Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

I NETWORK BAYESIANI NELLA GENETICA FORENSE

Relatore Prof. Marco Ferrante
Dipartimento di Matematica Pura ed Applicata

Laureando: Filippo Baldin
Matricola N 1034980

Anno Accademico 2013/2014

Sommario

In questa tesi ho studiato un problema di genetica forense utilizzando i network bayesiani. Estendendo un modello presente in letteratura ho considerato il caso in cui sulla scena di un crimine vengono rilevate tracce organiche miste di due individui. Non si conosce l'identità del colpevole, ma si sospetta di un individuo in particolare. Mentre della vittima si conosce il profilo genetico, del sospetto non si può ricavare questa informazione, e quindi si cerca per quanto possibile di risalire al suo DNA, raccogliendo il DNA del figlio naturale e della moglie co-genitrice.

Abstract

In this thesis I have studied a problem of forensic genetics by using bayesian networks. Starting from a model already studied in the literature, I have considered the case regarding a crime scene in which a two-individuals mixed trace is detected. The identity of the guilty is unknown, but an individual in particular is suspected of committing the crime. While the victim's profile is known, the suspect's one is unavailable, so, as far as possible, it is attempted to trace his DNA through his son's and his wife's (the other parent) DNA.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

La genetica forense è una branca della scienza forense che studia le problematiche relative all'accertamento della paternità o ad analisi del DNA su tracce biologiche. In questa tesi si parlerà solamente di analisi biologiche relative ad episodi criminali.
Un esempio semplice e classico è il delitto passionale: un individuo viene assassinato dal proprio partner ed il movente è l'infedeltà coniugale. Sulla scena del crimine vengono rinvenute tracce biologiche come sangue o saliva che possono poi essere confrontate con il DNA del sospetto.
Nell'esempio appena esposto, lo scienziato forense si trova a dover confrontare il DNA rinvenuto sulla scena del crimine (presumibilmente di un solo individuo), con il DNA del sospetto. In altri casi può essere che le tracce organiche rinvenute provengano da diversi individui e ci troviamo quindi di fronte ad una mistura di diversi profili genetici. In genere una mistura di DNA contiene tracce provenienti da diversi soggetti: dalla vittima, da uno o più colpevoli ed eventualmente da altri individui. Se da una supposizione o da una prova di natura diversa da quella biologica è ragionevole sospettare che un individuo abbia commesso il crimine, allora è utile confrontare il suo profilo genetico e quello della vittima con il DNA rinvenuto sulla scena del crimine.
Casi analoghi utilizzando i network bayesiani sono già stati ampiamente affrontati da [Cowell *e altri*(1999)] e da [Mortera *e altri*(2003)]. Il caso che cercherò di analizzare in questa tesi sarà un crimine in cui il DNA rinvenuto sulla scena del crimine appartiene a due persone, una vittima ed un colpevole.

Non si conosce ancora l'identità del colpevole, ma si sospetta di un individuo del quale non si riesce a ricavare il profilo genetico (è scappato all'estero o è irreperibile). Per supplire a questa carenza di informazione, si cercherà di confrontare il DNA rinvenuto con quelli di figlio e moglie o compagna del sospetto e madre del figlio. Il DNA del figlio serve perché possiede parte del patrimonio genetico del sospetto, mentre quello della compagna (assumendo che questa sia la co-genetrice) serve ad identificare con maggior precisione quale sia la parte di DNA che il figlio ha ereditato dal sospetto. Naturalmente non si può fare un'analisi precisa come in caso di informazione completa, e quindi lo scopo della mia tesi è quello di valutare la bontá del risultato con questa complicazione. Una breve introduzione sui principali concetti di genetica sarà affrontata nel Paragrafo 1.1, mentre nel Paragrafo 1.2 saranno introdotti i network bayesiani. Il network bayesiano costruito col pacchetto *gRain* del software **R** sarà mostrato nel Capitolo 3. Nel Capitolo 4 sarà analizzato il modello e nel Capitolo 5 verranno riportati i risultati ottenuti da un esempio e nel Capitolo 6 verrà presentato una variante al network che tiene conto della possibilità di errori di laboratorio.

1.1 Background genetico

In questa sezione vengono presentati alcuni concetti di base circa i profili DNA e una breve descrizione del processo di amplificazione (PCR) del DNA con i problemi che esso comporta. Una spiegazione più dettagliata per esempio si trova in [Butler(2005)].
Gli scienziati forensi decodificano un profilo genetico analizzando la composizione del DNA in varie posizioni dei cromosomi. La posizione di un gene all'interno di un cromosoma è chiamata *locus*. L'informazione di ogni locus consiste in una coppia non ordinata di *alleli* che forma il *genotipo* in quel locus. Si parla di coppia di cromosomi, in quanto uno viene dal padre e l'altro dalla madre. Si parla di coppia non ordinata, in quanto non si può sapere quale dei due cromosomi viene dal padre e quale dalla madre. Il DNA umano ha ventitre coppie di cromosomi, di cui una che identifica il sesso [Cowell *e altri*(2013)]. Nel modello che seguirà nei capitoli successivi, non si utilizzerà la

coppia di cromosomi sessuali.
Gli alleli di un locus sono sequenze di quattro basi azotate: *adenina*, *citosina*, *guanina* e *timina*, rappresentate rispettivamente dalle lettere A, C, G e T. Un allele è identificato da un *repeat number* che indica quante volte viene ripetuta una certa sequenza di basi. La sequenza ripetuta è diversa per ogni locus: per esempio l'allele 8 del locus TH01 ha otto ripetizioni consecutive della sequenza AATG. Si indica questa sequenza $[AAGT]_8$. Quando un locus ha gli alleli con lo stesso repeat number, allora è detto *omozigote*, in caso contrario *eterozigote*.
Il processo *polymerase chain reaction* (PCR) è una tecnica che consente la moltiplicazione di un frammento di DNA. Ciò avviene attraverso una prima reazione che permette la separazione dei due filamenti e con una seconda che permette di ricreare il filamento corrispondente. Queste due reazioni vengono ripetute solitamente ventotto volte. Quindi un laser illumina il DNA e ne misura la fluorescenza generata. Questa procedura permette di rivelare gli alleli presenti nella traccia su un *elettroferogramma* (EPG) come in Figura 1.1, cioè un diagramma che identifica gli alleli in base alla fluorescenza che emettono. Nell'EPG l'ascissa indica l'allele, mentre l'ordinata indica l'intensità della fluorescenza. Un picco nell'EPG indica al presenza di un allele e viene misurato in base all'altezza o all'area.

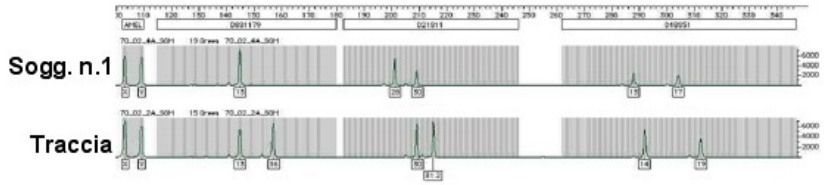

Figura 1.1: Un elettroferogramma (EPG) su quattro loci di due profili genetici, uno rilevato dalla scena del crimine, l'altro appartenente al sospetto. Questa immagine è fornita dal Laboratorio di Genetica Forense, Università degli Studi di Milano

1.1.1 Errori di laboratorio

La procedura di PCR può generare qualche errore. Innanzitutto nella fase di amplificazione del DNA può essere che non tutti gli alleli vengano copiati in ogni ciclo, e quindi le proporzioni dei vari alleli possono differire da prima a dopo il processo.
Inoltre può accadere che l'elettroferogramma presenti dei picchi con altezza minore, quindi l'esigenza di una soglia sotto cui scartare l'ipotesi di presenza di un allele rappresenta una soluzione. Se viene scartato un allele effettivamente presente nel DNA, allora si dice che si è verificato un *dropout*. Nel caso invece in cui ad un picco superiore alla soglia (e quindi non scartato) non corrispondesse un allele effettivamente presente nel DNA, si parla di *dropin*. Il *dropin* può avvenire in caso di contaminazione della traccia biologica.
Un altro problema comune è quello di *stutter* ed avviene quando, durante la fase di amplificazione del DNA, una sequenza di quattro basi non viene copiata, dando origine ad un picco minore con *repeat number* inferiore.

1.1.2 Assunzioni di base del modello

Il modello che sarà presentato nel prossimo capitolo è pensato per una popolazione che segua le ipotesi di base dell'equilibrio di Hardy-Weinberg:

- casualità degli accoppiamenti;
- popolazione di grandi dimensioni;
- nessun flusso genico, ovvero nessuna migrazione da una popolazione all'altra;
- nessuna mutazione, cioè gli alleli non si trasformano né se ne presentano di nuovi;
- gli individui con genotipi diversi hanno le stesse possibilità di sopravvivenza, ovvero non avviene la selezione naturale.

Queste assunzioni implicano una certa stazionarietà nella popolazione, in particolare per quanto riguarda la frequenza degli alleli.

Altre assunzioni più restrittive riguardano le fasi di laboratorio per la moltiplicazione del segmento di DNA da analizzare. Non verrà considerata la possibilità di *stutter*, ma verranno considerate quelle di *drop-in* e *drop-out*. In questo senso, la letteratura della genetica forense ha una bipartizione tra chi tratta l'altezza dei picchi da un punto di vista quantitativo e chi da un punto di vista qualitativo. In questa tesi si opterà per l'ambito qualitativo e quindi per ogni locus un allele verrà considerato una variabile booleana.

1.2 Breve introduzione sui network bayesiani

Il network bayesiano è un tipo di *modello grafico* che serve a rappresentare un insieme di eventi connessi tra di loro. Proprio per questo motivo la costruzione di un network bayesiano cerca di seguire il criterio di causalità, cioè si cerca di relazionare gli eventi secondo un ordine logico. Grazie a questo criterio, un network bayesiano risulta facile da costruire e da comprendere se si possiede un'adeguata conoscenza del problema.
Da un'ottica rigorosamente matematica, un network bayesiano è composto dai seguenti elementi:

- un insieme $\mathbf{X} = \{X_1, ..., X_n\}$ di variabili aleatorie che rappresenta l'insieme dei *nodi*;
- un insieme di *archi* $\mathbf{E} = \{E_{ij}\}$ che esprime le interdipendenze tra le variabili che compongono $\mathbf{X}$.

Ogni variabile X_i ha un insieme finito di stati. Le variabili e gli archi formano un *grafo aciclico diretto* (un grafo diretto è detto *aciclico* se non esiste nessun percorso $A_1 \to ... \to A_n$ tale che $A_1 = A_n$) ([Jensen e Nielsen(2001)]). Se un arco va da A a B come in Figura 1.2, allora A è *genitore* di B e B è *figlio* di A.

In generale, data una variabile A con genitori $B_1, ..., B_n$, si definisce:

$$pa(A) := \{B_1, ..., B_n\}.$$

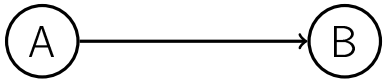

Figura 1.2: Un esempio di relazione causale

Ad ogni variabile A tale che $pa(A) = \{B_1, ..., B_n\}$ è associata una *tabella di probabilità* $\Pr(A \mid B_1, ..., B_n)$. Se A non ha genitori, allora la tabella delle probabilità corrispondente a quel nodo sarà la distribuzione incondizionata $\Pr(A)$. Per esempio, per quanto riguarda il network in Figura 1.3, le distribuzioni di probabilità associate ad ogni nodo sono $\Pr(A)$, $\Pr(B)$, $\Pr(C \mid A, B)$ e $\Pr(D \mid C)$. La distribuzione congiunta di tutte le variabili

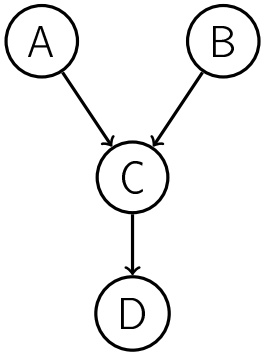

Figura 1.3: Un esempio di network bayesiano

del network in Figura 1.3 è data dal prodotto delle tabelle di probabilità di tutte le variabili presenti.

$$\Pr(A, B, C, D) = \Pr(A) \Pr(B) \Pr(C \mid A, B) \Pr(D \mid C)$$

In generale, dato $U = \{A_1, ..., A_n\}$,

$$\Pr(U) = \prod_{i=1}^{N} \Pr(A_i \mid pa(A_i)) \tag{1.1}$$

Per una dimostrazione si rimanda a [Jensen e Nielsen(2001)]. L'equazione 1.1 rappresenta un punto molto importante, in quanto semplifica notevolmente il calcolo della probabilità congiunta.
Esistono diversi tipi di relazioni tra variabili:

- le *relazioni seriali* (Figura 1.4);
- le *relazioni divergenti* (Figura 1.5);
- le *relazioni convergenti* (Figura 1.6).

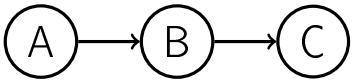

Figura 1.4: Relazione seriale

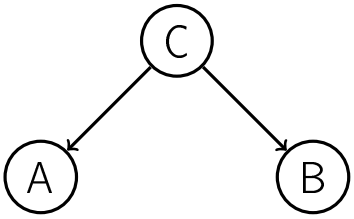

Figura 1.5: Relazione divergente

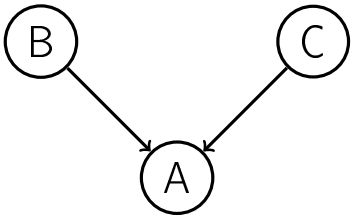

Figura 1.6: Relazione convergente

Un concetto fondamentale nel costruire network bayesiani è l'indipendenza condizionale [Taroni *e altri*(2006)]. Dato $U = \{A, B, C\}$, A e B sono

indipendenti condizionatamente a C se $\Pr(A \mid B, C) = \Pr(A \mid C)$. Per una relazione divergente come in Figura 1.5, si dimostra che A e B sono indipendenti condizionatamente a C. Per l'equazione 1.1 si ha che:

$$\begin{aligned} \Pr(A, B, C) &= \Pr(A \mid C) \Pr(B \mid C) \Pr(C) \\ &= \Pr(A \mid C) \Pr(B, C) \end{aligned} \tag{1.2}$$

Sfruttando la definizione di probabilità condizionata, si ha che:

$$\begin{aligned} \Pr(A \mid B, C) &= \frac{\Pr(A, B, C)}{\Pr(B, C)} \\ &= \frac{\Pr(A \mid C) Pr(B, C)}{\Pr(B, C)} \\ &= \Pr(A \mid C) \end{aligned} \tag{1.3}$$

Analogamene si può anche dimostrare l'indipenenza tra A e C condizionatamente a B.
Per quanto riguarda la relazione seriale (Figura 1.4), la dimostrazione dell'indipendenza tra A e C condizionatamente a B, è, cioé $\Pr(C \mid A, B) = \Pr(C \mid B)$, è la seguente:

$$\begin{aligned} \Pr(A, B, C) &= \Pr(A) \Pr(B \mid A) \Pr(C \mid B) \\ &= \Pr(A, B) \Pr(C \mid B) \end{aligned} \tag{1.4}$$

$$\begin{aligned} \Pr(C \mid A, B) &= \frac{\Pr(A, B, C)}{\Pr(A, B)} \\ &= \frac{\Pr(A, B) \Pr(C \mid B)}{\Pr(A, B)} \\ &= \Pr(C \mid B) \end{aligned} \tag{1.5}$$

Per quanto riguarda la relazione convergente, B e C sono stocasticamente indipendenti.

$$\Pr(A, B, C) = \Pr(A \mid B, C) \Pr(B) \Pr(C) \tag{1.6}$$

Vale però che

$$\Pr(A, B, C) = \Pr(A \mid B, C) \Pr(B, C) \tag{1.7}$$

da cui

$$\begin{aligned} \Pr(A \mid B, C) \Pr(B, C) &= \Pr(A \mid B, C) \Pr(B) \Pr(C) \\ \Pr(B, C) &= \Pr(B) \Pr(C) \end{aligned} \tag{1.8}$$

I network bayesiani sono utilizzati per calcolare la probabilità di un evento quando si ha un aumento o una diminuzione dell'informazione. Per esempio nel network in Figura 1.6, una nuova informazione nel nodo B (B=b) ha un effetto sul nodo A. Infatti la tabella di probabilità associata al nodo A diventa $\Pr(A \mid b, C)$ e quindi riducendo il numero di variabili che ancora possono condizionare A.

