Identificación de islas CpG en el genoma humano a través de las cadenas de Markov: Un modelo matemático basado en probabilidades

Quishpe Evelyn[2], Sánchez María Eugenia[1], Oleas-De la Carrera Gabriela[3], Paz-y-Miño César[1]

[1]Laboratorio de Genética Molecular y Citogenética Humana-Pontificia Universidad Católica del Ecuador

[2] Escuela Politécnica Nacional -Facultad de Ciencias-Ingeniería Matemática

[3] Escuela Politécnica del Ejército –Facultad de Ciencias Aplicadas- Ingeniería en Biotecnología

bevelyn_q@hotmail.com[2], cpazymino@puce.edu.ec[1]

Recibido 1 abril 2005, aprobado 15 junio 2005

RESUMEN. Los genes de importancia biológica conocidos como genes esenciales de mantenimiento o housekeeping genes suelen encontrarse rodeados por regiones denominadas Islas CpG, llamadas así debido a que contienen una cantidad mucho mayor de dinucleótidos CpG que el resto del genoma y, ya que del reconocimiento de tales islas se puede inferir la ubicación de los housekeeping genes, un modelo matemático de identificación de Islas CpG facilitaría su diferenciación del resto del genoma. El modelo matemático que se presenta en este artículo toma como ejemplo a una secuencia de 60 nucleótidos presente en el genoma del parvovirus canino y se basa en las cadenas de Markov para calcular la probabilidad de que un fragmento de dicha secuencia, en relación al resto de ella, corresponda o no a una Isla CpG. Este modelo puede ser utilizado en cualquier secuencia, independientemente de su número de nucleótidos sin embargo, la del parvovirus, escogida en este caso como una pequeña muestra de ejemplificación, sirvió para comparar y confirmar por simple inspección los resultados.

PALABRAS CLAVE. Cadenas de Markov, Islas CpG, Hosekeeping genes.

ABSTRACT. The biologically important genes known as essential genes or housekeeping genes are usually found surrounded by regions called "CpG Isles" . The CpG isles are named so because they contain a much larger quantity of dionucleotides CpG than the rest of the genome. Since during the recognition of such isles the location of the housekeeping genes can be inferred, a mathematical model of identification of CpG isles will make it easier to tell it apart from the rest of the genome. The mathematical model that is presented in this article uses as an example a sequence of 60 nucleotides present in the genome of the canine parovirus and is based on the Markov chains to calculate the probability that a fragment of this sequence, in relationship to the rest of it, corresponds or not to a CpG isle. This model can be used in any sequence, independently from its number of nucleotides. However the parovirus sequence, chosen in this case as a small sample, served to compare and confirm the results by simple inspection.

KEYWORDS. Markov chains, CpG isles, Housekeeping genes.

INTRODUCCIÓN

La importancia del estudio de los housekeeping genes radica en que de su expresión dependen funciones esenciales para la vida de la célula. Sin embargo, y más allá de lo complejo que podría llegar a ser en sí mismo el estudio del gen de nuestro interés, el simple hecho de identificar la región en la que se encuentran los housekeeping genes ya implica un gran avance debido a que, de la vastísima extensión del genoma de los eucariontes apenas una pequeña proporción corresponde a regiones codificantes.

Entre las herramientas que podemos usar para identificar las regiones en las que se encuentran los housekeeping genes están las Islas CpG.

Las Islas CpG son regiones que poseen altas concentraciones de pares de bases CpG y, a menudo, rodean los promotores de los genes expresados constitutivamente, aunque también se encuentran en los promotores de los genes regulados (1).

Las Islas CpG son un punto de partida para la localización de housekeeping genes; identificarlas de una forma rápida y precisa, se lo podría hacer con la utilización de modelos matemáticos.

