Modelado de alineamiento de secuencia múltiple: métodos y aplicaciones.

Abstracto

Esta revisión proporciona una visión general sobre el desarrollo de los métodos de alineación de secuencias múltiples (MSA) y sus principales aplicaciones.

Está enfocado en el progreso realizado en la última década. Las tres primeras secciones revisan desarrollos algorítmicos recientes para las alineaciones de proteínas, ARN / ADN y genómica. La cuarta sección trata sobre los puntos de referencia y explora la relación entre los datos empíricos y los simulados, junto con el impacto en los desarrollos del método. La última parte de la revisión brinda una descripción general de los estimadores de confiabilidad local de MSA disponibles y su dependencia de varias propiedades algorítmicas de los métodos disponibles.

Palabras clave: alineaciones de secuencias múltiples; alineaciones de proteínas; puntos de referencia de alineación; medidas de confiabilidad de alineación; RNA
alineaciones; Alineaciones de ADN

Introducción

Los métodos de alineación múltiple de secuencias (MSA) se refieren a una serie de soluciones algorítmicas para la alineación de secuencias evolutivamente relacionadas, teniendo en cuenta eventos evolutivos tales como mutaciones, inserciones, eliminaciones y reordenamientos bajo ciertas condiciones. Estos métodos se pueden aplicar a secuencias de ADN, ARN o proteínas. Un estudio reciente en Nature [1] revela que MSA es uno de los métodos de modelado más ampliamente utilizados en biología, con la publicación que describe ClustalW [2] que señala el # 10 entre los artículos científicos más citados de todos los tiempos. De hecho, una gran cantidad de análisis in silico dependen de los métodos MSA.

Estos incluyen análisis de dominio, reconstrucción filogenética, descubrimiento de motivo y toda una gama de otras aplicaciones, ampliamente descrita en [3-4].

MSA es de hecho una importante herramienta de modelado cuyo desarrollo ha requerido abordar una compleja combinación de problemas computacionales y biológicos. Hace tiempo que se sabe que el cálculo de una MSA precisa es un problema NP-completo, una situación que explica por qué se han desarrollado más de 100 métodos alternativos en las últimas tres décadas [4]. Los métodos originales de MSA (MSAM) y sus aplicaciones han sido extensamente cubiertos por varias revisiones [3-5]. Para evitar la redundancia, nos enfocaremos aquí en los principales desarrollos que han tenido lugar durante estos últimos 10 años y los pondremos en un contexto histórico más amplio cuando sea necesario. Las tres primeras secciones detallarán el marco algorítmico general de las MSAM y mostrarán cómo se relaciona con los métodos más nuevos y su aplicación a todo tipo de secuencias biológicas (proteínas, ARN, ADN). La cuarta parte será cubrir la validación del método y los puntos de referencia disponibles, con un énfasis especial en la generación más nueva diseñada para atender el modelado evolutivo y estructural. La última parte de esta revisión se ocupará de la cuantificación de la confiabilidad local dentro de MSAs. Esta tarea había sido identificada durante mucho tiempo como instrumental, y posiblemente más importante que el cálculo de los modelos necesariamente aproximados. Sin embargo, solo recientemente se han desarrollado enfoques sistemáticos con el objetivo explícito de cuantificar la confiabilidad local, lo que permite un filtrado y ponderación sistemáticos para el modelado aguas abajo. Revisaremos estos métodos a la luz de los últimos informes.

Marcos algorítmicos para el cálculo de MSA

A pesar de su gran diversidad, todas las MSAM comparten una propiedad clave: su confianza en la heurística aproximada y generalmente codiciosa, impuesta por la naturaleza completa del problema. Todas estas heurísticas dependen, más o menos explícitamente, de propiedades de datos específicos, como el tamaño, la naturaleza de la homología, la relación,
longitud y así sucesivamente. Como consecuencia, cualquier cambio, incluso menor, en el tipo de datos que se modela requiere el desarrollo de nuevas estrategias heurísticas. Dichos cambios han incluido recientemente la necesidad de realizar un aumento de escala bajo la presión de secuencia de alto rendimiento y la necesidad de descriptores de secuencia más complejos,
incluyendo ARN no codificante o secuencias genómicas no transcritas. Las necesidades cambiantes de modelado también pueden impulsar los desarrollos de heurísticas novedosas, un hecho bien ilustrado por el reciente desarrollo de alineadores conscientes de la filogenia. Otra fuerza impulsora detrás del desarrollo de nuevas heurísticas ha sido la creciente disponibilidad de datos estructurales que han impulsado el desarrollo de métodos híbridos capaces de tratar simultáneamente con secuencias y estructuras secundarias (ARN) o terciarias (ARN y proteínas). Del mismo modo, la explosión de datos genómicos disponibles ha ejercido una gran presión sobre el desarrollo de una nueva generación de alineadores de ADN no transcritos / no transcritos.

Algoritmos comúnmente utilizados

Dado un conjunto de secuencias biológicas (ARN, proteínas, ADN), el propósito de un método MSA es alinear las secuencias de una manera que refleje su relación evolutiva, funcional o estructural (Figura 1). Este propósito se logra insertando huecos de longitud variable dentro de las secuencias, permitiendo que las posiciones homólogas se alineen entre sí, de la misma manera que uno alinearía cuentas de color idéntico en un ábaco. En un contexto evolutivo, estas lagunas representan inserciones y deleciones (indels) dentro del genoma que se supone que ocurrieron durante la evolución desde un ancestro común. En una MSA correcta, los residuos alineados deberían ser lo más similares posible de acuerdo con algunos criterios especificados. Por ejemplo, si la alineación está destinada a la reconstrucción evolutiva, los residuos deben ser homólogos, es decir, corresponden al mismo residuo en el último ancestro único común de las secuencias consideradas. Si la alineación pretende ser un modelo estructural, los residuos alineados deberían tener posiciones comparables en sus respectivas estructuras 2D o 3D. Si la alineación es funcional, como puede suceder al analizar datos genómicos, se espera que las posiciones alineadas admitan funciones similares. Aunque es razonable esperar una superposición significativa entre estos criterios, se debe enfatizar que la complejidad de las fuerzas evolutivas es tal que su acuerdo total no puede darse por hecho. Por ejemplo, dos estructuras pueden ser similares como consecuencia de la evolución convergente pero no homólogas desde un punto de vista evolutivo.

Para construir un MSA, se necesita una función de puntuación (función objetivo) capaz de cuantificar los méritos relativos de cualquier alineación alternativa con respecto a la relación modelada. El MSA puede entonces estimarse calculando un modelo de puntaje óptimo. La función objetivo es un parámetro crítico, ya que define con precisión la precisión de modelado de un MSA y su capacidad predictiva. Cuando se trata de reconstrucciones evolutivas, las funciones objetivas más comúnmente utilizadas implican la maximización de las similitudes ponderadas (proporcionadas por una matriz de sustitución PAM o BLOSUM) al tiempo que se utiliza una penalización de hueco afín para estimar los costos indeles. El costo de sustitución puede ajustarse utilizando esquemas de ponderación basados en árboles que reflejen la contribución de información independiente de cada secuencia, y la puntuación de las columnas se estima considerando el costo total de sustitución de todos contra todos (sumas de pares). Es bien sabido que es poco probable que las funciones de suma de pares estén modelando las relaciones biológicas con la suficiente precisión [6], pero se ha demostrado que proporcionan una compensación razonable entre la corrección estructural y la computabilidad, es decir, la posibilidad para estimar rápidamente un MSA razonable.

