DISTRIBUCIÓN ASINTÓTICA DE LA TASA DE COBERTURA EN UN MODELO DE SECUENCIACIÓN GENÓMICA

GERARDO ANTONIO ALVARADO ESQUIVEL

TESIS
Presentada como requisito parcial
para obtener el grado de
Maestro en Ciencias
en Estadística Experimental

Universidad Autónoma Agraria
Antonio Narro

PROGRAMA DE GRADUADOS
Buenavista, Saltillo Coah.
Diciembre de 2005

UNIVERSIDAD AUTÓNOMA AGRARIA ANTONIO NARRO

SUBDIRECCIÓN DE POSTGRADO

DISTRIBUCIÓN ASINTÓTICA DE LA TASA DE COBERTURA EN UN MODELO DE SECUENCIACIÓN GENÓMICA

TESIS

Por

GERARDO ANTONIO ALVARADO ESQUIVEL

Elaborada bajo la supervisión del comité particular de asesoría y aprobada como requisito parcial para optar al grado de

MAESTRO EN CIENCIAS

EN ESTADÍSTICA EXPERIMENTAL

COMITÉ PARTICULAR

Asesor principal: _____
Dr. Rolando Cavazos Cadena

Asesor: _____
M.C. Luis Rodríguez Gutiérrez

Asesor: _____
M.C. Félix de Jesús Sánchez Pérez

_____
Dr. Jerónimo Landeros Flores
Subdirector de Postgrado

Buenavista, Saltillo, Coahuila. Diciembre de 2005.

AGRADECIMIENTOS

- A mi Familia, por respaldarme en ésta etapa tan importante de mi vida.

- Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por darme la posibilidad de cursar mis estudios de postgrado.

- A la Universidad Autónoma Agraria Antonio Narro, por los beneficios recibidos.

- Al Dr. Rolando Cavazos Cadena, por hacer posible este trabajo y muy en especial por el apoyo que me ha brindado siempre; Muchísimas gracias.

- Al M.C. Félix de J. Sánchez Pérez y al M.C. Luis Rodríguez Gutiérrez, por brindarme su tiempo en la etapa final de la tésis

- A mis amigos de la maestría: Jesús, Jessica y Soledad; Gracias por ser parte de mi vida.

DEDICATORIA

- A mi hermanos, Roberto, Ricardo e iris, por ser la fuente de inspiración más grande de mi vida.

- A mis Padres, Profesor Roberto Alvarado y Profesora Ma del Socorro Esquivel, por su amor incondicional que me fortalece en todo momento.

COMPENDIO

DITRIBUCIÓN ASINTÓTICA DE LA TASA DE COBERTURA EN UN MODELO DE SECUENCIACIÓN GENÓMICA

POR

GERARDO ANTONIO ALVARADO ESQUIVEL

MAESTRÍA EN CIENCIAS

en Estadística Experimental

UNIVERSIDAD AUTÓNOMA AGRARIA ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA. DICIEMBRE, 2005

Dr. Rolando Cavazos Cadena -Asesor-

Palabras claves: secuenciación genómica total, cobertura, proceso de Bernoulli, intervalo de credibilidad.

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

ABSTRACT

ASYMPTOTIC DISTRIBUTION OF COVERAGE RATE IN A GENOME SEQUENCING MODEL

BY

GERARDO ANTONIO ALVARADO ESQUIVEL

MASTER IN SCIENCE

in Experimental Statistics

UNIVERSIDAD AUTÓNOMA AGRARIA ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA. DECEMBER, 2005

Ph.D. Rolando Cavazos Cadena -Advisor-

Key Words: Whole genome shotgun sequencing, Bernoulli process, credibility interval, coverage.

This work concerns a shotgun sequencing model to determine the structure of a genome. It is proved that the proportion of nucleotides effectively analysed is asymptotically normal and using this result a credibility interval is determined.

Contenido

Capítulo 1

El Problema de Fragmentación

El propósito de este capítulo es presentar una descripción intuitiva del problema que se analizará en este trabajo. En términos generales, el objeto central de estudio es una sucesión $\mathcal{G}$ de cuyas componentes sólo se sabe que pertenecen a un conjunto finito $\mathcal{B}$; atendiendo a la motivación biológica, los miembros de $\mathcal{B}$ se denominan *bases*, mientras que la sucesión $\mathcal{G}$ será referida como *el genoma* de interés. Se supone que $\mathcal{G}$ es una sucesión 'larga', y en una primera aproximación es útil pensar que sus componentes conforman una muestra aleatoria (con sustitución) del conjunto de bases $\mathcal{B}$. El problema central es identificar los elementos del genoma bajo la siguiente condición: sólo es posible determinar las componentes de cadenas 'cortas', cuyas longitudes son mucho menores que la de $\mathcal{G}$. En estas condiciones, copias de $\mathcal{G}$ se someten a un proceso de segmentación que genera fragmentos lo suficientemente cortos para que sus componentes sean identificadas, y posteriormente dichos segmentos se acoplan para formar bloques mayores de componentes conocidas, obteniendo información sobre la estructura de $\mathcal{G}$. El tema de este trabajo es el estudio de un modelo para este proceso de fragmentación y acoplamiento, el cual permite abordar problemas como determinar la longitud esperada de un bloque, o encontrar la distribución asintótica de la proporción de componentes de $\mathcal{G}$ que serán analizadas en el proceso.

Introducción

Este trabajo trata sobre un modelo para el problema de determinar las componentes de una sucesión (cadena) $\mathcal{G}$, cuyas componentes ω_i pertenecen a un conjunto de 'bases' $\mathcal{B}$:

$$\mathcal{G} = (\omega_1, \omega_2, \ldots, \omega_g), \quad w_i \in \mathcal{B}, \quad i = 1, 2, \ldots, g, \tag{1.1.1}$$

Este problema se estudia para dar respuesta a diferentes interrogantes sobre la 'Secuenciación Genómica total', una técnica de laboratorio en la que se analiza el genoma de algún organismo para determinar su secuencia. En el contexto de ésta técnica, $\mathcal{B}$ consiste de las cuatro bases A, G, T y C (adenina, guanina, timina y citosina, respectivamente) y $\mathcal{G}$ es el genoma de interés, mientras que g es la longitud de $\mathcal{G}$, es decir, el número de bases que constituyen el genoma. Ahora bien, existe una limitante en la secuenciación: en forma directa, solo es posible determinar la secuencia de fragmentos de un máximo de 700 bases ($M.Pop, 2002$), por tal motivo es imposible secuenciar directamente algún genoma, pues el genoma más pequeño del que se tiene registro es el de *Mycoplasma genitalium* que consiste de 580,070 bases mientras que por ejemplo el de la especie humana está conformado por cerca de 3 billones de nucleótidos ($J.CraigVenter, 2001$). Para superar ésta limitante, la técnica se realiza mediante dos procesos, el primero denominado proceso de fragmentación, consiste en segmentar una gran cantidad de copias del genoma para generar una población de fragmentos. A partir de ésta población se selecciona una muestra de fragmentos de tamaño adecuado y se determina su secuencia. El siguiente proceso denominado proceso de acoplamiento, consiste en comparar la secuencia de los fragmentos muestreados, para irlos agrupando de tal manera que se formen bloques. Al final de este proceso se obtendrán una serie de bloques que servirán para determinar la secuencia total del genoma. En las siguientes secciones se presenta una versión simplificada del proceso biológico real de fragmentación y acoplamiento, y las ideas discutidas serán la base para formular el modelo probabilístico analizado en el resto del trabajo.

Fragmentos

Considere un gran número de copias $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N$ de la sucesión $\mathcal{G}$ en (1.1.1), las cuales se romperán, en fragmentos de longitud $L > 0$, donde L es 'mucho menor' que la longirud g del genoma $\mathcal{G}$. Por simplicidad, suponga que cada copia, al someterse al proceso de ruptura, genera un sólo fragmento de longitud L. En la práctica este proceso se lleva a cabo, sometiendo una gran cantidad de copias del genoma a altas presiones lo que ocasiona que se generen fragmentos de diversos tamaños. Posteriormente se efectúa una técnica llamada electroforésis, en la que los fragmentos se colocan en el extremo de un recipiente que contiene gel. A continuación se genera un flujo de corriente electrica que hace que los fragmentos sean atraídos hacia el otro extremo del recipiente. La magnitud del movimiento de cada fragmento depende de su número de bases, por lo cual los fragmentos quedan separados de acuerdo a su longitud, lo que hace posible que sean seleccionados solo aquellos que posean una longitud L (E. Myers, 1999). Los posibles fragmentos de longitud L son

$$\begin{aligned} F_1 &= (\omega_1, \omega_2, \ldots, \omega_L) \\ F_2 &= (\omega_2, \omega_3, \ldots, \omega_{1+L}) \\ F_3 &= (\omega_3, \omega_2, \ldots, \omega_{2+L}) \\ &\vdots \\ F_{g-L+1} &= (\omega_{g-L+1}, \omega_2, \ldots, \omega_g) \end{aligned} \tag{1.2.1}$$

Los fragmentos obtenidos pueden ser considerados como seleccionados al azar entre todas las alternativas posibles, por lo que cada vez que una copia de $\mathcal{G}$ se rompe, la probabilidad de obtener un fragmento F_i específico es

$$\rho = \frac{1}{g - L + 1} \approx \frac{1}{g}, \tag{1.2.2}$$

donde la aproximación se debe a que L es 'pequeño' comparado con g. Como la segmentación se realiza N veces, al final de la división el analista tiene ante sí una población de N fragmentos de longitud L, y en esa población cada segmento F_i

se repite, aproximadamente, $N\rho$ veces. Ahora deben seleccionarse algunos fragmentos para secuenciar. Los fragmentos no tienen marca alguna que, 'a primera vista', le indiquen al analista su posición dentro del genoma $\mathcal{G}$. Por esta razón, se seleccionan fragmentos al azar para determinar su contenido; debido a que, aproximadamente, hay $N\rho$ copias del fragmento F_i dentro de la población de N segmentos, la probabilidad de obtener F_i en una selección particular es

$$\frac{N\rho}{N} = \rho; \tag{1.2.3}$$

más aún, como N es 'grande', el proceso de muestreo puede considerarse con sustitución, al menos para tamaños de muestra moderados. Una pregunta que surge naturalmente en este momento, es ¿cómo reconocerá el analista que dos fragmentos analizados corresponden a la misma posición dentro del genoma? Desde luego, si dos segmentos analizados corresponden a las mismas posiciones dentro de $\mathcal{G}$, entonces sus contenidos son iguales. En adelante, se supondrá que el recíproco de esta afirmación es cierto, esto es, si dos segmentos de longitud L tienen las mismas componentes, entonces se ubican en la misma posición dentro del genoma. A continuación se verá que esta hipótesis es razonable, al menos si $\mathcal{G}$ puede considerarse como una muestra aleatoria de la población de bases $\mathcal{B}$. En efecto, en estas circunstancias, cada fragmento F_i es una muestra aleatoria de tamaño L de $\mathcal{B}$, y el hecho de que F_i y F_j tengan el mismo contenido cuando $i \neq j$ significa que dos muestras diferentes de tamaño L coinciden, lo que constituye un evento de probabilidad 'pequeñísima' aún para valores moderados de L. Por ejemplo, suponiendo que $\mathcal{B}$ tiene cuatro componentes, y que $i - j > L$, de manera que F_i y F_j no tienen posiciones en común, la probabilidad de que F_i y F_j tengan el mismo contenido es igual a $\sum_\sigma (1/4^L)^2 = 1/4^L$, donde la suma se extiende sobre todas las muestra ordenadas σ de tamaño L de la población $\mathcal{B}$; note que $1/4^L$ decrece muy rápidamente y, aún para valores moderados de L, es un número pequeño.

Aunque en las aplicaciones es díficil sostener que $\mathcal{G}$ constituya una muestra aleatoria del conjunto de bases $(M.Pop, 2002)$, este argumento muestra que el

supuesto de que si dos segmentos tienen el mismo contenido entonces se ubican en la misma posición dentro de $\mathcal{G}$ es una hipóteis razonable. La discusión precedente puede resumirse como sigue: Después de fragmentar un gran número N de copias de $\mathcal{G}$, se dispone de una población de segmentos en la que el número de fragmentos iguales a F_i es aproximadamente $N\rho \approx N/g$. Por lo tanto, al seleccionar un fragmento aleatoriamente, la probabilidad de que el fragmento seleccionado coincida con un segmento particular F_j es $\rho \approx 1/g$. Más aún, es razonable suponer que dos segmentos diferentes F_i y F_j, los cuales se ubican en posiciones distintas dentro de $\mathcal{G}$, no tienen la misma constitución. Note que la probabilidad $\rho \approx 1/g$ de que un segmento seleccionado al azar coincida con uno específico es bastante pequeña, pues la longitud g el genoma es grande. A continuacón se verá que la probabilidad de que el fragmento particular F_j sea seleccionado, y por lo tanto secuenciado, puede incrementarse siguiendo un plan de muestreo adecuado.

Probabilidad de Secuenciar un Fragmento

Suponga que el analista extrae al azar una muestra de la población de fragmentos hasta que obtiene m distintos segmentos. Como la elección es aleatoria, la probabilidad de que los fragmentos seleccionados conformen el conjunto $\{F_{i_1}, F_{i_2}, \ldots, F_{i_m}\}$ es la misma para todos los índices diferentes $i_1, i_2, \ldots, i_m$. De acuerdo a (1.2.1), el número de fragmentos diferentes es

$$M := g - L + 1, \tag{1.3.1}$$

y entonces, para cada conjunto $\{F_{i_1}, F_{i_2}, \ldots, F_{i_m}\}$ de m segmentos distintos se tiene que la probabilidad de que sea seleccionado es (Casella y Berger 2001, Dudewicz y Mishra, 1988)

$$P\{F_{i_1}, F_{i_2}, \ldots, F_{i_m}\} = \frac{1}{\binom{M}{m}}.$$

Por lo tanto, denotando mediante p_j a la probabilidad de que un fragmento es-

pecífico F_j pertenezca a la muestra, se obtiene que

$$\begin{aligned} p_j &= P[F_j \text{ pertenece a la muestra seleccionada}] \\ &= \sum_{j\in\{F_{i_1},F_{i_2},\ldots,F_{i_m}\}} P\{F_{i_1},F_{i_2},\ldots,F_{i_m}\} = \frac{\binom{M-1}{m-1}}{\binom{M}{m}} = \frac{m}{M}, \end{aligned}$$

esto es,

$$p_j = \frac{m}{g-L+1} = m\rho \equiv p, \quad j = 1,2,3\ldots; \tag{1.3.2}$$

vea (1.2.2) y (1.3.1). Por lo tanto, la probabilidad de analizar un segmento específico no depende de j y es proporcional al tamaño de la muestra m. Esto pone de manifiesto que el instrumento que el analista tiene para 'controlar' la probabilidad de secuenciar un segmento es el tamaño de la muestra: variando m se puede lograr que $p_j = m\rho$ alcance un nivel deseado.

Considere ahora la probabilidad $p_{i,\,j}$ de que dos fragmentos distintos F_i y F_j sean secuenciados, la cual está dada por

$$\begin{aligned} p_{i,\,j} &= P[F_i \text{ y } F_j \text{ pertenece a la muestra seleccionada}] \\ &= \sum_{i,j\in\{F_{i_1},F_{i_2},\ldots,F_{i_m}\}} P\{F_{i_1},F_{i_2},\ldots,F_{i_m}\} = \frac{\binom{M-2}{m-2}}{\binom{M}{m}} = \frac{m}{M}\times\frac{m-1}{M-1}, \end{aligned}$$

Puesto que $M = g - L + 1$ es 'grande', se tiene que

$$p_{i,\,j} \approx \frac{m}{M}\times\frac{m}{M} = p_i p_j,$$

de manera que la inclusión de un fragmento F_i en la muestra es, 'prácticamente', independiente del evento de que otro segmento F_j sea también analizado. Estas observaciones son la base para el modelo formal de fragmentación que se introducirá en el capítulo siguiente. De hecho, el modelo está basado en variables aleatorias X_i que asumen valores cero y uno, donde $X_i = 1$ (resp. $X = 0$) se interpreta como la inclusión (resp. ausencia) del fragmento F_i en la muestra, y se supondrá que $P[X_i = 1] = p$ para todo i, y que X_i y X_j son independientes para $i \neq j$; de hecho, por simplicidad en la argumentación, en el modelo se supondrá la independencia conjunta de *todas* las variables X_i, y no sólo la independencia por parejas.

