

Сравнительное изучение структуры и таксономии на примере 5 S и 16 S РНК бактерий

Овчинникова Ю.И.¹, Садовский М.Г.^{1, 2, 3*}

¹ Сибирский федеральный университет, Красноярск, Россия

² Институт вычислительного моделирования КНЦ СО РАН, Красноярск, Россия

³ СНКЦ ФМБА РФ, Красноярск, Россия

*msad@icm.krasn.ru

Ключевые слова: нуклеотидная последовательность; частота; упругая карта; кластер

Мотивация и цель: Выявление статистических свойства нуклеотидных последовательностей, отражающие связь структуры и таксономии исследуемых генов, является ключевой задачей молекулярной биологии и биоинформатики. Одним из классических объектов исследования в данной области являются последовательности 16S рРНК, однако существуют и другие ансамбли последовательностей генов рРНК, представляющие не меньший интерес. В рамках данного исследования мы изучали связь двух структур на примере последовательностей 5S РНК и 16S РНК бактерий.

Методы и алгоритмы: В настоящей работе выявлены подобию и различия структур в упомянутых ансамблях, определяемые триплетным составом генов 16S РНК и генов 5S РНК. Ранее [1] было показано, что триплетный состав генетических систем хорошо коррелирует с таксономическим положением носителей соответствующих генов. Однако зачастую видовой состав баз генетических данных (в нашем случае – базы SILVA) весьма смещён: численность генов различных таксономических групп организмов (бактерий) сильно отличается, что может приводить к искажению результатов. Поэтому необходимо провести индексирование баз данных, т. е. удаление части генов из сверхпредставленных групп. Тот факт, что таксономия бактерий тесно коррелирует с особенностями последовательностей рРНК бактерий, хорошо известен. Стандартный прием выявления такой связи использует технологию сравнения последовательностей генов, основанную на выравнивании (редакционном расстоянии). В нашей работе мы используем иные методы сравнения генов, не опирающиеся на идею выравнивания.

Результаты: перейдём к описанию самой работы и результатов. Сравнение структуры и таксономии индексированных баз генов 5S и 16S РНК бактерий проводилось с помощью метода упругих карт [2]. Для этого последовательности генов преобразовывались в частотные словари триплетов; частотный словарь (триплетов) – это список всех триплетов, встречающихся в последовательности гена, с указанием их частот встречаемости. Тем самым, каждая последовательность гена преобразуется в точку в многомерном метрическом пространстве частот триплетов, что позволяет задать расстояние между точками естественным образом (например, Евклидово расстояние) и применить множество методов анализа структуры распределения таких точек. В нашей работе изучалась кластеризация генов 5S РНК бактерий в сравнении с аналогичной кластеризацией 16S РНК бактерий. Структуры в распределении таких генов могут определяться

как с помощью классификации, так и с помощью кластеризации (метод упругих карт, метод динамических ядер, он же k-means). В нашей работе используется метод упругих карт для кластеризации генов на основе их частотных словарей триплетов. Частотные словари триплетов представляют собой представление генов в виде последовательностей из трех нуклеотидов (например, ACG, GTC и т. д.), и их частоты встречаемости в геноме. Кластеризация по частотным словарям триплетов проводилась с помощью ПО VidaExpert. Результатом настоящей работы является установление связи кластеризации генов 5S и 16S РНК. Так, возможными исходами являются полностью различные паттерны кластеризации, наблюдаемые на этих двух множествах генов, напротив, заметное их подобие. Ещё одним важным является ответ на вопрос: верно ли, что триплетный состав этих двух групп генов позволяет отличить гены 16S РНК и 5S РНК бактерий.

Comparative study of structure and taxonomy interplay over 5 S and 16 S ribosomal genes of bacteria

Ovchinnikova Y.¹, Sadovsky M.^{1, 2, 3*}

¹ Institute of computational modelling SB RAS, Krasnoyarsk, Russia

² FSR&CC of FMBA of Russia, Krasnoyarsk, Russia

³ SFU, Krasnoyarsk, Russia

* msad@icm.krasn.ru

Key words: nucleotide sequence; frequency; elastic map; cluster

Motivation and Goal: Revealing the statistical properties of nucleotide sequences that reflect the relationship between the structure and taxonomy of the genes under study is a key task in molecular biology and bioinformatics. One of the classic objects of research in this area is 16S rRNA sequences, but there are other ensembles of rRNA gene sequences that are of no less interest. In this study, we studied the relationship between the two structures using the example of bacterial 5S RNA and 16S RNA sequences. Here we studied the clustering of bacterial 5S RNA genes in comparison to the similar clustering of bacterial 16S RNA. Structures in the distribution of such genes can be determined using both classification and clustering (elastic map method, or k-means). Our work uses the elastic map method to cluster genes based on their triplet frequency triplet ensembles. Triplet frequency dictionaries are a representation of genes as a set of triplets (e.g., ACG, GTC, etc.) provided with their frequency of occurrence in a genome. Clustering by frequency dictionaries of triplets was carried out using VidaExpert software. The result of this work is the connection between the clustering of 5S and 16S RNA genes. Thus, possible clustering outcomes are completely different for these two sets of genes, noticeable similarity in some issues also has been observed. Another important point is whether it is possible to distinguish the 16S RNA and 5S RNA genes of bacteria. The answer is positive on this question.

Список литературы/References

1. Sadovsky M., Putintseva Yu., Chernyshova A., Fedotova V. Genome Structure of Organelles Strongly Relates to Taxonomy of Bearers. In: Ortuño F., Rojas I. (Eds.). Bioinformatics and Biomedical Engineering. IWBBIO 2015. Lecture Notes in Computer Science. Vol. 9043. Springer, 2015;481-490
2. Gorban A., Kégl B., Wunsch D.C., Zinovyev A. (Eds.). Principal Manifolds for Data Visualizations and Dimension Reduction. Springer, 2008