УДК 577.38 577.3.0 577.322.4

РАСПРЕДЕЛЕННАЯ СИСТЕМА RESTful-WEB-СЕРВИСОВ ДЛЯ РЕКОНСТРУКЦИИ И АНАЛИЗА ГЕННЫХ СЕТЕЙ

© 2012 г. **Н.Л. Подколодный[1], А.В. Семенычев[1], Д.А. Рассказов[1], В.Г. Боровский[1], Е.А. Ананько[1], Е.В. Игнатьева[1], Н.Н. Подколодная[1], О.А. Подколодная[1], Н.А. Колчанов[1, 2, 3]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: pnl@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия;
[3] НИЦ «Курчатовский институт», Москва, Россия

Поступила в редакцию 5 июля 2012 г. Принята к публикации 25 июля 2012 г.

В данной работе описывается распределенный программный комплекс на основе RESTful-Web-сервисов, который ориентирован на решение задач реконструкции генных сетей на основе интеграции данных из гетерогенных источников информации, включая базы данных о молекулярно-генетических взаимодействиях, метаболических и сигнальных путях, генных сетях и т. д.

Ключевые слова: распределенные системы, RESTful-Web-сервисы, генные сети, интеграция данных, анализ графов генных сетей.

ВВЕДЕНИЕ

Молекулярно-генетические системы характеризуются огромным разнообразием молекулярных механизмов, обеспечивающих их функционирование, включая транскрипцию, процессинг (созревание) РНК, трансляцию (синтез полипептидных цепей), процессинг белков, ДНК-белковые, РНК-белковые, белок-белковые, лиганд-белковые и другие взаимодействия, процессы метаболизма, передачи сигналов, транспорта, деградации и т. д.

К настоящему времени в области биоинформатики и системной биологии мировым сообществом разработано более 1400 баз данных, многие из которых полезны при описании молекулярно-генетических систем и генетических механизмов их функционирования, включая базы данных по молекулярным объектам (гены, РНК, белки), молекулярно-генетическим взаимодействиям и процессам, онтологиям, метаболическим путям и путям передачи сигналов в клетке, генетической регуляции молекулярных процессов и систем, генным сетям, экспрессии генов в различных клеточных условиях и под действием различных индукторов и т. д. (Galperin, 2012).

Интеграция данных из этих Интернет-доступных гетерогенных источников информации о молекулярно-генетических взаимодействиях и генных сетях и реконструкция на этой основе генных сетей являются важнейшей задачей биоинформатики и системной биологии.

Генные сети – это молекулярно-генетические системы, обеспечивающие формирование фенотипических характеристик организмов (молекулярных, биохимических, физиологических, морфологических, поведенческих и т. д.) на основе информации, закодированной в их геномах (Kolpakov *et al*., 1998; Ananko, 2005). Обычно генные сети состоят из сотен и тысяч элементов, объединенных сложными процессами взаимодействия.

Анализ структуры генных сетей, выявление закономерностей структурно-функциональной организации генных сетей, выделение подсистем, редукция описания молекулярно-генети-

ческих систем и построение на этой основе структурной модели генной сети являются важнейшими этапами в исследовании генных сетей и первым шагом в создании математических моделей динамики генных сетей (Newman, 2006).

Проблемы реконструкции и анализа структурно-функциональной организации генных сетей

Современные технологии реконструкции генных сетей основываются на:

1) использовании специализированных графических редакторов, обеспечивающих возможность пользователю вводить и редактировать информацию о молекулярно-генетических объектах, реакциях, генетических регуляциях, генных сетях;

2) интеграции данных о молекулярно-генетических взаимодействиях из различных источников информации (баз данных);

3) использовании методов теоретического предсказания молекулярно-генетических взаимодействий.

Для этих целей крайне важной информацией является онтологическое описание молекулярно-генетических объектов, систем и процессов, включая классификацию генов, белков, ферментов, метаболических путей и генных сетей.

Обычно процесс реконструкции генной сети или метаболического пути начинается с постановки задачи, которая включает описание проблемы, выявление целей, определение границ задачи, класса реконструируемой генной сети, формирование запроса на поиск начального множества элементов генов сети и т. п.

