

Н. Л. Подколотный^{1,2}, Д. А. Гаврилов³, Н. Н. Твердохлеб^{1,3}, О. А. Подколотная¹

¹ *Институт цитологии и генетики СО РАН
пр. Академика Лаврентьева, 10, Новосибирск, 630090, Россия*

² *Институт вычислительной математики и математической геофизики СО РАН
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия*

³ *Новосибирский государственный университет
ул. Пирогова, 1, Новосибирск, 630090, Россия*

pnl@bionet.nsc.ru

СYTOSCAPE – ПЛАГИН ДЛЯ ПОСТРОЕНИЯ СТРУКТУРНЫХ МОДЕЛЕЙ БИОЛОГИЧЕСКИХ СЕТЕЙ В ВИДЕ СЛУЧАЙНЫХ ГРАФОВ*

Современные методы экспериментальных исследований позволяют реконструировать различного типа биологические сети, включая генные и метаболические сети, сети интерактомики, сети коэкспрессии генов, сети заболеваний и т. д. В данной статье представлена разработанная нами система построения структурных моделей биологических сетей в виде набора случайных графов, структурные закономерности которых совпадают со структурными закономерностями исходной биологической сети. Такие структурные модели могут быть использованы для проверки различных статистических гипотез на сетях, в исследовании влияния структурных закономерностей в биологических сетях на их функцию и других задачах.

При генерации структурных моделей в случайных графах могут быть зафиксированы следующие характеристики: распределение степеней вершин, попарное распределение степеней вершин, средняя степень соседних вершин, коэффициент кластеризации, спектр кластеризации, частота структурных мотивов различных размеров и др. Разработанная система построена по архитектуре клиент-сервер и состоит из плагина Cytoscape и удаленного вычислительного сервиса. Взаимодействие между клиентом и сервером реализовано посредством фреймворка gRPC с применением протокола сериализации структурированных данных Protocol Buffers. Система позволяет асинхронно конструировать структурные модели заданных биологических сетей в виде случайных графов посредством программ Random Network Generator и GTrie Scanner. Результат построения может быть загружен для визуализации и анализа средствами пакета Cytoscape. С использованием разработанной системы проведен вычислительный эксперимент по реконструкции структурных моделей ряда биологических сетей, для которых удалось построить алгоритм предсказания времени расчетов структурных моделей.

Ключевые слова: биологические сети, случайные графы, структурные модели, Cytoscape.

Введение

Современные методы экспериментальных исследований в молекулярной биологии позволяют реконструировать различного типа биологические сети, включая генные сети, сети метаболических и сигнальных путей, сети интерактомики (сети взаимодействий белков, РНК,

* Разработка программного обеспечения выполнена в рамках государственного задания ИВМиМГ СО РАН по проекту № 0315-2016-0005. Подготовка данных для вычислительного эксперимента и расчеты с использованием вычислительных ресурсов ЦКП «Биоинформатика» проведены при поддержке бюджетного проекта № 0324-2018-0017.

ДНК), сети коэкспрессии генов, сети ассоциации заболеваний и генетических мутаций, сети кровеносных и лимфатических сосудов, сети контактов в макромолекулах, ассоциативные и семантические сети и др. [1–11]. Сетевое представление сложных биологических систем и процессов – достаточно продуктивный подход, который в настоящее время получает широкое распространение. Активно развиваются новые методы анализа сетей для решения различных задач, включая выявление ключевых элементов, структурных мотивов и кластеров в биологических сетях, приоритизацию на этой основе генов, белков, мутаций, поиск биомаркеров заболеваний и т. д. [2; 4; 6; 11–13].

Для проверки различных статистических гипотез на сетях в качестве «нулевых гипотез», как правило, используются различные модели случайных графов, которые могут быть описаны распределением вероятностей на графах или случайным процессом, который генерирует случайные графы.

Модели случайных графов, заданные структурные характеристики которых совпадают со структурными характеристиками анализируемой реальной биологической сети, можно рассматривать как структурные модели этой биологической сети. Исследование структурных моделей реальных биологических сетей позволяет выявлять внутренние зависимости между характеристиками сети. Для выявления таких зависимостей осуществляется генерация большого числа случайных графов, имеющих некоторую характеристику Y , и проверка гипотезы, что с большой вероятностью эти графы также имеют другую характеристику X .

Структурные модели позволяют также исследовать, от каких характеристик зависит целевой или функциональный признак биологической сети. Все структурные модели являются эквивалентными с точностью до структурных характеристик, по которым построены эти модели. Если имеется возможность оценки целевого или функционального признака биологической сети по ее структуре, то анализ структурных моделей позволяет выявить структурные закономерности в биологической сети, ответственные за проявление ее функций или целевых свойств. Например, сеть капиллярных сосудов характеризуется большой сложностью и может быть описана сетью, от структуры которой зависит кровоток [9; 10]. Другим примером такого рода целевых свойств могут быть робастность функционирования биологической сети (сохранение функций при удалении вершины сети или ребра) [14; 15].

Структурно-функциональные закономерности организации генной сети и сети белок-белковых взаимодействий могут быть важны для выделения главных компонент молекулярной системы и использоваться для построения математических моделей генных сетей или молекулярно-генетических систем.

Целью настоящей работы является разработка плагина-приложения системы Cytoscape для реконструкции различного уровня приближения структурных моделей исследуемых биологических сетей в виде случайных графов. Мы сосредоточились на одном из типов биологических сетей, описываемых неориентированным графом. К такого рода биологическим сетям относятся сети интерактомики, в частности сети белок-белковых взаимодействий (protein interaction network – PIN). Исследования в этой области привели к осознанию того, что топологические особенности сетей молекулярных взаимодействий могут дать новые знания об их природе и функциональных характеристиках, участвующих в этих взаимодействиях биологических макромолекул.

В статье представлены методы генерации структурных моделей сетей в виде случайных графов и описание распределенной системы генерации структурных моделей, реализованной с использованием клиент-серверной архитектуры (клиент-плагин Cytoscape и вычислительный сервис для реконструкции структурных моделей в виде случайных графов). В заключение представлены результаты вычислительного эксперимента по реконструкции структурных моделей реальных биологических сетей.

