Secuenciación de DNA

Bioinformática, 10-4-19

Elvira Mayordomo

Hoy …

- **Introducción a la secuenciación de DNA**
- El proyecto del genoma humano
  - El proyecto de los 1000 genomas
- Mapa físico
  - Sitios de restricción
  - Mapeo por hibridación
- Método shotgun
- Hibridación

Secuenciación de DNA

- El "**chain-termination method**" nos permite secuenciar fragmentos directamente de como máximo 1000 pb
- El genoma humano es de 3,5 Gbp

  = $3{,}5\ 10^9$ bp

Secuenciación del DNA: 2 enfoques

Mapa físico

- Varias copias se cortan en trozos grandes con solape (50-300 KBP) Se pierde el orden de los trozos
- Un **mapa físico** es un conjunto de regiones cortas (**marcadores**), junto con sus posiciones exactas en la secuencia
- El mapa físico se utiliza para reconstruir el orden de estos trozos grandes

Secuenciación del DNA: 2 enfoques

Mapa físico

- ….
- El mapa físico se utiliza para reconstruir el orden de estos trozos grandes
- Los trozos son demasiado grandes para secuenciar
- Proceso de copia y corte, se secuencian los trozos pequeños
- Ejemplo: El proyecto del genoma humano

Secuenciación del DNA: 2 enfoques

Método “shotgun”

- Omite el paso de trozos grandes y mapa físico
- Directamente varias copias se cortan en **trozos pequeños** con solape. Se pierde el orden de los trozos

Secuenciación del DNA: 2 enfoques

Método “shotgun”

- Directamente varias copias se cortan en **trozos pequeños** con solape
- Reconstrucción directa del orden, si el trozo es largo sólo se secuencia prefijo y sufijo
- Mucha computación intensiva
- Utilizado por Celera genomics en el proyecto del genoma humano

Hoy …

- Introducción a la secuenciación de DNA
- **El proyecto del genoma humano**
  - El proyecto de los 1000 genomas
- Mapa físico
  - Sitios de restricción
  - Mapeo por hibridación
- Método shotgun
- Hibridación

Proyecto del Genoma Humano

- Iniciado en los 80, declarado finalizado en 2003
- Carrera entre iniciativa pública y privada, universidades y laboratorios públicos vs Celera genomics
- Celera usó libremente los datos públicos pero no facilitó sus datos hasta mucho después
- La carrera la ganó la iniciativa pública …

Proyecto del Genoma Humano

- Todavía tiene muchos problemas:
  - Ambigüedad
  - Las diferencias entre los diferentes proyectos
  - Falta de información crítica: posición de los genes

Otros proyectos/retos

- El proyecto de los 1000 genomas
- El genoma de $1000

Hoy …

- Introducción a la secuenciación de DNA
- El proyecto del genoma humano
  - El proyecto de los 1000 genomas
- **Mapa físico**
  - Sitios de restricción
  - Mapeo por hibridación
- Método shotgun
- Hibridación

Mapa físico

- Sea D una secuencia de DNA. UN **mapa físico** es
  - un conjunto M de marcadores genéticos (secuencias cortas de DNA)
  - Una función p que asigna posiciones a cada marcador

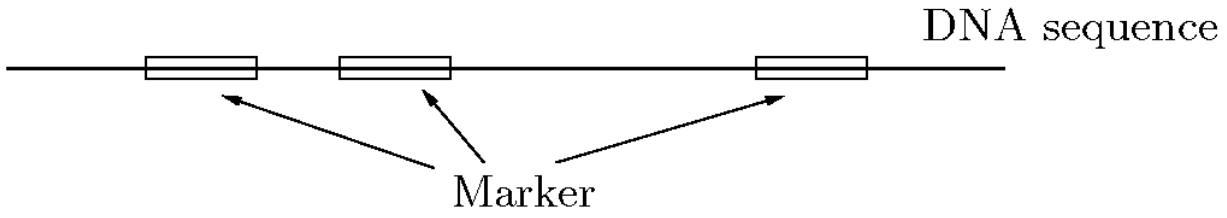


Fig. 7.1. Schematic view of a physical map

Mapa físico

- Utilizado también en la **búsqueda de genes**
- Dos enfoques para el mapa físico:
  - Los sitios de restricción
  - Hibridación

Hoy …

- Introducción a la secuenciación de DNA
- El proyecto del genoma humano
  - El proyecto de los 1000 genomas
- Mapa físico
  - **Sitios de restricción**
  - Mapeo por hibridación
- Método shotgun
- Hibridación

