Análisis de secuencias

Resumen

El análisis de secuencias es una herramienta clave en bioinformática que permite descifrar la información contenida en las secuencias de ADN, ARN y proteínas. A través de modelos computacionales y estadísticos, es posible estudiar patrones, predecir funciones y entender la relación (evolutiva) entre secuencias y su impacto biológico. El objetivo de este curso es entender cómo y por qué analizamos secuencias biológicas, enfatizando en el fundamento algorítmico y biológico de estas herramientas.

Sandra Mingo Ramírez

UAM - 2024/25

28 de septiembre de 2024 00:23

Universidad Autónoma de Madrid
Bioinformática y Biología Computacional

Código en Github

Índice general

Capítulo I

Modelos estadísticos en el análisis de secuencias

En biología existen tres tipos fundamentales de secuencias: ADN, ARN y aminoácidos. Estas secuencias representan el núcleo del dogma central de la biología molecular, ya que almacenan toda la información genética necesaria para perpetuar la vida. No obstante, el análisis de estas secuencias presenta una gran complejidad debido a su estructura y funcionalidad. Aunque son polímeros lineales formados por un número limitado de subunidades o monómeros, la disposición específica de estas subunidades a lo largo de la molécula es clave para su función. Debido a su estructura química, estas secuencias pueden representarse como cadenas de símbolos, lo que permite su estudio utilizando herramientas matemáticas y computacionales.

I.1. Modelo multinomial

Desde un enfoque matemático, las secuencias biológicas pueden verse como una concatenación de símbolos provenientes de un alfabeto definido: en el caso del ADN, el alfabeto es ATCG; en el ARN, AUCG; y en las proteínas, está compuesto por 20 aminoácidos. Formalmente, una cadena es una secuencia finita de símbolos de un alfabeto Σ. Así, Σ^3 representa todos los codones posibles dentro del código genético.

Ejemplo práctico: Consideremos un ejercicio de modelización estadística aplicado al ADN. En un experimento ChIP-seq (una técnica de secuenciación masiva que permite identificar sitios de unión de proteínas al ADN), se descubrieron 500 sitios de unión para un factor de transcripción. Dado que el genoma humano contiene entre 20,000 y 26,000 genes, estos 500 sitios parecen pocos. Sin embargo, la cuestión central es si esta cantidad es coherente con lo que se esperaría bajo un modelo estadístico. Los factores de transcripción se unen a subsecuencias específicas de ADN llamadas "motivos de respuesta". En este caso, el motivo de unión es RCGTG, donde R representa A o G. Aunque las moléculas biológicas interaccionan con cierta flexibilidad, este motivo es bastante restringido, ya que solo una posición es flexible. El genoma humano tiene alrededor de 3×10^9 bases, por lo que podemos calcular la cantidad esperada de sitios de unión basándonos en la probabilidad de que este motivo ocurra aleatoriamente. Asumiendo que los nucleótidos son independientes entre sí y tienen la

misma probabilidad de aparecer, la probabilidad de que aparezca la secuencia CGTG es $0,25^4$. Para la posición R, que puede ser A o G, la probabilidad es $0,5$. Por tanto, la probabilidad total de encontrar el motivo RCGTG es $0,25^4 \times 0,5 = \frac{1}{512}$, es decir, se esperaría encontrar esta secuencia una vez cada 512 posiciones. Con un genoma de 3×10^9 bases, se esperaría aproximadamente $\frac{3\times10^9}{512} \approx 6 \times 10^6$ sitios. Sin embargo, en el experimento solo se hallaron 500 sitios, lo que sugiere que el modelo experimental no refleja completamente la realidad biológica y es necesario recurrir a otros modelos, aunque sean simplificados. La secuencia por sí sola no es suficiente para que el factor de transcripción se una. Otros factores, como la accesibilidad de la cromatina, también juegan un papel crucial. No obstante, el modelo multinomial proporciona una referencia útil para evaluar los datos experimentales en un contexto aleatorio. En este modelo multinomial, se asume que los nucleótidos en cada posición son independientes y tienen la misma probabilidad de aparecer (hipótesis nula), lo que simplifica los cálculos:

$$\prod_{i=1}^{n} p(s_i)$$

Si bien este enfoque es sencillo, tiene limitaciones significativas, como la suposición de independencia entre nucleótidos. Sabemos que esto no es siempre cierto, por ejemplo, los dinucleótidos CG suelen ser menos frecuentes salvo en las "islas CpG", donde existe una gran concentración.

I.1.1. Frecuencia de dinucleótidos

Los dinucleótidos, que representan todas las combinaciones posibles de dos nucleótidos (Σ^2), deberían tener una frecuencia esperada de $\frac{1}{16}$ en el genoma humano. Al analizar las frecuencias observadas en el cromosoma 21, se encuentra que A y T aparecen con una frecuencia del 29.5 %, mientras que G y C con un 20.5 % (Figura I.1). Al recalcular las frecuencias de los dinucleótidos, se observa que, en general, la frecuencia observada coincide con la esperada, excepto para el dinucleótido CG, cuya frecuencia observada es tres veces menor a la esperada. Esto sugiere que los nucleótidos no son completamente independientes, y el modelo multinomial no es suficiente para describir esta dependencia.

