*Biotecnología Aplicada, Vol. 9, No. 1, 78-82 (1992)*

# SYNSOS: Paquete de programas de ayuda en el diseño de genes

**M. OCHAGAVÍA, V. JIMÉNEZ, J. FERNÁNDEZ DE COSSÍO, A. SUÁREZ, R. BRINGAS y R. RICARDO**

**Centro de Ingeniería Genética y Biotecnología (CIGB), Apartado 6162, La Habana 6, Cuba**

*Recibido en abril de 1991*
*Aprobado en octubre de 1991*

## RESUMEN

Este trabajo presenta un paquete de programas de computación de ayuda en el diseño de genes sintéticos.

Dada la secuencia de ADN que codifica para el gen que se desea obtener y la tabla de uso de codones apropiada, el programa introduce cambios en la secuencia de ADN que eliminan las subsecuencias repetidas, inversas repetidas y palindrómicas de mayor longitud, presentes en la secuencia original, con lo cual se disminuyen las uniones inespecíficas durante la síntesis enzimática y se eleva la eficiencia de la hibridación.

El paquete permite, además, editar secuencias, realizar análisis de restricción, editar tablas de uso de codones y, dada una proteína, obtener la secuencia de ADN formada por los codones de mayor uso basada en una tabla de uso de codones cualquiera que seleccione el usuario.

Este programa fue utilizado con buenós resultados en el diseño de varios genes sintéticos.

## SUMMARY

A computer tool for synthetic gene design is reported.

As input data the program requires the nucleotide sequence of the desired gene and a codon usage table. The algorithm eliminates repeats, inverse repeats and palindromic subsequences, substituting codons by their synonyms, taking into account the usage frequency of each codon, thus reducing non-specific joins during enzymatic synthesis which increases hybridization efficiency.

The software includes a powerful sequence editor and a restriction analysis option, and allows the update of any existing codon usage table as well as input of new tables. Taking the most frequently used codons for each aminoacid, the program can reverse translate a given protein sequence.

The program has been successfully used in the design of several genes.

## INTRODUCCION

La sustitución de algunos aminoácidos de una proteína (mediante mutagénesis dirigida o síntesis químico-enzimática del gen) puede determinar cambios sustanciales en su actividad biológica, lo que constituye un campo (ingeniería de proteínas) muy prometedor y en el cual el empleo de la computación desempeña un papel fundamental.

En el diseño de genes sintéticos, es una práctica común el empleo de programas de computación orientados al análisis de secuencias de ADN y proteínas, que ofrezcan

facilidades de edición de secuencias, análisis de restricción, predicción de estructura secundaria de ARN, etc. Programas tales como el CIBSOFT (Bringas y Pérez, 1986), o el potente PC/GENE™, permiten realizar estas tareas. No obstante, la selección adecuada de la secuencia nucleotídica a sintetizar requiere un análisis exhaustivo, imposible de realizar manualmente y no previsto en paquetes de programas de propósito general como los mencionados anteriormente.

La longitud máxima de los oligonucleótidos posibles de obtener en una máquina sintetizadora automática es limitada (oligonucleótidos de hasta 40 nucleótidos, aproximadamente). Esto implica que para lograr la síntesis de un gen se hace necesario combinar los fragmentos obtenidos en la máquina, mediante un proceso de ligazón.

En dicho proceso, es necesario que cada fragmento hibride con su complementario y sólo con este. Si existen oligonucleótidos con elevado porcentaje de complementariedad, con otros que no son su real complementario, la eficiencia del proceso de ligazón se afecta considerablemente.

En este trabajo describimos un programa que realiza los cambios necesarios en la secuencia nucleotídica dada por el usuario, para eliminar las subsecuencias repetidas de la secuencia original y de su complementaria, con lo cual se evita que durante el proceso de ligazón se produzcan uniones inespecíficas.

## MATERIALES Y METODOS

El paquete de programas fue desarrollado en Turbo Pascal®, para microcomputadoras IBM XT,AT,PS/2 y compatibles. Está constituido por diferentes opciones accesibles al usuario desde un menú principal (figura 1).