Случайные графы и методы генерации моделей

Случайные графы

Случайный граф – это граф, являющийся результатом случайного выбора из некоторого множества графов (генеральной совокупности графов) в соответствии с заданным на этом

множестве вероятностным распределением. Случайные графы можно описать как распределением вероятности на множестве графов, так и случайным процессом, создающим эти графы. Наиболее простыми и распространенными моделями случайных графов являются модели Эрдёша – Реньи [16] и Гилберта [17].

Для определения модели случайных графов Эрдёша и Реньи рассмотрим вероятностное пространство $G_{n,m}$ всех $\binom{N}{M}$ графов, имеющих n вершин и M ребер, где $N = \binom{n}{2}$ – максимальное число всех возможных ребер между n вершинами. Пусть $G(n, M)$ – случайный элемент пространства $G_{n,m}$ (т. е. $G(n, M) \in G_{n,m}$). Будем считать, что все случайные элементы пространства $G_{n,m}$ равновероятны с $p = M / \binom{N}{1}$. Таким образом, заданное вероятностное пространство случайных графов имеет четко выраженную интерпретацию. Другая классическая модель случайных графов введена Гилбертом одновременно с моделью Эрдёша и Реньи в 1959 г. [17]. В модели Гилберта для определения вероятностного пространства графов $G_{n,p}$ задается матрица инцидентности графа в виде массива случайных переменных $\{X_{ij} : 1 \leq i < j \leq n\}$, для которых $\Pr(X_{ij} = 1) = p$ и $\Pr(X_{ij} = 0) = 1 - p$. Случайный граф $G(n, p)$, в котором вершины i и j связаны, если случайная величина $X_{ij} = 1$, является случайным элементом вероятностного пространства $G_{n,p}$ (т. е. $G(n, p) \in G_{n,m}$). Вероятность получения случайного графа, имеющего n вершин и M ребер, равна $p^M (1 - p)^{\binom{n}{2} - M}$.

Для $M \approx pN$, где N – максимальное число возможных ребер, эти две наиболее широко используемые модели $G(n, M)$ и $G(n, p)$ почти взаимозаменяемы [18]. Поэтому многие теоретические результаты получены с использованием модели Гилберта. Однако такие простые модели случайных графов пригодны в качестве «нулевых моделей» далеко не для всех статистических гипотез. Возможным обобщением модели случайных графов Эрдёша и Реньи могут быть модели, в которых вероятности случайных элементов вероятностного пространства $G_{n,m}$ различны. Например, часто возникает необходимость использовать модели случайных графов, сохраняющих распределение степеней вершин [19; 20] или другие структурные характеристики, наблюдаемые в реальной биологической сети. Такие модели случайных графов можно рассматривать как структурные модели биологической сети.

Методы рандомизации биологической сети с использованием Марковских цепей

Пусть задана биологическая сеть, которую можно представить в виде простого неориентированного графа $G = (V, E)$, где V – это конечное множество вершин, а E – совокупность пар (v_i, v_j) , где $v_i, v_j \in V$. При этом в графе G нет петель, кратных и ориентированных ребер.

В данной работе используются следующие численные характеристики графов.

- 1 $d(v)$ – степень вершины v или число ребер вершины v .
- 2 Распределение вероятности степеней вершин графа задается вектором f , где $f(d)$ – вероятность того, что случайно выбранная вершина в графе будет иметь степень d .
- 3 Совместное распределение вероятности степеней вершин определяется матрицей J_{mn} , где $J(i, j)$ – вероятность того, что случайно выбранное ребро объединяет вершины со степенью i и j .

4 Коэффициент кластеризации $C_i = \frac{2n_i}{d(v_i)(d(v_i)+1)}$, где n_i – число связей, соединяющих всех соседей вершины v_i , иными словами, C_i есть вероятность того, что два ближайших соседа этой вершины сами есть ближайшие соседи.

5 Коэффициент ассортативности $r = \frac{M \sum_{i \in E} j_i k_i - (\sum_{i \in E} (j_i + k_i)/2)^2}{M \sum_{i \in E} (j_i^2 + k_i^2)/2 - (\sum_{i \in E} (j_i + k_i)/2)^2}$ – это коэффициент корреляции Пирсона ($0 \leq r \leq 1$) между степенями связанных вершин, где j_i и k_i – степень вершин на концах i -го ребра ($i \in E$), M – число ребер в графе. Для ассортативных сетей ($r > 0$) вершины с большим числом ребер чаще связаны с такого же типа вершинами, а вершины с небольшим числом связей, в свою очередь, чаще связаны с вершинами с небольшой степенью, для дисассортативных сетей ($r < 0$) вершины с большим количеством связей чаще связаны с вершинами с небольшим количеством связей [21].

6 Спектр кластеризации $c(k)$ – вероятность того, что две вершины, соседних с вершиной степени k , будут соседями [22].

7 Структурный мотив – неслучайный подграф, частота появления которого в сети значимо выше, чем по случайным причинам [23; 24].

Одним из подходов для получения случайного графа является рандомизация графа путем перестановки вершин в случайно выбранных парах ребер. В реализованной нами программе RNGmotifs для генерации случайных графов, сохраняющих заданные структурные характеристики, наблюдаемые в исходной биологической сети, используется метод рандомизации сети путем парной перестановки вершин в 2-х случайно выбранных ребрах графа. Процесс рандомизации соответствует Марковской цепи, которая стартует с $G_0 = G$, $E_0 = E$.

Для рандомизации графа с сохранением распределения степеней вершин на каждом k шаге выполняется выбор равновероятно двух непересекающихся ребер графа (v_1, v_2) и (v_3, v_4) , где $v_1 \# v_3$, $v_1 \# v_4$, $v_2 \# v_3$ и $v_2 \# v_4$. Далее делается равновероятно одна замена из двух вариантов замен ребер в графе:

$$E_k \leftarrow E_{k-1} \cup \{(v_1, v_4), (v_3, v_2)\} / \{(v_1, v_2), (v_3, v_4)\}$$

или

$$E_k \leftarrow E_{k-1} \cup \{(v_1, v_3), (v_2, v_4)\} / \{(v_1, v_2), (v_3, v_4)\}.$$

Если (V, E_k) простой граф (отсутствие петель и параллельных ребер), то с заданной вероятностью p принимается $G_k = (V, E_k)$, иначе $G_k = G_{k-1}$, $E_k = E_{k-1}$.