Bajo sus formulaciones más comunes, la optimización de los esquemas de evaluación de sumas de pares es NP-completa. Por lo tanto, uno necesita confiar en la heurística, la más común es el algoritmo de alineación progresiva inicialmente descrito por Hogeweg y Hesper [7]. Este algoritmo implica incorporar las secuencias de entrada una por una en el modelo final, siguiendo una orden de inclusión definida por un árbol de guía precalculado. En cada nodo, se lleva a cabo una alineación por pares entre un par de secuencias, una secuencia y un perfil o dos perfiles. Las alineaciones por pares que tienen lugar en cada nodo se estiman utilizando adaptaciones más o menos sofisticadas del algoritmo de alineación de programación dinámica global de Needlman y Wunsch [8]. La combinación entre una estrategia progresiva basada en árboles y un algoritmo de alineación por pares global forma la columna vertebral de la mayoría de los métodos disponibles (Figura 1), incluidos ClustalW [2], T-Coffee [9] y ProbCons [10]. También está particularmente bien adaptado para el diseño de estrategias iterativas (Figura 1), que implica reestimar árboles y alineamientos hasta que ambos convergen [11], como se implementa en MUSCLE [12], MAFFT [13] y Clustal Omega [14]. Además de la función objetivo, el componente algorítmico principal de la alineación progresiva es el procedimiento de estimación del árbol guía. Este árbol, que decide en qué orden se incorporarán las secuencias, se puede obtener usando una amplia variedad de métodos, el más estándar es Neighbor Joining (NJ) [15] y el Método de grupo de pares no ponderado con media aritmética (UPGMA) [16] . La interacción entre la función objetivo (esquema de sustitución y penalizaciones de hueco), el esquema de ponderación y el árbol es compleja y fue ampliamente explorada por Wheeler [17] quien mostró cómo el ajuste adecuado de estos diversos componentes puede tomar un método estándar hasta el nivel de los más precisos. Por lo tanto, no es sorprendente observar que los últimos desarrollos algorítmicos se han centrado en árboles de guía y mejoras de la función objetivo.

La principal advertencia del enfoque de alineación progresiva es la existencia de mínimos locales (alto nivel de similitud entre un subconjunto de secuencias que resulta de un artefacto). Por ejemplo, si el árbol de guía induce la alineación de dos secuencias distantemente relacionadas, a menudo sucede que la alineación óptima de estas dos secuencias no corresponderá a la proyección por pares que se obtendría del MSA óptimo de todo el conjunto de datos (iewithin the MSA la alineación de las dos secuencias será ligeramente inferior a la óptima para permitir la optimalidad global en el MSA
nivel). Esta situación es común cuando se trata de secuencias de baja identidad o baja complejidad. Cuando esto ocurre, el cálculo temprano de la primera alineación por parejas puede evitar el cálculo de una MSA óptima globalmente.

La estrategia más común para evitar los mínimos locales durante una alineación progresiva es el uso de consistencia, como se describió originalmente por [9]. El fundamento de la coherencia es relativamente sencillo: dado un conjunto de secuencias y sus alineaciones asociadas asociadas, tratadas como restricciones, los puntajes para pares coincidentes de residuos se vuelven a estimar a fin de entregar alineaciones por pares con mayor probabilidad de ser compatibles con una MSA óptima global. La primera estrategia que implicaba tal nueva estimación de los costos de los partidos fue reportada por Morgenstern como pesos superpuestos [18]. Este esquema inspiró más tarde el esquema de puntuación T-Coffee que se ha convertido en el alineador basado en la consistencia progresiva arquetípica [9]. La optimización de una alineación contra un conjunto de restricciones predefinidas se conoce como el problema de Rastreo máximo de peso. Es NP-completo en sus formulaciones más comunes y solo se puede resolver para casos pequeños [19, 20]. El algoritmo de T-Café es un enfoque heurístico que implica volver a estimar los costos iniciales de cada posible emparejamiento por pares teniendo en cuenta su compatibilidad con el resto de la alineación por pares. El esquema de puntuación resultante hace que sea más probable ensamblar subalineaciones consistentes durante el procedimiento MSA progresivo. La principal fortaleza de este enfoque es permitir el cálculo de MSA incluso cuando una función objetivo solo está disponible para ser optimizada a nivel de pares. Los métodos basados en la coherencia y sus relaciones han sido ampliamente revisados en [4]. Desde entonces, el enfoque basado

en la consistencia se ha convertido en uno de los marcos algorítmicos más populares para el desarrollo de métodos novedosos (Figura 1).
En un algoritmo basado en consistencia, el parámetro más crítico es la biblioteca primaria. Dado un conjunto de secuencias, la biblioteca primaria es una colección de todas las posibles comparaciones secuencia secuencial. Esta biblioteca se usa para definir la función objetivo basada en la coherencia. En el T-Coffee original [9], la biblioteca era una compilación de todos los pares de residuos alineados en todas las alineaciones locales y globales por pares. Estos pares de residuos se ponderaron de acuerdo con la fiabilidad estimada de sus alineaciones de origen. Más tarde, una variación del algoritmo T-Coffee ProbCons [10] estableció la superioridad de las bibliotecas basadas en HMM. En ProbCons, las bibliotecas se compilan utilizando un par HMM para estimar la probabilidad posterior de alinear todos los pares de residuos posibles (entre distintas secuencias).
El uso de un par HMM pronto se hizo popular entre otros métodos de alineación (Figura 1). Las principales características novedosas de ProbCons sobre T-Coffee fueron el uso de un marco probabilístico más formal, gracias al HMM y la implementación de una penalización de brecha bifásica al estimar alineaciones por pares.
Los algoritmos que dependen de una combinación similar a menudo se denominan algoritmos de coherencia probabilística; incluyen el alineador de genoma múltiple PECAN [21], que utiliza la alineación por pares Durbin [22] solo dividen y conquistan hacia adelante y MSAProbs [23], que se basa en una función de partición para lograr probabilidades posteriores más informativas al compilar la biblioteca. También se incorporó cierto grado de consistencia en el algoritmo MAFFT 'linsi'.
Cuando se compararon los alineamientos de referencia basados en la estructura, se ha demostrado que los alineadores basados en la consistencia arrojan los ASM más precisos [4, 23]. Esta precisión viene, sin embargo, a una memoria significativa y un costo de CPU, con la mayoría de las implementaciones que son cúbicas en la CPU y cuadráticas en la memoria con el número de secuencias. Se han propuesto tres estrategias para abordar este problema. El más simple implica un cómputo más rápido de la biblioteca. Por ejemplo, FM-Coffee, la rápida implementación de T-Coffee calcula su biblioteca utilizando tres alineadores rápidos, y finalmente extrae las proyecciones pairwise resultantes. La alta correlación entre las diversas proyecciones hace posible la reducción de la consistencia y la complejidad del tiempo y la memoria en un nivel casi cuadrático. Aunque las alineaciones resultantes no son tan precisas como las obtenidas mediante el procedimiento predeterminado, tienden a ser más precisas que las producidas individualmente por los métodos combinados.
La segunda estrategia implica paralelización. Dos de estos esquemas se han publicado recientemente Cloud-Coffee [24] y MSAProbs [23], que implican la paralelización del cálculo de la biblioteca y el paso de relajación durante el cual se reestiman los costos por pares cuando tiene lugar el ensamblaje de alineación progresiva. El último paso, que implica dividir el cálculo de acuerdo con la topología del árbol, es altamente dependiente de la simetría del árbol guía, logrando mejores resultados con árboles guía perfectamente equilibrados. La tercera estrategia es más sofisticada e implica afinar la granularidad de la biblioteca al considerar segmentos de secuencia en lugar de residuos únicos. Esta implementación, disponible en SeqAn [25], es especialmente adecuada para largas secuencias estrechamente relacionadas, en las que se pueden identificar segmentos idénticos largos.

Alineadores múltiples a gran escala

Incluso en sus formas más optimizadas, los métodos basados en consistencia no pueden tratar con más de unos cientos de secuencias. Este límite es bastante severo en un contexto donde la explosión de la disponibilidad de secuencias genómicas ha resultado en familias homólogas grandes sin precedentes que pueden requerir la alineación de hasta 1.5 millones de miembros (cantidad de transportadores ABC en Pfam) y pronto muchos
Más. Si bien la relevancia biológica de grandes MSA puede cuestionarse, los análisis recientes indican que se pueden establecer resultados importantes a partir de modelos tan grandes [26], lo que hace que

