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

1 de octubre de 2024 12:48

Universidad Autónoma de Madrid
Bioinformática y Biología Computacional

Código en Github

Índice general

Capítulo I

Modelos estadísticos en el análisis de secuencias

I.1. Secuencias biológicas como cadenas o strings

El ADN, el ARN y las proteínas son responsables del almacenamiento, mantenimiento y ejecución de la información genética, representando así el dogma central de la biología molecular. Estas moléculas están compuestas por miles de átomos dispuestos en complejas estructuras tridimensionales. Y lo que es más importante, la estructura de estas moléculas es clave para su función. Una característica notable común a estas biomoléculas es que, a pesar de su complejidad estructural, son **polímeros lineales de un número limitado de subunidades (monómeros)** y un gran número de pruebas experimentales indican que la secuencia de los monómeros en la estructura lineal de estas moléculas es el principal determinante de sus propiedades, incluidas la estructura y la función. Así pues, estas moléculas pueden conceptualizarse como cadenas de símbolos y este sencillo modelo capta sus propiedades más fundamentales. Sorprendentemente, esta abstracción coincide con la definición formal de una cadena en las herramientas matemáticas y computacionales.

I.1.1. Definición formal de una cadena

En los lenguajes formales, como los utilizados en matemáticas e informática, una cadena se define como una secuencia finita de símbolos de un alfabeto determinado. Sea Σ un conjunto finito no vacío de símbolos (caracteres), llamado alfabeto. Una cadena sobre Σ es cualquier secuencia finita de símbolos de Σ. El número total de símbolos de una cadena s se conoce como longitud de secuencia, o simplemente longitud, y se suele representar como $||s||$. Una palabra suele ser una cadena sobre Σ de longitud definida. El conjunto de todas las cadenas de longitud n sobre Σ, es decir, el conjunto de todas las palabras de tamaño n, se denomina Σ^n. Existen varias operaciones definidas para las cadenas, que también pueden representarse como nodos de un gráfico. En realidad, esto es clave para algunos métodos computacionales utilizados para ensamblar genomas completos a partir de estrategias de secuenciación shotgun.

I.1.2. ADN como cadena

Una molécula de ADN puede idealizarse como una cadena sobre el conjunto {A, C, G, T}, donde cada símbolo representa uno de los cuatro monómeros de nucleótidos del ADN, y una proteína como una cadena sobre el conjunto {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, donde cada símbolo representa cada uno de los 20 residuos de aminoácidos (monómeros) presentes en las proteínas naturales. Si $\Sigma = \{A, C, G, T\}$, entonces Σ^3 representa los codones del código genético.

I.2. Modelos estadísticos del ADN

Consideremos que queremos construir un dispositivo (podría ser un programa informático o un artefacto físico como una ruleta, véase más adelante) que pueda producir una secuencia de ADN (es decir, una cadena sobre el conjunto {A, C, G, T}) que sea una cadena que tenga las mismas propiedades (composición y distribución de nucleótidos) que las moléculas de ADN reales. Para ello podemos utilizar dos modelos: el modelo multinomial y el modelo de cadena de Markov.

I.2.1. Modelo multinomial

El modelo más simple de secuencias de ADN asume que los nucleótidos son independientes e idénticamente distribuidos (iid), es decir, la secuencia ha sido generada por un proceso que produce cualquiera de los cuatro símbolos en cada posición de secuencia i al azar, extrayéndolos independientemente de la misma distribución de probabilidad [1] sobre el alfabeto {A, C, G, T}.

Se puede generar una secuencia de ADN según el modelo multinomial [2] utilizando un dispositivo sencillo como el que se representa en la figura I.1. El modelo de secuencia multinomial es como tener una ruleta que se divide en cuatro partes diferentes etiquetadas como A, T, G y C, donde pA, pT, pG y pC son las fracciones de la ruleta ocupadas por los cortes con estas cuatro etiquetas. Si se hace girar la flecha situada en el centro de la rueda de la ruleta, la probabilidad de que se detenga en la porción con una etiqueta particular (por ejemplo, la porción etiquetada como "A") solo depende de la fracción de la rueda ocupada por esa porción (pA aquí).

