Секция «Биоинженерия и биоинформатика»

Динуклеотидный состав генома человека

Безменова Александра Васильевна
Студент
МГУ - Московский государственный университет имени М.В. Ломоносова,
Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: bsshka@yandex.ru

Адекватная статистическая модель распределения частот слов в геноме важна для разных задач: выявления недо- и перепредставленных слов, неоднородностей в геноме, поиска массовых коротких сигналов и др. Известно, что предсказание частот динуклеотидов в предположении о бернуллиевском распределении мононуклеотидов оказывается неудовлетворительным. Мы предлагаем модель, в которой учитывается эволюционное давление на динуклеотиды и неоднородность G+C-состава вдоль генома. Существование т. н. G+C-изохор – длинных участков, однородных по G+C-составу, – было показано ранее [1]. Для проверки предсказательной силы модели была разработана программа разделения генома на изохоры, основанная на методе, предложенном в работах [2,3]. На очередной итерации каждый полученный ранее участок разбивается на две части, наиболее различающиеся по G+C-составу; разбиение происходит, если различие достоверно. Ожидаемая частота динуклеотида XY на изохоре s определяется следующим образом: $f_s^{exp}(XY) = k(XY)f_s(X)f_s(Y)$, где $f_s(W)$ – частота моно- или динуклеотида W на изохоре s, а $k(XY)$ – коэффициент, характеризующий давление на нуклеотид XY в геноме. Для нахождения этого коэффициента строилось двумерное распределение, по оси абсцисс откладывалось произведение частот $f_s(X)f_s(Y)$ нуклеотидов X и Y, по оси ординат – наблюдаемое число динуклеотидов $f_s^{obs}XY$ (рис. 1, 2). Небольшое число очевидных выбросов считались исключительными изохорами и не учитывались; остальные точки аппроксимировались линейной функцией $y = k(XY)x$, $k(XY)$ - искомый коэффициент. Для каждой изохоры s можно определить свой коэффициент $k_s(XY)$. Проверено, что для большинства динуклеотидов k_s не зависит или мало зависит от G+C-состава, хромосомы и длины изохоры. Для динуклеотида CG коэффициент k_s растет с ростом G+C-состава, особенно выражен рост для исключенных выбросов, связанных с т.н. CpG островами. Полученные значения коэффициентов $k(XY)$ для всех динуклеотидов: AA - 1,11; TT - 1,12; AC - 0,83; GT - 0,83; AG - 1,19; CT - 1,19; AT - 0,86; CA - 1,20; TG - 1,21; CC - 1,20; GG - 1,20; CG - 0,27; GA - 0,99; TC - 0,99; GC - 0,98; TA - 0,73. На основе этих данных можно разделить динуклеотиды на четыре группы по характеру эволюционного давления:

- сильное давление против (CG)
- слабое давление против (AC, GT, AT, TA)
- отсутствие давления (GA, TC, GC)
- давление в пользу (AG, CT, CA, TG, AA, TT, CC, GG)

Интересной задачей является сравнение коэффициентов в геномах разных организмов, а также исследование изохор с коэффициентами, сильно отличающимися от средних.

Литература

1. Bernardi G. Isochores and the evolutionary genomics of vertebrates // Gene. 2000. 241, 3–17.

2. Jose L. Oliver et al. Isochore chromosome maps of the human genome // Gene. 2002. 300, 117–127.

3. Pedro Bernaola-Galvan, Ramon Roman-Roldan, Jose L. Oliver. Compositional segmentation and long-range fractal correlations in DNA sequences // Physical review. 1996. V. 53. No. 5.

Слова благодарности

Автор выражает благодарность научному руководителю А. В. Алексеевскому за руководство при выполнении работы.

