УДК 519.724

АЛГОРИТМ КОДИРОВАНИЯ ПРОЦЕССА ТРАНСЛЯЦИИ БЕЛКОВ В КЛЕТКЕ

М.А. ПРОТЬКО, О.Ф. БОРИСЕНКО

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь

Поступила в редакцию 19 марта 2023

Аннотация. Целью данной статьи является рассмотрение азотистых оснований в соотношении с кодируемыми ими аминокислотами, с дальнейшими поисками их взаимосвязи.

Ключевые слова: четверичный код, генетический код, трансляция.

Введение

Если поставить целью создание динамически развивающейся системы (системы, способной реагировать на условия, изначально не предусмотренные при ее проектировании, но потенциально возможные [1]), необходимо четко разграничить все ее параметры, или же, поставить строго структурированную формальную задачу.

Для достижения поставленной цели рассмотрим структуру генетического кода, свойственного ДНК и РНК, для выявления неких закономерностей.

В данной работе используются определения как из теории кодирования, так и из общей генетики. В начале следует определение из теории кодирования, затем, в скобках, из общей генетики.

Определение объекта

Рассмотрим предметную область. Генетическим кодом называется последовательность четырех азотистых оснований (аденин (А), тимин (Т) / урацил (У), гуанин (Г) и цитозин (Ц)) которым в соответствие ставиться 20 аминокислот (см. табл. 1).

Определим генетический код следующим образом, воспользовавшись [2]:

Положим, что существует некий источник, выдающий дискретное сообщение a (полипептиды и/или белки), которое можно рассматривать как последовательность элементарных сообщений a_i (аминокислоты). Эти элементарные сообщения – символы, и их совокупность $\{a_i\}$ – алфавит.

Пусть последовательность символов источника a заменяется последовательностью кодовых символов (триплетом, или же кодоном).

Элементарными символами кодовой комбинации в данном случае служат азотистые основания.

Общее число символов, составляющих кодовую комбинацию (длина кода) $n = 3,$ количество значений кодовых признаков (основание кода) $m = 4.$

Емкость кода $N_o = 64.$

Количество сообщений $N_a = 20.$

Кодовое расстояние $d_0 = 1.$

Относительная скорость кода $R_k = \log_2 N_a / \log_2 N_0 = 2,996 / 4,159 = 0,720.$

Избыточность $c_k = 1 - R_k = 0,280.$

Табл. 1. **Соответствия аминокислота – триплет**

Название	Частота	Кол-во кодонов	Кодоны	Мин. к.р.
Лейцин Leu	9,68	6	УУА УУГ ЦУУ ЦУЦ ЦУА ЦУГ	2 2
Аланин Ala	8,76	4	ГЦУ ГЦЦ ГЦА ГЦГ	1–2
Серин Ser	7,14	6	УЦУ УЦЦ УЦА УЦГ АГУ АГЦ	2 1
Глицин Gly	7,03	4	ГГУ ГГЦ ГГА ГГГ	1
Валин Val	6,73	4	ГУУ ГУЦ ГУА ГУГ	1
Глютаминовая кислота Glu	6,32	2	ГАА ГАГ	2
Аргинин Arg	5,78	6	ЦГУ ЦГЦ ЦГА ЦГГ АГА АГГ	2 1
Треонин Thr	5,53	4	АЦУ АЦЦ АЦА АЦГ	2
Асппарагиновая кислота Asp	5,49	2	ГАУ ГАЦ	2
Изолейцин Ile	5,49	3	АУУ АУЦ АУА	2
Лизин Lys	5,19	2	ААА ААГ	2
Пролин Pro	5,02	4	ЦЦУ ЦЦЦ ЦЦА ЦЦГ	2
Аспарагин Asn	3,93	2	ААУ ААЦ	2
Глютамин Gln	3,90	2	ЦАА ЦАГ	3
Фенилаланин Phe	3,87	2	УУУ УУЦ	1
Тирозин Tyr	2,91	2	УАУ УАЦ	3
Метионин Met	2,32	1	АУГ	3
Гистидин His	2,26	2	ЦАУ ЦАЦ	2
Цистеин: Cys	1,38	2	УГУ УГЦ	1
Триптофан Trp	1,25	1	УГГ	–

Генетический код (далее г.к.) является равномерным ($n = const$) и многопозиционным (хромосомный г.к.), однако, существует и неравномерный г.к. (митохондриальный г.к.).

По форме представления в канале передачи (процесс кодирования, или же трансляции, переход РНК – белок) – г.к. имеет параллельную форму.

