QÜESTIIÓ, vol. 24, 3, p. 555-576, 2000

CÁLCULO DE PROBABILIDADES EN EL ANÁLISIS DE PERFILES GENÉTICOS COMPATIBLES

MIGUEL SÁNCHEZ GARCÍA*
PEDRO CUESTA ALVARO**
Universidad Complutense de Madrid

Un hecho delictivo ha sido cometido por una o más personas. Para descubrir el número de presuntos delincuentes, se determina el perfil genético de la evidencia forense encontrada en el lugar del delito, mediante el estudio de marcadores de tipo STR. En el presente artículo se desarrollan diversos algoritmos. En el primero se considera la probabilidad de que n personas elegidas aleatoriamente de la población de referencia contengan un perfil genético compatible, en el segundo se calcula la evidencia del material genético y en el tercero se calcula el mínimo número de presuntos culpables seleccionado de un conjunto compatible. Finalmente, se analizan las evidencias en diversas situaciones particulares.

Calculation of evidence in the analysis of genetic compatible profiles

Palabras clave: STR, alelo, principio de inclusión-exclusión, multigrafo, problema de recubrimiento

Clasificación AMS (MSC 2000): 60C05,92D20

* Departamento de Estadística. Facultad de Medicina. Universidad Complutense de Madrid. 28040 Madrid. Tel.: 913941666. E-mail: mesegar@eucmos.sim.ucm.es
** Centro de Proceso de Datos. Servicios Informáticos. Avda. Complutense s/n. 28040 Madrid. Tel.: 913944755. E-mail: pedro@sim.ucm.es

– Recibido en marzo de 1998.
– Aceptado en octubre de 2000.

1. INTRODUCCIÓN

Se supone que se ha cometido un hecho delictivo sobre una o más personas, que se denotarán con el nombre genérico de *víctima*. Para extraer más información sobre el suceso, se analiza el material genético perteneciente a uno o más STR (iniciales de Short Tandem Repeat, según se expone en Foreman y otros (1997)), encontrado en el lugar donde se cometió el delito. Para cada STR, una persona aporta dos bandas alélicas A_i, A_j, que pueden ser iguales o distintas.

Suponiendo que se analiza un único STR, sea $\mathbf{B} = \{A_1, A_2, \ldots, A_q\}$ el conjunto de bandas alélicas diferentes que no pertenecen a la víctima y $\mathbf{V} = \{A_{q+1}, \ldots, A_k\}$ el conjunto de bandas alélicas pertenecientes a la víctima. $\mathbf{B} \cup \mathbf{V}$ es el conjunto de las k bandas alélicas diferentes encontradas en el lugar del suceso.

Es claro que cuando una persona tiene una banda alélica $A_s \notin \mathbf{B} \cup \mathbf{V}$ debe descartarse como presunto culpable. Además si $q = 2s$, el número de presuntos culpables debe ser al menos s; mientras que si $q = 2s + 1$, dicho número es al menos $s + 1$, debido a que cada uno aporta como máximo dos bandas alélicas.

En el artículo se supone que se cumple la ley de equilibrio de Hardy-Weinberg (que se explica en el apartado 3 y en Louis y Dempster (1987)) y que son conocidas las probabilidades de las bandas alélicas. Con estas hipótesis se proporcionan métodos de cálculo para resolver los tres problemas siguientes:

1) Calcular la probabilidad de que n personas seleccionadas aleatoriamente de una población, supuestamente infinita, puedan ser presuntas culpables por presentar compatibilidad con el material genético encontrado.

2) Calcular la probabilidad de que un conjunto de cualquier número de personas sea compatible. El inverso de esta probabilidad se llama *evidencia*.

3) Hallar un subconjunto compatible, de cardinal mínimo, de un conjunto compatible de n personas.

Respecto al cálculo de las probabilidades de compatibilidad, se realizan los oportunos comentarios para el caso de que no se cumpla la ley de Hardy-Weinberg, se analicen varios STR o se desconozcan las probabilidades alélicas y deban ser estimadas por las correspondientes frecuencias.

El artículo se estructura en seis apartados. En el segundo se formulan los tres problemas, que se resuelven en los apartados tercero, cuarto y quinto. En el sexto, se calculan y analizan las probabilidades relativas al cálculo de evidencias en diversas situaciones. Finalmente, se comentan las potenciales aplicaciones forenses de estos cálculos.

2. PLANTEAMIENTO DE LOS PROBLEMAS

Según lo descrito en la introducción, n personas seleccionadas aleatoriamente de la población de referencia aportan $2n$ bandas alélicas

$$\mathbf{L}_n = \{D_1, D_2, \ldots, D_{2n-1}, D_{2n}\}$$

$\mathbf{L}_n$ puede tener elementos repetidos, por lo que se utiliza la denominación de lista en lugar de la de conjunto.

Para que $\mathbf{L}_n$ sea compatible con la información genética recogida en $\mathbf{B}$ y $\mathbf{V}$, se deben cumplir las dos condiciones siguientes

c1) Todo D_j de la lista $\mathbf{L}_n$, es tal que $D_j \in \mathbf{B} \cup \mathbf{V}$

c2) Para todo $A_i \in \mathbf{B} = \{A_1, \ldots, A_q\}$, existe un D_j de $\mathbf{L}_n$ tal que $A_i = D_j$.

Si $\mathbf{L}_n^*$ es el conjunto que se obtiene cuando se quitan de $\mathbf{L}_n$ los elementos repetidos, las condiciones previas de compatibilidad equivalen a $\mathbf{B} \subset \mathbf{L}_n^* \subset \mathbf{B} \cup \mathbf{V}$.

Problema P1]. Sea $\mathbf{D}_n$ el suceso de que n personas, seleccionadas aleatoriamente de la población de referencia, aporten un material genético compatible con $\mathbf{B}$ y $\mathbf{V}$. Calcular $P(\mathbf{D}_n)$.

En el apartado tercero se presentan los algoritmos para el cálculo de la probabilidad de $\mathbf{D}_n$, para cada n. No obstante, muchos problemas genéticos sobre Medicina Forense consisten en calcular la evidencia del material genético encontrado en el lugar del suceso. Cuando sólo se dispone de la información aportada por $\mathbf{B}$ y $\mathbf{V}$, la *evidencia* es el inverso de la probabilidad de la unión de los $\mathbf{D}_n$; esto es:

$$1 \Big/ P\left(\bigcup_{n=1}^{\infty} \mathbf{D}_n\right)$$

lo que lleva a formular el segundo problema.

Problema P2]. Calcular la probabilidad del suceso $\bigcup_{n=1}^{\infty} \mathbf{D}_n$

Puesto que los sucesos $\mathbf{D}_n$ no son disjuntos, el cálculo, que se desarrolla en el apartado cuarto, es

$$P\left(\bigcup_{n=1}^{\infty} \mathbf{D}_n\right) = P(\mathbf{D}_1) + \sum_{n=2}^{\infty} P\left(\mathbf{D}_n \cap \left(\bigcup_{j=1}^{n-1} \mathbf{D}_j\right)^c\right)$$

El hecho de que n personas sean compatibles con el material genético encontrado en el lugar del delito, no quiere decir que no haya un subconjunto de estas personas que también sea compatible con el mismo suceso. En esta observación está la génesis del tercer problema, que se abordará en el quinto apartado de este trabajo, y que se enuncia en los siguientes términos.

Problema P3]. Sea un conjunto de n personas $\{I_1, I_2, \ldots, I_n\}$ elegidas aleatoriamente de la población de referencia, compatible con **B** y **V**. Hallar un subconjunto $\mathbf{F} \subset \{I_1, I2, \ldots, I_n\}$, tal que **F** también sea compatible y tenga cardinal mínimo.

