РЕФЕРАТ

Магістерська дисертація містить 90 сторінок, 25 рисунків, 2 таблиці, 1 додаток. Було використано 72 джерела.

Актуальність. В рамках даної роботи основна увага приділяється первинним структурам ДНК. Проблеми, пов’язані з аналізом первинної структури, в першу чергу стосуються задач розпізнавання білок-кодуючих областей вже відсеквенованих послідовностей нуклеотидів в молекулі ДНК.

З математичної точки зору поставлена проблема відноситься до задач розпізнавання, алгоритмів вирішення яких є доволі багато, більшість з яких базуються на методах машинного навчання. Та основна проблема полягає в тому, що існуючі об’єми даних неможливо обробити класичними методами машинного навчання, особливо враховуючи швидкість обчислень, збільшення якої в рамках окремої системи в наш час майже не є можливим. Незважаючи на те, що розподілені обчислення відкрили нові шляхи для вирішення задач, які потребують великих обчислювальних потужностей, а стрімкий ріст даних обумовив практично повсюдне використання розподілених обчислень, їх використання в зв’язці з методами машинного навчання досі залишається відносно новою і мало дослідженою задачею. Розпаралелювання обчислень та використання обчислювальних кластерів для вирішення подібного роду задач надасть змогу обробляти надвеликі масиви даних та дозволить значно підвищити швидкість обробки.

Зв’язок роботи з науковими програмами, планами, темами. Магістерська дисертація виконана у відповідності до плану відділу методів індуктивного моделювання та керування Інституту кібернетики імені В.М. Глушкова НАН України в рамках науково-дослідної теми «Розробка методів моделювання біологічних послідовностей та процесів на основі апарату активних частинок» (шифр ВФ.235.14, номер державної реєстрації 0114U000359, 2014-2018).

Мета і завдання дослідження. Метою роботи є пришвидшення процесу розшифровки геному за рахунок розподіленого пошуку кодуючих ділянок генів.

Для досягнення поставленої мети необхідно вирішити наступні задачі:

- провести аналіз відомих методів розпізнавання інтронів та екзонів в ДНК;
- розробити формальну постановку задачі розподіленого машинного навчання для розпізнавання інтронів та екзонів в ДНК;
- розробити алгоритмічне забезпечення задачі;
- впровадити розроблені алгоритми у вигляді програмного продукту;
- дослідити ефективність розроблених алгоритмів шляхом проведення обчислювального експерименту та зробити висновки.

Об’єкт дослідження – процес розшифрування геному.

Предмет дослідження – задача розпізнавання екзонів та інтронів в ДНК.

Методи дослідження базуються на методах машинного навчання та алгоритмах розподілених обчислень.

Наукова новизна отриманих результатів – розроблено розподілені методи машинного навчання, що базуються на використанні наївного баєсівського класифікатора та бінарної логістичної регресії, та адаптовані до вирішення задачі розпізнавання інтронів та екзонів в ДНК.

Публікації. Матеріали роботи опубліковані в рамках XXV міжнародної наукової конференції iScience «Актуальні виклики сучасної науки» [71], міжнародної науково-практичної конференції «Інноваційний розвиток науки нового тисячоліття» [72] та конференції IOT-2017.

МАШИННЕ НАВЧАННЯ, РОЗПІЗНАВАННЯ, КЛАСИФІКАЦІЯ, ЕКЗОН, ІНТРОН, ДНК, НАЇВНИЙ БАЄСІВСЬКИЙ КЛАСИФІКАТОР, БІНАРНА ЛОГІСТИЧНА РЕГРЕСІЯ, APACHE SPARK.