**Разработка программного модуля для нахождения тройных взаимодействий в биологических сетях с учётом временных задержек**

**Введение**

Анализ взаимодействия генов берет свое начало в 19 веке [1]. Однако только в конце 20 – начале 21 века развитие экспериментальных подходов и алгоритмов выявления генных взаимодействий позволило восполнить исследовательский интерес [REFERENCE]. Данная область позволяет найти взаимосвязь между биологическими процессами, протекающими в организме, с болезнями. Так, группа ученых обнаружила взаимодействующие гены, ответственные за развитие шизофрении [2].

Набор взаимодействующих генов удобно визуально представлять в виде сети, в которой узлами являются гены, а взаимодействия обозначаются направленными отрезками. Выявление структуры генной сети является одной из задач для анализа взаимодействий генов. При наличии в сети 1000 генов, количество возможных взаимодействий может достигать 1030. Прямое моделирование процесса взамодействий не всегда является возможным ввиду ограниченности экспериментальных данных полностью описывающих структуру, поведение и функции того или иного гена. Классом эффективных алгоритмов являются алгоритмы, работающие по принципу “обратной разработки”[REFERENCE]: зная набор взаимодействующих генов и результат их взаимодействия, можно восстановить структуру сети, которая приводит к наблюдаемым результатам. Результатом взамодействий генов в сети – это изменение концентрации продуктов генной сети во времени (белков, протеинов, и т.д.). Такой подход позволяет решать задачи восстановления сети без знания особенностей гена**,** снизить человеческие ресурсы, необходимые для ее решения.

В генных сетях присутсвуют как парные взаимодействия [REFERENCE], так и взаимодействия более высокого порядка [REFERENCE](согласованного взаимодействия более чем двух элементов) [3]. В данной работе, автор ограничивается рассмотрением тройных взаимодействий. Особый интерес представляет нахождение взаимодействий с учётом временных задержек, так как в реальных биологических процессах, реакция того или иного компонента сети может проявляться с задержкой[REFERENCE]. На текущий момент, существуют алгоритмы, учитывающие взаимодействия третьего порядка[REFERENCE], а также алгоритмы, учитывающие парные взаимодействия с временными задержками. Алгоритмов же, учитывающих тройные взаимодействия с временными задержками одновременно на данный момент не существует.

Целью данной работы является разработка алгоритма, учитывающего взаимодействия третьего порядка с временными задержками. Данный алгоритм должен использовать метод DTW компенсации временных задержек для учета различных вариаций наблюдаемых значений концентрации [REFERENCE]. Алгоритм должен восстанавливать структуру сети объемом 1000 элементов не более, чем за час.

За основу в работе был принят алгоритм MIDER, способный учитывать парные взаиомдействия с временными задержками[REFERENCE]. В работе показано, что доработка алгоритма MIDER учетом тройных взаимодействий, а также изменение алгоритма учета временных задержек на алгоритм DTW увеличивает точность алгоритма на [Insert] процентов. Данные для тестирования были взяты с открытой площадки Dream Challenge. Сложность алгоритма была уменшена в [INSERT] раз за счет применения регуляризации Шмидта, что позволило увеличить количество элементов в исследуемых сетях до [INSERT]. Данный алгоритм может быть эффективно применён как для исследования реальных биологических сетей, так и для других задач анализа взаимодействий, как, например, в задачах пространственной экономики.

Задача нахождения взаимодействий генов существует уже почти столетие [REFERENCE], но серьезных результатов, а именно большого количества рассмотренных взаимодействий, удалось достичь только в конце 20 – начале 21 века [REFERENCE]. В это время развитие компьютерных технологий и алгоритмов работы с большими данными позволили находить взаимодействие тысяч генов за время порядка недель, месяцев, в зависимости от количества генов и используемого алгоритма, что раньше не представлялось возможным.

Взаимодействующие гены, как правило, представляют в виде сети, где узлами сети являются гены, а стрелки между ними – направленные взаимодействия. Здесь используются направленные стрелки, так как продукт активации (например, белок) одного гена, может активировать работу другого гена [REFERENCE]. Такую информацию необходимо распознавать для восстановления хронологии работы сети. Представив взаимодействующие гены, как сеть, встает вопрос, каким образом восстановить структуру этой сети.

