“On the Entropy of DNA”

Universidad CEU – San Pablo

3º Ingeniería Sup. Telecomunicaciones

Raúl Conde Gago

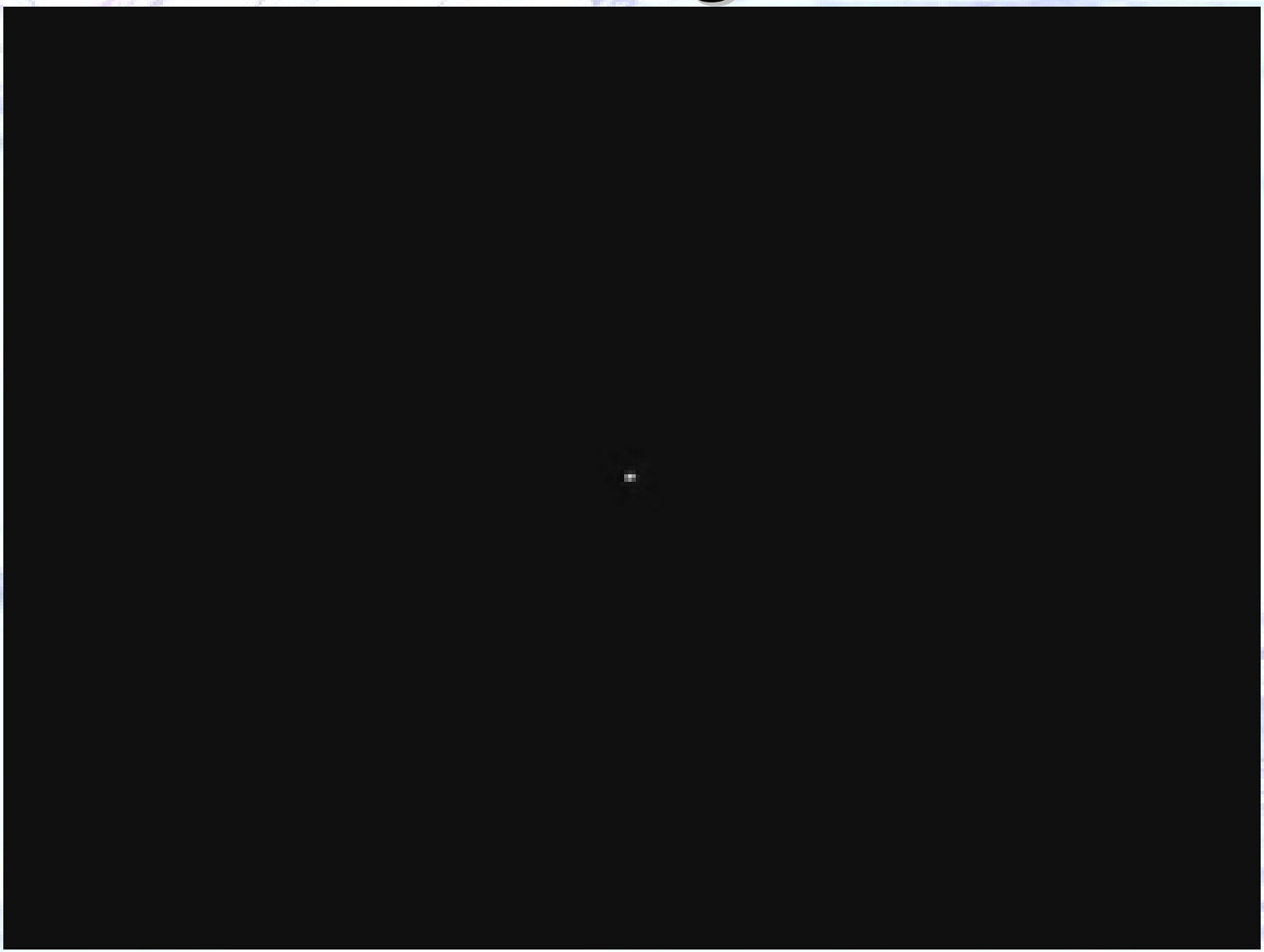
Indice

- 1- Introducción
- 2- Biología
- 3- Entropía
- 4- Métodos de estimación de entropía
- 5- Resultados de la estimación entropía
- 6- Detectando “splice junctions”
- 7- Patrones
- 8- Conclusiones

