

# Использование формальных грамматик для анализа вторичной структуры геномных и протеомных последовательностей

Семён Григорьев

6 марта 2019 г.

## 1 Сведения о проекте

### 1.1 Название проекта

**ru**

Использование формальных грамматик для анализа вторичной структуры геномных 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**

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 граммтики, синтаксический анализ, параллельные алгоритмы, вторичная структура, РНК, геномные последовательности, белки, протеомные последовательности, метагеномная сборка.

en

## 1.5 Аннотация проекта

ru

Проект посвящён исследованию применимости формальных граммтик для анализа вторичной структуры различных биологических последовательностей, например, геномных или протеомных.

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 граммтик, 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 граммтик и конкретные граммтики для конкретных задач

Применение на практике: классификация, поиск

en

## 2 Содержание проекта

### 2.1 Научная проблема, на решение которой направлен проект

ru

Формальные методы описания и анализа вторичной структуры

Способы и методы анализа вторичной структуры биологических последовательностей. Особенности задачи — особенности алгоритмов. Большой объём данных — дополнительные требования к алгоритмам: параллельные вычисления.

en

### 2.2 Научная значимость и актуальность решения обозначенной проблемы

ru

Формальные методы описания

Алгоритмы синтаксического анализа.

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 новых классов грамматик.

Предложение конкретных грамматик для конкретных важных задач (поиск маркерных последовательностей)

Разработка параллельных алгоритмов синтаксического анализа

Разработка алгоритмов и решений для анализа вторичной структуры с использованием методов теории формальных языков и машинного обучения.

en

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

**ru**

Текстовый анализ

Грамматики, описывающие первичную структуру.

Вторичная структура — через энергии и т.д.

Сложные элементы вторичной структуры.

Существование наборок, решающих демонстрационные задачи.

**en**

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

**ru**

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

Формальные грамматики для белков прямо сейчас (Витольд).

Формальные грамматики для вторичной структуры (Девушка с конфы)

**en**

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

**ru**

Предлагается построить алгоритмы для синтаксического анализа,

Сильно неоднозначные грамматики, что не характерно для языков программирования, для которых разрабатывались многие алгоритмы.

Предполагается применить для анализа РНК и белков. Подбор грамматик

2019-2020

2020-2021

**en**

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

**ru**

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**

**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**  
**ГПУ!**

**en**