UNIVERSITAS SCIENTIARUM
Revista de la Facultad de Ciencias
PONTIFICIA UNIVERSIDAD JAVERIANA

Vol. 8, 7-12

UNA NOTA SOBRE EL ESPACIO NULA

M. Alvarado[1], F. Novoa[2] y L. Rodríguez[3]

Departamento de Matemáticas, Pontificia Universidad Javeriana
1 e-mail: malvara@javeriana.edu.co
[2] e-mail: fernando.novoa@javeriana.edu.co
[3] e-mail: lrodrigu@javeriana.edu.co

RESUMEN

El número nula fue definido por L. Lareo y O. Acevedo para representar la información que las cadenas de nucleótidos contienen. Se desea determinar algunas propiedades algebraicas sobre el conjunto nula, dotando a éste con una operación y estudiando las propiedades que esta operación satisface.

Palabras clave: Número nula, ensamble, homomorfismo.

ABSTRACT

The nula number was introduced by L. Lareo and O. Acevedo in order to represent the information that sequences of nucleotides carry on. We want to identify some algebraic properties on the nula set, providing it with an operation and locking for the properties that this operation holds.

Key words: Nula number, assembling, homomorphism

INTRODUCCIÓN

El conjunto nula simbolizado por *N* está conformado por las ternas ordenadas de números reales correspondientes a la representación en tres dimensiones de la información que contiene una secuencia de nucleótidos (Adenina, Citosina, Guanina y Timina) en la forma como fue definida por Lareo y Acevedo (1999). Esta representación se obtiene asignando inicialmente a cada nucleótido un número primo distinto. Es usual hacer la siguiente asignación:

$$A \mapsto 2$$
$$C \mapsto 3$$
$$G \mapsto 5$$
$$T \mapsto 7$$

en donde A representa Adenina, C representa Citosina y así sucesivamente. Sea entonces *K* una cadena de nucleótidos de largo *n*. Se enumeran sus componentes de izquierda a derecha.

$$K = k_1 ... k_n$$

en donde cada $k_i \in \{A, C, G, T\}$ donde el largo de la cadena es $L(K) = n$. Adicionalmente se definen las siguientes cantidades:

a = número de veces que aparece *A* en *K*
b = número de veces que aparece C en *K*
c = número de veces que aparece G en *K*
d = número de veces que aparece T en *K*
α = suma de las posiciones del nucleótido A en K
β = suma de las posiciones del nucleótido C en K
γ = suma de las posiciones del nucleótido G en K
δ = suma de las posiciones del nucleótido T en K

Con esta visión y partiendo del hecho que una secuencia *K* de nucleótidos está totalmente determinada por la posición de cada nucleótido dentro de la secuencia, se definió el número nula de una secuencia *K* por medio de la terna ordenada

$$\eta(K)=\left(2^{\alpha}3^{\beta}5^{\gamma}7^{\delta},2^{a}3^{b}5^{c}7^{d},L(K)\right)$$

la cual se denomina el número nula de la secuencia *K*. Es importante notar que los números primos elegidos arbitrariamente pueden ser cuatro. Esta representación sin embargo presenta una desventaja y es que dos secuencias distintas pueden tener el mismo número nula.

Ejemplo 1. Para la cadena *K* = *CCTAAG* se tiene que

$$a=2 \quad b=2 \quad c=1 \quad d=1$$
$$\alpha=9 \quad \beta=3 \quad \gamma=6 \quad \delta=3$$

y por lo tanto

$$\eta(K)=(2^{9}3^{3}5^{6}7^{3},2^{2}3^{2}5^{1}7^{1},6)$$

se observa claramente que:

$$L(K)=n=a+b+c+d$$

$$\alpha+\beta+\gamma+\delta=\frac{n(n+1)}{2}$$

Ejemplo 2. Las secuencias

$K_1=ATTA$ y $K_2=TAAT$ tienen el mismo número nula

$$\eta(K_i)=\left(2^{5}7^{5},2^{2}7^{2},4\right) \text{ para } i=1,2.$$

Ejemplo 3. En el ejemplo 1 se tiene $\eta(K)=(2^{9}3^{3}5^{6}7^{3},2^{2}3^{2}5^{1}7^{1},6)$ para la secuencia *K* = *CCTAAG* ¿Cuántas secuencias tienen el mismo número nula? Es fácil comprobar que esta es la única secuencia que tiene ese número nula, por cuanto hay sólo una G y una T y además la suma de las posiciones de las C es 3 y la única forma de sumar esto con dos enteros positivos es 3=1+2.

En general, dado un número nula, el número de secuencias que corresponden a ese número nula es mayor que 1, por lo tanto surge la pregunta de cómo poder determinar esas secuencias, o al menos estimar su tamaño.

De esta forma, el conjunto nula *N* se define como el subconjunto de $\Re^3$ formado por las ternas (*x*,*y*,*z*) tales que existe una secuencia *K* con $\eta(K)=(x,y,z)$.

Operación ensamble

Se define sobre el conjunto nula una operación binaria * de la siguiente manera:

$$(x,y,z)*(e,d,f)=(xde^{z},ye,z+f)$$

Sean K_1 y K_2 dos secuencias tales que

$$\eta(K_1)=(x,y,z) \text{ y } \eta(K_2)=(d,e,f)$$

Entonces la secuencia $K=K_1K_2$ tendrá como largo

$$L(K)=L(K_1)+L(K_2)=z+f$$

y el número de veces que aparece cada nucleótido en *K* es exactamente la suma del número de veces que aparece ese nucleótido en las secuencias K_1 y K_2, por lo tanto si

$$y=2^{a_1}3^{b_1}5^{c_1}7^{d_1}$$
$$e=2^{a_2}3^{b_2}5^{c_2}7^{d_2}$$

entonces la segunda coordenada de $\eta(K)$ será

$$2^{a_1+a_2}3^{b_1+b_2}5^{c_1+c_2}7^{d_1+d_2}=ye$$

Finalmente, si un nucleótido aparece en K_2 en la posición $_s$, entonces ese nucleótido

aparecerá en K en la posición $L(K_1)+s=z+s$. Teniendo en cuenta esto, la suma de las posiciones en K del nucleótido A será entonces $\alpha_1+\alpha_2+za_2$, y así sucesivamente para los otros tres nucleótidos, por lo tanto la primera coordenada del número nula de K será:

$$2^{\alpha_1+\alpha_2+za_2}3^{\beta_1+\beta_2+zb_2}5^{\gamma_1+\gamma_2+zc_2}7^{\delta_1+\delta_2+zd_2}:$$
$$=\left(2^{\alpha_1}3^{\beta_1}5^{\gamma_1}7^{\delta_1}\right)\left(2^{\alpha_2}3^{\beta_2}5^{\gamma_2}7^{\delta_2}\right)\left(2^{a_2}3^{b_2}5^{c_2}7^{d_2}\right)^z$$
$$=xde^z.$$

De esta forma se tiene que la operación * está bien definida sobre el conjunto N.

Ejemplo 4. Si se consideran las secuencias $K_1=CC$ y $K_2=TAAG$. Sus respectivos números nula son:

$$\eta(K_1)=(3^3,3^2,2) \text{ y } \eta(K_2)=\left(2^5 5^4 7, 2^2 5^1 7^1, 4\right)$$

Entonces

$$\eta(K_1)*\eta(K_2)=\left(2^9 3^3 5^6 7^3, 2^2 3^2 5^1 7^1, 6\right)=\eta(K_1K_2)$$

Sean K_1 y K_2 son dos secuencias de nucleótidos. Si $K = K_1K_2$ entonces K se llama el ensamble de K_1 y K_2 y la operación * se llamará también operación ensamble.

