Математическая биология и биоинформатика. 2014. Т. 9. № 2. С. 534-542.
URL: http://www.matbio.org/2014/Petrovsky_9_534.pdf

========== **МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ** ==========

УДК: 577.21:577.29:004.42

Моделирование пространственного распределения эффекта нокаута генов, связанных с агрессивностью глиомы низкой степени злокачественности, в тканях мозга человека с помощью методов машинного обучения

Петровский Е.Д.[*1, 2], Колчанов Н.А.[1], Иванисенко В.А.[1]**

[1]Институт цитологии и генетики Сибирского отделения РАН, Новосибирск, 630090, Россия
[2]Институт "Международный томографический центр" Сибирского отделения РАН, Новосибирск, 630090, Россия

Аннотация. В настоящее время нашли широкое применение экспериментальные методы анализа транскриптомных данных, направленные на изучение особенностей экспрессии генов из различных тканей при воздействии разнообразных факторов внешней среды, а также внутренних факторов, включая полиморфизмы. В частности, существующие методы нокаута и нокдауна генов позволяют моделировать воздействие внешних факторов на экспрессию целевого гена. Имеющиеся в открытом доступе данные по экспрессии генов в различных частях организма и, в частности, в разных областях мозга позволяют построить статистические модели взаимной зависимости уровней экспрессии генов. База данных Allen Brain Atlas, например, содержит уникальные данные по пространственному распределению уровней экспрессии генов в тканях головного мозга человека и мыши. Впервые предложен подход к математическому моделированию пространственного распределения эффекта нокаута генов в тканях мозга человека с помощью методов машинного обучения и данных по экспрессии генов из Allen Brain Atlas. Показано, что нокаут центральных генов генной сети, связанной с агрессивностью глиомы низкой степени злокачественности, оказывает более значительный эффект на экспрессию других генов, по сравнению с генами, расположенными на периферии данной сети. При этом эффект имел выраженную неоднородность по локализации в пространстве.

Ключевые слова: *генные сети, мозг, экспрессия генов, микрочипы, база данных Allen Brain Atlas, база данных STRING, глиомы низкой степени злокачественности, пространственное распределение уровня экспрессии генов, методы машинного обучения.*

ВВЕДЕНИЕ

В настоящее время при изучении геномов различных организмов на первое место встают задачи выявления функциональных взаимосвязей между генами, а также молекулярных механизмов реакции организмов на воздействие внешних и внутренних факторов. Анализ транскриптомных данных (RNA-seq, микрочиповый анализ

[*]eugen_pt@bionet.nsc.ru
[**]salix@bionet.nsc.ru

экспрессии, ПЦР в реальном времени и др.) часто служит основным экспериментальным методом установления эффекта различных воздействий на экспрессию генов. Среди наиболее известных баз данных, содержащих информацию об уровне экспрессии генов в различных тканях человека и модельных животных, можно выделить Allen Brain Atlas [1], GENSAT (Gene Expression Nervous System Atlas) [2], BGEM (Brain Gene Expression Map) [3], GEO (Gene Expression Omnybus) [4]. Особое внимание заслуживает база данных Allen Brain Atlas, предоставляющая детальные данные об экспрессии генов различных тканей мозга человека и мыши, в зависимости от пространственной локализации вокселей или протяженных областей.

Для предсказания эффекта воздействия различного рода факторов на экспрессию генов часто используют математические модели. Существуют хорошо разработанные подходы к моделированию функционирования молекулярно-генетических систем и генных сетей, основанные на обыкновенных дифференциальных уравнениях, например, кинетические модели, булевы и байесовые сети [5–7]. Данные модели требуют знания о параметрах взаимодействий и константах реакций, число которых быстро растет при увеличении размера генной сети. Следует отметить, что подобная информация отсутствует для большинства генов.

В литературе описан также другой класс методов реконструкции и моделирования регуляторных генных взаимодействий, основанный на анализе данных экспрессии генов, которые не требуют информации о кинетических параметрах. Так в работах [8, 9] предлагаются байесовы методы описания механизмов регуляции по данным экспрессии генов, при том рассматриваются последовательности измерений (или моделированных данных) для разных временных точек. В работе [9] предлагается использование данных об изменении уровней экспрессии при нокауте некоторых генов для получения информации о регуляторных взаимодействиях и приводятся результаты соответствующего моделирования. В работе [10] приводится расширение стандартного метода булевых сетей для количественного моделирования регуляторных генных сетей, и, в частности, применение полученного метода к моделированию эффекта нокаута генов.

