

# Anotación de genomas

Enrique Blanco García

PID\_00165222



## Índice

<b>Introducción.....</b>	<b>5</b>
<b>Objetivos.....</b>	<b>7</b>
<b>1. Anotación del paisaje genómico.....</b>	<b>9</b>
<b>2. Modelos computacionales de señales y regiones.....</b>	<b>15</b>
<b>3. Arquitectura de los genes y sus regiones reguladoras.....</b>	<b>25</b>
<b>4. Predicción de genes <i>ab initio</i>.....</b>	<b>33</b>
<b>5. Predicción de genes por homología.....</b>	<b>45</b>
<b>6. Caracterización de regiones reguladoras.....</b>	<b>50</b>
<b>7. Impronta evolutiva de regiones reguladoras.....</b>	<b>59</b>
<b>8. Evaluación de las predicciones.....</b>	<b>65</b>
<b>Resumen.....</b>	<b>72</b>
<b>Actividades.....</b>	<b>73</b>
<b>Ejercicios de autoevaluación.....</b>	<b>74</b>
<b>Solucionario.....</b>	<b>75</b>
<b>Bibliografía.....</b>	<b>77</b>





## Introducción

La secuencia del genoma de cualquier organismo oculta astutamente distintos mensajes que involucran a la mayoría de las actividades celulares. La eficaz codificación de estas funciones mediante la oportuna combinación de diferentes señales es el resultado de millones de años de evolución natural. El correcto descifrado de toda esta información constituye la tarea fundamental de la genómica computacional. Una vez conocida la secuencia de nucleótidos de un genoma, es necesario realizar la identificación del catálogo de genes codificados en su interior. Como contenedores de la información necesaria para sintetizar los transcritos y las proteínas, no resulta extraño que los genes sean el objetivo primario de cualquier proceso de anotación de un genoma. No obstante, es importante resaltar que los genes únicamente ocupan un porcentaje ínfimo de la secuencia de los genomas eucariotas, por lo que no es posible omitir la anotación de otros elementos genómicos que también juegan un papel determinante en el funcionamiento celular.

El paisaje genómico es extremadamente rico en diferentes actores que gobiernan la activación de los genes en respuesta a numerosos condicionantes internos y externos. Desde los sitios de unión para factores de transcripción, hasta los microsatélites o los transposones, existe una vasta red de interacciones que regula el momento en el cual ciertos mensajes codificados en el genoma deben ser adecuadamente interpretados. Más allá de la secuencia de nucleótidos, la propia estructura de la fibra de cromatina del genoma posee información regulatoria adicional, empleando el intrigante código epigenético basado en la introducción de modificaciones químicas sobre las histonas de los nucleosomas que empaquetan el material hereditario.

La secuencia de cualquier genoma eucariota comprende varios millones de bases, mezclando en distintas proporciones miles de genes y otros elementos regulatorios dispersos a lo largo de grandes regiones genómicas. Dado que los mecanismos celulares de interpretación del código genético toleran un cierto grado de ambigüedad, resulta extremadamente difícil identificar con precisión el inventario de elementos funcionales. De hecho, cualquier aproximación experimental para efectuar estas búsquedas resulta prohibitiva debido a su excesivo coste y poca escalabilidad. Por tanto, para anotar con garantías cualquier genoma es necesario aprovechar la potencia de cálculo de los actuales ordenadores. Lógicamente, como contrapartida por acelerar el proceso de búsqueda, es necesario pagar un precio sobre la precisión de los resultados obtenidos computacionalmente. El grado de exactitud generalmente dependerá de la calidad de los modelos de predicción integrados en estas aplicaciones bioinformáticas. Para aumentar el porcentaje de éxito, la mayoría de estas aplicaciones permiten la integración de nuevas informaciones obtenidas a partir de búsquedas masivas en otras bases de datos o mediante comparaciones con

anotaciones existentes para otros genomas relacionados evolutivamente. En estos materiales exploraremos el problema de la caracterización del genoma desde una perspectiva computacional, enfatizando aquellos aspectos más relevantes sobre los modelos y algoritmos utilizados para identificar señales y regiones funcionales codificadas en su interior.

## Objetivos

El estudiante recibirá información sobre distintas facetas de la anotación de genomas, destacando especialmente aquellas materias relativas al análisis computacional:

- 1.** Conocer los distintos elementos del paisaje genómico de un organismo eucariota.
- 2.** Distinguir modelos para reconocer señales y regiones funcionales del genoma.
- 3.** Explicar los métodos más populares para obtener el catálogo de genes de un genoma.
- 4.** Conocer las técnicas de identificación de elementos regulatorios de los genomas.
- 5.** Introducir información sobre conservación evolutiva para mejorar las predicciones.
- 6.** Saber evaluar la calidad de las predicciones genómicas con un marco de referencia.



## 1. Anotación del paisaje genómico

La secuencia del genoma alberga toda la información suficiente para orquestar el conjunto de eventos biológicos que gobiernan la vida de nuestros organismos. Sin embargo, todos estos mensajes reconocibles por diferentes mecanismos celulares están celosamente escondidos dentro de la secuencia de nucleótidos de nuestros cromosomas. Distintos componentes funcionales codificados en el interior de la secuencia genómica están incrustados junto con otras secuencias aparecidas simplemente como resultado de los distintos procesos evolutivos que han moldeado el genoma a lo largo de millones de años. Toda esta amalgama de señales implementa en cada instante el programa genético que dota de respuesta inmediata a la célula ante cualquier suceso interno o externo.

Dentro del heterogéneo paisaje reconocible en la secuencia de cualquier genoma, el conjunto de genes codificado en su interior es el componente más relevante, dado que a partir de su identificación es posible reconstruir el catálogo de proteínas que ponen en funcionamiento toda la maquinaria celular. Existe, no obstante, una variedad significativa de otros elementos codificados en la secuencia del genoma que desempeñan también un papel relevante en diferentes procesos biológicos. Todo este abanico funcional actúa dinámicamente de forma coordinada conformando en definitiva el paisaje de cada genoma (ver figura 1):

- **Genes codificantes:** regiones genómicas con la información necesaria para sintetizar una proteína a partir de una molécula de ARN mensajero previamente transcrito.
- **Genes no codificantes:** regiones genómicas que son útiles para producir una molécula de ARN funcional que no necesariamente será traducida en una proteína.
- **Pseudogenes:** copias no funcionales de genes surgidas a partir de procesos evolutivos de intercambio de secuencias dentro del genoma.
- **Elementos regulatorios:** regiones reguladoras de la transcripción de los genes que contienen sitios de unión para diferentes proteínas.
- **Secuencias repetitivas:** repetición de un patrón de nucleótidos a lo largo de una región con fines estructurales.
- **Transposones:** secuencias genómicas incrustadas en un lugar concreto del genoma como producto de la extracción desde otra ubicación.

### Ved también

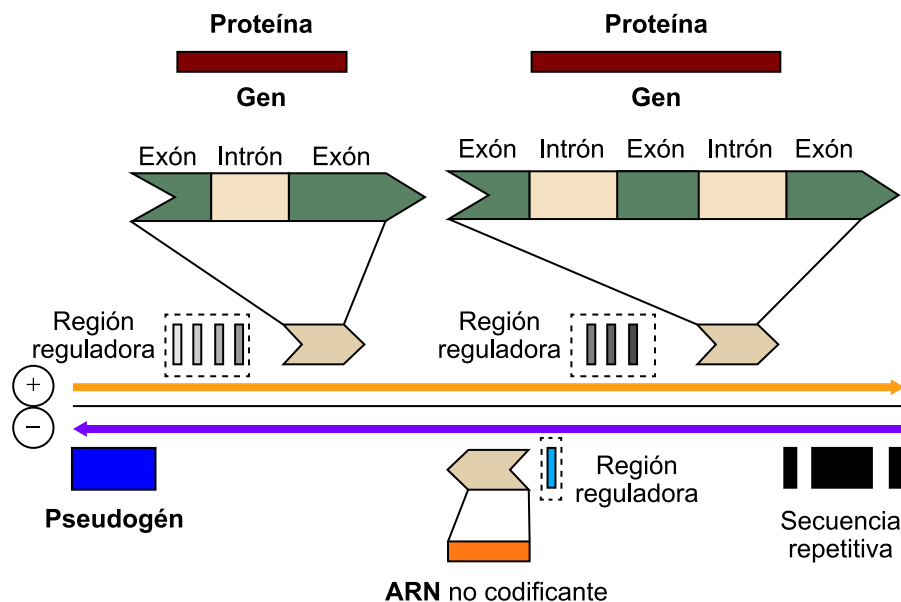
El estudiante encontrará un estudio biológico del genoma en la asignatura *Fundamentos de biología molecular*.

- **Características estructurales:** modificaciones químicas introducidas en las histonas que constituyen los nucleosomas, y que dotan a la cromatina de propiedades estructurales.

**Ved también**

Para más información sobre distribuciones de genomas y anotaciones, ver el capítulo "Servidores genómicos".

Figura 1. Elementos característicos del paisaje genómico eucariota.



Por su evidente complejidad y pese a los indudables esfuerzos internacionales de anotación, todavía desconocemos una parte sustancial del contenido de cada genoma. Para dificultar más este proceso, cada elemento funcional está estructurado internamente en varios componentes. Los genes, por ejemplo, están formados por exones e intrones, en función de la fracción de secuencia que es útil para sintetizar una proteína.

En otro orden de cosas, la comparación de la secuencia del genoma entre especies o entre individuos de la misma especie puede contribuir también a identificar características del genoma que no se han conservado universalmente (por ejemplo, las variaciones puntuales o los polimorfismos). En definitiva, para fotografiar este paisaje biológico con nitidez, serán necesarios enormes esfuerzos de análisis experimental y computacional que reconstruyan la voluminosa red de interacciones entre todos los elementos que, en tiempo real, mantienen a la célula activa.

No es posible comprender el vertiginoso progreso experimentado por la genómica en los últimos años sin la participación de las tecnologías de secuenciación o el desarrollo de potentes *pipelines* bioinformáticos. La mayoría de anotaciones elaboradas actualmente han sido obtenidas a partir del análisis computacional de predicciones sustentadas en resultados experimentales llevados a cabo en los últimos años. Esta hábil integración del conocimiento existente con los nuevos datos de secuenciación masiva mediante múltiples protocolos de anotación todavía requiere, no obstante, de la intervención humana para garantizar la calidad de los resultados.

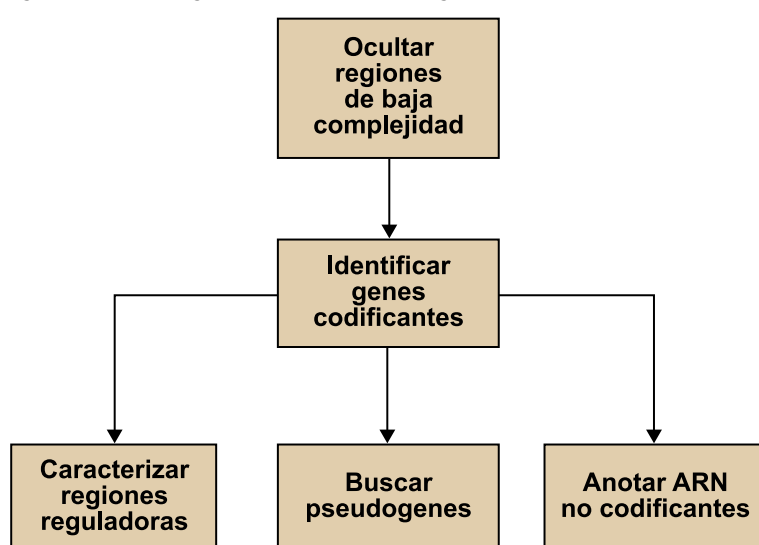
Dado que estamos hablando de miles de combinaciones posibles entre genes, regiones reguladoras, secuencias repetitivas, transposones y el resto de componentes genómicos, el uso intensivo de herramientas bioinformáticas resulta prácticamente imprescindible para interpretar esta amalgama de información. Anotar computacionalmente un genoma no es un problema sencillo: los genes ocupan únicamente el tres por ciento del genoma humano. La caracterización del resto de componentes del genoma que poseen funciones estructurales, regulatorias, evolutivas o meramente de mantenimiento de la cromatina también es fundamental para comprender los mecanismos de gobierno de la actividad génica. En resumen, distinguir regiones funcionales de meros artefactos en el interior de secuencias de millones de pares de bases es especialmente crítico.

### Lecturas complementarias

The Encode Project Consortium (2007). "Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project". *Nature* (núm. 447, págs. 799-816).

The modENCODE Project Consortium (2010). "Identification of functional elements and regulatory circuits by *Drosophila* modENCODE". *Science* (núm. 330, págs. 1787-1797).

Figura 2. Protocolo genérico de anotación de genomas.



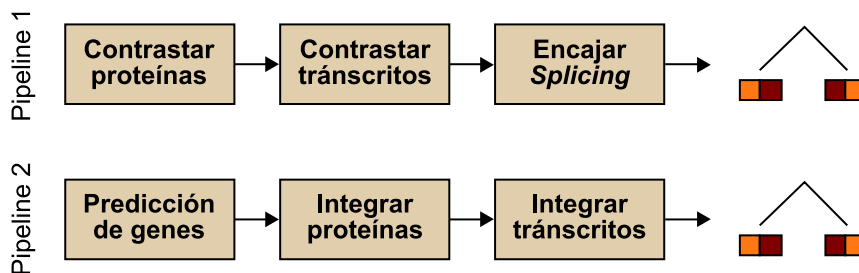
Los protocolos automáticos de anotación de genomas giran fundamentalmente en torno a la identificación del catálogo de genes. Para facilitar esta tarea, el procedimiento de anotación se inicia generalmente enmascarando aquellos elementos que presentan ciertos patrones regulares en su secuencia que no pertenecen a los genes (por ejemplo, regiones de baja complejidad o microsátélites). A continuación, puede ejecutarse el módulo de identificación de genes sobre el resto de la secuencia. Es importante disponer de dicho repositorio de referencia, dado que permite reconstruir el conjunto de proteínas de cada especie. Una vez obtenida esta primera colección de genes, otros sistemas secundarios de anotación procesan nuevamente el genoma en busca de otras características genómicas relevantes (ARN no codificantes, pseudogenes o transposones). La caracterización de las regiones regulatorias de los genes, en cambio, resulta más compleja por el reducido tamaño y la gran variabilidad de éstas. En todo caso, conocer la ubicación exacta de los genes permite refinar la búsqueda del resto de elementos (ver figura 2).

### Ved también

El estudiante encontrará abundante información sobre el diseño de protocolos automáticos en la asignatura *Fundamentos de informática en entornos bioinformáticos*.

Los *pipelines* de anotación automática realizan una primera caracterización relativamente conservadora del conjunto de genes codificados en la secuencia del genoma. La mayoría de servidores genómicos (como UCSC o ENSEMBL) posee un potente motor de búsqueda que efectúa el mapeado de las proteínas y los transcritos conocidos sobre el genoma. Una vez identificado el fragmento que es susceptible de codificar una proteína documentada, es preciso encajarlo sobre la estructura exónica más plausible. Esta anotación preliminar es capaz de identificar correctamente la mayoría de los genes conservados en otras especies. Sin embargo, esta aproximación es poco probable que encuentre aquella minoría de proteínas que no han sido todavía depositadas en esos bancos de datos. Para completar esta anotación será necesario identificar nuevos genes a partir del análisis informático de la secuencia del genoma sin acudir a colecciones externas. Esta estrategia permite identificar las posibles estructuras de ajuste codificadas en la secuencia, para validar después cuáles están soportadas por algún tipo de evidencia experimental (ver figura 3).

Figura 3. Protocolos alternativos de identificación de genes a gran escala.



### Lecturas complementarias

V. Curwen; E. Eyra; T. D. Andrews y otros (2004). "The Ensembl automatic gene annotation system". *Genome Research* (núm. 14, págs. 942-950).

F. Hsu; W. J. Kent; H. Clawson; R. M. Kuhn; M. Diekhans; D. Haussler (2006). "The UCSC Known Genes". *Bioinformatics* (núm. 22, págs. 1036-1046).

### Leyenda figura 3

En el primer protocolo buscamos por homología la ubicación de los genes para refinar después su estructura exónica. En el segundo protocolo identificamos los genes sobre la base de distintos modelos predictivos para ajustar posteriormente esta información con resultados experimentales.

Tras la obtención del catálogo inicial de genes codificado en la secuencia de un genoma recientemente ensamblado, estas anotaciones entran en un proceso de revisión manual. Existen varios consorcios internacionales de evaluadores preocupados en mejorar la calidad de esta información, cuya misión fundamental es filtrar posibles errores producidos por la anotación automática de los genes. Habitualmente, estos analistas verifican que un gen anotado posea todas las características esperables sin entrar en conflicto con otros datos, certificando además la existencia de una fuente suficiente de información experimental que proporcione soporte a dicha anotación.



El proyecto RefSeq realiza una evaluación manual de todas las anotaciones disponibles para cada gen y mantiene el mayor repositorio de transcritos génicos no redundantes existente actualmente. Este trabajo de búsqueda de transcritos de referencia en RefSeq se extiende a múltiples especies. Otros esfuerzos de anotación centran su interés en las secuencias útiles para sintetizar proteínas a partir de los genes. Por ejemplo, el consorcio CCDS mantiene la colección estándar de secuencias génicas codificantes humanas. Existen consorcios internacionales más interesados en ampliar esta colección de genes codificantes a otros grupos concretos de especies como en el caso de la colección de genes de mamíferos (en inglés, *mammalian gene collection* o MGC). En el marco del proyecto ENCODE, destinado a la anotación exhaustiva del genoma humano, también se ha producido una anotación génica de referencia denominada GENCODE. Este repositorio fue derivado a partir de la integración computacional, manual y experimental de nuevos hallazgos. Todas estas colecciones de referencia pueden ser consultadas mediante cualquiera de los navegadores genómicos convencionales. La visualización en forma de pistas permite el contraste de resultados para una fácil detección de inconsistencias entre cada sistema de anotación (ver figura 4).

Tabla 1. Consorcios internacionales de anotación manual.

Nombre	Referencia	Dirección
RefSeq	<i>Nucleic Acids Research</i> (núm. 37, págs. D32-D36) (2009)	<a href="http://www.ncbi.nlm.nih.gov/RefSeq/">http://www.ncbi.nlm.nih.gov/RefSeq/</a>
VEGA	<i>Nucleic Acids Research</i> (núm. 33, págs. D459-D465) (2005)	<a href="http://vega.sanger.ac.uk/">http://vega.sanger.ac.uk/</a>
CCDS	<i>Genome Research</i> (núm. 19, págs. 1316-1323) (2009)	<a href="http://www.ncbi.nlm.nih.gov/CCDS/">http://www.ncbi.nlm.nih.gov/CCDS/</a>
GENCODE	<i>Genome Biology</i> (núm. 7, págs. S4) (2006)	<a href="http://www.gencodegenes.org/">http://www.gencodegenes.org/</a>
MGC	<i>Genome Research</i> (núm. 19, págs. 2324-2333) (2009)	<a href="http://mgc.nci.nih.gov/">http://mgc.nci.nih.gov/</a>
ORFeome	<i>Genome Research</i> (núm. 14, págs. 2128-2135) (2004)	<a href="http://www.orfeomecollaboration.org/">http://www.orfeomecollaboration.org/</a>

### Lecturas complementarias

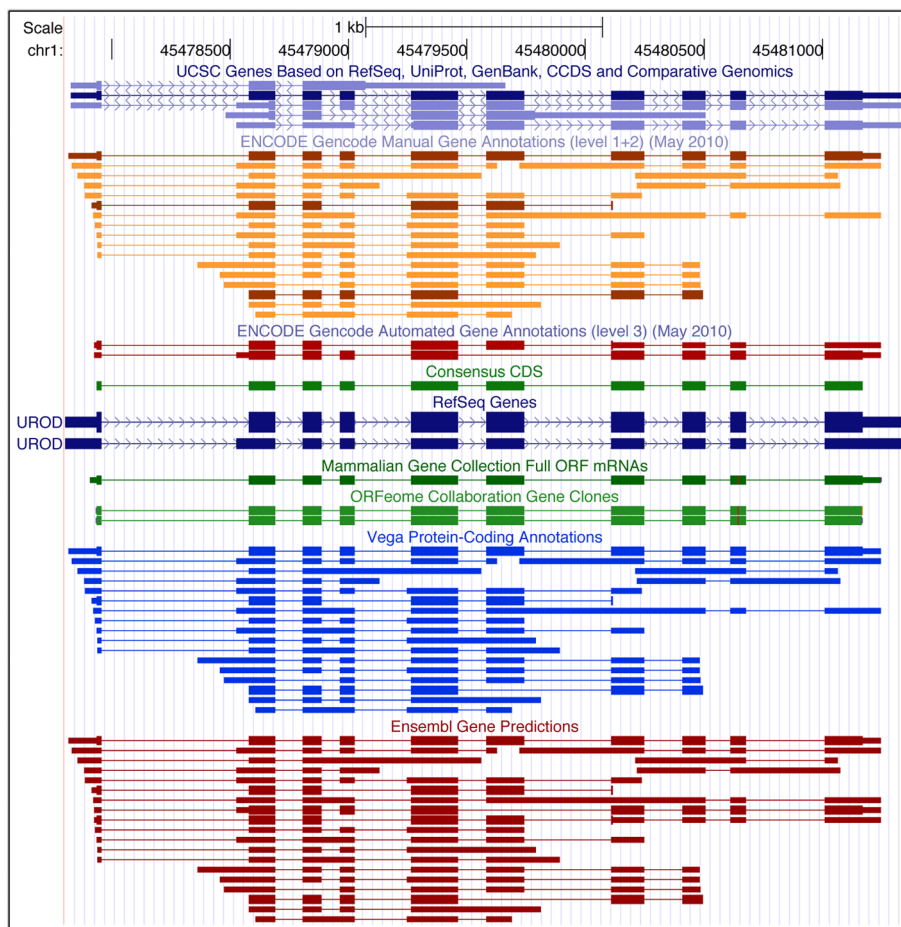
K. D. Pruitt; T. Tatusova; W. Klimke; D. R. Maglott (2009). "NCBI Reference Sequences: current status, policy and new initiatives". *Nucleic Acids Research* (núm. 37, págs. D32-D36).

K. D. Pruitt y otros (2009). "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes". *Genome Research* (núm. 19, págs. 1316-1323).

J. Harrow y otros (2006). "GENCODE: producing a reference annotation for ENCODE". *Genome Biology* (núm. 7, págs. S4).

MGC Project Team (2009). "The completion of the Mammalian Gene Collection (MGC)". *Genome Research* (núm. 19, págs. 2324-2333).

Figura 4. Anotaciones existentes sobre el gen humano UROD.



Dada la gran importancia de conocer el catálogo de genes que cualquier genoma codifica, pondremos especial énfasis en estos materiales en la caracterización computacional de las diferentes estructuras génicas. En particular, dado que es el componente esencial de cualquier protocolo automático de anotación genómica, estudiaremos con detalle cómo efectuar con garantías la identificación bioinformática de los genes que codifican la información necesaria para sintetizar proteínas. También prestaremos especial atención al modelado de las distintas señales genómicas que regulan tanto el reconocimiento como la activación de los genes en respuesta a un cúmulo de circunstancias ambientales. Únicamente analizaremos estos problemas en el contexto de los organismos eucariotas. En contraposición a los genomas procariotas, que poseen una organización genómica más sencilla, la correcta definición de estos mecanismos en eucariotas funciona gracias a un preciso solapamiento de distintos niveles de información.

#### Lectura complementaria

Para más información sobre análisis genómico en organismos procariotas:

**J. Collado-Vives y otros** (2009). "Bioinformatics resources for the study of gene regulation in bacteria". *Journal of Bacteriology* (núm. 191, págs. 23-31).

## 2. Modelos computacionales de señales y regiones

Cada célula en un organismo posee una copia idéntica del material hereditario. La interpretación que múltiples mecanismos celulares realizan de la secuencia de nucleótidos del genoma proporciona la respuesta adecuada durante el desarrollo vital del individuo. El genoma, en consecuencia, actúa como un repositorio de información que la célula utiliza gradualmente en función de los estímulos interiores y exteriores captados por diferentes sensores. Detrás de la mayoría de estos mecanismos encontramos un complejo entramado de interacciones entre genes y proteínas. Según las circunstancias, una combinación particular de proteínas puede delimitar el marco temporal y espacial en que otras proteínas deben ser sintetizadas. Para lograrlo, estas proteínas reguladoras controlan el proceso de transcripción de los genes que codifican aquellos productos péptidos necesarios. En este escenario regulatorio, los genes desempeñan un papel de simples dispositivos de almacén de la información. La maquinaria celular de transcripción, ajuste y traducción efectúa correctamente todas estas operaciones mediante múltiples mecanismos de señalización por afinidad entre los dominios de las proteínas regulatorias y ciertos fragmentos de la región genómica donde se ubican los genes en cuestión.

Una **señal** es un conjunto de nucleótidos reconocible en el interior de una secuencia genómica por distintos componentes de la maquinaria celular para activar determinados mecanismos biológicos.

La identificación del catálogo de señales que delimitan tanto los genes como sus estructuras regulatorias no es trivial. Para codificar el volumen de genes necesario para el desarrollo de las funciones celulares, es necesario incluir miles de señales en el interior de la secuencia de los cromosomas. Además, el gran tamaño del genoma, típicamente varios millones de pares de bases, propicia que los genes (y las señales que los definen) se encuentren extremadamente dispersos a lo largo de las secuencias. Para dificultar aún más su búsqueda, los distintos actores celulares que analizan la secuencia del genoma son capaces de tolerar errores de interpretación con gran eficacia. Este hecho se refleja en la existencia de variaciones en la secuencia de determinadas señales reconocidas en circunstancias similares.




Con todos estos contratiempos parece lógico que sea preciso el concurso de métodos computacionales intensivos para reconocer estas señales funcionales dentro de secuencias de ADN. El modelado estadístico apropiado de la variabilidad de las señales pertenecientes a una misma familia resulta fundamental para posteriormente reconocer con éxito esa clase de patrones en otras secuencias genómicas. Dentro del área de investigación en bioinformática se han desarrollado múltiples representaciones para sobreponerse a posibles cambios en la secuencia del mismo motivo funcional y extraer aquello que es significativamente común. Presentamos a continuación distintas estructuras de datos que permiten registrar cualitativa o cuantitativamente y con diferente nivel de detalle la composición de un determinado motivo asociado a una función genómica concreta (por ejemplo, definición de los exones de un gen o del sitio de unión en un promotor para un factor de transcripción).

### Lecturas complementarias

G. D. Stormo (2000). "DNA binding sites: representation and discovery". *Bioinformatics* (núm. 16, págs. 16-23).

A. Brazma y otros (1998). "Approaches to the automatic discovery of patterns in biosequences". *Journal of Computational Biology* (núm. 5, págs. 279-305).

