О распознавании кодирующих участков в ДНК последовательностях эукариот

В.А. Галинский, А.Н. Гайдук
Научно-исследовательский институт прикладных проблем математики и информатики, г. Минск,
email: GalinskijVA@bsu.by, GaidukAN@bsu.by

Актуальной задачей бионформатики является разработка эффективных математических моделей, методов, алгоритмов и программного обеспечения для распознавания кодирующих участков в ДНК последовательностях эукариот [1,2]. В настоящее время существует два основных подхода к распознаванию кодирующих участков в ДНК последовательностях [2]: подход, основанный на статистических методах и подход, основанный на сходстве ДНК последовательностей. Подход, основанный на сходстве последовательностей, является эффективным, если исследуемая последовательность имеется в базе данных ДНК последовательностей. Этот подход не применим для последовательностей, отсутствующих в базе данных. Поэтому активно развивается подход, использующий статистические свойства ДНК последовательностей. В настоящее время предложено большое число методов и разработан ряд программных комплексов для предсказания кодирующих участков ДНК последовательностей эукариот [2]. Основными недостатками предложенных методов и разработанных программ являются: пропуск и объединение кодирующих участков, ошибки в определении границ кодирующих участков.

В настоящем докладе для распознавания кодирующих участков предложен подход, основанный на марковских свойствах ДНК последовательностей эукариот [3,4,5]. В результате проведенных исследований марковских свойств ДНК последовательностей для моделирования кодирующих участков была выбрана цепь Маркова второго порядка. Параметры цепи Маркова оценивались по выборке h178 [6]. Точность распознавания кодирующих участков исследовалась с помощью вычислительного эксперимента. По результатам вычислительного эксперимента, полученным с использованием выборки Burg570 [7], вероятность правильного распознавания кодирующего участка составила 90%.

На основе предложенного подхода было разработано программное обеспечение для распознавания кодирующих участков ДНК последовательностей эукариот. На рисунках 1 и 2 приведены результаты работы программы для генов, содержащих один и три кодирующих участка.

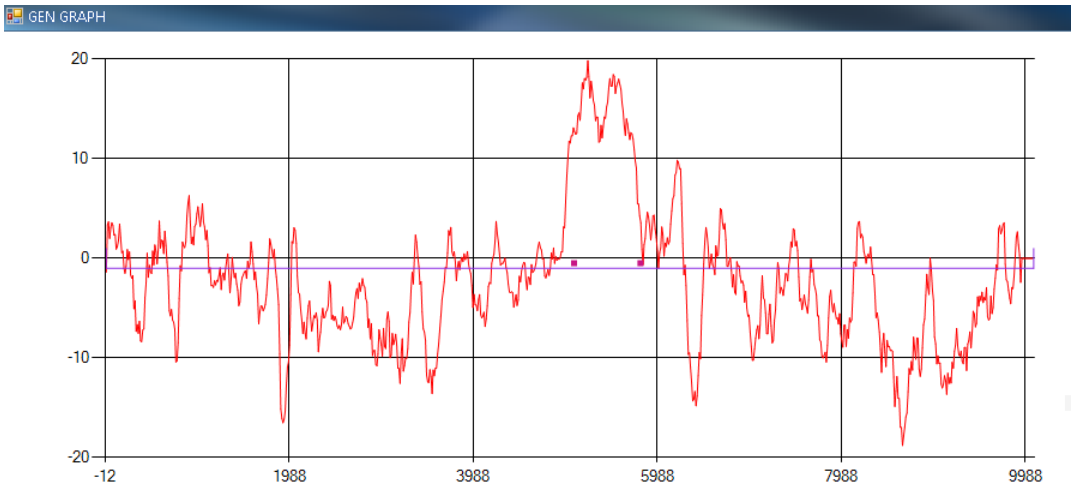

Рисунок 1. – Результат для гена с одним кодирующим участком (по оси абсцисс номера нуклеотидов, по оси ординат значение распознающей статистики)

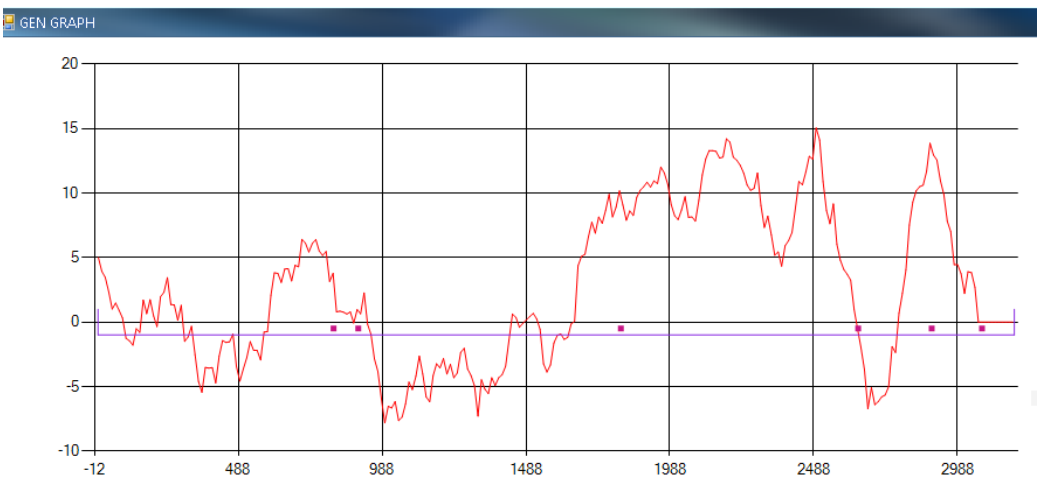

Рисунок 2. – Результат для гена с тремя кодирующими участками (по оси абсцисс номера нуклеотидов, по оси ординат значение распознающей статистики)

Полученные результаты позволяют сделать вывод о возможности использования предложенного подхода для распознавания кодирующих участков ДНК последовательностей эукариот.

Список использованных источников

1. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA Journal of Molecular Biologyю 1997; 268(1), p. 78-94.

2. Гельфанд М.С., Миронов А.А. Предсказание и компьютерный анализ экзон-интронной структуры генов человека. Молекулярная биология, 2004, том 38, №1, с. 82-91.

3. Харин Ю.С., Петлицкий А. И. Цепь Маркова с частичными связями ЦМ (s, r) и статистические выводы о ее параметрах. Дискретная математика. 2007, т. 19, № 2. С. 109–130.

4. Buhlmann P., Wyner A. Variable length Markov chains. The Annals of Statistics. 1999, vol. 27, № 2. P. 480–513.

5. Харин Ю. С., Мальцев М. В. Алгоритмы статистического анализа цепей Маркова с условной глубиной памяти. Информатика. 2011, №1. С. 34–43.

6. Evaluation of gene structure prediction programs / M. Burset, R. Guigo // Genomics 1996, V. 34. P. 353–367.

7. Burset/Guigo96 Dataset [Electronic resource]. – mode of access: http://www.imtech.res.in/raghava/genebench/datasets/Burset-Guigo96/. – Date of access: 05.10.2014.