Отличительной чертой перечисленных выше, а также большинства других существующих математических моделей, описывающих динамику функционирования генных сетей, является то, что в этих моделях предполагается идеальная однородность процессов в трехмерном пространстве. Такие модели напрямую не могут быть применены, в частности, для описания процессов, распределенных по разным тканям. Пространственная неоднородность экспрессии генов в мозге, в частности, хорошо наблюдается при анализе экспрессионных данных из базы данных Allen Brain Atlas.

В настоящей работе эта база данных использовалась для моделирования пространственного распределения эффекта нокаута генов, связанных с агрессивностью глиомы низкой степени злокачественности, в тканях мозга человека с применением методов машинного обучения. Глиомы представляют собой самую распространенную группу опухолей головного мозга и различаются по степени злокачественности, гистологическим признакам, возрасту манифестации, способности к инвазии и др. Глиомы низкой степени злокачественности (Low-Grade Gliomas, LGGs, I и II степени злокачественности согласно классификации Всемирной организации здравоохранения) являются наименее агрессивной формой заболевания и характеризуются бессимптомным развитием. Среди глиом низкой степени злокачественности выделяют различные гистологические типы, включая диффузную астроцитому, олигоастроцитому, олигодендроглиому и др.[11, 12].

В отличие от кинетических моделей методы машинного обучения позволяют строить статистические зависимости экспрессии генов на основе только экспрессионных данных и не требуют знаний о константах реакций. Использование информации о прямых и опосредованных связях между генами в реконструированной

генной сети с использованием системы STRING [13] значительно снизило размерность модели, что обеспечило возможность ее обучения на имеющихся экспериментальных данных.

С помощью построенной математической модели показано, что нокаут генов, являющихся центральными в генной сети, описывающей взаимосвязи между генами, экспрессия которых связана с агрессивностью глиомы низкой степени злокачественности, оказывает более значительный эффект на экспрессию других генов, по сравнению с генами, расположенными на периферии генной сети. При этом эффект имел выраженную пространственную неоднородность.

МАТЕРИАЛЫ И МЕТОДЫ

Для построения статистических моделей были выбраны экспрессионные данные, доступные из Allen Brain Atlas. Использовались данные донора H0351.2002 (чернокожий мужчина 39 лет, посмертное исследование, время после смерти – 10 часов), содержащие нормализованные данные измерений уровней экспрессии в наибольшем числе областей мозга человека (893). Данные содержали повторные измерения для ряда генов и областей мозга, а также каждой области мозга сопоставлены её пространственные координаты на МРТ-изображении, что было использовано для МРТ-визуализации. В ходе предварительной обработки данных было проведено усреднение повторных измерений и были получены нормализованные данные по уровням экспрессии 18242 гена в 893 областях мозга.

Для построения генных сетей, описывающих прямые и опосредованные взаимодействия между анализируемыми генами человека, использовалась система STRING. Это широко известная база данных по взаимодействиям между генами/белками, аккумулирующая в себе информацию из большого числа источников, включая фактографические и курируемые экспертами базы данных. В базе данных системы STRING каждой связи сопоставлен комбинированный вес (combined score), являющийся параметром, определяющим статистическую достоверность связей. В данной работе был выбран порог для комбинированного веса, равный 0.7, соответствующий высокой степени достоверности (high confidence).

