Genomas Ensamblado y Anotación

Rodrigo Lisperguier
Felipe Ramírez
Departamento de Informática
Universidad Técnica Federico Santa María

Introducción

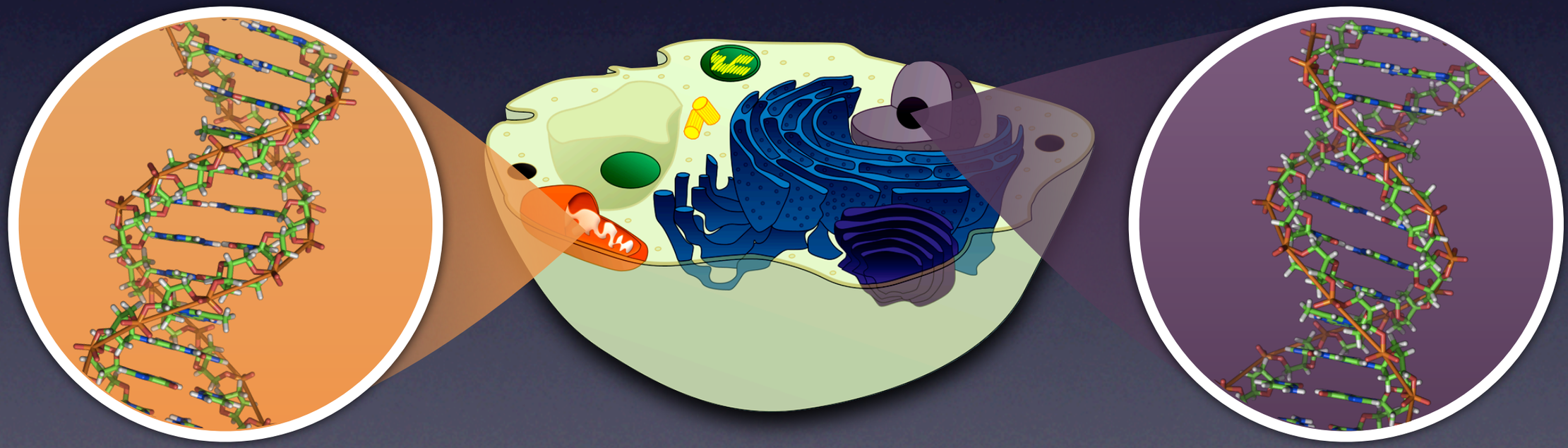
Introducción

Genoma

- Información genética de un organismo, contenida en su DNA.
- Los trozos inútiles ¿son parte de la información?

Genoma Mitocondrial

Genoma



Genoma = Secuencia de letras del DNA en el núcleo

Introducción

El genoma difiere entre individuos de una misma especie, pero no tanto como para que no se pueda hablar del ***genoma de la especie****.*

GENOMA1 ≈ GENOMA2 ≈ GENOMA3 ≈ GENOMA4 ≈ GENOMA5 ≈ GENOMA6 ≈ GENOMA7 ≈ GENOMA8
¡Pero no son iguales!

Introducción

Eucariotas

Genomas grandes
Baja densidad de genes
Intrones y exones
Identificación de genes es un problema complejo

Procariotas

Genomas pequeños
Alta densidad de genes
Sin intrones
Identificación de genes es relativamente simple

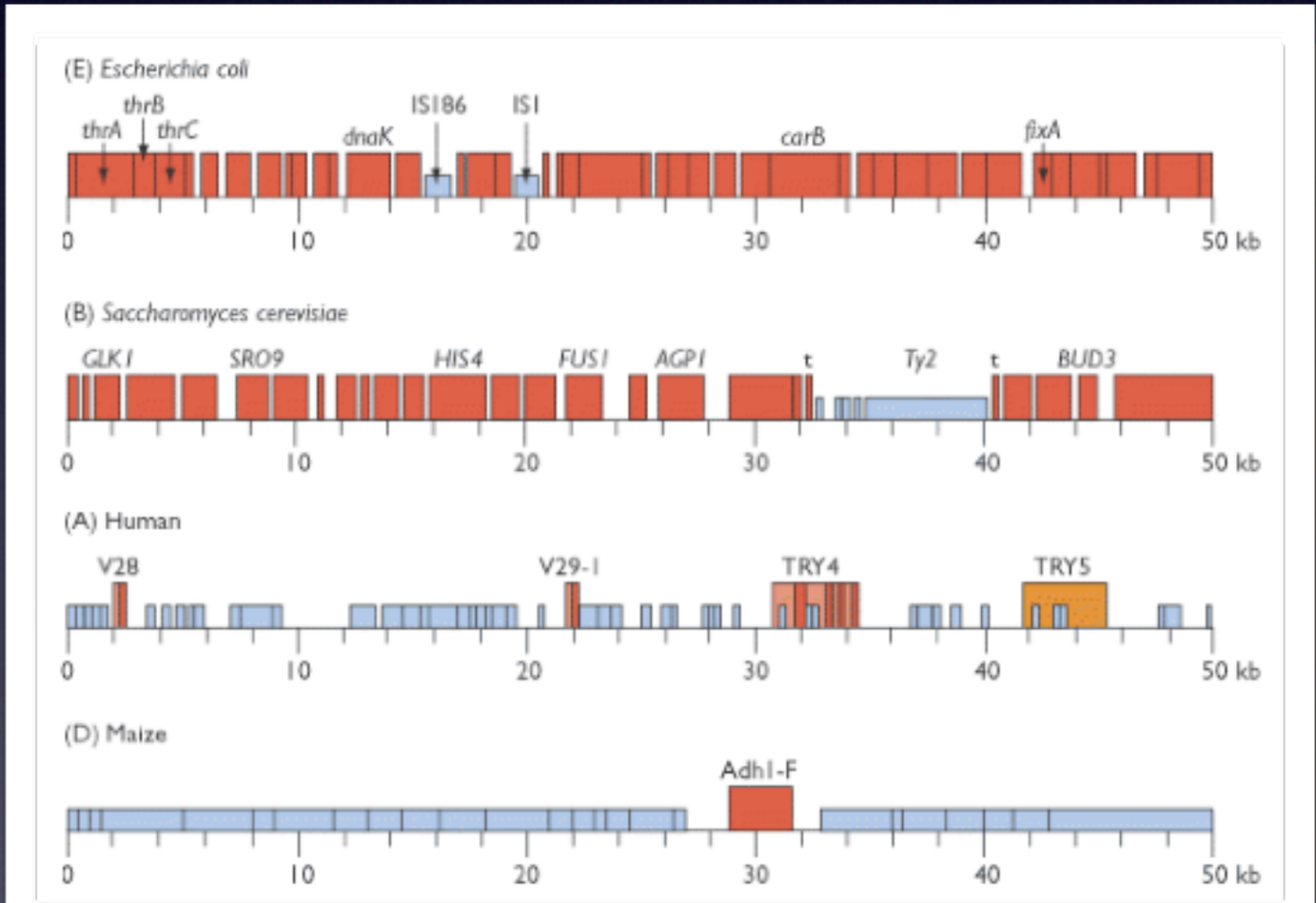
*A parte de esto, no hay relación entre el **tamaño del genoma** o la **cantidad de genes** y la complejidad de la especie*

Paradoja del valor C *"el tamaño del genoma en eucariotas no tiene relación con la complejidad del organismo"*

Introducción

¿Qué hay en el Genoma?

- **Genes** que codifican proteinas
- **Genes** que codifican RNAs estructurales y otros tipos de RNA
- Secuencias de control
- ¿Qué es **todo lo demás**?



E. Coli	90%
S. cerevisiae	50%
Human	2%
Maize	1%

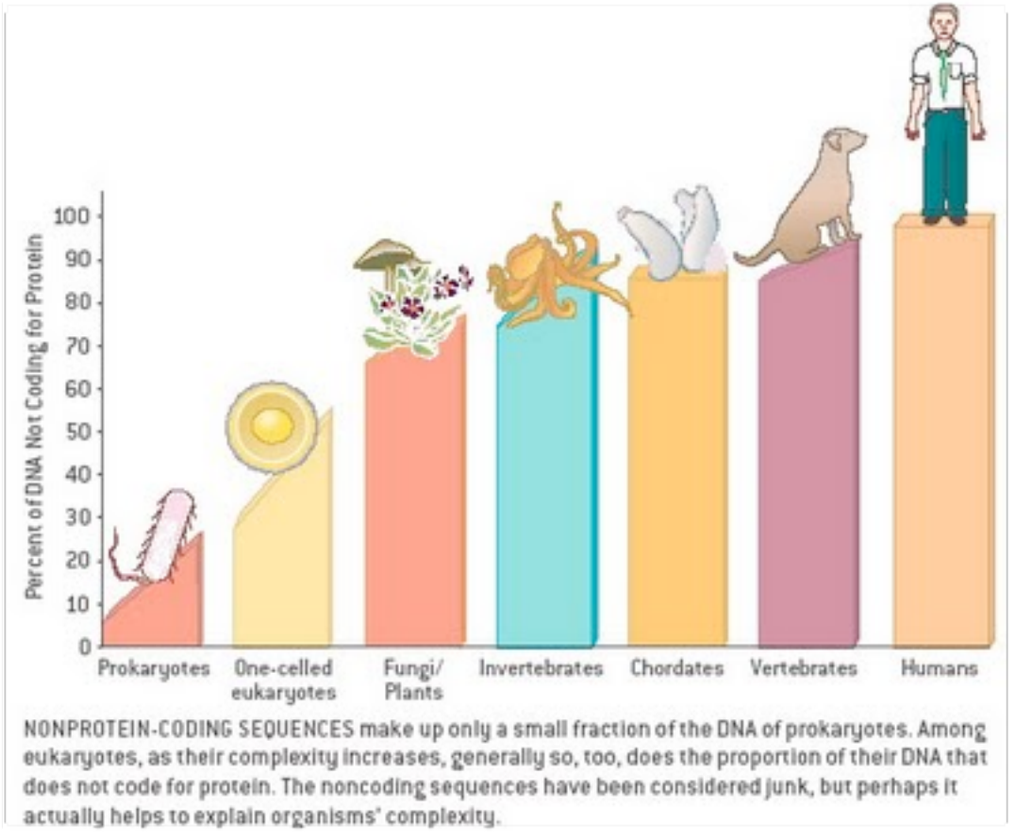
Introducción

*¿Qué es **todo lo demás**?*

- Secuencias repetitivas
- Elementos móbiles con instrucciones para replicarse
- Varias A y T (tienen relación con la estructura de la cromatina)
- Minisatélites / Microsatélites
- Nada reconocible
- Duplicados en desuso
- Etc, etc, etc...

Introducción

DNA Basura



NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to explain organisms' complexity.

"Las secuencias que no codifican proteínas sólo forman una pequeña fracción del DNA de las procariotas. Entre las eucariotas, a medida que su complejidad aumenta, también lo hace su proporción de DNA que no codifica proteinas.

Las secuencias que no codifican se consideran basura, pero tal vez en realidad ayudan a explicar la complejidad de los organismos."