В работе [25] показано, что если размер (n, M) двух графов и распределение степеней вершин совпадает, то один граф может быть трансформирован в другой с помощью конечного числа шагов данной Марковской цепи.

В процессе рандомизации кроме явно задаваемых критериев отбора графов (ограничений) используются разные алгоритмы реализации Марковских цепочек, которые обеспечивают получение равновероятно всех возможных графов $G(n, M) \in G_{n, m}$, которые сохраняют либо распределение степеней вершин, либо совместное распределение степеней вершин.

Для генерации случайных графов с заданными ограничениями можно задавать вероятность замены ребер, зависящую от функционала отклонения модели на очередном шаге Марковского процесса от исходной биологической сети по заданным структурным характеристикам, которые необходимо сохранить в модели, например: коэффициент кластеризации, спектр кластеризации, частота структурных мотивов и т. д.

Одним из наиболее распространенных является алгоритм, подобный оптимизации методом имитации отжига, когда вероятность отбора соответствует распределению $p = \exp\left(-\frac{W}{T}\right)$, где

W – функционал отклонения модели от исходной биологической сети, который можно интерпретировать как энергию системы при некоторой температуре T .

Например, в реализованной нами программе RNGmotifs для генерации случайных графов с частотой структурных мотивов, совпадающей с характеристиками исходной биологической

сети, используется $W = \sum_{i=1}^s \frac{(w_{i1} - w_{i0})^2}{s(w_{i1} + w_{i0})^2}$, где w_{i1} , w_{i0} – частоты наблюдения мотивов в био-

логической сети и модели соответственно, s – число мотивов в ограничении. Предполагается, что процесс протекает при постепенно понижающейся температуре $T(k)$ с увеличением шага k Марковской цепи. В этом случае в начале Марковского процесса используется более мягкий критерий отбора (принимаются варианты обмена рёбер в графе, которые могут привести к отклонению значения функционала качества), но далее критерий становится более жестким, что обеспечивает, в конечном счете, сходимость структурных характеристик модели к характеристикам исходного графа. Такой вариант управления сходимостью функционала качества обусловлен необходимостью выйти из возможного локального минимума функционала и обеспечить возможность достижимости всех графов, имеющих заданные характеристики.

Для отбора статистически значимых мотивов в качестве «нулевой» гипотезы могут быть использованы случайные графы, сохраняющие распределение степеней вершин. При другом подходе, предлагаемом нами, используется пошаговый алгоритм, когда для отбора мотивов степени $k+1$ в качестве нулевой гипотезы используются случайные графы, сохраняющие не только распределение степеней вершин, но и частоты неслучайных структурных мотивов размером k . В этом случае структурная модель строится последовательно, начиная от мотивов размером 3 и далее, пошагово включая в структурную модель мотивы большего размера.

Для генерации структурных моделей нами также используется подход, основанный на так называемых dk -сериях, реализованный в программной библиотеке Random Network Generator [26]. dk -серия является вложенной структурой: каждый следующий уровень $(d+1)k$ -распределения содержит тот же объем информации об исходном графе, что и dk -распределение, но при этом также предоставляет о нем дополнительные сведения посредством включения в список ограничений более строгих правил генерации. Так, нулевой элемент последовательности, $0k$ -распределение, фиксирует самую грубую из возможных характеристик графа – среднюю степень вершин, что дает наиболее слабое из правил генерации. Следующий элемент, $1k$ -распределение, сохраняет распределение степеней вершин – более строгое условие, чем сохранение средней степени вершин. $2k$ -распределение сохраняет уже совместное распределение степеней вершин, т. е. число подграфов размером 2 – другими словами, рёбер – между вершинами со степенями k_1 и k_2 . Таким образом, $2k$ -распределение обозначает попарную корреляцию степеней вершин, а также коэффициент ассортативности графа. $3k$ -распределение сохраняет 3-совместное распределение степеней вершин, т. е. подграфов в виде замкнутого треугольника и клик, состоящих из трех вершин со степенями k_1 , k_2 и k_3 , что определяет коэффициент кластеризации, и т. д. [20; 26].

Здесь используются два варианта рандомизации графа: (1) описанный выше алгоритм рандомизации графа с сохранением распределения степеней вершин и (2) алгоритм рандомизации графа с сохранением совместного распределения степеней вершин.

В последнем варианте на каждом k шаге Марковского процесса выполняются следующие операции.

1. Выбираем равновероятно одно ребро графа (v_1, v_2) . Далее выбираем равновероятно одну вершину в этом ребре. Пусть это будет v_2 .

2. Выбираем равновероятно другую вершину графа v_3 такая, что $d(v_2) = d(v_3)$. Далее выбирается равновероятно соседняя вершина v_4 , инцидентная v_3 . Это дает второе ребро (v_3, v_4) .

3. Замена рёбер в графе $E_k \leftarrow E_{k-1} \cup \{(v_1, v_3), (v_2, v_4)\} / \{(v_1, v_2), (v_3, v_4)\}$.

4. Если (V, E_k) – простой граф, то с заданной вероятностью p принимается $G_k = (V, E_k)$, иначе $G_k = G_{k-1}$, $E_k = E_{k-1}$.

Такая процедура позволяет выбирать равномерно случайно все возможные элементы $G(n, M) \in G_{n,m}$, имеющие заданное распределение степеней рёбер и совместное распределение степеней вершин [27].

Для определения критерия остановки используют различные подходы. Однако теоретические оценки дают слишком большие значения числа шагов, что практически трудно реализуемо для больших графов. В работе [28] предложен более практичный критерий остановки, который находится на основе сходимости распределения вероятностей ряда структурных характеристик графа (глобальный коэффициент кластеризации, диаметр графа, максимальное собственное значение Лапласиана графа) к стационарному распределению, т. е. если распределение структурных характеристик графа практически не меняется при достижении некотором числа шагов, то процесс можно останавливать. В результате рекомендовано использовать число шагов Марковского процесса $N = \alpha * M$, где M – число рёбер в графе, а $5 < \alpha < 30$.

Говоря о возможностях генерации случайных графов в целом, следует отметить, что Random Network Generator поддерживает работу с dk -серией вплоть до уровня $2k$, включая его расширения в виде $2.1k$ и $2.5k$ [26]. Эти уровни являются частными расширениями $2k$, полученными в результате добавления правил генерации в виде сохранения коэффициента кластеризации и спектра кластеризации исходного графа соответственно. Таким образом, Random Network Generator может применяться для пошаговой реконструкции структурных моделей согласно dk -серии.

