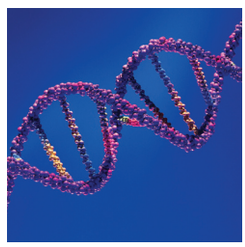
Genes, genomas y metagenomas: de Mendel a Venter

Rosa María Gutiérrez y Enrique Merino



Frecuentemente se relaciona al genoma con "el libro de la vida". Esta analogía se debe a que ambos pueden ser leídos secuencialmente, de principio a fin, una letra tras otra, y porque en el genoma se encuentra la información necesaria para hacer de cada organismo un ser vivo. El alfabeto con que está escrito este libro es extremadamente sencillo. Consta de tan sólo cuatro caracteres a los que se les han asignado las letras A, C, G, T, por el nombre de las bases nitrogenadas a las que hacen referencia: adenina, citosina, guanina y timina. Hasta la fecha hemos leído en su totalidad más de medio millar de genomas de organismos diferentes, desde los más sencillos, como el de las bacterias, hasta los más complejos, incluyendo el humano.

¿Qué tan bien entendemos lo que hemos leído en los genomas? ¿Qué tan complejo ha resultado el contenido del libro y cuáles son los elementos conceptuales que hemos desarrollado para su análisis? Sin duda alguna, una piedra angular en este proceso cognoscitivo fue el trabajo de Gregor Mendel, quien en 1865 elaboró las primeras bases matemáticas para la descripción de los procesos de la herencia y la predicción de las características observables de un organismo (fenotipo), a partir de lo que llamó elementos de información hereditaria (genotipo). Unas décadas más tarde, Walter Sutton demostró que los elementos de información hereditarios descritos por Mendel se encontraban en el interior de todas las células y que presentaban estructuras alargadas, a las que llamó "cromosomas". De manera análoga a los distintos capítulos que contiene un libro, el genoma de un organismo tiene desde uno hasta varios cromosomas. Pocos años después, Wilhelm Johannsen acuñaría el nombre de "genes" para la unidad de información hereditaria contenida en los cromosomas. No fue hasta 1928 cuando Fred Griffith realizó la primera transformación *in vitro* de un organismo, al convertir cepas no patógenas de *Diplococcus pneumoniae* en patógenas, utilizando un elemento transformador proveniente de la cepa infecciosa.

Después de una década y media, Oswald Avery, Colin MacLeod y Maclyn McCarty identificaron que este elemento transformador era el ácido desoxirribonucleico, también llamado ADN. Inesperadamente para Avery y sus colaboradores, la molécula donde residía la información genética, el ADN, era mucho más simple de lo imaginado. Solamente lo constituían cuatro elementos: un azúcar sencillo del tipo de la desoxirribosa, radicales fosfóricos y las cuatro bases nitrogenadas antes mencionadas:

A, C, G, T. Estudiando el ADN de distintos organismos, Erwin Chargaff demostró en 1950 que existía una relación equimolar en estas bases: por cada molécula de adenina existía una de timina y por cada molécula de citosina existía otra de guanina. A principios de la década de los cincuenta, un grupo de investigadores, bajo la dirección de Maurice Wilkins y Rosalind Franklin, se dio a la tarea de analizar el patrón de difracción que sufren los rayos X al pasar a través de un cristal de ADN, lo que les permitió sugerir que el ADN era una molécula helicoidal que se repetía cada 34 angströms (A) y tenía un ancho constante. Con esta información, en 1953 James Watson y Francis Crick finalmente propusieron la estructura de la doble hélice del ADN, con lo que se inició una nueva era en las ciencias biológicas.

Si bien el alfabeto del material genético había sido descifrado, aún quedaban por resolver varias interrogantes: ¿dónde radica la información necesaria para que el ADN se copie a si mismo? o más aún ¿cómo puede ser utilizada esta información para codificar macromoléculas más complejas, como las proteínas? Las aportaciones conceptuales de diversas investigaciones realizadas durante la segunda mitad de la década de los cincuenta y principios de los sesenta, permitieron a Francis Crick formular el dogma central de la biología molecular, que describe el flujo de la información genética. En este postulado se explica que el ADN puede dirigir su propia síntesis en un proceso llamado replicación, en el que cada una de las dos hebras del ADN sirve como molde para la síntesis de su correspondiente hebra complementaria. En un paso posterior, llamado *transcripción*, el ADN también servía como patrón para la síntesis del ARN, otro ácido nucleico similar al ADN que posee ribosa en lugar de desoxirribosa en su cadena lateral. Finalmente, la información contenida en el ARN podría ser empleada para dirigir la síntesis de las proteínas en un proceso llamado *traducción*.

