



ACADEMIA DE CIENCIAS  
DE LA  
REGIÓN DE MURCIA

## GENES ESQUIVOS

Discurso del Académico de Número  
Ilmo. Sr. D. Francisco J. Murillo Araujo  
leído en la Sesión Solemne de Apertura de Curso  
el día 28 de febrero de 2008

Murcia  
2008

GENES ESQUIVOS

Francisco J. Murillo Araujo

Todos los derechos reservados

Queda prohibida, salvo excepción prevista en la Ley, cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con autorización del titular de propiedad intelectual. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (arts. 270 y ss. del Código Penal).

© Academia de Ciencias de la Región de Murcia, 2008

I.S.B.N.: 978-84-612-2036-6

D.L.: MU-180-2008

# I

## **Unos pocos (pero extraordinarios) para comprender lo invisible**

En apenas un siglo, si descontamos los cuarenta años de olvido de los hallazgos de Mendel, la Genética ha alcanzado un grado de desarrollo formidable. Dos consideraciones pueden ofrecerse como claras pruebas de madurez de la Genética como ciencia. En el plano conceptual, los hallazgos más esenciales de la Genética impregnan hoy día a muchas otras disciplinas científicas. Desde luego, a todas aquellas empeñadas en entender el fenómeno biológico, desde el nivel molecular al nivel de sistemas, habiendo sido probablemente la Genética la disciplina que más ha contribuido a la visión integrada que tenemos hoy de la naturaleza, funcionamiento y evolución de los seres vivos (la famosa frase “nada tiene sentido en Biología si no es a la luz de la Genética”). Pero el conocimiento genético ha calado también otras disciplinas, relacionadas sobre todo con las ciencias agrarias y

veterinarias y, por supuesto, la ciencia médica, habiendo sido protagonista principal de variadas y novedosas estrategias en las investigaciones básicas o los esfuerzos de aplicación utilitaria de esas ciencias.

En el plano estrictamente tecnológico, la amplitud y solidez de los conocimientos genéticos han derivado en la aparición de numerosos métodos prácticos, algunos realmente precisos, a la par que ingeniosos, que han hecho posible un sinfín de nuevas investigaciones básicas y desarrollos aplicados. Vano intento sería tratar siquiera de resumir aquí unos y otros. Baste mencionar la enorme influencia de la Genética en la mejora animal y vegetal, o la contribución de la ingeniería genética molecular a la industria biotecnológica y, desde luego, a la comprensión, diagnóstico y tratamiento de enfermedades, y no sólo las hereditarias. La reciente consecución en dos laboratorios distintos de la transformación de células de ratón ya diferenciadas, adultas, en células embrionarias indiferenciadas, mediante la introducción artificial de unos genes concretos, es un excelente botón de muestra del prodigioso avance de nuestro conocimiento de la naturaleza y modo de acción de los genes, y de las fascinantes perspectivas abiertas por nuestra habilidad para manejarlos. Soslayo el (para algunos) espinoso asunto de los retos que el avance de la Genética ha supuesto para filósofos, moralistas y hasta políticos de una u otra índole, que constituye otra prueba adicional del grado de

desarrollo de una actividad científica, su manifiesta relevancia para los “asuntos humanos”

Me referiré aquí al reto que ha supuesto, y sigue suponiendo, entender el propio fenómeno genético, no a sus implicaciones, teóricas o prácticas, para otras áreas de la actividad humana. En contraste con el panorama de la investigación actual, al que me referiré más tarde, quisiera destacar el número relativamente escaso de investigadores que protagonizaron los primeros pasos de la aventura del conocimiento que desembocaría, transcurridos dos tercios del pasado siglo, en un modelo robusto y casi acabado, al menos en apariencia, sobre la naturaleza de los genes, sobre cómo actúan para determinar las características estructurales y funcionales básicas de los seres vivos, y sobre cómo evolucionan. Me estoy refiriendo, desde luego, a los investigadores a los que cupo el mérito de los grandes hitos, de avances conceptuales de tal magnitud que abrieron de golpe imprevistos y vastos territorios a la exploración, ahora sí, de legiones de otros investigadores.

Cronológicamente, y como es bien sabido, la lista de esos pocos, pero extraordinarios, aventureros, se inicia con Gregor Mendel, cuyos experimentos con guisantes (publicados en 1866) le llevaron a proponer la propia existencia de los genes, factores heredables los denominaría él, como entes discretos, responsables de las

características básicas de los seres vivos, depositados en los gametos (de acuerdo con ciertas reglas) para ser transmitidos de generación en generación, y que pueden encontrarse en estados alternativos, explicando así las variaciones heredables de unos individuos a otros. Reconocidos los hallazgos de Mendel, serían Thomas H. Morgan y sus seguidores, en particular Alfred H. Sturtevant, quienes, a principios del siglo pasado, trabajando con la mosca del vinagre, la famosa *Drosophila melanogaster*, dieron el siguiente gran salto y demostraron que los genes están asociados en grupos, en los que se ordenan en forma lineal, como las cuentas de un collar, grupos que Bárbara McClintock, la primera mujer en esta historia, demostraría que se corresponden con los cromosomas, unos corpúsculos alojados en el núcleo celular. Morgan y Sturtevant enseñaron además en qué forma, mediante análisis matemáticos sencillos de los resultados de cruzamientos controlados, podía identificarse la posición relativa de cada gen en el cromosoma correspondiente, su *locus*, enseñanza rápidamente aplicada a muchos otros organismos. Al demostrar que los rayos X provocan cambios en los genes, la *transmutación artificial* que llamó él, otro miembro del clan de Morgan, Hermann J. Muller, confirmó la naturaleza material de los genes, abriendo además el fructífero campo de la mutagénesis artificial.