Bloques

Después de seleccionar m fragmentos distintos, lo que el analista sabe sobre la estructura de $\mathcal{G}$ es la composición de m tramos del genoma, cada uno de longitud L. Sin embargo, no conoce la ubicación de dichos segmentos dentro de $\mathcal{G}$. Esto se debe a que al seleccionar un fragmento, digamos F_5, el analista no dispone de ninguna evidencia para saber que el segmento se inicia en la posición cinco, y lo único que puede hacer, y hace, es determinar las L componentes de dicho segmento. Así, después de secuenciar los m fragmentos es necesario buscar 'acoplarlos' para ampliar la porción de componentes consecutivas identificadas. Para clarificar esta idea, suponga que A_1 y A_2 son dos fragmentos cuyas componentes ya han sido determinadas. La pregunta que debe responderse es si hay evidencia de qué las posiciones ocupadas por A_1 y A_2 se traslapan en el genoma. Para abordar esta cuestión, suponga que ℓ (o más) componentes extremas de A_1 coinciden con ℓ componentes extremas de A_2 como se ilustra a continuación:

$$A_1\text{:} \quad \ldots\ldots\ldots\ldots\ldots \underbrace{**********}_{\ell \textit{ Componentes comunes}}$$

$$\overbrace{**********}\ldots\ldots\ldots\ldots\ldots\ldots : A_2$$

Si acaso A_1 y A_2 fueran dos segmentos disjuntos en el genoma, se tendría que el tramo común de ℓ elementos se repite en dos secciones *ajenas* de $\mathcal{G}$. Pensando en que las componentes de $\mathcal{G}$ fueron seleccionadas por medio de un mecanismo aleatorio a partir del conjunto de cuatro bases $\mathcal{B}$, la probabilidad de la repetición observada sería $1/4^\ell$, la cual es muy pequeña aún para valores moderados de ℓ; por ejemplo, $1/4^\ell < 0.001$ cuando $\ell \geq 5$. En este caso, sostener el supuesto de que A_1 y A_2 son disjuntos teniendo un tramo de ℓ componentes en común implica que se ha observado un evento 'muy raro', y siguiendo la filosofía aplicada en la teoría de pruebas de hipótesis, se rechaza el supuesto de que A_1 y A_2 ocupan posiciones disjuntas dentro del genoma, esto es, se afirma que las posiciones ocupadas por A_1 y A_2 no son disjuntas, y como las ℓ componentes extremas son comunes, se declara

que que A_1 y A_2 se intersectan precisamente en esas ℓ componentes, formando una región R de $2L - \ell$ posiciones en $\mathcal{G}$ cuyo contenido se conoce; en la figura anterior, R abarca desde la extrema izquierda de A_1 hasta la extrema derecha de A_2. A continuación, se considera otro segmento A_3 el cual se pegará a la región R si tiene ℓ o más componentes extremas en común con ella. Este proceso de agrandar R acoplando segmentos eventualmente termina, pues llega el momento en que ningún segmento restante se intersecta con la región R que se quiere incrementar en ℓ o más posiciones. Esa región de componentes ya conocidas que ya no se puede agrandar más se denomina *bloque*, y representa una región de posiciones consecutivas dentro de $\mathcal{G}$ cuyo contenido se conoce; en la literatura inglesa, lo que aquí se denomina bloque se llama *contig* (Ewens y Grant, 2005, Altshuler *et. al.*, 2000). Después de construir un bloque, se empieza con otro fragmento que no sea algún bloque existente, y se trata de agrandarlo, acoplándolo con otros segmentos de la manera descrita, hasta que se tiene una región que ya no se puede agrandar más, obteniendo otro bloque. Esta construcción continúa hasta que cada segmento se ha incorporado a un bloque. La familia de bloques que se obtiene representa la forma en que el analista obtiene información sobre $\mathcal{G}$ a partir de los fragmentos de longitud L de que se dispone originalmente; en general, el número de bloques es menor que el número original de fragmentos, mientras que la longitud de un bloque excede a la de los segmentos iniciales. Esta idea intuitiva de bloque se formaliza en el siguiente capítulo.

Los Problemas

El propósito de este trabajo es responder, en términos probabilísticos, algunas preguntas sobre el proceso de fragmentación y acoplamiento descrito de forma intuitiva en las secciones precedentes. Estas preguntas se refieren, primeramente, a las componentes individuales ω_i del genoma $\mathcal{G}$. Denote por α_n a la proporción

de posiciones del genoma con índice menor o igual a n que son analizadas (y, por lo tanto, determinadas) como parte de algún fragmento. Por ejemplo, suponga que $n = 5$, y que en la muestra no aparecen los fragmentos F_1, F_2 y F_3, mientras F_4 si aparece. En este caso, $\alpha_5 = 2/5$, pues las posiciones 4 y 5 se determinan al analizar F_4, mientras que las posiciones 1,2 y 3 no se analizan, de manera que, entre las primeras cinco posiciones, sólo dos de ellas son analizadas, y entonces $\alpha_5 = 2/5$. Desde luego, cada proporción α_n es una variable aleatoria, y entre los problemas que se analizarán se encuentran las siguientes:

(i) Determinar $E[\alpha_n]$, el valor esperado de α_n.

De particular interés en este problema es el caso $n = g$ =longitud del genoma, pues al investigador le interesa la proporción del total de bases que serán analizadas. Una fórmula para $E[\alpha_g]$ es importante pues, conociéndola, el analista sabrá como seleccionar la probabilidad p de que un segmento esté incluido en la muestra para alcanzar un nivel deseado de $E[\alpha_g]$. Por otro lado, α_n es una variable aleatoria, así que, aún conociendo su esperanza, para tener una idea de los valores que α_n puede tomar, o determinar un intervalo de credibilidad para esa proporción, es necesario conocer una medida de su variablidad, lo cual motiva el siguiente problema:

(ii) Determinar $\mathrm{Var}\,[\alpha_n]$, la varianza de α_n.

Al conocer la media y la varianza de α_n, la desigualdad de Chevichev permite establecer que

$$P\left[E[\alpha_n] - k\sqrt{\mathrm{Var}\,[\alpha_n]} \leq \alpha_n \leq E[\alpha_n] + k\sqrt{\mathrm{Var}\,[\alpha_n]}\right] \geq 1 - \frac{1}{k^2}$$

de tal manera que, antes de realizar el proceso de selección y secuenciación de fragmentos, para cada n de interés puede decirse que, con probabilidad $1-1/k^2$, α_n pertenecerá al intervalo $E[\alpha_n] \pm k\sqrt{\mathrm{Var}\,[\alpha_n]}$. La siguiente pregunta es interesante si se recuerda que, en las aplicaciones, la longitud g del genoma es un número bastante grande.

(iii) ¿Es válida una ley de los grandes números para la sucesión $\{\alpha_n\}$?. Más explícitamente, ¿existe una constante β tal que, en algún sentido, la sucesión $\{\alpha_n\}$ converge a β?

En esta dirección, se demostrará que la sucesión $\{\alpha_n\}$ si converge *con probabilidad uno* a una constante β, de manera que, debido a que la longitud g del genoma es'grande', es 'prácticamente' seguro que, α_g, la proporción de posiciones del genoma que son analizadas, sea un número muy cecano a β. Finalmente, con el propósito de establecer intervalos de credibilidad para la proporción de posiciones del genoma que será efectivamente analizada, es conveniente determinar si la sucesión $\{\alpha_n\}$ tiene una distribución asintótica. El siguiente resultado es una de las principales contribuciones de este trabajo:

(iv) Conforme n se incrementa, la distribución de $\sqrt{n}\,(\alpha_n - E[\alpha_n])$ converge a la distribución normal con media cero y varianza $v = \lim_{n\to\infty} n\mathrm{Var}\,[\alpha_n]$.

Por medio de este resultado, es posible obtener intervalos de credibilidad para α_g, los cuales son más cortos que los obtenidos mediante la desigualdad de Chebichev. Por otro lado, como ya se ha mencionado, la forma en que el investigador extrae información sobre la composición del genoma es por medio de la construcción de bloques delineada anteriormente: Por supuesto, los bloques son objetos aleatorios, y los siguientes problemas se refieren a propiedades de su distribución.

(v) Determinar la distribución de la longitud de un bloque, y encontrar su valor esperado.

Una respuesta a este problema permite al investigador diseñar el plan de muestreo para lograr obtener bloques 'grandes'. Por otro lado, entre menos bloques haya, se tendrá que el analista ha logrado una mejor identificación de los componentes del genoma, razón por la cual el siguiente problema es interesante.

(vi) Encontrar la distribución del número de bloques y su valor esperado.

Antes de concluir este capítulo e iniciar el desarrollo técnico, es conveniente establecer 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, y la construcción de bloques no impone una intersección mínima a la metodología de acoplamiento, mientras que en el desarrollo de este trabajo se emplea directamente un proceso de Bernoulli y si se requiere una intersección mínima de ℓ componente para poder acoplar dos fragmentos analizados.

Por otro lado, la principal contribución de este estudio, al cual se ubica en el Capítulo 3, consiste en establecer la normalidad asintótica de la sucesión de promedios de bases analizada $\{\alpha_n\}$. La idea detrás del análisis que condujó a ese resultado, proviene del área de series de tiempo estacionarias (Brockwell y Davis, 1998, 2002).

Capítulo 2

El Modelo

En este capítulo se formula un modelo probabilístico para el problema de secuenciación descrito anteriormente. La componente central del modelo es una sucesión de variables aleatorias de Bernoulli, cuyo papel es indicar si un fragmento es analizado o no. A partir de esta sucesión, se introducen las ideas de intensidad, variables y tasas de cobertura, así como la noción de bloque, conceptos alrededor de los cuales gira la parte técnica de este trabajo.

Introducción

El propósito de este capítulo es introducir el entorno probabilístico que se utilizará para analizar los problemas planteados previamente. La discusión inicia en la Sección 2 presentando los objetos básicos del modelo, a saber, una pareja de números L y ℓ, así como una sucesión $\{X_k\}$ de variables aleatorias de Bernoulli, las cuales se usan para indicar si el intervalo (segmento) de números naturales $\mathcal{I}_k$ que empieza en k y se prolonga L unidades hacia la derecha, es secuenciado o no. Posteriormente, en la Sección 3 se formula, la idea de intensidad del proceso de muestreo que se emplea para secuenciar el genoma, mientras que en la Sección 4

se introducen las nociones de variables y tasas de cobertura. Finalmente, en la Sección 5 se presenta la definición formal de bloque; el estudio de las propiedades de bloques y tasas de cobertura permitirá dar respuesta a los problemas planteados en el capítulo precedente.

Componentes Básicas

El modelo de fragmentación y acoplamiento tiene dos componentes básicas: una pareja de enteros positivos L y ℓ, que satisfacen $L > \ell$ y se denominan parámetros de *longitud y acoplamiento*, respectivamente, y una sucesión de variables aleatorias $\{X_n\}$ con las siguientes propiedades:

$$\begin{aligned} &X_1, X_2, X_3, \ldots, \text{ son independientes;} \\ &X_i \sim Ber(p) \text{ para todo } i; \end{aligned} \tag{2.2.1}$$

como es usual, en este enunciado $Ber(p)$ denota a la distribución de Bernoulli con parámetro p, de manera que la segunda de las condiciones en (2.2.1) expresa que, para cada $i = 1, 2, 3, \ldots$,

$$P[X_i = 1] = p = 1 - P[X_i = 0]. \tag{2.2.2}$$

Con relación al genoma $\mathcal{G}$ en (1.1.1), 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 hasta que se cubren L unidades a la derecha es analizado y secuenciado—esto es, sus componentes son identificadas—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 X_i un segmento de los números naturales como se hace a continuación.

Definition 2.2.1. Para cada $k = 1, 2, 3, \ldots$, el segmento $\mathcal{I}_k$ de números naturales está definido mediante

$$\mathcal{I}_k := \begin{cases} [\![k, k + L), & \text{si } X_k = 1, \\ \emptyset, & \text{si } X_k = 0, \end{cases}$$

donde se utiliza la siguiente notación: para enteros positivos a y b,

$$[\![a, b) := \{a, a+1, \ldots, b-1\}.$$

Con esta notación, los enteros dentro de $\mathcal{I}_k$ representan las posiciones de $\mathcal{G}$ cuyo contenido será determinado si $X_k = 1$. El parámetro de acoplamiento se utiliza e interpreta más adelante.

Intensidad de Cobertura

En esta sección se introduce una idea central en el estudio del modelo de fragmentación y acoplamiento, a saber, un indicador de la intensidad con que el genoma esta siendo analizado. Dicho índice es el número esperado de veces en que una posición específica del genoma es analizada como parte de uno de los segmentos $\mathcal{I}_k$.

Como punto de partida, note que el contenido de una posición del genoma puede ser, en general, determinado varias veces, tantas como el número de intervalos $\mathcal{I}_k$ a los que pertenezca. Por ejemplo, suponga que $L = 10$, y que $X_3 = X_8 = X_9 = X_{21} = 1$, mientras que $X_k = 0$ si $k \neq 3, 8, 9, 21$. En estas cirunstancias, los únicos intervalos $\mathcal{I}_s$ no vacíos son $\mathcal{I}_3, \mathcal{I}_8, \mathcal{I}_9$ e $\mathcal{I}_{21}$. Observando que

$$\mathcal{I}_3 = \{3, 4, 5, \ldots, 12\}$$

$$\mathcal{I}_8 = \{8, 9, 10, \ldots, 17\}$$

$$\mathcal{I}_9 = \{9, 10, 11, \ldots, 18\}$$

$$\mathcal{I}_{21} = \{21, 22, 23, \ldots, 30\}$$

se desprende que la posición 4 es identificada solamente una vez, como miembro del segmento $\mathcal{I}_3$, la posición 9 es determinada tres veces, como miembro de $\mathcal{I}_3$, de $\mathcal{I}_8$ y de $\mathcal{I}_9$, la posición 18 se identifica dos veces, pues 18 pertenece a $\mathcal{I}_9$ e $\mathcal{I}_{21}$, mientras que 24 se identifica solamente una vez, como miembro de $\mathcal{I}_{21}$. Por supuesto, las

posiciones 2 y 33 no se identifican vez alguna, pues ni 2 ni 33 pertenecen a alguno de los segmentos $\mathcal{I}_k$. El aspecto que se desea enfatizar con esta discusión, es que el número de veces en que el contenido de una posición del genoma se analiza y determina, es una *variable aleatoria*. Para cada entero positivo i, defina

$$N_i = \sum_{k \geq 1} I[i \in \mathcal{I}_k]. \tag{2.3.1}$$

En el lado derecho de esta igualdad, el término $I[i \in \mathcal{I}_k]$ es uno cuando i pertenece al segmento $\mathcal{I}_k$ y cero de otro modo, de tal manera que N_i es el número total de intervalos que contienen a la posición i, y por lo tanto es igual al número de ocasiones en que la posición i es secuenciada (determinada). Como se muestra a continuación, N_i tiene, siempre, una distribución binomial, y ésta es la misma para todo $i \geq L$.