Программная реализация системы

Система генерации структурных моделей построена по архитектуре «клиент-сервер», где в качестве клиента выступает плагин-приложение системы визуализации графов Cytoscape [29–31], а в качестве сервера – удаленный вычислительный сервис с модульным подключением программных средств построения структурных моделей.

Система Cytoscape¹ – это программная платформа с открытым исходным кодом для визуализации и анализа сложных биологических сетей с возможностью интеграции разного типа биоинформационных данных, таких как функциональная аннотация генов, уровень экспрессии генов и пр. Важной особенностью программной архитектуры Cytoscape является технология расширения функциональности системы путем подключения дополнительных модулей (плагинов), созданных на языке Java сторонними разработчиками. Большая часть плагинов (более 300) доступна на Cytoscape App Store², и их можно устанавливать с помощью менеджера приложений App Manager [32].

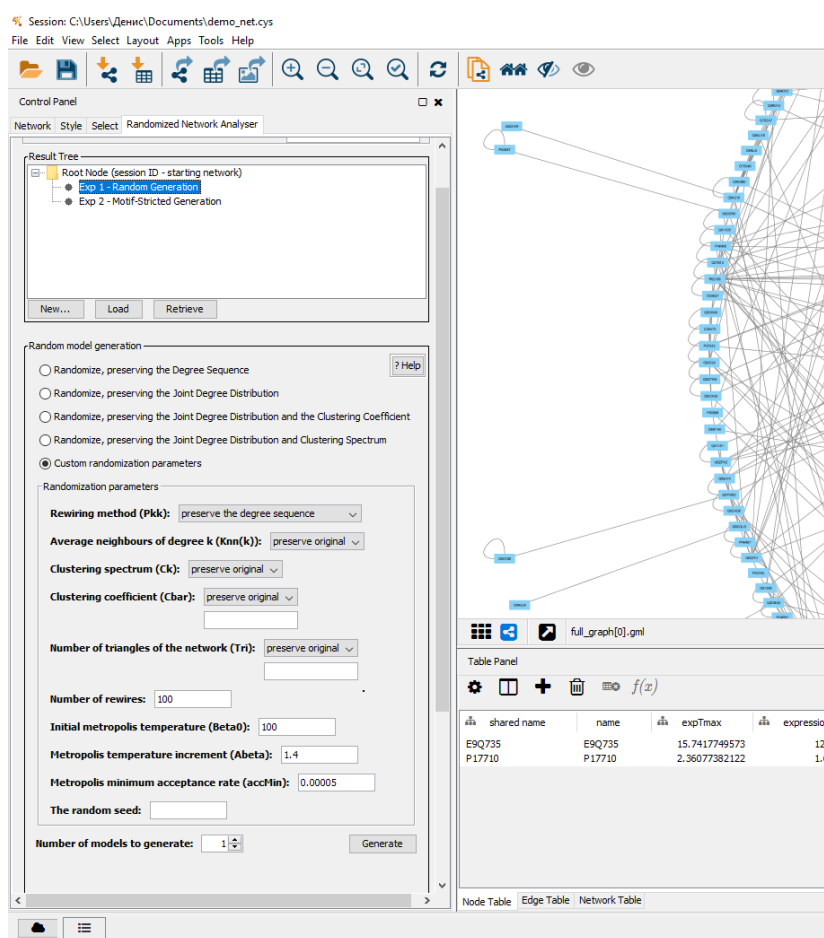
Большая часть этих плагинов Cytoscape разработаны для решения различных задач биоинформатики: в частности, загрузка биологических сетей из доступных баз данных, реконструкция биологических сетей на основе интеграции гетерогенной информации, визуализация, сравнение и анализ сетей, выявление функциональных модулей в сети, генерация атрибутов узлов сети с использованием доступных баз биологических данных, например базы данных по экспрессии генов, GO (Gene Ontology) аннотации и т. д. Система Cytoscape при реализации проекта дает возможность использовать этот функционал как для получения биологических сетей, так и для анализа структурных моделей.

¹ <http://www.cytoscape.org/>.

² <http://apps.cytoscape.org/>.

В рамках разрабатываемой системы, плагин-приложение является «тонким» клиентом, т. е. весь функционал по обработке загружаемых сетей ложится на удаленный вычислительный сервис. Плагин-приложение Cytoscape реализовано на языке Java, вычислительный сервис реализован на языке GoLang [33]. Взаимодействие между клиентом и сервером реализовано посредством фреймворка gRPC [34; 35] с применением протокола сериализации структурированных данных Protocol Buffers [36].

При разработке интерфейса использовались основные компоненты библиотеки Java Swing: JPanel, JScrollPane и JTree. В качестве основного элемента интерфейса используется древовидное представление текущей сессии работы с удаленным сервисом. В зависимости от выбранного пользователем элемента дерева, ему предлагается различный набор доступных опций. Например, при выборе узла верхнего уровня, соответствующего одному из доступных методов обработки исходной сети, пользователю предоставляется панель настройки параметров запускаемого алгоритма (см. рисунок).



Скриншот интерфейса системы реконструкции структурных моделей биологических сетей

Разработанная нами система позволяет асинхронно конструировать структурные модели заданных биологических сетей в виде случайных графов посредством программных библиотек Random Network Generator [26] и RNGmotifs, разработанной в ИЦиГ СО РАН на основе модификации пакета GTrie Scanner [37].

Пользовательский интерфейс плагина Cytoscape обеспечивает загрузку биологической сети для анализа, формирование запроса на удаленный вычислительный сервер для реконструкции различных структурных моделей в соответствии с их спецификацией, визуализацию реконструированных структурных моделей и их сравнительный анализ в пакете Cytoscape.

Вычислительный эксперимент: реконструкция и анализ структурных моделей биологических сетей

Целью вычислительного эксперимента является исследование того, как изменяется время вычислений при генерации различных структурных моделей в зависимости от структурных характеристик исходных биологических сетей.

