

Comparación de secuencias

Enrique Blanco García

PID_00165221

Índice

Introducción.....	5
Objetivos.....	7
1. Alfabetos, secuencias y alineamientos.....	9
2. Interpretación biológica de los alineamientos.....	15
3. Alineamientos globales o locales.....	19
4. Alineamientos simples o múltiples.....	22
5. Matrices de puntos.....	25
6. Alineamientos óptimos de secuencias.....	29
7. Alineamientos progresivos de secuencias.....	39
8. Identificación de motivos conservados.....	47
9. Búsquedas masivas en bases de datos.....	53
10. Alineamientos de genomas.....	62
Resumen.....	63
Actividades.....	65
Ejercicios de autoevaluación.....	65
Solucionario.....	67
Bibliografía.....	69

Introducción

En el último medio siglo hemos experimentado una auténtica revolución del conocimiento en múltiples disciplinas científicas. A nivel biológico, la actual tecnología de secuenciación de proteínas y ácidos nucleicos ha evolucionado espectacularmente, alcanzando cotas que parecían fuera de nuestro alcance no hace tanto tiempo. Actualmente, y gracias a la secuenciación masiva, se pueden llegar a producir diariamente varios *terabytes* de información a un precio asequible. Toda esta información se pone públicamente a disposición de la comunidad científica desde múltiples centros de investigación repartidos por el mundo. Naturalmente, el análisis comparativo de estos resultados permite extraer nuevo conocimiento que resulta muy valioso para complementar la información obtenida en los laboratorios.

A finales de los años sesenta del siglo pasado, varios pioneros en el campo de la biología molecular comenzaron a recopilar las primeras secuencias de proteínas. La construcción de árboles filogenéticos a partir de la comparación de estas secuencias demostró ser una importante herramienta para establecer relaciones evolutivas entre especies. Entonces surgió la necesidad de emplear métodos de trabajo más robustos que permitieran realizar sistemáticamente estos contrastes de parecido. Fue otro grupo de pioneros con sólidos conocimientos matemáticos quienes diseñaron los primeros algoritmos para calcular alineamientos óptimos de secuencias. Esta técnica es clave para el análisis bioinformático, dado que es fundamental poseer un método de comparación fiable para garantizar el mejor resultado posible independientemente de las secuencias utilizadas en cada caso.

El alineamiento de dos o más secuencias debe realizarse dentro de un estricto marco de trabajo que establece unos criterios de puntuación estrictos, lo que permite que en la mayoría de situaciones, siguiendo estas restricciones, averiguemos cuál es el alineamiento óptimo en un tiempo razonable. En otros casos que involucran comparaciones más complejas, es imprescindible realizar ajustes sobre los algoritmos básicos. Estas modificaciones generalmente consisten en aplicar criterios heurísticos basados en la experiencia para reducir el número de pasos necesarios para alinear las secuencias.

En este tipo de comparaciones es fundamental construir alineamientos sensatos desde el punto de vista biológico. Por regla general, encontraremos suficiente consenso entre las soluciones óptimas tanto en el plano algorítmico como en el biológico. Sin embargo, también es cierto que en algunas ocasiones la propuesta biológicamente más factible no concuerda necesariamente con la mejor solución en términos matemáticos. En este sentido, veremos en este módulo que la elección de la estrategia de alineamiento más apropiada será crítica para extraer las conclusiones más acertadas en cada contexto biológi-

co. En cualquier caso, debemos ser conscientes de la enorme importancia de interpretar correctamente los alineamientos resultantes. Presentaremos aquí, por tanto, un amplio abanico de técnicas de comparación de secuencias junto con un elenco de nociones básicas sobre su análisis bioinformático.

Objetivos

Mediante la adquisición de los contenidos establecidos para este módulo, el estudiante deberá poseer manejo suficiente en los siguientes aspectos relacionados con la comparación de secuencias biológicas:

1. Saber definir formalmente qué es un alineamiento de dos o más secuencias.
2. Aprender a interpretar biológicamente los alineamientos de secuencias.
3. Comprender los algoritmos fundamentales de comparación de secuencias.
4. Utilizar la clase de alineamientos más apropiada para cada problema biológico.
5. Manejar con soltura las búsquedas de secuencias en grandes bases de datos.
6. Explotar estos algoritmos básicos para comparar otros tipos de secuencias.

1. Alfabetos, secuencias y alineamientos

Para trabajar en nuestro ordenador con genes y proteínas es imprescindible previamente establecer un código que nos permita modelar formalmente el contenido de estas biomoléculas. Tanto los ácidos nucleicos como los péptidos son macromoléculas compuestas por pequeñas unidades elementales químicamente enlazadas (por ejemplo, nucleótidos en el caso del ADN/ARN, aminoácidos en las proteínas). Por tanto, antes de codificar estas estructuras biológicas debemos definir el alfabeto apropiado de símbolos para cada tipo de molécula.

Un **alfabeto** es un conjunto finito no vacío de símbolos. Un **lenguaje** es el conjunto de palabras construidas a partir de la combinación de los símbolos de un alfabeto.

Emplearemos un alfabeto que nos permita construir palabras con nucleótidos para almacenar secuencias de ADN en un entorno computacional. Asignaremos una letra distinta a cada nucleótido en función de las cuatro bases nitrogenadas existentes: adenina, citosina, guanina y timina (cuando analicemos moléculas de ARN deberá sustituirse timina por uracilo). Del mismo modo, para codificar proteínas construiremos un alfabeto con tantos símbolos como aminoácidos conocidos hay.

Figura 1. Alfabetos elementales en bioinformática.

$$\Sigma_{ADN} = \{A, C, G, T\}$$

$$\Sigma_{ARN} = \{A, C, G, U\}$$

$$\Sigma_{PRO} = \{A, C, D, E, F, G, H, I, L, K, M, N, P, Q, R, S, T, V, W, Y\}$$

Mediante la combinación apropiada de nucleótidos o aminoácidos codificaremos conceptualmente la estructura de estas moléculas con largas cadenas de símbolos que podrán ser almacenadas para su posterior análisis. El orden en que se dispongan estas unidades otorgará un significado biológico concreto a cada secuencia de caracteres, conformando un mensaje interpretable dentro de un determinado contexto celular. Solamente un número limitado de estas ordenaciones codifica mensajes funcionales (por ejemplo, genes y proteínas).

Ved también

La estructura química de estas moléculas se explica con detalle en la asignatura *Fundamentos de biología molecular*.

Lectura complementaria

J. E. Hopcroft; R. Motwani; J. D. Ullman (2007). *Introduction to automata theory, languages and computation*. Eaglewood Cliffs, NJ: Prentice Hall. ISBN: 0321462254.

Figura 2. Secuencia completa del genoma del virus VIH-1.

```
>gi|292658940|gb|GQ372990.1| HIV-1 isolate ES X2556-3 from Spain, complete genome
TTTCTTCAAGTTAGTGTGTGCCCGCTGTGTGTGACTCTGGTAAGTACAGAGATCCCTCAGACCCCTTTAGTCAGTGTGGAAAATCTCTAGC
AGTGGCGCCGCAAGCAGGAGCTTTGAAGCCGAAAGAGAAACCGGAGAAAGTCTCTCGACGCAGGACTCGGCTTCTGAAGCGCCGACGGCAG
AGGCGAGGGGGGGCGGACTGGTGAGTAGCGCAATTTTGTAGCTAGCGAGGCTAGAAAGGAGAGAGATGGTGGCAGGCGTCGATATTAAAGG
GGGGAAAATTAGATCCTCGGGAAGAATTTCGGTTAAGGCCAGGAGGAAAGAAACAAATATAGACTAAACACTTAGTATTGGCGAAGCAGGGA
GCTAGAACACTTCGCAAGTAACTCCTGGCCTTTAGAAAACAGCAGAGGGCTGTAAACAAATATAGAACAGCTACAACCCAGCCCTTCAGACA
GGACCGGAGAGCTTAGATCAATTATTAAACAGTAGTAGTCTCTATTGTGTACATCAAGAAATACAGGTAAAAGACCAAAAGAAAGCTT
TAGAGAAAGTAGAGGAAGCAAAACAAAGCAAGAAAAAGTACAGCAAGCAGCACTGGCAGCAGAAACAGCGCCAGGTCTAGCCAAAA
TTACCTTATAGTGCAGAACCTTCAGGGGCAGATGGTACATCAGGCTATATCACCTAGAACTTTAAATGCATGGGTAAAGTAGTAGAGGAG
AAGGCTTTCAGCCAGCAAGTAATACCATGTTTTTCAAGATTATCAGAAAGAGCCACCCCAAGATTAAACACCATGCTAAACACAGCTGG
GGGACATCAAGCAGCCATGCAATGTTAAAGAGAGCCATCAATGAGGAGGCTGCAAGATGGGATAGATTACATCCAGTGCATGAGGGCC
TATTGCACCCCGGAGATGAGAGAACCAAGGGGAAGTGACATAGCAGGAGTACTAGTACCCTTCAGGAACAAATAGGATGGATGACAAAT
AATCCACCTTATCCAGTAGAGAAATCTATAAAAGATGGATAATCATGGGATTAAATAAAATAGTAAGAATGTATAGCCCCCAAGCATTC
TGGATATAAAACAGGACCAAGGAACCCCTTAGAGATTATGTAGATCGGTTCTATAAAACCTTAAGAGCTGAGCAAGCTACACAGGAAGT
AAAAAATTGGATGACAGAACTTGTGGTCCAAATGCAAAACCCAGATTGTAAAGCTATTTTAAAGCAATTAGGACCAGCAGCTACACTA
GAGGAATGATGACAGCATGTGAGGAGTGGGAGGACCCACCCATAAGGCAAGGATTGTGCTGAAGCAATGAGCCAGTATCAAAACCAA
CCATAATGTGTGCAAGAGGCAATTTAAAAACCAAGAAAACTGTTAAGTGTTCATTGTGCAAGAGGCGACATAGCAAAAAATTG
CAGGGCCCTAGGAAAAAGGTTGTGGAAATGTGGACAGGAAGGACCAAAATGAAAGTTGTATGAAAGACAGCTAATTTTTTGGG
AAAACTGGCCCTCCCAAGAGGGAGGCGAGGAAATTCCTTCAGAGCAGACAGAGCCAAAGCCCCACAGAGGAGGCTTCAGGTTTG
GGAGGAGACAAACAGCCCTCTCAGAAAGCAGGAACGATAGACAGGACCTGTATCCCTCAGCTTCCCTCAAACTCACTTTTGGCAACGA
CCCTTGTGCACAAATA AGTATGAGAGGCACTAAAGGAAGCTCTATTAGATACAGGAGCAGATGATACAGTATTAGAAGACATAAATTTG
CAGGAAATGGAAACCAAAATGATAGGGGAATTGGCGGTTTTTCAAGTAGAGCAGTATGATCAGTACCATAGACATCTGTGGGCA
TAAAGCTATAGGTACAGTATTAGTAGGACCTACACCTGTCAACATAATAGGAAGAAACCTGTGACTCAGATTGGATGCATTTAAATTT
CCCATTAGTCTTATGAACATATACAGTAAAAATTAAAGCCAGGAATGGATGGCCAAAAGTTAAACAAATGGCCATGACAGAGAGAAAA
TAAAGCAATTAGTAGAAATTTGTACAGAAATGGAAAGGAAGGAAAAATTTCAAAAATTGGGCTGAAATCCATCAATCTCCAGTATT
TGCCTAAGAAAAAGCAGTACTAAGTGGAGAAAAATTAGTAGATTTCAGAGAACTTAATAAGAAACTCAAGACTTCTGGGAAGTCAA
TTAGGAATACACATCTCTGCAGGTTTAAAAAGAAAAAGTCACTAACAGTACTGGATGGGATGATCATCTTTTCAGTTCCCTAGATA
AAAACTTAGGAATATCTGCATTTACCATACCTAGTATAAACAATGAGACACCTGGAATTAGATACAGTCAATGTCTCCACAGG
ATGGAAGAGTACACAGCAATTTCCAAAGTAGCATGACAGAAATCTTAAAGCTTTTAGAAAAAATAATCAGACATAGTTATCTATCAA
TATATGGATGATTGTATAGGATCTGACTTAGATATAGAGCAGATAGAGCAAAAATAGAGAACTGAGACACATCTGTGGCATGGG
GATTTACACGCGCAGACAAAAACATCAAAAGAACCTCAATTTCTTGGATGGTTATGAACCTCACTGATAAATGGACAGTACAGCC
TATAGTGTGCGCAAGAAAGACAGCTGGACGCTCAATGACATACAGAACTAGTAGGAATACTAATGGGCAAGTCAATGGGCAAGTCAAGTATACGAGGG
ATTAAAGTAAGGCAACTATTGAATCCTTARAGGAGCCAAAGCACTGACAGAACTATAAACAATAACAAAGAACCAAACTAGAACTGG
CAGAAACAGGGAGATTTTAAAGAACAGCTACATGGAGTGTATTATGACCCATCAAAAGCTAATAGCAGAACTACAGAGCAGGGACA
AGGCCAGAGGACCTATCAAAATTATCAAGAGCCATTTAAAAATCTGAAACTGAAATAATGCAAAATAGGGGGTGGCCACACATATGAT
GTAAAACAATTAGTAGAGCAGTGTAAAAAATATCCACAGAAAGCATAGTAATATGGGAAGGACCTCAAAATTTAGACTACCCATACAAA
AGGAAACATGGGAAGCATGGTGACAGACATTTGGCAAGCCACCTGGATTCTGAGTGGGAGTTTGTAAATACCCCTCCCTTAGTGAATTT
ATGGTACAGCTTAGAAAGTAAACCAATTAATAGGAGCAGAACTTTCTATGTAGTAGGAGCAGTAAACAGGAGACTAAATAGGAAAGCA
GGATATGTTACTACAGAGGAGACAAAGGTTGTCTCCATACTGACACACAAATCAGAGAGCTAGTTACAGCAATTCATCTAGCTC
TGCAGGATTCCGGATCAGAACTAATCAATAGTAACAGACTACACAATGCAATTAGGAATCTTCAAGCACAACAGTAAAGTGAATCAGA
AATAGTCAGTCAATTAATAGAACATTTAATAAAAGGAAAGGGTCTACCTAGCATGGGTACAGCACAAAGGAATTTGGAGAAACGAA
CAAGTAGATAAATTAGTCAGTGTGGAATCAGGAAGTACTATTTTAGATGGGATAGTAGAGGCCAAGGAACATGAGAAATATACACA
ATAAATTGGAGAGCAATGGCTAGTGAATTTAACCTACCGCTATAGTAGCAAAAGAAATAGTGCCAGCTGTGACAAATGTACGCTAAAAG
AGAAGCCATACATGGACAAAGTGAGCTGTAGTCCAGGAATATGGCACTAGTTGTACACATTTAGAAGGAAAGTTATCCTAGTAGCAGTC
CATGTAGCCAGTGGATATAGAGACAGAGGTTATTCAGCAGAGACAGGACAGGAAACAGCATACTTTCTTAAATTAGCAGGAAGAT
GGCCAGTAAAAACATACATACAGACAAATGGCAGAAATTTACCTGTACTACAGTTAAGGCCGCTGTGTGGTGGCGGGGATCAACAGGA
ATTTGGGCTCCCTACAACTCCCAAGTCCAGGAGTAGTAGAATCCATGAATAAGAAATTAAGAAAAATATAGGACAGGTAAGAGATCAG
GCTGAACATCTTAAAGCAGCAGTACAATGGCAGTATTCTTACCAATTTAAGAGAAAGGGGGGATTTGGGGGTACAGTGCAGGGGAAA
GAATAATAGACATATAGCACAGACATACAACTAAAGAACTACAGAAACAAATACAAAAATCAAAATTTCCGGTTTATTACAGGGA
CAACAGAGACCCTGTGGAAGGAGGACGCAAGCTCCTCTGGAAGGTGAAGGGCAGTAGTAATACAAAGATAATAGTACATAAAGGTA
GTGCCAAGAAAGAAAGCAAGATCATTAGGATATAGGAAACAGATGGCAGGTGATGATTGTGTGGCAAGTGACAGGATGAGGATTAGA
ACATGGATAAAGCTAGTGAACATCATATATATTTCAAGAAAGCTAAAGGATGGTGTATAAACATCACTATGAACCCATCATCCAA
GAAAAGTTCAAAAGTACACATCCCACTAGGGGTTGCTAGGATGGTAATAACACATATTGGGGTCTGCATACAGGAGAAAGAGACTGGCA
TTTGGGTCAGGAGTCTCCGTAGAAATGCTGAAATGGAGGAAAGGAAATTAACACAAATAGACCTAACCTAGCAGACCAATTAATCCATACAT
TACTTTGATTGTTTTCAGAACTCTGTATAGAAAGCCATATTAGGACATATAGTTAGGCCCTAGTGTGAATATCGAGCAGGACATACAA
AGGTAGATCTCTACAGTACTTGGCAGTAAACAGCATTAATACACCAAAAGAGATAAGGCCACTTTTACCTAGTGTATAACAACTGACAGA
GGATAGATGAAACAGGCCCAAGAACCAAGGGCCACAGAGGTCACATCAATGAATGGACATTAGAATTTTGAAGAGCTTAAGAATG
AGGCTGTGAGACATTTTCTAGATAGTGGCTTCCATGGCTTAGGGCAATATATCTATGAACTTATGGAGATCTTTGGACAGGAGTGAAGC
CATTAATAAGAGTCTGCAACAACTGCTGTTTATTCTATTCAAAATTTGGTGCCGACATAGCAAAATAGGCATTAACTACAGAGGAGGACA
AGAAATGGAGCAGTAAGAACCTAGCTAGAGCCCTGGGAGCACCCAGGAGCGCGCTAGGACCGCTGTATCCCGTTGCTACTGCAAAAAA
TGTGCTGTTTATTGTCCAAAGTGTCTTTAACAAGAGGCTTAGGCATCTCTATGCGAGGAAGAGCGGAGACGACGAAAAATCTCCTC
AAGACAGTAAAGCATCATCAAGTTCTCTATCAAGCAGTGAATTAATACATGTAAATGCAATCTTTAATAATAGCATCAATAGTAGGATTAG
TAGTAGTAGGAATATAGCAATATAGTATAGTGTGCTATAGATTATCATGAATATAGGAATATTAAGCAAAAGAAAAATAGAAAGTTAAT
TGATAGAATAAGGAAAGAGCAGAGACAGTGGCAATGACAGTGAAGGGGATCAGGAAGAATATTTTCACTTATGGAATGGGCGACGAT
GCTCCTTGGGATTAATGATCTGTAGTGCTGCAGACAGTGTGGGTGCAGCTTATATGGGTCACCTGTGTGGAAAGAGCTACCACC
ACTTTATTTTGTGATAGTGTAAAGCATATAATACAGAAAGGCATAAATTTGGGCCACATGATGATGTACCCACAGAGCCCAAC
CACAAGAAATAAATTTGGAATGTGACAGAAAATTTTAAATATGGAAGAAATTAACATGTGAACAGATGATGAGGATATAATCAGGCT
ATGGGAGGAAGCTTTAAGCCATGTGTAAATTAACCTCACTCTGTGTTACTTTAAATTGCACTGATGTATAACCAATAGTAACCCGACA
AATACCATAGCGAATGGGGGAAAGATGGAGGAGGAGAAATAAAAAAGTGTCTTTCAAGACCAACCCAGTCAATAAGACAGAGGAA
AGGAACAATATGCAATTTGTTTAACTTTGATGTAGTACAAATGGAGGATAATGACTATAGACTGATAAGTTGTAACACCTCAGTCAT
TACACAGGCTGTGCAAGATTCCTTTGAGCCAAATCCAATACATATTGTGCCCCAGCTGGTTTGCATTTTAAAGTGTAAACAAAG
ACATTCAGTGGAAAGGACATGACAAATGTGACGACAGTACAAATGTACATAGGAATTAGGCCAGTAGTATCAACTCAATGCTGTTAA
ATGGCAGTGTAGCAAGAAAGATGACTAGTGTAGATCTGACAAATTTCTGAAACATGCTAAACCAATAATAGTACAGCTGAAACCCCTGT
AAATATTACTTGTTTAAGACCAACAATAATCAAGAAAGGATATACATATAGGGCCAGGAGAGCATTATACAAACAGGAGAAATATA
GGAGATATAAGCAAGACATTTGTAACATTAGTAGAGAACAAATGGAATGACACTTAAAGCAAGTGGCTGCCAAATAGGAAACAAATTTG
GGAATAGTAAACAAATAGCTTTAAGCAATCCTCAGGAGGGGACCAAGAAATGTATGATAGTATTTAATTGTGGAGGGGAATTTTCTA
CTGTAATACACAAACTGTTTAAATAGTACTTGGACTAATAGTACTTGGAAATAGAAATAGTCTGAAATATAATCATGATGAAACATCACA
CTCCAATGACAGATAAAACAAATTTAATAACCTGTGGCAAGAGTAGGAAAGCAATGTATGCCCTCCCATCCAAGGAATAATAGATGCA
CGTCAAACTTACAGGATACATTTAACAAGAGATGCTGGTAAATAACCAAAACGGGAGTGAATCTTACAGCTGGAGGAGAGATG
GAGAGCAATTTGAGAAAGTGAATATATAAATATAAGTAGTAAAAATTGAGCCATTAGGATAGCACCCCAAGGCAAGAGAGAGTG
GTGACAGAGAAAAAGAGCAGTGGGATTAGGAGCCCTGTTCTCGGGTCTTTGGGAGCMGACGGAAGCAGTATGGGCGCAGCGTCACTAA
CGCTGACGGTACAGCCAGAACTGCTGTTCTGGTATAGTGCAACAGCAGAAACATCTGCTGAGAGCTATTGAGGCGCAACACATATGTT
GCAACTCAGAGTCTGGGCAATTAACAGCTCCAGGCAAGAGCTCTGGCGGTGGAAGATACCTAAGGATCAACAGCTCCTGGGATTTGG
GGTGTCTGGAAACTCAATTGCACTGCTGTGCTGTTGGAATGCTAGTTGGAATATAAATCTCTAGAAAAAATTTGGGAAAAATGAC
CCTGGATGCAATGGGAAAGGAAATGCAATTTACACAGACATATAACACCTTACTTGAAGATTCGAAAAACGAGAAAGAAATGA
ACAAAGATTTTGGAAATTTGGATAAATGGCAAGTTTGTGGAGTTGGTTGACATACAAAATTTGGTGTGTTATATAAATATTTATAATG
ATAGTAGGAGGCTTAATAGGTTAAGAAATAGTTTTTACTGTACTTTCTATAGTAAACAGAGTTAGGCAGGATATTCCCAATATCGTTGC
AGACCCGCTCCCGAGCCAGAGGGGACCCGACAGGCGCGGAGGAATCGAAGAAAGGTTGGAGAGCAGACAGAGCAGATCCGCTCCCTT
AGTGATGATGATTCTTAGCAATTTATCTGGGTGACCTACGGAGCTGTTCTCTTCACTACCAACCGCTTGAGAGACTTAGTCTTGTATGTTG
ACGAGGACTGTGGAATCTTGGGACGAGGGGGTGGGAAGCCCTCAAGTATTGTTGAATCTCCTGCAATTTGGAGTCAGGAACATAAGA
ATAGTGTGTTAGCTTCTCAATACAGCCATTACAGTAGTACGAGGACAGATAGGTTATAGAAGGCTTACAAAGAGCTGGGAGAGC
TATACTTACATACCTACTAGAAATAGACAGGCTTGGAAAGGGCTTTACTATAAATGGGTGCAAAATGCTCAAAACCTAGCATAGAGG
ATGGCAGGCTGTAGGGAAGAAATGACAGCAGCAGGACCCAGATAGGCCAGCAGCAGGAGGGGTGGAGCAGCATCTCAAGACCTGGGA
AGCATGGAGCAATCAACAGCAGCAATACAGCAGTACCAATGCTGATTTGTGCTGGCTAGAAACACAAACAGAGGGGAGGAAGTGGGT
TTCCAGTCAGACCTCAGGTACCTTAAAGCAATGACCTTCAAGGAGGCTTGGATYTTAGCCACTTTTTAAGAGAAAAGGGGGAGCTGGA
AGGGTTAAATTTACTCCCAACAAAGACAGATATCCTTGATCTGGGTCTACACACACAAAGGCTACTTCCCTGATTGGCAGAACTACACA
CCAGGGCCAGGAGCAGATTTCCACTGACCTTTGGGTGGTGTCTCAAGCTAGTACCAAGTGAACAGAAAGATAGAGGCTGATGAAAGG
AGACAGCAGCTTGTACACCCCGTAAAGCTACATGGATGGATGATGACCCAGAGAAAGAGTTGACAGTGGAGTTTGAACGCGCTAGC
ATTTTCATCAGGTGGCCGAGAGCTGCATCCGAGTACTACAGGAGTCTGACATCGAGCTTTTACAAAGGACTTTCCCGCTGGGAGCTT
CCAGGGGAGCGGTACCGGG
```

Leyenda figura 2

La secuencia de uno de los nueve genes del virus del SIDA (el gen denominado gag) está subrayada para ilustrar su tamaño.