la construcción precisa y eficiente de grandes MSA sea uno de los grandes desafíos actuales de la biología moderna. Tres métodos son actualmente capaces
para tratar con conjuntos de datos tan grandes: el modo PartTree de MAFFT [13], Clustal Omega [14] y PASTA [27], la versión más reciente de SATé [28] (Figura 1). Estos métodos comparten una característica común: su dependencia de un paso rápido de pre-agrupamiento (sub-cuadrático en el tiempo) que hace posible determinar rápidamente el orden en el cual las secuencias deben estar alineadas.
En los métodos progresivos originales, el árbol guía se estimó comparando todas las secuencias entre sí para estimar una matriz de distancia. Esta comparación se puede basar en una alineación lenta de Needleman y Wunsch [8] o en una comparación rápida de vectores en k-tuple como se implementó en MAFFT [13], MUSCLE [29]
y T-Coffee [9]. Sin embargo, la comparación rápida no resuelve el problema de los requisitos de tiempo y espacio cuadráticos para el cálculo de la matriz, seguido de la complejidad del tiempo cúbico de la estimación del árbol cuando se utiliza UPGMA o NJ. Estos requisitos se vuelven prohibitivos al procesar más de 10 000 secuencias. Los métodos de agrupamiento recientes se han diseñado para abordar este problema. En Clustal Omega [14], el árbol guía se estima utilizando el método mBed [30]. El principio de mBed es primero estimar la distancia entre cada secuencia y un pequeño subconjunto de secuencias seleccionadas sobre la base de su longitud. Para cada secuencia, el resultado es un vector de distancia que se puede utilizar para ejecutar una agrupación k-means jerárquica (Figura 1), cuya complejidad relativamente baja (NlogN bajo las implementaciones heurísticas más comunes) permite grandes conjuntos de datos de 10 000 secuencias o más estar alineado. PartTree se basa en un procedimiento ligeramente diferente
eso también implica el uso de un pequeño conjunto de secuencias de semillas para pre-agrupar rápidamente las secuencias. Tanto en mBed como en PartTree, el paso de preclúster es seguido por el cálculo de subárboles que eventualmente se combinan para formar el árbol de guía. El enfoque PartTree se mejoró recientemente en el algoritmo SATé, que implica un paso iterativo adicional de refinamiento del árbol.
El último intento de alinear grandes conjuntos de datos es una versión adaptada del algoritmo T-Coffee que implica la combinación de clusters k-means con MSA basados en consistencia en un nivel inferior [26].
Estos enfoques se escalan bien, pero a costa de precisiones significativamente menores cuando se alinean> 1000 secuencias, como se muestra en el análisis de referencia de Clustal Omega [14]. Un posible efecto secundario de esta disminución de la precisión ha sido el informe de inconsistencias de alta alineación entre MAFFT, Clustal Omega y
T-Coffee cuando se trata de grandes conjuntos de datos de secuencias mitocondriales ortólogas relativamente similares. Al considerar conjuntos de datos completos, los autores informan niveles de acuerdo promedio tan bajos como 60% [26].

Alineadores múltiples sensibles a la filogenia
Un hito importante en el desarrollo de las MSAM ha sido la introducción de alineamientos de referencia basados en la estructura que pueden usarse para comparar las capacidades relativas de varios métodos para reconstruir alineamientos estructuralmente correctos a partir de secuencias solamente. La elección de la estructura parece bastante natural porque se sabe que las características 3D son más resistentes a la evolución que las secuencias subyacentes. Por otro lado, este enfoque se basa en la lógica no probada de que las alineaciones correctas estructural y evolutivamente son idénticas. No existe ninguna prueba de que esta suposición sea correcta, y un simple razonamiento sugiere que puede no ser el caso. De hecho, si bien puede haber una sola forma correcta de emparejar residuos homólogos, la que refleja perfectamente la historia evolutiva única de las secuencias y coincidencias consideradas, puede haber tantas alineaciones estructuralmente correctas como formas de superponer

las secuencias con 3D equivalente. compacidad. Otra importante discrepancia potencial entre las alineaciones estructurales y evolutivas resulta de la evolución convergente. Cuando un proceso de este tipo ha configurado algunas partes de un conjunto de datos de secuencia, el resultado
las regiones convergentes que coinciden con la alineación serán estructuralmente correctas y evolutivamente falsas, y recíprocamente.
Esta cuestión se ha abordado recientemente mediante una serie de trabajos que tienen como objetivo evaluar la precisión de alineadores múltiples en función de la calidad de los modelos filogenéticos que respaldan. Estos alineadores se conocen como alineadores de filogenia (Figura 1).
PRANK [31] fue uno de los primeros. Se basa en la idea de que los MSA correctos deben tener patrones indeles que reflejen adecuadamente el árbol filogenético subyacente. PRANK fue seguido rápidamente por SATé [28], un alineador múltiple iterativo derivado de MAFFT que intenta estimar el MSA que soporta el árbol de máxima verosimilitud de mayor puntuación. Un mérito importante de este enfoque es apartarse de la suposición largamente sostenida de que la mejor MSA es la que maximiza la similitud entre secuencias. En el contexto de alineadores conscientes de filogenia, la mejor MSA se define como la que produce el mejor modelo filogenético [32]. Las consecuencias en
las alineaciones resultantes son bastante significativas y fueron particularmente bien ilustradas en un análisis reciente por Blackburne y Whelan que encontraron que las MSAM basadas en similitud (por ejemplo, ClusalW, MUSCLE, ProbCons, MAFFT, T-Coffee) y 'evolution-
las 'MSAM' basadas (por ejemplo, PRANK y BAliPhy) tienden a formar grupos discretos bajo la escala multidimensional basándose en sus propias medidas de similitud entre pares de MSAs alternativas [33]. Estos autores encontraron que los alineadores seleccionados tienen un impacto sustancial en la inferencia filogenética aguas abajo e informan que las topologías de los árboles y la longitud de las ramas dependen de la categoría del alineador. Los alineadores también tienen un impacto claro al cuantificar la selección positiva, con diferentes lecturas asociadas con diversos alineadores según se informa en el análisis de varios genomas de Drosophila [34]. Morrison sugiere que los filogenéticos suelen estar insatisfechos con los alineamientos basados en la similitud y tienden a editar manualmente sus MSA para producir alineaciones que probablemente reflejen la homología desde un verdadero punto de vista evolutivo [32]. Esta observación también puede explicar por qué los resultados y la clasificación del método lograda en los conjuntos de datos simulados evolutivamente difieren significativamente de los medidos en datos empíricos basados en estructuras [4].
Sin embargo, un estudio reciente de Chang [35] muestra que se puede usar el mismo índice de confiabilidad para seleccionar tanto las posiciones más filogenéticamente informativas como las posiciones que con mayor probabilidad contienen residuos estructuralmente análogos. Vale la pena mencionar que este estudio dio resultados contrastantes con respecto a los alineadores de filogenia, y aunque SATé parece funcionar bien tanto en la estructura como en los puntos de referencia evolutivos, se encontró que PRANK devuelve alineamientos estructurales pobres, al tiempo que es capaz de producir alineamientos que soportan árboles con una precisión comparable con las otras MSAM-filogenia consciente y basada en la similitud.

MSA basados en estructura
Gracias a su alta capacidad de adaptación evolutiva, la información estructural puede ayudar a producir modelos de alta calidad, especialmente en situaciones en las que se pretende modelar relaciones estructurales y funcionales. Esta sección repasa brevemente algunos de los métodos capaces de combinar secuencia e información estructural cuando
alineando secuencias de ARN o proteína.

Secuencia proteica / estructura de alineaciones múltiples usando alineadores de proteínas basados en plantillas

Se sabe desde hace tiempo que la información estructural es más resistente que su contraparte de secuencia subyacente [36]. Sin embargo, es solo recientemente que el corpus de información estructural disponible ha hecho que valga la pena desarrollar métodos capaces de combinar secuencia e información estructural dentro de un único modelo. Mientras que la
La primera generación de métodos utilizados para confiar en el enhebrado de estructuras de proteínas y métodos relacionados, la nueva generación de alineadores aprovecha la disponibilidad de múltiples estructuras experimentales dentro de un número creciente de familias de proteínas. Eso
se ha convertido en una práctica común combinar la salida de alineadores estructurales (por pares o múltiples) usando un marco basado en la consistencia (Figura 1). El principio es bastante directo e implica asociar cada secuencia con una plantilla que puede ser una estructura bonafide o una secuencia con una estructura conocida relacionada de manera cercana a la secuencia de interés, de modo que no exista ambigüedad en la alineación de la secuencia plantilla / objetivo.
Cuando no hay estructura disponible, las secuencias se reemplazan por
perfiles desde los cuales uno puede generar un perfil de conservación (como se hace en PSI-Coffee o TM-Coffee [37]) o una predicción de estructura secundaria usando PSIPRED [38], o ambos como se hace en Prommals3D [39], 3D-Coffee [40] y Expresso [41]. Luego, la biblioteca se construye alineando las secuencias en pares, utilizando el método de pares más adecuado para las plantillas consideradas. De esta manera, los métodos alternativos se pueden combinar a la perfección. Este enfoque es especialmente conveniente cuando se trata de métodos de alineación estructural por pares que carecen de una implementación de alineación múltiple. La posibilidad de combinar varios alineadores estructurales alternativos también proporciona una manera simple de abordar la dificultad de distinguir objetivamente los modelos de alineación de secuencias alternativos basados en estructuras. En este contexto, el enfoque basado en la coherencia permite identificar la porción de un modelo mejor respaldada por todos los métodos considerados. Esta
el enfoque se ha implementado en el paquete Expresso, que admite tres de los alineadores estructurales más comúnmente utilizados y puede acomodar fácilmente cualquier otro alineador de terceros.

