Señales en el ADN

Bioinformática, 5-4-17

Kevin Yip-CSE-CUHK
(Universidad china de Hong-Kong)

Señales en el ADN

- Hemos visto cómo determinar y comparar secuencias de ADN
- ¿Cómo podemos averiguar el **significado biológico** de estas secuencias de ADN?
- Queremos encontrar zonas interesantes de ADN, las **señales**
- De esas señales hoy nos centraremos en las **posiciones de unión** de proteínas al ADN

Señales en el ADN

- Para encontrar señales nos fijaremos sobre todo en subcadenas frecuentes
- Estas cadenas frecuentes no tienen que ser idénticas en todas las apariciones
- Eso complica mucho el problema informático

Hoy …

1. **Definiciones generales**
2. Motivos en ADN

Posiciones de unión con proteínas

- Supongamos que una proteína se puede unir a ciertas posiciones de cada una de las siguientes secuencias:
  - AACCCGATACAGACGACCATTACGACC
  - GAGACGACATACATTACACCAA
  - CCGACTAAACCAGATACAGAGATTACAGCATAC
  - ACATCCATACAGACAAAAACATAGAGGGACGATT
- ¿Dónde se une la proteína?
  - Supuesto: La proteína reconoce un cierto patrón como posición de unión (debido a la forma y energía)

Posiciones de unión con proteínas

- Sería más fácil si se conoce la posición de unión en una de las secuencias:
  - AACCCGATACAGACGACCATTACGACC
  - GAGACGACATACATTACACCAA
  - CCGACTAAACCAGATACAGAGATTACAGCATAC
  - ACATCCATACAGACAAAAACATAGAGGGACGATT

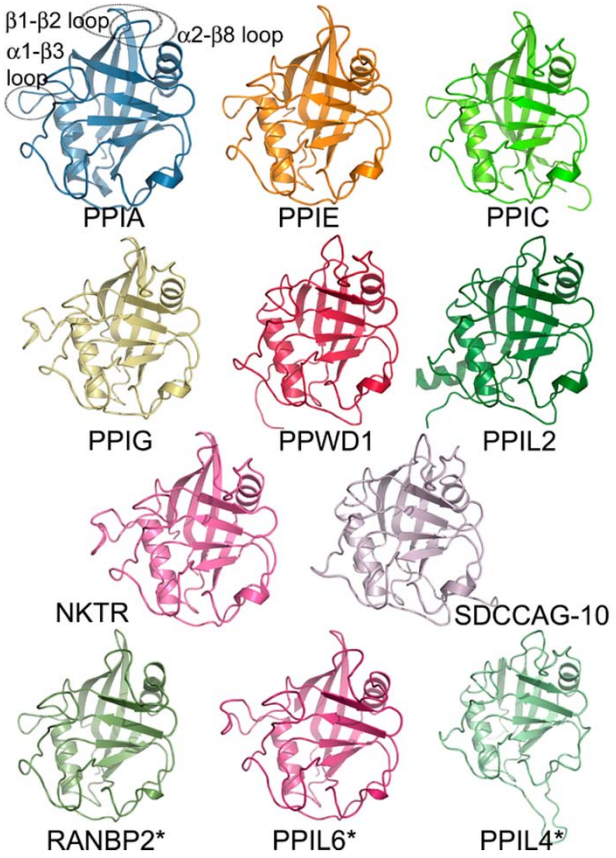
Posiciones de unión con proteínas

- A partir del ejemplo, puede utilizar la alineación de secuencias u otros métodos para averiguar la posición de unión de las otras secuencias:
  - AACCCGATACAGACGACCATTACGACC
  - GAGACGACATACATTACACCAA
  - CCGACTAAACCAGATACAGAGATTACAGCATAC
  - ACATCCATACAGACAAAAACATAGRAMOAGGGACGATT
- Notemos que:
  - Los diferentes sucesos podrían ser ligeramente diferentes
  - Puede haber múltiples sitios de unión en una secuencia

Motivos y dominios

- En general, definimos motivos / dominios como patrones que
  1. aparecen con frecuencia
     - puede no ser exactamente el mismo en diferentes apariciones
     - Es poco probable que se produzca “por casualidad”. En otras palabras, están “sobrerepresentados”
  2. Por lo general, sus roles funcionales se han conocido o predicho
  3. Están evolutivamente conservados
- Hay muchos tipos de motivos y dominios

Los dominios de proteínas:

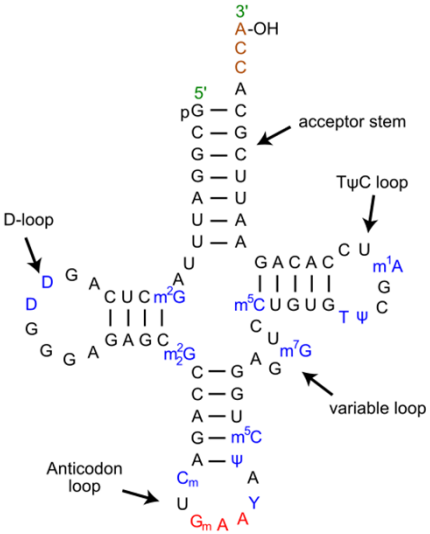


motivos de secuencia de ADN:

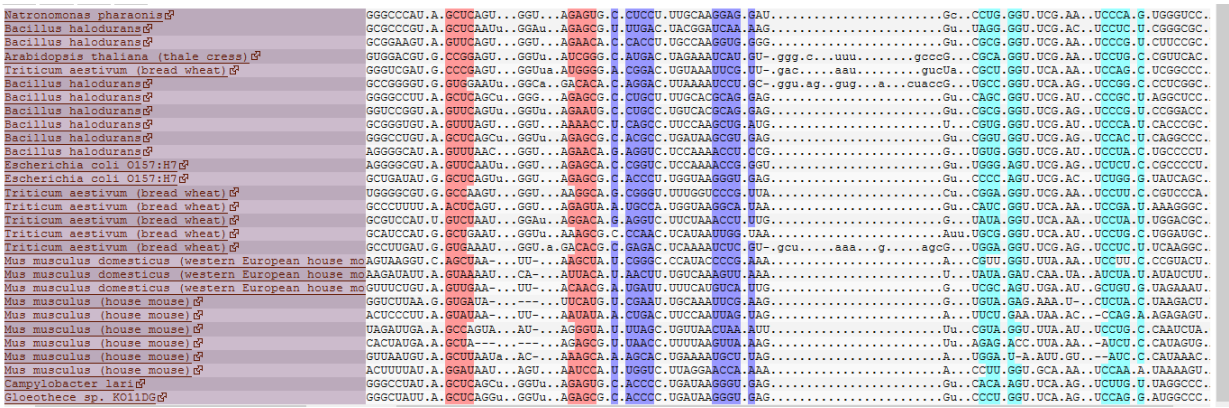
NAME	START	SITES	END	STRAND	MARGINAL SCORE
1	23	aaacgacgtaatgctacg	6	-	22.9
2	17	aaacgac.taat.ctacg	2	-	8.45
3	30	caacgaggta.tgcaacg	14	-	14.1
4	23	caaccacgtaatgcaacg	6	-	23.6
5	24	caaccacgtaatgcatag	7	-	17.5

Score: **51.1928**

logo with ssc



familias de ARN:



Fuentes de las imágenes: http://rosalind.info/media/problems/meme/logo1.png, Wikipedia, Rfam

motivos de ADN y dominios de proteínas

- Veremos:
  - Sitios de unión de factor de transcripción, que son secuencias reguladoras de ADN cortas que frecuentemente aparecen en ubicaciones genómicas específicas. Algunas de ellas se conservan en todas las especies.
  - Los dominios de proteínas, que son sub-secuencias similares en diferentes proteínas que tienen funciones particulares. Algunas de ellas están conservadas evolutivamente.

Motivos y dominios: Diferencias

- En la literatura, a veces motivos y dominios se distinguen por lo siguiente:
  - Un dominio se supone que poseer cierta independencia funcional o estructural. Un motivo puede que no.
  - Un dominio es generalmente más grande que un motivo.
- Hoy, utilizamos estos dos términos más o menos indistintamente.