Una **secuencia** es una cadena ordenada de símbolos que pertenecen a un cierto alfabeto. El orden relativo en que dichos caracteres se presentan a lo largo de la secuencia para definir las palabras permite codificar el mensaje biológico apropiado.

De acuerdo con este marco formal de trabajo podemos definir cualquier secuencia genérica (S), sea un gen (G) o una proteína (P), como una sucesión particularmente ordenada de símbolos (denotaremos el tamaño de una secuencia S como $|S|$):

Figura 3. Definición formal de una secuencia.

$$S = \langle s_1 s_2 \dots s_n \rangle; \quad \forall i: 1 \leq i \leq n: s_i \in \Sigma$$

$$G = \langle s_1 s_2 \dots s_n \rangle; \quad \forall i: 1 \leq i \leq n: s_i \in \sum_{ADN}$$

$$P = \langle s_1 s_2 \dots s_n \rangle; \quad \forall i: 1 \leq i \leq n: s_i \in \sum_{PRO}$$

La comparación de dos secuencias distintas permite, en ocasiones, inferir nuevo conocimiento sobre la evolución de ambas moléculas. Para evaluar el parecido entre dos cadenas, es necesario identificar los cambios en el orden en que aparecen los distintos símbolos que constituyen cada secuencia.

Un **alineamiento de dos secuencias** es una superposición exacta de los caracteres de ambas cadenas que arbitrariamente determinará el número de símbolos similares coincidentes en cada posición de éstas.

Formalmente, dadas dos secuencias $S = \langle s_1 \dots s_m \rangle$ y $S' = \langle s'_1 \dots s'_n \rangle$ pertenecientes a un mismo alfabeto finito Σ , definiremos un alineamiento como una correspondencia C entre los símbolos de ambas secuencias

$$C(S, S') = \left\{ (s_{i_1}, s'_{j_1}) \dots (s_{i_T}, s'_{j_T}) \right\}, \text{ tal que:}$$

Lecturas complementarias

D. Mount (2001). *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. ISBN: 0879696087.

S. Batzoglou (2005). "The many faces of sequence alignment". *Briefings in bioinformatics* (núm. 6, págs. 6-22).

Figura 4. Definición formal de un alineamiento de T símbolos.

1. Debe preservarse el orden relativo de los símbolos:

$$1 \leq i_1 \leq \dots \leq i_T \leq m, 1 \leq j_1 \leq \dots \leq j_T \leq n$$

2. Pueden existir símbolos no alineados:

$$\exists k: s_k \in S \wedge s_k \notin C, \quad \exists l: s'_l \in S' \wedge s'_l \notin C$$

3. Cada símbolo puede alinearse únicamente con otro:

$$(s_i, s'_j) \in C \Rightarrow \forall k: s'_k \in S' \wedge k \neq j: (s_i, s'_k) \notin C$$

4. No se permiten inversiones en el alineamiento:

$$(s_i, s'_j), (s_k, s'_l) \in C, \quad i < k \Rightarrow j < l$$

Los alineamientos resultantes pueden disponerse gráficamente en forma de matriz: utilizaremos las filas para presentar las secuencias y las columnas para denotar el tipo de parecido entre los símbolos alineados en cada posición. Por ejemplo, el siguiente alineamiento entre las secuencias $S = \langle AAGTTC \rangle$ y $S' = \langle AGCCG \rangle$ localiza dos coincidencias y una sustitución:

Figura 5. Ejemplo de alineamiento entre dos secuencias.

S =	A	A	G	T	T	C
S' =	A	-	G	C	C	G

Analizando el alineamiento mostrado en la figura 5 observamos que pueden ocurrir los siguientes emparejamientos en el interior de cada columna del alineamiento:

- Coincidencia (en inglés, *match*). Observamos el mismo símbolo en ambas secuencias. Lógicamente representa el máximo parecido biológico. Denotado indistintamente con los caracteres "|" o "*".
- Sustitución (en inglés, *mismatch*). Observamos un símbolo diferente en ambas secuencias. Dado que determinados nucleótidos o aminoácidos desempeñan un rol biológico similar, es factible permitir su alineamiento. Denotado con el símbolo "|", o con los caracteres ":" o ".", si existen diferentes grados de parecido.
- Inserción/delección (en inglés, *gap*, hueco). Observamos un símbolo en una secuencia que no ha podido ser alineado en la otra. Evolutivamente hablando, un *gap* puede reflejar la inserción de un símbolo en la primera secuencia o su delección en la segunda. Denotado con el carácter "-" en la secuencia donde no hubo coincidencia.

Es posible efectuar miles de posibles alineamientos para un par de secuencias concretas. Modificando la configuración de las inserciones/deleciones y seleccionando cuidadosamente qué coincidencias y sustituciones introducimos, podemos lograr resultados completamente diferentes. En la figura 6 se puede comprobar cómo varía el número de coincidencias en un alineamiento de dos secuencias de aminoácidos cuando permitimos la introducción de *gaps*:

Figura 6. Introducción de *gaps* en los alineamientos.

Alineamiento sin inserciones/deleciones:

M	E	A	N	K	Q	R	W	V	L	A	
											(3)
M	E	A	Q	R	T	F	V	L	C	C	

Alineamiento con *gaps* en la segunda secuencia:

M	E	A	N	K	Q	R	W	V	L	A			
											(5)		
M	E	A	-	-	Q	R	T	F	V	L	C	C	

Alineamiento con *gaps* en ambas secuencias:

M	E	A	N	K	Q	R	-	W	-	V	L	A	-	-	
															(7)
M	E	A	-	-	Q	R	T	-	F	V	L	-	C	C	

Leyenda figura 6

El número de coincidencias detectado en cada caso se muestra entre paréntesis. Para simplificar el ejemplo hemos omitido el uso de sustituciones.

Es fundamental, por tanto, establecer un sistema de evaluación que asigne diferentes puntuaciones a coincidencias, sustituciones y *gaps* para clasificar todos los posibles alineamientos. En la literatura general encontramos básicamente dos clases de sistemas para evaluar la bondad de un alineamiento: similitud y distancia.

La **similitud** entre dos secuencias evalúa el parecido entre éstas, recompensando las coincidencias (y en menor medida las sustituciones) y penalizando las inserciones/deleciones. En estos términos, un alineamiento óptimo permite identificar la máxima similitud posible entre dos secuencias.

La **distancia** entre dos secuencias se define como el mínimo número de cambios necesarios para transformar la primera secuencia en la segunda. Únicamente las diferencias son penalizadas. En estas condiciones, un alineamiento óptimo permite medir la mínima distancia existente entre dos secuencias.

En líneas generales, ambos sistemas de medición proporcionan resultados equivalentes (ver figura 7). Aunque el esquema de distancias resulta más apropiado para modelar el problema en términos evolutivos, el esquema de similitud proporciona mayor versatilidad para realizar diferentes tipos de comparaciones. Dado que esta última aproximación es la estrategia más popular entre los miembros de la comunidad científica, concentraremos en ella todo nuestro interés a lo largo de estos materiales. A menudo calcularemos el porcentaje de similitud de un alineamiento dividiendo el número de caracteres idénticos o parecidos por la longitud de éste. Cuando recompensemos exclusivamente las coincidencias exactas, hablaremos de porcentaje de identidad.

Lectura complementaria

T. F. Smith; M. S. Waterman (1981). "Comparison of bio-sequences". *Advances in Applied Mathematics* (núm. 2, págs. 482-489).

Figura 7. Puntuaciones con los esquemas de similitud y distancia.

S_1 :	T	A	T	A	A	A

S_2 :	T	A	T	A	T	A
---------	---	---	---	---	---	---

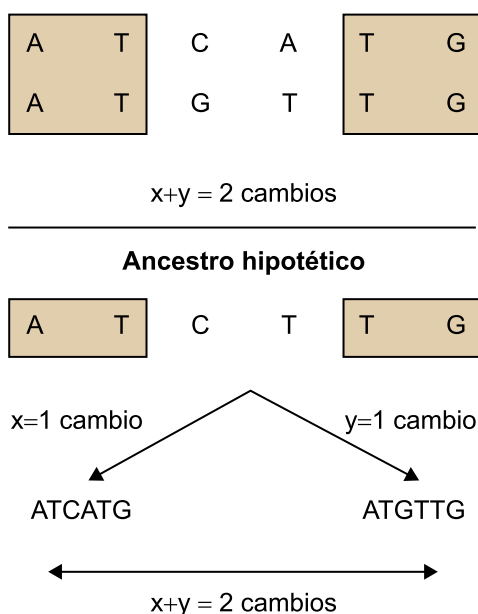
Similitud $(S_1, S_2) = 5$ (83 %)

Distancia $(S_1, S_2) = 1$

2. Interpretación biológica de los alineamientos

La comparación de secuencias es una excelente herramienta para descubrir parecidos causados posiblemente por la existencia de una función común. Una similitud alta entre dos secuencias puede ser indicativa de un rol biológico compartido. En ese caso, podríamos utilizar las propiedades conocidas sobre una secuencia para documentar otras menos estudiadas. El alineamiento puede ser útil para señalar aquellas regiones implicadas en el desempeño de la función conservada (por ejemplo, dominios de proteínas). Cuando evaluamos la similitud entre secuencias de diferentes especies, estamos investigando la existencia de una posible secuencia ancestral común a partir de la cual podrían haber derivado éstas en el pasado. Es a partir de la especiación de dicho ancestro cuando podríamos explicar biológicamente la similitud observada entre dos especies estudiadas. Esta constatación permitiría definir nuevas relaciones de homología. El alineamiento resultante representa, por tanto, un punto de partida adecuado para descubrir la sucesión de cambios ocurridos entre dos secuencias a lo largo de la evolución (ver el ejemplo mostrado en la figura 8).

Figura 8. Interpretación evolutiva de los alineamientos.



Ved también

La relación entre homología y similitud se explica en la asignatura *Fundamentos de biología molecular*.

Leyenda figura 8

Este ejemplo ilustra cómo la distancia entre dos secuencias calculada directamente a partir de su alineamiento representa una buena estimación de la distancia entre cada secuencia con el hipotético ancestro de ambas.

Inserción/delección

Una inserción de un carácter en una posición determinada de una secuencia equivale a una delección en otra alineada con ésta. Según cuál sea la característica mayoritaria en esa columna, este hecho será interpretado simplemente como una inserción o una delección.

En un contexto evolutivo, la acción de los distintos operadores que modelan la forma de los alineamientos posee un significado biológico propio:

- Las coincidencias involucran nucleótidos o aminoácidos con un posible rol funcional dada su conservación a lo largo de la evolución.
- Las sustituciones equivalen a mutaciones puntuales que no afectan a la funcionalidad de las moléculas.

- Las inserciones/deleciones representan cambios más drásticos en cada secuencia.

En resumen, en un alineamiento donde observamos un alto grado de similitud entre las secuencias de dos moléculas, podemos argumentar que éstas desempeñan en estas especies una función biológica parecida.

Para mejorar la precisión de las inferencias evolutivas derivadas del cálculo de alineamientos entre secuencias, es imprescindible incluir funciones de puntuación de éstos dotadas de un mayor sentido biológico. La mutación del material hereditario de las células es la base de la gran variabilidad que observamos en las formas de vida conocidas. Sin embargo, estos cambios puntuales en las secuencias de los genes resultan más o menos relevantes para la función de la proteína según el papel protagonizado por esa parte del péptido. Además, existen aminoácidos con propiedades similares cuyo intercambio no afecta necesariamente al producto final. Todo esto propicia que existan ciertas sustituciones con un mayor grado de aceptación en la selección natural de las proteínas a lo largo de un periodo evolutivo. Otras sustituciones, en cambio, aparentemente se observan con poca frecuencia en las proteínas, quizás porque afectan sustancialmente a la función de éstas.

A partir de un catálogo de proteínas homólogas, podemos estimar la puntuación $M(i, j)$ que deberíamos otorgar a la sustitución de un aminoácido i por otro aminoácido j en el interior de una proteína. Este valor debe calcularse en función de la frecuencia con la que observamos dicha sustitución $q(i, j)$ en esa colección de secuencias, contrastándose después con el valor esperable para éstas simplemente por azar ($p(i)$ y $p(j)$) en función de su abundancia dentro de las mismas secuencias. Aplicando logaritmos, logramos que el resultado de este cociente sea positivo si la sustitución es más frecuente de lo esperado (será negativa cuando el cambio es poco frecuente).

Figura 9. Cálculo de matrices de sustitución.

$$M(i, j) = \log \left(\frac{q(i, j)}{p(i)p(j)} \right)$$

PAM y BLOSUM son las dos familias más populares de matrices de sustitución de aminoácidos.

- Las matrices PAM (del inglés *point accepted mutation*, *mutación puntual aceptada*) se calcularon a partir del análisis de proteínas homólogas muy cercanas evolutivamente. Dado su alto grado de similitud, en estas secuencias están documentadas pocas sustituciones que, obviamente, no afectan a su función biológica (por ello se denominan mutaciones aceptadas). Una unidad PAM de divergencia evolutiva entre dos aminoácidos (PAM1) representa el número de sustituciones observadas normalizado por el número total de cambios cada 100 residuos (1% de mutaciones aceptadas). De

Ved también

En la asignatura *Fundamentos de biología molecular* se explica ampliamente las mutaciones y sus implicaciones genéticas.

Nota

Habitualmente, en estos cálculos no se distingue direccionalidad en el cambio:

$$M(i, j) = M(j, i).$$

forma analítica, mediante la sucesiva multiplicación de varias ocurrencias de la matriz PAM1 fue posible generar matrices para comparar secuencias con un grado mayor de divergencia. Por ejemplo, PAM250 (el resultado de elevar a la potencia 250 la matriz PAM1) equivale a introducir 250 sustituciones por cada cien aminoácidos, asumiendo que pueden concurrir múltiples sustituciones sobre una misma posición durante largos periodos de tiempo.

- Las matrices BLOSUM (del inglés *BLOcks of amino acid SUBstitution Matrix*, matriz de sustitución de bloques de aminoácidos) se calcularon más recientemente a partir del análisis de bloques conservados sin huecos entre proteínas con diferentes grados de parecido. Por ejemplo, para secuencias cuya similitud oscila alrededor del 45% es recomendable puntuar nuestros alineamientos con BLOSUM45. Del mismo modo, existen matrices BLOSUM62 y BLOSUM80, derivadas directamente de secuencias que poseían un máximo de 62% y 80% de similitud, respectivamente. Con estos nuevos esquemas de puntuación de coincidencias y sustituciones, podemos evaluar mejor los alineamientos, dotándolos de un mayor sentido biológico (figura 10).

Lecturas complementarias

M. O. Dayhoff y otros (1965). *Atlas of protein sequence and structure*. Silver Spring, Maryland: National Biomedical Research Foundation.

S. Henikoff; J. F. Henikoff (1992). "Amino Acid Substitution Matrices from Protein Blocks". *PNAS* (núm. 89, págs. 10915-10919).

Figura 10. Puntuación con matrices de sustitución de aminoácidos.

S_1 : T Q L P N

S_2 : T E E A N

Identities (S_1, S_2) = 2

PAM250 (S_1, S_2) = $3 + 2 + (-3) + 1 + 2 = 5$

BLOSUM62 (S_1, S_2) = $5 + 2 + (-3) + (-1) + 6 = 9$

Figura 11. Matrices de sustitución de aminoácidos.

PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
N	0	0	2	3	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1
H	-1	2	2	1	3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-2	0	9	-5	-3	-3	0	7	-1
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

BLOSUM62

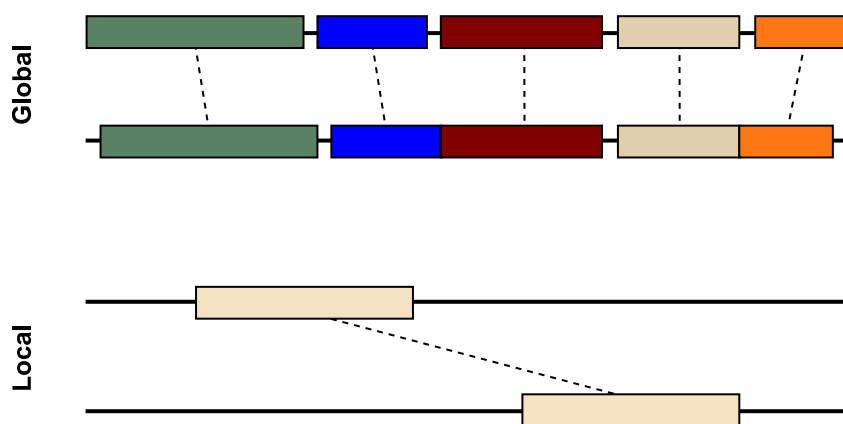
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-3	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

3. Alineamientos globales o locales

El alineamiento óptimo debe maximizar ciertos criterios convencionales de similitud. En líneas generales, el biólogo aprovecha estos alineamientos para inferir nuevo conocimiento a partir de la información previamente conocida sobre las secuencias. En función del contenido codificado en su interior, no obstante, es necesario modificar los criterios que rigen estas comparaciones para lograr un resultado óptimo. De este modo, el nivel de conservación estimado entre las secuencias permite acotar mejor la estrategia de comparación.

No es lo mismo alinear secuencias que desempeñan un rol biológico compartido, que buscar coincidencias en otras secuencias con una relación funcional más débil. El bioinformático debe conocer de antemano estas propiedades para elegir la estrategia de alineamiento más apropiada. Fundamentalmente, disponemos de dos clases de alineamientos, globales y locales, para efectuar la comparación de nuestras secuencias. Según la opción escogida, los resultados pueden ser diametralmente distintos, proporcionando puntos de partida diferentes para la elaboración de posteriores hipótesis científicas (ver figura 12).

Figura 12. Alineamiento global o local.



Aquellas secuencias que codifican elementos con una función biológica parecida probablemente poseen un tamaño y una estructura similares. Por ejemplo, las regiones genómicas codificantes del mismo gen pertenecientes a dos especies diferentes o las secuencias de aminoácidos de dos proteínas homólogas, presentarán necesariamente un alto grado de similitud. En función de la distancia evolutiva entre las especies comparadas, el parecido final resultante será más o menos acentuado. Para efectuar estas comparaciones, es preciso calcular un alineamiento global que recubra la práctica totalidad de las secuencias. Como se puede apreciar en la figura 13, el alineamiento global

Lecturas complementarias

D. Mount (2001). *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. ISBN: 0879696087.

A. D. Baxeavanis; B. F. Ouellette (2005). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Hoboken, NJ: John Wiley & Sons Inc. ISBN: 0471478784.

M. Zvelebi; J. O. Baum (2008). *Understanding bioinformatics*. Londres: Garland Science. ISBN: 0815340249.

Lectura complementaria

S. Batzoglou (2005). "The many faces of sequence alignment". *Briefings in bioinformatics* (núm. 6, págs. 6-22).

de dos proteínas homólogas permite asociar en la práctica a cada residuo su equivalente en el otro organismo. Dada la alta homología entre las dos proteínas, la longitud de ambas es idéntica.

Un **alineamiento global** efectúa la correspondencia entre las secuencias completas, maximizando el número total de caracteres coincidentes a lo largo de las cadenas.

