Использование формальных грамматик и искусственных нейронных сетей для анализа вторичной структуры геномных и протеомных последовательностей

Семён Григорьев

8 августа 2019 г.

1 Основные данные проекта

1.1 Название проекта

ru
Использование формальных грамматик и искусственных нейронных сетей для анализа вторичной структуры геномных и протеомных последовательностей

en

1.2 Основной код (по классификатору РФФИ)

Н3 Переход к персонализированной медицине, высокотехнологичному здравоохранению и технологиям здоровьесбережения, в том числе за счет рационального применения лекарственных препаратов (прежде всего антибактериальных)

1.3 Дополнительные коды (по классификатору РФФИ)

1.4 Ключевые слова (указываются отдельные слова и словосочетания, наиболее полно отражающие содержание проекта: не более 15, строчными буквами, через запятые)

ru

en

1.5 Аннотация проекта (кратко, в том числе – актуальность, уровень значимости и научная новизна исследования; ожидаемые результаты и их значимость; аннотация будет опубликована на сайте РФФИ, если проект получит поддержку)

ru

Графы широко используются для представления данных в таких областях, как социальные сети, графовые базы данных, верификация, semantic web, биоинформатика. Одной из часто решаемых задач является задача поиска путей в графе, удовлетворяющих заданным ограничениям на метки рёбер пути. Рассматривая метку одного ребра как символ в некотором алфавите, пути в графе можно сопоставить слово, а множеству путей — язык над некоторым алфавитом. Таким образом, искомое множество путей можно специфицировать с помощью грамматики. При этом выразительности регулярных грамматик часто оказывается недостаточно, поэтому при решении многих практических задач необходимо использовать контекстно-свободные грамматики (КС-грамматики). Например, такая задача биоинформатики, как поиск в метагеномной сборке цепочек с заданной вторичной структурой, которая описывается КС-грамматикой, тоже может быть сформулирована как поиск путей в графе. Также использование КС-грамматик актуально при обработке встроенных языков или динамически формируемого кода — ситуаций когда программа формирует в процессе своей работы код другой программы. Здесь граф представляет собой конечный автомат, порождающий все возможные значения динамически формируемого кода, а ограничения задаются грамматикой языка, код на котором должен генерироваться. Целью является проверка корректности исходной программы, а значит необходимо проверить, что все пути из стартовых в конечные состояния входного конечного автомата удовлетворяют заданным ограничениям, и выдать сообщение об ошибке, если это не так. При этом, в случае ошибки необходимо явно указать путь, приводящий к ней. В результате возникает задача поиска путей в графе, удовлетворяющих контекстно-свободным ограничениям: ограничениям, заданным в виде КС-грамматики. Исследования в данной области ведутся активно, однако ряд вопросов остаются открытыми, например, вопрос о возможности использования нисходящих алгоритмов синтаксического анализа для решения задачи поиска путей. В рамках данного исследования предполагается ответить на некоторые из этих вопросов.

en

1.6 Сроки реализации проекта

ru

3 года

2 Содержание проекта

2.1 Цель и задачи проекта

ru
Цель. Разработка эффективного подхода к поиску путей в графе, удовлетворяющих контекстно-свободным ограничениям.

Задачи.

- Разработка алгоритма поиска путей в графах, удовлетворяющих ограничениям, заданным в виде контекстно-свободной грамматики, основанного на алгоритме обобщённого нисходящего синтаксического анализа (GLL).
- Теоретическая оценка временной и пространственной сложности разработанного алгоритма.
- Программная реализация предложенного алгоритма.
- Экспериментальная проверка оценки сложности разработанного алгоритма.
- Разработка механизма обнаружения и диагностики ошибок для разработанного алгоритма.
- Разработка на основе предложенного алгоритма прототипа решения для поиска последовательностей с заданной вторичной структурой в метагеномных сборках.

2.2 Актуальность исследования

ru
Интерес к области растёт, работа показывает, что создание эффективного решения всё еще открытая проблема.

