Introducción a la Bioinformática

Gene prediction

Fernán Agüero

Bioinformática - curso de posgrado

Instituto de Investigaciones Biotecnológicas UNSAM

Gene prediction

- Qué significa buscar/predecir genes?

- Dada una secuencia de DNA no caracterizada, encontrar:
  - qué región codifica para una proteína
  - que hebra codifica el gen
  - cuál es el marco de lectura
  - donde comienza y termina el gen
  - donde comienza y termina un intron/exon (euk)
  - (opcional) donde se encuentran las regiones regulatorias del gen

Procariotas vs Eucariotas

- **Procariotas**

- Genomas pequeños
- Alta densidad de genes
- Sin intrones

- Identificación de genes relativamente simple (~99%)

- Problemas
  - **ORFs solapados**
  - **genes cortos**
  - **encontrar promotores y TSS**

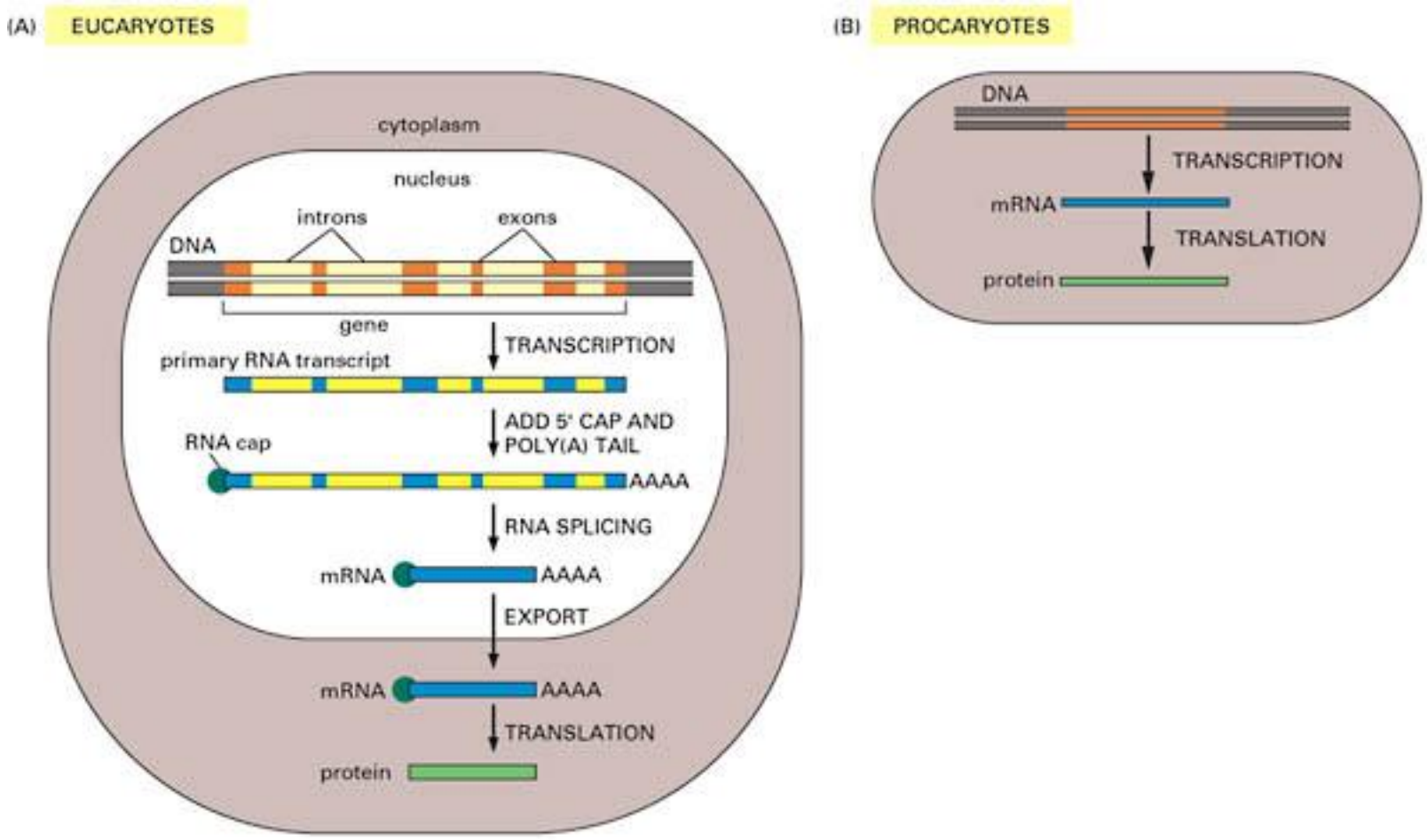
- **Eucariotas**

- Genomas grandes
- Baja densidad de genes
- Intrones y exones

- Identificación de genes es un problema complejo (~50%)

- Problemas
  - **muchos**

Estructura de los genes



Gene finding: distintas estrategias

- **Métodos basados en similitud de secuencias (extrínsecos)**
  - **Usan similitud con secuencias anotadas:**
    - proteínas
    - cDNAs
    - ESTs
- **Genómica comparativa**
  - **Alinear secuencias genómicas de distintas especies**
- ***Ab initio*** **gene finding (intrínseco)**
- **Estrategias que integran los anteriores**

Métodos basados en similitud

- Usan herramientas de alineamiento local (SW, BLAST, FASTA) para buscar proteínas, cDNA y ESTs
- No identifica genes que no estén en bases de datos (identifica sólo ~50%)
- Los límites de las regiones de similitud no están bien definidas

Similitud contra ESTs y cDNAs

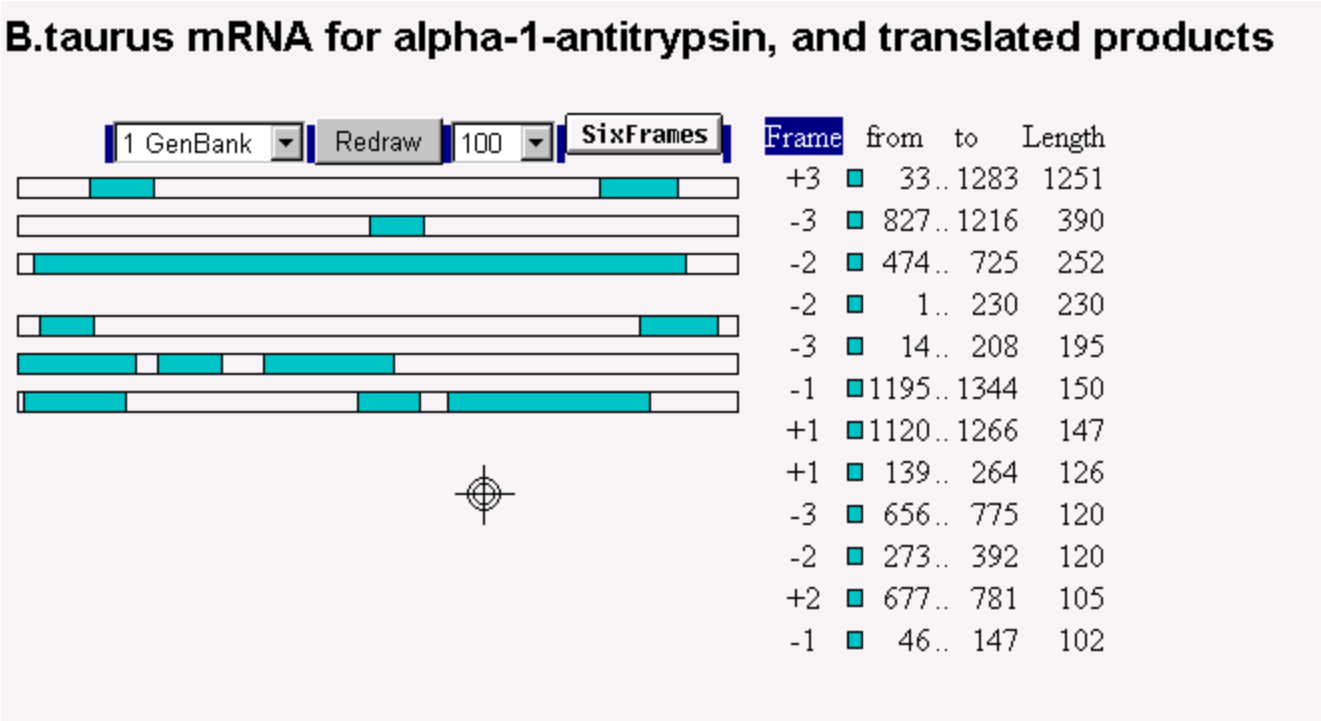
- Gran cantidad de ESTs disponibles. En vertebrados hay una gran cobertura
- Los cDNAs y algunos ESTs cubren más de un exon $\Rightarrow$ detección precisa de los límites intron/exon
- 1-5% de los EST's contienen intrones (splicing incompleto)

Bacterial gene prediction: ORFs

- Para genes procarióticos las técnicas más simples se basan en identificación de marcos de lectura abiertos (ORFs)
- Los ORFs se utilizan en búsquedas contra bases de datos de proteínas (blastx)
- Esto usualmente basta para cubrir densamente un genoma bacteriano
- Genes codificantes de tRNAs y rRNAs se detectan por separado usando tRNAscan o blastn

