Usando Modelos de Markov para buscar genes

Anotando un genoma

Una vez que tenemos la secuencia de un genoma, lo siguiente es ver qué es lo que está escrito ahí. A eso se le llama “anotar” el genoma.

Qué se busca?
•Secuencias que codifiquen proteínas
•Secuencias que codifiquen RNAs estructurales

En los eucariotas es mas complejo…

Anotación: Buscar genes

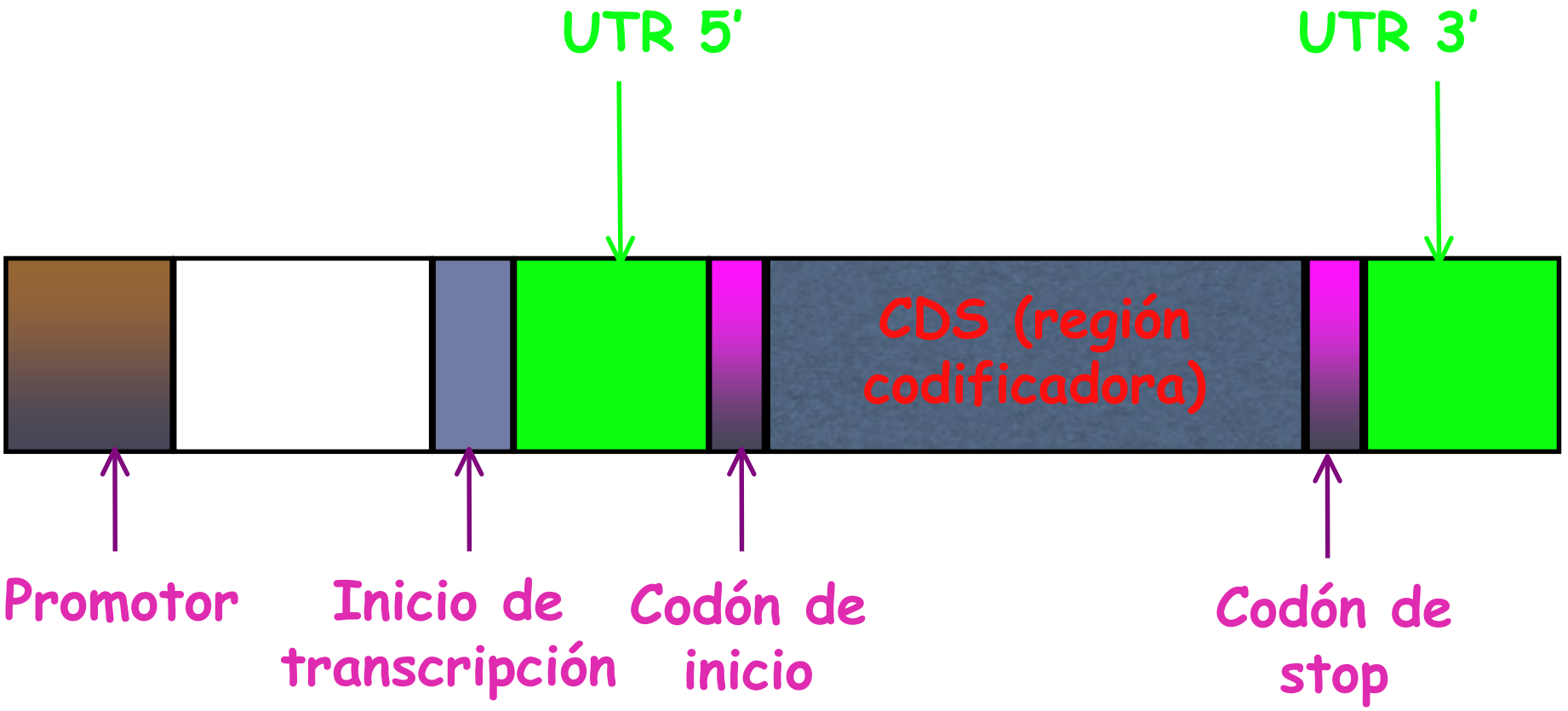
- Primero hablemos de genes que codifican proteínas.

- Recordatorio:
- un tramo de DNA se *transcribe* en un mRNA
- y eso se lleva al ribosoma, donde se traduce a proteína, siguiendo el código genético (y leyendo los nucleótidos de a tres → codones).

- En eucariotas, el mRNA además es editado. En procariotas no; la "anatomía del gen" es más simple. Por eso se usan métodos distintos de anotación, optimizados para cada caso.

