Estadística de secuencias genómicas

Rodrigo Santamaría

Estadística de secuencias genómicas

Objetivo

Modelos probabilísticos

Significación estadística

Modelos ocultos de Markov

Objetivo

- Encontrar “estructuras” de interés (p. ej. genes) en secuencias
  - Muy largas (millones de elementos en cada secuencia)
  - Sin información “relevante” desde el puntos de vista biológico
  - Difíciles de distinguir del ruido
- Necesidad de herramientas sofisticadas
  - Estadística: modelos probabilísticos
  - Minería de datos: algorítmica e inteligencia artificial

Estadística de secuencias genómicas

Introducción

Modelos probabilísticos

- Definiciones
- Modelo Multinomial
- Modelo de Markov
- Ejemplo

Significación estadística

Modelos ocultos de Markov

Modelos probabilísticos

Todos los modelos están equivocados,

pero algunos son útiles

George E. P. Box

Modelo

- Realidad biológica
  - ADN: molécula compleja tridimensional
- Modelo
  - ADN: secuencia unidimensional de símbolos de un alfabeto
    - A, C, G, T
  - Modelo muy poderoso: permite desarrollar gran cantidad de soluciones informáticas
  - Modelo incorrecto: simplifica la realidad.

Modelo

- Un modelo es en el fondo una propuesta que trata de encontrar un patrón en la forma en la que se distribuyen las secuencias reales
  - Se propone un modelo
  - Para una secuencia real dada, se ve si el modelo se ajusta a la realidad
    - Si se ajusta, nos da una explicación parcial de su comportamiento
    - Si no, bien se descarta, o se reajustan sus parámetros para ver si así se ajusta

Definición formal secuencia de ADN y genoma

- Una **secuencia de ADN** *s* es una cadena finita construida a partir de un alfabeto N={A,C,G,T} de nucleótidos
- Un **genoma** es el conjunto de todas las secuencias de ADN asociadas a un organismo
- Con este modelo podemos estudiar
  - La estructura interna de las secuencias
  - La similitud entre secuencias
  - La evolución en las secuencias

Definición formal elementos de una secuencia

- $s = s_1 s_2 \ldots s_n$
- Cada nucleótido está representado por s_i ($i=1 \ldots n$)
- Conjunto de posiciones
  - $K=\{i, j, k\} \rightarrow s(K) = s_i s_j s_k$
- Intervalo de posiciones
  - $K[i, j]$ o $K=(i:j) \rightarrow s(i:j) = s_i \ldots s_j$
- Ejemplos:
  - s=ATATGTCGTGCA
  - s(2,4,9)=TTT
  - s(3 : 6) =ATGT
  - s(8) = s_8= G

Definición formal alfabetos

- Nucleótidos
  - N_{ADN}={A,C,G,T} N_{ARN}={A,C,G,U}
  - Tamaño 4
- Aminoácidos
  - A={A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}
  - Tamaño 20
- Codones
  - C={AAA, AAC, AAG, AAT, ACA,ACC,…, TTT}
  - Tamaño 4^3=64

Modelo multinomial

- Asunción: Los nucleótidos son independientes y tienen la misma distribución
  - Las secuencias son generadas por un proceso estocástico (aleatorio) que produce cualquiera de los cuatro símbolos en N
  - Distribución de probabilidad
    - $p=(p_A, p_C, p_G, p_T)$
    - $p_A + p_C + p_G + p_T = 1$
    - No depende de la posición: $p_x = p(\mathbf{s}(i) = x)$
    - Todas las probabilidades son iguales: $p_A=p_C=p_G=p_T=0.25$

Modelo multinomial

- Probabilidad de que una secuencia *s* siga el modelo multinomial:

$$s = s_1 s_2 \ldots s_n$$

$$P(s) = \prod_{i=1}^{n} p(s_i)$$

$p_A = p_T = p_G = p_C = 0.25$

tamaño	3	5	10	15	20	25
P	0,0156	0,00098	9.53E-7	9.31E-10	9.09E-13	8,88E-16

Modelo multinomial

En vez de asumir probabilidades iguales, podemos obtener las probabilidades de un análisis simple del genoma

Haemophilus influenzae

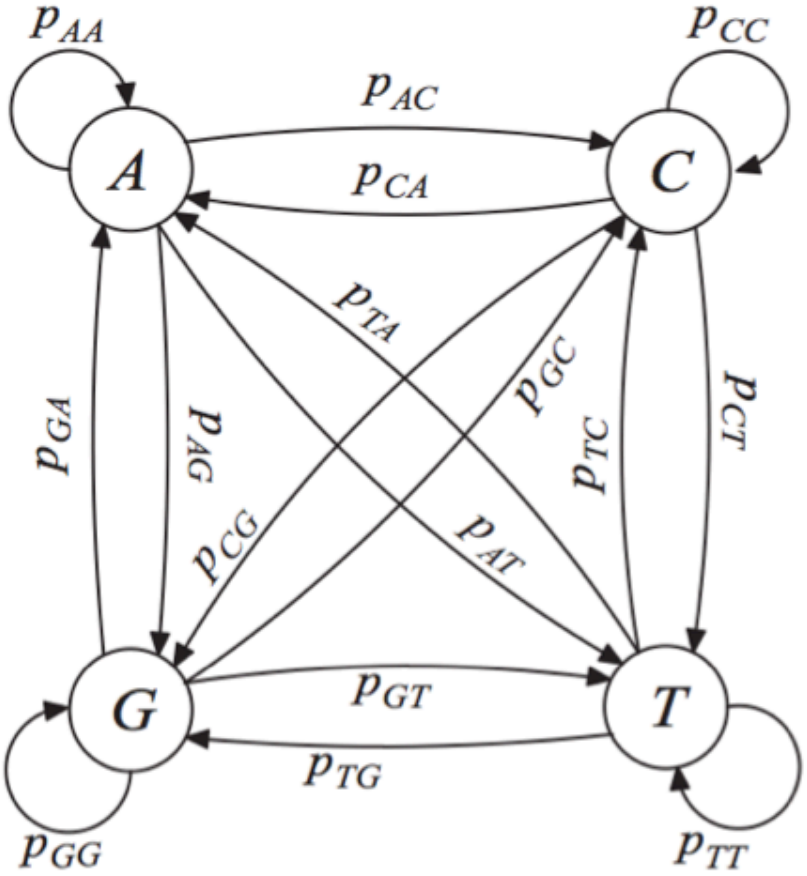
Base	Number	Frequency
A	567,623	0.3102
C	350,723	0.1916
G	347,436	0.1898
T	564,241	0.3083

Modelo de Markov

- Cadena de Markov
  - "La probabilidad de observar un símbolo depende de los símbolos precedentes que hay en la cadena"
  - De orden 1: la probabilidad de cada símbolo sólo depende del que le precede
  - De orden n: la probabilidad de cada símbolo sólo depende de los n que le preceden
- Modelo multinomial = Modelo de Markov de orden 0

Diagrama de transición de estados para orden 1

- Estados (A,C,G,T)
- Probabilidad de transición entre estados (p_{XY})



Matriz de transición

$$T = \begin{matrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CA} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{matrix}$$

$$\pi = \begin{matrix} \pi_A & \pi_C & \pi_G & \pi_T. \end{matrix}$$

- Estados (A,C,G,T)
- Probabilidad de transición entre estados (p_{XY})
- Probabilidad de estado de inicio (π_A,π_C,π_G,π_T)

Probabilidades condicionadas

- Probabilidad de pasar del estado x al y
  - Equivalente a la probabilidad de ver el estado y cuando va precedido por el estado x
  - $p_{xy} = p\,(s_{i+1} = y \mid s_i = x) = P\,(y \mid x) = P\,(x \cap y)/p(y)$

$$P(\mathbf{s}) = P(s_1 s_2 \cdots s_n)$$

$$P(\mathbf{s}) = P(s_n|s_{n-1})P(s_{n-1}|s_{n-2}) \cdots P(s_2|s_1)\pi(s_1)$$

$$P(\mathbf{s}) = \pi(s_1) \prod_{i=2}^{n} p(s_i|s_{i-1}) = \pi(s_1) \prod_{i=2}^{n} p_{s_{i-1}s_i}$$

matriz de transición y tabla de probabilidad

totales		*A	*C	*G	*T
0.3102	A*	0.1202	0.0505	0.0483	0.0912
0.1916	C*	0.0665	0.0372	0.0396	0.0484
0.1898	G*	0.0514	0.0522	0.0363	0.0499
0.3083	T*	0.0721	0.0518	0.0656	0.1189

Matriz de transición de dinucleótidos en *Haemophilus inluenzae*

$$P_{CA} = P(A \mid C) = \frac{P(C \cap A)}{P(C)} = \frac{0.0665}{0.1916}$$

	*A	*C	*G	*T
A*	0.3875	0.1628	0.1557	0.2940
C*	0.3469	0.1941	0.2066	0.2525
G*	0.2708	0.2750	0.1913	0.2629
T*	0.2338	0.1680	0.2127	0.3855

Tabla de probabilidad de transición entre estados en *Haemophilus inluenzae*

Resumen

Modelo	Asunción	Elementos independientes?	Probabilidades iguales?
Multinomial	Frecuencias iguales	Sí	Sí
Multinomial	Frecuencias distintas, basadas en alguna evidencia biológica	Sí	No
Markov	Cada nucleótido depende de los *n* anteriores (orden *n*)	No	No

Modelos: ejemplo

- Modelo multinomial con frecuencias iguales
- Modelo multinomial con frecuencias distintas
- Modelo de Markov de orden 1
  - Asumiendo frecuencias de inicio iguales

Modelo multinomial (frec. iguales)

	Probabilidad
A	0.25
C	0.25
G	0.25
T	0.25

Modelo multinomial

	Probabilidad
A	0.3102
C	0.1916
G	0.1898
T	0.3083

Modelo de Markov de orden 1

	*A	*C	*G	*T
A*	0.3875	0.1628	0.1557	0.2940
C*	0.3469	0.1941	0.2066	0.2525
G*	0.2708	0.2750	0.1913	0.2629
T*	0.2338	0.1680	0.2127	0.3855

Basados en el genoma de *H. influenziae*

Ejemplo

- ¿Cuál de estas dos secuencias sería más probable que se correspondiera a *H. influenziae*…
  - …según cada modelo?

