КОМПЬЮТЕРНАЯ СИСТЕМА "GENE DISCOVERY" ДЛЯ ПОИСКА ЗАКОНОМЕРНОСТЕЙ И ПРЕДСТАВЛЕНИЯ ЗНАНИЙ ПО РЕГУЛЯЦИИ ГЕННОЙ ЭКСПРЕССИИ В ИНТЕГРИРОВАННОЙ ЭЛЕКТРОННОЙ БИБЛИОТЕКЕ GENEEXPRESS

[1]Витяев Е.Е., Орлов Ю.Л., Поздняков М.А., Левицкий В.Г., Вишневский О.В., Подколодный Н.Л., Колчанов Н.А.

[1]Институт математики им. Соболева СО РАН,
пр-т Лаврентьева, 10, Новосибирск 630090, Россия
Эл. почта: vityaev@math.nsc.ru

Институт цитологии и генетики СО РАН,
ул. Коптюга, 4, Новосибирск 630090, Россия
Эл. почта: orlov@bionet.nsc.ru, mike@bionet.nsc.ru, oleg@bionet.nsc.ru, levitsky@bionet.nsc.ru, pnl@bionet.nsc.ru, kol@bionet.nsc.ru

Сбор экспериментальных данных, навигация, поиск информации, анализ данных и представление знаний в области регуляции генной экспрессии имеет важнейшее значение при решении широкого круга задач молекулярной биологии, молекулярной генетики, биотехнологии и медицины. 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. В рамках электронной библиотеки GeneExpress (http://wwwmgs.bionet.nsc.ru/mgs/gnw/) 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.

COMPUTER SYSTEM "GENE DISCOVERY" FOR REGULARITIES SEARCH AND KNOWLEDGE DISCOVERY ON GENE EXPRESSION REGULATION USING INTEGRATED DIGITAL LIBRARY GENEEXPRESS

[1]Vityaev E.E., Orlov Yu.L., Pozdnyakov M.A., Levitsky V.G., Vishnevsky O.V., Podkolodny N.L., Kolchanov N.A.

[1]Sobolev Institute of Mathematics SB RAS,
Acad. Koptyug prospect, 4, Novosibirsk, 630090, Russia.
E-mail: vityaev@math.nsc.ru

Institute of Cytology and Genetics SB RAS, Lavrentieva ave., 10, Novosibirsk, 630090, Russia.
E-mail: orlov@bionet.nsc.ru, mike@bionet.nsc.ru, oleg@bionet.nsc.ru, levitsky@bionet.nsc.ru, pnl@bionet.nsc.ru, kol@bionet.nsc.ru

Digital library **GeneExpress 2.1** is designed for accumulation of experimental data, data navigation, data analysis, and analysis of dependencies in the field of gene expression regulation. It integrates the databases and programs for processing the data about structure and function of DNA, RNA, and proteins and serves as background for data mining in the field of gene expression regulation. This paper presents implementation of Data Mining techniques for searching regularities in tables of context features of DNA sequences involved in transcription regulation. The goal is to discover regularities that interrelate nucleotide sequences with the functional class of these sequences. The search for regularities is implemented in a software system "Gene Discovery" which is based on first-order probabilistic logic. The "Gene Discovery" system provides a general scenario of functional annotation of an arbitrary nucleotide sequence. The method relies to original Knowledge Discovery approach that can be used for wide variety of complex bioinformatics problems.

Исследование регуляции генной экспрессии и представление знаний

Исследование механизмов регуляции транскрипции генов является одной из актуальных проблем современной молекулярной биологии и биоинформатики в наступившую после 2000 года "пост-геномную эпоху", в связи с полным секвенированием геномов в рамках масштабных международных программ. Такое исследование предполагает объединение имеющихся компьютерных ресурсов и баз данных по молекулярной биологии. 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 [1, 2]. Такое положение дает возможность широкомасштабного применения теории анализа данных и открытия знаний в биоинформатике [3]. Анализ данных в других научных областях связан с экстенсивной обработкой больших баз данных и поиском закономерностей и установлением новых знаний. Разработанные методики интеграции компьютерных ресурсов в области биоинформатики имею самостоятельную научную ценность, и представляют собой пример междисциплинарных исследований, объединяющих усилия биологов, математиков и специалистов в области информационных технологий.