The relationship between non-protein-coding DNA and eukaryotic complexity

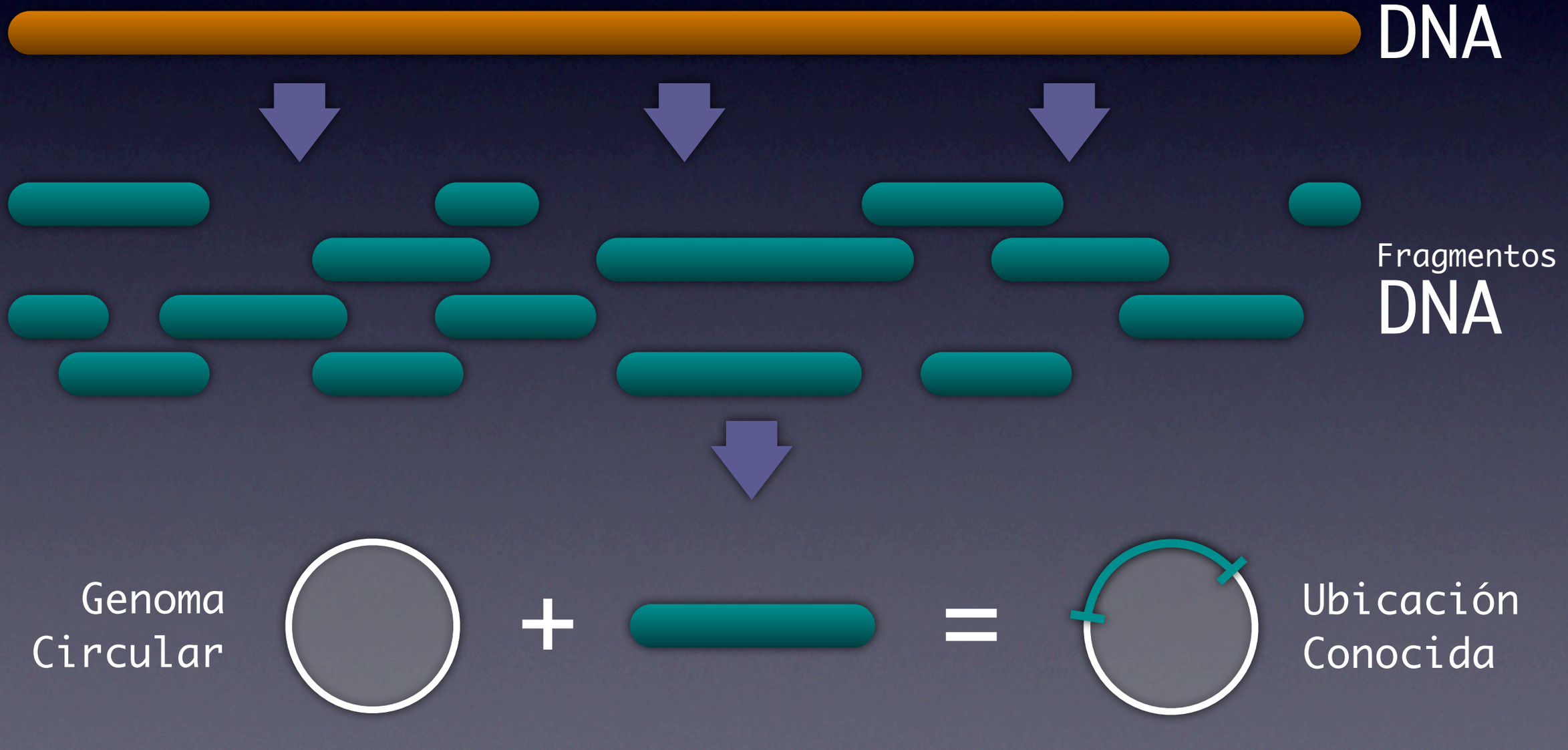
*Ryan J. Taft, Michael Pheasant, John S. Mattick **
doi:10.1002/bies.20544

Ensamblado

Ensamblado

¿Cómo se secuencia un genoma?

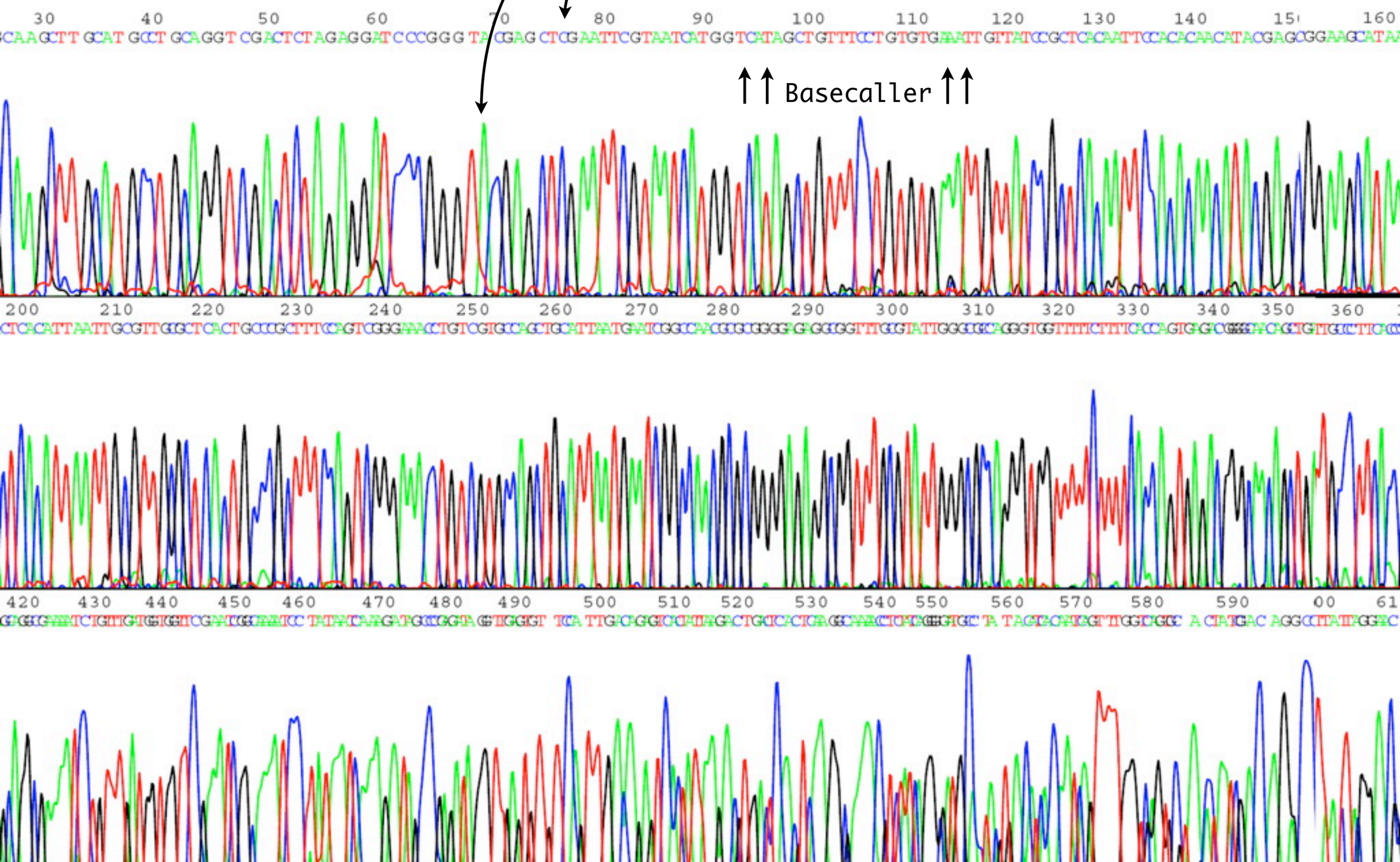
- La tecnología permite leer cadenas pequeñas (menor a 1Kbp).
- Hay que trozar los cromosomas, leer los trozos y reunirlos.
- Entremedio se amplifican vía clones en bacterias.



Ensamblado

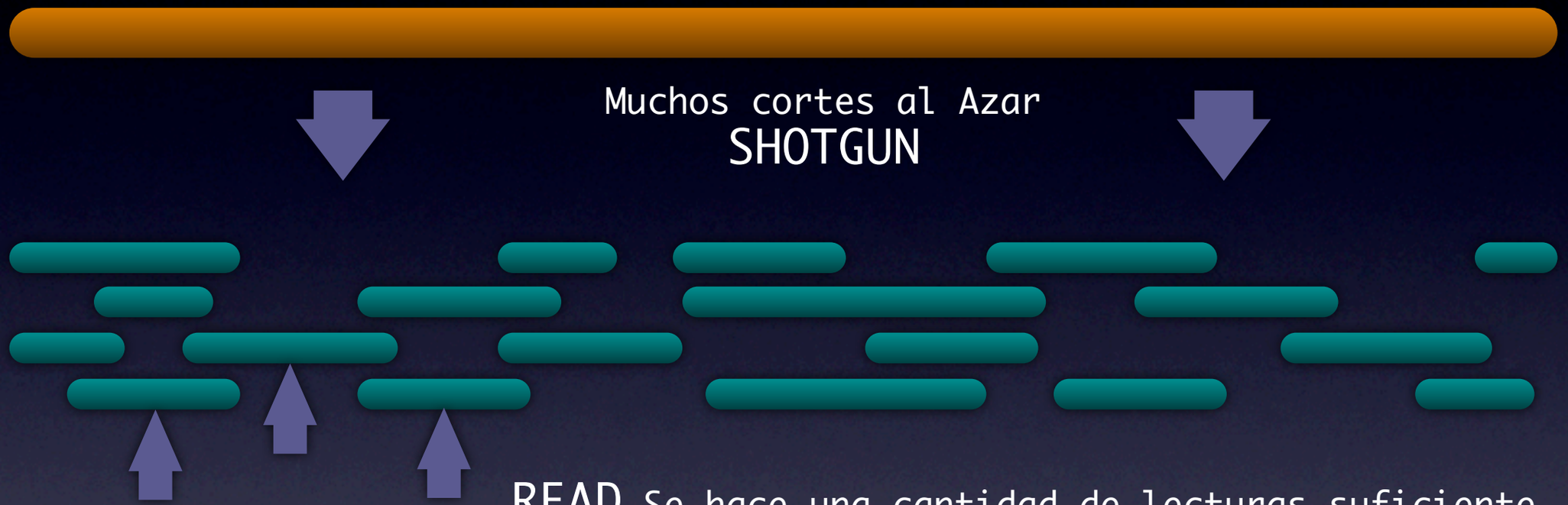
algo así entregan las máquinas

secuencia resultante de aplicar basecaller



Ensamblado

Para secuencias más largas que 500bp ***Shotgun***



READ Se hace una cantidad de lecturas suficiente para leer en promedio C veces la secuencia.

C~7 en general
C~12 para el Proyecto del Genoma Humano

- Luego se identifican traslapes entre trozos, para obtener la lectura de la secuencia completa.
- Problema de alineamiento (computacionalmente difícil (más difícil aún por presencia de *repeats*))

Ensamblado

Estrategias

- Whole Genome Shotgun: hacer shotgun sobre el cromosoma completo, hacer muchas lecturas, e ir pegándolas en segmentos cada vez mayores.
  - Costo y errores al pegar secuencias.
- Hierarchical Shotgun: cortar el cromosoma en trozos largos ("clones"), mapearlos con respecto al cromosoma, clonarlos en bacteria, y aplicar shotgun a cada trozo.
  - Problema de mapear bien los clones.

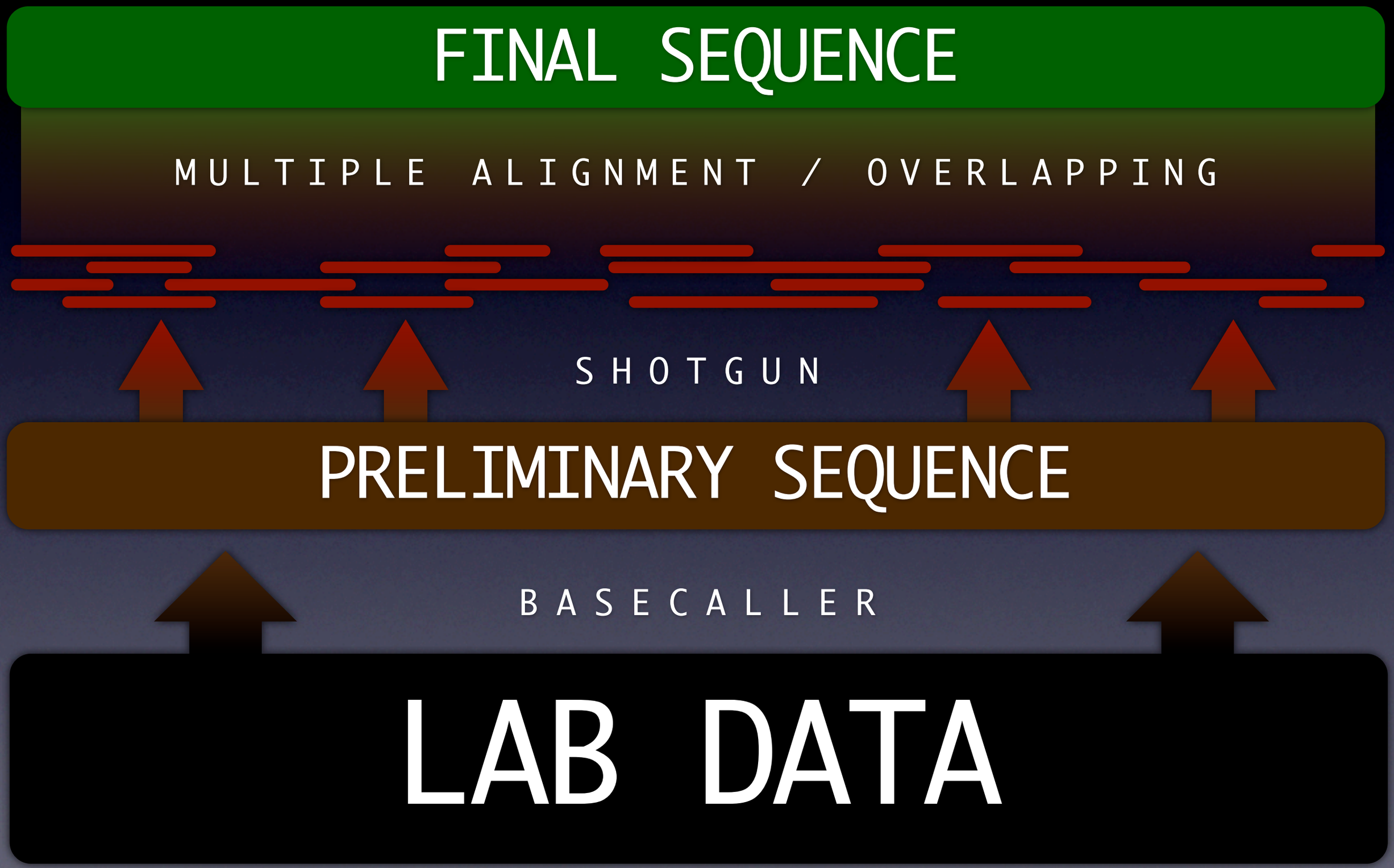
Para ambas estrategias en algún momento hay que fusionar lecturas. Ahí se trata de pegar trozos leídos, eventualmente con errores, y posiblemente con repeticiones en algunas partes. **K-tuplas.**

