Associative Network Discovery (AND) — компьютерная система для автоматической реконструкции сетей ассоциативных знаний о молекулярно-генетических взаимодействиях*

П. С. Деменков
Институт математики им. С.Л. Соболева СО РАН, Новосибирск, Россия
e-mail: demps@math.nsc.ru

Е. Э. Аман, В. А. Иванисенко
Институт цитологии и генетики СО РАН, Новосибирск, Россия

We created an Associative Network Discovery (AND) system enabling us to detect associations involving genes, proteins, ligands and diseases. The broad basis is the information we retrieved by automated analysis of PubMed abstracts and also the data from the databases about the interactions between objects.

Введение

Количество публикаций в области биологии, медицины и биотехнологии растет столь быстро, что имеющуюся информацию принципиально невозможно проанализировать для исследовательских и прикладных целей без автоматической обработки с использованием компьютерных средств.

Для решения задачи извлечения из текстов информации о взаимодействиях молекулярно-генетических объектов в мире было разработано несколько подходов, основанных на различных алгоритмах: от простейших (таких как поиск совместной встречаемости названий биологических объектов в текстах) до комплексных методов, включающих лингвистический и семантический анализ, а также методы машинного обучения.

Одним из самых простых алгоритмов реконструкции сетей ассоциаций молекулярно-генетических объектов является поиск совместной встречаемости имен генов, белков и других биологических объектов в текстах. Этот метод был использован в программе PubGene [1] для реконструкции так называемых “литературных сетей”, основанных на совместной встречаемости названий белков и генов человека в названиях и текстах рефератов статей из базы данных PubMed.

Дж. Купер [2] с коллегами создали простую систему для предсказания белок-белковых взаимодействий с использованием текстового анализатора, основанную на поиске

*Работа выполнена при поддержке Российского фонда фундаментальных исследований, гранты № 05-04-49283-а и № 08-04-91313-IND_а, CRDF Rup2-2629-NO-04 и № RUX0-008-NO-06, Междисциплинарных интеграционных проектов фундаментальных исследований СО РАН № 49 и 115, Государственного контракта с ФАНИ № 02.514.11.4065, а также гранта Поддержки ведущих научных школ № НШ-4413.2006.1.

специальных слов, описывающих взаимодействия, синонимов названий белков и простых правил встречаемости этих слов в реферате статьи. Эта система обладает высокой производительностью, но небольшой точностью (около 60 %).

В системах GeneScene [3] и MedScan [4] реализованы алгоритмы, основанные на глубоком лингвистическом анализе: разборе предложения по частям речи, синтаксическом и семантическом анализе. Эти системы позволяют достичь высокой точности распознавания фактов взаимодействий из текстов — 90 %, но их чувствительность, т. е. доля распознанных взаимодействий среди всех взаимодействий, описанных в тексте, невысока (около 20 %).

Знания о взаимодействиях молекулярно-генетических объектов содержатся не только в текстах научных публикаций. Созданы тысячи фактографических медико-биологических баз данных, содержащих разнообразную информацию о биологических объектах и их взаимодействиях на уровне геномов, клеток и организмов. Объемы этих баз данных чрезвычайно велики, например, база данных NCBI Gene [5] содержит 1 933 023 записей, количество которых постоянно увеличивается. В базах данных KEGG [6], EcoCyc [7], MetaCyc [8], GeneNet [9] и других представлены тысячи фактов о биомедицински и биотехнологически значимых генных сетях, метаболических путях, путях передачи сигналов и др.

В настоящей работе описана компьютерно-информационная система для автоматического извлечения и интеграции ассоциативных знаний из фактографических баз данных и текстовых источников информации. Под ассоциацией между молекулярно-генетическими объектами понимается прямое или опосредованное их взаимодействие, а также следственно-причинные связи между генами, белками и заболеваниями.