A partir de los años cuarenta, y en apenas dos décadas, se cumpliría el reto de penetrar en la naturaleza material de los genes,

comprender su modo de acción y explicar su capacidad de cambio, de evolución. Trabajando con organismos tan modestos como el hongo *Neurospora crassa* o la bacteria *Escherichia coli*, George W. Beadle y Edward L. Tatum por un lado, y Charles Yanofsky por otro, demostraron que en cada gen está cifrada en realidad la secuencia específica en que se ordenan las unidades básicas, los aminoácidos, de una enzima, un tipo de proteínas (polímeros de aminoácidos) encargadas de catalizar de forma específica alguna de las múltiples reacciones químicas que se producen en el interior de la célula. Extendida esta idea a las demás proteínas que actúan como herramientas para otras múltiples funciones celulares no enzimáticas, el problema de entender la misteriosa capacidad de los genes para determinar las características estructurales y funcionales básicas de un ser vivo se “reducía” a entender su capacidad de “información” sobre el número y la secuencia lineal específica en que se ordenan los 20 tipos de aminoácidos conocidos en cada una de las proteínas presentes en dicho ser vivo. Quedaba, desde luego, el enigma de la replicación de los genes, una propiedad imprescindible para su transmisión de generación en generación, sobre la que, en ese momento, no cabían más que especulaciones, aunque alguna tan sugestiva y premonitoria como la que aludía a algún tipo de complementariedad química entre moléculas, desarrollada sobre todo por Max Delbrück, un físico transmutado en biólogo y que ejerció una enorme influencia en el nacimiento y desarrollo de la moderna Biología Molecular.

Establecido que los genes están hechos de DNA - los experimentos del grupo de Oswald T. Avery, escasamente apreciados en su temprano momento de los años 40, y los posteriores de Alfred D. Hershey y Marta Chase, ahijados científicos de Delbrück - es bien conocido que el modelo de hélice de dos cadenas propuesto por James D. Watson, otro apadrinado de Delbrück, y Francis H. C. Crick, a principio de los 50, vino a aclarar de un golpe el modo en que debía producirse la replicación del material genético (la complementariedad química entre las dos cadenas de la hélice permitiría actuar a cada una como molde para la síntesis de la otra), y en qué forma podía actuar como banco de datos cifrados para informar de cuantas proteínas (secuencias de aminoácidos) distintas fuera necesario (los cuatro monómeros diferentes que forman el DNA, los denominados nucleótidos, podían disponerse en cualquier orden, sin restricción alguna, en cada sencilla de la doble hélice). Para completar su encanto irresistible, el modelo de W. y C. sugería también de inmediato una explicación fácil de la tercera propiedad fundamental de los genes, su capacidad de cambio, base de las variaciones genéticas entre individuos, incluidas nuestras enfermedades hereditarias, y base, en último término, de la aparición de nuevas especies y su evolución. Las mutaciones genéticas no serían más que cambios de mayor o menor envergadura en la secuencia de nucleótidos del DNA, quizás como consecuencia, entre otras posibles razones, de errores inevitables de

cualquiera que fuera la maquinaria encargada de replicarlo. El prematuro e invisible factor hereditario del Mendel del siglo XIX podía definirse en el siglo XX en términos moleculares precisos, asociarlo a un tramo delimitado y continuo de un polímero (el DNA), en cuya secuencia de monómeros (los nucleótidos), estaba cifrada, mediante alguna clave, la secuencia específica de otros monómeros distintos (los aminoácidos) de otros polímeros, las proteínas, que constituyen las herramientas básicas del metabolismo celular y la fisiología de los organismos.

Ahorro ahora detalles y nombres, y comentaré simplemente que, en pocos años, las sugerencias básicas del modelo de W. y C. se verían refrendadas por numerosos y variados datos experimentales. Así, se demostró que, durante la replicación del DNA, cada cadena sirve efectivamente de molde para la síntesis de su cadena complementaria, y se descifró la clave que conecta genes con proteínas, una clave de tres letras, el número mínimo necesario para que sólo cuatro nucleótidos determinen veinte aminoácidos diferentes. Por el camino, se haría el importante descubrimiento de que no es la información cifrada en el gen, en el propio tramo de DNA, la que es descifrada directamente por la célula para sintetizar una proteína, sino que la información de una de las cadenas de ese tramo de DNA es primero rescrita (transcrita) en otro polímero de ácido nucleico, también de una sola cadena, denominado RNA mensajero, hecho

también de cuatro nucleótidos casi idénticos químicamente a los que componen el DNA.

Sólo una breve mención también al tema de la regulación de la expresión de los genes. Durante la formación de un ser vivo, y en su vida adulta, los genes se expresan (generan el RNA mensajero y la proteína correspondientes) sólo cuando y donde resulta necesario, y el número de veces que conviene, de acuerdo a condicionantes endógenos, como ocurre en el complejo y asombrosamente reiterativo proceso de desarrollo embrionario de un organismo multicelular, o en respuesta a cambios ambientales. El trabajo pionero de los franceses Francois Jacob y Jacques Monod (años 60) despejó las vías de acceso al entendimiento de las bases moleculares de la regulación de la expresión génica, para la que se generalizó un modelo explicativo basado en la capacidad de ciertas proteínas para unirse a sitios determinados del DNA, en forma modulable por señales químicas específicas, para frenar o activar allí a una maquinaria general encargada de transcribir los genes.

En definitiva, a comienzos de los 70, el conocimiento de las bases moleculares de la herencia aparentaba ser completo, al menos en sus rasgos básicos (véase Figura 1).

