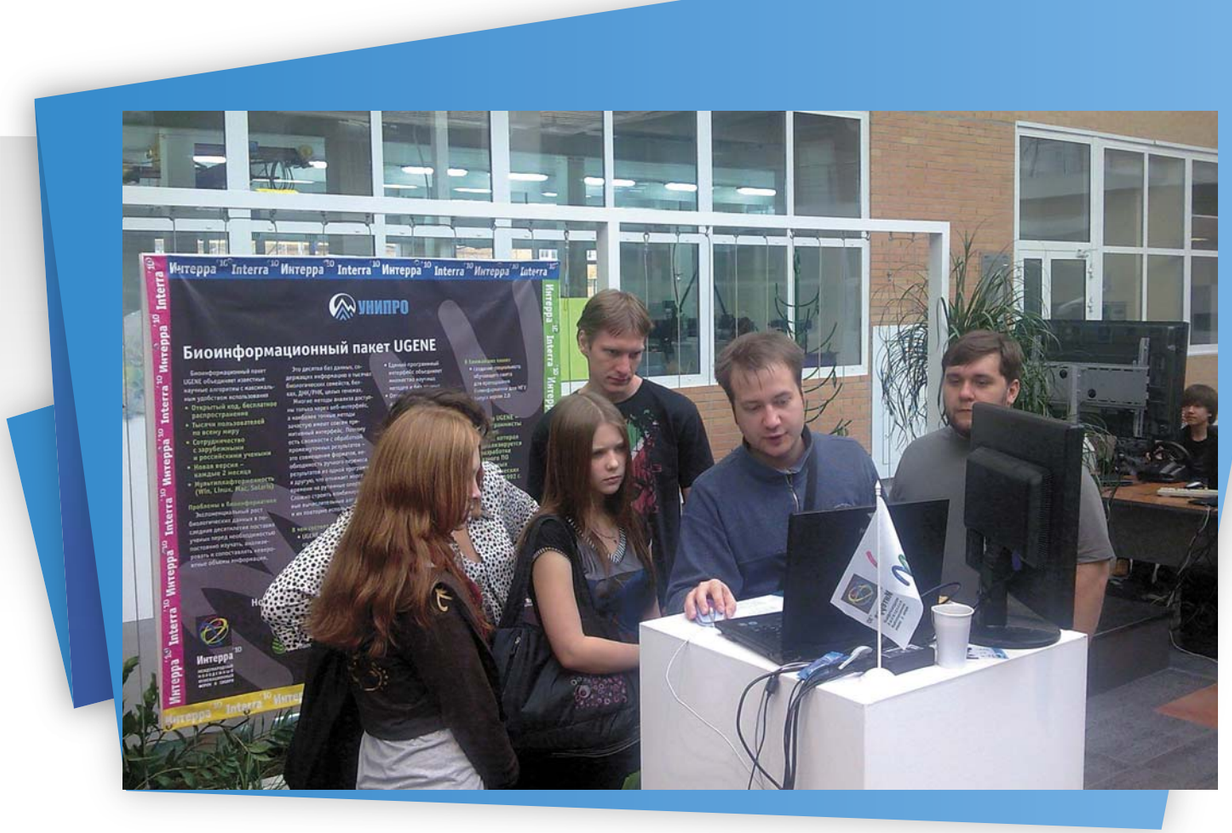
Поэтому в рамках UGENE предусмотрены различные специализированные рабочие среды. Например, если требуется анализировать только одну нуклеотидную последовательность, то открывается соответствующее окно с ее изображением, где предлагаются инструменты поиска схожих участков, конструирования молекулярных векторов и т.д. Для белковых аминокислотных последовательностей доступны визуальные трехмерные модели или алгоритмы предсказания вторичных структур.

Для манипулирования множественными последовательностями выделено отдельное рабочее пространство, где под рукой у биолога имеется свыше десятка известных алгоритмов выравнивания и сравнения последовательностей, а также построения филогенетических деревьев.

Довольно часто ученым приходится выполнять многостадийную обработку большого количества данных. Обычно в таких вычислительных «конвейерах» одни алгоритмы или программы используют результаты предварительной работы других, поэтому ученым часто приходится вручную осуществлять взаимодействие между ними. Встроенный в UGENE оригинальный *дизайнер вычислительных схем* позволяет быстро собирать конвейерные схемы обработки из вычислительных элементов, в которых данные передаются автоматически

Пример работы встроенного в UGENE дизайнера вычислительных схем: схема, позволяющая фильтровать любое количество нуклеотидных или аминокислотных последовательностей по заданной длине (отобранные последовательности записываются в разные файлы). Схему можно использовать многократно

Заместитель директора Института цитологии и генетики СО РАН (Новосибирск) С. В. Лаврюшев:
«Команду разработчиков Унипро знаем давно, регулярно и конструктивно общаемся. Мы видим их неуклонный прогресс в освоении новых технологий и областей биоинформатики. Наши ученые уверенно пользуются их софтом наряду с разработками Института и зарубежными пакетами. Надеемся, что наше сотрудничество в будущем будет только укрепляться»



по соединениям, заданным пользователем. Такие схемы хороши тем, что в дальнейшем их можно использовать многократно для обработки различных наборов входных данных.