Mapeo con sitios de restricción

- Aplicamos **enzimas de restricción:** reconocen subsecuencias cortas y cortan en esas posiciones (digestión)

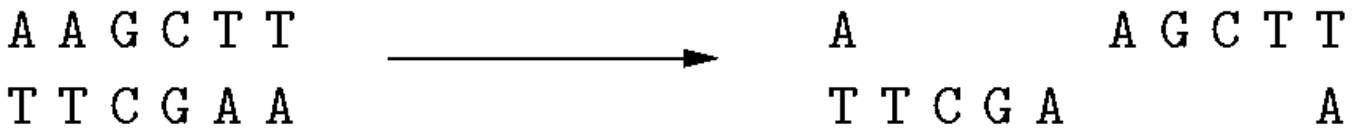


Fig. 7.2. Restriction site of restriction enzyme *HindIII*

- A partir de los fragmentos (desordenados) resultantes queremos reconstruir la secuencia

Enfoque digestión doble

- Utilizamos dos enzimas de restricción primero por separado y luego juntos.
- A partir de los tres conjuntos de fragmentos resultantes se reconstruye la secuencia (en realidad **a partir de las longitudes** de los fragmentos resultantes)

Enfoque digestión doble

Method 7.1 The double digest approach

Input: The considered DNA molecule $\mathcal{D}$ and two distinct restriction enzymes $\mathcal{A}$ and $\mathcal{B}$.

1. Generate three copies of $\mathcal{D}$.
2. Apply enzyme $\mathcal{A}$ to the first, enzyme $\mathcal{B}$ to the second, and both enzyme $\mathcal{A}$ and enzyme $\mathcal{B}$ to the third copy. We obtain a set of unordered fragment from each copy.
3. Determine the lengths of the fragments of $\mathcal{D}$ and obtain three multisets:
   - $\Delta(\mathcal{A})$: Contains the lengths of the fragments resulting from the digest of the (first) copy of $\mathcal{D}$ by enzyme $\mathcal{A}$.
   - $\Delta(\mathcal{B})$: Contains the lengths of the fragments resulting from the digest of the (second) copy of $\mathcal{D}$ by enzyme $\mathcal{B}$.
   - $\Delta(\mathcal{AB})$: Contains the lengths of the fragments resulting from the digest of the (third) copy of $\mathcal{D}$ by enzyme $\mathcal{A}$ and enzyme $\mathcal{B}$.

Output: The multisets $\Delta(\mathcal{A})$, $\Delta(\mathcal{B})$, $\Delta(\mathcal{AB})$.

