TÉCNICAS INTELIGENTES EN BIOINFORMÁTICA

Alineamiento múltiple de secuencias

Mario de J. Pérez Jiménez
Grupo de investigación en Computación Natural
Dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla

Máster Universitario en Lógica, Computación e Inteligencia Artificial

Curso 2014-15

Introducción (I)

Hasta ahora hemos visto:

- Cómo comparar pares de secuencias.
- Usando la el programa BLAST podemos comparar una secuencia con las secuencias de una cierta base de datos.
- Ahora se trata de comparar un conjunto de secuencias a fin de encontrar propiedades interesantes entre ellas.

Introducción (II)

Recordemos que las secuencias biológicas se agrupan en familias:

- Genes relacionados de un mismo organismo (parálogos).
- Genes relacionados de distintas especies (ortólogos).
- Secuencias dentro de una misma población (variantes polimórficas).

Introducción (II)

Recordemos que las secuencias biológicas se agrupan en familias:

- Genes relacionados de un mismo organismo (parálogos).
- Genes relacionados de distintas especies (ortólogos).
- Secuencias dentro de una misma población (variantes polimórficas).

Nota 1: Puede suceder que un par de secuencias no tengan un buen alineamiento entre ellas y, en cambio, sí los tenga con una tercera.

Introducción (II)

Recordemos que las secuencias biológicas se agrupan en familias:

- Genes relacionados de un mismo organismo (parálogos).
- Genes relacionados de distintas especies (ortólogos).
- Secuencias dentro de una misma población (variantes polimórficas).

Nota 1: Puede suceder que un par de secuencias no tengan un buen alineamiento entre ellas y, en cambio, sí los tenga con una tercera.

Nota 2: Las secuencias suelen evolucionar más rápidamente que sus funcionalidades o estructuras.

Alineamiento múltiple de secuencias (MSA)

Mecanismo de alineamiento simultáneo de un conjunto de secuencias: permite realizar análisis diversos que van desde la **filogenia** a la **búsqueda de motivos**.

Alineamiento múltiple de secuencias (MSA)

Mecanismo de alineamiento simultáneo de un conjunto de secuencias: permite realizar análisis diversos que van desde la **filogenia** a la **búsqueda de motivos**.

- Revela mucha más información que los alineamientos de pares.
- Problema de eficiencia (tiempo de cálculo y memoria requerida: exponencial en el tamaño de las secuencias).
- Recurrir a enfoques heurísticos.

Alineamiento múltiple de secuencias (MSA)

Mecanismo de alineamiento simultáneo de un conjunto de secuencias: permite realizar análisis diversos que van desde la **filogenia** a la **búsqueda de motivos**.

- Revela mucha más información que los alineamientos de pares.
- Problema de eficiencia (tiempo de cálculo y memoria requerida: exponencial en el tamaño de las secuencias).
- Recurrir a enfoques heurísticos.

Para realizar un alineamiento múltiple hay que seleccionar:

- Las secuencias homólogas a alinear.
- El software adecuado que utilice una "buena" función de puntuación.
- Los parámetros necesarios.

MSA: Formalización

Sean $s^{(1)}, \ldots, s^{(p)}$ una lista de p cadenas sobre un alfabeto Γ (proteínas, bases nucleótidas, etc.). Un **alineamiento múltiple** de las cadenas $s^{(1)}, \ldots, s^{(p)}$ es una matriz $\mathbf{A} = (a_{ij})$ de orden (p, q), con $q \geq \text{máx}\{|s^{(1)}|, \ldots, |s^{(p)}|\}$ tal que:

- Para cada i, j $(1 \leq i \leq p, 1 \leq j \leq q)$ se tiene que $a_{ij} \in \Gamma \cup \{-\}$.
- Para cada i $(1 \leq i \leq p)$ la sucesión $(a_{i1}, \ldots, a_{iq})_{|\Gamma}$ es la cadena $s^{(i)}$.
- Para cada j $(1 \leq j \leq q)$ se tiene que existe i $(1 \leq i \leq p)$ tal que $a_{ij} \in \Gamma$ (es decir, no existen columnas que sólo contenga el símbolo $-$).

MSA: Formalización

Sean $s^{(1)}, \ldots, s^{(p)}$ una lista de p cadenas sobre un alfabeto Γ (proteínas, bases nucleótidas, etc.). Un **alineamiento múltiple** de las cadenas $s^{(1)}, \ldots, s^{(p)}$ es una matriz $\mathbf{A} = (a_{ij})$ de orden (p, q), con $q \geq \text{máx}\{|s^{(1)}|, \ldots, |s^{(p)}|\}$ tal que:

- Para cada i, j $(1 \leq i \leq p, 1 \leq j \leq q)$ se tiene que $a_{ij} \in \Gamma \cup \{-\}$.
- Para cada i $(1 \leq i \leq p)$ la sucesión $(a_{i1}, \ldots, a_{iq})_{|\Gamma}$ es la cadena $s^{(i)}$.
- Para cada j $(1 \leq j \leq q)$ se tiene que existe i $(1 \leq i \leq p)$ tal que $a_{ij} \in \Gamma$ (es decir, no existen columnas que sólo contenga el símbolo $-$).

