Breve historia de la bioinformática

María Liliana Franco[1], Juan Fernando Cediel, MD[2], César Payán, MD[3]

RESUMEN

La bioinformática es el resultado de la unión indisoluble entre las tecnologías informáticas y las ciencias biológicas. Fue concebida en principio para resolver interrogantes como los siguientes: ¿cómo almacenar y organizar secuencias de ADN? ¿Cómo hallar intrones y exones en secuencias de ADN genómico? ¿Cuáles son las condiciones necesarias para la transcripción de un determinado gen? ¿Cómo conocer más acerca de la estructura de una proteína? ¿Cómo comparar secuencias de proteínas o predecir su estructura? En esta era postgenómica, la adquisición de nuevas y mejores herramientas computacionales ha hecho posible que la bioinformática se convierta en pieza clave para aplicaciones como filtro genético, diagnóstico molecular, hallazgo de nuevos fármacos y mejoramiento genético de cultivos.

Palabras clave: *Bioinformática; Biología computacional; Historia.*

Brief history of bioinformatics

SUMMARY

Bioinformatics appears as the result of the indissoluble marriage between informatics technology and life sciences. Although it was conceived in principle to resolve questions such as the following: How to store and organize DNA sequences? How to find introns and exons in genomic DNA sequences? What conditions are required for the transcription of a specific gene? How learn more about the structure of a protein? How does it compare sequences of proteins or predict their structures?, currently the acquisition of new and improved computational tools has made it possible that bioinformatics will be key piece in applications such as genetic screening, molecular diagnosis, drug discovery and crop genetic improvement.

Keywords: *Bioinformatics; Computational biology; History.*

Cuando en 1953 Watson y Crick propusieron el modelo de la doble hélice para explicar la estructura del ADN, no vislumbraron el formidable volumen de información que en forma exponencial se generaría a partir de ese momento[1] y que daría origen a problemas algorítmicos susceptibles de un manejo altamente cuidadoso y organizado. En forma venturosa, en las décadas siguientes hicieron su aparición herramientas computacionales que hicieron posible el análisis y la resolución de interrogantes que ya estaban planteados en la propia estructura del ADN, en la información genética codificante de las proteínas[2], en las propiedades estructurales de éstas y en los factores que las regulan[3,4], así como en los sucesos asociados con la regulación génica, las bases moleculares del desarrollo embrionario y la evolución de las vías metabólicas bioquímicas[5-7].

En forma contraria a lo que podría suponerse, las herramientas computacionales comenzaron a aplicarse en la biología molecular mucho antes del comienzo de la era de Internet o de los proyectos de secuenciación del genoma. Hacia 1960, la creciente cantidad de datos referentes a la química de las proteínas llevó a los científicos a combinar las estrategias de la biología mole-

1. Estudiante X semestre de Medicina, Universidad del Rosario, Bogotá DC, Colombia. e-mail: marialilianaf85@hotmail.com
2. Profesor Auxiliar, Escuela Ciencias de la Salud, Departamento de Ciencias Básicas, Unidad de Morfología, Universidad del Rosario, Bogotá DC, Colombia. e-mail: juan.cedielb@urosario.edu.co
3. Profesor Asistente, Escuela Ciencias de la Salud, Departamento de Ciencias Básicas, Unidad de Biología, Universidad del Rosario, Bogotá DC, Colombia. e-mail: cesar.payan33@urosario.edu.co
Recibido para publicación enero 18, 2008 Aceptado para publicación enero 31, 2008

cular, las matemáticas y los computadores, para enfrentar con éxito el desafío que ello representaba. Y en este punto aparecen la bioinformática y la biología computacional como disciplinas íntimamente relacionadas, donde la primera, de acuerdo con la definición de la NCBI (National Center for Biotechnology Information de los Estados Unidos de América), busca y utiliza patrones y estructura inherente en datos biológicos como secuencias génicas, así como el desarrollo de nuevas metodologías para acceso y búsquedas en bases de datos[8], mientras que la segunda se refiere a la simulación física y matemática de los procesos biológicos[9].

Brown en el año 2000, definió la bioinformática como «el uso de computadores para la adquisición, manejo y análisis de la información biológica», de modo que la contextualiza «en la intersección de la biología molecular, la biología computacional, la medicina clínica, las bases de datos informáticas, el Internet y el análisis de secuencia»[10].

Según el Weizmann Institute of Science de Israel, «aunque el término 'bioinformática' no puede ser bien definido, se podría afirmar que es el campo de la ciencia que se ocupa del manejo computacional de todas las clases de información biológica, bien sea de genes o sus productos, de organismos o aun de ecosistemas»[11].