Capitolo 2

La verosimiglianza

Essendo il profilo DNA di un individuo un insieme di loci e ad ogni locus corrisponde una coppia non ordinata di alleli, il confronto tra il profilo DNA del sospetto con quello ritrovato sulla scena del crimine si scompone in una verifica per ogni singolo locus. Maggiore è il numero di loci che viene analizzato, maggiore è la precisione. Un elevato numero di loci può generare problemi computazionali come un network bayesiano con un elevato numero di nodi. In particolare, se i loci sono dipendenti fra di loro, allora il grafo aumenta di dimensione. Si può superare questo problema con opportune assunzioni: se si ipotizza indipendenza degli alleli di un individuo tra loci ed entro i loci come in [Mortera *e altri*(2003)], allora basterà fare un network bayesiano per ogni locus e derivare il risultato congiunto come una composizione dei risultati per ogni singolo locus.
La verosimiglianza per una particolare ipotesi (H) è la probabilità di osservare tutte le prove sotto quest'ipotesi ed è quindi data da $\Pr(I \mid H, M)$. Con M si intende l'informazione e tutte le prove che si raccolgono circa la misture di DNA e con I si intende l'informazione circa i genotipi conosciuti degli individui implicati nel caso. Per confrontare due diverse ipotesi (H_0 contro H_1 per esempio) si usa il rapporto tra le due verosimiglianze, e cioè:

$$\frac{\Pr(H_0 \mid I, M)}{\Pr(H_1 \mid I, M)}$$

Per il teorema di Bayes,

$$\frac{\Pr(H_0 \mid I, M)}{\Pr(H_1 \mid I, M)} = \frac{\Pr(I \mid H_0, M)}{\Pr(I \mid H_0, M)} \cdot \frac{\Pr(H_0 \mid M)}{\Pr(H_1 \mid M)} \tag{2.1}$$

Il primo membro rappresenta l'*odd a posteriori*, mentre il secondo membro rappresenta il prodotto tra il *rapporto di verosimiglianza* (LR) e l'*odd a priori*. H_0 indica l'ipotesi di colpevolezza del sospetto, mentre H_1 indica quella di innocenza. In questo caso l'odd a posteriori confronta le probabilità di due diverse ipotesi data una determinata mistura ed un determinato background genetico. Il rapporto di verosimiglianza, invece, confronta le probabilità di avere una certa mistura date due diverse ipotesi. Infine l'odd a priori è il rapporto tra le probabilità di due ipotesi prima di conoscere la mistura. Il rapporto di verosimiglianza converte l'odd a priori in favore di H_0 nel'odd a posteriori in favore di H_1 [Taroni *e altri*(2006)].
Si prenderanno distribuzioni a priori non informative per H_0 e H_1 come $\Pr(H_0 \mid M) = \Pr(H_1 \mid M) = 0.5$. È ragionevole questa congettura per due motivi: da un lato si presuppone equiprobabilità, in quanto la distribuzione a priori si dovrebbe basare su prove non inerenti a quelle riguardanti il DNA (come le impronte digitali) che non abbiamo [Mortera *e altri*(2003)]; da un altro punto di vista, le assunzioni di equilibrio di Hardy-Weinberg fatte in precedenza permettono di pensare che M non influenzi H, ossia che il genotipo di due individui non implichi maggiore o minore probabilità che uno dei due commetta un crimine. L'equazione 2.1 quindi diventa

$$\begin{aligned}
\frac{\Pr(H_0 \mid I, M)}{\Pr(H_1 \mid I, M)} &= \frac{\Pr(I \mid H_0, M)}{\Pr(I \mid H_1, M)} \cdot \frac{\Pr(H_0 \mid M)}{\Pr(H_1 \mid M)} \\
&= \frac{\Pr(I \mid H_0, M)}{\Pr(I \mid H_1, M)} \cdot \frac{0.5}{0.5} \\
&= \frac{\Pr(I \mid H_0, M)}{\Pr(I \mid H_1, M)}
\end{aligned}$$

e così l'odd a posteriori diventa uguale al rapporto di verosimiglianza.
Per N loci, l'informazione I si suddivide in N componenti, $M = \{M_1, ..., M_N\}$ e $I = \{I_1, ..., I_N\}$, dove M_k e I_k sono l'informazione circa il k-esimo locus, il

rapporto di verosimiglianza è definito come:

$$\begin{aligned} LR &= \frac{\Pr(I \mid H_0, M)}{\Pr(I \mid H_1, M)} \\ &= \prod_{n=1}^{N} \frac{\Pr(I_n \mid H_0, M_n)}{\Pr(I_n \mid H_1, M_n)} \end{aligned}$$

L'ipotesi H_1 specificata finora prevede l'innocenza del sospetto. Tale proposizione è incompleta, in quanto non specifica nulla circa il vero colpevole. Sotto l'ipotesi H_1 il sospetto è innocente ed il vero colpevole è un individuo casuale della stessa popolazione. Si supponga ad esempio che per un determinato locus si siano trovati gli alleli A, B e C in una mistura, che la vittima (M_v) abbia genotipo AB e che il sospetto (M_s) abbia genotipo C. In tal caso

$$\Pr(I \mid H_0, M) = \Pr(I = ABC \mid H_0, M_v = AB, M_s = C) = 1$$

in quanto gli alleli presenti nella mistura sono il risultato degli alleli lasciati dai contributori, ossia il sospetto e la vittima. Sotto H_1, invece,

$$\begin{aligned} \Pr(I \mid H_1, M) &= \Pr(I = ABC \mid H_1, M_v = AB, M_s = C) \\ &= 2p_A p_C + 2p_B p_C + p_C^2 \end{aligned}$$

che è la somma delle probabilità che il colpevole sia un individuo eterozigote di genotipi AC e BC, o un omozigote con genotipo C. Infatti se la mistura ha gli alleli A, B e C e la vittima ha genotipo AB, allora il genotipo del colpevole deve contenere l'allele C e quindi può essere AC, BC o C. p_i è la frequenza dell'allele i nella popolazione. Quindi

$$\begin{aligned} LR &= \frac{\Pr(I = ABC \mid H_0, M_v = AB, M_s = C)}{\Pr(I = ABC \mid H_1, M_v = AB, M_s = C)} \\ &= \frac{1}{2p_A p_C + 2p_B p_C + p_C^2} \end{aligned}$$

Tutto ciò è valido sotto le ipotesi di equilibrio di Hardy-Weinberg.
Nel capitolo successivo si descriverà un network bayesiano per un singolo locus, poi nel Capitolo 5 si tratterà un esempio in cui si considerano

congiuntamente tre loci.

Capitolo 3

Il network bayesiano

In questo capitolo saranno descritti tutti i nodi, le relazioni che li legano ed infine sarà fatta una panoramica generale sul modello creato.

3.1 I principali nodi

I nodi utilizzati sono di diversi tipi a seconda degli stati che assumono.
Una relazione che si ripete più volte in questo modello è quella che lega i singoli alleli al genotipo di un locus. Gli alleli saranno identificati da lettere maiuscole, mentre i genotipi saranno identificati da coppie di lettere maiuscole. Le lettere *pg* e *mg* si riferiscono ai nodi "allele" ereditati rispettivamente dal padre e dalla madre e gli stati che possono assumere saranno identificati da lettere A, B, C, ... Per esempio il nodo *spg* si riferisce all'allele ereditato dal padre del sospetto.
Le lettere *gt* si riferiscono ai nodi "genotipo" e gli stati saranno identificati da due lettere maiuscole se il genotipo è eterozigote, una sola se è omozigote. Se l'individuo ha ereditato da entrambi i genitori l'allele A, allora il suo genotipo viene indicato con A (e non AA come si potrebbe pensare). Essendo gli alleli una coppia non ordinata, per ogni coppia di alleli diversi fra loro esiste un solo genotipo. Se ad esempio un individuo eredita dai propri genitori gli alleli A e B, allora il genotipo sarà indicato come AB invece di BA.
Bisogna distinguere due tipologie di individui: quelli di cui conosciamo l'i-

dentità e quelli che hanno effettivamente contribuito alla mistura. In generale, questi ultimi saranno indicati come *T1gt* e *T2gt* per quanto riguarda i genotipi, *T1pg*, *T1mg*, *T2pg* e *T2mg* per quanto riguarda gli alleli. Gli effettivi contributori vengono confrontati con i nodi *sgt* ed *vgt* (rispettivamente i genotipi di sospetto e vittima, cioè coloro di cui conosciamo l'identità) attraverso i nodi "test". Le lettere *T1* e *T2* indicano i nodi "test" rispettivamente per il sospetto e per la vittima. Per la loro natura, assumono gli stati "Yes" o "No".
Il nodo *Target* si riferisce alle ipotesi riguardanti congiuntamente sospetto e vittima. Assume uno tra questi quattro stati:

- *v & s* quando sia la vittima che il sospetto hanno contribuito alla mistura di DNA;
- *v & U* quando la vittima ed un individuo casuale della popolazione (ma non il sospetto) hanno contribuito;
- *s & U* quando il sospetto ed un individuo casuale della popolazione (ma non l'ipotetica vittima) hanno contribuito;
- *2U* quando a contribuire sono stati due individui casuali della popolazione.

Infine il nodo *mix* che indica gli alleli evidenziati dall'EPG. Analogamente ai noti *gt*, gli stati del nodo *mix* sono indicati da lettere maiuscole. Vale anche in questo caso che ogni gruppo di alleli evidenziati dall'EPG è identificato da un solo stato. Per esempio, se l'elettroferogramma evidenzia gli alleli A, B e C, la variabile *mix* assumerà lo stato ABC.

3.2 Le relazioni di causalità

Per costruire il network, bisogna creare dei legami tra i vari nodi. La logica del network bayesiano e la conoscenza del problema da affrontare permette di capire quali siano queste relazioni e quale sia la direzione di causalità.
Per quanto riguarda vittima, sospetto, moglie, figlio e i veri contributori, la

relazione tra i nodi degli alleli ereditati dai genitori (*pg* e *mg*) ed il nodo genotipo (*gt*) è la stessa. Un esempio per quanto riguarda il sospetto è in Figura 3.1. Per quanto riguarda il legame tra il figlio ed i propri genitori, i

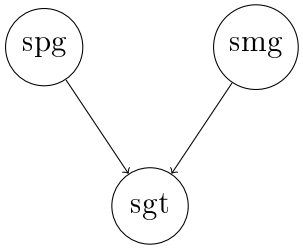

Figura 3.1: Il genotipo del sospetto

nodi relativi agli alleli del figlio (*cpg* e *cmg*) si collegano con i nodi relativi agli alleli dei propri genitori (*spg* e *smg* per il padre, *wpg* e *wmg* per la madre) come nella Figura 3.2.

Figura 3.2: Il figlio eredita un allele da ciascun genitore

Essendo la finalità del modello quella confrontare il DNA di due individui con quello trovato sulla scena del crimine, il nodo *mix* non sarà collegato direttamente con i genotipi o gli alleli dei vari individui. Il meccanismo logico è quello di collegare la presunta vittima (*vgt*) ed il sospetto (*sgt*) rispettivamente con la vera vittima (*T2gt*) ed il colpevole (*T1gt*), cioè gli individui che hanno effettivamente contribuito alla traccia biologica (Figura 3.3 e Figura 3.4).

I nodi *T1gt* e *T2gt* che si riferiscono ai veri contributori saranno poi collegati con il nodo *mix* come in Figura 3.5.

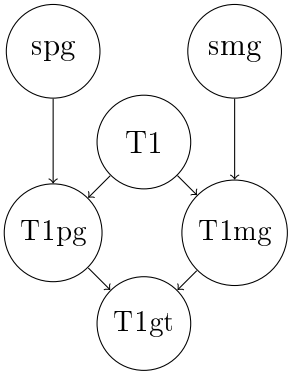

Figura 3.3: Verifica del genotipo del sospetto

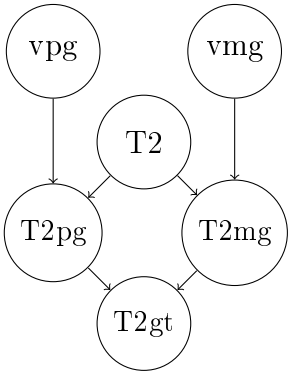

Figura 3.4: Verifica del genotipo della vittima

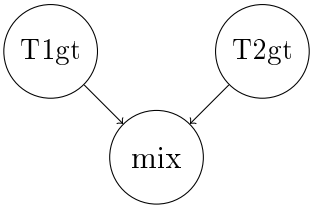

Figura 3.5: Legame tra *mix* e i veri contributori

Infine si considera il nodo *Target* come dipendente dai nodi *T1* e *T2*, come in Figura 3.6. Essendo il nodo *Target* un test più generico rispetto agli altri test, allora conterrà le informazioni di essi.

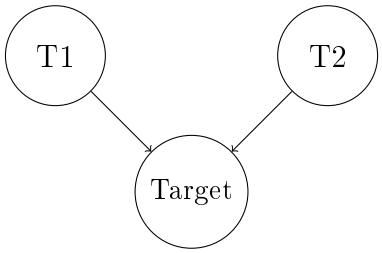

Figura 3.6: Legame tra nodi test

3.3 Le tabelle di probabilità

In questa sezione si parlerà delle tabelle di probabilità che legano i nodi secondo le relazioni esposte nella sezione 3.2.
Per quanto concerne la relazione alleli-genotipo in Figura 3.1, il genotipo sarà come logico quello corrispondente agli alleli ereditati. Si tratta di una relazione di tipo booleano, in quanto un determinato genotipo è possibile se e solo se gli alleli ereditati corrispondono. In questa tesi vengono considerati sei alleli diversi che generano ventuno diverse combinazioni genotipiche. La tabella delle probabilità del nodo *sgt* con tre alleli invece di sei, per esigenze di spazio, è riportato in Tabella 3.1. Naturalmente questa è assolutamente identica a quelle che riguardano i nodi *vgt*, *T1gt*, *T2gt*, *wgt* e *cgt*.

Tabella 3.1: Tabella delle probabilità condizionali per *sgt* dati *spg* e *smg*

smg:	A			B			C		
spg:	A	B	C	A	B	C	A	B	C
A	1	0	0	0	0	0	0	0	0
AB	0	1	0	1	0	0	0	0	0
AC	0	0	1	0	0	0	1	0	0
B	0	1	0	0	1	0	0	0	0
BC	0	0	0	0	0	1	0	1	0
C	0	0	0	0	0	0	0	0	1

La parte che mette in relazione il sospetto con il colpevole come in Figura 3.3, oltre a presentare una relazione tra genotipo ed alleli analoga a quella precedente, mostra il legame tra l'allele del sospetto e quello del colpevole, mediato da *T1*. Se *T1=Yes*, allora gli alleli *pg* e *mg* del sospetto e del colpevole coincidono. Se *T1=No*, allora il sospetto è innocente e gli alleli coincideranno solo se l'innocente ed il colpevole hanno (casualmente) lo stesso allele (Tabella 3.2).

Il genotipo del figlio dipende dagli alleli che eredita dai genitori. Essendo che la struttura genotipo-alleli distingue chiaramente tra allele ereditato dal padre o dalla madre, il nodo *cpg* ha come nodi condizionanti *spg* e *smg*.

Tabella 3.2: Tabella delle probabilità condizionali per *T1pg* dati *T1* e *spg*

T1:	Yes						No					
spg:	A	B	C	D	E	F	A	B	C	D	E	F
A	1	0	0	0	0	0	p_A	p_A	p_A	p_A	p_A	p_A
B	0	1	0	0	0	0	p_B	p_B	p_B	p_B	p_B	p_B
C	0	0	1	0	0	0	p_C	p_C	p_C	p_C	p_C	p_C
D	0	0	0	1	0	0	p_D	p_D	p_D	p_D	p_D	p_D
E	0	0	0	0	1	0	p_E	p_E	p_E	p_E	p_E	p_E
F	0	0	0	0	0	1	p_F	p_F	p_F	p_F	p_F	p_F

Tabella 3.3: Tabella delle probabilità per *cmg* dati *wpg* e *wmg*

wmg:	A			B			C		
wpg:	A	B	C	A	B	C	A	B	C
A	1	0.5	0.5	0.5	0	0	0.5	0	0
B	0	0.5	0	0.5	1	0.5	0	0.5	0
C	0	0	0.5	0	0	0.5	0.5	0.5	1

Analogamente *cmg* dipende da *wpg* e *wmg*. In caso di genitore omozigote, l'allele ereditato è certamente quello posseduto dal genitore, mentre in caso di genitore eterozigote, l'allele ereditato è uno dei due con uguale probabilità (Tabella 3.3).

Gli stati dei test *T1* e *T2* vengono assunti come equiprobabili (Tabella 3.4). Si tratta di un'assunzione che non ha influenza sul problema da affrontare, in quanto, dopo aver propagato le informzioni e le prove sul network, la risultante distribuzione a posteriori viene reinterpretata come rapporto di verosimiglianza [Mortera *e altri*(2003)].

Tabella 3.4: Tabella delle probabilità per *T1* o *T2*

Yes	No
0.5	0.5

Essendo *Target* dipendente da *T1* e *T2*, la probabilità condizionata dei vari stati assume valori 0 e 1. Data l'indipendenza tra le variabili *T1* e *T2*, è facilmente dimostrabile che la variabile *Target* ha una distribuzione

marginale uniforme discreta ed ogni stato ha probabilità 0.25. Vale la stessa considerazione fatta in precedenza per i nodi *T1* e *T2* sulla distribuzione a priori.

Tabella 3.5: Tabella delle probabilità per *Target* dati *T1* e *T2*

T1:	Yes		No	
T2:	Yes	No	Yes	No
s & v	1	0	0	0
s & U	0	1	0	0
v & U	0	0	1	0
2U	0	0	0	1

Per quanto concerne gli stati della variabile *mix*, questi hanno probabilità 0 o 1 a seconda che questi siano possibili o meno dati i genotipi condizionanti. Se *mix=ABC*, allora questo stato ha probabilità 1 se e solo se i genotipi *T1gt* e *T2gt* possiedono gli alleli A, B e C. Le combinazioni possibili sono:

- *T1gt=A* e *T2gt=BC*;
- *T1gt=AB* e *T2gt=AC*;
- *T1gt=AB* e *T2gt=BC*;
- *T1gt=AB* e *T2gt=C*;
- *T1gt=AC* e *T2gt=AB*;
- *T1gt=AC* e *T2gt=B*;
- *T1gt=AC* e *T2gt=BC*;
- *T1gt=B* e *T2gt=AC*;
- *T1gt=BC* e *T2gt=A*;
- *T1gt=BC* e *T2gt=AB*;
- *T1gt=BC* e *T2gt=AC*;

- *T1gt=C* e *T2gt=AB*.