Alineadores de secuencia múltiple de ARN

El alfabeto de baja complejidad de las moléculas de ARN hace que su alineación sea más desafiante que la de las secuencias de proteínas, con alineamientos biológicamente significativos difíciles de estimar <60% de identidad [42]. Siempre que las estructuras secundarias se conservan evolutivamente, la covariación a menudo se convierte en la más fuerte
señal disponible. Sin embargo, los alineadores estándar, como ClustalW, MAFFT o T-Coffee, asumen la independencia del sitio y no pueden tener en cuenta esta información, al menos en su uso predeterminado. De hecho, para estos alineadores estándar, la covariación es más un factor de confusión ya que disminuye la identidad de secuencia. Por lo tanto, se necesitan alineadores más especializados, capaces de reconocer simultáneamente la similitud en la secuencia y en el nivel de la estructura secundaria. Estos algoritmos son todas aproximaciones heurísticas, relacionadas más o menos explícitamente con el algoritmo de programación dinámica Sankoff [43], que simultáneamente pliega y alinea los ARN a un costo computacional prohibitivo O (N 3m), siendo m el número de secuencias y N su longitud . Se han informado varias implementaciones en bandas del algoritmo (Figura 1). Esto impone restricciones sobre el tamaño o la forma de las subestructuras; pueden ser alineadores por pares como Consan [44], Dynalign [45, 46], Stemloc [47] y Foldalign [48, 49] o alineadores múltiples como MXSCARNA [50], un alineador múltiple progresivo basado en SCARNA [51] , un método de alineación por pares basado en fragmentos de vástagos de longitud fija definidos por medio del algoritmo de McCaskill [52]. Murlet [53] es otro alineador de este tipo que estima

primero el emparejamiento de bases y las probabilidades de coincidencia antes de ejecutar el algoritmo de Sankoff con estas probabilidades para estimar la alineación final. En MARNA [54], la información estructural se utiliza para las comparaciones de ARN por pares antes de unirlas en un MSA con T-Coffee. PMcomp [55] es un método para alineamientos múltiples progresivos basado en un algoritmo de McCaskill para generar y luego comparar matrices de probabilidad de emparejamiento de bases, lo que permite un cálculo ligero. Para este propósito, utiliza un modelo de energía basado en un par de bases en lugar del modelo de energía basado en bucles original.

PMcomp simplifica el modelo de Sankoff al predecir solo una única estructura de consenso y ha sido una importante fuente de inspiración para el desarrollo de muchos alineadores de ARN, tales como LocARNA [56], FoldAlignM [57] y LocARNA-P [58] (Figura 1) que usan Heurística adicional para restringir aún más el espacio de plegado que se explorará, lo que resulta en una complejidad de tiempo O (n 4).

CARNA [59] extiende el modelo de PMcomp a las estructuras de pseudoknot (Figura 1). RAF [60] combinó las ideas de [61] y [55], dando como resultado una ligera variante de Sankoff con una velocidad basada en la secuencia. SPARSE [62] es uno de los algoritmos más nuevos en la categoría de estilo Sankoff. Es confiable. Se ejecuta en tiempo cuadrático gracias a su confianza en la predicción 'sparsified' y las alineaciones de RNA basadas en sus conjuntos de estructuras. En comparación con LocARNA, SPARSE logra una alineación similar y una mejor calidad de plegado en un tiempo significativamente menor (aceleración: 3.7). Otro enfoque que StrAl [63] implementa es un esquema de puntuación que combina la similitud de secuencia con la probabilidad de emparejamiento. Esta rápida heurística permite un tiempo de ejecución similar a ClustalW. T-Lara [64] implementa una representación basada en gráficos de alineaciones de secuencia-estructura modeladas usando programación lineal entera. Las alineaciones resultantes se integran luego en una biblioteca de estilo T-Coffee usando la relajación de Lagrange y finalmente se resuelven en un modelo MSA usando T-Coffee. RNAsampler es un algoritmo basado en muestras capaz de encontrar estructuras de ARN comunes en múltiples secuencias de ARN [65]. El programa muestrea probabilísticamente los vástagos de ARN alineados en función de las probabilidades de alineación de bases entre secuencias y la conservación del tallo calculada a partir de las probabilidades de apareamiento de bases intrasecuenciales. Otro ejemplo es RNAcast [66], que para cada secuencia predice perfiles de estructura dentro de un umbral de energía libre mínimo definido y luego calcula la estructura de consenso óptima que comparten todos los ARN.

Más recientemente, siguiendo a T-Lara, se hicieron intentos sistemáticos para aplicar el paradigma de consistencia a las predicciones de estructuras secundarias. Uno puede hacerlo considerando bibliotecas hechas de pares de residuos de emparejamiento. Este principio ha sido desarrollado en R-Coffee [67], que adopta un enfoque de pre-plegado, prediciendo con RNAplfold [68] la forma de las secuencias de ARN individuales en un primer paso. Posteriormente, el programa estima el MSA con la mayor concordancia entre estructuras y secuencias. Un enfoque similar se desarrolló más tarde en el

ARN versión compatible de MAFFT [69], donde se mide la consistencia mediante la combinación de pares de residuos emparejados a través de la combinación de trillizos. Ambos paquetes alcanzan niveles comparables de precisión, la principal fortaleza de R-Coffee es su capacidad para combinar alineadores complejos de RNA por pares como Consan en alineadores múltiples de alta precisión.

La escasez de información de ARN 3D probablemente explica por qué se ha prestado tan poca atención a la generación de alineaciones de ARN múltiples basadas en la estructura 3D precisa. La situación está cambiando lentamente con varios algoritmos nuevos recientemente descritos para tratar este problema. Las herramientas existentes incluyen alineadores por pares como ARTS [70], SARA [71], DIAL [72] y R3D Align [73], y múltiples como SARSA [74], LaJolla [75] y SARA- Coffee [76]. La naturaleza heurística de estos algoritmos tiende a hacerlos propensos a errores, de ahí la importancia de los editores de MSA específicos de ARN. Muchas de estas herramientas están disponibles (4SALE [77],

CONSTRUCT [78], JPHYDIT [79], RALEE [80], SARSE [81]) y pueden mostrar dinámicamente información secundaria y compensatoria mientras se editan RNA MSA.

Es importante observar que estos algoritmos solo funcionan bien cuando se trata de una estructura secundaria conservada evolutiva que contiene ARN. En su validación del algoritmo SARA-Café, Kemena et al. [76] informaron que cuando <70% de los nucleótidos están implicados en los pares de bases de Watson y Crick conservados evolutivamente, los alineadores sensibles a la estructura como los enumerados anteriormente tienden a degradar la precisión de la alineación. Esta degradación es una consecuencia mecánica del intento algorítmico explícito de buscar y unir estructuras secundarias bajo la suposición de que éstas deberían ser homólogas. En la práctica, sin embargo, las estructuras pueden no conservarse, o no predecirse adecuadamente, especialmente en los casos en que la interacción proteína / ARN desempeña un papel en el pliegue de ARN in vivo. Este problema es especialmente importante cuando se considera la cuestión de la alineación del ARN largo no codificante (lncRNA), la clase de genes de ARN descritos más recientemente [82]. Hasta el momento, no se ha informado de ninguna indicación de estructura secundaria ampliamente conservada para estos genes, lo que hace cada vez más probable que esta nueva categoría de transcripciones requiera una nueva generación de alineadores en los próximos años, posiblemente motivo sesgado y basado en el informe reciente que la información de dinucleótidos puede ayudar a mejorar las alineaciones de lncRNA [83, 84].

