INFORMÁTICA

DORCAS J. ORENGO FERRIZ

FUNDAMENTOS DE BIOLOGÍA MOLECULAR

EDITORIAL UOC

Fundamentos de Biología Molecular

Fundamentos de Biología Molecular

Dorcas J. Orengo Ferriz

EDITORIAL UOC

Director de la colección Bioinformática: Daniel Riera Terrén

Diseño de la colección: Editorial UOC

Primera edición digital: abril 2013

Rambla del Poblenou 156, 08018 Barcelona
www.editorialuoc.com

Realización editorial: Anglofort, S.A.

ISBN: 978-84-9029-919-7

Autora

Dorcas J. Orengo Ferriz
Doctora en Biología por la UB (Universitat de Barcelona). Investigadora en el grupo de Genética Molecular Evolutiva del Departament de Genètica de la Facultat de Biologia de la UB. Profesora del curso de postgrado Filogenias y Genealogías de DNA: Reconstrucción y Aplicaciones, del IRBio (Institut de Recerca de la Biodiversitat) de la UB. Consultora de la asignatura Fundamentos de Biología Molecular del Master en Bioinformática y Bioestadística de la UOC (Universitat Oberta de Catalunya).

Índice

Introducción

La **biología molecular** estudia la composición, estructura y función de las moléculas importantes para la vida, manteniendo estrechas relaciones con diversas disciplinas entre las que destacan la bioquímica y la genética.

El desarrollo de la biología molecular ha ido íntimamente ligado al desarrollo de las técnicas moleculares, propiciado principalmente por las aportaciones de físicos y químicos. Las primeras incursiones en este campo datan de los años 30-40 del siglo xx pero no es hasta la década de los 50, con la determinación de la secuencia de las primeras proteínas y la estructura del ADN, que adquiere un importante impulso. En los años 80 se empiezan a obtener secuencias de ADN de forma rutinaria y a finales del siglo xx la introducción de técnicas de secuenciación masiva permite abordar la secuenciación completa de diversos genomas. En la actualidad se sigue trabajando en la innovación y desarrollo de tecnologías que permitan una secuenciación fiable y rápida de genomas enteros.

Cada día se obtiene un mayor volumen de datos de secuencias en un menor tiempo. De hecho, actualmente, uno de los principales problemas con que se enfrenta la biología molecular es la necesidad de almacenar, procesar y gestionar un volumen de datos cada vez mayor y más complejo. Esto obliga a la utilización de herramientas informáticas cada vez más potentes y ha sido decisivo en la aparición y desarrollo de la bioinformática como una disciplina con entidad propia. A su vez, el avance bioinformático deberá permitir a la biología molecular experimentar un nuevo avance cualitativo fijándose nuevas metas.

El estudio de los seres vivos (la biología) es complejo y para hacerlo utilizamos conocimientos aportados por disciplinas tan diversas como la física, la química o la estadística. Además, aunque acotemos los límites de su estudio al nivel molecular, no debemos olvidar que en biología todo está conectado. Las moléculas no funcionan independientemente unas de otras, igual que tampoco lo hacen las células o los organismos. Por ello, aunque dedicaremos los siguientes capítulos al nivel molecular de la biología y en especial a los temas relacionados

con el estudio del ADN, también trataremos algunos aspectos de biología general, que creemos necesarios para obtener una visión global y comprender mejor las problemáticas que intenta resolver la biología molecular.

Objetivos

Este material pretende ofrecer al lector una visión global de los conocimientos básicos sobre biología molecular que necesitará al iniciarse en el estudio de la bioinformática.

Una vez finalizado el libro, el lector deberá ser capaz de:

1. Entender cuales son las características básicas de un organismo vivo.
2. Reconocer las diferencias entre células procariotas y eucariotas.
3. Conocer la química y la estructura de las moléculas esenciales para el funcionamineto de las células.
4. Comprender cómo se almacena la información hereditaria en las células y cómo se transmite a la generación siguiente.
5. Comprender los pasos de la síntesis proteica.
6. Conocer algunas de las estrategias utilizadas para la secuenciación de genomas.
7. Comprender los mecanismos de la evolución y cómo éstos alteran el contenido genético de las poblaciones.
8. Medir la tasa de cambio entre diferentes secuencias relacionadas por un antepasado común.
9. Reconstruir árboles filogenéticos e interpretarlos.
10. Identificar distintas áreas de la biología dónde las herramientas informáticas son especialmente necesarias.

Capítulo I

Biodiversidad: células, organismos y sus relaciones

La fabulosa diversidad de los seres vivos que habitan el planeta Tierra ha fascinado al hombre desde el momento en que se detuvo a observar la naturaleza allá en la noche de los tiempos. A pesar de esta aparente diversidad, cuando profundizamos en el estudio de los seres vivos comprobamos que las diferencias se consiguen a partir de las mismas estructuras y mecanismos básicos. Todos los seres vivos están formados por células que son portadoras de información genética que transmiten fielmente a sus células hijas.

Charles Darwin (1809-1882), que no conocía las leyes de la Genética ni la base molecular de la vida, ya observó las estrechas relaciones que existen entre las distintas especies y, en su libro *El origen de las especies* (1859), las explica como el resultado de descendencia con modificación. Hoy en día sabemos que la composición química de las células de los distintos organismos y las reacciones químicas que tienen lugar en su interior son básicamente las mismas en todas ellas. La información genética de las células de todos los organismos está contenida en el mismo tipo de molécula, el ADN, y el modo cómo se interpreta esta información también es común a todas las células. Esta universalidad de la molécula y el modo de transmisión de la información representan pruebas moleculares de la evolución de los seres vivos a la vez que nos proporcionan los medios para establecer relaciones de parentesco entre las distintas especies.

1.1. Biodiversidad

Se calcula que la vida apareció sobre la Tierra hace unos 3.600 millones de años. Desde entonces las formas vivas se han multiplicado y diversificado tanto que apenas quedan lugares donde no podamos encontrar organismos vivos. Incluso ambientes tan extremos como los desiertos, las profundidades marinas, o las fuentes termales, acogen gran cantidad de organismos especializados. De ahí que apareciera el concepto de **biosfera** refiriéndose al conjunto formado por la

totalidad de los seres vivos y el espacio del planeta en el que habitan, que comprende la zona más baja de la atmósfera, los mares y la zona más superficial del suelo. La existencia de la biosfera es un hecho singular entre los planetas que orbitan alrededor del Sol.

Los seres vivos pueden ser muy diversos: enormes como los elefantes o minúsculos como las diatomeas; longevos como las galápagos, o efímeros como las polillas; en movimiento perpetuo a merced de las aguas del mar, como los organismos del plancton, o firmemente enraizados en el suelo, como los robles; solitarios, como el tigre, o formando sociedades, como las abejas… Esta gran diversidad de seres vivos que habita sobre la Tierra se conoce como **biodiversidad** y es una característica de la vida, resultado de miles de millones de años de evolución biológica. Cada especie ocupa su lugar en la red de relaciones del ecosistema, de modo que su eliminación puede alterar considerablemente al resto del circuito. Por ello, ante la acción agresiva de la actividad humana que pone en peligro la supervivencia de numerosas especies, hay un creciente interés en el estudio de la biodiversidad y en la gestión de su conservación.

El conocimiento de la biodiversidad ha ido aumentando con el tiempo a medida que se exploraban nuevas regiones geográficas y se desarrollaban nuevas técnicas de estudio. Teofrasto en el siglo III AC catalogó unas 500 especies de plantas. Linné, en el siglo XVIII, describió más de 12.000 especies de plantas y animales y propició que sus discípulos participaran en viajes de exploración a lugares remotos en busca de nuevas especies. Sólo tras el perfeccionamiento del microscopio se descubrieron los microorganismos que, de hecho, constituyen la mayor parte de la biodiversidad del planeta. Por otro lado, el desarrollo de las técnicas de la biología molecular ha facilitado discernir entre especies crípticas que morfológicamente no podemos diferenciar.

Actualmente, el número de especies distintas descritas supera el millón y medio, pero se calcula que el número de especies actuales debe ser muy superior, entre 3 y 100 millones según los autores. Cada año, se describen nuevas especies de microorganismos, pero también de plantas y animales, incluyendo especies de algunos grupos que, como los mamíferos, han sido desde siempre especialmente bien estudiados (Ceballos & Ehrlich, 2009). El censo total de especies del planeta se ve dificultado por dos hechos: la dispersión de los datos en museos, colecciones y herbarios de todo el mundo, y la existencia de sinonimias. En la actualidad, el GBIF (*Global Biodiversity Information Facility*; http://www.gbif.org/), que es un proyecto internacional, intenta unificar toda la información disponible, digitaliza datos de colecciones de todo el mundo y promueve la in-

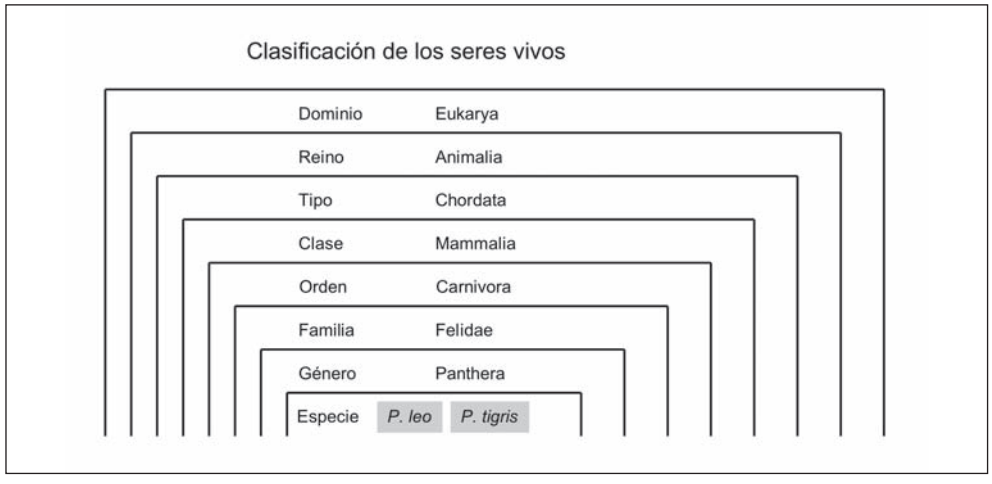
clusión digital de las nuevas observaciones, así como la creación de *software* especifico que permita manejar toda la información necesaria en la gestión de la biodiversidad.

1.2. Especies

El estudio de cualquier grupo de entidades conlleva implícita la necesidad de su ordenación. El estudio de los seres vivos no es una excepción y desde antiguo han sido clasificados. El esquema de clasificación de los seres vivos, o taxonomía, que utilizamos actualmente, se basa en el que propuso Carl von Linné (1707-1778) en su obra *Systema Naturae* (1735), con algunas modificaciones. La taxonomía tiene por objeto agrupar a los seres vivos según presenten semejanzas o diferencias entre sí, clasificándolos en categorías jerárquicas de modo que las categorías de nivel más bajo quedan englobadas en otras superiores (Figura 1.1).

Las semejanzas entre organismos pueden deberse a **analogías** o a **homologías**. Las estructuras análogas son el resultado de adaptaciones convergentes a unas necesidades funcionales concretas que, como en el caso del ala de las aves y de los insectos, tienen origen muy diverso. Por su parte, las estructuras homólogas son parecidas a causa de un origen común, como serían las extremidades

Figura 1.1. Grupos jerárquicos de clasificación de los seres vivos. Cada categoría está incluida en otra superior. La columna de la derecha muestra la clasificación del león y el tigre, que ilustra que, en realidad, cada grupo superior contiene diversos inferiores.



anteriores del topo, el murciélago, la ballena y el hombre, que aunque adaptadas a distinta función todas han evolucionado a partir de la misma estructura ósea ancestral. Para que la taxonomía tenga sentido biológico debe basarse en la filogenia, las relaciones evolutivas entre los organismos. Por tanto, debemos considerar sólo las homologías.

En la taxonomía biológica, el primer nivel de agrupamiento corresponde a la especie. Las especies más cercanas se agrupan en el mismo género. Los géneros se agrupan en familias, las familias en órdenes, los órdenes en clases, las clases en tipos (también conocidos como *Phylum* o divisiones) y éstos en reinos (Figura 1.1). Los nombres científicos de las especies siguen la nomenclatura binaria iniciada por Linné. Cada especie recibe su nombre científico en latín, que se escribe en cursiva y está compuesto por el nombre del género al que pertenece (en mayúscula) y su nombre específico (en minúscula). Por ejemplo, el nombre científico del león es *Panthera leo* y el del tigre *Panthera tigris* ambos del género Panthera.

El primer nivel de clasificación (la especie) es utilizado a diario en nuestra vida cotidiana. Todos somos capaces de reconocer que un gato, un canario, una hormiga y un pino, pertenecen a especies diferentes. Pero la clasificación en especies, que puede parecer simple, en ocasiones resulta muy compleja. De hecho, el concepto de especie es uno de los mas escurridizos en biología y tema de discusión frecuente entre biólogos (Hey, 2001).

> «Tampoco discutiré aquí las varias definiciones que se han dado del término especie. Ninguna definición ha satisfecho a todos los naturalistas; sin embargo, todo naturalista sabe vagamente lo que quiere decir cuando habla de una especie.» Darwin (1859) *El origen de las especies*.

Históricamente se utilizaba un **concepto morfológico** (o **tipológico**) de especie: a partir de un individuo tipo, todos los individuos que compartían con él una serie de características morfológicas, se consideraban pertenecientes a la misma especie. Pero este concepto es bastante subjetivo y la inclusión o no de un individuo en una especie determinada dependerá de qué caracteres consideremos para discriminar entre especies. Darwin, en su libro *El origen de las especies*, incide en esta problemática, mostrando la dificultad en decidir cuándo dos individuos pertenecen a especies diferentes o cuándo sus diferencias responden tan sólo a la variabilidad presente en las poblaciones y deben, por tanto, considerarse variedades de una misma especie. Precisamente esta dificultad le sugiere

que las diferencias entre individuos de una misma especie pueden ir acentuándose de generación en generación hasta originar nuevas especies.

A partir de Darwin, se formulan distintas definiciones de especie que incorporan la idea de identificar grupos que comparten procesos evolutivos. Así, Mayr (1942) formuló el **concepto biológico de especie** que muchos genetistas evolutivos utilizan normalmente. Según Mayr *una especie es un conjunto de poblaciones naturales cuyos individuos son capaces de reproducirse entre ellos produciendo descendencia fértil y que, a su vez, están aislados reproductivamente de otros grupos similares*. De todos modos, el concepto biológico de especie no puede ser aplicado universalmente. En ocasiones puede resultarnos bastante difícil realizar los cruzamientos oportunos, pero en otras es totalmente imposible, como en las especies de reproducción asexual y en las especies fósiles.

El concepto de especie en bacterias es especialmente difícil. No podemos utilizar el concepto biológico por reproducirse asexualmente y, por otro lado, cuentan con muy pocos caracteres morfológicos variables. Además, la gran velocidad con que se reproducen hace que los cambios genéticos en las poblaciones de bacterias sean rápidos, por lo que no es aconsejable distinguir especies basándose en sólo unos pocos caracteres. Por ello, en general se acepta que una especie bacteriana es un conjunto de clones que provienen del mismo linaje, que muestran gran semejanza fenotípica y que difieren de otros conjuntos por un gran número de caracteres independientes. Puesto que su morfología es poco variable, debe recurrirse a la observación de características bioquímicas, fisiológicas y ecológicas.

Cualquiera de los conceptos de especie aquí expuestos intentan explicar, con mayor o menor fortuna, una realidad: los individuos de una especie comparten los mismos genes. En la era postgenómica podemos decir que los individuos de una misma especie comparten su genoma y que las ligeras diferencias que presentan es lo que les confiere su personalidad individual. Siguiendo esta idea se están desarrollando nuevas herramientas para la identificación rápida de especies como es el código de barras de ADN *(DNA barcoding)*. El *Consortium for the Barcode of Life* (CBOL) es un consorcio internacional fundado en 2004 para determinar el código de barras de la mayoría de las especies, actuales. El código de barras de una especie se obtiene secuenciando un fragmento de ADN que esté muy conservado dentro de la especie pero que sea altamente variable entre especies. En animales se ha elegido como estándar un fragmento de 648 nucleótidos de un gen mitocondrial, el *citocromo oxidasa c I* (COI). La plataforma bioinformática BOLD (*Barcode of Life Data System*; www.barcodinglife.org)

recopila todos los códigos de barras disponibles permitiendo diagnosticar rápidamente a qué especie pertenece un individuo o decidir si se trata de una especie aún no catalogada (Ratnasingham y Hebert, 2007).

1.3. La célula: unidad funcional de todo ser vivo

El desarrollo del microscopio óptico, en el siglo XVII, supuso una gran revolución para la biología. Como ya hemos mencionado, permitió descubrir los microorganismos, pero más importante aun permitió descubrir que todos los seres vivos están formados por células.

La **célula** es la unidad estructural y funcional de todos los seres vivos. Encontramos seres unicelulares (formados por una única célula), que no podemos ver sin ayuda del microscopio, y organismos pluricelulares (formados por numerosas células). Las células de los organismos pluricelulares, que pueden ser desde unas pocas hasta millones, han perdido la capacidad de vivir independientemente y en cambio se organizan y reparten las funciones para crear un único organismo armónico.

Las distintas células de un organismo pluricelular pueden diferir mucho en forma y función pero todas ellas, al igual que los organismos unicelulares, se han originado por división de otras células. Toda célula está delimitada por una membrana que la separa del exterior pero que permite el intercambio de nutrientes, energía, información y materiales de desecho entre el interior y el exterior de la célula. En su interior tienen lugar procesos metabólicos dirigidos por las proteínas que se sintetizan a partir de la información contenida en el material genético, el ADN.

> En 1838 Schleiden y Schwann formulan la teoría celular que afirma que todos los tejidos animales y vegetales están formados por agregados celulares. En 1858 Virchow añade que toda célula proviene de otra célula.

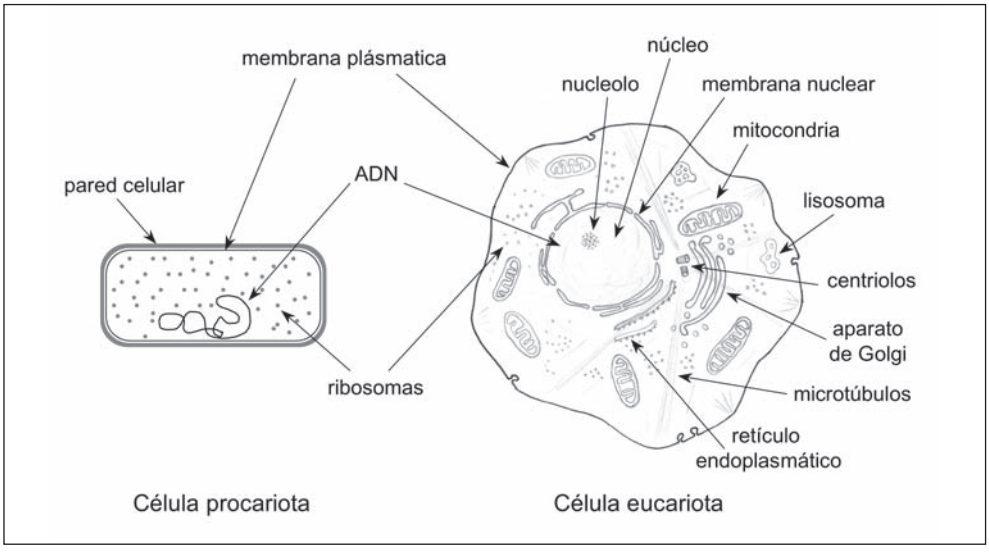
En los organismos pluricelulares encontramos distintos tipos de células según la función para la que se han especializado por diferenciación celular. En el cuerpo humano se han contabilizado más de 200 tipos celulares distintos. Las células nerviosas, musculares o hepáticas de un mismo individuo tienen forma y función muy distinta, pero todas ellas provienen de una primera célula huevo

o cigoto y, por tanto, tienen la misma información genética encerrada en su ADN. Sin embargo, durante el desarrollo se produce la diferenciación celular por la que los distintos linajes celulares expresan un conjunto distinto de genes, especializándose en funciones distintas. Unas pocas células se mantienen indiferenciadas y pueden seguir dividiéndose para reemplazar a las células que envejecen y mueren. Estas células se conocen como células madre.

1.3.1. Procariotas vs. eucariotas

Podemos clasificar a todos los seres vivos en 2 grupos, **procariotas** y **eucariotas**, según la estructura de sus células (Figura 1.2). **Procariotas** y **eucariotas** presentan diversas diferencias importantes que se resumen en la Tabla 1.1, pero la diferencia fundamental (y que da nombre a los dos grupos) es la ausencia o presencia de una membrana nuclear que separa el material genético del citoplasma.

Figura 1.2. Comparación de las estructuras básicas de las células procariotas y eucariotas.



La presencia, en los eucariotas, de orgánulos definidos por membranas que contienen su propio material genético (mitocondrias y cloroplastos) llevó a Lyn Margulis a formular la hipótesis del origen de la célula eucariota como una relación endosimbiótica entre procariotas. Con los años, y a medida que disponía de nuevos

Tabla 1.1. Diferencias entre **procariotas** y **eucariotas**

Procariotas	Eucariotas
Ausencia de membrana que separe el material hereditario del resto de la célula	Presencia de una membrana que delimita el núcleo donde se encuentra el material genético separándolo del resto de la célula o citoplasma
Generalmente un único cromosoma circular	Diversos cromosomas lineales
Genoma muy compacto	Genoma con regiones intergénicas e intrones
Ausencia de orgánulos definidos por membranas	Contiene orgánulos definidos por membranas que cumplen funciones esenciales, tales como las mitocondrias (producción de energía), el retículo endoplásmatico (síntesis de proteínas en los ribosomas adheridos), el aparato de Golgi o dictiosoma (síntesis y almacenamiento de glicoproteínas), y en vegetales, los cloroplastos (fotosíntesis)

datos, Margulis ha ido completando su teoría (Margulis et al. 2006). Según esta teoría, una Archaebacteria semejante a *Thermoplasma acidophilum* se asociaría a una Eubacteria similar a *Spirochaeta sp.* formando un «consorcio» con relaciones de sintrofia (una especie vive de lo que produce la otra especie), similar al actual Thiodendron. Con el tiempo, individuos de este consorcio se fusionarían formando un organismo indivisible con una estructura flagelar unida al núcleo por un conector nuclear. Este ser quimérico evolucionó hacia una célula nucleada estable o LECA (del inglés, *Last Eukaryotic Common Ancestor*). Posteriores eventos endosimbiónticos con otras Eubacteria darían origen a las mitocondrias (a partir de α-proteobacterias) y posteriormente a los cloroplastos (a partir de cianobacterias).

1.4. Árbol de la vida

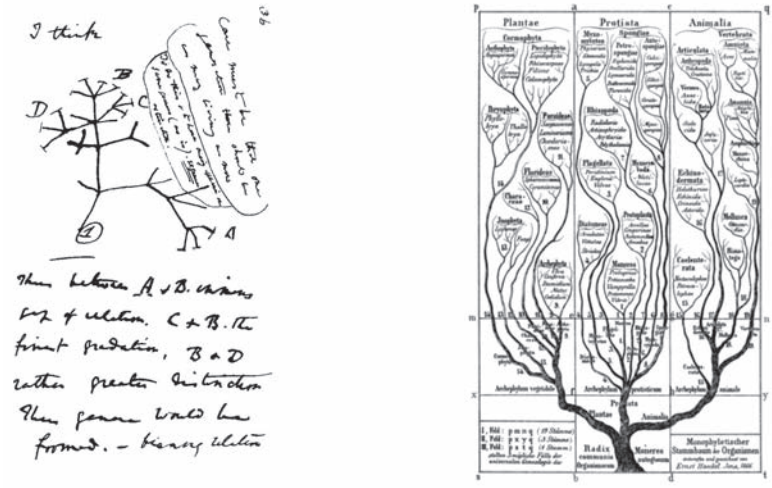
> «Las afinidades de todos los seres de la misma clase se han representado a veces por un gran árbol. Creo que este símil expresa mucho la verdad. Los vástagos verdes y en ciernes pueden representar las especies existentes, y las ramas producidas durante años anteriores pueden representar la larga sucesión de especies extinguidas [...] Así como los brotes dan origen, por crecimiento, a nuevos brotes, y éstos, si son vigorosos, se ramifican y sobrepujan por todos los lados a muchas ramas más débiles, así también a mi parecer, ha ocurrido en el gran árbol de la vida, que con sus ramas muertas y rotas llena la corteza terrestre y cubre su superficie con sus hermosas ramificaciones, siempre en constante bifurcación.» Darwin (1859) en *El origen de las especies*.

A pesar de la gran diversidad que observamos entre los seres vivos, también descubrimos suficientes afinidades que nos indican que todos ellos están relacionados por descendencia. Los árboles filogenéticos permiten representar gráficamente estas dos caras de una misma moneda: la diversidad biológica y la afinidad entre organismos por ascendencia común. En un libro de notas de 1837, Darwin ya esquematiza las relaciones entre organismos en forma arbórea (Figura 1.3) y la única ilustración que incluye en *El origen de las especies*, es un árbol mostrando la divergencia entre especies a lo largo de las generaciones.

Cuando en un único árbol incluimos a todos los seres vivos, hablamos del árbol filogenético universal o **árbol de la vida**. La raíz del árbol de la vida representa al ancestro común a todos los organismos vivos actuales y se conoce como LUCA (del inglés: *Last Universal Common Ancestor*).

Podemos considerar que el primero en dibujar un árbol de la vida fue Haeckel en 1866 en su obra *Generelle Morphologie der Organismen*, donde clasificaba a todos los seres vivos en 3 grandes Reinos: plantas, animales y protistas (Figura 1.3). A medida que se ha ido conociendo mejor la biodiversidad, el árbol

Figura 1.3. Primeras representaciones arbóreas de las relaciones entre los seres vivos. Árbol esbozado por Darwin (1837) y árbol de la vida de Haeckel (1866)



Imágenes obtenidas en Wikipedia.

de la vida ha ido creciendo y la posición de las ramas se ha ido modificando. Hasta no hace muchos años, los biólogos dibujaban 5 grandes ramas principales en el árbol de la vida, correspondientes a 5 Reinos: Monera, Protoctistas, Fungi, Plantae y Animalia. Éstos se agrupaban, según su estructura celular, en 2 grandes grupos: Procariotas y Eucariotas.

El desarrollo de la biología molecular y, en especial, la secuenciación del ADN ha permitido descifrar las verdaderas relaciones filogenéticas en muchos casos dudosos. El ADN es la molécula que encierra la información genética de todos los seres vivos y en ella quedan registrados fielmente los cambios que se han ido acumulando entre las especies. Por tanto, es la molécula idónea para realizar estudios filogenéticos y el único recurso para comparar seres tan alejados como puedan ser una bacteria, un roble y un ratón.

En base a la biología molecular, se ha comprobado que los procariotas (=Monera) son un grupo filogenéticamente heterogéneo que incluye las bacterias y las arqueobacterias. Las arqueobacterias se describieron inicialmente como bacterias que habitaban lugares inhóspitos, tales como salinas, fuentes termales, fuentes sulfurosas o profundidades oceánicas, pero también se las encuentra en ambientes habituales para otros organismos. Aunque su aspecto recuerda a las bacterias, a nivel molecular encontramos tanto características bacterianas como eucariotas. Así, mientras que los enzimas metabólicos y del sistema de transporte son típicamente bacterianos, toda la maquinaria implicada en la replicación del ADN y en la síntesis proteica (transcripción, procesamiento de ARN y traducción) es similar a la de los eucariotas. Por todo ello, la clasificación más aceptada actualmente divide a los seres vivos en Archaea, Bacteria y Eukarya (Tabla 1.2) y estos grupos reciben la categoría taxonómica de dominio o imperio que es superior a la de reino (Woese et al. 1990).

Tabla 1.2. Clasificación de los seres vivos en Dominios

Dominio	Reino	Ejemplos
Bacteria		eubacterias
Archaea		arqueobacterias
Eukarya	Protoctista	mixomicetos, protozoos
	Fungi	levaduras, mohos, basidiomicetes
	Plantae	musgos, confieras, plantas con flores
	Animalia	esponjas, corales, gusanos, moluscos, insectos, vertebrados

El origen endosimbióntico de la célula eucariota y de sus orgánulos, junto con la gran capacidad que los procariotas tienen de transferir parte de su genoma a otros organismos que no son sus descendientes (transferencia horizontal) hacen difícil el esclarecimiento de la continuidad entre las ramas más antiguas por lo que algunos autores representen al árbol de la vida con algunas de sus ramas basales fusionadas o entrelazadas.

Capítulo II

Composición química de la célula

Toda la materia que conocemos está formada por un centenar de elementos químicos distintos. Los organismos vivos están formados por elementos que también podemos encontrar en las rocas, la atmósfera y los océanos de nuestro planeta que, a su vez, son los mismos que encontramos en otros planetas y estrellas del Universo. Lo peculiar de la materia orgánica es la proporción y modo con que los elementos químicos se organizan en moléculas. Ello hace que la química de la materia viva se distinga fundamentalmente por tres características:

1. Se basa en compuestos del carbono. El hecho de que el átomo de carbono admita 4 enlaces con otros tantos elementos, o grupos de elementos, permite que puedan generarse multitud de moléculas distintas.
2. El agua es un compuesto primordial de la vida, siendo el principal componente de la célula. La mayoría de las reacciones químicas de la célula tienen lugar en medio acuoso.
3. Un sistema vivo encierra mucha más complejidad que cualquier otro sistema. La célula contiene multitud de moléculas diversas que constantemente están interactuando en procesos de transformación por los que generan moléculas más complejas pero también por los que se degradan en otras más sencillas.

La materia viva está sujeta a las leyes universales de la física y de la química pero presenta suficientes características especiales para que su química merezca estudiarse como un capítulo aparte: la **química orgánica**.

Podemos clasificar las moléculas que encontramos en cualquier célula en 3 grandes grupos: compuestos inorgánicos, pequeñas moléculas orgánicas y macromoléculas.

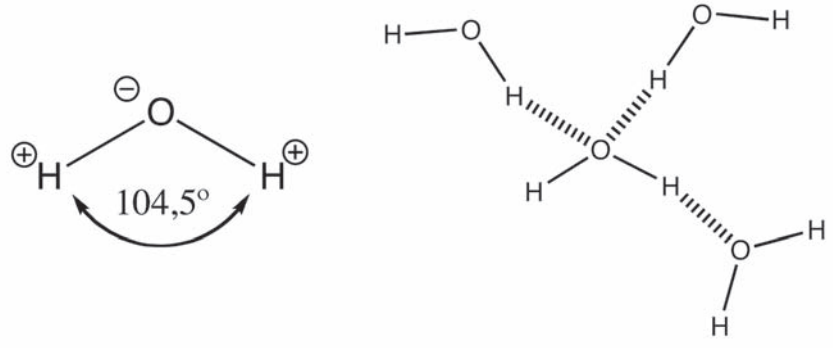
2.1. Compuestos inorgánicos

Los compuestos inorgánicos son moléculas compartidas con la materia inerte y que no pertenecen a la química del carbono o química orgánica. Básicamente se trata del **agua** y las **sales minerales**.

El **agua** es el componente mas abundante de los seres vivos, representando por término medio el 70 % del peso de la célula, aunque su proporción puede variar considerablemente dependiendo del organismo, tejido y edad.

En la molécula de agua (H_2O) los dos átomos de hidrógeno están unidos al átomo de oxígeno por enlaces covalentes de una forma polar. El núcleo de oxígeno atrae a los electrones con más fuerza que el núcleo de hidrógeno, con lo que se crean cargas positivas y negativas locales. Esto hace que la molécula tenga una constante dieléctrica elevada permitiendo que forme nuevas uniones con otras moléculas mediante puentes de hidrógeno (Figura 2.1). Esta estructura peculiar hace que el agua presente una serie de propiedades físico-químicas muy características que la convierten en indispensable para la vida.

Figura 2.1. Estructura molecular del agua



Las moléculas de H_2O se mantienen unidas por puentes de hidrogeno, confiriendo al agua importantes propiedades fisico-químicas

Las **sales minerales** pueden estar presentes de forma insoluble (básicamente formando parte de las estructuras esqueléticas) pero más comúnmente las encontramos disueltas (disociadas en iones). Algunas de las sales que encontramos disociadas en iones son: cloruros, fosfatos, sulfatos, bicarbonatos, carbonatos y nitratos (como aniones) de Na, K, Ca, Mg, Cu y Fe (como cationes).

2.2. Pequeñas moléculas orgánicas

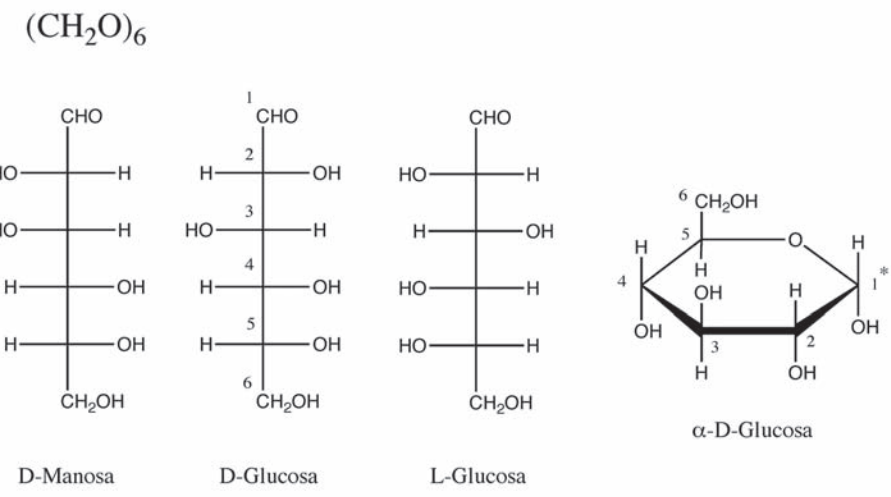
En la célula, encontramos multitud de pequeñas moléculas orgánicas que en ocasiones tienen función propia pero que más comúnmente constituyen las piezas básicas que, al unirse, originarán otras moléculas orgánicas mucho más complejas. Podemos agrupar estas pequeñas moléculas en 4 familias: **azúcares, ácidos grasos, aminoácidos y nucleótidos**.

Los **azúcares** más sencillos son los monosacáridos, que tienen como fórmula general $(CH_2O)_n$, donde *n* es un número entre 3 y 7. Los átomos de carbono se unen entre sí, formando una cadena, y cada uno de ellos está unido a un grupo hidroxilo (–OH), excepto uno que forma parte de un grupo carbonilo (>C=O). Hay que destacar que en cualquier monosacárido existe, como mínimo, un carbono asimétrico, es decir, un carbono que está unido a 4 grupos distintos.

Con esta estructura básica y el mismo número de átomos de cada tipo se consigue una gran variedad de moléculas distintas (isómeros). Para empezar, dependiendo de la posición del grupo carbonilo tenemos isómeros funcionales: un aldehído (carbonilo en un extremo de la cadena) o una cetona (carbonilo en una posición intermedia en la cadena de carbonos). Por otro lado, si una molécula tiene varios carbonos asimétricos, cambiando la disposición espacial relativa de los grupos se obtienen otros isómeros que también difieren en sus propiedades físico-químicas. Es el caso de la glucosa y la manosa que sólo difieren en la orientación de los grupos unidos al carbono 2 (Figura 2.2). Finalmente, para cada monosacárido podemos encontrar 2 enantiómeros, isómeros ópticos que son imágenes especulares uno del otro. Los enantiómeros tienen las mismas propiedades físicas y químicas excepto que, en disolución, desplazan la luz polarizada hacia planos distintos. Por convenio, aquellas moléculas cuyo último carbono asimétrico sitúa el grupo OH hacia la derecha se conoce como forma D y las que lo sitúan a la izquierda, como forma L. En la naturaleza, la gran mayoría de azúcares corresponden a la serie D (Figura 2.2).

Los monosacáridos que tienen más de 4 carbonos no son estables en forma de cadena lineal y, generalmente, forman estructuras en anillo por la unión de su grupo aldehído (o cetona) con un grupo alcohol dentro de la misma molécula. Además, el carbono que tiene el grupo carbonilo, puede reaccionar con cualquier grupo hidroxilo de otro monosacárido, formando un disacárido con la liberación de una molécula de H_2O.

Figura 2.2. Algunos isómeros de monosacáridos de 6 átomos de carbono



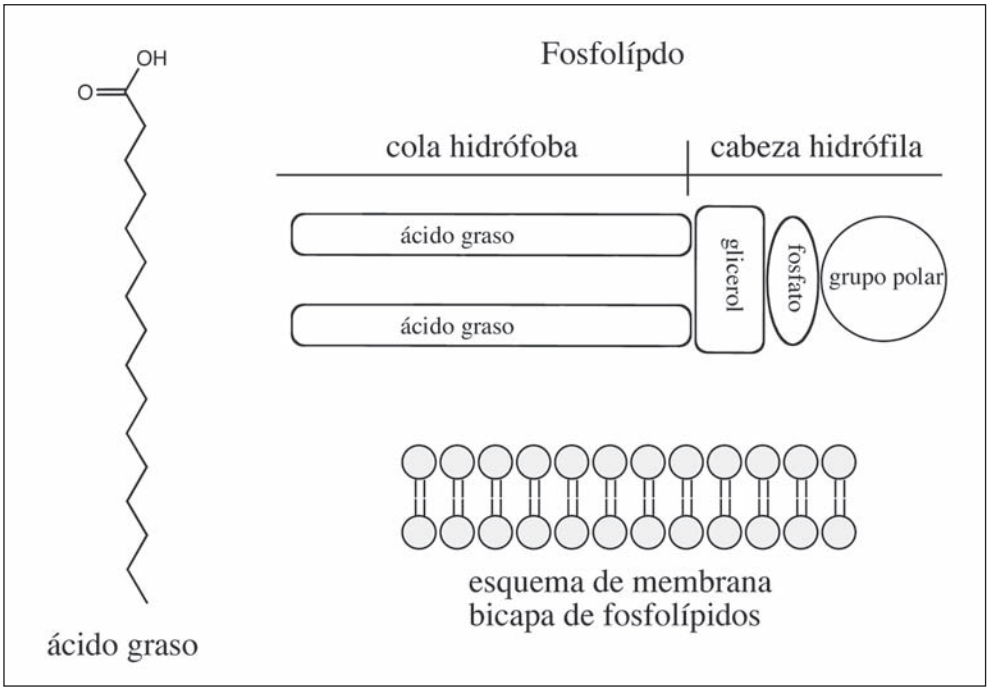
D-Glucosa y D-Manosa sólo difieren en la posición relativa de su grupo -OH sobre el carbono número 2. Cuando la molécula adopta estructura en anillo aparece un nuevo carbono asimétrico indicado por el asterisco.

Algunos monosacáridos especialmente importantes son: la glucosa, la ribosa y la desoxirribosa. La glucosa es una hexosa (6 átomos de C) que es el combustible de la célula. Ribosa y desoxirribosa son dos pentosas (5 átomos de C) que forman parte de los nucleótidos.

Los **ácidos grasos** son moléculas de 16 a 18 átomos de carbono formadas por una larga cadena hidrocarbonada, hidrófoba, unida a un grupo carboxilo (–COOH), hidrófilo, muy reactivo. Generalmente, los ácidos grasos están unidos covalentemente a otras moléculas por su cabeza hidrófila. Los fosfolípidos son especialmente importantes. Éstos se forman por la unión de 2 ácidos grasos a un glicerol que, además, está unido a un grupo fosfato que a su vez está unido a un grupo polar que, por lo general contiene nitrógeno (Figura 2.3). Los fosfolípidos son los componentes principales de las membranas celulares donde se disponen en una doble capa en las que sus colas hidrófobas se enfrentan.

Los **aminoácidos** son moléculas bastante heterogéneas pero que todas ellas tienen un carbono característico, el carbono α, que está unido a 4 grupos distintos (Figura 2.4). Estos 4 grupos son:

Figura 2.3. Fosfolípido

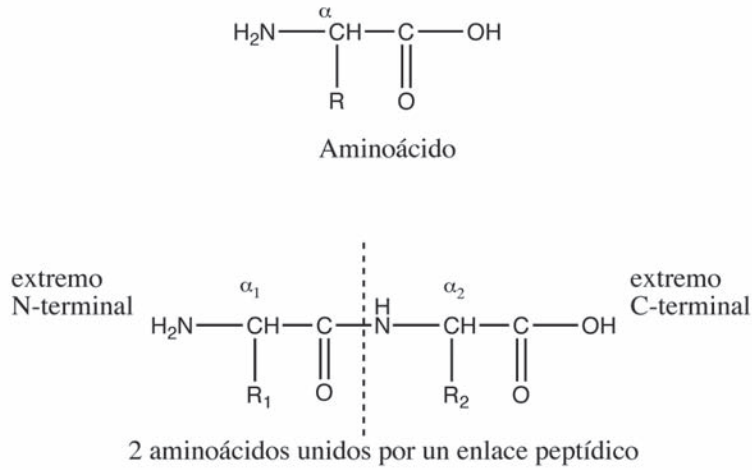


- un grupo amino ($-NH_2$), que tiene carácter básico
- un grupo carboxilo (–COOH), que tiene carácter ácido
- un hidrógeno
- una cadena lateral o grupo R (reactivo)

Dos aminoácidos pueden unirse mediante un enlace peptídico entre el grupo carboxilo de uno y el grupo amino del siguiente, con la liberación de una molécula de H_2O. De este modo, por la unión progresiva de aminoácidos se puede obtener una cadena polipeptídica que tiene polaridad estructural (Figura 2.4), con un grupo amino en uno de sus extremos (extremo N-terminal) y un grupo carboxilo en el otro extremo (extremo C-terminal).

Se conocen unos 50 aminoácidos pero las proteínas están formadas por combinaciones de tan sólo 20 aminoácidos distintos (Tabla 2.1). Unas pocas proteínas contienen un aminoácido número 21, la selenocisteina. Los distintos aminoácidos se diferencian por la estructura de su cadena lateral. En la glicina, la cadena lateral es un hidrógeno, pero en el resto de aminoácidos, R es una cadena más o menos compleja de carbonos, con lo que el carbono α es un carbono asi-

Figura 2.4. Estructura general de los aminoácidos y del enlace peptídico



métrico. Por tanto, en los aminoácidos, al igual que en los monosacáridos, hay isómeros ópticos. En este caso, todos los aminoácidos que forman parte de las proteínas son de la serie L.

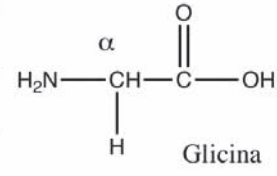
La identidad y propiedades características de cada aminoácido se las confiere su cadena lateral. Esta cadena puede estar formada por un simple hidrógeno, como en la glicina, o por una compleja estructura cíclica, como en el triptófano, pasando por cadenas más o menos complejas de carbonos. Las cadenas laterales de la metionina y la cisteina contienen átomos de azufre que en el caso de la cisteina pueden formar enlaces disulfuro entre las cadenas laterales de 2 cisteinas. La estructura de la prolina es muy característica, puesto que su grupo amino forma parte de un anillo (es una amina secundaria).

Unos pocos aminoácidos presentan carácter básico y otros pocos carácter ácido, de modo que, en disolución, forman iones con carga eléctrica. Otro grupo de aminoácidos son polares sin carga, comportándose como hidrofílicos. El resto son apolares hidrófobos.

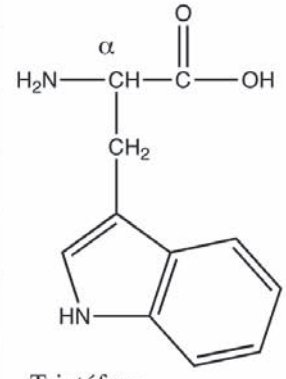
Los **nucleótidos** son moléculas compuestas de otras 3 más simples: un azúcar (ribosa o desoxirribosa), una base nitrogenada y ácido fosfórico (Figura 2.5). Los nucleótidos constituyen las piezas fundamentales de los ácidos nucleicos, pero algunos también tienen funciones específicas.

Tabla 2.1. Símbolos de los aminoácidos y carácter de sus cadenas laterales

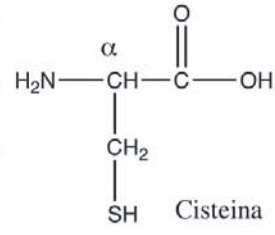
Aminoácidos			Carácter	Ejemplos
Ácido aspártico	Asp	D	Ácido	
Ácido glutámico	Glu	E	Ácido	
Alanina	Ala	A	No polar	
Arginina	Arg	R	Básico	
Asparagina	Asn	N	Polar sin carga	
Cisteina	Cys	C	No polar	
Fenilalanina	Phe	F	No polar	
Glicina	Gly	G	No polar	
Glutamina	Gln	Q	Polar sin carga	
Histidina	His	H	Básico	
Isoleucina	Ile	I	No polar	
Leucina	Leu	L	No polar	
Lisina	Lys	K	Básico	
Metionina	Met	M	No polar	
Prolina	Pro	P	No polar	
Serina	Ser	S	Polar sin carga	
Tirosina	Tyr	Y	Polar sin carga	
Treonina	Thr	T	Polar sin carga	
Triptófano	Trp	W	No polar	
Valina	Val	V	No polar	



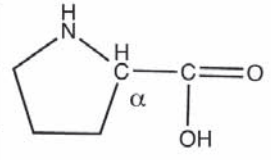
Glicina



Triptófano



Cisteina

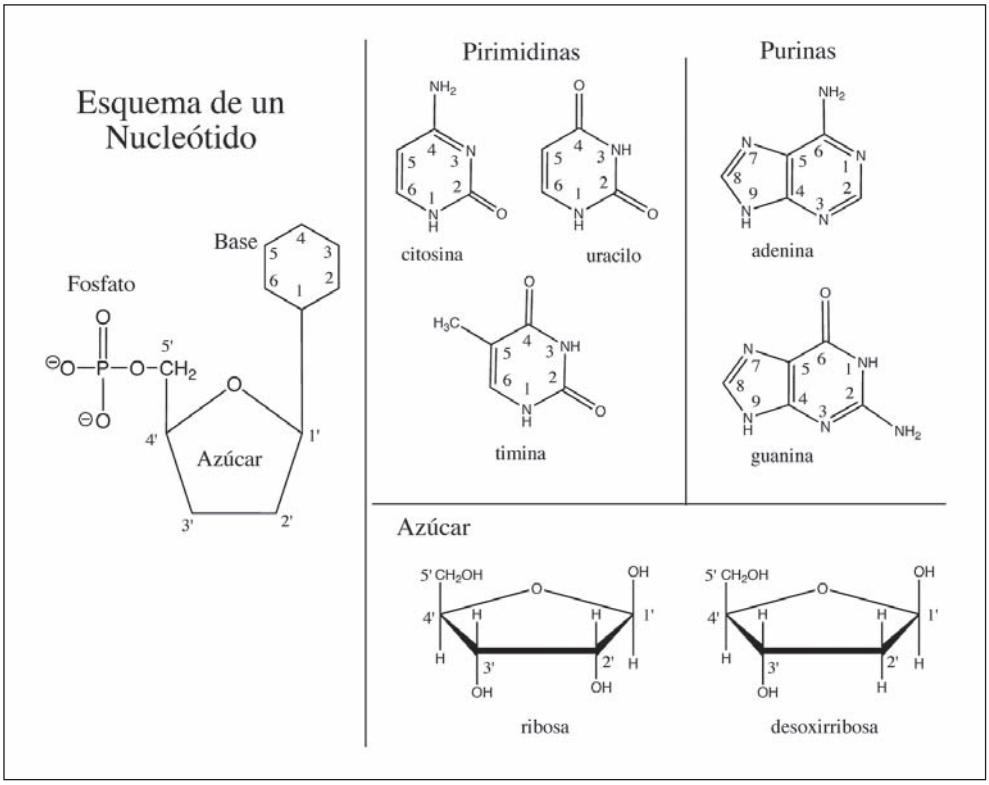


Prolina

Generalmente, los nucleótidos se nombran por la base nitrogenada que contienen (y, a menudo, directamente por su inicial): adenina (A), citosina (C), guanina (G), timina (T) y uracilo (U). No obstante, no debe olvidarse nunca que también pueden diferir por el azúcar. Así, al hablar, por ejemplo, de una adenina, debemos tener claro si nos estamos refiriendo a un ribonucleótido o a un desoxirribonucleótido.

Las bases nitrogenadas pueden ser purinas o pirimidinas (Figura 2.5). Las pirimidinas (C, T, U) son compuestos derivados de un anillo pirimidínico hexagonal. Las purinas (A, G) presentan un segundo anillo pentagonal unido al anillo hexagonal.

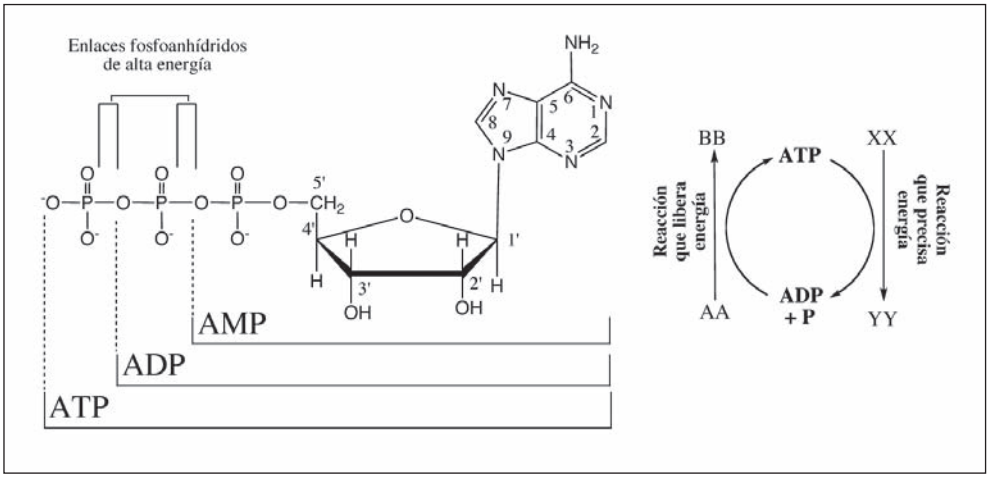
Figura 2.5. Esquema de un nucleótido y sus componentes



Los átomos que forman las estructuras cíclicas de las bases nitrogenadas se numeran para podernos referir fácilmente a cada uno de ellos. Los átomos de carbono del azúcar también se numeran y, para diferenciarlos de los de las bases, se les añade un símbolo de prima (').

El ATP (adenosina trifosfato) es un nucleótido de una importancia vital para la célula, donde actúa como transportador de energía. Sus 3 fosfatos están unidos en serie por enlaces muy ricos en energía. La rotura de uno de estos enlaces lo transforma en ADP (adenosina difosfato), liberando energía que puede ser utilizada en diferentes reacciones. Por su parte, el ADP puede absorber la energía liberada en otras reacciones uniéndose a un tercer fosfato y pasando de nuevo a ATP (Figura 2.6).

Figura 2.6. El ATP transporta energía química de unas reacciones a otras



2.3. Macromoléculas

Las macromoléculas que encontramos en la célula son polímeros de elevado peso molecular formados a partir de la unión de numerosas unidades (o monómeros) de pequeñas moléculas orgánicas. Los polisacáridos son polímeros de monosacáridos que pueden tener función de reserva (almidón, glucógeno) o estructural (celulosa, quitina). Los lípidos son derivados de los ácidos grasos y, aunque no se consideran como macromoléculas, comparten algunas características de éstas. Así, los fosfolípidos se asocian formando estructuras membranosas. Tanto polisacáridos como lípidos no tienen carácter específico, siendo comunes entre distintas especies de seres vivos.

Hay 2 tipos de macromoléculas que son de una importancia capital para la vida y su diversidad: **proteínas** y **ácidos nucleicos** (ADN y ARN). Las caracterís-

ticas físicas y metabólicas de un ser vivo dependen de las proteínas que presenta. Las proteínas tienen una gran especificidad, con diferencias que pueden ser grandes entre especies e incluso entre individuos. Los mecanismos de herencia genética aseguran la transmisión fiel, de una célula a sus células hijas, de la información necesaria para fabricar sus proteínas. Esta información se transmite por medio del ADN.

2.3.1. Proteínas

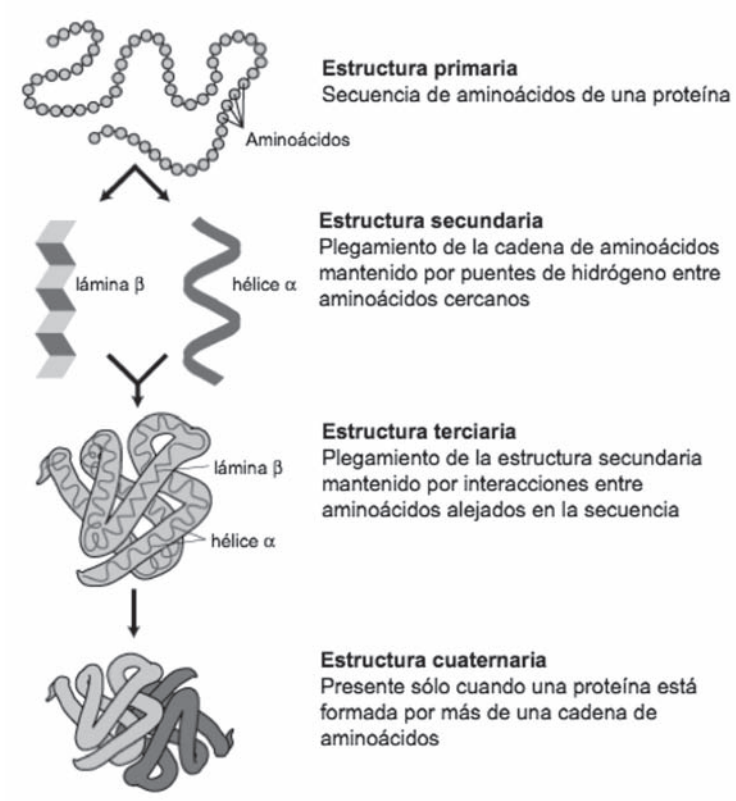
Las proteínas representan el 50% del peso seco de la célula y desempeñan funciones muy diversas: estructural, enzimática, transporte, reguladora, defensa... La gran diversidad de función se corresponde con la gran diversidad de formas proteicas presentes en la célula. Ello se debe a que las proteínas son largas cadenas en las que se combinan 20 aminoácidos distintos, pudiéndose obtener, por tanto, infinidad de moléculas distintas.

La identidad de una proteína viene definida por el número, tipo y orden en que se unen los aminoácidos que la componen, que se conoce como secuencia de aminoácidos de la proteína o **estructura primaria** de la proteína. La organización espacial de esta secuencia de aminoácidos puede determinar una estructura proteica muy compleja en la que podemos distinguir hasta 3 niveles superiores de estructura: secundaria, terciaria y, en algunos casos, cuaternaria (Figura 2.7).

En las proteínas, los distintos enlaces peptídicos (amino-carboxilo) se disponen linealmente formando un esqueleto polipeptídico y los grupos funcionales de las cadenas laterales quedan como ramificaciones. Los grupos N–H y C=O de dos enlaces peptídicos distintos pueden establecer uniones mediante puentes de hidrógeno haciendo que la proteína adopte ciertas disposiciones espaciales más estables. Esto hace que la cadena adquiera un cierto plegamiento que, principalmente, puede ser en forma de hélice α o de lámina β.

En una misma proteína pueden coexistir regiones con plegamientos de diversos tipos. A la disposición de estos plegamientos en la molécula se le conoce como **estructura secundaria** de la proteína. La estructura en anillo del grupo amino de la prolina, hace que el enlace peptídico en el que está implicada adopte un giro fijo y marcado en la estructura secundaria de la proteína. Por ello, la prolina a menudo aparece como disruptor de estructuras secundarias regulares y también suele ser el primer aminoácido de una hélice α.

Figura 2.7. Niveles estructurales de las proteínas



Modificado a partir de una figura de Wikipedia.

Las distintas regiones de la estructura secundaria, se disponen de una cierta forma en las 3 dimensiones del espacio y se estabilizan gracias a interacciones, más fuertes que los puentes de hidrogeno, entre residuos de aminoácidos que pueden estar muy alejados en la estructura primaria. Así aparece la **estructura terciaria** de la proteína.

En ocasiones varias cadenas polipeptídicas se asocian entre ellas para formar agrupaciones funcionales más grandes. Cada cadena polipeptídica será una subunidad de la proteína. Cada subunidad puede ser, o no, igual al resto de subunidades. Las proteínas que tienen interacciones de este tipo se dice que presentan **estructura cuaternaria**. Un ejemplo clásico de proteína con estructura cuaternaria es la hemoglobina, que está formada por 4 subunidades: 2 cadenas de α-globina y 2 de β-globina.

Por otro lado, en las proteínas grandes podemos encontrar distintas unidades estructurales locales que conocemos como **dominios proteicos**. Los dominios proteicos de una proteína se pliegan independientemente y generalmente están asociados a distintas funciones de la proteína.

El plegamiento de una proteína en el espacio depende de diversos factores. La estructura primaria es fundamental, puesto que es la que permitirá la formación de interacciones entre los distintos aminoácidos. En ocasiones, el sustituir un único aminoácido supone un cambio tan importante en la estructura de la proteína que puede llegar a alterar totalmente su función. Pero la mayoría de proteínas, además, necesitan unas condiciones de temperatura y pH determinadas dentro de unos límites muy precisos. Cuando se sobrepasan estos límites la proteína se desnaturaliza, sufre alteraciones que afectan a sus estructuras secundaria y terciaria.

Normalmente, cada proteína se pliega en una única conformación estable, a la cual llegaría espontáneamente, pero que suele acelerarse por la intervención de unas proteínas especiales, las chaperonas. Por otro lado, la conformación de una proteína, a menudo cambia ligeramente cuando interactúa con otras moléculas presentes en la célula. El ejemplo más ilustrativo es el de las enzimas alostéricas. Estas enzimas presentan además de la zona de unión al substrato (o centro activo) otras zonas donde se les unen factores reguladores (activadores o inhibidores). Según la proteína esté unida o no a estos reguladores, la enzima adopta una conformación espacial distinta permitiendo o no el acceso del substrato a su zona de unión o centro activo. Así pues, la conformación espacial de las proteínas depende de diversos factores y, en ocasiones, puede tratarse incluso de un estado bastante dinámico.

La determinación de la estructura tridimensional de las proteínas supone un trabajo experimental laborioso que requiere, además, el empleo de técnicas costosas. En 1960, se obtuvo la primera estructura tridimensional de una proteína (la mioglobina del cachalote) por medio de un estudio de cristalografía de rayos X. Esta técnica se aplica a una muestra de proteína pura cristalizada. El proceso de obtención de la proteína cristalizada es complejo y largo. Cada proteína requiere unas condiciones especiales de cristalización que deben determinarse por métodos de ensayo-error que conllevan una gran inversión de tiempo. Más recientemente, se están aplicando otras técnicas como la resonancia magnética nuclear (NMR, *nuclear magnetic resonance*) que permiten analizar la estructura de la proteína en disolución concentrada, sin necesidad de cristalizar. La limita-

ción de la técnica NMR está en que su resolución sólo es óptima para proteínas relativamente pequeñas.

Por ello, a menudo, se realiza el estudio de la estructura tridimensional de las proteínas por dominios.

Se podría pensar que, partiendo de una secuencia de aminoácidos, siempre podremos predecir su estructura terciaria correcta y que la única limitación es el tiempo computacional a invertir. Sin embargo, existen otros factores que pueden determinar que el plegamiento real de la proteína no sea el más estable a partir de su secuencia primaria. Algunas proteínas experimentan modificaciones después de su síntesis que pueden alterar las propiedades de algunos de sus aminoácidos. Por otro lado, la configuración biológicamente activa de algunas proteínas puede que no sea la más estable termodinámicamente. Por todo ello, la predicción de la estructura espacial de las proteínas sigue siendo uno de los grandes retos de la bioinformática y de la química teórica.

Actualmente, las bases de datos recogen la estructura tridimensional, determinada por rayos X o NMR, de más de 10.000 proteínas. Esto permite que cada día sea más fiable la predicción de la estructura tridimensional de una proteína problema (*query*) mediante modelado por homología. Estos métodos se basan en la idea de que, proteínas con secuencias de aminoácidos similares también tendrán estructura similar.

2.3.2. ADN

Los experimentos de Avery, MacLeod y McCarty (1944) demostraron que el ácido desoxirribonucleico (ADN) es la molécula que contiene toda la información genética de la célula. Desde entonces, muchos investigadores abordaron su estudio en busca de la clave que explicara cómo una molécula aparentemente tan simple podía ser la responsable de la transmisión de toda la información necesaria para construir la gran diversidad que observamos en los seres vivos. Finalmente, en 1953, James Watson y Francis Crick, utilizando los resultados experimentales de otros autores, dedujeron la estructura molecular del ADN que, como veremos más adelante, explica el mecanismo de la herencia. Su interpretación se basó en el conocimiento de 3 tipos de información:

1. El ADN es un polímero formado por bloques de fosfato, desoxirribosa y 4 bases nitrogenadas (adenina, citosina, guanina y timina).

2. Las observaciones de Edwin Chargaff sobre la relación cuantitativa entre las bases nitrogenadas del ADN, conocidas como reglas de Chargaff. Según estas reglas:
   - la cantidad de pirimidinas (T+C) siempre es igual a la de purinas (A+G).
   - la cantidad de A es igual a la de T y la cantidad de C es igual a la de G.
3. Las imágenes inéditas, obtenidas por Rosalind Franklin, de la difracción de rayos X del ADN.

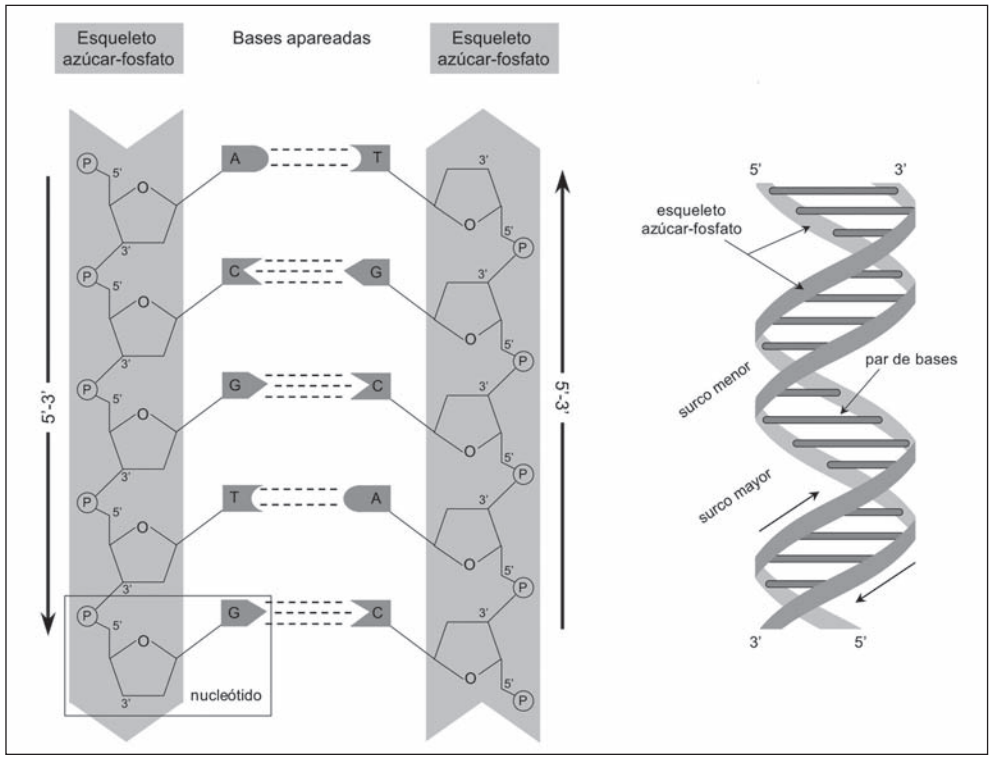
Encajando estas 3 piezas, Watson y Crick dedujeron que la estructura del ADN se compone de 2 cadenas helicoidales curvadas sobre el mismo eje. Cada cadena está formada por un esqueleto de enlaces fosfodiester que unen el átomo de carbono 3' de una desoxirribosa con el 5' de la siguiente (por lo que la cadena tiene polaridad 5' –> 3') y las dos cadenas discurren en sentidos opuestos (son antiparalelas). Las bases nitrogenadas, que están unidas al carbono 1' de la desoxirribosa, se disponen hacia el interior de la doble hélice. Cada una de las bases de una cadena establece puentes de hidrógeno con una base de la cadena complementaria consiguiendo que las dos cadenas se mantengan unidas. Los pares de bases sólo pueden ser A-T o C-G, de modo que, por ejemplo, una A en una cadena determina una T en la otra (Figura 2.8).

La sucesión de bases dentro de una cadena no parece tener ningún tipo de restricción. Sin embargo, debido a las reglas de complementación, si conocemos la secuencia de bases en una de las cadenas, la secuencia de bases de la otra queda automáticamente determinada.

El descubrimiento de la estructura del ADN supuso uno de los grandes avances de la biología del siglo xx. Por ello, en 1962, Watson y Crick recibieron, junto a Maurice Wilkins, el Premio Nobel. Aunque en su articulo de apenas una página Watson y Crick (1953) sólo describían la estructura del ADN, ya apuntaban que dicha estructura permitiría comprender la función del ADN como material genético.

«Queremos sugerir una estructura para la sal del ácido desoxirribonucleico (ADN). Esta estructura tiene características novedosas que son de considerable interés biológico [...] No nos ha pasado inadvertido que el apareamiento específico que hemos postulado sugiere inmediatamente un posible mecanismo de copia para el material genético.» Watson y Crick (1953).

Figura 2.8. Estructura en doble hélice del ADN



Las dos cadenas de polinucleótidos discurren en forma antiparalela (en sentido opuesto) y son complementarias. Las dos cadenas se unen por las bases de sus nucleótidos, que quedan en el interior de una doble hélice. La unión entre Adenina y Timina se realiza por 2 puentes de hidrógeno, mientras que Citosina y Guanina se unen mediante 3 puentes de hidrógeno.

La estructura de doble hélice le permite al ADN cumplir con las propiedades exigidas al material genético de la célula:

- Contiene la información para la síntesis de una gran diversidad de proteínas, codificada en su secuencia de bases nitrogenadas.
- Se transmite fielmente en cada división celular. La complementariedad de las bases de las dos cadenas permite que partiendo de una cadena se obtenga una copia fiel de la otra.
- Tiene capacidad de cambio, lo que ha originado toda la variabilidad que observamos. Durante el proceso de copia, pueden introducirse errores (mutaciones) que, desde ese momento, serán transmitidos en las subsiguientes divisiones celulares.

El ADN de la célula eucariota se encuentra en el núcleo, donde es el principal componente de los cromosomas. Cada cromosoma está formado por una molécula de ADN asociada a diversas proteínas que lo mantienen empaquetado. Los procariotas, generalmente, presentan un único cromosoma circular inmerso en el citosol.

Cada cadena de ADN puede estar formada por millones de nucleótidos. La longitud del ADN suele medirse en pares de bases (pb), puesto que, generalmente, el ADN está formado por dos cadenas complementarias, que sólo se diferencian por sus bases nitrogenadas. Podemos representar una molécula de ADN por la secuencia de bases de una de sus cadenas.

Por ejemplo:

```
5' .. ATGGGCCAAGAGGATCAGGAGCTATTAATTCGCGGAGGCAGC .. 3' 42pb
```

Por convenio, las moléculas de ADN se escriben, en sentido 5' –> 3'. En caso contrario debe especificarse claramente. Con esta información podemos reconstruir toda la molécula de doble cadena.

Siguiendo el ejemplo anterior:

```
5' .. ATGGGCCAAGAGGATCAGGAGCTATTAATTCGCGGAGGCAGC .. 3'  42pb
3' .. TACCCGGTTCTCCTAGTCCTCGATAATTAAGCGCCTCCGTCG .. 5'
```

Esto permite que en las bases de datos se guarde tan sólo la secuencia de una de las dos cadenas, sin por ello perder ninguna información.

2.3.3. ARN

El ácido ribonucleico (ARN) es otro polímero de nucleótidos. Los nucleótidos que forman el ARN (ribonucleótidos) son distintos de los que forman el ADN (desoxirribonucleótidos). Por un lado, el azúcar que contienen es la ribosa, en lugar de la desoxirribosa. El carbono 2' de la ribosa está unido a un grupo –OH que la desoxirribosa no presenta (Figura 2.5). Por otro lado, el ARN no contiene timina (T) y en su lugar se encuentra uracilo (U).

Existen diversos tipos de ARN entre los que destacan el ARN mensajero (ARNm), el ARN ribosómico (ARNr) y el ARN de transferencia (ARNt). El ARNm

es un intermediario entre el ADN y las proteínas. Los ARNt son los responsables de transportar el aminoácido correcto a la cadena de proteína que se está sintetizando y facilitar su incorporación a ésta. Los ARNr forman parte del ribosoma, un complejo macromolecular que guía la formación de la cadena proteica a partir de la información del ARNm. Algunos ARN se asocian a proteínas para formar complejos ribonucleoproteicos de función enzimática. Éste es el caso del ARN nuclear pequeño (snRNA, del inglés *small nuclear RiboNucleic Acid*) que forma parte del empalmosoma, de gran importancia para la maduración del ARNm. Además, en los últimos años, se ha descrito una nueva categoría de ARN, los ARN interferentes, de gran importancia para la regulación génica.

