ГРАФ-ОРИЕНТИРОВАННАЯ БАЗА ДАННЫХ ДЛЯ ИССЛЕДОВАНИЯ БАКТЕРИАЛЬНЫХ ПОПУЛЯЦИЙ

Рясик А.А., Темлякова Е.А., Джелядин Т.Р., Сорокин А.А.

ИБК РАН, Россия, 142290, Пущино, ул. Институтская 3, +7(4967)739404, arc7an@gmail.com

Зачастую, в биологических исследованиях приходится иметь дело с большим объемом разнородных данных, собранных из различных источников, затрачивая значительное время и компьютерные ресурсы. Это делает актуальной задачу эффективной интеграции информации и удобного представления полученных данных. В качестве решения этой проблемы недавно было разработано несколько интеграционных систем на основе реляционных СУБД [1, 2], агрегирующих данные из различных источников в едином хранилище. Однако, учитывая высокую вариабельность биологических данных, а также огромное количество связей между объектами – выбор граф-ориентированного хранилища представляется наиболее предпочтительным. Более того, такое хранилище обладает рядом существенных преимуществ: оно позволяет осуществлять простой и быстрый поиск "пути"между двумя объектами в базе данных, имеет более гибкую структуру и с легкостью поддается масштабированию.

Нами разработан программный комплекс для сбора сведений о генетических объектах, протеоме, транскриптоме, регуляторных и метаболических сетях, а также физико-химических свойствах молекулярных объектов микроорганизмов и т.п. взятых из таких авторитетных источников, как GenBank, MetaCyc, UniProt, PDB и др. Помимо этого представлен алгоритм присвоения связей подобия между объектами: гомология полипептидных последовательностей и сходства по физико-химическим и структурным свойствам.

Разработанная инфраструктура была использована для решения реальных биологических задач: изучение распределения физико-химических свойств функционально-значимых участков геномов бактерий *Escherichia coli*, *Bacillus subtilis*, *Chlamydia trachomatis* и *Corynebacterium glutamicum*, исследование основных ферментов биолюминесцентного метаболического пути светящихся бактерий.

База данных реализована на платформе Neo4j [3], наполнение и обновление осуществлялось на основе программных модулей реализованных на языке Python. Построение запросов, представление и визуализация результатов производится при помощи стандартного web-интерфейса.

Работа частично поддержана грантами РФФИ №14-04-31793-мол_а, №14-34-50215-мол_нр, №14-44-03679-р_центр_а.

Литература.

1. Swertz, M.A. et al, BMC Bioinformatics, 2010, 11 Suppl 12, S12.
2. Smith, R.N. et al. Bioinformatics (Oxford, England), 2012, 28(23), 3163–3165.
3. Neo4j web-site (Neo Technology): http://www.neo4j.org/