Se multiplican alineando secuencias no transcritas

La disponibilidad cada vez mayor de genomas completos hace que sea una necesidad apremiante desarrollar herramientas de alineación de secuencias intergénicas no transcritas (Figura 1). De hecho, estas secuencias presentan sus propios desafíos: longitud extrema, conservación deficiente, variaciones de orden (inversiones, translocaciones y duplicaciones) y la heterogeneidad extrema del reloj molecular resultante de la amplia gama de funciones soportadas de diferentes maneras por la parte no traducida del genoma . Es probable que este último tema sea cada vez más importante a medida que se informan nuevas funciones genómicas, a menudo asociadas con la epigenética [85, 86].

Alineamientos de genoma múltiple

Si bien los alineadores de secuencia estándar generalmente implican el modelado de tres operaciones evolutivas, inserción, eliminación y sustituciones, las alineaciones a escala del genoma deben incorporar al menos tres operaciones más: inversiones, translocaciones y duplicaciones. En general, los alineadores de genoma múltiple logran esto a través de dos pasos separados. En un primer paso, los fragmentos genómicos homólogos se clasifican en contenedores, y en un segundo paso, estos contenedores se convierten en modelos estándar de MSA. Este último paso generalmente depende de alineadores progresivos estándar, algorítmicamente similares a los descritos en la primera parte de esta revisión.

Si bien el primer paso de alineación podría basarse en un enfoque de agrupamiento / segmentación simple, dicho procedimiento produciría bloques de MSA desconectados, lo que daría pocas pistas sobre la evolución genómica. Por este motivo, la mayoría de los alineadores de genoma de nueva generación se basan en el algoritmo de clasificación por inversión para el paso de segmentación. La clasificación por inversión es un problema NP-completo que equivale a reconstruir la cadena mínima de eventos que editaría un genoma en otro utilizando una serie de translocaciones e inversiones [87]. No es necesario resolver este problema para alinear genomas, pero ayuda a cuantificar el costo evolutivo de alineaciones alternativas. En la práctica, la mayoría de los algoritmos comienzan buscando segmentos colineales, a menudo confiando en los puntos de anclaje (generalmente proteínas) reunidos mediante un procedimiento de BLAST todo-contra-todo. Los procedimientos más populares incluyen Mercator [88] que usa anclajes de proteínas, MUMS / Mems (Mugsy [89] o el sistema
uso tematico de alineaciones locales [90].

Se multiplican alineando secuencias no transcritas

La disponibilidad cada vez mayor de genomas completos hace que sea una necesidad apremiante desarrollar herramientas de alineación de secuencias intergénicas no transcritas (Figura 1). De hecho, estas secuencias presentan sus propios desafíos: longitud extrema, conservación deficiente, variaciones de orden (inversiones, translocaciones y duplicaciones) y la heterogeneidad extrema del reloj molecular resultante de la amplia gama de funciones soportadas de diferentes maneras por la parte no traducida del genoma . Es probable que este último tema sea cada vez más importante a medida que se informan nuevas funciones genómicas, a menudo asociadas con la epigenética [85, 86].

Alineamientos de genoma múltiple

Si bien los alineadores de secuencia estándar generalmente implican el modelado de tres operaciones evolutivas, inserción, eliminación y sustituciones, las alineaciones a escala del genoma deben incorporar al menos tres operaciones más: inversiones, translocaciones y duplicaciones. En general, los alineadores de genoma múltiple logran esto a través de dos pasos separados. En un primer paso, los fragmentos genómicos homólogos se clasifican en contenedores, y en un segundo paso, estos contenedores se convierten en modelos estándar de MSA. Este último paso generalmente depende de alineadores progresivos estándar, algorítmicamente similares a los descritos en la primera parte de esta revisión.

Si bien el primer paso de alineación podría basarse en un enfoque de agrupamiento / segmentación simple, dicho procedimiento produciría bloques de MSA desconectados, lo que daría pocas pistas sobre la evolución genómica. Por este motivo, la mayoría de los alineadores de genoma de nueva generación se basan en el algoritmo de clasificación por inversión para el paso de segmentación. La clasificación por inversión es un problema NP-completo que equivale a reconstruir la cadena mínima de eventos que editaría un genoma en otro utilizando una serie de translocaciones e inversiones [87]. No es necesario resolver este problema para alinear genomas, pero ayuda a cuantificar el costo evolutivo de alineaciones alternativas. En la práctica, la mayoría de los algoritmos comienzan buscando segmentos colineales, a menudo confiando en los puntos de anclaje (generalmente proteínas) reunidos mediante un procedimiento de BLAST todo-contra-todo. Los procedimientos más populares incluyen Mercator [88] que utiliza anclajes de proteínas, MUMS / Mems (Mugsy [89] o el uso sistemático de alineaciones locales [90].

TBA [91] fue uno de los primeros algoritmos para considerar una alineación de genoma múltiple (MGA) como un conjunto de bloques separados en lugar de una secuencia continua, por lo que el procesamiento de datos es un requisito previo necesario (Figura 1). En la generación más nueva de MGA, el preprocesamiento se ha integrado estrechamente con el proceso de alineación, como en Mercator-Mavid [88] o Enredo / Pecan [92], que utiliza estructuras de gráficos (Figura 1) para identificar los diferentes reordenamientos del genoma , divide los genomas múltiples en consecuencia y alimenta los contenedores resultantes de múltiples secuencias a Pecan, un alineador basado en la consistencia del espacio utilizando el procedimiento de programación dinámica de espacio lineal solo hacia adelante de Durbin [22]. Se han utilizado otras estructuras gráficas (por ejemplo, gráfico A-Brujin [93], gráfico de Cactus [94]) (véase Kehr et al. [95] para una comparación de diferentes estructuras gráficas).

Otra alternativa es llevar a cabo simultáneamente la alineación y la segmentación de forma progresiva. Este procedimiento desarrollado por Brudno [90] utiliza el equivalente de consistencia para identificar reordenamientos que probablemente sean respaldados por el conjunto de datos completo.

El desarrollo del método MGA, sin embargo, se ha visto obstaculizado por la dificultad de evaluar objetivamente los méritos relativos de cada alineador. En contraste con las proteínas o secuencias de ARN, no existe una estructura o su equivalente disponible para los genomas, y cuando el concurso Alignathon [96] propuso comparar las capacidades de los MGA en datos eucarióticos, la evaluación comparativa finalmente se llevó a cabo utilizando el PSAR. función objetivo [97], un estimador basado en secuencias que se basa en el muestreo probabilístico.

La función objetivo PSAR se desarrolló inicialmente para evaluar los MSA genómicos. Su principio es de alguna manera similar al enfoque basado en la consistencia de T-Coffee, aunque más completo y más exigente en términos de computación. En PSAR, dado un conjunto de datos, todas las secuencias se eliminan sucesivamente, las secuencias restantes se realinean y la secuencia eliminada se realinea a la subalineación. La estabilidad de la realineación con respecto a la entrada MSA se usa luego para estimar la confiabilidad de cada posicionamiento de residuo dentro del modelo de alineación final. Este procedimiento es genérico sin restricción limitándolo a las alineaciones de nucleótidos.
Sin embargo, hasta ahora solo se ha probado y evaluado en conjuntos de datos genómicos simulados.

El concurso de Alignathon sigue siendo el único intento genérico de comparar la fiabilidad de alineadores genómicos múltiples. Se compararon 13 paquetes de MGA en genomas de Drosophila o en genomas de mamíferos generados artificialmente. Como lo señalaron los propios autores, un tema importante en este trabajo es el diseño de un estándar de verdad aceptable. Los coordinadores de Alignathon tomaron la decisión de usar la función objetivo PSAR como un estándar de verdad. Dicha decisión viene con una advertencia importante, posiblemente reflejada en el claro dominio de PSAR-align-un paquete que optimiza explícitamente esta función-sobre la mayoría de los alineadores alternativos. De mayor relevancia es sin duda la medida tomada por los autores del acuerdo entre alineadores. En el análisis de índice de Jaccard correspondiente, encontraron que en los genomas de moscas no simulados, más del 50% de las posiciones alineadas a los niveles de nucleótidos son inconsistentes entre un par de métodos (Figuras 8B y C en el informe considerado). Tal dispersión debe tomarse como una medida de la complejidad que uno enfrenta al tratar de desarrollar un alineador de ADN genérico. En contraste, se puede usar un esfuerzo más enfocado en regiones genómicas bien definidas para entregar alineaciones de alta calidad de regiones funcionalmente homólogas. Este enfoque se ha desarrollado con éxito y se ha utilizado para estudiar promotores del genoma eucariótico.