Следует отметить, что отечественные разработки в области извлечения и интеграции знаний при одновременной работе с текстовыми и фактографическими базами данных, ориентированные на фармакологию, биотехнологию и биомедицину, отсутствуют, а имеющиеся за рубежом характеризуются низкой эффективностью.

Система AND позволяет пользователю быстро получать и анализировать большие объемы данных в форме графически визуализированных сетей молекулярно-генетических взаимодействий и их ассоциаций с заболеваниями.

1. Описание системы

Система реконструкции сетей ассоциативных знаний Associative Network Discovery [10] состоит из модуля анализа текста, базы знаний об ассоциативных взаимодействиях и программы визуализации.

Был разработан алгоритм выявления фактов ассоциаций между молекулярно-генетическими объектами на основе анализа текстов рефератов статей из базы данных PubMed и данных о функциональных и структурных характеристиках объектов. Схема этого алгоритма показана на рис. 1.

На основе информации из доступных в Интернете баз данных были составлены словари синонимов названий генов (база данных NCBI Gene), белков (SwissProt), метаболитов (ChEBI, KEGG), микроРНК (mirBase), заболеваний (PharmGKB) и организмов (NCBI Taxonomy). Из словарей были удалены все синонимы длиной менее трех символов, а также синонимы, пересекающиеся с английским словарем общей лексики. Суммарный объем неповторяющихся словосочетаний в построенных словарях составил более 2.5 млн записей.

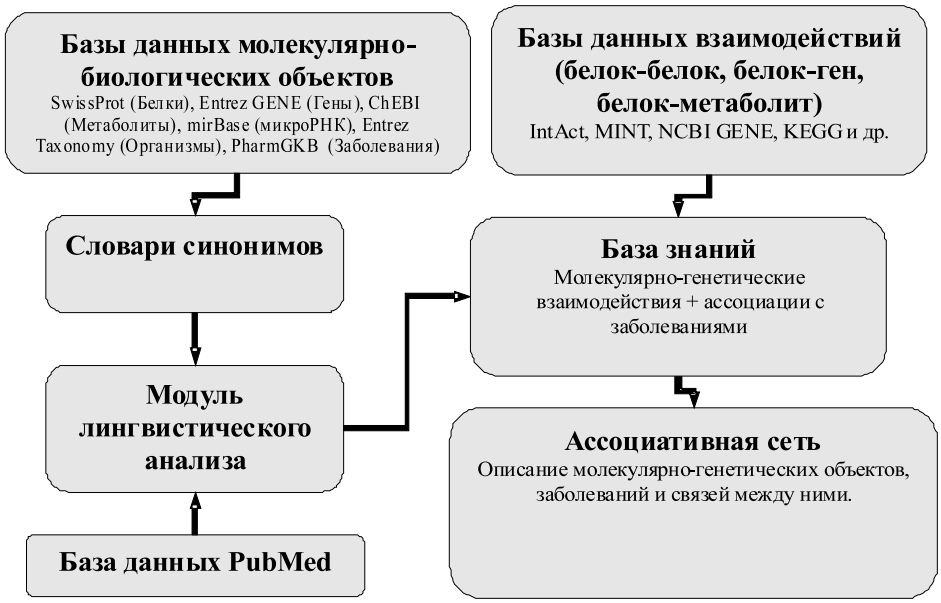

Рис. 1. Схема алгоритма автоматической реконструкции молекулярно-генетических взаимодействий

