CÁLCULO I

TRABAJO ACADÉMICO MATLAB

USO DE LOS MODELOS DE MARKOV OCULTOS PARA EL ANÁLISIS PROBABILÍSTICO DE REGIONES GENÓMICAS ASOCIADAS A RASGOS FENOTÍPICOS

Gabriel Vallet, Aritz Arribas, Jorge M. Pérez, Carles A. Irurita

ÍNDICE

1. Objetivo del modelo

El estudio busca identificar regiones genómicas (SNPs) relacionadas con un rasgo específico. El modelo evalúa la probabilidad de que un SNP pertenezca a diferentes categorías de asociación con genes responsables del rasgo.

2. Qué es un Modelo de Markov Oculto (HMM)

Se trata de un modelo probabilístico usado en la predicción de procesos estocásticos. Estos modelos permiten representar sistemas estocásticos que evolucionan con el tiempo. Están compuestos por secuencias de estados ocultos que cambian a lo largo del tiempo, llamados ocultos porque no podemos determinar en qué estado se encuentra el sistema en ningún momento. Tan sólo somos capaces de observar una secuencia de símbolos que se generan en función del estado en el que se encuentra el sistema.

La evolución en la secuencia de estados ocultos se modela con una cadena de Markov, con la propiedad de que llegar a un estado particular sólo depende del estado previo. Y la relación entre los estados ocultos y las observaciones se modela con una distribución de probabilidades condicionales que, al contrario que la cadena de Markov de los estados ocultos, esta relación sólo depende del estado actual y no de los estados previos o futuros.

Por ejemplo, podríamos observar síntomas (las señales visibles) y usar un HMM para inferir cuál es la enfermedad subyacente (el estado oculto).

Como se ha explicado antes, el funcionamiento de un HMM se basa en tres componentes principales: las probabilidades de transición, las probabilidades de emisión y las probabilidades iniciales. Las

probabilidades de transición describen la probabilidad de pasar de un estado a otro. Las probabilidades de emisión indican qué tan probable es observar un símbolo o señal dado un estado particular. Finalmente, las probabilidades iniciales definen cómo empieza el sistema, es decir, la probabilidad de que el modelo comience en cada uno de los estados posibles.

Los HMM utilizan algoritmos como el de Viterbi para encontrar la secuencia más probable de estados ocultos que explica las observaciones, o el algoritmo Forward para calcular la probabilidad total de una secuencia de señales. Esto los hace especialmente útiles en problemas donde las observaciones están influenciadas por procesos subyacentes complejos, como en el análisis de genes.

3. Qué son los SNPs

Cómo se distingue qué es un SNP

Un SNP (polimorfismo de un solo nucleótido) es una variación en una única base en una posición específica del genoma que está presente en una parte significativa de la población.

- Para que un cambio de base sea considerado un SNP, debe ser:
    - **Frecuente:** Aparecer en al menos un porcentaje mínimo de individuos en la población analizada (por ejemplo, 1%).
    - **Específico:** Ocurrir siempre en la misma posición genómica.

Si existen cambios de base aleatorios (mutaciones puntuales únicas no heredadas) o variaciones estructurales más grandes (como inserciones o deleciones), no es considerado un SNP

Durante el entrecruzamiento, un gen y un marcador se segregan de forma aleatoria. Si están localizados en el mismo cromosoma, se observa una desviación de la segregación aleatoria cuando ambos ocupan posiciones cercanas en el mismo cromosoma. Cuanta menos distancia, menos

probabilidad de que ocurra recombinación entre el gen y el marcador y más probabilidad de que exista la co-segregación.

De esta forma, los marcadores ligados al locus de un gen de interés, se segregarán de forma menos aleatoria que los que se encuentran desligados a ese locus. Más concretamente, la relación entre el SNP y el gen será mayor cuando la frecuencia de las bases observadas en los descendientes (A, T, C, G) en la posición específica del SNP concuerde más con la base que se encuentra en esa misma posición en la **cadena de referencia**. A este fenómeno se lo conoce como desequilibrio de ligamiento, cuyas implicaciones serán estudiadas más adelante.

Cadena de referencia: Es una secuencia genómica estándar, generalmente tomada de una población de referencia o de un individuo modelo. Cada posición en esta cadena tiene una base conocida, como una A en el SNP en cuestión.

Frecuencias observadas:

- Al analizar a los descendientes, se secuencia el ADN y se cuentan cuántas veces aparece cada base (A, T, C o G) en la posición del SNP. Estas frecuencias reflejan la **variabilidad genética** en esa posición.

- Por ejemplo, supongamos que el SNP en la referencia tiene una base **A**. En los descendientes, las frecuencias podrían ser: 70%A, 20%G, 5%T, 5%CEsto indica que la mayoría de los descendientes heredaron la **A** en esta posición, pero también hay una minoría que heredó otras bases, posiblemente debido a herencia del otro progenitor o mutaciones.

¿Por qué un mismo SNP puede variar de bases en los descendientes?

Un SNP puede presentar variabilidad en las bases observadas entre los descendientes. Esto ocurre por:

1. **Herencia biparental:**
    - Los descendientes reciben una copia del genoma de cada progenitor.
    - Si los progenitores tienen diferentes bases en la posición del SNP (por ejemplo, el padre tiene una **A** y la madre tiene una **G**), los descendientes pueden heredar cualquiera de estas variantes.
2. **Polimorfismo del SNP:**
    - Por definición, un SNP es una posición en el genoma donde existe **variación genética** en la población.
    - Esto significa que diferentes individuos pueden tener diferentes bases en esa posición.
    - En los descendientes, las frecuencias observadas reflejan esta variabilidad.
3. **Recombinación genética:**
    - Durante la formación de gametos, puede ocurrir recombinación, lo que puede afectar cómo se heredan variantes en regiones cercanas al SNP.
4. **Errores de secuenciación:**
    - Las frecuencias observadas también pueden incluir ruido técnico, como lecturas incorrectas de bases durante la secuenciación.

¿Qué implicaciones tiene esto en el modelo?

El modelo HMM utiliza las frecuencias de bases para calcular las probabilidades de emisión. Por ejemplo, si un SNP tiene 80% de bases

que coinciden con la referencia, hay una alta probabilidad de que el SNP esté en un estado asociado al progenitor con el rasgo.

Las bases menos frecuentes pueden indicar herencia del otro progenitor o incluso un estado "no asociado".

La variación en las bases entre los descendientes permite al modelo diferenciar entre estados (asociado, no asociado, etc.) al analizar patrones heredados.

4. ¿Qué es el mapeo genético?

El mapeo genético es el proceso de identificar la posición de genes en un cromosoma y la distancia relativa entre ellos. Este proceso es esencial para identificar regiones genómicas que influyen en rasgos fenotípicos o enfermedades, seleccionar variantes favorables en programas de cría o ingeniería genética…

En el caso de este modelo, el objetivo es identificar SNPs asociados con un gen específico como puede ser el de una enfermedad.

