Predicción computacional de genes

(Gene finding)

Predicción de genes (Gene finding)

- El genoma humano tiene 3.2 GB de nucleótidos, y poco más de 20.000 genes

ccgtacgtacgtagagtgctagtctagtcgtagcgccgtagtcgatcgtgtgggt
agtagctgatatgatgcgaggtaggggataggatagcaacagatgagcggatg
ctgagtgcagtggcatgcgatgtcgatgatagcggtaggtagacttcgcgcata
aagctgcgcgagatgattgcaaagragttagatgagctgatgctagaggtcagt
gactgatgatcgatgcatgcatggatgatgcagctgatcgatgtagatgcaataa
gtcgatgatcgatgatgatgctagatgatagctagatgtgatcgatggtaggtag
gatggtaggtaaattgatagatgctagatcgtaggtagtagctagatgcaggga
taaacacacggaggcgagtgatcggtaccgggctgaggtgttagctaatgatg
agtacgtatgaggcaggatgagtgacccgatgaggctagatgcgatggatgga
tcgatgatcgatgcatggtgatgcgatgctagatgatgtgtgtcagtaagtaagc
gatgcggctgctgagagcgtaggcccgagaggagagatgtaggaggaaggtt
tgatggtagttgtagatgattgtgtagttgtagctgatagtgatgatcgtag
……

¿Donde estan los genes que codifican las proteínas?


1 aacagggtgt atctcgcaca ttctcatcca ctagtataac tgctgctgac agtaatcgaa
61 ctagatagac tgttctggat gctatcattc gatattttga caacacggga gccatcctgt
121 tcgttgatcc gagattcgac gagtcatgca acaagatcca gaccgttgcc tgcaaacgcc
181 taggctgtga atgaacgact cgatcacgat cgctagtcgc acgtctgatc tcaccgattg
241 aagccgtatt ccacagagtg cgagaaccgg tcatttactg agtggttcgg ctctgtttaa
301 atacggaaag cccactcggg agagatatct ctccttaatg ggctatgaaa ggtatgaatg
361 gtggcggcga accgcgtttc ccagaggctc gcgcactcca gtactccccg gaacgctggt
421 gggcttatct tccgtgttcg ggatgggtac gggaggcaac cccaccgctg tggccgccta
481 acgtcgagtc acggaatcga accgcgatag taccagtctc gattaactct tccaccgtgt
541 gattacgtgc gatccagttt gcgcctggac tcgttcagcg acgagttaaa tcgatggtga
601 atgagtcaca gtgcgtatga atgatggctt tggtctgtta gtgctcgtgg gcttaacgtc
661 tcgttacctc gacgcgcaca ccccgagtct atcgaccgcg tcttgtacgc gggacctcgg
721 cggtgtctct tttccaagtg ggtttcgagc ttagatgcgt tcagctctta ccccgtgtgg
781 cgtggctacc cggcacgtgc tctctcgaac aaccggtaca ccagtggcca ccaaccgtag
841 ttcctctcgt actatacggt cgttcttgtc agacaccatt acacacccag tagatagcag
901 ccgacctgtc tcacgacggt ctaaacccag ctcacgacat cctttaatag gcgaacaacc
961 tcacccttgc ccgcttctgc acgggcagga tggagggaac cgacatcgag gtagcaagcc
1021 actcggtcga tatgtgctct tgcgagtgac gactctgtta tccctagggt agcttttctg
1081 tcatcaattg cccgcatcaa gcaggctaat tggttcgcta gaccacgctt tcgcgtcagc
1141 gttcctcgtt gggaagaaca ctgtcaagct taattttgct cttgcactct tcgccgggtc
1201 tctgtcccgg ctgagatagc catagggcgc gctcgatatc ttttcgagcg cgtaccgccc
1261 cagtcaaact gcccggctat cggtgtcctc ctcccggagt gagagtcgca gtcaccgacg
1321 ggtagtattt cactgttgac tcggtggccc gctagcgcgg gtacctgtgt agtgtctcct
1381 atgtatgctg cacatcggcg accacgtctc agcgacagcc tgcagtaaag ctccataggg
1441 tcttcgcttc cccctgggtg tctccagact ccgcactgga atgtacagtt caccgggccc
1501 aacgttggga cagtgaagct ctggttaatc cattcatgca agccgctact gatgcggcaa
1561 ggtactacgc taccttaaga gggtcatagt taccccgcc gttgacaggt ccttcgtcct
1621 cttgtacgag gtgttcagat acctgcactg ggcaggattc agtgaccgta cgagtccttg
1681 cggatttgcg gtcacctatg ttgttactag acagtccgag cttccgagtc actgcgacct
1741 gctccgttcc ggagcaggca tcccttcttc cgaaggtacg ggactaactt gccgaattcc
1801 ctaacgttgg ttgctcccga caggccttgg ctttcgccgc catggacacc tgtgtcggtt