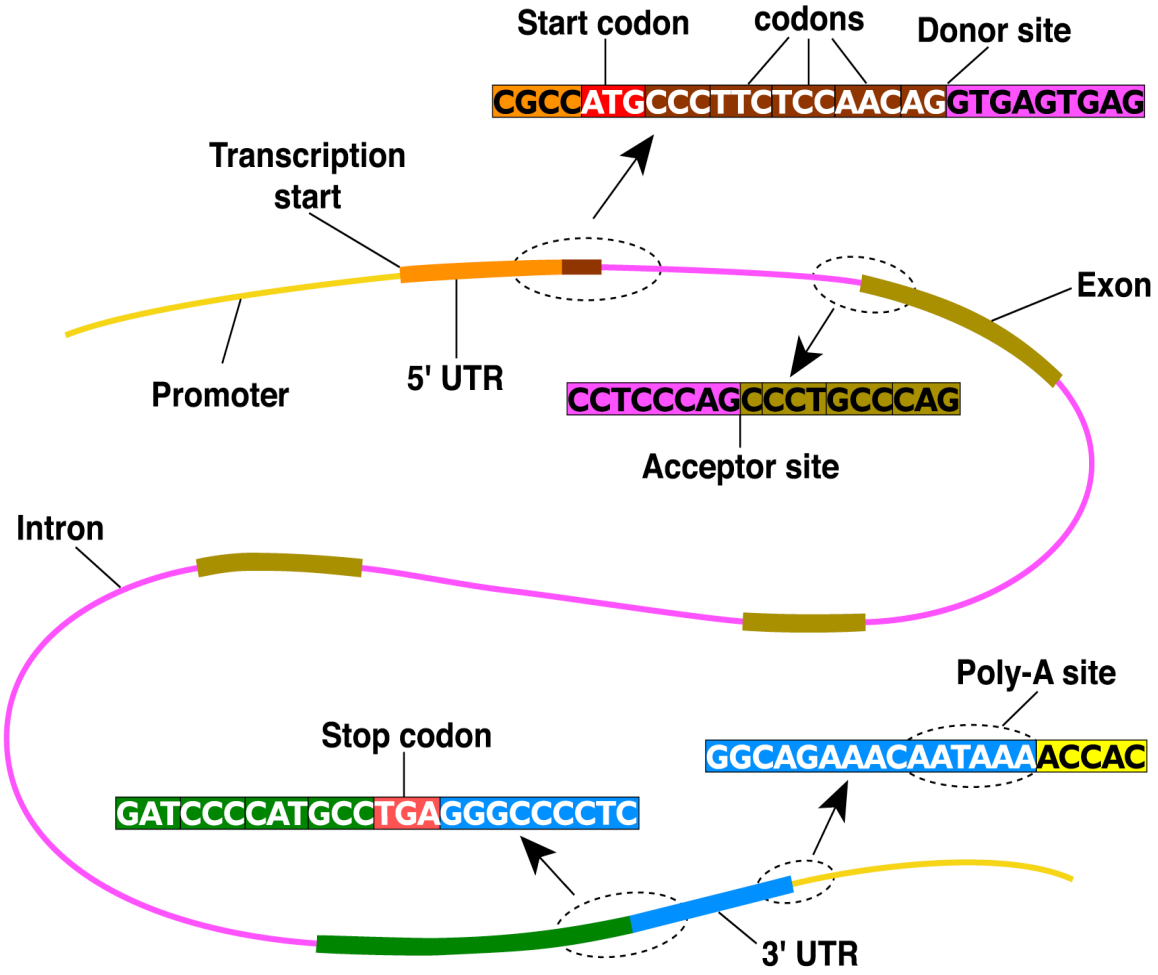
Anotación: Buscar genes

- Gen Procariota

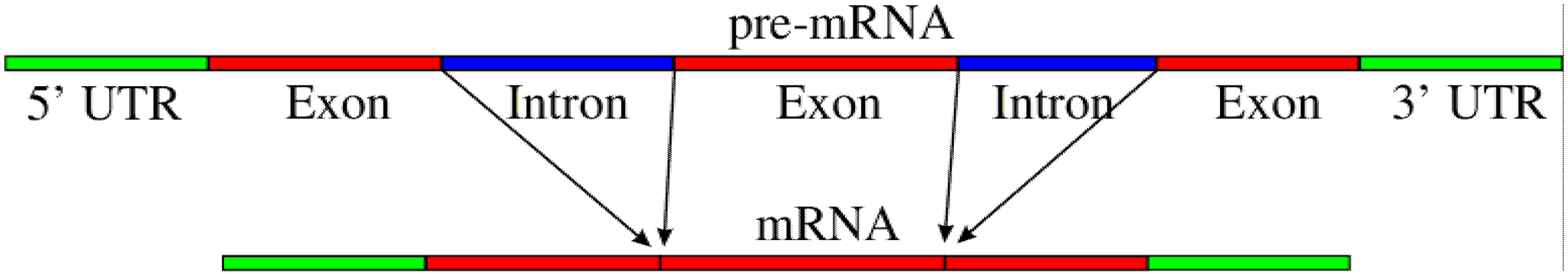


Anotación: Buscar genes

- Gen Eucariota



Anotación: Buscar genes

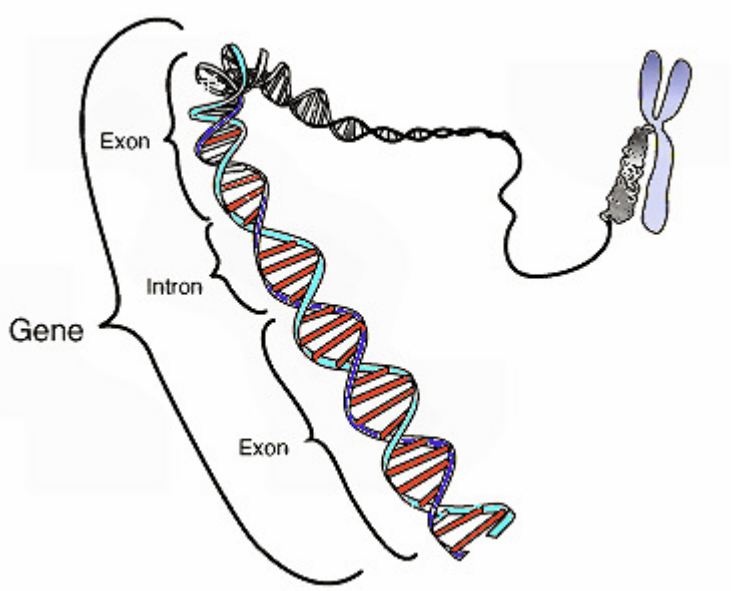


Anotación: Buscar genes

- Siempre se copia un poco más que lo que se traduce; esas son las UTR, “untraslated regions”.

- Un poco antes del punto en que comienza la transcripción, está el “promotor” (promoter), la secuencia donde la polimerasa se liga al DNA para comenzar a copiar.

- Los promotores siguen ciertos “motifs”, a veces dependientes de su función (el tejido o el momento en que el gen tenga que expresarse).
- Es también en esta zona donde intervienen los “factores de transcripción” (proteínas reguladoras).

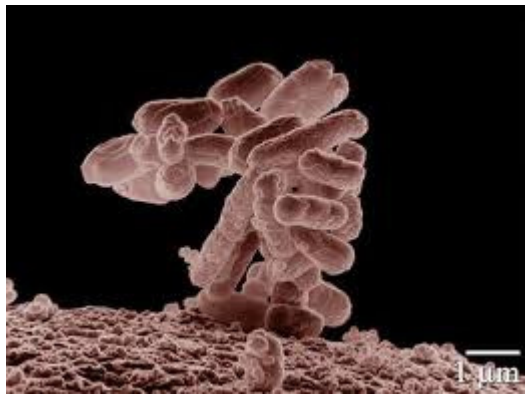


Anotación: Buscar genes

- ORF ("open reading frame"): un tramo largo de DNA, leído en alguno de los tres marcos de lectura posibles, en que no aparece ningún stop. Son candidatos a genes.

- Encontrando un stop, me devuelvo en buscar de un *start* (AUG). Si la longitud es razonable, puedo buscar el motif del promotor, si es que tengo información sobre eso.

- Por ejemplo, en *E. coli* la secuencia TTGACA y TATAAT aparecen 35 y 12 bases antes del inicio de la transcripción, respectivamente (eso, en promedio! Y con variaciones de secuencia!).



Anotación: Buscar genes

Otras cosas que hacen más probable que el ORF sea un gen:

•Homología con genes conocidos.

•Presencia de un periodo 3, detectable con transformada de Fourier. Aparece fuertemente en secuencias codificadoras, como consecuencia de la estructura del código genético.

•Uso de codones (para un aminoácido dado) de acuerdo al “estilo de uso de codones” de la especie [claro que eso requiere tener otros genes ya, como ejemplo para evaluar ese estilo].

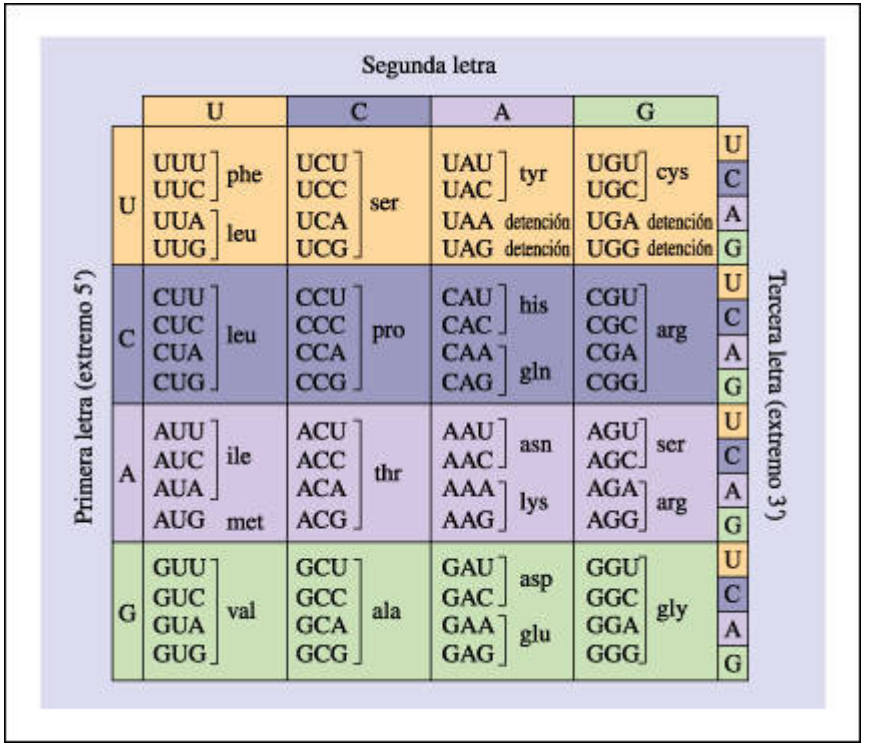
Anotación: Buscar genes

- USO DE CODONES:

- Los codones que codifican un mismo aminoácido debería aparecer, en principio con la misma frecuencia.

- Pero no. Las especies (y clados mayores) tienen estilos consistentes y característicos de codificación; para un aminoácido que admite 6 codones, puede que el 90% de las veces se limiten a dos de ellos.

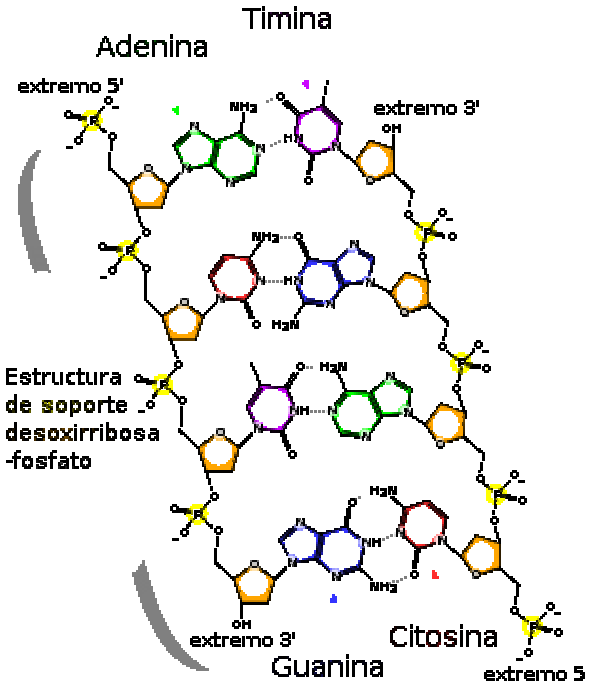
- Hay varias posibles mecanismos, y varios índices para medir estos sesgos.



“Estilos Genómicos”

- Uso de G+C:

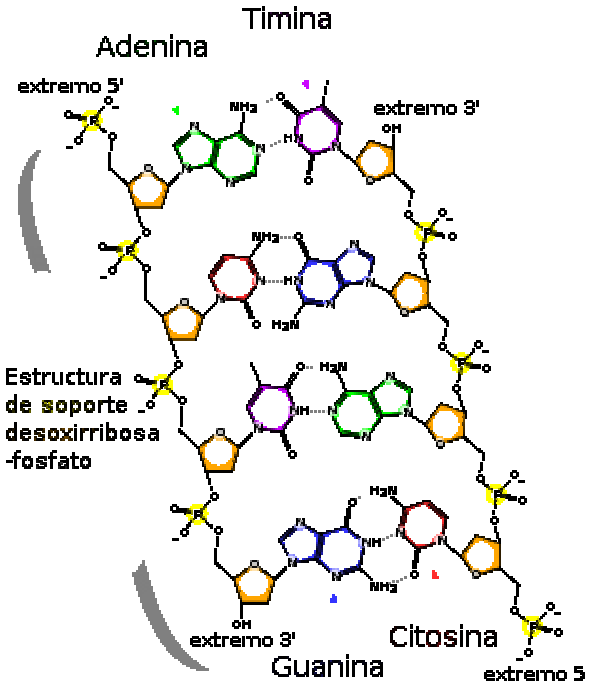
- Representa la cantidad de pares Guanina-Citosina en la molécula de ADN o genoma que está siendo investigado.
- Es una propiedad importante del DNA; determina la estabilidad, y por lo tanto también la temperatura a la que se denatura.
- Cada bacteria tiene un %GC característico; incluso se usa en la nomeclatura de algunos taxones.
- En eucariotas, existen tramos largos con %GC relativamente homogeneo *(isochores)*.
- El GC se puede medir por varios métodos, siendo uno de los más simples la temperatura de desnaturalización de la doble hélice del ADN con un espectrofotómetro.