В качестве исходных данных были выбраны следующие биологические сети:

1) сеть взаимодействия заболеваний человека ($N = 1\,419$, $M = 2\,738$), построенная на основе данных об известных ассоциациях «заболевание – ген» и указывает на общее генетическое происхождение многих заболеваний [38];

2) сеть белок-белковых взаимодействий у дрожжей ($N = 2\,361$, $M = 5\,375$) [39];

3) сеть белок-белковых взаимодействий в печени мыши для белков с циркадным изменением скорости трансляции ($N = 5\,753$, $M = 98\,813$) [40];

4) сеть белок-белковых взаимодействий в печени мыши для белков с циркадным изменением скорости трансляции и повышенной скоростью трансляции в начале суток ($T = 0$) ($N = 2\,702$, $M = 25\,893$) [40].

Сеть белок-белковых взаимодействий в печени мыши (PPI) была построена на основе базы данных IID (Integrated Interactions Database, версия от 2017-04) [41]. Далее было выбрано подмножество белков **Pc**, скорость трансляции которых в печени мыши имеет выраженное изменение в течение суток. С использованием этого подмножества взаимодействующих белков с циркадным изменением скорости трансляции была сформирована сеть белок-белковых взаимодействий № 3 (PIN_c) [40]. Если из белков **Pc**, гены которых имеют выраженную суточную динамику трансляции, выбрать для каждого момента времени суток подмножество белков со скоростью трансляции в это время больше, чем среднесуточное значение, то мы получаем множество сетей белок-белковых взаимодействий или динамическую сеть белок-белковых взаимодействий $PIN_c(T)$, зависящую от времени суток T [40]. Таким образом сформирована сеть № 4 $PIN_c(T_0)$, которая является подмножеством сети № 3. Исследование циркадных изменений динамической сети белок-белковых взаимодействий $PIN_c(T)$ имеет большое значение для выявления главных компонент структуры математической модели циркадного осциллятора.

Используя разработанный нами Cytoscape плагин и реконструированную сеть была проведена генерация структурных случайных моделей различного уровня точности: dk1.0 (сохранение распределения степеней вершин), dk2.0 (сохранение совместного распределения степеней вершин, т. е. распределения и корреляций степеней вершин), dk2.1 (сохранение совместного распределения степеней вершин и коэффициента кластеризации) и dk2.5 (сохранение совместного распределения степеней вершин и спектра кластеризации, т. е. вероятности того, что два узла, соседних с узлом степени k , будут соседями).

Расчеты выполнялись на рабочей станции HP Z800 Xeon 2x X5570QC 2.93ГГц. Сравнительные результаты времени расчетов указанных биологических сетей представлены в таблице.

Время вычислений реконструкции моделей биологических сетей (T/T^* , с)
и предсказание времени вычислений по структурным характеристикам сети

Модель	Качество предсказания времени расчета $R^2/p\text{-value}$	Сеть взаимодействия заболеваний человека	Сеть белок-белковых взаимодействий у дрожжей	Сеть белок-белковых взаимодействий в печени мыши	Сеть белок-белковых взаимодействий в печени мыши в начале суток
dk1.0	0,981/0,01	0,05/0,06	0,1/0,075	1/1,38	6/4,83
dk2.0	0.981/0.004	0,3/0,4	1/0,6	150/213	3240/2580
dk2.1	1/0.008	11,6/11,4	75/75.4	10510/11030	227000/219240
dk2.5	0.996/0.002	10,9/13,1	27/20,2	5879/8067	130000/105160

Как показывают результаты вычислительного эксперимента, скорость сходимости может сильно различаться даже на сетях с сопоставимым количеством вершин и рёбер.

Для исследования зависимости времени расчетов от структуры сети нами были оценены различные характеристики сети: размеры сети, распределение степеней вершин, коэффициенты кластеризации и ассортативности. Анализ показал, что распределение степеней вершин $f(d)$ во всех использованных в вычислительном эксперименте сетях имеет зависимость $f(d) = b * e^{-a*\sqrt{d}}$, где a и b – параметры распределения, которые были оценены для каждой сети с коэффициентом детерминации от 0,87 до 0,92 и уровнем значимости $p\text{-value} < 10^{-22}$.

Далее для предсказания времени вычисления каждой модели мы использовали метод пошаговой регрессии. Наиболее информативным показателем, учитывающим влияние структуры сети на логарифм времени вычислений, оказались характеристики распределения степеней вершин, в частности параметр a . В таблице представлены результаты предсказания времени расчета с указанием уровня достоверности и коэффициента детерминации (квадрат коэффициента корреляции (R^2) между предсказанным и истинным значением времени расчета).

Предсказание времени выполнения запросов для вычисления структурных моделей биологических сетей важно для планирования вычислительных экспериментов. Безусловно, полученные результаты можно использовать только для сетей, имеющих указанную зависимость распределения степеней вершин. Тем не менее предварительные результаты показали перспективность такого рода исследования для этого класса биологических сетей. При этом накопление вариантов проведенных расчетов моделей для различных биологических сетей такого рода могут автоматически использоваться для уточнения предсказания.

Заключение

Разработана система построения структурных моделей биологических сетей в виде набора случайных графов, структурные закономерности которых совпадают со структурными закономерностями исходной биологической сети. При генерации структурных моделей в случайных графах могут быть зафиксированы следующие характеристики: распределение степеней вершин, попарное распределение степеней вершин, средняя степень соседних вершин, коэффициент кластеризации, спектр кластеризации, частота заданных структурных мотивов различных размеров и т. д.

Система построена по архитектуре «клиент-сервер» и состоит из плагина-приложения Cytoscape и удаленного вычислительного сервиса. Взаимодействие между клиентом и сервером реализовано посредством фреймворка gRPC с применением протокола сериализации структурированных данных Protocol Buffers.

Система позволяет асинхронно конструировать структурные модели заданных биологических сетей в виде случайных графов посредством программных библиотек Random Network Generator и RNGmotifs, разработанной в ИЦиГ СО РАН на основе модификации пакета GTrie Scanner.

Пользовательский интерфейс плагина Cytoscape обеспечивает загрузку биологической сети для анализа, формирование запроса на удаленный вычислительный сервер для реконструкции различных структурных моделей в соответствии с их спецификацией, визуализацию реконструированных структурных моделей и их сравнительный анализ в пакете Cytoscape.

С использованием разработанной системы проведен вычислительный эксперимент по реконструкции структурных моделей ряда биологических сетей, для которых удалось построить алгоритм предсказания времени расчетов структурных моделей.