Teorema 2.3.1. Sea N_i la variable aleatoria definida en (2.3.1). En este caso

$$N_i \sim B\left(\min\{L, i\}, p\right), \tag{2.3.2}$$

donde $B\left(a, b\right)$ denota a la distribución binomial con parámetros a y b. Por lo tanto,

$$E[N_i] = ip, \quad i < L, \qquad E[N_i] = Lp, \quad i \geq L. \tag{2.3.3}$$

Demostración. Dados dos enteros positivos i y k, primero se demostrará que la función indicadora $I[i \in \mathcal{I}_k]$ puede expresarse como

$$I[i \in \mathcal{I}_k] = \begin{cases} X_k, & \text{si } i - L + 1 \leq k \leq i, \\ 0, & \text{si } k < i - L + 1 \text{ ó } k > i. \end{cases} \tag{2.3.4}$$

Con este propósito, note que $I[i \in \mathcal{I}_k] = 1$ significa que $i \in \mathcal{I}_k$—de manera que $\mathcal{I}_k$ es no vacío—lo que ocurre solamente cuando $X_k = 1$. A partir de esta observación, la Definición 2.2.1 permite establecer la siguientes equivalencias:

$$\begin{aligned} i \in \mathcal{I}_k &\iff X_k = 1 \text{ y además } i \in \{k, k+1, \ldots, k+L-1\} \\ &\iff X_k = 1, \quad \text{y} \quad k \leq i \leq k + L - 1; \end{aligned}$$

observando que la relación $k \leq i \leq k+L-1$ puede expresarse como $i-L+1 \leq k \leq i$ se desprende que

$$i \in \mathcal{I}_k \iff X_k = 1 \text{ y } i - L + 1 \leq k \leq i. \tag{2.3.5}$$

Por lo tanto, la inclusión $i \in \mathcal{I}_k$ no ocurre cuando la condición $i - L + 1 \leq k \leq i$ no se satisface, esto es cuando alguna de las desigualdades $k < i - L + 1$ ó $k > i$ se cumple, de manera que

$$I[i \in \mathcal{I}_k] = 0 \quad \text{si } k < i - L + 1 \text{ ó } k > i,$$

de conformidad con lo estipulado en el segundo caso de (2.3.4) Suponga ahora que $i - L + 1 \leq k \leq i$. En estas circunstancias, (2.3.5) muestra que $X_k = 1$ equivale a la inclusión $i \in \mathcal{I}_k$, la cual a su vez es equivalente a $I[i \in \mathcal{I}_k] = 1$. Por lo tanto, $X_k = 1$ si y sólo si $I[i \in \mathcal{I}_k] = 1$; como X_k y la función indicadora $I[i \in \mathcal{I}_k]$ toman solamente los valores cero y uno, se desprende que

$$I[i \in \mathcal{I}_k] = X_k \quad \text{si } i - L + 1 \leq k \leq i.$$

Combinando las dos últimas relaciones desplegadas se obtiene (2.3.4). Ahora la conclusiones del teorema pueden establecerse como sigue: A partir de (2.3.1) y (2.3.4) se desprende que

$$\begin{aligned} N_i &= \sum_{k \geq 1} I[i \in \mathcal{I}_k] \\ &= \sum_{1 \leq k, \;\; i-L+1 \leq k \leq i} X_k, \end{aligned} \tag{2.3.6}$$

y observando que las desigualdades $i - L + 1 \leq k$ y $1 \leq k$ equivalen a $\max\{i - L + 1, 1\} \leq k$, se desprende que

$$N_i = \sum_{\max\{1, i-L+1\} \leq k \leq i} X_k. \tag{2.3.7}$$

Considere ahora los siguientes casos:

Caso 1: $i < L$. Bajo esta condición se tiene que $i - L < 0$ de manera que $i - L + 1 \leq 0$, y entonces $\max\{1, i - L + 1\} = 1$. Por lo tanto, (2.3.7) implica que

$$N_i = \sum_{1 \leq k \leq i} X_k = \sum_{k=1}^{i} X_k, \quad i < L, \tag{2.3.8}$$

lo cual muestra que N_i es una suma de i variables aleatorias independientes con distribución de Bernoulli, de donde se desprende que

$$\text{para } i < L \text{ se tiene que } \; N_i \sim B\,(i, p), \quad \text{y entonces} \quad E[N_i] = ip; \tag{2.3.9}$$

(Koski 2002), Casella y Berger (2001).

Caso 2: $i \geq L$. En esta cirunstancia $i - L \geq 0$ de tal forma que $i - L + 1 \geq 1$, y entonces $\max\{1, i - L + 1\} = i - L + 1$, y a partir de (2.3.7) se obtiene

$$N_i = \sum_{i-L+1 \leq k \leq i} X_k = \sum_{k=i-L+1}^{i} X_k, \quad i \geq L, \tag{2.3.10}$$

mostrando que N_i es una suma de L variables aleatorias independientes con distribución de Bernoulli. Por lo tanto,

$$\text{si } i \geq L \text{ entonces } N_i \sim B\,(L, p), \quad \text{y consecuentemente} \quad E[N_i] = Lp.$$

Combinando este enunciado con (2.3.9) se obtienen las conclusiones establecidas en (2.3.2) y (2.3.3), concluyendo la demostración. □

De acuerdo al Teorema 2.3.1, para $i \geq L$, el número esperado de veces en que el i-ésimo elemento del genoma será cubierto por segmentos que serán secuenciados es Lp. Este número desempeña un importante papel en el desarrollo del trabajo, especialmente al obtener aproximaciones para diversas cantidades de interés.

Definition 2.3.1. *La intensidad de cobertura*, denotada mediante κ, se define mediante

$$\kappa := Lp.$$

Además de representar el número esperado de veces que una posición del genoma será analizada, κ también tiene otra interpretación interesante. Denote mediante $|\mathcal{I}_k|$ a la longitud del segmento $\mathcal{I}_k$ en la Definición 2.2.1, de manera que $|\mathcal{I}_k| = L$ si $X_k = 1$, mientras que $|\mathcal{I}_k| = 0$ si $X_k = 0$; por lo tanto, $E[|\mathcal{I}_k|] = pL$. La longitud total $\mathcal{L}$ de todos los segmentos $\mathcal{I}_k$, los cuales serán secuenciados, es $\mathcal{L} = \sum_{k=1}^{g} |\mathcal{I}_k|$, y su valor esperado es

$$E[\mathcal{L}] = \sum_{k=1}^{g} E[|\mathcal{I}_k|] = \sum_{k=1}^{g} pL = pLg,$$

y entonces

$$\frac{E[\mathcal{L}]}{g} = pL = \kappa.$$

Por lo tanto, κ es la razón de la longitud total analizada con relación a la longitud del genoma, mostrando, de nueva cuenta, que κ mide la intensidad con que se trata de analizar las posiciones de $\mathcal{G}$.

Variables y Tasas de Cobertura

En esta sección se introduce una sucesión $\{Y_k\}$ de variables indicadoras. Cada variable Y_k asume sólo los valores cero y uno, y por lo tanto es una variable de Bernoulli. Lo que el valor uno para Y_k indica es que la k–ésima posición del genoma $\mathcal{G}$ fue efectivamente secuenciado como miembro de alguno de los segmentos $\mathcal{I}_j$ en la Definición 2.2.1. Formalmente, esta idea se introduce a continuación.

Definition 2.4.1. Las *variables de cobertura* $Y_1, Y_2, Y_3, \ldots$ se definen como sigue: Para $i = 1, 2, 3, \ldots$,

$$Y_i = \begin{cases} 1, & \text{si } i \in \mathcal{I}_k \text{ para algún } k, \\ 0, & \text{si } i \notin \mathcal{I}_k \text{ para todo } k. \end{cases}$$

Una parte importante del análisis subsecuente se refiere a la proporción de posiciones del genoma que son efectivamente analizadas en el proceso de secuenciación. Estas proporciones (o tasas) de cobertura se introducen a continuación.

Definition 2.4.2. Para cada entero positivo, la tasa de cobertura hasta la posición n se denota mediate α_n y se define mediante

$$\alpha_n = \frac{\sum_{k=1}^{n} Y_k}{n}.$$

Puesto que la longitud del genoma g es 'grande', es interesante investigar el comportamiento de las tasas de cobertura α_n conforme n crece. Este trabajo se desarrolla en el siguiente capítulo, y está basado en las propiedades de las variables de cobertura establecidas a continuación.

Lemma 2.4.1. (i) Para cada $i < L$, $P[Y_i = 1] = 1 - (1-p)^i$, esto es,

$$Y_i \sim Ber(1 - (1-p)^i).$$

(ii) Si $i \geq L$, entonces $P[Y_i = 1] = 1 - (1-p)^L$, y por lo tanto

$$Y_i \sim Ber(1 - (1-p)^L).$$

(iii) Para $i \neq j$,

$$\text{Cov}\,[Y_i, Y_j] = P[Y_i = 0, Y_j = 0] - P[Y_j = 0]P[Y_i = 0].$$

(iv) Suponga que i y j son enteros positivos con $i \leq j$.

(a) Si $j \geq i + L$, entonces los vectores $(Y_1, Y_2, \ldots, Y_i)$ y $(Y_j, Y_{j+1}, Y_{j+2}, \ldots)$ son independientes.

(b) Si $j < i + L$, entonces

$$\text{Cov}\,[Y_i, Y_j] = (1-p)^{j-\max\{1,i-L+1\}+1} - (1-p)^{\min\{i,L\}+\min\{j,L\}}.$$

(c) Si $i + L \geq j \geq i \geq L$, entonces $\text{Cov}\,[Y_i, Y_j] = (1-p)^L[(1-p)^{j-i} - (1-p)^L]$.

(v) Dos vectores $(Y_i, Y_{i+1}, \ldots, Y_{i+d})$ y $(Y_j, Y_{j+1}, \ldots, Y_{j+d})$ son idénticamente distribuídos cuando $i \geq L$ y $j \geq i + d + L$.

Demostración. Como punto de partida, observe que Y_i es uno, sólo cuando i pertenece a alguno de los intervalos $\mathcal{I}_k$, y como N_i en (2.3.1) es el número de tales intervalos que contienen a i, se desprende que

$$Y_i = I[N_i > 0]. \tag{2.4.1}$$

Por lo tanto, $Y_i = 0$ si y sólo si $N_i = 0$, y utilizando la fórmula (2.3.7) se obtiene

$$\begin{aligned} Y_i = 0 &\iff \sum_{\max\{1,i-L+1\}\leq k\leq i} X_k = 0 \\ &\iff X_k = 0, \quad k = \max\{1, i-L+1\}, \ldots, i. \end{aligned} \tag{2.4.2}$$

(i) Suponga que $i < L$. En este caso $i-L+1 \leq 0$, de manera que $\max\{1, i-L+1\} = 1$, y entonces $Y_i = 0$ equivale a $X_k = 0$ para $k = 1, 2, \ldots, i$. Por lo tanto,

$$\begin{aligned} P[Y_i = 0] &= P[X_1 = 0, X_2 = 0, \ldots, X_i = 0] \\ &= \underbrace{(1-p) \times (1-p) \times \cdots \times (1-p)}_{i \ factores} \\ &= (1-p)^i; \end{aligned}$$

vea (2.2.1) y (X). Por lo tanto, $P[Y_i = 1] = 1 - (1-p)^i$ y, consecuentemente, $Y_i \sim Ber(1-(1-p)^i)$.

(ii) Cuando $i \geq L$, se tiene que $i-L+1 \geq 1$, y entonces $\max\{i-L+1, 1\} = i-L+1$, de donde se desprende que $Y_i = 0$ equivale a $X_k = 0$ para $k = i-L+1, i-L+2, \ldots, i$. Luego,

$$\begin{aligned} P[Y_i = 0] &= P[X_{i-L+1} = 0, X_{i-L+2} = 0, \ldots, X_i = 0] \\ &= \underbrace{(1-p) \times (1-p) \times \cdots \times (1-p)}_{L \ factores} \\ &= (1-p)^L, \end{aligned}$$

de tal manera que $P[Y_i = 1] = 1 - (1-p)^L$, de manera que

$$Y_i \sim Ber(1-(1-p)^L).$$

(iii) Defina $Z_i = 1 - Y_i$ y note que cada Z_i toma valores 0 y 1, asi que Z_i tiene distribución de Bernoulli, y que $Z_i = 1$ equivale a $Y_i = 0$. Por lo tanto $E[Z_i] =$

$P[Z_i = 1] = P[Y_i = 0]$. Además, Z_iZ_j toma sólo valores cero y uno, por lo que también tiene distribución de Bernoulli, y entonces $E[Z_iZ_j] = P[Z_iZ_j = 1]$; puesto que $Z_iZ_j = 1$ si y sólo si $Z_i = 1$ y $Z_j = 1$, se desprende que $E[Z_iZ_j] = P[Z_i = 1, Z_j = 1] = P[Y_i = 0, Y_j = 0]$, y entonces

$$\begin{aligned}\text{Cov}\,[Z_i, Z_j] &= E[Z_iZ_j] - E[Z_i]E[Z_j] \\ &= P[Y_i = 0, Y_j = 0] - P[Y_1 = 0]P[Y_j = 0].\end{aligned}$$

Combinando esta igualdad con

$$\text{Cov}\,[Z_i, Z_j] = \text{Cov}\,[1 - Y_i, 1 - Y_j] = \text{Cov}\,[Y_i, Y_j]$$

se obtiene la conclusión deseada.

(iv) (a) Observe que Y_k es una función de variables X_r con $r \leq k$. Por lo tanto,

• $(Y_1, Y_2, \ldots, Y_i$ es función de $X_1, X_2, \ldots, X_i)$.

Por otro lado, Y_k es función de variables X_r con $r > k - L$ de tal manera uqe

• $(Y_j, Y_{j+1}, Y_{j+2}, \ldots)$ es función de $(X_{j-L+1}, X_{j-L+2}, \ldots)$.

Luego, cuando $j \geq i + L$ los vectores $(Y_1, Y_2, \ldots, Y_i)$ y $(Y_j, Y_{j+1}, \ldots)$ dependen de grupos disjuntos de variables X_r, y la independencia de dichos vectores se desprende de (2.2.1).

(b) Suponga ahora que $0 < j - i < L$. Para calcular $\text{Cov}\,[Y_i, Y_j]$, es conveniente notar que $j - L < i$ y entonces $j - L + 1 \leq i$ y puesto que i es positivo se desprende que

$$\max\{j - L + 1, 1\} \leq i.$$

Además, como $j > i$, se tiene que

$$\max\{j - L + 1, 1\} \geq \max\{i - L + 1, 1\}.$$

de manera que

$$\max\{i - L + 1, 1\} \leq \max\{j - L + 1, 1\} \leq i < j. \tag{2.4.3}$$

Ahora observe que (2.4.2) implica que

$$Y_i = 0 \quad \text{y} \quad Y_j = 0$$

si y sólo si

$$X_k = 0 \quad si \quad \max\{1, i - L + 1\} \leq k \leq i \quad 0 \quad \max\{1, j - L + 1\} \leq k \leq j,$$

condición que, via (2.4.3), equivale a

$$X_k = 0 \quad si \quad \max\{1, i - L + 1\} \leq k \leq j.$$

Por lo tanto,

$$\begin{aligned} P[Y_i = 0, Y_j = 0] &= P[X_k = 0,\ \max\{1, i - L + 1\} \leq k \leq j] \\ &= \prod_{k=\max\{1,i-L+1\}}^{j} P[X_k = 0] \\ &= \prod_{k=\max\{1,i-L+1\}}^{j} (1 - p) \end{aligned}$$

donde las últimas dos igualdades se deben a (2.2.1), y entonces

$$P[Y_i = 0, Y_j = 0] = (1 - p)^{j - \max\{1, i-L+1\}+1}.$$

A continuación, observe que las partes (i) y (ii) previamente establecidas implican que

$$P[Y_i = 0] = (1 - p)^{\min\{i,L\}}, \quad P[Y_j = 0] = (1 - p)^{\min\{j,L\}}.$$

Combinando las dos últimas relaciones desplegadas con la fórmula establecida en la parte (iii) se obtiene que

$$\text{Cov}\,[Y_i, Y_j] = (1 - p)^{j - \max\{1, i-L+1\}+1} - (1 - p)^{\min\{i,L\}+\min\{j,L\}}.$$

(d) Cuand $j \geq i \geq L$ se tiene que

$$\max\{1, i - L + 1\} = i - L + 1, \quad \text{y} \quad \min\{i, L\} = \min\{j, L\} = L.$$

Sustituyendo estas expresiones en la fórmula obtenida en la parte (c) se desprende que

$$\text{Cov}\,[Y_i, Y_j] = (1-p)^{j-[i-L+1]+1} - (1-p)^{L+L} = (1-p)^{j-i+L} - (1-p)^{2L}$$

y entonces $\text{Cov}\,[Y_i, Y_j] = (1-p)^L[(1-p)^{j-i} - (1-p)^L]$, concluyendo la demostración.