Se nella traccia di DNA si dovessero rinvenire quattro alleli diversi, allora le combinazioni diminuirebbero sensibilemente. Dato che con sei alleli la mistura di DNA lasciata da due individui assume 56 stati, la Tabella 3.6 riporta tutti gli stati possibili per *mix* se si dovessero considerare solo tre alleli. Nel caso di sei, la tabella è analoga a questa.

Tabella 3.6: Tabella delle probabilità per *mix* dati *T1gt* e *T2gt*

T1gt:	A	A	A	A	A	A	AB	AB	AB	AB	AB	AB
T2gt:	A	AB	AC	B	BC	C	A	AB	AC	B	BC	C
A	1	0	0	0	0	0	0	0	0	0	0	0
AB	0	1	0	0	0	0	1	1	0	0	0	0
ABC	0	0	0	0	1	0	0	0	1	0	1	1
B	0	0	0	0	0	0	0	0	0	0	0	0
BC	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0

T1gt:	AC	AC	AC	AC	AC	AC	B	B	B	B	B	B
T2gt:	A	AB	AC	B	BC	C	A	AB	AC	B	BC	C
A	0	0	0	0	0	0	0	0	0	0	0	0
AB	0	0	0	0	0	0	1	1	0	0	0	0
ABC	0	1	0	1	1	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	0	0	1	0	0
BC	0	0	0	0	0	0	0	0	0	0	1	1
C	0	0	0	0	0	0	0	0	0	0	0	0

T1gt:	BC	BC	BC	BC	BC	BC	C	C	C	C	C	C
T2gt:	A	AB	AC	B	BC	C	A	AB	AC	B	BC	C
A	0	0	0	0	0	0	0	0	0	0	0	0
AB	0	0	0	0	0	0	0	0	0	0	0	0
ABC	1	1	1	0	0	0	0	1	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0
BC	0	0	0	1	1	1	0	0	0	1	1	1
C	0	0	0	0	0	0	0	0	0	0	0	1

3.4 Il modello finale

La composizione di tutti i nodi descritti nelle sezioni precedenti dà luogo al network bayesiano in Figura 3.7.
Il colore grigio scuro sta ad indicare un nodo "test", cioè i nodi che, data la loro natura, forniscono un supporto alle decisioni e per questo suscitano maggiore interesse.
I nodi in grigio chiaro sono quelli di cui si ha una prova o osservazione, ossia si sa quale stato assumono. Da questi nodi poi viene propagata l'informazione fino a dove possibile. Un esempio chiarirà meglio di che cosa si intende con "propagazione dell'informazione": si ipotizzi che *vgt* sia AB; allora *vpg* e *vmg* (che sono simmetrici) assumeranno A o B con probabilità 0.5 ciascuno, 0 altrimenti; nel caso in cui ci fosse la certezza circa l'identità della vittima (e quindi *T2=Yes*), allora anche i nodi *T2pg* e *T2mg* assumerebbero A o B con probabilità 0.5 ciascuno; in virtù di tutto questo, il nodo *T2gt* erediterebbe A o B da *T2pg* e A o B da *T2mg*; ma come visto nella sezione precedente, *T2gt* non potrebbe essere omozigote in quanto entrerebbe indirettamente in contrasto con il nodo *vgt*, quindi *T2gt* potrebbe essere solo eterozigote, ossia AB. Da questo semplice esempio emerge come un nodo possa influenzarne un altro a cui non è direttamente collegato.
L'evidenza può provenire da diversi nodi, ma deve essere coerente nel senso che le osservazioni non devono essere in contrasto fra di loro perché un modello possa funzionare. Un esempio banale di incoerenza è il caso in cui *vgt=AB* e *vpg=C*.
Talvolta l'informazione può essere superflua ai fini di un problema. Ad esempio, la conoscenza del nodo *sgt* rende inutile la conoscenza dei nodi *wgt* e *cgt*, in particolare per il problema di genetica forense che si desidera trattare. Questo esempio rappresenta il punto cruciale di questa tesi: quanto l'informazione data da *cgt* e da *wgt* può compensare l'assenza di informazione su *sgt*? A questo volgerà l'attenzione tutto il capitolo successivo.

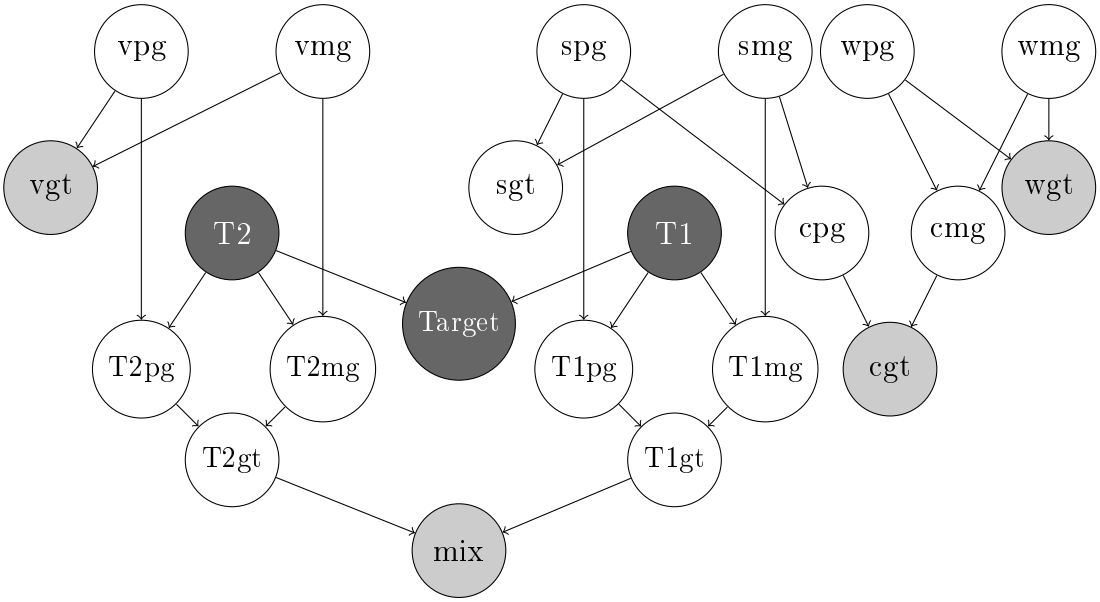

Figura 3.7: Il modello finale

Capitolo 4

Analisi

Il modello è stato applicato utilizzando sei alleli del locus TH01 per i nordamericani di razza caucasica [Budowle *e altri*(1999)] riportati nella Tabella 4.1. Sono stati scelti gli alleli con *repeat number* 6, 7, 8, 9, 9.3 e 10 in quanto sono i maggiormente frequenti (la loro probabilità somma a 1) e, per comodità, sono stati chiamati rispettivamente A, B, C, D, E e F.
L'implementazione del modello è stata effettuata utilizzando il pacchetto *gRain* di **R** [Højsgaard(2013)].
Seppur sia stato incluso il nodo *T2* ed il nodo *Target* preveda anche l'ipotesi che la vittima non sia tale, ossia che non abbia lasciato una traccia biologica, si procederà sempre con l'assunzione che il DNA della vittima (vgt) sia effettivamente nella traccia biologica rinvenuta. Le ipotesi sono quindi $H_0 = v\&s$ e $H_1 = v\&U$ e per confrontarle si usa il rapporto di verosimiglianza (LR) tra le due ipotesi [Mortera *e altri*(2003)] come nell'equazione 4.1, dove I indica l'informazione di cui si dispone.

$$LR = \frac{\Pr(I \mid H_0)}{\Pr(I \mid H_1)} \tag{4.1}$$

L'analisi è stata effettuata sia in caso di conoscenza del genotipo del sospetto (Sezione 4.1), che in caso di non conoscenza (Sezione 4.2).

Tabella 4.1: Locus TH01

Repeat number	6 (A)	7 (B)	8 (C)	9 (D)	9.3 (E)	10 (F)
Probabilità	0.2266	0.1724	0.1281	0.165	0.3054	0.0025

4.1 L'informazione include *sgt*

Essendo l'informazione completa, cioè si conosce il genotipo del sospetto, i genotipi di figlio e moglie diventano superflui. Quindi l'informazione essenziale include il genotipo della vittima, il genotipo del sospetto e gli alleli rilevati sulla scena del crimine. Questa conoscenza si traduce nei nodi *vgt*, *sgt* e *mix*. Inoltre l'assunzione di partenza sulla corrispondenza del profilo genetico della vittima con la traccia DNA implica anche *T2*=Yes.
Consideriamo il caso in cui la vittima ha genotipo AB e gli alleli rinvenuti sulla scena sono A, B, C e D: la logica afferma che il colpevole (e quindi anche il sospetto con una certa probabilità) abbia gli alleli C e D e quindi genotipo CD. Una volta propagata l'informazione circa i nodi *vgt*, *mix* e *T2*, l'output di **R** fornisce le sequenti probabilità per gli stati del nodo *sgt*:

```
> querygrain(setEvidence(pn,c("vgt","T2","mix"),c("AB","Yes",
+ "ABCD")))$sgt
sgt
          A           AB           AC           AD           AE
0.025673780 0.039065840 0.029027460 0.037389000 0.069203640
         AF            B           BC           BD           BE
0.000566500 0.014860880 0.022084440 0.028446000 0.052650960
         BF            C           CD           CE           CF
0.000431000 0.008204805 0.521136500 0.039121740 0.000320250
          D           DE           DF            E           EF
0.013612500 0.050391000 0.000412500 0.046634580 0.000763500
          F
0.000003125
```

L'output appena presentato indica la probabilità dei vari stati di *sgt* condizionatamente all'informazione sui nodi *vgt*, *T2* e *mix*. In termini più matematici, questa si traduce nella seguente formula:

Pr(sgt | cgt=AB, T2=Yes, mix=ABCD)

Lo stato CD risulta effettivamente il più probabile (circa il 52% di probabilità), mentre il secondo più probabile (AE) ha poco meno del 7% della probabilità. Nel caso in cui il genotipo del sospetto sia CD, **R** produce il seguente output per quanto riguarda i valori di H_0 e H_1:

```
> querygrain(setEvidence(pn,c("vgt","T2","mix","sgt"),c("AB",
+ "Yes","ABCD","CD")))$Target
Target
       v&s         v&U         s&U          2U
0.95944153 0.04055847 0.00000000 0.00000000
> LR
     v&s
23.65576
```

Il rapporto di verosimiglianza è chiaramente a favore dell'ipotesi nulla. Stimando invece con un genotipo diverso per *sgt* (ad esempio un caso di omozigosi), invece, si hanno i seguenti risultati:

```
> querygrain(setEvidence(pn,c("vgt","T2","mix","sgt"),c("AB",
+ "Yes","ABCD","C")))$sgt
Target
v&s v&U s&U  2U
  0   1   0   0
> LR
v&s
  0
```

Analizzando tutti i casi con AB come genotipo della vittima e gli alleli A, B, C e D nella variabile *mix*, le ipotesi H_0, H_1 e il rapporto di verosimiglianza riportati in Tabella 4.2 indicano chiaramente che l'unico genotipo possibile per il sospetto è CD e quindi il modello rifiuta certamente l'ipotesi nulla in

Tabella 4.2: H_0, H_1 e LR per ogni *sgt* dati *vgt*=AB e *mix*=ABCD

sgt:	H_0	H_1	LR
A	0	1	0
AB	0	1	0
AC	0	1	0
AD	0	1	0
AE	0	1	0
AF	0	1	0
B	0	1	0
BC	0	1	0
BD	0	1	0
BE	0	1	0
BF	0	1	0
C	0	1	0
CD	0.959	0.041	23.656
CE	0	1	0
CF	0	1	0
D	0	1	0
DE	0	1	0
DF	0	1	0
E	0	1	0
EF	0	1	0
F	0	1	0

caso il sospetto abbia un genotipo diverso. Un margine di incertezza è dato dalla possibilità che il colpevole sia un individuo con lo stesso genotipo del sospetto: infatti il modello rifiuta l'ipotesi di colpevolezza del sospetto con genotipo CD con una probabilità del 4% circa.

In generale, se in una traccia vengono rilevati quattro alleli diversi, i contributori possibili sono solo eterozigoti e c'è una relazione di biunivocità tra il genotipo della vittima e quello del sospetto. In altre parole, ad ogni genotipo eterozigote della vittima corrisponde uno ed uno solo genotipo (sempre eterozigote) del sospetto. Inoltre, ad ogni coppia di genotipi vittima-sospetto, corrisponde un particolare valore del rapporto di verosimiglianza (Tabella 4.3). I risultati evidenziano una certa prevalenza da parte dell'ipotesi nulla su quella alternativa, come era logico attendersi data l'informazione

a disposizione. Il caso con LR più elevato è quello con *vgt=AB* e *sgt=CD*. Questo risultato non è un caso in quanto gli alleli C e D sono i meno frequenti tra A, B, C e D, e quindi il risultato è da interpretare come una bassa probabilità di trovare un individuo casuale nella popolazione che abbia genotipo CD.

Tabella 4.3: H_0, H_1 e LR per ogni possibile coppia di *sgt* e *vgt* dato *mix*=ABCD

vgt	*sgt*	H_0	H_1	LR
AB	CD	0.959	0.041	23.656
AC	BD	0.946	0.054	17.577
AD	AB	0.958	0.042	22.64
BC	AD	0.93	0.07	13.373
BD	AC	0.945	0.055	17.225
CD	AB	0.928	0.072	12.799

Rifacendo l'analisi con *mix=CDEF* (Tabella 4.4), e quindi includendo anche gli alleli E ed F, rispettivamente il più frequente ed il meno frequente, i risultati confermano quanto detto finora, e cioè che ad un sospetto che ha alleli con bassa frequenza nella popolazione, corrisponde un LR più elevato.

Inoltre il rapporto di verosimiglianza in cui *sgt*=CD è 23.656, esattamente come nel caso precedente e, come si vede dalla Tabella 4.5, a parità di *sgt*, LR è uguale per tutte le combinazioni possibili di *mix* e *vgt*.

Se invece di quattro si hanno tre alleli nella traccia, i risultati cambiano leggermente. Se ad esempio nella traccia sono presenti solo gli alleli B, C e D e la vittima ha genotipo BC (Tabella 4.6), il sospetto può essere BD, CD

Tabella 4.4: H_0, H_1 e LR per ogni possibile coppia di *sgt* e *vgt* dato *mix*=CDEF

vgt	*sgt*	H_0	H_1	LR
CD	EF	0.998	0.002	654.879
CE	DF	0.999	0.001	1212.121
CF	DE	0.908	0.092	9.922
DE	CF	0.999	0.001	1561.28
DF	CE	0.927	0.073	12.781
EF	CD	0.959	0.041	23.656

Tabella 4.5: H_0, H_1 e LR per ogni possibile coppia di *mix* (con 4 alleli) e *vgt* dato *sgt*=CD

mix	*vgt*	*sgt*	H_0	H_1	LR
ABCD	AB	CD	0.959	0.041	23.656
ACDE	AE	CD	0.959	0.041	23.656
ACDF	AF	CD	0.959	0.041	23.656
BCDE	BE	CD	0.959	0.041	23.656
BCDF	BF	CD	0.959	0.041	23.656
CDEF	EF	CD	0.959	0.041	23.656

o D, cioè quelle combinazioni di alleli che hanno D e nessun allele diverso da B e C. Inoltre il rapporto di verosimiglianza è lo stesso per tutti e tre i profili genotipici del sospetto. Se invece si ha lo stesso *mix*, ma vittima omozigote come in Tabella 4.7, l'unico genotipo possibile per il sospetto è CD. Il rapporto di verosimiglianza corrispondente è 23.656, cioè lo stesso che si vede in Tabella 4.5. Il genotipo CD compare anche in Tabella 4.6, ma ha un LR diverso (7.912). Questa differenza è dovuta alla mancata relazione di esclusività di cui si è parlato nel caso con mistura a quattro alleli. Se prima (caso con quattro alleli) LR era 23.656 per ogni situazione con *sgt*=CD, ora (caso con tre alleli) ciò non avviene in quanto alla vittima BC corrisponde un sospetto che può avere diversi genotipi.

Una lista di esempi significativi riguardanti quanto detto finora è in Tabella 4.9. Emergono le seguenti considerazioni:

- LR è costante per ogni combinazione di *mix* e *vgt* che rende unico *sgt* (righe 1, 8 e 9 Tabella 4.9);

- se *sgt* non è unico, allora dati *mix* e *vgt*, LR dipende solamente dall'allele che deve essere presente (Tabella 4.9 righe 2, 3 e 4 per quanto *mix*=BCD e righe 5, 6 e 7 per *mix*=ADE).

