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

4 de octubre de 2024 22:39

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.

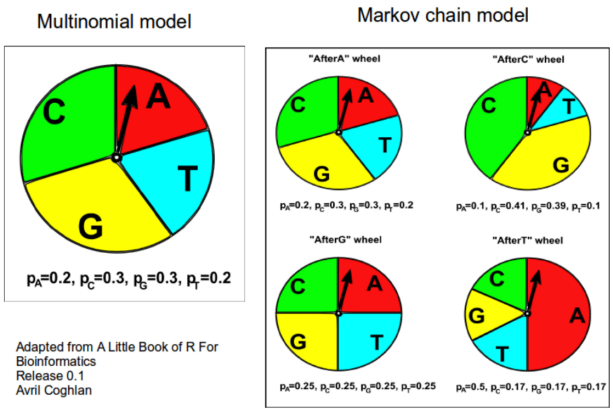


Figura I.1: *Comparación de los modelos de secuencia de ADN multinomial y cadena de Markov.*

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

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:

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

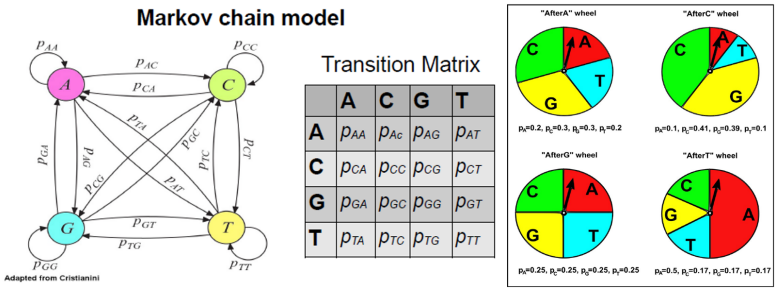


Figura I.3: *Representaciones gráficas de la cadena de Markov. En la matriz de transición, las filas corresponden a los nucleótidos de la posición anterior y las columnas los nucleótidos que les siguen.*

I.2.3. Ejercicios

Ejercicio 1: *Supongamos que el ADN humano puede dividirse en sólo dos tipos de regiones las ricas en C+G y el resto del ADN con una composición de bases no sesgada (no ricas en C+G). Suponiendo el modelo de independencia (la probabilidad de cada nucleótido en una posición dada es independiente de la identidad de los nucleótidos adyacentes) y que la secuencia es homogénea dentro de cada una de estas dos regiones, podemos representarlas mediante un modelo probabilístico multinomial. La región rica en G+C se define por los parámetros: pT=1/8, pC=3/8,pA=1/8 y pG=3/8. El resto del ADN por pT=pC=pA=pG=1/4. ¿Cuál es la probabilidad de observar la secuencia seg=CGACGCGCGCGTCG en una región rica en C+G? ¿Y en la no rica en G+C? Ahora bien, imaginemos que sólo el 1 % (¡me lo acabo de inventar!) del genoma es rico en C+G. Si tomamos un genoma de 14 pb al azar y resulta ser la secuencia CGACGCGCGCGTCG, ¿cuál sería la probabilidad de que proceda de una región rica en C+G?*

- Paso 1: Probabilidad de observar la secuencia en la región rica en C+G

  La probabilidad de observar una secuencia en una región rica en C+G, dada la independencia entre los nucleótidos, es el producto de las probabilidades de cada nucleótido en la secuencia. Las probabilidades en la región rica en C+G son las siguientes:

$$p_T = \frac{1}{8}, \quad p_C = \frac{3}{8}, \quad p_A = \frac{1}{8}, \quad p_G = \frac{3}{8}$$

  Dada la secuencia `CGACGCGCGCGTCG`, la probabilidad de observarla en la región rica en C+G es:

$$P(\text{CGACGCGCGCGTCG} \mid \text{C+G}) = p_C \cdot p_G \cdot p_A \cdot p_C \cdot p_G \cdot p_C \cdot p_G \cdot p_C \cdot p_G \cdot p_C \cdot p_G \cdot p_T \cdot p_C \cdot p_G$$

  Sustituyendo los valores de las probabilidades:

$$P(\text{CGACGCGCGCGTCG} \mid \text{C+G}) =$$

$$\left(\frac{3}{8}\right)\cdot\left(\frac{3}{8}\right)\cdot\left(\frac{1}{8}\right)\cdot\left(\frac{3}{8}\right)\cdot\left(\frac{3}{8}\right)\cdot\left(\frac{3}{8}\right)\cdot\left(\frac{3}{8}\right)\cdot\left(\frac{3}{8}\right)\cdot\left(\frac{3}{8}\right)\cdot\left(\frac{3}{8}\right)\cdot\left(\frac{1}{8}\right)\cdot\left(\frac{3}{8}\right)\cdot\left(\frac{3}{8}\right)$$

$$= 1,2 \cdot 10^{-7}$$

- Paso 2: Probabilidad de observar la secuencia en la región no rica en C+G

  En la región no rica en C+G, las probabilidades de cada nucleótido son iguales:

  $$p_T = p_C = p_A = p_G = \frac{1}{4}$$

  Por lo tanto, la probabilidad de observar la secuencia `CGACGCGCGCGTCG` es:

  $$P(\text{CGACGCGCGCGTCG} \mid \text{no C+G}) = \left(\frac{1}{4}\right)^{14} = 3,7 \cdot 10^{-9}$$