При реконструкции генной сети необходимо учитывать, что генная сеть – это сетевая модель взаимодействий функционирующих генов, которая описывает молекулярно-генетическую систему, обеспечивающую выполнение определенной функции клетки или ее подсистем при определенных условиях, состояниях организма или клетки, при воздействии внешних индукторов, взаимодействии с другими клетками или организмами, реализации определенных молекулярных событий на определенных стадиях молекулярно-генетических процессов и т. д.

Последовательность ДНК дает только комбинаторику возможных вариантов функционирования генов. Следующие уровни регуляции работы генов в клетке (состояние хроматина, различные модификации ДНК, гистонов, структура и локализация хромосомы и т. д.) задают специфические ограничения на эти возможности. Для каждой клетки эти ограничения в общем случае различаются и могут динамически изменяться с собственными характерными временами. Таким образом, работа генов существенно зависит от типа и состояния клетки. Поэтому при реконструкции генных сетей необходимо заранее определить тип клеток, тканей либо вид организма, в которых она реализуется. Информация, полученная на других типах клеток, тканях и организме, также может быть полезна при реконструкции генной сети. Однако необходимо обосновать такую возможность и оценить степень адекватности этой информации. Более того, внешние условия, при которых были получены экспериментальные данные, также должны быть строго определены. Только в этом случае можно оценить возможность их интеграции и совместного анализа.

Генные сети можно разделить на пять основных типов (Колчанов и др., 2000): контролирующие гомеостаз; регулирующие циклические процессы; обеспечивающие стрессовый ответ; контролирующие необратимые процессы и генные сети-интеграторы.

Разные типы генных сетей имеют различные структуру и состав, которые необходимо учитывать в процессе реконструкции. Например, генные сети, описывающие поддержание некоторого гомеостатического состояния, должны включать сенсорный элемент, отслеживающий состояние гомеостатируемого параметра, а также пути поступления и утилизации продуктов, определяющих состояние этого параметра.

Генная сеть, описывающая реакцию клетки на некоторое воздействие или сигнал, должна включать рецепторы, через которые этот сигнал передается компонентам генной сети, сам путь передачи сигнала и исполняющие элементы.

Одним из первых шагов в реконструкции генной сети может быть определение множества генов, которые задают ядро реконструируемой генной сети или метаболические пути, которые должны входить в генную сеть.

Проблемы анализа структурно-функциональной организации генных сетей

Реализация RESTful-Web-сервисов для реконструкции и анализа генных сетей REST (REpresentational State Transfer) представляет собой архитектурный стиль для создания ресурс-ориентированных распределенных программных систем, основанных на архитектуре клиент–сервер и, как правило, используется для построения Web-сервисов или RESTful-Web-сервисов (Richardson, 2007).

Архитектурный стиль REST включает ряд рекомендаций или ограничений, налагаемых на архитектуру, оставляя реализацию индивидуальных компонентов свободной (Richardson, 2007; Valverde, 2009; Subbu Allamaraju, 2010; Schreier, 2011).

Адресуемость. Основным понятием в REST-архитектуре являются ресурсы как источники конкретной информации, каждый из которых определяется ссылкой с глобальным идентификатором URI. Ресурсом является все, что имеет ссылку.

Отсутствие состояния сервиса. Вся информация, необходимая для выполнения запроса, содержится в самом запросе. Информация о предыдущих запросах сервером не сохраняется и не используется. Сервер не поддерживает сеанс и не фиксирует его состояние. Вся информация о состоянии сессии поддерживается при необходимости клиентом. Однако сам ресурс имеет определенное состояние, которое может изменяться в результате выполнения запросов клиента или по другим внешним причинам. Состояние на стороне сервера адресуемо через URI как ресурс. Это делает серверы не только более видимыми для мониторинга, но и более надежными в случае частичного отказа сети, а также дополнительно улучшает их масштабируемость.

Связность. Информация о связях между ресурсами может использоваться клиентом для обнаружения идентификаторов других связанных с запросом ресурсов, в том числе ссылок на автоматически созданные ресурсы, которые являются результатом обработки данных.

Унифицированный интерфейс. REST требует использования унифицированного интерфейса, включающего множество операций или методов с известной семантикой, которые изменяют состояние ресурса. Интерфейс зависит от схемы URI. Для http это методы GET, POST, PUT, DELETE, OPTIONS. Методы являются внешними по отношению к ресурсам и включают посылку стандартных сообщений Web-серверу, указывая URI запрашиваемого ресурса, метод, передаваемые данные или метаданные.