A diferencia del ADN y el ARN, las proteínas están formadas por aminoácidos. Hoy en día sabemos que la composición y orden secuencial de los aminoácidos determinan la estructura y función de las mismas, que a su vez están predefinidas por la secuencia nucleotídica del ADN. Por lo anterior, una de las primeras inferencias del análisis de la secuencia de un genoma consistió en determinar el posible conjunto de proteínas que tendría el organismo en cuestión. De ahí el interés de secuenciar el mayor número de genomas de la manera más completa.

Revolución genómica: la era de la secuenciación

Paradójicamente, los primeros genomas en secuenciarse totalmente no correspondieron a organismos vivos, sino a los virus bacterianos phi-X174 y lambda, secuenciados al final de la década de los ochenta. Pese a que el tamaño de estos genomas es de apenas algunos miles de pares de bases, constituyen un ejemplo de cómo la secuencia de un genoma completo puede "armarse" a partir de la de varios fragmentos de menor tamaño. A partir de entonces, las mejoras continuas en las técnicas de secuenciación han permitido que el crecimiento de las bases de datos de ácidos nucleicos en nuestros días sea exponencial. Como consecuencia de estos avances metodológicos, en 1995 se concluyó por primera vez la secuenciación de un organismo vivo: la bacteria patógena *Haemophilus influenzae*, agente causal de la gripe. Con este monumental logro se inició una nueva era en la ciencia: la era genómica.

En la actualidad se ha completado la secuencia de más de quinientos genomas. Éstos incluyen miembros de las tres ramas del árbol de la vida: eubacterias, arqueobacterias y eucariotes. Al inicio de la era genómica se observó una clara tendencia a secuenciar genomas pequeños de organismos patógenos, como el parásito *Mycoplasma genitalium,* relacionado con uretritis no gonococal. A la fecha, esta bacteria es el organismo con el menor genoma conocido, ya que solamente cuenta con medio millar de genes, esto lo hace un modelo interesante para entender las características metabólicas y genéticas de los primeros seres vivos. Un año después se publi-

có la secuencia del genoma de *Methanococcus jannaschii*, que resultaba de gran importancia por ser el primer representante del grupo de las arqueobacterias, con el cual se realizaron los primeros análisis comparativos entre genomas de arqueobacterias y eubacterias. A este par de genomas lo siguieron otros de organismos patógenos, como *Helicobacter pylori,* asociada con la gastritis crónica y carcinomas gástricos, *Neisseria meningitides, Treponema pallidum* y *Vibrio cholerae*, agentes causales de la meningitis, sífilis y cólera, respectivamente. El genoma de *Plasmodium falciparum*, agente causal de la malaria, y el de su organismo transmisor, el mosquito *Anopheles gambiae*, fueron secuenciados de manera casi simultánea en 2002, con lo que se abrió una nueva puerta para el desarrollo de vacunas y el diseño de nuevas medicinas para combatir esta grave enfermedad.

No todos los genomas que han sido secuenciados corresponden a organismos nocivos. En 1996 fue publicada la secuencia del genoma de un organismo que ha sido utilizado por siglos en la industria alimenticia: la levadura *Saccharomyces cerevisiae*. Este acontecimiento, de gran importancia científica, constituyó el primer ejemplo de la secuenciación total de un organismo eucariote. Los adelantos en las técnicas de secuenciación, así como el desarrollo de mejores programas de cómputo para su análisis, permitieron la determinación de genomas de mayor tamaño: la primera planta secuenciada *Arabidopsis thaliana*, dos variedades de arroz, *Oriza sativa japonica* y *Oriza sativa indica,* el nematodo *Caenorhabditis elegans*, la mosca *Drosophila melanogaster*, el pez cebra *Danio rerio* y el pez globo *Fugu rubripes*, por mencionar algunos ejemplos que sentaron las bases del siguiente gran paso en las ciencias genómicas: la secuenciación del genoma humano.

El genoma humano

Mucho se ha hablado sobre el proyecto de secuenciación del genoma humano. Su realización ha requerido la colaboración de varios centros científicos internacionales. Cabría preguntarnos si nuestro interés antropocéntrico ha tenido el fruto esperado. Con la esperanza de que la determinación de este genoma ayude a la implementación de nuevas terapias génicas y a la realización del diagnóstico molecular de diversas enfermedades hereditarias, la industria genómica mundial ha invertido más de 60 mil millones de dólares en la investigación del genoma humano. Paradójicamente, menos de cinco por ciento de los tres mil millones de bases del genoma humano codifica genes. El resto de la secuencia contiene un número inesperadamente alto de secuencias repetidas que dificultan la labor de identificación de regiones codificantes y, en consecuencia, reduce la utilidad de la secuenciación de nuestro genoma.