Но прежде, чем выбрать метод восстановления сети, нужно решить, какие взаимодействия внутри генной сети требуется восстановить. Этот вопрос обусловлен тем, что реальных биологических системах существуют как парные взаимодействия генов, так и взаимодействия более высокого порядка [REFERENCE]. Так, например, в компактных белковых структурах может присутствовать одновременное взаимодействие трёх и более генов, что необходимо учитывать при восстановлении пространственной формы такой структуры [REFERENCE]. Помимо различных вариантов взаимодействий, активация гена из-за работы другого гена может произойти с задержкой. Причиной этому может служить необходимая концентрация того или иного белка для активации, время, необходимое для протекания реакции и т.д. [REFERENCE]. Все эти факторы влияют на количество возможных взаимодействий в сети, набор используемых алгоритмов, что в свою очередь влияет на время, необходимое для решения задачи восстановления сети.

Определившись с границами задачи, нужно выбрать подход к решению этой задачи. На текущий момент существует несколько подходов:

1. Экспертное восстановление сети на основе трудов ученых области анализа взаимодействий генов [REFERENCE]
2. Автоматическое восстановление сети на основе трудов ученых области анализа взаимодействий генов на основе машинных систем [4]
3. Прямое моделирование взаимодействий генов [5]
4. Восстановление генной сети по принципу обратной разработки [REFERENCE]

Выбор подхода зависит от объема сети, временного ограничения для решения задачи. Ниже вкратце рассмотрены все подходы по порядку:

1. Метод заключается в ручном восстановлении сети. Эксперт ознакамливается с различными трудами, связанными с теми генами, взаимодейтсвие которых требуется восстановить. Эти труды представляют из себя публикации, в которой указаны результаты эксперимента взаимодействия генов. На основании анализа этих трудов, эксперт восстанавливает структуру сети. Данный подход даёт наиболее точные результаты восстановление сети, но плохо масштабируется на сети объемом более 10 элементов, поскольку требует больших человеческих трудозатрат. Это ограничение не позволяет рапозновать сети объемом порядка 100, 1000 элементов, которые встречаются в реальных биологических системах. К тому же, восстановление сетей с учетом взаимодействий высокого порядка затруднено, так как основная часть работ по анализу взаимодействий, на текущий момент, рассматривает только парное взаимодействие.
2. Подход заключается в автоматическом анализе текста публикаций по теме, интересующей исследователя, для выделения характеристик взаимодействия генов, а также особенностей генов, влияющих на структуру сети. Метод основан на использовании алгоритмов анализа биологического текста, таких как cTAKES [6], MetaMap [7] и т.д. Данный подход позволяет увеличить объем рассматриваемых генных сетей до тысяч, снижает человеческие ресурсы, позволяет рассматривать до десятков тысяч статей в час, что позволяет обработать базу данных с 106 записей за дни. Недостатками такого подхода является то, что как правило рассматривается только Abstract статей, в то время как ценная информация о характеристиках гена может содержаться в теле статьи. Во время написания статьи могут быть ошибки, что делает подход зависимым от ошибок других людей. Сбор информации для таких методов может занимать месяцы и годы, так как получение релавантной информации, не вносящей шум в результирующую сеть производится в полуавтоматическом режиме.
3. Прямое моделирование генных взаимодействий заключается в моделировании взаимодействия на основе характеристик гена. Выявив в ручном или полуавтоматическом режиме характеристики генов (таких как, форму, с какими генами взаимодействует и почему, локализацию и т.д.), задача заключается в формировании системы дифференциальных уравнений, на которую накладываются ограничения из полученных данных. Данный метод также позволяет увеличить объем рассматриваемой сети до сотен и тысяч, требует меньшей информации для начала обсчета. Но, к сожалению, не всегда существует подробная информацияя по генам в системе, и не всегда ее можно найти.
4. Моделирование по принципу обратной разработки заключается наблюдении изменения концентраций продуктов работы генов. Здесь и далее для обозначения изменения концентраций продуктов работы генов в отдельно взятом эксперименте будет использоваться термин *сигнал*. Сигналы различных генов в сети несут в себе информацию о взаимодействии, так как уменьшение или увеличение концентрации свидетельствует о связи. Чем более сигналы скоррелированы, тем более вероятно наличие взаимосвязи между соответсвующими генами. Основным шагом таких подходов нахождение корреляции сигналов для дальнейшего применения в алгоритмах максимизации или минимизации энтропии [8], [9] и других. Данный подход позволяет обрабатывать сети объемом тысяч и десятка тысяч элементов, не требует априорного знания характеристик генов, позволяет легко внедрить учет временных задержек за счет использования корреляционного подхода, позволяет полностью автоматизировать процесс распознавания генной сети. Также, экспериментальные данные для таких методов могут быть получены не только из текста, но и из современного эксперимента на ДНК-микрочипе, что позволяет получить информацию о тысячах сигналах генов меньше, чем за день [10].