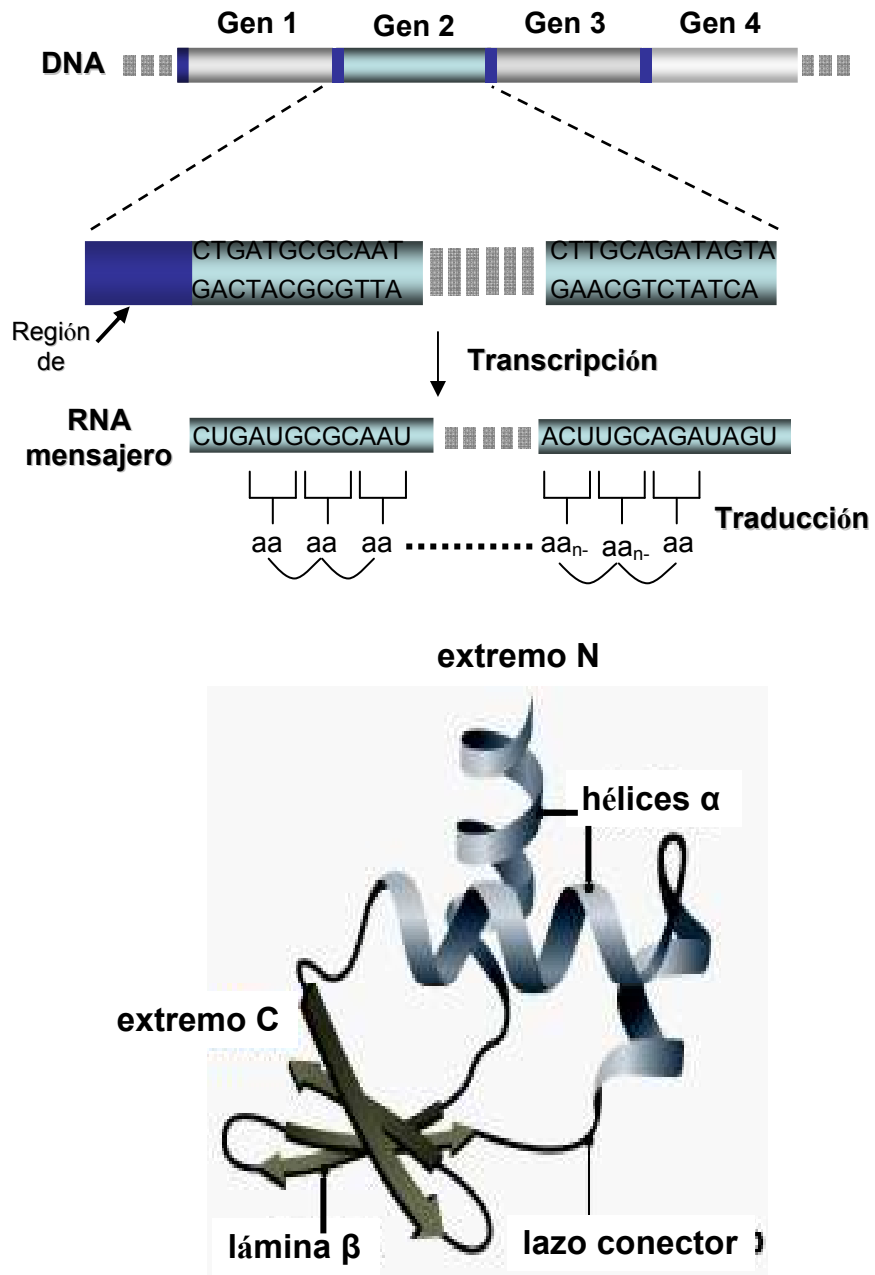


Figura 1 (leyenda en pág. 33)

No se vislumbraban, ni parecían necesarios, nuevos hallazgos esenciales. Cabía pensar, eso sí, en innovaciones técnicas, que realmente se han ido produciendo: métodos para analizar la expresión de cada gen, para clonarlos (reproducirlos) a voluntad, para modificarlos artificialmente, para transferirlos de forma efectiva de uno a otro organismo, para determinar su secuencia de nucleótidos, etc. etc. Tales instrumentos deberían servir tanto para impulsar los desarrollos aplicados que mencioné al principio como para acelerar el conocimiento de los detalles de la propia disciplina, para garantizar una provechosa y acelerada fase de investigación genético-molecular que, en cierto sentido, podía considerarse “de relleno”, bien que con retos de indudable interés, como, por citar algún ejemplo, identificar la función biológica de cada producto génico, o desentrañar en detalle los circuitos que regulan de la expresión génica durante el desarrollo embrionario. Como prueba de ese sentimiento de fin de ciclo, de haber establecido un paradigma definitivo, algunos de los protagonistas más inquietos, y todavía activos, de esta historia cambiaron de área de trabajo, y de forma radical, adentrándose, por ejemplo, en la Neurobiología, precisamente con la idea de que las nuevas herramientas y modos de aproximación experimental de la Genética moderna abrían nuevas posibilidades de exploración de un área todavía tan ignota, tan huérfana de paradigmas.