Es importante considerar que, tanto en el genoma humano como en el de otros organismos superiores, los genes transcritos en ARNs mensajeros pueden ser editados y procesados mediante la eliminación de regiones llamadas *intrones*. Los fragmentos restantes o *exones* son unidos para dar lugar a ARNs maduros. Dependiendo de este proceso de edición, un mismo gen puede dar origen a diferentes variantes de ARNs que, una vez traducidos, darán origen a sus correspondientes proteínas. Se estima que cerca de 38 por ciento de los genes en el humano son procesados de esta manera y que en promedio dan origen a 3.7 transcritos distintos por gen. Determinar correctamente los sitios de corte de intrones a partir del análisis de las secuencias nucleotídicas es, por tanto, de gran importancia. A pesar de que se han obtenido avances significativos en el desarrollo de métodos computacionales para determinar esos lugares de corte, un gran número de predicciones sobre la maduración de ARNs mensajeros no corresponde a los datos de secuenciación de fragmentos de ADN complementario (ADNc) o al de los ARNs mensajeros procesados *in vivo*. Actualmente se realizan microarreglos con oligonucleótidos sintéticos de secuencias que potencialmente corresponden a exones. Los resultados de tales estudios serán de gran utilidad para que sean elaborados mejores programas de cómputo y se logre

una mejor predicción *in silico* de tan complejo proceso biológico.

Sumada a la dificultad del reconocimiento de los genes y al de la predicción del procesamiento de sus correspondientes ARNs, la asignación de las funciones biológicas de las proteínas para las que codifican ha sido igualmente difícil. A la fecha existe un gran número de genes cuya función es aún desconocida por no tener similitud con genes de otros organismos que hayan sido caracterizados. Se estima que el número de genes en el genoma humano sea de 30 mil. Este número es considerablemente más pequeño de lo inicialmente esperado y corresponde solamente al doble del número que poseen los genomas de organismos menos complejos, como el del gusano *Caenorhabditis elegans* o el de la mosca *Drosophila melonogaster,* y tan sólo seis veces mayor al de la levadura, organismo eucariote unicelular. Por tanto, es claro que las grandes diferencias en términos de complejidad y capacidad que tenemos los humanos respecto al resto de los organismos está dado no tan sólo por el número de genes que poseemos. Al parecer las diferencias radican en el hecho de que los genes humanos son más complejos en su composición, ya que están formados por un mayor número de dominios funcionales y una serie de combinaciones nuevas de los mismos. La tarea de descifrar esta información, a través de su análisis, es uno de los nuevos retos a resolver en un futuro próximo.

La secuenciación del genoma humano, junto con la de otros cientos de genomas caracterizados, muestra claramente nuestra capacidad para secuenciar organismos aislados y perfectamente definidos, y establece las bases para un siguiente paso, que es la secuenciación simultánea de los genomas de organismos que pertenecen a un mismo ecosistema, lo que actualmente conocemos como *metagenoma.*

De los ecosistemas a los genes: secuenciación de metagenomas

El Instituto de Alternativas en Energía Biológica (IBEA), en Estados Unidos, ha empezado la secuenciación masiva de los genomas de miles de microorganismos que viven en el mar de los Sargazos. Sin duda alguna es uno de los proyectos genómicos más importantes de nuestros días, del que se espera obtener la secuencia de un *metagenoma,* o conjunto de genomas de organismos que constituyen un ecosistema. A pesar de que este proyecto fue realizado en tan sólo tres años, generó más información, en términos de secuencia genómica, de la que se contaba en su momento, ya que se pasó del orden de centenas de genomas parcialmente secuenciados al de decenas de miles. De acuerdo con el responsable principal de este proyecto, Craig Venter, puede existir mayor información genética en el conjunto de organismos que viven en un litro de agua de mar, que la existente en el genoma humano.