Figura 13. Alineamiento global de la proteína humana *LRRTM1* y su homóloga en el ratón.

```

CLUSTAL 2.0.12 multiple sequence alignment

Humano  MDFLLLGCLYWLLRRPSGVVLCLLGACFQMLPAAPSGCPQLCRCEGRLLYCEALNLTEA  60
Ratón   MDFLLLGCLHWLLRRPSGVVLCLLGACFQMLPAAPSGCPGQCRCEGRLLYCEALNLTEA  60
*****:*****

Humano  PHNLSGLLGLSLRYNSLSELRAGQFTGLMQLTWLYLDHNNHICSVQGDAFQKLRRVKELTL  120
Ratón   PHNLSGLLGLSLRYNSLSELRAGQFTGLMQLTWLYLDHNNHICSVQGDAFQKLRRVKELTL  120
*****

Humano  SSNQITQLPNTTFRMPNLRSDLSYNKLQALAPDLFHGLRKLTTLHMRANAIQFVPVRI  180
Ratón   SSNQITELANTTFRMPNLRSDLSYNKLQALAPDLFHGLRKLTTLHMRANAIQFVPVRI  180
*****:*****

Humano  FQDCRSKFLFDIGYNQLKSLARNSFAGLFKLTLEHNDLVKVNFAHFPRILSLHSLCL  240
Ratón   FQDCRSKFLFDIGYNQLKSLARNSFAGLFKLTLEHNDLVKVNFAHFPRILSLHSLCL  240
*****:*****

Humano  RRNKVAIVVSSLDVWVNLEKMDLSGNEIEYMEPHVFETVPHLQSLQSDNRLTYIEPRIL  300
Ratón   RRNKVAIVVSSLDVWVNLEKMDLSGNEIEYMEPHVFETVPHLQSLQSDNRLTYIEPRIL  300
*****:*****

Humano  NSWKSLSITLAGNLWDCGRNVCALASWLNQFQGRYDGNLQCASPEYAQGEDVLDVYAF  360
Ratón   NSWKSLSITLAGNLWDCGRNVCALASWLNQFQGRYDANLQCASPEYAQGEDVLDVYAF  360
*****.*****

Humano  HLCEDGAEPSTSGHLLS-AVTNRSDLGPPASSATTLADGGEGQHDGTFEPATVALPGGEHA  419
Ratón   HLCEDGAEPSTSGHLLSVAVTNRSDLTPESSATTLVDGGEG-HDGTFEPITVALPGGEHA  419
***** ** *****.*****

Humano  ENAVQIHKVVTGTMALIFSFLIVLVLYVSWKCFPASLRQLRQCFVTQRRKQKQKQTMHQ  479
Ratón   ENAVQIHKVVTGTMALIFSFLIVLVLYVSWKCFPASLRQLRQCFVTQRRKQKQKQTMHQ  479
*****

Humano  MAAMSAQEYYVDYKPNHIEGALVIINEYGSCSTCHQQPARECEV  522
Ratón   MAAMSAQEYYVDYKPNHIEGALVIINEYGSCSTCHQQPARECEV  522
*****

```

Los distintos elementos funcionales que componen el genoma de cada especie, a pesar de desempeñar un amplio abanico de diferentes funciones, están constituidos por los mismos bloques básicos. Las regiones reguladoras de la transcripción de cada gen, por ejemplo, poseen una configuración única de sitios de unión para algunos factores de transcripción que permite que la célula gradúe la actividad de dicho gen. Las proteínas también están divididas en pequeños dominios que proporcionan una función bioquímica específica. Todas estas regiones integran en su interior una mezcla diferencial de diversos elementos funcionales. La comparación global de esta clase de secuencias, sin

Lectura complementaria

S. Batzoglou (2005). "The many faces of sequence alignment". *Briefings in bio-informatics* (núm. 6, págs. 6-22).

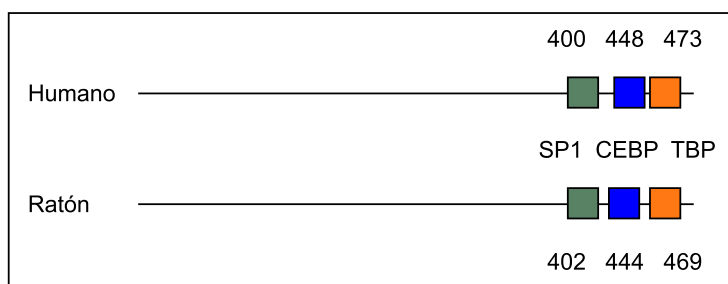
embargo, no arroja nuevo conocimiento, dado que, en general, su contenido no está altamente conservado. Para distinguir aquellos fragmentos más similares del resto es preciso efectuar un alineamiento local. Como mostramos en la figura 14, el alineamiento local entre la región promotora del gen humano *LEP* y su secuencia ortóloga en el genoma del ratón sólo refleja un fragmento parcialmente conservado cerca del inicio de transcripción, donde previamente había sido documentada la existencia de varios sitios de unión a ciertos factores de transcripción.

Un **alineamiento local** realiza exclusivamente la correspondencia entre aquellos fragmentos de las secuencias que poseen una coincidencia máxima de caracteres, descartando el resto de regiones a lo largo de dichas secuencias que no presentan una mínima similitud.

Frecuencia de alineamientos globales y locales

Siempre es posible obtener un alineamiento global entre dos secuencias (aunque en la mayoría de casos sea muy deficiente). Por el contrario, es poco frecuente identificar alineamientos locales dado que es necesario que exista cierta conservación funcional.

Figura 14. Alineamiento local de la región promotora del gen *LEP*.



NCBI Blast2seq

Score = 86.0 bits (94), Expect = 3e-21

Identities = 81/102 (80%), Gaps = 1/102 (0%)

Strand = Plus/Plus

Query 400 GGGCGGGGCGGGAGCTGGCGCTAGAAATGCGCCGGGGCCTGCGGGGCAGTTGCGCAAGTT 459

[illegible]

Sbjct 397 GGGTGGGGCGGGAGTTGGCGCTCGCAGGGA-CTGGGGCTGGCCGGACAGTTGCGCAAGTG 455

Query 460 GTGATCGGGCCGCTATAAGAGGGGCGGGCAGGCATGGAGCCC 501

Subject 456 GCACTGGGGCAGTTATAAGAGGGGCAGGCAGGCATGGAGCCC 497

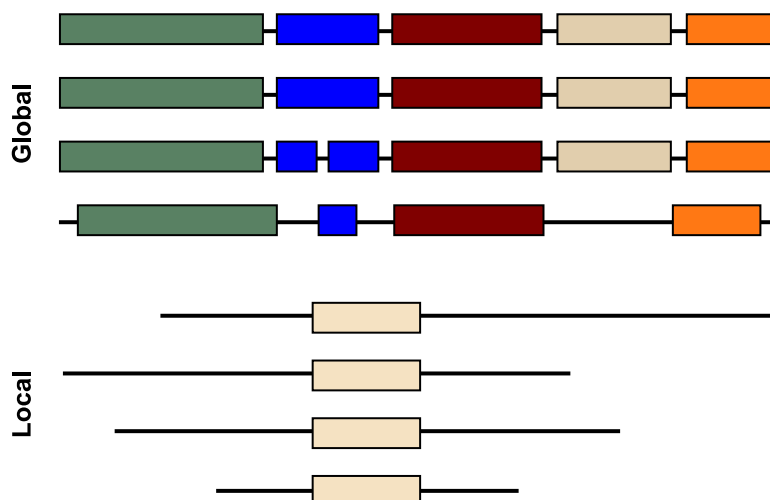
Leyenda figura 14

El inicio de transcripción de cada gen está ubicado en la posición 500 de ambas secuencias. Los sitios reconocidos por cada factor de transcripción están indicados con cajas de distintos colores a lo largo de los promotores. El alineamiento local mostrado en la parte inferior identifica precisamente esta región más cercana al gen, descartando el resto de las secuencias.

4. Alineamientos simples o múltiples

Empleando distintos bancos de datos, los investigadores pueden recopilar un número más elevado de secuencias relacionadas por algún vínculo funcional. Es posible inferir nuevo conocimiento a partir de estas comparaciones para clasificar diferentes familias de secuencias en función del grado de conservación. Para contrastar estas hipótesis sobre estos conjuntos de datos es preciso calcular alineamientos múltiples de secuencias (en inglés, *multiple sequence alignment* o MSA). Sin embargo, desde el punto de vista computacional no es posible abordar en la práctica el cálculo del alineamiento múltiple óptimo para un número elevado de secuencias debido a su excesiva complejidad. Para solventar esta limitación es preciso introducir determinados requisitos sobre el modo de obtener la solución en un tiempo razonable, llevando a cabo aproximaciones globales o locales en función del problema biológico. Cuando podemos asumir que las secuencias codifican elementos similares dentro de un mismo contexto biológico, el alineamiento múltiple global permite reconstruir la estructura de bloques conservados a lo largo de éstas (figura 15).

Figura 15. Alineamientos múltiples de secuencias.



Un **alineamiento múltiple global** de secuencias realiza la correspondencia entre todas las secuencias completas, maximizando el número de caracteres coincidentes en cada posición de éstas.

Desde los albores de la biología computacional existe una íntima relación entre el análisis de árboles filogenéticos y los alineamientos múltiples de secuencias. Aunque pueden crearse taxonomías evolutivas de forma independiente, la búsqueda de un alineamiento múltiple constituye un excelente punto de

Lectura complementaria

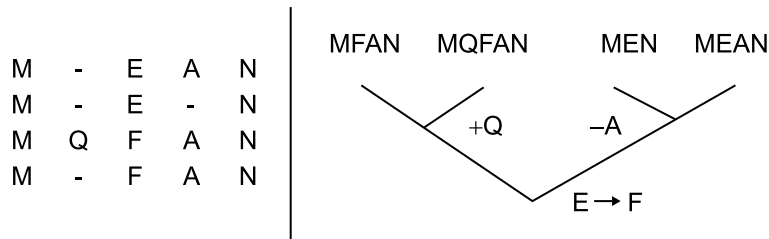
S. Batzoglou (2005). "The many faces of sequence alignment". *Briefings in bioinformatics* (núm. 6, págs. 6-22).

Ved también

En la asignatura *Fundamentos de biología molecular* se explican con detalle los métodos fundamentales de construcción de árboles filogenéticos.

partida para iniciar la construcción de jerarquías filogenéticas. En la figura 16 podemos observar el árbol resultante derivado de las diferencias existentes entre los residuos del alineamiento global calculado previamente.

Figura 16. Árboles evolutivos derivados de alineamientos múltiples globales.



Cuando los elementos codificados en el interior de las secuencias corresponden a pequeños bloques comunes, incrustados en otras regiones más grandes de menor parecido, es preciso obtener un alineamiento múltiple local de estas (figura 15). En terminología algorítmica este bloque altamente conservado de símbolos recibe la denominación de motivo o patrón (en inglés, *motif* o *pattern*). Es por ello por lo que esta familia de alineamientos ha sido rebautizada también en el campo de la inteligencia artificial como búsqueda de motivos o descubrimiento de patrones (en inglés, *motif finding* o *pattern discovery*).

Un **alineamiento múltiple local** efectúa únicamente la correspondencia entre aquellos fragmentos de varias secuencias que poseen una coincidencia relevante, descartando el resto de regiones a lo largo de todas las cadenas de caracteres que no exhiben esta similitud.

Para emular de un modo más realista ciertos contextos biológicos (como las regiones reguladoras de la transcripción), es necesario alterar algunas reglas de construcción de los alineamientos. Como hemos visto antes (figura 4), la disposición de los elementos coincidentes en los alineamientos debe preservar el orden original establecido en las respectivas secuencias comparadas. Esta propiedad, denominada colinearidad, permite restringir la búsqueda a aquellos elementos comunes a dos o más secuencias cuyo orden relativo se haya preservado también. La aplicación de esta regla beneficia al usuario, dado que se calculará el alineamiento final en menos tiempo. Sin embargo, en ocasiones es necesario tener en cuenta estas combinaciones para obtener el resultado final. De hecho, el alineamiento local puede producir de forma natural ordenaciones de fragmentos que presentan patrones no colineales de conservación (figura 17).

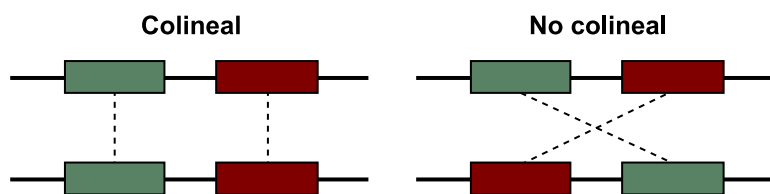
Lectura complementaria

A. Brazma y otros (1998). "Approaches to the automatic discovery of patterns in biosequences". *Journal of Computational Biology* (núm. 5, págs. 279-305).

Lectura complementaria

E. Blanco y otros (2007). "Multiple non-collinear TF-map alignments of promoter regions". *BMC Bioinformatics* (núm. 8, pág.138).

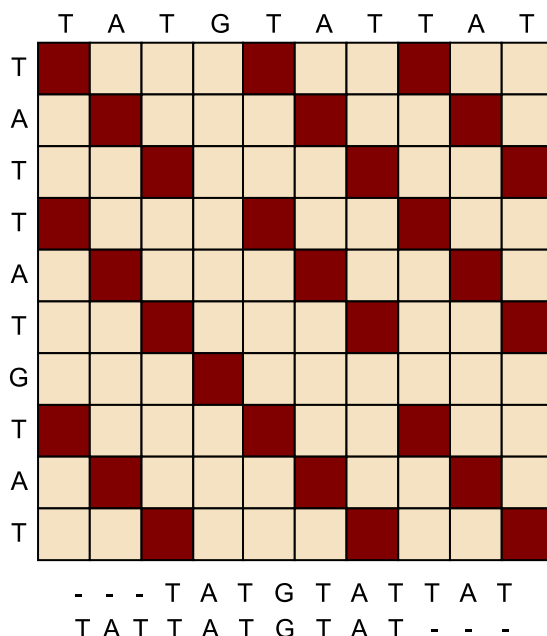
Figura 17. Colinealidad en los alineamientos.



5. Matrices de puntos

Las matrices de puntos (en inglés, *dot matrices* o *dot plots*) son un modo efectivo de visualizar gráficamente las regiones más similares entre dos secuencias. Para trasladar estas comparaciones sobre una matriz, ubicaremos la primera secuencia en el eje horizontal de la cuadrícula y la segunda en el eje vertical de la misma. De esta manera, cada posición de la matriz representa una posible correspondencia que debemos marcar con un punto cuando exista coincidencia en esos caracteres entre ambas secuencias. El gráfico resultante debe ser analizado en busca de aquellas diagonales con una longitud significativamente más grande (ver figura 18).

Figura 18. Matriz de puntos sobre dos secuencias genómicas.



La comparación entre dos secuencias con un alto grado de conservación se plasmará en un robusto solapamiento a lo largo de la diagonal de la matriz de puntos. En caso de existir, las inversiones aparecerán representadas como diagonales en sentido contrario. Si comparamos una secuencia contra sí misma, podemos descubrir regiones repetitivas en su interior. Es importante remarcar que con esta herramienta en ningún caso se procede efectivamente a calcular el alineamiento de las secuencias. Existen numerosos programas que implementan la generación de matrices de puntos, añadiendo nuevas opciones para configurar los gráficos resultantes (ver tabla 1).

Lectura complementaria

A. J. Gibbs; G. A. McIntyre (1970). "The diagram, a method for comparing sequences. Its use with amino acid and nucleotid sequences". *European Journal of Biochemistry* (núm. 16, págs. 1-11).

Leyenda figura 18

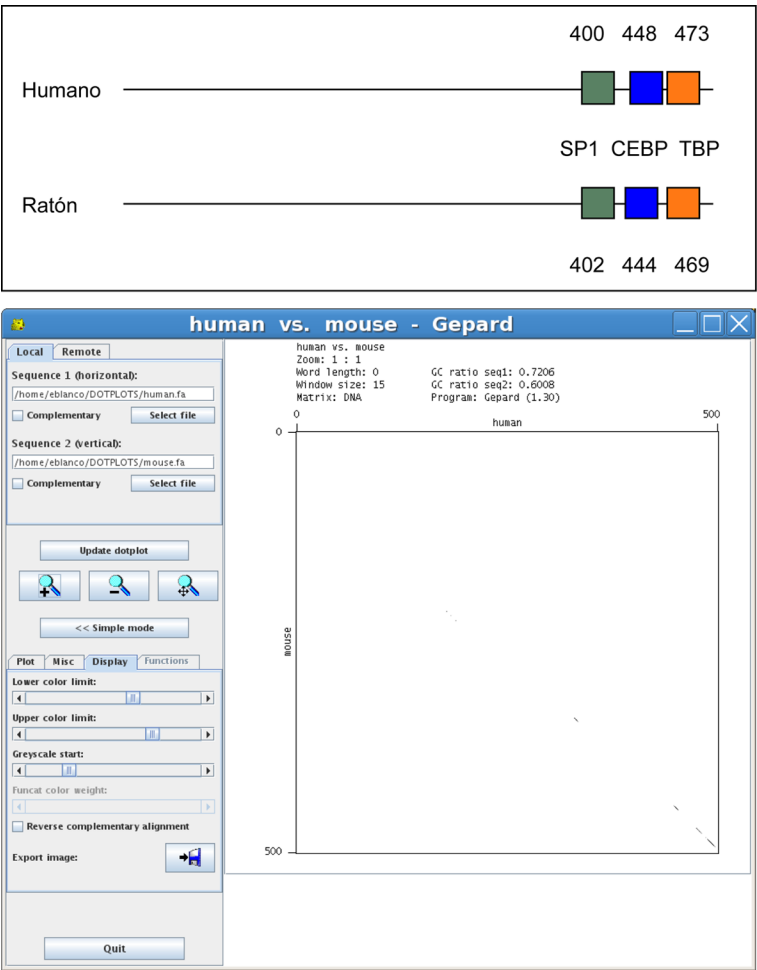
Mostramos solamente las diagonales constituidas por tres o más posiciones consecutivas idénticas. No se han permitido, en este caso, inversiones dentro de la matriz. El posible alineamiento resultante representado en la parte inferior corresponde a la diagonal de mayor tamaño.

Tabla 1. Programas de representación de dot plots.

Nombre	Referencia
Dotlet	<i>Bioinformatics</i> (núm. 16, págs. 178-179) (2000)
JDotter	<i>Bioinformatics</i> (núm. 20, págs. 279-281) (2004)
Gepard	<i>Bioinformatics</i> (núm. 23, págs. 1026-1028) (2007)

Para mostrar el funcionamiento de estas aplicaciones, vamos a comparar mediante el programa Gepard el promotor del gen humano *LEP* con su homólogo en el genoma de ratón. Codificados dentro de estas secuencias, conocemos de antemano la localización de tres sitios de unión –validados experimentalmente– que son reconocidos por factores de transcripción, relativamente cerca del inicio de transcripción de este gen (ubicado en la posición 500). En la figura 19 podemos observar la correlación entre la ubicación exacta de los tres factores de transcripción, la pantalla principal del programa y la matriz de puntos generada a partir de ambas secuencias. El estudiante puede apreciar la identificación de varias diagonales en la esquina inferior de la matriz generada por Gepard.

Figura 19. Matriz de puntos sobre dos regiones promotoras.



Nota

Recomendamos ejecutar alguno de estos programas para experimentar su funcionamiento sobre secuencias obtenidas de algún banco de datos.

Lecturas complementarias

A. D. Baxevanis; B. F. Ouellette (2005). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Hoboken, NJ: John Wiley & Sons Inc. ISBN: 0471478784.

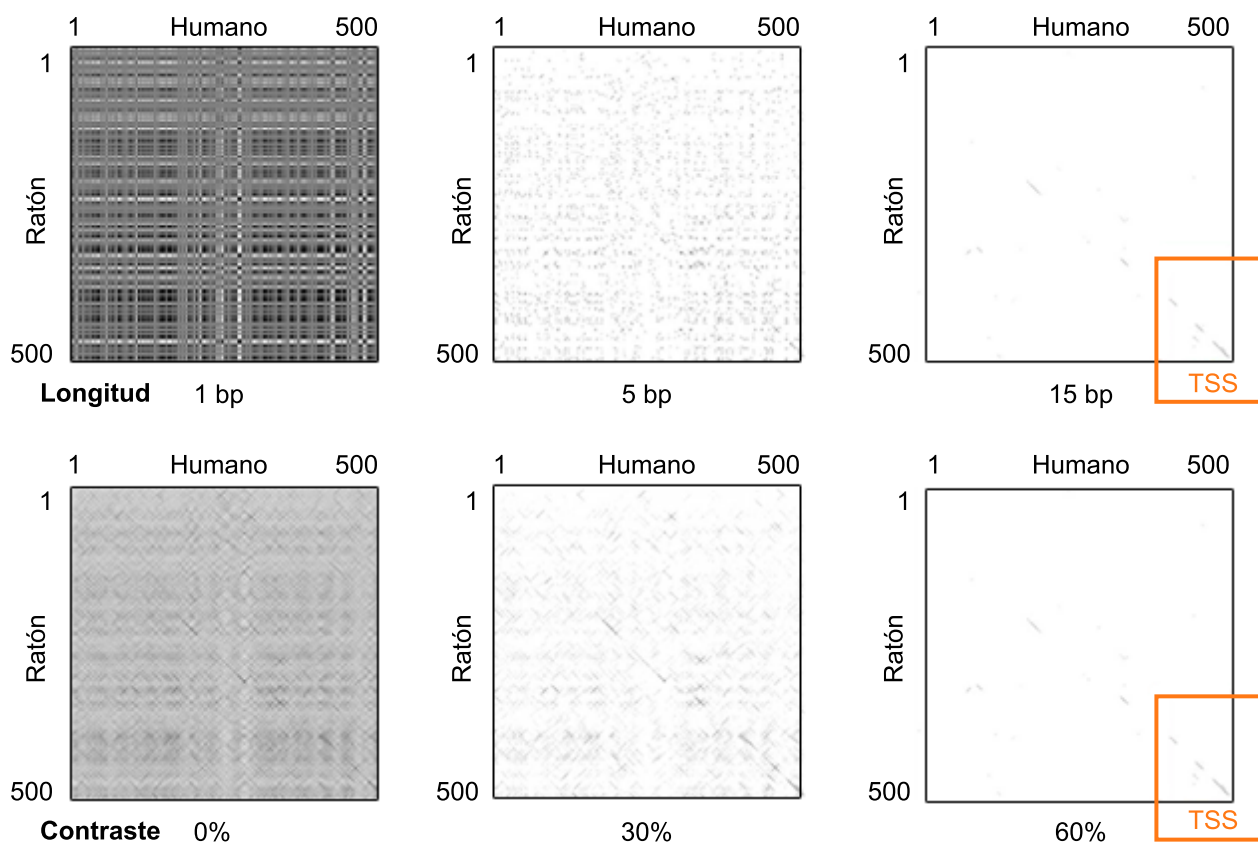
E. Blanco; D. Farre; M. Alba; X. Messeguer; R. Guigó (2006). "ABS: a database of Annotated regulatory Binding Sites from orthologous promoters". *Nucleic Acids Research* (núm. 34, págs. D63-D67).

Leyenda figura 19

Tanto las secuencias como las anotaciones han sido extraídas de la base de datos ABS.

El usuario de estos programas puede configurar numerosos parámetros y enriquecer los resultados. Para disminuir el ruido debido a las coincidencias espontáneamente distribuidas al azar a lo largo de las secuencias, es posible marcar en la matriz únicamente aquellos segmentos que contienen un mínimo número de aciertos consecutivos. Para mejorar la visibilidad y destacar gradualmente los mejores segmentos de la matriz, podemos modificar dinámicamente la gama de grises. El efecto de estos parámetros puede apreciarse claramente en la figura 20. Jugando con distintos contrastes vemos que las diagonales más nítidas se detectan precisamente en torno al inicio de transcripción de los genes.

Figura 20. Configuración de las matrices de puntos.



TSS, del inglés *transcription start site*, inicio de transcripción de los genes.