Como veremos más adelante, el ARN se sintetiza como copia monocatenaria (de una sola cadena) de regiones de ADN. De modo similar al ADN, los ribonucleótidos del ARN se unen entre sí por enlaces fosfodiester entre el carbono 3' de un nucleótido y el 5' del siguiente. Las bases del ARN siguen las mismas leyes de complementariedad que las del ADN, sustituyendo T por U pero, normalmente, el ARN se encuentra en forma de cadena sencilla. La molécula de ARN puede plegarse libremente sobre sí misma permitiendo que bases más o menos lejanas de la misma molécula se apareen. Este apareamiento de bases intramolecular es responsable de la estructura tridimensional del ARN, como por ejemplo la estructura en hoja de trébol de los ARNt (ver Figura 4.6).

Capítulo III

Transmisión de la Información Genética

Una de las características de la célula, como unidad fundamental de los seres vivos, es su capacidad de reproducirse, generar nuevas células semejantes a ella. La morfología y fisiología de la célula depende de la información genética que contiene su ADN. Por ello, es fundamental que la transmisión de esta información genética sea correcta entre una célula y sus células hijas. Para que esta transmisión sea correcta, primero debe realizarse una copia fiel de la información genética o replicación del ADN para luego repartirse correctamente durante la división celular entre las dos células hijas resultantes.

Tanto la replicación del ADN como la división celular son ligeramente distintas en procariotas y eucariotas. Debe recordarse que, en general, los procariotas estructuran su ADN en un único cromosoma circular inmerso en el citosol, mientras que en los eucariotas su ADN se distribuye en múltiples cromosomas lineales incluidos en el núcleo.

En los procariotas la duplicación de su cromosoma circular comienza en un lugar especial llamado origen de replicación. Una vez duplicado el cromosoma, las dos copias se separan hacia extremos opuestos y la pared celular crece separándolos y formando dos nuevas células. En condiciones óptimas, algunas bacterias pueden completar un ciclo de división cada 20 minutos. Esto nos permite, en el laboratorio, obtener colonias o suspensiones de células con millones de bacterias idénticas (o clones) en muy pocas horas pero también supone un problema cuando se quiere acabar con alguna infección.

El ciclo celular en los eucariotas, incluye los distintos estados por los que atraviesa la célula empezando tras la división celular que la ha originado y acabando en otra división celular que originará dos células hijas nuevas. Básicamente podemos dividir el ciclo celular en: Interfase y Mitosis. Durante la interfase, la célula crece y se desarrolla preparándose para entrar en una nueva división celular o mitosis.

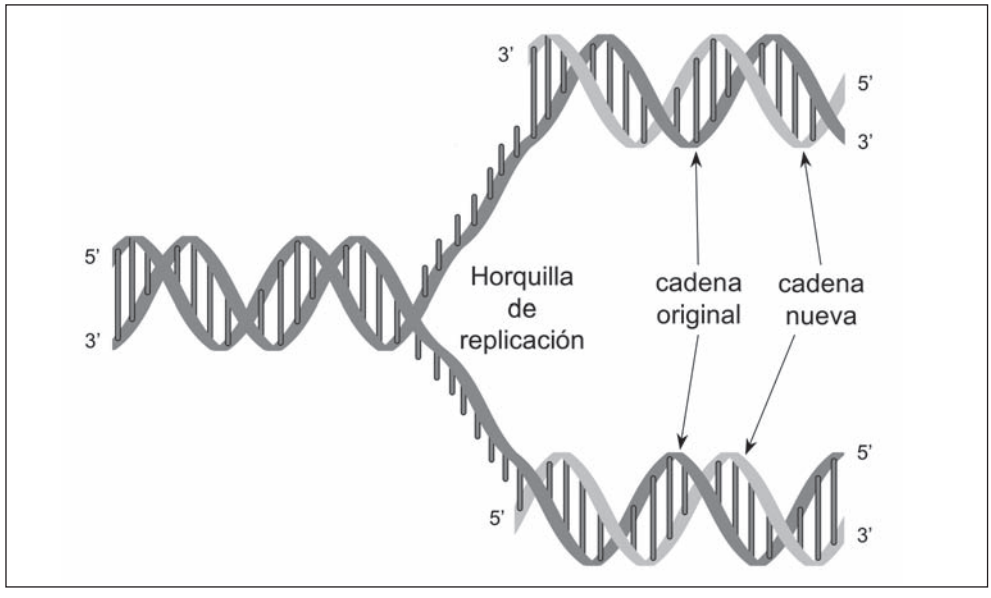
La transmisión de la información genética se inicia con el proceso de replicación del ADN durante la interfase y culminará con la división celular.

3.1. Replicación del ADN

Conocemos como replicación del ADN al proceso por el cual una molécula de ADN se duplica. La replicación del ADN se enfrenta a dos necesidades: la fidelidad de copia y la rapidez[1]. Para ello, la célula despliega una compleja maquinaria enzimática de alta precisión.

Tal como demostraron los experimentos de Meselson y Stahl (1958), la replicación del ADN es **semiconservativa**. Cada una de las dos cadenas de ADN sirve como molde para sintetizar su cadena complementaria. De este modo, se obtienen dos moléculas de ADN, cada una de las cuales estará formada por una cadena original y otra sintetizada de nuevo (Figura 3.1).

Figura 3.1. Replicación semiconservativa del ADN



1. *Escherichia coli* replica su ADN a una velocidad de 1000 pb por segundo introduciendo menos de un error cada mil millones de nucleótidos.

Para que el ADN pueda duplicarse, debe desenrollarse y separar sus dos cadenas complementarias como si fuera una cremallera que se abre. A medida que las dos cadenas se separan, las bases nucleotídicas de cada cadena quedan accesibles y sirven de molde a los nucleótidos que se van incorporando a la cadena sintetizada de nuevo. La zona de la doble hélice que separa sus dos cadenas y se va copiando, se conoce como horquilla de replicación.

El ADN puede estar estructurado de diversas formas: en cromosomas circulares como en las bacterias; en pequeñas moléculas circulares como en los plásmidos de algunas bacterias; o en cromosomas lineales como en los eucariotas. Cada uno de estos tipos de estructuración utiliza un modelo de replicación distinto: theta, círculo rodante o lineal.

La replicación del cromosoma circular de las bacterias se inicia por un único lugar especial llamado origen de replicación (oriC). A partir del oriC, la doble hélice se abre y aparece una burbuja de replicación, con lo que el cromosoma adopta una forma que recuerda la letra griega theta (θ). Por ello, a este modelo de replicación se le conoce como modelo theta de replicación. A partir del origen de replicación, la burbuja de replicación puede avanzar en sólo uno o en los dos sentidos. Las horquillas de replicación avanzan hasta que los dos extremos de la burbuja de replicación se encuentran.

El ADN circular de plásmidos y de algunos virus utiliza el modelo de replicación del círculo rodante. En este modelo, el proceso se inicia con el corte de una de las dos cadenas quedando dos extremos libres, 3' OH y 5' fosfato. La síntesis se inicia en el extremo 3' de la cadena escindida usando la cadena circular como molde. La incorporación de nucleótidos avanza a medida que el extremo 5' de la cadena escindida se va separando de la cadena circular como si fuera un hilo que se desenrolla de un ovillo. Al llegar al final del círculo se han obtenido una molécula circular de cadena doble y una molécula lineal de cadena sencilla. La molécula lineal puede ahora circularizarse y servir de molde para la síntesis de su cadena complementaria.

En los eucariotas, el ADN de cada cromosoma lineal está formado por millones de pares de bases. No obstante, la replicación se realiza en un tiempo récord si lo comparamos con los tiempos empleados en la replicación del ADN bacteriano. Ello se debe a que la replicación se realiza simultáneamente a partir de múltiples

orígenes de replicación en cada cromosoma. Además, en cada origen de replicación se generan dos horquillas que avanzan en ambos sentidos hasta alcanzar a las horquillas de replicación adyacentes.

Tabla 3.1. Diferencias entre los distintos modelos de replicación

Modelo	Theta	Círculo Rodante	Lineal Eucarionte
ADN molde	Circular	Circular	Lineal
Corte de la cadena de nucleótidos	NO	SI	NO
Origen de replicación	1	1	muchos
Direccionalidad de la replicación	Unidireccional o Bidireccional	Unidireccional	Bidireccional
Producto	2 moléculas circulares de doble cadena	1 molécula circular de doble cadena + 1 molécula lineal de cadena sencilla	2 moléculas lineales de cadena doble
Fragmentos de Okazaki	1000 – 2000 nt	----	100 – 200 nt

3.1.1. Detalles de la replicación

En la replicación intervienen numerosas enzimas con función diversa. La acción eficaz de todas ellas se consigue porque forman parte de un gran complejo nucleoproteico, el replisoma, que coordina su actividad en la horquilla de replicación.

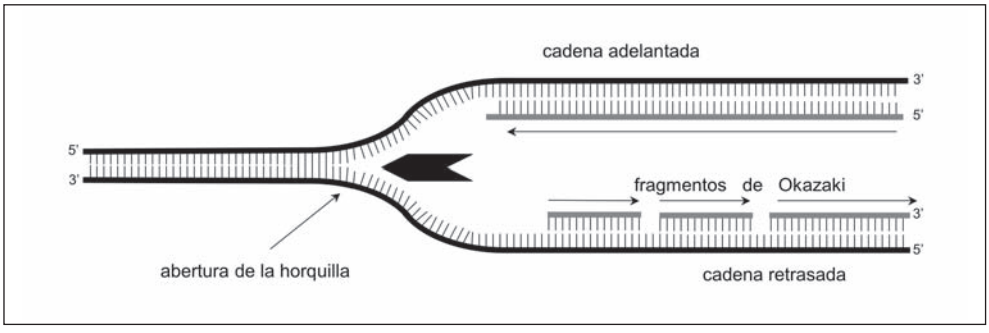
Las primeras enzimas del replisoma que actúan son las helicasas y las topoisomerasas que desenrollan la doble hélice, rompen los puentes de hidrógeno que unen las dos cadenas complementarias y evitan que se formen superenrollamientos de la hélice en zonas adyacentes.

Una vez que las bases de cada cadena quedan expuestas en la horquilla de replicación, las ADN polimerasas pueden utilizarlas como molde para incorporar los nucleótidos complementarios en la cadena que se está sintetizando. Dis-

tintas ADN polimerasas intervienen en momentos distintos de la replicación, pero todas ellas utilizan las formas trifosfato de los desoxirribonucleótidos (dATP, dCTP, dGTP y dTTP) para hacer crecer la nueva cadena en sentido 5' –> 3', y cada nucleótido que se incorpora libera dos fosfatos al unirse al nucleótido anterior. No obstante, las ADN polimerasas sólo pueden añadir nucleótidos a una cadena ya existente, no tienen la capacidad de iniciar la síntesis de una cadena. La síntesis se inicia a partir de unos **cebadores** o *primers* de ARN que son fragmentos de 8 a 12 ribonucleótidos sintetizados por una primasa que copia la cadena de ADN.

Debido a que las dos cadenas de la doble hélice son antiparalelas y que la incorporación de nucleótidos sólo se realiza en sentido 5' –> 3', a medida que va avanzando la horquilla de replicación, sólo una de las dos nuevas cadenas puede sintetizarse de forma continua[2]. La cadena que avanza en sentido contrario al que avanza la horquilla de replicación, se sintetiza de forma discontinua en fragmentos cortos que se conocen como **fragmentos de Okazaki**. La cadena que se sintetiza de forma continua se conoce como **cadena líder** o **adelantada** y la cadena que se sintetiza de modo discontinuo es la **cadena retrasada** (Figura 3.2).

Figura 3.2. Avance de la síntesis en la horquilla de replicación



Además de su actividad principal como polimerasa 5' –> 3', la ADN polimerasa, también tiene actividad exonucleasa 3' –> 5'. Esta actividad exonucleasa le permite corregir, sobre la marcha, la mayoría de errores que haya podido introducir durante la replicación. Cuando la ADN polimerasa detecta la inclusión reciente de un nucleótido erróneo, cambia su actividad a exonucleasa y

2. En YouTube existen numerosos videos explicativos de la replicación del ADN. Un buen ejemplo se encuentra en: http://www.youtube.com/watch?v=-EGKrYdQEHQ

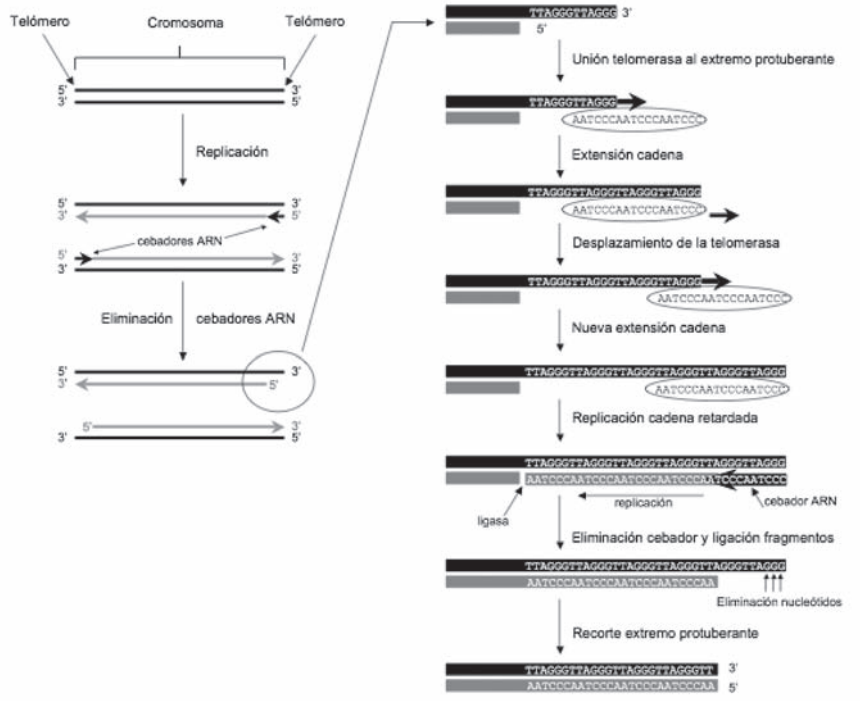
elimina los últimos nucleótidos para, posteriormente, retomar su actividad polimerasa.

Una ADN polimerasa especial tiene, además, función exonucleasa 5' –> 3', que utiliza para eliminar los nucleótidos de los cebadores de ARN. A continuación, esta misma enzima utiliza su función polimerasa para rellenar el hueco dejado, añadiendo desoxirribonucleótidos a partir del fragmento anterior. Finalmente interviene una ADN ligasa que une el extremo 3' de un fragmento de Okazaki con el extremo 5' del fragmento siguiente, creando un enlace fosfodiéster sin necesidad de añadir un nuevo nucleótido. Esta ligasa también se encarga de unir los fragmentos sintetizados a partir de distintas horquillas de replicación.

La replicación del ADN de los eucariotas presenta algunas dificultades adicionales. Por un lado, el ADN se enrolla alrededor de las histonas, formando los nucleosomas que mantienen el cromosoma compactado. Durante la replicación, la maquinaria enzimática del replisoma debe, primero, desensamblar el nucleosoma para, más tarde, ensamblarlo de nuevo en las moléculas hijas. Por otro lado, el ADN lineal presenta la complicación de la replicación de los extremos (Figura 3.3). La cadena líder puede extenderse hasta el último nucleótido de la cadena que está copiando, pero en la cadena retrasada, después de eliminar el cebador de ARN, queda un fragmento de ADN de cadena sencilla que puede degradarse fácilmente. Para evitar que en cada nueva replicación se pierdan unos pocos nucleótidos, los extremos de los cromosomas lineales, los telómeros, han desarrollado características especiales. Los **telómeros** están formados por múltiples repeticiones cortas en tándem de una secuencia característica. La **telomerasa** es una ribonucleoptroteína formada por una proteína con función polimerasa y una secuencia de ARN de 15-22 nucleótidos complementarios a la cadena protuberante. Esta secuencia de ARN se alinea en parte a las últimas bases de la cadena larga y el resto de bases actúan como molde para que la parte proteica de la telomerasa alargue aun más la cadena protuberante. A continuación la telomerasa se desplaza a una posición más adelantada de la cadena, propiciando la incorporación de nuevas repeticiones. Tras algunos saltos de la telomerasa, una primasa incorpora un nuevo cebador y se completa el fragmento del modo habitual. La eliminación de este cebador vuelve a dejar un fragmento de cadena sencilla pero ahora ya no es tan problemático puesto que la telomerasa ha alargado el cromosoma.

La telomerasa está presente en los organismos unicelulares, en las células germinales, en las células de los primeros estadios embrionarios y en algunos linajes celulares somáticos proliferativos. En el resto de células somáticas, la actividad telomerasa es prácticamente nula, por lo que estas células sólo pueden

Figura 3.3. Replicación de los telómeros



dividirse un número limitado de veces. En cada división, los cromosomas experimentan acortamientos y sobrepasado un valor umbral, los cromosomas son inestables y pueden experimentar alteraciones importantes que conducen a la muerte celular. Este acortamiento de los telómeros parece estar relacionado con los procesos de envejecimiento.

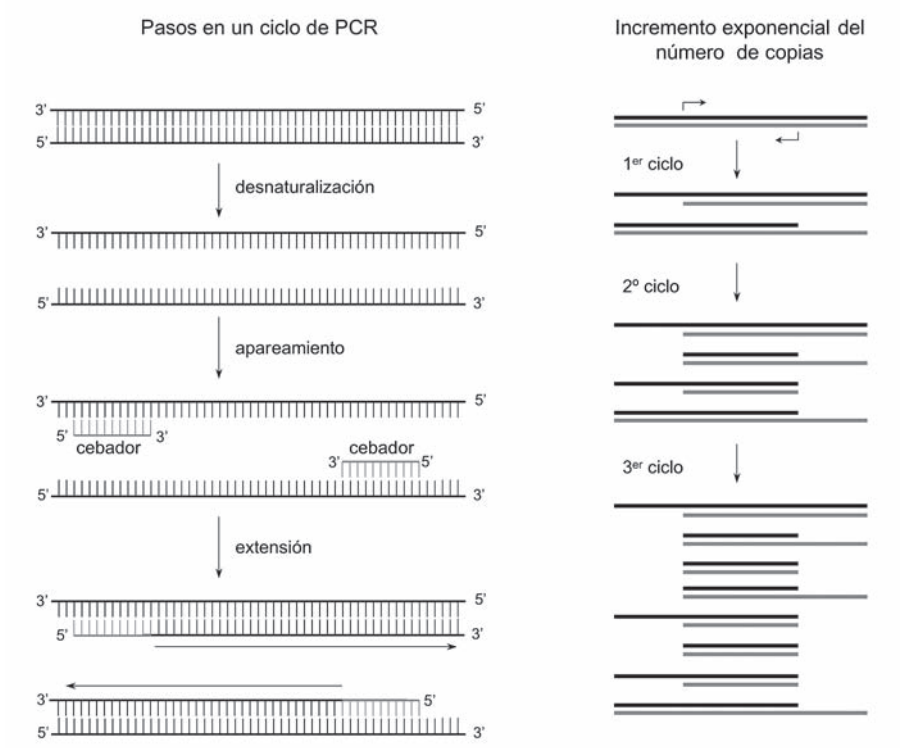
Al finalizar la replicación eucariota, las dos moléculas resultantes permanecen unidas por el centrómero, con lo que cada cromosoma está formado por dos moléculas de ADN denominadas **cromátidas hermanas**. Ahora la célula está en disposición de iniciar la división celular.

3.1.2. PCR: Replicación *in vitro*

El conocimiento de cómo la célula replica su ADN, ha permitido desarrollar una técnica de replicación *in vitro* del ADN que ha supuesto un avance revolu-

cionario en el campo de la biología molecular. Esta técnica es la reacción en cadena de la polimerasa o PCR *(Polymerase Chain Reaction)* con la que se obtienen millones de copias del fragmento de ADN deseado en muy poco tiempo (Figura 3.4).

Figura 3.4. Esquema de la reacción de PCR. Obsérvese que en cada nuevo ciclo el número de copias se dobla. A partir del tercer ciclo la mayoría de las moléculas tienen la longitud acotada por los cebadores



La reacción tiene lugar en un tubo que colocamos en un termociclador (o máquina PCR) que básicamente lo que hace es cambiar cíclicamente la temperatura de la reacción siguiendo un programa introducido por el usuario. La mezcla inicial del tubo de reacción debe contener el ADN a replicar, 2 cebadores (cadenas de unos 20 desoxirribonucleótidos), dNTPs (los 4 desoxirribonucleótidos: dATP, dCTP, dGTP y dTTP) y una enzima polimerasa termoestable (Taq). La secuencia de cada uno de los 2 cebadores debe ser complementaria a una cadena distinta del ADN de modo que delimiten el fragmento que deseamos amplificar.

La reacción consiste en una serie de ciclos (25-30) en los que la temperatura cambia en 3 pasos:

- Paso 1 (desnaturalización): temperatura elevada (>90°C) que provoca que las dos cadenas del ADN se separen dejando las bases accesibles.
- Paso 2 (apareamiento o *annealing*): temperatura baja (45-55°C) que permite a los cebadores unirse específicamente a su secuencia complementaria.
- Paso 3 (extensión o elongación): temperatura media (~60°C) en que, a partir de la secuencia del cebador unida a una cadena de ADN, la polimerasa incorpora nuevos nucleótidos complementarios a la cadena molde para construir una molécula de ADN de doble cadena.

Las condiciones de temperatura y tiempo para cada uno de los pasos de cada ciclo pueden variar considerablemente dependiendo de la longitud del fragmento a amplificar y de la secuencia de los cebadores. El diseño de la secuencia de los cebadores es crucial para que la reacción de PCR tenga éxito.

3.2. Reproducción celular. Mitosis

La **mitosis** es un proceso complejo por el cual la célula eucariota se reproduce. La importancia de la mitosis en la transmisión de la información genética radica en que las dos copias de cada molécula de ADN (llamadas cromátidas hermanas), que se han obtenido en la replicación, se separan repartiéndose equitativamente entre las células hijas. Para ello, primero desaparece la membrana nuclear y la célula despliega una matriz proteica de microtúbulos que arrastran a cada una de las cromátidas hermanas hacia polos opuestos de la célula. Cuando las 2 cromátidas se han separado, se las considera como cromosomas independientes. Los cromosomas que heredará cada célula hija se van separando hacia los polos opuestos de la célula y, una vez agrupados, se regenera la membrana nuclear. En este momento la célula muestra dos núcleos. La membrana citoplasmática crece estrangulando el citoplasma en dos porciones alrededor de cada núcleo hasta dividir la célula en dos nuevas células hijas que habrán heredado una copia idéntica de cada uno de los cromosomas de la célula original y aproximadamente la mitad del citoplasma con los distintos orgánulos que contenía.

Así, por mitosis se obtienen linajes de células genéticamente idénticas. El contenido genético no varía entre generaciones, exceptuando las pocas muta-

ciones que se hayan podido generar durante la replicación del ADN. Los primeros organismos eucariotas debieron reproducirse por mitosis, tal como hacen en la actualidad muchos eucariotas unicelulares y las células somáticas de los organismos pluricelulares.

Posteriormente apareció una innovación revolucionaria: la reproducción sexual. Con la reproducción sexual, en la que se mezcla la información genética de dos padres, los individuos generan descendencia genéticamente diversa. Esto permite a la especie experimentar la idoneidad de múltiples y nuevas combinaciones de genes en cada generación. La reproducción sexual consiste en dos procesos independientes: la meiosis y la fertilización.

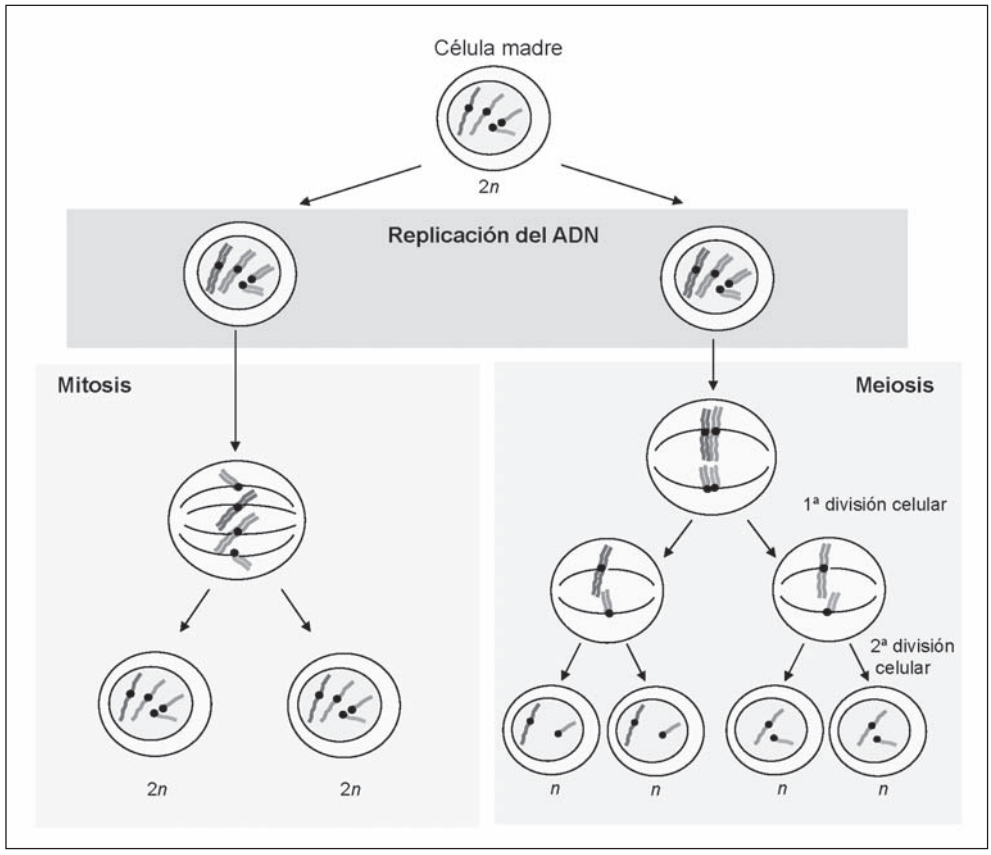
3.3. Reproducción sexual. Meiosis

En la reproducción sexual, dos células **(gametos)** provenientes de dos individuos se unen (fertilización) para originar una célula huevo **(cigoto)** a partir de la cual se construirá un individuo nuevo completo. Si las células que se unen para formar el cigoto fueran idénticas a las del resto del individuo, cada generación contaría con el doble de moléculas de ADN que la generación anterior. Para preservar el contenido genético de generación en generación, los organismos con reproducción sexual producen los gametos que son células con la mitad de cromosomas que el resto de células del individuo.

Los gametos se generan a partir de un tipo especial de división celular: la **meiosis**.

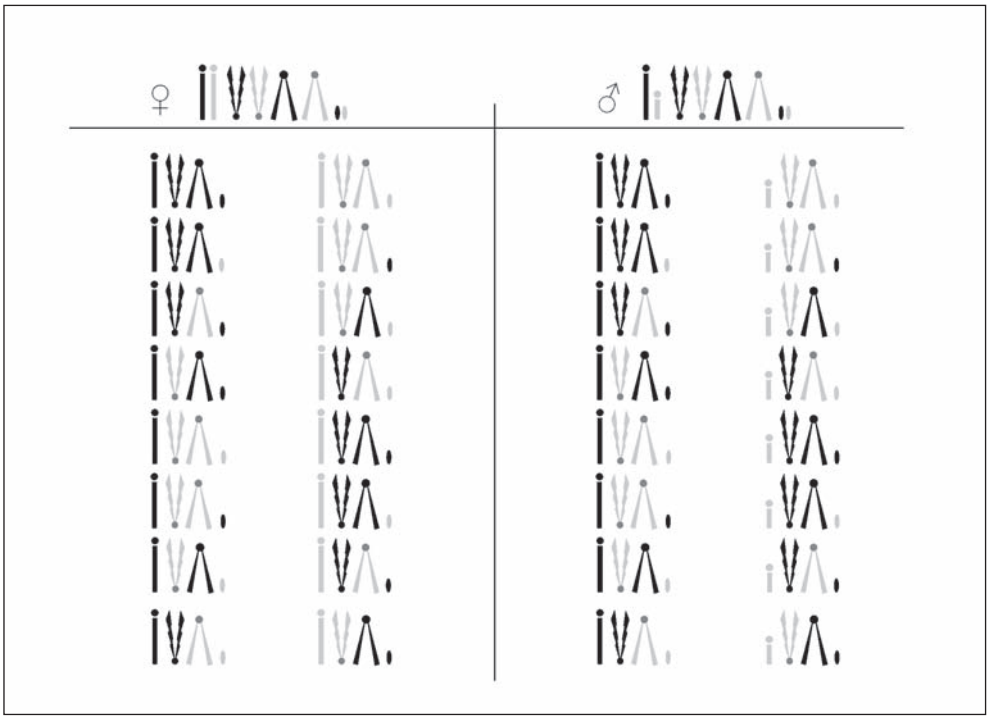
En un organismo diploide ($2n$), que contiene una copia materna y otra paterna de cada cromosoma, la meiosis genera a partir de una célula madre ($2n$), cuatro células hijas (n). Ello se debe a que el proceso se inicia con la replicación del ADN durante la interfase y a continuación tienen lugar 2 divisiones celulares consecutivas. En la primera división celular, los cromosomas homólogos (cada uno formado por dos cromátidas hermanas) se aparean situándose en el plano ecuatorial de la célula. La matriz proteica de microtúbulos arrastra a cada componente de la pareja de homólogos hacia un polo de la célula. De este modo se obtienen 2 células hijas con n cromosomas aunque $2n$ moléculas de ADN. La información genética de estas dos células hijas no es idéntica entre ellas ni respecto a la célula original. Cada célula hija ha heredado un único cromosoma de cada par homólogo. A continuación, y sin mediar una interfase, cada célula experimenta una nueva división celular. En esta división, las cromátidas hermanas se separan hacia los dos polos y al terminar se obtienen 4 células con n cromosomas y n moléculas de ADN cada una.

Figura 3.5. Comparación entre Mitosis y Meiosis



La reproducción sexual es una fuente importante de variabilidad genética entre los individuos de una población. Ya hemos visto cómo en una meiosis, a partir de una única célula se obtienen 4 células que, de no existir entrecruzamiento (ver más adelante), presentarán dos variantes distintas. Además, en la meiosis, los cromosomas de origen materno y paterno de distintos pares de homólogos se transmiten de forma independiente pudiendo generar diversas combinaciones de cromosomas. Así, cuanto mayor sea el numero de cromosomas que tiene la especie, mayor será el número de combinaciones distintas que se pueden obtener. Por ejemplo, una hembra de Drosophila que tiene 4 pares de cromosomas ($2n$=8) puede formar óvulos con 16 combinaciones cromosómicas distintas (Figura 3.6). Un macho de Drosophila también puede generar 16 tipos distintos de espermatozoides (Figura 3.6). Cuando ambos se apareen, su descen-

Figura 3.6. Combinaciones cromosómicas posibles en los gametos de un individuo



dencia podrá estar compuesta por individuos con 256 combinaciones cromosómicas distintas entre sí y que, además, serán distintas a las de sus progenitores.

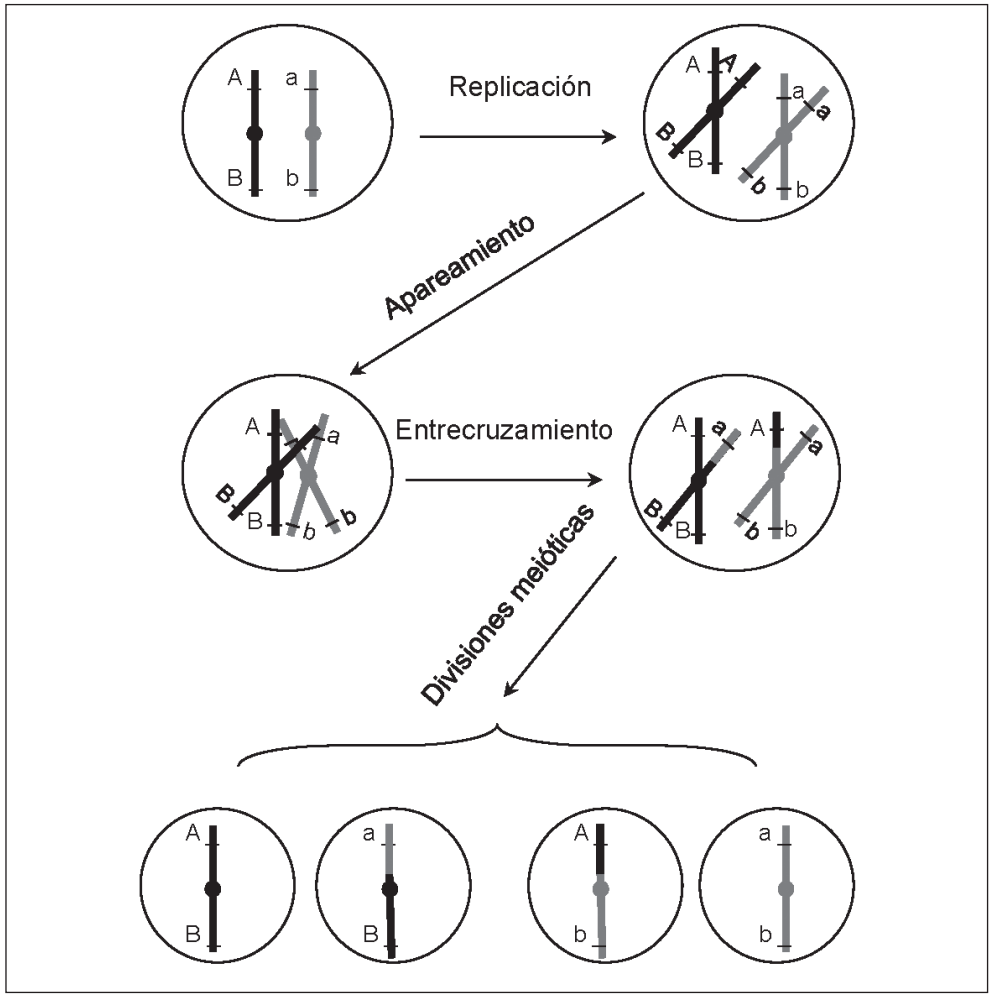
Pero la cosa no acaba aquí. Cada cromosoma contiene muchos genes distintos. Las combinaciones que hemos contemplado hasta ahora se refieren a cromosomas enteros. Pero durante la meiosis tiene lugar el **entrecruzamiento** entre cromátidas no hermanas de cromosomas homólogos, o recombinación intracromosómica, que es un fenómeno que enriquece todavía más la mezcla de genes aumentando el número de combinaciones posibles.

3.4. Recombinación intracromosómica

Se dice que los genes dispuestos sobre un mismo cromosoma están ligados puesto que tienden a heredarse conjuntamente. No obstante, y con cierta frecuencia, este ligamiento se rompe por recombinación intracromosómica. Al principio de la primera división celular de la meiosis, los cromosomas homólo-

gos permanecen apareados. En este momento pueden tener lugar entrecruzamientos, o intercambio de fragmentos de ADN, entre cromátidas de los cromosomas homólogos (Figura 3.7). El resultado es la posible aparición de multitud de nuevas combinaciones génicas.

Figura 3.7. Recombinación intracromosómica



Imaginemos que un cromosoma dispone de 2 genes marcadores (α y β) con 2 variantes o alelos cada uno: *A* y *a* para α y, *B* y *b* para β. Un individuo diploide, con 2 cromosomas homólogos, presenta la combinación alélica *AB* sobre el cro-

mosoma de origen paterno y la *ab* sobre su homólogo de origen materno. Por estar situados sobre el mismo cromosoma, o grupo de ligamiento, esperamos que las variantes *AB* se hereden conjuntamente así como las *ab*. No obstante, y con una frecuencia que será mayor cuanto más alejados físicamente se encuentren en el cromosoma, estas combinaciones se rompen y podemos encontrar gametos portadores de las combinaciones *Ab* y *aB* (Figura 3.7.).

El estudio de asociaciones entre marcadores tipo SNPs[3] o microsatélites y la manifestación de enfermedades tiene sentido debido a la existencia del ligamiento. En este caso se busca el ligamiento entre un determinado SNP y el gen causante de la enfermedad. En estos estudios es fundamental tener en cuenta que la recombinación puede romper dicho ligamiento. Debido precisamente a la recombinación, la asociación marcador-enfermedad será mayor cuanto más cercano se encuentre el marcador diagnóstico al gen causante de la enfermedad.

3. Los marcadores genéticos tipo SNPs y los microsatélites se verán en el capítulo VI sobre Genómica.

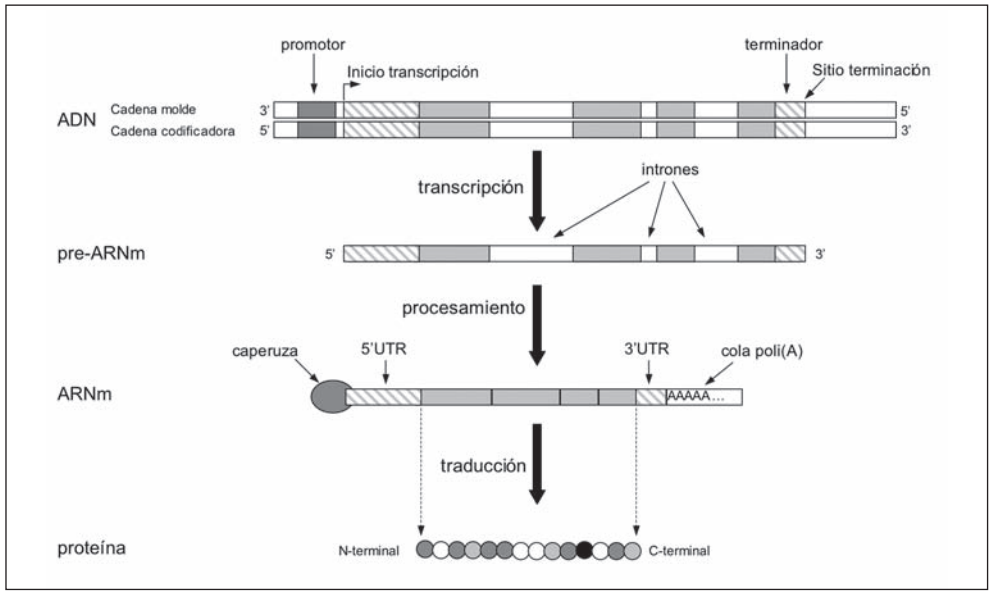
Capítulo IV

Síntesis proteica

A menudo se dice que la información es poder… pero si no sabemos gestionar la información es como si no dispusiéramos de ella. Si queremos descubrir el tesoro, no es suficiente con ser el feliz propietario del mapa del tesoro, heredado de generación en generación. Para conseguir nuestro objetivo, debemos conocer la clave que descifre el mapa y disponer de las habilidades que nos posibiliten llegar hasta su localización.

La información genética de la célula es fundamental para su supervivencia. De ahí la importancia en transmitirla fielmente a sus células hijas, como hemos visto en el capítulo anterior. Sin embargo, lo verdaderamente importante para cada célula es poder interpretar esta información que le permitirá decidir cómo actuar frente a cada situación ambiental.

Figura 4.1. Síntesis proteica en eucariotas



El ADN contiene, de forma codificada, la información (**genotipo**) necesaria para la síntesis de todas las proteínas que le confieren a la célula sus características (**fenotipo**) estructurales y metabólicas. La síntesis de las proteínas a partir de la información del ADN es un proceso complejo en el que interviene una tercera molécula cómo intermediaria, el ARN mensajero. En la síntesis proteica, podemos distinguir 3 pasos: **transcripción**, **procesamiento del ARN** y **traducción** (Figura 4.1). Cada uno de estos pasos presenta distintos niveles de regulación con lo que, al final, se obtiene la proteína adecuada en los niveles idóneos para cada situación concreta.

El funcionamiento y la reproducción de la célula precisan tanto de ADN como de proteínas. El ADN porta la información para la síntesis de las proteínas, pero las proteínas son indispensables en la replicación del ADN y en su descifrado. Por tanto, durante mucho tiempo existió el dilema de ¿qué fue primero, el ADN o las proteínas? El enigma se empezó a esclarecer en 1981, cuando se describieron los primeros ribozimas, unas moléculas de ARN que tienen capacidad catalítica.

4.1. Transcripción

Cada cromosoma está formado por una molécula de doble cadena de ADN que contiene la información de multitud de genes distintos, codificada en su secuencia de desoxirribonucleótidos. Muchos de estos genes codifican para la síntesis de proteínas. La maquinaria que interviene en la síntesis de proteínas, sin embargo, no lee esta información directamente del ADN, sino que utiliza una molécula intermediaria de ARN mensajero (ARNm). Cada molécula de ARNm contiene la información de un gen para la síntesis de una única proteína, codificada en su secuencia de ribonucleótidos. El código usado por ADN y ARN es básicamente el mismo, una secuencia de nucleótidos, por lo que el paso de ADN a ARN se conoce como transcripción.

El proceso de la transcripción tiene aspectos comunes con la replicación:

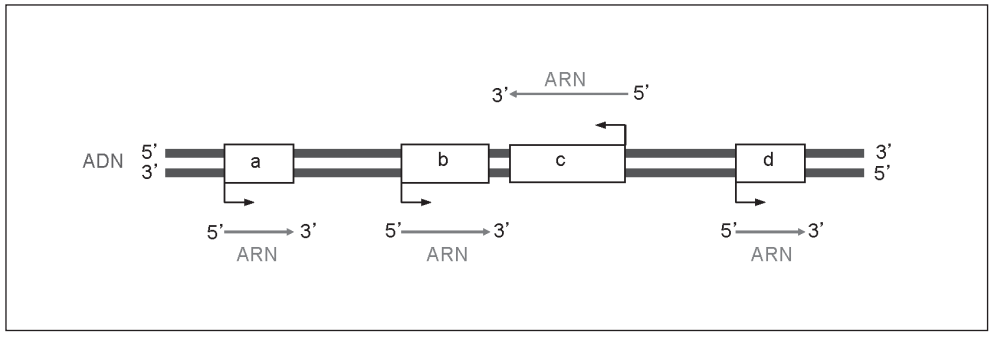
- Interviene un complejo enzimático que debe reconocer los lugares de inicio, desenrollar la doble hélice de ADN y separar sus bases complementarias para dejarlas expuestas a una polimerasa.
- La ARN polimerasa va colocando los nucleótidos trifosfato frente a sus complementarios de la cadena molde de ADN siguiendo las mismas reglas de complementariedad que el ADN (C frente a G, G frente a C, A frente T, pero

U frente a A) y cataliza la formación de los enlaces fosfodiester entre ellos, produciendo la elongación de la nueva cadena de ARN en sentido 5' –> 3'.

Pero la transcripción se diferencia en diversos aspectos de la replicación:

- La ARN polimerasa puede iniciar la síntesis de la nueva cadena de ARN sin necesidad de ningún cebador.
- Mientras que en la replicación se copia toda la cadena de ADN, en la transcripción sólo se copia a ARN un fragmento corto del ADN, correspondiente a un gen.
- Para cada gen, la ARN polimerasa sólo copia la información de una de las dos cadenas del ADN, la cadena molde.
- Los nucleótidos que se van incorporando son ribonucleótidos, contienen ribosa. Además, en lugar de Timina encontramos Uracilo.
- Los ribonucleótidos de la cadena de ARN que se va sintetizando no permanecen unidos por mucho tiempo a los nucleótidos de la cadena molde con puentes de hidrógeno. Así, a medida que la polimerasa avanza en la burbuja de trancripción, la nueva cadena de ARN se va separando de la cadena molde de ADN permitiendo que la doble hélice se recomponga.
- La transcripción a ARN no es tan fiel como la replicación del ADN. Se calcula que se comete un error cada 10000 nucleótidos. En este caso, los errores no son tan graves puesto que el ARN tiene una vida corta y, además, la célula cuenta con multitud de copias de ARNm para cada gen, traduciéndose simultáneamente.

Figura 4.2.



Ejemplo de localización de 4 genes sobre las dos cadenas del ADN, indicando la dirección de la transcripción. Los genes *a*, *b* y *d* están codificados por la misma cadena mientras que el gen c está codificado por la cadena complementaria.

Para cada gen, sólo se transcribe una de las dos cadenas de ADN. La cadena de ARN es complementaria a la cadena molde del ADN. Por tanto, la secuencia de bases del ARN es idéntica a la secuencia de bases de la cadena no molde del ADN, exceptuando que contiene Uracilo en lugar de Timina. Por ello, a la cadena no molde también se la conoce como cadena codificadora.

Ejemplo:

4.1.1. La unidad de transcripción

La unidad de transcripción es el fragmento de ADN que se transcribe a una molécula de ARN y las secuencias adyacentes necesarias para su transcripción. La unidad de transcripción consta de un promotor, la secuencia que se transcribe y un terminador. En los eucariotas, la unidad de transcripción incluye la información de un único gen, mientras que en los procariotas puede contener la información de varios genes. El **promotor** es una secuencia que es reconocida por la ARN polimerasa y le indica que cerca se encuentra la zona a transcribir. El promotor se sitúa en la región 5′, por delante del inicio de transcripción, por lo que se dice que se encuentra aguas arriba (*upstream*) respecto a éste. Por convenio, se indica al nucleótido correspondiente al inicio de la transcripción con el número +1. El resto de los nucleótidos de la unidad de transcripción se identifican por números negativos o positivos según su posición relativa al inicio de transcripción sea aguas arriba (*upstream*) o aguas abajo (*downstream*), respectivamente. El **terminador** es una secuencia de nucleótidos que indica a la polimerasa que la transcripción debe acabar. Generalmente, la transcripción finaliza una vez se ha transcrito el terminador.

La transcripción, generalmente se inicia bastante antes de la región que posteriormente será traducida a proteína. A esta porción de secuencia se la conoce como región 5′ no traducida o 5′ UTR (del inglés, *5′ Untranslated region*). Del mismo modo, la transcripción se prolonga más allá de la región codificadora obteniendo una región 3′ no traducida o 3′ UTR (del inglés, *3′ Untranslated region*).

Para que la ARN polimerasa inicie la transcripción, debe reconocer al promotor y unirse a éste. Aunque la secuencia del promotor de cada gen es distinta, presenta algunos fragmentos altamente conservados en que podemos reconocer una **secuencia consenso**. Una mutación en el promotor afecta la velocidad de la transcripción y un cambio en su posición altera la localización del lugar de inicio de la transcripción.

Una secuencia consenso es una secuencia que reconstruimos a partir de los nucleótidos que encontramos más frecuentemente en una posición determinada de distintas secuencias alineadas.

Por ejemplo, al comparar las siguientes secuencias:

```
5'..ATTCTATCGTACCAAT.. 3'
5'..ATTCTCTCGTACCCAA.. 3'
5'..AGTCTCTTGTACCGAT.. 3'
5'..ATTGTCTCTTACCTAT.. 3'
```

podemos construir la siguiente secuencia consenso:

```
5'..ATTCTCTCGTACCNAT.. 3'
```

En el ejemplo, la secuencia consenso no coincide con ninguna de las secuencias utilizadas para su construcción. Para cada posición nucleotídica se toma el nucleótido que aparece en más secuencias. En el caso del antepenúltimo nucleótido encontramos una N porque es una posición en que podemos encontrar cualquiera de los nucleótidos a frecuencias similares. Disponemos de otros símbolos de indeterminación para el caso de encontrar distintos nucleótidos a frecuencia similares (Tabla 4.1).

Tabla 4.1. Símbolos de indeterminación nucleotídica

Símbolo	Significado (origen)	Símbolo	Significado (origen)
R	G o A (*puRine*)	H	A o C o T (no G, H sigue a G)
Y	C o T (*pYrimidine*)	B	G o T o C (no A, B sigue a A)
M	A o C (*aMino*)	V	G o C o A (ni T ni U, V sigue a U)
K	G o T (*Keto*)	D	G o A o T (no C, D sigue a C)
S	G o C (*Strong interaction* 3 –H)		
W	A o T (*Weak interaction* 2 –H)	N	G o A o T o C (*aNy*)

Entre paréntesis se indica el origen del símbolo a partir de la terminología en inglés.

En líneas generales, la transcripción en procariotas y eucariotas es bastante similar, aunque en eucariotas es algo más compleja.

4.1.2 Transcripción en bacterias

En las bacterias interviene una única ARN polimerasa formada por diversas subunidades. Esta polimerasa se une transitoriamente a una subunidad llamada factor sigma (σ) que reconoce las secuencias consenso del promotor centradas en las posiciones –35 y –10. El factor sigma, además, interviene en el desenrollamiento de la doble hélice de ADN iniciando la burbuja de trancripción.

Una vez se ha iniciado la transcripción, el factor sigma se libera permitiendo que la polimerasa se desplace a lo largo de la burbuja alargando la nueva cadena de ARN. La cadena de ARN que se está sintetizando no permanece mucho tiempo unida al ADN molde. La burbuja de transcripción se va abriendo delante de la polimerasa y se va cerrando por detrás, permaneciendo sólo unos 8 nucleótidos de ARN emparejados a la cadena molde del ADN. Cuando la polimerasa detecta que el híbrido ARN-ADN es muy débil hace una parada y vuelve atrás para verificarlo. El proceso continúa hasta encontrar el terminador.

En las bacterias hay dos tipos de terminadores: los intrínsecos y los dependientes de rho. El terminador intrínseco suele ser una secuencia de unos 40 nucleótidos que contiene dos repeticiones invertidas de una secuencia rica en C y G[4] seguidas inmediatamente por una secuencia de al menos 6 A. Una vez que la polimerasa ha transcrito las repeticiones, la secuencia de ARN forma una estructura secundaria en horquilla emparejando las bases complementarias. Tras transcribir la tira de A a U, la polimerasa hace un paro pero la estructura en horquilla le impide volver atrás. Este paro momentáneo de la polimerasa, junto con el hecho de que las uniones A-U son más inestables que los otros pares de bases, favorece que la polimerasa se separe del ADN y que se acabe de liberar la cadena de ARN.

El terminador dependiente de rho posee una secuencia reconocida por el factor rho a unos 40-60 nucleótidos del final de la transcripción. Una vez que el factor rho se ha unido al ARN, éste se desplaza siguiendo a la polimerasa. Cuando la polimerasa llega a uno de los puntos en que realiza una pausa, es alcanza-

4. Debe recordarse que los pares C-G (con 3 puentes de hidrógeno) foman uniones más fuertes que los A-T o los A-U (con sólo 2 puentes de hidrógeno).

da por el factor rho que tiene función helicasa y provoca que la polimerasa se separe del ADN, terminando así la transcripción.

4.1.3 Transcripción en eucariotas

En los eucariotas, encontramos tres ARN polimerasas distintas, cada una de las cuales está especializada en la transcripción de distintos tipos de ARN. La ARN polimerasa I transcribe ARN ribosómico (ARNr), la ARN polimerasa II transcribe ARN mensajero (ARNm) y la ARN polimerasa III transcribe ARN de transferencia (ARNt) y otros pequeños ARN.

En los eucariotas, el ADN está unido a proteínas, formando la cromatina que le confiere una configuración muy compacta. Antes de la transcripción, esta estructura debe modificarse para que la cadena de ADN molde quede accesible. Para que la ARN polimerasa pueda iniciar la transcripción, requiere de la intervención de numerosas proteínas entre las que encontramos los factores de transcripción. Estas proteínas reconocen básicamente dos tipos de secuencia en el ADN: el promotor y los intensificadores *(enhancers)*. El promotor eucariota contiene una o más secuencias típicas. La más común de estas secuencias es la **caja TATA** (o *TATA box*) que tiene la secuencia consenso TATAAA y se localiza entre –25 y –30 pb del sitio de iniciación. La posición de los intensificadores es muy variable y pueden localizarse a mucha distancia del inicio de transcripción. Cada ARN polimerasa tiene un modo distinto de terminación. La ARN polimerasa I necesita de un factor de terminación que se une a la secuencia de ADN en una zona posterior a la de transcripción. La ARN polimerasa III finaliza la transcripción tras transcribir una sucesión de U, sin necesidad de la existencia de una estructura en horquilla. La terminación de la transcripción de la ARN polimerasa II viene determinada por una secuencia conservada (AAUAAA) que es reconocida por una enzima que corta el ARN naciente a unos 20 nucleótidos aguas abajo. La ARN polimerasa II sigue transcribiendo un fragmento de ARN pero pronto se separa del ADN, terminando la transcripción, y el fragmento extra de ARN es degradado.

La transcripción en las arqueas es muy similar a la de los eucariotas. Por ejemplo, tienen caja TATA y una proteína que se une a ésta, y también tienen un factor de transcripción fundamental para las polimerasas eucariotas que está ausente en las bacterias. Todo ello, pone de manifiesto que arqueas y eucariotas son grupos más cercanos entre sí que a las bacterias.

En los procariotas, al no tener un núcleo separado del citoplasma, a medida que el ARN se transcribe ya puede ser traducido por los ribosomas a proteína. En los eucariotas, el ARN transcrito debe atravesar la membrana nuclear antes de poder ser traducido en el citoplasma. Pero antes, aún debe de experimentar una serie de modificaciones.

4.2. Procesamiento del ARN

Aunque generalmente se habla de transcripción y posterior procesamiento del ARN, en realidad, el procesamiento es **cotranscripcional**. El ARN naciente va siendo modificado a medida que emerge del complejo de la ARN polimerasa. El procesamiento del ARN consiste en la modificación covalente de ambos extremos del ARN y en la eliminación de los intrones.

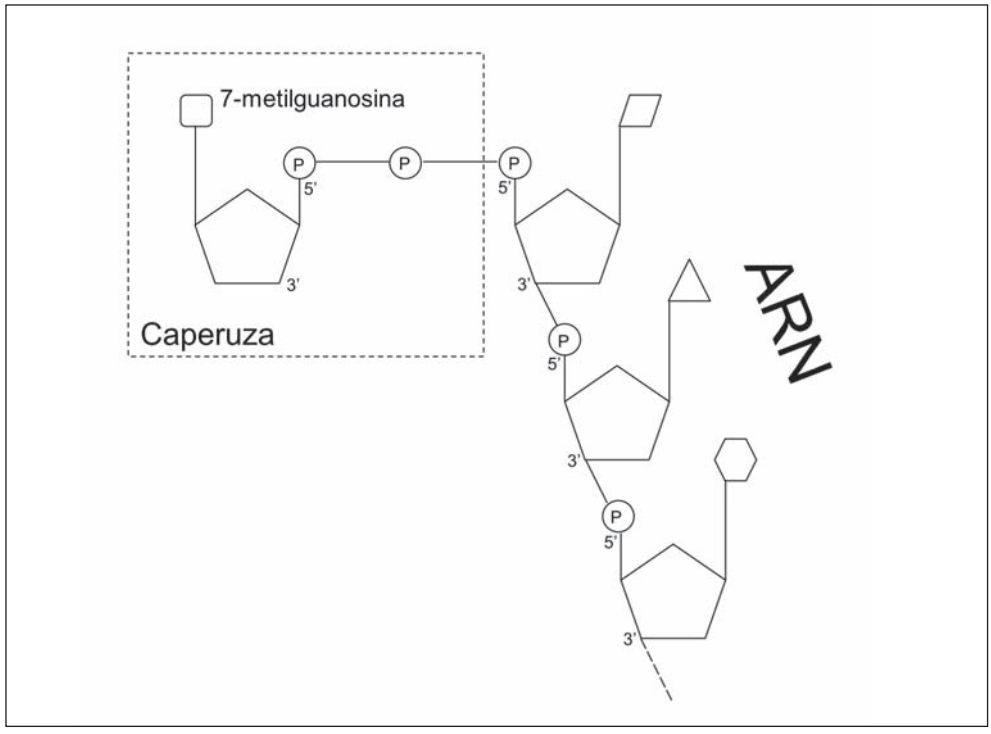
La modificación de los extremos del ARN consiste en la adición de una **caperuza** en el extremo 5' y la adición de una **cola de poli(A)** en el extremo 3'. Ambas modificaciones contribuyen a proteger al ARN contra la degradación y, además, constituyen señales para la célula de que la información que contiene la molécula de ARN está completa.

En los eucariotas, los genes incluyen secuencias que se expresan, los **exones**, separadas por otras secuencias que no se expresan, los **intrones**. En la maduración del pre-ARNm o ayuste (en inglés, *splicing*) se eliminan los intrones, para obtener un ARNm que podrá viajar al citoplasma y ser traducido.

En todo el procesamiento del ARN juega un papel muy importante el dominio C-terminal de la subunidad β de la ARN polimerasa II, conocido como dominio de la cola carboxilo o CTD (del inglés, *carboxil tail domain*). El CTD está ubicado cerca del sitio por donde el ARN recién sintetizado emerge de la ARN polimerasa II, situación que le permite ir interactuando con él. El CTD contiene una serie de repeticiones en tándem de 7 aminoácidos entre los que hay 2 serinas fosforilables. Según el grado de fosforilación de cada momento puede unirse a unas u otras proteínas que participan en la elongación de la transcripción y en el procesamiento del ARN.

Cuando la ARN polimerasa II ha unido ya unos 25 nucleótidos, una serie de enzimas unidas al CTD, añaden una caperuza al extremo 5'. Esta caperuza es un nucleótido modificado, la 7 metil guanosina, que se une al transcrito por un puente 5'-5' trifosfato (Figura 4.3).

Figura 4.3. Cadena de ARN con la adición de la caperuza en 5'



A medida que la transcripción avanza y antes de su terminación, los intrones van siendo eliminandos por el **empalmosoma** (en inglés, *spliceosome*). El empalmosoma es un complejo formado por 5 moléculas de snRNA (*small nuclear RNA*) y una serie de proteínas auxiliares que también está asociado al CTD y reconoce los extremos de los intrones. La gran mayoría de los intrones contienen la secuencia GU en su extremo 5' y la secuencia AG en su extremo 3', aunque también se conoce un buen número de intrones con los límites AU-AC. El empalmosoma corta el ARN por estas secuencias y empalma 2 exones consecutivos. Parece que existen dos complejos de empalmosoma, según los snRNA que lo componen. Cada uno de ellos reconocería principalmente, aunque no exclusivamente, una de las dos combinaciones de límites de los intrones (Sharp y Burge, 1997).

El CTD también porta proteínas que reconocen la secuencia de terminación de la transcripción. Cuando el ARN que emerge de la ARN polimerasa II contiene la secuencia de terminación, estas proteínas se unen al ARN permitiendo que una enzima corte el ARN y que, a continuación, una poli-A polimerasa añada uno a uno, y sin necesidad de ningún molde, unos 200 nucleótidos A.

4.2.1. Procesamiento alternativo

La existencia de intrones que deben ser eliminados tras la transcripción planteaba una cuestión práctica, ¿por qué los organismos mantenían los intrones? Aparentemente son un despilfarro energético para la célula, que se impone tanto la tarea de replicarlos como de transcribirlos inútilmente. Más tarde se observó que a partir de una única unidad de transcripción se pueden obtener distintas proteínas por procesamiento alternativo del ARN. Utilizando algunos fragmentos comunes y combinando otros fragmentos alternativos de una misma región de ADN se pueden obtener muchas proteínas distintas, aunque relacionadas. Al fin y al cabo, quizás a la célula le resulta más práctico mantener a los intrones que no a multitud de fragmentos de ADN, comunes a diversas proteínas, repetidos diversas veces. Además de la economía que le supone al compactar más la información de generación en generación, también le permite tener un mejor control sobre la regulación de la expresión de diversas proteínas en distintos momentos o lugares.

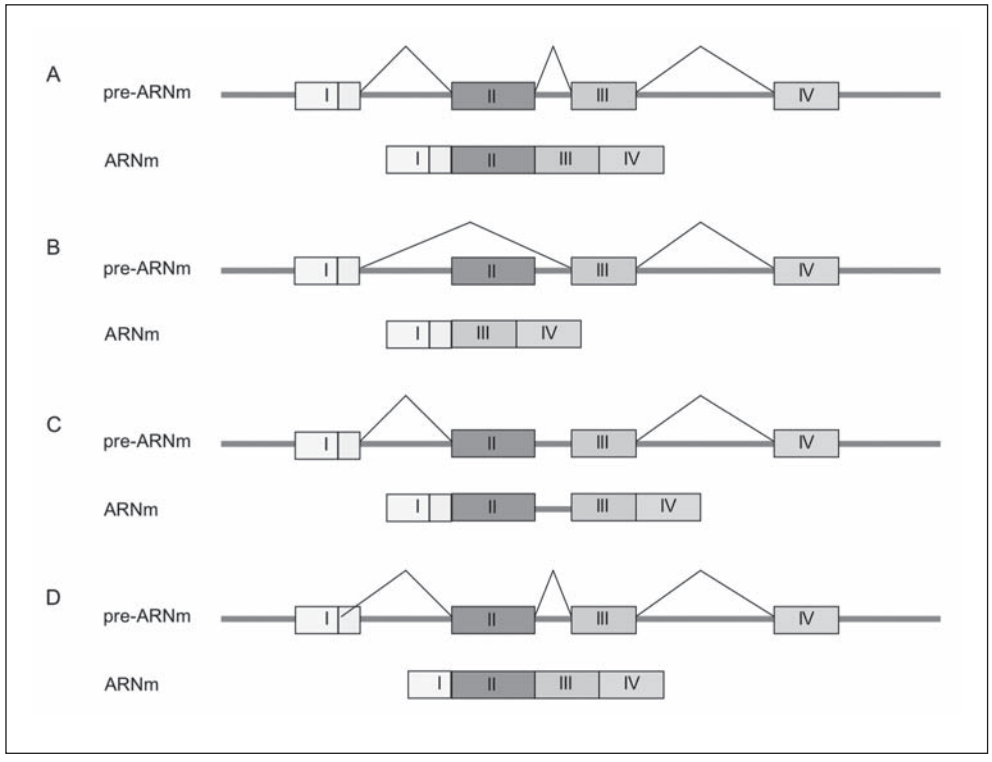
La importancia del ayuste alternativo (o *splicing* alternativo) se ha puesto especialmente de relieve tras la secuenciación de los primeros genomas. Se pensaba que los organismos más complejos, con más proteínas, debían contar proporcionalmente con más genes. Por ello, sorprendió que en el genoma humano sólo se anotaran aproximadamente 38.000 genes mientras que en *Caenorhabditis elegans* (un invertebrado muy simple) se anotaran aproximadamente 23.000. La solución a esta paradoja nos la proporciona el fenómeno del ayuste alternativo, que podría estar afectando entre un 35% y un 90% de los genes humanos, según las estimas de distintos autores.

La estructura variable del ARNm puede originarse por alteraciones en el sitio de inicio de la transcripción o en el sitio de finalización y poliadenilación pero también frecuentemente por distintos tipos de ayuste alternativo (Black, 2000). Los exones pueden ser incluidos o excluidos del ARNm y algunas secuencias que normalmente son intrones pueden permanecer en el ARNm. Las posiciones de corte pueden variar dando lugar a exones más o menos largos. Además, puede haber *trans-splicing* alternativo, unión alternativa de exones provenientes de distintas unidades de transcripción.

El ayuste alternativo permite controlar la inclusión de secuencias peptídicas particulares, seleccionando, para una posición exónica determinada, entre una serie de secuencias mutuamente excluyentes. Un ejemplo extremo lo encontramos en el gen *Dscam* de *Drosophila melanogaster* (Graveley, 2005). La secuencia

del ARNm de este gen se obtiene tras ensamblar 24 exones. Las secuencias de los exones 4, 6, 9 y 17 son elegidas entre 12, 48, 33 y 2 secuencias alternativas, respectivamente. Si todas las combinaciones posibles fueran usadas, el gen *Dscam* codificaría para 38.016 proteínas distintas, mientras que en el genoma de *D. melanogaster* sólo hay anotados ~14.000 genes.

Figura 4.4. Distintos ayustes de un mismo pre-ARNm



Las cajas indican los exones y la barra los intrones. Las líneas oblicuas indican los intrones escindidos en cada uno de los ayustes alternativos. En el ejemplo B, una secuencia normalmente exónica es escindida. En C, un intrón es retenido en el ARNm. En D, se utiliza un sitio de corte diferente para el primer intrón con lo que el exon I resulta más corto.

La producción de una u otra variante de ARNm depende de diversos factores, como pueden ser el ambiente genotípico, el linaje celular o el momento del desarrollo, pero no se conoce demasiado cómo se realiza la elección. Algunos sitios de corte no encajan totalmente con las secuencias consenso (GU-AG o AU-AC) mientras que otros sitios con secuencias consenso no son usados, posiblemente porque otras secuencias cercanas también deben ser importantes en su recono-

cimiento. Todo ello dificulta la predicción de exones y supone uno de los retos actuales para la bioinformática (Black, 2000).

4.3. Traducción

Una vez que el ARN ha sido procesado puede atravesar los poros de la membrana nuclear y llegar al citoplasma, donde será traducido. La traducción es la interpretación de la información codificada en la secuencia de nucleótidos del ARNm a secuencia de aminoácidos de una proteína. Podríamos decir que es el paso de genotipo a fenotipo.

En la traducción, la secuencia de nucleótidos del ARNm sirve de molde para la síntesis de una secuencia de aminoácidos. La traducción está dirigida por los ribosomas, un complejo ribonucleoproteico (ARNr y diversas proteínas) formado por dos subunidades. Además se necesita de la intervención de un tercer tipo de ARN, los ARN de transferencia. Los ARNt actúan de moléculas adaptadoras, que transportan aminoácidos particulares hasta el ribosoma, donde los aminoácidos se unen para formar cadenas polipeptídicas según la secuencia de nucleótidos del ARNm.

4.3.1. Código genético

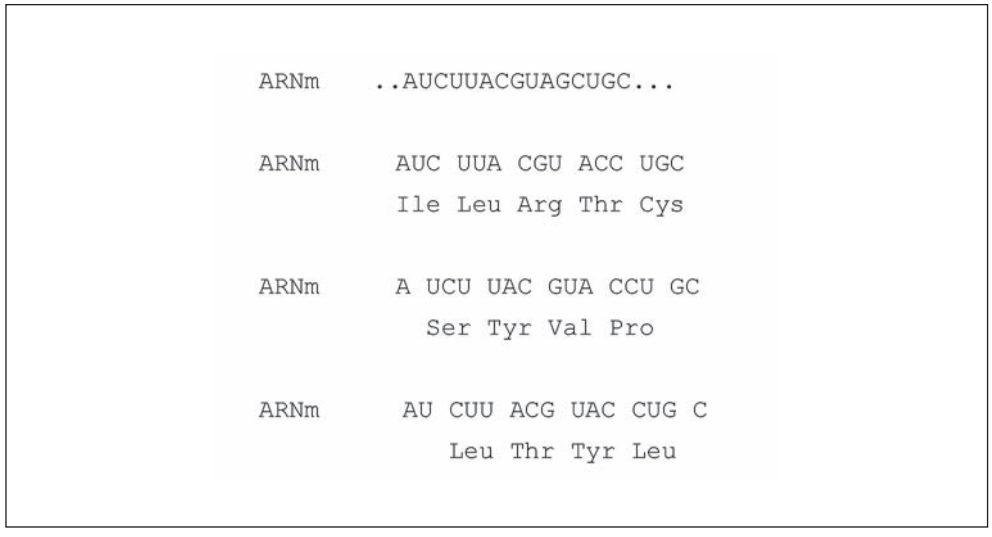
La unidad básica del código genético es el **codón**, la secuencia de nucleótidos que codifica para un aminoácido. Si una secuencia de 4 nucleótidos distintos codifica para 20 aminoácidos distintos, debe hacerlo, como mínimo, utilizando grupos de 3 nucleótidos, o tripletes, por aminoácido[5]. Con un código de tripletes se dispone de 4 x 4 x 4 = 64 codones distintos (Tablas 4.2 y 4.3). El código genético, cuenta con 3 codones de paro, o STOP, y 61 codones con sentido. Se dice que el código genético es degenerado porque algunos aminoácidos están codificados por diversos codones (hay codones sinónimos).

5. Combinaciones de 2 nucleótidos sólo permitirían codificar para 16 aminoácidos (4 x 4 = 16 combinaciones).

La primera evidencia experimental del código en tripletes la aportaron Crick y colaboradores (1961) trabajando con mutantes del gen *rII* del fago T4. Para ello utilizaron mutantes obtenidos por tratamiento con proflavina que provoca la inserción o deleción de un único par de bases. El fago T4 es capaz de crecer en 2 cepas distintas, B y K, de *Escherichia coli*. A partir de un mutante que era incapaz de crecer en la cepa K, obtuvieron nuevos mutantes, algunos de los cuales revertían a un fenotipo pseudosalvaje (casi normal). Esto se explica porque las dos mutaciones eran de signo contrario. Si una mutación era una deleción, la siguiente era una inserción, o viceversa, con lo que se recuperaba la pauta de lectura. A continuación, obtuvieron, por recombinación, mutantes triples y observaron que sólo recuperaban el fenotipo salvaje cuando las 3 mutaciones eran del mismo signo, indicando que se necesita la inserción (o deleción) de 3 nucleótidos para recuperar la pauta de lectura.

Como cada aminoácido viene codificado por un triplete, existen 3 **pautas de lectura** posibles. Las proteínas que se podrían sintetizar a partir de cada una de estas pautas son muy distintas (Figura 4.5). Así pues, la maquinaria encargada de la traducción debe reconocer un punto de inicio (codón de iniciación) para la síntesis de proteína que marcará qué pauta de lectura es la correcta. El codón de iniciación es el primer triplete del ARNm que especifica para un aminoácido. Generalmente se trata del codón AUG que codifica para metionina, aunque en ocasiones más raras también puede ser un codón GUG o UUG.

Figura 4.5. Las tres pautas de lectura potenciales para un fragmento de ARNm



Un cambio en un nucleótido (una mutación por substitución) puede comportar un cambio en un aminoácido. La pérdida o ganancia de un único nucleótido (mutación por deleción o inserción) conlleva consigo el cambio en la pauta de lectura a partir de la posición mutada. Así, una mutación por inserción o deleción acostumbra a tener un efecto mucho más drástico que una simple substitución.