Diferentes tipos de búsqueda:

- Genes RNA
  - tRNA, rRNA, snRNA, snoRNA, microRNA
- Genes codificadores de proteínas
  - Procariotas
    - No hay intrones, regulación más simple
  - Eucariotas
    - Exones-intrones
    - Regulación más compleja

Diferentes estrategias para la búsqueda computacional de genes

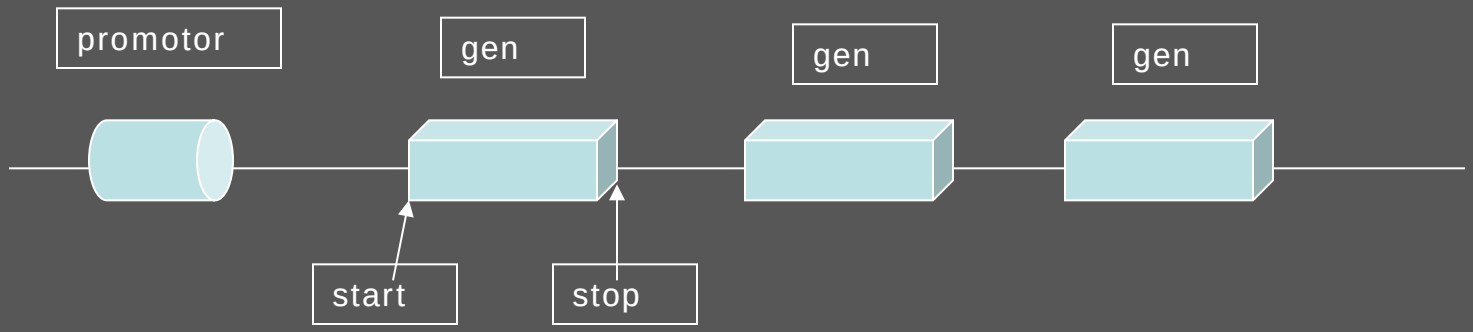
- Directa
  - Emparejamiento más o menos exacto con EST, cDNA o proteínas del mismo organismo o de otros relacionados

- Indirecta
  1. Homología con otros genes conocidos

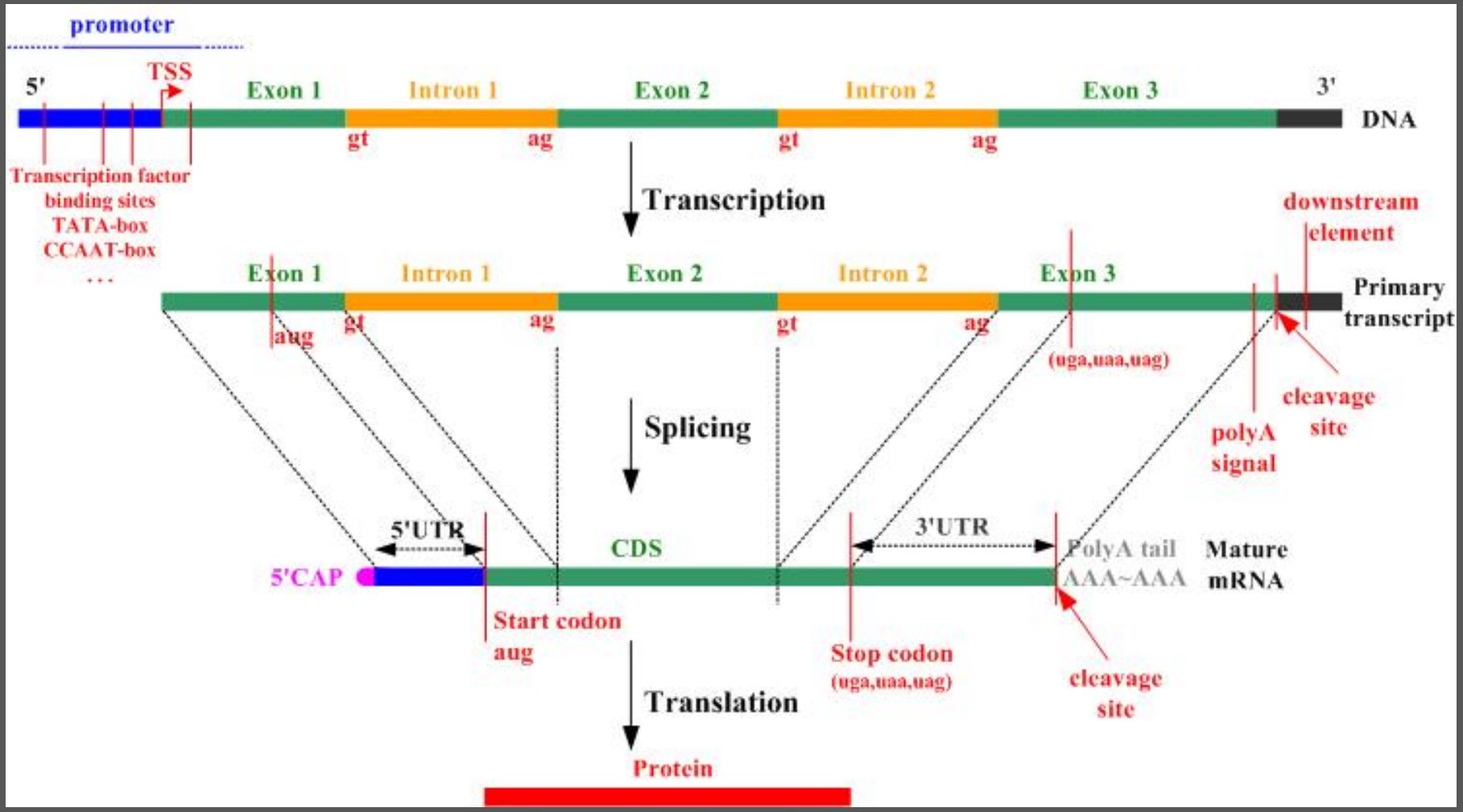
  2. Búsqueda de algo que se parece a un modelo teórico de gen (***ab initio***)