Визуализация полученных данных на МРТ-изображениях была проведена при помощи пакетов xjView [14] и Statistical Parametric Mapping (SPM8) [15], выполняемых в программной среде MATLAB (Mathworks Inc.). Координаты структур мозга в стандартном координатном пространстве (MNI, Montreal Neurological Institute) были взяты из базы данных Allen Brain Atlas и использованы для изображения данных структур поверх шаблонного T1-взвешенного МРТ-изображения головного мозга, входящего в пакет SPM8.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЯ

Построение статистических моделей, описывающих эффект действия нокаута генов

Выбор генов-мишеней для моделирования эффекта действия нокаутов проводился с использованием результатов работы [16], представленной в данном выпуске. Для анализа был выбран ген *spp1*, являющийся центральным в генной сети глиомы низкой степени злокачественности, а также *ddr1*, находящийся на периферии данной сети, что предполагает различную степень вовлеченности этих генов в процессы, характеризующие агрессивность опухолевого заболевания. Оба этих гена вовлечены в биологический процесс «адгезия клеток», который оказался представленным наибольшим числом генов анализируемой сети.

При построении модели, позволяющей предсказывать изменение уровня экспрессии заданного гена по изменению уровней экспрессии других генов, за основу была взята гипотеза о том, что значимое влияние на экспрессию гена может оказывать только экспрессия генов, непосредственно связанных с ним в генной сети. Таким образом, на первом шаге, проводили построение генной сети с использованием системы STRING, описывающей все прямые взаимодействия центральных генов *spp1* и *ddr1* с другими генами.

Известно, что «клеточная адгезия» является важным фактором развития опухоли. В связи с этим, для нас было интересным исследовать влияние нокаута генов *spp1* и *ddr1* на уровень экспрессии генов, участвующих в соответствующем биологическом процессе Gene Ontology «клеточная адгезия».

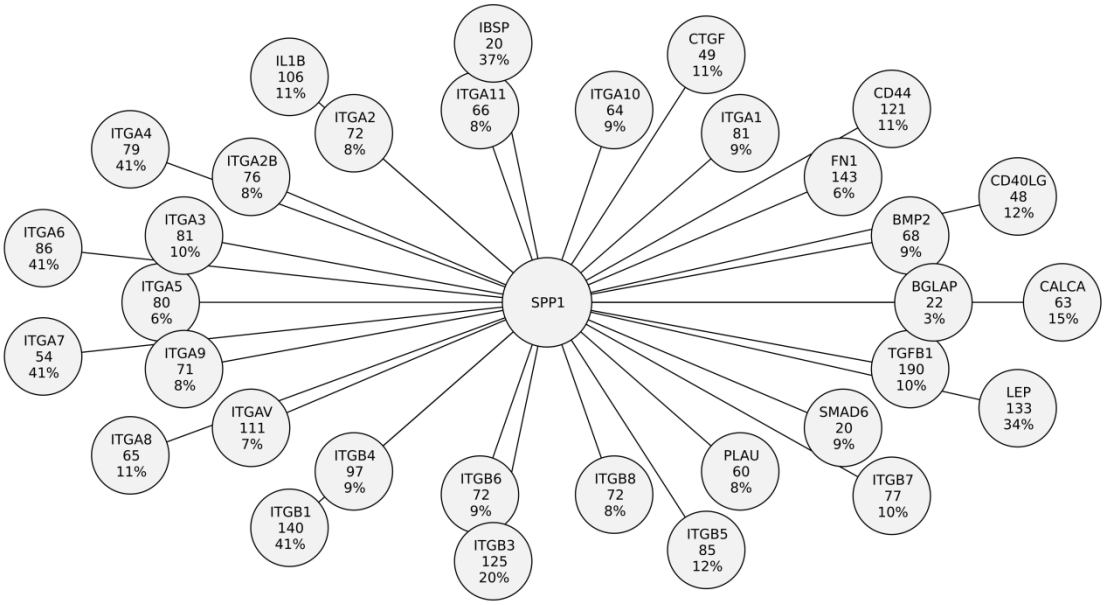

Рис. 1. Реконструированная генная сеть, описывающая непосредственные связи гена *spp1* с генами, участвующими в биологическом процессе Gene Ontology «клеточная адгезия». Под именем каждого целевого гена приведено число взаимодействующих с ним генов, учитываемых при построении модели, а также, в процентах, погрешность предсказания с помощью множественного линейного регрессионного анализа его уровня экспрессии.

