

**Diseño de herramientas matemático-computacionales para la
búsqueda de personas desaparecidas**

Por

Lic. Franco Leonel Marsico

Directora: Dra. Délida Inés Caridi

Co-director: Dr. Ariel Chernomoretz

Consejera: Dra. Leticia Betancort

Buenos Aires, Junio 2023

*A todas las personas que buscan,
Y en su búsqueda mantienen viva,
La Memoria de los desaparecidos.*

Diseño de herramientas matemático-computacionales para la búsqueda de personas desaparecidas

La búsqueda de personas desaparecidas es una tarea que requiere múltiples pasos, desde la recopilación de información básica sobre el caso, hasta la toma de decisiones respecto a las posibles hipótesis de identidad. El objetivo es comparar dos entidades: la de la persona desaparecida (identidad sin cuerpo) y la de la persona no identificada (cuerpo sin identidad). Durante el proceso, se recopilan diferentes líneas de evidencia, cada una con distintos grados de complejidad, que van desde la edad hasta las características genéticas de los individuos. El papel del científico forense es asignar probabilidades a las diferentes evidencias, considerando las proposiciones presentadas para explicarlas, siempre y cuando puedan ser formalizadas matemáticamente. Aunque varias líneas de evidencia pueden ser formalizadas, los datos genéticos se han convertido en la pieza clave para las identificaciones en los últimos años, dejando a otras evidencias relegadas. Cuando los datos genéticos inicialmente recolectados no son suficientes, ya sea por el mal estado de las muestras o la dificultad para genotipificar a los individuos, es necesario incorporar más evidencias y tomar decisiones con lo que se tiene disponible. En esta tesis, se abordan algunos de estos aspectos y se plantean tres problemas: formalización, optimización y priorización. El problema de formalización busca proponer modelos matemáticos basados en un enfoque bayesiano para evaluar estadísticamente el peso de la evidencia recolectada durante la investigación preliminar, como la edad y el sexo. En el problema de optimización, se formulan metodologías para la selección racional de casos que puedan ser potenciales identificaciones, incluso cuando la evidencia no es suficiente para llegar a una conclusión. Esta estrategia se basa en el cálculo de tasas de error en torno a la decisión de profundizar la investigación en un determinado caso o descartarla. En este punto, se utilizan elementos de teoría de la decisión. Por último, la tesis aborda el problema de la priorización de la incorporación de nuevas evidencias. La priorización es necesaria para equilibrar el costo de dicha incorporación y el beneficio en términos de mejora en la toma de decisiones. Este punto se enfoca desde dos marcos teóricos diferentes: uno similar al planteado en el problema de optimización y otro utilizando teoría de la información. Las herramientas generadas se encuentran implementadas en el lenguaje R y están disponibles en el repositorio CRAN.

Palabras clave: *Ciencias forenses, Personas desaparecidas, Genética humana, Estadística bayesiana, Teoría de la información, Teoría de la decisión.*

Development of computational and mathematical tools for missing person search

The search for missing persons is a task that requires multiple steps, from gathering basic information about the case to making decisions regarding identity hypotheses. The objective is to compare two entities: the missing person (bodyless identity) and the unidentified person (identityless body). Throughout the process, different lines of evidence are collected, each with varying degrees of complexity, ranging from age to genetic characteristics of individuals. The role of the forensic scientist is to assign probabilities to the different pieces of evidence, considering the propositions put forward to explain them as long as they can be mathematically formalized. While several lines of evidence can be formalized, genetic data has become the key piece for identifications in recent years, leaving other evidence relegated. When the initially collected genetic data is insufficient, either due to poor sample conditions or difficulty in genotyping individuals, it is necessary to incorporate more evidence and make decisions based on what is available. This thesis addresses some of these aspects and presents three problems: formalization, optimization, and prioritization. The formalization problem seeks to propose mathematical models based on a Bayesian approach to statistically evaluate the weight of the evidence collected during the preliminary investigation, such as age and sex. In the optimization problem, methodologies are formulated for the rational selection of cases that may be potential identifications, even when the evidence is not sufficient to reach a conclusion. This strategy is based on calculating error rates surrounding the decision to investigate a particular case further or discard it. At this point, elements of decision theory are utilized. Lastly, the thesis addresses the problem of prioritizing the incorporation of new evidence. Prioritization is necessary to balance the cost of such incorporation and the benefit of improved decision-making. This point is approached from two theoretical frameworks: one similar to that proposed in the optimization problem and the other using information theory. The generated tools are implemented in the R language and are available in the CRAN repository.

Keywords: *Forensic sciences, Missing persons, Human genetics, Bayesian statistics, Information theory.*

Agradecimientos

El principal agradecimiento es a mi directora, Inés, que durante estos años me orientó en el camino de la investigación. Me enseñó a tener paciencia, a profundizar en las preguntas y, sobre todo, a entender el trabajo científico como una forma de pensar y de actuar. Cuestionarse las creencias, ser autocríticos, explicitar los razonamientos, dudar de lo que es aparentemente obvio y buscar transparencia en los métodos son algunas de las tantas cosas que me transmitió.

A Ariel C., mi codirector, quien me acompañó desde épocas previas al inicio del doctorado en el mundo de las ciencias forenses, a través de distintas instituciones y contextos. En estos cambios y conversaciones junto a él pude entender que la investigación es bastante más que un trabajo que uno cumple en un lugar y lo termina al irse.

Al equipo, Ari S., Viqui y Tomi. Durante estos años aprendí mucho junto a Ariel, colaborando y estudiando juntos. Muchas de las líneas de esta tesis están motivadas en discusiones que tuvimos, siempre generando un clima ameno y lúdico. Con Viqui aprendí a corregir redundancias y ambigüedades, cosa que hoy considero sumamente útil y necesaria. Con Tomi colaboramos en proyectos en los cuales aprendí a contemplar siempre el potencial impacto de lo que se desarrolla.

A *los forenses*. De ellos y ellas aprendí lo fundamental del compromiso cotidiano con la Verdad y la Justicia. A Thore, con quien discutimos y desarrollamos herramientas, muchas que motivaron capítulos de esta tesis, sin perder de vista el gusto por el trabajo. A Pablo Gallo, quien me mostró el trabajo de búsqueda de personas desaparecidas, abriendo siempre la puerta a innovaciones, preguntas y desarrollos científicos. A Mariana, con quien nos realizamos muchas preguntas que motivaron desarrollos que se encuentran en la presente tesis, por su compromiso con una Ciencia al servicio de los Derechos Humanos. A Magnus por su vocación al desarrollo de código abierto y apertura para discutir ideas.

A Maca, que me acompañó todos estos años en este camino y en tantos otros, enseñándome a tener paciencia y disfrutar cada paso. A mi mamá y papá, que me motivaron desde muy temprana edad a cuestionarme e interrogar el mundo que me rodea. A mi hermano, Nico, con quien compartimos la vocación por investigar, y dimos nuestros primeros pasos juntos jugando. A mis amigos y amigas Mati, Salva, Romi, Pau G., Pau V., Carba, Guillli, Mary y muchos más con quienes comparto, ya desde hace varios años, las ganas de sorprendernos por un mundo inmensamente desconocido.

Índice general

Resumen	III
Agradecimientos	V
Lista de figuras	X
Lista de términos y abreviaciones	XV
Lista de términos y abreviaciones	XVI
1 Introducción	1
1.1 Ciencias Forenses	2
1.1.1 El rol del científico forense	2
1.1.2 Los desaparecidos	3
1.1.3 La búsqueda de personas desaparecidas	4
1.2 Inferencia probabilística	6
1.2.1 Perspectivas bayesiana y frecuentista	6
1.2.2 Redes bayesianas	12
1.2.3 Teoría de la información	16
1.3 La evidencia genética en casos forenses	18
1.3.1 La huella digital genética	19
1.3.2 Evaluación de la evidencia genética	21
1.4 Organización de la tesis	23
2 El problema de la formalización del peso estadístico de la evidencia: el caso de los datos recolectados durante la investigación preliminar	24
2.1 Introducción al capítulo	25
2.1.1 Información de la investigación preliminar	25
2.1.2 Redes complejas para encontrar relaciones no evidentes	27
2.1.3 Aprendiendo de los casos resueltos	28
2.1.4 La probabilidad a priori y la investigación preliminar	28
2.2 Métodos	30

2.2.1	Datos de la investigación preliminar	30
2.2.2	Modelo de verosimilitud para variables cualitativas dicotómicas	30
2.2.3	Modelo de verosimilitud para variables cualitativas politómicas	31
2.2.4	Modelo de verosimilitud para variables continuas	32
2.2.5	Cálculo de prior odds	33
2.2.6	La búsqueda en bases de datos	34
2.2.7	Evaluación del poder estadístico de la evidencia	35
2.2.8	Simulaciones computacionales	37
2.3	Resultados	37
2.3.1	Variable categórica dicotómica: el sexo biológico	37
2.3.2	Variable categórica politómica: el color de pelo	40
2.3.3	Variable continua: la edad	41
2.3.4	Combinando la evidencia	41
2.3.5	Análisis de sensibilidad con el prior odds	45
2.4	Discusión	45
2.5	Conclusiones del capítulo	46
3	El problema de la formalización del peso estadístico de la evidencia: análisis de parentesco genético	47
3.1	Introducción al capítulo	48
3.1.1	Parentesco genético	48
3.1.2	El nivel poblacional	51
3.1.3	El nivel del pedigrí	51
3.1.4	El nivel observacional	52
3.1.5	Algoritmos para el cómputo de la verosimilitud	52
3.2	Métodos	53
3.2.1	Datos genéticos	54
3.2.2	La evaluación estadística de la evidencia genética	54
3.2.3	Modelos poblacionales	54
3.2.4	Modelos para el pedigrí	56
3.2.5	Modelos observacionales	59
3.2.6	La expresión general de verosimilitud	60
3.3	Resultados	61
3.3.1	Estructura de la red	61
3.3.2	Análisis de probabilidades alélicas condicionadas	64

3.3.3	Probabilidades genotípicas y cociente de verosimilitud	65
3.3.4	Múltiples marcadores	67
3.4	Discusión	67
3.5	Conclusiones del capítulo	70
4	El problema de la optimización en la toma de decisiones: selección racional de umbrales para declarar potenciales identificaciones	73
4.1	Introducción al capítulo	74
4.1.1	La búsqueda mediante uso de bases de datos	74
4.1.2	La utilización de bancos de muestras para la reparación de crímenes de lesa humanidad	75
4.1.3	El poder estadístico	76
4.1.4	Umbral de cociente de verosimilitud	76
4.1.5	Distribuciones de cociente de verosimilitud	77
4.1.6	Toma de decisiones	77
4.2	Métodos	79
4.2.1	Base de datos de la investigación preliminar	79
4.2.2	Datos genéticos	79
4.2.3	Cálculo de prior odds	79
4.2.4	Análisis de parentesco genético	80
4.2.5	Cálculo de las posterior odds	80
4.2.6	Simulaciones computacionales	80
4.2.7	Evaluación del poder estadístico	81
4.2.8	Cálculo del umbral de decisión	83
4.2.9	Estrategia de búsqueda en la base de datos	83
4.3	Resultados	84
4.3.1	Proceso de identificación considerando solo datos genéticos	84
4.3.2	Incorporando datos de la investigación preliminar	91
4.4	Discusión	94
4.5	Conclusiones del capítulo	95
5	El problema de la priorización para la incorporación de nueva evidencia	96
5.1	Introducción al capítulo	97
5.1.1	El problema de priorización	97
5.1.2	Métodos basados en el cálculo del poder estadístico	98

5.1.3	Métodos basados en el cálculo del contenido de información	99
5.2	Métodos	100
5.2.1	Análisis de parentesco genético	100
5.2.2	Estrategia 1: métricas de poder estadístico	100
5.2.3	Estrategia 2: cálculo del contenido de información	101
5.2.4	Vinculando <i>CV</i> y la divergencia de Kullback-Leibler	102
5.3	Resultados	103
5.3.1	Estrategia 1: métricas de poder estadístico	103
5.3.2	Estrategia 2: cálculo del contenido de información	109
5.4	Discusión	114
5.5	Conclusiones del capítulo	115
6	Conclusiones y perspectivas	116
	Referencias	119
	Lista de publicaciones	129

Índice de figuras

1.1	Grafo Acíclico Dirigido del sistema de alarma contra robos. Los nodos representan las distintas variables, siendo terremoto, T , robo, B , radio R , alarma Al y llamado L . Las flechas representan las relaciones condicionalmente dependientes entre las variables.	14
1.2	Secuencia genética para el marcador STR autosómico CSF1PO. Se indican distintos alelos caracterizados por la cantidad de repeticiones de la serie AGAT. La orientación 5'-3' refiere a la orientación de la secuencia de ADN.	20
2.1	Esquema general de la búsqueda en bases de datos. Se muestra un evento de desaparición de una niña, a partir del cual la misma pasa a formar parte de un conjunto de personas no identificadas. Por otro lado, los familiares buscan a la niña desaparecida. A partir del primer paso, referido como investigación preliminar, se analizan datos que constituirán el prior odds. El segundo paso, el del test de parentesco genético permite el cálculo del cociente de verosimilitud. El último paso implica tomar la decisión en función del posterior odds obtenido para cada persona no identificada analizada.	35
2.2	Valores de $\text{Log}_{10}(CV_S)$ para cada sexo de PNI, considerando un MP femenino, en función de la proporción de femeninos en la población de referencia.	38
2.3	Frecuencia relativa de los valores de CV obtenidos en la simulación considerando H_1 y H_2 como verdaderas.	39
2.4	Métricas de rendimiento en función del valor de ϵ_S . En azul MCC, en rojo Recall y en negro precisión.	39
2.5	Frecuencia relativa de valores de $\text{Log}_{10}(CV)$ para el color de pelo, obtenidos en la simulación considerando que $C_{PD1} = 1$	40
2.6	Frecuencia relativa de valores de $\text{Log}_{10}(CV)$ para el color de pelo, obtenidos en la simulación considerando que $C_{PD1} = 5$	41
2.7	Frecuencia relativa de valores de CV para la edad, obtenidos en la simulación, dos PDs diferentes, con distintos rangos de edad.	42

2.8	Frecuencia relativa de valores de $\text{Log}_{10}(CV)$ para la combinación de variables obtenidos en la simulación considerando $C_{PD} = 1$, $S_{PD} = f$ y $E_{PD} = \{35, 45\}$. Se muestran los valores asumiendo H_1 y H_2 como verdaderas.	42
2.9	A y B. Distribución de probabilidad para la combinación de variables esperadas para PNI considerando H_2 como cierta, para PD_1 (izquierda) y PD_2 (derecha). D y E. Distribución de probabilidad para la combinación de variables esperadas para PNI considerando H_1 como cierta, para PD_1 (izquierda) y PD_2 (derecha). F y G. Valores de $\text{Log}_{10}(CV)$ para PD_1 (izquierda) y PD_2 (derecha) dada H_1 verdadera. Nótese que la figura E se obtiene diviendo C por A, y la F dividiendo D por B.	44
3.1	Caso de paternidad. Se comparan dos hipótesis, H_1 corresponde a la hipótesis de parentesco entre supuesto padre (SP) e hijo (H). H_2 es la hipótesis de no parentesco.	49
3.2	Concepto de niveles para los modelos de análisis de la evidencia genética. Se muestra el nivel poblacional, que describe las probabilidades genotípicas de los fundadores, el nivel del pedigrí que describe la herencia genética y el nivel observacional, que describe la probabilidad de obtener los genotipos considerando los métodos de laboratorio, entre otros. La Figura se basa en una presentada previamente por Egeland et al. (Egeland et al. 2015).	50
3.3	Red bayesiana de un caso simple con madre (Ma), padre (Pa) e hijo (Hi). El pedigrí utilizado para la construcción de la red se presenta a la izquierda. Cada nodo representa un alelo, y los conectores muestran las relaciones de dependencia condicional.	57
3.4	Pedigrí donde se presenta a un padre (1), una madre (2) y un hijo (3). La madre (en rayas) se encuentra genotipada.	62
3.5	Red bayesiana que surge a partir del análisis del pedigrí presentado en la Figura 3.4. En cada nodo, el nombre indica: individuo marcador alelo. Por ejemplo, 1 M1 p indica el individuo 1, marcador 1, alelo paterno. Cuando aparece S indica la variable selectora.	62
3.6	Pedigrí donde se indica un tío abuelo (8) genotipado. El individuo buscado es el 7. .	63
3.7	Estructura de la red bayesiana del caso del pedigrí presentado en la Figura 3.6. En cada nodo, el nombre indica: individuo marcador alelo. Cuando aparece S indica la variable selectora.	63

3.8	Probabilidades alélicas condicionadas para el individuo 3, marcador $M1$. A la izquierda se comparan la probabilidades para el alelo materno, tanto la condicionada por el pedigrí como la poblacional. A la derecha se comparan las probabilidades del alelo paterno con la poblacional.	65
3.9	Probabilidades alélicas condicionadas para el individuo 3, marcador $M2$. A la izquierda se comparan las probabilidades para el alelo materno, tanto la condicionada por el pedigrí como la poblacional. A la derecha se comparan las probabilidades del alelo paterno con la poblacional.	65
3.10	Probabilidades genotípicas del marcador $M1$ para el individuo 3 del pedigrí presentado en la Figura 3.4. A la izquierda se presentan las condicionadas por el pedigrí, en el centro las poblacionales y a la derecha el $\log_{10}(CV)$	66
3.11	Probabilidades genotípicas del marcador $M2$ para el individuo 3 del pedigrí presentado en la Figura 3.4. A la izquierda se presentan las condicionadas por el pedigrí, en el centro las poblacionales y a la derecha el $\log_{10}(CV)$	66
3.12	Configuración nuclear, resultados con la base de frecuencias de Argentina, 23 marcadores. Los cuadrados representan individuos masculinos, los círculos femeninos, en rojo se indica la persona desaparecida y con rayas los individuos cuyo genotipo está disponible. Por ejemplo, el pedigrí 1 cuenta con un hijo (3), su madre (2) y su padre (1). Solo la madre se encuentra genotipada.	68
3.13	Configuración medio-hermanos, resultados con la base de frecuencias de Argentina, 23 marcadores. Los cuadrados representan individuos masculinos, los círculos femeninos, en rojo se indica la persona desaparecida y con rayas los individuos cuyo genotipo está disponible.	69
3.14	Configuración avuncular, resultados con la base de frecuencias de Argentina, 23 marcadores. Los cuadrados representan individuos masculinos, los círculos femeninos, en rojo se indica la persona desaparecida y con rayas los individuos cuyo genotipo está disponible.	70
3.15	Configuración abuelos, resultados con la base de frecuencias de Argentina, 23 marcadores. Los cuadrados representan individuos masculinos, los círculos femeninos, en rojo se indica la persona desaparecida y con rayas los individuos cuyo genotipo está disponible.	71
4.1	Tabla de confusión donde se presenta el problema de clasificación binaria basado en el CV, considerando un umbral T	82

4.2 Izquierda: FPF y PFN para cada valor de T . Se marca en rojo cuando $T = UD$. Derecha: DEP en función de T , se marca en rojo la DEP mínima, o sea cuando $T = UD$.	83
4.3 Pasos durante un proceso de búsqueda de personas desaparecidas utilizando bases de datos. Se comienza por el paso 1, donde el poder estadístico es analizado para cada pedigrí. En caso de ser bajo, se buscan nuevos posibles miembros a ser incorporado (paso 2) o bien se continúa al cálculo de UD. Por último se realiza la búsqueda en la base de datos genética. El valor del umbral dependerá de si el pedigrí cuenta o no con suficiente poder estadístico.	85
4.4 Análisis del poder estadístico de los pedigríes. Se presentan 4 pedigríes simulados. En rojo se indica la persona desaparecida y en rayas los individuos genotipados. En el panel superior de la izquierda se muestran los valores de las métricas de poder estadístico para cada pedigrí. El valor óptimo (azul) se ubica en $PE = PI = 1$.	86
4.5 Análisis de la distribución de $\log_{10}(CV)$ considerando H_1 y H_2 como ciertas en un conjunto de casos con bajo poder estadístico. Se indica en línea violeta el valor de umbral de 10.000.	88
4.6 Curvas de tasas de falsos positivos y negativos para distintos valores de umbral. Se indica el UD en color rojo.	89
4.7 Cálculo de UD para el pedigrí F5. A Se muestra como UD es el valor de umbral que minimiza el DEP. B Se comparan las métricas de poder estadístico para el pedigrí F5 al considerar un umbral de 10.000 contra el umbral UD. Se observa la disminución de la probabilidad de falsos negativos a costo de un aumento de los falsos positivos.	90
4.8 Distribución de posterior odds considerando H_1 y H_2 como ciertas para el pedigrí F6. Se presentan los posterior odds calculados considerando el prior odds uniforme y los priors odds basados en datos de la investigación preliminar para cada caso, $PD1$ y $PD2$.	92
5.1 Análisis del poder estadístico para dos posibles combinaciones de genotipos en los familiares de referencia. (A) El pedigrí de referencia, donde Abuela y Tío son genotipados, y Tío 2 es una adición potencial. (B) Gráfico de poder estadístico. Los dos símbolos naranjas más grandes representan el poder promedio estimado para la incorporación del Tío 2. Cada punto pequeño consiste en el poder estimado para un Tío 2 específico.	104

5.2 Análisis estadístico de la incorporación de distintos posibles miembros al pedigrí de referencia (A). Se presentan las métricas PI y PE (B) y el CV_G promedio esperado acompañado del número de marcadores excluidos esperados (con H_2 cierta) (C)	106
5.3 Análisis estadístico del caso de los bisnietos. Se muestran pedigree de referencia en B, C y D. En A se muestran las distintas posibilidades de grupos de individuos a identificar. Si PNIs que son potenciales PD están disponibles (círculo blanco), el caso no dista de los previamente analizados en distintos ejemplos de este capítulo. En cambio, las otras opciones muestran combinaciones de posibles individuos a identificar, como por ejemplo un par de bisnietos (bis 1 + bis 2).	107
5.4 Análisis estadístico de la incorporación de nuevos sets de marcadores STRs autosómicos. Los puntos (triángulos o cuadrados) más pequeños muestran resultados de genotipos específicos simulados. Los puntos más grandes, cuadrado azul, triángulo rojo y círculo blanco, muestran los valores promedio para 23, 33 o el basal de 15 marcadores respectivamente.	108
5.5 (A) El flujo de análisis generado por fbnnet. (B) Estructura del pedigree, los individuos punteados están genotipados. (C) SE33 tabla de probabilidad condicionada (alelo materno). (D) SE33 tabla de probabilidad condicionada (alelo paterno). (E) CSF1PO tabla de probabilidad condicionada (alelo materno). (F) CSF1PO tabla de probabilidad condicionada (alelo paterno).	110
5.6 Ejemplo de pedigree con tres generaciones de ancestros de PD.	111
5.7 Análisis de distribución de métricas de informatividad para el tío, bisabuelo y abuelo. Izquierda - Distribuciones de Delta \mathcal{H} . Centro - Gráfico $KL - \overline{KL}$. Derecha - Gráfico $Cross\mathcal{H} - \overline{Cross\mathcal{H}}$	112
5.8 (A) Pedigrí analizado, donde el tío es genotipado. (B) Gráfico de $KL - \overline{KL}$ donde se indica el valor obtenido para el tío genotipado y las distribuciones obtenidas para los genotipos simulados de los posibles miembros a incorporar al pedigree.	113
5.9 (A) Pedigrí analizado, en rayas se muestra a los miembros genotipados. (B) Gráfico de $KL - \overline{KL}$ donde se indican las distribuciones obtenidas para los genotipos simulados de los posibles miembros a incorporar al pedigree.	114

Lista de términos y abreviaciones

PNI	Persona no identificada
PD	Persona desaparecida
DPD	Datos de la persona desaparecida
DPNI	Datos de la persona no identificada
KL	Divergencia de Kullback-Leibler
CV_G	Cociente de verosimilitud del paso genético
CV_{NG}	Cociente de verosimilitud del paso no genético
T_G	Umbral de selección del paso genético
T_{NG}	Umbral de selección del paso no genético
T	Umbral de selección general
PI	Probabilidad de inclusión
PE	Probabilidad de exclusión
PFP	Probabilidad de falso positivo
PFN	Probabilidad de falso negativo
PVP	Probabilidad de verdadero positivo
PVN	Probabilidad de verdadero negativo

Capítulo 1

Introducción

El término forense proviene de la palabra latina *forensis*. Esta era utilizada para describir aquello perteneciente al foro romano o *forum*. El mismo constituía un espacio físico donde convivían las diferentes instituciones de gobierno, legales, religiosas y comerciales. En el plano legal, el *forum* era el lugar donde la evidencia de un caso se presentaba. Con el advenimiento de los tiempos, el término forense se subscriptió al área legal y, particularmente, a un conjunto de disciplinas científicas relacionadas a la misma. Estas disciplinas se vinculan al análisis y presentación de la evidencia. La inferencia probabilística, dentro de las Ciencias Forenses, ha sido la herramienta encargada de evaluar el peso de la evidencia, desde una perspectiva estadística, siempre y cuando esta pueda ser formalizada matemáticamente. Con la disponibilidad actual de datos, el espacio de análisis forense se reconfigura, siendo los algoritmos computacionales, el diseño de bases de datos y los modelos probabilísticos parte del trabajo de búsqueda, recopilación, análisis e interpretación de la evidencia. Esto supone nuevos desafíos teóricos y computacionales en las Ciencias Forenses.

En este capítulo se introducen los aspectos fundamentales de las disciplinas relacionadas a un caso particular de las Ciencias Forenses, la búsqueda de personas desaparecidas. Describiremos el rol del científico forense en el marco general y en el caso particular planteado. Por su naturaleza investigativa, la búsqueda de personas desaparecidas se puede definir como un proceso que involucra múltiples pasos, desde la investigación preliminar, que implica recopilar datos relacionados a la desaparición, hasta la confirmación de la identificación. Proponemos que los distintos pasos poseen elementos que pueden ser matemáticamente formalizados, permitiendo la incorporación de modelos probabilísticos para su interpretación cuantitativa. Particularmente, se hace foco en la formalización matemática de distintas líneas de evidencia, como los datos recopilados durante la investigación preliminar. Además, se propone la formalización de distintos pasos en la toma de decisiones en casos de identificación, generando herramientas que asistan al investigador en la evaluación diagnóstica del sistema de identificación, y en la priorización de búsqueda de nuevas evidencias.

1.1 Ciencias Forenses

Las Ciencias Forenses comprenden un conjunto de disciplinas integradas por profesionales de distintas áreas. Su origen se remonta a la antigüedad, existiendo testimonios de procedimientos para el cuidado y tratamiento de la evidencia en civilizaciones como las de las antiguas Roma y Grecia (Brettell et al. 2005). Uno de los primeros registros de documentos escritos relacionados a la medicina forense corresponde a Song Ci, publicado en el siglo XIII en China (McKnight et al. 1983). El mismo da cuenta de los conocimientos de patología de la época e introduce guías y protocolos para la presentación de la autopsia, el cuerpo como evidencia, frente a la corte. A lo largo de los tiempos, con innovaciones tecnológicas y la generación de nuevas disciplinas las Ciencias Forenses fueron ampliando su campo de práctica. Desde sus inicios en la medicina (Brettell et al. 2005) y entomología (Amendt et al. 2011) hasta lo que hoy en día constituye una vasta cantidad de áreas, como la antropología forense (Schmitt et al. 2006), la física forense (Cross 2008), la genética forense (Amorim and Budowle 2016), economía forense (Zitzewitz 2012), entre otras.

1.1.1 El rol del científico forense

En trabajos pioneros, Evett y Weir describen los principios para la interpretación de la evidencia. Uno de los aspectos fundamentales que mencionan es el rol del científico forense en la asignación de probabilidades de observar los datos relacionados a los casos dadas las proposiciones consideradas (Evett and Weir 1998, Evett et al. 2000).

Este concepto puede ser explicado a través de un ejemplo sencillo: en un caso se acusa a un individuo de haber cometido un hecho delictivo. Se presenta una evidencia, llamada E , que podría vincular al acusado con el crimen. Ejemplos de evidencias son rastros de sangre, materiales biológicos, audiovisuales, etc. Existen dos posturas alternativas o explicaciones planteadas para que se haya producido E . Una es formulada por los fiscales, que son los encargados de llevar a cabo la investigación, y vincula al acusado con el delito. La otra es planteada por la defensa, y desvincula al acusado con el crimen. Al argumento de la fiscalía se lo denomina F , y al argumento de la defensa, D . En este ejemplo, según la perspectiva de Evett y Weir, el científico forense es aquel que evalúa la probabilidad de que se haya producido la evidencia E si F fuera cierta y, por otro lado, si D fuese cierta. Estas probabilidades son denominadas verosimilitudes, y los argumentos esbozados por la fiscalía y la defensa constituyen las hipótesis.

Como puede resultar intuitivo de pensar, existe evidencia cuya formalización matemática es compleja. Esta generalmente se encuentra relacionada con fuentes testimoniales, narraciones de carácter subjetivo, entre otras. Por otra parte, existen tipos de evidencia que cuentan con modelos probabilísticos robustos y validados por la comunidad científica. Este es el caso de la utilización del ADN (Ácido desoxi-ribonucleico) en el ámbito forense (Evett and Weir 1998). Existen también tipos de evidencia cuya formalización matemática es posible, pero no es ampliamente utilizada, generalmente por falta de consenso dentro de la comunidad en torno a la producción de modelos matemáticos (Budowle et al. 2011). Ejemplos de este último tipo son datos relacionados al caso, como el color de pelo de una persona desaparecida, la edad, etc.

1.1.2 Los desaparecidos

Un caso particular en Ciencias Forenses refiere a la búsqueda de personas desaparecidas (Puerto et al. 2021, Egeland et al. 2015). Los motivos de las desapariciones son diversos, pueden ser actuales o del pasado y se relacionan a catástrofes o desastres naturales (Corach 2010), conflictos bélicos (Parsons et al. 2019), desapariciones forzadas por terrorismo de Estado (Caridi et al. 2020, Puerto and Tuller 2017), migraciones (Reineke 2016, Citroni 2017, Doretti et al. 2017), tráfico humano (Bao et al. 2019, Aning and McIntyre 2004), entre otros. Existe una falta de estadísticas confiables acerca del número de personas que desaparecen anualmente en el mundo (Fyfe et al. 2015, Henderson et al. 2000). En parte, esto se debe a la heterogeneidad de contextos y procesos que derivan en las desapariciones.

Debido al fuerte impacto social, los conflictos bélicos y distintos sucesos de violencia política han sido paradigmáticos dentro de este campo. Por ejemplo, la guerra de los Balcanes, que tuvo en el eje del conflicto a distintos países del territorio yugoslavo durante la década de los 90', dejó un saldo de más de 40 mil personas desaparecidas (Stover and Shigekane 2002). Distintos esfuerzos derivaron en la identificación de más del 80 % de las víctimas al día de la fecha de la presente tesis.

Por otro lado, la violencia y discriminación racial ha generado casos de desaparición personas de grupos étnicos específicos. Un ejemplo proviene de Sudáfrica en el período entre 1960 y 1994. Una de las principales razones fue el programa conocido como *apartheid*, que implicaba definir zonas específicas para ser habitadas por personas *no-blancas*, y otras para *blancos* (Aronson 2011). La comunidad y distintas organizaciones no-gubernamentales locales e internacionales colaboraron con la búsqueda de los desaparecidos durante el proceso. Otro ejemplo se dio durante la Segunda Guerra Mundial, tanto por el conflicto bélico en sí mismo, como también dentro del marco conocido como el holocausto, donde políticas raciales llevaron al ejercicio de la violencia contra grupos étnicos específicos derivando en el asesinato y desaparición de miles de personas (Marjanović et al. 2015).

Con estos ejemplos se busca remarcar que en muchos casos las desapariciones se encuentran asociadas a programas que ponen a determinados grupos sociales, étnicos, culturales o políticos como blanco de una violencia específica. En estos casos la desaparición de personas es otra expresión de dicha violencia.

Un caso local se dio en el contexto de la última dictadura cívico-militar Argentina (1976-1983). En 1978, durante el período, se realizó el mundial de fútbol en Argentina. Marta Moreira de Alconada Aramburú, integrante de Madres de Plaza de Mayo, fue entrevistada por un medio holandés, y las siguientes fueron sus palabras:

“Nosotros solamente queremos saber dónde están nuestros hijos. Vivos o muertos, pero queremos saber dónde están. Ya no sabemos a quién recurrir: consulados, embajadas, ministerios, iglesias, (...) se nos han cerrado las puertas. Por eso les rogamos a ustedes, son nuestra última esperanza. Por favor, ayúdenos.”

Con *ustedes*, se refería a la prensa internacional, y la denuncia era por la desaparición de miles de jóvenes a causa de persecuciones políticas dentro de lo que se conoció como el *terrorismo de Estado* (Bosco 2006). La búsqueda de las Madres de Plaza de Mayo, una organización generada por familiares de desaparecidos por el *terrorismo de Estado*, impulsó avances científicos

cos y la creación de equipos de trabajo para la identificación de las víctimas. Este proceso tuvo fuertes repercusiones a nivel nacional e internacional, constituyendo un hito tanto dentro de los Derechos Humanos como en el campo de las Ciencias Forenses (Cordner and Tidball-Binz 2017, Fondebrider 2016, King 1991). En paralelo, la organización Abuelas de Plaza de Mayo se focalizó en la identificación de los niños y niñas robadas durante el terrorismo de Estado. Esta búsqueda derivó en la generación del *test* de abuelidad, una herramienta de genética estadística que permitía identificar nietos y nietas a partir de muestras obtenidas de sus abuelos y abuelas. Esto fue necesario debido a que en muchos casos no se contaba con información genética de los padres y madres de los niños robados, por encontrarse los mismos desaparecidos (King 1991). Aunque no se ahondará en detalles históricos referidos al caso argentino de búsqueda de personas desaparecidas, se listan referencias de la literatura a continuación: el Nunca Más es un libro fundamental que recopila los eventos relacionados al terrorismo de Estado ocurridos durante el período (*CONADEP, Nunca Más* 1984); Tumbas anónimas (Cohem Salama 1992) es el informe sobre la identificación de restos de víctimas de la represión ilegal, elaborado por el Equipo Argentino de Antropología Forense; y el libro de Gorini, La Rebelión de las Madres, cuenta la historia de las Madres de Plaza de Mayo (Gorini 2017).

1.1.3 La búsqueda de personas desaparecidas

Debido a la heterogeneidad de los contextos que derivan en desapariciones, en la literatura se encuentran diferentes definiciones y tipos de prácticas asociadas a la búsqueda de personas desaparecidas. Recientemente, Salado Puerto et al. (Puerto et al. 2021) la definieron de forma general en el marco de las Ciencias Forenses. En el artículo se describen conceptos fundamentales y los múltiples pasos que componen una búsqueda. Como aspecto general, se formaliza que el proceso de identificación se centra en el desafío de unir dos entidades. Por un lado la entidad PNI (*Persona No-Identificada*), refiere a un individuo o resto humano cuya identidad es desconocida. Por otro lado PD (*Persona Desaparecida*) se refiere a una identidad de la cual se desconoce el paradero. Durante la investigación se recopilan datos de ambas entidades. El cotejo de dichos datos constituirá la evidencia del caso. Generalmente, para PD, los datos se recopilan mediante entrevistas a familiares o seres queridos. En cambio para PNI, pueden tomarse directamente del lugar del hallazgo de los restos (en caso de tratarse de una identificación de restos humanos) o del individuo cuya identidad biológica se encuentra dubitada (por ejemplo en casos de tráfico humano).

El protocolo Minessota (Frey 2019), ha sido generado como una guía práctica de carácter internacional para la búsqueda de personas desaparecidas. Contiene recomendaciones legales y metodológicas. En una adaptación producida por el Equipo Argentino de Antropología Forense, se describen una serie de pasos que caracterizan al proceso asociado a la identificación de restos humanos. Los mismos se presentan y describen a continuación:

1. Investigación preliminar: este primer paso consiste en recopilar toda la información posible sobre la persona que se busca. Primordialmente se intenta obtener tres tipos de información: la información previa a la desaparición o muerte, la información biológica de la persona buscada y la información del grupo familiar del desaparecido. La información previa se utiliza para reconstruir los hechos en torno a lo sucedido con la víctima y puede

provenir de distintas fuentes como testimonios, llamadas telefónicas, redes sociales, entre otras. Las fuentes orales son fundamentales, los testigos primarios son aquellos que presenciaron los hechos directamente, los testigos secundarios suelen estar relacionados a la víctima. La información biológica consiste en recopilar datos como edad, estatura, tatuajes, fracturas y rasgos que puedan ser de utilidad para una identificación. Además, se toman muestras para el perfilamiento genético de los familiares directos. Se constituye también el árbol genealógico o pedigrí familiar, que concentra datos de los individuos relacionados al desaparecido (Puerto et al. 2021).

2. Etapa arqueológica: implica el trabajo sobre los restos humanos y está compuesta por dos procedimientos, por un lado la búsqueda de restos y por otro lado la recuperación. La búsqueda es un proceso complejo vinculado con la investigación preliminar, ya que a veces la localización del resto puede obtenerse mediante testimonios de testigos. La recuperación implica el trabajo con los restos una vez encontrados, a su vez requiere de una documentación precisa y trazable. Una vez recuperados los restos se transladan al lugar de análisis, el laboratorio.
3. Etapa de laboratorio: esta instancia consiste en analizar el material óseo mediante diferentes técnicas científicas. Un equipo multidisciplinario se encarga de caracterizar y analizar a los mismos. El laboratorio cuenta con una infraestructura de extracción de muestras y análisis genéticos con fines identificatorios. Se analizan distintas líneas de evidencia, como estimación de la edad, intervalo de muerte, etc.
4. Proceso de identificación: este paso implica la comparación de los datos recopilados para la persona desaparecida, PD, y los restos hallados no identificados, PNI. Se analizan distintas líneas de evidencia como edad, características físicas, vestimenta y el perfil genético, que generalmente implica la comparación entre el resto óseo y el grupo familiar de PD. La identificación constituye un proceso multidisciplinario y abarcativo de todas las líneas de evidencia. En caso de no ser concluyente, se procede a la búsqueda de nuevos datos que permitan arribar a un resultado fehaciente.
5. Notificación de resultados a los familiares y a la comunidad: una vez concluida la investigación, los resultados son comunicados a los familiares de las víctimas y a la comunidad afectada. Este último paso es muy sensible, y requiere un trato adecuado brindando el derecho de acceder a visualizar los restos y conocer distintos aspectos de la investigación llevada a cabo con el fin de arribar al resultado.

Cada uno de los pasos de la investigación es llevado a cabo por especialistas en estrecha comunicación. Asimismo, a pesar de presentarse los pasos como secuenciales, el proceso es dinámico en el sentido de que el resultado de un paso puede derivar en la necesidad de profundizar en otro. A modo de ejemplo, el proceso de identificación puede no derivar en resultados concluyentes por falta de suficientes muestras de ADN de familiares del desaparecido. Esto podría resultar en la necesidad de volver a la investigación preliminar, con el fin de reunir nuevos testimonios que permitan sostener una búsqueda de más restos humanos en la etapa arqueológica.

A lo largo de la tesis se analizará la información producida durante los distintos pasos y, cuando sea posible, su formalización matemática.

1.2 Inferencia probabilística

La inferencia probabilística ha acompañado al progreso científico y de la humanidad desde hace siglos. Su amplia contribución reside, entre otras cosas, en que permite vincular datos de la realidad, empíricos, con modelos ideales o teóricos. La inferencia probabilística ha sido interpretada como una ampliación de la lógica aristotélica a contextos en los cuales no se conoce el valor de verdad de una proposición, pero sí la posibilidad de la misma (Jaynes 2003). Respecto a la presente tesis, se asumirá que el lector conoce los fundamentos del cálculo de probabilidades. Existe extensa bibliografía que aborda el tópico, se recomienda el libro de McElreath (McElreath 2020) y el de Gelman (Kruschke 2010). Importantemente, ambas referencias abordan un aspecto relevante para esta tesis, el enfoque bayesiano.

1.2.1 Perspectivas bayesiana y frecuentista

La perspectiva frecuentista y la bayesiana se distinguen en la forma de interpretar a la probabilidad. La visión frecuentista considera a la probabilidad como la frecuencia límite de una serie de repeticiones de un experimento. Es decir, si las condiciones de un experimento se mantienen, y el mismo se repite una suficiente cantidad de veces, la frecuencia de los sucesos resultantes del experimento tenderá a asemejarse a la probabilidad. En casos forenses, pensar a la consumación de un crimen como un experimento repetible una determinada cantidad de veces no resulta adecuado. Desde la perspectiva bayesiana, la probabilidad de un suceso representa cuan creíble es que éste ocurra, dado un conjunto de parámetros asociados al experimento.

Supóngase un ejemplo sencillo, el lanzamiento de un dado de seis caras. Desde una perspectiva frecuentista, conocer si el dado está balanceado o no, es decir, conocer la probabilidad de que salga cada uno de los seis posibles valores, implicaría arrojar el dado un conjunto determinado de veces, y construir las probabilidades a partir de las frecuencias obtenidas. Desde una perspectiva bayesiana, se podría pensar a la probabilidad de los eventos condicionada al hecho de que el dado esté balanceado, y compararla por ejemplo con un escenario en el cual no lo estuviera. Esto permitiría, posteriormente, utilizar la información de los resultados de arrojar el dado (aunque sea solo un lanzamiento) para construir la probabilidad de que el mismo esté o no esté balanceado. Este último paso implica combinar la información del resultado obtenido al arrojar el dado con la creencia previa, subjetiva, acerca de si el mismo estaba o no balanceado. La relación entre la perspectiva bayesiana y la estadística forense se da desde hace tiempo (Fenton et al. 2016). La razón es entendible analizando el caso del dado, como se verá en mayor profundidad más adelante.

En esta parte se introduce un conjunto de notaciones que acompañarán a lo largo de la tesis. Para dicho objetivo, se continuará con el ejemplo del dado, describiendo cómo podría ser abordada su formalización matemática, de la misma manera que se hará más adelante con otras evidencias.

Se puede definir a los valores que puede tomar un dado de seis caras como una variable

aleatoria denominada X . Estos valores se organizan dentro de lo que se denomina el alfabeto o espacio muestral de la variable X . Al alfabeto se lo denomina A_X . Dependiendo el caso, A_X puede contener números discretos, continuos, palabras o bien valores binarios. En el ejemplo del dado A_X contiene valores de 1 a 6. Formalmente se dirá que $A_X = \{1, 2, 3, 4, 5, 6\}$, indicando dentro de los corchetes el conjunto de valores que pueden resultar de arrojar el dado. Por otro lado, se define a $P(X = x)$, con $x \in A_X$, como la probabilidad de que X tome el valor x . La probabilidad P va desde 0 (que representa un suceso imposible) hasta 1 (que representa un suceso cierto). Los valores del alfabeto A_x deben ser disjuntos, dicho de otro modo, en una tirada X puede tomar tan solo un valor del alfabeto (por ej. el resultado de una tirada no puede ser 1 y 3 a la vez). Se dirá que la instancia de X ocurre cuando $X = x \in A_X$.

A partir de este punto se definen un conjunto de reglas de la probabilidad que serán utilizadas a lo largo de la tesis.

- **Regla de la suma:** para dos posibles instancias de X , siguiendo el ejemplo del dado, $X = x_1 = 1$ o $X = x_2 = 2$, la probabilidad de que ocurra un evento u otro está determinada por

$$P(X \in \{x_1, x_2\}) = P(X = x_1) + P(X = x_2) = P(X = 1) + P(X = 2) \quad (1.1)$$

Es decir, la probabilidad de que ocurra un evento u otro es igual a la suma de la probabilidad de que ocurra cada uno de los eventos por separado.

- **Normalización:** es una propiedad intrínseca de la expresión de la probabilidad. Establece que la regla de la suma del conjunto del alfabeto, A_X debe ser igual a 1. Dicho en términos del ejemplo, la probabilidad de que al arrojar un dado se obtenga 1, 2, 3, 4, 5 o 6, debe ser igual a 1. Formalmente se expresa en la siguiente ecuación:

$$P(X \in A_X) = \sum_{x_i \in A_X} P(X = x_i) = 1 \quad (1.2)$$

- **Regla del producto:** si se consideran dos eventos independientes, como tiradas sucesivas del dado, se puede preguntar cuál es la probabilidad de obtener un determinado valor primero y luego otro. Por ejemplo, se requiere la probabilidad de arrojar un dado dos veces y obtener un 1 y un 2, o sea $X = 1$ y $Y = 2$. En este caso aparecen dos variables, dado que X es el resultado de arrojar un primer dado e Y el resultado del segundo dado, con las mismas características, es decir $A_Y = A_X$.

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y) = P(X = 1) \cdot P(Y = 2) \quad (1.3)$$

Se llama a esta expresión la probabilidad conjunta de X e Y . Toma esta forma sólo en casos en los cuales X e Y son variables independientes. Frecuentemente, se suele calcular la probabilidad de que un resultado se repita sucesivas veces. Por ejemplo, si se arroja un mismo dado 5 veces, la probabilidad de que se obtenga el valor 2 en todas las tiradas. Utilizar la expresión anterior implicaría multiplicar cinco veces la probabilidad $P(X = 2)$. Con el fin de simplificar la notación, puede ser útil definir una variable, llamada L_v ,

donde v es la cantidad de veces que se obtiene el valor especificado. Por ejemplo, $L_7 = 2$, define al evento de obtener siete veces el valor 2 al arrojar un determinado dado.

- **Marginalización:** al encontrarse con una probabilidad conjunta, como la definida previamente, $P(X, Y)$, se denomina marginalizar sobre Y a sumar todos los posibles valores que tome Y , considerando que $X = x$. Formalmente se define de la siguiente manera:

$$P(X = x) = \sum P(X = x, Y = y) \quad (1.4)$$

En el ejemplo del dado sería calcular la probabilidad de que en la tirada X se obtenga x y que luego, en la tirada del dado Y , se obtenga cualquier valor. Como se definió previamente por la regla de la normalización, la probabilidad de que en Y se obtenga alguno de los valores de A_Y es igual a uno.

Previamente a introducir las próximas definiciones es necesario ampliar el ejemplo del dado. Se mencionó que en un caso forense, generalmente distintas hipótesis son contrastadas para explicar los hechos. En este caso, la cuestión residía en torno a si el dado se encontraba o no balanceado. Se define, por lo tanto, una variable denominada B , tal que $A_B = \{0, 1\}$, siendo 0 un indicativo de que el dado está balanceado, y 1 de que no lo está. Se plantea además, que el desbalance favorece la obtención del valor 2 por encima del resto de los valores del dado. En caso de estar balanceado, lo que se espera es una equiprobabilidad para todos los valores posibles. Dicho esto, se define la probabilidad condicional.

- **Probabilidad condicional:** en ciertos casos, como en el planteo del dado, la información sobre el sistema bajo estudio puede modificar el conocimiento acerca de las probabilidades de los eventos, o dicho en términos bayesianos, puede modificar las creencias acerca de que ocurran ciertos eventos. La probabilidad condicional es una regla que permite estudiar la probabilidad de un evento condicionada en el conocimiento de otro. Por ejemplo, el valor que tome la variable previamente descrita B , condiciona la probabilidad $P(X)$. A continuación definimos la probabilidad condicional formalmente:

$$P(X = x|B = b) = \frac{P(B = b, X = x)}{P(B = b)} \quad (1.5)$$

En este caso se conoce que existe una dependencia condicional entre B y X , descrita por los modelos previamente explicados (dado balanceado o desbalanceado). Si no existiese dicha dependencia, como en el caso de X e Y , podría verificarse que se cumple lo siguiente:

$$P(X = x|Y = y) = \frac{P(Y = y, X = x)}{P(Y = y)} = \frac{P(Y = y) \cdot P(X = x)}{P(Y = y)} = P(X = x) \quad (1.6)$$

En el caso del dado es intuitivo pensar que obtener un resultado en una tirada no afecta al resultado que se obtendrá en la siguiente, es decir el valor de X no condiciona al de Y y viceversa.

- **Regla de Bayes:** en la ecuación 1.5 se describe el cálculo de la probabilidad de un evento específico al arrojar el dado, $X = x$ condicionada sobre el hecho de si el dado está balanceado o no $B = b$. La regla de Bayes es una potente herramienta estadística que permite obtener la probabilidad de $B = b$, en este caso diremos que es la probabilidad de la hipótesis en el juicio de los dados, condicionada por los resultados obtenidos, $X = x$, que es el valor resultante de arrojar el dado. Esto es posible explicitando una creencia previa, o probabilidad a priori, para B , y calculando la probabilidad del resultado $X = x$, dado B . A esta última probabilidad se la denomina verosimilitud. A continuación se enuncia la regla de Bayes:

$$P(B = b|X = x) = \frac{P(X = x|B = b) \cdot P(B = b)}{P(X = x)} \quad (1.7)$$

Por lo tanto, $P(X = x|B = b)$ es la verosimilitud y, en el caso de los dados, representa la probabilidad de obtener el valor x dada la condición del dado (balanceado a desbalanceado). Por otro lado, $P(B = b)$ representa un aspecto fundamental dentro de la perspectiva bayesiana y es, como se mencionó previamente, la probabilidad a priori de que el dado esté o no balanceado. Esta indica las creencias previas al lanzamiento del dado. Por último, $P(X = x)$ o constante de normalización, resulta de marginalizar X sobre B .

1.2.1.1 Prueba de hipótesis

Como se ha mencionado previamente, en un caso forense, la evidencia suele ser explicada por hipótesis alternativas. Por lo tanto, el área de la estadística que se encarga de poner a prueba las hipótesis resulta fundamental. Generalmente, una hipótesis es presentada por la fiscalía y otra por la defensa. Utilizando el ejemplo del dado, se dirá que existen dos hipótesis, H_1 : el dado está balanceado, por lo tanto, $B = 0$, y H_2 : el dado está desbalanceado favoreciendo el valor 2, por lo tanto, $B = 1$. Se sabe además, que se obtuvo el valor de 2 en un total de 6 tiradas consecutivas. Se abordará el problema de prueba de hipótesis desde dos perspectivas, la frecuentista y la bayesiana.

Desde la perspectiva bayesiana, se utiliza el *posterior odds* como métrica comparativa entre las probabilidades a posteriori de dos hipótesis contrastadas. El cálculo del mismo consiste en dividir el posterior obtenido para una hipótesis sobre el de la otra. Utilizando la ecuación 1.7 describimos a continuación el cálculo del posterior odds:

$$\frac{P(H_1|X = x)}{P(H_2|X = x)} = \frac{\frac{P(X=x|H_1) \cdot P(H_1)}{P(X=x)}}{\frac{P(X=x|H_2) \cdot P(H_2)}{P(X=x)}} \quad (1.8)$$

Despejando los términos de la ecuación se obtiene la siguiente expresión:

$$\frac{P(H_1|X = x)}{P(H_2|X = x)} = \frac{P(X = x|H_1)}{P(X = x|H_2)} \cdot \frac{P(H_1)}{P(H_2)} \quad (1.9)$$

Como puede verse, los componentes de la ecuación 1.7 se mantienen, el posterior, la verosimilitud y el prior, para cada hipótesis. Al estar dividiéndose, al cociente de los priors se le denomina prior odds, $O(H)$, al de la verosimilitud, cociente de verosimilitud, CV , y al del

posterior, posterior odds , $O(H|X)$. Nótese que la constante de normalización, $P(X = x)$, al ser igual para ambas hipótesis, se anula en el cociente.

Con esta expresión se procede a evaluar ambas hipótesis en función de la evidencia. Para este fin, se amplía el ejemplo. Primero, durante el juicio del dado, la defensa y la fiscalía llevan a cabo investigaciones. La defensa postula simplemente que el dado se encontraba equilibrado, siendo $\frac{1}{6}$ la probabilidad de obtener cada uno de los valores luego de arrojarse. La fiscalía recluta expertos en producción de dados, que dan cuenta de una serie específica de dados desbalanceados producidos a gran escala para apuestas fraudulentas. Estos poseían una probabilidad mayor de obtener el número 2. Luego de estudiar esta serie de dados, se obtiene la probabilidad de cada uno de los resultados:

Valor	1	2	3	4	5	6
H_1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
H_2	$\frac{15}{96}$	$\frac{1}{3}$	$\frac{1}{96}$	$\frac{1}{24}$	$\frac{15}{96}$	$\frac{1}{96}$

Tabla 1.1 Tabla de probabilidades de valores (del 1 al 6) para el dado balanceado, H_1 , y para el caso desbalanceado, H_2 .