- Paso 3: Probabilidad de que la secuencia provenga de una región rica en C+G

  Utilizamos el teorema de Bayes para calcular la probabilidad de que la secuencia provenga de una región rica en C+G. La fórmula de Bayes es:

  $$P(\text{C+G} \mid \text{secuencia}) = \frac{P(\text{secuencia} \mid \text{C+G}) \cdot P(\text{C+G})}{P(\text{secuencia})}$$

  Donde:

  - $P(\text{secuencia} \mid \text{C+G})$ es la probabilidad de observar la secuencia en una región rica en C+G (calculada en el Paso 1).
  - $P(\text{C+G}) = 0,01$ es la proporción del genoma que es rico en C+G.
  - $P(\text{secuencia})$ es la probabilidad total de observar la secuencia, que se calcula como:

  $$P(\text{secuencia}) = P(\text{secuencia} \mid \text{C+G}) \cdot P(\text{C+G}) + P(\text{secuencia} \mid \text{no C+G}) \cdot P(\text{no C+G})$$

  Donde $P(\text{no C+G}) = 1 - P(\text{C+G}) = 0,99$

  Sustituyendo todos los valores, podemos obtener la probabilidad de que la secuencia provenga de una región rica en C+G.

  $$\frac{1,2 \cdot 10^{-7} \cdot 0,01}{1,2 \cdot 10^{-7} \cdot 0,01 + 3,7 \cdot 10^{-9} \cdot 0,99} = 0,25$$

Ejercicio 2: *Ha secuenciado un fragmento de la cadena + de un nuevo organismo. Nosotros suponemos que es un fragmento representativo y que la composición es homogénea en todo el genoma. Las frecuencias absolutas de bases en este fragmento de secuencia se indican en la tabla siguiente. Estima los siguientes parámetros de un modelo de cadena de Markov para esta secuencia. ¿Qué sería la probabilidad de*

transición de T a A (PTA) y la probabilidad de y la probabilidad de transición de A a A (PAA)? ¿Cuál sería la probabilidad PTA para el modelo de cadena de Markov de la cadena - de este dsADN? ¿Y la probabilidad de transición de transición PAA de la cadena -? Teniendo en cuenta las siguientes probabilidades condicionales:

	To A	To C	To G	To T
From A	15	23	25	11
From C	9	38	35	8
From G	26	21	18	24
From T	25	8	10	3

$$P_{TA+} = \frac{25}{25+8+10+3} = 0,54$$

$$P_{AA+} = \frac{15}{15+23+25+11} = 0,20$$

En cuanto a la probabilidad de la cadena negativa, hay que tener en cuenta que las frecuencias están dadas en la cadena positiva. Por tanto, cuando se tiene en cuenta el cambio del segundo nucleótido de la pareja en la cadena negativa, el cambio en la cadena positiva se produce en el primero.

5' ————–TA————3'
3' ————–AT————5'

$$P_{TA-} = \frac{25}{25+15+26+9} = 0,33$$

$$P_{AA-} = \frac{3}{3+24+8+11} = 0,065$$

I.2.4. Problema práctico: islas CpG

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. La isla CpG tiene una longitud de 1.000 bases, mientras que la región genómica tendrá aproximadamente unos 40.000 nucleótidos.

Las islas CpG tienen una alta densidad de dinucleótidos de CG. Por tanto, hay que buscar una región genómica que tenga una alta densidad y que permita identificar la isla. Para ello, se debe emplear un sliding window, es decir, una ventana de una cierta cantidad de nucleótidos para calcular su frecuencia de CG. Como la isla CpG va a tener un tamaño de 1000, el tamaño razonable de ventana sería de 1000, y esta ventana se irá desplazando de nucleótido en nucleótido. En un gráfico que muestre la