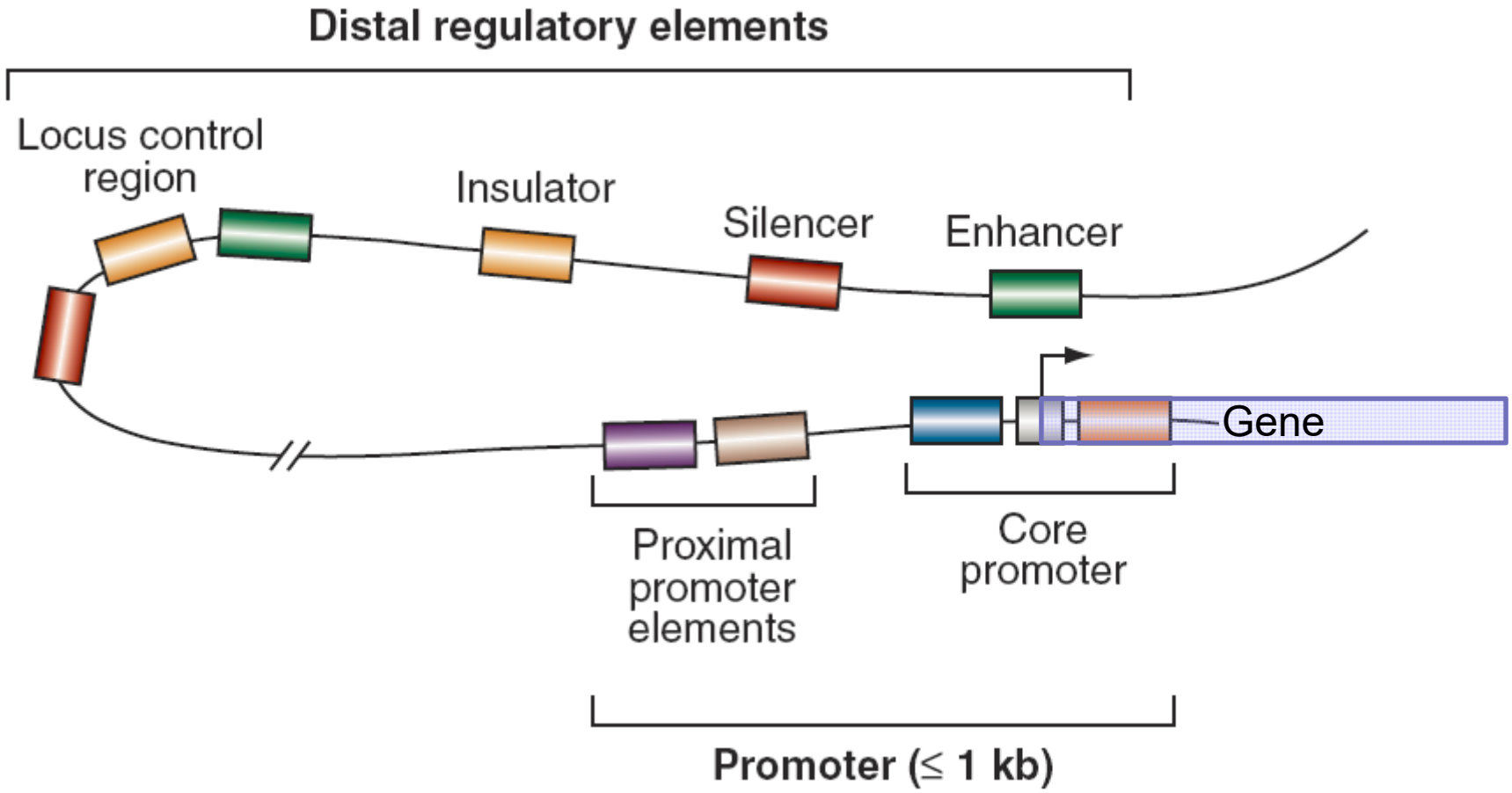
Hoy …

1. Definiciones generales
2. **Motivos en ADN**

Las regiones reguladoras

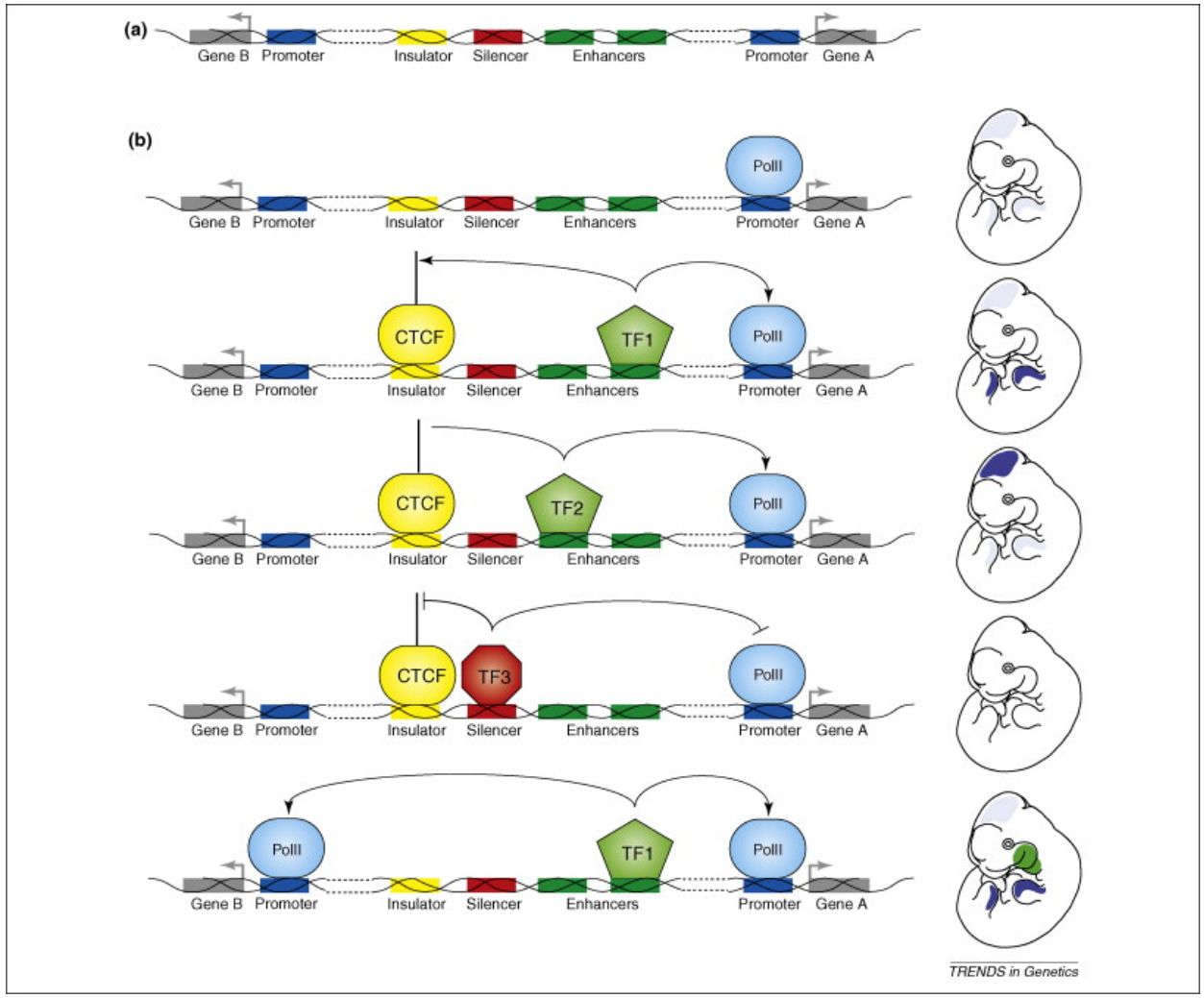
- Algunas proteínas (por ejemplo, los factores de transcripción) se unen a diferentes zonas de ADN para diferentes propósitos. Por ejemplo:
  - Promotores (alrededor del sitio de inicio de la transcripción): Para ayudar a la formación de la maquinaria de transcripción
  - Potenciadores (generalmente más lejos de un gen): Para mejorar la expresión de un gen
  - Silenciadores: para inhibir la expresión de un gen
  - Aisladores: Para marcar los límites de expresión

Las regiones reguladoras



Crédito de la imagen: Maston et al., *Annual Review of Genomics and Human Genetics* 7: 29-59, (2006)

Un ejemplo: circuito de regulación



Sin unión de TF: sólo expresión basal del gen A

Potenciador unido a TF1 (en las extremidades): elevación de la expresión del gen A

Potenciador unido a TF2 (en el cerebro): elevación de la expresión del gen A

Silenciador unido a TF3: la expresión del gen A está inhibida

Sin aislante unido a CTCF: la unión de TF1 en potenciador puede afectar tanto el gen A como al gen B

Crédito de la imagen: Sholtis y Noonan, *Trends in Genetics* 26 (3): 110-118, (2010)

Unión de factor de transcripción

- ¿Dónde se une un factor de transcripción?
  - Donde el ADN es accesible
  - Donde hay señales especiales en el ADN (por ejemplo, la falta de metilación) y las proteínas que rodean (por ejemplo, modificaciones de las histonas)
  - Donde la estructura de ADN es adecuada
  - Donde la secuencia de ADN es adecuada ← HOY
    - La región de ADN unida a un factor de transcripción se llama un sitio de unión a factor de transcripción (TFBS)
    - Por lo general, bastante corto (por ejemplo, 6-10bp)