Desafíos del mapeo genético tradicional

Resolución limitada: Los métodos tradicionales a menudo agrupan múltiples SNPs en bloques grandes, lo que dificulta identificar genes específicos.

Costos elevados: La evaluación individual de SNPs requiere muchos experimentos, lo que es ineficiente.

Ruido biológico: Las mutaciones, errores de secuenciación y desequilibrios de ligamiento pueden dificultar el análisis.

4.1 Distinción entre SNP y gen:

- **SNP (Single Nucleotide Polymorphism):** Es una variación en un único nucleótido (A, T, C o G) en una posición específica del genoma. Es el tipo más común de variación genética en las poblaciones. SNPs no son

necesariamente funcionales (es decir, no siempre afectan la expresión o función de un gen), pero actúan como marcadores de regiones genómicas específicas.

- **Gen:** Es una secuencia de ADN que codifica para una proteína o tiene una función reguladora. Los genes pueden abarcar cientos o miles de nucleótidos, lo que los hace regiones más grandes y complejas de analizar directamente.

4.2 Por qué identificar SNPs en lugar de genes?

SNPs son más específicos y fáciles de medir

Los SNPs son variaciones puntuales, por lo que se pueden identificar con precisión en una posición específica del genoma.

Los genes, al ser secuencias largas, tienen una mayor variabilidad interna, lo que dificulta una localización exacta.

Técnicas de análisis más directas:

Los SNPs se identifican fácilmente mediante tecnologías de secuenciación o microarrays mientras que los genes requieren análisis más complejos, como la evaluación de su estructura (exones, intrones, regiones reguladoras) y su expresión.

SNPs son marcadores de asociación con genes o rasgos

Los SNPs suelen estar en **desequilibrio de ligamiento (DEFINIR)** con genes funcionales. Esto significa que un SNP cercano a un gen puede actuar como marcador para ese gen sin necesidad de analizar directamente toda la región genómica.

Por ejemplo, si un SNP se asocia consistentemente con un rasgo, se puede inferir que un gen cercano podría ser el responsable del rasgo.

Reducción de ruido en los datos

Debido a menor variabilidad genética: un SNP es una variación puntual y

específica, mientras que un gen puede tener múltiples variaciones y mutaciones que complican su análisis. Por lo tanto, al trabajar con SNPs, se reduce la complejidad y el ruido en los datos.

Costos y tiempo

La identificación de SNPs mediante secuenciación de alto rendimiento o microarrays es más rápida y económica que realizar un análisis detallado de genes y su función.

Cómo se relacionan los SNPs y los genes

Los SNPs pueden estar

1. **Dentro de genes funcionales:** Si afectan directamente la secuencia codificante o reguladora.
2. **Cerca de genes:** Si están en desequilibrio de ligamiento con ellos (no se segregan de forma independiente).

Si un SNP tiene una alta probabilidad de estar en el estado "asociado al rasgo" en el modelo HMM, se analiza la región cercana al SNP para encontrar genes que puedan estar relacionados con el rasgo.

5. ¿Cómo optimiza un modelo de Markov el mapeo genético?

Los modelos de Markov ocultos (HMM) permiten abordar estos desafíos mediante consideración de dependencias**:** Un HMM captura las relaciones entre SNPs consecutivos a través de transiciones entre estados. Esto mejora la identificación de regiones genómicas relevantes al considerar las correlaciones entre SNPs cercanos. En lugar de clasificar cada SNP de forma determinista, un HMM calcula la probabilidad de que un SNP pertenezca a un estado particular que representa su asociación con un rasgo de interés. Una vez hecho, el HMM analiza las frecuencias observadas de SNPs en los descendientes para inferir regiones genómicas asociadas a un gen de interés.

Los algoritmos como Viterbi permiten hacen posible esto al ser capaces de analizar grandes volúmenes de datos de forma eficiente.

Qué significa la relación entre SNPs consecutivos

En el contexto de un HMM aplicado a SNPs, las relaciones entre SNPs consecutivos se refieren a cómo la probabilidad de un SNP de estar asociado un estado oculto (por ejemplo, "asociado al gen con el rasgo") depende del estado oculto del SNP inmediatamente anterior. La probabilidad de transición entre estados modela esta dependencia.

En otras palabras, el modelo asume que el estado de un SNP no es independiente del estado del SNP anterior, sino que existe una relación probabilística entre ellos. Esto refleja que los SNPs cercanos en el genoma están a menudo físicamente ligados. La razón biológica de esto se conoce como desequilibrio de ligamiento, el cual establece que los SNPs cercanos en el genoma tienden a heredarse juntos debido a su proximidad física en un cromosoma. Por ejemplo, si un SNP está en un estado "asociado al gen", es probable que los SNPs vecinos también estén asociados al mismo estado.

Cómo se detecta un SNP:

- **Comparación con la secuencia de referencia:**

  Cada posición del genoma se compara con la base correspondiente en una secuencia de referencia. Si una posición tiene una base diferente en un porcentaje significativo de individuos, se considera un SNP.

- **Agrupamiento de lecturas:**

  Las herramientas bioinformáticas agrupan lecturas de secuenciación en cada posición del genoma. Si una base diferente aparece con suficiente frecuencia, esa posición se clasifica como un SNP.

- **Ejemplo:**

  - Secuencia de referencia: **AAGCTG**.

- Individuo 1: **AAGCTG** (coincide con la referencia).
- Individuo 2: **AAGTTG** (variante en la posición 4: G → T).
- Individuo 3: **AAGCTG**.
- Con un 33% de variación en la posición 4, esta se clasificaría como un SNP.

- Ejemplo de identificación:
  - SNP1: Cromosoma 1, posición 100, referencia: **A**, variante: **G**.
  - SNP2: Cromosoma 1, posición 200, referencia: **T**, variante: **C**.

6. Construcción del modelo

Para empezar, se definieron los estados en los que se quería basar el modelo. Se consideraron 3 estados distintos de relación con el rasgo de interés en el gen para un SNP observado. En un escenario biológico, estos estados se asignarían dada la frecuencia con la que los descendientes heredan este estado ligado al SNP.

Diferenciación de estados:

- **Estado 1: "Asociado al progenitor con el rasgo":** Hay una base dominante que coincide con la del progenitor que presenta el rasgo.
- **Estado 2: "Asociado al progenitor sin el rasgo":** Hay una base dominante que coincide con la del progenitor que no tiene el rasgo.
- **Estado 3: "No asociado":** No hay una base dominante, y las frecuencias reflejan una herencia al azar.

Por ejemplo, si el progenitor con el rasgo tiene **A** en esa posición (que se hereda junto al rasgo) y la mayoría de los descendientes con el rasgo también presentan **A**, hay una alta probabilidad de concordancia entre el SNP y el rasgo (estado 1).

