УДК 577.214:004.822

ИНФОРМАЦИОННАЯ ПОДДЕРЖКА ИССЛЕДОВАНИЯ МЕХАНИЗМОВ РЕГУЛЯЦИИ ТРАНСКРИПЦИИ: ОНТОЛОГИЧЕСКИЙ ПОДХОД

© 2012 г. **Н.Л. Подколодный[1, 2], Е.В. Игнатьева[1], О.А. Подколодная[1], Н.А. Колчанов[1, 3, 4]**

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

Поступила в редакцию 15 июля 2012 г. Принята к публикации 31 августа 2012 г.

В настоящее время накоплен колоссальный объем данных в области регуляции транскрипции генов эукариот, которая контролируется при участии большого количества белков, выполняющих различные функции в зависимости от стадии процесса транскрипции, что создает возможность реализации большого разнообразия механизмов регуляции. В данной работе представлены подходы к построению онтологии предметной области, формализации описания механизмов регуляции транскрипции и разработке на этой основе методов интеграции гетерогенной информации об особенностях регуляции экспрессии генов эукариот и базы знаний по механизмам регуляции транскрипции. Описана пилотная версия базы знаний по регуляции транскрипции генов эукариот, которая включает понятия, связанные с процессом регуляции транскрипции; иерархическую классификацию регуляторов транскрипции; классификацию этапов и стадий транскрипции, а также базу данных транскрипционных регуляторов трех видов млекопитающих (человека, мыши, крысы) и словари по молекулярным процессам, обеспечивающим регуляцию транскрипции. База знаний предназначена для информационной поддержки исследования механизмов тканеспецифичной регуляции транскрипции генов. Рассмотрены подходы к построению гипотез о механизмах регуляции транскрипции генов эукариот с использованием информации из базы знаний.

Ключевые слова: биоинформатика, регуляция транскрипции, базы знаний, онтология.

ВВЕДЕНИЕ

Транскрипция генов эукариот – сложный процесс, который осуществляется при участии РНК-полимераз трех типов: Pol I, Pol II, и Pol III, каждая из которых обеспечивает транскрипцию определенного набора генов со специфическими механизмами регуляции (Carey, Smale, 2000; Kolchanov *et al*., 2002, 2008). Транскрипционная активность конкретного гена многоклеточного эукариотического организма зависит от типа клетки, ткани и органа, стадии развития организма, стадии клеточного цикла, этапа дифференцировки клеток, воздействия многочисленных индукторов и репрессоров и т. д. Кроме того, транскрипция генов эукариот контролируется большим количеством регуляторных белков, выполняющих различные функции на разных стадиях регуляции транскрипции и работающих в тесной кооперации в составе сложных комплексов (рис. 1).

Например, процесс инициации транскрипции, т. е. позиционирование РНК-полимеразы в районе старта транскрипции гена с последующим образованием короткой цепи РНК (2–9 оснований), осуществляется при участии ба-

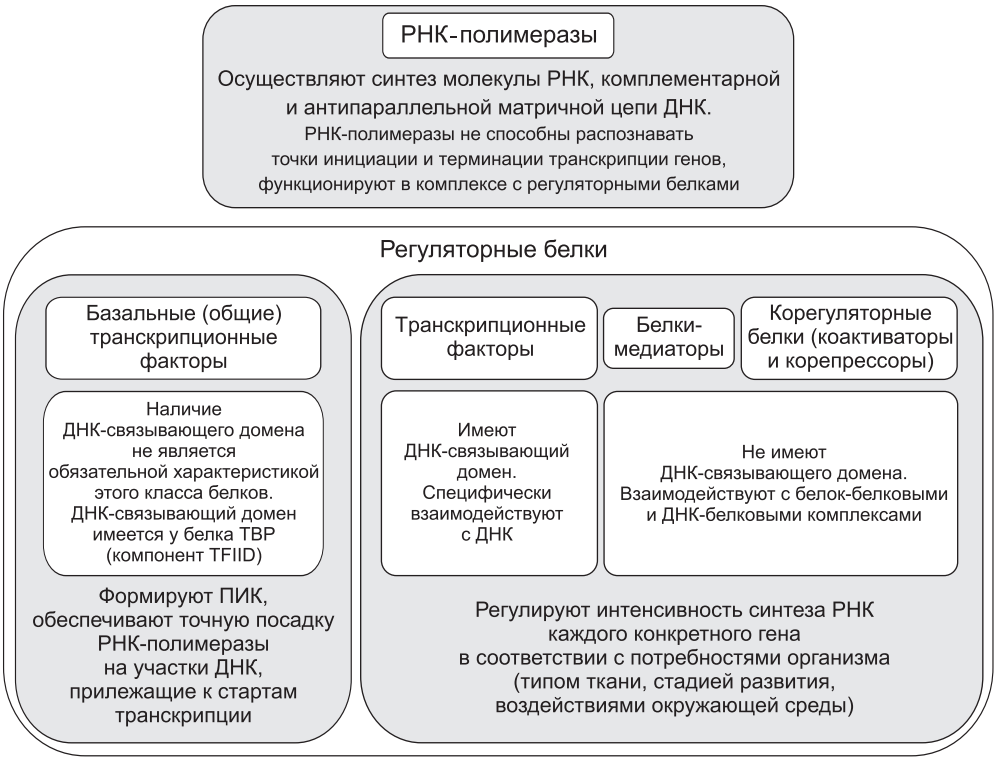

Рис. 1. Основные классы белков, участвующих в регуляции транскрипции генов эукариот.

ПИК – предынициаторный комплекс, включающий РНК-полимеразу и базальные транскрипционные факторы.

