El modelo de 3 paramétros de Kimura y la modelación de errores en la transmisión de datos

David Leonardo Luengas
Universidad del Rosario

1. Objetivos

Para realizar el proyecto final de la asignatura de teoría de números es necesario plantear los siguientes objetivos.

- Comprender el modelo de 3 parámetros y de 2 parámetros de Kimura.
- Identificar los elementos en común entre la evolución del ADN y la transposición de bits en el envío de mensajes.
- Evaluar el éxito de los modelos de de Kimura para analizar los efectos del error en la transmisión de un mensaje a través de una red.

2. Marco Teórico

2.1. Modelos de Kimura

Motoo Kimura propuso dos modelos en 1981 que analizan la evolución de pares homólogos de nucleotidos en distintas especies para estimar su distancia evolutiva. El modelo de 3 parámetros considera 3 tipos de sustituciones posibles en la evolución del ADN, como se ve en la siguiente figura:

Donde α es la tasa de sustitución de nucleotidos de tipo transición mientras que γ y β representan la tasa de transversiones de nucleotidos. Este modelo asume que las probabilidades de algún tipo de sustitución en cada par son independientes a cualquier otro tipo de sustitución que ocurra en la secuencia de ADN. Es decir, conforme va ocurriendo cada etapa de la evolución del ADN, los cambios que ocurren en etapas anteriores no afectan

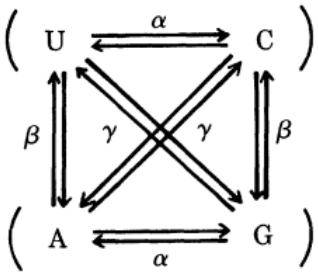

Figura 1: Transiciones del modelo de 3 parámetros, tomada de [3].

las probabilidades de que ocurran sustituciones en la siguiente etapa de evolución de la secuencia. De esta forma, es posible solo tener en cuenta las tasas en las que ocurren cada tipo de sustitución como únicos parámetros del modelo. Kimura define el siguiente sistema de ecuaciones diferenciales para modelar la evolución del ADN bajo estos supuestos.

$$\frac{dP}{dT} = 2\alpha - 2(2\alpha + \beta + \gamma)P - 2(\alpha - \gamma)Q - 2(\alpha - \beta)R$$

$$\frac{dQ}{dT} = 2\beta - 2(\beta - \gamma)P - 2(\alpha + 2\beta + \gamma)Q - 2(\beta - \alpha)R$$

$$\frac{dR}{dT} = 2\gamma - 2(\gamma - \beta)P - 2(\gamma - \alpha)Q - 2(\alpha + \beta + 2\gamma)R$$

Donde P representa la frecuencia relativa de transiciones de tipo α, Q la frecuencia relativa de transversiones de tipo β y R las

transversiones de tipo γ en un tiempo T. Con este modelo, es posible estimar la distancia evolutiva K entre secuencias de ADN con la siguiente ecuación.

$$K = \frac{-1}{4} \ln [(1-2P-2Q) \\ (1-2P-2R)(1-2Q-2R)]$$

Según [4], esta variante del modelo de Kimura tiene la particularidad de que las sustituciones de los nucleotidos definen un grupo con la siguiente tabla de multiplicación.

$\bigoplus$	0	P	Q	R
0	0	P	Q	R
P	P	0	R	Q
Q	Q	R	0	P
R	R	Q	P	0

Esta tabla es isomorfa al grupo $\mathbb{Z}_2 \times \mathbb{Z}_2$ donde 0 es la transformación nula, es decir, que la secuencia de ADN no haya sufrido cambios en el paso evolutivo en cuestión y el resto de los elementos se refieren a las transformaciones definidas anteriormente. La operación $\bigoplus$ se refiere a la concatenación de sustituciones durante la evolución de una cadena de ADN. Esto revela una estructura algebraica subyacente en los supuestos del modelo de Kimura.

El segundo modelo propuesto por Kimura simplifica el fenómeno de la evolución del ADN agrupando las 4 bases en dos grupos como se ve en la figura.

Donde α representa la tasa de sustituciones del grupo de nucleótidos A_1 al grupo A_2 y β la tasa de sustituciones del grupo A_2 al grupo A_1. Kimura propone el siguiente conjunto de ecuaciones diferenciales para modelar la evolución del ADN bajo estos supuestos.

$$\frac{dX}{dT} = -2\alpha X + \beta Z$$
$$\frac{dY}{dT} = -2\beta Y + \alpha Z$$
$$\frac{dZ}{dT} = 2\alpha X + 2\beta Y - (\alpha + \beta) Z$$

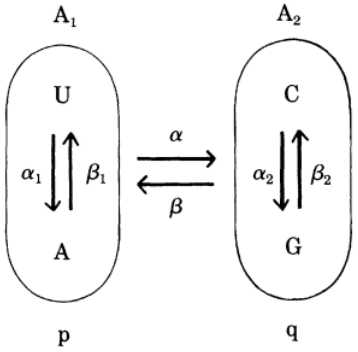

Figura 2: Transiciones del modelo de 2 parámetros, tomada de [3].