Figura 5. Modelos predictivos de señales biológicas.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
S <sub>1</sub>	-	C	T	T	A	A	A	T	T	T	G	A	-	-	-	-	-
S <sub>2</sub>	G	T	T	T	A	G	A	T	T	T	C	T	T	C	G	C	-
S <sub>3</sub>	C	T	T	T	A	T	A	T	A	T	C	T	T	-	-	-	-
S <sub>4</sub>	-	-	-	-	-	G	A	T	A	A	-	-	-	-	-	-	-
S <sub>5</sub>	A	T	G	C	T	G	A	T	A	A	C	A	T	C	C	C	C
S <sub>6</sub>	A	G	G	C	-	G	A	T	A	A	A	A	-	-	-	-	-
S <sub>7</sub>	-	-	-	-	-	-	-	T	A	T	A	A	-	-	-	-	-
S <sub>8</sub>	-	-	-	-	-	T	A	T	A	T	A	A	T	-	-	-	-
S <sub>9</sub>	-	-	-	T	A	T	A	T	A	A	A	A	G	A	T	G	T
S <sub>10</sub>	C	C	G	T	A	T	A	T	A	A	A	A	G	A	T	G	T
S <sub>11</sub>	-	-	-	-	-	T	A	T	A	A	A	-	-	-	-	-	-
S <sub>12</sub>	-	-	-	-	-	T	A	T	A	A	A	A	-	-	-	-	-
S <sub>13</sub>	-	-	-	-	-	T	A	T	A	A	A	A	-	-	-	-	-
S <sub>14</sub>	-	-	-	-	-	T	A	T	A	A	A	A	-	-	-	-	-
S <sub>15</sub>	A	C	G	T	G	T	C	T	A	G	A	A	-	-	-	-	-
S <sub>16</sub>	-	-	-	-	-	T	A	T	A	G	A	A	A	-	-	-	-
S <sub>17</sub>	-	-	-	-	-	T	A	T	A	A	A	A	-	-	-	-	-
<hr/>																	
	A	C	G	T	A	T	A	T	A	A	A	A	T	A	T	G	T
<hr/>																	
	M	Y	K	T	A	K	A	T	A	W	A	A	T	M	T	S	T
<hr/>																	
A	3	0	0	0	5	1	15	0	15	10	12	12	1	2	0	0	0
C	2	3	0	2	0	0	1	0	0	0	3	0	0	2	1	2	1
G	1	1	3	0	1	4	0	0	0	2	1	0	2	0	1	2	0
T	0	3	3	6	1	11	0	17	2	5	0	2	4	0	2	0	2
<hr/>																	
A	0.92	-4.32	-4.05	-4.64	1.46	-1.88	1.88	-5.64	1.79	1.21	1.53	1.75	-0.76	0.92	-3.47	-3.47	-3.05
C	0.39	0.74	-4.05	0.00	-4.32	-5.64	-1.88	-5.64	-5.64	-5.64	-0.41	-5.64	-4.32	0.92	0.00	0.92	0.39
G	-0.55	-0.76	0.92	-4.64	-0.76	0.00	-5.64	-5.64	-5.64	-1.06	-1.88	-5.64	0.18	-3.47	0.00	0.92	-3.05
T	-4.05	0.74	0.92	1.53	-0.76	1.45	-5.64	1.97	-1.06	0.22	-5.64	-0.78	1.14	-3.47	0.92	-3.47	1.30
<hr/>																	
2																	
1																	
0																	

### Leyenda figura 5

A partir del alineamiento múltiple de la secuencia de diferentes sitios de unión del factor de transcripción TBP, construimos distintos modelos predictivos: secuencia consenso, consenso extendido, matriz de frecuencias absolutas, matriz de pesos y pictograma realizado con el programa WEBLOGO. Para el cálculo de la matriz de pesos se empleó una pseudocuenta de 0.1 unidades. Delimitamos con líneas discontinuas la fracción más informativa de la señal.

Aunque representar una determinada señal genómica con un único motivo de secuencia parece poco realista, puede resultar efectivo para resaltar inicialmente las posiciones más conservadas. Podemos derivar la secuencia consenso de un conjunto de sitios de la misma familia a partir de un alineamiento múltiple de estos (ver figura 5). El consenso puede obtenerse fácilmente registrando la base que aparece más frecuentemente en cada posición de la señal. En función del número de coincidencias entre una nueva secuencia y el consenso, podemos evaluar la similitud del nuevo candidato respecto a esta familia de señales. No obstante, cuando el conjunto inicial de señales no es extremadamente uniforme, esta representación no registra información sobre aquellas posiciones en las que existe más dispersión. Para dotar de mayor flexibilidad a los consensos podemos emplear el código extendido de 15 símbolos de la IUPAC (ver tabla 2). Esta ampliación nos permite codificar con letras especiales las posiciones más ambiguas del motivo (ver figura 5). Esta mayor capacidad de codificación puede resultar perjudicial en ciertos casos al reconocer combinaciones de símbolos que no necesariamente estaban presentes en el conjunto inicial de señales. Las secuencias consenso, en definitiva, constituyen un modelo cualitativo para representar la variabilidad de un conjunto de señales biológicas y resultan útiles para construir representaciones fácilmente comprensibles para los investigadores bioinformáticos.

La información que el alineamiento múltiple nos proporciona sobre el grado de variabilidad de los sitios genómicos pertenecientes a la misma clase de señales podría ser el reflejo de alguna propiedad biológica relevante. Interpretando la frecuencia de cada cambio observado en las posiciones del motivo como una probabilidad gobernada por una determinada distribución, podemos asumir la existencia de un modelo estadístico que moldea esta familia de secuencias. En este contexto probabilista existen métodos para reconstruir la composición del ejemplo que deseamos representar y, posteriormente, identificar nuevas ocurrencias en otras secuencias.

Tabla 2. Alfabeto genómico extendido para permitir ambigüedades

Símbolo	Nucleótidos	Significado
A C G T	A C G T	Adenine Cytosine Guanine Thymine
R Y M K	A or G C or T A or C G or T	puRine pYrimidine aMino Keto
S W B D H V	C or G A or T C or G or T A or G or T A or C or T A or C or G	Strong interaction (3 H-bonds) Weak interaction (2 H-bonds) not A, B follows A not C, D follows C not G, H follows G not T (not U), V follows U
N	A or C or G or T	aNy

### Lecturas complementarias

**R. Staden** (1984). "Computer methods to locate signals in nucleic acid sequences". *Nucleic Acids Research* (núm. 12, págs. 505-519).

**K. Frech; G. Herrmann; T. Werner** (1993). "Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids". *Nucleic Acids Research* (núm. 21, págs. 1655-1664).

Las populares **matrices de pesos** (PWM, del inglés *position weight matrices*), que ponderan específicamente el distinto uso de cada base a lo largo de las posiciones de la señal, constituyen una aproximación más dúctil que las primitivas secuencias consenso para modelar señales. En primer lugar (ver figura 5), es necesario construir una matriz de frecuencias absolutas (PFM, del inglés *position frequency matrix*) a partir del alineamiento múltiple de las señales. Estas matrices registran el número de ocasiones en que cada nucleótido aparece en distintas posiciones del motivo funcional. Resulta más adecuado normalizar estos valores absolutos calculando la frecuencia relativa de las bases en cada posición para construir un modelo probabilista que intente explicar los cambios observados entre distintas ocurrencias del mismo motivo. Sea  $F$  una PFM y  $F(b, i)$  el número de ocasiones en que la base  $b$  aparece en la posición  $i$  del alineamiento, entonces la matriz de frecuencias normalizada  $P$  debe ser calculada del siguiente modo:

Figura 6. Matriz de frecuencias normalizada.

$$P(b, i) = \frac{F(b, i)}{\sum_{A,C,G,T} F(b, i)}$$

Para aumentar la resolución de esta aproximación, es posible contrastar el modelo inicial con una descripción alternativa que generalmente representa la distribución de fondo de distintos nucleótidos a lo largo del genoma de esa especie (o en su caso, una distribución uniformemente aleatoria de bases). La razón de verosimilitud entre ambos modelos resalta mejor las posiciones conservadas en el motivo original. Sea  $Q$  una tabla con la distribución de cada nucleótido en el genoma, podemos calcular la razón de verosimilitud  $M$  entre ambos modelos  $P$  y  $Q$  mediante el cociente de valores en cada posición de la señal:

Figura 7. Razón de verosimilitud entre dos modelos.

$$M(b, i) = \frac{P(b, i)}{Q(b)}$$

Estas representaciones matriciales que modelan nuestra familia de señales nos permiten catalogar con cierta probabilidad que una nueva secuencia  $S = \langle s_1 s_2 \dots s_n \rangle$  pertenezca hipotéticamente a la misma clase. Utilizando matrices de frecuencias normalizadas o razones de verosimilitud podemos clasificar esta secuencia fácilmente:

### Lecturas complementarias

**R. Staden** (1984). "Computer methods to locate signals in nucleic acid sequences". *Nucleic Acids Research* (núm. 12, págs. 505-519).

**P. Bucher** (1990). "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences". *Journal of Molecular Biology* (núm. 212, págs. 563-578).

### Composición de riqueza

Cada organismo posee una composición de riqueza en G +C específica.

Figura 8. Clasificación de señales con razones de verosimilitud.

$$P(S) = P(s_1, 1) P(s_2, 2) \dots P(s_n, n)$$

$$M(S) = M(s_1, 1) M(s_2, 2) \dots M(s_n, n)$$

La probabilidad  $P(S)$  nos permite conocer el grado de parecido entre nuestra secuencia y esta familia de señales. Cuando disponemos de dos modelos, el valor  $M(S)$  que obtenemos de la razón de verosimilitud determina con más precisión si esta secuencia se asemeja más a un modelo o a otro (en función de si el cociente resultante es superior o inferior a la unidad). Para simplificar el cálculo podemos aplicar logaritmos, trabajando entonces con el logaritmo de la razón de verosimilitud (en inglés, *log-likelihood ratio*). En adelante denominaremos matriz de pesos a esta representación final del modelo (ver figura 5). Formalmente, la conversión a partir de la razón de verosimilitud  $M$  es la siguiente:

Figura 9. Logaritmo de la razón de verosimilitud.

$$LM(b, i) = \log M(b, i) = \log \frac{P(b, i)}{Q(b)} = \log P(b, i) - \log Q(b)$$

$$LM(S) = LM(s_1, 1) + LM(s_2, 2) + \dots + LM(s_n, n)$$

Aceptaremos o rechazaremos la hipótesis de que una nueva secuencia pertenece a la familia de motivos estudiada en ese caso según si el valor final  $LM(S)$  es positivo o negativo. Para prevenir la aparición de valores indefinidos durante la creación de las distintas matrices podemos añadir un valor arbitrario  $\alpha$  a cada posición de la PFM que no distorsione el resultado final (en inglés, *pseudocounts*, pseudocuentas). Empleando un valor suficientemente pequeño evitaremos estos problemas (por ejemplo,  $\alpha = 0,1$  en la figura 5).

Figura 10. Utilización de pseudocuentas.

$$P(b, i) = \frac{F(b, i) + \alpha}{\sum_{A,C,G,T} F(b, i) + \alpha}$$

$$\alpha = 0,1$$

Bajo la hipótesis de que cada posición de la matriz de pesos realiza una contribución independiente a la señal final, podemos asumir que las posiciones con un grado superior de conservación representan precisamente aquellas zonas más relevantes para el normal desarrollo de la actividad biológica de esta. For-

#### Lectura complementaria

W. W. Wasserman; A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements". *Nature Reviews Genetics* (núm. 5, págs. 276-287).

#### Lectura complementaria

G. D. Stormo (2000). "DNA binding sites: representation and discovery". *Bioinformatics* (núm. 16, págs. 16-23).

malmente, la cantidad de información útil en una matriz de pesos se expresa en términos de su entropía. Podemos medir esta cualidad en número de bits por símbolo a lo largo de cada posición de la matriz de pesos:

Figura 11. Entropía de un evento probabilista.

$$H(i) = - \sum_{A,C,G,T} P(b, i) \log_2 P(b, i)$$

De este modo, un conjunto de motivos de la misma familia que presenten un gran parecido exhibirán una baja entropía, mientras que un grupo de secuencias completamente diferentes poseerán un valor de entropía superior. Encontraremos la entropía máxima de dos bits en una distribución uniforme de nucleótidos escogidos al azar (donde la probabilidad de aparecer cada base es de 0,25). A mayor entropía en una distribución de motivos, menor será la cantidad de información útil para construir nuestro modelo probabilista. La cantidad de información en una matriz PFM normalizada P es:

Figura 12. Cantidad de información de una matriz de pesos.

$$I(i) = H_{MAX} - \sum_{A,C,G,T} P(b, i) \log_2 P(b, i)$$

Cuantificando la información de las matrices de pesos podemos relacionar los cambios observados en los motivos con ciertas propiedades biológicas (por ejemplo, la energía de unión entre un factor de transcripción y un motivo genómico). Las posiciones más informativas de la matriz de pesos constituyen el núcleo de estos modelos predictivos. La cantidad de información presente en una matriz de pesos suele representarse gráficamente con un pictograma donde la altura de cada símbolo es inversamente proporcional a su grado de entropía en el conjunto de señales de la misma familia (ver figura 5).

Generalmente deseamos identificar potenciales ocurrencias de una cierta familia de señales dentro de largas secuencias genómicas previamente sin anotar. Para ello, es preciso evaluar el parecido entre el perfil de la señal modelada en la matriz de pesos y cada fragmento de nucleótidos de la misma longitud que existe en el interior de la secuencia. Los algoritmos de reconocimiento de patrones (en inglés, *pattern-matching*) implementan este procedimiento mediante el desplazamiento de una ventana de una determinada longitud sobre la secuencia para evaluar cada posible ocurrencia. Únicamente se registran aquellos sitios potenciales que superan un valor umbral T de puntuación. Esta estrategia de búsqueda resulta extremadamente eficiente con un orden de crecimiento asintótico lineal.

#### Lecturas complementarias

T. D. Schneider; R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences". *Nucleic Acids Research* (núm. 18, págs. 6097-6100).

G. E. Crooks y otros (2004). "WebLogo: A sequence logo generator". *Genome Research* (núm. 14, págs. 1188-1190).

#### Lecturas complementarias

G. D. Stormo (2000). "DNA binding sites: representation and discovery". *Bioinformatics* (núm. 16, págs. 16-23).

A. Brazma y otros (1998). "Approaches to the automatic discovery of patterns in biosequences". *Journal of Computational Biology* (núm. 5, págs. 279-305).



Figura 13. Algoritmo de reconocimiento de patrones con matrices de pesos.

```

PRE  $\equiv \{S: \text{secuencias}; M: \text{matriz de pesos}; T: \text{entero};\}$ 
POST  $\equiv \{L \text{ es la lista de motivos potenciales}\}$ 
 $i \leftarrow 1;$ 
(* Evaluar el modelo en cada ventana de  $|M|$  posiciones *)
mientras ( $i \leq (|S| - |M| + 1)$ ) hacer
     $j \leftarrow i + |M|;$ 
    puntos  $\leftarrow M(S_{i,j});$ 
    (* Filtrar los mejores candidatos *)
    si (puntos  $\geq T$ ) entonces
         $L \leftarrow \text{Insertar}(S_{i,j}, \text{puntos});$ 
    fsi
    (* Evaluar el siguiente candidato *)
     $i \leftarrow i + 1;$ 
fmientras
retorna ( $L$ )

```

Figura 14. Análisis de una secuencia en busca de motivos.

...CTATAAAAT...



	1	2	3	4	5	6	7
A	-1.88	1.88	-5.64	-1.79	1.21	1.53	1.75
C	-5.64	-1.88	-5.64	-5.64	-5.64	-0.41	-5.64
G	0.00	-5.64	-5.64	-5.64	-1.06	-1.88	-5.64
T	1.45	-5.64	1.97	1.06	0.22	-5.64	-0.78

Score = -13.49

...CTATAAAAT...



	1	2	3	4	5	6	7
A	-1.88	1.88	-5.64	1.79	1.21	1.53	1.75
C	-5.64	-1.88	-5.64	-5.64	-5.64	-0.41	-5.64
G	0.00	-5.64	-5.64	-5.64	-1.06	-1.88	-5.64
T	1.45	-5.64	1.97	-1.06	0.22	-5.64	-0.78

Score = 11.58

...CTATAAAAT...



	1	2	3	4	5	6	7
A	-1.88	1.88	-5.64	1.79	1.21	1.53	1.75
C	-5.64	-1.88	-5.64	-5.64	-5.64	-0.41	-5.64
G	0.00	-5.64	-5.64	-5.64	-1.06	-1.88	-5.64
T	1.45	-5.64	1.97	-1.06	0.22	-5.64	-0.78

Score = -9.41

**Leyenda figura 14**

Mostramos el procedimiento para desplazar una ventana de evaluación de cada candidato empleando la misma matriz de pesos.

El siguiente protocolo, que involucra la construcción y el uso de matrices de pesos, permite realizar con garantías la búsqueda de señales de una determinada familia en secuencias previamente sin anotar:

- Recopilar un conjunto amplio de ejemplos funcionales validados experimentalmente.
- Alinear esas secuencias y construir la correspondiente matriz de pesos.
- Estimar el valor umbral  $T$  adecuado sobre el conjunto de entrenamiento previo.
- Identificar nuevas ocurrencias de esa señal en otras secuencias.
- Inspeccionar posibles agrupaciones de señales para construir regiones funcionales.

Las señales genómicas delimitan con precisión regiones funcionales más extensas reconocidas por determinadas maquinarias celulares. No podemos asumir, en este caso, que la contribución de cada nucleótido a la función final de la región sea independiente del resto. En las regiones codificantes de los genes, por ejemplo, observamos cierta predisposición a que aparezca un determinado codón en función de aquellos otros codones colindantes a éste, reflejando las dependencias físico-químicas que existen entre aminoácidos vecinos en el péptido final. Para capturar este sesgo en la composición genómica de las regiones debemos utilizar, en consecuencia, modelos estadísticos más potentes.

Una **región** es una secuencia de nucleótidos delimitada por dos señales biológicas de inicio y terminación que posee una determinada composición característica.

Las cadenas de Markov son el modelo probabilista más utilizado para capturar dependencias entre símbolos en el interior de las secuencias biológicas. Una cadena de Markov  $M$  es una colección de estados  $X$  que permiten codificar el lenguaje de todas las palabras de  $k$  símbolos en un determinado alfabeto. A partir de un conjunto de observaciones reales se define la función de transición  $T$  entre dos estados  $a$  y  $b$  como la frecuencia de encontrar cierto símbolo justo después de otro grupo de  $k - 1$  símbolos cuando avanzamos a lo largo de una secuencia. La elección de este valor  $k$  –denominado orden– para modelar un problema obedece a criterios motivados por algún razonamiento biológico:

Figura 15. Función de transición en las cadenas de Markov.

$$T(a, b) = P(x_i = b | x_{i-1} = a)$$

#### Lectura complementaria

R. Durban; S. Hedí; A. Crogh; G. Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids*. Cambridge: Cambridge University Press. ISBN: 0521629713.

Al inicio de la secuencia no disponemos de suficientes nucleótidos para calcular la función de transición. En este caso, por tanto, es necesario calcular arbitrariamente este valor inicial  $I$  en función del primer grupo de  $k - 1$  símbolos que aparezca en nuestra secuencia:

Figura 16. Inicialización de las cadenas de Markov.

$$I(a) = P(x_1 = a)$$

Una vez construido el autómata de estados y establecidas las funciones de iniciación y transición para viajar de un estado a otro a partir de aquellos valores observados en casos reales, esta representación modela artificialmente la familia de secuencias analizadas en función de su contenido. Con este modelo estadístico podemos evaluar cuándo una nueva secuencia posee la misma composición característica, identificando el parecido con esta clase de regiones genómicas. De este modo, teniendo en cuenta las posibles dependencias de orden  $k$ , la probabilidad de que la secuencia  $S = \langle s_1 \dots s_n \rangle$  pertenezca a la familia de secuencias definida por el modelo de Markov  $M$  debe calcularse sobre la base de su contenido como la sucesión de transiciones entre estados desde un estado inicial.

Figura 17. Evaluación de candidatos con cadenas de Markov.

$$\begin{aligned} P(S|M) &= P(x_1 = s_1)P(x_2 = s_2 | x_1 = s_1) \dots P(x_n = s_n | x_{n-1} = s_{n-1}) \\ &= I(s_1)T(s_1, s_2) \dots T(s_{n-1}, s_n) \\ &= I(s_1)\prod_{i=2}^n T(s_{i-1}, s_i) \end{aligned}$$

#### Leyenda figura 17

En este caso mostramos el ejemplo cuando únicamente tenemos en cuenta la dependencia del nucleótido anterior ( $k = 1$ ).

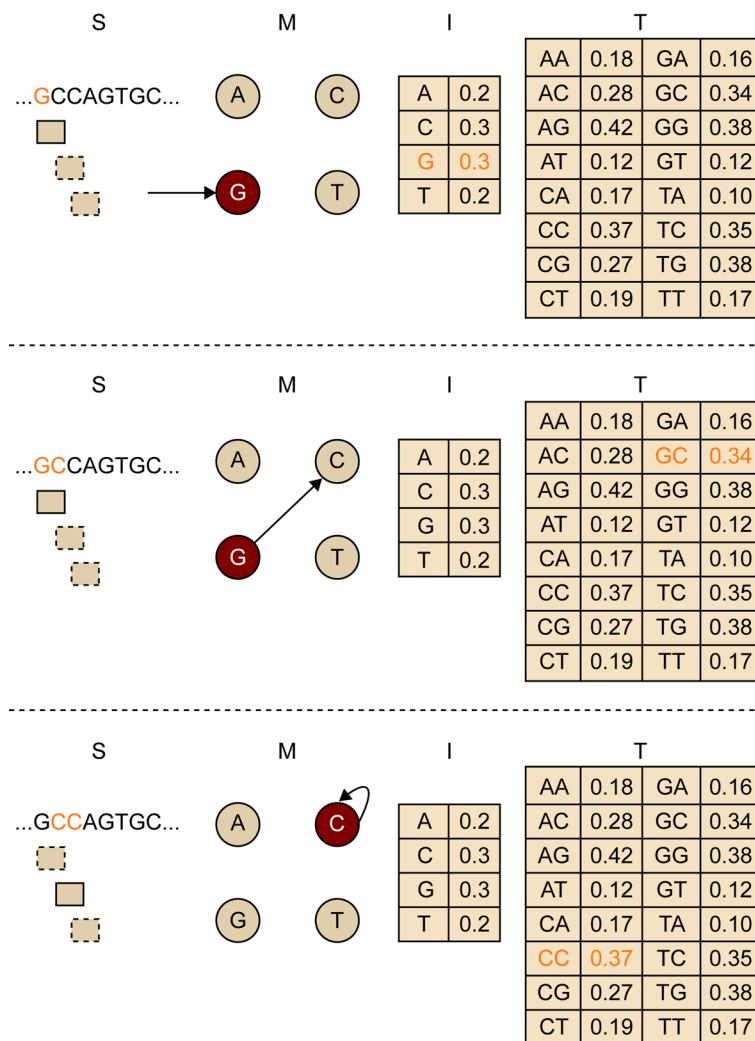
Es posible introducir un segundo modelo alternativo de composición derivado de otra clase de secuencias (por ejemplo, la frecuencia observable por azar de cada nucleótido a lo largo del genoma), añadiendo más contraste a dicho resultado. En consecuencia, el cálculo de la razón de verosimilitud entre ambos modelos ( $M+$  y  $M-$ ) puede orientarnos sobre el posible origen de nuestra secuencia. Dado que este cociente entre los parámetros de ambas cadenas está computado de antemano, aplicando logaritmos simplificaremos notablemente nuestros cálculos. En consecuencia, procederemos a recorrer la nueva secuencia, sumando los valores adecuados en función de cada transición establecida en las dos cadenas de Markov. De este modo, según el signo del valor final podemos discriminar objetivamente si la secuencia se asemeja más a uno o a otro tipo de región (ver figura 18).

Figura 18. Contraste de modelos de Markov.

$$\begin{aligned}
 \log \frac{P(s|M+)}{P(s|M-)} &= \log \frac{I^+(s_1) \prod_{i=2}^n T^+(s_{i-1}, s_i)}{I^-(s_1) \prod_{i=2}^n T^-(s_{i-1}, s_i)} \\
 &= (\log I^+(s_1) + \sum_{i=2}^n \log T^+(s_{i-1}, s_i)) - (\log I^-(s_1) + \sum_{i=2}^n \log T^-(s_{i-1}, s_i)) \\
 &= \log I^+(s_1) - \log I^-(s_1) + \sum_{i=2}^n (\log T^+(s_{i-1}, s_i) - \log T^-(s_{i-1}, s_i))
 \end{aligned}$$

En la práctica, cuando modelamos la composición de un grupo de secuencias con cadenas de Markov, podemos procesar una nueva secuencia deslizando una ventana de  $k$  nucleótidos, viajando en paralelo por el autómata de estados para obtener el resultado final. También es posible emplear esta estrategia para identificar la ubicación aproximada de determinadas regiones funcionales codificadas dentro de otras secuencias de mayor tamaño, siempre que existan diferencias significativas en su composición.

Figura 19. Análisis de regiones genómicas con cadenas de Markov.

**Lectura complementaria**

Mediante el uso de una variante denominada modelos ocultos de Markov es posible identificar con más precisión estas regiones. Para más información:

R. Durban; S. Hédí; A. Crogh; G. Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids*. Cambridge: Cambridge University Press. ISBN: 0521629713.

**Leyenda figura 19**

En este caso, hemos construido un modelo de Markov de orden 1 para detectar regiones ricas en G+C. Para contrastar la validez del resultado, puede reproducirse el mismo procedimiento con un segundo modelo para identificar regiones que no poseen esa composición.

### 3. Arquitectura de los genes y sus regiones reguladoras

Para identificar con precisión la ubicación de los genes a lo largo del genoma, es necesario definir previamente su arquitectura general. Ocultas dentro de la secuencia que contiene la información sobre un gen, se encuentran un conjunto de señales reconocibles por distintas maquinarias celulares. Únicamente la correcta interpretación de estos motivos permite el procesamiento del gen, eliminando los intrones para ensamblar la secuencia de exones correcta (ver figura 20).

La señal de inicio de la transcripción (en inglés, *transcription start site* o TSS) marca el lugar donde debe comenzar la síntesis del transcrito a partir de la secuencia genómica. El ajuste preciso de la molécula de ARN (en inglés, *splicing*) está guiado por dos tipos de señales, denominadas aceptadores y donadores, que resultan fácilmente reconocibles para el complejo de corte y empalme de los exones. En el contexto de la búsqueda computacional de genes, los métodos existentes concentran su atención fundamentalmente en el reconocimiento de las regiones codificantes para la proteína resultante (en inglés, *coding sequence* o CDS). Dentro de un transcrito maduro, la traducción de la proteína empleando el código genético comienza inexorablemente con la identificación del codón de inicio y finaliza con la detección del codón de parada. Es importante remarcar que no toda la secuencia del ARN mensajero es aprovechable para la síntesis de la proteína, existiendo en ambos extremos una región no traducible que posee funciones estructurales (en inglés, *untranslated region* o UTR). Es factible, por tanto, que los primeros/últimos exones no contribuyan en muchos casos a la región codificante del gen.