Разрабатываемая в Институте цитологии и генетики СО РАН Интернет-навигационная система GeneExpress 2.1 (http://wwwmgs.bionet.nsc.ru/mgs/gnw/) содержит информационные программные модули 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 [4]. Система GeneExpress-2.1 организована с учетом естественной иерархии уровней молекулярно-генетических систем организмов, и содержит четыре крупных модуля: (1) “Уровень ДНК”; (2) “Уровень РНК”; (3) “Уровень белков”; (4) “Уровень генных сетей”. Разбиение всех информационных и программных ресурсов, интегрируемых в рамках системы GeneExpress-2.1, на 4 принципиальных модуля обеспечивает возможность естественного расширения системы при присоединении к ней новых ресурсов без существенного изменения ее архитектуры и схемы интеграции.

Обработка данных на уровне нуклеотидных последовательностей - распознавание, поиск и разметка функциональных сайтов и регуляторных районов генов представляет собой область применения методов анализа данных и представления знаний. Процесс поиска закономерностей и представления знаний представлен в настоящей работе на примере анализа промоторных районов генов.

Регуляторные районы составляют лишь малую часть из 95% последовательностей генома позвоночных, не кодирующих белки, но они определяют уровень, порядок и хронологию экспрессии генов. Несмотря на важность этих некодирующих последовательностей, наши возможности в предсказании и определении функции этих участков ДНК чрезвычайно ограничены [2, 5].

Рассмотрим задачу исследования структуры промоторов генов эукариот (высших организмов, включая человека). Обязательным элементом, абсолютно необходимым для инициации транскрипции, является коровый (базальный) промотор, под которым понимают минимальную последовательность ДНК, необходимую для правильной инициации транскрипции гена in vitro (Рисунок 1). В коровый промотор входит старт транскрипции и область приблизительно от –60 до +40 п.о. по отношению к нему. Каждый регуляторный район содержит в своем составе сайты связывания определенных транскрипционных факторов (ССТФ).

Один ген может иметь множество промоторов, определяющих формирование различных белковых продуктов или обладающих различным уровнем специфической функциональной активности. Кроме того, для промоторов эукариот характерно отсутствие точной локализации контекстных сигналов, значимых для их функционирования и слабость этих сигналов.

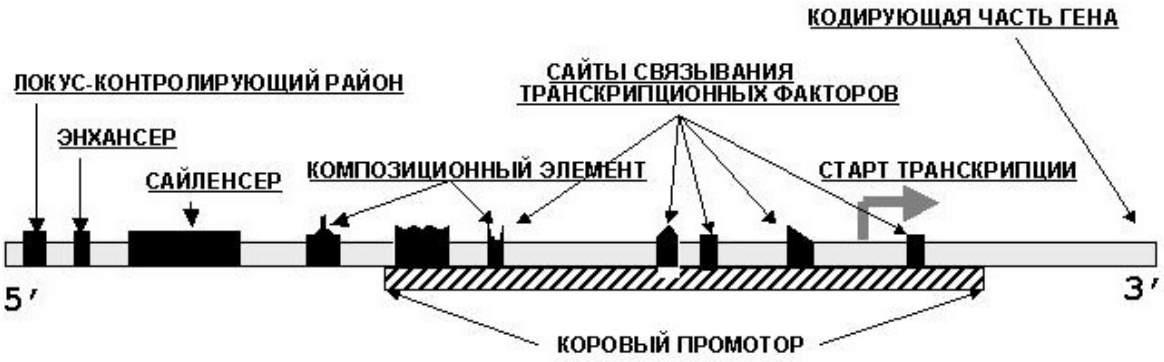

Рисунок 1. Модель структурно-функциональной организации регуляторного района. Схематически представлен участок ДНК, содержащий старт транскрипции (отсемен серой стрелкой)

Разнообразие строения промоторных районов генов создаёт наибольшие трудности для разработки программ распознавания промоторов. Несмотря на то, что создано большое количество методов распознавания промоторов РНК полимеразы II в геномах эукариот проблема повышения точности распознавания в целом остается нерешенной [6].

Статистическая задача распознавания функциональных сайтов в нуклеотидных последовательностях

Извлечение знаний является многоступенчатым интерактивным процессом, включающим создание выборки, предобработку данных, выделение априорных знаний. Разработан набор компьютерных средств для такого анализа. Программы распознавания функциональных сайтов и регуляторных районов в нуклеотидных последовательностях объединены в едином интерфейсе на Интернет-сервере ИЦиГ СО РАН (http://wwwmgs.bionet.nsc.ru/mgs/).

