Distribución Asintótica de la Tasa de Cobertura en un Modelo de Secuenciación Genómica

Gerardo Antonio Alvarado Esquivel* y Rolando Cavazos Cadena

Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro. Buenavista. 25315. Saltillo, Coah., México. Tel.: (844)4110334, 4110334 Fax: (844)4110228. E-mail: alvareski@yahoo.com.mx (*Autor responsable).

Abstract

Starting from a model of genomic sequencing, it is demonstrated that the asymptotic distribution of the coverage rate is normal and, on the basis of this, a credibility interval is established, in which the analyzed proportion of nucleotides is located.

Key words: Whole genom shotgun sequencing, coverage, Bernoulli process, interval of credibility

Resumen

A partir de un modelo de secuenciación genómica, se demostró que la distribución asintótica de la tasa de cobertura es normal y en base a esto se estableció un intervalo de credibilidad en el cual se ubica la proporción analizada de nucleótidos.

Palabras clave: Secuenciación genómica total, cobertura, proceso de Bernoulli, intervalo de credibilidad

Introducción

La secuenciación genómica es una técnica que consiste en descifrar el orden de los nucleótidos del genoma de un organismo en particular, y ha tenido una gran relevancia en los últimos años pues se ha utilizado para secuenciar el genoma de diferentes especies, incluyendo la humana.

En este trabajo se utiliza un modelo de esta técnica con la finalidad de estudiar el comportamiento asintótico de una variable denominada tasa de cobertura, la cual es un indicador de la eficiencia de la técnica. El objetivo es demostrar que esta variable se distribuye asintóticamente normal y, en base a esto, construir un intervalo de credibilidad. Cabe destacar que la principal diferencia entre el enfoque de este trabajo y el modelo comúnmente estudiado, por ejemplo en Ewens y Grant (2005), es que en la literatura se utiliza la aproximación de Poisson a procesos de Bernoulli, mientras que en el desarrollo de este trabajo se emplea directamente un proceso de Bernoulli.

El Modelo

El genoma se puede establecer como una sucesión G cuyas componentes w_i pertenecen a un conjunto finito.

$$G = (w_1, w_1, ...w_g),\ w_i \in B,\ i = 1,2,.....,\mathrm{g} \quad (1)$$

En donde B está constituido por los nucleótidos A, G, C y T (De acuerdo a la inicial de la base nitrogenada que poseen: adenina, guanina, citosina, y timina, respectivamente), mientras que g es la longitud de G, es decir el número de nucleótidos que constituyen el genoma.

Bajo este modelo, se asume a la sucesión *G* como una muestra aleatoria (con sustitución) del conjunto *B*.

La técnica de Secuenciación Genómica Total (*whole genome shotgun*) consiste en fragmentar varias copias del genoma y de los fragmentos generados, se selecciona una muestra de fragmentos constituidos por L bases. Posteriormente los fragmentos seleccionados son secuenciados y ensamblados en bloques de acuerdo a la similitud de sus extremos. A partir de los bloques se determina la secuencia del genoma.

El modelo de fragmentación tiene como componentes básicas: un parámetro de longitud y una sucesión de variables aleatorias $\{X_n\}$ con las siguientes propiedades

$X_1, X_2, X_3, ...$, son independientes

$$X_i \sim Ber(p) \text{ para todo } i \quad (2)$$

En donde $p \approx m/g$ y m es el número de fragmentos secuenciados, es decir que el parámetro es proporcional al tamaño de la muestra[1].

En la expresión (2), $Ber(p)$, denota a la distribución Bernoulli con parámetro, de manera que para cada i = 1,2,3,...

$$P[X_i = 1] = p = 1 - P[[X_i = 0]]$$

Con relación al genoma G en (1), la interpretación de X_i es la siguiente: Si X_i =1, entonces el fragmento que inicia en la posición i del genoma y se prolonga L unidades es analizado y secuenciado, mientras que si X_i = 0, no se analiza fragmento alguno cuyo extremo izquierdo sea la i-ésima posición. De esta manera es natural asociar con cada variable aleatoria un segmento de los números naturales como se hace en la siguiente definición.

