BioCódigos

María Alejandra Rodríguez Ríos.[1] **Edgar Santiago Ochoa Quiroga.**[2]
mrodriguezri@unal.edu.co eochoaq@unal.edu.co
Saul Alvarez Lazaro[3]
salvarezla@unal.edu.co

27 de febrero de 2025

Resumen

aa

1. Introducción

aaa

2. Preliminares

En esta sección, se abordarán los conceptos preliminares relacionados con el ADN y su proceso biológico de transcripción y traducción, así como una introducción a los códigos de corrección de errores, específicamente los códigos de Hamming y los códigos de Reed-Solomon.

2.1. ADN y el proceso de replicación, transcripción y traducción

El ácido desoxirribonucleico (ADN) es una molécula fundamental presente en el interior de las células tanto eucariotas como procariotas, que contiene la información genética necesaria para el desarrollo, funcionamiento y reproducción de los organismos. Esta molécula permite la transmisión de la información genética de una generación a la siguiente. Su estructura es una doble hélice, formada por enlaces débiles de hidrógeno que unen las bases nitrogenadas de los nucleótidos purínicos y pirimidínicos, los cuales se enrollan alrededor de un eje central. Cada nucleótido está compuesto por un esqueleto de azúcar desoxirribosa y grupos fosfato, conectados mediante las bases nitrogenadas.

En cuanto a la historia del descubrimiento de la estructura del ADN, se sabe que fue inicialmente identificada por la científica Rosalind Franklin, quien a través de la técnica de difracción de rayos X obtuvo fotografías que revelaban la forma helicoidal de la molécula. Sin embargo, su contribución no fue reconocida en su momento. Fue solo cuando James Watson y Francis Crick, científicos que trabajaban en el Laboratorio Cavendish, utilizaron esas imágenes y algunas de las deducciones previas para publicar, en 1953, el artículo que describía la estructura del ADN. Años después, Watson y Crick recibieron el Premio Nobel de Medicina por este descubrimiento, pero nunca se le otorgó el reconocimiento que le correspondía a Rosalind Franklin.

En cuanto a la estructura molecular, se sabe que el ADN está formado por cuatro elementos fundamentales: los nucleótidos adenina (A), guanina (G), timina (T) y citosina (C). Los dos primeros corresponden a los nucleótidos purínicos y los dos restantes son pirimidínicos.

Estas bases nitrogenadas se emparejan de manera específica: la adenina se empareja con la timina y la guanina con la citosina. Este emparejamiento es esencial para la estabilidad y la función del ADN, permitiendo una transmisión precisa de la información genética.

La importancia del ADN radica en que es esencial para los procesos que la célula utiliza para elaborar todas las proteínas que un ser vivo necesita para subsistir. Estos procesos de transmisión de información se realizan en tres etapas, a las cuales se denomina el dogma central de la biología molecular. Este concepto describe el flujo de información genética en una célula, estableciendo que la información genética fluye desde el ADN, a través de la síntesis de ARN, y luego a través de la síntesis de proteínas.

Para poder realizar el paso de ADN a ARNm o ácido ribonucleico mensajero, necesitamos pasar por el proceso de replicación. La replicación es el proceso mediante el cual una célula copia su ADN para asegurar que cada célula hija reciba una copia exacta de la información genética. Este proceso es esencial para la división celular y tiene lugar antes de que una célula se divida, durante la fase S del ciclo celular. Uno de los primeros pasos en la replicación es cuando se desenrolla la doble hélice del ADN, lo cual es realizado por una clase de enzimas conocidas como helicasas. Las helicasas tienen la función de romper los enlaces de hidrógeno que mantienen unidas las bases nitrogenadas de las dos cadenas complementarias de ADN, separándolas y formando lo que se conoce como la horquilla de replicación.