Proposición 1. La operación * es asociativa.

Aun cuando la prueba es directa es un poco tediosa y por eso se recurrió a una "prueba por computador"[1].
Se define la operación ensamble

```
> ensamble: = proc(A,B)
> return [A[1]*B[1]*B[2]^A[3],A[2]*B[2],
  A[3]+B[3]]
> end proc;
```

```
> ensamble([x,y,z],[d,e,f]);
```

$$[\,xde^z\,,\,ye\,,\,z+f\,]$$

Ahora se comprueba que es asociativa calculando primero

$$((x,y,z)*(d,e,f))*(g,h,i)$$

```
> ensamble(ensamble([x,y,z],[d,e,f]),[g,h,i]);
```

$$\left[xde^z gh^{(z+f)},\ yeh,\ z+f+i\right]$$

y después se calcula

$$(x,y,z)*((d,e,f)*(g,h,i))$$

```
> ensamble([x,y,z],ensamble([d,e,f],[g,h,i]));
```

$$\left[xdgh^f\,(eh)^z\,,\ yeh,\ z+f+i\right]$$

y se observa que los resultados son iguales.

Otras propiedades de la operación ensamble se resumen en la siguiente proposición:

Proposición 2: El conjunto N de los números nula es un monoide no conmutativo con la operación ensamble. Su elemento identidad es la terna (1,1,0) la cual es el número nula de la secuencia vacía:

$$\eta(\varnothing)=(1,1,0).$$

El término monoide significa que la operación es asociativa y tiene elemento identidad. También es evidente que la operación no es conmutativa.

Si se considera el conjunto de todas las secuencias de nucleótidos con la operación de ensamble[2], se observa que también es un monoide no conmutativo con identidad la secuencia vacía.

Note que utilizando la operación * es posible obtener el número nula de una secuencia dada por medio de los números nulas de subsecuencias de la secuencia original. Más aún, el número nula de la secuencia original no depende de las subsecuencias escogidas.

[1] Rutina desarrollada en Maple.

[2] La operación se ensamble también se suele llamar concatenación.

Ejemplo 5. Se consideran las secuencias $K_1 = CCT$ y $K_2 = AAG$. Esta es otra división de la misma secuencia dada en el ejemplo 4. Sus respectivos números nula son:

$$\eta(K_1) = (3^3 7^3, 3^2 7, 3) \text{ y } \eta(K_2) = \left(2^3 5^3, 2^2 5^1, 3\right)$$

Entonces

$$\eta(K_1) * \eta(K_2) = \left(\left(3^3 7^3\right)\left(2^3 5^3\right)\left(2^2 5^1\right)^3, \left(2^2 5^1\right)\left(3^2 7^1\right), 3+3\right)$$
$$= \left(2^9 3^3 5^6 7^3, 2^2 3^2 5^1 7^1, 6\right) = \eta(K_1 K_2).$$

Homomorfismos entre monoides

Se denota por (*H*,*N*) el monoide de secuencias de nucleótidos con la operación ensamble y (*N*,°) el espacio nula con su operación de ensamble definida anteriormente. Se define la siguiente función

$$\eta : (H, \circ) \to (N, *)$$
$$K \mapsto \eta(K),$$

donde por supuesto $\eta(K)$es el número nula de *K*.

Lema 1. η define un homomorfismo de monoides. Es decir, dados $K,P \in H$,

$$\eta(K \circ P) = \eta(K) * \eta(P).$$

Demostración: inmediata a partir de las propiedades para la operación de ensamble y la definición del número nula.

