Biotecnología Aplicada, Vol. 9, No. 3, 174-179 (1992)

PROFALIGN: una representación gráfica del alineamiento de dos secuencias biológicas

M. OCHAGAVIA, R. RICARDO, J. FERNÁNDEZ DE COSSÍO y R. BRINGAS

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

Recibido en enero de 1992
Aprobado en marzo 1992

RESUMEN

Se desarrolló un programa que calcula y muestra de forma gráfica en la pantalla de una computadora, un perfil del alineamiento de dos secuencias biológicas.

El perfil se obtiene evaluando en cada posición del alineamiento una función que tiene en cuenta el grado de similitud de las secuencias alineadas en las posiciones adyacentes. En el cálculo de la similitud se emplean la matriz de pesos y el valor de penalización por inclusión de un elemento nulo, utilizados en el alineamiento del par de secuencias.

El programa facilita la localización de zonas conservadas y permite conocer el grado de similitud de las secuencias en cualquier región de interés.

ABSTRACT

A computer program to calculate and graphically show the alignment profile of two biological sequences is described.

The program produces an alignment profile which is calculated using a previous two sequences alignment, a weight matrix and a gap penalty. The function which describes the profile is evaluated for each position taking into account the similarity score of its neighbour positions.

This program is useful to find the conserved regions and to evaluate the similarity level of the two sequences in every region.

INTRODUCCION

El alineamiento de secuencias biológicas [secuencias de ácidos desoxirribonucleicos (ADN), ácidos ribonucleicos (ARN) y proteínas], es uno de los procedimientos más usados para detectar regiones conservadas (regiones comunes a una o más secuencias biológicas que poseen determinados elementos funcionales o estructurales similares), en el proceso de evolución de las especies.

En términos generales, alinear dos secuencias consiste en colocar una debajo de la otra, insertando convenientemente en cada una, elementos nulos (Ø), como por ejemplo:

øATAAGCø
AAAAAøCG

De modo tal que alguna función de similitud sea "maximizada" (Needleman y Wunsch, 1970), o lo que es equivalente, que una determinada distancia evolutiva sea minimizada (Sellers, 1974).

En las últimas décadas se ha desarrollado una gran cantidad de programas de computación y métodos matemáticos de alineamiento, algunos de

los cuales han sido compilados en libros dedicados al análisis de secuencias de ácidos nucleicos y proteínas (Waterman, 1989; Kruskal *et al.*, 1983).

La forma usual en que estos programas presentan sus resultados es la siguiente: colocan las secuencias una debajo de otra y en una tercera fila ubican asteriscos para resaltar los elementos idénticos en cada columna. En ocasiones se usa, además, otro símbolo para destacar aquellos elementos que aunque no son iguales, tienen determinada similitud.

A partir de estos resultados, la detección de las regiones de mayor similitud es responsabilidad del usuario, el cual debe realizar una inspección visual minuciosa, en aras de localizar las zonas de mayor concentración de asteriscos.

En nuestro instituto se desarrolló un nuevo método de alineamiento de secuencias (Labarta *et al.*, 1992). Durante la puesta a punto de dicho método, fue necesario evaluar los resultados de una gran cantidad de alineamientos de pares de secuencias. Con el objetivo de facilitar la inspección visual de los resultados, desarrollamos un programa que calcula y muestra de forma gráfica en la pantalla de la computadora un perfil del alineamiento de dos secuencias.

El perfil se obtiene evaluando en cada posición del alineamiento una función que tiene en cuenta el grado de similitud de las secuencias alineadas en las posiciones adyacentes. En el cálculo de la similitud se emplean la matriz de pesos y el valor de penalización por inclusión de un elemento nulo, utilizados en el alineamiento del par de secuencias.

MATERIALES Y METODOS

Cálculo del perfil del alineamiento

Sean $\underline{a} = a_1,a_2,...,a_n$ y $\underline{b} = b_1,b_2,...,b_m$, dos secuencias biológicas de igual procedencia (ADN, ARN o proteínas) de longitudes n y m respectivamente, cuyos elementos pertenecen a un conjunto X (X es igual a {A,C,G,T}, a {A,C,G,U} o a {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}, dependiendo de que **a** y **b** provengan de ADN, ARN o proteínas, respectivamente).

Sea $\underline{\mathbf{M}}$ una matriz de pesos, en la cual cada elemento $m_{p,q}$ (p,q ∈ X, $m_{p,q} = m_{q,p}$), es un número entero positivo que representa el peso o costo de sustituir el elemento **p** por el elemento **q**.