зальных транскрипционных факторов, которые являются общими для всех генов, транскрибируемых конкретной РНК-полимеразой. Еще один класс регуляторных белков составляют транскрипционные факторы (ТФ), каждый из которых специфичным образом осуществляет регуляцию определенных групп генов в соответствии с клеточной ситуацией (Lemon, Tjian, 2000). ТФ взаимодействуют с определенными участками ДНК в регуляторных районах генов – сайтами связывания транскрипционных факторов (ССТФ) и влияют на интенсивность транскрипции. Обязательным атрибутом ТФ является наличие ДНК-связывающего домена, участвующего в распознавании специфических сигналов (сайтов связывания) в регуляторных районах генов, регулируемых конкретным ТФ.

Помимо ТФ к числу регуляторов транскрипции относятся белки-медиаторы и корегуляторные (кофакторные) белки, которые, как правило, не имеют ДНК-связывающих доменов и участвуют в регуляции транскрипции без непосредственного специфического взаимодействия с ДНК. Белки-медиаторы в составе медиаторного комплекса взаимодействуют с РНК-полимеразой II в области ее С-концевого домена и стабилизируют контакт РНК-полимеразы II с ДНК (Hahn, 2004). К числу корегуляторов транскрипции относятся белки, ковалентно модифицирующие гистоны и белки, осуществляющие АТФ-зависимую реорганизацию (ремоделирование) хроматина.

Процесс регуляции транскрипции можно разбить на этапы, каждый из которых характеризуется набором регуляторных событий и их участников.

В настоящее время накоплен колоссальный объем данных в области регуляции экспрессии генов эукариот, наблюдается их непрерывный рост. В связи с этим большую актуальность приобретают формализация описания механизмов регуляции транскрипции и разработка на этой основе методов интеграции гетерогенной информации об особенностях регуляции экспрессии генов.

Проблемы разработки онтологии регуляции экспрессии генов

Одним из основных этапов семантической интеграции гетерогенных данных является согласование понятий предметной области, их определений и атрибутов, отношений между ними, способов их описания и использования, а также связанных с ними аксиом и правил вывода. Такое согласованное описание конкретной предметной области называют онтологией (Smith *et al*., 2005).

В настоящее время онтологическое моделирование и построение онтологии становятся существенной частью современной биоинформатики и активно применяются при накоплении, сравнении, интеграции и анализе больших объемов гетерогенных данных, полученных с использованием высокопроизводительных экспериментальных исследований в масштабе генома (Подколодный, 2011).

В качестве примера можно привести онтологии, представленные в Open Biological Ontologies (OBO) (http://obo.sourceforge.net). Здесь содержится описание более 70 онтологий по различным направлениям, включая анатомию, биохимию, биологические процессы, биологические функции, биологические последовательности, здоровье, окружающую среду, экспериментальные доказательства, фенотип, белки, таксономии и др. (Schober *et al*., 2009). В рамках проекта OBO разрабатываются унифицированные подходы для разработки онтологий, методы интеграции онтологий, а также инструментальные средства для работы с онтологиями (Smith *et al*., 2007).

Одним из самых успешных проектов создания онтологии является Gene Ontology (GO) (http://www.geneontology.org/), которая включает 3 раздела: биологические процессы (biological process), биологические структуры (cellular component) и молекулярные функции (molecular function), которые выполняют гены, РНК или белки, локализованные в определенных клетках или клеточных структурах в том или ином биологическом процессе (Gene Ontology Consortium, 2010).

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:

$$\mathbf{O} = < \mathbf{T, R, F} >,$$

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, используемых для моделирования утверждений, которые всегда являются истинными, что ограничивает интерпретацию и обеспечивает корректное использование понятий.

Одним из наиболее продуктивных подходов к описанию и использованию знаний о предметной области являются дескриптивные логики (ДЛ), которые определяют формальный язык для описания понятий (концéпт, класс, категория или сущность) и отношений между понятиями (называемых ролями), утверждений о фактах и запросов к ним. Кроме этого, в ДЛ входят конструкторы (операции) для понятийных выражений, включающие конъюнкцию, дизъюнкцию и определение отношений.

Базы знаний предметной области с позиции дескриптивной логики подразделяются на общие знания о понятиях и их взаимосвязях (T-Box) и знания об индивидуальных объектах, их свойствах и связях с другими объектами (A-Box).

T-box (terminological knowledge) – это набор утверждений, описывающих множество классов понятий предметной области, их свойства и отношения между ними. Эти знания более стабильны и постоянны. Именно эти знания соответствуют онтологии предметной области.

A-box (assertional knowledge) содержит утверждения об экземплярах понятий, т. е. описывает предметную область на уровне конкретных данных (база данных). В базе знаний обе компоненты взаимосвязаны.

Разработка онтологии регуляции транскрипции является сложным и затратным процессом.

Первый этап этого процесса – онтологический анализ предметной области регуляции транскрипции генов эукариот, включая создание словаря терминов, точных их определений и взаимосвязей между ними, описание правил и ограничений, согласно которым на базе введенной терминологии формируются достоверные утверждения о состоянии системы.

С использованием стандартов OBO разрабатывается онтология регуляции генов Gene Regulation Ontology (GRO) (Beisswanger *et al*., 2008), которая включает 508 классов, в том числе и классы, описывающие процессы различных типов воздействия, биологические процессы, экспериментальные воздействия, молекулярные процессы, мутации, регуляторные процессы и т. д.

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

Механизм регуляции транскрипции удобно характеризовать с помощью таких понятий, как событие, действие, процесс. В этом случае 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, корегуляторные белки, белки-медиаторы и т. п.

Пространство описания понятий предметной области определяется необходимостью отвечать на вопросы: ЧТО? (описание ситуации или события, например уровень экспрессии генов в клетках конкретного биоматериала), ГДЕ и КОГДА? (локализация события во времени (возможно с точностью до отношения к моменту другого события) и в пространстве (возможно с точностью до компартмента), описание биоматериалов и клеточной ситуации: вид организма, состояние организма, индукторы, органы, ткани, клетки, их стадии развития), КАК (механизмы регуляции транскрипции и их нарушение) и ПОЧЕМУ? (множество событий, которые необходимы для понимания семантики конкретного события).