I.2. Cadena de Markov

Para modelar adecuadamente fenómenos como las islas CpG, es necesario utilizar un enfoque diferente: las cadenas de Markov. En estos modelos, cada posición en la secuencia de ADN se considera un "estado" que corresponde a uno de los cuatro nucleótidos, y la probabilidad de que un nucleótido aparezca en una posición depende del nucleótido anterior. Este tipo de dependencia se representa mediante una matriz de transición que contiene las probabilidades condicionales de pasar de un nucleótido a otro (Figura I.2). Matemáticamente, un modelo de Markov de primer orden puede expresarse de la siguiente manera:

Dinucl.	Observ	Expect	Diff	NormD
AA	9.77 %	8.69 %	+1.08	0.12
AC	5.08 %	6.02 %	-0.94	0.16
AG	6.92 %	6.05 %	+0.87	0.14
AT	7.71 %	8.72 %	-1.01	0.12
CA	7.29 %	6.02 %	+1.27	0.21
CC	5.1 %	4.17 %	+0.93	0.22
CG	1.15 %	4.19 %	-3.04	0.73
CT	6.88 %	6.04 %	+0.84	0.14
GA	6.04 %	6.05 %	-0.01	0.0
GC	4.25 %	4.19 %	+0.06	0.01
GG	5.15 %	4.21 %	+0.94	0.22
GT	5.08 %	6.07 %	-0.99	0.16
TA	6.39 %	8.72 %	-2.33	0.27
TC	5.98 %	6.04 %	-0.06	0.01
TG	7.3 %	6.07 %	+1.23	0.2
TT	9.9 %	8.75 %	+1.15	0.13

Figura I.1: *Cálculo de las frecuencias de los 16 dinucleótidos en el cromosoma 21 del ser humano. Los valores esperados y observados suelen coincidir en ± 1 % a excepción del dinucleótido CG.*

$$p(s_1) * \prod_{i=2}^{n} p(s_i|s_{i-1})$$

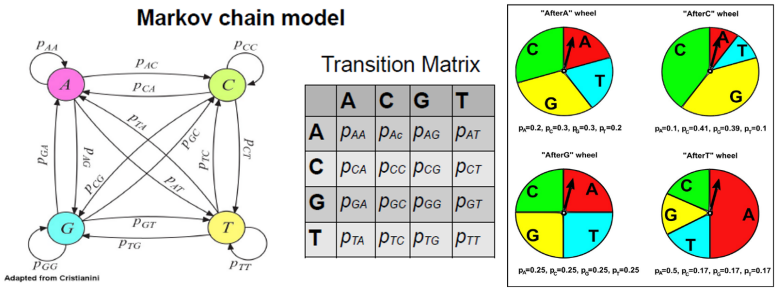

Figura I.2: *Representaciones gráficas de la cadena de Markov. En la matriz de transición, las filas corresponden a los nucleótidos de la posición anterior y las columnas los nucleótidos que les siguen.*

Por ejemplo, para calcular la probabilidad de encontrar la secuencia RCGTG utilizando este modelo, se deben considerar las probabilidades condicionales para cada posible combinación de nucleótidos. La probabilidad se calcula dividiendo la secuencia en dos casos, que luego se suman:

$$\begin{aligned} & 0,25 \times 0,3 \times 0,39 \times 0,25 \times 0,17 (ACGTG) \\ +& 0,25 \times 0,25 \times 0,39 \times 0,25 \times 0,17 (GCGTG) \\ &= 0,001243 + 0,001036 \\ &= 0,002279 \end{aligned}$$

Problema práctico: Un desafío interesante sería escribir un programa que identifique islas CpG en un fragmento del genoma humano. Los dinucleótidos CG

tienden a perderse debido a la metilación de la citosina, que, al desaminarse, se convierte en timina en lugar de regresar a citosina. Sin embargo, en regiones del genoma que no se metilan, como las regiones transcripcionalmente activas, las secuencias CG permanecen intactas, formando las llamadas islas CpG. El objetivo del programa sería localizar el inicio y el final de una de estas islas en una secuencia genómica.

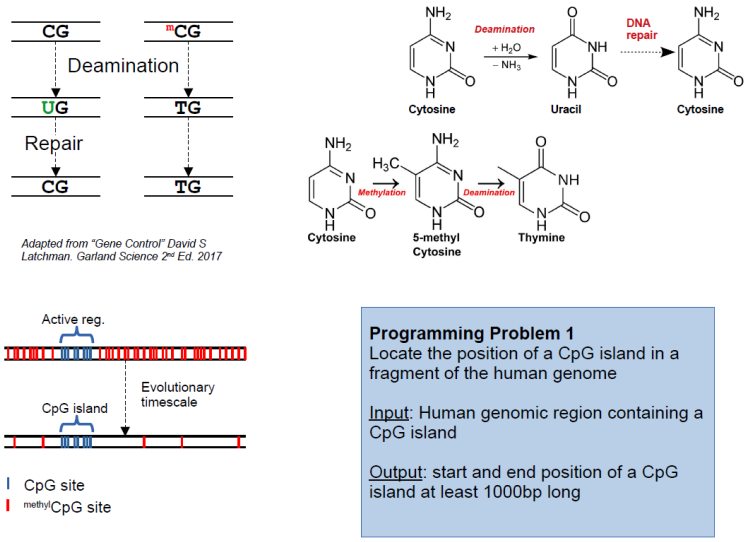

Figura I.3: *Explicación biológica gráfica de las islas CpG.*