¿Qué es un gen?

'Gene' is not a typical four-letter word. It is not offensive. It is never bleeped out of TV shows. And where the meaning of most four-letter words is all too clear, that of gene is not. The more expert scientists become in molecular genetics, the less easy it is to be sure about what, if anything, a gene actually is.

Helen Pearson: Genetics: What is a gene? (*Nature* 441, 398-401)

¿Qué es un gen?

Concepto mendeliano:
Unidad discreta de herencia (factor) que expresa un fenotipo dado
En 1909, W. Johanson acuño el termino ‘gen’

Morgan y CIA
Los genes se encuentran en loci específicos en los cromosomas

G.W. Beadle and E.L. Tatum
Confirmaron en 1941 la hipótesis de un gen una enzima postulado en 1908 por E.R. Garrod.
Los genes se heredan y pueden sufrir mutaciones.

S. Benzer (1957)

Acuño diferentes términos que finalmente daban al dogma central de la biología molecular:

gen (DNA) --> transcrito (RNA) --> proteína

La definición de un gen

A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions.

Helen Pearson: Genetics: What is a gene? (*Nature* 441, 398-401)

Genes RNA

- tRNA, snRNA, rRNA, snoRNA, lncRNA, vault RNA, piRNA, microRNA …

Funciones: Ribonucleoproteína, biogénesis de otros RNAs, traducción, regulación génica

Genes codificantes de proteínas

- **Procariotas:**
  - Sin intrones
  - Regulación más simple

- **Eucariotas**
  - Los genes contienen intrones,
  - Regulación compleja

La tarea de buscar los genes codificantes

El genoma humano tiene 3,200,000,000 pares de bases pero solo entre 20,000 y 25,000 genes codificantes que ocupan una proporción muy pequeña del genoma (aprox. 1,5%)

La tarea de buscar los genes codificantes

El genoma humano tiene 3,200,000,000 pares de bases pero solo entre 20,000 y 25,000 genes codificantes que ocupan una proporción muy pequeña del genoma (aprox. 1,5%)

¿Como podemos encontrar las islitas pequeñas que son los genes dentro del océano genómico?

La tarea de buscar los genes codificantes

El punto de partida: una secuencia de DNA

```
GGCCTACCATTTTCTACAGATGGCATTTATTCAGCAAGGGAATGTTGTAGGTATAGTAATTTATCTTTTTTCCATGAAACATGTGTT
ATCTTCTTGCTCTATTTTGAAGAGATATTATGATAAATGATGCCACTATTTAAAACTTGTAAAGCTTCTTGAGGTTCTCAGAAAAAA
GTACAAGGTATAGGTCAAAAAGAGTAATTTTGTCTCTCACAAAAGAATTTTCTGTTCAAAAAATGCTGTTAACTTGAAAAAGTCATA
TCTCCAAAAGAACGTTCTAATAGGTGGGATAATCATCATCATATTTCCAGCTATAATCCATCATATTTCCAGGTGTATCCCTTCCCT
TAAAAGTAAAAAAGGAGAAAATCACTTTGCTGTGTGGAAGTGCTTTGTGTGTTTTATAAAATCCACATAAAATGTTCGTGGTTCAAG
GAGTGGCTGCCGCATTTCTAAGGCAAATAAGCTTATCCCTGAAGAAGCTTGAAGATCTGACTCTCAAGAATCTTTGTGTTCTGGATG
AAAAAGGAATCATTATTTTTTAACTCCCCTGCTCAAGTCAGTGACTTCCACTGGCACCCAATCCATTGGTTATAAGTACCCATGAAT
TGCCTTGCCAATTATAGGCAGGATAACCTAGGAGGACCAACCAAAACCAAACTCCTGGTAAGCTTTATTAGCCTTTTCTACAATGTG
TCTCAGTTTCTCTTTCACACCTTGGAGTGGAAGACAATCTCTACTCACACTGTTGTGGTATGTGTCTGTCTGTCTCAGATGTGTGTA
TAAGGTGTATTTCTCAGCTCGGTATCTGAAGTAATTTCTTTTTCTTTTCTTTTTTTTAAAGACTGCTGTCTTAAGGCTTAAAAGTAA
TTTAAGGTTTTGTTTTTTGTTATTTTAAGAGTTGCAGTTTCTTTGGAGACAACAAAGTACTTTTTTTCCAAAGATGAAAAAGGCATA
TTCAGACTTTCTATGTTATGTTCCATAATTGTTTTGCTTATCACTGTATAACATTCTCATTATCATGGGCTACTTTAAGCTAATCAG
TTGCATTTTGGGGTCCTCATTTAGAAGTCAATCTGGAAAAAAGATGGCTGTAAAACTTAGATAAACCAGTATGCTATTTCTCCTTCC
TTGTCTGCCTTTACAGATGTTTCCTCTTTCTCTCTCTCTTCCCTGACTATCTCTCTCAATGATCATTGACAGCTCCCGCATCCCCTC
TTTCACTCAGCTTCACTCCACCATGACCAGAGCACCTCTCCTGCTACTATGTGTTGCCCTGGTGCTGCTTGGGCATGTGAATGGAGC
CACAGTAAGAAATGAGGACAAATGGAAGCCACTCAACAACCCCAGAAACAGAGATCTGGTAAGAACACACTTGGGGCCTGGGTTTCA
GGATAGAACTAATGCCTCTTTGGATTTACGCACACAGCATTCTGTTTTCCAGAACGTAGTGGAAAGGGGGCTACCAAGGGTTGAGAA
AATTCTCCAACAGGATTGGACCAGAGTTTTACTACCCTAGTGGTTGGCAACCTATTGTTTCTCTGGCCAGTTTACATTTGTTTCTGT
AAAGAGAAACAAATTTATATTTAGAAACAATCTTATACACACATAGTACAATTCATTTGTAACTGATCTTATATACACATTGACTTT
GTAGCCTTGATTTCTAATACATATAAAAAATAAAGAGTGTAACAGGGAATTGCTCGTGCATAGCTTAGATTGAACTGATGATAAAAC
AAGTTATCATATAACTAGAGATTGCCAAGCTCAGCTGCAAAAATTGAAAGTTGACCCTGCATGGGTATCAATTTACTGGACATATTT
TCAACATATAGTTATCTATAAAATAGAATTCAAAAGATCAGTAAACATTCTCTTTAAAAAAAATGCCAAATACAATAAACATGTTGA
CATGACTATTTTGAAATATGGTTTTTCTAAAGTGAATGATGGGAAATATAACAACATAGGATTAAAATAAACTGAGATAAGAGACCA
GCAGGATTTGTTTTCTGGTCACAGCTGCTGACGAAAACAGGATCTGGTC
```