Enfoque digestión doble

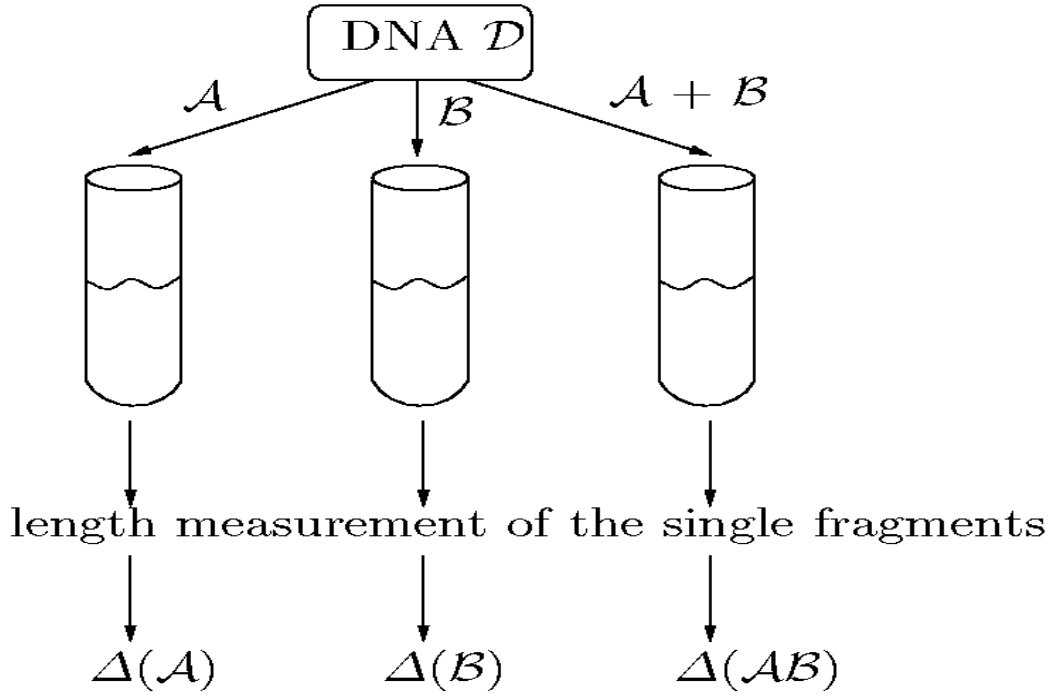


Fig. 7.3. Schematic view of the double digest approach

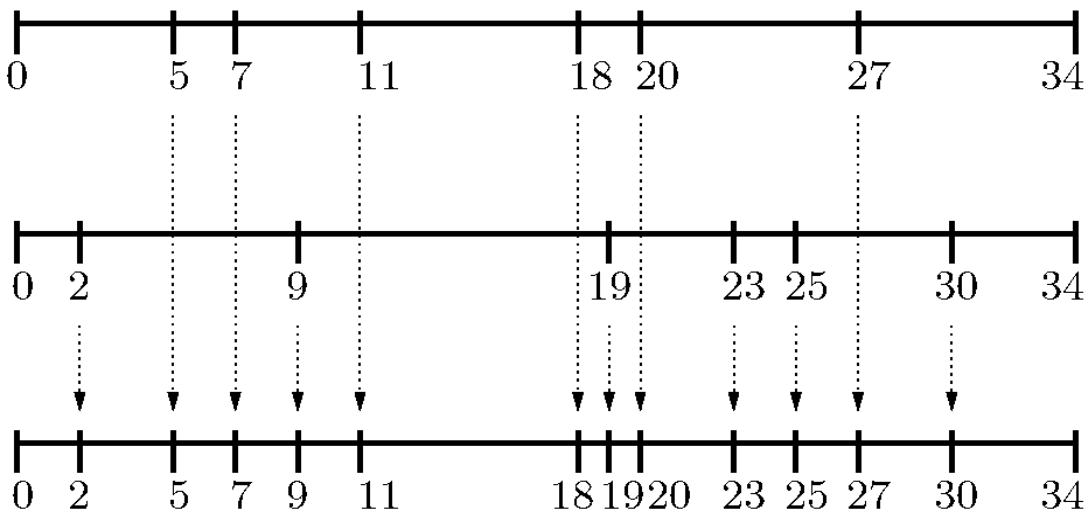
Enfoque digestión doble

- Observaciones:
  - La digestión es completa: el DNA se corta en todos los sitios posibles
  - Multiconjuntos: las longitudes de los fragmentos se pueden repetir
- Queremos reconstruir el orden de las secuencias (obteniendo de este modo un mapa físico con los marcadores de restricción)

Enfoque digestión doble

$A = \{2, 2, 4, 5, 7, 7, 7\}$
$\pi = (5, 2, 4, 7, 2, 7, 7)$

$B = \{2, 2, 4, 4, 5, 7, 10\}$
$\phi = (2, 7, 10, 4, 2, 5, 4)$



$C = \{1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 7\}$
$Pos(\pi) \cup Pos(\phi) = \{0, 2, 5, 7, 9, 11, 18, 19, 20, 23, 25, 27, 30, 34\}$

Fig. 7.4. Schematic view of the double digest problem. Find orderings π and ϕ of A and B, such that the multiset C is implied

Problema doble digestión (DDP)

- Dados 3 multiconjuntos de números A, B, C
- Encuentra una permutación de A y B que sea compatible con C

Problema doble digestión

- Es un problema NP-completo (su versión decisional)
- Existen múltiples soluciones (más difícil todavía)
- También podemos asignar coste a la distancia a una solución real y optimizar (sigue siendo difícil)

Problema doble digestión

- Resumen:
  - experimentos biológicos fáciles
  - problema combinatorio difícil

Enfoque digestión parcial

- Utilizamos un solo enzima de restricción
- Consideramos diferentes duraciones de la exposición al enzima (los fragmentos vuelven a contener sitios de restricción)

Enfoque digestión parcial

Method 7.2 The partial digest approach

Input: The considered DNA sequence $\mathcal{D}$ and an restriction enzyme $\mathcal{A}$.