(v) Note que a partir de (2.4.1) y () se desprende que, para cierta función f,

$$\begin{aligned} Y_i &= f(X_{i-L+1}, X_{i-L+2}, \ldots, X_i) \\ Y_{i+1} &= f(X_{(i+1)-L+1}, X_{(i+1)-L+2}, \ldots, X_{i+1}) \\ &\vdots \\ Y_{i+d} &= f(X_{(i+d)-L+1}, X_{(i+d)-L+2}, \ldots, X_{i+d}) \end{aligned}$$

de manera que

$$(Y_i, Y_{i+1}, \ldots, Y_{i+d}) = F(X_{i-L+1}, X_{i-L+2}, \ldots, X_{i+d})$$

para cierta función F. Similarmente,

$$(Y_j, Y_{j+1}, \ldots, Y_{j+d}) = F(X_{j-L+1}, X_{j-L+2}, \ldots, X_{j+d}).$$

Los vectores $X_{i-L+1}, X_{i-L+2}, \ldots, X_{i+d})$ y $(X_{j-L+1}, X_{j-L+2}, \ldots, X_{j+d})$ son idénticamente distribuidos, por (2.2.1), y cuando $j - L \geq i + d$, son independientes y, en este último caso, $(Y_i, Y_{i+1}, \ldots, Y_{i+d})$ y $(Y_j, Y_{j+1}, \ldots, Y_{j+d})$ también lo son. □

Acoplamiento

Como ya se ha mencionado, la forma en que el analista trata de obtener información sobre el genoma es acoplando segmentos que se intersectan, para formar bloques. Esta idea se introduce ahora formalmente y se utiliza la siguiente notación: Para cada conjunto A,

$$|A| := \text{Número de elementos de } A$$

Definition 2.5.1. Sea $\{\mathcal{I}_k\}$ la sucesión de intervalos en la Definición 2.2.1. Un bloque es una sucesión

$$B = (\mathcal{I}_{k_0}, \mathcal{I}_{k_2}, \ldots, \mathcal{I}_{k_{r-1}}) \tag{2.5.1))}$$

que satisface las siguientes condiciones (i)–(v):

(i) $k_0 < k_2 < \cdots < k_{r-1}$;

(ii) $\mathcal{I}_s = \emptyset$ si $k_0 \leq s \leq k_{r-1}$ y $s \neq k_0, k_2, \ldots, k_{r-1}$;

(iii) $\left|\mathcal{I}_{k_{i-1}} \cap \mathcal{I}_{k_i}\right| \geq \ell$ para $i = 1, 2, \ldots, r-1$;

(iv) $|\mathcal{I}_{k_0} \cap \mathcal{I}_{k_s}| < \ell$ para todo $s < k_0$;

(v) $\left|\mathcal{I}_{k_{r-1}} \cap \mathcal{I}_{k_s}\right| < \ell$ para todo $s > k_{r-1}$.

Note que los intervalos que conforman el bloque son consecutivos, en el sentido de que $\mathcal{I}_s$ es vacío para s entre dos indices k_i y k_{i+1}, por las condiciones (i) y (ii). Más aún, de acuerdo a la discusión presentada en el Capítulo 1, dos intervalos deben acoplarse cuando su intersección contenga ℓ o más elementos, y la condición (iii) muestra que dos intervalos consecutivos en el bloque B satisfacen ese requerimiento. Más aún, ningún intervalo 'anterior' a $\mathcal{I}_{k_0}$ puede acoplarse al bloque, pues su intersección con $\mathcal{I}_{k_0}$, y por lo tanto con los demás intervalos del bloque, consiste de menos de ℓ elementos, por la condición (iv). Similarmente, la condición (v) implica que ningún intervalo $\mathcal{I}_s$ con $s > k_{r-1}$ puede acoplarse al bloque. La definición anterior difiere de la proporiconada en Ewens y Grant (2005); en ésta referencia, ℓ se toma igual a 1, de manera que para acoplar dos fragmentos, basta que tengan una posición extrema en común.

La longitud del bloque B en (2.5.1)se denota mediante $\|B\|$ y se define como

$$\|B\| = \left|\bigcup_{i=0}^{r-1} \mathcal{I}_{k_i}\right|, \tag{2.5.2}$$

esto es, $\|B\|$ es el número de elementos en la unión de los intervalos que conforman al bloque B. El estudio de la longitud de un bloque y del número de bloques se lleva a cabo en el Capítulo 4.

Capítulo 3

Propiedades Límite de las Tasas de Cobertura

En este capítulo se estudia el comportamiento de las tasas de cobertura introducidas en la Definición 2.4.2 conforme la longitud n del genoma crece. El objetivo es demostrar que dichas tasas satisfacen una ley de los grandes números, y que su distribución asintótica es normal; éste último resultado, es una de las principales contribuciones de este trabajo.

Introducción

En este capítulo se estudia el comportamiento límite de la tasa de cobertura α_n conforme el número de bases n en el genoma crece. La exposición inicia en la Sección 2 calculando el valor esperado de α_n y su valor límite, α^*, el cual es posteriormente relacionado con la intensidad de cobertura κ, moostrando que $\alpha^* \approx 1 - e^{-\kappa}$, o equivalentemente, $\kappa \approx -\log(1-\alpha^*)$, relación que permite determinar el tamaño de muestra necesario para alcanzar una cobertura esperada; vea la Observación 3.2.1. Posteriormente, en la Sección 3 se estudia la varianza de α_n y se determina el límite $\lim_{n\to\infty} n\mathrm{Var}\left[\alpha_n\right]$, el cual es finito, resultado que, combinado con la desigualdad de Chebichev, implica que α_n converge en probabilidad a α^*, de tal forma que la ley débil de los grandes números es válida para

la sucesión $\{\alpha_n\}$, resultado que en la Sección 4 se refuerza, mostrando que una ley fuerte de los grandes números también es satisfecha por la sucesión de tasas de cobertura. Finalmente, el capítulo concluye en la Sección 5 mostrando que las tasas de cobertura tiene una distribución asintóticamente normal.

Valor Esperado

El propósito de esta sección es establecer el siguiente resultado referente a los valores esperados de las tasas de cobertura α_n.

Teorema 3.2.1. Sea $\{\alpha_n\}$ la sucesión de tasas de cobertura introducida en la Definición 2.4.2.

(i) Si $n \leq L$, entonces

$$E[\alpha_n] = 1 - \frac{1-p}{pn}[1-(1-p)^n]. \tag{3.2.1}$$

(ii) Para $n > L$,

$$E[\alpha_n] = \frac{L}{n} - \frac{1-p}{pn}[1-(1-p)^L] + \frac{n-L}{n}[1-(1-p)^L]. \tag{3.2.2}$$

Consecuentemente,

(iii) Conforme $n \to \infty$,

$$E[\alpha_n] \to \alpha^* := 1-(1-p)^L. \tag{3.2.3}$$

Demostración. Primeramente, note que a partir de las partes (i) y (ii) del Lemma 2.4.1 se desprende que

$$E[Y_i] = \begin{cases} 1-(1-p)^i, & i \leq L, \\ 1-(1-p)^L, & i > L. \end{cases} \tag{3.2.4}$$

(i) Si $n \leq L$, la primera parte de la anterior igualdad implica que

$$E\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} E\left[Y_i\right] = \sum_{i=1}^{n}[1-(1-p)^i] = n - \sum_{i=1}^{n}(1-p)^i;$$

y a partir de la fórmula $\sum_{i=1}^{n} r^i = r(1-r^n)/(1-r)$ con $1-p$ en lugar de r se sigue que

$$E\left[\sum_{i=1}^{n} Y_i\right] = n - \frac{1-p}{p}[1-(1-p)^n].$$

Dividiendo por n ambos lados de esta igualdad se obtiene la expresión (3.2.1) para $E[\alpha_n]$.

(ii) Suponga que $n > L$. Usando la segunda parte de (3.2.4) se tiene que

$$E\left[\sum_{i=L+1}^{n} Y_i\right] = (n-L)[1-(1-p)^L]$$

mientras que (3.2.1) con L en vez de n implica que

$$E\left[\sum_{i=1}^{L} Y_i\right] = L - \frac{1-p}{p}[1-(1-p)^L]$$

A partir de estas dos igualdades se desprende que

$$\begin{aligned} E\left[\sum_{i=1}^{n} Y_i\right] &= E\left[\sum_{i=1}^{L} Y_i\right] + E\left[\sum_{i=L+1}^{n} Y_i\right] \\ &= L - \frac{1-p}{p}[1-(1-p)^L] + (n-L)[1-(1-p)^L] \end{aligned}$$

de donde (3.2.2) se obtiene de inmediato.

(iii) Finalmente, tomando límite conforme n tiende a infinito en ambos lados de (3.2.2) se arriba a la igualdad $\lim_{n\to\infty} E[\alpha_n] = \alpha^*$, donde α^* es como se especifíca en (3.2.3). □

Como ya se ha mencionado, la longitud g del genoma $\mathcal{G}$ en (1.1.1) es 'grande', de manera que, de acuerdo al Teorema 3.2.1, la proporción esperada $E[\alpha_g]$ de posiciones del genoma que son analizadas como parte de algún fragmento es, prácticamente, α^*.

Observación 3.2.1. Recuerde que la tasa de cobertura κ está dada por $\kappa = Lp$. Por lo tanto

$$(1-p)^L = \left(1 - \frac{\kappa}{L}\right)^L.$$

Aplicando la relación

$$e^{-a-aO(1/L)} \leq \left(1-\frac{a}{L}\right)^{L} \leq e^{-a}$$

con κ en lugar de a se obtiene

$$e^{-\kappa-\kappa O(1/L)} \leq (1-p)^{L} \leq e^{-\kappa}.$$

Por lo tanto,

$$1-e^{-\kappa-\kappa O(1/L)} \geq 1-(1-p)^{L} \geq 1-e^{-\kappa},$$

esto es, $1-e^{-\kappa-\kappa O(1/L)} \geq \alpha^{*} \geq 1-e^{-\kappa}$. Esta relación se expresa de forma más compacta como

$$\alpha^{*} \approx 1-e^{-\kappa}, \tag{3.2.5}$$

y generaliza un resultado similar establecido en Ewens y Grant (2001) para un modelo de Poisson. La aproximación anterior puede usarse para determinar que tan intensamente debe secuenciarse el genoma para alcanzar un nivel dado de cobertura esperada límite α^{*}. Por otro lado, $\kappa = Lp$, mientras que, como se discutió en el Capítulo 1, $p \approx L(m/g)$. Por lo tanto, el tamaño de la muestra m y α^{*} se relacionan a través de

$$m \approx -\frac{g}{L}\log(1-\alpha^{*}).$$

Por ejemplo, se se desea una tasa de cobertura $\alpha^{*} = 0.95$, entonces se obtiene que $m = 1.3010(g/L)$.

Varianza

En esta sección se estudia la varianza de las tasas de cobertura $\{\alpha_n\}$, particularmente su comportamiento conforme n tiende a infinito. Como α_n es la suma $Y_1 + Y_2 + \cdots + Y_n$ dividida por n, la varianza de α_n se expresa, finalmente, en

términos de las covarianzas entre las variables Y_i, las cuales fueron calculadas en el Lema 2.4.1 (iv). Para el análisis subsecuente, es conveniente introducir la *función de covarianza*, definida como sigue: Para cada entero d,

$$\gamma(d) := \begin{cases} (1-p)^L[(1-p)^{|d|} - (1-p)^L], & \text{si } |d| < L \\ 0, & \text{si } |d| \geq L. \end{cases} \tag{3.3.1}$$

Comparando esta especificación de $\gamma(\cdot)$ con el Lemma 2.4.1(iv) se obtiene que

$$\text{Cov}\,[Y_i, Y_j] = \gamma(|i-j|), \quad i, j \geq L. \tag{3.3.2}$$

El principal resultado de esta sección relaciona esta función de covarianza con el comportamiento límite de $\text{Var}\,[\alpha_n]$, y las ideas detras del argumento de prueba estan motivadas en el análisis desarrollado en Brockell y Davis (1998, 2002).

Teorema 3.3.1. Defina v mediante

$$v := \sum_{d=-L}^{d=L} \gamma(d). \tag{3.3.3}$$

Con esta notación,

$$\lim_{n\to\infty} n\text{Var}\,[\alpha_n] = v. \tag{3.3.4}$$

A partir del Lemma 2.4.1 (iv) y (3.3.2), es claro que la fórmula para $\text{Cov}\,[Y_i, Y_j]$ es bastante simple cuando i y j son, ambos, mayores o iguales a L. Por esta razón, el argumento para establecer le Teorema 3.3.1 utiliza el siguiente lema.

Lemma 3.3.1. Para cada $n \geq L$, defina

$$\alpha'_n = \frac{1}{n} \sum_{i=L+1}^{n} Y_i. \tag{3.3.5}$$

En este caso

$$\lim_{n\to\infty} n\text{Var}\,[\alpha'_n] = \sum_{d=-L}^{L} \gamma(|d|) = v. \tag{3.3.6}$$

Demostración. Seleccione un entero $n > 2L$ y note que

$$\begin{aligned} \text{Var}\left[n\alpha'_n\right] &= \text{Var}\left[\sum_{i=L+1}^{n} Y_i\right] \\ &= \sum_{i=L+1}^{n}\sum_{j=L+1}^{n} \text{Cov}\left[Y_i, Y_j\right] \\ &= \sum_{i,j=L+1}^{n} \gamma(|i-j|) \end{aligned} \tag{3.3.7}$$

donde (3.3.2) se utilizó para establecer la última igualdad. Ahora se descompondrá la última sumatoria de acuerdo a los valores de las diferencias $i-j$. Con este fin, observe que cuando i y j varian entre $L+1$ y n, $i-j$ asume valores entre $-(n-L-1)$ y $n-L-1$, de tal manera que

$$\sum_{i,j=L+1}^{n} \gamma(|i-j|) = \sum_{d=-(n-L-1)}^{n-L-1} \sum_{i,\,j\,:\; i-j=d,\; L+1\leq i,j,\leq n} \gamma(|d|)$$

Para evaluar la suma en la extrema derecha note que, para $d \geq 0$, las parejas i, j con i y j entre $L+1$ y n y que satisfacen $i-j=d$ son $(L+1, L+1+d), (L+2, L+2+d), \ldots, (n-d, n)$ de las cuales hay un total de $(n-d-L) = (n-|d|-L)$. Por lo tanto,

$$\sum_{i,\,j\,:\; i-j=d,\; L\leq i,j,\leq n} \gamma(|d|) = \gamma(|d|)(n-|d|-L),$$

mientras que un análisis similar muestra que esta relación es también válida para $d < 0$, de tal manera que

$$\sum_{i,j=L}^{n} \gamma(|i-j|) = \sum_{d=-(n-L)}^{n-L} \gamma(|d|)(n-|d|-L) = \sum_{d=-L}^{L} \gamma(|d|)(n-|d|-L)$$

donde la segunda igualdad se debe a que $n > 2L$ y a que $\gamma(|d|) = 0$ cuando $|d| > L$; vea (3.3.1). Combinando esta expresión con (3.3.7) se desprende que $\text{Var}\left[n\alpha'_n\right] = \sum_{d=-L}^{L} \gamma(|d|)(n-|d|-L)$, y recordando que $\text{Var}\left[n\alpha'_n\right] = n^2\text{Var}\left[\alpha'_n\right]$ se obtiene

$$n\text{Var}\left[\alpha'_n\right] = \sum_{d=-L}^{L} \gamma(|d|)\left(1 - \frac{|d|+L}{n}\right),$$

y (3.3.6) se obtiene tomando límite conforme n tiende a infinito en ambos lados de esta igualdad. □