K~24

```
TACATAGATTACACAGATTACT GA
||   ||||||||||||||||||| ||
TAGTTAGATTACACAGATTACTAGA
```

Ver qué trozos tienen una k-tupla coincidente. En ese caso, extender el alineamiento. Si no da 95% de identidad, descartarlo

Ensamblado

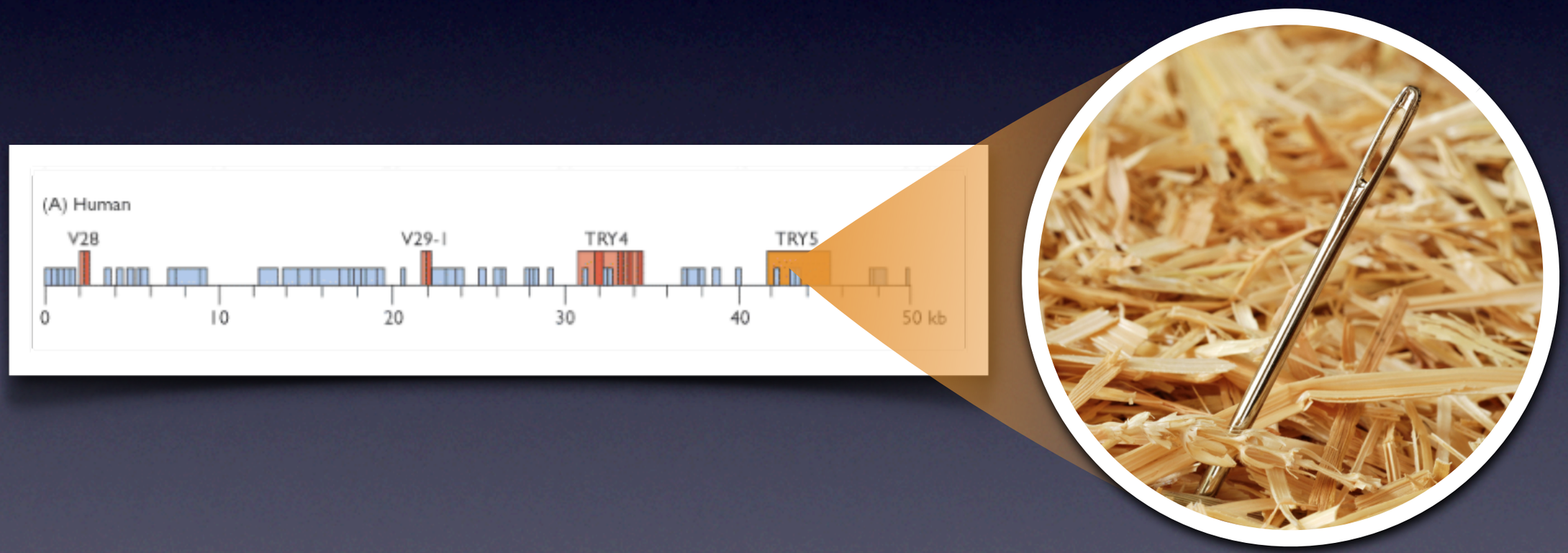

FINAL SEQUENCE
MULTIPLE ALIGNMENT / OVERLAPPING
SHOTGUN
PRELIMINARY SEQUENCE
BASECALLER
LAB DATA

Anotación

Anotación

Anotando un genoma = predecir genes

- Una vez que tenemos la secuencia de un genoma, lo siguiente es ver qué es lo que está escrito ahí.
- Se buscan secuencias que codifiquen proteínas y secuencias que codifiquen RNAs estructurales.

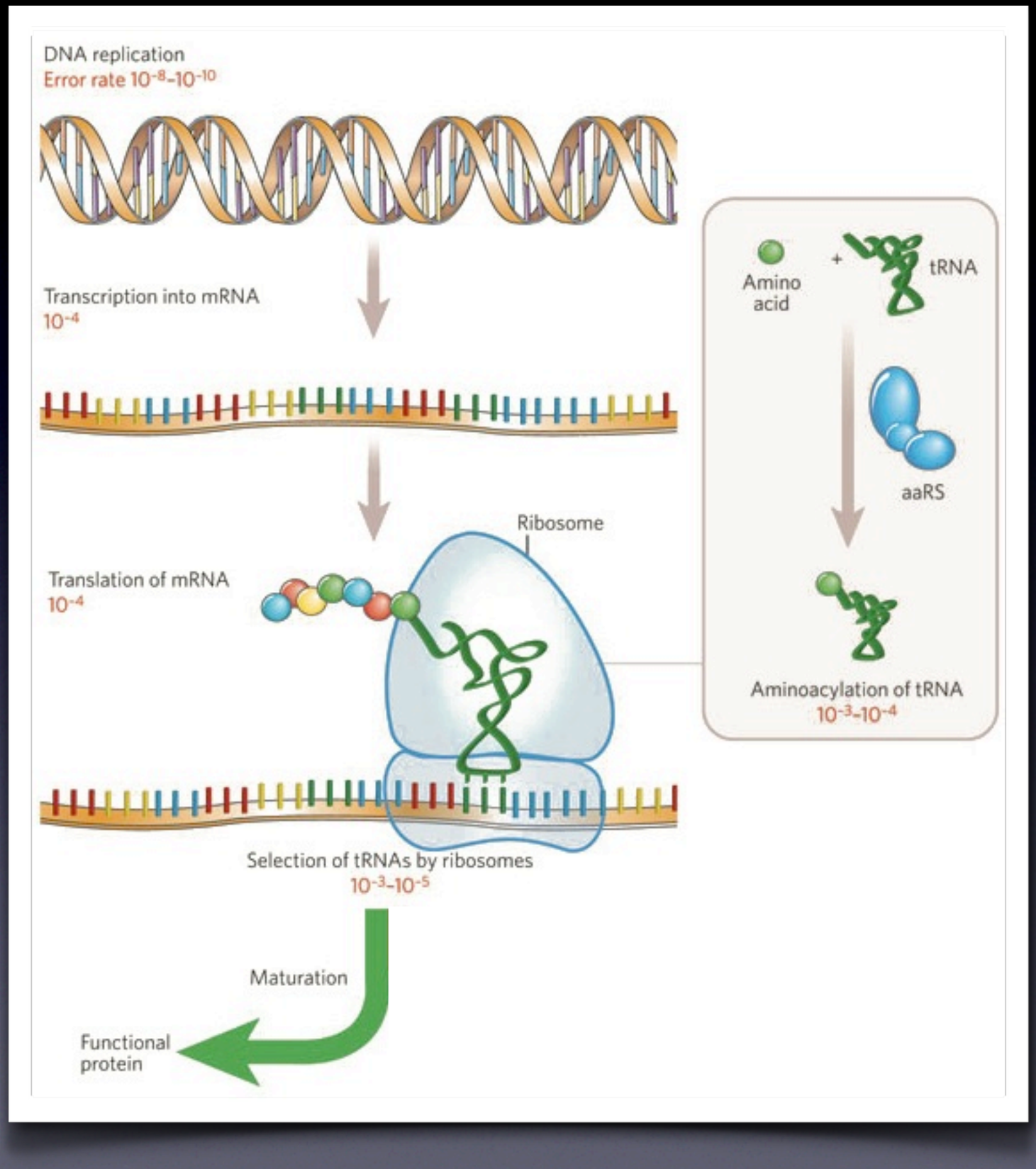


Difícil tarea en eucariotas con tanta basura...
Se usan diferentes estrategias en eucariotas y procariotas.

Anotación

Genes que codifican proteínas

- Recordatorio
- Diferencias entre **mRNA**
  - Procariota
  - Eucariota



Anotación

mRNA

- Procariota

- Eucariota (editado)

En regulación las eucariotas también son más complicadas...
Por eso se usan distintos métodos de anotación para cada caso.

Anotación

ORF (open reading frame)

- Es una sección del genoma de un organismo que posiblemente puede codificar una proteína y es la forma más simple de búscar estas secciones.
- El ORF se encuentra entre una secuencia inicial o codón de inicio (AUG para RNA) y la secuencia de término o codón de término.
- La existencia de un ORF relativamente largo, es un buen indicador de la existencia de un gen en ese sector.
- Seis posibles formas (3 en una hebra y otras 3 en la hebra complementaria)

UCUAAAAUGGGUGAC

UCUAAAAUGGGUGAC
CUAAAAUGGGUGAC
UAAAAUGGGUGAC

Anotación

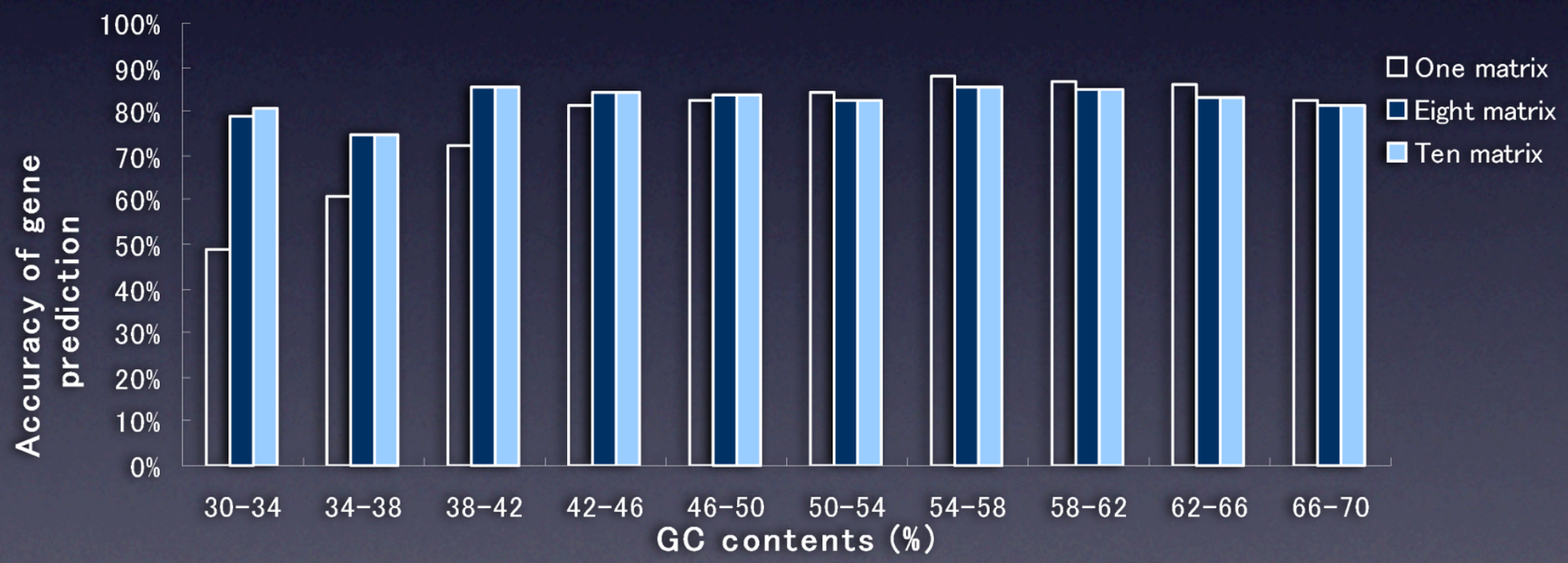
ORF (open reading frame)

- En procariotas el mayor ORF comenzando desde el primer codón de start hasta el primer codón de stop es una buena (pero no segura) predicción de una región que codifica proteínas.
- En eucariotas es algo más complejo debido a la presencia de intrones que suelen generar codones de stop que no necesariamente representan el término de la secuencia codificadora.

