

2 Organización de genomas.

El genoma es el conjunto de la información genética de un organismo, es decir, el conjunto de genes, entendiendo como tales sólo a los fragmentos codificantes del ADN, y no los elementos reguladores o espaciadores. Es muy fácil determinar la cantidad e, incluso, la secuencia del ADN, pero detectar en él los genes no es tarea fácil, y mucho más difícil es el estudio de la interacción génica.

Maximización del tamaño del genoma.

Los organismos con menos cantidad de ADN tienen que recurrir a estrategias para almacenar el máximo de información genética en el mínimo espacio. Esta necesidad se ve incrementada cuanto más sencillo es el organismo, mientras que los que son considerados más complejos no recurren a estrategias de almacenamiento porque les sobra ADN. Las estrategias de compresión de la información genética se pueden incluir en las siguientes:

- Genes solapados: Un mismo fragmento de ADN contiene información correspondiente a dos o más genes, es decir, comparten secuencia en determinados sitios. Suele darse en virus, en los que un mensajero tiene varios sitios de inicio de la traducción, pero siempre con la misma pauta de lectura.
- Pauta de lectura alternativa: Un mismo gen puede leerse en varias pautas de lectura para producir hasta tres proteínas distintas. También se encuentra en virus y otros organismos inferiores.
- Poliproteínas: Un solo producto de traducción de un solo mensajero se parte por sitios muy concretos en una serie de proteínas diferentes que, además, suele ser una proteasa que participa en el proceso. Muy común en virus; el VIH utiliza esta estrategia.
- Operones: Un solo elemento de regulación controla la expresión combinada de todos los genes de una vía o ruta metabólica, o de varias funciones metabólicas relacionadas. Permite una adaptación rápida y simultánea de la fisiología celular a los cambios en el entorno. El operón produce la síntesis de un ARNm policistrónico, en el que se encuentran varios genes concatenados, pero cada uno con su comienzo y su final, con lo que las proteínas se sintetizan independientemente. Es el sistema de organización encontrado en bacterias, mitocondrias y cloroplastos.
- Agrupación de genes de ARNr y ARNt: Con el mismo significado y estructura que los operones, al tener una regulación combinada, pero con la salvedad de que no hay productos proteicos. En este caso, también se encuentra en eucariotas, aunque en tal cantidad de copias que no parece tener como fin el almacenamiento, sino la expresión al máximo nivel.
- Procesamiento alternativo: No es un sistema de almacenaje muy económico, sino que está claramente orientado a la funcionalidad y la adaptación. Este sistema de almacenamiento de la información requiere la existencia de genes interrumpidos, es decir, con exones e intrones, y está presente en eucariotas, aumentando la complejidad con la del organismo. La economía de este sistema reside en que, mediante el procesamiento modular de un solo gen, se pueden conseguir varios productos génicos diferentes.

Tipos de genoma.

Hay mucha diferencia entre el número de genes de unos organismos y otros, pero existen dos entidades infectivas que no tienen genoma porque no tienen genes ni codifican para nada pero que son capaces de dirigir en metabolismo celular de uno u otro modo para que fabrique copias de ellos. Estas partículas son los priones y los viroides, de los que los primeros no tienen ácido nucleico, sino sólo una conformación proteica anómala, mientras que los segundos son un simple polirribonucleótido de 359 bases que no codifica nada, pero que fuerza a la célula a replicarlo. Son entidades autorreplicativas, aunque cada uno con un sistema diferente.

Pasando a considerar genomas reales de seres vivos, se observa que sólo los virus almacenan los genes en

cualquier forma de ácido nucleico, es decir, tanto ADN como ARN y de cadena doble o sencilla. Todos los demás organismos celulares mantienen la información genética en forma de ADNds. Como los virus infectan células, siempre pasa su desarrollo por la transcripción de su genoma en ADNds para poder incluirse en el ADN del hospedador.

Como ya se ha apuntado, la cuantificación de ácidos nucleicos es muy fácil, comparado con la detección de genes. Éstos sólo se pueden inferir de la secuencia del ácido nucleico, buscando las pautas de lectura abiertas (ORF; *open reading frames* en inglés). Actualmente se han secuenciado ya el ADN de *E. coli* y *S. cerevisiae*, de los que aún no se conocen los genes, pero se considera que las ORF aportan una sobreestimación de su número.

Tamaño del genoma.