Como se vio anteriormente dicho homomorfismo no es uno a uno. Es decir, es posible hallar $K_1 \neq K_2$ con $\eta(K_1) = \eta(K_2)$. Sin embargo claramente se tiene que si $\eta(K) = (1,1,0)$, entonces $K = \varnothing$. Por lo tanto el núcleo del homomorfismo es trivial aun cuando no sea uno a uno. Por el momento son muchas las preguntas que permanecen abiertas respecto al número nula, en particular su utilidad en la decodificación de genomas, o la determinación de ancestros comunes para distintos genomas, etc. Por esto mismo, su estudio provee de un ambiente interesante matemático y computacional para investigar posibles relaciones matemáticas, con la biología de la evolución.

Algunas preguntas por contestar

Una de las preguntas que se desea resolver es sobre el tamaño de la imagen inversa de un número nula. Así, si $(x, y, z) \in N$ se quiere hallar todas las secuencias *P* tal que $\eta(P) = (x, y, z)$. El cómputo explícito de tales secuencias puede realizarse por medio de las soluciones enteras de un número de ecuaciones.

Ejemplo 6: Para la secuencia

K = *ACCAGTTG* se tiene que

$$[5,5,13,13,2,2,2,2,8] = [\alpha, \beta, \gamma, \delta, a, b, c, d, z]$$

Por lo tanto, para hallar las secuencias que tienen el mismo nula hay que hallar las soluciones enteras para el sistema de ecuaciones

$$a_1 + a_2 = 5$$
$$c_1 + c_2 = 5$$
$$g_1 + g_2 = 13$$
$$t_1 + t_2 = 13$$

en donde los valores son distintos dos a dos, es decir, una A no puede estar en la posición de una C y así sucesivamente. De esta forma se llega a las particiones de 5 y 13 en dos partes distintas y tal que ninguna de las partes sea mayor a 9 porque el valor de z es 8. Particiones de 5 de esta clase son (4,1) y (3,2) y de 13 son (7,6) y (8,5). Por lo tanto después de hacer unos cuantos ensayos se tiene que las únicas secuencias que tienen el mismo nula

$$(2^5 3^5 5^{13} 7^{13}, 2^2 3^2 5^2 7^2, 8)$$

son:

$$K_1 = ACCAGTTG$$
$$K_2 = ACCATGGT$$
$$K_3 = CAACGTTG$$
$$K_4 = CAACTGGT$$

Sin embargo, hay secuencias de largo 8 que comparten el mismo nula con muchas más que en el ejemplo anterior o en caso extremo, ellas son las únicas que tienen ese número nula. La pregunta es si hay forma de determinar una cota superior para el tamaño de esas imágenes inversas en términos de algunos parámetros, por ejemplo el largo de las cadenas.

Dos cadenas que tienen el mismo nula necesariamente tienen el mismo largo y tienen el mismo número de cada uno de los nucleótidos. Por lo tanto ellas difieren en una permutación de sus nucleótidos. Teniendo ahora permutaciones, se puede definir una o varias "distancias" entre dos elementos con el mismo nula. ¿Qué tan "lejos" están unos de otros? ¿Cuál es la probabilidad que estando cerca dos secuencias éstas tengan el mismo número nula?

Como se dijo antes, dos secuencias con el mismo nula tienen los mismos nucleótidos, es decir, tienen el mismo contenido. Si se establece alguna relación entre cadenas de ADN y tablas de Young, éstas se pueden contar al hallar los coeficientes de ciertas expansiones de polinomios en términos de bases de polinomios simétricos y se podrá hallar al menos una cota superior para dichas imágenes inversas. Una tabla de Young es un arreglo de cajas numeradas por números naturales y existen varias formas de asociar una secuencia con una tabla estándar (o semiestándar o ambas), por ejemplo el algoritmo de inserción de Schensted.