Иллюстрации

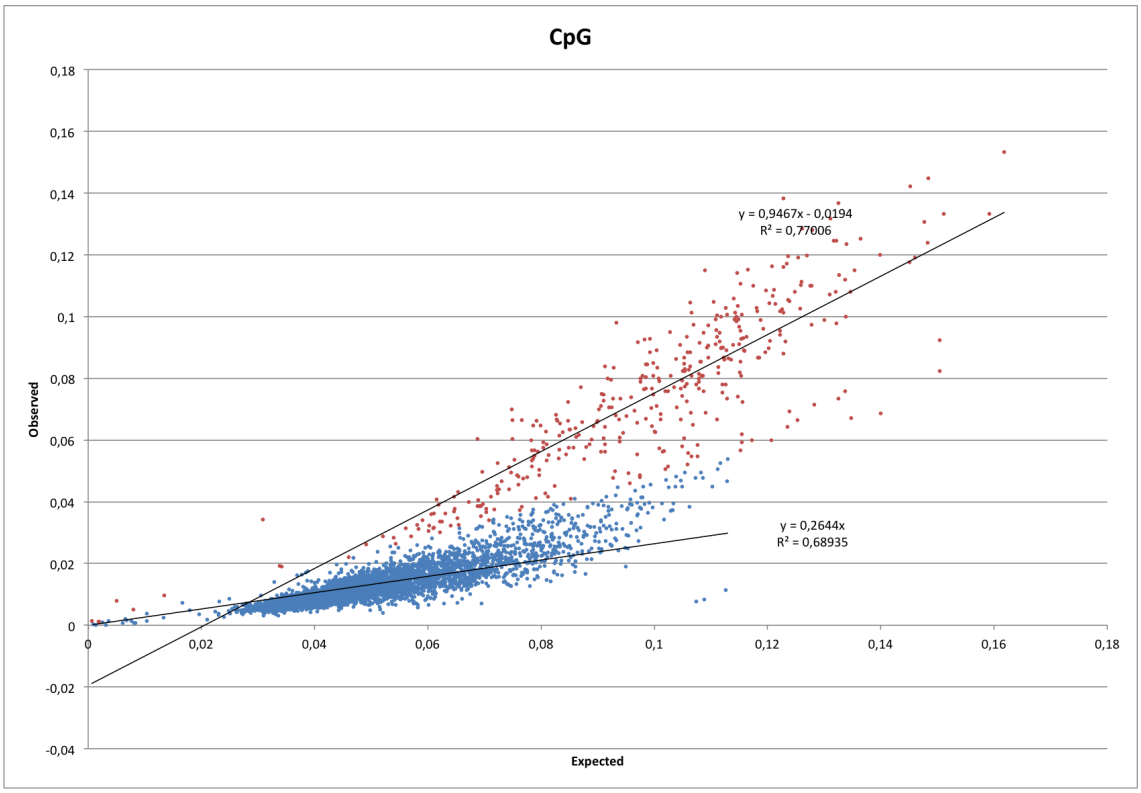

Рис. 1: Зависимость наблюдаемой частоты от ожидаемой для CpG

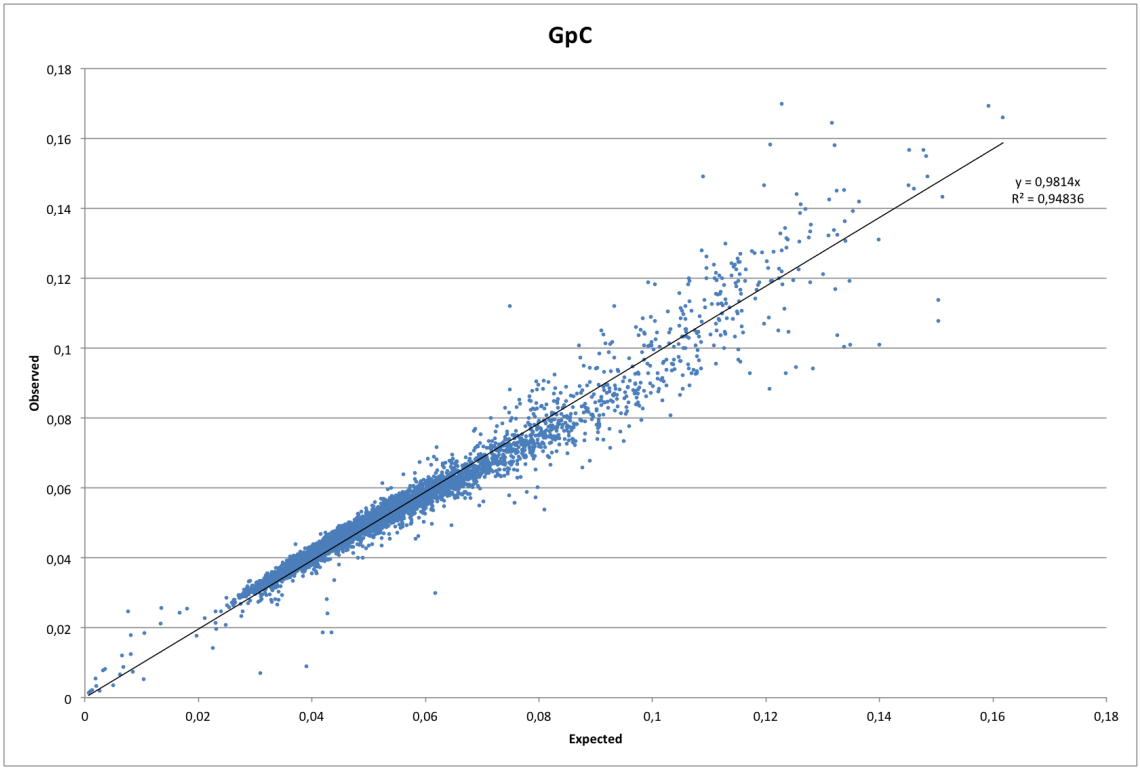

Рис. 2: Зависимость наблюдаемой частоты от ожидаемой для GpC