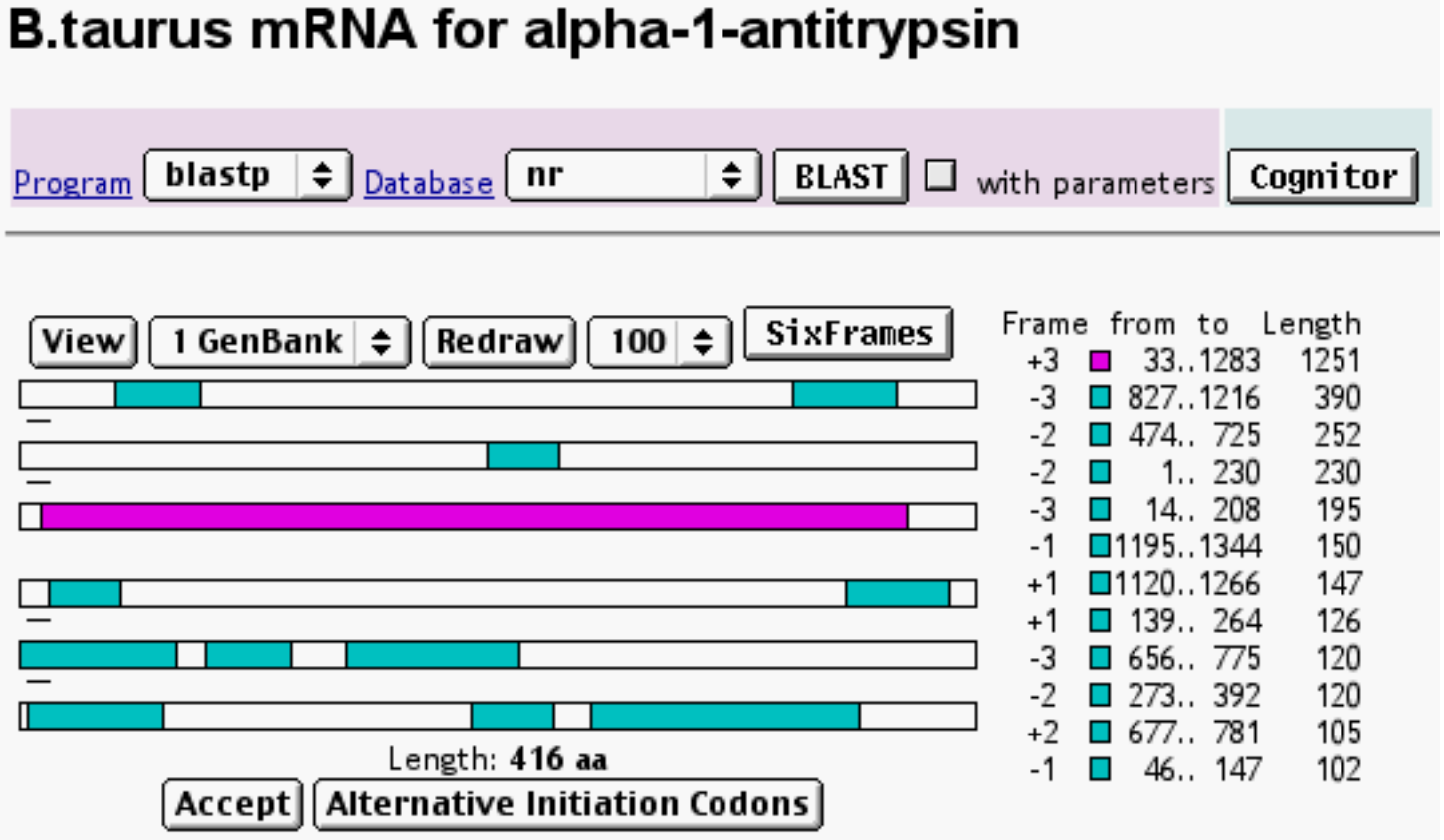
Gene prediction: ORFs

- **NCBI ORF Finder**
  - **http://www.ncbi.nlm.nih.gov/gorf/gorf.html**



Gene prediction: ORFs

- **NCBI ORF Finder**



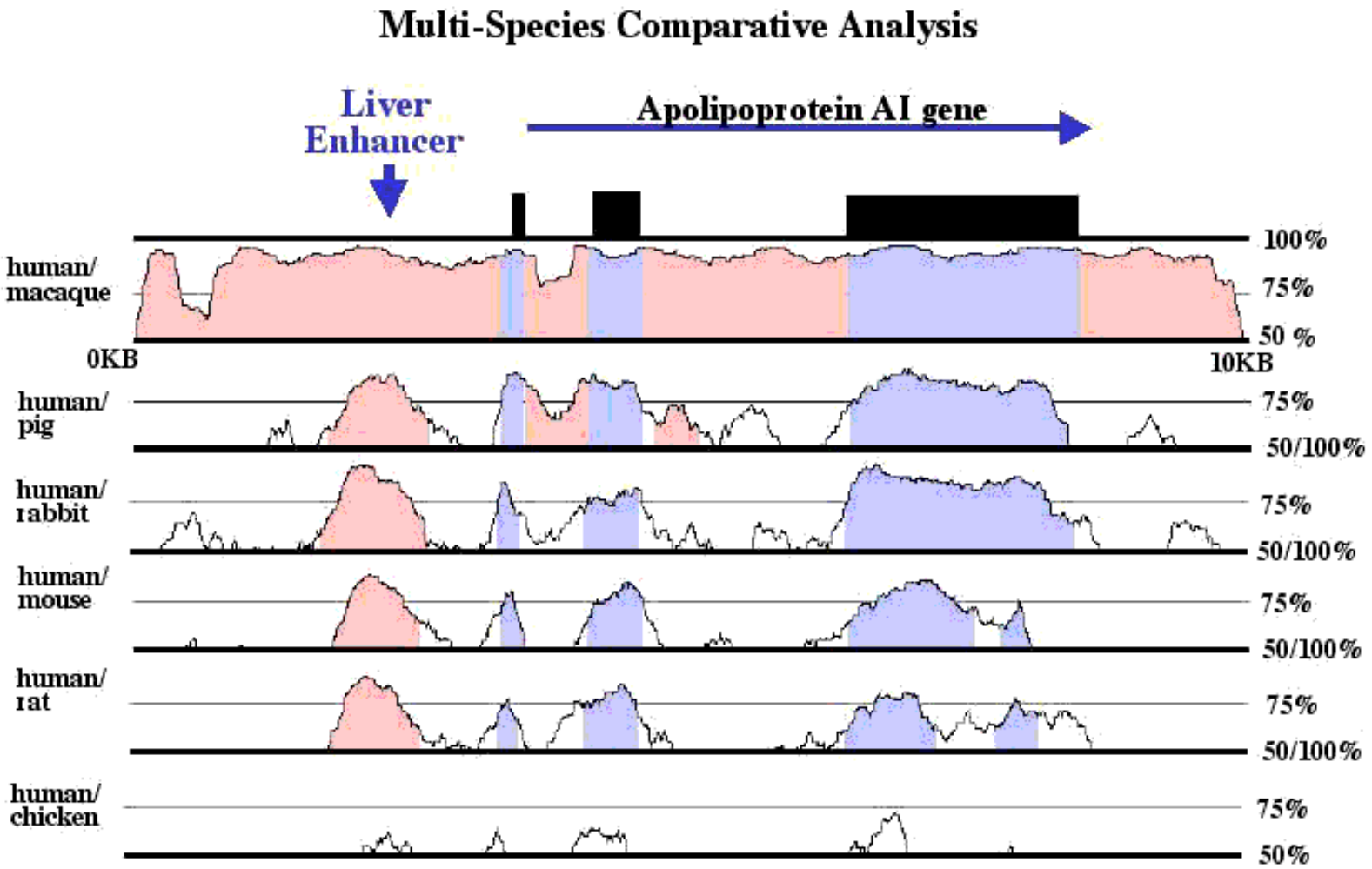
Genómica comparativa

- Se basa en la suposición de que las secuencias codificantes están más conservadas que las no-codificantes

- Dos estrategias:
  - intra-genómica (familias de genes)
  - inter-genómica (cross-species)

- Alineamiento de regiones homólogas
  - Difícil delinear los límites de similitud
  - Difícil definir una distancia evolutiva óptima (la conservación difiere entre loci)

Genómica comparativa

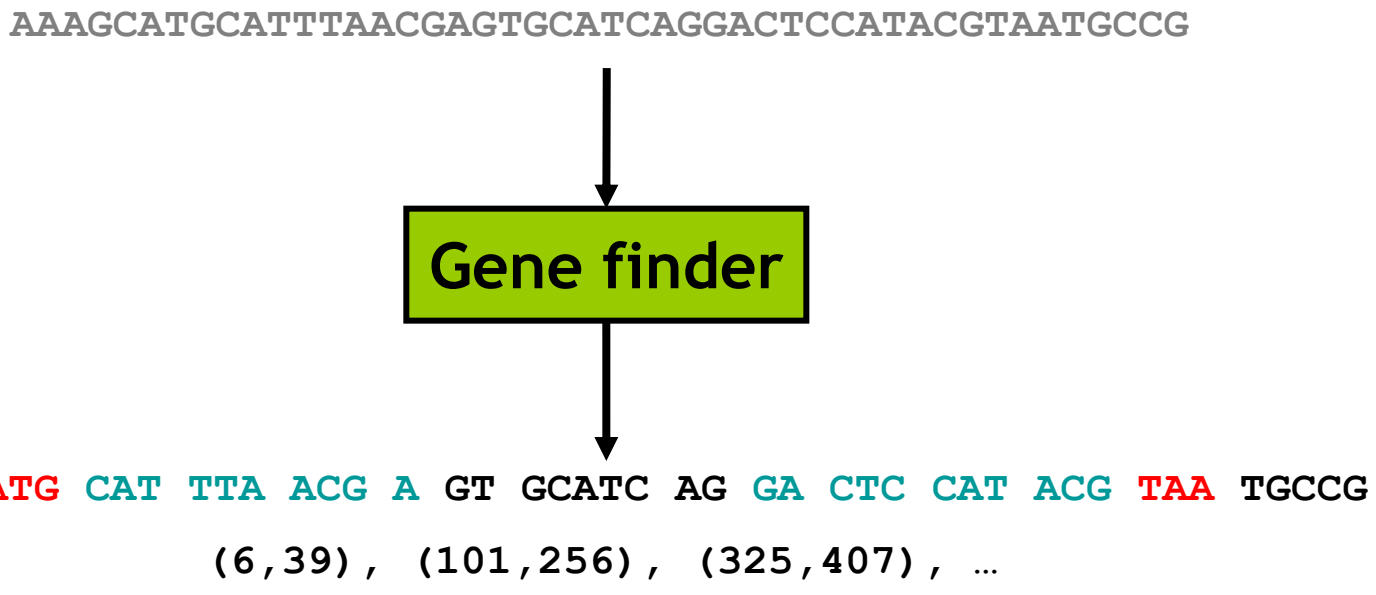


Resumen métodos extrínsecos

- **Pros**
  - **Se basan en información biológica pre-existente, deberían producir predicciones relevantes**
- **Contras**
  - **Limitado a información biológica pre-existente**
  - **Errores en las bases de datos**
  - **Difícil definir los límites de un gen en base a similitud**
  - **Es más rápido correr un programa de predicción *ab initio* que comparar contra GenBank usando blastx!**

ab initio gene finding

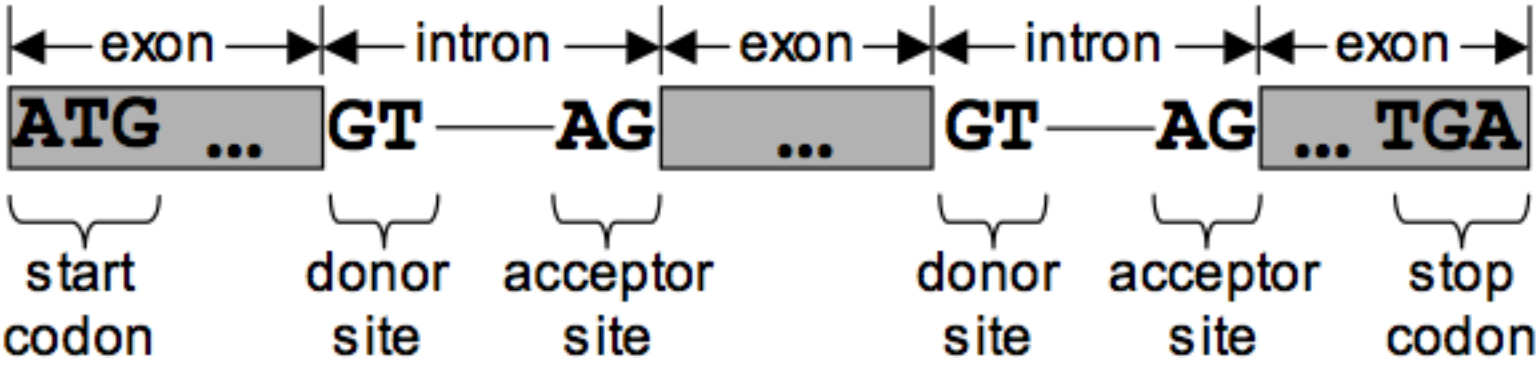
- Input: una cadena de DNA {A,C,G,T}
- Output: una anotación de la cadena que diga para cada nucleótido, si es codificante o no
- Usando sólo información de secuencia



ab initio gene finding

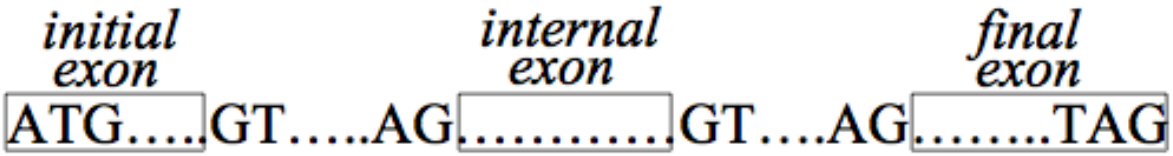
- Qué se busca?
- Información que permita identificar genes
  - Estructura regular
  - Longitud de exones/intrones
  - Composición de nucleótidos
  - Señales biológicas

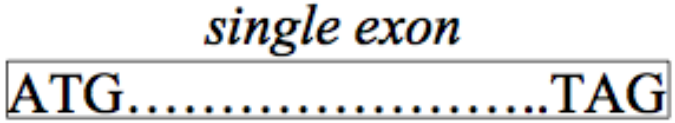
- Ejemplo de estructura regular:



Tipos de exones

- **Se definen 3 tipos de exones, por conveniencia:**
  - ***exones iniciales,*** **se extienden desde un codon de inicio hasta el primer sitio dador de splicing**
  - ***exones internos*** **se extienden desde un sitio aceptor hasta el próximo sitio dador de splicing**
  - ***exones finales*** **se extienden desde el último sitio aceptor hasta el codon stop**
  - ***single exons*** **(sólo ocurren en genes *intronless*) se extienden desde un codon de inicio hasta el codon stop**





Métodos estadísticos

- Se basan en medidas de distintos estimadores a partir de la secuencia

- Ejemplo:
  - Análisis de la secuencia en los 6 marcos de lectura
  - Distribución de codones de inicio y stop
  - Selección del marco con menor número (densidad) de stops

Secuencias codificantes: propiedades

- Una característica universal presente en cualquier genoma es el uso desigual de codones en las regiones codificantes
  - **uso desigual de aminoácidos en proteínas**
  - **uso desigual de codones sinónimos (se correlaciona con la abundancia de los tRNAs correspondientes)**

- Podemos usar esta característica para diferenciar entre regiones codificantes y no codificantes del genoma