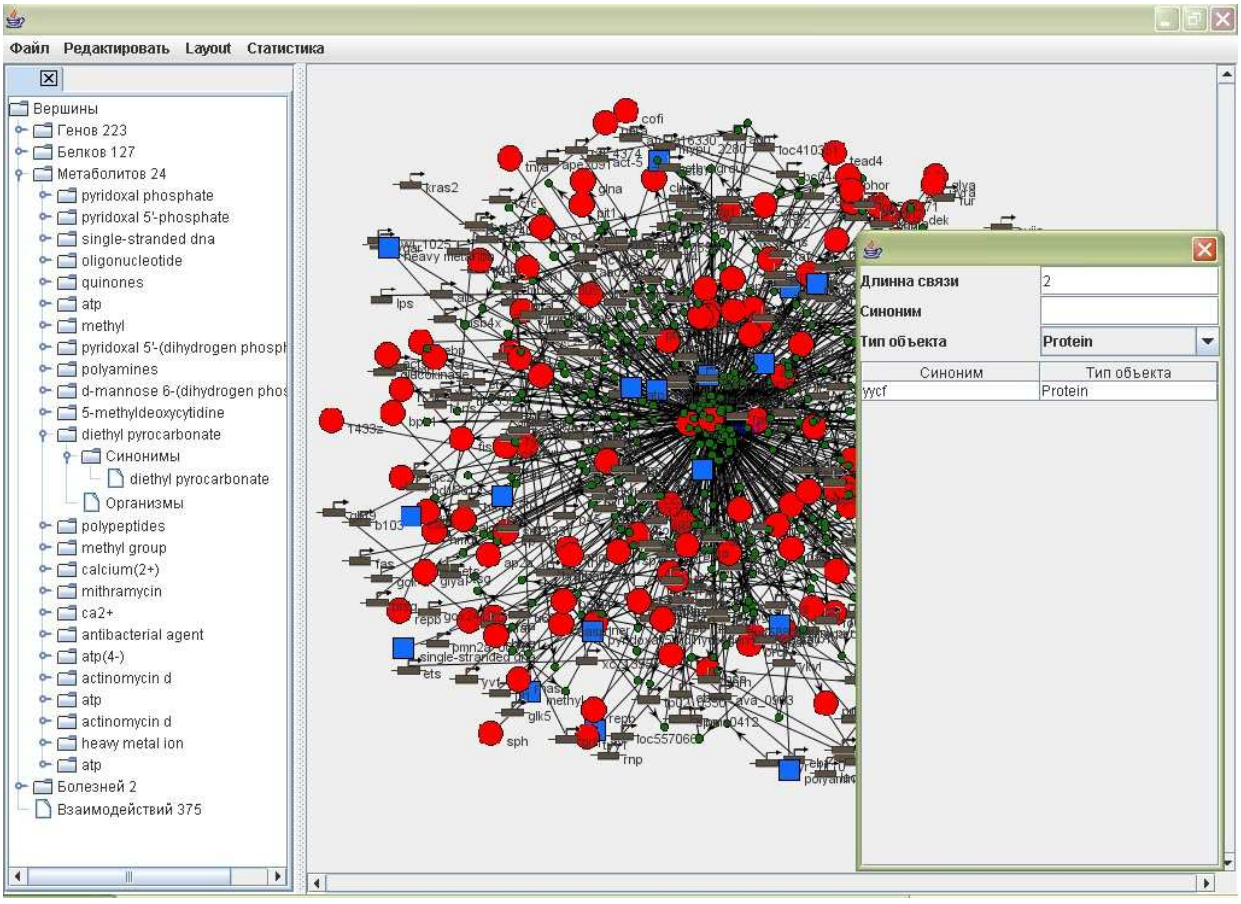

Рис. 2. Интерфейс программы визуализации ассоциативных сетей

Словарь слов-связок включает словоформы, встречающиеся в предложениях и описывающие взаимодействия объектов. Он был составлен вручную с помощью экспертного анализа рефератов из базы данных PubMed. На основании этого словаря были сформулированы правила (шаблоны) для извлечения из текста информации о взаимодействиях объектов.

С использованием словарей синонимов названий молекулярно-генетических объектов и правил, основанных на словаре слов-связок, производился анализ текстов рефератов из базы данных PubMed. Полученная информация об ассоциациях между белками, генами, метаболитами, микроРНК и заболеваниями заносилась в специализированную базу данных. Объектами в базе являются белки, гены, метаболиты, микроРНК, заболевания и ассоциации между ними. База данных содержит 18 таблиц и реализована с использованием СУБД MySQL 5.0.