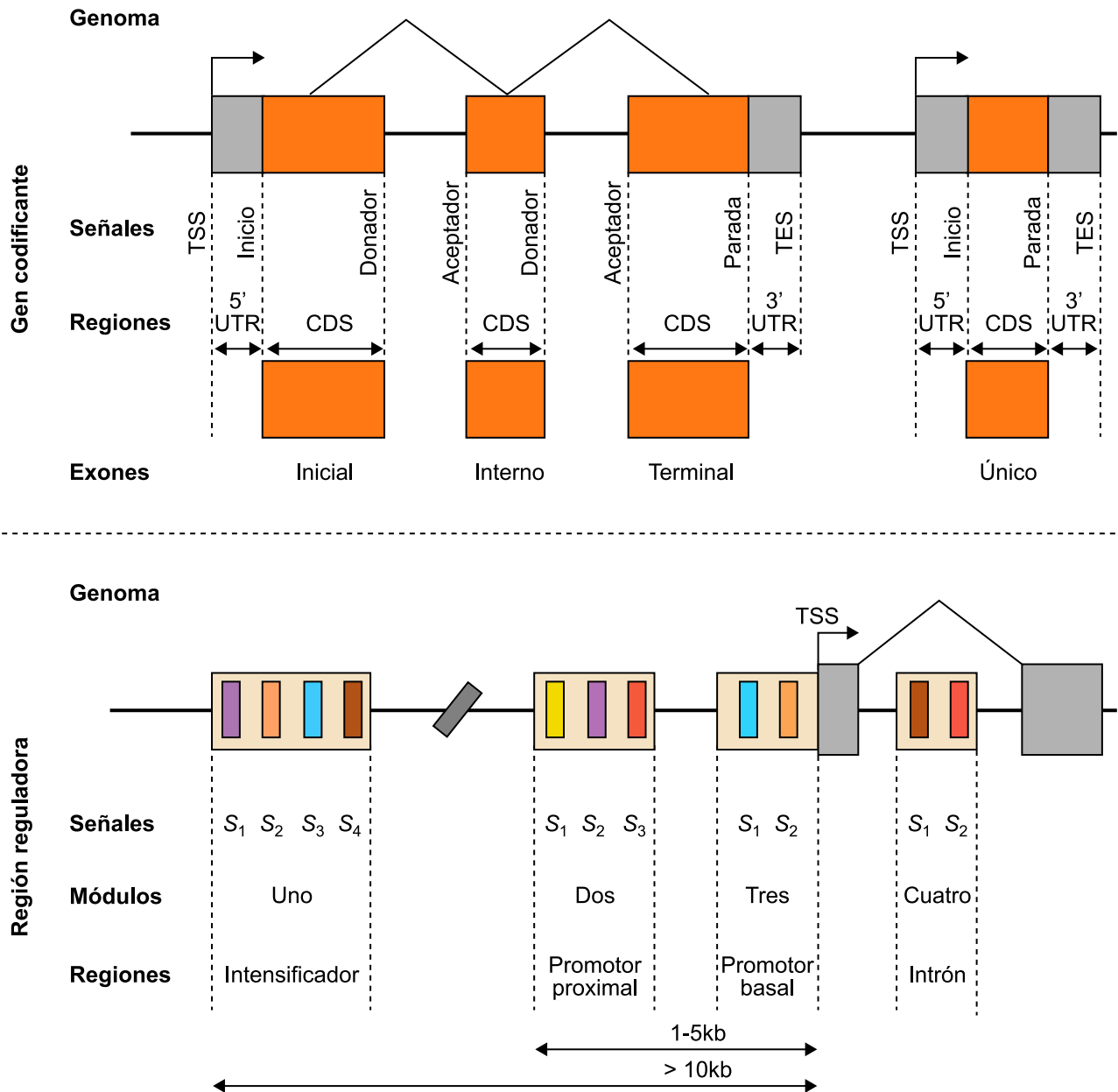
#### Ved también

El estudiante encontrará información más exhaustiva sobre la transcripción de los genes en la asignatura *Fundamentos de biología molecular*.

#### Lectura complementaria

M. Q. Zhang (2002). Computational prediction of eukaryotic protein-coding genes". *Nature Review Genetics* (núm. 3, págs. 698-709).

Figura 20. Anatomía de genes y regiones reguladoras.

**Leyenda figura 20**

Nótese que no existen limitaciones en cuanto a la cantidad de exones de un gen siempre que sea respetada su secuencia codificante. Las proteínas de unos pocos aminoácidos, por ejemplo, suelen codificarse en genes con un único exón.

En función de las señales que delimitan los distintos exones codificantes de un gen, estos pueden clasificarse en las siguientes categorías:

- **Iniciales** (en inglés, *first*): entre el codón de inicio de traducción y una señal donadora de ajuste.
- **Internos** (en inglés, *internal*): entre una señal aceptadora y otra donadora de ajuste.
- **Terminales** (en inglés, *terminal*): entre una señal aceptadora de ajuste y un codón de parada de traducción.
- **Únicos** (en inglés, *single*): entre una señal de inicio y otra de parada de traducción.

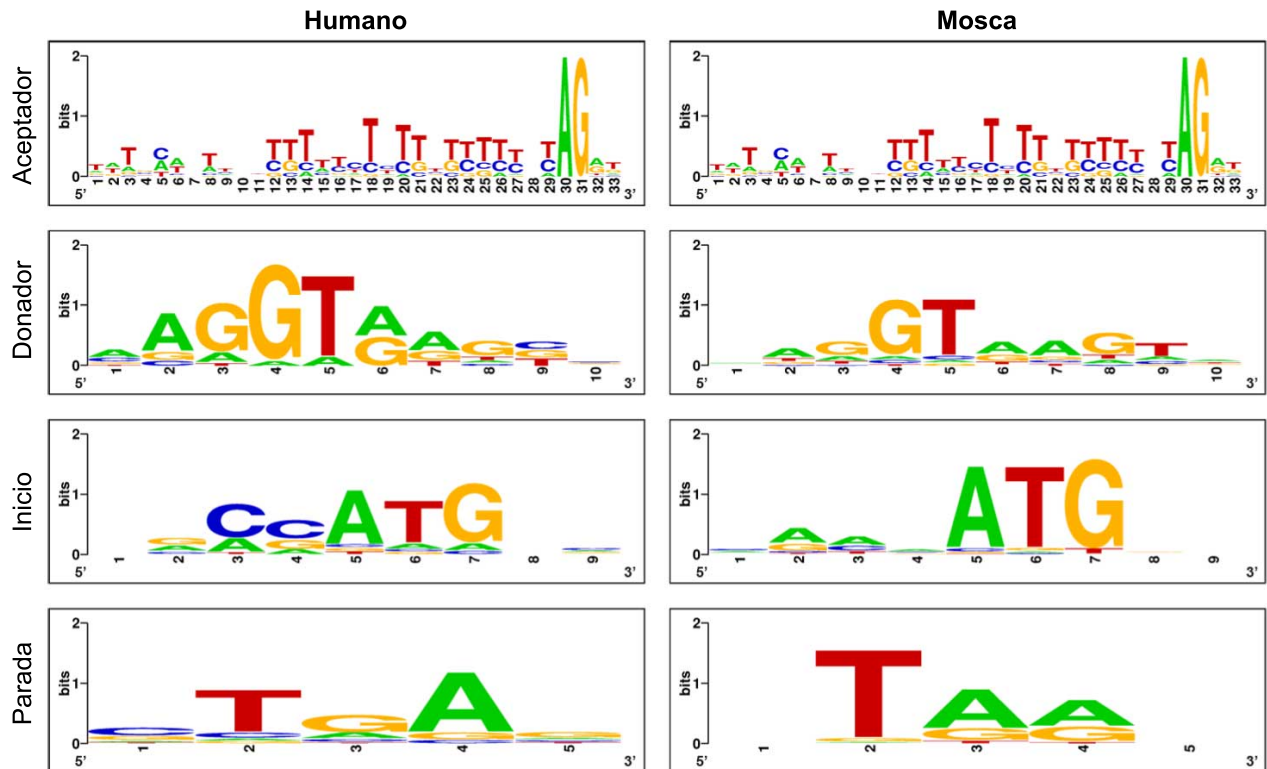
Dado que el elenco de señales de ajuste es el mismo para cualquier exón, parece lógico que para garantizar un reconocimiento más eficiente exista cierto grado de conservación en estos motivos. Comparando la secuencia de 100 exones humanos escogidos al azar (figura 21), apreciamos, por ejemplo, que la práctica totalidad de las señales aceptadoras muestran justo antes del inicio del exón el dinucleótido AG acompañado de una cola de pirimidinas (en inglés, *polypyrimidine tract*). Del mismo modo, para marcar el punto de corte entre el exón procesado y el siguiente intrón, la misma maquinaria celular reconoce el dinucleótido GT en la señal donadora de ajuste. Lógicamente, las señales de inicio y parada de traducción de estos mismos genes poseen los motivos esperables en este tipo de codones (ATG y TAA/TAG/TGA). Cuando reproducimos este estudio en 100 exones de la mosca de la fruta, observamos junto con el motivo canónico de cada señal una pequeña región flanqueante de nucleótidos conservados que suele ser específica de cada organismo.

#### Lectura complementaria

Existe documentada toda una casuística de señales de ajuste en la literatura:

**M. Bursat; A. Seledtsov; V. V. Solovyev** (2000). "Analysis of canonical and non-canonical splice sites in mammalian genomes". *Nucleic Acids Research* (núm. 28, págs. 4364-4375).

Figura 21. Señales de ajuste y traducción en 100 genes de dos organismos diferentes.



A partir de la combinación más acertada entre distintos tipos de señales génicas, podemos reconstruir el elenco de posibles exones de un gen. El número de posibilidades, no obstante, resultará generalmente prohibitivo. Afortunadamente, la secuencia codificante de los genes posee ciertas propiedades que favorecen su estudio analítico para priorizar las listas de candidatos. Por ejemplo, en la traducción de un posible exón interno que pertenece a una secuencia CDS no puede aparecer lógicamente un codón de parada que abortaría la traducción en curso. De hecho, de las tres pautas de lectura posibles (en inglés, *open reading frame* o ORF) en función de la agrupación en codones elegida para efectuar la traducción de cualquier gen, únicamente no observamos ningún codón de parada en la pauta correcta (ausencia de asteriscos sobre fondo rojo en la figura 22).

Figura 22. Traducciones de cada pauta de lectura de un mismo exón.

BQP • GEGS W O E S A R C L • • W P G S P G Q P O G H L C H T E • A A L • Q A A R G S • E L O  
 G N P K V K A H G K K V L G A F S D G L A H L D N L K G T F A T L S E L H C D K L H V D P E N F F  
 A T L R • R L M A R K C S V P L V M A W L T W T T S R A P L P H • V S C T V T S C T W I L R T S

Se ha observado también que para codificar un mismo aminoácido no se emplea cualquiera de los codones sinónimos apropiados en la misma proporción (ver columna Rel en la tabla 3). Este sesgo en la distribución de codones en el interior de regiones codificantes también se aprecia a nivel absoluto dado que existen ciertos trinucleótidos que aparecen con mayor frecuencia que otros en estas secuencias (ver columna Abs en la tabla 3). Resulta plausible pensar que estas desviaciones estadísticas son el reflejo de ciertos condicionantes bioquí-

#### Lectura complementaria

R. Guigó (1999). *DNA composition, codon usage and exon prediction*. Genetic Databases. Academic Press. ISBN: 0121016250.



micos en las proteínas resultantes. De hecho, algunos aminoácidos aparecen con más frecuencia que otros, existiendo restricciones que favorecen ciertas combinaciones entre dos aminoácidos colindantes.

Tabla 3. Tabla de uso de codones en el genoma humano.

Aa	Codón	Abs	Rel	Aa	Codón	Abs	Rel	Aa	Codón	Abs	Rel	Aa	Codón	Abs	Rel
Gly	GGG	17.08	23%	Arg	AGG	12.09	22%	Trp	TGG	14.74	100%	Arg	CGG	140.00	19%
Gly	GGA	19.31	26%	Arg	AGA	11.73	21 %	End	TGA	2.64	61 %	Arg	CGA	5.63	10%
Gly	GGT	13.66	18%	Ser	AGT	118.00	14%	Cys	TGT	9.99	42%	Arg	CGT	5.16	9%
Gly	GGC	24.94	33%	Ser	AGC	18.54	25%	Cys	TGC	13.86	58%	Arg	CGC	182.00	19%
Glu	GAG	38.82	59%	Lys	AAG	33.79	60%	End	TAG	73.00	17%	Gln	CAG	32.95	73 %
Glu	GAA	27.51	41%	Lys	AAA	22.32	40%	End	TAA	95.00	22%	Gln	CAA	11.94	27 %
Asp	GAT	21.45	44%	Asn	AAT	16.43	44%	Tyr	TAT	11.80	42%	His	CAT	9.56	41 %
Asp	GAC	27.06	56%	Asn	AAC	21.30	56%	Tyr	TAC	16.48	58%	His	CAC	14.00	59 %
Val	GTG	28.60	48%	Met	ATG	21.86	100%	Leu	TTG	11.43	12%	Leu	CTG	39.93	43 %
Val	GTA	6.09	10%	Ile	ATA	6.05	14%	Leu	TTA	5.55	6%	Leu	CTA	6.42	7%
Val	GTT	130.00	17%	Ile	ATT	15.03	35%	Phe	TTT	15.36	43%	Leu	CTT	11.24	12%
Val	GTC	15.01	25%	Ile	ATC	22.47	52%	Phe	TTC	272.00	57%	Leu	CTC	19.14	20 %
Ala	GCG	7.27	10%	Thr	ACG	6.80	12%	Ser	TCG	4.38	6%	Pro	CCG	7.02	11 %
Ala	GCA	15.50	22%	Thr	ACA	15.04	27%	Ser	TCA	196.00	15%	Pro	CCA	17.11	27 %
Ala	GCT	20.23	28%	Thr	ACT	13.24	23%	Ser	TCT	13.51	18%	Pro	CCT	18.03	29 %
Ala	GCC	28.43	40%	Thr	ACC	21.52	38%	Ser	TCC	17.37	23%	Pro	CCC	251.00	33 %

Abs: frecuencia absoluta de un codón por cada mil. Rel: porcentaje de uso relativo de cada codón sinónimo.

A la luz de los resultados arrojados por los nuevos experimentos de secuenciación masiva, parece imprescindible revisar la definición canónica de los genes. Más allá del concepto de un gen para una proteína, numerosos consorcios internacionales están anotando genes cuya estructura sobrepasa con creces los estándares arbitrariamente establecidos hace varias décadas (ver figura 23). Recientes experimentos apuntan que, por ejemplo, el ajuste alternativo de los genes sería la causa principal de la mayor complejidad de ciertos organismos en comparación con otros más simples. Aunque todavía estamos aprendiendo a conocer estos mecanismos, mediante un ajuste fino del mismo grupo de señales de ajuste, la célula puede inducir cambios sutiles que dotan a la proteína resultante de nuevas funcionalidades. En respuesta a diferentes estímulos, determinados exones serán tenidos en cuenta u omitidos durante la maduración de un transcrito. En consecuencia, podemos redefinir un gen como una colección de exones combinables de distintos modos para construir flexiblemente transcritos con propiedades específicas.

### Lecturas complementarias

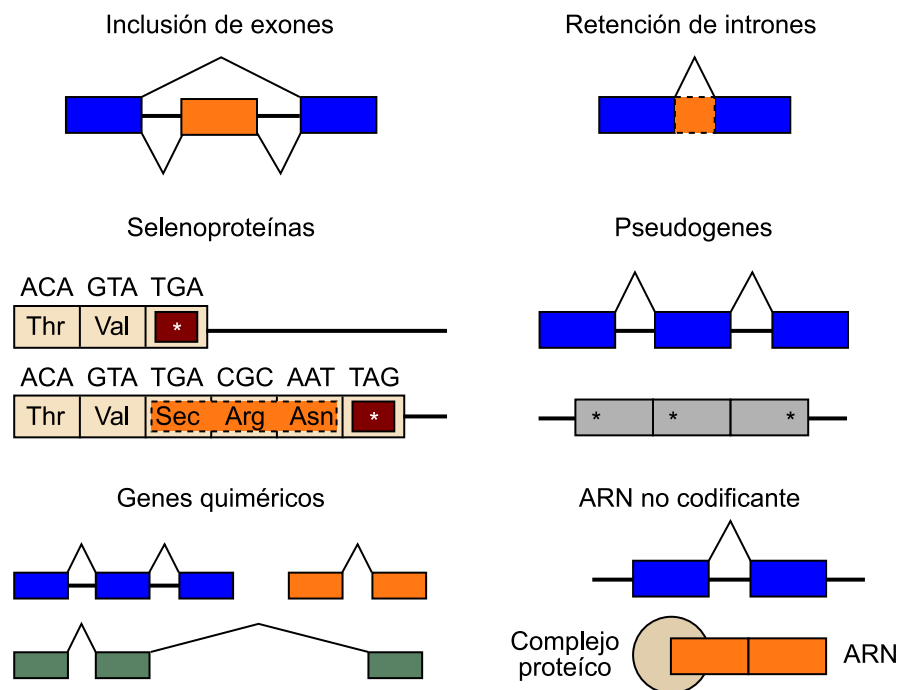
**M. Gerstein y otros (2007).** "What is a gene, post-ENCODE? History and updated definition". *Genome Research* (núm. 17, págs. 669-681).

**J. Harrow; A. Nagy; A. Raymond; T. Alioto; L. Patthy; S. Antonarakis; R. Guigó (2009).** "Identifying protein-coding genes in genomic sequences". *Genome Biology* (núm. 10, pág. 201).

**M. R. Brent (2008).** "Steady progress and recent breakthroughs in the accuracy of automated genome annotation". *Nature Reviews Genetics* (núm. 9, págs. 62-73).

Del mismo modo, se han identificado ciertas modificaciones del código genético que resultan útiles para sintetizar nuevas familias de proteínas. La recodificación del codón de parada TAG para ser traducido como el aminoácido selenocisteína confiere a las selenoproteínas funciones específicamente relacionadas con procesos de oxidación y reducción metabólicas. Incluso se ha observado cómo dos genes adyacentes pueden transcribirse de forma conjunta en ciertos casos para dar lugar a una proteína quimérica. También se conocen ejemplos de pseudogenes o copias no funcionales de genes distribuidas por el genoma que acumulan un porcentaje de mutaciones superior. La proporción del genoma potencialmente útil que conocemos ahora ha aumentado considerablemente, facilitando la identificación de numerosos genes que no necesariamente dan lugar a proteínas, sino a moléculas de ARN activas con funciones muy interesantes, como la interferencia de la transcripción de otros genes. El genoma, en definitiva, esconde numerosas sorpresas que dificultan la anotación correcta de los genes.

Figura 23. Estructuras génicas no canónicas.



La expresión de un gen produce la aparición de una característica observable que probablemente es debida al efecto de la proteína codificada por éste. La regulación transcripcional es el mecanismo primario que finalmente determina la cantidad de producto péptido que debe ser sintetizado. En cada instante, únicamente un subconjunto de genes se expresa en las células de un determinado tejido, introduciéndose drásticos cambios en esta configuración durante el ciclo de vida de éstas. El exquisito solapamiento de distintos niveles de control permite controlar con suma eficacia todo este proceso.

A nivel local, las regiones promotoras de un gen se comportan como interruptores accionados por distintos factores de transcripción para el reclutamiento de la ARN polimerasa precisamente en el punto donde debe iniciarse la transcripción del gen (ver figura 20). Los factores de transcripción son proteínas que poseen un dominio para detectar ciertos motivos de ADN que permiten la unión estable entre ambas moléculas. Estos sitios de unión son secuencias de 5 a 15 nucleótidos habitualmente reconocibles por más de una clase de factores de transcripción. De hecho, el mismo factor puede unirse con distinta afinidad a secuencias ligeramente diferentes (ver figura 5). La competencia entre dos factores por la misma región reguladora permite implementar mecanismos de control más complejos sobre el mismo gen. Por contra, la cooperación para activar sitios vecinos fomenta la formación de complejos proteicos entre varios factores que constituyen un módulo regulatorio. Pese a que los factores de transcripción contactan generalmente con regiones cercanas al inicio de transcripción del gen, están documentados numerosos casos donde esta actividad reguladora se desarrolla plenamente a miles de nucleótidos de distancia, gracias a plegamientos introducidos en la propia molécula de ADN.

### Lecturas complementarias

G. Gras; M. Hahn; E. Abouheif; J. Balhoff; M. Pizer; M. Rockman; L. Romano (2003). "The evolution of transcriptional regulation in eukaryotes". *Molecular Biology and Evolution* (núm. 20, págs. 1377-1419).

L. Barrera; B. Ren (2006). "The transcriptional regulatory code of eukaryotic cells—insights from genome-wide analysis of chromatin organization and transcription factor binding". *Current Opinion in Cellular Biology* (núm. 18, págs. 291-298).

Distintas regiones reguladoras gobiernan la transcripción de un gen:

- **Promotor basal** (en inglés, *core promoter*). Región ubicada inmediatamente antes del inicio de transcripción. Es la zona que reconoce directamente la ARN polimerasa mediante el enlace con factores generales de transcripción (por ejemplo, la caja TATA). Es responsable de un nivel basal de la transcripción del gen, precisando de factores adicionales para aumentar su productividad.
- **Promotor proximal** (en inglés, *proximal promoter*). Región genómica de unos pocos miles de bases adyacente al promotor basal del gen. Posee sitios de unión para ciertos factores de transcripción que resultan fundamentales en la activación del gen bajo el control de una determinada red regulatoria.
- **Intensificador** (en inglés, *enhancer*). Región del genoma ubicada en una localización distinta del gen, que posee una concentración superior de sitios de unión para factores de transcripción, potenciando la tasa de transcripción de éste.
- **Intrones**. Región interna del gen que en determinadas especies alberga sitios de unión para proteínas próximos a su punto de inicio de transcripción. Funciona de forma similar al promotor proximal, resultando en algunos casos más decisivo que la propia región promotora.

A nivel global, existen indicios que permiten sospechar de la existencia de una compleja red de interacciones regulatorias que trabaja por encima de los mecanismos de control locales. Los genomas eucariotas están altamente compartimentalizados: existen dominios cromosómicos que agrupan centenares de genes con una estructura de cromatina similar. En función del tipo celular y la etapa del desarrollo, cada configuración estructural permite un acceso más o menos restringido de los factores de transcripción a las regiones reguladoras de los genes. El posicionamiento de los nucleosomas resulta clave, por tanto, para explicar la expresión coordinada de un grupo de genes en un intervalo de tiempo concreto.

Sabemos que los nucleosomas pueden desplazarse para dejar paso a los factores de transcripción en determinadas circunstancias. Recientes descubrimientos muestran la existencia de un código de las histonas que constituyen los nucleosomas para implementar todos estos cambios sobre la estructura de la cromatina. Más fascinante aún resulta que determinados factores de transcripción parecen ser los responsables de inducir la protección o desprotección de estas regiones cromosómicas mediante la interacción con las histonas. Incluso más intrigante es la cada vez más reconocida simbiosis entre el proceso de transcripción y el de maduración de las moléculas de ARN. Desde la comunidad científica se postula mayoritariamente que todos estos procesos regulatorios están conectados gracias a la participación de los mismos actores en distintos contextos.

#### Lecturas complementarias

- E. Blanco; M. Pignatelli; S. Beltran; A. Punset; S. Perez-Lluch; F. Serras; R. Guigó; M. Corominas (2008). "Conserved chromosomal clustering of genes governed by chromatin regulators in *Drosophila*". *Genome Biology* (núm. 9, pág. R134).
- T. Kouzarides (2007). "Chromatin modifications and their function". *Cell* (núm. 128, págs. 693-705).
- A. Kornblihtt (2005). "Promoter usage and alternative splicing". *Current Opinion in Cell Biology* (núm. 17, págs. 262-268).

## 4. Predicción de genes *ab initio*

En cualquier genoma podemos apreciar diferencias notables en la estructura de sus genes (por ejemplo, número de exones, longitud de los intrones, tamaño de la proteína, distancia intergénica, etc.). Estas divergencias se acentúan cuando comparamos genes pertenecientes a distintas especies. No obstante y a pesar de que cada gen posea su propia arquitectura, podemos identificar numerosas características comunes en sus exones que resultan útiles para identificar otras regiones similares. Mediante el estudio analítico de una colección de exones podemos efectivamente modelar potentes sensores de señales y de contenido.

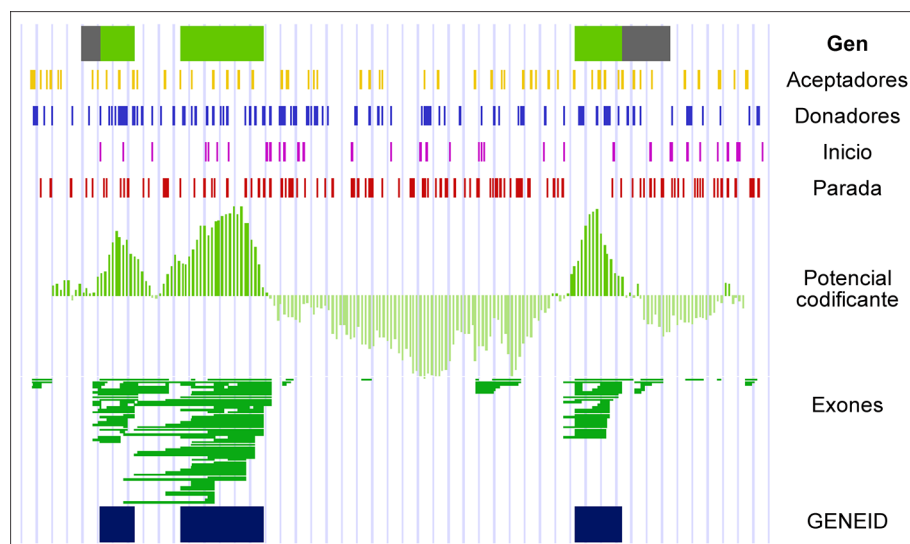
La **predicción *ab initio* de genes** (denominada también *de novo* o intrínseca) es una técnica computacional basada en la construcción de modelos estadísticos a partir del estudio de abundantes colecciones de exones reales para capturar la composición de las señales y las regiones codificantes internas sin emplear información externa derivada del uso de bases de datos de homologías.