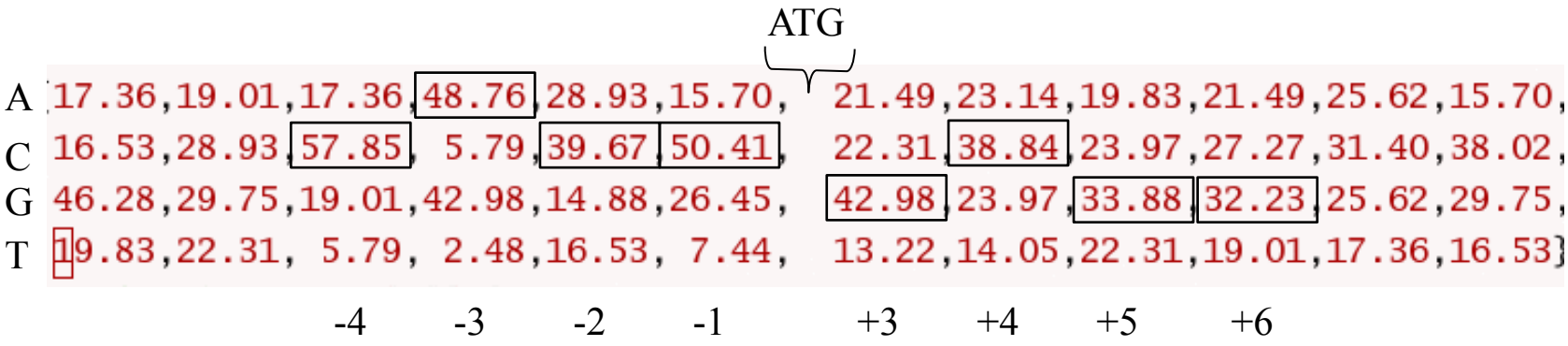
- **Coding statistics:** función que para una dada secuencia de DNA calcula la posibilidad de que la secuencia sea codificante

Coding statistics / Feature sensing

- **Señales / Signals**
  - **Cortas, de longitud definida: codones START, STOP, sitios de splicing**
- **Regiones con contenido biológico**
  - **Regiones de longitud variable: exones, intrones**
- **Como medir/sensar estos dos tipos de *features*?**
  - ***Señales: scanear la secuencia (sliding window)***
  - ***Contenido: asignar un score para regiones entre señales***
- **Cómo caracterizar regiones de longitud variable?**
  - **uso de codones (CUTG)**
  - **frecuencia de hexámeros (hexamer)**
  - **Azar/No-azar (testcode)**
  - **contenido de GC**
  - **periodicidad de nucleótidos**

Predicción del codon de inicio

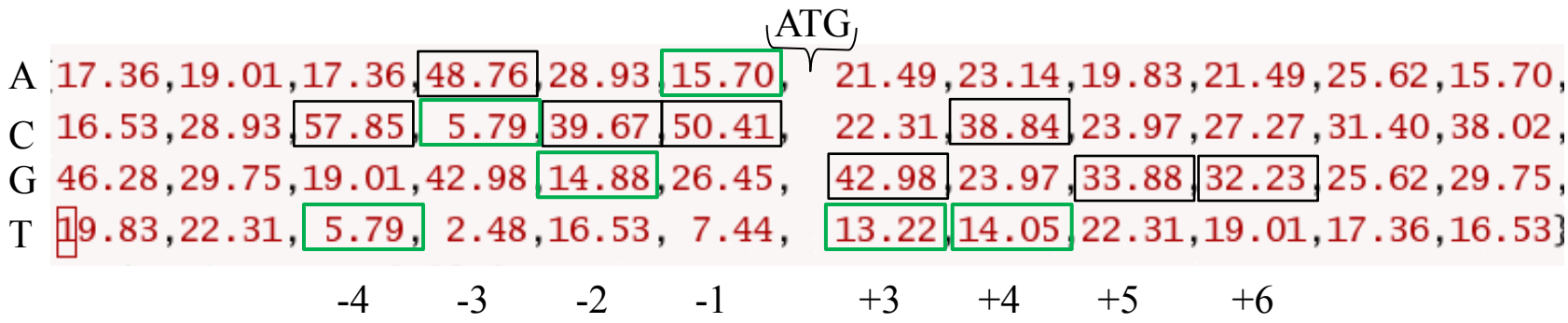
- Ciertos nucleótidos “prefieren” ciertas posiciones alrededor de un codon de inicio



- Esta desviación de la distribución al azar es información!
- Cuál de estas dos secuencias tiene mayor probabilidad de ser un codon de inicio?
  - **CACCATGGC**
  - **TCGAATGTT**

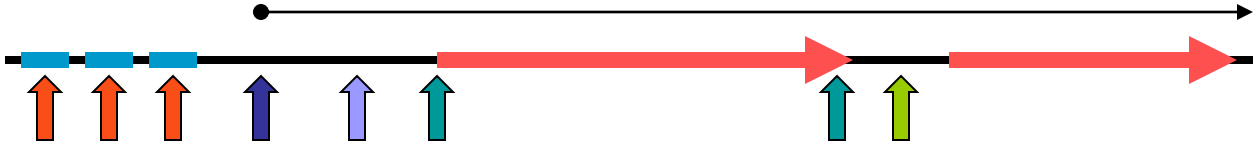
Como evaluar sitios informativos

- **Matemáticamente**
  - **F_i (X): frecuencia del residuo X (A,C,G,T) en la posición *i***
  - **El puntaje de una cadena de ADN (ej: CACCATGGC) se calcula como:**
    - $\sum \log (F_i/(X) / 0.25)$



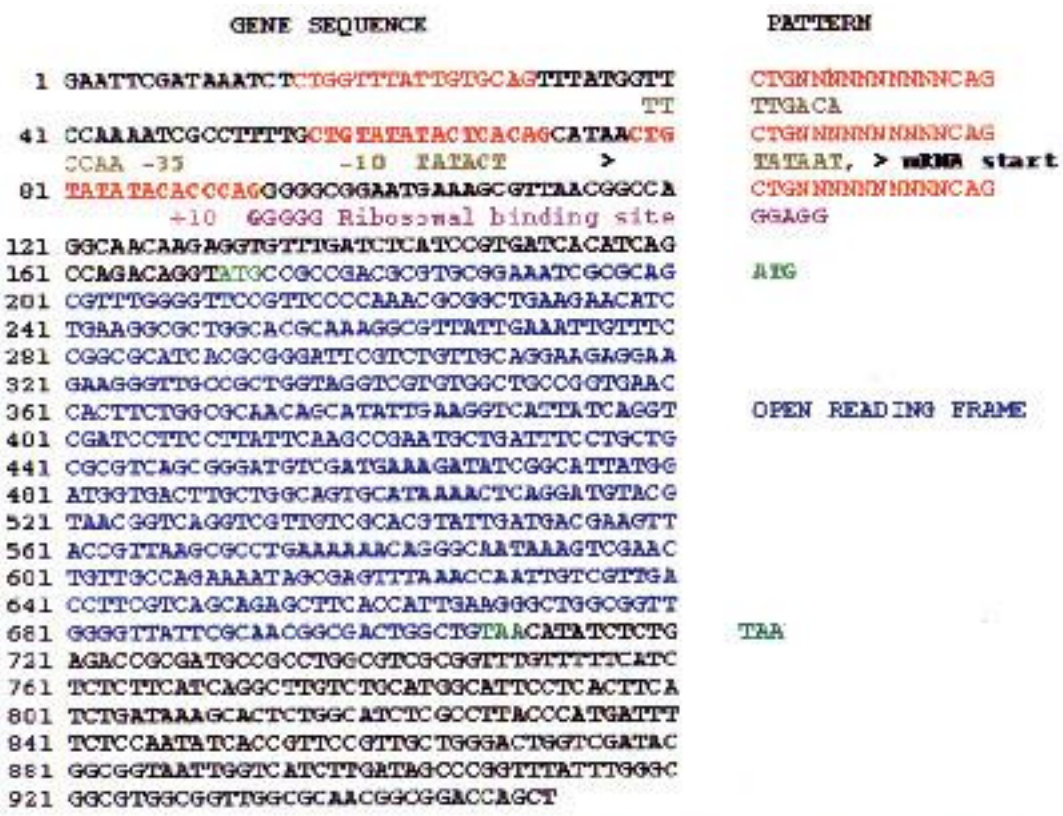
CACCATGGC: log(58/0.25) + log(49/0.25) + log(40/0.25) + log(50/0.25) + log(43/0.25) + log(39/0.25) = **13.59**
TCGAATGTT: log(6/0.25) + log(6/0.25) + log(15/0.25) + log(16/0.25) + log(13/0.25) + log(14/0.25) = **9.8**

Bacterial gene structure



- Transcription factor binding site.
- Promoters
  - 35 sequence ($T_{82}T_{84}G_{78}A_{65}C_{54}A_{45}$) 15-20 bases
  - 10 sequence ($T_{80}A_{95}T_{45}A_{60}A_{50}T_{96}$) 5-9 bases
- Start of transcription : initiation start: Purine (sometimes it's the "A" in CAT)
- Translation binding site (shine-dalgarno) 10 bp upstream of AUG (AGGAGG)
- One or more Open Reading Frames
  - start-codon (unless sequence is partial)
  - until next in-frame stop codon on that strand
  - Separated by intercistronic sequences
- Termination

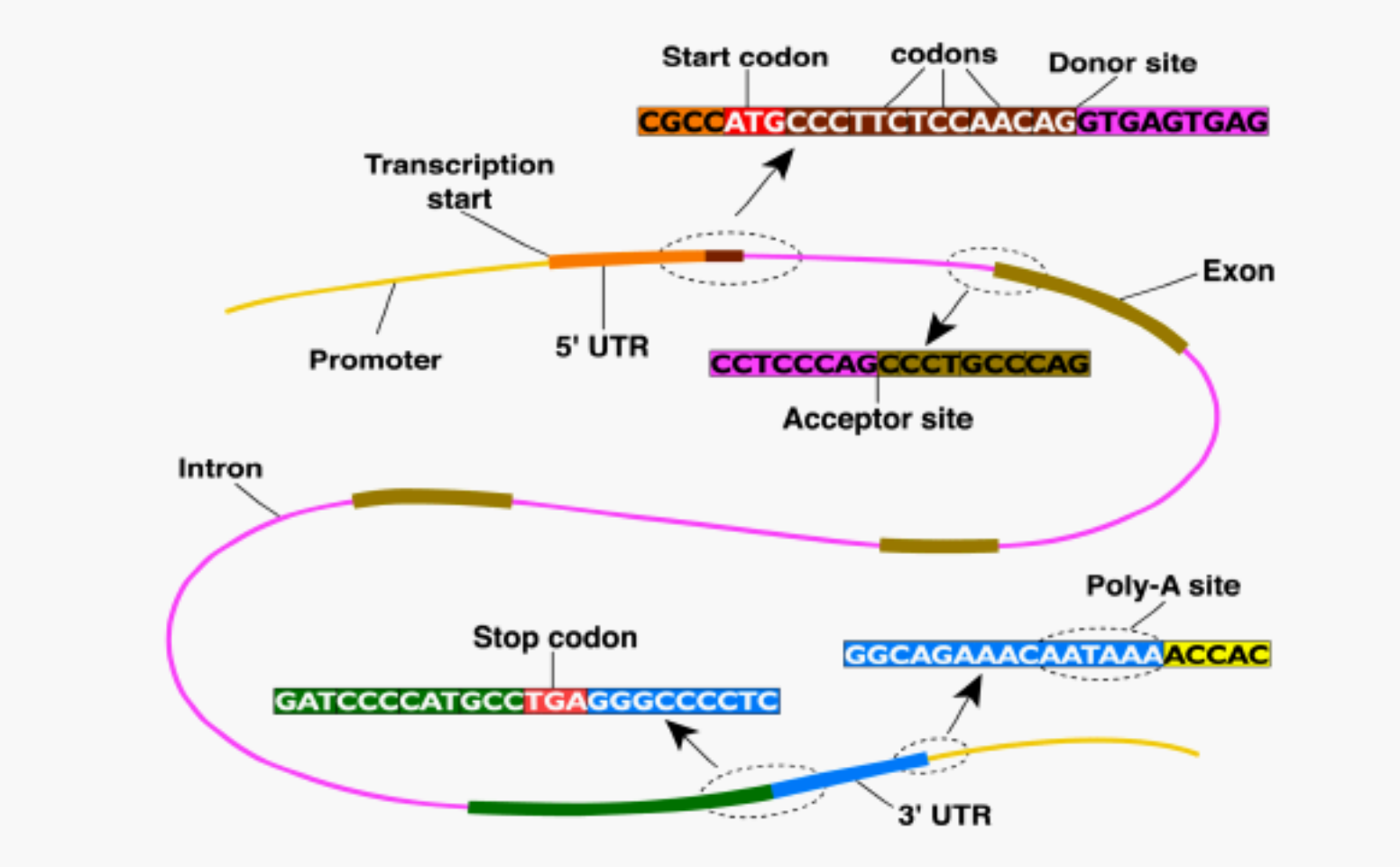
Bacterial gene structure



Shown are matches to approximate consensus binding sites for LexA repressor (CTGNNNNNNNNNNCAG), the -10 and -35 promoter regions relative to the start of the mRNA (TTGACA and TATAAT), the ribosomal binding site on the mRNA (GGAGG), and the open reading frame (ATG...TAA). Only the second two of the predicted LexA binding sites actually bind the repressor.