Demostración del Teorema 3.3.1. El argumento utiliza la siguente desigualdad para la varianza de una suma de variables aleatorias X y Y (Koski 2002):

$$\text{Var}\left[X+Y\right] \leq \left(\sqrt{\text{Var}\left[X\right]} + \sqrt{\text{Var}\left[Y\right]}\right)^2, \tag{3.3.8}$$

la cual es consecuencia de la desigualdad de Cauchy Schwarz (Casella y Berger 2001, Dudewicz i Mishra, 1998). Considere ahora un entero $n > L$ y observe que

$$n\alpha_n - n\alpha'_n = \sum_{i=1}^{L} Y_i =: F. \tag{3.3.9}$$

de manera que $n\alpha_n = n\alpha'_n + F$. Aplicando (3.3.8) con $n\alpha'_n$ y F en lugar de X y Y se obtiene que

$$\begin{aligned} n^2\text{Var}\left[\alpha_n\right] &= \text{Var}\left[n\alpha_n\right] \\ &= \text{Var}\left[n\alpha'_n + F\right] \\ &\leq \left(\sqrt{\text{Var}\left[n\alpha'_n\right]} + \sqrt{\text{Var}\left[F\right]}\right)^2 \\ &= \left(\sqrt{n^2\text{Var}\left[\alpha'_n\right]} + \sqrt{\text{Var}\left[F\right]}\right)^2 \end{aligned}$$

Por lo tanto,

$$n\text{Var}\left[\alpha_n\right] \leq \left(\sqrt{n\text{Var}\left[\alpha'_n\right]} + \frac{\sqrt{\text{Var}\left[F\right]}}{\sqrt{n}}\right)^2$$

y tomando límite cuando n tiende a infinito, (3.3.6) implica que

$$\lim_{n\to\infty} n\text{Var}\left[\alpha_n\right] \leq \left(\sqrt{\lim_{n\to\infty} n\text{Var}\left[\alpha'_n\right]}\right)^2 = v. \tag{3.3.10}$$

Similaremente, observando que $n\alpha'_n = n\alpha_n - F$, una aplicación de (3.3.8) con $n\alpha_n$ y $-F$ en lugar de X y Y permite concluir que

$$n\text{Var}\left[\alpha'_n\right] \leq \left(\sqrt{n\text{Var}\left[\alpha_n\right]} + \frac{\sqrt{\text{Var}\left[F\right]}}{\sqrt{n}}\right)^2$$

de manera que

$$v = \lim_{n\to\infty} n\text{Var}\left[\alpha'_n\right] \leq \left(\sqrt{\lim_{n\to\infty} n\text{Var}\left[\alpha_n\right]}\right)^2 = \lim_{n\to\infty} n\text{Var}\left[\alpha_n\right],$$

lo que combinado con (3.3.10) implica que $v = \lim_{n\to\infty} n\text{Var}\left[\alpha_n\right]$. □

En el Teorema 3.2.1(iii), se estableció que $E[\alpha_n]$ converge a α^* (vea (3.2.3)). Por lo tanto, para n suficientemente grande, $E[\alpha_n]$ es aproximadamente α^*, y como el tamaño g del genoma es es un número 'grande', el analista está seguro que $E[\alpha_g]$ se encuentra cerca de α^*. Sin embargo, aunque una información sobre el valor esperado de una cantidad aleatoria es importante, al investigador le gustaría saber si la porporción α_g, y no sólo su esperanza, se ubica cerca de α^*. Que esto realmente ocurre con una alta probabilidad, es el significado del siguiente resultado.

Corollary 3.3.1. La sucesión $\{\alpha_n\}$ converge en probabilidad a α^*, *i.e.*,

$$\alpha_n \xrightarrow{\text{P}} \alpha^*.$$

Más precisamente, para cada $\varepsilon > 0$

$$P[|\alpha_n - \alpha^*| > \varepsilon] \to 0 \text{ conforme } n \to \infty.$$

Demostración. Observe que a partir de (3.3.4) se desprende que $\text{Var}\left[\alpha_n\right] \to 0$. Dado $\varepsilon > 0$, la desigualdad de Chebichev implica que

$$P[|\alpha_n - E[\alpha_n]| > \varepsilon] \leq \frac{\text{Var}\left[\alpha_n\right]}{\varepsilon^2} \to 0,$$

de manera que $\alpha_n - E[\alpha_n] \xrightarrow{\text{P}} 0$. Combinando este hecho con la convergencia $E[\alpha_n] - \alpha^* \to 0$ demostrada en el Teorema 3.2.1, via el teorema de Slutsky establecido en Dudewicz y Mishra (1988, p. 332), se desprende que $\alpha_n - \alpha^* \xrightarrow{\text{P}} 0$, esto es, $\alpha_n \xrightarrow{\text{P}} \alpha^*$. □

Convergencia Casi Segura

En esta sección se analiza la convergencia con probabilidad uno, o casi segura, de la sucesión de tasas de cobertura $\{\alpha_n\}$. Antes de continuar, es conveniente recordar la formulación precisa de esta idea.

Definition 3.4.1. Una sucesión de variables aleatorias $\{W_n\}$ *converge casi seguramente* a una constante a, si

$$P[\lim_{n\to\infty} W_n = a] = 1, \tag{3.4.1}$$

y en este caso se escribe $W_n \xrightarrow{\text{c.s.}} a$.

Note que $W_n \xrightarrow{\text{c.s.}} a$ equivale a $W_n - a \xrightarrow{\text{c.s.}} 0$. Por otro lado, la convergencia casi segura—también conocida cono *convergencia con probabilidad uno*—es una noción más fuerte que la idea de convergencia en probabilidad. De hecho, si una sucesión converge casi seguramente, entonces converge en probabilidad, pero el recíproco de esta afirmación no es válido. Por otro lado, el estudio de la convergencia casi segura es, generalmente, mucho más técnico que el análisis de la convergencia en probabilidad, la cual usualmente se obtiene aplicando la desigualdad de Chebichev, como en la sección precedente. Un instrumento que con frecuencia es útil para establecer la convergencia casi segura es el siguiente resultado conocido como *lema de Borel-Cantelli*, cuya demostración puede encontrarse, por ejemplo, en Billingsley (1999) , o en Lange (2005).

Lemma 3.4.1. Suponga que $\{W_n\}$ es una sucesión de variables aleatorias definidas en un mismo espacio de probabilidad. Si para cada $\varepsilon > 0$ se tiene que

$$\sum_{n=1}^{\infty} P[|W_n| > \varepsilon] < \infty,$$

entonces $W_n \xrightarrow{\text{c.s.}} 0$.

El principal objetivo de esta sección es establecer el siguiente resultado.

Teorema 3.4.1. Conforme n tiende a infinto

$$\alpha_n \xrightarrow{\text{c.s.}} \alpha^*.$$

La demostración de este teorema utiliza el siguiente resultado preliminar, de acuerdo al cual las tasas de cobertura convergen casi seguramente a α^* a través de un subsucesión adecuada.

Lemma 3.4.2. Conforme k tiende a infinito, $\alpha_{k^2} \xrightarrow{\text{c.s.}} \alpha^*$.

Demostración. A partir de la convergencia $n\text{Var}\left[\alpha_n\right] \to v$ establecida en el Teorema 3.3.1 se desprende que existe una constante positiva C tal que $n\text{Var}\left[\alpha_n\right] \leq C$, y entonces

$$\text{Var}\left[\alpha_n\right] \leq \frac{C}{n}, \quad n = 1, 2, 3, \ldots.$$

A partir de este hecho con $n = k^2$ se obtiene que, para todo $\varepsilon > 0$,

$$P[|\alpha_{k^2} - \alpha^*| > \varepsilon] \leq \frac{\text{Var}\left[\alpha_{k^2}\right]}{\varepsilon^2} \leq \frac{C}{k^2\varepsilon^2}$$

de manera que

$$\sum_{k=1}^{\infty} P[|\alpha_{k^2} - \alpha^*| > \varepsilon] \leq \sum_{k=1}^{\infty} \frac{C}{k^2\varepsilon^2} = \frac{C}{\varepsilon^2} \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty,$$

y entonces, una aplicación del l Lema 3.4.1 con $W_k = \alpha_{k^2} - \alpha^*$ se permite concluir que $\alpha_{k^2} - \alpha^* \xrightarrow{\text{c.s.}} 0$, *i.e.*, $\alpha_{k^2} \xrightarrow{\text{c.s.}} \alpha^*$. □

Demostración del Teorema 3.4.1. Dado un entero positivo n, sea $k = k(n)$ el único entero positivo que satisface

$$k^2 \leq n < (k+1)^2; \tag{3.4.2}$$

note que k depende de n y que

$$\lim_{n\to\infty} \frac{k^2}{n} = \lim_{n\to\infty} \frac{(k+1)^2}{n} = 1. \tag{3.4.3}$$

A continuación observe que, debido a que las variables de cobertura Y_i son no negativas,

$$\sum_{i=1}^{k^2} Y_i \leq \sum_{i=1}^{n} Y_i \leq \sum_{i=1}^{(k+1)^2} Y_i$$

lo cual es equivalente a

$$k^2 \alpha_{k^2} \leq n\alpha_n \leq (k+1)^2 \alpha_{(k+1)^2};$$

vea la Definición . Por lo tanto,

$$\frac{k^2}{n}\alpha_{k^2} \leq \alpha_n \leq \frac{(k+1)^2}{n}\alpha_{(k+1)^2}.$$

de manera que tomando el límite cuando n tiende a ∞, (3.4.3) implica que

$$\lim_{n\to\infty} \alpha_{k^2} \leq \lim_{n\to\infty} \alpha_n \leq \lim_{n\to\infty} \alpha_{(k+1)^2}.$$

Para concluir note que $k = k(n)$ tiende a infinito cuando n lo hace, por (3.4.2), de tal manera que $\lim_{n\to\infty} \alpha_{k^2} = \lim_{k\to\infty} \alpha_{k^2}$ y $\lim_{n\to\infty} \alpha_{(k+1)^2} = \lim_{k\to\infty} \alpha_{(k+1)^2}$; además, estos últimos límites son igual a α^* con probabilidad 1, por el Lemma 3.4.2. Por lo tanto, la última relación desplegada equivale a

$$\alpha^* \leq \lim_{n\to\infty} \alpha_n \leq \alpha^* \quad \text{con probabilidad 1,}$$

esto es, $P[\lim_{n\to\infty} \alpha_n = \alpha^*] = 1$, lo cual significa que $\alpha_n \xrightarrow{\text{c.s.}} \alpha^*$. □

De acuerdo al Teroema 3.4.1 La proporción de bases efectivamente analizadas converge a α^* conforme n se incrementa. Como la longitud g del genoma es 'grande', el investigador puede estar seguro de que, a pesar de que no conozca

la posición exacta de las bases analizadas, la proporción de bases que efectivamente son identificadas es muy cercana a α^*. El siguiente paso, que se llevará a cabo en la siguiente sección, es determinar un intervalo en el cual se ubique la proporción analizada α_g con una probabilidad especificada de antemano.

Distribución Límite

El objetivo de esta sección es mostrar que, conforme n se incrementa, la tasa de cobertura α_n adecuadamente estandarizada tiene una distribución aproximadamente normal. Antes de continuar, es conveniente recordar que una sucesión $\{W_n\}$ de variables aleatorias *converge en distribución* a la distribución normal con media cero y varianza v—denotada por $\mathcal{N}(0, v)$—si para todo $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P[W_n \leq x] = \Phi(x/v) \tag{3.5.1}$$

donde

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

es la función de distribución normal estándar. Cuando (3.5.1) ocurre para todo $x \in \mathbb{R}$ se escribe

$$W_n \xrightarrow{\mathrm{d}} \mathcal{N}(0, v)\,.$$

El resultado principal de esta sección se formula a continuación.

Teorema 3.5.1. Conforme $n \to \infty$

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

El instrumento básico para establecer este resultado es el siguiente teorema clásico, conocido como Teorema Central de Límite, cuya demostración puede encontrarse, por ejemplo, en Dudewicz y Mishra (1988), Billingley (1999), o Lange (2005).

Teorema 3.5.2. Sean $T_1, T_2, T_3, \ldots$ variables aleatorias independientes e idénticamente distribuidas con media μ y varianza v. En estas circunstancias,

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

Recordando que la tasa de cobertura α_n está dada por

$$\alpha_n = \frac{\sum_{i=1}^{n} Y_i}{n},$$

una comparación de las conclusiones de estos dos resultados conduce naturalmente a preguntar por qué no puede obtenerse el Teorema 3.5.1 directamente del Teorema 3.5.2. Hay dos razones para que esto no pueda hacerse. Primeramente, las variables de cobertura Y_i no son idénticamente distribuidas. En efecto, $Y_i \sim Ber(1-(1-p)^i)$ para $i < L$, mientras que $Y_i \sim Ber(1-(1-p)^L)$ para $i \geq L$; vea las partes (i) y (ii) del Lemma 2.4.1. Sin embargo, este es realmente un punto menor, puesto que para $n \geq L$, todas las variables Y_i tienen la misma distribución. La razón realmente importante por la que el Teorema 3.5.1 es interesante, es que las variables de cobertura *no son independientes*, como lo demuestra el hecho de que la covarianza entre dos variables distintas Y_i y Y_j no es necesariamente cero; vea el Lemma 2.4.1(iv). La estrategia para demostrar el Teorema 3.5.1 usará el Teorema Central de Límite después de hacer las adecuaciones necesarias. El punto de partida, es introducir las siguientes variables aleatorias auxiliares.

Definition 3.5.1. Sea r un entero positivo fijo. Para $k = 1, 2, 3, \ldots$, defina

$$\delta_k := \frac{1}{L} \sum_{i=(k-1)(r+1)L+1}^{k(r+1)L+L} Y_i \qquad \text{y} \quad \beta_k := \frac{1}{rL} \sum_{i=(k-1)(r+1)L+L+1}^{k(r+1)L} Y_i. \tag{3.5.2}$$

Antes de continuar es conveniente ahondar en la especificación (3.5.2). Primeramente, lo que se está haciendo, es dividir a las variables de cobertura en grupos

sucesivos de tamaño $(r+1)L$; los primeros cuatro grupos se muestran a continuación:

$$\underbrace{Y_1, Y_2, \ldots, Y_L}, Y_{L+1}, \ldots, Y_{(r+1)L}$$

$$\underbrace{Y_{(r+1)L+1}, Y_{(r+1)L+2}, \ldots, Y_{(r+1)L+L}}, Y_{(r+1)L+L+1}, \ldots, Y_{2(r+1)L}$$

$$\underbrace{Y_{2(r+1)L+1}, Y_{2(r+1)L+2}, \ldots, Y_{2(r+1)L+L}}, Y_{2(r+1)L+L+1}, \ldots, Y_{3(r+1)L}$$

$$\underbrace{Y_{3(r+1)L+1}, Y_{3(r+1)L+2}, \ldots, Y_{3(r+1)L+L}}, Y_{3(r+1)L+L+1}, \ldots, Y_{4(r+1)L}$$

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 Lr 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 el promedio β_j, y a partir del Lemma 2.4.1(iv) se desprende que $\beta_1, \beta_2, \beta_3, \ldots$ son independientes, mientras que la parte (v) del mismo lemma implica que éstas variables tienen la misma distribución. Similarmente, puede establecerse que las variables δ_k son independientes. Estas conclusiones se establecen formalmente en el siguiente lema.

Lemma 3.5.1. Con la notación en la Definición 3.5.1,

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

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

Lemma 3.5.2. Para cada $n > 2(r+1)L$, $\sqrt{n}[\alpha_n - \alpha^*]$ se representa como

$$\sqrt{n}[\alpha_n - \alpha^*] = Z_n + D_n + R_n, \tag{3.5.3}$$

donde las variables aleatorias en el lado derecho satisfacen las siguientes porpiedades (i)–(iii):

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

(ii) $E[D_n] = 0$ y

$$\text{Var}\,[D_n] \leq \frac{L}{(r+1)};$$

(iii) $Z_n \xrightarrow{\text{d}} \mathcal{N}(0, \tau_r)$, y

$$\tau_r = (r+1)L\text{Var}\left[\alpha'_{(r+1)L}\right]; \tag{3.5.4}$$

vea (3.3.5) para la Definición de las variables α'_k.

