Reportes Científicos de la FACEN

Artículo Original

Reportes científicos de la FACEN, Enero - Junio 2024, 15(1):15–32

http://doi.org/10.18004/rcfacen.2024.15.1.015

Cadenas de Markov ocultas discretas y los experimentos mendelianos

Discrete hidden Markov chains and mendelian experiments

César Daniel Amarilla[1,*]

[1]Universidad Nacional de Asunción, Facultad de Ciencias Exactas y Naturales, Departamento de Matematica y Estadistica, Campus Universitario, San Lorenzo-Paraguay. *Email: ceamarilla@gmail.com.

Resumen: La naturaleza presenta innumerables fenómenos envueltos de incertidumbre. El modelado de estos fenómenos constituye uno de los pilares de la estadística matemática en la formulación de estrategias de estimación. Los modelos ocultos de Markov forman una clase muy variada de modelos que han sido muy estudiados y aplicados en muy diversas ramas de las ciencias, desde la Ingeniería en telecomunicaciones hasta la Biología. Desde el punto de vista de la Biología, los perfiles basados en las cadenas de Markov ocultas son extensiones no triviales de los modelos de perfil usuales. Este reporte científico estudia herramientas matemáticas necesarias para la solución de los problemas que presentan los modelos ocultos de Markov al ser ajustados en los experimentos mendelianos. Los resultados demuestran la validez de la base teórica matemática de estos modelos como estrategias de estimación de fenómenos biológicos.

Palabras clave: *incertidumbre, modelos ocultos de Markov, experimentos mendelianos, fenómenos biológicos.*

Abstract: Nature provides innumerable phenomena where uncertainty is involved. The modeling of these phenomena is the goal of statistics as a research field. Hidden Markov models form a very wide class of models that have been widely studied and applied in many different branches of science, from engineering to biology. From the viewpoint of biology, profiles based on hidden Markov chains are nontrivial extensions of the usual profile models. In this scientific report, we study some of the mathematical tools needed to solve problems posed when fitting hidden Markov models for subsequent application in mendelian experiments. Our results reinforce the validity of the theoretical basis of these models as mathematical estimation strategies for biological phenomena.

Keywords: *uncertainty, hidden Markov models, mendelian experiments, biological phenomena.*

Introducción

Esta investigación se ha basado en los trabajos de reseña de Forney (1973), Juang and Rabiner (1986) y (1990) y Cappé et al. (2007). Los experimentos mendelianos fueron desarrollados siguiendo los trabajos de Bilmes (1997), Thorvaldsen (2005) y Krogh el al. (1994). Las teorías necesarias para los cálculos de las probabilidades fueron tomadas de Bilmes (1997).

El comportamiento de la mayoría de los fenómenos físicos o dinámicos están gobernadas por la incertidumbre debido a la inestabilidad de las variables dentro de los procesos en el transcurso de su evolución. Esto obliga a los investigadores a desarrollar metodologías de estimación acerca del comportamiento de estos fenómenos, que les permiten conocer o al menos reducir la incertidumbre acerca de los mismos. Estas metodologías de estimación pueden consistir en la estructuración de modelos estadísticos matemáticos que permitan la descripción y comprensión del comportamiento interno de las propiedades de las variables involucradas.

Las características fundamentales en la formación de modelos estadísticos son las variables aleatorias intervinientes en el proceso estocástico en estudio y sus distribuciones de probabilidades asociadas, lo que hace necesario el conocimiento del espacio probabilístico del cual proviene. La teoría de la probabilidad es una de las teorías que ayuda a los investigadores en su afán de comprensión de la incertidumbre.

Las cadenas de Markov forman parte de los procesos estocásticos dependientes. Uno de los modelos estadísticos markovianos más utilizados en los últimos tiempos en diversas ramas de la ciencia, como ser en la Biología, constituyen las Cadenas de Markov Ocultas o sus siglas en inglés HMM

Editor responsable: Nery López Acosta* **Recibido:** 21/11/2023 **Aceptado:** 26/12/2023

*Universidad Nacional de Asunción, Facultad de Ciencias Exactas y Naturales, Dirección de Investigación, San Lorenzo, Paraguay.

(Hiddens Markov Models). Las características generales de estos modelos constituyen sus elementos y cada uno de los tres problemas fundamentales relacionados con ellos (evaluación, descodificación y aprendizaje). Cada uno de los tres problemas en uno modelo de Markov oculto puede ser estudiado mediante la utilización de ciertos algoritmos.

Materiales y métodos

Cadenas de Markov

En la realidad existen sistemas dinámicos que van evolucionando en el transcurso del tiempo y están gobernados por ciertas condiciones que pueden estar estructuradas sobre una base determinística. Por ello, las propiedades de los procesos físicos no cambian con el paso del tiempo, o sobre una base aleatoria, en cuyo caso los procesos se rigen por leyes probabilísticas. En general, un proceso estocástico es un modelo matemático que describe el comportamiento de un sistema dinámico sometido a un fenómeno de naturaleza aleatoria que hace que el sistema evolucione según un parámetro que normalmente es el tiempo, dado por la secuencia n = 1, 2, ..., t, ..., y que va cambiando probabilísticamente de un estado a otro.

Un tipo especial de proceso estocástico constituyen las cadenas de Markov, que fueron descubiertas por el matemático ruso Andrei Andreyevich Markov (1956 - 1922) alrededor de 1905, y que pueden aplicarse a una amplia gama de fenómenos científicos y sociales, y se cuenta con una teoría matemática extensa al respecto.

Espacio de estados y probabilidades de transición

Una cadena de Markov es un proceso estocástico a tiempo discreto dado por el conjunto $\{X_n : n = 0,1, ...\}$, con un espacio de estados discreto $S = \{S_1, S_2, ...\}$, y que satisface la propiedad de Markov. Esto es, para cualquier entero $n \geq 0$, y para cualesquiera estados $x_0, ..., x_n, x_{n+1}$, se cumple que:

$$p(x_{n+1} \mid x_0, ..., x_n) = p(x_{n+1} \mid x_n) \quad (1)$$

Si el tiempo $n + 1$ se considera como un tiempo futuro, el tiempo 0, 1, ... n − 1 como el presente y los tiempos 0, 1, ... n − 1 como el pasado, entonces la condición dada por las cadenas de Markov establece que la distribución de probabilidad del estado del proceso al tiempo futuro $n + 1$ depende únicamente del estado del proceso al tiempo n, y no depende de los estados en los tiempos pasados 0, 1, ... n − 1.

La probabilidad $p(X_{n+1} = j \mid X_n = i)$ se denota por $p_{ij}(n+1, n)$, y representa la probabilidad de transición del estado j al estado j en el tiempo $n + 1$. Son conocidas como las probabilidades de transición en un paso. Cuando los números n no dependen de n se dice que la cadena es estacionaria u homogénea en el tiempo. Por simplicidad se asume tal situación de modo que las probabilidades de transición en un paso se escriben como P_{ij}. De esta manera, dada una cadena de Markov $\{X_n : n = 0,1,...\}$ con espacio de estados S, la función P_{ij} con $i, j \in S$ es llamada función de transición de la cadena y está dada por:

$$p_{ij} = P(X_1 = j \mid X_0 = i) \quad (2)$$

La función de transición de la cadena cumple con las siguientes propiedades: $p_{ij} \geq 0, \forall \{i, j\} \in S$, esto es por definición de probabilidad y $\sum_j p_{ij} = 1$, para cada i.

Matriz de transición de probabilidades

Cuando el espacio de estados es finito, esto decir $S = \{0,1,..., N\}$, la función de probabilidad P_{ij} de dicha cadena puede ser expresada mediante la matricial cuadrada P de la siguiente manera.

$$P = \begin{pmatrix} p_{00} & \cdots & p_{0N} \\ \vdots & \ddots & \vdots \\ p_{N0} & \cdots & p_{NN} \end{pmatrix}$$

Esta matriz captura la esencia del proceso y determina el comportamiento de la cadena en cualquier tiempo futuro. La entrada (i, j) es la pro-

babilidad de transición P_{ij}, es decir la probabilidad de pasar del estado *i* al estado *j* en una unidad de tiempo. El índice *i* se refiere al renglón de la matriz y el índice *j* a la columna. Esta matriz cumple las siguientes condiciones: $p_{ij} \geq 0$ y $\sum_{j=1}^{N} p_{ij} = 1$. Si además satisface la condición $\sum_{i=1}^{N} p_{ij} = 1$ dice que es una matriz doblemente estocástica.

Distribución de probabilidad inicial

En general puede considerarse que una cadena de Markov inicia su evolución partiendo de un estado *i* cualquiera, o más generalmente considerando una distribución de probabilidad inicial sobre el espacio de estados.

Una distribución inicial para una cadena de Markov con espacio de estados dado por $S=\{0,1,...\}$ es simplemente una distribución de probabilidad sobre este conjunto, es decir una función π_i, que corresponde a la probabilidad de que la cadena inicie en el estado *i*, tal que:

$$\pi_i = P(X_0 = i),\ \forall i \in S \quad (3)$$

Si el espacio de estados es finito, es decir $S=\{0,1,2,...,N\}$, entonces la distribución inicial podría verse como una *n*-tupla aleatoria $\Pi=(\pi_0,...,\pi_N)$.