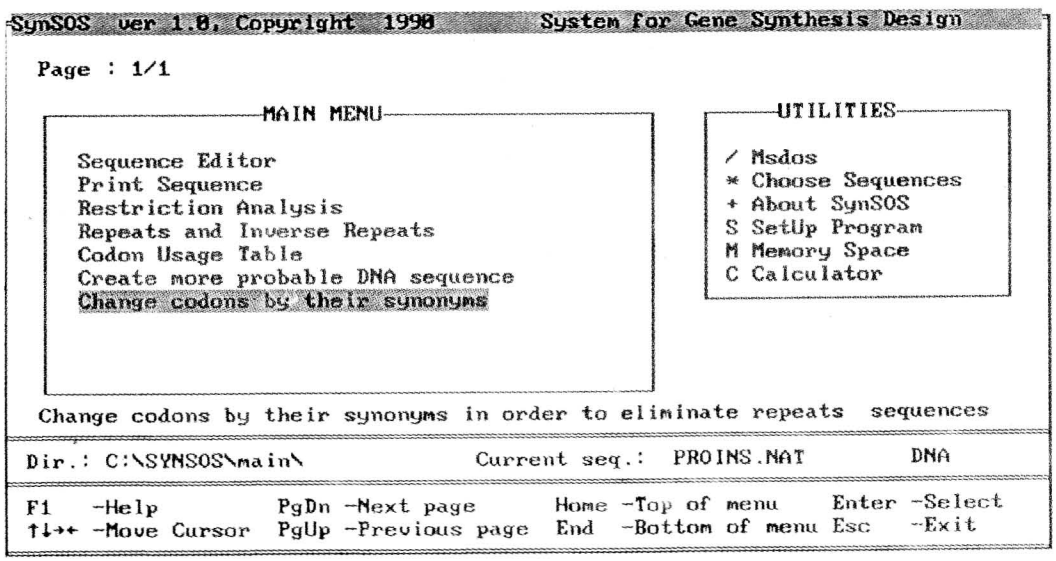


FIG. 1. Menú principal del SYNSOS.

El editor permite manipular hasta 9 secuencias simultáneamente, ofrece todas las facilidades de un editor de texto estándar y un extenso conjunto de opciones específicas para el tratamiento de secuencias biológicas*.

La opción de impresión de secuencias permite visualizar, imprimir o salvar secuencias en diferentes formatos.

La opción de análisis de restricción permite seleccionar las enzimas de restricción que se encuentran almacenadas en una base de datos, así como ofrece la posibilidad de incorporar nuevas enzimas a la base. Los resultados del análisis de restricción se presentan de diferentes formas, incluyendo una elegante representación gráfica de fácil comprensión.

La determinación de subsecuencias repetidas, inversas repetidas y palindrómicas se programó según un eficiente algoritmo que ejecuta en un espacio lineal (Martínez, 1983).

La opción de edición de tablas de uso de codones permite al usuario introducir o modificar tablas de uso de codones para diferentes organismos o especies.

La siguiente opción en el menú principal permite, dada una proteína, obtener la secuencia nucleotídica codificante para dicha proteína, formada por los codones de mayor uso (según una tabla de uso de codones previamente seleccionada).

La última opción del menú es el módulo más importante de este paquete de programas. Esta opción es la que permite eliminar las subsecuencias repetidas en la secuencia nucleotídica.

## Descripción del algoritmo de eliminación de repetidas

El programa solicita al usuario las siguientes informaciones:

1. Nombre de la secuencia nucleotídica fuente.
2. Nombre del fichero que contendrá la nueva secuencia nucleotídica.
3. Longitud máxima de las repetidas admisibles ($L$).
4. Nombre del fichero que contiene la tabla de uso de codones a tener en cuenta en las sustituciones.
5. Nombre del fichero donde se almacenarán los resultados del programa.

### *Pasos del algoritmo*

*A.* **Eliminación de subsecuencias repetidas:**

1. Se determinan las subsecuencias repetidas de mayor longitud.

   Sea $N$ la longitud máxima de las repetidas encontradas.

   Sean $C_i$ $i = 1,...,M$ los conjuntos que contienen las subsecuencias repetidas iguales entre sí y $m_i$, $i = 1,...,M$ la cantidad de repetidas dentro de cada conjunto $C_i$.
2. Para cada conjunto $C_i$, se intenta sustituir al menos un codón en $m_i$-1 subsecuencias. Si no todas las mi-1 subsecuencias pueden ser cambiadas, ello indica que la secuencia nucleotídica final contendrá repetidas de longitud $N$. Para cada subsecuencia se verifica que el posible cambio no introduzca nuevas repetidas de longitud mayor o igual que ella.
3. Se efectúan las sustituciones posibles, con lo cual se obtiene una secuencia nucleotídica nueva, en la que se deben haber eliminado total o parcialmente las repetidas de longitud $N$.
4. Se determinan las repetidas de mayor longitud $N'$, $N' < N$ y se sustituye $N$ por $N'$.
5. Se repiten los pasos (2), (3) y (4) mientras $N < L$.

*B.* **Eliminación de subsecuencias inversas repetidas:**

Se realiza como se describió en (A) para la eliminación de repetidas, excepto que en el punto (2) se verifica además que cada posible cambio de una subsecuencia no introduzca una nueva repetida de longitud mayor o igual que ella.