A	T	C	G	A	T	C	A	T	G	C	G
A	T	T	G	A	C	A	T	G	A	T	G

Multinomial ($p_A = p_T = p_C = p_G = 0.25$)

A	T	C	G	A	T	C	A	T	G	C	G	Probabilidad
0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	5.96E-8
A	T	T	G	A	C	A	T	G	A	T	G	?
0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	5.96E-8

Multinomial (Distribución calculada para *Haemophilus influenzae*)

A	T	C	G	A	T	C	A	T	G	C	G	
0.31	0.31	0.19	0.19	0.31	0.31	0.19	0.31	0.31	0.19	0.19	0.19	4.18E-8
A	T	T	G	A	C	A	T	G	A	T	G	
0.31	0.31	0.31	0.19	0.31	0.19	0.31	0.31	0.19	0.31	0.31	0.19	1.11E-7

Markov orden 1 (Probabilidades de transición calculadas para *Haemophilus influenzae*)

A	T	C	G	A	T	C	A	T	G	C	G	
0.25	0.29	0.17	0.21	0.27	0.29	0.17	0.35	0.29	0.21	0.28	0.21	4.31E-8
A	T	T	G	A	C	A	T	G	A	T	G	
0.25	0.29	0.39	0.21	0.27	0.16	0.35	0.29	0.21	0.27	0.29	0.21	9.00E-8

Ejemplo

- Distintos modelos nos dan distintos resultados
  - Cuanta más información contenga el modelo, mejor:
    - Frecuencias de cada nucleótido
    - Longitud de los n-gramas (o k-mers)
      - Secuencias de nucleótidos de tamaño n
      - Markov de orden n → (n+1)-gramas
      - La longitud no puede ser infinita, dependerá del
        - Coste computacional del cálculo
        - Longitud de las secuencias a comparar

Ejemplo

- Como vemos, un modelo siempre nos da un resultado (una probabilidad)
  - Necesitamos dar significado a ese resultado: ¿la cadena se corresponde o no con el modelo?
    - *Significado estadístico*:
      - determinar la probabilidad de que sí se corresponda
      - o visto de otra manera, determinar la probabilidad de que la correspondencia observada sea sólo fruto del azar.
    - *Significado biológico*: buscar razones biológicas para la adecuación (o no) del modelo a la secuencia

Estadística de secuencias genómicas

Objetivo

Modelos probabilísticos

Significación estadística

Contraste de hipótesis

Múltiples contrastes de hipótesis

Odds ratio

Modelos ocultos de Markov

Sentido estadístico

- Se busca determinar si el patrón (del tipo que sea) que hemos encontrado es o no fruto del azar
- Hay distintas opciones, pero básicamente se trata de **comparar la frecuencia observada con la frecuencia esperada** según un determinado modelo
  - Nos define la probabilidad de que lo encontrado se parezca al modelo o no.
    - Si el modelo es una distribución aleatoria, determinaremos si la frecuencia observada es fruto del azar o no
    - Si el modelo es una distribución definida, determinaremos si la frecuencia observada sigue dicho modelo o no.

Contraste de hipótesis

- Al encontrar un patrón en una secuencia, debemos considerar que lo podríamos haber encontrado **por casualidad**
  - En una secuencia aleatoria también lo podríamos haber encontrado
- **P-valor**: probabilidad (entre 0 y 1) de que lo hayamos encontrado por casualidad
  - Típicamente, se admiten entre un 5% y un 0.1% de fallos
    - El umbral de admisión se suele llamar α

Contraste de hipótesis

- Hipótesis nula (H_0): Un patrón ha sido encontrado al azar
- Hipótesis alternativa: no ha sido encontrado al azar
- P-valor p: probabilidad de encontrar el patrón debido al azar
  - $p < \alpha$: se rechaza la hipótesis nula → no es fruto del azar
  - $p > \alpha$: se acepta la hipótesis nula → es fruto del azar
  - α suele ser 0.05, 0.01, 0.001 (5%, 1% ó 0.1%)

Ejemplo: codón de paro

- Un codón es una secuencia de 3 nucleótidos
  - 4^3 → 64 codones distintos
- Open Reading Frame (ORF)
  - Secuencia de nucleótidos que comienza por el codón de inicio (ATG) y termina por un codón de paro (TAA, TAG, TGA)
  - Un ORF suficientemente largo puede ser un gen
    - Deberemos determinar si la probabilidad de encontrar una secuencia con un codón de inicio y uno de paro de su misma longitud *k* es probable en cadenas aleatorias o no.

Ejemplo: codón de paro

- En una secuencia aleatoria, conforme a un modelo multinomial (asumiendo distribución uniforme de codones) tenemos:
  - 3/64 = probabilidad de "elegir" un codón de paro
  - 61/64 = probabilidad de no "elegir" un codón de paro
- La probabilidad de tener por pura suerte una secuencia de *k* o más codones non-stop será:
  - $P(s_k \text{ non-stop})=(61/64)^k$
  - $(61/64)^{62}=0.051 \sim 5\%$
  - $(61/64)^{100}=0.0082 \sim 1\%$

Descartando los ORFs con k<64 (62+start+stop) eliminamos el 95% de ORFs falsos

Ejemplo: codón de paro

- Si consideramos una distribución no uniforme de codones (*M. genitalium*)
  - P(stop)=P(TAA)+P(TAG)+P(TGA)=0.039+0.016+0.021=0.076 > 3/64=0.047
  - P(k non-stop)=$[1-P(stop)]^k$
  - $(1-0.076)^{38}=0.0496$ ~5% $(1-0.076)^{58}=0.0102$~1%

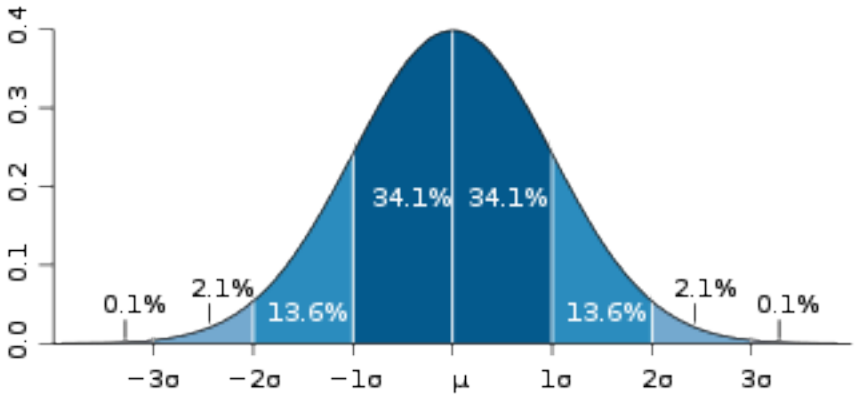
```
>> Mgenit = genbankread('NC_000908.gb');
>> seq=Mgenit.Sequence;
>> codoncount(seq)
AAA - 11510     AAC -  5295     AAG -  4627     AAT -  6645
ACA -  3229     ACC -  2325     ACG -   607     ACT -  3823
AGA -  3066     AGC -  2460     AGG -  1738     AGT -  3920
ATA -  3786     ATC -  3705     ATG -  3034     ATT -  6909
CAA -  5807     CAC -  1804     CAG -  2150     CAT -  2921
CCA -  2307     CCC -  1060     CCG -   299     CCT -  1680
CGA -   455     CGC -   494     CGG -   295     CGT -   634
CTA -  3199     CTC -  1201     CTG -  2035     CTT -  4781
GAA -  3961     GAC -   901     GAG -  1356     GAT -  3877
GCA -  2323     GCC -   594     GCG -   502     GCT -  2471
GGA -  1590     GGC -   612     GGG -  1095     GGT -  2213
GTA -  2136     GTC -   858     GTG -  1918     GTT -  5229
TAA -  7536     TAC -  2233     TAG -  3021     TAT -  3960
TCA -  3670     TCC -  1514     TCG -   426     TCT -  2775
TGA -  4124     TGC -  2085     TGG -  2209     TGT -  3104
TTA -  7561     TTC -  3646     TTG -  5264     TTT - 10793
```

Muestras aleatorias

- Para saber lo probable que es encontrar un patrón al azar, tenemos que diseñar una técnica para generar secuencias (o, en general, muestras) aleatorias
  - Podemos suponer una distribución aleatoria de nucleótidos
    - Bien suponiendo las mismas probabilidades para cada nucleótido o las frecuencias conocidas para un determinado organismo
  - O podemos reordenar aleatoriamente los nucleótidos (elementos) la secuencia (muestra) real
    - Asegura mantener las propiedades estadísticas de la secuencia real

