Стриж И. Г.
ЗНАЧЕНИЕ ОНТОЛОГИЙ ДЛЯ СОВРЕМЕННЫХ БИОМЕДИЦИНСКИХ ИССЛЕДОВАНИЙ И УЧЕБНОГО ПРОЦЕССА
Адрес статьи: www.gramota.net/materials/1/2008/11/45.html

Источник
Альманах современной науки и образования
Тамбов: Грамота, 2008. № 11 (18). С. 116-119. ISSN 1993-5552.
Адрес журнала: www.gramota.net/editions/1.html
Содержание данного номера журнала: www.gramota.net/materials/1/2008/11/

тат выполнения которых позволил бы студентам: оценивать эффективность своей работы; максимально реализовать свои особенности и способности; пережить «ситуацию успеха» в совместной деятельности и т.п.

Исходя из выше сказанного, методическая подготовка студентов педвуза должна опираться на следующие положения:

– методика обучения предмету - самостоятельная научная область с собственным предметам, методами исследования, понятийным аппаратом и концепциями;

– изложение материала ведется в контексте системного анализа и деятельностного подхода;

– разработка методических концепций осуществляется с учетом новых образовательных идей;

– изложение материала ведется с учетом результатов научных исследований по методике обучения предмету;

– студент не столько усваивает готовые факты, сколько принимает участие в их формулировке и обосновании.

ЗНАЧЕНИЕ ОНТОЛОГИЙ ДЛЯ СОВРЕМЕННЫХ БИОМЕДИЦИНСКИХ ИССЛЕДОВАНИЙ И УЧЕБНОГО ПРОЦЕССА

Стриж И. Г.
Московский государственный университет им. М. В. Ломоносова

Введение

В последнее десятилетие, в силу стремительного развития высоко-технологичных аналитических методов, приведших буквально к революции в области молекулярной биологии, биотехнологии, генной инженерии и биоорганической химии, мы наблюдаем экспоненциальный рост объема информации во всех областях биологии и медицины. Это, с одной стороны, дает ученым уникальную возможность исследовать и изучать динамические свойства любой биологической системы более подробно и тщательно, но, с другой стороны, создает проблему валового потока информации. Эффективное использование накопленной и получаемой информации становится насущной проблемой современной науки. С целью решения этой проблемы возникла новая область биологии – биоинформатика, предметом которой является анализ экспериментальных данных молекулярной биологи (секвенированных последовательностей биополимеров, экспериментально определенных пространственных структур биологических макромолекул, данных об экспрессии генов и т.д.). Однако, несмотря на активную работу биоинформаторов, способных грамотно обрабатывать рутинно полученные высокоинформативные экспериментальные данные, ответы на многие фундаментальные вопросы биологии до сих пор не получены. Успех и прогресс современной биологии будет зависеть от способности ученых воссоединить отдельные элементы процесса и создать целостную картину биологического явления. Очевидно, что решение этой задачи практически невозможно без применения современных информационных методов. В настоящей статье обсуждаются перспективы использования онтологий для информационной поддержки современных биомедицинских исследований и учебного процесса.

Проблема эффективного использования биологической информации

Принципиальным отличием биологии и медицины XXI века является то, что практически любое современное исследование не только приводит к появлению многочисленных данных, но и требует привлечения новых технологий, способствующих решению проблемы хранения, обработки и интеграции получаемых данных, а также позволяющих исследователям вычленить необходимую им информацию. Помочь ученым призваны методы биоинформатики, а именно методы организации молекулярно-биологической информации, широко понимаемые компьютерные методы, методы вычислительной математики и статистики. Для решения поставленных задач могут быть привлечены также такие информационные технологии как, например, Интернет-технологии и Грид-технологии [Jutchkov et al. 2005: 37].

На сегодняшний день основным местом хранения полученной и накопленной информации является множество баз данных. К концу 2007 года насчитывалось более 1000 молекулярно-биологическиех баз данных, причем за последние годы создавалось более 100 новых баз в год [Galperin 2007]. Основной задачей баз данных является интеграция и организация различной информации о биологических объектах, например, генах и их продуктах, что должно позволить пользователям ориентироваться в этом огромном объеме данных. Базы данных, поддерживаемые различными институтами и компаниями, зачастую сильно отличаются не только по содержанию, но и по используемым форматам и способам представления данных, что существенно затрудняет работу с ними. Исследователь, заинтересованный в конкретной информации, сталкивается с проблемой ее поиска по множеству разрозненных баз. Этот процесс, сам по себе, очень трудоемок и занимает огромное количество времени, поскольку поиск нужной информации сводится не только к поиску нужной базы, содержащей интересующий вас объект, но и к необходимости разобраться в структуре и устройстве найденной базы, а также в построении запросов к ней. Кроме того, зачастую, исследователей интересует информация, которая может храниться разрозненно в различных базах данных. К примеру, необходимо выяснить какой ген или генный продукт ответственен за формирование и развитие определенной структуры или органа в конкретном организме. Прямой поиск подобной информации может не скоро привести к успеху. Нередко исследователю необходимо расширить запросы и вести поиск, к примеру, генных продуктов выполняющих сходную функцию в различных организмах, либо, наоборот, узнать какую функцию выполняет гомологичный ген или его продукт, но в другом организме. Однако базы данных часто со-

