Descubrimiento de Motivos en Secuencias

Luis Garreta

Electiva de Bioinformática
MAESTRÍA EN INFORMÁTICA BIOMÉDICA
Universidad del Bosque
Bogotá-Colombia

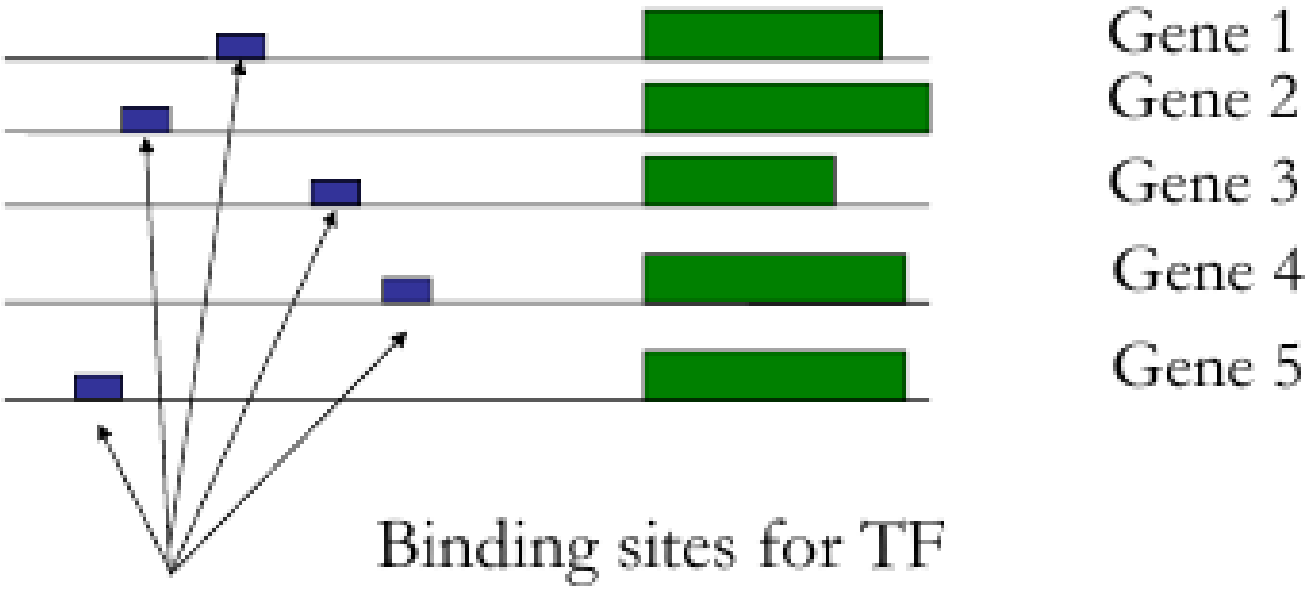
25 de septiembre de 2021

Motivo de secuencia: definiciones

- En bioinformática, un *motivo de secuencia* es un *patron de secuencia* de nucleótidos o de aminoácidos que *se repite* en varias partes de la misma secuencia o entre varias secuencias y ha sido probado o asumido que tiene un significado biológico.
- Una vez que conocemos el patrón de secuencia del motivo, podemos usar distintos métodos de búsqueda para encontrarlo en las secuencias (es decir, algoritmo de Boyer-Moore, Rabin-Karp, árboles de sufijos, etc.)
- El problema es descubrir los motivos, es decir, cuál es el *orden de las letras* que componen el motivo particular.

Ejemplos de motivos en el ADN

- *TATA BOX*: La secuencia del promotor TATA es un ejemplo de un motivo de secuencia de ADN altamente conservado que se encuentra en eucariotas.
- *Sitios de unión*: Otro ejemplo de motivos son los sitios de unión para factores de transcripción (TF) cerca de regiones promotoras de genes, etc.



