***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e51. doi: 10.17537/icmbb18.82*

Поиск паттернов ассоциации между функциональными элементами генома

Маткаримов О.О., Поливода Д.Э., Попцова М.С.

Национальный исследовательский университет «Высшая школа экономики»

moptsova@hse.ru, maria.poptsova@gmail.com

Технологии секвенирования следующего поколения сделали возможным картирование множества функциональных элементов генома. Так стало возможным определение расположения эпигенетических факторов, включая метилирование, модификации гистонов, места открытого хроматина, регуляторной РНК, а также места связывания транскрипционных факторов и других важных белков. Данные, генерируемые в результате NGS-экспериментов, хранятся на сайтах проектов в открытом доступе, обычно в формате *bed*-файлов. Актуальной является задача поиска взаимосвязей между различными функциональными аннотациями генома, как экспериментальными, так и теоретическими. Существующие программы поиска паттернов имеют существенные ограничения, большинство реализовано для работы в системе юникс, графический интерфейс отсутствует, а сами программы сложны в использовании. В данной работе мы представляем программу, запускаемую в браузере в любой операционной системе, с пользовательским графическим интерфейсом, которая принимает на вход два файла геномной аннотации в формате *.bed*, визуализирует распределение функциональных элементов в виде плотностей на уровне хромосомы и осуществляет поиск паттернов ассоциации между двумя исследуемыми геномными элементами. Найденные паттерны визуализируются, и информация об их расположении выдается в виде списка. Данная программа предназначается для решения широкого класса биоиформатических задач поиска паттернов ассоциации между различными функциональными аннотациями генома.

Ключевые слова: *поиск паттернов, полногеномная аннотация, файлы формата .bed, функциональные элементы генома, эпигенетика, ДНК-элементы, транскрипционные факторы, технологии секвенирования следующего поколения*.

Searching for patterns of association between functional genomic elements

Matkarimov O.O., Polivoda D.E., Poptsova M.S.

National Research University Higher School of Economics

Next-generation sequencing technologies made it possible to map numerous functional genomic elements. Thus, it became possible to define positions of epigenetic factors including methylation, histone modifications, sites of open chromatin, regulatory RNA, and also binding sites for transcription factors and other important proteins. Data, generated as a result of NGS-experiments, are stored at the project web-sites and is freely available usually in bed-format. The problem of finding associations between different functional genomic annotations, both experimental and theoretical, is very important. The existing programs for pattern search have significant limitations. most of them are developed to work in the Unix-like systems, they are lacking graphical interface, and programs are complex in usage. In the present work we present a program that can be run in a browser in any operation system, it has a graphical interface, and it accepts as an input two files of genome annotations in .bed format, visualize distribution of functional elements as densities at the level of chromosome and performs a search for patterns of association between different functional genomic annotations. The detected patterns are visualized and information about their position is given in a list. The presented program is designed to solve a broad class of bioinformatics problems of finding patterns of association between different functional genome annotations.

Key words: pattern search, full-genome annotation, .bed-files, functional genomic elements, epigenetics, DNA elements, transcription factors, next-generation sequencing technology.

1. Введение

В результате прогресса в технологиях секвенирования следующего поколения стало возможным осуществлять полногеномное картирование множества функциональных элементов генома. Так, в настоящее время на сайте консорциумного проекта по эпигеномике «The Roadmap Epigenomics» доступно около 3000 полногеномных аннотаций модификаций гистонов, сайтов открытого хроматина, метилирования, положений регуляторной РНК, а также транскриптомные данные для разных типов тканей, принадлежащих к трем основным классам – стволовые клетки, зародышевая ткань и взрослая ткань. Другой консорциумный проект «The Encode Project» (Энциклопедия ДНК-элементов) в настоящее время содержит около 14000 полногеномных картирований, основную часть которых занимают данные о сайтах связывания транскрипционных факторов. Консорциумные проекты по секвенированию раковых опухолей *TCGA* и *ICGC* также содержат несколько тысяч полногеномных аннотаций, полученных в результате технологий секвенирования следующего поколения. Помимо данных международных проектов, существует огромное количество данных более мелких проектов, как открытого доступа, так и находящихся в собственности клиник и госпиталей. Огромный объем полученных полногеномных данных требует специального программного обеспечения и алгоритмов для извлечения из них биологического смысла [1].

Проблема поиска паттернов в геномных аннотациях возникла сразу же вместе с возникновением данных, и ниже мы приводим краткий обзор наиболее успешных на сегодняшний день программ.

1.1. ChrommaSig

Программа ChrommaSig [2] может обрабатывать сырые экспериментальные данные. Заложенный в программу алгоритм позволяет находить коррелирующие между собой области хроматиновых сигнатур и создавать отчеты в виде иллюстрированного *pdf*-файла. Однако программа не имеет пользовательского интерфейса, в ней отсутствует возможность кластеризации и классификации. Программа не поддерживается с 2008 года.

1.2. ChrommHMM

Метод ChrommHMM [3] был разработан для определения состояний хроматина на основе метода скрытых цепей Маркова. Метод не доступен в виде программы для широкого пользования.

1.3. ChroModule

Метод ChroModule [4] был создан для анализа эпигеномных данных совместно с данными проекта Encode. Метод использует машинное обучение и находит паттерны в аннотациях гистонных модификаций, ассоциированные с регуляторными элементами, такими как промоторы и энхансеры.

