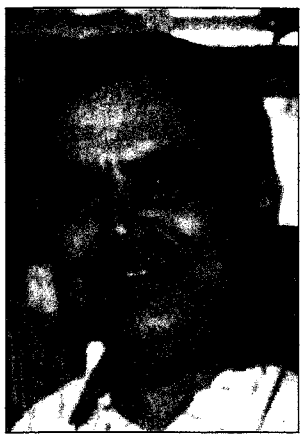
El ADN: de la Química a la Bioinformática



Juan A. Subirana

Juan A. Subirana
Departamento de Ingeniería Química
Escuela Técnica Superior de
Ingeniería Industrial de Barcelona

Xavier de la Cruz
Unidad de Modelado Molecular y Bioinformática
Departamento de Bioquímica
y Biología Molecular. Facultad de Química
Universidad de Barcelona

Xavier de la Cruz

En los últimos años el estudio del ADN ha conducido a resultados espectaculares, amplificados por los medios de comunicación. Ello es debido a que las técnicas desarrolladas en las últimas décadas han permitido secuenciar genomas completos, tanto de virus y bacterias, como de organismos superiores. Un ejemplo reciente es la secuenciación del genoma humano que pondrá a nuestro alcance unos tres mil millones de letras, cuyo orden determina la estructura de todas las proteínas e indirectamente de las demás moléculas que forman el cuerpo humano. De momento sólo se conoce la secuencia completa de un cromosoma (1), falta ordenar la mayor parte de la información disponible. Se espera que este trabajo se concluya en unos pocos años. A pesar de su indudable interés manejar, comprender y aplicar toda esta información requerirá un esfuerzo considerable, con resultados no inmediatos. Para ello ha surgido una nueva rama de la ciencia, la bioinformática, complemento esencial de las demás técnicas de caracterización de las macromoléculas biológicas.

Esta visión de la situación actual parece reducir el ADN a una serie de letras, un simple vehículo "lingüístico" de la información biológica. Tanto es así que en unas recientes declaraciones al "El País" (2) el premio Nobel Max Perutz manifestaba: "Lo único que hace el ADN es determinar el orden de los aminoácidos en las proteínas. Los genes no saben hacer otra cosa, son realmente muy torpes".

Sin embargo el ADN es una macromolécula compleja cuya función en la célula no se limita a ser únicamente una serie de letras, aunque éstas tengan una importancia esencial para la vida. En particular, la enorme complejidad estructural del ADN le permite realizar una amplia gama de funciones, esenciales para el buen funcionamiento del ser vivo. Por ello nos ha parecido oportuno dar aquí una visión general de estas dos aproximaciones contrapuestas y a la vez complementarias al ADN: la de la química y la de la bioinformática.

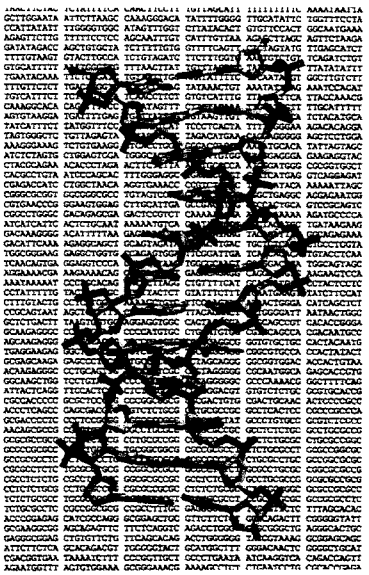


Figura 1. Estructura de la doble hélice del ADN superpuesta sobre la secuencia de un fragmento del cromosoma 21.

