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

Семён Григорьев

16 марта 2019 г.

1 Сведения о проекте

1.1 Название проекта

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

en
Utilization of formal grammars and artificial neural networks for secondary structure analysis of the genomic and proteomic sequences

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 использова-

нием формальных грамматик и искусственных нейронных сетей.

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

Часто при диагностике различных заболеваний необходимо проводить анализ сообществ микроорганизмнов. Один из подходов к такому анализу 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

The goal of the research is to find new methods for secondary structure analysis of biological sequences by using formal grammars and artificial neural networks.

The research includes the following tasks. First of all, it is necessary to formulate general principles for creation of the grammars which specify the secondary structure of different types of sequences. The second task is to create a parsing algorithm suitable for high-performance real data processing with respect to the grammar created at the previous step. The third task is to investigate the ways to compose parsing algorithms with artificial neural networks (ANN) in order to create a tool which can be used for sequences analysis such as finding sequences with similar secondary structure.

When these tasks are done, it will be possible to create solutions for microbiota analysis, genomic information analysis, and the development of new drugs.

Medical diagnosis often includes microbiota analysis. One of the possible ways to do it is to find marker sequences and use it for classification. It is important that some types of markers have a specific secondary structure.

Structural analysis of proteomic, genomic and other types of sequences plays an important role for drugs development. In some cases, drugs development is based on searching for the molecules which have the structure similar to the structure of the known drugs with specific properties. In other cases, analysis of the structure of the target sequence make the search more focused. These approaches can be combined.

1.4 Ключевые слова (приводится не более 15 терминов)

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

en
Formal grammars, syntax analysis, parsing, parallel algorithms, secondary structure, RNA, genomic sequences, proteins, proteomic sequences, metagenomic assembly, artificial neural network

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 (например, аминокислот или нуклеотидов). При этом, кроме последовательных связей между основаниями образуются также дополнительные — вторичные — связи, которые задают вторичную структуру цепочки. Вторичная структура некоторых цепочек обладает характерными особенностями. Классический пример — вторичная структура транспортной РНК: первичная структура (последовательность нуклеотидов) может сильно различаться даже у достаточно близких организмов, однако некоторые особенности вторичной структуры (характерный "крест") наблюдаются практически у всех организмов. Вторичная структура часто несёт существенную информацию о функциональной роли той или иной цепочки.

В связи с этим, важной задачей является разработка формальных методов для описания вторичной структуры и её особенностей. При этом важно, чтобы полученные формальные модели позволяли создавать эффективные решения для прикладных задач, требующих анализа вторичной структуры. Например, один из самых точных подходов к анализу вторичной структуры основан на анализе энергии межмолекулярного взаимодействия. Однако данный подход обладает высокой вычислительной сложностью, из-за чего он трудно применим на практике при анализе больших объёмов данных.

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, позволяющих строить применимые на практике решения.

Применение результатов теории формальных языков для анализа биологических последовательностей исследуется достаточно давно. В качестве примера можно привести результаты Шона Эдди (Sean Eddy) и инструмент Infernal. В основном, однако, грамматики применялись для описания первичной структуры цепочек: предпринимались попытки анализировать цепочки как текст над некоторым алфавитом (набором оснований). Применение формальных грамматик для описания вторичной структуры исследовано слабо. Вместе с этим, появились новые результаты в области формальных языков, предложены новые типы грамматик

(например, конъюнктивные), обладающие высокой выразительной силой и при этом позволяющие построение эффективных алгоритмов синтаксического анализа. Применимость данных типов грамматик для описания вторичной структуры исследована недостаточно. Таким образом, планируется получение новых результатов, связанных с применением новых типов грамматик для описания вторичной структуры цепочек.