El código genético (equivalencia entre tripletes de ácido nucleico y aminoácidos en las proteínas) es común en la mayoría de organismos. Por ello decimos que el código genético es universal (Tabla 4.2), aunque existen algunas excepciones. La mayoría de las excepciones se encuentran en los genes de las mitocondrias, aunque también hay en el ADN nuclear de algunos organismos. La mayoría de los codones implicados en las excepciones son codones de STOP. Así,

Tabla 4.2. Código genético universal (búsqueda por codones).

		Segunda Posición									
		U		C		A		G			
Primera posición	U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U	Tercera Posición
		UUC		UCC		UAC		UGC		C	
		UUA	Leu	UCA		UAA	Stop	UGA	Stop	A	
		UUG		UCG		UAG	Stop	UGG	Trp	G	
	C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U	
		CUC		CCC		CAC		CGC		C	
		CUA		CCA		CAA	Gln	CGA		A	
		CUG		CCG		CAG		CGG		G	
	A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U	
		AUC		ACC		AAC		AGC		C	
		AUA		ACA		AAA	Lys	AGA	Arg	A	
		AUG	Met	ACG		AAG		AGG		G	
	G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U	
		GUC		GCC		GAC		GGC		C	
		GUA		GCA		GAA	Glu	GGA		A	
		GUG		GCG		GAG		GGG		G	

Tabla de 3 entradas del Código Genético Universal. La columna de la izquierda indica el primer nucleótido, la fila superior indica el segundo y la columna de la derecha el tercero para un triplete o codón en cuestión.

en los Ciliados (un grupo de Protozoos entre los que encontramos a *Paramecium*) sólo hay 1 codón de STOP. En estos organismos, los codones UAA y UAG codifican para Glutamina.

Tabla 4.3. Código genético universal (búsqueda por aminoácidos)

Aminoácido			Codones
	Código		
Nombre	3 letras	1 letra	
Ácido Aspártico	Asp	D	GAC, GAU
Ácido Glutámico	Glu	E	GAA, GAG
Alanina	Ala	A	GCA, GCC, GCG, GCU
Arginina	Arg	R	AGA, AGG, CGA, CGC, CGG, CGU
Asparragina	Asn	N	AAC, AAU
Cisteina	Cys	C	UGC, UGU
Fenilalanina	Phe	F	UUC, UUU
Glicina	Gly	G	GGA, GGC, GGG, GGU
Glutamina	Gln	Q	CAA, CAG
Histidina	His	H	CAC, CAU
Isoleucina	Ile	I	AUA, AUC, AUU
Leucina	Leu	L	UUA, UUG, CUA, CUC, CUG, CUU
Lisina	Lys	K	AAA, AAG
Metionina	Met	M	AUG
Prolina	Pro	P	CCA, CCC, CCG, CCU
Serina	Ser	S	AGC, AGU, UCA, UCC, UCG, UCU
Tirosina	Tyr	Y	UAC, UAU
Treonina	Thr	T	ACA, ACC, ACG, ACU
Triptófano	Trp	W	UGG
Valina	Val	V	GUA, GUC, GUG, GUU
STOP			UAA, UAG, UGA

Relación de cada aminoácido con los tripletes que lo codifican así como con los códigos de 3 o 1 letra utilizados para designarlos.

4.3.2. Marco de lectura abierto

Un marco de lectura abierto u ORF (del inglés, *open reading frame*) es cualquier secuencia del ADN que potencialmente codifica para una proteína. Un ORF es una secuencia de codones con sentido, delimitada por un codón de inicio y un codón de parada de la traducción, y que, en general, tiene una longitud considerable. Tanto los intrones como las regiones 5'UTR y 3'UTR no forman parte del ORF.

La anotación de genes en el genoma por medios bioinformáticos se basa en la identificación de ORFs. Los algoritmos empleados para identificar genes deben tener en cuenta que el ADN puede contener ORFs en 6 pautas de lectura diferentes, 3 en cada una de sus cadenas y además reconocer los posibles intrones.

4.3.3. El ARNt: una molécula adaptadora

La secuencia de nucleótidos de una molécula de ARNm determina una secuencia de aminoácidos, pero ¿cuál es el mecanismo por el que los aminoácidos reconocen en qué orden deben unirse? En 1958, y sin ninguna evidencia experimental, Francis Crick especulaba lo siguiente:

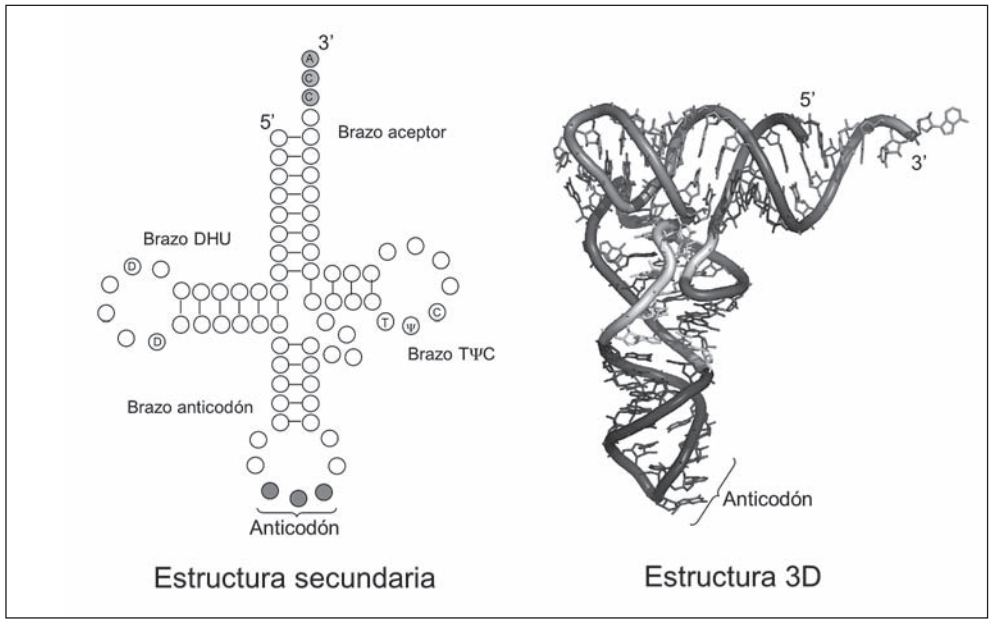
> «Por tanto, una hipótesis natural es que el aminoácido es transportado hasta el molde por una molécula «adaptadora», y que el adaptador es la parte que realmente encaja en el ARN. En su versión más simple se necesitan 20 adaptadores, uno por cada aminoácido [...] existe una posibilidad que parece inherentemente más probable que cualquier otra – que las moléculas adaptadoras deben contener nucleótidos. Esto les permitiría unirse al ARN molde según el mismo «apareamiento» de bases que se encuentra en el ADN, o en polinucleotidos.» Crick (1958).

Hoy en día sabemos que, efectivamente, intervienen moléculas adaptadoras que son los ARN de transferencia o ARNt. Cada molécula de ARNt está formada por una secuencia de unos 80 ribonucleótidos. Los ARNt son transcritos en el núcleo por la ARN polimerasa III y deben experimentar un procesamiento especial antes de viajar al citoplasma. Por un lado, el transcrito precursor del ARNt es más largo y en algunos casos incluye intrones, por lo que debe ser cortado. Por otro lado, muchos de sus nucleótidos son modificados después de la trans-

cripción.[6] Algunos de estos nucleótidos modificados intervienen en el reconocimiento del codón del ARNm y otros en la unión del aminoácido correcto.

La secuencia de nucleótidos del ARNt adopta una estructura secundaria especial en forma de hoja de trébol (Figura 4.6) por el apareamiento de bases alejadas en la cadena. Esta estructura se pliega sobre sí misma en forma de L. Los nucleótidos de los 2 extremos de esta L están desapareados y son muy importantes en el funcionamiento del ARNt. En uno de los extremos de la L hay el **anticodón**, secuencia de 3 nucleótidos que reconoce y se aparea con el codón complementario del ARNm. En el otro extremo de la L hay el **brazo aceptor**, que es donde se une al aminoácido correspondiente. El brazo aceptor tiene un fragmento de cadena sencilla de ARN que corresponde al extremo 3' del ARNt y acaba en la secuencia CCA. El grupo carboxilo del aminoácido transportado por el ARNt se une al grupo hidroxilo 2' o 3' de la adenina terminal del ARNt.

Figura 4.6. Estructura del ARN de transferencia



Sobre el esquema de la estructura secundaria en forma de hoja de trébol se indican los nucleótidos correspondientes al anticodón y al extremo CCA de unión al aminoácido. También se indican la posición de algunas bases características que dan nombre a los brazos TΨC y DHU. (Estructura 3D obtenida en Wikipedia, reproducida con permiso de N.R.Voss)

6. Se conocen más de 50 nucleótidos modificados distintos en el ARNt. Algunos de éstos son: la inosina (I), la dihidrouridina (D) y el pseudouracilo (Ψ).

Cada ARNt es específico para un aminoácido distinto, pero todos ellos unen su aminoácido a la secuencia CCA 3' terminal. La unión específica se consigue por la acción de las enzimas aminoacil-tRNA sintetasas. Existen 20 enzimas distintas, una por cada aminoácido, que reconocen tanto al aminoácido como a su correspondiente ARNt.

Como el código genético es degenerado, caben dos posibilidades: que un mismo ARNt reconozca diversos codones; o que exista más de un ARNt para muchos de los aminoácidos. En realidad las dos opciones no son excluyentes y ambas tienen lugar en la célula. Aunque en las proteínas normalmente sólo hay 20 aminoácidos, en la célula coexisten entre 30 y 50 ARNt distintos. Algunos ARNt distintos portan el mismo aminoácido, son los isoaceptores. Pero, por otro lado, existen 61 codones con sentido mientras que no hay tantos ARNt. Algunos ARNt reconocen, además de su codón complementario, otros codones alternativos. Estos ARNt reconocen y se unen fuertemente a las dos primeras bases del codon del ARNm pero toleran un emparejamiento débil con la tercera base del codón, permitiendo que no sea la complementaria en el anticodón. Este fenómeno que se conoce como **tambaleo** explica por qué la mayoría de codones sinónimos sólo varían en la tercera base. A menudo, el tambaleo es favorecido por la presencia en la primera base del anticodón (que se aparea a la tercera del codón) de un nucleótido modificado como la inosina (I).

4.3.4. El proceso de la traducción

La traducción tiene lugar en los ribosomas. Un ribosoma se une cerca del extremo 5' de una molécula de ARNm y se desplaza hacia el extremo 3' traduciendo los codones a medida que se desplaza. La síntesis de la proteína se inicia por su extremo amino y se alarga por la adición de aminoácidos al extremo carboxilo.

Los ribosomas están formados por 2 subunidades cuyos nombres indican su tamaño molecular[7] (30S y 50S en bacterias; 40S y 60S en eucariotas). Cada subunidad está, a su vez, constituida por ARN ribosómico (ARNr), que representa los 2/3 de su masa, y por proteínas. Primero se pensó que el ARNr tenia una

7. Las subunidades de los ribosomas se caracterizaron originalmente por su sedimentación diferencial tras someterlas a ultracentrifugación. Por ello se nombran a partir de su coeficiente de sedimentación en unidades de Svedberg (S).

función básicamente estructural pero ahora se piensa que interviene activamente en la síntesis proteica.

El ribosoma tiene un sitio de unión al ARNm, que se encuentra totalmente en su subunidad pequeña, y 3 sitios de unión al ARNt: sitio A (aminoacilo), sitio P (peptidilo) y sitio E (*exit*, salida en inglés). Mientras un ARNt permanece unido al ribosoma, hace de puente entre sus dos subunidades, apareando su anticodón con el codón correspondiente del ARNm en la subunidad pequeña, y manteniendo su extremo portador de aminoácido en la subunidad grande.

El inicio de la traducción precisa de una serie de elementos: el ARNm, las subunidades grande y pequeña del ribosoma desensambladas, los factores de iniciación, el ARNt iniciador y guanosina trifosfato (GTP).

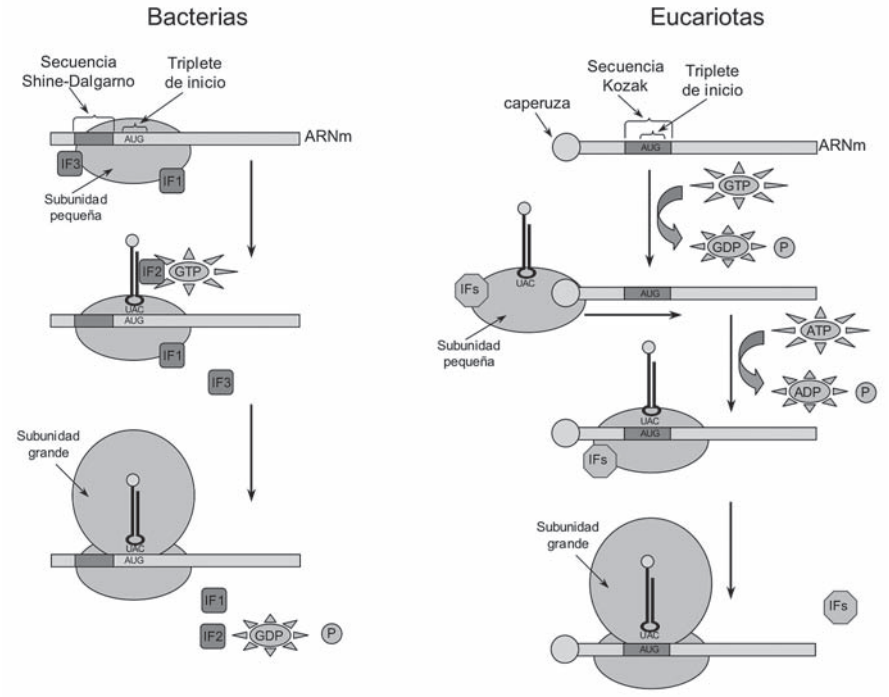
El ARNt iniciador es un ARNt especial que transporta metionina (en las bacterias, formilmetionina). Por ello, todas las proteínas tienen inicialmente una metionina en su extremo N-terminal, aunque más tarde puede ser eliminada por una proteasa específica. Este ARNt iniciador tiene una secuencia de nucleótidos distinta a la del ARNt que normalmente transporta la metionina. El ARNt iniciador se posiciona directamente en el sitio P del ribosoma, mientras que el resto de ARNt debe pasar primero por el sitio A.

En las bacterias, el extremo 3' del ARNr de la subunidad pequeña se une directamente a una secuencia especial del ARNm, la secuencia de Shine-Dalgarno, que precede al codón de iniciación (Shine y Dalgarno, 1975). De este modo, el codón de iniciación queda posicionado directamente en el sitio P, donde se unirá el ARNt iniciador. A continuación se liberan los factores de iniciación permitiendo la unión de las dos subunidades del ribosoma (Figura 4.7).

En los eucariotas, la subunidad pequeña del ribosoma, unida al ARNt de iniciación y a diversos factores de iniciación, se une a la caperuza del ARNm y empieza a desplazarse en dirección 5'–>3' en busca del codón de inicio (Figura 4.7). La identificación del codón AUG correcto se realiza por la presencia de una secuencia consenso, la secuencia Kozak, que rodea al codón de iniciación (Kozak, 1987). Una vez que el codón de iniciación se encuentra alineado con el ARNt iniciador, se liberan los factores de iniciación y se ensambla la subunidad mayor del ribosoma.

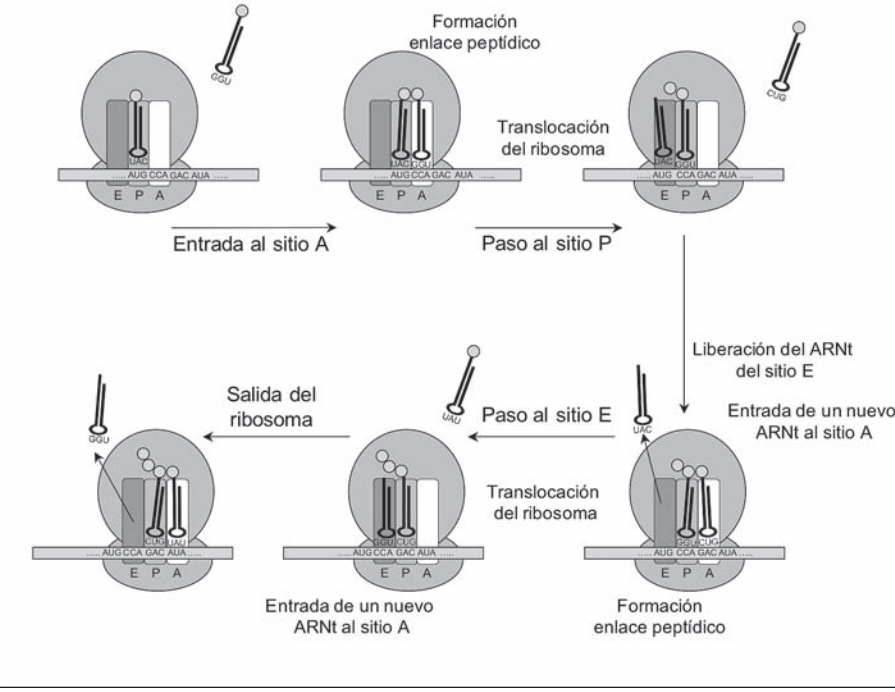
La elongación de la cadena peptídica se realiza en una serie de ciclos que constan de tres pasos (Figura 4.8). En un primer paso, el ARNt cargado con su aminoácido correspondiente entra en el sitio A. Este paso necesita la presencia de algunas proteínas que actúan como factores de elongación y la energía aportada por una molécula de GTP. En el segundo paso se crea un enlace peptídico

Figura 4.7. Inicio de la traducción. Diferencias entre Bacterias y Eucariotas



entre los aminoácidos portados por los ARNt que están situados en los sitios P y A. Esto provoca que el ARNt del sitio P se separe del aminoácido que portaba y que la cadena peptidíca quede unida al ARNt del sitio A. Durante mucho tiempo se pensó que la formación de este enlace peptídico se debía a la acción de alguna proteína de la subunidad mayor del ribosoma. Sin embargo, las evidencias actuales apuntan a que la actividad catalítica es responsabilidad del ARNr de la subunidad grande que actuaría como ribozima. El tercer paso es la **translocación** o desplazamiento del ribosoma para posicionarse sobre el siguiente codón del ARNm. Este paso también requiere un factor de elongación y la energía aportada por una molécula de GTP. Los ARNt de los sitios P y A permanecen unidos al ARNm por la unión codón-anticodón con lo que no se desplazan junto al ribosoma. El ARNt que ocupaba el sitio P pasa ahora al sitio E desde el que es liberado al citoplasma, donde podrá cargarse con un nuevo aminoácido. El ARNt que ocupaba el sitio A, se encuentra ahora en el sitio P. Así, el sitio A, queda vacío permitiendo la entrada de un nuevo ARNt cargado con el aminoácido correspondiente al siguiente codón.

Figura 4.8. Elongación de la cadena peptídica en el ribosoma. Seguimiento de un ARNt desde su entrada en el sitio A del ribosoma hasta su salida por el sitio E



La elongación continúa hasta que el ribosoma se transloca sobre un codón de terminación. Como no hay ningún ARNt con anticodón complementario al codón de terminación, el sitio A queda libre hasta que los factores de liberación se unen al ribosoma. Estos promueven la liberación del ARNt del sitio P y su separación de la cadena polipeptídica recién sintetizada.

4.4. Modificaciones postraduccionales

La mayoría de proteínas recién sintetizadas no son funcionales. Para ser funcional, toda proteína debe plegarse correctamente y, en muchos casos, también debe experimentar modificaciones diversas en la secuencia de sus aminoácidos.

Una proteína plegada correctamente se dice que presenta su **conformación nativa**, a diferencia de otra no plegada o plegada incorrectamente, que está en

una conformación no nativa. Muchas proteínas tienen un plegamiento nativo difícil de alcanzar en un ambiente acuoso como el de la célula y, para lograrlo, deben intervenir las proteínas chaperonas.

La estructura nativa puede perderse, como en la infección por priones. Un **prión** es una proteína con una secuencia de aminoácidos normal pero que ha adoptado un plegamiento anómalo transmisible. Es decir que su presencia induce a que otras moléculas de igual secuencia adopten la misma estructura terciaria anómala.

Entre las modificaciones en la secuencia de aminoácidos, la más frecuente es la eliminación de la metionina del extremo amino de la proteína. Además, algunas proteínas eliminan los primeros 15-30 aminoácidos de su extremo amino. Estos aminoácidos se conocen como **secuencia señal** y dirigen a la proteína hasta la zona de la célula dónde debe actuar. Una vez alcanzado este lugar, la secuencia señal es eliminada. Más raramente, algunas proteínas, se sintetizan como precursores mucho mayores que deben ser recortados por enzimas especiales antes de convertirse en funcionales.

Por otro lado, algunas proteínas deben unirse a hidratos de carbono para poder ser funcionales y otras muchas necesitan formar asociaciones de cadenas polipeptídicas (alcanzar una estructura cuaternaria). Es el caso de muchos enzimas diméricos (formados por dos polipéptidos) o el ejemplo tan conocido de la hemoglobina que para ser funcional necesita que 4 cadenas polipeptídicas estén unidas.

Otro grupo de modificaciones lo constituyen las modificaciones de las cadenas laterales de los aminoácidos, entre las que destacan la adición de grupos fosfato, carboxilo y metilo.

4.5. Selenoproteínas: un uso especial del código genético

Hasta ahora hemos considerado que las proteínas son cadenas en las que se combinan 20 aminoácidos diferentes. Estos 20 aminoácidos son los aminoácidos universales, presentes en todos los organismos y codificados directamente por el código genético universal. No obstante, existen otros aminoácidos raros, que sólo se encuentran en algunas proteínas, como la hidroxiprolina presente en el colágeno. Estos aminoácidos raros no se incorporan durante la traducción, sino que resultan de la modificación postraduccional de las cadenas laterales de los aminoácidos codificados por el ARNm.

Las proteínas que contienen el aminoácido especial selenocisteína (Sec) se conocen como selenoproteínas.[8] La selenocisteína se diferencia de la cisteína por tener un átomo de selenio (Se) en lugar de azufre. Estas proteínas son de gran importancia metabólica y se han encontrado tanto en arqueas como en bacterias y eucariotas.

Lo interesante de estas proteínas es que la presencia de Sec en la proteína no responde a una modificación postraduccional. La incorporación de selenocisteína es traduccional, siendo codificada por el codón UGA (Zinoni y colaboradores, 1986; Hatfield y Gladyshev, 2002), por lo que ha pasado a considerarse el aminoácido número 21. El codón UGA es un codón que normalmente es de parada de la traducción. En una misma célula, el uso del codón UGA es doble, como parada y como codificador de Sec. La maquinaria de traducción reconoce el significado de este codón por el contexto de la secuencia de ARNm. Para que el codón UGA sea traducido a Sec, es necesaria una secuencia especial de nucleótidos del ARNm que adopta una estructura secundaria en horquilla llamada elemento SECIS (del inglés, *Sec insertion sequence*). En eucariotas y arqueas el elemento SECIS se sitúa en la región 3' UTR, mientras que en bacterias se localiza en la región codificadora, inmediatamente aguas abajo del codón UGA que codifica para Sec (Hatfield y Gladyshev, 2002).

El ARNt encargado de incorporar la selenocisteína es muy particular y es la molécula clave que regula la expresión de todas las selenoproteínas. El gen del ARNt de la selenocisteína es de copia única y su secuencia primaria es de 87 nucleótidos. De modo postranscripcional se le añaden los nucleótidos CCA a su extremo 3', obteniéndose una secuencia final de 90 nucleótidos, la más larga de todos los ARNt eucariotas conocidos. Su transcripción también es peculiar puesto que empieza en el primer nucleótido de la región codificadora, con lo que éste conserva su grupo trifosfato. La selenocisteína que se incorpora a las proteínas en respuesta al codón UGA, se sintetiza a partir de una serina en el propio ARNt que la portará al ribosoma.

Más recientemente, se ha publicado el descubrimiento del aminoácido número 22 codificado por el código genético. Se trata de la pirrolisina (Pyl) que está presente en el sitio activo de una enzima necesaria para la producción de metano de algunas arqueas metanógenas (Srinivasan et al., 2002). Estas arqueas tienen un ARNt que carga con Pyl y cuyo anticodón corresponde al codón UAG, que normalmente también es de parada. En este caso, la interpretación del co-

8. En el genoma humano se han anotado 25 selenoproteínas (Kryukov et al., 2003).

dón UAG como Pyl respondería a un mecanismo distinto al utilizado para asignar UGA a Sec (Zhang et al, 2005) y en algunos casos podría incluso responder a una reasignación permanente.

Es interesante hacer notar que la presencia de codones duales supone un nuevo reto a la tarea bioinformática de rastrear genomas para predecir genes.

4.6. Regulación de la expresión génica

El ADN de una célula contiene información para miles de genes, pero no todos se expresan simultáneamente. En los seres unicelulares, los genes se expresan según las necesidades metabólicas de cada momento. En los seres pluricelulares, cada célula del organismo contiene todos los genes necesarios para la supervivencia del organismo completo. Un conjunto de estos genes, los **genes de mantenimiento** (en inglés, *housekeeping genes*), son necesarios para el mantenimiento de la célula y se expresan en todas ellas.[9] El resto de genes se expresan diferencialmente en cada tipo celular, según la función para la que se ha especializado durante el desarrollo.

La regulación de la expresión génica controla la expresión diferencial de los genes según el momento del desarrollo, las condiciones ambientales o el tipo celular. Esta regulación no consiste en una simple activación o desactivación de ciertos genes, sino que consigue controlar el nivel de expresión de un modo fino, según las necesidades de cada célula en cada momento. Esta precisión en la regulación se consigue gracias a la acción coordinada de diversos mecanismos que pueden intervenir a muy diferentes niveles de la síntesis proteica.

La síntesis proteica es un proceso complejo en el que se distinguen diversos pasos y donde intervienen multitud de proteínas. Cualquier alteración en la actividad de estas proteínas puede alterar la expresión génica y puede utilizarse en el proceso de regulación.

9. Algunos genes de la célula se expresan continuamente a unos niveles similares a lo largo del tiempo, son los **genes constitutivos**. La mayoría de genes de mantenimiento son también genes constitutivos.

Los mecanismos de regulación de la expresión génica pueden actuar:

- Antes de la transcripción alterando la estructura de la cromatina.
- En la transcripción
- En el procesamiento del ARN
- Variando la estabilidad del ARNm
- En la traducción
- En la introducción de modificaciones postraduccionales

El inicio de la transcripción requiere que el ADN a transcribir sea accesible y una maquinaria de transcripción compleja. Diversas proteínas intervienen en la modificación de la estructura de la cromatina para permitir la transcripción de determinadas regiones del ADN. Por otro lado, intervienen diversos factores de transcripción uniéndose al ADN a la vez que la polimerasa se asocia a diversas proteínas accesorias.

En cuanto al procesamiento alternativo confiere plasticidad en la expresión obteniendo, a partir de una única unidad de transcripción, proteínas distintas según la necesidad de la célula en cada momento. Un ejemplo notable de regulación por ayuste alternativo es el de la determinación del sexo en *Drosophila melanogaster*. En Drosophila, el sexo viene determinado por la proporción de cromosomas X y autosomas (A). Esta proporción provoca una cascada de regulación génica que acaba en el ayuste alternativo del pre-ARNm del gen *doublesex* (*dsx*). Así, el ambiente genotípico (relación X:A) determina la obtención de una u otra proteína por ayuste alternativo. Los individuos con una relación X:A = 1 producen una proteína Dsx^F específica de hembra, mientras que los individuos con una relación X:A = 0,5 producen un proteína Dsx^M específica de machos. El tipo de proteína Dsx que presenta un individuo determina que se desarrollen características de uno u otro sexo.

La cantidad de proteína que puede sintetizar una célula en un momento determinado depende de la cantidad de ARNm de que dispone. La cantidad de ARNm en la célula depende tanto de la velocidad de transcripción como de su estabilidad. Hay una gran variación en la estabilidad del ARN, pudiendo ser de horas, días o incluso meses. La estabilidad del ARNm se debe en gran parte a las proteínas que tiene unidas a su cola de poli(A) que le protegen de la degradación.

En la traducción intervienen multitud de proteínas y otras moléculas que pueden regular la expresión. En el caso de las selenoproteínas, vimos que el

ARNt transportador de Sec podía regular su expresión (la ausencia de esta molécula implica una parada prematura en la traducción).

Finalmente, variaciones en las enzimas que intervienen en las modificaciones postraduccionales pueden determinar que una proteína sea funcional o no.

4.6.1. Modificaciones en la estructura de la cromatina

El ADN de los eucariotas está asociado a proteínas histonas que lo empaquetan, formando la cromatina. Las histonas se agrupan en octámeros (2 moléculas de cada una de las 4 histonas: H3, H4, H2A y H2B), alrededor de los cuales el ADN se enrolla, formando los nucleosomas. Otra histona (la H1), interacciona con distintos nucleosomas consiguiendo un nivel mayor de compactación. Esta estructura compactada no permite la expresión de los genes por lo que antes de la transcripción debe modificarse.

Los extremos amino de las histonas sobresalen del octámero y se conocen como colas de las histonas. Algunos residuos de estas colas pueden modificarse covalentemente de modo reversible.[10] Las distintas modificaciones en una o más colas, actuando de un modo secuencial o simultáneo, constituyen un **código de histonas** que puede ser leído por otras proteínas que, a su vez, desencadenarán otros sucesos posteriores (Strahl y Allis, 2000). La adición de un grupo acetilo a determinados residuos de las histonas puede alterar su asociación con el ADN haciendo más probable que el nucleosoma cambie de posición. La acetilación de las histonas también influye en la unión de proteínas reguladoras al ADN.

Otra modificación de la estructura de la cromatina, relacionada con el control de la transcripción, es la metilación de las citosinas. Esta metilación es más frecuente en las citosinas adyacentes a una guanina en la misma cadena (CpG). Cerca de los inicios de transcripción se encuentran zonas con muchas secuencias CpG que denominamos islas CpG. La metilación CpG se asocia a la represión de los genes a largo plazo. Antes de poder empezar la transcripción deben eliminarse los grupos metilo.

10. Se conocen más de 150 modificaciones distintas de las histonas. La combinación de estas modificaciones puede encerrar una gran cantidad de información. De ahí que se hable de un código de las histonas.

4.6.2. Factores de transcripción

Los factores de transcripción o TF (del inglés, *transcription factor*) son proteínas que se unen específicamente a secuencias cortas (de unos 13 pb de media) del ADN denominadas sitios de unión a los factores de transcripción o TFBS (del inglés, *transcription factor binding site*). Los TFBS[11] se localizan cerca del gen diana, a menudo en el promotor o en el intensificador. La unión TF-TFBS controla la transcripción bien sea promoviéndola o bloqueándola.

Los factores generales de transcripción o GTF (del inglés, *general transcription factors*) son imprescindibles para el inicio de la transcripción. Los GTF se unen a todos los promotores utilizados por la ARN polimerasa II y, además, la mayoría interactúa directamente con la ARN polimerasa.

Gran parte de la complejidad en la regulación génica de los eucariotas se consigue por interacciones combinatorias entre TFs. Cuando múltiples TFs interactúan, se unen cooperativamente a una región del ADN conocida como módulo de regulación o CRM (del inglés, *cis-regulatory modules*). Un CRM es una secuencia del genoma de unos cientos de pb que contiene múltiples sitios de unión para diversos TFs. Un gen puede mostrar diversos patrones de expresión, temporal o espacial, regulados por distintos CRMs. Los distintos CRMs de un mismo gen o de genes distintos pueden variar considerablemente. Pueden diferir en el número de TFBS para un mismo o distintos TFs y en su posición relativa. Por ello, se habla de un **segundo código del genoma** que, localizado en el ADN *no codificador*, portaría información para la regulación génica.

La predicción de TFBS en el genoma es difícil, ya que la corta longitud de su secuencia puede ocasionar muchas predicciones falsas. La tendencia de los TFBS a presentarse agrupados en el genoma facilita enormemente la identificación de CRMs (Berman et al., 2002). A partir de éstos, la determinación de los TFBS individuales, puede realizarse de un modo mucho más preciso.

Los CRM forman los promotores y los intensificadores. Un intensificador típico tiene unos 500 pb y unos 10 sitios de unión a distintos factores de transcripción. Los intensificadores son capaces de afectar la transcripción de genes muy alejados. Además, la posición relativa del intensificador puede ser muy variable, aguas arriba, aguas abajo o incluso interna al gen, en un intrón. Esto es posible debido a que el ADN entre promotor e intensificador puede formar un

11. TRANSFAC (Matys et a., 2006) es una base de datos que recoge las secuencias de todos los TFBS determinados experimentalmente.

bucle que aproxima estas zonas facilitando la interacción entre las proteínas que tienen unidas. Del mismo modo, un intensificador puede influir sobre cualquier promotor cercano y su efecto está limitado por la presencia de un **aislador**.

4.6.3. Compensación de dosis

Las hembras tienen dos cromosomas X mientras que los machos solo tienen uno. En consecuencia, las hembras podrían generar, para los genes del cromosoma X, el doble de producto génico (proteína) que los machos. La concentración relativa de proteínas es crucial durante el desarrollo y estas diferencias entre machos y hembras podría ocasionar graves problemas. Por ello, los distintos organismos han desarrollado diversos mecanismos de compensación de dosis.

En Drosophila, los genes del cromosoma X de los machos se expresan aproximadamente el doble que en las hembras, para igualar sus niveles de expresión. Un complejo ARN-proteína en el que destaca una histona acetiltransferasa se une a lo largo de todo el cromosoma X de los machos, facilitando la activación de la cromatina.

En *Caenorhabditis elegans* los individuos XX son hermafroditas y los X0 son machos. En este caso, la compensación se consigue por la disminución de la expresión de los dos cromosomas X en los individuos hermafroditas. Esta expresión disminuida está dirigida por cambios estructurales en los cromosomas X de los hermafroditas.

En los mamíferos, las hembras inactivan los genes de uno de sus cromosomas X completo, en el que la cromatina adopta una estructura muy condensada llamada heterocromatina. La inactivación tiene lugar en los primeros momentos del desarrollo y cada célula inactiva un cromosoma X al azar. A partir de este momento, las células que descienden de una célula determinada mantiene inactivado el mismo cromosoma X. De este modo podemos observar fenotipos mosaico como el de los gatos carey. El color del pelaje de los gatos viene determinado por un gen ligado al cromosoma X. Este gen presenta dos variantes (alelos) uno para color naranja y otro para color negro. En las hembras heterocigotas para estos dos alelos se observa un pelaje a parches de colores naranja y negro. Esto se debe a que, en un estadio temprano del desarrollo, cada célula inactivó un cromosoma X aleatoriamente. Cada una de estas células originó un clon con el mismo cromosoma inactivado. Clones contiguos, con distinto cromosoma X inactivado, originan el fenotipo en parches característico de las gatas carey.

La inactivación del cromosoma X se produce por la acción del gen *Xist* (*X inactive-specific transcript*, transcrito específico del cromosoma X inactivado), localizado en el cromosoma X. El gen *Xist* es uno de los pocos que se expresan en el cromosoma inactivado. Su producto es un ARN que cubre al cromosoma X e induce su inactivación. Este ARN participa en la formación y extensión de la heterocromatina del cromosoma X. La heterocromatina del cromosoma X se caracteriza por tener el ADN hipermetilado y una serie de modificaciones en sus histonas. Entre estas modificaciones destacan la presencia de una variante específica de la histona H2A, la metilación de la histona H3 en la lisina 9 y la hipoacetilación de las histonas H3 y H4. Estas modificaciones determinan que los genes de este cromosoma no se expresen y además constituyen marcas epigenéticas que determinan la herencia de dichas modificaciones.

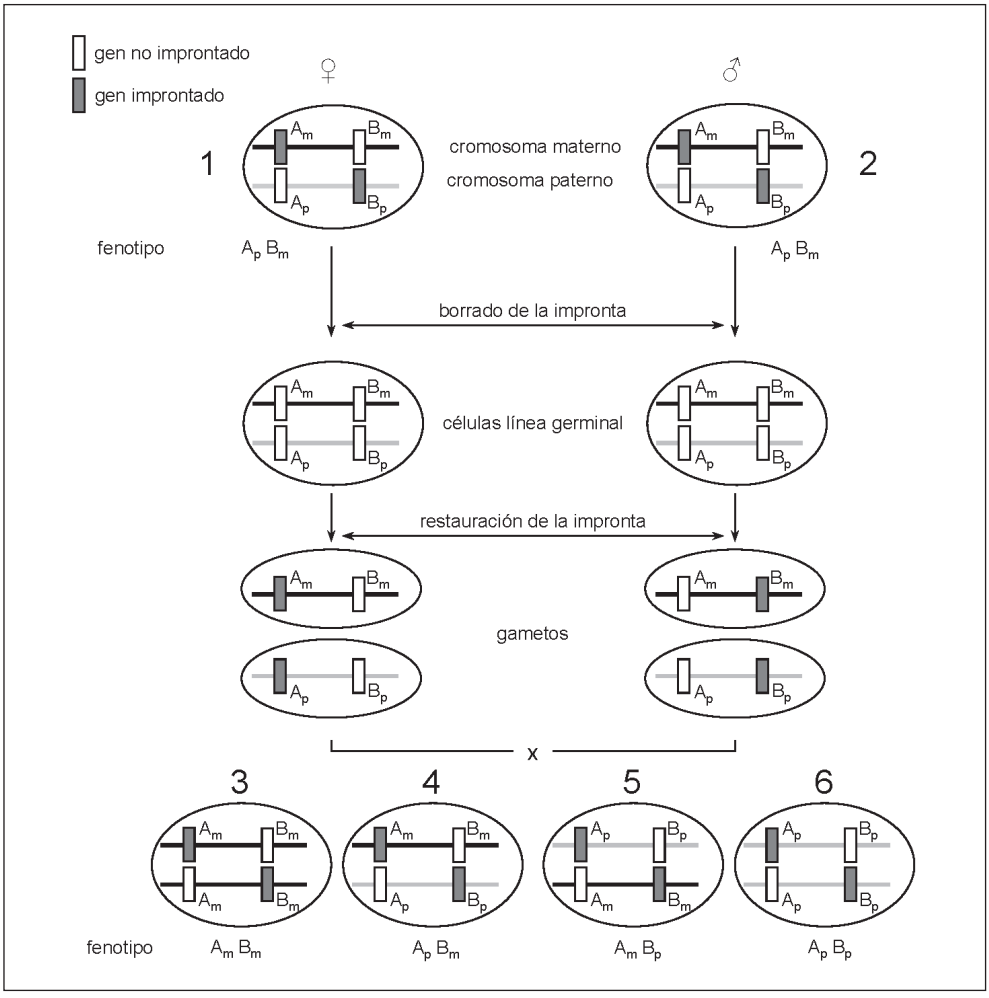
4.6.4. Impronta genética

En los organismos diploides, cada gen autosómico cuenta con dos copias que se expresan simultáneamente. No obstante, en los mamíferos se observa que una pequeña proporción de genes manifiestan impronta genética: sólo expresan una de las dos copias, dependiendo de su origen parental. Estos genes, por tanto, se comportan funcionalmente como si fueran haploides. Se ha visto que los genes improntados forman grupos en diversas regiones del genoma.

Los genes improntados responden a herencia epigenética, resultado de la metilación diferencial del ADN en cada sexo durante la formación de los gametos. Las marcas de metilación del ADN son heredadas de una generación celular a la siguiente pero son borradas en las células que originarán las células germinales. Una vez formados los gametos, los genes improntados reciben las marcas específicas de cada sexo, que decidirá si el gen será activo o no en la siguiente generación (Figura 4.9).

El fenómeno de impronta se ha estudiado extensamente en los genes *Igf2* y *H19* que forman parte de un grupo de genes improntados en el cromosoma 7 del ratón. El gen *Igf2* es un ejemplo de impronta materna (la copia derivada de la madre está inactiva). Por su parte el gen *H19* presenta impronta paterna (la copia de origen paterno permanece inactiva). Una región del ADN localizada entre los dos genes y muy cercana al gen *H19* está metilada en los gametos de los machos pero no en los de las hembras. Esta región se conoce como región de control de la impronta o ICR (del inglés, *imprinting control region*). En las células

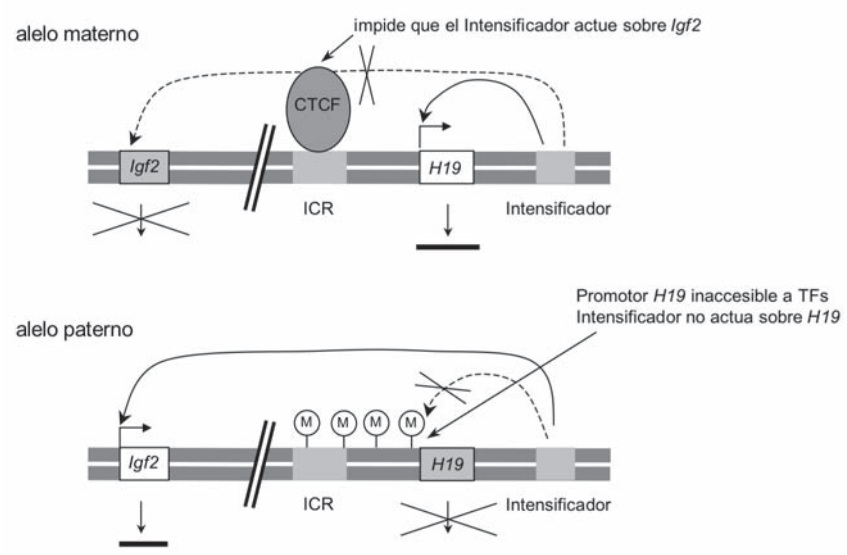
Figura 4.9. Herencia de genes improntados. Los individuos 1 y 2 comparten genotipo pero improntan diferencialmente sus gametos. Los individuos 4 y 5 comparten genotipo pero, debido a la impronta, no comparten fenotipo



somáticas, cuando esta región no está metilada, se une a la proteína CTCF que actúa como aislador. En el cromosoma de origen materno (ICR no metilado), CTCF puede unirse al ADN y actúa de aislador de un intensificador. Con ello, el intensificador sólo puede acceder al promotor de *H19* (que se expresa) pero no al promotor de *Igf2* (que entonces no se expresa). En el cromosoma de origen paterno (ICR metilado), la proteína CTCF no puede unirse y el intensificador puede actuar sobre el promotor de *Igf2* (ahora se expresa). En cambio, el intensifica-

dor ahora no puede activar el promotor de *H19*, puesto que la región metilada se extiende hasta el promotor impidiendo la unión de otros factores de transcripción (Figura 4.10).

Figura 4.10. Mecanismo de impronta genética de los genes *Igf2* y *H19* de ratón



4.6.5. Interferencia del ARN

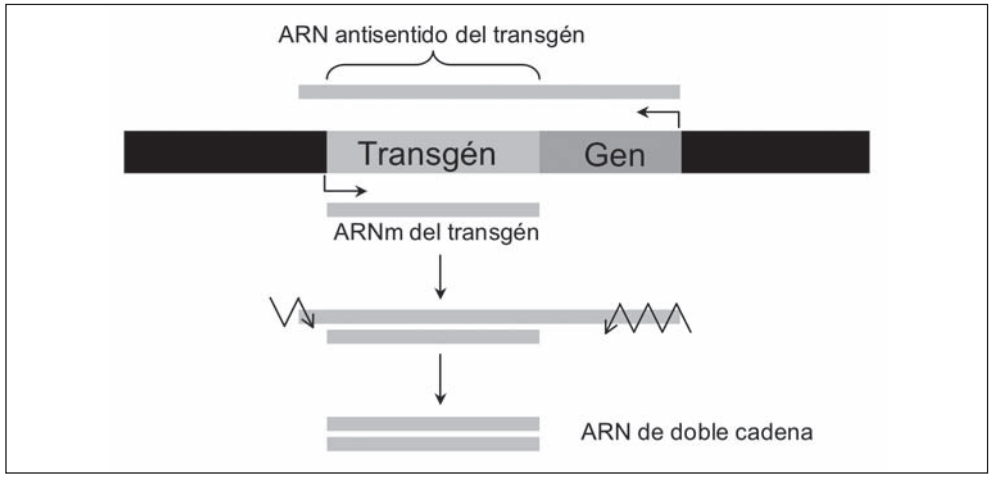
La expresión de algunos genes puede estar silenciada por la acción del ARN de interferencia. Los ARN de interferencia son pequeñas moléculas de ARN de unos 20 nucleótidos que, al aparearse con su secuencia complementaria en una molécula de ARNm desencadena la degradación de ésta.

La primera observación del efecto de los ARN de interferencia se obtuvo tras inyectar ARN de doble cadena en *Caenorhabditis elegans* (Fire et al., 1998). Se conocía que la introducción de ARN en una célula podía interferir la función normal de alguno de sus genes. En este experimento se midió el efecto de introducir por separado ARN sintetizado a partir de cada una de las dos cadenas de ADN (sentido o sin sentido), o una mezcla de ambas. Para sorpresa de los investigadores encontraron que el ARN de cadena doble tenía un mayor efecto que cualquiera de las dos cadenas individuales. Esta primera observación atrajo el

interés de muchos investigadores y en los últimos años se han realizado muchos avances en la comprensión del fenómeno.

Los ARN de interferencia se forman a partir de ARN de cadena doble que puede tener diversos orígenes. A menudo se distingue entre miRNA (del inglés, *micro RNA*) y siRNA (del inglés, *small interfering RNA*) según su origen y modo de acción. Los miRNA se originan a partir de secuencias endógenas de ADN con repeticiones invertidas, mientras que los siRNA provienen de secuencias nucleotídicas exógenas. Las secuencias exógenas que originan siRNA pueden ser de ADN, como el caso de los transgenes (Figura 4.11), pero también pueden ser de ARN como las provenientes de una infección vírica. Los miRNA actúan sobre genes distintos de los que han sido transcritos, mientras que los siRNA actúan sobre los mismos genes a partir de los cuales se transcriben.

Figura 4.11. Obtención de ARN de cadena doble a partir de un transgén

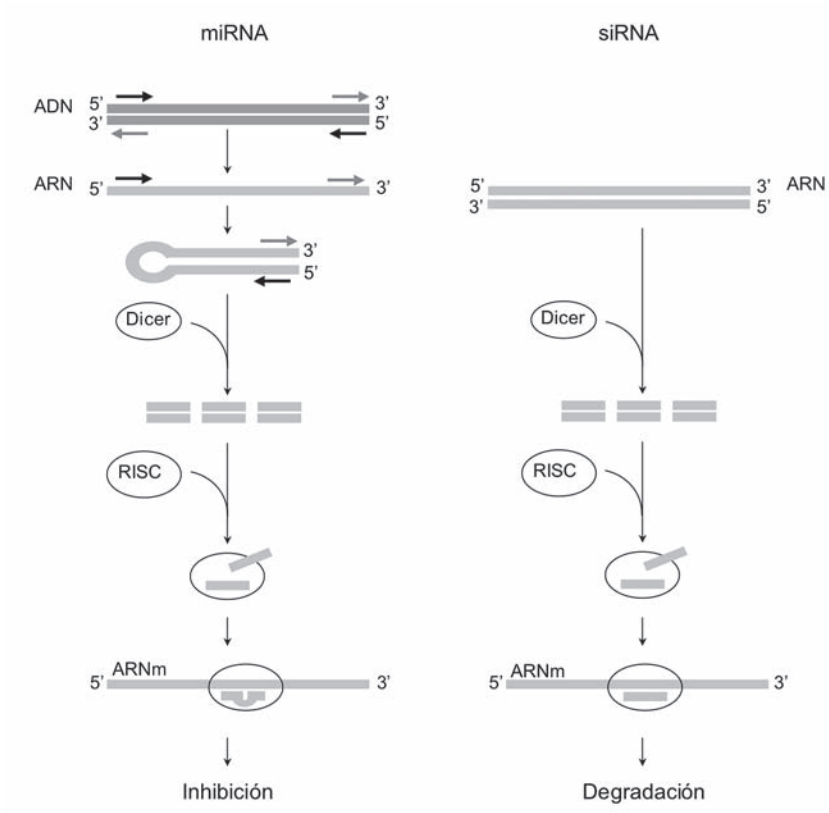


Si un transgén se inserta cerca de un gen y en dirección opuesta, el gen endógeno puede transcribir también el ADN del transgén. Cuando el ARN antisentido del transgén hibrida con el ARNm con sentido del transgén se forma el ARN de cadena doble.

El ARN de cadena doble es cortado por la proteína ribonucleasa Dicer (en castellano, *cortadora*) en moléculas de unos 20 pares de bases, típicas de los ARN de interferencia. Estas moléculas se unen a un complejo ribonucleasa denominado RISC (del inglés, *RNA-induced silencing complex*) que separa las dos cadenas del ARN interferente. La cadena antisentido, unida a RISC puede hibridar con un ARNm cuya secuencia sea complementaria. Esta unión del ARN de interferencia a un ARNm dirige al complejo RISC para que silencie específicamente

dicho ARNm. Parece que los siRNA tendrían una secuencia complementaria exacta a la secuencia diana y el complejo RISC degradaría estos ARNm, mientras que los miRNA podrían hibridar con secuencias no totalmente idénticas y se limitarían a impedir su traducción (Figura 4.12).

Figura 4.12. Diferencias entre los distintos tipos de ARN de interferencia



Los ARN de interferencia, y la maquinaria que los acompaña, se han encontrado en protozoos, hongos, plantas y animales, mostrando que es un mecanismo muy conservado en la evolución (Zamore, 2002). Esta conservación indica que su función es muy importante para los organismos. Las especulaciones sobre cual es realmente la función que ha mantenido a este mecanismo a lo largo de la evolución apuntan en dos sentidos: 1) la defensa viral y 2) la prevención a la expansión excesiva de transposones en el genoma.

Se está investigando intensamente en este nuevo aspecto de la regulación genética y se pronostica que, en un futuro próximo, el uso de los ARN de interferencia será habitual en diversas aplicaciones. Por ejemplo, podrán utilizarse para descubrir la función de un gen del cual sólo conocemos su secuencia o como tratamiento en enfermedades humanas. En los últimos años los ARN de interferencia han ido aportando nuevas sorpresas a los investigadores. Parece que en circunstancias especiales, esos pequeños ARN podrían incluso aumentar la traducción de su ARNm diana (Buchan y Parker, 2007).

4.7. El mundo de ARN

En 1958, Crick definió el **dogma central de la biología molecular** que, tal como afirma él mismo, era una idea aceptada por muchos biólogos, aunque no había sido expuesta explícitamente:

> *«El Dogma Central*
>
> Éste afirma que una vez que la «información» se ha transferido a proteína no puede transmitirse de nuevo. Con mayor detalle, la transferencia de la información es posible de ácido nucleico a ácido nucleico o de ácido nucleico a proteína, pero la transferencia de proteína a proteína o de proteína a ácido nucleico no es posible. Aquí, información significa la determinación precisa de la secuencia, bien sea de bases en el ácido nucleico o de residuos aminoácido en la proteína.» Crick (1958).

La definición del dogma central que dan los libros de texto de biología afirma que «la información genética pasa de ADN a proteína en una ruta unidireccional». En esta ruta se reconocen 3 posibles pasos de información:

– la replicación (ADN –> ADN),
– la transcripción (ADN –> ARN)
– la traducción (ARN –> proteína).

Del dogma central se desprende también la idea que el ADN contiene las instrucciones y que las proteínas son las herramientas capaces de realizar las funciones enzimáticas, mientras que el ARN es un mero intermediario. Esta visión, no obstante, plantea una paradoja. Las proteínas necesitan el ADN que

almacena la información necesaria para su síntesis, pero el ADN también necesita proteínas que son la maquinaria que realiza su replicación. Así, ¿qué fue primero, el ADN o las proteínas? ¿cómo pudo aparecer un tipo de molécula sin el otro?

Poco a poco, un mayor conocimiento del ARN nos ha empezado a desvelar la respuesta. Ahora se sabe que el ARN de algunos virus puede replicarse y en 1970 se descubrió una retrotranscriptasa inversa, que sintetiza ADN a partir de ARN (Baltimore, 1970). Esto convierte al ARN en una molécula capaz de almacenar información que puede transmitir a otras moléculas. Por otro lado, a principio de los años 80, dos grupos de investigadores, trabajando independientemente, descubrieron las primeras ribozimas, moléculas de ARN que son capaces de catalizar distintas reacciones (Kruger et al., 1982; Guerrier-Takada et al., 1982).

Así, el ARN ha pasado de ser considerado una mera comparsa a representar el papel protagonista.[12] En el ARN, descubrimos ahora funciones que anteriormente eran únicamente atribuidas al ADN (almacenar la información) o a las proteínas (catálisis de ciertas reacciones). Esto ha permitido especular en un mundo de ARN, un estado hipotético en el origen de la vida sobre la Tierra. En esta etapa las proteínas aun no intervendrían en las reacciones bioquímicas y el ARN supliría ambas funciones, almacenar la información y catalizar las reacciones.

12. La capacidad del ARN de replicarse y retrotranscribirse puede parecer que desmonta el dogma central. No obstante, es totalmente compatible con la definición original de Crick y, en todo caso, basta con redefinirlo de un modo más general.

Capítulo V

De genes, Genética...

El nacimiento de un bebé es una fuente inagotable de preguntas, sobre todo en los padres primerizos. ¿Crecerá fuerte y sano? ¿Sabré educarlo? ¿Ahora, por qué llora? ... Pero cuando llegan las visitas, la pregunta estrella es mucho más simple ¿Se parece al padre o a la madre? La experiencia nos dice que, aunque cada individuo es único y diferente, los hijos se parecen a los padres. Y, por ello, se emplea un buen rato en observar la carita redonda y sonrosada del bebé cuyos rasgos aun están por definir, intentado apreciar parecidos de una u otra familia.

El parecido entre padres e hijos se debe a la transmisión de caracteres heredables de una generación a la siguiente. Estos caracteres heredables pueden ser de diversa índole: aspecto físico, características fisiológicas, comportamiento… Estas características se transmiten a través de los genes. Pero ¿qué es un gen?

El gen puede definirse de distintos modos, según el nivel de estudio que estemos abordando. Para la comprensión de los conceptos expuestos en este capítulo basta con una definición algo abstracta pero sencilla de gen como: *cualquier factor que determina una característica particular y heredable de un organismo*. Al final del capítulo, no obstante, abordaremos la definición de gen como ente físico. Veremos que, en este sentido, la definición de gen ha ido cambiando a medida que hemos ido conociendo cómo se estructura la información genética en el genoma (Gerstein et al., 2007).

La genética estudia los genes, sus patrones de transmisión y cómo las poblaciones cambian su composición genética con el tiempo y el espacio. La genética es una ciencia relativamente joven. Podemos decir que nació en 1900 cuando fueron redescubiertos los experimentos publicados por Mendel en 1865. Pero la humanidad ha utilizado, más o menos conscientemente, los principios básicos de la genética desde hace miles de años.

El uso de la genética por el hombre se remonta a los orígenes de la agricultura (hace unos 12.000-15.000 años) y la domesticación de plantas y animales. En cada especie, el hombre seleccionó entre las variantes disponibles, aquellas que mejor se adaptaban a sus necesidades. Sin saberlo, estaba haciendo uso de la genética, favoreciendo la perpetuación de una serie de características y no otras. En

el Talmud, escritos del siglo II que recogen la tradición oral de los judíos, se encuentra la prohibición de practicar la circuncisión a bebés cuyos hermanos mayores (o primos hermanos, hijos de hermanas de su madre) hubieran muerto desangrados. Esto indica que eran conscientes que la aparición de estas hemorragias anómalas (la hemofilia[13]) tenía un origen hereditario que se transmitía vía materna. Pero el estudio sistemático de la herencia de caracteres no empieza hasta finales del siglo XIX con los trabajos de Gregor Mendel que llevaron a la formulación de las famosas Leyes de Mendel.

La herencia mendeliana es el patrón de herencia biológica más simple. Pero diversos fenómenos como el ligamiento, la interacción génica o la herencia aditiva pueden complicar mucho los patrones de herencia.

5.1. Genética clásica. Leyes de Mendel

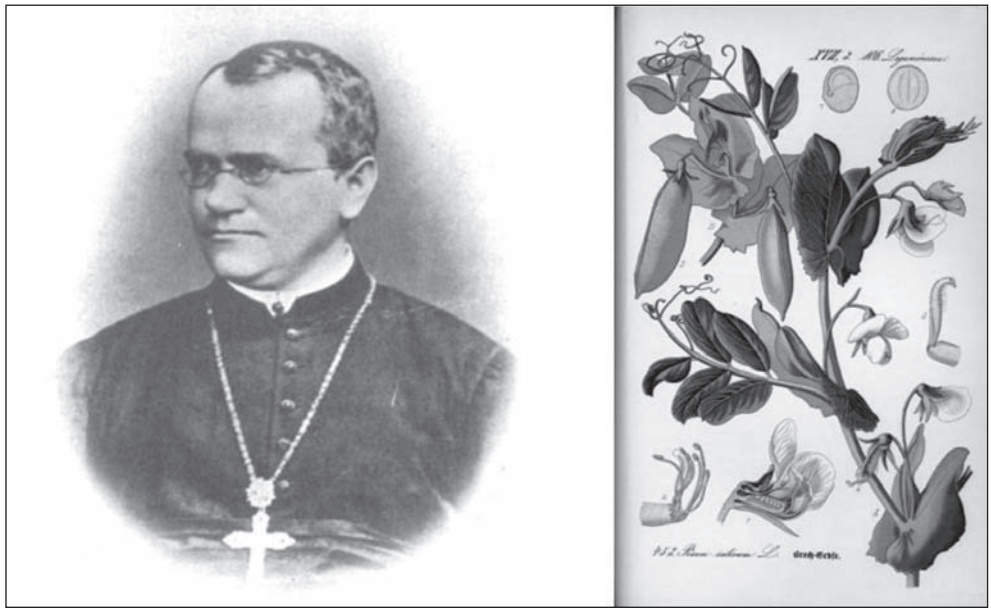
Gregor Mendel (1821-1884) es considerado el padre de la Genética debido a sus experimentos pioneros hibridando plantas de guisantes. Sus experimentos, publicados en alemán en 1865, pasaron inadvertidos hasta 1900 cuando fueron redescubiertos por Hugo de Vries, Carl Corrents y Erich von Tschermak. Las observaciones que realizó llevaron posteriormente a la formulación de las que se conocen como Leyes de Mendel.

Mendel eligió al guisante, *Pisum sativum*, como organismo modelo para realizar sus experimentos porque reunía una serie de características especiales: presenta variedades con características permanentes que son fácilmente reconocibles; aunque sus flores se autofecundan, los experimentos de fecundación artificial acostumbran a tener éxito; se cultivan fácilmente y el período de crecimiento es relativamente corto (Mendel, 1865). A esto debe añadirse la numerosa descendencia que se obtiene a partir de cada planta individual, lo que permite analizar matemáticamente las proporciones de las distintas características en la descendencia.

Los organismos modelo utilizados hoy en día en estudios genéticos, cumplen las características que ya buscó Mendel en los guisantes: existencia de variedades observables, fácil mantenimiento, tiempo de generación corto, control sobre

13. La hemofilia es una enfermedad hereditaria ligada al cromosoma X. El enfermo de hemofilia carece de un factor de coagulación de la sangre, lo que le supone que cualquier pequeña herida o golpe pueda ocasionar una grave hemorragia externa o interna.

Figura 5.1. Retrato de Gregor Mendel y lámina de la especie objeto de sus estudios *Pisum sativum*



Origen de las imágenes: Wikipedia.

los cruzamientos y numerosa descendencia. El guisante ha sido abandonado como organismo modelo porque, aunque Mendel consideró que su desarrollo era «relativamente» rápido, se trata de una planta anual. Por ello, Mendel tuvo que invertir 8 años para completar sus experimentos, algo totalmente impensable hoy en día.

Mendel partió de 34 variedades de guisantes que primero cultivó durante dos años consecutivos para comprobar que eran variedades puras (la descendencia era homogénea para las características analizadas). Eligió 7 caracteres morfológicos para los que las distintas variedades de guisantes presentaban una de dos posibles variantes: 1) forma de la semilla (lisa o rugosa), 2) color de la semilla (amarillo o verde), 3) color de las flores (violetas o blancas), 4) forma de la vaina madura (inflada o arrugada), 5) color de la vaina tierna (verde o amarilla), 6) posición de las flores en los tallos (a lo largo del tallo o en su extremo) y 7) longitud del tallo (matas altas o matas enanas).

Para cada uno de estos caracteres, realizó experimentos de hibridación cruzada entre variedades que se diferenciaban en dicho carácter. Los resultados que obtuvo para cada uno de los 7 caracteres que estudió, fueron muy similares. Por

ello y para simplificar, nos fijaremos en los resultados que obtuvo sólo en uno de ellos: la forma de la semilla (Tabla 5.1).

Tabla 5.1. Resultados obtenidos por Mendel al cruzar variedades que diferían en un único carácter del aspecto de la semilla

Generación parental (P)	1ª generación filial (F_1)	2ª generación filial (F_2)
Lisa x Rugosa	Todas lisas	5474 lisas 1850 rugosas
Amarillas x Verdes	Todas amarillas	6022 amarillas 2001 verdes

Al cruzar dos variedades (generación parental o P) que diferían en un carácter (p.e. forma de la semilla), vio que los híbridos (o generación filial F_1) presentaban siempre la misma variante de uno de los padres (semilla lisa). Cuando permitía la autopolinización de la F_1 para obtener la segunda generación filial o F_2, obtenía individuos con la variante de los híbridos F_1 (semilla lisa) y otros individuos en los que reaparecía la variante parental ausente en la F_1 (semilla rugosa). A la variante presente en la F_1 la llamó **dominante** y a la variante parental que reaparecía en la F_2 la llamó **recesiva**. La gran cantidad de descendientes obtenidos le permitió observar, además que, en la F_2 la proporción de variantes era aproximadamente de 3 dominante: 1 recesiva.

Estos resultados y observaciones se formularon posteriormente en la primera Ley de Mendel.

La Primera Ley de Mendel, o *Principio de Segregación*, establece que cada organismo diploide posee dos alelos[14] para un carácter determinado. Cuando se forman los gametos, estos alelos se separan yendo cada uno a un gameto distinto. Así mismo, el concepto de dominancia nos dice que cuando un organismo tiene dos **alelos** distintos para un carácter solo se observa el rasgo codificado por uno de ellos que es el **dominante**, quedando escondido el del otro que es el **recesivo**.

Mendel continuó sus experimentos durante algunas generaciones más de autofecundación y observó que los individuos con la variante recesiva se comportaban siempre como una raza pura, produciendo una descendencia homogénea

14. Alelo es el nombre que recibe cada una de las posibles variantes de un gen.

que mostraba dicha variante. En cambio, mientras que algunos individuos con la variante dominante también parecían una raza pura que producía descendencia homogénea, otros se parecían a los híbridos de la F_1 produciendo una descendencia heterogénea con proporciones 3:1. Estas diferencias en la descendencia de los híbridos de fenotipo dominante se explica porqué en realidad tienen genotipo distinto.

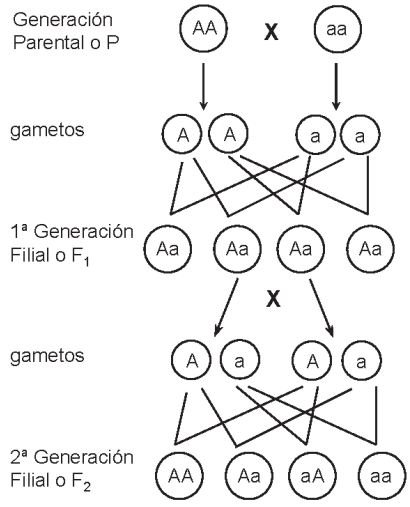
Es muy importante distinguir entre genotipo y fenotipo. El **genotipo** se refiere a la información genética presente en un individuo, es decir al conjunto de alelos que posee. El **fenotipo** se refiere al aspecto observable del individuo. Este aspecto o conjunto de caracteres que presenta es el resultado de la interacción entre su genotipo y el ambiente en que se ha desarrollado.

La Figura 5.2 muestra esquemáticamente la explicación a la Primera Ley de Mendel y su relación con la meiosis. Los genes se localizan en los cromosomas y, los individuos diploides cuentan con 2 copias de cada cromosoma (cromosomas homólogos). Así, cada individuo posee dos alelos para un gen determinado, localizados cada uno sobre un cromosoma homólogo distinto. Si al alelo dominante le llamamos *A* y al recesivo le llamamos *a*, el genotipo del híbrido de la F_1 es *Aa* y su fenotipo corresponde al del alelo dominante *A*. Como ya vimos en el capítulo III, los cromosomas homólogos se separan durante la meiosis. Así, cada gameto sólo contiene una de las dos copias del gen. Durante la fecundación se unen, al azar, un gameto de cada progenitor. De ahí las proporciones fenotípicas que observamos en la F_2, 3 individuos manifiestan la variante dominante (*A_*)[15] y 1 individuo manifiesta la variante recesiva (*aa*), que a su vez responden a las proporciones genotípicas de *1 AA : 2 Aa : 1 aa*.

Mendel fue aun más allá y realizó una serie de experimentos de dihibridismo, en los que consideraba simultáneamente distintos pares de características. Por ejemplo, cruzó una variedad que producía semillas lisas y amarillas con otra que las producía rugosas y verdes. Las semillas híbridas fueron todas lisas y amarillas, tal como esperaba por los resultados obtenidos anteriormente de los cruces monohíbridos (Tabla 5.1). En la siguiente generación, obtuvo distintos tipos

15. La notación *A_* indica que el individuo es portador de al menos un alelo dominante *A*, pudiendo ser homocigoto (*AA*) o heterocigoto (*Aa*).

Figura 5.2. Esquema que explica las observaciones de Mendel resumidas en la primera Ley de Mendel



A: alelo dominante
a: alelo recesivo

P	genotipo	AA		aa	
	fenotipo	A		a	
F_1	genotipo	Aa			
	fenotipo	A			
F_2	genotipo	AA	Aa	aA	aa
	fenotipo	A			a

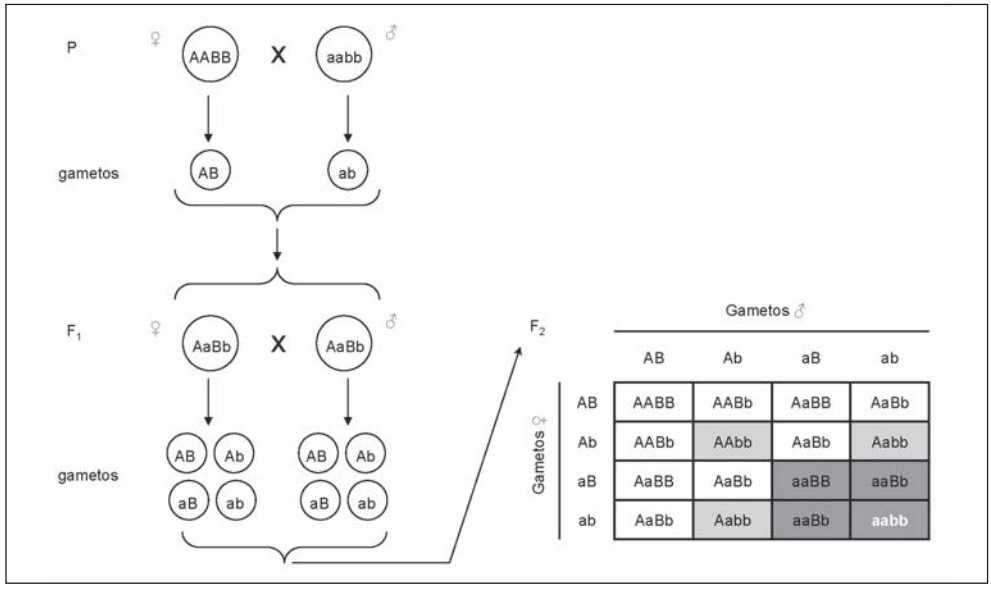
de semillas: 315 lisas amarillas, 101 rugosas amarillas; 108 lisas verdes y 32 rugosas verdes. Las variantes para los dos caracteres se entremezclan, apareciendo las 4 combinaciones posibles en unas proporciones cercanas a 9:3:3:1. Estos resultados son la base de la Segunda Ley de Mendel (Figura 5.3).

La **Segunda Ley de Mendel** o *Principio de la distribución independiente*, afirma que los alelos de **loci**[16] diferentes se separan de forma independiente.

Hasta aquí se ha considerado que un carácter tiene sólo dos posibles alelos, pero a menudo, un gen tiene alelos múltiples. La herencia de un sistema de alelos múltiples no difiere de la de un sistema de dos alelos, salvo en que existen más genotipos y fenotipos posibles. Un sistema de alelos múltiples bien conocido es el de los grupos sanguíneos AB0. Este sistema cuenta con 3 alelos: I^A, que codifica para el antígeno A; I^B, que codifica para el antígeno B; y el alelo *i* que no codifica para ningún antígeno (0). El alelo *i* es recesivo, mientras que los alelos

16. Locus (loci, en plural) se refiere al lugar que ocupa un gen determinado en el cromosoma. Por extensión también se refiere a dicho gen.

Figura 5.3. Segunda Ley de Mendel



Nótese que el cruzamiento de este ejemplo genera una F_2 en que los individuos pueden presentar 9 **genotipos** distintos pero sólo 4 **fenotipos** diferentes que esperamos encontrar en la siguiente proporción: 9/16 AB; 3/16 Ab; 3/16 aB y 1/16 ab.