Список литературы

1. Alm E., Arkin A. P. Biological networks // Current opinion in structural biology. 2003. Vol. 13. No. 2. P. 193–202.

2. Gosak M., Markovič R., Dolenšek J., Slak Rupnik M., Marhl M., Stožer A., Perc M. Network science of biological systems at different scales: A review // *Physics of Life Reviews*. 2018. Vol. 24. P. 118–135.
3. Ananko E. A., Podkolodnyy N. L., Stepanenko I. L., Podkolodnaya O. A., Rasskazov D. A., Miginsky D. S., Likhoshvai V. A., Ratushny A. V., Podkolodnaya N. N., Kolchanov N. A. GeneNet in 2005 // *Nucleic Acids Res.* 2005. Vol. 33. P. D425–D427.
4. Vella D., Zoppis I., Mauri G., Mauri P., Di Silvestre D. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data // *EURASIP Journal on Bioinformatics and Systems Biology*. 2017. Vol. 6.
5. Stuart J. M., Segal E., Koller D., Kim S. K. A gene-coexpression network for global discovery of conserved genetic modules // *Science*. 2003. Vol. 302 (5643). P. 249–255.
6. Podkolodnaya O. A., Podkolodnaya N. N., Podkolodnyy N. L. The mammalian circadian clock: Gene regulatory network and computer analysis // *Russian Journal of Genetics: Applied Research*. 2015. Vol. 5. No. 4. P. 354–362.
7. Emamjomeh A., Robat E. S., Zahiri J., Solouki M., Khosravi P. Gene co-expression network reconstruction: a review on computational methods for inferring functional information from plant-based expression data // *Plant Biotechnol. Rep.* 2017. Vol. 11. P. 71–86.
8. Hu J. X., Thomas C. E., Brunak S. Network biology concepts in complex disease comorbidities // *Nature Reviews Genetics*. 2016. Vol. 17. P. 615–629.
9. Novkovic M., Onder L., Bocharov G., Ludewig B. Graph Theory-Based Analysis of the Lymph Node Fibroblastic Reticular Cell Network // *Methods Mol. Biol.* 2017. Vol. 1591. P. 43–57. DOI: 10.1007/978-1-4939-6931-9_4.
10. Grebennikov D., Loon R. van, Novkovic M., Onder L., Savinkov R., Sazonov I., Tretyakova R., Watson D. J., Bocharov G. Critical Issues in Modeling Lymph Node Physiology // *Computation*. 2017. Vol. 5 (3).
11. Ivanisenko V. A., Saik O. V., Ivanisenko N. V., Tiys E. S., Ivanisenko T. V., Demenkov P. S., Kolchanov N. A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology // *BMC Syst. Biol.* 2015. Vol. 9. Suppl. 2:S2. DOI: 10.1186/1752-0509-9-S2-S2.
12. Luo T., Wu S., Shen X., Li L. Network cluster analysis of protein-protein interaction network identified biomarker for early onset colorectal cancer // *Mol. Biol. Rep.* 2013. Vol. 40 (12). P. 6561–6568.
13. Yan W., Xue W., Chen J., Hu G. Biological Networks for Cancer Candidate Biomarkers Discovery // *Cancer Informatics*. 2016. Vol. 15 (S3). P. 1–7. DOI: 10.4137/CIN.S39458.
14. Albert R., Jeong H., Barabási A.-L. Error and attack tolerance of complex networks // *Nature*. 2000. Vol. 406. P. 378–382.
15. Truong C.-D., Kwon Y.-K. Investigation on changes of modularity and robustness by edge-removal mutations in signaling networks // *BMC Systems Biology*. 2017. Vol. 11 (Suppl. 7). P. 125.
16. Erdős P., Rényi A. On random graphs I // *Publications Mathematicae Debrecen*. 1959. Vol. 6. P. 290–297.
17. Gilbert E. N. Random graphs // *Annals of Mathematical Statistics*. 1959. Vol. 30. P. 1141–1144. DOI:10.1214/aoms/1177706098.
18. Bollobas B., Riordan O. M. Mathematical results on scale-free random graphs // *Handbook of Graphs and Networks*. 1st ed. Eds. S. Bornholdt, H. G. Schuster. Wiley VCH, Weinheim, 2003. P. 1–34.
19. Orsini C., Dankulov M. M., Colomer-de-Simón P., Jamakovic A., Mahadevan P., Vahdat A., Bassler K. E., Toroczkai Z., Boguñá M., Caldarelli G., Fortunato S., Krioukov D. Quantifying randomness in real networks // *Nature Communications*. 2015. Vol. 6. P. 8627.
20. Orsini C., Mitrovic D. M., Jamakovic A., Mahadevan P., Colomer-de-Simon P., Vahdat A., Bassler K., Toroczkai Z., Boguñá M., Caldarelli G., Fortunato S., Krioukov D. How random are complex networks // *CoRR*. 2015. abs/1505.07503.
21. Newman M. E. J. Assortative Mixing in Networks // *Phys. Rev. Lett.* 2002. Vol. 89. P. 208701.

