Comparación de secuencias

Juan J. Nieto
Departamento de Análisis Matemático
Facultad de Matemáticas
Universidad de Santiago de Compostela
Correo electrónico: amnieto@usc.es

Resumen: El objetivo de la comparación de secuencias es el análisis de la similitud entre éstas para encontrar evidencias de homología entre ellas. El interés del análisis de la similitud entre varias secuencias se debe a que las bases de datos proteicas habitualmente se organizan por familias de proteínas. Cada familia proteica está formada por proteínas con similar estructura, similar función, o similar historial evolutivo. Una nueva proteína será admitida en la familia si tiene simultáneamente similitud con los miembros actuales de la familia y no sólo con una de ellas.

I INTRODUCCIÓN

Una *secuencia* consiste en un conjunto ordenado de letras seleccionadas de un alfabeto. Como ejemplos de alfabetos, podemos citar:

1. Castellano (27 letras)
2. Gallego (23 letras)
3. Inglés (26 letras)
4. Aminoácidos (20 letras)
5. ADN (4 letras): a , c , g , t .

Cada secuencia forma una palabra. Por ejemplo:

1. XYZSECW
2. GALLEGO
3. GATA
4. CAT

pero pueden tener distintos significados dependiendo del alfabeto. Así, CAT significa gato (en inglés), pero significa histidina en el alfabeto ADN.

La *complejidad* de un alfabeto se define como el número de letras diferentes que contiene.

Dispondremos de secuencias que suelen ser EST (Expressed Sequences Tags) o Marcas de Secuencias Expresadas. Cómo se obtienen es un problema tecnológico y normalmente implica el uso de un sistema laser fluorescente que lee los geles de secuenciación. Este proceso suele estar automatizado y a pesar de la sofisticación, a menudo es incapaz de tomar una decisión sobre qué base ocupa una determinada posición en la secuencia. Cuando esto sucede, el software de asignación de bases inserta una base ambigua que se suele denotar por n (ó X). Estas imprecisiones son muy importantes pero dificultan el análisis posterior y trataremos con ellas más adelante. Así se tiene el alfabeto EST

$$\{ \text{a} , \text{g} , \text{t} , \text{c} , \text{n} \}$$

de complejidad 5.

El software de asignación automática de bases esencialmente lo que hace es mirar los picos fluorescentes de las cuatro reacciones de secuenciación en una calle de gel de secuenciación. Al objeto de aumentar las posibilidades de encontrar picos y, por tanto, de asignar bases con precisión, se supone que las bases se asignan a intervalos regulares. Pero si las propiedades físicas del gel, u otras circunstancias afectan al flujo, la asignación de bases puede no ser realmente regular y a veces las bases son asignadas demasiado pronto (**IN**SERCIÓN), o quedan sin asignarse (ELIMINADAS - **DEL**ETED) y puede aparecer una inserción o una eliminación fantasma (INDEL fantasma). Esto habrá que tenerlo en cuenta en las comparaciones que vayamos a hacer.

Cabe esperar que los criterios de control de calidad usuales en un buen laboratorio mantengan la proporción de n's inferior a un tanto por ciento de la longitud total de una secuencia (normalmente menor o igual a un 5%). El principio y el final de la secuencia se pueden recortar para reducir aún más la ambiguedad.

En la comparación de secuencias se permitir el uso de huecos. Por ello se introduce una nueva letra en el alfabeto: - (hueco, o en inglés *gap*).

Se puede obtener una secuencia de ácidos nucleicos o de proteinas en el laboratorio o bien recurriendo a las numerosas bases de datos de acceso público y vía INTERNET.

Por ejemplo, el *Instituto Bioinformático Europeo* (EBI) que forma parte del *Laboratorio de Biología Molecular Euopeo* (EMBL) dispone de diversas bases de datos entre las que cabe citar EMBL-Bank en su versión 78 de Marzo 2004 y que puede verse en la dirección

http://www.ebi.ac.uk/embl/

El reciente proyecto *Ensembl* que trata de organizar la información de las secuencias de genomas completos puede consultarse en

http://www.ensembl.org/

GenBank se encuentra en la página

http://www.ncbi.nlm.nih.gov/Genbank/

del Centro Nacional para Información Biotecnológica (NCBI).