Introducción

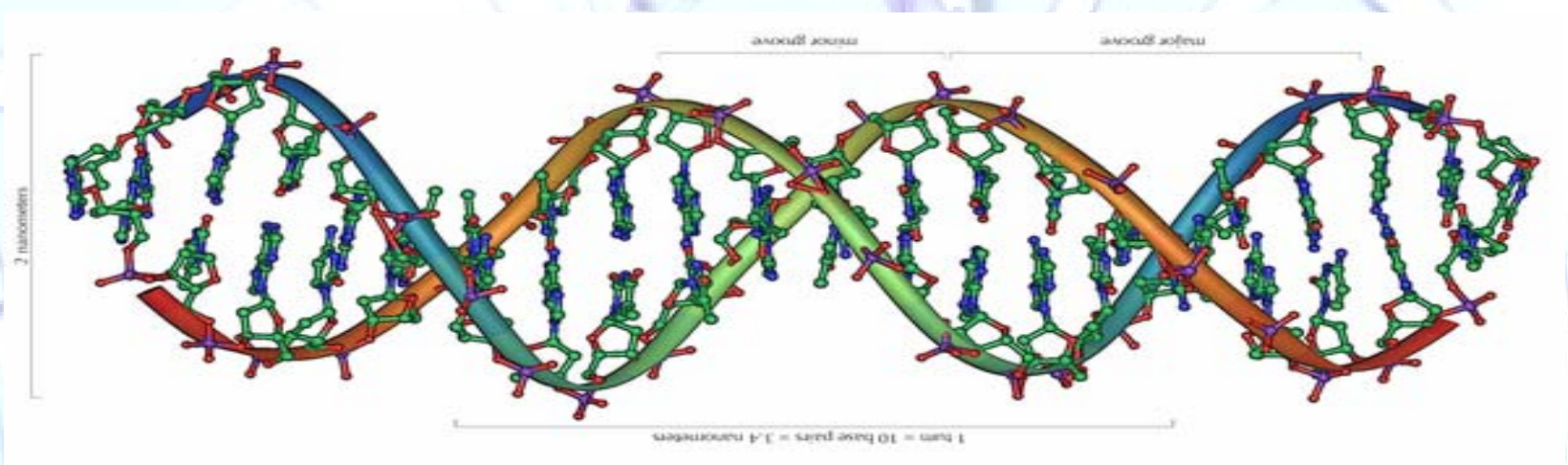
- Estructura DNA es de gran interés
- Investigadores han negado utilidad aplicación teoría de la información.
- A lo largo de esta exposición se intentará demostrar que no es cierto del todo que no sea de utilidad.

Biología



Biología

- **Exon:** cada una de las regiones de un gen que contiene la información para producir la proteína codificada en el gen.
- **Intron:** región del DNA que debe ser eliminada del transcrito primario del RNA