La tarea de buscar los genes codificantes

Diferentes aproximaciones

Búsqueda *ab-initio* (a partir de primeros principios)

Buscar signos o señales indicadores de genes codificantes de proteínas (promotores, composición de secuencia, existencia de una ORF (**O**pen **R**eading **F**rame, Marco de lectura abierto))

Aproximación extrínseca

- Se basa en evidencia externa, es decir se busca en el genoma objetivo secuencias similares a la de una secuencia externa
- Estas secuencias pueden provenir de experimentos de secuenciación del RNA total de la misma especie o de genes conocidos en otras especies

Las predicciones deben de ser confirmado y validado por experimentos en el laboratorio

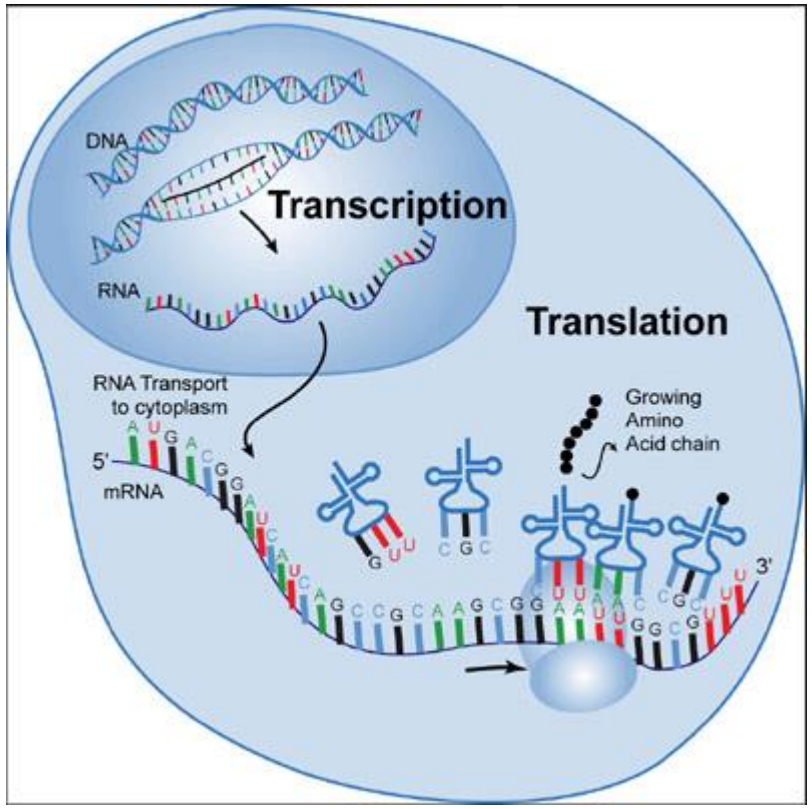
Confirmar que realmente se transcriben y traducen

Ab-Initio

¿Que propiedades de un gen codificante podemos buscar en una secuencia de DNA?

Ab-Initio

¿Que propiedades de un gen codificante podemos buscar en una secuencia de DNA?



Transcripción mediante polimerasa II
DNA --> RNA

Traducción en los ribosomas mediante RNAs de transferencia (tRNA/ARNt)

La traducción comienza en un codón de inicio y termina en el primer codón de parada que aparece en este marco de lectura

¡La secuencia del mRNA (ARNm) contiene un marco de lectura abierto!