Así, por ejemplo la bacteria *Mycobacterium Tuberculosis H37Rv* tiene el número de acceso BX 842583. Se puede obtener incluso el genoma completo, es decir, todo el material genético de los cromosomas del organismo. Tiene unos 4 411 000 pb, aproximadamente 4 000 genes y un alto contenido en guanina+citosina. Parte de su material genético se presenta a continuación:

1 atgtctatct gtgatccggc gctgcgtaat gcgctacgta ccctgaaact gtccggcatg
61 ctcgacaccc tcgacgcccg cctggcccaa acccgcaacg gcgacctggg gcatctggaa
121 ttcctgcaag cgttgcgtga agacgagatc gcccgccgcg agtccgccgc cctgacacga
181 cgattacgcc gcgccaagtt cgaagcccaa gccaccttcg aagacttcga cttcactgcc
241 aacccgaaac tgcccggtgc gatgttgcgc gatctggccg cgctgcgctg gctggatgcc
301 ggcgaatcgg tcatcctcca cggcccggtc ggcgtcggaa aaacccatgt agcacaagca
361 cttgtccacg ccgtggcccg ccgcggcggc gacgtgcgct tcgccaaaac ctcccgcatg
421 ctctccgacc tcgccggcgg gcacgccgac cgatcctggg gccaacgcat ccgcgaatac
481 accaagccgc tcgtgctcat tctggacgac ttcgcgatgc gtgagcacac cgccatgcac
541 gctgatgacc tctacgagct catcagcgac cgcgccatca ctggcaaacc gctgatcttg
601 accagcaacc gcgcaccgaa taactggtac ggcctgttcc ccaaccccgt cgtcgccgaa
661 tcactcctgg atcggctcat caacaccagc caccaaatcc tcatggacgg acccagctac
721 cgaccccgca agagacccgg ccgcaccacc agctag

con su correspondiente secuencia de aminoácidos

MSICDPALRNALRTLKLSGMLDTLDARLAQTRNGDLGHLEF
LQALREDEIARRESAALTRRLRRAKFEAQATFEDFDFTANPK
LPGAMLRDLAALRWLDAGESVILHGPVGVGKTHVAQALVHA
VARRGGDVRFAKTSRMLSDLAGGHADRSWGQRIREYTKPLV
LILDDFAMREHTAMHADDLYELISDRAITGKPLILTSNRAPNN
WYGLFPNPVVAESLLDRLINTSHQILMDGPSYRPRKRPGRTTS

Una vez que tenemos una secuencia, hay que analizarla y compararla con otras para:

1. Identificar los genes.
2. Determinar las funciones de dichos genes. Una manera es encontrar un gen (posiblemente de otro organismo) cuya función se conozca y tal que ambos genes (el conocido y el que estamos estudiando) tengan una cierta similitud en sus secuencias. Aquí topamos con el principal concepto y la cuestión de cómo comparar secuencias.
3. Identificar las proteinas que intervienen en la regulación de la expresión génica.
4. Determinar repeticiones o patrones en las secuencias.

5. Identificar otras regiones funcionales de la secuencia.

Nótese que implicitamente estamos suponiendo que cierta similitud de secuencias implica cierta similitud funcional, pero hay que tener presente que eso no siempre es cierto.

Todas las tareas que se acaban de señalar son *COMPUTACIONALES.* Dado que se están produciendo continuamente secuencias, para analizar toda esa ingente cantidad de datos, está claro que es necesario abordarlo desde una perspectiva interdisciplinar de las Ciencias de la Computación, la Inteligencia Artificial, la Informática, la Biología y las Matemáticas.

II ANÁLISIS DE UNA SECUENCIA

Aunque lo más relevante, o al menos uno de los aspectos más impotantes, es la comparación de secuencias, el análisis de una sola secuencia tambien puede aportar mucha información. Como ejemplo, se presentan a continuación las frecuencias correspondientes al *M. Tuberculosis* :

$$\begin{gathered} \mathrm{a} : 0.1693 \approx 17\ \% \\ \mathrm{c} : 0.3232 \approx 32\ \% \\ \mathrm{g} : 0.3304 \approx 33\ \% \\ \mathrm{t} : 0.1771 \approx 18\ \% \end{gathered}$$

Para un *humano*, tenemos las siguientes cifras:

$$\mathrm{a} \approx 31\ \%\ ,\ \mathrm{c} \approx 20\ \%\ ,\ \mathrm{g} \approx 20\ \%\ ,\ \mathrm{t} \approx 29\ \%.$$

El ADN es una molécula de doble cadena. En la naturaleza los pares de bases sólo se forman entre a y t y entre g y c. Por ello, para calcular las frecuencias sólo hemos usado una cadena de ADN ya que si secuenciamos ambas y debido a la complementariedad de las bases, las frecuencias habrían sido iguales, más que similirares.