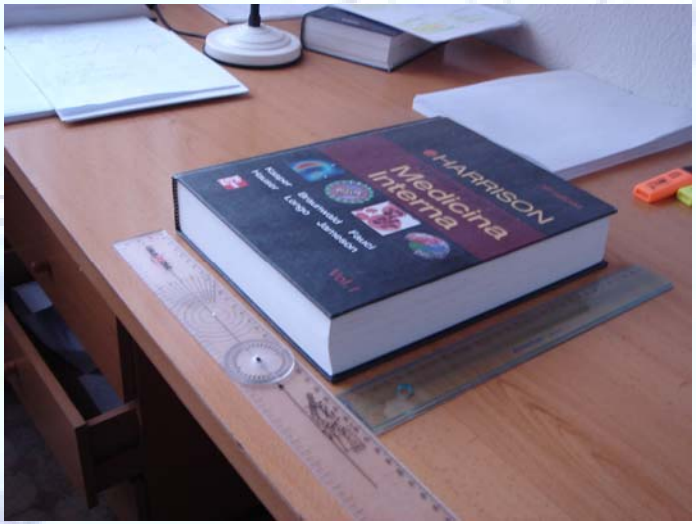


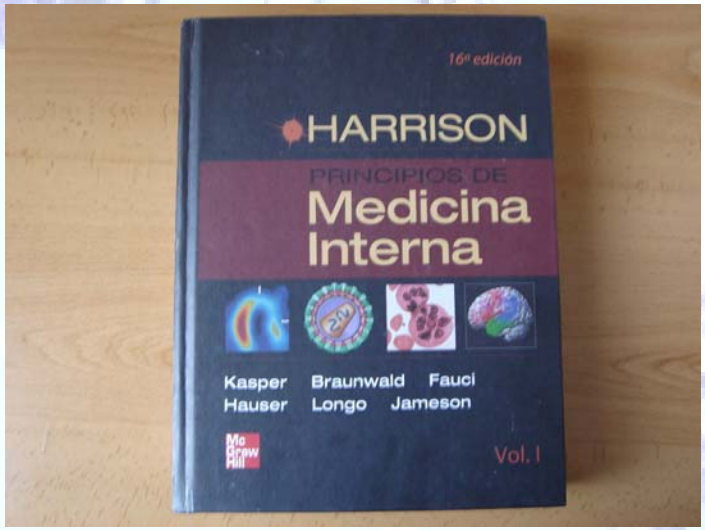
Biología

- Importante para algoritmos: intrones comienzan con GT y terminan con AG.

Biología

- Impresión en papel de secuencias de DNA = 120 volúmenes del libro “Principios de Medicina Interna de Harrison”





Problema…

¿GT pertenece a un intron o un exon?....

Problema…

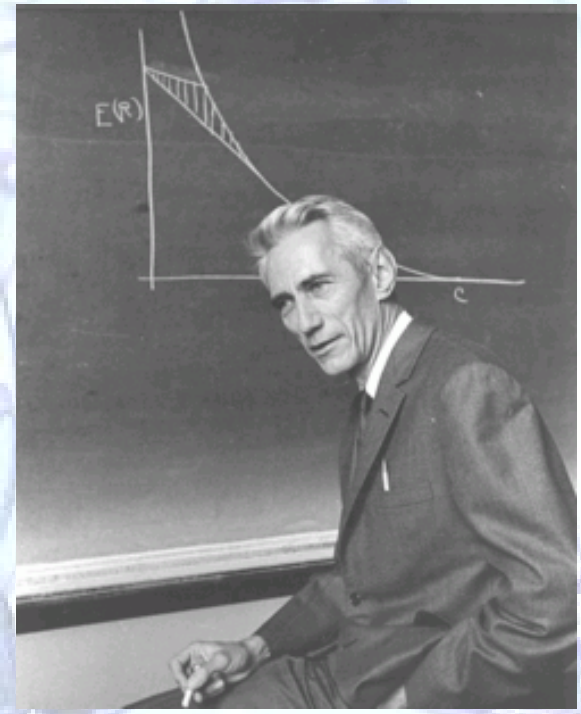
… Aquí es donde interviene la ENTROPÍA, interpretada como una medida de la aleatoriedad.

Entropía

- En Teoría de la Información:
  - Tiene mucho que ver con la incertidumbre que existe en cualquier experimento o señal aleatoria. (“ruido” ó “desorden”)
  - De esta forma podemos hablar de la cantidad de información que lleva una señal.

Entropía

- Entropía de Shannon
- Claude Elwood Shannon, ingeniero electrotécnico y matemático estadounidense.
- Teoría de la comunicación, conocida como Teoría de la información.



Entropía

- Definición según Shannon:
  - La medida de información debe ser proporcional (continua). Es decir, el cambio pequeño de una de las probabilidades de aparición de uno de los elementos de la señal debe cambiar poco la entropía.
  - Si todos los elementos de una señal son equiprobables a la hora de aparecer, entonces, la entropía será máxima.