  3. Híbrida, combinando homología y búsqueda *ab initio* (y quizas también evidencia experimental)

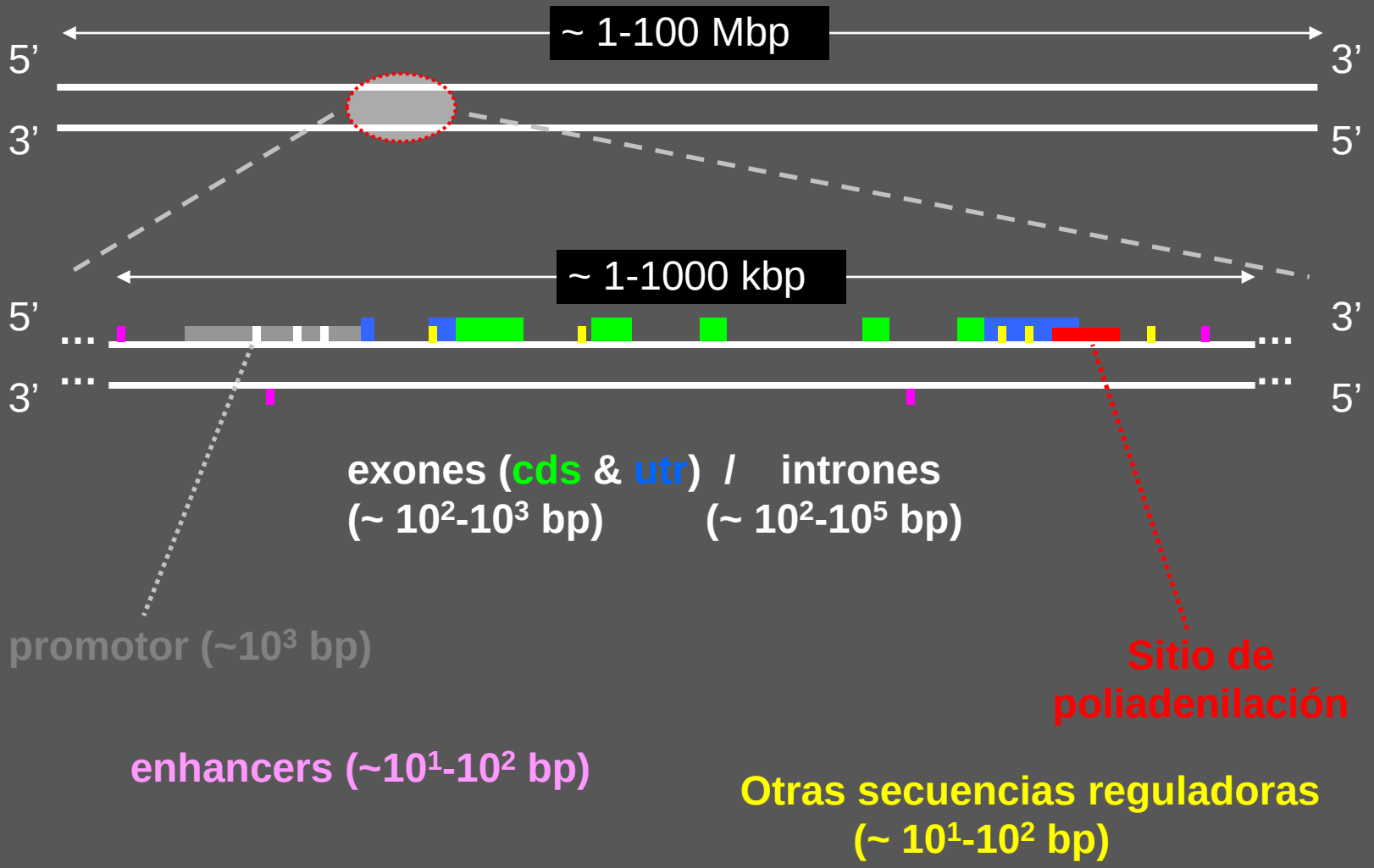
Gen procariótico



Gen eucariótico

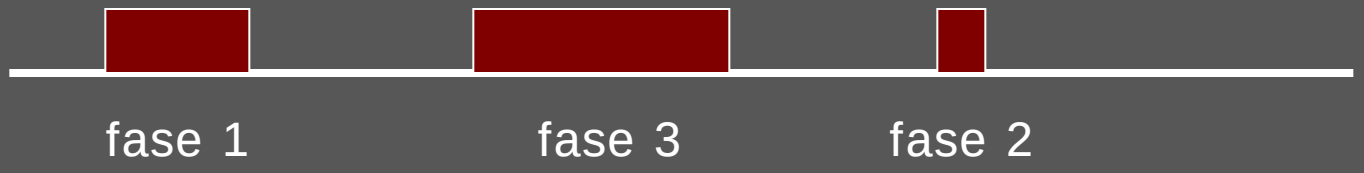


Gen eucariótico



Estructura génica

- Todos los exones de un gen están sobre la misma hebra, pero pueden estar en fases diferentes:



- Los exones de un gen tienen que tener una pauta de lectura consistente:

¿Qué podemos medir para predecir genes?