Motivo de secuencia: notaciones

Un ejemplo de un motivo en una proteína:

- N, seguido de cualquier cosa menos P, seguido de S o T, seguido de cualquier cosa menos P
- Una convención es escribir:

  N {P} [ST] {P}

  Donde
  - {X} significa cualquier aminoácido excepto X;
  - [XYZ] significa X o Y o Z.

Descubrimiento de motivos de secuencia a partir de la conservación

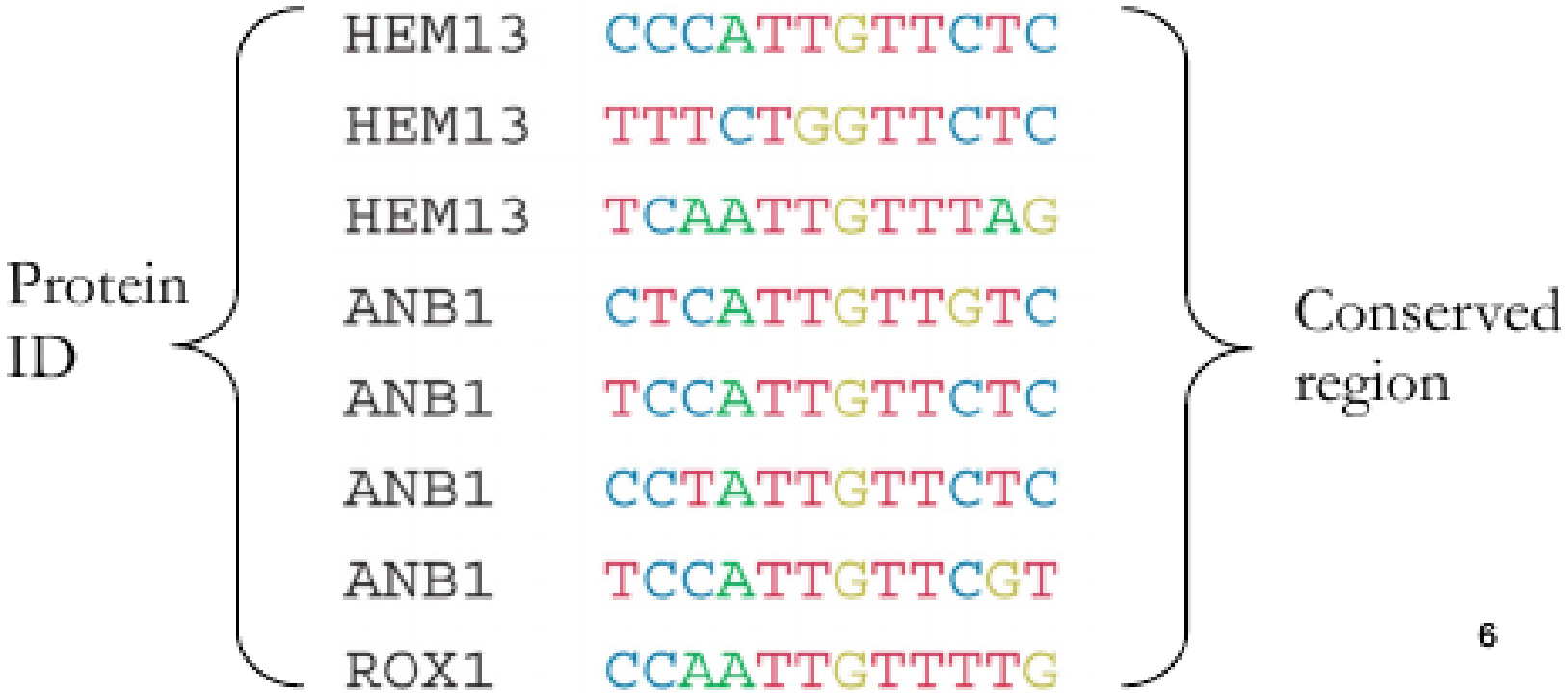
- Los motivos de secuencia son secuencias conservadas de patrones similares o idénticos que pueden ocurrir dentro de ácidos nucleicos (ADN, ARN) o proteínas, ya sea:
  - dentro de diferentes moléculas producidas por el mismo organismo, o
  - dentro de moléculas de múltiples especies de organismos