Muestras aleatorias

- Una vez elegido el método de aleatorización, generamos N muestras aleatorias y calculamos la similitud de nuestro patrón con respecto a las secuencias.
- Idealmente, las puntuaciones aleatorias seguirán una distribución normal o gaussiana
  - Tendremos una media de las puntuaciones aleatorias μ
  - Y una desviación estándar de dichas puntuaciones σ
  - Esto no siempre es así, como veremos en BLAST

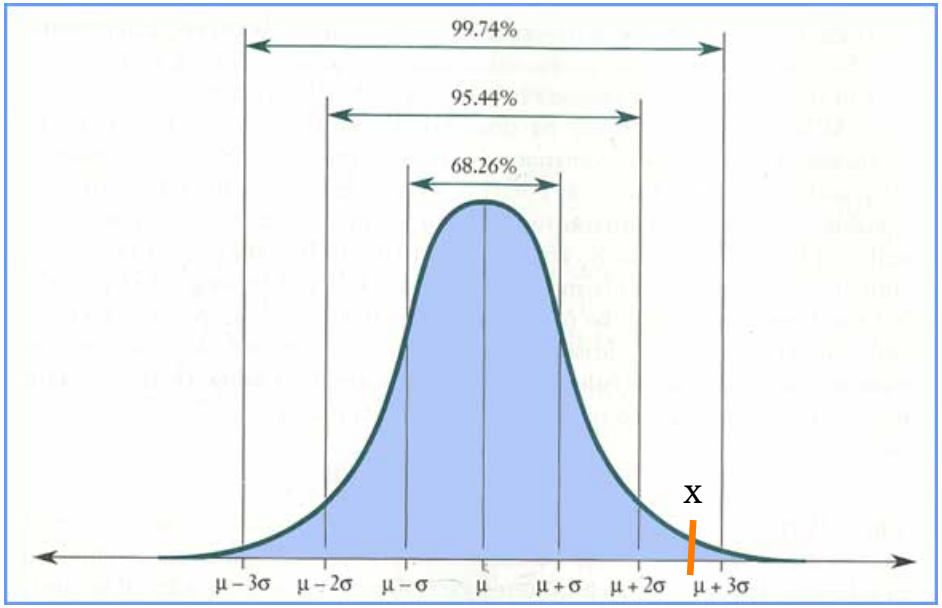


Muestras aleatorias

- Podemos calcular lo desviada que está la puntuación sobre la secuencia real (x) respecto a las puntuaciones sobre secuencias aleatorias, mediante el Z-score:

$$Z = \frac{x - \mu}{\sigma}$$

- Mediante tablas estadísticas, determinamos la probabilidad de que en una distribución aleatoria tengamos un valor como x → p-valor



Falsos positivos

- **Falso positivo**: aceptar un patrón que se debe al azar
- **Falso negativo**: rechazar un patrón que no se debe al azar
- **Sensibilidad**: capacidad de detectar todos los patrones que no se deben al azar
- **Especificidad**: capacidad de descartar todos los patrones que se deben al azar
- Necesidad de compromiso entre sensibilidad y especificidad

Sensibilidad vs especificidad

- Mi recomendación: *"El cementerio está lleno de héroes"*
  - Es decir, es mejor dar un falso negativo que dar un falso positivo (~es mejor ser conservador en nuestros resultados)
  - Por ejemplo: corrección a la baja del nº de genes en *H. sapiens*
    - Las primeras estimaciones hablaban de 2 millones de genes, que se terminaron reduciendo a 20.000 – 25.000
- $\alpha = 0.001$ es un buen valor
  - Cuidado con los múltiples contrastes de hipótesis

Múltiples contrastes de hipótesis

- Si comparamos un patrón con una secuencia, y tiene un p valor de 0.001, quiere decir que hay una probabilidad de un 0.1% de que se deba al azar → el patrón es bueno
- Si comparamos el patrón con un millón secuencias, tenemos un millón oportunidades de obtener un p-valor bajo
  - Necesidad de corregir los umbrales de significación α

Corrección de Bonferroni

- Es la corrección más simple y más conservadora
- Se divide el umbral α entre el número de tests
  - $\alpha = 0.05$, 10^6 tests → $\alpha(\text{corregido}) = 0.05/10^6 = 5 \cdot 10^{-8}$
- Gana especificidad a costa de reducir la sensibilidad
  - Muy conservador

Carlo Emilio Bonferroni
1892-1960

Familywise Error Rate (FWER)

- En este caso, α (llamado FWER) indica la probabilidad de tener al menos un falso positivo
  - Con Bonferroni α =0.05 en 1000 muestras indica que tendremos como mucho 50 falsos positivos
  - Con FWER, α =0.05 indica que tenemos un 5% de posibilidades de tener 1 o más falsos positivos.

Familywise Error Rate

- Formalmente, $\alpha = FWER = P(V \geq 1)$
  - Siendo V el número de falsos positivos
- Método de Holm-Bonferroni
  1. Se calculan los p-valores de nuestros N tests
  2. Se ordenan: $P_1 \leq P_2 \leq P_3 \ldots \leq P_N$
  3. P_k es significativo si $P_k < \alpha / (N - k)$

False Discovery Rate (FDR)

- Control directo sobre el número de falsos positivos en comparaciones múltiples
- Tasa de falsos positivos en nuestros tests
  - FDR=0.1 en 1000 tests quiere decir que 100 son falsos positivos
- Menos conservador que Bonferroni y FWER
- También conocida como corrección de Benjamini y Hochberg

False Discovery Rate (FDR)

- Método (tests independientes)
  1. Se calculan los p-valores de nuestros N tests
  2. Se ordenan: $P_1 \leq P_2 \leq P_3 \ldots \leq P_N$
  3. Para un valor α, se busca el valor más grande P_k tal que: $P_k < \frac{\alpha k}{N}$
  4. Se toma como FDR (umbral corregido) $Z=P_k$

Corrección de p-valores

Corrección	Significado (α=0.01)	Especificidad
Bonferroni	Como mucho un 1% serán falsos positivos	Muy alta
FWER	Probabilidad de un 1% de tener al menos un falso positivo	Alta
FDR	Exactamente un 1% serán falsos positivos	Moderada

Odds ratio

- El odds ratio es una manera de calcular la probabilidad
  - Similar a la terminología de las apuestas (20 a 1)
- Es una medida de lo probable que es un suceso coincidente (GC) teniendo en cuenta lo probables que son los sucesos por separado (G, C)

Odds ratio

$$odds\_ratio \approx \frac{N(xy)}{N(x)N(y)}$$

Frecuencia observada

Frecuencia esperada (dependerá del modelo de referencia, en este caso para el multinomial)

Odds ratio

base	frecuencia
A	0.3102
C	0.1916
G	0.1898
T	0.3083

	*A	*C	*G	*T
A*	0.1202	0.0505	0.0483	0.0912
C*	0.0665	0.0372	0.0396	0.0484
G*	0.0514	0.0522	0.0363	0.0499
T*	0.0721	0.0518	0.0656	0.1189

$$\frac{0.0522}{0.1916 * 0.1898} = \frac{0.0522}{0.0364}$$

AT en la tabla de frecuencias era el doble de frecuente que GT, pero si tenemos en cuenta que es más común que aparezca una A que que aparezca una G, la cosa cambia

	*A	*C	*G	*T
A*	1.2491	0.8496	0.8210	0.9535
C*	1.1182	1.0121	1.0894	0.8190
G*	0.8736	1.4349	1.0076	0.8526
T*	0.7541	0.8763	1.1204	1.2505

Estadística de secuencias genómicas

Objetivo

Modelos probabilísticos

Significación estadística

Modelos ocultos de Markov

- Definición
- HMM y secuencias
- Matrices de transición y emisión
- HMM y alineamientos

Modelo Oculto de Markov (HMM)

- Es un modelo de Markov en el que no podemos observar los estados directamente
  - Aunque sean conocidos, no podemos saber en qué estado estamos en cada momento
  - Pero podemos inferirlos a partir de observaciones
- Ejemplo: predicción del tiempo en Tokyo
  - Estados: 1 (soleado) y 2 (lluvioso)
    - No podemos observarlos directamente porque no estamos en Tokyo
  - Observación indirecta: un amigo en Tokio nos dice por teléfono que su perro ha salido al jardín