Construcción de las probabilidades de emisión

Las probabilidades de emisión reflejan cómo las frecuencias observadas de las bases (A, T, C, G) en un SNP coinciden o difieren de la base esperada en la secuencia de referencia.

Cada estado oculto (asociado/no asociado) tiene probabilidades de emisión definidas en función de la concordancia o discordancia con la secuencia de referencia.

Por ejemplo, si en un SNP, la referencia tiene la base **A** y las frecuencias observadas en los descendientes son: **80% A, 10% T, 5% C, 5% G**, se espera una alta concordancia con la base **A** si el SNP está en un estado "asociado al progenitor con el rasgo".

Por otra parte, el estado "asociado al progenitor sin el rasgo" en las probabilidades de emisión del modelo hace referencia a una situación en la que un SNP en la descendencia está más relacionado con la herencia del progenitor que no presenta el rasgo en estudio, en lugar de estar relacionado con el progenitor que sí lo tiene.

Esto significa que, para un SNP dado, la base más frecuente en los descendientes no coincide con la base que sería esperada si el rasgo estuviera presente. En términos del modelo, si un SNP está en este estado, las probabilidades de emisión reflejan que es más probable observar una base heredada del progenitor sin el rasgo, en lugar de una base asociada con el progenitor que sí presenta el rasgo.

Por ejemplo, supongamos que en la descendencia, el SNP muestra frecuencias como: **70% G, 20% A, 5% T, 5% C**. Si sabemos que el progenitor **con el rasgo** tenía una **A** en esta posición, pero el progenitor **sin el rasgo** tenía una **G**, entonces es probable que el SNP esté en este estado. Esto se debe a que los descendientes han heredado mayoritariamente la base del progenitor sin el rasgo.

Construcción de la matriz de emisión

Con esta información, se modeló una matriz de emisión sencilla suponiendo 20 SNPs distintos asignando 3 posibles estados a cada uno de ellos (*Figura 1*).

- El estado 1 favorece a los 10 primeros SNPs de la secuencia. De esta manera, la probabilidad de que los 10 primeros sean emitidos por este estado es más alta que en los 5 últimos.
- El estado 2 favorece a los últimos 15 SNPs
- El estado 3 favorece a los 20 por igual

Destaquemos que la probabilidad repartida en los 20 SNPs ha de sumar 1 para cada estado ya que estamos modelando qué tan probable es observar un SNP específico si el modelo está en un estado particular. Por ejemplo, para el Estado 1, los SNPs del 1 al 10 son más probables porque están relacionados con el gen de interés. Cada fila representando los estados es independiente de las otras dos, por eso las filas no suman 1, ya que se está modelando la probabilidad de observar un SNP emitido desde cada uno de los 3 estados posibles.

Qué representan las filas y columnas en la matriz de emisión:

- **Filas (estados):** Cada fila representa un estado oculto del modelo. Las probabilidades en cada fila indican qué tan probable es observar cada símbolo (SNP) si el modelo se encuentra en ese estado. Por eso, las filas deben sumar 1, asegurando que todas las probabilidades de emisión de un estado estén correctamente normalizadas.

- **Columnas (SNPs):** Cada columna representa las probabilidades de emisión de un SNP en todos los estados. Las probabilidades de una columna no necesariamente suman 1, ya que los estados son independientes en términos de emisión.

Construcción de las probabilidades de transición entre estados

Entenderíamos por este concepto la probabilidad de que haya un cambio de estado al saltar entre SNPs consecutivos en un gen cuando se lee la secuencia de SNPs en orden. Como ya se ha explicado estas probabilidades se modelan teniendo en cuenta que los SNPs cercanos tienden a heredarse juntos (con el mismo estado) en función del nivel de ligamiento que tengan.

Podemos observar en la *Figura 1* que esta matriz de transición cumple la regla de Markov, en la cual las probabilidades de transición entre estados sólo dependen del estado del SNP anterior.

Matrices de transición y emisión

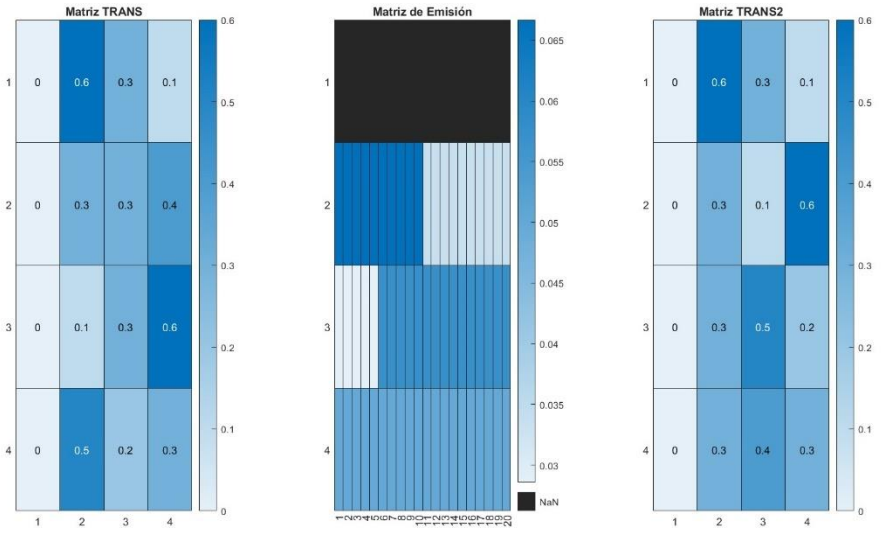

Figura 1: (izq, dcha) Dos matrices de transición distintas con distintas probabilidades. (mid) La matriz de emisión para 50 SNPs distintos. Fuente: (propia)

En este caso, tanto las probabilidades de transición como de emisión se modelaron de manera intuitiva, creando hasta dos modelos distintos para las matrices de transición, donde se tienen en cuenta distintos ligamientos entre SNPs. Normalmente las estas probabilidades se pueden modelar de manera más precisa con algoritmos como el de *Expectation-Maximization* (EM), el cual asigna probabilidades basándose en grandes cantidades de datos de secuencias de descendientes de una población en comparación con secuencias de referencia de los progenitores, comparando las frecuencias de bases entre estas secuencias genéticas y modelando las probabilidades que mayor se ajusten a estos datos (Yoon, B. J. 2009).

1. Explicación de los Algoritmos

Uno de los problemas más famosos en el contexto de los Modelos de Markov Ocultos es el de encontrar la probabilidad total de que se dé una secuencia de símbolos determinada dados los parámetros del modelo.

Para ello, se usa el *foward* algorithm, el cual busca responder: ¿Qué tan probable es observar una secuencia de SNPs específica (frecuencias de bases en descendientes) dado un modelo de Markov oculto que relaciona estados con estas observaciones?