- En el caso de la conservación de especies cruzadas, el motivo conservado indica que un patrón de secuencia particular puede haberse conservado durante la evolución para realizar una función vital en esas especies.
  - La conservación de motivos es la base del descubrimiento de motivos mediante el estudio de genes (o proteínas) similares en diferentes especies;

Descubrimiento de motivos basado en la alineación

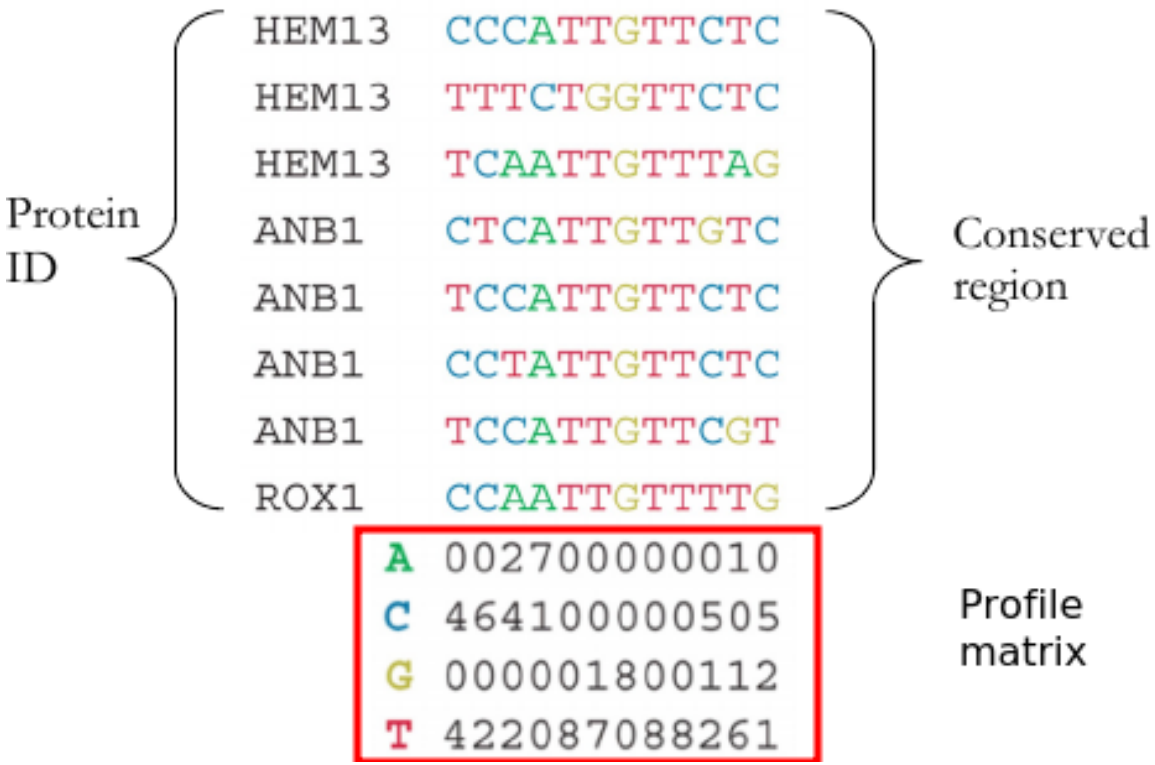
También se conoce como *análisis de perfil*. Los pasos son:

- Primero, se construye un *alineamiento local* de múltiples secuencias,
- Segundo, se aíslan las *regiones altamente conservadas*, en base a su alta puntuación de alineamiento



Descubrimiento de motivos basado en la alineación

- Después de aislar las regiones altamente conservadas, se utilizan para construir *matrices de perfil* para cada región conservada.
- La matriz de perfil para un motivo dado contiene *conteos de frecuencia* para cada letra en cada posición de la región conservada.