Под флагом UGENE

В Унипро не числится ни одного биолога, но поддерживать высокий научный уровень проекта позволяют постоянные контакты с учеными, среди которых не только пользователи UGENE, но и активные участники разработки алгоритмов.

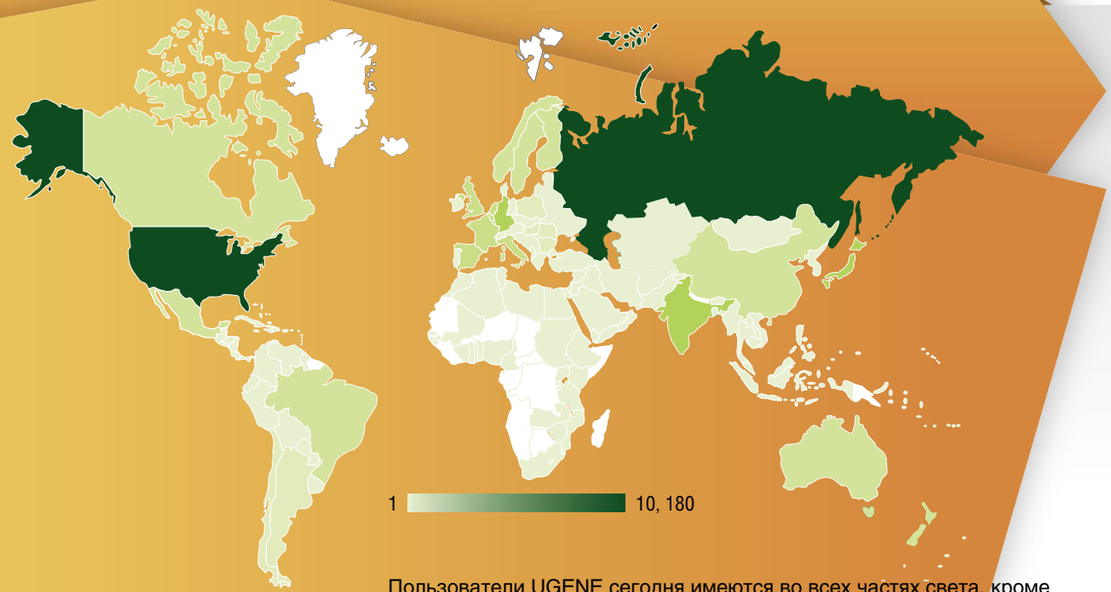
Например, в UGENE был встроен алгоритм поиска сайтов связывания транскрипционных факторов SITECON, разработанный к.б.н. Д. Ощепковым и его коллегами из Института цитологии и генетики СО РАН (Новосибирск). А по мнению д.ф.-м. н. Е.Е.Витяева из Института математики СО РАН, от интеграции с UGENE, несомненно, выиграла разработанная им система ExpertDiscovery для поиска комплексных сигналов в регуляторных районах генов. Благодаря такому «слиянию» этой программой могут пользоваться исследователи из разных стран.

Многие алгоритмы аннотирования и анализа геномов разрабатываются непосредственно под конкретные задачи биологов. Так, с 2004 г. Унипро сотрудничает с лабораторией молекулярно-генетических систем ИЦиГ СО РАН, где занимаются поиском мобильных элементов в геномах различных грибов. Алгоритмы поиска, разработанные в Унипро, позволили ускорить эти вычисления в десятки и сотни раз.

В 2011 г. Унипро совместно с НГУ и Intel-lab проводили школу-семинар «Вычислительные задачи молекулярной биологии и платформа UGENE», в ходе которого молодым ученым был задан вопрос: каким образом UGENE

Презентация биоинформационных разработок Унипро на выставке «Живая инновация».
Новосибирск, Технопарк, 2010

С 2010 г. для ученых-биологов время от времени проводятся ознакомительные и обучающие семинары по работе с пакетом UGENE. И хотя аудитория этих семинаров в сотни и даже тысячи раз меньше той, которую можно собрать в результате интернет-маркетинга, такая форма продвижения программного продукта достаточно эффективна. Ведь если хотя бы один человек из научного подразделения после посещения семинара начнет активно и успешно пользоваться программой, это неизбежно привлечет внимание и многих его коллег. Такой пиар очень значим для команды разработчиков, ведь именно «сарафанное радио» в научной среде является самым быстрым и надежным рекламным инструментом



Пользователи UGENE сегодня имеются во всех частях света, кроме Антарктиды. Его разработчики регулярно получают благодарные отклики : «Very good software for using… Many thanks!». Некоторые иностранные пользователи инициируют выпуск локализованных версий на своем родном языке: уже появились версии на чешском и китайском языках, а скоро, вероятно, к ним добавится испанская.
На карте – география пользователей UGENE по данным Google Analytics. Согласно этой информации, наибольшее число пользователей к настоящему времени приходится на Россию, страны СНГ и США

может быть полезен для Вашей научной работы? Среди ответов упоминались моделирование экспериментов по клонированию генов, построение филогенетических деревьев и другие задачи, уже успешно решающиеся в среде UGENE. Однако под специфические запросы некоторых участников имеющийся функционал пришлось дорабатывать. Например, по просьбе к.б.н. Н.Ю. Маценко из Института молекулярной биологии и биофизики СО РАМН (Новосибирск) был разработан специальный модуль по поиску в ДНК сайтов флексибильности, т.е. участков нуклеотидной цепочки с критическими параметрами гибкости и ломкости.