“Estilos Genómicos”

- Uso de G+C:

- Representa la cantidad de pares Guanina-Citosina en la molécula de ADN o genoma que está siendo investigado.
- Es una propiedad importante del DNA; determina la estabilidad, y por lo tanto también la temperatura a la que se denatura.
- Cada bacteria tiene un %GC característico; incluso se usa en la nomeclatura de algunos taxones.
- En eucariotas, existen tramos largos con %GC relativamente homogeneo *(isochores)*.
- El GC se puede medir por varios métodos, siendo uno de los más simples la temperatura de desnaturalización de la doble hélice del ADN con un espectrofotómetro.



“Estilos Genómicos”

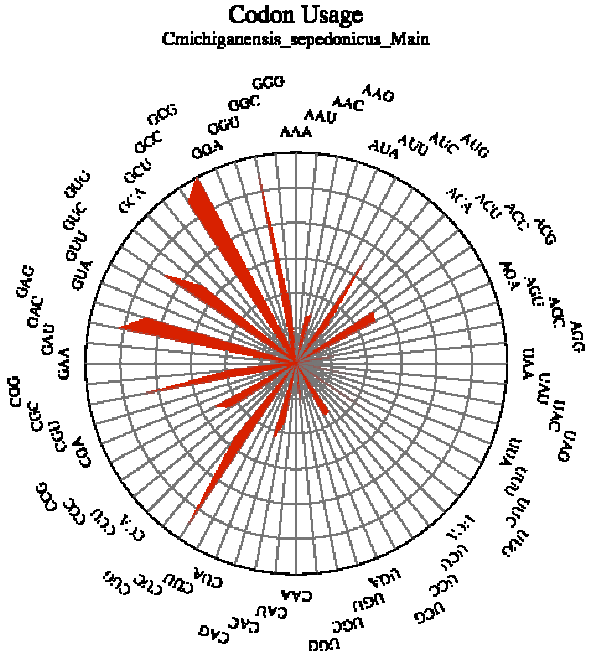
- Uso de G+C:

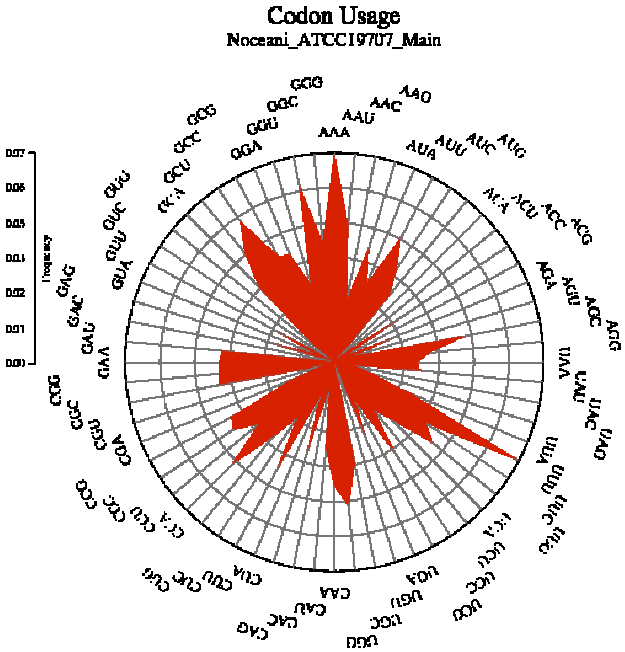
- Los genes suelen estar en regiones de %GC alto.

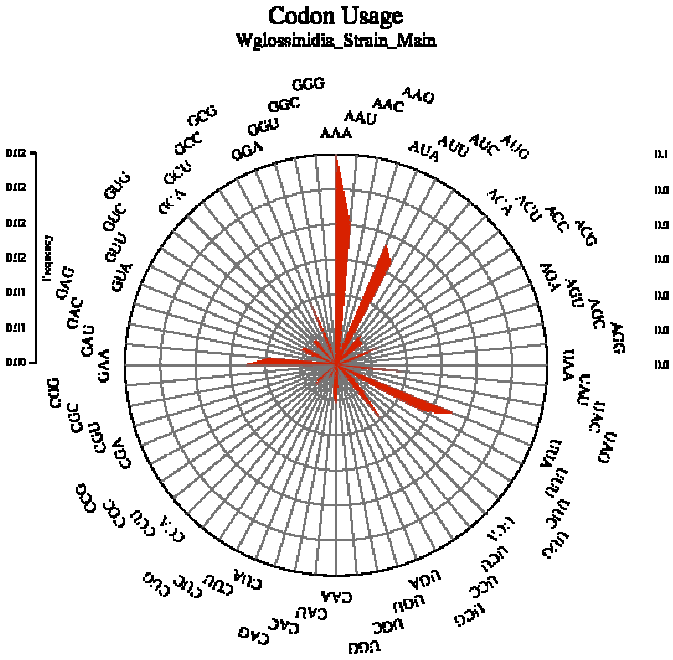
- Cuando hay sesgo hacia GC alto o bajo, ese sesgo es más fuerte en la tercera posición del marco de lectura.

- El uso de GC también ayuda a detectar transferencia horizontales recientes.

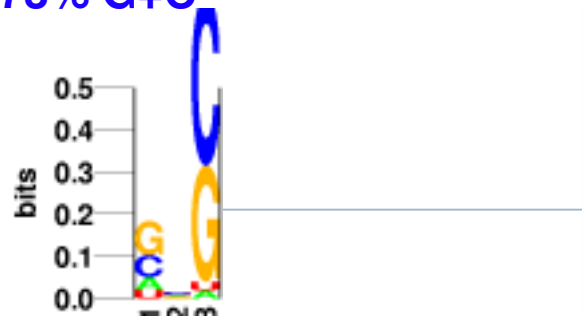
- %GC y uso de codones están relacionados:



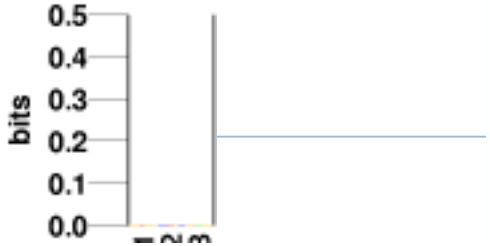




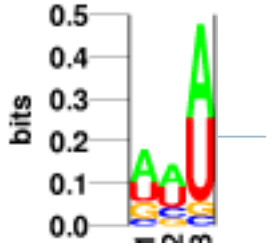
Clavibacter michiganensis
73% G+C



Nitrococcus oceani
50% G+C



Wigglesworthia glossinidia
22% G+C



Determinar Selección

- La existencia de posiciones “sinónimas” y otras que no lo son se usa para evaluar el nivel de selección al que una secuencia ha estado sometida.

- Se alinea con secuencias homólogas.

- Se calcula cuántas de las posiciones sinónimas han mutado (K_s).

- Se calcula cuántas de las posiciones *no* sinónimas han mutado (K_a).

- Se calcula K_a/K_s.
- [Hay variaciones, correcciones, etc, pero la idea es siempre la misma.]

Determinar Selección

- Si $K_a/K_s << 1$, la secuencia ha estado bajo fuerte selección negativa (purificadora): se han eliminado variantes que se alejen de ella.

- Si es $K_a/K_s \sim 1$, es probable que no haya mucha selección (aunque *puede* ser también que haya, pero pocos aminoácidos sean relevantes).