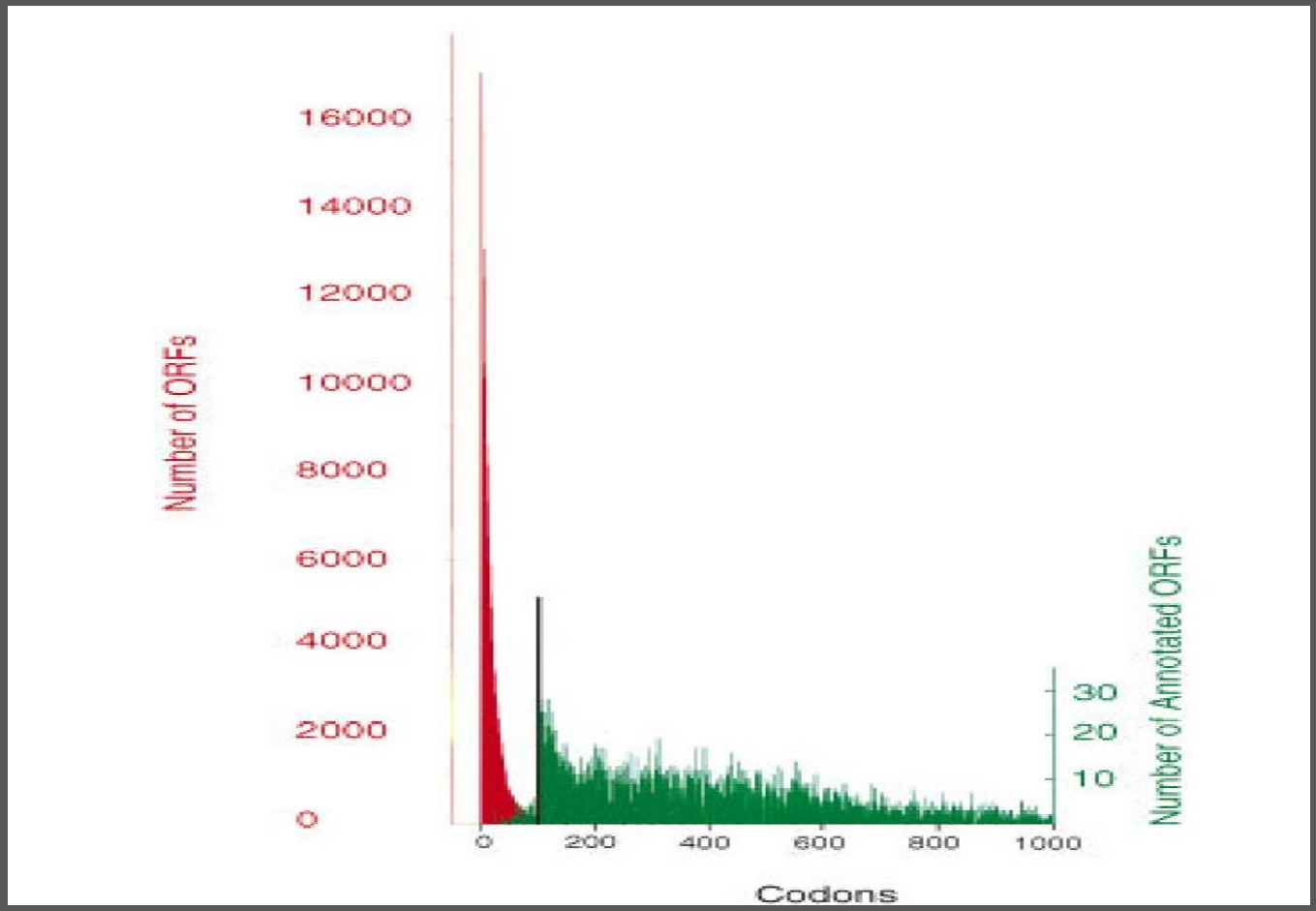
No existe aún la herramienta perfecta para predecir genes: todo se basa en ‘señales débiles’

- Genes codificadores de proteínas:
  - ORFs (Open Reading Frames)
  - Uso de codones
- Otros fenómenos
  - Frecuencias de nucleótidos y correlaciones
  - Sitios funcionales:
    - Sitios de splicing, promotores, UTRs, sitios de poliadenilación

¿Cómo identificar las ORFS que son exones?

- La distribución de longitudes no es aleatoria
  - Las ORFs largas tienen mayor probabilidad de ser exones (pero hay mini-exones!)
- ‘Firmas’ de los exones
  - Islas CpG
  - Sitios de splicing
  - Frecuencias de tetra- y hexa-nucleótidos
- ‘Firmas’ de los no-exones (elementos repetidos, ALUs, etc)
- Pauta de lectura ‘consistente’ entre los distintos exones de un gen