1. Generate several copies of $\mathcal{D}$.
2. Apply the enzyme $\mathcal{A}$ in separated experiments with different duration on the single copies and obtain a set of fragments from each experimental setup.
3. Determine the length of the resulting fragments of $\mathcal{D}$ and combine all these lengths into a single multiset $\Delta_p(\mathcal{A})$.

Output: The multiset $\Delta_p(\mathcal{A})$.

Advertencia: los fragmentos de diferentes digestiones se ponen juntos

Enfoque digestión parcial

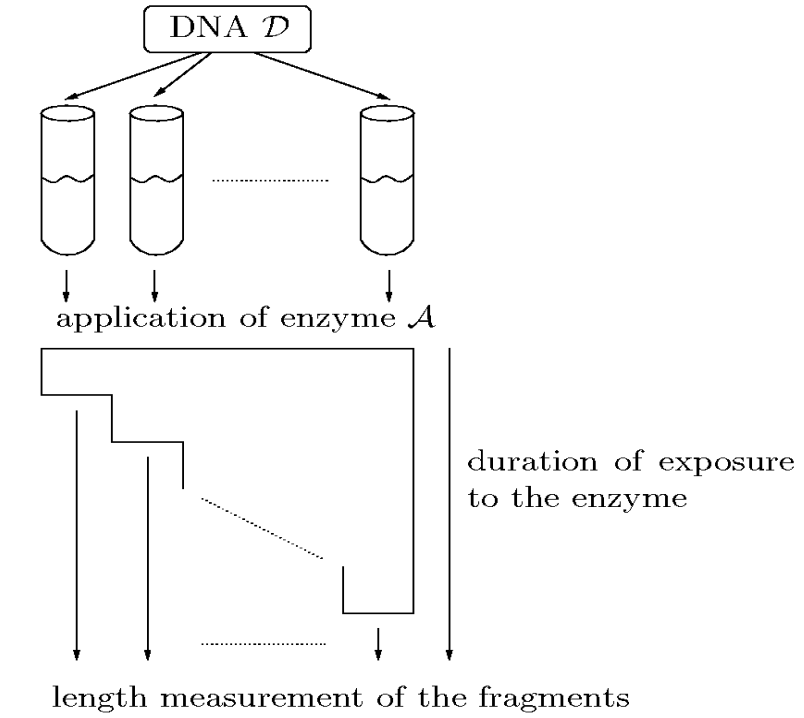


Fig. 7.6. The partial digest approach

Enfoque digestión parcial

- Simplificación: se supone que el resultado es *ideal*: Cada sitio se utiliza sólo una vez en todos los experimentos

Digestión parcial

Problema Digestión parcial (PDP)

- Dado un multiconjunto A (las longitudes de fragmentos resultantes de digestión parcial)
- Encontrar los sitios de restricción (lugares donde se ha cortado el DNA) $x_1 < x_2 \ldots < X_k$