Con esta nueva visión de las ciencias genómicas en la caracterización de ecosistemas se espera obtener un catálogo de la diversidad del mar. El principal desafío metodológico del proyecto radica en que la mayoría de los microorganismos de este ecosistema no crecen en condiciones de laboratorio y, por tanto, no es posible aislar individualmente a los organismos para su secuenciación. En este caso, la secuenciación de los genomas se realizará con muestras tomadas directamente del ecosistema y, posteriormente, mediante el análisis por computadora con algoritmos similares a los empleados en la secuenciación del genoma humano: cada fragmento de ADN secuenciado es identificado y ensamblado en su correspondiente genoma. Para entender la complejidad de este proceso habría que imaginar el armado de miles de rompecabezas diferentes, cuyas piezas se han mezclado entre sí. Evidentemente esta tarea implicaría mucho más trabajo que resolver el mismo número de rompecabezas individualmente. Actualmente se cuenta con la secuencia de más de medio centenar de metagenomas entre los que se encuentran los correspondientes a terrenos de cultivos, suelos de minas, comunidades del fondo del mar, entre otros.

¿Qué aprendemos de un proyecto metagenómico como los anteriormente mencionados? Posiblemente realizaremos la caracterización de nuevos genes que revelen el tipo de compuestos químicos utilizados por estos microorganismos como alimentos y el tipo de compuestos que pueden generar, así como describir las nuevas vías metabólicas que tienen e identificar cuáles de ellas intervienen en la conversión de la luz solar en energía. Se espera que este conocimiento sea empleado en nuestra vida diaria y que permita, entre otras cosas, el diseño de metodologías alternativas para la generación de energía, generación de nuevos fármacos así como la síntesis de nuevos compuestos.

Genómica comparativa

Las proteínas son los principales actores de las funciones celulares. Durante su proceso evolutivo pueden aceptar mutaciones y continuar llevando a cabo su misma función en diferentes organismos (*ortólogas*), o evolucionar para adquirir nuevas funciones dentro del mismo organismo (*parálogas*). En cualquier caso, ortólogas o parálogas, un par de proteínas es considerado como homólogo cuando comparten un ancestro común y sus secuencias, tal como las conocemos hoy, son el producto de un largo proceso evolutivo de mutaciones y selección, con base en la estructura y función de las mismas.

Las huellas de ese proceso evolutivo pueden ser observadas cuando las secuencias de algunos miembros de la familia son comparadas en pares, o bien, cuando son comparadas simultáneamente, a lo que se ha llamado "alineamientos múltiples de secuencias". Por lo común, tales comparaciones son efectuadas con secuencias de aminoácidos, en vez de las secuencias nucleotídicas de los genes que las codificaron, ya que las primeras son más informativas debido a las diferentes propiedades fisicoquímicas que poseen los aminoácidos. Dada la riqueza de información que guardan los alineamientos, han sido utilizados en bioinformática para realizar diferentes tipos de análisis, entre los que se encuentran los relacionados con procesos evolutivos dentro de la familia, y que generalmente son representados como árboles filogenéticos, o bien, en la determinación de la estructura secundaria de las proteínas mediante algoritmos de redes neurales que consideran las tendencias de cada aminoácido a formar parte de elementos estructurales de acuerdo con su entorno, o aun para identificar a miembros lejanos de la familia cuyo parentesco es indetectable cuando se compara individualmente con un elemento de la familia, pero que es identificado si se considera la información acumulada del alineamiento múltiple, utilizando modelos probabilísticos llamados "cadenas de Markov escondidas".

Adicionalmente, los aminoácidos conservados en un alineamiento múltiple indican que son esenciales para la estructura y función de los miembros de la familia, mientras que los residuos conservados en algunos miembros de la misma suelen estar relacionados con la función y la filogenia de una subfamilia (determinantes de árbol). Los residuos conservados en un alineamiento múltiple no son los únicos residuos informativos. También lo son aquellos residuos que varían de manera coincidente en dos familias de proteínas; es decir, aquellas mutaciones correlacionadas en las dos familias de proteínas. Esas mutaciones indicarían una co-evolución y se consideran una evidencia indirecta de que tales proteínas interaccionan físicamente.

Los análisis de secuencias antes mencionados consideran la información de la cadena peptídica como una secuencia lineal de aminoácidos. Sin embargo, sabemos que la estructura que adoptan las proteínas en el espacio es esencial para su función; por ello, considerar la estructura tridimensional de las proteínas en los diferentes análisis de secuencia resulta de gran valor. Actualmente se conoce la estructura tridimensional de un poco más de 50 mil proteínas, y existen diferentes proyectos genómicos cuyo objetivo es determinar, por métodos experimentales, la estructura de cada proteína que constituye el proteoma de un organismo. Estos proyectos forman parte de la genómica estructural. El trabajo que implica determinar

la estructura tridimensional de una proteína es muy superior al de obtener la secuencia del gen que la codifica. Por lo tanto, el número de proteínas con estructura desconocida excede en mucho al número de las conocidas. Las reglas que gobiernan el plegamiento de una proteína son muy complejas, por lo que la inferencia de la estructura de una proteína exclusivamente a partir de su secuencia se limita a algunas proteínas de tamaños muy pequeños.