Marco de lectura abierto (ORF)

La lectura de la información genética se lleva a cabo de triplete (codón) en triplete

- Un marco abierto de lectura empieza con un codón de inicio y termina con un codón de parada que se encuentran en la misma pauta de lectura
- No puede haber codones de parada en la pauta de lectura

Búsqueda de ORFs

Buscamos primero en la hebra directa en los 3 marcos posibles

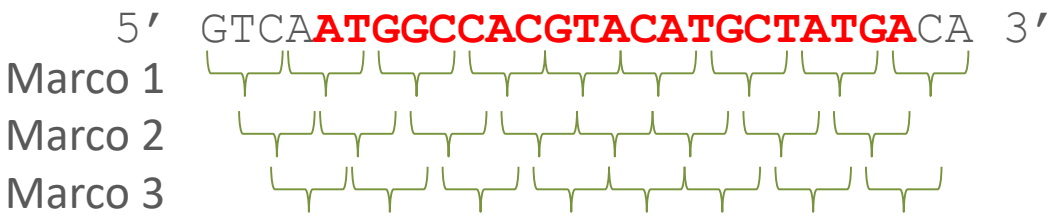
```
     5' GTCAATGGCCACGTACATGCTATGACA 3'
Marco 1 └┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘
Marco 2  └┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘
Marco 3   └┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘
```

Pero también tenemos que buscar en la secuencia inversa complementaria

```
     5' TGTCATAGCATGTACGTGGCCATTGAC 3'
Marco -1 └┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘
Marco -2  └┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘
Marco -3   └┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘└┬┘
```

Búsqueda de ORFs

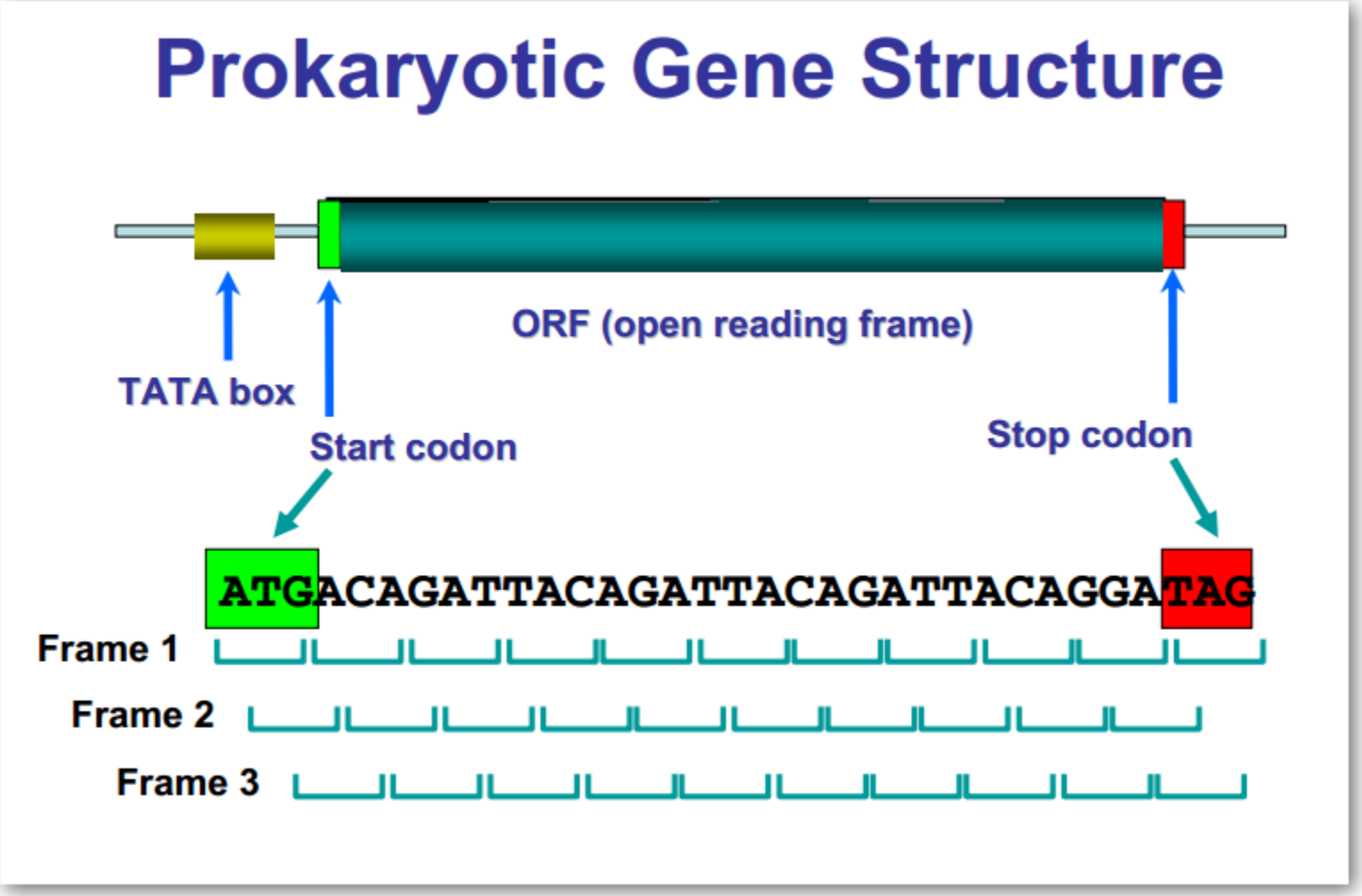
Buscamos primero en la hebra directa en los 3 marcos posibles