La bioinformática es pues una ciencia de naturaleza interdisciplinaria, cuya historia se partió en dos después que por vez primera se secuenció en forma completa una proteína, la insulina, por parte de Frederick Sanger y sus colegas en la Universidad de Cambridge, durante la década comprendida entre 1945 y 1955[12,13]. Sanger y su equipo, mediante un laborioso proceso analítico, separaron e identificaron los fragmentos de la degradación de la proteína y determinaron el orden de aparición de los aminoácidos, algo que nadie hasta ese momento había sido capaz de hacer.

Gracias al hallazgo de que cada proteína posee una estructura primaria única, Sanger obtuvo el Premio Nobel de química en 1958. Con posterioridad se desarrollaron otros métodos de secuenciación menos dispendiosos y más eficientes que el de Sanger, como la reacción de degradación de Edman, las columnas de intercambio iónico y la electroforesis, que contribuyeron a la automatización de la secuenciación y al desarrollo de librerías de aminoácidos[14,15]. Sin embargo, el logro alcanzado por Sanger fue el factor determinante en el rumbo que tomaría la bioinformática, pues hizo evidente la necesidad de interpretar la información contenida en las secuencias de ADN, ARN y proteínas. Por este motivo, se ha propuesto la existencia de dos eras consecutivas en la historia de la bioinformática: era pre-secuenciación y era post-secuenciación[16,17]. Pero la emergencia de la nueva ciencia no hubiera sido posible sin el concurso de los computadores digitales de alta velocidad. Inventados en el marco de programas de investigación para diseñar armamento bélico durante la segunda guerra mundial, los computadores sólo estuvieron al alcance de los investigadores a comienzos de la década de 1970, aunque con una disponibilidad muy limitada, 15% del total de centros de investigación y universidades de los Estados Unidos de América[14].

Dos hechos pertinentes fomentaron el desarrollo de la informática académica en la investigación biológica: por una parte el advenimiento de FORTRAN (del inglés *formula translation*), lenguaje de programación de alto nivel, de relativo fácil aprendizaje[15], y por otra los esfuerzos que efectuaron en tal sentido las agencias gubernamentales y la industria de los computadores de esa nación[18].

La difusión de las nuevas técnicas para secuenciar el ADN y las proteínas, así como el volumen cada vez mayor de secuencias almacenadas en los bancos de datos, hicieron necesaria la creación de algoritmos a fin de catalogar y comparar secuencias, en los que se reconoce como pionera a Margaret Oakley Dayhoff (1925-1983), connotada investigadora del Centro Médico de la Universidad de Georgetown. La doctora Dayhoff desarrolló métodos computacionales que le permitieron comparar secuencias proteicas y a partir de los alineamientos entre ellas investigar las relaciones y por tanto la historia evolutiva entre los diferentes reinos, phyla y taxa biológicos. Su monumental trabajo, que recopilaba todas las secuencias proteicas entonces conocidas, se publicó en 1965 en un pequeño libro titulado «Atlas de secuencia y estructura de proteínas»[19,20].

La primera edición del «*Atlas*» contenía las secuencias de 65 proteínas. Las siguientes ediciones se citan más de 4,500 veces y constituyen una fuente invaluable de referencia para científicos del mundo entero.

En 1980, la doctora Dayhoff creó la primera base de datos computadorizada de la que se tiene noticia, con secuencias de ácidos nucleicos y de proteínas, en un

computador casero al que los usuarios externos podían conectarse por vía telefónica. Para 1983 la Protein Sequence Database (PSD) era la base de datos más grande del mundo, con más de 2'000,000 de nucleótidos secuenciados, con sus respectivas referencias y anotaciones[19]. Sin embargo, este avance no hubiera sido posible sin la llegada de Internet. La red proveyó las facilidades de acceso para los usuarios así como también para el desarrollo del software necesario en el manejo y el análisis de inmensurables cantidades de datos[17].

Años después de la muerte de la doctora Dayhoff, su sueño de poner en funcionamiento un sistema en línea (online) consistente en programas y bases de datos accesibles a toda la comunidad científica mundial, comenzó a hacerse realidad. Mediante este sistema, conocido como Protein Identification Resource (PIR)[21], cualquiera podía identificar proteínas a partir de los datos de composición de aminoácidos o de secuencias, como también efectuar predicciones con base en éstas, o sencillamente buscar información[19]. A lo largo de más de 40 años de existencia, PIR provee acceso a muchas bases de datos de proteínas entre las que estaba incluida PSD. A partir del año 2002, PIR-PSD se asoció con EBI (European Bioinformatics Institute) y SIB (Swiss Institute of Bioinformatics), para dar origen a una única base de datos de secuencia y función de proteínas, conocida en la actualidad como UniProt[22,23].