Definición 1. Para cada k = 1,2,3,..., el segmento I_k de números naturales está definido mediante.

$I_k := [k, k+L)$, si $X_k = 1$.

$I_k := 0$, si $X_k = 0$.

La variable de estudio en este trabajo es la tasa de cobertura, la cual se refiere a la proporción de posiciones del genoma que son efectivamente analizadas en el proceso de secuenciación. Esta variable se expresa en términos de variables de cobertura, ambas se definen a continuación.

Definición 2. Las variables de cobertura $Y_1, Y_2, Y_3, ...$ se definen para i = 1,2,3,..., como: $Y_i = 1$, si $i \in I_k$, si para algún k

$Y_i = 0$, si $i \notin I_k$ para todo k

Definición 3. Para cada entero positivo, la tasa de cobertura hasta la posición n se denota mediante α_n y se define mediante $\alpha_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$

A continuación se enunciarán dos lemas que son importantes en el análisis subsecuente. En el lema 1 se enumeran las propiedades de las variables de cobertura., en el lema 2 se establece el comportamiento límite de la varianza de la tasa de cobertura.

Lema 1[2]

(i) Para cada $i < L$ $P[Y_i = 1] = 1 - (1-p)^i$ esto es $Y_i \sim Ber(1-(1-p)^i)$

(ii) Para cada $i \geq L$ $P[Y_i = 1] = 1 - (1-p)^L$ esto es ~

(iii) Para

(iv) Para para i , j enteros positivos. Si $j \geq i + L$. Entonces los vectores $(Y_1, Y_2, ..., Y_i)$ y

(v) Dos vectores $(Y_i, Y_{i+1}, ..., Y_{i+d})$ y $(Y_j, Y_{j+1}, ..., Y_{j+d})$ son idénticamente distribuidos.

cual $i \geq L$ y $j \geq i + d + L$ $(Y_j, Y_{j+1}, Y_{j+2} .....)$ son independientes

Lema 2 Para cada $n \geq L$ defina lo siguiente

$$\alpha_n^{'} = \frac{1}{n}\sum_{i=L+1}^{n} Y_i \quad v := \sum_{d=-L}^{d=L} \gamma(d)$$

en donde $\gamma(d) = (1-p)^L\left[(1-p)^{|d|} - (1-p)^L\right]$ si $|d| < L$ $\gamma(d) = 0$ si $|d| \geq L$

En este caso $\underset{n\to\infty}{Lim}\, n\, Var[\alpha_n^{'}] = \sum_{d=-L}^{L} \gamma(d) = v$

Distribución Límite

Una sucesión $\{W_n\}$ de variables aleatorias converge en distribución a la distribución normal con media 0 y varianza v denotada por $N(0,v)$, si para todo $x \in \Re$

$$\lim_{n\to\infty} P[W_n \leq x] = \Phi(x/v) \qquad (3)$$

donde $\Phi(x/v) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ es la función de distribución normal estándar. Cuando (3) ocurre para todo $x \in \Re$, se escribe $\{W_n\} \xrightarrow{d} N(0,v)$.

El resultado central de este trabajo se formula a continuación.

Teorema 1. Conforme $n \to \infty$

$$\sqrt{n}\,(\alpha_n - \alpha^*) \xrightarrow{d} N(0,v)$$

El instrumento básico para establecer este resultado es el siguiente teorema clásico, conocido como Teorema Central del Límite, cuya demostración puede encontrarse, por ejemplo en (Lange, 2003).

Teorema 2. Sean $T_1, T_2, T_3, \ldots\ldots$ variables aleatorias independientes e idénticamente distribuidas con media μ y varianza v. En estas circunstancias

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} T_i - \mu\right) \xrightarrow{d} N(0, v)$$

Este teorema no puede ser utilizado directamente para obtener el Teorema 1, debido a que como lo establece el Lema 1, las variables de cobertura Y_i no son idénticamente distribuidas y aún más importante no son independientes, como lo demuestra el hecho que la covarianza entre dos variables distintas Y_i y Y_j no son necesariamente 0. La estrategia para demostrar el Teorema 1 usará el Teorema Central del Límite después de hacer las adecuaciones necesarias. El punto de partida es introducir las siguientes variables aleatorias auxiliares.