Anotación

GC Content

- GC Content (%GC) es el porcentaje de bases Guanina y Citocina presentes en el DNA.
  - Se puede obtener el %GC de todo el genoma o de un trozo de DNA.
- En regiones de DNA con mayor %GC es más probable encontrar genes.



Improvement in the Accuracy of Gene Prediction in Human cDNA Sequences
Todokoro et al.

Genome Informatics 14: 454–455 (2003)

Anotación

Codon usage bias

- Recordando a Watson & Crick

The Genetic Code

The number of possible combinations of bases (A, T, C, G) taken two at a time. 16 combinations are possible.

AA	AT	AC	AG
TA	TT	TC	TG
CA	CT	CC	CG
GA	GT	GC	GG

Taken three at a time, 64 cominations are possible, which is enough to characterize the 22 amino acids plus a "stop".

TTT	phe	TCT	ser	TAT	tyr	TGT	cys
TTC	phe	TCC	ser	TAC	tyr	TGC	cys
TTA	leu	TCA	ser	TAA	stop	TGA	stop
TTG	leu	TCG	ser	TAG	stop	TGG	try
CTT	leu	CCT	pro	CAT	his	CGT	arg
CTC	leu	CCC	pro	CAC	his	CGC	arg
CTA	leu	CCA	pro	CAA	gln	CGA	arg
CTG	leu	CCG	pro	CAG	gln	CGG	arg
ATT	ile	ACT	thr	AAT	asn	AGT	ser
ATC	ile	ACC	thr	AAC	asn	AGC	ser
ATA	ile	ACA	thr	AAA	lys	AGA	arg
ATG	met	ACG	thr	AAG	lys	AGG	arg
GTT	val	GCT	ala	GAT	asp	GGT	gly
GTC	val	GCC	ala	GAC	asp	GGC	gly
GTA	val	GCA	ala	GAA	glu	GGA	gly
GTG	val	GCG	ala	GAG	glu	GGG	gly

Ala/A	GCU, GCC, GCA, GCG
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG
Asn/N	AAU, AAC
Asp/D	GAU, GAC
Cys/C	UGU, UGC
Gln/Q	CAA, CAG
Glu/E	GAA, GAG
Gly/G	GGU, GGC, GGA, GGG
His/H	CAU, CAC
Ile/I	AUU, AUC, AUA
Leu/L	UUA, UUG, CUU, CUC, CUA, CUG
Lys/K	AAA, AAG
Met/M	AUG
Phe/F	UUU, UUC
Pro/P	CCU, CCC, CCA, CCG
Ser/S	UCU, UCC, UCA, UCG, AGU, AGC
Thr/T	ACU, ACC, ACA, ACG
Trp/W	UGG
Tyr/Y	UAU, UAC
Val/V	GUU, GUC, GUA, GUG
START	AUG
STOP	UAA, UGA, UAG

Por ejemplo, la leucina se puede codificar de 6 formas distintas

Anotación

Codon usage bias

- Asumiendo uniformidad se pensaría que los codones que codifican un mismo aminoácido deberían aparecer con la misma frecuencia.
  - Si en un organismo se encuentra leucina, entonces la probabilidad de que tal o cual codón la haya codificado es

	1/6	1/6	1/6	1/6	1/6	1/6
Leu	UUA	UUG	CUU	CUC	CUA	CUG

- Esto en realidad no ocurre, hay organismos que prefieren determinados codones para codificar ciertas proteínas.

- Un ejemplo de esta idea
  - Estilo de codificación de leucina de cierto organismo hipotético.

	0.02	0.02	0.9	0.02	0.02	0.02
Leu	UUA	UUG	CUU	CUC	CUA	CUG

Anotación

Codon usage bias

- Este sesgo ayuda a predecir sitios de interés en secuencias de DNA.
- Métodos para predecir nivel de expresión de genes
  - Frequency of Optimal Codons
  - Codon Adaptation Index

Table 4. Performance of Several TIS Prediction Systems

Data Set	Met.	Sen	Spe	AA	OA
vert.	FirstAUG	64.31%	88.40%	76.36%	82.49%
	[19]	82.25%	87.80%	85.02%	86.44%
	[6]	0.24%	90.25%	45.24%	68.17%
	[20]	80.00%	58.11%	69.06%	63.57%
	CUB	**89.58%**	**96.61%**	**93.10%**	**94.89%**
Arab.	FirstAUG	72.85%	90.69%	81.77%	86.13%
	[19]	**97.32%**	88.79%	93.06%	90.97%
	[6]	0.57%	89.31%	44.94%	66.65%
	[20]	24.47%	76.66%	50.56%	63.33%
	CUB	91.78%	**97.18%**	**94.48%**	**95.80%**
TIS+50	FirstAUG	74.00%	97.04%	85.52%	94.68%
	[19]	**88.00%**	69.93%	78.97%	71.78%
	[6]	64.00%	**98.41%**	81.20%	94.89%
	[20]	85.71%	54.34%	70.03%	56.18%
	CUB	80.00%	97.72%	**88.86%**	**95.91%**

Effectiveness of Applying Codon Usage Bias for Translational Initiation Sites Prediction

Zeng et al.

doi:10.1109/BIBM.2008.30

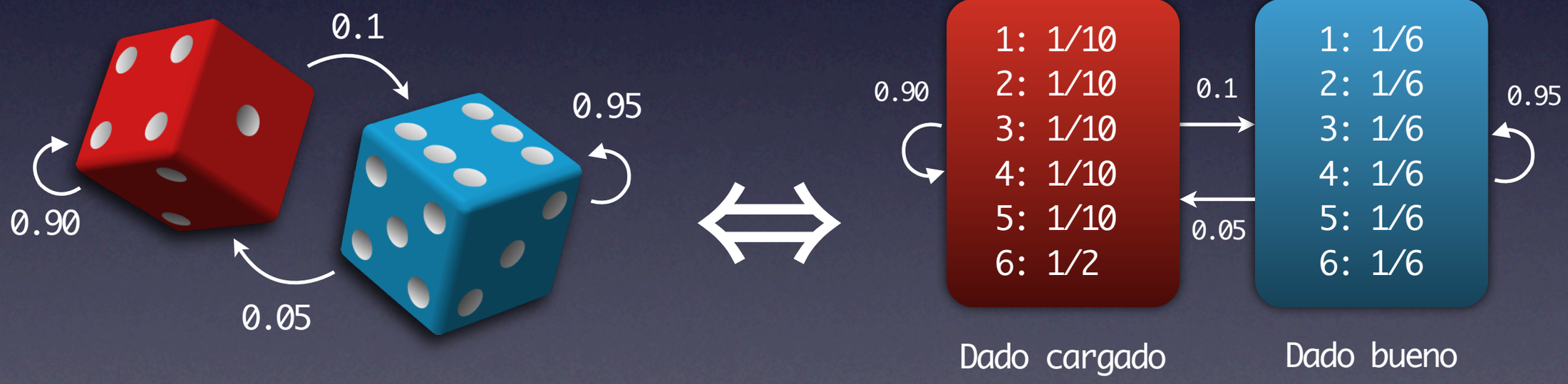
Anotación

Buscando genes mediante Modelos de Markov

- Para detectar motifs conocidos (promotor y otros).
- Para modelar los estados "dentro de un gen" y "fuera de un gen".
- Noción de cadenas de Markov

  Componentes

  - Set de estados
  - Distribución de probabilidad sobre los estados
  - Matriz de transición

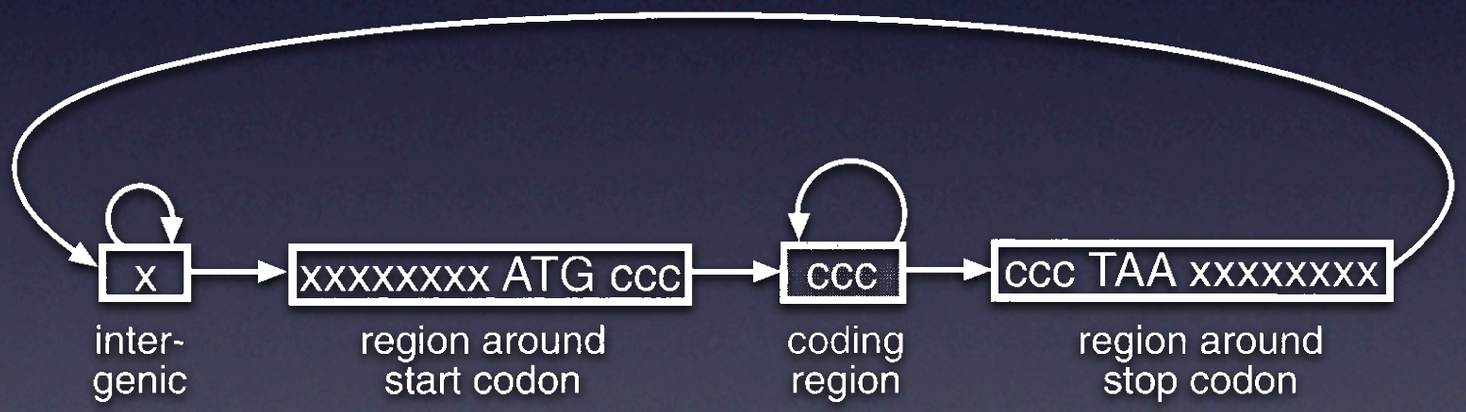


- ¿Cuál es la secuencia de lanzamientos más probable para obtener 312453666641 ?
  BBBBBBCCCCCC

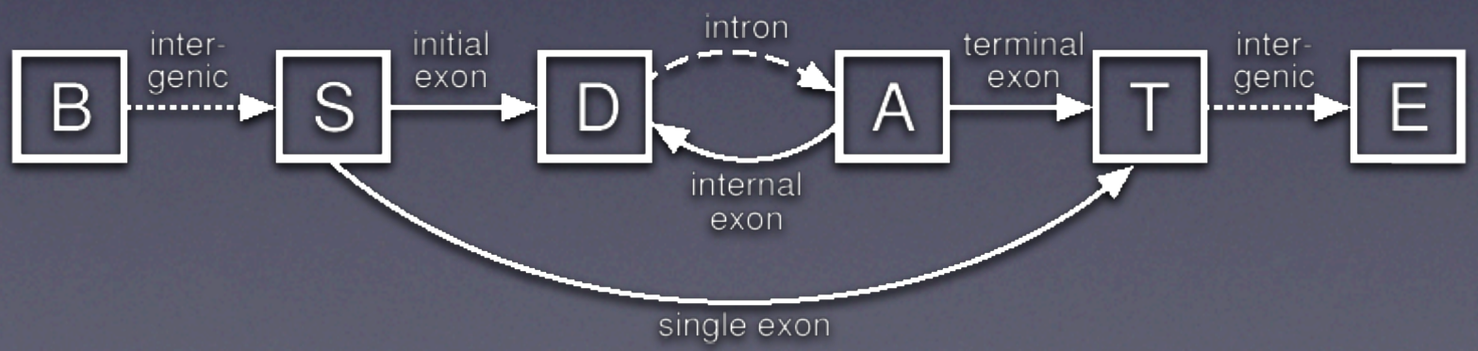
Anotación

Buscando genes mediante Modelos de Markov

- El modelo propuesto debe ser biológicamente consistente
- Se debe entrenar el modelo con datos conocidos para determinar los parámetros (entrenamiento por genoma)
- Con el modelo listo...
  - Leer secuencias de DNA y encontrar los genes más parecidos a lo que el modelo "conoce".