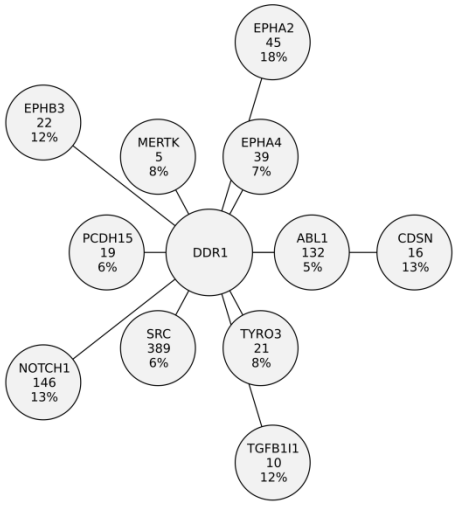

Рис. 2. Реконструированная генная сеть, описывающая непосредственные связи гена *ddr1* с генами, участвующими в биологическом процессе Gene Ontology «клеточная адгезия». Под именем каждого целевого гена приведено число взаимодействующих с ним генов, учитываемых при построении модели, а также, в процентах, погрешность предсказания с помощью множественного линейного регрессионного анализа его уровня экспрессии.

Выше (рис. 1 и 2)показаны две генные сети, описывающие непосредственные связи генов *spp1* и *ddr1* с генами, участвующими в биологическом процессе Gene Ontology «клеточная адгезия», которые являются целевыми генами для предсказания влияния нокаута на их экспрессию. Поскольку на уровень экспрессии целевых генов из процесса «клеточная адгезия» оказывают влияние и другие взаимодействующие с ними гены, то на втором шаге, проводили расширение каждой генной сети, т.е., в расширенную генную сеть включали новые гены, непосредственно взаимодействующие с целевыми генами. Следует заметить, что для центральных генов *spp1* и *ddr1* расширение не проводили. Таким образом, построенная расширенная генная сеть для каждого из генов *spp1* и *ddr1* включала 2697 и 844 гена, соответственно (выше (рис. 1, 2) для каждого целевого гена приведено число соседей в расширенной сети).

Далее для каждого целевого гена была построена регрессионная модель, связывающая его уровень экспрессии с уровнями экспрессии взаимодействующих с ним генов, включая гены *spp1* и *ddr1*, для которых изучался эффект нокаута. Следует отметить, что один из генов *spp1* и *ddr1* обязательно входил в состав каждой из рассмотренных моделей.

Для создания регрессионных моделей для каждого *i*-го целевого гена строилась таблица $\mathbf{X}_i$ уровней экспрессии его генов-соседей. Строки такой таблицы соответствовали структурам мозга (согласно базе данных Allen Brain Atlas), столбцы – генам, а ячейки содержали значения уровней экспрессии соответствующих генов для соответствующих структур мозга. Данные, представленные в этой таблице, рассматривались как независимые переменные. Зависимой переменной был вектор $\mathbf{y}_i$, элементами которого являлись уровни экспрессии целевого гена в соответствующих структурах мозга. В качестве типа регрессионной модели был выбран множественный линейный регрессионный анализ, описывающий связь уровня экспрессии целевого гена (зависимая переменная) с уровнями экспрессии генов-соседей (независимые переменные):

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{k}_i + k_i^0, \tag{1}$$

где $\mathbf{k}_i$ – векторный, а k_i^0 – скалярный параметры регрессионной модели, а количество компонент вектора равно количеству структур мозга.

Выбор линейной модели был обусловлен простотой такого класса моделей, построение которых требует оценки наименьшего количества параметров по сравнению с нелинейными моделями. Оценка точности предсказаний линейной модели проводилась при помощи перекрестной проверки. Для этого, из рассмотренной выше матрицы случайным образом удалялись 10% строк, которые использовались как тестовый набор данных для расчета точности предсказания модели. Обучение модели проводилось на оставшихся 90% строк. Процедура повторялась не менее 1000 раз. В качестве погрешности модели *i*-го целевого гена d_i вычислялась средняя по всем итерациям перекрестной проверки среднеквадратичная относительная разница между предсказанными и наблюдаемыми данными:

$$d_i = \frac{1}{M}\sum_{j=1}^{M}\sqrt{\frac{1}{Ns}\sum_{s_j=1}^{Ns}\left(\frac{y_i^{s_j} - Y_{i,j}^{s_j}}{y_i^{s_j}}\right)^2}\ , \tag{2}$$

где M – число повторений перекрестной проверки, Ns – число структур мозга в тестовом наборе данных, $y_i^{s_j}$ – уровень экспрессии *i*-го целевого гена в структуре s_j-ой

перекрестной проверки, $Y_{i,j}^{s_j}$ – предсказанный уровень экспрессии *i*-го целевого гена в структуре s_j-ой перекрестной проверки.