Es decir, para el caso desbalanceado es esperable que el número 2 salga una de cada tres veces que se arroja el dado. Por cuestiones físicas, el número 5 se ve particularmente desfavorecido (por encontrarse en la cara contraria al 2). El resto de los números presentan una equiprobabilidad. Con estos datos se puede calcular la probabilidad de obtener el número dos, condicionada por cada una de las hipótesis. El problema puede analizarse de la siguiente manera, considérese que el número total de tiradas del dado se denomina K , y la cantidad de veces que el resultado es un 2, se la denomina R . La probabilidad de $P(R)$ puede modelarse mediante una distribución binomial, donde R es la variable aleatoria a modelar, tal que:

$$R \sim B(k, w) \quad (1.10)$$

Siendo w la probabilidad de que se obtenga el valor 2. La esperanza, $E(R) = k \cdot w$, y la varianza, $VAR(R) = k \cdot w \cdot (1 - w)$. La función de densidad de probabilidad de la distribución binomial se presenta a continuación:

$$p(r|k, w) = \frac{k!}{k!(n-k)!} \cdot w^r (1-w)^{k-r} \quad (1.11)$$

Con el fin de simplificar la notación, se llamará L_6 al hecho de obtener un mismo valor, x , seis veces consecutivas al arrojarse el mismo dado. La siguiente expresión permite el cálculo del posterior odds:

$$\frac{P(H_2|L_6 = x)}{P(H_1|L_6 = x)} = \frac{P(L_6 = 2|H_2)}{P(L_6 = 2|H_1)} \cdot \frac{P(H_2)}{P(H_1)} = \frac{P(R \sim B(6, \frac{1}{3}))}{P(R \sim B(6, \frac{1}{6}))} \cdot \frac{P(H_2)}{P(H_1)} \quad (1.12)$$

El Juez estará encargado de definir los priors, dado que representan la subjetividad y el conocimiento previo en torno al caso. Comúnmente, si no hay evidencia previa el principio de

máxima incertidumbre es el escogido (Amorim and Budowle 2016). Esto implica que $P(H_1) = P(H_2) = 0,5$, es decir, no hay inclinación hacia ninguna de las hipótesis. Con estos valores se puede proceder a calcular el posterior odds:

$$\frac{P(H_2|X_6 = 2)}{P(H_1|X_6 = 2)} = \frac{\frac{1}{3}^6}{\frac{1}{6}^6} \cdot \frac{0,5}{0,5} = 64$$

El resultado obtenido, posterior odds, es de 64. La interpretación del mismo es que es 64 veces más probable observar los resultados (6 veces un 2) si H_2 es correcta, que si H_1 lo fuera. Es decir, la evidencia se inclina hacia el lado de que el dado se encuentra desbalanceado. Supóngase ahora un escenario diferente, donde se obtiene 6 veces el valor 1. Se sabe que la probabilidad de obtener 1 bajo H_2 es $\frac{15}{96}$, en cambio si H_1 es cierta, es de $\frac{1}{6}$. Se resuelve la ecuación 1.9 con los nuevos datos, manteniendo el principio de máxima incertidumbre en torno al prior.

$$\frac{P(H_2|X_6 = 1)}{P(H_1|X_6 = 1)} = \frac{\frac{15}{96}^6}{\frac{1}{6}^6} \cdot \frac{0,5}{0,5} = 0,67$$

El posterior odds obtenido en este caso es menor a 1, y se interpreta como: es 0,67 veces menos probable observar los resultados (6 veces un 2) si H_2 es correcta, que si H_1 lo fuera. Es decir, la evidencia se inclina hacia H_1 , con el dado balanceado. Nótese que una tercer hipótesis, donde el dado esté balanceado hacia el valor 1 podría alterar el análisis. Además, otra duda que puede surgir es el efecto del prior odds, es decir, el impacto que tienen los priors en el posterior final. Por ejemplo, si el prior odds es un valor muy grande, indica que a priori se cree que H_1 tiene más probabilidades de ser cierta, esto implicará que es necesario una fuerte evidencia para derivar en un posterior a favor de H_2 . Otro caso de toma de decisiones es la selección de un valor de posterior odds a partir del cual el Juez toma una decisión, por ejemplo, el valor 64 para este caso, ¿es suficiente para definir que el dado estaba desbalanceado hacia el 2?. Muchas de estas preguntas son objeto de debate en el campo de las Ciencias Forenses, y serán abordadas a lo largo de la tesis, proponiendo métodos y alternativas para la toma de decisiones.

Aunque el enfoque bayesiano y el frecuentista difieren en la concepción que le otorgan a la probabilidad, se presenta una manera frecuentista de abordar la prueba de hipótesis. En este caso, se seleccionará una de las dos hipótesis, que tendrá el carácter de hipótesis nula. Esta será la que la investigación pondrá a prueba, focalizándose en la probabilidad de observar los resultados dado que la misma es cierta. Un valor de corte, denominado valor de significancia, permitirá rechazar o aceptar la hipótesis en función de los resultados. En este caso, se selecciona como hipótesis nula a H_1 , que considera al dado balanceado. El modelo utilizado puede ser el mismo que el planteado para el cálculo de la verosimilitud en el planteo bayesiano. Como la hipótesis nula es H_1 , se tiene que:

$$E(R) = 6 \cdot \frac{1}{6}$$

$$VAR(R) = 6 \cdot \frac{1}{6} \cdot \frac{5}{6} = 0,83$$

La Esperanza, $E(R)$, indica que cada 6 tiradas del dado, una vez debería aparecer el valor 6, con una varianza de 0,83, considerando el modelo de H_1 . De la función de densidad de probabilidad se puede obtener que $P(6|6, \frac{1}{6}) = 0,00002$. De forma similar a lo observado para el

resultado de la comparación de hipótesis bayesiana, es necesario establecer un umbral a partir del cual tomar una decisión. ¿Es 0,00002 un valor suficientemente bajo como para rechazar la hipótesis de que el dado está balanceado? En el marco frecuentista, si el experimento se repite una infinita cantidad de veces, las frecuencias de los resultados se convierten en probabilidades. Por lo tanto, un enfoque podría ser tomar el dado en cuestión, y arrojarlo una suficiente cantidad de veces hasta obtener resultados convincentes de que el dado se encuentra balanceado, o no. En el ejemplo que se analiza, el dado no se encuentra disponible, es decir, no es evidencia. Solo los resultados del mismo, previamente obtenidos, pueden ser analizados. Este escenario emula una situación común en el campo forense. El enfoque bayesiano posee una ventaja respecto a este punto, y es que permite incorporar información del contexto en la probabilidad a priori. Es decir, permite la incorporación de otra información, subjetiva o de difícil formalización matemática. Además, como se verá más adelante, permite combinar líneas de evidencia de distintas naturaleza. La aplicación del enfoque frecuentista o bayesiano en el campo de las Ciencias Forenses fue, y continúa siendo, eje de debate y controversia (Taroni et al. 2016).

1.2.2 Redes bayesianas

En esta sección se hace hincapié en introducir aspectos de las redes bayesianas que son necesarios para entender el modelo presentado en el capítulo 3 sobre herencia genética. Existen distintas referencias en la literatura científica que ahondan en el tópico. Entre ellas está el libro de Darwiche (Darwiche 2009). En el mismo se introduce la notación, los modelos y su aplicación a distintos problemas reales. Por otra parte, el libro de Nagarajan et al. (Nagarajan et al. 2013) introduce al uso de las redes bayesianas con ejemplos prácticos, y es una buena referencia para la aplicación de la metodología en el campo de la biología de sistemas. En términos generales, las redes bayesianas corresponden a un tipo de modelo gráfico. Desde una perspectiva histórica, los primeros indicios de la utilización de modelos gráficos para la representación de información probabilística se dan en el campo de la física estadística en 1902, introducidos por Gibbs (Gibbs 1902). Años después, Wright (Wright 1921) los introduce en el campo de la biología. Además, la utilización de redes bayesianas ha sido propuesta como una herramienta robusta en Ciencias Forenses (Fenton et al. 2016).

La red bayesiana permite representar el conocimiento acerca de un mecanismo o situación particular dentro de un marco coherente (Darwiche 2009). Estas poseen dos componentes, por un lado uno cualitativo y por otro lado uno cuantitativo. El cualitativo corresponde a la definición de un Grafo Acíclico Dirigido (GAD) y determinará la *estructura* de la red. En este grafo, los nodos corresponden a variables, y los conectores explicitan relaciones condicionales entre las variables. Por otro lado, el componente cuantitativo servirá para reflejar las probabilidades que cuantifican las relaciones entre las variables unidas por conectores, es decir, las probabilidades condicionales. Definir el componente cuantitativo implica *parametrizar* la red. Uno de los aspectos fundamentales de la red bayesiana es que sólo las probabilidades que conectan directamente dos variables deben ser explicitadas, el resto se calcula automáticamente por algoritmos de inferencia. Por ejemplo, si tenemos tres variables, A, B y C tal que A se conecta con B, y B se conecta con C, solo habrá que explicitar las probabilidades condicionadas: $P(B|A)$ y $P(C|B)$. La $P(C|A)$ será determinada por los algoritmos. En redes con muchos nodos y conectores es-

te procedimiento puede dar lugar al cálculo de relaciones complejas y no evidentes entre las variables.

Se pueden mencionar al menos tres métodos para el modelado por redes bayesianas. Un primer método, fuertemente subjetivo, implica tomar el conocimiento personal, el de otros o el conocimiento experto, y reflejarlo en una red. Este podría ser el caso de la formalización del conocimiento experto de un médico especialista en una patología (Flores et al. 2011). Otro caso implica sintetizar en una red el conocimiento formal, por ejemplo los algoritmos clínicos, en los cuales se detalla el paso a paso de análisis para la realización de un diagnóstico. A estos dos tipos de enfoques se los conoce como *enfoques de representación del conocimiento*. Son de particular interés para esta tesis, dado que dentro de este grupo recae el problema de modelado de herencia genética (Fishelson and Geiger 2002). Un tercer enfoque es el propuesto por los algoritmos de *aprendizaje automático* donde la estructura de la red, la parametrización de la misma o ambas se definen a partir de un *set* de datos (Darwiche 2009). No ahondaremos en este último tipo de método en la tesis.

1.2.2.1 La estructura de la red

El enfoque en el cual se enmarcará la aplicación de redes bayesianas en esta tesis corresponde al de *representación del conocimiento*. Por este motivo, se sitúa un ejemplo acorde para introducir la notación y propiedades de las redes. Se supone un escenario en el cual hay una casa y sus dueños no se encuentran dentro. En esta hay una alarma, denominada Al , que puede ser activada por un robo, denominado B , o bien por el movimiento producido por un terremoto, llamado T . Una vez que se activa Al , se produce un fuerte ruido que genera que los vecinos llamen a los dueños de la casa, al llamado se lo denomina L . Por otro lado, si se produce T , el mismo será comunicado a través de la radio local, que se denomina R . Cada uno de los componentes mencionados constituyen las variables, y en todos los casos son proposicionales, es decir, pueden ser verdaderas o falsas. De forma general, se dirá que cuando estas variables toman el valor 1 son verdaderas, si toman un valor igual a 0, son falsas. Por ejemplo, $T = 1$ indica que se produjo un terremoto, y $Al = 0$ que no se activó la alarma. También se explicitan las dependencias condicionales. Por ejemplo, se sabe que si $T = 1$, la alarma debería activarse, por lo tanto $Al = 1$. Aunque se sabe que existe esa relación, todavía no se cuantificó la probabilidad de que ocurra un evento debido a que ocurrió el otro. Con la información con la que se cuenta ya se puede plantear la *estructura* de la red, presentada en la Figura 1.1.

A partir de esta estructura se puede definir un conjunto de notaciones dada una variable cualquiera, que se denominará V :

- Ancestros de V : aquellas variables, N , que tengan conectores de N a V .
- Descendiente de V : aquellas variables, M , que tengan a V como ancestro.
- No-descendientes de V : todas las otras variables que no son ni ancestros ni descendientes de V .

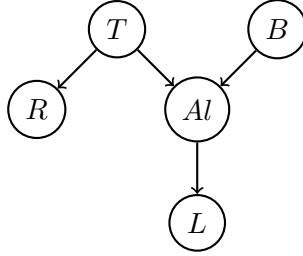


Figura 1.1 Grafo Acíclico Dirigido del sistema de alarma contra robos. Los nodos representan las distintas variables, siendo terremoto, T , robo, B , radio R , alarma Al y llamado L . Las flechas representan las relaciones condicionalmente dependientes entre las variables.

Para este caso puntual, donde se explicitan relaciones causales, podría decirse que los ancestros de V son las causas de V , y sus descendientes son las consecuencias. Importantemente, un atributo de la red bayesiana se basa en lo que se conoce como *asunción markoviana* o *condición de Markov*. Esta declara que todas las variables (nodos) de la red son condicionalmente independientes de sus no-descendientes, dado sus ancestros. En otros términos, esto implica que un nodo no tiene relación con aquellos que no descienden de él.

La terminología explicada se aplica al caso de la alarma de la siguiente manera: Al es descendiente de T , y como L es descendiente de Al , L es también descendiente de T . Los ancestros de Al son T y B . El ancestro de R es T . En este caso R no poseen descendientes, por lo tanto el resto de los nodos que no son T , son los no descendientes de R . La *asunción markoviana* implica que lo que se diga en la radio acerca del terremoto no tendrá efecto sobre la activación de la alarma, ni sobre la realización de un robo, ni sobre el llamado de los vecinos. Todo esto suponiendo que el valor T está dado, es decir, que se conoce si hay o no terremoto.

1.2.2.2 Las parametrización de la red

Como previamente se introdujo, parametrizar la red implica cuantificar mediante probabilidades, las relaciones de dependencia condicional descritas por los conectores. Más específicamente, para parametrizar el GAD lo que se requiere es: para cada variable V en el GAD, y sus ancestros N , se debe definir $P(v|n)$, para cada valor v de la variable V , y para cada instanciación n de las variables N . Volviendo al ejemplo presentado en la Figura 1.1, las probabilidades condicionadas a parametrizar son las siguientes:

$$P(l|a_l), P(r|t), P(a_l|t, b), P(t), P(b)$$

donde l , a_l , r , t y b son valores de las variables L , R , Al , T y B respectivamente. A modo de ejemplo, se determina para todos los posibles valores de las variables Al y L , la $P(l|a_l)$.

Dicho en palabras, si la alarma suena ($Al = 1$), hay un 0,80 de probabilidad de que los vecinos llamen ($L = 1$). Si la alarma no suena ($Al = 0$), hay sólo un 0,001 de probabilidad de que los vecinos llamen ($L = 1$). A diferencia de L , que se encuentra condicionada sólo por una variable, Al se encuentra condicionada por dos, que son T y B . La tabla de probabilidad de Al luce de la siguiente manera:

Como puede verse, si T y B ocurren en simultáneo, la probabilidad de que la alarma se active

Al	L	$P(l a_l)$
1	1	0,80
1	0	0,20
0	1	0,001
0	0	0,999

Tabla 1.2 Tabla de probabilidad para la variable llamado, L , condicionada por la variable alarma, Al .

B	T	Al	$P(a_l t, b)$
1	1	1	0,9999
1	1	0	0,0001
1	0	1	0,99
1	0	0	0,01
0	1	1	0,80
0	1	0	0,20
0	0	1	0,001
0	0	0	0,999

Tabla 1.3 Tabla de probabilidad para la variable alarma, Al , condicionada por las variables robo, B , y terremoto, T .

$$P(Al = 1|T = 1, B = 1) = 0,9999.$$

Ahora bien, el tamaño de la tabla de probabilidad para la variable L es de 4. En cambio, para la variable Al es de 8. Esto es un ejemplo de una propiedad de las tablas de probabilidad condicionada, y es que crecen exponencialmente con el número de variables que entran en juego. Considerando variables booleanas, como las analizadas en este caso, se puede decir que el tamaño de la tabla de probabilidad condicionada será de 2^{k+1} , siendo k la cantidad de ancestros de la variable para la cual se calcula la tabla.

1.2.2.3 Inferencia probabilística en la red

Por otro lado se supone que se instanciaron las variables B y T , siendo ambas iguales a 1. Interesa calcular la probabilidad de que los vecinos llamen, dada esta condición, es decir $P(L = 1|B = 1, T = 1)$. Esta probabilidad condicionada no se encuentra definida dentro de la parametrización de la red debido a que, como se mencionó, solo se parametrizan las relaciones entre las variables y sus ancestros. Aún así, es posible obtener dicha probabilidad a partir de inferir sobre la red, esto puede ser realizado manualmente en este sencillo ejemplo, y requiere de algoritmos en casos de mayor complejidad. En este caso, la siguiente expresión, aplicando la regla del producto, resuelve la probabilidad enunciada:

$$P(l|a_l, b, t) = P(l|a_l) \cdot P(a_l|t, b) \cdot P(t|b) = P(l|a_l) \cdot P(a_l|t, b) \cdot P(t) \cdot P(b) \quad (1.13)$$

Nótese que cuando dos variables son condicionalmente independientes, su probabilidad conjunta se obtiene a partir de la multiplicación de las probabilidades individuales. Es así, que $P(t|b) = P(t) \cdot P(b)$, esto sucede dado que T y B son condicionalmente independientes, con-

siderando las relaciones explicitadas en la Figura 1.1. En cambio, para $P(l|a_l)$, es necesario definir las probabilidades condicionadas dado que sí hay un relación de dependencia explicitada en la red. Ahora bien, si se quiere conocer la probabilidad $P(l|a_l, b, t)$, instanciando las variables B y T , lo que se debe hacer es marginalizar sobre el resto de los valores. Dado que B y T se encuentran instanciadas sus probabilidades serán igual a 1. De este modo la tabla 1.1 se reformula de la siguiente manera:

B	T	Al	L	$P(l a_l, b, t)$
1	1	1	1	0,79992
1	1	1	0	0,19998
1	1	0	1	0,0000001
1	1	0	0	0,0000999

Tabla 1.4 Tabla de probabilidad para la variable llamado, L , condicionada por la variable alarma, Al .

En resumen, si se marginaliza sobre Al queda que:

$$P(l = 1|a_l, b = 1, t = 1) = 0,7999201$$

$$P(l = 0|a_l, b = 1, t = 1) = 0,2000799$$

De esta forma se obtiene la probabilidad condicionada para L producto de que B y T sucedan (ambas igual a 1). En sí, las tablas de probabilidad condicionada constituyen distribuciones de probabilidad, que pueden ser analizadas con diferentes herramientas. A continuación se presentan métricas de teoría de la información, que serán empleadas en el capítulo 5 de la tesis.

1.2.3 Teoría de la información

En 1948, Claude Shannon publica el artículo *A Mathematical Theory of Communication* (Shannon 1948). En el mismo, se propone un formalismo matemático para abordar el problema de la transmisión de un mensaje a través de un canal ruidoso, es decir, a través de un medio que altera dicho mensaje. A lo largo del tiempo el enfoque trascendió el ámbito de la teoría de la comunicación, constituyendo un campo en sí mismo con aplicaciones en las Ciencias Físicas (Jaynes 1957a), económicas (Yang 2018), médicas (Benish 2020), ecología (Ulanowicz 2001) y biología de sistemas (Uda 2020). Ha sido propuesto también en el campo de la genética cuantitativa y poblacional (Galas et al. 2021). Dentro del campo de las Ciencias Forenses, Ramos et al., la analizaron como una opción para estudiar el *rendimiento* de los métodos de evaluación de la evidencia basados en el cociente de versosimilitud (Ramos et al. 2013).

En esta introducción se mencionan aspectos fundamentales de la teoría de la información, con especial énfasis en aquellas métricas que serán de utilidad. Para ahondar más en métodos y aplicaciones se recomienda el libro de Mackay (MacKay et al. 2003). Nuevamente, para introducir los conceptos y las métricas se utilizará el ejemplo del dado de seis caras.

1.2.3.1 Contenido de información y entropía

En este ejemplo, se supone que se poseen dos dados de seis caras, uno balanceado, que se llama Z y otro desbalanceado, que se llama W. A la distribución de probabilidades de Z, se la denomina $Z(x)$, y a la de W, $W(x)$, siendo x el valor obtenido, $A_x = \{1, 2, 3, 4, 5, 6\}$. Para simplificar, cuando se indique $W = 1$ o $Z = 2$, se hace referencia a que al arrojar el dado W o Z se obtuvieron dichos valores. La siguiente tabla muestra la probabilidad de obtener cada valor en una tirada:

Dado	1	2	3	4	5	6
Z	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
W	$\frac{15}{96}$	$\frac{15}{96}$	$\frac{2}{48}$	$\frac{1}{3}$	$\frac{15}{96}$	$\frac{15}{96}$

Tabla 1.5 Tabla de probabilidades de valores para un dado balanceado, Z, y para uno desbalanceado, W.

Se puede ver que la probabilidad de que salga cualquier valor al arrojar el dado Z es la misma, no así al arrojar el dado W. Este último se encuentra desbalanceado, favoreciendo que salga el valor 4. Debido a la organización de las caras del dado, el valor 3 es el menos probable, mientras que el resto de los 4 valores tienen la misma probabilidad. Si se arrojase el dado y obtuviese el valor 4 seguramente no sorprendería, dado que es el de mayor probabilidad. En cambio, si el valor obtenido fuera el 3, sería un resultado más sorprendente. Shannon introduce una métrica para cuantificar dicha sorpresa, y es denominada *contenido de información de Shannon*, generalmente simbolizada como $h(x)$, siendo x el valor que toma una variable X . Definimos matemáticamente a $h(x)$ de la siguiente manera:

$$h(x) = -\text{Log}_2 P(x) \quad (1.14)$$

El contenido de información, $h(x)$ comúnmente se expresa en *bits* (cuando la base del Log es 2), aunque también es posible expresarla en *Hartley* (cuando la base del Log es 10), o *nat* (cuando se aplica el logaritmo natural), entre otras opciones. Por ejemplo, para el dado Z se tendrá que

$$h(Z = 1) = h(Z = 2) = -\text{Log}_2\left(\frac{1}{6}\right) = 2.58$$

Lo mismo se cumple para el resto de los valores que puede tomar Z. En cambio, para W se tiene que

$$h(W = 1) = h(W = 2) = h(W = 5) = h(W = 6) = -\text{Log}_2\left(\frac{15}{96}\right) = 2,67$$

$$h(W = 3) = -\text{Log}_2\left(\frac{2}{48}\right) = 4,58$$

$$h(W = 4) = -\text{Log}_2\left(\frac{1}{3}\right) = 1,58$$

El mayor contenido de información para el dado W lo tiene el valor 3. Dicho de otro modo,

aquellos resultados más sorprendentes son los que otorgan mayor información. Por otro lado, el que menos contenido de información tiene es el valor 4, dado que es el más esperable. El resto de los valores, al tener igual probabilidad, poseen el mismo contenido de información. Se puede además promediar el contenido de información sobre todos los posibles valores que puede tomar cada dado. Esto lleva a la siguiente métrica, conocida como *Entropía de Shannon*, $\mathcal{H}(X)$. Se define formalmente como:

$$\mathcal{H}(X) = - \sum P(x) \log_2 P(x) \quad (1.15)$$

Entonces, queda que:

$$\begin{aligned} \mathcal{H}(Z) &= 6 \frac{1}{6} 2,67 = 2,67 \\ \mathcal{H}(W) &= 4 \frac{15}{96} 2,67 + \frac{2}{48} 4,58 + \frac{1}{3} 1,58 = 2,38 \end{aligned}$$

Como puede verse, aunque el dado W presenta un valor específico con alto contenido de información, en promedio los valores obtenidos por el dado Z tendrán más información. La entropía de Shannon puede interpretarse de la siguiente manera: la sorpresa esperada en promedio al obtener los valores de un determinado sistema. Realizando una interpretación del caso, diremos que es esperable que en promedio sea más informativo arrojar un dado balanceado, donde cualquier valor puede salir, que uno desbalanceado, donde ya se sabe que hay un valor con más posibilidades.

1.2.3.2 Divergencia Kullback-Leibler

Conocida como entropía relativa, la divergencia de Kullback-Leibler o divergencia KL es una medida de similitud entre dos distribuciones de probabilidad. Una definición formal de la divergencia KL es la siguiente:

$$D_{KL}(p(x)||q(x)) = \mathcal{H}(p, q) - \mathcal{H}(p) \quad (1.16)$$

$\mathcal{H}(p, q)$ es conocida como la entropía cruzada entre $p(x)$ y $q(x)$ mientras que $H(p)$ es la entropía de $p(x)$. La D_{KL} expresa la porción de la entropía cruzada que se debe a $q(x)$, quitando la propia entropía de la distribución $p(x)$. Otra forma de expresar la D_{KL} se muestra a continuación:

$$D_{KL}(p(x)||q(x)) = - \sum p(x) \log_2 \left(\frac{p(x)}{q(x)} \right) \quad (1.17)$$

Es importante notar que la divergencia KL tiene carácter asimétrico, es decir $D_{KL}(p(x)||q(x)) \neq D_{KL}(q(x)||p(x))$. Se ahondará en este aspecto cuando se apliquen estas métricas en el análisis forense.

1.3 La evidencia genética en casos forenses

En este apartado se introduce una línea de evidencia fundamental, la evidencia genética. El ácido desoxirribonucleico (ADN) es una de las moléculas fundamentales de la vida. Todas las células del organismo, salvo ciertas excepciones como los glóbulos rojos y la línea germinal

(células encargadas de transmitir el material genético a la progresión), poseen un programa completo de información genética. La estructura molecular del ADN fue descubierta en el año 1953 (Watson et al. 1953). El genoma humano constituye un código de cuatro tipos de nucleótidos (pieza constitutiva de la molécula de ADN). Estos son Adenina, Guanina, Citosina y Timina, simbolizados con las letras A,G,C y T respectivamente. La estructura molecular del ADN es una doble hélice, generada por dos cadenas complementarias de nucleótidos, que aparecen las bases C con G, y A con T. De esta manera, conocer la secuencia de nucleótidos de una cadena permite adivinar con exactitud la secuencia de la otra. A una posición específica de la secuencia se la denomina *locus*. El genoma se organiza en cromosomas, que constituyen secuencias independientes de nucleótidos. El humano cuenta con 23 pares de cromosomas (46 cromosomas en total). Estos se encuentran alojados dentro de lo que se conoce como el núcleo de la célula. Cada par está dotado por un cromosoma paterno y otro materno (Nurk et al. 2022). Esto implica que el humano posee la información genética duplicada, característica que distingue a los organismos diploides. Es decir, para un determinado *locus* se tendrá la información materna y la información paterna. La variante materna del locus es conocida como alelo materno, y la paterna como alelo paterno.

Esto último es correcto para gran parte del genoma, menos para los cromosomas sexuales, conocidos como cromosomas X e Y. El cromosoma Y se presenta en general en los individuos de sexo biológico masculino, y es de herencia paterna. Por otro lado, los individuos masculinos suelen presentar un solo cromosoma X, heredado por vía materna, mientras que los femeninos presentan dos cromosomas X, uno heredado por vía materna y otro por vía paterna. Por otra parte, una pieza importante de información genética se conoce como el ADN mitocondrial. La mitocondria es una organela, y su información genética se encuentra por fuera del núcleo celular (a diferencia del resto del genoma). Dicha organela es heredada por línea materna, y por lo tanto su información genética también lo es.

En su totalidad, el genoma humano contiene alrededor de tres mil millones de pares de bases nucleotídicas (considerando ambas cadenas complementarias, en una célula haploide). La variación, es decir las diferencias intra especie, hallada dentro de los humanos ronda en torno al 0,1 %. Esto implica que el 99,9 % del genoma es compartido para todos los humanos (McVean Gil A. et al. 2012). Para la identificación de individuos, lo que añade a una problemática específica en Ciencias Forenses, el 0,1 % de diferencias será fundamental. Es así que la caracterización de regiones específicas del genoma humano, denominadas como hipervariables, por su alta diversidad en la población, ha impulsado implementaciones del análisis de ADN en Ciencias Forenses (Jeffreys et al. 1985a).

1.3.1 La huella digital genética

La huella digital genética es el nombre que acuñó la estrategia del uso del análisis de diferencias en el genoma humano para la identificación de individuos (Jeffreys et al. 1985b). Fue propuesta por Alec Jeffreys en 1984 y luego postulada para su utilización en el campo forense (Gill et al. 1985). La estrategia consistió (y consiste) en caracterizar ciertos loci hipervariables del genoma humano con el fin de identificar individuos. A estos loci se los conoce como *marcadores moleculares*, y a los valores que pueden tomar dichos marcadores moleculares se los

conoce como *alelos*. La metodología molecular a partir de la cual se estudian estas regiones hipervariables ha ido cambiando a lo largo del tiempo (Martin et al. 2001).

Desde una definición matemática se dirá que para un marcador llamado M_1 , su alfabeto estará dado por todos los posibles valores que M_1 pueda tomar. Esto último se obtiene de la población general, o de referencia (se hará una explicación más detallada de este término más abajo). Entonces, si $A_{M_1} = \{A, C, T, G\}$, quiere decir que el marcador M_1 puede tomar cualquiera de las cuatro letras de los nucleótidos del genoma humano. Esta expresión suele utilizarse para los marcadores que se conocen como *Polimorfismos de nucleótido único* (o SNPs por sus siglas en inglés). Como su nombre lo indica, refiere a posiciones específicas en el genoma, cuya variación se centra en un nucleótido. La identificación de dichos polimorfismos resulta muy informativa para la identificación de personas, pero su aplicación en el campo forense data de los últimos años solamente, debido a que requirió avances previos en el campo de la biología molecular y la bioinformática para su establecimiento como una herramienta accesible y certera (Ballard et al. 2020). Los marcadores conocidos como *repeticiones cortas en tandem* o STRs (por sus siglas en inglés) han tenido un uso generalizado dentro de la comunidad forense, debido en gran medida a su bajo costo y alto desempeño en las identificaciones (Butler 2007)(Weber and May 1989, Edwards et al. 1991). Los STRs son marcadores con alta variabilidad poblacional, que como su nombre lo indica, se constituyen por repeticiones de una serie de nucleótidos. Supóngase un marcador STR denominado M_2 , con $A_{M_2} = \{4, 5, 7, 9, 11, 14\}$, siendo el valor asignado la cantidad de repeticiones de la serie de nucleótidos. Por ejemplo, para una persona denominada J , se tendrá que $J_{M_2} = \{5, 7\}$. Esto implica que, para el marcador M_2 , el individuo J posee un alelo con cinco repeticiones de la serie de nucleótidos, y otro con siete repeticiones. La técnica molecular por la cual se identifica la cantidad de repeticiones se conoce como reacción en cadena de la polimerasa (PCR). Aunque no se ahondará en detalles, puede encontrarse vasta literatura asociada al tema. A modo de referencia está el libro de Alberts que abarca tópicos generales de biología molecular y celular (Alberts et al. 2003). Más relacionado al campo de la genética forense, el libro de Siegel et al. describe distintas técnicas moleculares empleadas en el campo (Siegel and Saukko 2012). A continuación se muestra un ejemplo para el marcador CSF1PO, cuya serie nucleotídica característica es AGAT.

CSF1PO

5' - AGAT AGAT AGAT AGAT - 3' Alelo 5
5' - AGAT AGAT AGAT AGAT AGAT - 3' Alelo 6
5' - AGAT AGAT AGAT AGAT AGAT AGAT AGAT - 3' Alelo 8

Figura 1.2 Secuencia genética para el marcador STR autosómico CSF1PO. Se indican distintos alelos caracterizados por la cantidad de repeticiones de la serie AGAT. La orientación 5'-3' refiere a la orientación de la secuencia de ADN.

La alta variabilidad presente en estos marcadores moleculares se debe, entre otros factores,

a que es una zona proclive a mutaciones. Las mutaciones son errores en el procedimiento de copiado del ADN durante la división celular, que se da en muchas células del organismo, y particularmente en las que forman la línea germinal. Es decir, aquellas células que darán origen a la progenie. A modo de ejemplo, las mutaciones podrían permitir que dos individuos, $J1_{M2} = \{7, 7\}$ y $J2_{M2} = \{7, 7\}$, tengan un descendiente $J3_{M2} = \{7, 8\}$. La aparición del alelo 8, no presente en los progenitores, se debe a una mutación. En los análisis forenses esto cobra una gran relevancia, dado que si el parentesco biológico de los padres está en duda, y no se considerase la posibilidad de mutaciones, se podría derivar en falsas conclusiones. Más adelante se discutirán distintos modelos matemáticos para contemplar las mutaciones (Egeland et al. 2015).

1.3.2 Evaluación de la evidencia genética

La evaluación de la evidencia genética en el área forense se realiza mediante lo que se conoce como el *test de parentesco genético*. Este tema será abordado en un capítulo específico de la tesis donde se definirá el problema en términos matemáticos. Además el tópico es ampliamente discutido en distintas referencias, como la del libro de Amorim et al. (Amorim and Budowle 2016). Aun así, en esta sección se introducen algunas ideas generales y antecedentes teóricos necesarios para entender los fundamentos del análisis de parentesco.

Como se explicó previamente, en la sección 1.2.1, el caso del dado desbalanceado, el teorema de Bayes constituye una herramienta muy importante para la inferencia estadística en casos donde la probabilidad de las hipótesis dados los datos es lo relevante. Y también, como se analizó en un marco forense, se suelen plantear al menos dos hipótesis para la explicación de los hechos. En el caso del *test de parentesco genético* lo que está en duda es la identidad biológica de un individuo, o bien de los restos de un individuo. Por ejemplo, la evidencia que se presenta es el ADN de los restos de un individuo, PNI , y el ADN de sus potenciales padres biológicos, que son individuos que buscan a una persona desaparecida, llamada PD . Las hipótesis que buscan explicar la evidencia generalmente son dos:

H_1 : PNI es PD.

H_2 : PNI no es PD y es una persona tomada al azar de la población de referencia, no relacionada biológicamente con PD.

Al total de la evidencia genética recolectada, tanto de PNI como de los parientes biológicos de PD , se la denomina G . A continuación se expresa la fórmula de bayes utilizada para el *test de parentesco genético*:

$$\frac{P(H_1|G)}{P(H_2|G)} = \frac{P(G|H_1) \cdot P(H_1)}{P(G|H_2) \cdot P(H_2)} \quad (1.18)$$

1.3.2.1 Herencia mendeliana

Para el caso particular del *test de parentesco genético*, las hipótesis están asociadas a dos modelos con fuerte respaldo dentro del campo de la genética. Las leyes que establecen la relación probabilística entre los genotipos de los parientes de PD , con PD , son estudiadas desde hace siglos, teniendo sus fundamentos sobre las Leyes de herencia de Mendel. A lo largo del tiempo

se han descubierto diferentes fenómenos y comportamientos de la herencia que se desvían de lo planteado por Mendel. Es así, que al conjunto de genes y locus que encuadran dentro de la teoría, se los denomina genes mendelianos. Éste es un tópico extenso, y hay múltiples fuentes bibliográficas que lo abordan, como el libro de Hamilton (Hamilton 2021) o el de Falconer (Falconer and Mackay 1983). Aún así, dentro del área de la genética forense, los marcadores moleculares analizados son en su gran mayoría de carácter mendeliano. En este sentido el libro de Amorim vuelve a ser una buena referencia, específica al campo forense (Amorim and Budowle 2016). Más adelante en la tesis se dedicará espacio a la definición y formalización matemática de las herencias genéticas.

1.3.2.2 Genética de poblaciones

Por otra parte, para H_2 , la proposición establece que PNI no es PD , y es una persona tomada al azar de la población de referencia. Las reglas probabilísticas que permiten calcular los genotipos de los individuos a partir de las poblaciones de referencia son parte del campo de estudio de la genética de poblaciones. Esta tiene sus fundamentos sobre la teoría planteada inicialmente por Sewall Wright, Haldane y Ronald Fisher (Hamilton 2021) (Falconer and Mackay 1983). Particularmente se hace uso del principio de Hardy-Weinberg (HW) para explicar las probabilidades de observar determinados genotipos basándose en las frecuencias de la población de referencia. A modo de introducción resumida al tema, se plantea el siguiente ejemplo: se supone que existe un marcador llamado $M4$, cuenta con 4 alelos en la población, siendo $A_{M4} = \{4, 5, 7, 11\}$. Entonces, se considera que se cumplen los supuestos del modelo HW: panmixia (reproducción aleatoria), ausencia de mutaciones, migraciones, deriva génica y con un número alto de individuos. Las probabilidades de observar, por ejemplo, un genotipo de un individuo denominado $J4 = \{7, 7\}$, será $p(M4 = 7)^2$, siendo $p(M4 = 7)$ la frecuencia relativa del alelo 7 en la población de referencia. Por otro lado, si existe un individuo $J4 = \{7, 8\}$, la probabilidad será $2p(M4 = 7)p(M4 = 8)$, siendo $p(M4 = 8)$ la frecuencia del alelo 8 en la población de referencia. Nótese que el vector que describe la identidad de los alelos no es ordenado, por lo tanto $J4 = \{7, 8\} = \{8, 7\}$, motivo que determina el factor 2 en el cálculo de la probabilidad del genotipo heterocigota. Más generalmente, si ambos alelos son iguales, característica que se conoce como homocigosis, la probabilidad estará dada por

$$p_i^2$$

siendo i el alelo del individuo, y p_i la frecuencia de dicho alelo en la población de referencia. Por otro lado, si los alelos son diferentes, característica que se conoce como heterocigosis, la probabilidad estará dada por

$$2p_i p_j$$

siendo i y j dos alelos diferentes, y p_i y p_j sus frecuencias relativas en la población de referencia.

1.4 Organización de la tesis

En este capítulo se introdujeron tres aspectos fundamentales para la tesis, por un lado la problemática de la búsqueda de personas desaparecidas, por otro lado la inferencia probabilística y su rol en la Ciencias Forenses, y por último un tipo de evidencia específica que es la evidencia genética. Además, se definió a la búsqueda de personas desaparecidas como un proceso que involucra múltiples pasos. Se hizo hincapié en el rol del científico forense en la formalización matemática y producción de modelos probabilísticos para la evaluación estadística de la evidencia. A lo largo de la tesis se irán recorriendo distintos estadíos en el proceso de búsqueda, desde la investigación preliminar hasta la toma de decisiones en la identificación, proveyendo aportes teóricos, modelos estadísticos y herramientas computacionales.

Se comienza en el capítulo 2 por la formalización de distintas líneas de evidencia recolectadas durante la investigación preliminar en la búsqueda de personas desaparecidas. Se propone un modelo de cociente de verosimilitud para distintas evidencias, y se termina con una propuesta para una probabilidad a priori producida a partir de la información obtenida en la investigación preliminar. Se analiza el uso de este a priori para el siguiente paso, el de la evaluación de la información genética.

En el capítulo 3 se aborda un modelo basado en redes bayesianas para el cómputo del cociente de verosimilitud en el test de parentesco genético. Esto implica, a partir de un árbol familiar, o pedigrí, definir la estructura y parámetros del modelo de red bayesiana. Se analizan además los distintos eventos que modifican los modelos generales de genética de poblaciones y de herencia genética.

El capítulo 4 se centra en la toma de decisiones en casos de búsqueda de personas desaparecidas mediante uso de bases de datos con información genética y no-genética. Se analiza la evaluación del poder estadístico en las búsquedas. Se propone un modelo para la selección racional de valores de corte para el posterior odds, tal que se obtenga un *set* de posibles candidatos a ser la persona desaparecida. Se muestra la utilidad del modelo particularmente en casos donde la evidencia genética no es suficiente para arribar a una conclusión. Al final del capítulo, se indica la mejora producida en términos de poder estadístico, de la incorporación del prior odds propuesto en el capítulo 2.

El capítulo 5 aborda, mediante dos metodologías diferentes, un problema de priorización de nueva evidencia a recolectar. Para aquellos casos donde la información disponible no es suficiente, se propone un modelo para el diagnóstico y posterior selección de líneas de evidencia más informativas. El primer enfoque busca resolver el problema empleando simulaciones computacionales, mientras que el segundo lo hace mediante la aplicación de métricas de teoría de la información.

Al final de la tesis se realiza una conclusión general, mostrando avances y desafíos pendientes en la formalización de distintos eventos que ocurren durante un proceso de búsqueda de personas desaparecidas.

Capítulo 2

El problema de la formalización del peso estadístico de la evidencia: el caso de los datos recolectados durante la investigación preliminar

En este capítulo se analizan distintas líneas de evidencias recolectadas durante un paso fundamental del proceso de búsqueda de personas desaparecidas, la investigación preliminar. A pesar de su relevancia han habido pocos intentos de formalizar matemáticamente la evidencia recolectada. Esto deriva en que la misma sea tenida en cuenta sólo de forma cualitativa, sin un procesamiento probabilístico adecuado. Se propone, en la presente tesis, una posible formalización matemática para la evaluación estadística de un conjunto de datos asociados a la investigación preliminar. Primordialmente se estudian líneas de evidencia donde se espera una concordancia entre el dato de la persona desaparecida y la persona no identificada. Es decir que se esperaría que el atributo asociado a la persona desaparecida, por ejemplo sexo biológico femenino, sea igual al de la persona o los restos humanos no identificados, por ejemplo huesos pertenecientes a una persona de sexo femenino, considerando siempre que la persona no identifica es efectivamente la persona desaparecida. Se toman ejemplos genéricos de tipos de evidencia, como el color de pelo, la edad y el sexo biológico. Estos son tres tipos diferentes de variables, por un lado el pelo es analizada como variable cualitativa politómica, el sexo como cualitativa dicotómica y la edad como variable continua. Se formaliza matemáticamente a cada una y se proponen modelos para el cálculo del cociente de verosimilitud. Se incorpora al modelo la posibilidad de errores en la interpretación o consolidación de la evidencia, por ejemplo errores de tipeo. Mediante simulaciones computacionales de escenarios de búsqueda de personas desaparecidas se muestra la utilidad de los modelos propuestos. Esta estrategia se compara con distintos usos frecuentes de la evidencia recolectada durante la investigación preliminar, como por ejemplo para el filtrado de la base de datos de personas no identificadas en base a alguna característica de la persona buscada. Se propone, mediante un enfoque bayesiano, que la evaluación estadística de la evidencia recolectada durante la investigación preliminar pueda utilizarse para constituir una probabilidad a priori para el paso de análisis de la evidencia genética. Este aspecto es discutido en el marco de la literatura forense actual, en la cual la definición de la probabilidad a priori para este tipo de casos ha sido un tema de discusión. Se describe su potencial uso en el contexto de las búsquedas de personas desaparecidas mediante bases de datos.

2.1 Introducción al capítulo

La investigación preliminar es uno de los primeros pasos del proceso de búsqueda de personas desaparecidas (Hackman 2016). Esta se encuentra fuertemente vinculada con los distintos pasos del proceso de búsqueda proveyendo información valiosa para la toma de decisiones. A modo de ejemplo, las averiguaciones acerca del árbol genealógico de una persona desaparecida son esenciales para el *test* de parentesco genético. Así también para el paso arqueológico, en el cual se pueden buscar e identificar familiares del desaparecido con el fin de ampliar la evidencia genética. En el proceso de identificación, la información recopilada durante la investigación preliminar es utilizada para evaluar la coherencia de los datos obtenidos y concluir en torno al caso. En un trabajo reciente, Salado Puerto et al. detallan los pasos del proceso de investigación en la búsqueda de personas desaparecidas (Puerto et al. 2021). Tomando esta referencia, en esta sección se detallan las distintas fuentes de información y líneas de evidencia recopiladas durante la investigación preliminar. Además, se discute acerca de distintos usos que se le han dado a los datos recopilados durante este paso en la búsqueda de personas desaparecidas. Por último, se hace foco en el problema de la definición de la probabilidad a priori para el *test* de parentesco genético, donde diferentes autores lo han vinculado con datos provenientes de la investigación preliminar (Budowle et al. 2011, Biedermann et al. 2012).

2.1.1 Información de la investigación preliminar

Como se ha explicado previamente en la introducción general, la búsqueda consiste en vincular dos entidades, la de la persona desaparecida, que es una identidad sin cuerpo, y la de la persona no identificada, que es un cuerpo sin identidad. Durante la investigación se recolecta una amplia gama de evidencias relacionadas a los casos de desaparición como al contexto en el que se da la misma (Puerto et al. 2021). Las fuentes de información de las que provienen dichas evidencias son múltiples, y la certeza o confianza en las mismas es variable.

A continuación se listan un conjunto de fuentes usuales:

- Fuentes escritas: incluye documentación personal, cartas, documentación de escolaridad y trabajo, reportes militares y policiales, reportes de autopsias, investigaciones judiciales, datos de organizaciones no gubernamentales, información de hospitales y morgues, registros de cementerios, antecedentes médicos, entre otras.
- Fuentes orales: implica entrevistas con testigos del hecho, familiares de la víctima, informantes, conocidos del ámbito laboral o escolar, personal médico que haya atendido a los implicados, entre otros.
- Fuentes audio-visuales: incluye fotografías, audios, imágenes satelitales, geolocalización de centros de detención, rutas migratorias, campos de refugiados, entre otras.
- Redes sociales y comunicación cibernética: este aspecto cobró mayor relevancia en los últimos años y constituye una fuente de información valiosa debido al hecho de que amplía la red social de los individuos implicados, permitiendo una referencia temporal de los eventos próximos al hecho de desaparición.

Estas fuentes de información serán de utilidad para obtener datos tanto de la persona desaparecida como de la persona no identificada. En este sentido se explican por separado las líneas de evidencia y la forma de recolección de las mismas.

2.1.1.1 Datos de la persona desaparecida

Muchos de los datos recopilados mediante las fuentes previamente descritas pueden ser de utilidad para generar hipótesis del posible paradero de la persona desaparecida, pero más aún, para obtener información acerca de su identidad. En definitiva la pregunta fundamental que se busca responder durante la etapa de la investigación preliminar es: *¿Quién es la persona desaparecida?* En este sentido se listan tres aspectos importantes:

- Documentación oficial: cualquier documentación nacional de identidad, pasaporte, registro de conducir, certificados de nacimiento, entre otros.
- Aspectos biológicos y físicos: perfil biológico, por ejemplo edad, peso y altura, características físicas, fracturas, rasgos específicos, contextura, etc.
- Historia social y estilo de vida: antecedentes profesionales, académicos y políticos, relaciones con potenciales testigos, apodos, actividades deportivas y recreativas, entre otras.

Mucha de esta información puede ser utilizada no sólo para comparar con los datos de las personas no identificadas, si no que además permite la implementación de distintas herramientas matemáticas. Un ejemplo es la construcción de redes matemáticas que puedan demostrar relaciones no-evidentes entre personas para colaborar en la generación de hipótesis acerca del destino de las personas (Caridi et al. 2011, 2020, Baraybar et al. 2020).

2.1.1.2 Datos de la persona no identificada

En este caso hay dos escenarios bien demarcados (Puerto et al. 2021). Por un lado, si la persona está viva y duda de su identidad biológica se tiene acceso a entrevistarla. Ejemplo de este tipo de casos es el tráfico de personas o el robo de niños producto de conflictos bélicos y violencia política (King 1991). Por otro lado, si la persona ha fallecido, por lo tanto la identificación se debe realizar sobre restos humanos, los primeros datos a recopilar se realizan *in situ*, es decir, en el lugar donde se lo encuentra. Para los casos en los cuales la persona está viva, los datos recolectados suelen ser los siguientes:

- Revisión de las circunstancias que derivan en que la persona no conozca su identidad biológica.
- Entrevista personal por parte de un equipo especializado en comunicarse con las víctimas.
- Examinación forense de la documentación: documentos de identidad nacional, partidas de nacimiento, registros de conducir, etc.
- Aspectos biológicos y físicos: perfil biológico, edad, peso, algunas características físicas generales, toma de huellas dactilares.

Esta información puede resultar útil en varios aspectos: por un lado, para un análisis de compatibilidad con los individuos desaparecidos buscados, y por otro para caracterizar el proceso que pudo haber generado la pérdida o robo de identidad biológica de la persona en cuestión.

Por otro lado, si se trata con restos humanos, la información a recopilar es la siguiente:

- Revisión de las circunstancias que derivan en que sean restos no identificados.
- Información de la recuperación de los restos, incluyendo la localización.
- Examinación forense, incluyendo análisis de rasgos físicos y biológicos.
- Información de las circunstancias y forma de muerte.

Nuevamente, los rasgos físicos y biológicos pueden ser de utilidad para evaluar la compatibilidad con la persona desaparecida. Cabe destacar que la estimación de la edad, altura, e inclusive de rasgos más simples, como el color de pelo, requieren de expertos y de metodologías no exentas de errores e incertezas (Cunha et al. 2009). En el resto de las líneas de evidencia, la incertidumbre puede provenir de fuentes poco confiables, o inclusive contradictorias (Puerto et al. 2021). Aún así, el uso de esta información ha resultado útil en la búsqueda de personas desaparecidas. A continuación se presentan ejemplos concretos.

2.1.2 Redes complejas para encontrar relaciones no evidentes

Caridi et al. (Caridi et al. 2011) introducen la utilización del enfoque de redes complejas para el análisis de datos de la investigación preliminar en la búsqueda de personas desaparecidas. Como se ha mencionado en la introducción general, las redes constituyen un dispositivo gráfico donde un grupo de nodos conectados interactúan de diferentes maneras. Los nodos representan variables y las conexiones indican relaciones entre las variables. En el caso de las redes bayesianas dichas relaciones corresponden a una dependencia condicional (Darwiche 2009). En otros casos, los conectores representan similitud entre las variables.

Específicamente, el trabajo de Caridi et al (Caridi et al. 2011) se centra en la aplicación del método de redes complejas a la búsqueda de desaparecidos durante la última dictadura militar en Argentina, acontecida entre 1976 y 1983 (Gorini 2006). En la red constituida, los nodos representan individuos, y los conectores representan relaciones explícitas o implícitas entre los individuos hasta cierto punto. Aquellos individuos que posean una fecha cercana de desaparición, o pertenezcan a un mismo grupo político estarán vinculados. La elección de distintas variables para generar el vínculo puede alterar la estructura de la red, por lo tanto es necesaria la aplicación de conocimiento experto en torno a los casos para investigar variables relevantes y datos confiables. Una de las ventajas que posee el enfoque es que utiliza herramientas matemáticas para predecir patrones no-evidentes y conexiones perdidas por falta de información dentro de la red (Barlow 2003). En casos como el de búsqueda de personas desaparecidas la falta de información confiable es un fenómeno frecuente. La metodología desarrollada pudo ser aplicada a distintos contextos, como en el de desapariciones en casos de migraciones (Baraybar et al. 2020).

2.1.3 Aprendiendo de los casos resueltos

En un trabajo más reciente, Caridi et al. (Caridi et al. 2020), presentan un método para asistir al proceso de identificación en la búsqueda de personas desaparecidas. El mismo propone, mediante un enfoque bayesiano, el uso de variables no-genéticas recopiladas en la investigación preliminar para producir hipótesis de identificación. Es decir, dado un determinado caso, utilizando datos no-genéticos, producir un *ranking* de personas no identificadas, en función de la probabilidad de ser la persona desaparecida. Para la selección de las variables informativas y parametrización de las probabilidades, el enfoque propone utilizar los casos resueltos de un evento masivo, como un desastre natural o la desaparición forzada de un conjunto de personas, para encontrar patrones. Sitúa un ejemplo de búsqueda durante la última dictadura cívico-militar en Argentina (Gorini 2006). En la misma se analizan distintas variables, que son formalizadas matemáticamente y se estudia su capacidad predictiva para identificar una serie de casos resueltos, que son parte del *set* de datos de entrenamiento y validación. Cabe destacar que el enfoque de aprender de los casos resueltos puede arrojar pistas de propiedades del mecanismo de desaparición.

2.1.4 La probabilidad a priori y la investigación preliminar

En casos de búsqueda de personas desaparecidas e identificación de víctimas de catástrofes, la evidencia genética ha jugado un rol central. La evaluación estadística de la misma se realiza mediante la aplicación del teorema de Bayes (ecuación 1.6). La determinación del a priori para el paso genético en estos casos ha estado vinculada al número de víctimas (Prieto et al. 2022). En esta sección se introducen las expresiones comúnmente utilizadas dentro de la comunidad forense para la determinación de la probabilidad a priori. Ejemplos de estas son las aplicaciones de probabilidades a priori en los ejercicios llevadas a cabo por el Grupo de Habla Española y Portuguesa de la *International Society for Forensic Genetics* (*GHEP-ISFG*) en los ejercicios propuestos para los laboratorios forenses (Vullo et al. 2021, 2016). Se pueden encontrar también otros ejemplos de aplicación en la literatura forense (Brenner and Weir 2003, Zupanič Pajnič et al. 2010) y dentro de las recomendaciones de la *International Society for Forensic Genetics* (*ISFG*) Prinz et al. (2007). A continuación, se introduce mediante un ejemplo concreto, la metodología empleada para la determinación del a priori:

$$Prior = \frac{1}{N + 1} \quad (2.1)$$

Siendo N el número de víctimas. A modo de ejemplo, si en un caso en el cual se busca identificar huesos en una fosa común, y a partir de la reconstrucción de los mismos se determina que la cantidad de individuos a la que pertenecen es 125, la probabilidad a priori será:

$$Prior = \frac{1}{125 + 1} = 0,008$$

Budowle et al. (Budowle et al. 2011) analizan un caso en el cual distintas características físicas y rasgos biológicos pueden ser utilizadas para redefinir el grupo de víctimas. A modo de ejemplo, si de aquellas 125 víctimas, 4 tienen el color de pelo colorado, y la persona desaparecida buscada

posee también dicho color, la probabilidad a priori debería ser:

$$Prior = \frac{1}{4+1} = 0,20$$

En el mismo artículo se aclara que es necesario el establecimiento de protocolos claros y trazables para determinar dichas probabilidades a priori y combinar la información de la investigación preliminar con los datos genéticos. Uno de los motivos por los cuales no se profundiza en la formalización de los prior es porque algunos autores han sugerido que expresan el punto de vista subjetivo de quien analiza la evidencia y, por lo tanto, el establecimiento de reglas resulta complejo (Biedermann et al. 2012).