## II

### **Eso no era todo**

En verdad, no era tan sencillo. La exploración de los detalles, la investigación que se presumía de relleno, iba a servir en realidad para poner en entredicho al propio paradigma. El modelo de un gen como un tramo continuo de DNA, que ocupa un *locus* fijo de un cromosoma y cuyo texto cifrado es vertido, en un lenguaje químicamente similar, al de una molécula de RNA mensajero para ser finalmente traducido a la secuencia de aminoácidos de una proteína, el producto final funcional, se había basado, sobre todo, en experimentos realizados en unos pocos organismos sencillos, considerados como “modelos”, sobre todo en la bacteria *E. coli*. Para ser breve, diré que hoy pensamos que tal modelo no es aplicable en realidad, en términos generales, más que a las bacterias, en las que incluso no faltan ejemplos de algunos de los fenómenos que describiré a continuación y que han provocado la quiebra del ahora ya viejo paradigma de la Biología molecular. Tal quiebra se produciría con los estudios

moleculares de genes de distintos organismos multicelulares, desde la mosca *Drosophila* a los propios seres humanos, o de algunos de sus patógenos. Tales estudios avanzaron lentamente al principio, por el interés de cada grupo de investigación en uno o unos pocos genes implicados en fenómenos de su particular interés, y de forma explosiva después, con el acceso y análisis de la secuencia de nucleótidos del material genético completo (el *genoma*) de distintos organismos. Aparte de los cuantiosos datos que proporciona el análisis teórico de los “textos” genómicos mediante herramientas informáticas cada vez más incisivas, incluidos los interesantísimos datos derivados de la comparación entre textos distintos, la época actual de la *Genómica* ha propiciado el desarrollo de instrumentos para realizar observaciones experimentales masivas, como las derivadas del uso de los “micro-chips” de DNA. El resultado de todo ello ha sido la acumulación de numerosas pruebas experimentales en contra del que parecía arquetipo molecular definitivo del gen, avocado ahora a una profunda e inmisericorde revisión.

Pasan de una decena los novedosos fenómenos que contradicen el modelo molecular “clásico” del gen, pero, por brevedad, me limitaré a describir sólo dos de los que se producen de forma más general y resultan además especialmente “heréticos”. El primero se refiere al carácter continuo de la secuencia de nucleótidos de un gen. No es tal la norma de los genes de la mayoría de los

organismos. Muy al contrario, la secuencia de aminoácidos de una proteína concreta no deriva generalmente de una secuencia continua de material genético, sino de la yuxtaposición de distintos tramos de nucleótidos del DNA (denominados “exones”) separados entre sí por otros tramos (denominados “intrones”) que no “informan” sobre la secuencia de aminoácidos de dicha proteína. Para ello, una amplia región del DNA, que incluye intrones y exones, es transcrita ciertamente a una molécula primaria continua de RNA, pero ésta es sometida luego a un proceso que elimina los tramos correspondientes a los intrones y pega los que corresponden a los exones, dando lugar así a la molécula de RNA mensajero que será realmente traducida a proteína. No es raro que los tramos aparentemente no informativos de un gen, los intrones, ocupen mucho más DNA que los tramos informativos, los exones. Para mayor sorpresa, tampoco es infrecuente que se produzca un fenómeno conocido como procesamiento alternativo, o diferencial, por el que una única molécula de RNA primario genera distintos RNA mensajeros finales, formado cada uno de ellos por una combinación distinta de tramos separados de ese mismo RNA primario, pudiendo solaparse parcialmente las secuencias de los RNA mensajeros alternativos. Así pues, además de renunciar a la idea del gen como una secuencia continua de nucleótidos en el DNA, enfrentados al fenómeno del procesamiento alternativo del RNA, al hecho demostrado de que un mismo tramo de DNA determina más de un producto funcional (proteína) distinta, ¿debemos

aceptar que un mismo *locus* alberga varias unidades operativas hereditarias distintas, varios genes diferentes? Tal dilema, desde luego, es meramente antropocéntrico y a los seres vivos les importan poco nuestras limitaciones intelectuales. Lo que les importa, sin duda, son las posibilidades combinatorias que ofrece el procesamiento alternativo del RNA. A este respecto, quisiera insistir en que no se trata de un fenómeno infrecuente. En el caso del hombre, por ejemplo, la estimación actual es que cada RNA primario produce, como media, cinco RNA mensajeros finales distintos, que se traducirían en otras tantas proteínas diferentes (Figura 2).

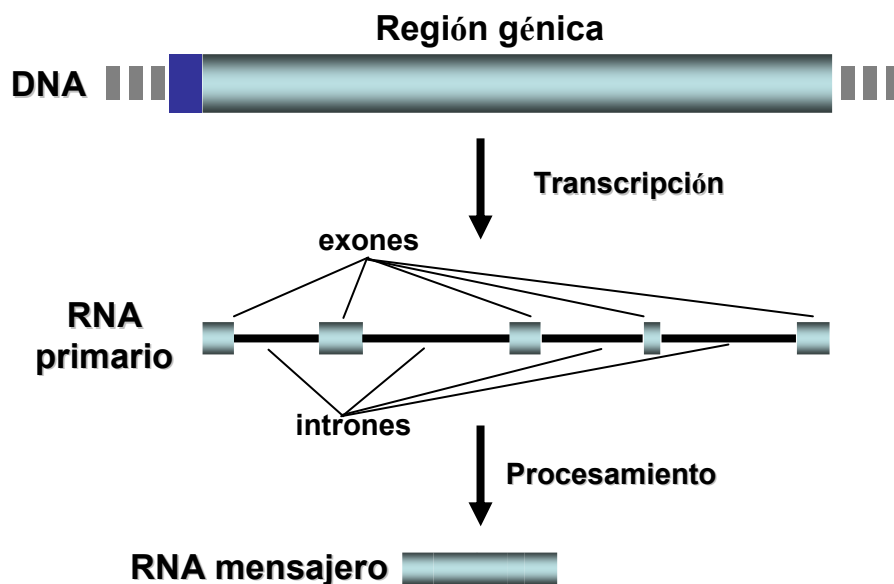


Figura 2 (leyenda en pág. 34)