Sabemos que cada símbolo toma una serie de valores del conjunto de observaciones $V(V_1, V_2, V_3 \cdots N)$ y que cada estado toma valores del conjunto de estados $S(S_1, S_2, S_3 \cdots M)$ donde *M* y *N* representan el número de distintas observaciones y estados del modelo. También conocemos los parámetros del modelo λ dada una secuencia de símbolos $x = (x_1, x_2, \cdots x_n)$ y de estados ocultos $y = (y_1, y_2, \cdots y_n)$, siendo:

- $t(i,j) = P\{y_{n+1} = j | y_n = i\}$, Probabilidad de transicionar del estado j al estado i
- $\pi_i = P\{y_1 = i\}$, Probabilidad de comenzar en el estado i
- $e\{x|i\} = P\{x_n = x | y_n = y\}$ Probabilidad de emitir el símbolo x desde el estado i

Dada una secuencia de observaciones (*x*) de *L*, cada una asociada a uno de los estados ocultos que forman la secuencia (*y*) dentro del conjunto de estados *M*, obedeciendo a los parámetros del modelo λ; una manera de calcular la probabilidad de que se dé la secuencia (*x)* sería computando todas las posibles secuencias *(y)* para la secuencia (*x)* y sumar las probabilidades como:

$$P\{x|\lambda\} = \sum_{y} P\{x, y|\lambda\}$$

Por desgracia, esto es muy caro computacionalmente dado que hay M^L secuencias de estados posibles. Es por ello que se usa el *foward algorithm* que

computa esta probabilidad de una manera más dinámica siguiendo tan sólo cuatro pasos:

- **Creación de la variable alfa:**

$$\alpha(n,i) = P\{x_1, x_2, \cdots x_n, y_n = i|\lambda\}$$

Siendo $\alpha(n,i)$ la probabilidad de observar una secuencia de símbolos $(x_1, x_2, \cdots x_n)$ y acabar en el estado y en el paso n dados los parámetros del modelo λ.

- **Inicialización:**

Consideramos al emisión del primer símbolo x asociado al estado i en el paso $n = 1$ dados los parámetros π_i y $e\{x|i\}$.

$$\alpha(1,i) = \pi_i \cdot e\{x|i\}$$

- **Recursión:**

Para cada paso siguiente $n = 2, \ldots L$ se calcula $\alpha(n,i)$ como una combinación de las probabilidades en el estado anterior.

$$\alpha(1,i) = \sum_k \alpha(n-1,k) \cdot t(i,j) \cdot e\{x|i\})$$

Siendo k el estado inmediatamente anterior a i, el programa evalúa en cada paso la probabilidad del estado actual teniendo en cuenta sólo en el estado anterior (todos los posibles estados M para k) y suma la probabilidad de llegar a i desde k. Esto es posible dado que el modelo sigue la regla de Markov antes mencionada.

Y así en cada paso siguiente, el programa usará los valores de α en el estado anterior.

- **Finalización:**

  Por último, quedaría calcular la probabilidad total, que no es más que la suma de las probabilidades de todos los pasos de la secuencia.

$$P\{x|\lambda\} = \sum_{i=1}^{n} \alpha(n,i)$$

De esta manera, la complejidad computacional se ve en gran parte reducida dado que si en cada paso tenemos M estados posibles para i y a su vez cada uno de ellos proviene de M estados de k posibles en $n-1$, el programa tiene que recorrer M^2 estados para cada paso, pues recordemos que sólo considera hasta el paso $n-1$ por la propiedad de Markov. Y si la suma de todos los α en el paso de finalización posee longitud L, la complejidad del cálculo de la probabilidad total es de $O(LM^2)$, significativamente más fácil de computar que con una complejidad de $O(M^L)$.

De manera aplicada, este algoritmo nos permite calcular la probabilidad total de observar una secuencia de SNPs bajo el modelo. Para identificar áreas genómicas que son consistentes con ciertos estados ocultos, el Forward Algorithm puede ser adaptado para analizar cómo evolucionan las probabilidades acumuladas para cada estado a lo largo de la secuencia de SNPs. Esto permite observar qué regiones tienen una mayor probabilidad de estar asociadas a un estado específico (por ejemplo, "asociado al rasgo").

Viterbi

Este algoritmo permite hallar la secuencia que se dará con la probabilidad máxima dados la secuencia de estados, símbolos y parámetros del modelo. Esto se conoce como el problema de decodificación global.

Su funcionamiento es similar al del *foward algorithm*:

- **Inicialización:**

Se crea una matriz $\gamma(n,i)$ con un tamaño de $L \times M$.

$\gamma(n,i)$ va a representar la probabilidad máxima de llegar a cada estado i en el paso n.

$$\gamma(n,i) = max_{y_1 \cdots y_{n-1}} P\{x_1 \cdots x_n, y_1 \cdots y_{n-1} y_n = i|\lambda\}$$

Y para el primer paso $n = 1$:

$$\gamma(1,i) = \pi_i \cdot e\{x|i\}$$

- **Recursión:**

La fórmula de la recursión es similar a la de $\alpha(n,i)$:

$$\gamma(n,i) = max_k(\gamma(n-1,k) \cdot t(i,j) \cdot e\{x|i\})$$

γ computará los caminos más probables para cada estado en cada paso (teniendo siempre en cuenta sólo el estado anterior). A su vez el programa almacenará cada estado en cada paso dentro de la variable δ para construir más tarde la ruta más probable. Cada vez que el algoritmo evalúa cuál es la ruta más probable hacia un estado i en el paso n, considera qué tan probable es cada $t(i,j)$ y escoge la más alta, y a su vez evalúa con $e\{x|i\}$ qué tan consistente es el estado actual i con la observación visible.

- **Terminación:**

Una vez se procesan todas las observaciones hasta $n = L$, el programa identifica al estado i que maximiza $\gamma(L,i)$

$$P_{máxima} = max_i \gamma(L,i)$$

- ***Backtracking:***

Con los estados almacenados en δ, se retrocede desde el último estado registrado hasta el primero para poder reconstruir la secuencia óptima.

Si bien γ presenta similitudes con α, esta variable no es acumulativa, sino que

se limita a considerar las rutas de estados más probables hasta un el último símbolo emitido de la secuencia.

Como resultado, a cada SNP en la secuencia se le asignará su estado más probable y en consecuencia, basándonos en la interpretación de los estados podemos predecir si el SNP está localizado en una región relacionada con un gen (o genes) responsables del rasgo de interés.

6. Resultados

El modelo se preparó para que generara una secuencia aleatoria de un número determinado de SNPs (símbolos) y, dados los parámetros fijados del modelo, los algoritmos hicieran su labor, representando los datos obtenidos de distintas maneras.

El hecho de comparar la misma secuencia de SNPs usando parámetros distintos puede ayudar a identificar regiones consistentes o divergentes dentro de la misma secuencia.