Es importante entender que el ADN es una molécula en la que sus dos cadenas tienen orientaciones opuestas. Una cadena se lee en la dirección 5′ a 3′, mientras que la otra se lee en la dirección opuesta, es decir, 3′ a 5′. Esto se refiere a los extremos de las cadenas de ADN: el extremo 5′ de una cadena de ADN tiene un grupo fosfato unido al quinto carbono del azúcar, mientras que el extremo 3′ tiene un grupo hidroxilo unido al tercer carbono del azúcar.Una vez que se forma la horquilla de replicación, las dos cadenas de ADN están abiertas y disponibles para ser copiadas. Sin embargo, la lectura y la síntesis del ADN no son simétricas, ya que la ADN polimerasa, la enzima encargada de sintetizar nuevas cadenas de ADN, solo puede agregar nucleótidos en la dirección 5′ a 3′.

Ahi inicia el segundo paso el cual es la transcripción, en este la información genética contenida en el ADN se transcribe a una molécula de ARN mensajero (ARNm). Este proceso ocurre en el núcleo de las células eucariotas y en el citoplasma de las procariotas y es esencial para la posterior síntesis de proteínas. La transcripción comienza cuando la enzima ARN polimerasa se une a una región específica del ADN conocida como el promotor. El promotor es una secuencia de bases que indica el inicio de un gen y es crucial para la correcta transcripción del ADN. Una vez que la ARN polimerasa se une al promotor, comienza a desenrollar la doble hélice del ADN y a leer la cadena de ADN en la dirección 3′ a 5′.

A medida que la ARN polimerasa avanza, sintetiza una cadena de ARN mensajero (ARNm) complementaria a la cadena molde del ADN. Durante este proceso, las bases del ADN se emparejan de forma específica con los ribonucleótidos, la adenina (A) del ADN se empareja con el uracilo (U) en el ARN , la citosina (C) con la guanina (G). El resultado es una nueva cadena de ARN que lleva la misma información genética que el ADN, pero en una forma que puede ser utilizada en la síntesis de proteínas.

La ARN polimerasa sigue leyendo el ADN hasta llegar a una secuencia de terminación que indica el final del gen. En este punto, la transcripción finaliza, y el ARN mensajero (ARNm) recién formado se separa del ADN. El ARNm, que ahora contiene la çopia"de la información genética, es transportado fuera del núcleo hacia los ribosomas en el citoplasma, donde se

llevará a cabo la siguiente etapa.

Por último tenemos la traducción, este es el proceso comienza cuando el ARNm se une a un ribosoma en el citoplasma. El ribosoma lee el ARNm en bloques de tres bases nitrogenadas consecutivas, llamados códones. Cada códon especifica un aminoácido particular, que es la unidad básica de las proteínas. Existen 64 posibles combinaciones de códones, que codifican para 20 aminoácidos diferentes, lo que permite una gran diversidad en la construcción de proteínas.

El ARN de transferencia ARNt tiene un anticódon, el cual es una secuencia de tres bases que es complementaria a un códon del ARNm en uno de sus extremos y un aminoácido específico en el otro extremo. A medida que el ribosoma lee el ARNm, el ARNt transporta el aminoácido correspondiente al códon leído y lo coloca en la cadena polipeptídica en crecimiento. Este proceso se repite a medida que el ribosoma avanza a lo largo del ARNm. La traducción continúa hasta que el ribosoma encuentra un códon de terminación en el ARNm, lo que indica el final de la síntesis de la proteína. En ese momento, la cadena polipeptídica recién formada se libera, y la proteína se pliega para adoptar una estructura funcional.

Sin embargo, este proceso no siempre se lleva a cabo sin errores, por lo cual se tienen mecanismos que buscan corregir la mayoría de los errores que pueden surgir puesto que el no corregirlos por ejemplo en los humanos puede verse reflejado en mutaciones genéticas, proteínas no funcionales o mal plegadas y a la pérdida de la integridad genética, lo que puede causar disfunción en todo el organismo.

2.1.1. Corrección de errores

La corrección de errores en los procesos celulares Los procesos de replicación, transcripción y traducción son fundamentales para la correcta expresión de la información genética. Sin embargo, a lo largo de estos procesos pueden ocurrir errores, como la inserción de bases incorrectas o la incorrecta traducción de los codones. Para eso existen mecanismos de corrección para asegurar la fidelidad genética y evitar que estos errores afecten a la célula.