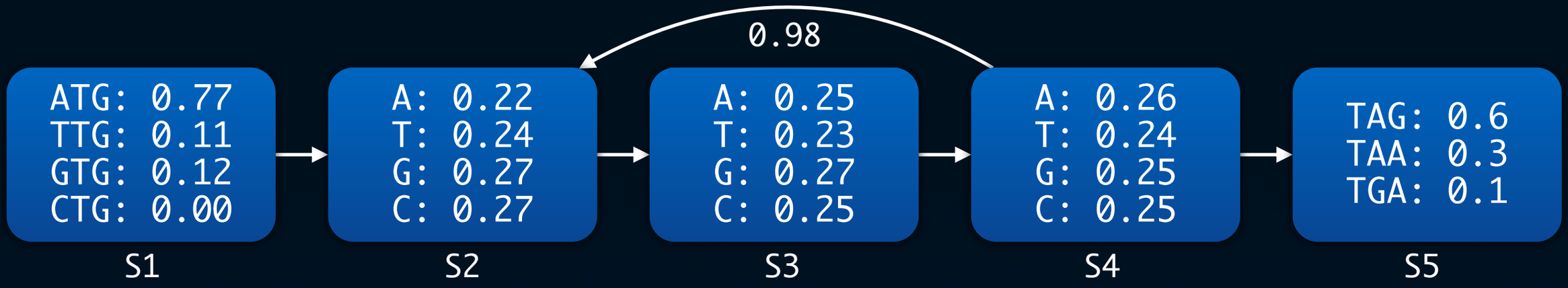


- Modelo simple sin Intrones

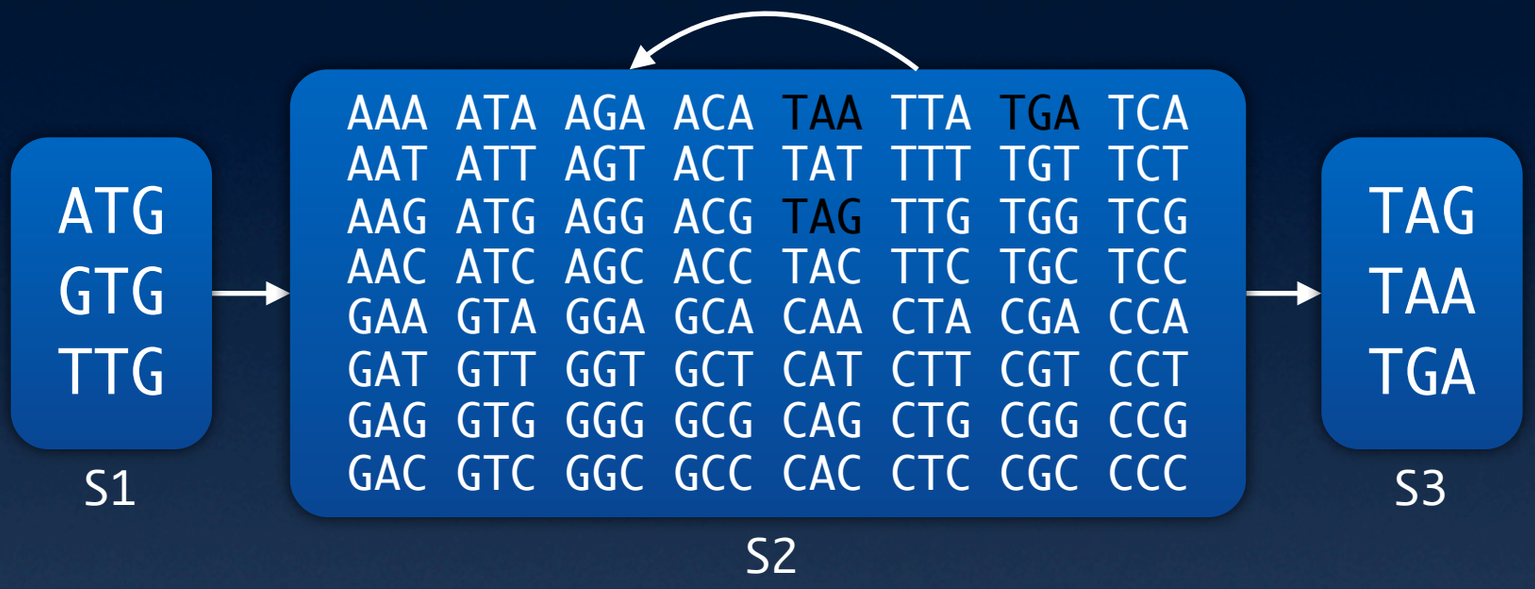


- Modelo simple con intrones exones y señales

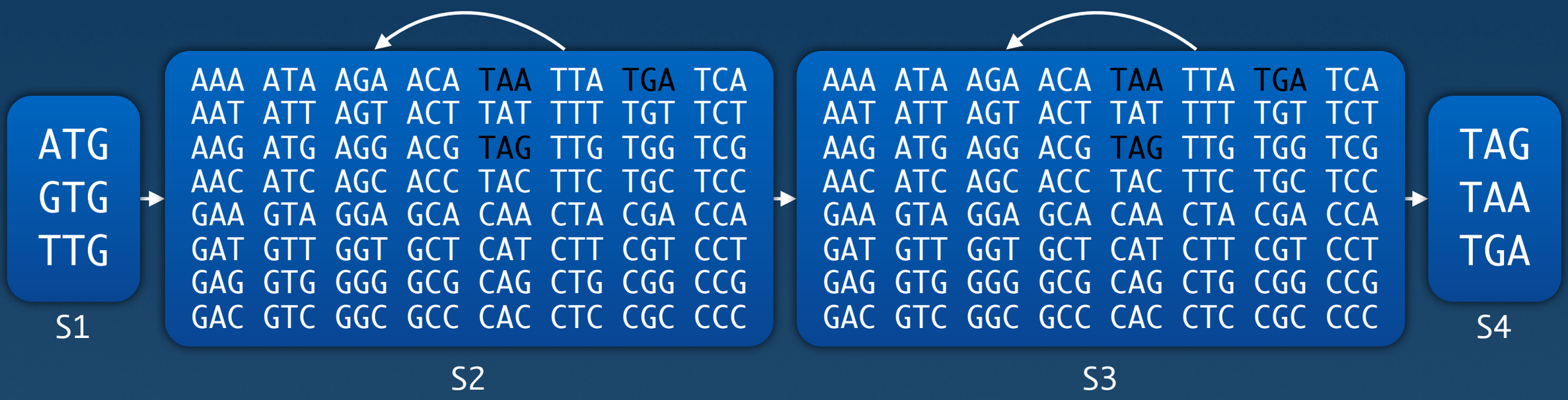
- Modelo simple para procariotas



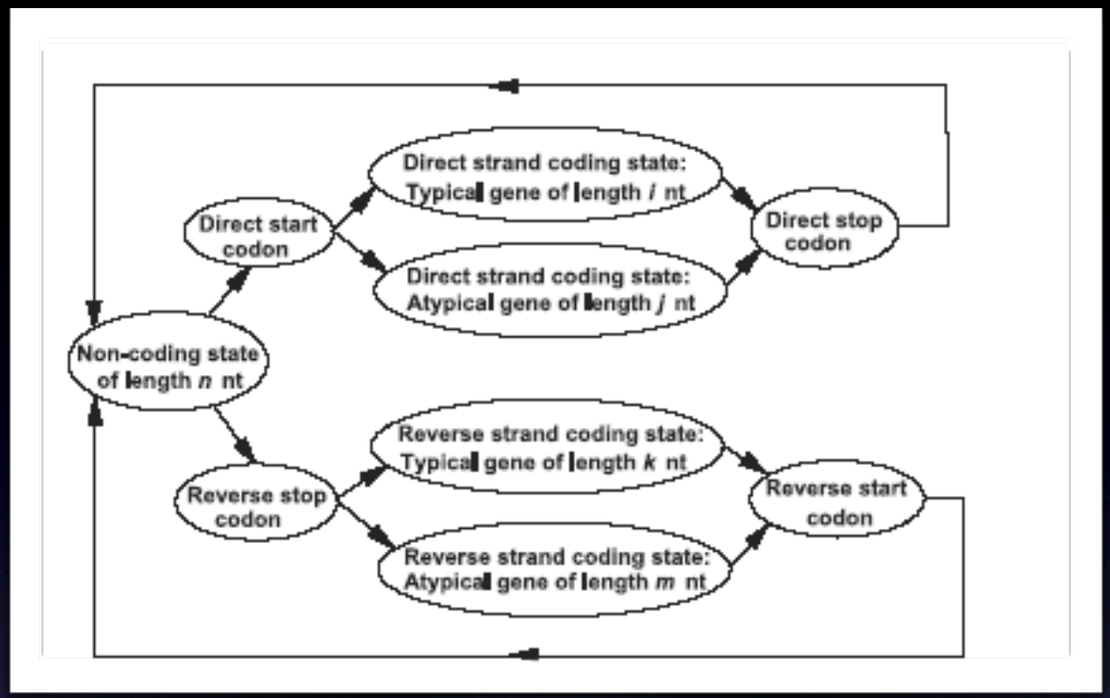
- Considerando uso de codones

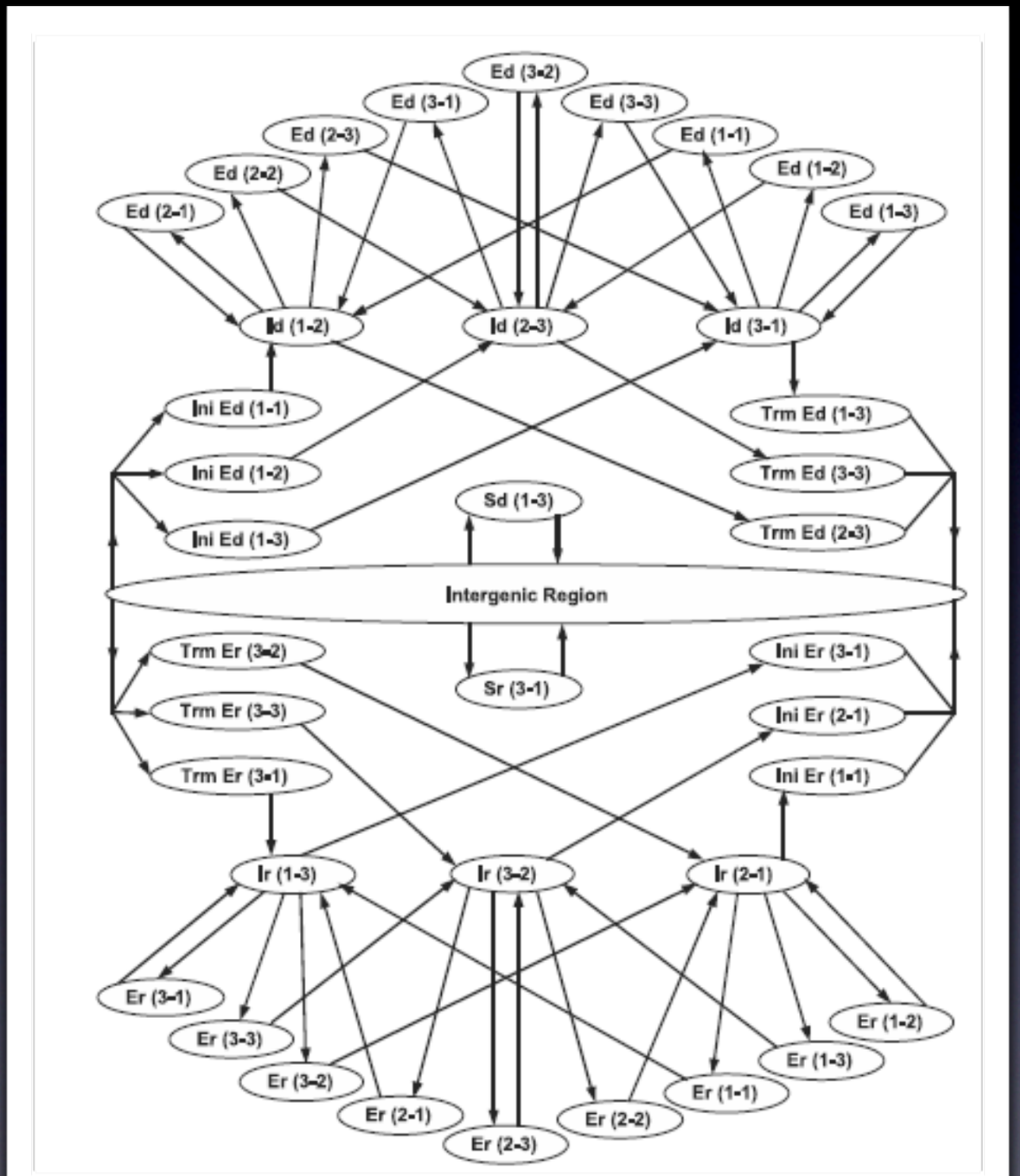


- Considerando dependencia entre codones vecinos



Anotación





GeneMark.HMM

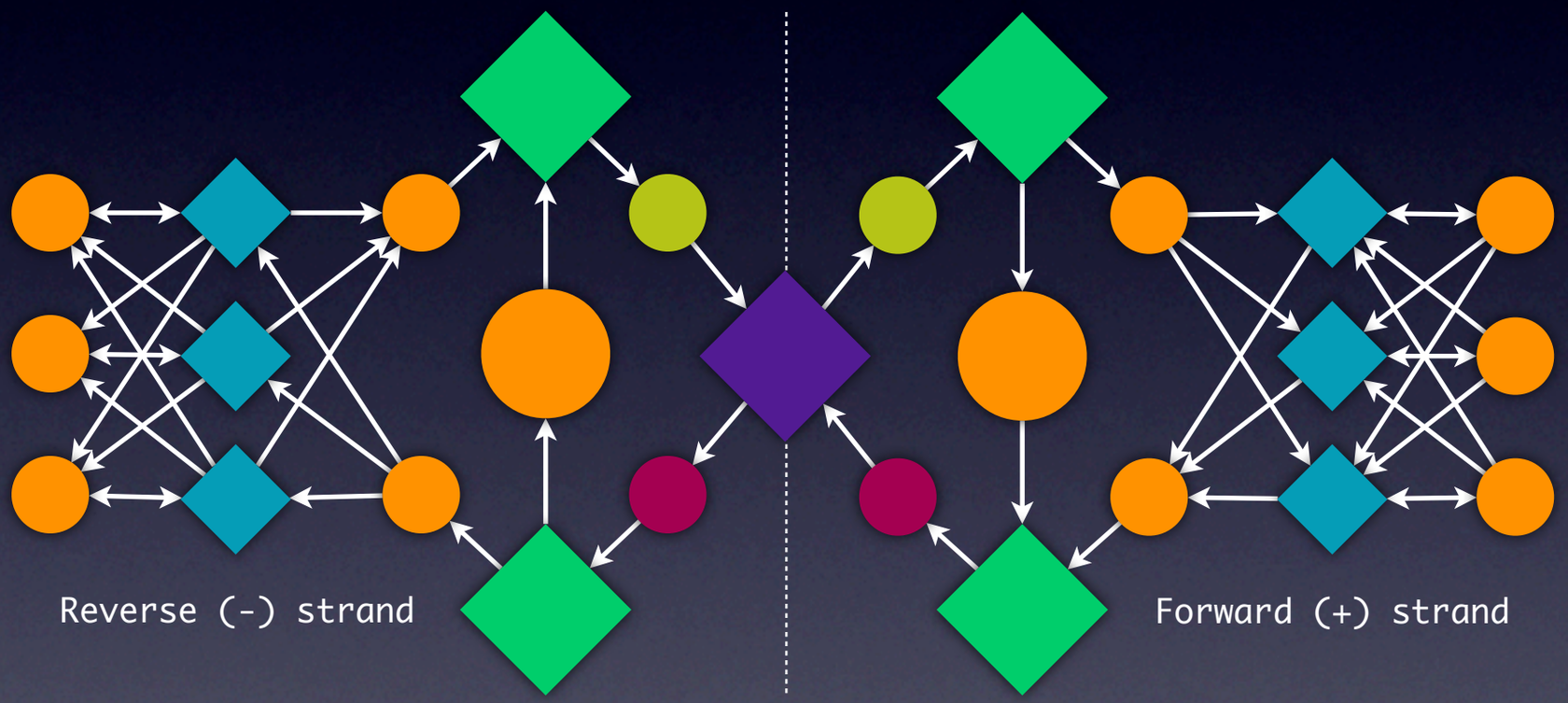
- Estrategia para Procariotas
- Estrategia para Eucariotas

¡La estrategia para procariotas en eucariotas falla!

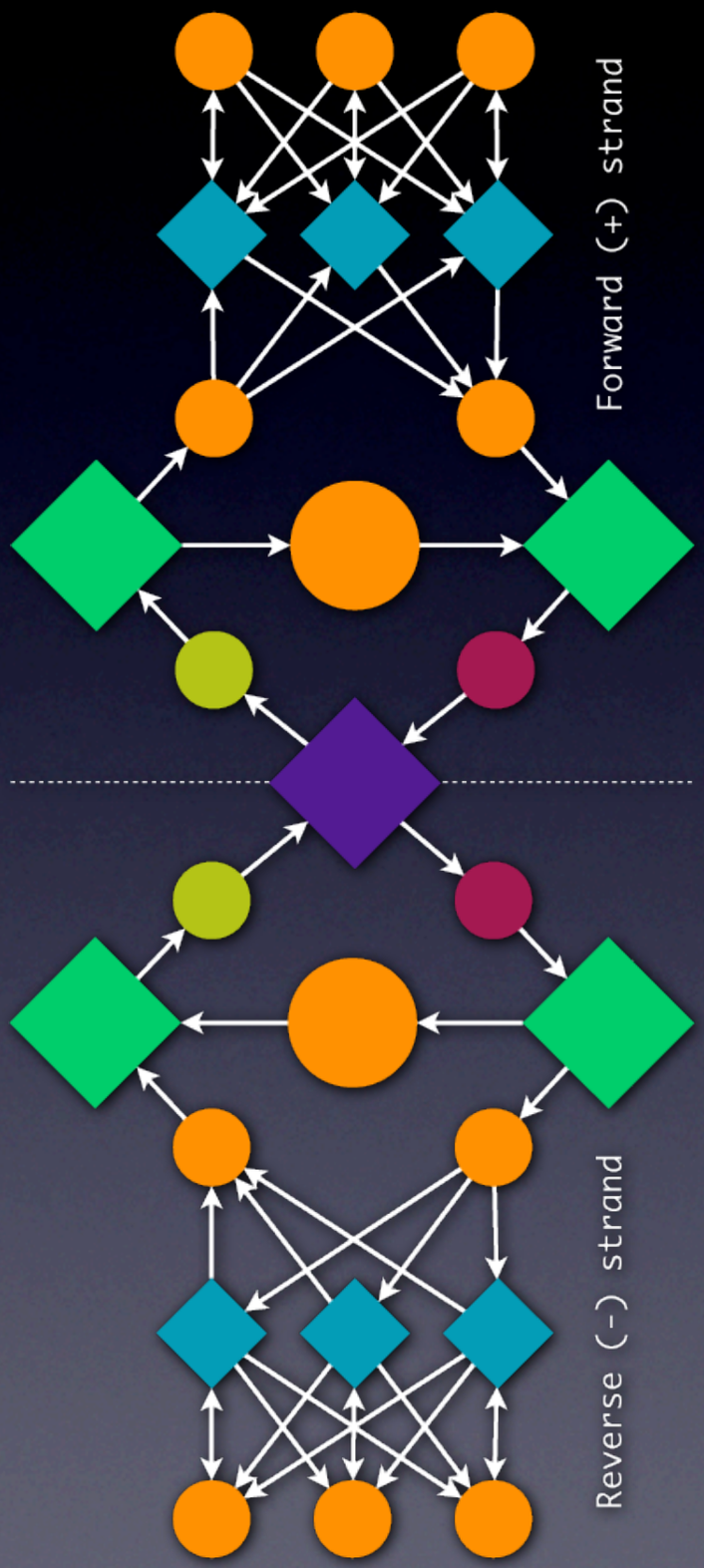
Anotación

GHMM, G for Generalized (Base de GenScan)

- Forma general de describir secuencias.
- Cada nodo emite secuencias de largo variable.



Anotación




```
62001 AGGACAGGTA CGGCTGTCAT CACTTAGACC TCACCCTGTG GAGCCACACC
62051 CTAGGGTTGG CCAATCTACT CCCAGGAGCA GGGAGGGCAG GAGCCAGGGC
62101 TGGGCATAAA AGTCAGGGCA GAGCCATCTA TTGCTTACAT TTGCTTCTGA
62151 CACAACTGTG TTCACTAGCA ACCTCAAACA GACACCATGG TGCACCTGAC
62201 TCCTGAGGAG AAGTCTGCCG TTACTGCCCT GTGGGGCAAG GTGAACGTGG
62251 ATGAAGTTGG TGGTGAGGCC CTGGGCAGGT 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 TTTACACAGT
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
```

