

Компьютерная система "Gene Discovery" для поиска закономерностей и представления знаний по регуляции генной экспрессии в интегрированной электронной библиотеке GeneExpress

¹ Витяев Е.Е., Колчанов Н.А., *Орлов Ю.Л., Подколотный Н.Л., Поздняков М.А.

¹Институт математики им. Соболева СО РАН, Новосибирск, Россия
Институт цитологии и генетики СО РАН, Новосибирск, Россия

Эл.почта: vityaev@math.nsc.ru, kol@bionet.nsc.ru, orlov@bionet.nsc.ru, pnl@bionet.nsc.ru,
mike@bionet.nsc.ru

* Автор для переписки

Ключевые слова: Открытие знаний, представление знаний, машинное обучение, анализ данных, электронные библиотеки, биоинформатика, экспрессия генов.

Резюме:

Электронная библиотека GeneExpress (<http://wwwmgs.bionet.nsc.ru/mgs/gnw/>) (Колчанов Н.А. и др., 2000), разрабатываемая в ИЦиГ СО РАН, предназначена для сбора экспериментальных данных, навигации, поиска информации, анализа данных и представления знаний в области регуляции генной экспрессии. Такого рода данные и знания имеют важнейшее значение при решении широкого круга задач молекулярной биологии, молекулярной генетики, биотехнологии и медицины. Растущие объемы данных по регуляции генной экспрессии дают возможность анализа структуры регуляторных районов (<http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/>) (Kolchanov N.A. et al., 2002). Предсказание регуляторных, и, прежде всего, промоторных районов требует интеграции разнородной информации, закодированной в последовательностях ДНК как на уровне ДНК-белкового связывания, так и взаимодействующих транскрипционных факторов. Необходимы новые программные средства для иерархического анализа регуляторных последовательностей ДНК, включающего как отдельные сайты связывания транскрипционных факторов, так и группы таких сайтов и промоторные районы в целом. Разработана компьютерная система "Gene Discovery" для поиска закономерностей контекстной организации и предсказания на этой основе сайтов связывания транскрипционных факторов (ССТФ) и регуляторных районов. Система позволяет: (1) находить локальные закономерности контекстной организации сайтов ССТФ; (2) определять закономерности распределения потенциальных ССТФ; (3) выявлять знания по иерархической организации промоторных районов и проводить на этой основе распознавание районов.

Исследование регуляции генной экспрессии и представление знаний

Последняя версия системы GeneExpress 2.1 содержит новые информационные программные модули (<http://www.mgs.bionet.nsc.ru/mgs/gnw/>) для продукции знаний, позволяющие анализировать информацию с целью выявления особенностей структурно-функциональной организации генетических макромолекул, значимых для их функции, уровня специфической активности, а также для их распознавания и классификации (Колпаков Ф.А. и др., 2000).

Извлечение знаний является многоступенчатым интерактивным процессом, включающим создание выборки, предобработку данных, выделение априорных знаний предметной области, визуализация (представление данных), выбор алгоритма. Процесс поиска закономерностей и представления знаний представлен в настоящей работе на примере анализа промоторных районов генов. Исследование структуры промоторов представляет большой интерес для понимания механизмов транскрипции генов эукариот.

Компьютерная система "Gene Discovery". Продукция знаний по структурно-функциональной организации регуляторных геномных последовательностей.

Метод машинного обучения и созданная на его основе система "Discovery" находит статистически значимые правила в логике первого порядка для функциональной аннотации регуляторных районов. Система "Discovery" успешно применялась к решению многих проблем в психологии, физике, медицине, финансах и других науках (Kovalerchuk B. & Vityaev E., 2000, 2001) (см. также раздел "comparison" www-сайта "Scientific Discovery": <http://www.math.nsc.ru/LBRT/logic/vityaev>). Также как и любая техника, основанная на логических правилах (Mitchell T., 1997), данная техника позволяет получить предсказывающие правила на естественном языке, которые интерпретируются с биологической точки зрения и обеспечивают предсказание промоторов (функциональную аннотацию). Эксперт-биолог может оценить корректность распознавания и значимость правил самих по себе. Научной проблемой в применении предсказывающих систем, основанных на данных, является обобщение. Система "Discovery" обобщает данные через обнаружение логических вероятностных правил-законов.