En representación de motivos y dominios

- ¿Cómo representar un TFBS?
- Si el patrón está muy conservado, puede utilizarse una representación exacta
  - Por ejemplo, secuencia de consenso
- En la mayoría de los casos, se necesita capturar las diferencias por representaciones estadísticas
  - Por ejemplo, matriz de pesos
- Con una representación más compleja,
  - Se puede capturar más información
  - Implica más parámetros
    - Necesita más datos y tiempo para estimar valores de los parámetros
    - Es más propenso a exceso de ajuste

Representaciones (patrones)

- Supongamos que tenemos las siguientes secuencias TFBS:
  - CACAAAA
  - CACAAAT
  - CGCAAAA
  - CACAAAA
- Secuencia de consenso:
  - CACAAAA
  - Problema: La pérdida de información
- Secuencia degenerada con código IUPAC:
  - CRCAAAW

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap (not used in motifs)

Ejemplo: http://conferences.computer.org/bioinformatics/CSB2003/NOTES/Liu_Color.pdf

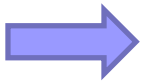
Representaciones (patrones)

- Supongamos que tenemos las siguientes secuencias TFBS:
  - CACAAAAA
  - CACAAA_T
  - CGCAAAAA
  - CACAAA_A
- Expresión regular (véase http://en.wikipedia.org/wiki/Regular_expression para la sintaxis)
  - P.ej, `C [AG] {3,4} CA [EN]`

Representaciones (patrones)

- Matrices de pesos por posición

```
ATGGCATG
AGGGTGCG
ATCGCATG
TTGCCACG
ATGGTATT
ATTGCACG
AGGGCGTT
ATGACATG
ATGGCATG
ACTGGATG
```



	1	2	3	4	5	6	7	8
A	0.9	0.0	0.0	0.1	0.0	0.8	0.0	0.0
C	0.0	0.1	0.1	0.1	0.7	0.0	0.3	0.0
G	0.0	0.2	0.7	0.8	0.1	0.2	0.0	0.8
T	0.1	0.7	0.2	0.0	0.2	0.0	0.7	0.2

- Pseudoconteo: A veces se suma una cantidad a todos para evitar valores pequeños

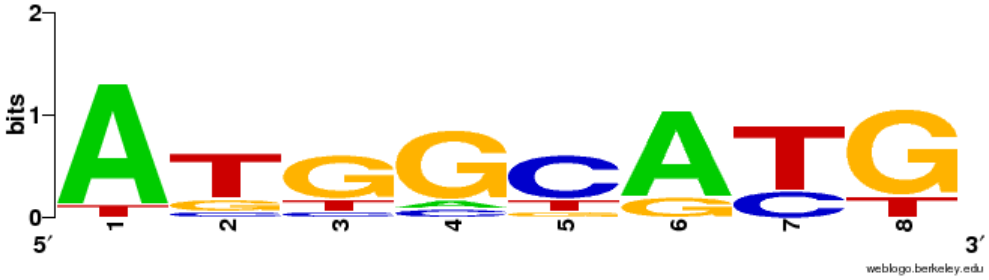
Ejemplo fuente: http://conferences.computer.org/bioinformatics/CSB2003/NOTES/Liu_Color.pdf

Representaciones (patrones)

- Secuencia logo
  - La altura de los nucleótidos depende de la probabilidad
  - La altura total de los nucleótidos en la posición *i*-ésima,

$$h_i = 2 + \sum_{x \in \{A,C,G,T\}} p_{i,x} \log_2 p_{i,x} - \frac{4-1}{2n \ln 2}$$

    - $p_{i,x}$: Probabilidad del carácter *x* en la posición *i*
    - *n*: Número de secuencias
  - Altura de nucleótido *x* = $p_{i,x}\, h_i$



	1	2	3	4	5	6	7	8
A	0.9	0.0	0.0	0.1	0.0	0.8	0.0	0.0
C	0.0	0.1	0.1	0.1	0.7	0.0	0.3	0.0
G	0.0	0.2	0.7	0.8	0.1	0.2	0.0	0.8
T	0.1	0.7	0.2	0.0	0.2	0.0	0.7	0.2