Entropía

- La entropía nos indica el límite teórico para la compresión.

- Medida de información que contiene el mensaje.

$$H(x) = \sum_{i=1}^{n} p(i) \log\left(\frac{1}{p(i)}\right)$$

Entropía

- ¿Por qué elegimos la entropía? Porque es una medida natural de complejidad, compresión, predicción y aleatoriedad.

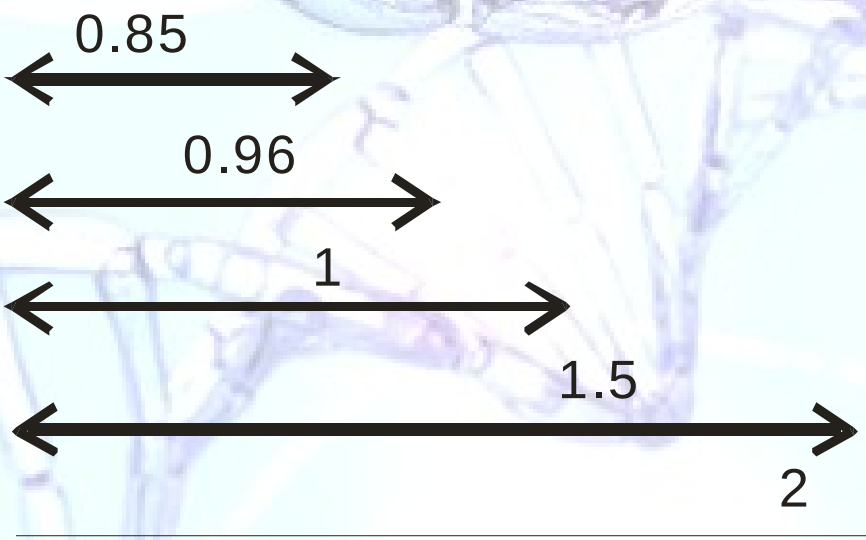
Entropía

- **Clave** para una buena utilización:

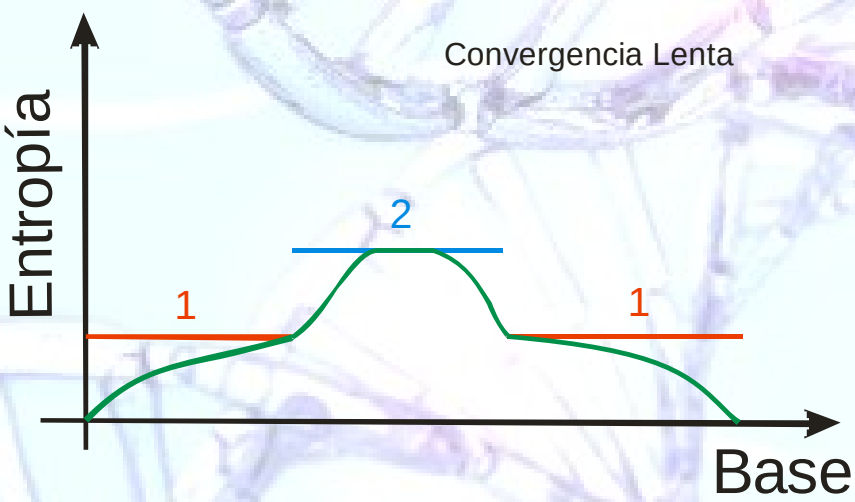
  - Shannon centró su estudio, en un contexto libre, sin prejuicios, por lo que debemos tratar los sucesos de manera independiente. Para ello se basó en secuencias de procesos estocásticos y estudió la entropía de la distribución o distribuciones.