Demo

GGTAATTAAAGCTAAATGAAGTTGTAAGAGTGGCCCTATCGCATAGGACTAGTGTCCCTATAAGAACACGAAGAAATCACCTTAGAAAGGCTGAGAAAGG
GAGTGCAGACTGAAAGATGCAGACCACTGGGCTTCTACTTCTGTTTCCATTTCTGATCCGGCCTGCATCTGCCTCCTTCCTGAACAGGCCAGAGAATTCA
CAGGCTGGGTGCTGTGGCTCACCTGTAATCCCAACACTTGGGAGGCCGAGGTGGGCAGATCACCTGAGGTCAGGAGTTCAAGGCTAGCCTAGCCAACATG
TACTAAAAAAATACAAAAATTAGCCAGGCATAGTGGCGCCTATAGTTCCAGCTACTTGGGGGCTGAGGTAGGAAGATCGCTAGAGCCTGGGAGGTTAAGG
ATTGTGCCACTGCACTCCAGCCTGGGTGACAGAGCAAGACCCTGCCTCAAAAATAAATAAATAAATAAATAAATAAAAATAAGAGTGCTTGGCAGCTTGA
AACCCATCTCTCAAGCAGCAGCTCTTCTCCTGTGCCATTGTCAGCTTTGTCCTGTCTGAGTCCATGGGACTCTTCTGTTTGATGGTGGTCTTCCTCATCC
GCTCCATGGAGATCACCTACCCATACCTGCTTCTGTGACCTCATGCCATTCCTGGTGTTGGAATGTGCCAAGGTTTGCCATTAAACACACATTTCTCATT
TATTATATATATGTGTGTGTGTGTGTGTTTATATATGCGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTATATATATATATATATATATATATATATAT
ATATAGGAAGAGGCACCAGAGAGCTCTCTGCATAGTCACAGAGGAAAGGTCATGTGAGGACAGCCAGAAGGCAGATGTCACAAGCCTCACCAGCAACCTA
ACCTCCATCCTGGAAGTCCAGCTTCTAAAATTAGAAGAAAATAGTCGGGTGTGGTGGCTCGCACCTATAATCCCAGCACTTTGGGAGGCTGATGTGGGAG
AGAGTTTGAAACCAGCCTAGGCAACATAGGGAGACCCTGTCTTTAAAAAAAATTTTTTTTTGTTTTAATTAGCTGGGTGTGATGGTGCACACCTGAGTCC
CTGAGGTAGGAGGATCCCCTGAGCCCAGGGAAGTGGAGGCTGCAGTGAGCCATGATCACACCACTGCAATACAGCCTGGGTGACAGAGCAAGACCTTATC
AAACAAAAAAGATGACAAAATAAATGTCTGTCGTTTAAGTCACCCATTCTGTGATATCTTGTTACGGCAGCCTGAACTGACCAATACACTTCCTCACCCA
TCAATCATAATCAGCCATTGCAATTACCCTCAACTGTATTATCAACCCTCAATTTGTATTAGTTGCTTGGCAAAACCCAAACCCTTGTGAAATCCAGTTC
GATGCTGCCGAATATGGCTGAAGAAAAGCAACTGTGTTGACTGGACTGCTTTAAATTCATGACCACTTACCTCAAGTGGGCACTTAACTTCCTGGCAATT
AGTCCATTAACTCTCCTCCTCTCTGAGTTAATTATTTCACAGCTTTTCCTCCCTCTTTATACATGTTCCATCCTAACTCTCTGCTGATGACCTTGTTTCT
GGCCACCAGGAGAGAACTCCCACAGCCATCAAATTCACCAAGCCAACAGCATCCTTACACAAATCCTCTGCCTTCTCTCTGGGCTGGCTGTGCCCTCTCT
CCCTAACTCTCCTATACTGTTGTTATTCACTCTCCAGTGGATAATCACCATCAGGATGCAAAGATGCTGTACTAGCTTCTGAACTCTCCAAAAACCCAGG
GGCTAAGCTTTTTCTTATTCCCCCTTCCAGCTATTGTACTGTTTCTCTGCTTTTAATTTATTTTTATTTATTTATTTATTTATTTATTTATTTTTTG
CTTGTTGCCCAGGCTGGAGCGCAATGGCGCGATCTCAGCTCACCGCAACCTCTACTTCCCGAATTCAAGTGATTGTCCTGCCTCAGCCTCCCGAGTAGCC
CGCCACCACGCCTGGCTAATTTTGTACTTTTAGTAGAGACGGGGTTTCTCCATGTTGCTCAGCCTGGTCACAAACTCCCGATCTCAGGTGATCTGCCTGC
TGCTGGGATTACAGGCGTGAGCCACCACGCCCCACCGTCTCTGTTCTCTTTTAAAGCACAATCCCTCAACACAAGTGTCTATACTCAGCGTCTCCACTTT
TCCCAGTGCCCCCTTGTCAGGTTTTCACCCCATGCTCCTCCAGGGCTAGTCTGCTCTTGCTTCCCGTCTTACTGGAAGACCAGCAGCATTTGACAGAGTT
CTTGGACACCTTTTCTTCACTTGGTTTCCAGAACAGCATTATCTCCTGCTTATTGTCTTCCTCAGTCTACCTCAGTGAAAAGCTTTACTGGTTCCTCCAC
AGTAATAACAGGAATGTACCATGCCATTGCTCTCTCTCTCTCCTTTTTTTTTTTTTTTTTTTTTTTTTTTGTTGAGACAGAGTCTCAATTTTATCACCCAGA
ATGATCATAGCTCATTGCAGTCTCGAACTCGTGGGCTCAAGCAATCCTCCCACCTCAGCCTCCTGAATAGCTGGGACTACAAGCAACACCACCATGCCCA
TTTATTTTTATTTTTTGTAGAGATGAGGTTTTACTATGTTGCCTAGGCTAGTCTTGAACTCCTGGGCCCAAATGATCCTCCCACCTTGGTCTCCCAAAGT
GTGAGCCACCGTGTCCAACTTCTCTTTCTTAATGGAATTTAGGCAAAAGTTATTACTCATGGCCTTGGAATGCTCTTTCCTCAGATAGCCACATGGCTCA
AGCTTTCTTCAAAGATCCACTTCTCAGTGAAGCTTTGTCCTGACCACCCAGCTGAAAATTGCAATCCTCTTCTGTCTACCATGTACATACTCTCTATTTG
TTCTCTCTGTAGGTGTGACCTAACATAACATATAATTTACTTCTGTACCTTGTTTGCTTTCTGTCTTCCCCTTTAGAACATAAGCTCCATGAGGGAAGGC
AGTCACTTTATCTCCAGCAACTACAACTATATGTATATATACACACACATATATATACACACACATATATATACACACACATATATATATACATATATAT
AATAAACATTCACTGAATGAATGAACAGTAATGCTCACTTGCCCATAAATACAAGTACCTCATCTTTTACCACAAAGGGTATTTGTAAATATTTAGGTTG
TGGCTTGGTAATTCTTTTTTTTTTTTTTCTAATTTTTATTTTTTTTCTAGGGACAGGGTCTCACTATGTTGCCCAGGATGGTCTTGAACTCCTGGGCTCAA
TGGCCTCCTAAAGTGCTGAGATTACAGGCATGAGCCACCGTGCCTGCCTTCATGTATGTTTTTAGAACACAGAGAAAATGTGTTCTAAATGTGCTCATTG
AGGCTTATGCAGTCACCACCAATCAAAAACTTTTTTTTTTTTTTTTTGAGACAAGATCTTGCTCTGTTGCCCAGGCTGGAGTGCAGTGGCAGGATCATAGC
CCTCATAGGCCTAAATCATCCTCCCACCTCAGCCTCACAAGTAGCTAAGACCACAGGTACAAGCCACCGTATCTAGCTAACTTTCAAAATTTTTTGAATT
TTTGAGGCCAGGCTGGCCTCAAACTCCTGAGCTCAAGCAATCCTCCCACCTTGGCTTCCCAAAGTGCTGGGATTATAGGCGTGAGCAACTGTACCTGGCA
CTTCGCTTCCAGGATTAGGCAACTTTAACCTTCAACAGTGATCATAACCCTTAGTTTTCAGATCCGATTAAGGGAAATGTGTAATGTCTTACTGACACAC

GGTAATTAAAGCTAAATGAAGTTGTAAGAGTGGCCCTATCGCATAGGACTAGTGTCCCTATAAGAACACGAAGAAATCACCTTAGAAAGGCTGAGAAAGG
GAGTGCAGACTGAAAGATGCAGACCACTGGGCTTCTACTTCTGTTTCCATTTCTGATCCGGCCTGCATCTGCCTCCTTCCTGAACAGGCCAGAGAATTCA
CAGGCTGGGTGCTGTGGCTCACCTGTAATCCCAACACTTGGGAGGCCGAGGTGGGCAGATCACCTGAGGTCAGGAGTTCAAGGCTAGCCTAGCCAACATG
TACTAAAAAAATACAAAAATTAGCCAGGCATAGTGGCGCCTATAGTTCCAGCTACTTGGGGGCTGAGGTAGGAAGATCGCTAGAGCCTGGGAGGTTAAGG
ATTGTGCCACTGCACTCCAGCCTGGGTGACAGAGCAAGACCCTGCCTCAAAAATAAATAAATAAATAAATAAATAAAAATAAGAGTGCTTGGCAGCTTGA
AACCCATCTCTCAAGCAGCAGCTCTTCTCCTGTGCCATTGTCAGCTTTGTCCTGTCTGAGTCCATGGGACTCTTCTGTTTGATGGTGGTCTTCCTCATCC
GCTCCATGGAGATCACCTACCCATACCTGCTTCTGTGACCTCATGCCATTCCTGGTGTTGGAATGTGCCAAGGTTTGCCATTAAACACACATTTCTCATT
TATTATATATATGTGTGTGTGTGTGTGTTTATATATGCGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTATATATATATATATATATATATATATATAT
ATATAGGAAGAGGCACCAGAGAGCTCTCTGCATAGTCACAGAGGAAAGGTCATGTGAGGACAGCCAGAAGGCAGATGTCACAAGCCTCACCAGCAACCTA
ACCTCCATCCTGGAAGTCCAGCTTCTAAAATTAGAAGAAAATAGTCGGGTGTGGTGGCTCGCACCTATAATCCCAGCACTTTGGGAGGCTGATGTGGGAG
AGAGTTTGAAACCAGCCTAGGCAACATAGGGAGACCCTGTCTTTAAAAAAAATTTTTTTTTGTTTTAATTAGCTGGGTGTGATGGTGCACACCTGAGTCC
CTGAGGTAGGAGGATCCCCTGAGCCCAGGGAAGTGGAGGCTGCAGTGAGCCATGATCACACCACTGCAATACAGCCTGGGTGACAGAGCAAGACCTTATC
AAACAAAAAAGATGACAAAATAAATGTCTGTCGTTTAAGTCACCCATTCTGTGATATCTTGTTACGGCAGCCTGAACTGACCAATACACTTCCTCACCCA
TCAATCATAATCAGCCATTGCAATTACCCTCAACTGTATTATCAACCCTCAATTTGTATTAGTTGCTTGGCAAAACCCAAACCCTTGTGAAATCCAGTTC
GATGCTGCCGAATATGGCTGAAGAAAAGCAACTGTGTTGACTGGACTGCTTTAAATTCATGACCACTTACCTCAAGTGGGCACTTAACTTCCTGGCAATT
AGTCCATTAACTCTCCTCCTCTCTGAGTTAATTATTTCACAGCTTTTCCTCCCTCTTTATACATGTTCCATCCTAACTCTCTGCTGATGACCTTGTTTCT
GGCCACCAGGAGAGAACTCCCACAGCCATCAAATTCACCAAGCCAACAGCATCCTTACACAAATCCTCTGCCTTCTCTCTGGGCTGGCTGTGCCCTCTCT
CCCTAACTCTCCTATACTGTTGTTATTCACTCTCCAGTGGATAATCACCATCAGGATGCAAAGATGCTGTACTAGCTTCTGAACTCTCCAAAAACCCAGG
GGCTAAGCTTTTTCTTATTCCCCCTTCCAGCTATTGTACTGTTTCTCTGCTTTTAATTTATTTTTATTTATTTATTTATTTATTTATTTATTTTTG
CTTGTTGCCCAGGCTGGAGCGCAATGGCGCGATCTCAGCTCACCGCAACCTCTACTTCCCGAATTCAAGTGATTGTCCTGCCTCAGCCTCCCGAGTAGCC
CGCCACCACGCCTGGCTAATTTTGTACTTTTAGTAGAGACGGGGTTTCTCCATGTTGCTCAGCCTGGTCACAAACTCCCGATCTCAGGTGATCTGCCTGC
TGCTGGGATTACAGGCGTGAGCCACCACGCCCCACCGTCTCTGTTCTCTTTTAAAGCACAATCCCTCAACACAAGTGTCTATACTCAGCGTCTCCACTTT
TCCCAGTGCCCCCTTGTCAGGTTTTCACCCCATGCTCCTCCAGGGCTAGTCTGCTCTTGCTTCCCGTCTTACTGGAAGACCAGCAGCATTTGACAGAGTT
CTTGGACACCTTTTCTTCACTTGGTTTCCAGAACAGCATTATCTCCTGCTTATTGTCTTCCTCAGTCTACCTCAGTGAAAAGCTTTACTGGTTCCTCCAC
AGTAATAACAGGAATGTACCATGCCATTGCTCTCTCTCTCTCCTTTTTTTTTTTTTTTTTTTTTTTTTTGTTGAGACAGAGTCTCAATTTTATCACCCAGA
ATGATCATAGCTCATTGCAGTCTCGAACTCGTGGGCTCAAGCAATCCTCCCACCTCAGCCTCCTGAATAGCTGGGACTACAAGCAACACCACCATGCCCA
TTTATTTTTATTTTTTGTAGAGATGAGGTTTTACTATGTTGCCTAGGCTAGTCTTGAACTCCTGGGCCCAAATGATCCTCCCACCTTGGTCTCCCAAAGT
GTGAGCCACCGTGTCCAACTTCTCTTTCTTAATGGAATTTAGGCAAAAGTTATTACTCATGGCCTTGGAATGCTCTTTCCTCAGATAGCCACATGGCTCA
AGCTTTCTTCAAAGATCCACTTCTCAGTGAAGCTTTGTCCTGACCACCCAGCTGAAAATTGCAATCCTCTTCTGTCTACCATGTACATACTCTCTATTTG
TTCTCTCTGTAGGTGTGACCTAACATAACATATAATTTACTTCTGTACCTTGTTTGCTTTCTGTCTTCCCCTTTAGAACATAAGCTCCATGAGGGAAGGC
AGTCACTTTATCTCCAGCAACTACAACTATATGTATATATACACACACATATATATACACACACATATATATACACACACATATATATATACATATATAT
AATAAACATTCACTGAATGAATGAACAGTAATGCTCACTTGCCCATAAATACAAGTACCTCATCTTTTACCACAAAGGGTATTTGTAAATATTTAGGTTG
TGGCTTGGTAATTCTTTTTTTTTTTTTTCTAATTTTTATTTTTTTTCTAGGGACAGGGTCTCACTATGTTGCCCAGGATGGTCTTGAACTCCTGGGCTCAA
TGGCCTCCTAAAGTGCTGAGATTACAGGCATGAGCCACCGTGCCTGCCTTCATGTATGTTTTTAGAACACAGAGAAAATGTGTTCTAAATGTGCTCATTG
AGGCTTATGCAGTCACCACCAATCAAAAACTTTTTTTTTTTTTTTTTGAGACAAGATCTTGCTCTGTTGCCCAGGCTGGAGTGCAGTGGCAGGATCATAGC
CCTCATAGGCCTAAATCATCCTCCCACCTCAGCCTCACAAGTAGCTAAGACCACAGGTACAAGCCACCGTATCTAGCTAACTTTCAAAATTTTTTGAATT
TTTGAGGCCAGGCTGGCCTCAAACTCCTGAGCTCAAGCAATCCTCCCACCTTGGCTTCCCAAAGTGCTGGGATTATAGGCGTGAGCAACTGTACCTGGCA
CTTCGCTTCCAGGATTAGGCAACTTTAACCTTCAACAGTGATCATAACCCTTAGTTTTCAGATCCGATTAAGGGAAATGTGTAATGTCTTACTGACACAC