Es obvio que $P3$ siempre tiene solución, aunque puede no ser única.

3. ALGORITMO PARA CALCULAR LAS PROBABILIDADES DE LOS SUCESOS $\mathbf{D_n}$

Sobre cada persona de la población de referencia Ω, se observan dos bandas alélicas $\{A_i, A_j\}$, que forman su sistema STR. Por tanto, sobre Ω se pueden definir dos variables ξ_1 y ξ_2 tales que $\xi_1(\omega) = A_i$ y $\xi_2(\omega) = A_j$. Si $G(\omega) = (A_i, A_j)$, la ley de Hardy-Weinberg dice que las variables ξ_1 y ξ_2 son probabilísticamente independientes, esto es

$$P[\omega | G(\omega) = (A_i, A_j)] = P[\omega | \xi_1(\omega) = A_i] \cdot P[\omega | \xi_2(\omega) = A_j]$$

Esta ley facilita el cálculo de probabilidades, pues sólo es necesario conocer las probabilidades de los alelos, cuyo número es menor que el de combinaciones genotípicas del STR.

Se supone que las bandas alélicas del STR analizado son $\{AL_1, AL_2, \ldots, AL_m\}$ y que $q_j = P(AL_j)$ son las correspondientes probabilidades.

Sea

$$\begin{aligned} \mathbf{B} &= \{A_1, A_2, \ldots, A_q\} = \{AL_{i_1}, AL_{i_2}, \ldots, AL_{i_q}\} \quad \text{y} \\ \mathbf{V} &= \{A_{q+1}, \ldots, A_k\} = \{AL_{i_{q+1}}, \ldots, AL_{i_k}\}. \end{aligned}$$

Se denota por $p_j = P(A_j) = P(AL_{i_j}) = q_{i_j}$ y $p_{\mathbf{v}} = P(\mathbf{V}) = \sum_{s=q+1}^{k} q_{i_s}$

Llamando $\mathbf{T} = \mathbf{B} \cup \mathbf{V}$ y $\mathbf{S}_j = \mathbf{T} - \{A_j\}$, para $j = 1, 2, \ldots, q$; como consecuencia del principio de inclusión-exclusión, Feller (1973 Cap. IV) se verifica la siguiente igualdad:

(1)

$$\begin{aligned} P(\mathbf{D}_n) &= P\left(\mathbf{T}^{2n} - \bigcup_{j=1}^{q} \mathbf{S}_j^{2n}\right) = P\left(\mathbf{T}^{2n}\right) - P\left(\bigcup_{j=1}^{q} \mathbf{S}_j^{2n}\right) = P(\mathbf{T}^{2n}) - \sum_{j=1}^{q} P(\mathbf{S}_j^{2n}) + \\ &+ \sum_{1 \leqslant i < j \leqslant q} P\left(\mathbf{S}_i^{2n} \cap \mathbf{S}_j^{2n}\right) - \sum_{1 \leqslant i < j < r \leqslant q} P\left(\mathbf{S}_i^{2n} \cap \mathbf{S}_j^{2n} \cap \mathbf{S}_r^{2n}\right) + \cdots + \\ &+ (-1)^{q-1} \sum_{j=1}^{q} P\left(\bigcap_{i \neq j}^{q} \mathbf{S}_i^{2n}\right) + (-1)^q P\left(\mathbf{S}_1^{2n} \cap \cdots \cap \mathbf{S}_j^{2n} \cap \cdots \cap \mathbf{S}_q^{2n}\right) \end{aligned}$$

Para facilitar el cálculo de probabilidades dadas por la fórmula (1), se tiene en cuenta la independencia de las bandas alélicas, tanto para una misma persona como para personas distintas:

i) $P\left(\mathbf{T}^{2n}\right) = [P(\mathbf{T})]^{2n} = [p_1 + p_2 + \cdots + p_q + p_{\mathbf{v}}]^{2n}$

ii) $P\left(\mathbf{S}_{i_1}^{2n} \cap \mathbf{S}_{i_2}^{2n} \cap \cdots \cap \mathbf{S}_{i_r}^{2n}\right) = [P\left(\mathbf{S}_{i_1} \cap \mathbf{S}_{i_2} \cap \cdots \cap \mathbf{S}_{i_r}\right)]^{2n}$

Se realiza el cálculo de la probabilidad de $\mathbf{D}_n$ en dos etapas: En la primera se calculan las probabilidades de los sucesos:

$$\mathbf{S}_{i_1,i_2,\ldots,i_r} = \mathbf{S}_{i_1} \cap \mathbf{S}_{i_2} \cap \cdots \cap \mathbf{S}_{i_r} = \mathbf{T} - \{A_{i_1}, A_{i_2}, \ldots, A_{i_r}\}$$

y en la segunda se calcula, iterativamente, la probabilidad de $\mathbf{D}_n$.

Cualquier suceso $\mathbf{S}_{i_1,i_2,\ldots,i_r}$ se identifica biunívocamente con el número binario de q cifras que tiene unos en las posiciones $i_1, i_2, \ldots, i_r$ y ceros en las restantes posiciones. Esta identificación facilita el diseño del algoritmo siguiente, que calcula las probabilidades de los sucesos $\mathbf{S}_{i_1,i_2,\ldots,i_r}$.

3.1. Algoritmo A1 de aplicación a las bandas alélicas B y V

Dados $p(1) = P(A_1), \ldots, p(q) = P(A_q)$; $p(q+1) = P(\mathbf{V}) = P\{A_{q+1}, \ldots, A_k\}$.
Inicializar $\mathbf{b}(1:q) = 0$;
Para $i = 1$ hasta $2^q - 1$
    Para $j = 1$ hasta q
        Si $b(j) = 0$, poner $b(j) = 1$ y Salir j
        En otro caso $b(j) = 0$
    Siguiente j
    $s(i) = s(0) - \sum_{j=1}^{q} b(j) \cdot p(j)$
Siguiente i
En $\mathbf{s}(0 : 2^q - 1)$ se almacena las probabilidades de los sucesos $\mathbf{S}_{i_1,i_2,\ldots,i_r}$.

3.2. Algoritmo A2 de aplicación a las bandas alélicas B y V

Se calculan las probabilidades de las sucesivas potencias cartesianas de los sucesos $\mathbf{S}_{i_1,i_2,\ldots,i_r}$, para evaluar las probabilidades de los sucesos $\mathbf{D}_n$ por la fórmula (1); para valores de n desde 1 a *nbio*.

Dados q; *nbio*; $\mathbf{s}(0:2^q-1)$
Para $i=0$ hasta 2^q-1; poner $sc(1-i)=s(i)\cdot s(i)$; Siguiente i
Poner $pD(1)=sc(1,0)$

Para $n=2$ hasta *nbio*
    Para $i=0$ hasta 2^q-1; poner $sc(n,i)=sc(n(1,i)\cdot sc(1,i)$; Siguiente i
    $pD(n)=sc(n,0)$
Siguiente n

Para $n=1$ hasta *nbio*
    $\mathbf{b}(1:q)=0$
    Para $i=1$ hasta 2^q-1
        Para $j=1$ hasta q
            Si $b(j)=0$, poner $b(j)=1$ y Salir j
            En otro caso $b(j)=0$
        Siguiente j
        $pD(n)=pD(n)+(-1)^{\text{suma}(\mathbf{b})}\cdot sc(n,i)$
    Siguiente i
    Escribe $n, pD(n)$
Siguiente n

Nota:

El entero *nbio* representa el número máximo de presuntos delincuentes, hasta el que se calculan las probabilidades de los sucesos $\mathbf{D}_n$.