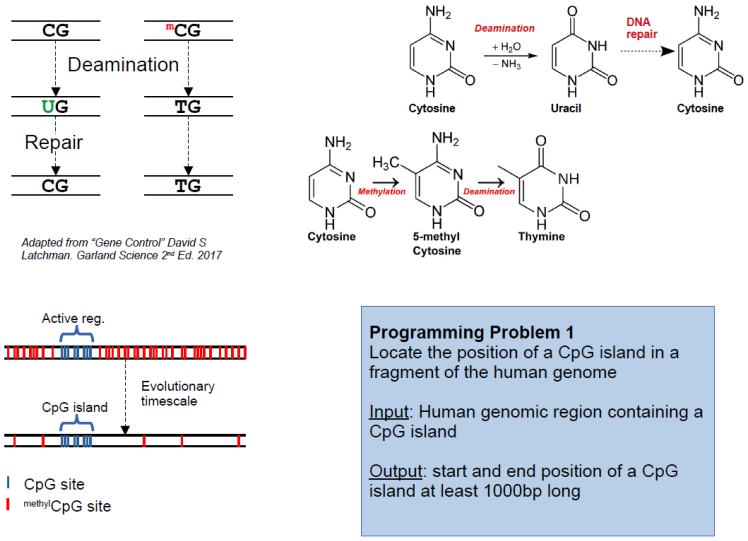


Figura I.4: *Explicación biológica gráfica de las islas CpG.*

densidad de CG, se observaría una frecuencia muy superior (un pico alto) donde se encuentre la isla. Como la gráfica real es algo ruidosa, hay que establecer un threshold para poder obtener la posición concreta de la isla. Se puede utilizar la frecuencia total de CG en la secuencia (contabilizar todas las apariciones de CG y dividir por la longitud para obtener la media), pero hay que tener en cuenta el margen de error. Se puede calcular el porcentaje de CG en todas las ventanas, calcular la media y la desviación estándar para poder tener la dispersión esperada de una ventana concreta. Una vez con eso, se puede dibujar la distribución de los porcentajes de CpG para poder establecer la frecuencia de fondo de los dinucleótidos y separarla de la frecuencia de las islas CpG. En caso de una distribución normal, se pueden establecer criterios arbitrarios como los criterios estadísticos del 5 % superior (one value t-test). Esto resulta en una distribución empírica, pero se puede utilizar una distribución binomial para obtener el mismo resultado más formalmente correcto. También se puede aproximar a una distribución de Poisson para cada ventana. La forma más correcta sería mediante los modelos ocultos de Markov, teniendo como etiquetas que una posición pertenezca o no a una isla CpG. Esto se verá más adelante en la asignatura.

Capítulo II

Alineamiento de secuencias por pares

El alineamiento de secuencias es la herramienta más fundamental de la bioinformática. Permite identificar secuencias relacionadas con una secuencia dada. Como veremos, el parentesco suele implicar que las secuencias pueden tener funciones comunes y esa es una de las principales aplicaciones del alineamiento de secuencias, inferir la función de una secuencia biológica.

II.1. Alineamiento de secuencias

La alineación de secuencias es el procedimiento de ordenar dos (alineación por pares) o varias (alineación de secuencias múltiples, MSA) secuencias intentando colocar el mayor número posible de residuos idénticos o similares en el mismo registro vertical (misma columna). Los residuos no idénticos pueden colocarse en la misma columna como una falta de coincidencia o frente a un hueco en la otra secuencia. El objetivo de la alineación es maximizar el número de coincidencias (residuos idénticos o similares en la misma columna) y minimizar el número de desajustes y huecos.

¿Por qué alineamos las secuencias de este modo? En la alineación de secuencias, el supuesto subyacente es que las **secuencias que se alinean proceden de un ancestro común**. Sin embargo, como consecuencia de las mutaciones acumuladas durante la evolución, las secuencias no serán idénticas. Así pues, el reto consiste en colocar los residuos que derivan de la **misma posición ancestral** en la misma columna del alineamiento. Sin embargo, sin información sobre la secuencia ancestral y su evolución, lo mejor que podemos hacer es maximizar el número de coincidencias y minimizar el número de discordancias. En las secuencias de proteínas, las sustituciones se producen cuando una mutación (mutación sin sentido o missense) en la secuencia ancestral hace que el codón de un aminoácido se cambie por el de otro. El resultado sería la alineación de dos aminoácidos no idénticos, es decir, un desajuste. Las inserciones y deleciones (normalmente abreviadas como INDEL) se producen cuando se añaden o eliminan residuos de la secuencia ancestral. Las inserciones o deleciones (incluso las de un solo carácter) se representan como huecos en el alineamiento. El número de mutaciones aumentará a medida que las dos secuencias diverjan de su

