Секция «Биоинженерия и биоинформатика»

Разработка алгоритма поиска нкРНК в геноме бактерий

Абдуллаев Эльдар Теймурович
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: abd.el@mail.ru

Некодирующие РНК (нкРНК) привлекают к себе огромный интерес ученых, т.к. они играют важную регуляторную роль у эукариот и прокариот. Также интерес вызван тем, что большая часть геномной ДНК эукариот транскрибируется, тогда как только несколько процентов ДНК кодируют белки (у человека примерно 1,2%), отсюда следует, что большинство транскриптов приходится на нкРНК [3],[4]. Существующие алгоритмы поиска нкРНК не универсальны и требуют больших вычислительных мощностей, что делает их непригодными для поиска в геноме [2]. Поэтому создание такого алгоритма является актуальной биоинформатической задачей.

Целью данной работы является создание алгоритма, способного находить нкРНК в геномах прокариот. Важными факторами при создании такого алгоритма являются универсальность и время его работы. За основу своего алгоритма я взял RNAplfold алгоритм [1]. Он определяет вторичную структуру РНК по нуклеотидной последовательности на основании поиска структур с минимальной энергией. В нашем алгоритме основным критерием для поиска нкРНК в геноме является поиск участков РНК, способных к образованию вторичных структур (предположительных нкРНК). Выдача RNAplfold не дает прямой информации о наличии генов нкРНК, отсюда следует, что требуется дальнейшая обработка.

Для анализа выдачи RNAplfold были использованы два подхода: анализ вероятностей нуклеотидных пар из выдачи и анализ статистических сумм (статсумм) нуклеотидов. В первом подходе я анализировал отклонение распределения вероятностей нуклеотидных пар в нкРНК от неверно определенных пар. Во втором подходе я рассчитывал статистические суммы (сумма вероятностей всех возможных пар данного нуклеотида). Для дальнейшего анализа статсумм я использовал следующие методы:

1) Поиск нуклеотидов с высоким значением статсумм

2) Поиск участков с высокими значениями статсумм (пики в графическом представлении)

3) Анализ отклонений распределения статсумм нуклеотидов нкРНК от остальных нуклеотидов

4) Поиск участков с высоким значением Z-score для статсумм

5) Поиск участков с высокими значениями статсумм, полученных при суммировании вероятностей пар с поощрением пар в составе структурных стеблей

В дальнейшем также будет опробован подход, в котором статсуммы будут рассчитываться для достаточно консервативных межгенных участков множественных выравниваний.

Литература

1. S. H. Bernhart, I.L. Hofacker, and P.F. Stadler (2006) Local Base Pairing Probabilities in Large RNAs. Bioinformatics 22: 614-615

2. J. Gorodkin, I.L. Hofacker (2009) De novo prediction of structured RNAs from genomic sequences Trends in Biotechnology 1: 9-19

3. Mattick, J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity

4. Mattick, J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms

Слова благодарности

Я благодарен Светлане Виноградовой, аспирантке ФББ, за полезные советы и конструктивную критику.