LA VARIABILIDAD DEL ADN

En la figura 1 se comparan las diferentes maneras de representar el ADN. Para el biólogo lo más importante es la secuencia de bases, la serie de letras que codifica la síntesis de proteínas. La naturaleza química de estas letras no tiene ningún interés aparente. Sin embargo, desde un punto de vista químico cada base tiene distintas posibilidades de interacción, que permiten a las proteínas reconocer secuencias distintas de ADN. Cada par de bases tiene una variabilidad local de conformación, determinada por el valor que adquieren catorce ángulos de torsión (ver encarte). En la forma B del ADN, que predomina en la célula, estos ángulos varían poco, aunque aparecen diferencias significativas en distintas zonas

del ADN. Pero existen zonas individualizadas del ADN en las que las distintas posibilidades conformacionales se manifiestan claramente. Es en estas zonas en las que el ADN manifiesta su versatilidad, así como las múltiples funciones adicionales que puede desempeñar y que en realidad no conocemos plenamente, de hecho son el objeto de investigaciones detalladas en numerosos laboratorios. Podríamos hablar de formas estructurales diversas, como la forma Z que gira a la izquierda, asociaciones entre distintas cadenas, plegamientos, etc. No parece apropiado hacer aquí una revisión completa, per sí vamos a dar algunos ejemplos de la versatilidad estructural de esta macromolécula.

EL ADN: UN POLIELECTROLITO

Un factor adicional a tener en cuenta es que el ADN es un polímero cargado eléctricamente, en el que cada grupo fosfato tiene una carga negativa asociada. Las cargas eléctricas ejercen su acción a larga distancia, por lo que su efecto neto no está claramente localizado. Los contraiones se disponen en forma más o menos desordenada en torno al ADN, neutralizando globalmente su carga. Sorprendentemente en las representaciones que se encuentran en prácticamente todos los libros de Bioquímica se ignora la existencia de estos contraiones, aunque son compañeros inseparables del ADN e influyen decisivamente en su comportamiento.

Recientemente, gracias a la utilización de técnicas de difracción en sincrotrón se han podido localizar los contraiones en algunos casos, tal como se muestra en la figura 2 para una sal de magnesio del ADN (3). Los cationes se disponen alrededor de la molécula en posiciones determinadas tanto por la disposición de los fosfatos en cada molécula como por la organización de las distintas moléculas de ADN en el cristal. En cualquier caso está claro que los fosfatos individuales no determinan la posición de los iones, viene determinada tanto por el ADN como por su entorno. De

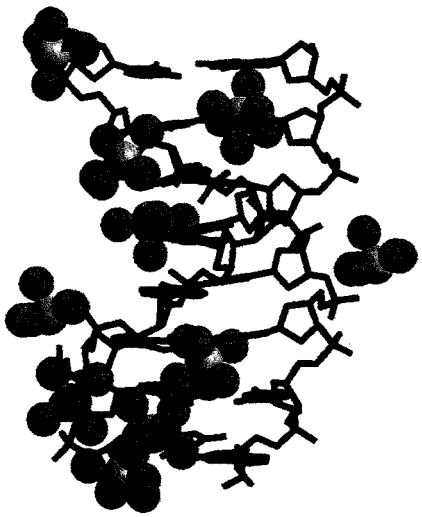


Figura 2. Fragmento de ADN rodeado de iones Mg^{2+} (rosa) en un cristal de oligonucleótido (3). Las aguas de hidratación de los iones se representan en color rojo.

hecho cuando se estudia el ADN a temperatura ambiente raramente puede precisarse la posición de sus contraiones, pues éstos presentan una movilidad apreciable y no tienen una posición fija.

En la célula el ADN está neutralizado no sólo por iones inorgánicos, sino, también por cadenas laterales de algunas proteínas cargadas positivamente. Algunas proteínas producen una neutralización global del ADN (histonas o protaminas, p. ej.), mientras que en muchos casos la interacción es más compleja e implica además de los residuos básicos otros residuos de la proteína capaces de reconocer una secuencia de bases específica. Un ejemplo son las proteínas que se unen a la "TATA box" que vamos a describir a continuación.