Hay un marco abierto de lectura en el segundo marco

ORFs y genes en procariotas

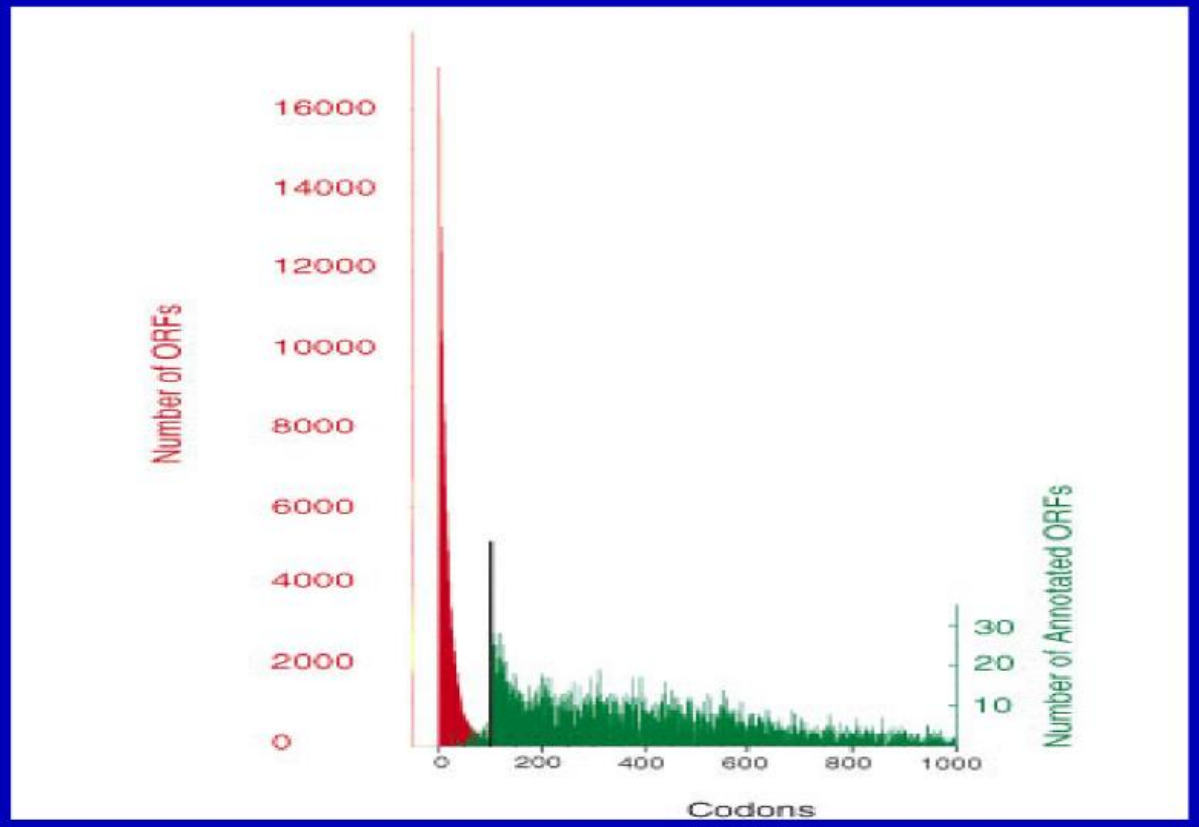
En procariotas el ORF está presente en la secuencia de DNA por la ausencia de intrones



ORF – CDS

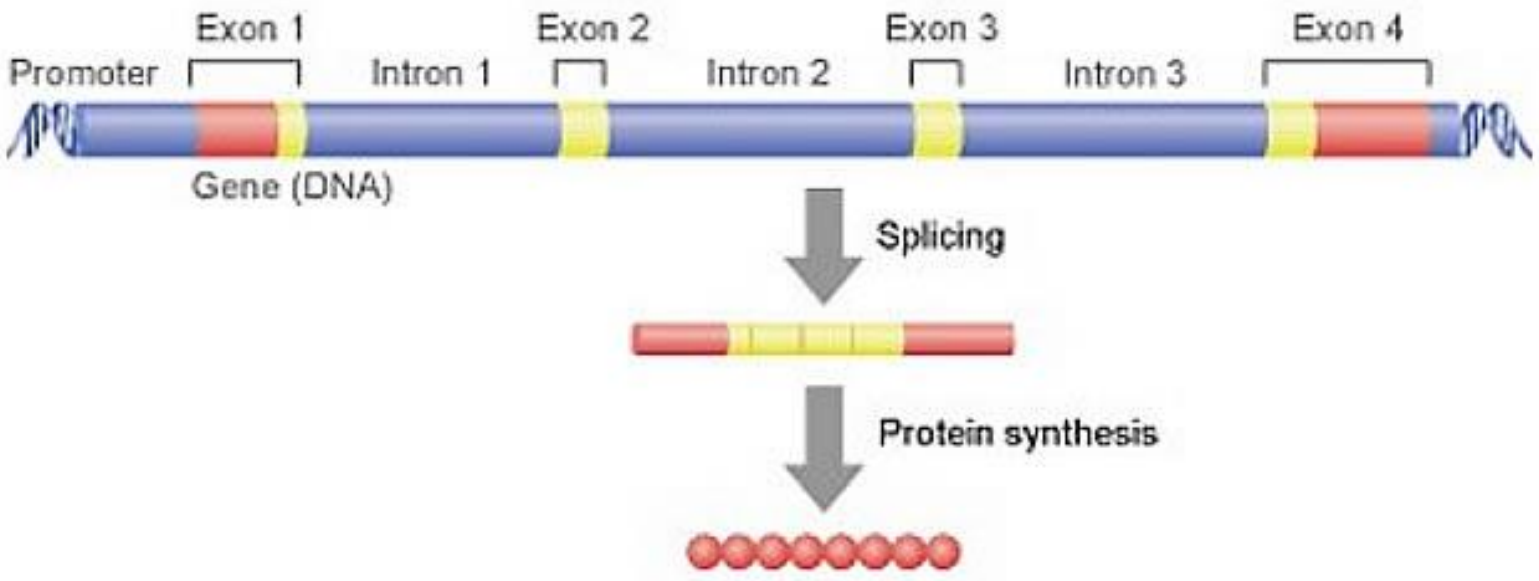
Que es la diferencia entre ORF y CDS?