La replicación del ADN es un proceso crítico para la división celular. Sin embargo, es común que durante la síntesis de las nuevas cadenas de ADN se produzcan errores en la incorporación de nucleótidos. Para corregir estos errores, la ADN polimerasa tiene una función de corrección por prueba de lectura. Esta actividad se realiza mediante la exonucleasa 3' a 5', una función de la propia enzima que le permite retroceder y eliminar los nucleótidos incorrectos que acaba de incorporar, reemplazándolos por los correctos.

De igual manera, en la transcripción para la corrección de errores el ARN polimerasa también posee mecanismos de corrección para detectar y corregir ciertos errores en el ARNm durante su síntesis, como lo es retirar el nucleótido equivocado y poner el correcto en su lugar

En la traducción también pueden ocurrir errores, como la incorporación de un aminoácido incorrecto debido a un códon mal leído o a un error en el emparejamiento entre el ARNm y el ARNt. Para esto, los ribosomas tienen mecanismos de verificación para asegurar que la secuencia de aminoácidos sea correcta, por ejemplo, el anticódon del ARNt debe coincidir exactamente con el códon del ARNm para que el aminoácido correcto se incorpore en la proteína en formación.

Otros errores que se pueden presentar en este proceso que no son por un cambio de nucleótido, sino por un daño en la estructura o un espacio sin nucleótido, para estos existe mecanismos como la reparación por escisión de bases y la reparación por escisión de nucleótidos, este es un mecanismo que se usa para detectar y eliminar ciertos tipos de bases dañadas mediante un grupo de enzimas llamadas glicosilasas, donde cada glicosilasa detecta y elimina un tipo específico de base dañada.

2.2. Códigos de Corrección de Errores.

De manera similar, cuando enviamos un mensaje a través de un canal, queremos que el receptor de aquel mensaje lo reciba de manera mas fidedigna posible, pero enviar el mensaje puede que la información se vea alterada por el canal, de esta manera la intención detrás de este tipo de códigos como lo indica su nombre es la de ser capaces de detectar los errores en el mensaje recibido, y de corregirlos, de esta manera el mensaje sera enviado con éxito completamente.
En esta sección daremos los hechos mas relevantes de los códigos que utilizaremos para realizar la codificación y decodificación de cadenas de ADN. No mostraremos las pruebas de los resultados teóricos usados, debido a que este no es el propósito, pero estos hechos pueden ser encontrados en (agregar bibliografia)

2.2.1. Codigos de Reed-Solomon

Como fue mencionado antes uno de los dos códigos que utilizaremos en este proyecto, son los códigos de Reed-Solomon. Primero debemos definirlos y para esto los introduciremos por medio de la definición dada en (insertar cita sarria)

Definición 2.2.1. Dado el cuerpo $GF(D)^n$, donde $k \leq n \leq D$ son enteros positivos. Definimos el código de dimensión k como

$$RS_D(\alpha, n, k) = \{(f(\alpha_1), \ldots, f(\alpha_n)) : f \in GD(D)[x], grad(f) \leq k-1\}.$$

Donde $\alpha = (\alpha_1, \ldots, \alpha_n) \in GF(D)^n$, con componentes distintas. La función de codificación esta dada por

$$(a_0, a_1, \ldots, a_{k-1}) \mapsto \left(\sum_{i=0}^{k-1} a_i \alpha_1^i, \ldots, \sum_{i=0}^{k-1} a_i \alpha_n^i \right).$$

Donde cada $a_i \in GF(D)$.

La idea detrás de la construcción de este tipo de código es hacer uso de que un polinomio se encuentra determinado por sus coeficientes. Para ver esto en acción, realicemos un ejemplo sencillo para ver esto en acción

Ejemplo. Consideremos un código con $k = 2, n = 3$ y $D = 3$. Tomamos $\alpha = (0, 1, 2)$, esto debido a que necesitamos que las entradas sean diferentes por definicion, Note que las tuplas las escribiremos como cadenas de simbolos de ahora en adelante, es decir $(0, 1, 2) = 012$. Luego el codigo esta dado por evaluar en todos los polinomios de grado 1 con coeficientes en $GF(3)$.

