Interferón y Biotecnología, Vol. 3, No. 2, 103-109 (1986)

Trabajos originales

Biología molecular y computadoras

DIETER SÖLL

Departamento de Biofísica y Bioquímica Moleculares
Universidad de Yale
New Haven, CT 06511, Estados Unidos

Recibido el 15 de octubre de 1985

RESUMEN

Este trabajo ofrece una visión general sobre las aplicaciones de la computación en la biología molecular y la organización de los sistemas de datos y fuentes de información.

SUMMARY

The present paper presents a general view about computers and their use in molecular biology, as well as data banks and main information sources.

INTRODUCCION

En este artículo desearía exponer una breve panorámica de la situación actual de la aplicación de la computación a las investigaciones que se llevan a cabo en el campo de la biología molecular. La discusión de estos temas refleja mi opinión personal, y aunque he contribuido en la realización de programas de propósitos generales para el análisis de secuencias de ácidos nucleicos y de proteínas (Sege *et al.*, 1981), no soy un experto en computación. Más bien, mis conceptos se basan en la edición de tres volúmenes sobre la *Aplicación de la computación a la investigación de los ácidos nucléicos* (Söll y Roberts, 1982; Söll y Roberts, 1984; Söll y Roberts, 1986). Estos numeros especiales de la revista *"Nucleic Acid Research"* presentan los principales avances en el software y las aplicaciones de la computación en esta esfera.

¿PARA QUE UTILIZA LA COMPUTACION EL BIOLOGO MOLECULAR?

El cuadro 1 presenta una lista de los objetivos primordiales para los que se utiliza la computación. Se mencionan algunos programas, y para más detalle, deben consultarse los artículos originales.

Copyright ©1986, Sociedad Iberolatinoamericana para Investigaciones sobre Interferón
Publicado por el Palacio de las Convenciones, La Habana, Cuba

Cuadro 1

PROGRAMA PARA EL BIOLOGO MOLECULAR*

Programas de objetivos generales para el análisis de secuencias de ácidos nucleicos y proteínas

Entrada y edición de información inicial
Búsqueda de homología, simetrías, etcétera, en una secuencia
Mapas de restricción enzimática a partir de secuencias de nucleótidos y aminoácidos
Traducción y retrotraducción
Uso de Codones
Búsqueda de marcos abiertos de lectura
Búsqueda de promotores
Búsqueda de inicialización de traducción
Comparaciones de secuencias (en especial, investigaciones sobre bases de datos)
Alineación de secuencias múltiples
Display y presentación de la información (por ejemplo, mapa circular de restricción, secuencia de ADN con señales de traducción y transcripción)
Geles de secuenciación (secuencia de disparo, "shotgun")
Construcción de mapas de restricción a partir de información de digestiones simples y dobles
Predicciones de la estructura secundaria del ARN
Predicciones de la estructura secundaria de la proteína

Programas de objetivos especiales

Síntesis del ADN con ayuda de la computación
Experimentos de hibridización
Electroforesis en gel, controlada por computación
Análisis de experimentos de rastreo ("foot print")
Generación de relaciones filogenéticas de las secuencias de ácidos nucleicos
Búsqueda de mutaciones de nucleótidos para crear zonas de restricción
Diseño de experimentos de clonaje
Cálculo de tamaño de los fragmentos (longitud), de acuerdo con la movilidad en el gel
Selección de las sondas de hibridación del ADN
Manejo de bancos de plásmidos
Probabilidad de codificación de los marcos de lectura de la secuencia de ADN
Ploteo de la estructura tridimensional del ADN, a partir de la secuencia primaria
Correlación de la secuencia de ARN de transferencia con la especificidad del aminoácido
Localización de repeticiones significativas en la secuencia
Localización de polimorfismo en los fragmentos largos de restricción (RFLP)

* Los artículos originales que describen los programas específicos pueden encontrarse en: Söll y Roberts (1982); Söll y Roberts (1984) y Söll y Roberts (1986)

Los principales programas de propósitos generales son útiles para operaciones que se ejecutan frecuentemente en las investigaciones de biología molecular. Una vez que las secuencias de ADN, ARN y proteínas son determinadas experimentalmente es necesario que se les analice.