Después de numerosos análisis de muchos organismos de diferentes filos se ha encontrado que existe un tamaño mínimo de genoma para cada uno de ellos, y que éste va en aumento con la complejidad del organismo, lo que en principio parece coherente. Sin embargo, al calcular la cantidad de ADN, dentro de los animales existe una gran diversidad de tamaños de genoma, dándose el caso sorprendente de que la mayoría de los anfibios tienen más que los demás filos, y que el animal con más ADN es el sapo americano *Xenopus laevis*, con una diferencia de casi un orden de magnitud respecto a los mamíferos (tiene 6 veces más ADN que un humano). Pero también se aprecia que la cantidad de ADN no es uniforme ni se aprecia la misma desviación en todos los filos, destacando de nuevo los anfibios, en los que las diferencias entre los dos extremos pueden ser de hasta dos órdenes de magnitud. Esta diferencia se acentúa en las plantas con flor, donde la diferencia es de más de tres órdenes de magnitud.

Estas apreciaciones dieron lugar a la paradoja del valor C (número de pares de bases del organismo), que apuntaba a que si la diferencia en número de pares de bases entre un procariota y un eucariota es de 106 pb, pero la diferencia en la cantidad de genes (aproximando por la cantidad de proteínas) no es de esa magnitud, eso debería significar que cuanto mayor es la complejidad, hay mucha más cantidad de ADN no codificante o espaciador. También plantea esta paradoja si es necesario ese ADN o si, por el contrario, los organismos superiores se han ido construyendo a base de repeticiones de nada o de ADN basura.

Complejidad genética.

La paradoja del valor C se puede estudiar midiendo la complejidad genética, que es un parámetro que relaciona el número de genes (cantidad de secuencias distintas) por unidad de ADN. El problema sigue siendo, sin embargo, el mismo: ¿cómo medir la cantidad de genes?.

La aproximación utilizada consiste en suponer que, como cada gen es diferente de los demás porque codifica para proteínas diferentes, el tiempo que tarde en renaturalizar una cantidad de ADN de un determinado organismo deberá estar en relación con la cantidad de secuencias distintas, cada una capaz de aparear sólo con su homóloga. Se puede representar, de hecho, la renaturalización como una reacción química de segundo orden:

$2\text{ADNss} \rightarrow \text{ADNds}$

Según la hipótesis planteada, si hay más genes en una molécula de ADN porque ésta es mayor, también hay más secuencias diferentes y la reasociación será más difícil y, por tanto, más lenta. Al partir de la misma cantidad de ADN de dos organismos distintos, de cada uno de ellos habrá más o menos copias, en relación inversamente proporcional a la cantidad de ADN que tengan por célula. Esto redundará en una mayor o menor concentración, de hecho, de ADNss que puede luego reasociar. Para seguir el experimento no hace falta más que ir cuantificando la cantidad de ADNss mediante métodos espectroscópicos. Recurriendo a la ecuación linealizada de la cinética de segundo orden sabemos que:

$$\frac{1}{C} = \frac{1}{C_0} + Kt$$

Esta representación genera una recta de pendiente negativa, pero no se utiliza directamente, sino a través del parámetro $Cot^{1/2}$, que realmente es el inverso de K, es decir, que cuanto menor sea K (reasociación más lenta), mayor será este parámetro (mayor complejidad). Este valor se halla experimentalmente de la curva que resulta de representar $1 - \frac{C}{C_0}$

(porción reasociada) frente a $\log(C_0 \cdot t)$, que es una curva sigmoidea. Al extrapolar el valor de $\log(C_0 \cdot t)$ cuando la reasociación es del 50% se obtiene este parámetro. A menor complejidad, menor será el $Cot^{1/2}$, pero para que sirva de referencia debe partirse siempre de la misma cantidad de ADN, que se ha establecido como 12 pg. De este modo, en esa cantidad de ADN hay unas 750 veces más copias de *E. coli* que de humano, por lo que el genoma de la bacteria reasocia mucho más rápido, pero no 750 veces más, como se verá más adelante.

El $Cot^{1/2}$ también puede dar una idea del número de secuencias distintas en un genoma, es decir, del número de genes. Para esto se estableció el patrón de *E. coli*, para relacionar todos los $Cot^{1/2}$ frente al suyo y determinar el número de pares de bases de complejidad. Es decir, la complejidad genética es un número de pares de bases que se obtiene de esta regla de tres:

$$\frac{Cot \text{ del organismo que se analiza}}{Cot \text{ de } E. coli} = \frac{\text{complejidad en n de pb}}{4,2 \cdot 10^6 \text{ pb de } E. coli}$$

Evidentemente, la complejidad y el número de pares de bases real no son el mismo número, aunque pueden coincidir. En el caso del patrón, todo su ADN está constituido por secuencias únicas, por lo que la complejidad sí coincide con el tamaño total de ADN. Hay que destacar que la complejidad entonces es **la longitud total de ADN de secuencia única**, y no cualquier otra cosa imaginable o más intuitiva.