A pesar de la bondad de los modelos estadísticos existentes, en el interior de cualquier secuencia genómica escogida al azar podemos localizar cientos de señales que no necesariamente son funcionales. Estudiemos a modo de ejemplo el gen humano *HBB* ilustrado en la figura 24. Dicho gen posee tres exones codificados en una secuencia de aproximadamente 2.000 nucleótidos. Empleando sensores de señal entrenados para capturar las secuencias que delimitan los exones, podemos detectar 61 aceptadores, 96 donadores, 34 codones de inicio de traducción y 103 codones de parada. La combinatoria nos permite construir miles de exones con estas señales, reduciendo esta cantidad hasta 538 exones empleando sensores de contenido basados en el uso sesgado de codones.

### Lecturas complementarias

**E. Blanco; R. Guigó** (2005). *Predictive methods using DNA sequences Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Nueva York: John Wiley & Sons Inc. ISBN: 0471478784.

**D. Haussler** (1998). "Computational genefinding". *Trends in Genetics (Trends guide to bioinformatics)* (págs. 12-15).

Figura 24. Identificación computacional del gen *HBB*.

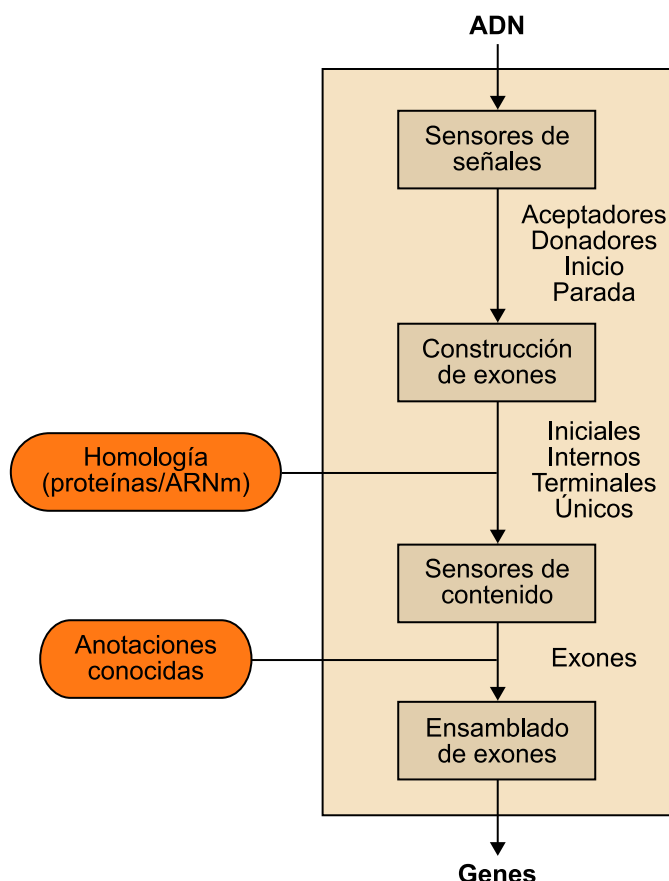
Disponemos ahora de un inventario de 70 exones iniciales, 278 exones internos, 179 exones terminales y 11 posibles genes sin intrones (por simplicidad, sólo mostramos los más plausibles en la figura 24). Para identificar correctamente el gen *HBB* deberíamos ser capaces de ensamblar precisamente los tres exones reales, descartando en consecuencia el resto de miles de posibles combinaciones. El estudiante debe ser consciente de que estos números se han contabilizado conociendo de antemano que dicho gen está codificado en la hebra positiva, lo cual simplifica notablemente el problema original. En un caso real no disponemos necesariamente de tantos datos, por lo que deben ser analizadas ambas hebras. También puede ocurrir que estemos trabajando con una secuencia que posee genes mezclados en ambos sentidos de la molécula de ADN, complicando aún más la búsqueda. En este contexto parece claro que resulta fundamental el concurso de aproximaciones computacionales para identificar automáticamente los genes codificados en una secuencia.

Para ilustrar el funcionamiento de un programa de predicción de genes *ab initio*, en estos materiales vamos a utilizar una aplicación bioinformática real. GENEID fue uno de los programas pioneros en el desarrollo de métodos estadísticos para identificar genes en secuencias anónimas. La arquitectura actual conserva reminiscencias de la versión original, aunque posee numerosas mejoras internas en su implementación que permiten anotar genomas completos en pocas horas. GENEID combina fundamentalmente la acción de sensores de señal para identificar los sitios de ajuste, sensores de contenido para evaluar el potencial codificante en base al uso de codones observado dentro de un exón y un potente algoritmo de ensamblado de genes basado en el uso de programación dinámica.

#### Lecturas complementarias

- R. Guigó; S. Knudsen; N. Drake; T. Smith (1992). "Prediction of gene structure". *Journal of Molecular Biology* (núm. 226, págs. 141-157).
- E. Blanco; G. Parra; R. Guigó (2003). *Using geneid to identify genes. Current Protocols in Bioinformatics*. Nueva York: John Wiley & Sons Inc. ISBN: 0471250937.
- G. Parra; E. Blanco; R. Guigó (2000). "GeneID in *Drosophila*". *Genome Research* (núm. 10, págs. 511-515).

Figura 25. Arquitectura del programa GENEID.



El fichero de parámetros de GENEID contiene los valores optimizados para una especie determinada de los distintos modelos estadísticos que identifican señales de ajuste/traducción y regiones codificantes. También incluye las reglas para ensamblar correctamente diferentes clases de exones dentro de un gen.

Los programas de predicción *ab initio* de genes buscan parecidos entre la composición de las regiones analizadas y un diccionario interno que contiene las distribuciones más frecuentes en exones conocidos de cada especie. Estos conjuntos de sensores de señal y de contenido están encapsulados dentro de un fichero de parámetros. Dado que el genoma de cada grupo taxonómico de especies posee su propia distribución estadística de señales y regiones codificantes, es preciso trabajar con el fichero de parámetros adecuado en cada caso. Lógicamente, todos estos modelos estadísticos son parametrizables, por lo que es necesario que los desarrolladores del software hayan entrenado el programa con conjuntos de exones almacenados en las bases de datos para capturar con precisión los rasgos específicos de cada genoma. Es fundamental que el usuario de estas aplicaciones seleccione el fichero más apropiado de acuerdo con la secuencia de trabajo (tabla 4).

Tabla 4. Varios ficheros de parámetros incluidos en la distribución de GENEID.

Nombre común	Especie	Fichero
Humano	<i>Homo sapiens</i>	human3iso.param
Mosca de la fruta	<i>Drosophila melanogaster</i>	dros.param
Abeja	<i>Apis mellifera</i>	amel.param
Ascidia	<i>Ciona intestinalis</i>	cintestinalis.param
Levadura	<i>Saccharomices cerevisiae</i>	yeast.param
Tomate	<i>Solanum lycopersicum</i>	slycopersicum.param
Arroz	<i>Oryza sativa</i>	rice.param
Arabidopsis	<i>Arabidopsis thaliana</i>	arabidopsis.param

Los sensores de señales (aceptador, donador, inicio y parada de traducción) en GENEID están implementados mediante cadenas de Markov. A partir de una muestra razonablemente extensa de exones anotados en cada especie, fueron calculados los valores óptimos para distinguir la distribución de nucleótidos característica alrededor de cada señal. Para aumentar el contraste, se escogió otra muestra suficientemente amplia de regiones no exónicas para identificar falsos positivos de cada clase. Estas secuencias poseen los motivos canónicos de cada señal (ver figura 21), pero carecen de un contexto funcionalmente conservado. En base a los resultados obtenidos en diferentes pruebas de optimización, los diseñadores del programa escogieron para cada tipo de señal una determinada longitud, introduciendo un orden diferente en el modelo de Markov en cada caso, para documentar dependencias entre nucleótidos colindantes (figura 26).

Figura 26. Sensores de señal de GENEID para *D. melanogaster*.

Aceptadores																												
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	
AA	-0.171	-0.072	-0.131	-0.179	-0.221	-0.245	-0.501	-0.304	-0.345	-0.508	-0.459	-0.514	-0.600	-0.903	-0.668	-0.624	-0.694	-0.892	-0.903	-0.477	0.604	-1.719	0.000	—	—	-0.253	0.039	
AC	0.524	0.397	0.585	0.503	0.508	0.560	0.442	0.526	0.558	0.641	0.511	0.403	0.614	0.572	0.631	0.643	0.597	0.824	0.750	0.694	0.482	1.062	—	—	—	0.104	0.143	
AG	-0.702	-0.769	-0.853	-0.690	-0.803	-0.709	-0.731	-0.810	-1.265	-1.490	-1.409	-1.810	-2.661	-2.291	-2.565	-2.869	-3.333	-3.444	-3.710	-3.599	-1.489	-4.427	—	0.000	—	-0.370	-0.279	
AT	0.300	0.345	0.322	0.326	0.391	0.357	0.609	0.445	0.627	0.673	0.697	0.944	0.855	0.936	0.872	0.869	0.915	0.866	0.855	0.762	-0.305	0.313	—	—	—	0.551	0.187	
CA	-0.185	-0.298	-0.334	-0.443	-0.416	-0.475	-0.577	-0.684	-0.819	-0.826	-1.139	-1.067	-1.313	-1.236	-1.140	-1.103	-1.027	-1.215	-1.367	-1.159	0.061	-1.935	0.000	—	—	-0.424	-0.221	
CC	0.057	0.078	0.093	0.089	0.047	0.114	0.160	0.084	0.257	0.252	0.291	0.135	0.154	0.129	0.178	0.284	0.382	0.417	0.541	0.484	0.101	0.431	—	—	—	-0.607	-0.238	
CG	-0.450	-0.347	-0.405	-0.501	-0.458	-0.345	-0.479	-0.365	-0.676	-0.691	-0.629	-0.528	-0.666	-1.046	-0.606	-0.919	-0.450	-1.277	-1.429	-1.280	0.322	-3.169	—	—	—	0.048	0.192	
CT	0.191	0.218	0.224	0.329	0.316	0.280	0.330	0.378	0.354	0.361	0.415	0.457	0.533	0.541	0.455	0.444	0.303	0.341	0.282	0.209	-0.307	0.503	—	—	—	0.574	0.287	
GA	0.028	-0.064	0.087	0.080	-0.058	-0.323	-0.291	-0.325	-0.494	-0.980	-1.079	-1.084	-1.093	-1.223	-1.147	-1.454	-1.337	-1.509	-1.405	-0.874	-0.184	-2.665	0.000	—	-0.177	-0.117	-0.152	
GC	0.147	0.194	0.113	0.226	0.250	0.276	0.260	0.260	0.292	0.282	0.243	0.209	0.338	0.136	0.390	0.356	0.573	0.630	0.216	0.228	0.252	1.081	—	—	—	-0.415	-0.115	0.121
GG	-0.457	-0.440	-0.655	-0.639	-0.409	-0.514	-0.613	-0.304	-0.641	-0.561	-0.479	-0.861	-0.888	-0.884	-0.950	-0.976	-0.792	-1.068	-0.913	-0.600	0.111	-4.305	—	—	—	-0.564	-0.085	-0.308
GT	0.307	0.320	0.395	0.299	0.245	0.480	0.537	0.343	0.644	0.776	0.789	0.932	0.872	0.982	0.890	0.955	0.805	0.801	1.035	0.810	-0.295	0.152	—	—	-0.714	0.336	0.375	
TA	-0.064	-0.051	-0.177	-0.069	-0.364	-0.701	-0.544	-0.490	-0.520	-0.744	-1.014	-1.070	-1.112	-1.411	-1.290	-0.995	-1.080	-1.067	-1.411	-1.106	-0.269	-1.494	0.000	—	—	—	-0.402	-0.278
TC	0.1817	0.044	0.169	0.158	0.211	0.219	0.329	0.224	0.253	0.238	0.360	0.333	0.327	0.357	0.546	0.464	0.634	0.569	0.623	0.113	0.210	0.794	—	—	—	-0.274	0.244	
TG	-0.239	-0.236	-0.208	-0.273	-0.113	-0.218	-0.363	-0.357	-0.440	-0.421	-0.453	-0.600	-0.555	-0.718	-0.819	-0.528	-0.567	-0.873	-1.141	-1.281	-0.061	-5.211	—	—	—	0.205	0.056	
TT	0.112	0.199	0.139	0.162	0.108	0.298	0.238	0.296	0.342	0.412	0.368	0.470	0.432	0.539	0.438	0.386	0.224	0.374	0.459	0.707	0.066	0.480	—	—	—	0.153	-0.176	

Donadores									
	1	2	3	4	5	6	7	8	9
A	0.365	0.837	-1.170	—	—	1.064	1.126	-1.202	-0.401
C	0.449	-0.612	-1.092	—	—	-2.047	-1.127	-1.701	-0.289
G	-0.287	-0.681	1.122	0.000	—	0.228	-0.620	1.228	-0.252
T	-0.917	-0.515	-1.604	—	0.000	-2.297	-1.329	-1.689	0.506

Codones de inicio														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	-0.520	0.029	-0.423	-0.186	-0.211	0.817	-0.173	-0.639	0.000	—	—	0.052	-0.023	-0.884
C	0.736	0.447	0.059	0.442	0.906	-1.434	0.864	0.720	—	—	—	-0.193	0.278	0.115
G	-0.237	-0.441	0.577	-0.092	-0.457	0.131	-0.577	-0.140	—	—	0.000	0.551	-0.170	0.577
T	-0.377	-0.227	-0.594	-0.296	-1.306	-2.711	-0.825	-1.322	—	0.000	—	-1.121	-0.107	-0.303



El cálculo de la razón de similitud entre los valores obtenidos para cada modelo (señales verdaderas y señales falsas) nos permite construir una distribución de valores que podemos utilizar para identificar señales de cada clase en secuencias previamente no anotadas. Dados dos modelos de Markov  $P$  y  $Q$  contruidos para capturar ejemplos positivos y negativos de la misma señal, construiremos la tabla de puntuaciones  $LM$  para evaluar cualquier secuencia de nucleótidos. En definitiva, estudiando la frecuencia observada para un nucleótido cuando aparece justo después de cada palabra de  $k$  pares de bases a lo largo de cada posición  $j$  de la señal en ambos modelos, podremos afirmar a qué modelo se asemeja en mayor medida.

Figura 27. Sensor de señal de GENEID.

$$LM^j(s_{k+1}|s_1 \dots s_k) = \log \frac{P^j(s_{k+1}|s_1 \dots s_k)}{Q^j(s_{k+1}|s_1 \dots s_k)}$$

Como hemos visto anteriormente, para analizar secuencias más largas debemos evaluar sistemáticamente cada grupo de  $k$  nucleótidos. Únicamente aquellos motivos que obtengan una puntuación superior a un cierto valor lindero (en inglés, *threshold* o *cutoff*) serán reportados. Para un sensor genérico  $LM$ , evaluaremos una secuencia  $S$  mediante la razón de verosimilitud calculada sobre cada posición  $i + k$  del posible candidato:

Figura 28. Utilización de un sensor de señal de GENEID.

$$L = \sum_{i=1}^{|S|-k} LM^{i+k}(s_{i+k}|s_1 \dots s_{i+k-1})$$

GENEID utiliza sensores específicos para detectar señales de inicio de traducción ( $L_B$ ), aceptadores ( $L_A$ ) y donadores ( $L_D$ ). Para las señales de terminación de la traducción ( $L_E$ ) todo se reduce a buscar ocurrencias exactas de los tres codones posibles (TAA, TAG y TGA). Durante el entrenamiento de GENEID, sobre cada genoma se establecen los valores óptimos que permiten capturar el máximo número de señales reales en los conjuntos de prueba (ver figura 26 para la mosca de la fruta).

Tabla 5. Parametrización de sensores de señal para GENEID.

Especie	Parámetro	$L_A$	$L_D$	$L_B$	$L_E$
Mosca	longitud	27	9	14	3
	dependencias	1	0	0	0
Humano	longitud	27	9	20	3
	dependencias	1	0	2	0
Arroz	longitud	24	9	18	3

### Lecturas complementarias

E. Blanco; G. Parra; R. Guigó (2003). *Using geneid to identify genes*. *Current Protocols in Bioinformatics*. Nueva York: John Wiley & Sons Inc. ISBN: 0471250937.

G. Parra; E. Blanco; R. Guigó (2000). "GeneID in *Drosophila*". *Genome Research* (núm. 10, págs. 511-515).

Especie	Parámetro	$L_A$	$L_D$	$L_B$	$L_E$
	dependencias	0	0	0	0

Cuando realizamos la predicción sobre una secuencia genómica  $S$  con GENEID, la aplicación de cada sensor sobre esta produce una lista de señales candidatas de cada clase. GENEID genera automáticamente las combinaciones adecuadas para construir los distintos tipos de exones (iniciales, internos, terminales y únicos), evitando incluir codones de parada dentro de la pauta de traducción abierta.

GENEID emplea el sensor de contenido para evaluar el potencial codificante de cada exón. A partir del mismo conjunto de exones reales utilizado para construir los sensores de señal, los desarrolladores también registran la frecuencia de cada codón en las mismas proteínas. Repitiendo el mismo procedimiento sobre la fracción de esas regiones genómicas que no pertenecían a ningún exón conocido (por ejemplo, los intrones), es posible construir un modelo negativo de contraste. La razón de verosimilitud entre ambos modelos (exónico e intrónico) permite identificar aquellas predicciones con un contenido codificante característico de dicha especie.

Está documentado que, dentro de una proteína, ciertos aminoácidos aparecen junto a otros con más frecuencia de la esperable. GENEID introduce cadenas de Markov de orden 5 en estos cálculos para evaluar dependencias entre grupos de seis nucleótidos en cada posible pauta de lectura. Para la inicialización es necesario registrar la frecuencia de cada pentámero en los exones de ambos conjuntos, mientras que para evaluar la dependencia entre dos codones consecutivos junto con el sesgo en la distribución general, es preciso conocer la frecuencia de cada hexámero en las mismas regiones. Únicamente para los exones reales deben tomarse estas medidas en las tres pautas  $j$  de lectura, ya que las regiones intrónicas no poseen marcos de traducción:

Figura 29. Sensor de contenido de GENEID basado en pentámeros y hexámeros.

$$LI^j(s_1 \dots s_5) = \log \frac{I^j(s_1 \dots s_5)}{I_0(s_1 \dots s_5)}$$

$$LT^j(s_1 \dots s_6) = \log \frac{T^j(s_1 \dots s_6)}{T_0(s_1 \dots s_6)}$$

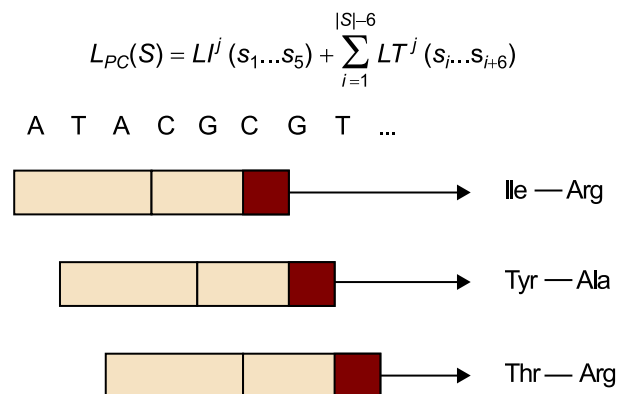
Para evaluar el potencial codificante de un exón  $S$  construido por GENEID, a partir del conjunto de señales detectadas en el paso previo, simplemente debe aplicarse la función  $L_{PC}$  sobre el pentámero inicial y la secuencia de hexámeros en la pauta de lectura que está siendo examinada. De este modo, el programa compara simultáneamente la composición de los codones de dicho exón y las dependencias entre aminoácidos colindantes respecto a aquellos valores

#### Lecturas complementarias

G. Parra; E. Blanco; R. Guigó (2000). "GeneID in *Drosophila*". *Genome Research* (núm. 10, págs. 511-515).  
J. Fickett; C. Tung (1992). "Assessment of protein coding measures". *Nucleic Acids Research* (núm. 20, págs. 6441-6450).

conocidos para un organismo concreto. Este procedimiento debe realizarse en los tres posibles marcos de lectura, siempre que no contengan codones de terminación:

Figura 30. Utilización del sensor de contenido de GENEID en tres pautas de lectura.



A partir de las puntuaciones obtenidas por los distintos sensores, GENEID asignará una puntuación  $L_E$  a cada exón que resulte proporcional a la probabilidad de que dicha secuencia fuera generada por un modelo estadístico derivado de exones conocidos para esa especie. Para cada tipo de exón debemos combinar las señales adecuadas ( $S_A$ ,  $S_D$ ,  $S_B$  y  $S_E$  representan las subsecuencias de las señales que definen el exón):

Figura 31. Puntuación final de un exón.

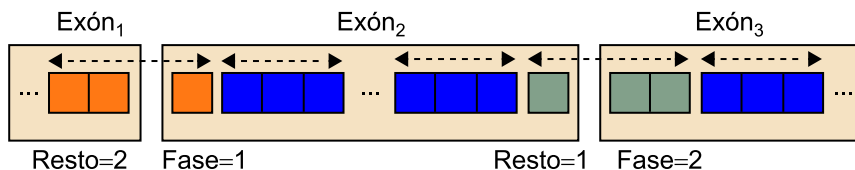
$$\begin{aligned}
 L_E(S) &= L_B(S_B) + L_D(S_D) + L_{PC}(S) \text{ (iniciales)} \\
 L_E(S) &= L_A(S_A) + L_D(S_D) + L_{PC}(S) \text{ (internos)} \\
 L_E(S) &= L_A(S_A) + L_E(S_E) + L_{PC}(S) \text{ (terminales)} \\
 L_E(S) &= L_B(S_B) + L_E(S_E) + L_{PC}(S) \text{ (únicos)}
 \end{aligned}$$

GENEID utiliza la puntuación obtenida por cada exón para filtrar aquellas secuencias que exhiben una baja probabilidad de contener regiones codificantes. Una vez se ha realizado esta criba, es necesario elegir aquellas predicciones que producen un ensamblado óptimo de exones. Durante la síntesis de estos genes, sin embargo, no todos los exones son susceptibles de combinarse para producir un fragmento traducible. Un aspecto clave que debe vigilarse en este proceso es cómo se unen los nucleótidos restantes del anterior exón con los sobrantes del posterior (denominados resto y fase, respectivamente). Internamente, GENEID verifica, por ejemplo, que la unión entre dos exones no produzca un codón de parada precisamente en el marco de lectura utilizado en el transcrito resultante. Debe garantizarse también que no se altera la pauta de traducción en estas uniones, por lo que únicamente se permiten estas uniones cuando la suma de los nucleótidos sobrantes en ambos exones sea múltiplo de tres (ver figura 32).

#### Lectura complementaria

G. Parra; E. Blanco; R. Guigó (2000). "GeneID in *Drosophila*". *Genome Research* (núm. 10, págs. 511-515).

Figura 32. Codones compartidos durante el ensamblado de exones.



El analista bioinformático puede definir las conexiones compatibles entre exones modificando adecuadamente el modelo de gen dentro del fichero de parámetros. Como podemos apreciar en la figura 33, cada regla de este modelo permite especificar las uniones legales entre exones de distintos tipos. En el interior de cada regla observamos tres columnas: clase de exón anterior, clase de exón posterior y distancias mínima/máxima entre ambas ocurrencias. Delimitando las combinaciones permitidas, el usuario indica al programa qué forma deben poseer los genes ensamblados a partir de los exones identificados computacionalmente. Es importante establecer las conexiones hábiles tanto en el interior de los genes como entre estos, añadiendo reglas específicas para cada hebra de la molécula de ADN (ver figura 33).

#### Sin solapamientos

Para producir proteínas válidas los exones ensamblados no deben solaparse.

Figura 33. Modelo de reglas para ensamblar exones.

# conexiones para construir un gen		
First+:Internal+	Internal+:Terminal+	20:25000
Terminal-:Internal-	First-:Internal-	20:25000
# conexiones entre genes		
Terminal+:Single+	Single+:First+	2000:100000
Terminal+:Single+	Single-:Terminal-	2000:100000
First-:Single-	Single+:First+	2000:100000
First-:Single-	Single-:Terminal-	2000:100000

La puntuación de un gen  $g$  se calcula sumando las puntuaciones de todos los exones que lo constituyen. Respetando el modelo de gen configurado anteriormente, el mejor ensamblado de exones es aquel que maximiza precisamente la puntuación final de éste. En condiciones normales, es posible identificar varios genes codificados sucesivamente en el interior de una región genómica, sumándose simplemente sus puntuaciones:

Figura 34. Puntuación final de un gen.

$$L_G(g) = L_E(e_1) + \dots + L_E(e_n) = \sum_{i=1}^n L_E(e_i)$$

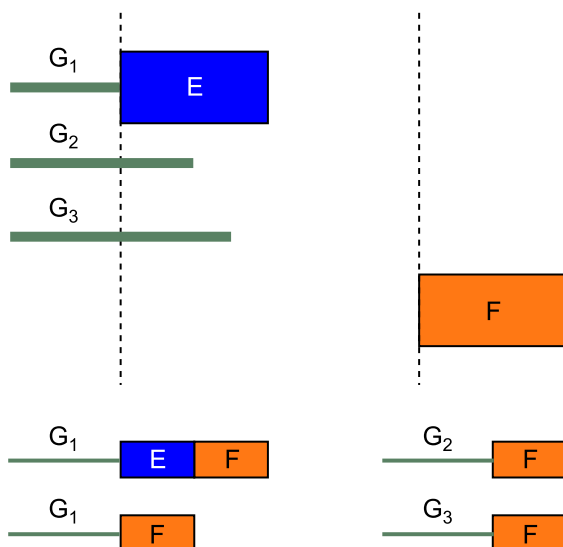
$$L_G(g_1, \dots, g_m) = \sum_{j=1}^m L_G(g_j)$$

La búsqueda del ensamblado óptimo optando por una aproximación basada en la fuerza bruta puede resultar impracticable. Para evitar un cálculo tan costoso, GENEID implementa internamente una versión del algoritmo GenAmic basado en el uso de programación dinámica. Esta estrategia no explora todas las posibles combinaciones de exones, limitando la búsqueda únicamente a regiones locales de la secuencia mediante hábiles ordenaciones del conjunto de exones de trabajo. GenAmic, en función del modelo de gen suministrado, trabaja con los exones agrupados en clases de equivalencia. Para el correcto funcionamiento del algoritmo de programación dinámica, el programa mantiene una lista organizada de los exones según el orden de aparición y de finalización (posiciones izquierda y derecha). Para cada exón de la entrada en el recorrido natural, GenAmic calcula el mejor ensamblado que finaliza precisamente en dicho exón. Para ello, si nos fijamos en dos exones consecutivos  $E$  y  $F$  cuya conexión está permitida en el modelo de gen predefinido, el algoritmo resuelve que el ensamblado óptimo que finaliza precisamente en  $F$  debe enriquecerse mediante la aplicación de una de estas tres reglas (ver figura 35):

- El mejor gen acabado en  $E$ , ensamblando  $F$  a continuación (si no hay solapamiento).
- El mejor gen acabado en  $E$  (excepto  $E$ ), ensamblando  $F$  a continuación.
- El mejor gen a elegir entre todos aquellos que terminan precisamente entre el inicio de  $E$  y el inicio de  $F$ , ensamblando  $F$  a continuación.

