Интегрированная система для информационной поддержки исследования механизмов регуляции транскрипции♣

© Н.Л. Подколодный[1,2], Е.В. Игнатьева[1], Д.А. Рассказов[1], О.А. Подколодная[1], Е.А. Ананько[1], Н.Н. Подколодная[1], Е.М. Залевский[1]

[1]Институт цитологии и генетики СО РАН, г. Новосибирск

[2]Институт вычислительной математики и математической геофизики СО РАН, г. Новосибирск

pnl@bionet.nsc.ru

Аннотация

В настоящее время накоплен колоссальный объем данных в области регуляции экспрессии генов эукариот. В данной работе представлены подходы к построению онтологии предметной области, формализации описания механизмов регуляции транскрипции и разработки на этой основе методов и системы интеграции гетерогенной информации об особенностях регуляции экспрессии генов. Результаты являются актуальными как для научных, так и прикладных исследований в области системной биологии и биоинформатики.

1 Введение

Исследование механизмов регуляции транскрипции является важной фундаментальной проблемой. Ее решение необходимо как для успешного предсказания особенностей экспрессии генов, так и для выполнения прикладных исследований, например, реконструкции регуляторных сетей, конструирования генетических конструкций с заданными свойствами, исследования механизмов заболеваний, поиск мишеней для лекарств, токсикологических исследованиях, выявлении ключевых биомаркеров и т.д.

У многоклеточных эукариотических организмов транскрипционная активность конкретного гена зависит от органа, ткани, типа клетки, стадии развития организма, стадии клеточного цикла или дифференцировки клеток, многочисленных индукторов либо репрессоров и т. д. Такая тонкая и сложная регуляция обеспечивается участием большого разнообразия регуляторных белков и механизмов их функционирования. Белки, участвующие в регуляции транскрипции, выполняют различные функции и работают в тесной кооперации, в составе сложных комплексов. Важную роль в контроле транскрипции играют транскрипционные факторы, специфически взаимодействующие с регуляторными районами генов и другими белками транскрипционной машины.

Интенсивность транскрипции гена в значительной степени определяется и другими обстоятельствами, к числу которых относятся состояние хроматина (открытый, закрытый), уровень метилирования ДНК, а также плотность нуклеосомной упаковки ДНК.

Коэкспрессирующиеся гены, имеющие сходный уровень транскрипции при определенных условиях в конкретном типе клеток, являются удобным объектом для исследования механизмов регуляции транскрипции. Исследования показывают, что регуляторные районы групп коэкспрессирующихся генов зачастую имеют общие черты организации, что выражается в наличии регуляторных паттернов (CRM – цис регуляторных модулей), состоящих из устойчивых сочетаний сайтов связывания траснкрипционных факторов различных типов и других мотивов [1]. Выявление и анализ регуляторных паттернов является основой для построения обобщенных моделей регуляторных районов группы коэкспрессирующихся генов [2] и обеспечивают понимание общего механизма регуляции транскрипции.

Наличие большого разнообразия тканей и типов клеток у животных организмов подразумевает наличие достаточно большого разнообразия механизмов регуляции транскрипции и, соответственно, большое разнообразие регуляторных паттернов, ответственных за их реализацию. В настоящее время накоплен колоссальный объем данных в области регуляции экспрессии генов эукариот, и наблюдается их непрерывный рост. В связи с этим, большую актуальность приобретают формализация описания механизмов регуляции транскрипции и разработка на этой основе методов интеграции гетерогенной

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

информации об особенностях регуляции экспрессии генов.

2 Постановка задач по исследованию механизмов регуляции транскрипции

2.1 Интеграция данных по регуляции транскрипции генов

Разрабатываемая система предназначена для компьютерной поддержки исследований механизмов регуляции транскрипции генов в различных типах клеток, тканей и органов, в частности, интеграции данных по регуляции транскрипции и экспрессионным паттернам генов из различных мировых ресурсов, накопления информации об известных механизмах регуляции транскрипции, выявления групп коэкспрессирующихся генов, обнаружения закономерностей организации районов регуляции транскрипции у групп коэкспрессирующихся генов, реконструкции гипотетических механизмов регуляции транскрипции с учетом информации о функциях и структурах регуляторных белков, присутствующих в заданных клетках или тканях на определенной стадии развития, а также закономерностей строения регуляторных районов коэкспрессирующихся генов.

