УДК 004.04 + 004.22 + 004.415.2 + 004.658.2

ПРОГРАММНАЯ СИСТЕМА И ИНФОРМАЦИОННЫЙ БАНК МЕТАДАННЫХ ТРЕТИЧНЫХ СТРУКТУР БЕЛКОВ

Т.А. Никитин, Ю.Б. Порозов

Приведена архитектура модели базы хранения метаданных о результатах проверок трехмерных структур белков. Построена концептуальная модель базы данных. Представлен сервис и процедура обновления базы данных, а также алгоритмы преобразования данных о записях структур белков и их качестве. Приведены наиболее важные сведения о записях и форме их представления для хранения, выборки и предоставления пользователям. Разработан комплекс программного обеспечения для реализации функциональных задач с использованием языка программирования Java в среде разработки NetBeans v.7.0, а также языка JQL для формирования запросов и взаимодействия с базой данных JavaDB. Проведено тестирование сервиса. Результаты работы системы показали ее эффективность при фильтрации записей структур белков PDB.

Ключевые слова: программная система, информационный банк, структура белка, PDB, качество структуры, сервер проверок, метаданные.

Введение

В настоящее время идет бурное развитие многих областей современной науки вследствие совершенствования техник и технологий, проведения экспериментов, требующих хранения огромного массива получаемых данных. Одним из таких направлений является биоинформатика – наука, занимающаяся разработкой методов анализа, управления и обработки данных о биологических последовательностях и их взаимосвязях. Эти данные хранятся в специально созданных информационных банках (ИБ), функционирующих на базе научно-исследовательских институтов или университетов, ведущих экспериментальные работы в области биохимии и молекулярной биологии [1].

В последние годы наблюдается взрывной рост количества записей в ведущих базах данных (БД) нуклеотидных последовательностей и структур белков. Как следствие, остро встает вопрос об администрировании, упорядочивании, фильтрации и периодической ревизии общедоступных некоммерческих БД, которые часто хранят неполные сведения, результаты, полученные на устаревшем оборудовании или имеющие ошибки. Наличие таких записей в результатах поиска приводит к неверным расчетам и выводам, ошибкам в постановке экспериментов и излишним временным и материальным затратам на ручную сортировку.

Важно отметить, что задача состоит не только в сборе информации и ее хранении, но и в предоставлении сервисов для выборки и анализа данных, выявлении в них закономерностей и взаимосвязей, построении регуляторных моделей и моделей для предсказания структуры, функции и клеточного расположения [2].

В связи с этим разработка нового программного обеспечения и совершенствование способов описания, хранения, простого и быстрого поиска информации о последовательностях и структурах становятся чрезвычайно актуальными [3].

Описываемая разработка имеет целью существенно облегчить работу специалистов в области структурной биологии белков, предоставляя только заранее проверенные и ранжированные записи из общедоступных БД, отсеивая или предупреждая о сомнительных, неполных или ошибочных данных.

Обзор существующих программных систем и БД хранения информации о последовательностях и структурах

В настоящее время существуют сотни веб-сайтов, предоставляющих данные по различным аспектам геномики, протеомики, системной биологии и средства их анализа. Многие из них имеют свой формат хранения данных, степень избыточности и связи с родственными или аналогичными БД. Каждый ресурс имеет свои средства работы с информацией – различные поисковые программы, программные средства визуализации, пополнения базы. Крупнейшие хранилища нуклеотидных и аминокислотных последовательностей пополняются аннотированными записями непосредственно исследователями и лабораториями. Однако в курируемых БД новая информация проверяется обслуживающим ее персоналом [4].

Необходимо подчеркнуть, что в отличие от традиционной библиографической научно-технической информации, собираемой и распространяемой на печатных носителях и в электронной форме, данные в биоинформатике являются фактографическими и гораздо более тесно привязаны к источникам их происхождения. В связи с этим большинство биоинформационных БД созданы и поддерживаются крупными центрами, например, European Molecular Biology Laboratories [5], European Bioinformatics Institute [6], National Center for Biotechnological Information [7], DNA DataBank of Japan. В то же время научные группы и лаборатории, ведущие работы в области биохимии, молекулярной и структурной биологии, также поддерживают свои БД. Как правило, в первом случае обеспечивается функционирование общих банков данных, содержащих информацию о последовательностях белков и нуклеиновых кислот, а во втором – специализированных банков данных.