$$RS(2, 3, 012) = \{000, 111, 222, 012, 120, 210, 021, 102, 210\}.$$

En particular si por ejemplo queremos codificar la palabra 12, que arroja una fuente triaria tenemos que

$$12 \mapsto (1 + 2 \cdot 0, 1 + 2 \cdot 1, 1 + 2 \cdot 2) = 102.$$

Con este ejemplo podemos enfatizar algunos conceptos

- Si uno quiere codificar palabras de longitud k, necesitamos que el campo base tenga mas elementos, es decir si queremos codificar una fuente $4-aria$, necesitamos trabajar mínimo con el cuerpo de finito de 4 elementos o mas.
- Note que en este caso la elección del α no es única ya que pudimos haber seleccionado 201, por lo que si bien el código bloque cumple la misma función, la codificación cambia. Por lo que seria bueno poder codificar independientemente del α escogido.
- El punto anterior tiene sentido al considerar que la codificación esta completamente determinada por el polinomio al que es asignado la palabra que emite la fuente.

En vista de eso, resulta natural preguntarse si podemos definir los códigos de Reed-Solomon sin considerar un α explicito, es decir, concentrarnos unicamente en la estructura polinomial. Para esto debemos hacer uso de algunos conceptos algebraicos.

Definición 2.2.2. Dado el cuerpo finito $GF(D)$, decimos que $\alpha \in GF(D) - \{0\}$, es un elemento primitivo, si α es un generador de $\alpha \in GF(D) - \{0\}$ visto como grupo bajo la operación de multiplicación.

Este concepto de elemento generador sera crucial, en el sentido de que si bien con la definición original podemos plantear una matriz generadora G y una de corrección H, ahora que trabajaremos con el elemento α en "abstracto". Generaremos el código por medio de un polinomio generador. Antes de eso mencionaremos los hechos algebraicos que sustentan esta construcción

Proposición 2.2.3. Dado el cuerpo $GF(p)$ con p un numero primo, podemos construir el cuerpo $GF(p^e)$ por medio de el cociente $GF(p)/\langle x^{p^e} - x\rangle$.

esto nos da una construcción por medio de clases de equivalencia, que podemos tomar por medio de residuos, pero resultaría engorrosa, por lo que estos para ejemplificar campos, los construiremos por medio de tomar un factor irreducible de ese polinomio sobre el cuerpo base. El siguiente hecho nos da una caracterización, de aquellos elementos primitivos que podemos ver como raíces del polinomio.

Proposición 2.2.4. Dado $\alpha \in GF(p^e)$, tenemos que $\alpha^{p^e} - \alpha = 0$, luego

$$x^{p^e} - x = \prod_{\alpha \in GF(p^e)} (x - \alpha).$$

Note que si excluimos el elemento 0, tenemos una factorización sobre los elementos no nulos de nuestro cuerpo finito. Con todos estos ingredientes procedemos a dar la definición que usaremos para la codificación

Definición 2.2.5. Dado α un elemento primitivo, el código de Reed-Solomon se define como

$$RS(D, k) = \{(f(1), f(\alpha), \ldots, f(\alpha^{n-1})) \in GF(D)^n : f \in GF(D)[x], grad(f) \leq k - 1\}.$$

Donde escogemos enteros positivos $n = D - 1$ y $k < D$.

Notemos que en primera instancia pareciera que no hay diferencia en los códigos, pero antes de proceder con la diferencia crucial, algunas observaciones.

- Note que bajo esta definición, por el uso del elemento primitivo α, a diferencia de la primera definición evaluación, ya no tenemos el elemento 0 considerado.
- Ejemplos pequeños como el realizado para la anterior definición, ya no son viables debido a la restricción de la evaluación en elementos primitivos.
- Note que antes había mas grados de libertad para el tamaño de la tupla, ahora la definimos directamente como $D - 1$, lo que nos da una cota superior para la longitud de nuestros mensajes.

Estas son pequeñas cosas que se pueden notar inmediatamente del código definido, pero el factor diferencial viene dado por el siguiente hecho que habíamos anticipado previamente

Teorema 2.2.6. Dado un código $RS(D, k)$, si la distancia mínima del código es $d = D - k$, entonces el polinomio generador del código esta dado por

$$g(x) = \prod_{i=1}^{d-1} (x - \alpha^i),$$

donde α es elemento primitivo.

