

SimProt: PROGRAMA COMPUTACIONAL PARA GENERAR SECUENCIAS DE PROTEÍNAS SIMPLIFICADAS

Jiménez Montaña M. A., Del Angel Ortiz R., Lucio García H. R. y Ramos Fernández A.

Facultad de Física e Inteligencia Artificial, Universidad Veracruzana. Sebastián Camacho # 5, Col. Centro, Xalapa, Ver., México, C.P. 91000, Tel. (228) 8172957, Fax. (228) 8172855. ajimenez@uv.mx

RESUMEN: A partir de una proteína funcional se infiere una gramática estocástica para generar variaciones de la secuencia de ARN original, de acuerdo al procedimiento descrito en un trabajo anterior¹. Se describe un programa de cómputo (SimProt) para implementar dicha gramática, que genera secuencias de ARN que codifican proteínas de un tipo especificado. Se ilustra su utilidad con dos ejemplos: 1) generando secuencias de proteínas helicoidales y 2) variantes de la enzima orotato fosforribosiltransferasa de *E. coli* (OPRTasa), que emplean un alfabeto reducido de 13 aminoácidos. SimProt consta de una clase en Java (código ejecutable) llamada CFGen la cual genera la secuencia estocástica de bases. Esta clase es accesible desde Matlab y se utilizan las funciones básicas de Matlab para traducir esta secuencia de acuerdo al código genético. Una vez que la secuencia es traducida a aminoácidos, se utiliza una consulta por medio del método POST para obtener, a través del modelo elegido de predicción de estructuras secundarias, una estimación del porcentaje de aproximación entre la secuencia original y la generada. El programa se puede utilizar tanto en modo automático como manual para hacer las consultas. Para los dos ejemplos estudiados se obtuvieron secuencias con una semejanza en la estructura secundaria, con respecto a la secuencia original, mayor del 80 % en promedio. Para evaluar la calidad de la predicción se generaron subrogados aleatorios, obteniéndose valores de Z, entre 9 y 11 desviaciones estándar.

ANTECEDENTES Y PROPÓSITOS: El diseño de proteínas *de novo* y los experimentos de mutagénesis dirigida *in vitro* son de gran utilidad para entender la evolución de las proteínas, estimar su robustez², diseñar enzimas con nuevas funciones y para construir proteínas simplificadas. La teoría de la evolución nos proporciona un poderoso algoritmo que puede implementarse en el laboratorio o simularse en una computadora para rediseñar proteínas. En este trabajo presentamos un programa de cómputo (SimProt) para implementar el método heurístico, para generar secuencias de ARN correspondientes a proteínas de una clase seleccionada de antemano, propuesto en un trabajo previo¹.

MATERIALES Y MÉTODOS. Siguiendo la estrategia de codificación binaria de aminoácidos (modelo HP) propuesta Kamtekar et al.³, para diseñar proteínas formadas por cuatro hélices en paquete (four-helix boundle), se generó una gramática que describiera, a nivel de ARN, el patrón binario (H, P). Este método emplea un procedimiento de simplificación en el cual los segmentos en que se divide la proteína se simplifican uno por uno, de manera independiente, y después se unen para formar un solo gene. Tras generar 100 secuencias de bases

nucleicas que cumplen las probabilidades de aplicación de las reglas, se eligieron 5 secuencias al azar; estas secuencias fueron traducidas a secuencias de aminoácidos utilizando el código genético dando origen a 74 residuos, posteriormente se realizó la predicción de su estructura secundaria por consenso con la herramienta NPS@⁵. Las estructuras secundarias de dichas secuencias fueron comparadas con la secuencia 86 del artículo de Wei et al.⁴ obteniéndose un grado promedio de similitud del 85%. Como control, se obtuvieron 10 subrogados aleatorios de cada una de las secuencias seleccionadas. Se eligieron cinco al azar y se predijo su estructura secundaria con la técnica descrita obteniéndose, como se esperaba, que no cumplen con la sucesión de hélices y vueltas de las proteínas de cuatro hélices en paquete.

Esta estrategia de simplificación no funciona para proteínas más complicadas, porque no toma en cuenta las correlaciones de largo alcance entre los aminoácidos. Para resolver este problema Akanuma et al.⁶ propusieron una estrategia de simplificación *secuencial*, la cual funciona porque la interferencia entre las mutaciones de diferentes segmentos puede evitarse a lo largo del proceso secuencial de variación de la proteína original. Realizamos una implementación computacional que simula este proceso, usando como ejemplo la misma enzima empleada en el trabajo experimental. Partiendo de la enzima E. Coli OPRTasa, inferimos una gramática estocástica para generar variaciones de la secuencia de ARN original. Las secuencias traducidas sólo emplean un alfabeto reducido de 13 aminoácidos, que son los siguientes: A,D,G,L,P,R,T,V,Y,E,F,S y K.

RESULTADOS Y CONCLUSIÓN: Para las proteínas helicoidales, obtuvimos secuencias que, de acuerdo los métodos de predicción de estructura secundaria empleados, difieren de la secuencia original en menos del 15 %. Para la enzima E. Coli OPRTasa obtuvimos variantes con diferencias menores al 17 %, que es la variación obtenida experimentalmente⁶. El procedimiento propuesto puede servir para ahorrar trabajo en el diseño de experimentos de mutagénesis dirigida *in Vitro*, porque permite escoger un conjunto de secuencias iniciales (primers) que, presumiblemente, corresponden a proteínas con estructuras semejantes a la enzima que se pretende modificar. En experimentos de evolución dirigida las clonas con un promedio de 5 mutaciones aleatorias pierden su estructura activa², mientras que con el procedimiento secuencial Akanuma et al.⁶ obtuvieron secuencias que toleraron 73 sustituciones de aminoácidos. Con nuestro procedimiento computacional obtuvimos secuencias con 35 substituciones que, presumiblemente, conservan la misma estructura.

¹ Jiménez-Montaña M. A., Lucio-García H. R., Ramos-Fernández A., 2005. *Periodicum Biologorum* 107 (4):397-402 .

² Kunichika K. , Hashimoto Y Imoto T., 2002. *Protein Engineering* vol.15 (10): 805–809

³ Kamtekar S., Schiffer J.M., Xiong H., Babik J.M., Hecht, M.H. 1993. *Science*. 262: 1680-1685

⁴ Wei Y., Hecht M.H. 2004. *Protein Engineering, Design & Selection* 17: 67-75

⁵ NPS@ Network Protein Sequence Analysis

http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html

⁶Akanuma S., Kigawa T., Yokoyama S. 2002. PNAS, 99 (21): 13549–13553

Agradecimiento: Al Fondo Sectorial de Investigación para la Educación SEP-CONACYT, Proyecto: Códigos Correctores de Errores en Biología Molecular, SEP-2003-CO2_44625 y al Sistema Nacional de Investigadores, por apoyo parcial.