I^A y I^B son codominantes entre sí. Podemos, pues, encontrar 6 genotipos distintos que determinan 4 fenotipos:

I^AI^A y I^Ai grupo sanguíneo A, I^BI^B y I^Bi grupo sanguíneo B, I^AI^B grupo sanguíneo AB y *ii* grupo sanguíneo 0.

Aunque en una población pueden existir multitud de alelos distintos para un gen cualquiera, un individuo sólo cuenta con dos alelos, que pueden ser iguales (**homocigoto**) o distintos (**heterocigoto**).

El patrón de herencia, puede ser bastante más complejo que el descrito por Mendel.

- En ocasiones, un carácter viene determinado por la interacción de diversos genes.
- En ocasiones, la manifestación de un carácter está supeditada a determinadas condiciones ambientales.

Además, debemos considerar matizaciones i/o extensiones para ambas Leyes de Mendel. Por un lado, no siempre se observa una relación de dominancia/recesividad entre alelos. En algunas ocasiones encontramos **dominancia incompleta**, observando que el heterocigoto presenta un fenotipo intermedio al de ambos homocigotos. En otras ocasiones hay **codominancia**, observándose el fenotipo de ambos homocigotos simultáneamente. Por otro lado, no todas las parejas de loci se heredan independientemente sino que puede existir **ligamiento**.

El modo de clasificar el tipo de relación de dominancia es un tanto arbitraria, pudiendo cambiar según el nivel al que estemos observando el fenotipo. Un ejemplo paradigmático de esto nos lo ofrece la anemia falciforme. La anemia falciforme es una enfermedad humana grave, determinada por el gen que codifica para la hemoglobina, la molécula responsable del transporte de oxígeno en la sangre. Existen 2 alelos Hb^A y Hb^S y la Tabla 5.2. recoge los fenotipos que determinan cada uno de los tres genotipos posibles. Si nos fijamos globalmente en el individuo, veremos que el alelo que determina la anemia falciforme es recesivo. Si centramos nuestra atención en los glóbulos rojos, observaremos que los heterocigotos presentan un fenotipo intermedio (dominancia incompleta). Si descendemos al nivel molecular veremos que los heterocigotos expresan los dos fenotipos simultáneamente, presentan los dos tipos de cadenas proteicas (codominancia).

Tabla 5.2. Distintos niveles de observación del fenotipo para el carácter anemia falciforme

Genotipo	Individuo	Glóbulos rojos	Hemoglobina
$Hb^A Hb^A$	Sano	Discoidales	Forma A
$Hb^A Hb^S$	Sano	Sólo se deforman en condiciones de poco oxígeno	Formas A y S
$Hb^S Hb^S$	Anemia	Deformados en forma de hoz	Forma S

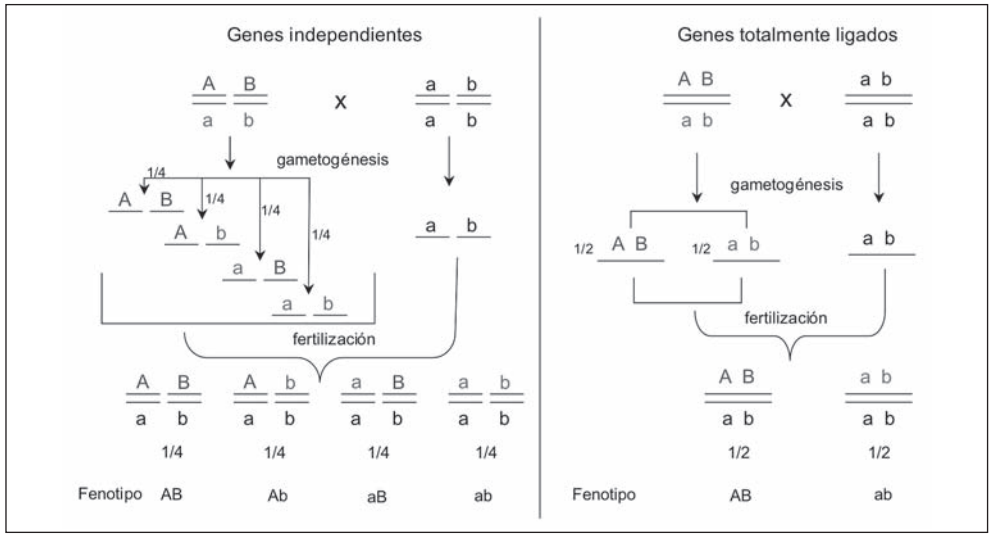
5.2. Ligamiento

Si sobre un cromosoma se encuentran numerosos genes, y los cromosomas se transmiten enteros durante la división celular, podemos concluir que dichos genes se transmiten conjuntamente: están ligados. Esta situación constituye una excepción a la segunda ley de Mendel, que, para ser más precisa debería especificar que son los genes localizados en cromosomas de distintas parejas los que se transmiten de forma independiente. De todos modos, los genes de un mismo

cromosoma también pueden separarse en la formación de los gametos debido al fenómeno de **recombinación intracromosómica**.

Podemos detectar el ligamiento mediante un **cruzamiento prueba**, cruzando un heterocigoto para ambos caracteres con el doble homocigoto recesivo. En caso de no estar ligados, esperamos encontrar 4 combinaciones fenotípicas equifrecuentes en la descendencia. En caso de caracteres totalmente ligados sólo

Figura 5.4. Resultado esperado de un cruzamiento prueba en caso de dos caracteres independientes o de dos caracteres totalmente ligados

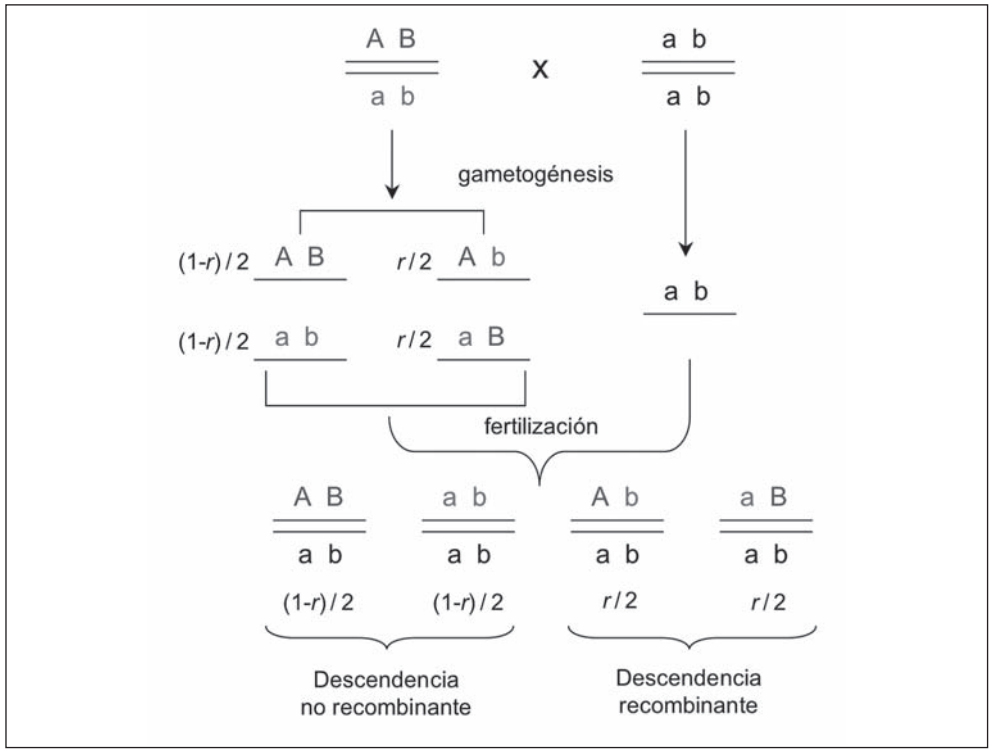


aparecerán 2 fenotipos en la descendencia (Figura 5.4).

En el caso de genes ligados entre los que hay recombinación con frecuencia *r*, encontramos 4 genotipos en la descendencia, pero estos genotipos no son equifrecuentes, siendo los más frecuentes los determinados por los gametos no recombinantes (Figura 5.5).

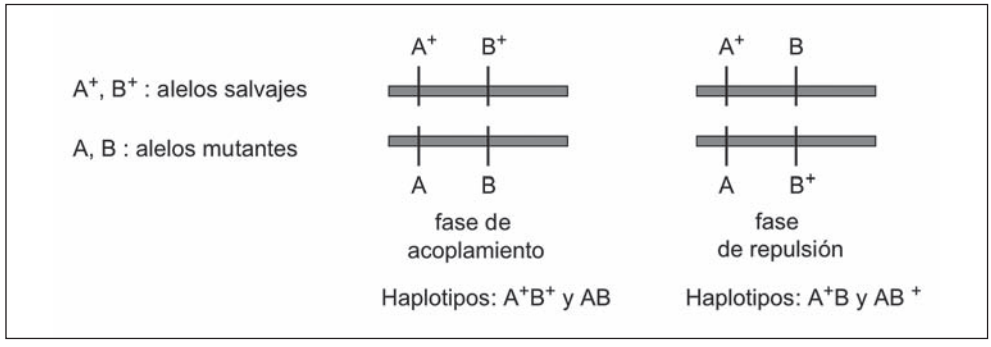
En el caso de un doble heterocigoto, es importante conocer cómo se disponen los alelos en cada cromosoma o **fase de los alelos**, que definirá el **haplotipo** del cromosoma o combinación de alelos de cada cromosoma. Cuando los dos alelos salvajes se encuentran sobre un mismo cromosoma y los mutantes sobre el otro se dice que se encuentran en fase de **acoplamiento o *cis***, en caso contrario se dice que están en fase de **repulsión o *trans*** (Figura 5.6). Dependiendo de la fase o haplotipo, la composición de la descendencia será una u otra ya que el haplotipo no recombinante se transmite con mayor frecuencia.

Figura 5.5. Resultado del cruce prueba para dos genes ligados con frecuencia de recombinación *r* entre ellos.



Cuando se secuencia el genoma de un organismo diploide, a menudo se secuencia simultáneamente la información portada por ambos cromosomas homólogos. En el caso de existir diversas posiciones heterocigotas, no es posible saber con certeza qué combinación de variantes es la portada por cada cromosoma.

Figura 5.6. Posibles fases de los alelos de un doble heterocigoto



HAPMAP (http://hapmap.ncbi.nim.nih.gov/) es un importante proyecto que consiste precisamente en determinar haplotipos en el genoma humano, es decir en determinar qué combinaciones de SNPs (acrónimo inglés para, *single nucleotide polymorphism*) están unidas, y cómo se distribuyen en distintas poblaciones del mundo. Esta información ayudará a los investigadores a encontrar asociaciones entre SNPs y ciertas enfermedades.

Un ejemplo especial de genes ligados lo constituyen los genes ligados al sexo, que son los que están en los cromosomas sexuales. El hecho de estar ligados a los cromosomas sexuales hace que la herencia de estos caracteres no sea uniforme entre los dos sexos.

5.2.1. Herencia ligada al sexo

La determinación del sexo (el que un individuo se desarrolle como hembra o como macho) se consigue por medios diversos según el tipo de organismo que consideremos. En muchos casos viene mediada por la combinación de un par de cromosomas: los **cromosomas sexuales**. A diferencia del resto de pares de cromosomas (o **autosomas**), los cromosomas sexuales no son totalmente homólogos. En Mamíferos y otros organismos estos cromosomas reciben el nombre de cromosomas X e Y. La combinación XX determina sexo femenino (♀) y la combinación XY, sexo masculino (♂). En otros organismos como Aves y Lepidópteros, los cromosomas sexuales se conocen como Z y W. En este caso, las hembras son **heterogaméticas**, presentan los dos cromosomas diferentes ZW, mientras que los machos son **homogaméticos** ZZ.

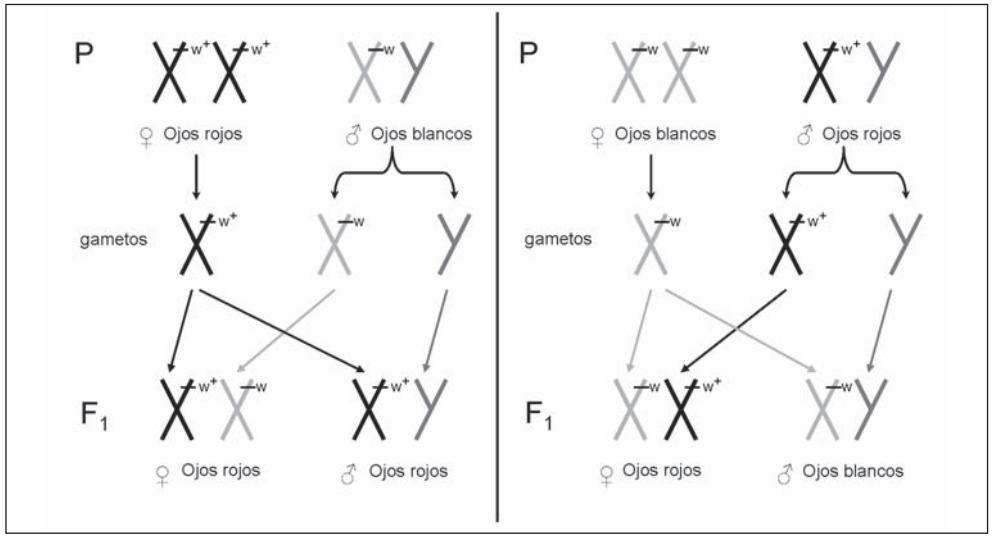
La herencia de los genes ligados a los cromosomas XY es similar a la de los genes ligados a los cromosomas ZW, cambiando sólo a qué sexo afecta. Por ello nos centraremos únicamente en el estudio de los genes ligados al sistema XY.

En general, el cromosoma Y es en gran parte **heterocromático**, o en otras palabras su cromatina está compactada impidiendo la expresión génica. Además, en comparación con otros cromosomas contiene pocos genes, algunos específicos del cromosoma Y y unos pocos compartidos con el cromosoma X. El cromosoma X contiene un mayor número de genes que el Y[17].

17. Se calcula que en el cromosoma Y de Drosophila sólo hay entre 10 y 20 genes de copia única codificadores de proteína mientras que en el cromosoma X hay más de 2500.

La herencia de los genes ligados al cromosoma X es un tanto peculiar debido a que machos y hembras tienen distinto número de copias. Mientras que las hembras cuentan con 2 copias (de origen materno y paterno) los machos sólo cuentan con 1 (de origen materno). Por ello, los cruzamientos recíprocos[18] muestran resultados distintos (Figura 5.7).

Figura 5.7. Herencia del gen *white* en *D. melanogaster*.



Esquema de cruzamientos recíprocos entre individuos de dos cepas puras de Drosophila respecto al color de sus ojos.
Nótese que:
- El fenotipo de la F1 es distinto en ambos cruzamientos.
- Cuando se cruzan hembras mutantes con machos de tipo salvaje, el fenotipo de la descendencia invierte su asociación con el sexo.

La herencia ligada al cromosoma X fue explicada por primera vez por Thomas H. Morgan[19] en 1910. Tras descubrir un macho de *D. melanogaster* con ojos blancos, lo cruzó con hembras de fenotipo salvaje, de ojos rojos. La descendencia resultó ser homogénea con ojos rojos, indicando que el alelo para ojos blancos era recesivo. La sorpresa llegó con el recuento de la F_2. Según la herencia

18. Los cruzamientos recíprocos son pares de cruzamientos en los que los machos utilizados en un cruzamiento presentan el fenotipo de las hembras del otro cruzamiento y viceversa.
19. T. Morgan recibió el Premio Nobel en 1933 por sus descubrimientos relacionados con el papel de los cromosomas en la herencia.

mendeliana esperaba que 1/4 de los individuos, indistintamente machos y hembras, tuvieran los ojos blancos. En cambio, obtuvo una descendencia en que todas las hembras y la mitad de los machos tenían los ojos rojos y sólo la mitad de los machos tenían los ojos blancos. Morgan supuso que el gen responsable del color de los ojos se encontraba en el cromosoma X. Puesto que las hembras tienen dos cromosomas X, pueden ser homocigotas o heterocigotas. En cambio, los machos al no poseer más que un solo cromosoma X se dice que son **hemicigóticos** y siempre expresan el único alelo que portan.

Morgan predijo que el cruzamiento entre hembras de ojos blancos y machos de ojos rojos generaría una descendencia en que el fenotipo de los individuos invertiría su asociación con el sexo: todas las hembras serían de ojos rojos y todos los machos tendrían ojos blancos. Efectivamente, cuando en las subsiguientes generaciones obtuvo hembras con ojos blancos comprobó experimentalmente que sus predicciones eran correctas.

5.2.2. Recombinación y mapas genéticos

En general, cuanto más alejados físicamente en el cromosoma se encuentren 2 genes, mayor será la frecuencia de recombinación (*r*) entre ellos y menos ligados estarán.

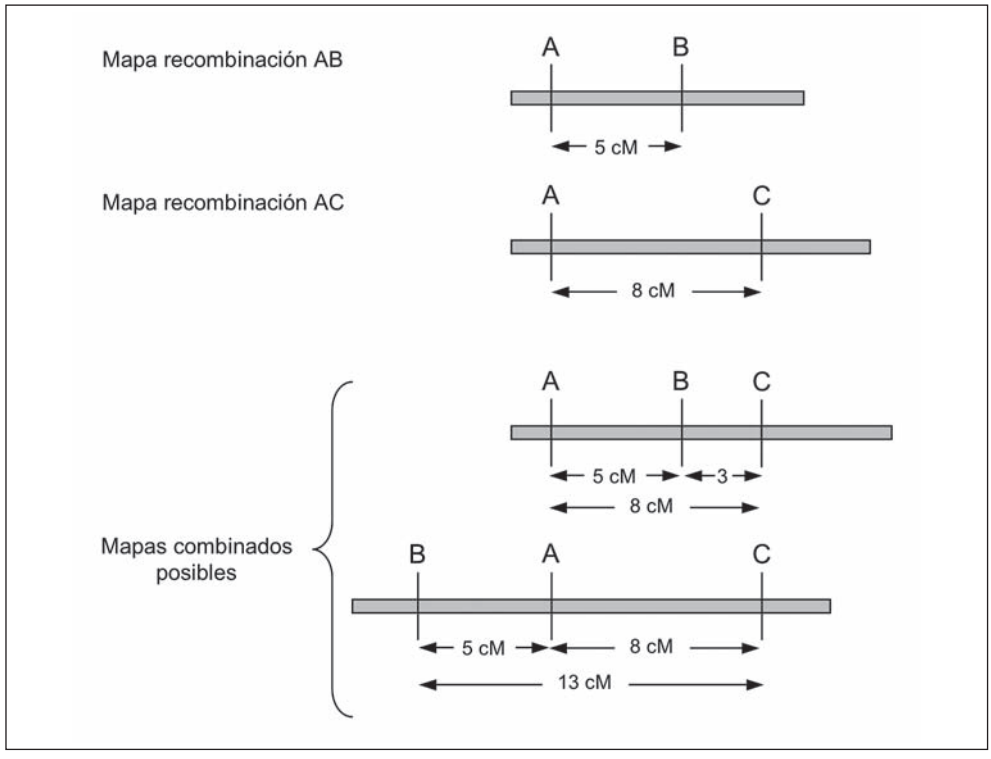
Morgan se dio cuenta que los genes ligados podían separarse y que lo hacían a frecuencias muy variadas dependiendo de los pares de genes que estuviera considerando. Partiendo de la teoría cromosómica de la herencia (los genes se localizan en los cromosomas), Morgan postuló que debía existir un mecanismo de intercambio de material cromosómico o entrecruzamiento. Además, consideró que los entrecruzamientos durante la meiosis se distribuían al azar sobre el cromosoma.

Estas ideas llevaron a Alfred Sturtevant, un joven estudiante en el laboratorio de Morgan, a pensar que se podía relacionar directamente las distancias físicas de los genes sobre el cromosoma con su frecuencia de recombinación. De este modo, dibujó el primer mapa de ligamiento para 6 genes del cromosoma X de *D.melanogaster* (Sturtevant, 1913).

En un mapa genético, las distancias entre genes se miden en centimorgans (cM). Un cM equivale a una recombinación del 1%. Las distancias genéticas

medidas a partir de la recombinación son casi aditivas y sus magnitudes nos dan pistas de las posiciones relativas de los genes. Imaginemos que tenemos tres genes ligados A, B y C. Si la distancia entre A y B es de 5 cM y la distancia entre A y C es de 8 cM, la distancia entre B y C puede ser de 3 cM o de 13 cM. En el primer caso, el gen B se localiza en una posición intermedia, mientras que en el segundo caso el gen intermedio es el A (Figura 5.8). Realizando una serie de cruzamientos dihíbridos para distintos genes podremos ir dibujando el mapa genético.

Figura 5.8. Propiedad aditiva de las distancias de mapa



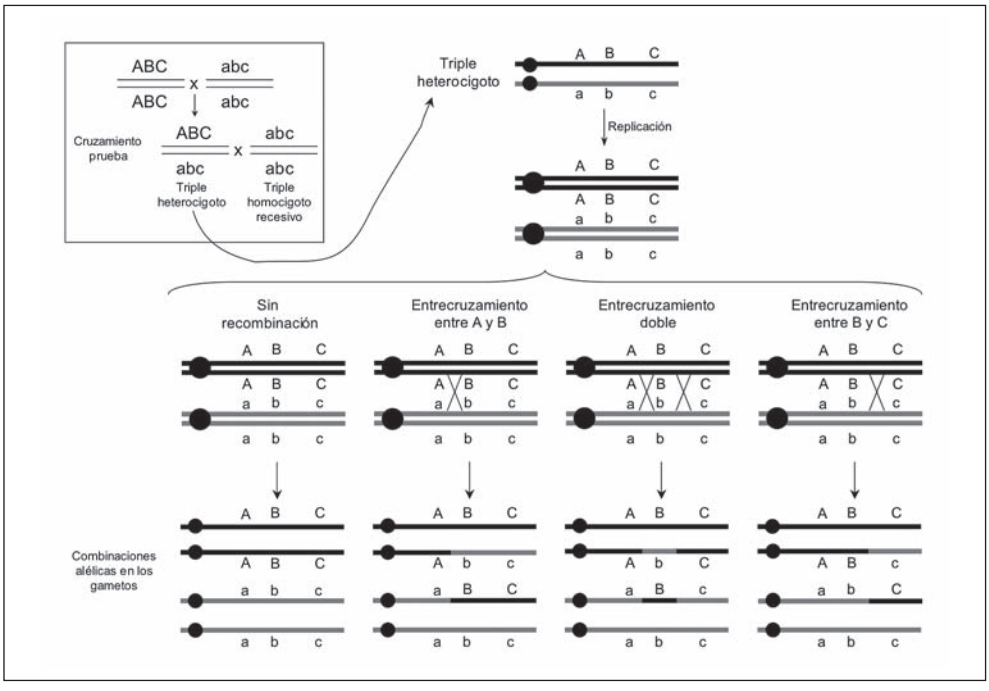
Al construir un mapa genético debe tenerse en cuenta que la recombinación máxima entre dos genes es del 50%, que es la que observamos entre genes situados en cromosomas distintos. Genes localizados en un mismo cromosoma, distantes más de 50 cM, no muestran más ligamiento que genes localizados en cromosomas distintos. Por otro lado, la recombinación que observamos entre dos genes distantes, es en realidad una subestima. Cuanto más alejados,

más probable es que exista recombinación, pudiendo existir dobles entrecruzamientos. El producto de un doble recombinante entre dos marcadores no difiere de su combinación original, lo cual conduce a una subestima de la frecuencia de recombinación.

Un modo más eficaz de reconstruir un mapa genético consiste en realizar un cruzamiento prueba de tres puntos. Un único cruzamiento nos proporciona toda la información para localizar espacialmente tres genes. Además, en un cruzamiento de este tipo podemos identificar dobles recombinantes entre los genes más distantes, que sin la información del gen central nos pasarían totalmente desapercibidos (Figura 5.9).

Este cruzamiento prueba consiste en utilizar como progenitores a un triple heterocigoto y un triple homocigoto recesivo. Aunque el homocigoto pueda experimentar entrecruzamiento en la meiosis, no apreciaremos recombinación, puesto que los marcadores usados son idénticos en ambos cromosomas. Además como siempre contribuye con un alelo recesivo, el fenotipo que observemos en la descendencia nos indicará directamente el alelo heredado del otro progenitor.

Figura 5.9. Combinaciones alélicas en los gametos de un triple heterocigoto. Entre 3 loci ligados podemos observar tres tipos de entrecruzaminetos



Esto nos permite observar los diferentes tipos de recombinantes y determinar su frecuencia. En estos cruzamientos podemos obtener 8 fenotipos distintos (correspondiendo a 8 genotipos). Los dos fenotipos más frecuentes nos indican la combinación parental original. Los dos fenotipos menos frecuentes nos indican los dobles recombinantes y nos proporcionan la pista de qué gen se localiza en el centro, que será aquel para el cual haya cambiado el alelo respecto a los de los otros dos genes. Para calcular la frecuencia de recombinación entre el gen central y uno de los marginales se deberá sumar las clases fenotípicas correspondientes a los recombiantes entre estos dos genes. Para conocer la frecuencia de recombinación entre los genes más externos debemos sumar las frecuencias de recombinación entre el gen central y cada uno de los marginales.

5.2.3. Transposones

Hasta aquí se ha considerado que la ubicación de los genes en el cromosoma es constante. Si bien existe recombinación, que permite el intercambio de material entre cromosomas homólogos, la posición relativa de este material en los cromosomas permanece inalterable. Sin embargo, existe un tipo de «gen saltarín» que es capaz de cambiar de posición dentro del cromosoma o incluso cambiar de cromosoma y que a menudo lo hace de un modo replicativo, expandiendo su número en el genoma. Estos genes son los **transposones**, elementos transponibles o TEs (del inglés, *transposable elements*).

Los transposones fueron descubiertos por Barbara McClintock[20] en 1950 cuando estudiaba la herencia de la coloración de los granos del maíz, *Zea mays*, y la composición génica de su cromosoma 9 (McClintock, 1950).

La mayoría de los granos tenían una pigmentación completa o eran amarillos. Unos pocos eran variegados, granos amarillos con manchas irregulares de pigmentación. Se pensaba que este patrón respondía a una mutación inestable. Una mutación en el gen salvaje producía los granos amarillos y la reversión de la mutación en algunas células era la responsable de los granos jaspeados.

McClintock encontró una línea de maíz en que el cromosoma 9 se rompía con mucha frecuencia por el locus *Ds* (Disociation). Además, *Ds* cambiaba frecuentemente de posición en el cromosoma sin provocar ninguna alteración visible en el cromosoma. Para que *Ds* se moviera era necesaria la presencia de otro

20. B. McClintock recibió el Premio Nobel en 1983 por su descubrimiento de los TEs.

gen no ligado *Ac* (Activator), que también cambiaba frecuentemente de posición. Vio que cuando *Ds* cambiaba su posición cerca del locus *C* (que determina la pigmentación del grano) desaparecía la acción normal del gen *C* apareciendo el fenotipo del alelo recesivo *c* (granos amarillos). Cuando *Ds* se movía de nuevo, abandonando esta posición, el fenotipo del alelo *C* quedaba totalmente restaurado. Refiriéndose a este ejemplo, McClintck apuntó:

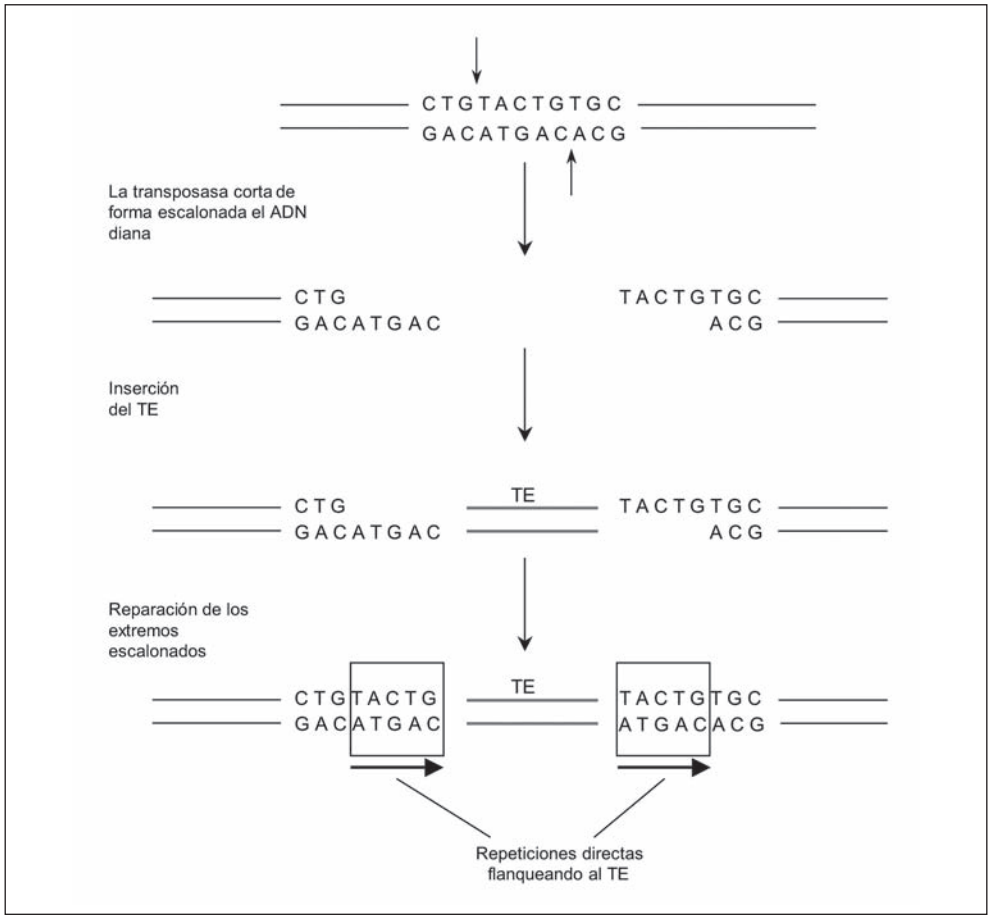
> «Un caso de transposición de *Ds* ha sido de particular importancia porque ilustra cómo pueden surgir nuevos loci mutables asociados con cambios en la expresión génica.» McClintock (1950).

Durante años, la comunidad científica pensó que la transposición se restringía al maíz y no se reconoció la importancia de los transposones hasta la década de 1970-80 en que se demostró su existencia en todos los organismos. Actualmente sabemos que representan un porcentaje importante de muchos genomas. Se calcula que el 50% del genoma humano lo constituyen secuencias móviles.

Existen muchos tipos de TEs que se clasifican de distintos modos. Algunos constan sólo de las secuencias necesarias para su propia transposición pero otros codifican, además, para otros genes. Unos son **autónomos** (no necesitan de ningún otro gen para transponerse, por ejemplo el elemento *Ac* del maíz) y otros son **no autónomos** (necesitan de otros genes para movilizarse, por ejemplo el elemento *Ds* del maíz que precisa de *Ac)*. El mecanismo de transposición puede ser **conservativo** o **replicativo**. En el conservativo, el transposón se escinde del cromosoma y se integra en una nueva localización. En el replicativo el transposón crea una copia que se integrará en un lugar nuevo, conservándose la inserción original.

Una característica común a los TEs es la presencia de repeticiones directas flanqueantes de 3-12 pb. Estas secuencias no pertenecen al TE y no se mueven con él, sino que se originan en el momento de insertarse en un lugar del genoma. Corresponden a la secuencia original dónde se produce una rotura escalonada y posterior inserción del TE con reparación del ADN cortado (Figura 5.10). Por ello, estas secuencias repetidas son muy variables, aunque su longitud es constante para cada tipo de TE. Estas repeticiones directas se mantienen una vez el transposón se escinde, quedando como huellas de su paso en esta posición del genoma. Si estas huellas han sido dejadas dentro o en las proximidades de un gen, pueden alterar la función de éste, incluso una vez ha desaparecido el TE.

Figura 5.10. Origen de las repeticiones directas que flanquean los TEs en el genoma



Los TEs de los eucariotas se clasifican en dos grupos: **retrotransposones** o TEs de clase 1 y **transposones de ADN** o TEs de clase 2.

Los retrotransposones utilizan un intermediario de ARN para movilizarse. Una transcriptasa inversa pasa el ARN transcrito a ADN que se insertará en un nuevo lugar. Contienen el gen de la transcriptasa inversa, y en ocasiones otros. Una vez que el ARNm llega al citoplasma, se traduce la transcriptasa inversa que se encarga a continuación de obtener una copia del ARNm a ADN de cadena doble. Este ADN volverá al núcleo donde podrá insertarse en un nuevo lugar del genoma. Los retrotransposones pueden ser de dos tipos, los que presentan repeticiones terminales largas o **LTR** (del inglés, *long terminal repeats*) y los que no las presentan o **non-LTR**. Los elementos LINE1 o L1 y Alu son retrotransposones

non-LTR presentes en gran cantidad en el genoma humano. Las secuencias Alu, de unos 300 pb, están representadas por cerca de un millón de copias en el genoma humano. Las secuencias L1, de unas 6 kb (kilobases) de longitud, representan aproximadamente el 15% del genoma humano.

Los transposones de ADN autónomos codifican para una transposasa y en ocasiones otros genes. La transposasa es necesaria para la escisión y la inserción del transposón. Los transposones de ADN poseen repeticiones terminales invertidas cortas de 9-40 pb que son secuencias invertidas y complementarias la una de la otra. Estas secuencias son reconocidas por la transposasa. Algunos transposones de ADN son el sistema *Ds/Ac* del maíz y los **elementos *P*** de Drosophila. Los elementos *P* parecen estar presentes en todas las poblaciones naturales de *D. melanogaster* pero su expansión ha debido tener lugar en los últimos 50 años, puesto que están ausentes en todas las líneas de laboratorio originadas a partir de capturas realizadas antes de 1945.

La movilización de los TEs puede conllevar la aparición de mutaciones tal como hemos visto en el complejo *Ds/Ac* del maíz. En ocasiones estas mutaciones son mucho más graves, como la inserción del retrotransposón L1 dentro del gen que codifica para el factor de coagulación VIII que provoca la hemofilia (Kazazian et a., 1988). Así, pues, un genoma con un alto porcentaje de transposones, debería ser altamente inestable y sujeto a multitud de mutaciones deletéreas. Para evitarlo, la mayoría de los transposones se encuentran silenciados, permaneciendo inactivos por la acción de diversos mecanismos como la metilación del ADN, el remodelado de la cromatina o la intervención de miRNAs. Algunos TE pueden codificar siRNAs que controlan su propio silenciamiento. Es el caso del siRNA[21] que interfiere con L1, que proviene de la secuencia 5'UTR de L1.

La capacidad de los transposones de transportar con ellos otros genes, ha sido explotada como herramienta para obtener **organismos transgénicos**. En Drosophila, Spradling y Rubin (1982), mostraron qué es posible obtener individuos transgénicos utilizando los elementos *P*. Para ello utilizaron embriones de un estadio temprano, cuando tan sólo es una célula plurinucleada, en qué los núcleos que originarán la línea germinal se agrupan en uno de los polos. A estos embriones se les inyecta una mezcla de dos plásmidos. Un plásmido porta un elemento *P* defectivo (sin transposasa) donde se ha insertado el gen de interés. El otro plásmido, el *helper* (ayudante, en inglés), contiene el gen de la transposa-

21. Los siRNAs y miRNAs y su modo de acción se explicaron en el apartado 4.6.5 «Interferencia del ARN».

sa pero carece de las repeticiones invertidas necesarias para insertarse, con lo que en las siguientes divisiones celulares se pierde. El elemento *P* puede insertarse en cualquier parte del genoma de los núcleos de la línea germinal, gracias a la transposasa expresada por el *helper*. El individuo adulto que se obtiene es normal pero en su descendencia aparecen individuos transgénicos, portadores del gen introducido con el elemento *P*. Cada individuo transgénico muestra el gen insertado en un lugar distinto y éste se hereda de modo estable y mendeliano.

5.3. Herramientas estadísticas básicas en el estudio genético

Cada patrón de herencia conlleva la observación de proporciones fenotípicas distintas. Cuando se aborda el estudio de un carácter, se contrastan las proporciones observadas en la descendencia con las esperadas según el modelo mendeliano (que es el más sencillo). En caso de no ajustarse al modelo, se plantea una hipótesis nueva, se calculan las proporciones esperadas bajo esta hipótesis y se contrasta de nuevo. Para este análisis se emplean una serie de herramientas estadísticas básicas.

En general, utilizamos la estadística en dos momentos: 1) cálculo de las proporciones esperadas y 2) contraste de hipótesis.

5.3.1. Cálculo de las proporciones esperadas

Las proporciones esperadas se calculan aplicando los principios básicos de probabilidad:

- La probabilidad de dos sucesos independientes es el producto de las probabilidades de estos sucesos.
- La probabilidad de un suceso unión de sucesos excluyentes, es la suma de las probabilidades de los sucesos individuales.

Así, si consideramos que, en un experimento de dihibridismo, los alelos de cromosomas homólogos tienen la misma probabilidad de formar parte de gametos viables y que todos los gametos tienen la misma probabilidad de intervenir en la formación de cigotos viables, podemos calcular las siguientes probabilidades:

Un gameto AB en la F1	P(AB) = P(A) x P(B) = 1/2 x 1/2 = 1/4
Un cigoto AABB en la F2	P(AABB) = P(AB) x P(AB) = 1/4 x 1/4 = 1/16
Un individuo A_bb en la F2	P(A_bb)= P(AAbb)+ P(Aabb) +P(aAbb)= 1/16 + 1/16 + + 1/16 = 3/16

En ocasiones, el cálculo de las proporciones esperadas se complica bastante. Por ejemplo, al considerar un mayor número de caracteres simultáneamente, o cuando los distintos gametos no se originan con la misma probabilidad. En estos casos puede resultar de gran ayuda el uso del **cuadrado de Punnet**. Para construirlo, dibujamos una cuadrícula en cuyo borde superior colocamos los gametos producidos por un progenitor y en el borde izquierdo los gametos producidos por el otro progenitor. En cada casilla escribiremos la combinación genotípica resultante de combinar los gametos correspondientes a las entradas superior e izquierda (Figura 5.3). La probabilidad de obtener un individuo correspondiente a una casilla determinada es el producto de las probabilidades de los gametos que lo originan.

5.3.2. Contraste de hipótesis

Las frecuencias que observamos siempre dependen de un muestreo y, como tal, están sujetas a un error. Por ello, las frecuencias observadas raramente coinciden exactamente con las esperadas. Para decidir si las desviaciones que observamos pueden ser debidas al muestreo o, por el contrario, son excesivas y debe buscarse otra explicación plausible, utilizamos la **prueba de bondad de ajuste de la chi-cuadrado** (χ^2).

La formula de la χ^2 es:

$$\chi^2 = \sum \frac{(observado - esperado)^2}{esperado}$$

Los valores que comparamos en una χ^2 siempre deben ser valores absolutos, nunca frecuencias relativas o porcentajes.

Veamos como lo aplicaríamos en un ejemplo concreto:

- En la tabla 5.1 vemos que, tras cruzar guisantes de semillas lisas con guisantes de semillas rugosas, Mendel obtuvo una F_2 de 7324 semillas: 5474

lisas + 1850 rugosas (proporción 2,96:1). Según la primera ley de Mendel esperamos una proporción 3:1, es decir: 5493 lisas y1831 rugosas. Para decidir si las diferencias observadas pueden ser debidas al azar, aplicamos el test de χ^2:

$$\chi^2 = (5474\text{-}5493)^2/\ 5493 + (1850\text{-}1831)^2/1831 = 0{,}0657 + 0{,}1972 = 0{,}2628$$

Para decidir si este valor de χ^2 indica un valor que se ajusta a nuestro modelo recurrimos a las tablas de χ^2. En estas tablas, encontramos los valores de χ^2 que marcan los valores límite de esta distribución para una probabilidad y número de grados de libertad determinados. Siempre que el valor obtenido en el test de χ^2 sea inferior al valor tabulado, aceptaremos la hipótesis testada como buena. Un valor de χ^2 superior al valor tabulado nos llevará a rechazar la hipótesis.

Los grados de libertad (gl) nos indican el número de modos en que pueden variar libremente las clases observadas. El número de grados de libertad en una prueba de bondad de ajuste de la χ^2 es uno menos que las clases observadas. En nuestro ejemplo tenemos 2 clases (lisas y rugosas) por lo que, partiendo de un número total de semillas, una vez sabemos el número de semillas lisas, el número de rugosas ya queda fijado. En general utilizamos un nivel de significación de $\varepsilon = 0{,}05$. En la Tabla de χ^2 el valor tabulado para 1 gl y $\varepsilon = 0{,}05$ es de 3,84. Dado que, en nuestro ejemplo, el valor obtenido en la prueba de χ^2 (0,26) es inferior al tabulado (3,84), aceptamos que nuestra observación se ajusta a los valores esperados según la hipótesis formulada. Las desviaciones obtenidas entre los valores observados y esperados pueden explicarse por azar en el muestreo.

Debe tenerse en cuenta que cuando decimos que la descendencia se ajusta a ciertas frecuencias, se refiere al promedio poblacional, no a un cruce concreto entre 2 individuos. Las frecuencias esperadas siguen las leyes de la probabilidad y, por tanto, debe considerarse un número elevado de individuos para no observar desviaciones totalmente aleatorias. En organismos con una descendencia considerable, como algunas plantas o incluso en Drosophila, podríamos contrastar las frecuencias observadas en un cruce concreto de 2 individuos con las frecuencias esperadas. Sin embargo, no podemos contrastarlas en la descendencia de una familia humana concreta que, por numerosa que sea, siempre contará con un número muy limitado de individuos.

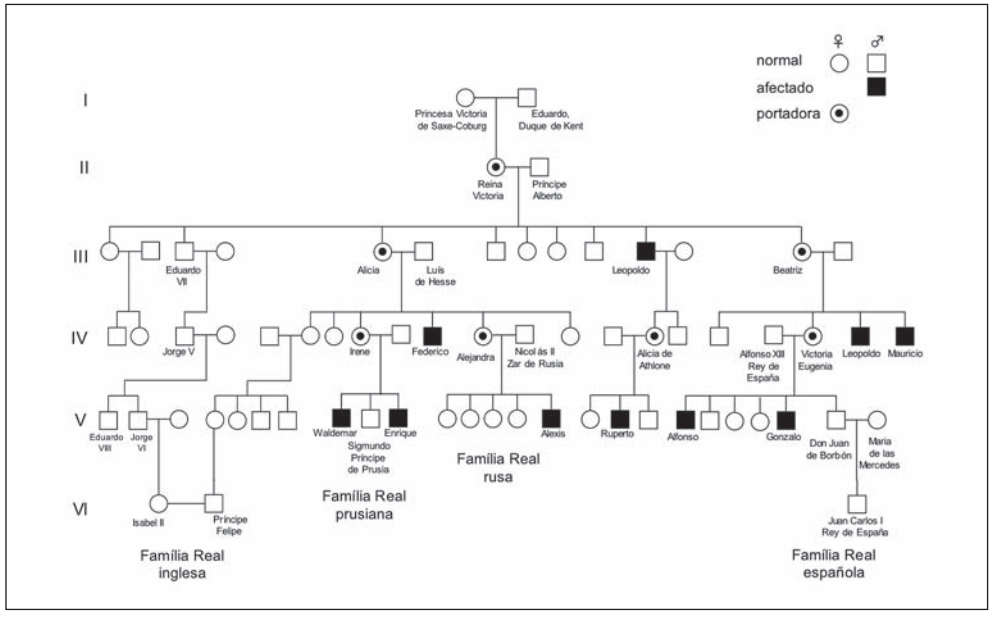
5.4. Estudio de pedigríes

La genética humana tiene un gran interés debido a que puede explicar la aparición de multitud de enfermedades. Sin embargo, el estudio de la genética humana se ve limitado por una serie de circunstancias. Por un lado, el investigador no puede recurrir a cruzamientos controlados que le permitan comprobar sus hipótesis. Por otro lado, de cada cruzamiento sólo se obtienen unos pocos descendientes que no permiten reconocer, sin ambigüedad, los patrones de herencia más simples.

Para solventar estas dificultades, los genetistas recurren habitualmente al análisis de pedigríes. Un **pedigrí** es una representación gráfica de la historia familiar para un carácter hereditario determinado, expresado en forma de árbol genealógico (Figura 5.11). En general, la reconstrucción de un pedigrí empieza con un individuo que presenta la característica que se desea estudiar, generalmente una enfermedad, y se le va añadiendo la información de todos los familiares conocidos.

En un pedigrí, cada generación se indica por números romanos y cada individuo dentro de una generación determinada se suele distinguir por números

Figura 5.11. Árbol genealógico (parcial) de las casas reales europeas donde se muestra la herencia de la hemofilia



arábigos. Los hombres se representan por cuadrados y las mujeres por círculos. Los individuos afectados por la característica se indican mediante símbolos llenos (cuadrados o círculos), mientras que los que no la presentan por símbolos vacíos. Las líneas que unen los distintos símbolos indican las relaciones entre los individuos: de pareja (línea horizontal directa) o ascendencia/descendencia (línea vertical).

El estudio de un pedigrí permite averiguar si un carácter determinado es dominante o recesivo y si su herencia es autosómica o ligada al sexo. A estas conclusiones no se llega contrastando frecuencias en la descendencia, sino reconociendo patrones asociados con los distintos tipos de herencia. Por ejemplo, si un individuo afectado desciende de un padre y una madre no afectados, la característica debe ser necesariamente recesiva (o más raramente haber aparecido como mutación dominante en uno de los gametos que originó a este individuo).

Para decidir qué tipo de herencia causa el patrón de fenotipos que observamos en un pedigrí determinado, debemos partir de una hipótesis y rastrear todo el pedigrí. Según esta hipótesis, asignamos un posible genotipo para cada individuo y nos cercioramos que no existe ningún caso incompatible con ella. Para ello, debemos tener en cuenta que el genotipo de los individuos que presentan el fenotipo dominante, en ocasiones, puede ser diverso (homocigoto o heterocigoto). Sólo podemos asegurar que el individuo es heterocigoto cuando desciende de un progenitor de fenotipo recesivo o tiene un hijo de fenotipo recesivo. Por ello, al comprobar nuestra hipótesis debemos asegurarnos de considerar todos los genotipos posibles para cada individuo. En la genealogía de la Figura 5.11 se muestra la transmisión de la hemofilia, un carácter recesivo ligado al cromosoma X. En el esquema, se han marcado como portadoras (X^AX^a) a las mujeres que tuvieron algún hijo hemofílico. Nótese, sin embargo, que una mujer sana pero heterocigota para la hemofilia puede, por azar, no pasar el alelo defectuoso a su progenie masculina. Por ello, no podemos confirmar el genotipo del resto de mujeres sanas.

Los pedigríes también se utilizan en consejo genético y en medicina forense. Una pareja con antecedentes familiares para alguna enfermedad hereditaria puede recurrir a un genetista en busca de consejo genético. El especialista estudia el pedigrí y, conociendo el tipo de herencia del carácter, establece la probabilidad de que un hijo de la pareja pueda padecer la enfermedad. En medicina forense ha sido habitual usar la herencia del grupo sanguíneo AB0 en pruebas de paternidad, conociendo el grupo sanguíneo de la madre, el hijo y el/los presuntos padres. Debe destacarse que estas pruebas permiten descartar a un posi-

ble padre pero nunca identificarlo al 100%. En la actualidad se utiliza una combinación de marcadores moleculares, tipo **microsatélites** o **SNPs**[22] que permiten una mayor definición del origen genético de un individuo.

5.5. Otros tipos de herencia

Existen otros tipos de herencia biológica menos generales o más complejos, como la herencia ligada al cromosoma Y, la herencia citoplasmática, la interacción génica, la herencia cuantitativa...

La **herencia holándrica** es la que presentan los poquísimos genes ligados al cromosoma Y. Estos genes se heredan por línea paterna, sin saltos generacionales, y son exclusivos de los machos.

La **herencia citoplasmática** hace referencia a todos aquellos genes localizados en los orgánulos citoplasmáticos (mitocondrias y cloroplastos). En principio, estos orgánulos son aportados al cigoto exclusivamente por el óvulo. Por tanto, la herencia de estos genes es por línea materna, pero en este caso son compartidos por machos y hembras.

Otro fenómeno que altera las proporciones respecto a la herencia mendeliana, durante la gametogénesis, es la **impronta genética**. Como vimos en el capítulo IV, algunos genes quedan marcados diferencialmente en machos y hembras. Esto lleva a que en la siguiente generación sólo se exprese el alelo heredado de uno de los progenitores.

La **interacción génica** tiene lugar cuando 2 o más genes actúan conjuntamente en la expresión de un carácter: el efecto de los alelos del primer locus depende de qué combinación alélica presente el segundo locus.

La **herencia cuantitativa** se observa en fenotipos de variación continua, muchas veces asociados a caracteres de interés en explotaciones agropecuarias (tamaño corporal, número de crías por camada, producción de leche, longevidad...). Estos caracteres están modelados por numerosos genes y una importante influencia ambiental, y dan origen a una variación continua en la población. Aunque la herencia de cada locus individual implicado en un carácter de variación continua sigue las leyes de Mendel, resulta difícil discernir la aportación de cada locus al fenotipo y para el estudio de caracteres cuantitativos debe recurrirse al uso de herramientas estadísticas especiales.

22. Veremos los marcadores moleculares tipo SNPs y los microsatélites en el apartado 6.3.

5.6. Evolución del concepto de gen

Al principio del capítulo ya apuntamos que la definición de gen puede ser diversa y más o menos compleja. Por ello, ofrecimos una definición sencilla que nos ha sido útil hasta aquí: un gen es *cualquier factor que determina una característica particular y heredable de un organismo*. Esta definición, un tanto abstracta, continúa siendo válida hoy en día. Sin embargo cuando queremos especificar qué es un gen con una visión más material, la cosa se complica y su definición ha sufrido una profunda evolución a medida que se conocía mejor la base molecular de la herencia.

Aunque el estudio sistemático de la genética y de los genes se remonta a mediados del siglo XIX con los trabajos de Mendel, la palabra genética no fue utilizada hasta 1905 por Bateson, quien también fue el primero en traducir al inglés los trabajos de Mendel en 1901. Por su parte, el vocablo gen fue utilizado por primera vez en 1909 por Johannsen quien lo definía como *las condiciones especiales presentes en los gametos de forma única e independiente y que especifican muchas características de los organismos*.

Con la teoría cromosómica de la herencia, el gen pasa a ocupar una posición física sobre el cromosoma. Por ello, Morgan y sus colaboradores explican las observaciones de ligamiento y recombinación por un modelo de *genes ordenados linealmente* como las cuentas de un collar.

Cuando se observó el efecto de algunas mutaciones sobre las vías metabólicas se formuló el principio de *un gen una enzima* que posteriormente se generalizó a *un gen una cadena polipéptidica*. Más tarde llegó el conocimiento del ADN como material genético, el descubrimiento de su estructura peculiar que asegura su fiel replicación, y el descifrado del código genético. La idea de gen se identifica con la de ORF, *una secuencia de ADN que codifica para proteína*.

Pero algunos ARN tienen función propia sin traducirse a proteína y, además, con la secuenciación de los primeros genes se descubrieron los intrones y el procesamiento del ARN. La idea de gen como ORF excluye a los genes de ARN y a los intrones. Esto llevó a la definición de gen como *un fragmento de ADN que se transcribe*. De nuevo, el descubrimiento del ayuste alternativo supuso un reto para la definición de gen. Un gen es *una secuencia de ADN que puede codificar para transcritos alternativos aunque éstos produzcan proteínas distintas*.

Una definición más actual afirma que un *gen es una porción de ADN compuesta de una región que se transcribe y de una región reguladora que hace posible la transcripción* (Griffiths el al, 2008)

El proyecto ENCODE[23] ha puesto de manifiesto la gran complejidad que pueden alcanzar las secuencias codificadoras y reguladoras:

- Muchos intensificadores (regiones reguladoras) se localizan muy alejadas de la secuencia que se transcribe. Además, un intensificador puede ser compartido por diversas unidades de transcripción.
- Existen numerosos genes que se solapan, bien sea compartiendo la misma cadena de ADN con otra pauta de lectura o utilizando la cadena complementaria. Es el caso del gen *Adh* de *D. melanogaster* que está incluido en un intrón enorme del gen *osp* y además utiliza la cadena complementaria.
- Existencia de trans-splicing, que es la unión en un ARNm de exones localizados en distintos pre-ARNm, que entra en conflicto con la vieja idea de un gen un locus.
- La gran cantidad de elementos móviles que también nos enseñan que la idea de una localización fija para cada gen no es válida.

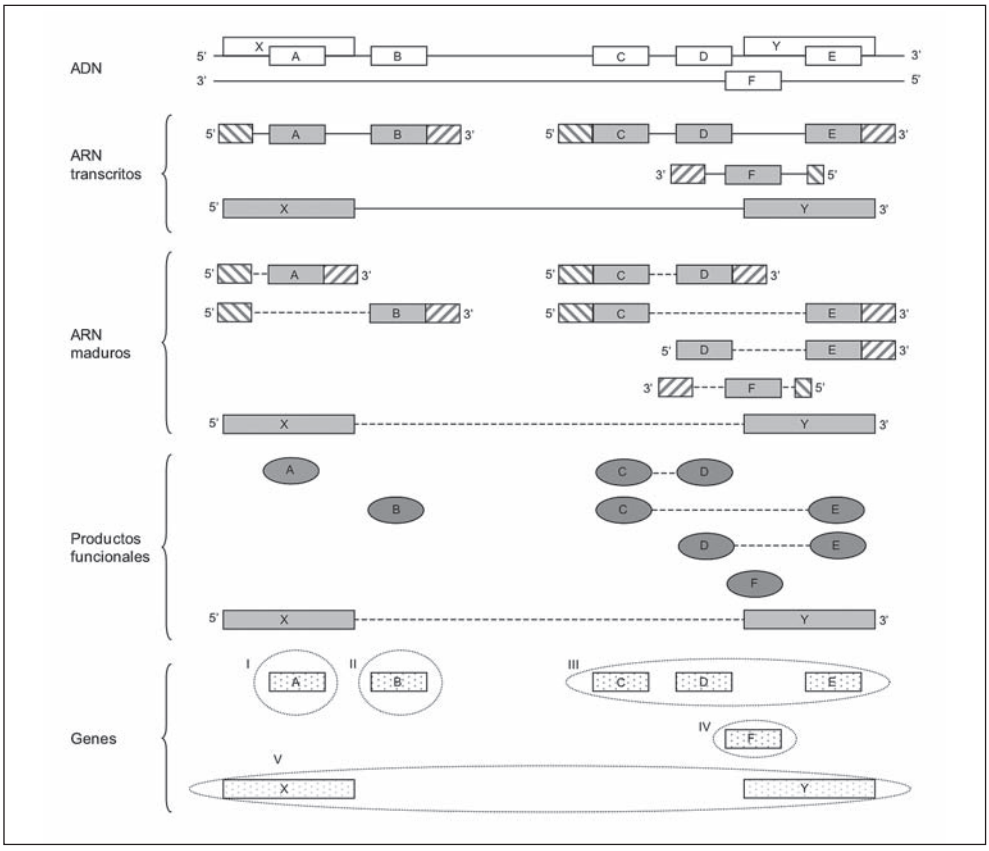
Por todo ello, se ha propuesto una nueva definición de gen cómo:

> El gen es una combinación de secuencias genómicas que codifican un conjunto coherente de productos funcionales potencialmente solapados. (Gerstein et al., 2007)

Según esta definición, lo que determina un gen son las secuencias de los productos funcionales, no las secuencias transcritas (Figura 5.12). Por ello, los distintos productos funcionales del mismo tipo (proteína o ARN) que están codificadas en secuencias de ADN solapadas, se agrupan en un mismo gen. Por el contrario, aquellos productos funcionales que no comparten secuencias del ADN aunque procedan de transcritos que comparten el promotor y el inicio de transcripción, se consideran genes independientes. Así, aunque son importantes, las regiones reguladoras y las secuencias UTR no se consideran en esta definición de gen propuesta.

23. El Proyecto ENCODE (ENCyclopedia Of DNA Elements) pretencde identificar todos los elementos funcionales en la secuencia del genoma humano. Ver apartado 6.5.

Figura 5.12. Ejemplo de una región de ADN con 5 genes según la definición post-ENCODE



Las cajas con letras indican los exones. Las cajas rayadas indican las regiones UTR en los genes de proteína. Los genes I y II comparten la región 5'UTR pero ningún exón.

El gen III está formado por un grupo de 3 exones (C, D y E) con corte y empalme alternativo, en el que cada exón es compartido por distintas proteínas.

El gen IV a pesar de estar solapado al gen III no comparte ningún exón (están en cadenas complementarias).

El gen V tiene el exón X que se solapa con el exón A del gen I y el exón Y que se solapa con el exón E del gen III. No obstante, como es un gen de ARN, mientras que los genes I y III son de proteína, representa un gen distinto.

Capítulo VI

... y de genomas, Genómica

La genética aborda el estudio de cada carácter fenotípico analizando el efecto de un gen o unos pocos genes sobre dicho carácter. Aunque este enfoque ha proporcionado conocimientos importantes sobre la herencia de muchos caracteres de interés (p.e. algunas enfermedades), sabemos que un gen no actúa de un modo independiente. Por un lado, la expresión de un gen viene determinada por la acción del producto de otros genes. Por otro lado, a menudo el producto de un gen interactúa con los productos de otros genes para determinar la manifestación de un carácter fenotípico. Por tanto, para resolver algunos de los retos planteados a la genética era preciso abordar conjuntamente el estudio de numerosos genes de los que, en la mayoría de las ocasiones, no conocíamos su ubicación y/o existencia.

A finales del siglo XX, con el avance tecnológico que permitía la obtención rápida y fiable de gran cantidad de secuencias de ADN, nació la **genómica** que es el estudio del genoma completo de los individuos. En un principio, puede parecer que la genómica tan sólo supone un salto cuantitativo de la **genética** en el modo de buscar respuestas. Sin embargo, en realidad, la genómica supone una revolución cualitativa tanto en el enfoque de las preguntas como en el modo de responderlas. La genómica permite que nos planteemos nuevas preguntas tales como: ¿Cuántos genes tiene la especie *x* y cuál es su localización en los cromosomas? (**genómica estructural**); ¿Qué genes se expresan en un tejido concreto en un momento concreto? (**genómica funcional**); ¿Qué genes comparten y en cuáles se diferencian las especies *x* e *y*? (**genómica comparada**). Para responderlas se precisa de técnicas de alto rendimiento (*high-throughput*) tanto en la obtención de los datos (secuenciación y/o genotipado) como en el análisis posterior de éstos (tratamiento bioinformático).

6.1. Genomas

El **genoma** de un individuo es el conjunto completo de su información genética, que incluye a todos sus genes y también sus zonas reguladoras e intergéni-

cas. En otras palabras, es la secuencia de todo su ADN. Generalmente, cuando se habla del genoma de un organismo eucariota, se refiere únicamente al ADN nuclear. La secuencia del ADN circular de los orgánulos celulares (mitocondrias y cloroplastos) se conoce como **genoma mitocondrial** y **genoma del cloroplasto** respectivamente.

Las células de un organismo pluricelular pueden tener morfología y función muy diversa pero todas ellas tienen el mismo genoma. La diferenciación celular se consigue por procesos epigenéticos a partir de la misma información del ADN. La totalidad de las células del individuo se han originado a partir del cigoto por sucesivas rondas de mitosis y, por tanto, son portadoras de la misma información genética. No obstante, en algunas divisiones mitóticas puede originarse alguna **mutación**, dando lugar a clones de células portadores de estos ligeros cambios. La mayoría de estas mutaciones no se manifiestan en el fenotipo y pasan totalmente desapercibidas (muchas corresponderán a regiones intergénicas o intrones o pueden ser cambios sinónimos). Otras mutaciones somáticas sí que conllevan un cambio fenotípico aparente de las células del clon resultante que en ocasiones es totalmente inocuo, como en el caso de muchos lunares de la piel, pero en otros casos puede ser patológico, como en el caso de muchos tumores.

El genoma secuenciado y publicado de una especie es un **genoma patrón** o de **referencia**. A menudo no corresponde a un único individuo sino que en su secuenciación se ha utilizado ADN de diversos individuos. Así, por ejemplo, el genoma humano es la secuencia mosaico del ADN de numerosos donantes anónimos, entre los que figuran individuos de distintas etnias, tanto hombres como mujeres. Por otro lado, aunque el organismo sea diploide (con 2 copias para cada gen), el genoma patrón sólo representa la secuencia de una de las dos copias.

6.1.2. Genomas secuenciados

En los años 80 del siglo XX se iniciaron diversos proyectos genoma. En cada uno de estos proyectos participaron numerosos investigadores y laboratorios que unieron esfuerzos para determinar la secuencia del genoma de uno o más organismos. La idea era proporcionar, a la comunidad científica mundial, una herramienta básica que permitiría acelerar considerablemente muchas investigaciones.

El Proyecto Genoma Humano[24] se inició en 1990 para obtener la secuencia del genoma humano, identificar todos sus genes, almacenar la información en una base de datos, mejorar las herramientas de análisis y tratar las cuestiones éticas, legales y sociales que pudieran surgir durante el desarrollo del proyecto. Como parte del Proyecto Genoma Humano se abordó también la secuenciación del genoma de algunos organismos modelo como la bacteria *Escherichia coli*, la levadura *Sacharomyces cerevisae*, el nematodo *Caenorhabditis elegans*, la mosca del vinagre *Drosophila melanogaster* y el ratón *Mus musculus*. Los genomas de estos organismos también son de gran interés para la comunidad científica y, dado su menor tamaño, resultaron ser un perfecto banco de pruebas de las técnicas necesarias para abordar la secuenciación de un genoma. Así, cuando a principios de 2001 se publica el genoma humano ya se dispone del genoma de diversos organismos modelo (Tabla 6.1). En la actualidad se dispone del genoma completo de unas 1000 especies y el número va aumentando rápidamente. Con el desarrollo de las técnicas de secuenciación de nueva generación que suponen mayor rapidez y menor coste, las previsiones son de que el número de genomas secuenciados siga aumentando a mayor ritmo.

Diversos consorcios han abordado la secuenciación de otros muchos organismos. La Tabla 6.1 destaca la obtención de algunos otros genomas de organismos modelo como la planta *Arabidopsis thaliana*, el hongo *Neurospora crassa* y la rata (*Rattus norvegicus*). También cabe destacar la obtención del genoma del chimpancé (*Pan troglodytes*) cuya proximidad con el hombre permitirá comprender mejor la evolución en los homínidos. También ha supuesto un hito destacable la publicación del genoma de 12 especies de Drosophila que ya han permitido realizar un gran número de estudios de genómica comparada entre especies muy cercanas.

También se ha abordado la secuenciación del ADN de especies extintas. En 2008 se publicó el genoma del mamut (*Mammuthus primigenius*) y en febrero de 2009 se anunció que ya se disponía de un primer borrador del genoma del Neandertal (*Homo neanderthalensis*). La secuenciación del genoma de estas especies presenta dificultades añadidas. La recuperación de ADN antiguo es escasa y generalmente ha experimentado inicios de degradación, como fragmentación o modificaciones moleculares. Por todo ello, es fácil que aparezca contaminación por ADN exógeno. En el caso del mamut, se ha podido extraer ADN del pelo de

24. El Proyecto Genoma Humano involucró el esfuerzo de laboratorios de 18 países. (http://www.ornl.gov/sci/techresources/Human_Genome/hg5yp/index.shtml).

Tabla 6.1. Algunos genoma secuenciados que han supuesto grandes hitos de la Genómica

Organismo	Dom.	kb	ORF	Publicación	
Homo sapiens	E	3080436	38612	15/02/2001	Nature **409**:860-921
Primer Genoma secuenciado de cada Dominio					
Haemophilus influenzae	B	1830	1774	28/07/1995	Science **269**:496-512
Methanocaldococcus jannaschii	A	1664	1878	23/08/1996	Science **273**: 1058-1073
Saccharomyces cerevisae	E	12069	5860	25/10/1996	Science **274**: 546-567
Principales Organismos Modelo					
Escherichia coli	B	4639	4612	5/09/1997	Science **277**: 1453-1474
Caenorhabditis elegans	E	100272	23209	11/12/1998	Science **282**: 2012-2018
Drosophila melanogaster	E	137000	14100	24/03/2000	Science **287**: 2185-2195
Arabidopsis thaliana	E	119707	26735	14/12/2000	Nature **408**: 796-815
Mus musculus	E	2644093	39625	5/12/2002	Nature **420**: 520-562
Neurospora crassa	E	37097	10082	24/04/2003	Nature **422**: 859-868
Rattus norvegicus	E	2750000	21166	1/04/2004	Nature **428**: 493-521
Pan troglodytes	E	3100000	48910	1/09/2005	Nature **437**: 69-87
Otros Genomas					
12 DROSOPHILA				8/11/2007	Nature **450**: 203-218
Mammuthus primigenius				20/11/2008	Nature **456**: 387-390
Homo sapiens (J.C. Venter)				21/09/2007	PLoS Biology **5**, e254
Homo sapiens (J.D. Watson)				17/04/2008	Nature **452**: 872-876
Homo sapiens (Nigerian male)				6/11/2008	Nature **456**: 53-59
Homo sapiens (YanHuang No.1)				6/11/2008	Nature **456**: 60-65
Homo sapiens (S.J. Kim)				24/06/2009	Genome Research **456**: 53-59
Homo sapiens (Korean male)				20/08/2009	Nature **460**:1011-1015

A 5/09/2009 se disponía de la secuencia completa de 1094 genomas (Archaea: 67; Bacteria: 911; Eukaria: 116) Según GOLD: Genomes OnLine Database (http://www.genomesonline.org/).

Dom., dominio taxonómico al que pertenece el organismo en cuestión: A: Arqueas; B, Bacterias; E: Eucariotas; kb, tamaño del genoma secuenciado en kilobases; ORF, número de *Open Reading Frames* anotados; Publicación, fecha y revista en que fueron publicados.

especímenes encontrados congelados en Siberia que datan de 18.500 años. Tras la secuenciación se han realizado comparaciones con el genoma del elefante (la especie viva más cercana) y se ha comprobado una gran homología, descubriendo sólo unos niveles muy bajos de contaminación por genomas bacterianos. En el caso del Neandertal, el ADN proviene de huesos que han estado expuestos a los agentes físicos y las bacterias del suelo durante miles de años. Además, encontrar homología con el genoma humano no supone ninguna garantía ya que podría ser debida a la contaminación involuntaria por los investigadores al manipular los huesos en el yacimiento o al obtener el ADN en el laboratorio. Por ello, deben tomarse precauciones especiales en la manipulación de las muestras y, en general, sus partes más superficiales no se utilizan para extraer ADN.

La disponibilidad del genoma de una especie facilita enormemente la obtención del genoma de otros individuos de la misma especie (**resecuenciación del genoma**). De momento, se han publicado 6 nuevos genomas humanos: 3 de personas identificadas (J.C. Venter, uno de los padres del Proyecto Genoma Humano; J.D.Watson, uno de los codescubridores de la estructura del ADN; y S.J. Kim, un investigador coreano), y otros 3 de individuos anónimos (un hombre nigeriano, un hombre chino, y un hombre coreano). También se ha publicado la secuenciación de dos genomas completos de una misma mujer: uno de células sanas y otro de células cancerosas (Ley et al., 2008). Todo ello ha creado gran expectativa sobre la posibilidad de la obtención del genoma personal que podría abrir las puertas a una medicina personalizada pero que también suscita mucha polémica de tipo ético.

Otro tipo de estudios que se ha iniciado son los de metagenómica. En estos estudios, se toma una muestra de un sistema y se secuencia todo el ADN presente. Básicamente se refiere al ADN de microorganismos presentes en un medio. Por ejemplo se ha secuenciado el metagenoma de sistemas tan dispares como pueden ser: las aguas residuales, las fuentes sulfurosas, el rumen de la vaca, la flora bacteriana de la boca humana, el agua de distintos mares, o los filtros de aire del interior de viviendas.

6.2. Secuenciación de Genomas

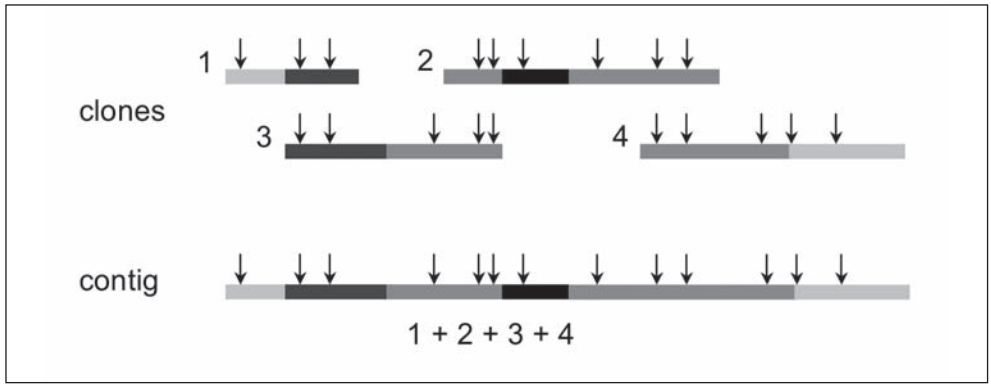
La secuenciación de *novo* del genoma de una especie requiere utilizar diversas estrategias y metodologías. Algunas técnicas actuales de secuenciación permiten leer hasta 800-1000 nucleótidos por reacción. El genoma humano

contiene 3000 Mb[25] distribuidos en 23 cromosomas. Esto significa que ha sido necesario secuenciar, ordenar y solapar correctamente como mínimo unos cuatro millones de fragmentos (de unos 800 nt) para obtener toda la secuencia del genoma.

