La Bioinformática: una perspectiva de la estructura de proteínas

Darrell Conklin*

Resumen

El Proyecto Genoma Humano está produciendo una inmensa cantidad de secuencias de nucleótidos que contienen genes que codifican proteínas nuevas. La tarea de la bioinformática *es interpretar esos datos. Aunque el código genético determine la secuencia de la proteína especificada por un gen, hasta la fecha no se ha podido solucionar el problema del plegamiento de las proteínas. Este trabajo presenta un repaso de la bioinformática como campo enfocado en una meta central: la predicción de la estructura de la proteína y el desarrollo de algoritmos rápidos para descubrir, solucionar e implementar partes del código de plegamiento de proteínas. La primera parte de este trabajo repasará los conceptos de genes y proteínas, y describirá las fuentes principales de secuencias de nucleótidos en bruto y derivadas. Se discutirán los problemas como el montaje de fragmentos cortos de secuencias de nucleótidos a secuencias de proteína más largas. La segunda parte planteará tres clases diferentes de métodos que se pueden usar para la predicción de estructura proteína: predicción de la estructura secundaria, predicción por semejanza y reconocimiento de plegamiento por enhebramiento.*

Abstract

The Human Genome Project is generating an enormous amount of raw nucleotide sequence data containing genes that code for novel proteins. It is the task of bioinformatics *to interpret these data. Although the genetic code determines the protein sequence specified by a gene, to date there has been no solution to the protein folding problem. This paper presents an overview of bioinformatics as a field driven by one central goal: the prediction of protein structure and the development of fast algorithms to uncover, solve, and implement portions of the protein folding code. The first part of this paper will review the concepts of genes and proteins, and will describe the main sources of raw and derived nucleotide sequence data. Problems such as the assembly of short nucleotide sequence fragments into longer protein coding sequences will be discussed. The second part of the paper will discuss three different classes of methods which can be used for the prediction of protein structure: secondary structure prediction, comparative modeling, and fold recognition by threading.*

1. Introducción

El Proyecto Genoma Humano es un plan multinacional para determinar la secuencia nucleica de cada gen en el genoma humano. Programada su finalización para el año 2005, el proyecto es el acontecimiento más revolucionario en la historia camino al entendimiento fundamental de la maquinaria de la biología humana a nivel molecular. Imprescindible para el éxito del proyecto es el campo nuevo de la *bioinformática*, una confluencia de Informática y Biología.

El conocimiento de todas las secuencias de gen implica, a través del código genético, el conocimiento de todas las secuencias de proteína en el cuerpo humano. Las proteínas son el material de importancia principal en la vida. El genoma humano cifra una gran diversidad de proteínas que funcionan, por ejemplo, como material estructural en el cuerpo, como enzimas que facilitan reacciones químicas en el cuerpo y como hormonas o mensajeros que facilitan la comunicación celular y que señalan el crecimiento de células. Algunas proteínas hormonales se usan para tratar enfermedades serias. Por ejemplo, la *insulina* se utiliza para el tratamiento de la diabetes, la *eritropoyetina* para el tratamiento de anemia y el *interferon* para el tratamiento de leucemia y esclerosis en placas. Varios laboratorios universitarios y empresas biotecnológicas están involucrados activamente en "minar" el genoma para descubrir nuevos candidatos de proteínas terapéuticas.

El código genético determina la secuencia de una proteína especificada por un gen. El "código de plegamiento de proteínas", que implicaría conocimiento de la estructura tridimensional de una proteína a través de su secuencia, no se conoce por completo. Además, es la estructura

* Profesor Titular de la Universidad Politécnica de Madrid
* Computational Biologist ZymoGenetics, Inc.

de una proteína la que en gran parte determina su función a nivel molecular.

Este trabajo pretende presentar la investigación de la bionformática como un campo enfocado en una meta central: la predicción de la estructura de una proteína y el desarrollo de algoritmos rápidos para descubrir, solucionar e implementar partes del código de plegamiento de proteínas. El énfasis estará en técnicas computacionales usadas para avanzar de secuencias nucleicas en bruto a un entendimiento de la estructura tridimensional de una proteína cifrada por un gen.