El presente trabajo relaciona ambos ámbitos, las Islas CpG y los modelos matemáticos a partir de que una cadena de Markov definida como una serie de eventos, en la cual la probabilidad de que ocurra un evento depende del evento inmediato anterior (2). Relacionando esta definición con el hecho de que la transición de una base a otra puede considerarse como un evento aleatorio es posible buscar las probabilidades de que un fragmento corresponda o no a una Isla CpG (Figura 1). El principio es simple: Dado un nucleótido C, cuál será la probabilidad de encontrar un nucleótido G a continuación de éste, es decir, la probabilidad buscada dependerá únicamente del estado anterior (C), situación perfectamente compatible con la definición de la cadena de Markov y aplicable no únicamente a un doblete de bases sino a toda una secuencia.

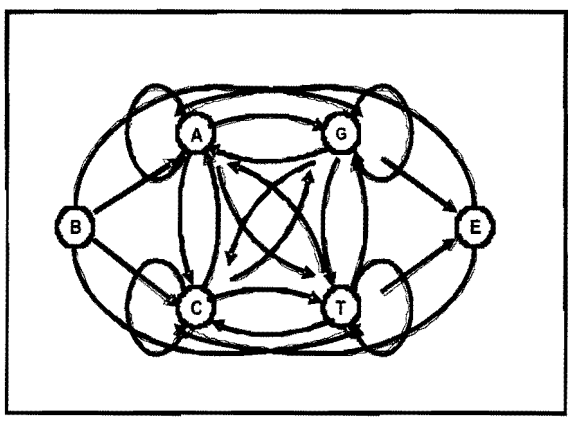

Figura 1.- Posibles cambios de estados (nucleótidos). B indica el punto de partida , una secuencia puede empezar en cualquiera de los nucleótidos, ya elegido un nucleótido, por ejemplo A, el siguiente puede ser una C o T o G o una misma A, ahora si el elegido es C, este se convierte en estado inicial y vuelve a iniciarse el proceso. (3)

MATERIALES Y MÉTODOS

La secuencia que se presenta a continuación corresponde a 60 nucleótidos tomados del genoma del parvovirus canino (3):

ATTCTTTAGAACCAACTGAC
CAAGTTCACGTACGTATGACG
TGATGACCCGCTGCGCGCG

Dicha secuencia fue dividida arbitrariamente en dos fragmentos:

Fragmento -: ATTCTTTAGAAC
CAACTGACCAAGTTCACG
Fragmento +: TACGTATGACGTGAT
GACCCGCTGCGCGCG

Para facilitar la representación de las fórmulas, el signo + fue aplicado a todas las secuencias concernientes a las islas CpG y el signo menos corresponderá a aquellas secuencias involucradas en el resto del genoma.

Ya que una cadena de Markov es un proceso aleatorio en el que

$$\mathbf{p_{ij} = P\,(X_{n+1} = s_{n+1} \mid X_n = s_n)}$$

donde X es el término que representa a los nucleótidos de la secuencia, *n el* número de nucleótidos y *S* las sucesiones posibles de los mismos; fue posible calcular las probabilidades de transición (el paso al azar de un nucleótido a otro), tanto para el fragmento – como para el fragmento + a través de las siguientes fórmulas:

$$p_{ij}^{+} = \frac{n_{ij}^{+}}{\sum_{K} n_{iK}^{+}} \qquad \sum n_{iK}^{*} \;, \qquad p_{ij}^{-} = \frac{n_{ij}^{-}}{\sum_{K} n_{iK}^{-}}$$

donde n^{*}_{ij} es el número de veces que el nucleótido i sigue al nucleótido j en el fragmento y es la suma de n^{*}_{ij} sobre A,C,G y T, es decir, la suma del número de veces que el nucleótido i sigue al nucleótido j sobre todas las cuatro bases. El asterisco (*) representa + o –, dependiendo del caso.