Из вновь добавленных в электронную библиотеку программ распознавания (http://wwwmgs.bionet.nsc.ru/mgs/gnw/regscan/) нужно отметить следующие: (1) программу MMSite, реализующую сложный сценарий автоматического распознаваиня сайтов связывания транскрипционных факторов (на основе использования множества различных методов, распознающих один и тот же сайт); (2) программу BinomSite распознавания сайтов связывания транскрипционных факторов на основе оценки гомологии с последовательностями из базы данных TRRD; (3) программу Recoq для распознавания промоторов; (4) программу NASCA для расчета корреляций свойств в фазированных выборках сайтов связывания; (4) программу Complexity для построения контекстных деревьев.

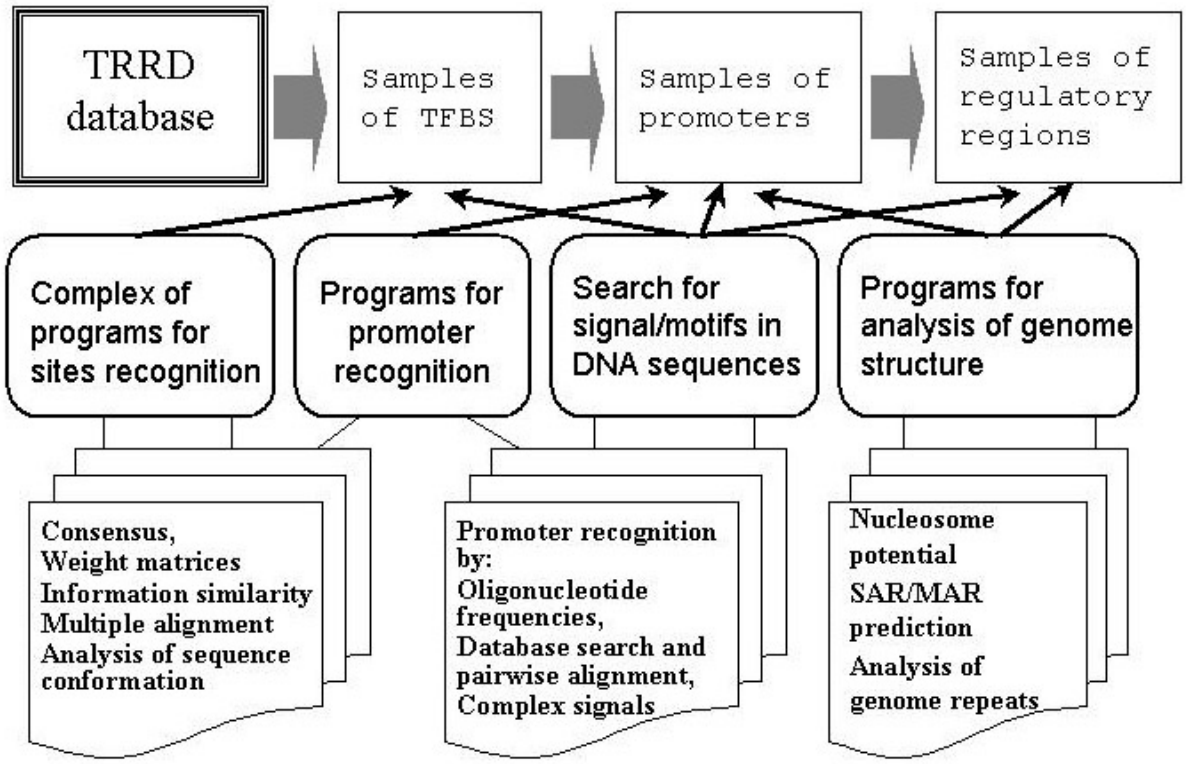

Рисунок 2. Схема анализа данных по регуляторным последовательностям ДНК, аннотированным в базе данных TRRD (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/) и наборы программных средств для такого анализа