Роль объекта определяется в контексте реализации конкретного события, которое изменяет значения атрибутов объектов, определяющих ситуацию. Поэтому для каждого события необходимо определить роли участников этих событий, которые необходимы для реализации события.

Таким образом, представление механизма регуляции транскрипции включает описание структуры системы и участников процесса регуляции транскрипции, множества возможных состояний системы, множества взаимосвязанных событий, которые определяют поведение системы и меняют состояние системы, а также роли, которые играют отдельные элементы системы в реализации тех или иных событий.

В общем случае ситуация может быть охарактеризована предикатом, который является истинным или ложным в зависимости от того, наблюдается или нет данная ситуация. Ситуации являются абстрактными сущностями и могут обладать различными свойствами. В этом отношении ситуации могут быть простыми (т. е. не иметь внутренней структуры и быть пределом точности описания в данной модели внешнего мира) и сложными, имеющими определенную структуру, включающими подмножество ситуаций.

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

– 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;

– регуляторные события, обусловливающие реализацию механизмов регуляции транскрип-

ции и роли, которые играют участники в этих событиях;

− 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. выполнении определенной роли на этой стадии;

− структурно-функциональные закономерности организации регуляторных районов генов (регуляторные структурные модули), обусловливающих особенности регуляции экспрессии генов, коэкспрессирующихся в разных клеточных ситуациях (Подколодный и др., 2010).

К отношениям верхнего уровня относятся базовые отношения (например is_a/has_subclass, part_of/has_part, part_for, instance_of/has_instance, includes/include_of, composed_of, consist_of), пространственные отношения (например located_in, contained_in, includes, composed_of, adjacent_to), временные, или темпоральные, отношения (например transformation_of, derives_from, preceded_by), отношения участия (например has_participant, has_agent, regulates и т. д. (Özgövde *et al*., 2010).

Ниже приведены определения и примеры использования некоторых базовых отношений между классами, которые применяются нами при описании предметной области:

Отношение ***is_a*** (класс–подкласс):

X ***is_a*** $Y =_{def} \forall x : x$ ***instance_of*** $X \Rightarrow y$ ***instance_of*** Y.

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

Отношение ***part_of***:

X ***part_of*** $Y =_{def} \forall x, t : x$ ***instance_of*** X *at* $t \Rightarrow \exists y : (y$ ***instance_of*** Y *at* t *&* x ***part_of*** y *at* $t)$.

Пример. P_1 ***part_of*** P_2 – для любого белкового комплекса из класса P_2 независимо от времени t 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.

На рис. 2 в качестве примера представлена схема фрагмента раздела «Biomaterials» онтологии регуляции транскрипции.

На рис. 3 представлен фрагмент раздела «Genome_Entity» онтологии регуляции транскрипции.

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

− связывание (bind);

− освобождение (release);

− расщепление (cleavage);

− модификации (modify), включая:

− модификации, связанные с появлением новых связей, например phosphorylate, glycosylate, methylate, hydroxylate, acetylate, acylate, ubiquitinize и др.;

− модификации, связанные с разрушением связей, например dephosphorylate, glycosylate, demethylate, dehydroxylate, deacetylate, deacylate, deubiquitinize и др.;

− транспорт (transport).

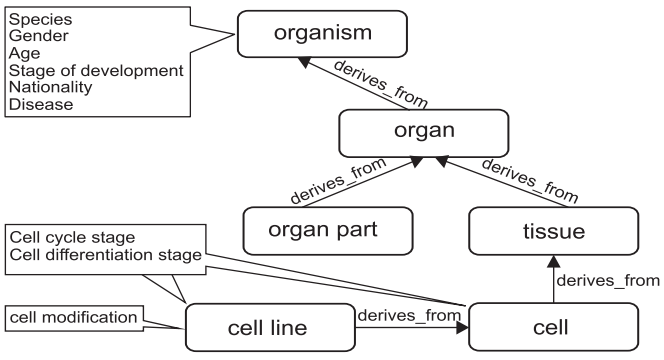

Рис. 2. Фрагмент раздела «Biomaterials» онтологии регуляции транскрипции.

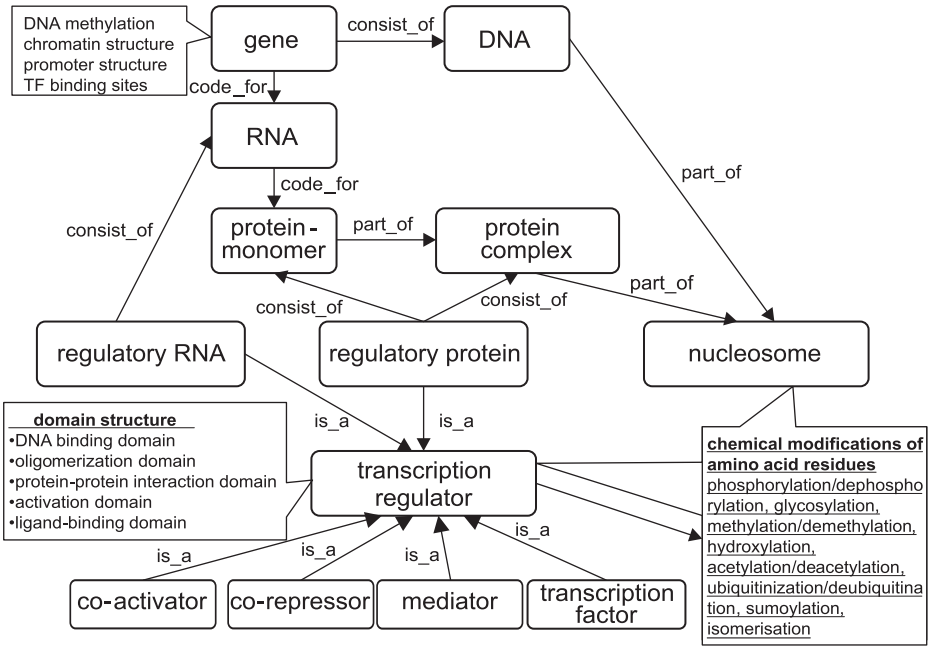

Рис. 3. Фрагмент раздела «Genome_Entity» онтологии регуляции транскрипции.

