

АНАЛОГОВОЕ ПРЕДСТАВЛЕНИЕ ЭЛЕМЕНТАРНЫХ ЗВЕНЬЕВ ДНК

Тверетин А.А., Подолян В.М.

Самарский государственный технический университет

Самара, Россия

Четыре органические молекулы (нуклеотиды) - аденин (А), гуанин (G), цитозин (С) и тимин/урацил (Т/U) являются алфавитом, с помощью которого осуществляется кодирование всей генетической информации. Известно, что генетический код избыточен, при этом, 61 триплет кодирует только 20 аминокислот (еще 3 триплета являются старт- и стоп-кодонами: Ochre, Amber, Opal).

Формальное представление АГТС-элементов и связей между двумя элементами комплементарного и последовательного характера ранжируется с учётом некоторых критериев. В связи с тем, что модель биологической структуры должна наиболее достоверно отражать её информационную сущность, желательно, чтобы критерий был не один, а физическая природа критериев не должна быть одинаковой. Ниже в качестве таковых используются особенности физического характера.

Все 64 триплета представлены в виде массивов $VF[3,1]$, элементами которых являются весовые эквиваленты сочетаний нуклеотидов вида X_1X_2 , которые зависят от таких физических параметров, как молекулярная масса и сила водородных связей между комплементарными нуклеотидами. Альтернативным путем является представление каждого нуклеотида в виде символов четырехзначного счисления, упорядоченных по убыванию их молекулярной массы.

Представив последовательность трех значений массивов в виде кривой, можно описать ее с помощью некоторой функции $Fx_1x_2x_3$, где $x_1x_2x_3$ - последовательность нуклеотидов в триплете, $i=1..3$.

Функции триплетов были получены в программе MathCAD 9. В данной программе реализована возможность выполнения линейной регрессии общего вида. При ней заданная совокупность точек приближается функцией вида $F(x,K1,K2,...,Kn)=K1F1(x)+K2F2(x)+...+KnFn(x)$. Таким образом, функция регрессии является линейной комбинацией функций $F1(x),F2(x),...,Fn(x)$, причем сами эти функции могут быть нелинейными, что резко расширяет возможности такой аппроксимации и распространяет ее на нелинейные функции. Для реализации линейной регрессии общего вида используется функция $linfit(VX,VY,F)$. Эта функция возвращает массив коэффициентов линейной регрессии общего вида K , при котором среднеквадратичная погрешность

приближения облака исходных точек, если их координаты хранятся в векторах VX и VY , оказывается минимальной. Массив F должен содержать базисные функции $F1(x)$, $F2(x)$, ..., $Fn(x)$, записанные в символьном виде, причем количество элементов массива F должно быть меньше чем количество элементов массивов VX и VY .

Для оценки точности аппроксимации использована нуклеотидная последовательность длиной 150 нуклеотидов или 50 триплетов 5'-концевого сегмента 16S-рРНК *E.coli*. Подсчитано, что точность аппроксимации составила 5,3%.

Применение числовых эквивалентов позволило представить маски аминокислот в виде функций с небольшой погрешностью. Это создает предпосылки для дальнейшего укрупнения элементарных нуклеотидных последовательностей, и формирования базы данных образов по их функциональной принадлежности.