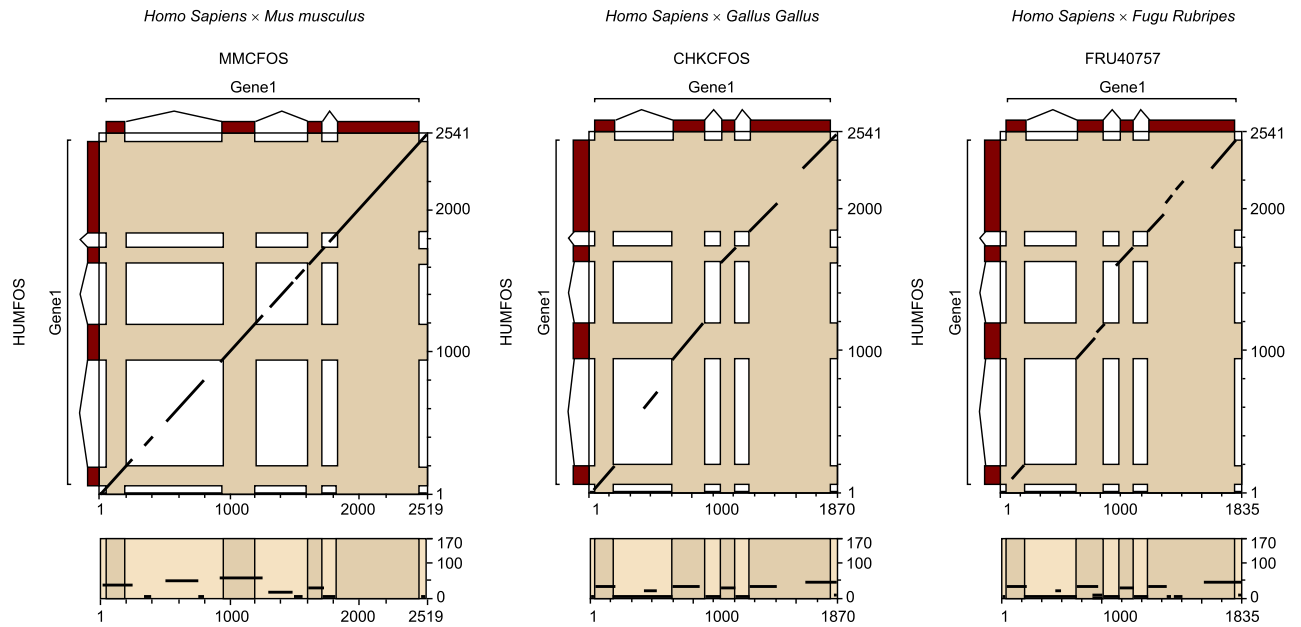
Con el objetivo de estudiar la conservación de los elementos funcionales del genoma, las matrices de puntos pueden enriquecerse empleando información adicional sobre las secuencias comparadas. El programa GFF2APLOT permite superponer una serie de anotaciones codificadas en el interior de las regiones similares de dos secuencias, utilizando cualquier programa de alineamiento externo. De este modo, el usuario puede estudiar el grado distinto de conservación de las secuencias dentro de la fracción funcional en comparación con el resto. Como se puede observar en la figura 21, mediante estas representaciones podemos superponer las regiones más similares sobre el diagrama de exones para un gen concreto. Observamos que a medida que nos alejamos evolutivamente en las especies introducidas para efectuar las comparaciones,

Lectura complementaria

J. F. Abril; R. Guigó; T. Wiehe (2003). "gff2aplot: Plotting sequence comparisons". *Bioinformatics* (núm. 19, págs. 2477-2479).

el área de mayor conservación coincide con precisión sobre los exones del gen (los intrones acumulan de forma natural más mutaciones dado que no codifican la secuencia de la proteína).

Figura 21. Comparación de la estructura exónica del gen humano *FOS* en otras especies.



Estas imágenes son cortesía del Dr. Josep F. Abril (Universidad de Barcelona).

6. Alineamientos óptimos de secuencias

Es posible efectuar miles de alineamientos posibles entre dos secuencias de caracteres. No obstante, una vez establecido un esquema de puntuaciones con cierto sentido biológico, resulta fundamental averiguar en un tiempo razonable cuál es el alineamiento óptimo que obtiene precisamente la máxima puntuación. De la bondad de este alineamiento resultante dependerá que podamos inferir correctamente información útil para nuestras investigaciones. Para llevar a cabo la búsqueda del mejor alineamiento –descartando el resto de combinaciones–, es imprescindible emplear la potencia de cálculo de los ordenadores. Es importante reseñar que vamos a estudiar varios algoritmos que en un reducido espacio de tiempo calculan la solución óptima desde un punto de vista algorítmico. Estas soluciones serán biológicamente aceptables cuando incorporem un esquema adecuado de recompensas y penalizaciones acorde con diferentes eventos evolutivos.

La explosión combinatorial provoca que no sea factible generar todos los alineamientos posibles para posteriormente ser puntuados y clasificados en función de su calidad. En lugar de explotar la fuerza bruta para analizar todas las combinaciones, generalmente haremos uso de una técnica algorítmica recursiva denominada **programación dinámica** para garantizar que efectuamos el número de comparaciones estrictamente necesario durante el proceso de identificación de la mejor solución posible. La programación dinámica funciona sólo cuando es factible solucionar el problema original descomponiéndolo recursivamente en pequeños subproblemas más simples que pueden aproximarse con más facilidad.

El **principio de optimalidad** necesario para aplicar **programación dinámica** exige que la solución a un cierto problema de optimización coincida con la combinación de las soluciones óptimas para problemas más reducidos derivados del original.

Como la solución óptima al problema del alineamiento de dos secuencias se puede formular en función de la combinación de las soluciones óptimas obtenidas para alinear fragmentos de éstas, el alineamiento global puede aproximarse mediante programación dinámica. Dado que es posible resolver con facilidad cuál es el mejor alineamiento para secuencias de tamaño reducido, resultará sencillo extrapolar con posterioridad cuál es el alineamiento óptimo de las secuencias completas.

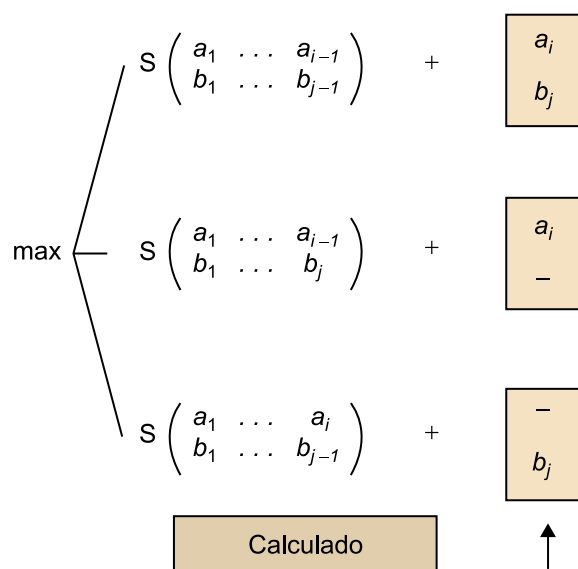
Lectura complementaria

S. R. Eddy (2004). "What is dynamic programming?". *Nature Biotechnology* (núm. 22, págs. 909-910).

Existen varias formulaciones para modelar este problema con programación dinámica. Optaremos por emplear dos estructuras de datos auxiliares: la función de puntuación s y la matriz de similaridad S . La clave de todo el proceso consiste en descomponer recursivamente cada secuencia en una serie de prefijos más simples, desde el más complejo que contiene la secuencia completa, hasta el más elemental que incluye el primer carácter de ésta. En consecuencia, podemos reformular la similaridad S del alineamiento óptimo entre dos secuencias A y B en función del valor de los mejores alineamientos entre los prefijos de éstas con un carácter menos, combinado con el alineamiento efectivo del último residuo (figura 22).

Figura 22. Recurrencias de programación dinámica: alineamiento global.

$$S(i, j) = \max \begin{cases} S(i-1, j-1) + s(a_i, b_j) & \text{Coincidencia} \\ S(i-1, j) + s(a_i, -) & \text{Deleción en B} \\ S(i, j-1) + s(-, b_j) & \text{Deleción en A} \end{cases}$$



El caso más simple, que puede solucionarse de forma trivial, ocurre cuando alineamos dos prefijos de un sólo carácter en cada secuencia. A partir de ese momento, asumiendo que conocemos el valor del mejor alineamiento formado por los tres posibles prefijos anteriores, podemos reconstruir el resto del alineamiento óptimo añadiendo el último carácter analizado. En este punto del proceso debemos valorar si resulta más beneficioso para el alineamiento actual introducir una coincidencia o un *gap* en esa posición.

Para evitar la repetición de cálculos, es necesario realizar todas estas operaciones en un determinado orden (definido por la generación de los prefijos), guardando simultáneamente los resultados en una estructura matricial visible durante todo el procedimiento. Como se puede observar en la figura 23, situaremos la primera secuencia A de m símbolos en el eje horizontal y la segunda secuencia B de n símbolos en el eje vertical. Para dotar de contenido la matriz de programación dinámica es preciso realizar un barrido en diagonal desde la

Lecturas complementarias

S. B. Needleman; C. D. Wunsch (1970). "A general method to search for similarities in the amino acid sequence of two proteins". *Journal of molecular biology* (núm. 48, págs. 443-453).

P. Sellers (1974). "On the theory and computation of evolutionary distances". *SIAM Journal of applied Mathematics* (núm. 26, págs. 787-793).

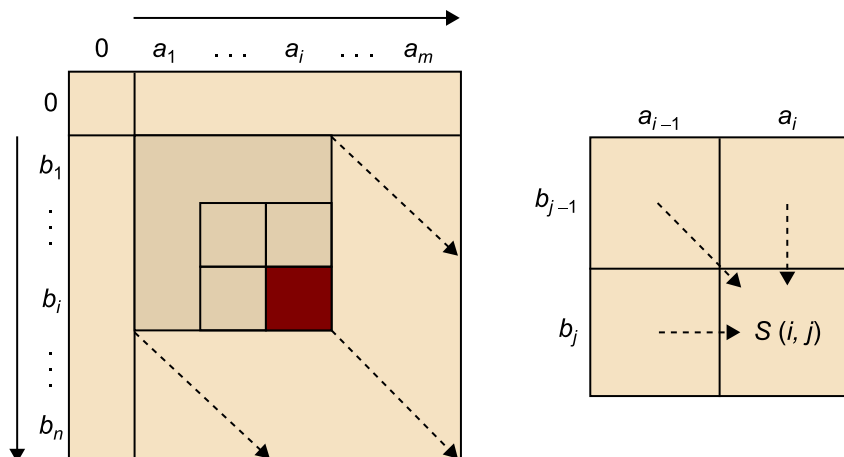
T. F. Smith; M. S. Waterman; W. M. Fitch (1981). "Comparative biosequence metrics". *Journal of Molecular Evolution* (núm. 18, págs. 38-46).

Barridos

Es posible efectuar barridos en horizontal o vertical de la matriz siempre que respetemos el orden establecido para los cálculos de las posiciones de ésta.

esquina superior izquierda (posición $S(0,0)$) hasta la esquina inferior derecha (posición $S(m,n)$). Durante este proceso, cada posición en esta matriz $S(i,j)$ obtiene su valor a partir de las tres posiciones adyacentes $S(i-1, j-1)$, $S(i-1, j)$ o $S(i, j-1)$, cuyo contenido había sido calculado anteriormente aplicando la recurrencia mostrada en la figura 22.

Figura 23. Representación gráfica de la matriz de programación dinámica.



Observamos que es preciso añadir arbitrariamente una fila y una columna (denotadas con el índice 0) para modelar los posibles *gaps* que puedan ser necesarios al principio o al final de alguna de estas secuencias. Estas posiciones deben inicializarse previamente multiplicando la penalización por introducir estos huecos por el número de *gaps* utilizados. Una vez obtenidos estos valores, el algoritmo procede efectivamente a iniciar el barrido de la matriz:

Figura 24. Recurrencias iniciales de programación dinámica.

$$S(i,0) = \sum_{k=1}^{i-1} s(a_k, -)$$

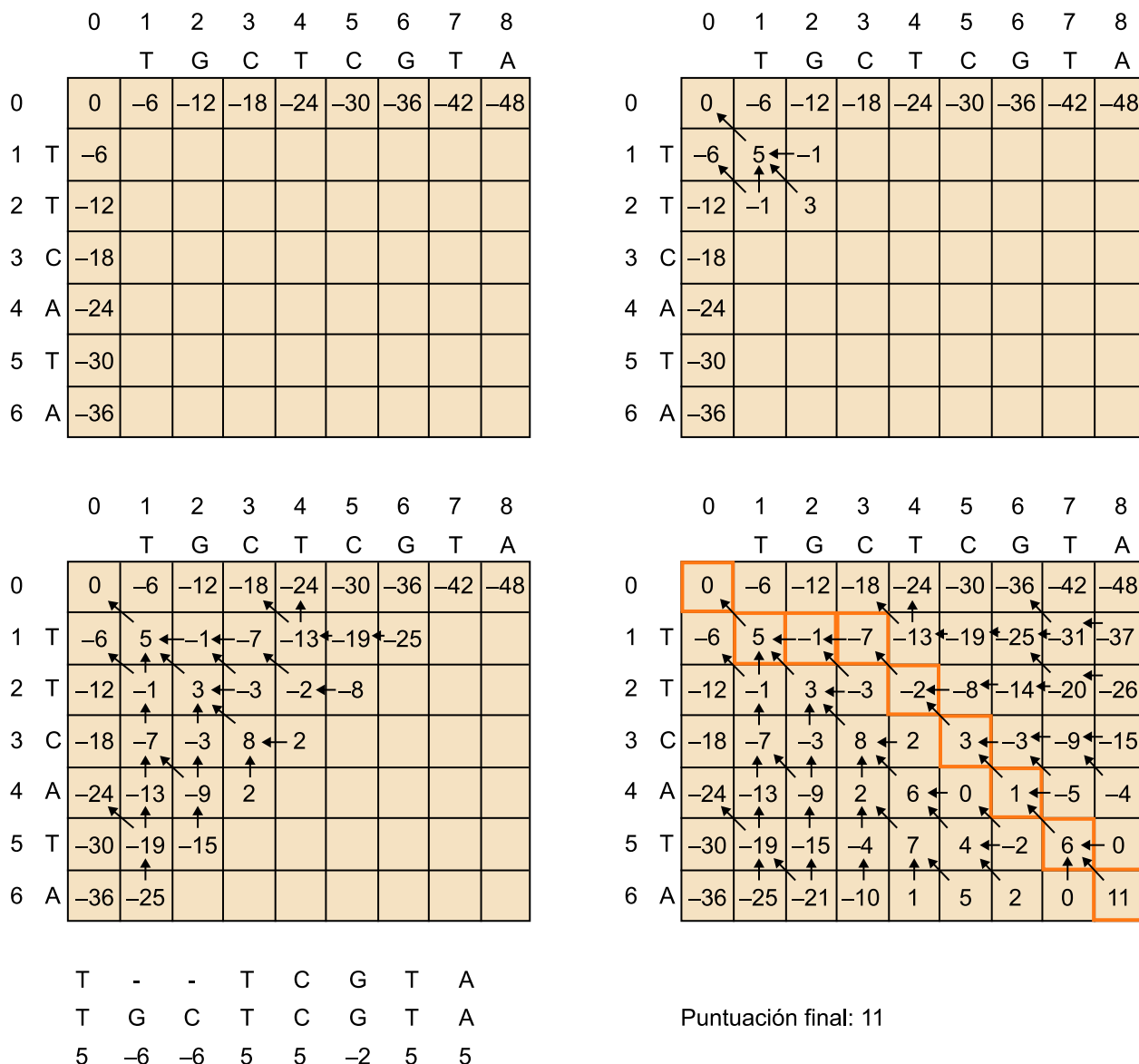
$$S(0,j) = \sum_{k=1}^{j-1} s(-, b_k)$$

Una vez finalizado el barrido de la matriz, la posición $S(m,n)$ contiene el valor del alineamiento global óptimo entre las dos secuencias A y B. Para recuperar exactamente las correspondencias entre los caracteres de ambas secuencias, es necesario llevar un registro del anterior prefijo utilizado para calcular el mejor alineamiento de la actual pareja de secuencias. De este modo, partiendo desde la posición $S(m,n)$, debemos desandar este camino en la matriz para reconstruir el alineamiento óptimo. En la figura 25 mostramos paso a paso el procedimiento de llenado de la matriz de programación dinámica aplicado sobre las secuencias TGCTCGTA y TTCATA.

Empate entre posiciones

En caso de existir conflicto por empate entre varias posiciones, puede reconstruirse arbitrariamente cualquiera de los posibles alineamientos óptimos.

Figura 25. Alineamiento global con programación dinámica paso a paso.

**Leyenda figura 25**

El esquema de puntuaciones empleado en este ejemplo es el siguiente:

- Coincidencia = +5.
- Sustitución = -2.
- Gap = -6.

Lectura complementaria

S. R. Eddy (2004). "What is dynamic programming?". *Nature Biotechnology* (núm. 22, págs. 909-910).

En la figura 26 se puede apreciar el código del algoritmo de alineamiento global de dos secuencias, popularmente conocido como Needleman y Wunsch. Este algoritmo utiliza una matriz S para registrar los valores de los alineamientos óptimos que finalizan en cada prefijo de las secuencias de entrada. Opcionalmente, una segunda matriz P registra el origen del valor previo seleccionado, entre los tres posibles vecinos, para calcular cada valor de S. Inicialmente, debe llenarse la primera fila y la primera columna con la penalización asociada a la

introducción de huecos en cualquier parte del alineamiento. Posteriormente, realizamos el barrido de la matriz por filas y columnas (emulando el recorrido en diagonal mostrado en la figura 23).

Cada posición de la matriz S contiene el valor del alineamiento óptimo entre esos dos prefijos de las secuencias originales. Para calcular ese valor es necesario seleccionar el máximo entre las tres alternativas posibles a cada paso (coincidencia o cualquiera de las dos inserciones/deleciones). Las coordenadas del valor anteriormente empleado para el cálculo de cada posición son almacenadas en la matriz P . Una vez finalizado el procesamiento completo de la matriz S , el valor del mejor alineamiento global entre las dos secuencias A y B queda almacenado en la posición $S(|A|, |B|)$. Para reconstruir el alineamiento es necesario extraer recursivamente el contenido de la matriz P , iniciando el recorrido desde esa última posición $P(m, n)$. Este algoritmo posee un coste de orden cuadrático, dado que precisa realizar tantas operaciones como posiciones contiene cualquiera de las dos matrices S y P . Si ambas secuencias poseen n caracteres, podemos afirmar que el coste asintótico está acotado superiormente por la función $O(n^2)$.

Nota

Para simplificar el código hemos introducido la penalización de los *gaps* dentro de la propia matriz de sustitución s .

Lectura complementaria

M. Zvelebil; J. O. Baum (2008). *Understanding bioinformatics*. Londres: Garland Science. ISBN: 0815340249.

Figura 26. Algoritmo de alineamiento óptimo global de dos secuencias.

```

PRE ≡ {A, B: secuencias; s: matriz de substitucion}
POST ≡ {S(|A|, |B|) es el valor del mejor alineamiento}
(* Inicializar la fila 0 y la columna 0 *)
S(0,0) ← 0;
para (i=1 hasta |A|) hacer
    S(i,0) ← i × s(ai, -);
fpara
para (j=1 hasta |B|) hacer
    S(0,j) ← j × s(-, bj);
fpara
(* Barrido de la matriz *)
para (i=1 hasta |A|) hacer
    para (j=1 hasta |B|) hacer
        (* A. Coincidencia *)
        max ← S(i-1, j-1) + s(ai, bj);
        P(i, j) ← (i-1, j-1);
        (* B. Gap en la secuencia B *)
        valor ← S(i-1, j) + s(ai, -);
        si (valor > max) entonces
            max ← valor;
            P(i, j) ← (i-1, j);
        fsi
        (* C. Gap en la secuencia A *)
        valor ← S(i, j-1) + s(-, bj);
        si (valor > max) entonces
            max ← valor;
            P(i, j) ← (i, j-1);
        fsi
        S(i, j) ← max;
    fpara
fpara

```

Varios estudios teóricos han intentado aproximar la programación dinámica original hacia una concepción más biológica. La recurrencia básica penaliza del mismo modo tanto un evento de inserción/delección de n símbolos como n eventos de inserción/delección de un único símbolo (figura 26). A pesar de que el resultado final es el mismo, esta simplificación es poco realista dado que el número final de eventos genéticos es distinto en cada caso. De hecho, es posible reformular este coste para penalizar levemente los bloques de *gaps* contiguos. Definimos la función $g(k)$, que penaliza la introducción de k huecos, de modo que $g(k) \leq kg(1)$. Para ello, generalizamos la recurrencia básica de programación dinámica, permitiendo ahora que el cálculo de la solución óptima de cualquier prefijo de las secuencias evalúe todos los elementos de la fila y la columna anteriores (ver figura 27).

Figura 27. Recurrencias de programación dinámica: bloques de *gaps*.

$$S(i, j) = \max \begin{cases} S(i-1, j-1) + s(a_i, b_j) & \text{Coincidencia} \\ \max_{1 \leq k \leq i} \{S(i-k, j) + g(k)\} & \text{Gap de tamaño } k \text{ en } B \\ \max_{1 \leq k \leq j} \{S(i, k-j) + g(k)\} & \text{Gap de tamaño } k \text{ en } A \end{cases}$$

$$S(k, 0) = g(k)$$

$$S(0, k) = g(k)$$

Lógicamente, esta generalización repercutirá sobre el número de operaciones efectuadas por el algoritmo resultante que ahora posee un coste cúbico $O(n^3)$. Este incremento en la complejidad del programa puede producir en la práctica un tiempo de ejecución prohibitivo. Podemos simplificar este esquema de puntuaciones modelando un bloque de *gaps* con dos componentes independientes: abrir el hueco inicial o extender el actual. El modelo lineal de *gaps* afines asigna una penalización inicial a a la iniciación del bloque y otra de menor cuantía por su extensión b (directamente proporcional a su tamaño). Podemos calcular la penalización por introducir un nuevo hueco en un bloque de k *gaps* empleando la fórmula mostrada en la figura 28. La aplicación de este esquema de puntuación en la recurrencia de programación dinámica posee un coste cuadrático.

Figura 28. Recurrencias de programación dinámica: modelo de *gaps* afines.

$$g(k) = \begin{cases} a & \text{if } k = 1 \\ a + bk & \text{if } k > 1 \end{cases} \text{ donde } a, b \geq 0$$

$$g(k+1) = a + b(k+1) = a + bk + b = g(k) + b$$

Lectura complementaria

M. S. Waterman; T. F. Smith; W. A. Beyer (1976). "Some biological sequence metrics". *Advances in Mathematics* (núm. 20, págs. 367-387).

Recurrencia básica

Con la recurrencia básica, obteníamos que $g(k) = kg(1)$.