Identificación de motivos sobrerrepresentados

- Las representaciones anteriores son para motivos conocidos: Sabemos las subsecuencias de ADN exactas de la TFBS
- En realidad, ¿cómo encontramos estas secuencias?
  - Hay experimentos que indican aproximadamente los lugares de unión de una proteína
    - Por ejemplo, inmuno-precipitación de cromatina seguida por microarray (ChIP-chip) o secuenciación (ChIP-seq / chip-exo)
  - Si se cree que un TF regula algunos genes al unirse a sus promotores, podemos recoger estas secuencias promotoras
  - En ambos casos, la resolución no es lo suficientemente alta

Problema de descubrimiento de motivos

- Entradas: un conjunto de secuencias, cada una contienen exactamente un TFBS
  - Hay otras variantes:
    - Cada secuencia contiene uno o más TBFS
    - Cada secuencia contiene cero o uno TFBS
    - Cada secuencia contiene cero o más TFBS
- Objetivo: Averiguar las ubicaciones en las secuencias de TFBS
- La idea principal: identificar patrones comunes en las secuencias

Identificación de motivos sobrerrepresentados

- Diferentes métodos:
  1. búsqueda exhaustiva de todas las palabras de tamaño hasta *k*
     - Éxito garantizado si el motivo tiene una tamaño ≤ *k*
     - El coste aumenta exponencialmente con respecto a *k*
       - La indexación ayuda hasta cierto punto
     - Más costoso si se permiten coincidencias inexactas
  2. Multialineamiento de secuencias
     - computacionalmente duro
     - Muchas heurísticas
  3. Se puede integrar información auxiliar para encontrar un sitio de unión activo
     - nivel de expresión génica (por ejemplo, la correlación de número de sitios de unión con el nivel de expresión)
     - Evidencia directa de unión (ChIP-chip o chip-ss)
     - Señales de la cromatina (por ejemplo, la accesibilidad del ADN)

Un ejemplo ilustrativo

- Supongamos que sabemos una proteína se une a algunas posiciones de las secuencias:
  - s_1= ACCGGCT
  - s_2= GTCAGCT
  - s_3= TCGGTAT
- enfoque 3-mer:
  - Por lo tanto, el sitio de unión puede ser alrededor CGG ó GCT

3-mer	Número de secuencias que la contienen
ACC	1
AGC	1
CAG	1
CCG	1
CGG	2
GCT	2
GGC	1
GGT	1
GTA	1
GTC	1
TAT	1
TCA	1
TCG	1

Significancia estadística

- Ahora, considera la siguiente situación:
  - Un cierto genoma contiene 80% C'S y G'S, y 20% A'S y T'S.
  - Tiene 100 secuencias que contienen los sitios de unión de una proteína
    - 90 de ellas contienen el patrón GCGC
    - 85 de ellas contienen el patrón ATAA
  - ¿Cuál crees que es el motivo de unión real?

Significancia estadística

- Estadísticamente significativo: algo es poco probable que suceda por casualidad
  - Puede sugerir importancia biológica (¿por qué?)
- Pasos para determinar la significación estadística:
  - Definir una hipótesis nula (modelo de fondo)
  - Probabilidad de ocurrencia dado el modelo nulo
    - Cálculo directo
    - Simulación (más caro, pero por lo general más realista)

Significancia estadística

- Ejemplo: Para una secuencia de ADN de longitud 4, asumiendo distribución independiente y uniforme de cada base, ¿probabilidad de de observar
  - 3 ó más A'S en una cadena?
    - Considere una hebra: $(0.25)^4 + 4\ (0,75)\ (0,25)^3 = 0,0508$
    - Considere ambas hebras: $0,0508 \times 2 = 0,1016$
  - 2 ó más AT'S en una cadena?
    - Considere una hebra: $(0.25)^4 = 0,0039$
    - Considere las dos hebras: 0,0039 (¿por qué no duplicado?)
- Hay que evitar el conteo redundante en general

Significancia estadística

- ¿Cuál sería un buen modelo nulo para una secuencia de ADN?
  - ¿Independiente, uniforme?
    - No es del todo cierto
  - ¿dependencia local, uniforme?
    - Mejor, pero todavía falta de distribución global
  - ¿dependencia local, no uniforme?
    - Bueno, pero más difícil de manejar
- En general, para conseguir una distribución nula más realista
  - Toma de muestras de datos permutados
  - Preservando al mismo tiempo algunas propiedades clave
    - Para el ADN, podemos querer conservar frecuencias de nucleótidos, frecuencias de dinucleótidos, etc.
  - Por último, obtener una distribución de las muestras y ver dónde está el número observado en la distribución

otras lecturas

- D'haeseleer, How does DNA Sequence Motif Discovery Work? *Nature Biotechnology* 24(8):959-961, (2006)