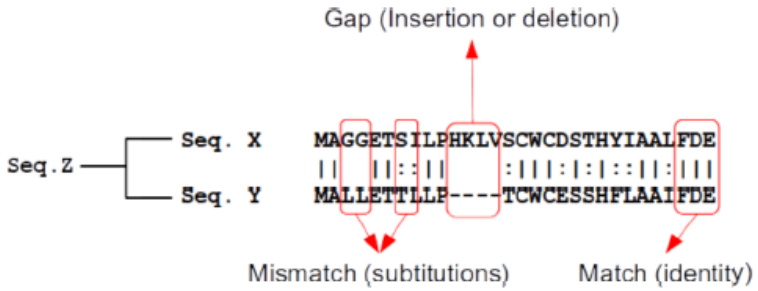


Figura II.1: ***Alineamiento por pares.*** *La alineación por pares modela la evolución de las dos secuencias a partir de un ancestro común. En un intento de colocar los residuos derivados de la misma posición ancestral en el registro vertical, el proceso de alienación maximiza las coincidencias y minimiza las diferencias debidas a mutación de la secuencia ancestral (desajustes y lagunas). En la representación, dos residuos iguales se muestran conectados por una línea vertical. En caso de discordancia, si a nivel biológico los residuos tienen una función similar, se denota con dos puntos, mientras que si la función es diferente, se deja en blanco, al igual que en el caso de inserciones y deleciones.*

ancestro común. Así, en general, el número de coincidencias disminuye con la distancia evolutiva. En general, se puede inferir que los residuos idénticos entre secuencias probablemente también estén presentes en el ancestro. En el caso de las sustituciones y deleciones/inserciones, no se puede determinar si el ancestro era de una forma u otra, es decir, si realmente se trata de un fragmento que se ha perdido (una deleción) o un fragmento que en algún momento se insertó.

II.2. Comparación de alineamientos

A la hora de alinear dos secuencias, se pueden producir todos los alineamientos posibles y escoger aquel que tenga el mayor número de coincidencias y el menor número de discordancias o gaps. Un alineamiento óptimo sería aquel que incluya en la misma columna residuos que derivan del mismo residuo original (es decir, alinear residuos ortólogos). Normalmente, no se conoce la secuencia del ancestro o la historia evolutiva, por lo que se intenta inferir al alinear todos los residuos idénticos o similares posibles. Para establecer el grado de similitud, se puede tener en cuenta las propiedades fisicoquímicas (hidrofobicidad, carga neta a pH fisiológico, flexibilidad de la cadena lateral) y la estructura (tamaño, presencia de anillos aromáticos). Además, si se permiten los gaps en el alineamiento, la cantidad de posibles alineamientos aumenta de forma astronómica. Por tanto, para encontrar el mejor alineamiento, se necesita:

- Una métrica cuantitativa que representa la similitud entre residuos.
- Un método de puntuación que produce un valor que resume lo buena que es la alineación teniendo en cuenta todas las posiciones y huecos.
- Un procedimiento capaz de producir todos los alineamientos posibles y puntuarlos eficazmente (y no evaluando todos los posibles alineamientos).

II.2.1. Matrices de sustitución

Como ya se ha dicho, el alineamiento consiste en reunir residuos idénticos o similares. Identificar los residuos idénticos es sencillo. Sin embargo, ¿qué entendemos por residuos similares? En el caso de los ácidos nucleicos, la función de un determinado nucleótido (su patrón de emparejamiento de bases) no suele poder sustituirse por ninguno de los demás nucleótidos. Por lo tanto, durante la alineación de secuencias de nucleótidos (normalmente) sólo nos preocupamos por las identidades [1]. Cualquier otro emparejamiento es un desajuste igualmente perjudicial. Sin embargo, en el caso de las secuencias de aminoácidos, ciertas sustituciones de aminoácidos tienen poco impacto, mientras que otras pueden abolir por completo la función/estructura de la proteína. Así, en el curso de la evolución, los residuos importantes para la función de la molécula tienden a permanecer inalterados o a ser sustituidos por un residuo similar, manteniendo así la estructura y/o la función. Por estas razones, algunas sustituciones particulares se encuentran comúnmente en proteínas relacionadas de diferentes especies. Así, para los alineamientos de proteínas asignamos una puntuación a cada par o aminoácidos que representa la probabilidad de observar la sustitución de uno por otro. Una tabla que contiene las puntuaciones de todos los posibles pares de residuos se denomina **matriz de sustitución**. Las puntuaciones de cada celda de una matriz de sustitución reflejan la probabilidad de que los dos residuos estén alineados porque son verdaderos homólogos en comparación con la probabilidad de que estén alineados en la misma posición por azar:

$$\frac{p(alineado|homlogo)}{p(alineado|aleatorio)}$$

Estas probabilidades pueden derivarse de **principios teóricos**, por ejemplo el número de mutaciones necesarias para convertir el codón de un aminoácido en el de otro o la similitud fisicoquímica entre los dos residuos comparados. Sin embargo, las puntuaciones de las matrices de sustitución más populares se han derivado de la **observación empírica** de las tasas de sustitución en alineaciones de proteínas homólogas. Dos matrices de sustitución populares derivadas empíricamente son PAM y BLOSUM.