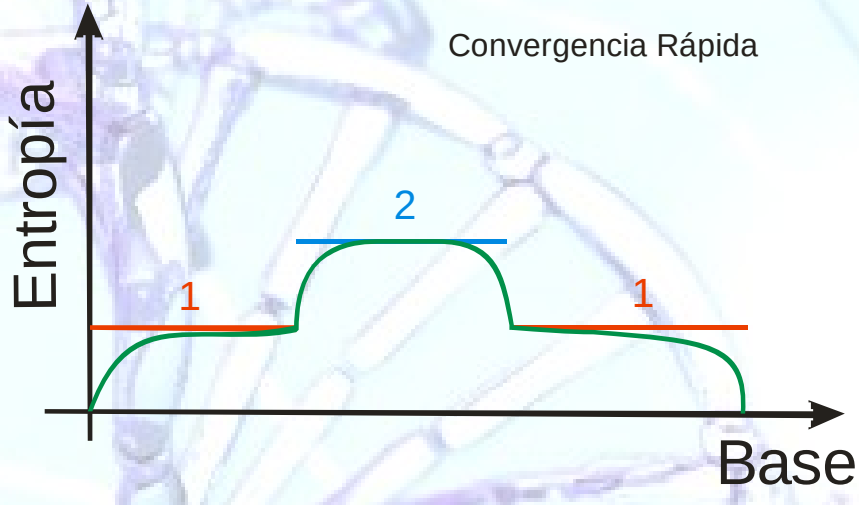
Convergencia

EXON	INTRON	EXON
H=1	**H=2**	**H=1**



Convergencia





Métodos de estimación de entropía

- Diversos métodos. Sólo explicaremos algunos.
- Lo más sencillo serían los programados.
- Recordar que son aproximaciones, para la real necesitamos una cantidad muy grande de datos.

Método de estimación de entropía

- Entropía: límite teórico de compresión … entonces usaremos métodos de compresión para aproximarnos a la entropía.
- Para ello el método debe ser universal. Al aplicarlo, el grado de compresión se acercará al valor de la entropía.
- Método que vamos a utilizar: **Lempel-Ziv**.

Método de estimación de entropía

- Lempel-Ziv:
  - Idea: no repetir “trozos” de texto ó secuencias.
  - Indicar la localización de la primera instancia de ese texto y la longitud.

Método de estimación de entropía

- **Ejemplo 1:**

  - Comienzo con …"*Es muy popular*"…. {1,14}

- **Ejemplo 2:**

  - En la **posición 200**… "*Es muy popular. Lo podemos*…"…. {200,26}

Método de estimación de entropía

- Ejemplo 3:

the o{1,3}r{4,2}n{3,2}is{3,1}{1,5}ld{3,1}{16,1}{1,1}

Método de estimación de entropía

- Con este método, la entropía se aproxima a:

$$\frac{C_n \log C_n}{n} \rightarrow H$$

- C_n es el número de símbolos que mando y n la longitud de la secuencia.

- Método universal. Lo malo es la convergencia lenta debido al gran número de observaciones para la construcción del diccionario.

Método de estimación de entropía

- Ahora veremos la estimación mediante ventana deslizante.
- Primero veremos ejemplo cuando el tamaño de la ventana es variante. (Ejemplo 1)
- Posteriormente veremos el caso donde el tamaño de la ventana es fijo y la ventana no es solapante (Ejemplo 2)

Método de estimación de entropía

- Ejemplo 1:

D_n = {AAGTCATTCAG}

X = TCATTG…

Método muy “cerrado”, poco útil, es decir, sólo aplicable en casos muy puntuales. Convergencia muy lenta.

Método de estimación de entropía

- Ejemplo 2: **N_w: 3**

Datos: {AAGCTAAGCC…}…{0,C}{1,A}{2,C}…

Diccionario: 0) AAG

1) CTA

2) AGC

Método de estimación de entropía

- La entropía estimada mediante el ejemplo de ventana deslizante, con ventana fija no solapante:

Método de estimación de entropía

- Aclaración antes de continuar:

  - 1) La medida de la entropía que estamos utilizando es sólo una aproximación de la medida real.

  - 2) DNA no es estacionario.

  - 3) DNA no es un proceso aleatorio.

