Секция «Биоинформатика»

# Система комплексного анализа для определения зависимости экспрессии от состава генной структуры

**Научный руководитель – Мустафаев Орхан Нариман**

***Мехдиев Шаиг Фаиг***

*Выпускник (магистр)*

Бакинский государственный университет, Биологический факультет, Баку, Азербайджан

*E-mail: bioprozium@gmail.com*

За последнее десятилетие резко увеличилось количество полностью или частично завершенных геномных и транскриптомных проектов с участием различных живых организмов. Это привело к созданию баз данных, содержащих обширную информацию о нуклеотидных последовательностях, что требует всестороннего исследования [1]. Яркие иллюстрации, полученные в результате структурно-функционального анализа нуклеотидных последовательностей, могут быть использованы для изучения функций транспозонов, функциональной структуры промоторов пользователем, открытия связанных с белками доменов и сходства генов у разных видов [2].

Как упоминалось выше, современная база данных содержит наиболее точную информацию для отдельного гена или для группы генов, но все они предназначены для крупномасштабного силикоанализа. Анализ размера транскриптов, изучение их мотивов, определение их нуклеотидного контекста, определение состава ГС и частоты кодонов, короче говоря, анализ всех потенциальных факторов, влияющих на эффективность экспрессии генов в живой клетке. Для выполнения таких расчетов исследователи вынуждены использовать существующие утилиты или создавать такие программы [3]. С учетом вышеперечисленных факторов создается комплексная система анализа образцов генома для исследователей, не обладающих специальными навыками в области биоинформатики [2, 3].

**Материалы и методы.** Определенная работа была проделана в этом направлении, изначально в прикладной программе «Microsoft Visual Studio 2017» было разработано ядро системы, выполняющее анализ образцов генома с помощью языка программирования C ++ 11, включая библиотеку инструментов «rapidJSON». Образцы генома для тестирования системы были загружены в формате FASTA с https://asia.ensembl.org/.

**Результаты - их обсуждение.** Информация о составе геномов разных организмов может быть получена этой системой в разных форматах (форматы FASTA и JSON). Полученная информация обрабатывается системой, разделяется на специальные классы и сохраняется в памяти. В этой системе вычисления выполняются двумя основными способами: сначала последовательность гена выбирается либо полностью, либо с использованием различных фильтров, например, по количеству нуклеотидов, в соответствии с областью последовательности гена, а затем несколько последовательностей гена на основе выбранных характеристик фильтра. Можно производить расчеты: количество нуклеотидов в последовательности гена, количество пар нуклеотидов, отношение нуклеотидов гуанина и цитозина к выбранной последовательности гена, «CpG-island», то есть отношение соединений цитозина и гуанина к выбранной последовательности гена, количество кодонов, количество аминокислот (Рис.1).

## Источники и литература

1) 1. Grote, A., Hiller, K., Scheer, M., Münch, R., Nörtemann, B., Hempel, D. C., & Jahn, D. // JCat: A novel tool to adapt codon usage of a target gene to its potential expression

host. Nucleic Acids Research, 2005, 33(SUPPL. 2), 526–531. https://doi.org/10.1093/nar/gki376

2) 2. Cosart, T., Beja-Pereira, A., & Luikart, G. // Exonsampler: A computer program for genome-wide and candidate gene exon sampling for targeted next-generation sequencing. Molecular Ecology Resources, 2014, 14(6), 1296–1301. https://doi.org/10.1111/1755-0998.12267

3) 3. Renaud, G., LaFave, M. C., Liang, J., Wolfsberg, T. G., & Burgess, S. M. // TrieFinder: An efficient program for annotating Digital Gene Expression (DGE) tags. BMC Bioinformatics, 2014, 15(1), 4–9. https://doi.org/10.1186/1471-2105-15-329

## Иллюстрации

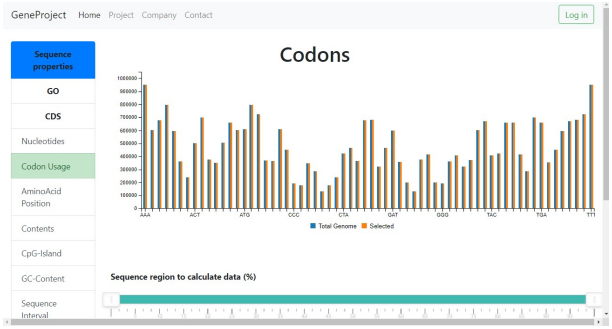


**Рис. 1.** Подсчет количества кодонов в геноме Arabidopsis thaliana