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

Семён Григорьев

15 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

НЗ 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 Шона Эдди (Sean Eddy) 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 с 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

Приект посвящён разработке формальных моделей для описания вторичной структуры бологичесикх цепочек и алгоритмов для решения задач анализа струкрты, на них основанных. Таким образом, проблемы, решаемые в проете, лежат в области биоинформатики.

Качественное решение прикладных задач невозможно без существования удачных формальных моделей, позволяющих не только успешно осуществлять поиск решения поставленной задачи, но и, с одной стороны, формализовать постановку задачи, с другой, получить механизмы оценки качества решения. При этом, удачная формальная модель должна совмещать в себе два важных качества: с одной стороны быть долстаточно выразительной, с другой, позволять эффективные реализации алгоритмов для решения прикладных задач. Необходимо учитывать, что одной из ключевых особенностей прикладных задач в данной области является большой объём обрабатываемых данных.

Поиск такой модели для описания структуры биологических цепочек (напрмиер, белковых или геномных) ведётся на протяжении длительного времени. С одной стороны, существуют модели, пытающиеся максимально полно учесть химические и физические законы взаимодействия между молекулами (например, модели, основанные на анализе энергии межмолекулярных связей). Данные модели точны, однако громоздки как с точки зрения формальных рассуждений, так и с точки зрения эффективной реализации соответствующих алгоритмов, которые оказываются очень ресурсоёмкими и малопригодными для обработки

реальных данных. С другой стороны, существуют модели, рассматривающие такие цепочки как последовательный набор оснований и трактующие их, например, как строки в некотором алфавите (например, {A,C,G,T}). Такие модели позволяют применять эффективные алгоритмы обработки строк, однако являются недостаточно точными для решения многих задач, так как не учитывают информацию о структурных особенностях цепочек (например, о вторичной структуре).

Один из подходов, активно исследуемых в настоящее время использует формальные граммтики для описания свойсв цепочек. Преимуществом является возможность привлечения общырных результатов теории формальных языков, развиающейся длительное время. Теория формальных языков, с одной стороны, может предложить богатую теоретическую базу, с другой, эффективные алгоритмы. При этом, выбирая класс граммтик можно подбирать баланс между выразительностью и возможность получить эффективную реализацию.

В рамках данного исследования планируется построение и изучение теоретических и практических свойств моделей, использующих формальные граммтики для описания вторичной структуры цепочек.

en

2.2 Научная значимость и актуальность решения обозначенной проблемы

ru

Удачные с теоретической точки зрения модели позволяют эффективно рассуждать о свойствах изучаемых объектов, выдвигать и проверять новые научные гипотезы, прогнозировать границы разрешимости прикладных задач. Например, таких важных задач, как поиск маркерных последовательностей для обнаруждения организмов, в том числе новых, ранее не изученных, или поиск лекарств (в том числе анитибактериальных).

При этом, необходимо найти такие модели, которые при должной выразительности будут позволять реалиовывать эффективные алгоритмические и прикладные решения. Построение таких моделей востребовано ввиду большого объёма данных, требующих обработки при решении прикладных задач.

Кроме этого, в ходе исследования в данной области могут возникнуть новые задачи в области алгоритмов синтаксического анализа и теории формальных языков, что будет способствовать развитию данных облатей.

en

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

ru

В рамках изучения формальных граммтик в качестве средства описания вторичной структуры планируется изучение применимости обыкновенных (не вероятностных) контекстно-свободных и конъюнктивных граммтик для анализа вторичной структуры геномных и белковых цепочек. В частности, планируется построение граммтик для конкретных задач, имеющих важное прикладное значение, таких как, например, поиск маркерных последовательностей.

Вместе с этим планируется построение алгоритмов, позволяющих проводить синтаксический анализ соответствующих классов языков и допускающих реализации, эффективно использующие возможности современного аппаратного обеспечения, такие как массовый парарлелизм и распределённые вычисления. Кроме того, разработываемые алгоритмы должны быть специализированы для работы с сильно неоднозначными граммтиками и решения специфичных задач, таких как поиск подстроки с заданной вторичной структурой.

Также планируется вести 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.

en

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

ru

Поиск эффективных моделей для описания вторичной структуры цепочек активно ведётся в настоящее время. Сформулированные задачи, с одной стороны, опираются на имеющиеся результаты, которые говорят о том, что формальные грамматики могут применяться для решения задач анализа биологических цепочек, с другой стороны, нацеены на улучшение существующих моделей. Это, с одной стороны, позволяет говорить о возможности получения запланированных результатов, а с другой, о том, что любое разумное улучшение, как в выразительном плане, так и в смысле возможности построения эффективных реализаций, будет новым результатом. Стоит отметить, что применение таких классов грамматик, как конъюнктивные, в данной области изучено крайне слабо, а у руководителя есть опыт применеиния таких граммтики и разработки алгоритмов стнтаксического анализа для них.

Кроме того, часть задач сформулирована ранее, представлена и обсуждалась на международных конференциях, что говорит об их актуальности и возможности гарантировать новизну полученных результатов.

У руководителя проекта есть опыт исследований в области формальных граммтик и алгоритмов синтаксического анализа, что должно помочь решить поставленне задачи, так как они лежат в этой же области. Кроме того, руководителем предложен метод совмещения формальных граммтик и методов машинного обучения для решения задач анализа вторичной структуры, который был успешно представлен на международной конференции.

en

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

ru

Применение формальных граммтик для анализа биологических цепочек — одна из активно развивающихся областей в биоинформатике.