Lectura complementaria

O. Gotoh (1982). "An improved algorithm for matching biological sequences". *Journal of Molecular Biology* (núm. 162, págs. 705-708).

El esquema afín de penalizaciones resulta sensato desde una óptica más biológica, aunque todavía muestra algunas debilidades en este aspecto. Por ejemplo, una vez abierto un primer *gap*, la penalización aplicada para el resto de espacios es la misma. A partir de observaciones empíricas llevadas a cabo sobre bloques de inserciones/delecciones en proteínas homólogas se ha determinado, en cambio, que quizás esta segunda penalización debería atenuarse a medida que el bloque de huecos es mayor, dado que, biológicamente, es bastante factible la inserción de otro *gap*. El modelo cóncavo de puntuaciones introduce un factor logarítmico para reducir esta penalización a medida que encontramos más *gaps* en una cierta región del alineamiento. La integración de esta solución en la recurrencia básica aumenta sensiblemente su coste, acotándose por la función $O(n^2 \log(n))$.

Figura 29. Recurrencias de programación dinámica: modelo cóncavo de *gaps*.

$$g(k) = \begin{cases} a & \text{if } k = 1 \\ a + b \log(k) & \text{if } k > 1 \end{cases} \quad \text{donde } a, b \geq 0$$

$$g(k+1) - g(k) \leq g(k) - g(k-1)$$

Para calcular los alineamientos locales óptimos entre dos secuencias hemos de modificar ligeramente la recurrencia básica definida en la figura 22. El alineamiento local debe reportar únicamente aquellos segmentos de las secuencias que presentan un alto grado de similitud, ignorando el resto del contenido de éstas. La reconstrucción de dicho alineamiento local debe comenzar necesariamente por la identificación de la parte final de éste. Para ello, es necesario registrar en el interior de la matriz *S* la solución que presenta el máximo valor de similitud. Una vez delimitado este punto, debemos iniciar la recuperación del alineamiento local óptimo que termina precisamente en dicha posición. Este procedimiento consiste en desplazarse hacia atrás por la matriz *S* (empleando la matriz de punteros *P* para desandar el camino), ensamblando simultáneamente el alineamiento parcial. Este recorrido nos conducirá a un cierto punto (que coincide con la parte izquierda de este alineamiento local) donde la calidad de la solución descenderá más allá de un límite preestablecido.

Para identificar fácilmente este umbral a partir del cual ya no es útil continuar añadiendo caracteres a la solución local, basta con añadir una nueva regla en la recurrencia de programación dinámica. En consecuencia, cuando el valor del alineamiento óptimo en una determinada posición caiga por debajo de un cierto límite (habitualmente inferior a 0), escribiremos en dicha posición una marca especial de final de alineamiento local.

Lectura complementaria

M. S. Waterman (1984). "Efficient sequence alignment algorithms". *Journal of Theoretical Biology* (núm. 108, págs. 333-337).

Lectura complementaria

T. F. Smith; M. S. Waterman (1981). "Identification of common molecular subsequences". *Journal of Molecular Biology* (núm. 147, págs. 195-197).

Figura 30. Recurrencias de programación dinámica: alineamiento local.

$$S(i, j) = \max \begin{cases} S(i-1, j-1) + s(a_i, b_j) & \text{Coincidencia} \\ S(i-1, j) + s(a_i, -) & \text{Gap en B} \\ S(i, j-1) + s(-, b_j) & \text{Gap en A} \\ 0 & \text{Final del segmento} \end{cases}$$

$$S(i, 0) = 0$$

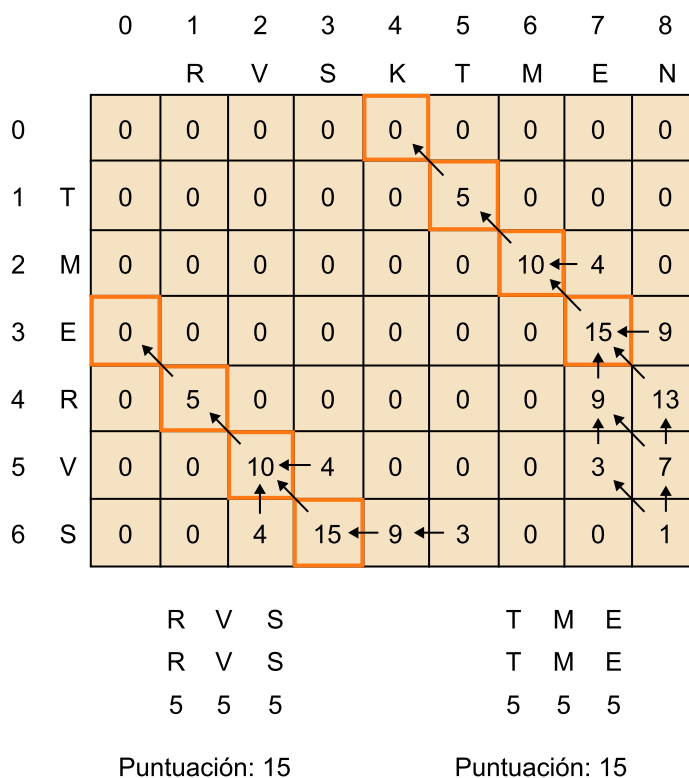
$$S(0, j) = 0$$

Como se puede apreciar en el alineamiento local de dos secuencias mostrado en la figura 31, el procedimiento de barrido de la matriz S para calcular los valores óptimos funciona de un modo similar al caso global. Durante este proceso, sin embargo, es preciso registrar dinámicamente las posiciones dentro de la matriz que obtienen los valores más altos. Una vez finalizado el llenado de la matriz S , para obtener el mejor alineamiento local óptimo entre las dos secuencias es preciso reconstruir el segmento local de ambas que finaliza precisamente en la posición de S que obtuvo la mayor puntuación. Esta reconstrucción termina en el momento en que alcancemos una posición que posea el valor umbral de marcaje (0 en este caso). Si el usuario desea obtener una lista con los mejores segmentos, debe repetir este procedimiento empleando el resto de las posiciones que exhiben los valores más altos. En el ejemplo de la figura 31, hemos recuperado dos segmentos locales de máxima similitud. Se puede apreciar que se ha invertido el orden en que estos aparecen dentro de cada secuencia, obteniéndose en este caso un alineamiento local no colineal.

Evitar segmentos solapantes

Para evitar reproducir segmentos solapantes podemos ocultar aquellas posiciones de la matriz S que ya hemos utilizado previamente en anteriores fragmentos.

Figura 31. Alineamiento local de dos secuencias.

**Leyenda figura 31**

El esquema de puntuaciones empleado en este ejemplo es el siguiente:

- Coincidencia = +5.
- Sustitución = -2.
- Gap = -6.

En la figura 32 presentamos el algoritmo para obtener el alineamiento local óptimo entre dos secuencias, popularmente conocido como Smith y Waterman. Esta variante emplea nuevamente una matriz S para almacenar los valores de los alineamientos óptimos que involucran cada prefijo de las secuencias, y una segunda matriz P para registrar su origen. Inicialmente, debe marcarse la primera fila y la primera columna con el símbolo seleccionado para indicar la terminación del alineamiento local. Para simplificar el código hemos introducido la penalización de los *gaps* dentro de la propia matriz de sustitución s. Posteriormente, realizamos el barrido de la matriz por columnas utilizando la recurrencia indicada en la figura 30 para calcular el valor óptimo del alineamiento de cada segmento.

Es importante mantener en memoria cuál es la posición (iMax,jMax) que obtuvo la máxima puntuación para recuperar después el segmento local de máxima similitud. Para reconstruir este alineamiento es suficiente con extraer recursivamente el contenido de la matriz P, iniciando el recorrido precisamente desde aquella posición. Para la producción de los mejores *n* alineamientos locales, podríamos gestionar una lista ordenada de las posiciones más prometedoras, repitiendo nuevamente la reconstrucción del siguiente alineamiento en cada caso. Dado que debemos acceder a todas las posiciones de las matrices S y P, este algoritmo realiza asintóticamente un número de operaciones de orden cuadrático ($O(n^2)$). En este sentido, la sutil modificación propuesta por Smith y Waterman no implica un incremento del tiempo de ejecución en comparación con el esquema global.

Lectura complementaria

T. F. Smith; M. S. Waterman (1981). "Identification of common molecular subsequences". *Journal of Molecular Biology* (núm. 147, págs. 195-197).

Figura 32. Algoritmo de alineamiento óptimo local de dos secuencias.

```

PRE  $\equiv \{A, B: \text{secuencias}; s: \text{matriz de substitucion}\}$ 
POST  $\equiv \{S(iMax, jMax) \text{ es el mejor alineamiento local}\}$ 
(* Inicializar la fila 0 y la columna 0 *)
 $S(0, 0) \leftarrow 0;$ 
para ( $i=1$  hasta  $|A|$ ) hacer
     $S(i, 0) \leftarrow 0;$ 
fpara
para ( $j=1$  hasta  $|B|$ ) hacer
     $S(0, j) \leftarrow 0;$ 
fpara
 $iMax \leftarrow 0;$ 
 $jMax \leftarrow 0;$ 
(* Barrido de la matriz *)
para ( $i=1$  hasta  $|A|$ ) hacer
    para ( $j=1$  hasta  $|B|$ ) hacer
        (* A. Terminacion del segmento *)
         $max \leftarrow 0;$ 
         $P(i, j) \leftarrow (0, 0);$ 
        (* B. Coincidencia *)
         $max \leftarrow S(i-1, j-1) + s(a_i, b_j);$ 
         $P(i, j) \leftarrow (i-1, j-1);$ 
        (* C. Gap en la secuencia B *)
         $valor \leftarrow S(i-1, j) + s(a_i, -);$ 
        si ( $valor > max$ ) entonces
             $max \leftarrow valor;$ 
             $P(i, j) \leftarrow (i-1, j);$ 
        fsi
        (* D. Gap en la secuencia A *)
         $valor \leftarrow S(i, j-1) + s(-, b_j);$ 
        si ( $valor > max$ ) entonces
             $max \leftarrow valor;$ 
             $P(i, j) \leftarrow (i, j-1);$ 
        fsi
         $S(i, j) \leftarrow max;$ 
        (* E. Registro del maximo absoluto *)
        si ( $max > S(iMax, jMax)$ ) entonces
             $iMax \leftarrow i;$ 
             $jMax \leftarrow j;$ 
        fsi
    fpara
fpara

```

7. Alineamientos progresivos de secuencias

Un alineamiento global múltiple organiza una correspondencia entre los caracteres de un conjunto de secuencias para maximizar el número de coincidencias en cada columna. Las variaciones en el interior de una columna pueden ser útiles para predecir las mutaciones puntuales ocurridas a lo largo de la historia evolutiva de una cierta secuencia, revelando el grado de conservación o divergencia en diferentes regiones (ver figura 16). Formalmente, sean $S_1 \dots S_k$ un conjunto de secuencias donde los símbolos de cada cadena $S_i = s_{i,1} \dots s_{i,|S_i|}$ pertenecen a un cierto alfabeto Σ de nucleótidos o aminoácidos. Definimos entonces el conjunto extendido de secuencias $S_1^* \dots S_k^*$ donde los símbolos de cada cadena $S_i^* = s_{i,1}^* \dots s_{i,|S_i^*|}^*$ pertenecen ahora al mismo alfabeto extendido con un *gap*, $\Sigma \cup \{-\}$. El alineamiento múltiple de las secuencias $S_1 \dots S_k$ consiste en una correspondencia C entre las secuencias extendidas representada como una matriz rectangular que necesariamente debe cumplir determinadas propiedades:

Figura 33. Definición formal de un alineamiento global múltiple.

$$C = \begin{pmatrix} s_{1,1}^* & s_{1,2}^* & \dots & s_{1,c}^* \\ s_{2,1}^* & s_{2,2}^* & \dots & s_{2,c}^* \\ \vdots & \vdots & \ddots & \vdots \\ s_{k,1}^* & s_{k,2}^* & \dots & s_{k,c}^* \end{pmatrix}$$

1. La longitud del alineamiento es exactamente c símbolos.
2. Si omitimos los *gaps*, recuperamos las secuencias originales.
3. Como mínimo un elemento de cada columna es diferente de un *gap*.

Cada columna de esta matriz $C(i) = (s_{1,i}^*, s_{2,i}^*, \dots, s_{k,i}^*)$ representa la coincidencia múltiple entre los respectivos símbolos de las secuencias originales junto con algún hueco dispuesto entre éstos. Lógicamente, para evaluar un alineamiento múltiple debemos calcular la suma de las puntuaciones obtenidas por las coincidencias de cada columna $C(i)$. Para puntuar una determinada columna del alineamiento múltiple, habitualmente se calcula un promedio de todos los posibles cambios de dos aminoácidos observados en ésta. Con esta aproximación, denominada suma de pares (en inglés, *sum of pairs*), no es necesario construir matrices multidimensionales de sustitución para puntuar k símbolos.

Suma de pares

Gracias a la técnica de la suma de pares, para evaluar una columna de n símbolos un programa realizará n^2 accesos a la respectiva matriz PAM o BLOSUM. Una aproximación por fuerza bruta necesitaría acceder a una matriz de 2^k combinaciones con un coste exponencial.

Aunque parezca el modo más simple de obtener el alineamiento múltiple de k secuencias, no es recomendable en este caso generalizar el mismo esquema recursivo empleado para alinear dos secuencias por su enorme coste computacional. Con programación dinámica sería necesario habilitar una matriz de similaridad de k dimensiones, precisando de un barrido de coste exponencial $O(n^k)$, que resulta difícilmente asumible para alinear más de cuatro o cinco secuencias. Una gran variedad de estrategias han sido propuestas para producir soluciones con un menor coste que resulten razonables en términos biológicos. De todas ellas, el alineamiento progresivo es uno de los métodos más efectivos, por lo que es ampliamente utilizado por numerosas aplicaciones bioinformáticas.

El **alineamiento progresivo** reconstruye incrementalmente el mejor alineamiento múltiple posible utilizando un árbol filogenético para guiar en qué orden deben incorporarse las restantes secuencias al resultado final.

En lugar de alinear simultáneamente todas las secuencias, el método progresivo emplea las dos secuencias más similares para constituir el germen del alineamiento múltiple. Progresivamente, el resto de secuencias se incorporan para refinar el resultado final con más información. Una vez identificadas cuáles son estas dos secuencias, en el siguiente paso debe decidirse si es más efectivo integrar una nueva secuencia dentro del alineamiento en curso, o crear otro alineamiento con otras dos secuencias más afines. Si fuera necesario, con posterioridad, es posible fusionar dos alineamientos parciales para generar un nuevo alineamiento múltiple más completo o, alternativamente, incorporar otras secuencias no utilizadas todavía a alguno de los alineamientos múltiples existentes en curso.

Lectura complementaria

H. Carrillo; D. Lipman (1988). "The multiple sequence alignment problem in biology". *SIAM Journal of Applied Mathematics* (núm. 48, págs. 1073-1082).

Para construir el alineamiento múltiple final es fundamental, por consiguiente, realizar alineamientos múltiples parciales entre dos subgrupos de secuencias en función de su mayor parecido. El orden en que serán seleccionadas las secuencias para efectuar alineamientos parciales define una determinada jerarquía entre éstas (por ello esta técnica también recibe el nombre de agrupación jerárquica o *hierarchical clustering* en inglés). Para guiar este procedimiento puede utilizarse un árbol filogenético prefijado inicialmente a partir del grado de similitud entre todas las secuencias con algún método de minimización de la suma total de las longitudes de cada rama del árbol. También es posible construir este árbol sobre la marcha, decidiendo dinámicamente cuál es el siguiente par de elementos más similares que deben ser alineados (secuencias o alineamientos parciales). En todo caso, como veremos a continuación, será necesario modificar las rutinas básicas de programación dinámica para acomodar alineamientos entre dos secuencias, alineamientos entre una secuencia y un alineamiento parcial, y alineamientos entre dos alineamientos previos.

Esta técnica no garantiza la obtención de la solución óptima. No obstante, mediante el uso cuidadoso de las matrices de sustitución apropiadas, modificando en tiempo de ejecución el esquema de penalizaciones de los *gaps* e introduciendo otro tipo de información (por ejemplo, arquitectura estructural, propiedades físico-químicas de los aminoácidos, etc.), los alineamientos resultan suficientemente aceptables en términos biológicos para comparar secuencias homólogas con garantías. En este sentido, es importante subrayar que el bioinformático debe ser consciente de que el resultado final depende significativamente del primer alineamiento escogido como referencia para incorporar el resto de las secuencias. Dicho alineamiento no necesariamente es representativo de la conservación global existente. Para amortiguar el impacto de cada elección pueden introducirse pesos que ponderan la importancia de cada secuencia. De este modo, también es posible corregir numéricamente un posible muestreo evolutivo desigual cuando deseamos alinear secuencias con distinto grado de conservación. Incrementando el peso de las secuencias que presentan más divergencias, podemos equilibrar el alineamiento reduciendo la relevancia de aquellas más similares. Gracias a la aplicación de numerosas reglas heurísticas, es posible obtener soluciones aceptables con este método en un reducido periodo de tiempo.

Lecturas complementarias

D. Fena; R. F. Doolittle (1987). "Progressive sequence alignment as a prerequisite to correct phylogenetic trees". *Journal of Molecular Evolution* (núm. 25, págs. 351-360).

N. Saitou; M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees". *Molecular Biology and Evolution* (núm. 4, págs. 406-425).

J. D. Thompson; D. G. Higgins; T. J. Gibson (1994). "CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic Acids Research* (núm. 22, págs. 4673-4680).

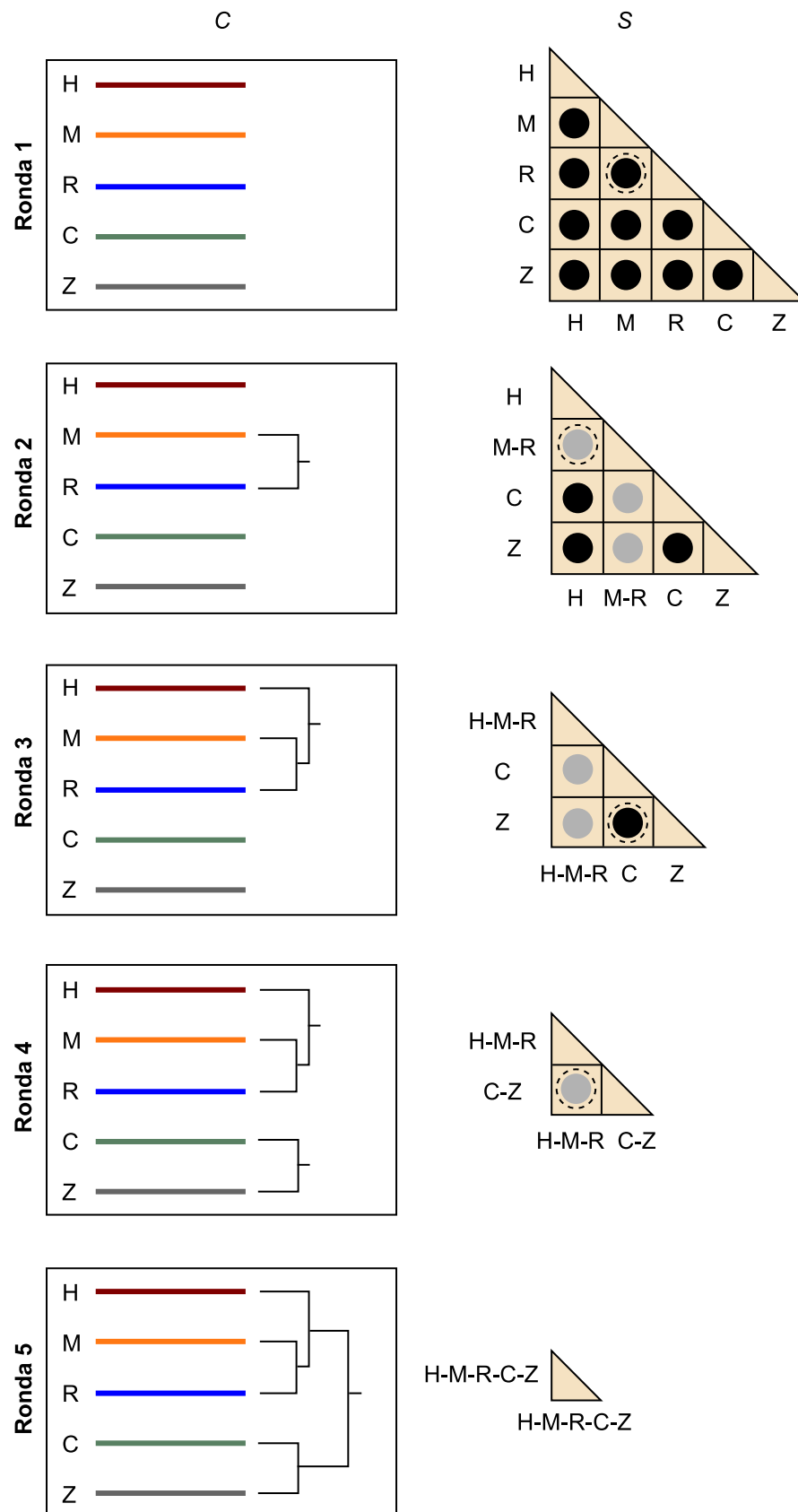
Ved también

Los métodos más extendidos de construcción de árboles filogenéticos se explican con detalle en la asignatura *Fundamentos de biología molecular*.

Lectura complementaria

J. D. Thompson; D. G. Higgins; T. J. Gibson (1994). "CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic Acids Research* (núm. 22, págs. 4673-4680).

Figura 34. Construcción del árbol guía del alineamiento progresivo.

**Leyenda figura 34**

Consideramos en este ejemplo el uso de cinco secuencias homólogas pertenecientes a humano (H), ratón (M), rata (R), pollo (C) y pez cebra (Z). C contiene la lista de secuencias empleadas y alineamientos parciales generados. S registra la similitud entre dos alineamientos/secuencias existentes. Dentro de S, utilizamos el color negro para reflejar los valores realmente calculados y el color gris para aquellos que han sido estimados. Con una línea discontinua indicamos los dos grupos seleccionados en cada ronda para fusionarse.