Ресурс может иметь множественное представление, которое соответствует стандартизованному формату или типу (MIME-type), и может предоставляться Web-сервером.

Понятие «RESTful» употребляют для описания сервисов, которые построены с учетом архитектуры REST и не нарушают ни одну из ее нотаций. Соблюдение этих ограничений и, следовательно, соответствие архитектурному стилю REST позволят любой распределенной системе иметь требуемые свойства, такие, как производительность, масштабируемость, простота, модифицируемость, видимость, мобильность и надежность.

Архитектура разработанной системы представляет собой модульную систему, основанную на центральном модуле, так называемом «ядре системы». Он представляет собой основу системы и содержит в себе модель данных, набор программных инструментов для работы с ней, а также прямой доступ к интегрированной базе данных (БД) и доступ к внешним БД. При сборке программного комплекса (ПК) ядро целиком помещается в программный компонент, тем самым делая его независимым приложением. Схематическое описание архитектуры ПК изображено на рис. 1.

Одним из основных преимуществ данного подхода является поддержка расширяемости системы. Добавление новых компонент в систему требует лишь использовать универсальное ядро системы для получения всего основного функционала, необходимого при создании новой компоненты.

Интегрированная база включает в себя словари, содержащие унифицированные имена сущностей для описания генных сетей (гены, белки, метаболиты и др.), и информацию о молекулярно-генетических взаимодействиях, реакциях и регуляциях в генных сетях из гетерогенных источников информации, которая представлена в виде нескольких уровней описания. Первый уровень – «сырые данные», представленные в

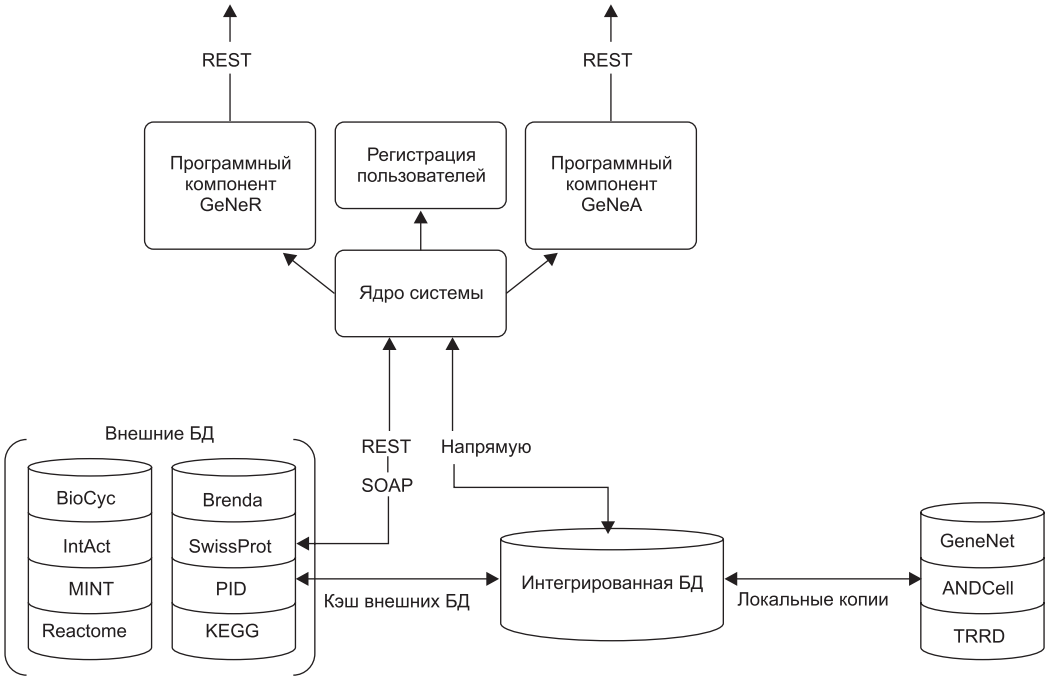

Рис. 1. Архитектура системы реконструкции и анализа генных сетей.

форматах источника данных. Второй уровень – привязанные к унифицированным именам и идентификаторам предварительно обработанные данные, включающие проекцию на схему интегрированной базы данных. БД по генным сетям реализована под управлением СУБД Oracle 11g (Гринвальд, 2009; Кайт, 2011).