По основным законам кодообразования, г.к. – комбинаторный код.

Рассмотрим табл. 1, являющуюся объединением информации из источников [3] и [4].

В табл. 1 в столбце «название» находится аминокислота с ее русским названием и английским сокращением, частота встречаемости основана на выборке из 7 555 843 062 аминокислот. Данные частоты являются усредненными показателями по всем царствам.

Положим за кодовое расстояние количество различающихся элементарных символов в кодоне.

Минимальное кодовое расстояние (Мин. к.р.) – разница между элементарными символами пары кодонов, рассматриваемых по их частоте встречаемости (Leu – Ala, Ala – Ser и т.д.). В случае, если аминокислота имеет кодоны, кодовое расстояние между которыми более 1, данные кодоны разделяются на группы. К каждой такой группе записано свое кодовое расстояние, рассчитанное по аналогичному принципу.

Как видно из табл. 1, количество кодовых групп (кодонов) никак не зависит от частоты встречаемости, кодируемой ими аминокислоты.

Кодовое расстояние между кодонами, кодирующими одно и то же основание, чаще всего, не превышает 1. Исключения: аргинин, серин и лейцин (6 кодовых последовательностей).

Кодовое расстояние между разными основаниями чаще всего – 2.

Никакой ярко выраженной закономерности между частотой встречаемости и кодовым расстоянием не наблюдается.

Из вышеописанного можно сделать вывод, что г.к. не является оптимальным (одно из свойств вырожденности).

Г.к. является помехоустойчивым кодом, поскольку позволяет при обнаружении ошибочной последовательности завершать трансляцию (таких последовательностей всего три, это «стоп» кодоны).

Г.к. очень близок к полному коду, согласно определению из [5].

Если положить, что «бессмысленные» последовательности («стоп» кодоны) являются разрешенными, то г.к. – полный код.

Рассмотрим пример части последовательности 11 хромосомы человека [6]:

АУГ ГУЦ ЦУГ ГГУ ГГЦ АУГ ГАГ ЦУЦ УУГ ЦАЦ ЦУЦ УАГ Г…

Met Val Leu Gly Gly Tyr Glu Leu Leu His Leu Стоп

Если предположить, что такая последовательность действительно кодирует некий белок, можно сделать следующий вывод: один и тот же кодон в случае, если он относится к группе избыточных, не будет повторяться.

Рассмотрим последовательность, преобразованную циклическим сдвигом:

ЦУУ ГЦУ ГГУ ГАА ЦГУ

Leu Ala Gly Glu Arg

УУГ ЦУГ ГУГ ААЦ ГУЦ

Leu Leu Val Thr Val

Можно заметить, что полученная таким образом комбинация все еще имеет смысл.

Если определить операции над множеством элементарных сигналов, а также матрицу строк, являющуюся аналогией матрицы полного кода, с линейной независимостью строк, можно сказать, что г.к. – циклический код.

Если же определить операции над множеством элементарных сигналов согласно свойствам, описанным в [7] и с учетом описанных алгоритмов в [6], получим, что г.к. – четверичный код.

При определении операции над множеством элементарных сигналов стоит отталкиваться от смысла изначального алфавита (аминокислоты и «текста» белков), поскольку от него будет зависеть образующая строка матрицы разрешенных последовательностей циклического кода.

Если предположить, что в нашей системе, где будет использоваться алгоритм г.к., существует процесс, подобный мутации и кроссинговеру, то количество запрещенных

последовательностей характеризует выраженность данного процесса. Чем меньше число «стоп» кодонов, тем больше избыточность, или же, тем меньше стабильность.

Пример на основании алгоритма Шенона-Фоне

Рассмотрим способ замены последовательности символов источника a (20 аминокислот) кодовыми символами с $n=3$ и $m=4$. Воспользуемся алгоритмом Шенона-Фоне. Положим, что частота встречаемости символов, как и сами символы источника, аналогичны представленным в табл. 1.

Результаты представлены в табл. 2.