Múltiples alineaciones de promotores

Los MGA apuntan a utilizar la información de reordenamiento del genoma para comprender mejor las relaciones evolutivas y posiblemente identificar las limitaciones funcionales asociadas con la conservación de la organización genética. En este contexto, las comparaciones múltiples entre promotores son probablemente el mejor ejemplo de alineamientos múltiples funcionales, con el objetivo de descubrir patrones regulatorios comunes entre secuencias relacionadas. Estos patrones se usan para revelar los sitios de unión del factor de transcripción (TFBS). Desde un punto de vista algorítmico, el problema se puede separar en dos categorías distintas: descubrimiento de motivo entre las secuencias desalineadas no homólogas (o secuencias lejanas relacionadas) y las MSA regulares. Las técnicas de motiffinding relevantes para el análisis de promotores se han revisado exhaustivamente en (ver, por ejemplo [98, 99] para revisiones), y su descripción está más allá del alcance de esta revisión.
Los métodos para el descubrimiento y comparación de regiones promotoras homólogas son más recientes. Inicialmente se informaron para el descubrimiento de TFBS, a través de un proceso que a menudo se denomina impresión evolutiva del pie. Se han descrito varios métodos para tal fin. Por ejemplo, en [100], los sitios potenciales de unión se predicen primero en secuencias únicas y luego se usan como anclajes durante el proceso de alineación. Otra estrategia es utilizar un esquema de puntuación alternativo en las posiciones dentro de una secuencia que se sabe que se ajusta a un elemento regulador [101]. Esto equivale más o menos a vestir una secuencia con matrices de peso de perfil que definen un esquema de puntuación de posición específica. La principal limitación, sin embargo, de estos métodos basados en motivos es su dependencia de conjuntos precomputados de motivos de referencia. Como alternativa, uno puede identificar los motivos simultáneamente y alinear las secuencias como se propone en [102, 103]. Otros métodos también pueden modelar inversiones y translocaciones, teniendo en cuenta la rotación de motivos rápidos informada en las regiones promotoras [106]. Todos estos métodos son computacionalmente demasiado intensivos para escalar en unas pocas (generalmente dos) secuencias, y se han propuesto alternativas escalables para el análisis de

secuencias múltiples [104]. También es posible afinar los métodos existentes para múltiples alineamientos de promotores, como lo muestran Erb et al. [105]. En este trabajo, los autores optimizaron tres métodos populares (MAFFT, Muscle, T-Coffee) por su capacidad para alinear de forma efectiva TFBS homóloga probada experimentalmente. La sintonización también tuvo en cuenta la capacidad discriminativa entre las alineaciones de las regiones de genes ortólogos y parálogos.

Evaluación comparativa de la precisión de alineadores múltiples

Cuantificar la precisión de alineadores múltiples es tan importante como alinear secuencias, especialmente cuando se considera la naturaleza aproximada de los alineadores. Este aspecto aparentemente obvio ha sido generalmente pasado por alto por la comunidad como se refleja en la relativa falta de correlación entre el uso general de los paquetes y su precisión informada. ClustalW, por ejemplo, cuyas 42 000 citas sugieren un nivel de uso global más alto que todos los demás paquetes reunidos, no se ha informado consistentemente como el método más preciso. Esta observación sorprendente probablemente se refleja en una combinación de factores. La más obvia es la relación entre las clasificaciones de los puntos de referencia y la usabilidad cotidiana. Es probable que ClustalW, a pesar de que no ocupa el lugar # 1 en todos los puntos de referencia, sea lo suficientemente preciso para muchas actividades de modelado, especialmente cuando se trata de conjuntos de datos ortólogos. También se puede especular sobre la existencia de una fuerte inercia metodológica dentro de la comunidad biológica, donde el uso de herramientas tiende a aumentar a través del protocolo de reciclaje.

El componente más crítico de un MSA es su función de puntuación / objetivo, la fórmula matemática que cuantifica el puntaje total y, por lo tanto, define la optimalidad, dado un conjunto de secuencias.

El resto del algoritmo es un procedimiento de optimización que intenta generar un modelo MSA que maximiza la función objetivo. Está bien establecido que incluso las mejores funciones objetivas son simplemente aproximaciones que intentan modelar el comportamiento de las secuencias biológicas [107]. Como consecuencia, no hay garantía de que un MSA perfectamente optimizado sistemáticamente dará como resultado el MSA biológicamente más significativo. Esta es la razón por la cual los alineadores múltiples también necesitan ser evaluados / comparados por su capacidad para producir alineaciones correctas. Un procedimiento de evaluación comparativa se basa en colecciones existentes de alineaciones de referencia consideradas como patrones oro. Estas MSA de referencia se usan rutinariamente como predictores de la precisión de un alineador dado en un tipo determinado de conjuntos de datos y han tenido una gran influencia en los desarrollos metodológicos. Las colecciones de referencia de proteína existentes se revisaron de forma extensa y crítica en [108] y [109] cuando los autores proponen agrupar puntos de referencia en cuatro categorías: basada en simulación, basada en consistencia, basada en estructura y basada en filogenia. Las últimas tres categorías cumplen el criterio de los conjuntos de datos de referencia, ya que pueden precompilarse y utilizarse para cuantificar los méritos relativos de un alineador sobre otro. Los puntos de referencia basados en la simulación, sin embargo, definen una función objetivo en lugar de un procedimiento de referencia y no pueden considerarse una medida de referencia en el mismo sentido que los demás.

Parámetros de proteína basados en estructura

La evaluación comparativa múltiple de alineadores ha sido impulsada en gran medida por el uso de MSA de referencia basadas en estructuras, BAliBASE [110] es el más utilizado. Todos estos puntos de referencia se basan en alineamientos de referencia basados en la estructura para evaluar cualquier alineador capaz de manejar sus secuencias (Figura 2). Se ha vuelto habitual informar nuevos alineadores junto con las lecturas de referencia establecidas en al menos dos conjuntos de datos de referencia basados en estructuras disponibles. Esto probablemente se debe a la sorprendentemente baja consistencia entre los puntos de referencia. De hecho, como se muestra en [4], las clasificaciones de alineador establecidas sobre la base de los puntos de referencia más comunes son en promedio <50% consistente. Esto significa que si un punto de referencia dado sugiere que el método A es más preciso que el método B, hay menos de una posibilidad de que el mismo ranking sea respaldado por otra

colección de referencia. En su análisis detallado de los puntos de referencia disponibles, Edgar sugirió SABmark [111] para ser el más completo e informativo, pero solo cuando se usa un subconjunto de SABmark hecho de alineamientos estructurales compatibles por pares.
La creciente necesidad de alineadores a gran escala ha resultado en el desarrollo de una nueva generación de referencia capaz de estimar la precisión de alineación al ensamblar grandes conjuntos de datos. El problema principal al hacerlo es la escasez de información estructural. De las 16 familias de 230 Pfam con información estructural experimental, aproximadamente el 50% solo tiene un miembro con una estructura tridimensional conocida, y el 25% solo tiene dos miembros. Para acomodar esta limitación, los conjuntos de datos de referencia se construyeron mediante incrustaciones de secuencias con una estructura conocida dentro de conjuntos de datos más grandes formados por secuencias con una estructura desconocida. Este enfoque ya utilizado en PREFAB [12] -con dos secuencias de estructura conocida integrada en un conjunto de datos de 50 secuencias- se ha extendido en HomFam [14, 112], para definir conjuntos de datos mucho más grandes de hasta 100 000 secuencias en el que se incrusta un promedio de 10 secuencias con estructuras conocidas. Al hacerlo, la precisión se estima alineando primero los grandes conjuntos de datos. Luego se extraen las proyecciones de secuencias con estructuras conocidas y se cuantifica la precisión comparando estas proyecciones con la referencia. En este procedimiento, la advertencia principal radica en la suposición de que la precisión de la secuencia de semillas refleja bien el conjunto de datos globales. Esta suposición, sin embargo, solo es correcta si las secuencias con estructuras conocidas están distribuidas uniformemente dentro del conjunto de datos considerado.
La evaluación comparativa basada en la estructura no depende necesariamente de una alineación de referencia, y también se han diseñado métodos alternativos que se basan en la superposición estructural en lugar de las alineaciones estructurales inducidas por la superposición. Estos desarrollos fueron principalmente consecuencia del trabajo de Lackner [114], que informó sobre situaciones en las que la superposición basada en la estructura es lo suficientemente ambigua como para soportar igualmente varias alineaciones de secuencia alternativas. Cuando esto ocurre, la alineación de referencia se convierte en la priorización arbitraria de una referencia sobre otra, lo que sesga el proceso de referencia. La mayoría de los puntos de referencia se ocupan de este problema especificando las regiones centrales en las que se espera que el alineamiento de referencia sea menos ambiguo, pero este procedimiento sigue dependiendo de la forma en que se definan las regiones centrales.
Existe una alternativa más general que implica comparar distancias intramoleculares entre pares de pares de residuos alineados. Esta medida, llamada iRMSD [115], permite cuantificar el ajuste estructural implicado por una alineación sin tener que depender de una referencia.