Интегрированная система включает следующие компоненты:

- база данных, интегрирующая информацию, необходимую для исследования механизмов регуляции транскрипции:

- структурно-функциональная организация районов регуляции транскрипции генов;

- локализация генов, стартов транскрипции, экзон/интронной структуры и других сигналов в полных геномах;

- уровень экспрессии генов в различных тканях и органах, полученный на основе ДНК-чиповых данных;

- компьютерная аннотация регуляторных районов генов, человека, мыши, крысы функциональными сигналами, значимыми для регуляции транскрипции, включая сайты связывания транскрипционных факторов, функциональные мотивы, CpG острова, нуклеосомный потенциал, структурные закономерности, значимые для регуляции транскрипции, и т. д.;

- экспрессия генов белков-регуляторов транскрипции в различных клеточных ситуациях;

- формальное описание механизмов регуляции транскрипции, включая стадии регуляции транскрипции;

- функциональные роли регуляторов транскрипции на различных стадиях;

- компонента поиска закономерностей и построения моделей структурно-функциональной организации регуляторных районов коэкспрессирующихся генов эукариот (человека, мыши, крысы);

- компонента реконструкции сетей регуляции транскрипции генов с использованием информации об уровнях экспрессии генов и структурных моделей регуляторных районов генов.

Семантическая интеграция гетерогенных данных основывается на результатах концептуального анализа предметной области, в рамках которого определяются множество понятий (терминов) предметной области, их определений и атрибутов, отношений между ними, способы их описания и использования, а также связанных с ними аксиом и правил вывода. Такое согласованное описание конкретной предметной области называют онтологией.

На рис. 1 представлена схема интеграции данных по регуляции транскрипции и экспрессионным паттернам генов из различных источников информации: баз данных EMBL/GenBank, UniGene, EntrezGenome, EntrezGene, SWISS-PROT, TRRD и др.

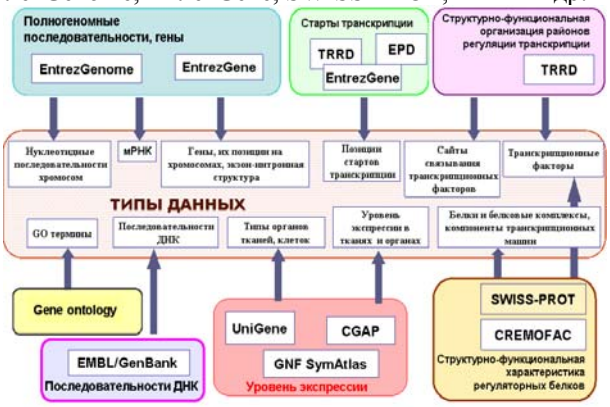

Рис. 1. Схема интеграции данных по регуляции транскрипции и экспрессионным паттернам генов из различных источников информации

2.2 Структура онтологии регуляции экспрессии генов

Одним из основных этапов семантической интеграции гетерогенных данных является согласование понятий предметной области, способов их описания и использования (сопоставления данных, обработки данных и т. д.).

Онтологии позволяют представить понятия в таком виде, что они становятся пригодными для машинной обработки и вследствие этого используются в качестве посредника между пользователем и информационной системой или между членами научного сообщества при обмене данными.

Формально онтология включает набор понятий (терминов) предметной области, их определений и атрибутов, а также связанных с ними аксиом и правил вывода.

Таким образом, формальная модель онтологии – это упорядоченная тройка конечных множеств

$$O = <T, R, F>,$$

где T – конечное и непустое множество классов и концептов (понятий, терминов) предметной области, которую описывает онтология O;

R – конечное множество отношений между концептами заданной предметной области;

F – конечное множество функций интерпретации, заданных на понятиях и/или отношениях онтологии O или аксиом, которые используются для моделирования утверждений, которые всегда являются истинными, что ограничивает интерпретацию и обеспечивает корректное использование понятий.