HMM: fundamentos

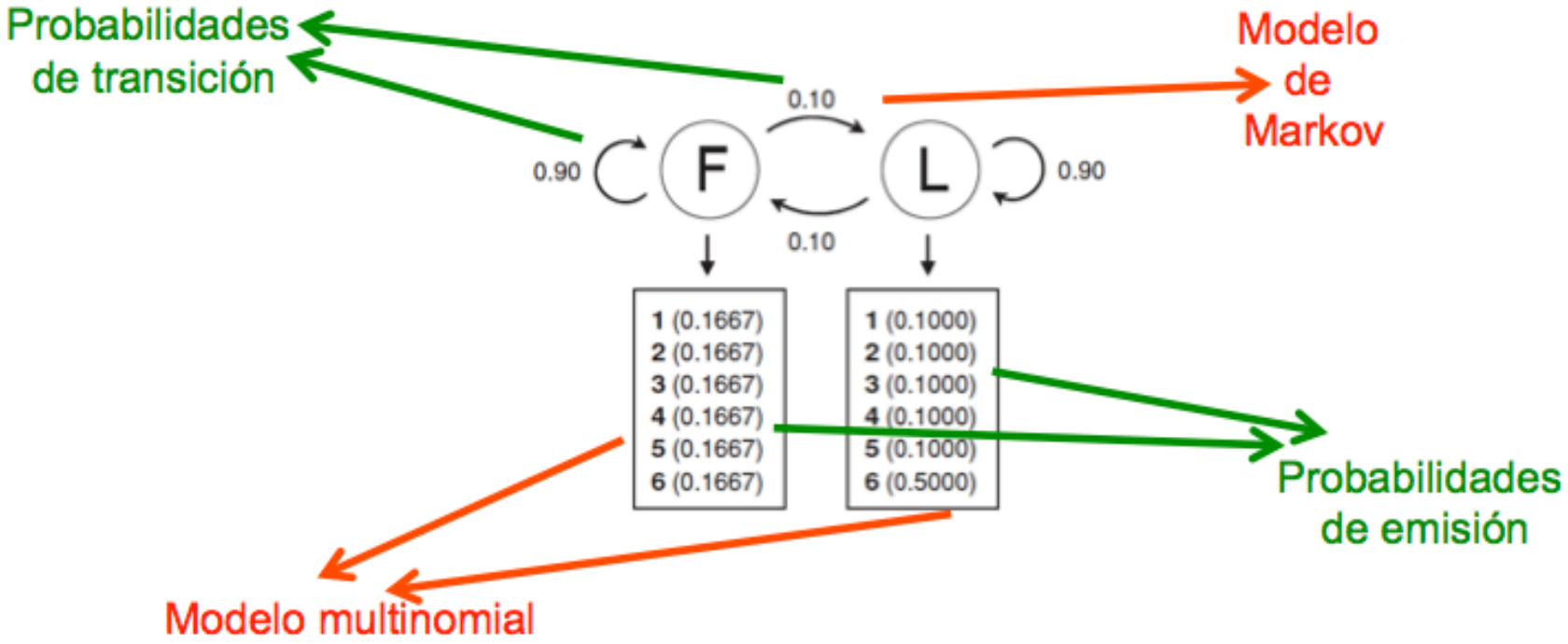
- La secuencia se modela como si fuera generada por una cadena de Markov
- En cada posición tenemos uno o más estados desconocidos (ocultos), lo único que observamos son los símbolos de la secuencia generados de acuerdo a una distribución multinomial que depende de dichos estados desconocidos
- Objetivo: a partir de la secuencia observada (ruidosa) inferir los estados ocultos

HMM: estados ocultos

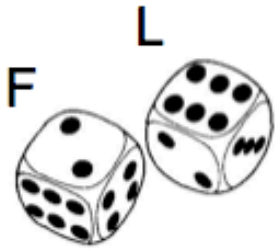
- Ejemplos de estados ocultos en bioinformática
  - Análisis sencillo de secuencias
    - "Rico en GC", "Rico en AT"
  - Análisis complejo de secuencias
    - "Región codificante", "intrón", "terminador", etc.
  - Alineamiento de secuencias
    - "inserción", "deleción" o "alineamiento"

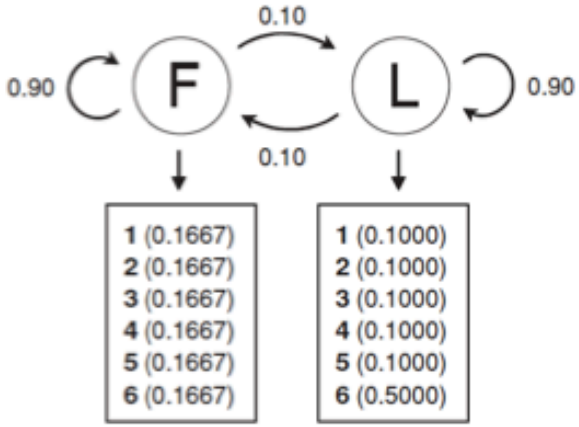
Ejemplo sencillo

- Determinar, a partir de varias tiradas de dados se ha usado un dado legal (F - fair) o uno trucado (L - liar)



Ejemplo sencillo





Secuencia observable

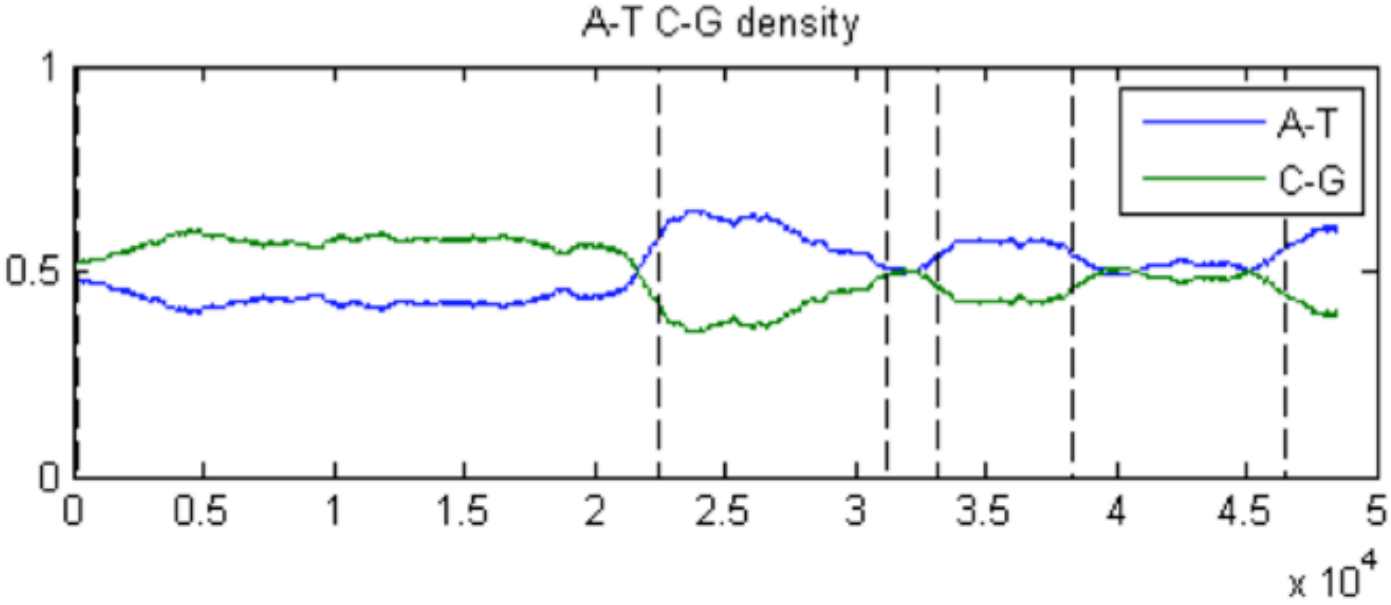
4553653163363555133362665132141636651666

?

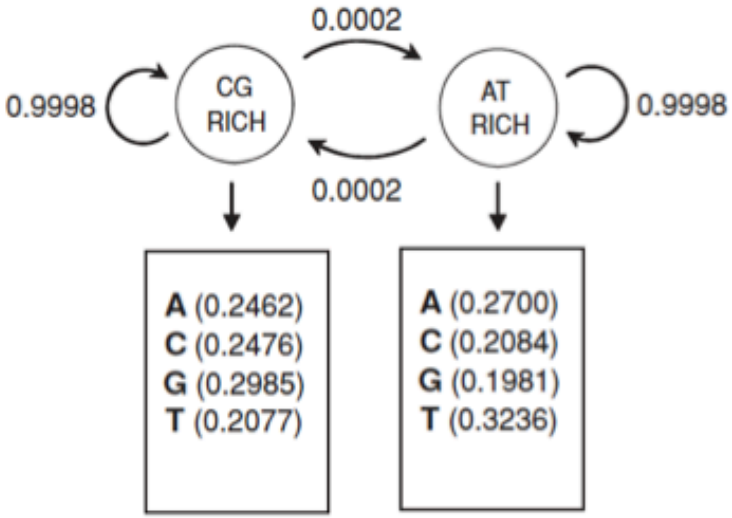
OCULTO FFFFFFFFFFFFFFFFFFFFLLLLFFFFFFFLLLLLLLLL

4553653163363555133362665132141636651666

Análisis sencillo



HMM para
Bacteriófago lambda



Matriz de transición y emisión

- La probabilidad de estar en el estado l, dado que se estaba en el estado k, es la entrada $T(k, l)$ de la matriz de transición
  - $T(k, l) = P(h_i = l \mid h_{i-1} = k)$
- La probabilidad de emitir la salida b desde el estado k viene dada por el modelo multinomial asociado al dicho estado k, y es la entrada $E(k, b)$ de la matriz de emisión
  - $E(k, b) = P(s_i = b \mid h_i = k)$

Probabilidades

- Sea h la secuencia oculta y $P(h)$ la probabilidad de que sea la secuencia correcta
- Sea s la secuencia observada y $P(s \mid h)$ la probabilidad de que se observe dicha secuencia siendo h la secuencia oculta

$$P(h) = P(h_1)\prod_{i=2}^{n} P(h_i \mid h_i - 1) = T(0, h_1)\prod_{i=2}^{n} T(h_i - 1, h_i)$$

$$P(s \mid h) = \prod_{i=1}^{n} P(s_i \mid h_i) = \prod_{i=1}^{n} E(h_i, s_i)$$

Teorema de la probabilidad total

- Determina la probabilidad de que se dé una secuencia observada *s*, si *h* es desconocida (el caso más frecuente)

$$P(\mathbf{s}) = \sum_{\forall \mathbf{h}_j \in \mathcal{H}^n} P(\mathbf{s}, \mathbf{h}_j) = \sum_{\forall \mathbf{h}_j \in \mathcal{H}^n} P(\mathbf{s}|\mathbf{h}_j) P(\mathbf{h}_j)$$