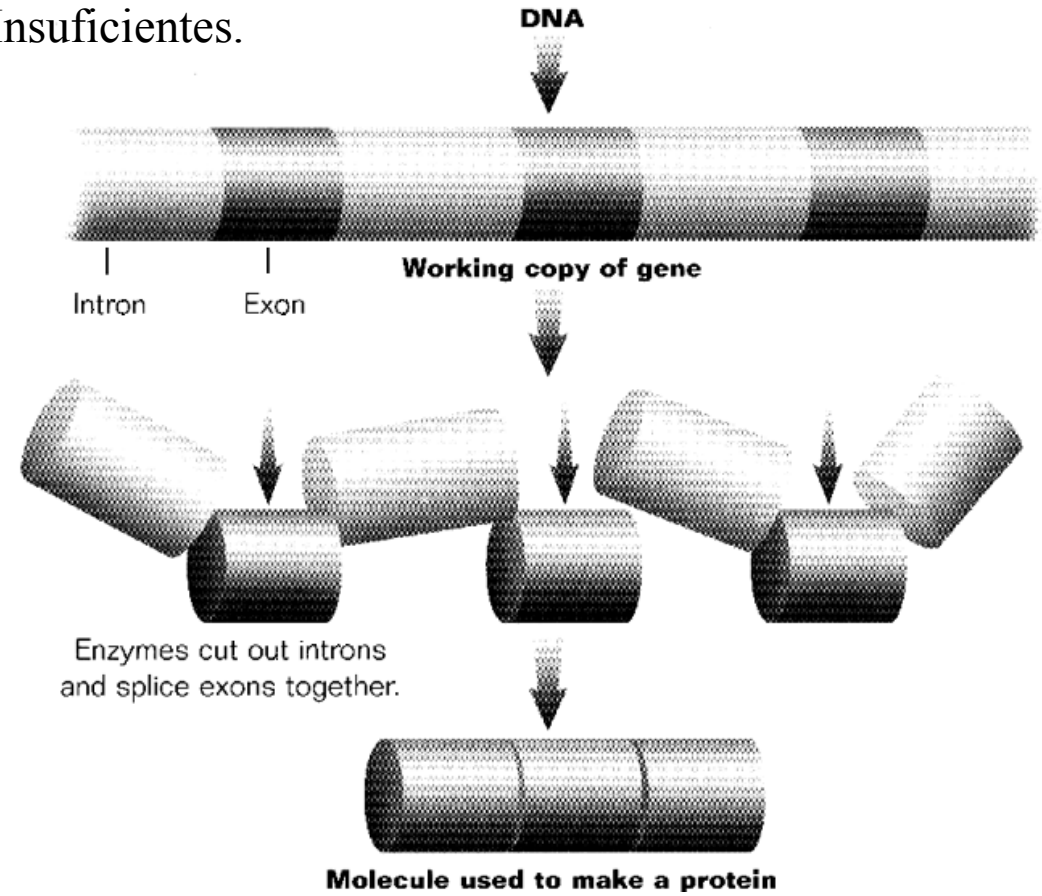
- Si es $K_a/K_s > 1$, ha habido selección positiva (se han "incentivado" los cambios relevantes).

- Más info en: http://selecton.tau.ac.il/overview.html**, donde se puede encontrar sw**

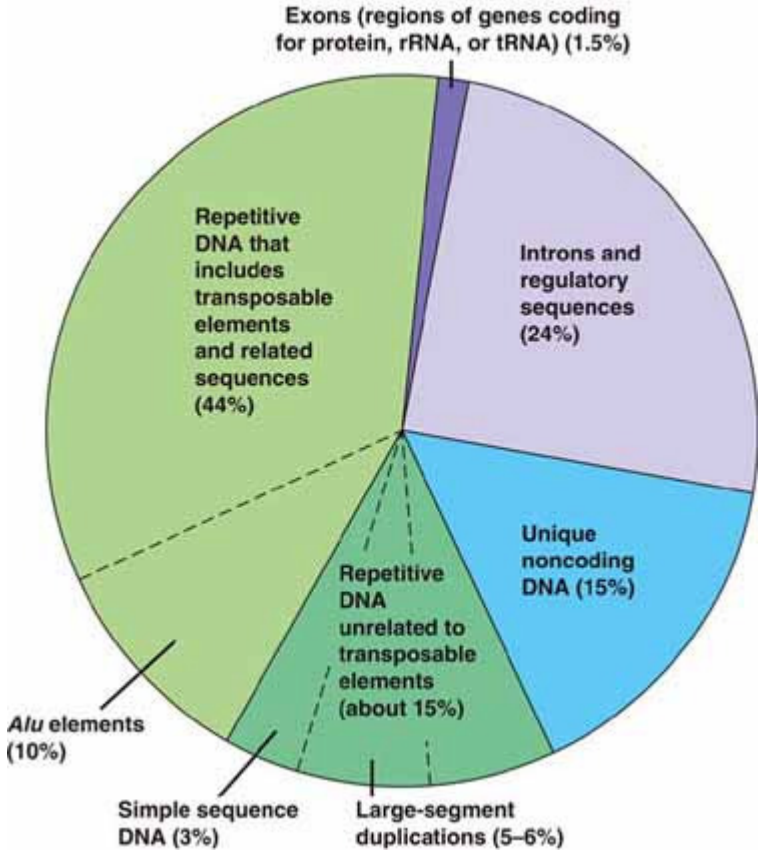
Buscar genes: tarea no trivial

Eucariotas: Aumenta dimensión de su genoma y presencia de Intrones.

Estrategias Insuficientes.



- Incluso a medida que su complejidad aumenta, también lo hace su proporción de DNA que no codifica proteínas.



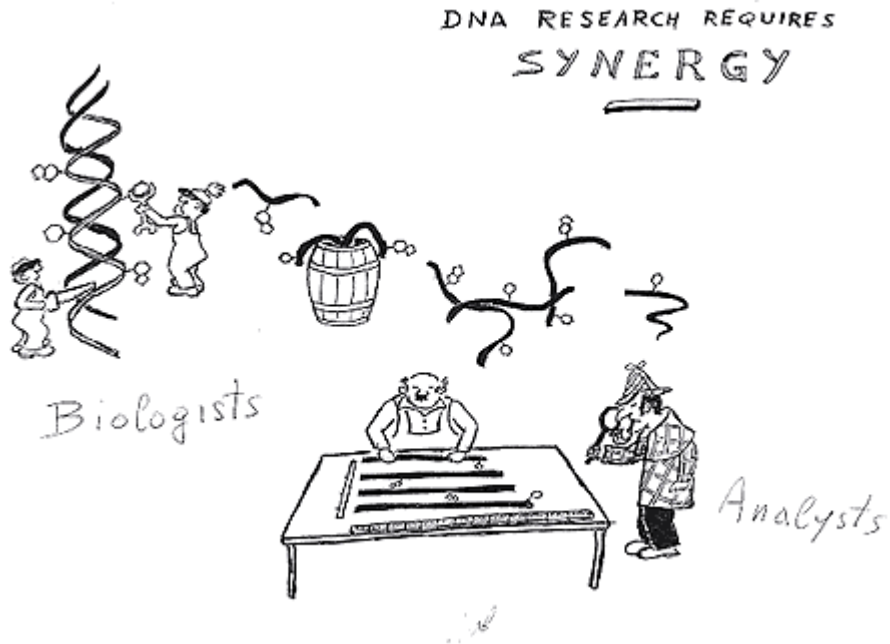
En humanos:
Promedio de 5 a 6 exones por gen.
Alrededor de 8% de genes sin intrones.

Buscar genes: permanente investigación

Idea: ser capaces de reconocer genes, intrones, exones, elementos regulatorios:

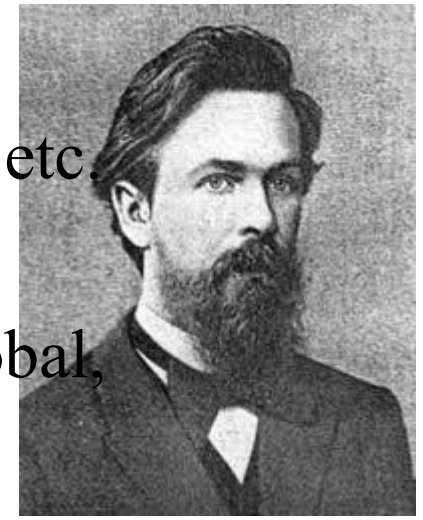
- Qué región codifica para una proteína.
- Qué hebra codifica el gen.
- Dónde comienza y termina el gen.
- Dónde comienza y terminan los intrones/exones.
- Dónde se encuentran las regiones regulatorias del gen.

Esta es un área de permanente investigación.



Buscar genes: aproximaciones

- Métodos “aislados”: Buscar motivos locales que indiquen presencia de algo (promotor, sitio de *splicing* [exón/intrón], etc.).



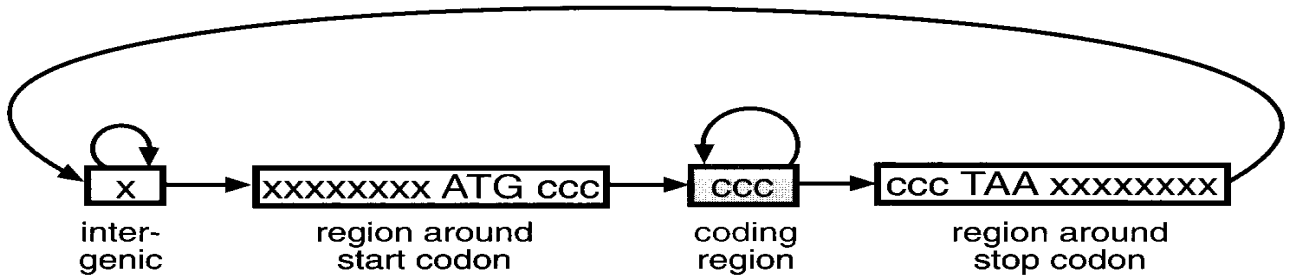
- Redes neuronales, HMM, Gibbs sampling, etc.
- Métodos “integrados”: Mirar estructura global, reconociendo zonas y su encadenamiento.
- GHMM

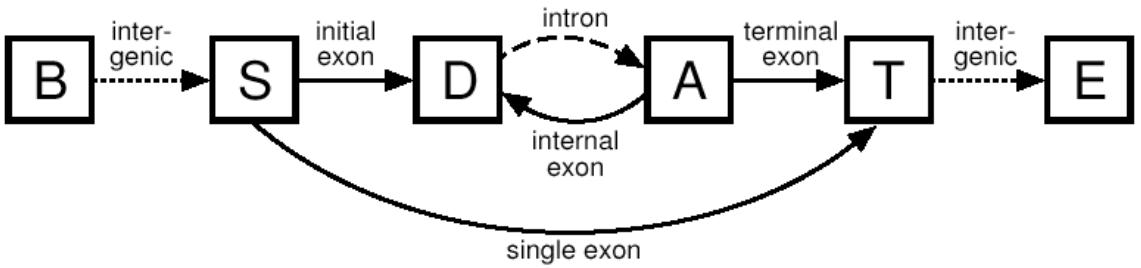
HMM

- Es utilizado para:

•Detectar motifs conocidos (promotor, y algunos otros que se conocen).

