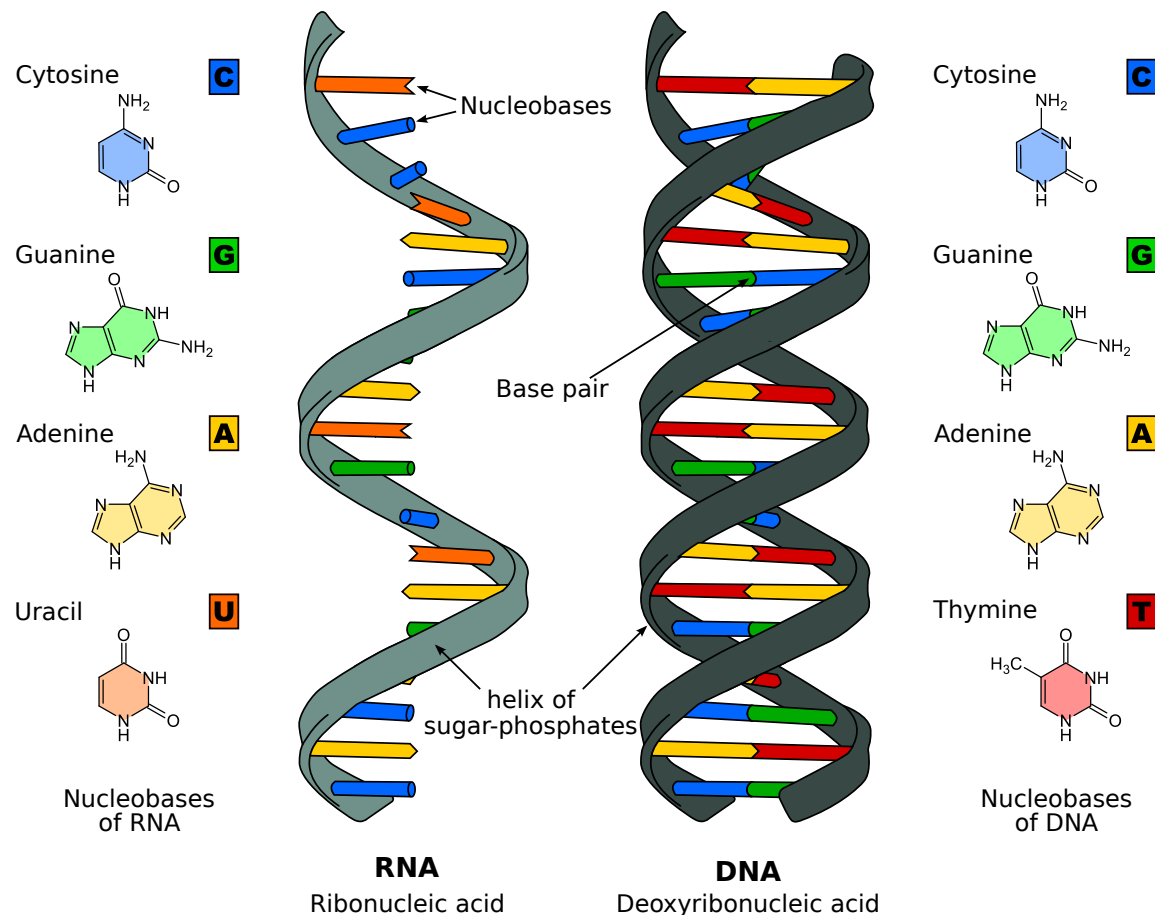


Anotación de genomas y HMMs

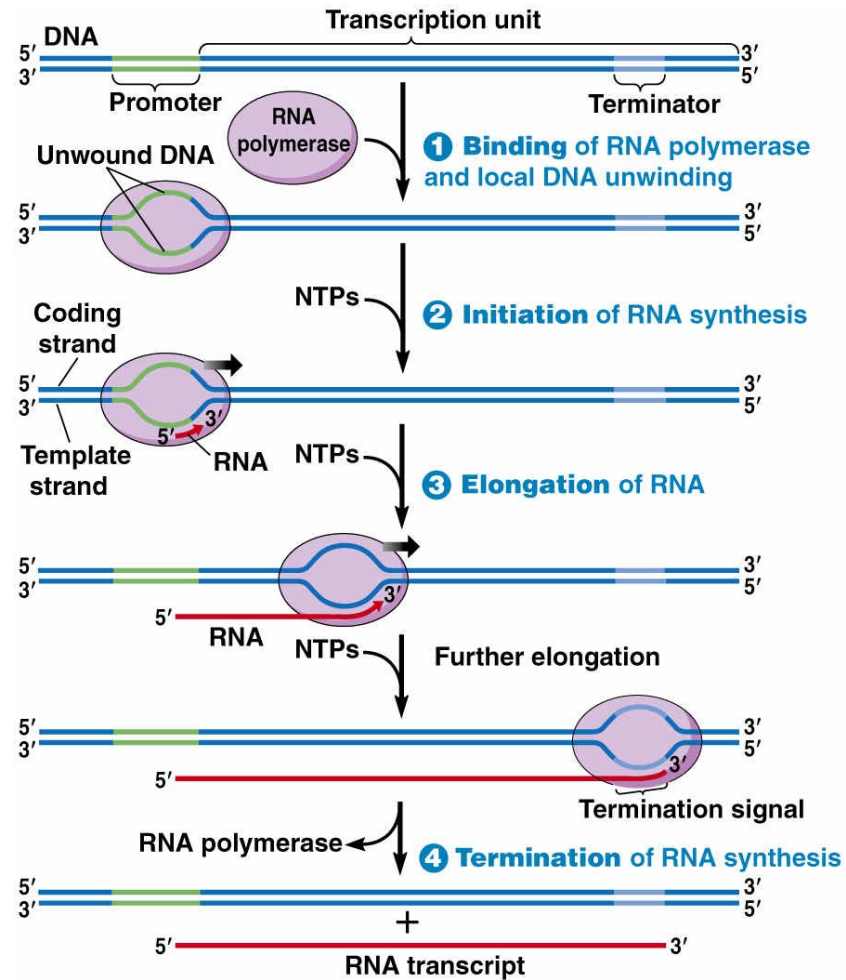
BCOM4104: Estadística en biología computacional
Jorge Duitama

Acido ribonucleico (RNA)