Afortunadamente se han desarrollado métodos computacionales para inferir la estructura de una proteína a partir de su secuencia mediante su comparación con la de aquellas proteínas similares, con estructura tridimensional determinada. No obstante, se espera que en un futuro próximo las bases de datos de estructura de proteínas crezcan significativamente y faciliten la asignación de la estructura, así como la función de proteínas identificadas en los proyectos de secuenciación genómica.

La forma también importa: curvatura del ADN

Como se mencionó antes, el primer modelo de la estructura del ADN propuesto por James Watson y Francis Crick suponía que la doble hélice de ADN ocupaba el espacio de un cilindro regular cuyas bases nitrogenadas se apilaban en planos paralelos. Hoy en día sabemos, por resultados experimentales y teóricos, que las bases nitrogenadas pueden presentar ciertos desplazamientos lineales y angulares que dan como resultado fragmentos curvos de ADN. La magnitud de esos desplazamientos puede evaluarse mediante el uso de matrices de rotación y traslación, cuyos valores han sido determinados experimentalmente. A este tipo de curvatura se le ha llamado "curvatura estática" o "curvatura intrínseca" para diferenciarla de la curvatura del ADN introducida a través de ciertas proteínas.

La relevancia de las regiones curvas en el ADN en diferentes procesos biológicos, como recombinación, replicación y regulación transcripcional, ha sido investigada por más de 30 años. En la mayoría de estos estudios experimentales las regiones de ADN curvo analizadas han sido pequeñas. El uso de algoritmos matemáticos en la predicción de la geometría del ADN abre la posibilidad de extender el estudio de curvatura del ADN de *loci* discretos a regiones de ADN de mayor longitud, incluyendo el análisis de las secuencias de genomas enteros. Recientemente, estudios *in silico* de los genomas totalmente secuenciados han establecido que al curvatura promedio de los genomas varía considerablemente de organismo a organismo, y que es una función directa de la frecuencia global de los dinucleótidos del genoma en cuestión. Cabe mencionar que la capacidad para evaluar el grado de curvatura del ADN a partir de su secuencia nucleotídica ha permitido corroborar su función como elemento activo del proceso de la regulación transcripcional, y se ha podido demostrar que esta función del ADN curvo puede estar conservada entre genes ortólogos de organismos filogenéticamente distantes, aunque estas regiones no presenten conservación de secuencia, lo que implica que la propiedad geométrica del ADN es biológicamente significativa.

Construcción de redes y modelos: biología integrativa

Una tendencia generalizada en las ciencias biológicas del siglo XX fue la de estudiar los componentes celulares y sus funciones de manera independiente, dentro de un esquema que podría llamarse "reduccionista". Con base en este enfoque se ha generado un conocimiento significativo dentro de cada una de las ramas de la biología, y se espera que el impacto de nuevas tecnologías permita que la velocidad con que se genera tal información sea notoriamente mayor. En las ciencias genómicas se encuentran claros ejemplos de ello con la secuenciación de organismos o con la cuantificación masiva de los transcritos del organismo bajo condiciones específicas de crecimiento. ¿Qué conocimiento puede ser generado con base en esta información? ¿Podemos avanzar a una nueva etapa que

integre y relacione a las *partes* para entender el *todo*? Hoy en día se empieza a reconocer a las células como sistemas que representan redes complejas de interacción entre sus productos genéticos y que tienen como resultado las funciones fisiológicas observables. ¿Es posible construir modelos computacionales que representen el estado fisiológico de un organismo?

Un modelo matemático es una representación de un conjunto de fenómenos o sucesos reales. Ese modelo se ha aplicado exitosamente en ciencias exactas, como la física. Sin embargo, dada la complejidad y el aún reducido conocimiento de los procesos biológicos, el modelado de la respuesta celular ha tenido alcances más limitados. Entre los principales enfoques que se han utilizado con este fin, se encuentran los basados en el análisis de distintos tipos de redes que representan las interacciones dentro de los elementos del sistema.