También es usual al trabajar con permutaciones hablar de la longitud de dicha permutación. Longitud en este caso no es lo mismo que largo de la cadena. Si se extiende la noción de permutación a una cadena de nucleótidos y se permiten repeticiones, una cadena de nucleótidos es una permutación

$$K = k_1 ... k_n$$

y se define la longitud de una permutación en la forma usual, es decir, contando el número total de inversiones dentro de la permutación, se pueden observar cosas interesantes. Primero, se está considerando la cadena como la permutación que asigna a cada $1 \leq i \leq n$ un nucleótido k_i el cual se representa por un número primo. Por lo tanto se puede suponer que se trabaja con secuencias numéricas o con palabras en un alfabeto sobre un conjunto totalmente ordenado de 4 símbolos. Se define una inversión en $K = k_1 ... k_n$ como una pareja $(i,j), i < j$ tal que $k_i > k_j$.

Ejemplo 7: Sea la cadena $K = ACGTTGCA$. Esta cadena se puede representar por la secuencia K = 23577532. La secuencia tiene en (2,8) una inversión pues $k_2 = 3 > k_8 = 2$. El número total de inversiones de esta secuencia es 12. Ahora bien, las secuencias[3] que tienen el mismo nula que la secuencia dada son

[[2, 3, 5, 7, 7, 5, 3, 2], [2, 3, 7, 5, 5, 7, 3, 2], [2, 5, 3, 7, 7, 3, 5, 2], [2, 7, 3, 5, 5, 3, 7, 2], [2, 5, 7, 3, 3, 7, 5, 2], [2, 7, 5, 3, 3, 5, 7, 2], [3, 2, 5, 7, 7, 5, 2, 3], [3, 2, 7, 5, 5, 7, 2, 3], [5, 2, 3, 7, 7, 3, 2, 5], [7, 2, 3, 5, 5, 3, 2, 7], [5, 2, 7, 3, 3, 7, 2, 5], [7, 2, 5, 3, 3, 5, 2, 7], [3, 5, 2, 7, 7, 2, 5, 3], [3, 7, 2, 5, 5, 2, 7, 3], [5, 3, 2, 7, 7, 2, 3, 5], [7, 3, 2, 5, 5, 2, 3, 7], [5, 7, 2, 3, 3, 2, 7, 5], [7, 5, 2, 3, 3, 2, 5, 7], [3, 5, 7, 2, 2, 7, 5, 3], [3, 7, 5, 2, 2, 5, 7, 3], [5, 3, 7, 2, 2, 7, 3, 5], [7, 3, 5, 2, 2, 5, 3, 7], [5, 7, 3, 2, 2, 3, 7, 5], [7, 5, 3, 2, 2, 3, 5, 7]]

y todas ellas tienen la misma longitud 12. Puede ser una casualidad pero se ha observado experimentalmente que si dos cadenas tienen el mismo número nula sus longi-

[3] Las secuencias se hallaron con rutina desarrollada en CoCoA.

tudes difieren por poco, en donde se espera poder definir ese poco en términos de algún parámetro conocido como puede ser, el largo de las cadenas. Note que esas casualidades se presentan independientemente de la representación de la cadena (puede elegir otros primos).

Estas y muchas otras preguntas (por ejemplo, relaciones entre el nula y cadenas complementarias, etc.) son las que se desea inicialmente solucionar para darle una fundamentación matemática al número nula.

CONCLUSIONES

El espacio nula dotado con la operación de ensamble resulta ser un monoide. El número nula define un homomorfismo de monoides entre el conjunto de las cadenas de nucleótidos y el conjunto nula, y de esta forma el ensamble de cadenas de nucleótidos es compatible con la operación de ensamble definida en el espacio nula.

LITERATURA CITADA

1. LAREO L. AND ACEVEDO O. *SEQUENCE MAPPING IN A THREE-DIMENSIONAL SPACE BY A NUMERIC METHOD AND SOME OF ITS IMPLICATIONS*. ACTA BIOTHEORETICA. 1999, 47: 123-128.

2. HERSTEIN I.N. *TOPICS IN ALGEBRA*. BLAISDELL PUBLISHING COMPANY.1964.

Recibido: 06-02-2003
Aceptado: 12-08-2003