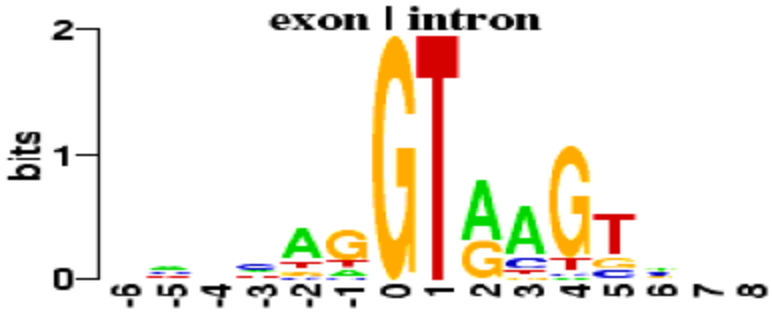
Signal sensors

Signal - una región en el ADN reconocida por la maquinaria celular



Signal sensors

- **Varios métodos de reconocimiento de patrones se utilizan para identificar estas señales**
  - secuencias consenso
  - matrices
  - HMMs
  - redes neurales



Position

%	-8	…	-2	-1	0	1	2	…	17
A	26	…	60	9	0	0	54	…	21
C	26	…	15	5	0	1	2	…	27
G	25	…	12	78	100	0	41	…	27
T	23	…	13	8	0	99	3	…	25

Secuencias consenso

- Ejemplo: obtenidas por selección de la base más frecuente en cada posición de un alineamiento múltiple
- Producen pérdida de la información
- Pueden producir muchos falsos positivos o falsos negativos

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT
TATAAT
TATRNT

Consenso
Consenso IUPAC

Matrices

- **Positional weight matrix**
  - **Se calcula midiendo la frecuencia de cada elemento para cada posición en el sitio**
  - **El score para cada sitio putativo es la suma de los valores de la matriz (convertidos en probabilidades) para esa secuencia**

- **Desventajas**
  - **Se necesita un cut-off value**
  - **supone independencia entre bases adjacentes**

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

	1	2	3	4	5	6
A	0	6	0	3	4	0
C	0	0	1	0	1	0
G	1	0	0	3	0	0
T	5	0	5	0	1	6

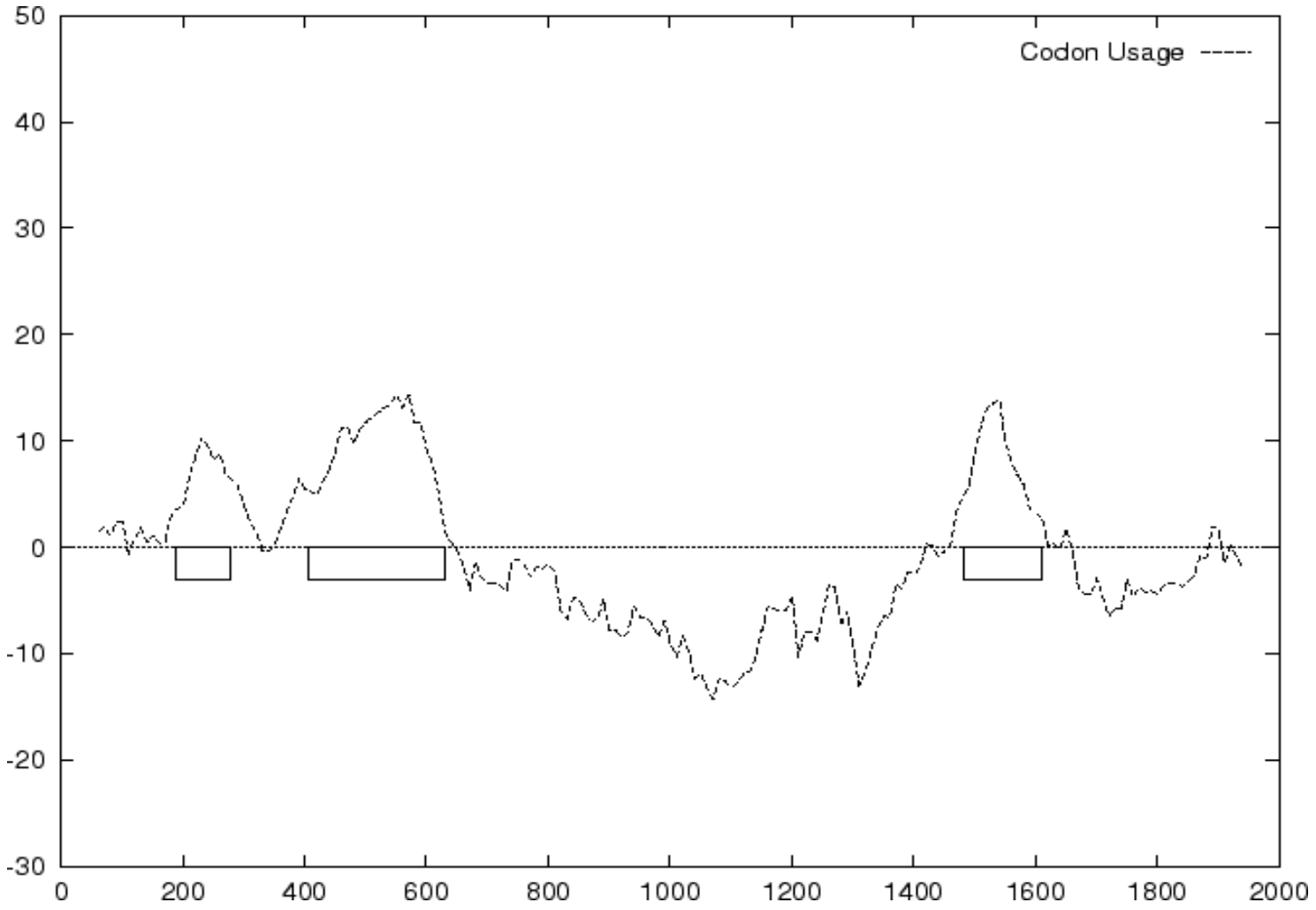
Codon usage

- **Tablas de uso de codones**

The Human Codon Usage Table

Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

Codon usage plots



β-globin gene

Codon Usage Database

- **Codon Usage Database**
  - http://www.kazusa.or.jp/codon/
  - **Derivada de secuencias codificantes de DDBJ/EMBL/GenBank**

Escherichia coli K12 **[gbbct]: 5054 CDS's (1603901 codons)**

fields: [triplet] [frequency: **per thousand**] ([number])

UUU 22.4(35930)	UCU 8.5(13633)	UAU 16.3(26180)	U
UUC 16.6(26609)	UCC 8.6(13783)	UAC 12.3(19675)	U
UUA 13.9(22279)	UCA 7.1(11438)	UAA 2.0(3244)	U
UUG 13.7(22000)	UCG 8.9(14305)	UAG 0.2(365)	U
CUU 11.0(17707)	CCU 7.0(11291)	CAU 12.9(20686)	C
CUC 11.0(17715)	CCC 5.5(8861)	CAC 9.7(15595)	C
CUA 3.9(6182)	CCA 8.5(13664)	CAA 15.5(24787)	C
CUG 52.8(84714)	CCG 23.3(37316)	CAG 28.8(46256)	C
AUU 30.4(48766)	ACU 8.9(14303)	AAU 17.6(28256)	A
AUC 25.0(40097)	ACC 23.4(37495)	AAC 21.7(34752)	A
AUA 4.3(6866)	ACA 7.0(11267)	AAA 33.6(53920)	A
AUG 27.8(44539)	ACG 14.4(23056)	AAG 10.2(16370)	A
GUU 18.4(29487)	GCU 15.3(24609)	GAU 32.2(51670)	G
GUC 15.2(24406)	GCC 25.5(40914)	GAC 19.1(30559)	G
GUA 10.9(17443)	GCA 20.3(32529)	GAA 39.6(63484)	G
GUG 26.2(42097)	GCG 33.7(53984)	GAG 17.8(28529)	G

Yersinia pestis CO92 **[gbbct]: 4069 CDS's (1289657 codons)**

fields: [triplet] [frequency: **per thousand**] ([number])

UUU 22.3(28716)	UCU 9.9(12830)	UAU 19.8(25541)	UGU 5.6(7252)
UUC 15.7(20289)	UCC 7.4(9576)	UAC 10.1(13065)	UGC 4.8(6161)
UUA 21.6(27867)	UCA 10.8(13872)	UAA 1.8(2293)	UGA 0.9(1196)
UUG 23.9(30882)	UCG 7.1(9127)	UAG 0.4(574)	UGG 13.6(17501)
CUU 9.5(12300)	CCU 10.1(12980)	CAU 14.1(18216)	CGU 20.2(26101)
CUC 9.1(11730)	CCC 7.6(9757)	CAC 8.7(11169)	CGC 16.8(21653)
CUA 7.9(10126)	CCA 12.4(15988)	CAA 23.7(30579)	CGA 4.1(5321)
CUG 36.6(47209)	CCG 13.5(17431)	CAG 24.8(31947)	CGG 7.9(10159)
AUU 30.7(39591)	ACU 10.6(13684)	AAU 24.0(30895)	AGU 12.9(16608)
AUC 23.4(30163)	ACC 21.6(27860)	AAC 16.5(21318)	AGC 14.4(18520)
AUA 7.7(9914)	ACA 9.6(12416)	AAA 32.1(41354)	AGA 3.1(4049)
AUG 26.3(33965)	ACG 12.1(15552)	AAG 12.1(15572)	AGG 2.0(2622)
GUU 19.2(24718)	GCU 19.7(25423)	GAU 36.6(47256)	GGU 26.2(33795)
GUC 15.5(20047)	GCC 26.9(34739)	GAC 14.5(18678)	GGC 23.6(30428)
GUA 11.1(14354)	GCA 21.0(27060)	GAA 34.7(44765)	GGA 6.2(8051)
GUG 22.8(29371)	GCG 23.9(30780)	GAG 20.5(26469)	GGG 15.7(20232)

Testcode

- Fickett, 1982
- Evalúa el azar posicional en una secuencia
  - en secuencias codificantes, la tercera base tiende a ser la misma con más frecuencia que la esperada por azar (non-random)
- Esto es debido al uso preferencial de ciertos codones
- Es una propiedad universal
- testcode (GCG), testcode (perl)

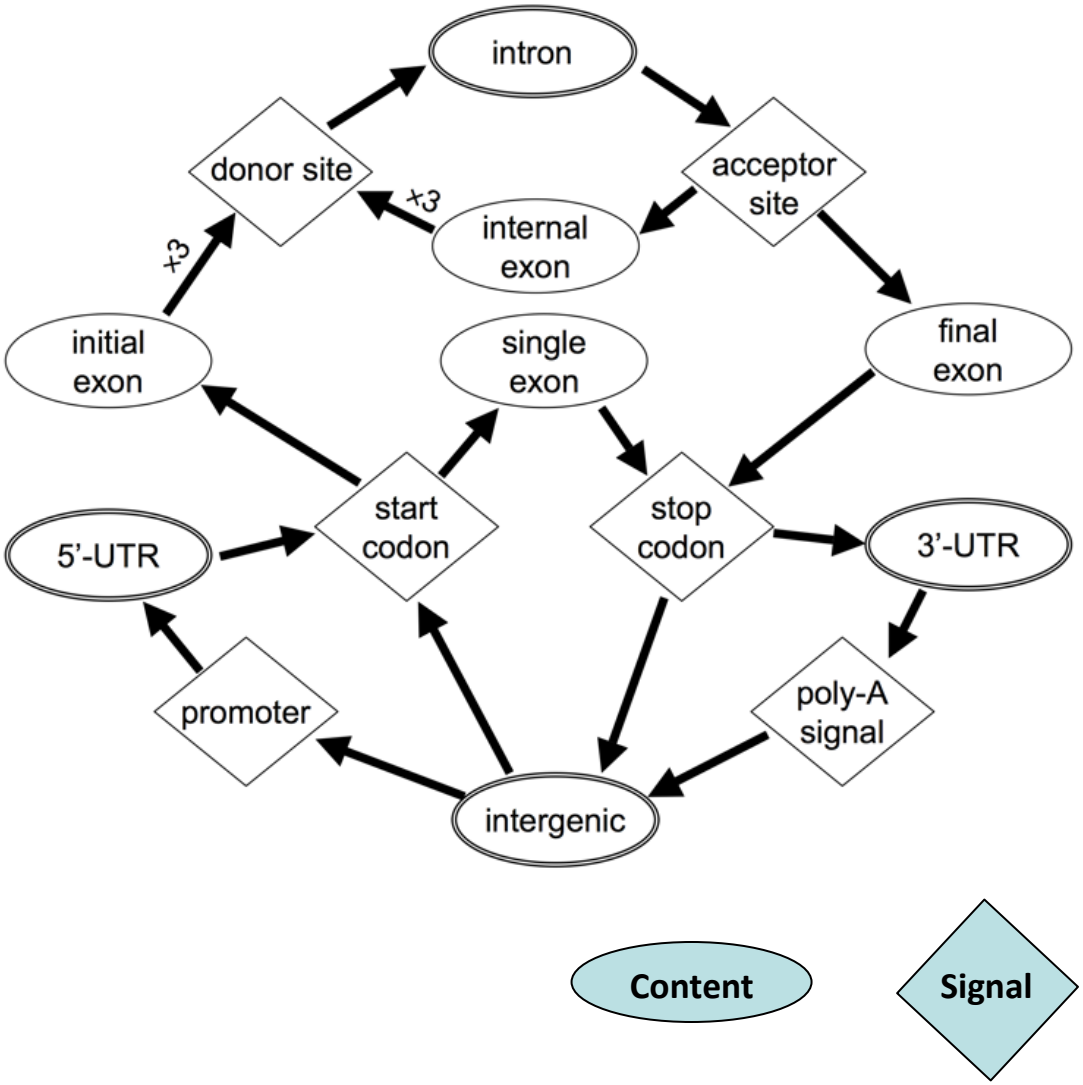
Cómo se integra todo esto?