Si el par $(a^*,b^*) = (a^*1,a^*2,...,a^*L,b^*1,b^*2,...,b^*L)$ (máx{m,n} ≤ L ≤ m + n, $a^*i,b^*j \in X + \emptyset$), es un alineamiento de **a** y **b** y extendemos la definición de **M** de modo que incluya:

$m_{p,\emptyset} = m_{\emptyset,p} = g$ ($g \leq 0$, para todo p, $p \in X$)

podemos calcular una función P(x) ($1 \leq x \leq L$), que llamaremos perfil del alineamiento (a^*,b^*), de la forma siguiente:

$P(x) = S - (V(x-[w/2]) - R(x-[/2]))$,
si $[w/2]+1 \leq x \leq L-w+[w/2]+1$,

$P(x) = P([w/2]+1)$,
si $1 \leq x < [w/2]+1$

$P(x) = P(L-w+[w/2]+1)$,
si $L-w+[w/2]+1 < x \leq L$,

donde:

w es una constante ($1 \leq w \leq L$),

$S = w^* \max\{m_{p,q}\}$,

$V(i) = \sum_{k=i}^{i+w-1} \max\{ m_{ak}{}^*{}_{,ak}{}^*, m_{bk}{}^*{}_{,bk}{}^*\}$,

$R(i) = \sum_{k=i}^{i+w-1} m_{ak}{}^*{}_{,bk}{}^*$.

Descripción del programa

El programa se escribió en Turbo Pascal v.6 para microcomputadoras IBM PC, XT, AT y compatibles que posean al menos una tarjeta gráfica EGA.

PROFALIGN puede ser ejecutado desde el sistema operativo o como parte del paquete de programas BioSOS (Bringas *et al.*, 1992).

La información de entrada es un fichero texto en formato ASCII (figura 1), similar al producido por el programa de alineamiento del BioSOS, el cual fue realizado a partir de un algoritmo que emplea un espacio lineal (Myers y Miller, 1988). Este programa utiliza la matriz unitaria (Doolittle, 1981) para alinear

secuencias de ADN. Las proteínas pueden ser alineadas usando, además de la matriz unitaria, las matrices de pesos conocidas por: Genetic Code Matrix (Sellers, 1974; Smith *et al.*, 1981), Structure-Genetic Matrix (Doolittle, 1979) o Dayhoff's Log-Odds Matrix (Dayhoff, 1972, 1978).

```
BioSOS v. 2.0.     Sequence Alignment
Sequences type     : PROTEINs
1st Sequence       : C:\BIO\SEQ\DSHUCZ
2nd Sequence       : C:\BIO\SEQ\SODL
Weight Matrix      : Dayhoff's Log-Odds Matrix
Gap cost           : 0
Indel cost         : 25
Start time         : 08:35:35
Elapsed time       : 00:00:01
Identity Score     : 68.21 %

MATKAVCVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTEGLHGFHVHEFGDNTAGCTS :  60
.  ******.* *   * . ***     .*     .**** * ****** ***** ** *
V_LKAVCVLRGAGETTGTVYFEQEGNANAVGKGIILKGLTPGEHGFHVHGFGDNTNGCIS :  59


AGPHFNPLSRKHGGPKDEERHVGDLGNVTADKDGVADVSIEDSVISLSGDHCIIGRTLVV : 120
******* *.**.*****.************ .*** . * *  ***.*   *****.*.
AGPHFNPASKKHAGPKDEDRHVGDLGNVTADANGVAKIDITDK_ISLTGPYSIIGRTMVI : 118


HEKADDLGKGGNEESTKTGNAGSRLACGVIGIAQ                          : 154
********.****** ********************  ..
HEKADDLGRGGNEESLKTGNAGSRLACGVIG_TE                          : 151
```

FIG. 1. Ejemplo de un fichero de entrada para FROFALIGN.