Por otra parte, dentro la comunidad forense se estableció que la probabilidad a priori es utilizada como análisis de sensibilidad (Amorim and Budowle 2016). Es decir, se toman un conjunto de valores posibles, se calcula el cociente de verosimilitud a partir de la evidencia genética y se obtienen un conjunto de probabilidades a posteriori. El resultado es concluyente cuando el peso de la evidencia genética es suficientemente grande como para obtener una probabilidad a posteriori también concluyente más allá del a priori seleccionado. Esto se refleja en la frase comúnmente aplicada: *la evidencia es concluyente cuando está por encima de toda sospecha lógica*. La sospecha ocupa el lugar de la probabilidad a priori (Amorim and Budowle 2016). Por este motivo es común encontrar tablas donde se evalúa, en cada fila, un prior odds diferente, que al ser multiplicado por el cociente de verosimilitud genético deriva en un posterior odds. Si todos los valores de posterior odds de la tabla, incluídos aquellos con muy bajos prior odds, superan un determinado umbral, se considera que la evidencia es suficientemente concluyente.

En este capítulo se formaliza un modelo de cociente de verosimilitud para la evidencia no-genética. A diferencia de lo planteado por Budowle et al. (Budowle et al. 2011) se considera que los datos de la investigación preliminar deben pasar por un paso de interpretación para luego, mediante el procedimiento bayesiano, poder convertirse en probabilidades a priori del paso genético. Se propone que se debe explicitar y formalizar un procedimiento para la asimilación de distintas líneas de evidencia, y que el mismo debe contemplar la formalización del cálculo de las probabilidades a priori en el paso genético. Por otro lado, como se verá más adelante, la asimilación de la evidencia recopilada durante la investigación preliminar también requiere de una probabilidad a priori. Es en este punto donde el Juez, o quien analice la totalidad de la evidencia reunida, puede colocar su punto de vista subjetivo en formato de probabilidad. A priori no significa necesariamente previo a, si no que es parte de la evidencia recolectada por fuera del análisis genético. Por último, respecto al rol central de la evidencia genética, es necesario considerar que la misma puede no ser suficiente para llegar a una conclusión, o bien, puede ser suficiente pero estar sujeta a procedimientos o errores que pueden derivar en pérdidas de identificaciones (Marsico et al. 2021).

2.2 Métodos

2.2.1 Datos de la investigación preliminar

Se define a los datos recolectados para PD como DPD (Datos de la Persona desaparecida), y aquellos recolectados para la PNI , como $DPNI$ (Datos de la Persona No Identificada). Además, se recolecta información relacionada al conjunto de casos analizados, cuando exista una vinculación entre los mismos. Por ejemplo, en caso de tratarse la identificación de un conjunto de restos humanos en una fosa común, se intentará constituir un número de total de víctimas presentes, o cantidad de $PNIs$. A este valor se lo define como N . Por otra parte, la cantidad de PDs es definida como K . De este modo se constituyen dos listados, uno para PDs , tal que $PDs = \{PD_1, PD_2, PD_3, \dots, PD_K\}$, y otro para $PNIs$, tal que $PNIs = \{PNI_1, PNI_2, PNI_3, \dots, PNI_N\}$. Puede ocurrir que $K \neq N$. Inclusive, puede suceder que estos valores no se conozcan con precisión. A su vez, cada PD contará con DPD , tal que $DPD_i = \{E_1, E_2, E_3, \dots, E_X\}$, siendo i un PD específico, y X la cantidad de líneas de evidencia recolectadas. Por otro parte, cada PNI tendrá $DPNI$ asociada, tal que $DPNI_j = \{E_1, E_2, E_3, \dots, E_X\}$, siendo j un PNI específico. Es importante tener en cuenta que en estos casos E_1 será una misma linea de evidencia tanto para PD como para PNI . Dicho de otro modo, de todas las líneas de evidencia disponible, se analizan aquellas que son comparables. Particularmente, se hace hincapié en aquellas que describan rasgos biológicos y físicos del individuo. A modo ilustrativo, para un determinado PNI , j , se tiene $DPNI_j = \{S, E\}$, siendo S el sexo biológico de PNI , y E su edad. Para el mismo caso, se cuenta con un PD i , tal que $PD_i = \{S, E\}$, siendo S y E el mismo tipo de evidencia descrita para PNI . Esto permitirá comparar, por ejemplo, el sexo biológico de PD_i con el de PNI_j .

A continuación se presentan los modelos para el cálculo de verosimilitud de tres tipos de variables.

- Sexo biológico, S , como variable categórica dicotómica. Puede tomar dos valores, tal que $A_S = \{m, f\}$ siendo m masculino y f femenino.
- Color de pelo, C , como variable categórica politómica. Puede tomar múltiples valores, tal que $A_C = \{1, 2, 3, 4, 5\}$, donde cada valor representa un color diferente: 1 castaño, 2 negro, 3 rubio, 4 blanco y 5 colorado.
- Edad, E , variable continua. Puede tomar valores enteros desde 0 hasta 100 tal que $A_E \in [0, 100]$.

Estas variables fueron seleccionadas a modo de ejemplo, cubriendo tres tipo diferentes de problemas para el planteo del modelo de verosimilitud. A continuación se presentan los modelos desarrollados para cada una.

2.2.2 Modelo de verosimilitud para variables cualitativas dicotómicas

Este tipo de variables se ejemplifican con el sexo biológico, S . Este puede ser determinado tanto para PNI (por análisis antropológicos o genéticos) como para PD (por testimonio de

conocidos). Sus respectivas probabilidades $P(s) = \{p_m, p_f\}$, suman 1. Al sexo de PNI se lo llama, S_{PNI} , y al de PD , S_{PD} . Las probabilidades p_m y p_f se obtienen a partir de la frecuencia relativa de m (masculino) y f (femenino) en la población de referencia. La población de referencia refiere al grupo o población en la cual se encuentran los individuos analizados, por ejemplo, podría indicar un país o un continente. El cociente de verosimilitud (CV), se expresa a continuación:

$$CV_S = \frac{P(S_{PNI}|S_{PD}, H_1)}{P(S_{PNI}|S_{PD}, H_2)} \quad (2.2)$$

Teniendo dos variables, S_{PD} y S_{PNI} , con dos valores (m y f), quedan cuatro posibles combinaciones. A continuación se presentan las expresiones y probabilidades para cada combinación:

$$CV_S = \frac{P(S_{PNI} = f|S_{PD} = f, H_1)}{P(S_{PNI} = f|S_{PD} = f, H_2)} = \frac{1 - \epsilon_S}{p(f)}$$

$$CV_S = \frac{P(S_{PNI} = m|S_{PD} = m, H_1)}{P(S_{PNI} = m|S_{PD} = m, H_2)} = \frac{1 - \epsilon_S}{p(m)}$$

$$CV_S = \frac{P(S_{PNI} = f|S_{PD} = m, H_1)}{P(S_{PNI} = f|S_{PD} = m, H_2)} = \frac{\epsilon_S}{p(f)}$$

$$CV_S = \frac{P(S_{PNI} = m|S_{PD} = f, H_1)}{P(S_{PNI} = m|S_{PD} = f, H_2)} = \frac{\epsilon_S}{p(m)}$$

Nótese que cuando $S_{PNI} = S_{PD}$, la $P(S_{PNI}|S_{PD}, H_1) = 1 - \epsilon_S$. Donde $\epsilon_S \in [0 : 1]$ es un parámetro que determina posibilidad de errores de tipo en la base de datos, en la determinación de S_{PNI} o de S_{PD} . En otras palabras, bajo H_1 el sexo biológico de PD y PNI debería ser el mismo, a menos que haya un error instrumental o humano en la determinación. Generalmente ϵ tendrá un valor bajo ($<0,05$). Se admite entonces, bajo H_1 , que $S_{PNI} \neq S_{PD}$, con una probabilidad ϵ_S . Por otra parte, bajo H_2 , como indica su enunciado H_2 : *PNI es un individuo tomado al azar de la población de referencia*, su probabilidad estará dada por la población de referencia, $P(S_{PNI}|S_{PD}, H_2) = p(S_{PNI})$.

De manera sintética, el modelo se describe a continuación:

$$CV_S = \frac{P(S_{PNI}|S_{MP}, H_1)}{P(S_{PNI}|S_{PD}, H_2)} = \begin{cases} \frac{1 - \epsilon_S}{P(S_{PNI})} & \text{si } S_{PNI} = S_{PD} \\ \frac{\epsilon_S}{P(S_{PNI})} & \text{si } S_{PNI} \neq S_{PD} \end{cases} \quad (2.3)$$

2.2.3 Modelo de verosimilitud para variables cualitativas politómicas

El color de pelo, C , hace referencia a un rasgo que es definido por fotos o testimonios para el PD y por inspección visual del PNI . Las probabilidades $P(C) = \{p_1, p_2, p_3, p_4, p_5\}$ suman 1 y se obtienen de la frecuencia relativa en la población de referencia. Se define el color de pelo de PD , como C_{PD} , y el de PNI como C_{PNI} . La ecuación de CV para esta variable es presentada a continuación:

$$CV_C = \frac{P(C_{PNI}|C_{PD}, H_1)}{P(C_{PNI}|C_{PD}, H_2)} \quad (2.4)$$

En este caso puede suceder que ϵ_C toma diferentes valores para distintas combinaciones de colores debido al grado de semejanza entre los mismos. Por ejemplo, en una inspección visual de restos óseos en mal estado de conservación es más factible confundir el color negro con el castaño que confundir el negro con el blanco. Esto da lugar a definir un ϵ_{ij} para cada par, siendo i el C_{PNI} y j el C_{PD} . En este caso se asume que $\epsilon_{ij} = \epsilon_{ji}$ para cada par ij .

		PNI				
		C_1	C_2	C_3	C_4	C_5
PD	C_1	λ_1	$\lambda_1 \cdot \epsilon_{12}$	$\lambda_1 \cdot \epsilon_{13}$	$\lambda_1 \cdot \epsilon_{14}$	$\lambda_1 \cdot \epsilon_{15}$
	C_2	$\lambda_2 \cdot \epsilon_{21}$	λ_2	$\lambda_2 \cdot \epsilon_{23}$	$\lambda_2 \cdot \epsilon_{24}$	$\lambda_2 \cdot \epsilon_{25}$
	C_3	$\lambda_3 \cdot \epsilon_{31}$	$\lambda_3 \cdot \epsilon_{32}$	λ_3	$\lambda_3 \cdot \epsilon_{34}$	$\lambda_3 \cdot \epsilon_{35}$
	C_4	$\lambda_4 \cdot \epsilon_{41}$	$\lambda_4 \cdot \epsilon_{42}$	$\lambda_4 \cdot \epsilon_{43}$	λ_4	$\lambda_4 \cdot \epsilon_{45}$
	C_5	$\lambda_5 \cdot \epsilon_{51}$	$\lambda_5 \cdot \epsilon_{52}$	$\lambda_5 \cdot \epsilon_{53}$	$\lambda_5 \cdot \epsilon_{54}$	λ_5

La variable λ_i se conoce como constante de normalización y es introducida con el fin de que todas las probabilidades para un determinado C_{PNI} sumen 1. A modo de ejemplo, si se tiene PNI cuyo $C_{PNI} = 1$, las probabilidades marginalizadas sobre PD deben sumar 1. Por lo tanto, habiéndose definido todos los valores de ϵ_{1j} , se puede obtener λ_1 a partir de la siguiente expresión:

$$\lambda_1 = \frac{1}{(1 + \epsilon_{12} + \epsilon_{13} + \epsilon_{14} + \epsilon_{15})} \quad (2.5)$$

De manera simplificada, se define la ecuación del cálculo de CV para la variable C a partir de la siguiente expresión:

$$CV_C = \frac{P(C_{PNI}|C_{PD}, H_1)}{P(C_{PNI}|C_{PD}, H_2)} = \begin{cases} \frac{\lambda_i}{P(C_{PNI})} & \text{si } C_{PNI} = C_{PD} \\ \frac{\lambda_i \epsilon_{ij}}{P(C_{PNI})} & \text{si } C_{PNI} \neq C_{PD} \end{cases} \quad (2.6)$$

2.2.4 Modelo de verosimilitud para variables continuas

La edad, E , puede ser también determinada tanto para el PNI como para el PD . En el primer caso, esta se estima a partir de los restos óseos implementando métodos de la antropología forense (Cunha et al. 2009, Schmitt et al. 2010). En el caso del PD suele conocerse a partir de testimonios de familiares de la víctima o bien a partir de documentación.

La estrategia para trabajar la variable continua A propuesta se basa en el abordaje conservador planteado por Crow et al. (Crow et al. 1996). El mismo se denomina *floating bin* y consiste en definir rangos equiprobables en los cuales se estima, con una determinada confianza, que se encuentra el valor de E . Por ejemplo, para E_{PNI} se definen los valores PNI_{min} y PNI_{max} tal que $PNI_{min} < E_{PNI} < PNI_{max}$. Del mismo modo se define un rango para A_{PD} tal que $PD_{min} < A_{PD} < PD_{max}$, siendo min el valor mínimo del rango de edad, y max el valor máximo. Una primera aproximación es declarar un *match directo* cuando exista

solapamiento entre los rangos. Dicho de otro modo, si se cumple que $PNI_{max} > PD_{min}$ y $PNI_{min} < PD_{max}$, se estará frente a un solapamiento de los rangos y un *match* será declarado. Con el fin de simplificar la notación, se define a esta condición como una variable booleana, M_E . Cuando $M_E = 1$, se cumple que $PNI_{max} > PD_{min}$ y $PNI_{min} < PD_{max}$, en cambio con $M_E = 0$ no se cumplen estas condiciones. Coloquialmente, se dirá que si $M_E = 1$, la ventana (o una porción de la misma) de valores estimados para E_{PNI} recae dentro de la ventana de valores estimados para E_{PD} . Utilizando la ecuación 1, se propone la siguiente formulación para el CV basado en la edad:

$$CV_E = \frac{P(M_E|E_{PNI}, E_{PD}, H_1)}{P(M_E|E_{PNI}, E_{PD}, H_2)} \quad (2.7)$$

Es decir, se contrasta la probabilidad de *match*, M , dados los datos E_{PNI} y E_{PD} bajo H_1 o H_2 . Teniendo una variable booleana M se definen las expresiones para sus dos posibles valores:

$$CV_E = \frac{P(M_E = 1|E_{PNI}, E_{PD}, H_1)}{P(M_E = 1|A_{PNI}, E_{PD}, H_2)} = \frac{1 - \epsilon_E}{p(M_E = 1)}$$

$$CV_E = \frac{P(M_E = 0|E_{PNI}, E_{PD}, H_1)}{P(M_E = 0|E_{PNI}, E_{PD}, H_2)} = \frac{\epsilon_E}{p(M_E = 0)}$$

$P(M_E = 1)$ y $P(M_E = 0)$ se obtienen a partir de la población de referencia, y estarán dados por la frecuencia de individuos cuya edad se encuentra dentro de la ventana de valores de E_{PD} (por lo tanto $M_E = 1$) y los que no se encuentran dentro (o sea $M_E = 0$). A continuación se define la expresión abreviada:

$$CV_E = \frac{P(M_E|E_{PNI}, E_{PD}, H_1)}{P(M_E|E_{PNI}, E_{PD}, H_2)} = \begin{cases} \frac{1 - \epsilon_E}{p(M_E = 1)} & \text{si } M_E = 1 \\ \frac{\epsilon_E}{p(M_E = 0)} & \text{si } M_E = 0 \end{cases} \quad (2.8)$$

Nótese que para esta variable hay dos fuentes de error: (i) la primera es debido a la incertezza en la estimación de la edad, en un caso relacionada a la metodología antropológica y en el otro a la precisión de los testimonios. Esta se expresa en el intervalo asignado para E_{PNI} y E_{PD} ; (ii) La segunda se encuentra asociada a errores de tipeo en la base de datos, testimonios equivocados (haciendo por ejemplo alusión a una persona que no es la PD) entre otros. Este último error se explicita en el valor de ϵ_E .

Respecto al error (i), como se ha mencionado, depende de la metodología empleada, habiendo muchas opciones disponibles (Cunha et al. 2009, Schmitt et al. 2010). La mayoría de estos modelos de estimación de la edad a partir de los restos otorgan la incertezza en términos de rangos. Con el fin de simplificar y dar espacio al investigador forense para plasmar la información de la incertidumbre, se opta por dejar las variables PNI_{min} y PNI_{max} disponibles de ser definidas.

2.2.5 Cálculo de prior odds

Como se mencionó previamente, en el enfoque bayesiano, la probabilidad a posteriori de una instancia del conocimiento puede ser la probabilidad a priori de otra instancia de conocimiento. En este caso, se propone utilizar la probabilidad a posteriori del paso no-genético, basado

en los datos de la investigación preliminar, como la probabilidad a priori del paso genético. Primeramente, el posterior no-genético es calculado aplicando la siguiente ecuación:

$$O(H | NG) = CV_{NG} \cdot O_{NG}(H) \quad (2.9)$$

El posterior, $O(H | NG)$, es obtenido mediante la multiplicación del CV_{NG} con el prior odds, $O_{NG}(H)$. El CV_{NG} es el cociente de verosimilitud no genético. Tomando de ejemplo las variables enunciadas, y considerando que son independientes, la expresión para CV_{NG} queda de la siguiente manera:

$$CV_{NG} = CV_S \cdot CV_C \cdot CVE \quad (2.10)$$

La definición de $O_{NG}(H)$ depende del caso, y múltiples opciones han sido discutidas (Biedermann et al. 2012). En este ejemplo se consideran distintos escenarios, utilizando priors uniformes. Uniformidad implica que habrá una misma probabilidad para cada par $PNI - PD$. Es importante considerar que este punto es donde quien analiza la evidencia en torno a los casos puede plasmar su subjetividad en términos de probabilidades. A partir de este punto, los siguientes pasos de actualización del conocimiento requerirán el procesamiento de nueva evidencia.

Una vez obtenido, el posterior $O(H | NG)$ puede volverse el prior del paso genético considerando dos puntos: por un lado las hipótesis a contrastar deben ser las mismas, y por el otro la evidencia del paso no-genético debe ser independiente de la recolectada en el paso genético. Ambos se cumplen para las variables analizadas como ejemplo y los marcadores genéticos utilizados habitualmente (Egeland et al. 2015). Para diferenciarlo del prior del paso no-genético, se llamará al prior del paso genético (basado en datos no genéticos) como $O_G(H)$.

En este caso, el prior dependerá de cada par PNI-PD, volviéndose no-uniforme. La ecuación presentada a continuación define el paso genético:

$$O(H | G) = CV_G \cdot O_G(H) \quad (2.11)$$

En el siguiente capítulo se ahondará en el cálculo del cociente de verosimilitud de los datos genéticos, aquí representado como CV_G .

2.2.6 La búsqueda en bases de datos

La búsqueda en bases de datos implica comparar la evidencia proveniente de diferentes fuentes. Por un lado, los datos de la investigación preliminar se alojan en la base DPD (para las personas desaparecidas), y $DPNI$ (para las personas no identificadas). Para cada par $PD - PNI$ se compara DPD y $DPNI$. En el ejemplo planteado ambos conjuntos de datos albergan información de las variables S , E y C . Como se introdujo previamente, el resultado de la comparación será un $O_{NG}(H)$ para cada par $PD - PNI$. Posteriormente, se comparan los datos genéticos. En este punto es importante remarcar que se obtiene un CV_G para cada par $PD - PNI$. De esta manera, el posterior odds del paso genético es obtenido multiplicando $O_{NG}(H)$ por CV_G .

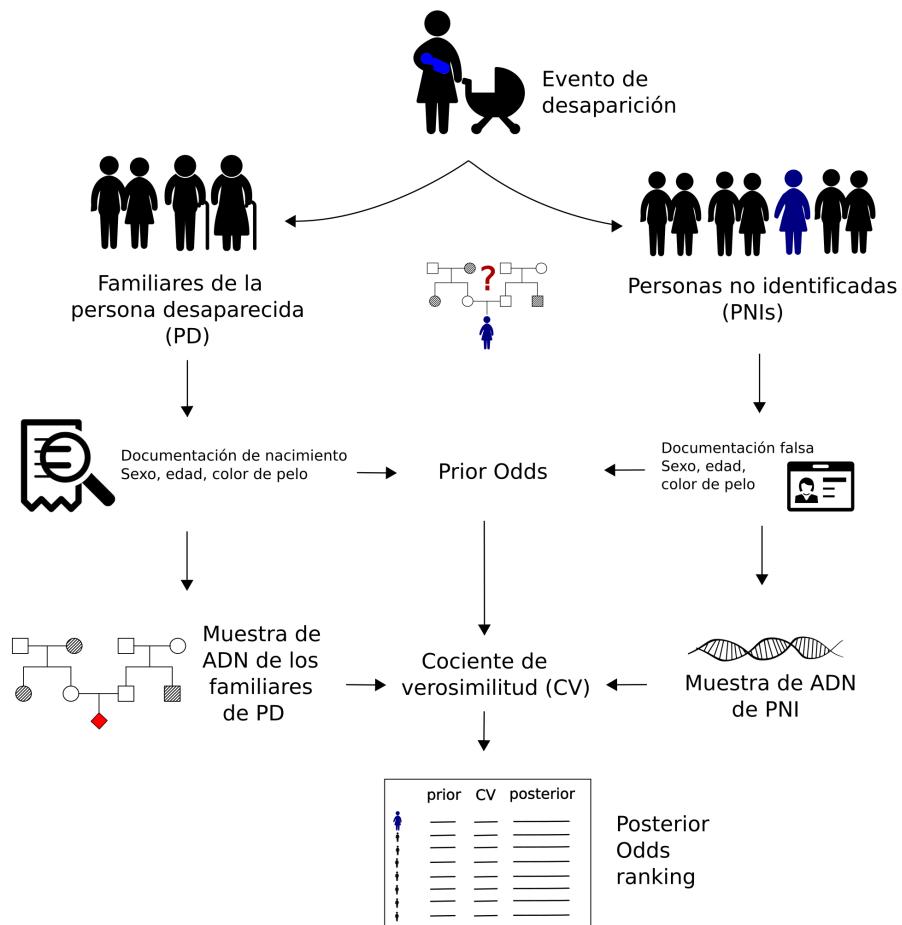


Figura 2.1 Esquema general de la búsqueda en bases de datos. Se muestra un evento de desaparición de una niña, a partir del cual la misma pasa a formar parte de un conjunto de personas no identificadas. Por otro lado, los familiares buscan a la niña desaparecida. A partir del primer paso, referido como investigación preliminar, se analizan datos que constituirán el prior odds. El segundo paso, el del test de parentesco genético permite el cálculo del cociente de verosimilitud. El último paso implica tomar la decisión en función del posterior odds obtenido para cada persona no identificada analizada.

2.2.7 Evaluación del poder estadístico de la evidencia

En esta sección se introduce un enfoque propuesto para evaluar el poder estadístico de la evidencia. El poder estadístico se refiere a la probabilidad de llegar a una conclusión correcta en función de la evidencia recolectada. Métricas para evaluar el poder estadístico en casos forenses de búsqueda de personas desaparecidas han sido previamente introducidas Kling et al. (2017). Los autores evalúan el poder estadístico de la evidencia genética, es decir, CV_G . En este caso se utiliza un enfoque similar para evaluar el poder estadístico de CV_{NG} . Este tema será abordado en detalle en el capítulo 4, donde se estudia el rol de estas métricas en la toma de decisiones.

2.2.7.1 Tasas de error

En la búsqueda en la base de datos se declara una potencial identificación cuando la CV supera un determinado umbral, el cual se denomina T . El problema puede plantearse como uno

de clasificación binaria. En este caso solo se considera el umbral T relacionado a CV_{NG} , por lo tanto se denomina T_{NG} . Si el resultado es considerado una potencial identificación, cuando PNI es PD (o sea H_1 es cierta), se obtiene un verdadero positivo (VP). En cambio, si el valor que se obtiene es menor al umbral, T_{NG} , considerando H_1 , se obtiene un falso negativo (FN).

En el caso contrario, se puede obtener un $CV_{NG} > T_{NG}$ cuando PNI no es PD (H_2 cierta). En este caso el resultado es un falso positivo (FP). Si sucede lo contrario, $CV_{NG} < T_{NG}$ con H_2 cierta, el resultado es un verdadero negativo (VN).

De esta manera, la tasa de falsos positivos (TFP) es la probabilidad de obtener $CV_{NG} > T_{NG}$, dado que H_2 es cierta, y se puede calcular con la siguiente expresión:

$$TFP = P(CV_{NG} > T_{NG}|H_2) \quad (2.12)$$

La tasa de verdadero negativos (TVN) es la probabilidad de obtener $CV_{NG} < T_{NG}$, dado que H_2 es cierta, y está relacionada con TFP de la siguiente manera:

$$TVN = 1 - TFP \quad (2.13)$$

La tasa de falsos negativos (TFN) es la probabilidad $CV_{NG} < T_{NG}$, cuando H_1 es cierta:

$$TFN = P(CV_{NG} < T_{NG}|H_1) \quad (2.14)$$

La tasa de verdaderos positivos (TVP) es la probabilidad de obtener $CV_{NG} > T_{NG}$, dado que H_1 es cierta, y se relaciona con TFN de la siguiente manera:

$$TVP = 1 - TFN \quad (2.15)$$

2.2.7.2 Medidas de rendimiento

Tomando de base a las definiciones previas, se presentan tres métricas que permitirán evaluar distintos aspectos del sistema de clasificación. Primero, el coeficiente de correlación de Matthews (MCC) (Chicco and Jurman 2020). El mismo toma valores de -1 a 1 , siendo sus valores extremos la clasificación completamente incorrecta (-1) o completamente correcta (1). Cuando $MCC=0$ implica que la clasificación es igual a un clasificador *random*.

$$MCC = \frac{VP \cdot VN - FP \cdot FN}{\sqrt{(VP + FP) \cdot (VP + FN) \cdot (VN + FP) \cdot (VN + FN)}} \quad (2.16)$$

La precisión es la proporción de los VP sobre el total de clasificados como positivos:

$$Precision = \frac{VP}{VP + FP} \quad (2.17)$$

Recall, es la fracción de VP sobre el total de verdaderos:

$$Recall = \frac{VP}{VP + FN} \quad (2.18)$$

2.2.8 Simulaciones computacionales

Un enfoque basado en simulaciones ha sido propuesto para explorar la distribución de CV_G (cociente de verosimilitud genético) en el contexto del *test* de parentesco genético en caso de búsqueda de personas desaparecidas (Egeland et al. 2016). El mismo evalúa específicamente la evidencia recolectada con datos genéticos CV_G . En este caso, se propone utilizar un enfoque similar para CV_{NG} por dos motivos: (i) el contexto de aplicación es el mismo, y por lo tanto podrá ser comparable con análisis hechos sobre datos genéticos, y (ii) permitirá evaluar la evidencia de la investigación preliminar por un lado, la evidencia genética por otro, y el resultado de combinar ambas evidencias. El pseudocódigo presentado a continuación muestra los pasos del algoritmo de simulación de evidencia.

Algoritmo 1 Simulaciones de evidencia

```

for  $j$  in  $(PD_1, PD_2, \dots, PD_K)$  do
    Simular  $N$  PNIs
    for  $i$  en  $(PNI_1, PNI_2, PNI_3, \dots, PNI_N)$  do
        Muestrear  $DPNI$  considerando  $P(DPNI|H)$ 
    end for
end for
```

Para cada PD_j en una base de datos con K PDs, se simulan N PNIs considerando H_1 o H_2 como ciertas. A modo de ejemplo, se describe cómo simular la variable S . En un caso hipotético, donde PD es femenina, por lo tanto $S_{PD} = f$, se considera un error $\epsilon_S = 0,05$. La simulación se realiza para obtener 20.000 PNIs. La salida de la simulación esta compuesta por dos listas: (i) positivos: 10.000 PNIs con sexo biológico adjudicado considerando H_1 verdadera, y (ii) negativos: 10.000 PNIs simulados considerando H_2 verdadera. En el primero, la proporción de f , con un margen de error debido a desviaciones por la incerteza del muestreo, debería ser en torno al 0,95. En la segunda lista, cuando H_2 es cierta, la proporción de f debería ser $P(f)$ (la frecuencia relativa de la población de referencia).

2.3 Resultados

En este apartado se presentan los resultados del planteo metodológico de la formalización de distintas líneas de evidencia recolectadas durante la investigación preliminar. Además, se analizan ejemplos puntuales para evaluar la contribución de la evidencia al proceso de identificación.

2.3.1 Variable categórica dicotómica: el sexo biológico

Para estudiar el comportamiento de la variable S , se considera un caso donde la persona desaparecida es de sexo femenino, es decir, $S_{PD} = f$. Tomando la ecuación 2.3 puede entenderse que el valor de CV_S dependerá de tres factores: (i) coincidencia de S_{PNI} y S_{PD} . En este caso, si $S_{PNI} = f$, habrá coincidencia, en cambio si $S_{PNI} = m$, no la habrá. (ii) $P(S_{PNI})$, es decir, la proporción del sexo del PNI comparado en la población de referencia. (iii) el valor de error, ϵ_S .

En la Figura 2.2 puede verse cómo el valor del $\text{Log}_{10}(CV_S)$ varía, tanto para un PNI masculino como para uno femenino, en función de $P(S = f)$ en la población de referencia. En el ejemplo se considera $\epsilon_S = 0,001$. A medida que $P(S = f)$ aumenta, la diferencia entre $\text{Log}_{10}(CV_S)$ obtenido para un PNI masculino y uno femenino disminuye (considereando siempre $SPD = f$). El caso extremo sucede cuando $P(S = f)$ se aproxima a 1, donde el CV_S para ambos sexos de PNI es igual. Esto puede resultar intuitivo, teniendo en cuenta que si PD es femenino, y se halla PNI también femenino, en una población con baja frecuencia de individuos femeninos, el peso de la evidencia será mayor que en una población con alta frecuencia de femeninos. En el caso extremo, si en la población de referencia todos los individuos fueran femeninos, el peso de la evidencia sería nulo.

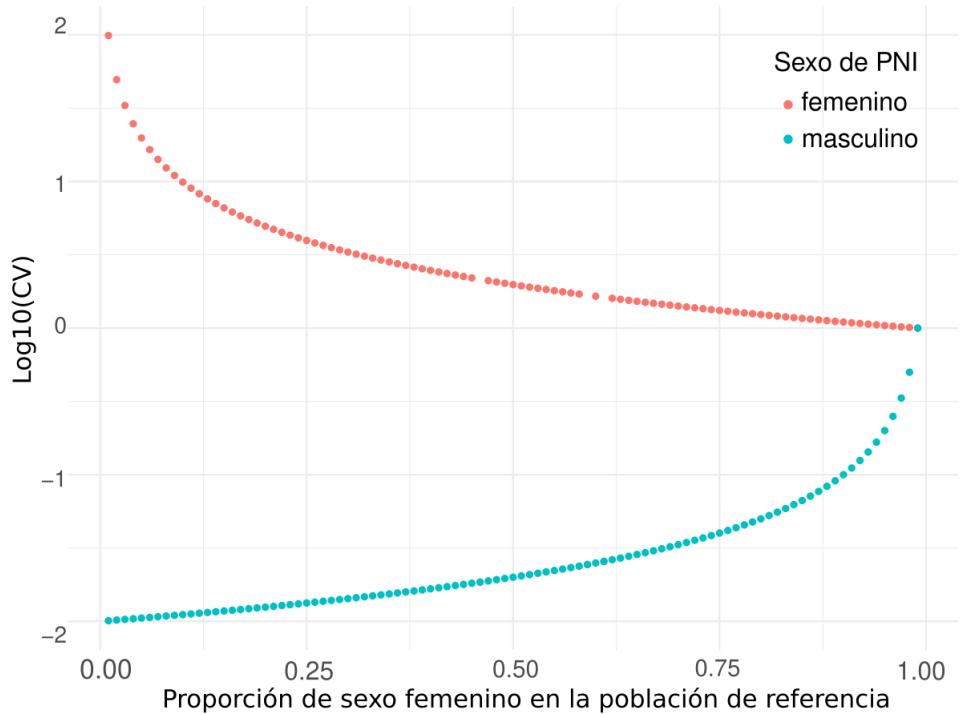


Figura 2.2 Valores de $\text{Log}_{10}(CV_S)$ para cada sexo de PNI, considerando un MP femenino, en función de la proporción de femeninos en la población de referencia.

En la Figura 2.3 se analiza otro ejemplo, donde $\epsilon_S = 0,05$ y $P(S = f) = 0,5$. En el mismo generaron 100.000 realizaciones de CV_S obtenidos para PNI de sexo masculino y femenino, considerando H_1 o H_2 ciertas. Puede verse la diferencia entre la distribución de valores entre ambas hipótesis. Para H_1 cierta, la proporción de realizaciones que tienen un valor de CV_S característico del sexo masculino corresponden a 0,05. Esto es esperable dado que coincide con el valor ϵ_S . La diferencia entre las distribuciones de H_1 y H_2 permitirá distinguir ambas poblaciones y, por lo tanto, llegar a identificaciones. Puede extraerse que valores mayores de ϵ_S , harán que la diferencia entre las distribuciones sea menor, y por lo tanto se pierda la capacidad identificatoria.

En la Figura 2.4 se analiza el comportamiento de distintas métricas de rendimiento de la identificación a medida que aumenta el valor de ϵ_S . En todos los casos se considera un valor de corte de $CV_S = 1$, dado que separa los valores obtenidos para ambos sexos (Figura 2.2). Puede

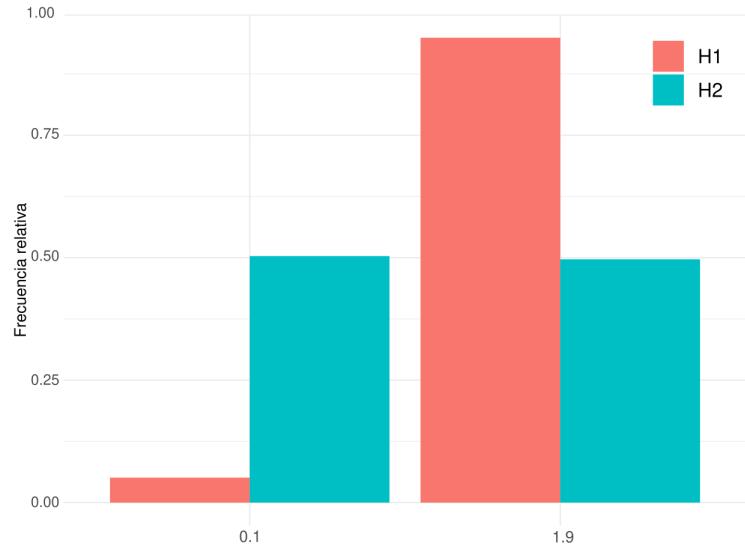


Figura 2.3 Frecuencia relativa de los valores de CV obtenidos en la simulación considerando H_1 y H_2 como verdaderas.

verse que todas las métricas (precisión, recall y MCC) caen a medida que aumenta ϵ_S . Con un $\epsilon_S = 0,5$ el MCC llega al valor de cero. Esto significa que la clasificación no se distingue por otra realizada al azar. Es entendible que, para una variable dicotómica, una tasa de error de 0,5 significa una capacidad asertiva esperable para una selección aleatoria. Lo mismo sucede con la precisión y el recall, donde los valores de 0,5 no se diferencian de los resultados esperados para un selector azaroso.

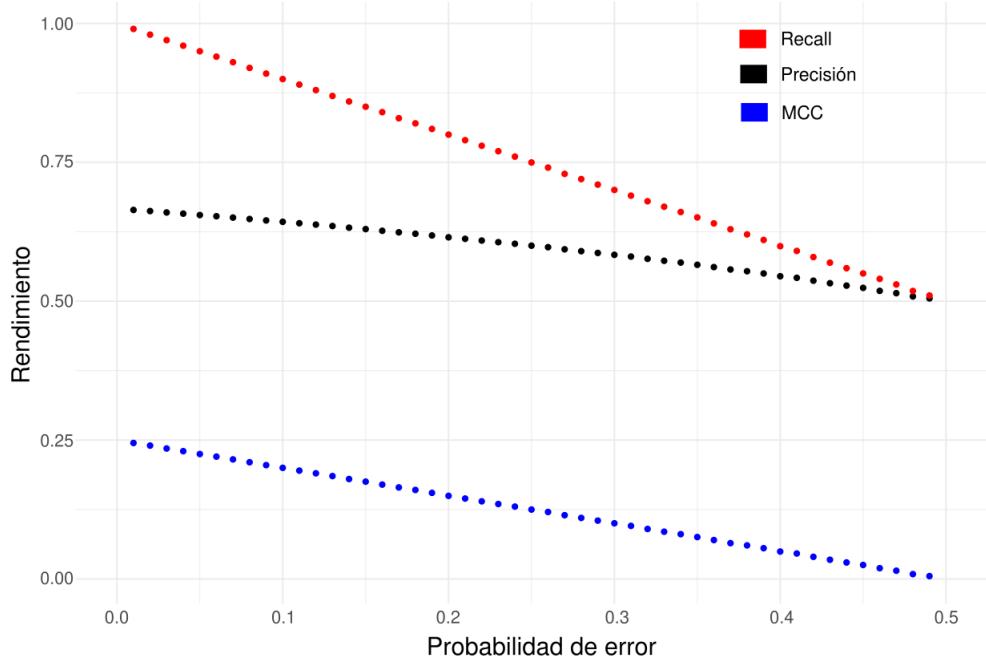


Figura 2.4 Métricas de rendimiento en función del valor de ϵ_S . En azul MCC, en rojo Recall y en negro precisión.

2.3.2 Variable categórica politómica: el color de pelo

Para la variable C se consideran cinco posibles colores, tipificados como $A_C = \{1, 2, 3, 4, 5\}$ y la siguiente distribución de probabilidades en la población, $P(C) = \{0, 3; 0, 25; 0, 20; 0, 15; 0, 10\}$. Los parámetros de error se presentan a continuación, considerando siempre que $\epsilon_{ij} = \epsilon_{ji}$.

ϵ_{12}	ϵ_{13}	ϵ_{14}	ϵ_{15}	ϵ_{23}	ϵ_{24}	ϵ_{25}	ϵ_{34}	ϵ_{35}	ϵ_{45}
0,04	0,04	0,01	0,01	0,01	0,01	0,01	0,03	0,04	0,02

Con estos valores, aplicando la ecuación 2.5 para obtener λ , queda la siguiente matriz:

		PNI				
		C_1	C_2	C_3	C_4	C_5
PD	C_1	0,909	0,036	0,036	0,009	0,009
	C_2	0,037	0,935	0,037	0,009	0,009
	C_3	0,036	0,009	0,893	0,027	0,036
	C_4	0,009	0,009	0,028	0,935	0,019
	C_5	0,009	0,009	0,037	0,019	0,926

Con estos parámetros, se procede a simular, mediante 100.000 realizaciones, la distribución de CV_C esperada considerando H_1 o H_2 verdadera para dos casos diferentes: (i) el $C_{PD1} = 1$, (ii) $C_{PD2} = 5$. Los resultados se pueden ver en las Figuras 2.5 y 2.6.

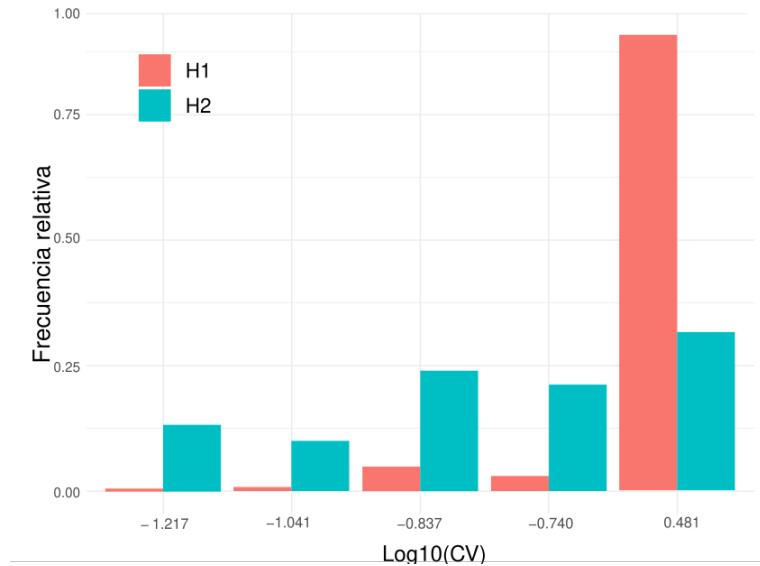


Figura 2.5 Frecuencia relativa de valores de $\text{Log}_{10}(CV)$ para el color de pelo, obtenidos en la simulación considerando que $C_{PD1} = 1$.

En la tabla 2.1 se presentan los valores de rendimiento para ambos casos, considerando un valor de corte $T_{CV} = 1$. Este valor de 1 indica que la verosimilitud de los datos es igual para ambas hipótesis. Puede verse como el color menos frecuente en la población presenta mejores métricas. Esto es esperable, dado que es más sencillo distinguirlo de la población general.

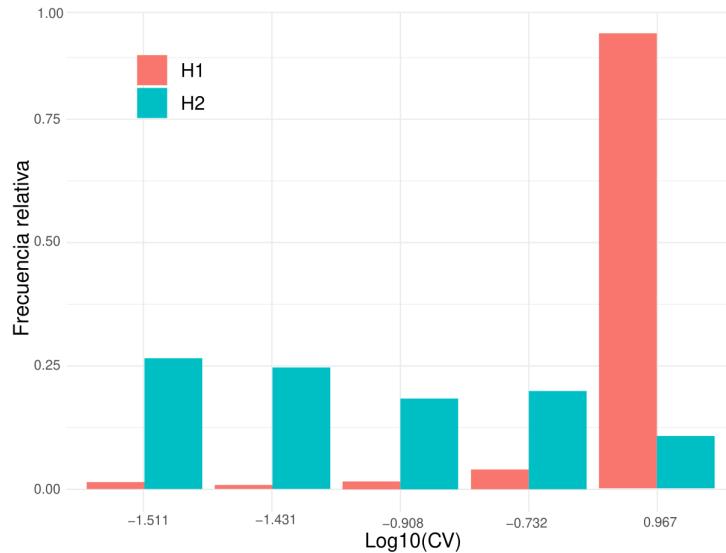


Figura 2.6 Frecuencia relativa de valores de $\text{Log}_{10}(CV)$ para el color de pelo, obtenidos en la simulación considerando que $C_{PD1} = 5$.

Métricas	$PD1$	$PD2$
Recall	0,913	0,926
Precision	0,774	0,908

Tabla 2.1 Métricas de rendimiento considerando un valor de umbral para CV igual a 1. Se indica para PD_1 y PD_2 donde los valores de los datos de la investigación preliminar son diferentes.

2.3.3 Variable continua: la edad

La variable edad, E , tal cual es descrita en la ecuación 2.7, recibe un tratamiento que permite caracterizar dos condiciones: (i) solapamiento de los intervalos de edad entre PD y PNI , con $M_E = 1$ y (ii) no solapamiento entre ambos intervalos, con $M_E = 0$. Esto hace que la variable M_E , que determina la CV_E , sea tratada como una variable categórica dicotómica (semejante a lo analizado para CV_S). La $P(M_E)$ se calcula a partir de la frecuencia de individuos que reportan el valor de M respecto a PD en la población de referencia. En la Figura 2.7 se comparan, para una misma edad, distintos intervalos para EPD , que podrían ser producto de testimonios con mayor o menor precisión o de herramientas de estimación de la edad. Puede verse que CV_E es menor para aquellos PD con un intervalo más grande. Como puede esperarse, el aumento de la incertezza deriva en una disminución del peso de la evidencia.

2.3.4 Combinando la evidencia

Una de las características que permite el enfoque bayesiano es combinar el aporte de las distintas líneas de evidencia multiplicando los valores de CV obtenidos. Esto es posible siempre y cuando las evidencias sean independientes, y las hipótesis contrastadas sean las mismas. Las tres variables analizadas cumplen con este criterio. En la Figura 2.8 se puede ver el valor de CVs obtenidos a partir de 100.000 realizaciones de una simulación donde a cada PNI se le asignaban

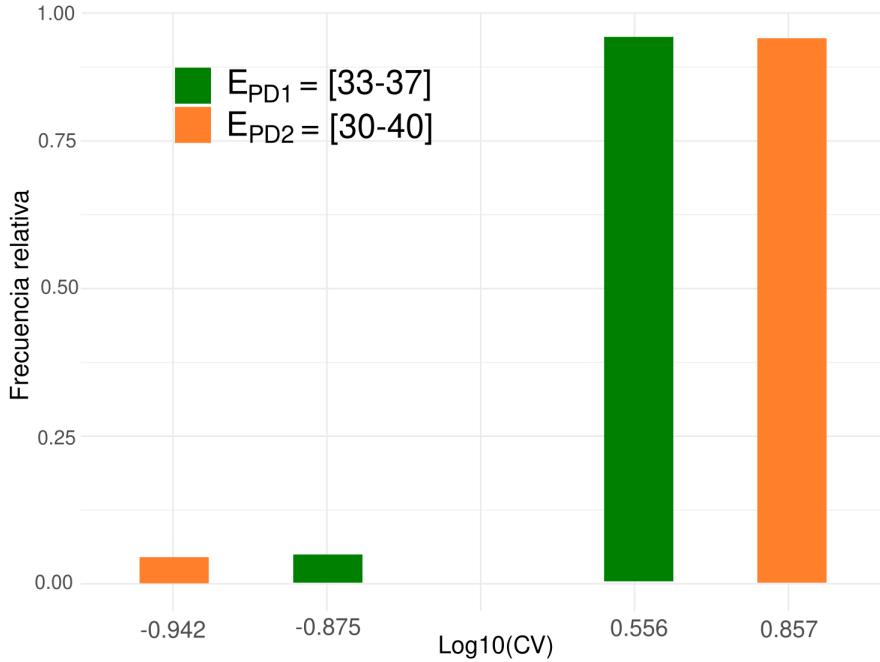


Figura 2.7 Frecuencia relativa de valores de CV para la edad, obtenidos en la simulación, dos PDs diferentes, con distintos rangos de edad.

las tres variables S , E y C tomando H_1 o H_2 como ciertas. Se consideró MP femenino ($S = f$), de 40 años de edad (intervalo 35 a 45, $E = \{35, 45\}$) y con color de pelo castaño ($C = 1$). Las métricas de rendimiento indican una precisión de 0,96 y un recall de 0,93.

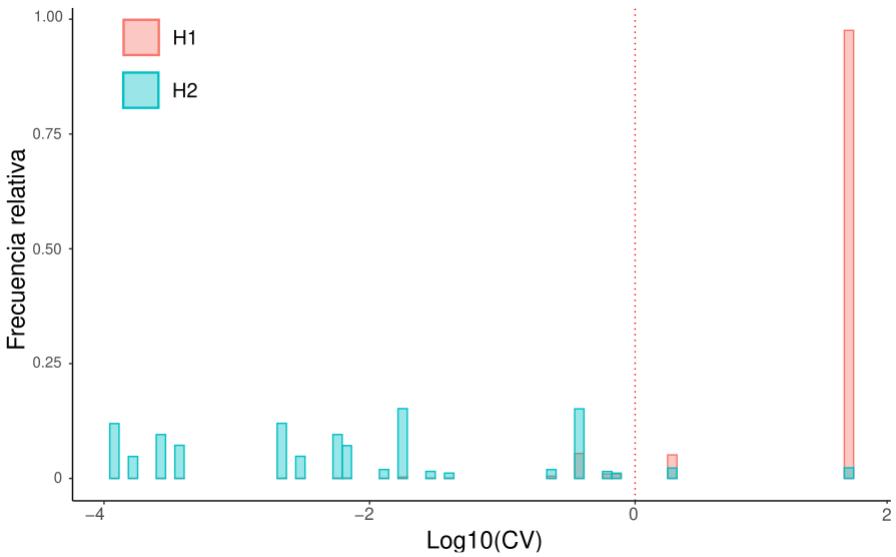


Figura 2.8 Frecuencia relativa de valores de $\text{Log10}(CV)$ para la combinación de variables obtenidos en la simulación considerando $C_{PD} = 1$, $S_{PD} = f$ y $E_{PD} = \{35, 45\}$. Se muestran los valores asumiendo H_1 y H_2 como verdaderas.

De forma más exhaustiva, se analizan dos ejemplos con las características especificadas en

la siguiente tabla:

	Sexo (S)	Color de pelo (C)	Edad (E)
PD_1	Femenino (F)	Castaño (1)	8 a 10
PD_2	Femenino (F)	Colorado (5)	18 a 22

Tabla 2.2 Tabla de datos recolectados durante la investigación preliminar para PD_1 y PD_2 . Se indica sexo biológico (S), color de pelo (C) y edad (E).

En la Figura 2.9 se muestra la distribución de probabilidades de las combinaciones de las variables considerando H_1 o H_2 como ciertas, además, se muestra la distribución de CVs .

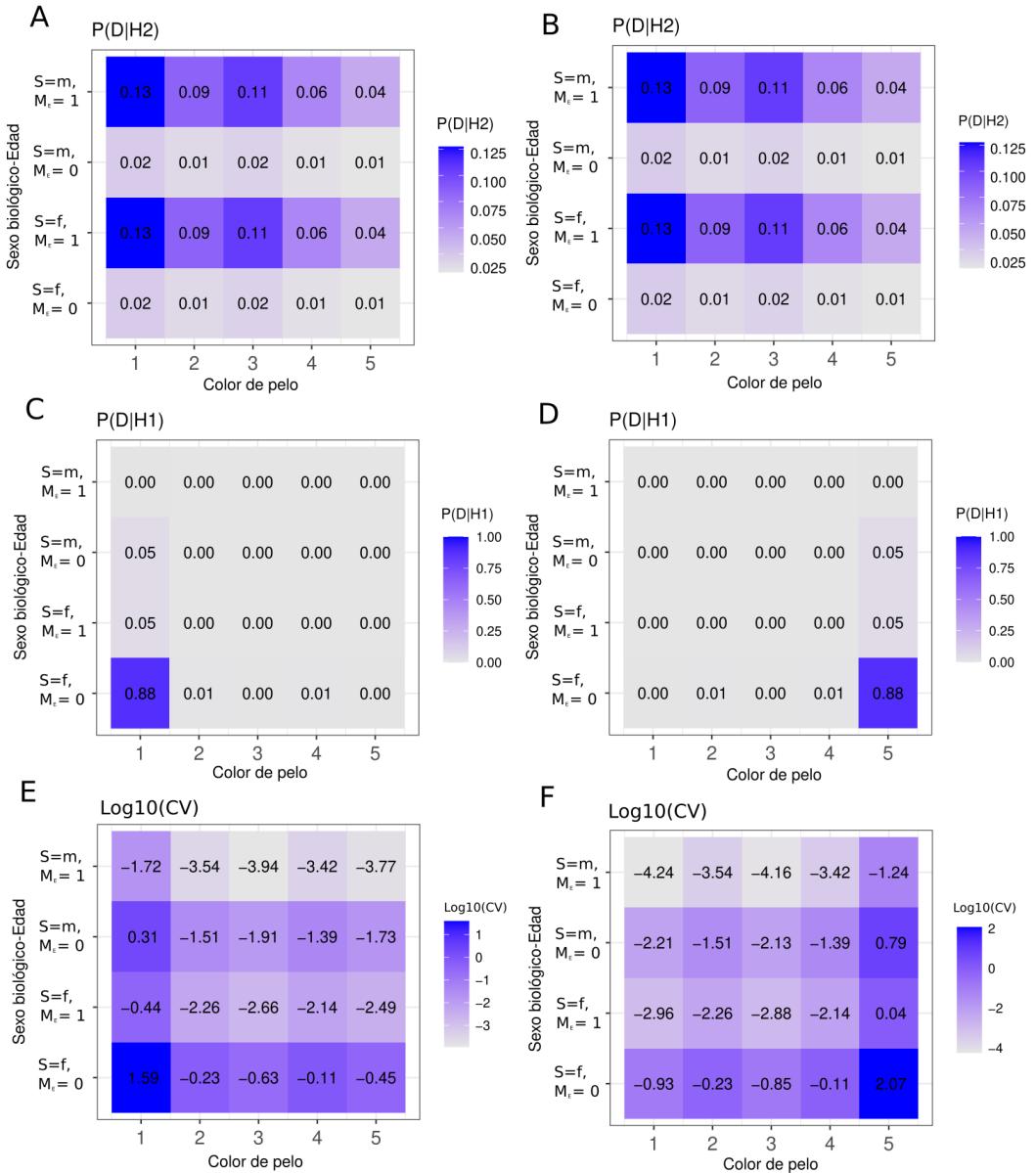


Figura 2.9 A y B. Distribución de probabilidad para la combinación de variables esperadas para PNI considerando H_2 como cierta, para PD_1 (izquierda) y PD_2 (derecha). D y E. Distribución de probabilidad para la combinación de variables esperadas para PNI considerando H_1 como cierta, para PD_1 (izquierda) y PD_2 (derecha). F y G. Valores de $\text{Log}_{10}(CV)$ para PD_1 (izquierda) y PD_2 (derecha) dada H_1 verdadera. Nótese que la figura E se obtiene dividiendo C por A, y la F dividiendo D por B.

2.3.5 Análisis de sensibilidad con el prior odds

Tomando como ejemplo el valor máximo de CV obtenido combinando las variables estudiadas, se realizará un análisis de sensibilidad probando distintos priors para H_1 y H_2 . En la siguiente tabla se enuncia solo el prior de H_1 entendiéndose que $P(H_2) = 1 - P(H_1)$.