База знаний по регуляции транскрипции

С целью систематизации данных о механизмах регуляции транскрипции генов эукариот нами разработана база знаний по регуляции транскрипции, включающая иерархически организованные словари (классификаторы) белков, регулирующих транскрипцию, стадий транскрипции, молекулярных процессов, обеспечивающих регуляцию транскрипции, а также сведения о транскрипционных регуляторах трех видов млекопитающих: человека, мыши, крысы (табл. 1) (Shipra *et al*., 2006; Podkolodnyy *et al*., 2008; Schaefer *et al*., 2011).

Таблица 1

Количество записей
по транскрипционным регуляторам
раздела A-box базы знаний

Организм	Человек	Мышь	Крыса
Транскрипционные факторы	1365	1301	1267
Транскрипционные кофакторы	536	521	508
Факторы, участвующие в ремоделировании хроматина	64	63	63

Построена классификация регуляторов транскрипции (рис. 4), имеющая иерархическую структуру (до 4-го уровня иерархии). Классы белков в классификации характеризуются через их функциональные роли в процессе транскрипции либо его регуляции. Понятия первого уровня иерархии соответствуют терминам, обозначающим активности РНК-полимераз (DNA-directed RNA polymerase activity) и транскрипционных регуляторов (transcription regulator activity). Понятия второго уровня описывают функциональные роли белков, либо указывая на тип РНК-полимеразы, в комплексе с которой работает белок, либо характеризуя белок как кофакторный (корегуляторный). Понятия третьего и четвертого уровней более детально характеризуют активности функциональных подклассов белков, регуляторов транскрипции.

Например, понятие второго уровня «транскрипционные кофакторы» (transcription cofactor activity) включает термин следующего, третьего, уровня, обозначающий активность белков, модифицирующих гистоны (histone modification activity). Данный термин, в свою очередь, имеет подчиненные понятия четвертого уровня иерархии, уточняющие механизм функционирования транскрипционных регуляторов конкретных подклассов: histone kinase activity, histone acetyl-

1. DNA-directed RNA polymerase activity
   1.1. DNA-directed RNA polymerase I activity
   1.2. DNA-directed RNA polymerase II activity
   1.3. DNA-directed RNA polymerase III activity
2. transcription regulator activity
   2.1. RNA polymerase I transcription factor activity
   2.2. RNA polymerase II transcription factor activity
      2.2.1. general RNA polymerase II transcription factor activity
      2.2.2. specific RNA polymerase II transcription factor activity
   2.3. RNA polymerase III transcription factor activity
   2.4. transcription cofactor activity
      2.4.1. histone modification activity*
         2.4.1.1. histone kinase activity
         2.4.1.2. histone acetyltransferase activity
         2.4.1.3. histone methyltransferase activity
         2.4.1.4. histone deacetylase activity
         2.4.1.5. histone demethylase activity
      2.4.2. ATP-dependent chromatin remodeling activity*

Рис. 4. Фрагмент иерархической классификации функциональных ролей белков, участвующих в процессе транскрипции и ее регуляции.

* Понятия, отсутствовавшие в GO и включенные на основе анализа научных публикаций.

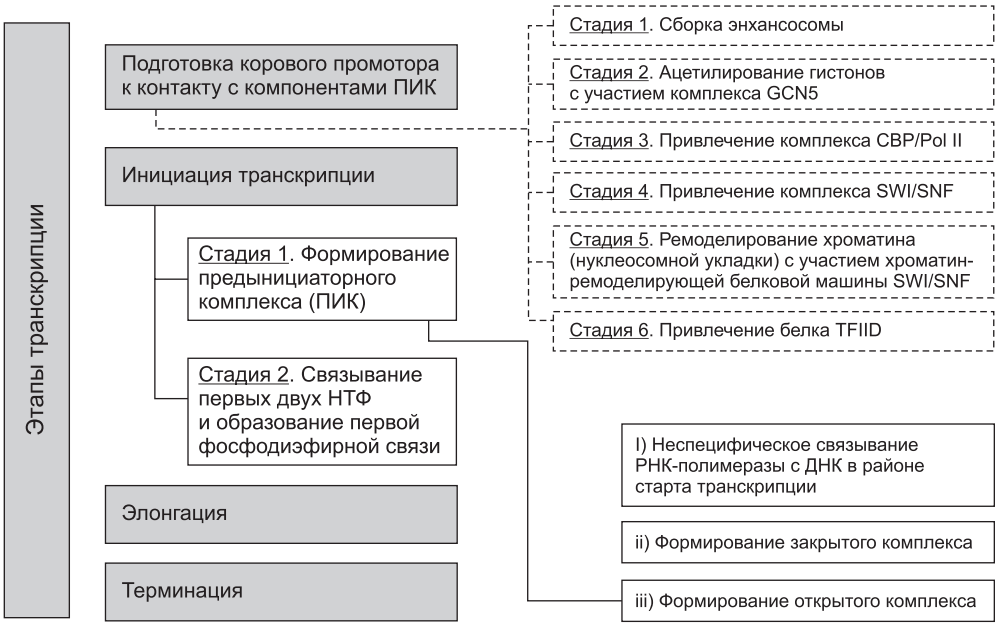

Рис. 5. Этапы и стадии процесса транскрипции генов эукариот.

Этапы обозначены серыми прямоугольниками. В качестве примера для этапа «инициация транскрипции» приведены понятия второго (стадии) и третьего уровней (процессы), которые обозначены белыми прямоугольниками со сплошной границей. Стадии, специфичные для этапа подготовки корового промотора конкретного гена (интерферона β человека), обозначены прямоугольниками, ограниченными пунктиром.

transferase activity, histone methyltransferase activity и др.

