

Обзор протеома археи *Acidilobus saccharovorans* 345-15

Аксенова М.А.

Факультет биоинженерии и биоинформатики МГУ им. М.В.Ломоносова

РЕЗЮМЕ

Данная работа была проведена с целью изучения протеома археи *Acidilobus saccharovorans* 345-15. Было изучено распределение длин белков археи, а также расположение белков и РНК в геноме, т.е. их положение на прямой и обратной цепях. Была проверена гипотеза случайного распределения генов на прямой и обратной цепях ДНК.

1. ВВЕДЕНИЕ

Архея *Acidilobus saccharovorans* 345-15 является облигатным анаэробом, термоацидофилом^{*}, растущим при pH 2,5-5,8 и температурах от 60 до 90°C.

Геном данной археи полностью секвенирован и содержит одну кольцевую хромосому. Его особенностью является отсутствие плазмид.

Цель данной работы состояла в исследовании распределения длин белков, описании особенностей генов археи и их расположения в геноме.

Также целью являлось овладение навыками работы с Microsoft Office Excel 2007 и обучение основным правилам написания статей и обзоров^{**}.

2. МЕТОДЫ

В работе были использованы файлы, скачанные из базы данных National Center for Biotechnology Information. Для обработки и анализа информации была использована программа Microsoft Office Excel 2007. При этом были применены следующие возможности программы:

- Импорт скачанных файлов NC_014374.ptt и NC_014374.rnt в Excel-файл;
- Функция «СЧЁТЕСЛИМН» – для подсчета частоты встречаемости белков в данном диапазоне длин, а также для подсчета количества генов на прямой и обратной цепях ДНК;
- Функция «МИН» и «МАКС» – для определения соответственно минимальной и максимальной длин генов;
- Функция «СРЗНАЧ» – для определения средней длины гена;
- Функция «МОДА» – для определения моды длин белков;
- Функция «БИНОМРАСП» – для подсчета статистической вероятности случайности расположения генов на прямой или обратной цепи ДНК;
- Построение гистограммы распределения длин белков по набору значений;

^{*}Экстремофильные археи, выживающие при высокой кислотности и содержании серы, а также высоких температурах.

^{**}Эти цели напрямую не относятся к данному обзору, но являются важной частью программы курса биоинформатики.

- Автозаполнение.

На листы Excel-файла “NC_014374.ptt” и “NC_014374.rnt” были импортированы соответствующие файлы, скачанные с NCBI, в качестве разделителя был использован знак табуляции.

Для подсчета количества генов на прямой и обратной цепях на листе “strands” была составлена таблица, содержащая направление гена, его идентификатор и тип (CDS или RNA, т.е. кодирующий белок или РНК ген).

На листе “proteins” была составлена таблица “идентификатор гена – длина гена”, где длины генов в аминокислотных остатках (а.о.) были отсортированы по возрастанию.

3. РЕЗУЛЬТАТЫ

В результате изучения протеома археи *Acidilobus saccharovorans* 345-15 было выявлено, что ее геном содержит всего одну кольцевую хромосому. Всего в геноме 1547 генов, из них 1499 (что составляет примерно 96,9%) кодируют белки, а 48 (т.е. 3,1%) кодируют тРНК и рРНК. Видно, что первых значительно больше.

Найдены кодирующие последовательности с длинами, не кратными трем.

3.1 Распределение длин белков

Длины белков представлены в диапазоне от 38 до 1384 а.о. Средняя длина белка составляет примерно 299,7 а.о., а мода – 185.

Распределение длин белков представлено в виде гистограммы (диаграмма 1). Сведения о белках, кодирующих РНК, не учитывались.

Диаграмма 1. Распределение длин белков в протеоме археи *Acidilobus saccharovorans* 345-15. По горизонтали указаны диапазоны длин белков в аминокислотных остатках, по вертикали – количества белков с данной длиной.



3.2 Распределение генов по прямой и обратной цепям

Данные по распределению генов по прямой и обратной цепям представлены в таблице 1.

Таблица 1. Распределение генов, кодирующих белки и РНК, по прямой и обратной цепям.

	CDS	RNA	Всего
Прямая цепь (+)	714	23	737
Обратная цепь (-)	785	25	810
Всего	1499	48	1547

Из таблицы видно, что гены распределены по цепям примерно поровну.

Для проверки гипотезы случайного (с вероятностью 0,5) распределения генов по цепям было подсчитано биномиальное распределение. Результаты подсчетов указаны в таблице 2.

Таблица 2. Р-значения для генов, кодирующих белки и РНК.

CDS	RNA
0,968550607	0,442716749

Из таблицы видно, что гипотеза верна и для генов белков, и для генов РНК.

4. ОБСУЖДЕНИЕ

Длины белков при распределении не образуют симметричного купола, поэтому среднее значение и мода так отличаются. Наибольшее количество белков имеют длину 100-150 а.о.

То, что в геноме археи есть кодирующие последовательности, длины которых не кратны трем, объясняется тем, что эти последовательности считываются рамками, не равными 0 (т.е. рамками считывания -2, -1, +1, +2,).

Полученные при проверке гипотезы случайного распределения генов по цепям Р-значения позволяют предположить, что такое распределение случайно и для генов, кодирующих белки, и для генов, кодирующих РНК.

5. СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

(1) Файл Excel

http://kodomo.fbb.msu.ru/~akmarina/term1/proteome_Acidilobus_saccharovorans_345-15.xlsx

6. БЛАГОДАРНОСТИ

Выражается благодарность преподавателям биоинформатики А. Алексеевскому и И. Русинову за обучение основам работы с Excel и предоставление шаблона для написания данного обзора