держат информацию только о генах или о генных продуктах и их функциях в конкретном отдельном организме. Это существенно затрудняет сопоставление данных полученных на разных объектах, а также анализ результатов эксперимента. Еще одной проблемой является то, что информация даже об одном объекте содержится в нескольких независимых базах. К примеру, существующая информация касающаяся *Arabidopsis thaliana*, уникального и удобного объекта биологии растений, заключена в десятках различных баз данных как, например, DDBJ, EMBL, GenBank, TAIR, MatDB, FLAGdb/FST, DatA, PPMdb и др. В результате работа с большими наборами данных, получаемыми как в результате собственной экспериментальной работы, так и уже с существующими и доступными через Интернет, является не простой задачей даже для опытных специалистов в области биоинформатики. Вместе с тем, работать с базами данных сегодня приходится ученым не зависимо от их специализации, причем зачастую это становится просто неотъемлемой частью современного исследования.

Рис. 1. *Сценарий работы современного биолога (с изменениями по [Strizh 2006: 199])*

Упрощенный сценарий работы современного биолога можно представить следующим образом (Рис. 1): (i) для решения поставленной задачи экспериментатор должен проанализировать большое количество информации из различных источников, в том числе и баз данных; (ii) результатом эксперимента с использованием современных высоко-технологичных методов является большой массив данных; (iii) анализ полученных данных требует обращения к разнородным базами данных, например с целью определения структуры и функции обнаруженных белков или экспрессирующихся генов; (iv) в случае успешно выполненной аналитической работы, результатом экспериментальной работы может явиться новая гипотеза или новый продукт [Strizh 2006: 199]. К сожалению, отсутствие отрегулированной связи между получением экспериментальных данных, их хранением и аналитической работой зачастую приводит к тому, что финальная и являющаяся наиболее значимой с точки зрения практической и интеллектуальной ценности часть работы остается не реализованной. Очевидно, что решение этой проблемы практически невозможно без привлечения соответствующего инструмента, способствующего эффективной навигации по базам данных, а также позволяющего извлекать и интегрировать необходимые исследователю данные. Следует отметить, что беспрецедентное влияние интернет-технологий на все сферы современной науки в последние несколько лет привело к переносу акцента с традиционных методов представления информации на представление информации в форме, удобной для ее использования в среде Интернет. Ярчайшим примером является база данных PubMed, содержащая аннотации и полнотекстовые электронные варианты статей, опубликованных в передовых научных изданиях. Следовательно, инструментарий для работы с информационными массивами должен удовлетворять требованиям той среды, в которой он будет работать, а именно среды Интернет. Как показывает практика, сближение с Интернет-технологиями практически любой научной сферы существенно ускоряет темпы ее развития, что подтверждается, в частности, вспышкой интереса к проблеме представления знаний [Кафтанников, Коровин 2003: 134]. Поскольку, знания являются неотъемлемой частью любого учебного процесса, возникает закономерный вопрос каким образом можно использовать результаты многочисленных исследований, которые «заключены» в базы данных не только в исследовательской работе, но и в обучении и подготовке современных биологов.

Онтологии – эффективный инструмент для работы с информацией

Одним из подходов, применяемых для интеграции данных и знаний в биологии и медицине, является использование онтологий. За последние десятилетия построение онтологий переросло из сугубо философской дисциплины в интенсивно развиваемые информационные технологии, которые уже получили признание в

том числе и в биологии [Strizh 2006: 199]. Сегодня онтологии являются эффективным средством навигации в огромных информационных массивах, таких как, например, геномные, траскриптомные и протеомные базы данных. Они также являются эффективным инструментом, который значительно облегчает получение, интеграцию и анализ разнородной биомедицинской информации [Жучков и др. 2004: 99]. Как инструмент для описания знаний, онтологии должны заинтересовать всех специалистов, в том числе и преподавателей, сталкивающихся в своей практике с проблемой представления и использования знаний.