Las probabilidades de transición fueron colocadas en matrices de transición a medida que se calculaban:

Cuadro 1. Matriz de transición para el fragmento –

→Sj+1

↓ Sj	Océano	A-	C-	G-	T-	Σf
	A-	0,333	0,444	0,111	0,111	1
	C-	0,375	0,245	0,125	0,25	1
	G-	0,666	0	0	0,333	1
	T-	0,142	0,285	0,142	0,428	1
	Σc	**1,516**	**0,979**	**0,378**	**1,122**	

(4)

Cuadro 2. Matriz de transición para el fragmento +

→Sj+1

↓ Sj	Océano	A-	C-	G-	T-	Σf
	A-	0	0,6	0	0,41	1
	C-	0	0,125	0,75	0,125	1
	G-	0,375	0,375	0	0,25	1
	T-	0,333	0	0,666	0	1
	Σc	**0,708**	**1,1**	**1,416**	**0,775**	

(4)

En ambas matrices de transición el símbolo Σf es la sumatoria de los valores de las filas, correspondientes a las probabilidades de transición para cada uno de los nucleótidos que, de ser correctas, siempre será 1.

El Símbolo Σc corresponde en cambio a la sumatoria de las columnas, cálculo necesario para determinar si el fragmento corresponde a una isla o al resto del genoma, para lo cual se calculó la probabilidad del fragmento + sobre el - y viceversa.

$$P(\text{fragmento+ / fragmento - }) = \prod_{(i,j)\in K^+} p_{ij} = 0{,}854655 \rightarrow 85\%$$

$$P(\text{fragmento - / fragmento +}) = \prod_{(i,j)\in K^-} p_{ij} = 0{,}629458 \rightarrow 63\%$$

Aquí Π representa el producto de la suma de las probabilidades de transición para cada columna de la matriz de transición.

RESULTADOS Y DISCUSIÓN

De los dos fragmentos en los que, arbitrariamente, fue divida la secuencia original se obtuvo que el primero de estos fragmentos presenta una probabilidad del 63% de corresponder a la secuencia ordinaria del ADN y no a una isla CpG. Del segundo fragmento se obtuvo una probabilidad del 85% de que dicho fragmento corresponde a una Isla CpG y no a la secuencia regular del genoma. Tales resultados, comparados con los de una simple inspección a ambos fragmentos de la secuencia original de 60 nucleótidos permiten confirmar que, efectivamente, el primer fragmento de la secuencia usada posee apenas un dinucleótido CG en comparación a los seis que saltan a la vista en el segundo fragmento. Si bien es cierto, la sección del genoma del parvovirus usada en este artículo como modelo de ejemplificación es sumamente pequeña, por este motivo el hecho de que la primera parte de la secuencia no sea una isla CpG y la segunda sí lo sea puede resultar evidente a simple vista, sin embargo tales consideraciones se dificultarían de manera proporcional al número de nucleótidos de la secuencia y, siendo este modelo perfectamente aplicable a *n* número de nucleótidos, las probabilidades basadas en este tipo de cálculos proporcionarían una forma más objetiva de evaluar cuál secuencia puede ser considerada como una Isla CpG y cuál no, en el caso de ser destinado a secuencias mucho más grandes.

Ahora bien, pese a la enorme ventaja que ofrece la flexibilidad de este modelo, siendo aplicable para una secuencia con cualquier número de nucleótidos, presenta en este mismo hecho su desventaja ya que, aunque los cálculos son relativamente simples, resultaría más laborioso el realizarlos para una secuencia de cientos o miles de nucleótidos, sin embargo, siendo en nuestros días tan sencillo preparar un programa que efectúe operaciones tales como las explicadas anteriormente, tal dificultad sería solventada a través de la elaboración de un programa a base del modelo matemático descrito en este artículo.

REFERENCIAS

1. LEWIN, B. GENES VII Ed. Marbán, Madrid 2001. página 682.
2. http://www.itlp.edu.mx/publica/tutoriales/investoper2/tema43.htm
3. NCBI. Nucleotide sequence and genome organization of canine parvovirus (Banco de Datos-secuencias de AND). www.ncbi.hlm-.nih.gov/entrez/query.fcgi?
4. KARLIN, S. A First Course in Stochastres Processes. USA.1975 Academic. Press. Pags. 58-59.