Использование формальных грамматик для анализа вторичной структуры генмных и протеомных последовательностей

Семён Григорьев

7 марта 2019 г.

1 Сведения о проекте

1.1 Название проекта

ru
Использование формальных грамматик 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
Анализ (детектирование микроорганизмов)

Анализ генетической информации

Поиск новых лекарственных препаратов (антибактериальных в том числе)

en

1.4 Ключевые слова (приводится не более 15 терминов)

ru
Формальные грамматики, синтаксический анализ, параллельные алгоритмы, вторичная структура, РНК, генмноые последовательности, белки, протеомные последовательности, метагеномная сборка.

en

1.5 Аннотация проекта

ru
Проект посвящён исследованию применимости формальных грамматик для анализа вторичной структуры различных биологических последовательностей, например, геномных или протеомных, и разработке соответствующих алгоритмов и решений.

Применение результатов теории формальных языков для анализа биологических последовательностей исследуется давно, однако появились новые результаты и требуется анализ.

Недостаточность современных методов: неточность, ресурсоёмкость.

Требуется применение новых классов грамматик, разработка новых алгоритмов. И даже подходов.

Поиск новых организмов, улучшение предсказания функций белков и т.д.

en

1.6 Ожидаемые результаты и их значимость

ru
Теоретические результаты — формальные методы описания и анализа вторичной структуры. Классы грамматик и конкртеные грамматики для конкретных задач

Применение на практике: классификация, поиск

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.

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
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
Предполагается исследовать !!!

Предлагается построить алгоритмы для синтаксического анлиза,

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

Предполагается применить для анализа РНК и белков. Подбор граммтик

Эксперименты на реальных данных — базах цепочек, имеющихся в открытом доступе.

2019-2020

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, в том числе в биологии, что подтверждается соответствующими статьями. Синтаксический анализ, статья по биологам. Выступления на биата.

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.

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 предпологатся

en