22. Nobari S. et al. Fast random graph generation // Proc. of the 14th International conference on extending database technology. ACM, 2011. P. 331–342.
23. Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U. Network motifs: simple building blocks of complex networks // Science. 2002. Vol. 298. P. 824–827.
24. Jamakovic A. et al. How small are building blocks of complex networks // arXiv preprint arXiv:0908.1143. – 2009.
25. Taylor R. Constrained switching in graphs // SIAM J. Algebraic Discrete Meth. 1982. Vol. 3 (1). P. 115–121.
26. Simon P. C. de. RandNetGen: a Random Network Generator. URL: <http://polcolomer.github.io/RandNetGen/> (дата обращения 25.05.2018).
27. Stanton I., Pinar A. Constructing and sampling graphs with a prescribed joint degree distribution // ACM J. Exp. Algor. 2012. Vol. 17 (3). Article 3.5 (August 2012). 25 p.
28. Ray J., Pinar A., Seshadhri C. A stopping criterion for Markov chains when generating independent random graphs // Journal of Complex Networks. 2015. Vol. 3 (2). P. 204–220.
29. Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B., Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks // Genome Res. 2003. Nov. Vol. 13 (11). P. 2498–2504.
30. Killcoyne S. et al. Cytoscape: a community-based framework for network modeling // Protein Networks and Pathway Analysis. 2009. P. 219–239.
31. Saito R. et al. A travel guide to Cytoscape plugins // Nature Methods. 2012. Vol. 9. No. 11. P. 1069–1076.
32. Lotia S., Montojo J., Dong Y., Bader G. D., Pico A. R. Cytoscape app store // Bioinformatics. 2013. Vol. 29 (10). P. 1350–1351.
33. Донован А. А., Керниган Б. У. Язык программирования Go: Пер. с англ. М.: ИД Вильямс, 2016. 432 с.
34. Cerami E. Web services essentials: distributed applications with XML-RPC, SOAP, UDDI & WSDL. O'Reilly Media, Inc., 2002. 304 p.
35. Seymour K. et al. Overview of GridRPC: A remote procedure call API for grid computing // International Workshop on Grid Computing. Springer, Berlin, Heidelberg, 2002. P. 274–278.
36. Varda K. Protocol buffers: Google's data interchange format // Google Open Source Blog, Jul. 2008. URL: <https://opensource.googleblog.com/2008/07/protocol-buffers-googles-data.html/>. (дата обращения 20.07.2018)
37. Ribeiro P. gtrieScanner – Quick Discovery of Network Motifs // CRACS & INESC-TEC, DCC/FCUP. URL: <http://www.dcc.fc.up.pt/gtries/> (дата обращения 25.05.2018).
38. Goh K. I. et al. The human disease network // Proc. of the National Academy of Sciences. 2007. Vol. 104. No. 21. P. 8685–8690.
39. Bu D. et al. Topological structure analysis of the protein-protein interaction network in budding yeast // Nucleic acids research. 2003. Vol. 31. No. 9. P. 2443–2450.
40. Подкольный Н. Л., Твердохлеб Н. Н., Подкольная О. А. Анализ циркадного ритма биологических процессов в печени и почках мыши // Вавиловский журнал генетики и селекции. 2017. Т. 21, № 8. С. 903–910.
41. Kotlyar M. et al. Integrated Interactions Database: Tissue-specific view of the human and model organism interactomes // Nucleic Acids Res. URL: <http://dcv.uhnres.utoronto.ca/iid/> (дата обращения 20.07.2018).

N. L. Podkolodnyy^{1,2}, D. A. Gavrilov³, N. N. Tverdokhle^{1,3}, O. A. Podkolodnaya¹

¹ Institute of Cytology and Genetics SB RAS
10 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

² Institute of Computational Mathematics and Mathematical Geophysics SB RAS
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

³ Novosibirsk State University
1 Pirogov Str., Novosibirsk, 630090, Russian Federation

pnl@bionet.nsc.ru

CYTOSCAPE PLUGIN FOR RECONSTRUCTION OF STRUCTURAL RANDOM GRAPH MODELS OF BIOLOGICAL NETWORKS

Modern experimental technologies in molecular biology allow reconstructing different types of biological networks, including gene and metabolic networks, networks of interatomic, gene co-expression networks, a network of diseases, etc. This article presents the program tool for reconstructing structural random graph models of biological networks, the structural regularities of which coincide with the structural regularities of the initial biological network. Such structural models can be used to test various statistical hypotheses on networks, to study the influence of structural regularities in biological networks on their function, and so on. Our tool generate the structural random graph models with the following fixed characteristics: the distribution of vertex degrees, the joint distribution of degrees of vertices, the average degree of neighboring vertices, the clustering coefficient, the clustering spectrum, the frequency of structural motifs of various sizes, etc.

The developed system is based on the client-server architecture and consists of the Cytoscape plug-application and remote computing service. The interaction between the client and the server is implemented through the gRPC framework using the Protocol Buffers (structured data serialization protocol).

The system allows to construct the structural random graph models of the given biological networks asynchronously through software Random Network Generator and GTrie Scanner. The result structural model can be loaded for visualization and analysis using the Cytoscape package. This article also presents the computational experiment for reconstruct the structural random graph models of a number of biological networks. The algorithm for estimating the time of calculations of structural models of this kind of biological networks was constructed.

Keywords: biological networks, random graphs, structural models, Cytoscape.

References

1. Alm E., Arkin A. P. Biological networks. *Current opinion in structural biology*, 2003, vol. 13, no. 2, p. 193–202.
2. Gosak M., Markovič R., Dolenšek J., Slak Rupnik M., Marhl M., Stožer A., Perc M. Network science of biological systems at different scales: A review. *Physics of Life Reviews*, 2018, vol. 24, p. 118–135.
3. Ananko E. A., Podkolodnyy N. L., Stepanenko I. L., Podkolodnaya O. A., Rasskazov D. A., Miginsky D. S., Likhoshvai V. A., Ratushny A. V., Podkolodnaya N. N., Kolchanov N. A. GeneNet in 2005. *Nucleic Acids Res.*, 2005, vol. 33, p. D425–D427.
4. Vella D., Zoppis I., Mauri G., Mauri P., Di Silvestre D. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP Journal on Bioinformatics and Systems Biology*, 2017, vol. 6.
5. Stuart J. M., Segal E., Koller D., Kim S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 2003, vol. 302 (5643), p. 249–255.
6. Podkolodnaya O. A., Podkolodnaya N. N., Podkolodnyy N. L. The mammalian circadian clock: Gene regulatory network and computer analysis. *Russian Journal of Genetics: Applied Research*, 2015, vol. 5, no. 4, p. 354–362.