2.2. Errores en la transmisión de datos

La transmisión de datos a través de un medio puede introducir errores en la transmisión del mensaje que pueden alterar el contenido del mensaje para el receptor. Esta es una propiedad que busca minimizarse al máximo para asegurar que el medio es óptimo para la comunicación. Una métrica muy usada para cuantificar la calidad del medio es el bit error rate, BER por sus siglas en ingles, que se calcula como la tasa de bits erróneos recibidos en un periodo de tiempo [1]. Varios factores afectan al BER incluyendo:

- Ruido del canal de transmisión
- Potencia de la señal
- Distancia de transmisión
- Interferencia del medio (i.e reflexión, refracción, etc)

Usualmente medios físicos, como los cables de fibra óptica, tienen un BER del orden de 10^{-9} a 10^{-12} dado que las condiciones de aislamiento de estos cables permiten una gran precisión y confiabilidad en la transmisión de mensajes. La comunicación inálambrica de alta calidad como el Wi-Fi tiene valores del BER más altos a causa de las condiciones de transmisión en las que se utilizan normalmente, en el rango de 10^{-6} a 10^{-9}. Otra métrica que se utili-

za con frecuencia para analizar los errores de transmisión en canales digitales es el signal to noise ratio, SNR por sus siglas en inglés. Esta cantidad se calcula como el cociente entre la potencia de la señal recibida y la potencia del ruido presente en el canal [2]. Esta métrica, expresada en decibeles, con un valor mayor a 30dB refleja un canal de comunicación de alta calidad mientras que, con valores cercanos a los 10dB, se considera un canal de comunicación poco confiable.

3. Planteamiento del modelo

Para el desarrollo del proyecto se busca poner a prueba el modelo de Kimura en una situación de transmisión de datos simulada. Utilizando MATLAB y el paquete de Comunicación incluido en el programa, se busca simular distintos canales de comunicación con varios tipos de ruido.La situación ideal que se va a simular es una red con varios nodos sucesivos. Entre cada nodo se introduce un ruido que afecte la señal y se envía por varios nodos sin corrección. Es decir, no se aplica ninguna las técnicas de corrección en el transporte del mensaje. Esto se puede pensar como la transmisión de un mensaje a larga distancia, donde cada nodo representa un amplificador que no corrige los errores del mensaje y solamente se asegura de darle la suficiente potencia a la señal para que pueda llegar al siguiente nodo hasta el receptor final. En este caso, solamente se considera el error debido al canal de comunicación.

De estas simulaciones es posible estimar los parámetros para el modelo de Kimura. Con cada iteración se calcula la probabilidad de un cambio de bit y se hace un promedio utilizando simulaciones de Monte Carlo. Para poder utilizar el modelo de 3 parámetros en esta situación, donde se tienen solamente dos acciones posibles sobre la cadena de bits, se igualan las probabilidades para las transiciones de tipo γ y β. De esta forma, el modelo tiene dos operaciones posibles, ya que las otras dos son equivalentes y puede ser aplicado al modelado del error en la transmisión de mensajes.

4. Implementación del Modelo

El modelo se ejecuta sobre un grafo sin pesos que representa una arquitectura de red en particular. Cada nodo corresponde a un router dentro de la red mientras que cada arista representa un posible camino que puede tomar un paquete. Los pesos de las aristas corresponden a la tendencia al error que tiene ese camino en particular; un valor mayor representando una poca confiabilidad de ese canal en mientras que un valor pequeño representa un canal de comunicación confiable y con una cantidad de errores mínima.

La implementación[1] del modelo requirió de varias funciones auxiliares para ejecutar la simulación y hacer las pruebas de su efectividad. Se construyó la función `sim_kimura_3()` que utiliza las soluciones analíticas al modelo propuestas en [3] para recuperar las probabilidades de un flip en algún bit del mensaje. Para encontrar los parámetros del modelo se implementó la función `get_params()` que lleva a cabo la simulación de Monte Carlo y envía N mensajes de prueba a través del canal, cuantificando la probabilidad de un flip de cada bit. La función `weight_kimura()` asigna el peso de una arista según el modelo de Kimura. Este resulta siendo la suma de las probabilidades de flip de cada uno de los bits. En las pruebas del modelo, se asume que la contribución al peso de la arista es equivalente para cada una de las probabilidades de cambio de bit. Por otro lado, la clase `Edge` codifica los parámetros relevantes del canal de comunicación como el BER, el SNR y el valor real del peso de esa arista en la red. Este peso corresponde a la suma de dos componentes dados por la siguiente función.

$$w_r = \frac{(BER + \frac{1}{6+SNR})}{2}$$

[1] El código completo de la implementación se puede acceder en el repositorio de GitHub.