Anotación

HMMs varios:

- *Se ve que el problema no es trivial*
- *Lo ideal sería poder reconocer áreas de interés con tanta fidelidad como lo hace la propia célula*
- *Existen otros enfoques*
  - *Redes Neuronales*
  - *Gibbs sampling*
- *GenMark es el más popular para anotar bacterias*
- *GenScan es muy utilizado.*
  - *Está entrenado con genomas ya secuenciados*
- *Lo existente no es perfecto: se necesita entrenamiento para organismos específicos.*
- *Área de permantente investigación*

Anotación

Buscando genes mediante Gramáticas

- Los algoritmos de análisis de secuencias tratan al DNA, RNA y a las proteínas como strings de nucleótidos o aminoácidos.

- La mayoría de estos algoritmos asume strings de elementos sin relación, donde el valor de un residuo en una posición no tiene efecto sobre el valor de otro residuo.

Esta suposición se rompe dramáticamente para el RNA

Anotación

¿La razón?

- La estructura secundaria del RNA pone restricciones sobre la secuencia del RNA. Por lo que no pueden modelar eficientemente utilizando HMM.
- Es necesario adoptar nuevos modelos que consideren las correlaciones a larga distancia entre pares de residuos (situación que los algoritmos anteriores no manejaban).
- Por esta razón, se utilizarán gramáticas y lenguajes formales.

Anotación

Buscando genes mediante Gramáticas

- Repasando algunas cosas de TALF

- Las gramáticas son un recurso bastante útil para modelar strings con símbolos correlacionados (como es el caso del RNA)
- Una gramática caracteriza un *lenguaje*
- Una gramática consiste de
  - **N:** Un conjunto de símbolos *no terminales*
  - **V:** Un conjunto de símbolos *terminales*
    (son los que realmente aparecen en el string)
  - **S:** Un símbolo no terminal de *start S*
  - **P:** Un conjunto de *producciones*

Anotación

Buscando genes mediante Gramáticas

- Un ejemplo para contextualizar

Lenguaje {UAA, UAG, UGA} (codones de stop)

N: {S, c_1, c_2, c_3, c_4}

S: s

V: {A, C, G, U}

P: $s \rightarrow c_1$ $c_1 \rightarrow Uc_1$ $c_2 \rightarrow Ac_1$ $c_3 \rightarrow A$

$c_2 \rightarrow Gc_4$ $c_3 \rightarrow G$

$c_4 \rightarrow A$

Anotación

Buscando genes mediante Gramáticas

- Jerarquía de Chomsky
  - La jerarquía de Chomsky es una clasificación de gramáticas, donde el grado de complejidad va en aumento.

Anotación

Buscando genes mediante Gramáticas

- Jerarquía de Chomsky
  - Gramáticas Regulares

$$u \rightarrow Xv \qquad u \rightarrow X$$

  - Gramáticas Libres de Contexto

$$u \rightarrow \beta$$

  - Gramáticas Sensibles al Contexto

$$\alpha_1 u \alpha_2 \rightarrow \alpha_1 \beta \alpha_2$$

  - Gramáticas Irrestrictas

$$\alpha_1 u \alpha_2 \rightarrow \gamma$$

Donde u y v son no terminales, X es un terminal, α y γ son cualquier secuencia de terminales / no terminales, excluyendo el string nulo, y β es cualquier secuencia de terminales / no terminales (incluyendo el string nulo).

Anotación

Buscando genes mediante Gramáticas

- Gramáticas probabilísticas
  - En donde cada producción tiene una probabilidad de ser aplicada (la probabilidad de cada producción suma 1).

P: $S \xrightarrow{1.0} c_1$ $\quad c_1 \xrightarrow{1.0} Uc_1$ $\quad c_2 \xrightarrow{0.7} Ac_1$ $\quad c_2 \xrightarrow{0.3} Gc_4$ $\quad c_3 \xrightarrow{0.2} A$ $\quad c_3 \xrightarrow{0.8} G$ $\quad c_4 \xrightarrow{1.0} A$

Anotación

Buscando genes mediante Gramáticas

- Palíndromos en RNA
  - No son los palíndromos que se conocen comúnmente.
  - Los palíndromos en el RNA se consideran en basea los complementos de los ácidos nucleicos.

AGAUUUCGAAAUCU

- Estos palíndromos tienen un largo variable por lo que deben ser modeladas por gramáticas libres de contexto o gramáticas más complejas.

$$S \rightarrow AsU \mid UsA \mid CsG \mid GsC \mid \varnothing$$

- El *parse tree* de la gramática representa la estructura primaria y secundaria del RNA.

Anotación

Buscando genes mediante Gramáticas

- Una CFG para RNA

Emisión de Pares

```
u → AvU | UvA | CvG | GvC
u → GvU | GvA (para pares no comunes)
```

Emisión de una sola letra

```
u → Av | Cv | Gv | Uv (izquierda)
u → vA | vC | vG | vU (der) (derecha)
```

Emisión de letras terminales

```
u → A | C | G | U
```

Bifurcaciones

```
u → vw
```

Saltos o eliminaciones

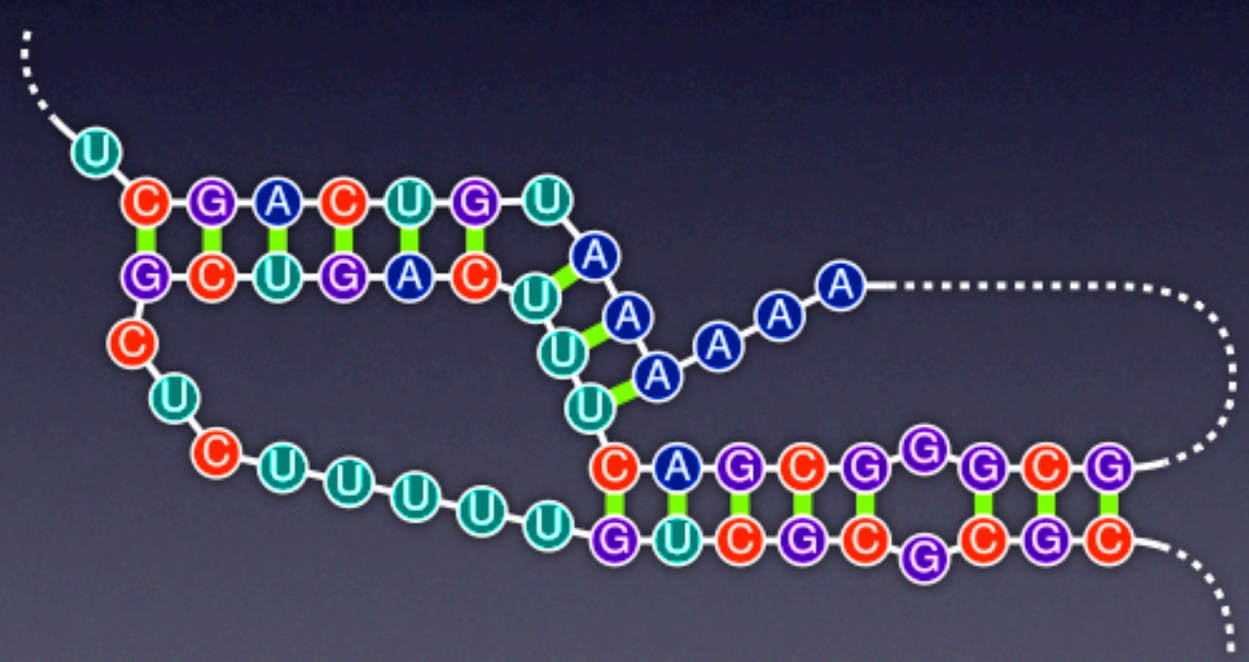
```
u → v
```

Se puede ver que a diferencia de las HMM, las CFG pueden emitir simultáneamente pares de letras

Anotación

Buscando genes mediante Gramáticas

- Sin embargo, la CFG mostrada en las slides anteriores no es aplicable a secuencias de RNA que posean "Pseudoknots".
- Un "Pseudoknot" ocurre cuando los nucleotidos de un loop en la cadena de RNA se enlazan con nucleotidos que se encuentran en otras partes de la secuencia.



- Por ello es necesario para estas situaciones, utilizar gramáticas sensibles al contexto.

Anotación

Buscando genes mediante Gramáticas

- Por ello es necesario para estas situaciones, utilizar gramáticas sensibles al contexto.
- PERO!, No se conocen algoritmos generales en tiempo polinomial para parsear gramáticas sensibles al contexto. Por lo que se transforma en una verdadera complicación resolver alguno de los problemas básicos. Por esta razón es necesario utilizar SCFG (Stochastic Context Free Grammar)

SCFG ~ HMM

Anotación

Buscando genes mediante Gramáticas

- Algunos alcances sobre las SCFG
  - La derivación de versiones más generales de SCFG, se realiza mediante métodos bastante similares al algoritmo de "forward propagation" para HMMs ("inside algorithm").
  - Para determinar el parse tree más probable de la gramática, se utiliza una versión de programación dinámica similar al algoritmo Viterbi para HMMs.
  - Las SCFG son entrenadas para determinar las probabilidades de las producciones. Posterior a esto se utilizan de manera similar a las HMMs.
  - Para el entrenamiento de las gramaticas se pueden utilizar los siguientes algoritmos de aprendizaje:
    - **The EM Algorithm**
    - **Gradient Descent and Viterbi Learning**

Anotación

Buscando genes mediante Gramáticas

- Los tres problemas básicos para SCFG
  - Problema de asignación de puntaje: ¿Cuan probable es una secuencia dado un SCFG parametrizado? – **Algoritmo Inside**
  - Problema de entrenamiento: Dado un conjunto de secuencias, ¿Cómo estimamos los parámetros de un SCFG? – **Algoritmo EM**
  - Problema de alineamiento: : ¿Cual es el parsing mas probable de una secuencia a un SCFG parametrizado? – **Algoritmo CYK**

También hay algoritmos que resuelven estos problemas con HMMs

¿Cuál es la equivalencia?

Anotación

Equivalencia entre HMM y SCFG

- Para la familia de problemas en predicción de genes se dan las siguientes equivalencias entre los algoritmos de los respectivos métodos.

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

Tarea

Tarea

Prediciendo genes con GeneMark.HMM

- Conseguir el Genoma de su procariota favorito de la tarea 1.

  http://cmr.jcvi.org/

  - Si no está disponible o no lo encuentra, usar algún otro genoma de su elección, por ejemplo el genoma de *Methanococcus jannaschii* de la tarea 2.
- Predecir genes que codifican proteínas en un trozo aleatorio de 100 kbp del genoma mediante GeneMark.HMM

  http://exon.biology.gatech.edu/hmmchoice.html

  - Utilice la versión para procariotas o eucariotas según corresponda
  - Ajuste el parámetro "Species" a su organismo
  - Seleccione la opción de traducir los genes encontrados en proteínas
- Elegir una de las secuencias de aminoácidos resultantes y hacer un ProteinBlast de esta.
  - El organismo donde está la proteína más cercana a la consulta ¿es cercano al del genoma original? Si son distintos ¿hay diferencias en la función de la proteína en los respectivos organismos?