2. Genes y proteínas

El genoma humano comprende dos conjuntos de 23 cromosomas, cada uno con una cadena muy larga de nucleótidos (A, C, T, G) enrollados en un doble hélice con su cadena complementaria. Incluyendo el cromosoma Y (una copia en varones), 24 cromosomas deberán estar secuenciados. El Proyecto Genoma Humano finalmente producirá la detallada secuencia genómica de cada cromosoma. Estos datos serán un compuesto de secuencias de varios individuos y estarán almacenados en la base de datos llamada GenBank (1998).

Se asume que el genoma humano contiene aproximadamente 100,000 genes, separados por regiones intergenéticas (Figura 1, superior). Todos los cromosomas juntos comprenden unos 3 billones de nucleótidos. Un gen puede residir en cualquiera de las dos cadenas de un cromosoma, como indican las flechas en la parte superior de la Figura 1. Los genes contienen *exones* e *intrones*. Un gen, a través del proceso de *transcripción*, expresa un *ARN mensajero* (ARNm) que es semejante a la secuencia del gen pero sin intrones, los cuales se extraen. El proceso de transcripción también produce variantes de ARNm por medio de extracciones alternativas de exones. Cada ARN comprende tres regiones: la región 5' no traducida, la región transcrita y la región 3' no traducida (UTR, Figura 1, tercero). Mediante el proceso de *traducción*, la región transcrita se traduce a una proteína por medio del *código genético*: cada tres bases sucesivas comprenden un *codón* que especifica un *aminoácido*. La región transcrita está flanqueada por dos señales especiales: un codón ATG (que es el código del aminoácido metionina) y un codón finalizador (un codón de secuencia TAA, TAG, TGA). El código genético se califica a veces *redundante* porque, aparte de los codones finalizadores, hay 61 codones pero sólo 20 aminoácidos diferentes.

Es la proteína, la cadena de aminoácidos especificada por un gen, la que es de mayor importancia para el biólogo. Sin embargo, hay dos problemas difíciles en la *predicción de genes* que necesitan solucionarse cuando se están descubriendo proteínas en secuencias genómicas (de cromosoma) en bruto: primero, ¿dónde están los límites de los exones? y segundo, ¿cuál es la cartografía entre dichos exones y los ARNms transcriptos? Se ha logrado considerable progreso con respecto a estos problemas usando algoritmos que buscan patrones específicos que indican los límites de genes y exones (Burge y Karlin 1997).

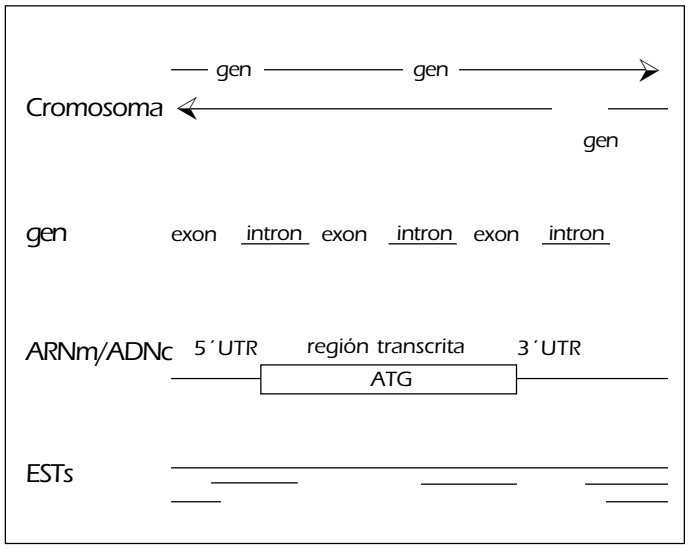

FIGURA 1. LOS TIPOS DE SECUENCIAS DE NUCLEÓTIDO.
Superior: secuencia de un cromosoma que contiene varios genes.
Segundo: secuencia de un gen individual que contiene intrones.
Tercero: un ADNc que contiene secuencia no traducida y la región transcrita de una proteína. Inferior: datos de ESTs, que son trozos cortos (de 200-300 nucleótidos) del ADNc. Cada línea representa un EST único que atraviesa la porción correspondiente del ADNc. Se ve que hay 4 grupos suerpuestos que se puede montar en 4 secuencias virtuales.

