

La secuenciación del ADN: consideraciones históricas y técnicas

DNA sequencing: historical and technical considerations

María Teresa Reguero Reza*

Editorial

Es innegable la importancia que tiene la generación de conocimiento alrededor de una molécula como el DNA en proyectos de investigación, desarrollo e innovación en áreas tan relevantes como la agricultura, la salud o el medio ambiente. Pero este gran avance científico ha sido el resultado de la conjunción de una serie de trabajos, que se inician con la identificación del ADN como una molécula importante, en la que se centra la transmisión de los determinantes genéticos de la herencia.

En el jardín de un monasterio, Gregor Mendel realizó sus experimentos de hibridación con plantas de arvejas (*Pisum sativum*), encontrando que las nuevas características que presentaban las plantas, eran el resultado de la mezcla de los caracteres iniciales que poseían las plantas. Hace 149 años Mendel dictó una conferencia acerca de la importancia de un “factor” de la herencia y ello podría considerarse como el punto de partida para lograr la identificación, aislamiento y caracterización de estos atributos y la forma como podrían ser fijados en una especie. Se le reconoce a Mendel ser el iniciador de las leyes fundamentales de la herencia, que impulsaron el desarrollo de la biología moderna.

Por primera vez, en 1871, Johann Meisher hace una descripción del ácido desoxirribonucleico (ADN) en el espermatozoide de la trucha, y en 1944 tres investigadores: Oswald Avery, Colin McLeod y Maclyn McCarty, demuestran que el ADN es la molécula en donde se encuentra la información genética.

En la década de 1940, Frederick Sanger, investigador del Departamento de Bioquímica de la Universidad de Cambridge, estaba interesado en el metabolismo de los aminoácidos y en la secuenciación de estos en una proteína de gran interés biológico: la insulina. En esa época no se conocía la estructura del ADN, por lo que su investigación se enfocaba en la elucidación estructural de las proteínas, compuestos que se creía conformaban el material genético.

La determinación de que las cantidades de adenina y timina y las de citosina y guanina son las mismas, es un gran aporte que hace Erwin Chargaff en 1950 y que, posteriormente, se constituyó en la “Regla de Chargaff”. Para complementar los aportes a la estructura del ADN, en 1952 Rosalind Franklin y Maurice Wilkins realizaron estudios de cristalografía de rayos X de esta molécula.

Como se ha mencionado, durante algún tiempo se consideró que el “factor” de la herencia de Mendel eran las proteínas. El 25 de abril de 1953, es un día importante para el conocimiento de la molécula que contiene la información genética que todo organismo. Este día se publica un artículo de tan solo dos páginas (737-738) en la revista *Nature* con el título de “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid” cuyos autores eran J.D. Watson y F.H.C. Crick. Fundamentalmente, Watson y Crick proponen la estructura de doble hélice del ADN, sustentada en el apareamiento específico que presentan las cuatro bases que la conforman: A-T y C-G. Esta publicación se convertiría en la piedra angular de la investigación genómica y con ello abren una senda, por la que la investigación y el desarrollo tecnológico han transitado en los últimos sesenta años.

En 1958 le fue otorgado a Frederick Sanger su primer Premio Nobel de Química “por su trabajo en la estructura de proteínas, particularmente de la insulina”, pues su trabajo de investigación llegó a la conclusión principal de que las dos cadenas polipeptídicas de la insulina tenía secuencias de aminoácidos precisos y, por extensión, que cada proteína tenía una secuencia única. Con este reconocimiento se pone de manifiesto la importancia de la secuenciación en el estudio de la relación entre la estructura y la función de las moléculas biológicas.

Otro hito relevante para el futuro de la investigación genómica es el que realizaron, entre 1957 y 1958, Matthew Meselson y Franklin Stahl al demostrar que el ADN se replica a través de un proceso semi-conservativo.

* Instituto de Biotecnología. Universidad Nacional de Colombia. Sede Bogotá. mtregueror@unal.edu.co

A principios de la década de 1970 Allan M. Maxam y Walter Gilbert adscritos al Departamento de Bioquímica y Biología Molecular de la Universidad de Harvard, investigan sobre la secuenciación del ADN y en febrero de 1977 publican, en *Proceedings of the National Academy of Sciences of the United States of America*, un artículo que titularon “A New method for sequencing ADN” en el que afirman que el ADN puede ser secuenciado por un procedimiento químico, mediante el cual es posible determinar la secuencia de nucleótidos del ADN. La propuesta se resume en la realización de cuatro reacciones diferentes de síntesis de ADN utilizando un dideoxinucleótido (ddNTP) distinto en cada tubo y nucleótidos normales, lo que permite la síntesis de diversos fragmentos que terminan con un ddNTP específico y posteriormente son separados por electroforesis, de acuerdo con su tamaño. Utilizando esta metodología es posible secuenciar ADN en término de horas.