Referencia de alineación de ARN basada en la estructura
Los puntos de referencia estructurales también se han desarrollado para la evaluación de la alineación del ARN (Figura 2). Tres de tales puntos de referencia existen. BAliBASE [116] es el más comúnmente utilizado. Permite evaluar la precisión de un alineador múltiple en secuencias de ARN al considerar la capacidad de modelado del alineador evaluado con respecto a alguna estructura secundaria de referencia. Esta dependencia de la secuencia (en la que se basa la estimación de la estructura secundaria) limita ligeramente su alcance, ya que implica dependencias comunes entre la compilación de referencia y el procedimiento de evaluación. BraliDart [76], un conjunto de datos más reciente, que se basa únicamente en información estructural y contiene conjuntos de familias homólogas de ARN con estructuras experimentales conocidas, ha sido reportado recientemente. Este conjunto de datos está limitado por la relativa escasez de estructuras experimentales de ARN 3D. Otra especificidad de BraliDart es su falta de confianza en un alineamiento estructural de referencia, sino más bien en el ajuste estructural implícito en la alineación de la secuencia utilizando una medida de distancia RMSD, tal como se define en el método iRMSD. La tercera categoría principal de referencia de ARN está hecha de alineaciones de referencia de ARN ribosómico [113].

No se han ensamblado con fines de evaluación comparativa, sino más bien como consecuencia de la importancia de las alineaciones precisas del ARN ribosomal (ARNr) al estimar el árbol de la vida.
Estas alineaciones se han realizado manualmente, teniendo en cuenta las estructuras secundarias de ARNr altamente conservadas que desempeñan papeles críticos en las capacidades funcionales de los ribosomas. En el momento en que escribimos esta revisión, aún no se ha publicado ningún conjunto de datos de referencia para validar los MSA de ARN largos no codificantes, un documento recientemente publicado de población descrita de transcripciones.

Conjuntos de datos simulados para el análisis evolutivo

Aunque los puntos de referencia de datos empíricos son las estrategias más utilizadas para evaluar los métodos de alineación, siguen siendo limitados por su dependencia de los datos estructurales y la falta de tales datos para la evaluación de ciertos tipos de alineaciones, como el ADN no transcrito. Además, queda por establecer hasta qué punto se puede esperar que las alineaciones basadas en la estructura sean evolutivamente correctas. Esta pregunta es especialmente crítica teniendo en cuenta que el modelado filogenético es una de las principales aplicaciones del modelado de MSA. Un problema importante de los métodos alineadores más populares es su dependencia sistemática, y la posibilidad de ajuste en alineamientos de secuencia estructuralmente correctos. Sin embargo, estos métodos a menudo se usan para llevar a cabo una reconstrucción filogenética. Esta inconsistencia ha sido señalada por la comunidad evolutiva, que rutinariamente se basa en conjuntos de datos simulados en lugar de los empíricos [117].
Los conjuntos de datos simulados se basan en modelos que imitan la evolución para generar secuencias cuya diversidad se espera que represente un verdadero proceso evolutivo. La principal fortaleza de este enfoque es proporcionar un modelo perfectamente rastreable, en el que se conoce explícitamente la relación entre nucleótidos o aminoácidos.
Su inconveniente más obvio es confiar en modelos evolutivos que se supone que son correctos, mientras que la verdadera medida en que representan escenarios biológicamente realistas sigue siendo desconocida. En cualquier caso, estos enfoques son útiles al estimar el impacto de condiciones extremas en la capacidad de modelado, por ejemplo, la evolución acelerada, la atracción de ramas largas y efectos similares que pueden confundir el análisis estándar. Varios paquetes han sido diseñados para generar conjuntos de datos simulados (Figura 2), siendo los más utilizados Rose [118], Seq-Gen [119], Dawg [120] o INDELible [121].
Al usar estos paquetes, las alineaciones simuladas se consideran alineación "verdadera", lo que permite utilizar el mismo sistema de puntuación (puntuación de suma de pares, SP o puntuación de columna, CS [122]) como para los puntos de referencia empíricos. Vale la pena señalar que cada vez que se utilizan conjuntos de datos de referencia simulados y estructurados para validar algoritmos similares para la precisión de alineación, se encontró que los rankings difieren significativamente entre estos dos grupos de puntos de referencia, una clara indicación de que se están evaluando diferentes características de alineación [4, 123].
Todos los alineadores con reconocimiento de filogenia se evalúan actualmente utilizando estos conjuntos de datos simulados. Al hacerlo, la evaluación se realiza a menudo en la capacidad de modelado de árboles en lugar de hacerlo en el propio MSA. Tales algoritmos incluyen [117, 124-126].
Resolver las aparentes discrepancias entre los conjuntos de datos de referencia simulados y basados en estructuras probablemente requerirá una mejor comprensión de la relación compleja entre la precisión de alineación y la reconstrucción filogenética confiable.
Avanzando un paso en esta dirección, Dessimoz y Gil introdujeron recientemente pruebas basadas en árboles de precisión de alineación, que no solo utilizan muestras grandes y representativas de datos biológicos reales, sino que también permiten evaluar el efecto de la colocación de brechas en la inferencia filogenética [127] . En un trabajo no relacionado [35], Chang y sus coautores propusieron el uso de conjuntos de datos empíricos obtenidos mediante el enriquecimiento de colecciones de genes

ortólogos en familias que probablemente apoyen el Árbol de la Vida. Al usar tales conjuntos de datos, la discrepancia entre la corrección filogenética y estructural parece ser menos marcada.

Índices de calidad para la estimación de la confiabilidad de MSA

Dada la naturaleza aproximada de todos los alineadores disponibles, identificar las partes confiables de una alineación es probablemente de mayor importancia práctica que conocer la precisión general esperada. En los últimos años, se han reportado nuevos métodos apuntando precisamente a esto (Figura 3). Se pueden dividir aproximadamente en tres categorías: los que usan información estructural para evaluar la exactitud de proteína o ARN, los que dependen de un índice de conservación para identificar las posiciones más correctas y los métodos que dependen de alguna forma de inestabilidad numérica local para identificar los más estables porciones de un modelo MSA.

Índices de conservación estructural

Con los datos estructurales cada vez más disponibles, el uso sistemático de la información en 3D para el control de la precisión de MSA se está convirtiendo lentamente en una perspectiva realista. Los primeros métodos de este tipo [41, 128] se diseñaron utilizando la precisión estructural medida en todos los pares posibles de secuencias con una estructura 3D conocida como proxy para la precisión global. Este enfoque es útil, pero adolece de la principal limitación: el número limitado de familias de proteínas y ARN para el que hay más de una estructura disponible (aproximadamente el 25% de todas las familias de PFAM con estructura conocida y <1% en RFAM). Por lo tanto, los esfuerzos recientes se centraron en el uso de estructuras únicas para estimar la precisión de MSA. La matriz de sustitución de contacto CAO [129] es uno de los primeros trabajos en esta dirección. El principio es incrustar una secuencia con una estructura conocida en el MSA. Esta estructura luego se usa para identificar putativos contactos de aminoácidos y las columnas correspondientes son reevaluadas utilizando la CAO, una 400? Matriz de sustitución 400 que asigna un puntaje a cada posible sustitución de contacto. Desafortunadamente, la estimación de esta matriz está limitada por la falta de datos disponibles. Este problema fue abordado por el algoritmo STRIKE [130], en el que la matriz de sustitución de contacto se reemplaza por una métrica de potencial de contacto que considera la puntuación de todos los contactos potenciales, tal como se obtiene de los datos estructurales. Al usar esta matriz para evaluar un MSA, los contactos de columna, como lo implica al menos una estructura incrustada, se evalúan sumando el puntaje de contacto que se encuentra en la matriz de registro de contacto impar. Este enfoque demostró ser significativamente superior al CAO como un medio para discriminar entre alineamientos alternativos.