Se una traccia ha due soli alleli allora si hanno i seguenti casi:

- la vittima è eterozigote (e quindi ha gli stessi alleli della mistura) e quindi il sospetto può essere sia eterozigote che omozigote;

Tabella 4.6: H_0, H_1 e LR per ogni *sgt* dati *vgt*=BC e *mix*=BCD

sgt:	H_0	H_1	LR
A	0	1	0
AB	0	1	0
AC	0	1	0
AD	0	1	0
AE	0	1	0
AF	0	1	0
B	0	1	0
BC	0	1	0
BD	0.888	0.112	7.912
BE	0	1	0
BF	0	1	0
C	0	1	0
CD	0.888	0.112	7.912
CE	0	1	0
CF	0	1	0
D	0.888	0.112	7.912
DE	0	1	0
DF	0	1	0
E	0	1	0
EF	0	1	0
F	0	1	0

Tabella 4.7: H_0, H_1 e LR per ogni *sgt* dati *vgt*=B e *mix*=BCD

sgt:	H_0	H_1	LR
A	0	1	0
AB	0	1	0
AC	0	1	0
AD	0	1	0
AE	0	1	0
AF	0	1	0
B	0	1	0
BC	0	1	0
BD	0	1	0
BE	0	1	0
BF	0	1	0
C	0	1	0
CD	0.959	0.041	23.656
CE	0	1	0
CF	0	1	0
D	0	1	0
DE	0	1	0
DF	0	1	0
E	0	1	0
EF	0	1	0
F	0	1	0

Tabella 4.8: H_0, H_1 e LR per ogni possibile coppia di *mix* (con 3 o 4 alleli) e *vgt* dato *sgt=CD*

mix	*vgt*	H_0	H_1	LR
ACD	A	0.959	0.041	23.656
ABCD	AB	0.959	0.041	23.656
ACDE	AE	0.959	0.041	23.656
ACDF	AF	0.959	0.041	23.656
BCD	B	0.959	0.041	23.656
BCDE	BE	0.959	0.041	23.656
BCDF	BF	0.959	0.041	23.656
CDE	E	0.959	0.041	23.656
CDF	F	0.959	0.041	23.656
CDEF	EF	0.959	0.041	23.656

Tabella 4.9: Esempi riassuntivi

mix	*vgt*	*sgt*	H_0	H_1	LR
BCD	B	CD	0.959	0.041	23.656
BCD	BC	BD	0.888	0.112	7.912
BCD	BC	CD	0.888	0.112	7.912
BCD	BC	D	0.888	0.112	7.912
ADE	AE	AD	0.791	0.209	3.78
ADE	AE	D	0.791	0.209	3.78
ADE	AE	DE	0.791	0.209	3.78
ABCD	AB	CD	0.959	0.041	23.656
BCDE	BE	CD	0.959	0.041	23.656

Tabella 4.10: H_0, H_1 e LR per ogni possibile coppia *sgt* e *vgt* dato *mix=CD*

mix	*vgt*	*sgt*	H_0	H_1	LR
CD	C	CD	0.935	0.065	14.3889
CD	C	D	0.935	0.065	14.3889
CD	CD	C	0.921	0.079	11.64041
CD	CD	CD	0.921	0.079	11.64041
CD	CD	D	0.921	0.079	11.64041
CD	D	C	0.945	0.055	17.041
CD	D	CD	0.945	0.055	17.041

- la vittima è omozigote e quindi il sospetto può essere sia omozigote che eterozigote, ma deve possedere l'allele presente nella mistura che la vittima non ha.

Dalla Tabella 4.10 si vede la conferma di quanto detto prima, e cioè che LR dipende dall'allele che il sospetto deve possedere per poter essere considerato un contributore. Inoltre il caso in cui la vittima è eterozigote evidenzia un solo LR per ogni stato (possibile) di *sgt*.

Il caso in cui la traccia evidenzia un solo allele è banale in quanto sospetto e vittima possono essere solamente omozigoti ed avere lo stesso allele presente nella traccia (Tabella 4.11). L'allele con LR maggiore è F che è quello meno frequente nella popolazione. I rapporti di verosimiglianza sono ordinati con ordine inverso rispetto alle loro probabilità.

Tabella 4.11: H_0, H_1 e LR con un solo allele in *mix*

mix	H_0	H_1	LR
A	0.951	0.049	19.457
B	0.971	0.029	33.645
C	0.984	0.016	60.94
D	0.973	0.016	36.731
E	0.915	0.085	10.722
F	1	0	∞

4.2 Nessuna informazione su *sgt*

In caso di assenza di informazione circa il genotipo del sospetto, diventano molto importanti quelli di moglie e figlio. Essendo il genotipo umano una coppia non ordinata di alleli, non è possibile risalire con certezza a quale sia l'allele che proviene dal padre utilizzando solamente il profilo genetico del figlio. Per questa ragione, quello della madre permette di intuire o almeno avere un'idea di quale dei due alleli il figlio ha ereditato dal padre (Figura 4.1).

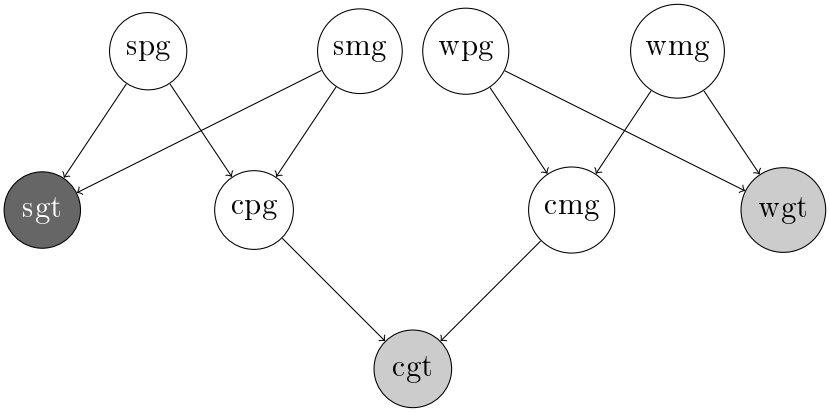

Figura 4.1: Il genotipo del figlio

Se si intuisce quale dei due alleli proviene dalla madre, si sa anche quale proviene dal padre. Se per esempio la madre ha genotipo AB ed il figlio ha genotipo AC, allora è logico che A è l'allele che proviene dalla madre e C dal padre. Il problema avviene quando la madre ed il figlio hanno lo stesso genotipo, in tal caso non si ha certezza sull'ereditarietà degli alleli.
In caso di genotipi diversi, il modello implementato con **R** fornisce il seguente

output per quanto riguarda l'allele del figlio proveniente dal sospetto ed il genotipo del sospetto:

```
> querygrain(setEvidence(pn,c("cgt","wgt"),c("AC","AB")))$cpg
cpg
A B C D E F
0 0 1 0 0 0

> querygrain(setEvidence(pn,c("cgt","wgt"),c("AC","AB")))$sgt
sgt
     A      AB      AC      AD      AE      AF       B      BC
0.0000 0.0000 0.2266 0.0000 0.0000 0.0000 0.0000 0.1724
    BD      BE      BF       C      CD      CE      CF       D
0.0000 0.0000 0.0000 0.1281 0.1650 0.3054 0.0025 0.0000
    DE      DF       E      EF       F
0.0000 0.0000 0.0000 0.0000 0.0000
```

L'output indica C come unico possibile allele ereditato dal padre dati i genotipi di moglie e figlio. I genotipi possibili evidenziati sono tutti quelli che contengono l'allele C e le loro probabilità sono uguali alle frequenze alleliche nella popolazione, cioè la probabilità di sgt=CD è 0.165 che è uguale alla probabilità marginale che un allele sia D, la probabilità di sgt=AC è 0.2266, come la probabilità marginale che un allele sia A e così via.
In caso di genotipi identici tra madre e figlio, **R** fornisce il seguente output:

```
> querygrain(setEvidence(pn,c("cgt","wgt"),c("AB","AB")))$cpg
cpg
        A         B         C         D         E         F
0.5679198 0.4320802 0.0000000 0.0000000 0.0000000 0.0000000

> querygrain(setEvidence(pn,c("cgt","wgt"),c("AB","AB")))$sgt
sgt
          A           AB           AC           AD           AE
0.128690627 0.195818747 0.072750526 0.093706767 0.173442707
```

```
         AF           B          BC          BD          BE
0.001419799 0.074490627 0.055349474 0.071293233 0.131957293
         BF           C          CD          CE          CF
0.001080201 0.000000000 0.000000000 0.000000000 0.000000000
          D          DE          DF           E          EF
0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
          F
0.000000000
```

L'allele che proviene dal padre è A oppure B con probabilità proporzionale alle frequenze degli alleli stessi nella popolazione. Infatti

$$\begin{aligned}\Pr(\text{cpg=A} \mid \text{cgt=AB, wgt=AB}) &= \frac{\Pr(A)}{\Pr(A)+\Pr(B)} \\ &= \frac{0.2266}{0.2266+0.1724} \\ &= 0.5679\end{aligned}$$

$$\begin{aligned}\Pr(\text{cpg=B} \mid \text{cgt=AB, wgt=AB}) &= \frac{\Pr(B)}{\Pr(A)+\Pr(B)} \\ &= \frac{0.1724}{0.2266+0.1724} \\ &= 0.4321\end{aligned}$$

I genotipi possibili in questo caso sono tutti quelli che hanno almeno uno tra gli alleli A e B. La probabilità di ogni stato di *sgt* è proporzionale alla frequenza dei singoli alleli nella popolazione e al risultato di *cpg*.

$$\begin{aligned}\Pr(\text{sgt=A} \mid \text{cgt=AB, wgt=AB}) &= \Pr(\text{cpg=A} \mid \text{cgt=AB, wgt=AB}) \cdot \Pr(A) \\ &= 0.5679 \cdot 0.2266 \\ &= 0.1287\end{aligned}$$

$$\Pr(\text{sgt=BC} \mid \text{cgt=AB, wgt=AB}) = \Pr(\text{cpg=B} \mid \text{cgt=AB, wgt=AB}) \cdot \Pr(C)$$

$$\begin{aligned} &= 0.4321 \cdot 0.1281 \\ &= 0.0553 \end{aligned}$$

Per quanto riguarda invece il genotipo AB, entrambi i gli alleli possono essere trasmessi da qualunque dei genitori, e quindi:

$$\begin{aligned} \Pr(\text{sgt=AB} \mid \text{cgt=AB, wgt=AB}) &= \Pr(\text{cpg=A} \mid \text{cgt=AB, wgt=AB}) \cdot \Pr(B) \\ &+ \Pr(\text{cpg=B} \mid \text{cgt=AB, wgt=AB}) \cdot \Pr(A) \\ &= 0.5679 \cdot 0.1724 + 0.4321 \cdot 0.2266 \\ &= 0.1958 \end{aligned}$$

Naturalmente se figlio e moglie hanno lo stesso genotipo e questo è omozigote, si può dedurre con certezza *cpg* e quindi si ritorna al caso precedente.

Dal punto di vista del network, abbiamo appena visto che il genotipo del sospetto può ricevere informazioni dai nodi *cgt* e *wgt*. Tuttavia dell'informazione può arrivare anche da un'altra direzione, e cioè dal nodo *mix*. Se ad esempio al caso in cui moglie e figlio hanno genotipi diversi si aggiunge la prova della traccia di DNA, si ottengono le seguenti probabilità per *sgt*:

```
> querygrain(setEvidence(pn,c("cgt","wgt","mix"),c("AC","AB",
+ "CE")))$sgt
sgt
           A           AB           AC           AD           AE
0.0000000000 0.0000000000 0.0564768255 0.0000000000 0.0000000000
          AF            B           BC           BD           BE
0.0000000000 0.0000000000 0.0429682467 0.0000000000 0.0000000000
          BF            C           CD           CE           CF
0.0000000000 0.2397699108 0.0411239020 0.6190380255 0.0006230894
           D           DE           DF            E           EF
0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000
           F
```

```
0.0000000000
```

I genotipi possibili sono gli stessi del caso precedente, come si vede dalla Tabella 4.12, ma cambiano le probabilità: quelle di C e CE aumentano, mentre diminuiscono tutte le altre. Essendo gli alleli presenti nel *mix* C ed E, aumentano le possibilità che il genotipo contenga questi alleli.

Tabella 4.12: *mix* come informazione aggiuntiva

sgt	Senza *mix*=CE	Con *mix*=CE
A	0	0
AB	0	0
AC	0.2266	0.0565
AD	0	0
AE	0	0
AF	0	0
B	0	0
BC	0.1724	0.043
BD	0	0
BE	0	0
BF	0	0
C	0.1281	0.2398
CD	0.165	0.0411
CE	03054	0.619
CF	0.0025	0.0006
D	0	0
DE	0	0
DF	0	0
E	0	0
EF	0	0
F	0	0

Va ribadito che *sgt* non è il genotipo del colpevole, bensì quello del sospetto; pertanto le probabilità di genotipi come AC o BC non si annullano, in quanto il modello prevede ancora la possibilità che il sospetto ed il colpevole siano due individui differenti. Togliendo questa possibilità, i soli genotipi possibili diventano C e CE come si vede dal seguente output:

```
> querygrain(setEvidence(pn,c("cgt","wgt","mix","T1"),c("AC",
```

```
+ "AB","CE","Yes")))$sgt
sgt
        A         AB         AC         AD         AE         AF
0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
        B         BC         BD         BE         BF          C
0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.2768417
       CD         CE         CF          D         DE         DF
0.0000000 0.7231583 0.0000000 0.0000000 0.0000000 0.0000000
        E         EF          F
0.0000000 0.0000000 0.0000000
```

Sull'entità del cambiamento nelle probabilità influisce certamente la frequenza degli alleli nella popolazione ed altri fattori quali la presenza di un altro contributore (la vittima).
Se invece si annulla la possibilità che il sospetto sia il colpevole (*T1* =No), allora si annulla l'informazione proveniente dal *mix* e si torna al caso in cui le uniche informazioni utili ad intuire il genotipo del sospetto si conoscono solamente i genotipi di moglie e figlio.

```
> querygrain(setEvidence(pn,c("cgt","wgt","mix","T1"),c("AC",
+ "AB","CE","No")))$sgt
sgt
     A      AB      AC      AD      AE      AF       B      BC      BD
0.0000 0.0000 0.2266 0.0000 0.0000 0.0000 0.0000 0.1724 0.0000
    BE      BF       C      CD      CE      CF       D      DE      DF
0.0000 0.0000 0.1281 0.1650 0.3054 0.0025 0.0000 0.0000 0.0000
     E      EF       F
0.0000 0.0000 0.0000
```

Aggiungendo ulteriore informazione come ad esempio il genotipo della vittima, aumenta la precisione nell'intuizione degli alleli presenti nella mistura che provengono dal colpevole. Ad esempio il seguente output si riferisce ad una mistura CE ed una potenziale vittima (non c'è alcuna informazione sul nodo *T2*) con genotipo E. *T1gt*, ossia il genotipo del colpevole può essere solamente una combinazione degli alleli E e C.

```
> querygrain(setEvidence(pn,c("vgt","mix"),c("E","CE")))$T1gt
T1gt
         A          AB          AC          AD          AE
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        AF           B          BC          BD          BE
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        BF           C          CD          CE          CF
0.00000000 0.15887730 0.00000000 0.76816192 0.00000000
         D          DE          DF           E          EF
0.00000000 0.00000000 0.00000000 0.07296078 0.00000000
         F
0.00000000
```

Essendo il sospetto ed il colpevole due individui potenzialmente diversi, allora *sgt* può assumere ancora tutti gli stati.

```
> querygrain(setEvidence(pn,c("vgt","mix"),c("E","CE")))$sgt
sgt
          A           AB           AC           AD           AE
0.025673780 0.039065840 0.029027460 0.037389000 0.069203640
         AF            B           BC           BD           BE
0.000566500 0.014860880 0.022084440 0.028446000 0.052650960
         BF            C           CD           CE           CF
0.000431000 0.087643453 0.021136500 0.423202702 0.000320250
          D           DE           DF            E           EF
0.013612500 0.050391000 0.000412500 0.083114970 0.000763500
          F
0.000003125
```

In particolare gli alleli A, B, D ed F possono essere presenti nel genotipo del sospetto solo se questo non è il colpevole. Il seguente output indica le probabilità marginali di ogni genotipo, ossia le frequenze di ogni genotipo per un individuo casuale nella popolazione, e aiuterà a fare qualche considerazione.

```
> querygrain(setEvidence(pn))$sgt
sgt
         A          AB          AC          AD          AE
0.05134756 0.07813168 0.05805492 0.07477800 0.13840728
        AF           B          BC          BD          BE
0.00113300 0.02972176 0.04416888 0.05689200 0.10530192
        BF           C          CD          CE          CF
0.00086200 0.01640961 0.04227300 0.07824348 0.00064050
         D          DE          DF           E          EF
0.02722500 0.10078200 0.00082500 0.09326916 0.00152700
         F
0.00000625
```

Come si può vedere dalla Tabella 4.13 che contiene i risultati dei due precedenti output, tutti i genotipi che contengono un allele diverso da C ed E dimezzano rispetto alle probabilità marginali. Ciò avviene in quanto, in caso di coincidenza tra sospetto e colpevole (che ha probabilità del 50% in quanto *mix* e *vgt* non influenzano in alcun modo il nodo *T1*), allora è impossibile che *sgt* assuma questi genotipi, mentre in caso contrario il sospetto non c'entra col crimine e quindi il suo genotipo avrà legge come la marginale in Tabella 4.13. Le probabilità condizionali di C, CE ed E cambiano diversamente in quanto considerano i casi in cui il sospetto sia il colpevole o meno e che la vittima contribuisca o meno alla mistura di DNA. In particolare:

- se il sospetto è il colpevole e la vittima ha contribuito alla traccia, allora il sospetto avrà certamente l'allele E;
- se il sospetto è il colpevole e la vittima non ha contribuito alla traccia, allora il sospetto avrà certamente almeno uno tra gli alleli C ed E;
- se il sospetto ed il colpevole sono due persone diverse, allora il suo genotipo seguirà la distribuzione marginale in Tabella 4.13.

