ГРАФОВАЯ БАЗА ДАННЫХ КАК НОВЫЙ ИНСТРУМЕНТ ДЛЯ ИССЛЕДОВАНИЯ БАКТЕРИАЛЬНЫХ ПОПУЛЯЦИЙ

Темлякова Е.А., Рясик А.А., Сорокин А.А.

Институт биофизики клетки РАН, Россия, 142290, Пущино, ул. Институтская, д.3,
+7(4967)739165 evgenia.teml@gmail.com

В рамках данной работы была создана граф-ориентированная база данных на платформе Neo4j [1], предназначенная для хранения разнородной информации о больших бактериальных популяциях. Такая архитектура хранилища была выбрана по целому ряду причин, важных для работы с биологическими объектами: возможность использования арсенала методов теории графов без дополнительной подготовки данных, возможность поиска “цепочек” объектов с известным начальными и конечными узлами, но неизвестным числом промежуточных шагов, совместное хранение противоречивых данных. Помимо этого, графовая база данных характеризуется гибкостью структуры при внесении новых типов данных, характеристик и категорий, простой системой построения запросов и высокой скоростью работы. Для наполнения хранилища было создано несколько программных модулей, позволяющих осуществлять загрузку и интеграцию данных из следующих внешних источников: GenBank, MetaCyc, RegulonDB, ChEBI, NCBI Taxonomy.

На данный момент в базу данных загружена информация о 240 видах организмов-симбионтов человека, являющихся представителями микробиомов ротовой полости, воздухоносных путей и кишечника человека [2]. Для рассмотренных организмов представлены следующие данные: генетическая информация в виде набора контигов и плазмид с разной степенью детализации описания, метаболические пути и компартментализация конкретных биохимических реакций, положение на таксономическом дереве, функции белков и их участие в регуляции клеточных процессов, положение транскрипционных единиц, генов, ДНК-маркирующих сайтов, мобильных элементов и протяженных повторов. Проанализированы характеристики, влияющие на производительность базы данных как в режиме загрузки информации о новых организмах, так и при обработке запросов пользователя. Разработаны программные модули автоматически отслеживающие ошибки содержимого, дубликаты и корректно обрабатывающие противоречивые данные, возникающие на этапе загрузки. Предложены методы, позволяющие исправить эти ошибки и предотвратить их возникновение в дальнейшем.

Помимо этого проведен первичный анализ содержимого базы данных: исследовано наличие информации в свободном доступе для указанных 240 бактерий и изучен ряд биологических закономерностей.

Литература.

1. Neo4j web-site (Neo Technology): http://www.neo4j.org/
2. *J. Peterson et al.* The NIH human microbiome project// *Genome research* **19**, 12, 2009. Pp. 2317–2323.