Logotipo de secuencia y secuencia consenso

- Podemos extraer la llamada *secuencia de consenso*, es decir, la cadena de letras más frecuentes:

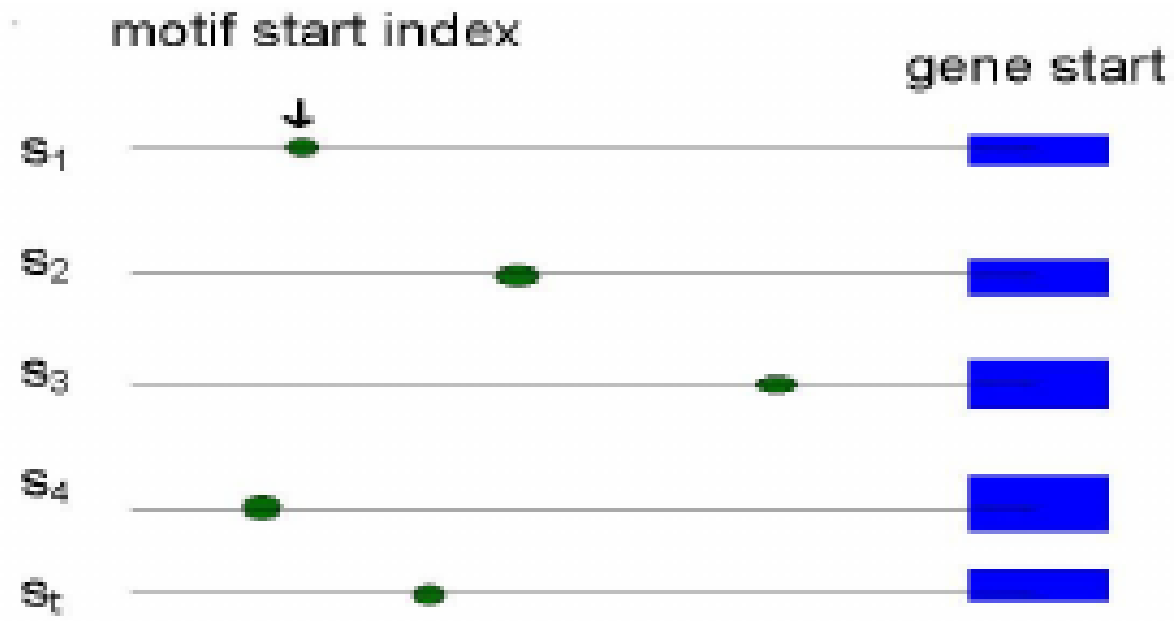
YCHATTGTTCTC

- Una representación gráfica de la secuencia de consenso se denomina *logotipo de secuencia*:

- La altura de las diferentes letras en la misma posición es proporcional a su frecuencia en los motivos: *cuanto mejor sea la conservación de la base en esa posición, más altas serán las letras*.

Descubriendo motivos sin alineación

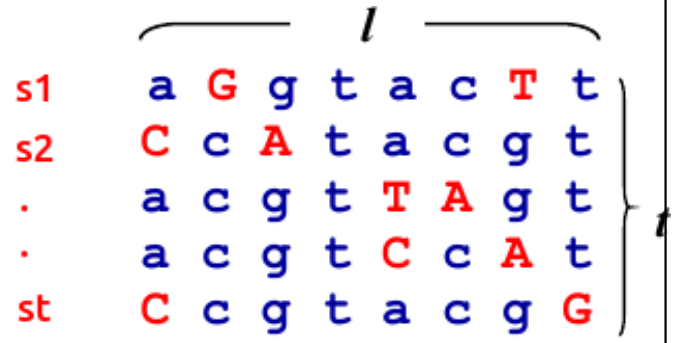
- Primero, supongamos que sabemos dónde comienza el motivo en el conjunto de secuencias de referencia.
- Las posiciones de inicio del motivo en secuencias se pueden representar como el conjunto $s = (s_1, s_2, s_3, ..., s_t)$donde s_i es el índice de posición