Infatti, secondo le regole di probabilità condizionata e, considerando $I = \{\text{vgt=C, mix=CE}\}$, si ha che:

$$\Pr(\text{sgt=CE} \mid \text{I}) \;=\; \Pr(\text{sgt=CE} \mid \text{I, T2}) \cdot \Pr(\text{T2} \mid \text{I})$$

Pr(T2 | I) è data dal seguente output:

```
> querygrain(setEvidence(pn,c("vgt","mix"),c("C","CE")))$T2
T2
      Yes        No
0.8668427 0.1331573
```

Pr(sgt=CE | I, T2) che è la probabilità di *sgt* condizionata allo stato di *T2* è la seguente:

```
> querygrain(setEvidence(pn,c("vgt","mix","T2"),c("C","CE",
+ "Yes")))$sgt
sgt
          A           AB           AC           AD           AE
0.025673780 0.039065840 0.029027460 0.037389000 0.069203640
         AF            B           BC           BD           BE
0.000566500 0.014860880 0.022084440 0.028446000 0.052650960
         BF            C           CD           CE           CF
0.000431000 0.008204805 0.021136500 0.267220031 0.000320250
          D           DE           DF            E           EF
0.013612500 0.050391000 0.000412500 0.318536289 0.000763500
          F
0.000003125

> querygrain(setEvidence(pn,c("vgt","mix","T2"),c("C","CE",
+ "No")))$sgt
sgt
          A           AB           AC           AD           AE
0.025673780 0.039065840 0.029027460 0.037389000 0.069203640
         AF            B           BC           BD           BE
```

```
0.000566500 0.014860880 0.022084440 0.028446000 0.052650960
         BF           C          CD          CE          CF
0.000431000 0.061617411 0.021136500 0.318167698 0.000320250
          D          DE          DF           E          EF
0.013612500 0.050391000 0.000412500 0.214176017 0.000763500
          F
0.000003125
```

Di conseguenza

$$\begin{aligned}
\Pr(\text{sgt=CE} \mid \text{I}) &= \Pr(\text{sgt=CE} \mid \text{I, T2}) \cdot \Pr(\text{T2} \mid \text{I}) \\
&= \Pr(\text{sgt=CE} \mid \text{I, T2=Yes}) \cdot \Pr(\text{T2=Yes} \mid \text{I}) \\
&+ \Pr(\text{sgt=CE} \mid \text{I, T2=No}) \cdot \Pr(\text{T2=No} \mid \text{I}) \\
&= 0.26722 \cdot 0.86684 + 0.31817 \cdot 0.13316 \\
&= 0.27400
\end{aligned}$$

Lo stesso vale per i genotipi C ed E.

In questa sezione ho finora considerato l'influenza dell'informazione sul genotipo del sospetto proveniente da due fonti diverse, ossia quella che arriva dai genotipi di figlio e moglie, e quella che arriva dalla mistura di DNA. Il passo successivo è quello di unire le due "correnti" di informazione e fare delle opportune analisi. Va inoltre detto che finora ci si è limitati a concentrare l'analisi sul nodo relativo al genotipo del sospetto, tuttavia non va dimenticato lo scopo dello studio, ossia di fornire un modello che verifichi che l'individuo sospettato sia o meno il colpevole di un crimine. La ragione degli output relativi al nodo *sgt* mostrati fino a questo momento è quello di far vedere come l'informazione possa aiutare a fornire una cerchia più ristretta di genotipi possibili (o almeno con una disceta probabilità). L'interesse principale va comunque alla distribuzione a posteriori del nodo *Target*.
In questa sezione si è parlato della vittima come contributrice potenziale e non certa, in modo da poter entrare più facilmente nella logica del network bayesiano. D'ora in avanti, invece, si considererà *T2*=Yes.
L'informazione ora comprende i seguenti nodi: *mix*, *vgt*, *T2*, *cgt* e *wgt*. Al-

Tabella 4.13: Probabilità condizionate a *mix*=CE e *vgt*=E e marginali

sgt	Condizionali	Marginali
A	0.02567	0.05135
AB	0.03907	0.07813
AC	0.02923	0.05805
AD	0.03739	0.07478
AE	0.0692	0.13841
AF	0.00057	0.00113
B	0.01486	0.02972
BC	0.02208	0.04417
BD	0.02845	0.05689
BE	0.05265	0.1053
BF	0.00043	0.00086
C	0.01532	0.01641
CD	0.02114	0.04227
CE	0.274	0.07824
CF	0.00032	0.00064
D	0.01361	0.02723
DE	0.05039	0.10078
DF	0.00041	0.00083
E	0.30463	0.09327
EF	0.00076	0.00153
F	0	0.00001

l'inizio di questa sezione si è parlato del problema generato da moglie e figlio aventi lo stesso genotipo. Si consideri il caso in cui *mix*=ABCD, *vgt*=AB, *cgt*=DE e *wgt*=EF. Da una parte si intuisce facilmente che il colpevole ha genotipo CD, dall'altra si capisce che il sospetto ha l'allele D nel suo profilo. Infatti, dall'output che segue, si vede che gli unici stati possibile per *sgt* sono i genotipi che contengono l'allele D e che CD, grazie all'informazione che arriva dalla mistura e dalla vittima, ha probabilità che supera il 78%.

```
> querygrain(setEvidence(pn,c("T2","vgt","mix","cgt","wgt"),
+ c("Yes","AB","ABCD","DE","EF")))$sgt
sgt
           A            AB            AC            AD            AE
0.0000000000 0.0000000000 0.0000000000 0.0562240602 0.0000000000
```

```
          AF           B           BC           BD           BE
0.0000000000 0.0000000000 0.0000000000 0.0427759398 0.0000000000
          BF           C           CD           CE           CF
0.0000000000 0.0000000000 0.7836639098 0.0000000000 0.0000000000
           D           DE           DF            E           EF
0.0409398496 0.0757759398 0.0006203008 0.0000000000 0.0000000000
           F
0.0000000000
```

I test risultanti e LR danno i seguenti risultati:

```
> querygrain(setEvidence(pn,c("T2","vgt","mix","cgt","wgt"),
+ c("Yes","AB","ABCD","DE","EF")))$Target
Target
      v&s       v&U       s&U       2U
0.7518797 0.2481203 0.0000000 0.0000000
> LR
[1] 3.030303
```

Se il genotipo della moglie cambia variando l'allele che non trasmette al figlio, naturalmente LR non cambia. Lo stesso vale anche per il figlio, se a variare è l'allele che riceve dalla madre (e contestualmente l'allele della madre).

```
> querygrain(setEvidence(pn,c("T2","vgt","mix","cgt","wgt"),
+ c("Yes","AB","ABCD","DE","CE")))$Target
Target
      v&s       v&U       s&U       2U
0.7518797 0.2481203 0.0000000 0.0000000

> querygrain(setEvidence(pn,c("T2","vgt","mix","cgt","wgt"),
+ c("Yes","AB","ABCD","AD","AE")))$Target
Target
      v&s       v&U       s&U       2U
0.7518797 0.2481203 0.0000000 0.0000000
```

Stesso risultato si ha inoltre nei casi in cui cambia la mistura, ma rimane inalterato (e presente nella mistura) l'allele che il figlio eredita dal sospetto. Alcuni esempi sono presenti nella Tabella 4.14: dalla prima alla seconda riga cambia la mistura, dalla prima alla terza cambiano sia la mistura che (il dedotto) genotipo del colpevole, LR in nessun caso cambia.

Tabella 4.14: LR dipende dall'allele che il figlio eredita dal sospetto

mix	*vgt*	*cgt*	*wgt*	$\Pr(H_0)$	LR
ABCD	AB	DE	EF	0.7519	3.0303
CDEF	EF	DE	EF	0.7519	3.0303
ABDF	AB	DE	EF	0.7519	3.0303

All'inizio di questa sezione si è parlato del problema generato da figlio e moglie del sospetto quando questi hanno lo stesso genotipo: si vedranno ora alcuni esempi per vedere l'effetto che ha sul nodo *Target*. Si prenda ad esempio *mix*=ABCD, *vgt*=AB, *cgt*=*wgt*=DE. La probabilità condizionata circa i vari genotipi del sospetto è data dal seguente output:

```
> querygrain(setEvidence(pn,c("T2","vgt","mix","cgt","wgt"),
+ c("Yes","AB","ABCD","DE","DE")))$sgt
sgt
           A            AB            AC            AD            AE
0.0000000000 0.0000000000 0.0000000000 0.0385294724 0.0713145507
          AF             B            BC            BD            BE
0.0000000000 0.0000000000 0.0000000000 0.0293136851 0.0542569662
          BF             C            CD            CE            CF
0.0000000000 0.0000000000 0.5370326669 0.0403150660 0.0000000000
           D            DE            DF             E            EF
0.0280554411 0.1038561418 0.0004250824 0.0961141385 0.0007867890
           F
0.0000000000
```

Da una parte il modello intuisce che il genotipo più probabile per il sospetto è CD (che è l'unico possibile per *T1gt*, ossia il colpevole), dall'altra capisce che il sospetto ha un allele tra D ed E, anche se non sa quale. Infatti dal

prossimo output, che si riferisce alle probabilità condizionate di *cpg*, ossia dell'allele che il figlio eredita dal sospetto, si vede che *cpg* è D oppure E:

```
> querygrain(setEvidence(pn,c("T2","vgt","mix","cgt","wgt"),
+ c("Yes","AB","ABCD","DE","DE")))$cpg
cpg
        A         B         C         D         E         F
0.0000000 0.0000000 0.0000000 0.6852844 0.3147156 0.0000000
```

La probabilità che *cpg* sia D è maggiore rispetto a quella di E e tiene conto dell'informazione che arriva dal nodo *mix* e della distribuzione a priori degli alleli.
Per quanto riguarda il nodo *Target*, invece:

```
> querygrain(setEvidence(pn,c("T2","vgt","mix","cgt","wgt"),
+ c("Yes","AB","ABCD","DE","DE")))$Target
Target
      v&s       v&U       s&U        2U
0.5152514 0.4847486 0.0000000 0.0000000

> LR
[1] 1.062925
```

LR è decisamente inferiore rispetto al solito a causa dei genotipi identici di *cgt* e *wgt*. Mentre nel caso con moglie e figlio con genotipi differenti LR è lo stesso a parità di *cpg*, in questo caso l'informazione non permette di dedurre con certezza lo stato di *cpg* e quindi LR dipende da entrambi gli alleli di *cgt*. Nella Tabella 4.15 sono riportati alcuni esempi di come, con *mix*=ABCD e *vgt*=AB e al variare di *cgt*=*wgt*, cambia LR. Se *cgt* = *wgt*∈{AC, BC, CE, CF}, allora LR cambia proporzionalemente alla frequenza degli alleli A, B, E ed F. Infatti LR di *cgt*=*wgt*=CF è il maggiore in quanto F è il meno frequente. Coerentemente, LR per *cgt*=*wgt*=CF è il minore essendo E l'allele più frequente.
Discorso diverso per *cgt* = *wgt*∈{C, CD}. Nel caso di omozigosi ci si riconduce al caso in cui si riesce a dedurre con certezza lo stato di *cpg* e quindi

LR è più alto. Nel caso invece di *cgt*=*wgt*=CD, LR è più alto in quanto gli alleli di *cpg* sono gli stessi che si deducono dalla mistura. Ricapitolando:

- se *cgt* = *wgt*∈{AC, BC, CE, CF}, allora *cpg* ha come stati possibili un allele che appartiene alla scena del crimine ed uno che non ne appartiene;
- se *cgt*=*wgt*=CD, allora *cpg* ha come stati possibili due alleli ed entrambi appartengono alla scena del crimine;
- se *cgt*=*wgt*=C, allora *cpg* ha come unico stato possibile C e questo appartiene alla scena del crimine.

Tabella 4.15: LR cambia al variare dei genotipi di moglie e figlio

mix	*vgt*	*cgt*	*wgt*	$\Pr(H_0)$	LR
ABCD	AB	AC	AC	0.585	1.4097
ABCD	AB	BC	BC	0.6246	1.6639
ABCD	AB	C	C	0.7961	3.9032
ABCD	AB	CD	CD	0.7733	3.4118
ABCD	AB	CE	CE	0.5356	1.1534
ABCD	AB	CF	CF	0.7929	3.8283

Passando al caso in cui una mistura ha tre alleli diversi, è opportuno distinguere due casi:

- la vittima è omozigote, in tal caso le considerazioni sono analoghe al caso della mistura con quattro alleli;
- la vittima è eterozigote, e quindi il colpevole può avere tre genotipi diversi.

Per quanto concerne il secondo caso, si prenda come esempio *mix*=BCD e *vgt*=BC: il colpevole ha certamente un genotipo tra BD, CD e D. Perché possa essere presa in considerazione l'ipotesi che il sospetto possa essere l'effettivo colpevole, *cpg* deve assumere uno tra i seguenti stati: B, C e D. Di conseguenza possono essere presi in considerazione solamente le situazioni in cui da figlio e moglie, si ottiene che almeno uno tra gli stati B, C e D

abbia una probabilità non nulla di essere assunto da *cpg*. Dunque se moglie e figlio hanno genotipi diversi, allora l'allele del figlio che si deduce essere stato trasmesso dal padre deve essere uno tra B, C e D. Se invece moglie e figlio hanno lo stesso genotipo, allora almeno uno dei due alleli deve essere uno tra B, C e D. I seguenti output dimostreranno quanto detto finora.

```
> querygrain(setEvidence(pn,c("T2","vgt","mix","cgt","wgt"),
+ c("Yes","BC","BCD","CF","EF")))$Target
Target
      v&s       v&U       s&U        2U
0.5662514 0.4337486 0.0000000 0.0000000

> LR
[1] 1.370931

> querygrain(setEvidence(pn,c("T2","vgt","mix","cgt","wgt"),
+ c("Yes","BC","BCD","CF","CF")))$Target
Target
      v&s       v&U       s&U        2U
0.5614983 0.4385017 0.0000000 0.0000000

> LR
[1] 1.280493
```

Da un punto di vista quantitativo, LR è più alto se l'allele dedotto di *cpg* è quello che non appartiene alla vittima, D nell'esempio precedente. La Tabella 4.16 dà una dimostrazione di quanto appena detto, infatti LR nel caso in cui *cpg*=D con probabilità 1 (terza riga), è maggiore rispetto agli altri.

Nel caso in cui *cgt*=*wgt*, LR è generalmente più basso, in quanto non c'è certezza sull'allele che il figlio eredita dal sospetto. Vanno distinti quattro casi:

- degli allele di *cgt*, solo uno appartiene alla mistura;

Tabella 4.16: LR cambia al variare dei genotipi di moglie e figlio

mix	*vgt*	*cgt*	*wgt*	$\Pr(H_0)$	LR
BCD	BC	BF	EF	0.5663	1.3055
BCD	BC	CF	EF	0.5663	1.3055
BCD	BC	DF	EF	0.7865	3.683

- degli allele di *cgt*, solo uno appartiene alla mistura e questo è quello che deve appartenere al colpevole;
- entrambi gli alleli di *cgt* appartengono alla mistura e al genotipo della vittima (quindi *cgt*=*wgt*=*vgt*);
- entrambi gli alleli di *cgt* appartengono alla mistura e uno di questi appartiene certamente al genotipo del colpevole.

Da un caso all'altro LR è sensibilmente diverso, come si vede dalla Tabella 4.17.

Tabella 4.17: LR nel caso in cui la mistura ha tre alleli e *cgt*=*wgt*

mix	*vgt*	*cgt*	*wgt*	$\Pr(H_0)$	LR
BCD	BC	AB	AB	0.3606	0.5641
BCD	BC	AD	AD	0.6081	1.5518
BCD	BC	BC	BC	0.5663	1.3055
BCD	BC	BD	BD	0.7117	2.4682

I casi in cui LR è più grande sono quelli in cui *cpg* assume l'allele che deve appartenere al sospetto con probabilità non nulla.

Se la mistura contiene solo due alleli, allora bisogna nuovamente distinguere i casi in cui la vittima è omozigote o eterozigote e i casi in cui moglie e figlio hanno lo stesso genotipo o un genotipo differente.
Se il genotipo è differente e la vittima è eterozigote (e quindi *vgt*=*mix*), allora l'allele del figlio che si deduce essere ereditato dal padre deve appartenere alla mistura. LR è lo stesso qualsiasi sia l'unico stato possibile di *cpg* (Tabella 4.18).

Tabella 4.18: LR nel caso in cui sia *mix* che *vgt* hanno due alleli

mix	*vgt*	*cgt*	*wgt*	$\Pr(H_0)$	LR
BC	BC	CD	BD	0.7689	3.3278
BC	BC	BD	AD	0.7689	3.3278

Se la vittima è omozigote e moglie e figlio hanno genotipi differenti, allora l'allele del figlio che si deduce essere ereditato dal sospetto deve appartenere alla mistura. LR stavolta cambia: se l'allele ereditato dal padre è quello che deve necessariamente appartenere al colpevole, allora questi LR è maggiore (Tabella 4.19).

Tabella 4.19: LR nel caso in cui *mix* ha due alleli e *vgt* ha un allele

mix	*vgt*	*cgt*	*wgt*	$\Pr(H_0)$	LR
BC	B	BD	AD	0.6789	2.1146
BC	B	CD	AD	0.8322	4.9605

Se la vittima è eterozigote e moglie e figlio hanno lo stesso genotipo, allora almeno uno degli alleli di *cgt* deve appartenere alla mistura. Se entrambi ne appartengono, allora LR aumenta (Tabella 4.20).

Tabella 4.20: LR nel caso in cui sia *mix* che *vgt* hanno due alleli e *cgt*=*wgt*

mix	*vgt*	*cgt*	*wgt*	$\Pr(H_0)$	LR
BC	BC	BD	BD	0.6297	1.7004
BC	BC	CD	CD	0.5926	1.4544
BC	BC	BC	BC	0.7689	3.3278

Infine, se la vittima è omozigote e moglie e figlio hanno lo stesso genotipo, allora almeno uno dei due alleli di *cgt* deve appartenere alla mistura. Il caso con LR maggiore è quello in cui entrambi gli alleli del figlio appartengono alla mistura, mentre il caso con LR minore è quello in cui l'unico allele di *cgt* che appartiene alla mistura è l'unico contenuto in *vgt* (Tabella 4.21).