II.2.1.1. Matrices de sustitución PAM

Para construir una matriz de sustitución a partir de la observación de los reemplazos ocurridos durante la evolución, sólo necesitamos alinear las proteínas y contar el número de cambios de cada tipo. Sin embargo, generar una matriz de sustituciones a partir de alineamientos de proteínas es un problema circular: se necesita el alineamiento para contar el número de sustituciones observadas pero, para generar un buen alineamiento, se necesitan las puntuaciones de cada par de residuos. Para sortear este problema, Margaret Dayhoff (la primera bioinformática en la historia) y su equipo idearon una estrategia inteligente. Utilizaron secuencias muy similares de homólogos bien conocidos

[1] De hecho, dado que las transiciones (es decir, las sustituciones entre las purinas A y G o entre las pirimidinas C y T) son más frecuentes que las transversiones (sustituciones entre purina y pirimidina o viceversa), existen algunos esquemas de puntuación específicos para la alineación de residuos de nucleótidos no idénticos.

para poder generar alineaciones fácilmente y con gran confianza incluso en ausencia de matrices de sustitución. A continuación, a partir de estos alineamientos generaron árboles filogenéticos que les permitieron inferir la secuencia ancestral de cada par de proteínas alineadas. Por último, a partir de estos árboles calcularon las probabilidades de que cualquier aminoácido mutara en cualquier otro. Así, Dayhoff y sus colegas construyeron árboles filogenéticos a partir de familias de proteínas estrechamente relacionadas y calcularon la probabilidad de que dos residuos alineados derivaran del mismo residuo ancestral (véase la figura II.2). En este proceso definieron una **mutación puntual aceptada** (abreviada como PAM) como la sustitución de un residuo original por otro que ha sido aceptado por la selección natural (de lo contrario no estaríamos observando estas secuencias). Como ya se ha mencionado, el conjunto original de proteínas que utilizaron para derivar la matriz de sustitución era muy similar y tenía 1 mutación puntual aceptada por cada 100 residuos de aminoácidos. En consecuencia, esta matriz se denomina PAM1.

Sin embargo, por definición, esta matriz es óptima para puntuar secuencias estrechamente relacionadas, pero no secuencias distantes (está sesgada a secuencias muy próximas evolutivamente). Para generar matrices que reflejaran relaciones más distantes, Dayhoff y sus colegas extrapolaron sus datos observados multiplicando PAM1 por sí mismo varias veces. Cuanto mayor era el número de veces que se multiplicaba el PAM1 por sí mismo, mayor era la distancia que representaba. Por ejemplo, PAM250, derivado de multiplicar PAM1 por sí mismo 250 veces, se utiliza habitualmente para comparar proteínas distantemente relacionadas.

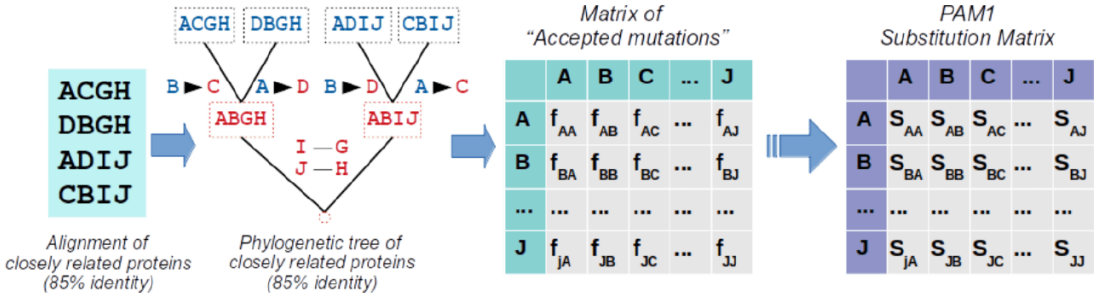


Figura II.2: ***Generación de la matriz de sustitución PAM1.*** *A partir de alineaciones de secuencias estrechamente relacionadas (> 85 % de identidad), Margaret Dayhoff y sus colegas derivaron el árbol filogénico que representaba la evolución de la familia que requería el menor número de mutaciones. A partir de estos árboles contaron el número de veces que cada residuo fue sustituido por cualquier otro y registraron los valores en la matriz de mutaciones aceptadas. Por último, a partir de los datos de esta matriz generaron la matriz de sustitución PAM1 que representa la relación entre la probabilidad de la sustitución observada en el modelo evolutivo (suponiendo homología) y la probabilidad en el modelo aleatorio.*

II.2.1.2. Matrices de sustitución BLOSUM

Más recientemente, el matrimonio Henikoff utilizó una familia de proteínas más alejada para poder inferir la frecuencia de sustitución en una matriz BLOSUM.