En los eucariotas la situación se complica, porque las curvas no son sencillas como las de los procariotas, sino que presentan tres escalones, cada uno con un $Cot^{1/2}$ característico. Según la hipótesis de trabajo, el $Cot^{1/2}$ más bajo debe representar secuencias de ADN muy repetidas, el $Cot^{1/2}$ intermedio se correspondería con secuencias moderadamente repetidas y el $Cot^{1/2}$ mayor pertenecería a ADN de secuencia única, es decir, a los genes. Dado que la representación se realiza frente al % de ADN reasociado, se puede calcular el porcentaje de cada una de estas tres especies de ADN eucariótico, y corresponden aproximadamente al 25, 30 y 45%, respectivamente. También se puede hallar la complejidad de cada parte, pero teniendo en cuenta que la C_0 va variando según se van completando reasociaciones de los fragmentos más repetidos, de modo que la C_0 del ADN de secuencia única será la inicial menos la que ya se ha reasociado en las etapas anteriores. De esta aproximación se obtienen para un organismo hipotético que podría ser un mamífero los valores de 340, $6 \cdot 10^5$ y $3 \cdot 10^8$ pb de secuencias distintas en cada clase. De la relación entre la cantidad de ADN que corresponde a cada clase y la complejidad se puede calcular el número de copias de cada una, y se obtiene que el ADN muy repetido consta de unas 500000 copias, el ADN moderadamente repetido consta de unas 350 copias y el ADN de secuencia única no está repetido, evidentemente.

De estos estudios de complejidad de genomas se pueden deducir dos cosas:

- La mayor cantidad de ADN implica una mayor complejidad cinética debida a una mayor proporción de secuencias únicas y, por tanto, de genes.
- Los genes no constituyen todo el ADN, hay secuencias que no lo son y sólo hacen aumentar el tamaño de la molécula sin contribuir a la complejidad genética efectiva (la del ADN de secuencia única).

ADN codificante.

El ADN no repetido se corresponde con los genes estructurales (que codifican para proteínas), que son los que dan características fenotípicas. Éste constituye el 100% del genoma de bacterias, el 80% en plantas, anfibios y eucariotas sencillos y entre el 40 y el 50% en animales. La localización de los genes en este tipo de ADN se determinó de modo análogo a los experimentos de complejidad genética; mediante ensayos de asociación de trazas de ARNm marcado sobre ADN, se vio que había complementariedad con el ADN moderadamente repetido (un 10% del ARNm), pero sobre todo con la parte de ADN de secuencia única (el 90% restante). Los genes para ARNr y ARNt se encuentran en la fracción de ADN moderadamente repetido.

Con un planteamiento inverso, se puede determinar cuánto ADN de secuencia única se expresa en la célula en un momento o tejido dado, al hacer hibridar ADN frente a un exceso de ARNm. Raramente se supera el 2%, obteniéndose como media un 1,3% de ADN expresado. Parte de estos genes expresados son comunes a todas las células de un mismo organismo, o bien están siempre constitutivamente activados porque codifican para enzimas del metabolismo basal, pero parte de ellos dependen del tejido y la situación de la célula individual. El resultado de estos experimentos es que sólo se expresan unos 10000 genes por célula en un organismo eucariota.

El mismo experimento, pero con un enfoque distinto, permite determinar la complejidad del repertorio de ARNm. Los resultados son coherentes con la expresión de unos 100 genes, cada uno en unas 1000 ó 10000 copias, procedentes de ADN moderadamente repetido, y unos 10000 genes, cada uno con 5 a 10 copias, procedentes de los genes estructurales.

Familias génicas.

Una familia génica es un conjunto de genes con un alto grado de homología que, sin llegar a ser alelos (variantes de un mismo gen) porque codifican para productos distintos, poseen un parentesco evolutivo que es consistente con un proceso de copia, diversificación y evolución independiente. Es importante no confundirlo con ADN repetido. Generalmente las familias génicas se encuentran agrupadas, y los genes que las constituyen distan unos pocos Kpb unos de otros, pero son independientes, cada uno con sus propios elementos de regulación que, a veces, actúan de manera totalmente independiente. Sin embargo, se conoce desde hace mucho tiempo que familias génicas como, por ejemplo, la de las hemoglobinas, aun constando de proteínas independientes, posee una regulación conjunta de todos los genes durante el desarrollo, de modo que se puede considerar la existencia de un elemento de regulación de la familia entera. Se han descubierto elementos de ese tipo a 5' del locus de la hemoglobina, denominados LCR (*locus controlling regions*: regiones controladoras del locus), pero no se sabe cómo actúan, sobre todo porque controlan la expresión de genes que están a Mpb de distancia.