Классификация этапов и стадий транскрипции включает понятия нескольких иерархических уровней, соответствующих упорядоченным по времени (т. е. следующих друг за другом) процессам. Понятия верхнего уровня иерархии соответствуют основным этапам, посредством которых осуществляется транскрипция генов эукариотических организмов (рис. 5). Необходимо отметить, что этап подготовки корового промотора к контакту с компонентами предынициаторного комплекса (ПИК) специфичен для процесса транскрипции генов эукариот, в силу того что геномная ДНК эукариот находится в комплексе с белками (хроматин), что затрудняет взаимодействие белков транскрипционной машины с ДНК в области промотора (Разин, 2007; Berger, 2007).

Понятия следующего уровня иерархии соответствуют стадиям, посредством которых реализуется конкретный этап. Например, этап инициации транскрипции включает две стадии. На первой стадии происходит формирование предынициаторного комплекса, включающего РНК-полимеразу и вспомогательные белки (базальные транскрипционные факторы), на второй стадии осуществляются связывание первых двух нуклеотидтрифосфатов и образование первой фосфодиэфирной связи вновь синтезированного транскрипта РНК (Hahn, 2004).

В свою очередь, стадия формирования ПИК включает следующие процессы:

– первоначально РНК-полимераза связывается с двухцепочечной ДНК неспецифически;

– затем РНК-полимераза в составе ПИК связывается с ДНК в районе промотора специфически, благодаря электростатическим взаимодействиям и формирует закрытый комплекс, в котором ДНК сохраняет двухспиральную структуру.

– далее закрытый комплекс превращается в открытый, в котором РНК-полимераза расплетает двойную спираль ДНК в районе точки инициации транскрипции (Hahn, 2004).

Классификация этапов и стадий транскрипции включает понятия двух типов:

1) общие для всех генов;

2) специфичные для конкретного гена либо группы генов и выполняющиеся в каждом конкретном случае в определенном порядке.

Например, подготовка корового промотора к взаимодействию с машиной может осуществляться различными механизмами, комбинации которых обеспечивают разнообразие паттернов экспрессии генов, специфичных для определенной стадии развития организма, ткани либо типа клеток (Blanchette *et al.*, 2006). Для гена интерферона β человека этот этап включает 6 стадий (рис. 5) (Agalioti *et al*., 2000).

Раздел A-box базы знаний включает также данные по транскрипционным факторам, транскрипционным кофакторам и белкам, участвующим в ремоделировании хроматина (табл. 1). В настоящее время в базе знаний содержатся сведения по транскрипционным регуляторам трех видов млекопитающих (человека, мыши, крысы) (Ignatieva, 2012).

Словарь по молекулярным процессам, обеспечивающим регуляцию транскрипции, составлен на основе анализа терминов из системы Gene Ontology (раздел «biological_process»), а также анализа научных публикаций и включает около 40 терминов, распределенных по 4 иерархическим уровням. Например, термин, обозначающий «регуляцию транскрипции путем реорганизации хроматина», является одним из понятий верхнего уровня иерархии. Термины следующих двух уровней иерархии представляют более детальное описание возможных механизмов реализации процесса (рис. 6).

Представление и использование знаний о механизмах регуляции транскрипции

В нашей онтологии механизм регуляции транскрипции является классом, который соответствует потенциально всем возможным вариантам реализации процессов регуляции. В зависимости от контекста, предусловия и источника реализуется конкретный вариант процесса.

На рис. 7 представлена упрощенная схема зависимостей между компонентами базы знаний T-box и A-box, включающая описание понятий Process, Event, Role, Type_of_object, конкретные реализации (экземпляры): process, event, object соответственно.

Процесс регуляции транскрипции является комплексным процессом, который состоит из множества взаимосвязанных подпроцессов и/или элементарных событий.

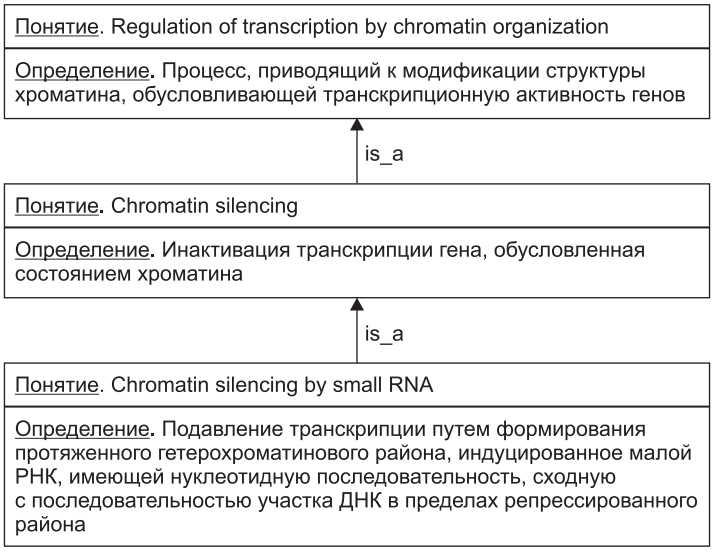

Рис. 6. Группа иерархически подчиненных терминов из словаря по молекулярным процессам, обеспечивающим регуляцию транскрипции.

Для описания механизма регуляции транскрипции необходимо выделить основные подпроцессы и регуляторные события, из которых складывается это биологическое явление; описать основных участников этих процессов и их ролевые функции.

Событие считается неделимым, и все процессы описываются с точностью до события. Для каждого класса события Event указывается набор ролей объектов (Role), которые необходимы для реализации этого события или могут участвовать в нем.

За основу описания понятия «процесс» нами взяты стандартные ситуативные роли Дж. Совы (http://www.jfsowa.com/ontology/) и работа N. Baumgartner с соавт. (2006). В описание понятия «процесс» входят следующие компоненты.

– **Контекст**. Какая клеточная ситуация (состояние системы) может обусловливать реализацию данного механизма? Где (гены, виды, органы, ткани, типы клеток и т. д.), когда (стадии клеточного цикла, стадии развития и т. д.) может возникать такая клеточная ситуация.