Для удобного пользователю графического представления данных о взаимодействиях на языке Java была разработана программа визуализации ассоциативных сетей. Программа позволяет реконструировать молекулярно-генетические сети по запросу к базе данных, извлекать информацию о свойствах объектов сети, редактировать сети с помощью применения фильтров или добавления/удаления объектов по желанию пользователя, получать доступ к базам данных молекулярно-генетических объектов или текстов PubMed (рис. 2).

2. Результаты

Было проанализировано 8 114 444 рефератов из базы данных PubMed за период с 1990 по 2006 год, на основе этих текстов выявлено 2 497 567 ассоциаций.

Наряду с информацией о взаимодействиях, выявленных из текстов, в базу данных сетей ассоциативных знаний были интегрированы данные о взаимодействиях, извлеченных из фактографических баз данных, таких как KEGG, IntAct, TRRD [11], MirBase и др.

На сегодняшний день объем базы данных составляет более 60 млн записей.

Для оценки точности распознавания фактов взаимодействий из текстов нами было проведено сравнение генной сети, реконструированной экспертом (активация NF-kB), с ассоциативной сетью NF-kB, построенной на основе информации, извлеченной из текстов рефератов. В этих сетях было выявлено 89 % общих объектов, 59 % общих связей.

Система AND может быть полезна при решении широкого спектра задач системной биологии, биомедицины и биотехнологии, таких как расширение и дополнение генных сетей, реконструированных экспертом, выявление ассоциации генных сетей с заболеваниями, поиск возможных молекулярных механизмов ассоциаций между патологиями, выявление генов-кандидатов для генотипирования, мутации в которых приводят к возникновению заболеваний.

Список литературы

[1] Jenssen T.K., Laegreid A., Komorowski J., Hovig E. A literature network of human genes for high-throughput analysis of gene expression // Nat. Genet. 2001. Vol. 28, N 1. P. 21–28.

[2] Cooper J.W., Kershenbaum A. Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information // BMC Bioinformatics. 2005. N 6. P. 143.

[3] Leroy G., Chen H. Filling preposition-based templates to capture information from medical abstracts // Pac. Symp. Biocomput. 2002. P. 350–361.

[4] Novichkova S., Egorov S., Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts // Bioinformatics. 2003. Vol. 19, N 13. P. 1699–1706.

[5] Maglott D., Ostell J., Pruitt K.D., Tatusova T. Entrez Gene: gene-centered information at NCBI // Nucleic Acids Res. 2005. Vol. 33 (Database Issue). P. D54—D58.

[6] Kanehisa M., Goto S., Hattori M. et al. From genomics to chemical genomics: new developments in KEGG // Nucleic Acids Res. 2006. N 34. P. D354–D357.

[7] Keseler I.M., Collado-Vides J., Gama-Castro S. et al. EcoCyc: A comprehensive database resource for Escherichia coli // Nucleic Acids Res. 2005. N 33. P. D334–D337.

[8] Caspi R., Foerster H., Fulcher C.A. et al. MetaCyc: A multiorganism database of metabolic pathways and enzymes // Nucleic Acids Res. 2006. N 34. P. D511–D516.

[9] Ananko E.A., Podkolodny N.L., Stepanenko I.L. et al. GeneNet in 2005 // Nucleic Acids Res. 2005. N 33 (Database issue). P. D425–D427.

[10] Aman E.E., Demenkov P.S., Pintus S.S. et al. A computer system for the automated reconstruction of molecular-genetic interaction networks // Proc. BGRS-2006. Novosibirsk, 2006. Vol. 3. P. 15–18.

[11] Kolchanov N.A., Ignatieva E.V., Ananko E.A. et al. Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucleic Acids Res. 2002. Vol. 30, N 1. P. 312–317.

Поступила в редакцию 7 марта 2008 г.