Использование формальных грамматик и скусственных нейронных сетей для анализа вторичной структуры генмных и протеомных последовательностей

Семён Григорьев

14 марта 2019 г.

1 Сведения о проекте

1.1 Название проекта

ru
Использование формальных грамматик и скусственных нейронных сетей для анализа вторичной структуры генмных и протеомных последовательностей

en

1.2 Направление из Стратегии НТР РФ

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

1.3 Обоснование соответствия тематики проекта направлению из Стратегии НТР РФ: необходимо кратко сформулировать научную проблему (проблемы) и конкретные задачи в рамках выбранного направления, решению которых будет посвящен проект, обосновать соответствие проекта направлению

ru
Проект посвящён разработке методов анализа вторичной структуры цепочек с использованием формальных граммтик и искусственных нейронных сетей.

Врамках проекта ставятся следующие задачи. Во-первых, необходимо сформулировать общие принципы построения формальных граммтик, описывающих вторичную структуру различных типов цепочек. Во-вторых, разработать алгоритмы синтаксического анализа, пригодные для высокопроизводительной обработки реальных цепочек на основе построенных граммтик. В-третьих, необходимо исследовать возможности совмещения алгоритмов синтаксического анализа с искусственными нейронными сетями (ИНС) для решения таких прикладных задач, как, напрмиер, поиск цепочек с аналогичными вторичными структурами.

Решение этих задач позволит создавать решения, применимые в таких областях, как анализ сообществ микроорганизмов, анализ генетической информации, поиск новых лекарственных препаратов.

При анализе сообществ микроорганизмов, что часто необходимо при диагностике различных заболеваний, один из подходов заключается в поиске маркерных цепочек, некоторые из которых обладают характерной вторичной структурой (например, 16s РНК), с последующей их классификацией.

Вместе с тем, анализ структурных особенностей белковых, геномных и других последовательностей необходим для эффективного поиска новых лекарственных препаратов, в том числе антибактериальных. Так, например, поиск новых антибактериальных препаратов часто основан на поиске соединений, структурно аналогичных уже известным, обладающим антибактериальными свойствами. В других же случаях требуется анализ структуры цепочки-мишени для более прицельного поиска препарата. Данные подходы могут совмещаться.

en

1.4 Ключевые слова (приводится не более 15 терминов)

ru
Формальные граммтики, синтаксический анализ, параллельные алгоритмы, вторичная структура, РНК, генмноые последовательности, белки, протеомные последовательности, метагеномная сборка, искусственные нейронные сети.

en

1.5 Аннотация проекта

ru
Различные молекулярные соединения, такие как белковые молекулы или ДНК/РНК-молекулы, часто рассматривают как цепочки, состоящие из последовательно соединённых более простых элементов-оснований (напрмиер, аминокислот или нуклиотидов). При этом, кроме последовательных связей между основаниями образуются также дополнительны — вторичные — связи, которые задают вторичную структуру цепочки. Известно, что вторичная структура

некоторых цепочек обладает характерными особенностями. Классический пример — вторичная структура транспортной РНК: первичная структура (последовательность нуклиотидов) может сильно различаться даже у достаточно близких орагизмов, однако некторые особенности вторичной структуры (характерный "крест") наблюдаются практически у всех орагнизмов.

Также выяснено, что часто вторичная структура несёт существенную информацию о функциональной роли той или иной цепочки.

Всвязи с этим, важной задачей является разработка формальных методов для описания вторичной структры и её особенностей. При этом важно, чтобы полученные формальные модели позволяли создавать эффективные решения для прикладных задач, требующих анализа вторичной структруры. Напрмиер, один из самых точных подходов к анализу вторичной структуры основан на анализе энергии межмолекулярного взаимодействия. Однако данный подход трудно применим на практике при анализе больших объёмов данных ввиду его высокой вычислительной сложности.

Проект посвящён исследованию применимости формальных грамматик в качестве формальной модели для описания вторичной структуры различных типов цепочек, например, геномных или белковых, а также разработке соответствующих алгоритмов, позволющих строить применимые на практике решения.

С одной стороны, применение результатов теории формальных языков для анализа биологических последовательностей исследуется достаточно давно. В качестве примера можно привести результаты Сина Эдди и инстумент Infernal. С другой, граммтики, в основном, применялись для описания первичной структуры цепочек: предпринимались попытки анализировать цепочки как текст над некоторым алфавитом (набором оснований). Применение фрмальных граммтик для описания вторичной структуры исследовано слабо. Вместе с этим, появились новые результаты в области формальных языков, предложены новые типы граммтик (например, конъюнктивные), обладающие высокой выразительной силой и при этом позволяюще построение эффективных алгоритмов синтаксического анализа. Применимость данных типов граммтик для описания вторичной структуры исследовано недостаточно. Таким образом, планируется получение новых результатов, связанных с применением новых типов граммтик для описания вторичной структуры цепочек.