– **Предусловие** – условия, необходимые для старта процесса, реализующего механизм: наличие участников процесса, внешние сигналы и т. д.

– **Сценарий процесса** – сеть взаимосвязанных событий с частичным порядком по времени. Фиксированный источник порождает дерево событий с частичным порядком. Корнем этого дерева будет **Источник**.

– **Инициатор** – детерминирующий участник (участник, определяющий направление процесса или цель).

– **Источник** – внешний сигнал, запуск процесса, начало процесса (должен присутствовать в начале процесса, но не обязан принимать участие во всем процессе).

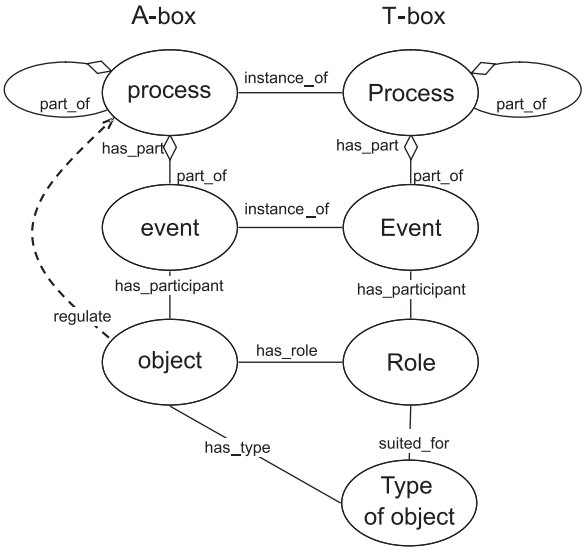

Рис. 7. Схема зависимостей между компонентами базы знаний.

Таблица 2

Формальное описание механизма активации эукариотического гена на примере интерферона β человека

1. Характеристика объекта		
Ген (G_k)	Вид организма	Клеточная ситуация (C_i)
IFNB1	Homo sapiens	Virus-infected HeLa cells

2. Характеристика стадий транскрипции и событий		
Этапы транскрипции (S_i)	Стадии и события (E_i)	Регуляторные белки (комплексы) (S_i, K_i), функционирующие на данной стадии
S_1: Подготовка корового промотора к контакту с компонентами ПИК	E_1 – сборка энхансосомы	Транскрипционные факторы ATF2, NF-KB, IRF1, HMGI(Y)
	E_2 – ацетилирование гистонов с участием комплекса GCN5	Комплекс GCN5
	E_3 – привлечение комплекса CBP/ Pol II	Комплекс CBP/ Pol II
	E_4 – привлечение комплекса SWI/SNF	Комплекс SWI/SNF
	E_5 – ремоделирование хроматина (нуклеосомной укладки) с участием хроматин-ремоделирующей белковой машины SWI/SNF	Комплекс SWI/SNF
	E_6 – привлечение белка TFIID	TFIID

3. Выборочная характеристика функциональных ролей регуляторных белков – участников определенной стадии транскрипции (на примере двух белков)

Регуляторный белок (P_i)	Функциональный класс	Стадия процесса транскрипции	Событие	Функциональная роль регуляторного белка (R_i)
HMG(Y)	Транскрипционный фактор	S_1	E_1 – сборка энхансосомы	Связывание с ДНК и белок-белковые взаимоедйствия в пределах энхансосомы
Комплекс GCN5	Корегулятор транскрипции	S_1	E_2 – ацетилирование гистонов с участием комплекса GCN5	Ацетилирование гистонов

− **Продукт** – продукт (может появляться в конце процесса, но не обязан принимать участие во всем процессе).

− **Постусловие** – условие окончания процесса.

Описание элементарного события (Event) включает описание ролей (Role) участников этого события, которые могут иметь определенный тип (Type of object). Например, множество участников события типа (класса) «Метаболическая реакция» включает типы объектов или роли: субстраты, продукты, ферменты, коферменты и регуляторы. Конкретная реакция (реализация или экземпляр класса) идентифицируется набором конкретных субстратов, продуктов и ферментов. Регуляторы реакции влияют на скорость реакции, т. е. меняют значения параметров этого события.

Иерархические классификаторы и сведения о транскрипционных регуляторах, накопленные в базе знаний, обеспечивают возможность формализованного описания механизмов регуляции транскрипции.

Например, механизм активации гена интерферона β человека, реконструированный на основе экспериментальных данных в работе Agalioti с соавт. (2000), может быть представлен в виде стадий регуляции транскрипции и набора регуляторных событий (табл. 2). Такое описание включает:

– характеристику объекта и клеточной ситуации;

– характеристику этапов (стадий) процесса транскрипции, а также регуляторных событий с указанием их участников (регуляторных белков);

– характеристику функциональных ролей регуляторных белков, участвующих в регуляторных событиях на конкретном этапе.

Для данных, представленных в формате OWL/RDF, можно сделать запрос на языке SPARQL (SPARQL Query Language for RDF, 1998).

```
SELECT ?stadia ?event ?role ?objectType ?objectName
WHERE {
      ?gene rdf:has_type gene.
      ?gene rdf:species ?species.
      ?gene rdf:name ?geneName.
      ?process rdf:has_type «regulation of transcription».
      ?process rdf:has_context ?context.
      ?context rdf:gene ?gene.
      ?process rdf:has_part ?stadia.
      ?stadia  rdf:has_part ?event.
      ?event   rdf:has_participant ?object.
      ?object  rdf:has_type ?objectType.
      ?object  rdf:has_role ?role.
      ?object  rdf:has_name ?objectName.
       FILTER (?gene = «IFN beta», ?species = «human»)
      }
```

В результате запроса будет выдана таблица, включающая список всех стадий регуляции транскрипции гена интерферона β человека, с указанием всех событий, которые происходят на каждой стадии, участников этих событий с указанием роли, типа и имени объекта.