- Moléculas que guían el funcionamiento celular

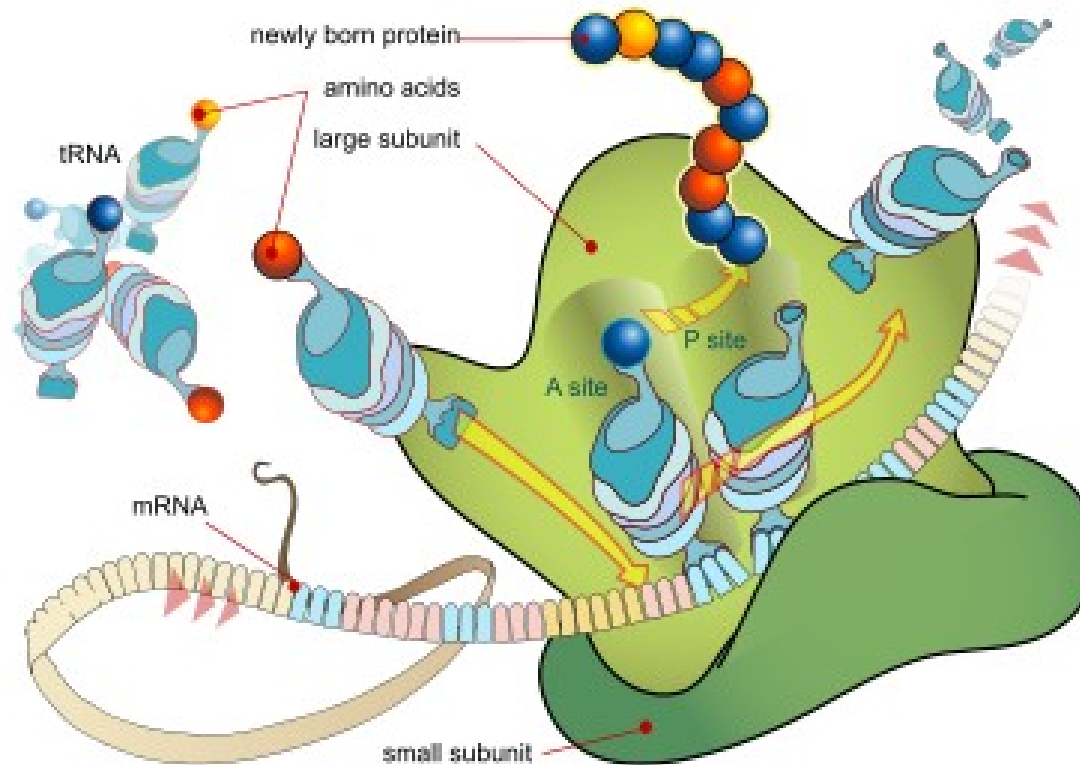


Transcripción

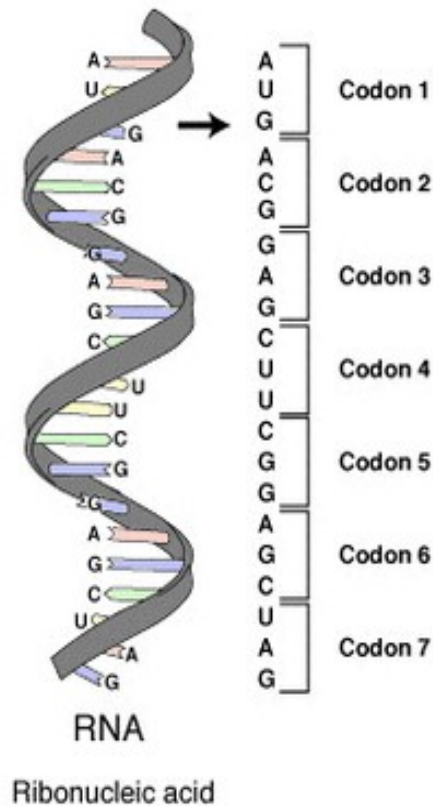


Traducción

- Síntesis de proteínas a partir de ARN mensajero



Código genético



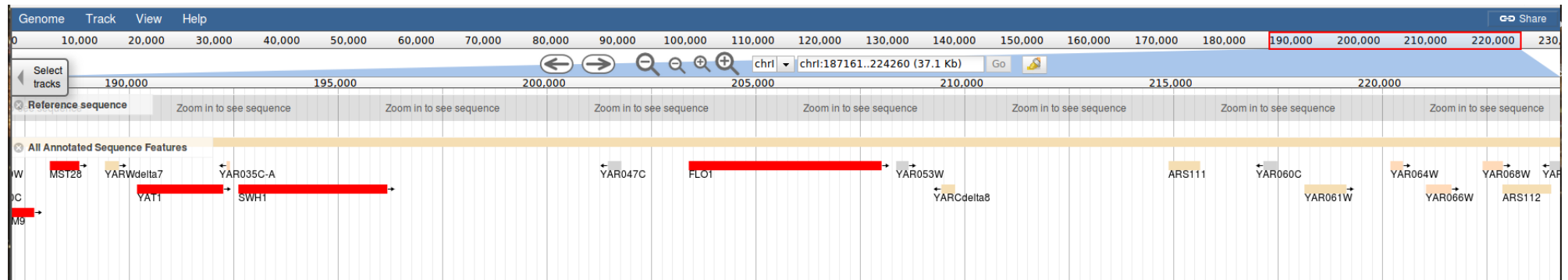
		Second Letter				
		U	C	A	G	
1st letter	U	UUU Phe UUC UUA Leu UUG	UCU Ser UCC UCA UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	3rd letter
	C	CUU Leu CUC CUA CUG	CCU Pro CCC CCA CCG	CAU His CAC CAA Gln CAG	CGU Arg CGC CGA CGG	
	A	AUU Ile AUC AUA AUG Met	ACU Thr ACC ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	
	G	GUU Val GUC GUA GUG	GCU Ala GCC GCA GCG	GAU Asp GAC GAA Glu GAG	GGU Gly GGC GGA GGG	

https://en.wikipedia.org/wiki/Genetic_code

<http://biology.kenyon.edu/courses/biol114/Chap05/code.gif>

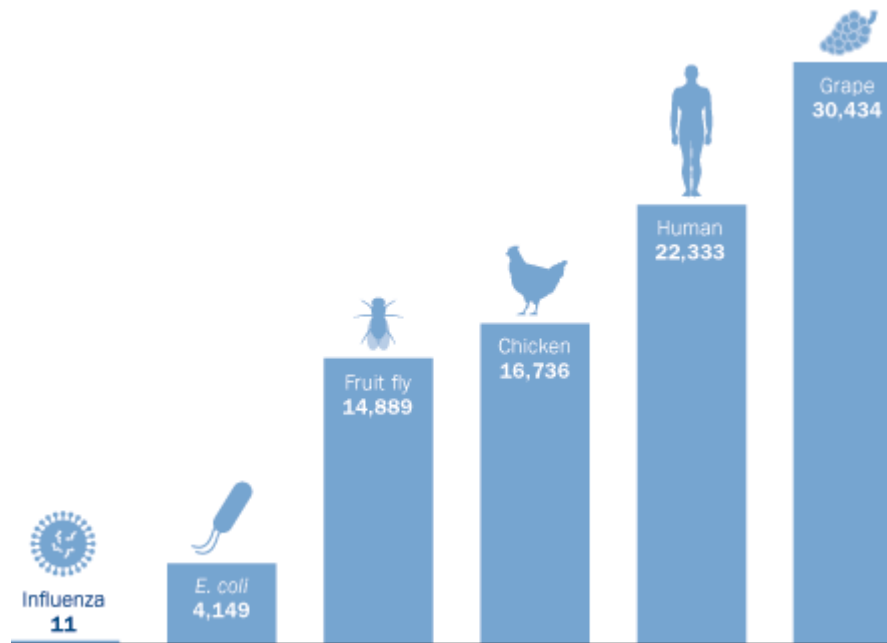
Anotación de genes

- Localización de genes en el genoma



Numero de genes por especie

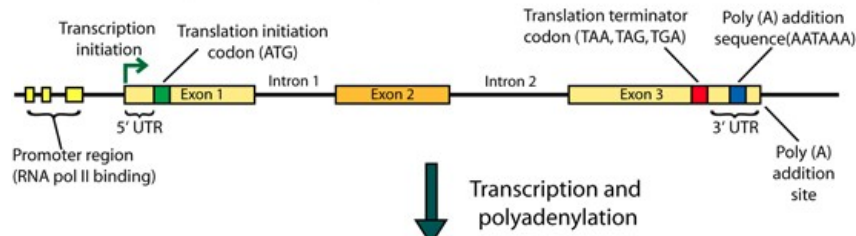
- La complejidad del organismo no correlaciona con el número de genes



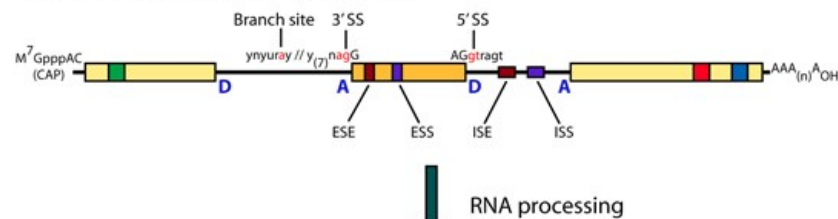
<https://www.sciencenews.org/article/more-chicken-fewer-grape>

Transcripcion en eucariotas

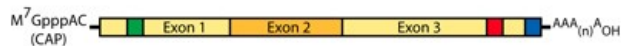
Double-stranded genomic DNA template



Single-stranded pre-mRNA (nuclear RNA)



Mature mRNA

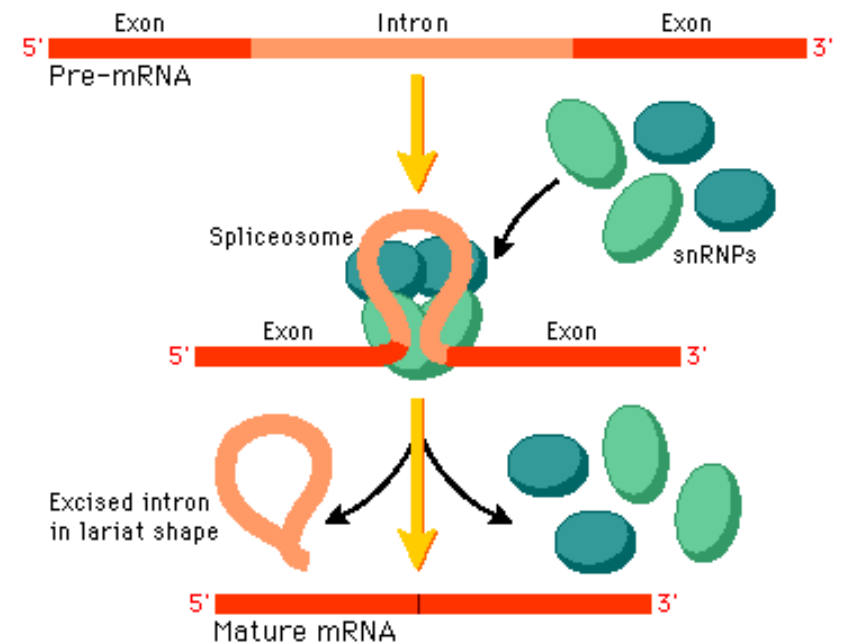
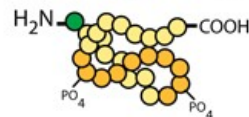