Una medida simple: comparación de las longitudes de ORFs anotadas y espúreas en *S. cerevisiae*



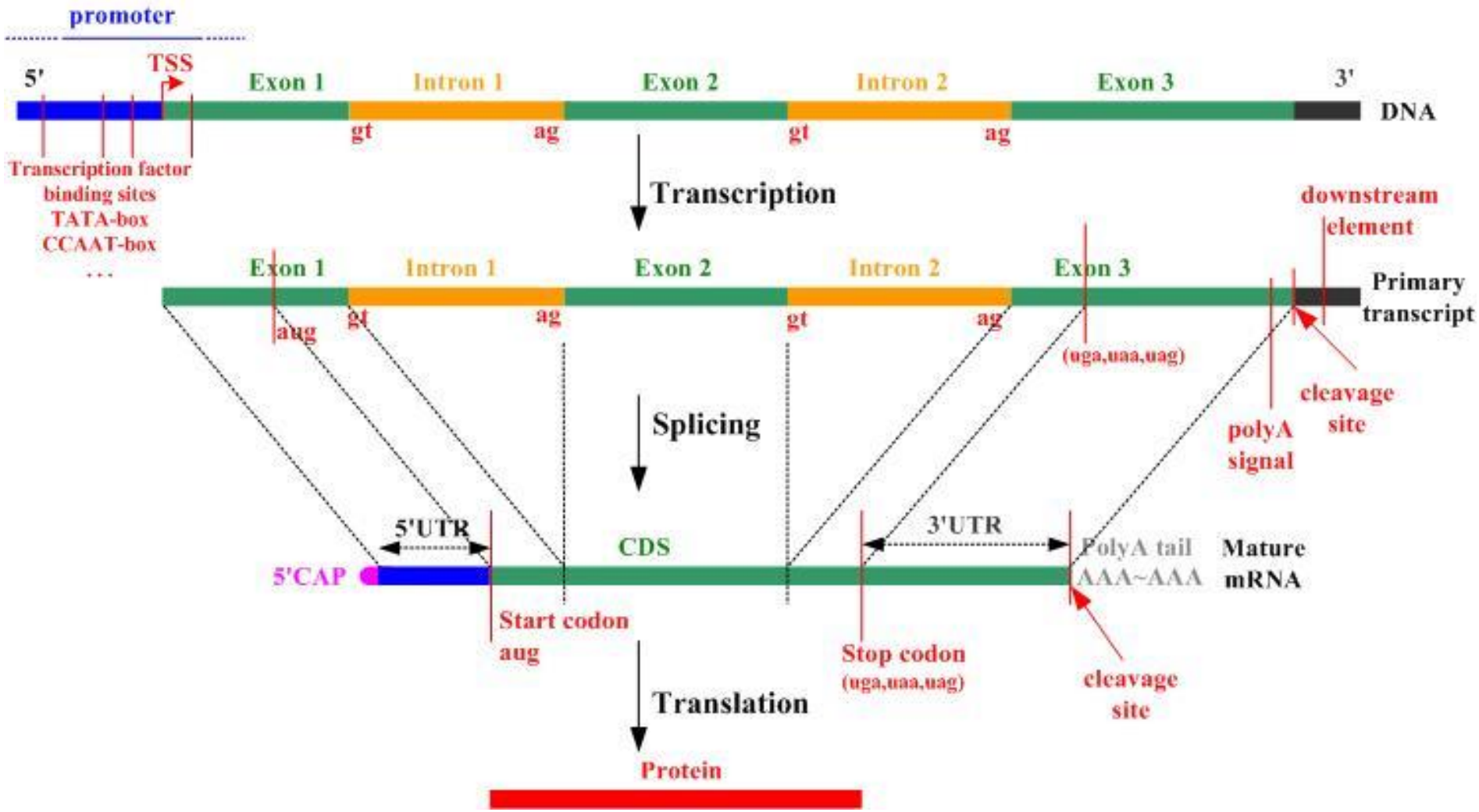
Basrai MA, Hieter P, and Boeke J Genome Research 1997 7:768-771

Predicción de genes en eucariotas



¡En genes con intrones, el ORF no se presenta en la secuencia de DNA!

Detectar los genes



Predicción de genes en eucariotas

¿Como podemos buscar/predecir genes que contienen intrones?