Полученные оценки погрешности предсказания уровня экспрессии для каждого целевого гена выражены в процентах и приведены выше (рис. 1, 2). Из рисунков видно, что погрешность варьировалась от 3% до 41%, при этом больше половины построенных моделей обладали погрешностью, не превышающей 10% (18 из 33 и 6 из 11 для генов *spp1* и *ddr1* соответственно).

Пространственное распределение предсказанных эффектов нокаута генов

Для моделирования эффекта нокаута генов *spp1* и *ddr1* уровень их экспрессии принимался равным нулю. Таким образом, в матрице $\mathbf{X}_i$, содержащей входные данные для регрессионной модели целевого вектора y_i, столбцы, соответствующие генам *spp1* и *ddr1*, заполнялись нулями. Для каждой структуры **s** изменение уровня экспрессии целевых генов до и после нокаута оценивалось как среднеквадратичное по целевым генам относительное изменение уровня экспрессии, деленное на погрешность модели, определенную методом перекрестной проверки:

$$E^s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\left(Y_i^{s,0} - y_i^s\right)}{y_i^s} / d_i\right]^2}, \qquad (3)$$

где суммирование идёт по всем генам-соседям гена-мишени, то есть по целевым генам, y_i^s – значение уровня экспрессии *i*-го целевого гена в структуре мозга **s** по данным Allen Brain Atlas, $Y_i^{s,0}$ – предсказанное значение уровня экспрессии *i*-го гена-соседа в структуре мозга **s** при нокауте гена-мишени, d_i – среднеквадратичная относительная погрешность модели *i*-го гена-соседа, определенная методом перекрестной проверки.

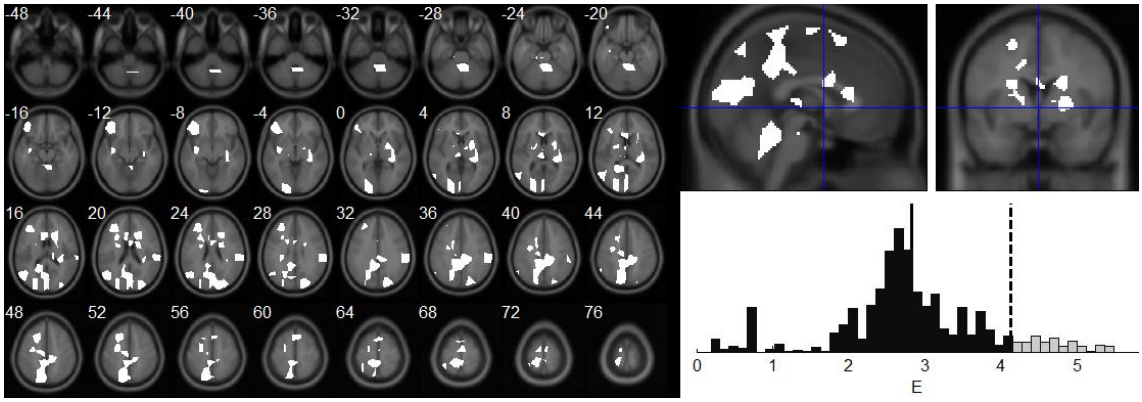

Рис. 3. Пространственное распределение смоделированного эффекта нокаута гена *spp1*. Слева белым отмечены 10% объема мозга, имеющие наибольшие значения эффекта, области изображены на серии аксиальных срезов МРТ-изображения головного мозга, цифрами отмечены *z*-координаты срезов в MNI-пространстве. Справа сверху – то же, на среднесаггитальном и фронтальном срезах, проходящих через начало координат MNI-пространства. Справа снизу – гистограмма распределения значения величины эффекта по объему головного мозга, сплошной вертикальной черной линией отмечено среднее значение, пунктирной – 10% квантиль, белым изображена часть распределения, соответствующая областям, отмеченным белым на МРТ-изображениях.