Принципиальная схема системы "Gene Discovery" для анализа нуклеотидных последовательностей представлена на Рисунке 1.

"Gene Discovery" (Vityaev E.E. et al., 2002) состоит из трех основных модулей: (1) модуль для интерактивного представления контекстных сигналов в стандартной таблице данных; (2) модуль "Discovery" для поиска закономерностей; (3) модуль для распознавания класса последовательности, используя найденные закономерности. Программа написана на языке C++ и предназначена для интерактивного использования (Рисунок 1).

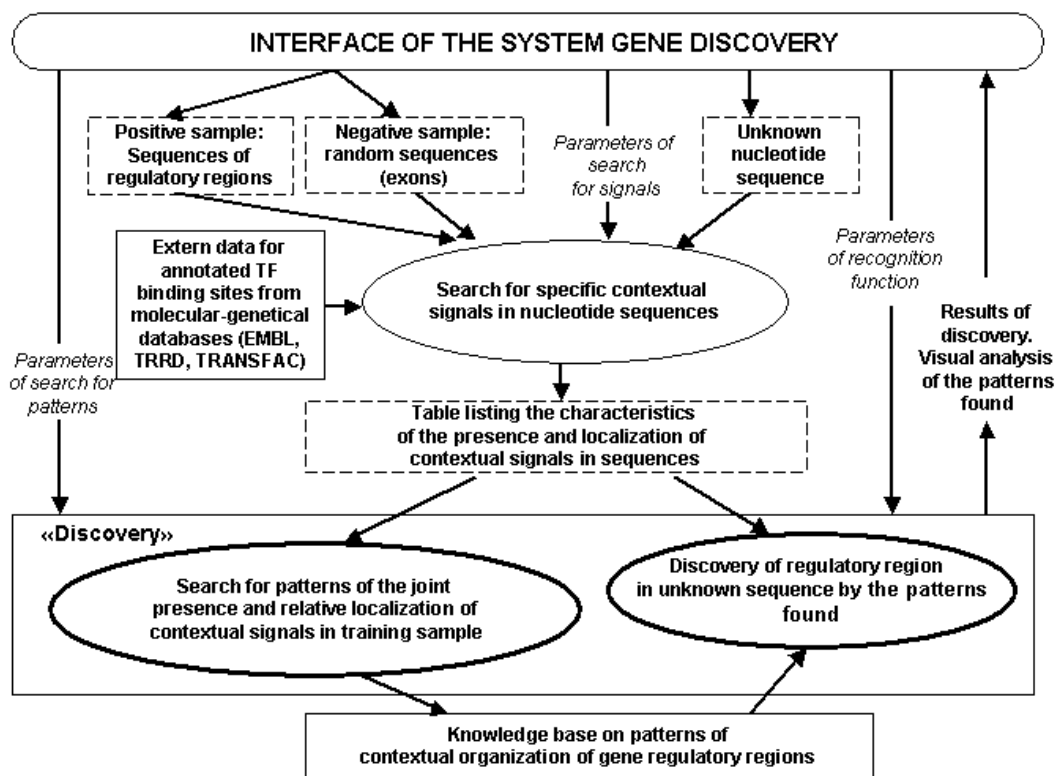


Рисунок 1. Блок-схема системы.

Сигнал может быть:

- контекстным (короткое олигонуклеотидное слово, функциональный сайт и т.д.),
- конформационным (участок ДНК, характеризующийся особенностями конформационных или физико-химических свойств, например, легкоплавкие участки ДНК, сильно изогнутая ДНК и т.д.),
- структурным (например, Z-ДНК или шпилька вторичной структуры РНК и др.).

Все эти сигналы могут быть установлены с использованием знаний о свойствах ДНК и консенсусных схемах, на основе экспериментальной информации из специализированных баз данных.