Caracterización de la propiedad markoviana en un proceso estocástico

En todo proceso estocástico que cumpla con la propiedad de Markov es posible calcular la distribución conjunta de cualquier secuencia finita de variables aleatorias en el proceso. Esto es, si $X_0, X_1, ..., X_n$ constituye una secuencia de variables aleatorias se tiene que:

$$p(x_0\, x_1, ..., x_n) = \pi_0 \prod_{i=1}^{n} p(x_i \mid x_{i-1}) \quad (4)$$

Dígrafos de transición

Un dígrafo o grafo dirigido es una tripleta (S, E, I), donde *S* es un conjunto cuyos elementos son llamados vértices, *E* otro conjunto cuyos elementos son las aristas y finalmente *I* es una función que le asocia a cada arista $e \in E$ un par ordenado de vértices llamados extremo de *e*. El primer vértice es llamado la cola y el segundo la cabeza de *e*. Generalmente un dígrafo es dibujado en forma tal que cada vértice queda representado por un punto en el plano, y cada arista por una curva que une los representantes de sus extremos. Para distinguir cabeza de cola, dibujamos una flecha en la cabeza de la arista (Fig. 1).

Las cadenas de Markov pueden ser representadas mediante gráficos denominados dígrafos de transición o grafos dirigidos, tal como lo muestra la Fig. 2. Estas cadenas suelen ser analizadas mediante la utilización de grafos dirigidos con pesos, que están constituidos por una terna (S, E, I), donde los elementos de *S* (conjunto de nodos) son los estados de la cadena de Markov. Los elementos del conjunto *E* (conjunto de aristas) constituyen arcos, simbolizados $e=(i,j)$, a los cuales son asignadas probabilidades no nulas de transición, $p_{ij} > 0$. Las

Figura 1: Dígrafo totalmente dirigido.

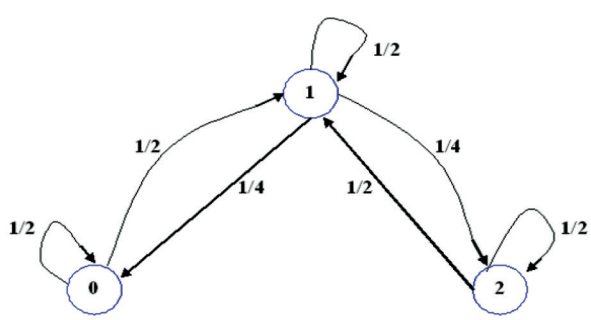

Figura 2: Dígrafo de transición.

aristas pueden también presentarse como bucles o lazos, situaciones en las cuales los arcos empiezan y terminan en un mismo nodo.

Evolución de las cadenas de Markov

Una cadena de Markov con espacio de estados finito $S=\{0,1,...,N\}$ y distribución inicial $\Pi=(\pi_0,...,\pi_N)$ puede dividirse en sectores sobre el espacio de estados de acuerdo a un criterio fácilmente observable; pero esta división en subconjuntos no es estable ya que va evolucionando en el tiempo y el proceso queda establecido en *N* sectores o estados del proceso y en un instante dado la condición de la situación es

$$V_k=\left(v_k(0),v_k(1),...,v_k(N)\right),\forall k\in\{0,1,...\},$$

donde cada $v_k(i)$, $\forall i\in\{0,1,...,N\}$, indica el tamaño del sector *i*-ésimo en el instante *k*, ya sea en magnitud absoluta, en porcentaje del total, o en porción de unidad.

Al utilizar porciones de la unidad para representar las componentes $V_k(i)$ de V_k, se tendrá que cada V_k es una distribución de probabilidad, sus componentes son no negativas y suman 1. Como la observación no se realiza en todo momento, sino al término de lapsos iguales de tiempos, se tendrá que V_k depende del período en que se realiza la observación durante el proceso, es decir es una función de variable que va tomando valores en los números naturales 0,1,... y arroja una sucesión de distribuciones Π, V_1, V_2, … , V_k, V_{k+1}, ..., donde Π indica cómo es cada sector en el inicio del proceso y se supone que la distribución V_k depende linealmente de la distribución anterior V_{k+1}, es decir existe una matriz de probabilidades de transición *P* de orden de $(n\times n)$ tal que:

$$V_k=V_{k-1}.P=\Pi\ .P^{(k)} \quad (5)$$

Con todo lo expresado, se puede concluir que cualquier secuencia observada en una cadena de Markov queda totalmente especificada por la distribución inicial y la matriz de probabilidades de transición.

Probabilidades de transición en n pasos

En una cadena de Markov $\{X_n:n=0,1,...\}$, con espacio de estados *S* y función de transición dada por $p_{ij}=P(X_1=j\mid X_0=i)$, $P(X_{m+n}=j\mid X_m=i)$ corresponde a la probabilidad de pasar del estado *i*, en el tiempo *m*, al estado *j*, en el tiempo *m* + *n*. Dado el supuesto de la condición de homogeneidad en el tiempo, esta probabilidad no depende realmente del tiempo *m*, por lo que coincide con la probabilidad $P(X_n=j\mid X_0=i)$, y se denota por $p_{ij}(n)$ o $p_{ij}^{(n)}$, donde el número de pasos *n* se escribe entre paréntesis para distinguirlo de algún posible exponente, y se llama probabilidad de transición en *n* pasos.

Si el espacio de estados es finito, las probabilidades de transición pueden ser expresadas mediante la matriz de transición de probabilidades *P* y por lo tanto, la función de transición en *n* pasos es igual a la potencia *n*-ésima de la matriz *P*.

Comunicación

Para que dos estados en una cadena de Markov se comuniquen necesariamente debe darse la propiedad de accesibilidad entre ellos. De esta forma, un estado *j* es accesible desde un estado *i* cuando se tenga un entero $n\geq 0$ tal que $p_{ij}^{(n)}\geq 0$. Por lo tanto, los estados *i* y *j* se comunican si *i* es accesible desde el estado *j* y *j* es accesible desde el estado *i*. La comunicación entre dos estados constituye la posibilidad de pasar de un estado a otro en un número finito de transiciones e indica una partición del espacio de estados de la cadena dada por los subconjuntos de estados comunicantes. Esto es, dos estados pertenecen al mismo elemento de la partición si y solo si, tales estados se comunican.

De este modo el espacio de estados de una cadena de Markov se subdivide en clases de comunicación.

Como la propiedad de accesibilidad constituye una relación de equivalencia, se cumple que: todo estado *i* es accesible desde si mismo, si el estado *i*

es accesible desde el estado *j* se tiene que el estado *j* es accesible desde el estado *i* y si el estado *i* es accesible desde el estado *j* y el estado *j* es accesible desde el estado *k* entonces el estado *i* es accesible desde el estado *k*.

Estados absorbentes, recurrentes y transitorios

En una cadena de Markov, un estado *i* es absorbente si $P_{ij} = 1$. Mientras que un estado *i* es recurrente si la probabilidad de eventualmente regresar a *i*, partiendo de *i*, es uno, es decir si:

$$P(X_n = i \text{ para alguna } n \geq 1 \mid X_0 = i) = 1 \quad (6)$$

Todo estado absorbente es un estado recurrente. Un estado que no sea recurrente recibe el nombre de transitorio, y en tal caso la probabilidad anterior es menor a uno.

Estados límites y de equilibrio

Dada una cadena de Markov cuya matriz de probabilidades de transición es *P*, sea *V* una distribución de probabilidad. Diremos que *V* es la distribución de equilibrio del proceso, si se verifica que

$$V = V\,P \quad (7)$$

Si un proceso llegara a una distribución de equilibrio, se haría constante, es decir, todas las distribuciones posteriores serían iguales.

Por otra parte, sea $\{X_n : n = 0, 1, ...\}$ una cadena de Markov con espacio de estado finito que tiene una distribución inicial Π y matriz de transición *P*. De acuerdo a la noción de convergencia, cabe preguntarse si existe:

$$V_\infty = \lim_{k\to\infty} V_k = \Pi \lim_{k\to\infty} P^{(k)} = \Pi\, P^{(\infty)} \quad (8)$$

En caso de existir, se dice que V_∞ es la distribución límite o de equilibrio de la cadena de Markov.

Regularidad

Las cadenas de Markov regulares son cadenas finitas que cumplen con la propiedad Markoviana y que a partir de un cierto momento un vector de estado pasa a otro cualquiera con probabilidades estrictamente positivas en un paso. Esto es, una cadena de Markov finita tiene matriz de transición regular si existe un entero no negativo, $n \geq 0$, tal que $p_{ij}(n) > 0$, para cualesquiera estados *i* y *i*. Entonces, una cadena de Markov será regular si alguna potencia de su matriz de probabilidades de transición tiene todas sus entradas estrictamente positivas.