En $s(i)$ se almacena la probabilidad del suceso $\mathbf{S}_{i_1,i_2,\ldots,i_r}$ asociado con el número i representado en binario.

En $sc(n,i)$ se almacena $s(i)^{2n}$.

$\mathbf{b}(1:q)$ es una referencia a los elementos $b(j)$, $j=1,\ldots,q$. En $\mathbf{b}(1:q)$ se almacena iterativamente los desarrollos binarios de los números que van desde 0 hasta 2^q-1.

La expresión suma(**b**) es el número de unos en $\mathbf{b}(1:q)$.

Finalmente en $pD(n)$ se almacena la probabilidad del suceso $\mathbf{D}_n$.

OBSERVACIONES

a) Cuando no se puede admitir la ley de Hardy-Weinberg, se deben conocer las probabilidades conjuntas de las dos bandas alélicas (A_i, A_j). El cálculo de la probabilidad de $\mathbf{D}_n$ se realiza de forma idéntica al descrito, sustituyendo en la fórmula (1) el conjunto $\mathbf{T}$ por los pares de bandas alélicas pertenecientes ambas a $\mathbf{B} \cup \mathbf{V}$, cada suceso $\mathbf{S}_j$, $j \in \mathbf{B}$, por los pares de bandas alélicas que pertenecen a $\mathbf{B} \cup \mathbf{V} - j$, la potencia $2n$ por n y las probabilidades de las bandas alélicas por las de los pares de bandas.

b) En el caso de medir varios STR, si éstos son independientes se calcularía la probabilidad por la regla del producto, mientras que si son dependientes se utilizaría la regla del producto condicional.

c) Sea $\hat{p}_i$ la estimación por máxima verosimilitud de las probabilidades de las bandas alélicas, cuando éstas se estiman por las frecuencias observadas en una muestra de la población de referencia.

Al sustituir en el cálculo de $P(\mathbf{D}_n), p_i$ por $\hat{p}_i$ se obtiene el estimador $\hat{P}(\mathbf{D}_n)$.

Si (a_i, b_i) son intervalos de confianza de $\hat{p}_i$, al tomar muestras aleatoriamente tal que $(p_1^0, p_2^0, \ldots, p_k^0) \in \prod_{i=1}^{k} (a_i, b_i)$, se obtienen distintos valores de $\hat{P}(\mathbf{D}_n)$, con los que se pueden construir intervalos de confianza para $P(\mathbf{D}_n)$.

EJEMPLO

En la tabla 3.1 se muestran las probabilidades $P(\mathbf{D}_n), n = 1, \ldots, 10$; en diferentes supuestos. $P_{\mathbf{B}}$ es la suma de las probabilidades de los alelos no pertenecientes a la víctima; probabilidades que se muestran en las columnas $p_j, j = 1, \ldots, q$; para diferentes valores de q.

El valor m que maximiza $P(\mathbf{D}_n)$ es el número de personas con mayor probabilidad de genotipos compatibles. En el primer supuesto de la tabla es $m = 1$, en el segundo $m = 2$ y en el tercero $m = 3$.

4. ALGORITMO PARA EL CÁLCULO DE EVIDENCIAS

Puesto que los sucesos $\mathbf{D}_n$ no son incompatibles, no se puede aplicar la fórmula de sucesos disjuntos para calcular la probabilidad de $\mathbf{E} = \bigcup_{n=1}^{\infty} \mathbf{D}_n$. La descomposición en sucesos disjuntos se consigue por el procedimiento:

$$\mathbf{E} = \bigcup_{n=1}^{\infty} \mathbf{D}_n = \mathbf{D}_1 \cup \left[\bigcup_{n=2}^{\infty} \left(\mathbf{D}_n \cap \left(\bigcup_{j=1}^{n-1} \mathbf{D}_j \right)^c \right) \right]$$

Tabla 3.1. *Valores de $P(\mathbf{D}_n)$ en diversos supuestos*

							$P(\mathbf{D}_n)$									
$p_{\mathbf{V}}$	$p_{\mathbf{B}}$	q	p_1	p_2	p_3	p_4	1	2	3	4	5	6	7	8	9	10
0,10	0,35	2	0,205	0,145			0,0595	0,0289	0,0073	0,0016	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000
0,10	0,35	3	0,072	0,105	0,174		0,0000	0,0086	0,0038	0,0011	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000
0,10	0,35	4	0,075	0,062	0,092	0,122	0,0000	0,0012	0,0015	0,0006	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000
0,10	0,45	2	0,176	0,274			0,0965	0,0663	0,0245	0,0080	0,0025	0,0008	0,0002	0,0001	0,0000	0,0000
0,10	0,45	3	0,148	0,142	0,161		0,0000	0,0262	0,0161	0,0064	0,0022	0,0007	0,0002	0,0001	0,0000	0,0000
0,10	0,45	4	0,110	0,143	0,091	0,106	0,0000	0,0036	0,0060	0,0036	0,0015	0,0006	0,0002	0,0001	0,0000	0,0000
0,10	0,60	2	0,235	0,365			0,1716	0,1809	0,1061	0,0553	0,0278	0,0137	0,0068	0,0033	0,0016	0,0008
0,10	0,60	3	0,293	0,199	0,108		0,0000	0,0603	0,0586	0,0384	0,0219	0,0117	0,0061	0,0031	0,0015	0,0008
0,10	0,60	4	0,164	0,131	0,148	0,157	0,0000	0,0120	0,0301	0,0277	0,0185	0,0108	0,0058	0,0030	0,0015	0,0008
0,20	0,30	2	0,099	0,201			0,0397	0,0302	0,0108	0,0032	0,0009	0,0002	0,0001	0,0000	0,0000	0,0000
0,20	0,30	3	0,120	0,105	0,075		0,0000	0,0079	0,0052	0,0020	0,0006	0,0002	0,0001	0,0000	0,0000	0,0000
0,20	0,30	4	0,083	0,094	0,054	0,069	0,0000	0,0007	0,0013	0,0008	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000
0,20	0,40	2	0,185	0,215			0,0796	0,0796	0,0384	0,0154	0,0058	0,0021	0,0008	0,0003	0,0001	0,0000
0,20	0,40	3	0,136	0,108	0,157		0,0000	0,0220	0,0191	0,0102	0,0045	0,0018	0,0007	0,0003	0,0001	0,0000
0,20	0,40	4	0,070	0,087	0,134	0,109	0,0000	0,0021	0,0052	0,0043	0,0025	0,0012	0,0005	0,0002	0,0001	0,0000
0,20	0,60	1	0,600				0,6000	0,4080	0,2621	0,1678	0,1074	0,0687	0,0440	0,0281	0,0180	0,0115
0,20	0,60	2	0,375	0,225			0,1686	0,2691	0,2201	0,1547	0,1032	0,0674	0,0435	0,0280	0,0180	0,0115
0,20	0,60	3	0,274	0,120	0,206		0,0000	0,0812	0,1148	0,1040	0,0798	0,0567	0,0387	0,0258	0,0169	0,0110
0,20	0,60	4	0,098	0,177	0,155	0,170	0,0000	0,0110	0,0424	0,0572	0,0543	0,0439	0,0324	0,0228	0,0155	0,0104

De esta forma los sucesos $\mathbf{D}_1$ y $\mathbf{D}_n \cap \left(\bigcup_{j=1}^{n-1} \mathbf{D}_j\right)^c$, para $n \geqslant 2$, son disjuntos.