1.4. HMMSeg

Программа HMMSeg [5] осуществляет сегментацию непрерывных геномных данных на основе скрытых цепей Маркова и *wavelet*-анализа. Программа может принимать на вход одновременно несколько аннотаций в формате *.bed* и находить многотипные функциональные домены. Программа не имеет пользовательского интерфейса и запускается исключительно с командой строки.

Все перечисленные программы имеют множество ограничений для использования в задачах поисках паттернов для любых двух полногеномных аннотаций. Таким образом, необходимость разработки программы, которая бы имела графический пользовательский интерфейс и позволяла широкому кругу пользователей, в том числе без знания среды *Unix*, осуществлять анализ полногеномных аннотаций остается актуальной задачей. Мы разработали программу Genomic Pattern Recognition System, имеющую графический пользовательский интерфейс, что позволяет загружать файлы через графическое меню открытия файлов, с разномасштабной визуализацией исследуемых данных, а также производить поиск паттернов между двумя полногеномными аннотациями.

2. Программа поиска паттернов Genomic Pattern Recognition System

2.1. Пользовательский интерфейс

Главное окно программы изображено на рисунке 1. Пользователь может загрузить два файла с аннотацией в формате *.bed через стандартное диалоговое окно. Загруженная аннотация пересчитывается в плотности на заданном пользователем интервале (например, 1 Mb). С помощью линейки управления пользователь может увеличивать/уменьшать масштаб выводимых графиков, перемещать график вправо или влево, сохранять выводимый на экран участок в файл в виде графического файла.

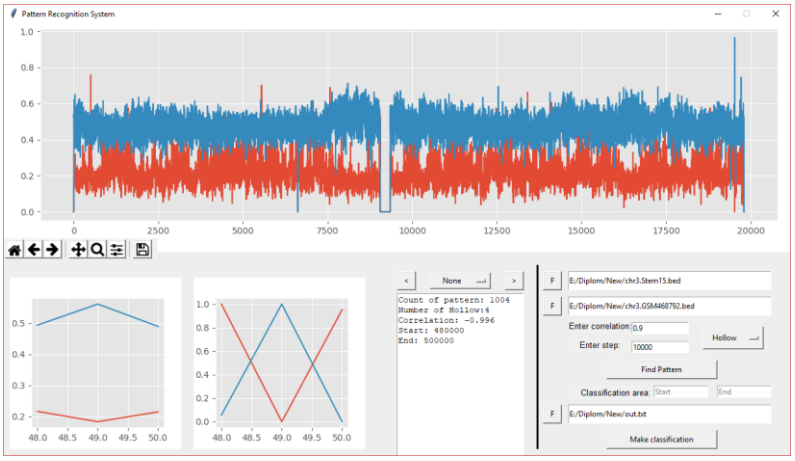

Рис. 1. Главное окно программы Genomic Pattern Recognition System.

Пример увеличенного масштаба графика с помощью линейки управления представлен на рисунке 2.

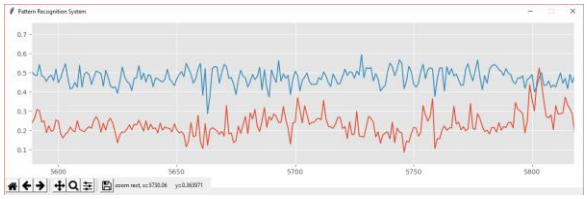

Рис. 2. Две сравниваемые функциональные аннотации в увеличенном масштабе на выделенном отрезке.

2.2. Поиск паттернов

После нахождения распределения плотностей хромосом в гене, для каждого из файлов был реализован алгоритм [6] для дальнейшего поиска корреляции между двумя аннотациями.

В результате работы алгоритма в отдельные окна выводятся ненормализованные и нормализированные участки двух графиков с корреляцией выше пороговой (задаваемой пользователем), а также значения корреляции для этого участка, его номера из всех участков и его тип (рис. 1). Все найденные участки паттернов делятся на четыре типа: 1) пик первого графика совпадает с пиков второго графика; 2) пик первого графика совпадает с впадиной второго графика; 3) впадина первого графика совпадает с пиком второго графика; 4) впадина первого графика совпадает с впадиной второго графика.

Предусмотрена возможность пролистывания всех найденных паттернов. Возможно сохранить список всех найденных паттернов с координатами в геноме.

Программа доступна для скачивания на сайте: https://cs.hse.ru/dnapunctuation/.

3. Благодарности

Работа выполнена в рамках проведения исследования № 18-05-0038 в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2018 г. и в рамках государственной поддержки ведущих университетов Российской Федерации «5-100».

4. Список литературы

1. Berger B., Peng J., Singh M.. Computational solutions for omics data. *Nat Rev Genet.* 2013. V. 14. № 5. P. 333–346.
2. Hon G., Ren B., Wang W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput Biol*. 2008. V. 4. № 10. P. e1000201.
3. Ernst J., Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*. 2010. V. 28. № 8. P. 817–825.
4. Won K.J., Zhang X., Wang T., Ding B., Raha D., Snyder M., Ren B., Wang W. Comparative annotation of functional regions in the human genome using epigenomic data. *Nucleic Acids Res*. 2013. V. 41. № 8. P. 4423–4432.
5. Day N., Hemmaplardh A., Thurman R.E., Stamatoyannopoulos J.A., Noble W.S. Unsupervised segmentation of continuous genomic data. *Bioinformatics*. 2007. V. 23. № 11. P. 1424–1426.
6. Gao W., Brown C., Grossman R., Ma L., Slattery M., White K., Yu P. Discovering geometric patterns in genomic data. In: *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. ACM, 2012. P. 147–154.