Dado que el BER puede tener valores negativos, esta función codifica que los valores más negativos del BER tienen mayor peso y permite que los valores positivos del BER siempre contribuyan al peso de la arista, así sea mínimamente. Además esta función está normalizada, permitiendo que los pesos determinados a través del modelo de Kimura La función `weight_edges()` construye dos grafos de la misma red. El primer grafo utiliza los valores reales como los pesos de la red mientras que el segundo grafo asigna los pesos dados por el modelo de Kimura. Aquí vale la pena aclarar el procedimiento para determinar los parámetros del BER y el SNR para cada arista de la red. El BER se muestrea de una variable aleatoria $X \sim \mathcal{U}(0, 1)$ mientras que el SNR se muestrea de una variable aleatoria $Y \sim \mathcal{U}(-5, 5)$. El intervalo escogido para la variable Y asegura la mayor varianza en los resultados del error en las aristas de las redes. Esta elección también determina la función del peso para este parámetro en la clase `Edge`.

El último método que vale la pena resaltar a profundidad es la función `eval_performance()`. Esta función toma los grafos con los pesos y usa el algoritmo de Dijkstra para encontrar el camino más corto de un lado de la red a otro usando como función de distancia los pesos del grafo. El éxito del algoritmo se mide a través de la similitud entre los caminos que devuelve el algoritmo; se suman los aciertos en tanto el camino real comparta las mismas aristas en la misma posición con el camino más corto según los pesos dados por el algoritmo de Kimura.

5. Resultados

Para evaluar el desempeño del modelo en la tarea de asignar los pesos correctamente a los grafos de las redes se diseñaron las tres arquitecturas de red vistas en la Figura 3. Se realizaron 1000 pruebas para cada una de las arquitecturas en las cuales se produjeron parámetros aleatorios para cada una de las aristas en cada iteración siguiendo el procedimiento detallado en la implementación del modelo. Para la simulación de Monte Carlo que determina los parámetros del modelo de Kimura se realizan 100 simulaciones y se toma el promedio. En particular, fue necesario mantener este parámetro en este rango para mantener el tiempo de ejecución controlado.

Después de realizar la simulación en cada una de las arquitecturas de prueba se obtuvieron los siguientes resultados.

Arquitectura	**Precisión (%)**
Red 1	43.46
Red 2	40.84
Red 3	43.62

Cuadro 1: Resultados de la Simulación

5.1. Análisis de Resultados

El desempeño final del modelo para determinar el peso de las aristas en el grafo de la red no es muy bueno en vista de las métricas de precisión que arroja el modelo. Para analizar este resultado se pueden proponer varias posibles hipótesis que revelan las debilidades de este modelo para enfrentar la situación propuesta.

Por un lado, la manera en la que se estiman los parámetros del modelo de Kimura no permiten estimar correctamente la contribución de los parámetros del BER y el SNR al peso final de la arista. El modelo toma una muestra y estima la probabilidad de un cambio de bit. Después de agregar un ruido blanco gaussiano, resulta muy difícil para el modelo acercarse al valor real del BER y el SNR. De esta forma, la estimación es muy rudimentaria para aproximarse al comportamiento de los parámetros. Incluso, como la contribución que hace cada uno de los parámetros al peso real de la arista no sigue una relación lineal, la expresividad del modelo está fuertemente limitada por la linealidad en la que estima los pesos del grafo.