Cadenas de Markov ocultas

La idea general de un Modelo Oculto de Markov (HMM) también llamada cadena de Markov oculta está en el hecho de que, al medir o cuantificar cierto fenómeno, la lectura que genera la secuencia observable de símbolos no necesariamente es el proceso real del fenómeno, ya que el instrumento de medición podría estar introduciendo un ruido en la verdadera señal. Por esta razón, este tipo de modelamiento introduce dos procesos: uno observable y otro oculto. En este modelo el proceso oculto es una cadena de Markov homogénea con espacio de estados finito y no se requiere que el proceso observado cumpla con la propiedad de Markov, sino que sea simplemente un proceso estocástico.

Los modelos ocultos de Markov (HMM) fueron introducidos a finales de los sesenta y principios de los setenta por Leonard E. Baum en un artículo, el cual propuso este modelo como un método estadístico de estimación de las funciones probabilísticas de las cadenas de Markov.

Desde finales de 1970 cuando los modelos ocultos de Markov fueron aplicados a sistemas de reconocimiento de voz, se desarrollaron técnicas para estimar probabilidades sobre estos sistemas en específico. Estas técnicas permiten a estos modelos llegar a ser eficientes, robustos y flexibles de manera computacional.

Un modelo oculto de Markov constituye una técnica de modelización de datos secuenciales aplicada inicialmente en el campo del reconocimiento automático del habla (Rabiner,1989), donde actualmente es una herramienta casi imprescindible. Este modelo puede ser pensado como una máquina de estados finito donde las transiciones entre los estados dependen de la ocurrencia de algún símbolo.

En general, son muy útiles para el reconocimiento de patrones y en particular son substancialmente efectivos para modelar el comportamiento del hombre o procesos en los que existe la intervención del hombre.

Formalmente, una cadena de Markov oculta es un proceso binario (doblemente estocástico) $\{X_n,\ Y_n\},\ \forall n \in \{0,1,2,...\}$, donde $\{X_n\}$ es una cadena de Markov homogénea con espacio de estados finito y $\{Y_n\}$ es un proceso estocástico cuya distribución condicional solo depende de $\{X_n\}$.

Se entiende entonces que los modelos ocultos de Markov son modelos estadísticos que juntan de acuerdo a leyes probabilísticas una colección de variables aleatorias $\{Y_1, Y_2, ..., Y_T, , X_1, X_2, ..., X_T\}$, con T finito. Las variables Y_t pueden ser observaciones continuas o discretas y las variables X_t son variables ocultas y discretas. Los HMM pueden particularizarse de acuerdo a las leyes probabilísticas que gobiernan a las variables aleatorias que intervienen en el modelo. En este tipo de modelos probabilísticos se asumen dos condiciones de independencia que son:

la t-ésima variable aleatoria: dada la (t–1)-ésima variable aleatoria del proceso oculto, es independiente de variables previas, es decir que X_t depende solamente de la X_{t-1} con lo que:

$$P(X_t \mid X_{t-1}, Y_{t-1}, ..., X_1, Y_1) = P(X_t \mid X_{t-1}) \quad (9)$$

la t-ésima observación: dada la t-ésima variable oculta, es independiente de cualquier otra variable, es decir que Y_1 depende solo de X_1, con lo que

$$P(Y_t \mid X_T, Y_T, ..., X_{t+1}, Y_{t+1}, X_t, Y_t, , X_{t-1}, Y_{t-1}, ..., X_1, Y_1) = P(Y_t \mid X_t) \quad (10)$$

La Fig. 3 es un modelo probabilístico gráfico que muestra un conjunto de variables aleatorias y sus condiciones de dependencia. Esta gráfica es llamada Red Bayesiana dinámica simple y refleja las dos condiciones de independencia impuestas a las variables aleatorias intervinientes en los HMM. Los círculos en azul representan a las variables aleatorias del proceso oculto y los rectángulos en rojo a las del proceso visible. Estas variables aleatorias pueden tomar ciertos valores en cada instante t, además cada símbolo X_t representa una variable aleatoria del proceso markoviano discreto oculto en el instante t, mientras que los símbolos X_t constituyen las variables aleatorias del proceso observable en cada instante t. Las flechas indican las formas de dependencias condicionales entre las variables aleatorias.

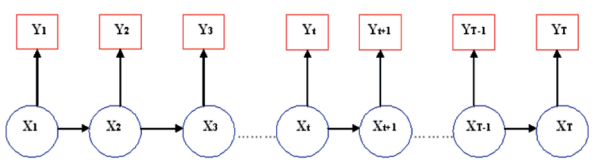

Figura 3: Red Bayesiana dinámica simple que refleja las condiciones de independencia en los HMM.

Elementos de los HMM discreto

Si se tiene que las observaciones Y_t son discretas con N posibles valores, entonces estamos ante una cadena de Markov oculta discreta. En los HMM discretos intervienen los siguientes elementos:

- $S = \{1, 2, ..., N\}$ es el conjunto finito de estados del proceso oculto, en donde N es la cantidad total de estados en el modelo.
- V es el conjunto de valores o símbolos diferentes que se pueden observar en cada uno de los estados, puede considerarse también un alfabeto finito. Cada uno de los símbolos que un estado puede emitir se denota como $\{v_1, v_2, ..., v_M\}$, donde M M es el número de símbolos del alfabeto y cada v_k se refiere a un símbolo diferente.
- $\boldsymbol{\Pi} = (\pi_1, \pi_2, ..., \pi_N)$ es la distribución de probabilidad inicial de los estados del proceso oculto, con $\pi_i = P(X_1 = i), \forall i \in \{1, 2, ..., N\}$ y $\sum_{I=1}^{N} \pi_i = 1$. Cada π_i representa la proba-

bilidad de que el proceso esté en el estado i en el instante en que se inicia.

- $\mathbf{A} = \{p_{ij}\}$ constituye la matriz de probabilidades de las transiciones entre los estados del proceso estocástico oculto, es decir: $p_{ij} = P(X_{n+1} = j \mid X_n = i)$, $\forall \{i, j\} \in \{1, 2, ..., N\}$ y $\forall n \in \{1, 2, ...\}$. Por lo cual, cada p_{ij} constituye la probabilidad de que el sistema se encuentre en el estado j en el tiempo $n+1$ dado que se encuentra en el estado i en el tiempo n.

- $\mathbf{B} = \{b_i(k)\}$ constituye la matriz de probabilidad de emisión de un estado dado. Constituye la distribución de probabilidad de los estados en el proceso observable, en donde cada $b_i(k) = P(Y_n = v_k \mid X_n = i), \forall i \in \{1, 2, ..., N\}$, $\forall n \in \{1, 2, \ldots\}$ y $\forall k \in \{1, 2, ..., M\}$ con $\sum_{I=1}^{N} b_i(k) = 1$. Cada $b_i(k)$ constituye la probabilidad de que el sistema, estando en el instante n en el estado i, genere la observación v_k.

Cualquier secuencia de observaciones generadas dentro del proceso se simboliza por $Y = (Y_1, Y_2, ..., Y_k)$, donde cada $Y_i \in V$, $\forall i \in \{1, 2, ..., k\}$, representa una observación y k es la cantidad de observaciones en la secuencia.

La Fig. 4 muestra el modelo gráfico de un HMM con cuatro estados ocultos y dos tipos de observaciones. Las flechas continuas indican las transiciones entre los estados del proceso markoviano oculto ($p_{ij} > 0$), y las flechas punteadas constituyen las probabilidades de emisión de observación en cada estado oculto ($b_i(k) > 0$).

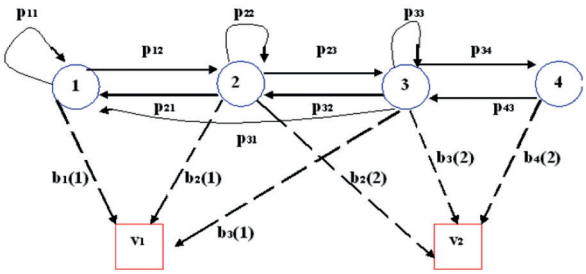

Figura 4: Grafica de un HMM con cuatro estados ocultos y dos tipos de observaciones.

Experimentos mendelianos

En 1866, se publicaron los famosos experimentos de Mdel en la hibridación de las plantas, el cual es considerado a menudo el trabajo de rompimiento del hielo de la genética moderna. Mendel no tenía ningún conocimiento anterior de la naturaleza dual de los genes, pero a través de una serie de experimentos en el jardín de su convento, pudo detectar la presencia de gen oculto y nombrarlo “Elemente”.

El mundo biológico es bastante complejo y a menudo demasiado complicado para el modelado matemático. En comparación con la física, solo recientemente han atacado problemas biológicos mediante el uso de métodos matemáticos enfocados a lo computacional. La más nueva y más desafiante interacción entre la biología y la matemática viene de la biología molecular moderna y de la bioinformática.

La genética ha tenido una explosión en número de descubrimientos y se ha desarrollado a la par del avance computacional. Experimentos como los de Gregor Mendel (1822-1884), titulado “los experimentos en los híbridos de las plantas (1866)”, pueden ser revisitados y las leyes de la herencia redescubiertas usando las técnicas y los datos de su publicación original. La genética mendeliana se refiere a la transmisión de los rasgos biológicos discretos de una generación a otra y los modos de expresión de los genes.