En una cadena generada por un modelo multinomial, la probabilidad de observar el símbolo (nucleótido en el caso del ADN y aminoácido en el caso de la proteína) x en la posición i de la secuencia se denota por $p_x, i = p(s(i) = x)$ y no depende de la posición i. Por lo tanto, podemos calcular la probabilidad de observar la cadena s donde n = ||s|| como:

[1] Una distribución de probabilidad es una lista de los posibles resultados con sus correspondientes probabilidades que cumple tres reglas: 1. los resultados deben ser disjuntos; 2. cada probabilidad debe estar comprendida entre 0 y 1; 3. las probabilidades deben sumar 1.

[2] La distribución binomial describe la probabilidad de obtener un número determinado de éxitos en n experimentos independientes. Fundamentalmente, la distribución binomial se aplica sólo cuando el experimento tiene sólo dos resultados posibles. La distribución multinomial es una generalización de la distribución binomial donde cada variable aleatoria puede tomar más de dos valores.

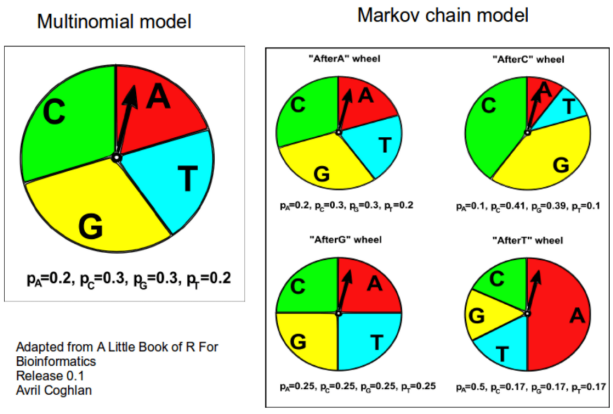

Figura I.1: *Comparación de los modelos de secuencia de ADN multinomial y cadena de Markov.*

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

Ejemplo práctico: 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 pueden parecer 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. 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.2.1.1. 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.2). 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.

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.2: *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.*

I.2.2. Cadena de Markov

El modelo multinomial es una herramienta sencilla e intuitiva que representa con precisión muchas secuencias biológicas de ADN. Sin embargo, se supone que la probabilidad de que aparezca un nucleótido en una posición determinada es independiente de la identidad de los residuos cercanos, lo que no siempre es así. Por ejemplo, si quisiéramos modelar un tramo de ADN que comprende una isla CpG, la probabilidad de observar una G estaría estrictamente condicionada a la identidad del residuo anterior, es decir, la probabilidad de observar una G después de una C sería probablemente más alta que después de cualquier otro residuo de nucleótido. Las cadenas de Markov pueden modelar correlaciones locales entre símbolos en una cadena. Para ello utilizan probabilidades condicionales. Por lo tanto, mientras que en el modelo multinomial se suponía que pG era constante a lo largo de la secuencia, en el modelo de cadena de Markov pG después de C $p(G|C)$ no es necesariamente igual a pG después de A $p(G|A)$. Se puede generar una secuencia de ADN según el modelo de Markov utilizando un dispositivo sencillo como el que se muestra a la derecha

en las figuras I.1 y I.3. En este caso tenemos cuatro ruletas, cada una de las cuales representa las probabilidades de los nucleótidos del ADN. Para generar un residuo en cualquier posición determinada usando este modelo, elegiríamos una de estas cuatro ruedas de ruleta dependiendo del residuo que obtuviéramos en la posición anterior. Se podría representar todas estas probabilidades usando una matriz donde las filas representan el nucleótido encontrado en la posición anterior de la secuencia, mientras que las columnas representan los nucleótidos que podrían encontrarse en la posición actual de la secuencia. En la tabla I.1 se muestra una representación de la ruleta a la derecha de la figura I.1 en forma de matriz.