Las regiones consistentes o divergentes se refiere a cómo las diferentes configuraciones del modelo (por ejemplo, distintas matrices de transición o emisión) afectan las secuencias de estados predichos en ciertas partes del genoma. Son partes del genoma donde las diferentes configuraciones del modelo predicen los mismos estados ocultos para los SNPs.

Esto indica que el modelo tiene una alta confianza, con estabilidad en esas predicciones aunque los parámetros utilizados muestren ligeras diferencias, o bien que esa determinada secuencia de SNPs normalmente tenderá a estar asociada con un cierto estado independientemente de los parámetros.

Después de ejecutar el modelo se obtuvieron los siguientes resultados:

Viterbi

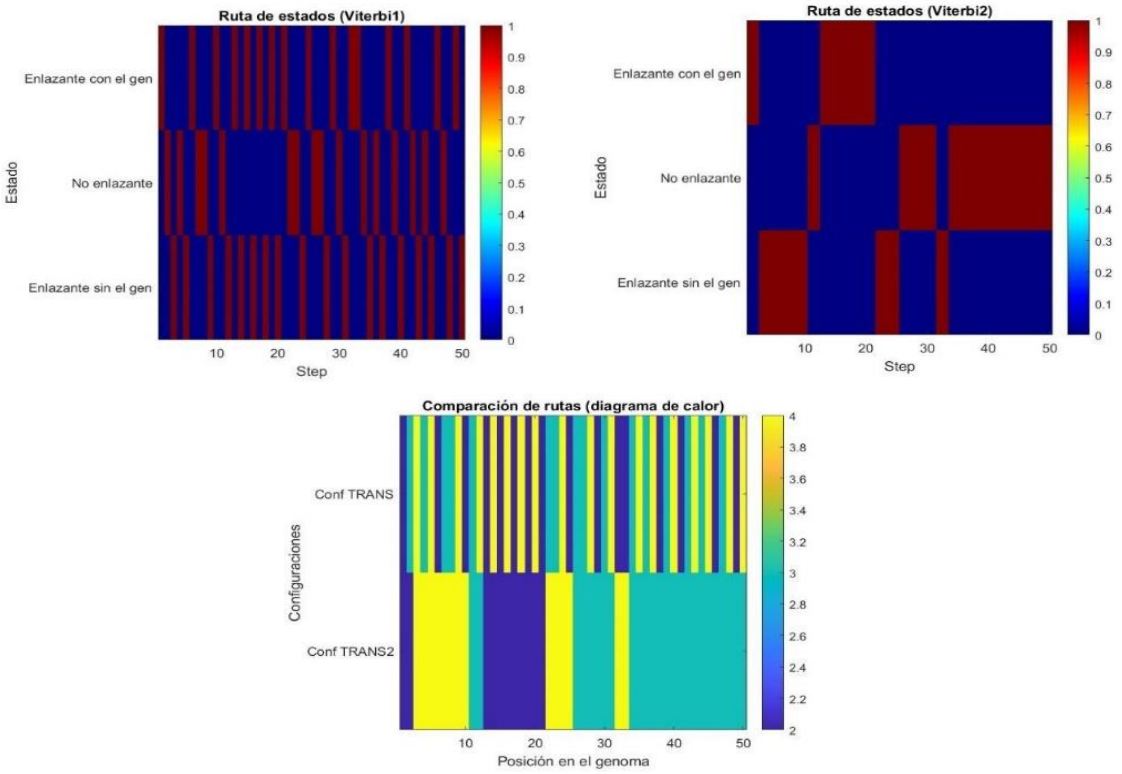

Figura 2: Comparación entre las rutas óptimas encontradas en una secuencia de 50 SNPs usando 2 matrices de transición distintas (TRANS y TRANS2)

Se escogió representar las rutas óptimas de la secuencia generada con matrices de calor. Se compararon los resultados de dos modelos con matrices de transición distintas.

Podemos observar que las configuraciones de cada una de las matrices de transición afectan significativamente a la ruta óptima encontrada por el algoritmo para la misma secuencia de SNPs. TRANS2, indica a que hay ciertas regiones de SNPs adyacentes que poseen el mismo estado, lo que biológicamente se puede asociar con un fuerte equilibrio de cruzamiento mientras que en TRANS se observa una ruta más caótica, pudiéndose asociar a un débil equilibrio de ligamiento.

Aunque podemos observar que en ciertas regiones se descartan estados para ambas configuraciones de transición (de los SNPs del 10 al 20 en

ninguna de los dos presentas estados “no enlazantes”), el hecho de que ambos algoritmos estén indicando predicciones distintas para una misma secuencia confirma la importancia de un modelo con parámetros precisos dada la alta sensibilidad a los cambios de parámetros que pueden presentar ciertas secuencias, como la estudiada en este trabajo.

Foward Algorithm

Recuérdese que este algoritmo proporciona una probabilidad acumulada para cada estado a lo largo de los pasos de la secuencia, correspondientes a los SNP.

Al analizar esta probabilidad en diferentes regiones genómicas, podemos identificar áreas que son más consistentes con ciertos estados ocultos, como estar asociados a un rasgo. Para identificar áreas genómicas que son consistentes con ciertos estados ocultos, el Forward Algorithm puede ser adaptado para analizar cómo evolucionan las probabilidades acumuladas para cada estado a lo largo de la secuencia de SNPs. Esto permite observar qué regiones de SNPs tienen una mayor probabilidad de estar asociadas a un estado específico (por ejemplo, "asociado al rasgo).