Para evitar la incertidumbre en los alineamientos, Dayhoff utilizó un conjunto de secuencias extremadamente relacionadas para derivar la PAM1. Sin embargo, las matrices PAM para proteínas más distantes se extrapolaron a partir de PAM1 en lugar de derivarse de la observación directa de los alineamientos reales. La acumulación de secuencias de proteínas en bases de datos a lo largo de los años permitió a Henikoff y Henikoff desarrollar un nuevo conjunto de matrices de sustitución a principios de los 90. Estas matrices, denominadas BLOCKS [2] amino acid SUbstitution Matrices (BLOSUM), se generaron al registrar cada posible sustitución de aminoácidos observada en los alineamientos de bloques. Utilizando alineamientos de proteínas que mostraban diferentes porcentajes de identidad, derivaron matrices BLOSUM que representaban la tasa de sustitución observada para diferentes grados de divergencia (figura II.3). Para ello, eliminan del bloque todas las secuencias que son idénticas en más de un x % de posiciones, dejando una única secuencia representativa (por ejemplo, en BLOSUM62 se eliminaron las secuencias que compartían un 62 % de identidad o más).

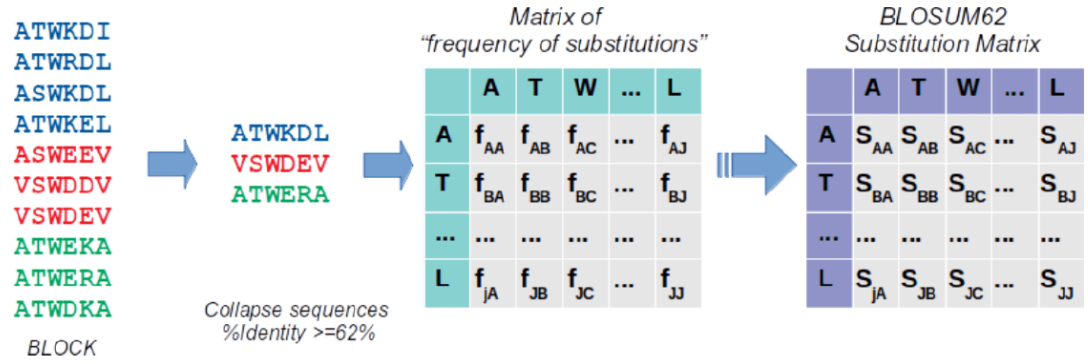


Figura II.3: ***Generación de la matriz de sustitución BLOSUM.*** *Partiendo de alineaciones sin colapsar de familias de proteínas (BLOCKS), Henikoff y Henikoff derivaron alineaciones que representaban diferentes distancias evolutivas colapsando todas las secuencias del bloque que compartían un umbral, C, de identidad. En la figura C = 62 %, todas las secuencias que comparten un porcentaje de identidad igual o superior al 62 % se muestran en el mismo color de fuente (primera columna), y luego se colapsan en un consenso que representa el clúster (segunda columna). A partir de estos alineamientos, contaron el número de veces que cada residuo i fue sustituido por cualquier otro y registraron los valores en la matriz de frecuencia de sustituciones. Por último, a partir de los datos de esta matriz generaron la matriz de sustituciones BLOSUM62 que representa la relación entre la probabilidad de la sustitución observada en el modelo evolutivo (suponiendo homología) y la probabilidad en el modelo aleatorio.*

Nótese que existen algunas diferencias importantes entre las matrices PAM y BLOSUM. En primer lugar, todas las matrices BLOSUM se derivan de la observación directa de alineamientos, mientras que sólo PAM se deriva de datos y el resto son extrapolaciones. En segundo lugar, mientras que PAM1 se generó a partir de alineaciones de secuencias estrechamente relacionadas (85 % de identidad), las matrices BLOSUM derivan de alineaciones que (pueden) incluir secuencias con un bajo porcentaje de identidad. Por último, para la construcción de PAM se infirieron

[2] un BLOCK se define como una región no superpuesta en el alineamiento de secuencias múltiples de menos de sesenta residuos de aminoácidos

sustituciones a partir de árboles filogenéticos derivados de los alineamientos. En BLOSUM no se construyó ningún árbol filogenético y las sustituciones se contaron a partir de la observación directa de los residuos alineados. Sin embargo, no se trata de sustituciones reales porque las secuencias alineadas evolucionaron a partir de un ancestro común y entre sí.