LA "TATA BOX"

Para utilizar la información codificada en el ADN es preciso que tenga lugar un proceso de transcripción preciso en que se lea exactamente la secuencia de las bases que interesan para sintetizar las moléculas de ácido ribonucleico (ARN). La célula utiliza estas nuevas moléculas de ARN como vehículo para obtener las proteínas. La síntesis del ARN es llevada a cabo por unos enzimas denominados polimerasas de ARN. ¿Cómo reconocen estos enzimas el lugar dónde han de empezar a actuar?

En los organismos superiores (eucariotas), es decir, que tienen un núcleo celular, la iniciación de la transcripción requiere varias proteínas. Entre ellas una de las más importantes es la que se une a la "TATA box". La "TATA box" es una secuencia de bases del tipo TATAXAX, donde X es A ó T. Esta secuencia se encuentra siempre unas 25 bases antes de las zonas que codifican para el ARN e indirectamente para las proteínas. Al unirse al ADN produce una distorsión acusada de la molécula, como puede verse en la Figura 3. El ADN sufre un cambio de dirección de 90° aproximadamente como resultado de la distorsión inducida por la proteína (4,5).

CONFORMACION DEL ADN

El ADN está formado por dos cadenas de fosfodiésteres, en las que grupos fosfato unen sendos residuos de desoxiribosa (pentágonos en la Fig. 1) al nivel de los oxígenos 3 y 5 del anillo. Así cada residuo tiene seis ángulos conformacionales (α, β, γ, δ, ε y ζ) cuyo valor define la estructura que adopta la cadena fosfodiester. Cada azúcar lleva unida una base nitrogenada en el átomo C1 mediante el denominado enlace glicosídico que define otro ángulo de torsión (χ). En el ADN hay cuatro bases posible, dos purinas (guanina, G y adenina, A) y dos pirimidinas (timina, T, y citosina, C) cuyo orden determina la información genética. Estas bases se aparean entre las dos cadenas, siempre A con T y C con G. Los pares de bases aparecen de perfil como lineas horizontales sobre el eje de la hélice de la Fig. 1. En la forma normal, denominada forma B, son como los peldaños de una escalera de caracol.

Figura 3. Distorsión inducida en el ADN (color verde) al interactuar con una proteína TBP (Tatabox Binding Protein, en color granate).

Esta distorsión local es reconocida por otras proteínas, denominadas factores de iniciación, y por la ARN polimerasa, que puede así iniciar el proceso de transcripción en la señal que se encuentra a unas 25 bases de distancia a lo largo del ADN.

Estos cambios estructurales, representados en la Figura 3, son un ejemplo claro de la interacción mutua entre el ADN y las proteínas, mostrando como los cambios conformacionales del ADN permiten regular el funcionamiento de la célula.

LOS TELOMEROS

En el extremo de los cromosomas aparecen unas estructuras peculiares denominadas telómeros (6). Estas estructuras son fundamentales para el funcionamiento e integridad de los cromosomas, pues impiden que se fusionen entre si para generar cromosomas de doble tamaño. Por otra parte estas zonas terminales se acortan en cada división celular y constituyen un reloj interno de la célula: después de un cierto número de divisiones la célula muere al haber perdido sus telómeros. La fisiología de los telómeros está pues relacionada con la longevidad de la célula. De hecho una de los factores que contribuyen a transformar células normales en cancerosas parece residir precisamente en la activación de un enzima, la telomerasa, que impide el acortamiento progresivo de los cromosomas durante la división celular. Estas consideraciones han conducido a una gran actividad en la investigación de la estructura y función de las zonas teloméricas del cromosoma, que son un objetivo posible para el desarrollo de fármacos que interfieren en su actividad (7).

Las secuencias de ADN que forman los telómeros son muy simples (8) y se repiten muchas veces, como mínimo veinte veces, pero pudiendo llegar a varios miles de repeticiones, tal es el caso en el ratón. Una característica general de estas secuencias es la presencia de varias guaninas seguidas. Algunos ejemplos son los siguientes:

-AGGGTT- -TCCCAA-	(mamíferos)
-AGGGTTT- -TCCCAAA-	(plantas)
-GGGGTT- -CCCCAA-	(algunos protozoos)

etc.