Также будет исследована возможность применения обыкновенных, не вероятностных, граммтик для описания вторичной структуры. Современные подходы предполагают использование верочтностных грамматик для описания цепочек, что связано ст тем, что реальные данные содержат большое количество мутаций и привнесённых шумов, что делает невозможным построение точных моделей. В данном исследовании предлагается изучить вопрос использования обыкновенных грамматик, а в качестве вероятностной модели использовать искусственную нейронную сеть, что является новым подходом к использованию граммтик.

en

1.6 Ожидаемые результаты и их значимость

ru

В результате изучения применимости различных типов граммтик для описания вторичной структуры будут, вопервых, выявлены основные принципы построения грамматик для конкретнх типов цепочек, а также предложены конкретные граммтики для некторых типов цепочек и некторых задач.

Предполагается, что будут разработаны новые алгоритмы синтаксического анализа, учитывающие особенности решаемой задачи, такие как, с одной стороны, свойства используемых граммтик (сильная неоднозначность), и с другой стороны возможности современного аппаратного обеспечения, такие как массовый парарллелизм.

Также будет сформулирован метод совмещения обыкновенных граммтик и ИНС для решения задач анализа вторичной структуры цепочек.

В совокупности данные результаты должны позволить создавать прикладные решения, применимые как в исследовательских, так и в прикладных задачах биологии и медицины.

en

2 Содержание проекта

2.1 Научная проблема, на решение которой направлен проект

ru

Создание формальной модели для описания и изучения вторичной структуры последовательносткй, обладающей хорошими формальными свойствами, но при этом позволяющей создавать эффективные прикладные решения на своей основе.

Разработка алгоритмов стнтаксического анализа, учитывающих особенности решаемых задач. Сильно неоднозначные, болшой объём данных, поиск подстроки.

Большой объём данных, возникающий в прикладных задачах, выдвигает дополнительные требования к алгоритмичеким решениям, касающиеся, в первую очередь, необходимости получать высокопроизводительные решения. Разработка параллельных алгоритмов, эффективно использующих возможности современной вычислительной техники, может решить эту проблему.

en

2.2 Научная значимость и актуальность решения обозначенной проблемы

ru

Поиск маркерных последовательностей для обнаруждения организмов, в том числе новых, ранее не изученных, поиск лекарств (анитибактериальных) — актуальные вопросы. Современные методы решения во многом основываются на анализе вторичной структруры различными методами. Формальные методы описания

Алгоритмы стнтаксического анализа. Постановка новых задач в области алгоритмов синтаксического анализа и теории формальгных языков.

Классификация, обнаружение и т.д.

en

2.3 Конкретная задача (задачи) в рамках проблемы, на решение которой направлен проект, ее масштаб и комплексность

ru

Изучение применимости обыкновенных (не вероятностных) контекстно-свободных и конъюнктивных граммтик для анализа вторичной структ грмматик. Предполагается, что будет вестись поиск новых подходов, позволяющих построить не только обозримые формальные модели, но и эффективные на практике решения по анализу вторичной структуры. Одним из направлений будет совмещение методов теории формальных языков и синтаксического анлиза с подходами машинного обучения.

Также планируется построение граммтик для конкретных задач, имеющих важное прикладное значение, таких как, например, поиск маркерных последовательностей.

Кроме того, планируется разработка параллельных алгоритмов синтаксического анализа, специализированных для работы с сильно неоднозначными граммтиками и решения специфичных задач, таких как поиск подстроки с заданной вторичной структурой. Предполагается, что разработанные алгоритмы будут эффективно использовать возможности современного аппаратного обеспечения, такие как массовый парарллелизм.

en

2.4 Научная новизна исследований, обоснование достижимости решения поставленной задачи (задач) и возможности получения запланированных результатов

ru

Новые алгоритмы и новые типы граммтик. Подход с описанием вторичной, а не первичной

структуры.

Текстовый анализ

Грамматики, описывающе первичную струкруту.

Метод совмещения синтаксического анализа и искусственных нейронных сетей для анализа вторичной структуры не изучен.

Вторичная структура — через энергии связи — точный, но очень ресурсоёмкий подход.

Сложные элементы вторичной структуры, такие как псевдоузлы, не выразимы в терминах хороши изученных классов (контекстно-сводобных и регулярных).

Существование наработок, решающих демонстрационные задачи. Существование активных исследований в данной области в настоящий момент.

en

2.5 Современное состояние исследований по данной проблеме, основные направления исследований в мировой науке и научные конкуренты

ru
Тренировка вероятностных грамматик — да. Переложить это на искуственные нейронные сети — нет.

Грамматики для работы с первичной структурой — да. Граммтики для описания вторичной структуры — нет.

Применение конъюнктивных граммтик исследовано крайне слабо, но активно развивается (2? работы).

Использование формальных граммтик и алгоритмов синтаксического анализа для изучения вторичной структуры белков в настоящее время активно иследуется группой (Витольд).