Разработка онтологии регуляции транскрипции является сложным и затратным процессом. Первым этап этого процесса – онтологический анализ предметной области регуляции транскрипции генов эукариот, результатом которого являются:

- словарь терминов, точных их определений и взаимосвязей между ними;
- описание правил и ограничений, согласно которым на базе введенной терминологии формируются достоверные утверждения, описывающие состояние системы;
- модель, которая на основе существующих утверждений позволяет сделать соответствующие выводы, позволяющие вносить изменения в систему для повышения эффективности её функционирования.

Структура онтологии регуляции экспрессии генов, которая разрабатывается нами, включает следующие разделы:

(1) Онтология верхнего уровня или онтология базовых знаний. В этом разделе онтологии описываются наиболее общие концепты и отношения, которые не зависят от конкретной проблемы или области.

(2) Понятия предметной области. В этом разделе онтологии описываются такие понятия предметной области, как ген, РНК, белок, геномная последовательность, район регуляции транскрипции, промотор, сайт связывания транскрипционного фактора, структура и функция белка, механизм регуляции транскрипции, путь передачи сигнала, метаболические пути, генная сеть и т. д.

(3) Онтология экспериментальных исследований и доказательств. Этот раздел онтологии включает описание экспериментальных методик, методов трансформации и интерпретации данных, обоснования и оценки достоверности получаемых научных результатов.

(4) Онтология представления знаний или **терминологическая и информационная онтология** включает *тезаурусы* и *метаописание* существующих баз данных, например, схему баз данных, описание полей, их интерпретацию в терминах онтологии предметной области. Цель – концептуализация формализмов представления знаний.

(5) Онтология задач включает описания задач, методов и программных средств решения задач. Описания задач выполняются в терминах предметной области. Описания методов решения конкретных задач могут включать такие характеристики, как эффективность, ограничения метода, точность решения задачи, вычислительные затраты, параметры программы, значения которых наиболее адекватны при решении конкретной задачи, и т. д. К этому разделу также относится и метаописание, в котором представлены способы доступа к тем или иным программам, обеспечивающим решение конкретных прикладных задач заданным методом; протоколы обращения; форматы и состав входных и выходных данных и т. д.

2.3 Онтология верхнего уровня

Пусть X – некоторый класс, а x – конкретные экземпляры этого класса, которые могут принимать различные значения. Например, будем обозначать через G, R, P, C классы РНК, генов, белков, белковых комплексов, а через g, r, p, c – конкретные гены, РНК, белки и белковые комплексы, соответственно.

В качестве базовых отношений верхнего уровня нами используются следующие отношения:

- foundational relations – is_a (has_subclass), part_of (has_part), part_for, instance_of (has_instance), includes (include_of), composed_of, consist_of;
- spatial relations – located_in, contained_in, includes, composed_of , adjacent_to;
- temporal relations – transformation_of, derives_from, preceded_by;
- participation relations – has_participant, has_agent, regulates.

Введем определения некоторых отношений:

$$x \textbf{ instance\_of } X =_{\text{def}} x \subseteq X.$$

Например: p ***instance_of*** P – белок p входит в класс белков P.

Введем отношения между классами, которые используются нами при описании предметной области:

$$X_1 \textbf{ is\_a } X_2 =_{def} \forall x: x \textbf{ instance\_of } X_1 \Rightarrow x \textbf{ instance\_of } X_2.$$

Например: P_1 ***is_a*** P_2 – любой белок из класса P_1 входит в класс P_2.

$$A \textbf{ has\_subclass } B =_{def} B \textbf{ is\_a } A.$$

Например: P_1 ***has_subclass*** $P_2 =_{def}$ P_2 ***is_a*** P_1 – любой белок из класса P_2 входит в класс P_1

$$X \textbf{ part\_for } Y =_{def} \forall x: x \textbf{ instance\_of } X \Rightarrow \exists y: (y \textbf{ in-}$$

$$\textbf{stance\_of } Y \,\&\, x \textbf{ part\_of } y).$$

Например: для любого белка из класса P_1 существует белковый комплекс из класса P_2, в который входит этот белок.