Export to cytoplasm and translation

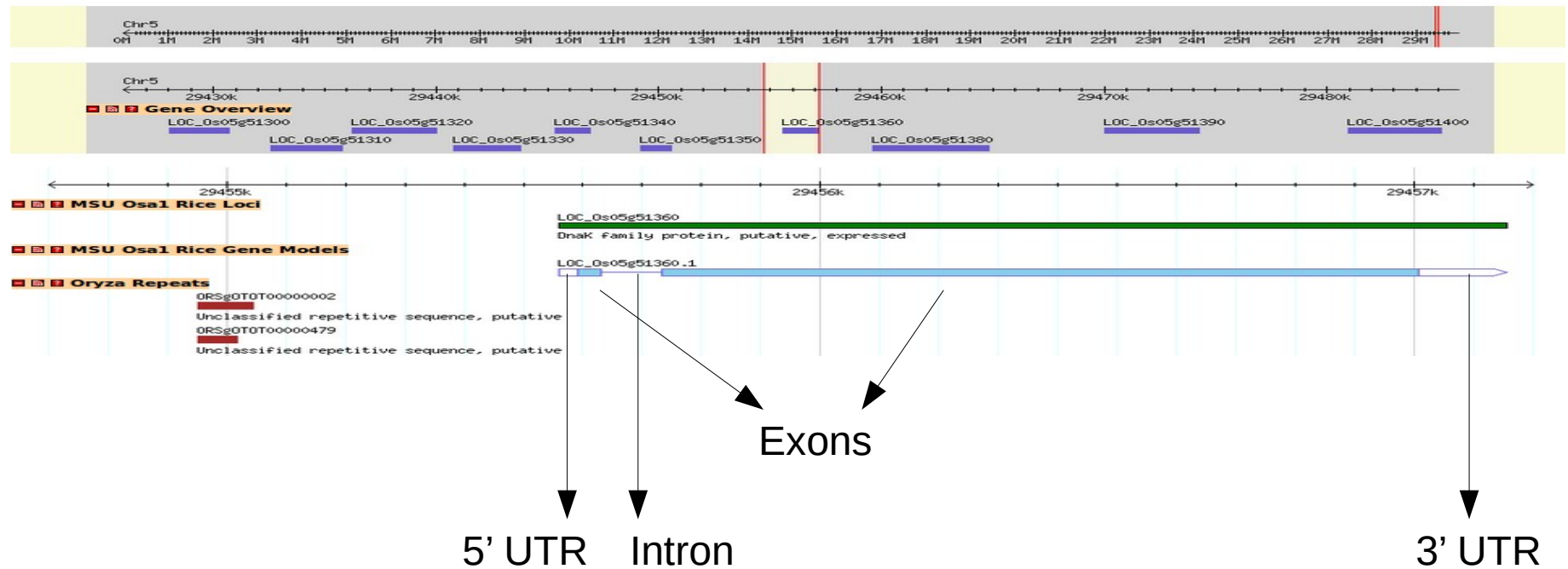
Protein (amino acid sequence)



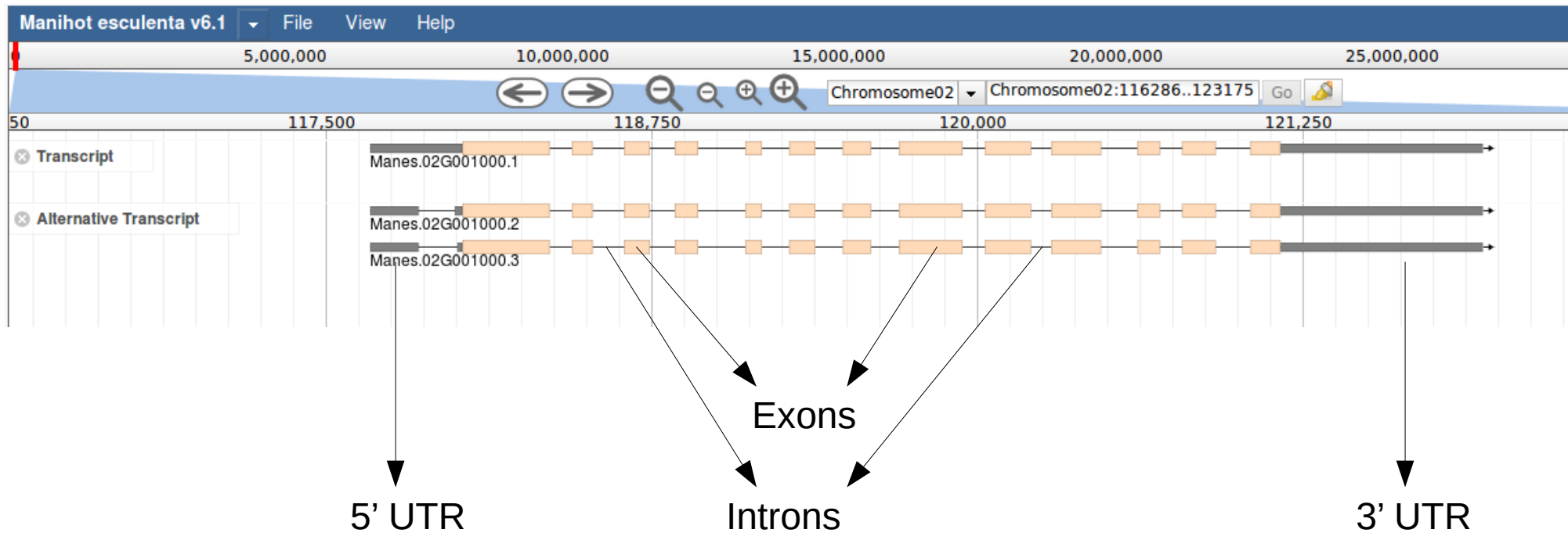
Folding, posttranslational modification, subcellular localization, etc.