Las redes de interacción entre proteínas dentro de las cascadas de señalización y redes metabólicas son, sin duda, temas centrales en la elaboración de modelos celulares, por lo que resulta necesario determinar cuáles, de las miles de proteínas dentro del organismo, pueden interactuar formando un complejo funcional. En esta dirección se ha desarrollado una variedad de metodologías experimentales y computacionales para dilucidar la red de interacciones proteínicas que ocurren en una célula. Experimentos masivos con sistemas de dos híbridos, de aislamiento sistemático de complejos multienzimáticos, así como de correlación de la expresión de mensajeros, han empezado a describir estas interacciones. La confiabilidad con que los distintos métodos, tanto teóricos como experimentales, predicen una interacción física real varía, pero en general la evaluación de las predicciones para cualquier método ha demostrado que los niveles de error aún son altos y su cobertura tiende a ser baja. Pese a lo anterior, la certeza de las predicciones aumenta considerablemente cuando más de un método independiente arroja los mismos resultados.

Para representar el metabolismo celular de una manera precisa sería necesario, entre otras cosas, conocer los parámetros cinéticos y la concentración de cada uno de los elementos que intervienen en las vías metabólicas. Aun para los organismos mejor caracterizados esa información es limitada y, en muchos casos, inexistente. Por ello, el éxito de los modelos cinéticos dentro de la ingeniería de vías metabólicas ha sido modesto. No obstante, en ausencia de información cinética se han elaborado modelos metabólicos con base en la distribución de flujos bajo el supuesto de un estado en equilibrio. A diferencia de los modelos cinéticos, los modelos estequiométricos no plantean encontrar el comportamiento preciso de la red metabólica, sino determinar un espacio de soluciones posibles, con base en las restricciones de estequiometría impuesta a la red. De este conjunto de soluciones se determina aquella para la cual es máxima alguna función específica.

Por otro lado, las redes de regulación de la expresión genética representan los diferentes elementos, por lo que los genes de una célula son transcritos en la cantidad y tiempo requeridos para contender con los estímulos externos o con base en un programa de desarrollo predeterminado. Se han utilizado distintos enfoques para realizar esa representación, desde los más sencillos, que consideran estados binarios, hasta los más complejos, que representan estados discretos. Una parte esencial en la construcción de este tipo de redes es la identificación de los efectores transcripcionales, en general proteínas reguladoras, y el de sus correspondientes blancos de reconocimiento en el genoma. En el caso de genomas pequeños, como el de algunos bacteriófagos con tan sólo algunos miles de pares de bases, esta tarea es sencilla y se han construido modelos de regulación muy completos. En genomas bacterianos y organismos eucariotes unicelulares, con tamaños promedios de algunos millones de pares de bases, la complejidad es notoriamente mayor; sin embargo, el hecho de que la mayoría de las señales de regulación se encuentren inmediatamente anteriores a los genes que regulan simplifica considerablemente su identificación.

Nuestra visión

A través del breve recorrido histórico que hemos presentado, se puede notar que cada nuevo descubrimiento o avance científico pone al ser humano en un contexto histórico tal que le permite resolver nuevas interrogantes. La capacidad de determinar la secuencia genómica de diversos organismos constituye un ejemplo de ello. En este sentido, el objetivo de nuestro grupo es el de entender el significado biológico de la información contenida en los genomas y de cómo dicha información se genera, evoluciona y expresa. Las preguntas fundamentales que competen a nuestro grupo de investigación ubicadas en el campo de la genómica computacional, han girado alrededor de la identificación y compresión del papel que juegan las señales que afectan a la regulación de la expresión genética, con miras a tener una visión integrativa que nos lleve a entender, no sólo el papel actual de estos elementos de manera individual, sino la participación global dentro de las redes de regulación de los sistemas biológicos.

En general, nuestra estrategia de análisis, como se ejemplifica en la **figura 1**, parte de una simple premisa: "Aquellas propiedades biológicas o señales observadas con frecuencias de aparición *estadísticamente significativas* pueden ser también *biológicamente significativas"*. Evidentemente, la anterior premisa no implica que todos los procesos tengan que presentarse en frecuencias estadísticamente altas para ser relevante en el contexto biológico. Dado lo anterior, nuestro grupo ha tratado de identificar características de regulación que se presentan de una manera común en el conjunto de los cientos de genomas actualmente secuenciados. En particular, nuestros estudios se han centrado en la identificación de señales de regulación dentro de alguno de los siguientes niveles: i) las contenidas en la secuencia primaria del ADN; ii) las que que dependen la curvatura intrínseca del ADN, y iii) las relacionadas con la estructura secundaria del ARN transcrito.