$P(H_1)$	Prior odds	CV	Posterior odds
0,003	0,003	38,9	0,13
0,01	0,01	38,9	0,39
0,05	0,05	38,9	2,04
0,10	0,11	38,9	4,32
0,15	0,18	38,9	6,86
0,20	0,25	38,9	9,72
0,25	0,33	38,9	13,00
0,30	0,43	38,9	16,67
0,35	0,54	38,9	20,95
0,40	0,67	38,9	25,93
0,45	0,82	38,9	31,83
0,50	1	38,9	38,90

Tabla 2.3 Análisis de sensibilidad de prior odds. Se muestra la probabilidad a priori para la hipótesis H_1 , $P(H_1)$, el prior odds, el CV y el posterior odds obtenido en cada caso. Nótese que el CV es igual en todos los casos, lo único que varía es el prior odds.

Para el prior de 0,003 (primera línea de la tabla) se considera un total de 300 víctimas. El posterior odds por debajo de 1 indica que la evidencia recolectada no es suficiente para inclinar la balanza en favor de la H_1 . Por el contrario, H_2 continúa siendo más probable. Puede verse que a partir de un prior odds de 0,05, el posterior comienza a favorecer a la H_1 .

2.4 Discusión

La información recolectada durante la investigación preliminar es fundamental para el proceso de búsqueda (Puerto et al. 2021). Distintas propuestas han demostrado la utilidad de contemplar esta información para priorizar hipótesis de identificación de personas desaparecidas (Caridi et al. 2020, 2011). Por otro lado, se ha discutido acerca de la necesidad de generar protocolos para la definición de prior odds en identificaciones basadas en *test* de parentesco genético (Budowle et al. 2011). En dicho trabajo, se utilizan datos de la investigación preliminar para definir el prior odds. Específicamente se contemplan datos fenotípicos y se hace referencia a la necesidad de incorporar incertezza en la evaluación de dichos datos.

Desde el planteo propuesto en esta tesis, la observación de Budowle et al. refiere correctamente a la necesidad, y utilidad, de incorporar otra información, más allá de la genética, en el proceso de búsqueda. Además de utilizarla, se observa que es necesario llegar a una aproximación cuantitativa y comunicar en conjunto con los resultados del *test* de parentesco genético.

A lo largo de este capítulo se profundizó en modelos para la interpretación de evaluación del peso de la evidencia recolectada durante la investigación preliminar. A diferencia de lo analizado previamente por Budowle et al. (Budowle et al. 2011), se propone un paso de análisis

de la evidencia con modelos de cociente de verosimilitud, previamente a incorporar los datos en el prior odds del paso genético. Esta estrategia permite incorporar la incertezza asociada a la observación o registro de los datos, además de combinar distintas líneas de evidencia. Por otro lado, definiendo un prior para el paso de los datos de la investigación preliminar, se puede obtener un posterior. El mismo contiene el conocimiento de las creencias en torno al caso, y el de la evidencia aportada por la investigación preliminar. Este posterior luego puede ser utilizado como prior del paso genético. En este esquema, la evidencia que se utiliza es siempre evaluada estadísticamente, y un peso se le es asignado. Esto es acorde al esquema bayesiano.

Como se ha mencionado previamente, hay ejemplos de búsquedas en los cuales no se cuenta con información genética, y es necesario priorizar utilizando sólo los datos disponibles (Baraybar et al. 2020, Caridi et al. 2020). Este aspecto abre la puerta a la incorporación de estos modelos a los motores de búsqueda, con el fin de seleccionar casos donde la información genética sea confirmatoria. Asimismo, los modelos pueden ser utilizados para complementar la información genética cuando esta aporte poco poder estadístico (por ejemplo con muestras degradadas de ADN). Estos aspectos serán abordados en el capítulo 4, que trata de la toma de decisiones en casos de búsqueda de personas desaparecidas mediante el uso de bases de datos.

2.5 Conclusiones del capítulo

En este capítulo se propusieron distintos modelos para la evaluación estadística de líneas de evidencia recolectadas durante la investigación preliminar. Se discutió además las implicaciones que esto tiene dentro de la búsqueda de personas desaparecidas, permitiendo cuantificar el impacto de la formalización en la búsqueda. El enfoque propuesto mantiene las mismas hipótesis que se presentan con la evidencia genética. De este modo, es posible combinar dichas líneas de evidencia. Se volverá sobre este punto en el capítulo 4, donde además se evaluará el impacto de la evidencia no-genética en un contexto de identificación mediante análisis genéticos.

Capítulo 3

El problema de la formalización del peso estadístico de la evidencia: análisis de parentesco genético

El perfil genético ha sido, desde las últimas décadas, la pieza principal de evidencia para las identificaciones de individuos. Una de las características que le permitió ganarse este lugar es el alto nivel discriminatorio que posee, gracias a regiones específicas hipervariables presentes en el genoma humano y en el de muchas otras especies. La evidencia genética es utilizada tanto en contextos de casos forenses típicos, como en el hallazgo de una muestra con material genético en una escena de crimen, como en contextos de búsqueda de personas desaparecidas. En este último caso, un tipo específico de análisis es necesario para la interpretación estadística del peso de la evidencia. El mismo es conocido como *test* de parentesco genético. Al igual que en el modelo planteado para líneas de evidencia recolectadas en las investigación preliminar, dos hipótesis son contrastadas, por un lado se considera que la persona no identificada, PNI , es la persona desaparecida, PD , presentada como H_1 . Por otro lado se considera que PNI no es PD , hipótesis denominada H_2 . Los modelos de verosimilitud surgidos de estas hipótesis remontan sus conceptos fundamentales a dos grandes áreas de la biología, por un lado la genética de poblaciones, y por otro lado las leyes de la herencia genética. Estos modelos están ampliamente discutidos y validados dentro de la comunidad científica. Asimismo, la incertidumbre en torno a sus parámetros y supuestos ha sido descrita en detalle. La extensión del uso de la evidencia genética a lo largo del mundo derivó en la creación de los primeros bancos de datos genéticos con fines forenses hace ya más de dos décadas. La informatización de los procesos, el desarrollo de nuevos algoritmos y el almacenamiento de grandes volúmenes de datos han traído aparejados nuevas oportunidades, pero también nuevos desafíos. En este capítulo se busca hacer especial énfasis en los eventos que pueden violar los supuestos de los modelos, induciendo errores en la interpretación de la evidencia, lo cual deriva en errores en la asignación de identidades biológicas. Se analizan los modelos matemáticos para la evaluación de la evidencia genética, describiéndolos y contextualizando su aplicación en casos de identificación de personas. Más aún, se introducen tres metodologías empleadas para el cómputo de la verosimilitud de los datos genéticos en los pedigríes. En particular, se explica en detalle la inferencia utilizando redes bayesianas, a partir de las cuales se generó una herramienta de código abierto llamada *fbnet*, y disponible en el repositorio CRAN.

3.1 Introducción al capítulo

En esta sección se introducen los modelos para la evaluación del peso de la evidencia genética en casos de búsqueda de personas desaparecidas. Se profundiza en el análisis de los marcadores autosómicos, por tratarse de los más utilizados en los laboratorios forenses, dejando aquellos no autosómicos y sexuales por fuera del objeto de estudio (Jeffreys et al. 1985b). Aún así, existen desafíos pendientes no solo en la interpretación de la evidencia genética de estos marcadores no autosómicos, si no también en la combinación de las líneas de evidencia. Para una lectura más profunda se recomienda el trabajo de Amorim et al. (Amorim and Budowle 2016). Más generalmente, el análisis genético en la búsqueda de personas desaparecidas se engloba dentro de lo que se conoce como análisis de parentesco genético (Egeland et al. 2015). Como se mencionó en el capítulo 1, dos modelos que cuentan con amplio consenso dentro de la comunidad científica entran en juego a la hora de evaluar la evidencia. Por un lado el modelo de genética poblacional y por otro el de herencia mendeliana (Falconer and Mackay 1983). Estos serán analizados en dos niveles diferentes, el poblacional y el del pedigrí. Además, se introduce un tercer nivel que, continuando la nomenclatura adoptada por Egeland et al., se denominará nivel observacional (Egeland et al. 2015). El mismo da cuenta de distintos fenómenos que pueden ocurrir durante la tipificación de la muestra de ADN, y tienen que ver primordialmente con el laboratorio que realiza el análisis, o bien con el estado y proveniencia de las muestras (Amorim and Budowle 2016). La integración de todos estos niveles requiere de altas capacidades de cómputo, además de estrategias de simplificación del problema. Sobre todo, en casos complejos donde se presentan pedigríes con muchos miembros, o se analizan muchos marcadores a la vez.

Existen distintos algoritmos generados para el análisis de parentesco genético. Estos algoritmos representan distintas soluciones al problema planteado de alto costo computacional para el cálculo de la verosimilitud (Egeland et al. 1997, Elston and Stewart 1971, Lander and Green 1987, Allen and Darwiche 2008, Chernomoretz et al. 2020). Una de las ventajas de la implementación de soluciones que optimizan la capacidad de cómputo de la valoración estadística de la evidencia es que han funcionado como bloque inicial de la masificación de búsqueda de personas desaparecidas mediante el empleo de bases de datos genéticas (Puerto and Tuller 2017).

3.1.1 Parentesco genético

Un análisis de parentesco genético para la búsqueda de personas desaparecidas se basa en el contraste de dos hipótesis H_1 y H_2 que se describen a continuación. H_1 : PNI es PD. Esto, para el caso particular de la evidencia genética, se puede interpretar como que PNI se encuentra biológicamente relacionado al pedigrí de referencia, que posee la información genética de los familiares de PD, denominado P1 (Figura 3.1). Por otro lado H_2 , PNI no es PD. Esto se puede interpretar como que PNI no se encuentra relacionado con el pedigrí de referencia, siendo un individuo tomado al azar de la población, esto se representa en el pedigrí P2 (Vigeland 2021).

El resultado se expresa en un cociente de la probabilidad de los datos genéticos dadas las hipótesis, siendo el método recomendado por la Sociedad Internacional de Genética Forense (Gjertson et al. 2007). El pedigrí se utiliza para representar las relaciones familiares entre individuos. El mismo puede analizarse como una grafo acíclico dirigido en el cual los nodos repre-

sentan individuos, y los conectores relaciones de parentesco directas (padre-hijo y madre-hijo) (Egeland et al. 2015). A continuación se utiliza como ejemplo un caso típico de paternidad, donde la persona desaparecida es un supuesto padre (SP). En la Figura 3.1 se muestran los pedigríes alternativos.

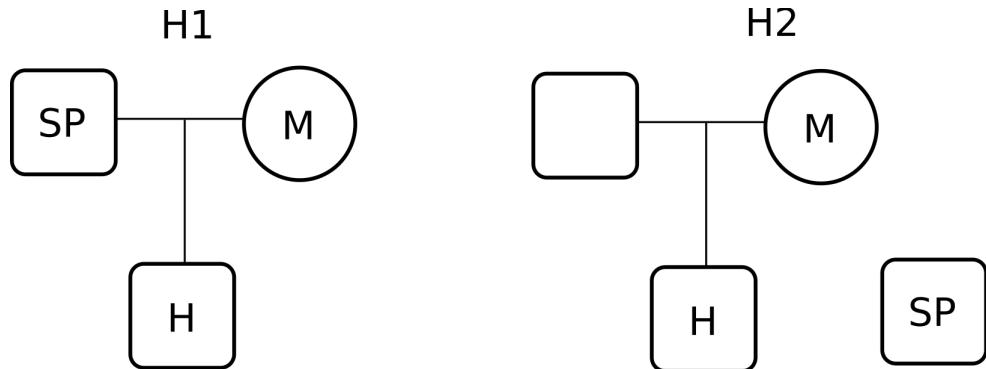


Figura 3.1 Caso de paternidad. Se comparan dos hipótesis, H_1 corresponde a la hipótesis de parentesco entre supuesto padre (SP) e hijo (H). H_2 es la hipótesis de no parentesco.

En el pedigrí se cuenta con información genética de la madre y del hijo, y se analiza la compatibilidad de SP, que es la persona no identificada, del cual se adquiere la muestra. Un pedigrí, P_1 , involucra a SP como padre del hijo, en conjunto con la madre. El pedigrí alternativo, P_2 , excluye a SP, colocando en la posición de padre del hijo a un individuo desconocido. En general, el pedigrí suele ser más complicado que el analizado en este ejemplo, incorporando información de tíos, abuelos, bisabuelos, etc. Además los pedigríes evaluados pueden diferir a los presentados. Por ejemplo, el pedigrí de H_2 podría considerar que SP es el tío del hijo, y no el padre. El único requisito en la generación de pedigríes para el contraste de hipótesis es que todos los miembros genotipados deben encontrarse en ambas hipótesis, es decir estar en una parte del pedigrí (inclusive desvinculados del resto, como SP en el ejemplo analizado).

El análisis de parentesco genético difiere del contraste de hipótesis clásico (Vigeland 2021). Una de las características principales es que las hipótesis, en el caso de análisis de parentesco, se describen verbalmente, como sentencias declarativas. En contraposición, el contraste de hipótesis clásico es paramétrico, donde se evalúa el valor de un parámetro específico, ej. $\mu = 0$ versus $\mu \neq 0$. Otra diferencia es la ausencia de hipótesis nula en el caso forense. Esto genera que muchas de las herramientas desarrolladas para el contraste de hipótesis clásicas no se puedan aplicar, o carezcan de sentido, en el análisis de parentesco. Ejemplos de esto son los errores tipo I o tipo II, p-valor, etc. García-Magariños et al. (García-Magariños et al. 2015) han explorado la formulación paramétrica de las hipótesis para los casos forenses. Esto trae una serie de ventajas, que implican que la hipótesis de no-parentesco, generalmente H_2 , como se ve en el ejemplo, puede ser formulada de forma mucho más general. Aún así, el alto número de parámetros a ser evaluados dificulta el enfoque.

El cociente de verosimilitud surge a partir de evaluar la probabilidad de los datos genéticos obtenidos considerando H_1 como cierta, sobre la verosimilitud de los mismos datos considerando H_2 como cierta. Como han conceptualizado Egeland et al. (Egeland et al. 2015) el cálculo de verosimilitud tiene en cuenta tres niveles: el observacional, el del pedigrí y el poblacional.

Por un lado el nivel *observacional*, que da cuenta de la probabilidad de obtener los datos en

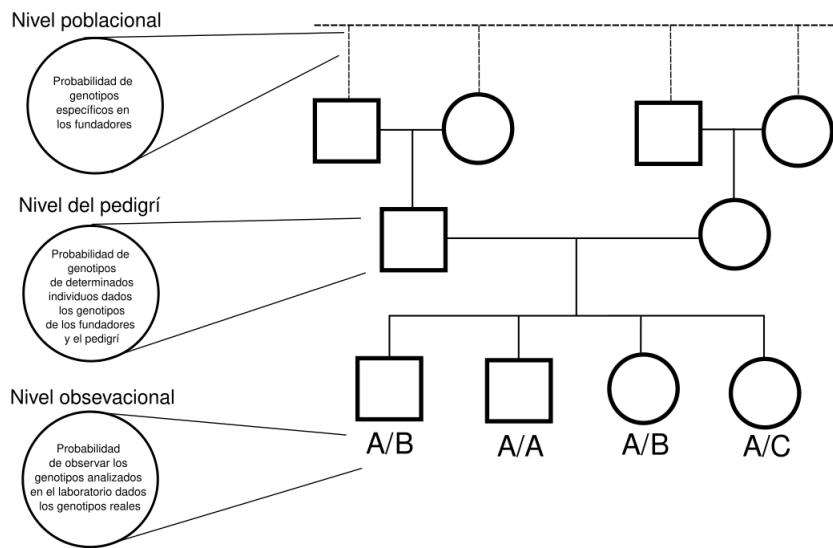


Figura 3.2 Concepto de niveles para los modelos de análisis de la evidencia genética. Se muestra el nivel poblacional, que describe las probabilidades genotípicas de los fundadores, el nivel del pedigrí que describe la herencia genética y el nivel observacional, que describe la probabilidad de obtener los genotipos considerando los métodos de laboratorio, entre otros. La Figura se basa en una presentada previamente por Egeland et al. (Egeland et al. 2015).

función de distintas características del laboratorio (procesamiento de la muestra o estado de la misma). El nivel del pedigrí da cuenta de la probabilidad de observar los datos en función de las relaciones de herencia brindadas. Este nivel se encuentra fuertemente relacionado al mecanismo de herencia genética. Por último el nivel *poblacional* permitirá describir la probabilidad de observar genes de aquellos individuos cuya procedencia no es declarada, es decir, no cuentan con ancestros en el pedigrí (Hamilton 2021). Por ejemplo en el pedigrí de la Figura 3.1 (H_1), los individuos SP y M no poseen ancestros declarados. A este tipo de individuos se los conoce como *fundadores*.

Para analizar estos conceptos puede utilizarse el ejemplo de la Figura 3.1. En el mismo, SP y M son fundadores en ambos pedigríes. Esto significa que la probabilidad de observar sus genotipos estará explicada por el modelo de genética de poblaciones que depende de las frecuencias poblacionales. En cambio, la probabilidad del genotipo del hijo, H, estará dada por el mecanismo de herencia al nivel del pedigrí. Por ejemplo, para un determinado marcador, llámese $M1$, SP es homocigota para el alelo a , es decir, $M1_{SP} = \{a, a\}$, y el de la madre es $M1_M = \{b, b\}$. Si H_1 fuera cierta, el genotipo del hijo debería ser $M1_H = \{a, b\}$, con una probabilidad igual a 1 (por lo tanto cualquier otra posibilidad sería cero, asumiendo que no se permiten mutaciones ni errores de laboratorio de diferente tipo durante el genotipado). Es decir, el hijo debe haber heredado un alelo paterno (solo a) y uno materno (solo b). En cambio, en el pedigrí alternativo, donde se evalúa H_2 , el genotipo esperado para H es $M1_H = \{b, -\}$, donde $-$ indica que cualquier alelo disponible en la población de referencia podría ser compatible. Esto se debe a que se desconoce el genotipo paterno. La probabilidad de heredar el alelo por vía paterna estará dada por el modelo poblacional. Este ejemplo es una simplificación, ya que en

casos reales intervienen múltiples fenómenos afectando las probabilidades de los genotipos. A continuación se introducen las principales características de cada nivel.

3.1.2 El nivel poblacional

El modelo poblacional se utiliza para explicar la probabilidad de los genotipos de los *fundadores* (Egeland et al. 2015). Es decir, aquellos individuos cuyos genotipos no se encuentran condicionados por otro individuo (ancestro) en el pedigrí. El principal modelo considera el siguiente escenario: para un marcador genético autosómico existen un conjunto de k alelos considerados posibles, y los mismos son observados con probabilidades, o frecuencias alélicas, que pueden ser especificadas en términos de probabilidades. Estas probabilidades son fijas y tienen un valor positivo, que va de cero a uno. La suma de los valores de frecuencias alélicas de un marcador es igual a 1. Los diferentes alelos observados en distintas o en la misma persona son independientes. Una de las consecuencias directas de este modelo es el conocido equilibrio de Hardy-Weinberg (Edwards 2008). El mismo declara que la probabilidad de observar un genotipo A/B será $2p_a p_b$, y la de A/A, p_a^2 , siendo p_a la frecuencia alélica de a y p_b la frecuencia alélica de b . Este modelo básico ha sido extensamente utilizado dentro del campo de la genética, aún así, son muchos los contextos en los cuales sus asunciones no se cumplen. Por ejemplo, para mantener una frecuencia alélica fija es necesario que se den un conjunto de eventos, por ejemplo *ausencia de mutaciones*, la *ausencia de migraciones* o la *ausencia de presión de selección*. Por otro lado, la subestructuración poblacional, donde los individuos no se aparean de forma aleatoria, afecta el cálculo de las frecuencias genotípicas, como se explicará en detalle más adelante. Debido a la naturaleza de los marcadores moleculares utilizados para el análisis forense se asume ausencia de presión de selección (Amorim and Budowle 2016). Por esta razón en las siguientes secciones tan solo se hace hincapié primordialmente en los aspectos relacionados al campo forense y que tienen efecto significativo en el cálculo del peso de la evidencia genética.

Existe amplia bibliografía sobre el tipo de estudios que se realizan sobre las bases de datos poblacionales de referencia y es un protocolo estándar en laboratorios forenses (Amorim and Budowle 2016). En la sección Métodos se hará hincapié acerca del tratamiento de los alelos que aparecen con muy baja frecuencia en la base de datos por su particular interés en el proceso de evaluación de la evidencia genética.

3.1.3 El nivel del pedigrí

Los mecanismos de herencia genética han sido desarrollados a lo largo de los últimos siglos con el aporte fundamental de Mendel (Bateson and Mendel 2013). La principal característica del modelo de herencia que entra en juego es la segregación genética. El modelo básico es el siguiente: dado un parente de genotipo A/B para un determinado marcador, el hijo heredará A o B con la misma probabilidad, es decir 0,5. Lo mismo sucede para la herencia por vía materna. Este principio se cumple para los marcadores autosómicos, además de considerar que la transmisión de los cromosomas es independiente, y por lo tanto, los marcadores moleculares analizados son independientes entre sí. Entonces, si por ejemplo un parente, que posee genotipo A/A, y una madre, genotipo B/B, el hijo tendrá un genotipo A/B con probabilidad igual a 1. Esto implica además que cualquier otro genotipo para el hijo será incompatible, y por lo tanto

su probabilidad será cero. Los marcadores STRs han sido seleccionados para la identificación humana debido a su gran variabilidad. La misma reside en que estas partes del genoma son proclives a eventos de mutación (Jeffreys et al. 1985a). La mutación es un proceso que ocurre durante la replicación del genoma, en el cual las nuevas copias de ADN sufren errores que las diferencian del molde original. Esto impacta directamente en la herencia, debido a que existe la posibilidad de que durante el proceso de generación de gametas el alelo A se convierta en un B, y viceversa. La frecuencia con la que ocurren las mutaciones varía enormemente entre diferentes tipos de marcadores. Los datos de 18 marcadores disponibles en STRbase (Ruitberg et al. 2001) muestran tasas de mutación desde 0,0001 hasta 0,0064. Los marcadores de *polimorfismos de nucleótido único*, o SNP, tienden a tener tasas de mutación en el rango de 10^{-9} a 10^{-8} . Si solo se usan algunos de estos marcadores, puede ser razonable asumir la ausencia de mutaciones. Sin embargo, los marcadores STR populares en los casos de genética familiar pueden tener tasas de mutación de alrededor del 0,005. Distintos modelos matemáticos para la incorporación de mutaciones han sido propuestos (Egeland et al. 2014). Algunos buscan captar con mayor precisión el mecanismo molecular por el cual ocurre la mutación, siendo muy detallados a la hora de definir las probabilidades de los distintos tipos de mutación. Otros, más simples, se utilizan por su bajo costo computacional, dando una descripción muy general de las posibles mutaciones.

3.1.4 El nivel observacional

Este último nivel abarca un conjunto de fenómenos que van desde la toma de muestra, procesamiento y análisis hasta la interpretación. Sencillamente se centra en los eventos que pueden alterar la identidad del alelo o genotipo analizado, asignándole una identidad errónea. Como es de esperarse, errores en el genotipado pueden tener fuertes implicancias en la interpretación de la evidencia, impactando en el resultado final. A continuación se describen algunos eventos comunes dentro del campo forense: (i) alelo silente: en ciertos contextos, la amplificación mediada por PCR de los marcadores falla, debido a variaciones genéticas en las regiones lindantes a los STRs. En estos casos, si el error no es debidamente detectado, el genotipo asignado al individuo es homocigota para el alelo que sí se detecta. (ii) drop in y drop out: en casos de muestras de baja calidad o contaminadas se pueden dar fenómenos en los cuales la PCR falla, impidiendo que se reconozcan alelos presentes en la muestra (drop out), o inclusive que se detecten alelos de forma errónea (drop in), (iii) errores de ingreso de datos: un último punto, no trivial, refiere a la probabilidad de un error técnico durante el ingreso de datos de genotipos. Esto puede volverse frecuente en casos en los cuales los laboratorios se ven obligados por grandes volúmenes de muestras a procesar y (iv) errores de asignación alélica por fallas en asociadas a fallas de EC (electroforesis capilar) y/o errores de interpretación del analista que está analizando y validando el perfil genético obtenido.

3.1.5 Algoritmos para el cómputo de la verosimilitud

A nivel de cómputo, la complejidad del cálculo de verosimilitud estará estrechamente relacionada a los datos faltantes en el pedigrí. Por ejemplo, en el caso de la Figura 3.1, mientras que H_1 es relativamente sencilla de calcular, con todos los miembros genotipados, H_2 enfrenta el

problema del padre, SP, cuyo genotipo no se conoce. La verosimilitud para H_1 estará dada por el producto entre la verosimilitud del genotipo del supuesto padre, la madre (obtenidos a partir del modelo poblacional, dado que son fundadores) y el del hijo (obtenido a partir del modelo del pedigrí, considerando el genotipo de la madre y el supuesto padre). En cambio, para H_2 , el supuesto padre y la madre se continúan calculando a partir del modelo poblacional, pero para el hijo es necesario integrar sobre todos los posibles valores que podría tener el padre no genotipado. Para el marcador $M1$ la madre es homocigota en b, por lo tanto el hijo puede ser a/b , b/b , c/b , ..., l/b , siendo l , el último alelo de dicho marcador. La verosimilitud de los pedigríes planteados en ambas hipótesis estará dada por la suma de la probabilidad de todos aquellos posibles genotipos. Aunque en términos computacionales el ejemplo pueda resultar sencillo, el problema escala con la incorporación de individuos en el pedigrí. La cantidad de términos a computar en un pedigrí estará dada por el número de posibles genotipos elevado a la cantidad de miembros del pedigrí (en la sección métodos se detallará la expresión matemática a partir de la que se obtiene esto). En casos complejos, donde es necesario realizar identificaciones mediante familiares distantes esto no es trivial. Distintos enfoques se han propuesto a la hora de abordar el problema computacional. Los dos principales algoritmos propuestos para el cómputo exacto de la verosimilitud del pedigrí son el de Elston-Stewart (Elston and Stewart 1971) y el de Lander-Green (Lander and Green 1987). Ambos radican en un método por eliminación de variables que depende de distintas estrategias para encontrar el orden de eliminación. El algoritmo de Elston-Stewart resuelve un grupo nuclear familiar a la vez. El algoritmo de Lander-Green se basa en la aplicación del Modelo Oculto de Markov, y resuelve un locus a la vez. La complejidad de Lander-Green es lineal al número de loci, pero exponencial con el número de individuos no-fundadores en el pedigrí. En cambio, la complejidad de Elston-Stewart es lineal al número de individuos no fundadores en el pedigrí, pero exponencial con el número de loci. Es así que Elston-Stewart se utiliza primordialmente en casos con pedigríes compuestos por múltiples individuos, pero con bajo número de marcadores genéticos a ser analizados. Lander-Green, en cambio, puede lidiar mejor con pedigríes pequeños, pero con un gran número de marcadores. Con los años y los avances en la capacidad de cómputo, los límites de ambos algoritmos se han ido extendiendo. Aún así, está claro que cada algoritmo se encuentra relacionado a diferentes problemas de cálculo de verosimilitud. En este capítulo se aborda el problema de cómputo mediante una estrategia diferente, desarrollada en la presente tesis (Chernomoretz et al. 2022). El enfoque consiste en adaptar el marco de redes bayesianas, originalmente desarrollado para análisis de ligamiento genético Fishelson and Geiger (2002), Darwiche (2009), con el fin de abordar el problema del cálculo de verosimilitud del pedigrí para la identificación de personas. Este enfoque permite hacer uso de diferentes estrategias para ordenar la eliminación por variables, dando un marco más general y adaptable a distintos contextos.

3.2 Métodos

En esta sección se formalizan matemáticamente los modelos asociados a los distintos niveles previamente introducidos. Además, se aborda el problema del cómputo del cociente de verosimilitud y su resolución mediante el enfoque de redes bayesianas.

3.2.1 Datos genéticos

Los datos están constituidos por un set de marcadores genéticos, denominados STRs. A su vez, cada uno puede tomar un conjunto de valores alélicos, simbolizados por un número que muestra la cantidad de repeticiones de la secuencia nucleotídica. Por ejemplo, el marcador M_1 , puede tomar valores tal que $M_1 = \{2, 3, 8, 11, \dots, 16\}$. Para un determinado individuo, i , su genotipo para un dado marcador será $G_i = \{a_f, a_m\}$, siendo a_f el alelo de procedencia materna, y a_p el de procedencia paterna.

3.2.2 La evaluación estadística de la evidencia genética

Similarmente a lo planteado en el capítulo 2, el contraste de hipótesis mediante el cálculo de posterior odds es la forma consensuada de evaluación de la evidencia (Gjertson et al. 2007). Las hipótesis contrastadas son las mismas, $H_1 : \text{PNI es PD}$, y $H_2 : \text{PNI no es PD}$, y corresponde a un individuo tomado al azar de la población de referencia. Utilizando de molde un pedigrí con la misma estructura del de la Figura 3.1, pero donde el supuesto padre corresponde a una persona desaparecida, las hipótesis contrastadas implican la evaluación genética de dos pedigríes diferentes. El que se vincula a H_1 , posiciona a PNI (el supuesto padre) en el lugar de la persona desaparecida, PD, mientras el pedigrí de H_2 lo coloca por fuera, no estando relacionado con el resto de los miembros del pedigrí de referencia. De forma más general, múltiples pedigríes pueden ser evaluados, y se llamará φ_j , al *j*esimo, siendo K el número de pedigríes alternativos evaluados. La expresión de Bayes se muestra a continuación:

$$P(\varphi_j | G) = \frac{p_j P(G|\varphi_j)}{\sum_{k=1, k=K} p_k P(G|\varphi_k)} \quad (3.1)$$

Donde p_j es la probabilidad a priori del pedigrí y G son los todos genotipos observados. Para el cálculo se debe especificar $P(G|\varphi_j)$, o verosimilitud del pedigrí. En las siguientes secciones se describe cada uno de los componentes de dicha verosimilitud.

3.2.3 Modelos poblacionales

Para un determinado marcador, dígase m , los posibles valores alélicos que pueden tomar se listan, $A_m = \{a_1, a_2, a_3, a_4, \dots, a_n\}$, siendo n el número de alelos. Su respectivas probabilidades constituyen un vector de forma $P_m = \{p_{a_1}, p_{a_2}, p_{a_3}, p_{a_4}, \dots, p_{a_n}\}$. La suma de los valores del vector es igual a 1. El modelo general propone que frente a la ausencia de mutaciones, migraciones y selección natural, asumiendo apareamiento aleatorio y un tamaño infinito de la población, las probabilidades genotípicas pueden ser calculadas directamente a partir de las frecuencias alélicas utilizando la distribución multinomial, para un marcador con un número n de alelos, considerando un organismo diploide, para un determinado alelo i , la probabilidad del genotipo homocigota estará dada por:

$$p(a_i/a_i) = p_i^2 \quad (3.2)$$

Y su correspondiente heterocigota, junto a un alelo q será

$$p(a_i/a_q) = 2p_ip_q \quad (3.3)$$

3.2.3.1 Frecuencias mínimas

Una frecuencia es, por definición, computada desde una serie de observaciones específicas de un evento, dividido el total de observaciones realizadas. La ecuación que describe la frecuencia es la siguiente:

$$p = C/N \quad (3.4)$$

Siendo C la cantidad de conteos del evento, y N la cantidad total de observaciones. El evento analizado en este caso es la presencia del alelo, y la cantidad de observaciones corresponde a la cantidad de alelos analizados en la base de datos. Si existen L individuos, $N = 2L$. Esto sucede dado que se consideran independientes los alelos de cada individuo. En ciertos casos puede suceder que C sea muy bajo. Distintos autores abordan las complejidades que esto produce en la evaluación estadística de la evidencia (Balding and Steele 2015, Gill and Buckleton 2004). La principal problemática radica en que obtener probabilidades a partir de eventos de baja frecuencia tiene una mayor incertezza. Una propuesta definida en Butler et al. (Butler 2005) consiste en adjudicar una frecuencia de $5/N$ a todos los alelos para los cuales se conozca su existencia y no hayan sido vistos en la base de datos, o bien tengan una cantidad de observaciones menor a 5. Esto se conoce como criterio de frecuencia mínima. La razón por la cual asignar 5 y no 1 proviene de un enfoque conservativo, para disminuir el valor de cociente de verosimilitud en caso de una identificación. Recuérdese que la hipótesis H_2 evalúa la evidencia genética utilizando el modelo poblacional, por lo tanto una probabilidad muy baja de observación de un alelo derivará en un valor bajo en el denominador, aumentando el valor del cociente de verosimilitud, CV . También se han propuesto estrategias bayesianas para el cálculo de las frecuencias alélicas, para profundizar en el problema de la estimación de las frecuencias se recomienda el libro de Egeland et al. (Egeland et al. 2015).

3.2.3.2 Estructuración poblacional

La idea de que los alelos de los fundados de un pedigrí son aleatoriamente muestreados de una población, con una frecuencia alélica definida, es obviamente una simplificación. Contextos socioeconómicos, redes sociales, bagaje cultural, entre muchos otros factores afectan al hecho de que el apareamiento no sea aleatorio inclusive dentro de un mismo territorio. Sin embargo, modelar esto no es sencillo.

La estratificación de la población y la relación entre individuos pueden invalidar algunos cálculos previos. Por ejemplo, el Equilibrio Hardy-Weinberg puede no ser aplicable, como en el caso en que los fundadores de un pedigrí puedan estar relacionados de forma remota, aunque sin saberlo, simplemente porque pertenecen a la misma subpoblación. Balding y Nichols (Balding and Nichols 1995) propusieron una forma práctica de manejar estos problemas, y ha sido adoptada dentro de la comunidad forense (Egeland et al. 2015). A continuación se describe el enfoque.

El efecto de la estratificación de la población se modela mediante el coeficiente de coancestría $\theta \in [0, 1]$. El valor 0 corresponde a la Equilibrio Hardy-Weinberg, mientras que los valores positivos aumentan la probabilidad de homocigotas. Se supone que los alelos se muestrean secuencialmente. La probabilidad de muestrear el alelo i como el primer alelo es p_i . Supongamos que i se muestrean como el j -ésimo alelo y sea b_j el número de alelos del tipo i entre los $j-1$ alelos muestreados previamente. Para lograr que el j -ésimo alelo se muestree considerando los alelos ya muestreados se utiliza $\frac{\theta b_j + (1-\theta)p_i}{\theta(j-1) + (1-\theta)}$ como la probabilidad de muestreo del alelo i . Reordenando, esto queda la fórmula de muestreo.

$$p'_i = \frac{b_j \theta + (1 - \theta)p_i}{1 + (j - 2)\theta} \quad (3.5)$$

De la misma se extraen las siguientes expresiones para las frecuencias genotípicas, para un determinado alelo i , la probabilidad del homocigota estará dada por

$$p(a_i/a_i) = \theta p_i + (1 - \theta)p_i^2 \quad (3.6)$$

Y su correspondiente heterocigota, junto a un alelo q será

$$p(a_i/a_q) = 2p_i p_q (1 - \theta) \quad (3.7)$$

Nótese que se trata de una corrección al modelo de Hardy-Wainberg, a mayor estratificación, θ será más alto y por lo tanto la población se enriquecerá de genotipos homocigotas.

3.2.4 Modelos para el pedigrí

El modelo para el pedigrí se propone desde un enfoque de redes bayesianas. Como se ha mencionado en la introducción, las redes bayesianas cuentan con dos componentes, por un lado la estructura de la red, y por otro lado la parametrización, que implica definir las probabilidades condicionales, o bien las que se denominan probabilidades de transmisión. La estructura de la red se obtiene directamente del pedigrí, mientras que la probabilidades condicionales de la herencia se definen por las leyes de Mendel (Bateson and Mendel 2013). Al igual que para el nivel poblacional, el modelo general es una simplificación a partir del cual se extienden distintas correcciones con el fin de contemplar eventos que puedan afectar significativamente la evaluación de la evidencia. En este nivel, las mutaciones constituyen un evento central a tener en cuenta.

3.2.4.1 Del pedigrí a la red bayesiana

En consonancia con Allen et al (Allen and Darwiche 2008), se consideran dos variables aleatorias, $g_{i,j}^p$ y $g_{i,j}^m$. Estas permiten modelar los valores alélicos maternos y paternos para el marcador j de la persona i . Para los individuos no fundadores, dos variables adicionales, denominadas selectores, $s_{i,j}^p$ y $s_{i,j}^m$, son consideradas. Estas son utilizadas para describir la herencia paterna o materna de los genotipos. El pedigrí familiar es utilizado como estructura para definir el Grafo Acíclico Dirigido (GAD), que codifica la influencia probabilística directa entre las variables conectadas. Utilizando esta información, para cada variable aleatoria, se computan

las tablas de probabilidad genotípica condicionada. Estas son utilizadas para calcular la verosimilitud de los genotipos de interés. En la Figura 3.3 se ejemplifica la construcción de una red bayesiana a partir de un pedigrí sencillo (con madre, padre e hijo). Se analizan dos marcadores genéticos, 1 y 2. Puede verse que para el padre y la madre hay dos nodos por cada marcador, representando sus respectivos alelos de herencia materna y paterna. Para el hijo, el alelo paterno del marcador 1 posee como ancestros el alelo materno y paterno del marcador 1 de su padre, además de la variable selectora s_1^p . Dicha variable indicará la probabilidad de que el estado del alelo paterno del hijo sea igual al estado del alelo paterno del padre o la madre (ambas probabilidades sumarán 1). En los casos analizados en el contexto forense $s_{i,j}^p$ y $s_{i,j}^m$ valdrán 0,5, asignando la misma probabilidad de herencia del alelo materno y paterno de los individuos. Lo mismo que es explicado para el alelo paterno del marcador 1 del hijo puede aplicarse al alelo materno del mismo marcador, o a ambos alelos del marcador 2. De este modo, es esperable que en un pedigrí exista una cantidad de grupos de nodos interconectados igual al número de alelos presentes (dos por el número de marcadores).

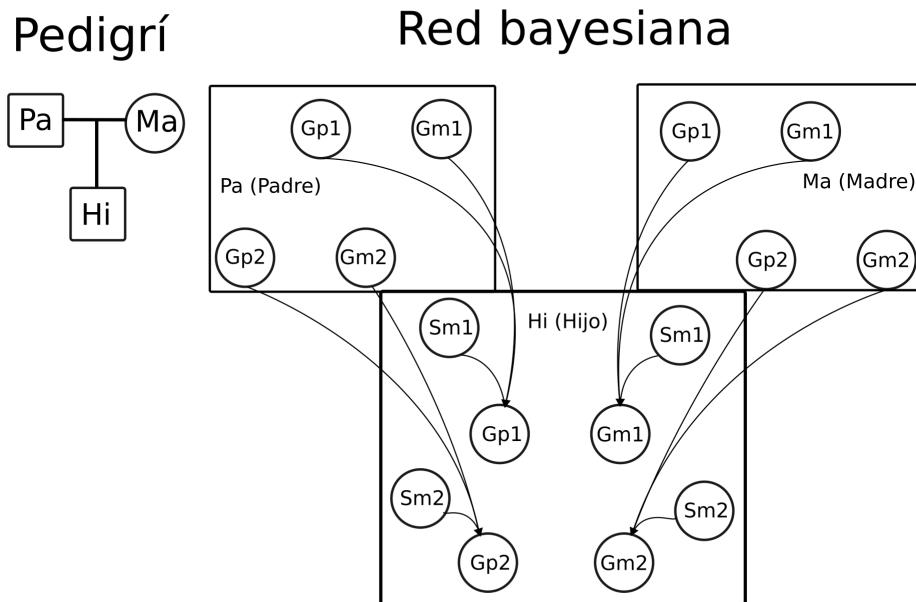


Figura 3.3 Red bayesiana de un caso simple con madre (Ma), padre (Pa) e hijo (Hi). El pedigrí utilizado para la construcción de la red se presenta a la izquierda. Cada nodo representa un alelo, y los conectores muestran las relaciones de dependencia condicional.

Para definir las probabilidades de transmisión (parametrización de la red), sea g_A el genotipo del individuo A. Si A es un no fundador del pedigrí, con padre F y madre M, se define la probabilidad de transmisión de g_A como

$$TR(g_A) = P(g_A|g_F, g_M, \Theta). \quad (3.8)$$

Los componentes de Θ relevantes para la transmisión incluyen el modo de herencia, las tasas de mutación/error y el pedigrí. En el caso más simple, sin mutaciones ni errores, $TR(g_A)$ sigue directamente las leyes de Mendel.

3.2.4.2 Mutaciones

Para describir las probabilidades de mutación, se utiliza la siguiente notación: se asume que hay n alelos posibles en el marcador, y que una mutación cambia el alelo del tipo i al tipo j . Se define a m_{ij} como la probabilidad de que el alelo i sea transmitido al hijo como el alelo j . Así, m_{ii} es la probabilidad de que el alelo i no mute al ser transferido de un ancestro a su descendencia directa. Se utiliza R para denotar la tasa de mutación y se asume que la probabilidad de mutación es $m_{ii} = 1 - R$. Los valores de m_{ij} para $i \neq j$ especifican las probabilidades de mutación particulares mencionadas anteriormente. Se pueden recopilar todos estos valores en la matriz de mutación que especifica el modelo mutacional.

$$\begin{pmatrix} m_{11} & m_{12} & m_{13} & \cdots & m_{1n} \\ m_{21} & m_{22} & m_{23} & \cdots & m_{2n} \\ m_{31} & m_{32} & m_{33} & \cdots & m_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \cdots & m_{nn} \end{pmatrix}$$

Nótese que cada fila representa la probabilidad del estado del alelo i en la descendencia, por lo que estos valores deben ser no negativos y sumar 1. Es decir, para cada i , se tiene que $m_{i1} + m_{i2} + \cdots + m_{in} = 1$. Encontrar una matriz de mutación que represente de forma exacta la probabilidad de que un alelo transicione a otro durante la descendencia es muy difícil. Esto puede depender de muchísimos factores, desde genéticos hasta ambientales. Por este motivo es necesario utilizar modelos paramétricos de mutación, donde se explicitan los mecanismos de mutación de base para construir la matriz.

3.2.4.3 Modelo mutacional uniforme

Este primer modelo es el más sencillo y consiste en asumir que la probabilidad de mutación es igual para todos los alelos. Por este motivo el modelo recibe el nombre de "uniforme". Su matriz de mutación es:

$$\begin{pmatrix} 1 - R & R/(n - 1) & \cdots & R/(n - 1) \\ R/(n - 1) & 1 - R & \cdots & R/(n - 1) \\ \vdots & \vdots & \ddots & \vdots \\ R/(n - 1) & R/(n - 1) & \cdots & 1 - R \end{pmatrix}$$

donde R es la tasa de mutación y n es el número total de alelos posibles. El modelo también se puede expresar como

$$m_{ij} = \begin{cases} 1 - R, & \text{si } i = j, \\ R/(n - 1), & \text{si } i \neq j. \end{cases}$$

El modelo *uniforme* es adecuado para situaciones en las que no se sepa que un tipo de mutación es más probable que otro. Por ejemplo, este modelo es razonable para marcadores

SNPs. Sin embargo, para marcadores STRs, se conoce cómo suceden las mutaciones. Más aún, se sabe que algunos tipos de mutación son más probables que otros. Para tales marcadores, en su lugar, se debe considerar el modelo que se describe en la siguiente sección.

3.2.4.4 Modelo mutacional de a pasos

Los alelos *STR* consisten en una serie de repeticiones de una secuencia corta de nucleótidos. Es decir, para un marcador STR el alelo 9 tiene 9 repeticiones de una secuencia de, por ejemplo, 3 nucleótidos, mientras que el alelo 10 tiene 10 repeticiones de la misma secuencia. Cuando se copian durante la meiosis transfiriendo el alelo de padre a hijo, puede ocurrir una especie de “*desplazamiento*”. Este error de copia consiste principalmente en que se agrega o elimina generalmente una repetición. Con menor frecuencia, se agregan o eliminan dos o más. El modelo *de a pasos* busca reflejar estas diferencias.

Una forma sencilla de construir el modelo es decir que una adición o sustracción de $k + 1$ repeticiones es r veces más probable que una adición o sustracción de k repeticiones, para cualquier $k > 0$, donde $r < 1$ es un parámetro, llamado “*rango mutacional*” (Egeland et al. 2015). Considerando esto, la matriz de mutación es la siguiente:

$$\begin{pmatrix} 1 - R & k_1 r_1 & k_1 r_2 & \dots & k_1 r_{n-1} \\ k_2 r_1 & 1 - R & k_2 r_1 & \dots & k_2 r_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ k_{n-1} r_1 & k_{n-1} r_2 & k_{n-1} r_3 & \dots & 1 - R \end{pmatrix}$$

Aquí, R es la tasa de mutación y k_1, k_2, \dots, k_n son constantes elegidas de tal manera que cada línea suma 1. También se puede expresar como:

$$m_{ij} = \begin{cases} 1 - R, & \text{si } i = j \\ k_i r^{|i-j|}, & \text{si } i \neq j \end{cases}$$

3.2.5 Modelos observacionales

Los modelos observacionales dan cuenta de la probabilidad de observar los datos, dado el genotipo, considerando distintos errores experimentales, o biológicos. Como se ha introducido previamente, estos consisten en el Drop-out y el alelo silente. Los mismos son eventos de baja frecuencia, cuyos modelos se presentan a continuación.

3.2.5.1 Drop-out

Cuando la prueba genética se basa en pequeñas cantidades de ADN o en ADN de baja calidad, puede haber una probabilidad no despreciable de errores en la observación de genotipos erróneos. Por ejemplo, un alelo presente en el genotipo puede no ser observado en absoluto, o incluso puede haber casos de inserciones, donde se observan alelos que no están presentes en el genotipo, debido a problemas en el proceso de reacción en cadena de la polimerasa o

contaminación de ADN. El modelo se resume a continuación: cada alelo tiene una probabilidad fija, llamada d , de no ser observado. Si el genotipo observado es heterocigota, por ejemplo a/b , el verdadero genotipo debe ser heterocigoto a/b . Sin embargo, si el genotipo observado es $a/-$, el resultado es compatible con tres genotipos posibles a/a y a/b . Por lo tanto, las probabilidades de observar $a/-$ dado cada uno de los genotipos se resumen a continuación:

$$P(a/-|a/a) = 1 - d^2$$

$$P(a/-|a/b) = (1 - d)d$$

$$P(a/-|b/a) = d(1 - d)$$

3.2.5.2 Alelo silente

En este modelo simplificado, se considera que un homocigota con genotipo a/a y un heterocigoto con genotipo a/S , siendo S un alelo silente, generan datos indistinguibles que se denotan como $a/-$ (Egeland et al. 2015). Claramente, si los datos observados corresponden a un heterocigoto a/b , el verdadero genotipo debe ser a/b . En términos de cálculo de verosimilitudes, $a/-$ consistiría en los tres posibles genotipos ordenados a/a , a/S y S/a para cada observación $a/-$. De esta manera, se agrega una frecuencia para el alelo silente S , y se trata como un alelo más en la base de datos de referencia.

3.2.6 La expresión general de verosimilitud

En este punto se integran todos los niveles previos, que según los parámetros que se hayan seleccionado, considerando o no mutaciones, drop-out, estructuración poblacional, etc, se verán afectadas tanto las probabilidades genotípicas de los fundadores, como las de los no-fundadores.

Sea G el conjunto de datos de genotipo para los miembros de un pedigree, Pe . Se busca calcular la probabilidad $P(G|Pe)$. Para ser precisos, la probabilidad de G generalmente depende de varios factores además de Pe , que incluyen frecuencias alélicas, el modo de herencia, tasas de mutación, estructuración poblacional, entre otras. Todos estos factores en conjunto se denotan con σ , la probabilidad que buscamos es entonces:

$$P(G|Pe, \sigma) \tag{3.9}$$

Generalmente se suprime σ de la notación, a menos que algunos de sus componentes sean particularmente relevantes para la discusión.

Los cálculos de verosimilitud más simples son aquellos en los que cada miembro de Pe tiene un solo marcador. En tales casos, solo se requieren dos componentes:

- Probabilidades de fundadores, es decir, las probabilidades de observar los genotipos de los fundadores.
- Probabilidades de transmisión de los padres a cada hijo.

La probabilidad de los fundadores se define por el modelo poblacional, y la probabilidad de transmisión por el modelo del pedigrí.

En general, para un pedigrí cuyos fundadores están etiquetados como 1, 2, ..., n_0 , y los no fundadores como $n_0 + 1, \dots, n$, la probabilidad (dado un solo marcador) puede escribirse como

$$P(G|Pe) = \sum_{g_1 \in G_1} \dots \sum_{g_n \in G_n} P(g_1) \dots P(g_{n_0}) TR(g_{n_0+1}) \dots TR(g_n) \quad (3.10)$$

Donde G_i es el conjunto de posibles genotipos para el individuo i , siendo g_i un genotipo específico. Cada término de la ecuación es usualmente fácil de calcular (considerando los distintos fenómenos y las ecuaciones de verosimilitud previamente presentadas). El problema surge dado que la cantidad de términos crece exponencialmente con n . Para exemplificar esto, se denomina $|G_i|$ el número de elementos en G_i . En general, cada G_i puede contener el conjunto completo de genotipos en el locus. Con L alelos, esto equivale a $|G_i| = \gamma$, donde

$$\gamma = \frac{1}{2} (L^2 + L) \quad (3.11)$$

Por lo tanto, la fórmula de verosimilitud puede tener hasta γ^n términos. Como se ha mencionado, distintas estrategias pueden utilizarse para disminuir la cantidad de términos. Una de estas es la eliminación por variables, y depende del orden en la cual las mismas se van quitando de la red. Este proceso es explicado en detalle por Darwiche et al. (Darwiche 2009, Allen and Darwiche 2008).

3.3 Resultados

En esta sección se presentan los resultados de un conjunto de ejemplos utilizados para mostrar la evaluación de la evidencia genética. Los ejemplos buscan resaltar las características metodológicas, dejando el tratamiento de ejemplos más realistas para los próximos capítulos.

3.3.1 Estructura de la red

En esta primera sección se analiza la estructura de la red producida para dos pedigríes, cada uno con tan solo un miembro genotipado. Ambos pedigríes varían en complejidad.

Primero se analiza el caso presentado en la Figura 3.4. En el mismo se presenta un pedigrí con tres individuos: madre, padre e hijo. En rayas se muestra a la madre, que es la única genotipada. Siguiendo la nomenclatura presentada la sección métodos de este capítulo, el pedigrí contiene dos fundadores (individuos 1 y 2) y un no-fundador, cuyas probabilidades genotípicas estarán condicionadas por los genotipos parentales. A partir del pedigrí se construye la red bayesiana. La misma se presenta en la Figura 3.5. Pueden verse cuatro estructuras por separado, y en cada una se presentan cuatro nodos o variables, y sus respectivos conectores. El primer número del código simboliza al individuo del cual refiere el alelo, a continuación se expresa el marcador y por último la procedencia, p es paterno y m materno. En el centro de cada una de las cuatro estructuras se ubica el individuo 3, o sea el hijo. Dos corresponden al marcador $M1$, y otras dos al marcador $M2$. Como es de esperar, para cada marcador un nodo refiere al alelo de procedencia paterna, p y otro al de procedencia materna m . El cuarto nodo de cada una de las estructuras refiere a la variable selector.

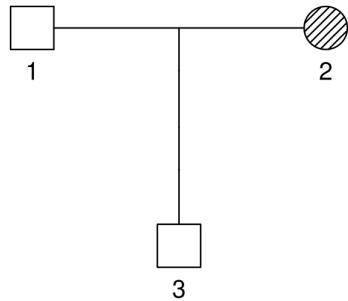


Figura 3.4 Pedigrí donde se presenta a un parente (1), una madre (2) y un hijo (3). La madre (en rayas) se encuentra genotipada.

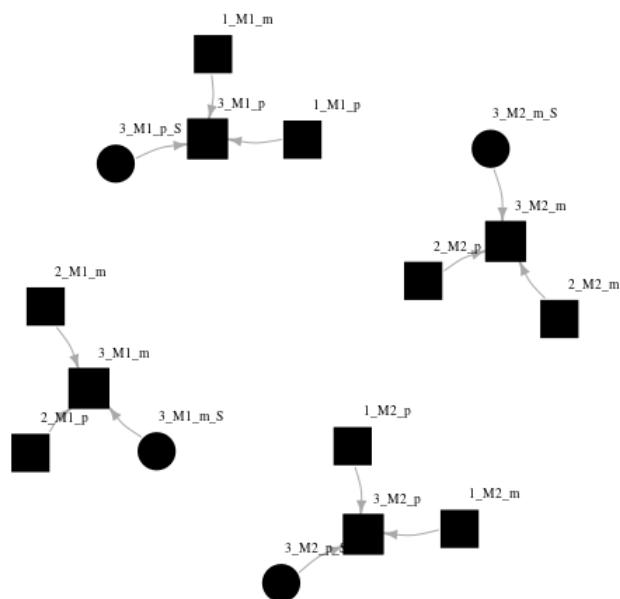


Figura 3.5 Red bayesiana que surge a partir del análisis del pedigrí presentado en la Figura 3.4. En cada nodo, el nombre indica: individuo marcador alelo. Por ejemplo, 1_M1_p indica el individuo 1, marcador 1, alelo paterno. Cuando aparece S indica la variable selectora.

Por otro lado, se analiza un caso más complejo, donde la estructura del pedigrí llega hasta los bisabuelos, siendo un tío abuelo el único genotipado (Figura 3.6).