$$\{x_j - x_i \mid 1 \leq i < j \leq k\} = A$$

Digestión parcial

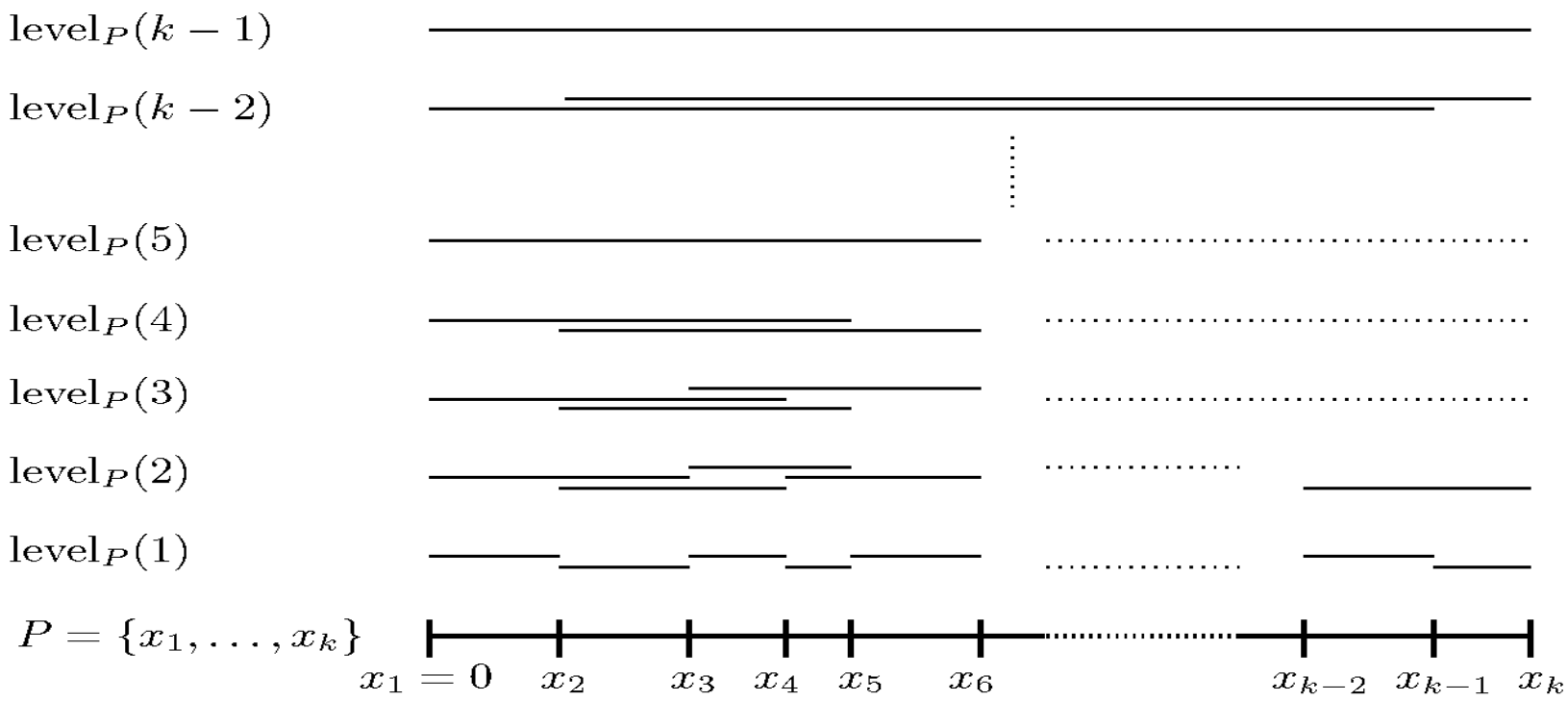


Fig. 7.7. Schematic view of the PDP

Digestión parcial

- Se resuelve con backtracking
- El peor caso es tiempo exponencial, pero si hay poco backtracking es muy eficiente
- No hay resultados en NP-completitud
- Los experimentos biológicos son simples, el problema combinatorio es razonable

Hoy …

- Introducción a la secuenciación de DNA
- El proyecto del genoma humano
  - El proyecto de los 1000 genomas
- Mapa físico
  - Sitios de restricción
  - **Mapeo por hibridación**
- Método shotgun
- Hibridación

Hibridación

- La hibridación es la unión de las cadenas complementarias
- En este método se nos da la “huella” del resultado de la hibridación con "sondas“ (probes)
- Los fragmentos con solape tienen huellas similares

Method 7.3 Mapping by Hybridization

Given: The DNA sequence $\mathcal{D}$ under investigation.

1. Cut, according to an arbitrary method (restriction enzymes, sonic waves, vibration), several copies of $\mathcal{D}$ into a set of fragments. This set will include overlapping fragments with high probability.
2. For each of the fragments generate copies that will be used within a hybridization experiment. The fragments are called clones.
3. Let $C = \{c_1, \ldots, c_n\}$ be the set of clones (also called clone library), and choose a set of probes $P = \{p_1, \ldots, p_m\}$.
4. Perform all hybridization experiments (c_i, p_j), $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$, (for example, by means of a DNA chip).

Output: An $(n \times m)$-*hybridization matrix* $\mathcal{H}$, where

$$\mathcal{H}(i,j) = \begin{cases} 1 & \text{if } c_i \text{ and } p_j \text{ hybridize, and} \\ 0 & \text{otherwise.} \end{cases}$$

Hibridación

- A partir de la matriz de hibridación queremos reconstruir el orden original de los clones en el DNA D
- Nótese que no podemos distinguir si una sonda se produce una o varias veces en un clon