Los inicios, las cuentas y los experimentos realizados por Mendel

Mendel fue miembro del monasterio de Santo Tomás en Brünn, República Checa, a la edad de 21 años y estudió teología entre los años 1844 y 1848. Fue profesor de ciencias, física y matemática. Estudió en la Universidad de Viena entre 1851 y 1853. Su maestro más influyente fue Andreas Von Ettingshausen. Escribió un libro en combinatoria ("Die combinatorische Análisis", Viena, 1826), y había recogido claramente algunos de los métodos utilizados allí, más tarde en su serie de experimentos con las plantas. Resultó ser un buen experimentador en el jardín del monasterio, a pesar de que sólo tenía acceso a equipos bastantes simples. Hizo sus clásicos experimentos en la cría de plantas entre los años 1856 y 1863, donde cultivó y experimentó con al menos 28.000 plantas.

Los guisantes son la cuarta parte más importante en el cultivo de legumbres del mundo. Mendel eligió un guisante común del jardín que eran arvejas de la especie Pisum sativum para sus primeros experimentos en la hibridación por motivos prácticos (eran baratos, fáciles de obtener y tienen un tiempo de generación relativamente corto) y por otras dos razones que fueron cruciales para el éxito de sus experimentos, ya que las semillas de las arvejas presentaban variedades de formas y colores fáciles de identificar y analizar. Además, estas plantas, en forma silvestre se autofecundan y pueden ser manipuladas por el experimentador de modo tal que se puede dirigir la reproducción entre dos individuos (fecundación cruzada).

Las arvejas son plantas que se autofecundan de manera natural y exhiben caracteres que se presentan de manera simple y en distintas formas. Un carácter es un cierto rasgo o característica específica de un organismo y en general posee diferentes variantes, formas en que puede presentarse. Para lograr arvejas con ciertos caracteres, Mendel las cultivó durante muchas generaciones obteniendo líneas puras, es decir, subpoblaciones que producían descendencia homogénea para cada carácter elegido. Obtuvo siete parejas de líneas puras para siete caracteres, diferenciando cada pareja sólo respecto de un carácter. Los siete caracteres, junto con sus respectivas variantes, estudiados por Mendel en las arvejas se presentan en la Tabla 1.

Tabla 1: Caracteres versus variantes en las arvejas.

Carácter	Variante	
Color de la flor	Rojo	Blanco
Color de la semilla	Amarillo	Verde
Textura de la semilla	Lisa	Rugosa
Forma de la vaina	Hinchada	Hendida
Color de la vaina	Verde	Amarillo
Longitud del tallo	Largo	Corto
Tipo de floración	Terminal	Axial

Color de las flores de las plantas

El primer carácter estudiado por Mendel fue el color de las flores de las plantas. Como primer paso fecundó una planta de flores rojas con el polen de una planta de flores blancas, estos ejemplares de líneas puras constituyen la generación parental (P). Este cruzamiento originó la primera generación filial (F1), cuyos integrantes tenían todos flores rojas. Luego tomó plantas de flores blancas y las fecundó con polen de plantas de flores rojas y obtuvo los mismos resultados.

En su segundo experimento, tomó los ejemplares de la primera generación filial (F1) y las autofecundó, obteniendo como resultado 929 semillas en la segunda generación filial (F2), que luego sembró y cuando comenzaron a dar flores, se observó que algunas plantas tenían flores blancas, con lo cual había reaparecido la otra variante. Con estos resultados, primero contó la cantidad de individuos de cada variante hallada en F2 y calculó la proporción entre ellos, encontró que de un total de 929 plantas de F2, 705 tenían flores rojas y 224 tenían flores blancas. Las proporciones eran de aproximadamente 3:1, de cada 4 flores 3 eran de color rojo.

Mendel dedujo que F1 recibe la capacidad de producir flores rojas o blancas. Sin embargo, una de las variantes (flor blanca) no se expresa. Para denominar este fenómeno, utilizó el término do-

minante para designar a la variante que se expresa en F1 y el término recesivo, para la variante que queda enmascarada y reaparece en F2.

Color de las semillas

Siguiendo con sus experimentos, Mendel estudió el color de las semillas. La ventaja de usar este carácter era que no tenía que esperar a que cada uno de los individuos de cada generación floreciera. Para este carácter, dedujo que el color amarillo es dominante y que el color verde era recesivo. Además, pensó que, si a partir de las plantas con semillas amarillas de F1 surgieron, en F2, plantas con semillas verdes, era posible que, a su vez, las plantas amarillas de F2 también "escondieran" el carácter verde.

Para analizar su idea, Mendel autofecundó 519 plantas con semillas amarillas de F2 y vio que 166 de ellas sólo producían arvejas con semillas amarillas. Las llamó "amarillas puras". Por otra parte, 353 plantas producían arvejas con semillas amarillas y verdes, nuevamente en una proporción 3:1. De cada cuatro semillas tres eran de color amarillo, llamadas "amarillas impuras" (Fig. 5). Por lo tanto, de las plantas con semillas amarillas en F2, alrededor de los $1/3$ eran como el parental amarillo de línea pura y los $2/3$ eran como los amarillos de F1, es decir, producían semillas amarillas y verdes en una proporción 3:1.

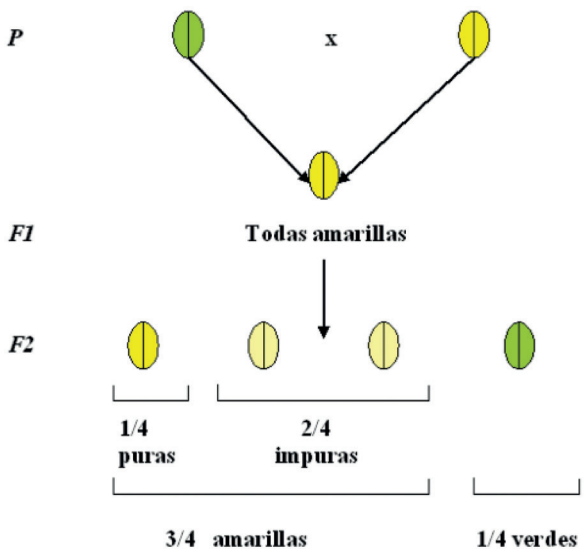

Figura 5: Proporciones observadas durante el proceso de fecundación de la semilla híbrida.

Por otra parte, las plantas con semillas verdes siempre daban arvejas con semillas verdes, por lo que eran puras siempre. Con estos resultados, Mendel se dio cuenta que era preciso distinguir la constitución genética de un individuo, su genotipo, de la expresión visible del mismo, su fenotipo.

Entonces, en F1 se tiene al fenotipo amarillo y en F2 a los fenotipos amarillo y verde. Las plantas con fenotipos amarillos en F2 no son todas iguales. Utilizando la notación A, para la variante dominante (amarilla) y a para la variante recesiva (verde), podemos representar a los genotipos de F2 de la siguiente manera: el genotipo AA corresponde a las plantas con semillas "amarillas puras", el genotipo aA corresponde a las plantas con semillas "amarillas impuras" y el genotipo aa corresponde a las plantas con semillas "verdes".

Según los resultados, en F2 se contemplan dos tipos de fenotipos, amarillo y verde, pero tres tipos de genotipos, denominadas homocigota dominante (AA), heterocigota (Aa) y homocigota recesiva (aa). En su documento, Mendel informó que realizó sus experimentos para cuatro a seis generaciones, sin dar más cifras, indicando solamente los resultados generales y fórmulas matemáticas.

Textura y color de las semillas

A fin de responder a la pregunta: ¿cómo serían las proporciones cuando se consideran dos caracteres simultáneamente? Mendel consideró en su siguiente experimento los caracteres "textura de la semilla", cuyas categorías son lisa (B) y rugosa (b), y "color de las semillas" con categorías amarillo (A) y verde (a), utilisando dos líneas puras: plantas con semillas lisas y verdes cuyos genotipos son de la forma $aaBB$ y plantas con semillas rugosas y amarillas con genotipos $AAbb$.

En el cruzamiento de las dos líneas puras ($AAbb \times aaBB$), tal como lo muestra la Fig. 6, se observó que en F1 todas las semillas eran lisas y amarillas con genotipos $AaBb$, es decir que es

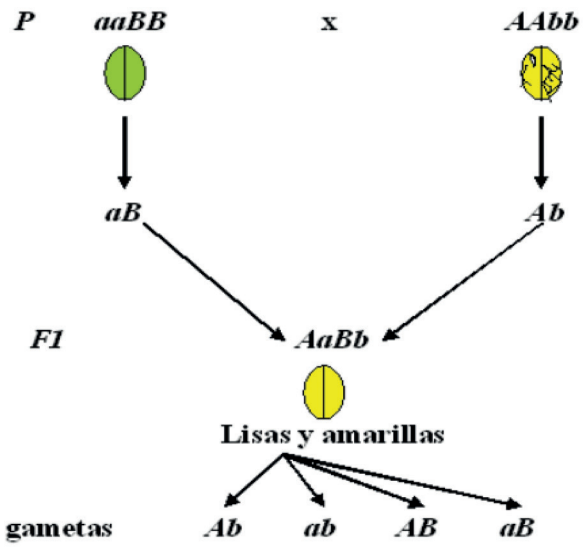

Figura 6: Resultados obtenidos en la primera generación filial (F1) al tomar los caracteres textura y color de las semillas simultáneamente.

homogénea y su fenotipo es el correspondiente a la forma dominante de cada carácter, en este caso liso y amarillo. En F2, Mendel obtuvo las cuatro combinaciones fenotípicas posibles (liso y amarillo, liso y verde, rugoso y amarillo, y rugoso y verde) en proporción 9:3:3:1, tal como lo muestra la Tabla 2.