Lectura complementaria

S. Batzoglou (2005). "The many faces of sequence alignment". *Briefings in bioinformatics* (núm. 6, págs. 6-22).

En estos materiales vamos a estudiar con detalle una implementación específica del esquema progresivo que construye dinámicamente el árbol guía a medida que vamos añadiendo más secuencias a los alineamientos existentes

(ver figura 34). En este contexto, definimos un grupo como el alineamiento múltiple de un subconjunto de secuencias. Para trabajar con esta aproximación necesitamos mantener en memoria una lista –que denominaremos C – de todos los grupos generados. Inicialmente, cada secuencia que debe contribuir al alineamiento múltiple final constituye su propio grupo. Posteriormente, el método progresivo selecciona reiteradamente los dos grupos más similares para realizar su alineamiento y crear un nuevo grupo en el que participan todas estas secuencias. Mediante la aplicación de esta regla, este procedimiento termina cuando reducimos sucesivamente el conjunto inicial de secuencias a un único grupo que albergará el alineamiento final resultante. Para decidir el orden en que agrupamos las secuencias necesitamos definir una matriz de similitud S que almacene en cada momento la similitud entre los grupos existentes. Cuando creamos un nuevo alineamiento entre dos grupos, es necesario recalcular la similitud entre este alineamiento y el resto de elementos de la matriz. Es posible estimar este valor tomando en cuenta la similitud entre los grupos que originalmente dieron lugar a este alineamiento y el resto. Con esta técnica, solamente calcularemos el alineamiento múltiple de dos grupos concretos cuando dicha combinación sea escogida como la mejor posible.

Proponemos una posible implementación de esta estrategia de alineamiento progresivo presentada anteriormente (ver figura 36). Este algoritmo gestiona la ejecución ordenada de todos los pasos detallados anteriormente. En primer lugar cada secuencia es asignada a un nuevo grupo en C , y se calcula después el grado de similitud entre todas las secuencias con una variante de la rutina básica de programación dinámica para alinear dos secuencias que denominaremos AlineamientoOptimo. A partir de estos alineamientos, cuya similitud es almacenada en la matriz S , escogemos las dos secuencias más parecidas para crear el primer alineamiento. Ahora debemos sustituir las secuencias utilizadas en C por el nuevo grupo resultante. Será necesario modificar las posiciones de S donde estas dos secuencias estaban involucradas. Para evitar calcular en cada momento los alineamientos entre este nuevo grupo y el resto, podemos estimar la similitud entre estos con la función `EstimarSimilitud`. Esta rutina básicamente implementa el método denominado WPGMA (de inglés *weighted pair group method with arithmetic mean*, método de agrupación ponderada de pares utilizando media aritmética):

Figura 35. Formulación del método WPGMA.

$$S(C_{i-j}, C_k) = \frac{|i| \cdot S(C_i, C_k) + |j| \cdot S(C_j, C_k)}{|i| + |j|}$$

Una vez el conjunto de grupos ha sido reducido en una unidad, este proceso continua iterativamente hasta que todas las secuencias resultan agrupadas en un solo alineamiento múltiple construido progresivamente. El coste aproximado del algoritmo en términos asintóticos es $O(k^2 n^2)$, dado que cada alineamiento entre dos grupos cualquiera requiere de n^2 pasos y la incorporación

Lectura complementaria

J. D. Thompson; D. G. Higgins; T. J. Gibson (1994). "CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic Acids Research* (núm. 22, págs. 4673-4680).

Nota

Con esta estimación es posible equilibrar la contribución de cada elemento al nuevo grupo en función del número de secuencias alineadas dentro de éste.

progresiva de k secuencias realizará como máximo tantas agrupaciones como combinaciones de dos secuencias sean posibles (k^2 para k secuencias). Este resultado es notablemente menor al coste exponencial $O(n^k)$ esperable en el caso de que recurriéramos a la aproximación directa por programación dinámica.

Figura 36. Algoritmo general de alineamiento múltiple progresivo.

```

PRE  $\equiv \{S_1 \dots S_k$ : secuencias;  $s$ : matriz de substitucion}
POST  $\equiv \{C$  es el mejor alineamiento progresivo}
(* Crear un grupo nuevo con cada secuencia *)
 $C \leftarrow \emptyset$ ;
para ( $i=1$  hasta  $k$ ) hacer
     $C_i \leftarrow S_i$ ;
fpara
    (* Crear la matriz inicial de comparaciones *)
     $iSim \leftarrow 0$ ;
     $jSim \leftarrow 0$ ;
     $maxSim \leftarrow -\infty$ ;
    para ( $i=1$  hasta  $k$ ) hacer
        para ( $j=i+1$  hasta  $k$ ) hacer
             $S(C_i, C_j) \leftarrow \text{AlineamientoOptimo}(C_i, C_j, s)$ ;
            (* Registrar las dos secuencias mas similares *)
            si ( $S(C_i, C_j) > maxSim$ ) entonces
                 $maxSim \leftarrow S(C_i, C_j)$ ;
                 $iSim \leftarrow i$ ;
                 $jSim \leftarrow j$ ;
            fsi
        fpara
    fpara
    (* Crear un nuevo grupo con estas dos secuencias *)
     $C_{iSim-jSim} \leftarrow \text{AlineamientoOptimo}(C_{iSim}, C_{jSim}, s)$ ;
    (* Estimar la similaridad con el resto de grupos *)
     $S \leftarrow \text{EstimarSimilaridad}(C, iSim, jSim)$ ;
    (* Alineamiento progresivo de los grupos restantes *)
    (* Agrupar hasta reducir el numero de grupos *)
mientras ( $|C| > 1$ ) hacer
     $iSim \leftarrow 0$ ;
     $jSim \leftarrow 0$ ;
     $maxSim \leftarrow -\infty$ ;
    para ( $i=1$  hasta  $|C|$ ) hacer
        para ( $j=i+1$  hasta  $|C|$ ) hacer
            (* Registrar los dos grupos mas similares *)
            si ( $S(C_i, C_j) > maxSim$ ) entonces
                 $maxSim \leftarrow S(C_i, C_j)$ ;
                 $iSim \leftarrow i$ ;
                 $jSim \leftarrow j$ ;
            fsi
        fpara
    fpara
    (* Crear un nuevo grupo con estos dos grupos *)
     $C_{iSim-jSim} \leftarrow \text{AlineamientoOptimo}(C_{iSim}, C_{jSim}, s)$ ;
    (* Estimar la similaridad con el resto de grupos *)
     $S \leftarrow \text{EstimarSimilaridad}(C, iSim, jSim)$ ;
fmientras

```

La función AlineamientoOptimo es una variante de la recurrencia básica de programación dinámica para dos secuencias. Resulta particularmente interesante su estudio dado que es necesario generalizar la valoración de las coincidencias para permitir el alineamiento de dos alineamientos parciales. Podemos observar en la figura 37 que la matriz de programación dinámica no cambia excesivamente su disposición habitual (ver figura 23). El primer alineamiento debe colocarse horizontalmente –la comparación entre ratón y rata–, mientras que el segundo será representado verticalmente –en este caso, la proteína humana. Cada alineamiento por separado puede contener *gaps* que deben ser respetados en el alineamiento final. Cada posición de esta matriz contendrá el valor del mejor alineamiento entre los prefijos de las proteínas de ratón y rata junto con la secuencia homóloga humana.

Para proceder a realizar el barrido de la matriz, valoraremos nuevamente tres posibilidades diferentes en cada posición de ambos alineamientos: introducir un hueco en alguno de los dos alineamientos o asignar una coincidencia entre dichas posiciones. En consecuencia, la recurrencia básica presentada anteriormente en la figura 22 resulta igualmente válida en esta nueva variante del problema. Únicamente debe modificarse el modo en que se evalúan las coincidencias entre dos posiciones concretas de cada alineamiento múltiple. Para ello, es necesario calcular el promedio de todas las sustituciones posibles entre los caracteres del primer alineamiento y los caracteres del segundo en las columnas de los respectivos alineamientos (ver figura 37).

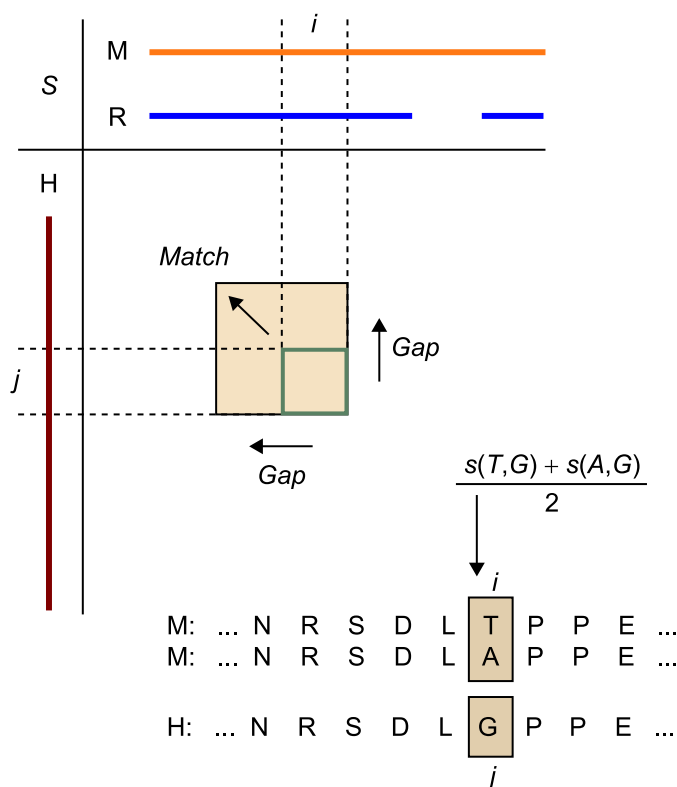
Nota

Los programas modifican diferentes parámetros de esta rutina según los datos de entrada: matrices de sustitución, penalizaciones de huecos, etc.

ClustalW

El programa ClustalW introduce los pesos de cada secuencia en este paso para realizar la ponderación de todas las sustituciones.

Figura 37. Alineamiento con programación dinámica de dos alineamientos múltiples.



Leyenda figura 37

Consideramos en este ejemplo el uso de tres secuencias homólogas pertenecientes a humano (H), ratón (M) y rata (R).

8. Identificación de motivos conservados

En la práctica, la identificación directa de pequeños fragmentos conservados en varias secuencias biológicas por aplicación de las técnicas de programación dinámica convencionales no resulta factible. Sin embargo, dada la relevancia de este problema, han sido presentadas numerosas alternativas en el campo de la localización de motivos regulatorios de la transcripción génica o de la caracterización de subdominios de proteínas.

Para evitar la posible explosión combinatoria, estos programas no realizan ningún tipo de alineamiento local entre los fragmentos de cada secuencia, sino que identifican aquellos motivos comunes mediante refinamientos iterativos de un modelo estadístico que representa el patrón definitivo. En algunas aproximaciones, puede introducirse además información sobre la filogenia de las secuencias para ponderar la relevancia de los motivos detectados en cada secuencia. Para un número razonable de regiones, estos programas pueden identificar una serie de motivos conservados en cuestión de pocos minutos. Aunque no existe garantía de encontrar la solución óptima, los motivos resultantes pueden ser biológicamente plausibles. El método de la maximización de la esperanza o método EM (del inglés *expectation maximization*) es una de las técnicas de reconocimiento de patrones más populares.

El **método EM** estima los parámetros de un modelo probabilista que representa la composición ideal de los motivos conservados en un conjunto de secuencias.

El programa MEME es la herramienta estándar empleada por la comunidad bioinformática para llevar a cabo la búsqueda de patrones cortos conservados en múltiples secuencias biológicas con el método EM. El propósito fundamental de esta aplicación es averiguar cuáles son los parámetros más adecuados de un modelo probabilista para distinguir los motivos del resto de símbolos de cada secuencia (denominado *background* en inglés). Para el correcto funcionamiento del algoritmo debe definirse una función de evaluación que contraste la diferencia entre el contenido del motivo en construcción y el resto de las secuencias (en inglés, función de *fitness*). Refinando iterativamente la representación del patrón, el programa intenta aproximarse a la localización exacta de los motivos conservados. En el instante en que no se observa mejora alguna utilizando esta función de evaluación, el programa debe finalizar y reportar los motivos conservados. En cierto modo, para fomentar el entrenamiento de este tipo de aproximaciones, el modelo final logrado por la técnica EM pue-

Lectura complementaria

T. L. Bailey; C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers". *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology (ISMB)* (págs. 28-36).

Lectura complementaria

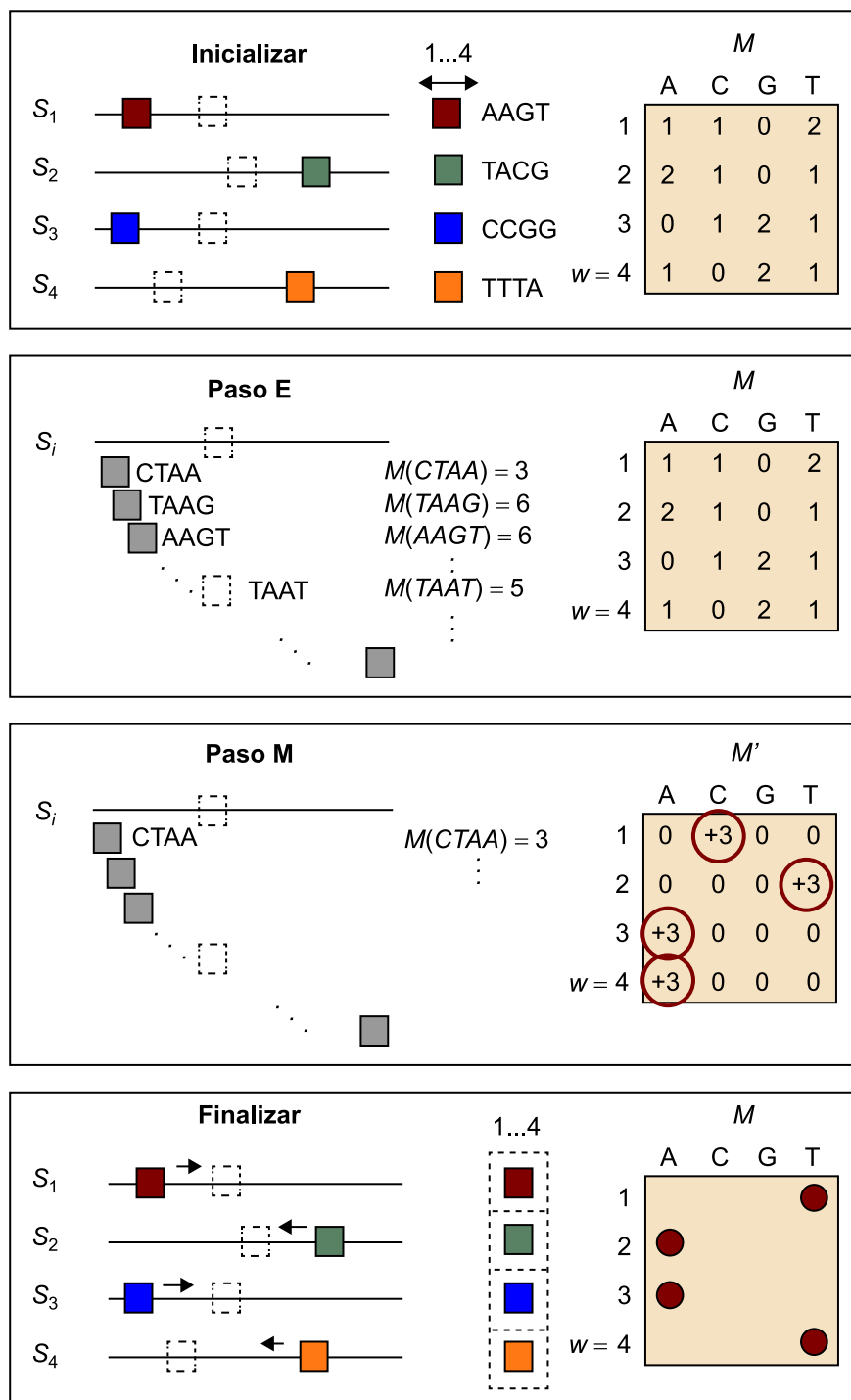
T. L. Bailey; C. Elkan (1994). "MEME Suite: tools for motif discovery and searching". *Nucleic Acids Research* (núm. 37, págs. W202-W208).

de utilizarse para generar aleatoriamente secuencias que poseen en su interior ocurrencias del motivo conservado incrustadas dentro de regiones con la misma distribución observada en este *background*.

En comparación con las técnicas clásicas de alineamiento, la principal ventaja del método EM es su reducido tiempo de ejecución. Existen numerosas opciones que permiten modificar la configuración básica de estos programas. Por ejemplo, podemos instruir al programa MEME para trabajar con nuestras secuencias en distintos escenarios:

- Asumir que algunas secuencias no poseen el motivo conservado en el resto.
- Esperar que el motivo conservado pueda aparecer en más de una ocasión dentro de la misma secuencia.
- Determinar probabilísticamente la longitud para el motivo conservado.
- Establecer a priori la longitud de motivo conservado y el número esperable de motivos.

Figura 39. Funcionamiento del algoritmo EM para identificar motivos.

**Leyenda figura 39**

En este ejemplo, marcamos con líneas discontinuas la ubicación del motivo TAAT conservado en cuatro secuencias genómicas. Indicamos con cajas de colores la apuesta inicial sobre dicha localización. En la parte derecha representamos gráficamente el contenido del motivo conservado durante la progresión del algoritmo (no mostramos explícitamente el *background*).

Nota

Para evitar caer en máximos locales, el algoritmo básico selecciona internamente diferentes puntos de partida.

A continuación, vamos a detallar brevemente los componentes de la variante más sencilla de este método para localizar motivos de w símbolos (ver figura 39). En este caso asumiremos que cada secuencia únicamente posee una ocurrencia de este motivo, conociendo de antemano que este contiene cuatro nucleótidos. Para partir de una representación inicial del motivo, el algoritmo escoge al azar una localización distinta en cada una de las secuencias. Contabilizando la frecuencia de cada nucleótido en cada posición de estas ocurrencias, el programa puede construir una primera versión de la matriz de pesos M que alberga la primera representación del motivo conservado.

Figura 40. Algoritmo genérico EM.

```

PRE  $\equiv \{S = S_1 \dots S_n$ : secuencias;  $T$ : entero $\}$ 
POST  $\equiv \{M$  es el motivo;  $L$  es la lista de sitios $\}$ 
(* Elegir un lugar al azar en cada secuencia *)
(* Construir el modelo inicial con esos sitios *)
InicializarModelo ( $M, S$ );
 $i \leftarrow 1$ ;
convergencia  $\leftarrow$  FALSO;
mientras ( $i \leq \text{MAXITERACIONES}$  y no convergencia) hacer
  para cada ( $S_i$  en  $S$ ) hacer
    (* Paso E: evaluar todas las posiciones *)
    (* de  $w$  símbolos a lo largo de  $S_i$  *)
    Puntos  $\leftarrow$  EvaluarCandidatos ( $S_i, M$ );
    (* Paso M: actualizar el modelo *)
    (* con esas puntuaciones para *)
    (* detectar los motivos frecuentes *)
     $M' \leftarrow$  ActualizarModelo ( $M$ , Puntos);
  fpara
  si ( $\text{calidad}(M') \leq \text{calidad}(M)$ ) entonces
    convergencia  $\leftarrow$  CIERTO;
  sino
     $M \leftarrow M'$ ;
     $i \leftarrow i+1$ ;
  fsi
fmientras
(* Identificar los sitios de este modelo en  $S$  *)
 $L \leftarrow$  Analizar Secuencias ( $S, M, T$ );
retorna ( $L$ )

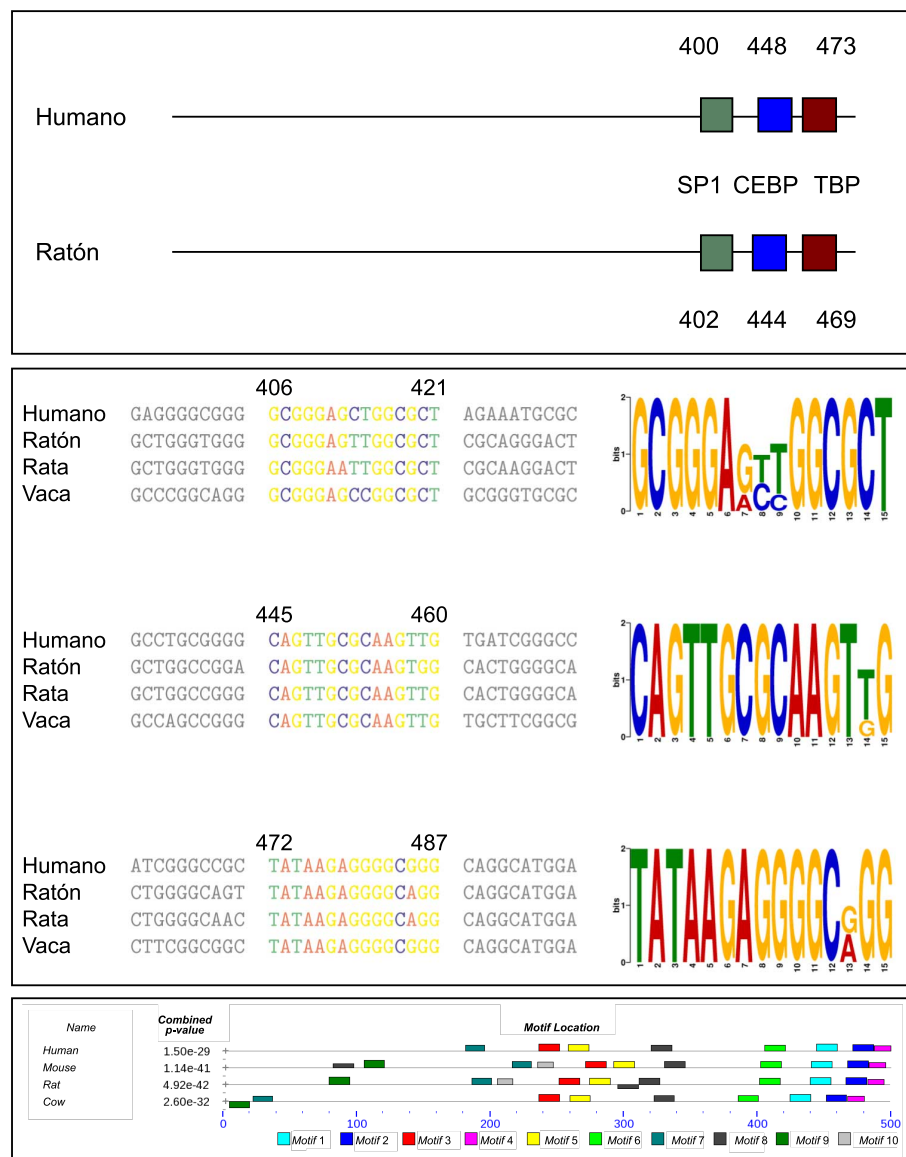
```

Esta primera selección (a no ser que introduzcamos otros criterios) probablemente generará un modelo de motivo con un contenido prácticamente aleatorio. A partir de este momento, el algoritmo EM mostrado en la figura 40 realiza un tratamiento iterativo, basado en la sucesiva aplicación de dos pasos denominados E y M, para refinar el motivo modelado mientras se observa ganancia en la calidad de este:

- Paso E: Utilizar el modelo M para puntuar el parecido con cada candidato de w letras de las secuencias. Para ello debemos buscar en la posición de la matriz M qué valor corresponde al nucleótido existente en una determinada posición de cada posible motivo de la secuencia (por ejemplo, el candidato CTAA recibe la valoración 3 en la figura 39). Esta evaluación produce un valor que asociaremos a cada subsecuencia de w símbolos existente dentro de nuestras secuencias.
- Paso M: Para generar una nueva versión del motivo modelado con una matriz de pesos (que denominaremos M'), procedemos a depositar el valor evaluado para cada candidato precisamente en las posiciones de la nueva matriz que coinciden con su contenido (por ejemplo, el candidato CTAA debe acumular su valoración en las posiciones (1,C), (2,T), (3,A) y (4,A) de M , figura 39). Este procedimiento de actualización de la matriz debe reproducirse con todas las secuencias de nuestro conjunto de datos.