	A	B	C	D
c_1	1	1	0	0
c_2	1	1	1	1
c_3	1	0	0	1
c_4	0	0	0	1

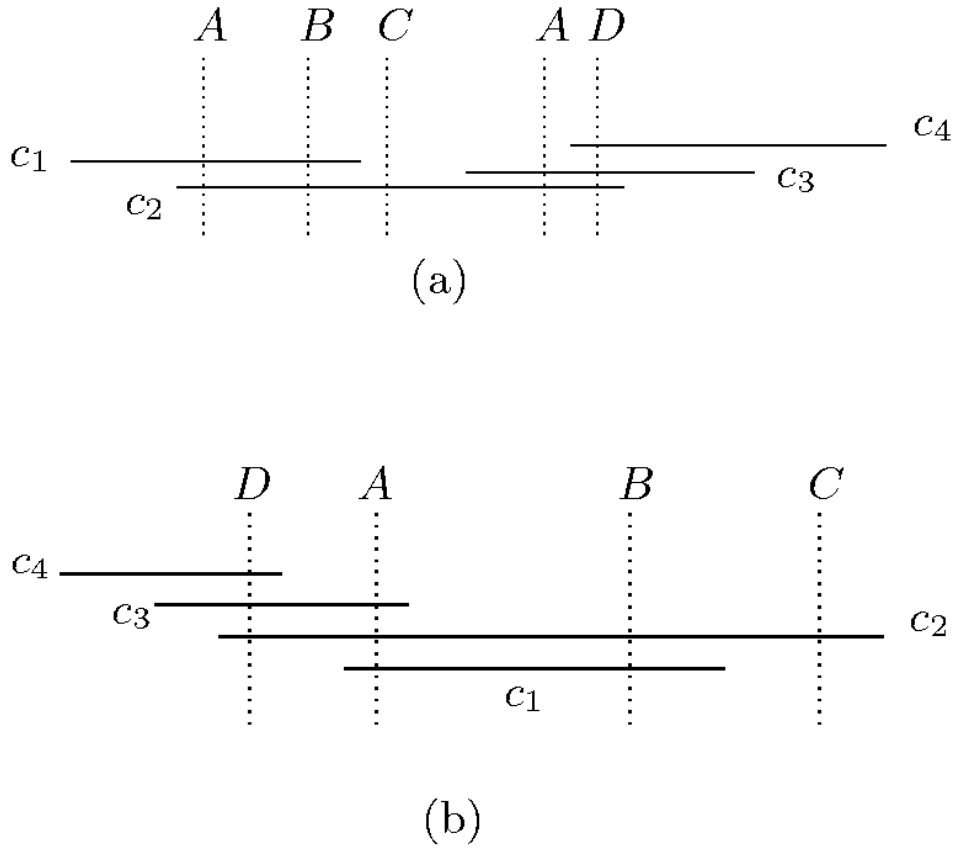


Fig. 7.10. Hybridization matrix and two derived orderings of clones or probes. Both orderings will lead to the given hybridization matrix. Note that in ordering (a) the probe A occurs twice and in ordering (b) the clone c_1 is completely covered by clone c_2

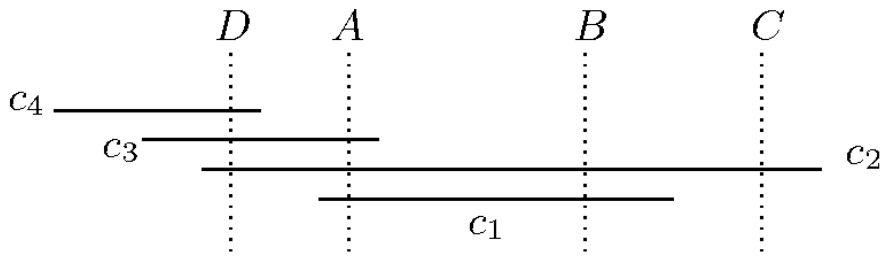
Mapeo por hibridación

- La matriz puede contener errores
- El problema es más fácil si cada sonda puede ocurrir sólo una vez en cada clon (factible experimentalmente)
- Consideramos:
  - Mapeo con sondas únicas
  - Mapeo con sondas únicas y errores
  - Mapeo con sondas no únicas

Mapeo con sondas únicas

- La matriz está libre de errores
- Cada sonda puede ocurrir sólo una vez en un clon