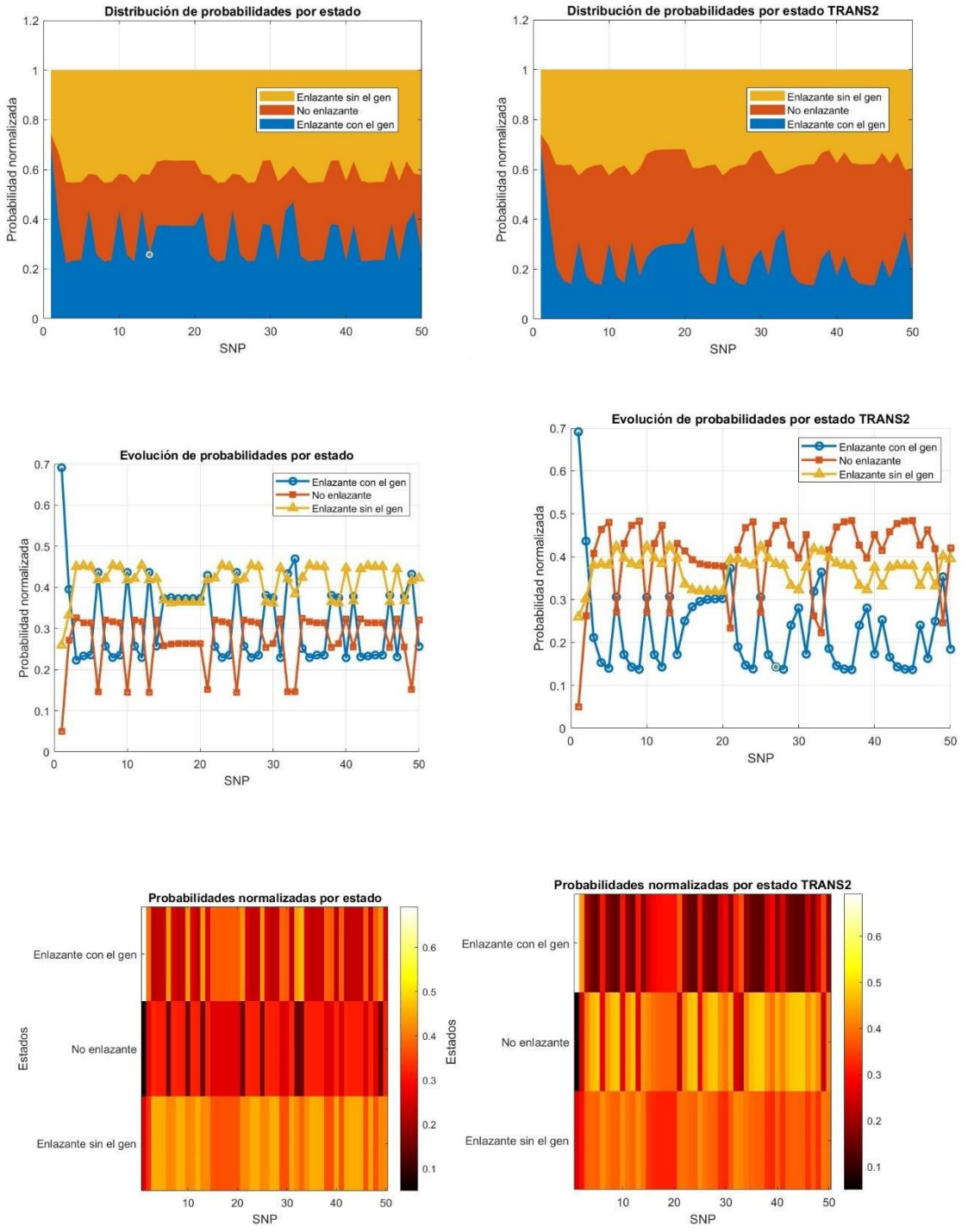

Figura 3: Representaciones de las probabilidades acumuladas de cada estado (Enlazante con el gen, No enlazante, Enlazante sin el gen) calculadas con el foward algorithm para dos configuraciones de transición distintas en la misma secuencia de SNPs. (sup)Gráfico de áreas apiladas para cada estado, (mid) Gráfico de líneas individuales para cada estado, (inf)Diagrama de calor para cada estado. (izq)TRANS. (dcha)TRANS2

Con la representación de este algoritmo, a diferencia de con Viterbi, se puede apreciar las probabilidades acumuladas de cada estado en cada SNP, lo cual nos permite analizar los estados dominantes en cada SNP a su vez que se observa la evolución de las probabilidades de todos los estados según se avanza por la secuencia.

En el gráfico de áreas apiladas, el área total nos muestra cómo los estados contribuyen conjuntamente a la probabilidad total en cada paso. Las áreas dominantes indican estados con alta probabilidad en ciertos pasos. En el caso de TRANS2 el área dominante es la naranja, el estado no enlazante; mientras que en TRANS es el amarillo, el estado enlazante con el gen. Se puede confirmar de igual manera en los gráficos de líneas individuales y diagramas de calor.

Esto coincide con las predicciones de secuencias óptimas en el algoritmo de Viterbi. No obstante, en los diagramas de calor se puede visualizar con facilidad una región consistente de los SNPs 10 a 20, en los cuales está marcada en ambas configuraciones con probabilidades ligeramente similares entre los tres estados, presentando un color anaranjado muy similar en los tres. Esto es imposible de captar con Viterbi ya que sólo devolverá la probabilidad máxima pero en las representaciones del *foward algorithm* se puede identificar un patrón con esa región específica, en el que los 3 estados tienden a volverse equiprobables, aunque lejos de ello en este caso. Esto se aprecia con más detalle en el gráfico de líneas individuales, donde se ve que los tres estados tienden a converger (más pronunciadamente en TRANS2) y a quedarse estacionarios.

Esto nos puede dar pistas de la existencia de secuencias de SNPs concretos que en un gen, pueden ser consistentes y siempre llevar los mismos estados asociados.

7. Conclusiones

Esto nos puede dar pistas de la existencia de secuencias de SNPs concretos que en un gen, pueden ser consistentes y siempre llevar los mismos estados asociados. En este trabajo, se ha desarrollado y aplicado un modelo probabilístico basado en Cadenas de Markov ocultas (HMM, por sus siglas en inglés) para analizar la relación entre variantes genómicas individuales (SNPs) y un rasgo de interés en una población de estudio. La metodología presentada permite desentrañar las regiones del genoma más consistentes con estados genotípicos particulares, proporcionando una herramienta valiosa para estudios de asociación genética y mapeo de loci. El uso de HMMs en este contexto se justifica por su capacidad para modelar da os secuenciales con dependencia entre observaciones adyacentes, lo cual es esencial en el análisis de genomas, ya que la información genética suele estar correlacionada a lo largo de las regiones cromosómicas. Los resultados obtenidos a partir del algoritmo de Viterbi y la Forward Algorithm destacan cómo las probabilidades de transición y emisión pueden combinarse para identificar rutas óptimas y asignar probabilidades acumuladas a estados ocultos en cada posición genómica. Estas herramientas permiten no solo inferir las regiones más probablemente asociadas con un rasgo, sino también medir la confianza en estas inferencias.

Al observar gráficas como las de áreas apiladas, se logró identificar estados dominantes en diferentes regiones del genoma. Este análisis visual ofrece una forma intuitiva de explorar los datos y resaltar diferencias genómicas entre individuos con y sin el rasgo estudiado.

La comparación de rutas óptimas mediante el algoritmo de Viterbi también reveló que las configuraciones del modelo (por ejemplo, parámetros iniciales y distribución de estados) tienen un impacto directo en la interpretación de los resultados. La visualización de rutas genómicas divergentes permite identificar regiones de especial interés donde las hipótesis del modelo podrían necesitar ajustes o donde la señal genómica sugiere la acción de procesos biológicos más complejos.

En un contexto más amplio, este enfoque puede ser aplicado a una variedad de estudios genómicos, desde el mapeo de enfermedades hasta la selección genética en agricultura o ganadería. Su flexibilidad permite ajustar los parámetros del modelo y las probabilidades de emisión para adaptarse a distintos organismos o datos experimentales. Asimismo, el análisis automatizado de rutas genómicas podría integrarse con otras herramientas de inteligencia artificial, permitiendo descubrimientos más rápidos y precisos en estudios de gran escala.