Por tanto, el cálculo de $P(\mathbf{E})$, cuyo inverso es la evidencia de $\mathbf{B}$ y $\mathbf{V}$, se reduce al cálculo iterativo de las probabilidades de los sucesos $\mathbf{D}_n \cap \left(\bigcup_{j=1}^{n-1} \mathbf{D}_j\right)^c$ y a su posterior suma.

Para el cálculo de estas probabilidades se debe tener en cuenta que la persona seleccionada en el lugar n-ésimo debe aportar las bandas alélicas de $\mathbf{B}$ que no pertenezcan a las primeras $n-1$ personas seleccionadas. Como una persona aporta como máximo dos bandas alélicas, en $\mathbf{L}_{n-1}$ tienen que estar todos los elementos de $\mathbf{B}$, salvo uno o dos como máximo. El uno o dos que faltan deben ser aportados por la n-ésima persona.

Se denota por $P(n,-A_i)$, $1 \leqslant i \leqslant q$ la probabilidad de que n personas, elegidas aleatoriamente de la población, sean compatibles con las bandas alélicas $\mathbf{B}-\{A_i\}$ y $\mathbf{V}$ y por $P(n,-\{A_i,A_j\})$, $1 \leqslant i < j \leqslant q$ la probabilidad de que sean compatibles con las bandas alélicas $\mathbf{B}-\{A_i,A_j\}$ y $\mathbf{V}$.

Las probabilidades de los sucesos $\mathbf{D}_n \cap \left(\bigcup_{j=1}^{n-1} \mathbf{D}_j\right)^c$ se calculan por la fórmula:

(2)

$$pI(n) = P\left(\mathbf{D}_n \cap \left(\bigcup_{j=1}^{n-1} \mathbf{D}_j\right)^c\right) = \sum_{i=1}^{q} P(n-1,-A_i)\left[2P(A_i)P(\mathbf{B}\cup\mathbf{V}) - P(A_i)P(A_i)\right] + \\ + \sum_{1\leqslant i<j\leqslant q} P\left(n-1,-\{A_i,A_j\}\right)\left[2P(A_i)P(A_j)\right]$$

Los cálculos de $P(n-1,-A_i)$ y $P(n-1,-\{A_i,A_j\})$ se realizan por los algoritmos A1 y A2 aplicados a las bandas alélicas $\mathbf{B}-\{A_i\},\mathbf{V}$ y $\mathbf{B}-\{A_i,A_j\},\mathbf{V}$, respectivamente.

Con esta terminología el algoritmo para calcular evidencias se diseña como sigue.

Algoritmo A3

Dado *nmax*, valor máximo de n para el que se calculan las probabilidades $pI(n)$.
Inicializar $\mathbf{pI}(1:nmax)=0$
Calcular $pI(1)=pD(1)$ aplicando A1 y A2

Para $n=2$ hasta *nmax*

Para $i=1$ hasta q
Calcular $P(n-1,-A_i)$ aplicando los algoritmos A1 y A2 para

$nbio = n-1$ y las bandas alélicas $\mathbf{B} - \{A_i\}$ y $\mathbf{V}$.

$pI(n) = pI(n) + P(n-1, -A_i) \cdot [2P(A_i)P(\mathbf{B}(\mathbf{V}) - P(A_i)P(A_i)]$

Siguiente i

Para $i = 1$ hasta $q-1$

Para $j = i+1$ hasta q

Calcular $P(n-1, -\{A_i, A_j\})$ aplicando los algoritmos A1 y A2 para $nbio = n-1$ y las bandas alélicas $\mathbf{B} - \{A_i, A_j\}$ y $\mathbf{V}$.

$pI(n) = pI(n) + P(n-1, -\{A_i, A_j\}) \cdot [2P(A_i)P(A_i)]$

Siguiente j

Siguiente i

Escribe $n, pI(n)$

Siguiente n

Nota:

Se pueden evaluar y almacenar, por los algoritmos A1 y A2, las probabilidades:

$P(n, -A_i), n = 1, 2, \ldots, nmax,$ fijando cada banda alélica A_i.

$P(n, -\{A_i A_j\}), n = 1, 2, \ldots, nmax,$ fijando cada par A_i, A_j, con $i < j$.

Y después evaluar $pI(n)$, para $n = 2, \ldots, nmax$ por la fórmula (2).

Con esta forma de proceder se necesita más memoria para guardar datos intermedios, aunque se gana rapidez computacional.

5. PLANTEAMIENTO Y RESOLUCIÓN DEL PROBLEMA P3

El problema que consiste en calcular el mínimo número de personas, tal que el material genético aportado por ellas sea compatible, se formula como sigue:

Problema P3]. Sea un conjunto de n personas $\{I_1, I_2, \ldots, I_n\}$ elegidas aleatoriamente de la población de referencia, cuyo material genético es compatible con el encontrado en el lugar del delito. Hallar un subconjunto $\mathbf{F} \subset \{I_1, I_2, \ldots, I_n\}$, tal que $\mathbf{F}$ también sea compatible y tenga cardinal mínimo.

El problema se resuelve hallando $\mathbf{F} \subset \{I_1, I_2, \ldots, I_n\}$ tal que $|\mathbf{F}|$ sea mínimo y se cumpla la condición de compatibilidad c2), ya que la c1) se cumple obviamente.

Para facilitar su resolución se plantea $P3$ como un problema de recubrimiento. Para ello se utilizan n variables binarias; de tal forma que si $x_j = 1$ entonces $I_j \in \mathbf{F}$, mientras que

si $x_j = 0$ entonces $I_j \notin \mathbf{F}$. La matriz $\mathbf{A}$ del problema de recubrimiento es de dimensión $q \times n$ y se construye por el siguiente procedimiento.

Para cada i y cada j, $1 \leqslant i \leqslant q$, $1 \leqslant j \leqslant n$, el elemento a_{ij} de $\mathbf{A}$ es:

$$\boldsymbol{a}_{ij} = \begin{cases} 1 & \text{si} \quad A_i \in \{D_{2j-1}, D_{2j}\} \\ 0 & \text{si} \quad A_i \notin \{D_{2j-1}, D_{2j}\} \end{cases}$$

siendo $A_i \in \mathbf{B}$ y $\{D_{2j-1}, D_{2j}\}$ las bandas alélicas de la persona I_j.

Con estos datos, el problema de recubrimiento equivalente a $P3$, se formula como:

PR1] $\quad \min x_1 + x_2 + \cdots + x_n \qquad \text{con } x_j \in \{0,1\}$

$\quad$ Sujeto a $a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \geqslant 1 \qquad i = 1, 2, \ldots, q$

Cada solución de $PR1$ es compatible con $\mathbf{B}$ y $\mathbf{V}$ y tiene cardinal mínimo, por lo que es solución de $P3$. Recíprocamente cada solución de $P3$ es una solución de $PR1$, de lo que se deduce que ambos problemas son equivalentes.

5.1. Planteamiento del problema de recubrimiento por un grafo

La información suministrada por $\mathbf{B}$, $\mathbf{V}$ y $\mathbf{L}_n = \{D_1, D_2, \ldots, D_{2n-1}, D_{2n}\}$, bandas alélicas aportadas por $\{I_1, I_2, \ldots, I_n\}$ posibilita la construcción del multigrafo $\mathbf{G}^* = (\mathbf{h}, \mathbf{a}^*)$, donde $\mathbf{h} = \{1, 2, \ldots, q+1\}$ son los $q+1$ vértices de $\mathbf{G}^*$. El vértice $q+1$ se identifica con las bandas alélicas de $\mathbf{V} = \{A_{q+1}, \ldots, A_k\}$ y cada vértice restante i se identifica con la banda alélica A_i. Cada arista se corresponde con el par de bandas alélicas de cada persona de $\{I_1, I_2, \ldots, I_n\}$ mediante el siguiente proceso.