Muchas veces, durante la evolución de una copia de un gen determinado, el producto no es funcional, o impide el buen funcionamiento en uno u otro sentido, por lo que se selecciona negativamente y se inactiva permanentemente. Este gen nunca transcrito se llama pseudogén, porque tiene todos los elementos necesarios para poder llegar a sintetizarse, excepto alguno esencial (el promotor, el codón de iniciación, etc.). A pesar de que es un gen inútil, sobre él pueden desarrollarse fenómenos de mutación al azar que deriven en un gen funcional y reviertan la inactivación o, a corto plazo, es parte activa en procesos de recombinación genética porque su secuencia es altamente homóloga con la de los genes funcionales de su alrededor. Este hecho permite la aparición de mutantes híbridos, como es el caso de la mayoría de las talasemias que conciernen al locus de la hemoglobina.

Los ARNr son un ejemplo de familia génica; no por las posibles analogías entre cada uno de ellos, sino en que en una célula eucariota suele haber cientos de repeticiones en tándem del conjunto de genes que los codifican. En el humano son unas 200 copias repartidas en cinco cromosomas. Cada una de estas copias tiene su propio promotor, pero en todas es el mismo, claro.

ADN repetido.

Repeticiones de di- y trinucleótidos.

- Islas CG: Se encuentran aproximadamente 1 Kpb a 5' de muchos genes, antes del promotor. Las repeticiones suelen ser de decenas, con un patrón de distribución del dinucleótido característico. Tanto la rareza como esta especial distribución y concentración, que no se da en otras partes del genoma, hacen suponer una misión reguladora a estas secuencias. De hecho, se han encontrado elementos conocidos de regulación solapando con ellas (cajas), por lo que probablemente modulen la unión a ese elemento.
- Agrupaciones de AT y de CA, sobre todo, que abundan en el ADN espaciador entre genes.
- Repeticiones de CA en el ADN centromérico.
- ADN telomérico: Consta de secuencias altamente repetidas (cientos de veces) de 4 a 8 nt. Favorecen la replicación del ADN en el extremo lineal del cromosoma. Parece que, además, ayuda a estabilizar los cromosomas y evitar su degradación por esos extremos, pero también parece limitar el número de potenciales divisiones celulares. La enzima que añade este polinucleótido (diferente para cada especie) es la telomerasa, que sintetiza ADN independientemente de molde.
- Expansión de trinucleótidos: En ciertos genes hay repeticiones de un trinucleótido entre 10 y 100 veces, tanto en la secuencia codificante como en los intrones y las 3' y 5' UTR (*untranslated terminal region*: región no traducida terminal; hay dos presentes en todos los ARNm y que se encuentran antes y después de los codones de inicio y terminación, respectivamente, con las secuencias de adhesión y disociación del ribosoma y con implicaciones en la expresión génica). Este fenómeno se conoce muy bien en la enfermedad de la corea de Huntington, en la que el gen que codifica para la proteína responsable, la huntingtina, posee entre 10 y 30 repeticiones del trinucleótido CAG cuando es funcional, y la enfermedad se desarrolla cuando se sobrepasa cualquiera de los dos límites, porque la proteína pierde su función, y la gravedad es proporcional al número de repeticiones de más.

Repeticiones de secuencias largas.

- SINE (secuencias cortas dispersas entre genes): Son secuencias de unos cientos de pb, presentes en un muy elevado número de copias. Las que mejor se conocen son las de la familia Alu, cada una con 300 pb y distribuidas en 500000 copias, pero en cualquier caso no se sabe qué son ni por qué permanecen en el genoma, cuando no parecen tener ninguna función. Se sabe que estos elementos son capaces de incluirse en otro sitio del genoma porque se han detectado casos de mutaciones en las que un exón quedaba interrumpido por una de estas secuencias que antes no estaba ahí. Se dice entonces que la mutación es *knockin* (mutación por inclusión de material genético). No se conoce el mecanismo de replicación de estas secuencias, pero se supone que deben, en algún momento, traducirse (no se sabe si con un promotor propio o con otro cercano) para después retrotranscribirse (como los retrovirus; posibles implicaciones evolutivas) en ADNds y posteriormente incluirse al azar en el genoma. Lo que es evidente es que juegan un papel importante en la construcción de genomas y que, probablemente, lo han hecho a lo largo de la evolución generando una gran variabilidad.
- LINE (secuencias largas dispersas entre genes): Son secuencias de unos 7000 pb que se encuentran flanqueando genes. Por su tamaño forman parte del ADN moderadamente repetido. Se cree que favorecen la recombinación y el sobrecruzamiento, por lo que su principal misión sería favorecer la diversificación genética. También tiene importantes implicaciones evolutivas, ya que alguna vez la recombinación puede no ser homóloga y provocar, por ejemplo, la unión o separación de cromosomas.

Número de genes.