1) Potencial codificador de los cdsExons

- Los exones tienen otra composición de secuencia que las regiones no-exónicas (frecuencias de hexámeros).
- Esta información puede ser usado por métodos de aprendizaje automatizada

2) Señales como donadores y aceptores

- Los 2 primeros nucleótidos de un intron son en la inmensa mayoría de los casos 'GT' (Donador)
- Los 2 últimos nucleótidos de un intron son en la inmensa mayoría de los casos 'AG' (Aceptor)

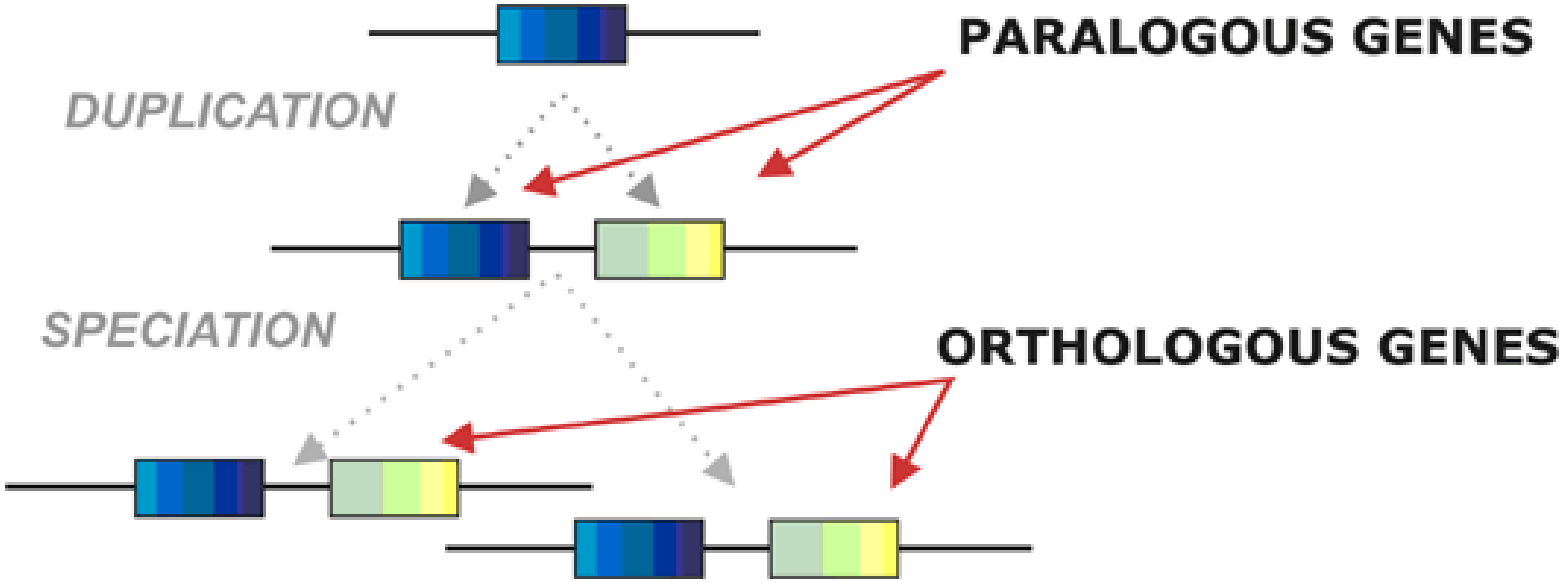
3) Promotores (islas CpG, composición de secuencia)

- La búsqueda de genes se puede refinar mediante la detección de regiones o motifs asociados al promotor (TFBS, islas CpG)

Usar información extrínseca

La idea:

Alinear secuencias de mRNAs conocidos (de otras especies o obtenidas en estudios de expresión génica) frente al genoma objetivo



http://biol.lf1.cuni.cz/ucebnice/en/glossary.htm

Usar información extrínseca

La idea:

Alinear secuencias de mRNAs conocidos (de otras especies o obtenidas en estudios de expresión génica) frente al genoma objetivo

Por ejemplo:

- **Tenemos una secuencia mRNA de humano y buscamos si esta secuencia también se encuentra en el chimpancé**
- **Si encontramos la secuencia y si además observamos que no hay mutaciones que cambien el ORF, podemos concluir que es probable que este gen es funcional también en el chimpancé**

Pero:

¡La prueba definitiva que el posible gen se transcribe y traduce tiene que venir por parte de experimentos de laboratorio (Northern Blot, Western Blot)!

Alineamiento: similitud de secuencia

$$\text{Similitud de secuencia} = \frac{\text{Número de bases emparejadas}}{\text{Número de bases alineadas}}$$

```
ACGTATAGCG
|||||   |||
ACGTA--GCG
```

Ungapped: 100%

Gapped: 80%

```
ACGTATAGCG
|||||   |
ACGTAGCG
```

75%

Usar información extrínseca

- El BLAT de ADN está diseñado para encontrar de forma rápida una secuencia en un genoma dado
- Optimizado para las secuencias con mayor similitud de 95% y mayor longitud que 25 bp → probablemente falla detectar alineamientos mas divergentes o cortos

Kent WJ.
BLAT--the BLAST-like alignment tool.
Genome Res. 2002

BLAT

```
>panTro4_refGene_NM_001252546
GGGTGGGAGCGCGTGCTGTTGGGAGTTGCTTGGAGGTTGGCGGCGCGGGG
CTGAAGGCTAGCAAACCGAGCGATCATGTCGCACAAACAAATTTACTATT
CGGACAAATACGACGACGAGGAGTTTGAGTATCGACATGTCATGCTGCCC
AAGGACATAGCCAAGCTGGTCCCTAAAACCCATCTGATGTCTGAATCTGA
ATGGAGGAATCTTGGCGTTCAGCAGAGTCAGGGATGGGTCCATTATATGA
TCCATGAACCAGAACCTCACATCTTGCTGTTCCGGCGCCCACTACCCAAG
AAACCAAAGAAATGAAGCTGGCAAGCTACTTTTCAGCCTCAAGCTTTACA
CAGCTGTCCTTACTTCCTAACATCTTTCTGATAACATTATTATGTTGCCT
TCTTGTTTCTCACTTTGATATTTAAAAGATGTTCAATACACTGTTTGAAT
GTGCTGGTAACTGCTTTGCTTCTTGAGTAGAGCCACCACCACCATAGCCC
AGCCAGATGAGTGCTCTGTGGACCCACAGCCTCAGCTGAGTGTGACCCCA
G
```