No resisto la tentación de citar algún ejemplo especialmente llamativo (hay muchos donde escoger) de procesamiento diferencial del RNA. Citaré el caso del locus de la mosca *Drosophila* conocido como *Dscam*. Por cierto, la región correspondiente de DNA, además de estar presente en todas las especies de insectos cuyo genoma ha sido secuenciado, resulta similar a un *locus* del genoma humano relacionado con el síndrome de Down (*Dscam* corresponde, en realidad, a las siglas de *Down syndrome cell adhesion molecule*). El análisis experimental de *Dscam* demuestra que su RNA primario puede generar nada menos que 38.016 RNA mensajeros finales distintos y, por tanto, un número similar de proteínas diferentes, aunque éstas compartan entre sí bastantes dominios aminoacídicos. Tan vasto repertorio de isoformas proteicas juega, al parecer, un papel importante en el proceso de reconocimiento entre neuronas, tan importante para la formación de los intrincados circuitos del sistema nervioso de la mosca. Si acudimos a la clásica comparación, que casi todos los profesores de Genética hemos hecho alguna vez, entre información genética y texto literario, cabe admirarse de que para la célula resulten tener sentido tan numerosas combinaciones distintas, casi al azar, de las palabras de un breve escrito. De nuevo, es la comparación antropocéntrica la que resulta impropia. Obviamente, en el terreno de las funciones génicas, el “significado” de cada palabra

(aminoácido) es mucho más independiente de contexto que el de cada uno de nuestros vocablos. En todo caso, el procesamiento diferencial del RNA complica extraordinariamente el intento de establecer el número de genes de una especie, tema que tanto eco ha despertado en los medios de comunicación general, sobre todo al difundirse algunas cifras, derivadas normalmente de prematuros análisis bioinformáticos teóricos, que adjudican, por ejemplo, el mismo grado de complejidad genética a nuestra especie que a un gusano. Ciertamente, la diversidad de funciones de un organismo depende, sin duda, del número de sus unidades génicas operativas distintas, de la variedad de herramientas moleculares cifradas en su genoma, y del potencial de éstas para interactuar entre sí, aspecto éste de enorme relevancia si hablamos de la generación de complejidad y en el que no tengo tiempo de entrar. Pero el ejemplo de la región *Dscam* de *Drosophila* ilustra la dificultad actual para establecer de forma teórica cuantos y cuales son realmente los mensajes operativamente distintos escondidos en un genoma.

La segunda irreverencia con el dogma clásico a la que quisiera aludir aquí se refiere al papel de las moléculas de RNA. El paradigma de la moderna Biología molecular, tan tempranamente obsoleto, establecía que el “DNA hace RNA que, a su vez, hace proteína”, relegando al RNA a un mero papel de recadero. Ciertamente, ya en los albores de esta historia se habían descrito algunos tipos de RNA con funciones propias, aunque normalmente inespecíficas, como es el caso

de las moléculas de RNA que participan en el mecanismo celular que traduce todos los RNA mensajeros, u otras descubiertas más tarde que participan en el mecanismo que se encarga de la eliminación de intrones en los RNA primarios. Pero a partir de investigaciones iniciadas en plantas y hongos a primeros de los 90s, hoy estamos convencidos de que la información de muchos tramos del genoma de los organismos superiores, incluidos los animales y nosotros mismos, está destinada a generar pequeñas moléculas de RNA que operan per se, sin ser traducidas a proteína alguna, y que lo hacen de forma muy específica. Les llamaré aquí con el término genérico de mini-RNAs, aunque en la literatura científica se distinguen distintas clases de ellos, con distintos nombres.

Como les ocurre a las cadenas sencillas del DNA respecto a sus compañeras en la hélice doble, una determinada molécula de RNA, de una cadena sola, posee el atributo de la complementariedad, la capacidad para reconocer y unirse estrechamente a una segunda cadena de nucleótidos, sea en realidad de DNA o RNA, aunque sólo hace tal cosa si la secuencia de nucleótidos de esa segunda cadena resulta ser su “media naranja” química. Es esa propiedad la que propicia que un mini-RNA se asocie a una molécula específica de RNA mensajero de la misma célula, bloqueando su función normal, la de ser traducida para dar lugar a una proteína concreta. Según el caso, la unión del mini-RNA impide que el RNA mensajero sea reconocido

adecuadamente por la maquinaria general de traducción, o activa un mecanismo para degradarlo. Los mini-RNAs resultan, pues, de enorme importancia en la regulación del juego de proteínas y su concentración en el interior de las células. Refiriéndonos a los mamíferos, las investigaciones actuales indican que los mini-RNAs participan en la regulación de casi cada proceso en que ha sido investigada su presencia, incluyendo algunos de tanta importancia como la proliferación y diferenciación celular durante el desarrollo embrionario, la respuesta inmune o la oncogénesis. La estimación, también en mamíferos, es que los mini-RNAs regulan la expresión del 30 %, aproximadamente, de los genes que determinan proteínas. Debemos hablar, pues, de una clase nueva de genes, y muy numerosa, en los que, en último término, no está cifrada la secuencia de aminoácidos de una proteína sino la de nucleótidos de un mini-RNAs. También aquí se da el caso de que el producto final que actúa como mini-RNA, de poco más de 20 nucleótidos, se genera a partir de una molécula de RNA primaria más larga, que es luego procesada convenientemente. Se conocen mini-RNAs que proceden de la transcripción de un sitio del DNA (un gen), que le es propio, pero ocurre también que algunos mini-RNAs proceden de uno de los intrones eliminados de un RNA primario. Así, el mismo *locus*, el mismo tramo de DNA, resulta ser responsable tanto de una proteína, o varias (si su RNA primario es sometido a procesamiento alternativo), como de uno o varios mini-RNAs.