В зависимости от содержания различают несколько типов биологических БД. Первый тип – архивные: GenBank [8], EMBL, Protein Data Bank (PDB) [9]. Второй тип – курируемые, к которым относится Swiss-Prot – наиболее качественная БД, содержащая аминокислотные последовательности белков. Третий тип – производные БД, получающиеся в результате обработки данных из архивных и курируемых баз: SCOP (структурной классификации белков), PFAM (семейств белков), Gene Ontology, ProDom (белковых доменов). Благодаря такому разделению обеспечивается высокий уровень экспертной оценки поступающих и сохраняемых данных.

Вместе с тем в настоящее время остро стоит проблема интеграции и фильтрации информации, содержащейся в различных базах, так как большое количество записей содержит фрагменты белков, устаревшие данные, дублирующие записи (структура одного и того же белка в комбинации с разными лигандами или фрагментами других молекул), записи низкого разрешения с ошибками структур различного рода или не соответствующие спецификации PDB, что затрудняет работу с ними.

Материалы и методы

Разработанное программное обеспечение собирает данные о результатах проверок записей белковых структур и поддерживает в актуальном состоянии БД с ними. Созданный сервис позволяет предоставить отфильтрованные по критерию качества данные, соответствующие заданным параметрам поиска. Потенциал сервиса проявляется в комплексном взаимодействии с банками данных, хранящими информацию о трехмерных структурах белков, сервисами их проверок и учете запросов в соответствии с индивидуальными потребностями пользователей сервиса.

Выбранная для работы сервиса мировая база PDB предоставляет информацию о структурах в форматах PDB и mmCIF – координаты атомов и дополнительную информацию (ротамеры, элементы вторичной структуры, ссылки на статьи и т.д.) [10].

Чтобы сравнить записи конкретного белка и организовать работу системы, необходимо иметь все индексы записей PDB. Актуальные индексы представлены в файле на специализированном сервисе WHY_NOT [11], который индексирует все записи в банке PDB и составляет списки доступных записей, предоставляя тем самым удобный способ получения информации и статистики. Сами записи белковых структур хранятся в различных форматах, например, в сжатом виде (*.gz), и доступны для загрузки [11].

Проверки на сервере What If. Анализ качества белковых структур

Для качественного сравнения трехмерных структур белков был использован онлайн-сервер What If [12], предоставляющий результаты следующих проверок:

- общих сведений в заголовке записи (имя, номенклатура, потерянные атомы);
- сведений о симметрии (количество молекул и состав объекта);
- параметров геометрии (межатомные расстояния, длины ковалентных связей, параметры планарных и торсионных углов);
- структурных параметров (ошибки вторичной структуры в соответствии с DSSP, качество упаковки, положение атомов, углов и аномальных связей, анализ *B*-фактора).

Результаты проведенных проверок находятся на сервере What If в открытом доступе. Для каждой белковой записи хранятся результаты трех проверок: оригинальной записи из PDB, записи после консервативной и полной оптимизации. Авторами были отобраны 10 оценок, отражающих качество анализируемых структур: разрешение структуры (Å); общая оценка качества; качество структуры 1 и 2 уровней; распределение по диаграмме Рамачандрана; характеристики углов χ^1 и χ^2; конформация основной цепи; длины ковалентных связей; планарные углы; число столкновений атомов; распределение структуры.

Информация об используемых данных, доступных в явном виде, помогающая находить объект, называется метаданными, которые особо важны, поскольку их использование позволяет повысить качество поиска. При этом хранимые данные не дублируют исходные ресурсы, а производится сбор и хранение лишь выборочной информации для получения ускоренного доступа к оригиналу [13]. Сохраняемые метаданные содержат название исходной записи, ее уникальный идентификатор PDB, дополнительную информацию и результаты проверок. Главный параметр записи – ее идентификатор, представляющий собой четырехсимвольную последовательность, первый символ которой является цифрой, а последующие могут быть буквами или цифрами. Названием белковой записи обычно является одно или несколько слов, определяющих название структуры. Дополнительная информация берется из раздела заголовка PDB-файла TITLE. Ссылки на адреса самой структуры и дополнительной информации о ней, серверы проверок и основных используемых ресурсов в сети Интернет хранятся в виде URL-ссылок.