- Coding statistics y signal sensors se integran en un modelo global usando
  - machine learning (HMMs, árboles de decisión, redes neurales)
  - discriminant analysis (distintas funciones: lineales, cuadráticas)

- Son capaces de predecir
  - genes en ambas hebras simultáneamente
  - genes parciales o muchos gernes en una secuencia
  - exones subóptimos

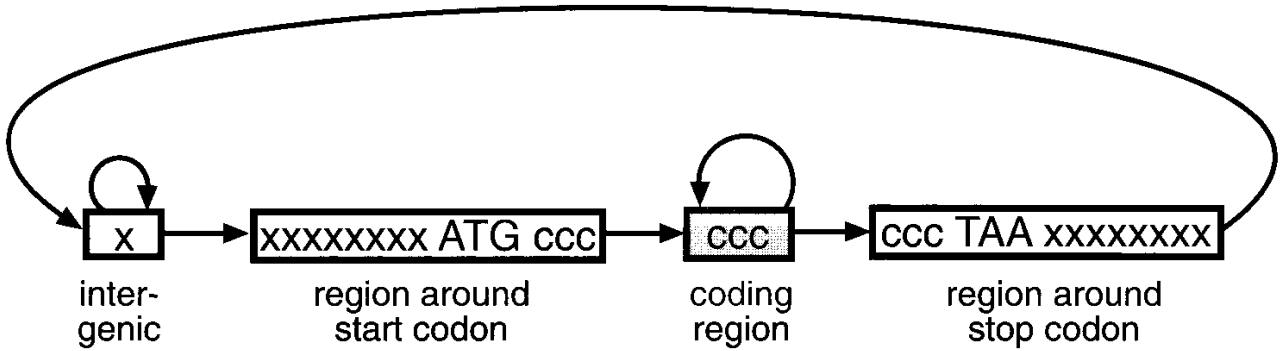
Generalized Hidden Markov Model

- En cada uno de los estados hay un ***sensor***
- El sensor evalúa la probabilidad de que la secuencia de ADN se encuentre en ese estado.



HMMs

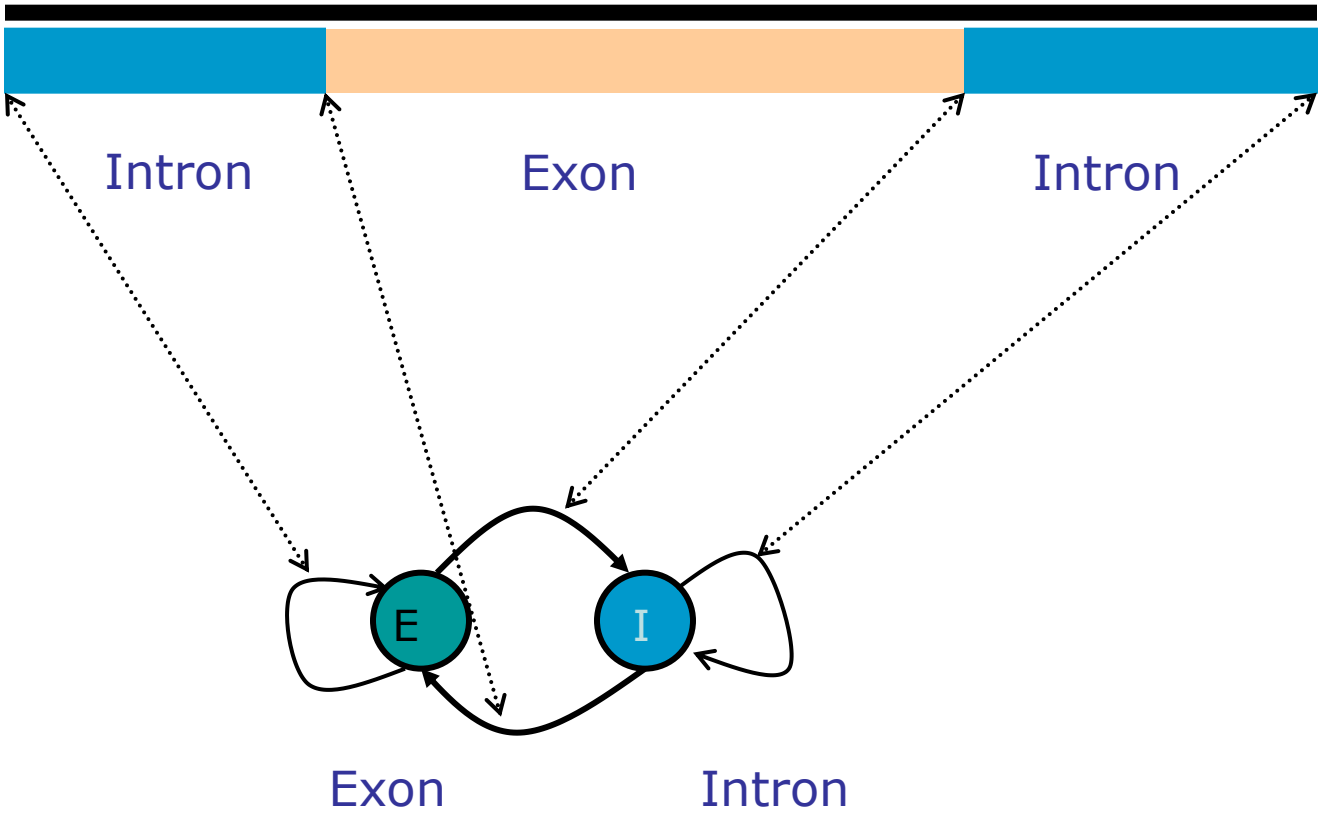
- Nucleótidos {A,C,G,T} son las observaciones
- Diferentes estados generan nucleótidos con distintas frecuencias
- Un HMM simple para genes sin intrones:



AAAGC **ATG** CAT TTA ACG AGA GCA CAA GGG CTC **TAA** TGCCG

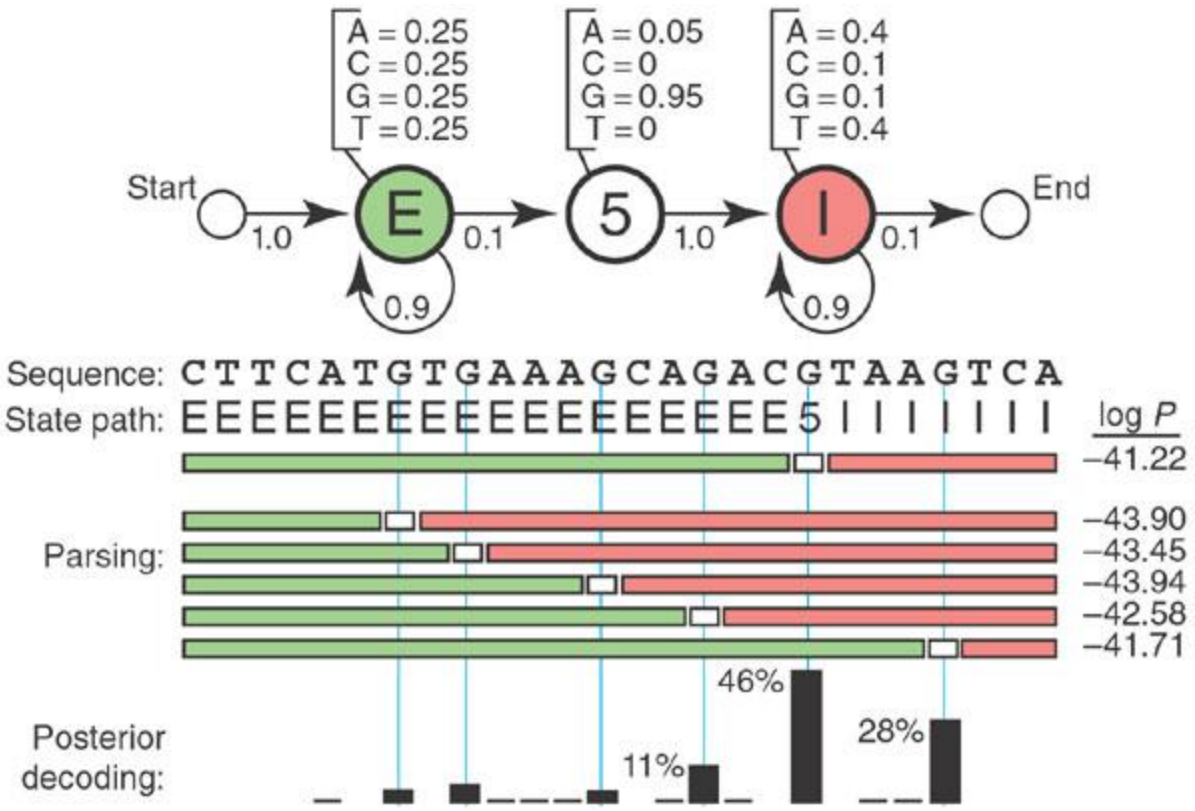
La secuencia de estados es una anotación de la cadena generada. Cada nucleótido se genera en un estado: **intergénico**, **start/stop** o **codificante**.

- **Estructura exon/intron modelada por un HMM**
  - Modelo simple que no incluye estados para señales de splicing, etc



Un ejemplo simple de HMM

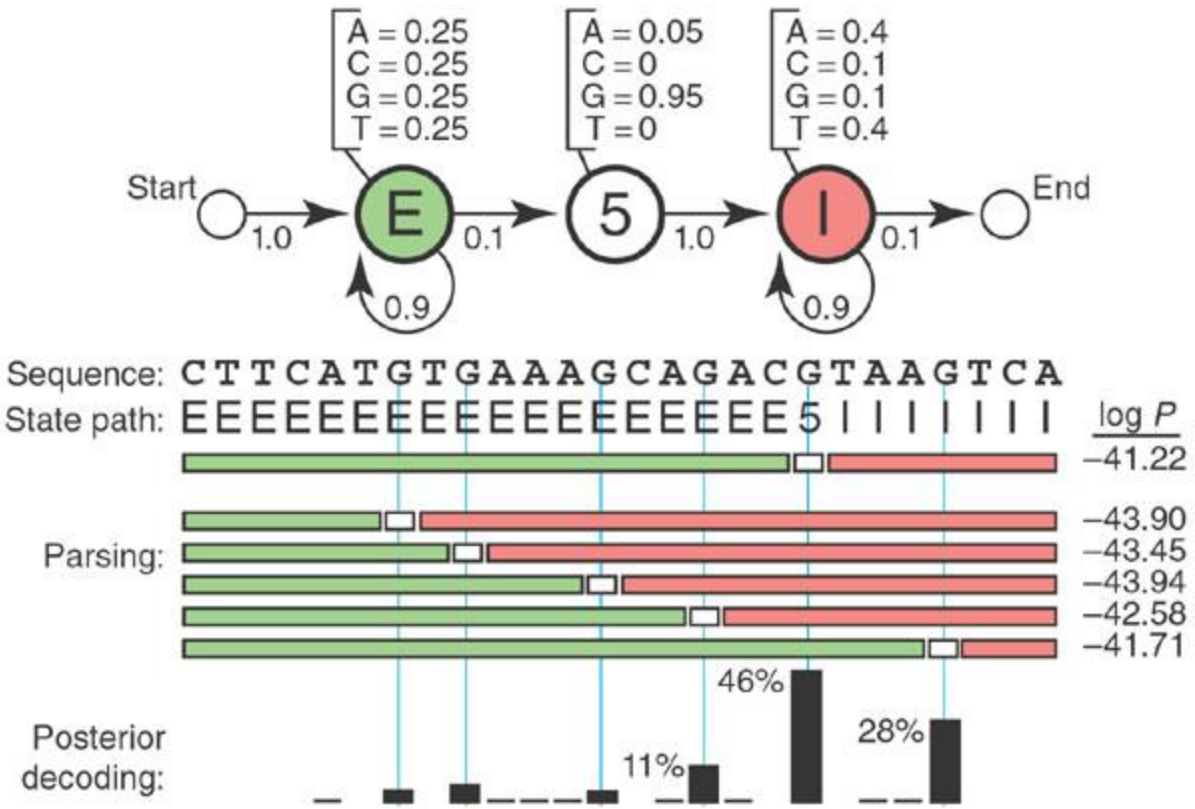
- **Reconocimiento de sitio de splicing aceptor 5'**
  - **3 estados**
  - **Probabilidades de emisión**
  - **Probabilidades de transición**
- **Modo emisión**
  - **Genera secuencia**
  - **El camino está escondido**
- **Modo analítico**
  - **Dada una secuencia**
  - **Encontrar el mejor camino**
  - **P(S,∏|HMM,Θ)**
    - La probabilidad de que un HMM con Θ parámetros y un camino ∏ genere una secuencia S dada es el producto de todas las probabilidades usadas