Demostración. Primeramente, dado $n > 2(r+1)L$, define $m = m(n)$ mediante

$$m = \left[\frac{n}{(r+1)L}\right] \tag{3.5.5}$$

donde $[a]$ denota la parte entera de a. Con esta notación se tiene que n puede representarse como

$$n = m(r+1)L + s, \qquad 0 \leq s < (r+1)L. \tag{3.5.6}$$

Por otro lado, note que (3.5.2) permite escribir

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

mientras que a partir de la Definición 2.4.2 y (3.5.6) se obtiene que

$$\begin{aligned}
n[\alpha_n - \alpha^*] &= \sum_{i=1}^{n}(Y_i - \alpha^*) \\
&= \sum_{k=1}^{m} \sum_{i=(k-1)(r+1)L+i}^{k(r+1)L} (Y_i - \alpha^*) + \sum_{i=m(r+1)L+1}^{k(r+1)L+s} (Y_i - \alpha^*) \\
&= \sum_{k=1}^{m} L[(\delta_k - \alpha^*) + r(\beta_k - \alpha^*)] + \sum_{i=m(r+1)L+1}^{k(r+1)L+s} (Y_i - \alpha^*) \\
&= Lr\sum_{k=1}^{m}(\beta_k - \alpha^*) + L\sum_{k=2}^{m}(\delta_k - \alpha^*) + L(\delta_1 - \alpha^*)r \\
&\qquad + \sum_{i=m(r+1)L+1}^{k(r+1)L+s} (Y_i - \alpha^*)
\end{aligned}$$

de manera que

$$\sqrt{n}[\alpha_n - \alpha^*] = Z_n + D_n + R_n,$$

donde

$$Z_n = \frac{Lr}{\sqrt{n}} \sum_{k=1}^{m} (\beta_k - \alpha^*), \qquad D_n = \frac{L}{\sqrt{n}} \sum_{k=2}^{m} (\delta_k - \alpha^*), \tag{3.5.7}$$

y

$$R_n = \frac{1}{\sqrt{n}} L(\delta_1 - \alpha^*) + \frac{1}{\sqrt{n}} \sum_{i=m(r+1)L+1}^{k(r+1)L+s} (Y_i - \alpha^*). \tag{3.5.8}$$

A continuación se probará que estas variables tienen las propiedades especificadas.

(i) Como las variables de cobertura Y_i y α^* pertenecen al intervalo $[0, 1]$, se desprende que δ_1, siendo un promedio de variables de cobertura, también se ubica entre cero y uno, y entonces $|\delta_1 - \alpha^*| \leq 1$ y $|Y_i - \alpha^*| \leq 1$, Por lo tanto,

$$\begin{aligned} |R_n| &\leq \frac{1}{\sqrt{n}} L|\delta_1 - \alpha^*| + \frac{1}{\sqrt{n}} \sum_{i=m(r+1)L+1}^{k(r+1)L+s} |Y_i - \alpha^*| \\ &\leq \frac{1}{\sqrt{n}} L + \frac{1}{\sqrt{n}} \sum_{i=m(r+1)L+1}^{k(r+1)L+s} 1 \\ &= \frac{L+s}{\sqrt{n}} \end{aligned}$$

Recordando que $0 \leq s < (r+1)L$ (vea (3.5.6)), se concluye que $|R_n| \leq L(r+2)/\sqrt{n}$.

(ii) Usando que $E[\delta_k] = \alpha^*$ para $k \geq 2$ (vea el Lemma 3.5.1), se desprende que $E[D_n] = 0$. Por otro lado, como δ_k se ubica entre cero y uno, se sigue que $\text{Var}\,[\delta_k] \leq 1$, de manera que la independencia de $\delta_2, \delta_3, \ldots$ permite establecer que

$$\begin{aligned} \text{Var}\,[D_n] &= \text{Var}\left[\frac{L}{\sqrt{n}} \sum_{k=2}^{m} (\delta_k - \alpha^*)\right] \\ &= \frac{L^2}{n} \sum_{k=2}^{m} \text{Var}\,[\delta_k] \leq \frac{L^2}{n} \sum_{k=2}^{m} 1 < \frac{L^2 m}{n}. \end{aligned}$$

Note ahora que

$$\frac{m}{n} \leq \frac{1}{(r+1)L},$$

por (3.5.6), de donde se desprende que $\text{Var}\,[D_n] \leq L^2/[(r+1)L] = L/(r+1)$.

(iii) Observe que

$$Z_n = \frac{Lr\sqrt{m}}{\sqrt{n}}\sqrt{m}\left[\frac{1}{m}\sum_{k=1}^{m}\beta_k - \alpha^*\right] \tag{3.5.9}$$

Por otro lado, por el Lemma 3.5.1, las variables β_k son independientes e idénticamente distribuidas con media α^* y varianza $\text{Var}\,[\beta_1]$, de manera que el Teorema 3.5.2 implica que

$$\sqrt{m}\left[\frac{1}{m}\sum_{k=1}^{m}\beta_k - \alpha^*\right] \xrightarrow{\text{d}} \mathcal{N}\left(0, \text{Var}\,[\beta_1]\right).$$

Además, la expresión (3.5.6) muestra que n tiende a infinito si y sólo si m lo hace, y en ese caso, $m/n \to 1/[L(r+1)]$, de manera que

$$\frac{Lr\sqrt{m}}{\sqrt{n}} \to r\sqrt{\frac{L}{r+1}}$$

Combinando las dos últimas relaciones desplegadas con la expresión (3.5.9) para Z_n, el Teorema de Slustky en Dudewicz y Mishra (1980, p.323) permite concluir que

$$Z_n \xrightarrow{\text{d}} \mathcal{N}\left(0, \frac{Lr^2}{r+1}\text{Var}\,[\beta_1]\right) = \mathcal{N}\left(0, \tau_r\right),$$

donde

$$\tau_r = \frac{Lr^2}{r+1}\text{Var}\,[\beta_1].$$

Para concluir, es suficiente demostrar que τ_r es como se especifica en (3.5.4). Con este fin, observe que a partir de (3.3.1) y la Definición 3.5.1 se desprende que

$$\beta_1 = \frac{1}{rL}\sum_{i=L+1}^{(r+1)L} Y_i, \quad \text{y} \quad \alpha'_{(r+1)L} = \frac{1}{(r+1)L}\sum_{i=L+1}^{(r+1)L} Y_i.$$

Por lo tanto,

$$\beta_1 = \frac{r+1}{r}\alpha'_{(r+1)L}$$

y entonces

$$\text{Var}\,[\beta_1] = \left(\frac{r+1}{r}\right)^2 \text{Var}\left[\alpha'_{(r+1)L}\right]$$

Por lo tanto,

$$\tau_r = \frac{Lr^2}{r+1}\left(\frac{r+1}{r}\right)^2 \text{Var}\left[\alpha_{(r+1)L}\right] = L(r+1)\text{Var}\left[\alpha'_{(r+1)L}\right],$$

concluyendo la demostración. □

El siguiente lemma es la última etapa antes de la demostración del Teorema 3.5.1.

Lemma 3.5.3. Dado $\varepsilon \in (0,1)$, sea n tal que

$$n > \left(\frac{(r+2)L}{\varepsilon}\right)^2, \tag{3.5.10}$$

y descomponga $\sqrt{n}[\alpha_n - \alpha^*]$ como en (3.5.3). En este caso, las siguientes desigualdades son válidas:

$$P[\sqrt{n}[\alpha_n - \alpha^*] \leq x] \leq P[Z_n \leq x + 2\varepsilon] + \frac{L}{(r+1)\varepsilon^2}. \tag{3.5.11}$$

$$P[\sqrt{n}[\alpha_n - \alpha^*] \leq x] \geq P[Z_n \leq x - 2\varepsilon] - \frac{L}{(r+1)\varepsilon^2}. \tag{3.5.12}$$

Demostración. Como punto de partida, note que (3.5.10) implica

$$(r+2)L/\sqrt{n} \leq \varepsilon,$$

de donde se desprende que

$$|R_n| \leq \varepsilon. \tag{3.5.13}$$

Ahora, dado $x \in \mathbb{R}$, observe que

$$\begin{aligned}\sqrt{n}[\alpha_n - \alpha^*] \leq x &\iff Z_n + D_n \leq x - R_n \\ &\implies Z_n + D_n \leq x + \varepsilon\end{aligned}$$

donde se utilizó (3.5.13) para establecer la implicación. Por lo tanto

$$\begin{aligned}
\left[\sqrt{n}[\alpha_n - \alpha^*] \leq x\right] &\subset [Z_n + D_n \leq x + \varepsilon] \\
&= [Z_n + D_n \leq x + \varepsilon, \ \ |D_n| \leq \varepsilon] \\
&\qquad \cup [Z_n + D_n \leq x + \varepsilon, \ \ |D_n| > \varepsilon] \\
&= [Z_n \leq x + \varepsilon - D_n, \ \ |D_n| \leq \varepsilon] \cup [|D_n| > \varepsilon] \\
&\subset [Z_n \leq x + \varepsilon + \varepsilon, \ \ |D_n| \leq \varepsilon] \cup [|D_n| > \varepsilon] \\
&\subset [Z_n \leq x + 2\varepsilon] \cup [|D_n| > \varepsilon]
\end{aligned}$$

Por lo tanto

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

Observe ahora que la desigualdad de Chebichev y el Lemma 3.5.2(ii) implican que

$$P[|D_n| > \varepsilon] \leq \frac{\mathrm{Var}\left[D_n\right]}{\varepsilon^2} \leq \frac{L}{(r+1)\varepsilon^2} \tag{3.5.14}$$

y combinando estas dos últimas relaciones se obtiene

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

estableciendo (3.5.11). El argumento para demostrar (3.5.12) es similar. Note que a partir de (3.5.13) se obtiene

$$Z_n \leq x - 2\varepsilon \Longrightarrow Z_n + R_n \leq x - \varepsilon.$$

Por lo tanto,

$$\begin{aligned}
[Z_n \leq x - 2\varepsilon] &\subset [Z_n + R_n \leq x - \varepsilon] \\
&= [Z_n + R_n + Dn \leq x - \varepsilon + D_n] \\
&= [Z_n + R_n + Dn \leq x - \varepsilon + D_n, \ |D_n| \leq \varepsilon] \\
&\qquad \cup [Z_n + R_n + Dn \leq x - \varepsilon + D_n, |D_n| > \varepsilon] \\
&\subset [Z_n + R_n + Dn \leq x, \ |D_n| \leq \varepsilon] \\
&\qquad \cup [|D_n| > \varepsilon] \\
&\subset [Z_n + R_n + Dn \leq x] \cup [|D_n| > \varepsilon]
\end{aligned}$$

Por lo tanto,

$$[Z_N \leq x - 2\varepsilon] \subset [\sqrt{n}[\alpha_n - \alpha^*] \leq x] \cup [|D_n| > \varepsilon],$$

y entonces, via (3.5.3) se desprende que

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

y (3.5.12) se desprende a partir de esta relación usando (3.5.14). □

Después de los resultados en los lemas anteriores, el resultado del Teorema 3.5.1 puede establecerse como sigue.

Demostración del Teorema 3.5.1. Dado $\varepsilon \in (0,1) > 0$, sea $n > (r+2)^2 L^2/\varepsilon^2$, de manera que las desigualdades (3.5.11) y (3.5.12) en el Lemma 3.5.3 son válidas. Tomando límite conforme n tiende a ∞ en las mencionadas desigualdades, se desprende que

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

Recordando que $Z_n \xrightarrow{\mathrm{d}} \mathcal{N}(0, \tau_r)$, por el Lemma 3.5.2(iii), se tiene que

$$P[Z_n \leq x + \varepsilon] \to \Phi((x+\varepsilon)/\tau_r),$$

y similarmente, $P[Z_n \leq x - \varepsilon] \to \Phi((x-\varepsilon)/\tau_r)$. Por lo tanto,

$$\Phi\left(\frac{x-\varepsilon}{\tau_r}\right) - \frac{C}{(r+1)\varepsilon^2} \leq \lim_{n\to\infty} P[\sqrt{n}[\alpha_n - \alpha^*] \leq x] \leq \Phi\left(\frac{x+\varepsilon}{\tau_r}\right) + \frac{C}{(r+1)\varepsilon^2}. \tag{3.5.15}$$

Por otro lado, de acuerdo al Lemma 3.3.1, $\lim_{n\to\infty} n\mathrm{Var}\,[\alpha'_n] = v$, de manera que

$$(r+1)L\mathrm{Var}\left[\alpha'_{(r+1)L}\right] \to v \quad \text{si } r \to \infty,$$

de manera que a partir de la fórmula (3.5.4) 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 límite cuando r tiende a ∞ en (3.5.15) se obtiene que

$$\Phi\left(\frac{x-\varepsilon}{v}\right) \leq \lim_{n\to\infty} P[\sqrt{n}[\alpha_n - \alpha^*] \leq x] \leq \Phi\left(\frac{x+\varepsilon}{v}\right).$$

y tomando límite en esta relación conforme ε tiende a cero por la derecha, se arriba a

$$\Phi\left(\frac{x}{v}\right) \leq \lim_{n\to\infty} P[\sqrt{n}[\alpha_n - \alpha^*] \leq x] \leq \Phi\left(\frac{x}{v}\right),$$

de manera que $P[\sqrt{n}[\alpha_n - \alpha^*] \leq x] \to \Phi(x/v)$; como esta convergencia es válida para todo $x \in \mathbb{R}$, se desprende que $\sqrt{n}[\alpha_n - \alpha^*] \xrightarrow{\mathrm{d}} \mathcal{N}(0, v)$. □

A partir del Teorema es posible establecer un intervalo de credibilidad para la tasa de cobertura α_n: Denotando por $z_{c/2}$ al percentil derecho de orden $c/2$ para la distribución normal estándar, se tiene que, para n 'grande',

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

y como el tamaño del genoma es 'grande', se tiene que α_g se ubica, con probabilidad aproximadamente $1 - c$, dentro del intervalo $\alpha^* \pm vz_{c/2}/\sqrt{g}$.

Capítulo 4

Bloques

En este capítulo se abordan los dos problemas que se plantearon inicialmente sobre los bloques de segmentos acoplables, a saber, encontrar la longitud esperada de un bloque, y el número esperado de bloques que se generan como producto del proceso de muestreo.

Introducción

Este capítulo trata sobre el problema de determinar el número esperado $\mathcal{N}$ de bloques y la longitud media $\mathcal{L}$ de cada uno de ellos. Para contar el número de bloques, en la sección 2 se introduce una familia de variables indicadoras, las que al tomar el valor uno señalan el inicio de un bloque. Posteriormente, en la Sección 3 estas variables son utilizadas para encontrar el número esperado de bloques $\mathcal{N}$, mientras que el problema de determinar su tamaño esperado $\mathcal{L}$ se aborda en las secciones 4 y 5. Primeramente, en la Sección 4 se formula una expresión para la longitud de un bloque, la cual muestra que ésta puede espresarse como la suma de un número aleatorio de variables independientes con distribución geométrica común. Este resultado se utiliza para conluir el capítulo en la Sección 5 encontrando la longitud esperada de un bloque

Indicadores de Inicio

El objetivo de esta sección es definir las variables que indican el inicio de un bloque, las cuales se utilizarán posteriormente para contar el número esperado de bloques y su longitud esperada. Sea

$$B = (\mathcal{I}_{k_1}, \mathcal{I}_{k_2}, \ldots, \mathcal{I}_{k_r})$$

un bloque en el sentido de la Definición 2.5.1. En este caso, el intervalo inicial $\mathcal{I}_{k_1}$ es no vacío, de modo que

$$X_{k_1} = 1 \tag{4.2.1}$$

y el inicio del bloque se ubica en la posición k_1. Esto es así porque I_{k_1} no puede acoplarse con ningún intervalo a su izquierda, esto es, $|\mathcal{I}_s \cap \mathcal{I}_{k_1}| < \ell$ para $s < k_1$. Puesto que $\mathcal{I}_s = \emptyset$ si $X_s = 0$, se desprende que si $X_s = 1$ para $s < k_1$, entonces $[\![s, s+L) \cap [\![k_1, k_1+L)$ tiene menos de ℓ elementos, lo que significa que $s+L-k_1 < \ell$, esto es $s < k_1 - (L - \ell)$. Por lo tanto, si s es un entero positivo tal que $k_1 > s \geq k_1 - (L-\ell)$, entonces debe tenerse que $X_s = 0$, lo cual puede expresarse de manera más compacta como

$$X_s = 0, \quad k_1 - (L - \ell) \leq s < k_1, \quad s \geq 1. \tag{4.2.2}$$

Las condiciones (4.2.1) y (4.2.2) aseguran que $\mathcal{I}_{k_1}$ es no vacío y que este intervalo no puede acoplarse con ninguno de los que le preceden. Una vez que esto se tiene, $\mathcal{I}_{k_1}$ se acoplará con el primer intervalo $\mathcal{I}_{k_2}$ que le suceda y que se intersecte con él en por lo menos ℓ elementos. El proceso de acoplamiento se repite con k_2 en lugar de k_1, y se continúa hasta que no sea posible acoplar más intervalos, dando por finalizada la construcción del bloque. Por lo tanto, la ocurrencia simultánea de (4.2.1) y (4.2.2) señala el inicio de un bloque, por lo cual es razonable la siguiente definición.