$$P_2 \textbf{ has\_part } P_1 =_{def} \forall y: y \textbf{ instance\_of } P_2 \Rightarrow \exists x:$$

$$(x \textbf{ instance\_of } P_1 \,\&\, x \textbf{ part\_of } y)$$

Например: для любого белкового комплекса из класса P_2 существует белок из класса P_1, который входит в этот белковый комплекс.

$$P_1 \textbf{ part\_of } P_2 =_{def} P_1 \textbf{ part\_for } P_2 \,\&\, P_2 \textbf{ has\_part } P_1$$

Например: для любого белкового комплекса из класса P_2 существует белок из класса P_1, который входит в этот белковый комплекс, а также для любого белка из класса P_1 существует белковый комплекс из класса P_2, в который входит этот белок. Таким образом, P_1 – класс белков, которые образуют комплексы, а P_2 – класс белковых комплексов, которые образуют белки из P_1.

Отношение ***part_of*** может иметь различный смысл. В частности, выделяют следующие типы отношений ***part_of:***

- ***part_of _Place*** – Area;
- ***part_of _Stuff*** – Object;
- ***part_of _Portion*** – Mass;
- ***part_of _Member*** – Collection;
- ***part_of _Component*** – Integral object.

$$P_1 \textbf{ part\_of } P_2 =_{def} P_1 \textbf{ part\_for } P_2 \,\&\, P_2 \textbf{ has\_part } P_1$$

Верхний уровень понятий предметной области регуляции транскрипции включает классы с отношениями ***is_a***:

- **Thing**
  - **a. Abstract_Entity**
    - **i. Quantity**
    - **ii. Proposition**
    - **iii. Attribute**
    - **iv. Relation**
    - **v. Role**
  - **b. Physical_Entity**
    - **i. Biomaterials**
    - **ii. Genome_Entity**
    - **iii. …**
  - **c. Occurrence**
    - **i. Process**
    - **ii. Event**
    - **iii. State**
    - **iv. Situation**
    - **v. Causation**

2.4 Понятия предметной области

Основой для формирования онтологии регуляции транскрипции генов является формальное представление следующих понятий:

- Основные понятия предметной области - физические сущности (**Physical_Entity**), в частности, ген, рнк, белок, белковый комплекс, геномная последовательность, район регуляции транскрипции, промотор, сайт связывания транскрипционного фактора, нуклеосома, транскрипционный фактор, регулятор транскрипции и т. д.
- механизм регуляции транскрипции;
- стадии регуляции транскрипции с ролями, которые играют участники регуляции;
- регуляторные события, которые обуславливают реализацию механизма;
- описание клеточных ситуаций, в которых получены экспериментальные данные по экспрессии генов;
- свойства регуляторов транскрипции, которые коррелируют с их функциональными возможностями; компьютерное предсказание этих свойств позволяет, например, делать выводы о возможности участия конкретного белка в регуляции транскрипции на определенной стадии, т. е. выполнение определенной роли на этой стадии;
- структурно-функциональные закономерности организации регуляторных районов генов (регуляторные структурные модули), обуславливающих особенности регуляции экспрессии генов, коэкспрессирующихся в разных клеточных ситуациях.

На рис. 2 представлена схема фрагмента раздела “Biomaterials” онтологии регуляции транскрипции.

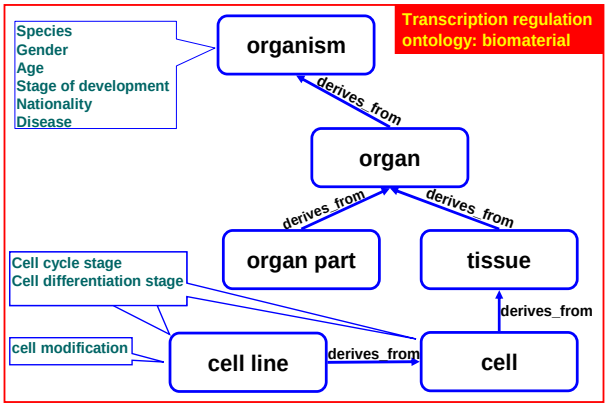

Рис. 2 Фрагмент раздела “Biomaterials” онтологии регуляции транскрипции

На рис. 3 представлен фрагмент раздела “Genome_Entity” онтологии регуляции транскрипции.