Se in una mistura c'è un solo allele, allora per poter prendere in considerazione l'ipotesi di colpevolezza del sospetto, questo deve essere omozigote (con lo stesso allele della mistura) e il figlio deve ereditare quell'allele. Se

Tabella 4.21: LR nel caso in cui *mix* ha due alleli, *vgt* ne ha uno e *cgt*=*wgt*

mix	*vgt*	*cgt*	*wgt*	$\Pr(H_0)$	LR
BC	B	AB	AB	0.3606	0.9137
BC	B	AC	AC	0.6081	1.7915
BC	B	BC	BC	0.5663	3.3278

l'allele viene dedotto con certezza, allora LR avrà un certo valore, altrimenti sarà minore (Tabella 4.22).

Tabella 4.22: LR nel caso in cui *mix* ha un allele

mix	*vgt*	*cgt*	*wgt*	$\Pr(H_0)$	LR
B	B	AB	AD	0.8529	5.8005
B	B	AB	AB	0.7148	2.5063

Capitolo 5

Applicazione ad un caso particolare

Per l'esempio di questo capitolo capitolo si è preso spunto da [Clayton *e altri*(1998)]. I dati disponibili sono quelli di una mistura di DNA e dei genotipi di due individui, una vittima ed un sospetto, per sette diversi loci di cui uno per il sesso (l'*amelogenina*). Ne sono stati usati solo cinque e sono: *D8S1179*, *D18S51*, *FGA*, *TH01* e *vWA*. I dati *Clayton* riportano anche le aree per tutti i picchi evidenziati nell'elettroferogramma. Poiché in questa tesi gli alleli presenti nella mistura saranno trattati da un punto di vista qualitativo, si prenderanno queste informazioni come indicazioni di presenza dell'allele nella mistura. Le frequenze nella popolazione dei vari alleli sono state prese da [Kupferschmid *e altri*(1999)], [Budowle *e altri*(1999)], [Corp.(2002)]. Per quanto riguarda il locus TH01 sono stati scelti alleli diversi da quelli del capitolo precedente. Essendo il network bayesiano creato per un numero massimo di sei alleli, sono stati selezionati i sei più frequenti (Tabella 5.1) e le loro frequenze sono poi state normalizzate a 1. Poiché tra i dati a disposizione c'è anche il genotipo del sospetto, allora è possibile fare un analisi con informazione completa. Comunque per poter confrontare questo caso con quello ad informazione incompleta, occorre disporre dei genotipi del figlio o della moglie della vittima e quindi, dato che questi non sono a disposizione, sono stati creati (assumendo che il sospetto sia effettivamente coniugato ed abbia

Tabella 5.1: Frequenze alleliche per locus

locus	A	B	C	D	E	F
repeat number	10	11	12	13	14	15
D8S1179	0.1020	0.0587	0.1454	0.3393	0.2015	0.1097
repeat number	12	13	14	15	16	18
D18S51	0.1440	0.1170	0.1680	0.1480	0.1140	0.084
repeat number	20	21	22	23	24	25
FGA	0.0894	0.1921	0.1854	0.1722	0.1225	0.1093
repeat number	5	6	7	8	9	10
TH01	0.0072	0.2392	0.1411	0.1196	0.1531	0.3349
repeat number	14	15	16	17	18	19
vWA	0.132	0.079	0.228	0.232	0.225	0.089

un figlio naturale): per ogni locus, dapprima si è creato il genotipo del figlio partendo da uno dei due alleli del sospetto, poi si è creato il genotipo della moglie partendo dall'allele del figlio non ereditato dal sospetto. Per esempio, per quanto riguarda il locus *vWA*, il sospetto ha genotipo BC e quindi i possibili genotipi di figlio e moglie sono riportati in Tabella 5.2. Per il sistema di ipotesi $H_0 : s\&v$ contro $H_1 : v\&U$, si calcoleranno i rapporti di verosimiglianza sia nel caso di informazione completa (e cioè col genotipo del sospetto) che nel caso di incompletezza (cioè senza il genotipo del sospetto, ma con quelli di figlio e moglie). I dati relativi a mistura e genotipi sono nella Tabella 5.3.

Nel caso di completezza di informazione, i rapporti di verosimiglianza sono quelli rappresentati in Tabella 5.4 ed indicano prevalenza dell'ipotesi nulla su quella alternativa: gli odd a posteriori di ogni locus vanno da un minimo di 6.4396 ad un massimo di 45.043 volte l'odd a priori. L'odd a posteriori congiunto è di quasi 800000 volte l'odd a posteriori a conferma dell'ipotesi di colpevolezza del sospetto.

Tabella 5.2: I possibili genotipi di figlio (*cgt*) e moglie (*wgt*) dato *sgt*=BC

Allele ereditato	*cgt*	*wgt*	Allele ereditato	*cgt*	*wgt*
B	AB	A	B	BC	AC
B	AB	AB	B	BC	BC
B	AB	AC	B	BC	C
B	AB	AD	B	BC	CD
B	AB	AE	B	BC	CE
B	AB	AF	B	BC	CF
C	AC	A	C	C	AC
C	AC	AB	C	C	BC
C	AC	AC	C	C	C
C	AC	AD	C	C	CD
C	AC	AE	C	C	CE
C	AC	AF	C	C	CF
B	B	AB	B	BD	AD
B	B	B	B	BD	BD
B	B	BC	B	BD	CD
B	B	BD	B	BD	D
B	B	BE	B	BD	DE
B	B	BF	B	BD	DF
C	BC	AB	C	CD	AD
C	BC	B	C	CD	BD
C	BC	BC	C	CD	CD
C	BC	BD	C	CD	D
C	BC	BE	C	CD	DE
C	BC	BF	C	CD	DF

Allele ereditato	*cgt*	*wgt*	Allele ereditato	*cgt*	*wgt*
B	BE	AE	B	BF	AF
B	BE	BE	B	BF	BF
B	BE	CE	B	BF	CF
B	BE	DE	B	BF	DF
B	BE	E	B	BF	EF
B	BE	EF	B	BF	F
C	CE	AE	C	CF	AF
C	CE	BE	C	CF	BF
C	CE	CE	C	CF	CF
C	CE	DE	C	CF	DF
C	CE	E	C	CF	EF
C	CE	EF	C	CF	F

Tabella 5.3: Dati *Clayton*

Locus	Mistura	Sospetto	Vittima
D8S1179	CDE	DE	C
D18S51	CDEF	CD	EF
FGA	CD	CD	D
TH01	AC	C	AC
vWA	BCDF	BC	DF

Tabella 5.4: Verosimiglianze di H_0 e H_1 e LR

	D8S1179	D18S51	FGA	TH01	vWA	Congiunta
$s\&v$	0.8700	0.9235	0.8854	0.9783	0.9642	
$v\&U$	0.1300	0.0765	0.1146	0.0217	0.0358	
LR	6.4396	12.0771	7.7231	45.043	26.9305	728592.9

5.1 L'esempio *Clayton* con informazione incompleta

Nel caso di informazione incompleta si calcolano i rapporti di verosimiglianza per tutte le possibili combinazioni genotipiche di figlio e moglie del sospetto. In Tabella 5.5 sono elencati i LR per tutte le combinazioni genotipiche di moglie e figlio del sospetto dato il genotipo di quest'ultimo (sgt=DE). Da notare che cgt=wgt=DE è ripetuto due volte in quanto sono due casi distinti a seconda di quale allele il figlio erediti dal sospetto: da sgt=DE, il figlio può ereditare D (e di conseguenza eredita E dalla madre) oppure E (e di conseguenza eredita D dalla madre). Considerazioni su LR ed il suo legame con l'allele che il figlio eredita dal sospetto sono state già fatte nel Capitolo 4, quindi riproporremo solamente un riassunto per quanto riguarda il locus in questione:

- se il figlio eredita D dal sospetto e la moglie ha un genotipo diverso dal figlio (ad esempio cgt=AD, wgt=AB), allora LR=1.4096;

- se il figlio eredita E dal sospetto e la moglie ha un genotipo diverso dal figlio (ad esempio cgt=AE, wgt=AB), allora LR=2.3742;

- se figlio e moglie hanno lo stesso genotipo, allora LR è inferiore rispetto al caso con genotipi diversi (ad esempio con *cgt*=AE e *wgt*=AE, LR=1.5763).

I rapporti di verosimiglianza vanno da un minimo di 0.9868 (*cgt*=*wgt*=CD) ad un massimo di 2.3742 (i casi in cui il si riesce a dedurre che il figlio eredita E dal sospetto). Il valore minimo indica che l'odd a posteriori è simile a quello a priori, e cioè che i dati a disposizione confermano l'ipotesi di equiprobabilità tra H_0 e H_1. Il valore massimo, invece, indica che i dati a disposizione propendono per l'ipotesi di colpevolezza, seppur non nettamente. Infatti LR=2.3742 implica una probabilità a posteriori di H_0 di:

$$1 - \frac{1}{1 + 2.3742} = 0.7036$$

La media di tutti i possibili valori di LR per il locus D8S1179 è di 1.8082. Questo dato è calcolato pesando ogni valore di LR con le probabilità marginali del genotipo della madre (Tabella A.1 in appendice). Non bisogna pesare anche per le probabilità marginali dei genotipi del figlio in quanto questi, una volta noti i profili dei genitori, dipendono solo dagli alleli di questi. A titolo di esempio, il valore 2.3742 in cui *cgt*=EF e *wgt*=CF della Tabella 5.5 avrà peso 0.0349 (colonna D8S1179, riga CF della Tabella A.1).
Se per quanto riguarda un solo locus la difficoltà a prendere una decisione rimane, proviamo a vedere cosa accede considerando cinque loci differenti. La Tabella 5.6 riporta valori massimi, minimi e medie dei rapporti di verosimiglianza per ogni locus e complessivamente. Ogni locus ha undici diversi valori di LR, tranne il locus TH01 che ne ha soli cinque, in quanto il sospetto è omozigote. Essendo il rapporto di verosimiglianza complessivo il prodotto dei singoli, sono state calcolate tutte le $11^4 \cdot 5$ combinazioni possibili di LR. Alcuni percentili sono stati riportati nella Tabella 5.7 e si può vedere il quantile al 5% è di 6.2751, e cioè i dati di quella particolare combinazione implicano che l'odd a posteriori è circa sei volte l'odd a priori. I dati fanno quindi propendere per l'ipotesi di colpevolezza del sospetto.

L'odd a posteriori nel caso di informazione completa è di circa 728592 ed implica una notevole evidenza a favore dell'ipotesi di colpevolezza. Quello nel

Tabella 5.5: Locus D8S1179: genotipi possibili per figlio e moglie dato *sgt*=DE

cgt	*wgt*	LR	*cgt*	*wgt*	LR
AD	A	1.4096	D	AD	1.4096
AD	AB	1.4096	D	BD	1.4096
AD	AC	1.4096	D	CD	1.4096
AD	AD	1.0839	D	D	1.4096
AD	AE	1.4096	D	DE	1.4096
AD	AF	1.4096	D	DF	1.4096
AE	A	2.3742	DE	AD	2.3742
AE	AB	2.3742	DE	BD	2.3742
AE	AC	2.3742	DE	CD	2.3742
AE	AD	2.3742	DE	D	2.3742
AE	AE	1.5763	DE	DE	1.769
AE	AF	2.3742	DE	DF	2.3742
BD	AB	1.4096	DE	AE	1.4096
BD	B	1.4096	DE	BE	1.4096
BD	BC	1.4096	DE	CE	1.4096
BD	BD	1.2016	DE	DE	1.769
BD	BE	1.4096	DE	E	1.4096
BD	BF	1.4096	DE	EF	1.4096
BE	AB	2.3742	DF	AF	1.4096
BE	B	2.3742	DF	BF	1.4096
BE	BC	2.3742	DF	CF	1.4096
BE	BD	2.3742	DF	DF	1.0652
BE	BE	1.8382	DF	EF	1.4096
BE	BF	2.3742	DF	F	1.4096
CD	AC	1.4096	E	AE	2.3742
CD	BC	1.4096	E	BE	2.3742
CD	C	1.4096	E	CE	2.3742
CD	CD	0.9868	E	DE	2.3742
CD	CE	1.4096	E	E	2.3742
CD	CF	1.4096	E	EF	2.3742
CE	AC	2.3742	EF	AF	2.3742
CE	BC	2.3742	EF	BF	2.3742
CE	C	2.3742	EF	CF	2.3742
CE	CD	2.3742	EF	DF	2.3742
CE	CE	1.3789	EF	EF	1.537
CE	CF	2.3742	EF	F	2.3742

Tabella 5.6: Statistiche di LR per ogni locus

	D8S1179	D18S51	FGA	TH01	vWA	Complessivo
Minimo	0.9868	1.2421	0.777	1.9897	1.0707	2.0289
Media	1.8082	2.3236	2.2706	6.1490	3.9600	232.2917
Massimo	2.3742	2.6181	3.1708	6.7114	6.2344	824.6699

Tabella 5.7: Percentili dell'odd a posteriori complessivo

0%	5%	10%	15%	25%	50%	75%	100%
2.0289	6.2751	8.1752	9.8475	13.1276	23.1104	42.2429	824.669

caso di incompletezza di informazione va da 2.0289 a circa 825 ed indicano rispettivamente una probabilità di colpevolezza ($\Pr(H_0 \mid I)$) del circa 66% e quasi 100%. Tra questi valori, vi è ad esempio il quantile al 15% a cui corrisponde $\Pr(H_0 \mid I) = 90.7\%$. Di conseguenza, seppur in assenza di informazione, il modello può avere una certa affidabilità.

I risultati per quanto riguarda questo esempio evidenziano che il modello in assenza di informazione prevede, a seconda dei genotipi di moglie e figlio, la possibilità che il modello sia sufficientemente affidabile come dimostrato da alcuni rapporti di verosimiglianza evidenziati.

Capitolo 6

Modello con artefatti

In questo capitolo si parlerà degli artefatti (traduzione dall'inglese "artefacts"), cioè dei problemi che si possono manifestare in laboratorio durante il processo di PCR e si farà un leggero aggiustamento al network bayesiano in modo da prevedere queste eventualità. Gli errori considerati sono il *dropin* ed il *dropout*. Il problema di *stutter* non è stato considerato perché ci farebbe modificare il network bayesiano di partenza. Ricordiamo che il dropin è ciò che avviene quando nell'elettroferogramma compare un allele che non effettivamente presente nella mistura, mentre il dropout è l'opposto e cioè avviene quando un allele presente nella mistura non compare nell'elettroferogramma. Come in [Balding e Buckleton(2009)] si assumerà che possa avvenire al più un solo dropin per ogni locus e si escluderà la possibilità di un dropout ed un dropin sullo stesso allele. Tipicamente, il fenomeno del dropin (che chiameremo C) è abbastanza raro e quindi si assumerà $\Pr(C) = 0.05$ [Balding e Buckleton(2009)]. Anche per quanto riguarda il dropout (che indicheremo con D) verrà assunto la possibilità della perdita di al più un solo allele. La probabilità di dropout di un allele non è costante, bensì aumenta proporzionalmente al proprio repeat number e all'altezza del picco. Tuttavia si assumerà che il fenomeno di dropout avverrà con la stessa probabilità per ogni allele di ogni locus. $\Pr(D)$ sarà assunto essere 0.05 come nel caso *Garside-Bates* presentato in [Balding e Buckleton(2009)]. In virtù delle assunzioni elencate finora, le possibili differenze tra la vera mistura e quella

misurata in laboratorio sono le seguenti:

- la mistura nell'elettroferogramma presenta un allele in più rispetto a quella originale e quindi è avvenuto un dropin;
- la mistura nell'elettroferogramma presenta un allele in meno rispetto a quella originale e quindi è avvenuto un dropout;
- la mistura nell'elettroferogramma presenta un allele diverso rispetto a quella originale e quindi sono avvenuti sia un dropin che un dropout.

Non sarà inoltre considerato il caso in cui la mistura originale ha un solo allele e quest'ultimo subisce un dropout, in quanto l'elettroferogramma evidenzia almeno un allele.

6.1 Il network bayesiano con artefatti

Rispetto al network in Figura 3.7 viene aggiunto il nodo *mix.af* che si riferisce al profilo della scena del crimine misurato dall'elettroferogramma. *mix.af* dipende dal nodo *mix* e la relazione che li lega è che la misurazione della mistura originale (*mix*) ha come risultato la mistura evidenziata dall'elettroferogramma (*mix.af*). La direzione di causalità di conseguenza è come quella in Figura 6.1.

Per capire le probabilità di transizione da uno stato all'altro, si consideri il caso in cui *mix*=ABC:

- può accadere che avvenga un dropin con probabilità $\Pr(C) = 0.05$ e quindi *mix.af* assumerebbe con la stessa probabilità uno stato tra $\{ABCD, ABCE, ABCF\}$;

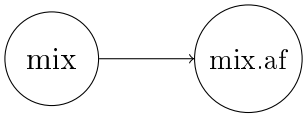

Figura 6.1: La relazione tra *mix* e *mix.af*

- può accadere che avvenga un dropout con probabilità $\Pr(D) = 0.05$ e quindi *mix.af* assumerebbe con la stessa probabilità uno stato tra $\{AB, AC, BC\}$;

- può accadere che avvengano sia un dropin che un dropout con probabilità $\Pr(C) \cdot \Pr(D) = 0.05^2$ e quindi *mix.af* assumerebbe con la stessa probabilità uno stato tra $\{ABD, ABE, ABF, ACD, ACE, ACF, BCD, BCE, BCF\}$;

- può infine accadere non avvenga alcun dropin o dropout e che di conseguenza il *mix.af*=ABC, con probabilità complementari a quelle elencate nei punti precedenti.