Aparte de la entrada habitual por teclado, el uso de digitalizadores para "leer" la secuencia de los geles ha tenido gran éxito. Recientemente se ha añadido una entrada oral (Arnold *et al.*, 1986), este sistema es capaz de responder a la voz de operadores diferentes que suministren la información.

Se ha puesto mucho énfasis en la búsqueda de señales biológicas en secuencias de ADN y ARN. Se han descrito programas para encontrar promotores (Mulligan y McClure, 1986), zonas de iniciación de la traducción (Stormo *et al.*, 1984), zonas potenciales de conformación Z-ADN (Vass y Wilson, 1984) y secuencias significativas (Clift *et al.*, 1986). Además, se evalúa la significación de estas señales.

La alineación de las secuencias múltiples, así como la predicción de la estructura secundaria del ARN, tareas que en la computación resultan muy costosas, se han visto beneficiadas con el desarrollo de nuevos algoritmos (Sobel y Martínez, 1986; Williams y Tinoco, 1986).

A la vez se desarrolla un amplio rango de programas de objetivos especiales. Puede dirigirse la síntesis de grandes lotes de deoxioligonucléotidos sobre filtros de celulosa (Frank *et al.*, 1983; Matthés *et al.*, 1984) con un programa de computación que mantenga la información de las operaciones manuales (Tolstoshev *et al.*, 1986). Los experimentos de dactilogramas pueden cuantificarse y evaluarse por medio de computadoras (Dabrowiak *et al.*, 1986).

Resulta interesante el programa que localiza los nucleótidos de la secuencia de ADN que pueden ser alterados por mutaciones para introducir nuevos sitios de reconocimiento de enzimas de restricción, sin cambiar la secuencia de aminoácidos de la proteína codificada por el gen (Arentzen y Ripka, 1984). Algunos programas (por ejemplo, Lonsdale *et al.*, 1984) entre los análisis con Lotus 123 (Campione-Piccardo y Ruben, 1986) se han elaborado con el fin de manipular los bancos de clones y mantener información sobre el origen y vía utilizada en la construcción de los clones. Se han desarrollado nuevos programas para plotear la estructura tridimensional del ADN a partir de la secuencia primaria (Shapiro *et al.*, 1986; Tung y Harvey, 1986). Estos programas se basan en reglas relacionadas con ciertas distorsiones del ADN debidas al contexto de la secuencia y revelan la variación debido a secuencias específicas del ADN en su estructura terciaria, lo que puede resultar importante en el reconocimiento de la secuencia específica de las proteínas. Se ha elaborado un programa del ADN basado en el análisis electroforético de las digestiones con enzimas de restricción el cual permite el análisis del polimorfismo de los tamaños de los fragmentos de restricción (MacMurray, 1986). Finalmente se ha propuesto la representación de secuencias de ADN en forma musical, que pueden además ser leídas por la máquina (Cowin *et al.*, 1986).

¿Cuál es el objetivo de los Bancos de Datos y los Recursos Nacionales?

En el cuadro 2 se muestra un resumen de las principales instalaciones existentes. El objetivo principal del GenBank y del EMBL es la recopilación, organización y distribución de bases de datos de secuencias de ácidos nucleicos. *"Protein Identification Resource"* (PIR) tiene la misma función para las secuencias de proteínas. En todos los casos, la información se distribuye de forma regular a los suscriptores que han abonado la cuota correspondiente. En la actualidad, el GenBank y el EMBL tienen entre 6 y 8 meses de retraso en relación con la fecha de la publicación y la inclusión de una nueva secuencia. Se espera que puedan actualizar su colección de datos para principios de 1986.

Por supuesto, la secuencia de aminoácidos de las proteínas puede derivarse de la secuencia del ADN del gen. Así, en PIR, cerca del 80 por ciento de las secuencias de proteínas se derivan de la información del ADN, y las restantes se determinan por análisis proteico.

Aunque las instalaciones de los Bancos de Datos permiten la búsqueda y manipulación de las secuencias recopiladas, el desarrollo y la adquisición de software para el análisis no es uno de los objetivos primordiales del GenBank y del EMBL. Más bien, el análisis de la información de la secuencia puede realizarse localmente o por acceso remoto con instalaciones como BIONET.