Puntuación de motivos

- Queremos calcular la puntación del motivo $s = «ACGTACGT»$

- Dado $s = (s_1, s_2, s_3, ..., s_t)$ alineamos t motivos de longitud l de todas las secuencias:



- Construimos una matrix de perfiles:

A	3	0	1	0	3	1	1	0
C	2	4	0	0	1	4	0	0
G	0	1	4	0	0	0	3	1
T	0	0	0	5	1	0	1	4

- Calculamos el puntaje del motivo $s = «ACGTACGT»$

$$Score(s) = \sum_{i=1}^{l} \max_{k \in \{A,C,G,T\}} [count(k,i)]$$

Score=3+4+4+5+3+4+3+4

Score = 30

Problema de búsqueda de motivos

- El problema es encontrar las posiciones iniciales $s = (s_1, s_2, s_3, ..., s_t)$ para maximizar la puntuación (s) de la matriz de perfil resultante.
- Se utilizan varios tipos de matrices de perfiles:
  - *Una matriz de frecuencia de posición (PFM)* registra la frecuencia dependiente de la posición f de cada letra, es decir, cuántas veces aparece una letra en una posición determinada en N secuencias.
  - *Una matriz de probabilidad de posición (PPM)*. Cuando se normaliza a 1, esta frecuencia se convierte en una probabilidad, es decir, P = f / N.
  - Una matriz de peso de posición (PWM):

$$\sum_{\beta \in \{A,C,G,T\}} f_{\beta k} \log \frac{f_{\beta k}}{q_\beta}$$

Corrección en las Frecuencias Genómicas del ADN de Fondo

Se asume una frecuencia igual de ocurrencia de las bases:

- f(A) = f(C) = f(G) = f(T) = 1/4 = 0.25
- Pero el contenido G+C de los genomas cambia por especie:
  - E.coli (51 % GC), human (41 %) : razonable
  - S. cerevisiae (38 %), Caenorhabditis elegans (36 %) : de cuidado
  - Plasmodium falciparum (19 %), Streptomyces coelicolor (72 %): extremo

Contenido de Información

$$I_i = 2 + \sum_b f_{b,i} log_2 f_{b,i}$$

$$I_i = -\sum_b f_{b,i} log_2 \frac{f_{b,i}}{q_b}$$

Donde
I_i: Contenido de información de la posición i
$f_{b,i}$: Frecuencia de la base b en la posición i.
qpb_b : Frecuencia de fondo de la base b en el genoma

Empecemos de nuevo: Problema de búsqueda de motivos

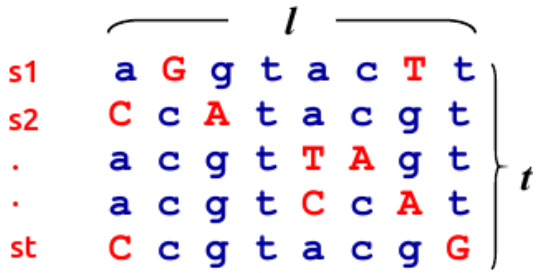
- Dada una lista de t secuencias cada una de longitud n, encuentre el "mejor" patrón de longitud l que aparece en cada una de las t secuencias.

- Sea $s = (s_1, s_2, s_3, ..., s_t)$ el conjunto de posiciones iniciales para *l-meros* en nuestras secuencias t.

- Las cadenas correspondientes a estas posiciones iniciales formarán:
  - Matriz de perfil P de dimensiónes 4xl para DNA, y
  - Matriz de perfil P de dimensiones 20xl para proteínas
  - La matriz de perfil P se definirá en términos de *probabilidad de letras*, y no como conteo de letras.