Puede observarse que la proporción de a es similar a la proporción de t y que la de g es muy similar a la de c. Si se consideran las bases púricas (a , g), se puede ver que es muy similar a la de bases pirimidínicas (c , t). Así se tienen las reglas de Chargaff:

$$\begin{gathered} \mathrm{a} \equiv \mathrm{t}\ , \\ \mathrm{g} \equiv \mathrm{c}\ , \\ \mathrm{a} + \mathrm{g} \equiv \mathrm{c} + \mathrm{t}\ . \end{gathered}$$

La asimetría es el cociente de a + t entre g + c . Para el *M. Tuberculosis* vale 0.54 y en el caso de un *humano* vale 1.50 . Como dato interesante, señalamos que para el *cangrego de mar* vale 15.67 .

En la tabla siguiente se presentan el número total de nucleótidos en las tres bases (sólo en las regiones codificantes) del *M. Tuberculosis.* Los detalles pueden verse en los artículos [6,7]:

	t	c	a	g
primera base	216051	409011	228244	470868
segunda base	269638	416457	233472	404607
tercera base	217803	458256	210892	437223

Se tiene que en la primera base hay un total de 216 051 t de un total de 1 324 174 bases. Por tanto, la fracción de t en la primera base es de

$$\frac{216051}{1324174} = 0.1632 = 16.32\%.$$

De esta manera se tiene la siguiente tabla de frecuencias:

	t	c	a	g
primera base	0.1632	0.3089	0.1724	0.3556
segunda base	0.2036	0.3145	0.1763	0.3056
tercera base	0.1645	0.3461	0.1593	0.3302

Esta última tabla se puede pensar como una matriz 3x4 y proporciona una interesante información sobre el organismo considerado. Más adelante veremos como comparar dos organismos usando esta información.

Otro concepto muy interesante es el de entropía de una secuencia. Con el alfabeto ADN, si las cuatro letras fuesen igualmente probables, se tendría la distribución uniforme y cada letra tendría una probabilidad de 0.25. Curiosamente, esa es la distribución para, por ejemplo, el *Aspergillus*. La fórmula matemática que nos da el valor de la entropía, suponiendo que las bases que aparecen en distintas posiciones son independientes, es

$$\sum_i f_i \log_2(\frac{f_i}{u_i})$$

donde i varía en las letras del alfabeto, f_i es la frecuencia corrspondiente y $u_i = 0.25$, la frecuencia de la distribución uniforme. Se tiene, por ejemplo, $f_a = 0.1693$ para el *M. Tuberculosis*. De esta forma, la entropía del *M. Tuberculosis* es igual a 0.0693 . En términos de información significa que en una secuencia de longitud 100, hay 6.93 bits de información extra. Asimismo, si se toma una secuencia con una distribución similar a la del organismo en cuestión, entonces la probabilidad de que dicha secuencia provenga del *M. Tuberculosis* es $2^{6.93} = 121.9$ veces más de que haya sido generada aleatoriamente a partir de una distribución uniforme.

III COMPARACIÓN DE SECUENCIAS

El proceso de comparación de dos o más secuencias es consustancial a la bioinformática. Antes de nada, recordaremos un ejemplo médico real.

La esclerosis múltiple se caracteriza por una afectación multifocal de la mielina, principal componente que recubre los axones nerviosos del Sistema Nervioso Central (SNC), produciendo alteraciones neurológicas crónicas, generalmente progresivas y de evolución imprevisible.

Se conjeturó que algunas proteinas del SNC debian ser similares a ciertas proteinas de bacterianas presentes en una infección anterior. A fin de corroborar dicha hipotesis, se llevaron a cabo los siguientes pasos:

- Fueron secuenciadas las proteinas de las vainas de mielina
- Se buscaron en bases de datos de proteinas secuencias de bacterias y virus similares
- Se hicieron experimentos en laboratorios para ver si los linfocitos T atacaban esas proteinas.