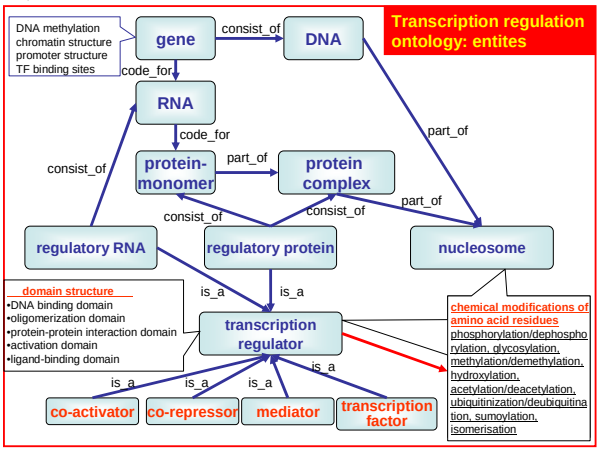

Рис. 3. Фрагмент раздела “Genome_Entity” онтологии регуляции транскрипции

2.5 Представление знаний о механизмах регуляции транскрипции

Под механизмом функционирования молекулярно-генетической системы будем понимать описание структуры этой системы и множества взаимосвязанных событий, которые определяют поведение

системы и роли, которые играют отдельные элементы системы в реализации тех или иных событий.

Механизм регуляции транскрипции можно описывать на разном уровне детальности, и полнота описания зависит от наших знаний и возможностей.

Следует отметить, что знания о механизме регуляции транскрипции генов основываются на интеграции гетерогенных знаний о биологических объектах (белках, генах, рнк и др.), вовлеченных в регуляторный процесс, их структурно-функциональной организации и ролях, которые они играют на различных стадиях регуляции.

Механизм регуляции транскрипции удобно характеризовать с помощью таких понятий, как *событие*, *действие*, *процесс*.

Таким образом, для описания механизма регуляции транскрипции необходимо выделить основные подпроцессы, из которых складывается это биологическое явление, описать основных участников этих процессов и их ролевые функции.

В качестве участников процесса регуляции транскрипции выступают гены и регуляторы транскрипции различного типа, включая транскрипционные факторы, активаторы, медиаторы.

Пространство описания понятий предметной области определяется необходимостью отвечать на вопросы ЧТО? (уровень экспрессии генов в клетках конкретного биоматериала), ГДЕ и КОГДА? (описание биоматериалов и клеточной ситуации: вид организма, состояние организма, индукторы, органы, ткани, клетки, их стадии развития), КАК и ПОЧЕМУ? (механизмы регуляции транскрипции и их нарушение).

В качестве основных типов молекулярно-генетических событий, которые играют важную роль в регуляции транскрипции, можно выделить:

- связывание (bind);
- освобождение (release);
- расщепление (cleavage);
- модификации (modify):
  - модификации, связанные с появлением новых связей, например, phosphorylate, glycosylate, methylate, hydroxylate, acetylate, acylate, ubiquitinize;
  - модификации, связанные с разрушением связей, например, dephosphorylate, glycosylate, demethylate, dehydroxylate, deacetylate, deacylate, deubiquitinize;
- транспорт (transport).

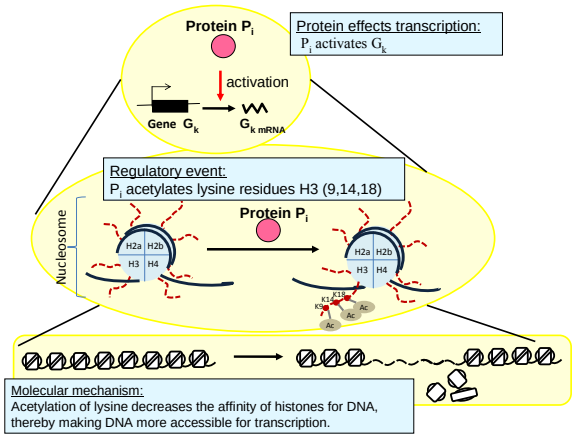

Рис. 4. Пример описания регуляторных событий, определяющих механизм регуляции транскрипции

Так, например, ковалентные модификации гистонов являются одним из механизмов регуляции транскрипции. В частности, знание о том, что белок $\mathbf{P_i}$ активирует экспрессию гена $\mathbf{G_k}$ путем ацетилирования аминокислоты лизина в позициях 9, 14 и 18 гистона H3 может быть формально представлено в виде (см. рис. 4):

$\mathbf{P_i}$ activates $\mathbf{G_k}$ through Event($\mathbf{P_i}$ acetylate lysine.H3(9,14,18)).