Este trabajo no solo resalta el poder de los HMMs para el análisis genómico, sino que también subraya la importancia de un enfoque interdisciplinario para resolver preguntas biológicas complejas. La integración de modelos probabilísticos con técnicas de visualización y herramientas computacionales representa un paso significativo hacia una mejor comprensión de la base genética de los rasgos fenotípicos y su relación con variantes genómicas específicas.

8. Bibliografía

Cohen, A. (1998, November). Hidden Markov models in biomedical signal processing. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286)* (Vol. 3, pp. 1145-1150). IEEE.

Hur, P., Shorter, K. A., Mehta, P. G., & Hsiao-Wecksler, E. T. (2012). Invariant density analysis: Modeling and analysis of the postural control system using markov chains. *IEEE Transactions on Biomedical Engineering*, *59*(4), 1094-1100.

Khalifa, Y., Mandic, D., & Sejdić, E. (2021). A review of Hidden Markov models and Recurrent Neural Networks for event detection and localization in biomedical signals. *Information Fusion*, *69*, 52-72.

Yoon, B. J. (2009). Hidden Markov models and their applications in biological sequence analysis. *Current genomics*, *10*(6), 402-415.

9. Anexo

```
%Queremos tres estados: Enlazante con el gen(1), No enlazante(2),
Enlazante sin el gen(3)

%Probabilidades iniciales

PI = [0, 0.6, 0.1, 0.3]; %Si las distribuciones iniciales son iguales, al
normalizar no habrá diferencia entre EMIS y EMIS_ADJ

%Matrices de transición para cada
caso TRANS = [0.3, 0.3, 0.4;

         0.1, 0.3, 0.6;

         0.5, 0.2, 0.3];

TRANS_2 = [0.5, 0.2, 0.3;

           0.2, 0.5, 0.3;

           0.1, 0.4, 0.5];

%%

%Matrices de emisión

%Se hace que cada uno de los estados (filas) tenga una probabilidad de


%emisión en ciertos SNP dados los 20 que hay. Todas han de sumar 20, de ahí

%que se haga de 1, 10 a 1, 10 etc

EMIS = [0.1*ones(1, 10) 0.05*ones(1, 10);  % Estado 1 favorece primeros 10
SNPs
0.05*ones(1, 5) 0.1*ones(1, 15);% Estado 2 favorece los últimos 15

0.05*ones(1, 20)];                 % Estado 3 emite todos los SNPs igual
```

```
%% %Los estados son independientes entre sí en términos de emisión, por eso
%las columnas no suman 1. Por ejemplo, el SNP nº1, tiene una probabilidad
%de 0.1 de emitirse en el estado 1, y un 0.9 de no emitirse en ese estado,
%todo independiente de su probabilidad de emisión en el estado 2.
%%
%Creamos matrices ampliadas para integrar PI
TRANS_HAT = [PI; zeros(size(TRANS,1),1) TRANS]; %Añado una fila con PI y una
columna de ceros para que sea cuadrada
%Para TRANS_2
TRANS_HAT2 = [PI; zeros(size(TRANS_2,1),1) TRANS_2];
%Para EMIS
EMIS_HAT = [zeros(1,size(EMIS,2)); EMIS]; %Añado una fila de ceros para que
sea compatible con TRANS_HAT
% Normalizar las matrices ampliadas
EMIS_HAT = EMIS_HAT ./ sum(EMIS_HAT, 2); % Normalizar la matriz EMIS_HAT
TRANS_HAT = TRANS_HAT ./ sum(TRANS_HAT, 2); % Normalizar la matriz
TRANS_HAT TRANS_HAT2 = TRANS_HAT2 ./ sum(TRANS_HAT2, 2); % Normalizar la
matriz TRANS_HAT2
%Comprobación de que las matrices ampliadas son
correctas sum(TRANS_HAT, 2);  % Debe devolver 1 para
todas las filas
sum(EMIS_HAT, 2); % Debe devolver 1 para todas las filas (excepto la
fila inicial)
%%
%Gráficos de calor para visualizar las matrices
figure;
subplot(1,3,1);
```

```
%% heatmap(TRANS_HAT, 'Title', 'Matriz TRANS');
subplot(1,3,2);

heatmap(EMIS_HAT, 'Title', 'Matriz de Emisión');
subplot(1,3,3);

heatmap(TRANS_HAT2, 'Title', 'Matriz TRANS2');

%%

%Generamos secuencia de observaciones

n = input('Número de observaciones: '); %Longitud de la secuencia de
observaciones (steps)

[seq, states] = hmmgenerate(n, TRANS_HAT, EMIS_HAT); %Generamos la secuencia

%La función hmmgenerate me va a generar una secuencia aleatoria de 1000
emisiones y estados

%teniendo en cuenta las matrices TRANS y EMIS.

%seq es la secuencia de 1000 observaciones (SNPS con valores del 1 al 20

%states es la secuencia de 1000 estados con valores del 2 al 4 porque el 1

%es el estado ficticio

%La función comienza en el estado 1 en el step 0 y transiciona al estado 1
en el step 0 por deafult

%%

%Calculamos la ruta más óptima con Viterbi

Vit1 = hmmviterbi(seq, TRANS_HAT, EMIS_HAT); %No usa PI porque asume que
el modelo comienza en el estado 1

%Hay que multipilicar la matriz de EMIS por PI para quitar el sesgo del
programa (ver más arriba)

Vit2 = hmmviterbi(seq, TRANS_HAT2, EMIS_HAT);

%%

%Calculamos con la foward variable la probabilidad de observar esta
```

```
%% %secuencia de observaciones generada. Recuerda que el foward modela la
%probabilidad de observar un símbolo en un estado t. De tal forma que si
%itera eso para todos los t, me da la probabilidad de una secuencia
%completa.
% En escala logarítmica dado que a secuencias muy grandes, la probabilidad
tiende a 0
potito = pr_hmmMINE2(seq, TRANS_HAT, EMIS_HAT, PI); % Escala lineal con
secuencias cortas
potito
% Calcular la matriz de probabilidades Forward
% Llamar a la función ajustada
[p, m] = pr_hmmMINE2(seq, TRANS_HAT, EMIS_HAT, PI);


% Excluir el estado ficticio
m_real = m(:, 2:end); % Eliminar la columna correspondiente al estado
ficticio
%Gráfico de áreas apiladas de los distintos estados
%Definimos un alpha
[~, alpha] = pr_hmmMINE2(seq, TRANS_HAT, EMIS_HAT, PI);
alpha_normalized = alpha ./ sum(alpha, 2); %Lo normalizamos
%Se puede identificar regiones donde un estado domina analizando alpha a lo
%largo de los SNPs. Recuerda que alpha en foward me da una probabilidad