2.3 Направление из Стратегии научно-технологического развития Российской Федерации (при наличии) (выбор из справочника)

ru

2.4 Анализ современного состояния исследований в данной области (приводится обзор исследований в данной области со ссылками на публикации в научной литературе)

ru
Задача поиска путей с контекстно-свободными ограничениями разрешима за полиномиальное

время от размера графа (Barrett C., Jacob R., Marathe M. Formal-language-constrained path problems. 2000 г.). В настоящий момент предложены алгоритмы со сложностью $O(|G| * n^5)$ (Lange M. Model checking propositional dynamic logic with all extras. 2006.), $O(n^3 * m^2)$ (Sevon P., Eronen L. Subgraph queries by context-free grammars. 2008.), $O((|G|*m)+(|G|*n)^3)$ (Hellings J. Conjunctive context-free path queries. 2014.). Здесь $|G|$ — константа, зависящая от грамматики, n — количество вершин во входном графе, m — количество рёбер во входном графе. При этом в качестве алгоритма синтаксического анализа используются такие алгоритмы как CYK или Earley, вычислительная сложность которых в лучшем случае $O(n^3)$ и $O(n^2)$ соответственно, что хуже $O(n)$, достигаемого такими алгоритмами, как GLR, на однозначных грамматиках. Кроме этого, например, CYK требует преобразования грамматики к нормальной форме Хомского, что приводит к её значительному разрастанию и негативно влияет на производительность алгоритма (увеличивая константу |G|).

В работах Annamaa A. et al. "An interactive tool for analyzing embedded SQL queries" (2010г) и E. Verbitskaia, S. Grigorev, and D. Avdyukhin. "Relaxed parsing of regular approximations of string-embedded languages" (2015г), посвящённых обработке встроенных языков, используются алгоритмы семейства LR. Они не требуют преобразования грамматик в нормальную форму Хомского и демонстрируют линейную сложность на однозначных грамматиках. Однако известно, что GLR и RNGLR, используемые в указанных работах, в худшем случае могут демонстрировать более чем кубическую сложность. Кроме того, отсутствуют теоретические оценки сложности предложенных на их основе решений.

В рамках данного проекта предполагается разработка алгоритма поиска путей с KC-ограничениями на основе GLL-алгоритма, который для LL-грамматик показывает линейную сложность и кубическую в худшем случае. Также будет выполнена теоретическая оценка временной и пространственной сложности предложенного алгоритма.

Существенной проблемой также является сложность отладки запросов и анализа результатов их исполнения. Одно из возможных решений данной проблемы предложено в работе P. Hofman and W. Martens "Separability by short subsequences and subwords" (2015).

Предложенный в работе E. Verbitskaia, S. Grigorev, and D. Avdyukhin. Relaxed parsing of regular approximations of string-embedded languages (2015г) алгоритм строит лес вывода для всех корректных путей — структурное представление, которое может быть использовано для отладки и анализа структуры результата. В данной работе предполагается использовать эти результаты и адаптировать их к разрабатываемому алгоритму.

Для работы с метагеномными сборками существуют такие инструменты как Xander, EMIRGE и Reago, однако они не обладают достаточной производительностью и точностью. Кроме того, не все инструменты обрабатывают сборки, представленные в виде графа, и не используют грамматики для описания вторичных структур искомых цепочек. Использование грамматик для описания вторичной структуры достаточно изучено и распространено (Eddy S. R. "Homology searches for structural RNAs: from proof of principle to practical use". 2015; Anderson J. W. J. et al. "Evolving stochastic contextfree grammars for RNA secondary structure prediction". 2012) и используется, например, в инструменте Infernal, обладающем высокой точностью и скоростью работы. Однако данный инструмент не применим к метагеномным сборкам.

В рамках данного исследования предполагается разработать на основе предложенного

алгоритма инструмент, применимый к метагеномным сборкам и использующий грамматики для задания вторичной структуры искомых цепочек.

2.5 Научная новизна проекта (формулируется научная идея, постановка и решение заявленной проблемы)

ru
!!!