Задача иерархического поиска специфичных паттернов для сайтов связывания тесно связана с задачей анализа структуры регуляторных районов. В этом случае представление закономерностей носит двухуровневый характер - сначала в нуклеотидной последовательности распознаются потенциальные сайты связывания, а затем группы таких сайтов, формирующие регуляторные комплексы либо композиционные элементы. Поиск специфичных паттернов и построение на их основе иерархии уровней генных кодов включает анализ и распознавание основных элементов структуры гена - кодирующих частей, сайтов сплайсинга, промотора, 5'UTR, сайта полиаденилирования.

Применение системы для анализа регуляторных районов генов функциональных систем организма

Были проанализированы последовательности промоторов генов нескольких функциональных систем, в частности эндокринной системы, и соответствующие им по частотам олигонуклеотидов случайные последовательности из базы данных TRRD (<http://wwwmgs.bionet.nsc.ru/>). Для выделения олигонуклеотидных сигналов, специфичных к данной группе промоторов, использовалась программа ARGO (<http://wwwmgs.bionet.nsc.ru/mgs/programs/argo/>) (Babenko V.N. et al., 1999), (Вишневский О.В., Витяев Е.Е., 2001)

Отобранные контекстные сигналы (вырожденные олигонуклеотиды) были локализованы в исследуемых последовательностях ДНК и представлены в виде таблицы данных "объект-признак". В этой таблице объектами являются последовательности ДНК, признаками – присутствие контекстных сигналов и их локализация относительно экспериментально определенного старта транскрипции.

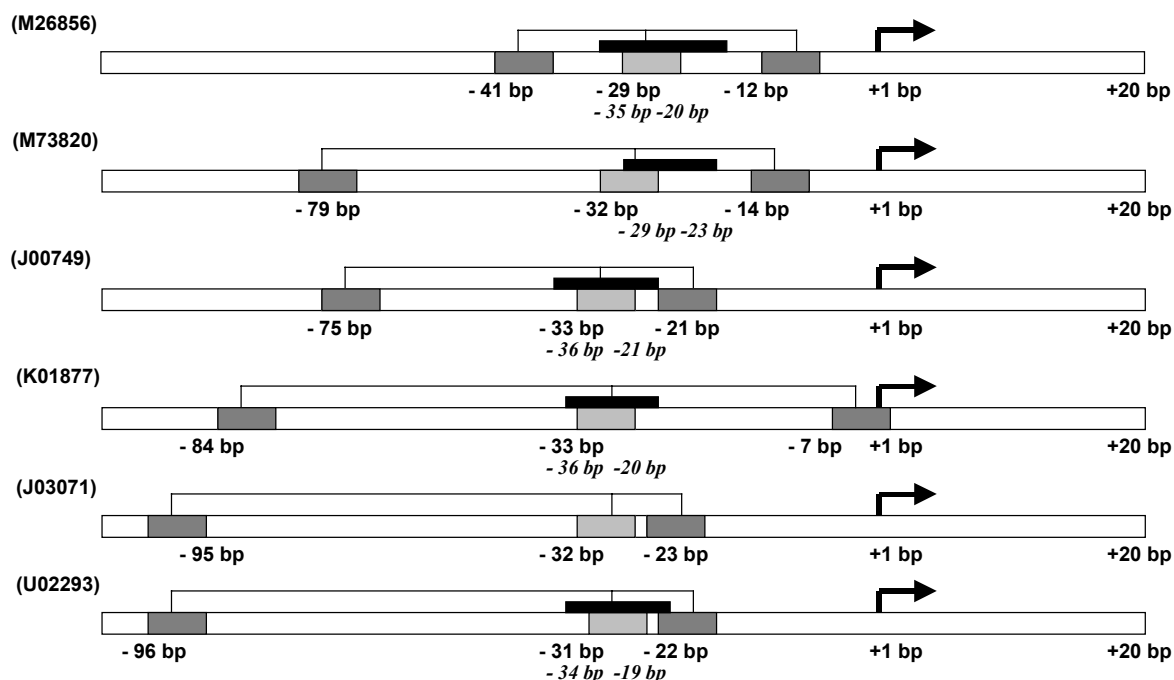


Рисунок 2. Схема расположения комплексного сигнала CWGNRGCN<NGSYMTAM<MAGKSHCN в промоторах генов эндокринной системы. Последовательности промоторов сфазированы относительно старта транскрипции (позиция +1 п.о.), выделенного стрелкой. Идентификатор банка данных EMBL исследуемой последовательности указан слева в скобках. Входящие в комплексный сигнал олигонуклеотидные мотивы длиной 8 п.о. отмечены черными прямоугольниками, указана позиция первого нуклеотида относительно старта транскрипции. Положение ТАТА-бокса, проиндексированное в базе данных TRRD, отмечено заштрихованными прямоугольниками.