2.1 Datos EST

A pesar de los avances en la predicción de genes y el aumento de datos genómicos en bruto en GenBank, existe un camino más rápido al descubrimiento de ARNms expresados y sus proteínas traducidas. Hace veinte años, se lograron varios avances tecnológicos que les permitieron a los investigadores capturar y determinar la secuencia de ARNm (más específicamente, una forma complementaria llamada ADNc) directamente. La técnica consiste en purificar los millones de trozos de ARNm en bruto contenidos en un cultura celular, clonar el ADNc copiado metiéndolo en genomas de bacteria y determinar la secuencia del ADNc clonado. Más recientemente, la automatización ha permitido determinación rápida de secuencias parciales de grandes cantidades de clones escogidos al azar de aquellas genotecas. Cada secuencia parcial se llama un EST (Expressed Sequence Tag) (Figura 1). El repositorio principal de datos de EST se llama dbEST (Boguski et al., 1993).

En teoría la técnica de EST hace innecesario el proceso de descubrir proteínas a través de secuencias genómicas y predicción de genes. En la práctica hay varias limitaciones serias con esta técnica. Aunque ARNm padece de secuencia intrónica, las genotecas de ADNc pueden contener *pre*-ARNm que son genes transcritos incompletamente y por lo tanto contienen regiones intrónicas. También, algunos genes, debido a su baja tasa de transcripción, estarán presentes en muy pocos clones (por ejemplo, uno en cada cien mil) y sin secuenciar cada clon en una genoteca (una tarea costosa) nunca se encontrará en una base de datos de EST. Algunos genes se expresan solamente en pocos tipos de células y es posible que ni siquiera se encuentre un ADNc específico en una genoteca dada. En general, los ESTs muestran tasas altas de errores en la secuencia, de los cuales el más serio es el de nucleótidos perdidos que producen errores de *marco de lectura* en la proteína traducida. Como es de esperar, esto confunde el análisis enormemente. Finalmente, los ESTs no son nada más que trozos cortos del ARNm completo y miden sólo alrededor de 250 nucleótidos. Algunos de estos problemas se pueden corregir por medio del agrupamiento y montaje de los ESTs.

2.2 Agrupamiento y montaje de datos EST

Hay alrededor de 1.5 millones de ESTs disponibles en bases de datos públicas y por lo menos la misma cantidad en bases de datos privadas y propietarias. Se puede encontrar un nivel de redundancia y superposición, especialmente en los ARNms que son abundantemente expresados por las células. Una técnica poderosa para manejar esta redundancia es montar grupos de ESTs que superponen en secuencias largas "virtuales". Además de lograr la reducción de redundancia, la ventaja principal de usar montajes es que a menudo la región transcrita completa de un ARNm se puede reconstruir si hay bastante cobertura del gen en los ESTs. También, con suficiente cobertura, los errores en la secuencia de los ESTs individuales se pueden identificar y corregir con un algoritmo de consenso de montaje (Yee y Conklin 1998, Sutton et al., 1995). En ambos casos se puede extraer más secuencia de proteína transcrita de los datos en bruto de EST.

3. Predicción de la estructura de proteínas

Mientras progresa el Proyecto Genoma Humano, será necesario determinar la función biológica de un número creciente de secuencias de proteínas. La función de una proteína está estrechamente vinculada a su estructura, es decir, cómo se tuerce y pliega la cadena de aminoácidos en tres dimensiones. La estructura determina la posición espacial de los aminoácidos individuales, algunos de los cuales están involucrados en la actividad bioquímica de la proteína. Por ejemplo, los aminoácidos catalíticos de una enzima están fijados por el andamiaje tridimensional de la proteína. Aunque la información que determina la estructura de una proteína está contenida en su secuencia, el "código de plegamiento de la proteínas" no se entiende completamente. Por lo tanto, un papel importante de la bioinformática es predecir la estructura de nuevas proteínas y desarrollar teorías nuevas y algoritmos eficientes para la tarea.