Número máximo de iteraciones

Puede establecerse alternatively un número máximo de iteraciones para efectuar la parada.

Figura 41. Identificación de motivos reguladores de la transcripción del gen *LEP*.**Leyenda figura 41**

El programa MEME fue instruido para identificar los diez motivos más relevantes conservados en estos promotores. Mostramos en primer lugar tres motivos conservados en dos especies que han sido validados experimentalmente y, a continuación, la salida gráfica con todas las predicciones.

Tras cientos de iteraciones, aquellas subcadenas dentro de cada secuencia que contienen el motivo dejarán su huella con mayor intensidad dentro de la matriz *M*. En consecuencia, este modelo mostrará progresivamente los valores más altos en las posiciones que corresponden con la composición del motivo real. Bajo la hipótesis de que el resto de candidatos presentan una distribución irregular de nucleótidos, observaremos con cierta nitidez en la mayoría de ocasiones una diferencia evidente entre el motivo y el *background* (este último no está representado explícitamente en esta versión simplificada del algoritmo EM).

Finalmente, la propia matriz de pesos *M* registrará las posiciones de los candidatos que más semejanza exhiben con este patrón ideal. Podemos observar en la figura 41 el resultado de utilizar el programa MEME sobre las regiones reguladoras del gen humano *LEP* en comparación con su homólogo en el ratón, la rata y la vaca (longitud del motivo prefijada entre 5 y 15 pares de bases). Este programa identifica correctamente los tres sitios de unión a factores de

transcripción conservados a lo largo de la evolución en estas cuatro especies. Se puede apreciar la diferencia entre cada motivo coloreado y las secuencias flanqueantes marcadas en color gris.

9. Búsquedas masivas en bases de datos

A partir de las diferencias y los rasgos comunes observados en el alineamiento de dos secuencias, obtenemos suficiente información para establecer las posibles relaciones evolutivas entre estas. Junto con la reconstrucción de la filogenia de nuestras secuencias, estas comparaciones nos proporcionan más datos interesantes. Imaginemos que para una de estas secuencias averiguamos qué rol biológico desempeña en la célula, mientras que de la otra molécula no sabemos nada a priori. En este caso, podemos inferir las funciones que hipotéticamente puede poseer la segunda secuencia a partir de aquello conocido para la primera.

Este proceso de inferencia es factible cuando el porcentaje de similaridad entre ambas secuencias supera un cierto límite (40%-50%). Este valor umbral establece, cuando es plausible, que dos secuencias puedan ser homólogas o no en base a las características estructurales de su alineamiento. Dado que continuamente la comunidad científica realiza anotaciones más precisas de los genomas y sus proteínas, la búsqueda de posibles homologías entre nuestras secuencias de estudio y el cuerpo de conocimiento almacenado en los bancos de datos biológicos es una operación muy común en el análisis bioinformático habitual (ver figura 42). Ello implica que, aunque el cálculo del alineamiento óptimo (global o local) sea posible en un tiempo aceptable cuando comparamos únicamente dos secuencias, su coste cuadrático resulta inaceptable en la práctica si deseamos contrastar miles de secuencias anotadas para descubrir homologías con nuestra secuencia de estudio. Obviamente, los usuarios de estos servicios necesitan obtener una rápida respuesta a sus preguntas para acelerar el proceso de investigación científica.

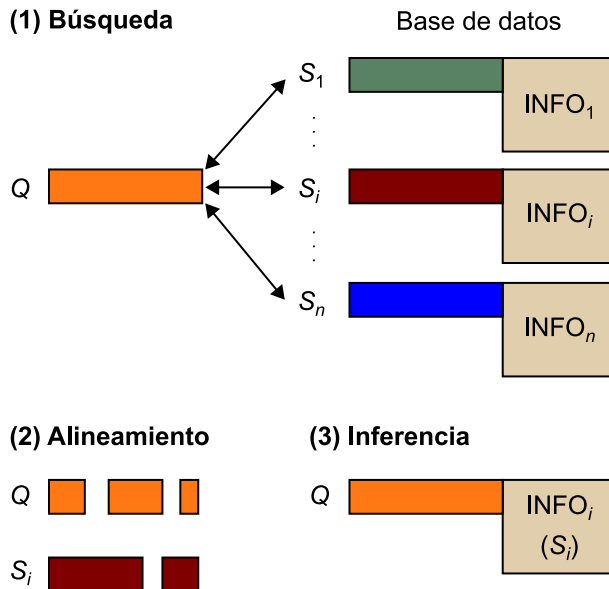
Lecturas complementarias

D. Mount (2001). *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. ISBN: 0879696087.

A. D. Baxevanis; B. F. Ouellette (2005). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Hoboken, NJ: John Wiley & Sons Inc. ISBN: 0471478784.

M. Zvelebil; J. O. Baum (2008). *Understanding bioinformatics*. Londres: Garland Science. ISBN: 0815340249.

Figura 42. Estrategia de las búsquedas por homología.



Varias técnicas heurísticas, generalmente basadas en el uso de alineamientos locales, han surgido como alternativa a la programación dinámica para realizar miles de comparaciones en pocos segundos. Pese a que no existe garantía de que los resultados obtenidos con estos métodos sean óptimos desde el punto de vista algorítmico, existen ciertos requisitos que aseguran en la mayoría de ocasiones un resultado plausible desde el punto biológico. Para evitar muchas operaciones de comparación innecesarias, los programas de búsqueda de homología realizan un preprocesamiento previo de todas las secuencias de la base de datos, dando formato a un diccionario interno.

La familia de programas BLAST (del inglés *basic local alignment search tool*, herramienta de búsqueda de alineamientos locales básicos) es el representante más popular de estos procedimientos heurísticos. La ganancia en términos de velocidad de ejecución lograda por las búsquedas implementadas con BLAST está causada por una aproximación diferente al problema del alineamiento. En lugar de realizar comparaciones con las secuencias completas, BLAST identifica primero aquellos fragmentos de nuestra secuencia de interés (denominada *query* o interrogación) que podrían generar alineamientos prometedores con ciertas secuencias de la bases de datos. Una vez localizadas estas semillas, BLAST hace crecer cada posible alineamiento en ambas direcciones evaluando simultáneamente la calidad del resultado.

Un concepto relevante en la terminología de los programas BLAST es el de vecindad. Estos programas trabajan con palabras relativamente cortas (por ejemplo, tres aminoácidos). El preprocesamiento de una base de datos con BLAST calcula, para cada palabra que pueda existir en alguna secuencia, el conjunto de palabras de la misma longitud que obtienen una puntuación relativamente similar. Como cada palabra está internamente enlazada con las secuencias donde se encuentra, para descubrir qué secuencias de la base de datos son más útiles y extender los alineamientos, es suficiente comparar las palabras que

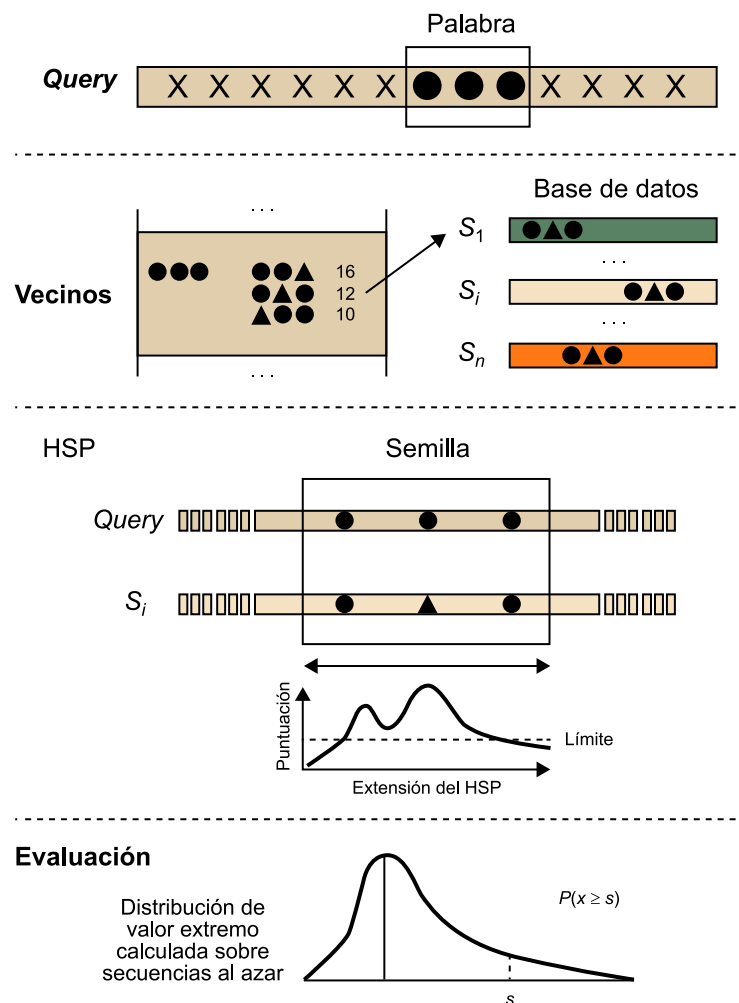
Lectura complementaria

S. F. Altschul; W. Gish; W. Mfller; E. W. Myers; D. J. Lipman (1990). "Basic local alignment search tool". *Journal of Molecular Biology* (núm. 215, págs. 403-410).

constituyen nuestra *query* con el diccionario de vecinos o sinónimos de la base de datos que estamos interrogando. Para evaluar los sinónimos de cada posible palabra debe emplearse una matriz de sustitución, recompensando además aquellas palabras menos frecuentes dado que resultan enormemente útiles para acelerar la búsqueda. El protocolo de búsqueda de BLAST está dividido en los siguientes pasos (ver figura 43):

- Analizar la secuencia *query* para enmascarar regiones repetitivas que podrían producir resultados erróneos en la exploración.
- Generar el vecindario de cada palabra de k símbolos de la secuencia de entrada. Clasificar las palabras parecidas que contienen sustituciones sinónimas según su similaridad.
- Buscar en la base de datos aquellas secuencias que contienen en su interior alguna de esas palabras más similares para generar los alineamientos.
- Construir la semilla del futuro alineamiento local empleando cada palabra de nuestra *query* que posea algún vecino en una secuencia de la base de datos.
- Extender las semillas en ambas direcciones, evaluando la similaridad de los nuevos símbolos alineados con una matriz de sustitución. El proceso de crecimiento del nuevo alineamiento finaliza cuando la puntuación global de éste cae por debajo de cierto umbral.
- Evaluar la bondad del nuevo alineamiento en función de la probabilidad que existe de encontrar por azar un alineamiento de parecida calidad en la base de datos.

Figura 43. El algoritmo de alineamiento de BLAST.



El alineamiento resultante que supera positivamente la fase de extensión recibe el nombre de HSP (del inglés *high-scoring segment pair*, par de segmentos con alta puntuación). Lógicamente, antes de iniciar las exploraciones, la colección de secuencias utilizada como base de datos debe ser preprocesada para elaborar el diccionario de palabras. Esta operación únicamente debe llevarse a cabo una vez, proporcionando una enorme ganancia de velocidad durante las futuras búsquedas. Dado que estamos trabajando con miles de secuencias, es factible que muchos resultados se obtengan puramente por azar. Para evaluar sólidamente la bondad de uno de nuestros alineamientos, es necesario que BLAST genere una muestra suficiente, entre secuencias escogidas aleatoriamente que posean la misma composición. Con esta es posible obtener la distribución estadística de puntuaciones de alineamientos en función de nuestra base de datos. En los problemas de optimización (en este caso buscamos la máxima similitud), estos conjuntos de valores suelen seguir la distribución de valores extremos (en inglés, *extreme value distribution*).

Para valorar la significación estadística de nuestro alineamiento bajo un cierto valor de confianza, es necesario contar cuántos alineamientos escogidos al azar lograron obtener una puntuación igual o mejor. Esta distribución depende obviamente del tamaño de nuestra secuencia (m), de las secuencias alma-

Lectura complementaria

S. Karlin; S. F. Altschul (1990). "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes". *Proceedings of the National Academy of Sciences* (núm. 87, págs. 2264-2268).

cenadas en la base de datos (N) y de dos constantes (k y l) relacionadas con la normalización de las puntuaciones calculadas con las matrices de sustituciones en los HSP. Para un alineamiento que logra una determinada puntuación s , BLAST calcula esta probabilidad como:

Figura 44. Probabilidad de la similaridad en la distribución de valores extremos.

$$P(x \geq s) = kmNe^{-\lambda s}$$

En el área de la genómica computacional, los analistas bioinformáticos habitualmente realizan contrastes entre secuencias de nucleótidos y aminoácidos para identificar homologías en regiones codificantes de los genes. Bajo el paraguas de la filosofía de la técnica de búsquedas en grandes bases de datos proporcionado por BLAST, existen varios programas de esta familia para cada clase de comparación que sea preciso llevar a cabo (ver figura 46 para una representación general).

Los dos modos más sencillos involucran búsquedas con secuencias de la misma clase por homología a nivel genómico (BLASTN) o a nivel de proteínas (BLASTP). No obstante, cuando trabajamos con genes podemos asumir ciertas propiedades sobre las secuencias que facilitan su posterior análisis. Por ejemplo, si sospechamos que una secuencia genómica codifica una proteína en su interior, es posible solicitar al programa que previamente obtenga la traducción en las seis pautas de lectura de la secuencia de ADN (tres en cada hebra de la molécula) para compararlas después directamente con las proteínas de algún banco de datos (BLASTX). Mediante esta traducción, si efectivamente nuestra secuencia codifica una proteína (o un fragmento de esta), una de las seis comparaciones arrojará resultados positivos. A la inversa, cuando estamos comparando una proteína contra una colección de transcritos, podemos solicitar realizar la traducción del banco de datos genómico completo para evaluar la similaridad con una cierta proteína (TBLASTN). En caso de que tanto la secuencia *query* como la base de datos sean de procedencia genómica, cuando conozcamos a priori que todas estas secuencias codifican proteínas en su interior, podemos requerir la traducción en paralelo de ambos conjuntos para ejecutar la búsqueda a nivel proteómico (TBLASTX). Un alineamiento entre dos codones sinónimos puede no ser significativo a nivel genómico, pero obtener una puntuación positiva a nivel de proteína empleando cualquier matriz de sustitución. Con esta última opción, por tanto, podrían identificarse homologías más alejadas evolutivamente que difícilmente serían reconocibles estudiando cadenas de ADN.

Fácil memorización

La nomenclatura de los programas facilita su memorización: BLASTN o BLASTP indican que la *query* es genómico (N) o proteína (P), la (X) indica que ésta debe ser traducida (como en BLASTX), mientras que la (T) se utiliza para indicar la traducción de la base de datos.

Lectura complementaria

Existen versiones más sofisticadas de BLAST que permiten reconocer patrones lejanos de similaridad (como PSI-BLAST). Para más información:

S. F. Altschul; T. L. Madden; A. A. Schaffer; J. Zhang; Z. Zhang; W. Millery; D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Research* (núm. 25, págs. 3389-3402).

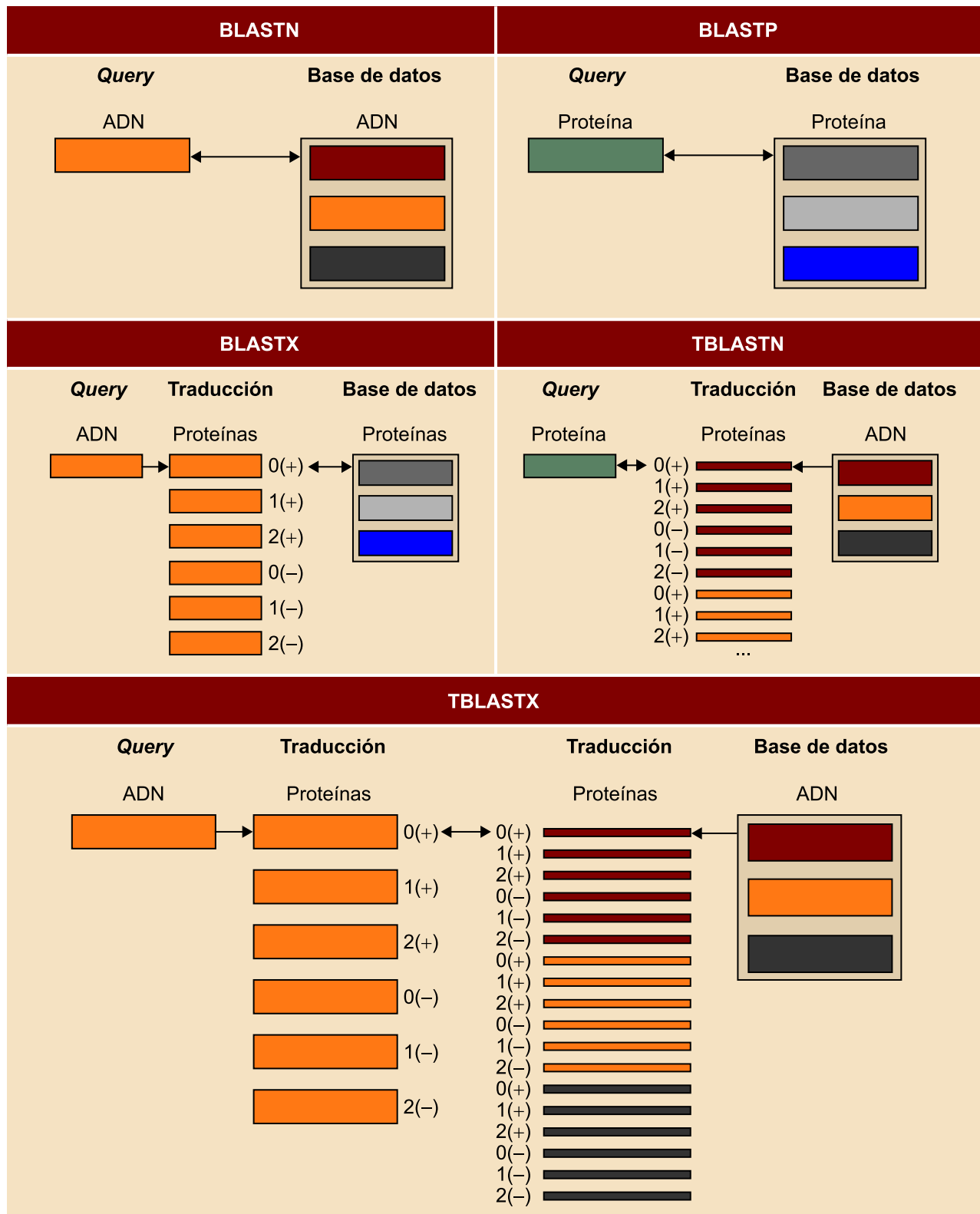
Figura 45. Capturando homología en regiones codificantes.

$S_1 =$	A	G	G	T	A	C	T	T	A	C	C	G
$S_2 =$	C	G	A	T	A	T	A	T	C	C	C	T
<hr/>												
$S_1 =$	A	G	G	T	A	C	T	T	A	C	C	G
Σ_{PRO}		R			Y			L			P	
Σ_{PRO}		R			Y			I			P	
$S_2 =$	C	G	A	T	A	T	A	T	C	C	C	T

Leyenda figura 45

Observamos que el alineamiento de las dos secuencias de ADN sólo detecta la mitad de las coincidencias entre las respectivas proteínas.

Figura 46. La familia de programas BLAST.

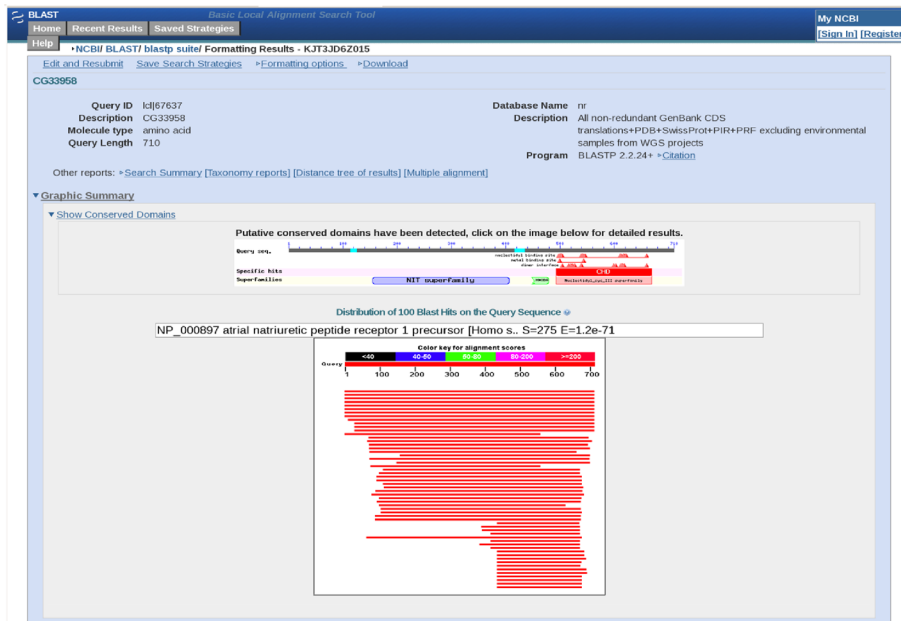


Para finalizar esta sección vamos a estudiar brevemente el funcionamiento en un caso real con el servidor web de BLAST ubicado en el National Center for Biotechnology Information (NCBI, Centro Nacional para la Información Biotecnológica de Estados Unidos). El objetivo primario de nuestra búsqueda es la localización de homólogos remotos en otras especies de la proteína CG33958 perteneciente a la mosca *Drosophila melanogaster*. Como podemos observar en

la figura 47, debemos seleccionar el formulario para ejecutar la versión BLASTP (proteína contra una base de datos de proteínas), copiar nuestra secuencia en la caja de texto de la *query* y elegir la base de datos más apropiada (la colección de proteínas nr). Tras un breve lapso de tiempo aparecerán en nuestra pantalla los resultados de la exploración. BLAST realiza una caracterización de los dominios de la proteína y, a continuación, muestra gráficamente las proteínas del banco de datos más similares a nuestra secuencia, clasificadas de mayor a menor similitud. Para observar uno de los alineamientos debemos pulsar sobre alguno de los segmentos coloreados. El resultado mostrado aquí corresponde a una proteína humana que posee una similitud aceptable con el segundo dominio de nuestra proteína de *Drosophila*.