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

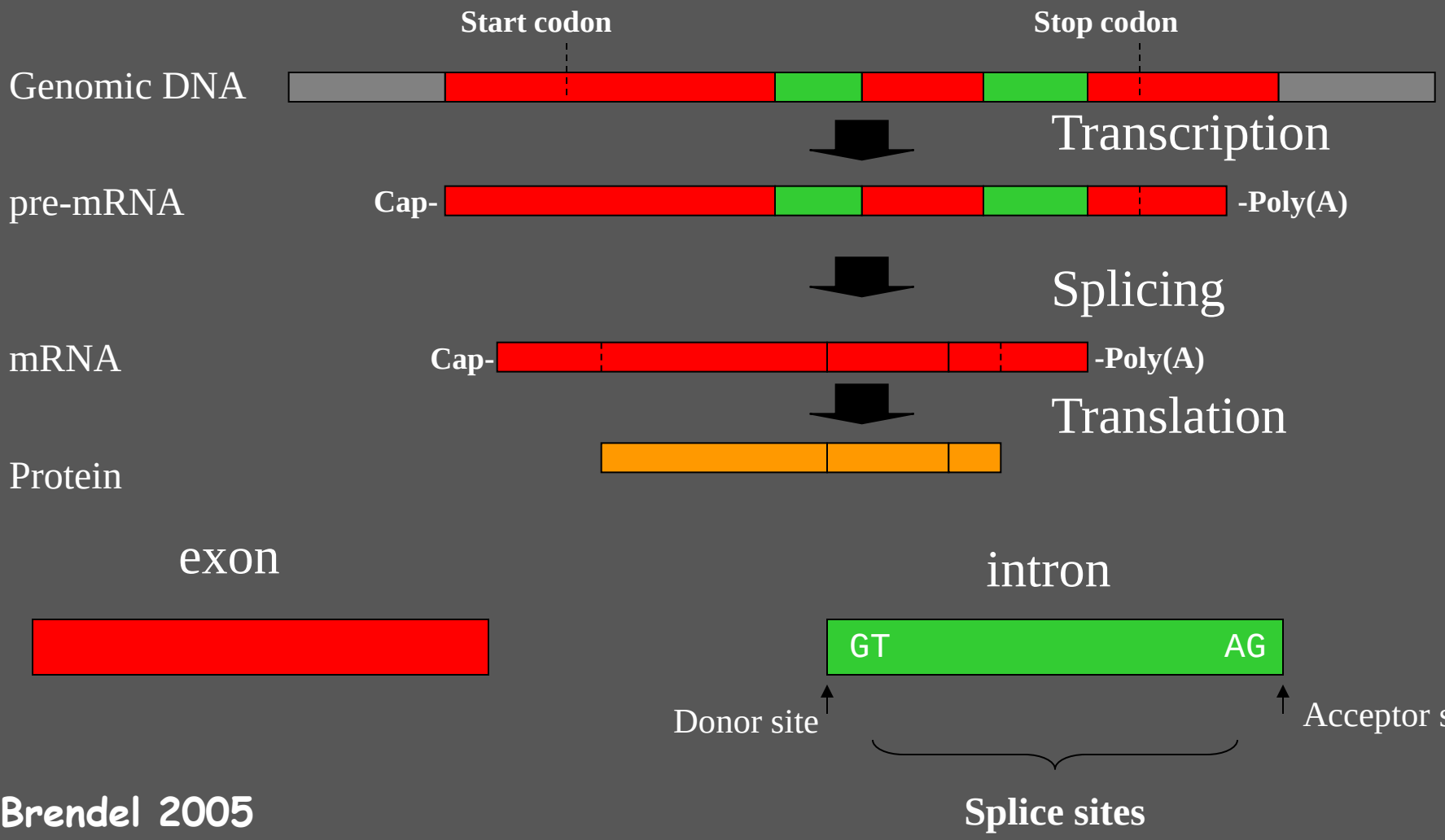


Basrai MA, Hieter P, and Boeke J Genome Research 1997 7:768-771

Islas CpG

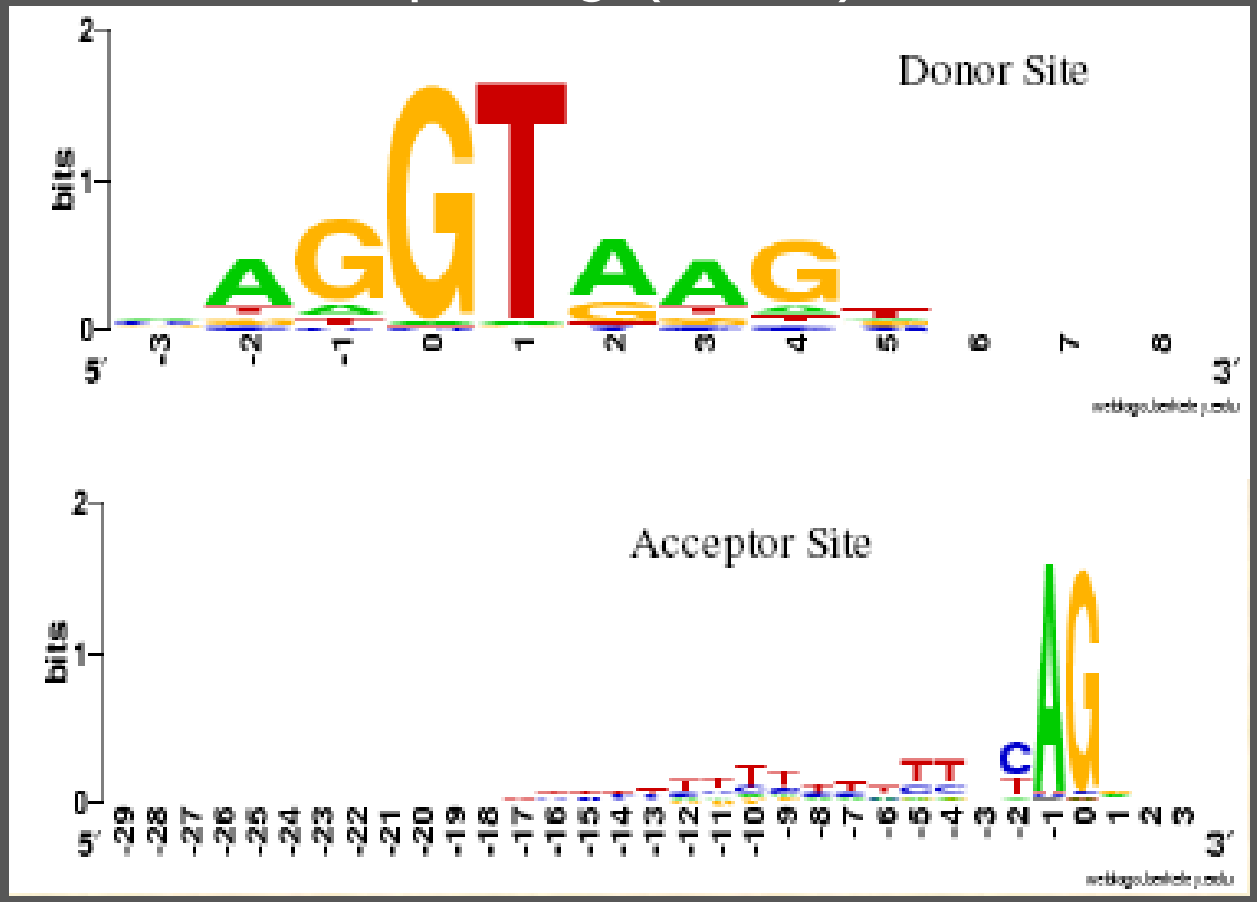
- Son regiones del genoma con una proporción relativamente alta (la que cabe esperar por azar) de dinucléotidos CpG
  - Solapan el promotor y los exones de un 50% de los genes de mamíferos
  - El resto del genoma contiene muy pocos CpGs y suelen estar metilados
- Definición clásica: secuencias >500 bp con
  - G+C > 55%
  - Observados(CpG)/Esperados(CpG) > 0.65
- Otra definición: clusters de CpGs estadísticamente significativos (CpGcluster)