	To A	To C	To G	To T
From A	0,20	0,30	0,30	0,20
From C	0,10	0,41	0,39	0,10
From G	0,25	0,25	0,25	0,25
From T	0,50	0,17	0,17	0,17

Tabla I.1: *Matriz de transición de cadena de Markov.*

En la jerga de los modelos de Markov, esta matriz se denomina **matriz de transición**. La razón es que una cadena de Markov generadora de secuencia de ADN se puede idealizar como una estructura con cuatro estados diferentes, que representan cada uno de los cuatro nucleótidos, y la secuencia se produce por la transición de un estado a otro. Las transiciones entre estados no son igualmente probables, sino que ocurren con las probabilidades indicadas en los bordes que unen cada estado, que en conjunto son las probabilidades de transición y pueden representarse como una matriz de transición (véase figura I.3). Las entradas en la matriz de transición corresponden a probabilidades condicionales. Por ejemplo, pCG es la probabilidad de G en la posición i dado que hay una C en la posición i-1, es decir $pG = p(s_i = G|s_{i-1} = C)$. Por tanto, la probabilidad de la secuencia s según este modelo podría calcularse como $p(s) = \prod p(s_i|s_{i-1})$. Sin embargo, vale la pena señalar que, para representar una molécula de ADN lineal, también necesitaríamos un conjunto de parámetros que representen las probabilidades del primer nucleótido en la secuencia (dado que no hay uno anterior, podríamos obtener esta probabilidad de la matriz de transición). Si definimos estas probabilidades iniciales como $\pi(A), \pi(C), \pi(G), \pi(T)$, entonces la probabilidad de una secuencia lineal según este modelo se puede calcular como:

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

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:

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

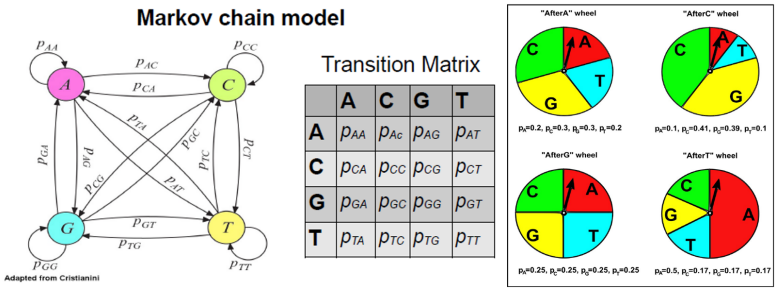

Figura I.3: *Representaciones gráficas de la cadena de Markov. En la matriz de transición, las filas corresponden a los nucleótidos de la posición anterior y las columnas los nucleótidos que les siguen.*

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.

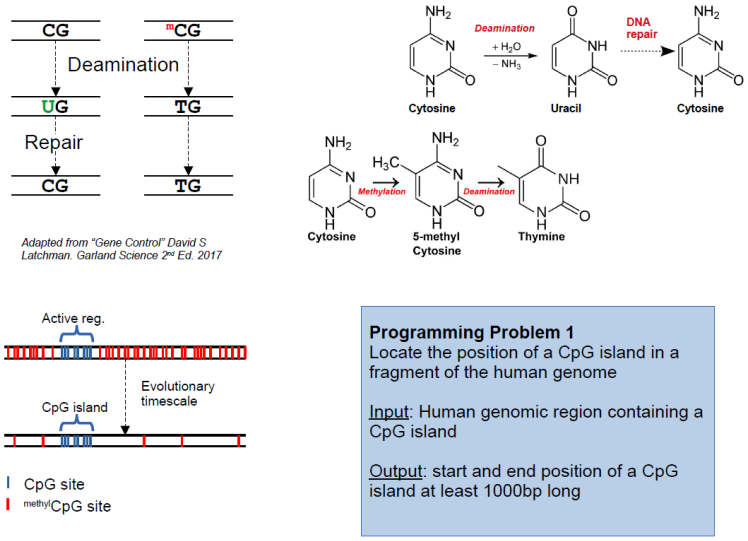

Figura I.4: *Explicación biológica gráfica de las islas CpG.*