Интерпретация полученных данных с учетом известных механизмов функционирования транскрипционных факторов позволит сформировать новые гипотезы механизмах регуляции транскрипции, обеспечивающих координированную регуляцию групп коэкспрессирующихся генов.

2.5 Анализ данных и построение множества непротиворечивых гипотез о механизмах регуляции траснкрипции генов

Анализ данных используется в системе как для извлечения знаний из слабоструктурированных данных, так и для вывода новых знаний и их интерпретации.

Часть информации, которая должна быть интегрирована в систему, представлена в неформализованном виде в текстовых источниках. В частности, структура белковых комплексов, участвующих в регуляции транскрипции, описана в текстовых полях базы данных SWISS-PROT. Для работы с такого рода источниками информации нами были разработаны специальные методы извлечения знаний с использованием методов text-mining.

Нами построены правила, использующие текстовые шаблоны, для извлечения информация о белковых комплексах и их составе, взаимодействии белков, участии в регуляции транскрипции. На рис. 5 представлен пример структуры шаблона для извлечения знаний о составе белкового комплекса.

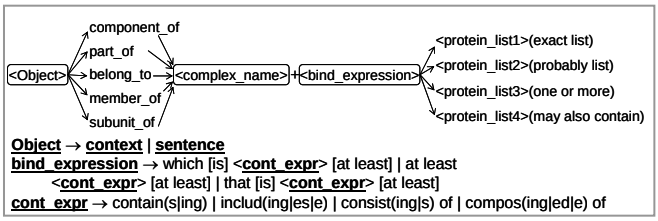

Рис. 5. Пример структуры и фрагмент описания одного из текстовых шаблонов для извлечения из текстовых источников информации о структуре белкового комплекса

Как правило, информация о составе и функции белковых комплексов фрагментарна, и требуется проводить логический анализ всей совокупности информации для вывода знаний о структуре и функции белкового комплекса.

Приведем простой пример логического анализа информации о структуре белковых комплексов. В поле “Subunit structure” описания белка “TFIIH basal transcription factor complex helicase subunit” в базе данных SWISS-PROT (Имя белка – ERCC2_HUMAN) приводится следующая информация: «One of the **six subunits** forming the **core-TFIIH basal transcription factor** which associates with the **CAK complex** composed of **CDK7**, **CCNH/cyclin H** and **MNAT1** to form the **TFIIH basal transcription factor**. …».

Первая фаза анализа включает выявление имен комплексов и белков с применением правил в виде текстовых шаблонов и сравнения имен со словарем белков. В результате из этого поля извлечены следующие знания:

1. “core-TFIIH basal transcription factor” includes (“ERCC2”);
2. number_of_components(“core-TFIIH basal transcription factor”) := (6,6); // интервал значений
3. “CAK complex” composed_of(CDK7, “CCNH/cyclin H”, MNAT1);
4. “TFIIH basal transcription factor” composed_of (“core-TFIIH basal transcription factor”, “CAK complex”);

Дальнейший анализ информации из базы данных SWISS-PROT выявил 5 других белков, которые также входят в комплекс:

5. “core-TFIIH basal transcription factor” includes (GTF2H1);
6. “core-TFIIH basal transcription factor” includes (GTF2H2);
7. “core-TFIIH basal transcription factor” includes (GTF2H3);
8. “core-TFIIH basal transcription factor” includes (GTF2H4);
9. “core-TFIIH basal transcription factor” includes (ERCC3);

Здесь:

A includes B, C $\equiv_{def}$ (A includes B) & (A includes C).