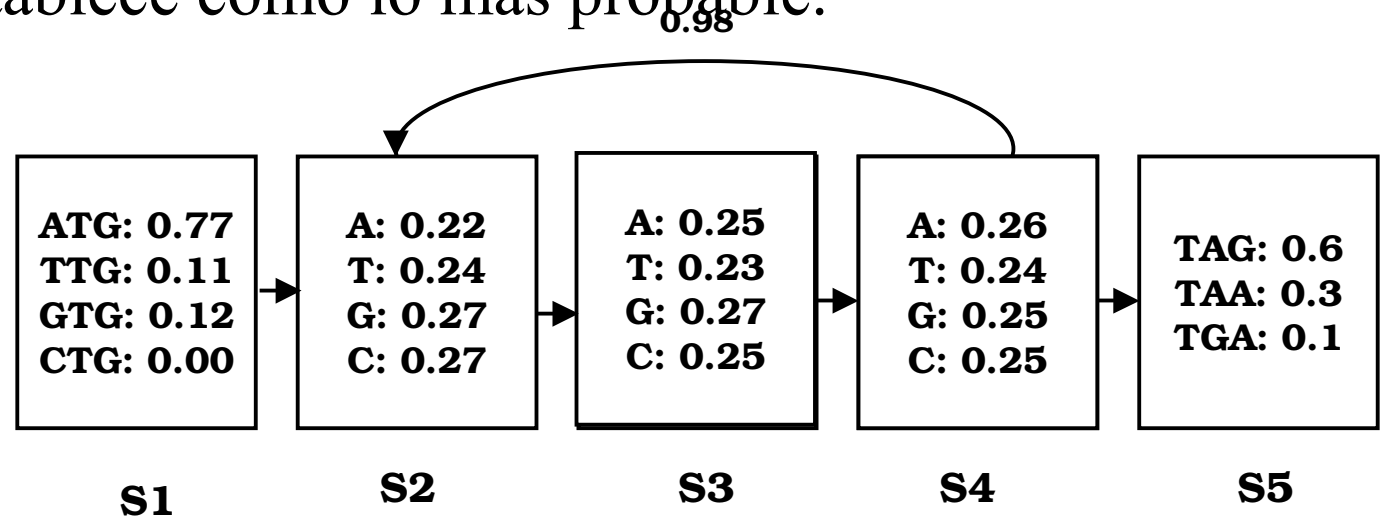
•Para modelar los estados “dentro de un gen” y “fuera de un gen”; al estar dentro de un gen se agrega además un modelo de los codones.





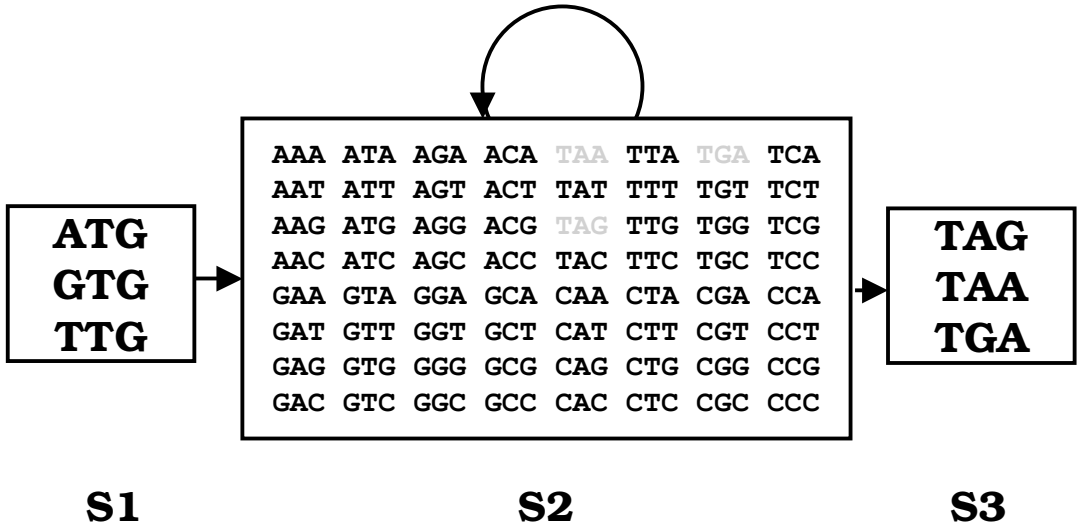
HMM

- Es necesario entrenar el modelo para cada genoma con genes conocidos.
- Luego con el modelo listo, leer secuencias de DNA y encontrar los genes más parecidos, según lo que el modelo establece como lo más probable.