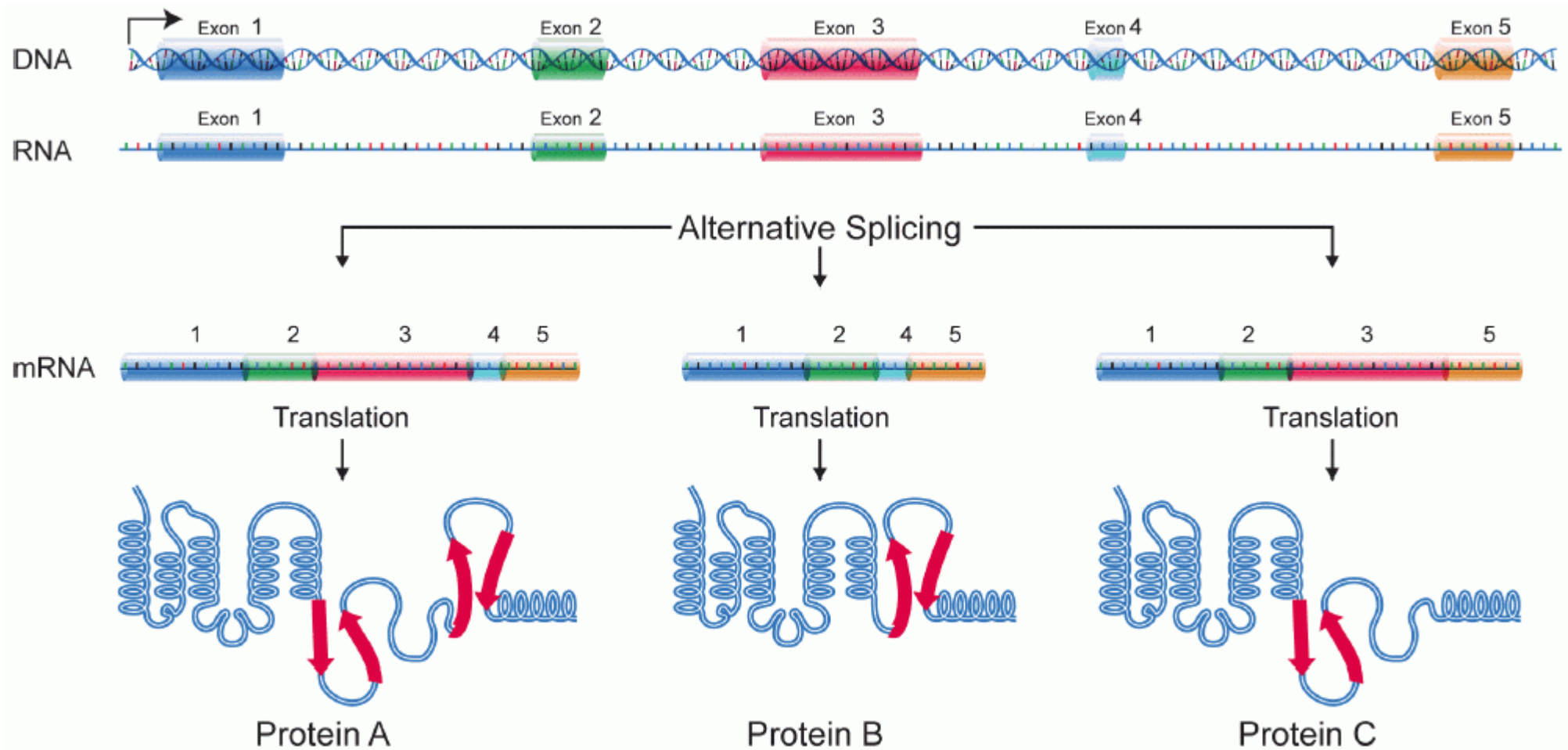
Anotación de genes



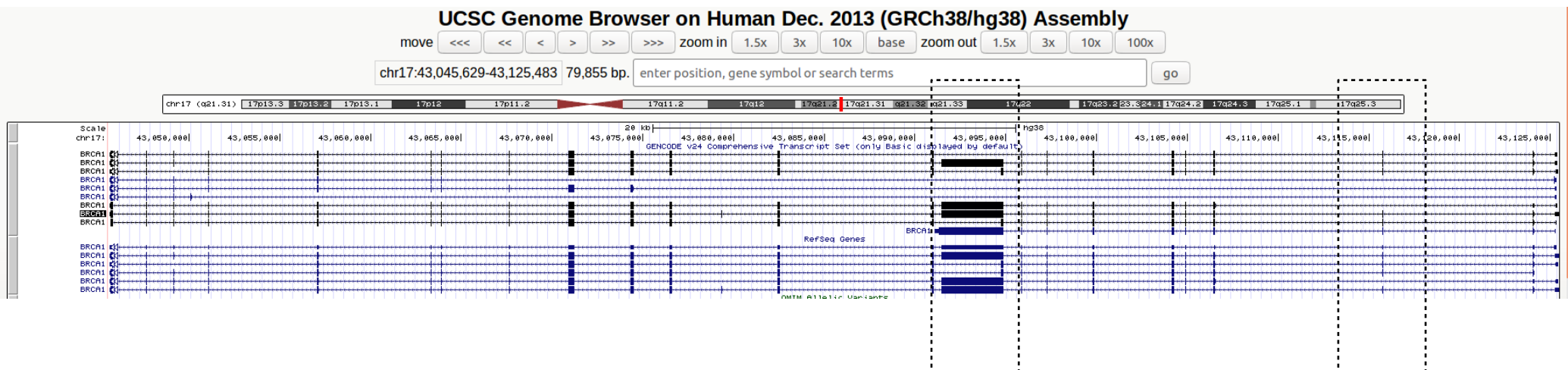
Anotación de genes



Alternative splicing

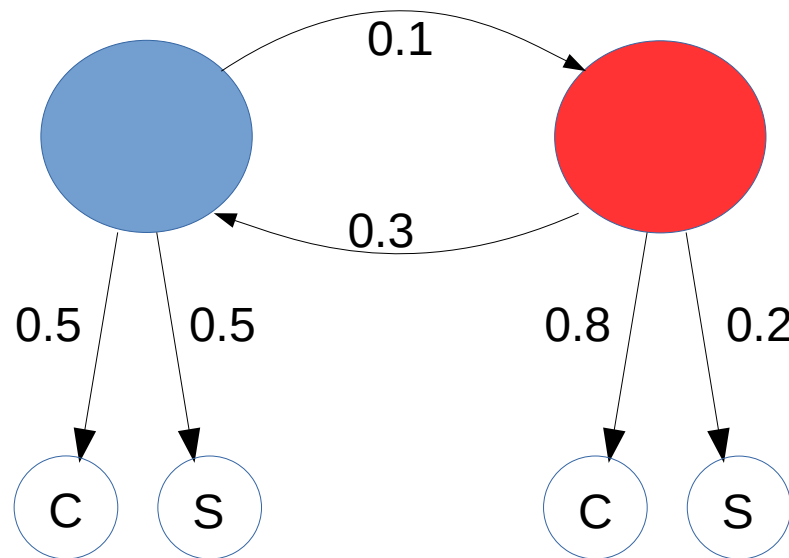


Anotación de genes



Modelos ocultos de Markov

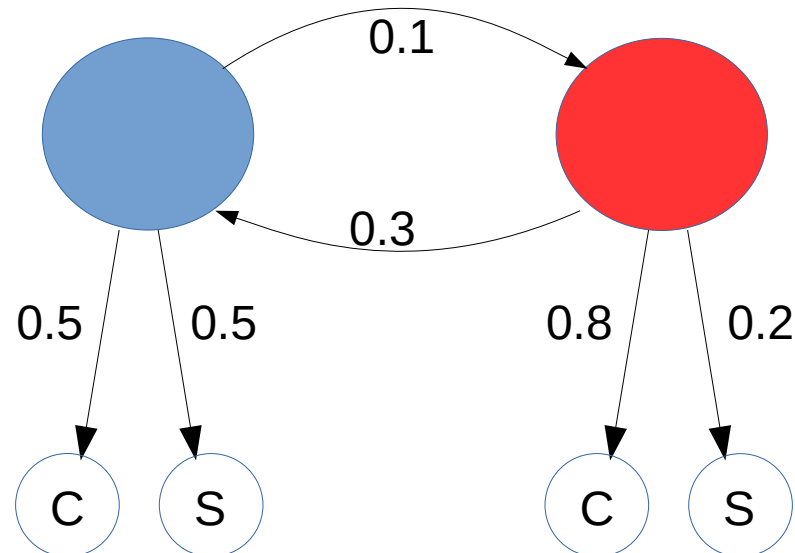
- Los estados de la cadena ya no son visibles
- Cada estado “emite” los datos observados con una cierta probabilidad



$$P(\text{CCCCCSC}) = ?$$

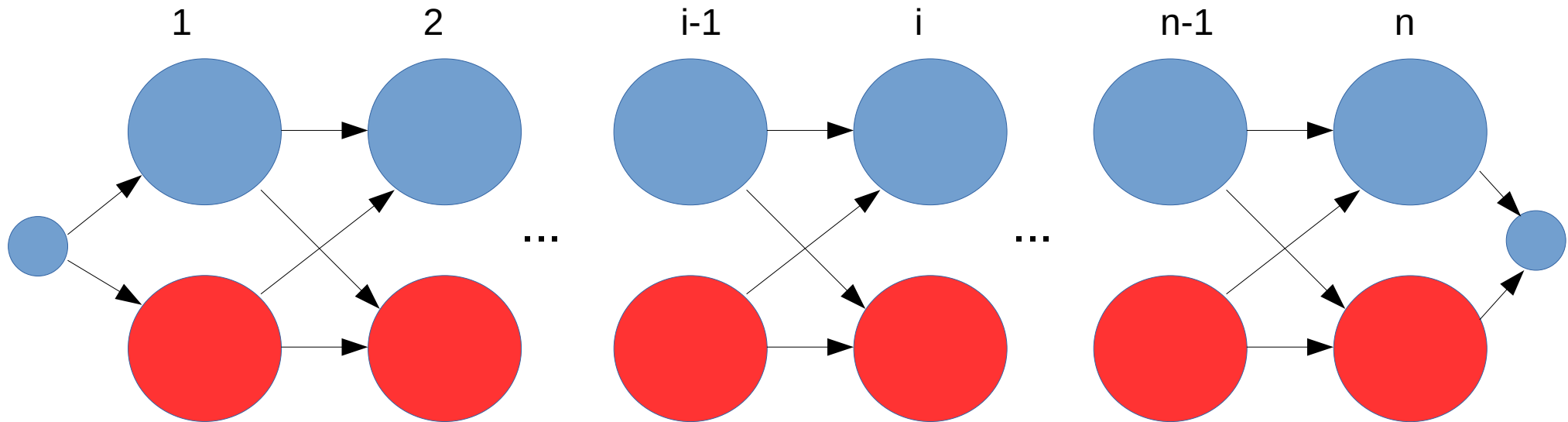
Modelos ocultos de Markov

- Probabilidad de una salida
- Estado más probable en algún momento
- Secuencia de estados más probable que genera una salida
- Probabilidad de una salida en un momento específico
- Generación aleatoria de salidas siguiendo el modelo