Analizando primeramente el carácter textura de la semilla, que está compuesta de 423 semillas lisas y 133 semillas rugosas, se puede aprecia que guardan entre sí una relación 3:1. De la misma manera analizando el carácter color de las semillas se observa que 416 semillas son amarillas y 140 verdes, también guardan entre sí una relación 3:1.

Entonces, al considerarse los dos caracteres simultáneamente, los diferentes fenotipos resultantes deberían guardar entre sí una combinación de las proporciones de cada carácter por separado. Al considerar el carácter textura de la semilla, se tiene que: lisa tiene una probabilidad 3/4 y rugosa 1/4. De la misma manera, en el carácter color de las semillas, amarillo tiene una probabilidad de 3/4 y verde 1/4. Como los genes para distintos caracteres se heredan de forma independiente, se tiene:

$$P(lisa\ y\ verde) = P(liso)\,P(verde) = \frac{3}{4}.\frac{1}{4} = \frac{3}{16}$$

$$P(rugosa\ y\ amarillo) = P(rugosa)\,P(amarillo) = \frac{1}{4}.\frac{3}{4} = \frac{3}{16}$$

$$P(rugosa\ y\ verde) = P(rugosa)\,P(verde) = \frac{3}{4}.\frac{3}{4} = \frac{9}{16}$$

Extendiendo este razonamiento a n caracteres calculando las probabilidades combinadas de cada caso (Tabla 3).

Cantidades	Proporciones
315 semillas lisas y amarillas	9
108 semillas lisas y verdes	3
101 semillas rugosas y amarillas	3
32 semillas rugosas y verdes	1
556 semillas en total	16

Tabla 2: Resultados obtenidos en la segunda generación filial (F2) al tomar los caracteres textura y color de las semillas simultáneamente.

Tabla 3: Genotipos de la F2 de un cruzamiento $AaBb\,x\,AaBb$.

Gametas	*AB*	*Ab*	*aB*	*ab*
AB	*AABB*	*AABb*	*AaBB*	*AaBb*
Ab	*AABb*	*AAbb*	*AaBb*	*Aabb*
aB	*AaBB*	*AaBb*	*aaBB*	*aaBb*
ab	*AaBb*	*Aabb*	*aaBb*	*aabb*

Leyes de Mendel

Las leyes de Mendel son un conjunto de reglas básicas sobre la transmisión por herencia de las características de los organismos de los padres a sus hijos.

Primera ley de Mendel: Ley de la uniformidad de los híbridos de la primera generación filial (F1): Si se cruza una línea pura de arvejas cuya semilla es de color amarilla con otra cuya semilla es de color verde, los individuos de la primera generación filial (F1) son todos uniformes. En este caso se parece a uno de los progenitores, el de la semilla con color amarilla.

Segunda ley de Mendel: Ley de la segregación: Los dos factores (genes) para cada carácter no se mezclan ni se fusionan de ninguna manera, sino que se segregan (separan) en el momento de la formación de los gametos.

Tercera ley de Mendel: Ley de la transmisión independiente: Los genes para distintos caracteres se heredan de forma independiente

Resultados y discusión

Experimentos mendelianos como cadenas de Markov

Los experimentos genéticos mendelianos sobre la base de la libre fertilización, pueden representarse mediante simples modelos de Markov.

Diploide con un solo par de genes

Mendel en sus primeros experimentos estudió el color de las plantas y luego el color de las semillas de las plantas. En ambos casos tomó un solo carácter, las cuales tienen dos tipos de variantes (genes).

Considerando uno de los dos primeros experimentos mendelianos, por ejemplo, el experimento realizado sobre el carácter "color de las semillas", en la cual las variantes (genes) son "amarillo" al que se le llamó gen dominante (A) y "verde" al que denominó gen recesivo (a). Recordando que en cada generación se pudo notar que era preciso distinguir la constitución genética de un individuo de la expresión visible del mismo (aspecto), es decir su genotipo (homocigota dominante (AA), heterocigota (aA) ó homocigota recesivo (aa)) de su fenotipo (amarillo (A) ó verde (a)). Entonces, el genotipo de las semillas puede ser de un nivel sobre el conjunto de estados posibles

$$S=\left\{AA\left(estado\,1\right),aA\left(estado\,2\right),aa\left(estado\,3\right)\right\}$$

en el tiempo t, con probabilidades de $v_t\left(1\right)$, $v_t\left(2\right)$ y $v_t\left(3\right)$, donde $v_t\left(1\right)+v_t\left(2\right)+v_t\left(3\right)=1$.

Si X_t es la variable aleatoria que muestra el estado en la que se encuentra el genotipo del individuo en el tiempo t, la distribución de probabilidad de los estados del proceso cuando evolucione en el tiempo será $V_t=\left(v_t\left(1\right),v_t\left(2\right),v_t\left(3\right)\right)$, donde $v_t\left(i\right)=P\left(X_t=i\right)$, $\forall t\in\left\{1,2,3,...\right\}$ y $\boldsymbol{\Pi}^1=\left(0,1,0\right)$. Como al principio de su experimento Mendel tomó semillas de colores amarillas y las fecundó con semillas de colores verdes, comenzó este proceso tomando una población fundada en las primeras plantas de heterocigotas y con ello, la distribución de probabilidad inicial viene dada por $\boldsymbol{\Pi}^1=\left(0,1,0\right)$.

Por otra parte, Mendel observó que si tomaba una población de plantas con semillas de colores amarillas puras (homocigotas dominantes) y las autofecundaba obtenía siempre plantas con semillas de colores amarillas. De manera análoga, si tomaba una población de plantas con semillas verdes puras obtenía siempre plantas con semillas de colores verdes. Ahora bien, si tomaba una población de plantas heterocigotas la relación de los genotipos era de 1:2:1, con la cual las proporciones genotípicas eran de $1/4$, $1/2$ y $1/4$ (Fig. 7).

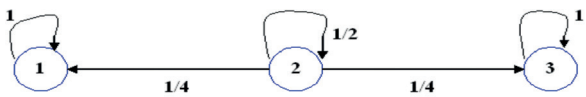

Figura 7: Dígrafo de transición (Diploide con un solo par de genes).

Considerando que la segregación cromosómica (distribución de los cromosomas) es al azar y que el tiempo de transición es homogéneo, se puede definir a la matriz de transición de probabilidades de la cadena de Markov para la autofertilización como:

$$P_1=\begin{pmatrix}1 & 0 & 0\\ 1/4 & 1/2 & 1/4\\ 0 & 0 & 1\end{pmatrix}$$

Las p_{ij}, $\forall i,j\in\left\{1,2,3\right\}$, de la matriz de transición de probabilidades de un paso corresponden a la probabilidad de que cada célula realice la transición de un cierto estado j a un cierto estado

j. Es decir, la probabilidad del paso de un cierto genotipo a otro genotipo de un tiempo t a otro tiempo $t+1$. En este proceso de Markov, el estado 2 es transitorio y los estados 1 y 3 son absorbentes. Tomando en cuenta la ecuación (5), la distribución de probabilidad de la variable aleatoria X_{t+1} (de los estados) en el tiempo $t+1$ es:

$$V_{t+1} = \Pi^1.P^{(t)} = \left(\frac{1}{2}-\left(\frac{1}{2}\right)^{t+1}, \left(\frac{1}{2}\right)^{t}, \frac{1}{2}-\left(\frac{1}{2}\right)^{t+1}\right) \quad (11)$$

Calculando el límite de V_{t+1} cuando t tiende al infinito, se obtiene:

$$V_{\infty} = \lim_{k\to\infty} V_k = \left(\frac{1}{2}, 0, \frac{1}{2}\right) \quad (12)$$

Resultado que muestra que el proceso será absorbido rápidamente por alguno de los dos estados absorbentes intervinientes en la cadena. Entonces, con la autofertilización, una población se divide en una serie de líneas que rápidamente se vuelven muy homocigótica y que asintóticamente produce la autofecundación de dos genotipos puros, lo que hace que todos los descendientes sean del mismo tipo.

Debido a que en la reproducción incluye tanto la reproducción celular (de una célula madre se obtienen dos células hijas) como la procreación (a partir de dos células de los individuos progenitores, se produce un individuo hijo), se toma un modelo con la fertilización al azar en una población infinita diploide sin superposición entre las generaciones en lugar de la reproducción de individuos del mismo linaje (endogamia), es decir que al mezclarse dos individuos de sexos opuestos (progenitores) de la población en cada paso dan origen a un nuevo ser vivo (hijo) y este nuevo ser vivo hereda un gen de manera aleatoria de cada uno de sus progenitores. El genotipo del hijo es nuevamente de un nivel sobre el espacio de estados $S = \{AA(estado\,1), aA(estado\,2), aa(estado\,3)\}$.