Otra característica importante de las secuencias teloméricas es que la hebra del ADN que contiene las guaninas se extiende más allá del extremo del cromosoma, formando una extensión de unas 15-20 bases que no están apareadas con una hebra complementaria: en lugar de formar una hélice doble típica se asocia con proteínas, probablemente en forma de hélice cuádruple (9) tal como se muestra en la Figura 4. Las guaninas son capaces de aparearse entre sí en forma de cuadrupletes estabilizados por puentes de hidrógeno: cada guanina está directamente asociada con otras dos. Las cuatro guaninas definen una estructura prácticamente plana, apilándose tres planos (Figura 4). Las bases intermedias (TTA en los mamíferos) definen las curvas que permiten el apareamiento de las guaninas presentes en la hebra terminal del cromosoma.

Como decíamos al empezar este apartado, los telómeros definen los puntos terminales de los cromosomas. En cada división celular esta estructura peculiar ha de generarse nuevamente en cada uno de los cromosomas hijos. En este proceso la larga secuencia de bases repetidas se va acortando progresivamente. Para mantener su longitud es precisa la activación de un enzima, la telomerasa, que normalmente se encuentra reprimido. La estructura en hélice cuádruple que hemos mostrado en la Figura 4 juega un papel importante en este proceso. Esta estructura es un ejemplo más de la versatilidad conformacional del ADN, esencial para llevar a cabo sus múltiples funciones biológicas.

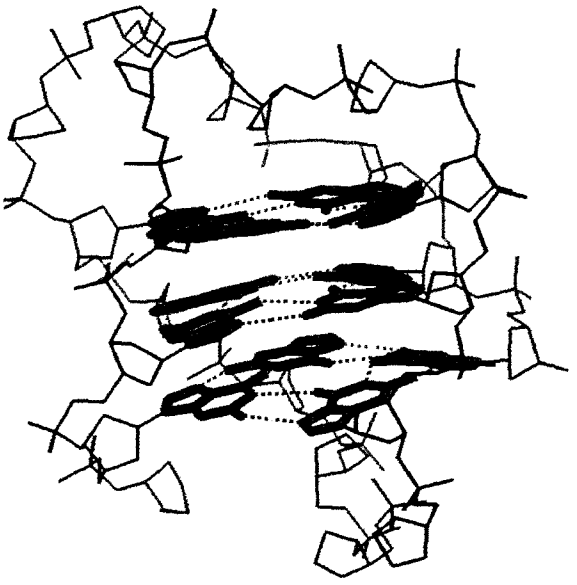


Figura 4. Cuadruplete formado por la secuencia telomérica del ADN. Únicamente se indican la cadena fosfodiester del ADN (en verde) y las guaninas asociadas (en rojo).