В текущей работе необходимо распознавать генную сеть объемом порядка тысяч элементов за времена поряда часов, при этом затраты на получение данных по сети необходимо максимально минимизировать. Исходя из приведенного рассмотрения классов алгоритмов, а также требований, для дальнейшей работы был выбран класс алгоритмов обратной разработки. Поэтому, следующая часть главы будет посвящена этим методам.

An external file that holds a picture, illustration, etc.
Object name is fgene-03-00071-g002.jpgКак было отмечено, методы обратной разработки работают с сигналами генов. Визуально, эти сигналы можно представить в виде графика:

Рис. 1.1 Изменение концентрации продуктов генов YRO2–YHP1–MRH1, полученная на ДНК-микрочипе [3]. Ось x - время, ось y - уровень концентрации.

Эти сигналы используются для нахождения количественной характеристики уровня взаимосвязи между сигналами. Так, например, для нахождения этой величины может использоваться корреляционный коэффициент Пирсона [11]:

1‑1

, где – сигналы генов, – значение сигналов генов в определенный момент времени. Данный коэффициент прост и известен, но его применение на практике зачастую невозможно по следующей причине. Предположим, что в исслеудемой генной сети сожержатся 3 гена:

Рис. 1.2 Пример генной сети, в которой ген A активирует работу генов B и C.

Ген A активирует работу генов B и C, что в свою очередь приводит к похожему изменению их концентраций. В этом случае, коэффициент Пирсона покажет, что сигналы генов B и C скоррелированы, а значит гены взаимодействуют друг с другом, но на самом деле это не так. Для решения этой проблемы используется условная корреляция:

1‑2

В этом случае, коэффициент корреляции между г енами B и C не покажет взаимосвязи между ними. Аналогично, можно посчитать корреляцию двух сигналов генов при условии фиксированного значения нескольких сигналов генов.

Корреляция генов B и C при условии фиксированного значения сигнала гена A

Другой способ решения этой проблемы предложен Самойловым [12], [13] и используется в алгоритме MIDER восстановления генных сетей [9]. Он основан на методе уменьшения энтропии [14], характеризуемой выражением:

1‑3

, где – случайные вектора, а – возможные их значения. Идея заключается в итеративном выделении значимого подпространства генов , которое максимально минимизирует энтропию настолько, что добавление нового не изменяет энтропию:

1‑4

Таким образом, представляет из себя непосредственно те гены, которые взаимосвязаны с текущим.

На моменте подсчета *параметра взаимосвязи*(поскольку техник подсчета связи много: корреляция, условная корреляция, подход минимизации/максимизации энтропии – здесь и далее все эти техники будут называться этим словосочетанием), как правило, вводят и учет временных задержек.

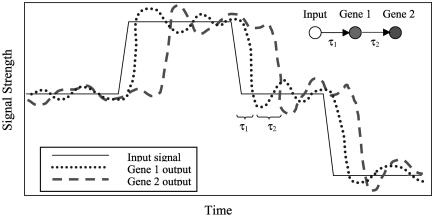


Рис. 1.3 Пример изменения концентрации гена 2 с задержкой, вызванной активацией гена 1 [16]

Упомянутые формулы нахождения параметра взаимосвязи (1‑1, 1‑4) считаются на основании последовательных сигналов, которые, как показано рисунке выше, могут быть сдвинуты во времени, что может дать неправильный результат. Следовательно, перед подсчетом параметра взаимосвязи, нужно воспользваться алгоритмом выравнивания сигналов.