База данных

Разработанная БД представляет собой ИБ метаданных, организованных в логическую структуру формального вида для эффективного поиска и обработки данных. В данном случае целесообразно использование централизованной БД для минимизации затрат ресурсов и времени на получение данных [14].

Для обеспечения минимальной логической избыточности БД при проектировании отношений учитывалась совокупность требований – нормальные формы. БД создана в соответствии с тремя базовыми требованиями классической реляционной модели данных, которая является фактическим стандартом для современных систем управления базами данных (СУБД) [15]. Она была взята за основу благодаря своим достоинствам, таким как упрощение схемы БД (таблицы), логическая и физическая независимость, доступ к БД на языках высокого уровня, улучшение целостности и защиты данных. Поскольку реляционная структура концептуально проста, она позволяет реализовывать небольшие и простые БД [16]. Концептуальная модель хранилища данных представлена главными сущностями и отношениями между ними, используемыми при разработке модели (рис. 1).

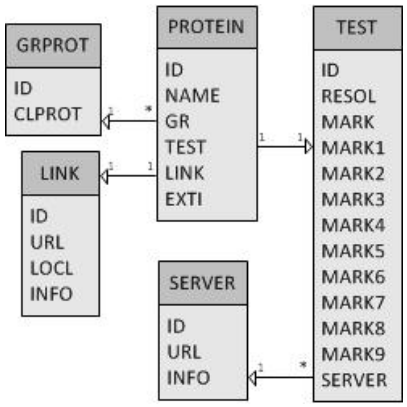

Рис. 1. Концептуальная модель

Модель содержит пять основных сущностей, используемых для обеспечения минимальных требований, предъявляемых к работе системы и ее функциям: PROTEIN, GRPROT, LINK, TEST, SERVER, главной из которых является таблица PROTEIN, содержащая записи метаданных белковых структур: идентификатор записи, имя и дополнительную информацию о ней. Связь каждой записи с ее группой по первичному ключу, находящемуся в таблице GRPROT, является связью «многие-к-одной», поскольку записи входят в состав группы (принадлежат определенной группе). Для каждой записи белка в таблице PROTEIN имеется связь с записями хранения ссылок по первичному ключу в таблице LINK, являющаяся отношением «один-к-одному», поскольку в ней хранятся уникальные ссылки.

Основная таблица PROTEIN содержит шесть полей (уникальный идентификатор PDB, имя белковой записи, идентификаторы для связи с таблицами хранения группы, тестов, ссылок и дополнительную информацию). Таблица LINK содержит ссылки на локальные и внешние источники, а также дополнительные сведения о них. Таблица GRPROT содержит информацию о группах записей белковых структур и их названия.

В таблице TEST хранятся результаты проверок, связанных первичным ключом с соответствующей записью таблицы PROTEIN (отношение «один-к-одному»). Записи о проверках качества на различных серверах связаны с записями о самом сервере по ключу таблицы SERVER (отношение «многие-к-одному»), хранящей адреса серверов проверок и дополнительную информацию о них.

Типы полей разработанной БД обусловлены хранимой в них информацией.

Доступ к базе данных

При интеграции приложения с БД решалась задача организации максимально прозрачного доступа к ней. Один из подходов к разрешению этой задачи заключается в использовании так называемых тонких клиентов. В этом случае клиентское приложение обеспечивает реализацию презентационной логики, а сервер объединяет логику доступа к ресурсам и бизнес-логику. Интерфейс пользователя разрабатывался при помощи языка гипертекстовой разметки HTML с использованием HTML-форм, которые являются наиболее удобным механизмом передачи запросов.

Доступ к БД был обеспечен на стороне веб-сервера, который динамически формирует представление, используя классы Java. Существует универсальный способ доступа к реляционным данным – разработанный компанией Sun ряд стандартных Java-классов [17]. Он позволяет выполнять запросы к БД и ее модификацию непосредственно из Java-программ: классов связанности баз данных – пакетов JDBC (Java Database Connectivity), которые имеют описание всех свойств СУБД, доступных с помощью драйверов JDBC, и обеспечивают абсолютную независимость приложения от БД благодаря инструментам администрирования СУБД в Java [18].