Также будет исследована возможность применения обыкновенных, не вероятностных, грамматик для описания вторичной структуры. Современные подходы предполагают использование вероятностных грамматик для описания цепочек: 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 грамматик.

en

Some types of molecules, such as RNA/DNA or proteins, are often treated as sequences of the elements called bases (amino acids or nucleotides, for example). Within sequential connections between bases, there are also additional — secondary — connections which specify the secondary structure of the sequence. The secondary structure of some types of sequences contains specific features. The well-known example is a structure of tRNA: primary structures (the sequence of nucleotides) vary from one organism to another, but all of them share a typical cloverleaf structure. Moreover, structure often characterizes the sequences functions.

We can conclude, that the development of formal methods for secondary structure description and analysis is an important task (problem). Note that it is necessary to develop formal methods which can be a base for efficient applied solutions. Although the energy-based approach to secondary structure analysis is the most accurate method, it requires huge computational resources for real data analysis which makes it inapplicable for the real-world problems.

Application of the Formal language theory to biological sequences analysis has a long history. The most prominent are the results of Sean Eddy and the Infernal tool. Most tools only use formal grammars for primary structure description: sequences are treated as a text over the alphabet of bases. How to use formal grammars for secondary structure description is not yet well understood. Among the recent results in formal language theory are the new types of grammars (such as conjunctive grammars) which are more expressive than the context-free grammars, and there are also efficient parsing algorithms with respect to these classes. However the applicability of such grammars for secondary structure specification requires more research. We plan to get new results on the application of these types of grammars for secondary structure description.

Also, we plan to investigate the applicability of ordinary grammars (as opposed to probabilistic) for secondary structure analysis. Well-known approaches utilize probabilistic grammars for sequences analysis because real data contains a lot of mutations and noise. This renders the exact modeling impossible and this is why we should use probabilistic approaches. In this research, we plan to utilize ANNs as a probabilistic model but use ordinary grammars. This is a new way to use grammars for structure analysis.

1.6 Ожидаемые результаты и их значимость

ru

В результате изучения применимости различных типов грамматик для описания вторичной структуры будут выявлены основные принципы построения грамматик для конкретных типов цепочек, а также предложены конкретные грамматики для некоторых типов цепочек и задач.

Предполагается, что будут разработаны новые алгоритмы синтаксического анализа, учитывающие особенности решаемой задачи, а именно свойства используемых грамматик (сильная неоднозначность) и возможности современного аппаратного обеспечения, такие как массовый параллелизм.

Также будет сформулирован метод совмещения обыкновенных грамматик 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

We expect to get the following results.

We will formulate general principles for creation of the grammars for specific sequences.

We will create new parsing algorithms. These algorithms will be tuned for the problem in question (for example, they should handle highly ambiguous grammars). Also, they will utilize modern massively-parallel hardware effectively.

We will propose the method to compose ordinary grammars and ANNs for secondary structure analysis.

These results as a whole should provide a solid base for solutions suitable both for research and applied tasks in biology and medicine.

2 Содержание проекта

2.1 Научная проблема, на решение которой направлен проект

ru

Проект посвящён разработке формальных моделей для описания вторичной структуры биологических цепочек и алгоритмов для решения задач анализа структуры, на них основанных. Таким образом, проблемы, решаемые в проекте, лежат в области биоинформатики.

Качественное решение прикладных задач невозможно без удачных формальных моделей. Хорошая модель позволяет не только искать решение поставленной задачи, но и формализовывать постановку задачи и предоставлять механизм оценки качества решения. Удачная формальная модель должна совмещать в себе два важных качества: быть достаточно выразительной и позволять эффективные реализации алгоритмов для решения прикладных

задач. Необходимо учитывать, что большой объём обрабатываемых данных является одной из ключевых особенностей прикладных задач в данной области.

Поиск хорошей модели для описания структуры биологических цепочек (например, 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, но громоздки как с точки зрения формальных рассуждений, так и 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 грамматики для описания свойств цепочек. Преимуществом является возможность привлечения обширных результатов теории формальных языков, развивающейся длительное время. Теория формальных языков может предложить как богатую теоретическую базу, так и эффективные алгоритмы. Выбирая класс грамматик можно подбирать баланс между выразительностью и эффективностью реализации.