Таким образом, для всех структур мозга для каждого из нокаутируемых генов *ddr1* и *spp1* были получены значения величины *E*, характеризующей эффект нокаута, усредненный по целевым генам, то есть по генам, взаимодействующим с нокаутируемым геном, или геном-мишенью. Величина *E* выражена в единицах

погрешности соответствующих моделей, и может характеризовать, таким образом, статистическую значимость результата.

Для каждой структуры **s** в базе данных Allen Brain Atlas имеются её пространственные координаты в стандартном пространстве MNI, что позволило визуализировать пространственное распределение получаемых величин E_s на МРТ-изображениях реального мозга. Ниже (рис. 3, 4) изображены пространственные распределения эффекта нокаута E_s генов *spp1* и *ddr1* соответственно.

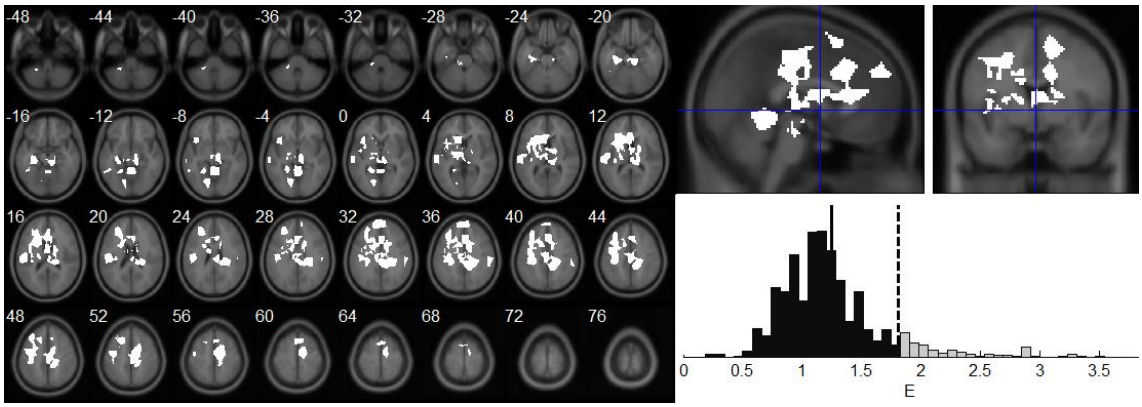

Рис. 4. Пространственное распределение смоделированного эффекта нокаута гена *ddr1*. Расположение частей рисунка, форма представления информации и пространственное расположение срезов МРТ-изображений соответствуют таковым на рис. 3.

Можно видеть, что смоделированные распределения изменения уровней экспрессии целевых генов при нокауте генов *spp1* и *ddr1* имеют качественно различный характер. Области максимального эффекта нокаута гена *spp1* сосредоточены в основном в теменной и затылочной долях коры головного мозга, тогда как аналогичные области для гена *ddr1* в основном сосредоточены во фронтальной доли коры головного мозга и захватывают также часть объема лимбической системы. При этом средняя по всему объему мозга величина эффекта более чем в 2 раза больше для гена *spp1*, который располагается в центральной позиции в сети коэкспрессии, связанной с агрессивностью опухолевого заболевания, согласно работе [16]. Важно отметить, что измерялся средний эффект по соседям гена-мишени по генной сети, и ген *spp1* имеет в 3 раза больше таких соседних генов. Также стоит заметить, что оба полученных средних значения превосходят единицу, что соответствует величине эффекта, большей, чем погрешность модели, и может служить показателем статистической достоверности получаемых результатов.

ЗАКЛЮЧЕНИЕ

В данной работе был предложен метод моделирования пространственного распределения эффекта нокаута, основанный на методах машинного обучения и информации, доступной из базы данных Allen Brain Atlas. В качестве объекта исследования были выбраны гены, входящие в реконструированную ранее генную сеть коэкспрессии, связанную с агрессивностью глиомы низкой степени злокачественности [16]. Было проведено сравнительное моделирование эффектов для двух генов, занимающих различное положение в этой генной сети, и показано различие, как в пространственном распределении эффекта нокаута, так и в средней величине эффекта.