Преимущество использованного J2EE (Java 2 Platform, Enterprise Edition) состоит в том, что модель приложения заключает уровни функциональных возможностей в определенные типы компонентов [19]. Взаимодействие с клиентом осуществляется посредством обычных веб-страниц с использованием

апплетов на основе Java-технологии – Java Server Pages (JSP), или посредством автономных Java-приложений. Веб-приложение Java создает интерактивные веб-страницы, содержащие различные типы языков разметки, например HTML, а динамическое содержимое формируется веб-компонентами – страницами JSP [20], сервлетами и средствами JavaBeans, которые позволяют изменять данные, осуществлять их временное хранение, взаимодействовать с БД и веб-службами.

Для предоставления качественного сервиса пользователям, содержимое БД должно быть актуальным и своевременно обновляться, а также иметь удобный интерфейс и продуманную форму представления.

Сервис актуализации состояния базы данных

Раз в неделю по четвергам на сервере PDB обновляются базы хранения белковых структур. По этой причине целесообразно также раз в неделю запускать сервис обновления. Работа данной функции начинается со сканирования файлов на сервере PDB, содержащих информацию о новых записях. Сервис обращается к европейскому зеркалу мировой базы белковых структур (наиболее близкому, а, следовательно, и быстрому по отношению к нашему серверу) и сканирует ее на предмет наличия новых записей. После этого происходит обращение к серверу проверок What If и просмотр информации о наличии результатов проверок новых записей белковых структур. В случае обнаружения новых записей сервис считывает необходимую информацию о новой белковой структуре, результатах ее проверок и записывает их в БД (рис. 2). После успешного обновления и вывода идентификаторов новых записей, пользователь может найти их по поиску или просмотреть подробную информацию о них (информация о записи 2EGX приведена на рис. 3).

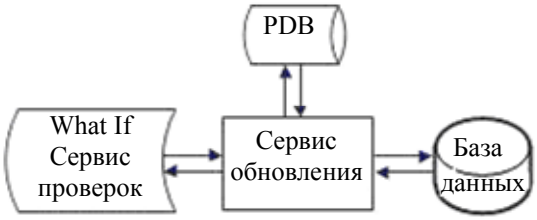

Рис. 2. Процедура обновления состояния базы данных

ID	null
Name	CRYSTAL STRUCTURE OF THE PUTATIVE ACETYLGLUTAMATE KINASE FROM THERMUS THERMOPHILUS
Classification	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
SpaceGroup	P 65
Test service	PDBe
Resolution	1.919
R-value	0.2249
1st generation packing quality	0.823
2nd generation packing quality	-0.0689
Ramachandran plot appearance	-0.725
Chi-1/Chi-2 rotamer normality	-0.4029
Backbone conformation	0.2249
Bond length RMS Z-score	0.2569
Bond angle RMS Z-score	0.5849
Total number of bumps	0.2311
Inside/Outside distribution	70

Рис. 3. Подробная информация о белке 2EXG

Сервис обновления не дублирует записи, а сохраняет их в виде метаданных в локальной БД и использует их, ссылаясь на расширенные отчеты результатов проверок [21] и сами записи в PDB. Ссылки формируются на страницах представления благодаря использованию шаблонов адресов.

Тестирование системы и результаты

Для подтверждения работоспособности разработанной программной системы была проведена проверка обновления БД с последующим контролем появления новых записей в ней и их соответствия реальным данным. Проверка работоспособности системы включала в себя тесты на корректность представления метаданных и соответствующей информации из БД на всех страницах веб-модуля и тест работоспособности ее EJB-модуля.

Была выполнена выборка данных, удовлетворяющих критериям поиска по определенным параметрам на поисковой странице системы. В качестве теста был произведен поиск в локальной базе по записям, содержащим в своем имени строку «calmodulin». Поисковый запрос по ключевому слову «calmodulin» в базе PDB находил 377 записей. В отличие от нее, в разработанной и обновленной базе данных программной системы поиск выдает 90 записей, поскольку, как уже говорилось ранее, не все записи проходят проверку качества и имеют оценки, многие из этих записей устарели или имеют неправильный формат данных. Результатом поиска являются представленные в табличной форме отфильтрованные записи, содержащие поля «идентификатор PDB», «разрешение», «z-score» (кумулятивная оценка качества) и ссылки на внешние источники (рис. 4). Для таблиц вывода записей в системе существует возможность их сортировки по полям и ограничению количества записей на одной странице с указанием максимального количества записей по конкретному запросу в целом и смещения этого интервала относительно найденных записей.