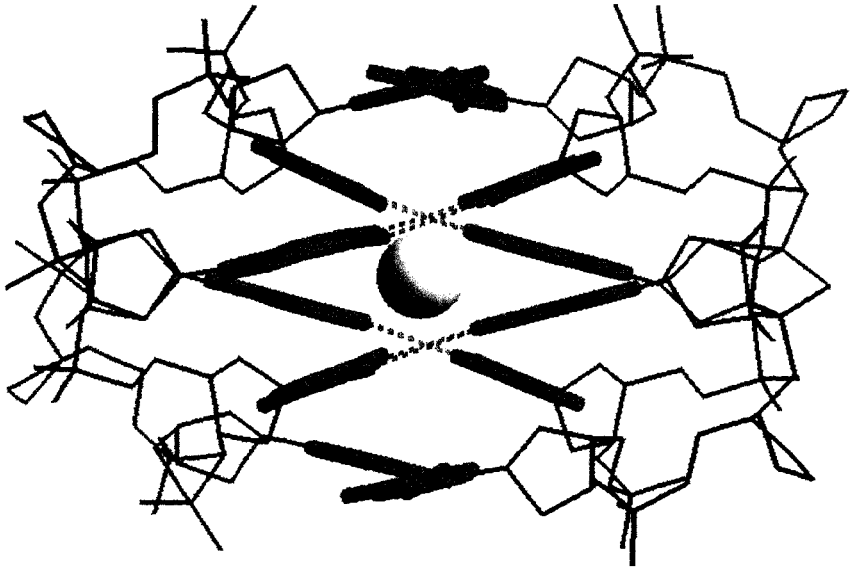


Figura 5. Asociación entre dos moléculas cíclicas de ADN (cadena principal en color verde), estabilizada por un ion Na^+ (en amarillo). Las bases (en rojo) forman puentes de hidrógeno.

ADNS CICLICOS

Las estructuras que hemos visto se han determinado por técnicas tales como la cristalografía de rayos X o la resonancia magnética nuclear (NMR), técnicas que requieren una estructura estable y que no sufra modificaciones importantes durante el tiempo de medida. Para estudiar otras conformaciones posibles que aparecen transitoriamente puede modificarse la estructura química del ADN dándole una mayor rigidez. Aquí es donde los químicos orgánicos juegan un papel esencial y han permitido grandes avances en el conocimiento del ADN gracias a la introducción de diversas modificaciones químicas. Vamos a describir sólo un ejemplo en el que han participado varios grupos de investigación españoles (10, 11).

Por métodos de síntesis orgánica es posible obtener ADNs de una sola hebra, pero que sean circulares. Al tener una sola hebra no puede formarse una doble hélice como las que hemos visto en las figuras 1-3. Sin embargo las bases del anillo pueden aparearse con sus bases complementarias. Como ejemplo en la figura 5 se muestra la estructura que adquiere un anillo sintético formado por ocho bases del ADN. La estructura en anillo estabiliza una conformación atípica, que no se observa normalmente en el ADN. Cada anillo forma dos pares de bases internos idénticos a los que se encuentran en el ADN nativo (Fig. 1), pero además estos pares de bases pueden formar un cuadruplete estabilizado por un ion sodio, tal como se muestra en la Figura 5. Estos cuadrupletes son distintos de los que forman las secuencias teloméricas (Fig. 4), pues aparecen entre pares de bases convencionales: un par AT con otro par AT, como se muestra en la Figura 5. En otros casos (12) es un par CG que puede asociarse con otro par CG. Esta tendencia a interactuar secuencias idénticas podría tener gran importancia en algunos procesos de recombinación genética.

El descubrimiento de esta estructura sólo ha sido posible gracias a disponer de un ADN cíclico sintetizado en el laboratorio, que no existe como tal en la célula. Sin embargo nos ha permitido descubrir nuevas posibilidades estructurales del ADN que pueden permitir aclarar aspectos nuevos en el comportamiento biológico del ADN.

BIOINFORMATICA Y PROYECTOS GENOMA: EL ADN COMO SOPORTE FISICO DE LA INFORMACION GENÉTICA

Para concluir vamos a ver brevemente cómo se puede extraer información de tipo químico y biológico a partir de los resultados que proporciona diariamente la secuencia ordenada de más de un millón de bases.

El objetivo del proyecto genoma humano es el de aumentar nuestro conocimiento sobre la base genética de las enfermedades humanas a través de la obtención de la secuencia de los 50000 a 80000 genes del ser humano. Sin embargo, la simplicidad con la que se formula este objetivo oscurece las enormes dificultades técnicas asociadas a su consecución. En particular, y sobre todas ellas, destacan dos: el tipo y gran volumen de información obtenida en la secuenciación masiva del ADN. Ello es debido a que la materia prima generada por el proyecto genoma no es más que una larguísima tira de letras -unos 3000 millones-, correspondientes a la secuencia lineal de las bases del ADN en los diferentes cromosomas, cuyo tamaño descarta cualquier tipo de análisis manual. La resolución de los problemas mencionados ha dado lugar al nacimiento de la bioinformática, una ciencia que se sitúa en la interfase entre la biología molecular y la informática.