Definition 4.2.1. La sucesión $\{B_i\}$ de variables aleatorias *indicadoras de bloque* está dada por

$$B_i = I[X_i = 1, X_s = 0,\ \max\{1, i-(L-\ell)\} \leq s < i], \quad i = 1, 2, 3, \ldots.$$

Observación 4.2.1. Suponga que $j > (L-\ell)+i$. De acuerdo a la definición anterior, B_i es una función de variables aleatorias X_s con $s \leq i$, mientras que B_j depende de $X_{j-(L-\ell)}, X_{j-(L-\ell)+1}, \ldots, X_j$. Luego, la condición $j > (L-\ell)+i$ asegura que B_i y B_j dependen de grupos disjuntos de variables X_s, y entonces B_i y B_j son independientes, por (2.2.1). Verbalmente, si los índices de dos variables indicadoras de bloque difieren por más de $L-\ell$ unidades, entonces son independientes.

Los valores esperados y las covarianzas entre las variables indicadoras de bloque se determinan a continuación.

Lemma 4.2.1. Para enteros positivos i y j

$$E[B_i] = \mu_i := \begin{cases} p(1-p)^i, & \text{si } i \leq L-\ell \\ p(1-p)^{L-\ell}, & \text{si } i > L-\ell, \end{cases}$$

mientras que

$$\text{Cov}\,[B_i, B_j] = \begin{cases} \mu_i(1-\mu_i), & \text{si } i = j \\ -\mu_i\mu_j, & \text{si } 0 < |i-j| \leq L-\ell \\ 0, & \text{si } |i-j| > L-\ell. \end{cases}$$

Demostración. De acuerdo a la Definición 4.2.1, si $i < L-\ell$ entonces

$$B_i = I[X_i = 1, X_s = 0, \quad 1 \leq s < i]$$

de manera que $E[B_i] = p(1-p)^{i-1}$, por (2.2.1). Cuando $i > L-\ell$, se tiene que

$$B_i = I[X_i = 1, X_s = 0, \quad i-(L-\ell) \leq s < i]$$

y entonces $E[B_i] = p(1-p)^{L-\ell}$, estableciendo el resultado para el valor esperado de las variables B_i. Por otro lado, como B_i es una variable de Bernoulli con media

μ_i, se tiene que $\text{Cov}\,[B_i, B_i] = \text{Var}\,[B_i] = \mu_i(1-\mu_i)$, mientras que a partir de la Observación 4.2.1 se desprende que $\text{Cov}\,[B_i, B_j] = 0$ cuando $|i-j| > L-\ell$. Para concluir la demostración sólo resta establecer la fórmula para $\text{Cov}\,[B_i, B_j]$ cuando $|i-j| \leq L-\ell$. Con este fin, suponga que $|i-j| \leq L-\ell$ y, sin pérdida de generalidad, que $j \geq i$, de manera que

$$0 < j - i \leq L - \ell. \tag{4.2.3}$$

Se mostrará que en este caso se tiene que $B_i B_j = 0$. En efecto, si $B_i = 0$ este producto es claramente nulo, mientras que si $B_i = 1$, entonces $X_i = 1$, por la Definición 4.2.1, y debido a que (4.2.3) implica que $j-(L-\ell) \leq i < j$, se desprende que el evento $[X_j = 1, X_s = 0,\ L-\ell \leq s < j]$ no ocurre y por lo tanto $B_j = 0$, y entonces $B_i B_j = 0$. Por lo tanto, $E[B_i B_j] = 0$, de manera que $\text{Cov}\,[B_i, B_j] = E[B_i B_j] - E[B_i]E[B_j] = -\mu_i \mu_j$.

Número Esperado de Bloques

Para cada n positivo, defina el número total de bloques hasta la posición n mediante

$$T_n = \sum_{i=1}^{n} B_i, \quad n = 1, 2, 3, \ldots. \tag{4.3.1}$$

El propósito de esta sección es establecer la siguiente fórmula para la esperanza de T_n.

Teorema 4.3.1. El valor esperado de T_n está dado por

$$E[T_n] = \begin{cases} 1-(1-p)^n, & \text{si } n \leq L-\ell, \\ 1-(1-p)^L + [n-(L-\ell)]p(1-p)^{L-\ell}, & \text{si } n > L-\ell. \end{cases}$$

Demostración. Considere el caso $n \leq L-\ell$. De acuerdo al Lemma 4.2.1, $E[B_i] = p(1-p)^i$ para $i \leq n$, y entonces $E[T_n] = \sum_{i=1}^{n} E[B_i] = \sum_{i=1}^{n} p(1-p)^{i-1}$; usando la fórmula $\sum_{i=1}^{n} r^{i-1} = (1-r^i)/(1-r)$ con $(1-p)$ en lugar de r, se desprende

que $E[T_n] = [1-(1-p)^n]$. Ahora suponga que $n > L-\ell$. En estas circunstancia $T_n = T_{L-\ell} + \sum_{i=L-\ell+1}^{n} B_i$, y como $E[B_i] = p(1-p)^{L-\ell}$, se tiene que

$$\begin{aligned} E[T_n] &= E[T_{L-\ell}] + \sum_{i=L-\ell+1}^{n} E[B_i] \\ &= (1-(1-p)^{L-\ell}) + \sum_{i=L-\ell+1}^{n} p(1-p)^{L-\ell} \\ &= (1-(1-p)^{L-\ell}) + [n-(L-\ell)]p(1-p)^{L-\ell} \end{aligned}$$

concluyendo el argumento. □

Observación 4.3.1. Defina la fracción de acoplamiento f mediante

$$f = \frac{\ell}{L}.$$

y recuerde que la intensidad de cobertura esta dada por $\kappa = Lp$. Observando que

$$(1-p)^{L-\ell} = \left(1 - \frac{Lp}{L}\right)^{L(1-f)} = \left(1 - \frac{\kappa}{L}\right)^{L(1-f)} \approx e^{-\kappa(1-f)},$$

se obtiene la siguiente aproximaxión para $E[T_n]$:

$$E[T_n] \approx 1 - e^{-\kappa(1-f)} + [np - \kappa(1-f)](e^{-\kappa(1-f)}.$$

Una Fórmula Para la Longitud

En esta sección se determina una expresión para la longitud de un bloque, la cual tome en cuenta que lo intervalos que lo conforman son aleatorios y se utilizará, posteriormente, para calcular la longitud esperada. En adelante, se supone que $\mathcal{I}_{k_0}$ es el intervalo inicial del bloque, esto es, $X_{k_0} = 1$ y $X_s = 0$ si $k_0 - (L-\ell) \leq s < k_0$. Como punto de partida, es conveniente introducir la siguiente notación.

Definition 4.4.1. La sucesión de variables aleatorias $T_1, T_2, T_3, \ldots$, se define como sigue:

$$\begin{aligned}
T_1 &:= \min\{n > 0\, 1\ X_{k_0+n} = 1\};\\
T_2 &:= \min\{n > 0\, 1\ X_{k_0+T_1+n} = 1\};\\
T_3 &:= \min\{n > 0\, 1\ X_{k_0+T_1+T_2+n} = 1\};\\
&\vdots\\
T_k &:= \min\{n > 0\, 1\ X_{k_0+T_1+\cdots+T_{k-1}+n} = 1\};\\
&\vdots
\end{aligned}$$

Pensando en que $X_i = 1$ corresponde a 'éxito', $T_1, T_2, \ldots$ es la sucesión de tiempos de espera entre éxitos sucesivos posteriores a k_0, y entonces se tiene el siguiente resultado, cuya demostración puede verse en Dudewicz y Mishra (1988).

Lemma 4.4.1. (i) Las variables aleatorias $T_1, T_2, T_3, \ldots$ son independientes e idénticamente distribuídas, y su distribución común es geométrica con parámetro p, esto es,

$$P[T_k = m] = (1-p)^{m-1}p, \quad m = 1, 2, 3, \ldots$$

(ii) La sucesión S_k (de tiempos sucesivos de éxito después de k_0) está dada

$$S_1 := T_1, \quad S_k = T_1 + \cdots T_k, \quad k \geq 2. \tag{4.4.1}$$

Note que, dentro de los intervalos posteriores a $\mathcal{I}_{k_0}$, los que son diferentes del vacío son

$$\mathcal{I}_{k_0+S_1}, \mathcal{I}_{k_0+S_2}, \mathcal{I}_{k_0+S_3}, \ldots.$$

A continuación, se encuentra la longitud de una unión de intervalos consecutivos, para los cuales la intersección de intervalos sucesivos es no nula.

Lemma 4.4.2. Suponga que la sucesión

$$(\mathcal{I}_{k_0}, \mathcal{I}_{k_0+S_1}, \ldots, \mathcal{I}_{k_0+S_{r-1}}) \tag{4.4.2}$$

es tal que

$$\mathcal{I}_{k_0} \cap \mathcal{I}_{k_0+S_1} \neq \emptyset, \quad \mathcal{I}_{k_0+S_1} \cap \mathcal{I}_{k_0+S_2} \neq \emptyset, \ \cdots, \mathcal{I}_{k_0+S_{r-2}} \cap \mathcal{I}_{k_0+S_{r-1}} \neq \emptyset.$$

En este caso, la unión

$$U = \mathcal{I}_{k_0} \cup \mathcal{I}_{k_0+S_1} \cup \cdots \cup \mathcal{I}_{k_0+S_{r-1}}) \tag{4.4.3}$$

está dada por

$$U = [\![k_0, k_0 + S_{r-1} + L)$$

y por lo tanto, su longitud es

$$|U| = S_{r-1} + L.$$

Demostración. Sólo recuerde que

$$\begin{aligned}
\mathcal{I}_{k_0} &= [\![k_0, k_0 + L), \\
\mathcal{I}_{k_0+S_1} &= [\![k_0 + S_1, k_0, K_0 + S_1 + L) \\
\mathcal{I}_{k_0+S_2} &= [\![k_0 + S_2, k_0, K_0 + S_2 + L) \\
&\vdots \\
\mathcal{I}_{k_0+S_{r-1}} &= [\![k_0 + S_{r-1}, k_0, K_0 + S_{r-1} + L).
\end{aligned}$$

Puesto que dos intervalos sucesivos en la sucesión (4.4.2) se intersectan, se desprende que

$$k_0 \leq k_0 + S_1 \leq k_0 + L,$$

y

$$k_0 + S_{i+1} \leq k_0 + S_i + L \leq k_0 + S_{i+1} + L, \quad i = 0, 1, \ldots, r-1.$$

lo cual implica que la unión (4.4.3) es igual a $[\![k_0, k_0 + S_{r-1} + L)$, y por lo tanto su longitud es $(k_0 + S_{r-1} + L) - k_0 = S_{r-1} + L$. □

Un bloque B en el sentido de la Definición 2.5.1 es, ante todo, una sucesión de fragmentos consecutivos, tales que dos segmentos sucesivos se intersectan, así que la longitud de B puede calcularse a partir del Lemma 4.4.2, siempre y cuando se disponga del valor exacto de r, el número de intervalos que componen al bloque. Para caracterizar este número, considere la sucesión

$$B = (\mathcal{I}_{k_0}, \mathcal{I}_{k_0+S_1}, \ldots, \mathcal{I}_{k_0+S_{r-1}}).$$

Si B es un bloque se tiene que (i) dos segmentos sucesivos se acoplan, esto es, su intersección contiene ℓ o más elementos, y (ii) el segmento $\mathcal{I}_{k_0+S_{r-1}}$ no puede acoplarse con el siguiente fragmento no nulo, el cual está dado por $\mathcal{I}_{k_0+S_r}$, pues la intersección de ambos contiene *menos* de ℓ elementos. Estas dos condiciones permiten caracterizar r como sigue:

(i) Puesto que $\mathcal{I}_{k_0} \cap \mathcal{I}_{k_0+S_1} = [\![k_0 + S_1, k_0 + L)$ y $\mathcal{I}_{k_0+S_{i-1}} \cap \mathcal{I}_{k_0+S_i} = [\![k_0 + S_i, k_0 + S_{i-1} + L)$, el hecho de que estas intersecciones tengan ℓ o más elementos siginifica que

$$k_0 + L - (k_0 + S_1) \geq \ell$$

$$k_0 + S_{i-1} + L - (k_0 + S_i) \geq \ell, \quad i = 2, 3, \ldots, r-1$$

esto es, $S_1 \leq L - \ell$, y $S_i - S_{i-1} \leq L - \ell$; como $S_1 = T_1$ y $S_i - S_{i-1} = T_i$, se desprende que

$$T_i \leq L - \ell, \quad i = 1, 2, \ldots, r-1. \tag{4.4.4}$$

(ii) Como $\mathcal{I}_{k_0+S_{r-1}} \cap \mathcal{I}_{k_0+S_r} = [\![k_0 + S_r, k_0 + S_{r-1} + L)$ tiene menos de ℓ elementos, se desprende que $k_0 + S_{r-1} + L - (k_0 + S_r) < \ell$, de tal forma que

$$T_r = S_r - S_{r-1} > L - \ell.$$

Esta desigualdad, conjuntamente con (4.4.4), caracteriza a r como el primer entero k tal que T_k excede a $L - \ell$. En símbolos,

$$r = \min\{k \,|\, T_k > L - \ell\}. \tag{4.4.5}$$

Note que r es *una variable aleatoria*, y que la desigualdad $r > i$, significa que $T_1, T_2, \ldots, T_i$ son todos menores o iguales a $L - \ell$, esto es,

$$[r > k] = [T_s \leq L - \ell,\ s = 1, 2, \ldots k]. \tag{4.4.6}$$

El resultado central de esta sección es el siguiente.

Teorema 4.4.1. Sea

$$B = (\mathcal{I}_{k_0}, \mathcal{I}_{k_0+S_1}, \ldots, \mathcal{I}_{k_0+S_{r-1}}),$$

un bloque en el sentido de la Definición 2.5.1. En este caso, su longitud es

$$\|B\| = L + \sum_{k=1}^{\infty} T_k I[T_s \leq L - \ell,\ s = 1, 2, \ldots k]. \tag{4.4.7}$$

Demostración. La longitud de B es $\|B\| = L + S_{r-1}$, por el Lemma 4.4.2, esto es,

$$\|B\| = \sum_{k=1}^{r-1} T_k = \sum_{k=1}^{\infty} T_k I[r > k];$$

note que en la última sumatoria, $I[r > k]$ se anula cuando $k = r, r+1, r+2, \ldots$. De acuerdo a la discusión previa, r esta dado por (4.4.5), y sustituyendo (4.4.6) en la última sumatoria se obtiene la fórmula (4.4.7). □

Tamaño Esperado de un Bloque

El propósito de esta sección es establecer la fórmula para la longitud esperada de un bloque dada en el siguiente teorema.