En este ejemplo, los individuos 1, 2, 4 y 6 son los fundadores, mientras que 3, 5, 7 y 8 son no fundadores. Supóngase que el caso trata de identificar al individuo 7 a partir de la información genética de 8. El individuo 7 no es descendiente directo de quien se encuentra genotipado. Los fundadores, 1 y 2, se encuentran mediando el vínculo de dichos individuos. Este tipo de estructuras llevan a redes más complejas, como se muestra en la Figura 3.7.

Pueden verse un total de seis conjuntos de nodos. Por los nombres puede evidenciarse que en el pedigrí se analizan un total de 3 marcadores, denominados D21S11, D8S1179 y D7S820. Las tres estructuras más simples colocan a los alelos de procedencia materna de estos tres marcadores correspondientes al individuo 7. En todos los casos dicho nodo es conectado por nodos

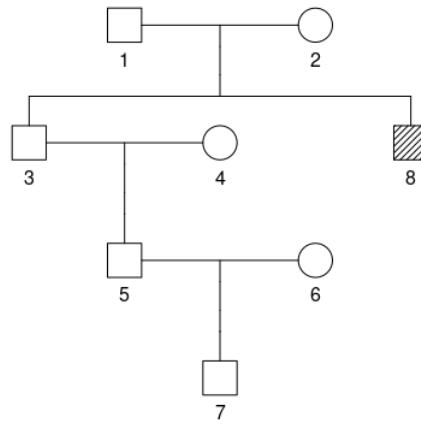


Figura 3.6 Pedigrí donde se indica un tío abuelo (8) genotipado. El individuo buscado es el 7.

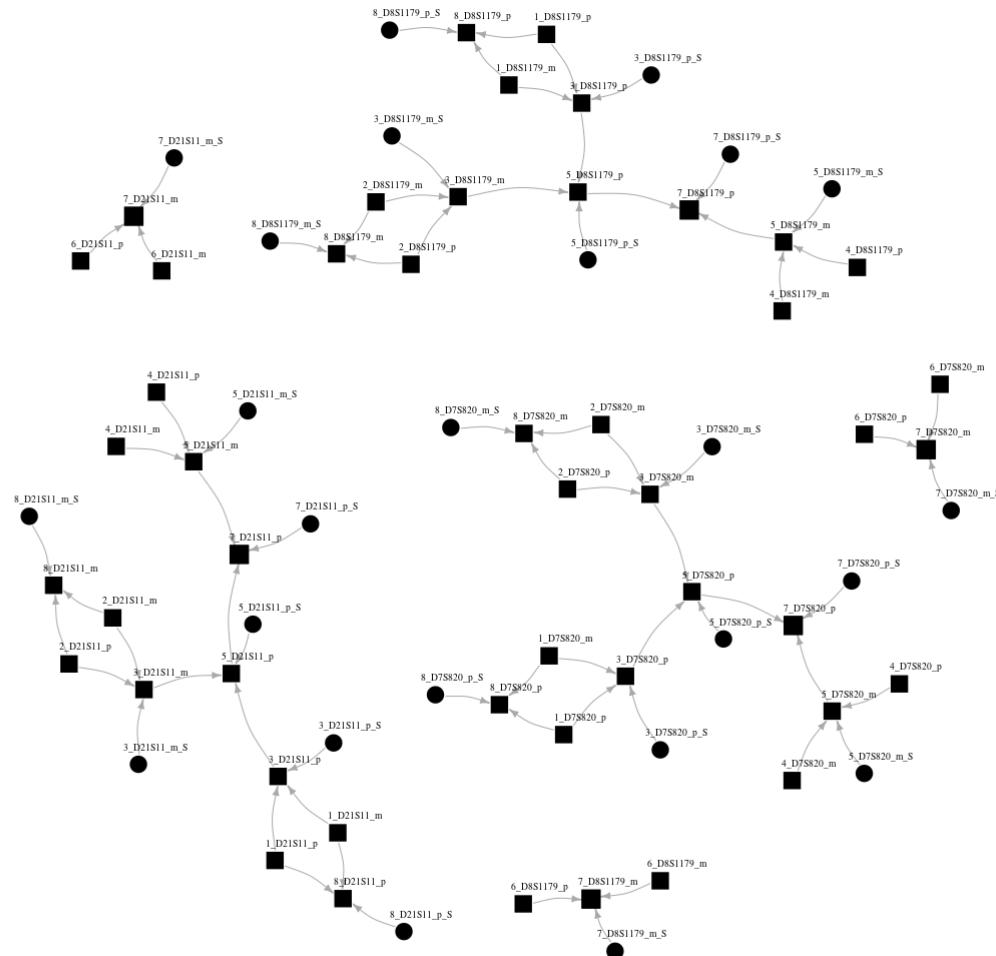


Figura 3.7 Estructura de la red bayesiana del caso del pedigrí presentado en la Figura 3.6. En cada nodo, el nombre indica: individuo marcador alelo. Cuando aparece S indica la variable selectora.

que representan a la madre (individuo 6), indicando el mismo marcador, de procedencia materna o paterna. Un cuarto nodo se refiere a la variable selectora. Esto es lógico, debido a que desde la perspectiva de 7, su madre es una fundadora, escenario similar al de la Figura 3.5. Un panorama muy diferente se ve para las otras tres estructuras, dado que el padre, 5, no es un miembro fundador del pedigrí, y se cuenta con mucha información de su procedencia. Al seguir cualquiera de las estructuras más complejas, por ejemplo la de D21S11, puede verse cómo el alelo paterno del individuo 7 es condicionado por los alelos paterno y materno del individuo 5, o sea su padre, y por una variable selectora. El alelo materno de 5 está solo condicionado por los alelos paterno y materno de 4 y la variable selectora. Mientras que los alelos de 4 no son condicionados, esto es esperable dado que se trata de una fundadora. Por vía paterna, 5 es condicionado por los alelos de 3, que a su vez son condicionados por los de 1 y 2 (bisabuelos, fundadores). Los individuos 1 y 2 condicionan a su vez a los alelos de 8 (tío abuelo). De esta manera queda claro como la red describe el entramado de relaciones familiares. Como se ha mencionado previamente, una vez establecida la estructura, es necesaria la parametrización, es decir, la asignación de valores a las probabilidades condicionadas. En la siguiente sección se volverá a un ejemplo más sencillo para el análisis de probabilidades genotípicas.

3.3.2 Análisis de probabilidades alélicas condicionadas

Una vez producida la red, quedan claras las relaciones condicionales presentes en el pedigrí. Esto permite la construcción de probabilidades alélicas condicionadas que darán probabilidades genotípicas condicionadas. Para analizar esto se toma al ejemplo ilustrado en la Figura 3.4. Se asumen los modelos generales, sin extensiones. Es decir, para el modelo poblacional, se cumple el equilibrio de Hardy-Wainberg, habiendo ausencia de estructuración y con frecuencias alélicas por encima de la mínima. En el modelo del pedigrí no se consideran mutaciones, por lo tanto la probabilidad de transmisión del genotipo materno es igual a la del paterno, 0,5 en ambas. Las probabilidades de drop out y alelo silente también son cero. Los genotipos de los individuos son los siguientes:

Individuos	<i>M1</i>	<i>M2</i>
1	-/-	-/-
2	D/D	B/C
3	-/-	-/-

Donde el único individuo genotipado es la madre. Las frecuencias poblacionales de ambos marcadores son las siguientes:

Marcador	A	B	C	D
<i>M1</i>	0,1	0,2	0,3	0,4
<i>M2</i>	0,24	0,26	0,23	0,27

Las probabilidades alélicas de los fundadores, 1 y 2, provendrán de la población de referencia para el padre bajo el modelo poblacional básico, y estarán definidas con certeza para la madre, dado que las probabilidades de alelo silente y drop out son iguales a cero. Las probabilidades para el hijo se definen a partir del modelo de herencia básico, y se muestran en la Figura 3.8 para el marcador 1.

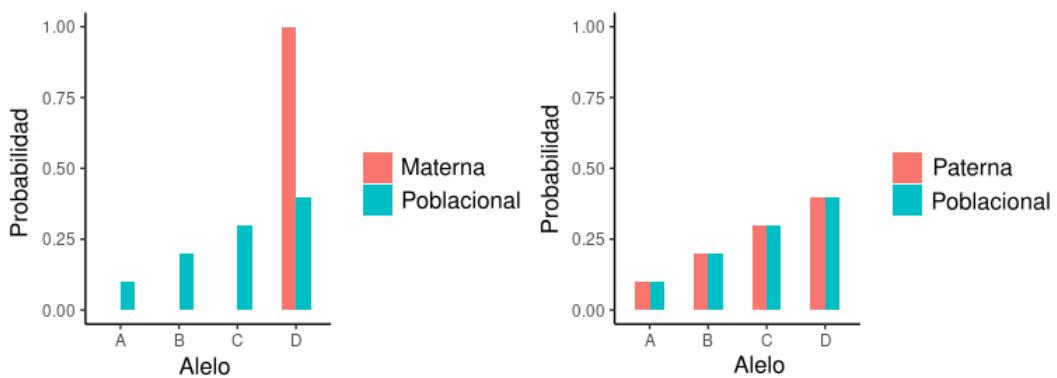


Figura 3.8 Probabilidades alélicas condicionadas para el individuo 3, marcador $M1$. A la izquierda se comparan la probabilidades para el alelo materno, tanto la condicionada por el pedigree como la poblacional. A la derecha se comparan las probabilidades del alelo paterno con la poblacional.

Al comparar la probabilidad condicionada del alelo materno, con la de la población general, se observa una diferencia marcada. Por vía materna, el individuo 3 solamente puede heredar el alelo D. En cambio, por vía paterna todas las probabilidades alélicas son iguales a la de la población de referencia. Esto se debe a la ausencia de evidencia genética en el padre. Lo mismo sucede para el marcador $M2$ (Figura 3.9).

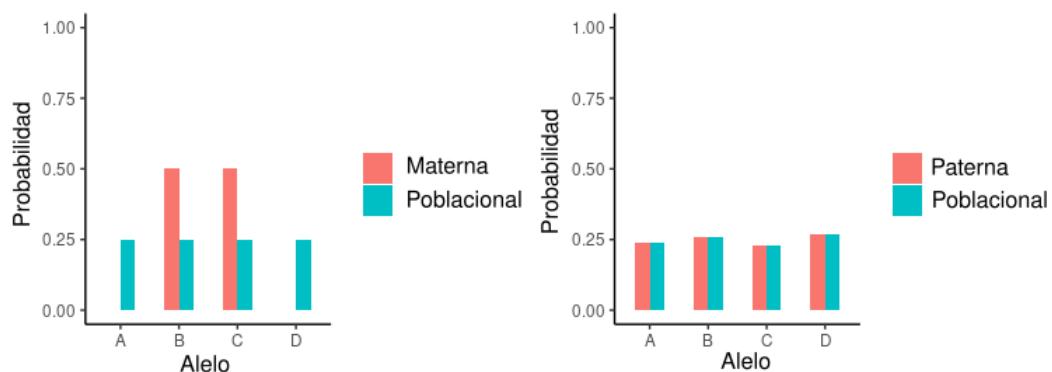


Figura 3.9 Probabilidades alélicas condicionadas para el individuo 3, marcador $M2$. A la izquierda se comparan las probabilidades para el alelo materno, tanto la condicionada por el pedigree como la poblacional. A la derecha se comparan las probabilidades del alelo paterno con la poblacional.

En este caso, por vía materna, el individuo 3 puede heredar los alelos B y C con igual probabilidad. Las distribuciones de probabilidades alélicas son el insumo para el cálculo de las probabilidades genotípicas.

3.3.3 Probabilidades genotípicas y cociente de verosimilitud

En este caso, utilizando los mismos ejemplos que en la sección anterior, se computan las probabilidades genotípicas para el individuo 3, y los cocientes de verosimilitud para cada uno de los posibles genotipos. En la Figura 3.10 se muestran los resultados para $M1$. Puede verse

cómo el panel de probabilidades condicionadas permite solamente genotipos que contemplen un alelo materno D, mientras que el resto son todos permitidos (probabilidad diferente de cero). Más aún, el resto continúan con la probabilidad poblacional, siendo $P(d/i) = P(i)$, con $P(G)$ como la probabilidad del genotipo y $P(i)$ como la probabilidad de heredar el alelo i por vía paterna. El genotipo d/d , para el marcador M_1 , es el más frecuente en la población. Por otro lado, al analizar el CV_G , puede verse que todos los genotipos que poseen el alelo d por vía materna, cuentan con el mismo $CV_G = 2,5$. El resto de las opciones poseen un $CV_G = 0$. Esto es consistente con el hecho de que la única condición que rige sobre los genotipos de 3 es la necesidad de tener un alelo d por herencia materna.

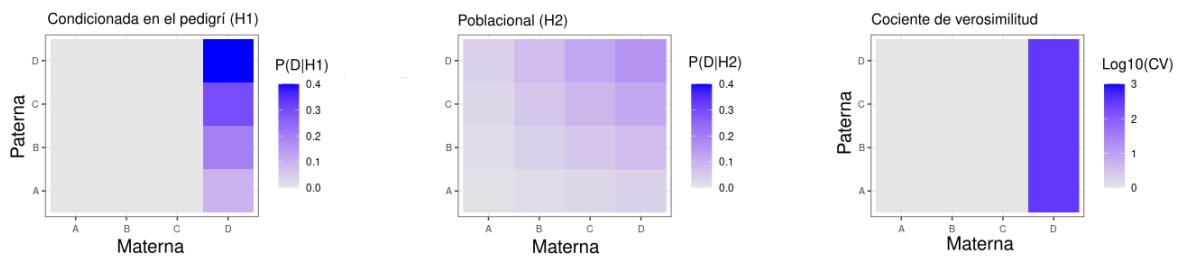


Figura 3.10 Probabilidades genotípicas del marcador M_1 para el individuo 3 del pedigrí presentado en la Figura 3.4. A la izquierda se presentan las condicionadas por el pedigrí, en el centro las poblacionales y a la derecha el $\text{Log}_{10}(CV)$.

A continuación (Figura 3.11) se presenta el mismo análisis para M_2 . A diferencia del caso anterior, dos alelos son posibles de ser heredados por vía materna, el b y el c , ambos con la misma probabilidad. Esto deriva en que en el panel donde se presenta la probabilidad condicional, las dos columnas que hacen referencia a estos alelos tengan las mismas probabilidades. Por otro lado, en el panel poblacional se observa una mayor homogeneidad que para el marcador M_1 . Esto se debe a que las frecuencias alélicas en la población de referencia son más similares. Por último, y consistente con el análisis anterior, cualquier genotipo que contenga los alelos b o c por vía materna tendrá el mismo $CV_G = 2$.

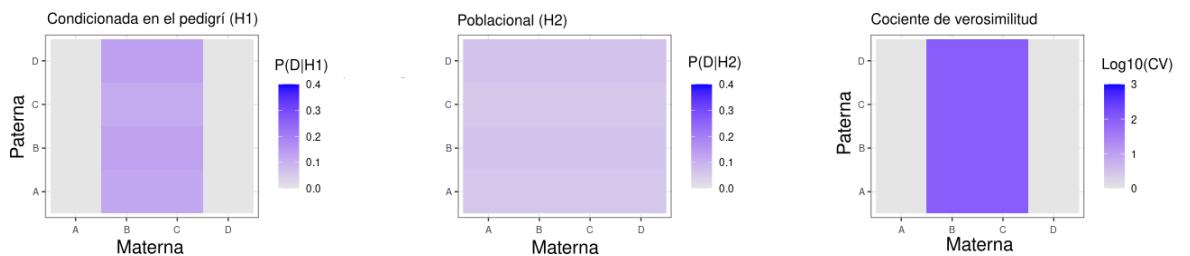


Figura 3.11 Probabilidades genotípicas del marcador M_2 para el individuo 3 del pedigrí presentado en la Figura 3.4. A la izquierda se presentan las condicionadas por el pedigrí, en el centro las poblacionales y a la derecha el $\text{Log}_{10}(CV)$.

Intuitivamente, podría pensarse que el marcador M_1 presenta mejores condiciones para la identificación de individuos, porque permite excluir (asignar probabilidad igual a cero) a 12 de 16 genotipos. Mientras que el marcador M_2 excluye solo a 8 de 16 genotipos. Este concepto será analizado de forma general en el capítulo 4, con el análisis del poder estadístico, y en mucha mayor profundidad en el capítulo 5, con las medidas de información.

Debido a que los marcadores son independientes, los CV_G pueden combinarse multiplicándose. Esto da un total de 256 posibles combinaciones genotípicas. A medida que aumentan los marcadores, el tamaño de la tabla de CV_G crece, dificultando una representación directa. Aún así, se podría calcular el CV_G promedio para una identificación en este caso que será:

$$\langle CV_G \rangle = CV_{M1} \cdot CV_{M2} = 2 \cdot 2,5 = 5$$

El promedio se realiza sobre todos los genotipos posibles de cada marcador.

3.3.4 Múltiples marcadores

En esta sección se analizan los valores esperados de CV_G para 23 marcadores STRs, utilizando un modelo mutacional uniforme, con una tasa R de 0,002. Además, la probabilidad de drop-out y alelo silente son iguales a cero. La base de datos de referencia considerada es la de Argentina, en la misma se considera ausencia de estructura poblacional (Borosky et al. 2014). La metodología utilizada se basa en simulaciones, estas son no condicionadas y siguen el mismo procedimiento que el indicado en el capítulo 2 para las variables no-genéticas (Kling et al. 2017). El mismo se basa en la asignación aleatoria de alelos, considerando las frecuencias poblacionales, en el pedigrí de referencia, sin incorporar la información previamente disponible. En el capítulo siguiente se hará especial énfasis en distintos tipos de simulaciones de datos genéticos, por este motivo no se aborda el tema en detalle en esta sección.

El objetivo es inspeccionar distintas configuraciones de pedigríes y sus posibles valores. Para esto se dividen los pedigríes en subgrupos. Por un lado los del grupo *nuclear* (Figura 3.12) poseen información genética de familiares directos de la persona desaparecida. El subgrupo *hemi* (Figura 3.13), implica a medio-hermanos, el grupo *avuncular* (Figura 3.14) las relaciones tío-sobrino o primos, y el *abuelos* (Figura 3.15) implica un salto generacional. Nótese que a medida que los familiares se alejan de la persona desaparecida, o bien la cantidad de familiares se reduce, disminuye el valor de CV_G esperado en una identificación. En el siguiente capítulo se abordará esta problemática.

3.4 Discusión

La evidencia genética es considerada como el estándar de oro en la identificación forense (Amorim and Budowle 2016). Aún así, como se analizó en el presente capítulo, distintos eventos pueden alterar los modelos básicos, y si no se tienen en cuenta, pueden derivar en errores en la interpretación de la evidencia (Egeland et al. 2015). La extensión de los modelos clásicos ha permitido incorporar distintos fenómenos y considerarlos en la evaluación general. Esta extensión resulta en un incremento de la complejidad del cálculo (Vigeland 2021). Aún así, distintos algoritmos propician la posibilidad de analizar pedigríes complejos en un tiempo razonable (Chernomoretz et al. 2020). Esto ha abierto las puertas a la extensión del uso del ADN como herramienta identificatoria, propiciando las bases para las búsquedas a gran escala (Kruijver et al. 2014).

Cabe destacar que esto último ha tensionado la utilización del material genético para identificación (Kaye 2001). Uno de los problemas centrales se asocia a que errores de baja probabilidad

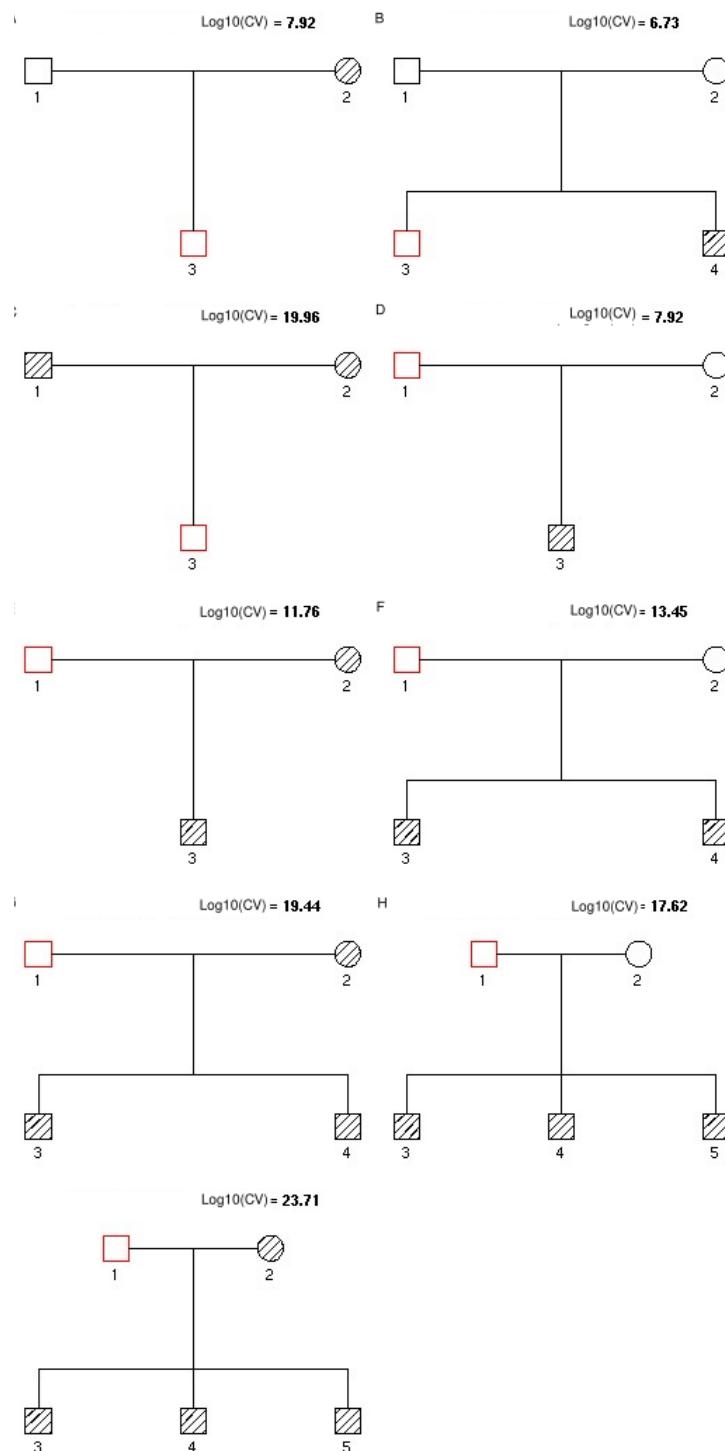


Figura 3.12 Configuración nuclear, resultados con la base de frecuencias de Argentina, 23 marcadores. Los cuadrados representan individuos masculinos, los círculos femeninos, en rojo se indica la persona desaparecida y con rayas los individuos cuyo genotipo está disponible. Por ejemplo, el pedigrí 1 cuenta con un hijo (3), su madre (2) y su padre (1). Solo la madre se encuentra genotipada.

dad, pueden volverse frecuentes en bases de datos de gran tamaño (Marsico et al. 2021, Marsico and Caridi 2023). Esto produjo que se requiera la reconsideración de esquemas conservadores para la evaluación de la evidencia, utilizando umbrales de cociente de verosimilitud altos para

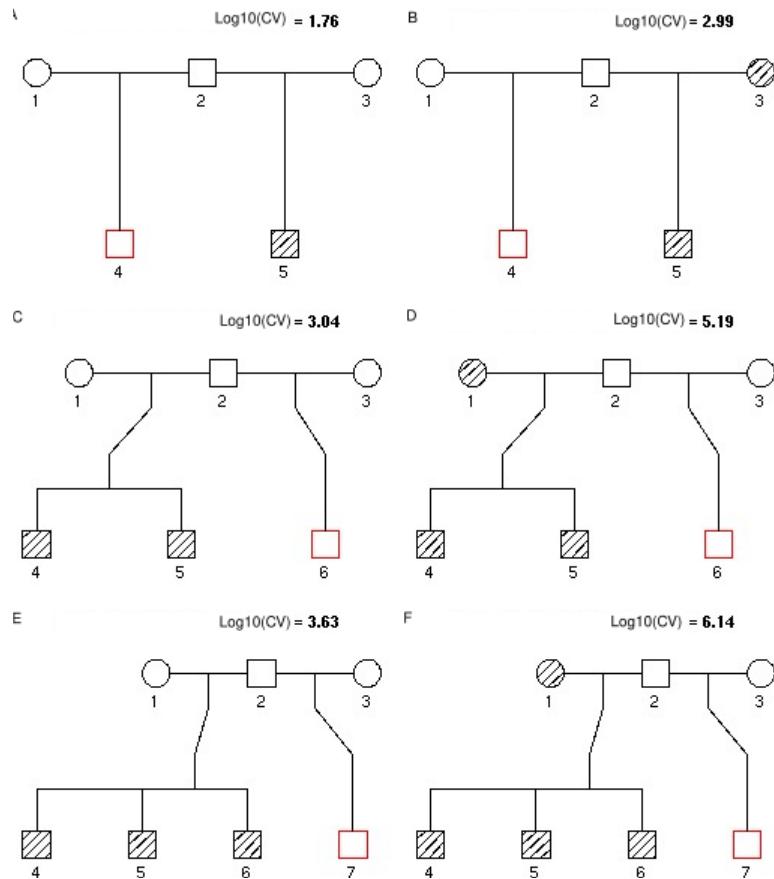


Figura 3.13 Configuración medio-hermanos, resultados con la base de frecuencias de Argentina, 23 marcadores. Los cuadrados representan individuos masculinos, los círculos femeninos, en rojo se indica la persona desaparecida y con rayas los individuos cuyo genotipo está disponible.

definir identificaciones (Amorim and Budowle 2016) y así evitar falsos positivos. Otros autores han esbozado que el ADN constituye una evidencia más, pero debe ser fuertemente contextualizada, y considerada en coherencia junto a otras líneas de evidencia para llegar a conclusiones (Puerto et al. 2021).

En el capítulo se abordó el cálculo del cociente de verosimilitud, las recomendaciones indican que este valor debe ser reportado al Juez (Gjertson et al. 2007). Algunos laboratorios o agencias de investigación, además del cociente de verosimilitud, reportan una probabilidad a posteriori. Esto implica definir un a priori. Como se discutió en el capítulo anterior, este hecho continúa siendo un tema de debate (Budowle et al. 2011, Biedermann et al. 2012).

En los capítulos que restan se analizarán contextos en los cuales la evidencia genética es escasa, dejando múltiples resultados inconcluyentes. Se verán distintas estrategias asociadas a la toma de decisiones, formalizándolas matemáticamente. Además, se contextualizan errores que pueden surgir del uso indebido de la evidencia genética, o de decisiones *ad hoc* frecuentes en los laboratorios forenses.

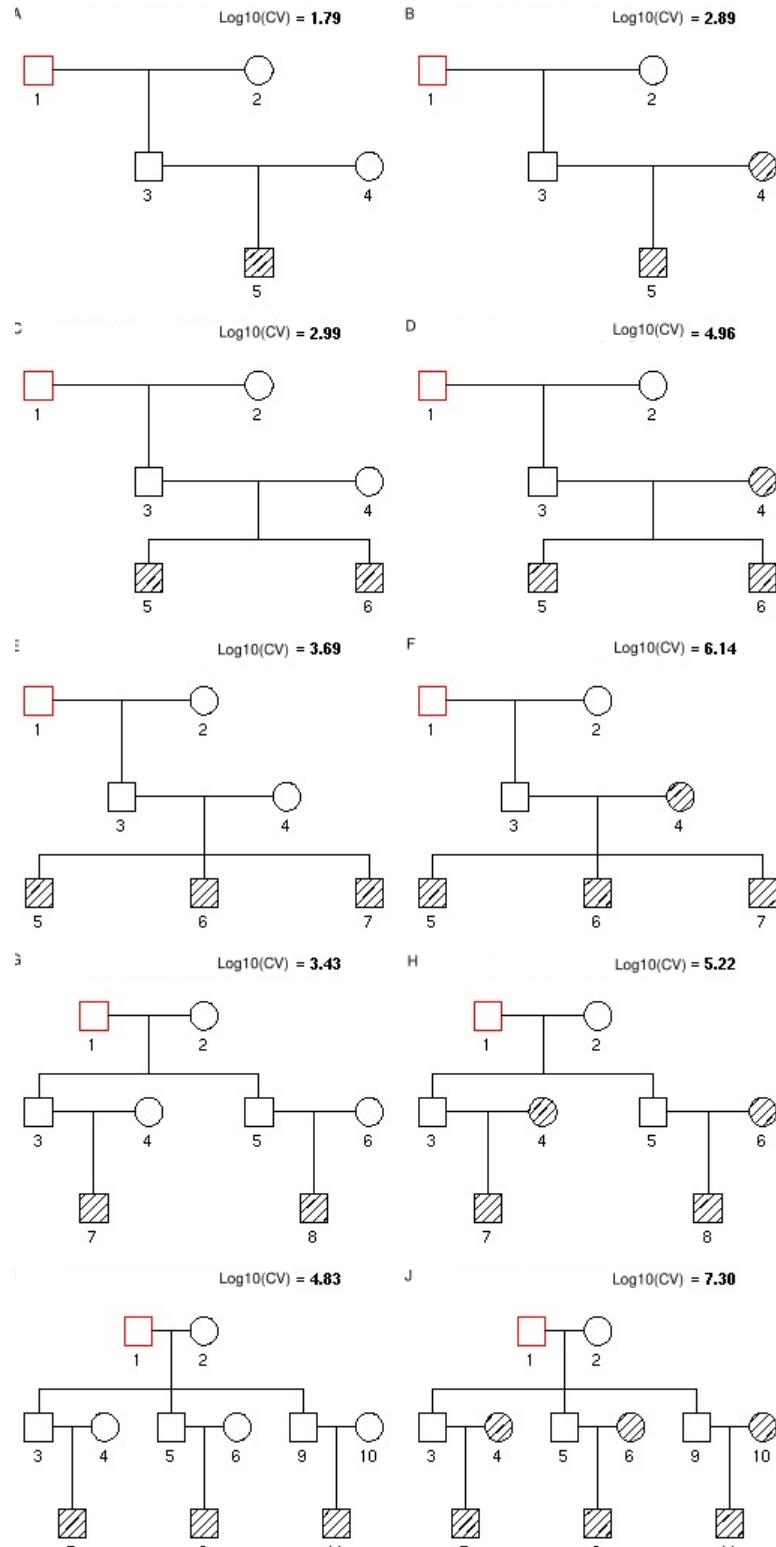


Figura 3.14 Configuración avuncular, resultados con la base de frecuencias de Argentina, 23 marcadores. Los cuadrados representan individuos masculinos, los círculos femeninos, en rojo se indica la persona desaparecida y con rayas los individuos cuyo genotipo está disponible.

3.5 Conclusiones del capítulo

En este capítulo se analizaron distintos modelos de evaluación estadística de la evidencia genética. Más aún, se propone un enfoque diferente a los algoritmos más utilizados en el cam-

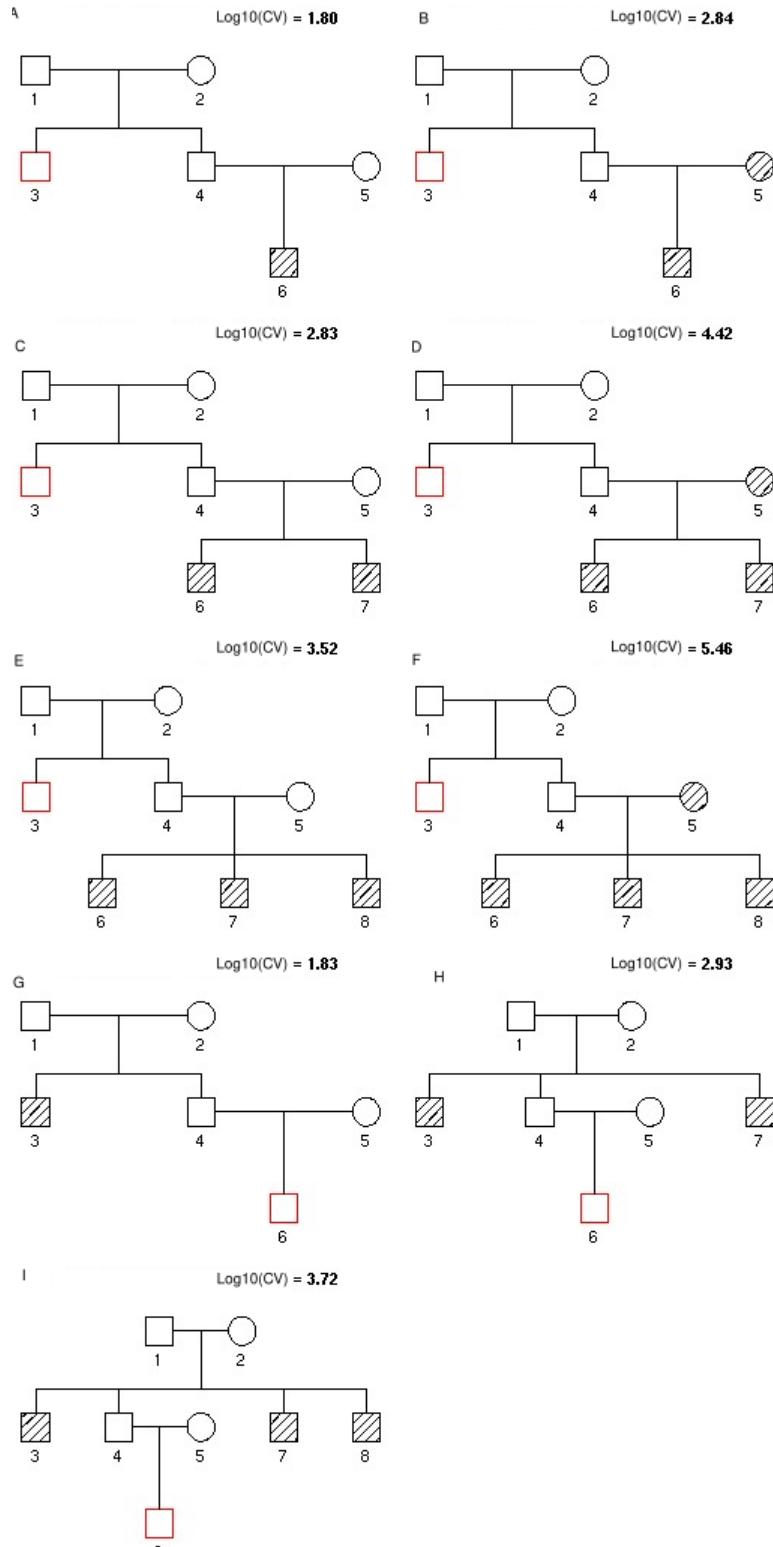


Figura 3.15 Configuración abuelos, resultados con la base de frecuencias de Argentina, 23 marcadores. Los cuadrados representan individuos masculinos, los círculos femeninos, en rojo se indica la persona desaparecida y con rayas los individuos cuyo genotipo está disponible.

po. El mismo fue validado recientemente y es utilizado como motor de cálculo del software GENis, actualmente presente en distintos laboratorios forenses de Argentina y otros países de

latinoamerica (Chernomoretz et al. 2022).

Capítulo 4

El problema de la optimización en la toma de decisiones: selección racional de umbrales para declarar potenciales identificaciones

En este capítulo se aborda la problemática de la toma de decisiones en casos de búsqueda de personas desaparecidas. Particularmente, se analizan búsquedas realizadas mediante el uso de bases de datos, tanto con información genética como también con la recolectada durante la investigación preliminar. A pesar de los avances en los últimos años en las tecnologías de secuenciación del ADN y los *softwares* para análisis de parentesco genético, los científicos forenses afrontan problemáticas cotidianas en la búsqueda. Estas pueden deberse a errores en la tipificación de muestras, presencia de mutaciones en los perfiles genéticos y pedigríes con pocos miembros o alejados de la persona de referencia. Este último aspecto deriva en búsquedas realizadas con poco poder estadístico. Una definición simplificada del poder estadístico en el contexto de la búsqueda de personas desaparecidas es considerarlo como la probabilidad de abordar a una conclusión fehaciente, dadas las proposiciones consideradas y la evidencia evaluada. Donde fehaciente significa la obtención de un resultado concluyente. Por lo tanto, realizar búsquedas con bajo poder estadístico mediante el uso de grandes volúmenes de datos puede derivar en una gran cantidad de resultados erróneos o no concluyentes. En estos contextos, el científico forense se ve forzado a tomar decisiones. Descartar los casos no concluyentes como potenciales candidatos de ser identificaciones puede derivar en la pérdida de identificaciones. Por el contrario, considerar todos los casos no concluyentes como candidatos, y por lo tanto incorporar más análisis y evidencias para llegar a una conclusión, puede derivar en un gran gasto de recursos de distinto tipo (económicos, de tiempo, de recursos humanos, etc). En este contexto, la pregunta que se realiza el científico es *¿cómo seleccionar un subconjunto de casos para los cuales incorporar más análisis?*. Esto implica considerar un balance entre la probabilidad de perder una identificación, y la probabilidad de incorporar análisis a casos que luego podrían descartarse. La propuesta de este capítulo es la utilización de herramientas estadísticas y simulaciones computacionales con el fin de diagnosticar el poder estadístico de los pedigríes involucrados en una búsqueda de personas desaparecidas. Como se explicó en el capítulo 3, en cada análisis de parentesco se obtiene un valor de cociente de verosimilitud genético. Esto da cuenta del peso de la evidencia. En este capítulo se propone que para cada pedigrí es posible obtener valores de cociente de verosimilitud genético esperados para los casos donde el individuo analizado es la persona desaparecida, y otros valores esperados para los casos en los cuales dicho individuo no es la persona desaparecida. También es posible obtener los valores esperados de posterior odds para ambos casos, incorporando los datos de la investigación preliminar con la metodología propuesta en el capítulo 2.

4.1 Introducción al capítulo

La identificación ha sido descrita como el último paso dentro del proceso de búsqueda (Frey 2019). La misma implica reunir y comparar toda la evidencia recopilada con el fin de llegar a una conclusión en torno al caso. Generalmente, la comparación entre PNI y PD puede derivar en: (i) identificación, cuando PNI es PD, (ii) descarte, cuando se concluye que PNI no es PD y (iii) no concluyente, cuando la evidencia recolectada no es suficiente para concluir en ninguna de las dos direcciones previas. Como se ha mencionado, los pasos del proceso de búsqueda se encuentran fuertemente vinculados, haciendo que el resultado de un paso posterior promueva una mayor investigación en pasos previos (Puerto et al. 2021). En los casos no concluyentes, se puede promover más investigaciones dentro de la etapa preliminar o bien en el laboratorio. Esto se realiza para obtener más evidencia con el fin de abordar a una conclusión (Marsico et al. 2021). En contextos donde se utilizan grandes bases de datos para la búsqueda, el trabajo con casos que presentan escasa evidencia (genética y no genética) puede derivar en un gran número de resultados no concluyentes (Puerto and Tuller 2017).

En esta sección se introduce la búsqueda de personas desaparecidas mediante el uso de bases de datos. Además, se resumen diferentes aproximaciones para la evaluación del poder estadístico. Por último, se hace foco en las problemáticas que enfrenta el científico forense en contextos de bajo poder estadístico utilizando bases de datos.

4.1.1 La búsqueda mediante uso de bases de datos

Desde la propuesta del uso de la huella digital genética para la individualización de personas a partir de muestras de ADN (Jeffreys et al. 1985b), la comunidad forense internacional adoptó distintas metodologías para el estudio de las regiones hipervariables del genoma humano (Martin et al. 2001). Distintos laboratorios forenses desarrollaron sus propias herramientas, generando una diversidad que posteriormente tuvo que ser homogeneizada. Particularmente, con el descubrimiento de la reacción en cadena de la polimerasa (PCR) y los marcadores conocidos como STRs (Edwards et al. 1991, Weber and May 1989), la comunidad encontró consensos que permitieron detectar y almacenar la variación genética de los individuos. Aún así, no fue hasta el año 1995 donde se crea la primera base de datos a nivel nacional de inteligencia criminal en Reino Unido (Martin et al. 2001). Luego de 4 años, para 1999, dicha base de datos contaba con perfiles de más de 700.000 individuos. Su éxito en la identificación de personas, primordialmente en el contexto forense, derivó en que otros países adopten políticas similares. En 1999 Países Bajos, Alemania, Austria, Finlandia y Noruega implementaron también bases de datos a nivel nacional. Para ese momento, estas almacenaban información de entre 6 y 7 marcadores STRs. En paralelo, el sistema CODIS (*Combined DNA Index System*) era postulado para homogeneizar la nomenclatura y almacenamiento de los datos genéticos en Estados Unidos (Budowle et al. 1998) mientras el desarrollo de los *software* para el cómputo del test de parentesco se encontraba en marcha (Egeland et al. 1997). Hoy en día, la utilización de bases de datos para la identificación humana ha cobrado gran popularidad, llegándose inclusive a la postulación de bases de datos universales, donde se almacena información de toda la población de un país (Hazel et al. 2018). Más allá de estos planteos, su uso extendido abrió el

campo a nuevas estrategias de identificación, como la búsqueda familiar (Maguire et al. 2014). Los conceptos fundamentales de la búsqueda familiar fueron introducidos en el capítulo 3. La búsqueda consiste en la comparación del ADN obtenido, por ejemplo, de una evidencia en una escena de crimen, contra una base de datos de referencia. A diferencia del enfoque clásico en las identificaciones forenses, donde se busca evaluar si dicho ADN se encuentra en la base de datos, en la búsqueda familiar el abordaje es mediante el test de parentesco genético. Es decir, se pregunta si el ADN en la escena del crimen pertenece a un familiar de alguno de los individuos cuyo ADN se encuentra en la base de datos. Esto permite generar hipótesis de identificación que colaboran con el proceso judicial y policial. Esta estrategia deja al descubierto el potencial uso de las bases de datos genéticas, ampliando el rango de la población a la cual se podría acceder (Haimes 2006, Machado et al. 2020). Por otra parte, la utilización de bases de datos genéticas en el campo de la búsqueda de personas desaparecidas ha crecido en los últimos años (Junior et al. 2022, Laurent et al. 2022). Distintos *software* como el Familias (Kling et al. 2014), Bonaparte (van Dongen et al. 2011), CODIS (Budowle et al. 1998) y GENis (Chernomoretz et al. 2022) cuentan con módulos específicos para este tipo de casos. Además, algunos de ellos incorporan el almacenamiento y resguardo de los datos, tanto genéticos como no-genéticos.

4.1.2 La utilización de bancos de muestras para la reparación de crímenes de lesa humanidad

Como explica Cordner et al. (Cordner and Tidball-Binz 2017), la acción humanitaria forense pone a la Argentina como el país donde nacieron los primeros exponentes de la disciplina. Como acción humanitaria se entiende al quehacer orientado a aliviar el sufrimiento humano y proteger la dignidad de todas las víctimas en cualquier conflicto armado o desastre natural. En este punto, la organización no-gubernamental denominada Equipo Argentino de Antropología Forense se creó como institución pionera en la búsqueda de los desaparecidos durante la última dictadura militar ocurrida en la Argentina, entre 1976 y 1983 (Cordner and Tidball-Binz 2017). Por otro lado, en el mismo período, la organización Abuelas de Plaza de Mayo se centró en la búsqueda de los hijos e hijas de los desaparecidos (Berra et al. 1986). Particularmente relacionado a esta última búsqueda, se constituyó en 1987 el Banco Nacional de Datos Genéticos como el encargado de almacenar la información genética de los familiares de las personas desaparecidas y de individuos cuya identidad biológica se encuentre duditada (Penchaszadeh 1997). Dicha creación constituye el primer registro a nivel internacional de la generación de un archivo sistemático de muestras utilizadas para el perfilamiento genético de individuos con fines identificatorios. Por otro lado, en el año 1991 se creó en Argentina el Servicio de Huellas Digitales Genéticas (SHDG) de la Facultad de Farmacia y Bioquímica de la Universidad de Buenos Aires. Dicho servicio utilizó herramientas de genética molecular para la identificación de individuos en distintos contextos, tanto de catástrofes, atentados, como de desapariciones específicas (Corach et al. 1995). Cabe realizar una distinción entre un banco de datos genéticos y una base de datos genética. Como indican Budowle et al. (Budowle et al. 1998), en el marco forense, el banco de datos es un registro de muestras, y la base de datos con fines identificatorios es la información genética organizada para la ejecución de algoritmos de búsqueda. Es por este motivo que un mismo banco de muestras puede tener distintas bases de datos de información genética,

ya sea por la forma de almacenamiento de la información, por la tecnología de secuenciación genética aplicada o bien por la parte de genoma que se analiza (ADN mitocondrial, cromosoma X o cromosoma Y). Contar con el banco de muestras permite, entre otras cosas, llevar a cabo análisis complementarios al estándar de STRs cuando con el mismo no alcanza para arribar a una conclusión.

4.1.3 El poder estadístico

Como se ha mencionado, el poder estadístico en el contexto del test de parentesco genético puede definirse como la probabilidad de abordar a una conclusión, dadas las proposiciones evaluadas (Kling et al. 2017). En el contexto de búsqueda de personas desaparecidas, las proposiciones evaluadas son H_1 : PNI es PD, y H_2 : PNI no es PD. Generalmente, cada PNI es analizado contra cada pedigrí, que cuenta con parientes biológicos de PD. El test de parentesco genético implica el cálculo del previamente introducido cociente de verosimilitud (CV), que indica el peso de la evidencia genética (Prinz et al. 2007). Una identificación potencial es declarada cuando el cociente de verosimilitud obtenido supera un determinado umbral (Kling et al. 2017). El establecimiento del valor del umbral depende de distintos factores, como el número de marcadores genéticos utilizados, el tamaño de la base de datos, etc (Kruijver et al. 2014). En ciertos casos, donde el poder estadístico es muy bajo, puede suceder que dicho umbral nunca sea alcanzado, inclusive estando frente a una identificación real (es decir PNI es PD). Dos métricas han sido propuestas para la evaluación del poder estadístico (Vigeland et al. 2020). Por un lado el *poder de inclusión* (PI) es la probabilidad de obtener un cociente de verosimilitud por encima del umbral seleccionado, considerando que PNI es PD. Por otro lado, el *poder de exclusión* (PE) es la probabilidad de obtener un cociente de verosimilitud igual a cero, considerando ausencia de mutaciones, cuando PNI no es PD. Importantemente, para la evaluación del poder de exclusión se considera una tasa mutacional igual a cero. Si en un determinado caso se supera el umbral, la potencial identificación es confirmada incorporando distintos análisis genéticos (ADN mitocondrial, cromosoma X, etc) y cotejando con los datos no-genéticos. Esto se realiza con el fin de evitar falsos positivos. En cambio, los casos que quedan por debajo del umbral no suelen ser re-analizados con mayor profundidad, convirtiendo a los falsos negativos en errores silenciosos (Marsico et al. 2021). Silenciosos se refiere al hecho de que no reciben habitualmente un segundo chequeo. Como ha sido mencionado por Kling et al. (Kling et al. 2017), el establecimiento de umbrales altos implica una menor cantidad de falsos positivos, a costo de un aumento de la probabilidad de falsos negativos. Por el contrario, disminuir el umbral puede implicar mayor cantidad de falsos positivos, pero con el beneficio de disminuir la probabilidad de perder una identificación por un falso negativo. El costo que tienen estos errores depende fuertemente del contexto, el tamaño de las bases de datos y los recursos (más marcadores genéticos, información de la investigación preliminar, etc) para concluir en torno a los casos.

4.1.4 Umbral de cociente de verosimilitud

Seleccionar un umbral óptimo para el cociente de verosimilitud ha sido un desafío en muchas aplicaciones forenses. En la práctica, los laboratorios de análisis de ADN seleccionan umbrales *ad hoc*, basados en los tipos de casos con los que se trabaja (búsqueda de personas

desaparecidas, test de paternidad, etc) (Kruijver et al. 2014). En los casos de búsqueda de personas desaparecidas se suelen seleccionar umbrales altos, con el objetivo de evitar falsos positivos (Kling et al. 2017). Esto es particularmente cierto en los análisis masivos, donde muchos PNIs, por ejemplo más de 10.000, son analizados contra muchos PDs, por ejemplo más de 400. En este caso se obtendría un total de 4.000.000 de resultados de cociente de verosimilitud (uno por cada par PD-PNI). Por lo tanto, seleccionar umbrales bajos con una mayor tasa de falsos positivos puede derivar en una cantidad inimaginable de posibles identificaciones. Kling et al. (Kling et al. 2017) han propuesto un umbral para el cociente de verosimilitud genético de 10.000. Para llegar a este valor se evaluaron la ausencia de falsos positivos para un conjunto de pedigríes. El resultado se consiguió mediante la realización de simulaciones computacionales, donde se estudió la distribución obtenida, para un pedigrí dado, de los cocientes de verosimilitud considerando H_1 : PNI es PD o H_2 : PNI no es PD. En el trabajo (Kling et al. 2017) se describe que pedigríes con bajo poder estadístico pueden no llegar a cocientes de verosimilitud mayores al umbral de 10.000, inclusive cuando PNI es PD.

4.1.5 Distribuciones de cociente de verosimilitud

Trabajos previos han explorado distintas estrategias para seleccionar umbrales de cociente de verosimilitud basándose en la distribución de valores obtenidos considerando ciertas ambas hipótesis, H_1 y H_2 (Cho et al. 2017, Li et al. 2019, Laurent et al. 2022). Estos enfoques analizan distintos tipos de parentesco (tío-sobrino, abuelo-nieto, etc) con el fin de determinar umbrales a partir de los cuales declarar una identificación. Como mencionan Slooten et al. (Slooten 2020), se debe hacer una distinción entre las nociones de evaluación del peso de la evidencia y la de toma de decisiones. El cociente de verosimilitud tiene un valor e interpretación por sí mismo. Por lo tanto, el estudio de su distribución para el establecimiento de tasas de falsos positivos y falsos negativos no debe confundirse con la interpretación inicial. A modo de ejemplo, supóngase un caso donde un cociente de verosimilitud en un test de parentesco genético arroja un valor de 100. Esto indica que es 100 veces más probable observar los datos genéticos analizados si H_1 es cierta, que si H_2 lo fuera. Si esas 100 veces son suficientes o no para concluir en torno al caso o para analizar más evidencia es una decisión que queda en manos del encargado y responsable de interpretarla. Si por ejemplo la tasa de falsos positivos para el valor de cociente de verosimilitud de 100 fuera muy baja (0,0001) no indica que dicho valor es suficiente para llegar a una conclusión. Es decir, las tasas de error no indican cuán correcta es H_1 frente a H_2 . Esto sucede debido a que las tasas de error no aplican al cociente de verosimilitud, si no a las decisiones que se toman utilizando al mismo.

4.1.6 Toma de decisiones

Kruijver et al. (Kruijver et al. 2014) analizan las distintas estrategias utilizadas para el establecimiento de umbrales del cociente de verosimilitud para la toma de decisiones en casos de identificación mediante test de parentesco genético. Particularmente, definen cuatro enfoques, que serán introducidos en esta sección y comparadas con la metodología propuesta en el resto del capítulo:

- **Top-N.** La primera estrategia selecciona como potenciales identificaciones a los primeros N casos ordenados de mayor a menor en el valor obtenido del cociente de verosimilitud: La cantidad N dependerá de los recursos del laboratorio o inclusive de consideraciones técnicas. Por ejemplo, Myers et al. (Myers et al. 2011) reportan que el Estado de California implementa una estrategia donde $N = 168$. Esto se debe a que es posible realizar análisis de cromosoma Y utilizando plaquetas que permiten el ingreso de dicho número de muestras.
- **Umbral fijo:** La segunda estrategia propone seleccionar un umbral fijo de cociente de verosimilitud a partir del cual se consideran las potenciales identificaciones. Este tipo de estrategia es descrita por Kling et al. (Kling et al. 2017), donde el umbral seleccionado fue de 10.000.
- **Centrada en los perfiles:** esta estrategia también propone el uso de umbrales de verosimilitud, pero que se establecen para cada pedigrí. En este caso lo que queda fijo entre los pedigríes son las tasas de falsos positivos o de falsos negativos. Como las distribuciones de cociente de verosimilitud varían entre los pedigríes, y las tasas dependen de estas, distintos umbrales son seleccionados (Slooten and Meester 2014).
- **Condisional:** esta estrategia se basa en el establecimiento de umbrales pero sobre los posterior odds (Slooten and Meester 2014). Para este fin, es necesario definir los prior odds. Si los prior odds son uniformes, los resultados obtenidos no diferirá de los planteados con una estrategia de "umbral fijo" para el cociente de verosimilitud. En cambio, con prior odds no-uniformes, los resultados difieren.

En este capítulo, se propone una alternativa a los modelos previamente descritos que permite definir un valor umbral específico para cada pedigrí. Basándose también en las distribuciones del cociente de verosimilitud, se indica un umbral llamado Umbral de Decisión (UD) (Marsico et al. 2021). Específicamente se muestra la utilidad de dicho umbral en casos con bajo poder estadístico, donde es necesario llegar a un balance entre falsos positivos y falsos negativos. Más aún, se incorporan prior odds no uniformes, con el fin de evaluar la estrategia utilizada para el cociente de verosimilitud, sobre el posterior odds. Con el fin de ilustrar la aplicación se plantea un caso hipotético de búsqueda de personas desaparecidas. Más específicamente, se simulan escenarios en los cuales el padre y la madre de la persona desaparecida no se encuentran disponible para ser genotipados. Este escenario es similar al que ocurre en la búsqueda de los nietos de las "Abuelas de Plaza de Mayo" (Penchaszadeh 1997, Gorini 2006), donde los padres de la persona buscada (nieto o nieta) se encuentran desaparecidos.