La pregunta que se formulaba anteriormente acerca de cuántos genes tiene un organismo se puede contestar de varios modos:

- Mediante secuenciación directa y análisis de ORF: Es el método más directo pero da una sobreestimación, al poder incluirse los pseudogenes como genes funcionales.
- Estimaciones indirectas:
- Estructura y organización de los genes.
- Tamaño y complejidad del genoma.
- Expresión de genes mediante sondas de ARNm.

Sin embargo, con estas técnicas no se puede llegar a saber la distribución ni la densidad de genes en el genoma.

Las estimaciones de cada abordaje difieren en muchos miles de genes (entre 50000 y 150000 para el ser humano), por lo que la precisión a la que se puede llegar es muy vaga.

Genoma de orgánulos.

Viendo que había casos de segregación somática de caracteres (durante la mitosis) que no seguían las leyes de Mendel, así como caracteres que se heredaban de manera atípica (herencia materna), se llegó a la conclusión de que debía haber genes fuera de los núcleos. Efectivamente, las mitocondrias y los cloroplastos son orgánulos que tienen su propio material genético, con determinados genes exclusivos y toda la maquinaria de regulación, transcripción y traducción necesaria (con código genético propio en las mitocondrias). La actividad y replicación del genoma extranuclear son totalmente autónomas, así como la regulación, aunque en determinados casos está coordinada con el núcleo.

Se pueden establecer árboles filogenéticos para cada orgánulo, habiéndose considerado una tasa de mutación de un 2 a 4% cada millón de años, que representa unas diez veces más que la correspondiente al genoma nuclear. La tasa total de divergencia del ADN mitocondrial humano es de 0,57%, y proviene de una mujer que vivía en África hace unos 200000 años.

Mitocondrias.

El genoma mitocondrial humano (ADNm) consta de 5 a 10 copias de un solo cromosoma circular cerrado de 16596 pb, que codifica para 13 proteínas de la cadena de transporte electrónico y todos los ARNr y ARNt necesarios para la traducción.

Es un genoma extremadamente compacto, hasta tal punto que no suele haber separación entre la terminación de una proteína y el comienzo de otra, con sólo dos promotores, uno en cada sentido, que transcriben sendos mensajeros policistrónicos: uno casi tan largo como todo el ADN, y otro sólo para la subunidad 6 de la ATPasa y la subunidad 3 de la citocromo oxidasa; en cualquier caso, todas las proteínas son mitocondriales. La ARN polimerasa mitocondrial se sintetiza en el núcleo desde un gen específico (es distinta a las demás nucleares) y luego se importa. Los dos ARNr de los que constan los ribosomas mitocondriales son propios, pero las 75 proteínas que se necesitan también se sintetizan en el núcleo. Todos los ARNt se sintetizan en la mitocondria, pero las aa-ARNt sintetizadas mitocondriales provienen también del núcleo. En resumen, todos los ARN propios de la mitocondria se conservan en ella, pero las proteínas que participan en la transcripción y la traducción son todas nucleares. Esta situación responde a una gran economía de recursos, además de un ejemplo de maximización del almacenamiento de la información, pero obliga a la expresión coordinada de todos los genes implicados.

Las mitocondrias de levaduras no presentan tal cantidad de compactación en su genoma, sino que tienen varios promotores y ADN espaciador, e incluso genes interrumpidos con intrones que se autoescinden (tipo II), y conservan muchas más funciones que las mitocondrias de eucariotas más evolucionados.

Cloroplastos.

Los cloroplastos son orgánulos mucho mayores que las mitocondrias, y conservan muchos más genes propios que ellas. La referencia es *Euglena gracilis*, cuyo genoma cloroplasmático consta de unas 50 copias de una molécula de ADN circular cerrada de 143172 pb, pero acerca de la estructura fina se sabe muy poco, porque es sabido que los vegetales y los microorganismos no patógenos siempre son menos atractivos de estudiar que los animales.

Teoría endosimbiótica (opcional).

Desde un punto de vista estático, por un lado, el genoma extranuclear permite que el núcleo pueda delegar la tarea de fabricar determinadas proteínas específicas de ese orgánulo, con lo que se ahorran recursos al no tener que transportar los componentes hasta su destino. Por otro lado, se hace necesario un mecanismo de control de la autonomía del orgánulo, así como una regulación coordinada de determinados genes cuyos productos deben interaccionar para formar partículas o productos.