La aplicación de esta recurrencia de programación dinámica garantiza, cuando han sido realizadas todas las ordenaciones según las diferentes clases de equivalencias entre exones, que se logrará construir el ensamblado óptimo con un coste asintótico lineal respecto al número total de exones.

Figura 35. Funcionamiento del algoritmo GenAmic.



### Lectura complementaria

R. Guigó (1998). "Assembling genes from predicted exons in linear time with dynamic programming". *Journal of Computational Biology* (núm. 5, págs. 681-702).

La aplicación de todos los sensores y el posterior ensamblado de los exones predichos por el programa GENEID sobre la secuencia del gen *HBB*, produce el siguiente resultado (mostrado gráficamente en la figura 24) en unos pocos segundos:

Figura 36. Salida original del programa GENEID.

```
## date Fri Feb 25 11:44:44 2011
## source-version: geneid v 1.2 - - geneid@imim.es
# Sequence hg18_dna - Length = 2001 bps
# Optimal Gene Structure. 1 genes. Score = 3.08
# Gene 1 (Forward). 3 exons. 148 aa. Score = 3.08
First      186  277 -1.88 - 0 2 1.43 1.30  4.96 0.00 AA   1: 31 hg18_dna_1
Internal   408  630  4.76 - 1 0 3.52 3.60 14.97 0.00 AA  31:105 hg18_dna_1
Terminal 1481 1609  0.20 - 0 0 5.54 0.00  5.93 0.00 AA 106:148 hg18_dna_1
>hg18_dna_1|geneid_v1.2_predicted_protein_1|148_AA
MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAMVGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFTLSELHCDKHLHVDPENFRLLGNVLVCVLAHHFG
KEFTTPPVQAAAYQKVVAGVANALAHKYH*
```

1 Tipo de exón	6 Fase	11 Puntuación de homología
2 Posición inicial	7 Resto	12 Aminoácidos del exón
3 Posición final	8 Puntuación de señal inicial	13 Identificador del exón
4 Puntuación del exón	9 Puntuación de señal final	
5 Hebra (+/-)	10 Potencial codificante	

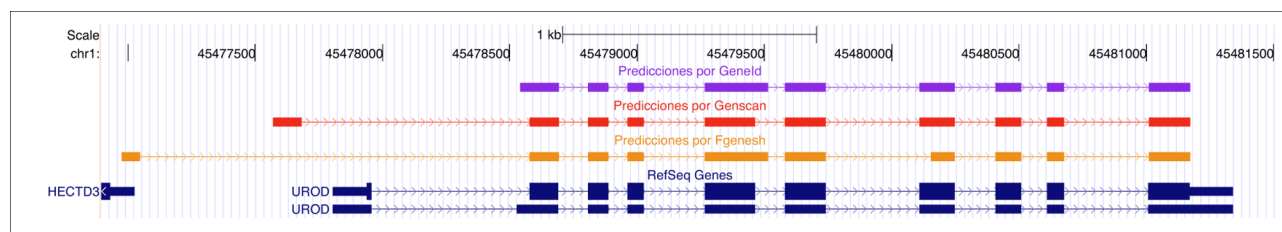
Cada programa implementa su propio motor de predicción, entrenándolo con un conjunto de secuencias diferente. En consecuencia, aunque por regla general las predicciones no han de diferir excesivamente, es posible que exista alguna discrepancia en la anotación computacional de un gen cuando empleamos varios sistemas. Para estudiar con detalle cómo atacar este problema, vamos a centrarnos en la identificación del gen humano *UROD*. Este gen posee 10 exones según el consorcio RefSeq (ver figura 37). Utilizaremos los programas GENEID, GENSCAN y FGENESH para evaluar la calidad de este tipo de aplicaciones precisamente sobre esta secuencia de aproximadamente 5.000 nucleótidos. Como se puede apreciar en la figura 37, detectamos cierta disparidad entre las predicciones, pese a que los tres programas parecen coincidir en la parte central del gen:

#### Lecturas complementarias

C. Burge; S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA". *Journal of Molecular Biology* (núm. 268, págs. 78-94).

A. Salamas; V. Solovyev (2000). "Ab initio Gene Finding in *Drosophila* Genomic DNA". *Genome Research* (núm. 10, págs. 516-522).

Figura 37. Predicciones existentes para el gen humano *UROD*.



Para analizar cuidadosamente estas predicciones vamos a intentar clasificar aquellos exones reportados por cada sistema, identificando los rasgos comunes de los resultados. Podemos tabular esta información para asociar los exones de cada programa que claramente presentan regiones en común:

## Predictores génicos

Para el genoma humano, cada navegador genómico ofrece varias pistas con las predicciones precalculadas de cada programa. Sólo recomendamos ejecutar el servidor web de los predictores génicos para estudios más sofisticados sobre el ajuste de una estructura génica concreta.

Tabla 6. Predicciones de GENEID, GENSCAN y FGENESH.

Exón	Tipo	GENEID	GENSCAN	FGENESH
1	Inicial			144-214
2	Interno		739 851	
3	Interno	1710-1860	1748-1860	1748-1860
4	Interno	1976-2055	1976-2055	1976-2055
5	Interno	2132-2194	2132-2194	2132-2194
6	Interno	2434-2682	2434-2631	2434-2682
7	Interno	2749-2910	2749-2910	2749-2910
8	Interno	3279-3416	3279-3416	3324-3416
9	Interno	3576-3676	3576-3676	3576-3676
10	Interno	3780-3846	3780-3846	3780-3846
11	Terminal	4179-4340	4179-4340	4179-4340

La estrategia más apropiada en este caso es comenzar a dibujar el gen a partir de aquellos exones que, como mínimo, aparecen en dos de los tres sistemas de predicción. El alineamiento global de las tres predicciones a nivel de proteína resulta suficientemente informativo (figura 38): hasta nueve de esos exones han sido simultáneamente detectados por dos o más programas, formando el núcleo inicial de nuestra anotación. Existen más reservas sobre los exones putativos detectados al inicio de la predicción por un único programa. En la próxima sección veremos cómo enriquecer esta primera anotación empleando información obtenida mediante búsquedas por homología en bancos de proteínas.

### Lectura complementaria

Existen también aplicaciones que realizan automáticamente este análisis global de las predicciones para ensamblar el conjunto de exones más razonable:

**K. Howe; T. Chothia; R. Durbin** (2002). "GAZE: a generic framework for the integration of gene-prediction data by dynamic programming". *Genome Research* (núm. 12, págs. 1418-1427).

Figura 38. Comparación de las predicciones a nivel de proteína.

GENEID	-----HTDTYPHPH--LIAR	PQGFPELKNDTFLRAAWGEETD	35
FGENESH	-----MSQLARPRTELPTTFFPAFG--QPLP	PQGFPELKNDTFLRAAWGEETD	46
GENSCAN	VQAIVVWTLDKTVGIIVGTCAKLRIPRLSDENKFLMSPPQGFPELKNDTFLRAAWGEETD		60
*****			
GENEID	YTPVWCMRQAGRYLPEFRETRAADFFSTCRSPEACCELTQLPLRRFLDAAIIFSDILV		95
FGENESH	YTPVWCMRQAGRYLPEFRETRAADFFSTCRSPEACCELTQLPLRRFLDAAIIFSDILV		106
GENSCAN	YTPVWCMRQAGRYLPEFRETRAADFFSTCRSPEACCELTQLPLRRFLDAAIIFSDILV		120
*****			
GENEID	VPQALGMEVTMVPKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQLAGRV		155
FGENESH	VPQALGMEVTMVPKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQLAGRV		166
GENSCAN	VPQALGMEVTMVPKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQLAGRV		180
*****			
GENEID	PLIGFAGAFVWMDRAGTRGAGRSLWK	WTLMTYMVEGGGSSTMAQAKRWLYQRPQASHQLL	215
FGENESH	PLIGFAGAFVWMDRAGTRGAGRSLWK	WTLMTYMVEGGGSSTMAQAKRWLYQRPQASHQLL	226
GENSCAN	PLIGFAGAF-----WTLMTYMVEGGGSSTMAQAKRWLYQRPQASHQLL		223
*****			
GENEID	RILTDALVPYLVGQVVAGAACALQLFESHAGHLGPQLFNKFALPYIRDVAKQVKARLREAG		275
FGENESH	RILTDALVPYLVGQVVAGAAC-----LFNKFALPYIRDVAKQVKARLREAG		271
GENSCAN	RILTDALVPYLVGQVVAGAACALQLFESHAGHLGPQLFNKFALPYIRDVAKQVKARLREAG		283
*****			
GENEID	LAPVPMIIFAKDGHFALEELAQAQYEVVGLDWTVPKKKARECVGKTVTLQGNLDPCALYA		335
FGENESH	LAPVPMIIFAKDGHFALEELAQAQYEVVGLDWTVPKKKARECVGKTVTLQGNLDPCALYA		331
GENSCAN	LAPVPMIIFAKDGHFALEELAQAQYEVVGLDWTVPKKKARECVGKTVTLQGNLDPCALYA		343
*****			
GENEID	SEEEIGQLVKQMLDDFGPHRYIANLGHGLYPDMDPEHVGA	FVDAVHKHSRLLRQN	390
FGENESH	SEEEIGQLVKQMLDDFGPHRYIANLGHGLYPDMDPEHVGA	FVDAVHKHSRLLRQN	386
GENSCAN	SEEEIGQLVKQMLDDFGPHRYIANLGHGLYPDMDPEHVGA	FVDAVHKHSRLLRQN	398
*****			

Leyenda figura 38

Gran parte de las predicciones son comunes. Enmarcamos las tres regiones conflictivas donde se producen discrepancias en algún sistema.



## 5. Predicción de genes por homología

La **predicción por homología de genes** (denominada también *extrínseca*) es una técnica computacional basada en la comparación de secuencias para reconstruir la estructura exónica de un gen a partir de la identificación de proteínas homólogas conservadas en otras especies, empleando información externa derivada de búsquedas en bases de datos o mediante procedimientos de genómica comparativa.

La anotación del catálogo de genes de un genoma es el resultado de integrar conocimiento derivado de fuentes diversas. Para guiar la predicción *ab initio* a gran escala, generalmente se identifica primero el conjunto de proteínas conocidas para otras especies que resultan estar conservadas en dicho genoma. Para ello, es posible ubicar información de genes conocidos directamente sobre nuestra secuencia de trabajo utilizando distintas estrategias de alineamiento. A continuación, los programas de predicción *ab initio* se pueden usar para caracterizar regiones donde previamente no había sido reconocida ninguna anotación por homología. Cuando no disponemos de la anotación completa de un genoma, podemos emplear la información de homología para reforzar aquellas predicciones obtenidas inicialmente con estrategias *de novo* (ver figura 2).

Existen tres posibles fuentes de información por homología:

- Bases de datos de proteínas conocidas.
- Bases de datos de expresión de transcritos.
- Comparación de la misma región en otros genomas.

La búsqueda masiva de información útil en bases de datos externas puede arrojar nuevos datos sobre nuestra secuencia de trabajo. Generalmente podemos localizar fragmentos de transcritos o proteínas de otras especies que encajan significativamente en alguna región de esta secuencia. Para llevar a cabo este tipo de comparaciones más sofisticadas, existe una familia de programas de alineamiento capaces de superponer una proteína conocida sobre una secuencia genómica, evaluando la bondad de los posibles sitios de ajuste. Asimismo, los programas de predicción *ab initio* también permiten la introducción directa de información obtenida por homología entre secuencias para refinar las predicciones iniciales (ver figura 25). Por ejemplo, si utilizamos la secuencia de aminoácidos de cada exón predicho por GENEID para realizar búsquedas en bases de datos de proteínas con el programa BLAST, recuperaremos fragmentos de numerosas formas homólogas de la proteína UROD (figura 39). Dado que

### Lectura complementaria

E. Birney; R. Durbin (2000). "Using GeneWise in the *Drosophila* annotation experiment". *Genome Research* (núm. 10, págs. 547-548).

todos los resultados apuntan en la misma dirección, podríamos concluir que cada exón predicho por GENEID está soportado por esta clase de evidencias (tabla 7).

Figura 39. Validación por homología de una predicción.

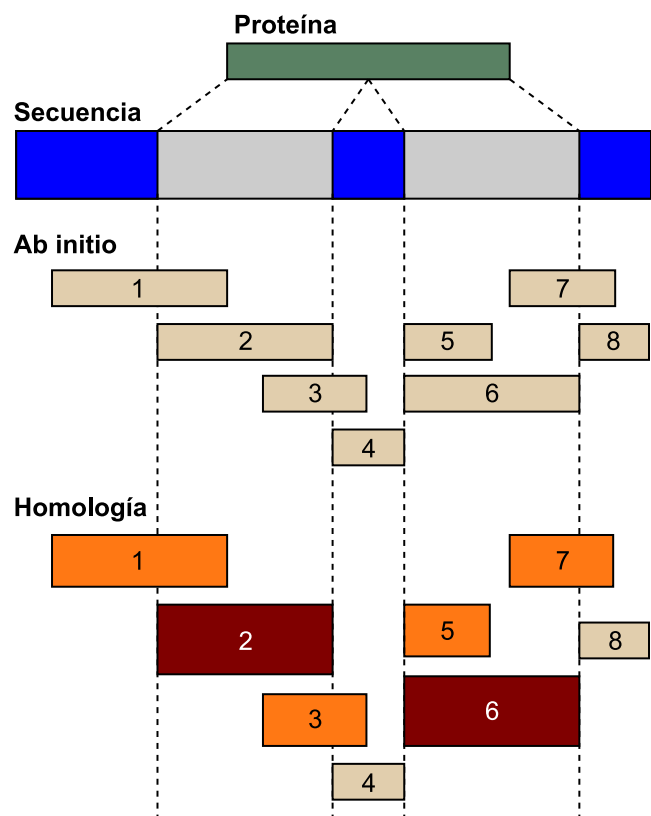
```
GENEID
Terminal 4179 4340 4.55 + 0 0 2.65 0.00 19.89 0.00 54
EEIGQLVKQMLDDFGPHRYIANLGHGLYPMDPEHVGAFVDAVHKHSRLLRQN

BLAST
>ref|XP_001153755.1|PREDICTED: similar to Human Urod [Pan troglodytes]
Score = 119 bits (298), Expect = 1e-25, Method: Composition-based stats.
Identities = 53/53 (100%), Positives = 53/53 (100%), Gaps = 0/53 (0%)

Query 1 EEIGQLVKQMLDDFGPHRYIANLGHGLYPMDPEHVGAFVDAVHKHSRLLRQN 53
      EEIGQLVKQMLDDFGPHRYIANLGHGLYPMDPEHVGAFVDAVHKHSRLLRQN
Sbjct 277 EEIGQLVKQMLDDFGPHRYIANLGHGLYPMDPEHVGAFVDAVHKHSRLLRQN 329
```

Tabla 7. Predicciones de GENEID soportadas por homología con la proteína UROD.

Exón	GENEID	Proteína
3	1710-1860	NP_001012341, UROD [ <i>Ovis aries</i> ]
4	1976-2055	XP_002919490, UROD-like [ <i>Ailuropoda melanoleuca</i> ]
5	2132-2194	AAC50482 UROD [ <i>Homo sapiens</i> ]
6	2434-2682	BAF98769 [ <i>Homo sapiens</i> ]
7	2749-2910	BAG57136 [ <i>Homo sapiens</i> ]
8	3279-3416	XP_001154586 UROD isoform 5 [ <i>Pan troglodytes</i> ]
9	3576-3676	XP_002750795 UROD [ <i>Callithrix jacchus</i> ]
10	3780-3846	XP_002808261 UROD-like [ <i>Macaca mulatta</i> ]
11	4179-4340	XP_001153755 Human UROD [ <i>Pan troglodytes</i> ]

Figura 40. Mejora por homología de las predicciones *ab initio*.**Leyenda figura 40**

El grosor de los exones es proporcional a su puntuación. Las predicciones ubicadas en las regiones de la secuencia donde encaja la proteína verán aumentada su valoración.

**Bancos de datos**

Para proteínas: *non-redundant protein sequences (nr)*, *swiss-prot*, *RefSeq*, *Protein Data Bank*.  
Para genómico: *nucleotide collection (nr/nt)*, *RefSeq mRNA*, *ESTs*, *ALU repeats*.

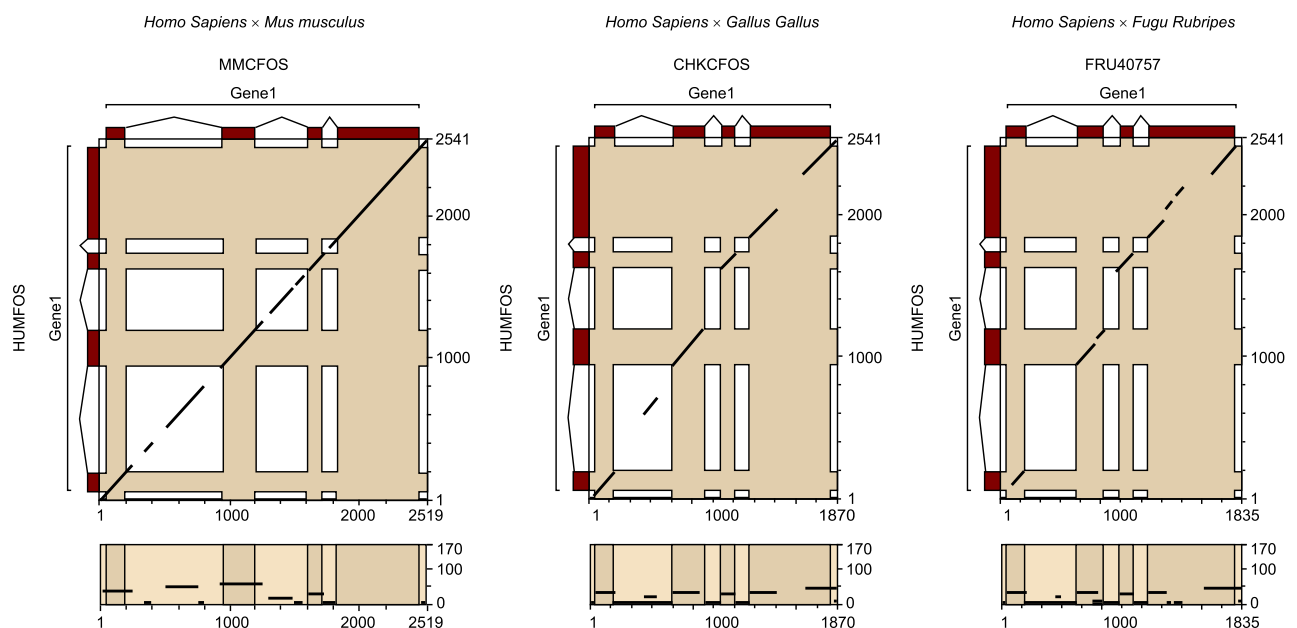
Para mejorar la anotación inicial, podemos informar al programa de predicción *ab initio* de genes sobre la existencia de homología con proteínas conocidas. Cuando estamos anotando secuencias de gran tamaño, la introducción a gran escala de estos datos permite integrar automáticamente la predicción *de novo* con este tipo de conocimiento. Con GENEID, aumentando la puntuación de aquellos exones de la anotación preliminar que presentan coincidencia con fragmentos de proteína de forma proporcional al grado de solapamiento, esta predicción puede beneficiarse de una mejora sustancial. En el caso genérico mostrado en la figura 40, logramos distinguir con más precisión los verdaderos exones introduciendo información sobre homología con una determinada proteína. Para refinar las predicciones es posible también buscar homologías con fragmentos de transcritos obtenidos experimentalmente (EST o *expressed sequence tag*, etiqueta de secuencia expresada) o con secuencias completas de ARN mensajeros conocidos (ADNc o ADN complementario). En otro orden de cosas, antes de llevar a cabo la anotación génica en una secuencia es bastante común también buscar por homología regiones de baja complejidad para enmascarar su contenido. Aunque no codifican proteínas en su interior, estas zonas pueden confundir a los programas de predicción génica.

A nivel evolutivo, debido a su relevancia, la secuencia de las regiones funcionales suele presentar un menor grado de cambios respecto a otras zonas del genoma. Dado que este hecho es observable para todas las especies, la comparación de secuencias homólogas entre dos organismos ubicados en distintos lugares del árbol filogenético puede proporcionarnos importantes pistas sobre la localización de elementos funcionales en ambos genomas. La elección de las especies comparadas resulta, no obstante, fundamental para obtener resultados positivos. Lógicamente, a mayor distancia evolutiva, identificaremos una cantidad superior de cambios entre regiones sinténicas de los cromosomas. Como podemos apreciar en la figura 41, la secuencia completa del gen humano *FOS* parece claramente conservada en el genoma de ratón. Sin embargo, a medida que introducimos especies más lejanas, la comparación de las secuencias comienza a resaltar únicamente los exones. En definitiva, la genómica comparativa puede resultar beneficiosa para anotar una región poco caracterizada de un genoma en función de las anotaciones conocidas para otro genoma (denominado informante) del que poseemos un mayor grado de conocimiento.

### Lectura complementaria

M. R. Brent (2008). "Steady progress and recent breakthroughs in the accuracy of automated genome annotation". *Nature Reviews Genetics* (núm. 9, págs. 62-73).

Figura 41. Comparación de la estructura exónica del gen humano *FOS* en otras especies.



Estas imágenes son cortesía del Dr. Josep F. Abril (Universidad de Barcelona).

Es posible explotar la conservación de secuencia entre dos especies para depurar los catálogos de genes previamente anotados en ambos genomas. En un primer paso es necesario llevar a cabo una comparación exhaustiva a nivel de secuencia entre todos los cromosomas de los dos organismos, prestando especial atención a aquellas regiones susceptibles de codificar proteínas. Aunque resulte extremadamente costoso en tiempo de computación, la herramienta habitualmente empleada para realizar esta tarea con garantías es el alineamiento a nivel de proteína de las seis pautas posibles de traducción para

cada fragmento comparado de los dos genomas. Posteriormente, aquellas regiones claramente enriquecidas con más fragmentos conservados son tratadas para resumir toda esta información.

De un modo similar a cuando trabajamos con datos obtenidos a partir de búsquedas en colecciones de proteínas y transcritos (ver figura 40), la puntuación de aquellos exones identificados *de novo* puede resultar potenciada si existe solapamiento con aquellas regiones conservadas en ambos genomas. Esta estrategia ha sido generalizada también para análisis simultáneo de múltiples genomas aumentando notablemente la calidad de las anotaciones finales. El usuario de estos métodos debe ser consciente de que puede existir una minoría de genes presente específicamente en algunas especies que no sería detectable en otros genomas mediante estas técnicas.

En cualquier caso, estas aproximaciones redundan en la mejora de las anotaciones producidas informáticamente, disminuyendo la necesidad de intervención humana para refinar estos resultados. De hecho, la arquitectura de la mayoría de programas de predicción *ab initio* contiene determinados módulos para incorporar automáticamente toda esta información. Podemos apreciar en la figura 42 el poder de resolución de estas iniciativas comparando la región genómica que alberga el gen humano *UROD* con la zona homóloga en el ratón. En la imagen hemos resaltado el grado de conservación precisamente en el último exón del gen. De hecho, cualquier usuario puede emplear este procedimiento relativamente simple para reafirmar el conjunto de predicciones iniciales en una secuencia de menor longitud.

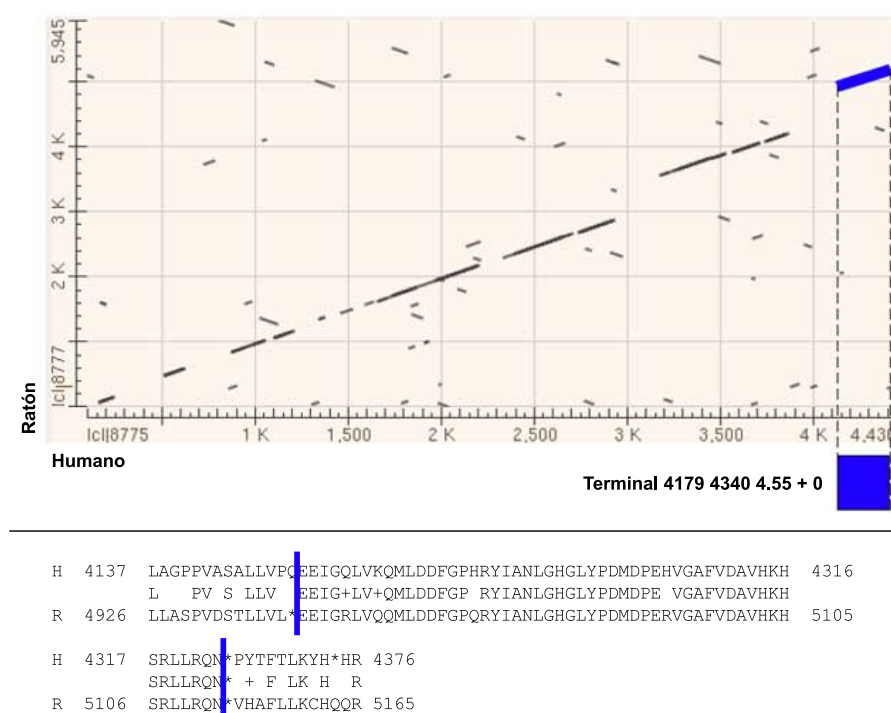
### Lecturas complementarias

G. Parra; P. Agarwal; J. Abril; T. Wiehe; J. Fickett; R. Guigó (2003). "Comparative gene prediction in human and mouse". *Genome Research* (núm. 13, págs. 108-117).

I. Korf; P. Flicek; D. Duan; M. Brent (2001). "Integrating genomic homology into gene structure prediction". *Bioinformatics* (núm. 17, págs. S140-S148).

S. Gross; M. Brent (2006). "Using multiple alignments to improve gene prediction". *Journal of Computational Biology* (núm. 13, págs. 379-393).

Figura 42. Comparación mediante TBLASTX del gen humano *UROD* con el del ratón.



## 6. Caracterización de regiones reguladoras

La complejidad de la vasta red regulatoria de interacciones entre diferentes factores de transcripción que gobiernan la expresión del conjunto de genes del genoma no debe ser menospreciada. Si pensamos en una célula humana en términos de una caja negra con aproximadamente 20.000 entradas asociadas a otros tantos genes, el número de posibles combinaciones de genes activados en un determinado momento alcanza la nada desdeñable cifra de  $2^{20000}$  de estados. Pero esta cantidad aumenta dramáticamente si tenemos en cuenta que los genes se expresan con diferentes intensidades, condicionados por las necesidades de la célula en un momento dado. El estudio computacional de las distintas regiones reguladoras de los genes puede arrojar luz sobre determinados componentes de esa arquitectura génica (figura 20). Dada la particular naturaleza del proceso de reconocimiento de un motivo en una secuencia genómica por parte de los factores de transcripción, el análisis de esas regiones en busca de elementos reguladores no resulta trivial.