En el análisis de la información genómica, la bioinformática distingue dos aspectos fundamentales (Figura 6): identificación de los genes y predicción de la estructura y función de las proteínas.

La identificación de los genes es un problema complejo debido a que sólo un 3% de la secuencia del ADN humano es codificante para alguna proteína. El resto del genoma corresponde a las áreas necesarias para la compactación y almacenamiento del ADN, control de la transcripción, etc., aunque no está totalmente claro cuál es la función de este 97% del ADN que juega un papel aparentemente secundario. El problema principal reside en el hecho de que no hay una señal clara que nos permita distinguir la secuencia de los genes entre el ruido de fondo de la secuencia total del ADN. Adicionalmente, en muchos casos las partes codificantes de un gen -los exones- se pueden combinar para dar lugar a diferentes proteínas, en

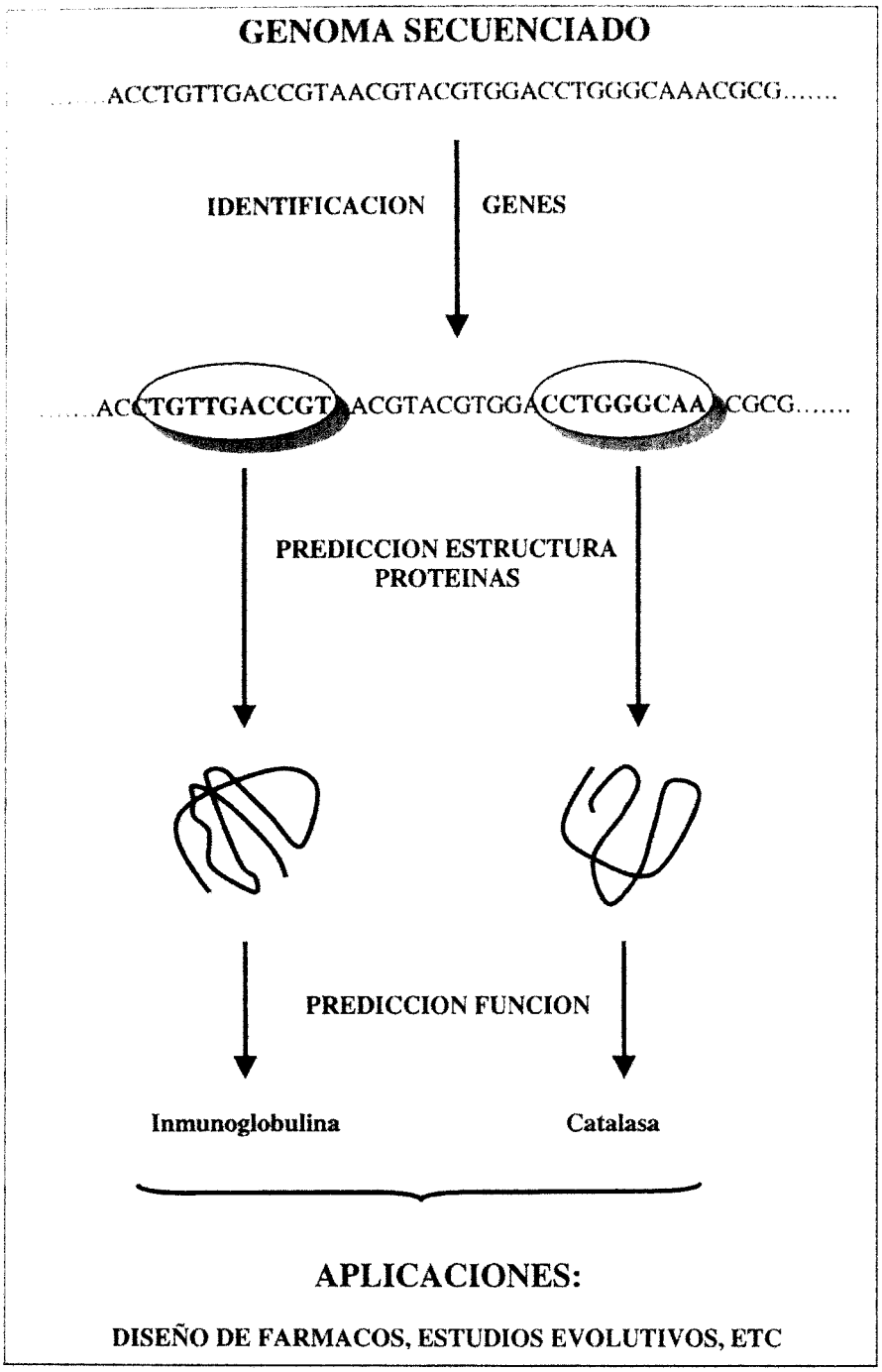


Figura 6. Esquema de los métodos de análisis de la información proporcionada por la secuencia de bases del genoma. Se trata de identificar los genes para predecir a continuación la estructura y función de las proteínas codificadas.

un proceso conocido como "splicing" alternativo. Para solventar estos problemas se han desarrollado una serie de estrategias basadas en el estudio de las propiedades estadísticas y composicionales de la molécula de ADN (13) y el uso de diferentes algoritmos -e.g. programación dinámica- (14). Aunque los resultados obtenidos son prometedores (en la actualidad se puede identificar correctamente un 40% de los genes presentes en el genoma), todavía falta un largo camino por recorrer hasta la identificación totalmente automática de los genes.

Una vez analizada la secuencia de ADN e identificados los genes, el segundo paso en el análisis de la información genómica consiste en la predicción de la estructura y función de la proteína expresada por cada gen. Este paso es crucial para el desarrollo de cualquiera de las aplicaciones biomédicas asociadas al proyecto genoma, ya que sólo la comprensión de la base molecular de las patologías puede dar lugar al desarrollo de terapias adecuadas. Sin embargo, la predicción de la estructura tridimensional de una proteína a partir del conocimiento de su secuencia de aminoácidos -derivada directamente de la secuencia del gen- es probablemente el problema más difícil de resolver en biología molecular estrucural (15). Para abordarlo se han desarrollado dos familias de métodos: *ab initio* y "threading fold recognition". La primera familia de metodos se basa en la construcción de modelos físico-químicos de las proteínas y cálculos de minimización de energía. Lamentablemente, los resultados obtenidos mediante estos métodos no son todavía lo bastante buenos como para permitir su uso generalizado en el marco de los proyectos genoma. Por ejemplo, ni siquiera utilizando las mejores predicciones *ab initio* del momento (16) es posible definir correctamente la geometría adecuada del centro activo de un enzima.

Los métodos de la segunda familia son totalmente empíricos y se basan en comparar la secuencia de la proteína "incógnita" con la de aquellas proteínas de estructura y función conocidas. Los resultados proporcionados por los métodos de esta segunda familia se han mostrado muy adecuados, siendo capaces de encontrar la estructura y función de un proteína "incógnita" en un 40% - 60% de los casos problema (17). Sin embargo, la mejora en la calidad de los modelos estructurales obtenidos todavía constituye un desafío importante.

CONCLUSIONES

A lo largo de este artículo hemos intentado demostrar que el ADN no es una simple serie de letras que "únicamente determina el orden de los aminoácidos en las proteínas". Se trata de una macromolécula compleja con múltiples conformaciones estructurales. Su estabilidad química le ha permitido ocupar un lugar central como

vehículo de la información genética. Pero evidentemente esta no es su única función. Aquí hemos pasado revista solamente a unos pocos aspectos de la estructura del ADN y del análisis de sus secuencias. El lector interesado puede ampliar su información acudiendo a las citas referenciadas o también a textos especializados sobre el ADN (18-21). En cualquier caso hay aún muchas propiedades del ADN que no conocemos plenamente y que pueden tener aplicaciones insospechadas. Por ejemplo, un aspecto que no hemos tratado es la posible utilización práctica de la propiedad de reconocimiento que tienen las secuencias de ADN, que se ha demostrado (22) puede utilizarse incluso como una computadora molecular!!. AQ