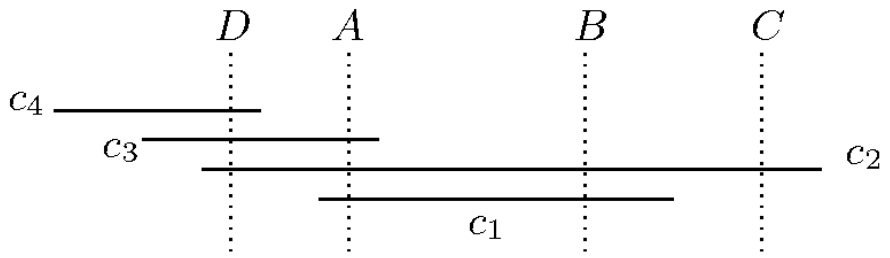
	A	B	C	D
c_1	1	1	0	0
c_2	1	1	1	1
c_3	1	0	0	1
c_4	0	0	0	1



Mapeo con sondas únicas

- Buscamos a una reordenación de las columnas de la matriz de modo que todos los 1 son adyacentes

	A	B	C	D
c_1	1	1	0	0
c_2	1	1	1	1
c_3	1	0	0	1
c_4	0	0	0	1



Mapeo con sondas únicas

Problema de los unos consecutivos

- Dado una matriz (n x m) con entradas en {0,1}
- Encontrar una permutación de las columnas de manera que en cada fila NO hay 0 entre dos 1s

Mapeo con sondas únicas

- El Problema de los unos consecutivos puede ser resuelto en el tiempo O(n+m+k), donde k es el número de 1s en la matriz

- Algoritmo muy sofisticado ...

Mapeo con sondas únicas y errores

- Un error de falso negativo es cuando hay un 0 en la matriz que debe ser 1
- Un error de falso positivo significa que hay un 1 que debe ser un 0
- También podemos tener dos filas mezcladas en una (porque dos clones se han unido en uno)

Mapeo con sondas únicas y errores

- Nuestro enfoque sigue el *principio de parsimonia*: que haya el menor número de errores posible que explique nuestros datos
- Llamamos hueco a un bloque de 0s consecutivos bordeado por 1s dentro de una fila de la matriz

Mapeo con sondas únicas y errores

Problema de minimización de huecos

- Dada una matriz (n x m) con entradas en {0,1}
- Encuentra una permutación de las columnas que minimiza el número total de huecos

Mapeo con sondas únicas y errores

Problema de minimización de huecos

- NP-completo
- Muy cerca de TSP euclídeo
- Varios algoritmos de aproximación y heurísticas conocidas

Mapeo con sondas no únicas

- Cada sonda puede ocurrir varias veces en un clon, pero con poca frecuencia
- Queremos minimizar el número de repeticiones

	A	B	C	D	E	F	G
c_1	1	1	0	1	0	0	0
c_2	0	1	0	1	0	0	1
c_3	1	0	0	1	1	0	1
c_4	1	1	0	1	1	1	0
c_5	0	1	1	1	1	1	0
c_6	0	0	1	0	1	0	1
c_7	0	1	0	0	0	1	1

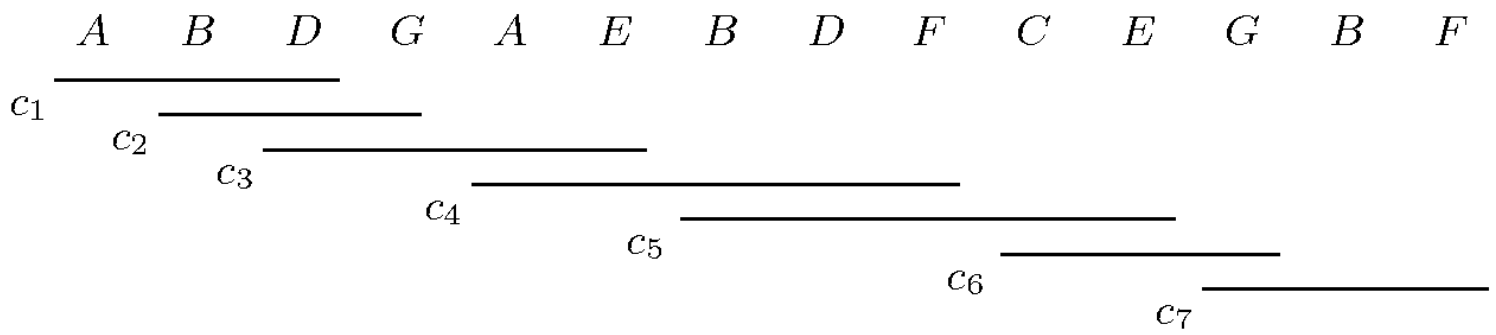


Fig. 7.30. Hybridization matrix and a feasible solution of the shortest covering string problem