Signals: Pre-mRNA Splicing



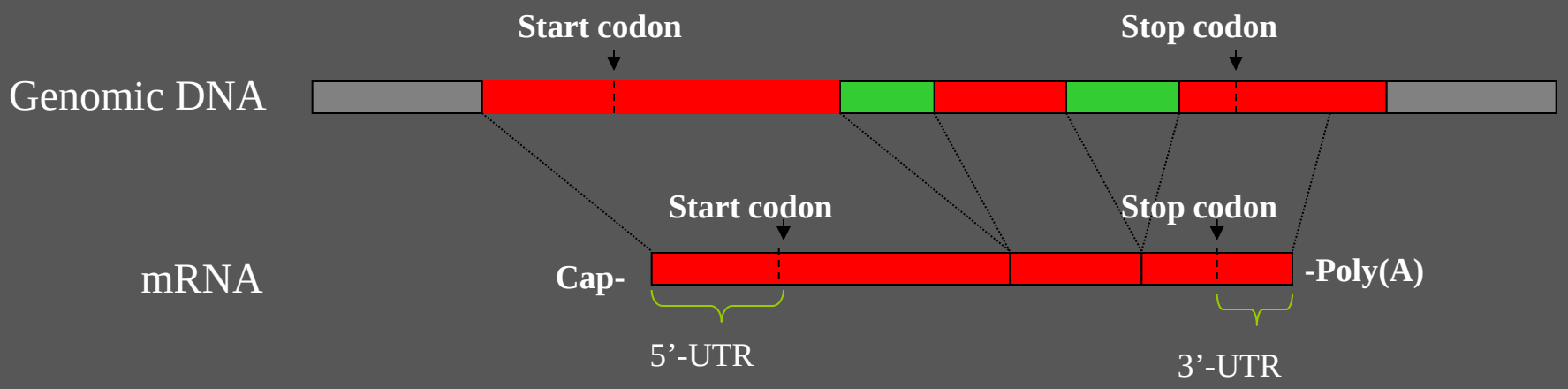
Brendel 2005

Señales de splicing (ratón)