Generalmente, en búsquedas de personas desaparecidas mediante el empleo test de parentesco genético se cuenta con dos bases de datos genéticas: una para pedigríes de referencia, compuesta por familiares de los desaparecidos, y otra para personas no identificadas, PNIs. Nótese que los PNIs pueden ser restos no identificados de personas, o como en el caso de *Las Abuelas de Plaza de Mayo*, corresponden a individuos cuya identidad biológica se encuentra duditada (Kling et al. 2017).

4.2 Métodos

4.2.1 Base de datos de la investigación preliminar

La base de datos de la investigación preliminar está constituida por distintas fuentes de información recabadas durante el proceso de búsqueda. Se cuenta con dos bases de datos de este tipo, una para PDs, $PDs = \{PD_1, PD_2, PD_3, \dots, PD_N\}$, y otro para PNIs, tal que $PNIs = \{PNI_1, PNI_2, PNI_3, \dots, PNI_K\}$. Siendo $K = 10.000$, el número de PNIs y $N = 300$, el número de PDs. Ambas bases de datos cuentan con información relacionada a tres variables: sexo, S , que puede tomar valores $A_S = \{f, m\}$, color de pelo C , que puede tomar valores $A_C = \{1, 2, 3, 4, 5\}$, y edad, E , que puede tomar valores de 0 a 100. Para PNI se calcula a partir de la fecha de nacimiento, y para PD a partir de la fecha de nacimiento conocida (en caso contar con información precisa) o a partir de la fecha de nacimiento estimada (en caso de no conocerse con certeza). La edad se expresa mediante rangos, siendo E_{min} la cota inferior y E_{max} la superior. Como se ha introducido en el capítulo 2, la variable M_E indica el solapamiento entre el rango de edad de PD y el de PNI . Toma un valor de 1 si hay solapamiento y de 0 si no lo hay.

4.2.2 Datos genéticos

Todas las muestras de PNIs y miembros de los pedigríes son analizadas con 23 STRs autosómicos. Cada marcador cuenta con información de dos alelos, uno de cada cromosoma, para cada individuo. Cuando es necesario para concluir en torno a un caso, se realizan análisis adicionales, como la expansión de 23 a 30 marcadores STRs, análisis de cromosoma X, Y, ADN mitocondrial, entre otros (Vigeland et al. 2020).

4.2.3 Cálculo de prior odds

El prior odds se calcula en función de la metodología descrita en el capítulo 2 de la presente tesis. En este apartado se retoman las ecuaciones fundamentales, haciendo hincapié en los parámetros seleccionados. Las hipótesis contrastadas, tanto en los datos genéticos, como en los no genéticos, son dos: H_1 (PNI es PD) y H_2 (PNI no es PD).

Para definir el prior odds basado en datos de la investigación preliminar es necesario realizar una evaluación estadística del peso de la evidencia. Para ello se aplica la siguiente expresión:

$$O(H | NG) = CV_{NG} \cdot O_{NG}(H) \quad (4.1)$$

Donde $O_{NG}(H)$ es el prior odds, CV_{NG} el cociente de verosimilitud no genético y $O(H|NG)$ el posterior odds no genético. La definición de $O_{NG}(H)$ depende del caso, y múltiples opciones han sido discutidas (Biedermann et al. 2012). En este caso, como se están comparando dos hipótesis, H_1 y H_2 , se selecciona $O_{NG}(H) = 0,5$. Esto se fundamenta en el hecho de que no hay información previa, subjetiva o de evidencia, que incline las probabilidades hacia alguna de las hipótesis, por lo tanto quedan equiprobables. Respecto a CV_{NG} , el mismo se obtiene a partir de la siguiente expresión:

$$CV_{NG} = CV_S \cdot CV_C \cdot CV_E \quad (4.2)$$

Se considera que las variables analizadas son condicionalmente independientes. Las tasas de error son: $\epsilon_S = \epsilon_c = \epsilon_E = 0,01$. Asumiendo que hay un error de 0,01 en el ingreso de datos o algún otro tipo de error que de lugar a que el dato ingresado no sea el real. Los datos poblacionales de las distintas características fenotípicas se obtienen de la base de datos de la población de referencia. Es decir la población del país en que se analiza el caso.

Nótese que el $O_{NG}(H)$ generalmente será uniforme para todos los pares $PD - PNI$, debido a que no hay, a este punto, datos que permitan distinguir a los casos. Cuando se introducen los datos de la investigación preliminar, el posterior odds, $O(H | NG)$, ya no es uniforme para cada par $PD - PNI$.

4.2.4 Análisis de parentesco genético

En este paso, el cociente de verosimilitud refiere a la interpretación de los datos genéticos, y se calcula de la siguiente manera:

$$CV_G = \frac{P(G|H_1, \theta)}{P(G|H_2, \theta)} \quad (4.3)$$

G refiere a la totalidad de los datos genéticos observados y CV_G es el cociente de verosimilitud de los pedigríes analizados en el caso. Como se explica en el capítulo 3, distintos parámetros fijos condicionan la probabilidad de los genotipos dadas las hipótesis. Estos parámetros hacen referencia a propiedades de los marcadores moleculares, de los modelos considerados para la herencia genética, a la población de referencia, entre otras. En este caso se considera un modelo mutacional *uniforme* (Egeland et al. 2015), con una tasa mutacional de 0,002. La razón de la selección del modelo uniforme se basa en su menor costo computacional, permitiendo un cálculo más rápido en la base de datos en comparación a otros modelos de mayor complejidad. Se considera ausencia de estructuración poblacional, con probabilidad de drop-in y drop-out iguales a cero. La base de datos de referencia utilizada es la de Argentina publicada por Borosky et al. (Borosky et al. 2014).

4.2.5 Cálculo de las posterior odds

Como se ha introducido en el capítulo dos, el posterior odds de una instancia puede convertirse en el prior odds de la instancia siguiente de actualización del conocimiento (MacKay et al. 2003). En este caso, el posterior odds no genético es utilizado como prior odds del paso genético. Por lo tanto, $O_G(H) = O(H | NG)$. El posterior odds del paso genético se calcula de la siguiente manera:

$$O(H | G) = CV_G \cdot O_G(H) \quad (4.4)$$

Este posterior combina información de la instancia de investigación preliminar, y de la instancia del análisis genético.

4.2.6 Simulaciones computacionales

Las simulaciones para el cálculo del poder estadístico en el *test* de parentesco genético fueron descritas por Egeland et al. (Egeland et al. 2016). Kling et al. (Kling et al. 2017) presentaron

los cálculos basados en simulaciones condicionadas, actualmente implementados en la librería forrel disponible en el repositorio CRAN (Vigeland et al. 2020). Dichas simulaciones se realizan asumiendo, a menos que sea declarado de otra forma, que los datos genéticos son consistentes, es decir, que los genotipos observados poseen probabilidad diferente a cero en el pedigree de referencia. El enfoque planteado por Egeland et al. (Egeland et al. 2016) permite estudiar la distribución de $P(CV_G|H)$, a partir de simular los datos genéticos de PNI considerando H_1 o H_2 como ciertas. Dado un determinado pedigree, simular los datos genéticos para PNI considerando H_1 cierta implica asumir que PNI es PD, por lo tanto está relacionado con los miembros del pedigree. Una vez realizado, se aplican las leyes de segregación genética que permiten estudiar la distribución de posibles genotipos de PNI. En el caso contrario, cuando H_2 se considera cierta, los posibles genotipos para PNI se obtienen a partir de la base de datos de frecuencias alélicas, aplicando los modelos de genética de poblaciones basados en el equilibrio de Hardy-Weinberg.

Para CV_{NG} se simulan los datos no genéticos, del mismo modo en el cual se introduce en el capítulo 2, para ambas hipótesis. Con los datos no genéticos simulados es posible calcular un posterior odds no genético, asumiendo un prior. El posterior odds no genético sirve como prior odds para el paso genético.

En cada simulación se obtienen perfiles de datos genéticos y no genéticos para 20.000 PNIs. La mitad, 10.000, serán producidos considerando H_1 cierta, la otra mitad lo serán considerando H_2 cierta. Esa cantidad de realizaciones ha demostrado ser suficiente para explorar la distribución de CV_G y CV_{NG} (Kling et al. 2017). De esta manera, con cada PNI generado, se tendrá información genética de 23 marcadores STRs autosómicos, sexo, color de pelo y edad. Estos datos son utilizados para calcular CV_{NG} , CV_G y el posterior odds del paso genético. A continuación se presenta el algoritmo general que produce las simulaciones.

Algoritmo 2 Simulaciones de evidencia

```

for  $j$  in  $(PD_1, PD_2, \dots, PD_K)$  do
    Simular  $N$  PNIs
    for  $i$  en  $(PNI_1, PNI_2, PNI_3, \dots, PNI_N)$  do
        Muestrear  $NG_{PNI}$  considerando  $P(NG_{PNI}|H)$ 
        Calcular  $CV_{NG}$  y  $O(H | NG)$ 
        Muestrear  $G_{PNI}$  considerando  $P(G_{PNI}|H)$ 
        Calcular  $CV_G$  y  $O(H | G)$ 
    end for
end for

```

4.2.7 Evaluación del poder estadístico

Las simulaciones computacionales permiten obtener las distribuciones de CV_G , CV_{NG} y $O(H | G)$ tanto considerando H_1 como H_2 como verdaderas. Esto permite tomar a estos cocientes de verosimilitud y posterior odds como variables aleatorias a partir de las cuales computar otras métricas. Considerando sólo la evidencia genética, Kling et al. (Kling et al. 2017), definen dos métricas. Por un lado el poder de inclusión, cuya expresión se muestra a continuación:

$$PI_{10000} = P(CV_G \geq 10.000 | H_1) \quad (4.5)$$

El poder de inclusión se determina como la probabilidad de que CV_G supere un umbral de 10.000 cuando H_1 es cierta. Por otra parte, el poder de exclusión se define a continuación:

$$PE = P(CV_G = 0 | H_2) \quad (4.6)$$

El poder de exclusión refiere a la probabilidad de obtener un valor de $CV_G = 0$, cuando H_2 es correcta. La combinación de ambas métricas fue postulada para diagnosticar la capacidad de identificación basada en ADN de los pedigríes. El valor ideal para un pedigrí es $PI_{10.000} = PE = 1$. Esto indica que frente a un caso donde PNI es PD, el resultado superará un CV_G de 10.000, declarándose como potencial identificación a confirmar. Por otro lado, $PE = 1$ implica que todos los PNIs que no correspondan al pedigrí tendrán un valor de CV_G igual a cero.

Aunque estas métricas han demostrado ser útiles (Kling et al. 2017), dependen de umbrales de cociente de verosimilitud fijos, por ejemplo, 10.000. Esto no permite evaluar el rendimiento de la elección de distintos umbrales en la toma de decisiones. Por lo tanto, es necesario introducir otras métricas que permitan evaluar diferentes valores del umbral. En este caso la variable es llamada umbral genético, T_G (Marsico et al. 2021). A continuación se introduce la tasa de falsos positivos:

$$PFP_G = P(CV_G \geq T_G | H_2) \quad (4.7)$$

Es decir, la probabilidad de que CV_G supere el valor T_G considerando que H_2 es cierta, o sea que PNI no es PD. Por otro lado, la tasa de falsos negativos se define a continuación:

$$PFN_G = P(CV_G < T_G | H_1) \quad (4.8)$$

Esta métrica representa la probabilidad de obtener un valor CV_G menor que T_G considerando H_1 cierta. La misma también se puede definir considerando un umbral para el posterior odds, $O(H | G)$. En este caso, se denominará al umbral sencillamente como T , y a las métricas como PFP y PFN . Análogamente a lo planteado para las métricas PE y PI, existe con PFP y PFN un punto óptimo. El mismo es en el cual $PFP = PFN = 0$. Es decir, un pedigrí con dichos valores para un determinado umbral, tendrá probabilidades de falsos positivos y falsos negativos iguales a cero. En la siguiente matriz de confusión se muestra cómo se clasifican los distintos falsos positivos y negativos para constituir las probabilidades previamente mencionadas:

		Clasificación verdadera	
		PNI es PD	PNI no es PD
Predicción	$CV \geq T$	Verdadero positivo	Falso positivo
	$CV < T$	Verdadero negativo	verdadero negativo

Figura 4.1 Tabla de confusión donde se presenta el problema de clasificación binaria basado en el CV, considerando un umbral T .

4.2.8 Cálculo del umbral de decisión

El umbral de decisión, UD se define como el valor de T_G o T que minimiza la distancia euclidiana pesada (DEP) al punto óptimo, donde $PFP = PFN = 0$, en un gráfico donde el eje x corresponde a PFN y el eje y a PFP .

$$DEP = \sqrt{(w_1 PFP)^2 + (w_2 PFN)^2} \quad (4.9)$$

En la Figura 4.2 se resumen, a modo de ejemplo, gráficos de PFP , PFN y DEP en función del umbral, T .

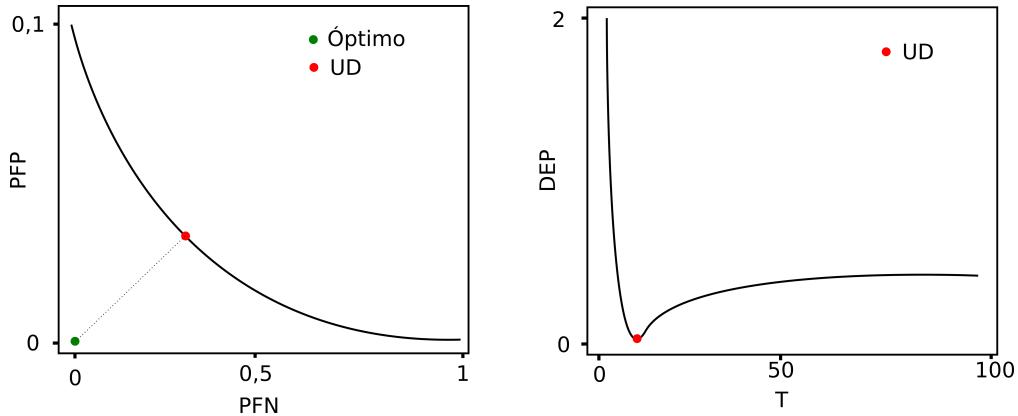


Figura 4.2 Izquierda: PFP y PFN para cada valor de T . Se marca en rojo cuando $T = UD$. Derecha: DEP en función de T , se marca en rojo la DEP mínima, o sea cuando $T = UD$.

La selección de los pesos w_1 y w_2 refleja la importancia relativa de los falsos positivos y negativos definidas por los científicos forenses. Los pesos dependen del tamaño de la base de datos, de la capacidad del laboratorio para realizar nuevos análisis y de otros factores del contexto. Una razón de 10 a 1, correspondiente a $W_1 = 10$ y $W_2 = 1$, entre PFP y PFN permite obtener UDs que derivan en un número manejable de falsos positivos esperados en el contexto que se exemplifica (Marsico et al. 2021). Aún así, estos pesos podrían variar según el caso. Un enfoque similar se utiliza para la selección de umbrales óptimos cuando se aplica la metodología de curvas de ROC (Rota and Antolini 2014).

4.2.9 Estrategia de búsqueda en la base de datos

En la búsqueda mediante bases de datos genéticas se considera una serie de pasos para cada pedigrí (Marsico et al. 2021). Análogamente, pueden realizarse los mismos pasos cuando se combinan datos genéticos y no genéticos. Los mismos son resumidos en la Figura 4.3 y listados a continuación:

1. **Evaluación del poder estadístico:** el primer paso constituye la evaluación del poder estadístico del pedigrí. Si el mismo es suficiente, es decir $PI = PE = 1$, se pasa a la comparación mediante bases de datos (paso 4), utilizando un umbral predefinido por el laboratorio. Generalmente este debe ser alto, para evitar falsos positivos, considerando que la probabilidad de falsos negativos sea cero. En caso contrario se continúa al paso 2.

2. **Incluir más evidencia:** si hay más miembros del grupo familiar cuya muestra puede obtenerse, se los busca. Una vez incorporada la muestra se vuelve al paso 1, evaluando el poder estadístico. En caso de haber múltiples opciones para incorporar miembros se pueden utilizar herramientas para priorizar aquellos que potencialmente agreguen más información (Vigeland et al. 2020). Este problema es específicamente abordado en el capítulo 5 de la presente tesis.
3. **Cálculo del umbral de decisión:** se calcula el UD según la metodología previamente explicada.
4. **Búsqueda en la base de datos:** los casos que quedan por debajo del umbral son descartados. Los que quedan por encima pasan a la etapa de confirmación.
5. **Confirmación:** se reúne toda la evidencia disponible, tanto formalizada matemáticamente como no formalizada, con el fin de identificar inconsistencias y evitar falsos positivos.

4.2.9.1 Implementación

Para la realización de los cálculos de poder estadístico se utilizó la librería disponibles en CRAN forrel (Vigeland 2021). Para el resto, tanto cálculo de los prior odds como selección de UD, se desarrolló una librería específica, publicada en el repositorio CRAN, llamada mispitools (Marsico and Cardi 2023).

4.3 Resultados

En este apartado se presentan los resultados del planteo metodológico de la formalización de distintas líneas de evidencia recolectadas durante la investigación preliminar. Además, se analizan ejemplos puntuales para evaluar la contribución de la evidencia al proceso de identificación.

4.3.1 Proceso de identificación considerando solo datos genéticos

Primero, se analiza un caso donde solamente se utilizan los datos genéticos para la búsqueda en la base de datos. Se introducen 10 pedigríes distintos simulados a modo de ejemplo. Las estructuras de los mismos se pueden ver en las Figuras 4.4 y 4.5. Se deja a la información recolectada durante la investigación preliminar para el paso 5, de confirmación de las potenciales identificaciones. Nótese que este enfoque, comúnmente utilizado en los laboratorios forenses (Marsico et al. 2021), no permite el uso de esta información en los casos descartados, donde existen potenciales falsos negativos.

4.3.1.1 Análisis de poder estadístico de los pedigríes

Pedigríes con solo parentescos de segundo o tercer grado a la persona desaparecida son comunes en casos búsqueda de personas desaparecidas (Kling et al. 2017, Marsico et al. 2021). Algunos de estos pueden arrojar valores altos de CV_G al compararse contra PNIs. Esto puede deberse a la presencia de alelos de baja frecuencia en la población de referencia, que aportan

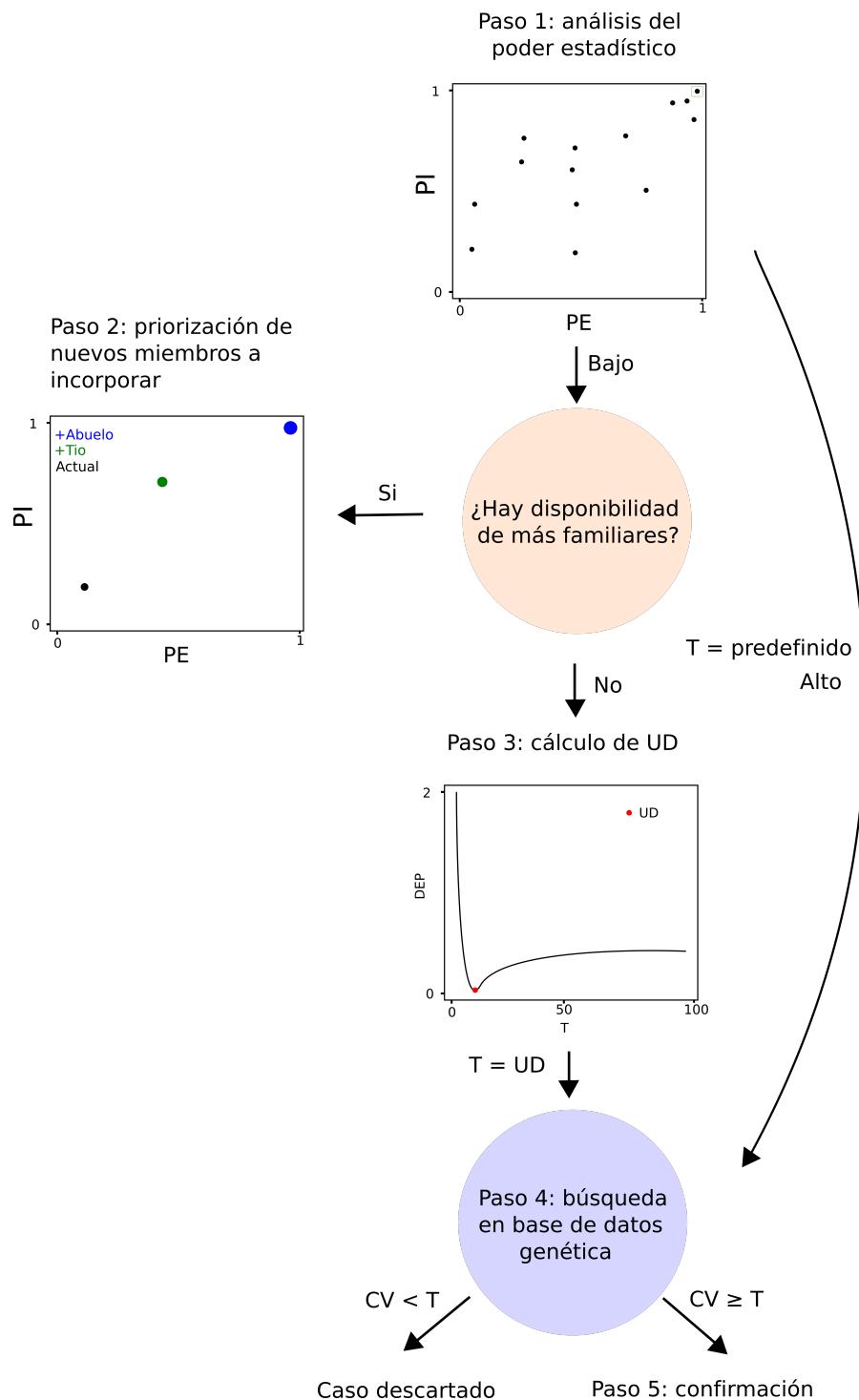


Figura 4.3 Pasos durante un proceso de búsqueda de personas desaparecidas utilizando bases de datos. Se comienza por el paso 1, donde el poder estadístico es analizado para cada pedigrí. En caso de ser bajo, se buscan nuevos posibles miembros a ser incorporado (paso 2) o bien se continúa al cálculo de UD. Por último se realiza la búsqueda en la base de datos genética. El valor del umbral dependerá de si el pedigrí cuenta o no con suficiente poder estadístico.

valores altos de CV_G en caso de coincidencias. Aún así, esta característica convierte a estos grupos familiares en fuente de falsos positivos. Esto es particularmente cierto en pedigríes con alto PI y bajo PE (por ejemplo F3).

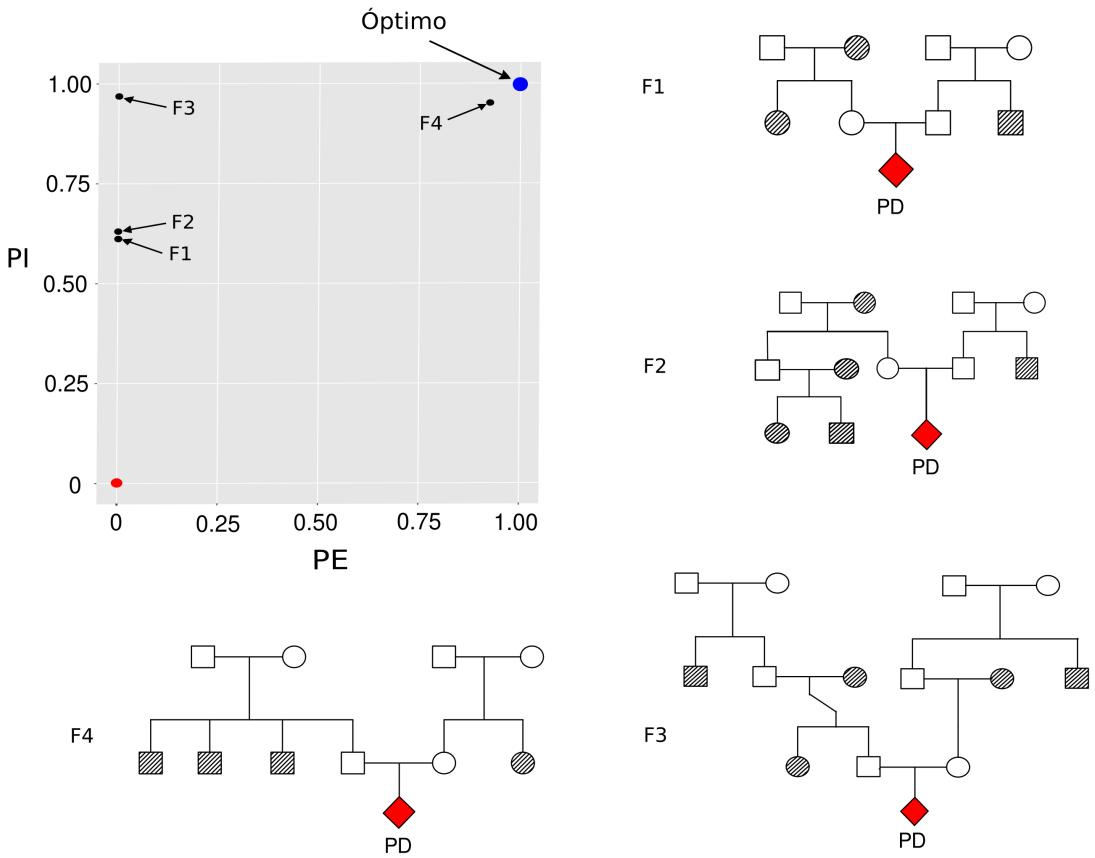


Figura 4.4 Análisis del poder estadístico de los pedigríes. Se presentan 4 pedigríes simulados. En rojo se indica la persona desaparecida y en rayas los individuos genotipados. En el panel superior de la izquierda se muestran los valores de las métricas de poder estadístico para cada pedigrí. El valor óptimo (azul) se ubica en $PE = PI = 1$.

El impacto en el PI y PE de incorporar más marcadores genéticos al pedigrí depende de la estructura del mismo. Como se muestra en la Tabla 4.1, se analizan distintos ejemplos donde se pasa de 15 marcadores STRs a un total de 23. Las estructuras de los mismos se encuentran en la Figura 4.4. La incorporación de más marcadores no mejora el PE en pedigríes con muy pocos parientes o muy distantes a PD. Esto puede verse en F1, F2 y F3. Es interesante comparar F1 y F2, donde la presencia del perfil del tío materno diferencia a ambos. Para F2, a pesar de no contar con la información del tío, la incorporación de sus dos hijos y el de la madre de los mismos alcanza para llegar a un valor de PI similar al de F1. El incremento del número de marcadores permite a F3 aumentar el PI, no así el PE, que continúa siendo cero. En el caso de F4, ambas métricas mejoran.

4.3.1.2 Análisis de la distribución de CV_G

En este apartado se analizan las distribuciones de CV_G de una serie de pedigríes con bajo poder estadístico. Los valores de las métricas se presentan en la siguiente tabla, y la estructura de los mismos en la Figura 4.5.

Como puede observarse, F5 y F10 son los que presentan las peores métricas. Mientras que

Nro marcadores	PI		PE	
	15	23	15	23
F1	0,31	0,57	0	0
F2	0,22	0,63	0	0
F3	0,41	0,95	0	0
F4	0,37	0,91	0	0,89

Tabla 4.1 Métricas de poder estadístico, PE y PI para los pedigríes F1 a F4. Se muestran los resultados al contar con 15 y con 23 marcadores STRs autosómicos.

	PI	PE
F5	0,01	0
F6	0,23	0
F7	0,54	0
F8	0,72	0
F9	0,56	0,86
F10	0	0

Tabla 4.2 Métricas de poder estadístico, PE y PI para los pedigríes F5 a F10. Se muestran los resultados al contar con 23 marcadores STRs autosómicos.

F8 posee el mayor PI, F9 conserva el mayor PE. En la siguiente Figura se analizan las distribuciones de CV_G considerando H_1 o H_2 como verdaderas.

En la Figura puede verse, como es de esperar, que la distribución H_1 cierta (relacionados), se encuentra desplazada hacia valores más altos que la de H_2 cierta (no relacionados). La línea violeta representa el valor de umbral estándar, de 10.000. Puede verse como la curva de H_2 no supera, en ninguno de los casos, dicho umbral. Esto se asocia con el hecho de una baja probabilidad de falsos positivos para el mismo. Aún así, se ve como la curva de H_1 , dependiendo del caso, supera en distintas proporciones al umbral. Nótese que por definición, la proporción de casos que superan dicho umbral de 10.000 se encuentra descrita por el PI, en la tabla 4.2. Disminuir el umbral aumentará la proporción de casos donde PNI es PD (H_1 verdadera) que lo superen (verdaderos positivos), a costa de la aparición de casos donde PNI no es PD (H_2 verdadera) que también lo superen (falsos positivos).

4.3.1.3 Cálculo del Umbral de Decisión

Como se explica en la Figura 4.5, para aquellos pedigríes con bajo poder estadístico, en los cuales no hay posibilidad de incorporar más miembros, se procede al cálculo de UD. La Figura 4.6 muestra las curvas de PFP-PFN utilizadas para la obtención del mismo.

Se muestra como, en todos los casos, disminuir el valor del umbral produce un descenso en la PFN, a costo de un incremento de la PFP. Se marca el valor de DT, que surge de minimizar la distancia al punto óptimo. La siguiente tabla resume (tabla 4.6) los valores de UD, PFP y PFN para cada pedigrí.

Puede verse que la metodología propuesta establece un $UD = 10$ para F7, con $PFN = 0,03$ y $PFP = 0,002$. En una base de datos de 10.000 PNIs, se esperan obtener 20 individuos no

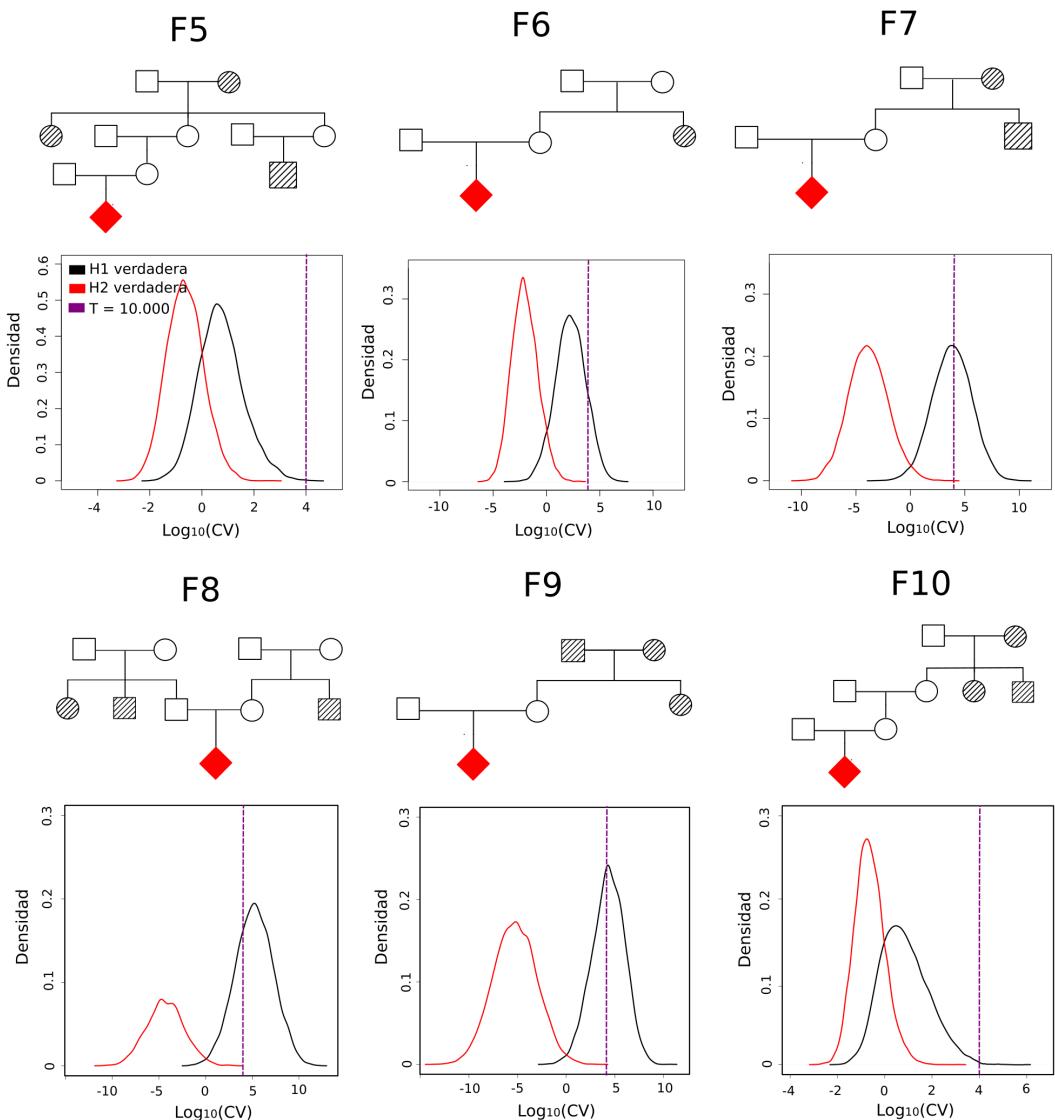


Figura 4.5 Análisis de la distribución de $\text{Log}_{10}(\text{CV})$ considerando H_1 y H_2 como ciertas en un conjunto de casos con bajo poder estadístico. Se indica en línea violeta el valor de umbral de 10.000.

relacionados (falsos positivos) que superen dicho umbral. Previamente, con un umbral de 10.000 (tabla 4.2) los falsos positivos esperados eran cero, pero el PFN (1-PI) era de 0,46. En el marco del uso del UD, el laboratorio puede evaluar con mayor profundidad estos 20 casos, con el beneficio de un decrecimiento considerable en la probabilidad de perder una identificación. Por otra parte, para F6, el valor de UD derivaría en 120 falsos positivos esperados en una base de 10.000. Los casos más complejos, F5 y F10, con UD's muy bajos, implican una cantidad alta de falsos positivos esperados (más de 200). Además, la probabilidad de falsos negativos continúa siendo alta. En términos generales, la UD mejora considerablemente la búsqueda en los casos F7, F8 y F9. El caso F6 es de una complejidad intermedia, y los casos F5 y F10 continúan siendo de compleja resolución. Aún así, en estos últimos, se observa una clara mejoría, ilustrada en la

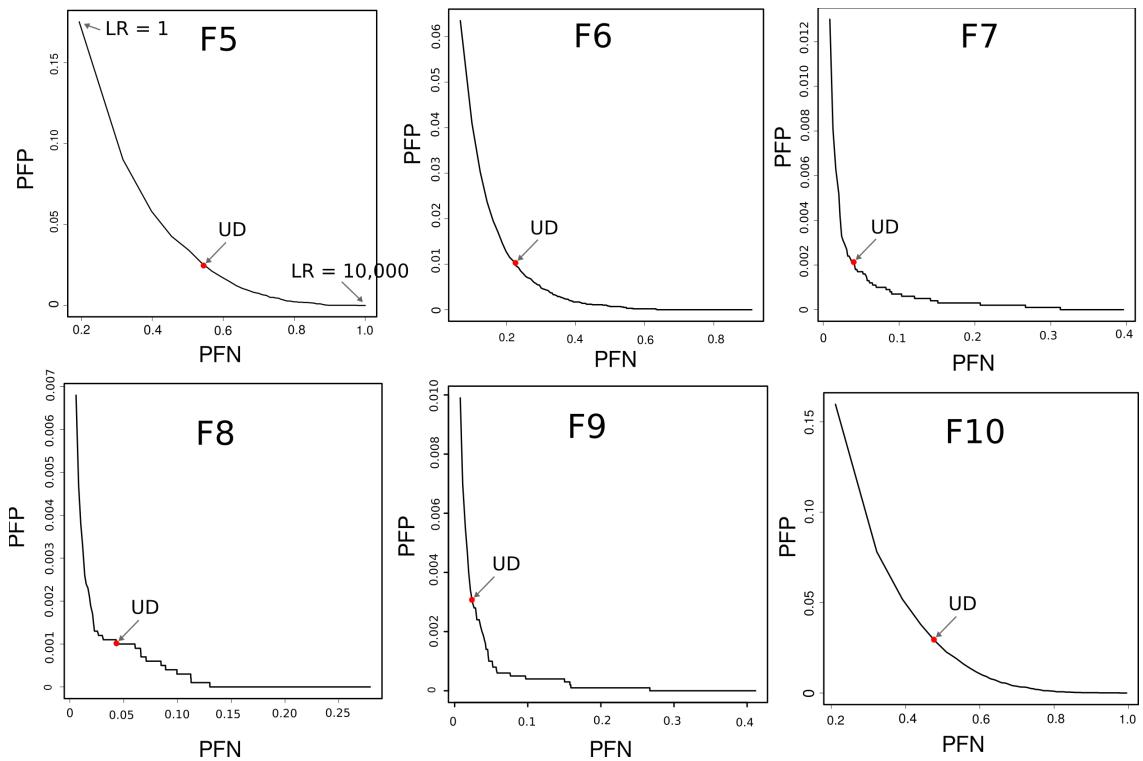


Figura 4.6 Curvas de tasas de falsos positivos y negativos para distintos valores de umbral. Se indica el UD en color rojo.

Pedigree	UD	PFP	PFN
F5	6	0,021	0,53
F6	8	0,012	0,21
F7	10	0,002	0,03
F8	12	0,001	0,02
F9	7	0,003	0,02
F10	6	0,024	0,47

Tabla 4.3 Valores de umbral de decisión, UD y las tasas de error de falsos positivos, PFP y falsos negativos PFN para cada pedigrí de F5 a F10. En todos los casos se analizaron 23 marcadores STRs autosómicos.

Figura 4.7.

Para aquellos casos con una cantidad de falsos positivos esperados muy alta el laboratorio forense puede verse impedido de aumentar masivamente el número de análisis genéticos a realizar. Es en este punto donde la formalización de los datos provenientes de la investigación preliminar se vuelve fundamental en la toma de decisiones, como se verá en una sección posterior. Previamente, se compara el enfoque del UD con los otros previamente descritos en la bibliografía científica y comúnmente utilizados en los laboratorios.

4.3.1.4 Comparación con otras estrategias de selección del umbral T_G

En esta sección se compara el enfoque UD, para la selección de T_G , con otros métodos previamente descritos por Kruijver et al. (Kruijver et al. 2014). Las propuestas son:

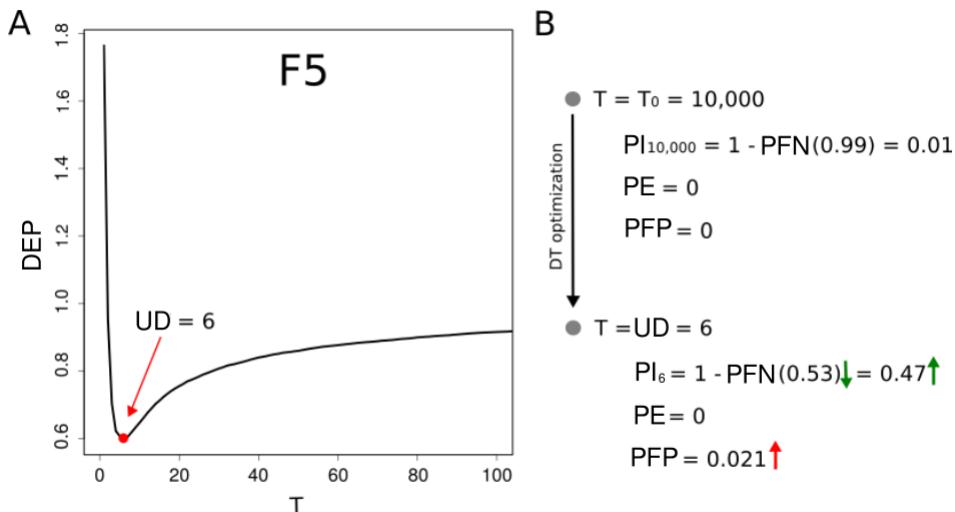


Figura 4.7 Cálculo de UD para el pedigrí F5. A Se muestra como UD es el valor de umbral que minimiza el DEP. B Se comparan las métricas de poder estadístico para el pedigrí F5 al considerar un umbral de 10.000 contra el umbral UD. Se observa la disminución de la probabilidad de falsos negativos a costo de un aumento de los falsos positivos.

1. PFP fijada: cuando una PFP es fijada para todos los pedigríes y lo que varía es el umbral T_G y la PFN.
2. PFN fijada: cuando una PFN es fijada para todos los pedigríes y lo que varía es el umbral T_G y la PFP.
3. T_G fijada: cuando un T_G es fijado para todos los pedigríes y lo que varía es PFP y PFN.

Las primeras dos son versiones de la estrategia centrada en los perfiles, descrita en la introducción de este capítulo. La tercera es la de umbral fijo. La estrategía Top-N no se la considera, por tratarse de un caso particular de la PFP fijada (donde la cantidad de falsos positivos depende fuertemente de los recursos del laboratorio). La estrategia condicional será analizada en el apartado siguiente, con la incorporación de los prior odds basados en los datos no-genéticos.

Para la fijación de valores se seleccionan los siguientes parámetros: PFP = 0,01 para (i); PFN = 0,03 para (ii); y $T_G = 10$ para 3 (iii). Se selecciona un umbral diferente a 10.000 debido a que este ya fue analizado mediante los cálculos de PI y, como se comentó, puede variar según el laboratorio. Los resultados se presentan en la siguiente tabla (tabla 4.4):

Se pueden observar diferencias entre los valores propuestos por los distintos métodos. Pedigríes con muy bajo poder estadístico, como F5 y F10, muestran mayor divergencia con el enfoque UD. Particularmente, el método PFN fijo da resultados poco útiles para los casos F5 y F10, derivando en una PFP igual a uno (es decir, incluir a todos los PNIs como posibles identificaciones). También el método PFP fijado da valores más altos de PFN que el enfoque UD en los casos F5, F6, F8 y F10. En los casos F7 y F9, con el enfoque UD, un leve aumento en la PFN permite una mucho menor cantidad de falsos positivos. El enfoque T_G fijado da mayores valores de PFN en todos los casos, menos en F7, donde el UD = 10 coinciden los valores de umbral.

Pedigrí	$T_G = 10$		$PFP = 0,01$		$PFN = 0,03$	
	PFP	PFN	T_G	PFN	T_G	PFP
F5	0.011	0.61	11	0.64	0 (<0.1)	1
F6	0.010	0.23	11	0.22	24	0.005
F7	0.002	0.03	1	0.01	10	0.002
F8	0	0.03	1	0.06	12	0
F9	0.001	0.04	3	0.01	9	0.002
F10	0.014	0.64	12	0.67	0 (<0.1)	1

Tabla 4.4 Análisis de umbrales y tasas de error propuestas por otras metodologías alternativas al umbral UD , presentes en la literatura científica Kruijver et al. (2014). Se muestra, para cada pedigrí de F5 a F10, las tasas de error, PFP y PFN, considerando un umbral fijo de $T = 10$. Luego el valor de umbral y la tasa PFN, considerando una PFP fija de 0,01 para todos los pedigríes. Por último, se muestra el valor de umbral y la tasa PFP, considerando un PFN fijo de 0,03 para todos los pedigríes.

4.3.2 Incorporando datos de la investigación preliminar

En esta sección se analizan dos contextos para evaluar el impacto de la incorporación de los datos de la investigación preliminar, tanto en casos con buen poder estadístico, como en casos con bajo poder estadístico.

4.3.2.1 En casos de bajo poder estadístico

En los casos F5, F6 y F10, el UD continúa con una tasa de falsos positivos suficientemente grande como para ser un desafío para el laboratorio forense. En este apartado se analizará el caso de F6, debido a que la búsqueda por parte de solamente un tío es un evento frecuente. Con 23 marcadores, se esperan alrededor de 120 posibles falsos positivos considerando solo la información genética. La posibilidad de incorporar más análisis, por ejemplo ADN mitocondrial, cromosoma X, o más marcadores STRs, implica estudiar las muestras del pedigrí (sólo del tío) y también la de las potenciales identificaciones (todas las esperadas). En este escenario hipotético, se considera que dicho tío está buscando a dos sobrinas desaparecidas, denominadas PD_1 y PD_2 . A continuación se resumen sus características:

	Sexo biológico (S)	Color de pelo (C)	Edad (E)
PD_1	femenino (F)	castaño (1)	entre 8 a 10
PD_2	femenino (F)	colorado (5)	entre 18 a 22

Tabla 4.5 Datos recolectados durante la investigación preliminar para las personas desaparecidas PD_1 y PD_2 .

Debido a que la relación de parentesco (tío-sobrina) es la misma, la simulación de datos genéticos arrojará el mismo resultado para ambas PDs . No así cuando se consideran los prior odds basados en datos de la investigación preliminar.

Como puede observarse, la incorporación de prior odds no-uniformes, basados en los datos de la investigación preliminar, permiten una mejor separación de las distribuciones conseguidas

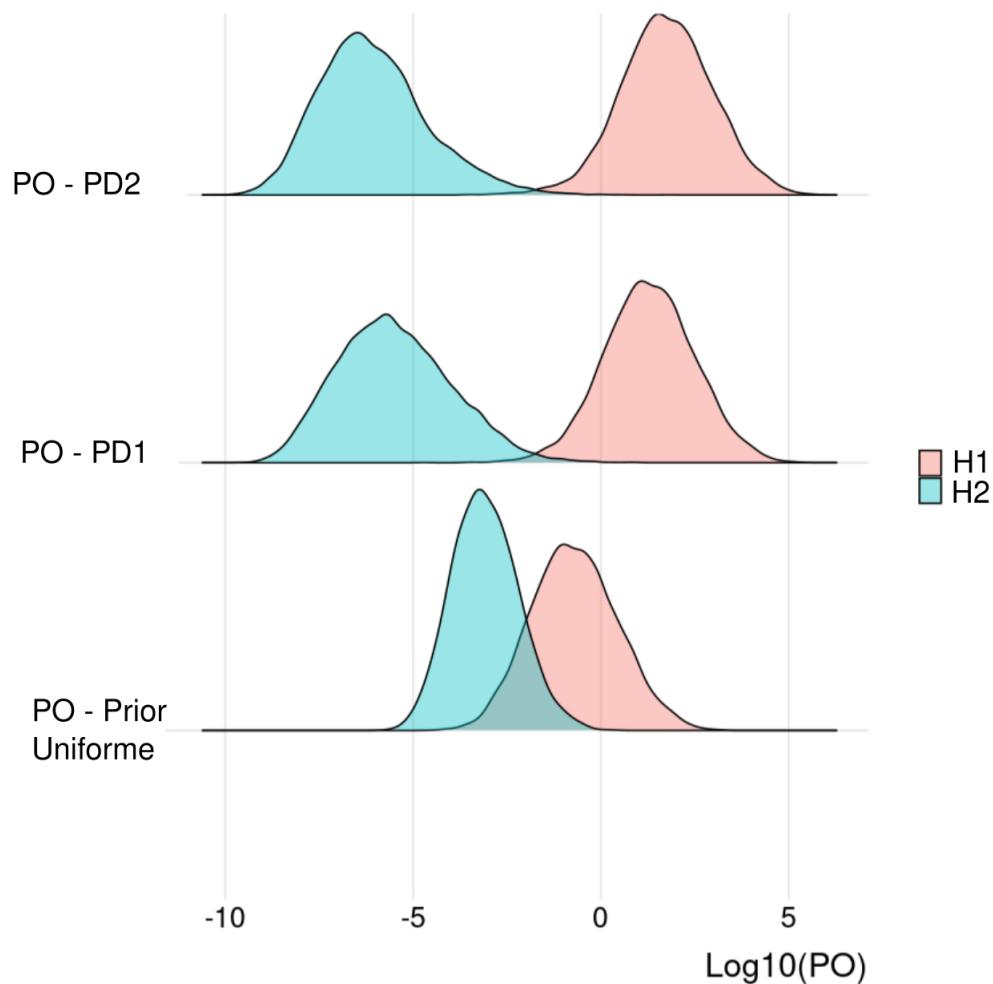


Figura 4.8 Distribución de posterior odds considerando H_1 y H_2 como ciertas para el pedigrí F6. Se presentan los posterior odds calculados considerando el prior odds uniforme y los priors odds basados en datos de la investigación preliminar para cada caso, PD_1 y PD_2 .

considerando H_1 y H_2 como ciertas (Figura 4.8). Más aún, puede apreciarse cómo la separación es aún mayor para PD_2 . Esto se debe, en parte, a que el color de pelo de la misma es menos frecuente en la población. La siguiente tabla (tabla 4.3) resume las métricas obtenidas para los casos. Recuérdese que en este caso el umbral es sobre el posterior odds, por lo tanto lo denominamos T , a diferencia de T_G que se basa sólo en la información genética:

Para los casos considerados, las métricas de rendimiento para PD_1 y PD_2 son mejores al incorporar los datos no genéticos. Con el prior odds uniforme permanecen iguales a considerar solo la evidencia genética. Esto último es esperable, debido a que el prior no aporta información para discernir entre los casos relacionados y no relacionados. En este caso, en una base de datos de 10.000 se pasaría a seleccionar solo 10, al incorporar la información no-genética.

Pedigree	UD_T	PFP	PFN
F6-Uniforme	0,12	0,021	0,53
F6- PD_1	1	0,001	0,01
F6- PD_2	0,8	0,001	0,02

Tabla 4.6 Valores de UD, PFP y PFN para cada pedigrí considerando los distintos prior odds utilizados para el pedigrí F6. En el caso uniforme el mismo prior odd se calcula para cada para $PD - PNI$, en el caso de $F6 - PD_1$ se utiliza el prior odd basado en los datos de la investigación preliminar para este individuo. En el caso de $F6 - PD_2$ se utilizan los datos de PD_2 .

4.3.2.2 En casos con alto poder estadístico

En casos con alto poder estadístico, donde $PI = PE = 1$, se espera que al analizar a todos los PNIs en una base de datos produzca valores mayores al umbral de 10.000 tan solo si se encuentra a PD dentro de la misma. El resto de los resultados esperados es igual a cero.

En ciertos contextos, debido al gran tamaño de las bases de datos, o bien porque no se encuentran todas las muestras de ADN analizadas, los investigadores forenses tienden a filtrar los casos utilizando datos de la investigación preliminar, como por ejemplo el sexo biológico (Vigeland and Egeland 2021, Marsico and Caridi 2023). Tal es así, que algunos softwares como el Familias (Egeland et al. 2015), otorgan esta opción de filtrado antes de realizar el análisis masivo de parentesco genético. Este enfoque, aunque puede parecer razonable, fuerza, implícitamente, una correspondencia perfecta entre las variables del PD y las de PNI. Dicho de otro modo, filtrar significa establecer un criterio de concordancia perfecta entre los valores que toman las variables no-genéticas de PNI y PD . Como se menciona en la metodología del presente capítulo, y con más detalle en el capítulo dos, la posibilidad de errores en la carga de datos, y/o errores en la asignación de los forenses a las distintas variables pueden generar problemas a la hora de realizar el filtrado. Para este ejemplo, se asume un error de un 0,01 para todas las variables no-genéticas cargadas en la base de datos.

Se analiza un caso con alto poder estadístico, llamado F11, donde ambos padres de la persona desaparecida están tipificados, y por lo tanto $PI = PE = 1$. Con el fin de simplificar, se considera que la pareja busca dos hijas, con los mismos atributos que los mencionados en la Tabla 4.5.

En la tabla 4.7 se muestra el resultado de analizar todos los PNIs sin realizar un filtrado. Esto implícitamente implica asumir un umbral para el posterior del paso no genético igual a 0 (considerando solo los datos de la información de la investigación preliminar). Al tratarse de tres variables, permite hasta 3 no-concordancias entre el PNI y PD. El extremo opuesto se simboliza con el nombre CP, concordancia perfecta, y emula la decisión de filtrar por variables no-genéticas. En este caso no se permiten no concordancias y se compara un número mucho menor de PNIs, sólo 225 para PD_1 y 75 para PD_2 (recuérdese que PD_2 contaba con rasgos menos frecuentes en la población). Un resultado similar se obtiene para un $T_{NG} = 10$. Interesantemente, un $T_{NG} = 1$ permite una no-concordancia entre las variables, dejando 3125 casos a analizar para PD_1 y 425 para PD_2 .