BIBLIOGRAFIA

1. M. Hattori, A. Fujiyama, T.D. Taylor, et al., The DNA sequence of human chromosome 21, Nature, 405, 311-319 (2000)

2. J. Sampedro, "Max Perutz. Premio Nobel de Química", El País, 14 marzo 2000.

3. M. Soler-López, L. Malinina y J.A. Subirana, Solvent organization in an oligonucleotide crystal. The structure of d(GCGAATTCG)2 at atomic resolution, J. Biol. Chem., 275, 230-340 (2000)

4. D. B. Nikolov, H. Chen, E. D. Halay, A. Hoffmann, R. G. Roeder and S. K. Burley, Crystal structure of a human TATA box-binding protein/TATA element complex, Proc. Natl. Acad. Sci. USA, 93, 4862-4867 (1996)

5. Z. S. Juo, T. K. Chiu, P. M. Leiberman, I. Baikalov, A. J. Berk and R. E. Dickerson, How proteins recognize the TATA Box, J. Mol. Biol., 262, 239-254 (1996)

6. T. R. Cech, Life at the End of the Chromosome: Telomeres and Telomerase, Angew. Chem. Int. Ed., 39, 34-43 (2000)

7. L. H. Hurley, R. T. Wheelhouse, D. Sun, et al., G-quadruplexes as targets for drug design, Pharmacology & Therapeutics, 85, 141-158 (2000)

8. V. A. Zakian, Telomeres: Beginning to understand the end, Science 270, 1601-1607 (1995)

9. Y. Wang y D. J. Patel, Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex, Structure 1, 263-282 (1993)

10. S. A. Salisbury, S. E. Wilson, H. R. Powell, et al., The bi-loop, a new general four-stranded DNA motif, Proc. Natl. Acad. Sci. USA, 94, 5515-5518 (1997)

11. C. González, N. Escaja, M. Rico and E. Pedroso, NMR structure of two cyclic oligonucleotides. A monomer-dimer equilibrium between dumbbell and quadruplex structures, J. Amer. Chem. Soc., 120, 2176-2177 (1998)

12. G. A. Leonard, S. Zhang, M. R. Peterson, et al., Self-association of a DNA loop creates a quadruplex: crystal structure of d(GCATGCT) at 1.8 Å resolution, Structure 3, 335-340 (1995)

13. T.K. Attwood and D.J. Parry-Smith, "Introduction to Bioinformatics", Longman, Dorset, 1999.

14. R. Guigó, Assembling genes from predicted exons in linear time with dynamic programming, J. Comp. Biol., 5, 681-702 (1998)

15. K.A. Dill, Folding Proteins. Finding a Needle in a Haystack, Curr. Opin. Struct. Biol., 3, 99-103 (1993)

16. L. P. Wei, E. S. Huang and R. B. Altman, Are predicted structures good enough to preserve functional sites?, Struct. Fold. Design, 7, 643-650 (1999)

17. X. de la Cruz and J. M. Thornton, Factors limiting the performance of prediction-based fold recognition methods, Protein Sci., 4, 750-759 (1999)

18. S. Neidle (ed.), "Oxford Handbook of Nucleic Acid Structure" Oxford University Press, Oxford (1999)

19. G. M. Blackburn and M. J. Gait, "Nucleic Acids in Chemistry and Biology" 2nd Ed., Oxford University Press, Oxford, 1996

20. W. Saenger, "Principles of Nucleic Acid Structure", Springer-Verlag, New York, 1984

21. J. A. Subirana, "Estructura del ADN", Editorial Alhambra, Madrid, 1985

22. K. Sakamoto, H. Gouzu, K. Komiya et al., Molecular computation by DNA hairpin formation, Science 288, 1223-1226 (2000)