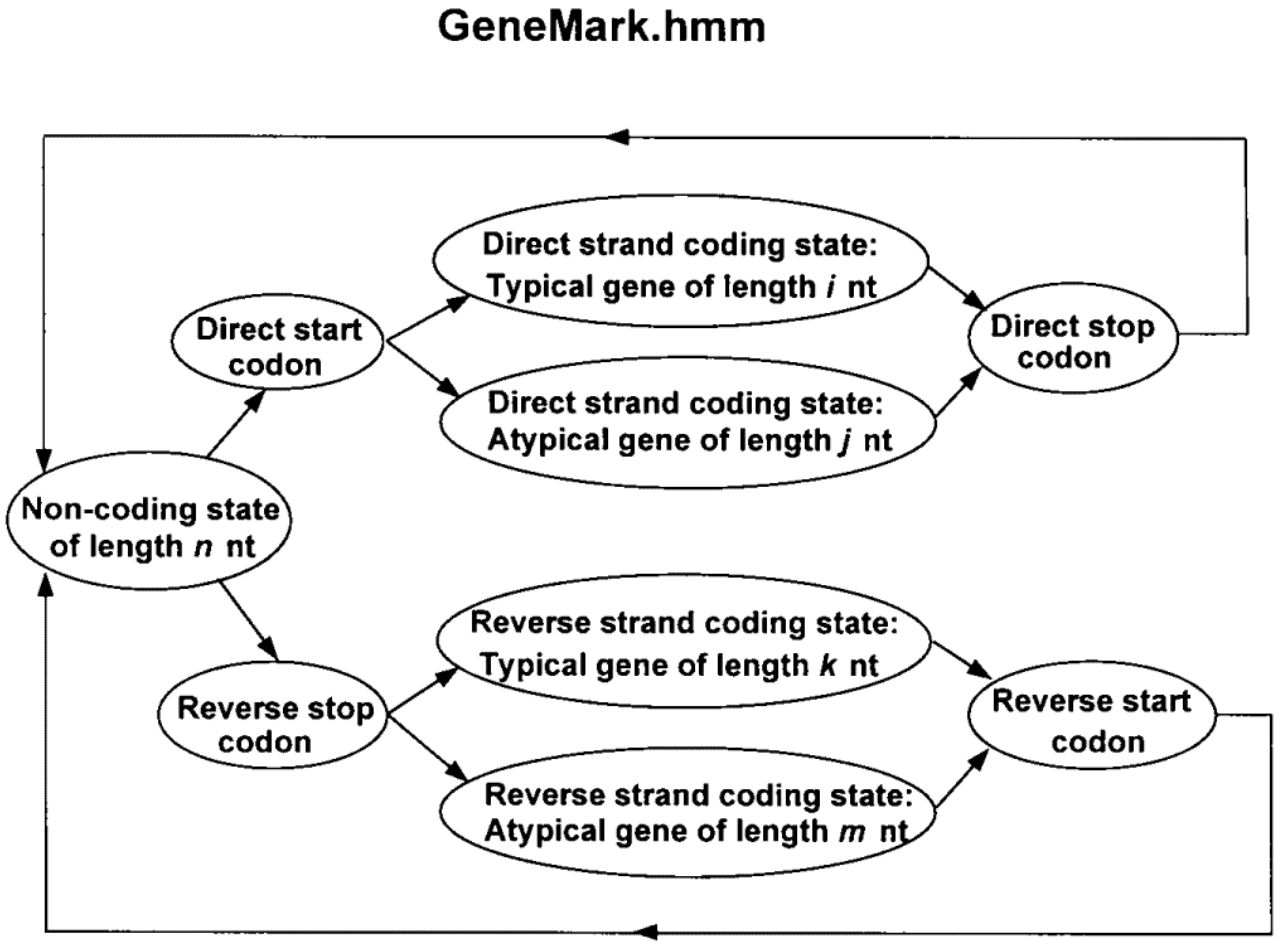
HMM

- Agregando modelo de codones dentro del gen



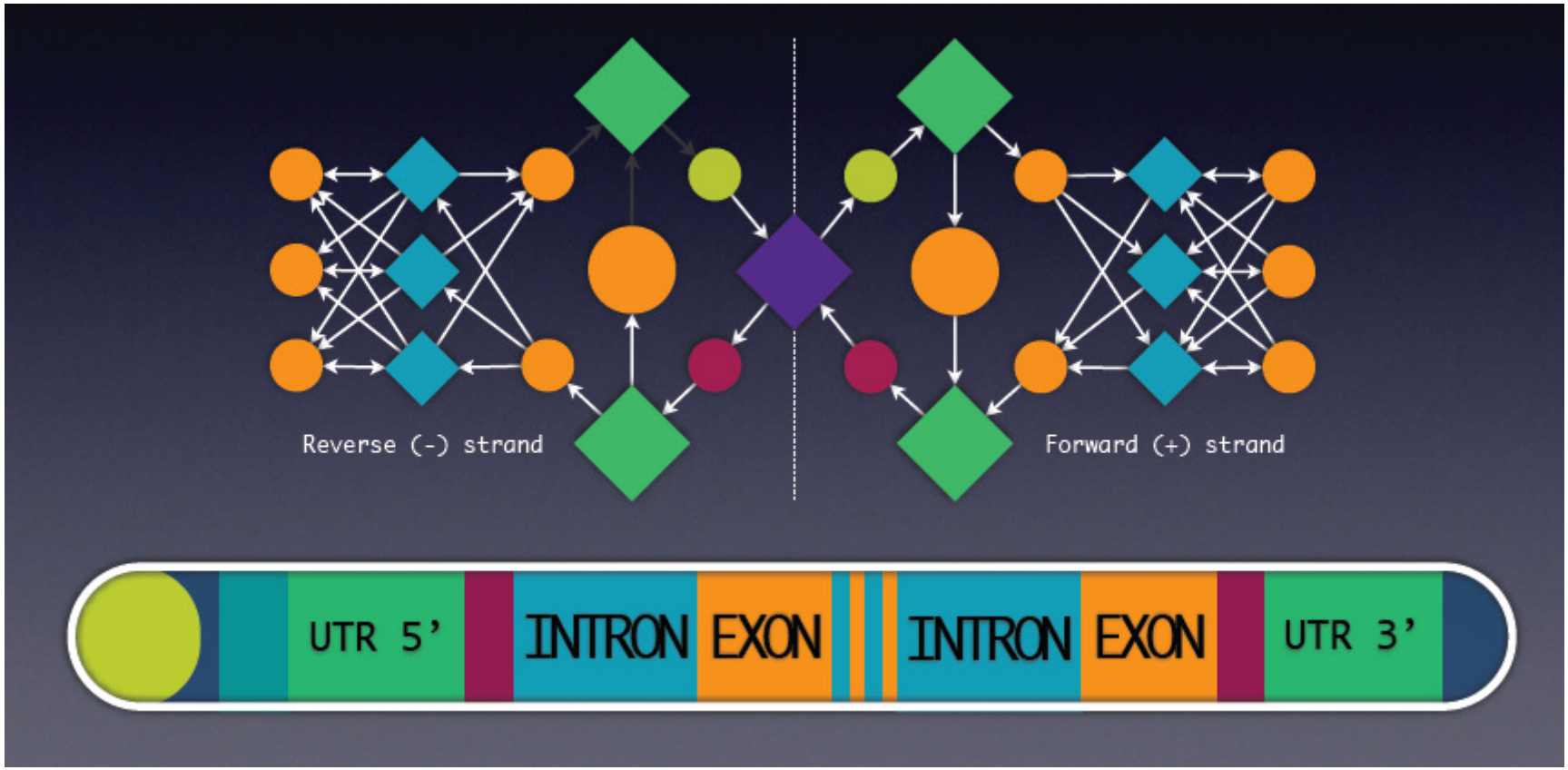
HMM: GenMark

- software más popular para anotar bacterias

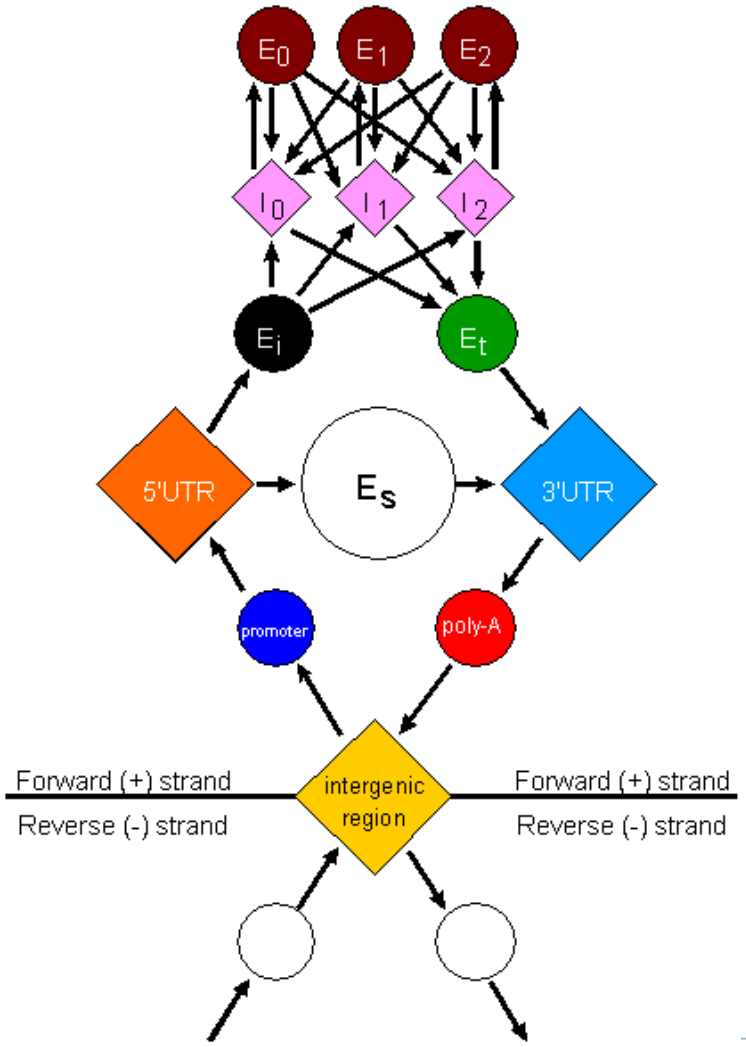


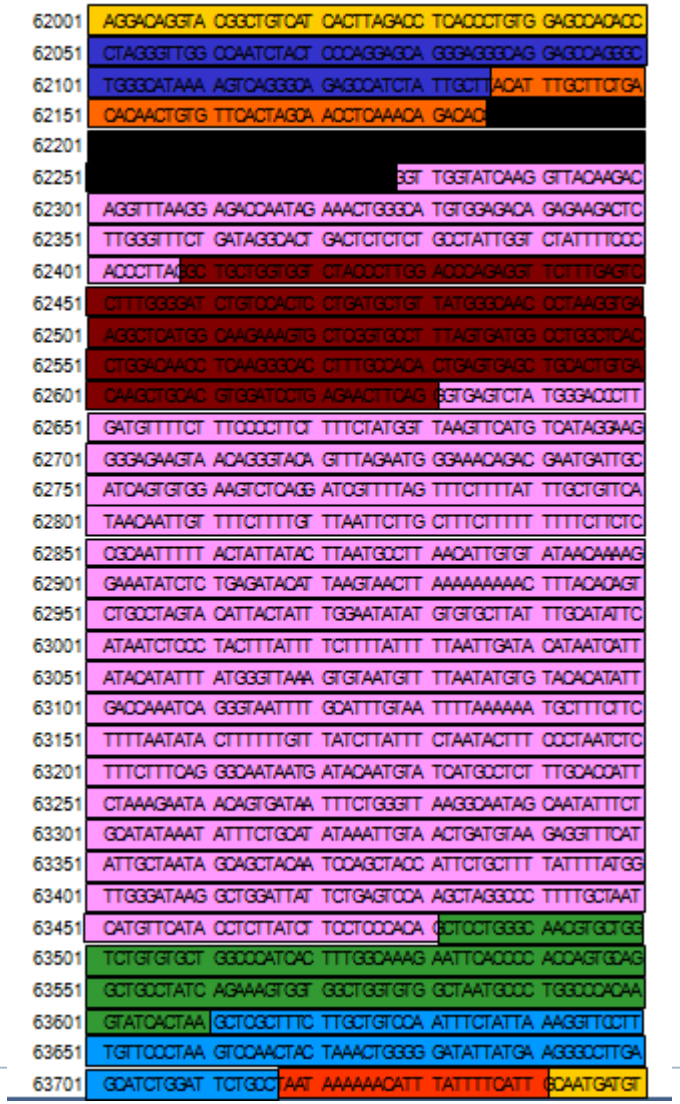
GHMM

- Forma general de describir secuencias.
- Cada nodo corresponde a una región.