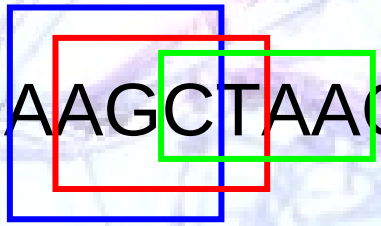
Resultado de estimación entropía

- Aplicando LZ a secuencias de DNA obtengo que la variabilidad de los intrones es mayor que la de los exones.
- Los métodos usados hasta ahora tienen convergencia lenta.
- Los métodos de convergencia rápida son más fidedignos.
- Para obtener convergencia rápida, introducimos una modificación en el método de ventana deslizante, que ahora será solapante.

Resultado de estimación entropía

- Ejemplo: $N_w = 4$

Secuencia: AAGCTAAG…{0,A}{1,G}…



Diccionario: 0) AAGC

1) AGCT

Esta variación en el método de ventana deslizante, nos permite conocer con más exactitud la redundancia, y por consiguiente la entropía. Por supuesto, la convergencia es más rápida.

Resultado de estimación entropía

- Aplicando este método, realizo un experimento, quiero ver si las entropías de los intrones y los exones es igual.
- Hago la suposición de que conozco los límites, por lo que puedo crear secuencias puras de exones e intrones.
- Este proceso nos da la variable aleatoria:

$$L^{type(exon,\mathrm{int}\,ron)}_{i(indice\_gen),\,j(indice\_ocurrencia)}$$

Resultado de estimación entropía

- $N_w = 16$
- Realizo el test dos veces sobre el mismo dato.
- Realizo el test del rango con signo.
- Condición H(intro)=H(exon) -> $E[Y_{i,j}]=0$.
- Realizamos el test, y en 303 comparaciones, nos da un error del 73 %.
- Conclusión: Entropías son distintas y longitudes también son distintas.

Detectando “splice junctions”

- “Splice junction” = punto donde empieza y termina un intron.
- Intron empieza con GT y termina con AG.
- Ahora queremos ver si GT es el comienzo de un intron o un par de bases cualquiera.

Detectado “splice junction”

- Conjunto de datos: 39439 cadenas que empiezan con GT.
- 579 son intrones.
- Uso reglas de la entropía condicional.

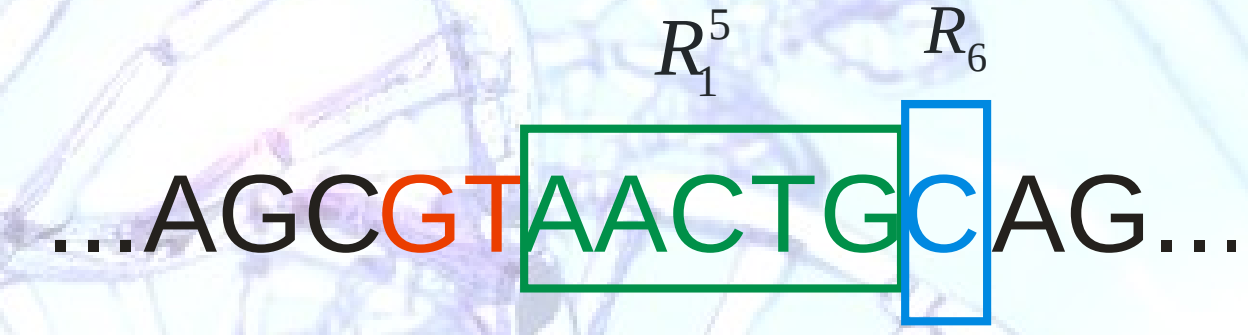
$$H(U \mid v) = -\sum_{u \in U} prob(u \mid v) * \log_2 \, prob(u \mid v)$$

$$H(U \mid V) = E_v[H(U \mid v)]$$

Detectando “splice junction”

1) R_n

2) R_1^{n-1}



1) Letra en la posicíón n de un conjunto dado.