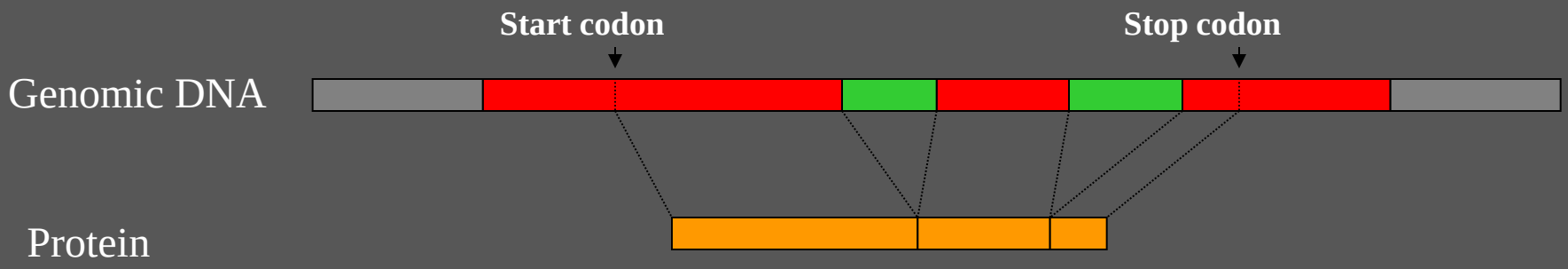
- Hay mucha variación (consenso estadístico)
- Muchos GT y AG en el genoma no son señales de splicing

Brendel - Spliced Alignment I:
Compare with cDNA or EST probes



Brendel 2005

Brendel - Spliced Alignment II:
Compare with protein probes



Brendel 2005

Aspectos a tener en cuenta acerca del software de predicción de genes

- En general es especie-específico
- Funciona mejor con genes que son razonablemente similares a otros conocidos previamente
- Hace falta información externa para identificar los ensamblados alternativos de un gen (splicing alternativo)
- Es imperfecto! (es biología despues de todo, donde la complejidad y las excepciones son la regla).

Retos pendientes en la predicción de genes (eucariotas)

- Splicing alternativo
  - Variantes de splicing, variantes Start/stop
- Genes solapados
  - La mayoría UTRs o intrónicos, pero también codificadores
- Elementos funcionales no-canónicos
  - Alternativas a la regla GT-AG
- Predicción de UTRs
  - Especialmente con intrones
- Exones pequeños (mini): hay exones con 3 bp!

Retos pendientes en la predicción de genes (procariotas)

- Predicción del codón de inicio
  - La mayoría de los algoritmos son ‘greedy’ (avaros), tendiendo a tomar la ORF más larga

- Genes solapados
  - Muy problemático, especialmente con los algoritmos de programación dinámica usados habitualmente

Herramientas para la búsqueda computacional de genes mediante homologías

- BLAST, FASTA, etc.
  - Pros: rápidos, bien fundamentados estadísticamente
  - Cons: no se tiene en cuenta la estructura génica

- BLAT, Sim4, EST_GENOME, etc.
  - Pros: tienen en cuenta la estructura génica
  - Cons: splicing no-canónico, más lentos que Blast

Programas y servidores web para la predicción de genes en eucariotas

- Genscan (ab initio), GenomeScan (hybrid)
  - (http://genes.mit.edu/)
- Twinscan (hybrid)
  - (http://genes.cs.wustl.edu/)
- FGENESH (ab initio)
  - (http://www.softberry.com/berry.phtml?topic=gfind)
- GeneMark.hmm (ab initio)
  - (http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi)
- MZEF (ab initio)
  - (http://rulai.cshl.org/tools/genefinder/)
- GrailEXP (hybrid)
  - (http://grail.lsd.ornl.gov/grailexp/)
- GeneID (hybrid)
  - (http://www1.imim.es/geneid.html)

Programas para la predicción de genes en procariotas

- Glimmer
  - http://www.tigr.org/~salzberg/glimmer.html
- GeneMark
  - http://opal.biology.gatech.edu/GeneMark/gmhmm2_prok.cgi
- Critica
  - http://www.ttaxus.com/index.php?pagename=Software
- ORNL Annotation Pipeline
  - http://compbio.ornl.gov/GP3/pro.shtml

Software para la predicción de genes no-codificadores

- tRNA
  - tRNA-ScanSE
    - http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
  - FAStRNA
    - http://bioweb.pasteur.fr/seqanal/interfaces/fastrna.html
- snoRNA (small nucleolar RNAs)
  - snoRNA database
    - http://rna.wustl.edu/snoRNAdb/
- miRNA (microRNA)
  - Sfold
    - http://www.bioinfo.rpi.edu/applications/sfold/index.pl
  - SIRNA (small interfering RNA)
    - http://bioweb.pasteur.fr/seqanal/interfaces/sirna.html