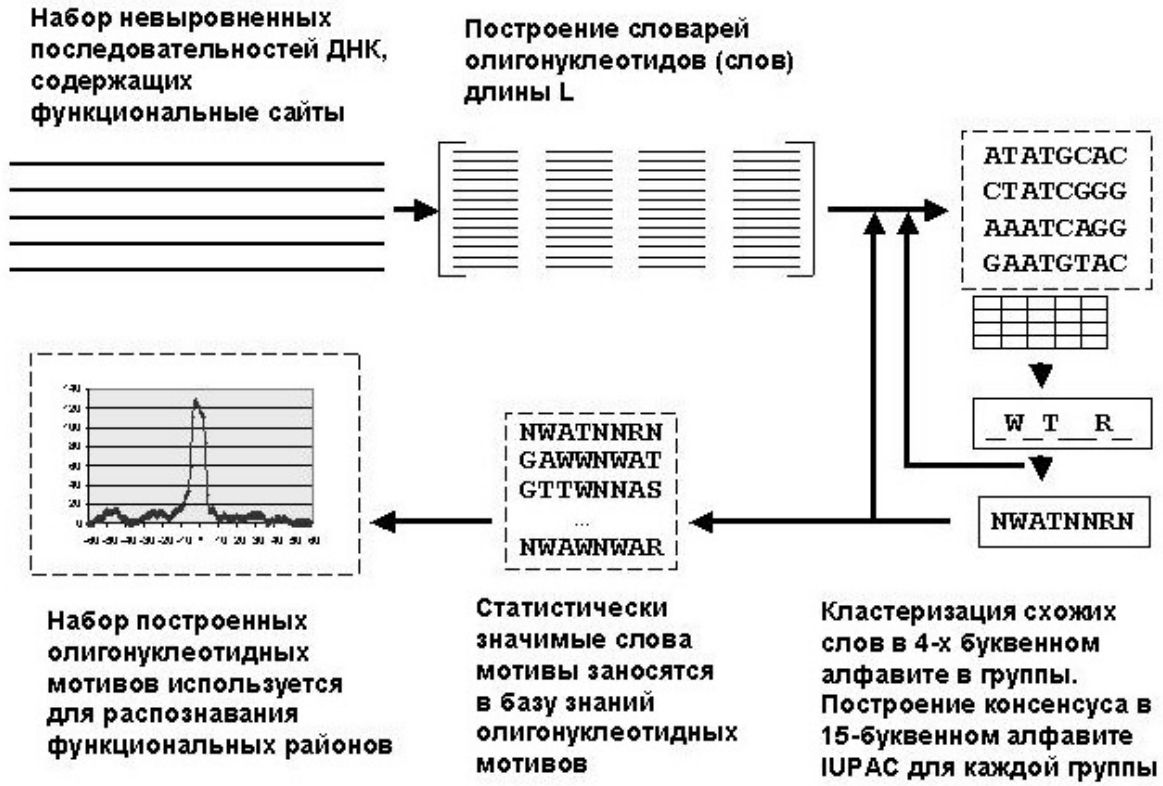

Рисунок 3. Блок схема алгоритма ARGO для распознавания функциональных сайтов в последовательностях ДНК на основе статистически значимых олигонуклеотидных мотивов

Продукция знаний по структурно-функциональной организации регуляторных геномных последовательностей

Представляет особый интерес методология объединения результатов компьютерного распознавания и иерархический анализ таких результатов для максимально полной функциональной аннотации генома. Задача иерархического поиска специфичных паттернов для сайтов связывания в целом 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. Сайты могут быть предсказаны независимо различными методами, с достаточно большим уровнем ошибок 1-го и 2-го рода, что не дает возможности статистически достоверно определить тип регуляторного района (промотора) и механизм экспрессии гена. Для повышения точности предсказания необходимо объединения результатов распознавания сайтов с помощью статистически значимых паттернов, соответствующих совместно функционирующим, биологически значимым группам сайтов.

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, и использование знаний о структуре ДНК более высокого порядка, связанной с позиционированием нуклеосом и упаковкой ДНК в структуру хроматина.

Компьютерная система "Gene Discovery"

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 ранее к решению многих проблем в психологии, физике, медицине, финансах и других науках [7, 8] (см. также www-сайт "Scientific Discovery": http://www.math.nsc.ru/LBRT/logic/vityaev/, раздел "comparison"). Также как и любая техника, основанная на логических правилах [9], 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 4.
"Gene Discovery" [10,11] 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.

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.).