A la vista de la disminución del genoma mitocondrial según se asciende en la complejidad de los organismos, además de la propia estructura de los orgánulos, su fisiología, el tipo de genoma (circular cerrado) y la propia existencia de ADN distinto del del núcleo, con funciones propias y evolución separada (código genético propio), hicieron suponer la teoría endosimbiótica. Esta teoría propone que tanto mitocondrias como cloroplastos son formas reducidas de una bacteria aerobia y de un alga cianofícea, respectivamente, que fueron fagocitadas por una primitiva célula eucarionte anaerobia, pero no digeridas, con lo que se benefició de la energía que le proporcionaba. Este modelo de eucarionte debió de tener tal éxito que eliminó a todos los demás y, hoy día, no se conocen otros modelos. En un principio las bacterias endosimbiontes (quizás la situación fuera la de una bacteria con un periparásito, ya que el realmente beneficiado fue el eucarionte) habrían mantenido todo su genoma y sus potencialidades, pero a medida que los organismos evolucionaban hacia grados mayores de complejidad lo fueron perdiendo, adquiriendo las funciones el núcleo mediante evolución de genes nucleares. El desarrollo por parte del núcleo de funciones análogas a las que le proporcionaba el orgánulo permitió que posteriormente se pudieran dar mutaciones y eliminaciones de genes esenciales en el genoma extranuclear, sin que supusieran una merma en la actividad vital del organismo, así se podría explicar la compactación del genoma mitocondrial a lo largo de la evolución.

El operón bacteriano.

El operón es un tipo de organización genómica presente en bacterias, y que incluye todos los mecanismos y sistemas de regulación de la expresión a los niveles de transcripción, estabilización de ARNm y traducción, así como una disposición y estructuración de los genes característica. Se descubrieron mediante estudios de expresión génica de Jacob y Monod. Los sistemas que se dan en eucariotas son distintos (genes individuales interrumpidos) y corresponden con el temario de Genética molecular II, en el próximo curso.

Los operones son el sistema al que recurren los organismos más simples, fisiológicamente hablando, porque permite una respuesta inmediata y coordinada de muchos genes relacionados con rutas necesarias en ese momento, en lugar de recurrir a la homeostasia, que tiene una regulación más complicada. Los eucariotas no regulan las tasas de transcripción como adaptación a corto plazo, sino que actúan a otros niveles posttranscripcionales y posttraduccionales, ya que el núcleo difumina el efecto de inmediatez de este tipo de regulación.

La regulación del operón, así como de todos los sistemas genéticos, puede darse en *cis* o en *trans*, refiriéndose a elementos adyacentes a los genes, o bien mediante productos de expresión de genes distintos. Los primeros incluyen regiones situadas cerca de los genes y que modifican directamente, por proximidad, la tasa de transcripción, pero los segundos son necesariamente proteínas elaboradas a partir de otros genes (distantes), con sus propios elementos de regulación.

El modelo básico. Regulación de la transcripción.

Los operones cuentan en principio con tres niveles de transcripción: basal, reprimido e inducido, que se corresponden con una actividad media, ninguna o mucha, respectivamente. La regulación de la actividad de los operones se realiza mediante un sistema de unión de distintas proteínas a las secuencias que se encuentran en 5' de los genes que se van a transcribir (aguas arriba; elementos en *cis*). Estas proteínas incluyen la ARN polimerasa y los distintos elementos de regulación en *trans*.

Los operones se estructuran como una serie de genes contiguos que reciben una regulación coordinada mediante un solo promotor, situado al comienzo de la transcripción, y la transcripción cursa con una cadena de ARNm policistrónico. Los elementos adyacentes a él constituyen la región operadora o el operador, a los que se unen las diversas proteínas reguladoras. Los estados de transcripción basal, reprimida o inducida se deben siempre a la unión de una proteína reguladora a la región del operador, con lo que se habla de proteínas inductoras o represoras, respectivamente.

Sin embargo, las proteínas reguladoras no deben ser siempre 100% activas, ya que se perdería la característica principal del sistema, que es la versatilidad. Existen pequeñas moléculas, que pueden ser metabolitos internos o sustancias del exterior, que modulan la acción de inductores y represores, activándolos o desactivándolos. Estas moléculas simples se llaman efectores, y combinando todos los sistemas posibles se obtienen estos cuatro modelos de regulación del operón:

- **Estimulación de la represión:** La proteína represora es inactiva y la unión con el efector la activa y promueve su unión al operador, con lo que se reprime la transcripción.
- **Inhibición de la represión:** La proteína represora es activa y la unión con el efector la inactiva y la separa de la región operadora, con lo que la transcripción vuelve al nivel basal.
- **Estimulación de la inducción:** La proteína inductora es inactiva y la unión con el efector la activa y la une al operador, con lo que la transcripción se induce.
- **Inhibición de la inducción:** La proteína inductora es siempre activa y la unión con el efector la inactiva, provocando su separación del operador y devolviendo la transcripción a un nivel basal.

Los operones reales siguen este modelo, pero combinando múltiples elementos de regulación, tanto represores como inductores.

El operón *lac*.

Como ejemplo de regulación combinada de un represor y un inductor está el operón que contiene los genes necesarios para catabolizar la lactosa. El modelo *lac* es el que siguen muchos operones relacionados con las vías catabólicas.