La arista asociada con las bandas alélicas $\{D_{2j-1}, D_{2j}\}$ de I_j es (i,s) si $D_{2j-1} = A_i$ y $D_{2j} = A_s$, para $1 \leqslant i, s \leqslant q$; es $(i, q+1)$ si $D_{2j-1} = A_i$ y $D_{2j} \in \mathbf{V}$ ó $D_{2j} = A_i$ y $D_{2j-1} \in \mathbf{V}$, $1 \leqslant i \leqslant q$; y es el bucle $(q+1, q+1)$ si $D_{2j-1} \in \mathbf{V}$ y $D_{2j} \in \mathbf{V}$.

Ejemplo

Sea $\mathbf{B} = \{A_1, A_2, A_3\}$, $\mathbf{V} = \{A_4, A_5\}$ y siete personas $\{I_1, I_2, \ldots, I_7\}$ que aportan el material genético

$$\mathbf{L}_7 = \{A_4, A_1;\ A_2, A_2;\ A_4, A_4;\ A_1, A_3;\ A_2, A_3;\ A_3, A_4;\ A_1, A_3\}$$

El multigrafo asociado $\mathbf{G}^* = (\mathbf{h}, \mathbf{a}^*)$ tiene cuatro vértices; esto es $\mathbf{h} = \{1, 2, 3, 4\}$ y 7 aristas $\mathbf{a}^* = \{(1,4)(2,2)(4,4)(1,3)(2,3)(3,4)(1,3)\}$. Este multigrafo se representa en la figura 5.1

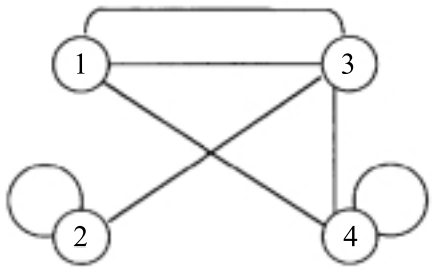

Figura 5.1.

En términos del multigrafo $\mathbf{G}^*$, la solución del problema *PR*1 consiste en hallar el mínimo número de aristas que recubran los vértices $\{1,2,3\}$ del grafo $\mathbf{G}^*$.

5.2. Resolución del problema

Es claro que, para hallar la solución del problema *PR*1, las aristas repetidas son redundantes, pues recubren los mismos vértices (alelos) repetidamente. Por ello se deben suprimir, en un primer paso, todas las aristas y bucles que estén repetidos. Aplicando este proceso al grafo de la figura 5.1 se obtiene el grafo de la figura 5.2.

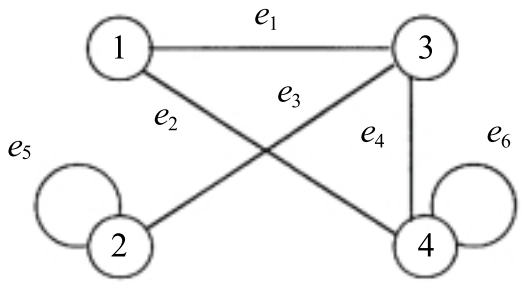

Figura 5.2.

Eliminadas las aristas repetidas del multigrafo $\mathbf{G}^* = (\mathbf{h}, \mathbf{a}^*)$ queda el unigrafo $\mathbf{G} = (\mathbf{h}, \mathbf{a})$, con $\mathbf{a} \subset \mathbf{a}^*$.

Por construcción del grafo **G**, cada uno de sus vértices está recubierto por al menos una arista de **a**, y por tanto tiene solución el problema de recubrimiento de los vértices **h** de **G** por sus aristas.

Si **G** no es un grafo conexo, la resolución del problema de mínimo recubrimiento de los vértices por aristas se realiza resolviendo el problema para cada una de las componentes conexas de **G**, por lo que, sin perder generalidad, se puede suponer que el grafo **G** es conexo.

El problema de recubrimiento de los vértices de un grafo por sus aristas, se resuelve por programación lineal continua, ya que se ha logrado caracterizar el politopo del problema de recubrimiento entero asociado. Esta caracterización está desarrollada en Nemhauser y Wolsey (1999). Debido a su sencillez y brevedad se expone a continuación.

Si $\mathbf{U} \subset \mathbf{h}$ es un subconjunto de vértices del grafo **G** sean

$$E(\mathbf{U}) = \{a \in \mathbf{a} / \text{ Si } a = (v_i, v_j) \qquad \text{entonces } v_i, v_j \in \mathbf{U}\}$$
$$\delta(\mathbf{U}) = \{a \in \mathbf{a} / \text{ Si } a = (v_i, v_j) \qquad \text{entonces } v_i \in \mathbf{U} \text{ y } v_j \in \mathbf{h} - \mathbf{U} \text{ ó viceversa}\}$$

Es claro que para recubrir un conjunto **U** de vértices de cardinal impar, se necesitan al menos $\left\lfloor \frac{|\mathbf{U}|}{2} \right\rfloor + 1$ aristas. De esta forma, son desigualdades válidas para el politopo recubrimiento las siguientes.

Para todo subconjunto de vértices **U** de cardinal impar:

$$\sum_{a \in E(\mathbf{U})} x_a + \sum_{a \in \delta(\mathbf{U})} x_a \geqslant \left\lfloor \frac{|\mathbf{U}|}{2} \right\rfloor + 1$$

Estas desigualdades, unidas a las restricciones de grado y no negatividad de las variables, definen el politopo de recubrimiento, que es la envoltura convexa de los recubrimientos por aristas. Estas ideas se sintetizan en el teorema siguiente.

Teorema 5.2 (Nemhauser-Wolsey)

La envoltura convexa del problema del recubrimiento de los vértices de un grafo $\mathbf{G} = (\mathbf{h}, \mathbf{a})$ *por sus aristas, se caracteriza por las desigualdades:*

$$\sum_{e \in \delta(v)} x_e \geqslant 1 \qquad \forall v \in \mathbf{h}$$
$$\sum_{e \in E(\mathbf{U}) \cup \delta(\mathbf{U})} x_e \geqslant \left\lfloor \frac{|\mathbf{U}|}{2} \right\rfloor + 1$$
$$0 \leqslant x_e \leqslant 1 \qquad \forall e \in \mathbf{a}$$

para todos los subconjuntos **U** *de vértices de cardinal impar.*

□

Se finaliza este apartado hallando la solución del correspondiente problema del grafo de la figura 5.2. En este ejemplo sólo hay una faceta de cardinal impar, que corresponde a $\mathbf{U} = \{1,2,3\}$.

Para la formulación del politopo de recubrimiento, se renumeran las aristas, llamando $e_1 \equiv (1,3)$, $e_2 \equiv (1,4)$, $e_3 \equiv (2,3)$, $e_4 \equiv (3,4)$, $e_5 \equiv (2,2)$ y $e_6 \equiv (4,4)$. Con esta notación la formulación del problema es:

$$
\begin{array}{ll}
\text{Mín} & x_{e_1} + x_{e_2} + x_{e_3} + x_{e_4} + x_{e_5} + x_{e_6} \\
& x_{e_1} + x_{e_2} \geqslant 1 \\
& x_{e_3} + x_{e_5} \geqslant 1 \qquad\qquad 0 \leqslant x_{e_j} \leqslant 1 \quad 1 \leqslant j \leqslant 6 \\
& x_{e_1} + x_{e_3} + x_{e_4} \geqslant 1 \\
& x_{e_1} + x_{e_2} + x_{e_3} + x_{e_4} + x_{e_5} \geqslant \left\lfloor \frac{3}{2} \right\rfloor + 1 = 2
\end{array}
$$

El problema tiene varias soluciones, cada una de ellas formada por dos aristas. Son soluciones $\{e_1, e_5\}$, $\{e_1, e_3\}$, $\{e_2, e_3\}$.