La caracterización computacional de una región reguladora en términos del inventario de posibles sitios de unión para diversos factores de transcripción permite reconstruir la arquitectura regulatoria que gobierna la expresión del gen. En dicho proceso se utilizan modelos predictivos contruidos a partir de casos verificados experimentalmente.

La identificación de sitios de unión a factores de transcripción es llevada a cabo empleando la misma clase de sensores de reconocimiento de señales utilizados en la predicción génica. La comparación de distintas ocurrencias del motivo regulador en diferentes regiones promotoras es útil para construir un modelo predictivo aprovechable para buscar ese tipo de secuencias en otras regiones genómicas (ver figura 5). Pese a que estas caracterizaciones del elenco de motivos asociados a distintos factores de transcripción son todavía precarias en algunos casos, podemos echar mano de aquellas secuencias identificadas experimentalmente para construir matrices de pesos más específicas. Existen a disposición de la comunidad científica diferentes repositorios de información regulatoria que puede ser aprovechada para construir modelos de predicción sobre aquellos factores de transcripción involucrados en un determinado problema biológico.

Los catálogos de información sobre sitios de regulación validados experimentalmente recopilan información a partir de la bibliografía existente. Sin embargo y a diferencia de lo que ocurre en el campo de la predicción computacional de genes, el volumen de anotaciones conocidas es notablemente escaso. En algunos casos, junto con el inventario de secuencias reconocidas por deter-

### Lectura complementaria

E. Davidson (2006). *The regulatory genome*. San Diego: Elsevier, Academic Press. ISBN: 9780120885633.

### Lectura complementaria

E. Blanco (2011). *Computational characterization of regulatory regions. Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications*. Hoboken, NJ: Wiley-Blackwell (John Wiley & Sons Ltd). ISBN-13: 978-0470505199.

minados factores, estos recursos bioinformáticos proporcionan una colección de matrices de pesos útiles para buscar esa clase de motivo en nuestras propias secuencias reguladoras.

TRANSFAC fue uno de los recursos pioneros en reunir datos sobre factores de transcripción y motivos de reconocimiento en regiones de ADN. Actualmente, aunque es mayormente conocida por su vasta colección de matrices de pesos, TRANSFAC también alberga información sobre distintos componentes de las redes regulatorias más estudiadas. Aparecida más recientemente, JASPAR es otra colección de modelos predictivos derivados de datos experimentales que incorpora nuevas herramientas para su análisis bioinformático. En ambos recursos, cada entrada de la base de datos contiene las referencias bibliográficas a partir de las cuales se han extraído los motivos para derivar la matriz de pesos (ver figura 43). ABS es otro repositorio de información experimental sobre conjuntos de motivos regulatorios conservados filogenéticamente. Para cada gen, esta base de datos aporta información sobre la ubicación exacta de determinados sitios reguladores en el genoma de estas especies (ver figura 44). OREGANNO es una base de datos de anotaciones reunidas por reconocidos expertos que incluye datos sobre la técnica experimental de validación utilizada en cada caso (figura 45).

Está documentada la existencia de un grado de redundancia significativo entre todos estos modelos predictivos, causado fundamentalmente por la versatilidad de las proteínas para reconocer hábilmente sutiles modificaciones del mismo motivo. Este hecho provoca una cantidad excesiva de predicciones, por lo que resulta especialmente problemática su concentración en determinadas posiciones de las secuencias.

### Lecturas complementarias

E. Wingender (2008). "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation". *Briefings in Bioinformatics* (núm. 9, págs. 326-332).

E. Portales-Casamar y otros (2010). "JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles". *Nucleic Acids Research* (núm. 38, págs. D105-D110).

E. Blanco; D. Farre; M. Alba; X. Messeguer; R. Guigó (2006). "ABS: a database of Annotated regulatory Binding Sites from orthologous promoters". *Nucleic Acids Research* (núm. 34, págs. D63-D67).

The Open Regulatory Annotation Consortium (2008). "OREGAnno: an open-access community-driven resource for regulatory annotation". *Nucleic Acids Research* (núm. 36, págs. D107-D113).

Figura 43. Ficha de TRANSFAC y JASPAR para el mismo factor TBP.

AC M00252  
 ID V\$TATA\_01  
 DE cellular and viral TATA box elements  
 BF T00796 TBP; Species: mouse, *Mus musculus*.  
 BF T00794 TBP; Species: human, *Homo sapiens*.  
 BF T00797 TBP; Species: fruit fly  
 PO     A       C       G       T  
 01     61     145     152     31     S  
 02     16     46     18     309     T  
 03     352     0     2     35     A  
 04     3     10     2     374     T  
 05     354     0     5     30     A  
 06     268     0     0     121     A  
 07     360     3     20     6     A  
 08     222     2     44     121     W  
 09     155     44     157     33     R  
 10     56     135     150     48     N  
 11     83     147     128     31     N  
 12     82     127     128     52     N  
 13     82     118     128     61     N  
 14     68     107     139     75     N  
 15     77     101     140     71     N  
 BA 389 TATA box elements  
 RX PUBMED: 2329577.  
 RA Bucher P.  
 RT Weight matrix descriptions of four eukaryotic  
    RNA polymerase II promoter elements derived  
    from 502 unrelated promoter sequences  
 RL J. Mol. Biol. 212:563-578 (1990).

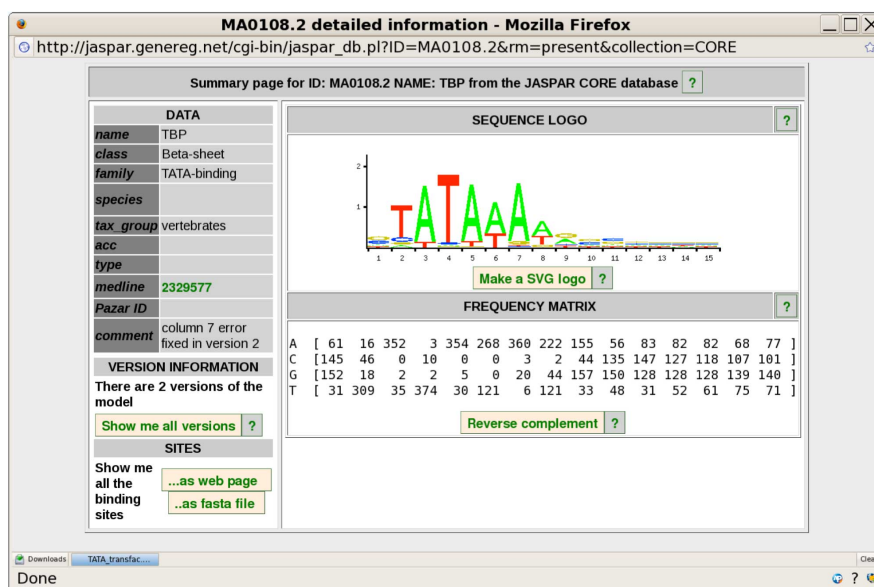




Figura 44. Ficha de ABS para el promotor del gen *LEP*.

**ABS database v 1.0**

ENTRY: A0010  
Leptin

REFSEQ

-NM\_000230 (human)  
-NM\_008493 (mouse)

PUBMED

- B. Lenhard, A. Sandelin, L. Mendoza, P. Engström, N. Jareborg and W.W. Wasserman.  
"Identification of conserved regulatory elements by comparative genome analysis."  
Journal of Biology 2:13 (2003).  
[PMID: 12760745]

Organism	GenBank Accession	Promoter
Homo sapiens	U43589	[500 bps]
<b>Site</b>	<b>Position</b>	<b>Sequence</b>
SP1	400	GGGCGG
CEBP	448	GTTGCGCAAG
TBP	473	TATAAG
Mus musculus	U36238	[500 bps]
<b>Site</b>	<b>Position</b>	<b>Sequence</b>
SP1	402	GGGCGG
CEBP	444	GTTGCGCAAG
TBP	469	TATAAG

Figura 45. Ficha de OREGANNO para un sitio de unión de TBP.

**Record View**

**Target Gene name:** PMF1

**Target Gene version:** homo\_sapiens\_core\_41\_36c

**TF Source:** USER DEFINED


**TF Gene ID:** TBP

**TF Name:** TBP


**Species:** Homo sapiens

**Sequence:** GTTTATA

**Sequence with Flank:** ctgctcaggactcagctggcctgccccgccagcctccagcactGTTTATAccctctgggc  
tgtgccagc



BUILD:  
HG18



BUILD:  
HG19

**Record Details: Record Evidence**

**Evidence class:** Transcription regulator site (OREGEC00001)

**Evidence type:** Mutagenesis (OREGET00009)

**Evidence subtype:** Site-directed (OREGES00054)

**Evidence comment:** Mutating the genomic non-consensus TATA box to the consensus TATA box increases TBP basal activity 1.7-fold.

#### Protocolo para la caracterización de una región reguladora:

- Uso de matrices de pesos almacenadas en catálogos regulatorios.
- Análisis a ciegas de motivos conservados entre secuencias.
- Identificación de módulos regulatorios formados por motivos.

El primer paso para elaborar la descripción regulatoria de una secuencia genómica es la obtención de un listado inicial de sitios candidatos mediante la aplicación del algoritmo de reconocimiento de patrones mostrado en la figura 13. Para esta aproximación es necesario utilizar las matrices de pesos compiladas en las colecciones de TRANSFAC o JASPAR. Afortunadamente, si determinados factores son sospechosos de regular hipotéticamente la expresión de un gen, es posible delimitar un grupo restringido de matrices que nos proporcionarán resultados más específicos. En caso contrario, iniciaremos la búsqueda con el catálogo completo de estos modelos predictivos.

Como podemos apreciar en la figura 46, es imprescindible establecer de antemano un punto de corte para filtrar aquellos motivos más degenerados. De este modo, únicamente informaremos de aquellas ocurrencias con un parecido óptimo (por ejemplo, un 85% o superior). Este valor puede obtenerse empíricamente evaluando el comportamiento de la matriz de pesos sobre un conjunto de sitios reales de la misma familia. El segundo paso de este protocolo genérico permite reafirmar las predicciones iniciales mediante el contraste entre secuencias con un patrón de regulación similar (por ejemplo, secuencias reguladoras de genes ortólogos). Mediante algoritmos de descubrimiento de motivos conservados en regiones genómicas es posible identificar con más solidez los sitios de unión para factores de transcripción que ayuden a explicar precisamente la existencia de estos mecanismos de regulación comunes. Aquellos motivos identificados con este procedimiento pueden ser contrastados con repositorios de matrices conocidas para deducir a qué factor de transcripción corresponden posiblemente (ver figura 47).

El tercer componente de este *pipeline* de anotación regulatoria es el análisis de la distribución espacial de los sitios resultantes a lo largo de las secuencias de estudio. Está documentado que los motivos de determinadas familias tienden a agruparse en el genoma para favorecer el éxito del reclutamiento del factor hacia la zona potencialmente reguladora de los genes. En consecuencia, la identificación de grupos de sitios cercanos con una probabilidad superior a aquella esperable por azar puede ser sumamente indicativa de actividad reguladora. Más intrigante todavía es la formación de ciertos módulos regulatorios (en inglés, *cis-regulatory modules* o CRM), cuya composición invariablemente consta de los mismos motivos en distintos promotores. Estas agrupaciones, en algunos casos, son un reflejo de la interacción necesaria a nivel de proteínas entre los propios factores de transcripción una vez han reconocido positivamente la secuencia promotora. De hecho, la arquitectura de estos módulos

#### Lecturas complementarias

G. D. Stormo (2000). "DNA binding sites: representation and discovery". *Bioinformatics* (núm. 16, págs. 16-23).

J. Turatsinze; M. Thomas-Chollier; M. Defranco; J. van Helden (2008). "Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules". *Nature protocols* (núm. 3, págs. 1578-1588).

T. L. Bailey y otros (2009). "MEME Suite: tools for motif discovery and searching". *Nucleic Acids Research* (núm. 37, págs. W202-W208).

#### Lectura complementaria

T. Werner (2000). "Identification and functional modeling of DNA sequence elements of transcription". *Briefings in bioinformatics* (núm. 1, págs. 372-380).

regulatorios puede explicarse en términos de cooperación entre varios factores para potenciar la transcripción, o de competencia por los mismos sitios de unión para inhibir la acción reguladora de otros factores que desempeñan un rol distinto en el control de la expresión del gen.

Figura 46. Caracterización con matrices de pesos del promotor del gen humano *LEP*.

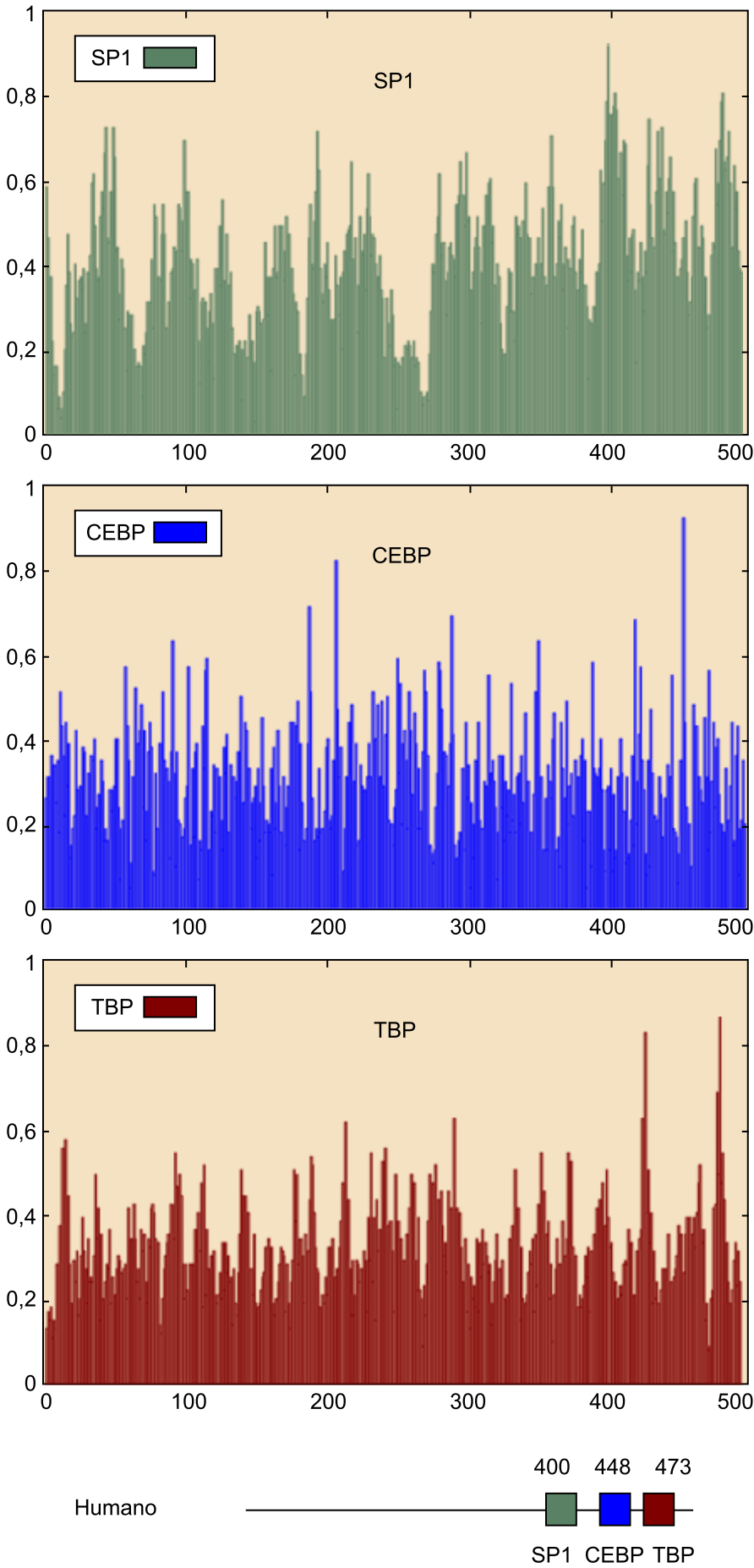
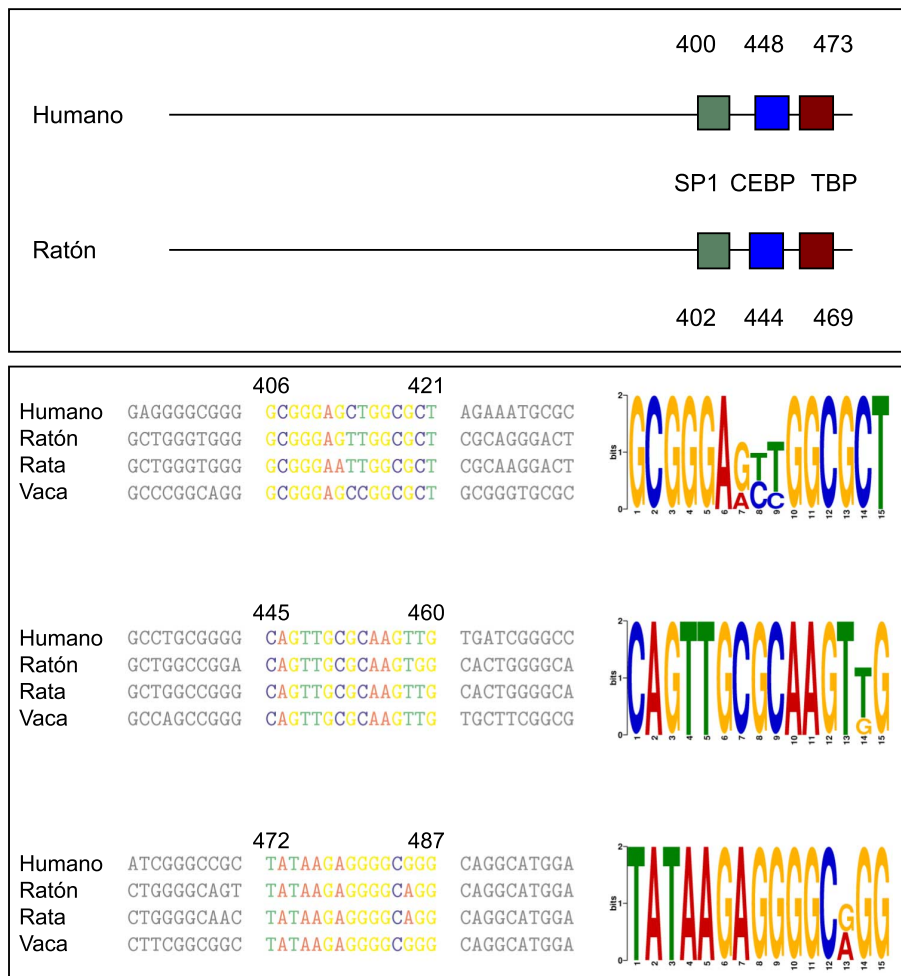


Figura 47. Identificación con el programa MEME de motivos reguladores del gen *LEP*.

Los promotores suelen ser las regiones reguladoras mejor estudiadas para comprender la expresión de un gen. Cuando disponemos de una sólida anotación del catálogo de genes de un genoma podemos deducir fácilmente, a partir de la ubicación del primer exón de estos, cuál es la posible región promotora de la transcripción. Sin embargo, no existe un protocolo bien definido para delimitar con exactitud la longitud de estas secuencias, dado que cada gen posee su propia región reguladora. Para garantizar el éxito en la caracterización de los sitios de unión asociados al control de un determinado gen es imprescindible, no obstante, reconocer correctamente la ubicación de su promotor. Aunque están documentadas algunas características que permiten predecir promotores sobre la base de cierto sesgo de nucleótidos en su composición, es más fiable extraer arbitrariamente suficiente secuencia aguas arriba del inicio de transcripción. En este caso, siempre debemos tener en cuenta la distancia en promedio entre genes en esa especie, evitando solapamientos con la secuencia exónica de los genes vecinos. Para genomas más voluminosos será mayor la secuencia promotora que normalmente deberemos analizar en cada gen. Por ejemplo, para la mosca de la fruta pueden extraerse aproximadamente 1.000 nucleótidos justo antes del inicio del gen, mientras que para el genoma humano es normal trabajar con secuencias de hasta 10.000 pares de bases. Inevitablemente, es preciso disponer de la anotación correcta del inicio de transcripción para nuestros genes. Recientes estudios sobre numerosos catálogos de

### Lecturas complementarias

J. Fickett; A. Hatzigeorgiou (1997). "Eukaryotic promoter recognition". *Genome Research* (núm. 7, págs. 861-878).

Perier, R. C. y otros (2000). "The Eukaryotic Promoter Database (EPD)". *Nucleic Acids Research* (núm. 28, págs. 302-303).

R. Davuluri; I. Grosse; M. Q. Zhang (2001). "Computational identification of promoters and first exons in the human genome". *Nature Genetics* (núm. 29, págs. 412-417).

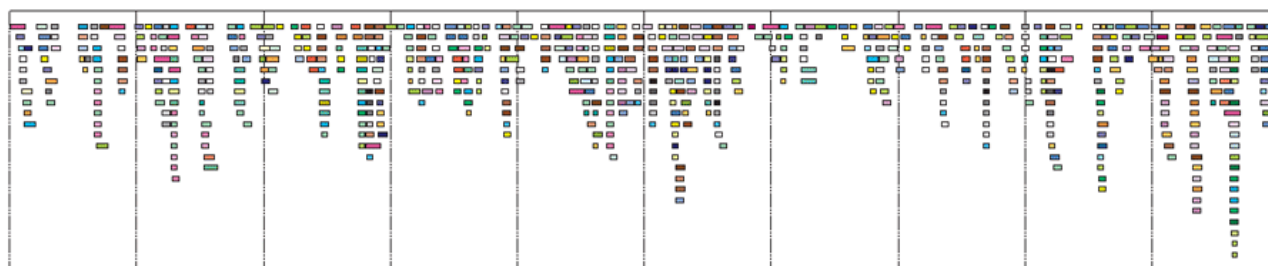
T. Abeel; Y. Van de Peer; Y. Saeys (2009). "Toward a gold standard for promoter prediction evaluation". *Bioinformatics* (págs. i313-i320).

genes han revelado, sin embargo, que la calidad de estas anotaciones es todavía deficiente. Por esta razón, existen varias bases de datos que recopilan información sobre inicios de transcripción génica validados experimentalmente.

## 7. Impronta evolutiva de regiones reguladoras

Dado que un factor de transcripción puede reconocer variaciones de un motivo degenerado y que un motivo concreto puede asociarse a distintos factores, la caracterización inicial de una región promotora generalmente produce una cantidad abundante de predicciones. Como podemos observar en la figura 48, para una secuencia reguladora humana de 2000 bases, una búsqueda que emplee un catálogo amplio de matrices de pesos (indicadas con cajas de distinto color a lo largo de la secuencia, que no se muestra explícitamente aquí) detecta posibles coincidencias en la práctica totalidad de la secuencia. Junto con esta pobre especificidad, apreciamos también que en determinadas posiciones se acumulan más resultados, denotando que ese motivo es reconocible por diferentes factores de transcripción.

Figura 48. Excesivo número de predicciones en una región promotora humana.



Para reducir sustancialmente el ruido en estas representaciones podemos emplear la comparación con otras secuencias relacionadas, aportando nueva información sobre la conservación regulatoria. Existen diversas fuentes de conocimiento que nos permiten identificar regiones de genes que hipotéticamente comparten una colección de sitios de unión:

- Genes que desempeñan funciones similares en el organismo.
- Genes que en experimentos de expresión a gran escala poseen patrones parecidos de activación.
- Genes ortólogos pertenecientes a múltiples especies.

En todos los casos, asumiendo que funciones similares deben ser implementadas mediante combinaciones de motivos comunes, la comparación de secuencias nos ayuda a focalizar nuestro interés únicamente sobre ciertas regiones conservadas para reforzar aquellas predicciones obtenidas con matrices de pesos. De todas estas alternativas, el análisis de regiones reguladoras que incluye información sobre conservación filogenética es el método que arroja resultados más prometedores.

### Lecturas complementarias

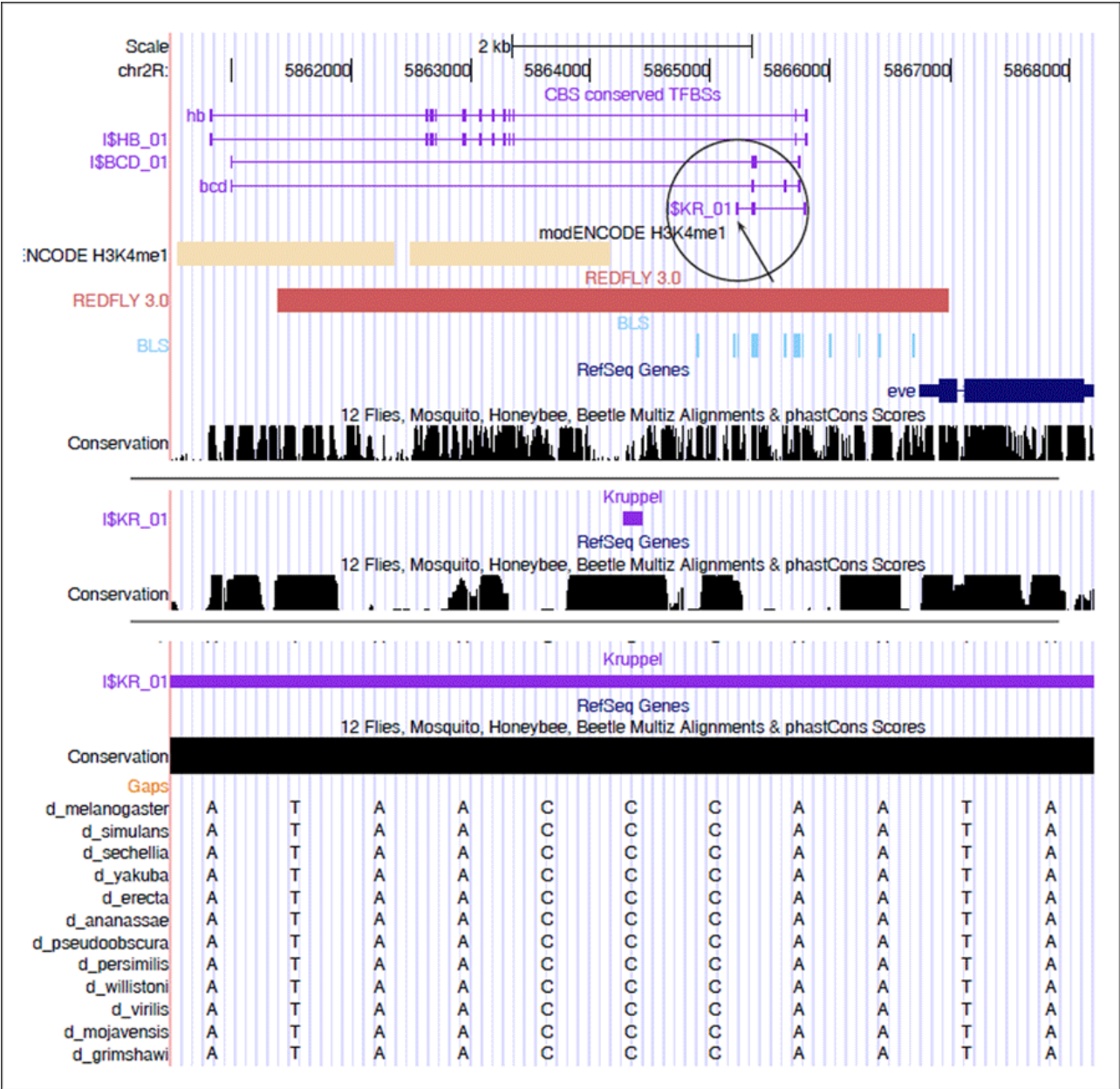
W. W. Wasserman; A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements". *Nature Reviews Genetics* (núm. 5, págs. 276-287).