Los objetivos de BIONET aparecen en el cuadro 2. BIONET fue fundada por la División de Recursos para la Investigación, de los Institutos Nacionales de Salud y trabaja con instalaciones de computación, el software central, y el respaldo de la División de IntelliGenetics, de IntelliCorp. El acceso a BIONET está limitado a organizaciones académicas no lucrativas. En la actualidad, cerca de 500 investigadores renombrados tienen acceso a esta instalación. La demanda esperada ha conducido a un esfuerzo en el establecimiento de satélites BIONET, instalaciones de computación que explotan el uso del software de BIONET localmente, pero que pueden comunicarse e intercambiar correspondencia electrónica con otros usuarios e instalaciones de BIONET.

Cuadro 2

BANCOS DE DATOS Y RECURSOS NACIONALES*
(Agosto 1985)

GENBANK	(Los Alamos, Boston) Recopilación, organización y distribución de datos sobre secuencias de nucleótidos (5544 secuencias, 5.038 x 10^6 bases)
EMBL	(European Molecular Biology Laboratory, Heidelberg) Recopilación, organización y distribución de datos sobre secuencia de nucleótidos (4835 secuencias, 4.5 x 10^6 bases)
PIR	(Protein Identification Resource, Washington) Recopilación, organización y distribución de datos sobre secuencia de aminoácidos (3311 secuencias; 740,030 aminoácidos)
BIONET	(Palo Alto) Brindar ayuda en computación en el campo del análisis de las secuencias de las proteínas y los ácidos nucleicos Servir de centro para el desarrollo e intercambio de nuevo software Promover el rápido intercambio de información entre científicos participantes (correspondencia electrónica y boletines)

* Hay una breve descripción de estas instalaciones en Söll y Roberts (1986).

¿Qué computadoras se utilizan para correr estos programas de software?

La ingeniosidad de los programadores, combinada con las necesidades específicas, ha conducido al desarrollo del software para máquinas de diversos tamaños, desde la calculadora programable manualmente hasta la supercomputadora. Muchos de los primeros programas se desarrollaron para las grandes computadoras de los centros de computación institucionales en condiciones de tiempo compartido. A causa de los avances tecnológicos y el descenso de los precios, las minicomputadoras han sustituido a estas computadoras y se han convertido en el principal recurso de la computación para la biología molecular.

Además, muchos investigadores y laboratorios individuales cuentan con microcomputadoras en su equipamiento. La disponibilidad inmediata de la computadora compensa generalmente el mayor tiempo de ejecución de las microcomputadoras. Asimismo, estas unidades locales sirven con frecuencia como terminales de computadoras mayores, en las que se pueden efectuar tareas de computación más exigentes. En virtud de estos acontecimientos, actualmente existen

muchos programas para minicomputadoras VAX y microcomputadoras IBM-PC. Aunque los programas aparecen en muchos lenguajes, los más utilizados son Fortran, Pascal, C y Basic.

¿Qué tipo de computadora es la más adecuada para nuestras necesidades?

La elección de la mejor computadora depende de muchos factores: las necesidades individuales, la disponibilidad de equipo o instalaciones y los conocimientos sobre computación que posea el investigador. En la actualidad, no todos los laboratorios tienen (o usan) computadoras modernas, acceso a Banco de Datos o comunicación electrónica con otros científicos. Por supuesto, muchas de las operaciones de la biología molecular pueden realizarse sin la computadora. No obstante, para analizar en detalle la información de la secuencia (en especial comparaciones de secuencia) o para utilizar el conocimiento de la estructura tridimensional de las proteínas en la planificación de futuros experimentos, son necesarios modernos instrumentos de computación.

¿De qué programas disponemos?

Altualmente no es fácil elegir entre muchos y diversos programas que han sido elaborados. Muchos de ellos se han publicado en *Nucleic Acid Research* (ver, por ejemplo, Söll y Roberts, 1982; Söll y Roberts, 1984 y Söll y Roberts, 1986) o en otras publicaciones más especializadas como *Computer Applications in the Biosciences* (CABIOS). Otros programas solo reciben la publicidad y gestión de ventas de organizaciones comerciales. Es muy probable que existan programas que puedan satisfacer las necesidades de todo el mundo. Para ayudar a resolver esta situación, GenBank ha creado recientemente un Centro Distribuidor de Software de Base de Datos, que contiene la información (alcance del programa, computadora utilizada, precio y existencia, etcétera), sobre muchos programas. De forma similar BIONET puede servir como una fuente de información sobre software y nuevos avances en el software.

