ИССЛЕДОВАНИЕ ЗАКОНОМЕРНОСТЕЙ, ОБНАРУЖЕННЫХ В ТАБЛИЦАХ ИСПОЛЬЗОВАНИЯ КОДОНОВ.

Кузьмина О.В.

Тюменский государственный университет

e-mail: Angel_RT@qip.ru

Семёнов Д.А.

e-mail: dasem@mail.ru

Одной из характеристик геномов является набор частот встречаемости определенных кодонов в кодирующей части генов. Представление данных в такой форме имеет сходство с традицией, существующей при статистическом анализе литературных текстов. Частота использования определенных букв может характеризовать очень общие свойства языка. Использовать эту информацию для анализа авторства текста или жанра текста уже не перспективно. О содержании текста такая информация просто ничего не говорит. Сама по себе она даже ничего не говорит о ближнем порядке в расположении букв: хотя и позволяет выдвигать некоторые гипотезы на сей счет, но проверка гипотез требует уже привлечения дополнительных данных.

С учетом всего сказанного, неочевидно, что таблицы частот использования кодонов (codon usage) будут содержать важные закономерности, носящие общий характер для всех организмов. Тем не менее, ряд закономерностей выявляется, и можно даже сказать: «бросается в глаза». Так, почти общий характер носят закономерности: 1) Доминирование гуанина над цитозаном в первой позиции кодона (G>C для 100% исследованных геномов) 2) Доминирование аденина над урацилом в среднем по кодирующей части мРНК (для 80% геномов) и для первой позиции кодона (86% геномов). 3) Доминирование урацила над аденином в третьей позиции кодона (более чем для 80% геномов). 4) «Вымывание» аргининовых кодонов CGA и CGG вследствие AT-обогащения генома (замена на AGA и AGG). 5) Пороговый характер встречаемости цитозина и урацила в третьей позиции кодона, указывающий на неустойчивость процесса.

Из всех вышеперечисленных закономерностей в данной работе важны пункты 1-3. То есть предметом нашего рассмотрения будет асимметрия распределения нуклеотидов в среднем по кодирующей части, а также в первой и третьей позиции кодона. Именно эти явления будут нами рассмотрены совместно.

Материалом для анализа первоначально служила выборка из базы данных (http://www.kazusa.or.jp/codon/) , содержащей данные о частоте встречаемости кодонов для различных организмов. Выборка была сформирована путем упорядочения всех данных по объему и отбора наиболее представительных данных. Таким образом, в выборку вошли 1100 «геномов» представленные более чем 50000 триплетов каждый.

На первом этапе были построены распределения для интересующих нас величин, характеризующих в каждом случае асимметрию встречаемости нуклеотидов: 1) доля аденина, по отношению к сумме частот аденина и урацила A/(A+U), 2) то же самое для первой позиции кодона A1/(A1+U1), 3) то же самое для третьей позиции кодона

A3/(A3+U3), 4) доля гуанина по отношению к сумме частот гуанина и цитозина G/(G+C). Исследование полученных распределений привело к выбору инструмента для последующей работы: так как было показано, что распределения значимо отличаются от нормального, то для исследования использовались непараметрические критерии, в частности критерий Колмогорова-Смирнова.

Как известно, в ДНК содержится попарно одинаковое количество аденина и тимина; цитозина и гуанина. Для синтеза мРНК выбирается толька одна из цепей двойной спирали и, как оказалось не случайно. На этот факт указывает исследованная нами асимметрия. На втором этапе была продемонстрирована статистическая значимость обнаруженных закономерностей. Что каждый раз сводилось к демонстрации значимости смещения математического ожидания от значения 0,5.

Накопленное в современной биологии количество фактических данных позволяет выстраивать направление исследования от общего к частному. Первоначально закономерности были найдены в данных по всему геному. Существование возможной специфики на уровне отдельных генов заставило сформировать выборку на основе кодирующей последовательности случайно выбранных генов. На уровне отдельных генов все три закономерности оказались статистически значимы.

На третьем этапе была сформулирована гипотеза, объясняющая все три закономерности на основе представления о совместном влиянии AT-обогащения генома и структуры генетического кода. Характеристикой обнаруженной асимметрии будем считать отношение количества аденина в кодирующей части к суммарному количеству аденина и урацила A/(A+U). Считая критерием AT-обогащения, процент AT-пар приходящийся на кодирующую часть генома, построим зависимость асимметрии от AT-обогащения. В выбранных координатах построили линейную регрессию.

Усредненная асимметрия аденина и урацила действительно хорошо согласуется с этой гипотезой (коэффициент корреляции 0,4). Для асимметрии в первой и второй букве кодона гипотеза не подтвердилась, это можно сформулировать так: даже если аденина мало, то в первой позиции кодона его больше, чем урацила; даже если урацила мало, то в третьей позиции его больше, чем аденина.

На четвертом этапе работы нас интересовало последовательное сочетание кодонов, поэтому одних данных по частоте встречаемости было не достаточно. Выборка, составленная на основе отдельных генов, позволяет проверить простую гипотезу: накопление урацила в третьей позиции кодона и аденина в первой позиции можно объяснить из предположения о наличии взаимодействия между третьей и первой позициями соседних последовательных кодонов. Иными словами, ожидается не случайное сочетание кодонов вида (NNU)(ANN). Повышенная частота таких сочетаний могла бы обеспечивать большую частоту встречаемости стоп-кодонов при сдвиге рамки считывания на две позиции вправо.

Вопреки ожиданиям, проверка данной гипотезы продемонстрировала крайне низкое число сочетаний данного типа. Даже при сдвиге рамки считывания стоп-кодоны избегаются, что противоречит работе [1]. Более того, хотя одновременно существует доминирование аденина в первой и доминирование урацила в третьей позиции кодона, но величины

характеризующие эти процессы коррелируют значимо отрицательно (-0,7). То есть сильная асимметрия в первой позиции не сопровождается сильной асимметрией по третьей позиции, что может интерпретироваться как следствие избегания комбинаций (NNU)(ANN).

Одновременно, обращает на себя внимание значительная частота сочетаний кодонов вида: (NNU)(GNN). Можно предположить, что данное сочетание влияет на фолдинг мРНК, что позволяет рассматривать данный результат в русле работ [2, 3].

Гуанин, также как и аденин значимо накапливается в первой позиции кодона (причем в 100% исследуемых геномов). На материале отдельных генов продемонстрировано накопление пар UG, где урацил находится в третьей, а гуанин в первой позиции соседних кодонов. Возникает вопрос: насколько взаимное влияние урацила и аденина позволяет объяснить общую тенденцию накопления этих нуклеотидов в данных позициях? Для ответа на этот вопрос исследуем связь двух величин: 1) разность частот гуанина в первой позиции и в среднем по геному 2) разность частот урацила в третьей позиции и в среднем по геному. Две эти величины оказываются линейно связаны (коэффициент корреляции 0,678). То есть гипотеза о взаимном влиянии урацила и гуанина в этих позициях хорошо объясняет закономерности накопления урацила в третьей, а гуанина в первой позициях кодона.

1. Г.Г. Малинецкий, С.А. Науменко, А.В. Подлазов Об экстремальных свойствах разметки генетического кода. ДАН 2007 т.414 №6 с.831-835
2. T. Warnecke L.D. Hurst GroEL dependency affects codon usage—support for a critical role of misfolding in gene evolution. Molecular Systems Biology 6:340
3. D.A. Semenov Evolution of the genetic code. Emergence of stop codons arXiv:0710.5825 2007