Факултет по Математика и Информатика

ПРОЕКТ

за курса

Линукс и езиците за програмиране

Преподавател: Д. Василев, И. Михайлов

REST API using Ensembl

Изготвено от:

Мартин Георгиев

04.02.2021г. БМИ, 1ви курс, фак.№ 26353

Проектът реализира RESTfull API на Python, което позволява на потребител да извлича информация от базата данни на геномния браузър Ensembl (https://www.ensembl.org/).

Описание на проекта

За реализирането на проекта е използвана web базираната платформа Flask. Реализирани са GET операции, които се обръщат директно към API на Ensembl за извличане последователни и допълнителна информация по подаден ID номер на ген.

Функционалност

Посочен е списък с endpoint-и:

- /v1/sequence/gene/id/
  - приема id на ген и да връща като резултат цялата секвенция на гена и всички екзони като отделни свойства
- /v1/sequence/gene/id/?gc_content=true&swap=A:T
  - приема id на ген и да връща като резултат секвенцията с пресметнато процентното GC съдържание и разменени бази аденин с тимин в конкретния пример
- /v1/sequence/id/?content-type=fasta
  - приема ID и връща секвенция в специфичен формат. Поддържаните формати са: fasta, multi-fasta, x-fasta

Стартира се на локалния сървър на потребителя и в браузъра се изписва URL с IP на сървъра + една от желание функционалности

Използвани методи и реализация

- check_status(r) – приема като входен параметър върнатият резултат от **get** заявка от **requests** библиотеката; ако състоянието на резултата не е валидно, приключва програмата с вградена функция exit(), спираща процеса

- swapSeq (seq, first, second) – приема секвенция като стринг и 2 символа, които да бъдат разменени; връща като резултат стринг с разменените азотни бази <**first**> и <**second**>

- getSeqData (id) – входният параметър е ID номер на желания ген; изпълнява GET заявка към API на Ensemble; връща резултата от заявката в удобен обект в JSON формат чрез **json()** метод от библиотеката **requests**

- getSeqDataLookup(id) – прави GET заявка в lookup директорията на API на Ensembl; съдържа информация за транскрипта на гена, включително и подробна информация за екзоните

- getExons(id, sequence) – за вече извлечена секвенция, извлича с GET заявка транскрипта и предоставя последователността на екзоните в транскрипта на секвенцията, отделяйки информацията чрез операции върху стрингове; връща списък с екзоните, представени чрез структура от данни **dictionary**

- fasta(seq_data), x_fasta(seq_data), multi_fasta(id) – методите за форматиран изход на заявката; първите два приемат резултата от GET заявката към gene директорията на API на Ensemble и връщат **dictionary** с информацията; третият метод приема id, за да може да се обърне и към lookup директорията, за да извлече информация за id на екзоните, да ги конкатенира с последователността от оригиналната секвенция и да върне стринга като multi-fasta формат

- gene_data(id) е маршрутизираният метод за първите 2 крайни точки (endpoints); проверява дали има подаден content-type заглавен параметър за GC съдържание и swap опция (работи единствено когато и двата параметъра са подадени) и изпълнява първите 2 endpoint-а според наличието или липсата на заглавни параметри; връща **dictionary** с желаната информация

- seq_data(id) е маршрутизираният метод за третият endpoint; проверява дали подаденият заглавен параметър е в списъка от разрешените формати - **formats = ['fasta', 'x-fasta', 'multi-fasta']** и изпълнява метода format_output(id, content_type), който извиква правилния метод за форматиране; ако подаденият параметър е сбъркан или не е разрешен, се връща като резултат **'Format not supported!'**