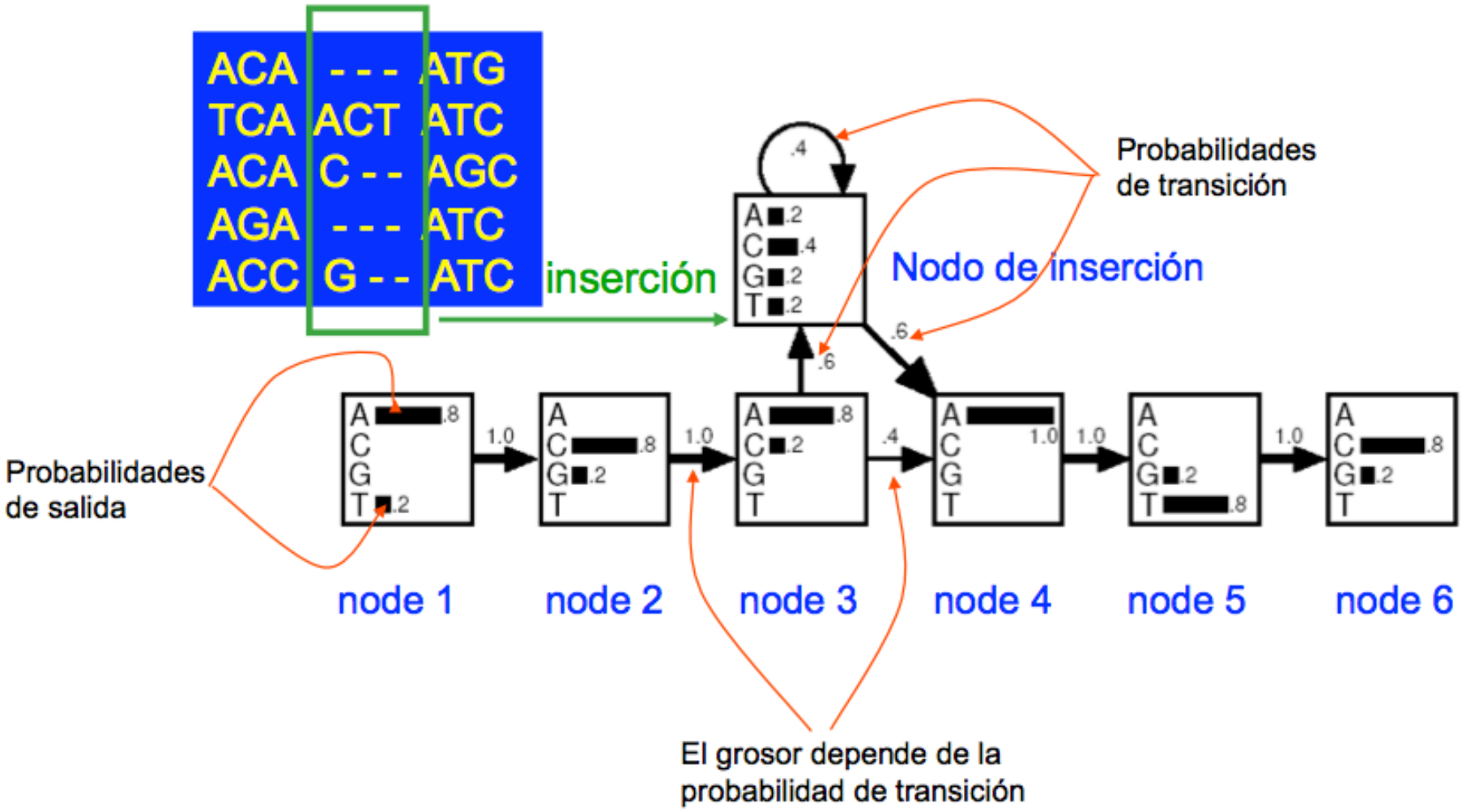
$\mathcal{H}^n$ es el conjunto de todas las posibles cadenas ocultas de longitud igual a la longitud de la secuencia observada

Algoritmo de Viterbi

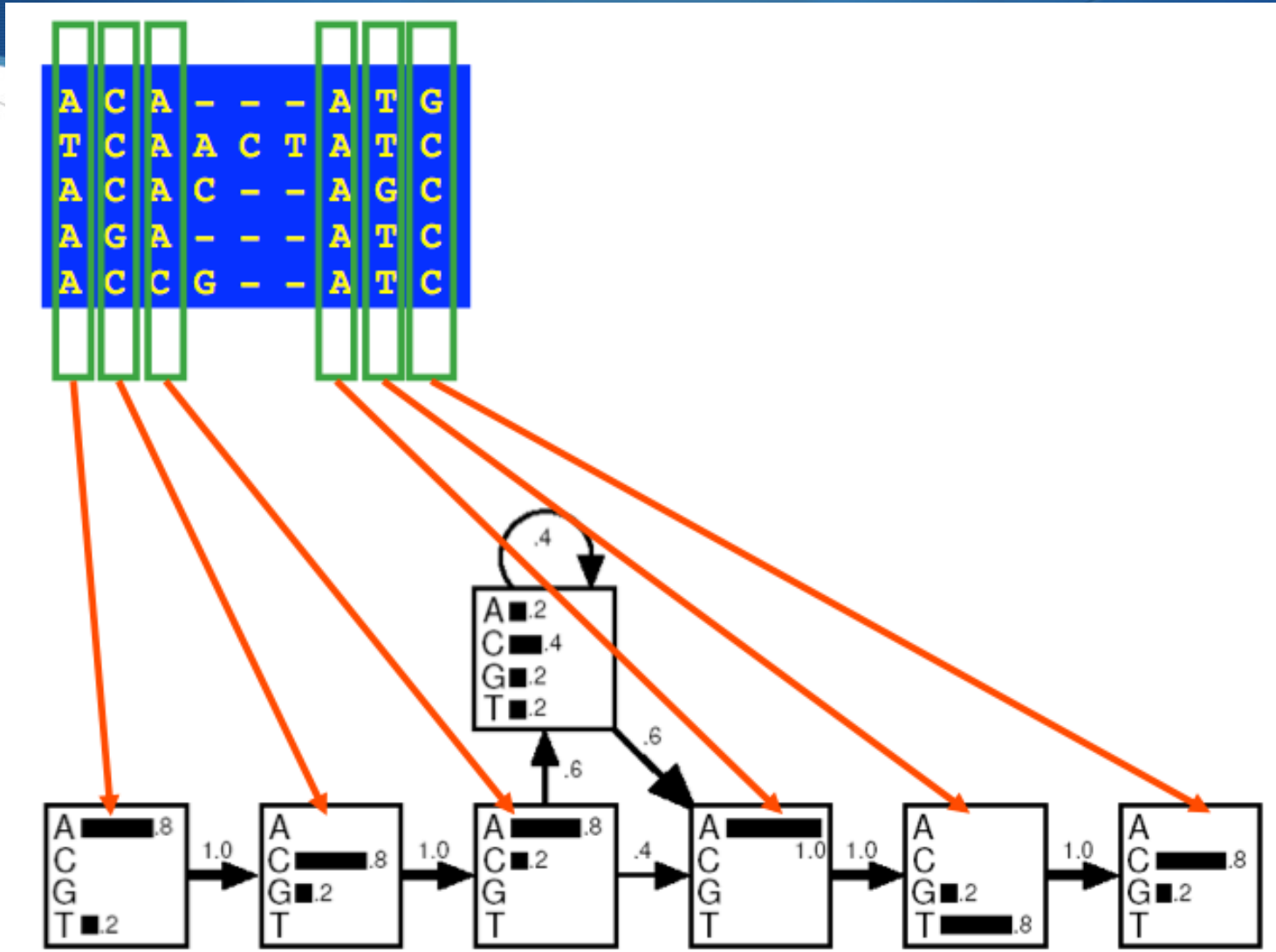
- Calculando todos los $P(s, h)$ podemos determinar cuál es el camino de los estados más probables (o camino de Viterbi):

$$\mathbf{h}^* = \arg \max_{\forall \mathbf{h} \in \mathcal{H}^n} P(\mathbf{s}, \mathbf{h})$$