Definición 4. Sea r un entero positivo fijo, para k = 1,2,3,..... defina

$$\delta_k := \frac{1}{L}\sum_{i=(k-1)(r+1)L+1}^{(k-1)(r+1)L+L} Y_i$$

$$\beta_k := \frac{1}{L}\sum_{i=(k-1)(r+1)L+L+1}^{k(r+1)L} Y_i \qquad (4)$$

Con esto se divide a las variables de cobertura en grupos sucesivos de tamaño (r + 1). Después de formar los grupos, el promedio de las primeras L variables se usa para construir δ_k, mientras que el promedio de las restantes L r variables genera β_k. Note que para cada $j < k$, las variables de cobertura involucradas en el promedio β_k tienen índice que supera en, por lo menos, L unidades al índice de cualquier variable aleatoria Y_i que aparece en el promedio β_j, y a partir del lema 1(iv) se desprende que $\beta_1, \beta_2, \beta_3, \ldots\ldots$ son independientes, mientras que la parte (v) del mismo lema implica que estas variables tienen la misma distribución. Similarmente, puede establecerse que las variables δ_k son independientes. Estas conclusiones se establecen formalmente en el siguiente lema.

Lema 3. Con la notación en la definición 3.5.1

(i) $\delta_1, \delta_2, \delta_3, \ldots$ son independientes

(ii) $\beta_1, \beta_2, \beta_3, \ldots$ son independientes e idénticamente distribuidas

Lema 4. Para cada $n > 2(r+1)L$, $\sqrt{n}[\alpha_n - \alpha^*]$ se representa como $\sqrt{n}[\alpha_n - \alpha^*] = Z_n + D_n + R_n$

En donde $\alpha^* = 1 - (1-p)^L$ y las variables del lado derecho satisfacen las siguientes propiedades

(i) $|R_n| \le (r+2)L/\sqrt{n}$

(ii) $E[D_n] = 0$ y $Var[D_n] \le \frac{L}{(r+1)}$

(iii) $Z_n \xrightarrow{d} N(0, \tau_r)$ y $\tau_r = (r+1)\,L\,Var\left[\alpha'_{(r+1)L}\right]$

El siguiente lema es la última etapa antes de la demostración del Teorema 1.

Lema 5. Dado $\varepsilon \in (0,1)$ sea n tal que

$$n > \left(\frac{(r+2)L}{\varepsilon}\right)^2 \quad (5)$$

y $\sqrt{n}[\alpha_n - \alpha^*] = Z_n + D_n + R_n$. En este caso, las siguientes desigualdades son válidas:

$$P\left[\sqrt{n}[\alpha_n - \alpha^*] \le x\right] \le P[Z_n \le x + 2\varepsilon] + \frac{1}{(r+1)\varepsilon^2} \quad (6)$$

$$P\left[\sqrt{n}[\alpha_n - \alpha^*] \le x\right] \ge P[Z_n \le x + 2\varepsilon] - \frac{1}{(r+1)\varepsilon^2} \quad (7)$$

Demostración. Como punto de partida note que (5) implica que $(r+2)L/\sqrt{n} \le \varepsilon$

de donde se desprende que $|R_n| \le \varepsilon$ (8)