Con *mix*=ABC, i casi di dropin, come quelli di dropout, sono tre ed hanno probabilità 0.05, mentre quelli in cui avvengono entrambi sono nove ed hanno probabilità $0.05^2 = 0.0025$. Quindi

$$\begin{aligned} \Pr(\text{mix.af=ABC} \mid \text{mix=ABC}) &= 1 - 3 \cdot 0.05 - 3 \cdot 0.05 - 9 \cdot 0.05^2 \\ &= 0.6775 \end{aligned}$$

Se invece *mix*=A, allora può solo avvenire un dropin. Perciò si ha che da $mix = A$, $mix.af$ può assumere uno stato tra $\{AB, AC, AD, AE, AF\}$ con probabilità 0.05 e di conseguenza:

$$\begin{aligned} \Pr(\text{mix.af=A} \mid \text{mix=A}) &= 1 - 5 \cdot \Pr(C) \\ &= 1 - 5 \cdot 0.05 \\ &= 0.75 \end{aligned}$$

È prevista la possibilità che da una mistura di quattro alleli risulti nell'elettroferogramma una mistura di cinque alleli, e cioè che avvenga un dropin.

Le probabilità di transizione che identificano un dropin, un dropout o entrambi sono, qualora possibile, rispettivamente 0.05, 0.05 e 0.0025. Le probabilità che la mistura misurata in laboratorio sia quella originale dipende

dallo stato di partenza: se *mix* ha un solo allele, allora

$$\Pr(mix.af = k \mid mix = k) = 1 - 5 \cdot 0.05$$

Se *mix* ha due, tre o quattro alleli, identificando con N il numero di alleli presenti nella mistura originale, si ha che

$$\begin{aligned} \Pr(mix.af = k \mid mix = k) &= 1 - (6 - N)\Pr(C) - N\Pr(D) \\ &- (6 - N)N\Pr(C)\Pr(D) \end{aligned}$$

Gli stati possibili di *mix.af* sono sessantadue, sei in più rispetto a *mix* in quanto è prevista la possibilità che l'elettroferogramma evidenzi un allele in più oltre ai quattro già presenti. Sarebbe ragionevole pensare in caso si mistura con cinque alleli che ci siano stati più di due contributori, tuttavia in questa tesi ci si è sempre posti nella condizione di sapere che i contributori alla traccia rilevata sono due.
La tabella delle probabilità ha 56 colonne e 62 righe e quindi, per problemi di spazio, solo una piccola parte è riportata (Tabella 6.1). Il network bayesiano con la nuova variabile *mix.af* è quello in Figura 6.2.

A differenza del network descritto nel Capitolo 3, questo modello è un po' meno rigido in quanto prevede la possibilità che il sospetto sia effettivamente il colpevole, pur non essendo i suoi alleli presenti in *mix.af*. Infatti, con l'usuale assunzione che la vittima contribuisca effettivamente alla traccia trovata, **R** fornisce il seguente output per un caso ad informazione completa:

```
> querygrain(setEvidence(pn,c("mix.af","T2","vgt","sgt"),
+ c("ABCD","Yes","AB","CE")))$Target
Target
      v&s       v&U       s&U        2U
0.0269266 0.9730734 0.0000000 0.0000000
```

Si vede che, con *sgt*=CE, $\Pr(H_0)$, seppur bassa, non è nulla come sarebbe accaduto col network nei capitoli precedenti. Si prenda come esempio il locus

Tabella 6.1: Tabella delle probabilità per le misture A, ABC e ABCD

	A	ABC	ABCD		A	ABC	ABCD
A	0.75	0	0	B	0	0	0
AB	0.05	0.05	0	BC	0	0.05	0
AC	0.05	0.05	0	BD	0	0	0
AD	0.05	0	0	BE	0	0	0
AE	0.05	0	0	BF	0	0	0
AF	0.05	0	0	BCD	0	0.0025	0.05
ABC	0	0.6775	0.05	BCE	0	0.0025	0
ABD	0	0.0025	0.05	BCF	0	0.0025	0
ABE	0	0.0025	0	BDE	0	0	0
ABF	0	0.0025	0	BDF	0	0	0
ACD	0	0.0025	0.05	BEF	0	0	0
ACE	0	0.0025	0	BCDE	0	0	0.0025
ACF	0	0.0025	0	BCDF	0	0	0.0025
ADE	0	0	0	BCEF	0	0	0
ADF	0	0	0	BDEF	0	0	0
AEF	0	0	0	BCDEF	0	0	0
ABCD	0	0.05	0.68	C	0	0	0
ABCE	0	0.05	0.0025	CD	0	0	0
ABCF	0	0.05	0.0025	CE	0	0	0
ABDE	0	0	0.0025	CF	0	0	0
ABDF	0	0	0.0025	CDE	0	0	0
ABEF	0	0	0	CDF	0	0	0
ACDE	0	0	0.0025	CEF	0	0	0
ACDF	0	0	0.0025	CDEF	0	0	0
ACEF	0	0	0	D	0	0	0
ADEF	0	0	0	DE	0	0	0
ABCDE	0	0	0.05	DF	0	0	0
ABCDF	0	0	0.05	DEF	0	0	0
ABCEF	0	0	0	E	0	0	0
ABDEF	0	0	0	EF	0	0	0
ACDEF	0	0	0	F	0	0	0

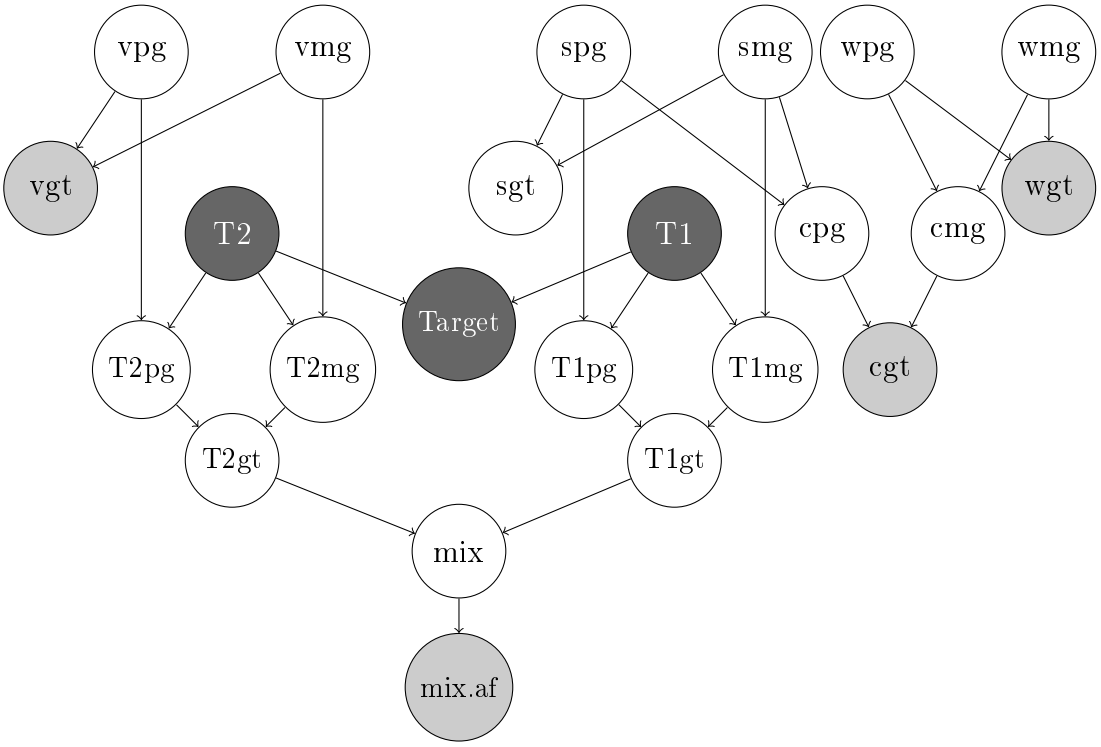

Figura 6.2: Il modello con artefatti

vWA ed il caso in cui l'elettroferogramma rivela la presenza degli alleli A, B e C. Allora è più probabile che, dato *vgt*=AB, il colpevole abbia un genotipo tra AC, BC e C. LR è lo stesso in tutti e tre i casi come già spiegato nel Capitolo 4 (Tabella 6.2). I casi in cui $sgt \in \{AD, E\}$ sono quelli in cui l'unica possibilità che il sospetto sia colpevole è che siano avvenuti un dropout ed un dropin ed hanno LR=0.0209. Se il sospetto ha genotipo CD o CE, allora deve essere avvenuto solo un dropout perché il sospetto possa essere il colpevole con LR=0.4184. Nel caso analogo, ma con solo dropin, LR è lo stesso. Quindi LR non fa distinzione tra dropin e dropout, a parità di *vgt*. Solo se avvengono entrambi, allora LR cambia.

Rispetto all'analisi nel Capitolo 4, il genotipo della vittima influisce notevolmente LR: se prima aveva solamente il ruolo di identificare gli alleli potenzialmente appartenenti al genotipo del colpevole, ora consente anche di avere una maggiore precisione nell'identificazione della mistura originale. Nella

Tabella 6.2: Esempi riassuntivi

mix	*vgt*	*sgt*	LR
ABC	AB	AC	5.6690
ABC	AB	BC	5.6690
ABC	AB	C	5.6690
ABC	AB	AD	0.0209
ABC	AB	E	0.0209
ABC	AB	CD	0.4184
ABC	AB	CE	0.4184
ABC	AB	A	0.4184
ABC	AB	B	0.4184

Tabella 6.3: Esempi riassuntivi

mix	*vgt*	*sgt*	LR
ABC	AB	AC	5.6690
ABC	AD	BC	19.3607
ABC	AD	BD	0.9680
ABC	A	B	1.5061
ABC	A	AB	1.5061
ABC	B	AB	1.0049

Tabella 6.3 sono riportati alcuni casi che evidenziano il ruolo del genotipo nel riconoscimento della traccia di DNA originale. La seconda riga (*cgt*=AD, *sgt*=BC) indica che c'è stato certamente un dropout dell'allele D; la certezza deriva dal fatto che la vittima è sicuramente contributrice alla mistura (*T2*=Yes). La terza riga (*cgt*=AD, *sgt*=BD) indica che c'è stato sicuramente un dropout (dell'allele D) e probabilmente anche un dropin (dell'allele C). Infine le ultime righe indicano un probabile dropin dell'allele C.

6.2 L'esempio *Clayton* con artefatti

Si consideri ancora l'esempio *Clayton* considerando la mistura come quella misurata in laboratorio, anziché quella originale come fatto in precedenza. Ripetendo le stesse procedure risulta che, in caso di disponibilità del genotipo del sospetto, i valori dei rapporti di verosimiglianza diminuiscono leggermente

Tabella 6.4: Dati *Clayton*: confronto tra il modello con artefatti e quello senza in caso di informazione completa

	D8S1179	D18S51	FGA	TH01	vWA	Congiunta
Senza artefatti	6.4396	12.0771	7.7231	45.043	26.9305	728592.9
Con artefatti	5.6876	9.4996	6.5784	17.7850	17.4473	110290.2

Tabella 6.5: Massimo, minimo e media per ogni locus in caso di informazione incompleta

	D8S1179	D18S51	FGA	TH01	vWA	Complessivo
Minimo	1.0627	1.1675	0.7734	1.5622	1.0859	1.6278
Media	1.7221	2.1484	2.1267	3.4921	3.1793	87.35546
Massimo	2.1755	2.3829	2.9859	3.7928	4.5833	266.9498

(Tabella 6.4), comunque LR complessivo spinge nettamente all'accettazione dell'ipotesi nulla.

Per quanto riguarda il caso con informazione incompleta, ripetendo la stessa analisi fatta in precedenza, LR diminuisce ancora (Tabella 6.5). Il rapporto di verosimiglianza complessivo minimo è basso e quindi una decisione basata su questo potrebbe rivelarsi un azzardo, mentre LR medio sembra sufficientemente alto da permettere di dare un giudizio di colpevolezza del sospetto. La Tabella 6.6 riporta alcuni percentili della distribuzione del LR complessivo: se i quantili del 10% e del 20% non danno generalmente garanzie, i quantili 30% e 40% cominciano a fornire maggiore evidenza a favore dell'ipotesi nulla. Se i rapporti di verosimiglianza sono ancor più grandi, allora sempre maggiori sono le prove a favore dell'ipotesi di colpevolezza del sospetto.

Tabella 6.6: Percentili della distribuzione del LR complessivo

0%	10%	20%	30%	40%	50%	75%	100%
1.6278	4.9067	6.5941	8.2305	10.0073	12.0714	20.1778	266.9498

Seppur più limitatamente rispetto al caso senza artefatti, i risultati ottenuti dal modello applicato ai dati *Clayton* indicano che, anche considerando l'eventualità di problemi di laboratorio, il network bayesiano può fornire uno strumento di supporto alle decisioni.

Capitolo 7

Conclusioni

L'obiettivo di questo lavoro era quello di fornire un network bayesiano flessibile al cambiamento dell'informazione in un problema di genetica forense. Il cambiamento in questione è l'assenza del profilo genetico del sospetto, parzialmente compensata dai profili della moglie e del figlio. Si è creato un network bayesiano per un singolo locus e poi, sfruttando l'assunzione di indipendenza tra i vari loci, si è calcolato il risultato congiunto come il prodotto dei risultati di ogni singolo locus.
Si è seguito un approccio di tipo qualitativo per cui un allele è presente nella mistura oppure assente, senza contare quanto esso incida. Mentre nell'approccio quantitativo si considera quanto un individuo contribuisce alla traccia, nell'approccio qualitativo tutti gli individui implicati sono presenti o meno. Che un allele sia stato lasciato sulla scena del crimine da un solo individuo o da due individui poco importa in quanto viene considerato solamente presente.
Il modello applicato ai dati *Clayton* ha dimostrato che un network bayesiano può essere utilizzato anche in caso di assenza del profilo del sospetto. In particolare si è dimostrato che in un caso in cui si sarebbe accettata l'ipotesi di colpevolezza del sospetto in presenza del genotipo di quest'ultimo, buona parte delle possibili combinazioni genotipiche di moglie e figlio portano a prendere la stessa decisione.
Siccome i profili DNA di moglie e figlio permettono normalmente di identi-

ficare uno solo degli alleli del sospetto, c'è il rischio di commettere l'errore di accettare l'ipotesi di colpevolezza del sospetto, anche se innocente, che ha il profilo DNA simile a quello del vero colpevole. Tuttavia assumendo l'indipendenza tra loci, aumentando il numero di questi si dovrebbe superare questo problema a meno che i due individui non abbiano genotipi simili per diversi loci.
Si è inoltre visto che, in caso di stesso genotipo tra moglie e figlio del sospetto, si abbassa il valore del rapporto di verosimiglianza rischiando di portare ad una decisione che, disponendo del genotipo del sospetto, non si prenderebbe. Tuttavia va considerato che se per molti loci moglie e figlio hanno lo stesso genotipo, questo può essere notato da chi deve prendere una decisione ed è quindi ragionevole dubitare del risultato, cercando di rifare le analisi aggiungendo altri loci.
Resta quindi molto importante fare un'analisi tra molti loci, in modo che una situazione sfavorevole su di uno (come ad esempio genotipo identico per la moglie ed il figlio del sospetto), sia meno pesante in termini complessivi.
Per quanto riguarda il modello con gli artefatti, le considerazioni appena fatte continuano a valere anche se le assunzioni potrebbero sembrare troppo restrittive. È stata esclusa la possibilità di *stutter*, si è assunto indipendenza tra le possibilità di dropin e quelle di dropout e si sono scelti dei valori fissi per le probabilità di questi ultimi. Non c'è tuttavia un valore generico delle probabilità di stutter, dropin e dropout in quanto questo dipende dal laboratorio e dalla precisione del processo di PCR.
Sulle assunzioni di base del modello come quello dell'equilibrio di Hardy-Weinberg o l'indipendenza tra loci sono già stati fatti altri studi, mentre per quanto riguarda le assunzioni circa gli artefatti, la letteratura continua ad evolvere.

Appendice A

Tabelle per dati *Clayton*

In questa appendice sono elencate alcune tabelle:

- la tabella con le frequenze dei vari genotipi per ogni locus (Tabella A.1);
- le tabelle con tutte le combinazioni genotipiche di moglie e figlio dato il profilo genetico del sospetto per vari locus senza considerare gli artefatti (Tabella A.2, Tabella A.3, Tabella A.4 e Tabella A.5);
- le tabelle con tutte le combinazioni genotipiche di moglie e figlio dato il profilo genetico del sospetto per vari locus considerando gli artefatti (Tabella A.6, Tabella A.7, Tabella A.8, Tabella A.9 e Tabella A.10).