Al ver las métricas, puede entenderse que la estrategia de filtrado es similar a colocar un

PD	T_{NG}	# MM	N_t testeados	PFP	PFN
PD_1	PM	0	225	0	0,13
PD_2	PM	0	75	0	0,13
PD_1	10	0	225	0	0,13
PD_2	10	0	75	0	0,13
PD_1	1	1	3125	0	0,05
PD_2	1	1	425	0	0,05
PD_1	0	3	10000	0	0
PD_2	0	3	10000	0	0

Tabla 4.7 Métricas de rendimiento. T_{NG} indica el umbral para el paso no genético utilizado, PM indica la concordancia perfecta entre las variables de PD y PNI , o sea si PD es femenino, PNI debe serlo también. MM es el número de incompatibilidades (por ejemplo, sexo biológico diferente) permitidos entre PD y PNI . N_t es el número de PNIs cuyo valor CV queda por encima del umbral basado en datos no genéticos T_{NG} . Estos casos son posteriormente analizados mediante test de parentesco genético. Las tasas, PFP y PFN, analizan el error general considerando ambos pasos, el genético y el de la investigación preliminar. Nótese que umbrales muy altos de T_{NG} o el requisito de concordancia perfecta PM equivalen a filtrar la base de datos de PNI buscando solo aquellos cuyas características coincidan completamente con las de PD . A esto se lo denomina estrategia de filtrado.

umbral muy alto, derivando en $PFN = 0,13$. Este error es implícito si solo se analiza la información genética en el proceso de toma de decisiones. El caso contrario, realizar el análisis de parentesco genético para todos los pares posibles $PNI - PD$ pareciera ser la mejor decisión pero, como se mencionó, puede encontrarse imposibilitado por el tamaño de la base de datos o bien por no contar con información genética de todos los PNIs.

4.4 Discusión

En este capítulo se introdujo una estrategia general para lidiar con casos de búsqueda de personas desaparecidas en bases de datos. Específicamente, se analizó la problemática de pedigríes con poco poder estadístico, donde el análisis masivo puede derivar en un alto número de falsos positivos (Marsico et al. 2021).

Enfoques de teoría de decisión bayesiana se han propuesto como forma de optimización de la toma de decisiones (Tillmar and Mostad 2014). Estos requieren que se especifique un costo para los falsos positivos y negativos, además del uso de prior odds. Este enfoque resultó útil en laboratorios que trabajaron con casos de paternidad, como se explica en (Hedell et al. 2018), y puede aplicarse a otros problemas de genética forense. Otros enfoques han explorado el uso de redes bayesianas y teoría de la decisión para la toma de decisiones en bases de datos (Gittelson et al. 2012, Biedermann and Taroni 2012).

La estrategia propuesta en este capítulo, a diferencia de los otros enfoques, busca hacer uso de la información genética ya disponible y de la información recolectada durante la investigación preliminar. En los casos de poco poder estadístico, a pesar de no ser suficiente para llegar a una conclusión, la evidencia recolectada puede ser utilizada para seleccionar un subconjunto de PNIs donde realizar más análisis. Más aún, el método propuesto por Tillmar et al. (Tillmar and

Mostad 2014) permite seleccionar un conjunto de marcadores de forma óptima con el fin de llegar a una conclusión. Comparaciones entre distintos *sets* de marcadores se basan en una función que relaciona el valor de la nueva información esperada, adherida a la información disponible.

El enfoque de UD arroja un valor manejable de falsos positivos reduciendo la probabilidad de falsos negativos en la serie de casos analizados. Al agregarse la información de la investigación preliminar, se permite un resultado elocuente inclusive con casos de mayor complejidad, como por ejemplo una identificación con tan sólo un tío.

Como se analiza en la literatura (Kling et al. 2017, Kruijver et al. 2014), en los casos con buen poder estadístico, la toma de decisiones se ve simplificada debido a que se esperan valores altos de CV_G solo en aquellos casos donde H_1 es cierta. En el resto de los casos, el valor esperado es la exclusión ($CV_G = 0$). Aún así, en el capítulo se analiza qué decisiones previas a la comparación genética, como el filtrado de la base de datos mediante el uso de variables de la investigación preliminar, puede afectar al rendimiento general de la búsqueda. Se proponen métodos alternativos, que permiten reducir la probabilidad de falsos negativos, relajando las condiciones de filtrado. Esto implica incorporar más individuos a la comparación genética.

El balance entre individuos a analizar y la probabilidad de falsos negativos tolerada para los casos implica un costo asociado a falsos negativos o falsos positivos. En este enfoque, dicho costo se materializa en los pesos asignados a PFP y PFN (w_1 y w_2) cuando se calcula la DEP.

4.5 Conclusiones del capítulo

En este capítulo se propuso un modelo general para la optimización de la toma de decisiones en la búsqueda de personas desaparecidas mediante bases de datos. Se contextualiza su aplicación en un escenario de casos simulados donde el padre y la madre de la persona a identificar se encuentran desaparecidos, por lo tanto, no es posible obtener muestras de ADN de los mismos. Es así que familiares de segundo o tercer grado son incorporados en el pedigrí para realizar la búsqueda. En este contexto, se analizó el rol de la formalización de los datos recolectados durante la investigación preliminar. Se pudo evaluar el impacto de incorporar los mismos como prior odds del paso genético, utilizando los modelos descritos en el capítulo 2. En el capítulo siguiente se analizarán distintas herramientas para la toma de decisiones en la incorporación de nueva evidencia.

Capítulo 5

El problema de la priorización para la incorporación de nueva evidencia

En este capítulo se aborda una problemática específica en la búsqueda de personas desaparecidas: la priorización en la recolección de nueva evidencia genética en casos con poco poder estadístico. Como se ha visto en el capítulo 4 de la presente tesis, el bajo poder estadístico en búsquedas mediante bases de datos genéticas puede llevar a un alto número de falsos positivos y una mayor probabilidad de falsos negativos. Cuando un pedigrí presenta bajo poder estadístico, el investigador forense suele buscar posibles nuevos miembros de referencia a incorporar. En ciertos casos, múltiples opciones de incorporación de nuevos miembros son posibles y el tomador de decisiones debe evaluar en función del balance entre el costo y el beneficio qué opción elegir. Por ejemplo, la búsqueda de un tío de una persona desaparecida para incorporar al pedigrí de referencia puede ser una tarea sencilla, dado que el mismo se encuentra vivo y es de fácil acceso. En cambio, la incorporación de un abuelo fallecido puede involucrar una exhumación para dar con sus restos y, por lo tanto, un procedimiento judicial que trae aparejado mayores costos. En este contexto, el investigador forense puede encontrarse con la disyuntiva de utilizar los recursos disponibles para buscar al tío vivo, o bien ordenar la exhumación, o ambas. En muchos casos, no solo el tiempo es un recurso limitado, si no también otros recursos (económicos, entrevistas, procedimientos judiciales, etc) que se requieren para realizar la investigación preliminar con el fin de localizar a los parientes del desaparecido. En este capítulo se exploran diferentes herramientas cuantitativas con el fin de predecir el impacto en el poder estadístico de la incorporación de nuevos miembros al pedigrí de referencia. Específicamente, se comparan dos estrategias. Por un lado un enfoque basado en simulaciones computacionales, que hace uso de las métricas de poder de inclusión y exclusión, descritas en el capítulo 4. Por otro lado, se propone una metodología basada en la teoría de la información, que hace uso del potencial de las redes bayesianas (introducidas en el capítulo 3), para computar el impacto esperado de la incorporación de nuevos miembros al pedigrí sin necesidad de realizar simulaciones, reduciendo drásticamente el costo computacional y de tiempo. Finalmente, se discuten las ventajas introducidas por el abordaje de la teoría de la información, sus antecedentes en el campo de las Ciencias Forenses, y sus posibles usos para responder a otras problemáticas.

5.1 Introducción al capítulo

El éxito en la búsqueda de personas desaparecidas depende críticamente del contenido de la información recolectada durante el proceso (Marsico et al. 2021). Como se ha discutido a lo largo de la tesis, tanto aquella información recolectada durante la investigación preliminar (Caridi et al. 2020), como la información genética, contribuyen a llegar a conclusiones en torno a los casos durante la búsqueda. En distintos escenarios, es posible contar con más de una opción de nuevas líneas de evidencia a incorporar. Es ahí donde el científico forense debe priorizar entre alguna de ellas, debido a limitaciones de distintos recursos disponibles (Vigeland et al. 2020). Este capítulo se centra en explorar herramientas que contribuyan a una selección racional de nuevas evidencias con el fin de incrementar el poder estadístico en la búsqueda.

5.1.1 El problema de priorización

El capítulo 4 de la presente tesis se sumerge en la toma de decisiones en casos de búsqueda de personas desaparecidas mediante el uso de bases de datos. Particularmente se aborda un problema de optimización mediante el establecimiento de umbrales. Estos son utilizados para definir posibles identificaciones y pueden aplicarse utilizando solamente el cociente de verosimilitud o bien el posterior odds. Los casos que superan dicho umbral, es decir que el cociente de verosimilitud o el posterior odds dan valores más altos que el umbral establecido, son considerados candidatos para la incorporación de nuevos análisis genéticos, con el fin de arribar a una conclusión. Distintos valores de umbral harán que una mayor o menor cantidad de pares *PNI – PD* sean considerados potenciales identificaciones. La selección de dichos pares debe ser racional, debido a que los análisis genéticos (por ejemplo cromosoma X, cromosoma Y o ADN mitocondrial) se deben realizar tanto sobre los miembros del pedigrí, como sobre cada PNI. Esto puede suponer un gran gasto (en insumos para nuevos análisis genéticos, tiempo para investigación preliminar con el fin de confirmar datos, etc) si la cantidad de casos potenciales a confirmar es muy alta (Vigeland et al. 2020). Esta estrategia se propone como alternativa, en caso de no contar con más familiares para incorporar a un pedigrí con bajo poder estadístico. Esto sucede debido a que la incorporación de miembros a un pedigrí permite aumentar el poder estadístico de forma general, es decir, en el análisis del pedigrí contra todos los PNIs (sin tener que realizar análisis extra sobre los mismos). Intuitivamente podría pensarse que la incorporación de más familiares de referencia es siempre una mejor opción. Aún así, la búsqueda e incorporación de nuevos miembros puede verse dificultada por múltiples motivos, como: (i) registros escasos del árbol genealógico del desaparecido, (ii) desconocimiento del paradero de los restos del familiar, en caso de haber fallecido, (iii) exhumaciones complejas debido al mal estado, o mal registro en los cementerios, (iv) negación por parte del familiar a dar una muestra para ser analizada, entre otros (Vigeland et al. 2020). En algunos casos para dar con los familiares será necesario profundizar en la investigación preliminar (Puerto et al. 2021), o bien llevar a cabo medidas judiciales para la obtención de la muestra. En este sentido, nuevamente el investigador forense se encuentra frente a un problema de toma de decisiones. Debe evaluar entre el costo (en términos de tiempo, economía y el impacto de la producción de expectativas sobre los familiares de las personas desaparecidas) y el beneficio (en términos de la mejora de poder

estadístico para la identificación genética) de incorporar miembros al pedigrí. El hecho se torna aún más complejo cuando hay múltiples opciones de incorporación, por ejemplo la decisión de incorporar dos primo-hermanos que se encuentran vivos y de fácil acceso, o realizar la exhumación de un abuelo fallecido (Vigeland et al. 2020). Formalmente, el problema planteado es de *priorización*, y el principal objetivo del capítulo es introducir formas prácticas para asistir a la toma de decisiones frente a dicho problema. Se busca una respuesta robusta a la pregunta: *¿Qué miembro adicional debe incorporarse al pedigrí para mejorar la capacidad de identificación?* El costo temporal y económico es complejo de formalizar y sistematizar, debido a que depende de diferentes contextos. En cambio, el beneficio en términos de poder estadístico puede abordarse de distintas maneras. En trabajos previos se han discutido problemas similares relacionados a la identificación de personas mediante el test de parentesco genético (Pinto et al. 2019, Ge et al. 2011). Estos enfoques se basan en casos específicos, utilizando simulaciones computacionales para estudiar la distribución de los cocientes de verosimilitud genéticos esperados con la incorporación de nuevos miembros. Los enfoques planteados en el presente capítulo difieren de los estudios previos en dos aspectos: por un lado, la propuesta de Pinto et al. (Pinto et al. 2019) y la de Ge et al. (Ge et al. 2011) se centran en la probabilidad de obtener valores altos de CV_G cuando H_1 es cierta (es decir PNI es PD). En ambos métodos se analizan casos específicos, mientras que no se generaliza una metodología para la respuesta a la pregunta de priorización. En los modelos planteados en este capítulo, las metodologías propuestas consideran la importancia de llegar a conclusiones en ambas direcciones, tanto para la obtención de valores altos de CV_G cuando H_1 es correcta, como la de valores bajos de CV_G cuando H_2 es correcta. Por otro lado, se propone una estrategia general, para lidiar con diferentes tipos de configuraciones de pedigríes.

5.1.2 Métodos basados en el cálculo del poder estadístico

Esta estrategia se basa en simulaciones computacionales, introducidas en el capítulo 4, utilizando las métricas de *poder de inclusión* (PI) y *poder de exclusión* (PE). En enfoques previos se utilizó la estrategia de simulaciones computacionales no condicionadas (Ge et al. 2011). Estas implican generar perfiles genéticos de individuos que potencialmente se incorporarían al pedigrí, sin utilizar la información ya disponible de los perfiles analizados. Las simulaciones condicionales, en cambio, permiten hacer uso de esta información, y han demostrado ser más exactas en el cálculo del poder estadístico (Kling et al. 2017). Uno de los resultados más llamativos de estas simulaciones es que pedigríes con la misma estructura, por ejemplo, cada uno con un tío genotipado, pueden resultar en un poder estadístico diferente. Las características genéticas de menor frecuencia en la población de referencia colaborarán con la obtención de CV_G s altos, y por lo tanto un aumento del PI. En cambio, en el caso del PE la interpretación se vuelve más compleja. El PE puede verse afectado por los genotipos específicos cuando múltiples miembros se encuentran genotipados en un mismo pedigrí. Por ejemplo, si dos hermanos de la persona desaparecida son muy similares entre sí, aportarán menos PE que dos hermanos muy diferentes (Egeland et al. 2014). En las simulaciones no condicionales, estos efectos específicos no son tenidos en cuenta, y solo se presenta el promedio de valores obtenidos de las simulaciones de posibles genotipos. En este sentido, ambas métricas, PE y PI, resultan informativas para priorizar

miembros a incorporar al pedigrí. Por lo tanto, una respuesta práctica al problema de priorización radica en brindar herramientas que permitan visualizar ambas métricas para posibles opciones de incorporaciones de familiares. En este capítulo se hace especial énfasis en visualizaciones de ágil interpretación, asociadas al hecho de que los forenses, en ciertos contextos, deben tomar decisiones en el campo de trabajo.

5.1.3 Métodos basados en el cálculo del contenido de información

La teoría de la información es un área del conocimiento propuesta a mediados del siglo XX (Shannon 1948). Esta introduce un marco teórico para la medición del contenido de información, inicialmente en los mensajes transmitidos. Hoy en día, dicho marco ha sido utilizado en una amplia cantidad de problemáticas (Uda 2020, Jaynes 1957b). Trabajos previos han explorado el enfoque de teoría de la información en Ciencias Forenses, como los de Ramos et al. (Ramos et al. 2020, 2013). Estos se centran en líneas de evidencia como el reconocimiento de voces, utilizando también el cociente de verosimilitud como estrategia para evaluar el peso de la evidencia. Aún así, no se ha explorado en profundidad la utilización de la teoría de la información para el análisis de la evidencia genética.

Como se ha introducido en el capítulo 1, uno de los conceptos fundamentales dentro de la teoría de la información refiere a la entropía de Shannon (Shannon 1948). Esta mide la información aportada por una variable, en promedio. La información se computa como el logaritmo de la probabilidad. Si el logaritmo utilizado es en base 2, su unidad son los *bits*, en cambio, si su base es el logaritmo natural, su unidad son los *nats*.

Para un determinado pedigrí, es posible obtener los valores de CV_G esperados considerando H_1 y H_2 ciertas. A partir de estos valores se pueden analizar sus respectivas distribuciones. Como se ha visto en el capítulo 4, a medida que las distribuciones de H_1 y H_2 se separan, el rendimiento de la búsqueda aumenta. Por lo tanto, las métricas de la diferencia de ambas distribuciones pueden ser de utilidad. Distintas métricas provenientes de la teoría de la información permiten medir la diferencia de información entre dos distribuciones de probabilidad. A modo de ejemplo, cuando un pedigrí cuenta solo con poca información genética (10 marcadores STRs) de un individuo lejano a la persona desaparecida (ej. un tío-abuelo), se espera que la diferencia entre las distribuciones de CV_G obtenidas cuando H_1 o H_2 son ciertas sea baja. Esto llevará a un solapamiento considerable de ambas distribuciones que, como se observó en el capítulo anterior, puede derivar en errores como falsos positivos y negativos (Marsico et al. 2021). El ejercicio en el que se centra el método propuesto basado en teoría de la información es cuantificar la diferencia entre las distribuciones de probabilidades genotípicas cuando nuevos potenciales miembros son incorporados al pedigrí. Para poder calcular la diferencia entre distribuciones de probabilidad mencionadas, es necesario computar de forma exhaustiva la distribución de probabilidades genotípicas para cada miembro del pedigrí, condicionada en la información disponible. Esto es posible mediante la implementación de las redes bayesianas (Chernomoretz et al. 2022, Darwiche 2009). Se estudia la relación entre la distancia de dichas distribuciones y la capacidad identificatoria de un pedigrí, proponiendo herramientas cuantitativas diagnósticas.

Dicho enfoque se utiliza para analizar el impacto de la incorporación de nuevos miembros al grupo familiar, restringiendo el espacio de posibles genotipos basándose en la información

disponible. De forma análoga a lo planteado mediante el método de simulaciones computacionales (Vigeland et al. 2020), se evalúa la informatividad en ambos sentidos, tanto para evaluar como ciertas H_1 y H_2 .

Ambos enfoques se encuentran implementados, y libremente disponibles, en CRAN. El método de teoría de la información se aplica mediante el paquete *fbnet* (Chernomoretz et al. 2022). El método basado en simulaciones se encuentra disponible en el paquete *forrel* (Vigeland et al. 2020) y *mispitools* (Marsico and Caridi 2023). Importantemente, el análisis de parentesco es heterogéneo, en tipos de casos y en estructuras de grupos familiares involucrados en la búsqueda (Puerto and Tuller 2017). En este sentido, ambas metodologías son fácilmente adaptables a distintos contextos.

5.2 Métodos

5.2.1 Análisis de parentesco genético

Como se explica en el capítulo anterior, y con mayor detalle en el capítulo 3, el análisis de parentesco genético implica el cálculo de CV_G . En esta sección se enfatizará solo en los parámetros seleccionados para el cálculo del mismo, mientras que las ecuaciones se dan por introducidas previamente. Para todos los cálculos realizados se utilizó la base de datos de frecuencias alélicas de la Argentina, con un total de 23 marcadores STRs autosómicos. Cuando aplique, se utiliza el modelo mutacional *uniforme*, con una tasa de 0,002. Para el cálculo de PE ningún modelo mutacional es considerado. Se considera que la población de referencia se encuentra en equilibrio de Hardy-Weinberg, por lo tanto no se aplica parámetro de corrección por estructura poblacional. Se asume que la probabilidad de *drop-in* y *drop-out* son iguales a cero.

A continuación, se describen los dos posibles enfoques que serán abordados durante el capítulo.

5.2.2 Estrategia 1: métricas de poder estadístico

Considerando el flujo de toma de decisiones presentado en el capítulo 4, las métricas de PE y PI resultan centrales para la evaluación de los pedigríes (Vigeland et al. 2020, Kling et al. 2017). En este caso, la estrategia 1 consiste en analizar el PE y PI esperados para la incorporación de nuevos miembros mediante simulaciones computacionales. El algoritmo 4, presenta el pseudocódigo utilizado para la simulación. El algoritmo involucra, para cada PD analizado, definir un conjunto denominado PR , o de referencia. Estos son posibles miembros nuevos a incorporar en el pedigrí, y deben ser diferentes a los miembros ya genotipados. Para cada R (posición específica en el pedigrí), se simula un conjunto de M genotipos de PRs , o parientes de referencia. Es decir posibles genotipos compatibles con esa posición en el pedigrí. Posteriormente, para cada realización de PR , se generan N realizaciones de $PNIs$, considerando H_1 o H_2 como ciertas. Esto permite, para cada PR , calcular un PI y un PE. De esta manera, la distribución de PIs y PEs obtenida caracteriza a la posición indicada en R .

Supóngase que un determinado pedigrí cuenta con dos tíos paternos genotipados. En el mismo, se debe decidir sobre la incorporación de un tío materno, o de un abuelo paterno. Por lo

Algoritmo 3 Algoritmo de priorización

```
for  $j$  in ( $PD_1, PD_2, \dots, PD_K$ ) do Definir un conjunto R Simular  $M$  PBs
    for  $w$  in ( $PR_1, PR_2, \dots, PR_L$ ) do
        Simular  $N$  PNIs
        for  $i$  in ( $PNI_1, PNI_2, PNI_3, \dots, PNI_N$ ) do
            Muestrear  $g$  considering  $P(g|H)$ 
            Calcular PI
            Calcular PE
        end for
        Cálculo PI promedio
        Cálculo PE promedio
    end for
end for
```

tanto, $R = \{R_1, R_2\}$, siendo R_1 el tío materno y R_2 el abuelo paterno. Para cada pariente, se simula un total de 1000 realizaciones, obteniéndose 1000 genotipos compatibles con R_1 y 1000 compatibles con R_2 . Para cada uno de los genotipos compatibles, se generan 1000 realizaciones de PNIs considerando H_1 verdadera, y 1000 considerando H_2 verdadera. Este procedimiento produce 4 millones de genotipos de PNIs en total. Se obtiene un total de 1000 valores de PIs y 1000 de PEs para el tío materno, y otros 1000 PIs y PEs para el abuelo paterno. Una de las principales ventajas de este enfoque es que permite estudiar la distribución de estas métricas obtenidas a partir de cada realización. Además, el promedio de PI y PE permite caracterizar a cada R . Dentro de R se pueden definir subconjuntos de más de un miembro permitiendo comparar, por ejemplo, la incorporación de dos tíos maternos, contra la del abuelo paterno. Este tipo de preguntas emula condiciones de toma de decisiones reales (Marsico et al. 2021), permitiendo adaptarse a diferentes contextos.

5.2.3 Estrategia 2: cálculo del contenido de información

En esta sección se introducen las métricas de teoría de la información para el análisis de la evidencia genética.

5.2.3.1 Entropía

La entropía es una medida del grado de incertezza de una variable (Shannon 1948). En este caso las variables corresponden a los marcadores genéticos, STRs autosómicos, y los alelos son los posibles valores que estos pueden tomar. Para el análisis en cuestión, la entropía se define de la siguiente manera:

$$\mathcal{H}(G) = \sum p(g_i) \ln(p(g_i)) \quad (5.1)$$

Donde g_i es un genotipo específico. Si, para un determinado individuo, se conoce el genotipo (tipificado), se dirá que la incertezza es mínima, y que $p(g_i) = 1$, siendo g_i el valor del genotipo del individuo. \ln es el logaritmo con base natural. Considerando la ecuación, $H(G) = 0$. En casos en los cuales las variables son independientes, como lo son los marcadores genéticos, la entropía es aditiva sobre los distintos marcadores.

En casos de búsqueda de personas desaparecidas, la probabilidad de observar un genotipo,

considerando H_2 cierta, es la frecuencia relativa del genotipo en la población de referencia. En cambio, al incorporar información del pedigrí y evidencia genética, la probabilidad se ve condicionada, es decir, considerando H_1 . La probabilidad de la evidencia bajo cada hipótesis puede analizarse como dos distribuciones de probabilidades diferentes. A continuación se presenta la variación de entropía como una métrica para comparar el contenido de incertezas de ambas distribuciones, ΔH :

$$\Delta H = \mathcal{H}_G|H_1 - \mathcal{H}_G|H_2 = \sum p(g_i) \ln(p(g_i)) - \sum q(g_i) \ln_2(q(g_i)) \quad (5.2)$$

Donde $p(g_i)$ es la probabilidad del genotipo i considerando H_1 cierta, y $q(g_i)$ la probabilidad del mismo genotipo considerando H_2 cierta.

5.2.3.2 Divergencia Kullback-Leibler

La divergencia de Kullback-Leibler, o divergencia KL, mide la diferencia entre dos distribuciones de probabilidad. Esta es una medida asimétrica, siendo $|D_{KL}(x, y)| \neq |D_{KL}(y, x)|$. A continuación se presenta la ecuación.

$$KL = D_{KL}(P(g_i)|Q(g_i)) = \sum_i p(g_i) \ln\left(\frac{p(g_i)}{q(g_i)}\right) \quad (5.3)$$

y

$$\overline{KL} = D_{KL}(Q(g_i)|P(g_i)) = \sum_i q(g_i) \ln\left(\frac{q(g_i)}{p(g_i)}\right) \quad (5.4)$$

En la primera ecuación se presenta la divergencia entre el genotipo condicionado por el pedigrí y la población. En la segunda se introduce la divergencia inversa, o sea entre la población y la condicionada por el pedigrí. Por simplicidad, se adquiere la siguiente notación, $D_{KL}(P(g_i)|Q(g_i))$ es llamada KL , y $D_{KL}(Q(g_i)|P(g_i))$, \overline{KL} . En ambos casos se utiliza el logaritmo natural, por lo tanto las unidades serán los nats.

5.2.3.3 Entropía cruzada

La entropía cruzada mide también la diferencia entre dos distribuciones, pero incorporando la entropía de la propia distribución desde la cual se plantea. Está directamente relacionada con las dos métricas previamente presentadas, siendo su expresión:

$$Cross\mathcal{H} = \mathcal{H}(G) + D_{kl}(P(g_i)|Q(g_i)) \quad (5.5)$$

y en la otra dirección:

$$\overline{Cross\mathcal{H}} = \mathcal{H}(G) + D_{kl}(Q(g_i)|P(g_i)) \quad (5.6)$$

5.2.4 Vinculando CV y la divergencia de Kullback-Leibler

En esta sección se retoman expresiones matemáticas ya introducidas con el fin de mostrar su intrínseca relación. Por un lado CV_G representa el cociente entre dos probabilidades, $P(G|H_1)$ y $P(G|H_2)$. Generalmente, el CV_G se presenta como su logaritmo en base diez con el fin de

obtener una visualización más clara de las distribuciones, siendo:

$$\text{Log}(CV_G) = \text{Log}_{10} \left(\frac{P(G|H_1)}{P(G|H_2)} \right) \quad (5.7)$$

En esta misma sección, $q(g)$ y $p(g)$ fueron introducidas también como distribuciones de probabilidad de los genotipos bajo H_1 y H_2 verdaderas. Por lo tanto, redefiniendo el logaritmo de CV_G con esta nomenclatura se puede escribir de la siguiente manera:

$$\text{Log}(LR) = \text{Log}_{10} \left(\frac{p(g)}{q(g)} \right) \quad (5.8)$$

Donde $q(g)$ y $p(g)$ pueden obtenerse mediante simulaciones, o de forma directa mediante el empleo de redes bayesianas (Chernomoretz et al. 2022). La distribución de valores de CV_G para ambas hipótesis ha sido analizada en profundidad previamente (Marsico et al. 2021). Una métrica relevante es el promedio de valores de $\text{Log}_{10}(CV_G)$ obtenidos considerando H_1 o H_2 como ciertas. Con la nomenclatura introducida se redefine el promedio de $\text{Log}_{10}(CV_G)$ esperado considerando H_1 como verdadera de la siguiente manera:

$$\langle \text{Log}(CV_G) \rangle = p(g_i) \sum_{G=g_i} \text{Log}_{10} \left(\frac{p(g_i)}{q(g_i)} \right) = p(g_i) \sum_{G=g_i} \text{Log}_{10}(CV_G) \quad (5.9)$$

Importantemente, esta expresión está directamente vinculada con la de KL (ecuación 5.3), siendo la única diferencia la base del logaritmo. Por otro lado, la expresión de la media de $\text{Log}_{10}(\overline{CV_G})$ esperada con H_2 como verdadera es presentada a continuación:

$$\langle \text{Log}(CV_G^{-1}) \rangle = q(g_i) \sum_{G=g_i} \text{Log}_{10} \left(\frac{q(g_i)}{p(g_i)} \right) = q(g_i) \sum_{G=g_i} \text{Log}_{10}(CV_G^{-1}) \quad (5.10)$$

Esta expresión es similar a la de \overline{KL} (ecuación 5.4). Además del cambio de base, en dicha ecuación el logaritmo se aplica sobre CV_G^{-1} .

Esto permite no solo abordar el problema desde un enfoque de teoría de la información, si no también obtener una interpretación directa de su significado.

5.3 Resultados

En esta sección se muestran los resultados de ambos métodos planteados. Se utilizan escenarios simulados concretos que emulan condiciones reales de toma de decisiones.

5.3.1 Estrategia 1: métricas de poder estadístico

Se presentan una serie de casos donde se aplicó la estrategia 1, con el objetivo de priorizar posibles incorporaciones de miembros a pedigree con bajo poder estadístico (Kling et al. 2017, Vigeland et al. 2020). Todas las simulaciones genotípicas, a menos que se aclare otra cosa, consisten en un set de 23 marcadores STRs autosómicos, actualmente empleados como estándar

de los laboratorios forenses (Kling et al. 2017, Marsico et al. 2021).

5.3.1.1 La importancia de las simulaciones condicionales

En este primer caso se demuestra la importancia del empleo de simulaciones condicionales, producidas mediante la estrategia 1. En el pedigrí de referencia de la Figura 5.1, los genotipos ya disponibles incluyen a una abuela y a un tío. El poder estadístico para identificar a PD depende de los genotipos de dichos dos miembros. Si estos poseen alelos de baja frecuencia en la población, el poder de inclusión será mayor que si tuviesen alelos muy frecuentes. En la Figura 5.1 B, en blanco, se muestra el poder estadístico considerando dos posibles configuraciones genotípicas de los miembros ya tipificados (tío y abuela). El círculo muestra la configuración genotípica con alelos más frecuentes en la población de referencia, y por lo tanto, de menor poder estadístico, y el cuadrado una de mayor poder. En color naranja se muestra el efecto de la incorporación de un segundo tío, Tío 2, los círculos o cuadrados más grandes representan el centroide, mientras que los más pequeños, valores específicos para posibles genotipos del mismo. Con el objetivo de mostrar cómo estas diferencias de base impactan en el problema de priorización, se generaron 20 genotipos posibles para Tío 2. Se estimó el poder estadístico para cada genotipo, y se analizó el promedio obtenido. Esto se muestra en las Figuras 5.1 B. Puede verse cómo, en el caso de los genotipos de base con menor poder estadístico, la incorporación del Tío 2 continúa quedando por debajo de un valor aceptable de CV_G esperado en una identificación (panel B, PI menor a 0,5). En cambio, en el caso de los genotipos con mejor poder estadístico (cuadrado), la incorporación del tío deja al pedigrí en mejores capacidades (PI cercano a 0,75). En ambos casos el PE aumenta considerablemente.

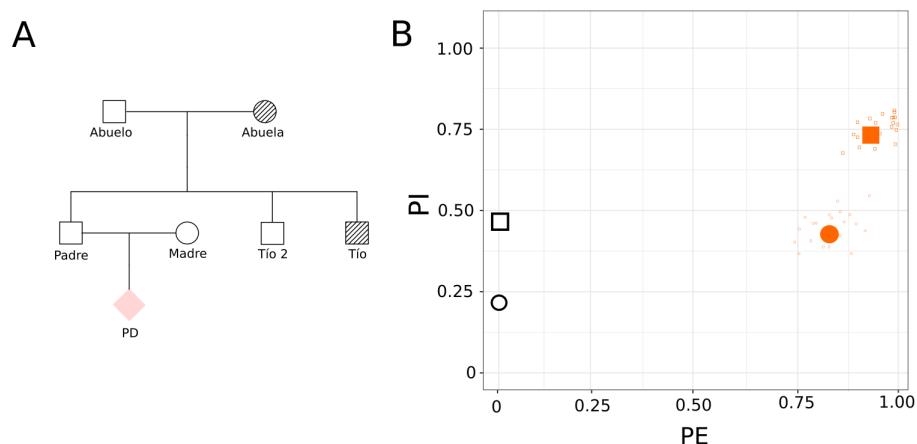


Figura 5.1 Análisis del poder estadístico para dos posibles combinaciones de genotipos en los familiares de referencia. (A) El pedigrí de referencia, donde Abuela y Tío son genotipados, y Tío 2 es una adición potencial. (B) Gráfico de poder estadístico. Los dos símbolos naranjas más grandes representan el poder promedio estimado para la incorporación del Tío 2. Cada punto pequeño consiste en el poder estimado para un Tío 2 específico.

5.3.1.2 Un problema de priorización más complejo

La Figura 5.2 A muestra un pedigrí con los mismos miembros genotipados que el caso de la Figura 5.1, pero con una extensión de posibles miembros a incorporar. Como se mostró previamente, la línea basal incluye solo información de un tío y una abuela genotipados, pero en este caso, mediante simulaciones, se estudia la incorporación de cinco posibles miembros o combinaciones de miembros considerando un abuelo, bisabuelos, tía abuela y otro tío. La pregunta a responder es la siguiente: ¿Cuál de estos miembros o combinación de miembros debería ser genotipado? La respuesta intuitiva podría ser que el abuelo es quien aporta más información, por su cercanía a la persona desaparecida. Aún así, otros aspectos pueden generar dudas:

- ¿Son los bisabuelos, en combinación, tan informativos como el abuelo?
- ¿Cuál es la importancia relativa de la incorporación de un bisabuelo o de un tío abuelo?
- ¿Cuál es la ganancia de incorporar ambos bisabuelos y la tía abuela, en vez de tan sólo una de ambas alternativas?
- ¿Cuán informativo es el Tío 2, comparado con las alternativas previas?

Puede observarse, además, que con ambos parientes genotipados, no es posible excluir PNIs no relacionados, debido a que no se conocen con certeza ambos alelos paternos en ningún marcador. Considerando lo discutido en el capítulo 4, en ciertos contextos, de búsqueda en bases de datos genéticas, el poder de exclusión resulta fundamental en el proceso de toma de decisiones para disminuir el número de falsos positivos. Esto puede derivar en una priorización en términos de incrementar el poder estadístico.

Con el objetivo de responder las preguntas previamente planteadas, se prosigue con la estrategia 1. Para cada combinación de miembros a incorporar se simulan 20 genotipos de *PR*. Para cada *PR* se simulan 1000 PNIs considerando H_1 como cierta, y 1000 PNIs considerando H_2 como cierta. Los resultados se resumen en la Figura 5.2 B. Por ejemplo, puede observarse que el Abuelo solo es sustancialmente más informativo que incluir ambos bisabuelos. De hecho, la Figura 5.2 B, indica que la adición de tan solo el abuelo transforma al pedigrí de ser uno con muy poco poder estadístico, a uno con un PI igual a 0,87 y PE igual a 1. Más aún, puede verse que la incorporación del bisabuelo y de la tía abuela dan valores muy similares en términos de PI, pero solo el bisabuelo permite la exclusión. Esto muestra la necesidad de observar ambos valores para la priorización. Finalmente, el tío aporta solo un poco más de información que el bisabuelo.

5.3.1.3 Búsqueda de bisnietos

La inspiración de este ejemplo proviene de una serie de casos emergentes en la búsqueda comenzada por las Abuelas de Plaza de Mayo (Penchaszadeh 1997), y consiste en la identificación de los bisnietos (Vigeland et al. 2020). En un escenario hipotético planteado a modo de ejemplo, la posible nieta desaparecida tuvo un total de cuatro hijos con el mismo padre. Esta persona fallece, y sus restos son cremados, por lo tanto no es posible acceder a muestras de

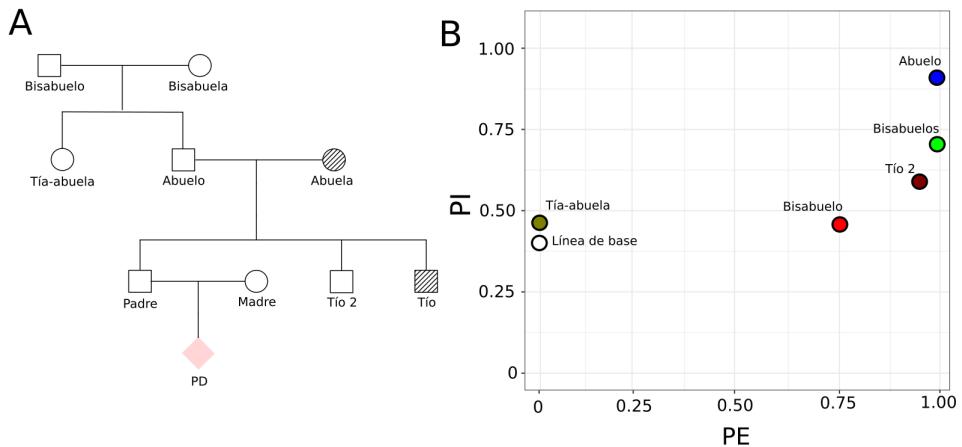


Figura 5.2 Análisis estadístico de la incorporación de distintos posibles miembros al pedigrí de referencia (A). Se presentan las métricas PI y PE (B) y el CV_G promedio esperado acompañado del número de marcadores excluidos esperados (con H_2 cierta) (C).

ADN. Cualquiera de sus hijos podrían ser bisnietos en los casos no-resueltos de las Abuelas. Las Figuras 5.3 B-D investigan distintos pedigríes, en los cuales se evalúa el poder estadístico frente a distintas configuraciones de hijos y pareja de la posible nieta desaparecida. Además, se incorpora una configuración hipotética (círculo verde) que indica el resultado posible si se contase con el material genético de la posible nieta.

La estrategia de simulación utilizada difiere levemente de la planteada en la metodología. Esto sucede porque los posibles miembros a incorporar, cuyos genotipos se simulan, son parientes de PNI, y no de PD, como en los otros casos. Aún así, el procedimiento es exactamente el mismo en el resto de los aspectos. En cada uno de los 3 pedigríes analizados, un conjunto de genotipos simulados es asignado a los familiares que se consideran ya tipificados, emulando casos reales de búsqueda. Para cada combinación de parientes de PNI posibles de ser genotipados se generó una simulación con 1000 realizaciones, considerando H_1 como cierta, y 1000 considerando H_2 como cierta. Utilizando estos datos, se calcularon las métricas de poder estadístico.

La Figura 5.3 B muestra un caso donde el pedigrí de referencia incluye solo a dos abuelos de PD. Como se muestra con el círculo verde, si se contase con la información genética correspondiente a la posible nieta, se obtendría un buen PI (cercano a 1) y un PE estandarizado de 0,5. Si se contase con datos de solo un hijo de la posible nieta, el poder estadístico daría muy bajo, cercano a $PE = PI = 0$ (Vigeland et al. 2020). La clave para obtener un poder estadístico mayor consiste en incorporar a la pareja de la nieta desaparecida. Aunque no posee un vínculo biológico con la misma, permite discernir entre aquellos alelos pertenecientes a sus hijos, que provienen de la madre. El mejor poder estadístico se obtiene con datos de la pareja y de sus 4 hijos.

La Figura 5.3 C muestra un pedigrí con dos tíos paternos y dos tíos maternos. El poder estadístico regular, con la potencial desaparecida genotipada, sería bueno (PE cerca de 0,8 y PI cerca de 0,9). Si solo un hijo de la posible nieta se incorpora, el poder estadístico decae drásticamente, llegando a una situación muy mala (PE y PI cercanos a cero). Como es de esperarse, a medida que más miembros se incorporan el poder estadístico crece. Al igual que en el caso previo, la incorporación de la pareja de la potencial desaparecida resulta fundamental (Vigeland

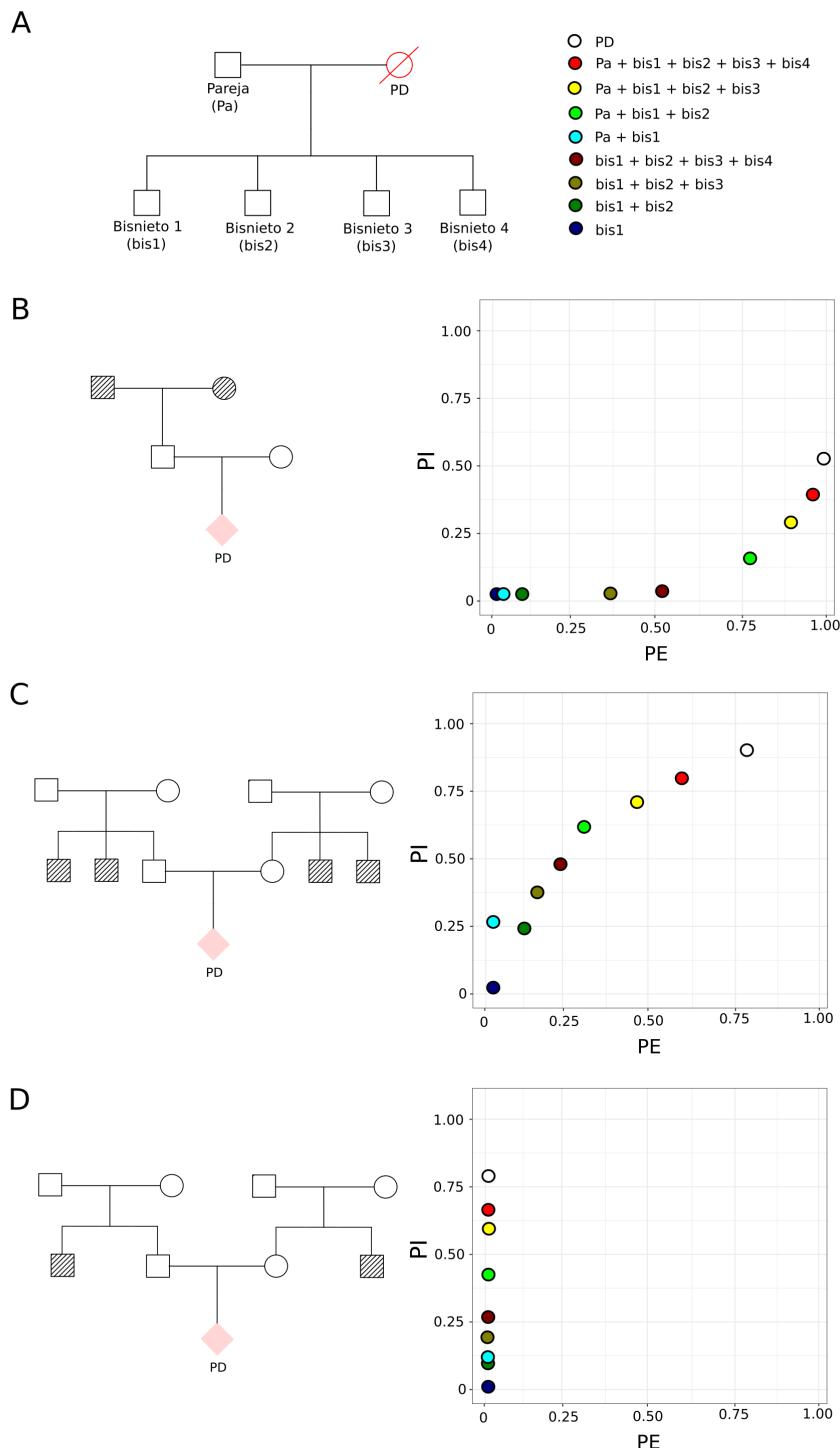


Figura 5.3 Análisis estadístico del caso de los bisnietos. Se muestran pedigree de referencia en B, C y D. En A se muestran las distintas posibilidades de grupos de individuos a identificar. Si PNIs que son potenciales PD están disponibles (círculo blanco), el caso no dista de los previamente analizados en distintos ejemplos de este capítulo. En cambio, las otras opciones muestran combinaciones de posibles individuos a identificar, como por ejemplo un par de bisnietos (bis 1 + bis 2).

et al. 2020).

La Figura 5.3 D difiere a las dos anteriores en el hecho de que no es posible excluir (PE igual a cero) en todos los contextos. Por lo tanto la incorporación de más miembros solo mejora el PI, llegando a valores cercanos a 0,7. Este caso muestra que sería necesario un esquema de priorización, inclusive contando con los datos de todos los hijos de la potencial desaparecida.

Es interesante notar que aunque en este último caso no se cuenta con PE, distintas estrategias genéticas podrían utilizarse para ganar capacidad de discernimiento, como la incorporación de análisis de ADN mitocondrial, por la presencia de la tía materna, y considerando que la persona desaparecida es mujer, y por lo tanto transmite su ADN mitocondrial a la descendencia.

5.3.1.4 Expandiendo el *set* de marcadores

Una estrategia común para incrementar el poder estadístico en el problema de la búsqueda de personas desaparecidas implica la retipificación de individuos de referencia con un mayor número de marcadores STRs. En este proceso, es importante estimar la ganancia, considerando que implica un costo, tanto por el análisis de los miembros del pedigrí, como por parte de los PNIs que se analicen contra el mismo. Además, puede suceder que se requiera comparar distintos *kits* disponibles.

Para ejemplificar este tipo de análisis, se utiliza como referencia al grupo familiar presentado en la Figura 5.4 A. En la misma, tres individuos están genotipados. Se cuenta inicialmente con un total de 15 marcadores STRs. Se analiza la posibilidad de incorporar nuevos marcadores hasta llegar a 23, o bien llegar a 33. Como se observa en la Figura 5.4 B, la condición inicial da un PI muy bajo, cercano a 1, y un PE levemente mayor a 0,5.

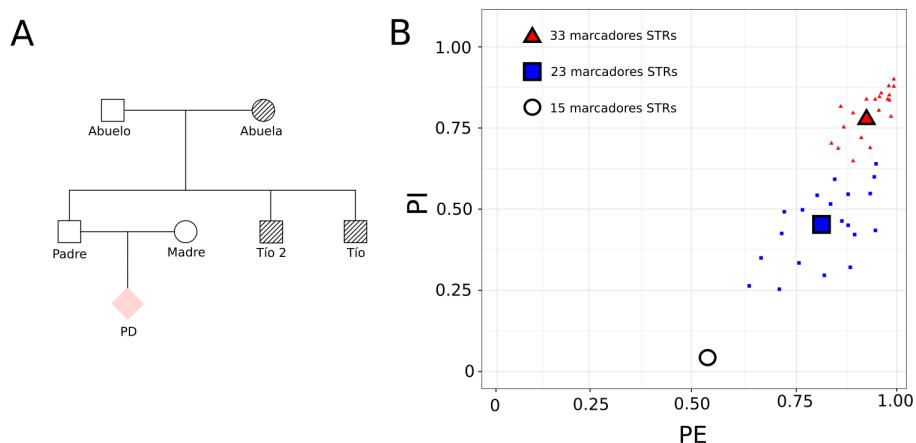


Figura 5.4 Análisis estadístico de la incorporación de nuevos sets de marcadores STRs autósomicos. Los puntos (triángulos o cuadrados) más pequeños muestran resultados de genotipos específicos simulados. Los puntos más grandes, cuadrado azul, triángulo rojo y círculo blanco, muestran los valores promedio para 23, 33 o el basal de 15 marcadores respectivamente.

La estrategia de simulación es la misma que la planteada en la sección metodología, excepto en el hecho de que los individuos simulados permanecen siendo los mismos y lo que varía es el número de marcadores. Los resultados se muestran en las Figuras 5.4 B. Como puede verse, hay

muchas variaciones dependientes de los genotipos que pueden obtenerse a partir de la incorporación de nuevos marcadores. Aún así, es claro que extender el *set* a un conjunto mayor de marcadores deriva en un incremento del poder estadístico. Aún así, este método permite cuantificar la mejora. Para este caso, un PI de 0,5 con 23 marcadores puede resultar insuficiente para una identificación. Contrariamente, el PE cercano a 0,8 sí podría ser muy útil para funcionar como un primer filtro con PNIs de 23 marcadores, y luego extender a 33 (utilizando por ejemplo un umbral de decisión, como lo planteado en el capítulo 4), en un conjunto de casos que merezcan mayor atención.

5.3.2 Estrategia 2: cálculo del contenido de información

En esta sección se muestran los resultados de una estrategia de priorización diferente. El objetivo es hacer hincapié en el enfoque teórico, más que en la solución práctica de los problemas, que se entiende presentada en la Estrategia 1. Con este fin se introducen ejemplos distintos, mostrando en detalle lo que implica el cálculo del contenido de información. Al finalizar se compara esta metodología con la previamente descrita.

5.3.2.1 Analizando el contenido de información de distintos marcadores

Previamente se mencionó que la incorporación de más marcadores implica un aumento del poder estadístico. Aún así, no se hizo foco en el aporte de marcadores específicos, o acerca de la diferencia entre los mismos. En este análisis se estudia el contenido de información de marcadores específicos. Se toma un ejemplo concreto, presentado en la Figura 5.5 B, donde se cuenta con información genética de la abuela paterna y del tío paterno. En la Figura 5.5 A se muestra el esquema general empleado por el análisis mediante redes bayesianas (explicado en detalle en el capítulo 3). La información presente en el pedigree proporciona un esquema para condicionar la distribución de probabilidad de los genotipos de la persona desaparecida. Si esta evidencia no está presente, las probabilidades genotípicas se obtienen de la población de referencia. En las Figuras 5.5 E y F puede verse al marcador genético STR autosómico denominado CSF1PO, y las probabilidades condicionadas para PD. El mismo cuenta con doce posibles alelos presentes en la población de referencia. En el panel C se ve como la distribución de probabilidades condicionada en el genotipo de la madre (alelo materno) no difiere de la de la población de referencia. Esto sucede debido a que no hay evidencia por la vía materna. En cambio, la Figura 5.5 D muestra la distribución de probabilidades para el alelo paterno, mientras que la poblacional continúa siendo la misma, la condicionada en el pedigree por vía paterna varía. Esto sucede gracias al aporte de la evidencia genética de la abuela paterna y el tío paterno. Las Figuras 5.5 C y D exploran lo mismo, pero para el marcador SE33. Este cuenta con una mayor cantidad de posibles valores alélicos, además de múltiples valores de mayor frecuencia, lo que es indicativo de diversidad genética en la población. El panel E, en consonancia con el C, muestra que no hay diferencia entre la distribución de probabilidades asociada al alelo de herencia materna, y aquél de la población de referencia. En cambio, para el alelo paterno, panel D, se observa una marcada preferencia hacia valores específicos. Puede observarse como las distribuciones en el panel D difieren entre sí en mayor medida que aquellas distribuciones del panel F. En la tabla 5.1 se ahonda, de forma cuantitativa, este aspecto. Para ambos marcadores puede verse cómo

las métricas de información dan por encima de cero, demostrando el incremento del contenido de información producido por la incorporación de evidencia genética. En concordancia con los resultados previamente descritos, SE33 presenta valores mayores para todas las métricas. Esto implica que el mismo resulta más informativo. El resultado es intuitivo, debido a la mayor cantidad de alelos posibles para SE33 y su distribución probabilística.

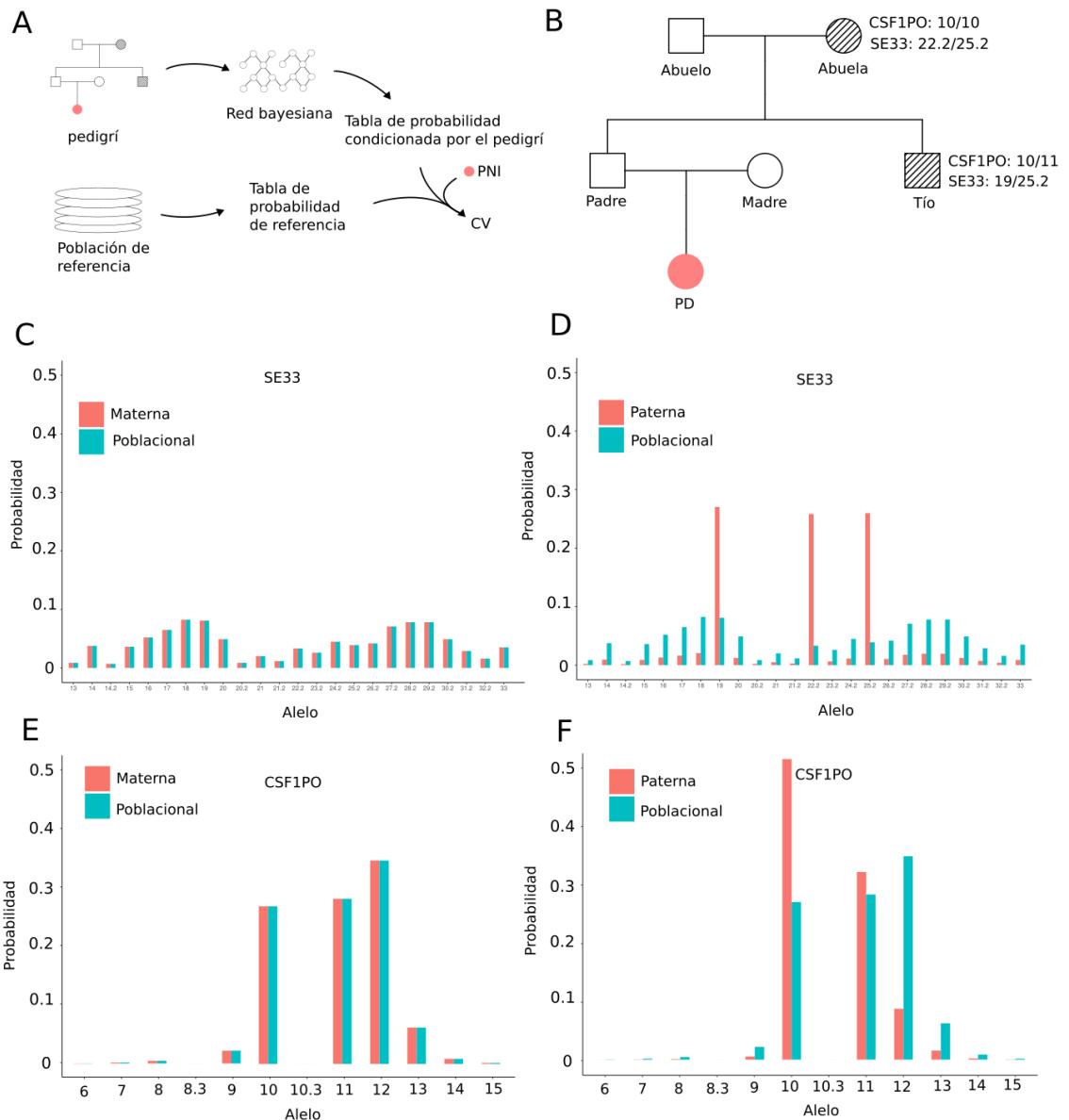


Figura 5.5 (A) El flujo de análisis generado por fbnet. (B) Estructura del pedigrí, los individuos punteados están genotipados. (C) SE33 tabla de probabilidad condicionada (alelo materno). (D) SE33 tabla de probabilidad condicionada (alelo paterno). (E) CSF1PO tabla de probabilidad condicionada (alelo materno). (F) CSF1PO tabla de probabilidad condicionada (alelo paterno).