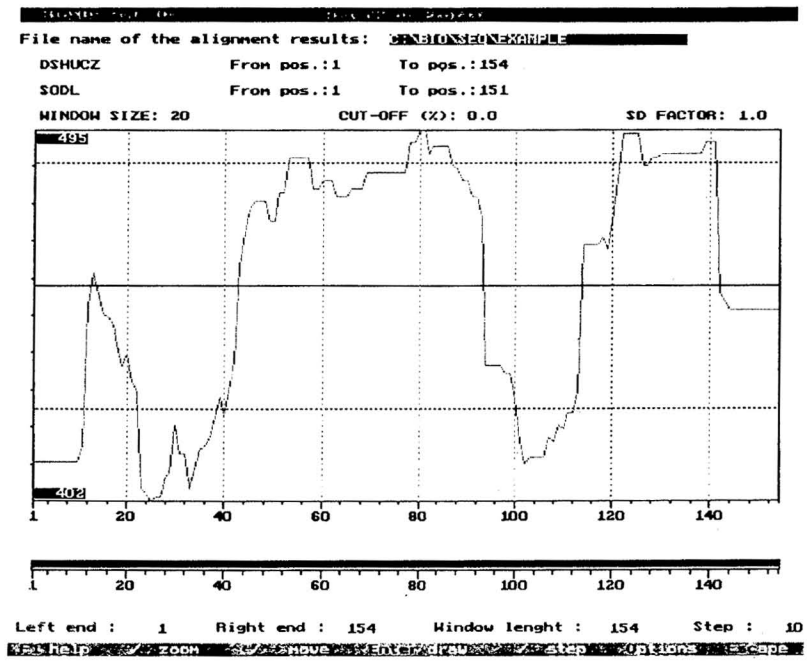

FIG. 2. Gráfico producido por PROFALIGN.

Descripción del gráfico

En la pantalla de la computadora se presenta un gráfico, tal y como se muestra en la figura 2.

En la parte superior de la pantalla se encuentran los nombres de las secuencias, seguidos por las posiciones iniciales y finales de cada secuencia de acuerdo con la región del alineamiento.visualizada en el gráfico.

La curva describe el valor que toma la función **P(x)** para cada posición del alineamiento (eje x), usando como valor de **w** el indicado en la pantalla como **Window size**. A este parámetro se le asigna un valor inicial igual a 20 y puede ser cambiado por el usuario.

La línea horizontal continua representa la media aritmética P de la función P evaluada en todas las posiciones.

$$\overline{P} = \frac{\sum_{k=1}^{L} P(k)}{L},$$

Las líneas horizontales discontinuas por encima y por debajo de la línea horizontal continua, representan dos valores D^+ y D^-, respectivamente, calculados de la forma siguiente:

$D^+ = P + c * DS$ y

$D^-, = P - c * DS$,

donde:

c es una constante, y

DS es la desviación estándar.

En el gráfico, **c** es llamada **SD Factor**. Este **SD Factor** se asume inicialmente igual a 1 y puede ser modificado por el usuario.

El programa ofrece la posibilidad de visualizar los valores que toma la función P en todas las posiciones del alineamiento o solamente aquellos que sean mayores o iguales que un valor de corte **VC** definido de la siguiente forma:

$VC = S * CO / 100$,

donde:

CO es una constante, y

S es el valor máximo que puede tomar la función P (ver el subepígrafe Cálculo del perfil).

En el gráfico, **CO** es llamado **Cut-Off**. Este parámetro se asume inicialmente igual a 0 y puede ser modificado por el usuario.

El parámetro **Cut-Off** ofrece la posibilidad de representar en el gráfico solamente las regiones más conservadas, eliminando las de bajo nivel de similitud (figura 3).

En el extremo inferior de la pantalla se presenta un menú (figura 2), que contiene las siguientes opciones:

←/→: Permite reducir o ampliar la región a visualizar.

Esta región se refleja en la barra horizontal.

^←/^→: Permite mover la barra horizontal.

Enter: Dibuja el perfil de la región definida por la barra horizontal.

+/-: Incrementar o disminuir el paso (en potencias de 10), para extender, reducir y mover la barra horizontal.

Options : Permite el acceso a otro menú (figura 3) que cuenta con las opciones siguientes:

Print: Imprimir el perfil.

Numeric info : Muestra en una ventana las regiones de cada secuencia para las cuales el valor del perfil es mayor o igual que **VC**.

Alignment: Muestra en una ventana el fichero de entrada del programa que contiene el alineamiento de las secuencias.

Window: Permite modificar el parámetro **Window size**.

Cut-Off: Permite modificar el parámetro **Cut-Off**.

SD Factor: Permite modificar el parámetro **SD Factor**.

Escape: Retorna al menú anterior.

Escape: Termina la ejecución del programa.

RESULTADOS Y DISCUSION

El programa fue utilizado en nuestro instituto para la evaluación de los alineamientos de más de 300 pares de secuencias producidos por dos métodos de alineamiento (Myers y Miller, 1988; Labarta *et al.*, 1992). Su empleo permitió evaluar de una forma más efectiva los resultados.

PROFALIGN puede ser usado como herramienta auxiliar de cualquier programa de alineamiento de dos secuencias biológicas. El único requerimiento es crear el fichero texto de entrada de la forma adecuada, para que pueda ser correctamente leído por nuestro programa.