Al concentrarse en uno de los progenitores, por decir el padre, la distribución de probabilidad inicial está dada por $\boldsymbol{\Pi}^1 = (\pi_1, \pi_2, \pi_3)$, donde $\pi_i = P(X_0 = i) = q_i$ indica la probabilidad de que el individuo padre escoja una madre que tenga como par de genes los correspondientes al estado i, con $i \in \{1,2,3\}$. Por otra parte, la probabilidad de que la madre tenga algún gen dominante o recesivo en el estado i está dada por q_{Di} o q_{Ri}. En este modelo, los elementos intervinientes en los cálculos para la obtención de las probabilidades de transición de un paso tienen las siguientes interpretaciones:

a) Si el padre tiene ambos genes dominantes y escoge a una madre para cruzarse, la probabilidad de que el hijo tenga:

- ambos genes dominantes (AA que se puede simbolizar por D) es:

$$P(D_{t+1}|D_t) = p_{11} = q_1 q_{D1}\, p_D + q_2 q_{D2}\, p_D + q_3 q_{D3}\, p_D$$

$$p_{11} = \pi_1 \cdot 1 \cdot 1 + \pi_2 \cdot \frac{1}{2} \cdot 1 + \pi_3 \cdot 0 \cdot 1 = \pi_1 + \frac{\pi_2}{2}$$

- un gen dominante y otro recesivo (aA que se puede simbolizar por H) es:

$$P(H_{t+1}|D_t) = p_{12} = q_1 q_{R1}\, p_D + q_2 q_{R2}\, p_D + q_3 q_{R3}\, p_D$$

$$p_{12} = \pi_1 \cdot 0 \cdot 1 + \pi_2 \cdot \frac{1}{2} \cdot 1 + \pi_3 \cdot 1 \cdot 1 = \frac{\pi_2}{2} + \pi_3$$

- ambos genes recesivos (aa que se puede simbolizar por R) es:

$$P(R_{t+1} \mid D_t) = p_{13} = 0$$

Como el individuo hereda un gen dominante del padre con probabilidad 1, entonces es imposible que llegue a tener ambos genes recesivos.

b) Si el padre tiene un gen dominante y el otro recesivo, y escoge a una madre para cruzarse, la probabilidad de que el hijo tenga:

- ambos genes dominantes (D) es:

$$P(D_{t+1}|H_t) = p_{21} = q_1 q_{D1}\, p_D + q_2 q_{D2}\, p_D + q_3 q_{D3}\, p_D$$

$$p_{21} = \pi_1 \cdot 1 \cdot \frac{1}{2} + \pi_2 \cdot \frac{1}{2} \cdot \frac{1}{2} + \pi_3 \cdot 0 \cdot \frac{1}{2} = \frac{\pi_1}{2} + \frac{\pi_2}{4}$$

- un gen dominante y otro recesivo (H) es:

$$P(H_{t+1}|D_t) = p_{22} = q_1 q_{D1}\, p_R + q_2 \left(q_{D2} p_R + q_{R2} p_D\right) p_D + q_3 q_{R3}\, p_D$$

$$p_{22} = \pi_1 \cdot 1 \cdot \frac{1}{2} + \pi_2 \cdot \left(\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}\right) + \pi_3 \cdot 1 \cdot \frac{1}{2} = \frac{1}{2}$$

- ambos genes recesivos (R)

$$P(R_{t+1}|H_t) = p_{23} = q_1 q_{R1}\, p_R + q_2 q_{R2}\, p_R + q_3 q_{R3}\, p_R$$

$$p_{23} = \pi_1 \cdot 0 \cdot \frac{1}{2} + \pi_2 \cdot \frac{1}{2} \cdot \frac{1}{2} + \pi_3 \cdot 1 \cdot \frac{1}{2} = \frac{\pi_2}{4} + \frac{\pi_3}{2}$$

c) Si el padre tiene ambos genes recesivos, y escoge a una madre para cruzarse, la probabilidad de que el hijo tenga

- ambos genes dominantes (D) es:

$$P(D_{t+1}|R_t) = p_{31} = 0$$

Puesto que el individuo hereda un gen recesivo del padre con probabilidad 1, entonces es imposible que llega a tener ambos genes dominantes.

- un gen dominante y otro recesivo (H) es:

$$P(H_{t+1}|R_t) = p_{32} = q_1\, q_{D1}\, p_R + q_2\, q_{D2}\, p_R + q_3 q_{D3}\, p_R$$

$$p_{22} = \pi_1 \cdot 1 \cdot 1 + \pi_2 \cdot \frac{1}{2} \cdot 1 + \pi_3 \cdot 0 \cdot 1 = \pi_1 + \frac{\pi_2}{2}$$

- ambos genes recesivos (R)

$$P(R_{t+1}|R_t) = p_{33} = q_1 q_{R1}\, p_R + q_2 q_{R2}\, p_R + q_3 q_{R3}\, p_R$$

$$p_{33} = \pi_1 \cdot 0 \cdot 1 + \pi_2 \cdot \frac{1}{2} \cdot 1 + \pi_3 \cdot 1 \cdot 1 = \frac{\pi_2}{2} + \pi_3$$

Con los resultados anteriores, la matriz de transición de probabilidades es:

$$P = \begin{pmatrix} \pi_1 + \frac{\pi_2}{2} & \frac{\pi_2}{2} + \pi_3 & 0 \\ \frac{\pi_1}{2} + \frac{\pi_2}{4} & \frac{1}{2} & \frac{\pi_2}{4} + \frac{\pi_3}{2} \\ 0 & \pi_1 + \frac{\pi_2}{2} & \frac{\pi_2}{2} + \pi_3 \end{pmatrix}$$

y la distribución de probabilidad de la variable aleatoria X_t (de los estados) en el tiempo t, $\forall t \in \{1,2,3,\ldots\}$, es:

$$V_t = \Pi^1 . P^{(t-1)} = \left(\left(\pi_1 + \frac{\pi_2}{2}\right)^2, 2\left(\pi_1 + \frac{\pi_2}{2}\right)\left(\frac{\pi_2}{2} + \pi_3\right), \left(\frac{\pi_2}{2} + \pi_3\right)^2\right) \quad (13)$$

Esto nos dice que esta cadena de Markov alcanza ya su condición estable en el tiempo $t = 1$, es decir, que con la primera generación de descendientes ya se alcanza el equilibrio entre los genotipos.

Diploide con dos pares de genes independientes

En sus anotaciones Mendel dejó constancia de que también estudió en detalle el diploide con dos y tres pares de genes. En estos estudios dedujo que las elecciones de los gametos de cada carácter en cada generación eran independientes. El caso del diploide con dos pares de genes consiste en tomar los caracteres textura de las semillas cuyas categorías son lisa (A) y rugosa (a), y color de las semillas con categorías amarilla (B) y verde (b), simultáneamente. En el genotipo, las semillas pueden estar según la Tabla 3 en un nivel sobre el siguiente conjunto de estados:

$$S = \{AABB, aABB, aaBB, AAbB, aAbB, aabB, AAbb, aAbb, aabb\}$$

Si X_t es la variable aleatoria que representa el estado sobre la cual se encuentra el genotipo en el instante t, se tendrá que la distribución de probabilidad de los estados de este proceso de autofecundación en el tiempo t será $V_t = \left(v_t(1), v_t(2), \ldots, v_t(9)\right)$, con $i \in \{1,2,...,9\}$, $t \in \{1,2,...\}$ y $v_t(i) = P(X_t = i)$. En este caso, Mendel notó que si tomaba semillas con líneas puras y las autofecundaba, en la próxima generación todas las semillas eran de la misma línea pura tomada. Además, si consideraba semillas que no eran de líneas puras se daban las siguientes situaciones.

- Si las semillas consideradas tienen genotipos $aABB$ y se autofecundan se tienen tres tipos posibles de genotipos que son: $aaBB$, $aABB$ y $AABB$.

Tabla 4: Genotipos resultantes de un cruzamiento $aABB\,x\,aABB$.

Gametas	*aB*	*AB*
aB	*aaBB*	*aABB*
AB	*aABB*	*AABB*

Tomando en consideración la Tabla 4 se tiene:

$$P\left(AABB|aABB\right)=p_{21}=\frac{1}{4},P\left(aABB|aABB\right)=p_{22}$$

$$=\frac{1}{2},P(aaBB\mid aABB)=p_{23}=\frac{1}{4}$$

- Si las semillas consideradas tienen genotipos $AAbB$ y se autofecundan se tienen tres tipos posibles de genotipos que son: $AAbb, AAbB$ y $AABB$ (Tabla 5).

$$P\left(AABB|AAbB\right)=p_{41}=\frac{1}{4},P\left(AAbB|AAbB\right)=$$

$$=p_{44}=\frac{1}{2},P(AAbb\mid aABB)=p_{73}=\frac{1}{4}$$

Tabla 5: Genotipos resultantes de un cruzamiento $AAbB\,x\,AAbB$.