Matriz de Perfil P

- Dado $s = (s_1, s_2, s_3, ..., s_t)$ alineamos t *l-meros* de todas las secuencias:



- Construcción del perfil *P:*
  - Cada posición de la matriz tiene un peso *w* o probabilidad.
  - Se normaliza a 1 estas frecuencias y se vuelve una probabilidad, es decir, $P = f/N$

PWM

	i1	i2	i3	i4	i5	i6	i7	i8
A	0.6	0.0	0.2	0.0	0.6	0.2	0.2	0.0
C	0.4	0.8	0.0	0.0	0.2	0.8	0.0	0.0
G	0.0	0.2	0.8	0.0	0.0	0.0	0.6	0.2
T	0.0	0.0	0.0	1.0	0.2	0.0	0.2	0.8

La probabilidad de los *l-meros*

- $Pr(a|P)$ se define como la probabilidad de que un *l-mero* a fuera creado por el perfil P.
- Si a es muy similar a la cadena de consenso (es decir, motivo), entonces $Pr(a|P)$ será alto.
- Si $P_{ai,k}$ es la probabilidad de la letra a_i en la posición k, entonces la probabilidad de un *l-mero* a es igual al producto de las probabilidades individuales $P_{ai,k}$, es decir:

$$\Pr(\mathbf{a} \mid \mathbf{P}) = \prod_{k=1}^{l} P_{a_i,k}$$

Puntuación de *l-mero* con un perfil

Dado un perfil P =

A	1/2	7/8	3/8	0	1/8	0
C	1/8	0	1/2	5/8	3/8	0
T	1/8	1/8	0	0	1/4	7/8
G	1/4	0	1/8	3/8	1/4	1/8

$Pr(\mathbf{aaacct}|\mathbf{P})$ = ???

Puntuación de *l-mero* con un perfil (continuación)

Dado un perfil P =

A	**1/2**	**7/8**	**3/8**	0	1/8	0
C	1/8	0	1/2	**5/8**	**3/8**	0
T	1/8	1/8	0	0	1/4	**7/8**
G	1/4	0	1/8	3/8	1/4	1/8

$Pr(\mathbf{aaacct}|\mathbf{P})$ = 1/2 x 7/8 x 3/8 x 5/8 x 3/8 x 7/8 = .033646

Puntuación de *l-mero* con un perfil (continuación)

Dado un perfil P =

A	**1/2**	7/8	**3/8**	0	**1/8**	0
C	1/8	0	1/2	**5/8**	3/8	0
T	1/8	**1/8**	0	0	1/4	7/8
G	1/4	0	1/8	3/8	1/4	**1/8**

Prob(**aaacct**|**P**) = 1/2 x 7/8 x 3/8 x 5/8 x 3/8 x 7/8 = .033646

Prob(**atacag**|**P**) = 1/2 x 1/8 x 3/8 x 5/8 x 1/8 x 1/8 = .001602

Motivo - el *l-mero P*-más probable

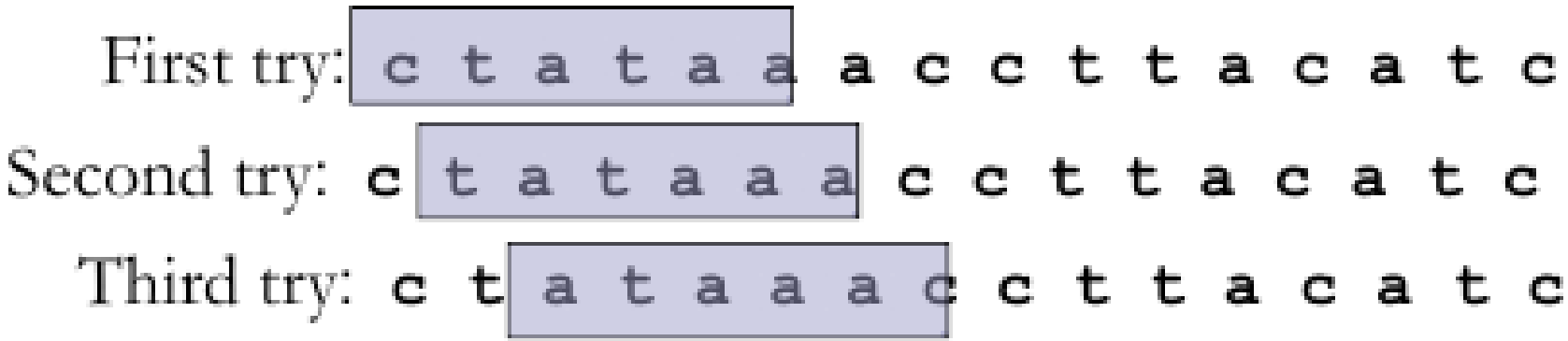
- Defina el *l-mero P-más probable* de una secuencia como un *l-mero* en esa secuencia que tiene la mayor probabilidad de ser creado a partir del perfil *P*.
- Tarea: dada una secuencia ctataaaccttacatc y el perfil conocido *P*, encuentre el *P-6-mero* más probable:

$\mathbf{P} =$

A	1/2	7/8	3/8	0	1/8	0
C	1/8	0	1/2	5/8	3/8	0
T	1/8	1/8	0	0	1/4	7/8
G	1/4	0	1/8	3/8	1/4	1/8

l-mero P-más probable

A	1/2	7/8	3/8	0	1/8	0
C	1/8	0	1/2	5/8	3/8	0
T	1/8	1/8	0	0	1/4	7/8
G	1/4	0	1/8	3/8	1/4	1/8



Deslice la ventana para evaluar cada 6-mer posible

- Enfoque de fuerza bruta

l-mero P-más probable

Calcule $Pr(a|P)$ para cada 6-mero posible:

Window, Highlighted Red	Calculations	$Pr(a\|P)$
ctataaaccttacat	1/8 x 1/8 x 3/8 x 0 x 1/8 x 0	0
ctataaaccttacat	1/2 x 7/8 x 0 x 0 x 1/8 x 0	0
ctataaaccttacat	1/2 x 1/8 x 3/8 x 0 x 1/8 x 0	0
ctataaaccttacat	1/8 x 7/8 x 3/8 x 0 x 3/8 x 0	0
ctataaaccttacat	1/2 x 7/8 x 3/8 x 5/8 x 3/8 x 7/8	.0336
ctataaaccttacat	1/2 x 7/8 x 1/2 x 5/8 x 1/4 x 7/8	.0299
ctataaaccttacat	1/2 x 0 x 1/2 x 0 1/4 x 0	0
ctataaaccttacat	1/8 x 0 x 0 x 0 x 0 x 1/8 x 0	0
ctataaaccttacat	1/8 x 1/8 x 0 x 0 x 3/8 x 0	0
ctataaaccttacat	1/8 x 1/8 x 3/8 x 5/8 x 1/8 x 7/8	.0004

P-6-mero más probable en la secuencia es *aaacct*:

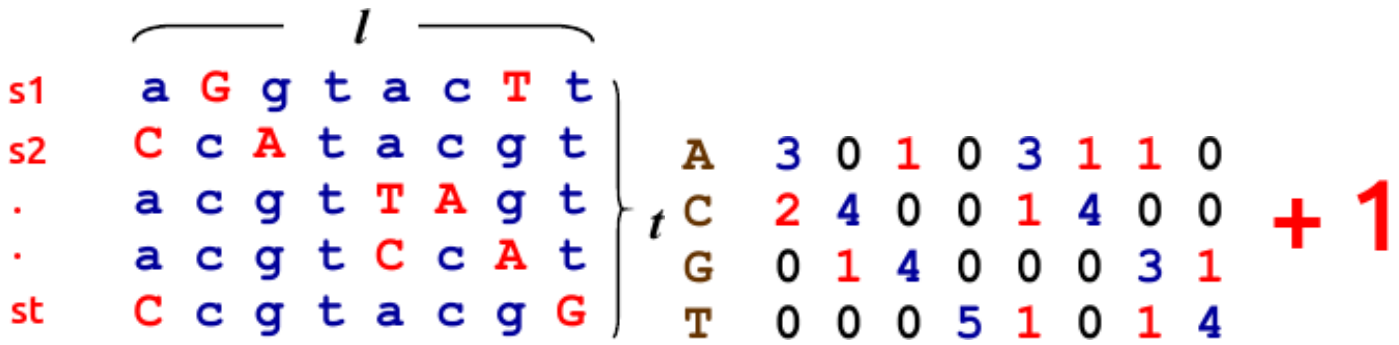
Algunas correcciones: Pseudoconteos en la Matriz de Frecuencias