Figura 47. Utilizando NCBI BLAST para buscar homología.

The screenshot shows the NCBI BLAST web interface. The top navigation bar includes links for Home, Recent Results, Saved Strategies, and My NCBI. The main heading is 'Basic Local Alignment Search Tool'. Below this, there are tabs for 'blastn', 'blastp', 'blastx', 'tblastn', and 'tblastx'. The 'blastp' tab is selected. The page title is 'BLASTP programs search protein databases using a protein query. more...'. The 'Enter Query Sequence' section has a text area with the sequence: 'MLNGKAHKSGENPDSPCPAGGSRMPTPSVKIDISTCQSNRPMESTRAPD'. Below the text area is a 'Browse...' button. The 'Job Title' field contains 'CG33958'. The 'Choose Search Set' section has a 'Database' dropdown set to 'Non-redundant protein sequences (nr)'. The 'Organism' field is empty. The 'Exclude' section has checkboxes for 'Models (XM/XP)', 'Uncultured/environmental sample sequences', and 'Entrez Query'. The 'Program Selection' section has a radio button selected for 'blastp (protein-protein BLAST)'. The 'BLAST' button is at the bottom left. The bottom right shows the search parameters: 'Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)' and a checkbox for 'Show results in a new window'.



>ref|NP_000897.3| atrial natriuretic peptide receptor 1 precursor [Homo sapiens]
 Score = 275 bits (704), Expect = 1e-71, Method: Compositional matrix adjust.
 Identities = 136/245 (56%), Positives = 179/245 (74%), Gaps = 4/245 (1%)

```

Query  434  LVKNAAATIQLYALNLSQKAKELKR----EKRKSDSLFQMLPPSVAMQLKQTQQVPAEL  489
      ++ N  + ++ YA NL + + E  +   EKRK+++LL+Q+LP SVA QLK+ + V AE
Sbjct  812  ILDNLLSRMEQYANNLEELVEERTQAYLEEKRKAEALLYQILPHSVAEQLKRGETVQAEA  871

Query  490  YEAVTIYFSDIVGFTEIAADCTPLEVVTFLNSIYRVFDERIECYDVYKVVETIGDSYMVAS  549
      +++VTIYFSDIVGFT ++A+ TP++VVT LN +Y  FD  I+ +DVYKVVETIGD+YMV S
Sbjct  872  FDSVTIYFSDIVGFTALSAESTPMQVVTLLNDLYTCFDAVIDNFDVYKVVETIGDAYMVVS  931

Query  550  GLPVKNGNKHISEIATMALDLLDASSVFRIPRAGDEFVQIRCGVHTGPVVAGIVGTMKMPR  609
      GLPV+NG H E+A MAL LLDA FRI E +++R G+HTGFPV AG+VG KMPR
Sbjct  932  GLPVNRNRLHACEVARMALALLDAVRSFRIRHRPQQLRLRIGIHTGFPVCAGVVGLKMPR  991

Query  610  YCLFGDVTNTASRMESTGEAQKIHITTEMHDSLQOVGGFRTEHRGLIDVKGKGLMSTYWL  669
      YCLFGDVTNTASRMES GEA KIH++ E L++ GGF E RG +++KGKG + TYWL
Sbjct  992  YCLFGDVTNTASRMESNGEALKIHLSSSETKAVLEEFGGFELELRGDVEMKKGKVRTYWL  1051

Query  670  TCKDG 674
      + G
Sbjct  1052  LGERG 1056
  
```

10. Alineamientos de genomas

Actualmente, gracias a los importantes avances en la secuenciación masiva de los genomas, es posible efectuar fácilmente análisis a gran escala entre las anotaciones de múltiples especies. Dado que estamos hablando de alineamientos que involucran millones de nucleótidos, es preciso emplear técnicas heurísticas generalmente basadas en el uso de diccionarios de palabras para producir resultados aceptables en un tiempo ajustado. A nivel de secuencia genómica, por ejemplo, es posible elaborar directamente comparaciones entre el elenco de cromosomas de dos o más especies con el objetivo de identificar translocaciones de bloques sinténicos entre varios genomas (ver figura 48). Resultados similares han sido calculados para generar mapas de sintenia entre los cromosomas humanos y del ratón, reportando grandes bloques de alta similitud entre estas especies.

En otro campo del análisis bioinformático, los experimentos de secuenciación a gran escala de diferentes elementos funcionales de los genomas requieren de potentes algoritmos capaces de descubrir en fracciones de segundo la ubicación en el genoma de pequeñas secuencias de escasamente 30 o 40 nucleótidos, pertenecientes a cualquier cromosoma, que generalmente poseen millones de pares de bases. En consecuencia, el alineamiento entre genomas está demostrando ser una herramienta cada vez más efectiva y necesaria para procesar grandes volúmenes de datos producidos por las biotecnologías emergentes.

Lecturas complementarias

A. Darling; B. Mau y otros (2004). "Mauve: Multiple alignment of conserved genomic sequence with rearrangements". *Genome Research* (núm. 14, págs. 1394-1403).

N. Bray; I. Dubchak; L. Pachter (2003). "AVID: A global alignment program". *Genome Research* (núm. 13, págs. 97-102).

M. Blanchette; W. J. Kent; C. Riemer y otros (2004). "Aligning multiple genomic sequences with the Threaded Blockset Aligner". *Genome Research* (núm. 14, págs. 708-715).

Figura 48. Identificación de translocaciones en varios microorganismos con MAUVE.



Resumen

A lo largo de este capítulo el estudiante ha adquirido las nociones básicas para comprender la importancia de la correcta comparación de secuencias biológicas dentro de la investigación en genética y biología molecular. Hemos especificado formalmente los principales algoritmos para el cálculo del alineamiento óptimo entre dos secuencias, introduciendo diferentes esquemas de puntuación biológicamente más realistas. Para emplear más secuencias en nuestros alineamientos, hemos introducido las técnicas progresivas más populares junto con la identificación de motivos conservados. Finalmente, hemos indagado sobre el funcionamiento de la herramienta BLAST para efectuar búsquedas a gran escala en bases de datos y hemos repasado brevemente la importancia de otra clase de alineamientos basados en la comparación de genomas completos.

Actividades

1. Escoged dos proteínas homólogas, efectuad su comparación mediante algún programa de generación de *dot plots* y contrastad el resultado con un programa de alineamiento convencional. Configurad los parámetros de la matriz de puntos para reproducir el resultado del alineamiento global.
2. Realizad distintas pruebas con varias matrices de sustitución de la familia BLOSUM sobre un conjunto de proteínas homólogas de varias especies y estudiad cómo afecta el uso de cada matriz al resultado final.
3. Diseñad la rutina de recuperación recursiva del alineamiento óptimo dentro del esquema convencional de programación dinámica para alinear globalmente dos secuencias.
4. Implementad en Perl el algoritmo completo de alineamiento óptimo global basado en el uso de programación dinámica mostrado en los materiales (denominado también Needleman y Wunsch).
5. En la figura 25, recuperad el mejor alineamiento que finaliza en la posición (4,5) de la matriz de programación dinámica, cuyo valor es 7. Repetid con la posición (6,6) que contiene el valor 2.
6. Demostrad formalmente que el coste de generalizar el esquema de programación dinámica para alinear tres secuencias es cúbico. Indicad los cambios que se deberían producir en las estructuras de datos y los algoritmos para adaptarse a esta nueva situación.
7. Modificad la implementación propuesta para alinear localmente dos secuencias de modo que pueda reportar los N mejores alineamientos locales en cada ejecución.
8. Estudiad el algoritmo de alineamiento global de dos secuencias basado en programación dinámica, para proponer una versión basada en el uso de la función de distancia en lugar de la función de similitud típicamente utilizada.
9. Realizad la misma aproximación para proponer una adaptación del algoritmo de Smith y Waterman al modelo de distancias.
10. T. L. Bailey y C. Elkan describieron el uso de la técnica de maximización de la esperanza para identificar motivos conservados en:

"Fitting a mixture model by expectation maximization to discover motifs in biopolymers". *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology (ISMB)* (págs. 28-36).

Estudiadlo cuidadosamente y especificad formalmente el problema de partida y el modelo probabilista que proponen con las posibles opciones.
11. Buscad dos proteínas homólogas remotas y localizad algún fragmento de estas que únicamente sea posible identificar a nivel de secuencia de aminoácidos (es decir, que a nivel genómico, BLASTN no funcionaría apropiadamente en ese caso).

Ejercicios de autoevaluación

1. Definid en pocas palabras los conceptos de alfabeto, lenguaje y secuencia.
2. Enumerad las cuatro propiedades básicas de los alineamientos representados como correspondencias entre los caracteres de dos secuencias.
3. ¿Qué tres operaciones podemos utilizar para colocar un carácter en un alineamiento?
4. Describid qué diferencias existen entre evaluar un alineamiento mediante un esquema de similitud y uno basado en el uso de distancias.
5. Reflexionad sobre el diferente rango de valores adoptados por un esquema de similitud o uno definido sobre distancias matemáticas.
6. Explicad en términos evolutivos el significado del alineamiento entre dos secuencias.
7. ¿Qué concepto biológico podemos cuantificar con el uso de las matrices de sustitución de aminoácidos?

8. Describid las principales diferencias en el proceso de generación de las matrices PAM y BLOSUM.
9. En términos de alineamientos resultantes, explicad la principal diferencia entre las estrategias globales y locales para comparar secuencias.
10. ¿En términos biológicos, para qué problemas resulta más adecuado utilizar una aproximación global o una comparación local?
11. ¿Qué denominación específica recibe el alineamiento múltiple de secuencias en un contexto local?
12. Describid en pocas palabras la utilidad de una matriz de puntos (*dot plot*).
13. ¿Qué características debe poseer un problema de optimización para ser abordado mediante la técnica de programación dinámica?
14. Describid en pocas palabras qué representa el valor almacenado en una posición $S(i,j)$ de la matriz de programación dinámica, una vez ha sido completado el barrido de ésta.
15. ¿Qué tres opciones tenemos en programación dinámica para calcular el mejor alineamiento entre dos prefijos de las secuencias originales?
16. Explicad para qué sirven las recurrencias de inicialización de la primera fila y la primera columna de la matriz de programación dinámica.
17. Estamos alineando dos secuencias de m y n símbolos, respectivamente. Definid exactamente qué posición de la matriz de programación dinámica albergará la similaridad entre ambas secuencias.
18. Describid en pocas palabras la modificación del algoritmo básico de alineamiento global propuesta por Smith y Waterman para calcular alineamientos locales de dos secuencias.
19. Explicad por qué para realizar alineamientos múltiples, la aproximación progresiva es más eficiente en la práctica que aquella basada en la extensión de las recurrencias de programación dinámica.
20. ¿Cuál es la ventaja fundamental de la estimación WPGMA sobre el cálculo efectivo de todas las posibles combinaciones de alineamientos durante las rondas del alineamiento progresivo?
21. Decidid si el método EM realiza alineamientos múltiples para identificar motivos conservados en secuencias biológicas.
22. Discutid cómo efectúa BLAST las búsquedas masivas en bases de datos sin proceder explícitamente a realizar todos los alineamientos posibles.
23. Describid qué método es a priori más adecuado para identificar homologías entre dos secuencias genómicas codificantes: BLASTN o TBLASTX.
24. Explicad el tipo de secuencia *query* y la clase de base de datos que necesitaríais para ejecutar una búsqueda con el programa TBLASTN.
25. Enumerad dos aplicaciones de los programas de alineamiento que operan a nivel de genomas completos.

Solucionario

1. Un alfabeto es una colección de símbolos, que combinados de cierta forma constituyen las palabras que definen un lenguaje. Una secuencia es una sucesión ordenada de palabras con un determinado significado.

2. Propiedades básicas de los alineamientos:

- Es obligatorio no romper el orden relativo entre los elementos de una misma secuencia dentro del alineamiento con otra cadena.
- No es necesario incluir todos los elementos de una secuencia en el alineamiento final.
- Cada carácter puede alinearse exclusivamente con otro elemento de la otra secuencia.
- No está permitido violar la colinearidad del alineamiento.

3 En una posición de un determinado alineamiento puede ocurrir una coincidencia, una sustitución o un hueco (*gap*).

4. El esquema basado en similaridad permite identificar el grado de parecido entre dos secuencias, recompensando básicamente los caracteres coincidentes. La medición de la distancia entre dos secuencias posee un significado más biológico en términos de cambios evolutivos aparecidos hipotéticamente a lo largo de miles de años.

5. Mientras la similaridad permite un rango de valores negativos y positivos, la distancia es siempre positiva, tomando en el mejor de los casos el valor de cero (por ejemplo, el alineamiento de dos secuencias idénticas).

6. El alineamiento identifica las diferencias entre dos secuencias. En términos evolutivos, en el caso de que ambas secuencias hubieran surgido a partir de una secuencia ancestral común, estaríamos reconstruyendo idealmente su contenido.

7. La matriz de sustitución nos permite evaluar con mayor sentido biológico aquellos cambios que no afectan a la estructura y función de las proteínas involucradas.

8. Mientras las matrices PAM se derivaron analíticamente elevando a múltiples potencias los resultados obtenidos en una primera comparación de proteínas homólogas, las matrices BLOSUM de distintos valores se derivaron directamente de bloques conservados en el interior de proteínas que poseían distintos grados de parecido.

9. El alineamiento global realiza una correspondencia general de los caracteres a lo largo de las secuencias comparadas. El alineamiento local, en cambio, únicamente reporta como resultado aquellas regiones dentro de las secuencias con un parecido por encima de un cierto límite.

10. El alineamiento global es idóneo para la comparación de genes y proteínas homólogos o que podemos esperar a priori que desempeñen un rol biológico similar. El alineamiento local es apropiado para identificar elementos de pequeño tamaño conservados dentro de un contexto general diverso. Estos fragmentos recuperados posiblemente pertenezcan a una cierta configuración de dominios de proteínas o sitios de unión a factores de transcripción.

11. Búsqueda de motivos o descubrimiento de patrones (en inglés, *motif finding* o *pattern discovery*).

12. Una matriz de puntos permite explorar gráficamente la existencia de regiones similares entre dos secuencias. Posteriormente, en caso de resultar positiva la prueba, será necesario calcular efectivamente el alineamiento de ambas secuencias.

13 Debe ser posible descomponer el problema original en subproblemas más pequeños, de modo que la solución óptima al problema principal pueda expresarse en términos de las mejores soluciones para estos problemas más simples.

14. La posición $S(i,j)$ contiene el valor del mejor alineamiento global posible entre el prefijo de i caracteres de la primera secuencia y el prefijo de j caracteres de la segunda. En un esquema de puntuaciones basado en la similaridad, este valor representa la máxima similaridad entre esas dos subsecuencias.

15. Siempre tenemos tres opciones: emparejar el último carácter de ambos prefijos, introducir un hueco en el final del primero o introducir un hueco en el final del segundo.

16. La primera fila y columna de la matriz son útiles para introducir *gaps* de cualquier longitud al inicio y final del alineamiento.

17. En un alineamiento global, la posición $S(m,n)$ de la matriz alberga la máxima similaridad entre las dos secuencias de trabajo, obtenida a partir del cálculo del mejor alineamiento posible.

18. Esta modificación consiste en fijar un valor mínimo que cualquier alineamiento entre los prefijos de las secuencias deberá alcanzar. Durante la reconstrucción del mejor alineamiento, el procedimiento finalizará en el momento en que dicha marca sea alcanzada (la similaridad del fragmento resultante no mejoraría si continuáramos la extensión).

19. La técnica progresiva siempre calcula alineamientos que involucran dos elementos (ya sean secuencias u otros alineamientos previamente calculados), mientras que la aproximación de programación dinámica genera dinámicamente el alineamiento múltiple empleando todas las secuencias simultáneamente. Este último caso no es factible en la práctica cuando estamos trabajando con varias secuencias, a causa de su abundante gasto de memoria y excesivo tiempo de cálculo.

20. Con la estimación WPGMA nos ahorramos calcular efectivamente ciertos alineamientos que posiblemente no serían seleccionados durante el proceso progresivo a causa de su deficiente calidad.

21. El método EM no produce alineamientos efectivos que involucren las secuencias relacionadas. Por el contrario, estimando los parámetros de un modelo probabilístico que reproduce el contenido del motivo mejor conservado, realiza iterativamente barridos por las secuencias para identificar aquellas cadenas de w símbolos más frecuentes.

22. BLAST utiliza un diccionario de palabras para identificar qué secuencias de la base de datos serán posiblemente más similares a la nuestra. Únicamente calcula los alineamientos entre aquellos fragmentos de las secuencias que contienen palabras comunes entre nuestra *query* y el banco de información utilizado.

23. TBLASTX es más adecuado dado que procesa todas las posibles proteínas codificadas dentro de cada secuencia, empleando una matriz de sustitución para recuperar aquellos aminoácidos que producen un cambio sinónimo.

24. El programa TBLASTN realiza búsquedas entre una secuencia genómica y todas las secuencias de ADN de una base de datos. Previamente a la comparación, tanto la secuencia *query* como el banco de datos genómico deben ser traducidos en las seis pautas de lectura. Este tipo de exploración, aunque lenta y más costosa, resulta especialmente efectiva cuando estamos comparando regiones codificantes de genes.

25. Los programas de comparación de genomas completos pueden identificar translocaciones de bloques conservados de secuencia entre dos especies y también pueden averiguar la localización exacta de un pequeño fragmento de centenares de bases dentro de la secuencia completa de los cromosomas de un organismo.

Bibliografía

Abril, J. F.; Guigó, R.; Wiehe, T. (2003). "gff2aplot: Plotting sequence comparisons". *Bioinformatics* (núm. 19, págs. 2477-2479).

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. (1990). "Basic local alignment search tool". *Journal of Molecular Biology* (núm. 215, págs. 403-410).

Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Research* (núm. 25, págs. 3389-3402).

Bailey, T. L. y otros (2009). "GMEME Suite: tools for motif discovery and searching". *Nucleic Acids Research* (núm. 37, págs. W202-W208).

Bailey, T. L.; Elkan, C. (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers". *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology (ISMB)* (págs. 28-36).

Batzoglou, S. (2005). "The many faces of sequence alignment". *Briefings in bioinformatics* (núm. 6, págs. 6-22).

Baxevanis, A. D.; Ouellette, B. F. (2005). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Hoboken, NJ: John Wiley & Sons Inc. ISBN: 0471478784.

Blanchette, M.; Kent, W. J.; Riemer, C., y otros (2004). "Aligning multiple genomic sequences with the Threaded Blockset Aligner". *Genome Research* (núm. 14, págs. 708-715).

Blanco, E. y otros (2007). "Multiple non-collinear TF-map alignments of promoter regions". *BMC Bioinformatics* (núm. 8, págs. 138).

Blanco, E.; Farre, D.; Alba, M.; Messeguier, X.; Guigó, R. (2006). "ABS: a database of Annotated regulatory Binding Sites from orthologous promoters". *Nucleic Acids Research* (núm. 34, págs. D63-D67).

Bray, N.; Dubchak, I.; Pachter, L. (2003). "AVID: A global alignment program". *Genome Research* (núm. 13, págs. 97-102).

Brazma, A. y otros (1998). "Approaches to the automatic discovery of patterns in biosequences". *Journal of Computational Biology* (núm. 5, págs. 279-305).

Carrillo, H.; Lipman, D. (1988). "The multiple sequence alignment problem in biology". *SIAM Journal of Applied Mathematics* (núm. 48, págs. 1073-1082).

Darling, A.; Mau, B. y otros (2004). "Mauve: Multiple alignment of conserved genomic sequence with rearrangements". *Genome Research* (núm. 14, págs. 1394-1403).

Dayhoff, M. O. y otros (1965). *Atlas of protein sequence and structure*. Silver Spring, Maryland: National Biomedical Research Foundation.

Eddy, S. R. (2004). "What is dynamic programming?". *Nature Biotechnology* (núm. 22, págs. 909-910).

Feng, D.; Doolittle, R. F. (1987). "Progressive sequence alignment as a prerequisite to correct phylogenetic trees". *Journal of Molecular Evolution* (núm. 25, págs. 351-360).

Gibbs, A. J.; McIntyre, G. A. (1970). "The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences". *European Journal of Biochemistry* (núm. 16, págs. 1-11).

Gotoh, O. (1982). "An improved algorithm for matching biological sequences". *Journal of Molecular Biology* (núm. 162, págs. 705-708).

Henikoff, S.; Henikoff, J. F. (1992). "Amino Acid Substitution Matrices from Protein Blocks". *PNAS* (núm. 89, págs. 10915-10919).

Hopcroft, J. E.; Motwani, R.; Ullman, J. D. (2007). *Introduction to automata theory, languages and computation*. Englewood Cliffs, NJ: Prentice Hall. ISBN: 0321462254.

Karlin, S.; Altschul, S. F. (1990). "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes". *Proceedings of the National Academy of Sciences* (núm. 87, págs. 2264-2268).

Mount, D. (2001). *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. ISBN: 0879696087.

Needleman, S. B.; Wunsch, C. D. (1970). "A general method to search for similarities in the amino acid sequence of two proteins". *Journal of molecular biology* (núm. 48, págs. 443-453).

Saitou, N.; Nei, M. (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees". *Molecular Biology and Evolution* (núm. 4, págs. 406-425).

Sellers, P. (1974). "On the theory and computation of evolutionary distances". *SIAM Journal of applied Mathematics* (núm. 26, págs. 787-793).

Smith, T. F.; Waterman, M. S. (1981). "Comparison of biosequences". *Advances in Applied Mathematics* (núm. 2, págs. 482-489).

Smith, T. F.; Waterman, M. S. (1981). "Identification of common molecular subsequences". *Journal of Molecular Biology* (núm. 147, págs. 195-197).

Smith, T. F.; Waterman, M. S.; Fitch, W. M. (1981). "Comparative biosequence metrics". *Journal of Molecular Evolution* (núm. 18, págs. 38-46).

Thompson, J. D.; Higgins, D. G.; Gibson, T. J. (1994). "CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic Acids Research* (núm. 22, págs. 4673-4680).

Waterman, M. S. (1984). "Efficient sequence alignment algorithms". *Journal of Theoretical Biology* (núm. 108, págs. 333-337).

Waterman, M. S.; Smith, T. F.; Beyer, W. A. (1976). "Some biological sequence metrics". *Advances in Mathematics* (núm. 20, págs. 367-387).

Zvelebil, M.; Baum, J. O. (2008). *Understanding bioinformatics*. Londres: Garland Science. ISBN: 0815340249.