Teorema 4.5.1. La esperanza de la longitud del bloque

$$B = (\mathcal{I}_{k_0}, \mathcal{I}_{k_0+S_1}, \ldots, \mathcal{I}_{k_0+S_{r-1}}),$$

está dada por

$$E[\|B\|] = L + \frac{(1-p)^{-(L-\ell)} - 1}{p} - \frac{L-\ell}{1-p}. \tag{4.5.1}$$

Demostración. Aplicando el Teorema 4.4.1 se tiene que

$$E[\|B\|] = L + \sum_{k=1}^{\infty} E\left[T_k I[T_s \leq L-\ell,\ s = 1, 2, \ldots k]\right].$$

Por otro lado, como las variables T_i son independientes y tienen distribución común,

$$\begin{aligned} &E\left[T_k I[T_s \leq L-\ell,\ s = 1, 2, \ldots k]\right] \\ &\quad = E\left[T_k I[T_k \leq L-\ell] I[T_s \leq L-\ell,\ s = 1, 2, \ldots k-1]\right] \\ &\quad = E\left[T_k I[T_k \leq L-\ell]\right] E[I[T_s \leq L-\ell,\ s = 1, 2, \ldots k-1]] \\ &\quad = E\left[T_1 I[T_1 \leq L-\ell]\right] P[T_1 \leq L-\ell]^{k-1} \end{aligned}$$

y entonces

$$\begin{aligned} E[\|B\|] &= L + E\left[T_1 I[T_1 \leq L-\ell]\right] \sum_{k=1}^{\infty} P[T_1 \leq L-\ell]^{k-1} \\ &= L + \frac{E\left[T_1 I[T_1 \leq L-\ell]\right]}{1 - P[T_1 \leq L-\ell]}. \end{aligned}$$

Además, debido a que las variables T_i tienen distribución geométrica, se desprende que

$$1 - P[T_1 \leq L-\ell] = P[T_1 > L-\ell] = (1-p)^{L-\ell},$$

mientras que

$$\begin{aligned} E\left[T_1 I[T_1 \leq L-\ell\right] &= \sum_{k=1}^{L-\ell} pk(1-p)^{k-1} \\ &= p \sum_{k=1}^{L-\ell} k(1-p)^{k-1} \\ &= p\left[\frac{1-(1-p)^{L-\ell}}{p^2} - \frac{(L-\ell)(1-p)^{L-\ell-1}}{p}\right] \end{aligned}$$

donde en el último paso se utilizó la fórmula $\sum_{k=1}^{n} r^{k-1} = (1-r^n)/(1-r)^2 - nr^{n-1}/(1-r)$ con $L-\ell$ y $1-p$ en lugar de n y r, respectivamente. Combinando las tres últimas relaciones desplegadas se obtiene la fórmula (4.5.1). □

Observación 4.5.1. En términos de la tasa de cobertura $\kappa = Lp$ y de la fracción de acoplamiento $f = \ell/L$, la longitud esperada de un bloque está dada, aproximadamente, por

$$E[\|B\|] \approx L + L\frac{e^{\kappa(1-f)} - 1}{\kappa} - \frac{L^2(1-f)}{L-\kappa}.$$

LITERATURA CITADA

B. Derrida, T. Fink. (2002). Sequence determination from overlapping fragments: a simple model of whole-genome shotgun sequencing, *Physical Revision Letter*, **88**, No. 6, 068106.

D. Altshuler, V. Pollara y C. R. Pollara (2000). A map of the human genome generated by reduced representation shotgun sequencing, *Nature*, **407**, 513–516.

E. Lander , M. Waterman (1998). Genomic mapping by fingerprinting random clones: a mathematical analysis, *Genomics*, **2**, No.3, 231-239.

E. Dudewicz y S. N. Mishra (1988). *Mothern Mathematical Statistics*, *John WIley & Sons*, New York.

E. Myers (1999). Whole-Genome DNA Sequencing, *IEEE Computational Engineering and Science*, **3**, No. 1, 33-43

G. Karp (1998) Biologa Celular y Molecular, *McGraw-Hill Interamericana.*

G. Casella y R. Berger (2001). *Statistical Inference*, *Duxbury Press*, Boston.

I. Dunham (2003). *Genome Mapping and Sequencing*, *Horizon Scientific Press*,

Cambridge.

J.Venter (1996). A new strategy for genome sequencing, *Nature*, **381**, No. 6581, 364-366.

J.Venter (2001). The Sequence of the Human Genome *Science*, **291**, 1304-1351.

J. Weber,E. Myers. (1997). Human whole-genome shotgun sequencing, *Genome Research*, **7**, No. 5, 401-409.

K. Lange (2003). *Applied Probability*, *Springer–Verlag*, New York.

M. Pop (2002). Genome Sequence Assembly: Algorithms and Issues, *IEEE Computational Engineering and Science*, **35**, No. 7, 47-54.

P. Green. (1997). Against a whole-genome shotgun, *Genome Research*, **7**, No. 5, 410-407.

P. Billingsley (1999). *Probability and Measure*, *John Wiley & Sons*, New York.

P. J. Brockwell y R. A. Davis (1998). *Time Series: Theory and Methods*, *Springer–Verlag*, New York.

P. J. Brockwell y R. A. Davis (2002). *Introduction to Time Series and Forecasting*, *Springer–Verlag*, New York.

R. Weaver (2005) Molecular Biology, Third Edition, *McGraw-Hill*.

R. H. Waterson, E. S. Lander y J. E. Suston (2002). On the sequencing of human genome,*Proceedings of the National Academy of Sciences , USA*, **99**, No. 6, 3712–3716.

S. Batzoglou ,D. Jaffe ,K. Stanley,J. Butler, (2002). ARACHNE: a whole-genome shotgun assembler, *Genome Research*, **12**, No. 1, 177-189

S. Istrail, G.G. Sutton L. Florea (2004). Whole-genoma shotgun assembly and comparison of human genome assemblies, *Proceedings of the National Academy of Science, USA*, **101**, No. 7, 1916–1921.

S. Schbath (1997). Coverage processes in physical mapping by anchoring random clones, *Journal of computational biology*, **4**, No. 1, 61-82.

T. Koski, (2002). *HiddenMarkov Models in Bioinformatics*, *Kluwer Academic*, New York.

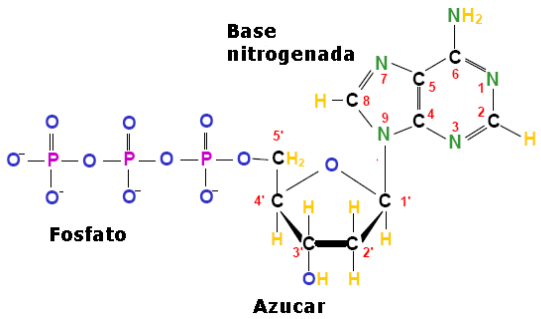
W. J. Ewnes y G. Grant (2005). *Statistical Methods in Bioinformatics: An introduction*, Second Edition, *Springer–Verlag*, New York.

Apéndice

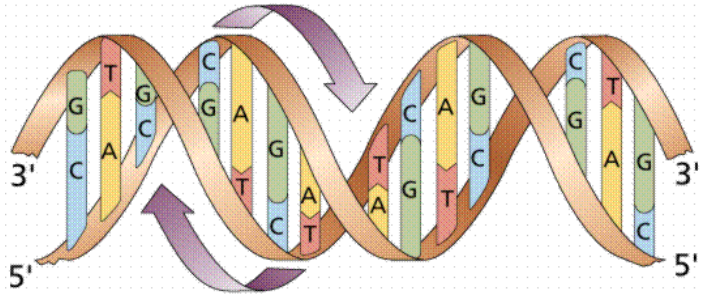
Este apéndice tiene como finalidad dar un panorama general sobre la estructura del genoma y su importancia para los seres vivos, además se amplía la información respecto al fundamento biológico de la técnica de secuenciación genómica total

Estructura y Función del Genoma

El Genoma es la secuencia total de ácido desoxirribonucleico (DNA) de un organismo y se localiza en el núcleo de cada una de sus células. La molécula de DNA es una doble cadena en forma de espiral, cuyas unidades son los nucleótidos constituidos por un azúcar, un fosfato y una base nitrogenada (G Karp, 1998).



Existen 4 tipo de nucleótidos que conforman el DNA, la diferencia entre ellos es el tipo de base nitrogenada que presentan: adenina, guanina, timina o citosina. En la figura 1 se muestra un esquema del DNA, la forma de la molécula se puede visualizar como una escalera en espiral en la que el azúcar y el fosfato de cada nucleótido son el contorno y las bases nitrogenadas son los escalones.



Las cadenas se diferencian por el extremo terminal que presentan, una cadena se sitúa en dirección 3' a 5' y la otra en dirección 5' a 3'.
Existe una unión específica entre los nucleótidos de ambas cadenas, la adenina de una cadena se une a la timina de la otra cadena y la guanina se une con la citosina de la cadena opuesta, este tipo de unión es importante en la técnica de secuenciación, pues solo se requiere conocer la secuencia de una de las cadenas para inferir la opuesta..

La importancia de conocer la secuencia completa de DNA, es decir el genoma, radica en el papel fundamental de ésta molécula en los seres vivos.

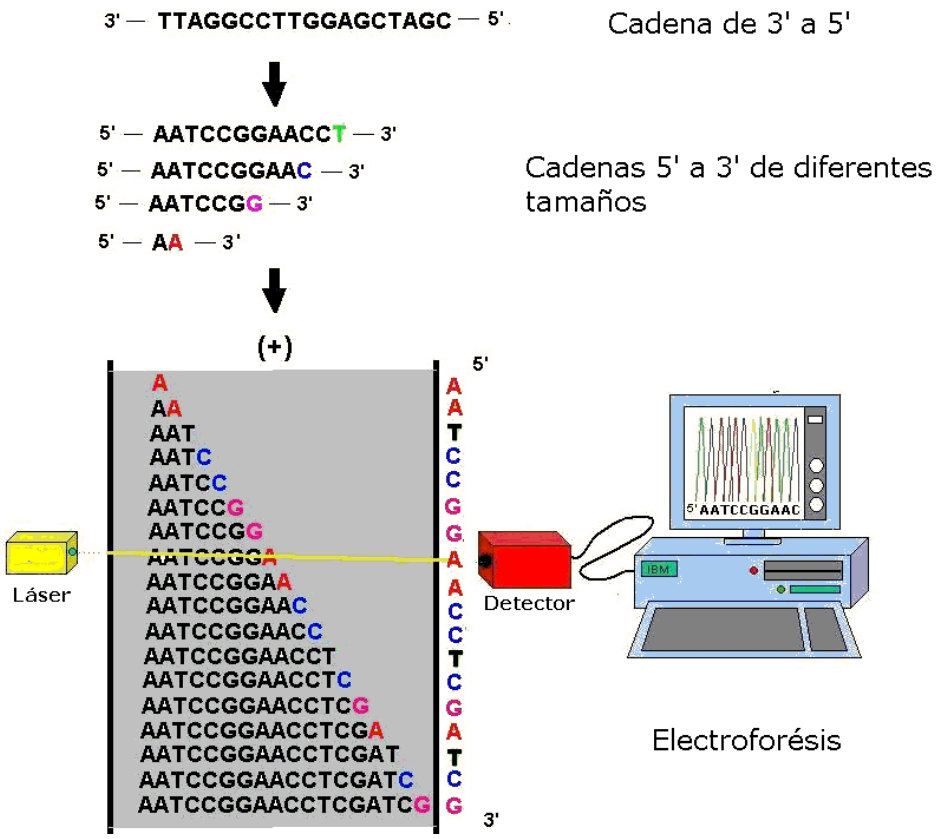
Esta secuencia de letras contiene las instrucciones para producir las proteínas que se requieren en los diferentes procesos biológicos, de tal manera que existe un mecanismo de lectura que se activa cada vez que se necesita. Además la información hereditaria está contenida en el DNA, en el caso de los organismos de reproducción sexual, el genoma está conformado por DNA proveniente de ambos progenitores, de tal forma que al constituirse un nuevo organismo, este va a poseer una combinación única de características establecidas en el genoma que se manifiestan en la etapa embrionaria y en el desarrollo del individuo.

Una vez que se conoce la secuencia es posible delimitar las áreas en las que se encuentran los genes denominadas "áreas codificantes" y distinguirlas de las áreas no codificantes que sirven de resguardo. Esta información marca el punto de partida para posteriormente identificar el límite de cada uno de los genes y su función

Secuenciación de Sanger

La técnica de secuenciación fue creada en 1975 por Frederick Sanger, tiene como limitante que solo puede utilizarse para secuenciar fragmentos de entre 500 y 700 nucleótidos.
Esta técnica inicia con el aislamiento del fragmento de DNA que se desea secuenciar, posteriormente se separan las cadenas de la molécula, generándose 2 cadenas simples. A continuación se selecciona solamente la cadena 3' a 5' y se induce un proceso denominado síntesis de DNA que ocurre comúnmente en las células y que consiste en tomar como molde una cadena 3' a 5', para construir su cadena complementaria 5'a 3' uniendo nucleótidos libres, en la naturaleza este mecanismo permite obtener una gran cantidad de copias de la molécula, en el caso de la técnica se modifica este mecanismo ya que se agregan una cierta cantidad de nucleótidos marcados con colorante (un color diferente para cada tipo de nucleótido), que además están modificados en su estructura de tal manera que al ser incorporados en la síntesis de DNA, provocan que se detenga el proceso, resultando un fragmento más corto que el original. Esto se efectúa muchas veces, produciéndose fragmentos de DNA de todos las longitudes posibles.

Posteriormente se efectúa una electroforesis, que consiste en colocar los fragmentos 5' a 3' en un recipiente con gel. En un extremo del recipiente se aplica una carga positiva y en el otro una carga negativa, generándose una corriente eléctrica lo que provoca que las cadenas se desplacen hacia el polo positivo. La velocidad del desplazamiento será inversamente proporcional al número de nucleótidos, de tal forma que las cadenas se acomodarán de menor a mayor tamaño a lo largo del gel. El siguiente paso es utilizar un rayo láser y un detector para determinar el tipo de base que posee el último nucleótido de cada cadena (nucleótido marcado), con este procedimiento se registra la secuencia de bases en una computadora



Secuenciación Genómica Total

Debido a que la secuenciación de sanger está limitada a fragmentos "cortos", no es posible utilizarla para obtener la secuencia de un genoma, con este fin se desarrolló una técnica denominada secuenciación por disparo (Shotgun sequencing), en la que se lleva a cabo un proceso de fragmentación que consiste en someter el DNA a presiones altas provocando su ruptura en fragmentos de diversos tamaños, posteriormente se selecciona una muestra de fragmentos de un cierto tamaño y se procede a obtener la secuencia de cada uno de ellos con la técnica de Sanger. Las secuencias obtenidas pasan por un proceso de acoplamiento en el cual son comparadas para localizar regiones en común que permiten ensamblar los fragmentos en bloques. Apartir de un conjunto de bloques se obtiene la secuencia del genoma.

Existen dos variantes de la secuenciación por disparo, que se diferencian tanto en el proceso de segmentación como en el de acoplamiento. En la primera técnica denominada **secuenciación clón por clón**, el proceso de segmentación se realiza en base a un mapa físico, que permite separar el genoma en diferentes secciones llamadas clones, para posteriormente secuenciar cada clón con la técnica de sanger. En el proceso de acoplamiento se cuenta con una serie de "marcas" en los fragmentos que permiten ensamblarlos con mayor precisión al momento de formar los bloques.. La segunda técnica, la cual es el tema de estudio de este trabajo, recibe el nombre

de **secuenciación genómica total** porque el proceso de fragmentación se realiza empleando la secuencia completa del genoma, de tal manera que no existen puntos de referencia que indiquen en donde ensamblar los fragmentos. En los primeros años en que surgieron éstas técnicas, la secuenciación genómica total era utilizada únicamente para genomas "pequeños" de organismos simples como las bacterias y los virus (S. Strail 2004), posteriormente al iniciar el proyecto humano, el investigador Craig Venter propuso ésta técnica para ser utilizada en este proyecto (Venter 1996), en ese entonces se desató un debate respecto a la confiabilidad de las secuencias obtenidas mediante ésta técnica (Weber 1997), (Green 1997), sin embargo en la actualidad se ha convertido en la técnica más utilizada gracias a los adelantos tecnológicos que permiten obtener secuencias de gran calidad y con sistemas altamente automatizados.

En (E. Myers, 1999) y (Weaver, 2005) se puede encontrar una explicación detallada de ambas técnicas. Y en (E. Lander, 1998) y (S. Schbath 1997), se realiza un análisis probabilístico de la técnica clón por clón.