Что же представляют собой онтологии? Из основ философии известно, что термин онтология используется для обозначения системы знаний, относящихся к окружающему нас миру. Другими словами, онтология, в философском понимании, это наука о бытии, наука о природе вещей и взаимосвязях между ними. С развитием информационных технологий и биоинформатики, в частности, этот термин вновь прочно вошел в обиход людей, занятых проблемой интеграции информационных ресурсов. В контексте информационных технологий представления знаний, термином онтология можно определить некоторый механизм или способ, используемый для описания некоторой области знаний (предметной области), в частности базовых понятий этой области и связей между ними [Gruber 1995: 907]. Онтологии создаются с двумя первичными и прагматичными целями: первая – ускорить общение между людьми и организациями; вторая – улучшить взаимосвязь между системами. Важное условие при создании онтологий – они должны быть достаточно удобны всем участникам взаимодействия в сети Интернет – и людям, и программным системам. Таким образом, онтология является, можно сказать, «машиночитаемой» моделью некоторой предметной области. Онтологии позволяют не только и не столько структурировать содержимое информационных массивов (они уже структурированы, эту функцию выполняют модели баз данных и/или метаданные), сколько получить полное представление о содержимом информационного ресурса. Следует отметить, что поскольку созданием онтологий ученые занялись относительно недавно, дебаты о том, что является «правильной» онтологией и как она должна быть выполнена и реализована продолжаются и по сей день [Rubin et al. 2008: 75].

Современные биомедицинские онтологии, как правило, представляют собой словарь терминов-концептов, формирующих описание определенной области знаний. Наиболее известной и широко используемой биоонтологией является Генная онтология [Gene Ontology Consortium 2001: 1425]. Она представляет собой совокупность трех огромных контролируемых словарей терминов, касающихся молекулярной функции, биологического процесса и клеточной структуры. Каждый термин в этих трех онтологиях связан с конкретными примерами – генами, определяющими ту или иную функцию, процесс или структуру. Следует подчеркнуть, что эти словари могут быть использованы для описания генных продуктов в любом организме, поэтому многие базы данных используют Генную Онтологию для навигации внутри них. В результате, выбор определенного термина-концепта Генной Онтологии позволяет сразу же выйти на соответствующие гены и их продукты, обнаруженные в различных организмах. Онтологический анализ представляет собой высокоэффективную альтернативу стандартно применяемому поиску в различных поисковых системах, который зачастую приводит к избыточности информации. Удобство таких «биологических словарей», понимаемых компьютером бесспорно не только с точки зрения формализации биологического знания и автоматизации процесса поиска, но и для изучения практически любой биологической дисциплины или отдельного биологического процесса.

Помимо Генной онтологии разрабатываются специализированные онтологии для отдельных организмов. Например, с целью детального аннотирования данных биологии растений в настоящее время активно разрабатываются онтологии по анатомии, морфологии и стадиям роста покрытосеменных растений [Avraham et al. 2008: D499]. В частности, онтология стадий роста растений позволяет аннотировать и интегрировать данные, полученные не только с использованием *Arabidopsis,* но и многих злаковых [Pujar et al. 2006: 414]. Например, используя эту онтологию можно найти и сопоставить известные на сегодняшний день ключевые гены, участвующие в регуляции перехода к цветению у арабидопсиса и у риса, что является несомненным подспорьем при изучении этой проблемы. Следует отметить, что использование онтологий позволяет гораздо быстрее сориентироваться при изучении такой дисциплины как биология развития растений, особенностью которой является экспоненциально растущий поток молекулярно-биологической информации.

Любая онтология, как правило, является авторским набором концептов и отношений, поэтому разные эксперты-пользователи в зависимости от интересующей их проблемы могут создавать различные онтологии для одного и того же авторского набора данных. Перспективой развития онтологического анализа может быть создание частных, доменных онтологий, которые могут позволить пользователю по-новому взглянуть на имеющиеся данные и рассматривать их, возможно, даже в ином контексте. Иными словами, вместо одного общего «содержания», по которому пользователь пытается найти необходимую ему информацию, в результате создания доменных онтологий мы можем иметь дело с несколькими детализированными «подразделами». Таким образом, применяя ту или иную онтологию к информационным ресурсам, исследователи смогут получать разные знания. Предложенная и протестированная нами ранее модель совместной работы с помощью онтологий [Strizh et al. 2007: 428] позволяет говорить, что использование и построение частных биоонтологий может являться не только эффективным звеном между валовым получением экспериментальных данных и структурированным научным знанием, но и необходимым элементом биологического образования в постгеномном веке (Рис. 2).