В алгоритме MIDER [9] используется простой итеративный подход, основанный на последовательном смещении рассматриваемой зоны сигнала с начала и хвоста сигнала для нахождения минимальной энтропии:

1‑5

, где , , – параметр сдвига. Таким образом, алгоритм MIDER [9] учитывает сдвиги сигнала. Но помимо сдвигов, сигнал может быть растянут или сжат. Для учета этого может быть использован алгоритм DTW [15], принцип работы которого основан на составлении матрицы трансформации (на основе матрицы расстояний между узлами сигналов) и последующего прохода по ней для построения оптимального пути, минимизирующего стоимость пути. Так, на рисунке ниже показан пример обычного сопоставления сигналов и с использованием алгоритма DTW.

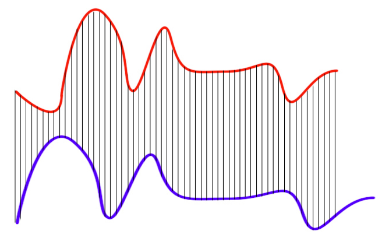
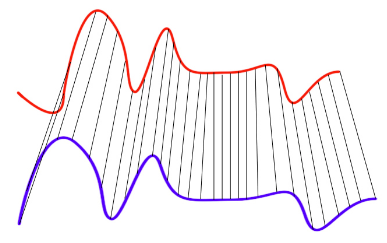


Рис. 1.4 Выравнивание сигналов. Слева – евклидово выравнивание, справа – DTW выравнивание [17]

После нахождения матрицы параметоров взаимосвязи (возможно, с учетом временных задержек), дальнейшие действия варьируются для различных алгоритомов.

Так, рассмотрим алгоритм MIDER [9], который на текущий момент является одним из признанных алгоритмов нахождения генной сети с учетом только двойных взаимодействий и временных задержек. Алгоритм MIDER основан на концепции минимизации энтропии. Идея такого подхода взята из второго закона термодинамики, следствием которого является утверждение, что энтропия закрытой системы сходится к константе [16]. Таким образом, рассматривая генную сеть, как закрытую систему, условием выгодного состояния сети будет минимум энтропии. Как было сказано, энтропия взаимодействия двух элементов описывается уравнением 1‑3, но также было упомянуто, что на значение энтропии, как параметра взаимосвязи, может влиять общий активирующий ген. Для учета такого поведения, в алгоритме MIDER, используется как принцип, предложенный Самойловым [12], [13], так и условная энтропия, определяемая выражением:

1‑6

Для получения более точных результатов, используется не только энтропия гена при условии фиксированного значения одного гена, но и при условии фиксированного значения нескольких генов(максимум - четырех). Очевидно, что с увеличением фиксируемых генов, возрастает и вычислительная сложность данного этапа алгоритма как , где – количество условных генов. Такая процедура проделывается для каждого сигнала. Эта информация используется для подсчета взаимодействующих генов по формуле 1‑4. Также, для подсчета матрицы переносной энтропии, используемой для восстановления направления взаимодействия:

1‑7

Данная мера является показателем воможности уменьшения вариантов возможного значения в зависимости от прошлых значений и . Посколька она является несимметричной, выбор направления зависит от максимума значения при том или ином направлении [17]. Помимо энтропии и матрицы энтропий переноса, подсчитываются также расстояния между сигналами генов по формуле:

, где – мера, количетсвенно характеризующяя насколько связаны , а именно, с какой уверенностью можно судить о значении , если известно значение .

**Заключение**

В данной работе был разработан алгоритм для нахождения тройных взаимодействий в биологических сетях с учётом временных задержек на основе алгоритма MIDER. На текущий момент, этот алгоритм является первым, одновременно учитывающим тройные взаимодействия, а также временные задержки.