El resultado final fue la identificación de ciertas proteinas de virus y bacterias que el sistema inmunológico confundía con las de la proteinas de las vainas de mielina.

Aunque la etiología de esta enfermedad continua siendo desconocida, tanto las características epidemiológicas como las patogénicas sugieren la implicación de agentes infecciosos e inmunes. Actualmente se sabe que ciertas bacterias y virus contienen proteinas con estructura similar a las que existen en el SNC. Por ello, se postula que la infección por estos microorganismos activaría el sistema inmunológico de forma que éste destruiría no sólo a los invasores, sino tambien a sustancias análogas (mielina) del propio organismo.

Por tanto, si tenemos una nueva secuencia, la podemos comparar con otras ya existentes para ver si son similares y pueden tener las mismas funciones. Si no sabemos nada sobre la nueva secuencia, se puede iniciar una busqueda de secuencias análogas para saber, o al menos intuir, de qué se trata. Tambien se puede hacer alguna conjetura sobre la funcionalidad de la nueva secuencia

y buscar secuencias que compartan esas propiedades. La similitud entre secuencias puede servir para predecir la estructura 3-dimesional de una secuencia de proteinas.

El BLAST (Basic Local Alignment Search Tool) es un conjunto de programas de búsqueda de similitud que exploran bases de datos de secuencias. Por ejemplo el **blastn** compara una secuencia de nucleótidos contra una base de datos de secuencias de nucleótidos.

Uno de los métodos más básicos para comparar secuencias es el *gráfico de puntos* (dotplot). Dadas dos secuencias se construye una matriz poniendo una letra, por ejemplo una x, si los residuos coinciden, y nada en caso contrario. Es fácil de visualizar. Nótese que en un gráfico de este tipo, dos secuencias identicas se caracterizan por una linea diagonal. Por contra, dos secuencias semejantes se caracterizan por una diagonal rota, indicando la región interrumpida la localización de desemparejamientos de las secuencias. Un ejemplo de gráfico de puntos se presenta a continuación:

	g	c	t	g	a	a	c	g
c		x					x	
t			x					
a					x	x		
t			x					
a					x	x		
a					x	x		
t			x					
c		x					x	

IV ALINEAMIENTO DE SECUENCIAS

El alineamiento de secuencias consiste en establecer un segmento entre ellas donde el número de coincidencias sea máximo. El alineamiento puede ser simple (dos secuencias) o múltiple (más de dos secuencias), y global o local.

Para centrar las ideas, vamos a considerar dos secuencias:

Primera secuencia: g c t g a a c g

Segunda secuencia: c t a t a a t c

que como puede observarse tienen dos coincidencias (en la quinta y sexta posiciones). Si se permite la introduccin de huecos, pero sin que coincidan dos huecos, otros posibles alineamientos serian:

-	-	-	-	-	-	-	-	g	c	t	g	a	a	c	g
c	t	a	t	a	a	t	c	-	-	-	-	-	-	-	-

g	c	t	g	a	-	a	-	-	c	g
-	-	c	t	-	a	t	a	a	t	c

-	-	-	-	g	c	t	g	a	a	c	g
c	t	a	t	a	a	t	c	-	-	-	-

g	c	t	g	-	a	a	-	c	g
-	c	t	a	t	a	a	t	c	-

De estos cuatro alineamientos, claramente el primero es muy malo (0 acoincidencias y además se han introducido muchos huecos), el segundo es malo (sólo se han introducido tres huecos pero el número de coincidencias es 0), y el tercero no es bueno (hay una coincidencia, pero se han introducido 4 huecos), y cuando menos son peores que el alineamiento original que nos daba dos coincidencias sin hacer nada. Por último, el cuarto sólo tiene dos huecos y 5 coincidencias, por lo que parace bastante bueno. ¿Será el óptimo? Puede parecer que un alineamiento es mejor (en nuestro ejemplo el cuarto) que otro, pero ¿qué será mejor? ¿Un alineamiento sin huecos y con menos coincidencias, u otro con más coincidencias pero en el que hemos necesitado introducir más modificaciones?