A finales de la década de 1980 y comienzos de la de 1990, el trabajo de Margaret Oakley Dayhoff impulsó la generación de bases de datos primarias como GenBank, FASTA y BLAST (Basic Local Alignment Tool). Mientras GenBank almacenaba y catalogaba las secuencias de ADN y de proteínas, BLAST permitía comparar con mayor rapidez que su predecesor FASTA las secuencias de interés contra cada una de las secuencias contenidas dentro de la enorme base de datos[24]. Estuvo pues la bioinformática caracterizada en la década de 1990 por la utilización de bases de datos primarias que contenían información experimental en gran escala en las áreas de genómica y proteómica, lo que permitió comprender las funciones de los genes y de las proteínas.

En la actualidad, existen bases secundarias, llamadas también bases de conocimiento porque contienen el conocimiento biológico acumulado necesario para comprender el funcionamiento y la utilidad en todos los niveles de organización de un ser vivo (molecular, celular, organismo). Así por ejemplo, estas bases incluyen todas las familias de proteínas con sus dominios funcionales y sus estructuras tridimensionales, así como también las diferentes vías de señalización[24].

Para el futuro, se espera disponer de una representación computacional completa de la célula y del organismo con el fin de entender los principios que determinan el elevado nivel de complejidad de los sistemas biológicos[17].

REFERENCIAS

1. GenBank. (fecha de acceso diciembre 3 de 2007). Disponible en**:** http://www.ncbi.nlm.nih.gov/GenBank/index.html
2. Gamow G, Rich A, Ycas M. The problem of information transfer from nucleic acids to proteins. *Adv Biol Med Phys.* 1956; *4:* 23-68.
3. Pauling L, Corey RB, Branson HR. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA.* 1951; *37:* 205-11.
4. Szent-Györgyi AG, Cohen C. Role of proline in polypeptide chain configuration of proteins. *Science.* 1957; *126:* 697.
5. Britten RJ. Davidson EH. Gene regulation for higher cells: a theory. *Science.* 1969; *165:* 347-57
6. Turing AM. The chemical basis for morphogenesis. *Phil Trans R Soc London B.* 1952; *237:* 37-72.
7. Horowitz NH. On the evolution of biochemical syntheses. *Proc Natl Acad Sci USA.* 1945; *31:* 153-7.
8. NCBI Bioinformatics. (fecha de acceso diciembre 3 de 2007). Disponible en: http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html
9. Yu U, Lee SH, Kim YJ, Kim S. Bioinformatics in the post-genome era. *J Biochem Mol Biol.* 2004; *37:* 75-82.
10. Brown SM. Get your bioinformatics on the Web! *Biotechniques.* 2000*; 28:* 244-6.
11. Bioinformatics & Biological Computing. (fecha de acceso diciembre 3 de 2007). Disponible en: http://bip.weizmann.ac.il/
12. Sanger F, Thompson EO. The aminoacid sequence in the glycyl chain of insulin. *Biochem J.* 1952; *52:* iii.
13. Sanger F. Chemistry of insulin. *Science.* 1959; *129:* 1340-4.
14. Hagen JB. The origins of bioinformatics. *Nature.* 2000; *1:* 231-6.
15. Piast M, Kustrzeba-Wójcicka I, Matusiewicz M, Krzystek-Korpacka M, Banas T. Bioinformatics: From arduous beginnings to molecular databases. *Adv Clin Exp Med.* 2007; *16:* 85-93.
16. Roberts RJ. The early days of bioinformatics publishing. *Bioinformatics.* 2000; *16:* 2-4.
17. Kanehisa M, Bork P. Bioinformatics in the post-sequence era. *Nat Genet.* 2003; *33:* 305-10.
18. Ledley RS. Digital electronic computers in biomedical sciences. *Science.* 1959; *130:* 1225-34.
19. Dr. Margaret Oakley Dayhoff 1925-1983 Biography. (fecha de acceso: enero 6 de 2008). Disponible en: http://www.

dayhoff.cc/
20. Dayhoff M. *Atlas of protein sequence and structure.* Vol. 4 Silver Spring: National Biomedical Research Foundation; 1969.
21. PIR Protein Information Resource. (fecha de acceso diciembre 3 de 2007). Disponible en: http://pir.georgetown.edu.home.shtml
22. UniProt/Swiss-Prot. (fecha de acceso diciembre 3 de 2007). Disponible en: http://www.ebi.ac.uk/swissprot
23. Wu CH , Yeh LSL, Huang H, Arminski L, Castro-Alvear J, Chen Y, *et al*. The protein information resource. *Nucl Acids Res.* 2003; *31:* 345-7.
24. Martínez-Barreneche J. La bioinformática como herramienta para la investigación en salud humana. *Salud Publica Mex.* 2007; *49:* 64-6.