El operón consta de tres genes necesarios para el catabolismo de la lactosa: *lacZ* para la β -galactosidasa, *lacY* para la permeasa y *lacA* para la transacetilasa, además de uno (*lacI*) que codifica para la proteína reguladora en *trans*. En este caso el gen de la proteína reguladora está cercano, pero tiene su promotor propio, por lo que se transcribe y regula de manera independiente al resto, que constituyen el operón propiamente dicho. El resto de elementos reguladores se aprecia en el esquema.

En este caso, y en el de muchos otros operones relacionados con el catabolismo, la proteína reguladora sintetizada constitutivamente por el gen *lacI* es un represor activo tetramérico que se une al operador, impidiendo de este modo la unión de la ARN polimerasa al promotor. La molécula efectora es, en este caso, la propia lactosa (el isómero alolactosa, que es al que se transforma inmediatamente en la célula) que pudiera haber en el medio y hubiera entrado en la célula a través de sistemas de difusión facilitada o selectiva. La unión al represor (a cada subunidad) provoca su inactivación y, por tanto, la transcripción vuelve a niveles basales (inhibición de la represión).

El operón *lac* también es inducible mediante el sistema CAP, que tienen muchos operones del catabolismo. El mecanismo consiste en la unión al sitio correspondiente de la proteína activadora del catabolismo (CAP), que es un inductor estimulado por la unión de AMPc, que dirige su dimerización. Este metabolito interno se produce en respuesta a niveles bajos de glucosa, principalmente, y la unión a CAP (proteína del tipo hélice–bucle–hélice que produce un curvamiento enorme en el ADN) para que se induzca la transcripción de genes relacionados con el catabolismo, es decir, la obtención de energía, responde a la preferencia de los sistemas vivos por la glucosa, antes que por cualquier otro sustrato metabolizable que se encuentre en el medio.

Por la combinación de estos dos elementos de regulación, la transcripción del operón y de los genes para el metabolismo de la lactosa estará totalmente reprimida en ausencia de lactosa, independientemente de la concentración de glucosa; a niveles basales con altas concentraciones de ambos sustratos; e inducida con altos niveles de lactosa, pero bajos de glucosa.

El resultado de esta regulación compleja de la tasa de transcripción de los genes que metabolizan lactosa se resume en que:

- No se transcriben si no hay lactosa.
- Cuando está presente, se prefiere la glucosa porque es energéticamente más favorable.
- Cuando se acaba la glucosa, entonces se induce realmente la síntesis de las enzimas para catabolizar la lactosa.

Atenuación. Operón *trp*.

Un mecanismo más de regulación genética se da a nivel de la traducción. Principalmente se da en operones relacionados con el anabolismo de aminoácidos, y de modo bastante directo. Si en el caso del operón anterior se trataba de calibrar las concentraciones de lactosa y glucosa, es decir, de los sustratos de las rutas metabólicas, la regulación de este operón es contraria, ya que se reprime cuando hay producto en el medio. En condiciones de abundancia de Trp, éste se une a la proteína represora, activándola y evitando la transcripción de los genes que codifican para la ruta que lleva a su síntesis. Sin embargo, el represor de *trp* tiene poca afinidad por el aminoácido (por cualquier aminoácido, en realidad, ya que muchos operones parecidos lo comparten), por lo que puede darse el caso de que con alta concentración de Trp, aunque no suficiente para unirse al represor, se comience la transcripción de los genes estructurales.

Este problema se solventa, en la mayoría de los operones de aminoácidos, mediante la atenuación. Para entender cómo funciona hay que tener en cuenta que en procariotas la traducción está acoplada a la transcripción, que el ARNm policistrónico del operón *trp* codifica para un pequeño péptido al comienzo, antes de los verdaderos genes de la ruta, que es muy abundante en triptófano, y que el ARNm saliente es capaz de plegarse sobre sí mismo en su extremo 5' para formar dos horquillas a partir de sólo tres secuencias, es decir, si las secuencias son 1, 2 y 3, pueden aparear 1 con 2 o bien 2 con 3 para formar una horquilla de cada vez, pero no las dos a la vez.

Con estas premisas supongamos que la cantidad de triptófano es alta, pero que se ha comenzado la transcripción. Inmediatamente un ribosoma se une al extremo 5' y comienza la síntesis de este péptido rico en Trp (*trpL*). Como hay suficiente de este aminoácido, el ribosoma lo incorpora sin problemas y se posiciona rápidamente sobre la secuencia 1, con lo que la horquilla se genera entre las secuencias 2 y 3. Esta horquilla es una señal para la ARN polimerasa (recordar que todo esto está muy junto) para que termine la transcripción (es el terminador), antes de que se haya llegado a los genes estructurales. Si la concentración de Trp es baja, entonces el ribosoma tiene problemas en sintetizar el primer péptido, y se detiene en esa primera secuencia, con lo que da tiempo a que se forme la horquilla entre las secuencias 1 y 2 (antiterminador). Esta estructura deja libre la secuencia 3, que no forma estructura secundaria, con lo que la ARN polimerasa continúa sintetizando el ARNm completo, con lo que se comienza la síntesis de las enzimas de la ruta de síntesis de

Trp.