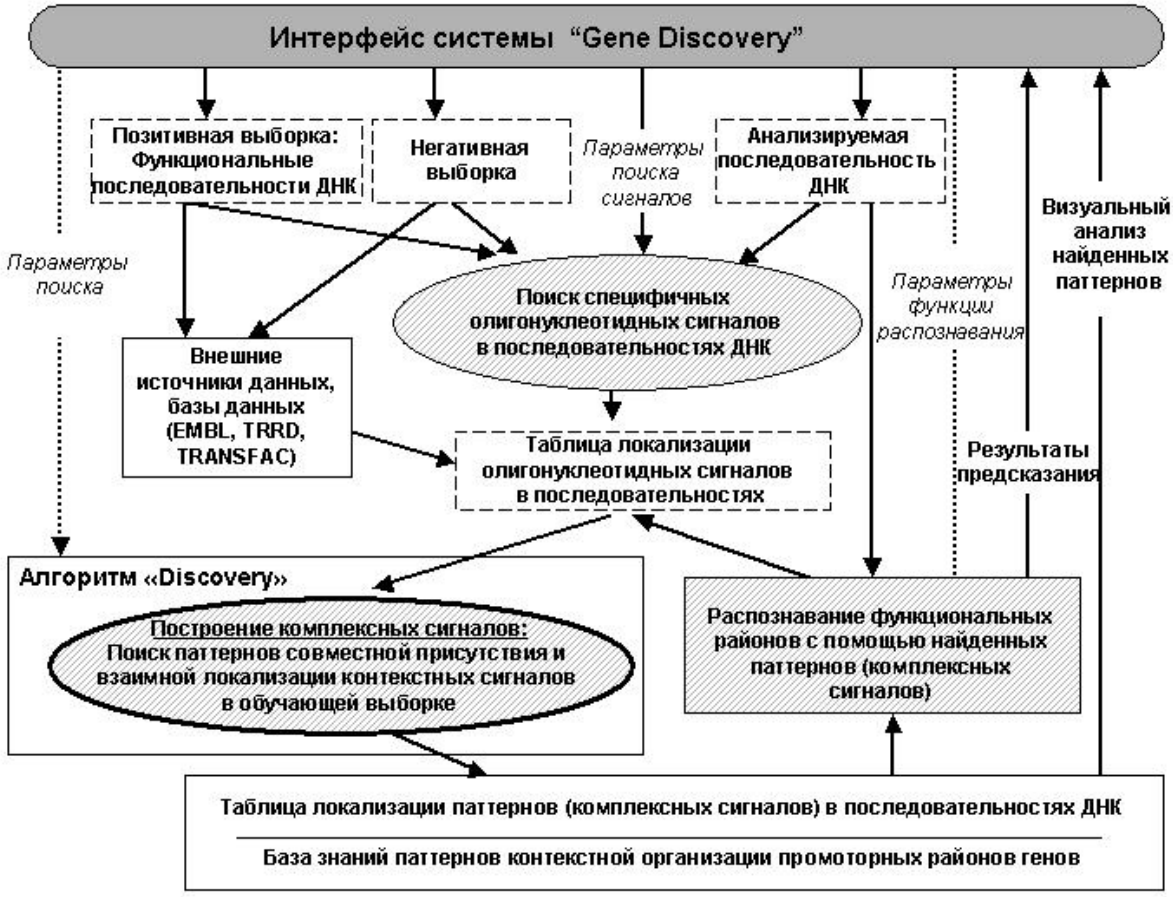

Рисунок 4. Блок-схема системы

Такие сигналы могут быть установлены с использованием знаний о свойствах ДНК, на основе экспериментальной информации из баз данных и с помощью программ распознавания функциональных сайтов (ССТФ).

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/) [12, 13].

Пример расположения комплексного сигнала для генов эндокринной системы представлен на Рисунке 5. Отобранные контекстные сигналы (вырожденные олигонуклеотиды) были локазизованы в исследуемых последовательностях ДНК и представлены в виде таблицы данных "объект-признак".