Para secuenciar un genoma completo pueden utilizarse dos aproximaciones diferentes: secuenciación basada en el mapa, y secuenciación por fragmentos elegidos al azar (*shotgun*). La secuencia del genoma humano se obtuvo utilizando ambos enfoques: el consorcio público internacional utilizó métodos basados en el mapa genético mientras que la empresa privada Celera utilizó el método de *shotgun*.

Para realizar una secuenciación basada en el mapa, primero debe obtenerse un mapa genético o físico que localiza multitud de distintos marcadores sobre cada cromosoma. A continuación se trocea el genoma por digestión parcial de alguna enzima de restricción. Los fragmentos obtenidos se clonan en un vector tipo cósmido, YAC (*Yeast Artificial Chromosome*) o BAC (*Bacterial Artificial Chromosome*). Con la ayuda de los marcadores genéticos, se seleccionan los clones necesarios para obtener una **genoteca** genómica, o colección de fragmentos de ADN que representan al genoma completo. Algunos clones contienen más de un marcador y esto permite seleccionar clones que solapen y por tanto se obtenga una secuencia continua (o *contig*) de parte del genoma (Figura 6.1).

Figura 6.1. Obtención de un *contig*



Las flechas indican la posición de marcadores. El solapamiento de distintos clones permite construir un *contig*.

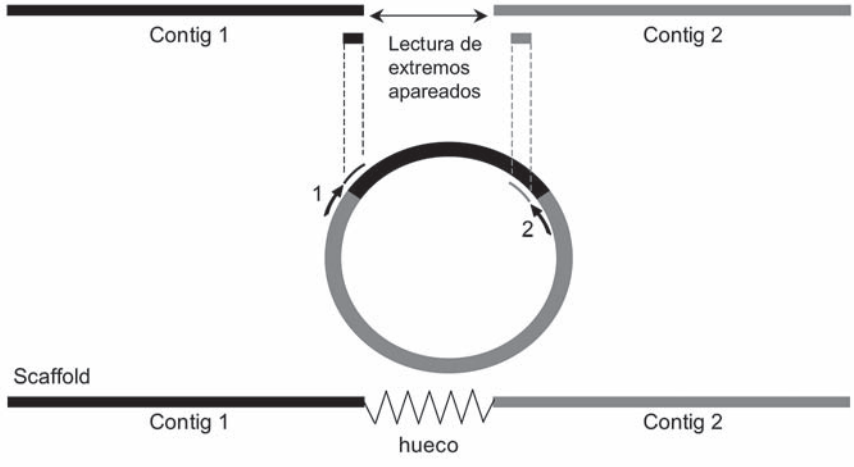
25. 1 Mb (1 megabase) equivale a 10^6 pb.

En la secuenciación de fragmentos al azar, el genoma se trocea en fragmentos pequeños que se clonan en plásmidos, que son vectores de clonación más fáciles de manipular. Tras la secuenciación, los programas informáticos se encargan de localizar la posición relativa de los clones identificando las posibles superposiciones y obtener así los *contig*. La necesidad de obtener clones que se superpongan exige que, en realidad, el genoma se secuencie más de una vez y se crea una redundancia de la mayor parte de la secuencia. Cada genoma tiene asociado un valor de redundancia. Cuando, por ejemplo, se dice que un genoma es 10x significa que, por termino medio, cada posición nucleotídica ha sido secuenciada 10 veces (a partir de 10 fragmentos distintos). Cuanto mayor sea el grado de redundancia, mejor calidad obtendremos en el ensamblado final.

Las genotecas que se utilizan para secuenciar genomas son colecciones de fragmentos de un tamaño conocido (por ejemplo, 2, 10 o 150 Kb) insertados en unas posiciones concretas de los vectores de clonación. Estas posiciones están cercanas a unas secuencias que son reconocidas por unos cebadores diseñados especialmente. La secuenciación se realiza a partir de estos cebadores en ambos extremos del ADN clonado, obteniéndose dos secuencias de 800-1000 pb. Así, la longitud de la secuencia obtenida por clon sólo es una pequeña parte del ADN clonado.

Algunos fragmentos del genoma son especialmente difíciles de secuenciar y/o ensamblar. En los genomas eucariotas son frecuentes las regiones de ADN repetitivo. Las regiones de secuencia única del genoma se pueden ensamblar fácilmente en *contigs* pero estos se interrumpen donde empieza una zona repetitiva. Para poder ordenar los distintos *contigs* entre sí se puede utilizar las lecturas de los dos extremos de un clon individual o lectura de extremos apareados (Figura 6.2). Si la secuencia de un extremo de un clon pertenece a un *contig* y el otro extremo a otro *contig*, se deduce que éstos son contiguos. Además, obtenemos información de la orientación relativa de estos *contigs* y de la longitud de la secuencia que falta entre ellos o *gap*, puesto que conocemos la longitud que separa a las dos secuencias de los extremos de un clon. De este modo podemos construir un *scaffold* (palabra inglesa que significa andamio). Un *scaffold* es una colección de *contigs* ordenados entre los que existen huecos o *gaps*. A partir de los clones que tienen secuencias en dos *contigs* contiguos puede intentarse completar la secuencia del *gap*.

Figura 6.2. Construcción de *scaffolds*



Las flechas marcadas como 1 y 2 indican la posición de los cebadores universales. Se indican las dos secuencias obtenidas a partir del ADN clonado y su posición relativa en los *contigs*.

6.2.1. Secuenciación del ADN

Los primeros métodos de secuenciación rápida se desarrollaron a finales de la década de 1970. El método que obtuvo más aceptación y que posteriormente ha pervivido con modificaciones y mejoras es el método Sanger[26] o didesoxi (Sanger et al., 1977). El método didesoxi se basa en la replicación del ADN y para ello se prepara una solución que contiene:

- el fragmento de ADN que queremos secuenciar
- un cebador específico
- una polimerasa
- desoxirribonucleótidos (dATP, dTTP, dCTP y dGTP)
- didesoxirribonucleótidos (ddATP, ddGTP, ddCTP o ddTTP).

La polimerasa incorpora desoxirribonucleótidos a partir de la secuencia de un cebador y, siguiendo las reglas generales de complementación de bases, copia

26. Frederick Sanger recibió el premio Nobel de química en 1980 por su contribución a la determinación de la secuencia de bases de los ácidos nucleicos. Era su segundo Nobel. En 1958 lo recibió por su trabajo sobre la estructura de las proteínas y en especial de la insulina.

la secuencia de la cadena de ADN. Sólo utilizamos 1 cebador porque en cada reacción de secuencia sólo podemos leer una de las 2 cadenas de ADN.

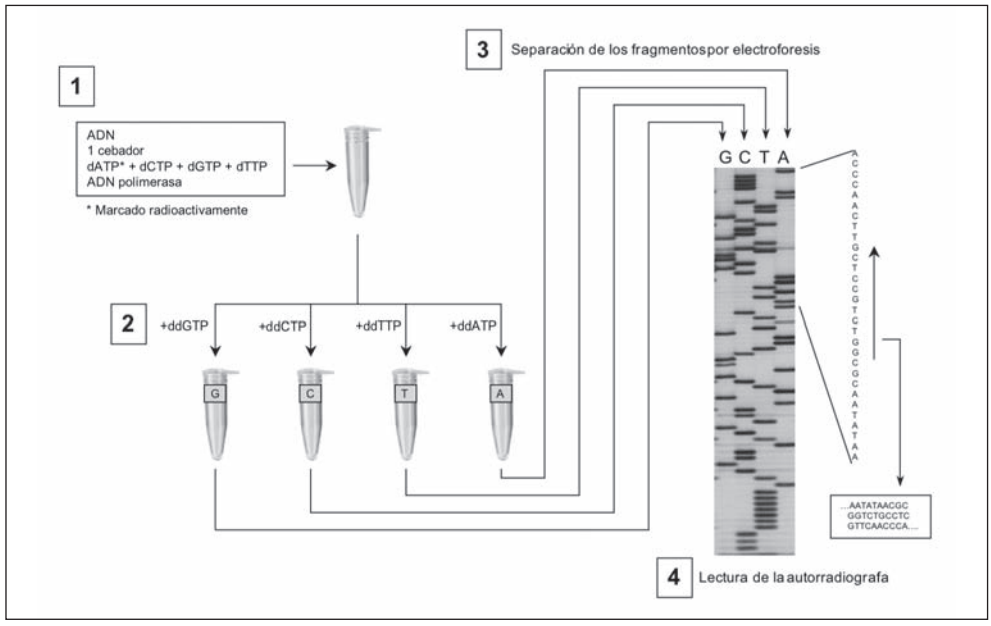
Nos interesa obtener una colección de fragmentos que, empezando en el mismo punto marcado por el cebador, tengan distintas longitudes. Esto se consigue añadiendo terminadores didesoxirribonucleótidos (ddNTPs). Estos terminadores son nucleótidos modificados (no llevan el grupo OH en el carbono 3') de modo que una vez incorporados impiden que la cadena siga creciendo. La adición a la cadena que se está sintetizando de cada uno de los dos tipos de nucleótidos normal/modificado es al azar. De este modo, obtenemos una colección de fragmentos de ADN de cadena sencilla de distintas longitudes, según en qué punto se haya incorporado un didesoxirribonucleótido.

El paso siguiente a la reacción de secuenciación es la separación por electroforesis de la colección de fragmentos obtenidos. Dos bandas contiguas en el gel de electroforesis corresponden a dos fragmentos que se diferencian por la adición de un único nucleótido. De algún modo, debemos determinar qué terminador específico corresponde a cada fragmento, y esto nos indicará el nucleótido que existe en una posición concreta de la secuencia de ADN.

Originariamente, para distinguir el didesoxi incorporado a cada fragmento se utilizaba un rodeo ingenioso. Se preparaba una mezcla con el ADN, el cebador, la polimerasa y los desoxirribonucleótidos. En esta mezcla, estaba marcado radiactivamente uno de los nucleótidos (o el cebador) y se dividía la reacción en cuatro tubos independientes. Cada uno de estos 4 tubos recibía, además, un único terminador correspondiente a cada uno de los 4 nucleótidos distintos. Una vez finalizada la reacción, los fragmentos de cada uno de los tubos se separaban en carriles contiguos de un gel de acrilamida y se obtenía una autorradiografía (Figura 6.3). Las bandas en el gel indican las posiciones hasta las que ha migrado cada fragmento, que dependen de su tamaño. Bandas de tamaños consecutivos indican posiciones de nucleótidos contiguos. Leyendo simultáneamente los 4 carriles de una reacción de secuencia se iba determinando la secuencia del ADN.

En la actualidad el proceso es mucho más rápido, simple y automatizado. Para empezar, no se trabaja con radioactividad sino que cada uno de los terminadores incorpora un fluorocromo distinto. Esto permite realizar la reacción de secuenciación en un único tubo. La reacción de secuencia se realiza como una PCR ligeramente modificada, con lo que en cada reacción se obtiene una gran cantidad de fragmentos. Por otro lado, la electroforesis se realiza en unas máquinas especiales, los secuenciadores. El ADN se inyecta en unos capilares rellenos

Figura 6.3. Secuenciación manual por el método didesoxi

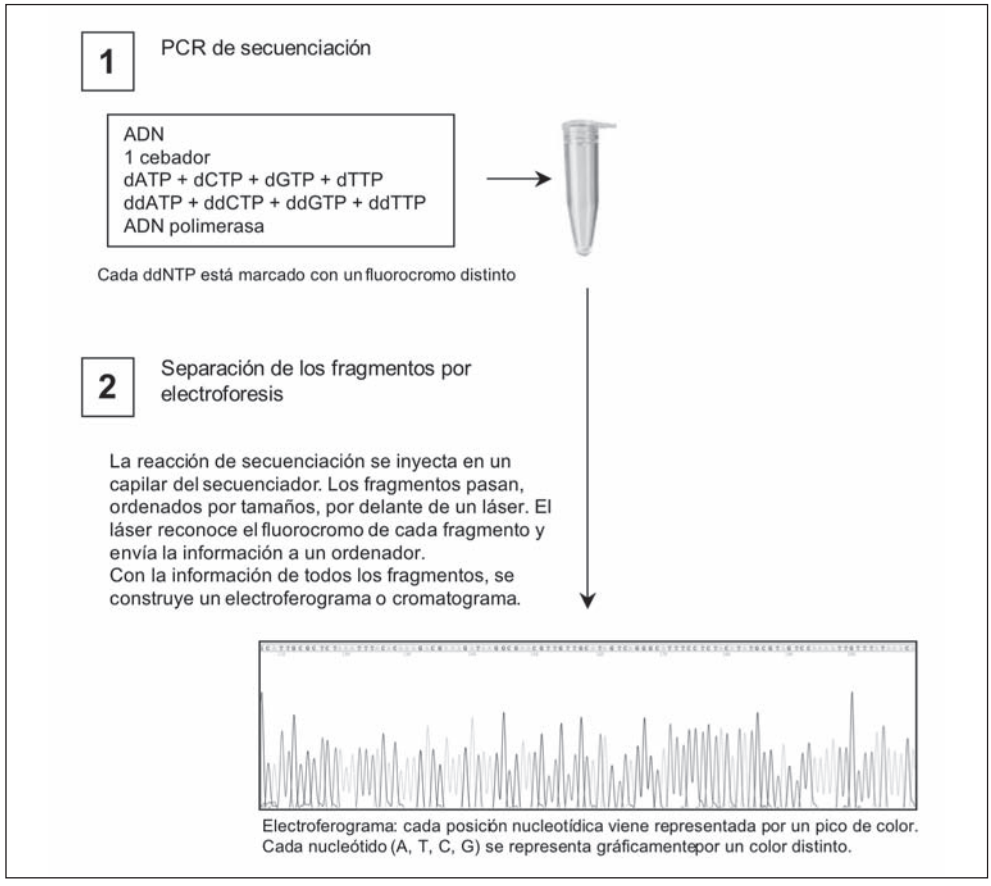


de una matriz que separa electroforéticamente los fragmentos por tamaño. De este modo, los fragmentos van pasando ordenadamente de menor a mayor por delante de un láser. El láser capta la fluorescencia asociada al fluorocromo del ddNTP correspondiente a cada fragmento y lo almacena en un archivo informático que, más tarde, nos devolverá traducido como una secuencia de ADN en un formato gráfico denominado electroferograma o cromatograma (Figura 6.4).

En los últimos años, diversas empresas han desarrollado distintas plataformas de secuenciación conocidas como *next-generation sequencing* o la nueva generación de secuenciación. Todas ellas utilizan técnicas *high-throughput* y persiguen obtener secuencias fiables en menor tiempo y reduciendo costos respecto a la secuenciación tradicional basada en el método Sanger. Una de las plataformas que parece tener más aceptación es la plataforma 454 de Roche. La tecnología 454 permite la automatización de todo el proceso de secuenciación de un genoma, desde la generación de la genoteca y la amplificación de los fragmentos hasta la obtención de las secuencias (Margulies et al., 2005).

Para empezar, se obtiene una genoteca de fragmentos de ADN de cadena sencilla a los que se han unido dos secuencias adaptadoras distintas, cada una específica para el extremo 5' o 3' del fragmento. La secuencia de estos adaptadores se utiliza en los posteriores pasos de purificación, amplificación y secuencia-

Figura 6.4. Secuenciación automática por el método didesoxi



ción. La amplificación de los fragmentos de ADN se realiza in vitro, evitando el proceso de clonación. Todos los fragmentos se amplifican simultáneamente en el mismo tubo de reacción, mediante una PCR especial. El adaptador unido al extremo 5' facilita que el fragmento se adhiera a una microesfera especial. Cada fragmento queda adherido a una microesfera distinta y a continuación se añade una emulsión que mantiene las microesferas separadas. De este modo pueden amplificarse multitud de fragmentos simultáneamente y, al finalizar la PCR, los productos de amplificación de un fragmento individual se mantienen adheridos a la misma microesfera.

El método de secuenciación que utiliza la plataforma 454 se basa en la piro-secuenciación (Ronagui et al, 1996), que unifica los pasos de síntesis de ADN y detección de los nucleótidos, de modo que los nucleótidos van siendo leídos a

medida que se incorporan a la cadena de ADN. La reacción de secuenciación tiene lugar en una placa especial con múltiples pocillos. Cada pocillo permite alojar una única microesfera, portadora de múltiples copias de un fragmento de ADN. La secuenciación se realiza paralelamente en todos los pocillos, iniciándose a partir de un cebador que reconoce la secuencia del adaptador unido al extremo 3' del fragmento. Para ello, los distintos dNTPs se van inyectando secuencialmente en un orden fijo a toda la placa. Cada vez que uno o más nucleótidos son incorporados a la cadena de ADN de uno de los pocillos se genera una señal lumínica que es captada por una cámara, y un ordenador registra la incorporación del nucleótido en este punto de la placa.

El método presenta la ventaja de una mayor automatización del proceso, obteniendo los datos de un modo más rápido y económico que el método tradicional. Su inconveniente es que sólo permite resolver secuencias de hasta 500-600 nucleótidos frente a los ~1000 del método tradicional. De todos modos, el resto de nuevas plataformas sólo permiten obtener secuencias mucho más cortas. Esto supone la necesidad de manejar un mayor volumen de fragmentos de secuencias, dificultando su ensamblaje y requiriendo el desarrollo de herramientas bioinformáticas más potentes.

Las plataformas de nueva generación se están usando más y más frecuentemente en la secuenciación de genomas completos y, por ejemplo, el genoma de J.D.Watson se obtuvo con la tecnología 454 (Wheeler, et al, 2008).

6.2.2. Ensamblado de secuencias

Las técnicas tradicionales de secuenciación de ADN permiten leer un máximo de 800 nucleótidos. Por tanto, la secuenciación de cromosomas enteros que constan de varios millones de pares de bases, debe abordarse de forma troceada. Por otro lado, las enzimas utilizadas en la secuenciación pueden introducir errores con una cierta frecuencia. Para detectar estos errores, es conveniente obtener la secuencia de las dos cadenas de la molécula de ADN. Todo ello hace que, al secuenciar un genoma, manejemos un número muy elevado de secuencias que debemos ordenar y ensamblar correctamente para obtener una lectura completa del texto del genoma. Es como si tuviéramos un montón de piezas de un puzzle que debemos recomponer para que tenga sentido.

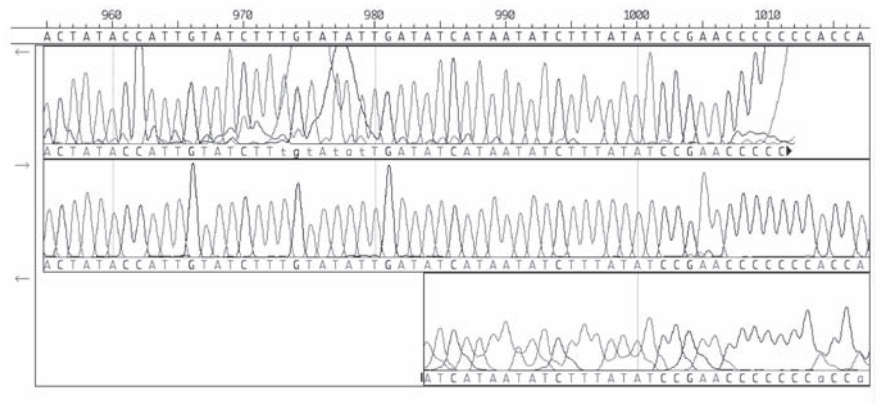
El ensamblaje se realiza con ayuda de potentes programas informáticos que encuentran homologías entre múltiples secuencias. Estos programas tienen en

cuenta que pueden encontrar secuencias correspondientes a ambas cadenas del ADN. La secuenciación, dado que se basa en la replicación, siempre es en el sentido 5' –> 3'. Cuando secuenciamos las dos cadenas de ADN, debemos transformar la secuencia de una de ellas en su complementaria y reversa para poder ensamblarlas. A partir de las secuencias ensambladas, estos programas generan una secuencia *consensus*.

En la Figura 6.5 se muestran los **cromatogramas** de 3 secuencias ensambladas. Las flechas de la izquierda indican el sentido en que se ha obtenido cada secuencia. Así, el segundo cromatograma corresponde a un fragmento central de una secuencia que se ha obtenido en el mismo sentido en que la vemos en el gráfico. Por otro lado, el primer y el tercer cromatograma corresponden al inicio y final de dos secuencias que el programa ha tenido que transformar en su cadena complementaria y reversa antes del ensamblaje. Por último, la secuencia en la parte superior es la secuencia *consensus*.

Un programa de ensamblaje debe permitir, además, editar los datos. En la Figura 6.5. podemos ver diversos ejemplos que son aquellas letras de las secuencias que aparecen en minúscula. El programa que devuelve los datos de la secuenciación no puede determinar con precisión qué nucleótido es el presente en determinadas posiciones y, por tanto, escribe Ns u otros códigos de indeterminación. El investigador, tras una comprobación visual de los picos de los nucleótidos en el cromatograma, puede optar por editar manualmente dichas posiciones nucleotídicas. En el ejemplo, la secuencia central, de muy buena calidad, confirma la validez de los nucleótidos editados manualmente. Para que un pro-

Figura 6.5. Ejemplo de cromatogramas de 3 secuencias ensambladas



yecto a gran escala, como la secuenciación entera de un genoma, sea factible en el menor tiempo posible requiere que la intervención humana sea lo menor posible. Por ello, la decisión sobre qué nucleótido ocupa una posición determinada de la secuencia, al menos en una primera fase, debe tomarlas el propio programa a partir de valores indicadores de la calidad de las diversas secuencias de que dispone para dicha posición.

6.3. Marcadores genéticos moleculares

En el capítulo anterior hemos visto que se pueden construir mapas genéticos midiendo la recombinación entre distintos genes. Sin embargo, esta metodología tiene algunas limitaciones. Por un lado, los genes utilizados deben tener variantes con un efecto fácilmente identificable. Por otro lado, en el genoma existen vastas zonas sin ningún gen. Por ello, y a partir del gran avance de las técnicas moleculares, se ha desarrollado la utilización de distintos marcadores genéticos moleculares.

Un marcador molecular es cualquier característica del ADN que presente variación, pudiendo ser cartografiada en el cromosoma. Podemos encontrar marcadores moleculares para cualquier zona del genoma y además tienen la ventaja de ser codominantes. Los marcadores moleculares más utilizados en la actualidad responden a dos tipos: los polimorfismos de un solo nucleótido o **SNPs** (acrónimo inglés de *Single Nucleotide Polymorphism* y pronunciado esnip) y la variación en el número de copias del ADN repetitivo, entre los que destacan los **microsatélites**.

Si bien los marcadores moleculares localizados sobre un mapa ayudan a posicionar las secuencias para ensamblar un genoma, la secuencia del genoma también facilita el descubrimiento de nuevos posibles marcadores moleculares.

6.3.1. SNPs

Los SNPs son variaciones de un único nucleótido entre dos secuencias de ADN homólogas. Para considerarse un polimorfismo en la población, cada variante debe estar presente con una frecuencia mínima del 1%.

Ejemplo de SNP:

Secuencia 1	**Secuencia 2**
5′ ACT**A**GGATCC 3′	5′ ACT**G**GGATCC 3′
3′ TGA**T**CCTAGG 5′	3′ TGA**C**CCTAGG 5′

Cada SNP se ha originado como una mutación al azar en una secuencia de ADN que posteriormente se ha extendido por la población. Los SNPs son muy estables y la gran mayoría muestran tan solo dos variantes (de los 4 posibles nucleótidos). Esto se explica porque la probabilidad que una misma posición nucleotídica cambie más de una vez por azar es muy baja.

Encontramos SNPs a lo largo de todo el genoma. La mayoría de los SNPs son neutros y no afectan al fenotipo, aunque ocasionalmente sí que pueden ser la causa directa de una variante fenotípica. Un individuo diploide puede presentar dos variantes distintas de un SNP que siempre son codominantes.

Todas estas características los convierten en perfectos marcadores genéticos que permiten construir mapas con SNPs a intervalos regulares a lo largo de todo el genoma.

Para localizar la posición de los SNPs deben secuenciarse fragmentos de ADN de distintos individuos. Se calcula que el genoma humano cuenta con unos 3 millones de SNPs, mostrando un SNP cada 300-1000 nucleótidos. En abril de 1999 se creó el TSC (*The SNP Consortium*) que pretendía descubrir 300.000 SNPs en el genoma humano. Al final del proceso de búsqueda se habían obtenido 1,8 millones.

Una vez obtenido un mapa de SNPs podemos utilizar métodos *high-trougput* para obtener el genotipo de un individuo para un conjunto grande de SNPs. En este caso no secuenciamos fragmentos de ADN en busca de variantes, sino que miramos directamente las posiciones que ocupan SNPs conocidos para comprobar qué genotipo tiene el individuo. Para ello pueden utilizarse, por ejemplo, microchips de ADN (o *microarrays*).

La información del genotipo nos permite realizar **estudios de asociación**. En éstos se compara el patrón de SNPs de un grupo de individuos afectados por un fenotipo determinado (por ejemplo, una enfermedad) con el patrón de un grupo de individuos no afectados (sanos). Los SNPs con variantes compartidas por todos los afectados pero no por los sanos nos dan pistas sobre la localización del gen causante de la enfermedad. Del mismo modo se puede obtener información sobre la tolerancia o las reacciones inmunológicas adversas a fármacos. Esta in-

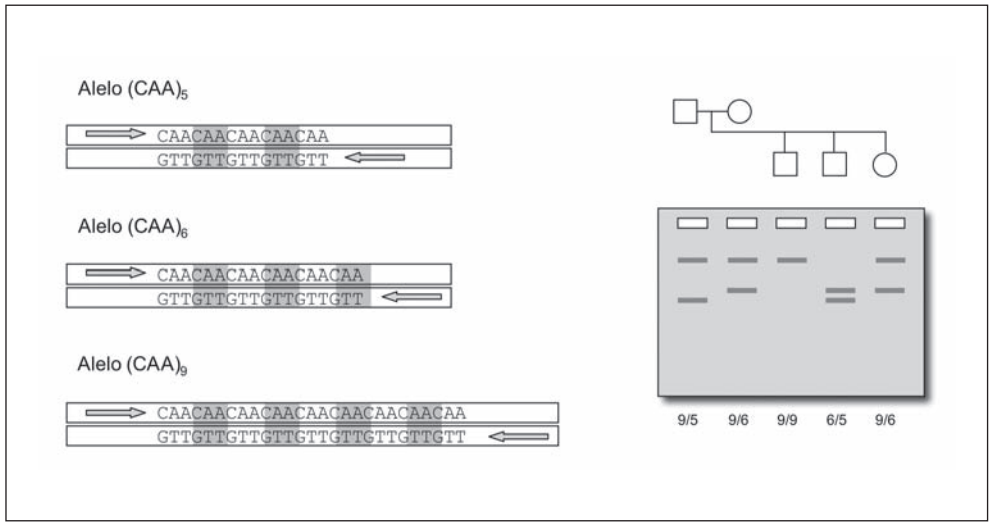
formación abrirá las puertas a una medicina personalizada más eficiente y menos agresiva.

6.3.2. Microsatélites

Los microsatélites son un tipo de VNTR (acrónimo inglés para: *Variable Number of Tandem Repeats*). En concreto son secuencias de 1 a 6 nucleótidos repetidas en tándem cuyo número de repeticiones es muy variable. Un organismo suele tener numerosos grupos de repeticiones de una misma unidad, dispersas por todo su genoma y flanqueadas por secuencias únicas. Los microsatélites son muy abundantes, pudiéndose encontrar un grupo de repeticiones cada 10 kb aproximadamente.

Las secuencias que flanquean las repeticiones permiten analizar un locus microsatélite individual bien sea utilizando enzimas de restricción (*RFLP,* del inglés *Restriction Fragment Length Polymorphism*) o diseñando cebadores específicos y amplificando el fragmento por PCR (AFLP, del inglés *Amplified Fragment Length Polymorphism*). En ambos casos, se determina la longitud del fragmento por electroforesis (Figura 6.6).

Figura 6.6. Detección por AFLP del genotipo para un locus microsatélite en una familia



En este ejemplo hay 3 alelos para el locus estudiado. Las flechas indican la posición y dirección de los cebadores utilizados para amplificar el microsatélite. Los números debajo del gel de electroforesis indican el genotipo de cada individuo.

Los microsatélites son marcadores moleculares con numerosos alelos en la población. Esta gran variabilidad responde a una alta tasa de mutación que parece ser debida a un deslizamiento o *slipagge* entre las cadenas de ADN durante la replicación (ver capítulo VII, Figura 7.4). La tasa de mutación, no obstante es suficientemente baja como para no observar diferencias en una generación. Estas características convierten los microsatélites en huellas dactilares del ADN o *fingerprinting* y, por ello, son muy utilizados como marcadores en biología forense. Así, desde 1997, el FBI (*Federal Bureau of Investigation*, Oficina Federal de Investigación de los Estados Unidos) utiliza una combinación de 13 loci microsatélite para identificar criminales o personas desaparecidas.

6.4. Genómica comparada

Con la secuenciación de los primeros genomas ya se vislumbró la gran potencialidad que ofrecía el estudio comparado de las secuencias. Por un lado, la comparación de secuencias dentro de un mismo genoma ha permitido, entre otras muchas cosas, identificar familias génicas y descubrir regiones repetidas en distintos cromosomas que pueden responder a inserciones de transposones. Por otro lado, la comparación de genomas de especies más o menos cercanas, ofrece la oportunidad de ver cómo ha actuado la evolución a nivel molecular y nos permite identificar elementos muy conservados entre especies alejadas filogenéticamente que deben responder a alguna funcionalidad.

Al realizar estudios de genómica comparada debemos asegurarnos que las secuencias que utilizamos en las comparaciones son realmente homólogas. La similitud entre secuencias homólogas no se debe al azar sino a que tienen un origen común.

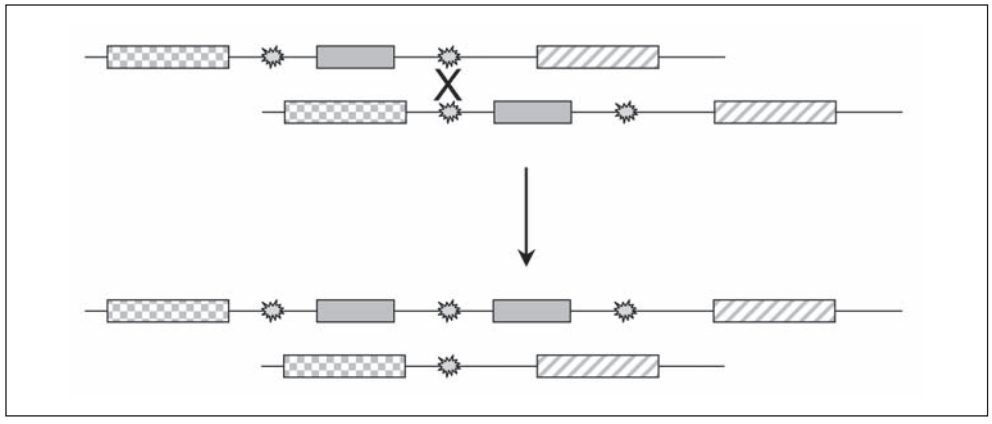
La genómica comparada es una ciencia muy joven que, además, destaca por la rapidez con que se está desarrollando. Esto ha propiciado que aparecieran términos nuevos que no siempre se han utilizado correctamente. Entre estos términos destacan por su importancia los de **ortología**, **paralogía** y **bloques de sintenia**.

6.4.1. Famílias génicas

En un genoma determinado, una familia génica es un grupo de genes cuya secuencia de nucleótidos es similar porque derivan de una secuencia común ancestral. Un gen ancestral se duplicó y posteriormente cada una de las copias ha evolucionado independientemente pudiendo acumular diferencias entre ellas. Las duplicaciones génicas son una importante materia prima que permite la aparición de nuevas funciones. El tamaño de las familias génicas es muy variable, pudiendo encontrar desde familias con sólo 2 genes a familias con centenares de ellos.[27].

El análisis de los genomas secuenciados demuestra que un porcentaje elevado de los genes se ha generado por duplicación génica (Zhang, 2003). La duplicación génica puede surgir por un emparejamiento desigual entre secuencias total o parcialmente idénticas seguido de un entrecruzamiento (Figura 6.7).

Figura 6.7. Duplicación génica por entrecruzamiento desigual



El entrecruzamiento desigual genera un cromosoma con duplicaciones y otro con deleciones.

Cuando un gen completo se duplica, la copia es idéntica al original. Con el tiempo, la evolución de estas secuencias duplicadas puede llevar a distintas situaciones:

- **Conservación de la función**. Las copias pueden permanecer practicamente iguales entre sí reteniendo su función, de forma que se consigue un aumento del producto génico. Este es el caso de los genes de ARNr, ARNt, histonas, etc.

27. La mayor familia génica en los mamíferos es la de los receptores olfativos que cuenta con alrededor de 1000 genes.

- **Subfuncionalización**. En este caso, cada una de las copias del gen adopta parte de las funciones del gen ancestral. Cada una de las copias puede, por ejemplo, expresarse en un tejido o en un momento del desarrollo distinto.
- **Neofuncionalización**. Una de las copias puede acumular cambios nucleotídicos (mutaciones) y adquirir una función nueva. La mayoría de las veces se trata de una función relacionada, más que de una función enteramente nueva.
- **Pseudogenización**. En ocasiones, las mutaciones que se acumulan en una de las copias provocan que ésta pierda totalmente la función, pasando a ser lo que llamamos un **pseudogén**.

En general, las familias génicas aparecen en el mismo cromosoma y a corta distancia, pero también es posible que se encuentren en cromosomas distintos. También podemos contemplar la existencia de superfamilias génicas, que incluirían los genes de familias muy relacionadas entre sí, aunque la diferenciación de sus secuencias podría ser mayor. Por ejemplo, los genes de las α-globinas (sobre el cromosoma 16 humano) constituyen una familia génica, al igual que los genes de las β-globinas (sobre el cromosoma 11 humano). Estas dos familias junto con la familia de las mioglobinas (sobre el cromosoma 22 humano) constituyen la superfamilia de las globinas.

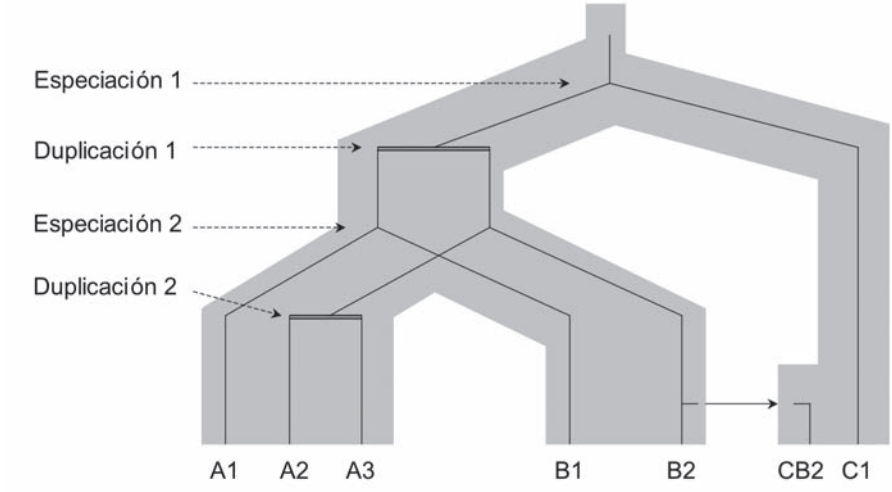
Cuando comparamos los genes de una familia génica en diversos genomas, estamos tratando con genes que han divergido por dos procesos evolutivos distintos: especiación y duplicación génica. Así, aunque todos ellos son genes homólogos, puesto que se han originado a partir de una misma secuencia ancestral, debemos distinguir entre **genes ortólogos** y **genes parálogos**.

Los genes ortólogos son aquellos que se han diferenciado a partir de un proceso de especiación. Los genes parálogos son los que se han originado por un proceso de duplicación.

A menudo, los conceptos de ortología y paralogía se han utilizado erróneamente. Por un lado, existen algunas confusiones al interpretar las definiciones. Por otro lado, las relaciones entre los genes de una familia génica pueden ser bastante complejas, dificultando la identificación de los distintos tipos de homólogos. Esto ha motivado la publicación de diversas revisiones (p.e. Fitch, 2000; Jensen, 2001; Sonnhammer y Koonin, 2002).

Dos genes homólogos en una misma especie siempre serán parálogos, porque deben provenir de una duplicación. Pero dos genes homólogos en especies distintas pueden ser ortólogos o pueden ser parálogos. Si tras la duplicación tiene lugar un evento de especiación, encontraremos genes parálogos en distintas especies (es el caso de los genes A1 y B2 en la Figura 6.8).

Figura 6.8. Relaciones de homología entre genes de una familia génica



Esquema de las relaciones entre los genes de una familia génica aparecidos por dos procesos de especiación, dos procesos de duplicación y uno de transferencia horizontal. El fondo gris muestra el árbol filogenético de las especies.

Es común también la idea de que un gen determinado sólo tiene un gen ortólogo en cada especie. En la Figura 6.8 se muestra que el gen B2 tiene 2 ortólogos en la especie A (A2 y A3). Si ascendemos en el árbol de los genes, vemos que el ancestro común de todos ellos es anterior a un proceso de especiación (especiación 2). Así, los genes A2 y A3 son **co-ortólogos** de B2, pero A2 y A3 son parálogos entre sí porqué se separaron en una duplicación (duplicación 2).

En una misma especie podemos encontrar diversos tipos de genes parálogos. Algunos de estos genes pueden ser co-ortologos de un gen en otra especie y otros no. En el ejemplo de la Figura 6.8 los genes A1, A2 y A3 son parálogos entre sí. Los genes A2 y A3 son co-ortólogos del gen B2 pero el gen A1 es parálogo del gen B2. Para distinguir entre estos dos tipos de parálogos dentro de una especie, Soonhammer y Koonin (2002) introdujeron los conceptos de inparálogos y

outparálogos. Los inparálogos son genes parálogos que han evolucionado por duplicación posterior a la especiación que separó las dos especies que estamos considerando. En nuestro ejemplo A2 y A3 son inparalogos puesto que aparecieron por la duplicación 2, que es posterior a la especiación 2. Los outparálogos son genes parálogos, en una misma especie, que surgieron por una duplicación anterior a la especiación. En el ejemplo, A1 es outparálogo tanto de A2 como de A3.

Otra posible relación de homología entre genes dentro de un genoma aparece al considerar genes que se han adquirido por transferencia horizontal. En este caso hablamos de **xenología** y de **genes xenólogos**. En el ejemplo de la Figura 6.8 el gen CB2 ha aparecido por transferencia horizontal del gen B2 (de la especie B a la especie C) y, por tanto, es un xenólogo de C1. La inclusión de estos genes xenólogos en la reconstrucción de la filogenia de una familia génica puede distorsionar notablemente los resultados. De hecho, frecuentemente, los resultados chocantes son los que dan pistas sobre un evento de transferencia horizontal reciente.

Los grandes estudios de genómica comparada se enfocan principalmente a identificar ortólogos que, generalmente, son equivalentes a nivel funcional y evolutivo. Por otro lado, el estudio de parálogos permiten detectar adaptaciones específicas de linaje y la aparición de nuevas funciones.

Los conceptos de ortología y paralogía, normalmente se refieren a genes pero pueden generalizarse a otros niveles de estudio. Así, podemos descender de nivel y considerar la homología de las secuencias correspondientes a distintos dominios proteicos o ascender y fijarnos en la homología de bloques de genes sobre un cromosomas.

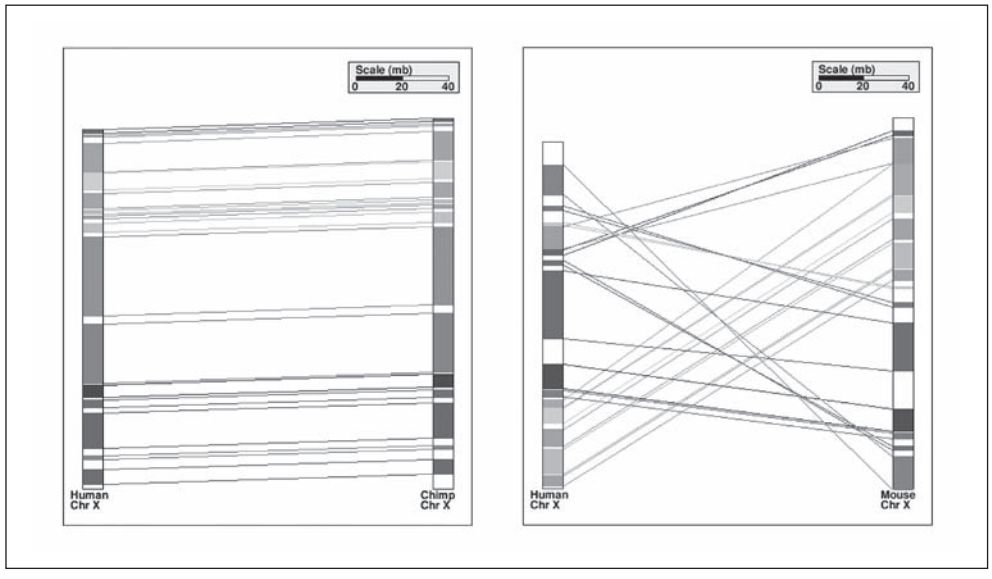
6.4.2. Bloques de sintenia

El término **sintenia** se utiliza para indicar que dos o más genes se encuentran en el mismo cromosoma. La localización de genes y otros marcadores sobre los cromosomas de dos especies cercanas muestra que la mayoría de genes sinténicos en una especie siguen siéndolo en la otra (conservan la sintenia) y que en muchos casos mantienen, además, el orden, son **colineares**. En ocasiones, no obstante, se observa que el orden se rompe de modo que todo un bloque de genes aparece en otra posición del mismo o diferente cromosoma (Figura 6.9).

Con la secuenciación de genomas completos se ha comprobado que estas reordenaciones son muy frecuentes. Se han realizado numerosos estudios sobre

estas roturas de la colinearidad que permiten ver cómo se han ido diferenciando las especies entre sí. La colinearidad de los genes es más alta cuanto más cercanas son las especies que estamos comparando. Esto se debe a que, cuanto más cercanas son las especies, menos tiempo han tenido para evolucionar de forma separada. La Figura 6.9 muestra que entre el cromosoma X humano y el del chimpancé no se aprecia ninguna reordenación de los marcadores, mientras que entre el cromosoma X humano y el de ratón ha habido diversas reordenaciones.

Figura 6.9. Bloques de sintenia entre el cromosoma X humano y el del chimpancé y entre el humano y el del ratón



La colinearidad se mantiene para los bloques de sintenia entre el cromosoma humano y el de chimpanzé pero se observa multitud de reordenamientos entre el humano y el del ratón.
Imagenes obtenidas en http://cinteny.cchmc.org/

A pesar de la oposición de algunos (Passarge et al., 1999), cada vez está más ampliamente aceptado el término de **bloque sinténico** como sinónimo de un fragmento de cromosoma en que se mantiene la colinearidad entre dos especies. También se usan frecuentemente en este sentido los términos de sintenia conservada y de microsintenia.

Sea como fuere, el estudio de estos bloques sinténicos y cómo varían entre distintas especies es muy importante para inferir las relaciones filogenéticas entre las especies. También permite inferir las relaciones de ortología para los genes de una familia génica que han divergido mucho entre especies. Además, la

observación de bloques de sintenia conservados de forma excepcional durante la evolución pueden revelar relaciones funcionales importantes entre determinados genes.

6.4.3. Los 12 genomas de Drosophila

Drosophila melanogaster es el organismo modelo favorito de muchos genetistas. Su estudio fue impulsado por los trabajos de Morgan a principios del siglo XX y desde entonces se ha estudiado su biología, desarrollo, genética, ecología... Por ello, fue elegida como una de las primeras especies eucariotas para secuenciar su genoma completo (Adams et al., 2000).

El genero Drosophila contiene unas 1500 especies. Algunas de estas especies como *D. pseudoobscura* en América y *D.subobscura* en Europa son muy utilizadas en estudios de genética de poblaciones. Por ello, la comunidad drosofilista estaba muy interesada en obtener el genoma de más especies de Drosophila. Así, en 2005 se publica el genoma de *D. pseudoobscura* (Richards et al., 2005) y en 2003 ya se impulsa la secuenciación del genoma de otras 10 especies.

Las 10 especies adicionales fueron elegidas de modo que abarcaran un amplio rango de distribución geográfica y de variación en el tiempo de divergencia (Figura 6.10). Un primer grupo de especies (*D. sechellia*, *D. simulans, D. yakuba, D. erecta,* y *D. ananassae*) más cercanas a *D. melanogaster* divergieron entre 8 y 15 millones de años atrás. La comparación de estos genomas debía permitir encontrar genes o motivos génicos que evolucionan rápidamente durante la especiación. Otro grupo de especies (*D. willistoni, D. grimshawi, D. virilis*, y *D. mojavensis*) que ha evolucionado independientemente del grupo melanogaster durante 35-50 millones de años, debía permitir encontrar elementos conservados por una fuerte presión selectiva. Finalmente, *D. persimilis*, muy cercana a *D. pseudoobscura*, permitiría estudiar la especiación entre estas dos especies. Además, estas dos especies permiten disponer de un grupo cuyo tiempo de divergencia respecto a *D. melanogaster* es intermedio al de los otros dos grupos de especies, cubriendo mejor todo el espectro temporal.

En noviembre de 2007 se publica en la revista Nature la secuenciación de los 12 genomas de Drosophila (Drosophila 12 Genomes Consortium, 2007). Simultáneamente, aparecen unos 40 artículos en Nature y otras revistas (Genome Research, Genetics, Molecular Biology and Evolution, PLoS) con los resultados de los primeros análisis de genómica comparada entre estas especies. Algunos

Figura 6.10. Relaciones filogenéticas de las 12 especies de Drosophila cuyo genoma está completamente secuenciado

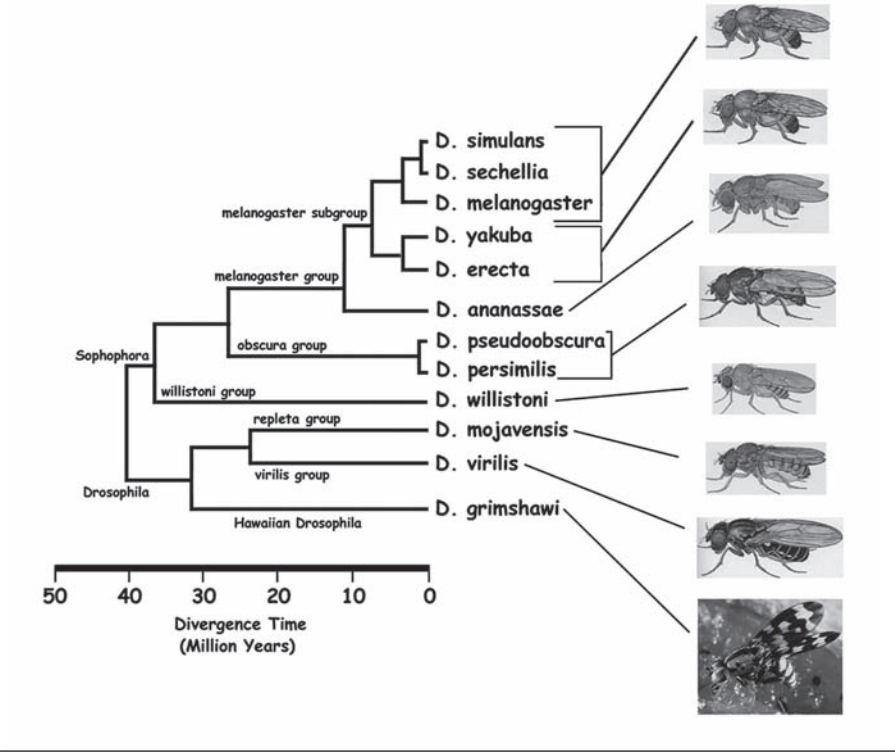


Imagen tomada de Fybase (http://flybase.bio.indiana.edu/)

de los aspectos que han analizado estos trabajos son: las tasas evolutivas de ortólogos y parálogos y la evolución de familias génicas concretas, los bloques de sintenia entre distintas especies, cómo ha evolucionado el cromosoma X en el género, la caracterización y evolución de elementos reguladores, micro ARN, transposones...

Por otro lado, el genoma de *D. simulans* se obtuvo como una secuencia consenso del genoma de 7 líneas de la especie (Begun et al., 2007). En este caso se dispone de las secuencias independientes de 5 de las líneas (con agujeros o *gaps* distribuidos de forma distinta en cada línea) y del genoma patrón. La comparación de estos cinco genomas parciales ha permitido realizar los primeros estudios de genómica poblacional en una especie (Begun et al., 2007). En este mismo sentido se está completando un proyecto para secuenciar 50 genomas de *D. melanogasster* (parte de un proyecto aún más ambicioso de secuenciar

300 genomas) que permitirá realizar numerosos estudios a nivel poblacional (http://www.dpgp.org/).

6.5. Proyecto ENCODE

La secuenciación del genoma completo de una especie es sólo un primer paso para comprender la información genética que encierra. A continuación deben anotarse (localizar sobre el genoma) las secuencias que corresponden a los genes. Sin embargo, una vez identificados los genes, aun queda mucha información por descubrir. Por un lado, en genomas complejos como el humano, las secuencias que corresponden a los genes sólo representan una pequeña porción del genoma. Por otro lado, la expresión de los distintos genes responden, entre otros factores, a la presencia de elementos reguladores en las secuencias no codificadoras. Por ello, el siguiente paso lógico es el de intentar descifrar la información que encierran estas zonas que, durante mucho tiempo, se consideraron ADN basura pero que, cada vez más, están demostrando tener diversas funciones.

Así, en septiembre de 2003 nace el proyecto ENCODE (**Enc**iclopedia **O**f **D**NA **E**lements) con el propósito de identificar todos los elementos funcionales de la secuencia del genoma humano (http://www.genome.gov/ENCODE/). El proyecto es muy ambicioso y por ello se inició como un proyecto piloto cuyo objetivo era anotar una región de unas 30 Mb del genoma que representa un 1% del total. En esta fase, científicos experimentalistas y computacionales interactuaron para evaluar distintos métodos de anotación del genoma y en 2007 publicaron los primeros resultados (The ENCODE Project Consortium, 2007).

Paralelamente se está desarrollando otro proyecto modENCODE (model organism Enciclopedia of DNA Elements). Este proyecto es similar a ENCODE pero para los organismos modelo *Caenorhabditis elegans* y *Drosophila melanogaster*, (http://www.modencode.org/).

Capítulo VII

Variabilidad genética y evolución

En el año 2009 se celebró el año Darwin en conmemoración del 200 aniversario del nacimiento de Charles Darwin (1809-1882) y del 150 aniversario de la publicación de su libro más emblemático «*El origen de las especies*». En su libro, Darwin aporta multitud de pruebas que demuestran la evolución de las especies y, además, expone un mecanismo que explica cómo tiene lugar la evolución: la selección natural.

El pensamiento de Darwin revolucionó el modo de pensar no sólo de los naturalistas o biólogos sino de toda la sociedad. La idea de la evolución por selección natural de Darwin es una idea simple que no siempre ha sido bien comprendida y que en su tiempo fue origen de gran controversia. Controversia que regularmente surge de nuevo. Es cierto que la selección natural no lo explica todo, pero sigue siendo un modelo válido para muchas situaciones observadas en la naturaleza.

Para que una población pueda evolucionar, debe existir previamente variabilidad. En este capítulo trataremos del origen de esta variabilidad en la población (**mutación**) y de cómo evolucionan las poblaciones, cambiando para adaptarse al medio (**selección natural**) o totalmente al azar (**deriva genética**).

7.1. Mutación

Antes que nada hay que apuntar que, en ocasiones, el término mutación puede llevar a confusión. Ello se debe a que usamos la misma palabra para referirnos a cosas muy relacionadas pero no intercambiables. Por mutación podemos entender, como mínimo tres cosas distintas:

1. Una característica fenotípica que difiere del patrón normal.
2. Un cambio hereditario en el material genético, sea o no observable en el fenotipo.
3. El mecanismo que provoca este cambio genético.

Por ello, las mutaciones también pueden clasificarse de diversas maneras según nos centremos en su efecto sobre el fenotipo (1), su grado de afectación del genoma (2) o su origen molecular (3).

Aunque la línea divisoria puede ser muy fina, en este apartado vamos a tratar principalmente de la segunda acepción de la palabra mutación: los cambios hereditarios en el material genético. Así, considerando la extensión del genoma que se ve afectada, podemos clasificar las mutaciones en: (i) mutaciones génicas, (ii) mutaciones cromosómicas.

Debemos diferenciar entre mutaciones somáticas y mutaciones en la línea germinal. Las **mutaciones somáticas** aparecen durante la proliferación celular (por mitosis) en los tejidos somáticos del organismo (aquellos que no forman gametos). Por ello, el efecto de estas mutaciones se limita al individuo dónde se han originado pero no se transmiten a la descendencia. Por el contrario, las **mutaciones en la línea germinal** no suelen afectar al individuo dónde se ha originado pero pueden transmitirse a la descendencia.

El efecto de una mutación somática será más o menos grave dependiendo del tipo celular y del momento del desarrollo en que aparezca la mutación. Cualquier mutación que aparezca en una célula, es heredada por todas las células que desciendan de ella, formando un **clon** de células portadoras de la mutación. La mayoría de las mutaciones somáticas no tienen mayor importancia, puesto que el producto de los genes alterados suele ser suplido por el del resto de células normales. Una excepción significativa lo constituye aquellas mutaciones que estimulan la división celular, con lo que la proporción de células con la mutación aumenta dramáticamente; estas mutaciones son la base de los distintos tipos de cáncer.

Por su parte, las mutaciones en la línea germinal son la materia prima que permite la evolución. Por un lado son de gran importancia para la supervivencia de la especie, ya que son el origen de la variabilidad que observamos en las poblaciones. Esta variabilidad permite que las poblaciones se adapten, por selección natural, al medio cambiante. Por otro lado, la acumulación diferencial de mutaciones entre distintas poblaciones es el origen de la mayoría de los procesos de especiación.

También podemos clasificar las mutaciones en beneficiosas, neutras o deletéreas según cómo afecten a la **eficacia biológica**[28] del individuo que la porta.

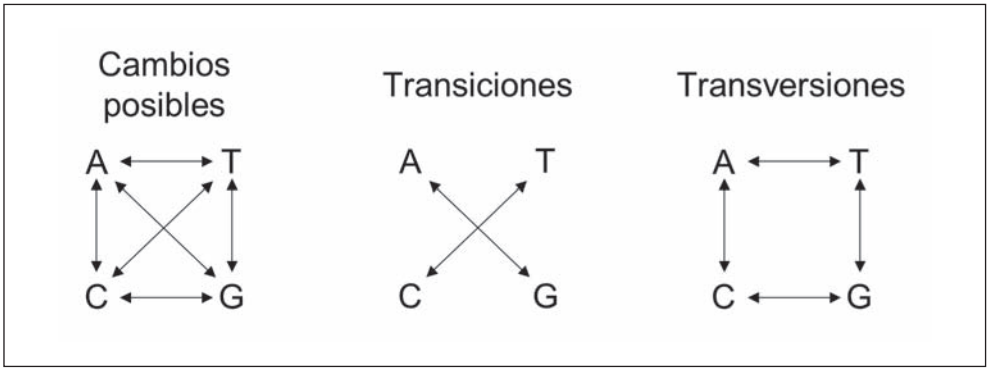
7.1.1. Mutaciones génicas

Las mutaciones génicas son alteraciones de uno o unos pocos nucleótidos que en general afectan a un solo gen. Las mutaciones génicas pueden deberse a substitución de bases o a inserción-deleción.

Sustituciones de Bases. Las sustituciones de bases son el tipo de mutación más sencilla. Consisten en el cambio de un único nucleótido por otro distinto. Cuando en la población existen dos variantes para una posición nucleotídica a frecuencias apreciables se dice que existe un polimorfismo o SNP (*single nucleotide polymorphism*).

Las bases nitrogenadas de los nucleótidos del ADN pueden ser purinas (A, G) o pirimidinas (T, C). El cambio de una base nitrogenada por otra del mismo tipo (purina por purina, o pirimidina por pirimidina) es una **transición**. Cuando el cambio es de purina a pirimidina o viceversa es una **transversión**. Aunque existen más posibles transversiones que transiciones (Figura 7.1), normalmente se observa una mayor frecuencia de transiciones.

Figura 7.1. Posibles cambios nucleotídicos



28. La eficacia biológica o *fitness* mide la capacidad que tiene un individuo de sobrevivir y dejar descendencia.

Además, dadas las características del código genético, las sustituciones de bases en la región codificadora pueden originar tres tipos de mutaciones: sinónimas, no sinónimas y sin sentido (Figura 7.3). Las **mutaciones sinónimas** son las que no suponen un cambio en el aminoácido codificado debido a la redundancia del código genético. Las **mutaciones no sinónimas** son las que sí determinan un cambio de aminoácido. Las **mutaciones sin sentido** son aquellas sustituciones que provocan la aparición de un codón de STOP.

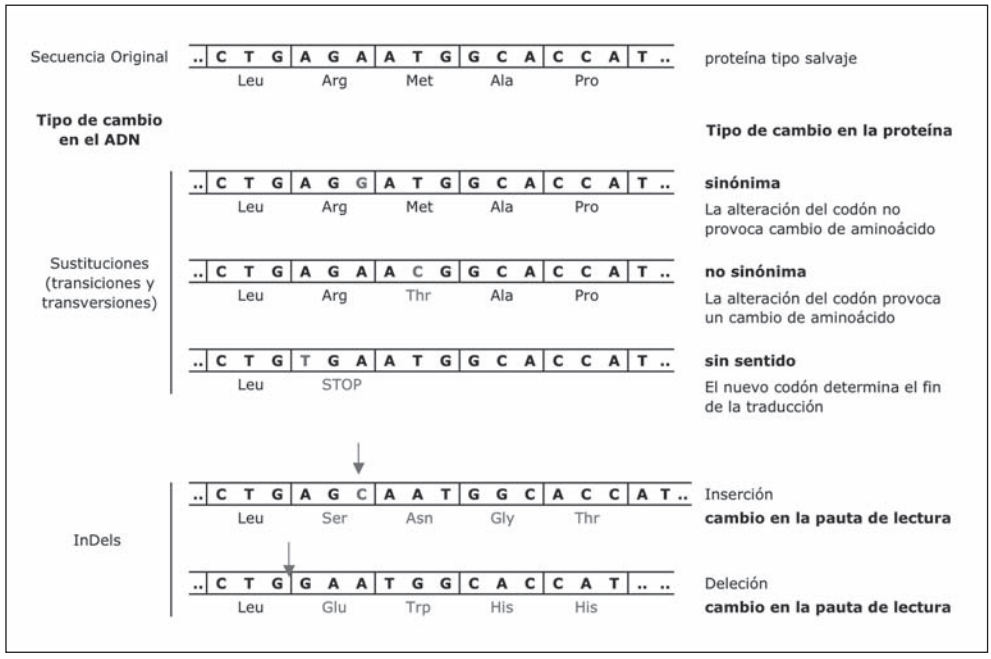
InDels. La **inserción** y la **deleción** de uno o más nucleótidos puede agruparse en una única clase denominada InDel. Ello se debe a que la diferencia en longitud entre dos secuencias alineadas puede deberse tanto a la ganancia de nucleótidos por parte de una de las secuencias como por pérdida por parte de la otra.

Cuando alineamos secuencias entre las que existen InDels debemos introducir espacios en blanco o *gaps* (Figura 7.2) para mantener alineados los nucleótidos homólogos. Estas discontinuidades en las secuencias son sólo un artefacto gráfico puesto que, en la realidad, las secuencias de ADN son continuas.

Figura 7.2. Ejemplo de 2 secuencias alineadas entre las que se observa una substitución nucleotídica y un InDel.

Las inserciones y deleciones dentro de las secuencias codificadoras pueden causar disfunciones muy importantes del gen afectado, debido a un **cambio en la pauta de lectura** (Figura 7.3), que lleva a la selección natural a eliminarlas rápidamente. Por ello, observamos una mayor frecuencia de inserciones/deleciones en las regiones no codificadoras (intrones y ADN intergénico). En las regio-

Figura 7.3. Efecto de distintos tipos de mutaciones en la región codificadora

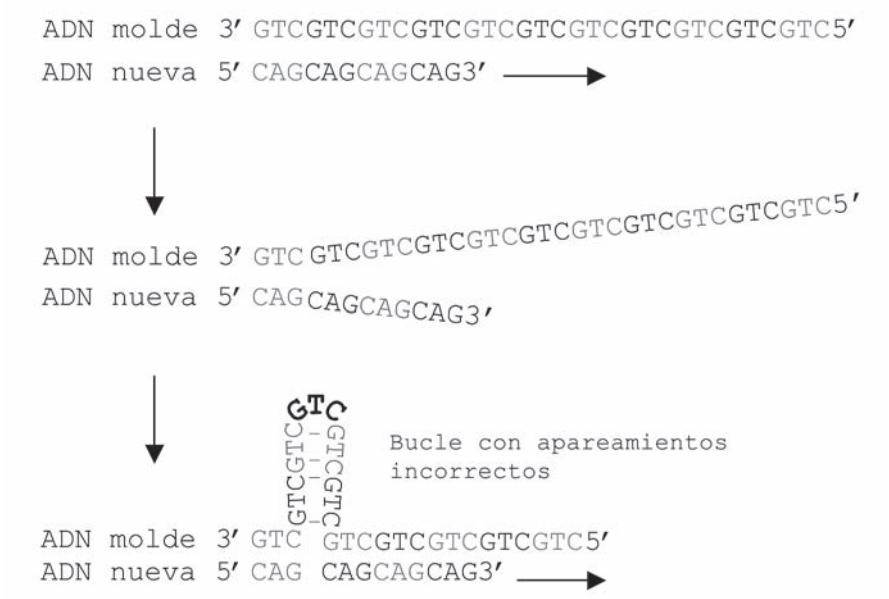


nes codificadoras podemos observar InDels correspondientes a un número de nucleótidos múltiplo de 3, que no alteran la pauta de lectura. Si el aminoácido afectado (ganado o perdido) no es crucial para la determinación de la estructura tridimensional y la función de la proteína, un Indel de 3 nucleótidos siempre será menos grave que otro de 1 o 2 nucleótidos.

Un tipo especial de InDels son los microsatélites, que son inserciones/deleciones de repeticiones cortas de nucleótidos. La frecuencia con que varía el número de repeticiones en los microsatélites es superior a la de mutación por substitución de bases. Esto se debe a los mecanismos que generan esta variación de número: el entrecruzamiento desigual y el deslizamiento de las cadenas de ADN (en inglés, *slippage*) durante la replicación (Figura 7.4). Ambos fenómenos son capaces de generar regiones repetidas a partir de secuencias no repetidas pero son más frecuentes cuando ya existen repeticiones.

El deslizamiento de las cadenas se produce porque, durante la replicación, la polimerasa puede separarse de la cadena que está copiando junto al último fragmento de nucleótidos que acaba de incorporar en la nueva cadena. Si la secuencia de esta zona presenta repeticiones, cuando las dos cadenas vuelven a unirse para continuar con la replicación pueden hacerlo en una posición equivocada,

Figura 7.4. Deslizamiento de las cadenas durante la replicación del ADN que genera variación en el número de repeticiones de un microsatélite



con lo que el número de repeticiones en la nueva cadena puede crecer o disminuir. Entonces, la cadena más larga puede formar estructuras en bucle, que impiden que se corrijan estos errores de la replicación. En la siguiente ronda de replicación, las dos cadenas de ADN harán de molde y las moléculas de ADN de cadena doble que se formen, diferirán en este InDel.

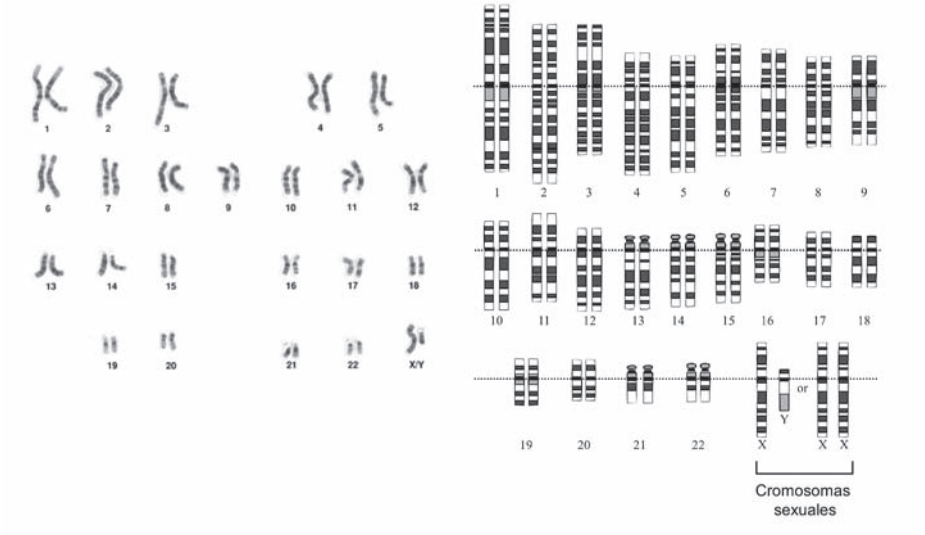
En el genoma humano, son frecuentes las repeticiones de 3 nucleótidos. Algunas de estas repeticiones están asociadas a ciertas enfermedades. Es el caso de la enfermedad de Huntington que se debe a la repetición de la secuencia CAG dentro del gen *HD* (*Huntington disease*). El gen normal presenta un número variable de repeticiones con 10-34 copias de CAG que se traducen en una cadena de poliglutaminas en la proteína. Los individuos afectados presentan una expansión de este trinucleótido de 42-121 copias (Sinden, 1999), produciendo una proteína nociva. En otras enfermedades asociadas a la expansión de trinucleótidos, el microsatélite se localiza fuera de la región codificadora. Así, la distrofia miotónica de tipo 1 se debe a la repetición del trinucleótido CTG en la región 3'UTR del gen *DM*. En este caso, el número de copias en el gen normal es de 5-37 mientras que en las personas afectadas es de 80 a 1000 (Sinden, 1999). La secuencia larga de $(CUG)_n$ del preARN captura una proteína de unión que impide el correcto procesamiento del ARN.

7.1.2. Mutaciones cromosómicas

Las mutaciones cromosómicas son aquellas que afectan a fragmentos grandes del ADN de un cromosoma o incluso a cromosomas enteros, de modo que en muchas ocasiones pueden visualizarse en el **cariotipo** del individuo que las porta.

El cariotipo[29] de un individuo es el conjunto de cromosomas que posee cada una de sus células somáticas. Para su estudio se utilizan células en mitosis, que es cuando los cromosomas están en su máxima contracción y se hacen visibles al microscopio. Además, se utilizan diversas técnicas de tinción que revelan distintos patrones de bandas en los cromosomas. Generalmente, un cariotipo se representa con los cromosomas ordenados por parejas según su tamaño, forma y patrón de bandas (Figura 7.5). Todos los individuos de una especie comparten el mismo cariotipo, exceptuando algunas aberraciones cromosómicas, resultado de mutaciones cromosómicas.

Figura 7.5. Cariotipo humano



La imagen de la izquierda muestra el cariotipo de un hombre, obtenido a partir de fotografías microscópicas de los cromosomas, recortadas y ordenadas. A la derecha, esquema representativo del cariotipo humano. (Imágenes adaptadas a partir de Wikipedia).

29. El cariotipo humano consta de 23 pares de cromosomas: 22 pares de autosomas más el par de cromosomas sexuales.

Las mutaciones cromosómicas pueden ser de dos tipos: mutaciones que alteran la estructura de los cromosomas y mutaciones que alteran el número de cromosomas.

Los cambios en la estructura de los cromosomas pueden ser debidos a **duplicaciones, deleciones, inversiones o translocaciones**. Todos estos cambios suponen la aparición de reordenamientos cromosómicos (nuevas ordenaciones de los genes de uno o más cromosomas). Las mutaciones de este tipo pueden explicarse por sucesos de rotura y reunión de los cromosomas o por entrecruzamiento entre segmentos repetitivos vía recombinación homóloga no alélica (NAHR, del inglés *Non Allelic Homologous Recombination*). Los reordenamientos cromosómicos son los responsables de que observemos diferencias entre los bloques de sintenia de distintas especies (Figura 6.9).

Duplicaciones. En ocasiones, parte de un cromosoma se duplica. La copia puede aparecer contigua a la región duplicada (duplicación en tándem) o alejada de ésta (duplicación desplazada). Como ya vimos en el capítulo anterior, las duplicaciones pueden originarse por entrecruzamiento desigual (Figura 6.7).