Un recurso imprescindible para el científico de la bioinformática es el Protein Data Bank (PDB, 1998), que

es una base de datos de estructuras solucionadas de proteínas, es decir, cartografías de cada átomo en una proteína al espacio tridimensional. Se hace mediante de cristalografía a rayos X, un método que interpreta la estructura de difracción de cristales moleculares para crear mapas de densidad de electrones. La técnica lleva mucho tiempo y requiere la disponibilidad de proteínas cristalizadas. Por esta razón, el número de secuencias de proteínas conocidas excede enormemente el número de secuencias que tienen estructuras solucionadas en el PDB.

3.1 Clase de proteína y topología

Los elementos básicos de la estructura de la proteína son *las estructuras secundarias*: el *hélice alfa* (de unos 15 aminoácidos de largo) y la *cadena beta* (de unos 5 aminoácidos). Estos elementos están enlazados por regiones sin estructura definitiva (*espirales*) y cambios en la dirección de la cadena de la proteína (*curvas*). Estas estructuras secundarias se forman con piezas cortas de la secuencia, pero las proteínas se pliegan de tal manera que aminoácidos que se encuentran lejos en el secuencia resultan estar cerca en la estructura. Es por esta razón que el problema de predicción de estructuras de proteínas es tan difícil.

En 1976 Levitt y Chothia, en un trabajo clásico, propusieron una clasificación de estructuras de proteínas que consiste en cuatro clases: *alfa*, *beta*, *alfa/beta*, y *alfa+beta*. Hasta la fecha estos conceptos se han usado para organizar bases de datos de estructuras de proteínas (Murzin et al., 1995). Además, Levitt y Chothia reconocieron que dentro de cada clase estructural existen varias topologías: órdenes y direcciones de elementos *alfa* o *beta* principales. Diseñaron una representación diagramática para la topología de proteínas que se emplea hasta hoy día para subclasificar proteínas en grupos de topología (Westhead et al., 1998). Estas topologías se pueden representar en el ordenador como gráficas matemáticas y se pueden comparar para detectar semejanzas con algoritmos de correspondencia gráfica.

3.2 Predicción de estructura secundaria

Aunque hay una fuerte relación entre la composición de aminoácidos de una proteína y su clase estructural, usualmente una predicción de clase es demasiado general para ser útil. El conocimiento, por ejemplo, que una proteína es de clase alfa no sirve para distinguirla entre *globino*, *annexin* o *interferon*, todas proteínas de clase alfa con topologías (cantidad y conexiones de hélices) diferentes y por lo tanto funciones bioquímicas muy diferentes. La técnica computacional que proporciona el científico con información adicional de topología se llama *predicción de estructura secundaria*. Esta predicción de estructura secundaria es simplemente una secuencia de estados de estructura secundaria, por ejemplo, H significa hélice, E significa cadena, que corresponden a la secuencia de la proteína.

La mayoría de algoritmos diseñados usan el PDB para derivar sus modelos, usando relaciones observadas entre trozos cortos de secuencia contigua y estructura secundaria. Estas relaciones no son desgraciadamente específicas y los trozos cortos de secuencias idénticas pueden mostrar estructuras secundarias enteramente diferentes (Sternberg y Islam 1990). Además, el PDB no es lo suficientemente grande para contener estadísticas suficientes sobre trozos más largos y estructuralmente espec{ificos que se pueden encontrar en una secuencia nueva. Los mejores algoritmos (Rost y Sander 1993) actuales logran una precisión de alrededor del 70% (el número de estados de estructura secundaria asignados correctamente divididos por el número de aminoácidos de la proteína). Hay acuerdo general de que los métodos corrientes han alcanzado un límite superior de precisión, porque ninguno de ellos pueden modelar de manera aceptable las interacciones a larga distancia en la secuencia.