Используя правило:

IF(A includes $(B_1, \ldots, B_n)$) & number_of_components(A)=(,n) THEN (A composed_of $(B_1, B_1, \ldots, B_n)$),*

получаем точное описание структуры белкового комплекса:

“TFIIH basal transcription factor” composed_of (GTF2H1, GTF2H2, GTF2H3, GTF2H4, ERCC2, ERCC3, CDK7, “CCNH/cyclin H”, MNAT1).

Такой результат с четко определенной структурой и функцией белкового комплекса не всегда возможно получить.

Знания, полученные из гетерогенных источников, могут быть неполными, нечеткими, косвенными и противоречивыми. В частности, может оказаться известным только то, что белок в составе некоторого неизвестного комплекса участвует в регуляции транскрипции. Знания о составе белкового комплекса могут быть неполными. Например, не все субъединицы комплекса известны или неизвестно, сколько всего субъединиц входит в комплекс.

В случае неполных, нечетких, косвенных или противоречивых данных нами используется метод генерации правдоподобных гипотез, которые не противоречат известным фактам. Такого рода гипотетические знания с указанием относительного уровня достоверности полезны при дальнейшем анализе и построении непротиворечивой картины мира.

В некоторых случаях имеется возможность усилить уровень достоверности гипотезы путем привлечения дополнительной информации, которая также не противоречит этой гипотезе. Это позволяет задать частичный порядок на множестве гипотез по уровням относительной достоверности.

Пусть, например, известно, что некоторый белок в составе неизвестного комплекса участвует в регуляции транскрипции. Среди множества белковых комплексов, в состав которых входит этот белок, те комплексы, в состав которых входят другие белки, имеющие транскрипционную активность, с большой вероятностью могут быть транскрипционными факторами.

Примером косвенных знаний могут быть знания о взаимодействии между субъединицами, участвующими в регуляции транскрипции. Эти знания дают основание предположить, что участие обоих этих белков в регуляции транскрипции может осуществляться через образование транскрипционного комплекса, в который входят оба белка. Это предположение становится более правдоподобным, если известно, что действие этих белков на транскрипцию одинаково (либо подавление, либо усиление транскрипции).

В ряде случаев можно распространять свойства через мереологические иерархии. В качестве примера вывода гипотетических свойств белкового комплекса по свойствам субъединиц можно привести связывание с ДНК (DNA_binding). Наличие ДНК связывающего домена в субъединице позволяет сделать предположение о возможности связывания белкового комплекса, в который входит эта субъединица:

$$\forall x, y, z : rel(x, y) \vee part\text{-}of(y, z) \rightarrow rel(x, z).$$

Безусловно, это предположение может рассматриваться только как гипотеза, и только экспериментальная проверка может подтвердить этот факт.

3 Результаты и выводы

С целью информационной поддержки исследования механизмов тканеспецифичной регуляции транскрипции генов нами разработана система RETRA [6], интегрирующая информацию, необходимую для исследования механизмов регуляции транскрипции, включая:

- структурно-функциональную организацию районов регуляции транскрипции генов; локализации генов, стартов транскрипции, экзон/интронной структуры и других сигналов в полных геномах (EntrezGene, RefSeq, TRRD [7]);
- уровень экспрессии генов в различных тканях и органах на основе ДНК-чиповых данных и EST;
- функциональную аннотацию генов (Gene Ontology);
- компьютерную аннотацию регуляторных районов генов, человека, мыши, крысы функциональными сигналами, значимыми для регуляции транскрипции, включая сайты связывания транскрипционных факторов, функциональных мотивов, CpG острова, нуклеосомный потенциал, структурные закономерности, значимые для регуляции транскрипции, и т. д.;
- экспрессию генов белков-регуляторов транскрипции в различных клеточных ситуациях;
- формальное описание механизмов регуляции транскрипции, включая стадии регуляции транскрипции, функциональные роли регуляторов транскрипции на различных стадиях.

Одной из функций системы являются анализ экспериментальных данных, поиск закономерностей и построение моделей структурно-функциональной организации регуляторных районов коэкспрессирующихся генов, предсказание ролей белков-регуляторов на различных стадиях регуляции транскрипции, реконструкции сетей регуляции транскрипции генов с использованием информации об уровнях экспрессии генов и структурных моделей регуляторных районов генов, генерации гипотез о механизмах регуляции транскрипции; интерпретация экспериментальных данных по экспрессии генов в терминах механизмов регуляции транскрипции. На рис. 6 приведен типовой сценарий исследования механизмов регуляции коэкспрессирующихся генов.