G. Bejerano y otros (2004). "Ultraconserved elements in the human genome". *Science* (núm. 304, págs. 1321-1325).

Protocolo para la introducción de información sobre conservación evolutiva:

- Elegir adecuadamente las especies para la comparación.
- Identificar las secuencias ortólogas apropiadas.
- Comparar secuencias para extraer bloques de conservación.
- Caracterizar las regiones conservadas con distintas matrices de pesos.
- Clasificar los sitios candidatos en función de su conservación.

Figura 49. Conservación filogenética de sitios de unión a factores de transcripción.



Leyenda figura 49

Mostramos con distintos enfoques el grado de conservación del primer sitio de unión predicho para el factor Krüppel en el promotor del gen *eve* de *D. melanogaster*.



La disponibilidad de la secuencia de múltiples genomas eucariotas representa un avance sustancial hacia la caracterización óptima de las regiones reguladoras de los genes. Existe la creciente convicción entre los miembros de la comunidad científica de que la diversidad animal observada en la naturaleza podría explicarse en términos de distintos grados de sofisticación de la expresión de los genes en cada especie. Las secuencias funcionales tienen tendencia a preservarse mejor que aquellas sin un rol aparente a lo largo de la evolución, acumulando menos mutaciones en su secuencia para evitar daños irreversibles en el organismo. Las comparaciones entre especies pueden resultar enormemente útiles para la identificación de secuencias regulatorias comunes a un grupo de especies.

Tagle y otros acuñaron el término de *impronta filogenética* (en inglés, *phylogenetic footprinting*) para describir esta clase de comparaciones que revelan elementos conservados en regiones genómicas homólogas. La elección del rango de especies apropiadas para el estudio de cada gen en particular resulta crucial dado que cada segmento del genoma evolucionó a una velocidad distinta. No obstante, la existencia de elementos funcionales específicos de cada especie debe ser tomada en cuenta también al realizar estos análisis comparativos. En cualquier caso, esta técnica está ampliamente extendida por su eficacia a la hora de mejorar la anotación de elementos regulatorios. En la figura 49 se puede apreciar el potencial de estas comparaciones; se muestra la caracterización de sitios de unión de distintos factores validados experimentalmente sobre el promotor del gen *eve* de la mosca de la fruta. Puede apreciarse cómo la secuencia genómica del sitio elegido por el factor Krüppel está perfectamente conservada en todas las especies de *Drosophila*.

Figura 50. Reforzamiento de predicciones mediante genómica comparativa.

#### Secuencia



#### Predicciones



#### Conservación entre especies



#### Reforzamiento de predicciones



#### Lectura complementaria

M. Levine; R. Tijan (2003). "Transcriptional regulation and animal diversity". *Nature* (núm. 424, págs. 147-151).

#### Lecturas complementarias

D. Tagle y otros (1988). "Embryonic  $\hat{\imath}$  and  $\hat{g}$  globin genes of a prosimian primate, nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints". *Journal of Molecular Biology* (núm. 203, págs. 439-455).

L. Duret; P. Bucher (1997). "Searching for regulatory elements in human noncoding sequences". *Current Opinion in Structural Biology* (núm. 7, págs. 399-406).

E. T. Dermitzakis; A. G. Clark (2002). "Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover". *Molecular Biology and Evolution* (núm. 7, págs. 1114-1121).

#### Leyenda figura 50

El grosor de las predicciones es proporcional a su puntuación. Las predicciones ubicadas en las regiones de la secuencia conservadas en otras especies verán aumentada su valoración.

No existe una única forma de implementar el análisis comparativo de regiones reguladoras. Es posible introducir la información sobre conservación evolutiva, obtenida a partir de alineamientos entre secuencias genómicas, para reforzar las predicciones iniciales (ver figura 50), o bien llevar a cabo esta predicción computacional sólo sobre esas regiones conservadas. Básicamente, la mayoría de aproximaciones suelen diferir en el método de alineamiento de los genomas involucrados, de modo que es altamente recomendable aprovechar las comparaciones entre especies precalculadas en los navegadores genómicos de propósito general como UCSC (ver las pistas de conservación en la figura 49).

Las moléculas cuya secuencia es similar desempeñan generalmente funciones similares. Sin embargo, a menudo funciones parecidas son llevadas a cabo mediante secuencias distintas que presentan en su interior configuraciones de elementos codificados comunes. Por ejemplo, un mismo factor de transcripción reconoce un amplio abanico de motivos que exhiben cierta variabilidad a nivel de secuencia. Por consiguiente, las regiones promotoras de los genes que presentan patrones de expresión similares podrían ocultar dichos rasgos de conservación a pesar de estar reguladas por parecidas configuraciones de factores de transcripción. La técnica del metaalineamiento transforma las secuencias de promotores en cadenas de símbolos pertenecientes a un nuevo alfabeto en el que cada símbolo representa un factor de transcripción distinto. El alineamiento de estos mapas de factores es capaz de identificar módulos regulatorios comunes que no necesariamente están conservados a nivel de secuencia. En la figura 51 podemos apreciar como el meta-alineamiento del promotor del gen *MMP13* en nueve especies de vertebrados es capaz de rescatar varios sitios de unión conservados en este árbol filogenético que no serían detectables empleando el alineamiento convencional de secuencias.

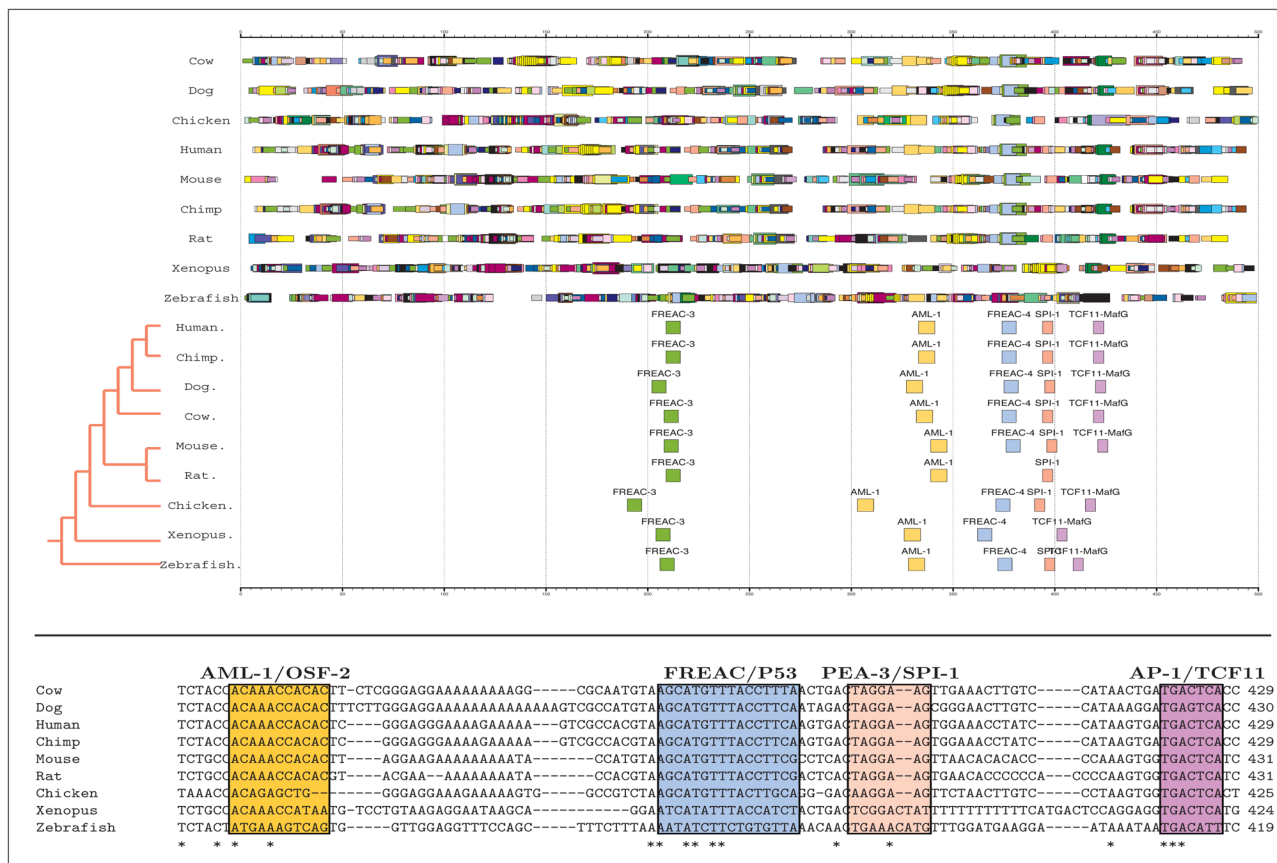
### Lectura complementaria

W. W. Wasserman; A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements". *Nature Reviews Genetics* (núm. 5, págs. 276-287).

### Lecturas complementarias

E. Blanco; X. Messeguer; T. Smith; R. Guigó (2006). "Transcription factor map alignments of promoter regions". *PLoS Computational Biology* (núm. 2, pág. e49).

E. Blanco; R. Guigó; X. Messeguer (2007). "Multiple non-collinear TF-map alignments of promoter regions". *BMC Bioinformatics* (núm. 8, pág. 138).

Figura 51. Metaalineamiento múltiple de la región promotora del gen *MMP13*.

El siguiente código implementa una versión simplificada del metaalineamiento de dos mapas de factores de transcripción. Para cada coincidencia entre los elementos de los mapas, es necesario discernir mediante programación dinámica cuál es el mejor alineamiento calculado previamente que finaliza en una pareja anterior de sitios de unión pertenecientes a ambas secuencias. Para ponderar todos los posibles alineamientos existen determinadas penalizaciones relativas a la conservación de la posición entre dos sitios de unión coincidentes y al número de elementos integrados en el resultado final:

Figura 52. Algoritmo de metaalineamiento de secuencias.

```

PRE ≡ { $S_1, S_2$ : secuencias;  $M$ : catálogo;  $T$ : entero;}
POST ≡ { $A$  es el meta-alineamiento de sitios de union}
(* Construir los mapas de sitios usando las matrices *)
Mapa1 ← reconocimiento_patrones ( $S_1, M, T$ ) ;
Mapa2 ← reconocimiento_patrones ( $S_2, M, T$ ) ;
(* Visitar cada sitio para buscar coincidencias *)
para cada ( $i$  en Mapa1) hacer
    para cada ( $j$  en Mapa2) hacer
        si ( $TF(i) = TF(j)$ ) entonces
            puntos ← 0.0;
            (* Analizar las coincidencias anteriores *)
            para cada ( $(i', j')$  en  $A$ ) hacer
                puntos' ← Alineamiento ( $i, j, i', j'$ );
                si (puntos' > puntos) entonces
                    puntos ← puntos';
            fsi
        fpara
            RegistrarCoincidencia ( $i, j, A, puntos$ );
    fsi
fpara
fpara
ReportarAlineamientoOptimo ( $A$ );
retorna ( $A$ )

```

## 8. Evaluación de las predicciones

Para garantizar la eficacia de cualquier programa de identificación de genes o regiones reguladoras es imprescindible evaluar previamente la calidad de sus predicciones sobre un conjunto estable de referencia. Dado que en muchas ocasiones utilizaremos varias aplicaciones bioinformáticas distintas para caracterizar la misma secuencia, resulta necesario definir un marco teórico con un amplio abanico de parámetros cuantitativos que nos permita contrastar todas las predicciones obtenidas por cada sistema. Existe un área de la bioinformática que define formalmente los parámetros de calidad que deben ser calculados en este procedimiento de evaluación de la precisión de estos sistemas.

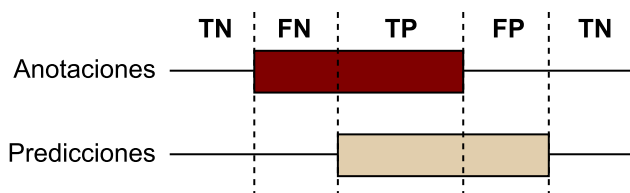
- **Anotaciones:** Conjunto de segmentos genómicos identificados en el genoma por sus coordenadas exactas cuya existencia ha sido validada experimentalmente (por ejemplo, exones, sitios de unión a factores de transcripción o inicios de transcripción).
- **Predicciones:** Conjunto de segmentos genómicos identificados en el genoma por sus coordenadas exactas que han sido suministrados por una aplicación bioinformática después de analizar una secuencia de nucleótidos.

La evaluación consiste básicamente en calcular el grado de solapamiento entre las anotaciones y las predicciones de cada sistema bioinformático sobre las mismas secuencias. Para realizar de modo más efectivo estas comparaciones, el conjunto de anotaciones empleado durante este procedimiento no debe guardar relación con aquellos elementos utilizados para el entrenamiento del programa bioinformático. Dado que en ambos casos podemos definir formalmente el problema como una división en distintas partes de la secuencia de entrada, tanto la predicción de genes como la detección de elementos reguladores comparten la mayoría de las medidas de evaluación genéricas que presentamos continuación.

- **Sensibilidad:** Proporción de los elementos conocidos de las secuencias que resultan efectivamente capturados por las predicciones de la aplicación bioinformática.
- **Especificidad:** Proporción de las predicciones de la aplicación bioinformática que permite la detección correcta de los elementos conocidos en las secuencias.

Los programas de predicción deben lograr un exquisito equilibrio entre sensibilidad y especificidad. Un excesivo número de predicciones puede producir una sensibilidad máxima con una pobre especificidad. Del mismo modo, una pronunciada carencia de predicciones sobre una secuencia posiblemente obtendrá una alta especificidad con una sensibilidad muy escasa. Idealmente, los diseñadores de estas aplicaciones buscan el punto de coincidencia óptimo tal que predicciones y anotaciones encajan perfectamente. Para medir con precisión estos dos parámetros es necesario registrar para cada secuencia los diferentes puntos de intersección entre ambos conjuntos. Si representamos gráficamente anotaciones y predicciones como cajas ubicadas en ciertas coordenadas a lo largo de la misma secuencia, es posible medir distintos parámetros que nos conducirán al cálculo final de la sensibilidad y la especificidad de nuestras predicciones (figura 53):

Figura 53. Parámetros en la evaluación de predicciones bioinformáticas.



#### Lectura complementaria

M. Burset; R. Guigó (1996).  
"Evaluation of gene structure prediction programs".  
*Genomics* (núm. 34, págs. 353-367).

Basándonos en el análisis comparativo entre una región genómica conocida y la predicción de ésta realizada por una aplicación bioinformática, podemos definir las siguientes **métricas de precisión** (incluimos su correspondiente denominación original en inglés):

- **Verdadero positivo** (TP, *true positive*): aquellas posiciones de la región genómica real correctamente incluidas en la predicción.
- **Verdadero negativo** (TN, *true negative*): fragmento de la secuencia que no correspondiendo a ninguna región conocida no resulta incluido en la predicción.
- **Falso positivo** (FP, *false positive*): fragmento de la secuencia que no correspondiendo a ninguna región conocida resulta incluido en la predicción.
- **Falso negativo** (FN, *false negative*): aquellas posiciones de la región genómica real incorrectamente no incluidas en la predicción.

A partir de estos parámetros básicos, ahora podemos definir formalmente la sensibilidad (en inglés, *Sn* o *sensitivity*) y la especificidad (*Sp* o *specificity*) de un conjunto de predicciones respecto a las anotaciones documentadas dentro de una región genómica. La sensibilidad es el cociente entre el número de aciertos respecto a la cantidad total de anotaciones, mientras que la especificidad es el cociente entre el número de aciertos y la cantidad total de predicciones. Numéricamente, por tanto, el rango de valores para las dos medidas de precisión está entre 0 y 1 (considerando 1 el valor máximo):

Figura 54. Sensibilidad y especificidad de las predicciones.

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

$$Sp = \frac{TN}{TN + FP}$$

La evaluación por separado de la sensibilidad y la especificidad únicamente nos proporciona información parcial sobre la calidad de un conjunto de predicciones. Para medir el balance entre ambas medidas, la media aritmética resulta poco precisa. Existe cierto consenso entre los miembros de la comunidad bioinformática para evaluar este equilibrio mediante el coeficiente de correlación (en inglés, *CC* o *correlation coefficient*), que puntúa positivamente las posiciones correctamente identificadas, penalizando negativamente los errores de predicción (rango numérico de -1 hasta +1, correlación de Pearson):

#### Lectura complementaria

M. Burset; R. Guigó (1996). "Evaluation of gene structure prediction programs". *Genomics* (núm. 34, págs. 353-367).

Figura 55. Correlación entre sensibilidad y especificidad de las predicciones.

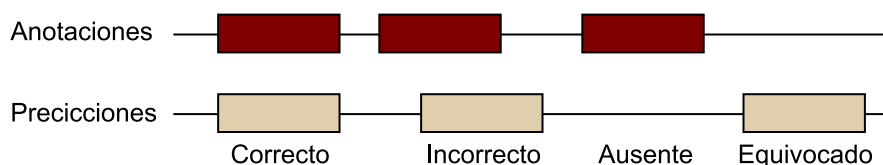
$$CC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

Aunque la terminología básica es la misma para la predicción de genes y la caracterización de regiones reguladoras, cada campo de investigación posee determinadas medidas específicas debido a las características particulares de estos elementos biológicos. Ya que se ha invertido un enorme esfuerzo en conseguir anotar el catálogo de genes del genoma, los procedimientos de medición de la calidad de predicciones génicas han sido desarrollados más ampliamente. A la hora de contrastar el grado de parecido entre los exones conocidos de un gen y las predicciones bioinformáticas, podemos distinguir hasta tres niveles de evaluación: nucleótido, exón y transcrito completo. Mientras el cálculo de la sensibilidad y la especificidad a nivel de nucleótido está basado simplemente en la aplicación estricta de las fórmulas anteriores sobre el conjunto de posiciones de la secuencia de trabajo, a nivel de exón es preciso definir una nomenclatura adicional para abarcar toda la casuística posible.

A partir del grado de solapamiento entre un exón conocido y las predicciones disponibles, podemos catalogar el resultado final del siguiente modo (ver figura 56):

- Exón correcto, cuando las dos señales de ajuste han sido identificadas correctamente y la predicción encaja perfectamente sobre todo el exón conocido.
- Exón parcialmente correcto, cuando efectivamente existe solapamiento entre un exón conocido y una predicción, pero no encajan perfectamente las señales de ajuste.
- Exón equivocado (en inglés, *wrong exon*), cuando no existe nucleótido alguno en el exón predicho que pueda superponerse a ningún exón conocido.
- Exón ausente (en inglés, *missing exon*), cuando para cualquier posición del exón conocido no ha sido identificado ningún exón predicho que presente cierto grado de solapamiento.

Figura 56. Evaluación de exones conocidos.



Aunque es posible calcular la sensibilidad y especificidad a nivel de exón utilizando el número de exones correctamente identificados, es habitual mencionar simplemente el porcentaje de exones acertados. Generalizando estas definiciones para todos los exones de un transcrito conocido, afirmaremos que un gen está correctamente identificado cuando todos sus exones han sido per-



fectamente predichos. A nivel de gen, según el caso, también podemos clasificar las predicciones como ausentes o equivocadas. Dado que en este punto es importante ensamblar correctamente los exones dentro del gen apropiado, también es posible hablar de genes innecesariamente fragmentados (en inglés, *split genes*) o genes incorrectamente fusionados (*joined genes*). Por la propia definición de cada jerarquía (bases, exones y genes), cualquier evaluación entre anotaciones y predicciones proporcionará valores más positivos a nivel de nucleótido, decreciendo a medida que tenemos en cuenta únicamente exones o genes correctamente identificados.

Tomando como ejemplo la predicción obtenida por el programa GENEID sobre la secuencia del gen humano *UROD* (ver tabla 6) y las coordenadas de sus exones conocidos, podemos calcular a nivel de nucleótido los distintos valores de las métricas estudiadas anteriormente en la nueva tabla 8. Tanto la sensibilidad como la especificidad de las predicciones a este nivel son óptimas (95% o más). Además, de los diez exones que conforman este gen, GENEID identifica correctamente siete de ellos, ignorando uno completamente y equivocándose parcialmente en otros dos (70% de exones correctos). En consecuencia, durante un procedimiento sistemático de evaluación de la precisión del programa a nivel de transcrito completo no podría darse por satisfactoria esta predicción. Analizando el alineamiento global entre todas las predicciones obtenidas anteriormente y la proteína conocida (ver figura 57), podemos identificar claramente el rendimiento de cada programa de predicción génica en este caso particular.

Tabla 8. Evaluación de las predicciones de GENEID y el gen *UROD*.

Tipo	GENEID	UROD	nTP	nFP	nFN	Exón
Inicial	-	1107-1126	0	0	20	ausente
Interno	1710-1860	1748-1860	113	39	0	incorrecto
Interno	1976-2055	1976-2055	80	0	0	correcto
Interno	2132-2194	2132-2194	63	0	0	correcto
Interno	2434-2682	2434-2631	198	52	0	incorrecto
Interno	2749-2910	2749-2910	162	0	0	correcto
Interno	3279-3416	3279-3416	138	0	0	correcto
Interno	3576-3676	3576-3676	101	0	0	correcto
Interno	3780-3846	3780-3846	67	0	0	correcto
Terminal	4179-4340	4179-4340	162	0	0	correcto
Totales			1084	91	20	

La evaluación sistemática de la calidad de las predicciones suministradas por distintos programas bioinformáticos debe llevarse a cabo seleccionando cuidadosamente el conjunto de secuencias de prueba. Con la progresiva mejora de las técnicas de secuenciación y el incremento del número de genes anotados, nuevos estudios sobre la precisión de las predicciones han actualizado el estado del arte de la identificación génica. No obstante, aún quedan determinados aspectos que necesitan de más atención en el futuro, como el contraste de las distintas formas alternativas del gen. Actualmente, según los últimos estudios realizados sobre el genoma humano, la predicción *ab initio* garantiza buenas predicciones a nivel de nucleótido (CC alrededor del 80%), aumentando significativamente este valor (cerca del 95%) cuando introducimos información de homología con proteínas conocidas o genómica comparativa. Sin embargo, los resultados a nivel de exones empeoran, logrando acertar entre el 50% y el 80% de los exones cuando empleamos en este último caso información derivada de búsquedas en bases de datos externas. Según el conjunto de referencia, estos mismos estudios afirman que el grado de acierto sobre estructuras génicas completas oscila alrededor del 30% y el 50% de éxito. En el caso de introducir en la evaluación criterios más restrictivos referentes a la identificación de isoformas del mismo gen, estos valores podrían descender sensiblemente.

### Lecturas complementarias

M. Burset; R. Guigó (1996). "Evaluation of gene structure prediction programs". *Genomics* (núm. 34, págs. 353-367).

M. Reese; G. Hartzell; N. Harris; U. Ohler; J. Abril; S. Lewis (2000). "Genome annotation assessment in *Drosophila melanogaster*". *Genome Research* (núm. 10, págs. 483-501).

S. Rogic; A. Mackworth; F. Ouellette (2001). "Evaluation of gene-finding programs on mammalian sequences". *Genome Research* (núm. 11, págs. 817-832).

R. Guigó y otros (2006). "EGASP: the human ENCODE genome annotation assessment project". *Genome Biology* (núm. 7, pág. S2).

Figura 57. Comparación de la proteína UROD y las predicciones bioinformáticas.

GENSCAN	VQAIVWTWLDKTVGIIVGTCALRIIPRLSDENKFLMSHPQGFPELKNDTFLRAAWGEETD	60
UROD	-----MEANGLG-----PQGFPELKNDTFLRAAWGEETD	29
GENEID	-----HTDTYPHPHLIAR-----PQGFPELKNDTFLRAAWGEETD	35
FGENESH	-----MSQLARPR---TELPTTFPAFGQP---LPQGFPELKNDTFLRAAWGEETD	46
	*****	
GENSCAN	YTPVWCMRQAGRYLPEFRETRAADFFSTCRSPEACCELTLOPLRRFLLDAAIIFSDILV	120
UROD	YTPVWCMRQAGRYLPEFRETRAADFFSTCRSPEACCELTLOPLRRFLLDAAIIFSDILV	89
GENEID	YTPVWCMRQAGRYLPEFRETRAADFFSTCRSPEACCELTLOPLRRFLLDAAIIFSDILV	95
FGENESH	YTPVWCMRQAGRYLPEFRETRAADFFSTCRSPEACCELTLOPLRRFLLDAAIIFSDILV	106
	*****	
GENSCAN	VPQALGMEVTMPVGKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQRLAGRV	180
UROD	VPQALGMEVTMPVGKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQRLAGRV	149
GENEID	VPQALGMEVTMPVGKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQRLAGRV	155
FGENESH	VPQALGMEVTMPVGKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQRLAGRV	166
	*****	
GENSCAN	PLIGFAGAP-----WTLMTYMVEGGGSSTMAQAKRWLYQRPQASHQLL	223
UROD	PLIGFAGAP-----WTLMTYMVEGGGSSTMAQAKRWLYQRPQASHQLL	192
GENEID	PLIGFAGAPVMWDRAGTRGAGRSLWKWTLMTYMVEGGGSSTMAQAKRWLYQRPQASHQLL	215
FGENESH	PLIGFAGAPVMWDRAGTRGAGRSLWKWTLMTYMVEGGGSSTMAQAKRWLYQRPQASHQLL	226
	*****	
GENSCAN	RILTDALVPYLVGQVVAGACALQLFESHAGHLGPQLFNKFALPYIRDVAQVVKARLREAG	283
UROD	RILTDALVPYLVGQVVAGACALQLFESHAGHLGPQLFNKFALPYIRDVAQVVKARLREAG	252
GENEID	RILTDALVPYLVGQVVAGACALQLFESHAGHLGPQLFNKFALPYIRDVAQVVKARLREAG	275
FGENESH	RILTDALVPYLVGQVVAGACALQLFESHAGHLGPQLFNKFALPYIRDVAQVVKARLREAG	271
	*****	
GENSCAN	LAPVPMIIFAKDGHFALEELAQAQGYEVVGLDWTVPKKKARECVGKTVTLQGNLDPCALYA	343
UROD	LAPVPMIIFAKDGHFALEELAQAQGYEVVGLDWTVPKKKARECVGKTVTLQGNLDPCALYA	312
GENEID	LAPVPMIIFAKDGHFALEELAQAQGYEVVGLDWTVPKKKARECVGKTVTLQGNLDPCALYA	335
FGENESH	LAPVPMIIFAKDGHFALEELAQAQGYEVVGLDWTVPKKKARECVGKTVTLQGNLDPCALYA	331
	*****	
GENSCAN	SEEEIGQLVKQMLDDFGPHRYIANLGHGLYPMDPEHVGAFVDAVHKHSRLLRQN	398
UROD	SEEEIGQLVKQMLDDFGPHRYIANLGHGLYPMDPEHVGAFVDAVHKHSRLLRQN	367
GENEID	SEEEIGQLVKQMLDDFGPHRYIANLGHGLYPMDPEHVGAFVDAVHKHSRLLRQN	390
FGENESH	SEEEIGQLVKQMLDDFGPHRYIANLGHGLYPMDPEHVGAFVDAVHKHSRLLRQN	386
	*****	

A diferencia de la predicción computacional de genes, el campo de la caracterización de regiones reguladoras no ha reunido todavía un catálogo amplio de anotaciones experimentales para poder evaluar diferentes métodos bioinformáticos. Existe una enorme variedad de componentes de cualquier red regulatoria que desconocemos completamente, de modo que no podemos garantizar que disponemos de la ubicación de todos los elementos que efectivamente gobiernan un promotor en distintas situaciones. De hecho, con la tecnología actual no es posible siquiera delimitar cuál es la región promotora de un determinado gen debido al gran número de interacciones con factores no necesariamente cercanos al inicio de transcripción. Ante esta carencia de anotaciones reales, resulta poco apropiado realizar mediciones de sensibilidad. En estudios de este tipo es habitual formular la especificidad de un sistema de predicción simplemente como el cociente entre el número de sitios de unión predichos y la longitud total de la secuencia. Esta amalgama de distintas limitaciones, sin embargo, no ha impedido que mediante la creación de juegos de pruebas basados en la plantación de motivos reales en secuencias generadas artificialmente sobre colecciones de promotores parcialmente conocidos, podamos evaluar el grado de precisión de distintas aplicaciones bioinformáticas. En líneas generales, no obstante, estas pruebas no han arrojado resultados excesivamente satisfactorios, indicando que el margen de mejora es enorme en esta área de conocimiento.