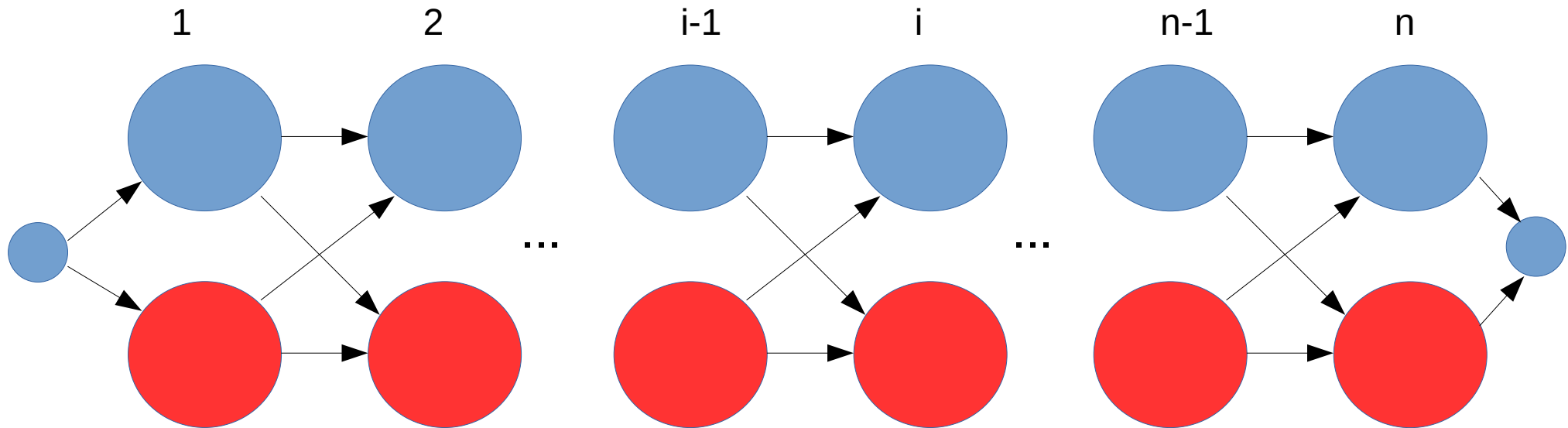
Algoritmo forward

- Teóricamente se debe sumar sobre todos los posibles caminos para calcular la probabilidad de una secuencia S de tamaño n dadas las probabilidades iniciales $i(A)$, $i(R)$, las transiciones t y las emisiones e
- $F(A,1) = i(A) e(A,s[1])$, $F(R,1) = i(R) e(R,s[1])$
- $F(A,i) = F(A, i-1) t(A \rightarrow A) e(A, s[i]) + F(R, i-1) t(R \rightarrow A) e(A,s[i])$, $2 \leq i \leq n$
- $F(R,i) = F(A, i-1) t(A \rightarrow R) e(R, s[i]) + F(R, i-1) t(R \rightarrow R) e(R,s[i])$, $2 \leq i \leq n$
- $P(s) = F(A, n) + F(R, n)$



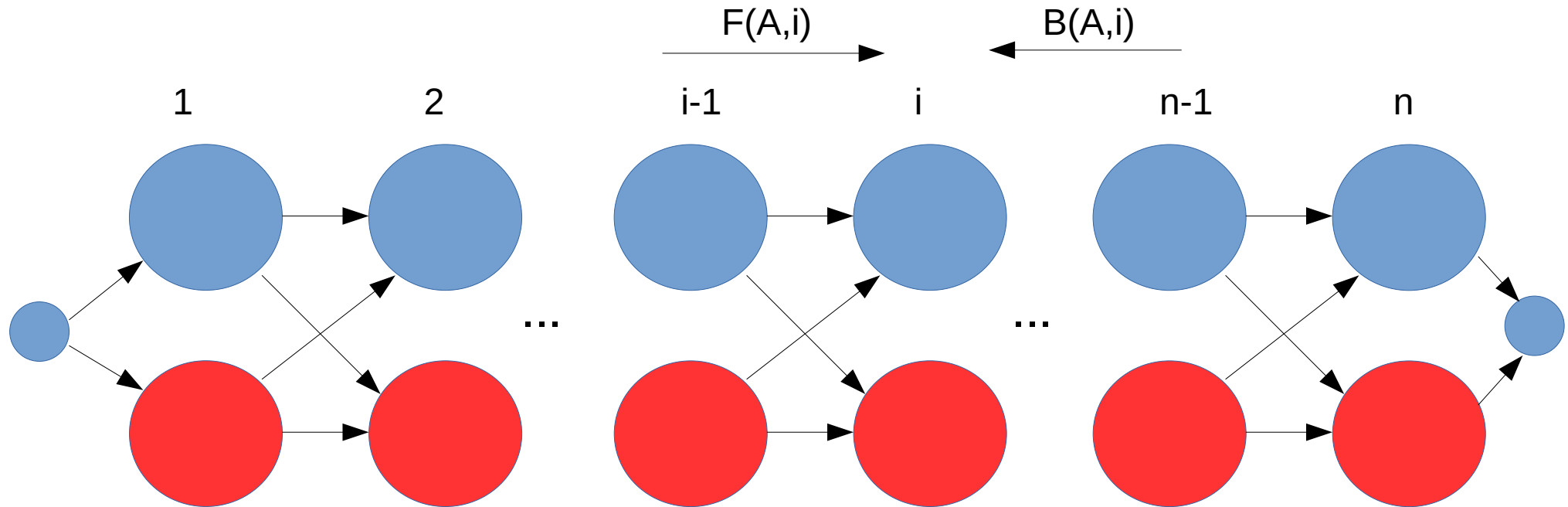
Algoritmo backward

- Se puede sumar también de adelante hacia atrás
- $B(A,n) = B(R,n) = 1$
- $B(A,i) = B(A, i+1) e(A, s[i+1]) t(A \rightarrow A) + B(R, i+1) e(R,s[i+1]) t(A \rightarrow R), 1 \leq i \leq n-1$
- $B(R,i) = B(A, i+1) e(A, s[i+1]) t(R \rightarrow A) + B(R, i+1) e(R,s[i+1]) t(R \rightarrow R), 1 \leq i \leq n-1$
- $P(s) = B(A, 1) e(A, s[1]) i(A) + B(R, 1) e(R, s[1]) i(B)$



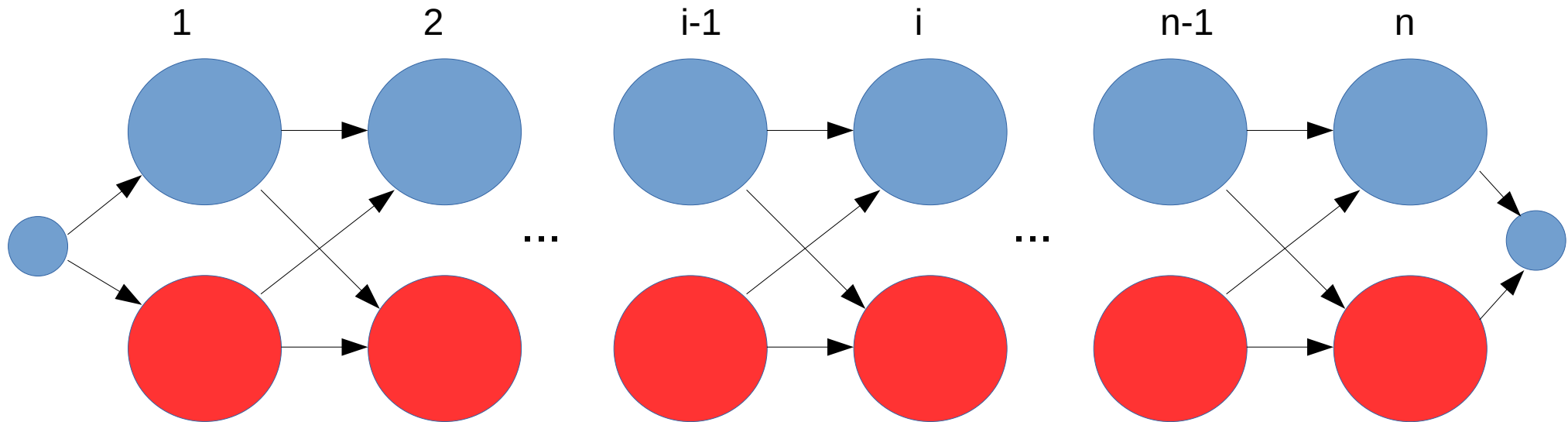
Probabilidad posterior

- La probabilidad posterior de un estado específico en un momento específico i dada toda la secuencia se puede calcular a partir de las probabilidades forward y backward:
- $P(A \mid i, S) = F(A, i) B(A, i)$



Algoritmo Viterbi

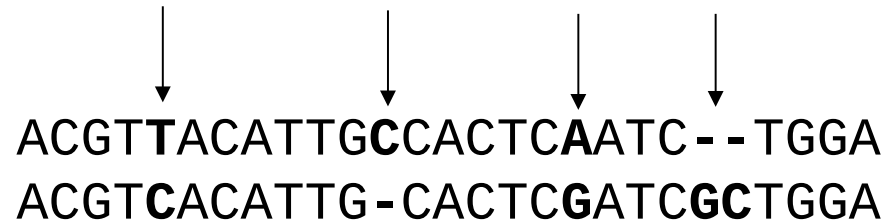
- Parecido al algoritmo forward pero se maximiza en lugar de sumar
- $V(A,1) = i(A) e(A,s[1])$, $V(R,1) = i(R) e(R,s[1])$
- $V(A,i) = \max (V(A, i-1) t(A \rightarrow A), V(R, i-1) t(R \rightarrow A)) e(A,s[i])$, $2 \leq i \leq n$
- $V(R,i) = \max (V(A, i-1) t(A \rightarrow R), V(R, i-1) t(R \rightarrow R)) e(R,s[i])$, $2 \leq i \leq n$
- $V(s) = \max (V(A, n), V(R, n))$



Haplotipado

- Las células de muchos organismos son diploides, es decir, contienen dos copias casi iguales de cada cromosoma