Si mediante el mero análisis informático de un genoma resulta difícil predecir cuantos y cuales son los genes que determinan proteínas, por culpa, sobre todo, del procesamiento del RNA y las complicaciones adicionales del procesamiento alternativo, predecir cuantos y cuales son los genes que determinan mini-RNAs resulta aún más arduo. No obstante, la identificación de ciertas pautas comunes en las secuencias de los cada vez más numerosos mini-RNA identificados y analizados experimentalmente, está sirviendo para generar programas informáticos cada vez mejor entrenados en predecir la presencia de genes de mini-RNA en los genomas conocidos, incluido el humano.

### III

#### **Un ejército (bien pertrechado) para entender un libro abierto**

Precedido de la secuenciación del genoma de otros organismos menos complejos, el reto técnico de determinar la secuencia completa de nucleótidos del genoma humano (3.000 millones de nucleótidos en 13 años de trabajo) produjo en su momento una gran admiración, bien que acompañada de una enorme decepción, por la aparente impenetrabilidad del texto genómico. El genoma humano podría ser un soberbio depósito de información, pero críptico casi por completo. En las fases históricas a las que me referí al principio, sin haberlo visto nunca, y aún antes de saber en que idioma estaba escrito, unos pocos pioneros habían penetrado en muchos y fundamentales secretos del libro de la vida. Ahora, abiertas todas y cada una de sus páginas, nadie era capaz de entender más que unas escasísimas frases de aquí y de allá. Uniendo los datos de los trabajos previos interesados en

analizar experimentalmente distintas regiones del genoma y los obtenidos al aplicar al texto completo los procedimientos informáticos predictores más inteligentes, incluida la comparación con todos los genes ya conocidos de cualquier organismo, resultaba que no entendíamos más allá del 1 % del texto genómico humano. Y lo que es peor, el resto parecía un completo galimatías.

Algunos emplearon el término “DNA basura” para referirse al inmenso volumen de información aparentemente inútil del DNA humano (y de nuestros congéneres en el mundo de los vertebrados), pero la mayoría se inclinaba por el reconocimiento crítico de nuestra deficiente capacidad de lectura de los mensajes cifrados en nuestro propio texto genético. Como ilustra el hallazgo reciente del papel y abundancia de los mini-RNAs que acabo de mencionar, podrían existir en el DNA otras muchas unidades funcionales a cuya lógica operativa seamos completamente ajenos todavía. La cuestión esencial es cómo reconocerlas. Aunque con dificultades, la clave que relaciona una secuencia de nucleótidos con la de aminoácidos de una proteína, o los aspectos que mencioné anteriormente reconocibles en los mini-RNA, permiten predecir de manera teórica las unidades operativas correspondientes del genoma, al menos provisionalmente, a la espera del necesario análisis experimental, que al menos podemos dirigir hacia sitios concretos. Pero ¿cómo identificar y catalogar esas otras posibles unidades funcionales ignotas? Tal reto es el que se ha

propuesto un proyecto internacional conocido como ENCODE (*ENCyclopedia Of Dna Elements*) y cuyo objetivo es nada menos que catalogar todos los elementos funcionales del genoma humano.

El proyecto ENCODE, auspiciado y financiado en gran parte por el “US National Human Genome Research Institute”, se inició el año 2003, con una primera fase piloto concentrada en el estudio del 1 % de todo el genoma humano, unos 30 millones de nucleótidos. Esa cifra corresponde a la suma de 44 regiones concretas dispersas por todo el genoma. La mitad de ellas se escogieron por el criterio de ser regiones ya relativamente bien caracterizadas, como se dice en el argot genómico, con muchas “anotaciones” sobre funciones génicas, resultado de trabajos experimentales previos, mientras que la otra mitad se escogió de forma más o menos aleatoria.

Los responsables del proyecto ENCODE hicieron un llamamiento internacional para seleccionar grupos de investigación de todo el mundo que demostraran su capacidad para llevar a cabo algún tipo de análisis relevante, experimental o bioinformático, aplicable al DNA, y que pudieran hacerlo en forma suficientemente masiva y eficiente como para cubrir esos 30 millones de nucleótidos en un tiempo razonable. Se seleccionó así a un buen número de grupos de investigación de 10 países distintos, formados en su conjunto por 308 científicos, tanto analistas experimentales como expertos en computa-

ción biológica. No voy entrar desde luego en la descripción de las diversas formas de asalto, hasta 200 distintas, que tan numeroso ejército, equipado con las más poderosas armas de análisis experimental y los más modernos programas informáticos, ha lanzado contra la playa de “Normandía” del territorio genómico humano, contra la puerta de entrada al castillo del conocimiento que más ansiamos conquistar. Sólo como ilustración de los ataques experimentales, diré que se ha buscado y caracterizado cada molécula de RNA que se genera en el DNA del territorio ENCODE, tanto RNAs primarios como RNAs ya procesados, empleando además varias líneas celulares de uso frecuente en el laboratorio o cualquiera de una decena de tejidos humanos normales distintos. O como ejemplo del ataque informático, diré que se han realizado exhaustivas comparaciones del texto seleccionado del genoma humano con el texto correspondiente, también conocido, de 9 especies de primates, 14 de otros mamíferos y 5 especies de otros vertebrados (1 ave, 1 anfibio y 3 peces). Uno de los objetivos de esas comparaciones ha sido detectar tramos de secuencia de DNA particularmente parecidos de unos organismos a otros (“resistentes a la evolución” en el argot de los especialistas), confiando que tal conservación a lo largo de millones de años denote la presencia de alguna unidad operativa esencial.