5.3.2.2 Analizando la informatividad de los parientes

En esta sección se analiza un caso hipotético, presentado en la Figura 5.6. Se consideran un total de 23 marcadores STRs. Para cada familiar, se simula un genotipo específico, emulan-

Marker	ΔH	$CrossH$	\overline{CrossH}	KL	\overline{KL}
CSF1PO	0.27	2.17	2.47	0.10	0.12
SE33	0.64	5.45	6.11	0.73	0.75

Tabla 5.1 Métricas de información para los marcadores STRs CSF1PO y SE33 para el pedigrí de la Figura 5.5 B.

do a un caso real. Se computan posteriormente las métricas de información para los distintos miembros o combinaciones de miembros (Tabla 5.2). Nótese que en este caso no es necesario realizar simulaciones de genotipos de PNIs considerando H_1 y H_2 ciertas. Las métricas de información solo comparan la distribución de probabilidades genotípicas previa y posteriormente a incorporar la evidencia.

Para las métricas ΔH , \overline{CrossH} , KL y \overline{KL} se observan mayores valores al incorporarse los parientes 3 y 4, como era de esperarse, por tratarse de los más cercanos a PD (exceptuando al padre y la madre). Estos son por lo tanto los más informativos. No se observan diferencias al incorporar 1 y 3, versus incorporarse solo 3. Esto marca la redundancia de 1 en este contexto. Incorporar 1 y 2 resulta en mayor información que incorporar solo 1, como también era de esperarse. En comparación, 8 incorpora menos información que 3. En cambio, para $CrossH$, los mayores valores se obtienen con 1, seguido de 1 y 2. El tío, 8, presenta mayores valores que 3, y por último, los abuelos 3 y 4 presentan los menores valores.

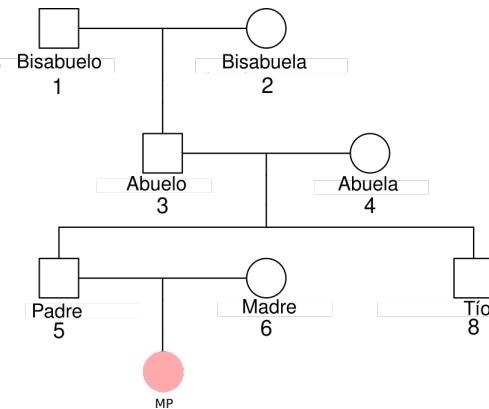


Figura 5.6 Ejemplo de pedigrí con tres generaciones de ancestros de PD.

Familiares	ΔH	$CrossH$	\overline{CrossH}	KL	\overline{KL}	$Log_{10}(CV)$	$Log_{10}(\overline{CV})$
3	5.27	72.09	76.92	4.43	4.00	1.93	1.77
1, 3	5.27	72.09	76.92	4.43	4.00	1.93	1.77
3, 4	11.79	71.60	206.28	10.46	133.35	5.54	13.41
1, 2	3.00	72.74	75.34	2.81	2.41	1.34	0.03
8	4.79	72.18	76.65	4.04	3.72	1.75	1.57
1	1.32	73.12	74.04	1.52	1.12	0.84	0.85

Tabla 5.2 Métricas de contenido de la información para el pedigrí presentado en la Figura 5.6. Se indica el familiar o combinación de familiares genotipados.

Al comparar con $E(Log_{10}(CV_G))$ y $E(Log_{10}(\overline{CV}_G))$ el abuelo y la abuela (3,4) presentan

los mayores valores esperados, seguido de solo el abuelo (3) o del abuelo con el bisabuelo (1,3). El tío (8) también muestra valores menores que el del abuelo, 3. Por último, los menores valores son obtenidos para el bisabuelo. Este patrón es el mismo que el observado para las métricas ΔH , \overline{CrossH} , KL y \overline{KL} . Este resultado es concordante con la vinculación directa planteada entre CV_G y KL .

5.3.2.3 Incorporando parientes al pedigrí

En este caso, se plantea la posibilidad de incorporar nuevos parientes, sin contar con genotipos previos en el pedigrí. Las posibilidades son tres: un abuelo, un tío o un bisabuelo. Para cada PR , se genera una simulación de 10.000 posibles genotipos. Se calculan las métricas de informatividad para cada uno, obteniendo por lo tanto una distribución. Nuevamente, en este caso no es necesario simular, para cada PR , múltiples genotipos de PNIs, dado que el cálculo se realiza directamente sobre la distribución de probabilidades. Como es de esperarse, tanto el abuelo como el tío aportan mayor informatividad que el bisabuelo, con las métricas ΔH , \overline{CrossH} , KL y \overline{KL} . En contraposición, $CrossH$ deriva en valores similares para todas las combinaciones.

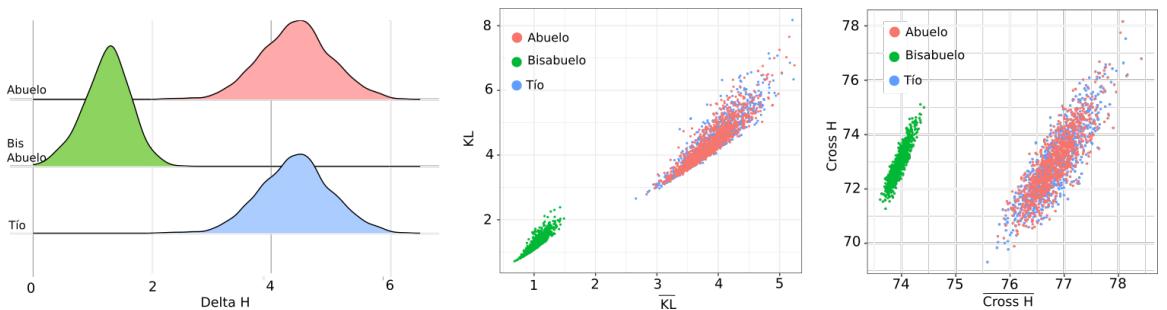


Figura 5.7 Análisis de distribución de métricas de informatividad para el tío, bisabuelo y abuelo. Izquierdo - Distribuciones de ΔH . Centro - Gráfico $KL - \overline{KL}$. Derecha - Gráfico $CrossH - \overline{CrossH}$

5.3.2.4 Tomando decisiones: ¿Más marcadores o más parientes?

Mientras los últimos dos ejemplos fueron ilustrativos, yendo a casos de baja complejidad en términos de toma de decisiones, esta sección se adentra en casos más complejos. Con el objetivo de simplificar, se analiza el gráfico $KL - \overline{KL}$. Estas métricas permitieron distinguir efectivamente los casos más informativos, además de vincularse directamente con el CV , por lo tanto con una interpretación intuitiva para la comunidad forense. Aún así, como se explica en métodos, las distintas métricas se encuentran fuertemente relacionadas entre sí.

En el primer ejemplo, analizado en la Figura 5.8, se muestra un caso donde una persona desaparecida se intenta identificar mediante una muestra obtenida de un tío. Previo a analizar el material genético con un *kit* de marcadores STRs, se estudia, simulando 10.000 posibles perfiles del tío, la distribución de valores de informatividad esperados, con 15 marcadores. Observando una gran dispersión, los forenses deciden genotipificar al tío con 15 marcadores, para dejar abierta la posibilidad a un resultado informativo (un tío con alelos de muy baja frecuencia por ejemplo que otorgue un buen poder estadístico). Una vez analizado (Figura 5.8 B), los forenses

se enfrentan al hecho de que el tío no posee alelos de muy baja frecuencia, sino más bien se ubica dentro de los valores promedio esperados para este familiar. Con este resultado, deben decidir entre gastar los recursos en aumentar el número de marcadores de 15 a 23, o bien realizar una exhumación y obtener el perfil genético del abuelo paterno, pero manteniendo el número de marcadores en 15. Para este fin, realizan 10.000 simulaciones condicionales del abuelo paterno, y de la posible adición de marcadores al tío ya genotipado.

Como puede verse en la Figura 5.8, tanto la incorporación de nuevos marcadores, como la del abuelo paterno manteniendo el número de marcadores, mejora las métricas KL y \overline{KL} . Importantemente, el agregado del abuelo mejora aún más \overline{KL} , mientras que mantiene similares valores de KL respecto al tío con 23 marcadores. Esto sugiere que la decisión de incorporar al abuelo con 15 marcadores podría ser la correcta.

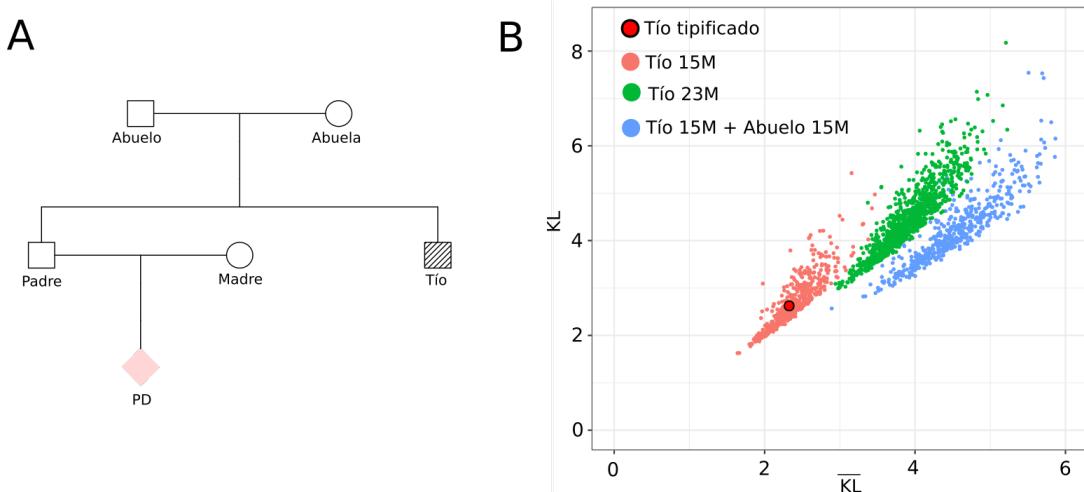


Figura 5.8 (A) Pedigrí analizado, donde el tío es genotipado. (B) Gráfico de $KL - \overline{KL}$ donde se indica el valor obtenido para el tío genotipado y las distribuciones obtenidas para los genotipos simulados de los posibles miembros a incorporar al pedigrí.

En un escenario más complejo, presentado en la Figura 5.9, se analizan distintas posibilidades de incorporación de nuevos miembros, considerando en todos los casos 23 marcadores STRs autosómicos. Como se muestra en el panel B, cuatro parientes diferentes son analizados para su potencial incorporación. En cada caso se construyen 10.000 posibles genotipos para los *PRs*. Las simulaciones para dar con los genotipos son condicionadas, dado que ya se cuenta con información de miembros tipificados. Como es esperado, ambos bisabuelos son los que aportan la menor información. La abuela aporta más información que el abuelo. Esto puede entenderse debido a que ya se encuentra genotipado un tío abuelo, que aporta información parcial del abuelo. O dicho de otra manera, como se ha analizado en un ejemplo anterior en esta misma sección, el tío abuelo se vuelve redundante frente a incorporación del abuelo. Esto no sucede si se incorpora la abuela. Por último, el que peores métricas presenta es la incorporación de un tío abuelo más.

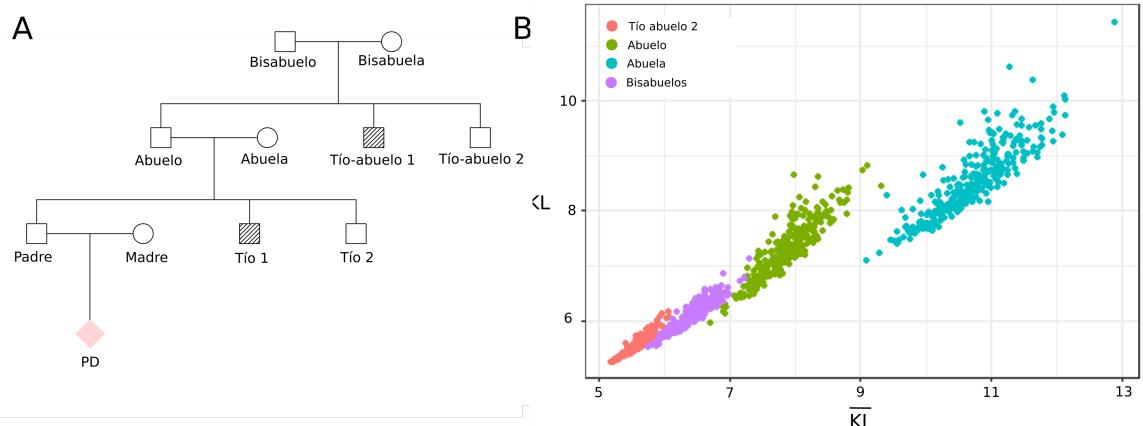


Figura 5.9 (A) Pedigrí analizado, en rayas se muestra a los miembros genotipados. (B) Gráfico de $KL - \overline{KL}$ donde se indican las distribuciones obtenidas para los genotipos simulados de los posibles miembros a incorporar al pedigrí.

5.4 Discusión

En este capítulo se han propuesto dos estrategias diferentes para el problema de priorización en búsqueda de personas desaparecidas. El problema en sí es multifactorial (Puerto et al. 2021), dependiendo de aspectos que pueden ser formalizados y generalizados, como la mejora en capacidad identificatoria basada en ADN, y otros que no, como la accesibilidad a restos de familiares del desaparecido. En este sentido, basándose en el aspecto genético, las estrategias buscan medir qué miembros pueden mejorar la capacidad de identificar a PD, o bien de descartar a un individuo que no es PD. De este modo, la principal pregunta que se busca responder es: ¿Qué miembro aporta mayor información genética para la identificación en un pedigrí dado? Esta es una subparte de la pregunta de priorización.

En consonancia con lo discutido en el capítulo 4, la toma de decisiones en casos de búsqueda de personas desaparecidas puede ser vista como un problema de dos direcciones. Por un lado, la definición de sí PNI es PD, es decir, concluir H_1 . Por otro lado definir si PNI no es PD, o sea, concluir H_2 . Aunque puede no ser intuitivo, dado que ambas opciones son mutuamente excluyentes, la capacidad de un pedigrí para concluir H_1 o H_2 es diferente. En la estrategia 1 esto se materializa mediante el PI, que hace referencia a H_1 , y el PE, que hace referencia a H_2 . Como se ha observado en los resultados, existen casos donde PI es alto y PE es bajo, y viceversa. Un PI alto indica un valor esperado alto en una identificación, es decir un verdadero positivo, pero también puede encontrarse relacionado con valores altos cuando PNI no es PD, o sea un falso positivo. En contraposición, aquellos pedigríes con valores altos de PE, pero bajos de PI, podrán excluir de forma eficiente aquellos PNIs que no son PD, pero no obtendrán valores altos de CV_G frente a una identificación (Vigeland et al. 2020).

Por distintos motivos, las simulaciones computacionales presentan escenarios simplificados de la realidad. Esto se debe a un conjunto de asunciones que deben ser tenidas en cuenta a la hora de interpretar los datos (Kling et al. 2017). Por un lado, se asume que las frecuencias alélicas de la población de referencia son suficientemente fehacientes. Esta asunción se vuelve difícil de cumplir en casos donde la frecuencias de ciertos alelos son muy bajas (Egeland et al. 2015).

Como se ha mencionado, alelos de frecuencia baja son además características valiosas dentro del cálculo del PI. Por otra parte, la estrategia basada en simulaciones computacionales para la priorización implica que para cada PR debe muestrearse una elevada cantidad de genotipos, y para cada uno de estos genotipos, es necesario simular PNIs. Esto deriva en una rápida escalada en la cantidad de realizaciones a producirse en la simulación, elevando el costo computacional y de tiempo (Vigeland et al. 2020).

La segunda estrategia difiere en varios aspectos de la primera. Por un lado, las métricas de información propuesta no requieren la definición de un valor de umbral para CV_G . Además, no son necesarias las simulaciones de PNIs para su cálculo. Este punto no es trivial, debido a que disminuye fuertemente el costo computacional, haciendo posible orientar el mismo a una mayor exploración de los posibles genotipos de los familiares a incorporar. En sintonía con la estrategia 1, nuevamente las decisiones se toman en dos direcciones, hacia H_1 y H_2 . Estas son captadas por dos métricas diferentes, por un lado KL , hacia H_1 , y \overline{KL} , para H_2 . Aunque de una interpretación más compleja, estas métricas dan cuenta de los valores de CV_G esperados en ambas direcciones. Además, esta metodología puede utilizarse para seleccionar entre la incorporación de más marcadores, y medir la informatividad de los mismos.

El problema de la priorización de la incorporación de nueva evidencia no se suscribe solo al área de la genética. La evidencia recolectada durante la investigación preliminar cumple un rol fundamental en el proceso de búsqueda (Caridi et al. 2020, 2011, Puerto et al. 2021). Similarmente a lo planteado en este capítulo, distintas líneas de evidencia pueden resultar clave para la resolución de casos. También, la investigación preliminar puede involucrar un arduo trabajo, tanto de recursos humanos, económicos como de tiempo debido al acceso a la documentación, entrevistas, etc. Una de las ventajas de la formalización matemática de la evidencia no genética, presentada en el capítulo 2, puede ser su inclusión en un esquema de priorización, utilizando metodologías muy similares a las planteadas para la evidencia genética.

5.5 Conclusiones del capítulo

En este capítulo se propuso un modelo general para la priorización de nuevos parientes a incorporar en los pedigríes de referencia. Se contextualiza utilizando escenarios simulados que buscaron emular condiciones reales de búsqueda. En este caso, se abordó la problemática mediante dos estrategias que implican marcos teóricos diferentes. Por un lado una basada en el cálculo de poder estadístico mediante simulaciones computacionales, y por otro una basada en la teoría de la información. Se pudo evaluar el impacto de incorporar miembros a los pedigríes, demostrando su utilidad en el contexto de priorización.

Capítulo 6

Conclusiones y perspectivas

En este apartado se repasan las principales conclusiones de cada capítulo de la tesis, y luego se realiza una conclusión general. En sí, la tesis se centra en el estudio de un mismo proceso, la búsqueda de personas desaparecidas, y la principal pregunta que busca responder es *¿es posible formalizar matemáticamente elementos de las distintas etapas del proceso de búsqueda con el fin de proponer modelos probabilísticos que integren la información recolectada?*

En el capítulo 2 se propusieron distintos modelos para la evaluación estadística de líneas de evidencia recolectadas durante la investigación preliminar. Se contextualiza además la información aportada en término de contraste de hipótesis bayesiano, comparando la hipótesis de identificación, es decir el individuo analizado es la persona buscada, contra la hipótesis de descarte, el individuo analizado no es la persona buscada. El enfoque propuesto deja abierta la puerta a la combinación de líneas de evidencia mediante la formalización de una probabilidad a priori para los datos genéticos, basada en un paso previo, de análisis de la evidencia no genética.

En el capítulo 3 se propone un modelo para el análisis de la evidencia genética basado en redes bayesianas. El mismo consiste en dos etapas, primero utiliza la información disponible de un pedigrí para condicionar la distribución de probabilidades genotípicas de la persona buscada. Se observa que dicho condicionamiento, en caso de ser informativo, diferencia la distribución de probabilidades respecto a la de la población de referencia, tomada como línea de base. En el segundo paso, estas distribuciones son utilizadas para el contraste de hipótesis bayesiana, en el cual se evalúa la verosimilitud de los datos observados para cada una de las hipótesis.

En el capítulo 4 se propone un modelo general basado en simulaciones computacionales para la optimización de la toma de decisiones en la búsqueda de personas desaparecidas mediante bases de datos. Además, se analiza el rol de la formalización de los datos recolectados durante la investigación preliminar, permitiendo cuantificar el impacto que tiene dicha línea de evidencia en la toma de decisiones. Se discuten distintas estrategias para el establecimiento de umbrales de decisión que permitan seleccionar al subgrupo de casos donde incorporar más análisis, teniendo en cuenta el balance entre la probabilidad de falsos negativos y falsos positivos.

En el capítulo 5 se propone un modelo para la priorización de nuevos parientes a incorporar en los pedigríes de referencia. La problemática se aborda mediante dos estrategias que implican marcos teóricos diferentes. Por un lado una estrategia similar a la desarrollada en el capítulo 4, basada en simulaciones computacionales. Por otro lado se estudia el empleo de métricas de teoría de la información para cuantificar el impacto de la incorporación de nueva evidencia.

En su conjunto, estos capítulos abordan distintos aspectos en el proceso de búsqueda de personas desaparecidas. Particularmente, se estudió la formalización matemática tanto de la

evaluación de la evidencia, como de la toma de decisiones realizada por los científicos forenses. Respecto a la formalización de la evidencia, el capítulo 2 se centró en los datos recolectados durante la investigación preliminar, y el 3 en aquellos datos genéticos. Más aún, se evaluó la posibilidad de combinar ambas líneas de evidencia. Respecto a la toma de decisiones, el capítulo 4 analiza el proceso realizado por el forense en el empleo de búsquedas mediante bases de datos. Luego, en el capítulo 5, se analiza el problema de priorización de nueva evidencia a recolectar para la resolución de casos. De esta forma, la tesis busca resaltar la complejidad del proceso de búsqueda, en el cual interfieren múltiples factores que hacen al análisis de la evidencia, pero también a la interpretación y contextualización de la misma llevada a cabo por el científico forense. Para contribuir a una Ciencia Forense que permita brindar trazabilidad y transparencia, es un requisito explorar en profundidad la formalización de cada paso, incluyendo aquellas decisiones tomadas *ad-hoc* por los investigadores basados en conocimiento experto y punto de vista subjetivo. El paradigma bayesiano permite otorgar un lugar específico a aquella subjetividad, conocida como la probabilidad a priori. Asimismo, permite integrar distintos pasos de ampliación de conocimiento y de interpretación de la evidencia. Esto se demuestra en los capítulos 2 y 3, dándose un contexto funcional en el capítulo 4. También, herramientas de la teoría de la decisión y teoría de la información permiten colaborar, tanto en la decisión particular como en la formalización de la toma de decisiones por parte del científico forense. Este aspecto resulta particularmente relevante para problemáticas modernas, donde el avance de la computación y del desarrollo de algoritmos permiten masificar y acelerar procesos judiciales. Esto conlleva a múltiples oportunidades, pero también desafíos. Los procesos de automatización muchas veces poseen inherentemente decisiones implícitas tomadas por los científicos forenses. Esto puede llevar a errores no-evidentes o cuya dimensión puede parecer despreciable hasta que es cuantificada. A modo de ejemplo, en el capítulo 2 se discuten las repercusiones de optar por una estrategia de filtrado de la base de datos basada en la información de la investigación preliminar. Otro ejemplo es el establecimiento de umbrales para la selección de casos considerar potenciales identificaciones o descartarlas como posibles. Aquellos casos que superan el umbral son sometidos a más análisis, mientras que los otros son apartados. Distintas estrategias para la selección de umbrales se proponen en la literatura, pero en este caso, el capítulo 4 propone una forma de cuantificar la probabilidad de error, y optimizar el umbral para minimizar el riesgo en cada pedigrí. Se discute cómo metodologías de selección de umbral previas podrían generar situaciones muy desfavorables para algunos casos (como colocar valores de corte que nunca podrían ser alcanzados en una identificación), sobre todo en contextos con grupos familiares heterogéneos. La toma de decisiones en torno a la priorización de nuevas evidencias recolectadas tiene un fin práctico en la resolución de los casos. Es importante tener en cuenta que en este tipo de problemáticas el tiempo juega un rol central, debido a que el robo de una identidad o bien a la imposibilidad de los familiares de despedir los restos de sus seres queridos es un fenómeno que genera un fuerte daño en los individuos y en la comunidad. Considerando esto, la generación de herramientas para guiar la investigación puede colaborar a evitar que la misma se vea retrasada.

Quedan líneas pendientes de investigación y desarrollo. La investigación preliminar recopila datos espacio temporales en torno a los hechos de desaparición que pueden ser sumamente informativos. Aún así, estas características pueden depender del contexto y de los casos de des-

aparición analizados, por lo tanto su generalización significa un desafío teórico y práctico. Por otro lado, se comentó acerca de otros tipos de marcadores genéticos, como el ADN mitocondrial, el cromosoma X o el cromosoma Y, que son de uso frecuente en identificaciones humanas. Aún así, existe una falta de modelos que permitan integrar la información recopilada por distintos tipos de marcadores. También, la aparición de nuevas tecnologías de análisis genético, como la secuenciación masiva paralela, permiten una mayor potencia estadística para las identificaciones. La ampliación del uso de estas tecnologías se encuentra aún en desarrollo.

Por último, durante la tesis múltiples desarrollos matemático-computacionales se han llevado a cabo. Salvo excepciones, la totalidad de los mismos derivaron en librerías actualmente disponibles en lenguaje R, en el repositorio CRAN (Marsico and Caridi 2023, Chernomoretz et al. 2022). Estas son de carácter de código abierto y con licencia GPL-3, lo que indica que pueden ser utilizadas, copiadas y modificadas, buscando colaborar con la transparencia y trazabilidad metodológica en el campo de las Ciencias Forenses.

Bibliografía

- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. et al. (2003), ‘Molecular biology of the cell’, *Scandinavian Journal of Rheumatology* **32**(2), 125–125.
- Allen, D. and Darwiche, A. (2008), ‘Rc_link: Genetic linkage analysis using bayesian networks’, *International journal of approximate reasoning* **48**(2), 499–525.
- Amendt, J., Richards, C. S., Campobasso, C. P., Zehner, R. and Hall, M. J. (2011), ‘Forensic entomology: applications and limitations’, *Forensic science, medicine, and pathology* **7**, 379–392.
- Amorim, A. and Budowle, B. (2016), *Handbook of forensic genetics: biodiversity and heredity in civil and criminal investigation*, Vol. 2, World Scientific.
- Aning, K. and McIntyre, A. (2004), ‘From youth rebellion to child abduction: The anatomy of recruitment in sierra leone’, *Invisible stakeholders: Children and war in Africa* pp. 67–86.
- Aronson, J. D. (2011), ‘The strengths and limitations of south africa’s search for apartheid-era missing persons’, *International Journal of Transitional Justice* **5**(2), 262–281.
- Balding, D. J. and Nichols, R. A. (1995), ‘A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity’, *Genetica* **96**, 3–12.
- Balding, D. J. and Steele, C. D. (2015), *Weight-of-evidence for Forensic DNA Profiles*, John Wiley and Sons.
- Ballard, D., Winkler-Galicki, J. and Wesoły, J. (2020), ‘Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects’, *International Journal of Legal Medicine* **134**(4), 1291–1303.
- Bao, X., Galiani, S., Li, K. and Long, C. (2019), Where have all the children gone? an empirical study of child abandonment and abduction in china, Technical report, National Bureau of Economic Research.
- Baraybar, J. P., Caridi, I. and Stockwell, J. (2020), ‘A forensic perspective on the new disappeared: migration revisited’, *Forensic science and humanitarian action: interacting with the dead and the living* pp. 101–115.
- Barlow, J. (2003), ‘Nexus small worlds and the groundbreaking theory of networks’, *Interface: The Journal of Education, Community and Values* **3**(8).

- Bateson, W. and Mendel, G. (2013), *Mendel's principles of heredity*, Courier Corporation.
- Benish, W. A. (2020), 'A review of the application of information theory to clinical diagnostic testing', *Entropy* **22**(1), 97.
- Berra, J., Grinspon, D., Liwski, N. and Binz, M. T. (1986), Genetical identification of "missing" children in argentina, in '11th Congress of the Society for Forensic Haemogenetics (Gesellschaft für forensische Blutgruppenkunde eV) Copenhagen, August 6–10, 1985', Springer, pp. 443–448.
- Biedermann, A. and Taroni, F. (2012), 'Bayesian networks for evaluating forensic dna profiling evidence: a review and guide to literature', *Forensic Science International: Genetics* **6**(2), 147–157.
- Biedermann, A., Taroni, F. and Margot, P. (2012), 'Reply to budowle, ge, chakraborty and gill-king: use of prior odds for missing persons identifications', *Investigative Genetics* **3**(1), 1–2.
- Borosky, A., Toscanini, U., Gómez, A., Parolín, M. L., Basso, N. and Vullo, C. (2014), 'Forensic population data for 20 str loci in argentina', *Forensic Science International: Genetics* **13**, e20–e21.
- Bosco, F. J. (2006), 'The madres de plaza de mayo and three decades of human rights' activism: Embeddedness, emotions, and social movements', *Annals of the Association of American Geographers* **96**(2), 342–365.
- Brenner, C. H. and Weir, B. S. (2003), 'Issues and strategies in the dna identification of world trade center victims', *Theoretical population biology* **63**(3), 173–178.
- Brettell, T., Butler, J. and Saferstein, R. (2005), 'Forensic science', *Analytical chemistry* **77**(12), 3839–3860.
- Budowle, B., Ge, J., Chakraborty, R. and Gill-King, H. (2011), 'Use of prior odds for missing persons identifications', *Investigative genetics* **2**(1), 1–6.
- Budowle, B., Moretti, T. R., Niezgoda, S. J. and Brown, B. L. (1998), Codis and pcr-based short tandem repeat loci: law enforcement tools, in 'Second European symposium on human identification', Vol. 7388, Promega Corporation, Madison, Wisconsin.
- Butler, J. M. (2005), *Forensic DNA typing: biology, technology, and genetics of STR markers*, Elsevier.
- Butler, J. M. (2007), 'Short tandem repeat typing technologies used in human identity testing', *Biotechniques* **43**(4), Sii–Sv.
- Caridi, I., Alvarez, E. E., Somigliana, C. and Puerto, M. S. (2020), 'Using already-solved cases of a mass disaster event for prioritizing the search among remaining victims: a bayesian approach', *Scientific reports* **10**(1), 1–11.

Caridi, I., Dorso, C. O., Gallo, P. and Somigliana, C. (2011), ‘A framework to approach problems of forensic anthropology using complex networks’, *Physica A: Statistical Mechanics and its Applications* **390**(9), 1662–1676.

Chernomoretz, A., Belparda, M., La Grutta, L., Calabrese, A., Martinez, G., Escobar, M. S. and Sibilla, G. (2020), ‘Genis, an open-source multi-tier forensic dna information system’, *Forensic Science International: Reports* **2**, 100132.

Chernomoretz, A., Marsico, F., Iserte, J., Piñero, M. H., Escobar, M. S., Belparda, M. and Sibilla, G. (2022), ‘Bayesian networks for dna-based kinship analysis: Functionality and validation of the genis missing person identification module’, *Forensic Science International: Genetics Supplement Series* **8**, 131–132.

Chicco, D. and Jurman, G. (2020), ‘The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation’, *BMC genomics* **21**(1), 1–13.

Cho, S., Shin, E. S., Yu, H. J., Lee, J. H., Seo, H. J., Kim, M. Y. and Lee, S. D. (2017), ‘Set up of cutoff thresholds for kinship determination using snp loci’, *Forensic Science International: Genetics* **29**, 1–8.

Citroni, G. (2017), ‘The first attempts in mexico and central america to address the phenomenon of missing and disappeared migrants’, *International Review of the Red Cross* **99**(905), 735–757.

Cohem Salama, M. (1992), *Tumbas Anonimas*.

CONADEP, Nunca Más (1984).

Corach, D. (2010), Mass disaster victim identification assisted by dna typing, in ‘Molecular Diagnostics’, Elsevier, pp. 407–415.

Corach, D., Sala, A., Penacino, G. and Sotelo, A. (1995), ‘Mass disasters: rapid molecular screening of human remains by means of short tandem repeats typing’, *Electrophoresis* **16**(1), 1617–1623.

Cordner, S. and Tidball-Binz, M. (2017), ‘Humanitarian forensic action—its origins and future’, *Forensic science international* **279**, 65–71.

Cross, R. (2008), ‘Forensic physics 101: Falls from a height’, *American Journal of Physics* **76**(9), 833–837.

Crow, J. F., Berger, M., Diamond, S., Kaye, D., Kazazian, H., Motulsky, A., Nagylaki, T., Nei, M., Sensabaugh, G., Siegmund, D. et al. (1996), ‘The evaluation of forensic dna evidence’, *National Re-899 search Council* **900**.

Cunha, E., Baccino, E., Martrille, L., Ramsthaler, F., Prieto, J., Schuliar, Y., Lynnerup, N. and Cattaneo, C. (2009), ‘The problem of aging human remains and living individuals: a review’, *Forensic science international* **193**(1-3), 1–13.

Darwiche, A. (2009), *Modeling and reasoning with Bayesian networks*, Cambridge university press.

Doretti, M., Osorno, C. and Daniell, R. (2017), ‘The border project: towards a regional forensic mechanism for the identification of missing migrants’, *Fatal Journeys* 3(Part I), 99–117.

Edwards, A. (2008), ‘Gh hardy (1908) and hardy–weinberg equilibrium’, *Genetics* 179(3), 1143–1150.

Edwards, A., Civitello, A., Hammond, H. A. and Caskey, C. T. (1991), ‘Dna typing and genetic mapping with trimeric and tetrameric tandem repeats.’, *American journal of human genetics* 49(4), 746.

Egeland, T., Kling, D. and Mostad, P. (2015), *Relationship inference with familias and R: statistical methods in forensic genetics*, Academic Press.

Egeland, T., Kling, D. and Mostad, P. (2016), Chapter 8 - making decisions, in T. Egeland, D. Kling and P. Mostad, eds, ‘Relationship Inference with Familias and R’, Academic Press, San Diego, pp. 203–228.

URL: <https://www.sciencedirect.com/science/article/pii/B9780128024027000084>

Egeland, T., Mostad, P. F. and Olaisen, B. (1997), ‘A computerised method for calculating the probability of pedigrees from genetic data’, *Science and Justice* 37(4), 269–274.

Egeland, T., Pinto, N. and Vigeland, M. D. (2014), ‘A general approach to power calculation for relationship testing’, *Forensic Science International: Genetics* 9, 186–190.

Elston, R. C. and Stewart, J. (1971), ‘A general model for the genetic analysis of pedigree data’, *Human heredity* 21(6), 523–542.

Evett, I. W., Jackson, G., Lambert, J. and McCrossan, S. (2000), ‘The impact of the principles of evidence interpretation on the structure and content of statements.’, *Science and justice: journal of the Forensic Science Society* 40(4), 233–239.

Evett, I. W. and Weir, B. S. (1998), *Interpreting DNA evidence: statistical genetics for forensic scientists*, Vol. 244, Sinauer Associates Sunderland, MA.

Falconer, D. S. and Mackay, T. F. (1983), *Quantitative genetics*, Longman London, UK.

Fenton, N., Neil, M. and Berger, D. (2016), ‘Bayes and the law’, *Annual Review of Statistics and Its Application* 3, 51.

Fishelson, M. and Geiger, D. (2002), ‘Exact genetic linkage computations for general pedigrees’, *Bioinformatics* 18(suppl_1), S189–S198.

Flores, M. J., Nicholson, A. E., Brunskill, A., Korb, K. B. and Mascaro, S. (2011), ‘Incorporating expert knowledge when learning bayesian network structure: a medical case study’, *Artificial intelligence in medicine* 53(3), 181–204.

- Fondebrider, L. (2016), The application of forensic anthropology to the investigation of cases of political violence: perspectives from south america, in ‘Handbook of Forensic Anthropology and Archaeology’, Routledge, pp. 107–116.
- Frey, B. A. (2019), Using the minnesota protocol to investigate disappearance cases, in ‘Disappearances in the Post-Transition Era in Latin America’, British Academy, pp. 225–233.
- Fyfe, N. R., Stevenson, O. and Woolnough, P. (2015), ‘Missing persons: the processes and challenges of police investigation’, *Policing and society* **25**(4), 409–425.
- Galas, D. J., Kunert-Graf, J., Uechi, L. and Sakhanenko, N. A. (2021), ‘Toward an information theory of quantitative genetics’, *Journal of Computational Biology* **28**(6), 527–559.
- García-Magariños, M., Egeland, T., López-de Ullibarri, I., Hjort, N. L. and Salas, A. (2015), ‘A parametric approach to kinship hypothesis testing using identity-by-descent parameters’, *Statistical applications in genetics and molecular biology* **14**(5), 465–479.
- Ge, J., Budowle, B. and Chakraborty, R. (2011), ‘Choosing relatives for dna identification of missing persons’, *Journal of forensic sciences* **56**, S23–S28.
- Gibbs, J. W. (1902), *Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics*, C. Scribner’s sons.
- Gill, P. and Buckleton, J. (2004), Biological basis for dna evidence, in ‘Forensic DNA Evidence Interpretation’, CRC Press, pp. 15–40.
- Gill, P., Jeffreys, A. J. and Werrett, D. J. (1985), ‘Forensic application of dna ‘fingerprints’’, *Nature* **318**(6046), 577–579.
- Gittelson, S., Biedermann, A., Bozza, S. and Taroni, F. (2012), ‘The database search problem: A question of rational decision making’, *Forensic science international* **222**(1-3), 186–199.
- Gjertson, D. W., Brenner, C. H., Baur, M. P., Carracedo, A., Guidet, F., Luque, J. A., Lessig, R., Mayr, W. R., Pascali, V. L., Prinz, M. et al. (2007), ‘Isfg: recommendations on biostatistics in paternity testing’, *Forensic Science International: Genetics* **1**(3-4), 223–231.
- Gorini, U. (2006), *La rebelión de las Madres: historia de las Madres de Plaza de Mayo*, Vol. 1, Grupo Ed. Norma.
- Gorini, U. (2017), *La rebelión de las Madres*.
- Hackman, L. (2016), Forensic anthropology and missing persons investigations, in ‘Handbook of missing persons’, Springer, pp. 415–425.
- Haimes, E. (2006), ‘Social and ethical issues in the use of familial searching in forensic investigations: insights from family and kinship studies’, *The Journal of Law, Medicine and Ethics* **34**(2), 263–276.
- Hamilton, M. B. (2021), *Population genetics*, John Wiley and Sons.

- Hazel, J. W., Clayton, E. W., Malin, B. A. and Slobogin, C. (2018), ‘Is it time for a universal genetic forensic database?’, *Science* **362**(6417), 898–900.
- Hedell, R., Hedman, J. and Mostad, P. (2018), ‘Determining the optimal forensic dna analysis procedure following investigation of sample quality’, *International Journal of Legal Medicine* **132**, 955–966.
- Henderson, M., Henderson, P. and Kiernan, C. (2000), *Missing persons: incidence, issues and impacts*, Australian Institute of Criminology Canberra, ACT.
- Jaynes, E. T. (1957a), ‘Information theory and statistical mechanics’, *Physical review* **106**(4), 620.
- Jaynes, E. T. (1957b), ‘Information theory and statistical mechanics’, *Physical review* **106**(4), 620.
- Jaynes, E. T. (2003), *Probability theory: The logic of science*, Cambridge university press.
- Jeffreys, A. J., Wilson, V. and Thein, S. L. (1985a), ‘Hypervariable ‘minisatellite’regions in human dna’, *Nature* **314**(6006), 67–73.
- Jeffreys, A. J., Wilson, V. and Thein, S. L. (1985b), ‘Individual-specific ‘fingerprints’ of human dna’, *Nature* **316**(6023), 76–79.
- Junior, R. C. S., Bezerra, L. S., Matte, C. H., Sales, S. L., Oliveira, E. C., Beltrami, L. S., Morais, B. D., Altmann, V., Mallmann, P. B., Figueiredo, E. T. et al. (2022), ‘Dna databases as a tool to improve the search for missing persons in brazil’, *Forensic Science International: Genetics Supplement Series* **8**, 167–169.
- Kaye, D. H. (2001), ‘Two fallacies about dna data banks for law enforcement’, *Brook. L. Rev.* **67**, 179.
- King, M. C. (1991), ‘An application of dna sequencing to a human rights problem’, *Mol Genet Med* **1**, 117–31.
- Kling, D., Egeland, T., Piñero, M. H. and Vigeland, M. D. (2017), ‘Evaluating the statistical power of dna-based identification, exemplified by ‘the missing grandchildren of argentina’’, *Forensic Science International: Genetics* **31**, 57–66.
- Kling, D., Tillmar, A. O. and Egeland, T. (2014), ‘Familias 3–extensions and new functionality’, *Forensic Science International: Genetics* **13**, 121–127.
- Kruijver, M., Meester, R. and Slooten, K. (2014), ‘Optimal strategies for familial searching’, *Forensic Science International: Genetics* **13**, 90–103.
- Kruschke, J. K. (2010), ‘Bayesian data analysis’, *Wiley Interdisciplinary Reviews: Cognitive Science* **1**(5), 658–676.
- Lander, E. S. and Green, P. (1987), ‘Construction of multilocus genetic linkage maps in humans.’, *Proceedings of the National Academy of Sciences* **84**(8), 2363–2367.

- Laurent, F.-X., Fischer, A., Oldt, R. F., Kanthaswamy, S., Buckleton, J. S. and Hitchin, S. (2022), ‘Streamlining the decision-making process for international dna kinship matching using world-wide allele frequencies and tailored cutoff log10lr thresholds’, *Forensic Science International: Genetics* **57**, 102634.
- Li, L., Ye, Y., Song, F., Wang, Z. and Hou, Y. (2019), ‘Genetic structure and forensic parameters of 30 indels for human identification purposes in 10 tibetan populations of china’, *Forensic Science International: Genetics* **40**, e219–e227.
- Machado, H., Granja, R., Machado, H. and Granja, R. (2020), ‘Dna technologies in criminal investigation and courts’, *Forensic Genetics in the Governance of Crime* pp. 45–56.
- MacKay, D. J., Mac Kay, D. J. et al. (2003), *Information theory, inference and learning algorithms*, Cambridge university press.
- Maguire, C. N., McCallum, L. A., Storey, C. and Whitaker, J. P. (2014), ‘Familial searching: A specialist forensic dna profiling service utilising the national dna database® to identify unknown offenders via their relatives—the uk experience’, *Forensic Science International: Genetics* **8**(1), 1–9.
- Marjanović, D., Hadžić Metjahi, N., Čakar, J., Džijan, S., Škaro, V., Projic, P., Madžar, T., Rod, E. and Primorac, D. (2015), ‘Identification of human remains from the second world war mass graves uncovered in bosnia and herzegovina’, *Croatian Medical Journal* **56**(3), 257–262.
- Marsico, F. and Caridi, I. (2023), ‘Incorporating non-genetic evidence in large scale missing person searches: A general approach beyond filtering’, *Forensic science international: genetics (Aceptado)*.
- Marsico, F. L., Vigeland, M. D., Egeland, T. and Piñero, M. H. (2021), ‘Making decisions in missing person identification cases with low statistical power’, *Forensic science international: genetics* **54**, 102519.
- Martin, P. D., Schmitter, H. and Schneider, P. M. (2001), ‘A brief history of the formation of dna databases in forensic science within europe’, *Forensic science international* **119**(2), 225–231.
- McElreath, R. (2020), *Statistical rethinking: A Bayesian course with examples in R and Stan*, Chapman and Hall/CRC.
- McKnight, B. E. et al. (1983), ‘The washing away of wrongs: forensic medicine in thirteenth-century china’, *VRÜ Verfassung und Recht in Übersee* **17**(1), 114–115.
- McVean Gil A., R. A. D. H. . K. C. . L. S. . L. L. . M. D. . R. J. . W. M. ., of MIT, B. I., 3, H. L. P. I. E. S. . A. D. M. . . G. C.-C. S. B. . G. N., 12, E. B. I. F. P. I. P. C. L. . L. R. . S. R. E. . Z.-B. X., 8, I. B. P. I. D. R. . G. R. . H. S. . J. T. . K. Z., of Health Sherry (Principal Investigator) Stephen T., U. N. I., of Oxford McVean (Principal Investigator) Gil A. 2, U. et al. (2012), ‘An integrated map of genetic variation from 1,092 human genomes’, *Nature* **491**(7422), 56–65.

- Myers, S. P., Timken, M. D., Piucci, M. L., Sims, G. A., Greenwald, M. A., Weigand, J. J., Konzak, K. C. and Buoncristiani, M. R. (2011), ‘Searching for first-degree familial relationships in California’s offender DNA database: Validation of a likelihood ratio-based approach’, *Forensic Science International: Genetics* **5**(5), 493–500.
- Nagarajan, R., Scutari, M. and Lèbre, S. (2013), ‘Bayesian networks in r’, *Springer* **122**, 125–127.
- Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A. et al. (2022), ‘The complete sequence of a human genome’, *Science* **376**(6588), 44–53.
- Parsons, T. J., Huel, R. M., Bajunović, Z. and Rizvić, A. (2019), ‘Large scale DNA identification: The ICMP experience’, *Forensic Science International: Genetics* **38**, 236–244.
- Penchaszadeh, V. B. (1997), ‘Genetic identification of children of the disappeared in Argentina’, *J Am Med Womens Assoc* **52**(1), 16–21.
- Pinto, N., Simões, R., Amorim, A. and Conde-Sousa, E. (2019), ‘Optimizing the information increase through the addition of relatives and genetic markers in identification and kinship cases’, *Forensic Science International: Genetics* **40**, 210–218.
- Prieto, L., Ruiz, Y., Hernandis, E. and Carracedo, Á. (2022), ‘DNA test evaluation in large-scale identification cases of missing persons’, *Spanish Journal of Legal Medicine* .
- Prinz, M., Carracedo, A., Mayr, W., Morling, N., Parsons, T., Sajantila, A., Scheithauer, R., Schmitter, H. and Schneider, P. M. (2007), ‘DNA commission of the international society for forensic genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI)’, *Forensic Science International: Genetics* **1**(1), 3–12.
- Puerto, M. S., Abboud, D., Baraybar, J. P., Carracedo, A., Fonseca, S., Goodwin, W., Guyomarc’h, P., Jimenez, A., Krenzer, U., Mendez, M. D. M. et al. (2021), ‘The search process: Integrating the investigation and identification of missing and unidentified persons’, *Forensic Science International: Synergy* **3**, 100154.
- Puerto, M. S. and Tuller, H. (2017), ‘Large-scale forensic investigations into the missing: Challenges and considerations’, *Forensic science international* **279**, 219–228.
- Ramos, D., Gonzalez-Rodriguez, J., Zadora, G. and Aitken, C. (2013), ‘Information-theoretical assessment of the performance of likelihood ratio computation methods’, *Journal of forensic sciences* **58**(6), 1503–1518.
- Ramos, D., Meuwly, D., Haraksim, R. and Berger, C. E. (2020), Validation of forensic automatic likelihood ratio methods, in ‘Handbook of forensic statistics’, Chapman and Hall/CRC, pp. 143–162.
- Reineke104, R. (2016), ‘Missing persons and unidentified remains at the United States–Mexico border’, *Fatal Journeys* **75**.

- Rota, M. and Antolini, L. (2014), ‘Finding the optimal cut-point for gaussian and gamma distributed biomarkers’, *Computational Statistics and Data Analysis* **69**, 1–14.
- Ruitberg, C. M., Reeder, D. J. and Butler, J. M. (2001), ‘Strbase: a short tandem repeat dna database for the human identity testing community’, *Nucleic acids research* **29**(1), 320–322.
- Schmitt, A., Cunha, E. and Pinheiro, J. (2006), *Forensic anthropology and medicine*, Springer.
- Schmitt, A., Saliba-Serre, B., Tremblay, M. and Martrille, L. (2010), ‘An evaluation of statistical methods for the determination of age of death using dental root translucency and periodontosis’, *Journal of forensic sciences* **55**(3), 590–596.
- Shannon, C. E. (1948), ‘A mathematical theory of communication’, *The Bell system technical journal* **27**(3), 379–423.
- Siegel, J. A. and Saukko, P. J. (2012), *Encyclopedia of forensic sciences*, Academic Press.
- Slooten, K. (2020), ‘Likelihood ratio distributions and the (ir) relevance of error rates’, *Forensic Science International: Genetics* **44**, 102173.
- Slooten, K. and Meester, R. (2014), ‘Probabilistic strategies for familial dna searching’, *Journal of the Royal Statistical Society: Series C: Applied Statistics* pp. 361–384.
- Stover, E. and Shigekane, R. (2002), ‘The missing in the aftermath of war: When do the needs of victims’ families and international war crimes tribunals clash?’, *International Review of the Red Cross* **84**(848), 845–866.
- Taroni, F., Biedermann, A. and Bozza, S. (2016), ‘Statistical hypothesis testing and common misinterpretations: Should we abandon p-value in forensic science applications?’, *Forensic science international* **259**, e32–e36.
- Tillmar, A. O. and Mostad, P. (2014), ‘Choosing supplementary markers in forensic casework’, *Forensic Science International: Genetics* **13**, 128–133.
- Uda, S. (2020), ‘Application of information theory in systems biology’, *Biophysical reviews* **12**(2), 377–384.
- Ulanowicz, R. E. (2001), ‘Information theory in ecology’, *Computers and chemistry* **25**(4), 393–399.
- van Dongen, C., Slooten, K., Slagter, M., Burgers, W. and Wiegerinck, W. (2011), ‘Bonaparte: Application of new software for missing persons program’, *Forensic Science International: Genetics Supplement Series* **3**(1), e119–e120.
- Vigeland, M. D. (2021), *Pedigree analysis in R*, Academic Press.
- Vigeland, M. D. and Egeland, T. (2021), ‘Joint dna-based disaster victim identification’, *Scientific Reports* **11**(1), 13661.

- Vigeland, M. D., Marsico, F. L., Pinero, M. H. and Egeland, T. (2020), ‘Prioritising family members for genotyping in missing person cases: a general approach combining the statistical power of exclusion and inclusion’, *Forensic Science International: Genetics* **49**, 102376.
- Vullo, C. M., Catelli, L., Rodriguez, A. A. I., Papaioannou, A., Merino, J. C. Á., Lopez-Parra, A., Gaviria, A., Baeza-Richer, C., Romanini, C., González-Moya, E. et al. (2021), ‘Second ghep-isfg exercise for dvi:“dna-led” victims’ identification in a simulated air crash’, *Forensic Science International: Genetics* **53**, 102527.
- Vullo, C. M., Romero, M., Catelli, L., Šakić, M., Saragoni, V. G., Pleguezuelos, M. J. J., Romanini, C., Porto, M. J. A., Prieto, J. P., Castro, A. B. et al. (2016), ‘Ghep-isfg collaborative simulated exercise for dvi/mpi: lessons learned about large-scale profile database comparisons’, *Forensic Science International: Genetics* **21**, 45–53.
- Watson, J. D., Crick, F. H. et al. (1953), ‘A structure for deoxyribose nucleic acid’.
- Weber, J. L. and May, P. E. (1989), ‘Abundant class of human dna polymorphisms which can be typed using the polymerase chain reaction.’, *American journal of human genetics* **44**(3), 388.
- Wright, S. (1921), ‘Correlation and causation’.
- Yang, J. (2018), ‘Information theoretic approaches in economics’, *Journal of Economic Surveys* **32**(3), 940–960.
- Zitzewitz, E. (2012), ‘Forensic economics’, *Journal of Economic Literature* **50**(3), 731–69.
- Zupanič Pajnič, I., Gornjak Pogorelc, B. and Balažic, J. (2010), ‘Molecular genetic identification of skeletal remains from the second world war konfin i mass grave in slovenia’, *International Journal of Legal Medicine* **124**, 307–317.

LISTA DE PUBLICACIONES

1. Marsico, F., Cardi, I. (2023). Incorporating Non-Genetic Evidence in Large Scale Missing Person Searches: A General Approach Beyond Filtering. *Forensic science international: genetics*. Aceptado.
2. Chernomoretz, A., Marsico, F., Iserte, J., Piñero, M. H., Escobar, M. S., Belparda, M., Sibilla, G. (2022). Bayesian networks for DNA-based kinship analysis: Functionality and validation of the GENis missing person identification module. *Forensic Science International: Genetics Supplement Series*, 8, 131-132.
3. Marsico, F. L., Vigeland, M. D., Egeland, T., Piñero, M. H. (2021). Making decisions in missing person identification cases with low statistical power. *Forensic science international: genetics*, 54, 102519.
4. Vigeland, M. D., Marsico, F. L., Pinero, M. H., Egeland, T. (2020). Prioritising family members for genotyping in missing person cases: a general approach combining the statistical power of exclusion and inclusion. *Forensic Science International: Genetics*, 49, 102376.
5. Marsico, F., Sibila, G., Chernomoretz, A. (2023). The Missing Person problem through the lens of information theory. En desarrollo.