GHMM


E0
E1
E2
I0
I1
I2
Ei
Et
5'UTR
Es
3'UTR
promoter
poly-A
Forward (+) strand
Reverse (-) strand
intergenic region
Forward (+) strand
Reverse (-) strand


62001 AGGACAGGTA CGGCTGTCAT CACTTAGACC TCACCCTGTG GAGCCACACC
62051 CTAGGGTTGG CCAATCTACT CCCAGGAGCA GGGAGGGCAG GAGCCAGGGC
62101 TGGGCATAAA AGTCAGGGCA GAGCCATCTA TTGCTTACAT TTGCTTCTGA
62151 CACAACTGTG TTCACTAGCA ACCTCAAACA GACAC
62201
62251 GGT TGGTATCAAG GTTACAAGAC
62301 AGGTTTAAGG AGACCAATAG AAACTGGGCA TGTGGAGACA GAGAAGACTC
62351 TTGGGTTTCT GATAGGCACT GACTCTCTCT GCCTATTGGT CTATTTTCCC
62401 ACCCTTAGGC TGCTGGTGGT CTACCCTTGG ACCCAGAGGT TCTTTGAGTC
62451 CTTTGGGGAT CTGTCCACTC CTGATGCTGT TATGGGCAAC CCTAAGGTGA
62501 AGGCTCATGG CAAGAAAGTG CTCGGTGCCT TTAGTGATGG CCTGGCTCAC
62551 CTGGACAACC TCAAGGGCAC CTTTGCCACA CTGAGTGAGC TGCACTGTGA
62601 CAAGCTGCAC GTGGATCCTG AGAACTTCAG GGTGAGTCTA TGGGACCCTT
62651 GATGTTTTCT TTCCCCTTCT TTTCTATGGT TAAGTTCATG TCATAGGAAG
62701 GGGAGAAGTA ACAGGGTACA GTTTAGAATG GGAAACAGAC GAATGATTGC
62751 ATCAGTGTGG AAGTCTCAGG ATCGTTTTAG TTTCTTTTAT TTGCTGTTCA
62801 TAACAATTGT TTTCTTTTGT TTAATTCTTG CTTTCTTTTT TTTTCTTCTC
62851 CGCAATTTTT ACTATTATAC TTAATGCCTT AACATTGTGT ATAACAAAAG
62901 GAAATATCTC TGAGATACAT TAAGTAACTT AAAAAAAAAC TTTACAGAGT
62951 CTGCCTAGTA CATTACTATT TGGAATATAT GTGTGCTTAT TTGCATATTC
63001 ATAATCTCCC TACTTTATTT TCTTTTATTT TTAATTGATA CATAATCATT
63051 ATACATATTT ATGGGTTAAA GTGTAATGTT TTAATATGTG TACACATATT
63101 GACCAAATCA GGGTAATTTT GCATTTGTAA TTTTAAAAAA TGCTTTCTTC
63151 TTTTAATATA CTTTTTTGTT TATCTTATTT CTAATACTTT CCCTAATCTC
63201 TTTCTTTCAG GGCAATAATG ATACAATGTA TCATGCCTCT TTGCACCATT
63251 CTAAAGAATA ACAGTGATAA TTTCTGGGTT AAGGCAATAG CAATATTTCT
63301 GCATATAAAT ATTTCTGCAT ATAAATTGTA ACTGATGTAA GAGGTTTCAT
63351 ATTGCTAATA GCAGCTACAA TCCAGCTACC ATTCTGCTTT TATTTTATGG
63401 TTGGGATAAG GCTGGATTAT TCTGAGTCCA AGCTAGGCCC TTTTGCTAAT
63451 CATGTTCATA CCTCTTATCT TCCTCCCACA GCTCCTGGGC AACGTGCTGG
63501 TCTGTGTGCT GGCCCATCAC TTTGGCAAAG AATTCACCCC ACCAGTGCAG
63551 GCTGCCTATC AGAAAGTGGT GGCTGGTGTG GCTAATGCCC TGGCCCACAA
63601 GTATCACTAA GCTCGCTTTC TTGCTGTCCA ATTTCTATTA AAGGTTCCTT
63651 TGTTCCCTAA GTCCAACTAC TAAACTGGGG GATATTATGA AGGGCCTTGA
63701 GCATCTGGAT TCTGCCTAAT AAAAAACATT TATTTTCATT GCAATGATGT

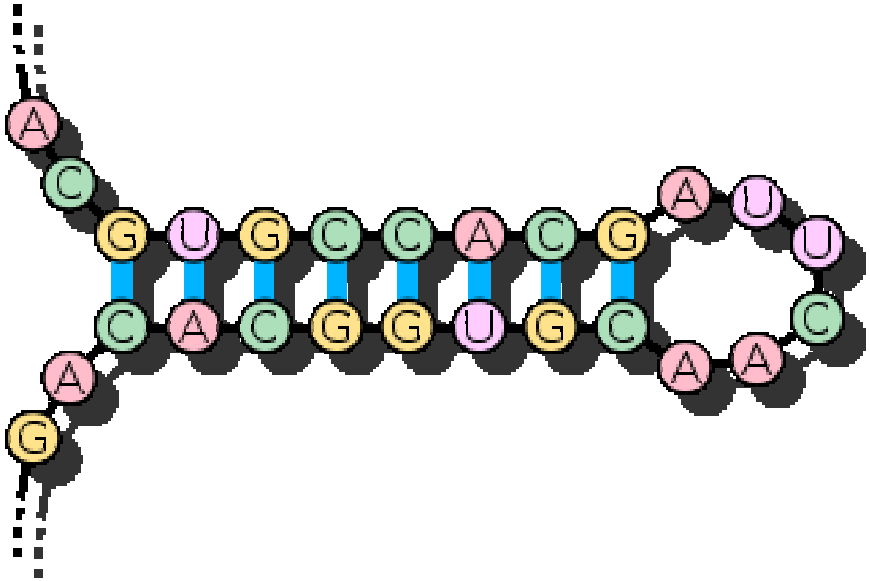
GHMM: GenScan

- GenScan, softwae muy utilizado. (se usó en el Proyecto Genoma Humano).

Usando Gramáticas Formales para anotar secuencias

Anotación de RNA

- El RNA es generalmente una secuencia de una hebra que puede plegarse sobre si misma generando lo que se conoce como estructura secundaria

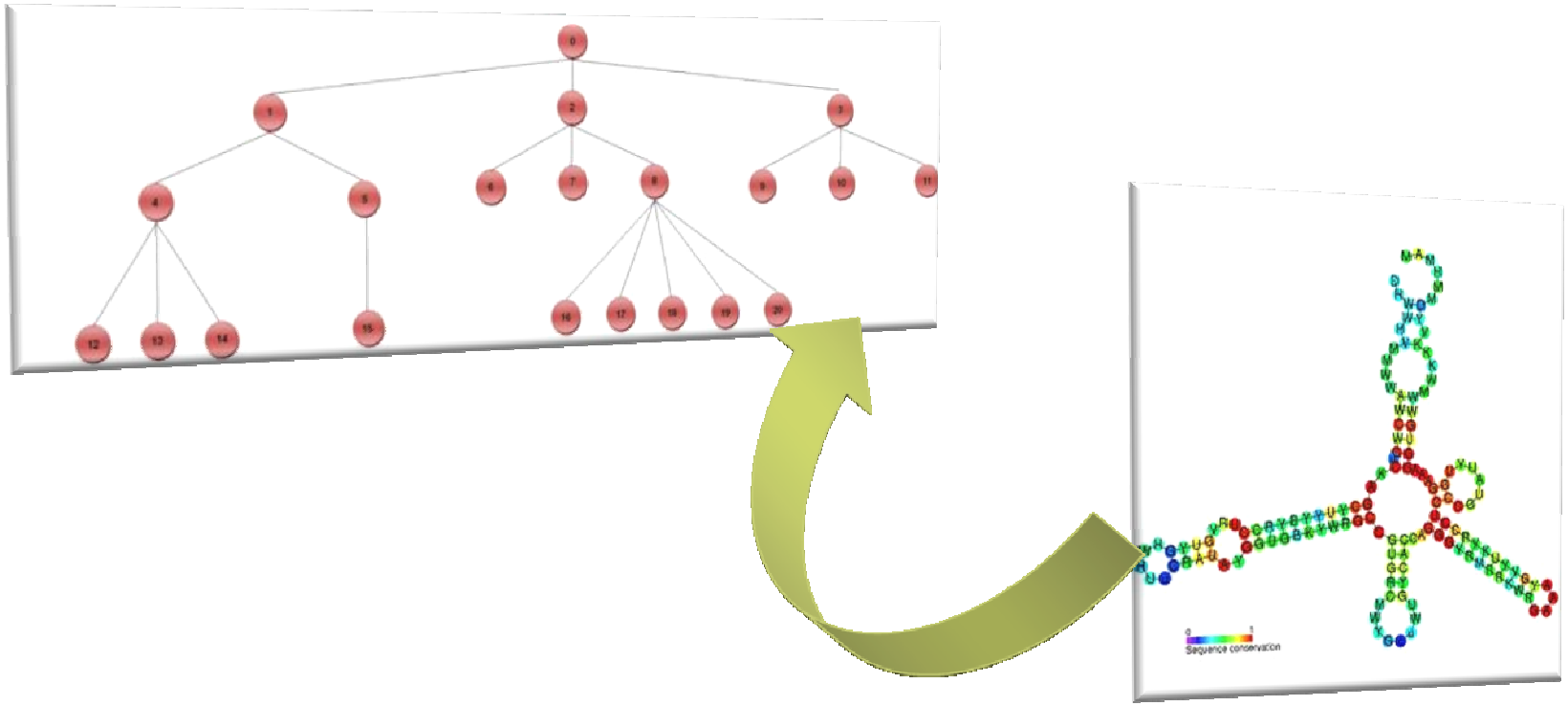


Anotación de RNA

- Debido a su capacidad de pliegue es mucho más complicado modelar la estructura de un RNA que la del DNA
- Se observa que lo que se conserva más entre RNA es la estructura secundaria
- Los HMM no son capaces de modelar esta estructura de forma eficiente
  - Emiten sólo una letra por estado
- Se deben considerar las correlaciones entre pares de residuos

Gramáticas Formales

- La estructura que tiene el RNA puede ser representada mediante una un árbol n-ario, y éste a su vez puede ser representado por una gramática



¿Qué es una gramática?