Índices de conservación de secuencia

La conservación de secuencias es una de las formas más sencillas de estimar la precisión de MSA. Se ha desarrollado una gran cantidad de herramientas para este propósito que se dividen aproximadamente en dos categorías principales: estructural (es decir, estimaciones estructurales que usan información de secuencia) e índices evolutivos. Los índices evolutivos apuntan a identificar dentro de un MSA todas las posiciones que puedan obstaculizar la reconstrucción filogenética. Estos índices generalmente se centran en la eliminación de diversas columnas o regiones indelente enriquecidas.

Las herramientas más comúnmente usadas son Gblocks [131,132] y trimAl [136], una reimplementación de Gblocks usando un procedimiento de parametrización automatizado para ajustar el nivel de filtrado. Si bien estas herramientas son extremadamente populares y forman parte de muchas tuberías filogenéticas a gran escala, el valor real del filtrado de columnas sigue siendo un punto de discusión. Dos informes recientes sugieren que el filtrado podría disminuir el potencial de modelado filogenético de MSA [28, 35]. Se han desarrollado herramientas similares para estimar la corrección estructural de las MSA proteicas. Los más simples como AL2CO [133] simplemente miden la conservación de acuerdo con varios criterios fisicoquímicos. A las columnas y residuos eventualmente se les asigna un valor de índice que se puede usar al modelar. Las variaciones más sofisticadas incluyen RASCAL [137], un procedimiento de escaneo de MSA destinado a identificar regiones espurias dentro de un MSA.

Índices de estabilidad de alineación

Los paquetes MSA más utilizados se basan en una combinación entre el algoritmo progresivo y las implementaciones de programación dinámica más o menos sofisticadas, lo que permite alineamientos por parejas de secuencias o perfiles. Estas dependencias hacen que estos algoritmos sean inherentemente inestables. En los últimos años, el desarrollo de métodos capaces de cuantificar esta inestabilidad para estimar la confiabilidad local se ha convertido en una tendencia de rápido crecimiento. La idea de usar la solidez como un indicador de precisión biológica no es nueva y ya se había utilizado ya en 1996 [138] en un procedimiento que incluía eliminar a su vez cada par de aminoácidos en un par de secuencias antes de realinearlos, a fin de evaluar la estabilidad de alineación local. Más tarde, se usó la función objetivo T-Coffee [107] para mostrar el poder predictivo de la consistencia. En general, cualquier procedimiento que pueda usarse para perturbar una alineación se presta a la definición de un índice de robustez.

Dichos índices se pueden evaluar para su correlación con el potencial de modelado estructural o filogenético. El procedimiento Head o Tail (HoT) [134] es un buen ejemplo de un método simple (las secuencias simplemente se invierten), produciendo información útil a costa de una sobrecarga computacional moderada. Se han descrito otros procedimientos similares, aunque más costosos. PSAR es uno de ellos [97].

Es un método que implica la generación de varias MSA alternativas al tiempo que elimina cada secuencia. Otro procedimiento de este tipo se denomina ORIENTACIÓN [135, 139], donde el MSA se vuelve a estimar varias veces utilizando árboles de guía estimados a partir de réplicas de arranque del MSA original. El problema principal con

estos dos enfoques es su costo computacional relativamente alto. Sin embargo, estos métodos son mucho más informativos que sus alternativas de conservación de secuencias. En un informe reciente, se ha demostrado que el puntaje de consistencia de T-Coffee [35] supera a HoT y GUIDANCE para la identificación de porciones estructuralmente correctas dentro de MSA y trimAL y Gblocks para la construcción de árboles filogenéticos precisos.

Conclusión

Esta revisión es un intento de poner en contexto y cubrir los desarrollos que han tenido lugar en el campo de las MSA en la última década más o menos. El ritmo de desarrollo sin precedentes hace que sea difícil ser realmente exhaustivo. No obstante, hemos intentado proporcionar al lector una visión general de los aspectos principales y cómo se conectan entre sí. Como se muestra en la Figura 1, el marco de alineamiento progresivo (alineando las secuencias siguiendo un orden de árbol) es la principal heurística algorítmica que ha sido adoptada por casi todos los métodos de alineación existentes. Además, podemos observar una agrupación clara de los métodos en función del tipo de secuencias para las que están diseñados para alinearse (ARN, ADN / genomas o proteínas). También vale la pena señalar que la inflación actual en el número de métodos disponibles simplemente refleja el ritmo creciente de la acumulación de datos. El modelado de MSA es una de las formas más poderosas para dar sentido a las secuencias biológicas. Las MSAM, por su naturaleza aproximada, están condenadas a seguir una estrategia evolutiva de reina roja y deberán evolucionar, cada vez más rápido, para mantenerse al día con el procesamiento de datos biológicos estándar.

Figura 1. Principales componentes algorítmicos de los alineadores múltiples más ampliamente utilizados. En el mapa de calor, las entradas de color naranja indican una característica implementada en el método. Tanto los alineadores como los componentes se agruparon por similitud utilizando el paquete R.

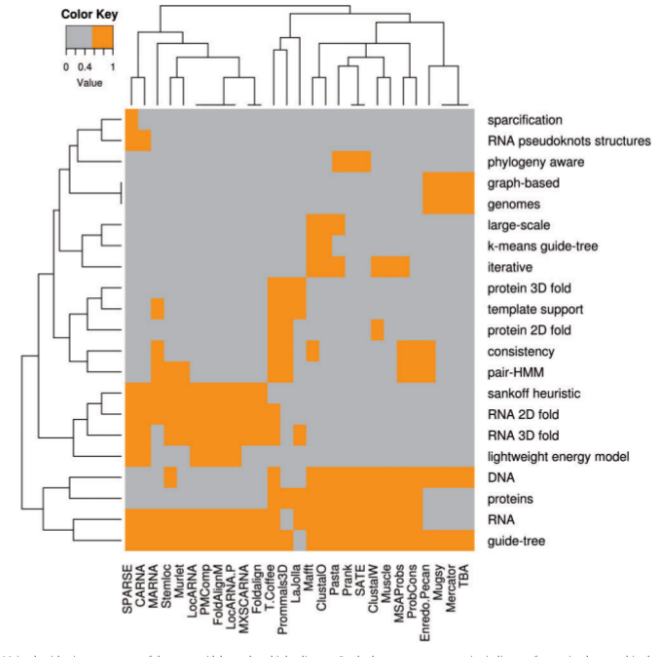

Figura 2. Principales métodos de referencia y sus propiedades más relevantes. En el mapa de calor, las entradas de color naranja indican una propiedad que describe un método determinado. Ambas propiedades y puntos de referencia se agruparon por similitud.

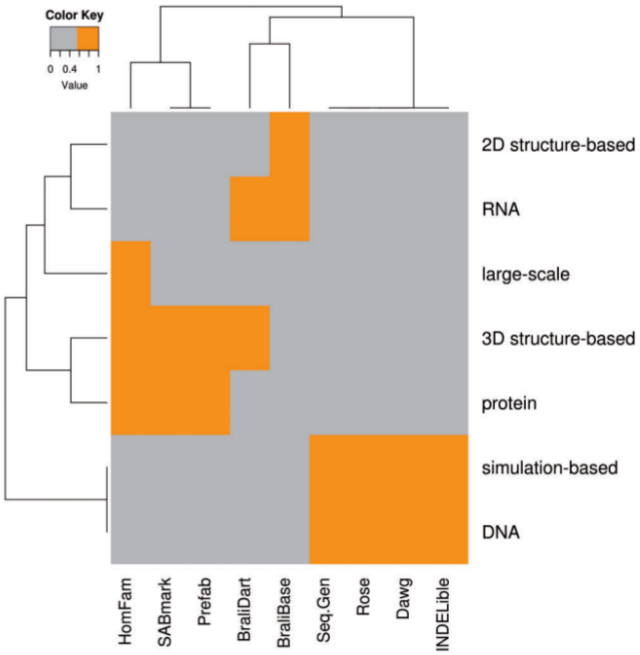

Figura 3. Índices de calidad de MSA y sus características. Las características con cero no son utilizadas por el índice de calidad específico.