В рамках данного исследования планируется построение и изучение теоретических и практических свойств моделей, использующих формальные грамматики 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 обнаружения организмов, в том числе новых, ранее не изученных, или поиск лекарств (в том числе антибактериальных).

При этом, необходимо найти такие модели, которые при должной выразительности будут позволять реализовывать 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 спо-

собствовать развитию данных областей.

en

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

ru

В рамках изучения формальных грамматик в качестве средства описания вторичной структуры планируется изучение применимости обыкновенных (не вероятностных) контекстно-свободных и конъюнктивных грамматик для анализа вторичной структуры геномных и белковых цепочек. В частности, планируется построение грамматик для конкретных задач, имеющих важное прикладное значение, таких как 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 параллелизм и распределённые вычисления. Кроме того, разрабатываемые алгоритмы должны быть специализированы для работы с сильно неоднозначными грамматиками 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 анализа 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 опираются на имеющиеся результаты, которые говорят о применимости формальных грамматик для решения задач анализа биологических цепочек. Также они нацелены на улучшение существующих моделей. С одной стороны, это 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 применения таких грамматик и разработки алгоритмов синтаксического 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 грамматик и алгоритмов синтаксического анализа, что поможет решить поставленные задачи. Кроме того, руководителем предложен метод совмещения формальных грамматик 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 грамматик для анализа биологических цепочек — одна из активно развивающихся областей в биоинформатике.

Большое количество результатов, многие из которых 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. Данная группа является ведущей в этой области и активно занимается исследованиями в этом направлении и в настоящее время.

Применение конъюнктивных грамматик для описания структуры геномных последовательностей исследовано крайне слабо. В настоящий момент опубликована одна работа 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 грамматик и алгоритмов синтаксического анализа для изучения вторичной структуры белков в настоящее время активно исследуется группой под руководством Витольда Дирки (Witold Dyrka) в Польше.

Изучение применимости формальных грамматик 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 грамматик 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 определить, какие типы формальных грамматик необходимо использовать и 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 грамматик. При этом необходимо учитывать вычислительную сложность алгоритмов синтаксического анализа и возможность реализации параллельных алгоритмов.

Далее планируется разработать параллельный алгоритм синтаксического анализа для выбранного класса грамматик. Предполагается, что будут исследованы возможности эффективного использования массово-параллельных архитекур. Алгоритм должен быть адаптирован к использованию сильно неоднозначных грамматик и к решению задачи поиска подстроки с заданной структурой в строке. Необходимо также будет учесть и другие особенности грамматики, такие как небольшой терминальный алфавит и небольшое количество правил, по сравнению с грамматиками, 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 грамматиками.

Следующий шаг будет посвящён развитию метода совмещения синтаксического анализа и искусственных нейронных сетей для анализа вторичной структуры, предложенного руководителем проекта. Планируется изучить различные типы и архитектуры искусственных нейронных сетей с целью выявления наиболее подходящей. 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 упросить решение задачи нормировки данных за счет применения стандартных решений из области цифровой обработки изображений. Также необходимо изучить битовые сети, предназначенные для обработки битовых векторов — наиболее естественного представления результатов синтаксического анализа. 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 с целью проверить практическую применимость разработанных методов и алгоритмов. Предполагается, что будут решаться задачи классифика-

ции цепочек по различным признакам. Например, белковые последовательности планируется классифицировать по функциям, а маркерные РНК-последовательности — по тому, является ли цепочка химерой. 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.

2019-2020

Разработка грамматик 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 грамматики 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 различных 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 грамматик 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 грамматик и ИНС для анализа вторичной структуры, который предполагается 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), Фонда 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 выполнения 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

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

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

en

2.10 Ожидаемые в конце первого года конкретные научные результаты

ru

Предложены обыкновенные (не вероятностные) грамматики для описания особенностей вторичной структуры маркерных цепочек. Проведён ряд экспериментов по использованию данных грамматик 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, индексируемых в scopus.

Разработан параллельный алгоритм синтаксического анализа, адаптированный для решения задачи поиска подстроки с заданной вторичной структурой 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