Subconjuntos compatibles de personas de tamaño mínimo son $\{I_4, I_2\}$, $\{I_7, I_2\}$, $\{I_4, I_5\}$, $\{I_7, I_5\}$ y $\{I_1, I_5\}$.

6. CÁLCULO DE EVIDENCIAS EN DIVERSOS SUPUESTOS PRÁCTICOS

En la tabla 6.1 se calcula el valor de $P(\mathbf{E})$, inverso de la evidencia, para diferentes valores de $P_{\mathbf{V}}$, $P_{\mathbf{B}}$ y q, número de alelos de **B**. Se asignan valores aleatorios a las probabilidades de los q alelos de **B** de tal forma que $\sum_{j=1}^{q} p_j = P_{\mathbf{B}}$.

Se supone que éstas son las probabilidades alélicas conocidas en la población, que se cumple la ley de Hardy-Weinberg y que se analiza un único STR.

Se ha calculado $P(\mathbf{E})$, para un valor de n máximo $nmax = 50$. Si el número de alelos crece, es necesario aumentar $nmax$ para estimar $P(\mathbf{E})$ con más precisión.

Puede observarse lo siguiente:

1. Fijado q, al aumentar $P_{\mathbf{B}}$ aumenta $P(\mathbf{E})$, disminuyendo la evidencia.
2. Fijado $P_{\mathbf{B}}$ y q al aumentar $P_{\mathbf{V}}$ aumenta $P(\mathbf{E})$, disminuyendo la evidencia, aunque no exageradamente. Como es lógico, $P(\mathbf{E})$ aumenta con la suma $P_{\mathbf{V}} + P_{\mathbf{B}}$, llegando a ser $P(\mathbf{E}) = 1$ cuando $P_{\mathbf{V}} + P_{\mathbf{B}} = 1$.

Tabla 6.1. *Valores de $P(\mathbf{E})$ en diversos supuestos*

$P_{\mathbf{V}}$	$P_{\mathbf{B}}$	$P_{\mathbf{V}}+P_{\mathbf{B}}$	q	$P(\mathbf{E})$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
0.10	0.05	0.15	1	0.01262626	0.050							
0.10	0.05	0.15	2	0.00130111	0.021	0.029						
0.10	0.05	0.15	3	0.00001279	0.011	0.024	0.015					
0.10	0.05	0.15	4	0.00000061	0.007	0.015	0.013	0.015				
0.10	0.05	0.15	5	0.00000001	0.009	0.007	0.009	0.012	0.013			
0.10	0.05	0.15	6	0.00000000	0.008	0.006	0.010	0.011	0.007	0.008		
0.10	0.05	0.15	7	0.00000000	0.008	0.005	0.009	0.005	0.005	0.009	0.009	
0.10	0.05	0.15	8	0.00000000	0.003	0.008	0.007	0.005	0.006	0.008	0.009	0.004
0.10	0.20	0.30	1	0.08080808	0.200							
0.10	0.20	0.30	2	0.01879462	0.144	0.056						
0.10	0.20	0.30	3	0.00144784	0.046	0.064	0.090					
0.10	0.20	0.30	4	0.00020712	0.058	0.068	0.037	0.037				
0.10	0.20	0.30	5	0.00001764	0.033	0.047	0.027	0.048	0.045			
0.10	0.20	0.30	6	0.00000191	0.030	0.045	0.044	0.029	0.030	0.022		
0.10	0.20	0.30	7	0.00000015	0.039	0.034	0.023	0.037	0.031	0.020	0.017	
0.10	0.20	0.30	8	0.00000001	0.015	0.022	0.026	0.035	0.014	0.032	0.023	0.033
0.10	0.35	0.45	1	0.19444444	0.350							
0.10	0.35	0.45	2	0.07476469	0.221	0.129						
0.10	0.35	0.45	3	0.01119406	0.115	0.164	0.070					
0.10	0.35	0.45	4	0.00315265	0.090	0.097	0.083	0.079				
0.10	0.35	0.45	5	0.00040041	0.097	0.040	0.097	0.040	0.077			
0.10	0.35	0.45	6	0.00009305	0.080	0.080	0.038	0.051	0.061	0.040		
0.10	0.35	0.45	7	0.00001786	0.049	0.027	0.047	0.060	0.041	0.062	0.064	
0.10	0.35	0.45	8	0.00000307	0.023	0.034	0.049	0.034	0.061	0.044	0.061	0.045
0.10	0.50	0.60	1	0.35353534	0.500							
0.10	0.50	0.60	2	0.18294196	0.207	0.293						
0.10	0.50	0.60	3	0.05507759	0.159	0.141	0.200					
0.10	0.50	0.60	4	0.01630503	0.074	0.102	0.135	0.188				
0.10	0.50	0.60	5	0.00533386	0.128	0.078	0.129	0.100	0.065			
0.10	0.50	0.60	6	0.00192446	0.080	0.082	0.092	0.095	0.089	0.063		
0.10	0.50	0.60	7	0.00046448	0.045	0.066	0.109	0.060	0.050	0.073	0.097	
0.10	0.50	0.60	8	0.00013885	0.054	0.036	0.050	0.097	0.077	0.053	0.082	0.051
0.10	0.65	0.75	1	0.55808079	0.650							
0.10	0.65	0.75	2	0.37073513	0.288	0.362						
0.10	0.65	0.75	3	0.16721177	0.139	0.252	0.259					
0.10	0.65	0.75	4	0.08777120	0.136	0.201	0.188	0.125				
0.10	0.65	0.75	5	0.04187640	0.105	0.159	0.158	0.128	0.099			
0.10	0.65	0.75	6	0.01409525	0.077	0.165	0.063	0.057	0.146	0.140		
0.10	0.65	0.75	7	0.00898127	0.082	0.117	0.073	0.071	0.085	0.132	0.090	
0.10	0.65	0.75	8	0.00461954	0.076	0.086	0.080	0.103	0.066	0.076	0.079	0.085