7. Emamjomeh A., Robat E. S., Zahiri J., Solouki M., Khosravi P. Gene co-expression network reconstruction: a review on computational methods for inferring functional information from plant-based expression data. *Plant Biotechnol. Rep.*, 2017, vol. 11, p. 71–86.
8. Hu J. X., Thomas C. E., Brunak S. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 2016, vol. 17, p. 615–629.
9. Novkovic M., Onder L., Bocharov G., Ludewig B. Graph Theory-Based Analysis of the Lymph Node Fibroblastic Reticular Cell Network. *Methods Mol. Biol.*, 2017, vol. 1591, p. 43–57. DOI: 10.1007/978-1-4939-6931-9_4.
10. Grebennikov D., Loon R. van, Novkovic M., Onder L., Savinkov R., Sazonov I., Tretyakova R., Watson D. J., Bocharov G. Critical Issues in Modeling Lymph Node Physiology. *Computation*, 2017, vol. 5 (3).
11. Ivanisenko V. A., Saik O. V., Ivanisenko N. V., Tiys E. S., Ivanisenko T. V., Demenkov P. S., Kolchanov N. A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst. Biol.*, 2015, vol. 9, suppl. 2:S2. DOI: 10.1186/1752-0509-9-S2-S2.
12. Luo T., Wu S., Shen X., Li L. Network cluster analysis of protein-protein interaction network identified biomarker for early onset colorectal cancer. *Mol. Biol. Rep.*, 2013, vol. 40 (12), p. 6561–6568.
13. Yan W., Xue W., Chen J., Hu G. Biological Networks for Cancer Candidate Biomarkers Discovery. *Cancer Informatics*, 2016, vol. 15 (S3), p. 1–7. DOI: 10.4137/CIN.S39458.
14. Albert R., Jeong H., Barabasi A.-L. Error and attack tolerance of complex networks. *Nature*, 2000, vol. 406, p. 378–382.
15. Truong C.-D., Kwon Y.-K. Investigation on changes of modularity and robustness by edge-removal mutations in signaling networks. *BMC Systems Biology*, 2017, vol. 11 (suppl. 7), p. 125.
16. Erdős P., Rényi A. On random graphs I. *Publications Mathematicae Debrecen*, 1959, vol. 6, p. 290–297.
17. Gilbert E. N. Random graphs. *Annals of Mathematical Statistics*, 1959, vol. 30, p. 1141–1144. DOI:10.1214/aoms/1177706098.
18. Bollobas B., Riordan O. M. Mathematical results on scale-free random graphs. In: *Handbook of Graphs and Networks*. 1st ed. Eds. S. Bornholdt, H. G. Schuster. Wiley VCH, Weinheim, 2003, p. 1–34.
19. Orsini C., Dankulov M. M., Colomer-de-Simón P., Jamakovic A., Mahadevan P., Vahdat A., Bassler K. E., Toroczka Z., Boguñá M., Caldarelli G., Fortunato S., Krioukov D. Quantifying randomness in real networks. *Nature Communications*, 2015, vol. 6, p. 8627.
20. Orsini C., Mitrovic D. M., Jamakovic A., Mahadevan P., Colomer-de-Simon P., Vahdat A., Bassler K., Toroczka Z., Boguñá M., Caldarelli G., Fortunato S., Krioukov D. How random are complex networks. *CoRR*, 2015, abs/1505.07503.
21. Newman M. E. J. Assortative Mixing in Networks. *Phys. Rev. Lett.*, 2002, vol. 89, p. 208701.
22. Nobari S. et al. Fast random graph generation. *Proc. of the 14th International conference on extending database technology. ACM*, 2011, p. 331–342.
23. Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U. Network motifs: simple building blocks of complex networks. *Science*, 2002, vol. 298, p. 824–827.
24. Jamakovic A. et al. How small are building blocks of complex networks. *arXiv preprint arXiv:0908.1143*. – 2009.
25. Taylor R. Constrained switching in graphs. *SIAM J. Algebraic Discrete Meth.*, 1982, vol. 3 (1), p. 115–121.
26. Simon P. C. de. RandNetGen: a Random Network Generator. URL: <http://polcolomer.github.io/RandNetGen/> (дата обращения 25.05.2018).
27. Stanton I., Pinar A. Constructing and sampling graphs with a prescribed joint degree distribution. *ACM J. Exp. Algor.*, 2012, vol. 17 (3), article 3.5 (August 2012). 25 p.
28. Ray J., Pinar A., Seshadhri C. A stopping criterion for Markov chains when generating independent random graphs. *Journal of Complex Networks*, 2015, vol. 3 (2), p. 204–220.

29. Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B., Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 2003, Nov., vol. 13 (11), p. 2498–2504.
30. Killcoyne S. et al. Cytoscape: a community-based framework for network modeling. *Protein Networks and Pathway Analysis*, 2009, p. 219–239.
31. Saito R. et al. A travel guide to Cytoscape plugins. *Nature Methods*, 2012, vol. 9, no. 11, p. 1069–1076.
32. Lotia S., Montojo J., Dong Y., Bader G. D., Pico A. R. Cytoscape app store. *Bioinformatics*, 2013, vol. 29 (10), p. 1350–1351.
33. Donovan A. A., Kernigan B. U. Yazyk programmirovaniya Go. Transl. from Engl. Moscow, Viliyams Publ., 2016, 432 p. (in Russ.)
34. Cerami E. Web services essentials: distributed applications with XML-RPC, SOAP, UDDI & WSDL. O'Reilly Media, Inc., 2002, 304 p.
35. Seymour K. et al. Overview of GridRPC: A remote procedure call API for grid computing. *International Workshop on Grid Computing*. Springer, Berlin, Heidelberg, 2002, p. 274–278.
36. Varda K. Protocol buffers: Google's data interchange format. *Google Open Source Blog*, Jul. 2008. URL: <https://opensource.googleblog.com/2008/07/protocol-buffers-googles-data.html/>. (accessed 20.07.2018)
37. Ribeiro P. gtrieScanner – Quick Discovery of Network Motifs. *CRACS & INESC-TEC, DCC/FCUP*. URL: <http://www.dcc.fc.up.pt/gtries/> (accessed 25.05.2018).
38. Goh K. I. et al. The human disease network. *Proc. of the National Academy of Sciences*, 2007, vol. 104, no. 21, p. 8685–8690.
39. Bu D. et al. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic acids research*, 2003, vol. 31, no. 9, p. 2443–2450.
40. Podkolodnyy N. L., Tverdokhleby N. N., Podkolodnaya O. A. Analiz tsirkadnogo ritma biologicheskikh protsessov v pecheni i pochkakh myshi. *Vavilovskiy zhurnal genetiki i seleksii*, 2017, vol. 21, no. 8, p. 903–910. (in Russ.)
41. Kotlyar M. et al. Integrated Interactions Database: Tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.* URL: <http://dcv.uhnres.utoronto.ca/iid/> (accessed 20.07.2018).

For citation:

Podkolodnyy N. L., Gavrilov D. A., Tverdokhleby N. N., Podkolodnaya O. A. Cytoscape Plugin for Reconstruction of Structural Random Graph Models of Biological Networks. *Vestnik NSU. Series: Information Technologies*, 2018, vol. 16, no. 3, p. 37–50. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-37-50