La regulación de la expresión genética de cualquier organismo se lleva a cabo en una primera instancia al inicio de la transcripción de los genes. Dicho evento involucra fundamentalmente proteínas que, al unirse al ADN, favorecen o inhiben la expresión de los genes. Las proteínas reguladoras antes mencionadas, reconocen comúnmente secuencias específicas en el ADN de tal forma que sólo aquellos genes que requieren ser activados o reprimidos son coordinadamente modulados. Estas secuencias de unión pueden ser determinadas mediante la comparación de las regiones de regulación de genes ortólogos en diversos organismos. En nuestro grupo hemos llevado a cabo la identificación de dichas secuencias primarias para diversas proteínas reguladoras mediante el análisis estadístico de la sobre-representación de ciertas "palabras" o motivos en las regiones de regulación.

Adicionalmente a la secuencia primaria del ADN, existe un componente estructural del ADN que es reconocido por las proteínas reguladoras. Con el objetivo de identificar el papel de esta propiedad en los genomas totalmente secuenciados, hemos conducido un estudio *in silico* que mostró que la curvatura estática del ADN es un elemento de regulación que puede ser compartido en diferentes grupos de genes evolutivamente relacionados y funcionalmente equivalentes (genes ortólogos) involucrados en procesos celulares fundamentales, como la expresión genética, división celular, biosíntesis de flagelo y motilidad. Con el fin de verificar los resultados obtenidos en nuestro análisis teórico, hemos llevado a cabo experimentos en el laboratorio para introducir mutaciones en sitios específicos y alterar la geometría curva del ADN presente en ciertas regiones reguladoras y evaluar su efecto en la expresión de los genes. Dichos experimentos demostraron que es factible predecir la forma que una molécula puede adoptar en el espacio, y que ciertas conformaciones del ADN (curvo en nuestro caso) tienen una implicación biológica.

Adicionalmente a la regulación transcripcional efectuada a nivel del ADN, existen procesos de regulación que se llevan a cabo en las moléculas de ARN. En este caso, el ARN puede adoptar estructuras en el espacio que modu-