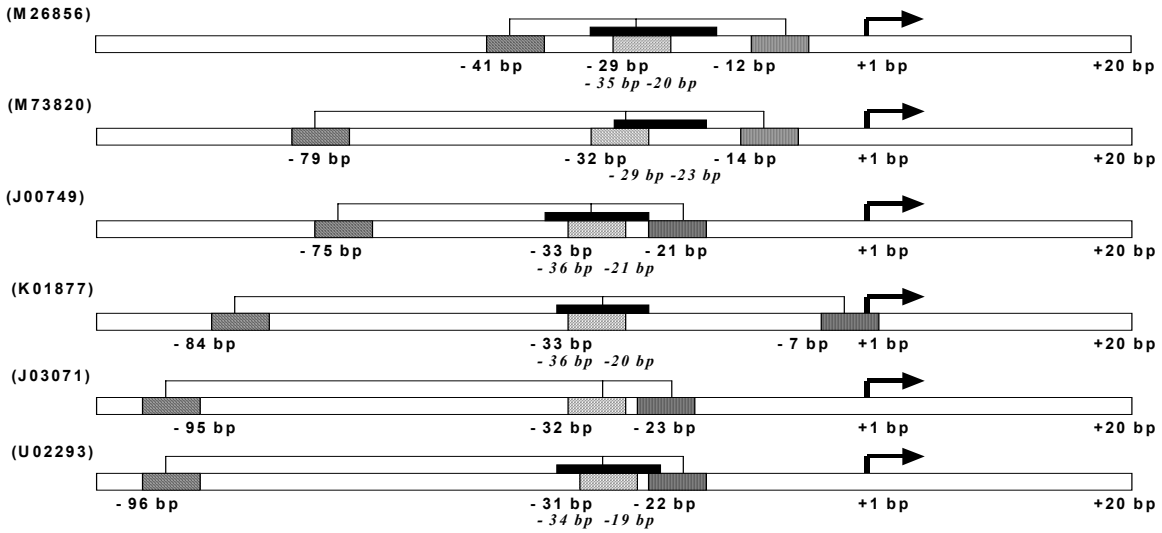

Рисунок 5. Схема расположения комплексного сигнала CWGNRGCN<NGSYMTAM<MAGKSHCN в промоторах генов эндокринной системы [11]. Последовательности промоторов сфазированы относительно старта транскрипции (позиция +1 п.о.), выделенного стрелкой. Идентификатор банка данных EMBL исследуемой последовательности 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 TATA-бокса, проиндексированное в базе данных TRRD, отмечено заштрихованными прямоугольниками

В этой таблице объектами являются последовательности ДНК, признаками – присутствие контекстных сигналов и их локализация относительно экспериментально определенного старта транскрипции. В итоге для анализа данных были построены таблицы, содержащие до нескольких тысяч строк.

Найденные закономерности (знания) имеют смысл комплексных сигналов, регулирующих транскрипцию посредством связывания с ДНК белков, специфичных к данному типу районов. Запись комплексного сигнала "CWGNRGCN<NGSYMTAM<MAGKSHCN" означает, что три олигонуклеотида в 15-буквенном алфавите имеют заданное взаимное расположение относительно старта транскрипции.

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 [11]. Исследование закономерностей совместной встречаемости и взаимного расположения сайтов с помощью системы "Gene Discovery" открывает путь для создания компьютерных методов поиска потенциальных композиционных элементов.

Представленная методика поиска комплексных сигналов имеет большое практическое значение. Информация, распределенная по научной литературе и сосредоточенная в молекулярно-биологических базах данных, содержит тысячи экспериментальных результатов о последовательностях ДНК, вовлеченных в регуляцию транскрипции генов. Такого рода данные и знания имеют важнейшее значение при решении широкого круга задач молекулярной биологии, биотехнологии, медицины и генной инженерии. Повышение точности распознавания промоторных районов, их функциональная аннотация важны для понимания механизмов работы генов и управления их работой с помощью генной инженерии, фармацевтических средств. Электронная библиотека GeneExpress (http://wwwmgs.bionet.nsc.ru/mgs/gnw/), разрабатываемая в ИЦиГ СО РАН [14], является базой не только для поиска информации, но и для разработки новых методов анализа данных и представления знаний в области молекулярной биологии.

Работа частично поддержана РФФИ (гранты 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90355, 02-07-90359, 00-04-49229, 00-04-49255), Министерством промышленности, науки и технологий Российской Федерации (№ 43.073.1.1.1501), СО РАН (Интеграционные проекты № 65, № 66), INTAS (YSF 00-178).

Литература

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

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

[3] Jakobsen I.B. et al., TreeGeneBrowser: phylogenetic data mining of gene sequences from public databases, Bioinformatics 17 (2001) 535-540.

[4] Колпаков Ф.А., Подколодный 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.

[5] Hardison R.C. Conserved non-coding sequences are reliable guides to regulatory elements, Trends Genet. 16 (2000) 369-372.

[6] J.W. Fickett and A.G. Hatzigeorgiou, Eukaryotic promoter recognition, Genome Res. 7 (1997) 861-878.

[7] 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.

[8] 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.

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

[10] 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/>

[11] 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.

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

[13] 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.

[14] 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.