Si la duplicación es grande puede observarse en el cariotipo, donde aparece un cromosoma más largo de lo normal que puede presentar bandas extras. Durante la meiosis, los cromosomas homólogos de los individuos heterocigotos por duplicación, forman bucles para permitir la alineación correcta de las regiones homólogas (Figura 7.6)

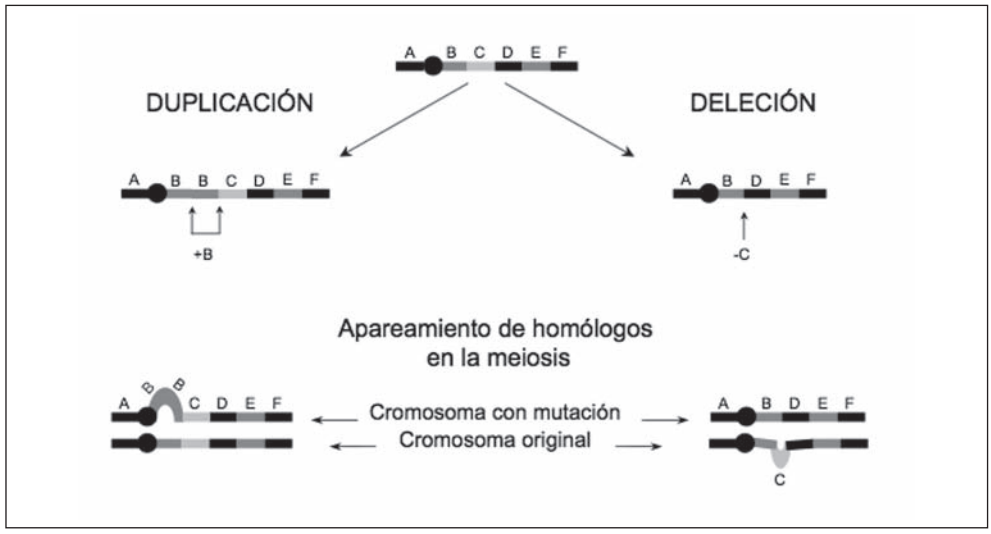
Las duplicaciones pueden ocasionar problemas durante el desarrollo porque provocan desequilibrios entre las cantidades de distintas proteínas. Por otro lado, las duplicaciones son el origen de las familias génicas y son la materia prima que permite la aparición de nuevas funciones.

Deleciones. Una deleción supone la pérdida de un segmento de cromosoma y, como vimos en el capítulo anterior, puede originarse simultáneamente a una duplicación en el cromosoma homólogo por entrecruzamiento desigual (Figura 6.7).

Si la deleción es grande puede observarse en el cariotipo, por la aparición de un cromosoma más corto de lo normal. Durante la meiosis, los cromosomas homólogos de un heterocigoto por deleción forman un bucle para poder aparearse correctamente (Figura 7.6).

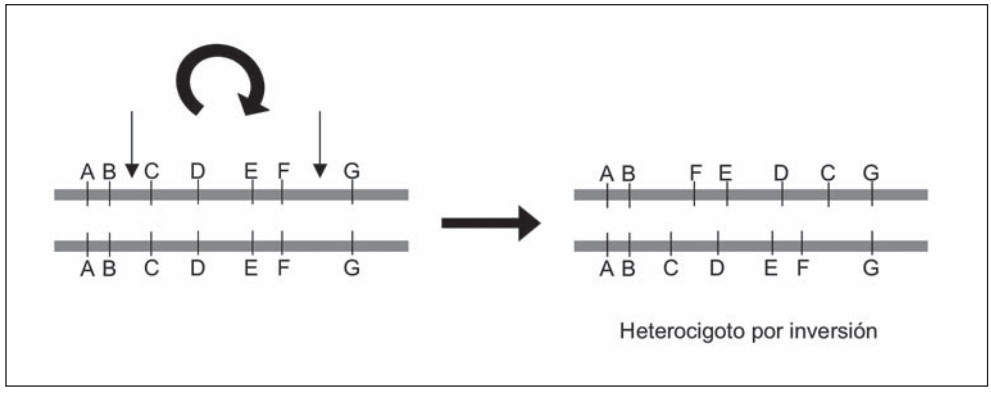
Una deleción, como pasaba con una duplicación, puede acarrear un desequilibrio entre los productos génicos de distintos genes. Además, en un heterocigoto por deleción pueden expresarse los alelos recesivos presentes en el cromosoma intacto, originando fenómenos de pseudodominancia.

Figura 7.6. Duplicaciones y Deleciones cromosómicas



Inversiones. Una inversión se produce por la rotura del cromosoma en 2 regiones y la posterior unión del ADN en sentido inverso. El cromosoma afectado no gana ni pierde genes, sólo presenta una nueva ordenación de éstos (Figura 7.7).

Figura 7.7. Inversión cromosómica



Algunas inversiones pueden generar defectos graves sobre el fenotipo. Por ejemplo, si el punto de rotura de una inversión cae dentro de un gen, éste dejará de ser funcional. En otras ocasiones la inversión provoca la separación del gen de su zona reguladora. En todos estos casos, las inversiones son eliminadas rápidamente de la población.

En las poblaciones naturales de Drosophila son frecuentes las inversiones cromosómicas (Figura 7.8). Los individuos portadores de las distintas ordenaciones cromosómicas por inversión no presentan diferencias fenotípicas aparentes. El principal efecto de las inversiones es la reducción, total o parcial, de la recombinación entre cromosomas homólogos de distinta ordenación.

Figura 7.8. Par de cromosomas homólogos en un heterocigoto para una inversión cromosómica

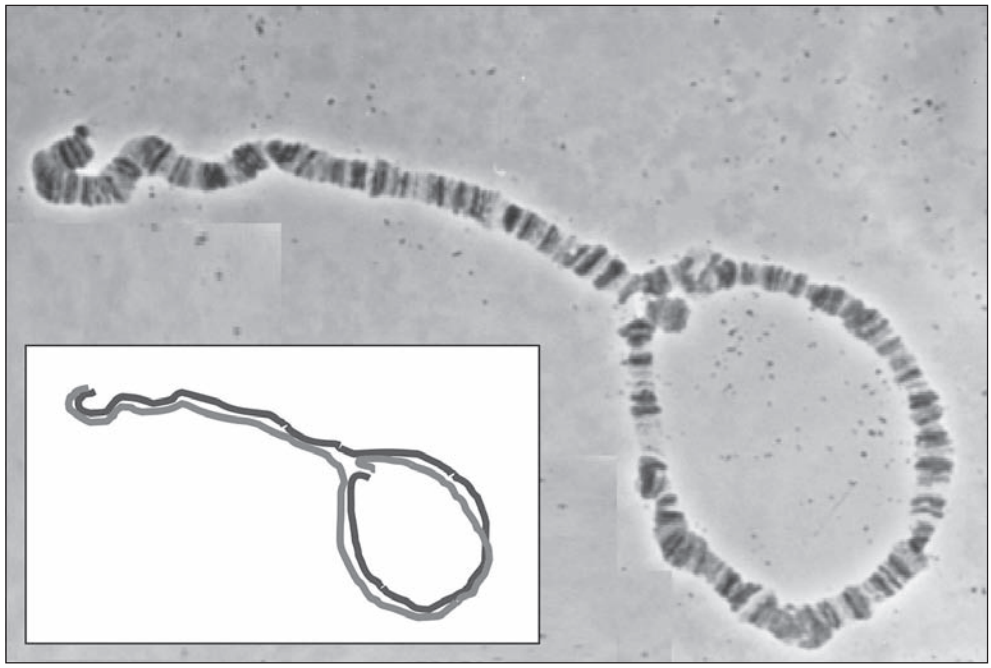


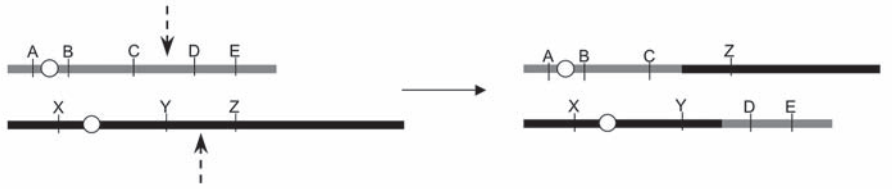
Imagen tomada al microscopio óptico de un par de cromosomas politénicos[30] de *Drosophila subobscura.* Se observa el asa que se forma al aparearse dos cromosomas homólogos con distinta ordenación cromosómica por inversión. El esquema de la izquierda indica cómo se disponen los dos cromosomas homólogos para emparejarse en la zona invertida.

Translocaciones. Una translocación es el intercambio de material genético entre cromosomas no homólogos. El tipo más común es la **translocación recíproca**, dos cromosomas no homólogos experimentan una rotura y a continua-

30. En algunos tejidos de Drosophila existen unos cromosomas especiales, los cromosomas politénicos. Estos cromosomas acumulan copias del ADN replicado sin que la célula se divida. Además, los 2 cromosomas homólogos permanecen unidos. Esto facilita su observación al microscopio óptico y la detección de mutaciones cromosómicas.

ción intercambian los fragmentos acéntricos (Figura 7.9). No obstante, también puede haber **translocación no recíproca**, en que sólo existe paso de ADN en una dirección.

Figura 7.9. Translocación recíproca



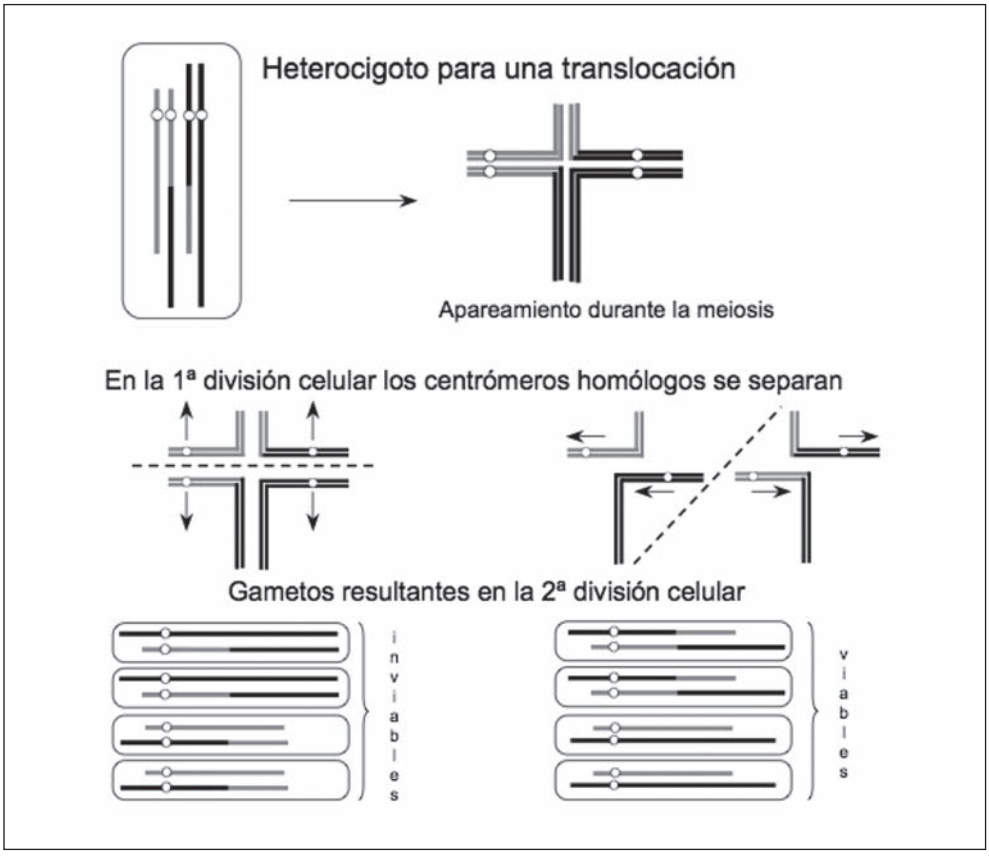
Las traslocaciones pueden afectar al fenotipo por diversas razones. Como sucedía con las inversiones, si la rotura del cromosoma se localiza dentro de un gen, la funcionalidad de éste queda directamente afectada. Por otro lado, las relaciones de ligamiento a otros genes también se alteran, pudiendo afectar a la expresión de los genes. Además, aunque un individuo heterocigoto para una translocación puede ser totalmente normal, la repartición de cromosomas durante la meiosis puede generar gametos con duplicaciones y deleciones (Figura 7.10). La mayoría de estos gametos son inviables pero en caso de pasar a la descendencia pueden causar distintos síndromes.

Los **cambios en el número de cromosomas** pueden ser de dos tipos: pérdida o ganancia de uno o más cromosomas individuales **aneuploidías** o cambio en el número de juegos completos de cromosomas **poliploidías**.

Las **aneuploidías** pueden producirse por diversos fallos durante la repartición de los cromosomas en la meiosis. Si los gametos que forman un cigoto contienen cromosomas extra o les falta algún cromosoma, producirán individuos aneuploides. En humanos, la formación de gametos con un número anómalo de cromosomas parece ser un fenómeno bastante frecuente. Algunos estudios indican que el 30% de las concepciones sufre un aborto espontáneo y, de éstos, el 50% parecen ser debidos a fenómenos de aneuploidía. Algunas aneuploidías son viables pero suelen provocar distintos síndromes. La aneuploidía humana más conocida es la trisomía (presencia de 3 cromosomas homólogos) del cromosoma 21 que provoca el síndrome de Down primario.

La **poliploidía** o presencia de más de 2 juegos de cromosomas en un individuo se origina porque el número de cromosomas no se ha reducido durante la

Figura 7.10. Parte de los gametos de un individuo heterocigoto para una traslocación son inviables porque presentan importantes duplicaciones y deleciones



meiosis. Se pueden encontrar individuos triploides (3n), tetraploides (4n)... La poliploidía es bastante corriente en el reino vegetal y a menudo supone un mecanismo que conlleva fenómenos de especiación. Así mismo, parece que el genoma ancestral de los Vertebrados debió generarse por dos rondas de duplicación (2R) del genoma completo (Dehal y Boore, 2005).

7.2. Evolución

Una de las características inherentes a cualquier entidad viva es su capacidad de cambio. La célula es una unidad dinámica en la que continuamente se están realizando multitud de reacciones bioquímicas. A nivel de individuo, observa-

mos cambios continuos en su aspecto, fisiología y comportamiento. Del mismo modo, las poblaciones experimentan cambios en las características predominantes que presentan los individuos que las componen. A un nivel superior también podemos observar cambios en la biodiversidad del ecosistema y en las relaciones entre los seres que lo componen.

La evolución es el proceso de cambio en la constitución genética de las poblaciones. No debe confundirse la evolución con los cambios que pueda experimentar un individuo en particular a lo largo de su vida. Los cambios en un individuo son el resultado de su desarrollo natural o la respuesta al ambiente cambiante, que no se transmiten a la descendencia.

La unidad evolutiva es la población. En biología, una población es un grupo de individuos de la misma especie que generalmente habitan un área más o menos delimitada y que se reproducen entre ellos compartiendo un mismo acervo genético[31] (en inglés, *gene poo*l).

Para que observemos evolución, la población debe tener diversidad genética. La evolución puede expresarse como el cambio en las frecuencias de los distintos alelos (frecuencias génicas) en la población. El origen de la diversidad genética es la mutación. A partir de la diversidad genética existente en una población, la frecuencia de los distintos alelos puede variar debido a distintas fuerzas entre las que destacan la selección natural y la deriva genética.

7.2.1. Selección natural, la gran idea de Darwin

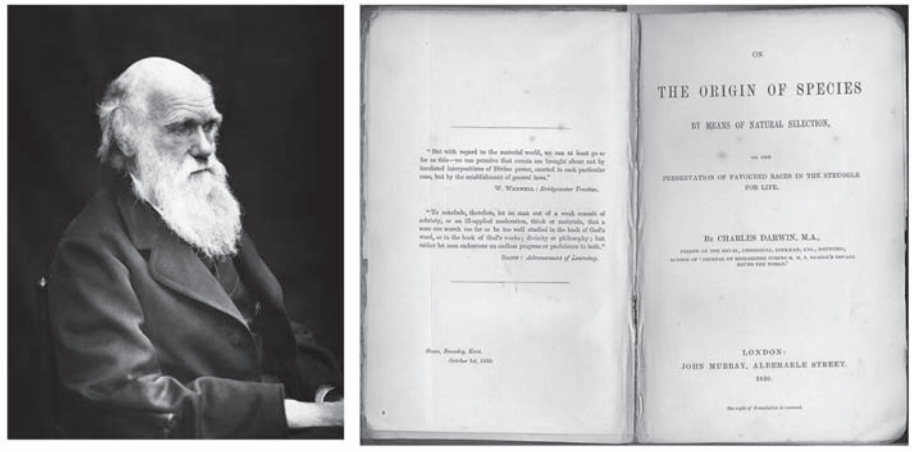
En 1859, Charles Darwin (1809-1882) publica su libro *El origen de las especies*. En éste, Darwin aporta multitud de observaciones que indican que las distintas especies tienen un origen común (han evolucionado a partir de un ancestro común).

«Se comprende fácilmente que un naturalista que aborda el estudio del origen de las especies y que observa las afinidades mutuas de los seres orgánicos, sus relaciones embriológicas, su distribución geográfica, su sucesión geológica y otros hechos aná-

31. El acervo genético es el conjunto completo de alelos distintos presentes en una especie o población.

logos, llegue a la conclusión de que las especies no fueron creadas independientemente unas de otras sino que, como las variedades, descienden de otras especies.» Darwin (1859) *El origen de las especies*.

Figura 7.11. Retrato de Darwin y ejemplar de la primera edición de *El origen de las especies*


ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

Fuente de las imágenes: Wikipedia, foto Darwin tomada por J. Cameron en 1869.

La idea de la evolución de las especies no era una idea nueva. Entre otros, autores como Erasmus Darwin (su abuelo paterno) y Lamarck ya pensaban que las especies aparecían por evolución. Lo novedoso en *El Origen de las Especies* de Darwin es que va más allá y también propone un mecanismo que explica cómo pueden evolucionar unas especies a partir de otras: la **selección natural**.

La idea de la selección natural de Darwin es el resultado de la observación y reflexión profunda durante más de 20 años. Darwin embarcó en el *Beagle* como naturalista, en una expedición que, durante 5 años (1831-1836), lo llevó a dar la vuelta al mundo. Sus experiencias durante el viaje, junto a la lectura de libros de disciplinas muy diversas y sus experimentos en la cría de animales y plantas, hicieron que fuera forjando su idea de la evolución por selección natural. Durante años escribe diversos cuadernos de notas y trabaja en una obra minuciosa que no se decide a concluir. En 1858, Alfred R. Wallace le envía un artículo donde describe la selección natural tal como el propio Darwin la concebía. Aconsejados por C. Lyell, Darwin y Wallace presentan conjuntamente sus artículos ante la *Linnean Society of London* el 1 de julio de 1858. Además, esto será

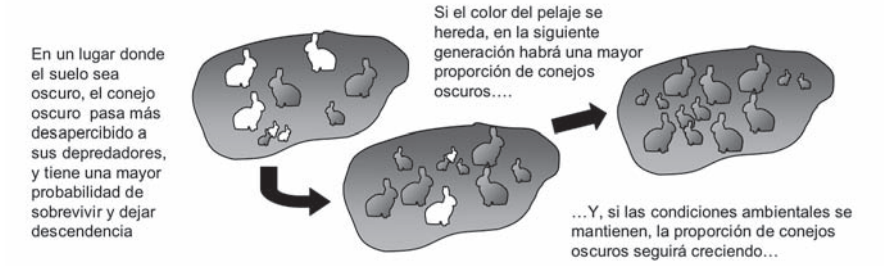
el revulsivo para que Darwin se decida a dar por acabada su gran obra, el *Origen de las especies* que finalmente publicará el 24 de noviembre de 1859.

Darwin fundamenta su idea de la selección natural en 3 observaciones.

- **Abundancia excesiva de descendencia.** Cada generación produce más individuos de los que, con los recursos disponibles, pueden subsistir.
- **Existencia de diversidad.** Los individuos de una especie no son todos iguales para todas las características.
- **Herencia.** Parte de la variación fenotípica es heredable. Algunas características se transmiten a la siguiente generación.

De ahí concluye que debe producirse una **selección de los mejor adaptados.** Puesto que no todos los individuos podrán subsistir, aquellos que tengan alguna ventaja para descubrir y utilizar los recursos o para evitar a los depredadores tendrán más probabilidad de sobrevivir y dejar un mayor número de descendencia. Si las características que los hace más aptos son heredables, en la siguiente generación, la proporción de individuos con la característica beneficiosa aumentará (Figura 7.12).

Figura 7.12. Efecto de la selección natural sobre una población



> «A esta conservación de las variaciones y diferencias individualmente favorables y la destrucción de las que son perjudiciales la he llamado selección natural o supervivencia de los más aptos.» Darwin (1859) *El origen de las especies*.

La idea de evolución fue ampliamente aceptada por los científicos pero el mecanismo de la selección natural encontró bastante oposición. El principal problema estaba en que no se conocía cómo se transmitían los caracteres here-

dables. Ya en el siglo XX, y tras redescubrir las leyes de Mendel, se produjo la integración de una serie de conceptos que dieron lugar al neodarwinismo, también conocido como síntesis moderna o teoría sintética de la evolución[32]. El neodarwinismo se basa en:

- La teoría de la evolución por selección natural de Darwin.
- Las Leyes de Mendel.
- La teoría cromosómica de la herencia. Los caracteres heredables están controlados por fragmentos de ADN que se encuentran en los cromosomas y pueden transmitirse a la descendencia.
- Las mutaciones son cambios al azar del material genético que originan la variabilidad que se encuentra en las poblaciones.
- La genética de poblaciones y el desarrollo de modelos matemáticos que explican cómo varía la composición genética de la población a lo largo del tiempo.

Uno de los argumentos esgrimidos contra la selección natural es que los cambios que genera son tan graduales que sólo podríamos observar su efecto después de muchísimas generaciones. Existen, no obstante, algunos ejemplos en que hemos observado el efecto de la selección natural actuando en pocos años. Quizás el caso más conocido es el de la polilla del abedul, *Biston beularia*, y el fenómeno conocido como melanismo industrial. Esta polilla que durante el día reposa en los troncos de abedules, presenta dos coloraciones, una clara y otra oscura. Gracias a las colecciones de museos, se conoce la proporción relativa de sus dos formas en Inglaterra desde hace más de 200 años. A principios de siglo XIX la forma predominante era la clara pero hacia mediados del siglo XIX empieza a abundar la oscura, hasta el extremo que a finales de siglo, el 99% de los ejemplares eran oscuros. La explicación está en que antes de la revolución industrial, los troncos de los abedules y los líquenes que los cubrían presentaban una coloración clara que proporcionaban un buen camuflaje a las polillas claras, mientras que las polillas oscuras destacaban sobre el fondo claro y eran depredadas rápidamente por los pájaros. Al incrementar la actividad industrial, con la consiguiente contaminación ambiental, los troncos de los abedules ennegrecieron por el hollín. En esta situación las polillas claras destacan y son depredadas mientras que las polillas oscuras quedan camufladas. Lo más interesante del caso es que a partir de los años

32. Algunas de las principales figuras que propiciaron la Síntesis Moderna son: R.A. Fisher, T. Dobzhansky, J.B.S. Haldane, S.Wright, E. Mayr, G. Simpson, y G.L. Stebbins.

60 del siglo xx, con la legislación inglesa anti-polución, los troncos de los árboles empiezan a estar más limpios y poco a poco se ha observado una recuperación de la forma clara de la polilla frente a la oscura.

El ejemplo de *B. betularia* nos ilustra que, aunque el cambio producido por la selección natural puede ser direccional, no persigue una dirección evolutiva predeterminada. La selección natural es oportunista, elige entre las variantes disponibles, aquella mejor adaptada a las condiciones ambientales existentes. Si las condiciones ambientales cambian, la selección natural seguirá eligiendo la variante mejor adaptada, pero muy posiblemente será otra distinta a la seleccionada con anterioridad.

7.2.2. Deriva genética, el factor azar

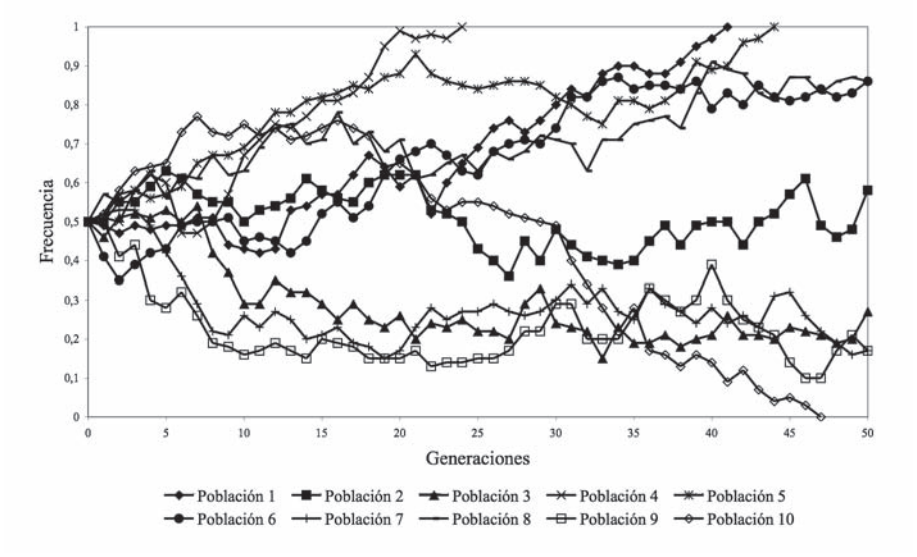
La deriva genética (en inglés, *genetic drift*) es otra de las fuerzas evolutivas que actúan provocando que la población cambie a lo largo del tiempo. Se trata de un cambio aleatorio en las frecuencias génicas de una generación a otra, debido a que los gametos que se unen para formar los individuos de la siguiente generación son una muestra reducida de todos los existentes en la población. Dado que los cambios que genera son aleatorios, no son necesariamente adaptativos. Los efectos de la deriva son más acusados en poblaciones de tamaño pequeño como en el caso de poblaciones que han experimentado una disminución drástica en el número de sus individuos (cuello de botella) o cuando unos pocos individuos colonizan una nueva área (efecto fundador).

Normalmente se produce una pérdida de los alelos menos frecuentes y una fijación de los más frecuentes, provocando una disminución en la diversidad genética de la población. Por otro lado, la tendencia que seguirá la población es impredecible. Si analizamos diversas poblaciones independientes sometidas a deriva genética, cada una fijará, por azar, alelos distintos (Figura 7.13). En otras palabras, se observa una disminución de la variabilidad dentro de la población y un aumento de la diferenciación genética entre poblaciones.

La **teoría neutralista de la evolución**[33] afirma que la mayor parte de la variación molecular existente en la población es neutra (no está sujeta a la selección natural). Por tanto, el principal motor de cambio evolutivo de la población debe ser la deriva genética.

33. El principal artífice de la Teoría neutralista fue Mooto Kimura (1924-1994).

Figura 7.13. Simulación del efecto de la deriva genética sobre la frecuencia del alelo *A* en 10 poblaciones independientes. En el ejemplo, cada población está formada por 50 individuos e inicialmente presentaban 2 alelos distintos a frecuencia 0,5. Después de 50 generaciones, se observa que el alelo *A* se fijó en 3 poblaciones, se perdió en 1 y se mantiene a frecuencias intermedias en las otras 6.



7.3. Genética de poblaciones

La genética de poblaciones es la rama de la genética que estudia la composición genética de las poblaciones (sus frecuencias génicas y genotípicas) y cómo varía esta composición genética a lo largo de las generaciones. Las **frecuencias génicas** se refieren a las frecuencias de los distintos alelos de un locus determinado y las **frecuencias genotípicas** a las frecuencias de los distintos genotipos posibles a partir de estos alelos.

Como ya hemos apuntado anteriormente, la unidad evolutiva es la población. De ahí la importancia de comprender cómo actúan los distintos factores que modelan la composición genética de la población. Para ello, la genética de poblaciones ha desarrollado un importante cuerpo teórico-matemático partiendo de la hipótesis nula en la que no tiene lugar ningún cambio.

7.3.1. Equilibrio Hardy-Weinberg

En 1908, Hardy (matemático británico) y Weinberg (médico alemán) aplicando las leyes de Mendel demostraron independientemente que en una población grande con reproducción sexual y apareamiento aleatorio sobre la que no actúe ningún mecanismo evolutivo (mutación, migración, selección o deriva), las frecuencias génicas y genotípicas se mantienen constantes de generación en generación. Esta constancia en las frecuencias génicas y genotípicas se conoce como ley de Hardy-Weinberg (Stern, 1943) y es la base del desarrollo de la genética de poblaciones.

La Ley de Hardy-Weinberg parte de los siguientes supuestos:

- Los organismos son diploides con reproducción sexual y generaciones discretas.
- El apareamiento entre los individuos es al azar.
- La población es suficientemente grande como para despreciar el efecto de la deriva genética.
- La población no experimenta mutación, migración ni selección.

En este escenario, consideremos un locus con dos alelos A_1 y A_2 presentes en la población con unas frecuencias $f(A_1) = p$ y $f(A_2) = q$, de modo que $p + q = 1$. Podemos obtener las frecuencias genotípicas de la siguiente generación usando un cuadro de Punnett como:

		Hembras	
		$A_1\,(p)$	$A_2\,(q)$
Machos	$A_1\,(p)$	$A_1A_1\,(p^2)$	$A_1A_2\,(pq)$
	$A_2\,(q)$	$A_1A_2\,(pq)$	$A_2A_2\,(q^2)$

$$f(A_1A_1) = P = p^2$$
$$f(A_1A_2) = H = 2pq$$
$$f(A_2A_2) = Q = q^2$$
$$P + H + Q = 1$$

A partir de las frecuencias genotípicas podemos calcular las frecuencias génicas de esta generación:

$$p' = (P + 1/2\,H) = p^2 + pq = p\,(p+q) = p$$
$$q' = (Q + 1/2\,H) = q^2 + pq = q\,(q+p) = q$$

Comprobando que son las mismas que en la generación anterior.

La ley de Hardy-Weinberg muestra que la reproducción sexual es un factor conservador que, por sí solo, no altera las frecuencias génicas y que las frecuencias genotípicas se estabilizan en las proporciones $p^2 : 2pq : q^2$, tras una generación de apareamiento aleatorio.

Ejemplo:

Fundamos una población con 100 moscas con la siguiente proporción: 75 *AA*, 10 *Aa* y 15 *aa*. (frecuencias genotípicas $P = 0{,}75 : H = 0{,}10 : Q = 0{,}15$). Las frecuencias génicas podemos calcularlas de dos formas:

p = alelos *A*/ total alelos = ((75x2)+10)/200 = 0,8 o $p = P + ½\ H = 0{,}75 + 0{,}05 = 0{,}8$

q = alelos *a*/ total alelos = ((15x2)+10)/200 = 0,2 o $q = Q + ½\ H = 0{,}15 + 0{,}05 = 0{,}2$

Las frecuencias genotípicas en la siguiente generación serán $p^2 : 2pq : q^2$,

$P' = 0{,}64 : H' = 0{,}32 : Q' = 0{,}04$.

A partir de las que podremos calcular las frecuencias génicas como:

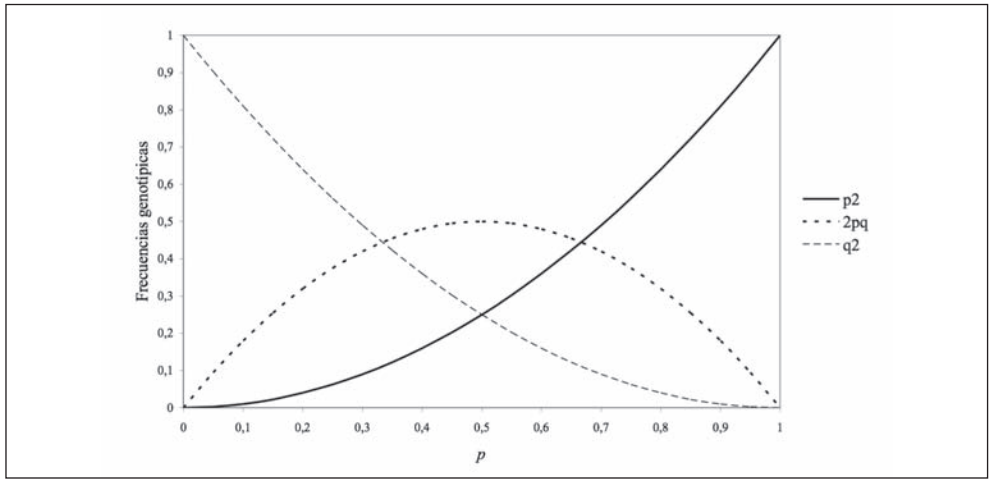
$p' = P' + ½\ H' = 0{,}64 + 0{,}16 = 0{,}8$
$q' = Q' + ½\ H' = 0{,}04 + 0{,}16 = 0{,}2$

Comprobando que no han cambiado respecto a la generación anterior. En la siguiente generación, las frecuencias genotípicas tampoco variarán puesto que sólo dependen de las génicas.

Cuando, en una población, las frecuencias génicas no varían de generación en generación y los genotipos están en una proporción $p^2 : 2pq : q^2$, decimos que la población está en equilibrio Hardy-Weinberg. La Figura 7.14 muestra cómo varían las frecuencias genotípicas en función de las génicas en una población en equilibrio Hardy-Weinberg. En este gráfico podemos ver que cuando la frecuencia de un alelo es baja, la mayoría de las copias de este alelo se encuentran en los heterocigotos.

Por ejemplo: cuando la frecuencia del alelo *A* es de 0,1 la frecuencia del homocigoto *AA* será de 0,01 mientras que la del heterocigoto será de 0,18. También podemos ver que la mayor proporción de heterocigotos se obtiene cuando los dos alelos son equifrecuentes ($p = 0{,}5$).

Figura 7.14. En una población en equilibrio, las frecuencias genotípicas quedan definidas por las frecuencias génicas



La evolución es el proceso de cambio en la constitución genética de las poblaciones, y podemos observar su efecto a través del cambio en las frecuencias génicas. Hemos visto que una población en equilibrio Hardy-Weinberg mantiene su acervo genético estable y, por tanto, no experimenta evolución. Sin embargo, debe recordarse que el equilibrio Hardy-Weinberg se consigue sólo bajo unas condiciones ideales que raramente se cumplen.

En genética de poblaciones, la ley de Hardy-Weinberg sirve para dos propósitos. Por un lado, comprobando si la población está en equilibrio Hardy-Weinberg sabremos si la población evoluciona o es estable. Por otro lado, a partir de su formulación podemos incorporar el efecto teórico de las distintas fuerzas evolutivas para predecir cómo será el cambio que provoquen en las poblaciones. Las desviaciones del equilibrio Hardy-Weinberg pueden ser debidas a 3 tipos de sucesos:

a) Sucesos sistemáticos, de los que podemos predecir tanto la magnitud como la dirección del cambio. Incluye a mutación, migración y selección.
b) Sucesos dispersivos, de los que podemos predecir la magnitud del cambio pero no la dirección. Se refiere a la deriva genética.
c) Sucesos no recurrentes, no podemos predecir ni la magnitud ni la dirección del cambio. Se trata de sucesos raros (poco frecuentes) en la historia de la población, como la aparición de mutaciones raras, hibridación, cuellos de botella…

A continuación veremos cómo afectan a la formulación de Hardy-Weinberg las dos fuerzas principales de evolución: la selección natural (como suceso sistemático) y la deriva (un suceso dispersivo).

7.3.2. Desviaciones del equilibrio

Tanto mutación como migración son dos factores que introducen variación en la población. Aunque la introducción de nuevos alelos se mantenga en cada generación (sucesos sistemáticos), los cambios que generan tanto mutación como migración en las frecuencias génicas suelen ser muy lentos. La selección natural es otro suceso sistemático cuyo efecto sobre las frecuencias génicas es mucho más acusado.

Para analizar el efecto de la selección natural utilizaremos dos variables relacionadas: la eficacia biológica (*w*) y el coeficiente de selección (*s*). La eficacia biológica (en inglés, *fitness*) es el éxito reproductivo relativo de un genotipo. El coeficiente de selección es la intensidad relativa de selección contra un genotipo $s = 1 - w$. El adjetivo «relativo» en estas dos definiciones se refiere a que su valor se calcula en relación al genotipo más eficaz. Por ello, ambas variables varían entre 0 y 1 aunque en sentido inverso. Veámoslo en el siguiente ejemplo:

Genotipo	A_1A_1	A_1A_2	A_2A_2
Nº medio de descendientes viables	4	10	9
w	$w_{11} = 4/10 = 0,4$	$w_{12} = 10/10 = 1$	$w_{22} = 9/10 = 0,9$
s	$s_{11} = 1 - 0,4 = 0,6$	$s_{12} = 1 - 1 = 0$	$s_{22} = 1 - 0,9 = 0,1$

En general, para calcular el efecto de la selección natural partiremos de una tabla que nos indica las frecuencias iniciales y sus eficacias relativas:

Genotipo	A_1A_1	A_1A_2	A_2A_2	Total
Frecuencia inicial	p^2	$2pq$	q^2	1
w	w_{11}	w_{12}	w_{22}	
Contribución a la siguiente generación	$p^2\, w_{11}$	$2pq\, w_{12}$	$q^2\, w_{22}$	$p^2\, w_{11} + 2pq\, w_{12} + q^2\, w_{22} = T$ $(T < 1)$
Frecuencia normalizada	$p^2\, w_{11} / T$	$2pq\, w_{12} / T$	$q^2\, w_{22} / T$	1

A partir de aquí podemos calcular la frecuencia del alelo A_1 en la siguiente generación p' como:

$$p' = \frac{p^2 w_{11} + pqw_{12}}{p^2 w_{11} + 2pqw_{12} + q^2 w_{22}} \qquad \text{y} \qquad q' = 1 - p'$$

y calcular la magnitud del cambio de las frecuencias génicas en una generación cómo:

$$\Delta p = p' - p = \frac{pq\left(p\left(w_{11} - w_{12}\right) + q\left(w_{12} - w_{22}\right)\right)}{p^2 w_{11} + 2pqw_{12} + q^2 w_{22}}$$

El resultado de la selección será distinto según cómo sean las magnitudes relativas de las eficacias biológicas de los distintos genotipos. No obstante, tanto la magnitud como la dirección del cambio quedan determinados por los valores de eficacia biológica de los distintos genotipos.

Por comodidad del cálculo, la eficacia biológica suele expresarse en función de *s*. Veámoslo en el siguiente ejemplo de selección en contra de uno de los homocigotos:

Genotipo	A_1A_1	A_1A_2	A_2A_2	Total
Frecuencia inicial	p^2	$2pq$	q^2	1
w	1-*s*	1	1	
Frecuencia tras la selección	$p^2\,(1\text{-}s)$	$2pq$	q^2	$1 - (p^2 s)$

$$q' = \frac{q^2 + pq}{p^2(1-s) + 2pq + q^2} = \frac{q(q+p)}{p^2 - p^2 s + 2pq + q^2} = \frac{q}{1 - p^2 s}$$

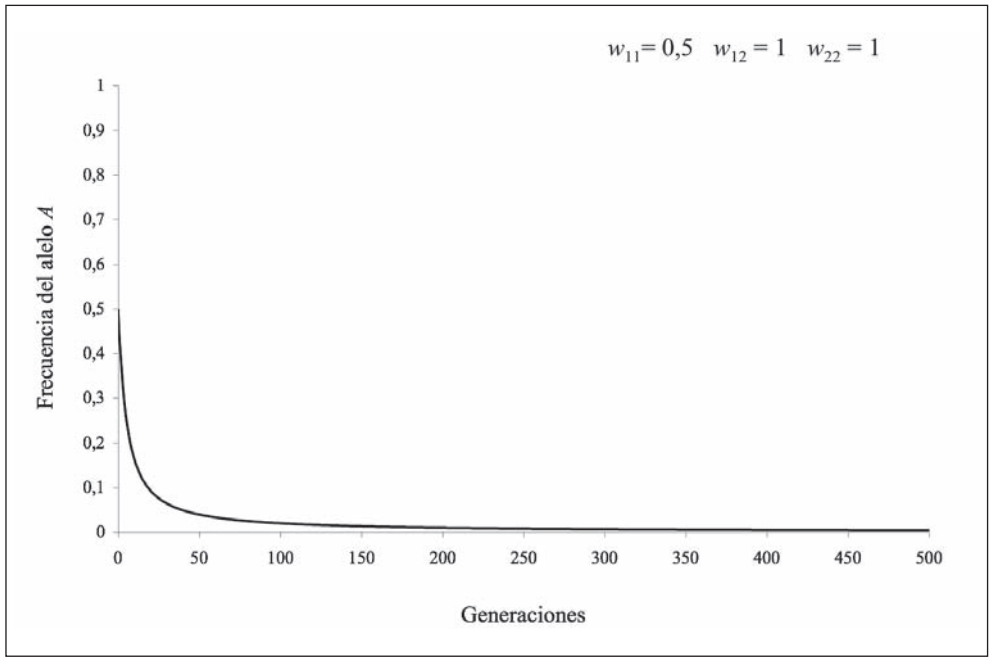
$p' = 1 - q'$

de ahí el cambio experimentado en la generación siguiente será:

$$\Delta q = q' - q = \frac{q}{1 - p^2 s} - q = \frac{q - \left(1 - p^2 s\right)q}{1 - p^2 s} = \frac{q - q + qp^2 s}{1 - p^2 s} = \frac{qp^2 s}{1 - p^2 s}$$

En este caso, el equilibrio ($\Delta q = 0$) se alcanza cuando el numerador es 0 (no hay selección o uno de los alelos ha desaparecido). Por tanto, si no cambia la dirección de la selección, la frecuencia del alelo A_2 irá aumentando en cada generación hasta que el alelo A_1 quede eliminado (Figura 7.15).

Figura 7.15. Efecto de realizar selección en contra del homocigoto A_1A_1



La selección biológica puede conducir a la eliminación de uno de los alelos pero en algunos casos, como en el de la selección a favor del heterocigoto, se llega a un estado de equilibrio (Figura 7.16). Un ejemplo bien conocido de selección a favor del heterocigoto es el de la anemia falciforme[34] en poblaciones humanas africanas afectadas por malaria (Allison, 1954). Los homocigotos para el alelo de la anemia falciforme padecen una serie de patologías graves que provocan que tengan una eficacia biológica muy baja. Los heterocigotos, en cambio sobreviven mejor a la malaria que los homocigotos para el alelo normal. En esta situación se alcanza un equilibrio dirigido por la acción de la selección natural.

34. La herencia y fenotipo de la anemia falciforme ya se comentó en el apartado 5.1 sobre Genética clásica (ver Tabla 5.2).

Una vez se ha alcanzado el equilibrio, las frecuencias génicas permanecen constantes a lo largo de las generaciones. Veámoslo:

Genotipo	A_1A_1	A_1A_2	A_2A_2	Total
Frecuencia inicial	p^2	$2pq$	q^2	1
w	$1–s_1$	1	$1–s_2$	
Frecuencia tras la selección	$p^2\ (1–s_1)$	$2pq$	$q^2\ (1–s_2)$	$1–p^2s_1–q^2s_2$

$$p' = \frac{p^2(1-s_1)+pq}{p^2(1-s_1)+2pq+q^2(1-s_2)} = \frac{p^2(1-s_1)+pq}{1-p^2s_1-q^2s_2}$$

y desarrollando la formula $\Delta p = p' - p$ podemos encontrar que las frecuencias en el equilibrio ($\Delta p = 0$) sólo dependen de los coeficientes de selección:

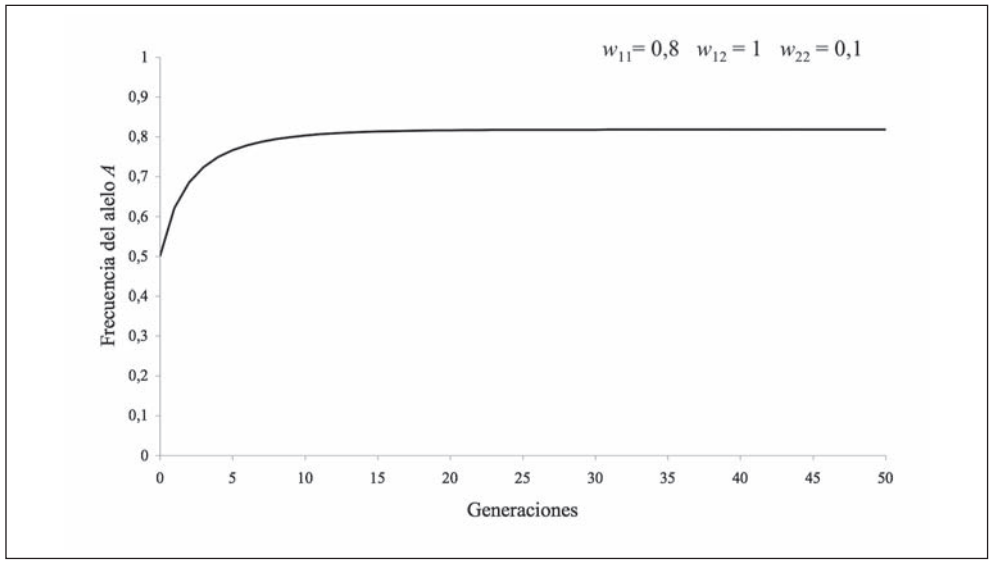
$$p = s_2 / (s_1 + s_2) \qquad \text{y} \qquad q = s_1 / (s_1 + s_2)$$

Considerando que $s_1 = 0{,}2$ y $s_2 = 0{,}9$ (datos para la anemia falciforme, tomados de Dobzhansky et al, 1983) se alcanza un equilibrio con las frecuencias de:

$$p = 0{,}9 /(0{,}2 + 0{,}9) = 0{,}82 \qquad \text{y} \qquad q = 0{,}2 / (0{,}2 + 0{,}9) = 0{,}18$$

como podemos ver en la figura 7.16.

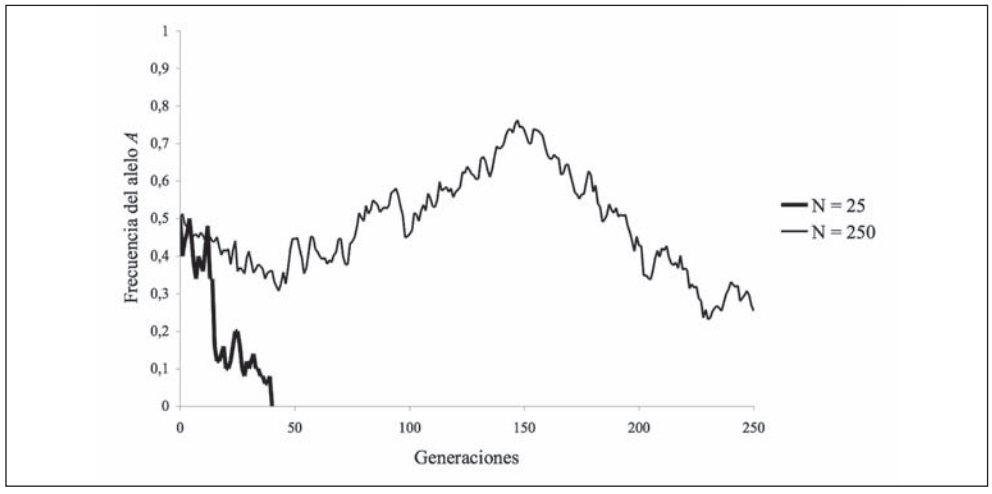
Figura 7.16. Equilibrio obtenido al realizar selección a favor del heterocigoto



En ocasiones, asignamos arbitrariamente una eficacia de 1 a un determinado genotipo. Entonces, la eficacia de los otros genotipos dependerá de su relación con la de éste como 1–*s* o 1+*s*. Un valor de *s* negativo indica una disminución en la eficacia biológica (desventaja selectiva) respecto al genotipo de referencia. En cambio, un valor positivo de *s* indica un aumento de la eficacia (ventaja selectiva) respecto al genotipo de referencia.

El otro gran motor de la evolución es la deriva genética. Éste es un fenómeno dispersivo, no podemos predecir su dirección puesto que actúa al azar debido al error de muestreo en el momento de elegir los gametos que generarán la siguiente generación. Imaginemos una población con *N* individuos diploides. Cada individuo forma gametos en gran cantidad pero en el conjunto de la población sólo se toman 2*N* gametos para crear la siguiente generación de *N* individuos. Si la frecuencia del alelo *A* es *p*, esperamos que en la siguiente generación su frecuencia sea también *p* pero con un error de muestreo de $\sqrt{[pq/2N]}$. Esto nos indica que cuanto menor sea el tamaño de la población (*N*) mayor será la desviación respecto a la frecuencia inicial (Figura 7.17).

Figura 7.17. La magnitud del cambio por deriva genética sólo depende del tamaño de la población



También significa que si consideramos diversas poblaciones con las mismas frecuencias iniciales *p*, *q* y tamaño poblacional *N*, la frecuencia media de los dos alelos en el conjunto de poblaciones se mantienen como *p* y *q*, aunque cada población habrá divergido independientemente pudiendo haber fijado o eliminado un alelo distinto (Figura 7.13).

Capítulo VIII

Evolución molecular

La evolución es el cambio en las características genéticas de los individuos que forman una población a lo largo de las generaciones. El estudio de la evolución se inició con la observación a nivel morfológico de la variación entre los individuos de una población **(polimorfismo)** o las diferencias entre individuos de distintas especies **(divergencia)**. Esta aproximación morfológica al estudio de la evolución obedecía a que no se disponía de las herramientas para observar otro tipo de variabilidad. Sin embargo, ésta no es la mejor aproximación posible. Por un lado, la variabilidad morfológica, aunque responda mayoritariamente a factores genéticos también está modelada en parte por factores ambientales y, por lo tanto, no se transmitirá totalmente a la descendencia. Por otro lado, tanto la elección de los caracteres a considerar, como la cuantificación de las diferencias entre individuos están sujetas a criterios más o menos subjetivos.

A partir de la década de los años sesenta del siglo XX, la obtención cada vez más generalizada de datos de variación molecular, permitió iniciar el estudio de la evolución a nivel molecular. Ambos niveles de variabilidad (morfológica y molecular) están relacionados. La variabilidad morfológica responde en última instancia a la variabilidad de las proteínas y ésta, a la del ADN. La variabilidad molecular, sin embargo, tiene la ventaja de ser totalmente heredable y permitir un tratamiento objetivo de las diferencias observadas.

El estudio de la variabilidad molecular llevó a Mooto Kimura a enunciar la teoría neutralista de la evolución molecular, según la cual el principal motor de la evolución es la deriva genética. El desarrollo de esta teoría ha permitido un gran avance en el estudio de la evolución molecular tanto a nivel poblacional como filogenético.

Como iremos viendo, para el estudio de la evolución molecular se hace imprescindible el uso de numerosas herramientas informáticas. Desde el alineamiento de secuencias, a la simulación por coalescencia, pasando por el cálculo de estadísticos que rastrean las diferencias entre múltiples secuencias son tareas

que no se hubieran podido emprender sin la asistencia de programas informáticos.

8.1. Substitución génica

La substitución génica se define como el proceso por el que una variante alélica sustituye al alelo predominante (o salvaje) en una población. En este proceso, una variante nueva aparece por mutación en una única copia y acaba fijándose después de un número de generaciones más o menos grande.

Como hemos visto en el capítulo VII **(Variabilidad genética y evolución)** tanto selección como deriva suelen llevar a la fijación de un alelo, pero no todas las nuevas mutaciones acaban fijándose. Las mutaciones deletéreas serán, en general, eliminadas rápidamente por la selección purificadora, pero también las neutras e incluso las beneficiosas se pueden perder fácilmente por azar durante las primeras generaciones, en que su frecuencia es muy baja. Por otro lado, mientras que las mutaciones beneficiosas tenderán a fijarse por la acción de la selección positiva, cualquier mutación (incluso una fracción de las deletéreas) puede fijarse por efecto de la deriva (Tabla 8.1). Por ello debemos considerar la probabilidad de fijación, *P*.

Tabla 8.1. Para cada tipo de mutación, posibilidad (+) o no (–) de acabar fijándose por la acción de la deriva o de la selección. En la columna *Factor* se muestra el principal factor que rige el destino de cada tipo de mutación.

	Deriva	Selección	Factor
Mutación neutra	+	–	Equilibrio mutación-deriva
Mutación beneficiosa	+	+	Selección positiva
Mutación deletérea	(+)*	–	Selección purificadora

* Sólo una pequeña fracción de las mutaciones deletéreas llegará a fijarse por acción de la deriva.

La probabilidad de fijación de un alelo A_2 depende de su frecuencia inicial ($f(A_2)$), de su ventaja o desventaja selectiva (s) y del tamaño efectivo de la población (N_e)[35]. Para simplificar el problema, consideraremos que $N_e = N$, siendo N el número de individuos de la población.

35. El tamaño efectivo de la población, N_e, viene determinado por el número de individuos reproductores.

En una población formada por N individuos diploides tenemos $2N$ alelos. Así, cualquier mutación que aparezca de nuevo tendrá una frecuencia inicial de $1/(2N)$. Si la mutación es neutra, su fijación dependerá sólo del azar (deriva genética) y por tanto su probabilidad de fijarse es igual a su frecuencia:

$$P = 1/(2N)$$

En caso de una mutación seleccionada positivamente ($s > 0$) y en poblaciones grandes, se llega a la expresión:

$$P = 2s$$

Por otro lado, la substitución génica es un proceso continuo. Como en la población aparecen continuamente nuevas mutaciones, una mutación puede substituir a otra y más tarde también podrá ser sustituida. Por ello, podemos hablar de una tasa de substitución génica, k, como el número de mutaciones que se fijan por unidad de tiempo. Si las nuevas mutaciones aparecen a una velocidad de u por gen por generación, en toda la población tendremos $2Nu$ mutaciones por locus por generación y la tasa de substitución será:

$$k = 2NuP$$

Consideremos primero las mutaciones neutras. Substituyendo el valor de P tenemos que la tasa de substitución es igual a la tasa de mutación:

$$k = 2NuP = 2\ Nu(1/[2NJ) = u$$

Si consideramos las mutaciones beneficiosas, tenemos que la tasa de substitución es:

$$k = 2NuP = 2\ Nu\ (2s) = 4Nsu$$

Otro aspecto a tener en cuenta al considerar la substitución génica es el tiempo necesario para alcanzar la fijación. Este tiempo depende de la frecuencia inicial de la mutación (q_0), del tamaño poblacional y de si está, o no, sometida a selección. En el caso de una nueva mutación, $q_0 = 1/2N$. Kimura y Otha (1969) calcularon que el tiempo medio necesario para alcanzar la fijación de una mutación neutra es $\bar{t} = 4N$ generaciones y en el caso de una mutación sometida a un coeficiente de selección s, el tiempo necesario es $\bar{t} = (2/s)\ \ln(2N)$ generaciones.

Supongamos una población de *Drosophila* de tamaño poblacional 10^6. Una mutación neutra tardará por término medio $\bar{t} = 4N = 4 \times 10^6$ generaciones en fijarse. Si consideramos que la mutación experimenta selección positiva con $s = 0{,}005$, tardará, por término medio, $\bar{t} = (2/s)\ln(2N) = 5.803$ generaciones en fijarse.

Durante este tiempo han podido aparecer otras nuevas mutaciones. Muchas desaparecerán por azar mientras permanecen a baja frecuencia, pero otras pueden incrementar su frecuencia e iniciar un nuevo proceso de substitución. Desde que aparece una nueva mutación en un locus y hasta que no se alcanza su fijación o su pérdida, en la población coexisten dos o más variantes y podemos decir que el locus presenta polimorfismo.

8.1.1. El reloj molecular

En la década de los años sesenta del siglo XX se obtuvieron las secuencias de aminoácidos de la hemoglobina y del citocromo *c* de diversos organismos. Comparando la divergencia observada (*K*, número de diferencias aminoacídicas por posición) con el tiempo de divergencia entre los linajes (*t*, inferido a partir del registro fósil), se observó que cuanto más tiempo de divergencia separa a dos linajes determinados, más diferencias acumulan sus secuencias de aminoácidos y que la relación entre *K* y *t* era una relación lineal:

$$K/t = \alpha$$

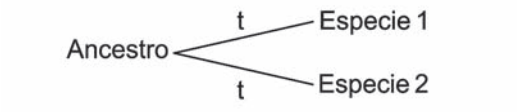
Esto llevó a Zuckerkandl y Pauling[36] a proponer que, para una proteína determinada, la tasa de evolución molecular es aproximadamente constante en todos los linajes o, en otras palabras, que las moléculas se comportan como un reloj molecular. La idea del reloj molecular tiene interesantes implicaciones en el estudio de la evolución. Conociendo el ritmo del reloj y la divergencia entre dos especies podríamos inferir el tiempo transcurrido desde su divergencia.

$$t = K/\alpha$$

36. Linus C. Pauling (1901-1994) fue un físico y químico norteamericano considerado uno de los científicos más influyentes del siglo xx. En 1954 recibió el Premio Nobel de Química por su contribución al conocimiento de los enlaces químicos y en 1962 recibió el Premio Nobel de la Paz.

Entonces, la tasa de substitución (*k*) puede calcularse como la divergencia observada entre dos especies (*K*) dividida por dos veces el tiempo (*t*) transcurrido desde su divergencia. Cada especie ha tenido un tiempo *t* de divergir independientemente, por lo que la divergencia que observamos corresponde a la suma de los cambios acumulados en los dos linajes desde su separación.

$$k = K/2t$$



Por otro lado, aunque las proteínas evolucionan a un ritmo constante, cada proteína lo hace a un ritmo distinto (cada proteína es un reloj distinto). Estas diferencias se explican porqué unas proteínas tienen mayor limitación funcional que otras. La limitación funcional de una proteína viene determinada por la proporción de aminoácidos que pueden cambiar libremente sin que la proteína pierda su función. A mayor limitación, menos aminoácidos pueden cambiar. La mayoría de los cambios que alteran la función son deletéreos y, por tanto, serán eliminados rápidamente por la selección purificadora. Las proteínas con mayor limitación funcional experimentarán una mayor proporción de mutaciones deletéreas, con lo que observaremos una menor tasa de substitución.

8.1.2. Teoría neutralista de Kimura

La observación del reloj molecular es uno de los pilares que hicieron que Mooto Kimura (1924-1994) llegara a la conclusión de que la mayoría de las mutaciones que se fijan deben ser mutaciones neutras. Cuando las mutaciones que se fijan son neutras, para observar una tasa de substitución ($k = u$) constante, basta con que la tasa de mutación sea constante. En cambio, si la mayoría de mutaciones que se fijan son beneficiosas, para que la tasa de substitución ($k = 4Nsu$) sea constante, además de la tasa de mutación, también debe ser constante el producto Ns.

El otro pilar sobre el que Kimura apoya su teoría es el elevado polimorfismo existente en las poblaciones. Los datos moleculares a nivel poblacional mostraban que el nivel de polimorfismo en las poblaciones era mucho más elevado de lo que se esperaba a partir de los datos morfológicos. Sin embargo, los modelos

teóricos a partir de la teoría selectiva, que ven el polimorfismo como un estado estático, no permiten explicar la existencia de tanta variabilidad en las poblaciones. Kimura reinterpreta el polimorfismo como un estado transitorio que acabará con la fijación al azar de una de las variantes neutras.

Kimura, sin negar la importancia de la selección, pone de manifiesto que la deriva genética debe tener un papel muy importante en la evolución (Kimura, 1968, 1969). El desarrollo de las ideas de Kimura se conoce como teoría neutralista de la evolución molecular.

> «Ninguno de los argumentos anteriores pretende implicar que la selección natural no sea importante en la evolución. Lo que hemos postulado es que, acompañando a los cambios adaptativos que ocurren por selección, hay una gran cantidad de ruido aleatorio debido a cambios casi-neutros al azar.» Kimura, (1969).

A menudo se han contrapuesto las teorías neodarwinista (selección) *versus* neutralista (deriva). Sin embargo, ninguna de las dos teorías niega la acción tanto de la selección como de la deriva. La controversia radica en la importancia relativa que se otorga a una u otra fuerza evolutiva. Por otro lado, la teoría neutralista proporciona una hipótesis nula que permite testar el posible efecto de la selección natural.

8.2. Estimas de la substitución nucleotídica

Como ya hemos visto, la substitución génica es el proceso por el que un alelo nuevo, aparecido en la población por mutación, va aumentando su frecuencia y desplaza el alelo salvaje hasta fijarse. Las diferencias entre los distintos alelos estriban en su secuencia de nucleótidos. Así, podemos considerar la substitución génica como la substitución de una secuencia por otra. Además, podemos considerar que el cambio en la secuencia es un proceso continuo. Estudiando cómo cambia una secuencia a lo largo del tiempo podemos estimar el número de sustituciones entre dos secuencias que nos permitirá determinar su distancia genética.

Para analizar las relaciones entre distintas especies (divergencia), es necesario conocer la distancia genética que las separa. Un buen estimador de la distancia genética debe estar relacionado con el tiempo. A partir de las observaciones del reloj molecular, se pensó que un buen estimador de la distancia genética podría

ser el número de sustituciones por posición nucleotídica (*K*), ya que su relación con el tiempo es constante.

Para medir el valor de *K* se comparan secuencias de nucleótidos homólogas que previamente hemos alineado. La obtención del alineamiento es un paso crucial para la obtención del valor correcto de *K*. Cuanto mayor sea el tiempo que separa a dos secuencias, mayor será el número de diferencias acumuladas entre ellas, dificultando la identificación de las posiciones homólogas. Las diferencias entre las secuencias pueden ser por substitución nucleotídica, pero también por InDels, que dificultan aún más su correcto alineamiento. La complejidad del alineamiento de secuencias es un problema que debe ser resuelto con herramientas informáticas.

Como ya se ha mencionado, cada proteína evoluciona según el ritmo de un reloj distinto. Esto nos permite resolver las relaciones entre especies de tiempos de divergencia muy diversos. Si queremos comparar un grupo de especies que han divergido recientemente, deberemos utilizar proteínas que evolucionen rápidamente; de lo contrario, apenas detectaremos ningún cambio que nos pueda dar información de su evolución. En cambio, si deseamos resolver las relaciones entre especies muy alejadas filogenéticamente, deberemos utilizar proteínas de evolución lenta, de lo contrario no seremos capaces de descubrir similitudes entre las secuencias que nos permitan alinearlas.

Una vez hemos obtenido el alineamiento, podremos medir el número de diferencias por número de posiciones (*p*), descontando las posiciones con InDels (o *gaps*).

Ejemplo:

```
             1         10         20         30         40
secuencia 1  ATCGAGC--GGTATTACAGAGCCTCGGTTGAATCTCGGTA  38nt
             |||  |||   |||  |||||||||  ||  ||||||||||||||
secuencia 2  ATC-AGCTTGGTGTTACAGAGCTTCAGTTGAATCTCGGTA  39nt
```

El alineamiento anterior cuenta con 40 posiciones nucleotídicas. Cada una de las dos secuencias alineadas tienen 38 y 39 pb respectivamente, pero el número de nucleótidos alineados es inferior (37) debido a la existencia de InDels entre las secuencias. Para el cálculo de *p* sólo se consideran las posiciones alineadas sin *gaps*.

p = 3 diferencias /37 posiciones = 0,081

El número de diferencias observadas entre dos secuencias es una subestima del número real de sustituciones que han tenido lugar. Ello se debe a que una misma posición nucleotídica ha podido experimentar sustituciones múltiples (en inglés, *multiple hits*) a lo largo del tiempo y en cualquiera de los dos linajes que han originado las secuencias que estamos comparando (Tabla 8.2).

Tabla 8.2. Relación entre los cambios reales y los observados entre los nucleótidos que ocupan una misma posición en dos secuencias que han divergido a partir de una secuencia ancestral común.

Ancestro → secuencia 1	Ancestro → secuencia 2	Cambios reales	Cambios observados	Tipo de substitución
A → A	A → A	0	0	No hay substitución
A → A	A → C	1	1	Simple
A → C	A → G	2	1	Coincidente
A → A	A → T → G	2	1	Secuencial
A → C	A → C	2	0	Paralela
A → C	A → G → C	3	0	Convergente
A → A	A → C → A	2	0	Reversa

Se han descrito diversos modelos para estimar *K* a partir de *p* (Li, 1997, Graur y Li, 2000). Los dos modelos más sencillos y también más utilizados son los de Jukes y Cantor (1969), y el de Kimura de dos parámetros (Kimura, 1980).

8.2.1. Modelo de Jukes y Cantor

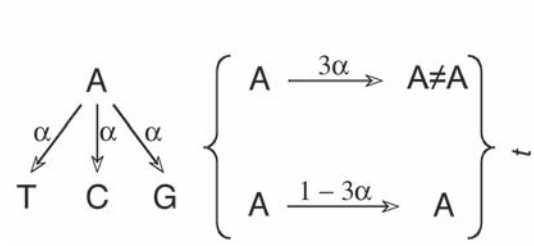
El modelo de Jukes y Cantor (JC) supone que:

- Las distintas posiciones nucleotídicas evolucionan independientemente.
- La tasa de substitución es constante en el tiempo y entre los linajes.
- La probabilidad de cambio de un nucleótido por otro distinto es siempre igual.

Así, este modelo sólo depende de un parámetro α, que es la tasa de substitución de un nucleótido concreto por otro nucleótido concreto por unidad de tiempo.

Primero veamos cómo cambia una secuencia a lo largo del tiempo, según este modelo.

Supongamos que inicialmente ($t = 0$) el nucleótido de una posición determinada en una secuencia de ADN es una A. El nucleótido de esta posición tiene la misma probabilidad (α) de cambiar a C, que a G, que a T. Por tanto, la probabilidad de cambiar es 3α, y la de no cambiar es $1 - 3\alpha$.



	t+1 A	T	C	G
A	1–3α	α	α	α
T	α	1–3α	α	α
C	α	α	1–3α	α
G	α	α	α	1–3α

La probabilidad de que esta misma posición siga ocupada por una A en el momento $t = 1$ es:

$$P_{A(1)} = 1 - 3\alpha$$

La probabilidad de que en el momento $t = 2$ sea una A, será la probabilidad de no cambiar si ya había una A, más la probabilidad de cambiar hacia A si no había una A:

$$P_{A(2)} = (1 - 3\alpha)\, P_{A(1)} + \alpha\, (1 - P_{A(1)})$$

y en general, para t:

$$P_{A(t)} = (1 - 3\,\alpha)\, P_{A(t\text{-}1)} + \alpha\, (1 - P_{A(t\text{-}1)})$$

Esta expresión puede aproximarse a un modelo continuo y, resolviendo una ecuación diferencial de primer orden, se obtiene que:

$$P_{A(t)} = 1/4 + (P_{A(0)} - 1/4)e^{-4\alpha t}$$

Podemos partir de dos situaciones distintas en el momento inicial ($t = 0$):

1. Que la posición estuviera ocupada por una A, con lo que $P_{A(0)} = 1$:

$$P_{AA(t)} = 1/4 + 3/4\ \mathrm{e}^{-4\alpha t}$$

2. Que estuviera ocupada por un nucleótido cualquiera distinto de A (por ejemplo T), con lo que $P_{A(0)} = 0$:

$$P_{TA(t)} = 1/4 - 1/4\ e^{-4\alpha t}$$

Puesto que en el modelo de Jukes y Cantor todos los nucleótidos son equivalentes, podemos generalizar estas probabilidades como:

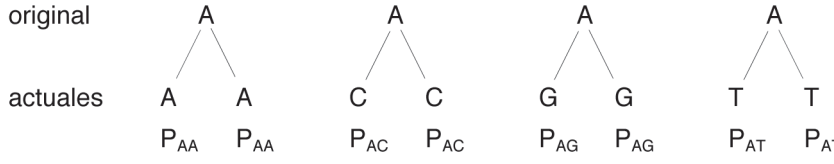
$$P_{ii(t)} = 1/4 + 3/4\ e^{-4\alpha t} \qquad \text{y} \qquad P_{ij(t)} = 1/4 - 1/4\ e^{-4\alpha t}$$

donde *i* y *j* son nucleótidos distintos. Nótese que $P_{ii(t)}$ es la probabilidad de que no haya cambio mientras que $3P_{ij(t)}$ es la probabilidad de cambio para una posición nucleotídica determinada. Cuando $t \to \infty$, la probabilidad de no cambiar es 1/4, y la de cambio 3/4.

A continuación, consideremos cómo divergen dos secuencias a partir de una secuencia ancestral común.

Ahora necesitamos conocer $I_{(t)}$, que es la probabilidad de que en un momento *t* el nucleótido de una posición concreta sea el mismo en las dos secuencias.

Partiendo de una secuencia original donde teníamos una A, la probabilidad que las dos secuencias sean idénticas debe contemplar las siguientes posibilidades:



En la primera situación tenemos que ninguna de las dos secuencias ha experimentado ningún cambio, por lo que consideraremos P_{ii}. En las otras tres situaciones, sí que ha habido cambio, por lo que consideraremos P_{ij}. Así tenemos que:

$$I = P_{ii}^{\,2} + 3P_{ij}^{\,2} = (1/4 + 3/4e^{-4\alpha t})^2 + 3(1/4 - 1/4\ e^{-4\alpha t})^2$$

$$I = 1/4 + 3/4\ e^{-8\alpha t}$$

Puesto que I es el número de posiciones idénticas por posición, $p = 1 - I$

$$p = 1 - (1/4 + 3/4e^{-8\alpha t})$$

de donde podemos fácilmente llegar a la expresión:

$$8\alpha t = -\ln (1 - 4/3p)$$

Recordemos que:

- queremos averiguar la relación entre K (el número de sustituciones por posición) y p (el número de diferencias por posición).
- Y que: $k = K / 2t$.

En el modelo de Jukes y Cantor $k = 3\alpha$, por lo que $K = 6\alpha t$.

Combinando las expresiones:

$$8\alpha t = -\ln (1 - 4/3p) \qquad \text{y} \qquad K = 6\alpha t$$

obtenemos:

$$K = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$$

De este modo obtenemos una estima del número de sustituciones (K) a partir del número de cambios observados (p). Cuando p es muy pequeño, K es muy parecido a p. Cuando p es muy grande, la diferencia entre p y K es muy grande. Cuando $p \geq 0{,}75$, no puede aplicarse la corrección de Jukes y Cantor, puesto que deberíamos calcular el logaritmo de 0 o de un número negativo. Recuérdese que el límite cuando $t \to \infty$ de la probabilidad de cambio era $3/4 = 0{,}75$. Por tanto, para este modelo, no tiene sentido una probabilidad de cambio superior a 0,75.