What is a Hidden Markov Model? S Eddy. Nature Biotechnology 22: 1315 (2004)

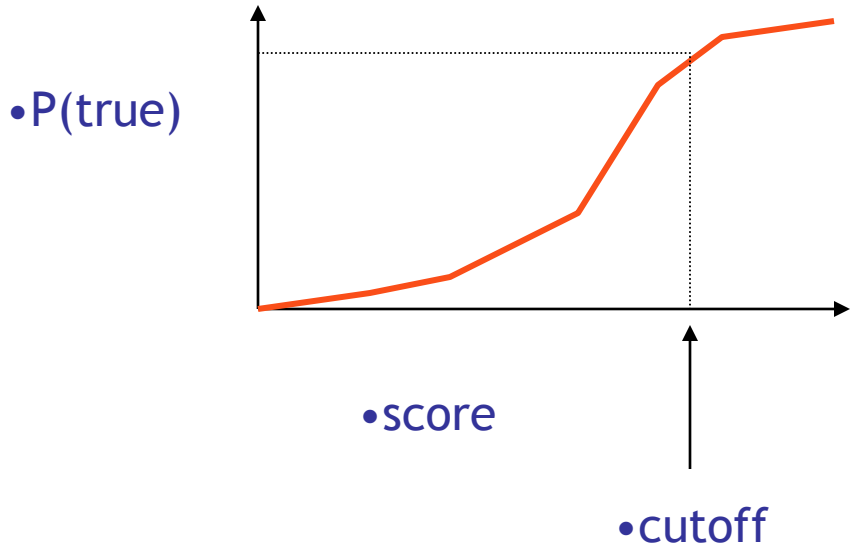
Un ejemplo simple de HMM: cont.

- **Encontrar el mejor camino**
  - **Enumerarlos todos, calculando las probabilidades correspondientes**
    - Computacionalmente intensivo
  - **O usar un método más inteligente**
    - Dynamic Programming
    - Viterbi algorithm

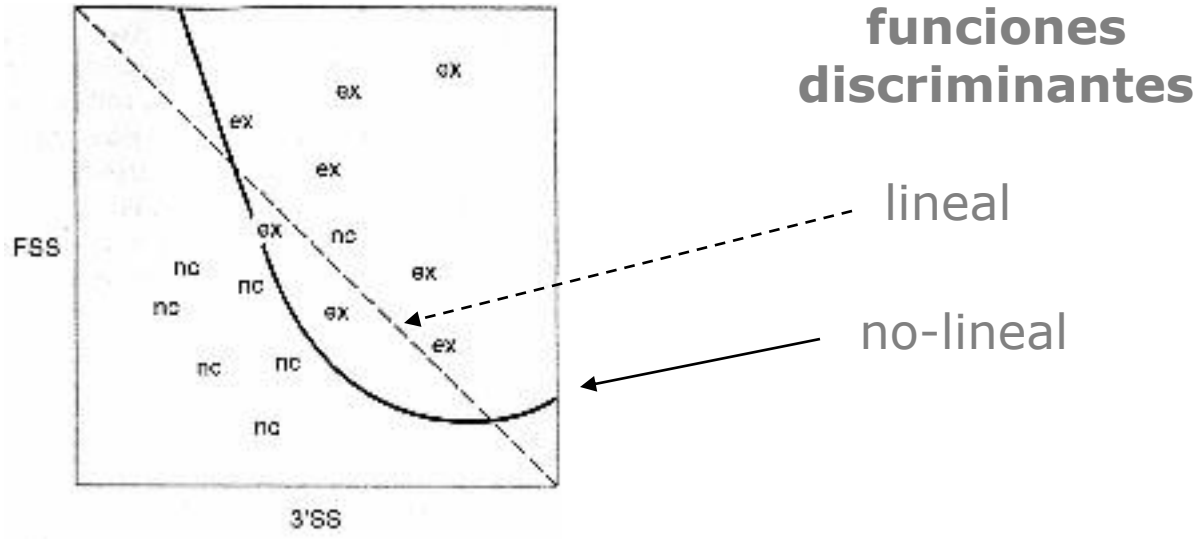


Combinar varios scores

- **Discriminant analysis**
- **Linear discriminant analysis: simplemente suma todos los scores y produce un score único**
- **O una probabilidad de que la predicción sea correcta dado un determinado score**
- **En general se ponderan diferencialmente los scores, para obtener mejores predicciones**

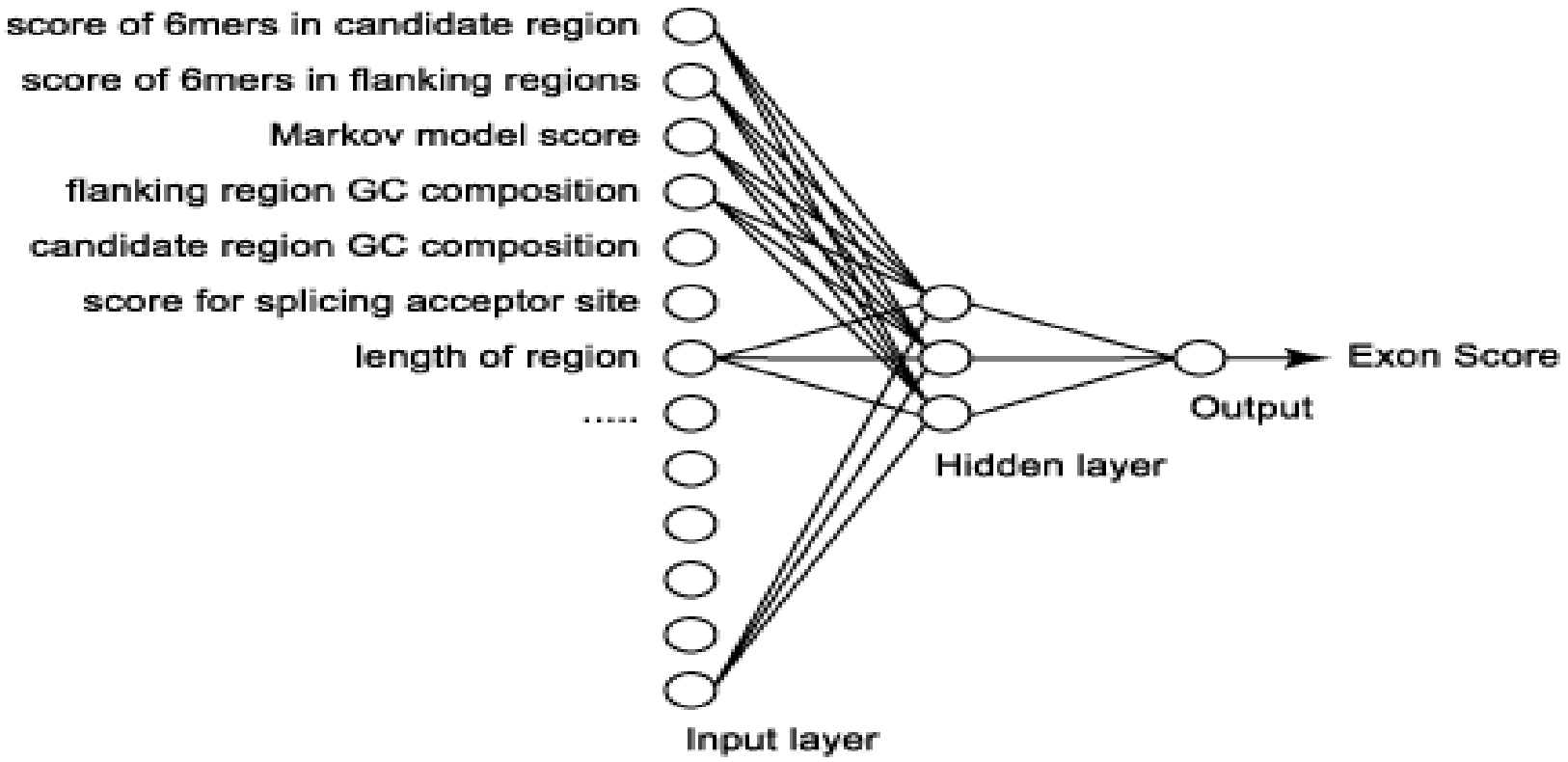


- Quadratic Discriminant analysis (usado en MZEF)



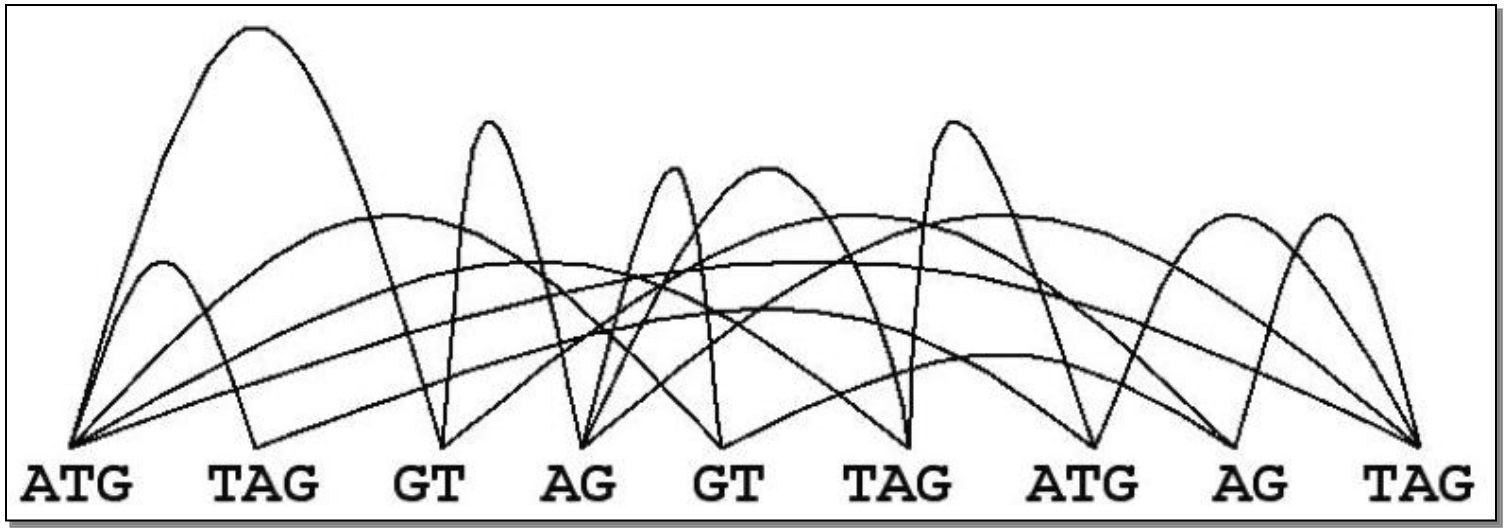
Combinar varios scores

- Usando una red neural (Grail)