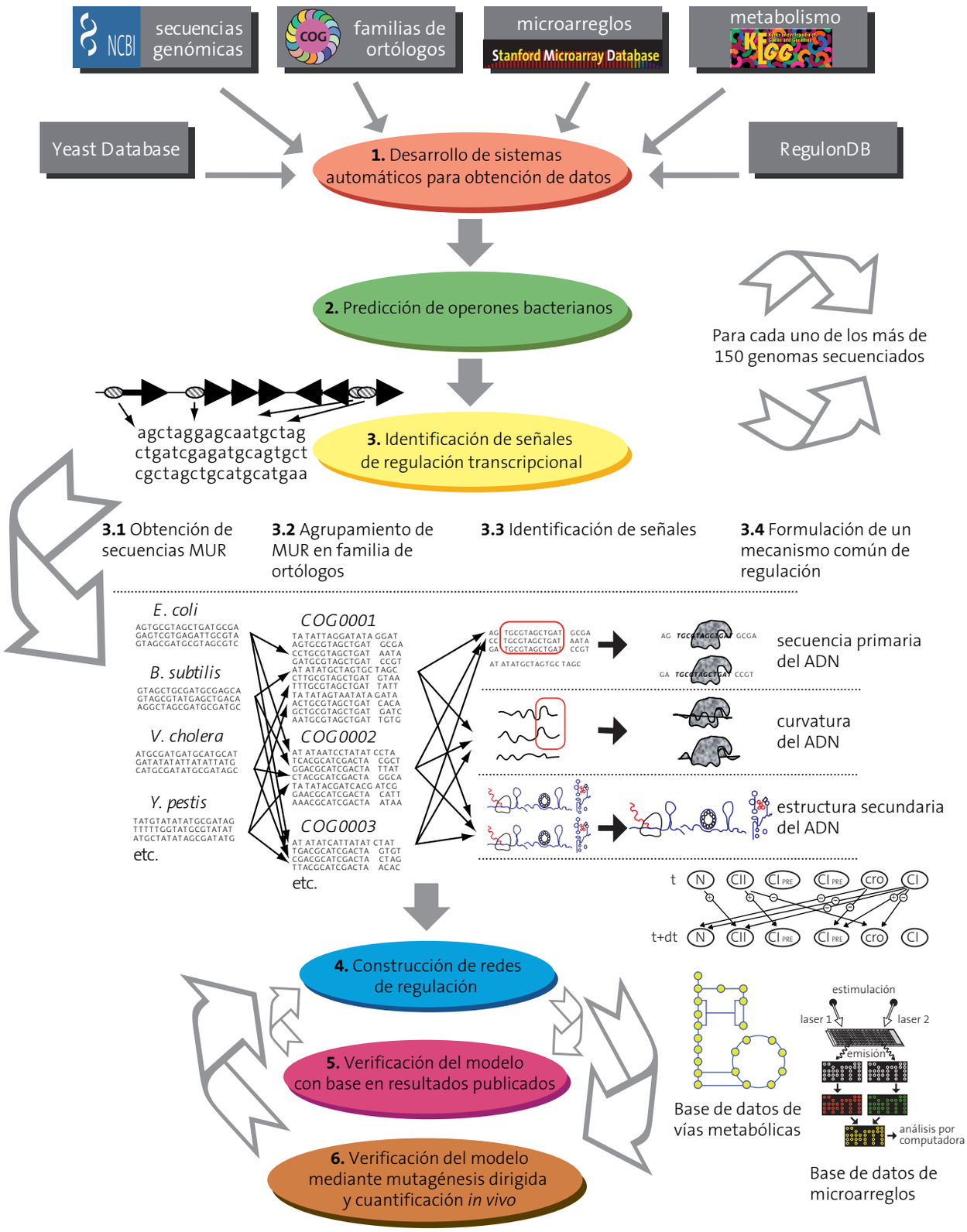


Figura 1.
Esquema general del procedimiento de análisis desarrollado para identificar señales de regulación mediante genómica comparativa y su integración para la construcción de modelos matemáticos.

lan los procesos de transcripción y traducción del mensaje. A estas estructuras del ARN se les ha llamado *atenuadores*. En nuestro grupo hemos realizado programas de cómputo que nos permiten identificar atenuadores en los diversos genomas bacterianos disponibles públicamente. Dicha reconocimiento es realizado primordialmente con base a la energía libre del conjunto de estructuras secundarias del ARN que pueden formarse en la región líder del ARN mensajero, y algunas de sus propiedades en términos de distancia y composición de la secuencia, entre otras. Nuestro análisis identificó la gran mayoría de los atenuadores reportados en la literatura, así como un gran número de nuevos atenuadores conservados en distintas familias de genes ortólogos.

Recientemente se ha descrito que ciertos tipos de ARN tienen la capacidad de formar estructuras tales que reconocen, con gran afinidad y gran especificidad, metabolitos y otras pequeñas moléculas. Dicho reconocimiento es a su vez utilizado para coordinar el encendido y apagado de los genes de acuerdo a condiciones metabólicas específicas, por lo que a estos elementos de ARN se les ha llamado *riboswitches*. A través de la genómica comparativa hemos podido desarrollar un método que permite identificar a cualquier tipo de riboswitch en las secuencias genómicas. Nuestro método ha demostrado ser exitoso en la identificación de todos y cada uno de los riboswitches previamente caracterizadas por otros grupos (riboswitches de tiamina, riboflavina y vitamina B12, T-box, etc.,), así como de nuevas secuencias altamente conservadas que pudieran corresponder a nuevos tipos de riboswitches.

Finalmente, una parte esencial de nuestro estudio consiste en la elaboración de modelos matemáticos que nos permitan integrar nuestros resultados de señales de regulación, así como otra información públicamente disponible de las redes metabólicas, además de la cuantificación masiva de transcrito (*transcriptoma*), concentración de proteínas intracelulares (*proteoma*), e interacciones proteína-proteína (*interactoma*). Modelos de redes de regulación que permitan diferenciar los distintos estados de expresión genética a partir de un conjunto de reglas epigenéticas para un conjunto de condiciones definidas, y cuya precisión será evaluada en función con la congruencia que tenga con los datos antes mencionados.

Comprendemos que las metas a alcanzar son difíciles, y que cada pequeño resultado nos lleva siempre a nuevas interrogantes, pero también confiamos en que cada nuevo descubrimiento nos permitirá integrar una nueva pieza de información al rompecabezas que a nuestros ojos son los sistemas biológicos. Creemos que para que dicha tarea sea lograda, el análisis genómico deberá de madurar significativamente, con procedimientos que posiblemente incluyan nuevos métodos estadísticos y el auxilio de la inteligencia artificial (y la propia).