%para cada
estado. figure;
area(alpha_normalized(:, 2:end), 'LineStyle', 'none'); % Excluye el estado
ficticio
```

```
%% xlabel('SNP');

ylabel('Probabilidad normalizada');
title('Distribución de probabilidades por
estado');

legend({'Enlazante con el gen', 'No enlazante', 'Enlazante sin el gen'},
'Location', 'Best');

colormap('parul
a'); grid on;

%Gráfico de
líneas figure;

hold on;

plot(alpha_normalized(:, 2), '-o', 'LineWidth', 2, 'DisplayName',
'Enlazante con el gen');

plot(alpha_normalized(:, 3), '-s', 'LineWidth', 2, 'DisplayName', 'No
enlazante');

plot(alpha_normalized(:, 4), '-^', 'LineWidth', 2, 'DisplayName',
'Enlazante sin el gen');

xlabel('SNP');

ylabel('Probabilidad normalizada');
title('Evolución de probabilidades por
estado'); legend show;

grid
on;
hold
off;

%Gráfico de calor
compacto figure;

imagesc(alpha_normalized(:,
2:end)'); colormap('hot');

colorbar;
```

```
%% xlabel('SNP');
ylabel('Estados
');
yticks(1:3);

yticklabels({'Enlazante con el gen', 'No enlazante', 'Enlazante sin el
gen'}); title('Probabilidades normalizadas por estado');

%%

%Lo mismo para TRANS2

% Calcular la matriz de probabilidades Forward

% Llamar a la función ajustada

[p, m] = pr_hmmMINE2(seq, TRANS_HAT2, EMIS_HAT, PI);

% Excluir el estado ficticio

m_real = m(:, 2:end); % Eliminar la columna correspondiente al estado
ficticio

%Gráfico de áreas apiladas de los distintos estados

%Definimos un alpha

[~, alpha2] = pr_hmmMINE2(seq, TRANS_HAT2, EMIS_HAT, PI);
alpha_normalized = alpha2 ./ sum(alpha2, 2); %Lo normalizamos

%Se puede identificar regiones donde un estado domina analizando alpha a lo

%largo de los SNPs. Recuerda que alpha en foward me da una probabilidad

%para cada
estado. figure;

area(alpha_normalized(:, 2:end), 'LineStyle', 'none'); % Excluye el estado
ficticio

xlabel('SNP');

ylabel('Probabilidad normalizada');

title('Distribución de probabilidades por estado TRANS2');
```

```
%% legend({'Enlazante con el gen', 'No enlazante', 'Enlazante sin el
gen'}, 'Location', 'Best');

colormap('parul
a'); grid on;

%Gráfico de
líneas figure;

hold on;

plot(alpha_normalized(:, 2), '-o', 'LineWidth', 2, 'DisplayName',
'Enlazante con el gen');

plot(alpha_normalized(:, 3), '-s', 'LineWidth', 2, 'DisplayName', 'No
enlazante');

plot(alpha_normalized(:, 4), '-^', 'LineWidth', 2, 'DisplayName',
'Enlazante sin el gen');

xlabel('SNP');

ylabel('Probabilidad normalizada');

title('Evolución de probabilidades por estado
TRANS2'); legend show;

grid
on;
hold
off;

%Gráfico de calor
compacto figure;

imagesc(alpha_normalized(:,
2:end)'); colormap('hot');

colorbar;
xlabel('SNP');
ylabel('Estados
');
yticks(1:3);
```

```
%% yticklabels({'Enlazante con el gen', 'No enlazante', 'Enlazante sin el
gen'}); title('Probabilidades normalizadas por estado TRANS2');

%%

%Ploteamos Vit1 con diagrama de calor

%

figure;
subplot(1,1,1);

heatmap(Vit1, 'Title', 'Matriz de Vit1');

%Vit2
figur
e;

subplot(1,1,1);

heatmap(Vit2, 'Title', 'Matriz de Vit2');

%Hay que buscar una manera de comparar Viterbi con otras secuencias

%plotteando la probabilidida de cada step en el diagrama

%Diagrama para la secuencia de estados Vit1

% Crear matriz para el diagrama de calor

n_steps = length(Vit1); % Longitud de la secuencia

heatmap_data = zeros(3, n_steps); % Tres estados reales (excluye el estado
ficticio)

for t = 1:n_steps

    heatmap_data(Vit1(t)-1, t) = 1; % Resta 1 para mapear estados 2, 3, 4 a
    filas

1, 2, 3

end

% Crear el gráfico de calor con
imágenes figure;
```

```
%% imagesc(heatmap_data);



% Cambiar el esquema de colores

colormap('jet'); % Esquema de colores más vistoso
colorbar; % Agrega una barra de colores para
referencia

% Ajustar los
ejes
xlabel('Step');
ylabel('Estado'
);

title('Ruta de estados (Viterbi1)');

% Ajustar las etiquetas del eje Y
yticks(1:3); % Fila 1 a 3

yticklabels({'Enlazante con el gen', 'No enlazante', 'Enlazante sin el
gen'}); ax = gca; % Obtener el objeto del eje

ax.YAxis.TickLabelInterpreter = 'none'; % Asegurar que se respeten las
etiquetas

%Diagrama para la secuencia de estados Vit2

% Crear matriz para el diagrama de calor

n_steps = length(Vit2); % Longitud de la secuencia

heatmap_data = zeros(3, n_steps); % Tres estados reales (excluye el estado
ficticio)

for t = 1:n_steps

    heatmap_data(Vit2(t)-1, t) = 1; % Resta 1 para mapear estados 2, 3, 4 a
    filas

1, 2, 3

end
```

```
%% % Crear el gráfico de calor con
imágenes figure;

imagesc(heatmap_data);

% Cambiar el esquema de colores

colormap('jet'); % Esquema de colores más vistoso
colorbar; % Agrega una barra de colores para
referencia

% Ajustar los
ejes
xlabel('Step');
ylabel('Estado'
);

title('Ruta de estados (Viterbi2)');

% Ajustar las etiquetas del eje Y
yticks(1:3); % Fila 1 a 3

yticklabels({'Enlazante con el gen', 'No enlazante', 'Enlazante sin el
gen'}); ax = gca; % Obtener el objeto del eje

ax.YAxis.TickLabelInterpreter = 'none'; % Asegurar que se respeten las
etiquetas

%Comparación de rutas

% Generar posiciones ficticias

positions = 1:length(Vit1); % Suponemos que ambas rutas tienen la misma
longitud

% Diagrama de calor para comparar
rutas comparison_data = [Vit1;
Vit2]; figure;

imagesc(comparison_da
ta);
colormap('parula');
colorbar;
```

```
%%  xlabel('Posición en el
    genoma');
    ylabel('Configuraciones');
    yticks(1:2);

    yticklabels({'Conf TRANS', 'Conf TRANS2'});
    title('Comparación de rutas (diagrama de calor)');
```