Mapeo con sondas no únicas

Problema del string más corto que cubre

- Dada una matriz (n x m) con entradas en {0,1}
- Encontrar repeticiones de columnas y una permutación de las columnas de manera que en cada fila NO hay 0 entre dos 1s, reducir al mínimo el número total de columnas

Mapeo con sondas no únicas

- El Problema del string más corto que cubre es NP-completo
- No está claro si es un buen modelo de sondas no únicas, debido a que algunas sondas pueden ocurrir de forma natural con más frecuencia que otras.

Resumen mapa físico

- Sitios de restricción
  - Digestión doble: experimentos simples, problema combinatorio difícil
  - Digestión parcial: experimentos simples, problema combinatorio razonable
- Mapeo por hibridación:
  - Sondas únicas: algoritmo eficiente
  - Sondas únicas y errores: aproximación y heurísticas
  - Sondas no únicas: problema difícil, modelización dudosa

Hoy …

- Introducción a la secuenciación de DNA
- El proyecto del genoma humano
  - El proyecto de los 1000 genomas
- Mapa físico
  - Sitios de restricción
  - Mapeo por hibridación
- **Método shotgun**
- Hibridación

Método Shotgun

Method 8.1 Shotgun Sequencing

Input: A DNA molecule $\mathcal{D}$.

1. Generate a set $C = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ of copies of $\mathcal{D}$.
2. Cut each copy from C randomly into smaller fragments. This results in a set of overlapping fragments $F = \{f_1, \ldots, f_n\}$.
3. Determine the sequence of all the fragments (or their starting sequences) by means of direct sequencing. This gives us a set of strings $S = \{s_1, \ldots, s_n\}$ over the alphabet $\Sigma_{\text{DNA}} = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$.

Output: The set of strings $S = \{s_1, \ldots, s_n\}$ corresponding to the sequences of the fragments in F.

Seguido por la solución a partir de los trozos …

Shotgun: SuperSeqComun

- Dados s_1, …, s_k strings, m un entero
- ¿Existe un string t con |t|=m de forma que s_i es subsecuencia de t para todo i?

- Este problema es NP-completo (Middendorf 1994)

Shotgun: SuperSeqComun

- Dados s_1, …, s_k strings, m un entero
- ¿Existe un string t con |t|=m de forma que s_i es subsecuencia de t para todo i?

- Podemos asumir que ningún s_i es substring de otro s_j

Una solución voraz

Algorithm 8.1 Greedy Superstring

Input: A set of strings $S = \{s_1, \ldots, s_n\}$.

while $|S| > 1$ **do**

1. Find $s_i, s_j \in S$, $s_i \neq s_j$, that have a maximum overlap among all strings in S.
2. Let $s' = \langle s_i, s_j \rangle$ be the merge of strings s_i and s_j.
3. Delete s_i, s_j from S and insert s' into S.

Output: The only remaining string $w_{\text{greedy}} \in S$.

Hoy …

- Introducción a la secuenciación de DNA
- El proyecto del genoma humano
  - El proyecto de los 1000 genomas
- Mapa físico
  - Sitios de restricción
  - Mapeo por hibridación
- Método shotgun
- **Hibridación**

Hibridación

Method 8.2 Sequencing by Hybridization — SbH

Input: A DNA molecule $\mathcal{D}$ and a natural number l.

1. Generate a DNA chip that contains all different probes of length l, where each probe is assigned to a specific position on the chip.
2. Generate a number of copies of $\mathcal{D}$.
3. Expose the generated copies to the DNA chip and detect the positions where hybridizations between probes and DNA occurred.

Output: The set of strings $S = \{s_1, \ldots, s_n\} \subseteq \Sigma^l_{\text{DNA}}$ of length l that are contained as substrings in $\mathcal{D}$.

tarea restante: reconstruir D de S
(Se puede hacer en tiempo lineal)

Secuenciación de DNA

- Mapa físico
  - Los sitios de restricción
    - doble digestión
    - digestión parcial
  - Mapeo por hibridación
    - sondas únicas
    - sondas únicas y errores
    - sondas no únicas
- Método shotgun
- Hibridación

Problemas de los métodos clásicos de secuenciación …

- Coste y tiempo
- Tasa de errores
- Para hibridización hay que tener información a priori …