Queda por demostrarse concluyentemente que la estructura secundaria implica una topología única, o sea, que no existen dos secuencias de estados de estructura secundaria que estén derivadas de proteínas con topologías distintas. Di Francesco et al. (1997) describen un algoritmo que intenta predecir la topología de una proteína a través de una estructura secundaria predicha y observan

cierto éxito con proteínas de clase alfa. Aun si la hipótesis de una topología única resulta falsa, la secuencia correcta de estados de estructura secundaria limitarán dramáticamente las topologías que se necesitan considerar como posibilidades para una secuencia dada.

3.3 Modelado por semejanza

A lo largo de las épocas el genoma humano ha sido el objetivo de procesos evolutivos mayores como la duplicación, fusión, reorganización y supresión de genes. El gen individual ha sido sometido al proceso más sutil de mutaciones de bases, que a menudo cambia la secuencia del producto del gen. Los genes han evolucionado sustancialmente a la vez que han conservado la estructura tridimensional de su proteína. Esto ocurre porque mutaciones que alteran considerablemente el pliegue de una proteína suelen destruir su función normal y no persisten a lo largo de las generaciones. Además, los aminoácidos con características semejantes en características hidropáticas con frecuencia se pueden sustituir uno por otro sin un cambio apreciable en su estructura.

Por esta razón, el primer paso en la predicción del pliegue de una proteína nueva es determinar si está relacionada evolutivamente a cualquier secuencia en el PDB. La técnica usada para determinar la relación evolutiva entre dos secuencias es *alineamiento por pares* con un algoritmo de *programación dinámica* (e.g., Smith y Waterman 1981). Hoy en día es abundantemente claro que cuando un par de secuencias tiene un alto porcentaje de aminoácidos en posiciones alineadas, suelen tener pliegues muy similares. Sander y Schneider (1991) cuantificaron la relación entre porcentaje de identidad, largo de alineamiento y semejanza estructural estudiando proteínas con estructuras conocidas en el PDB. Más o menos, cuando un par de secuencias muestran por lo menos 25% de identidad en por lo menos 80 aminoácidos consecutivos, hay una probabilidad alta de que las secuencias tengan la misma estructura en la región alineada. Es muy importante recordar que el converso no funciona (Figura 2): debido a la divergencia evolutiva, es posible que un par de secuencias relacionadas no tengan un alineamiento significativo. En otras palabras, con alta probabilidad se podría encontrar, entre dos secuencias escogidas al azar, un alineamiento de calificación parecida.

Secuencias de proteínas que están relacionadas por antiguos acontecimientos evolutivos forman parte de una *familia* de secuencias. Para delinear una familia de proteínas es necesario, por lo tanto, tener acceso a todas las secuencias ancestrales, lo cual es imposible. No obstante, es un hecho conveniente que familias de proteínas se comporten como una relación de equivalencia. De interés particular es la *transitividad*: si una secuencia A está relacionada a B, y B está relacionada a C, se puede inferir que A está relacionada a C, aunque A y C no tengan un alineamiento significativo. Si cualquier secuencia en una clase equivalente tiene una estructura asociada en el PDB, se puede inferir que todas las secuencias en la familia tienen una estructura similar. Este método se usa rutinariamente para predecir la estructura de secuencias nuevas.

3.4 Motivos en secuencia y estructura

Aunque una familia de proteínas sea divergente, sería posible identificar cortas regiones que parecen tener se-

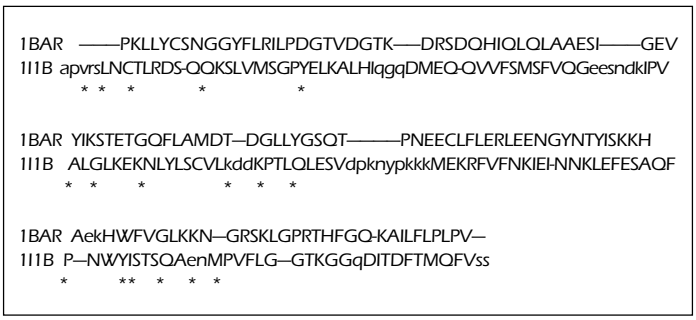

FIGURA 2. UN EJEMPLO DE UNA ALINEAMIENTO.