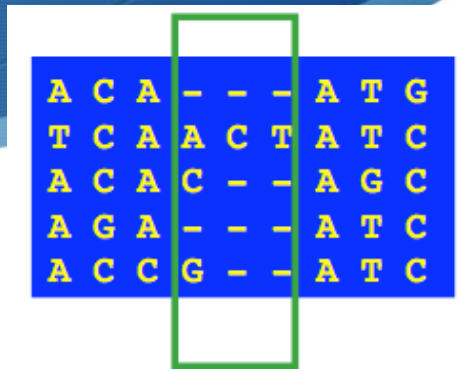
HMM a partir de un alineamiento



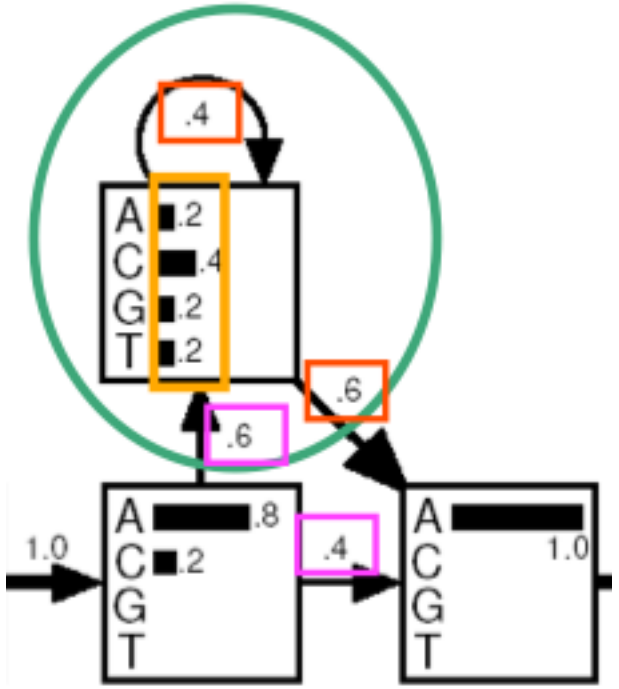
Primeras y últimas columnas



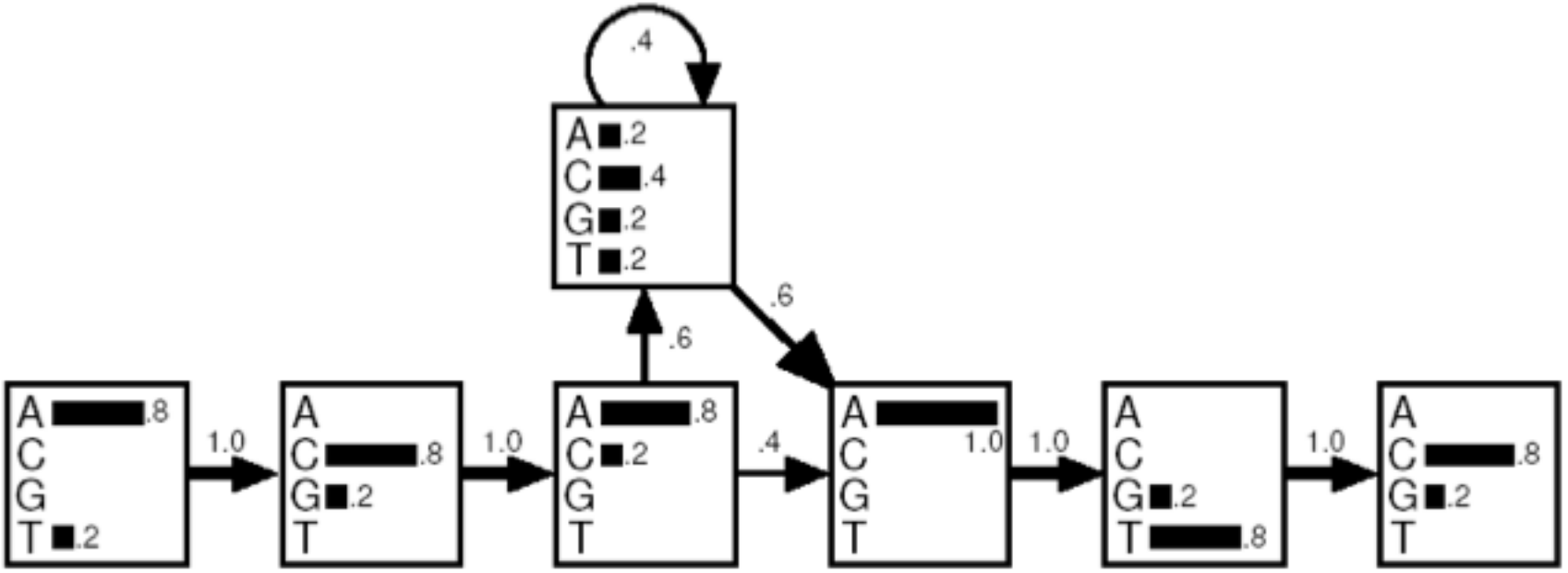
Nodo de inserción



- Las columnas 4,5,6 contienen inserciones en la cadena base de 6 nucleótidos
  - 3 de las 5 columnas tienen alguna inserción (algún nucleótido en las columnas 4,5,6)
    - Probabilidad de inserción es 3/5=0.6
- En el nodo de inserción hay 1A, 2C, 1G y 1T
  - Probabilidades 0.2, 0.4, 0.2, 0.2
- 3 de las 5 cadenas terminan tras una inserción
  - 3/5=0.6 probabilidades de salir de la inserción



Probabilidad de una secuencia



- P(ACACATC)=(0.8·1)·(0.8·1)·(0.8·0.6)·(0.4·0.6) ·(1·1) ·(0.8·1) ·(0.8) ~0.047

Probabilidad de algunas secuencias

	Secuencia	Prob %
Consenso	**ACAC–ATC**	4.7
Secuencia 1	**ACA---ATG**	3.3
Secuencia 2	**TCAACTATC**	0.0075
Secuencia 3	**ACAC--AGC**	1.2
Secuencia 4	**AGA---ATC**	3.3
Secuencia 5	**ACCG--ATC**	0.59
Excepcional	**TGCT--AGG**	0.0023

Problemas con las probabilidades

- Sesgadas por la longitud de la secuencia
  - P(ACAC--ATC) = 0.047
  - P(TCAACTATC) = 0.000075
- Normalización por la longitud de la secuencia L
  - Odd ratio: dividimos la probabilidad por la probabilidad de distribución multinomial estándar: $(0.25)^L$
  - Tomamos el logaritmo: log-odds score

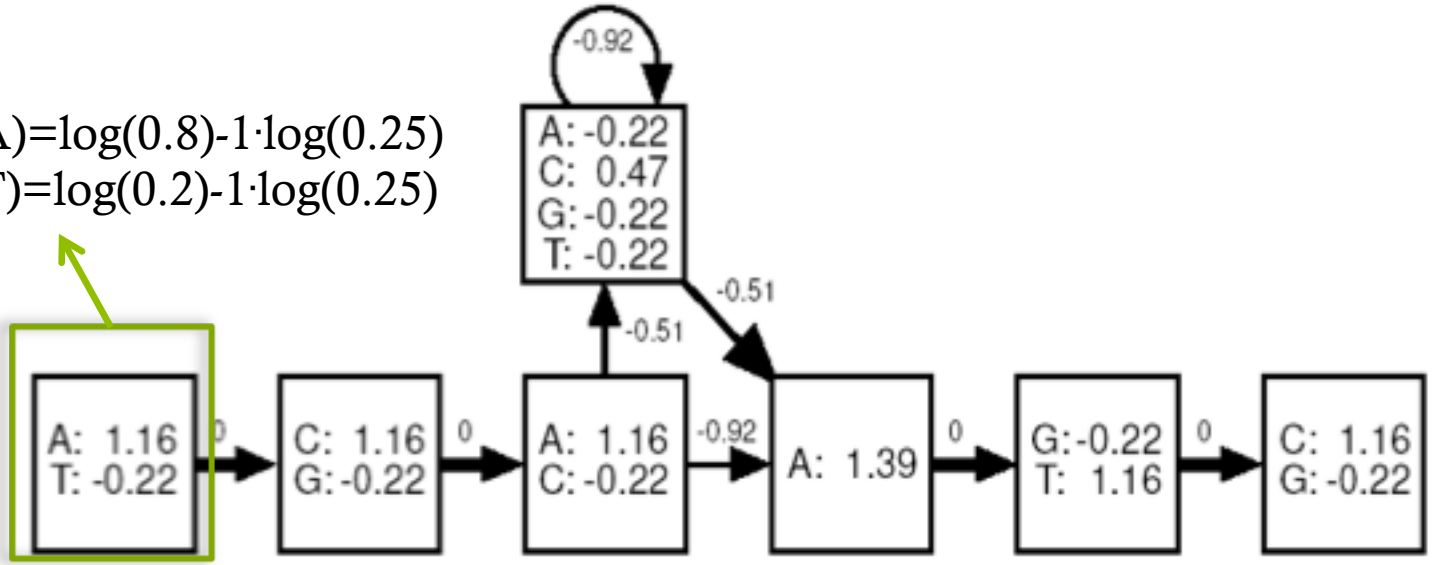
$$\log - odd(S) = \log \frac{P(S)}{0.25^L} = \log P(s) - L \log 0.25$$

Probabilidad y log-odds

	Secuencia	Prob · 100	log-odds
Consenso	ACAC--ATC	4.7	6.7
Secuencia 1	ACA---ATG	3.3	4.9
Secuencia 2	TCAACTATC	0.0075	3.0
Secuencia 3	ACAC--AGC	1.2	5.3
Secuencia 4	AGA---ATC	3.3	4.9
Secuencia 5	ACCG--ATC	0.59	4.6
Excepcional	TGCT--AGG	0.0023	-0.97