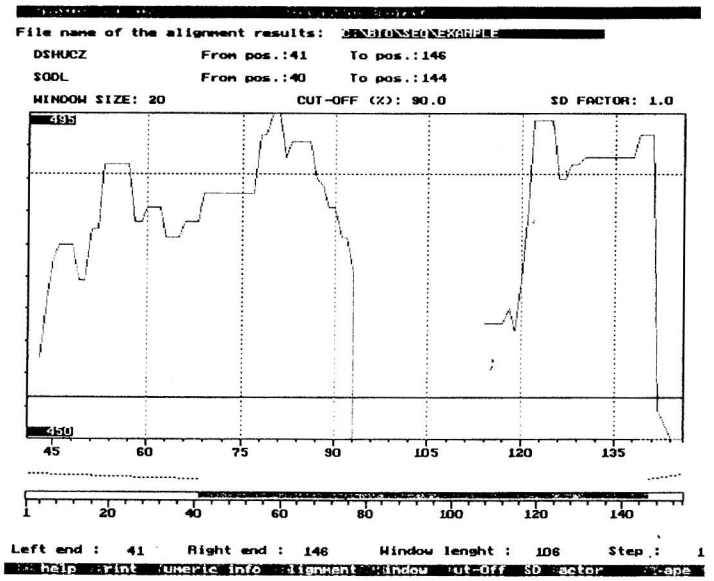

FIG. 3. Regiones conservadas encontradas por PROFALIGN, usando un valor de CUT-OFF igual a 90 %.

Su utilidad inmediata radica en las facilidades que brinda para la detección de regiones conservadas y para la evaluación del nivel de similitud de cualquier otra zona de interés.

Como utilidad derivada de la información que el usuario pueda extraer del gráfico, el programa podría contribuir a la selección del método de alineamiento que mejor se ajusta para un determinado par de secuencias, así como, de los valores más adecuados de los parámetros del programa de alineamiento.

En la actualidad se trabaja en una nueva versión que permita al usuario introducir la matriz de pesos que desee para el cálculo del perfil.

Futuras versiones de PROFALIGN podrían incluir otras funciones para el cálculo del perfil que tengan en cuenta, por ejemplo, información acerca de la estructura secundaria, para el caso de perfiles de alineamiento de proteínas.

REFERENCIAS

BRINGAS, R.; R. RICARDO; J. FERNANDEZ DE COSSIO; M. OCHAGAVIA; A. SUAREZ y R. RODRIGUEZ (1992). BioSOS: Un paquete de programas para el análisis de secuencias biológicas. *Biotecnología Aplicada* **9**:(2) (180-185).

DAYHOFF, M.O. (1972). "A model of evolutionary change in proteins. Detecting distant relationships: computer methods and results". In: Atlas of protein sequence and structure. Ed. M.O. Dayhoff. National Biomedical Research Foundation, Washington DC., **5**: 89-110.

DAYHOFF, M.O. (1978). "A model of evolutionary change in proteins. Matrices for detecting distant relationships". In: Atlas of Protein Sequence and Structure. Ed. M.O. Dayhoff. National Biomedical Research Foundation, Washington DC., **5**: 345-358.

DOOLITTLE, R.F. (1979). "Protein evolution". In: The proteins. Eds. H. Neurath, R.L. Hill. Academic Press, New York., **4**: 1-118.

DOOLITTLE, R.F. (1981). Similar amino acid sequences: chance or common ancestry?. *Science* **214**: 149-159.

KRUSKAL, J.B. y D. SANKOFF (Eds) (1983). *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison. Addison-Wesley, London.*

LABARTA, V.; M. OCHAGAVIA; R. RICARDO; J. FERNANDEZ DE COSSIO Y R. BRINGAS (1992). Alineamiento de secuencias previa selección de regiones homólogas. (En preparación).

MYERS, E.W. y W. MILLER (1988). Optimal alignments in linear space. *CABIOS* **4:** 11-17.

NEEDLEMAN, S.B. y C.D. WUNSCH (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J.Mol.Biol.* **48:** 443-453.

SELLERS, P.H. (1974). Evolutionary distances. *SIAM J. Appl. Math.* **26:** 787-793.

SMITH, T.F.; M.S. WATERMAN y W.M. FITCH (1981). Comparative biosequence metrics. *J. Mol. Evol.* **18:** 38-46.

WATERMAN, M.S. (1989). "Sequence alignments. In: Mathematical Methods for DNA sequences. Ed. M.S. Waterman. CRC Press, Inc., Boca Raton, Florida, pp. 53-92.