Знания, полученные из гетерогенных источников, могут быть неполными, фрагментарными, 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 тоже могут быть неполными. Например, не все субъединицы комплекса известны, или неизвестно, сколько всего субъединиц входит в комплекс.

В некоторых случаях имеется возможность генерации правдоподобных гипотез, которые не противоречат известным фактам. Такого рода гипотетические знания с указанием относительного уровня достоверности полезны при дальнейшем анализе и построении непротиворечивых знаний (Ponomaryov *et al*., 2011).

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, обладающие способностью регулировать транскрипцию, с большой вероятностью могут быть транскрипционными факторами.

Примером косвенных знаний могут быть знания о взаимодействии между субъединицами белков, участвующих в регуляции транскрипции. Эти знания дают основание предположить, что участие обоих этих субъединичных белков в регуляции транскрипции может осуществляться через образование транскрипционного комплекса, в который входят оба белка.

Ниже приводится пример логического вывода новых знаний о регуляции транскрипции на основании фактов о связывании белков и образовании белкового комплекса и участии их в регуляции транскрипции.

∀ proteinA, proteinB
// 1. Белок proteinA участвует в регуляции транскрипции через образование
// неизвестного регуляторного комплекса proteinX
$\exists e_1$: e_1 ***instance_of*** TranscriptionRegulationProcess and e_1 ***has_participant*** proteinX,
∃proteinX : proteinA ***part_of*** proteinX,

// 2. Белок proteinB участвует в регуляции транскрипции через образование
// неизвестного регуляторного комплекса proteinY
$\exists e_1 : e_2$ ***instance_of*** TranscriptionRegulationProcess,
$\exists$proteinY : e_2 ***has_participant*** proteinY and proteinB ***part_of*** proteinY,
// 3. Белок proteinA связывается с белком proteinB, образуя белковый комплекс ProteinAB.
proteinA ***part_of*** proteinAB & proteinB ***part_of*** proteinAB
Вывод (гипотеза):
proteinX ≡ proteinY ≡ proteinAB & $e_1 \equiv e_1 \equiv e$ & e ***has_participant*** proteinAB.

Это предположение становится более правдоподобным, если известно, что действие этих белков на транскрипцию одинаково (подавление либо усиление транскрипции). Это позволяет задать частичный порядок на множестве гипотез по уровням относительной достоверности.

В ряде случаев можно распространять свойства через мереологические иерархии (часть–целое). 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. Безусловно, это предположение может рассматриваться только как гипотеза, и только экспериментальная проверка может подтвердить этот факт.

Анализ реализаций механизмов регуляции транскрипции для конкретных генов в конкретной клеточной ситуации дает информацию о возможных этапах регуляции транскрипции, множестве возможных элементарных событий, составляющих их, и типах участников событий.

Аналогичные этапы могут присутствовать в регуляции транскрипции других генов. Можно предположить, что в реализации элементарных событий, возможно, будут участвовать другие объекты, но того же типа, т. е. играть ту же роль в процессе. Поэтому рассуждение по аналогии позволяет делать выводы о возможных вариантах реализации механизма регуляции конкретного гена, перебирая на роль участника процесса все объекты соответствующего типа, которые могут присутствовать в данной клеточной ситуации.

Предположим, что нам не известен механизм регуляции транскрипции некоторого гена G, т. е. не известны этапы, стадии регуляции транскрипции, составляющие их события, и участники этих событий. Однако известно, что белок P участвует в регуляции транскрипции этого гена. Тогда возможен следующий вариант логического анализа и вывода гипотез о возможных механизмах регуляции транскрипции для этого гена:

Шаг 1. Анализ ролей $\{R_i\}$ белка P, которые этот белок имел (has_role) в регуляторных событиях (Event).

Шаг 2. Выделение классов событий $\{E_i\}$, для которых необходимы участники с ролями $\{R_i\}$.

Шаг 3. Выявление других ролей $\{Rs_k\}$, которые необходимы для реализации этих классов событий $\{E_i\}$.

Шаг 4. Поиск в базе знаний других объектов $\{Ps_i\}$, участников событий этих классов $\{E_i\}$ и анализ возможности участия этих объектов в регуляции гена G.

Шаг 5. Выявление класса процессов (этапов и стадий регуляции транскрипции), которые включают (part_of) эти события.

Шаг 6. Реконструкция гипотетического механизма регуляции транскрипции с типами стадий регуляции транскрипции, которые включают события $\{E_i\}$ с участниками $\{Ps_i\}$, выполняющими роли $\{Rs_k\}$.

Если роли объекта не известны, то для их предсказания возможно использование знаний о типе объекта и его свойствах. Предполагается, что класс Type_of_object включает описание типов объектов и их свойств, которые важны для предсказания возможности выполнения определенных ролей.

Необходимо отметить, что большинство этапов логического анализа может быть выполнено путем запроса к базе знаний.

ЗАКЛЮЧЕНИЕ

Разработаны подходы к построению онтологии регуляции транскрипции. На основе разработанной онтологической модели создана

пилотная версия базы знаний по регуляции транскрипции генов эукариот, включающей знания (T-Box) об онтологических понятиях и их взаимосвязях в области регуляции транскрипции (иерархически организованные словари (классификаторы) белков, регуляторов транскрипции, этапов транскрипции, молекулярных механизмов) и базу данных (A-box) по регуляторам транскрипции человека, мыши и крысы. На примере описания этапов и стадий активации конкретного эукариотического гена продемонстрирована применимость онтологической модели и базы знаний для формализованного описания механизмов регуляции транскрипции генов и построения гипотез о механизмах регуляции транскрипции генов эукариот с привлечением информации из базы знаний. Таким образом, предложенные подходы могут использоваться при реконструкции гипотетических механизмов регуляции транскрипции с учетом информации о строении регуляторных районов генов и функциях регуляторных белков, присутствующих в заданных клетках или тканях на определенной стадии развития.