En lo relativo al cálculo de la precisión de aplicaciones bioinformáticas sobre conjuntos de genes y regiones reguladoras conocidas, debemos mencionar que existen herramientas experimentales para realizar la validación de determinadas predicciones que han superado distintos filtros de calidad durante el procesamiento bioinformático. Lógicamente, este paso del método científico resulta más costoso y delicado en términos de tiempo y personal empleados para este fin. Varias metodologías experimentales han sido desarrolladas para efectuar estas pruebas de forma eficiente, generalizándose algunas de ellas dentro de plataformas de procesamiento de secuencias biológicas a gran escala (por ejemplo, chips de expresión o secuenciación masiva de muestras de inmunoprecipitaciones). Aunque los resultados de estos experimentos requieren de un tratamiento bioinformático específico, la combinación transversal de todas estas aproximaciones promete revelar en los próximos años nuevas facetas de la arquitectura regulatoria que gobierna con precisión los genes codificados en el genoma.

### Lecturas complementarias

J. Fickett; A. Hatzigeorgiou (1997). "Eukaryotic promoter recognition". *Genome Research* (núm. 7, págs. 861-878).

T. Abeel; Y. Van de Peer; Y. Saeys (2009). "Toward a gold standard for promoter prediction evaluation". *Bioinformatics* (págs. i313-i320).

M. Tompa y otros (2005). "Assessing computational tools for the discovery of transcription factor binding sites". *Nature Biotechnology* (núm. 23, págs. 137-144).

### Lecturas complementarias

R. Guigó y otros (2003). "Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes". *PNAS* (núm. 100, págs. 1140-1145).

L. Elnitski; V. Jin; P. Farnham; S. Jones (2006). "Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques". *Genome Research* (núm. 16, págs. 1455-1464).

## Resumen

El estudiante ha aprendido durante este módulo cómo es conceptualmente la estructura de los genes codificados en el genoma y qué componentes poseen las regiones que regulan su expresión. Mediante el modelado estadístico de su composición, es posible construir herramientas computacionales para identificar nuevas ocurrencias de miembros de una determinada familia de señales en secuencias previamente sin analizar. Con estos conocimientos, el bioinformático puede elaborar hipótesis sobre la estructura exónica de un gen o la arquitectura de un promotor, introduciendo información sobre conservación evolutiva para reforzar las predicciones iniciales. Finalmente, hemos especificado cuál es el marco matemático empleado para evaluar la precisión de estas predicciones informáticas en comparación con un conjunto de anotaciones reales de referencia.

## Actividades

1. Para el gen *UROD*, investigad todas las anotaciones disponibles en las pistas de datos servidas por UCSC, ENCODE, GENCODE, VEGA, CCDS y RefSeq (ver figura 4). Analizad el origen de las isoformas incluidas por cada consorcio en estas anotaciones.
2. Implementad en Perl el algoritmo de reconocimiento de patrones con matrices de pesos (figura 13). Verificad su correcto funcionamiento con los datos suministrados para el promotor del gen humano *LEP*.
3. Extraed 500 bases de la región promotora de todos los genes humanos anotados en el navegador genómico UCSC, según la pista RefSeq. Realizad la predicción de cajas TATA con la matriz de pesos mostrada en la figura 5 y estimad qué proporción de genes puede contener en su promotor este motivo.
4. Para contrastar con los datos mostrados para el genoma humano y la mosca de la fruta (ver figura 21), intentad averiguar las características básicas de las señales de ajuste en el genoma del ratón.
5. Extraed del genoma humano una secuencia que codifique en su interior cualquier gen conocido. Escribid un programa en Perl que cargue en memoria la tabla de uso de codones incluida en estos materiales. Posteriormente, emplead dicha tabla dentro del programa para analizar la región extraída estudiando el comportamiento de los resultados en comparación con la ubicación de los exones conocidos.
6. Realizad un trabajo de investigación bibliográfica sobre la predicción computacional de selenoproteínas. Enfocad el análisis desde el punto de vista de las proteínas conocidas de esa familia, su función biológica, las modificaciones introducidas en el programa GENEID para su detección, el uso de información sobre estructuras secundarias del ARN de dichos transcritos y el proceso de entrenamiento. Finalizad el trabajo con el estudio del catálogo de selenoproteínas conocidas actualmente.
7. Profundizad en las distintas implicaciones de la existencia del denominado código de las histonas. Realizad un estudio de las asociaciones entre modificaciones post-traduccionales de éstas y la detección de regiones promotoras. Investigad sobre la existencia de intensificadores activos o pasivos en función de la organización espacial de la cromatina.
8. Conectaos a la página web del programa GENEID y obtened una copia gratuita de su versión 1.2. Acceded a la información de soporte y, posteriormente, ejecutad los distintos tests sobre los conjuntos de prueba incluidos en esta distribución. Estudiad los ficheros de parámetros para varias especies. Finalmente, explorad el código fuente para identificar los diferentes componentes de su arquitectura, mostrada en la figura 25.
9. Reproducíd la predicción del gen *UROD* mediante los programas GENEID, GENSCAN y FGENESH. Contrastad las diferentes predicciones con la proteína real. Enriqueced estos resultados iniciales buscando información sobre la conservación de cada exón predicho en otras especies (por ejemplo, ratón, vaca, pez cebra).
10. Repetid el mismo procedimiento con otro gen humano (predicción *ab initio*, genómica comparativa y contraste con la proteína real).
11. Analizad el algoritmo GenAmic y estudiad el artículo original para comprender todas las ordenaciones necesarias para reducir el coste. Prestad atención especial al modo en que la técnica de programación dinámica es utilizada para solucionar este problema.
12. Analizad la integración derivada de alineamientos entre el genoma humano y del ratón implementada en las publicaciones de los programas SGP y TWINSKAN.
13. Localizad el promotor ortólogo del gen *LEP* en distintas especies de vertebrados (observad el ejemplo del gen *MMP13*). Emplead el programa MEME para identificar el grado de conservación de los tres sitios de unión conocidos en todas estas secuencias.
14. Analizad las mismas secuencias promotoras de la pregunta anterior con la técnica del metaalineamiento, empleando el servidor web implementado por sus autores.
15. Explorad el genoma humano y escoged un gen que como mínimo posea cinco exones. Seleccionad su anotación según RefSeq como referencia y realizad la predicción con GENEID y GENSCAN. Finalmente, evaluad la precisión de sus predicciones en comparación con la anotación de referencia establecida anteriormente.
16. Diseñad un protocolo automático de medición de la precisión de las predicciones de un programa. Este procedimiento debe generar las predicciones a partir de un conjunto de

secuencias previamente seleccionadas y, posteriormente, evaluar todas las métricas presentadas a lo largo de este capítulo.

## Ejercicios de autoevaluación

1. Enumerad cinco componentes del paisaje genómico en eucariotas.
2. ¿Cuál es el primer paso en un protocolo de anotación de un nuevo genoma?
3. En la anotación de genes, ¿en qué suele consistir la primera fase?
4. Describid cómo interactúan el proceso de anotación computacional y el refinamiento manual de anotaciones.
5. Definid los conceptos de señal y región biológica.
6. Definid una matriz de pesos.
7. Identificad las diferencias más relevantes entre secuencias consenso y matrices de pesos.
8. Definid qué es una razón de verosimilitud en el contexto de las matrices de pesos.
9. ¿Para qué resulta útil aplicar logaritmos en estos cálculos? ¿Y las pseudocuentas?
10. Describid en pocas palabras el algoritmo de *pattern-matching* de motivos.
11. ¿En qué circunstancias resultan útiles las cadenas de Markov?
12. Justificad qué equivalencia existe entre una matriz de pesos y una cadena de Markov.
13. Enumerad las cuatro clases de exones posibles.
14. Enumerad las cuatro clases de señales que pueden delimitar las regiones exónicas.
15. Explicad brevemente en qué consiste el sesgo de codones en regiones codificantes.
16. Enumerad los cuatro tipos de regiones reguladoras de la transcripción génica.
17. ¿Cuál es la principal diferencia entre predicción de genes intrínseca y extrínseca?
18. Describid el contenido general de un fichero de parámetros del programa GENEID.
19. Explicad en pocas palabras la estrategia que se debe seguir para anotar un gen empleando diferentes programas de predicción y otras bases de datos.
20. ¿Qué dos tipos de información por homología se emplean habitualmente?
21. Explicad brevemente qué información podéis explorar en TRANSFAC o JASPAR.
22. Describid qué es el *phylogenetic footprinting*.
23. ¿Cuál es la principal ventaja de emplear metaalineamientos de mapas de señales?
24. Definid en pocas líneas las dos métricas más representativas en el campo de la predicción genómica: sensibilidad y especificidad.
25. ¿Por qué no es recomendable medir directamente la especificidad de una serie de predicciones de sitios de unión a factores de transcripción?

## Solucionario

1. Genes, moléculas de ARN no codificante, pseudogenes, elementos regulatorios y transposones.
2. Reconstruir el catálogo de genes codificados en su secuencia.
3. Inicialmente se recuperan aquellas proteínas similares a ejemplos ya conocidos mediante búsquedas masivas en distintas bases de datos.
4. Tras una primera anotación a gran escala mediante procedimientos computacionales, diferentes expertos humanos en estructuras génicas proceden a depurar las primeras anotaciones en función de la calidad de las evidencias que soporta cada gen.
5. Una señal es una ubicación concreta de la secuencia que posee funcionalidad biológica para ser reconocible por determinados factores que procesan el genoma. Una región genómica que codifica cierta información está delimitada por una señal en cada flanco.
6. Una matriz de pesos es una representación numérica que resume la composición de nucleótidos en cada posición de un motivo reconocible en diferentes situaciones.
7. El consenso representa únicamente aquellas tendencias más claras dentro de un conjunto de motivos relacionados. La matriz de pesos permite discriminar esas diferencias más sutilmente, mediante su composición numérica.
8. El cálculo de una razón de verosimilitud entre dos modelos permite inferir numéricamente a cuál de los dos puede pertenecer una nueva secuencia, según el resultado de evaluar el cociente de la puntuación obtenida por separado para ésta.
9. El uso de logaritmos reduce la magnitud de los cálculos, simplificando las operaciones de comparación entre modelos. Las pseudocuentas se introducen para evitar cálculos de logaritmos sobre valores nulos.
10. Este algoritmo se basa en el desplazamiento de una ventana ficticia a lo largo de la secuencia para aplicar una matriz de pesos sobre cada motivo fijado en su interior.
11. Los modelos de Markov son herramientas útiles para detectar dependencias entre grupos de símbolos consecutivos de una secuencia.
12. Una matriz de pesos convencional equivale a una cadena de Markov de orden cero.
13. Inicial, interno, terminal y único.
14. Aceptor, donador, inicio y parada de traducción.
15. Por su propia naturaleza, existen codones cuya presencia en regiones que codifican para proteínas es más frecuente de lo esperado. Además, entre dos aminoácidos consecutivos dentro del mismo péptido también existen dependencias que pueden capturarse analizando dos codones adyacentes de la secuencia genómica.
16. Promotores basales, proximales, intensificadores e intrones.
17. La predicción intrínseca (*ab initio*) únicamente emplea información estadística recopilada de colecciones de exones conocidos, utilizando estos modelos para buscar nuevas ocurrencias de esa familia de señales o exones en secuencias sin anotar previamente. La predicción extrínseca emplea comparaciones explícitas entre nuestra secuencia de trabajo y bancos de proteínas conocidas para identificar secuencias similares.
18. Contiene los modelos estadísticos para identificar señales que pueden flanquear exones codificantes junto con las propiedades del sesgo de codones típico de esa especie (todo implementado con cadenas de Markov de diferentes órdenes). También posee un modelo de reglas para ensamblar determinados exones únicamente cuando la conexión de esas clases está permitida.
19. La estrategia básica consiste en extraer primero aquellas predicciones consistentemente producidas por la mayoría de programas, para después reforzarlas mediante información de homología con otras proteínas y otros genomas.
20. Proteínas y transcritos conocidos junto con regiones ortólogas en otros genomas.

21. Ambas son bases de datos sobre regulación génica. Poseen descripciones de miles de factores de transcripción, junto con modelos predictivos para reconocer los sitios de unión más característicos.
22. Es el reforzamiento de un sitio de unión predicho a partir del grado de conservación de éste en regiones ortólogas de otras especies para el mismo gen.
23. La principal ventaja de efectuar el alineamiento de mapas de predicciones de sitios de unión entre ADN y factores de transcripción es que permite reconocer coincidencias entre motivos equivalentes que presentan cierta variabilidad debido al proceso evolutivo. Es posible que estas señales no fueran capturadas empleando una comparación a nivel de secuencia.
24. La sensibilidad resulta útil para evaluar qué porción de la anotación conocida acierta un sistema de predicciones. La especificidad, en cambio, permite cuantificar el porcentaje de las predicciones que efectivamente puede usarse para reconocer elementos reales.
25. Porque no podemos garantizar con el estado actual de las anotaciones existentes que conozcamos todos los sitios funcionales de una región promotora. De este modo, podríamos certificar que una predicción es errónea aunque simplemente ocurra que no existe suficiente conocimiento sobre esa red regulatoria.



## Bibliografía

**Abeel, T.; Van De Peer, Y.; Saeys, Y.** (2009). "Toward a gold standard for promoter prediction evaluation". *Bioinformatics* (págs. i313-i320).

**Bailey, T. L. y otros** (2009). "MEME Suite: tools for motif discovery and searching". *Nucleic Acids Research* (núm. 37, págs. W202-W208).

**Barrera, L.; Ren, B.** (2006). "The transcriptional regulatory code of eukaryotic cells—insights from genome-wide analysis of chromatin organization and transcription factor binding". *Current Opinion in Cellular Biology* (núm. 18, págs. 291-298).

**Bejerano, G. y otros** (2004). "Ultraconserved elements in the human genome". *Science* (núm. 304, págs. 1321-1325).

**Birney, E.; Durbin, R.** (2000). "Using GeneWise in the *Drosophila* annotation experiment". *Genome Research* (núm. 10, págs. 547-548).

**Blanco, E.; Parra, G.; Guigó, R.** (2003). *Using geneid to identify genes*. *Current Protocols in Bioinformatics*. Nueva York: John Wiley & Sons Inc. ISBN: 0471250937.

**Blanco, E.; Guigó, R.** (2005). *Predictive methods using DNA sequences Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Nueva York: John Wiley & Sons Inc. ISBN: 0471478784.

**Blanco, E.; Farre, D.; Alba, M.; Messeguer, X.; Guigó, R.** (2006). "ABS: a database of Annotated regulatory Binding Sites from orthologous promoters". *Nucleic Acids Research* (núm. 34, págs. D63-D67).

**Blanco, E.; Messeguer, X.; Smith, T.; Guigó, R.** (2006). "Transcription factor map alignments of promoter regions". *PLoS Computational Biology* (núm. 2, pág. e49).

**Blanco, E.; Guigó, R.; Messeguer, X.** (2007). "Multiple non-collinear TF-map alignments of promoter regions". *BMC Bioinformatics* (núm. 8, pág. 138).

**Blanco, E.; Pignatelli, M.; Beltran, S.; Punset, A.; Perez-Lluch, S.; Serras, F.; Guigó, R.; Corominas, M.** (2008). "Conserved chromosomal clustering of genes governed by chromatin regulators in *Drosophila*". *Genome Biology* (núm. 9, pág. R134).

**Blanco, E.** (2011). *Computational characterization of regulatory regions. Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications*. Hoboken, NJ: Wiley-Blackwell (John Wiley & Sons Ltd). ISBN-13: 978-0470505199.

**Brazma, A. y otros** (1998). "Approaches to the automatic discovery of patterns in biosequences". *Journal of Computational Biology* (núm. 5, págs. 279-305).

**Brent, M. R.** (2008). "Steady progress and recent breakthroughs in the accuracy of automated genome annotation". *Nature Reviews Genetics* (núm. 9, págs. 62-73).

**Bucher, P.** (1990). "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences". *Journal of Molecular Biology* (núm. 212, págs. 563-578).

**Burge, C.; Karlin, S.** (1997). "Prediction of complete gene structures in human genomic DNA". *Journal of Molecular Biology* (núm. 268, págs. 78-94).

**Burset, M.; Guigó, R.** (1996). "Evaluation of gene structure prediction programs". *Genomics* (núm. 34, págs. 353-367).

**Crooks, G. y otros** (2004). "WebLogo: A sequence logo generator". *Genome Research* (núm. 14, págs. 1188-1190).

**Curwen, V.; Eyra, E.; Andrews, T. D. y otros** (2004). "The Ensembl automatic gene annotation system". *Genome Research* (núm. 14, págs. 942-950).

**Davuluri, R.; Grosse, I.; Zhang, M.** (2001). "Computational identification of promoters and first exons in the human genome". *Nature Genetics* (núm. 29, págs. 412-417).

**Davidson, E.** (2006). *The regulatory genome*. San Diego: Elsevier, Academic Press. ISBN: 9780120885633.

**Dermitzakis, E. T.; Clark, A. G.** (2002). "Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover". *Molecular Biology and Evolution* (núm. 7, págs. 1114-1121).

**Durbin, R.; Eddy, S.; Crogh, A.; Mitchison, G.** (1998). *Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids*. Cambridge: Cambridge University Press. ISBN: 0521629713.

**Duret, L.; Bucher, P.** (1997). "Searching for regulatory elements in human noncoding sequences". *Current Opinion in Structural Biology* (núm. 7, págs. 399-406).

**Elnitski, L.; Jin, V.; Farnham, P.; Jones, S.** (2006). "Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques". *Genome Research* (núm. 16, págs. 1455-1464).

**Fickett, F.; Tung, C.** (1992). "Assessment of protein coding measures". *Nucleic Acids Research* (núm. 20, págs. 6441-6-450).

**Fickett, F.; Hatzigeorgiou, A.** (1997). "Eukaryotic promoter recognition". *Genome Research* (núm. 7, págs. 861-878).

**K. Frech, K.; Herrmann, G.; Werner, T.** (1993). "Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids". *Nucleic Acids Research* (núm. 21, págs. 1655-1664).

**Gerstein, M. y otros** (2007). "What is a gene, post-ENCODE? History and updated definition". *Genome Research* (núm. 17, págs. 669-681).

**Gross, S.; Brent, M.** (2006). "Using multiple alignments to improve gene prediction". *Journal of Computational Biology* (núm. 13, págs. 379-393).

**Guigó, R.; Knudsen, S.; Drake, N.; Smith, T.** (1992). "Prediction of gene structure". *Journal of Molecular Biology* (núm. 226, págs. 141-157).

**Guigó, R.** (1998). "Assembling genes from predicted exons in linear time with dynamic programming". *Journal of Computational Biology* (núm. 5, págs. 681-702).

**Guigó, R.** (1999). *DNA composition, codon usage and exon prediction*. Genetic Databases. Academic Press. ISBN: 0121016250.

**Guigó, R. y otros** (2003). "Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes". *PNAS* (núm. 100, págs. 1140-1145).

**Guigó, R. y otros** (2006). "EGASP: the human ENCODE genome annotation assessment project". *Genome Biology* (núm. 7, pág. S2).

**Harrow, J.; Nagy, A.; Reymond, A.; Alioto, T.; Patthy, L.; Antonarakis, S.; Guigó, R.** (2009). "Identifying protein-coding genes in genomic sequences". *Genome Biology* (núm. 10, pág. 201).

**Harrow, J. y otros** (2006). "GENCODE: producing a reference annotation for ENCODE". *Genome Biology* (núm. 7, pág. S4).

**Haussler, D.** (1998). "Computational genefinding". *Trends in Genetics (Trends guide to bioinformatics)* (págs. 12-15).

**Hsu, F.; Kent, W. J.; Clawson, H.; Kuhn, R. M.; Diekhans, M.; Haussler, D.** (2006). "The UCSC Known Genes". *Bioinformatics* (núm. 22, págs. 1036-1046).

**Korf, I.; Flicek, P.; Duan, D.; Brent, M.** (2001). "Integrating genomic homology into gene structure prediction". *Bioinformatics* (núm. 17, págs. S140-S148).

**Kornblihtt, A.** (2005). "Promoter usage and alternative splicing". *Current Opinion in Cell Biology* (núm. 17, págs. 262-268).

**Kouzarides, T.** (2007). "Chromatin modifications and their function". *Cell* (núm. 128, págs. 693-705).

**Levine, M.; Tijan, R.** (2003). "Transcriptional regulation and animal diversity". *Nature* (núm. 424, págs. 147-151).

**Mgc Project Team** (2009). "The completion of the Mammalian Gene Collection (MGC)". *Genome Research* (núm. 19, págs. 2324-2333).

**Parra, G.; Agarwal, P.; Abril, J.; Wiehe, T.; Fickett, J.; Guigó, R.** (2003). "Comparative gene prediction in human and mouse". *Genome Research* (núm. 13, págs. 108-117).

**Parra, G.; Blanco, E.; Guigó, R.** (2000). "GeneID in *Drosophila*". *Genome Research* (núm. 10, págs. 511-515).

**Perier, R. C. y otros** (2000). "The Eukaryotic Promoter Database (EPD)". *Nucleic Acids Research* (núm. 28, págs. 302-303).

**Portales-Casamar, E. y otros** (2010). "JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles". *Nucleic Acids Research* (núm. 38, págs. D105-D110).

**Pruitt, K. D.; Tatusova, T.; Klimke, W.; Maglott, D. R.** (2009). "NCBI Reference Sequences: current status, policy and new initiatives". *Nucleic Acids Research* (núm. 37, págs. D32-D36).

**Pruitt, K. D. y otros** (2009). "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes". *Genome Research* (núm. 19, págs. 1316-1323).

**Reese, M.; Hartzell, G.; Harris, N.; Ohler, U.; Abril, J.; Lewis, S.** (2000). "Genome annotation assessment in *Drosophila melanogaster*". *Genome Research* (núm. 10, págs. 483-501).

**Rogic, S.; Mackworth, A.; Ouellette, F.** (2001). "Evaluation of gene-finding programs on mammalian sequences". *Genome Research* (núm. 11, págs. 817-832).

**Salamov, A.; Solovyev, V.** (2000). "Ab initio Gene Finding in *Drosophila* Genomic DNA". *Genome Research* (núm. 10, págs. 516-522).

**Schneider, T.; Stephens, R. M.** (1990). "Sequence logos: a new way to display consensus sequences". *Nucleic Acids Research* (núm. 18, págs. 6097-6100).

**Staden, R.** (1984). "Computer methods to locate signals in nucleic acid sequences". *Nucleic Acids Research* (núm. 12, págs. 505-519).

**Stormo, G. D.** (2000). "DNA binding sites: representation and discovery". *Bioinformatics* (núm. 16, págs. 16-23).

**The Encode Project Consortium** (2007). "Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project". *Nature* (núm. 447, págs. 799-816).

**The modENCODE Project Consortium** (2010). "Identification of functional elements and regulatory circuits by *Drosophila* modENCODE". *Science* (núm. 330, págs. 1787-1797).

**The Open Regulatory Annotation Consortium** (2008). "OREGAnno: an open-access community-driven resource for regulatory annotation". *Nucleic Acids Research* (núm. 36, págs. D107-D113).

**Tagle, D. y otros** (1988). "Embryonic  $\beta$  and  $\gamma$  globin genes of a prosimian primate, nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints". *Journal of Molecular Biology* (núm. 203, págs. 439-455).

**Tompa, M. y otros** (2005). "Assessing computational tools for the discovery of transcription factor binding sites". *Nature Biotechnology* (núm. 23, págs. 137-144).

**Turatsinze, J.; Thomas-Chollier, M.; Defrance, M.; Van Helden, J.** (2008). "Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules". *Nature protocols* (núm. 3, págs. 1578-1588).

**Wasserman, W. W.; Sandelin, A.** (2004). "Applied bioinformatics for the identification of regulatory elements". *Nature Reviews Genetics* (núm. 5, págs. 276-287).

**Werner, T.** (2000). "Identification and functional modelling of DNA sequence elements of transcription". *Briefings in bioinformatics* (núm. 1, págs. 372-380).

**Wingender, E.** (2008). "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation". *Briefings in Bioinformatics* (núm. 9, págs. 326-332).

**Wray, G.; Hahn, M.; Abouheif, E.; Balhoff, J.; Pizer, M.; Rockman, M.; Romano, L.** (2003). "The evolution of transcriptional regulation in eukaryotes". *Molecular Biology and Evolution* (núm. 20, págs. 1377-1419).

**Zhang, M. Q.** (2002). "Computational prediction of eukaryotic protein-coding genes". *Nature Review Genetics* (núm. 3, págs. 698-709).