Alinear frente al genoma mediante BLAT

BLAT

Longitud y coordenadas de alineamiento de la secuencia de entrada

Región en el genoma
Cromosoma, hebra , inicio, final

Score/puntuación

Similitud de secuencia

Longitud de la región en el cromosoma

BLAT Search Results

ACTIONS		QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser	details	panTro4_refGene_NM_001252546	547	1	551	551	99.9%	1	+	154947147	154951489	4343
browser	details	panTro4_refGene_NM_001252546	513	13	551	551	97.6%	14	+	96689382	96689920	539
browser	details	panTro4_refGene_NM_001252546	503	1	551	551	95.3%	11	+	82920366	82920915	550
browser	details	panTro4_refGene_NM_001252546	502	32	551	551	98.3%	8	-	81556932	81557451	520
browser	details	panTro4_refGene_NM_001252546	502	14	551	551	96.3%	X	+	30635509	30636045	537
browser	details	panTro4_refGene_NM_001252546	470	76	551	551	99.4%	5	-	61807834	61808309	476
browser	details	panTro4_refGene_NM_001252546	468	76	551	551	99.2%	10	+	29986864	29987339	476
browser	details	panTro4_refGene_NM_001252546	388	13	551	551	91.3%	5	+	143818706	143819276	571
browser	details	panTro4_refGene_NM_001252546	93	14	130	551	89.8%	19	-	9115818	9115934	117
browser	details	panTro4_refGene_NM_001252546	26	492	534	551	71.5%	2	+	7182785	7182816	32
browser	details	panTro4_refGene_NM_001252546	24	452	480	551	76.0%	4	-	24220330	24220354	25
browser	details	panTro4_refGene_NM_001252546	24	292	316	551	100.0%	1	-	62669892	62669936	45
browser	details	panTro4_refGene_NM_001252546	23	485	507	551	100.0%	11	+	907550	907572	23
browser	details	panTro4_refGene_NM_001252546	20	396	415	551	100.0%	3	-	155686371	155686390	20

BLAT: Resuelve la estructura génica

Dos exones de NM_001252546 (chimpancé) alineados en el genoma humano (bases en azul)

```
gtgctccgtg gggctcccat tatgttgtag ataatcaact agaagactat  154950196
aaggtccatg aaggcggaga aaatgtgcct tgaacttaag tgctcattca  154950246
ctggaattat tcactaattt gtatttttaa tatactctcg gaatttatat  154950296
aacatagcct gtatttctag ctgctgctac cccacaataa aattatataa  154950346
actctgacat caaaaaccag tttttctttc ctttcttggc cttgtaaaca  154950396
ggcatttgtc taagagactg tatctggtac taacataaat tccccacttc  154950446
cccgtttctg ttacagACAT GTCATGCTGC CCAAGGACAT AGCCAAGCTG  154950496
GTCCCTAAAA CCCATCTGAT GTCTGAATCT GAATGGAGGA ATCTTGGCGT  154950546
TCAGCAGAGT CAGGGATGGG TCCATTATAT GATCCATGAA CCAGgtcagt  154950596
gcactggcta aaaacaacca tatagaactg ctacactgag agaatgaaag  154950646
aataagattg tataacccaa atagggagat aggaaatggt ttactggttc  154950696
cttccccctc cagtcgtggg ggatttttt aaaaaaaaac tagtgaccaa  154950746
aaataagact aaaatatctg ggaagttcag agacaacctg tcactgaaaa  154950796
acctcctgta atctttcatt caatcagagg gtattctttt taaggccaca  154950846
tatagcctga tcatagcccc tgcctcattc tccatcgaaa acattcttgg  154950896
atgtgttcta aataagcaaa ggaaagtata tttattgata agacaccaga  154950946
cacccagctg ccaggcaaaa ctaataaggg acaccctggg gctgtataaa  154950996
catagcaaaa gaactgatat taacaattct gtacttggca gacagtccag  154951046
acttctgggt ctgcttctaa ggccatatgc ttaagtcttt atttagttat  154951096
aaagatctga gtgggtgata gctggggagg tggtagtgga atacactata  154951146
ggttgtacat agaatggtgt gagcttgtct cttagatttc cctcactctt  154951196
tcagAACCTC ACATCTTGCT GTTCCGGCGC CCACTACCCA AGAAACCAAA  154951246
GAAATGAAGC TGGCAAGCTA CTTTTCAGCC TCAAGCTTTA CACAGCTGTC  154951296
CTTACTTCCT AACATCTTTC TGATAACATT ATTATGTTGC CTTCTTGTTT  154951346
CTCACTTTGA TATTTAAAAG ATGTTCAATA CACTGTTTGA ATGTGCTGGT  154951396
AACTGCTTTG CTTCTTGAGT AGAGCCACCA CCACCATAGC CCAGCCAGAT  154951446
GAGTGCTCTG TGGACCCACA GCCTaAGCTG AGTGTGACCC CAGaagccac  154951496
gatgtgctct gtatccagaa cacacttggc agatggagga agcatctgag  154951546
tttgagacca tggctgttac agggatcatg taaacttgct gtt
```