В дальнейшем нами планируется развитие базы знаний по регуляции транскрипции генов эукариот с целью ее использования для интерпретации закономерностей строения регуляторных районов коэкспрессирующихся генов, функциональной интерпретации микрочиповых и протеомных данных, отражающих уровни экспрессии генов, и выявления регуляторных составляющих генных сетей, контролирующих фенотипические признаки организма.

БЛАГОДАРНОСТИ

Работа выполнена при частичной поддержке Президиума РАН (проекты А.II.6.8, 30.29), СО РАН (проект фундаментальных исследований VI.50.1.2. «Биоинформатика и системная биология молекулярно-генетических систем и процессов»), Совета по грантам Президента Российской Федерации (НШ-5278.2012.4).

ЛИТЕРАТУРА

Подколодный Н.Л. Онтологическое моделирование в биоинформатике и системной биологии // Онтологическое моделирование. ИПИ РАН, 2011. С. 233–269.

Подколодный Н.Л., Игнатьева Е.В., Рассказов Д.А. и др. Интегрированная система для информационной поддержки исследования механизмов регуляции транскрипции // Тр. 12-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010. Казань, Россия, 2010. С. 69–75.

Разин С.В. Хроматин и регуляция транскрипции // Молекуляр. биология. 2007. Т. 41. № 3. С. 387–394.

Agalioti T., Lomvardas S., Parekh B. *et al*. Ordered recruitment of chromatin modifying and general transcription factors to the IFN-b promoter // Cell. 2000. V. 103. P. 667–678.

Baumgartner N., Retschitzegger W. A survey of upper ontologies for situation awareness // Proc. of the 4th IASTED Intern. Conf. on Knowledge Sharing and Collaborative Engineering, St. Thomas, US VI. 2006. P. 1–9.

Beisswanger E., Lee V., Kim Jung J. Gene regulation ontology (GRO): design principles and use cases // II Proc. 21st. Intern. Congr. of the Europ. Federation for Med. Inform. (MIE 2008). 2008. P. 9–14.

Berger S.L. The complex language of chromatin regulation during transcription // Nature. 2007. V. 447. No. 7143. P. 407–412.

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.

Carey M., Smale S.T. Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. 2000. 639 p.

Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements // Nucl. Acids Res. 2010. V. 38. P. D331–335.

Gene Regulatory aontology (GRO), version 0.5, 1.09.2011 – http://bioportal.bioontology.org/ontologies/1106.

Hahn S. Structure and mechanism of the RNA Polymerase II transcription machinery // Nat. Struct. Mol. Biol. 2004. V. 11. No. 5. P. 394–403.

Ignatieva E.V. TrDB: a database of the human, mouse, and rat transcriptional regulators and its potential applications in systems biology // The Eighth Intern. Conf. on Bioinformatics of Genome Regulation and Structure / Systems Biology (BGRS/SB'12). Novosibirsk, Russia, June 25–29. 2012. P. 125.

Kolchanov N.A., Ignatieva E.V., Ananko E.A. *et al*. Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucl. Acids Res. 2002. V. 30. No. 1. P. 312–317.

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.

Lemon B., Tjian R. Orchestrated response: a symphony of transcription factors for gene control // Genes Dev. 2000. V. 14. No. 20. P. 2551–2569.

Özgövde A., Grüninger M. Foundational process relations in bio-ontologies // Proc. of the Sixth Intern. Conf. on Formal Ontology in Information Systems (FOIS 2010). IOS Press Amsterdam, The Netherlands, 2010. P. 243–256.

Podkolodnyy N.L., Nechkin S.S., Ignatieva E.V. *et al*. A database for analysis of the organizational features 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.

Ponomaryov D., Omelianchuk N., Mironova V. *et al*. From published expression and phenotype data to structured knowledge: The Arabidopsis gene net supplementary database and its applications // Lecture Notes in Artificial Intelligence. 2011. P. 101–120.

Schaefer U., Schmeier S., Bajic V.B. TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins // Nucl. Acids Res. 2011. V. 39. P. D106–D110.

Schober D., Smith B., Lewis S. *et al*. Survey-based naming conventions for use in OBO foundry ontology development // BMC Bioinformatics. 2009. 10(125). P. 1–9.

Shipra A., Chetan K., Rao M.R.S. CREMOFAC – a database of chromatin remodeling factors // Bioinformatics. 2006. V. 22. No. 23. P. 2940–2944.

Smith B., Ashburner M., Rosse C. *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration // Nat. Biotech. 2007. 25(11). P. 1251–1255.

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

SPARQL Query Language for RDF. 1998 – http://www.w3.org/TR/rdf-sparql-query/

INFORMATION SUPPORT OF RESEARCH ON TRANSCRIPTIONAL REGULATORY MECHANISMS: AN ONTOLOGICAL APPROACH

N.L. Podkolodnyy[1, 2], E.V. Ignatieva[1], O.A. Podkolodnaya[1], N.A. Kolchanov[1, 3, 4]

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

Summary

By now, a huge body of experimental data on gene transcription regulation has been accumulated. Transcription is controlled by a great number of proteins acting at various steps of the process; thus, a diversity of regulatory mechanisms can be realized. This paper presents approaches to building knowledge domain ontology, formalized description of the mechanisms of transcriptional regulation and the development of methods for integration of heterogeneous information on the features of the regulation of gene expression on this base. The pilot version of the knowledge base on the transcriptional regulation of eukaryotic genes includes: (1) description of basic terms related to transcription regulation and relationships between them; (2) hierarchical classification of transcription regulators; (3) classification of phases and steps of transcription; (4) a database of transcriptional regulators of three mammalian species (human, mouse, and rat); and (5) dictionaries for molecular processes involved in transcriptional regulation. The knowledge base is designed for information support of computer analysis of transcriptional regulatory mechanisms. Approaches to reconstruction of eukaryotic transcriptional regulatory mechanisms with the new knowledge base are presented.

Key words: bioinformatics, transcription regulation, knowledge base systems, ontology.