Программный компонент GeNeR для реконструкции генных сетей

Программный компонент GeNeR отвечает за реконструкцию генных сетей в ПК. Схематическое описание архитектуры модуля представлено на рис. 2.

Программный компонент GeNeR предоставляет три типа сервисов в стиле REST для внешнего доступа:

– Сервис доступа к внешним источникам данных по молекулярно-генетическим взаимодействиям.

– Сервис интеграции данных и реконструкции генных сетей.

– Сервис доступа к интегрированной БД (есть по умолчанию во всех компонентах).

Данный компонент активно использует менеджер доступа к внешним базам данных по молекулярно-генетическим взаимодействиям, метаболическим путям и генным сетям. Для каждой базы данных пишется свой уникальный драйвер, который обрабатывает унифицированные запросы от системы и запрашивает данные во внешнем источнике. Каждый драйвер несет в себе подробную метаинформацию по базе данных, такую, как: название БД, краткое описание, список поддерживаемых методов доступа и форматов. Указав соответствующий заголовок запроса к сервису на получение метаинформации, можно получить ответ в формате XML.

Программный комплекс позволяет выполнять унифицированные запросы к внешним базам данных по молекулярно-генетическим взаимодействиям, включая базы данных GeneNet (Ananko, 2005), TRRD (Kolchanov, 2008), KEGG (Wrzodek *et al*., 2011), SWISS-PROT, Pathway Interaction Database (Schaefer, 2009), IntAct (Aranda, 2009), REACTOME (Croft, 2011), MINT (Chatr-Aryamontri, 2007), bioCyc (Caspi *et al*., 2010), BRENDA (Scheer, 2011).

Реализованы средства унифицированных запросов к интегрированной базе данных, данным из внешних источников, сервису интеграции данных и реконструкции генных сетей.

Общий формат запроса к сервису доступа к данным из внешних источников: https://host/rws/extdbs/{dbname}/{entity}?{query}.

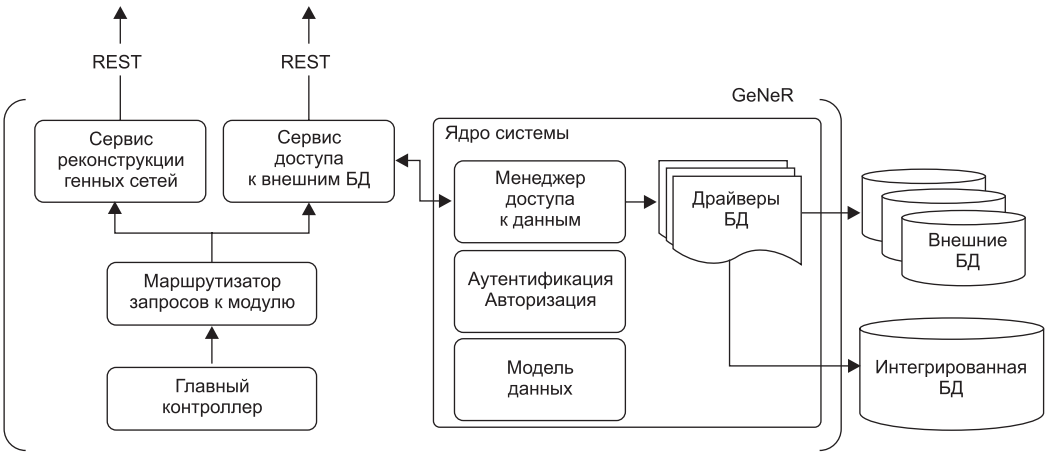

Рис. 2. Архитектура модуля GeNeR.

Приведем некоторые примеры использования сервиса в формате REST запросов:

GET https://host/rws/extdbs/ – выдает список всех внешних источников, представленных в системе.

GET https://host/rws/extdbs/{dbname}/ – выдает информацию по конкретному внешнему источнику данных. В набор информации входят: название БД, краткое описание, список поддерживаемых сущностей, организмов и типов данных.

GET https://host/rws/extdbs/{dbname}/{entity}?{query} – выдает список конкретных записей из указанной БД. Дополнительные параметры для фильтрации записей передаются в {query}.

Общий формат запроса к сервису доступа к интегрированной базе данных: https://host/rws/idb/pathways/{pathway}?{query}.

Подробное описание использования сервиса в формате REST запросов:

GET https://host/rws/idb/ – выдает информацию по интегрированной БД.