ID	Name	Res	Z-score	SpaceGroup	Links
1exr	THE 1.0 ANGSTROM CRYSTAL STRUCTURE OF CA+2 BOUND CALMODULIN	1.0	0.134	P 1	
2o5g	CALMODULIN-SMOOTH MUSCLE LIGHT CHAIN KINASE PEPTIDE COMPLEX	1.08	0.142	P 1 21 1	
2pq3	N-TERMINAL CALMODULIN ZN-TRAPPED INTERMEDIATE	1.299	0.143	P 43 21 2	
3gof	CALMODULIN BOUND TO PEPTIDE FROM MACROPHAGE NITRIC OXIDE SYNTHASE	1.45	0.1449	P 1 21 1	
2jam	CRYSTAL STRUCTURE OF HUMAN CALMODULIN-DEPENDENT PROTEIN KINASE I G	1.7	0.155	P 1	
3qp2	CALMODULIN BOUND TO PEPTIDE FROM CALMODULIN KINASE II (CAMKII)	1.46	0.1599	C 1 2 1	
2vwi	CRYSTAL STRUCTURE OF SU6656-BOUND CALCIUM/CALMODULIN- DEPENDENT PROTEIN KINASE II DELTA IN COMPLEX WITH CALMODULIN	1.899	0.1606	P 21 21 21	
2vz6	STRUCTURE OF HUMAN CALCIUM CALMODULIN DEPENDENT PROTEIN KINASE TYPE II ALPHA (CAMK2A) IN COMPLEX WITH INDIRUBIN E804	2.299	0.1633	P 21 21 21	
1vrk	THE 1.9 ANGSTROM STRUCTURE OF E84K-CALMODULIN RS20 PEPTIDE COMPLEX	1.899	0.171	P 1 21 1	
3gwr	CRYSTAL STRUCTURE OF PUTATIVE CALCIUM/CALMODULIN-DEPENDENT PROTEIN KINASE TYPE II ASSOCIATION DOMAIN (YP_315894.1) FROM THIOBACILLUS DENITRIFIC	2.009	0.1729	P 21 21 21	

Рис. 4. Вывод ограниченного количества записей, отсортированных по оценке качества структур z-score

Заключение

Основной целью проекта являлась разработка программного обеспечения и базы данных, предоставляющих отсортированные по критерию качества структуры записи PDB. Известно, что при поиске по ключевым словам в PDB пользователь обычно получает несколько записей (зачастую несколько десятков), выбрать лучшую из которых непросто. Разработанная система актуализации представления метаданных и сервис динамического пополнения новыми записями и данными о результатах проверки качества структур помогает пользователю быстро находить несколько наиболее качественных структур среди всех записей белка (и его частей) в PDB. Была разработана архитектура и создана база данных и служба ее актуализации. В качестве фундамента была выбрана база данных JavaDB. Связь с ней была организована посредством JDBC, а языком программной реализации был выбран язык высокого уровня Java для обеспечения совместимости приложения с любыми операционными системами на уровне исполняемого кода.

Проект выгодно отличается по простоте использования от аналогичных сервисов – в нем нет сложной структурной организации возможностей использования и поиска по многочисленным критериям выборки. Разработанная программная система использует результаты работы сервиса онлайн-проверки белковых структур, что обеспечивает точность оценки качества. В то же время она достаточно просто организована для облегчения понимания результатов и ускорения принятия решений о выборе той или иной структуры для работы. Созданная система является своего рода «фильтром», помогающим в поиске наиболее качественных структур среди многих, зачастую одинаковых, неполных, устаревших или ошибочных данных. Интуитивно понятный графический и простой текстовый интерфейсы позволяют пользователю быстро разобраться с основными принципами работы сервиса и начать его эксплуатацию в повседневной работе совместно с сервисами PDB, что способствует сокращению временных затрат на поиск и выборку необходимой информации из PDB.