Доработки алгоритма MIDER учетом тройных взаимодействий, алгоритмом DTW позволили улучшить его точность, что было продемонстрировано на площадке DREAM Challenge, на испытаниях [REFERENCE]. На каждом из испытаний, алгоритм восстанавливал структурные гены сети с точностью [INSERT]%, а также взаимодействия между ними с точностью [INSERT]%, что на [INSERT]% в среднемвыше, чем расммотренных в Главе 1 алгоритмов. Данный алгоритм может быть применён как для исследования реальных биологических сетей, так и для других задач анализа взаимодействий, на выборках объёмом не более [INSERT] элементов. Предложенные методы оптимизации, такие как регуляризация Шмидта, выделение значимого подпространства, позволяют уменьшить вычислительную сложность алгоритма минимум в [INSERT] раза, и дальнейшие изыскания в области оптимизации алгоритма позволят увеличить производительность в 10-100 раз [INSERT].

Открытая реализация алгоритма[REFERENCE] в среде MatLab позволяет уменьшить трудозатраты генетиков при восстановлении сети на основе экспериментальных данных.

Нахождение взаимодействий более высокого порядка(четырёх и более) является перспективной задачей, поскольку учёт таких взаимодействий позволяет получить более точную картину сети. В особенности, если учитываются временные задержки. Дальнейшая работа в этой области будет направлена на увеличение порядка учитываемых взаимодействий, а также на оптимизацию существующих методов.

# Bibliography

|  |  |
| --- | --- |
| [1] | G. Mendel, «Versuche über Pflanzen-Hybriden,» *Proceedings of the Natural History Society of Brünn,* 1866. |
| [2] | J. C. B. X. T. S. B. J. A. G. M. K. &. D. V. Sarah R Gilman, «Schizophrenia Gene Networks Found, with Link to Autism,» NATURE NEUROSCIENCE, 2012. |
| [3] | Y.-C. H. C.-M. C. G. S. S. Ci-Ren Jiang, «Inferring Genetic Interactions via a Data-Driven Second Order Model,» National Center for Biotechnology Information, 2012. |
| [4] | O. A. A. A. S. G. T. Ö. D. D. a. R. A. Gabriela Jurca, «Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends,» BMC Research Notes, 2016. |
| [5] | G. P. W. Thomas F. Hansen, «Modeling Genetic Architecture: A Multilinear Theory of Gene Interaction,» ScienceDirect, New Haven, Connecticut, 2001. |
| [6] | J. J. M. P. V. O. J. Z. S. S. K. C. K.-S. a. C. G. C. Guergana K Savova, «Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,» Journal of the American Medical Informatics Association : JAMIA, 2010. |
| [7] | A. R. Aronson, «Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.,» Proceedings of the AMIA Symposium. , 2001. |
| [8] | O. G. T. Jian Zhou, «Global Quantitative Modeling of Chromatin Factor Interactions,» Public Library of Science, 2014. |
| [9] | J. R. F. M. J. R. B. Alejandro F. Villaverde, «MIDER: Network Inference with Mutual Information Distance and Entropy ReductionAlejandro F. Villaverde , John Ross, Federico Morán, Julio R. Banga,» Public Library of Science, 2014. |
| [10] | F. F. a. H. A. B.-S. Victor Trevino, «DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research,» Molecular Medicine, 2007. |
| [11] | H. M. K. C. Y. E. E. Nicholas A. Furlotte, «Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity,» APPLIED BIOINFORMATICS, 2011. |
| [12] | S. M, «Reconstruction and functional analysis of general chemical reactions and reaction networks,» Stanford, 1997. |
| [13] | A. A. R. J. Samoilov M, «On the deduction of chemical reaction pathways from measurements of time series of concentrations,» 2001. |
| [14] | S. C, «A mathematical theory of communication,» 1948. |
| [15] | L. Gupta, D. L. Molfese, R. Tammana и P. G. Simos, «Nonlinear alignment and averaging for estimating the evoked potential,» IEEE Transactions on Biomedical Engineering, 1996. |
| [16] | J. R. M. R. a. G. S. William A. Schmitt, «Elucidation of Gene Interaction Networks Through Time-Lagged Correlation Analysis of Transcriptional Data,» Genome Research, 2004. |