Frederick Sanger, tomando en consideración lo reportado por Meselson y Stahl, desarrolla un método de secuenciación de pequeños fragmentos de ADN cuyo enfoque diferencial, frente a la propuesta de Maxam y Gilbert que era eminentemente química y tomaba al menos un día, fue la utilización de síntesis enzimática de la cadena complementaria del fragmento de ADN a secuenciar. Utilizando esta estrategia, Sanger y colaboradores, en 1978, realizan la secuenciación total del bacteriófago Φ -X174 y desde entonces es el método más utilizado para determinar secuencias de ADN. En 1980 Sanger y Gilbert fueron laureados con el Premio Nobel en Química, el segundo para Sanger, “por sus contribuciones en la determinación de la secuencia de las bases en los ácidos nucleicos”. Este mismo año le fue otorgado el Premio Nobel a Paul Naim Berg, profesor de la Escuela de Medicina de la Universidad de Washington y Profesor Emérito de la Universidad de Stanford, “por sus estudios sobre la bioquímica de los ácidos nucleicos en particular por la técnica del ADN-recombinante”. En 1981 se reporta la secuencia del ADN mitocondrial humano y Robert A. Weinberg (MIT), Michael Wigler (Cold Spring Harbor Laboratory) y Mariano Barbacid (NIH), líderes de tres grupos de investigación, informan por primera vez que los oncogenes son responsables de un cáncer humano.

En cuanto a los aspectos técnicos en los que se fundamenta la secuenciación, es relevante mencionar el gran aporte de Kary Mullis quien, en abril de 1990, publica un artículo en *Scientific American*, el que describe la reacción en cadena de la polimerasa (PCR, por sus siglas en inglés) que permite amplificar millones de veces fragmentos específicos de ADN.

Hasta este momento la secuenciación era un proceso manual, pero en 1982 Marvin Caruthers y Leroy Hood desarrollan el primer método automatizado para secuenciar ADN, el cual era capaz de secuenciar fragmentos de 5 a 75 pares de bases (pb) y, en 1986, Leroy Hood y Lloyd Smith diseñan el primer secuenciador automático que utiliza rayos láser que reconocen marcadores de fluorescencia en el ADN.

Analicemos lo acaecido en la evolución de los procesos de secuenciación del ADN, al pasar del método de Sanger, que era manual, al advenimiento de los secuenciadores automáticos diseñados por Leroy Hood y un grupo de investigadores del Instituto Tecnológico de California (CALTECH). Básicamente, la finalidad que perseguían era automatizar los procesos que se llevaban a cabo en el laboratorio y hacerlo en colaboración con el sector empresarial, para posteriormente comercializar sus innovaciones. En los proyectos de investigación del grupo de Hood en CALTECH se empleaba de manera rutinaria la secuenciación manual de proteínas y decidieron que la automatización sería una excelente oportunidad para comercializar el equipo. Con este propósito en mente buscaron financiación para el proyecto y crearon una empresa *start-up* que denominaron Applied Biosystems, la cual se encargaría de la comercialización de lo que desarrollaran. Dado que uno de los problemas para la automatización del equipo era la lectura de la secuencia, cuya interpretación la realizaba visualmente el investigador, idearon la inclusión de un computador y sustituyeron las bandas negras que se obtenían en las tiras, por bandas de un color diferente para cada uno de los nucleótidos detectados por láser, generando información digital que el computador era capaz de procesar.

Estas técnicas son básicas en la realización del Proyecto Genoma Humano (PGH). El primero de octubre de 1990 el Instituto Nacional de Salud (NIH) y el Departamento de Energía de los Estados Unidos (DOE), informan oficialmente el lanzamiento del Proyecto Genoma Humano (PGH) y nombran a James Watson director del Centro Nacional de Investigación del Genoma Humano. Se inicia la ejecución del proyecto en 1991 y al año siguiente nace “The Institute for Genomic Research” (TIGR). En 1994, se tiene el primer mapa genético y en 1996, el mapa físico (ESTs); se crea la a corporación privada Celera Genomics (1998) y para 1999 se obtiene la secuencia completa del primer cromosoma: el número 22. Posteriormente, en el año 2000, se dispone del primer borrador y el lunes 26 de junio de ese año, el doctor J. Craig Venter anunciaba, en una conferencia de prensa en la Casa Blanca y con la presencia del presidente de los Estados Unidos, Bill Clinton, y el primer ministro de Gran Bretaña, Tony Blair, que era un día histórico para la humanidad al poder leer, por primera vez, las letras químicas del código genético humano. Así mismo resaltaba que este resultado era el esfuerzo conjunto de múltiples investigadores en una alianza de instituciones públicas y privadas, que había iniciado cinco años antes, The Institute for Genomic Research in Rockville, Maryland.