Частью предложенного метода является процедура построения статистических моделей взаимной зависимости уровней экспрессии генов. Такая процедура может представлять самостоятельную научную ценность и применяться для построения широкого круга моделей, помимо продемонстрированных моделей эффектов нокаута

генов. Представляется возможным также расширение модели на различные виды генных сетей и углубление модели с использованием более совершенных видов зависимости, отличных от линейных, что позволит проводить моделирование сложных регуляторных взаимодействий и более подробный анализ получаемых результатов.

Разработка методов и анализ экспрессионных данных осуществлялись при поддержке гранта РНФ № 14-24-00123. Создание параллельных версий программ и суперкомпьютерные вычисления выполнялись при поддержке проекта VI.61.1.2.

ЛИТЕРАТУРА

1. Hawrylycz M.J., Lein E.S., Guillozet-Bongaarts A., Shen E.H., Ng L., Miller J.A., van de Lagemaat L.N., Smith K.A., Ebbert A., Riley Z.L. et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*. 2012. V. 489. № 7416. P. 391–399.
2. Heintz N. Gene expression nervous system atlas (GENSAT). *Nature Neuroscience*. 2004. V. 7. № 5. P. 483–483.
3. Magdaleno S., Jensen P., Brumwell C.L., Seal A., Lehman K., Asbury A., Cheung T., Cornelius T., Batten D.M., Eden C. et al. BGEM: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system. *PLoS Biology*. 2006. V. 4. № 4. Article No. e86.
4. Edgar R., Domrachev M., Lash A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*. 2002. V. 30. № 1. P. 207–210.
5. De Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*. 2002. V. 9. № 1. P. 67–103.
6. Someren E.V., Wessels L.F.A., Backer E., Reinders M.J.T. Genetic network modeling. *Pharmacogenomics*. 2002. V. 3. № 4. P. 507–525.
7. Hecker M., Lambeck S., Toepfer S., van Someren E., Guthke R. Gene regulatory network inference: data integration in dynamic models – a review. *Biosystems*. 2009. V. 96. № 1. P. 86–103.
8. Husmeier D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*. 2003. V. 19. № 17. P. 2271–2282.
9. Rogers S., Girolami M.A. Bayesian regression approach to the inference of regulatory networks from gene expression data. *Bioinformatics*. 2005. V. 21. № 14. P. 3131–3137.
10. Steggles L.J., Banks R., Shaw O., Wipat A. Qualitatively modelling and analysing genetic regulatory networks: a Petri net approach. *Bioinformatics*. 2007. V. 23. № 3. P. 336–343.
11. Duffau H., Capelle L. Preferential brain locations of low-grade gliomas. *Cancer*. 2004. V. 100. № 12. P. 2622–2626.
12. Pouratian N., Schiff D. Management of low-grade glioma. *Current Neurology and Neuroscience Reports*. 2010. V. 10. № 3. P. 224–231.
13. Franceschini A., Szklarczyk D., Frankild S., Kuhn M., Simonovic M., Roth A., Lin J., Minguez P., Bork P., von Mering C., Jensen L.J. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*. 2013. V. 41 № D1. P. 808–815.
14. Cui X., Li J. *XjVie, a viewing program for SPM*. URL: http://people.hnl.bcm.tmc.edu/cuixu/xjView/ (дата обращения: 11.09.2014).
15. Penny W., Friston K., Ashburner J., Kiebel S., Nichols T. *Statistical Parametric Mapping: The Analysis of Functional Brain Images.* 2007. 656 p.

16. Иванисенко Н.В., Губанова Н.В., Колчанов Н.А., Иванисенко В.А. Предсказание экспрессионных маркеров агрессивности глиомы низкой степени злокачественности по экспрессионным данным TCGA. *Математическая биология и биоинформатика*. 2014. Т. 9. № 2. С. 527–533. URL: http://www.matbio.org/2014/Ivanisenko_9_527.pdf.

Материал поступил в редакцию 27.11.2014, опубликован 18.12.2014.