La fase piloto del proyecto ENCODE, que ha durado unos tres años, ha tenido un coste de algo más de 42 millones de dólares. Como producto de las 200 técnicas experimentales o aproximaciones informáticas distintas empleadas, se han generado nada menos que un total de 603 millones de datos puntuales, 400 millones de ellos de tipo experimental. Como cualquiera puede figurarse fácilmente, la integración, estudio analítico e interpretación de tal volumen de datos no es un asunto baladí. Aún así, ha sido ya objeto de varias publicaciones, una que resume el conjunto completo de los resultados y recoge las conclusiones más fundamentales, aparecida en la revista *Nature* el pasado 14 de junio de 2007, y varias otras que presentan y analizan distintos subconjuntos de datos, aparecidas en el mismo mes de junio en un número especial de la revista *Genome Research*.

¿Y qué se dice en esas publicaciones? ¿Qué se ha descubierto? Personalmente creo que nada especialmente relevante en cuanto al problema esencial de ampliar nuestra capacidad general de comprensión del mensaje genómico humano. Desde luego, el número de descripciones y correlaciones es abrumador, pero la impenetrabilidad original de vastas extensiones de nuestro DNA permanece intacta. Aunque meramente descriptivos, algunos hallazgos

resultan ciertamente curiosos. Por citar algún ejemplo de resultado experimental, diré que el proyecto ENCODE ha demostrado que prácticamente cada nucleótido del DNA está representado en alguna molécula de RNA primario. Ello incide en que no hay tal cosa como el DNA basura. ¿Porqué molestarse la célula en transcribir un mensaje sin sentido? Pero cual sea ese sentido sigue siendo una incógnita. Como también es una incógnita que, como ha descubierto el proyecto ENCODE, un número muy significativo de las moléculas de RNA primario sean mucho mayores de lo que cabría esperar, mayores que las más largas conocidas hasta ahora. Así, es frecuente detectar moléculas de RNA primario que cubren las que creíamos varias regiones génicas de transcripción independiente, y cada una de las cuales ya nos parecía bastante compleja. Es más, algunas de esas largas moléculas de RNA primario conectan entre sí dos genes muy alejados que determinan proteínas conocidas, de manera que tal RNA primario genera, al procesarse, un RNA mensajero final que combina regiones informativas, exones, de esos dos genes, que no sólo están situados a gran distancia, sino que flanquean a varios otros genes que también determinan proteínas conocidas (Figura 3).

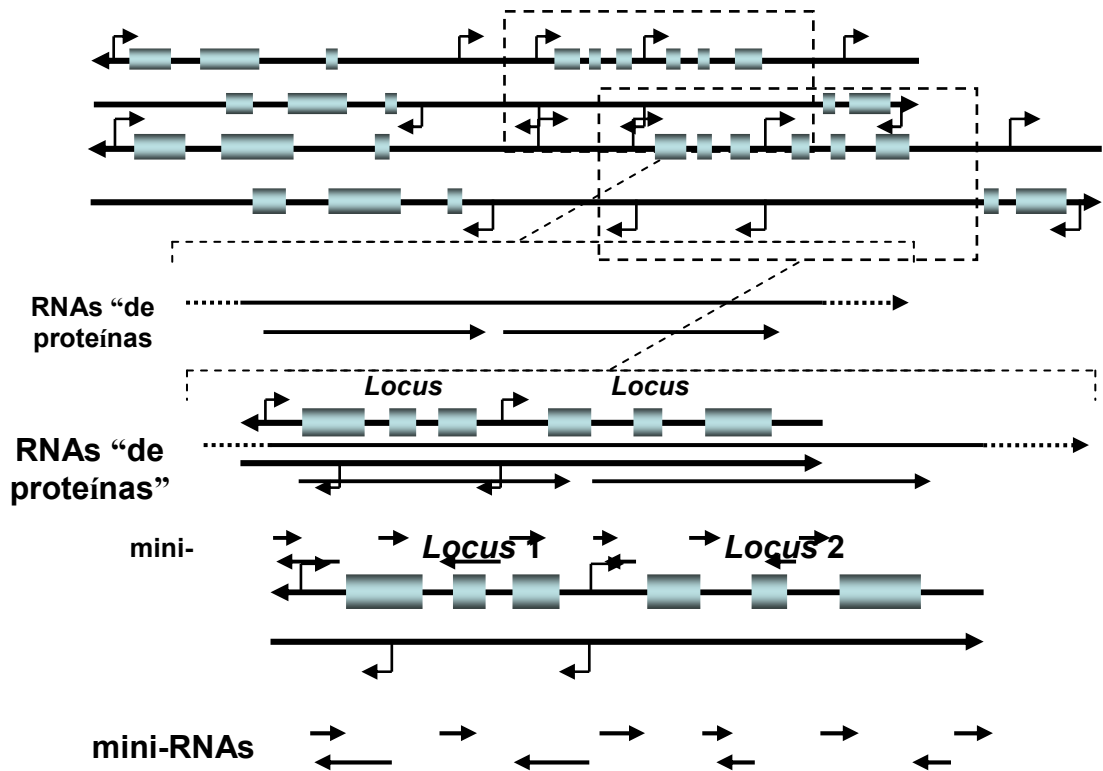


Figura 3 (leyenda en pág. 34)

Por supuesto, el proyecto ENCODE ha permitido identificar una numerosa colección de novedosas moléculas de RNA ya procesadas, incluyendo una buena batería de nuevos RNA pequeños cuya función queda por desvelar.

Como ejemplo de resultado de los análisis informáticos, diré que el proyecto ENCODE ha desvelado que del orden del 5 % del genoma humano corresponde a regiones que se han resistido de forma especial a los cambios mutacionales desde que se produjo la separación evolutiva de los distintos tipos de vertebrados. El 60 % de esa fracción de DNA de evolución constreñida puede relacionarse con alguna anotación funcional previa (40%) o al menos con alguno de los nuevos elementos anotados por el propio proyecto ENCODE, aunque sea de función desconocida (20 %). Pero nada sabemos del 40 % que resta de ese DNA especialmente conservado (unos 40 millones de nucleótidos).

