DOI: 10.25205/978-5-4437-1843-9-14

ОПТИМИЗАЦИЯ ДАННЫХ В БИОИНФОРМАТИКЕ

DATA OPTIMISATION IN BIOINFORMATICS

Р. З. Габбасов, Г. Т. Закирьянова, С. И. Хажина

Башкирский государственный медицинский университет, Уфа

R. Z. Gabbasov, G. T. Zakiryanova, S. I. Khazhina

Bashkir State Medical University, Ufa

✉ vessel228@yandex.ru

Аннотация

Современная биомедицина столкнулась с беспрецедентным ростом объемов данных, требующих обработки: от полногеномного секвенирования до сложных медицинских изображений. По данным NIH, ежегодный прирост геномной информации составляет до 40 экзабайт, при этом значительная часть этих данных остается неструктурированной и неиспользованной, что обусловлено множеством факторов.

Abstract

Modern biomedicine is faced with an unprecedented increase in data volumes that require processing, from genome-wide sequencing to complex medical images. According to the NIH, the annual increase in genomic information is up to 40 exabytes, while a significant part of this data remains unstructured and unused, due to many factors.

Биоинформатика — одна из самых динамических наук в наше время. В условиях стремительного роста объемов геномной информации, развития высокопроизводительного секвенирования и появления новых экспериментальных технологий актуальность методов оптимизации становится особенно очевидной. Нелинейность, высокая структурная и морфологическая сложность моделей на фоне общего роста слабоструктурированной исходной информации и относительного роста ее нечеткости приводят к необходимости предварительного преобразования и интеграции данных. Современные исследования требуют не только обработки колоссальных массивов данных, но и обеспечения высокой точности, скорости и интерпретируемости результатов.

Стохастические методы оптимизации (SDG) являются основными для обучения крупных моделей на больших данных в совокупности с обновлением параметров на мини-батчах и возможностью параллелизации, именно поэтому данные алгоритмы в докинге и предсказании 3D-структур белков.

Активно применяются распределенные алгоритмы в оптимизации данных. В проекте 1000 Genomes оптимизированный алгоритм выравнивания позволяет анализировать 2500 геномов одновременно, при этом обрабатывая 200 ГБ данных за 3 часа вместо 24. В криминалистике пользуется успехом SPR Opt, оптимизирующий судебно-медицинские маркеры для получения максимального количества информации при минимальном количестве анализов. В качестве входных данных он принимает полную или частичную последовательность генома.

В условиях работы с неразмеченным данными крайне эффективно себя проявляет обучение без надзора (Unsupervised learning), позволяя выявлять скрытые структуры и закономерности данных без необходимости ручной разметки. Условно можно выделить два основных метода: кластеризация и метод уменьшения данных. Кластеризация — процесс объединения информации в кластеры, используемый в биоинформатике для анализа генов, белков и других биомолекул. Она нашла применение в протеомике, структурной биологии, анализе биомаркеров и др. Наиболее известные алгоритмы включают k-means, DBSCAN и иерархическую кластеризацию. Метод понижения размерности позволяет сохранить наиболее важную информацию, уменьшив количество вводных данных.

Применение машинного обучения обширно. Расчет стабильности комплексов может быть оптимизирован за счет предсказания термодинамических параметров ДНК-белковых взаимодействий.

Несмотря на имеющиеся успехи, существуют проблемы, связанные с передачей данных, требующих возрастающих вычислительных мощностей. Возникают сложности с интерпретируемостью некоторых моделей и высоким уровнем шума, часто встречающимся в биологических экспериментах.

В результате проведенного анализа представлен обзор современных методов оптимизации данных в биоинформатике. Установлено, что применение усовершенствованных алгоритмов обработки геномной информации позволяет существенно повысить точность интерпретации биологических данных. Полученные результаты демонстрируют значительный потенциал для развития персонализированной медицины и оптимизации доклинических исследований за счет внедрения эффективных вычислительных решений.