Los estructuras tridimensionales de factor acídico de crecemiento de fibroblasto bovino (código PDB 1BAR) e interleuken humano 1-beta (código PDB 1I1B) se han superpuesto automáticamente dando origen a este alineamiento de secuencias (Holm y Sander, 1998). Los aminoácidos idénticos están señalados con asteriscos. A pesar de la alta similitud de sus estructuras y de la relación probable evolutiva, este alineamiento no se puede detectar a través de algoritmos de alineamiento de secuencias por pares. Además, el porcentaje de identidad es tan bajo (menos de 15%), que este alineamiento no se le consideraría significativo estadísticamente.

cuencias conservadas y por lo tanto estructuras conservadas. Cada región puede ser descrita por un *motivo* que señala, para cada posición, la variación permitida en aminoácidos con una calificación particular para cada uno. Algoritmos de programación dinámica se usan para alinear motivos y secuencias. Los motivos se pueden entender como expresiones compactas de una familia de proteínas, una alternativa a la representación de la familia como una lista de su miembros. Además, una coincidencia entre un motivo y una secuencia, aunque no sea estadísticamente significativa, puede ser biológicamente significativa porque calificaciones altas pueden no ocurrir cuando se alinea el motivo a otras familias de proteínas. Si un motivo coincide con una secuencia de estructura desconocida se puede hacer la inferencia de que la secuencia tiene la misma estructura que la familia.

Las técnicas para crear motivos se subdividen en cuatro clases. La técnica normal crea el motivo basado en la variación observada en columnas de un *alineamiento múltiple* de la familia. Hay varios métodos para computar motivos usando un alineamiento múltiple (Gribskov 1987, Tatusov et al., 1994). Otras técnicas son algoritmos de *aprendizaje de máquinas* que intentan crear motivos sin usar un alineamiento múltiple (Brazma et al., 1998). Los *modelos Markov* tratan de encajar datos de secuencias disponibles con una secuencia de distribuciones de probabilidad usando un algoritmo de optimizacion local. Los más recientes son algoritmos *iterativos* que generalizan un motivo a través de búsquedas repetidas de una base de datos de secuencias (Altschul et al., 1997; Tatusov et al., 1994) usando un motivo cambiante. Esto se puede ver como la implementación de la transitividad de familias de proteínas.

Un trabajo importante de Unger et al. (1989) demostró sin duda que la columna estructural de la proteína puede ser descrita con el uso de cien elementos básicos hexamétricos (motivos estructurales de seis aminoácidos consecutivos). Ese hecho ha motivado a algunos investigadores a preguntar lo siguiente: ¿contienen los motivos estructurales cortos alguna especificidad en secuencia? y ¿existen motivos generales de secuencia que predicen la estructura local y que atraviesan fronteras de familias de proteínas? (Conklin 1995, Unger et al., 1989, Han y Baker 1995; Rooman et al., 1990). Dichos motivos podrían ser usados, en principio, para la predicción directa de estructuras tridimensionales de segmentos cortos de una nueva secuencia de proteína. Resulta importante que para un motivo estructural específico es necesario considerar un *conjunto* de motivos de secuencias. Es decir, varios motivos de secuencia podrían existir que predicen el mismo motivo estructural, pero un motivo en particular para esa estructura sería demasiado general y no serviría para la predicción. Así como el código genético es redundante, así lo será el código de plegamiento de proteínas.

3.5 Reconocimiento de plegamiento por enhebramiento

A menudo una nueva secuencia no contiene motivos reconocibles, ni tampoco se puede deducir su estructura modelándola por semejanza. En estos casos, la bioinformática recurre a métodos de *reconocimiento de plegamiento*. La tarea de reconocimiento de plegamiento es definida fácilmente pero es notoriamente difícil de solucionar: por una secuencia dada se determina cuáles estructuras en el PDB, si las hay, son compatibles con la secuencia.