Использование формальных граммтик в качестве теоретической модели для описания вторичной структуры РНК активно исследуется (Девушка с конфы)

БОльшое количество исследований, в том числе практические инструменты, использующие грамматики.

en

2.6 Предлагаемые методы и подходы, общий план работы на весь срок выполнения проекта и ожидаемые результаты

ru
На первом этапе планируется выявить общие принципы построения формальных грамматик

для описания вторичной структуры геномных и протеомных последовательностей. Предстоит ответить на такое вопросы, как какие типы формальных грамматик необходимо использоавть, какие особенности вторичной структуры необходимо учитывать при решении прикладных задач и, следовательно, описывать. Принципы основаны на особенностях вторичной структуры и способе их задания в виде граммтики. Классы грамматик и особенности вторичной структуры.

Разработать алгоритмы синтаксического анализа. Сильно неоднозначные гаммтики, что не характерно для языков програмирования, для которых разрабатывались многие алгоритмы.

Развить метод совмещения синтаксического анализа и искусственных нейронных сетей для анализа вторичной структуры, предложенный руководителем проекта. Планируется изучить разлчные типы и архитектуры искусственных нейронных сетей, с целью выявления наиболее подходящей для рассматриваемого применения. Среди типов особый интерес представляют свёрточные сети, позволяющие обрабатывать результат синтаксического анализа, представленный в виде изображения, что позволит, напрмиер, упросить ршение задачи нормировки данных, применив стандартные решения из области цифровой обработки изображений. Также необходимо изучить битовые сети, так как битовый вектор — наиболее естественное представление результатов синтаксического анализа, а данный тип сетей предназначен для обработки таких данных.

Далее планируется провести ряд экспериментов на реальных данных — базах цепочек, имеющихся в открытом доступе. Цель экспериментов — проверить практическую применимость разработанных методов и алгоритмов. Предполагается, что будут решаться задачи классификации цепочек по различным признакам. Например, по функциям для белковых последовательностей, или по тому, является ли цепочка химерой, для маркерных РНК-последовательностей. На данном шаге также будет вестись подбор грамматик для конкретных задач: несмотря на то, что предполагается наличие общих принципов построения таких грамматик, решение конкретной задачи может потребовать значительных уточнений грамматики для получения наилучшего результата.

В результате работы будут получены !!! Разработан алгоритм, подход, границы применимости.

2019-2020

Предплоагается исследовать границы применимости подхода, сформулированного руководителем проекта в работе

Разработка гармматик для анализа вторичной структуры РНК-последовательностей.

Эксперименты по обнаружению маркерных цепочек.

2020-2021

Белки.

Предсказание вторичной структуры.

en

2.7 Имеющийся у руководителя проекта научный задел по проекту, наличие опыта совместной реализации проектов

ru

Руководитель проекта обладает опытом в разработке и исследовании алгоритмов синтаксического анализа, и их применении в различнах областях, что подтверждается соответствующими статьями (Grigorev, Ragozina, "Context-free path querying with structural representation of result SECR-2017; Azimov, Grigorev, "Context-free path querying by matrix multiplication GRADES-NDA-2018; Verbitskaia, Kirillov, Nozkin, Grigorev, "Parser combinators for context-free path querying Scala-2018)

В том числе, у руководителя имеется опыт применения формальных граммтик и алгоритмов синтаксического анализа для решения задач в области биологии (биоинформатики), что подтверждается выступлениями на тематических конференциях Biata-2017/2018, BIOINFORMATICS-2019.

Кроме того, руководителем был предложен метод совмещения формальных граммтик и ИНС для анализа вторичной структруры, который аредполагается развивать в рамках данного исследования. Метод был изложен в статье "The Composition of Dense Neural Networks and Formal Grammars for Secondary Structure Analysis"и представлен на конференции BIOINFORMATICS-2019.

Руководитель принимал успешное участие в совместной работе над проектами в рамках грантов РФФИ (15-01-05431 и 18-01-00380), Фондоа содействия развитию малых форм предприятий в технической сфере (программа УМНИК, проекты N 162ГУ1/2013 и N 5609ГУ1/2014), а также является руководителем научной группы, в соавторстве с участниками которой опубликованы указанные выше и некоторые другие работы.

en

2.8 Перечень оборудования, материалов, информационных и других ресурсов, имеющихся у руководителя проекта для выполнения проекта

ru

Ресурсы, необходимые для выплнения проекта, такие как базы данных с биологическими последовательностями (базы маркерных цепочек, базы белковых последовательностей) имеются в открытом доступе в сети Интернет. Использование иных особых ресурсов не предполагается.

en

2.9 План работы на первый год выполнения проекта

ru

Эксперименты с 16s и химерами. Эксперименты с белками. Конъюнктивные граммтики. Работа над агоритмами синтаксического анализа

en

2.10 Ожидаемые в конце первого года конкретные научные результаты

ru

Граммтики

Алгоритм.

Парсер.

en

2.11 Перечень планируемых к приобретению руководителем проекта за счет гранта Фонда оборудования, материалов, информационных и других ресурсов для выполнения проекта

ru

Не более !!! тыс. рублей ежегодно будет тратиться на поездки с докладами на конференции. Расходов на оборудование не предполагается.

en