Cabe señalar que, concomitantemente con el desarrollo del proyecto del genoma humano, en 1990 tres grupos de investigación diseñaron en forma simultánea, el método de electroforesis capilar que sería de gran utilidad en la secuenciación. El primer genoma totalmente secuenciado de una célula procarionta fue el de *Haemophilus influenzae* en 1995; un año después se reporta el genoma de una levadura: *Saccharomyces cerevisiae*. En 1998 se termina la secuenciación del geno-

ma de un animal *Caenorhabditis elegans* y en el siguiente año se realiza la secuenciación del primer cromosoma humano, el número 22. Al disponer de equipos de secuenciación automática cada año aparecen nuevos genomas, y es así como en el año 2000 se dispone del primer genoma de un vegetal: *Arabidopsis thaliana*. En 2003 se completó la secuenciación del genoma humano, que si se colocara en papel ocuparía alrededor de 100 volúmenes, cada uno con 1000 páginas escrito con los cuatro caracteres A, T, C y G. Por ello, para que la información obtenida, pudiese ser consultada en la red, hubo necesidad de desarrollar sistemas de cómputo capaces de almacenar y procesar los datos.

Merece la pena mencionar que el costo total del proyecto del genoma humano ascendió a US\$ 2700 millones, y en la actualidad el costo por genoma secuenciado oscila entre US\$ 5000 y US\$ 3000, lo que muestra la necesidad de contar con secuenciación de bajo costo y ello ha dado lugar a las distintas tecnologías de secuenciación de alto rendimiento. Para lograr disminuir más los costos de la secuenciación de un genoma humano, dos empresas: Life Technology® e Illumina® se encuentran diseñando secuenciadores cuya finalidad es poder secuenciar por US\$ 1000 o menos, fundamentado en el avance de la nanotecnología y particularmente con la utilización de nanoporos para una sola molécula. Esta aplicación de los nanoporos fue descrita y patentada ante la Oficina de Patentes de los Estados Unidos, por George Church y colaboradores, cuya solicitud data de 1995 y que fue otorgada en 1998.

Es innegable el gran potencial que tiene el conocimiento de los factores genéticos, asociados a enfermedades o desordenes genéticos, para su detección y posible tratamiento. En este contexto la secuenciación de nueva generación (NGS – Next Generation Sequence) de genomas, es una tecnología ideal para ser utilizada, debido a su rapidez y economía pero adicionalmente porque permite obtener datos en forma rápida, confiable y de buena calidad para el entendimiento de sistemas biológicos a nivel de genoma, transcriptoma, metaboloma o epigenoma. La NGS de genomas implica la confluencia de equipos automatizados, en los que los avances experimentales en química, ingeniería, biología molecular, nanotecnología se entrelazan con la computación de alto rendimiento, para incrementar la velocidad con la cual se obtienen los datos, su almacenamiento, así como su análisis y procesamiento.

A partir de 2011 se dispuso de varias plataformas para realizar la secuenciación, un ejercicio de comparación del desempeño de tres plataformas de secuenciación de nueva generación: Ion Torrent®, Pacific Biosciences®, Illumina MiSeq® fue realizado por Michael A. Quail y colaboradores del Instituto Sanger (2012) para determinar la secuencia de cuatro genomas bacterianos en el que se evalúan parámetros tales como cantidad de muestra (3 mg - 50 ng); tiempo de procesamiento (30 min. - 11 días); pares de bases por corrida (52.5 kb - 600 Gb). En 2012 Illumina® lanzó una nueva versión de su secuenciador HiSeq® 2500, que es un sistema de nueva generación de secuenciadores, cuyas características principales son la rapidez, pues es capaz de secuenciar un genoma en tan solo 24 horas y el número de datos, ya que puede generar 120 gigabases de datos en 27 horas o 600 Gb por corrida.

Los datos de secuenciación obtenidos con esta tecnología disminuyen tiempo y costo y, al integrarlos con los datos biológicos, generan información genómica relevante para la clínica, ya que pueden ser utilizados para la predicción de sets de datos que se utilizan en el diagnóstico y tratamiento de enfermedades, para establecer su incidencia en una determinada patología o como blanco terapéutico. Esta estrategia innovadora permite su utilización en áreas como la agricultura, nutrición y salud. Particularmente, se pueden evidenciar biomarcadores y diseñar nuevas entidades moleculares en oncología o en la caracterización de enfermedades genéticas orientadas hacia la consolidación de la medicina personalizada o medicina de precisión, entendida como la habilidad de diseñar tratamientos terapéuticos acordes con las distintas características de cada paciente, incorporando la información genómica para hacer más asertiva la toma de decisiones.