El gen eucariota.

Esta parte no es más que un pequeño avance de lo que es la estructura genómica en eucariotas, que se estudiará con más detalle en Genética II.

Los eucariotas mantienen organizados sus genes de modo independiente, es decir, cada uno de ellos tiene sus propios elementos de regulación de la transcripción. Además, el gen eucariota típico es el gen interrumpido, que consta de exones e intrones, y que sufren los procesos de maduración nuclear que ya se han descrito en apartados anteriores.

La regulación en *cis* de un gen eucariota suele concentrarse en el extremo 5', cerca del promotor, pero también se han descrito elementos de regulación activos en 3', es decir, al final del gen y también en intrones, sobre todo, dentro del primero. Estos elementos de regulación tienen una organización mucho más compleja que la de los operones bacterianos y no se sabe bien cómo actúan, aunque se han descrito los factores que interaccionan con muchos de ellos e, incluso, qué secuencias protegen. Los elementos reguladores en *trans* son esa multitud de factores de transcripción, sintetizados a partir de otro gen muy distante y, en la mayoría de los casos, en otro cromosoma.

No todo el ARNm se traduce, sino que hay secuencias al principio y al final (UTR) que influyen en la estabilidad del mensajero (en 3') y en el compartimento de destino (en 5'), y que contienen elementos de regulación de la traducción (la interacción de ambas, es decir, la circularización del ARNm influye en la traducibilidad). Además, hay ocasiones en los que los genes no se traducen desde el primer exón, sino desde el segundo, porque el elemento de regulación activo es el que está en el primer intrón.

Los intrones se descubrieron al comparar el ARN nuclear con el ARN citosólico en estudios de inducción de genes. Se comprobó que los primeros eran más largos, más heterogéneos y menos estables que los segundos. Posteriormente se estudió la homología de las secuencias en distintos eucariotas y se comprobó que los exones son los únicos responsables de la homología, mientras que los intrones varía enormemente incluso dentro de una misma especie. Además, la disposición y el orden, así como el tamaño de los exones de genes homólogos en especies distintas siguen un patrón casi idéntico, mientras que el tamaño de los intrones no muestra homología alguna. A la vez que esto último, se comprobó que los genes están interrumpidos, pero no desordenados, es decir, que los exones son correlativos en la secuencia tal y como se disponen en el gen. Una característica de los intrones es que se eliminan, por lo que nunca pueden llegar a traducirse, pero si por alguna razón el sistema fallara, todos los intrones, independientemente del gen en el que estén, tienen señales de terminación de la transcripción en su comienzo y, además, en las tres pautas de lectura posibles. De este modo se interrumpe la producción de una proteína sin sentido o que pudiera causar algún daño al organismo, además de ahorrar gasto de energía inútil.

Los intrones no se eliminan ordenadamente (empezando por el primero y acabando con el último), pero tampoco totalmente al azar, porque los ARNhn parecen seguir rutas preferentes de maduración que no se conocen todavía. En una electroforesis no se observan todos los productos teóricos de una maduración al azar, ni tampoco sólo los que se producirían en una maduración secuencial, sino una especie de mezcla. Probablemente sea la propia estructura del ARNhn la que dirija la maquinaria de corte y empalme preferentemente sobre unos intrones antes que sobre otros.

Los intrones son mucho más largos que los exones, que aproximadamente tienen 150 pb (50 a 60 aa), que es el tamaño medio de un dominio proteico. El primer pensamiento que se ocurre es que las proteínas se han podido ir ensamblando desde unidades estructurales y funcionales básicas, cada una con una actividad diferente, y reorganizadas en el nivel del ADN mediante copias y translocaciones. El tamaño medio de los intrones es de 1 ó 2 Kpb.

El grado de interrupción génica aumenta con la complejidad del organismo (5% en *S. cerevisiae* y 94% en mamíferos). Como media, pero con una gran dispersión cada gen eucariota mide entero unas 15 Kpb, mientras que los ARNm no suelen pasar de las 2 Kb.

Para acabar, se han encontrado casos de empalme transmolecular, es decir, mensajeros de genes distintos se reúnen para formar un único elemento de traducción y, por tanto, una proteína híbrida.

Densidad y distribución de pares CG en una isla.

lacI

lacZ

lacY

lacA

Operador

Promotores

Sitio de unión de la proteína activadora del catabolismo (CAP)

Genes estructurales