Variantes heterocigotas



ACGTTACATTGCCACTCAATC--TGGA
ACGTCACATTG-CACTCGATCGCTGGA

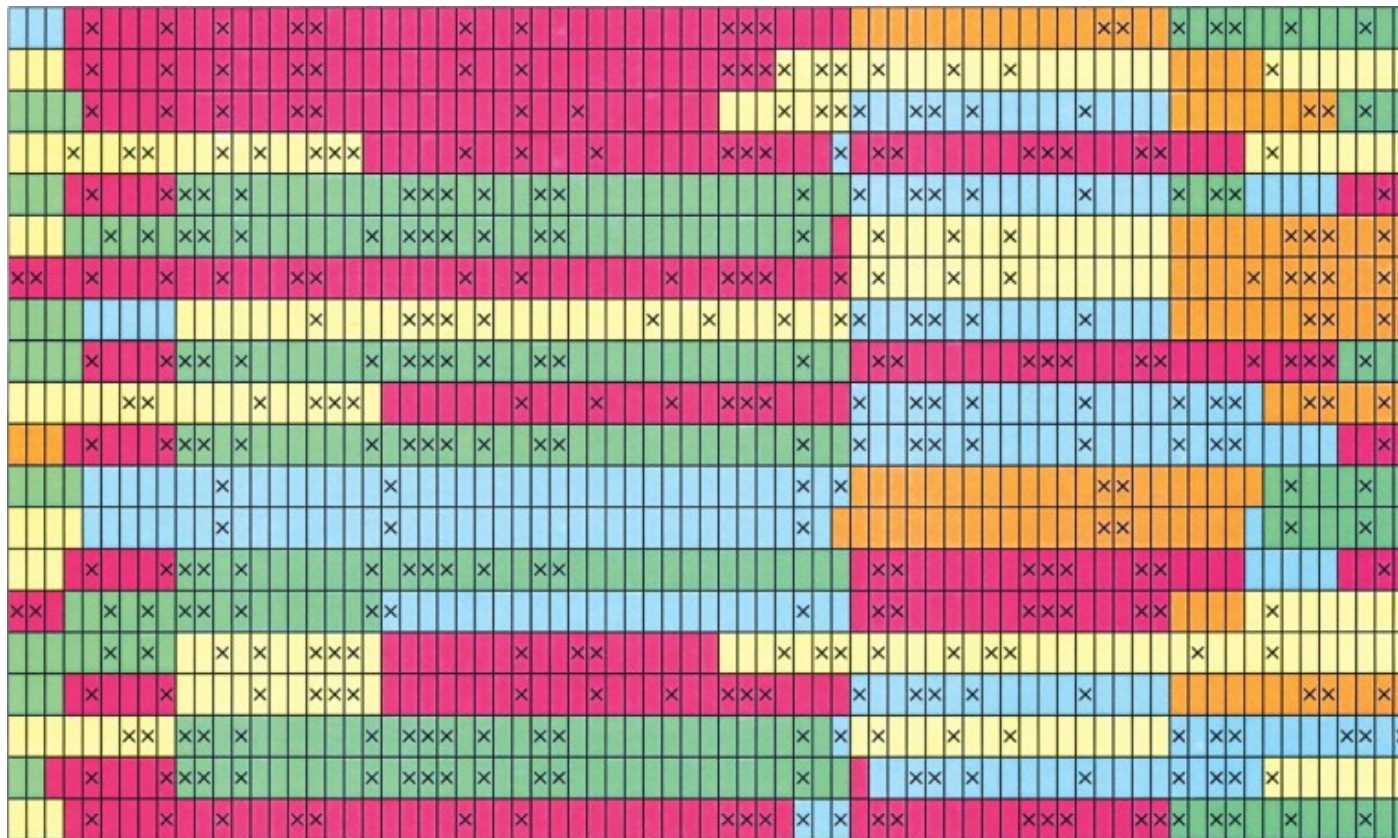
Haplotipado estadístico

- Se usan genotipos o haplotipos de otros miembros de la población a la que pertenece el individuo para adivinar la mejor configuración

Sitio	1	2	3	4
Alelos	C,T	A,C	G,T	A,G
P1	CC	CC	GT	AG
P2	TT	AA	GG	GG
P3	TT	AA	GG	AG
P4	CC	CC	GT	GG
Haplotipo 1	C	C	t	?
Haplotipo 2	T	A	g	?

Haplotipado estadístico

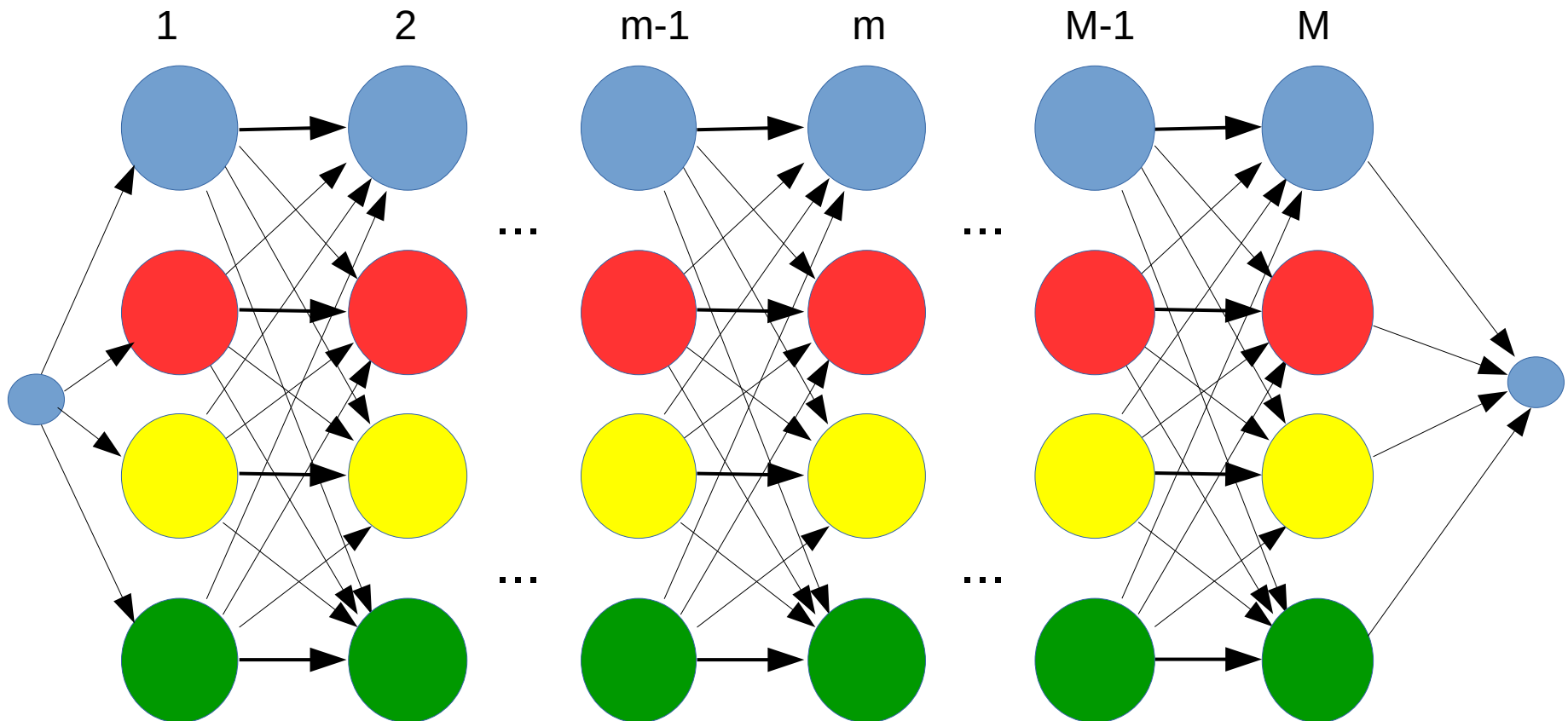
- FastPHASE implementa clusters locales de haplotipos



Scheet P and Stephens M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. The American Journal of Human Genetics 78 (4): 629 – 644, 2006.

