Генератор нуклеотидных последовательностей для хранения и передачи небиологической информации в молекулах ДНК

О.Ю. Кирьянова[1], И.И. Кирьянов[2], Р.Р. Гарафутдинов[3], А.В. Чемерис[3], И.М. Губайдуллин[1,4]

[1]*Уфимский государственный нефтяной технический университет, Космонавтов 1, Уфа, Россия, 450062*
[2]*ООО «Корнинг СНГ», Шателена, 26а, Санкт-Петербург, Россия, 194021*
[3]*Институт биохимии и генетики, обособленное структурное подразделение УФИЦ РАН, проспект Октября 71, Уфа, Россия, 450054*
[4]*Институт нефтехимии и катализа, обособленное структурное подразделение УФИЦ РАН, проспект Октября 141, Уфа, Россия, 450075*

Аннотация
В работе представлены результаты разработки генератора нуклеотидных последовательностей с целью применения молекул ДНК для кодирования небиологической информации и ее стеганографической передачи. Программа позволяет генерировать случайные последовательности с учетом таких параметров как длина, процентное содержание гуанина и цитозина, наличие/отсутствие гомополимерных участков. Полученные последовательности являются прообразами реальных молекул ДНК, которые необходимо синтезировать в лабораторных условиях.

Ключевые слова
Нуклеотидная последовательность, хранение небиологической информации в ДНК, стеганография, генерация случайных последовательностей, Python

1. Введение

Молекула ДНК представляет собой природный биополимер, содержащий четыре азотистых основания – аденин (A), цитозин (C), гуанин (G), тимин (T). Порядок расположения нуклеотидов в том или ином фрагменте ДНК может быть установлен с помощью секвенирования ДНК. Кроме того, молекулы ДНК с заданными последовательностями нуклеотидов могут быть искусственно синтезированы химическим или ферментативным способом. Данные особенности молекулы ДНК позволяют кодировать и хранить в ней информацию любого типа. Известны разные способы цифровой кодировки отдельных азотистых оснований в цепочках ДНК, применяющие различные системы счисления [1]. При этом существующие способы не используют некоторую вырожденность цифрового кода азотистых оснований в ДНК, что на самом деле крайне важно для оцифровки небиологических данных.

2. Система конструирования различных NYRN-олигонуклеотидов

Предлагаемый авторами работы способ кодирования больших объемов небиологической информации заключается в применении протяженных последовательностей ДНК, состоящих из NYRN-олигонуклеотидов. Информационная часть (YR) является кодирующей и несет в себе информацию о бинарном коде символа таблицы ASCII (1 байт). Таким образом, каждый символ представляет собой набор из восьми бит информации и кодируется бинарным представлением чисел от 0 до 255. Каждый бит информации соответствует азотистому основанию: «1» – A, G (R - пурины), «0» – C, T (Y - пиримидины). Последовательность,

кодирующая символы таблицы ASCII, должна удовлетворять следующим условиям: 1) длина – 8 нуклеотидов, 2) GC-состав 35-65%, 3) гомополимерные участки длиной более 3 нуклеотидов должны быть исключены, 4) не должно быть многократно повторяющихся мотивов (например, AGAGAGA). Служебная часть (N--N) должна удовлетворять следующим требованиям: 1) длина – 12 нуклеотидов, 2) GC-состав 35-65%, 3) не должно быть мотивов, состоящих из четырех G и C (GGGG, GCGC, CGCG и т.д.), 3) гомополимерные участки длиной более 3 нуклеотидов должны быть исключены. Кроме того, необходима генерация маскирующей последовательности, которая необходима при стеганографической передаче данных. К ней предъявляются следующие требования: 1) варьируемая длина порядка 1000 нуклеотидов, 2) GC-состав 35-65%, 3) отсутствие гомополимерных участков.

Существующие на данный момент генераторы случайных нуклеотидных последовательностей достаточно просты и не позволяют учитывать представленные выше требования в комплексе [2, 3]. Поэтому для этих целей была разработана программа, позволяющая генерировать последовательности с варьируемыми начальными условиями. Программа написана на языке программирования Python. Разработанный генератор дает более широкие возможности для исследования нуклеотидных последовательностей не только в рамках кодирования небиологической информации, но и для решения других задач в области исследования ДНК. Входными данными являются: необходимое количество последовательностей, GC-состав, наличие/отсутствие гомополимерных участков, палиндромов, повторов. Сначала формируется случайная строка с помощью модуля random. Далее она «дорабатывается». В случае наличия гомополимерных участков, палиндромов, проводится замена символов. Проверяется доля символов G и C. Если строка удовлетворяет заданным требованиям, она сохраняется в списке. Действия повторяются до тех пор, пока не будет получено нужное количество строк. После чего производится вывод результатов на экран или запись в текстовый документ.

3. Заключение

Разработанная программа представляет собой генератор нуклеотидных последовательностей, который позволяет получать различные последовательности нуклеотидов в зависимости от заданных требований. Планируется разработка данного генератора в формате web-приложения.

4. Благодарности

Работа выполнена при финансовой поддержке гранта РФФИ № 20-07-00222.

5. Литература

[1] Сахабутдинова, А.Р. Небиологическое применение молекул ДНК / А.Р. Сахабутдинова, К.И. Михайленко, Р.Р. Гарафутдинов, О.Ю. Кирьянова, М.А. Сагитова, А.М. Сагитов, А.В. Чемерис // Биомика – 2019. – Т. 11, № 3. – С. 344-377. DOI: 10.31301/2221-6197.bmcs.2019-28.

[2] Medina-Rivera, A. RSAT 2015: Regulatory Sequence Analysis Tools / A. Medina-Rivera, M. Defrance, O. Sand, C. Herrmann, J. Castro-Mondragon, J. Delerce, S. Jaeger, C. Blanchet, P. Vincens, C. Caron, D. Staines, B. Contreras-Moreira, M. Artufel, L. Charbonnier-Khamvongsa, C. Hernandez, D. Thieffry, M. Thomas-Chollier, J. van Helden // Nucleic Acids Research. – 2015. – Vol. 43(W1). – P. W50-W56.

[3] Random DNA Sequence Generator [Электронный ресурс]. – Режим доступа: http://www.faculty.ucr.edu/~mmaduro/random.htm (10.12.2020).