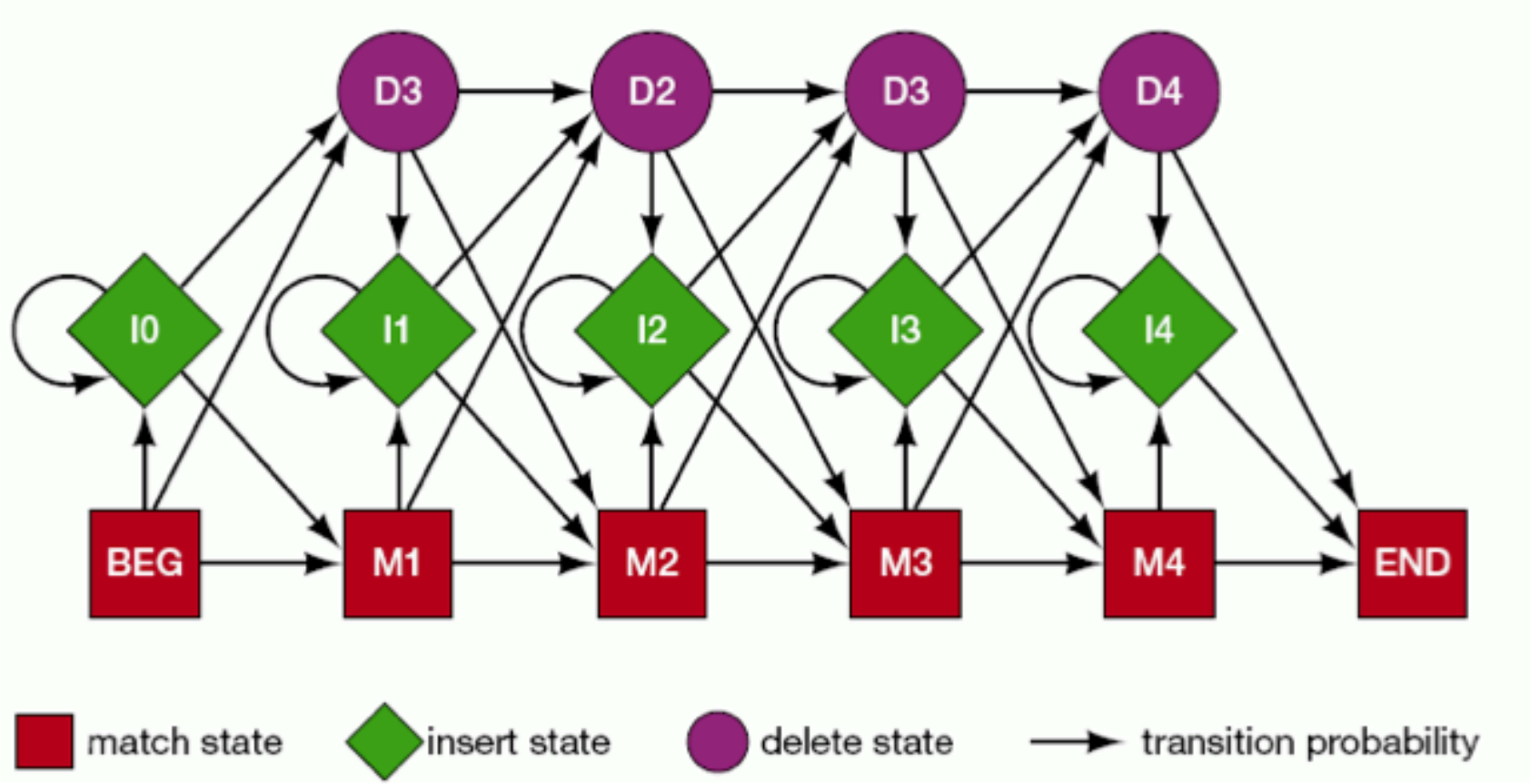
Log-odd de una secuencia

log-odd(A)=log(0.8)-1·log(0.25)
log-odd(T)=log(0.2)-1·log(0.25)



- P(ACACATC)=(1.16+0)+(1.16+0)+(1.16-0.51)+(0.47-0.51)+(1.39+0)+(1.16+0)+1.16

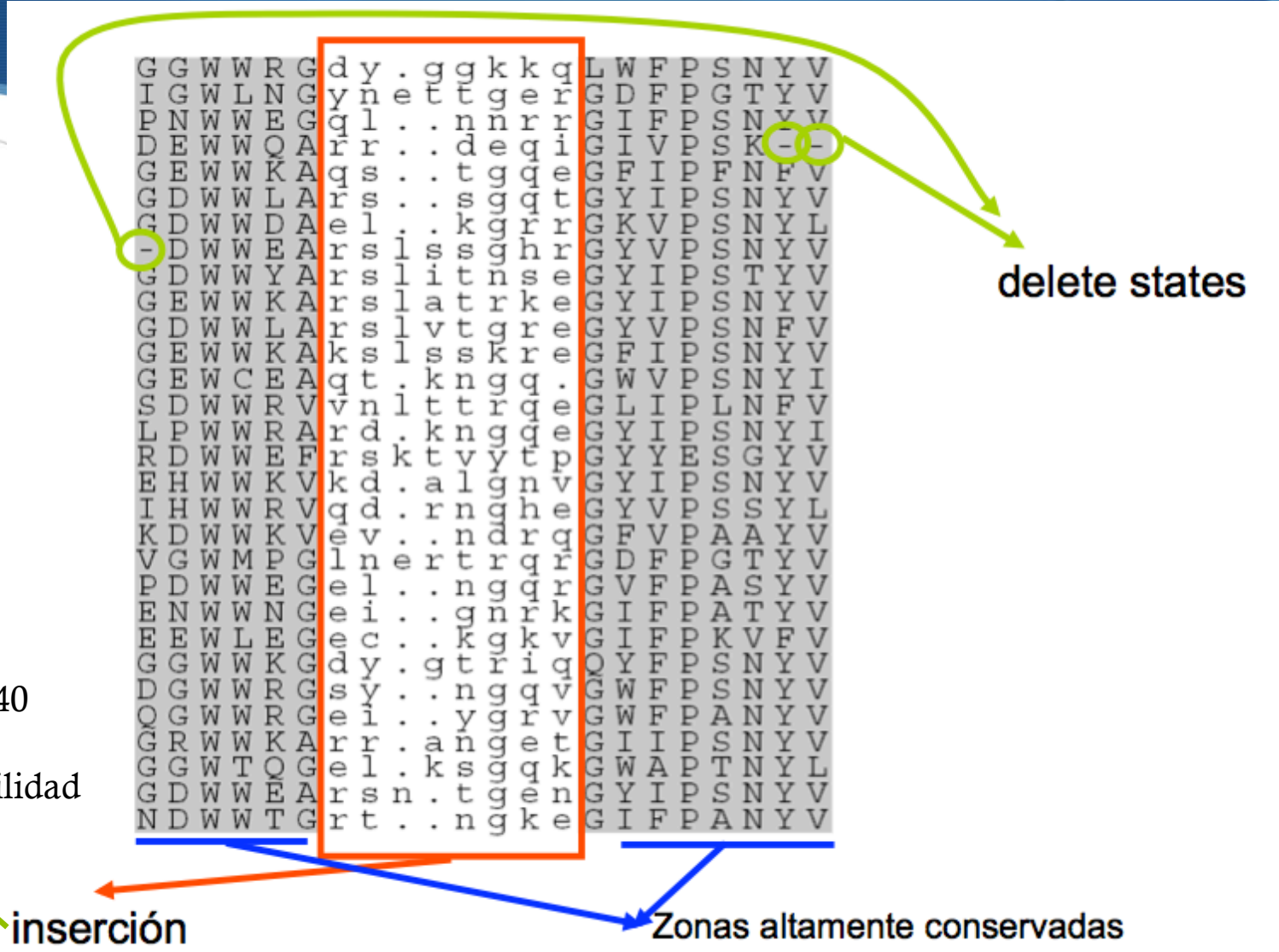
Perfil HMM


D3
D2
D3
D4
I0
I1
I2
I3
I4
BEG
M1
M2
M3
M4
END
match state
insert state
delete state
transition probability

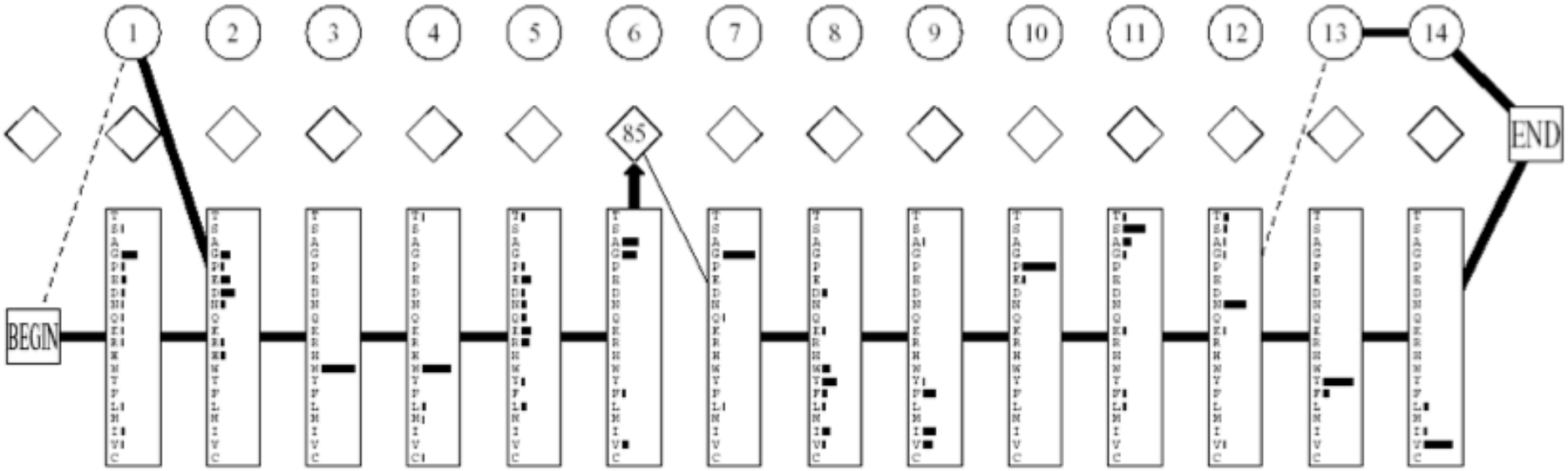
Perfil HMM

- Modelo para alineamientos múltiples de secuencia
- Match state (estado principal o de coincidencia)
  - Modela las regiones conservadas en el alineamiento
  - Probabilidad de distribución: la observada en el MSA
- Insert state (estado de inserción)
  - Modela regiones muy variables en el alineamiento
  - Probabilidad de distribución: basada en el MSA o usar una distribución fija de residuos
- Delete state (estado de deleción)
  - Modela situaciones con pocos huecos