2) Conjunto de letras desde la posición 1 hasta n-1

Detectando “splice junction”

- Mediante teoremas llegamos a:

$$H(R_n \mid R_1^{n-1}) \leq H(R_n) \leq 2$$

- Dos afirmaciones: 1º) n letras que siguen a GT deben ser equiprobables 2º) n letras de GT son independientes a las letras previas.

$$H(R_n \mid R_1^{n-1}) \approx 2$$

Detectando “splice junction”

- Realizamos el experimento.
- Obtenemos que el principio de un intron aparece unido a unos ciertos patrones de la forma {*xxx*GT*xxxx*}.
- Desafortunadamente no nos da la secuencia que indica claramente que es un intron.

Patrones

- Ya tenemos como es la forma del comienzo de una secuencia de intron.
- Ahora queremos obtener la secuencia exacta.
- 7 tuplas cercanas a GT.
- Discriminador con test de “Neyman-Pearson Criterion” y decidir si GT es una pareja que indica el comienzo o no de un intron.

Patrones

- Condiciones de “Neyman-Pearson criterion”:
- Z = siete letras alrededor de GT.
- H_0: GT no marca comienzo de intron.
- H_1: GT marca comienzo de intron.
- $U^{(7)}$: conjunto de 4^7 = 16384 posibles combinaciones de siete letras.

Patrones

- Particionamos U en dos conjuntos, Z_0 y Z_1, siguiendo dos propiedades:

  - 1ª) cada elemento de z perteneciente a U será un elemento exacto de de uno de los dos conjuntos Z_0 y Z_1.
  - 2ª) si z pertenece a Z_0, escogemos la hipótesis H_0, sino, cogemos la hipótesis H_1.

Patrones

- Regla de decisión llevada a cabo por dos criterios:

  - 1º) probabilidad de detección:

$$pD = \sum_{z \in Z_1} pr(z \mid H_1)$$

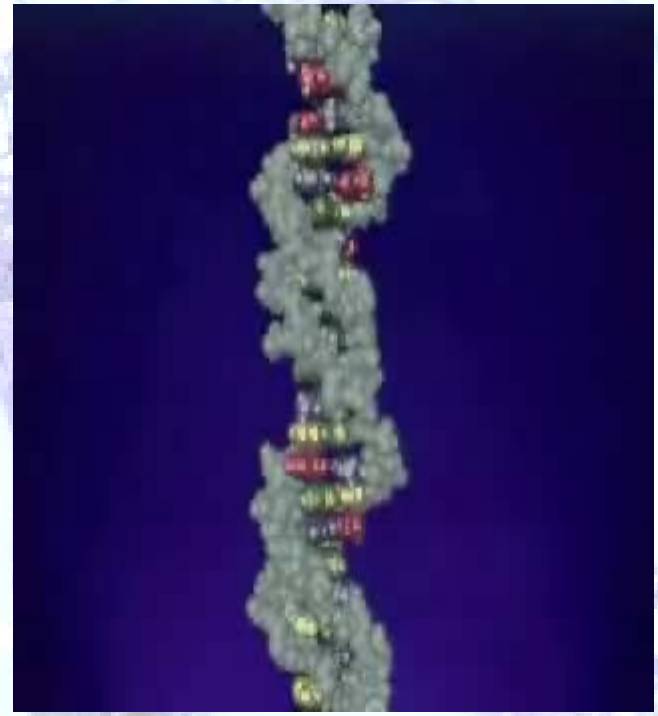
  - 2º) probabilidad de fallar:

$$pF = \sum_{z \in Z_1} pr(z \mid H_0)$$

Patrones

- Llevando a cabo el experimento, llegamos a la conclusión de que hay cuatro patrones que diferencian un par GT comienzo de un intron de un par GT cualquiera:

AAGGTAAGT
AAGGTGAGT
CAGGTAAGT
CAGGTGAGT



Conclusion

1) Las medidas de entropía anteriores han dado resultados no válidos debido a estimadores con convergencia lenta.

2) La entropía de los intrones es mayor que la de los exones.

3) Las mutaciones en un intron no pasa nada a no ser que se produzcan en el punto donde comienza el intron.

4) Las modificaciones en un exon puede tener resultados terribles.

5) Los diseñadores de algoritmos deben confiar en algunas propiedades de la entropía pero deben tener cuidado en no hacer un mal uso de los estimadores.