Gene parses: ORF graphs

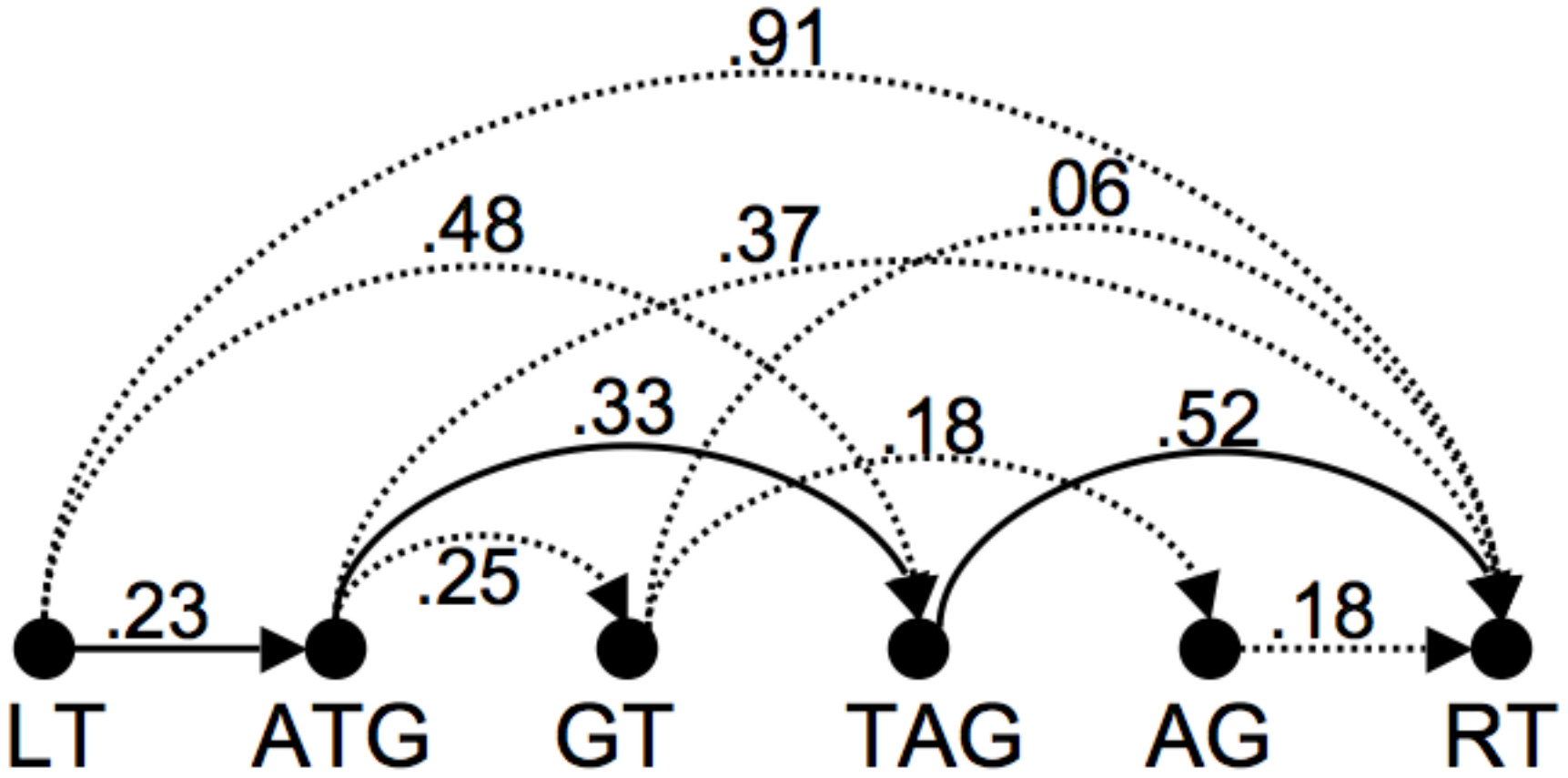
- Parse = analizar sintácticamente
- Luego de identificar señales en una región genómica, podemos usar las reglas sintácticas para construir un gen
- Conectando señales en un ***ORF graph***



- El grafo representa todos los ***gene parses*** posibles (c/u tiene un score)
- Un camino a través del grafo es un ***gene parse***

Como encontrar el mejor *gene parse*?

- Encontrar el mejor camino entre varios posibles
- Algoritmos?



Algunos ejemplos

- FGENES
  - función discriminante lineal para contenido y signal sensors y dynamic programming para encontrar la combinación óptima de exones
- GeneMark
  - http://genemark.biology.gatech.edu/GeneMark/
  - HMMs combinados con reconocimiento de RBS
- Genie
  - http://www-hgc.lbl.gov/projects/genie.html
  - redes neurales para splicing, HMMs para coding sensors. La estructura final se modela con un HMM
- Genscan
  - http://CCR-081.mit.edu/GENSCAN.html
  - weight matrix y árboles de decisión como signal sensors. HMMs como sensores de contenido. HMM para el modelo final
- MZEF
  - http://sciclio.cshl.org/genefinder
  - función discriminante cuadrática, predice sólo exones internos

Genscan

- **Desarrollado en 1997 por Chris Burge (MIT)**
- **Uno de los gene finders (*ab* initio) más precisos**
- **Modela en forma explícita la duración dentro de los estados del HMM (distintas longitudes de exones)**
- **El modelo tiene distintos parámetros para regiones con distinto contenido de GC**
- **HMMs para exones, intrones e intergénicos**
- **Weight Matrix para sitios de splicing (acceptor, branch point), polyA y promotores**
- **Decision trees para sitio donor de splicing**

Predecir genes *ab initio* es difícil

- Genes separados por regiones intergénicas largas
- Genes no son continuos, están partidos en regiones codificantes pequeñas, separadas por regiones no codificantes más largas
- Las señales (secuencias) esenciales para la identificación de la estructura de un gen son degeneradas y altamente inespecíficas
- Splicing alternativo
- Elementos repetitivos: algunos contienen regiones codificantes

Problemas

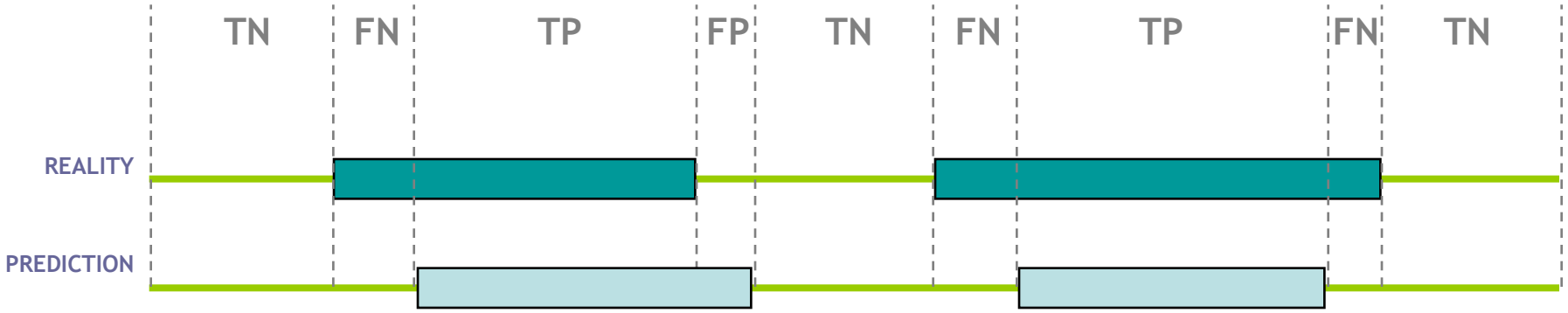
- No cuentan con evidencia biológica
- En secuencias largas, puede haber muchos falsos positivos (overprediction)
- La precisión de las predicciones es alta, pero no es suficiente

Evaluación de los resultados

- Evaluar la precisión de las predicciones
- Varios estudios
  - Burset & Guigó (1996), genes de vertebrados
  - Pavy et al. (1999), Arabidopsis
  - Rogic et al. (2001), genes de mamíferos
- Todos necesitan un set de datos (test) validado experimentalmente
  - genes para los cuales se conoce exactamente la estructura (promotor/exones/intrones) y formas de splicing

Evaluación de los resultados

- **Al nivel de la secuencia**



Sensibilidad

$$Sn \equiv \frac{TP}{TP + FN}$$

$$\frac{\text{No de exones correctos}}{\text{No total de exones reales}}$$

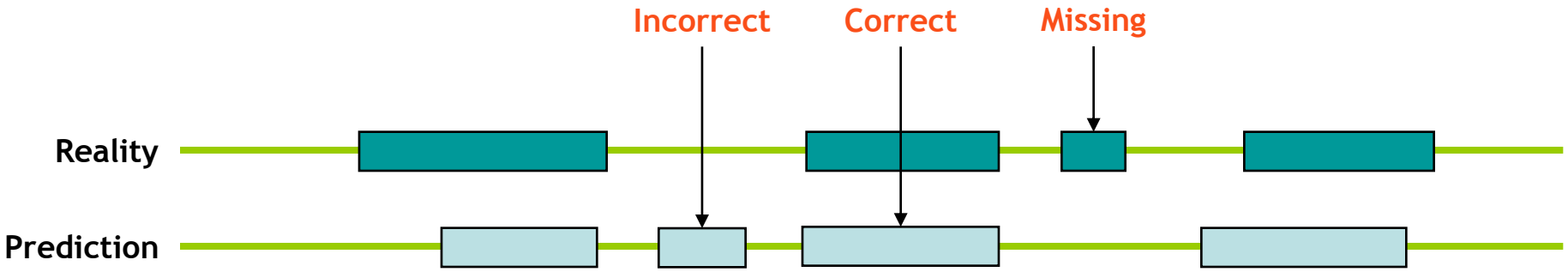
Especificidad

$$Sp \equiv \frac{TP}{TP + FP}$$

$$\frac{\text{No de exones correctos}}{\text{No total predicciones de exones}}$$

Evaluación de los resultados

- Al nivel de los exones



Sensibilidad

$$ESn \equiv \frac{C}{ER}$$

$$\frac{\text{No de exones correctos}}{\text{No total de exones reales}}$$

Especificidad

$$ESp \equiv \frac{C}{TP}$$

$$\frac{\text{No de exones correctos}}{\text{No total predicciones de exones}}$$

Evaluación de resultados

- **Rogic et al., 2001**
  - **Generación de un nuevo set de datos para validación**
  - **HMR195**

  - **Características de las secuencias**
    - human - mouse - rat
    - DNA genómico relativamente cortos tomados de GenBank
    - Un gen por secuencia
    - Se excluyeron secuencias que fueron utilizadas para entrenar a los distintos programas

Evaluación de los resultados

- **Filtrado**
  - **Codones START y STOP canónicos**
  - **Sitios de splicing canónicos (AG - GT)**
- **Dataset no redundante: secuencias similares eliminadas**
- **Confirmación de localización de exones por alineamiento con mRNA**

Resultados

Programs	# of sequences	Nucleotide accuracy				Exon accuracy							
		Sn	*Sp*	*AC*	*CC*	*ESn*	*ESp*	*(ESn+Esp)/2*	*ME*	*WE*	*PCa*	*PCp*	*OL*
FGENES	195 (5)	0.86	0.88	0.84 ± 0.19	0.83	0.67	0.67	0.67 ± 0.32	0.12	0.09	0.20	0.17	0.02
GeneMark.hmm	195 (0)	0.87	0.89	0.84 ± 0.18	0.83	0.53	0.54	0.54 ± 0.36	0.13	0.11	0.29	0.27	0.09
Genie	195 (15)	0.91	0.90	0.89 ± 0.16	0.88	0.71	0.70	0.71 ± 0.30	0.19	0.11	0.15	0.15	0.02
Genscan	195 (3)	0.95	0.90	0.91 ± 0.12	0.91	0.70	0.70	0.70 ± 0.32	0.08	0.09	0.21	0.19	0.02
HMMgene	195 (5)	0.93	0.93	0.91 ± 0.13	0.91	0.76	0.77	0.76 ± 0.30	0.12	0.07	0.14	0.14	0.02
Morgan	127 (0)	0.75	0.74	0.70 ± 0.21	0.69	0.46	0.41	0.43 ± 0.26	0.20	0.28	0.28	0.25	0.07
MZEF	119 (8)	0.70	0.73	0.68 ± 0.21	0.66	0.58	0.59	0.59 ± 0.28	0.32	0.23	0.08	0.16	0.01

Verificación adicional

- **Evaluación de los resultados en función de la secuencia y de las características de la predicción**
  - **contenido de GC**
  - **longitud de exones**
  - **tipo de exones**
  - **tipo de exones y señales presentes**
  - **probabilidad de exones y scores**
  - **especificidad filogenética**

Estrategias integradoras

- **Algunos programas integran análisis de similitud con métodos *ab initio***
  - **GenomeScan, FGENESH+, Procrustes**
- **Algunos programas utilizan la sintenía entre organismos (comparative genomics)**
  - **Rosetta, SLAM**
- **Combinar predicciones de diferentes programas (combination of experts, wisdom of crowds)**