Después de haber descubierto la existencia de un programa particular, el usuario potencial se enfrenta a una difícil elección.

Usualmente no se dispone de revisiones que detallen el producto. Korn y Queen (1984) publicaron una breve reseña comparativa de muchos de los programas existentes. Más recientemente, CABIOS dio inicio a una sección informativa de estos productos. Se espera que esta sección se amplíe para que abarque un rango mucho mayor.

Como mucho de estos programas son altamente comerciables, su precio puede cambiar desde un costo insignificante hasta miles de dólares. Actualmente es posible adquirir programas comerciales mediante una demostración para que el usuario pueda hacer una evaluación de ellos, no obstante, esto no constituye una norma.

Una de las dificultades en el anuncio de los nuevos progresos del software es la condición de especialización de las publicaciones. Los nuevos programas o las modificaciones de los ya existentes son frecuentemente considerados como inapropiados por la mayoría de las publicaciones que leen de forma habitual los biólogos moleculares. Además, el potencial para la distribución comercial del software puede impedir su publicación, ya que algunas revistas sólo editan trabajos sobre software si estas se hacen libremente accesibles.

¿Cuál es el futuro?

Actualmente parece que se trabaja mucho en la consolidación de diversos programas de manera que puedan llevarse a la práctica de forma más sencilla, para incorporarles mejores

cualidades y para facilitar el intercambio de información con las bases de datos (Modelevsky, 1985). Aunque es difícil predecir el futuro, puede darse por sentado que la introducción del software para la actualización automática y la incorporación automática de nueva información a sistemas mayores será un hecho.

Por ejemplo, si se ha buscado en una base de datos la homología con una secuencia de interés, se deseará tener las secuencias adicionales que se añadirán a la base de datos en algún momento futuro, cotejadas con la secuencia en cuestión. Podrán prepararse subrutinas (R. Roberts, comunicación personal), que de forma regular transferirán automáticamente la actualización de la base de datos a las computadoras individuales, se llevarán a cabo las búsquedas apropiadas (u otras rutinas) y la información saldrá impresa para el usuario. Además, el avance de la comunicación electrónica permitirá al biólogo molecular promedio, la utilización de la computadora como una interfase con la comunidad exterior. No obstante, el énfasis en los programas futuros deberá radicar en sencillez para el usuario, de manera que sean accesibles para la mayoría de los científicos e investigadores interesados que no sean expertos en las complejas operaciones de computación.

RECONOCIMIENTOS

Desearía extender mi agradecimiento al Doctor Richard Roberts por todos nuestros debates. Este manuscrito se basa en la conferencia impartida en la Reunión Anual de la Sociedad Norteamericana de Microbiología.

REFERENCIAS

ARENTZEN, R. y W. C. RIPKA (1984). *Introduction of Restriction Enzyme Sites in Protein-coding DNA Sequences by Site-specific Mutagenesis not Affecting the Amino Acid Sequence: A Computer Program.* Nucleic Acids Res. **12**: 777-787.

ARNOLD, J.; V. K. ECKENRODE; K. LAMKE; G. J. PHILLIPS y S. W. SCHAEFFER (1986). *A Comprehensive Packge for DNA Sequence Analysis in FORTRAN IV for the PDP-11.* Nucleic Acids Res. **14**: in press.

CAMPIONE-PICCARDO, J. y M. RUBEN (1986). *An Integrated Software System for Microcomputer Management of Recombinant DNA.* Nucleic Acids Res. **14**: in press.

CLIFT, B.; D. HAUSSLER; R. Mc CONNELL; T. D. SCHNEIDER y G. D. STORMO (1986). *Sequence Landscapes.* Nucleic Acids Res. **14**: in press.

COWIN, J. E.; C. H. JELLIS y D. RICKWOOD (1986). *A New Method of Representing DNA Sequences which Combines Ease of Visual Analysis with Machine Readability.* Nucleic Acids Res. **14**: in press.