Ya que la función de una proteína es determinada por la estructura tridimensional, mutaciones de aminoácidos que alteran enormemente la estructura por lo general apagan su función y serán suprimidas por los procesos evolutivos. Es por esta razón que, a pesar del inmensa cantidad de secuencias de proteína exploradas por la evolución a lo largo del tiempo, probablemente existan sólo unas miles de topologías únicas de proteínas (Chothia 1992). Mientras el PDB sigue su expansión con estructuras nuevamente solucionadas, la probabilidad de que un nuevo producto genético se pliegue como una estructura conocida continuará en aumento.

El reconocimiento de plegamiento a través de *enhebramiento* es un enfoque nuevo y poderoso. Estos métodos están basados en la suposición de que las estructuras

de proteínas se pliegan de manera que obtienen un estado de mínima energía libre y que esta energía se puede calcular para cualquier estructura. Dicha computación toma en cuenta la compatibilidad de diferentes aminoácidos en cada posición de la estructura. La compatibilidad suele reflejar la preferencia de aminoácidos hidrofóbicos en el ambiente nucleico de la proteína y la energía potencial creada cuando dos aminoácidos se acercan.

Dado que existe una función que puede medir y evaluar la compatibilidad de una secuencia con un plantilla estructural que no contiene la secuencia nativa, los algoritmos de enhebramiento intentan minimizar esta función al considerar varios alineamientos entre secuencia y estructura. La tarea es enormemente compleja puesto que un número exponencial (como función del tamaño de la secuencia y estructura) de alineamientos son posibles y la existencia de un número arbitrario de interacciones por pares en la estructura impide el uso de algoritmos de programación dinámica para sacar soluciones óptimas. Frente a este complejo problema, hay dos algoritmos heurísticos interesantes para obtener por lo menos una solución viable. El primero es el método de Jones et al. (1992) que usa una variante del algoritmo establecido de programación dinámica. El segundo es el método de muestro estadística de Madej et al. (1995) que repetidamente modifica un alineamiento subóptimo hasta encontrar una mínima local de energía. Ambos han tenido cierto éxito en predicción del plegamiento de proteínas problema, aunque la baja selectividad (proteínas de estructuras diferentes que parecen ser compatibles) seguirá siendo una problema importante.

4. Conclusiones

Este trabajo ha proporcionado un repaso de la bioinformática desde la perspectiva de la estructura de proteínas. Se discutieron las fuentes de datos en bruto de secuencias nucleicas con un enfoque particular en bases de datos de ESTs. Es posible derivar bases de datos más útiles y menos redundantes montando fragmentos cortos de ESTs. Sin embargo, algunos ARNms no pueden ser montados a través de datos de ESTs, debido a su baja tasa de transcripción. Las proteinas cifradas por estos ARNms esencialmente quedarán invisibles hasta que (o a menos que) algoritmos de predicción de genes los detecten en secuencias genómicas en bruto.

Cuando la región transcrita completa de un gen es conocida, es importante desempeñar alguna clasificación según su estructura predicha. Una vez que una estructura se predice, los científicos pueden comenzar a inferir rasgos centrados en proteínas con estructuras similares. Este trabajo ha proporcionado un resumen de varios tipos de predicciones estructurales que se hacen según la clase y la topología de una proteína al plegamiento. Esta tarea será cada vez más importante mientras el Proyecto Genoma Humano siga generando más secuencias nuevas de proteínas

Bibliografía

ALTSCHUL, S., MADDEN, T., SCHAFFER, A., ZHANG, J., ZHANG, Z., MILLER, W., and LIPMAN, D., 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, 25(17):3389-3402.

BOGUSKI, M., LOWE, T. and TOLSTOSHEV, C., 1993, dbEST-database for Expressed Sequence Tags, *Nature Genetics*, 4:332-333.

BRAZMA, A., JONASSEN, I., EIDHAMMER, I., and GILBERT, D., 1998, Approaches to the automatic discovery of patterns in biosequences, *Computational Biology*, 5(2):279.