¿Cuántas formas hay de comparar dos secuencias? En el ejemplo que estamos considerando se pueden considerar todos los 265 729 posibles alineamientos e incluso decidir cuál es mejor con respecto a un criterio fijado. Teoricamente, se puede decidir por medio de un sencillo algoritmo:

1. Considerar todos los alineamientos posibles
2. Determinar un valor para cada alineamiento (por ejemplo el número de coincidencias; o dar +3 puntos por cada coincidencia, -2 por cada no coincidencia y -1 por la introducción de un hueco)
3. Guardar el valor máximo

Sin embargo, el tiempo necesario para este algoritmo crece de una forma exagerada. Una secuencia, moderada en longitud, contiene 200-500 pares de base. Una proteina puede ser del orden de 200–400 aminoácidos (sobre 600–1200 letras). Nótese que para secuencias con longitudes superiores a 107, el número de posibles alineamientos es superior a 10^{80} , número estimativo de la cantidad de protones del universo [5].

Para poder encontrar un alineamiento global en un tiempo razonable, se han desarrollado numerosos algoritmos. El más famoso de ellos quizás sea el ya clásico algoritmo de Needleman & Wunsch que utiliza Programación Dinámica. Esencialmente consiste en:

1. Se define una función de similitud $s_{i,j}$ entre los elementos a alinear
2. Los indels se penalizan con un cierto peso w
3. Se construye una matriz $(a_{i,j})$ de una fila y una columna más que las secuencias
4. Se inicializa la primera fila y la primera columna con ceros
5. Se va llenando la matriz, dónde los valores dependen de los inmediatamente anteriores:

$$a_{ij} = \max\{a_{i-1,j-1} + s_{i,j}\ ,\ a_{i-1,j} - w\ ,\ a_{i,j-1} - w\}$$

6. Una vez completa la matriz, se recupera la solución

Vamos a volver al ejemplo. Cada coincidencia se valora con +1 punto (de manera que $s_{i,j} = 1$) y cada inserción o eliminación, es decir cada hueco no se penaliza ($w = 0$), de manera que sólo valoramos las coincidencias. En la primera fila colocamos la primera secuencia y en la primera columna, la segunda secuencia. El algoritmo nos lleva a la siguiente matriz:

		g	c	t	g	a	a	c	g
	0	0	0	0	0	0	0	0	0
c	0	0	1	1	1	1	1	1	1
t	0	0	1	2	2	2	2	2	2
a	0	0	1	2	2	3	3	3	3
t	0	0	1	2	3	3	3	3	3
a	0	0	1	2	3	4	4	4	4
a	0	0	1	2	3	4	5	5	5
t	0	0	1	2	3	4	5	5	5
c	0	0	1	2	3	4	5	x	y

Para calcular una entrada $a_{i,j}$ se consideran los valores que están inmediatamente encima $a_{i-1,j}$, a su izquierda $a_{i-1,j}$ y en la diagonal superior izquierda $a_{i-1,j-1}$. Se mira si coinciden las letras (+1 punto) o no (0 puntos) y se aplica la regla del algoritmo. Así, vamos a calcular los dos valores que faltan en la matriz. Si $x = a_{i,j}$ entonces $a_{i-1,j-1} = a_{i-1,j} = a_{i,j-1} = 5$. Dado que las letras (c y c) coinciden, se tiene que

$$x = a_{i,j} = \max\{5+1, 5, 5\} = 6\,.$$

Igualmente, se puede ver que $y = 6$.

Ahora se elige un camino: se toma la última entrada (en este ejemplo, la que tiene el valor y) y se comienza a buscar un camino que maximice la función y que nos dará un alineamiento óptimo con respecto a los parámetros elegidos. El algoritmo recorre los vecinos de la entrada (arriba, izquierda y diagonal superior izquierda) y se selecciona el que presente el valor más alto (si hay un empate, es posible obtener diferentes caminos y por tanto diferentes alineamientos para las mismas secuencias). En nuestro caso, un posible camino es el que se presenta a continuación (señalado con *):

		g	c	t	g	a	a	c	g
	0*	0*	0	0	0	0	0	0	0
c	0	0	1*	1	1	1	1	1	1
t	0	0	1	2*	2	2	2	2	2
a	0	0	1	2	2*	3	3	3	3
t	0	0	1	2	3*	3	3	3	3
a	0	0	1	2	3	4*	4	4	4
a	0	0	1	2	3	4	5*	5	5
t	0	0	1	2	3	4	5*	5	5
c	0	0	1	2	3	4	5	6*	6*

Empezando por la esquina superior izquierda y hasta llegar a la esquina inferior derecha, si vamos hacia la derecha ($\rightarrow$), se incluye una letra de la primera secuencia y un hueco en la segunda secuencia; si vamos hacia abajo ($\downarrow$), se pone un hueco en la primera y la letra correspondiente de la segunda; si se va en diagonal ($\searrow$), se ponen las letras correspondientes de ambas secuencias. De esta manera se obtiene el alineamiento

g	c	t	g	-	a	a	-	c	g
-	c	t	a	t	a	a	t	c	-

que, con nuestros parámetros, y según el algoritmo de N&W es óptimo (5 coincidencias).