DABROWIAK, J. C.; A. SKOROBOGATY; N. RICH; C. P. H. VARY y J. N. VOURNAKIS (1986). *Computer Assisted Microdensitometric Analysis of Footprinting Autoradiographic Data.* Nucleic Acids Res. **14**: in press.

FRANK, R.; W. HEIKENS; G. HEISTERBER-MOUTSIS y H. BLÖCKER (1983). *A New General Approach for the Simultaneous Chemical Synthesis of Larges Numbers of Oligonucleotides: Segmental Solid Supports.* Nucleic Acids Res. **11**: 4365-4377.

KORN, L. J. y C. QUEEN (1984). *Analysis of Biological Sequences on Small Computers.* DNA, **3**: 421-436.

LONSDALE, D. M.; T. P. HODGE y P. J. STOEHR (1984). *A computer Program for the Management of Small Cosmid Banks.* Nucleic Acids Res. **12**: 665-673.

MacMURRAY, A. (1986). *DIGICALC: A Restriction Fragment Analysis Program.* Nucleic Acids Res. **14**: 529-536.

MATTHES, H. W. D.; W. M. ZENKE; T. GRUNDSTRÖM; A. STAUB; M. WINTZERITH y P. CHAMABON (1984). *Simultaneous Rapid Chemical Synthesis of Over one Hundred Oligonucleotides on a Microscale.* EMBO J. 3: 801-805.

MODELEVSKY, J. L.; F. NORRIS y G. GRIESINGER (1986). *GENEVIEW and the DNACE Data Bus: Computational Tools for Analysis, Display and Exchange of Genetic Information.* Nucleic Acids Res. **14**: 387-403.

MULLIGAN, M. E. y McCLURE (1986). *Analysis of the Occurrence of Promoter-sites in DNA.* Nucleic Acids Res. **14**: 109-126.

SEGE, R.; D. SÖLL; F. H. RUDDLE y C. QUEEN (1981). *A Conversational System for the Computer Analysis of Nucleic Acid Sequences.* Nucleic Acids Res. **9**: 437-447.

SHAPIRO, B.; R. NUSSINOV; C. E. LIPKIN y J. V. MAIZEL (1986). *A Sequence Analysis System Encompassing Rules for DNA Helical Distortion.* Nucleic Acids Res. **14**: 75-86.

SOBEL, E. y H. MARTINEZ (1986). *A Multiple Sequence Alignment Program.* Nucleic Acids Res. **14**: 363-374.

SÖLL, D. y R. J. ROBERTS (1982). *The Applications of Computers to Research on Nucleic Acids.* IRL Press, Oxford, p. 457.

SÖLL, D. y R. J. ROBERTS (1984). *The Applications of Computers to Research on Nucleic Acids II.* IRL Press, Oxford. p. 856.

SÖLL, D. y R. J. ROBERTS (1986). *The Applications of Computers to Research on Nucleic Acids III.* IRL Press, Oxford, p. 5XX.

STORMO, G. D; T. D. SCHNEIDER; L. GOLD y A. EHRENFEUCHT (1981). *Use of the "Perceptron" Algorithm to Distinguish Translational Initiation Sites in E. Coli.* Nucleic Acids Res. **10**: 2997-3011.

TOLSTOSHEV, C. M.; H. W. D. MATTES y P. OUDET (1986). *Computer Management of Oligonucleotide Synthesis on Cellulose Filters.* Nucleic Acids Res. **14**: 405-415.

TUNG, C. y S. C. HARVEY (1986). *Computer Graphics Program to Reveal the Dependence of the Gross Three-dimensional Structure of the B-DNA Double Helix on Primary Structure.* Nucleic Acids Res. **14**: 381-387 in press.

VASS, J. K. y R. H. WILSON (1984). *"ZSTATS" - A Statistical Analysis for Potential Z-DNA Sequences.* Nucleic Acids Res. **12**: 825-833.

WILLIAMS, A. L. y I. TINOCO (1986). *A Dynamic Programming Algorithm for Finding Alternate RNA Secondary Structures.* Nucleic Acids Res. **14**: 299-315.