II.2.1.3. Construcción de matrices de sustitución

En las secciones anteriores vimos dos estrategias diferentes para determinar la frecuencia de cambios a partir de la observación empírica de alineamientos de proteínas homólogas. Dejando a un lado los detalles, ambos métodos producen una **matriz de frecuencia de mutación** [3], donde las entradas $q_{a,b}$, representan la **probabilidad observada** de encontrar los residuos a y b **alineados en proteínas homólogas**. En otras palabras, $q_{a,b}$, corresponde al término $p(alineado|homlogo)$. Ahora, para obtener el valor de la entrada para los residuos a y b en la matriz de sustitución correspondiente, necesitamos calcular el término $p(alineado|aleatorio)$, que sería la **probabilidad esperada**. En el modelo aleatorio suponemos que las dos proteínas alineadas no están relacionadas y no existen restricciones estructurales o funcionales que puedan causar correlación entre los residuos en una posición dada. Así, en este modelo la probabilidad de encontrar los residuos a y b alineados sólo depende de su frecuencia en las proteínas. En el modelo aleatorio no existe correlación alguna entre los residuos alineados en una posición dada, por lo que la probabilidad de observar a en una secuencia y b en la otra son independientes de modo que:

$$p(alineado|aleatorio) = p(a \cap b) = p_a p_b$$

donde p_a, y p_b, son las frecuencias de a y b respectivamente. La probabilidad de observar a y b alineados en estos dos modelos puede compararse tomando el cociente de las probabilidades, denominado **odds ratio**: $q_{a,b}/(p_a p_b)$. Cuando la probabilidad en el modelo evolutivo es mayor que en el modelo aleatorio, el odds-ratio toma cualquier valor entre 1 e infinito. Sin embargo, cuando la probabilidad en el modelo aleatorio es mayor, la odds-ratio está entre 0 y 1. Para evitar esta asimetría, se suele tomar el logaritmo de la odds-ratio para obtener la **log-odds ratio**. Como veremos más adelante, tomar el logaritmo del odds-ratio también facilita el cálculo de la puntuación total de la alineación. Por lo tanto, la entrada en la matriz de sustitución correspondiente a a y b se calcula como:

$$s_{a,b} = log\frac{q_{a,b}}{p_a p_b} = log\frac{p(cambio|modeloevolutivo)}{p(cambio|aleatorio)} = log\frac{p(observado)}{p(esperado)}$$

La figura II.4 muestra las matrices de sustitución PAM250 y BLOSUM62. Dado que la puntuación del alineamiento a sobre b es la misma de b sobre a, estas matrices son simétricas. Por este motivo, normalmente sólo se representa la mitad de la matriz. Los números positivos significan que se han observado más veces el cambio de residuos que lo que cabría esperar por azar, por lo que debe haber alguna presión positiva para que

[3] la suma de todas las entradas de la matriz da 1

se mantenga. En el caso de los números negativos, se debe a una selección negativa. Cuando es 0, el ratio es 1 y por tanto la frecuencia es la observada por azar, no hay ninguna presión.

Un ejemplo: El triptófano tiene una frecuencia de mutación observada muy pequeña, pero en la tabla BLOSUM, su número es el más alto. Esto significa que es un aminoácido muy importante que no se puede cambiar por ningún otro. Así, la tabla de frecuencias per se no refleja el parecido entre residuos, ya que hay que tener en cuenta la frecuencia. Sin embargo, la tabla BLOSUM sí refleja el parecido entre los residuos.

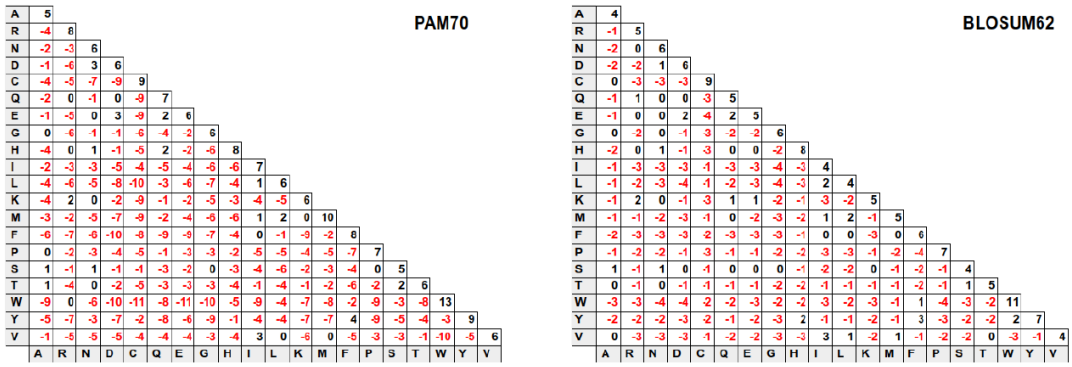


Figura II.4: ***Matrices de sustitución de aminoácidos.*** *La figura muestra las matrices de sustitución PAM250 (izquierda) y BLOSUM62 (derecha). Los valores negativos (en rojo) indican las sustituciones que tienen más probabilidades de observarse en el modelo aleatorio que en el evolutivo.*