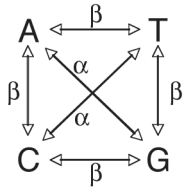
Ejemplo:

p	*K*
0,1	0,10733
0,5	0,82396
0,6	1,20708
0,74	3,23812
≥ 0,75	No aplicable

El modelo de Jukes y Cantor es muy utilizado por su sencillez pero no se ajusta totalmente a la realidad. Por ello, se han desarrollado otros modelos más complejos que intentan explicar mejor la dinámica de la substitución nucleotídica.

8.2.2. Modelo Kimura de dos parámetros

En el capítulo VII **(Variabilidad genética y evolución)** ya comentamos que, aunque existen más posibles cambios por transversión (8) que por transición (4), normalmente se observa un mayor número de cambios por transición. Esto significa que la probabilidad de una transición es mayor que la de una transversión. El modelo propuesto por Kimura (1980) tiene en cuenta esta observación y considera dos parámetros α y β, donde la tasa de substitución por transición es α y la tasa de substitución por transversión es β.



t \ *t*+1	A	T	C	G
A	1−(2β+α)	β	β	α
T	β	1−(2β+α)	α	β
C	β	α	1−(2β+α)	β
G	α	β	β	1−(2β+α)

Con este modelo se obtiene una probabilidad P_{ii} de no cambiar de nucleótido, pero la probabilidad de cambio P_{ij} se obtiene en dos partes: la probabilidad de substitución por transición y la probabilidad de substitución por transversión. De igual modo, los cambios observados (p) se dividen en dos grupos: por transición (P) y por transversión (Q). De modo que $p = P + Q$.

Bajo este modelo, Kimura (1980) obtiene que:

$$K = -\frac{1}{2}\ln\left[(1-2P-Q)\sqrt{1-2Q}\right]$$

y en los libros de texto (Li, 1996, Graur y Li, 2000) solemos encontrar la expresión reorganizada como:

$$K = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right)+\frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

8.2.3. Otros modelos de substitución

El modelo de substitución puede complicarse aún más. El propio Kimura propuso otro modelo de tres parámetros donde desdoblaba la probabilidad de cambio por transversión en dos tipos (A↔T C↔G / A↔C T↔G). Existen otros modelos en los que el número de parámetros puede llegar a 12.

Todos estos modelos predicen que los cuatro nucleótidos serán equifrecuentes en la secuencia. Sin embargo, normalmente no se observa esta equifrecuencia. Así, por ejemplo, las regiones codificadoras son ricas en C y G, mientras que en las regiones intergénicas se observa una mayor abundancia de A y T.

Esta observación hizo que se consideraran modelos que tienen en cuenta la composición de la secuencia, como el de Tajima y Nei (1984). Este modelo es una generalización del JC según el cual tenemos que $K = -b \ln(1 - p/b)$ donde $b = 1 \Sigma q_i^2$, siendo q_i la frecuencia del nucleótido i (i = 1, 2, 3, 4, para A, C, G, T). Se comprueba fácilmente que para el caso de equifrecuencia $q_i = 1/4$, $b = 3/4$ y se obtiene la expresión para Jukes y Cantor.

El modelo de Tamura (1992) considera la corrección de Kimura y el porcentaje de CG. El de Hasegawa, Kishino y Yano (1985) también tiene en cuenta la corrección de Kimura 2 parámetros, pero además considera la proporción de los 4 nucleótidos.

Todos estos modelos consideran que la probabilidad de substitución es uniforme a lo largo de la secuencia. Esta suposición puede ser válida para las regiones no codificadoras, pero no es en nada realista para las regiones codificadoras.

8.2.4. Sustituciones en regiones codificadoras. Método Nei-Gojobori

Los cambios en las regiones codificadoras están limitados por constricciones funcionales de las proteínas para las que codifican. Por un lado, unas proteínas son más variables que otras. Por otro lado, los aminoácidos de algunas regiones de las proteínas son prácticamente invariables entre especies, mientras que otras regiones son mucho más variables.

Además, en las regiones codificadoras podemos distinguir dos tipos de mutaciones: sinónimas y no sinónimas. Las mutaciones sinónimas no provocan ningún cambio de aminoácido, con lo que podemos considerarlas neutras, y su fijación dependerá de la deriva.

En cambio, las mutaciones no sinónimas generan un cambio de aminoácido (que puede afectar a la funcionalidad de la proteína), por lo que, en general, en su fijación, además de la deriva, también interviene la selección natural.

En principio son más probables las mutaciones no sinónimas que las sinónimas. En una secuencia codificadora cualquiera, la mayoría de posiciones son posiciones que al cambiar de nucleótido originan un cambio de aminoácido. Sin embargo, la mayoría de las mutaciones no sinónimas son deletéreas, siendo rápidamente eliminadas por la selección purificadora. Esto hace que las sustituciones sinónimas sean mucho más frecuentes que las no sinónimas. Debemos, por tanto, considerar dos tasas de substitución independientes: la sinónima (k_S) y la no sinónima (k_A).

El método de Nei-Gojobori (1986) es uno de los más utilizados para calcular frecuencias de substitución sinónimas y no sinónimas. El método consiste en realizar los siguientes pasos:

- Cálculo del número de posiciones sinónimas (S) y del número de posiciones no sinónimas (N).
- Recuento del número de diferencias sinónimas (S_d) y del número de diferencias no sinónimas (N_d).
- Obtención de los valores de p_S y p_A.
- Corrección por las posibles sustituciones múltiples (obtención de K_S y K_A).

Tanto para el cómputo del número de posiciones como del de diferencias sinónimas y no sinónimas se excluyen los codones de inicio y de STOP.

Número de posiciones sinónimas y no sinónimas

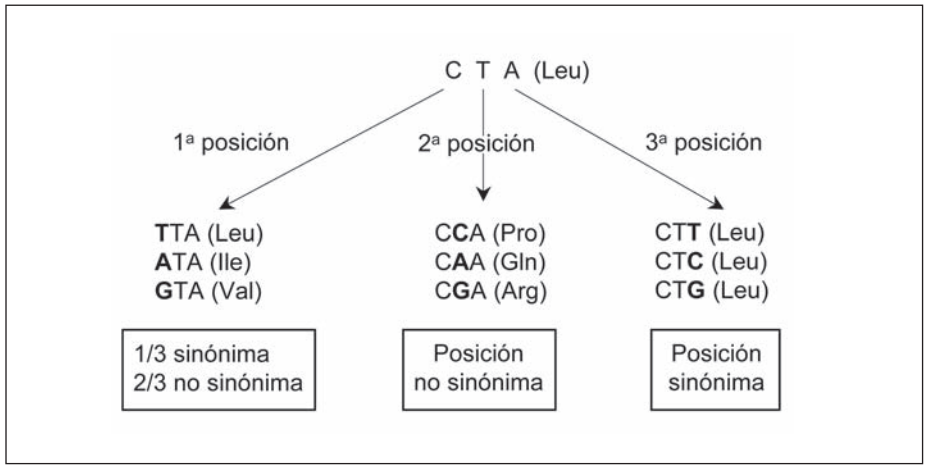
Para el cálculo de las posiciones sinónimas (s) y no sinónimas (n) de una secuencia, debemos recorrerla codón a codón y comprobar si un cambio potencial de nucleótido representaría, o no, un cambio en el aminoácido codificado por el codón.

Tomemos como ejemplo el codón CTA que codifica para el aminoácido Leu.

- Si en la primera posición del codón cambiamos la C por una T, no supone ningún cambio de aminoácido. Un cambio de C → A o de C → G supone cambiar a Ile o Val, respectivamente. Por ello, esta posición no es totalmente sinónima ni totalmente no sinónima, sino que es 1/3 sinónima y 2/3 no sinónima.

- La segunda posición es totalmente no sinónima, cualquier cambio de nucleótido supone un cambio de aminoácido (Figura 8.1). De hecho, todas las segundas posiciones de los codones son siempre no sinónimas.
- La tercera posición de este codón es totalmente sinónima, cambiar la A por T, C o G no supone cambiar de aminoácido Leu.

Figura 8.1. Cálculo del número de posiciones sinónimas y no sinónimas.



Así, de las tres posiciones de este codón, tenemos 1,33 posiciones sinónimas y 1,67 no sinónimas.

$$s_{CTA} = 1/3 + 0 + 1 = 4/3 = 1{,}33$$
$$n_{CTA} = 2/3 + 1 + 0 = 5/3 = 1{,}67 \qquad \text{o} \qquad n_{CTA} = 3 - s_{CTA} = 3 - 1{,}33 = 1{,}67$$

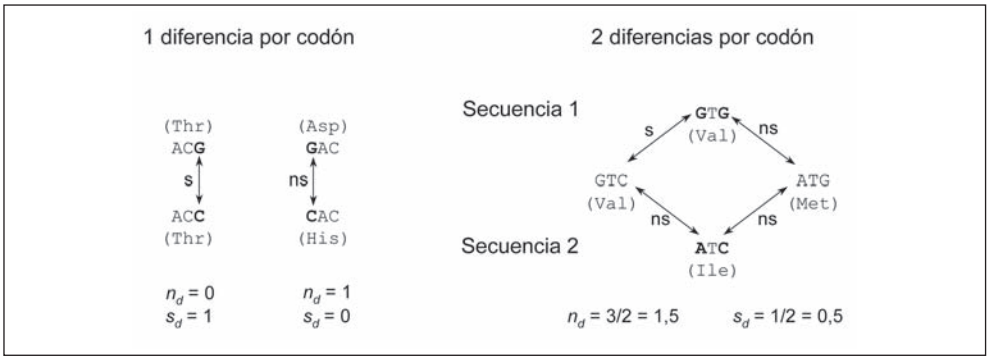
Este cálculo de posiciones sinónimas y no sinónimas se realiza para todos los codones de la secuencia y se obtiene la suma total de posiciones de cada tipo. Se repite el cálculo para la segunda secuencia a comparar y, finalmente, se obtienen las medias para ambas (S y N).

Número de diferencias sinónimas y no sinónimas

Para calcular el número de diferencias sinónimas (S_d) y no sinónimas (N_d) entre un par de secuencias homólogas debemos tenerlas alineadas. Así, podremos calcular las diferencias entre cada par de codones alineados (s_d y n_d). Cuando en un codón observamos sólo una diferencia en una de sus 3 posiciones, el cómputo es directo, una diferencia sinónima cuando ambos codones codifican para un

mismo aminoácido o una diferencia no sinónima si el aminoácido codificado por cada codón es distinto (Figura 8.2).

Figura 8.2. Cálculo del número de diferencias sinónimas y no sinónimas.



El cálculo se complica cuando observamos más de una diferencia entre un par de codones (Figura 8.2). Si al comparar dos codones vemos diferencias en dos de sus nucleótidos, por ejemplo entre GTG y ATC, estas diferencias se han podido alcanzar de dos formas distintas:

I GTG (Val) ↔ GTC (Val) ↔ ATC (Ile)[37],

o

II GTG (Val) ↔ ATG (Met) ↔ ATC(Ile)

Para calcular el número de diferencias sinónimas y no sinónimas, debemos recorrer los distintos caminos posibles y contar en cada paso del camino cómo es la diferencia generada. En este caso concreto, el camino I implica la existencia de un cambio sinónimo (GTG ↔ GTC) y otro no sinónimo (GTC ↔ ATC), mientras que siguiendo el camino II han debido producirse dos cambios no sinónimos. Para obtener el número de diferencias sinónimas, calcularemos el número medio de diferencias sinónimas al seguir cualquiera de los caminos posibles. En otras palabras, sumamos el total de diferencias sinónimas que se podrían haber producido siguiendo ambos caminos y dividimos entre 2, que son los caminos posibles a seguir. Para conocer el número de diferencias no sinónimas podemos proceder de un modo similar o lo calculamos como $n_d = 2 - s_d$.

37. Nótese que dibujamos las flechas en ambos sentidos, puesto que no sabemos cuál fue el estado ancestral.

Del mismo modo, si al comparar dos codones observamos tres diferencias, éstas se han podido generar por 6 caminos posibles. En este caso también recorreremos todos los caminos posibles contando por separado el número total de cambios sinónimos y no sinónimos y, en este caso, dividiremos entre 6.

Una vez comparados todos los pares de codones de las dos secuencias alineadas, se suman por separado las diferencias sinónimas y no sinónimas observadas en el total de codones S_d y N_d.

Cálculo de *p* y de *K*

Una vez conocido el número de posiciones y de diferencias tanto sinónimas como no sinónimas, podemos calcular la proporción de las diferencias de cada tipo como:

$$p_S = S_d/S \qquad \text{y} \qquad p_A = N_d/N$$

Finalmente, podremos estimar el número de sustituciones sinónimas K_S y no sinónimas K_A utilizando cualquiera de las correcciones expuestas en los apartados anteriores.

Puesto que la tasa de substitución sinónima es mucho más elevada que la no sinónima y es similar para muchos genes, el número de sustituciones sinónimas K_S puede utilizarse como distancia genética para reconstruir la filogenia de las especies.

8.3. Polimorfismo

Según la visión neodarwinista de la evolución, la substitución génica se produce por la selección de las mutaciones beneficiosas y el polimorfismo en las poblaciones se mantiene por selección balanceadora. Así, substitución y polimorfismo serían fenómenos distintos dirigidos por fuerzas distintas.

Para la teoría neutralista, substitución y polimorfismo son aspectos de un mismo fenómeno. La substitución es un proceso gradual y lento por el que las frecuencias de los alelos varían aleatoriamente hasta que se fijan o se pierden. En el intervalo, durante el cual la frecuencia de los alelos varía, el locus es polimórfico. El polimorfismo, pues, es un estado transitorio de la evolución molecular, y la tasa de evolución está correlacionada positivamente con el polimorfismo.

Según el modelo de equilibrio mutación-deriva, se obtiene que la heterocigosidad nucleotídica (θ) es $\theta = 4Nu$. La heterocigosidad nucleotídica es el parámetro poblacional del polimorfismo y nos indica la probabilidad de que dos alelos tomados al azar en una población sean distintos. Aunque no conocemos el valor real de θ en la población, podemos obtener distintas estimas de él.

8.3.1. Estimas de polimorfismo

Para medir el polimorfismo de las secuencias nucleotídicas de un locus en una población debemos partir de las secuencias de diversos individuos alineadas. Como ocurría al analizar la substitución entre secuencias, no se tendrán en cuenta aquellas posiciones que correspondan a InDels. Partiendo de *n* secuencias de longitud *m*, disponemos de diversos estadísticos que recogen distintos tipos de información sobre el polimorfismo. Normalmente, para estimar el polimorfismo se considera un modelo de sitios infinitos y panmixia. El **modelo de sitios infinitos** supone que la secuencia es tan larga que cualquier nueva mutación ocurre en una posición nucleotídica que no había experimentado una mutación anteriormente. La **panmixia** se refiere a que los individuos de la población se cruzan entre ellos totalmente al azar.

Una primera estima del polimorfismo entre un grupo de secuencias podría ser el número de posiciones variables, o polimórficas, *S*. Sin embargo, esta estima está condicionada al número de secuencias (*n*) que estemos analizando y a la longitud de la secuencia considerada (*m*). Si aumentamos la muestra o la longitud analizada, será más fácil que observemos más posiciones variables. Para poder comparar el nivel de polimorfismo en secuencias diversas, sin importar el tamaño de la muestra (*n*) ni la longitud de la secuencia (*m*), debemos corregir para estos dos parámetros.

Watterson (1975) demostró que:

$$E(S) = a\theta \qquad \text{donde} \qquad a = \sum_{i=1}^{n-1} \frac{1}{i}$$

Por ello, propuso la siguiente estima de theta:

$$\theta_w = S/a$$ (que se conoce como *theta de Watterson*.)

Esta estima, que está corregida para *n*, nos da un valor de θ_W por locus. Para poder comparar el polimorfismo de loci de distinta longitud de secuencia, podemos dividir por *m* para obtener el valor de θ_w por posición.

Otra estima del polimorfismo nos la proporciona *K*, que es el número medio de diferencias nucleotídicas al comparar las secuencias 2 a 2.

$$K = \frac{\sum_{i<j} K_{ij}}{\frac{n(n-1)}{2}}$$

donde K_{ij} indica el número de diferencias entre las secuencias *i* y *j*.

Bajo el modelo de sitios infinitos y panmixia también se puede demostrar que:

$$E(K) = \theta$$

por lo que el valor de *K* es una estima de θ que se conoce como θ_T (theta Tajima).

El estadístico *K* corrige para el número de secuencias, pero no para su longitud. La diversidad nucleotídica (π) proporciona una estima del polimorfismo corregida tanto para *n* como para *m*, indicando el número medio de diferencias por posición que esperamos observar al comparar dos secuencias cualquiera de la muestra.

$$\pi = \frac{n}{n-1} \sum_{i\,i} x_i x_j \pi \qquad i \neq j$$

donde:

- x_i es la frecuencia de la secuencia *i*.
- x_j es la frecuencia de la secuencia *j*.
- π_{ij} es la proporción de diferencias entre las secuencias *i* y *j* (S_{ij}/m).

También podemos ver que $\pi = K/m$.

Ejemplo:

Tenemos un fragmento de 40 pb secuenciados en 5 cromosomas de una población.

```
Secuencia 1          AATCGCTCGATCGTAGCTGGGTACTGATCGATCTGCTACG
Secuencia 2          AATCGGTCGACCGTAGCTGGGCACTGATCAATCTGCTACG
Secuencia 3          AATCGGTCGGTCGTAGCTGGGTACTGATCAATCTGCTACG
Secuencia 4          AATCGCTCGATCGTAACTGGGTACTGATCGATCTGCTACG
Secuencia 5          AATCGCTCGGTCGTAACTGGGTACTGATCAATCTGCTACG
```

$S = 6$ $K = 3{,}2$ $\pi = 0{,}08$

8.3.2. Coalescencia

La teoría de la coalescencia es un modelo retrospectivo de genética de poblaciones que permite analizar estadísticamente los datos de polimorfismo. Partiendo de la composición actual de la población, rastrea hacia el pasado la historia de la población hasta llegar al ancestro común de todos los individuos actuales (MRCA, del inglés *Most Recent Common Ancestor*).

Si consideramos que la secuencia de ADN que analizamos no experimenta recombinación, las distintas variantes están relacionadas por un árbol genealógico simple. Para dibujar este árbol partimos de las n secuencias actuales y se va hacia atrás hasta que 2 de las secuencias se unen, quedando $n - 1$ secuencias ancestrales. En una siguiente fase quedarán $n - 2$ y así sucesivamente hasta llegar al MRCA.

En una población de N individuos diploides, la probabilidad de que dos secuencias sean idénticas por descendencia (deriven de la misma secuencia de la generación anterior) es de 1/(2N), y de que no lo sean es 1 – 1/(2N). La probabilidad de que dos secuencias deriven de una misma secuencia de hace dos generaciones es: $[1 - 1/(2N)]\,[1/(2N)]$. En general, la probabilidad que dos secuencias provengan de una misma secuencia ancestral de hace $t + 1$ generaciones es:

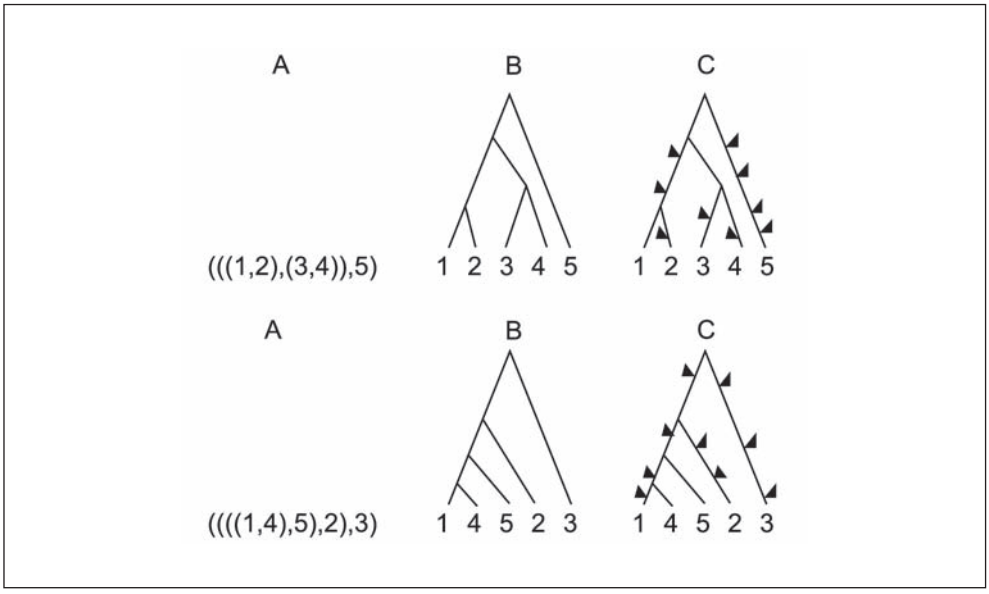
$$P_c(t+1) = [1 - 1/(2N)]^t\,[1/(2N)] \approx [1/(2N)]\exp[-t/(2N)]$$

A partir de esta distribución deducimos que la media del tiempo de coalescencia es $2N$.

En las simulaciones por coalescencia, primero generamos la topología de un árbol (el orden por el que se van uniendo las secuencias), a continuación defini-

mos los tiempos de coalescencia (longitud de las ramas) y finalmente introducimos el proceso mutacional (Figura 8.3). La tasa de mutación por secuencia por generación es *u* y el número de mutaciones en una secuencia en un periodo de tiempo *t* sigue una Poison de parámetro *ut*.

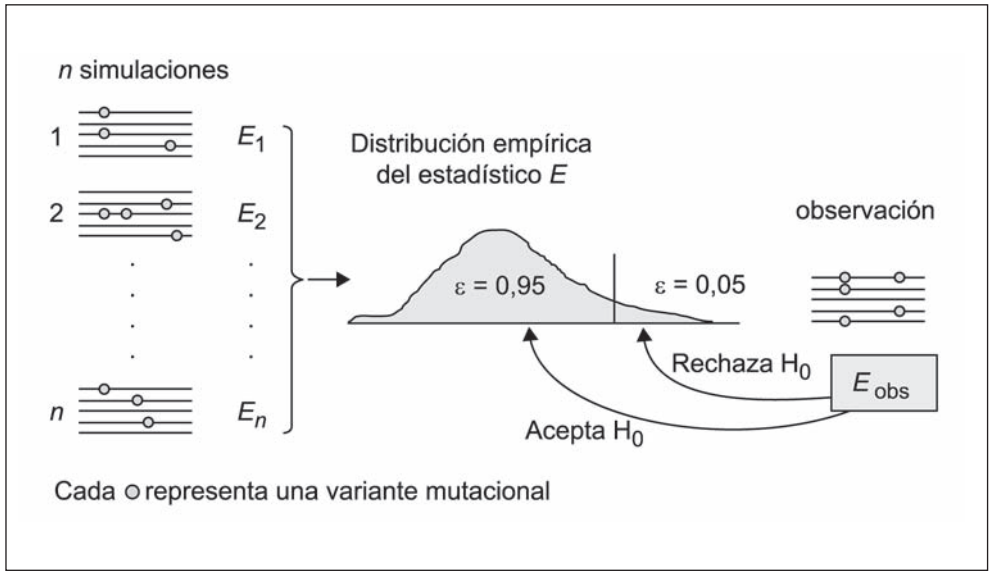
Figura 8.3. Ejemplo de 2 réplicas de simulaciones por coalescencia para una muestra de $n = 5$ y $S = 9$.



En la construcción de cada réplica se siguen tres pasos: A) topología, B) tiempos de coalescencia y C) introducción de las mutaciones.

La coalescencia permite simular muestras poblacionales según la teoría neutralista que nos permitirá testar si los valores observados en una muestra de una población real se ajustan a una genealogía neutra. Para ello, se realiza un número elevado de simulaciones (por ejemplo 1.000). A continuación, para cada una de las genealogías simuladas, se calcula el estadístico que deseamos contrastar, y con estos valores obtenemos la distribución empírica del estadístico. Comparando el valor observado del estadístico en la muestra real, con la distribución empírica del estadístico podemos asignarle su nivel de significación (Figura 8.4), que nos dirá si nuestra observación se ajusta o no a la hipótesis nula (aquella bajo la que se han realizado las simulaciones).

Figura 8.4. Esquema del proceso seguido para testar, mediante simulaciones, el ajuste de una observación a una hipótesis.



Se realizan *n* simulaciones bajo la hipótesis nula H_0 y, para cada una, se calcula el estadístico *E*. Con los *n* estadísticos se construye su distribución empírica. El estadístico observado E_{obs} se compara con la distribución empírica para aceptar o rechazar la hipótesis nula H_0.

La teoría de la coalescencia se gestó a principios de los años ochenta del siglo pasado (Kingman, 2000). Desde entonces, se han desarrollado numerosos programas para simular genealogías por coalescencia incorporando nuevos elementos, como la recombinación o distintos factores demográficos y selectivos. Entre los genetistas de poblaciones, el uso de la coalescencia se popularizó a partir del trabajo de Richard Hudson y la mayoría de los programas disponibles se basan en el código de su programa ms (Hudson, 2002).

8.4. Tests de neutralidad y detección de la selección

Como hemos visto, la mayoría de las sustituciones nucleotídicas pueden explicarse por (1) aparición de nuevas mutaciones, (2) deriva genética y (3) selección purificadora en contra de las mutaciones deletéreas.

En algunos casos, sin embargo, también puede actuar la selección positiva a favor de mutaciones beneficiosas.

Existe, pues, un gran interés por detectar estos casos de selección positiva.

Por un lado, sigue abierta la polémica sobre cuán importante es la relación selección/deriva. Por otro lado, los cambios selectivos han debido ser muy importantes para la evolución de la población porque suponen una ventaja adaptativa que, en muchos casos, puede estar relacionada con la aparición de nuevas funcionalidades. Por ello, se han desarrollado numerosos tests que intentan detectar la acción de la selección natural. Estos tests utilizan diversos aspectos de la variación neutra como hipótesis nula y, en principio, cualquier desviación de los datos indicaría la acción de la selección.

Un primer grupo de tests compara dos estimas distintas de θ. Cada estima de θ captura un aspecto diferente de la variabilidad[38], pero bajo un patrón de variabilidad neutra deberían ser intercambiables. Sin duda el test más utilizado de este grupo es el de la *D* de Tajima (Tajima, 1989), que compara las estimas de θ_T con la de θ_W.

$$D = \frac{\theta_T - \theta_W}{\sqrt{V(\theta_T - \theta_W)}}$$

En una situación de evolución neutra *D* adoptará valores cercanos a 0. Valores negativos nos indica un exceso de variantes a frecuencias extremas (que puede ser debido a selección positiva direccional). Valores positivos indican un exceso de variantes a frecuencias intermedias (que puede deberse a selección positiva balanceadora). No obstante, diversos procesos demográficos, como los cuellos de botella o la mezcla de poblaciones, podrían explicar también estos resultados. Para discernir entre los efectos de la selección y de los procesos demográficos, debemos recurrir a estudios multilocus. Mientras que la demografía debe afectar por igual a todo el genoma, el efecto de la selección se limita a la zona cercana al locus seleccionado (la diana de selección).

Otro grupo de tests se basan en la predicción de la teoría neutralista de la proporcionalidad entre polimorfismo y divergencia. El test de Hudson-Kreitman-Aguadé (HKA, Hudson et al, 1987) utiliza los datos, tanto de polimorfismo como de divergencia, de como mínimo dos loci. El test de McDonald y Kreitman (1991), compara polimorfismo y divergencia para dos tipos de posiciones de una misma región, uno de los cuales puede considerarse neutro y sirve de referencia

38. Existen diversos programas que analizan la variabilidad de las secuencias. Quizás el más utilizado sea el DnaSP (http://www.ub.edu/dnasp/), que calcula numerosos estadísticos y tests de neutralidad (Librado y Rozas, 2009).

para detectar selección en el otro tipo. Normalmente utiliza las posiciones sinónimas y no sinónimas de un mismo gen. Con estos datos obtenemos una tabla de contingencia que nos permite calcular un estadístico que se ajusta a una χ^2.

Otro test muy utilizado es el de K_A/K_S, también llamado dN/dS o ω. El valor de $\omega = dN/dS$ mide las diferencias entre las tasas de substitución sinónima y no sinónima. Cuando consideramos un único codón podemos comprender fácilmente que si un cambio de aminoácido es neutro, su tasa de fijación será la misma que la de una mutación sinónima y $\omega = 1$. Si el cambio de aminoácido es deletéreo, la selección purificadora reducirá drásticamente su tasa de substitución, con lo que $\omega < 1$. Sólo cuando el cambio de aminoácido suponga una ventaja selectiva, se fijará a una velocidad superior que las mutaciones sinónimas, y tendremos una $\omega > 1$. Este estadístico es muy conservativo puesto que los distintos codones pueden estar sometidos a distintos tipos de selección. Mientras sobre un codón puede actuar la selección positiva, otros codones pueden tener una elevada limitación funcional con lo que actúa una fuerte selección purificadora. Al considerar todos los codones del gen conjuntamente, unos valores pueden contrarrestar a otros y el valor de ω no diferirá significativamente de 1. Por ello, Nielsen y Yang (1998) desarrollaron un modelo que contempla que ω pueda variar entre codones. Este y otros modelos complejos están implementados en el paquete informático PAML 4 (*Phylogenetic Analysis by Maximum Likelihood,* Yang 2007) muy utilizado en la actualidad.

Capítulo IX

Reconstrucción filogenética

La filogenia tiene como objetivo reconstruir la historia evolutiva de los seres vivos. Para ello, determina las relaciones evolutivas entre los organismos, estima sus tiempos de divergencia y representa esta historia en forma de árbol filogenético (o genealógico).

En el primer capítulo de este libro ya comentamos que todos los seres vivos están relacionados por descendencia y que sus relaciones las podemos representar mediante lo que denominamos el árbol de la vida. El estudio de la evolución molecular nos ha dotado de herramientas para cuantificar objetivamente las diferencias genéticas entre los organismos. Estas diferencias genéticas nos permiten reconstruir el árbol que relaciona a los organismos de un modo cada vez más preciso. Así, de las primeras representaciones meramente esquemáticas de las relaciones entre especies muy alejadas (Figura 1.3, hemos pasado a poder dibujar árboles en que las longitudes de sus ramas representan los tiempos de divergencia entre especies más cercanas (Figura 6.10).

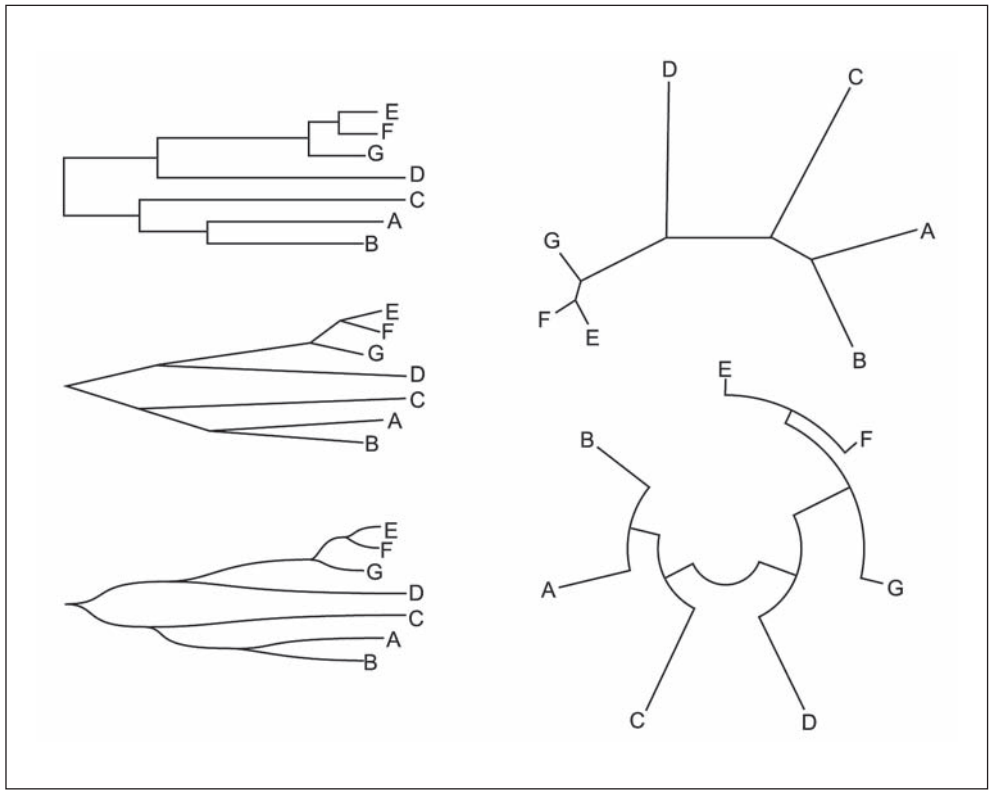
La reconstrucción filogenética a nivel molecular nos permite trazar las relaciones entre secuencias. Cada secuencia se identifica con una unidad taxonómica que puede representar a una especie, una población, un individuo o incluso un gen. Dependiendo de qué tipo sean las unidades taxonómicas que estamos comparando, el árbol que obtengamos nos indicará distintos aspectos de la evolución. Por ejemplo, podemos representar cómo se ha producido la diferenciación en especies (especiación), cómo se han originado las secuencias presentes en la población (polimorfismo), o cómo ha sido la evolución de una familia génica (genes parálogos).

9.1. Árboles filogenéticos. Generalidades

Un árbol filogenético es una representación gráfica de las relaciones evolutivas entre distintas especies, u otro tipo de entidades, que han divergido a partir

de un origen común (Figura 9.1). A nivel molecular, estas relaciones se infieren a partir de las semejanzas y diferencias entre secuencias de nucleótidos del ADN o de aminoácidos de las proteínas.

Figura 9.1. Distintas formas de dibujar el mismo árbol filogenético que sólo difieren por su estética.



Árboles obtenidos con el programa MEGA.[39]

En la representación de cualquier árbol filogenético distinguimos los **nodos** y las **ramas** (Figura 9.2). Los nodos corresponden a las unidades taxonómicas y las ramas definen las relaciones entre los nodos, en términos de ancestralidad y descendencia.

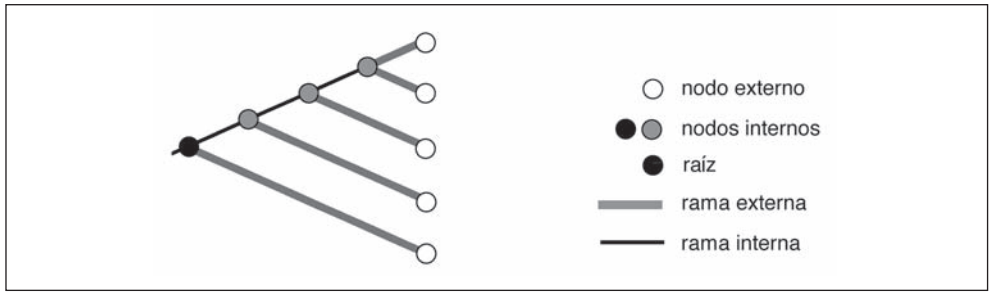
39. El programa MEGA (Molecular Evolutionary Genetics Analysis) es uno de los más usados para reconstrucción filogenética (Tamura et al., 2011)

Los nodos se dividen en nodos externos y nodos internos (Figura 9.2). Los **nodos externos** son las unidades taxonómicas operacionales u OTUs (del inglés, *Operational Taxonomic Unit*) y corresponden a las secuencias actuales, que conocemos y estamos comparando. Los **nodos internos** representan las unidades taxonómicas ancestrales que no conocemos, por lo que en ocasiones se denominan unidades taxonómicas hipotéticas o HTUs (del inglés, *Hypothetical Taxonomic Units*).

En algunos árboles hay un nodo interno particular, la **raíz**, que representa el ancestro común del resto de nodos.

Las ramas también se dividen en ramas externas y ramas internas. Las **ramas externas** son las que relacionan un nodo externo con su ancestro y las **ramas internas** unen dos unidades taxonómicas ancestrales (Figura 9.2). La longitud de las ramas suele indicar el número de cambios que han tenido lugar entre las unidades taxonómicas que relacionan.

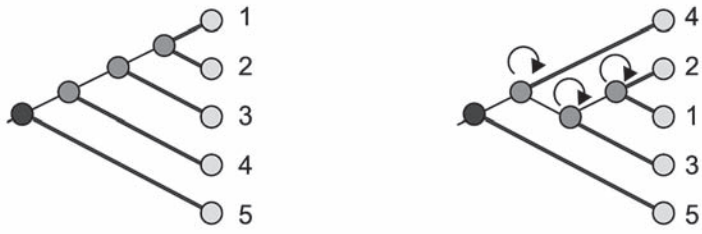
Figura 9.2. Nodos y ramas de un árbol filogenético.



Un árbol queda perfectamente definido por dos características: su topología y la longitud de sus ramas.

La **topología** de un árbol es su patrón de ramificación, que define las relaciones entre las unidades taxonómicas. Las ramas de un árbol pueden rotar libremente sobre cualquiera de sus nodos internos. En este caso, aunque el aspecto pueda parecernos muy diferente (Figura 9.3), se trata de un mismo árbol con la misma topología. La topología indica el orden en que se han ido diferenciando las unidades taxonómicas. A partir de un nodo concreto, indica qué dos unidades taxonómicas divergieron de él, siendo indiferente cómo situemos estas unidades en el espacio (en el ejemplo de la Figura 9.3, arriba o abajo).

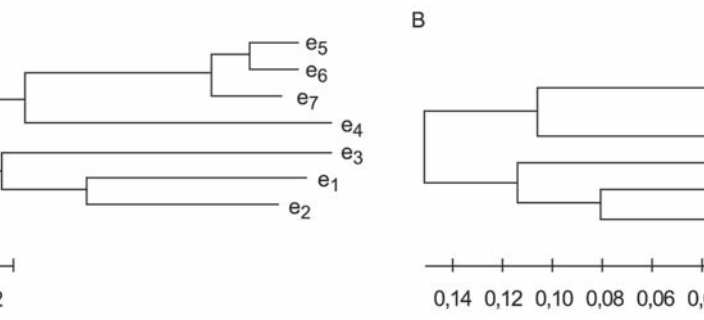
Figura 9.3. Dos representaciones de un mismo árbol. Tanto la longitud de las ramas como el patrón según el que se agrupan las unidades taxonómicas son idénticos. Las diferencias de aspecto resultan de rotar las ramas sobre los nodos internos.



Podemos distinguir 3 tipos de árboles filogenéticos según el tipo de información que nos proporcionen: cladogramas, árboles aditivos y árboles ultramétricos.

- En un **cladograma**, las ramas sólo definen las relaciones entre las unidades taxonómicas, y sus longitudes no son informativas.
- En un **árbol aditivo**, la longitud de las ramas es proporcional al número de cambios que se han producido en cada linaje particular (Figura 9.4 A). Este tipo de árboles también se conoce como filograma o árbol métrico.

Figura 9.4. Árbol aditivo (A) y ultramétrico (B) a partir de las mismas secuencias.

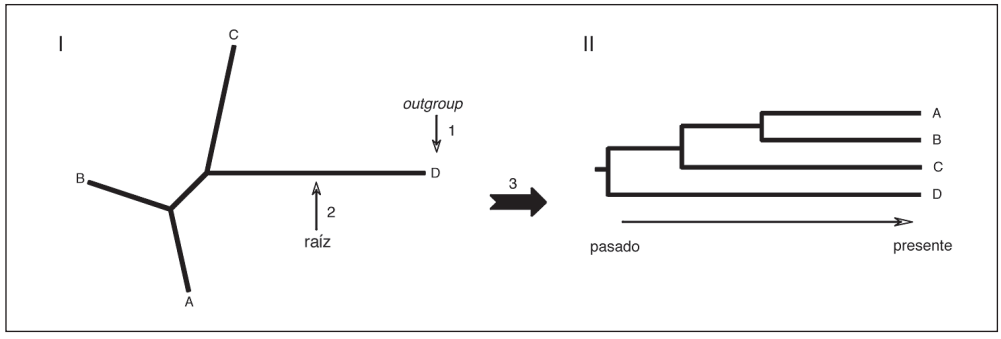


Árboles obtenidos con el programa MEGA.

- En un **árbol ultramétrico**, se supone que la tasa de substitución es constante en todos los linajes. En este caso, los OTUs quedan alineados, sobre una línea imaginaria que representa el presente, y los nodos internos se sitúan según una escala temporal (Figura 9.4. B). También se conocen como dendrogramas.

Los **árboles con raíz** (o enraizados) definen un camino evolutivo, mientras que los **árboles sin raíz** sólo ilustran las relaciones entre los OTUs sin definir ninguna direccionalidad (Figura 9.5). La mayoría de los métodos de reconstrucción de árboles generan árboles sin raíz. No obstante, podemos forzar la obtención de una raíz introduciendo en el análisis un OTU de referencia u *outgroup*. Un *outgroup* es un OTU del cual sabemos, por información independiente, que divergió con anterioridad del resto de OTUs analizados. De este modo, una vez obtenido el árbol, podemos situar la raíz en algún punto de la rama que une el *outgroup* al resto de OTUs (Figura 9.5).

Figura 9.5. Conversión de un árbol sin raíz (I) en uno enraizado (II).



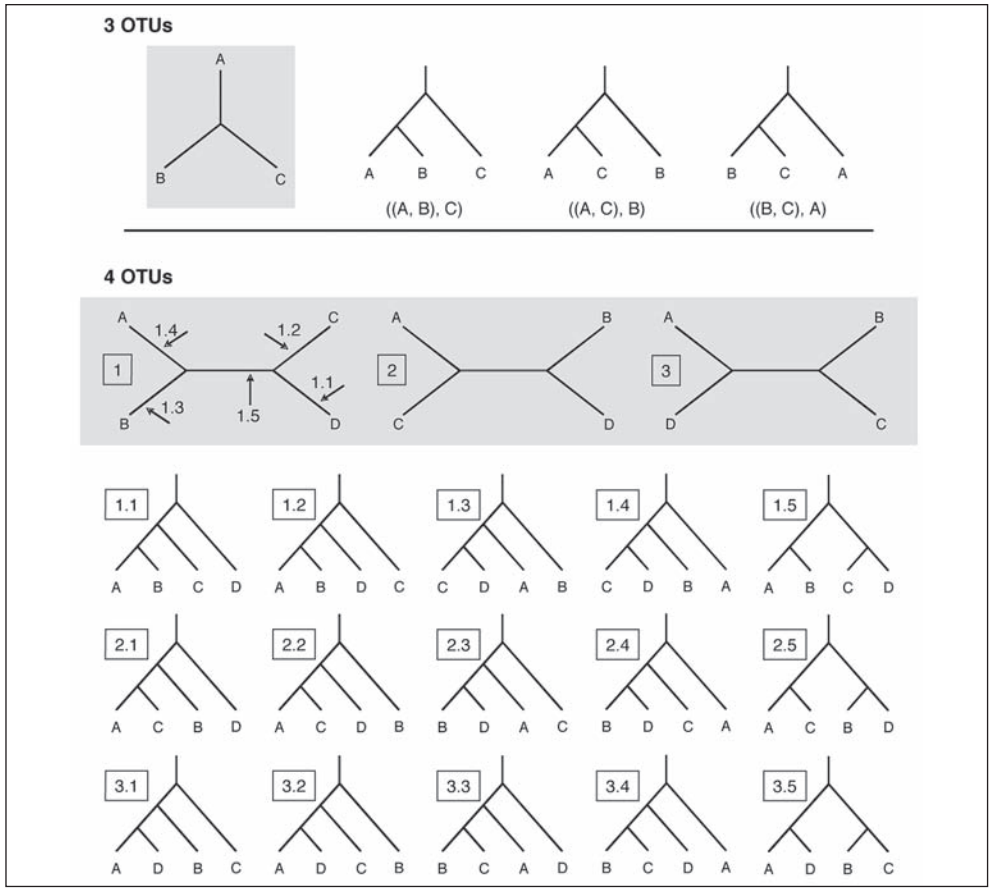
Si consideramos que el OTU D es un *outgrup* (1), podemos situar la raíz sobre la rama exterior que conduce a D (2), obteniendo así la componente temporal del árbol (3).

Considerando 3 OTUs, sólo podemos dibujar un árbol sin raíz con 3 ramas externas, pero al aumentar el número de OTUs, el número de árboles posibles crece muy rápidamente (Tabla 9.1). Así, con 4 OTUs ya podemos dibujar 3 árboles sin raíz de 5 ramas, y con 5 OTUs podemos dibujar 15 árboles de 7 ramas. Para transformar un árbol sin raíz en enraizado, podemos situar la raíz en cualquiera de sus ramas. De este modo, por cada árbol sin raíz con *m* ramas podemos obtener *m* árboles enraizados. Así, para 3 OTUs podemos obtener 3 árboles enraizados (1 árbol por cada rama del único árbol sin raíz) y para 4 OTUs podemos obtener un total de 15 árboles enraizados o, lo que es lo mismo, 5 árboles posibles para cada uno de los 3 sin raíz (Figura 9.6).

Tabla 9.1. Número de árboles posibles.

	Ramas			Árboles	
OTUs (*n*)	externas	internas árbol sin raíz	internas árbol enraizado	sin raíz (N_{sR})	enraizados (N_R)
3	3	0	1	1	3
4	4	1	2	3	15
5	5	2	3	15	105
6	6	3	4	105	945
n	*n*	*n* – 3	*n* – 2	$(2n-5)! / [2^{n-3}(n-3)!]$	$(2n-3)! / [2^{n-2}(n-2)!]$

Figura 9.6. Posibles árboles para tres y cuatro OTUs.



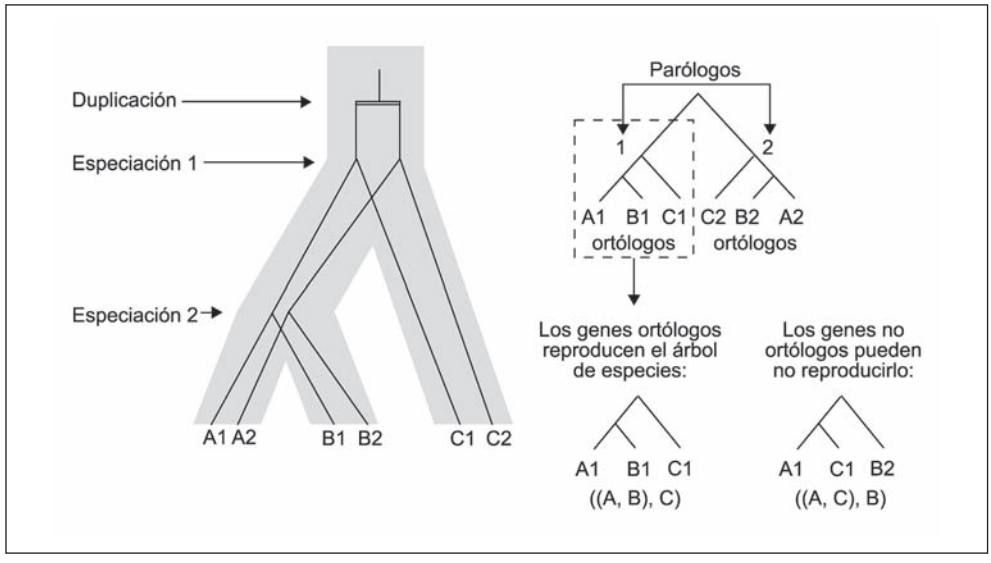
Sobre fondo gris se muestran los árboles sin raíz. Las flechas sobre el primer árbol sin raíz de 4 OTUs indican las ramas sobre las que se sitúa la raíz en el correspondiente árbol enraizado.

La historia evolutiva de los seres vivos ha sido única. Por ello, sólo existe un árbol filogenético verdadero. Sin embargo, no disponemos de toda la información necesaria para poder reconstruir fielmente esta historia que empezó hace millones de años. En algunos casos, los paleontólogos pueden reconstruir algunas series evolutivas a partir de restos fósiles datados geológicamente, pero la mayoría de las especies ancestrales nos son totalmente desconocidas.

El estudio de la evolución molecular ha favorecido el desarrollo de diversas metodologías que permiten inferir la historia evolutiva a partir de los datos moleculares (secuencias de nucleótidos o de aminoácidos) en las especies actuales. Cada método de reconstrucción elige un árbol de entre todos los posibles como el árbol que mejor explica los datos observados según un determinado modelo evolutivo. El árbol que nos devuelve es un árbol inferido, que puede coincidir o no con el árbol verdadero.

Podemos distinguir también entre el árbol de especies y un árbol de genes. El árbol de especies representa cómo han divergido un grupo de especies (sería el árbol verdadero). Un árbol de genes es un árbol inferido a partir de la información aportada por un único gen, secuenciado en cada una de las especies. Debe tenerse especial cuidado en no comparar genes parálogos que nos pueden llevar a inferir árboles de especie incorrectos (Figura 9.7).

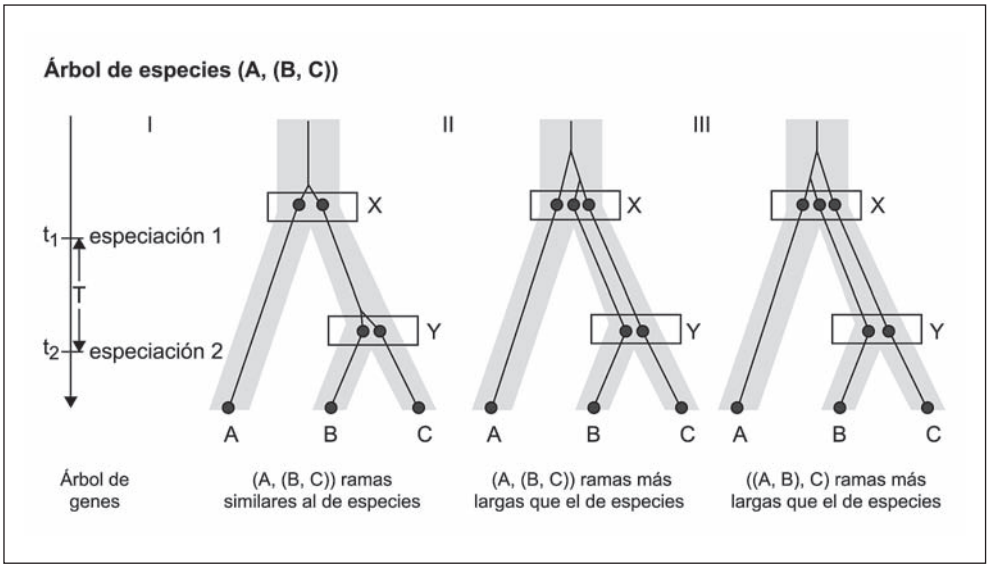
Figura 9.7. Árbol de especies y árbol de genes cuando existen familias génicas.



En el ejemplo tenemos tres especies (A, B y C) y dos genes (1 y 2).

Además, incluso comparando sólo genes ortólogos, podemos encontrarnos situaciones en las que un árbol de genes puede diferir del árbol de especies, debido a la existencia de polimorfismos intraespecíficos en las especies ancestrales y pérdida diferencial de alelos durante la especiación. Así, si la divergencia de los genes es anterior a la divergencia de las especies, el árbol de genes puede sobreestimar la longitud de la rama que separa a dos especies (Figura 9.8, II). Por otro lado, también la topología del árbol de genes puede variar respecto a la del árbol de especies (Figura 9.8, III). Cuando los 2 eventos de especiación que originarán a 3 especies se producen con poca diferencia de tiempo, es más fácil que obtengamos una topología errónea a partir del árbol de genes. Esta dificultad, por ejemplo, aparecía cuando se intentaba desentrañar la historia evolutiva entre humanos, chimpancés y gorilas (Miyamoto et al., 1987). Dependiendo del gen que se utilice para la reconstrucción filogenética, se puede obtener un orden u otro de agrupación entre los distintos linajes. Este problema sólo se solventa reconstruyendo el árbol a partir de la secuencia de más genes independientes y más individuos por especie.

Figura 9.8. Posibles relaciones entre el árbol de especies y el árbol de genes cuando el gen de cada una de las especies desciende de alelos distintos que coexistieron en una especie ancestral.



I: el árbol de genes coincide con el de especies. II: El árbol de genes muestra ramas más largas que el de especies (la diferencia es más acusada entre B y C). III: el árbol de genes, además, presenta una topología distinta a la del árbol de especies.

9.2. Métodos de reconstrucción filogenética

Para la reconstrucción de cualquier árbol filogenético basado en datos moleculares, debemos partir de un alineamiento de secuencias (nucleótidos o aminoácidos). Las secuencias elegidas en las diferentes especies deben corresponder a genes ortólogos.

Existen numerosos métodos de reconstrucción de árboles, entre los que destacan los métodos basados en matrices de distancias, los métodos de máxima parsimonia y los métodos de máxima verosimilitud.

Los métodos basados en distancias parten de la matriz de distancias genéticas entre los OTUs para ir agrupándolos secuencialmente. Estas distancias indican el número de substituciones de nucleótidos (o de aminoácidos) entre secuencias, y pueden calcularse utilizando cualquiera de los métodos que vimos en el aparado 8.2. (Estimas de la subsitución nucleotídica). Entre los métodos basados en distancias encontramos el UPGMA y el *Neighbor-Joining*.

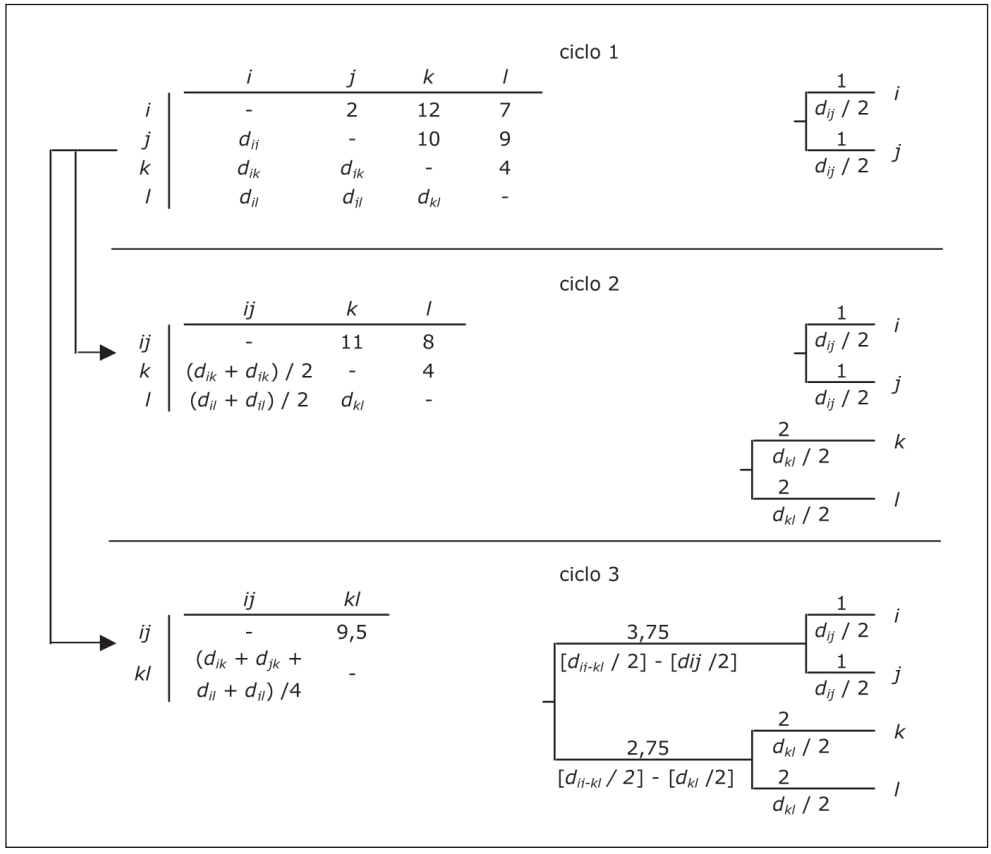
9.2.1. UPGMA (*Unweighted Pair-Group Method with Arithmetic Mean)*

El método UPGMA es el método de reconstrucción más sencillo y, aunque no se utiliza demasiado, tiene su importancia histórica. El método asume constancia de la tasa de substitución en todos los linajes. Con este método se obtiene un árbol enraizado y ultramétrico (todos los OTUs son equidistantes al nodo raíz).

El método UPGMA utiliza un algoritmo de agrupamiento secuencial en el que las relaciones de topología se infieren en orden decreciente de similitud de las secuencias. Además, las longitudes de las ramas se van determinando simultáneamente a la topología (Figura 9.9).

Si queremos construir un árbol que relacione n OTUs, dispondremos de una semimatriz con $n(n\text{-}1)/2$ distancias. Para empezar el proceso de reconstrucción, se elige aquel par de OTUs (i, j) con una distancia menor y se colocan sobre la misma línea temporal que representa el presente. La longitud de las ramas que unen cada uno de estos OTUs con su ancestro común es la mitad de la distancia entre ellos (puesto que se asume constancia en las tasas de substitución).

Figura 9.9. Ejemplo de la obtención de un árbol de 4 OTUs (*i*, *j*, *k*, *l*) que se unen en el orden ((*ij*) (*kl*)) por el método de UPGMA.



La semimatriz inferior y los valores bajo las ramas indican las fórmulas generales utilizadas en el algoritmo. La semimatriz superior y los valores sobre las ramas muestran un ejemplo numérico concreto.

A partir de este momento, consideraremos a los OTUs *ij* como una unidad taxonómica única. Por tanto, debemos construir una nueva matriz de distancias que contará con (*n*-1)(*n*-2)/2 distancias (Figura 9.9). Las distancias entre el nuevo OTU *ij* y cualquier otro OTU *x* debe recalcularse como la media de las distancias entre el OTU *x* y cada uno de los OTUs agrupados *ij*, $d_{x(ij)} = (d_{xi} + d_{xj})/2$. A continuación se procede como en el ciclo anterior, se eligen los dos OTUs separados por la distancia menor (*kl*), se sitúan en la línea del presente y se calcula la longitud de las ramas hasta el ancestro común, como la mitad de la distancia entre los OTUs.

El proceso de elegir un par de OTUs y recalcular la nueva matriz de distancias se repite *n* – 1 veces, hasta colocar todos los OTUs sobre el árbol. Debe tenerse en

cuenta que cada nueva matriz de distancias se calcula a partir de las distancias de la matriz original. Además, cuando alguno de los OTUs que debemos unir corresponde a un OTU compuesto (*ij*), la distancia al nuevo nodo interno debe desglosarse en dos ramas: la distancia de los OTUs individuales al nodo interno común más próximo (que ya calculamos anteriormente) y la distancia entre éste y el nuevo nodo. El ejemplo de la Figura 9.9 muestra cómo se unen 2 OTUs compuestos, *ij* y *kl*.

9.2.2. Cálculo de la longitud de las ramas

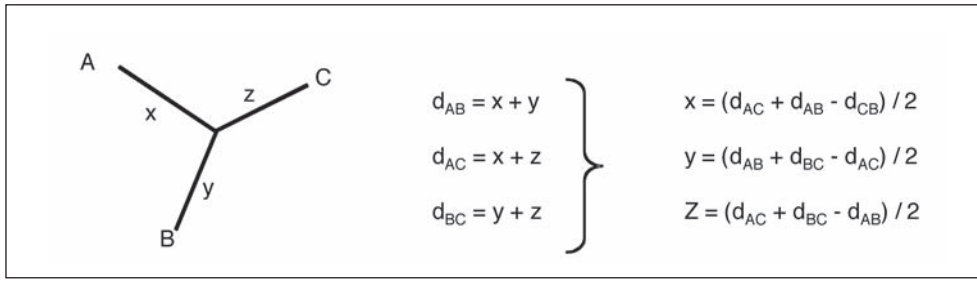
Al reconstruir la historia de un grupo de OTUs debemos resolver dos preguntas: 1) cuál es la topología del árbol que los relaciona y 2) determinar la longitud de las distintas ramas.

Muchos de los métodos de reconstrucción filogenética resuelven primero la topología y posteriormente calculan la longitud de las ramas utilizando diversos algoritmos, como pueden ser el método de Fitch y Margoliash o el método de los cuadrados mínimos.

Como ya hemos visto, el método UPGMA resuelve ambas preguntas simultáneamente, asumiendo constancia en la tasa de substitución. También se han desarrollado métodos para, una vez obtenida la topología, inferir la longitud de las ramas sin asumir constancia en la tasa de substitución.

El método de Fitch y Margoliash (1967) toma los OTUs de tres en tres y utiliza la información de las distancias entre ellos para calcular la longitud de las ramas (Figura 9.10).

Figura 9.10. Cálculo de la longitud de las ramas por el método de Fich y Margoliash.



Cuando tenemos más de 3 OTUs, también se comparan por tríos. En este caso, el primer trío está formado por los 2 OTUs que se incorporaron primero al

árbol (A y B) y un tercer OTU (C) resultado de considerar al resto de OTUs conjuntamente. En el siguiente paso, los OTUs AB forman un OTU compuesto que se utilizará junto a otros 2 nuevos OTUs para seguir calculando las longitudes de las ramas.

El método de los cuadrados mínimos calcula las longitudes de las ramas de modo que se obtenga el mínimo para la expresión:

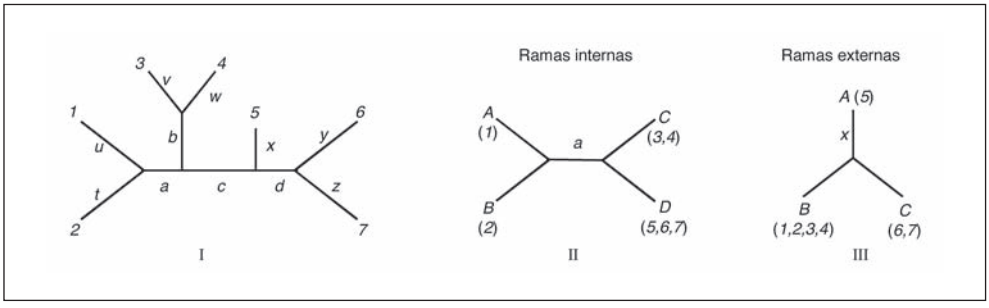
$$\sum_{ij}(d_o - d_i)^2$$

donde d_o es la distancia observada entre los OTUs *ij*, es decir, los valores calculados a partir de los datos moleculares y tabulados en la matriz de distancias. Por su parte, d_i es la distancia inferida a partir de la longitud de las ramas.

Tómese de ejemplo el árbol I de la Figura 9.11. Para calcular la longitud de una rama interna cualquiera (*a*-*d*), se reduce el árbol a uno de 4 grupos de OTUs, como en el ejemplo del árbol II para la rama *a*, y se calcula su distancia como:

$$a = \tfrac{1}{2}\,[\gamma(d_{AC} + d_{BD}) + (1\text{-}\gamma)(d_{BC} + d_{AD}) - d_{AB} - d_{CD}]$$

Figura 9.11. Ejemplo del agrupamiento de OTUs para el cálculo de las longitudes de las ramas de un árbol por el método de los cuadrados mínimos.



donde $\gamma = (n_B\, n_C + n_A\, n_D) \,/\, [(n_A + n_B)(n_C + n_D)]$ y n_A, n_B, n_C y n_D son el número de OTUs en los grupos A, B, C y D.

Para calcular la longitud de las ramas externas se reduce el árbol a uno de 3 grupos de OTUs, como en el árbol III de la Figura 9.11 para el cálculo de la rama *x*. En este caso, la longitud de *x* se calcula de modo similar al método de Fitch y Margoliash como:

$$x = (d_{AB} + d_{AC} - d_{BC}) \,/\, 2$$

Debe tenerse en cuenta que A, B o C pueden tratarse de OTUs compuestos y los valores de d_{AB}, d_{AC}, d_{BC} deben calcularse como la media de las distancias entre los pares de OTUs que pertenecen a dos OTUs compuestos distintos:

$$d_{XY} = \sum_{i,j} d_{ij} / (n_X n_Y)$$

En el ejemplo de la Figura 9.11 tenemos que:

$d_{AB} = (d_{51} + d_{52} + d_{53} + d_{54}) / 4$
$d_{AC} = (d_{56} + d_{57}) / 2$
$d_{BC} = (d_{16} + d_{17} + d_{26} + d_{27} + d_{36} + d_{37} + d_{46} + d_{47}) / 8$

9.2.3. *Neighbor-Joining* (NJ)

El método *Neighbor-joining* (Saitou y Nei, 1987) o de agrupamiento de vecinos[40], busca secuencialmente los vecinos que minimicen la longitud de las ramas del árbol.

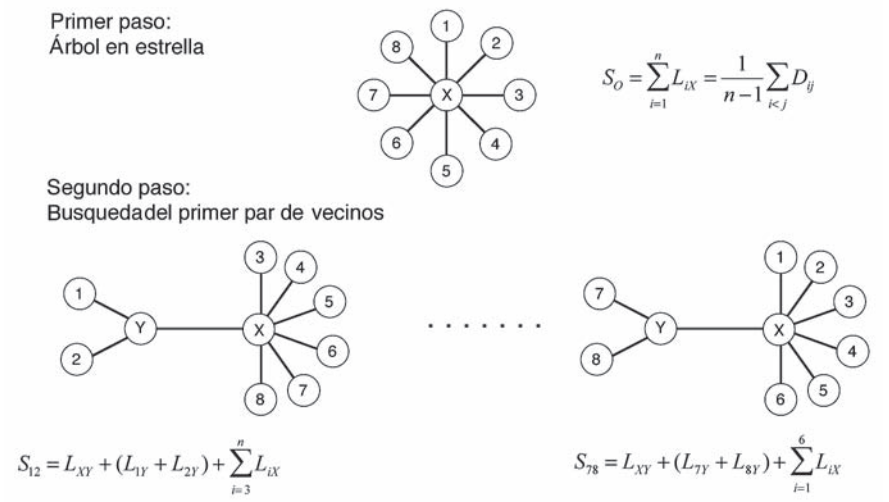
Para la reconstrucción de un árbol NJ se parte de un árbol en estrella donde todos los OTUs están unidos a un único nodo interno, *X*, (Figura 9.12). A partir de la matriz de $[n(n-1)]/2$ distancias podemos calcular la longitud de las ramas de este árbol. Si definimos D_{ij} como la distancia entre los OTUs *ij* y L_{ab} como la longitud de la rama entre los nodos *a* y *b*, podemos calcular la suma de las ramas como:

$$S_O = \sum_{i=1}^{n} L_{iX} = \frac{1}{n-1} \sum_{i<j} D_{ij}$$

Cada distancia D_{ij} de la matriz equivale a la suma de dos ramas del árbol estrella (L_{iX} y L_{jX}), por lo que al sumar todas las distancias tomamos *n*-1 veces cada rama.

40. Se dice que 2 OTUs son vecinos cuando están conectados a través de un único nodo interno.

Figura 9.12. Primeros pasos en la construcción de un árbol por *Neighbor-Joining.*



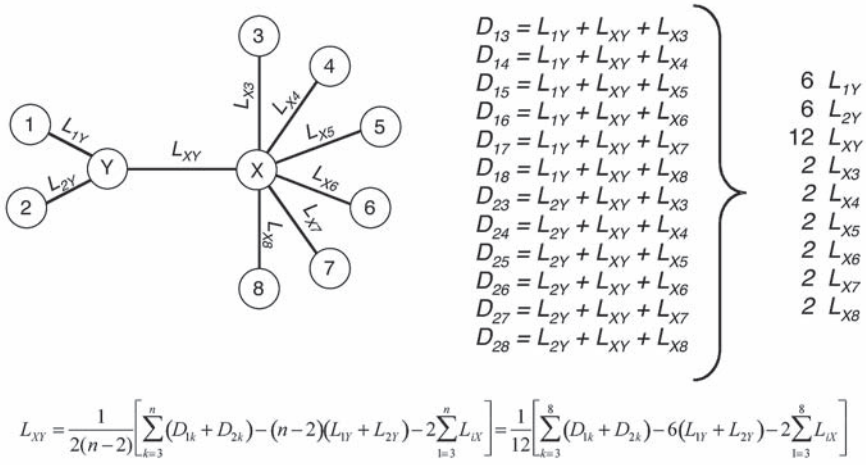
Queremos encontrar el árbol que presente las ramas con longitud mínima. Empezaremos por considerar un árbol en el que agrupamos dos vecinos. Este árbol presentará una única rama interna uniendo dos nodos internos X e Y. De uno de estos nodos internos derivan los dos OTUs considerados vecinos y del otro nodo interno derivan el resto de OTUs, en una topología en estrella (Figura 9.12). Para *n* OTUs existen [*n*(*n*-1)]/2 parejas de posibles vecinos. Cada pareja de vecinos define un posible árbol que diferirá del resto de posibles árboles en la longitud total de sus ramas. Debemos calcular la suma de las longitudes para todos los posibles árboles y nos quedaremos con aquel de menor longitud de ramas. Veamos cómo calcularíamos la suma de las ramas para el árbol que considera como primeros vecinos a los OTUs 1 y 2:

$$S_{12} = L_{XY} + (L_{1Y} + L_{2Y}) + \sum_{i=3}^{n} L_{iX}$$

Para calcular L_{XY} debemos basarnos en las distancias entre los pares de OTUs en que cada uno está unido a un nodo interno distinto. Al sumar todas estas distancias estamos incluyendo 2(*n*-2) veces la longitud que queremos calcular y además estamos incluyendo ramas extra que deberemos eliminar (Figura 9.13).

$$L_{XY} = \frac{1}{2(n-2)} \left[\sum_{k=3}^{n} (D_{1k} + D_{2k}) - (n-2)(L_{1Y} + L_{2Y}) - 2 \sum_{1=3}^{n} L_{iX} \right]$$

Figura 9.13. Cálculo de la longitud de la rama interna en el primer ciclo NJ.



donde $L_{1Y} + L_{2Y} = D_{12}$ y el sumatorio de L_{iX} corresponde a la longitud total de un árbol en estrella de n-2 OTUs y puede calcularse de modo similar a S_O:

$$\sum_{i=3}^{n} L_{iX} = \frac{1}{n-3} \sum_{3 \le i < j} D_{ij}$$

Sustituyendo adecuadamente todos estos valores en la fórmula de S_{12}, vemos que podemos calcular la suma de longitudes de las ramas utilizando los valores de distancia, que conocemos a partir de los datos moleculares:

$$S_{12} = \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{n-2} \sum_{3 \le i < j} D_{ij}$$

Una vez analizados los posibles árboles con un primer par de vecinos, elegimos el de menor longitud de ramas. Ahora consideraremos este primer par de vecinos, *ij*, como un único OTU compuesto y calcularemos la nueva matriz de distancias para los n-1 OTUs. En esta matriz, la distancia entre un OTU cualquiera k y el OTU compuesto *ij* se calcula como la media de las distancias entre k y cada uno de los componentes (i, j) del OTU compuesto: $D_{(ij)k} = (D_{ik} + D_{jk})/2$.