Gametas	*Ab*	*AB*
Ab	*AAbb*	*AAbB*
AB	*AAbB*	*AABB*

- Si las semillas consideradas tienen genotipos $aAbB$ y se autofecundan se tienen nueve tipos posibles de genotipos que son: $aabb, aabB, aAbb, aAbB, aaBB, aABB, AAbb, AAbB$ y $AABB$ (Tabla 6).

$$P(AABB\mid aAbB)=p_{51}=\frac{1}{16},P(aABB\mid aAbB)=p_{52}$$

$$=p_{52}=\frac{1}{8},P(aaBB\mid aAbB)=p_{53}=\frac{1}{16}$$

$$P(AAbB\mid aAbB)=p_{54}=\frac{1}{8},P(aAbB\mid aAbB)=p_{55}$$

$$p_{55}=\frac{1}{4},P(aabB\mid aAbB)=p_{56}=\frac{1}{8}$$

$$P(AAbb\mid aAbB)=p_{57}=\frac{1}{16},P(aAbb\mid aAbB)=$$

$$p_{58}=\frac{1}{8},P(aabb\mid aAbB)=p_{59}=\frac{1}{16}$$

Tabla 6: Genotipos resultantes de un cruzamiento $aAbB\,x\,aAbB$.

Gametas	*ab*	*aB*	*Ab*	*AB*
ab	*aabb*	*aabB*	*aAbb*	*aAbB*
aB	*aabB*	*aaBB*	*aAbB*	*aABB*
Ab	*aAbb*	*aAbB*	*AAbb*	*AAbB*
AB	*aAbB*	*aABB*	*AAbB*	*AABB*

- si las semillas consideradas tienen genotipos $aabB$ y se autofecundan tendremos tres tipos posibles de genotipos que son: $aabb, aabB$ y $aaBB$ (Tabla 7).

$$P(aaBB\mid aabB)=p_{63}=\frac{1}{4},P(aabB\mid aabB)=$$

$$p_{66}=\frac{1}{2},P(aabb\mid aabB)=p_{69}=\frac{1}{4}$$

Tabla 7: Genotipos resultantes de un cruzamiento $aabB\,x\,aabB$.

Gametas	*ab*	*aB*
ab	*aabb*	*aabB*
aB	*aabB*	*aaBB*

- si las semillas consideradas tienen genotipos $aAbb$ y se autofecundan tendremos tres tipos posibles de genotipos que son: $aabb, aAbb$ y $AAbb$ (Tabla 8).

Tabla 8: Genotipos resultantes de un cruzamiento $aabB\,x\,aAbb$.

Gametas	*ab*	*Ab*
ab	*aabb*	*aAbb*
Ab	*aAbb*	*AAb*b

$$P(AAbb \mid aAbb) = p_{87} = \frac{1}{4}, P(aAbb \mid aAbb) =$$

$$p_{88} = \frac{1}{2}, P(aabb \mid aABB) = p_{89} = \frac{1}{4}$$

Como al inicio del estudio del diploide con dos pares de genes, se cruzó las líneas puras liso y verde cuyos genotipos eran $AAbb$ con líneas puras rugoso y amarillo con genotipos $aaBB$, obteniendo en F1 semillas lisas y amarillas con genotipos $aAbB$, se tiene que la distribución inicial del proceso está dada por $\boldsymbol{\Pi}^2 = (0,0,0,0,1,0,0,0,0)$.

Al igual que en el caso del diploide con un solo par de genes, considerando que la distribución de los cromosomas es al azar, que las elecciones de los genes de carácter a carácter son independientes y que el tiempo de transición es homogéneo, se obtiene el dígrafo de transición mostrado en la Fig. 8 y la matriz de probabilidades de transición está dada por P_2. En esta cadena de Markov los estados 1, 3, 7 y 9 son absorbentes, y los demás estados son transitorios. Realizando un procedimiento análogo al caso del diploide con un par de genes, y teniendo en cuenta el caso límite $V_\infty = \left(\frac{1}{2}, 0, \frac{1}{2}\right)$ y que

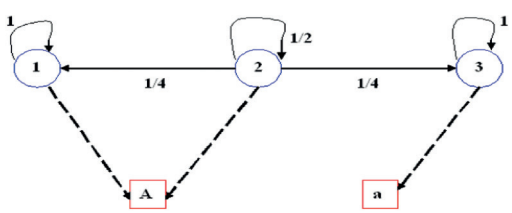

Figura 8: Modelo gráfico del Diploide con un solo par de genes.

$0 = (0,0,0)$, se obtiene:

$$V_t^2 = \left(\frac{1}{4} - \frac{2^t - 1}{4^t}, \frac{2^{t-1} - 1}{2^{2t-1}}, \frac{1}{4} - \frac{2^t - 1}{4^t}, \frac{2^{t-1} - 1}{2^{2t-1}}, \frac{1}{4^{t-1}}, \frac{2^{t-1} - 1}{2^{2t-1}}, \frac{1}{4} - \frac{2^t - 1}{4^t}, \frac{2^{t-1} - 1}{2^{2t-1}}, \frac{1}{4} - \frac{2^t - 1}{4^t}\right) \quad (14)$$

$$V_\infty^2 = \lim_{t\to\infty} V_t^2 = \left(\frac{1}{4}, 0, \frac{1}{4}, 0, 0, 0, \frac{1}{4}, 0, \frac{1}{4}\right) = \frac{1}{2}(V_\infty, 0, V_\infty) \quad (15)$$

Según el resultado obtenido, esta cadena markoviana será absorbida por alguno de los estados absorbentes intervinientes y, la población de semillas será rápidamente homogénea. Por lo tanto, la forma de dicha población corresponderá a alguna de las cuatro líneas puras tomadas.

Diploide con k pares de genes

Al analizar el diploide con k pares de genes todos independientes, se tiene que los genotipos pueden estar sobre un total de 3^k estados. La matriz de probabilidades de transición de este proceso markoviano es de orden $3^k \times 3^k$ y estará dada por:

$$P_k = \begin{pmatrix} P_{k-1} & O & O \\ \frac{1}{4}P_{k-1} & \frac{1}{2}P_{k-1} & \frac{1}{4}P_{k-1} \\ O & O & P_{k-1} \end{pmatrix}$$

siendo P_{k-1} la matriz de probabilidades de transición de orden $3^{k-1} \times 3^{k-1}$ que se obtiene en el caso del diploide con $k-1$ pares de genes y O es la matriz nula de orden $3^{k-1} \times 3^{k-1}$. La distribución inicial será $\boldsymbol{\Pi}^k = (0,\ldots,1,\ldots,0)$ en dirección del eje $\frac{3^{k+1}}{2}$, con $\Pi^k \in \mathrm{R}^{3k}$.

Al realizar procedimientos análogos a los anteriores para cada $k \in \{3,4,5,6,7\}$ se obtiene que

la distribución límite está dada por:

$$V_{\infty}^{k}=\frac{1}{2}\left(\ V_{\infty}^{k-1},0,\ V_{\infty}^{k-1}\right)\ (16)$$

donde V_{∞}^{k-1} es la distribución límite para el caso del diploide con $k-1$ pares de genes independientes uno del otro.

Experimentos mendelianos como cadenas de Markov

Mendel, mediante experimentos, dedujo que era preciso distinguir la constitución genética (genotipo) de un individuo de su característica visible (fenotipo). Este hecho muestra que en los experimentos mendelianos interviene un proceso visible consistente en los fenotipos observados en los individuos en cada generación y un proceso oculto basado en los genotipos de los individuos en cada generación.

Modelo HMM para el diploide con un solo par de genes

En la sección anterior, se ha visto que el proceso oculto es una cadena de Markov con espacio de estados $S=\left\{AA\left(estado1\right),aA\left(estado2\right),aa\left(estado3\right)\right\}$, matriz de transición de probabilidades de un paso

$$\mathbf{A}_1=\begin{pmatrix}1 & 0 & 0\\ 1/4 & 1/2 & 1/4\\ 0 & 0 & 1\end{pmatrix}$$

y distribución de probabilidad inicial $\boldsymbol{\Pi}_1=\left(0,1,0\right)$.

El proceso visible consiste en un proceso estocástico con dos estados, los fenotipos "color amarillo (A)" y color "verde (a)". Para determinar la probabilidad de que se observe el fenotipo k dado que el individuo se tiene genotipo del tipo i, $b_i\left(k\right)$, $\forall i\in\left\{1,2,3\right\}$ y $\forall k\in\left\{a,A\right\}$ se sigue el siguiente procedimiento:

- Si el individuo se encuentra en el estado 1, su genotipo es homocigota dominante (AA), las plantas con semillas de color amarilla siempre dan semillas de color amarilla y por lo tanto, $b_1\left(A\right)=1$ y $b_1\left(a\right)=0$.
- Si el individuo se encuentra en el estado 2, su genotipo es heterocigota (aA), entonces, las semillas se verán de color amarilla debido a que este color amarillo (A) es dominante y por lo tanto, $b_1\left(A\right)=1$ y $b_1\left(a\right)=0$.
- Si el individuo se encuentra en el estado 3, su genotipo es homocigota recesivo (aa), las plantas con semillas de color verde siempre dan semillas de color verde y por lo tanto, $b_1\left(A\right)=0$ y $b_1\left(a\right)=1$.