Como ocurrió en su momento con el resultado del propio proyecto de secuenciación del genoma humano, los resultados del proyecto ENCODE parecen haber generado más preguntas que respuestas. En realidad, los propios responsables del proyecto reconocen su incapacidad para extraer el máximo partido posible de una información ciertamente tan voluminosa y diversa. Confían por ello en que el acceso completo a los datos crudos de ENCODE, libremente disponibles en una página web global del proyecto y en otras páginas web más parciales, agrupados por categorías y referidos a cada posición estudiada del genoma, atraiga a otros investigadores

de cualquier lugar del mundo. Tales investigadores pueden utilizarlos de referencia en conexión con una investigación ya en marcha sobre alguna zona del DNA del territorio ENCODE, interesarse en caracterizar en detalle algunos de los nuevos elementos detectados en el proyecto, o aplicar quizás nuevos y más inteligentes instrumentos bioinformáticos que desvelen algún atributo o propiedad general de la información genética que quizás está ya en el conjunto global de los datos de ENCODE, ahí a la vista, esperando a ser descifrado, a ser leído correctamente.

## Referencias

Se incluyen sólo algunas citas de especial interés sobre los aspectos que se consideran más llamativos de este discurso.

### Sobre el procesamiento alternativo del RNA

Graveley (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genetics*, **17**: 100-107.

Zipursky *et al* (2006) Got diversity? Wiring the fly brain with Dscam. *TrendsBiochem. Sci.*, **31**: 581-588.

### Sobre los RNA “pequeños”

Brodersen y Voinnet (2006) The diversity of RNA silencing pathways in plants. *Trends Genetics*, **22**: 268-280.

Filipowicz *et al.* (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Rev. Genetics*, **9**: 102-114.

### Sobre el proyecto ENCODE

The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**: 636-640.

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

Número especial de la revista *Genome Reseach*. El número de esta revista (volumen 17) correspondiente al mes de junio de 2007 recoge varios artículos sobre distintos aspectos del proyecto ENCODE.

## Leyendas de las figuras

**Figura 1.-** El paradigma molecular clásico del gen. Como se representa en la parte superior de la figura, los genes serían tramos continuos de DNA, alineados unos junto a otros, sin solapamientos y sin apenas separación entre ellos. Debajo se representa una ampliación del hipotético “gen 2”, en la que se muestra la secuencia de nucleótidos de cada cadena de la doble hélice. Un tramo de secuencia (no indicado) de uno de sus extremos (región de control) es reconocido por proteínas reguladoras que, de forma modulada por señales químicas, controlan la transcripción del resto de la secuencia en una molécula de RNA mensajero. La secuencia de nucleótidos de este RNA es luego traducida, de acuerdo con una clave que asigna cada uno de los veinte aminoácidos que componen las proteínas a cada tres nucleótidos. Las condiciones fisiológicas de la células, incluyendo la posible acción de otras proteínas, determinan que una secuencia concreta de aminoácidos adquiera una conformación tridimensional concreta. Esta conformación otorga a la proteína su condición de herramienta celular para llevar a cabo alguna función, sea catalizar una reacción química, incorporarse a la membrana para actuar de receptor de alguna señal externa, participar en la propia maquinaria de transcripción, o de traducción, reconocer específicamente algún sitio del DNA para regular la expresión de algún gen, etc. etc. Debajo se representa la configuración

tridimensional de una proteína hipotética, en la que se aprecian dos elementos estructurales típicos de las proteínas, como son las “hélices alfas” o las “láminas beta”.

**Figura 2.-** Procesamiento del RNA. Con mucha frecuencia, en los organismos no bacterianos, el producto inicial de la transcripción de un gen (RNA primario) es luego procesado, cortándose de él varios tramos intermedios, llamados **intrones**, y pegándose los otros tramos, denominados **exones**. Sólo estos forman parte del RNA mensajero final, cuya secuencia de nucleótidos será traducida a la de aminoácidos de una proteína. Los intrones no “informan”, pues, sobre dicha secuencia y, por tanto, el gen, como unidad hereditaria operativa, no se corresponde con un tramo continuo de DNA.

**Figura 3.-** Representación esquemática de una región genómica humana. Arriba se representan las dos cadenas de DNA de una región hipotética del genoma humano. Los bloques representan los tramos de DNA que se corresponden con “exones” de algún RNA mensajero, y las flechas indican puntos de inicio de la transcripción. El proyecto ENCODE ha desvelado que estos puntos son mucho más frecuentes de lo que se pensaba y que algunos de ellos generan moléculas de RNA primario extraordinariamente largas. Debajo se representa una ampliación de la zona recuadrada. Sobre las dos cadenas del DNA se indican algunas moléculas de RNA primario

destinadas a generar proteínas. Como se aprecia, además de las dos moléculas de RNA que cubren los *loci* 1 y 2, cada uno de los cuales determinaba, por ejemplo, una proteína conocida (o varias, si el RNA primario es objeto de procesamientos alternativos), aparece una molécula de RNA mucho más larga, cuya transcripción se inició en algún sitio a la izquierda del *locus* 1 y terminará en algún sitio a la derecha del *locus* 2. Las pequeñas flechas que aparecen debajo representan múltiples moléculas de mini-RNAs detectadas experimentalmente en el proyecto ENCODE, muchas de ellas previamente desconocidas (la distinta dirección de la flecha indica si la secuencia del RNA se corresponde con la de una o la otra cadena del DNA).