2.6 Предлагаемые подходы и методы, их обоснование для реализации цели и задачи проекта (Развернутое описание; форма изложения должна дать возможность эксперту оценить соответствие подходов и методов поставленным целям и задачам проекта)

ru
— Алгоритм обобщённого нисходящего синтаксического анализа GLL. Ранее данный алгоритм не применялся для решения задачи поиска путей в графе. — Построение конечной структуры данных для представления результатов запроса. Данный подход к представлению результата был впервые предложен участниками данной исследовательской группы. Планируется обобщение данного подхода для применения в целях отладки запросов и анализа их результатов. — Применение разработанного алгоритма поиска путей с контекстно-свободными ограничениями для поиска цепочек с заданной с помощью грамматики вторичной структурой в метагеномных сборках. Использование грамматик для описания вторичной структуры широко распространено, однако алгоритм поиска путей с КС-ограничениями ранее не применялся.

2.7 Ожидаемые результаты реализации проекта и их научная и прикладная значимость

ru
— Алгоритм поиска путей в графе, удовлетворяющих КС-ограничениям. Результатом работы алгоритма является конечная структура данных, представляющая результат выполнения запроса. — Теоретическая оценка временной и пространственной сложности предложенного алгоритма. — Программная реализация разработанного алгоритма. — Механизм диагностики ошибок для разработанного алгоритма. — Прототипы программных средств, использующих разработанный алгоритм, для решения прикладных задач. — Оценка применимости разработанного алгоритма к решению задачи биоинформатики по поиску цепочек в метагеномных сборках. — Программные компоненты для решения прикладных задач, использующие разработанный алгоритм.

2.8 Общий план реализации проекта (форма представления информации должна дать возможность эксперту оценить реализуемость заявленного исследования; общий план реализации проекта даётся с разбивкой по годам)

ru

- Первый год
  - Собран и проанализирован набор данных для экспериментального исследования
  - Проведён сравнительный анализ алгоритмов
  - Разреженные матрицы
  - Кратчайший путь
- Второй год
  - Собран и проанализирован набор данных для экспериментального исследования
  - Проведён сравнительный анализ алгоритмов
  - Разреженные матрицы
  - Кратчайший путь
- Третий год
  - Собран и проанализирован набор данных для экспериментального исследования
  - Проведён сравнительный анализ алгоритмов
  - Разреженные матрицы
  - Кратчайший путь

2.9 Ожидаемые научные результаты за первый год реализации проекта (форма изложения должна дать возможность провести экспертизу результатов)

ru

- Собран и проанализирован набор данных для экспериментального исследования
- Проведён сравнительный анализ алгоритмов
- Разреженные матрицы
- Кратчайший путь
- Подготовлены публикации. Результаты представлены на тематических конференциях

2.10 Имеющийся у коллектива научный задел по проекту (указываются полученные результаты, разработанные программы и методы, экспериментальное оборудование, материалы и информационные ресурсы, имеющиеся в распоряжении коллектива для реализации проекта)

ru

- Разработан алгоритм ослабленного синтаксического анализа графов на основе обобщённого LR-анализа (RNGLR), решающий задачу поиска путей, удовлетворяющих контекстно-свободным ограничениям. Алгоритм RNGLR впервые применён для поиска путей.
- Разработана модификация алгоритма обобщённого LL-анализа, использующая таблицы вместо явной генерации кода, что необходимо для его применения в задаче поиска путей. Ранее рассматривались версии обобщённого LL алгоритма, использующие явную генерацию кода.
- Разработан алгоритм поиска путей с контекстно-свободными ограничениями, основанный на алгритме обобщённого LL-анализа (GLL). Алгоритм GLL впервые применён для поиска путей.
- Разработано решение для поиска путей с контекстно-свободными ограничениями, основанное на парсер-комбинаторах. Особенностями данного решения являются: прозрачная интеграция запросов в язык программирования общего назначения, возможность снабжать запросы дополнительными семантическими действиями (напрмиер, дополнительная фильтрация путей). Решение с подобными свойствами предложено впервые.
- Разработан алгоритм поиска путей с контекстно-свободными ограничениями, основанный на опреациях над матрицами. Данное решение позволяет использовать для поиска путей с контекстно-свободными ограничениями массово-переллельные архитектуры (напрмиер, GPGPU). Ранее подобный подход к решению задачи представлен не был.
- Предложена архитектура программного средства, позволяющая использовать разработанный алгоритм для создания инструментов синтаксического анализа встроенных языков. Разработана соответствующая программная платформа. Ранее предлагались инструменты, предназначенные для решения специфичных задач; платформа, позволяющая создавать собственные инструменты, предложена впервые.
- Также нашим коллективом выполнены и успешно защищены кандидатская диссертация и ряд дипломных работ и магистерских диссертаций по тематике проекта.

