**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.

**Глава 1**

Задача нахождения взаимодействий генов существует уже почти столетие [REFERENCE], но серьезных результатов, а именно большого количества возможных 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.

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 реакци 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. Забегая вперед, отмечу, что в данной работе предлагается алгоритм восстановления генной сети с учетом тройных взаимодействий и временными задержками.

Определившись с границами задачи, наступает момент выбора подхода решения этой задачи:

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]

Рассмотрим каждый вариант 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 публикацию, в которой указаны результаты эксперимента взаимодействия двух 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 информпции, 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 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 час [10].

Исходя из приведенного рассмотрения классов алгоритмов, для дальнейшей работы был выбран класс алгоритмов обратной разработки. Поэтому, следующая часть главы будет посвщена этим методам.

**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.

# Works Cited

|  |  |
| --- | --- |
| [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. |