Ejemplo de cálculo de matriz puntuación y odds ratio: Cambio D-L
La matriz de frecuencia de mutación observada indica que el cambio D-L se ha observado 15 veces en 10000, es decir, 15/10000. La frecuencia de los bloques es 0,054 para D y 0,099 para L. Esto representa la frecuencia esperada. La puntuación se calcularía siguiendo la fórmula:

$$s = 2 \cdot \log_2(oddsratio) = 2 \cdot \log_2(\frac{observado}{esperado})$$

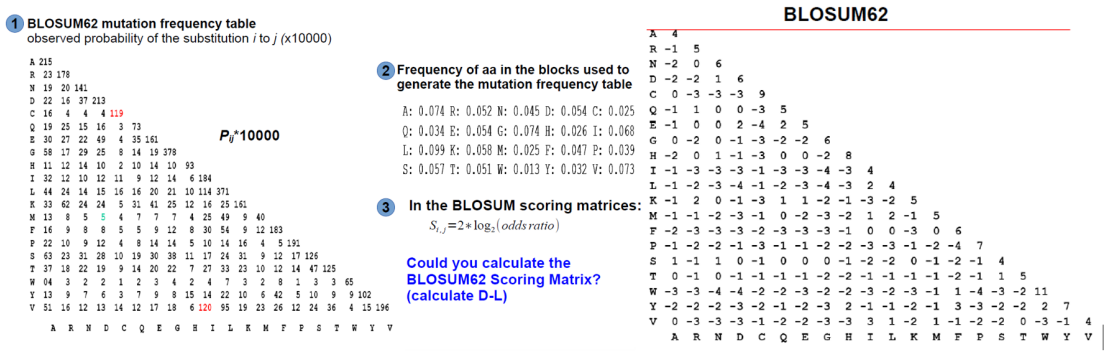
Sustituyendo los valores:

$$s = 2 \cdot \log_2(\frac{15/10000}{0,054 \cdot 0,099}) = -3,66 \approx -4$$

Cuando hay números decimales, se redondea al siguiente número entero. Al comprobar el valor en la matriz BLOSUM, el resultado efectivamente es -4.

II.2.2. Alineamientos de puntuación (scoring alignments)

Las matrices de sustitución ofrecen un método para puntuar posiciones individuales. Sin embargo, para comparar diferentes alineaciones, necesitamos un único valor que represente la puntuación combinada de todas las posiciones. Para calcular dicha



puntuación, suponemos que cada posición del alineamiento es independiente de las demás [4] y calculamos la puntuación del alineamiento S como la suma de las puntuaciones individuales de cada una de las n posiciones, siendo s la entrada de la matriz de sustitución para los residuos a y b en la posición i.

$$S = \sum_{i=1}^{n} (s_{a,b})_i$$

En otras palabras, se pueden sumar los valores de las matrices BLOSUM de cada posición al haber utilizado el logaritmo. Ahora, esta función de puntuación sólo funciona para coincidencias y discordancias pero no tiene en cuenta los INDELs. Para representar los INDEL, un residuo o una serie de residuos en una secuencia de la alineación se empareja con guiones («-») en la otra secuencia. Durante la puntuación, la presencia de un hueco en el alineamiento da lugar a una penalización por hueco que se resta de la puntuación total. Hay dos razones para penalizar los huecos. En primer lugar, un hueco implica una diferencia entre las secuencias comparadas y, por tanto, reduce nuestra certeza sobre su origen común. Los huecos corresponden a eventos de inserción/deleción que ocurrieron durante la evolución desde el ancestro común en uno de los linajes. Por lo tanto, en general, cuanto mayor sea el número de huecos, mayor será la distancia evolutiva entre las secuencias. La segunda razón es que, introduciendo un número ridículo de huecos, podríamos aumentar artificialmente el número de coincidencias y, como consecuencia, aumentar la puntuación del alineamiento, aunque el alineamiento resultante no tendría sentido desde el punto de vista biológico. Así, las penalizaciones por huecos actúan limitando la introducción de huecos. Por lo general, el usuario establece la penalización por hueco a partir de un conjunto de valores predefinidos [5] que se han determinado empíricamente a partir de la observación de su efecto en los alineamientos.

[4] Nótese que esto es probablemente una simplificación excesiva porque en las proteínas reales a menudo existe una correlación entre residuos adyacentes. Por ejemplo, en una hélice anfipática los residuos polares e hidrófobos se distribuyen en caras opuestas. Por lo tanto, habrá cierta correlación entre los residuos en ciertas posiciones.

[5] En algunos programas, la penalización por hueco varía en función del tipo de residuo con el que se alinea el hueco. La razón es que algunos residuos tienden a estar fuertemente conservados debido a su impacto en la estructura/función. Por lo tanto, es más probable que la supresión de esos residuos altere la estructura/función y, por lo tanto, sufra selección negativa.