BLAT: Alineamientos

```
000000135 acatgtcatgctgcccaaggacatagccaagctggtccctaaaacccatc 000000184
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
154950463 acatgtcatgctgcccaaggacatagccaagctggtccctaaaacccatc 154950512

000000185 tgatgtctgaatctgaatggaggaatcttggcgttcagcagagtcaggga 000000234
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
154950513 tgatgtctgaatctgaatggaggaatcttggcgttcagcagagtcaggga 154950562

000000235 tgggtccattatatgatccatgaaccag 000000262
>>>>>>>>> |||||||||||||||||||||||||||| >>>>>>>>>
154950563 tgggtccattatatgatccatgaaccag 154950590
```

```
000000263 aacctcacatcttgctgttccggcgcccactacccaagaaaccaaagaaa 000000312
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
154951201 aacctcacatcttgctgttccggcgcccactacccaagaaaccaaagaaa 154951250

000000313 tgaagctggcaagctacttttcagcctcaagctttacacagctgtcctta 000000362
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
154951251 tgaagctggcaagctacttttcagcctcaagctttacacagctgtcctta 154951300

000000363 cttcctaacatctttctgataacattattatgttgccttcttgtttctca 000000412
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
154951301 cttcctaacatctttctgataacattattatgttgccttcttgtttctca 154951350

000000413 ctttgatatttaaaagatgttcaatacactgtttgaatgtgctggtaact 000000462
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
154951351 ctttgatatttaaaagatgttcaatacactgtttgaatgtgctggtaact 154951400

000000463 gctttgcttcttgagtagagccaccaccaccatagcccagccagatgagt 000000512
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
154951401 gctttgcttcttgagtagagccaccaccaccatagcccagccagatgagt 154951450

000000513 gctctgtggacccacagcctcagctgagtgtgaccccag 000000551
>>>>>>>>> |||||||||||||||||||| ||||||||||||||||||| >>>>>>>>>
154951451 gctctgtggacccacagcctaagctgagtgtgaccccag 154951489
```

```
gtgctccgtg gggctcccat tatgttgtag ataatcaact agaagactat  154950196
aaggtccatg aaggcggaga aaatgtgcct tgaacttaag tgctcattca  154950246
ctggaattat tcactaattt gtatttttaa tatactctcg gaatttatat  154950296
aacatagcct gtatttctag ctgctgctac cccacaataa aattatataa  154950346
actctgacat caaaaaccag tttttctttc ctttcttggc cttgtaaaca  154950396
ggcatttgtc taagagactg tatctggtac taacataaat tccccacttc  154950446
cccgtttctg ttacagACAT GTCATGCTGC CCAAGGACAT AGCCAAGCTG  154950496
GTCCCTAAAA CCCATCTGAT GTCTGAATCT GAATGGAGGA ATCTTGGCGT  154950546
TCAGCAGAGT CAGGGATGGG TCCATTATAT GATCCATGAA CCAGgtcagt  154950596
gcactggcta aaaacaacca tatagaactg ctacactgag agaatgaaag  154950646
aataagattg tataacccaa atagggagat aggaaatggt ttactggttc  154950696
cttccccctc cagtcgtggg ggatttttt aaaaaaaaac tagtgaccaa  154950746
aaataagact aaaatatctg ggaagttcag agacaacctg tcactgaaaa  154950796
acctcctgta atctttcatt caatcagagg gtattctttt taaggccaca  154950846
tatagcctga tcatagcccc tgcctcattc tccatcgaaa acattcttgg  154950896
atgtgttcta aataagcaaa ggaaagtata tttattgata agacaccaga  154950946
cacccagctg ccaggcaaaa ctaataaggg acaccctggg gctgtataaa  154950996
catagcaaaa gaactgatat taacaattct gtacttggca gacagtccag  154951046
acttctgggt ctgcttctaa ggccatatgc ttaagtcttt atttagttat  154951096
aaagatctga gtgggtgata gctggggagg tggtagtgga atacactata  154951146
ggttgtacat agaatggtgt gagcttgtct cttagatttc cctcactctt  154951196
tcagAACCTC ACATCTTGCT GTTCCGGCGC CCACTACCCA AGAAACCAAA  154951246
GAAATGAAGC TGGCAAGCTA CTTTTCAGCC TCAAGCTTTA CACAGCTGTC  154951296
CTTACTTCCT AACATCTTTC TGATAACATT ATTATGTTGC CTTCTTGTTT  154951346
CTCACTTTGA TATTTAAAAG ATGTTCAATA CACTGTTTGA ATGTGCTGGT  154951396
AACTGCTTTG CTTCTTGAGT AGAGCCACCA CCACCATAGC CCAGCCAGAT  154951446
GAGTGCTCTG TGGACCCACA GCCTaAGCTG AGTGTGACCC CAGaagccac  154951496
gatgtgctct gtatccagaa cacacttggc agatggagga agcatctgag  154951546
tttgagacca tggctgttac agggatcatg taaacttgct gtt
```