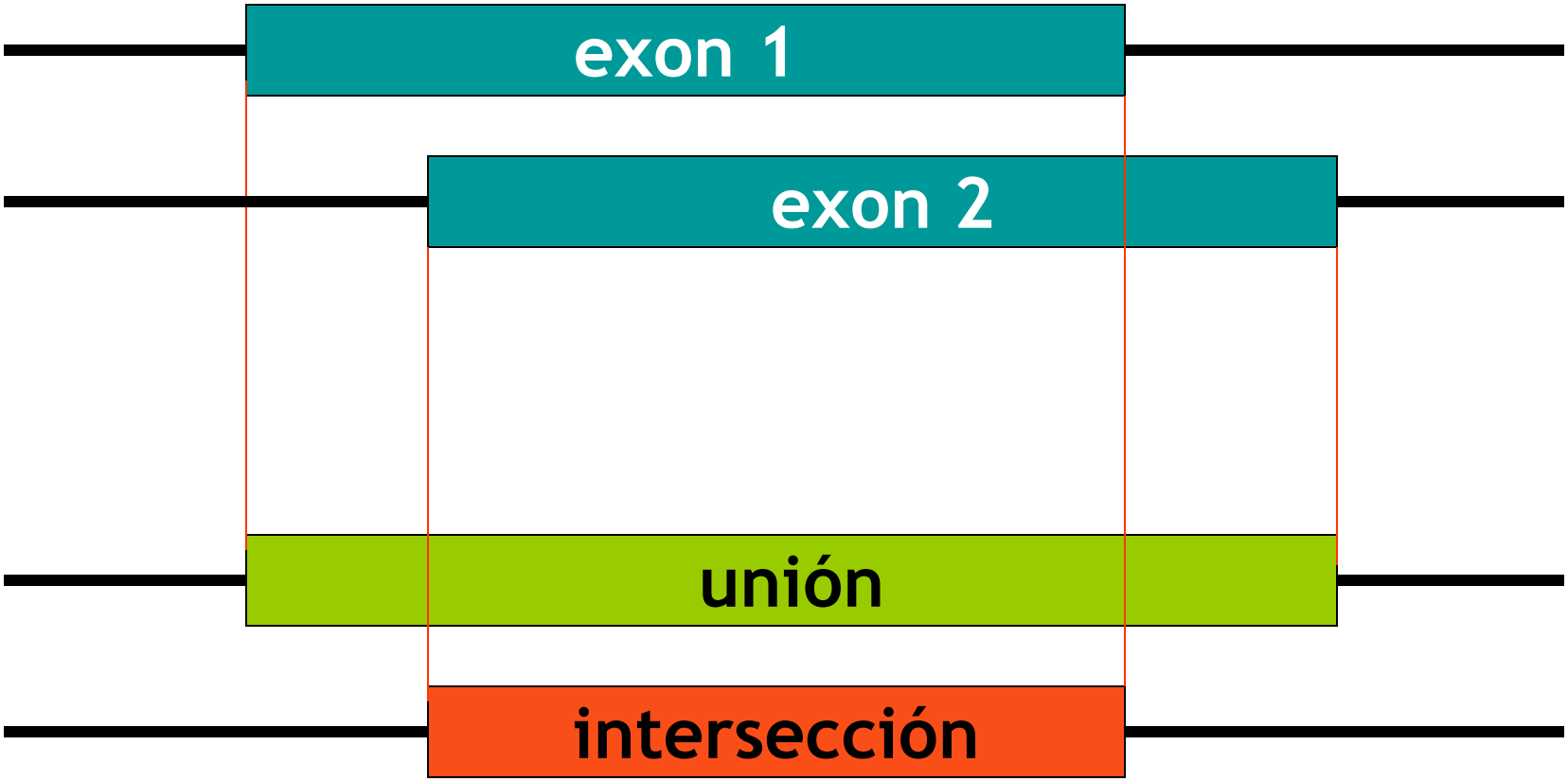
Cómo combinar las predicciones?

- Hay que usar un método

- Burset & Guigó (1996)
  - Investigaron la correlación entre 9 programas de gene finding
  - 99% de los exones encontrados por todos los programas eran correctos
  - 1% de los exones no fueron detectados por ningún programa

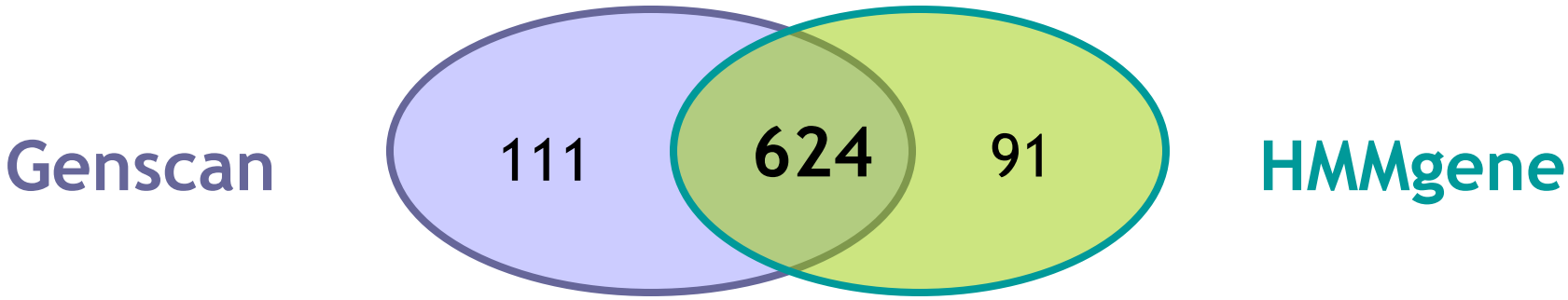
- Murakami & Tagaki (1998)
  - 5 métodos para combinar las predicciones de 4 programas

Métodos: AND vs OR



Combinar Genscan y HMMgene

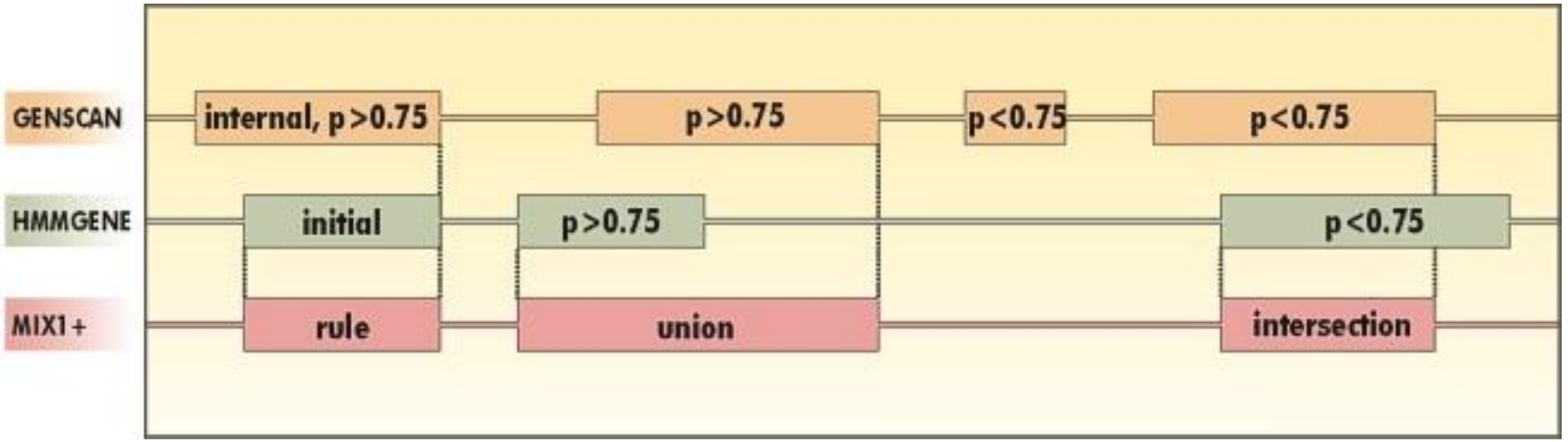
- Son los mejores candidatos: alta precisión de las predicciones



- Genscan predice el 77% de los exones correctamente
- HMMgene el 75%
- Ambos el 87%

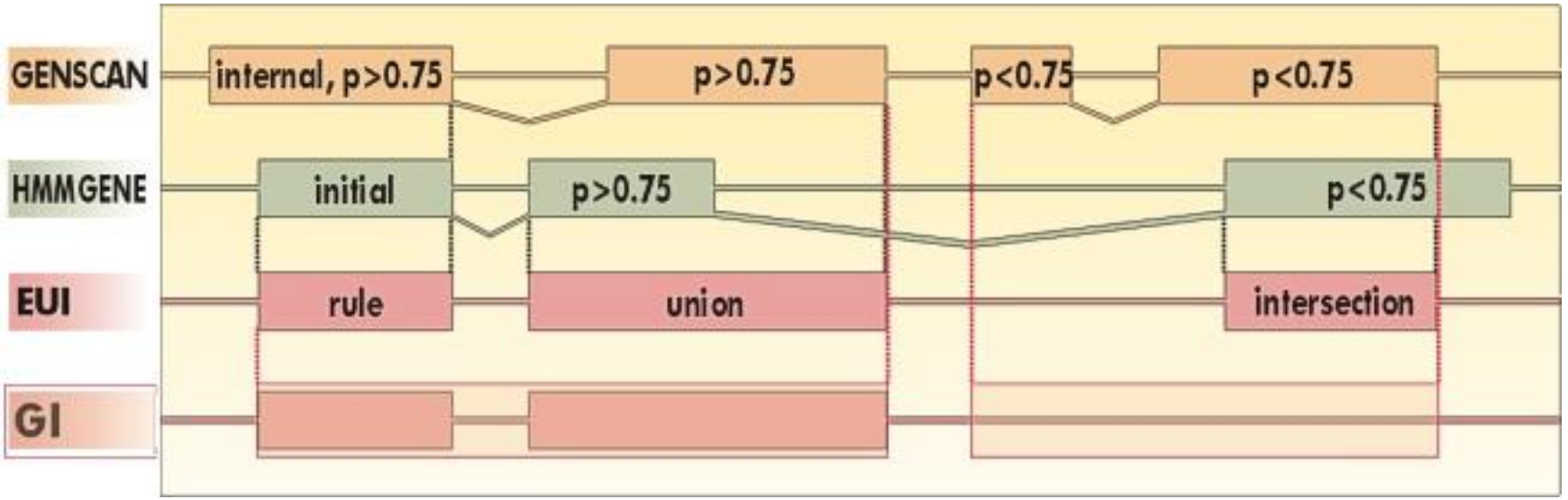
Métodos: EUI (exon union/intersection)

- Unión en exones con $p \geq 0.75$
- Intersección en exones con $p < 0.75$
- Regla especial para exones iniciales



Métodos: GI (gene intersection)

- Aplicar método EUI a exones que pertenezcan en forma completa a genes GI



Métodos: EUI frame

- EUI + reading frame consistency
- Asigna probabilidades a los genes GI. Determina la posición de sitios aceptores y donores en un marco de lectura
- El gene GI con la más alta probabilidad impone el marco de lectura. Elige los exones EUI contenidos en genes GI que se encuentran en el marco de lectura elegido

Resumen métodos de integración

- **Para el dataset HMR195**
  - **Sp incrementada 3.2%**
  - **Esn incrementada 2.6%**
  - **Esp incrementada 11.7%**
  - **El número de exones incorrectos decrece significativamente!**

METHODS	*#no prediction*	*Nucleotide accuracy*			*Exon accuracy*				
		Sn	*Sp*	*AC*	*ESn*	*ESp*	*(ESn+Esp)/2*	*ME*	*WE*
Genscan	3	0.95	0.90	0.91	0.70	0.70	0.70	0.08 (76)	0.09 (104)
HMMgene	5	0.93	0.93	0.91	0.76	0.77	0.76	0.12 (128)	0.07 (81)
EUI	3	0.94	**0.95**	**0.93**	**0.79**	**0.83**	**0.81**	0.10 (104)	**0.04** (55)
GI	15	0.91	**0.96**	**0.92**	**0.78**	**0.86**	**0.82**	0.19 (149)	**0.03** (43)
EUI_frame	3	0.93	**0.95**	**0.93**	**0.78**	**0.83**	**0.80**	0.11 (115)	**0.03** (46)

Recordar

- La mayoría de los métodos *ab initio* se entrenan sobre secuencias particulares
  - ⇒ **van a funcionar mejor en la predicción de genes similares a los del set de entrenamiento**
- Muchos métodos tienen un requerimiento absoluto de predicción de un comienzo y fin concreto para un gen
  - ⇒ **van a cometer errores frente a genes truncados o multiples genes**
- Exsiten genes que no tienen una estructura canónica
  - ⇒ **NTT (non-coding transcript in T cells), IPW (involucrada en imprinting y asociada al síndrome Prader-Willi)**
  - ⇒ **no pueden ser detectados por ningún método actual**

Recordar: Bases/Suposiciones de los métodos

- No se detectan genes solapados
- No se detectan genes anidados
- No se aceptan frame shifts o errores de secuencia
- Generalmente se anota/predice el ***parse*** óptimo
  - Dificil predecir splicing alternativo
- No se aceptan codones de inicio partidos (**ATGT..AGG**)
- No se aceptan codones de stop partidos (**TGT..AGGA**)
- No se acepta TGA como codon para selenocisteina
- No se aceptan bases ambiguas (Y, R, N, etc.)