2.11 Публикации (не более 15) участников коллектива, включая руководителя коллектива, наиболее близко относящиеся к проекту за последние 5 лет (для каждой публикации, при наличии, указать ссылку в сети Интернет для доступа эксперта к аннотации или полному тексту публикации)

ru

1. Nikita Mishin, Iaroslav Sokolov, Egor Spirin, Vladimir Kutuev, Egor Nemchinov, Sergey Gorbatyuk, and Semyon Grigorev. 2019. Evaluation of the Context-Free Path Querying Algorithm Based on Matrix Multiplication. In Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) (GRADES-NDA'19), Akhil Arora, Arnab Bhattacharya, and George Fletcher (Eds.). Ссылка: https://dl.acm.org/citation.cfm?id=3328503

2. Ekaterina Verbitskaia, Ilya Kirillov, Ilya Nozkin, and Semyon Grigorev. 2018. Parser combinators for context-free path querying. In Proceedings of the 9th ACM SIGPLAN International Symposium on Scala (Scala 2018). Ссылка: https://dl.acm.org/citation.cfm?id=3241655

3. Rustam Azimov and Semyon Grigorev. 2018. Context-free path querying by matrix multiplication. In Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) (GRADES-NDA '18), Akhil Arora, Arnab Bhattacharya, George Fletcher, Josep Lluis Larriba Pey, Shourya Roy, and Robert West (Eds.). Ссылка: https://dl.acm.org/citation.cfm?id=3210264

4. Semyon Grigorev and Anastasiya Ragozina. 2017. Context-free path querying with structural representation of result. In Proceedings of the 13th Central & Eastern European Software Engineering Conference in Russia (CEE-SECR '17). Ссылка: https://dl.acm.org/citation.cfm?id=31661

5. Marina Polubelova, Sergey Bozhko, Semyon Grigorev. 2016. Certified Grammar Transformation to Chomsky Normal Form in F*. Proceedings of the Institute for System Programming. Ссылка: `http://www.ispras.ru/en/proceedings/isp_28_2016_2/isp_28_2016_2_127/`

6. Polubelova M. I., Grigor'ev S. V. Lexical analysis of dynamically generated string expressions //Sistemy i Sredstva Informatiki [Systems and Means of Informatics]. – 2016. – Т. 26. – №. 2. – С. 43-62. Ссылка: `http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=ssi&paperid=461&option_lang=eng`

7. Verbitskaia E., Grigorev S., Avdyukhin D. Relaxed Parsing of Regular Approximations of String-Embedded Languages //International Andrei Ershov Memorial Conference on Perspectives of System Informatics. – Springer International Publishing, 2015. – С. 291-302. Ссылка: `http://link.springer.com/chapter/10.1007/978-3-319-41579-6_22`

8. Semen Grigorev, Ekaterina Verbitskaia, Andrei Ivanov, Marina Polubelova, and Ekaterina Mavchun. 2014. String-embedded language support in integrated development environment.

In Proceedings of the 10th Central and Eastern European Software Engineering Conference in Russia (CEE-SECR '14). ACM, New York, NY, USA, , Article 21 , 11 pages. Ссылка: http://dl.acm.org/citation.cfm?id=2687247&CFID=498663671&CFTOKEN=45521518

9. Grigor'ev S. V., Ragozina A. K. Generalized table-based LL-parsing //Sistemy i Sredstva Informatiki [Systems and Means of Informatics]. – 2015. – Т. 25. – №. 1. – С. 89-107. Ссылка: http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=ssi&paperid=395&option_lang=eng