- Definición tipo TALF
  - Conjunto de reglas de formación que permiten generar cadenas de caracteres a partir de un alfabeto dado. El conjunto de todas las cadenas formadas por este medio se llama lenguaje formal. La gramática define una forma y no un significado
  - Una gramática tiene 4 componentes
    - Alfabeto (símbolos terminales, hojas del árbol…)
    - Producciones (set de reglas)
    - Carácter de inicio S
    - Símbolos no terminales (forman las producciones)

¿Qué es una gramática?

- Definición más humana
  - Una gramática es una maquinita que recibe como entrada letras y genera cadenas con estas letras siguiendo algún tipo de regla

Ejemplo de gramática

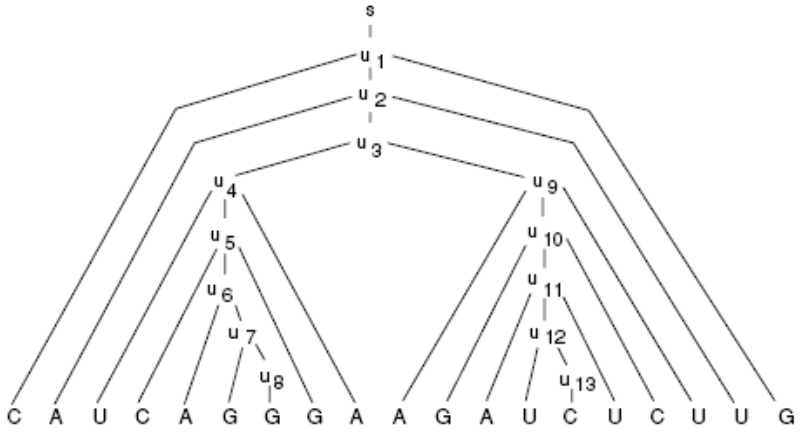
a. Productions

$$
\begin{array}{llll}
P = \{ & s \rightarrow u_1, & u_7 \rightarrow \mathsf{G}\, u_8, & \\
 & u_1 \rightarrow \mathsf{C}\ u_2\ \mathsf{G}, & u_8 \rightarrow \mathsf{G}, & \\
 & u_1 \rightarrow \mathsf{A}\ u_2\ \mathsf{U}, & u_8 \rightarrow \mathsf{U}, & \\
 & u_2 \rightarrow \mathsf{A}\ u_3\ \mathsf{U}, & u_9 \rightarrow \mathsf{A}\ u_{10}\ \mathsf{U}, & \\
 & u_3 \rightarrow u_4\ u_9, & u_{10} \rightarrow \mathsf{C}\ u_{10}\ \mathsf{G}, & \\
 & u_4 \rightarrow \mathsf{U}\ u_5\ \mathsf{A}, & u_{10} \rightarrow \mathsf{G}\ u_{11}\ \mathsf{C}, & \\
 & u_5 \rightarrow \mathsf{C}\ u_6\ \mathsf{G}, & u_{11} \rightarrow \mathsf{A}\ u_{12}\ \mathsf{U}, & \\
 & u_6 \rightarrow \mathsf{A}\ u_7, & u_{12} \rightarrow \mathsf{U}\ u_{13}, & \\
 & u_7 \rightarrow \mathsf{U}\ u_7, & u_{13} \rightarrow \mathsf{C} & \}
\end{array}
$$

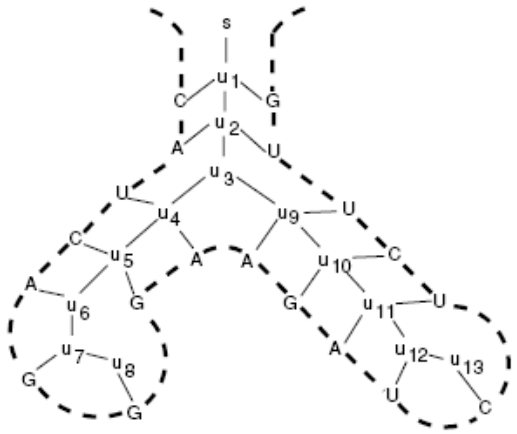
b. Derivation

$$
\begin{aligned}
s &\Rightarrow u_1 \Rightarrow \mathsf{C}u_2\mathsf{G} \Rightarrow \mathsf{CA}u_3\mathsf{UG} \Rightarrow \mathsf{CA}u_4u_9\mathsf{UG} \\
&\Rightarrow \mathsf{CAU}u_5\mathsf{A}u_9\mathsf{UG} \Rightarrow \mathsf{CAUC}u_6\mathsf{GA}u_9\mathsf{UG} \\
&\Rightarrow \mathsf{CAUCA}u_7\mathsf{GA}u_9\mathsf{UG} \Rightarrow \mathsf{CAUCAG}u_8\mathsf{GA}u_9\mathsf{UG} \\
&\Rightarrow \mathsf{CAUCAGGGA}u_9\mathsf{UG} \\
&\Rightarrow \mathsf{CAUCAGGGAA}u_{10}\mathsf{UUG} \\
&\Rightarrow \mathsf{CAUCAGGGAAG}u_{11}\,\mathsf{CUUG} \\
&\Rightarrow \mathsf{CAUCAGGGAAGA}u_{12}\,\mathsf{UCUUG} \\
&\Rightarrow \mathsf{CAUCAGGGAAGAU}u_{13}\,\mathsf{UCUUG} \\
&\Rightarrow \mathsf{CAUCAGGGAAGAUCUCUUG}
\end{aligned}
$$

c. Parse tree

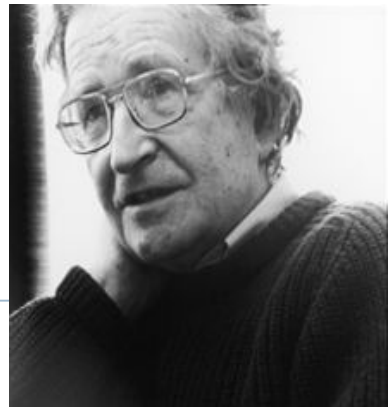


d. Secondary Structure



Tipos de gramáticas

- Existen clasificaciones para las gramáticas según las reglas que utilizan para crear sus producciones. Fueron creadas por Noam Chomsky
  - Gramática tipo 0 (sin restricciones)
    - Generan todos los lenguajes reconocibles por una Máquina de Turing
  - Gramáticas Sensibles al Contexto
    - Cada producción depende del contexto
    - $\alpha A \beta \rightarrow \alpha \mu \beta$
  - Gramáticas de Contexto Libre
    - Producciones simples: $A \rightarrow \alpha$
  - Gramáticas Regulares
    - Para expresiones regulares



Selección de una gramática

- Hemos visto que el RNA puede generar nudos debido a su autoplegado, pero las gramáticas de contexto libre, regulares, sensibles al contexto y de tipo 0 no pueden representar esta situación
  - Debemos utilizar gramáticas de contexto libre con probabilidades para las reglas de producción!!!
    - Stochastic Context Free Grammar
  - Esta solución extiende la funcionalidad de los HMM

SCFG

- A cada producción se le agrega una probabilidad, y la probabilidad de una derivación es el producto de las probabilidades de cada una de las producciones que la componen
- La gramática debe ser entrenada para determinar las probabilidades
  - Algoritmo Esperanza Maximización
    - Encontrar estimadores de máxima verosimilitud de parámetros en modelos probabilísticos
  - Gradient Descent
    - Busca mínimos locales en una función al avanzar en dirección opuesta al gradiente
  - Viterbi

Equivalencia HMM - SCFG

	HMM	SCFG
Alineamiento	*Viterbi*	*CYK*
Puntajes	*Forward*	*Inside*
Entrenamiento	*Baum-Welch*	*EM*

- CYK (Cocke-Younger-Kasami) determina si una cadena puede ser generada por una CFG y si es posible, de que forma
- Inside/Outside permite re estimar probabilidades en una SCFG y es una generalización del algoritmo Forward/Backward de los HMM

Tarea

- Investigue como funciona el algoritmo CYK y de que forma puede ser extendido para las gramáticas de contexto libre probabilísticas. Debe entregar un reporte de una página con sus conclusiones y opcionalmente una página extra para anexos (imágenes, gráficos, etc…)
  - Links de Ayuda
    - http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/papers/ijcnlp04.pdf
    - link2