Se inicia un nuevo ciclo donde consideraremos de nuevo todos los posibles pares de vecinos y el proceso se repite hasta obtener una matriz con 3 OTUs, en la que sólo existe un árbol posible.

Una vez obtenida la topología, se calcula la longitud de las ramas por el método de Fitch y Margoliash (1967) o por el de los cuadrados mínimos. De este modo, el árbol que se obtiene es un árbol aditivo, es decir, que la longitud de sus ramas indican las distancias entre las secuencias que relacionan.

9.2.4. Máxima parsimonia

El método de máxima parsimonia compara estados de un carácter y fue desarrollado inicialmente para la reconstrucción filogenética a partir de caracteres morfológicos. Posteriormente, se ha adaptado a la comparación de secuencias y, para aplicarlo, partimos directamente de las alineaciones de las secuencias. El método de máxima parsimonia utiliza un criterio de evolución mínima para buscar el árbol que requiere del menor número de cambios nucleotídicos para explicar las diferencias observadas entre las secuencias que estamos analizando.

Al comparar secuencias en un alineamiento múltiple, encontramos distintos tipos de posiciones nucleotídicas. Las posiciones invariables, aquellas en que todas las secuencias analizadas presentan el mismo nucleótido, no precisan de ningún cambio y, por tanto, son irrelevantes para decidir entre posibles árboles. Dentro de las posiciones variables podemos distinguir entre posiciones informativas y no informativas. Una posición será informativa si nos permite elegir unos árboles como más probables que otros. Para que una posición sea informativa debe mostrar un mínimo de dos variantes que, además, deben estar presentes cada una en, al menos, dos secuencias distintas.

Supongamos que queremos comparar las 4 secuencias de la Figura 9.14. Sobre el alineamiento clasificamos las posiciones en invariables, variables no informativas (sobre fondo sonbreado en el ejemplo) e informativas (marcadas con un asterisco). Las únicas posiciones que nos interesan son las informativas. No obstante, en la Figura 9.14 también se muestran los cambios mínimos necesarios para explicar cada uno de los árboles posibles para una posición no informativa. Obsérvese que dicha posición no aporta ninguna información (de ahí su nombre), puesto que, para esa posición, todos los árboles posibles necesitan el mismo número mínimo de cambios.

Con 4 secuencias podemos obtener tres árboles sin raíz distintos (árboles I, II y III en la Figura 9.14). Para cada posición informativa, se determina qué nucleótido presenta cada una de las secuencias, indicándolo sobre

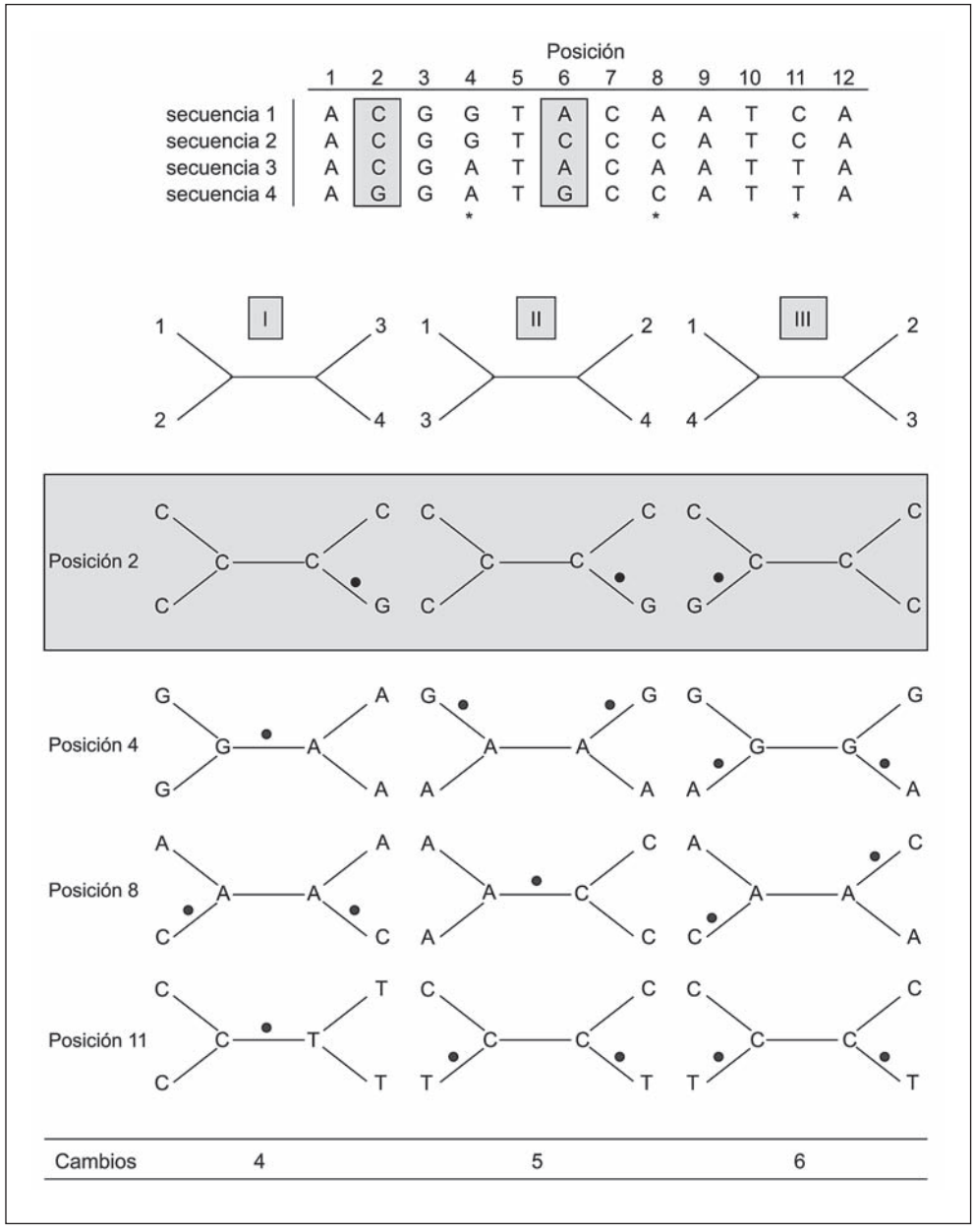
el nodo externo correspondiente a la secuencia, en cada uno de los árboles posibles. A continuación, se infiere qué nucleótido debería estar presente en la secuencia de cada nodo interno (que no conocemos) para explicar las secuencias actuales con un número menor de substituciones. En algunos de los árboles, los nucleótidos en los nodos internos sólo son una posible alternativa entre otras igual de plausibles. Por ejemplo, en el árbol II para la posición 4 se ha considerado que cada nodo interno era una A, aunque hubiéramos obtenido el mismo resultado si consideramos que ambos son una G. En ambos casos son necesarias 2 substituciones nucleotídicas, si bien sobre distintas ramas. Sin embargo, cualquier otra combinación de nucleótidos en estos nodos internos requeriría un mayor número de substituciones para explicar las secuencias actuales.

Una vez analizadas todas las posiciones informativas se suma, para cada árbol posible, el número de substituciones necesarias y nos quedamos con el que precise un menor número. En el ejemplo de la Figura 9.14, el árbol más parsimonioso es el I, puesto que sólo precisa de 4 cambios frente a los 5 y 6, necesarios para los árboles II y III. La situación mostrada por la posición 8, en que el árbol I necesita un mayor número de substituciones que el árbol II, se explicaría por la existencia de substituciones múltiples (*multiple hits*) en esta posición.

Cuando el número de OTUs es superior a 4, la cosa se complica considerablemente, porque el número de árboles posibles aumenta muy rápidamente (Tabla 9.1.). Para un número pequeño de OTUs se pueden utilizar algoritmos que rastreen todos los árboles posibles en una búsqueda exhaustiva. Se empieza por un árbol con 3 secuencias y se van añadiendo una a una el resto de secuencias de forma ordenada. En el primer paso tendremos 3 árboles, resultantes de añadir la nueva secuencia en cada una de las tres ramas del árbol de 3 secuencias. En el siguiente paso obtendremos 5 árboles por cada uno de los 3 de 4 secuencias... De este modo, cuando se haya introducido la última secuencia se habrán obtenido todos los árboles posibles.

Cuando se estudian más de 12 OTUs, la búsqueda exhaustiva se hace inviable (más de 650 millones de árboles) y debe recurrirse a otras estrategias. Una posibilidad es utilizar el método de *branch-and-bound*. En este método se parte de un árbol que puede ser totalmente arbitrario, o que se ha obtenido por un método rápido como el NJ, y para el cual calculamos el número mínimo de substituciones *L*. Este valor *L* se considerará un valor umbral crítico. A continuación se inicia un proceso similar al de la búsqueda exhaustiva. Se parte de un árbol

Figura 9.14. Ejemplo de reconstrucción por máxima parsimonia para 4 secuencias.



En el alineamiento múltiple se observan 5 posiciones variables: 2 no informativas (sobre fondo gris) y 3 informativas (marcadas con un asterisco). Los puntos sobre las ramas indican la necesidad de una substitución nucleotídica en esa rama. Sobre fondo gris se ilustran los cambios necesarios para explicar cada uno de los posibles árboles para una de las posiciones no informativas.

con 3 secuencias y se inicia un camino añadiendo progresivamente una nueva secuencia. Para cada nuevo árbol intermedio se calcula el número de substituciones y todo aquel que devuelva un valor superior a *L* es desechado y no es utilizado para añadir ninguna secuencia más en este camino. De este modo, no se exploran todos los caminos posibles para llegar al árbol que contenga todas las secuencias, ahorrándose mucho tiempo computacional. La lógica de esta estrategia se basa en que al añadir una secuencia sólo podemos obtener un árbol con igual o mayor número de substituciones, pero nunca menor. Por ello, cuando encontramos un árbol intermedio que necesita un número mayor de substituciones que *L* no continuamos el rastreo por los árboles que proceden de añadir secuencias a éste. En este caso, volvemos atrás y empezamos un nuevo camino. Si al llegar al final del camino el valor del árbol que contiene todas las secuencias es inferior a *L*, tomamos el valor de este árbol como nuevo valor umbral y continuamos con el proceso de rastrear un nuevo camino.

Para más de 20 OTUs, incluso el *branch-and-bound* es ineficiente (más de 220×10^{18} árboles posibles) y es preciso utilizar métodos heurísticos. Estos métodos, en lugar de realizar búsquedas completas, sólo examinan un subconjunto manejable de los árboles posibles. La mayoría de los métodos heurísticos parten de un árbol obtenido por un método rápido como el NJ y exploran una serie de árboles con una topología similar.

9.2.5. Máxima verosimilitud

Podemos definir la **verosimilitud** (en inglés, *likelihood*) de una hipótesis como la probabilidad de observar los datos (D) si se cumple la hipótesis (H), y normalmente se expresa como $L = P(\text{D/H})$. En el caso de la reconstrucción filogenética a partir de datos moleculares, la verosimilitud de un árbol es la probabilidad de observar un alineamiento múltiple determinado (los datos), dado un árbol y un modelo evolutivo concreto.

El objetivo de los métodos de máxima verosimilitud, pues, es encontrar el árbol con una mayor verosimilitud. Para ello, debe calcularse la probabilidad, para cada uno de los árboles posibles, de haber generado los datos observados bajo un modelo evolutivo de acúmulo de substituciones concreto (Jukes y Cantor, Kimura 2 parámetros, Tamura,...) y se elige el que tenga una mayor verosimilitud.

La Figura 9.15 muestra el cálculo de la verosimilitud de uno de los 3 posibles árboles (I) para una posición concreta. En el caso de cuatro secuencias, desco-

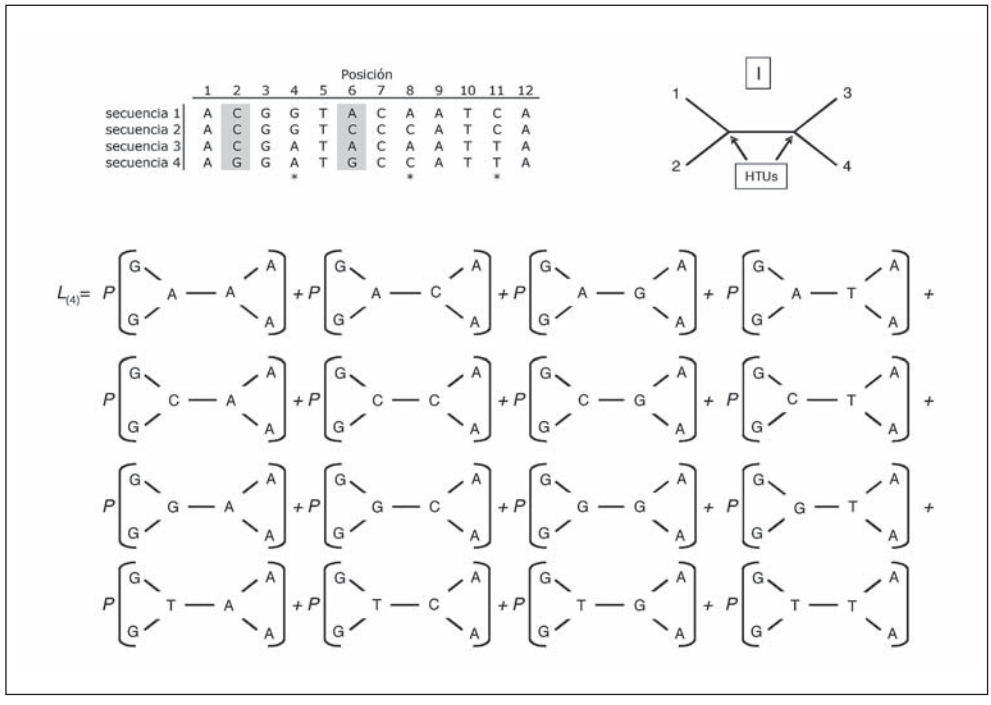
nocemos el nucleótido presente en dos nodos internos (HTUs) y debe calcularse la probabilidad, según el modelo evolutivo elegido, para cada una de las 16 posibles combinaciones de nucleótidos (4 x 4). Para el cálculo de la verosimilitud, se tienen en cuenta todas las posiciones (informativas y no informativas). Si suponemos que cada posición nucleotídica evoluciona independientemente, podemos calcular la probabilidad para cada posición y después combinarlas. Así, la probabilidad del árbol es el producto de las probabilidades individuales para el total de posiciones:

$$L = L_{(1)} \times L_{(2)} \times \ldots \times L_{(n)} = \prod_{i=1}^{n} L_{(i)}$$

Generalmente, la probabilidad del árbol se expresa como una log probabilidad:

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ldots + \ln L_{(n)} = \sum_{i=1}^{n} \ln L_{(i)}$$

Figura 9.15. Cálculo de la verosimilitud de un árbol (I) para una posición concreta (4) de un alineamiento de cuatro secuencias.



Puesto que la verosimilitud de un árbol depende del modelo evolutivo que consideremos, el árbol más verosímil para un modelo concreto puede no serlo al considerar otro modelo. Así, la elección del modelo evolutivo que se aplica es de crucial importancia. Para esta elección, también puede utilizarse un método de máxima verosimilitud[41]. En este caso, se fija un árbol y se comprueba la bondad de ajuste de los modelos, mediante el estadístico δ que usa una razón de verosimilitudes.

$$\delta = 2\log\Lambda \qquad \text{siendo} \qquad \Lambda = \frac{\max\left(L_0|D\right)}{\max\left(L_1|D\right)}$$

y L_0, L_1 es la verosimilitud para el modelo nulo y alternativo, respectivamente.

9.3. Soporte estadístico. *Bootstrap*

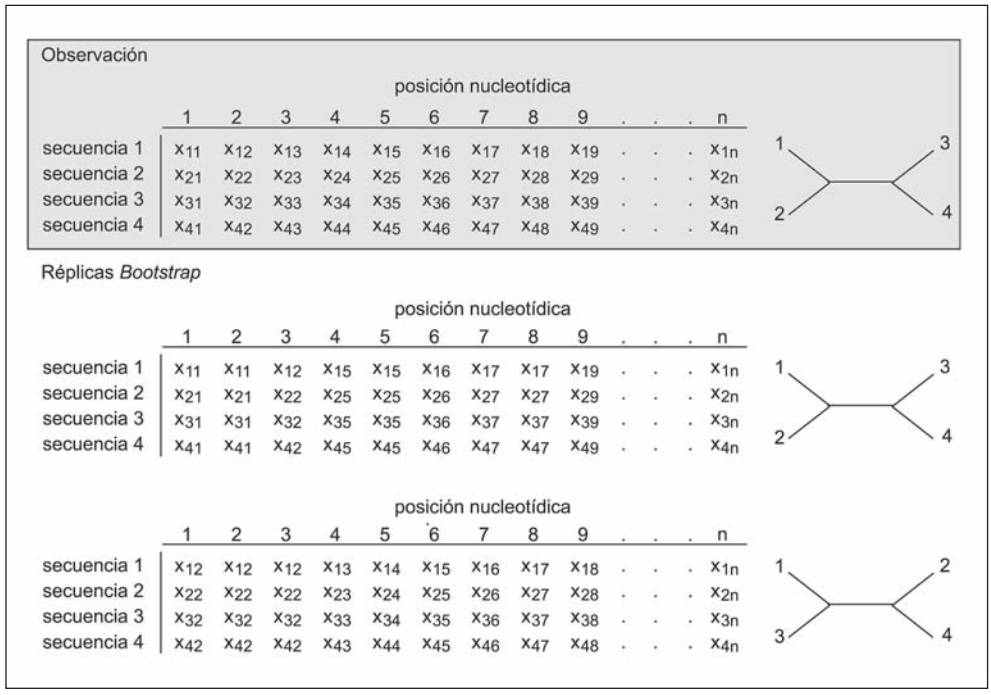
Una vez obtenido un árbol por reconstrucción filogenética, también es importante conocer su soporte estadístico. Para asociar un intervalo de confianza a un árbol, podemos realizar simulaciones por *bootstrap*.

Para ello, construiremos una serie de replicas aleatorias a partir de nuestros datos observados. Los datos observados corresponden a un alineamiento múltiple que puede considerarse como una tabla de *m* secuencias (filas) x *n* posiciones nucleotídicas (columnas). Cada réplica simulada consistirá en un alineamiento múltiple de igual número de secuencias y posiciones nucleotídicas que la muestra original. Para construirla, tomaremos al azar posiciones nucleotídicas (columnas enteras) con reemplazo (Felsenstein, 1985). Esto significa que en una muestra simulada, una posición nucleotídica de la muestra original puede aparecer varias veces, mientras que otras estarán ausentes (Figura 9.16).

Para cada muestra simulada se infiere un árbol, utilizando el mismo método que se usó con la muestra de datos observados. A continuación, se compara la topología del árbol obtenido para la simulación con la del árbol original. Cada rama interior que es idéntica a la del árbol original recibe un valor de 1, y cada rama interior distinta recibe un valor de 0.

41. Hay aplicaciones informáticas como el MODELTEST (Posada y Crandall, 1998) que realizan una búsqueda del modelo más verosímil, testando múltiples modelos evolutivos de un modo jerarquizado.

Figura 9.16. Esquema del proceso de *bootstrap.*



Sobre fondo gris se representa los datos observados y el árbol inferido a partir de éstos. A continuación se muestran 2 ejemplos de réplicas obtenidas por *bootstrap*.

El proceso de construir una réplica, inferir el árbol y compararlo al original, se repite un número elevado de veces (en general, un mínimo de 500 veces). A continuación se contabiliza el número de veces que cada rama interior ha recibido un 1 (ha resultado ser idéntica a la original) y se calcula el porcentaje de árboles simulados que representa. Este porcentaje es el valor de *bootstrap* y cuanto mayor sea, indicará un mayor soporte estadístico a la rama. A menudo, se considera que un límite de confianza aceptable es por encima del 95% y, en general, son deseables por encima del 80%. Valores inferiores representan un soporte estadístico muy bajo.

Resumen

Todos los seres vivos que habitan la Tierra provienen de un antepasado común. La gran diversidad que muestran se ha alcanzado por evolución biológica a través de millones de años. Pruebas de ello son: que comparten las características fundamentales de su unidad funcional y estructural (la célula), y que utilizan la misma molécula para codificar y transmitir la información genética que necesitan para su correcto funcionamiento (el ADN).

La estructura del ADN de dos cadenas complementarias de polinucleótidos supone que: 1) pueda contener información codificada (según sea la secuencia de nucleótidos); 2) que esta información se transmita intacta de generación en generación (por la replicación semiconservativa, que usa cada cadena como molde de su complementaria); 3) que, de vez en cuando, se introduzcan errores en la replicación (mutaciones) que serán el origen de la biodiversidad que observamos.

El ADN contiene la información necesaria para la síntesis de todas las proteínas de un individuo. Las proteínas son cadenas, por lo general bastante largas, de distintas combinaciones de unos 20 aminoácidos distintos y desempeñan funciones muy diversas. Las proteínas son las moléculas que dan especificidad a los seres vivos: las distintas especies presentan diferencias en sus proteínas y también los individuos de una misma especie pueden mostrar diferencias en algunas de sus proteínas.

El genoma de una especie es la secuencia completa de su ADN. Esta secuencia contiene: 1) la información para la síntesis de sus proteínas (genes de proteínas); 2) la información para la síntesis de ARNt y ARNr (genes de ARN); 3) señales de reconocimiento por diversas proteínas y 4) regiones intergénicas de las que, en muchos casos, se desconoce su función. La secuenciación del genoma completo de numerosas especies ha permitido abordar el estudio de nuevas cuestiones y ha puesto de manifiesto nuevos retos que sólo podrán resolverse con la ayuda de herramientas informáticas.

Las características del ADN lo convierten en la molécula idónea para realizar estudios de evolución molecular. Al comparar dos secuencias homólogas (que

provienen de un ancestro común) observamos diferencias que son debidas a la mutación. Cuanto mayor sea el tiempo de divergencia entre estas secuencias, más diferencias habrán podido acumular entre ellas. Las diferencias que observamos son sólo una subestima de las mutaciones que han tenido lugar. Podemos aplicar distintos modelos evolutivos para obtener una estima del número real de substituciones que han tenido lugar entre dos secuencias (distancia genética).

La comparación entre secuencias permite establecer las relaciones filogenéticas entre los individuos de un modo objetivo. Estas relaciones las podemos representar mediante árboles filogenéticos.

Actividades

1. Averigua a qué organismos corresponden los nombres científicos siguientes:

- *Mus musculus*
- *Arabidopsis thaliana*
- *Drosophila melanogaster*
- *Caenorhabditis elegans*
- *Pan troglodytes*
- *Neurospora crassa*
- *Saccharomyces cerevisae*
- *Rattus norvegicus*
- *Escherichia coli*

Puedes encontrar la información en muchos sitios de la web. Para los organismos más comunes basta con entrar en Wikipedia y teclear el nombre en la caja de búsqueda. Otra posibilidad más específica es entrar a la página de EOL (Encyclopedia Of life): http://www.eol.org/. Introduciendo el nombre científico en la caja de búsqueda obtendréis directamente toda la información disponible para dicho organismo.

2. Entrad en la página de GOLD (http://.genomesonline.org) y comprobad cuántos genomas completos se han publicado en la actualidad y cuántos genomas se están secuenciando. Entre los eucariotas que se están secuenciando, ¿qué organismos despiertan más interés?

3. Analizad el efecto de la deriva genética sobre las frecuencias génicas de una población. Para ello seguid los siguientes pasos:

– Bajad el programa gratuito PopG de la página web: http://evolution.gs.washington.edu/popgen/popg.html

– En la barra de herramientas del programa seleccionad Run/ NewRun. Introducid los siguientes parámetros:

- *Population size* (100)
- *Fitness* de los distintos genotipos (1)
- *Mutation from A to a* y *from a to A* (0)
- *Migration rate between populations* (0)
- *Initial freq. of allele A* (0.5)
- *Generations to run* (100)
- *Number of populations evolving* (10)

Observad los resultados y razonad su explicación.

4. Suponed que disponemos de un fragmento de ADN secuenciado en 2 especies A y B:

```
A: CCGATGGAC
B: CCAATGCTT
```

Calcula el número de substituciones según el modelo de Jukes y Cantor. Imaginemos ahora que esta secuencia corresponde a 3 codones. Calcula el número de substituciones sinónimas por posición sinónima y el número de substituciones no sinónimas por posición no sinónima.

5. Construid un árbol UPGMA para las especies A, B, C, D, E, F a partir de la siguiente matriz de distancias:

	A	B	C	D	E	F
A	–					
B	3,56	–				
C	0,48	4,24	–			
D	3,34	1,50	2,20	–		
E	4,24	1,75	4,02	2,88	–	
F	3,00	5,02	3,20	5,00	5,23	–

Ejercicios de autoavaluación

1. ¿Cómo se ha originado la Biodiversidad actual?

2. Enumera las diferencias entre ADN y ARN.

3. ¿Cómo son las uniones entre las dos cadenas de una molécula de ADN?

4. ¿Qué es un cebador?

5. ¿Cuál es la función de la telomerasa?

6. En la replicación, ¿por qué encontramos una cadena adelantada y otra de retrasada?

7. Si una especie de mamífero tiene $2n = 20$ cromosomas, ¿cuántos cromosomas tendrán los gametos que produzcan sus individuos?

8. ¿Qué tres procesos podemos distinguir en la síntesis proteica de los eucariotas?

9. ¿En qué consiste la transcripción?

10. ¿Qué tienen en común la replicación y la transcripción, en eucariotas?

11. ¿En qué consiste la maduración de ARN?

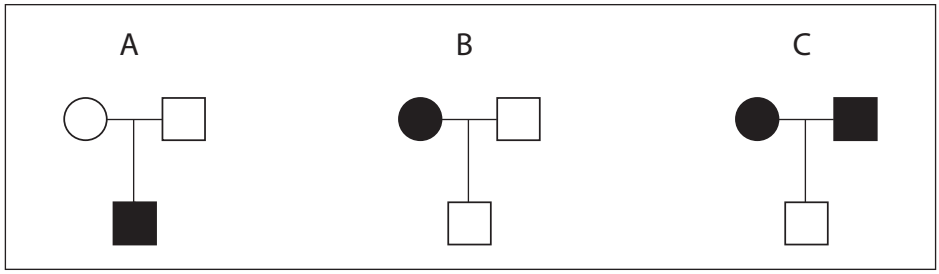
12. ¿En qué consiste el ayuste alternativo?

13. ¿Qué es el código genético universal? ¿Por qué decimos que es degenerado o redundante?

14. ¿Qué son las selenoproteínas?

15. La antigua afirmación «un gen, una proteína» no es válida. ¿Por qué?

16. Se estudiaron 3 caracteres fenotípicos (A, B, C) en una misma familia compuesta de padre, madre e hijo. A continuación se representa el pedigrí para cada carácter.



¿Cómo es la herencia de cada uno de estos caracteres y cuál es el genotipo de los distintos individuos?

17. ¿Qué supone la recombinación para los estudios de asociación?

18. ¿Qué entendemos por genoma de una especie?

19. ¿Cuál es el origen de las familias génicas?

20. En una especie, ¿pueden existir 2 genes ortólogos?

21. Dos genes homólogos en 2 especies distintas son ¿ortólogos o parálogos?

22. ¿Por qué la inserción de un nucleótido en una región codificadora tiene, en general, un efecto más grave que la substitución de un nucleótido?

23. ¿Qué tipo de mutaciones podemos detectar analizando el cariotipo de un individuo?

24. ¿Cuál es el efecto de las inversiones cromosómicas que observamos en las poblaciones naturales?

25. ¿A qué nos referimos cuando afirmamos que la selección natural es oportunista?

26. ¿Qué frecuencia de heterocigotos hay en una población en equilibrio?

27. ¿De qué tipo pueden ser las mutaciones que afectan la estructura de los cromosomas?

28. ¿Por qué la poliploidía puede llevar a una especiación rápida?

29. ¿Qué diferencia hay entre una mutación neutra y una mutación sinónima?

30. Si la secuencia ACGAGG corresponde a 2 codones, cuantas posiciones sinónimas y no sinónimas contiene?

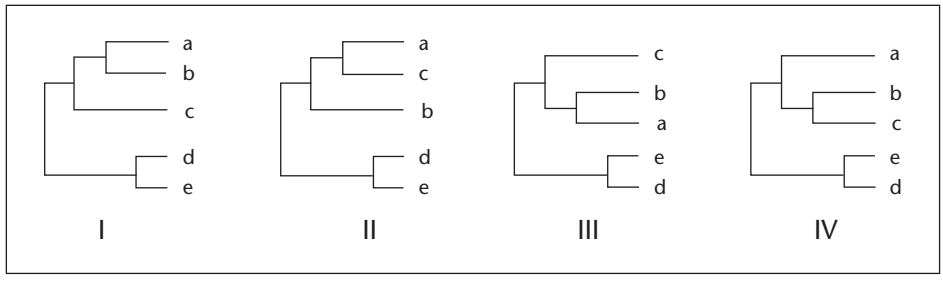
31. ¿Qué conseguimos con correcciones del estilo Jukes y Cantor o Kimura 2 parámetros?

32. ¿Qué representa la longitud de las ramas de un cladograma?, ¿de un filograma? ¿de un dendrograma?

33. Para solucionar las relaciones filogenéticas entre especies muy alejadas, ¿podemos utilizar los mismos genes que utilizamos para analizar las relaciones entre especies muy cercanas?

34. Si tenemos 8 OTUs, ¿cuántos árboles distintos sin raíz podemos dibujar?

35. ¿Cuáles de estos árboles son iguales?



Solucionario

1. Por evolución biológica durante millones de años.

2.

ADN	ARN
Contiene desoxirribosa	Contiene ribosa
Contiene Timina	Contiene Uracilo
Normalmente doble cadena	Normalmente cadena sencilla

3. Las dos cadenas de una molécula de ADN se mantienen unidas por puentes de hidrógeno entre sus bases complementarias. Entre Adenina y Timina hay 2 puentes de hidrógeno y entre Citosina y Guanina 3 puentes de hidrógeno.

4. Un cebador es una secuencia corta de nucleótidos que, al emparejarse a una cadena de ADN complementaria, permite que la ADN polimerasa inicie la replicación.

5. La protección de los telómeros durante la replicación. En cada replicación, la telomerasa dirige la incorporación de secuencias repetidas en los extremos de los cromosomas lineales. Estas repeticiones protegen al cromosoma frente a la tendencia a acortarse en cada ronda de replicación. La transcripción se inicia por la incorporación de un cebador de ARN y avanza en sentido 5' –> 3'. Posteriormente este cebador es eliminado, y en el extremo del cromosoma deja un fragmento de ADN de cadena sencilla que es fácilmente degradable.

6. Porqué las dos cadenas son antiparalelas y la replicación en ambas es en el mismo sentido 5' –> 3'.

7. Los gametos de esta especie tendrán $n = 10$ cromosomas. Durante la meiosis que originará los gametos, se produce la reducción de cromosomas a la mitad.

8. Transcripción, Procesamiento del ARN y Traducción.

9. La transcripción es el paso de la información genética codificada en el ADN (desoxirribonucleótidos) a ARN (ribonucleótidos).

10. En ambos procesos:

    – Se requiere que la doble hélice se desenrolle y separe las bases de sus dos cadenas.
    – La polimerasa incorpora nucleótidos en sentido 5' –> 3' que son complementarios a una cadena molde.

11. La maduración del ARN de los eucariotas consiste en 3 procesos: adición de una caperuza en el extremo 5', adición de una cola de poli(A) en el extremo 3' y la eliminación de los intrones.

12. El ayuste alternativo consiste en la eliminación diferencial de intrones a partir de una misma unidad de transcripción. Un mismo gen permite la traducción de múltiples proteínas dependiendo del tejido o del momento del desarrollo.

13. El código genético universal es la correspondencia entre tripletes de nucleótidos y aminoácidos. Es universal porque es compartido por la gran mayoría de organismos. Es redundante o degenerado porque existen codones sinónimos, distintos codones que codifican para un mismo aminoácido.

14. Las selenoproteínas son proteínas que contienen un aminoácido especial, la selenocisteína. La incorporación de este aminoácido está mediada por un ARN de transferencia específico que reconoce el codón UGA, que normalmente es de paro.

15. Por 2 razones:

    – Por un lado, existen genes que no codifican ninguna proteína (los que codifican para ARNr o ARNt).
    – Por otro lado, la existencia del ayuste alternativo permite que un único gen codifique para diversas proteínas.

16. El carácter A es recesivo. El hijo presenta el carácter mientras que ninguno de sus padres lo presenta. El genotipo de ambos padres debe ser Aa mientras que el del hijo será aa.

    Con los datos que disponemos no podemos saber si el carácter B es dominante o recesivo. Tampoco podemos saber cómo es el genotipo de ninguno de los individuos.

    El carácter C es dominante. Aunque lo presentan ambos progenitores, el hijo no lo presenta. El genotipo de ambos progenitores debe ser Cc mientras que el del hijo será cc.

17. La existencia de la recombinación supone que la asociación entre marcadores es más acusada cuando más cercanos se encuentran en la secuencia de ADN.

18. El genoma de una especie es el conjunto de toda la información genética que contiene sus células.

19. Las familias génicas se originan por duplicación génica.

20. No. Dos genes homólogos en una misma especie sólo han podido divergir a partir de un proceso de duplicación. Por eso, deben ser necesariamente parálogos.

21. Dos genes homólogos en 2 especies distintas pueden ser ortólogos o parálogos. Si los dos genes divergieron de un gen ancestral por duplicación, serán parálogos, si por el contrario su divergencia de un gen ancestral se originó por especiación serán ortólogos.

22. Porqué, en una región codificadora, la adición de un nucleótido representa alterar la pauta de lectura de la secuencia posterior a la adición, alterando toda la secuencia de aminoácidos. En cambio, una substitución nucleotídica, en caso de ser no sinónima, sólo representa el cambio de un aminoácido concreto.

23. Analizando el cariotipo de un individuo podemos detectar las mutaciones que afectan al número de los cromosomas y a los cambios estructurales de los cromosomas cuando afectan a fragmentos grandes.

24. Las inversiones cromosómicas provocan un cambio en la localización de los genes sobre el cromosoma y pueden tener efectos diversos. Aquellas que son deletéreas serán rápidamente eliminadas. Las inversiones que normalmente se mantienen en las poblaciones naturales tienen como efecto principal la eliminación total o parcial de la recombinación.

25. La selección natural es oportunista porqué toma, de aquellas variantes disponibles en cada momento, la variante mejor adaptada a las condiciones existentes. Si las condiciones ambientales cambian, la selección natural puede cambiar su preferencia por otra variante de entre las disponibles.

26. La frecuencia genotípica de los heterocigotos, para un gen con dos alelos, en una población en equilibrio Hardy-Weinberg es de $2pq$, donde p y q son las frecuencias génicas de cada uno de los alelos.

27. Las mutaciones que afectan la estructura de los cromosomas pueden ser: duplicaciones, deleciones, inversiones o traslocaciones.

28. La poliploidía puede conducir a una especiación rápida porque el número distinto de cromosomas entre los individuos salvajes y los mutantes supone una barrera para la obtención de híbridos. Esto hace que se vayan diferenciando dos poblaciones según el número de cromosomas que iran acumulando cada vez más diferencias entre ellas.

29. Mutación neutra y mutación sinónima son categorias de mutaciones que, en ocasiones pueden coincidir pero que, responden a conceptos disitntos. Una mutación neutra es la que no afecta a la eficacia biológica del individuo que la porta. Una mutación sinónima es aquella que, en una región codificadora, no supone un cambio de aminoácido. Una mutación sinónima puede ser neutra o cuasi neutra. Una mutación neutra puede ser sinónima, no sinónima o estar afectando a región no codificadora.

30. En total hay 4,33 posiciones no sinónimas y 0,66 sinónimas.

    ACG –> Thr. Este codón tiene 2 posiciones no sinónimas y 1 sinónima. Aunque cambiemos la G (3ª posición) por cualquier otro nucleótido sigue codificando para Thr. Por el contrario, cualquier cambio en A (1ª posición) o C (2ª posición) representa cambiar de aminoácido.

AGG -> Arg. Este codón tiene 2,33 posiciones no sinónimas y 1,66 sinónimas. Al cambiar A (1ª posición) por C sigue codificando para Arg, pero cualquier otro nucleótido hace cambiar el aminoácido. Un cambio en el nucleótido de la segunda posición representa un cambio del aminoácido codificado. Al cambiar la G (3ª posición) por A no hay cambio de aminoácido pero al cambiar por T o por G sí.

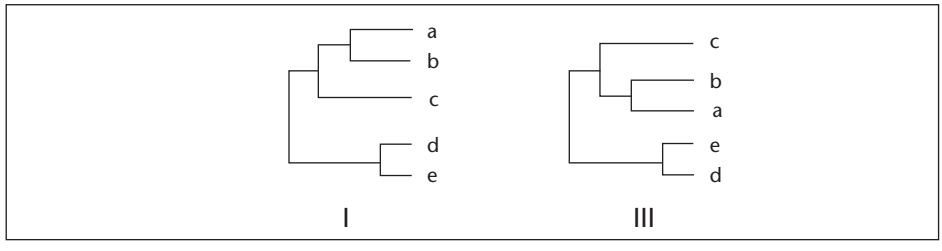
31. Los cambios que observamos entre dos secuencias sólo son una fracción de todos los que han tenido lugar. Con las correciones Jukes y Cantor y similares obtenemos una estima del número real de cambios que han tenido lugar entre dos secuencias.

32. La longitud de las ramas de un cladograma no representa nada. En un filograma, representa el número de diferencias entre las secuencias y en un dendrograma representa el tiempo de divergencia entre las unidades taxonómicas que relacionan.

33. No. Para poder observar diferencias entre especies cercanas necesitamos usar la secuencia de genes que evolucionan rápidamente. Por el contrario, para resolver las relaciones entre especies muy alejadas debemos utilizar la secuencia de genes de evolución lenta. De lo contrario, no seríamos capaces de alinear las secuencias.

34. Con 8 OTUs podemos dibujar 10395 árboles sin raíz distintos.

35. Los árboles I y III son iguales, comparten topología y longitud de las ramas. Un árbol puede obtenerse a partir del otro simplemente rotando las ramas sobre los nodos internos.



Glosario

ADN *m* Ácido desoxirribonucleico. Es la molécula que contiene la información genética de las células.

aislador *m* Elemento regulador que situado entre el intensificador y el promotor impide la acción del primero sobre el segundo.

alelo *m* Cada una de las posibles variantes de un gen.

alelo dominante *m* Alelo cuyo efecto en el fenotipo se manifiesta tanto en homocigotos como en heterocigotos.

alelo recesivo *m* Alelo cuyo efecto en el fenotipo sólo se observa en los homocigotos.

aminoácido *m* Molécula orgánica sencilla que es la unidad básica de las proteínas. En las proteínas, podemos encontrar unos 20 aminoácidos distintos.

analogía *f* Similitud observada entre dos estructuras o dos secuencias que no tienen un origen común. El parecido se alcanza por convergencia.

anticodón *m* Triplete de nucleótidos del ARNt que, en el ribosoma, se unen a un codón del ARNm facilitando la unión del aminoácido transportado por dicho ARNt a la cadena proteica creciente.

árbol de especies *m* Árbol que representa las relaciones filogenéticas entre determinadas especies.

árbol de genes *m* Árbol inferido a partir de las secuencia de un gen.

árbol de la vida *m* Árbol que representa las relacionas filogenéticas entre la totalidad de los seres vivos.

ARN *m* Ácido ribonucleico.

ATP (Trifosfato de adenina) *m* nucleótido que, en la célula, transfiere energía entre distintas reacciones.

autosoma *m* Cualquiera de los cromosomas exceptuando los cromosomas sexuales.

ayuste *m* Proceso por el cual se eliminan los intrones de un pre-ARNm.

biodiversidad *f* Diversidad total que presentan el conjunto de seres vivos del planeta.

bloque de sintenia *m* Grupo de genes ordenados de forma similar en distintas especies.

burbuja de transcripción *f* Lugar en que la doble hélice de ADN se abre permitiendo que la ARNpolimerasa efectúe la transcripción.

cadena codificadora *f* Cadena de ADN que contiene la información genética para un gen determinado.

cadena molde *f* Cadena de ADN utilizada para obtener una molécula de ARN determinada.

caja TATA *f* Secuencia característica incluida en el promotor de los eucariotas, cuya secuencia consensus es TATAAA.

caperuza *f* Es un nucleótido especial (7-metil guanosina) unida al extremo 5' del ARNm mediante un puente 5'-5' trifosfato durante el procesamiento del ARNm de los eucariotas.

cariotipo *m* Conjunto completo de cromosomas que contiene una célula somática. En general se representa con los cromosomas ordenados por parejas según su tamaño y morfología.

cebador *m* Cadena corta de nucleótidos que al emparejarse a una cadena de ADN permite iniciar su replicación.

célula *f* Unidad estructural y funcional de los seres vivos.

célula eucariota *f* Célula que tiene un núcleo definido por una membrana que contiene el material genético de la célula.

célula procariota *f* Célula en la que no existe membrana nuclear.

cigoto *m* Célula resultante de la fusión de 2 gametos y que por mitosis dará lugar a todas las células de un individuo pluricelular.

cladograma *m* Árbol que representa las relaciones filogenéticas entre los OTUs pero en el qué la longitud de sus ramas no aporta ningún tipo de información.

código genético *m* Correspondencia entre tripletes de nucleótidos (codones) y aminoácidos. La gran mayoría de organismos comparten el código genético, por lo que decimos que el código genético es universal.

codón *m* Grupo de 3 nucleótidos que codifican para un aminoácido.

coeficiente de selección *m* Intensidad relativa de selección en contra (o a favor) de un genotipo.

cola de poli(A) *f* Tira de Adeninas que se añade al extremo 3' del ARNm de los eucariotas durante su maduración.

conformación nativa *f* Estructura tridimensional de una proteína que le permite realizar su función normal.

contig *m* Fragmento de genoma cuya secuencia se ha obtenido por el ensamblado de diversos fragmentos de ADN solapados.

cromátida *f* Cada una de las dos copias que resultan de la duplicación del cromosoma, mientras se mantienen unidas.

cromosoma *m* Estructura empaquetada que adopta una molécula de ADN asociada a distintas proteínas.

dendrograma *m* Árbol filogenético cuyas ramas indican el tiempo de divergencia entre las unidades taxonómicas que unen.

deriva genética *f* Cambio en la frecuencia génica de los alelos de una población debido al azar introducido por el error de muestreo.

Dicer *m* Complejo enzimático que reconoce moléculas grandes de ARN de doble cadena y las corta en trozos pequeños.

divergencia *f* Proceso por el que dos o más poblaciones de una especie ancestral acumularon mutaciones diferenciales entre ellas. También se aplica a la acumulación de diferencias entre secuencias que tienen un ancestro común.

duplicación *f* Mutación cromosómica que consiste en la existencia de más de una copia de un segmento cromosómico determinado.

eficacia biológica *f* Éxito reproductivo relativo de un genotipo.

empalmosoma *m* Complejo enzimático formado 5 moléculas de snRNA y distintas proteínas que dirige el ayuste del pre-ARNm.

entrecruzamiento *m* (ver recombinación intracromosómica).

especiación *f* Proceso de formación de nuevas especies.

especie *f* Grupo de individuos que pueden reproducirse entre ellos y que están aislados reproductivamente de otros grupos similares.

eucariota *m* Organismo cuyas células son eucariotas.

evolución *f* Cambio en las características genéticas de una población.

exón *m* Fragmento de la secuencia de un gen que se mantiene en el ARNm tras eliminar los intrones.

fenotipo *m* Características observables de un organismo, tanto morfológicas, fisiológicas como de comportamiento.

filograma *m* Árbol filogenético cuyas ramas indican el número de cambios que se han producido en cada linaje particular.

gameto *m* Célula sexual obtenida por meiosis. La unión de dos gametos produce el cigoto.

gen constitutivo *m* Gen que está siempre activo, expresándose a una tasa constante.

gen de mantenimiento *m* Gen que se expresa en todas las células porqué su producto realiza una función básica para la célula.

genoma *m* Conjunto completo de la información genética de un organismo.

genoteca *f* Colección de clones que contienen todos los fragmentos de ADN de una fuente determinada.

genotipo *m* Conjunto de genes que tiene un individuo.

heterocigoto *m* Individuo que, para un gen determinado, tiene dos alelos distintos.

homocigoto *m* Individuo que, para un gen determinado, tiene los dos alelos iguales.

homología *f* Similitud observada entre dos estructuras o dos secuencias debido a que tienen un origen común.

InDel *m* Diferencia observada entre dos secuencias debido a la existencia de una Inserción o una Deleción entre ellas.

intensificador *m* Elemento regulador que actúa sobre el promotor favoreciendo la transcripción.

intrón *m* Fragmento de la secuencia de un gen que se transcribe a ARN pero que es eliminado durante la maduración de éste.

locus (pl. loci) *m* Lugar que ocupa un gen determinado en el cromosoma. Por extensión, también se refiere a dicho gen.

meiosis *f* Proceso de división celular eucariota que genera los gametos. Tras la replicación de los cromosomas tiene lugar dos divisiones celulares, con lo que se obtienen 4 células con la mitad de cromosomas que las células somáticas del organismo.

microsatélite *m* Secuencia de ADN en que un fragmento de 1 a 6 nucleótidos se repite en tándem. La variación en el número de repeticiones puede ser muy grande, por lo que distintas combinaciones de microsatélites son usadas como marcadores moleculares diagnóstico.

mitosis *f* División celular eucariota por la que, a partir de una célula, se obtienen 2 células idénticas.

mutación *f* Cambio permanente en el material genético.

mutación beneficiosa *f* Mutación que supone una mayor eficacia biológica al individuo que la porta.

mutación deletérea *f* Mutación que supone una pérdida en la eficacia biológica del individuo que la porta.

mutación neutra *f* Mutación que no supone ningún cambio en la eficacia biológica del individuo que la porta.

mutación no sinónima *f* Mutación en la región codificadora que supone un cambio de aminoácido en la proteína.

mutación sin sentido *f* Mutación en la región codificadora que supone la aparición de un codón de STOP prematuro.

mutación sinónima *f* Mutación en la región codificadora que no supone un cambio de aminoácido en la proteína.

neodarwinismo (o Teoría sintética de la evolución) *m* Teoría evolutiva que se basa en la idea de la selección natural de Darwin y la teoría de la herencia de Mendel. Considera la selección natural como el principal motor de la evolución.

nodo *m* En un árbol filogenético, unidades taxonómicas.

nucleótido *m* Molécula formada por un azúcar (ribosa o desoxirribosa), una base nitrogenada y un fosfato. Son las unidades cuya unión forman los ácidos nucleicos (ARN y ADN).

ortología *f* Relación de homología entre dos genes que se han diferenciado por un proceso de especiación.

paralogía *f* Relación de homología entre dos genes que se han diferenciado por un proceso de duplicación.

pauta de lectura *f* Cada una de las 3 posibles formas de leer una secuencia en grupos de 3 nucleótidos.

PCR (Polymerase Chain Reaction) *f* Reacción in vitro por la que se obtiene un número muy elevado de copias de un fragmento de ADN delimitado por dos cebadores determinados.

polimerasa *f* Enzima que incorpora nucleótidos a una cadena polinucleótida. Distintas polimerasas intervienen en la replicación y en la transcripción de los distintos tipos de ARN.

polimorfismo *m* Para una característica determinada, coexistencia de 2 o más variantes en la población.

procariota *m* Organismo cuyas células son procariotas.

procesamiento de ARN *m* Modificaciones que experimenta el pre-ARNm antes de abandonar el núcleo de los eucariotas. Consiste en la adición de una caperuza, una cola de poli(A) y la eliminación de los intrones.

promotor *m* Región reguladora cercana al inicio de la transcripción, donde se une la ARN polimerasa para iniciar la transcripción.

proteína *f* Molécula orgánica formada por la unión de aminoácidos. Su composición, que es muy diversa y específica de especie o incluso de individuo, viene determinada por la información genética que contiene el ADN de la célula. Su función es muy diversa pudiendo ser estructural, de transporte, catalítica, de defensa...

pseudogén *m* Secuencia nucleotídica similar a la de un gen pero que no se expresa. Muchos pseudogenes son el resultado de mutaciones sin sentido en una de las copias de los genes duplicados.

raíz *f* En un árbol filogenético, nodo que representa el ancestro común a todas las unidades taxonómicas.

recombinación *f* Aparición de nuevas combianciones de alelos durante la reproducción sexual.

recombinación intracromosómica (Entrecruzamiento) *f* Intercambio de un fragmento de ADN entre cromátidas homólogas no hermanas durante la meiosis.

replicación *f* Proceso por el cual una molécula de ADN se duplica, manteniendo ambas moléculas resultantes la misma información genética.

ribosoma *m* Orgánulo citoplasmático formado por dos subunidades constituidas por ARNr y distintas proteínas. Los ribosomas dirigen la síntesis de proteínas.

ribozima *f* ARN con actividad catalítica.

secuencia consenso *f* Secuencia que indica los nucleótidos más frecuentes para unas posiciones concretas.

selección natural *f* Mecanismo evolutivo resultante del hecho que los genotipos mejor adaptados a las condiciones ambientales existentes tienen una mayor probabilidad de dejar descendencia que los genotipos peor adaptados.

selenoproteína *f* Proteína que contiene Selenocisteína, el aminoácido número 21. La inclusión de este aminoácido la realiza un ARNt que reconoce un codón que generalmente es de STOP.

SNP (polimorfismo de un único nucleótido) *m* Variación natural, en uno o más individuos, del nucleótido que ocupa determinada posición en el genoma.

tambaleo *m* Propiedad de algunos ARNt por la que la primera posición del anticodón puede unirse a distintos nucleótidos de la tercera posición de un codón. Esto explica que la mayor parte de los codones sinónimos sólo difieran en su tercera posición.

taxonomía *f* Clasificación de los seres vivos en una jerarquía de taxones anidados.

teoría neutralista *f* Teoría evolutiva, cuyo principal impulsor fue M. Kimura. Esta teoría afirma que la mayoría de los cambios evolutivos a nivel molecular están impulsados por la deriva genética, que fija aleatoriamente unas u otras variantes neutras.

topología *f* Patrón de ramificación de un árbol filogenético.

traducción *f* Paso de la información genética del ARNm a la cadena de aminoácidos (proteína).

transcripción *f* Paso de la información genética del ADN a ARN.

transición *f* Mutación debida al cambio de un nucleótido por otro del mismo tipo (purina por purina o pirimidina por pirimidina).

transposón *m* Fragmento de ADN móvil, que generalmente contiene genes que codifican para enzimas que facilitan su movilidad. Debido a su movilidad, la localización de los trasnposones en el genoma es múltiple y diversa.

transversión *f* Mutación debida al cambio de un nucleótido por otro de distinto tipo (purina por pirimidina o viceversa).

unidad de transcripción *f* Está formada por la secuencia de ADN que codifica para la transcripción de una molécula de ARN y las secuencias necesarias para su transcripción. Geneneralmente incluye un promotor, la secuencia codificante del ARN y un terminador.

Lista de Abreviaturas

ADN *m*	ácido desoxirribonucleico
AFLP *m*	*Amplified Fragment Length Polymorphism*
ARN *m*	ácido ribonucleico
ATP *m*	adenosina trifosfato
BAC *m*	*Bacterial Artificial Chromosome*
BOLD *m*	*Barcode Of Life Data system*
CBOL *m*	*Consortium for the Barcode Of Life*
CRM *m* pl	*Cis-Regulatory Modules*
CTD *m*	*Carboxil Tail Domain*
ENCODE *f*	*ENCyclopedia Of DNA Elements*
FBI *m*	Federal Bureau of Investigation
GBIF *f*	*Global Biodiversity Information Facility*
GTP *f*	guanina trifosfato
HTU *f*	*Hypothetical Taxonomic Unit*
HW *m*	Hardy-Weinberg
ICR *f*	*Imprinting Control Region*
JC *m*	Jukes y Cantor
LECA *m*	*Last Eukaryotic Common Ancestor*

LTR *m* pl *Long Terminal Repeats*
LUCA *m* *Last Universal Common Ancestor*
miRNA *m* *micro_RNA*
MRCA *m* *Most Recent Common Ancestor*
NAHR *f* *Non Allelic Homologous Recombination*
NJ *m* Neighbor-Joining
ORF *m* *Open Reading Frame*
OTU *f* *Operational Taxonomic Unit*
PCR *f* *Polymerase Chain Reaction*
RFLP *m* *Restriction Fragment Length Polymorphism*
RISC *m* *RNA-induced Silencing Complex*
SECIS *f* *Sec Insertion Sequence*
siRNA *m* *small interfering RNA*
SNP *m* *Single Nucleotide Polymorphism*
snRNA *m* *small nuclear RNA*
TE *m* *Transposable Element*
TF *m* *Transcription Factor*
TFBS *m* *Transcription Factor Binding Site*
TSC *m* *The SNP Consortium*
UPGMA *m* *Unweighted Pair-Group Method with Arithmetic Mean*
UTR *f* *UnTranslated Region*
VNTR *m* *Variable Number of Tandem Repeats*
YAC *m* *Yeast Artificial Chromosome*

Bibliografía

Adams, M.D. y 194 coautores. (2000). «The genome sequence of *Drosophila melanogaster*». *Science* (vol. 287, pág. 2185-2195).

Alberts, B.; Jonson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. (2004). *Biología Molecular de la Célula* (4ª. ed.) Barcelona: Ediciones Omega, S.A.

Allison, A.C. (1954). «Protection afforded by sickle-cell trait against subtertian malarial infection». *British Medical journal* (vol. 1, pág. 290-294).

Avery, O.T.; MacLeod C.M.; McCarty M. (1944). «Studies on the chemical nature of the substance inducing transformation of Pneumococcal types». *The Journal of Experimental Medicine* (vol. 79, pág. 137-159).

Baltimore, D. (1970). «Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumor virusess». *Nature* (vol. 226, pág. 1209-1211).

Begun, D.J.; Holloway, A.K.; Stevens, K.; Hiller, L.W.; Poh, Y.-P.; Hahn, M.W.; Nista, P.M.; Jones, C.D.; Kern, A.D., Dewey, C.N.; Pachter, L.; Myers, E.; Langley, C.H. (2007). «Population genomics: Whole-genome análisis of polymorphism and divergente in *Drosophila simulans*». *PLoS Biology* (vol. 5, e310).

Berman, B.P.; Nibu, Y.; Pfeiffer, B.D.; Tomancak, P.; Celniker, S.E.; Levine, M. (2002). «Exploiting transcription factor binding site clustering to identify cis-regulatory modules envolved in pattern formation in the Drosophila genome» *Proc. Natl. Acad. Sci. USA* (vol. 99, pág. 757-762).

Black, D.L. (2000). «Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology». *Cell* (vol. 103, pág. 367-370).

Buchan, J.R.; Parker, R. (2007). «The two faces of miRNA». *Science* (vol. 318, pág. 1877-1878).

Ceballos, G.; Ehrlich, P.R. (2009). «Discoveries of new mammal species and their implications for conservation and ecosystem services» *Proc. Natl. Acad. Sci. USA* vol. 106, pág. 3841-3846).

Crick, F.H.C. (1958). «On protein synthesis». *The Symposium of the Society of Experimental Biology* (vol. 12, pág. 138-163).

Crick, F.H.C.; Leslie Barnet, .F.R.S.; Brenner, S.; Watts-Tobin, R.J. (1961). «General nature of the genetic code for proteins». *Nature* (vol. 192, pág. 1227-1232).

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in struggle for life*. London: John Murray.*

Dehal, P.; Boore, J. (2005) «Two rounds of whole genome duplication in the ancestral vertebrate». *PLoS Biology* (vol. 3, pág. 1700-1708).

Dobzhaznsky, T.; Ayala, F.J.; Stebbins, G.L.; Valentine, J.W. (1983). *Evolución*. Barcelona: Ediciones Omega.

Drosophila 12 Genomes Consortium. (2007). «Evolution of genes and genomes on the Drosophila phylogeny». *Nature* (vol. 450, pág. 203-218).

Felsenstein, J. (1985). «Confidence limits on phylogenies: an approach using the bootstrap». *Evolution* (vol. 39, pág. 783-791).

Fire, A.; Xu, S.; Montgomery, M.K.; Kostas, S.A.; Driver, S.E.; Mello, C.C. (1998). «Potent and specific genetic interferente by double-stranded RNA in *Caenorhabditis elegans*». *Nature* (vol. 391, pág. 806-811).

Fitch, W.M. (2000). «Homology a personal view on some of the problems». *TRENDS in Genetics* (vol. 16, pág. 227-231).

Fitch, W.M.; Margoliash, E. (1967). «Construction of phylogenetic trees». *Science* (vol. 155, pág. 279-284).

Gerstein, M.B.; Bruce, C.; Rozowsky J.S.; Zheng, D.; Du, J.; Korbel, J.O.; Emanuelsson, O.; Zhang, Z.D.; Weissman, S.; Zinder, M. (2007). «What is a gene, post-ENCODE? History and updated definition». *Genome Research* (vol. 17, pág. 669-681).

Graur, G.; Li, W.-H. (2000). *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates, Inc., Publishers.

Graveley, B.R. (2005). «Mutually exclusive splicing of the insect *Dscam* premRNA directed by competing intronic RNA secondary structures». *Cell* (vol. 123, pág. 65-73).

Griffiths, A.J.F.; Wessler, S.R.; Lewontin, R.C.; Carroll, S.B. (2008) *Genética*. Novena edición. Madrid: McGraw-Hill/Interamericana de España, S.A.U.

Guerrier-Takada, C.; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S. (1982). «The RNA moiety of ribonuclease P is the catalytic subunit of the enzime». *Cell* (vol. 35, pág. 849-857).

Haeckel, E. (1866). *Generelle Morphologie der Organismen*. Berlín: Reimer

Hassegawa, M.; Kishino, H.; Yano, T. (1985). «Dating of the Human-Ape splitting by a molecular clock of mitochondrial DNA». *Journal of Molecular Evolution* (vol.22, pág. 160-174).

Hatfield, D.L.; Gladyshev, V.N. (2002). «How selenium has altered our understanding of the genetic code». *Molecular and Celular Biology* (vol. 22, pág. 3565-3576).

Hey, J. (2001). «The mind of the species problem». *Trends in Ecology and Evolution* (vol.16, pág. 326-329).

* Darwin publicó 6 ediciones del *Origen de las especies* entre 1895 y 1876, cada una revisada y ampliada. En la página Darwin-online se encuentran todas ellas, además pueden encontrarse numerosas traducciones.

Hudson, R.R. (2002). «Generating simples under a Wright-Fisher neutral model of genetic variation». *Bioinformatics* (vol. 18, pág. 337-338).

Hudson, R.R.; Kreitman, M.; Aguadé, M. (1987). «A test of neutral molecular evolution based on nucleotide data». *Genetics* (vol.116, pág. 153-159).

Jensen, R.A. (2001). «Orthologs and paralogs –we need to get right». *Genome Biology* (vol. 2, Interactions 1002.1-1002.3).

Jukes, T.H.; Cantor, C.R. (1969). «Evolution of protein molecules» In Munro (ed.) *Mamalian Protein Metabolism*. (pág. 21-132). Academic Press, New York.

Kazazian, H.H.; Wong, C.; Youssoufian, H.; Scout, A.F.; Phillips, D.G.; Antonarakis, S.E. (1988). «Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man». *Nature* (vol. 332, pág. 164-166).

Kimura, M. (1968), «Evolutionary rate at the molecular level». *Nature* (vol. 217, pág. 624-626).

— (1969). «The rate of molecular evolution considered from the standpoint of population genetics». *Genetics* (vol. 63, pág. 1181-1188).

— (1980). «A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences». *Journal of Molecular Evolution* (vol. 16, pág. 111-120).

Kimura, M.; Ohta, T. (1969). «The average number of generations until fixation of a mutant gene in a finite population». *Genetics* (vol. 61, pág. 763-771).

Kingman, J.F.C. (2000). «Origins of coalescent: 1974-1982». *Genetics* (vol. 156, pág. 1461-1463).

Kozak, M. (1987). «An análisis of 5' –noncoding sequences form 699 vertebrate Messenger RNAs». *Nucleic Acid Research* (vol. 15, pág. 8125-8148).

Kruger, K.; Grabowski, P.J.; Zaug, A.J.; Sands, J.; Gottschling, D.E.; Cech, T.R. (1982). «Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena». *Cell* (vol. 31, pág. 147-157).

Kryukov, G.V.; Castellano, S.; Novoselov, S.V.; Lobanov, A.V.; Zehtab, O.; Guigó, R.; Gladyshev, V.N. (2003). «Characterization of mamalian selenoproteomes». *Science* (vol. 300, pág. 1439-1443).

Ley, T.J.; y 47 coautores (2008). «DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome». *Nature* (vol. 456, pág. 66-72).

Li, W.-H. (1997). *Molecular Evolution*. Sundderland, MA: Sinauer Associates, Inc., Publishers.

Librado, P.; Rozas, J. (2009). «DnaSP v5: a software for comprehensive análisis of DNA polymorphism data». *Bioinformatics* (vol. 25, pág. 1451-1452).

Linnaei, C. (1735) *Sistema Naturae sive regna tria naturae systematice proposita per secundum classes, ordines, genera, & species, cum characteribus, differentiis, synonymis, locis*. Leiden.*

* Linné publicó 13 ediciones de *Sistema Naturae* entre 1735 y 1770, todas en latín. Cada nueva edición supone una revisión y ampliación considerable de la anterior. La décima edición de 1758 introduce la nomenclatura binomial para los animales que ya había utilizado para las plantas.

Margulies, M. y 55 coautores. (2005). «Genome sequencing in microfabricated high-density picolitre reactors». *Nature* (vol. 437, pág.376-380).

Margulis, L.; Chapman, M.; Guerrero, R.; Hall, J. (2006). «The last eukaryotic common ancestro (LECA): Acquisition of cytoskeletal motility from aeerotolerant spirochetes in the Proteozoic Eon». *Proc. Natl. Acad. Sci. USA* (vol. 103, pág. 13080-13085).

Matys, V.; Kel-Margoulis, O.V.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; Voss, N.; Stegmaier, P.; Lewicki-Potapov, B.; Saxel, H.; Kel, A.E.; Wingender, E. (2006). «TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukariotes». *Nucleic Acids Research* (vol. 34, D108-D110).

Mayr, E. (1942). *Systematics and the Origin of Species*. New York: Columbia University Press.

McClintock, B. (1950). «The origin and behavior of mutable loci in maize». *Proc. Natl. Acad. Sci. USA* (vol. 36, pág. 344-355).

McDonald, J.H.; Kreitman, M. (1991). «Adaptive protein evolution at the *Adh* locus in Drosophila». *Nature* (vol. 351, pág. 652-654).

Mendel, G. (1865). «Versuche über Plflanzenhybriden». Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865. (Abhandlungen, 3-47). Brünn.*

Meselson, M.; Stahl, F.W. (1958). «The replication of DNA in *Escherichia coli*». *Proc. Natl. Acad. Sci. USA* (vol. 44, pág. 671-682).

Miyamoto, M.M.; Slightom, J.L.; Goodman,M. (1987). «Phylogenetic relations of humans and African apes from DNA sequences in the ψη-Globin region». *Science* (vol. 238, pág. 369-372).

Nei, M.; Gojobori, T. (1986). «Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions». *Molecular Biology and Evolution* (vol. 3, pág. 418-426).

Nielsen, R.; Yang, Z. (1998). «Likelihood models for detecting positively selected amino acid sites and aplications to the HIV-1 envelope gene». *Genetics* (vol. 148, pág. 929-936).

Passarge, E.; Horsthemke, B.; Farber, R.A.; (1999). «Incorrect use of the term synteny». *Nature Genetics* (vol. 23, pág. 387).

Pierce, B.A. (2006). *Genética. Un enfoque conceptual*. Madrid: Editorial Médica Panamericana.

Posada, D.; Crandall, K.A. (1998). «MODELTEST: testing the model of DNA substitution». *Bioinformatics* (vol. 14, pág. 817-818).

Ratnasingham, S.; Hebert, P.D.N. (2007). «Barcoding BOLD: The Barcote of Life Data System (www.barcodinlife.org)». *Molecular Ecology Notes* (doi:10.1111/j.1471-8286.2006.01678.x)

* También existe una traducción de Bateson (1910) al inglés: «Experiments in plant hybridization (1865) read at the Februay 8th, and March, 1865 meetings of the Brün Natural History Society».

Richards, S.; y 51 coautores. (2005). «Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and cis-element evolution». Genome Research (vol. 15, pág. 1-18).

Ronaghi, M; Karamohamed, S.; Pettersson, B.; Uhlén, M.; Nyrén, P. (1996). «Real-Time DNA sequencing using detection of pyrophosphate release» *Analytical Biochemistry* (vol. 242, pág. 84-89).

Saitou, N.; Nei, M. (1987). «The Neighbor-joining method: A new method for reconstructing phylogenetic trees». *Molecular Biology and Evolution* (vol. 4, pág. 406-425).

Sanger, F.; Nicklen, S.; Coulson, A.R. (1977). «DNA sequencing with chain-terminating inhibitors». *Proc. Natl. Acad. Sci. USA* (vol. 74, pág. 5463-5467).

Sharp, P.A.; Burge, C.B. (1997). «Classification of introns: U2-type or U12-type». *Cell* (vol. 91, pág. 875-879.

Shine, J.; Dalgarno, L. (1975). «Determinant of cistron specificity in bacterial ribosomes». *Nature* (vol. 254, pág. 34-38).

Sinden, R.R. (1999). «Human genetics'99: trinucleotide repeats. Biological implications of the DNA structures associated with disease-causing triplet repeats». *The American Journal of Human Genetics* (vol.64, pág. 346-353).

Sonnhammer E.L.; Koonin, E.V. (2002). «Orthology, paralogy and proposed classification for paralogs subtypes. *TRENDS in Genetics* (vol. 18, pág. 619-620).

Spradling, A.C.; Rubin, G.M. (1982) «Transposition of cloned P elements into Drosophila germ line chromosomes». *Science* (vol. 218, pág. 341-347).

Srinivasan, G.; James, C.M.; Krzycki, J.A. (2002). «Pyrrolysine encoded by UAG in Archaea: Charging of a UAG-decoding specialized tRNA». *Science* (vol. 296, pág. 1459-1462).

Stern, C. (1943). «The Hardy-Weinberg Law». *Science* (vol.97, pág. 137-138.

Strahl, B.D.; Allis, C.D. (2000). «The language of covalent histone modifications». *Nature* (vol. 403, pág. 41-45).

Sturtevant, A.H. (1913). «The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association». *Journal of Experimental Zoology* (vol. 14, pág. 43-59).

Tajima, F. (1989). «Statistical method for testing the neutral mutation hipótesis by DNA polymorphism». *Genetics* (vol. 123, pág. 585-595).

Tajima, F.; Nei, M. (1984). «Estimation of evolutionary distances between nucleotide secuences». *Molecular Biology and Evolution* (vol. 1, pág. 269-285).

Tamura, K. (1992) «The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA». *Molecular Biology and Evolution* (vol. 9, pág. 814-825).

Tamura, K; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S. (2011). «MEGA5: Molecular Evolutionary Genetics Analysis using maximumlikelihood, evolutionary distance, and maximun parsimony methods». *Molecular Biology and Evolution* (vol. 2, pág. 2731-2739).

The ENCODE Project Consortium. (2007). «Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot Project». *Nature* (vol. 447, pág. 799-816).

Watson, J.D.; Crick, F.H.C. (1953). «Molecular structure of nucleic acids: A structure for Deoxyribose Nucleic acid». *Nature* (vol. 171, pág. 737-738).

Watterson, G.A. (1975). «On the number of segregating sites in genetical models without recombination». *Theoretical Population Biology* (vol. 7, pág. 256-276).

Wheeler, D.A.Y 26 coautores. (2008). «The complete genome of an individual by massively parallel DNA sequencing». *Nature* (vol. 452, pág. 872-876).

Woese, C.R.; Kandler, O.; Wheelis, M.L. (1990). «Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria , and Eucarya». *Proc. Natl. Acad. Sci. USA* (vol. 87, pág. 4576-4579).

Yang, Z. (2007). «PAML 4: Phylogenetic Analysis by Maximum Likelihood». *Molecular Biology and Evolution* (vol. 24, pág. 1586-1591).

Zamore, (2002). «Ancient pathways programmed by small RNAs» *Science* (vol.296, pág. 1265-1269).

Zhang, J. (2003). «Evolution by gene duplication: an update». *Trends in Ecology and Evolution* (vol. 18, pág. 292-298).

Zhang, Y.; Baranov, P.V.; Atkins, J.F.; Gladyshev, V.N. (2005). «Pyrrolysine and Selenocysteine use dissimilar decoding strategies». *The Journal of Biological Chemistry* (vol. 280, pág. 20740-20751).

Zinoni, F.; Birkmann, A.; Stadman, T.C.; Böck, A. (1986). «Nucleotide sequence and expresión of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from *Escherichia coli*». *Proc. Natl. Acad. Sci. USA* (vol. 83, pág. 4650-4654).