*C.* **Eliminación de subsecuencias palindrómicas:**

Se realiza como se describió en (A) para la eliminación de repetidas, excepto que en el punto (2) se chequea, además, que cada posible cambio de una subsecuencia no introduzca una nueva repetida, ni una nueva inversa repetida de longitud mayor o igual que ella.

### *Resultados del programa*

Al concluir la ejecución del programa se obtiene una nueva secuencia. En la pantalla se muestran ambas secuencias y se señalan con asteriscos las bases sustituidas (figura 2). Los resultados se almacenan en los ficheros propuestos inicialmente por el usuario.

*Bringas R., *et al.*, resultados no publicados.

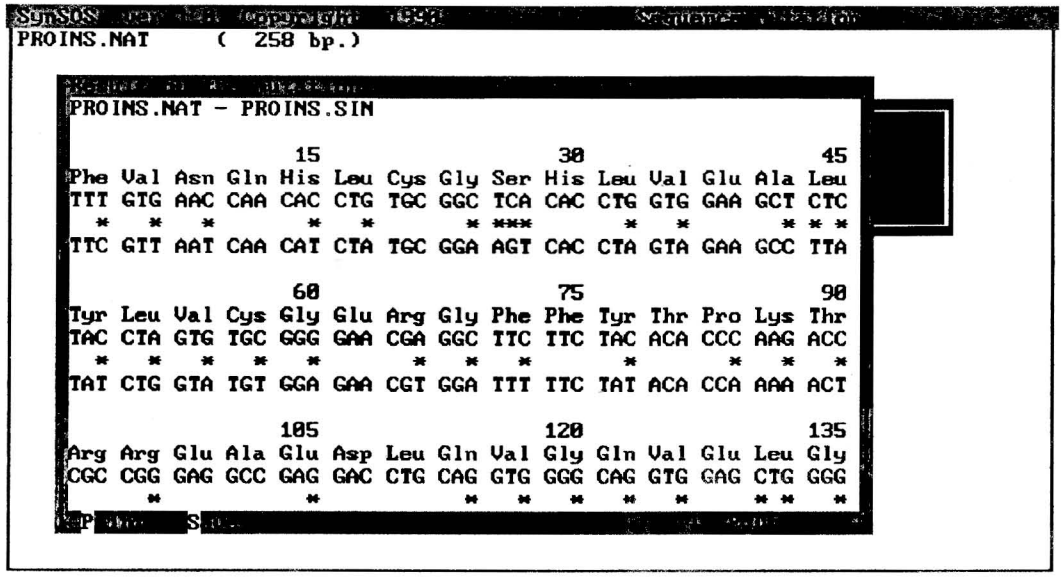


FIG. 2. Pantalla que muestra los resultados del programa. Fragmento del gen natural de la proinsulina humana y secuencia obtenida después de realizadas las sustituciones.

## RESULTADOS Y DISCUSION

El algoritmo presentado para la eliminación de subsecuencias repetidas no analiza todas las posibles combinaciones de codones que generen la secuencia óptima de menor cantidad de repetidas. Si consideramos que como promedio se pueden emplear tres codones sinónimos en cada sustitución, para una proteína de 200 aminoácidos, la cantidad de cambios posibles asciende aproximadamente a $2{,}7 \times 10^{95}$. Evaluar todas estas posibles combinaciones resulta impracticable en tiempo, en las microcomputadoras actuales. No obstante, el programa descrito garantiza que la secuencia resultante no presente repetidas de mayor longitud que la secuencia fuente.

En la actualidad se trabaja en la introducción de nuevas condiciones a tener en cuenta en el programa de eliminación de repetidas, tales como evitar la introducción de sitios de restricción indeseables o la eliminación de otros de interés como resultado de las sustituciones de codones, así como evitar la abundancia de citosinas y guaninas en la secuencia nucleotídica final.

El programa ha sido utilizado en la síntesis químico-enzimática de varios genes en el CIGB, entre ellos, el gen de la proinsulina humana (Jiménez *et al.*,1990), en el cual se obtuvo una alta disminución de la cantidad de subsecuencias repetidas gracias a la sustitución de 67 bases de la secuencia nucleotídica original.

## REFERENCIAS

BRINGAS, R. y S. PEREZ (1986). CIBSOFT: Un paquete de programas para el análisis de ácidos nucleicos y proteínas. *Interferón y Biotecnología* 3(3): 225-228.

JIMENEZ, V.; R. GUIMIL; R. UBIETA; M. OCHAGAVIA; A. SILVA; G. VILLEGAS y L. HERRERA (1990). Síntesis químico-enzimática del gen de la proinsulina humana. *Biotecnología Aplicada* 7(2): 142-152.

MARTINEZ, H. (1983). An efficient method for finding repeats in molecular sequences. *Nucleic Acids Res.* 11: 4629.