Hay que notar que el algoritmo de N&W es global. Un algoritmo local es, por ejemplo, el de Smith-Waterman que puede consultarse en los libros de Lesk (2002) y de Attwood y Parry-Smith (2002).

No podemos terminar sin mencionar que hoy en día prácticamnete todo análisis de secuencias involucra un alineamiento múltiple de aquellas secuencias que presumiblemente son homólogas a la secuencia problema. Algunos programas bien conocidos para el alinemaiento múltiple son CLUSTAL-V y CLUSTAL-W. A pesar de que el alineamiento múltiple se puede ver como una generalización del alineamineto de dos secuencias, la complejidad crece exponencialmente con el número de secuencias por lo que hay que recurrir a nuevos métodos y todos ellos requieren una gran demanda computacional.

V COMPARACIÓN DE GENOMAS

La comparación de genomas completos no es una cuestión trivial ya que cada organismo posee un número diferente de bases. Como ya dijimos, el del *M. Tuberculosis* tiene unos 4 millones de bases y el de un *humano* unos 3 000 millones.

La comparación se ha de hacer con una medida cuantitativa de su similitud o su diferencia: A mayor similitud, menor diferencia y, reciprocamente, a menor similitud, mayor diferencia. Vamos a comparar las frecuencias correspondientes al *M. Tuberculosis* con las del *Escherichia coli* que se detallan a continuación (ver [6]):

	t	c	a	g
primera base	0.1605	0.2420	0.2600	0.3374
segunda base	0.3116	0.2286	0.2846	0.1752
tercera base	0.2619	0.2568	0.1831	0.2981

Para ello lo más adecuado es recurrir al concepto matemático de *distancia* o *métrica*. Recientemente se ha introducido una nueva métrica (ver la demostración en el artículo [4]) para la comparación de genomas completos:

$$d(A,B) = \frac{\sum_{i,j} |a_{i,j} - b_{i,j}|}{\sum_{i,j} \max\{a_{i,j}, b_{i,j}\}},$$

donde $a_{i,j}$ es la entrada ij de la matriz de frecuencias correspondientes. Por ejemplo, para los dos organismos citados, A = *M. Tuberculosis* y B = *E. Coli*,

$a_{11} = 0.1632$, $b_{24} = 0.1752$, $|a_{11} - b_{11}| = 0.0027$, $\max\{a_{32}, b_{32}\} = 0.3461$,

de forma que

$$d(M.Tuberculosis, E.Coli) = d(A,B) = \frac{0.8516}{3.4253} = 0.2483\,.$$

Se puede hacer un análisis similar con los dinucleótidos.

REFERENCIAS:

[1] T.K. Attwood y D.J. Parry-Smith, *Introducción a la Bioinformática.* Prentice Hall, Madrid, 2002.

[2] P. García Barreno, *Cincuenta años de ADN. La doble hélice.* Espasa Forum, Madrid, 2003.

[3] A.M. Lesk, *Introduction to Bioinformatics.* Oxford University Press, Oxford, 2002.

[4] J.J. Nieto, A. Torres, y M.M. Vázquez-Trasande, A metric space to study differences between polynucleotides, *Applied Mathematics Letters* Vol. **16** (2003), pp. 1289-1294.

[5] A.Torres, A. Cabada y J.J. Nieto, An exact formula for the number of alignments between two DNA sequences. *DNA Sequence* Vol. **14** (2003), pp. 427-430.

[6] A. Torres y J.J. Nieto, The fuzzy polynucleotide space: Basic properties. *Bioinformatics* Vol. **19** (2003), pp. 587-592.

[7] J.J. Nieto, A. Torres, D.N. Georgiou y T. Karakasidis, Fuzzy polynucleotide spaces and metrics. *Bulletin of Mathematical Biology* (2005) (en prensa).