Es también importante reflexionar sobre la necesidad de contar con talento humano experto, no solo en secuenciación, sino en bioinformática y en aspectos clínicos, para implantar y apoyar su inminente utilización en forma rutinaria, como parte de las determinaciones que ofrecerán los laboratorios clínicos y de las cuales ya se pueden mencionar algunos ejemplos: cáncer de seno, de pulmón, de hígado y colorrectal.

Para apoyar este tipo de aplicaciones de la secuenciación, el 4 marzo de 2014 se realizó en Londres una reunión de los integrantes de la Global Alliance for Genomics and Health, en la que también participaron entidades públicas y privadas de investigación biomédica, entre las que se pueden citar Wellcome Trust, los Institutos de Salud de los Estados Unidos, las universidades de Berkeley, Oxford, Stanford Texas, John Hopkins, la Sociedad Americana de Oncología Clínica y Genética Humana, el Centro Nacional de Análisis Genómico de España, industrias farmacéuticas como Sanofi y Merck y empresas de computación como IBM, de secuenciadores Illumina® y Google, entre otras. El propósito de esta magna reunión fue, como en el caso del PGH, aunar esfuerzos para recopilar la información sobre datos clínicos y genéticos con el propósito de avanzar hacia la medicina de precisión. Ese mismo día Craig Venter anunció la creación de la empresa Human Longevity Inc. (HLI®) y de un nuevo proyecto cuyo propósito es recopilar y analizar el genoma y los parámetros bioquímicos de miles de personas, que voluntariamente deseen participar en el proyecto. Se espera poder

secuenciar 40 000 genomas al año, con la finalidad de desarrollar kits de diagnóstico, predecir el riesgo de padecer una determinada patología y comercializar los productos que este proyecto genere para el diseño de una terapia personalizada. Otra vertiente de la secuenciación masiva de genomas es la que abanderada California Life Company (Calico®), empresa cuyos desarrollos se orientan a mitigar y curar el envejecimiento y las enfermedades asociadas.

Finalmente y en honor a la molécula del ADN por su importancia para las ciencias de la vida, el día 25 de abril se ha designado como el “día del ADN” coincidiendo con el día que apareció la publicación de Watson y Crick y tomando en consideración que el conocimiento sobre el ADN ha sido factor clave para el desarrollo de la industria biotecnológica orientada a sectores como salud, agropecuario, ambiental y bioenergético, entre otros.

Referencias bibliográficas

- Church G.M., Deamer D.W., Branton D., Baldarelli R., Kasianowicz J. 1998. Characterization of individual polymer molecules based on monomer-interface interaction. *US PATENT 5.795.782*. Solicitada el 17 de marzo de 1995; otorgada el 18 de agosto de 1998.
- De Necochea Campion R., Canul Tec J.C. 2004. *Métodos Físicoquímicos en Biotecnología: Secuenciación de Ácidos Nucleicos*. Universidad Nacional Autónoma de México. Instituto de Biotecnología. Consultado en: http://pt7mdv.ceingebi.unam.mx/computo/pdfs/met/secuenciacion_acidos_nucleicos.pdf
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature*. 409: 860-921. doi:10.1038/35057062
- Maxam A.M., Gilbert W. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*. 74(2): 560-564.
- Mayr E. 1982. The growth of Biological Thought: Diversity, evolution, and inheritance. Harvard University Press. Pp. 681 - 726.
- Mullis K.B. 1990. The unusual origin of the polymerase chain reaction. *Scientific American*. 262(4): 56-61.
- Olson M. 1993. The human genome project. *Proceedings of the National Academy of Sciences*. 90(10): 4338-4344.
- Quail M.A., Smith M., Coupland P., Otto T.D., Harris S.R., Connor T.R., Bertoni A., Swerdlow H.P., Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*. 13(1): 341-353. doi:10.1186/1471-2164-13-341.
- Sanger F., Nicklen S., Coulson A.R. 1977. DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Sciences*. 74(12): 5463-5467.
- Schlenk F. 1988. Early nucleic acid chemistry. *Trends in Biochemical Sciences*. 13(2): 67-69.
- Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., ... , & Zhu X. 2001. The sequence of the human genome. *Science*. 291(5507):1304-1351.
- Watson J.D., Crick F.H.C. 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*. 171: 737 - 738. doi:10.1038/171737a0