GET https://host/rws/idb/pathways/ – выдает список генных сетей, содержащихся в интегрированной БД.

POST https://host/rws/idb/pathways/ – создает новую генную сеть и возвращает ссылку на ее ресурс.

GET https://host/rws/idb/pathways/{pathway}?{query} – выдает конкретную генную сеть из интегрированной БД.

В качестве примеров можно привести следующие запросы к ПК.

Пример. Получить список биохимических реакций у человека, в которых участвует нитрооксид («Nitric oxide»).

Запрос: GET https://host/rws/dbs/kegg/hsa/reaction?metabolite=”Nitric oxide”.

Результаты запроса представлены на рис. 3.

Программный компонент GeNeA для анализа графов генных сетей

Программный компонент GeNeA для анализа графов генных сетей представляет собой набор Web-сервисов, реализованных на языке Java и обеспечивающих запуск прикладных программных модулей анализа структуры генных сетей. Основными данными, которыми оперирует прикладной программный модуль, являются графы генных сетей, представленные как список ребер (соединенных вершин), векторы и матрицы.

Для передачи входных и выходных данных при запуске модуля используются стандартные потоки ввода/вывода. Это обеспечивает возможность быстрого подключения новых программных модулей в систему анализа графов генных сетей. Прикладной программный модуль для анализа графа генных сетей реализован на языке Java и поддерживает унифицированный интерфейс вызова процедур анализа графов, который позволяет подключать внешние библиотеки, реализованные на различных языках программирования, и использовать различные форматы представления графа. В частности,

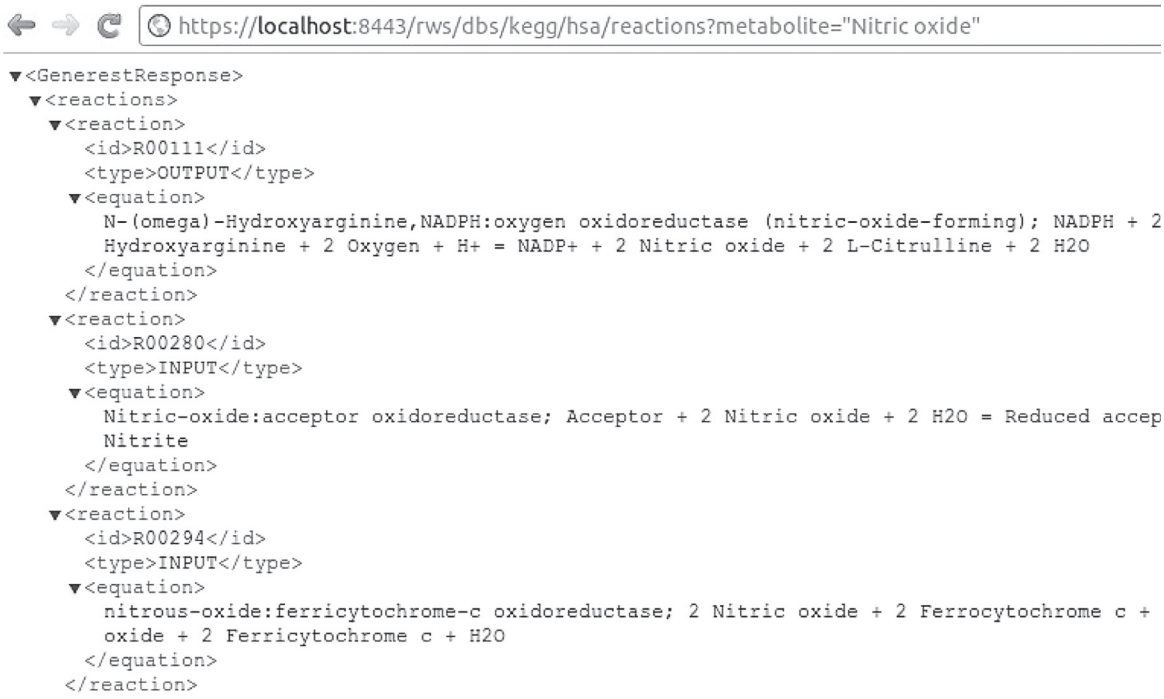

Рис. 3. Результат выполнения запроса списка реакций у человека, в которых участвует нитрооксид («Nitric oxide») (3 реакции).