BURGE, C. and KARLIN, S., 1997, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, 268:78-94.

CHOTHIA, C., 1992, One thousand families for the molecular biologist, *Nature*, 357:543-544.

CHOU, K., 1995, A novel approach to predicting protein structural classes, *PROTEINS: Structure, Function, and Genetics*, 21:319-344.

CONKLIN, D., 1995, Machine discovery of protein motifs, *Machine Learning*, 21:125-150.

DANDEKAR, T. and KONIG, R., 1997, Computational methods for the prediction of protein folds, *Biochimica et Biophysica Acta*, 1343:1-15.

DI Francesco, V., GARNIER, J. and MUNSON, P. , 1997, Protein topology recognition from secondary structure sequences: application of the hidden Markov models to the Alpha class proteins, *J. Mol. Biol.*, 267:446-463.

GenBank, 1998, http://www.nlm.nih.gov.

GRIBSKOV, M., MCLACHLAN, A., and EISENBERG, D. , 1987, Profile analysis: detection of distantly related proteins, *Proc. Natl. Acad. Sci. U.S.A.,* 84:4355.

HAN, K. and BAKER, D., 1995, Recurring local sequence motifs in proteins, *J. Mol. Biol.*, 251:176-187.

HOLM, L. and SANDER, C., 1998, Touring protein fold space with DALI/FSSP., *Nucleic Acids Research*, 26(1):316-319.

JONES, D., TAYLOR, W., and THORNTON, J., 1992, A new approach to protein fold recognition, *Nature*, 358:86-89.

KROGH, A., BROWN, M., MIAN, I., SJOLANDER, K., and HAUSSLER, D., 1994, Hidden Markov models in computational biology, *J. Mol. Biol.*, 235:1501-31.

LEVITT, M. and CHOTHIA, C., 1976, Structural patterns in globular proteins, *Nature*, 261:552-557.

MADEJ, T., GILBRAT, J. and BRYANT, S., 1995, Threading a database of protein cores, *PROTEINS: Structure, Function, and Genetics*, 23:356-369.

MURZIN, A., BRENNER, S., HUBBARD, T., and CHOTHIA, C., 1995, SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247:536-540.

Protein Data Bank, 1998, http://www.pdb.bnl.gov.

ROOMAN, M., RODRIGUEZ, J., and WODAK, S., 1990, Relations between protein sequence and structure and their significance, *J. Mol. Biol.*, 213:337-350.

ROST, B. and SANDER, C., 1993, Prediction of protein structure at better than 70% accuracy, *J. Mol. Biol.*, 232:584-599.

SANDER, C. and SCHNEIDER, R., 1991, Database of homology-derived protein structures and the structural meaning of sequence alignment, *PROTEINS: Structure, Function, and Genetics*, 9:56-68.

SMITH, T. and WATERMAN, M., 1981, Identification of common molecular subsequences, *J. Mol. Biol.*, 147:195-197.

STERNBERG, M. and ISLAM, S., 1990, Local protein sequence similarity does not imply a structural relationship, *Protein Engineering*, 4:125-131.

SUTTON, G., WHITE, O., ADAMS, M., and KERLAVAGE, A., 1995, TIGR assembler: a new tool for assembling large shotgun sequencing projects, *Genome Science and Technology*, 1, 1:9-19.

Swiss-Prot., 1998, http://expasy.hcuge.ch/sprot.

TATUSOV, R., ALTSCHUL, S., and KOONON, E., 1994, Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks, *Proc. Natl. Acad. Sci. USA*, 91:12091-12095.

UNGER, R., HAREL, D. WHERLAND, S., and SUSSMAN, J., 1989, A 3D building blocks approach to analyzing and predicting structure of proteins, *PROTEINS: Structure, Function, and Genetics*, 5(4):355-373.

WESTHEAD, D., HATTON, D. and THORNTON, J., 1998, An atlas of protein topology cartoons available on the world-wide web, *Trends in Biochemical Sciences*, 23:35-36.

YEE, D. and CONKLIN, D., 1998, Automated clustering and assembly of large EST collections. *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press. 203-211.