Y dado que $x \in \Re$, observe que

$$\sqrt{n}[\alpha_n - \alpha^*] \le x \Leftrightarrow Z_n + D_n \le x - R_n$$

$$\Rightarrow Z_n + D_n \le x + \varepsilon$$

donde se utilizó (8) para establecer la implicación. Por lo tanto

$$\left[\sqrt{n}[\alpha_n - \alpha^*] \le x\right] \subset [Z_n + D_n \le x + \varepsilon]$$

$$= [Z_n + D_n \le x + \varepsilon,\ |D_n| \le \varepsilon] \cup [Z_n + D_n \le x + \varepsilon,\ |D_n| > \varepsilon]$$

$$= \left[Z_n \leq x + \varepsilon - D_n,\ |D_n| \leq \varepsilon\right] \cup \left[|D_n| > \varepsilon\right]$$

$$\subset \left[Z_n \leq x + \varepsilon + \varepsilon,\ |D_n| \leq \varepsilon\right] \cup \left[|D_n| > \varepsilon\right]$$

$$\subset \left[Z_n \leq x + 2\varepsilon,\ \right] \cup \left[|D_n| > \varepsilon\right]$$

Por lo tanto

$$P\left[\sqrt{n}\left[\alpha_n - \alpha^*\right] \leq x\right] \leq P\left[Z_n \leq x + 2\varepsilon,\ \right] + P\left[|D_n| > \varepsilon\right]$$

Observe que la desigualdad de Chebichev y el Lema 4 (ii) implican que

$$P\left[|D_n| > \varepsilon\right] \leq \frac{Var\left[D_n\right]}{\varepsilon^2} \leq \frac{L}{(r+1)\varepsilon^2} \quad (9)$$

y combinando éstas dos últimas relaciones se obtiene

$$P\left[\sqrt{n}\left[\alpha_n - \alpha^*\right] \leq x\right] \leq P\left[Z_n \leq x + 2\varepsilon\ \right] + \frac{L}{(r+1)\varepsilon^2}$$

estableciendo (6). El argumento para demostrar (7) es similar. Note que a partir de (8) se obtiene:

$Z_n \leq x - 2\varepsilon \Rightarrow Z_n + R_n \leq x - \varepsilon$. Por lo tanto

$$\left[Z_n \leq x - 2\varepsilon\right] \subset \left[Z_n + R_n \leq x - \varepsilon\right]$$

$$= \left[Z_n + R_n \leq x - \varepsilon + D_n\right]$$

$$= \left[Z_n + R_n + D_n \leq x - \varepsilon + D_n,\ |D_n| \leq \varepsilon\right] \cup \left[Z_n + R_n + D_n \leq x - \varepsilon + D_n,\ |D_n| > \varepsilon\right]$$

$$\subset \left[Z_n + R_n + D_n \leq x,\ |D_n| \leq \varepsilon\right] \cup \left[|D_n| > \varepsilon\right]$$

$$\subset \left[Z_n + R_n + D_n \leq x\ \right] \cup \left[|D_n| > \varepsilon\right]$$

Por lo tanto $\left[Z_n \leq x - 2\varepsilon\right] \subset \sqrt{n}\left[\alpha_n - \alpha^* \leq x\right] \cup \left[|D_n| > \varepsilon\right]$

y entonces vía $\sqrt{\mathrm{n}}\left[\alpha_n - \alpha^*\right] = \mathrm{Z_n} + \mathrm{D_n} + \mathrm{R_n}$ se desprende que

$$P\left[Z_n \leq x - 2\varepsilon\right] \leq P\left[\sqrt{n}\left[\alpha_n - \alpha^* \leq x\right]\right] + P\left[|D_n| > \varepsilon\right]$$

y empleando (9)

$$P\left[\sqrt{n}\left[\alpha_n - \alpha^*\right] \leq x\right] \geq P\left[Z_n \leq x - 2\varepsilon\right] - \frac{1}{(r+1)\varepsilon^2}$$

Demostración del Teorema

Dado $\varepsilon \in (0,1)$ sea n tal que $n > (r+2)^2 L^2 / \varepsilon^2$, de manera que las desigualdades (6) y (7) en el Lema 5 son válidas. Tomando como límite conforme n tiende a ∞ en las mencionadas desigualdades, se desprende que

$$\lim_{n\to\infty} P\left[Z_n \leq x - 2\varepsilon\right] - \frac{C}{(r+1)\varepsilon^2} \leq \lim_{n\to\infty} P\left[\sqrt{n}\left[\alpha_n - \alpha^*\right] \leq x\right]$$

$$\lim_{n\to\infty} P\left[\sqrt{n}\left[\alpha_n - \alpha^*\right] \leq x\right] \leq \lim_{n\to\infty} P\left[Z_n \leq x + 2\varepsilon\right] + \frac{C}{(r+1)\varepsilon^2}$$