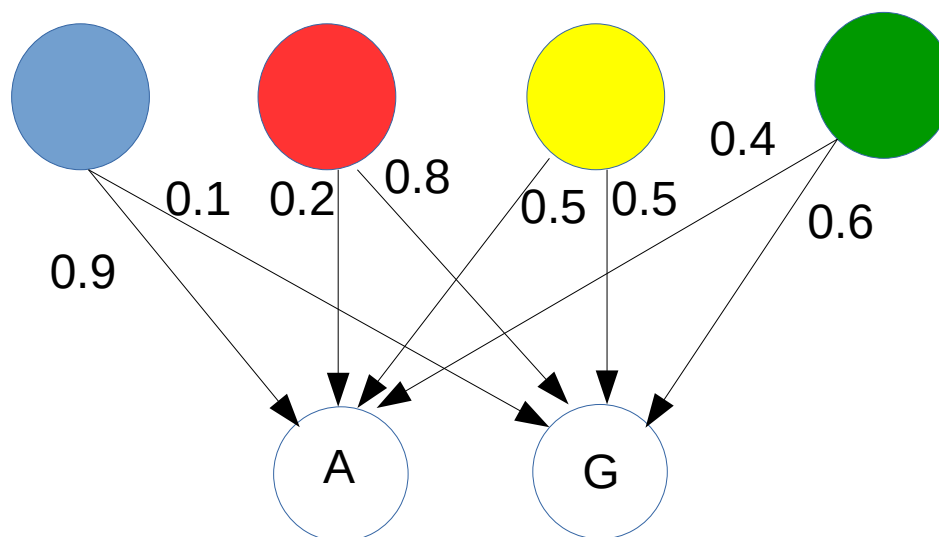
Haplotipado estadístico

- Cada cluster de haplotipos es un estado del modelo
- Las transiciones están relacionadas con la probabilidad de recombinación



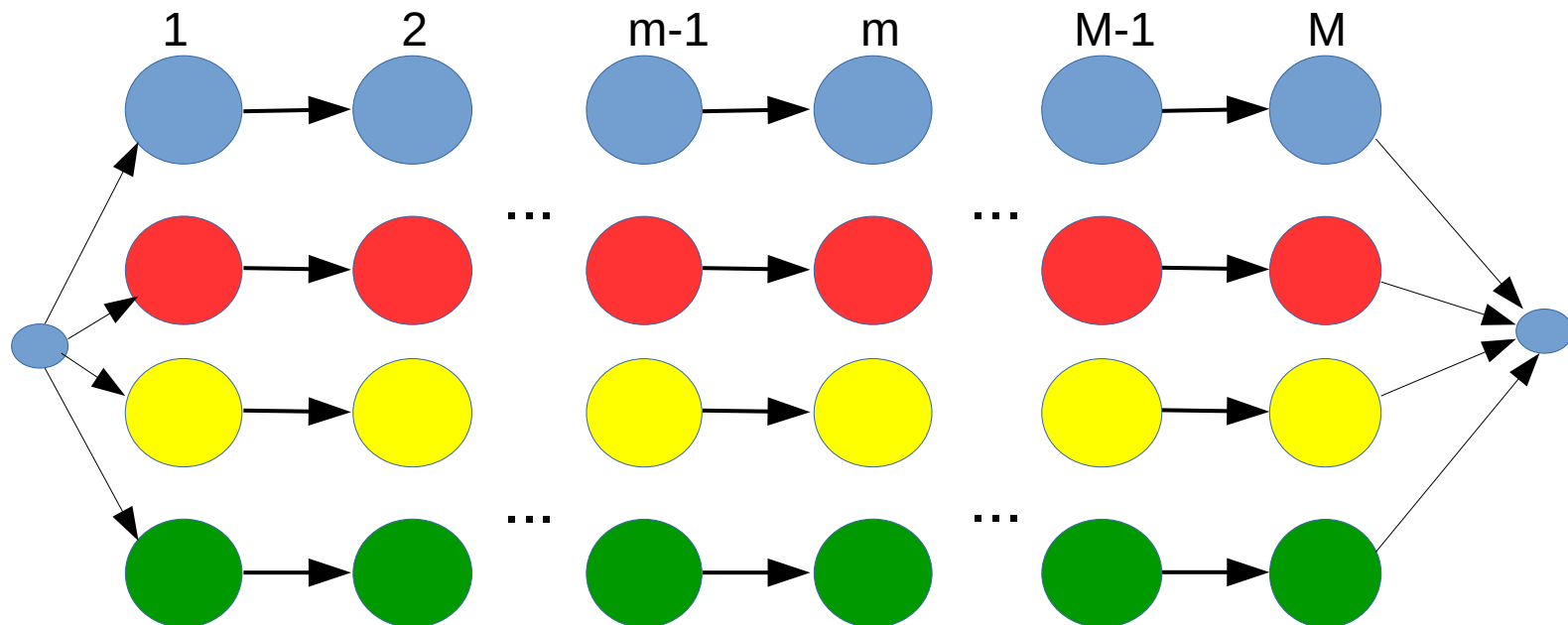
Haplotipado estadístico

- Cada cluster de haplotipos emite los dos alelos con diferentes probabilidades



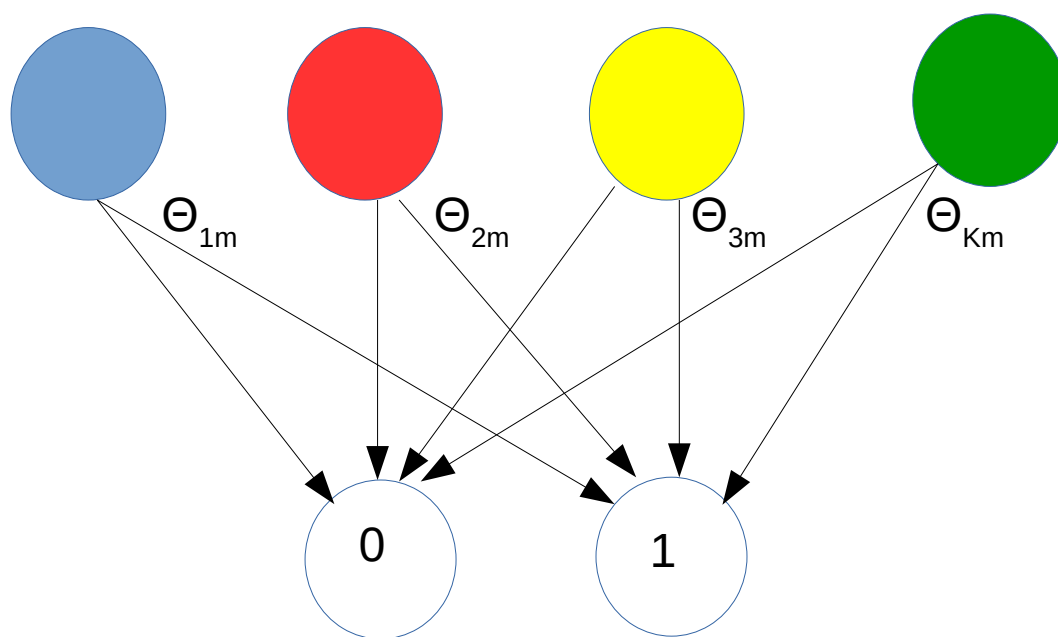
Haplotipado estadístico

- Notación K clusters, $1 \leq i \leq n$ haplotipos, $1 \leq m \leq M$ SNPs bialélicos
- $1 \leq z_i \leq K :=$ Cluster al que pertenece el haplotipo h_i
- $0 \leq a_k \leq 1 :=$ Frecuencia relativa del cluster k . $p(z_i = k \mid \mathbf{a}) = a_k$