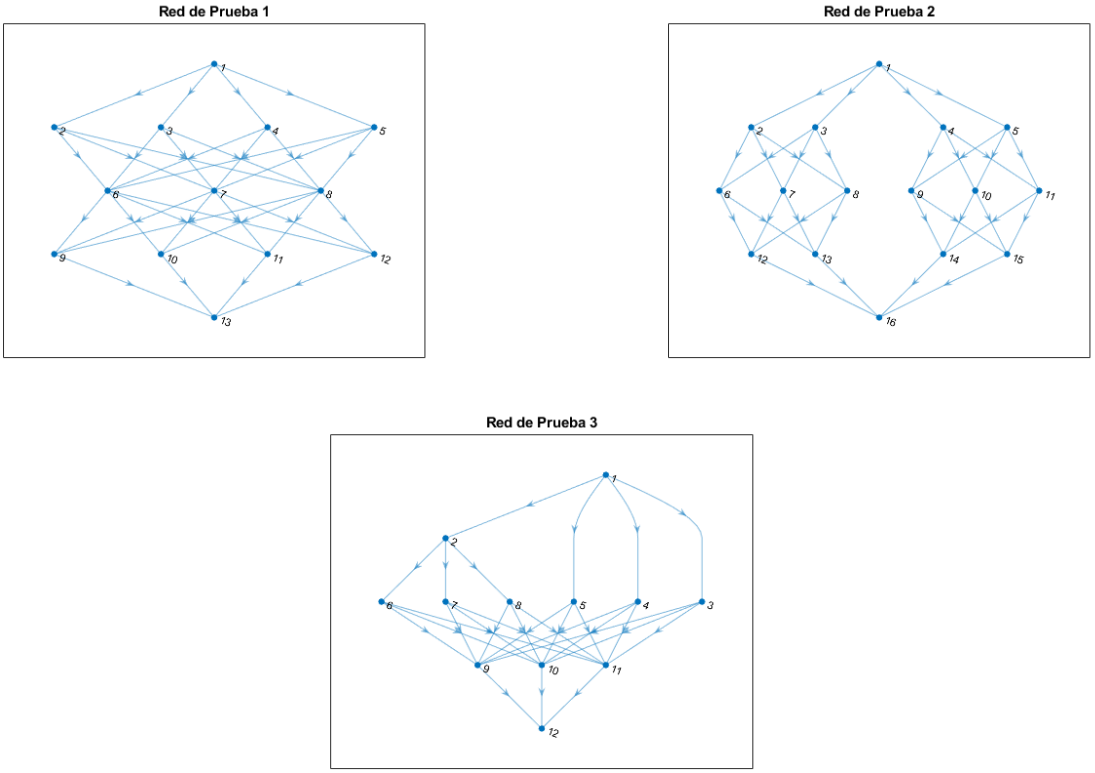

Figura 3: Redes de Prueba para el modelo

Después de varios experimentos individuales, se notó que el peso máximo que puede otorgar el modelo de Kimura a una arista está alrededor del 0.5 mientras que el valor real del peso de la arista puede llegar a tener un valor de 1. Este detalle en particular parece ser el determinante a la hora de evaluar el fracaso del modelo; el fenómeno de la evolución del ADN bajo los supuestos del modelo de Kimura es muy simple con respecto a la interferencia del ruido en la transferencia de mensajes en una red.

En [4] encontramos explícitamente los supuestos que gobiernan el funcionamiento del modelo de Kimura done uno de los más importantes se refiere a la independencia de cada sitio a las transposiciones de bases. Este supuesto no se cumple en su totalidad en la situación problema en tanto el flip de los bits es un proceso aleatorio que no depende de que tipo de bit sea. Es decir, los bits que sufren un cambio en la cadena no tienen una probabilidad única asociada según el tipo de bit que sea. Esto contrasta con la suposición de que cada tipo de base nitrogenada en el ADN tiene asociada una probabilidad independiente de las otras bases donde ciertos tipos de cambios tienen asociados distintos tipos de probabilidades.

Referencias

[1] GIGALIGHT. *What is the Bit Error Rate(BER)? — gigalight.medium.com.* https://gigalight.medium.com/what-is-the-bit-error-rate-ber-869cf7b9bb7d. [Accessed 22-04-2024]. 2023.

[2] Robert Kieser, Pall Reynisson y Timothy J. Mulligan. «Definition of signal-to-noise ratio and its critical role in split-beam measurements». En: *ICES Journal of Marine Science* 62.1 (ene. de 2005),

págs. 123-130. ISSN: 1054-3139. DOI: 10.1016/j.icesjms.2004.09.006. URL: http://dx.doi.org/10.1016/j.icesjms.2004.09.006.

[3] M Kimura. «Estimation of evolutionary distances between homologous nucleotide sequences.» En: *Proceedings of the National Academy of Sciences* 78.1 (ene. de 1981), págs. 454-458. ISSN: 1091-6490. DOI: 10.1073/pnas.78.1.454. URL: http://dx.doi.org/10.1073/pnas.78.1.454.

[4] Mike Steel et al. «A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model». En: *New Zealand Journal of Botany* 31.3 (jul. de 1993), págs. 289-296. ISSN: 1175-8643. DOI: 10.1080/0028825x.1993.10419506. URL: http://dx.doi.org/10.1080/0028825X.1993.10419506.