нами используется библиотека igraph, в которой включено большое число типовых алгоритмов анализа графов (Csárdi, Nepusz, 2006a, b).

Схематическое описание архитектуры модуля GeNeA изображено на рис. 4. Здесь представлены следующие типы сервисов в стиле REST для внешнего доступа:

– Сервис анализа графов генных сетей.

– Сервис визуализации генных сетей.

– Сервис отправки заданий на вычислительный кластер.

– Сервис доступа к интегрированной БД (есть по умолчанию во всех компонентах).

В рамках программного компонента реализован унифицированный доступ к проблемно-ориентированным вычислительным сервисам для анализа графов генных сетей, которые обеспечивают следующие возможности:

– поддержка интроспекции (получение описания сервисов по запросу клиента);

– поддержка асинхронной обработки запросов, требующих длительных вычислений;

– поддержка передачи параметров запроса и результатов в виде файлов;

– использование архитектурного стиля REST и распространенных форматов представления данных XML, JSON.

ЗАКЛЮЧЕНИЕ

Разработан распределенный программный комплекс на основе RESTful-Web-сервисов, который ориентирован на решение задач реконструкции генных сетей на основе интеграции данных из гетерогенных источников информации, включая базы данных о молекулярно-генетических взаимодействиях, метаболических и сигнальных путях, генных сетях. Программный комплекс включает модули для расчета различного рода характеристик графа генных сетей, в частности: распределение степеней вершин, коэффициенты кластеризации, диаметр графа, плотность графа, индекс центральности, индекс Боначича, индекс Фримана, спектр графа, поиск структурных мотивов, поиск циклов в графе генных сетей и др.

Анализ структуры генных сетей, выявление закономерностей структурно-функциональной организации генных сетей, выделение подсистем, редукция описания молекулярно-генетических и построение на этой основе структурной модели генной сети являются важнейшими этапами исследования генных сетей и первым шагом в создании математических моделей динамики генных сетей.

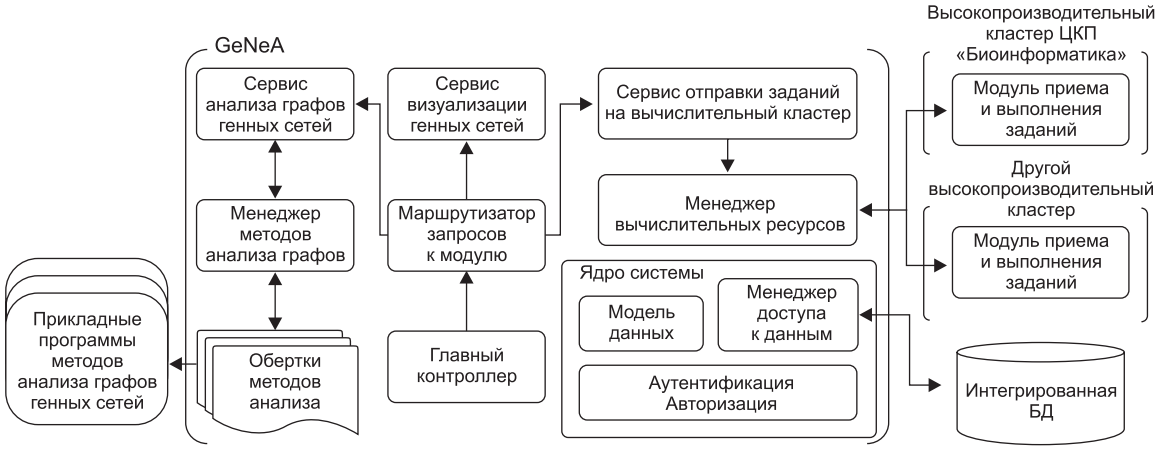

Рис. 4. Архитектура модуля GeNeA.

Работа поддержана Министерством образования и науки РФ (Госконтракт № 07.514.11.4023 по теме «Проектирование и разработка RESTful-Web-сервисов для создания распределенной инфраструктуры, ориентированной на решение задач реконструкции и анализа генных сетей»).

ЛИТЕРАТУРА

Гринвальд Р., Стаковьяк Р., Стерн Д. Oracle 11g. Основы. СПб.: Символ-плюс, 2009. 464 с.

Кайт Т. Oracle для профессионалов: архитектура, программирование и особенности версий 9i, 10g и 11g. «ВИЛЬЯМС», 2011. 848 с.