Найденные закономерности (знания) имеют смысл комплексных сигналов, регулирующих транскрипцию посредством связывания с ДНК белков, специфичных к данному типу районов. Пример расположения комплексного сигнала для генов эндокринной системы представлен на Рисунке 2.

Таким образом, компьютерная система "Gene Discovery" позволяет выявлять как индивидуальные значимые мотивы (вырожденные квазиинвариантные олигонуклеотиды), так и комплексные сигналы. Проведенный анализ показал, что промоторы генов эндокринной системы и эритроид-специфичные промоторы характеризуются высокой насыщенностью такими сигналами (Витяев Е.Е. и др., 2001).

Информация, распределенная по научной литературе и сосредоточенная в молекулярно-биологических базах данных, содержит тысячи экспериментальных результатов о последовательностях ДНК, вовлеченных в регуляцию транскрипции. В настоящее время в мире существует около 300 молекулярно-биологических баз данных, доступных через Интернет. (Baxeavanis A.D., 2002), что требует интеграции данных и разработки новых программных средств для анализа данных и представления знаний в биоинформатике.

Работа была частично поддержана РФФИ (но. 01-07-90376, 00-07-90337, 02-07-90355, 00-04-49229), Министерством науки (43.073.1.1.1501), СО РАН (65), INTAS (YSF 00-178).

Литература

Babenko V.N. *et al.*, Investigating extended regulatory regions of genomic DNA sequences, *Bioinformatics* **15** (1999) 644-653.

Baxeavanis A.D., The Molecular Biology Database Collection: 2002 update. *Nucleic Acids Res.*, **30** (2002) 1-12.

Kolchanov N.A. *et al.*, Transcription Regulatory Regions Databases (TRRD): its status in 2002, *Nucleic Acids Res.* **30** (2002) 312-317.

Kovalerchuk B. and Vityaev E., Data Mining in finance: Advances in Relational and Hybrid Methods. (Kluwer international series in engineering and computer science; SECS 547), Kluwer Academic Publishers, 2000, 308 p.

Kovalerchuk B. *et al.*, Consistent Knowledge Discovery in Medical Diagnosis, *IEEE Engineering in Medicine and Biology Magazine* (Special issue: "Medical Data Mining", July/August) 2000, pp. 26-37.

Mitchell T., Machine Learning. New York: McGraw Hill, 1997.

Vityaev E.E. *et al.*, Computer system "Gene Discovery" for promoter structure analysis, *In Silico Biol.* **2** (2002) 0024
<<http://www.bioinfo.de/isb/2002/02/0024/>>

Витяев Е.Е., Орлов Ю.Л., Вишневский О.В., Беленок А.С., Колчанов Н.А. Компьютерная система "GENE DISCOVERY" для поиска закономерностей организации регуляторных последовательностей эукариот. *Молекулярная биология*, 2001, 35(6), 952-960

Вишневский О.В., Витяев Е.Е. Анализ и распознавание промоторов эритроид – специфичных генов на основе наборов вырожденных олигонуклеотидных мотивов. *Молекулярная биология*, 2001, 35(6), 979-986.

Колпаков Ф.А., Подколотный Н.Л., Лаврюшев С.В., Григорович Д.А., Пономаренко М.П., Колчанов Н.А. Методы интеграции неоднородных информационных ресурсов по регуляции генной экспрессии в электронной библиотеке GeneExpress. *Программирование*, 2000, **3**, 72-80.

Колчанов Н.А. и др. Анализ данных и продукция знаний в система GeneExpress - электронной библиотеке по структуре и функции ДНК, РНК и белков. *Вторая Всероссийская научная конференция "Электронные библиотеки: перспективные методы и технологии, электронные коллекции"* 26-28 сентября 2000 г., Протвино, 154-161.