0.1. *Пестова А.С.* **Кластерный анализ ДНК на основе пиковой характеристики символьного разнообразия**

В последние 20-25 лет усилия научного сообщества были направлены на накопление секвенированных последовательностей ДНК. В настоящее время в основных молекулярно-генетических банках данных (EMBL, GenBank, DDBJ) уже накоплена огромная информация о полностью секвенированных геномах микроорганизмов и геномах эукариот, включая геном человека, причем объем расшифрованных последовательностей стремительно растет. Полученная информация обрабатывается и изучается различными математическими аппаратами, одним из которых является аппарат комбинаторики слов. Ряд результатов на основе этого подхода получен в [1]. Пиковая характеристика символьного разнообразия строки, основанная на вычислении максимума конечной разности функции энтропии сдвигов предложена в [2] в аспекте исследования временных рядов. Данная работа посвящена применению данной оценки к изучению характеристик символьного разнообразия нуклеотидного состава ДНК, определению способности геномов к рекомбинации генов.
В основе исследования лежат вычисления значений функции энтропии сдвигов для входного генома по следующей формуле:

$$F(m) = -\Sigma_{i=1}^{4^m} \frac{C_i}{n-m+1} \cdot \log_{4^m} \frac{C_i}{n-m+1}, \quad (1)$$

где n - длина всей строки, m – размер текущего окна (сдвига), C_i - счетчик по всем позициям окна в исходном слове для подслова длины m с номером "i". Верхний предел суммирования — 4^m определяется общим числом всех возможных подслов длины m в алфавите ДНК. Значение функции $F(m)$ вычисляется последовательно, начиная с $m = 1$ с шагом 1 до $m = \lceil 2\log_4 n \rceil$, так как значения функции $F(m)$ на последующих значениях m будут близки к 0, что не имеет принципиального значения для анализа символьного разнообразия ДНК. Для получения пиковой характеристики символьного разнообразия важен лишь резкий скачок вниз функции $F(m)$, который будет находиться в диапазоне от 1; до $\lceil 2\log_4 n \rceil$.
Полученные значения функции энтропии сдвигов в дальнейшем были использованы для определения оценок символьного разнообразия исследуемых ДНК, рассматриваемых как слова в четырехсимвольном алфавите. На основе тестового исследования 20 геномов из банков данных EMBL, GenBank, DDBJ, была выявлена тенденция их объединения в отдельные группы по пиковым значениям характеристик символьного разнообразия. В ходе последующего кластерного анализа в пространстве (пиковое значение конечной разности функции энтропии сдвигов, аргумент пикового значения) выяснилось, что в таком кластерном пространстве геномы объединяются по семействам биологических организмов. Развитие данных исследований предполагает как расширение осей кластерного пространства за счет применения новых характеристик символьного разнообразия, так и расширение набора исследуемых геномов организмов из других семейств.

Список литературы

[1] Орлов Ю. Л. Анализ регуляторных геномных последовательностей с помощью компьютерных методов оценок сложности генетических текстов: дис. ... канд. биол. наук — 2004 — 180 с.

[2] Сметанин Ю. Г. , Ульянов М. В. Мера символьного разнообразия: подход комбинаторики слов к определению обобщенных характеристик временных рядов // Бизнес-информатика. — 2014. — № 2(28). С. 15—21.