Активные пользователи UGENE есть не только в Сибири, но и в европейской части России. Среди них – специалисты из санкт-петербургского Всероссийского научно-исследовательского института сельскохозяйственной микробиологии РАСХН, занимающиеся анализом почвенных микробиомов*, и московского Центрального НИИ эпидемиологии, интересующиеся обработкой результатов высокопроизводительного секвенирования ДНК микроорганизмов. А недавно сотрудники Унипро стали желанными гостями на Всероссийском семинаре по современным методам индикации возбудителей инфекционных болезней для сотрудников институтов системы Роспотребнадзора, где провели практическое обучение за компьютерами. Это свидетельствует о том, что сегодня UGENE используется уже не только в фундаментальной науке, но и в прикладных исследованиях, направленных на улучшение здоровья людей и окружающей среды.

В 2012 г. был заключен контракт с американским Национальным институтом аллергических и инфекционных болезней (NIAID), согласно которому на базе дизайнера вычислительных схем UGENE предстоит построить особую инфраструктуру по анализу данных высокопроизводительного секвенирования.

Дело в том, что обычно такие вычисления производят на мощных компьютерных кластерах, однако в некоторых случаях (например, когда пользователь находится в экспедиции в удаленных районах) возникает необходимость использовать только локальный компьютер. Все результаты этой работы также должны войти в открытую бесплатную версию UGENE, т.е. станут доступны для всего мирового сообщества.

* Подробнее на с. 68–75

Обучение молодых ученых на школе-семинаре в лаборатории НГУ-Intel, *Новосибирск, май 2011 г.*

Биоинформатика сразу стала в Унипро самым молодежным направлением. Костяк фирмы в последние годы остается стабильным, однако приток начинающих программистов постоянен. Поэтому над проектом UGENE успели потрудиться не менее двадцати человек – студенты, магистранты и недавние выпускники вузов, для которых проект стал настоящей профессиональной школой. Открытый код проекта, с одной стороны, предъявляет к его участникам повышенные требования, но с другой – служит им отличной рекомендацией.

Сейчас биоинформационный отдел компании, насчитывающий 13 человек, превратился в настоящую научно-исследовательскую лабораторию. Здесь ведется постоянная работа по исследованию и разработке новых вычислительных алгоритмов, оптимизации уже существующих методов, использованию новых программных технологий. Сюда ежегодно на преддипломную практику приходят студенты НГУ и НГТУ, постепенно становясь профессиональными программистами, а сотрудники регулярно выступают с докладами на конференциях и семинарах.

Два года назад UGENE был удостоен звания «Лучший свободный проект России – 2011», а биоинформационная команда Унипро стала единственным российским участником международного конкурса «Illumina IDEA Challenge 2011», объявленного мировым производителем секвенаторов. В 2012 г. разработчики UGENE опубликовали результаты своей работы в «Bioinformatics» – одном из самых уважаемых международных журналов в этой области, что упрочило позиции пакета в научной среде.

За девять лет работы в области биоинформатики в компании Унипро только укрепились в мысли, что хорошие программисты нужны биологам, и что биоинформатика – нетривиальный и интересный путь, ведущий к пониманию окружающего нас сложного органического мира.

Заместитель директора НЦИТ «Унипро» М.Ю. Фурсов: «Нам часто задают вопрос: открытый и такой серьезный проект – это здорово, но за чей счет вы его делаете, неужели он окупается?
Особенность нашего проекта в том, что он не начался с нуля, а родился и развивался внутри стабильной компании, имеющей долгосрочные проекты в других технологических областях программирования.
Именно поэтому мы пошли на риск, сделав UGENE бесплатным, и долгое время развивали его без какой-либо финансовой отдачи. Лишь в 2011 г., через три года после выпуска первой версии, стали поступать первые заказы на биоинформационные разработки. Это дает основания надеяться, что уже в ближайшем будущем проект станет окупаемым»

Литература

Lesk A. Introduction to Bioinformatics. Third Edition. 2008.

Vaskin Y.Y., Khomicheva I.V., Ignatyeva E.V. Vityaev E.E. Expert Discovery and UGENE integrated system for intelligent analysis of regulatory regions of genes // Silico Biol. 2011–2012;11(3–4):97–108. doi: 10.3233/ISB-2012-0448. PMID: 22935964.

Okonechnikov K., Golosova O., Fursov M. Unipro UGENE: a unified bioinformatics toolkit // Bioinformatics. 2012; doi:10.1093/bioinformatics/bts091.

Авторы выражают благодарность команде UGENE за помощь в подготовке статьи