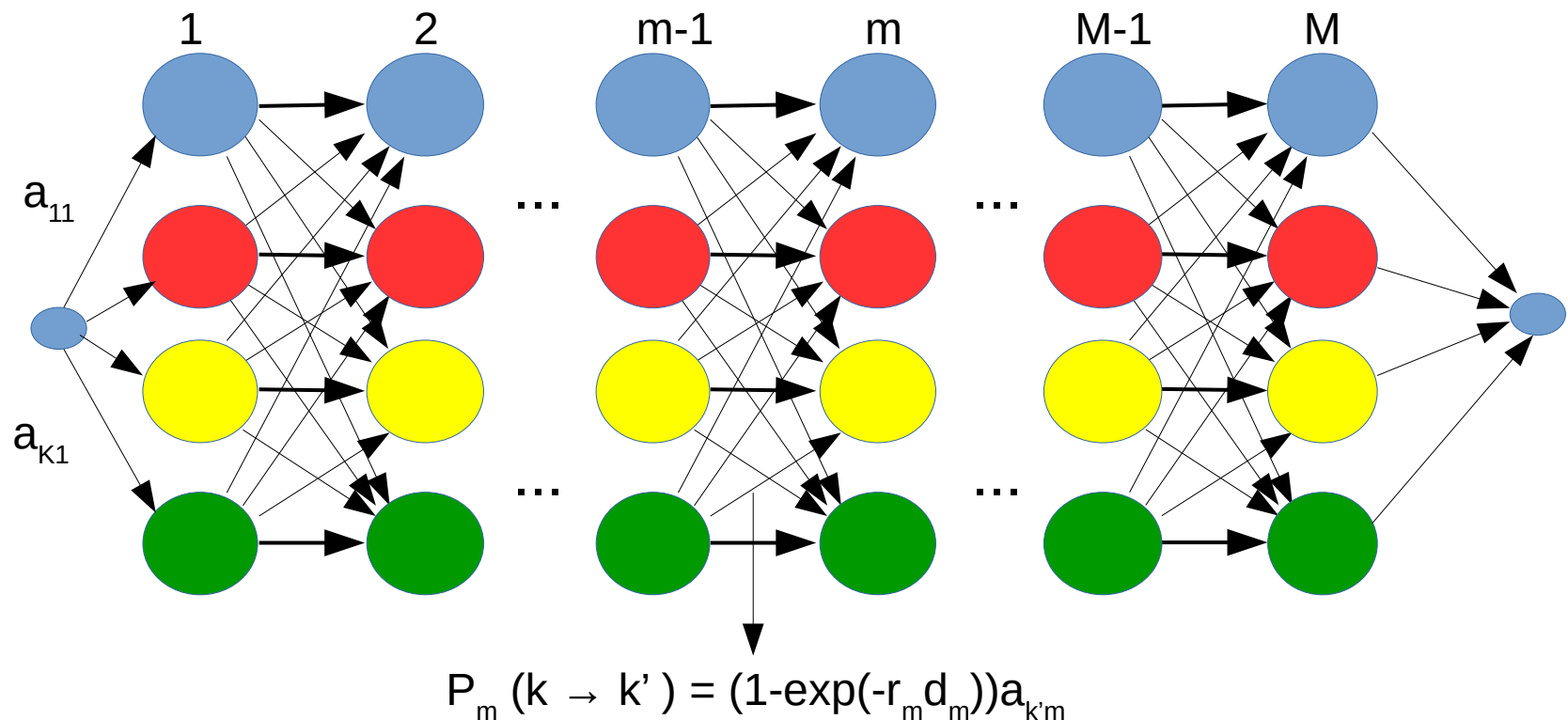
Haplotipado estadístico

- $0 \leq \Theta_{km} \leq 1$:= Frecuencia del alelo 1 del SNP m en el cluster k
- $P(h_i | z_i = k, \Theta)$: Producto de emisiones por sitio
- $P(h_i | a, \Theta)$: Suma sobre los posibles clusters de $P(h_i | z_i = k, \Theta)$



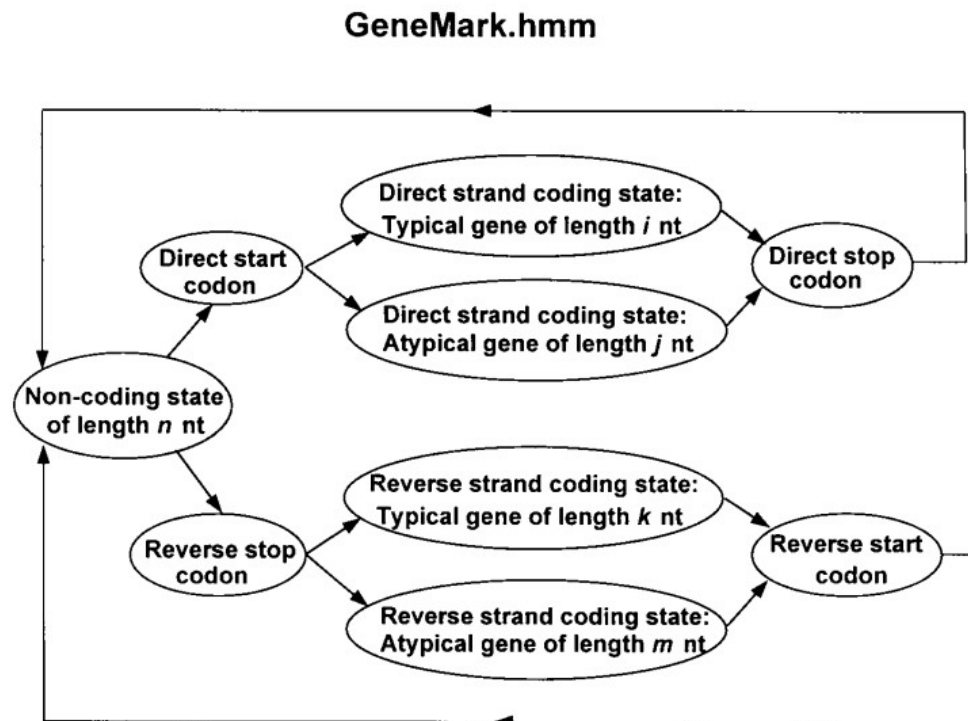
Haplotipado estadístico

- Para permitir transiciones, z_i se convierte en z_{im}
- $1 \leq z_{im} \leq K$: Cluster de origen de h_{im} . Estados ocultos del modelo
- Las transiciones están relacionadas con una “tasa de recombinación local” r_m y la distancia física d_m



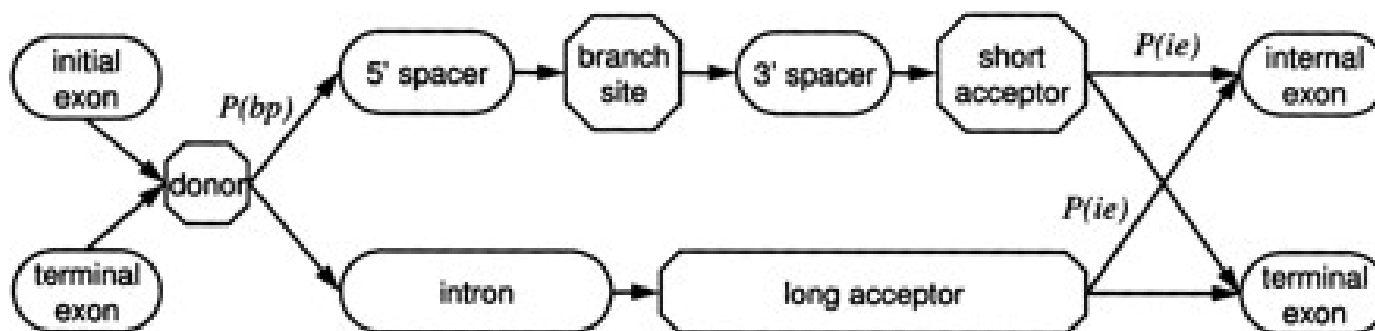
Anotación de genes: GeneMark

- Estado inicial “Non-coding”
- Estados para codones de inicio y parada
- Estados separados para strand forward y reverse
- Dos tipos de genes



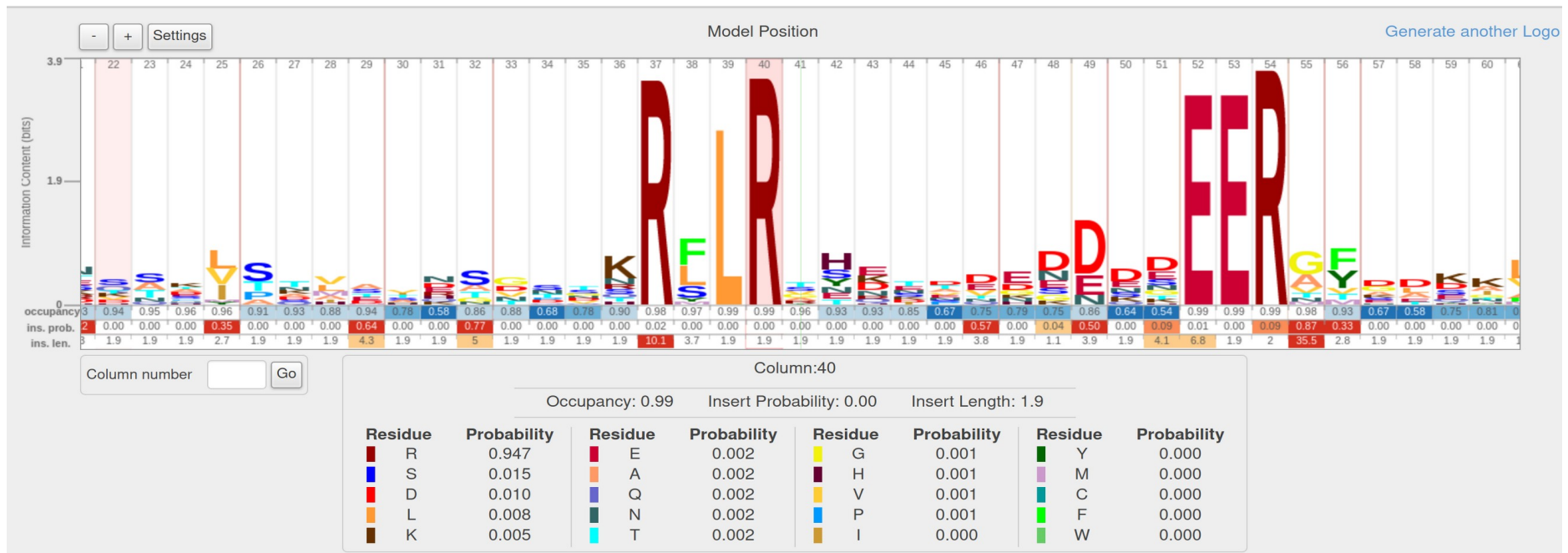
Anotación de genes: GeneMark

- Modelo de detección de intrones



Detección de homología

- Dado un conjunto de secuencias relacionadas, determinar si una secuencia nueva pertenece a la familia
- En ausencia de indels, la pertenencia de una secuencia a una familia se podría determinar con una matriz de abundancias de aminoácidos



Detección de homología

- Se utiliza un modelo de Markov para tener en cuenta posibles inserciones y borrados