Del procedimiento anterior, la matriz de probabilidades de emisión de las observaciones está dada por:

$$\mathbf{B}_1=\begin{pmatrix}1 & 0\\ 1 & 0\\ 0 & 1\end{pmatrix}$$

y el modelo oculto de Markov obtenido queda totalmente especificada por el conjunto de parámetros $\lambda=\left(\mathbf{A}_1,\mathbf{B}_1,\boldsymbol{\Pi}_1\right)$.

Modelo HMM para el diploide con dos pares de genes independientes

Mendel al considerar los caracteres textura y color de las semillas obtuvo nueve tipos diferentes de genotipos y cuatro de fenotipos, por lo que el proceso oculto, consistente en la autofecundación de las semillas en su genotipo, es una cadena de Markov con espacio de estados S compuesto por los estados $AABB\left(1\right)$, $aABB\left(2\right)$, $aaBB\left(3\right)$, $AAbB\left(4\right)$, $aAbB\left(5\right)$, $aabB\left(6\right)$, $AAbb\left(7\right)$, $aAbb\left(8\right)$ y $aabb\left(9\right)$ (Fig. 9), distribución inicial $\boldsymbol{\Pi}_2=\left(0,0,0,0,1,0,0,0,0\right)$

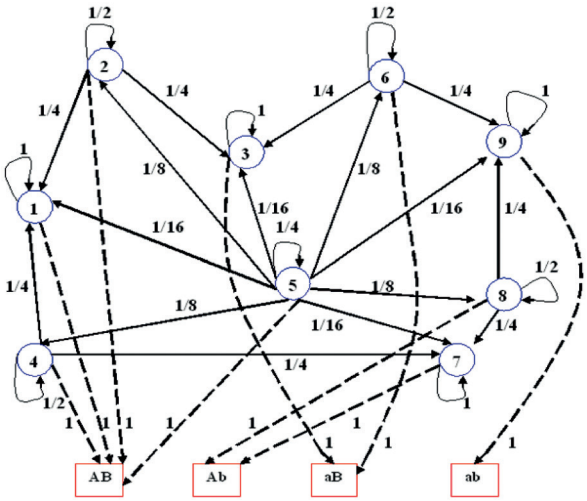

Figura 9: Modelo gráfico del Diploide con dos pares de genes independientes.

y matriz de probabilidades de transición

$$\mathbf{A}_2 = \begin{pmatrix} \mathbf{A}_1 & 0 & 0 \\ \frac{1}{4}\mathbf{A}_1 & \frac{1}{2}\mathbf{A}_1 & \frac{1}{4}\mathbf{A}_1 \\ 0 & 0 & \mathbf{A}_1 \end{pmatrix}$$

En el proceso observable tenemos cuatro categorías que constituyen los tipos posibles de fenotipos que se observan en cada generación filial. Estas categorías son: liso y amarillo (AB), liso y verde (Ab), rugoso y amarillo (aB), rugoso y verde (ab). Considerando los genes dominantes liso (A) y amarillo (B) en los caracteres textura y color de las semillas respectivamente, se procede a formar la matriz de probabilidades de emisión de observaciones y con ello, verificar que observaciones podrán emitir los estados del proceso oculto. En primer lugar, si el proceso oculto se encuentra en los estados $AABB, aABB, AAbB$ y $aAbB$ con probabilidad 1 se observan semillas lisas y amarillas (AB). Por otro lado, si se consideran los estados $AAbb$ y $aAbb$ se tendrán semillas lisas y verdes (Ab). En el caso del estado $aabb$ se tienen semillas rugosas y verdes (ab). Finalmente, estando en los estados $aaBB$ y $aabB$ se verán semillas rugosas y amarillas (aB). Con estos resultados, se tiene:

$$\mathbf{B}_2 = \begin{pmatrix} 1 & 00 & 0 \\ 1 & 00 & 0 \\ 0 & 01 & 0 \\ 1 & 00 & 0 \\ 1 & 00 & 0 \\ 0 & 01 & 0 \\ 0 & 10 & 0 \\ 0 & 10 & 0 \\ 0 & 00 & 1 \end{pmatrix}$$

El modelo oculto de Markov obtenido, bajo estas condiciones, queda totalmente especificada por el conjunto de parámetros $\lambda = (\mathbf{A}_2, \mathbf{B}_2, \boldsymbol{\Pi}_2)$.

Conclusiones

Según las revisiones de los trabajos de reseña analizados, Mendel no tenía ningún conocimiento anterior de la naturaleza dual de los genes, pero a través de una serie de experimentos en el jardín de su convento pudo detectar la presencia de gen oculto y nombrarlo “Elemente”. La genética mendeliana se refiere a la transmisión de los rasgos biológicos discretos de una generación a otra y los modos de expresión de los genes.

La estadística y la teoría de probabilidad, constituyen las piezas fundamentales en la formación del marco teórico de los modelos ocultos de Markov. Bajo este contexto, la matemática proporciona las herramientas necesarias para el planteamiento de soluciones a los distintos problemas que se presentan al momento de la formulación de un modelo oculto de Markov.

La matemática y las leyes de la uniformidad de los híbridos de la primera generación filial, de la segregación y de la transmisión independiente constituyen los principales pilares en la formulación de los distintos modelos de Markov aplicados a los experimentos mendelianos. En este sentido, mediante cadenas de Markov y cadenas ocultas de Markov, considerando una cierta cantidad de genes, se han

estudiado y modelado los experimentos genéticos mendelianos sobre la base de la libre fertilización obteniéndose muy buenos resultados.

En la formulación de la cadena de Markov realizada sobre un par de genes, se ha podido constatar que el proceso será absorbido rápidamente por alguno de los dos estados absorbentes intervinientes. Entonces, con la autofertilización, una población se divide en una serie de líneas que rápidamente se vuelven muy homocigóticas y asintóticamente, ésta produce la autofecundación de dos genotipos puros lo que hace que todos los descendientes sean del mismo tipo. En el caso del diploide con dos pares de genes independientes, el proceso será absorbida por alguno de los cuatro estados absorbentes intervinientes y la población de semillas será de la forma correspondiente a alguna de estas cuatro líneas puras.

Por su parte, en la formulación de los modelos ocultos de Markov realizada sobre los casos del diploide con un solo par de genes y el de dos pares de genes independientes se ha podido constar que quedan totalmente especificados por el conjunto de parámetros; matriz de probabilidades de transición $\mathbf{A}$, matriz de probabilidades de emisión de observaciones $\mathbf{B}$ y distribución inicial $\boldsymbol{\Pi}$.

Con todo lo expresado, se ha podido comprobar que, mediante el uso del álgebra, del cálculo, la teoría de probabilidad y las tres leyes de Mendel se puede aplicar de manera satisfactoria y eficiente la teoría de Markov, sobre la base de la libre fertilización, en el campo de la biología.

Agradecimientos

Al Altísimo, dueño eterno del reino, del poder y la gloria desde que el tiempo no tenía memoria. A este Divino Creador, cuya fuente de luz y de fuerza me impulsa a seguir adelante y me levanta cuando tropiezo por los caminos de la vida.

A mis padres y a mi amada familia por el apoyo de siempre. En especial a mi abuela Avelina por haberme guiado desde siempre por el largo sendero de la vida, a Francisca Ofelia, mis hijas Jessica Larissa y Lisset Fiorella, y mis hermanas Máxima y Nancy por el apoyo y confianza incondicional de siempre para el logro de mis metas y seguir mis ideales.

A la Dra. Ana Georgina Flesia por su orientación, paciencia y comprensión durante la realización de esta investigación.

Fuente de financiamiento

Fuente de financiamiento propia.

Literatura citada

Bilmes, J.A. (1997). Gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden Markov models. *International Computer Science Technical Report*, 21: i + 13 pp.

Breiman, L. (1992). *Probability*. Filadelphia: Society for Industrial and Applied Mathematics. *Classics in applied mathematics*, 7: 435 pp.

Forney, G. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3): 263–278.

Thorvaldsen, S. (2005). A tutorial on Markov models based on Mendel's classical experiments. *Journal of Bioinformatics and Computational Biology*, 3(6): 1441–1460.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *Journal of Molecular Biology*, 235(5): 1501– 1531.

Juang, B.H. & Rabiner, L.R. (1990). The segmental k means algorithm for estimating the papramеters of hidden Markov models. *IEEE Transactions in Accoustics, Speech and Sgnal Processing*, 38(9): 1639–1641.

Cappé, O., Moulines, E. & Ryden, T. (2007). *Inference in hidden Markov models*. New York: Springer Verlag. *Springer Series in Statistics*. xvii + 653 pp.

Rabiner, L.R. & Juang, B.H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1): 4–15.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected publications in speech recognition. *Proceedings of the IEEE*, 77(2): 257–286.