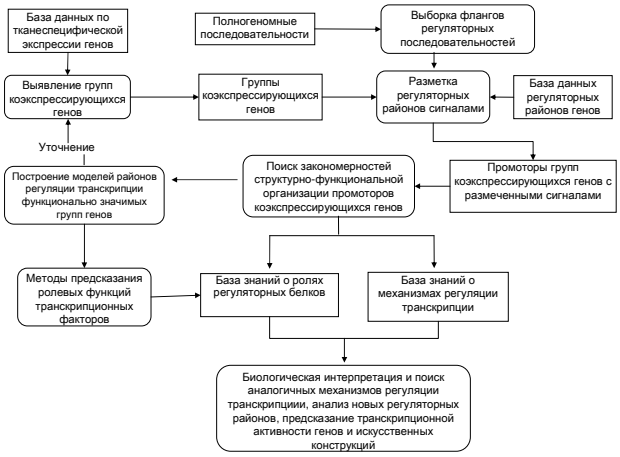

Рис. 6. Типовой сценарий использования системы при исследовании механизмов регуляции коэкспрессирующихся генов

В качестве примеров запросов, на которые ориентирована система, можно привести следующие запросы:

- экстракция определенных участков генов (промоторных областей, интронов, экзонов, 5'-НТП, 3'-НТП и др.);
- выявление групп генов, расположенных определенным образом в геноме (гены с определенного участка хромосомы, вложенные гены, перекрывающиеся гены);
- поиск множеств коэкспрессирующихся генов (ткане-, стадиеспецифичных и др.);
- поиск структурно-функциональных закономерностей организации промоторов коэкспрессирующихся генов;
- реконструкция сетей регуляции транскрипции.

Литература

[1] Blanchette M., Bataille A.R., Chen X et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression// Genome Res. – 2006. – V. 16, No 5. – P. 656-668.

[2] Krivan W., Wasserman W.W. A predictive model for regulatory sequences directing liver-specific transcription// Genome Res. – 2001. – V. 11, No 9. – P. 1559-1566.

[3] Smith B., Ceusters W., Klagges B. et al. Relations in biomedical ontologies // Genome Biology. – 2005. – V. 6. – No R46.

[4] Rzhetsky A., Koike T., Kalachikov S. et al. A knowledge model for analysis and simulation of regulatory networks // Bioinformatics. – 2000. – V. 16, No 12. – P. 1120-1128.

[5] Hoehndorf R., Kelso J., Herre H. The ontology of biological sequences // BMC Bioinformatics. – 2009. – V. 10, No 377.

[6] Podkolodnyy N.L., Nechkin S.S., Ignatieva E.V. et al. A database for analysis of the organizational fea-

tures of the promoter regions in the co-expressed groups of genes // Proc. of the Sixth Int. Conf. on Bioinformatics of Genome Regulation and Structure, 2008.
[7] Kolchanov N.A. et al. Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucl. Acids Res. – 2002. – V. 30. – P. 312-317.

Integrated system for information support of research on transcription transcription regulation mechanisms

N.L. Podkolodnyy, E.V. Ignatyeva, D.A. Rasskazov, O.A. Podkolodnaya, E.A. Ananko, N.N. Podkolodnaya, E.M. Zalevsky

Now the huge volume of experimental data in the field of gene expression regulation has been accumulated. This paper describes the approaches to construction of ontology of subject domain, formalization of the description of mechanisms of regulation of a transcription and developing on this basis the methods and systems of integration of the heterogeneous information on features of regulation of an expression of genes. The integrated system for study the mechanism of gene transcription regulation was developed using ontology based approach.

These workings out are actual as for scientific, and applied researches in the field of system biology and bioinformatics.

♣ Работа выполнялась при поддержке Министерства образования и науки РФ (госконтракты П721, П857), СО РАН (Междисциплинарные интеграционные проекты 119, 26)