Ceros en la matriz de frecuencias pueden generar problemas en los calculos:

- Se presentan cuando hay pocas secuencias $s = (s_1, s_2, s_3, ..., s_t)$
- Resulta en probabilidades iguales a 0
- Resulta en logaritmos inválidos (log de 0 es infinito)

Solución:

- Sumarle 1 valor que no afecte el conteo



Algunas correcciones: logaritmos en vez de Probabilidades

$$P(S/M) = Score(s) \prod_{i=1}^{l} f_{b,i}$$

$$P(S/M) = Score(s) \sum_{i=1}^{l} log_2 f_{b,i}$$

Donde:

$b = A, C, G, T$

$i = 1...l$, l *es la longitud de la sequencia*

Ejemplo:

P(«ACGTACGT»/M)

PWM

log2

	i1	i2	i3	i4	i5	i6	i7	i8
A	0.6	0.0	0.2	0.0	0.6	0.2	0.2	0.0
C	0.4	0.8	0.0	0.0	0.2	0.8	0.0	0.0
G	0.0	0.2	0.8	0.0	0.0	0.0	0.6	0.2
T	0.0	0.0	0.0	1.0	0.2	0.0	0.2	0.8

$$P = log_2 0{,}6 + log_2 0{,}8 + log_2 0{,}8 + log_2 1{,}0 + log_2 0{,}6 + log_2 0{,}8 + log_2 0{,}6 + log_2 0{,}8$$

Algoritmo voraz para búsqueda de motivos de perfil

Utilice los l-meros P-más probables para ajustar nuevas posiciones de inicio hasta que alcancemos el "mejor" perfil; esto será el motivo.

- Seleccione posiciones iniciales aleatorias, luego:
  1. Cree un perfil P a partir de los *l-meros* en estas posiciones iniciales.
  2. Encuentre el *P-l-mero a* más probable en cada secuencia y cambie las posiciones iniciales a las posiciones iniciales de las *a*s.
  3. Vaya al paso 1 y repita hasta que no podamos aumentar más la puntuación.

Resumen del descubrimiento de motivos algoritmo voraz

- Dado que elegimos posiciones iniciales al azar, hay pocas posibilidades de que nuestra suposición esté cerca de un motivo óptimo, lo que significa que llevará mucho tiempo encontrar el motivo óptimo.
- En la práctica, este algoritmo se ejecuta muchas veces con la esperanza de que las posiciones iniciales aleatorias estén cerca de la solución óptima simplemente por casualidad.
- El algoritmo puede mejorarse mediante el conocimiento heurístico, donde aproximadamente deberíamos comenzar o mediante técnicas estadísticas más sofisticadas, como el *muestreo de Gibbs* que estima las posiciones iniciales más probables para los motivos, donde deberíamos comenzar nuestro proceso iterativo de descubrimiento de motivos.

Tarea: Muestreo de Gibbs

Investigar sobre el algoritmos de «Gibbs sampling», específicamente crear un ensayo corto en español o inglés plano que trate los siguientes puntos:

- En que consiste el algoritmos o método de «Gibbs sampling»
- Cómo soluciona este algoritmo el procedimiento voraz de búsqueda de mótivos.
- Los dos algortimos usados comunmente en la búsqueda de motivos son «Gibbs sampling» y «expectation-maximization», qué diferencias existen entre estos dos algoritmos.