Perfil HMM (ejemplo)



Perfil HMM (ejemplo)



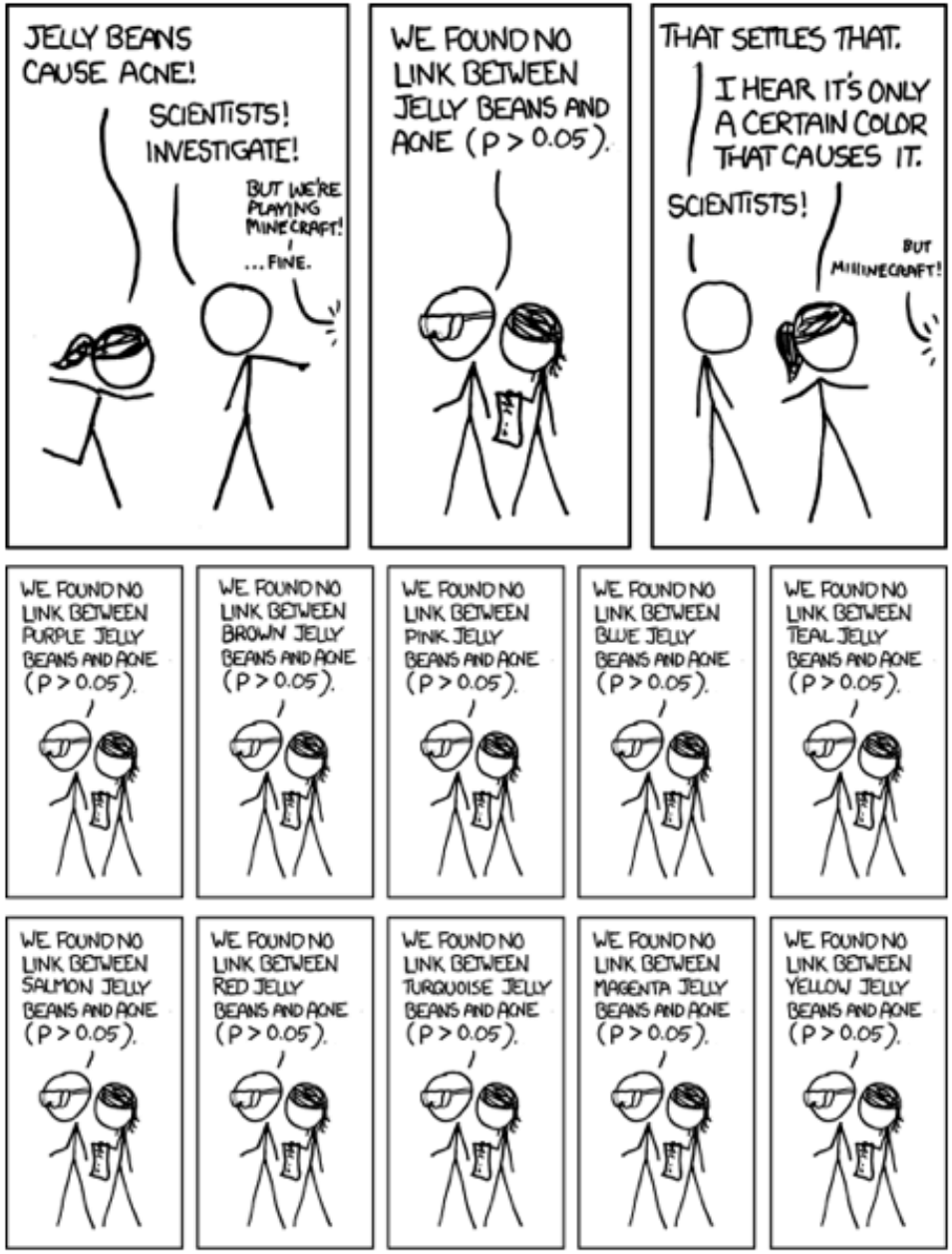
- Transiciones
  - Sin flecha = de izquierda a derecha
  - De un estado de inserción a sí mismo no se muestra
- Probabilidades
  - Grosor de la línea
  - En los estados de inserción, dentro del rombo

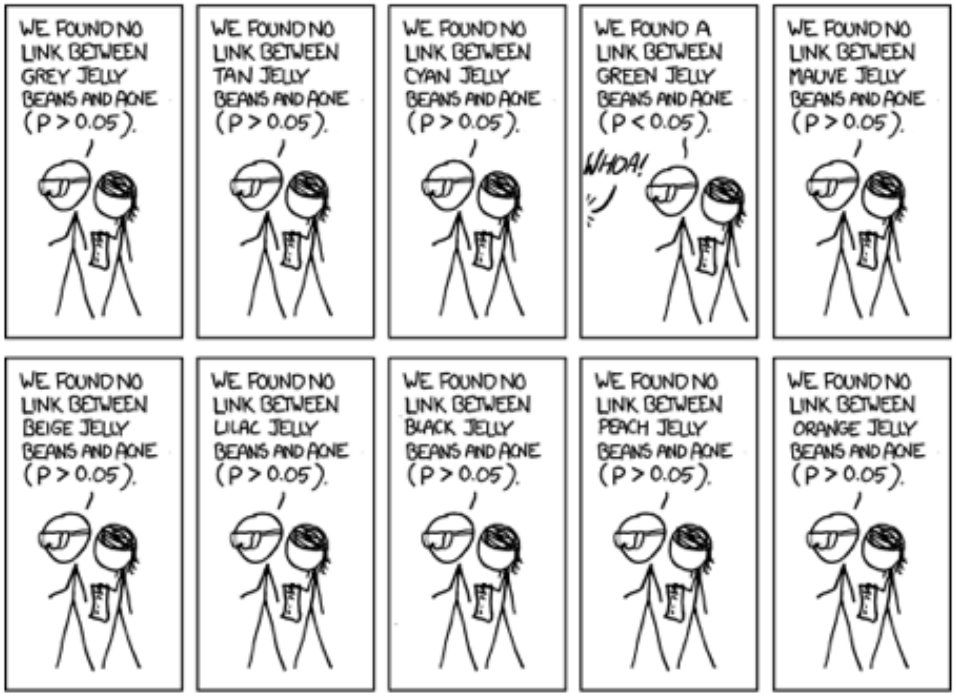
Resumen

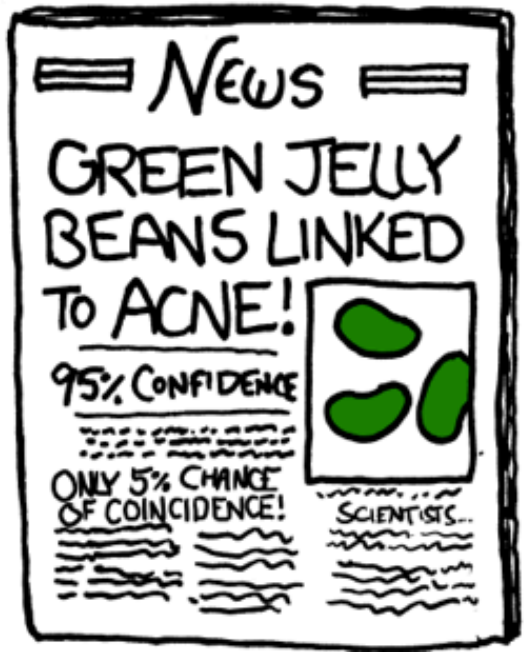
- El uso de modelos distorsiona (simplifica) la realidad, pero nos ayuda a entenderla en parte. Es importante un equilibrio en la complejidad del modelo. Un modelo demasiado simple puede no ser útil, pero uno muy complicado puede estar muy influenciado por datos externos
- Un modelo es siempre una guía, y la adecuación o no a un patrón debe siempre estimarse según su significado estadístico, y corroborarse según evidencias biológicas. Nunca es una prueba irrefutable de algo.
- La significación estadística debe ser rigurosa para minimizar el número de falsos positivos y negativos. Ante la duda, suele ser recomendable ser conservadores en nuestras afirmaciones. Es importante tener en cuenta el número de pruebas (si hay más de una) para realizar correcciones a los estadísticos
- Los modelos ocultos son un tipo de modelos bastante utilizados en bioinformática para determinar el patrón cuando no se pueden hacer asunciones del modelo a priori. Son muy utilizados en alineamientos.

Preguntas a debate

- ¿Crees que el modelado de sistemas es útil? ¿Lo ves como una herramienta complementaria al laboratorio o como algo independiente?
- ¿Qué opináis de la significatividad estadística? ¿Daríais por válido un resultado soportado por la estadística pero no por la biología? ¿Y al revés, soportado por la biología pero no por la estadística?







http://xkcd.com/882/