Дальнейшая работа над комплексом предполагает решение таких задач, как обеспечение взаимодействия с другими сервисами и расширение набора хранимых данных для более гибкого задания параметров выборки посредством подключения иных сервисов онлайн-проверок, информационных банков хранения информации о белковых структурах и увеличения количества атрибутов записей, расширив их представление в базе метаданных.

Исследование поддержано грантом Министерства образования и науки Российской Федерации (ГК № 07.514.11.4163).

Литература

1. Леск А. Введение в биоинформатику. – М.: Бином. Лаборатория знаний, 2009. – 324 с.
2. Игнасимуту С. Основы биоинформатики. – М.: Ижевск: НИЦ «Регулярная и хаотическая динамика». Институт компьютерных исследований, 2007. – 320 с.
3. Афанасьева Г. Биоинформатика: виртуальный эксперимент в шаге от реальности // Наука и жизнь. – 2004. – № 11. – С. 20–24.

4. Молекулярно-биологические базы данных // Объединенный центр вычислительной биологии и биоинформатики [Электронный ресурс]. – Режим доступа: http://jcbi.ru/baza/index.shtml, свободный. Яз. рус. (дата обращения 16.05.2012).
5. Stoesser G., Baker W., van den Broek A. et al. The EMBL Nucleotide Sequence Database // Nucleic acids research. – 2002. – V. 30. – № 1. – P. 21–26.
6. Emmert D.B. The European Bioinformatics Institute (EBI) databases // Nucleic acids research. – 1994. – V. 22. – № 17. – P. 3445–3449.
7. Geer L.Y., Marchler-Bauer A., Geer R.C. et al. The NCBI BioSystems database // Nucleic acids research. – 2010. – V. 38 (Database issue). – D492–496.
8. Benson D.A., Karsch-Mizrachi I., Lipman D.J. et al. GenBank // Nucleic acids research. – 2005. – V. 33. – suppl. 1. – D34–38.
9. Bernstein F.C. The Protein Data Bank: a computer-based archival file for macromolecular structures // Journal of molecular biology. – 1977. – V. 112. – № 3. – P. 535–542.
10. Henrick K., Feng Z., Bluhm W. et al. Remediation of the Protein Data Bank Archive // Nucleic acids research. – 2008. – V. 36. – Suppl. 1. – D426–D433.
11. Joosten R.P., Beek T.A.H., Krieger E. et al. A series of PDB related databases for everyday needs // Nucleic acids research. – 2011. – V. 39. – Suppl. 1. – D411–D419.
12. Vriend G. WHAT IF: a molecular modeling and drug design program // Journal of Molecular Graphics. – 1990. – V. 8. – Is. 1. – P. 52–56.
13. Башмаков А.И., Старых В.А. Систематизация информационных ресурсов для сферы образования: классификация и метаданные. – М.: Европейский центр по качеству, 2003. – 384 с.
14. Дейт К.Дж. Введение в системы баз данных. – 8-е изд. – М.: Вильямс, 2005. – 1328 с.
15. Горев А., Ахаян Р., Макашарипов С. Эффективная работа с СУБД. – СПб: Питер, 2006. – 704 с.
16. Кузнецов С.Д. Базы данных: языки и модели: Учебник. – М.: Бином. Лаборатория знаний, Интернет-университет информационных технологий, 2008. – 720 с.
17. Монахов В. Язык программирования Java и среда NetBeans. – 2-е изд. – СПб: БХВ-Петербург, 2009. – 720 с.
18. Хейк Б. JDBC: Java и базы данных. – М.: Лори, 1999. – 320 с.
19. Ноутон П., Шилдт Г. Java 2. – СПб: БХВ-Петербург, 2008. – 1072 с.
20. Steelman A., Murach J. Murach's Java Servlets and JSP. – 2nd ed. – Fresno, CA, USA: Mike Murach & Associates, 2008. – 729 p.
21. Hooft R. A WHAT IF check report: what does it mean. – 2007 [Электронный ресурс]. – Режим доступа: http://swift.cmbi.ru.nl/gv/pdbreport/checkhelp/explain.html, свободный. Яз. рус. (дата обращения 16.05.2012).

Никитин Тимофей Александрович – Санкт-Петербургский государственный политехнический университет, студент, tim04k@gmail.com

Порозов Юрий Борисович – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кандидат медицинских наук, зав. лабораторией, porozov@ifc.cnr.it