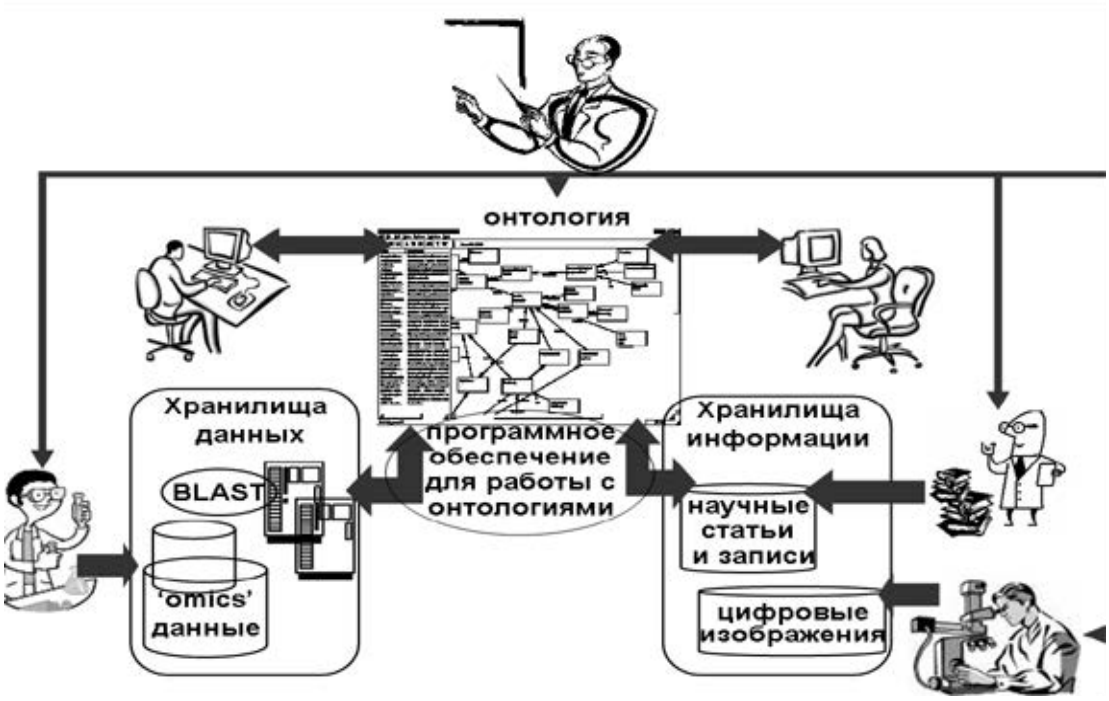

Рис. 2. *Онтология как связующее звено между экспериментаторами, биоинформаторами и преподавателями (с изменениями по [Strizh et al. 2007: 428])*

Список использованной литературы

1. **Жучков А. В., Арнаутов С. А., Твердохлебов Н. В., Голицын С. В., Стриж И. Г.** Использование онтологий при работе с гетерогенными федеративными массивами данных в распределенных информационных системах // Сборник научных трудов «Распределенные вычисления и грид-технологии в науке и образовании». – Дубна, 2004. - С. 99-103.
2. **Кафтанников И. Л., Коровин С. Е.** Перспективы использования web-онтологий в учебном процессе // Educational Technology & Society. - 2003. – Т. 6. - Н. 3. - С. 134-138.
3. **Avraham S., Ilic K., Jaiswal P. et al.** The Plant Ontology Database: a Resource for Plant Structure and Developmental Stages Controlled Vocabulary and Annotations // Nucleic Acids Research. - 2008. - V. 36. – D 449–D 454.
4. **Galperin M. Y.** The Molecular Biology Database Collection: 2008 Update // Nucleic Acids Research. - 2008. – V. 36. – D 2-D 4.
5. **Gene-Ontology-Consortium. Creating the Gene Ontology Resource: Design and Implementation** // Genome Res. - 2001. - V. 11. - P. 1425-1433.
6. **Gruber T. R.** Toward Principles for the Design of Ontologies Used for Knowledge Sharing // Int. J. Hum. Computer Stud. – 1995. - V. 43. - P. 907-928.
7. **Jutchkov A., Tverdokhlebov N., Strizh I., Arnautov S., Golityin S.** Grid-Based Onto-Technologies Provide an Effective Instrument for Biomedical Research // Studies in Health Technology and Informatics. - 2005. - V. 112. - P. 37-46.
8. **Pujar A., Jaiswal P., Kellogg E. A. et al.** Whole-Plant Growth Stage Ontology for Angiosperms and its Application in Plant Biology // Plant Physiology. - 2006. - V. 142. - P. 414–428.
9. **Rubin D. L., Shah N. H., Noy N. F.** Biomedical Ontologies: a Functional Perspective // Briefings in Bioinformatics. – 2008. - V. 9. – № 1. - P. 75–90.
10. **Strizh I., Jutchkov A., Tverdokhlebov N., Golitsyn S.** Systems Biology and Grid Technologies: Challenges for Understanding Complex Cell Signaling Networks // FGCS. - 2007. - V. 23. - P. 428–434.
11. **Strizh, I. G.** Ontologies for Data and Knowledge Sharing in Biology: Plant ROS Signaling as a Case Study // BioEssays. - 2006. - V. 28. - № 2. - P. 199–210.