Большое количество результатов, многие из которыз уже стали классическими, и воплощёны в широко распространённом инструменте Infernal, использующем вероятностные грамматики для анализа структуры РНК-последовательностей, получены группой под руководством Шона Эдди (Sean Eddy) в США. Даннвя группа является ведущей в этой области и активно занимается исследованиями в этом направлении и в настоящее время.

Применение конъюнктивных граммтик для описания структуы геноных последовательностей исследовано крайне слабо. В настоящий момент опубликована однаработа Райана Цир-Фогеля (Ryan Zier-Vogel, "RNA pseudoknot prediction through stochastic conjunctive grammars"), в которой для предсказания вторичной структуры РНК используются вероятностные конъюнктивные грамматики.

Исследование возможностей использования формальных граммтик и алгоритмов синтаксического анализа для изучения вторичной структуры белков в настоящее время активно иследуется группой под руководством Витольда Дирки (Witold Dyrka) в Польше.

Изучение применимости формальных граммтик в качестве теоретической модели для описания вторичной структуры РНК и исследование теоретических свойств этой модели активно ведётся Микелой Квадрини (Michela Quadrini) в Италии.

При этом, наиболее исследованным является применение вероятностных грамматик и алгоритмов их построения и работы с ними. Использование обыкновенных граммтик в сочетании с ИНС в качестве вероятностной модели исследовано крайне слабо.

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 такие вопросы, как какие типы формальных граммтик необходимо использоавть, какие особенности вторичной структуры необходимо учитывать при решении прикладных задач и, следовательно, описывать. Для этого будут привлечены методы теории формальных языков, позволяющие рассуждать о выразительных свойствах граммтик. При этом необходимо учитывать вычислительную сложность алгоритмов синтаксического анализа и возможность реализации параллельных алгоритмов.

Далее планируется разработать параллельный алгоритм синтаксического анализа для выбранного класса граммтик. Предполагается, что будут исследованы возможности эффективного использования массово-параллельных архитекур. При этом, алгоритм должен быть адаптирован к использованию сильно неоднозначных гаммтики, и к решению задачи поиска подстроки в строке с заданной структурой. Необходимо также будет учеть и другие особенности граммтики, такие как сравнительно небольшой терминальный алфавит и небольшое количество правил в самой граммтике, по сравнениею с граммтиками, возникающими в лингвистике или при анализе языков программирования. Работа на данном этапе потребует привлечения методов параллельного программирования, теории построения и анализа алгоритмов. Построение алгоритма будет вестись на основе алгоритмов синтаксического анализа, использующих матричные операции, так как такие операции хорошо поддаются распараллеливанию на современном аппаратном обеспечении. Ключевым является алгоритм Валианта, а также его модификация, предложенная Охотиным и адаптированная для работы с конъюнктивными граммтиками.

Следующий шаг будет посвящён развитию метода 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 ИНС особый интерес представляют свёрточные сети. Предполагаемый алгоритм синтаксического анализа на выходе будет строить набор булевых матриц, которые можно трактовать как слои изображения. Это позволит обрабатывать результат синтаксического анализа как 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. Возможно, применение такого типа ИНС позволит уменьшить объём необходимой памяти и упростить процедуру обучения.

После этого 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 граммтик, решение конкретной задачи может потребовать значительных уточнений грамматики для получения наилучшего результата.

2019-2020

Разработка гармматик для анализа вторичной структуры РНК-последовательностей.

Разработка параллельного алгоритма синтаксического анализа для решения задачи поиска подстроки с заданной в виде грамматики структурой.

Проведение экспериментов по обнаружению маркерных цепочек.

2020-2021

Исследование применимости различных типов ИНС для анализа результатов синтаксического анализа.

Доработка метода совмещения синтаксического анализа и ИНС для анализа вторичной структуры биологических цепочек с учётом результатов предыдущего шага.

Разработка граммтики для анализа вторичной структуры белков с использоваием предложенного метода, проведение соответствующих экспериментов на реальных данных.

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

Сопставление особенностей вторичной структуры маркерных цепочек с классами формальных граммтик, необходимых для выражения таких особенностей. Построение обыкновенных граммтик, в том числе конъюнктивных, описывающих характерные особенности вторичной структры маркерных цепочек. Формулирование основных принципов построения таких грамматик. Проведение ряда экспериментов на реальных данных для оценки практической значимости полученных граммтик. Для этого планируется провести анализ и предварительную подготовку данных, имеющихся в открытом доступе.

Разработка параллельного агоритма синтаксического анализа, адаптированного для решения задачи поиска подстроки с заданной вторичной срукткрой в строке. Оформление результатов и их публикация.

en

2.10 Ожидаемые в конце первого года конкретные научные результаты

ru

Предложены обыкновенные (не вероятностные) граммтики для описания особенностей вторичной структуры маркерных цепочек. Проведён ряд экспериментов по использованию данных граммтик для решения прикладных задач, таких как поиск и классификация цепочек. По итогам данного этапа, две работы будут представлены на конференциях и опубликованы в сборниках докладов, индексируемом в scopus.

Разработан параллельный алгоритм синтаксического анализа, адаптированный для решения задачи поиска подстроки с заданной вторичной струтурой в строке. Разработанный алгоритм представлен на конференции и опубликован в сборнике материалов конференции, индексируемом в scopus.

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