Recordando que $Z_n \xrightarrow{\mathrm{d}} N(0, \tau_{\mathrm{r}})$, por el Lema 4 (iii), se tiene que

$$\lim_{n\to\infty} P\left[Z_n \leq x + 2\varepsilon\right] \to \Phi\left((x+\varepsilon)/\tau_r\right)$$

y similarmente $\lim_{n\to\infty} P\left[Z_n \leq x - 2\varepsilon\right] \to \Phi\left((x-\varepsilon)/\tau_r\right)$. Por lo tanto:

$$\Phi\left(\left(x-\varepsilon\right)/\tau_r\right)-\frac{C}{(r+1)\varepsilon^2} \leq \lim_{n\to\infty} P\left[\sqrt{n}\left[\alpha_n-\alpha^*\right]\leq x\right] \leq \Phi\left(\left(x+\varepsilon\right)/\tau_r\right)+\frac{C}{(r+1)\varepsilon^2} \quad (10)$$

Por otro lado de acuerdo al Lema 2 $\lim_{n\to\infty} n\ Var\left[\alpha_n^{'}\right]= v$ de manera que

$(r+1)LVar\left[\alpha_{(r+1)L}^{'}\right]= v$ si $r\to\infty$,

de manera que a partir de la fórmula para τ_r se desprende que $\lim_{r\to\infty}\tau_r = v$.

Combinando este hecho con la continuidad de $\Phi(\cdot)$, después de tomar el límite cuando r tiende a ∞ en (10) se obtiene que $\Phi\left(\left(x-\varepsilon\right)/\tau_r\right)\leq \lim_{n\to\infty} P\left[\sqrt{n}\left[\alpha_n-\alpha^*\right]\leq x\right] \leq \Phi\left(\left(x+\varepsilon\right)/\tau_r\right)$.

Y tomando el límite en ésta relación conforme ε tiende a cero por la derecha, se arriba a

$\Phi\left(x/v\right)\leq \lim_{n\to\infty} P\left[\sqrt{n}\left[\alpha_n-\alpha^*\right]\leq x\right] \leq \Phi\left(x/v\right)$ de manera que $\sqrt{n}\left[\alpha_n - \alpha^*\right]\xrightarrow{d} N(0,v)$

A partir del Teorema 1 se establece el siguiente intervalo de credibilidad para la tasa de cobertura. Denotando a $z_{c/2}$ como el percentil derecho de orden $c/2$ para la distribución normal estándar, se tiene que para 'n grande',

$$P\left[-vz_{c/2}\leq\sqrt{n}\left[\alpha_n-\alpha^*\right]\leq vz_{c/2}\right]\approx \Phi\left(vz_{c/2}/v\right) - \Phi\left(-vz_{c/2}/v\right)= 1 - c$$

Conclusiones

La tasa de cobertura α_n estandarizada, se distribuye asintóticamente normal, y como el tamaño del genoma es grande, se tiene que α_g se ubica, con probabilidad aproximadamente1-c, dentro del intervalo $\alpha^* \pm vz_{c/2}/\sqrt{g}$.

Como se indicó anteriormente, el parámetro p es proporcional al número de fragmentos secuenciados en la técnica, por lo tanto este intervalo indica los valores entre los que se ubica la proporción de nucleótidos analizados en función del tamaño de la muestra de fragmentos secuenciados

Literatura Citada

Ewnes, W. y G. Grant. 2005. Statistical Methods in Bioinformatics: An introduction, 2nd ed., Springer–Verlag, New York. 588 p.

Lange, K. 2003. Applied Probability, Springer–Verlag, New York, USA. 367 p.