Tabla 6.1. *Valores de $P(\mathbf{E})$ en diversos supuestos (Cont.)*

$P_{\mathbf{V}}$	$P_{\mathbf{B}}$	$P_{\mathbf{V}}+P_{\mathbf{B}}$	q	$P(\mathbf{E})$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
0.20	0.05	0.25	1	0.02343750	0.050							
0.20	0.05	0.25	2	0.00149686	0.031	0.019						
0.20	0.05	0.25	3	0.00002957	0.014	0.018	0.019					
0.20	0.05	0.25	4	0.00000096	0.018	0.009	0.013	0.010				
0.20	0.05	0.25	5	0.00000002	0.010	0.005	0.012	0.013	0.009			
0.20	0.05	0.25	6	0.00000000	0.009	0.006	0.011	0.007	0.007	0.010		
0.20	0.05	0.25	7	0.00000000	0.010	0.010	0.004	0.010	0.006	0.006	0.005	
0.20	0.05	0.25	8	0.00000000	0.004	0.006	0.005	0.007	0.007	0.006	0.007	0.007
0.20	0.20	0.40	1	0.12500000	0.200							
0.20	0.20	0.40	2	0.02547343	0.134	0.066						
0.20	0.20	0.40	3	0.00255608	0.049	0.096	0.055					
0.20	0.20	0.40	4	0.00030535	0.035	0.077	0.058	0.031				
0.20	0.20	0.40	5	0.00003621	0.027	0.032	0.057	0.034	0.050			
0.20	0.20	0.40	6	0.00000363	0.024	0.031	0.048	0.047	0.032	0.017		
0.20	0.20	0.40	7	0.00000047	0.034	0.027	0.031	0.032	0.015	0.030	0.032	
0.20	0.20	0.40	8	0.00000004	0.019	0.027	0.020	0.038	0.016	0.038	0.013	0.029
0.20	0.35	0.55	1	0.27343749	0.350							
0.20	0.35	0.55	2	0.09856715	0.136	0.214						
0.20	0.35	0.55	3	0.02251314	0.102	0.133	0.115					
0.20	0.35	0.55	4	0.00543509	0.098	0.097	0.062	0.093				
0.20	0.35	0.55	5	0.00104799	0.087	0.037	0.068	0.067	0.090			
0.20	0.35	0.55	6	0.00021431	0.088	0.036	0.052	0.035	0.059	0.079		
0.20	0.35	0.55	7	0.00005886	0.037	0.036	0.059	0.059	0.051	0.062	0.046	
0.20	0.35	0.55	8	0.00001011	0.049	0.032	0.060	0.051	0.033	0.053	0.052	0.021
0.20	0.50	0.70	1	0.46874999	0.500							
0.20	0.50	0.70	2	0.24653688	0.292	0.208						
0.20	0.50	0.70	3	0.08907496	0.214	0.177	0.109					
0.20	0.50	0.70	4	0.03167564	0.163	0.172	0.092	0.073				
0.20	0.50	0.70	5	0.01271238	0.138	0.118	0.104	0.078	0.062			
0.20	0.50	0.70	6	0.00481702	0.050	0.104	0.085	0.111	0.072	0.077		
0.20	0.50	0.70	7	0.00173363	0.053	0.071	0.072	0.045	0.087	0.067	0.105	
0.20	0.50	0.70	8	0.00070759	0.052	0.071	0.044	0.062	0.071	0.060	0.059	0.081
0.20	0.65	0.85	1	0.71093748	0.650							
0.20	0.65	0.85	2	0.48876294	0.215	0.435						
0.20	0.65	0.85	3	0.30504636	0.142	0.225	0.283					
0.20	0.65	0.85	4	0.17157925	0.082	0.211	0.153	0.203				
0.20	0.65	0.85	5	0.09504163	0.190	0.075	0.084	0.193	0.107			
0.20	0.65	0.85	6	0.05957792	0.083	0.054	0.113	0.127	0.133	0.140		
0.20	0.65	0.85	7	0.03566955	0.055	0.108	0.116	0.120	0.060	0.089	0.103	
0.20	0.65	0.85	8	0.02045982	0.065	0.078	0.105	0.076	0.043	0.119	0.081	0.082

3. Fijado $P_{\mathbf{B}}$, destaca el fuerte decrecimiento de $P(\mathbf{E})$, al aumentar el número de alelos q; como consecuencia el aumento de la evidencia es grande.

4. Incluso para probabilidades grandes, $P_{\mathbf{V}} = 0.2$ y $P_{\mathbf{B}} = 0.65$, si el número de bandas alélicas es superior a 4, la probabilidad $P(\mathbf{E})$ es pequeña y por tanto la evidencia es grande.

5. Como se infiere de 4., supuesto la independencia de los sistemas, sería necesario observar pocos sistemas polimórficos para obtener evidencias mayores que un millón, cifra superior a la evidencia que las leyes de cualquier país fijan, para que sea tenida en cuenta a efectos legales.

7. CONSIDERACIONES SOBRE APLICACIONES PRÁCTICAS

El artículo tiene una gran relevancia para las aplicaciones prácticas, debido a que permite calcular, de forma rápida, la evidencia del material genético encontrado en el lugar donde se cometió el hecho delictivo; bien para un solo sistema STR, o para más de uno; utilizando, en este último supuesto, la regla de la probabilidad producto condicional o independiente, según los casos. Además, debido al carácter general de la formulación de los problemas, sus posibilidades de aplicación son muy amplias.

Como consecuencia de los cálculos, cuando la evidencia es pequeña, el material genético tiene escasa relevancia en el esclarecimiento del caso, y se puede poner en duda la realización de pruebas genéticas a los presuntos culpables. Sin embargo, ante una gran evidencia del material genético, sería recomendable la realización de dichas pruebas.

La utilización, en el cálculo de la evidencia, de las probabilidades alélicas estimadas, en función de las frecuencias de los datos muestrales disponibles, no supone ningún inconveniente en la práctica, puesto que se suele mantener un cierto orden entre las magnitudes de la verdadera evidencia y de la evidencia estimada. Por ello, los métodos de cálculo expuestos en el artículo son también de gran utilidad práctica, en este supuesto.

La mayor, y principal, ventaja de la aplicación de los métodos de la genética a casos delictivos es la de decidir sobre la inocencia de posibles sospechosos, cuando su ADN no coincide con el hallado en el lugar del delito. Salvo en el caso de mutaciones o errores en las medidas, la decisión sobre la inocencia sería, en estos casos, segura.

Uno de los mayores inconvenientes es precisar el espacio poblacional de referencia, en el que se encuentran los verdaderos culpables.

También puede ser un verdadero inconveniente la utilización de la independencia entre sistemas en el cálculo de probabilidades. Este hecho suele dar lugar, muy frecuente-

mente, a una excesiva sobreestimación de la verdadera evidencia. Por otra parte, no se suele disponer, en las bases de datos genéticas, de información suficiente para calcular, con cierta precisión, las probabilidades que se necesitan para estimar el producto condicional. En nuestra opinión, estos son los mayores inconvenientes con que podemos encontrarnos en la práctica.

AGRADECIMIENTOS

Los autores están muy agradecidos al Dr. Ruíz de la Cuesta, así como al Dr. Arroyo, ambos pertenecientes al Departamento de Medicina Legal de la Universidad Complutense de Madrid, pues el primer problema surgió de las conversaciones que los autores mantuvieron con ellos. También están muy agradecidos a los valiosos comentarios de dos revisores anónimos, que han servido para mejorar notablemente la redacción de este artículo.

BIBLIOGRAFÍA

Feller, W. (1973). *Introducción a la Teoría de las Probabilidades y sus Aplicaciones*. Limusa-Wiley.

Foreman, L.A.; Smith, A.F.M. and Evett, I.W. (1997). «A Bayesian Approach to validating STR Multiplex Databases for use in Forensic Casework». *Int. J. Legal Med.*, 110, 244-250.

Louis, E.J. and Dempster, E.R. (1987). «An Exact Test for Hardy-Weinberg and Multiple Alleles». *Biometrics*, 43, 805-811.

Nemhauser, G.L. and Wolsey, L.A. (1999). *Integer and Combinatorial Optimization*. John Wiley.