Un ejemplo: 6 secuencias $s^{(1)}, \ldots, s^{(6)}$ tales que $|s^{(1)}| = 11, |s^{(2)}| = 10$, $|s^{(3)}| = 12, |s^{(4)}| = 12, |s^{(5)}| = 12, |s^{(6)}| = 11$ (p=6 y q=14).

$$\mathbf{A} = \begin{pmatrix} A & T & - & A & A & C & C & T & - & C & G & - & T & G \\ G & A & A & - & A & G & - & T & - & C & C & C & A & - \\ A & T & - & A & A & C & T & T & G & A & G & - & T & T \\ C & T & G & A & A & C & C & T & - & C & G & C & - & G \\ A & G & T & A & A & G & A & - & A & - & C & C & T & G \\ T & T & T & A & A & C & - & T & - & C & G & - & G & C \end{pmatrix}$$

MSA: Aplicaciones

Entre las aplicaciones del alineamiento múltiple de secuencias destacan:

- Dar información acerca de la estructura, función y evolución de una secuencia.
- Encontrar miembros distantes de una familia de proteínas con funcionalidades comunes.
- Generación de bases de datos de proteínas, una vez secuenciado el genoma completo de un organismo vivo.
- Predicción de estructuras secundarias y terciarias de proteínas.
- En un primer y muy importante paso en la generación de árboles filogenéticos.

MSA: aproximaciones algorítmicas (I)

- Métodos exactos.
  - ⋆ Se basa en la técnica de programación dinámica.
  - ⋆ Aseguran un alineamiento óptimo.
  - ⋆ Coste computacional: para n secuencias de longitud media m el coste en tiempo es $O(2^n \cdot m^n)$.

MSA: aproximaciones algorítmicas (I)

- Métodos exactos.
  - ⋆ Se basa en la técnica de programación dinámica.
  - ⋆ Aseguran un alineamiento óptimo.
  - ⋆ Coste computacional: para n secuencias de longitud media m el coste en tiempo es $O(2^n \cdot m^n)$.
- Alineamiento progresivo.
  - ⋆ Calcula alineamientos de pares entre las secuencias consideradas.
  - ⋆ Elige el mejor alineamiento de entre ellos.
  - ⋆ Añade progresivamente más secuencias al alineamiento.
  - ⋆ El programa de alineamiento progresivo más usado es **ClustalW**: coste computacional $O(n^4 + m^2)$.

MSA: aproximaciones algorítmicas (II)

- Aproximaciones iterativas.
  - ⋆ Calculan una solución subóptima mediante un alineamiento progresivo y luego modifican el alineamiento mediante programación dinámica hasta que la solución converge.
  - ⋆ En un alineamiento progresivo, una vez que cometemos un error, no lo podemos corregir (aquí, sí).
  - ⋆ El programa **MUSCLE**.

MSA: aproximaciones algorítmicas (II)

- Aproximaciones iterativas.
  - ⋆ Calculan una solución subóptima mediante un alineamiento progresivo y luego modifican el alineamiento mediante programación dinámica hasta que la solución converge.
  - ⋆ En un alineamiento progresivo, una vez que cometemos un error, no lo podemos corregir (aquí, sí).
  - ⋆ El programa **MUSCLE**.
- Métodos basados en la consistencia.
  - ⋆ Incorpora la información de las distintas secuencias enla creación de cada alineamiento de pares.
  - ⋆ Los programas **ProbCons** y **T-Coffee**.

MSA: aproximaciones algorítmicas (II)

- Aproximaciones iterativas.
  - ⋆ Calculan una solución subóptima mediante un alineamiento progresivo y luego modifican el alineamiento mediante programación dinámica hasta que la solución converge.
  - ⋆ En un alineamiento progresivo, una vez que cometemos un error, no lo podemos corregir (aquí, sí).
  - ⋆ El programa **MUSCLE**.
- Métodos basados en la consistencia.
  - ⋆ Incorpora la información de las distintas secuencias enla creación de cada alineamiento de pares.
  - ⋆ Los programas **ProbCons** y **T-Coffee**.
- Métodos basados en la estructura.
  - ⋆ Utilizan la información sobre estructuras 3D de una o más de las secuencias.
  - ⋆ Los programas **PRALINE**, **PipeAlign** y **Expresso**.

CLUSTAL

Es un programa que se usa para el método de **alineamiento progresivo**.
Implementa el **algoritmo de FENG y DOOLITTLE progresivo**

1. Se alinea localmente todos los posibles pares de secuencias por el algoritmo NW.
2. A partir de las puntuaciones de similaridad se calcula una matriz de distancias.
   - Sea $S(s^{(i)}, s^{(j)})$ la similaridad entre las secuencias $s^{(i)}$ y $s^{(j)}$.
   - Sea $S_{max}(s^{(i)}, s^{(j)}) = \frac{S(s^{(i)},s^{(i)})+S(s^{(j)},s^{(j)})}{2}$ la máxima similaridad posible.
   - Sea $S_{rand}(s^{(i)}, s^{(j)})$ la media de las similaridades de esas secuencias permutadas muchas veces.
   - Se considera $S_{eff}(s^{(i)}, s^{(j)}) = \frac{S(s^{(i)},s^{(j)})-S_{rand}(s^{(i)},s^{(j)})}{S_{max}(s^{(i)},s^{(j)})+S_{rand}(s^{(i)},s^{(j)})}$.
   - La distancia $d(s^{(i)}, s^{(j)})$ se define así $d(s^{(i)}, s^{(j)}) = -\log S_{eff}(s^{(i)}, s^{(j)})$.
3. Se elabora un árbol guía (clustering).
   - La longitud de las ramas depende de las distancias.
   - Las ramas de las secuencias con distancias más cortas, se unen.
4. Se va construyendo el alineamiento múltiple mediante alineamiento de pares según distancias.
   - Se seleccionan las dos secuencias más cercanas según el árbol.
   - Se implementa un algoritmo de alineamiento para las mismas.
   - Se prosigue el proceso con las dos secuencias más cercanas hasta llegar a la raíz del árbol.