Note que este polinomio divide a el polinomio por el que se realiza el cociente del cuerpo $GF(D)$, por lo que desde un punto de vista algebraico resulta lógico que sea así. Ademas podemos observar que por fin vemos el concepto de distancia mínima que habíamos esquivado hasta el momento, pero que era inevitable evitarlo mas, debido a su rol crucial en la capacidad de un código para detectar y corregir patrones de errores.

3. Códigos de Hamming

El otro código corrector de errores con el que trabajaremos sera el código de Hamming.

El Código de Hamming es un esquema de detección y corrección de errores diseñado por Richard Hamming en 1950.
El principio fundamental del Código de Hamming consiste en agregar bits de redundancia a los datos originales, de manera que los errores introducidos en la transmisión puedan ser detectados e incluso corregidos. En particular, el código Hamming $(7, 4)$ permite la corrección de un único bit erróneo y la detección de hasta dos errores en un bloque de datos, pero el problema es que no puede diferenciar entre uno y dos errores entonces para esto se utiliza el codigo de Hamming extendido donde agregamos un bit de paridad mas, el cuales la paridad de todo los datos.

4. Fundamentos Matemáticos del Código de Hamming

El Código de Hamming se fundamenta en la teoría de códigos lineales y hace uso de matrices generadoras y de comprobación de paridad para la codificación y la detección de errores.

4.1. Matriz Generadora

Para el código Hamming $(7, 4)$, la matriz generadora G se define como:

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Dado un mensaje de 4 bits $m = (m_1, m_2, m_3, m_4)$, el código resultante se obtiene mediante la multiplicación matricial:

$$c = mG$$

El resultado es un código de 7 bits que incluye tanto los bits de información como los bits de paridad.

4.2. Matriz de Comprobación de Paridad

Para detectar y corregir errores en la transmisión, se usa una matriz de comprobación de paridad H, definida como:

$$H = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Dado un código recibido $r = (r_1, r_2, ..., r_7)$, el síndrome S se calcula como:

$$S = H \cdot r^T$$

Si $S = 000$, significa que no hay errores en la transmisión. Si el resultado es distinto de cero, indica la posición del bit erróneo en el código recibido, permitiendo su corrección.

4.3. Ejemplo de Codificación, Corrección y Decodificación

Supongamos que queremos codificar el mensaje $m = (1, 0, 1, 1)$.
Multiplicamos por la matriz G:

$$c = (1, 0, 1, 1)G = (1, 0, 1, 1, 0, 1, 0)$$

El código transmitido es 1011010. Supongamos que se introduce un error en la posición 3 y se recibe $r = 1001010$.
Calculamos el síndrome:

$$S = H \cdot r^T = (0, 1, 0)$$

El síndrome indica que el error está en la posición 3. Corrigiéndolo, obtenemos el código correcto 1011010, que decodificamos extrayendo los bits de datos 1011.

5. Codigos y ADN

En esta sección estudiaremos la manera en la que aplicaremos los códigos, cuales serán nuestras pautas de partida, por que las tomamos así, y algún ejemplo pequeño para ilustrar lo que sera aplicado posteriormente.

5.1. Reed-Solomon aplicado al ADN

Dado el hecho que un mensaje puede ser visto por medio de ASCII, es natural empezar a preguntarnos por por un campo finito con 256 elementos, es decir nuestro punto de partida sera $GF(D)$, con $D = 2^8$. Recordemos que este cuerpo se puede construir consiguiendo un polinomio irreducible de grado 8 sobre $GF(2)$. Luego como cada elemento esta dado por el residuo, tenemos polinomios de grado 7 o menos, así podemos escribir estos residuos simplemente como cadenas de los coeficientes.

$$GF(D) = \{a_0 a_1 \dots a_7 : a_i \in GF(2), 0 \leq i \leq 7\}.$$

Referencias

[1] L Grafakos. *Classical Fourier Analysis*. Springer, 2008.

[2] O. Riaño. *Notas de clase : Series de Fourier*. UNAL, 2024.

[3] R. J. Iorio Jr. and V. d. M. A. Iorio. *Fourier Analysis and Partial Differential Equations*, volume 70 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2001.