Колчанов Н.А., Ананько Е.А., Колпаков Ф.А. и др. Генные сети // Молекуляр. биология. 2000. Т. 34. № 4. С. 533–544.

Ananko E.A., Podkolodny N.L., Stepanenko I.L. *et al*. GeneNet in 2005 // Nucl. Acids Res. 2005. V. 33. D425–D427.

Aranda B., Achuthan P., Alam-Faruque Y. *et al.* The IntAct molecular interaction database in 2010 // Nucl. Acids Res. 2009. V. 38. D525–D531.

Caspi R., Altman T., Dale J.M. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases // Nucl. Acids Res. 2010. V. 38. P. 473–479.

Chatr-Aryamontri A., Ceol A., Palazzi L.M. *et al.* MINT: the Molecular INTeraction database // Nucl. Acids Res. 2007. V. 35. P. 572–574.

Croft D., O'Kelly G., Wu G., Haw R. *et al.* Reactome: a database of reactions, pathways and biological processes // Nucl. Acids Res. 2011. V. 39. P. 691–697.

Csárdi G., Nepusz T. The igraph software package for complex network research // Intern. J. Complex Syst. 2006a. V. 1695.

Csárdi G., Nepusz T. igraph Reference Manual // 29-33 Konkoly-Thege Miklys road, Budapest H-1121, Hungary, 509 p. 2006b. http://igraph.sourceforge.net/doc/igraph-docs.pdf

Galperin M.Y., Fernández-Suárez X.M. The 2012 nucleic acids research database issue and the online molecular biology database collection // Nucl. Acids Res. 2011. V. 40. D1-D8.

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A. *et al.* TRRD: Technology for extraction, storage, and use of knowledge about the structural-functional organization of the transcriptional regulatory regions in the eukaryotic genes // Intell. Data Anal. 2008. V. 12. No. 5. P. 443–461.

Kolpakov F.A., Ananko E.A., Kolesov G.B., Kolchanov N.A. GeneNet: a database for gene networks and its automated visualization // Bioinformatics. 1998. V. 14. No. 6. P. 529–537.

Newman M.E.J. Finding community structure in networks using the eigenvectors of matrices // Phys. Rev. 2006. E 74, 036104.

Richardson L., Ruby S. RESTful Web Services. O'Relly Media. Inc., 1005 Gravensein Highway North, Sebastopol, CA, 2007. 420 c.

Schaefer C.F., Anthony K., Krupa S. *et al.* PID: the pathway interaction database // Nucl. Acids Res. 2009. V. 37. P. 674–679.

Scheer M., Grote A., Chang A. *et al.* BRENDA, the enzyme information system in 2011 // Nucl. Acids Res. 2011. V. 39. P. 670–676.

Schreier S. Modeling RESTful applications // Proc. WS-REST '11 Proceedings of the Second Intern. Workshop on RESTful Design. ACM, NY, USA, 2011. P. 15–21.

Subbu Allamaraju RESTful Web Servises Cookbook. O'Relly Media, Inc. 2010. 293 c.

Valverde F., Pastor O. Dealing with REST Services in Model-driven Web Engineering Methods // V Jornadas Científico-Técnicas en Servicios Web y SOA, JSWEB. 2009.

Wrzodek C., Drager A., Zell A. KEGG translator: visualizing and converting the KEGG PATHWAY database to various formats // Bioinformatics. 2011. V. 27. No. 16. P. 2314–2315.

DISTRIBUTED RESTful WEB SERVICES FOR RECONSTRUCTION AND ANALYSIS OF GENE NETWORKS

N.L. Podkolodnyy[1], A.V. Semenychev[1], D.A. Rasskazov[1], V.G. Borowsky[1], E.A. Ananko[1], E.V. Ignatieva[1], N.N. Podkolodnaya[1], O.A. Podkolodnaya[1], N.A. Kolchanov[1, 2, 3]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: pnl@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia;
[3] National Research Centre «Kurchatov Institute», Moscow, Russia

Summary

This paper describes a RESTful Web service-based distributed software system, which focuses on the reconstruction of gene networks by integrating data from heterogeneous data sources, including databases of molecular-genetic interactions, metabolic and signaling pathways, gene networks, etc.

Key words: distributed systems, RESTful-Web services, gene networks, data integration, gene network graph analysis.