Табл. 2. **Соответствия при *n*=3**

Название	Кодоны	Кол-во кодонов	Мин. к.р.
Лейцин Leu	ЦЦ-	4	1
Аланин Ala	ЦГ-	4	1
Серин Ser	ЦУ-	4	2
Глицин Gly	ГЦ-	4	1
Валин Val	ГГ-	4	1
Глютаминовая кислота Glu	ГУ-	4	1
Аргинин Arg	ГА-	4	2
Треонин Thr	УЦ-	4	1
Асппарагиновая кислота Asp	УГ-	4	1
Изолейцин Ile	УУ-	4	1
Лизин Lys	УАЦ	1	1
Пролин Pro	УАГ	1	2
Аспарагин Asn	АЦ-	4	1
Глютамин Gln	АГ-	4	1
Фенилаланин Phe	АУЦ	1	1
Тирозин Tyr	АУГ	1	2
Метионин Met	ААЦ	1	1
Гистидин His	ААГ	1	1
Цистеин: Cys	ААУ	1	1
Триптофан Trp	ААА	1	–

Получаем код из 56 значащих последовательностей и 8 запрещенных. Стоит учитывать то, что минимальное кодовое расстояние в данном случае у большинства символов источника – 1 (посчитанное по тому же принципу, что и в табл. 1).

Пример на основании кодового расстояния

Рассмотрим способ замены последовательности символов источника кодовыми символами с учетом кодового расстояния.

Пусть имеется последовательность символов источника a (Leu, Ala, Ser, Gly, Val).

Сохраняя соотношения, между емкостью кода и количеством символов источника, получим емкость кода $N_o = 16$.

Таким образом, длина кода $n=2$ с основанием $m=4$.

Полный код при данных условиях:

$$\pi_4^2 = \begin{vmatrix} А\,А\,А\,А\,Г\,Г\,Г\,Г\,У\,У\,У\,У\,Ц\,Ц\,Ц\,Ц \\ А\,Г\,У\,Ц\,А\,Г\,У\,Ц\,А\,Г\,У\,Ц\,А\,Г\,У\,Ц \end{vmatrix}.$$

Выбор последовательности кодовых символов из полного кода для символов источника будем совершать таким образом, чтобы кодовое расстояние между двумя соседствующими по частоте встречаемости символами источника было максимальным.

Таким образом, получим соотношения, представленные в табл. 3.

Табл. 3. **Соответствия при *n*=2**

Leu	Ala	Ser	Gly	Val
АЦ	ГЦ	АА	ГГ	УУ
АУ	ГА	АГ	ГУ	УГ
ЦЦ	УЦ	ЦА	ЦГ	УА

В табл. 3 представлен один из вариантов кода, получаемого из полной последовательности. В данном случае, самые отличающиеся по характеристикам варианты – те, у которых разное число запрещенных последовательностей, (число «стоп» кодонов – от 1 до 6).

Заключение

Ключевое свойство, приводящее одновременно и к стабильности, и к изменчивости систем на основе генетического кода – это его избыточность. Причем избыточность такого рода, что при возникновении ошибки (мутации) вероятность критических изменений смысла сообщения остается достаточно малой (кодон либо переходит в свой эквивалент, либо в иную существующую аминокислоту). Т.е., для построения системы с подобным свойством достаточно выбора подходящего веса, рассматривая ошибки как благо. Разрядность кода не играет столь существенной роли в достижении этого свойства.

Дальнейший анализ полученных кодовых последовательностей заключается в сборе статистики при использовании данных кодовых последовательностей в генетических алгоритмах, по таким параметрам как количество популяций на алгоритм кодирования (отсчет популяции заканчивается при нахождении решения или вырождении).

ALGORITHM FOR ENCODING THE PROCESS OF PROTEIN TRANSLATION IN A CELL

M.A. PROTSKO, O.F. BORISENKO

Abstract. The aim of this article is to find correlation between nitrogenous bases in relation to the amino acids they encode.

Keywords: quaternary code, genetic code, translation.

Список литературы

1. Протько М.А., Борисенко О.Ф. // Простейшие шифры и генетический алгоритм. 2023.
2. Кузьмин И.В., Кедрус Н.А. Основы теории информации и кодирования. 1986.
3. Каминская Э.А. Общая генетика. Минск, Вышэйшая школа. 1992.
4. Kozlowski L.P. // Proteome-pI: proteome isoelectric point database. Nucleic Acids Research. 45 (D1): D1112–D1116. doi:10.1093/nar/gkw978. PMC 5210655. PMID 27789699
5. Варакин Л.Е. Системы связи с шумоподобными сигналами. Москва, Радио и связь.1985.
6. Фрагмент генетичекого кода из 11 хромосомы человека [Электронный ресурс]. URL: http://www.ensembl.org/Homo_sapiens/Info/Index?db=core
7. Зиновьев Д.В., Соле П. // Пробл.передачи информ. 2004, Т. 40, В. 2. С. 50–62.