ENGLISH SUMMARY

CALCULATION OF EVIDENCE IN THE ANALYSIS OF GENETIC COMPATIBLE PROFILES

MIGUEL SÁNCHEZ GARCÍA*
PEDRO CUESTA ALVARO**
University Complutense of Madrid

One or more people have committed an offence. To discover the number of presumed delinquents, the genetic profile of the forensic evidence found in the place of the offence has been analysed, by means of the study of STR markers systems. Several algorithms are developed for that purpose in this paper. In the first one we consider the probability that n individuals, chosen at random from the reference population, contain a compatible genetic profile. In the second, the evidence of the genetic material is calculated, and in the third, the minimum number of presumed guilties, selected from a compatible group, is found. Finally we show and analyse evidences in several particular situations.

Keywords: STR, allele, inclusion-exclusion principle, multigraph, set covering problem

AMS Classification (MSC 2000): 60C05,92D20

---

* Departamento de Estadística. Facultad de Medicina. Universidad Complutense de Madrid. 28040 Madrid. Tel.: 913941666. E-mail: mesegar@eucmos.sim.ucm.es
** Centro de Proceso de Datos. Servicios Informáticos. Avda. Complutense s/n. 28040 Madrid. Tel.: 913944755. E-mail: pedro@sim.ucm.es

– Received March 1998.
– Accepted October 2000.

A crime has been committed against one or more persons, who will be referred to as *victim*. The genetic profile belonging to one o more STR (Short Tandem Repeat), found in the crime scene, is analyzed. For each STR, a person supplies two alleles Ai,Aj, that can be either equal or distinct.

For a single STR, let $\mathbf{V} = \{A_{q+1}, \ldots, A_k\}$ the set of different alleles that belong to the victim and $\mathbf{B} = \{A_1, A_2, \ldots, A_q\}$ the alleles that do not. $\mathbf{B} \cup \mathbf{V}$ is the set of k different alleles found in the scene of the crime.

A group of n persons, randomly selected from a population, supplies a list of $2n$ alleles, $\mathbf{L}_n = \{D_1, D_2, \ldots, D_{2n-1}, D_{2n}\}$. $\mathbf{L}_n$ is *compatible* with the genetic information picked up in $\mathbf{B}$ and $\mathbf{V}$, when the following two conditions are matched.

c1) Any D_j of the list $\mathbf{L}_n$, is such as $D_j \in \mathbf{B} \cup \mathbf{V}$

c2) For any $A_i \in \mathbf{B} = \{A_1, \ldots, A_q\}$, there is D_j of $\mathbf{L}_n$ such as $A_i = D_j$

If $\mathbf{L}_n^*$ is the set obtained removing from $\mathbf{L}_n$ the repeated elements, the previous compatibility conditions are equivalent to $\mathbf{B} \subset \mathbf{L}_n^* \subset \mathbf{B} \cup \text{`}\mathbf{V}$.

In the present article the following three problems are formulated and solved.

Problem P1]. Let $\mathbf{D}_n$ be the event that n persons, randomly selected from a population, provide a genetic profile that is compatible with $\mathbf{B}$ and $\mathbf{V}$. Calculate $P(\mathbf{D}_n)$.

Problem P2]. Calculate the probability of the event $\bigcup_{n=1}^{\infty} \mathbf{D}_n$, whose inverse is the evidence.

Problem P3]. Let $\{I_1, I_2, \ldots, I_n\}$ be a set of n persons, whose genetic information is compatible with $\mathbf{B}$ and $\mathbf{V}$. Find a subset $\mathbf{F} \subset \{I_1, I_2, \ldots, I_n\}$, such as $\mathbf{F}$ is the smallest compatible subset.

To solve $P1$ the Probability Theory Inclusion-Exclusion principle is used. Denoting $\mathbf{T} = \mathbf{B} \cup \mathbf{V}$ and $\mathbf{S}_j = \mathrm{T} - \{A_j\}$, for $j = 1, 2, \ldots, q$; the calculations are performed through the formula:

$$
\begin{aligned}
P(\mathbf{D}_n) &= P\left(\mathbf{T}^{2n} - \bigcup_{j=1}^{q} \mathbf{S}_j^{2n}\right) = P\left(\mathbf{T}^{2n}\right) - P\left(\bigcup_{j=1}^{q} \mathbf{S}_j^{2n}\right) = P(\mathbf{T}^{2n}) - \sum_{j=1}^{q} P(\mathbf{S}_j^{2n}) + \\
&+ \sum_{1 \leqslant i < j \leqslant q} P\left(\mathbf{S}_i^{2n} \cap \mathbf{S}_j^{2n}\right) - \sum_{1 \leqslant i < j < r \leqslant q} P\left(\mathbf{S}_i^{2n} \cap \mathbf{S}_j^{2n} \cap \mathbf{S}_r^{2n}\right) + \cdots + \\
&+ (-1)^{q-1} \sum_{j=1}^{q} P\left(\bigcap_{i \neq j}^{q} \mathbf{S}_i^{2n}\right) + (-1)^q P\left(\mathbf{S}_1^{2n} \cap \cdots \cap \mathbf{S}_j^{2n} \cap \cdots \cap \mathbf{S}_q^{2n}\right)
\end{aligned}
$$

In order to make calculations easy, two algorithms have been developed to evaluate $P(\mathbf{D}_n)$ in a fast way, taking into account the existing correspondence between the terms of the previous formula and the binary decomposition of the numbers lower than 2^q.

The problem $P2$ can be solved bearing in mind that

$$
\mathbf{E} = \bigcup_{n=1}^{\infty} \mathbf{D}_n = \mathbf{D}_1 \cup \left[\bigcup_{n=2}^{\infty} \left(\mathbf{D}_n \cap \left(\bigcup_{j=1}^{n-1} \mathbf{D}_j\right)^c\right)\right]
$$

This way the events $\mathbf{D}_1$ and $\mathbf{D}_n \cap \left(\bigcup_{j=1}^{n-1} \mathbf{D}_j\right)^c$, for $n \geqslant 2$, are disjoint.

The probabilities are calculated through the formula

$$
\begin{aligned}
P\left(\mathbf{D}_n \cap \left(\bigcup_{j=1}^{n-1} \mathbf{D}_j\right)^c\right) &= \sum_{i=1}^{q} P(n-1, -A_i)\left[2P(A_i)\,P(\mathbf{B} \cup \mathbf{V}) - P(A_i)\,P(A_i)\right] + \\
&+ \sum_{1 \leqslant i < j \leqslant q} P\left(n-1, -\{A_i, A_j\}\right)\left[2P(A_i)\,P(A_j)\right]
\end{aligned}
$$

$P(n, -A_i)$, $1 \leqslant i \leqslant q$ is the probability that n persons, chosen at random from the population, are compatible with the alleles $\mathbf{B} - \{A_i\}$ and $\mathbf{V}$.

$P(n, -\{A_i, A_j\})$, $1 \leqslant i < j \leqslant q$ is the probability that n persons are compatible with the alleles $\mathbf{B} - \{A_i, A_j\}$ and $\mathbf{V}$.

The calculations of these probabilities are performed by the algorithms for the first problem, respectively applied to $\mathbf{B} - \{A_i\}$, $\mathbf{V}$ and $\mathbf{B} - \{A_i, A_j\}$, $\mathbf{V}$.

Problem $P3$ is modeled by a multigraph. Vertices are identified with the alleles of $\mathbf{B}$, to which a new vertex identifying the alleles of $\mathbf{V}$ is added. Each person, with alleles A_i, A_j, has correspondence with an edge in the multigraph. The goal is to solve a set covering problem, stated as finding the minimum number of edges that covers all hipergraph vertices.

The problem is solved by set covering problem polytopes characterization, developed by Nemhauser and Wolsey.

A computer program has been developed for the algorithms of the problems $P1$ and $P2$. Several tables of results are shown in the article, together with examples of each problem and comments about their practical advantages and disadvantages.