Tabella A.1: Probabilità genotipiche marginali per ogni locus dei dati *Clayton*

	D8S1179	D18S51	FGA	TH01	vWA
A	0.0114	0.0345	0.0105	0.0001	0.0180
AB	0.131	0.0561	0.0453	0.0035	0.0215
AC	0.0324	0.0805	0.0437	0.0020	0.0620
AD	0.0756	0.0710	0.0406	0.0017	0.0631
AE	0.0449	0.0547	0.0289	0.0022	0.0612
AF	0.0245	0.0403	0.0258	0.0048	0.0242
B	0.0038	0.0228	0.0487	0.0578	0.0064
BC	0.0187	0.0655	0.0939	0.0682	0.0371
BD	0.0436	0.0577	0.0872	0.0578	0.0378
BE	0.0259	0.0444	0.0621	0.0740	0.0366
BF	0.0141	0.0327	0.0554	0.1618	0.0145
C	0.0231	0.0470	0.0453	0.0201	0.0536
CD	0.1078	0.0828	0.0842	0.0341	0.1090
CE	0.0640	0.0638	0.0599	0.0436	0.1057
CF	0.0349	0.0470	0.0534	0.0954	0.0419
D	0.1258	0.0365	0.0391	0.0144	0.0555
DE	0.1494	0.0562	0.0556	0.0370	0.1076
DF	0.0814	0.0414	0.0496	0.0809	0.0426
E	0.0444	0.0216	0.0198	0.0237	0.0522
EF	0.0483	0.0319	0.0353	0.1036	0.0413
F	0.0132	0.0117	0.0157	0.1132	0.0082

Tabella A.2: Locus D18S51: genotipi possibili per figlio e moglie dato *sgt*=CD

cgt	*wgt*	LR
AC	A	2.3065
AC	AB	2.3065
AC	AC	1.2421
AC	AD	2.3065
AC	AE	2.3065
AC	AF	2.3065
AD	A	2.6181
AD	AB	2.6181
AD	AC	2.6181
AD	AD	1.3271
AD	AE	2.6181
AD	AF	2.6181
BC	AB	2.3065
BC	B	2.3065
BC	BC	1.3596
BC	BD	2.3065
BC	BE	2.3065
BC	BF	2.3065
BD	AB	2.6181
BD	B	2.6181
BD	BC	2.6181
BD	BD	1.4621
BD	BE	2.6181
BD	BF	2.6181
C	AC	2.3065
C	BC	2.3065
C	C	2.3065
C	CD	2.3065
C	CE	2.3065
C	CF	2.3065
CD	AD	2.3065
CD	BD	2.3065
CD	CD	2.4524
CD	D	2.3065
CD	DE	2.3065
CD	DF	2.3065

cgt	*wgt*	LR
CD	AC	2.6181
CD	BC	2.6181
CD	C	2.6181
CD	CD	2.4524
CD	CE	2.6181
CD	CF	2.6181
CE	AE	2.3065
CE	BE	2.3065
CE	CE	1.3741
CE	DE	2.3065
CE	E	2.3065
CE	EF	2.3065
CF	AF	2.3065
CF	BF	2.3065
CF	CF	1.5377
CF	DF	2.3065
CF	EF	2.3065
CF	F	2.3065
D	AD	2.6181
D	BD	2.6181
D	CD	2.6181
D	D	2.6181
D	DE	2.6181
D	DF	2.6181
DE	AE	2.6181
DE	BE	2.6181
DE	CE	2.6181
DE	DE	1.479
DE	E	2.6181
DE	EF	2.6181
DF	AF	2.6181
DF	BF	2.6181
DF	CF	2.6181
DF	DF	1.6702
DF	EF	2.6181
DF	F	2.6181

Tabella A.3: Locus FGA: genotipi possibili per figlio e moglie dato *sgt*=AD

cgt	*wgt*	LR	*cgt*	*wgt*	LR
AC	A	3.1708	CD	AC	1.6441
AC	AB	3.1708	CD	BC	1.6441
AC	AC	2.139	CD	C	1.6441
AC	AD	3.1708	CD	CD	2.4357
AC	AE	3.1708	CD	CE	1.6441
AC	AF	3.1708	CD	CF	1.6441
AD	A	1.6441	CE	AE	3.1708
AD	AB	1.6441	CE	BE	3.1708
AD	AC	1.6441	CE	CE	1.9091
AD	AD	1.082	CE	DE	3.1708
AD	AE	1.6441	CE	E	3.1708
AD	AF	1.6441	CE	EF	3.1708
BC	AB	3.1708	CF	AF	3.1708
BC	B	3.1708	CF	BF	3.1708
BC	BC	1.5572	CF	CF	1.9949
BC	BD	3.1708	CF	DF	3.1708
BC	BE	3.1708	CF	EF	3.1708
BC	BF	3.1708	CF	F	3.1708
BD	AB	1.6441	D	AD	1.6441
BD	B	1.6441	D	BD	1.6441
BD	BC	1.6441	D	CD	1.6441
BD	BD	0.777	D	D	1.6441
BD	BE	1.6441	D	DE	1.6441
BD	BF	1.6441	D	DF	1.6441
C	AC	3.1708	DE	AE	1.6441
C	BC	3.1708	DE	BE	1.6441
C	C	3.1708	DE	CE	1.6441
C	CD	3.1708	DE	DE	0.9605
C	CE	3.1708	DE	E	1.6441
C	CF	3.1708	DE	EF	1.6441
CD	AD	3.1708	DF	AF	1.6441
CD	BD	3.1708	DF	BF	1.6441
CD	CD	2.4357	DF	CF	1.6441
CD	D	3.1708	DF	DF	1.0057
CD	DE	3.1708	DF	EF	1.6441
CD	DF	3.1708	DF	F	1.6441

Tabella A.4: Locus TH01: genotipi possibili per figlio e moglie dato *sgt*=C

cgt	*wgt*	LR
AC	A	6.7114
AC	AB	6.7114
AC	AC	6.7114
AC	AD	6.7114
AC	AE	6.7114
AC	AF	6.7114
BC	AB	6.7114
BC	B	6.7114
BC	BC	2.49
BC	BD	6.7114
BC	BE	6.7114
BC	BF	6.7114
C	AC	6.7114
C	BC	6.7114
C	C	6.7114
C	CD	6.7114
C	CE	6.7114
C	CF	6.7114
CD	AD	6.7114
CD	BD	6.7114
CD	CD	3.6324
CD	D	6.7114
CD	DE	6.7114
CD	DF	6.7114
CE	AE	6.7114
CE	BE	6.7114
CE	CE	3.2184
CE	DE	6.7114
CE	E	6.7114
CE	EF	6.7114
CF	AF	6.7114
CF	BF	6.7114
CF	CF	1.9897
CF	DF	6.7114
CF	EF	6.7114
CF	F	6.7114

Tabella A.5: Locus vWA: genotipi possibili per figlio e moglie dato *sgt*=AB

cgt	*wgt*	LR	*cgt*	*wgt*	LR
AB	A	6.2344	BE	AE	6.2344
AB	AB	2.3343	BE	BE	1.6202
AB	AC	6.2344	BE	CE	6.2344
AB	AD	6.2344	BE	DE	6.2344
AB	AE	6.2344	BE	E	6.2344
AB	AF	6.2344	BE	EF	6.2344
AC	A	2.1598	BF	AF	6.2344
AC	AB	2.1598	BF	BF	2.9308
AC	AC	1.368	BF	CF	6.2344
AC	AD	2.1598	BF	DF	6.2344
AC	AE	2.1598	BF	EF	6.2344
AC	AF	2.1598	BF	F	6.2344
B	AB	6.2344	C	AC	2.1598
B	B	6.2344	C	BC	2.1598
B	BC	6.2344	C	C	2.1598
B	BD	6.2344	C	CD	2.1598
B	BE	6.2344	C	CE	2.1598
B	BF	6.2344	C	CF	2.1598
BC	AC	6.2344	CD	AD	2.1598
BC	BC	3.2082	CD	BD	2.1598
BC	C	6.2344	CD	CD	1.0707
BC	CD	6.2344	CD	D	2.1598
BC	CE	6.2344	CD	DE	2.1598
BC	CF	6.2344	CD	DF	2.1598
BC	AB	2.1598	CE	AE	2.1598
BC	B	2.1598	CE	BE	2.1598
BC	BC	3.2082	CE	CE	1.0872
BC	BD	2.1598	CE	DE	2.1598
BC	BE	2.1598	CE	E	2.1598
BC	BF	2.1598	CE	EF	2.1598
BD	AD	6.2344	CF	AF	2.1598
BD	BD	1.5838	CF	BF	2.1598
BD	CD	6.2344	CF	CF	1.5533
BD	D	6.2344	CF	DF	2.1598
BD	DE	6.2344	CF	EF	2.1598
BD	DF	6.2344	CF	F	2.1598

Tabella A.6: Locus D8S1179: genotipi possibili per figlio e moglie dato *sgt*=DE con artefatti

cgt	*wgt*	LR	*cgt*	*wgt*	LR
AD	A	1.4164	D	AD	1.4164
AD	AB	1.4164	D	BD	1.4164
AD	AC	1.4164	D	CD	1.4164
AD	AD	1.0919	D	D	1.4164
AD	AE	1.4164	D	DE	1.4164
AD	AF	1.4164	D	DF	1.4164
AE	A	2.1755	DE	AD	2.1755
AE	AB	2.1755	DE	BD	2.1755
AE	AC	2.1755	DE	CD	2.1755
AE	AD	2.1755	DE	D	2.1755
AE	AE	1.4484	DE	DE	1.6992
AE	AF	2.1755	DE	DF	2.1755
BD	AB	1.4164	DE	AE	1.4164
BD	B	1.4164	DE	BE	1.4164
BD	BC	1.4164	DE	CE	1.4164
BD	BD	1.2092	DE	DE	1.6992
BD	BE	1.4164	DE	E	1.4164
BD	BF	1.4164	DE	EF	1.4164
BE	AB	2.1755	DF	AF	1.4164
BE	B	2.1755	DF	BF	1.4164
BE	BC	2.1755	DF	CF	1.4164
BE	BD	2.1755	DF	DF	1.0732
BE	BE	1.6871	DF	EF	1.4164
BE	BF	2.1755	DF	F	1.4164
CD	AC	1.4164	E	AE	2.1755
CD	BC	1.4164	E	BE	2.1755
CD	C	1.4164	E	CE	2.1755
CD	CD	1.0627	E	DE	2.1755
CD	CE	1.4164	E	E	2.1755
CD	CF	1.4164	E	EF	2.1755
CE	AC	2.1755	EF	AF	2.1755
CE	BC	2.1755	EF	BF	2.1755
CE	C	2.1755	EF	CF	2.1755
CE	CD	2.1755	EF	DF	2.1755
CE	CE	1.363	EF	EF	1.4126
CE	CF	2.1755	EF	F	2.1755

Tabella A.7: Locus D18S51: genotipi possibili per figlio e moglie dato *sgt*=CD con artefatti

cgt	*wgt*	LR	*cgt*	*wgt*	LR
AC	A	2.1559	CD	AC	2.3829
AC	AB	2.1559	CD	BC	2.3829
AC	AC	1.1675	CD	C	2.3829
AC	AD	2.1559	CD	CD	2.2622
AC	AE	2.1559	CD	CE	2.3829
AC	AF	2.1559	CD	CF	2.3829
AD	A	2.3829	CE	AE	2.1559
AD	AB	2.3829	CE	BE	2.1559
AD	AC	2.3829	CE	CE	1.3995
AD	AD	1.2149	CE	DE	2.1559
AD	AE	2.3829	CE	E	2.1559
AD	AF	2.3829	CE	EF	2.1559
BC	AB	2.1559	CF	AF	2.1559
BC	B	2.1559	CF	BF	2.1559
BC	BC	1.2766	CF	CF	1.5322
BC	BD	2.1559	CF	DF	2.1559
BC	BE	2.1559	CF	EF	2.1559
BC	BF	2.1559	CF	F	2.1559
BD	AB	2.3829	D	AD	2.3829
BD	B	2.3829	D	BD	2.3829
BD	BC	2.3829	D	CD	2.3829
BD	BD	1.3371	D	D	2.3829
BD	BE	2.3829	D	DE	2.3829
BD	BF	2.3829	D	DF	2.3829
C	AC	2.1559	DE	AE	2.3829
C	BC	2.1559	DE	BE	2.3829
C	C	2.1559	DE	CE	2.3829
C	CD	2.1559	DE	DE	1.4701
C	CE	2.1559	DE	E	2.3829
C	CF	2.1559	DE	EF	2.3829
CD	AD	2.1559	DF	AF	2.3829
CD	BD	2.1559	DF	BF	2.3829
CD	CD	2.2622	DF	CF	2.3829
CD	D	2.1559	DF	DF	1.6233
CD	DE	2.1559	DF	EF	2.3829
CD	DF	2.1559	DF	F	2.3829

Tabella A.8: Locus FGA: genotipi possibili per figlio e moglie dato *sgt*=AD con artefatti

cgt	*wgt*	LR
AC	A	2.9859
AC	AB	2.9859
AC	AC	2.0502
AC	AD	2.9859
AC	AE	2.9859
AC	AF	2.9859
AD	A	1.5103
AD	AB	1.5103
AD	AC	1.5103
AD	AD	1.0316
AD	AE	1.5103
AD	AF	1.5103
BC	AB	2.9859
BC	B	2.9859
BC	BC	1.524
BC	BD	2.9859
BC	BE	2.9859
BC	BF	2.9859
BD	AB	1.5103
BD	B	1.5103
BD	BC	1.5103
BD	BD	0.7734
BD	BE	1.5103
BD	BF	1.5103
C	AC	2.9859
C	BC	2.9859
C	C	2.9859
C	CD	2.9859
C	CE	2.9859
C	CF	2.9859
CD	AD	2.9859
CD	BD	2.9859
CD	CD	2.2724
CD	D	2.9859
CD	DE	2.9859
CD	DF	2.9859
CD	AC	1.5103
CD	BC	1.5103
CD	C	1.5103
CD	CD	2.2724
CD	CE	1.5103
CD	CF	1.5103
CE	AE	2.9859
CE	BE	2.9859
CE	CE	1.842
CE	DE	2.9859
CE	E	2.9859
CE	EF	2.9859
CF	AF	2.9859
CF	BF	2.9859
CF	CF	1.9197
CF	DF	2.9859
CF	EF	2.9859
CF	F	2.9859
D	AD	1.5103
D	BD	1.5103
D	CD	1.5103
D	D	1.5103
D	DE	1.5103
D	DF	1.5103
DE	AE	1.5103
DE	BE	1.5103
DE	CE	1.5103
DE	DE	0.9286
DE	E	1.5103
DE	EF	1.5103
DF	AF	1.5103
DF	BF	1.5103
DF	CF	1.5103
DF	DF	0.9668
DF	EF	1.5103
DF	F	1.5103

Tabella A.9: Locus TH01: genotipi possibili per figlio e moglie dato *sgt*=C con artefatti

cgt	*wgt*	LR
AC	A	3.7628
AC	AB	3.7628
AC	AC	3.7628
AC	AD	3.7628
AC	AE	3.7628
AC	AF	3.7628
BC	AB	3.7628
BC	B	3.7628
BC	BC	1.7163
BC	BD	3.7628
BC	BE	3.7628
BC	BF	3.7628
C	AC	3.7628
C	BC	3.7628
C	C	3.7628
C	CD	3.7628
C	CE	3.7628
C	CF	3.7628
CD	AD	3.7628
CD	BD	3.7628
CD	CD	2.198
CD	D	3.7628
CD	DE	3.7628
CD	DF	3.7628
CE	AE	3.7628
CE	BE	3.7628
CE	CE	2.0106
CE	DE	3.7628
CE	E	3.7628
CE	EF	3.7628
CF	AF	3.7628
CF	BF	3.7628
CF	CF	1.5622
CF	DF	3.7628
CF	EF	3.7628
CF	F	3.7628

Tabella A.10: Locus vWA: genotipi possibili per figlio e moglie dato *sgt*=AB con artefatti

cgt	*wgt*	LR	*cgt*	*wgt*	LR
AB	A	4.5833	BE	AE	4.5833
AB	AB	1.7286	BE	BE	1.2059
AB	AC	4.5833	BE	CE	4.5833
AB	AD	4.5833	BE	DE	4.5833
AB	AE	4.5833	BE	E	4.5833
AB	AF	4.5833	BE	EF	4.5833
AC	A	2.1376	BF	AF	4.5833
AC	AB	2.1376	BF	BF	2.3665
AC	AC	1.3612	BF	CF	4.5833
AC	AD	2.1376	BF	DF	4.5833
AC	AE	2.1376	BF	EF	4.5833
AC	AF	2.1376	BF	F	4.5833
B	AB	4.5833	C	AC	2.1376
B	B	4.5833	C	BC	2.1376
B	BC	4.5833	C	C	2.1376
B	BD	4.5833	C	CD	2.1376
B	BE	4.5833	C	CE	2.1376
B	BF	4.5833	C	CF	2.1376
BC	AC	4.5833	CD	AD	2.1376
BC	BC	2.7669	CD	BD	2.1376
BC	C	4.5833	CD	CD	1.2613
BC	CD	4.5833	CD	D	2.1376
BC	CE	4.5833	CD	DE	2.1376
BC	CF	4.5833	CD	DF	2.1376
BC	AB	2.1376	CE	AE	2.1376
BC	B	2.1376	CE	BE	2.1376
BC	BC	2.7669	CE	CE	1.0859
BC	BD	2.1376	CE	DE	2.1376
BC	BE	2.1376	CE	E	2.1376
BC	BF	2.1376	CE	EF	2.1376
BD	AD	4.5833	CF	AF	2.1376
BD	BD	1.4626	CF	BF	2.1376
BD	CD	4.5833	CF	CF	1.6496
BD	D	4.5833	CF	DF	2.1376
BD	DE	4.5833	CF	EF	2.1376
BD	DF	4.5833	CF	F	2.1376

Bibliografia

[Balding e Buckleton(2009)] Balding D. J.; Buckleton J. (2009). Interpreting low template dna profiles. *Forensic Science International: Genetics*.

[Budowle *e altri*(1999)] Budowle B.; Moretti T. R.; Baumstark A. L.; Defenberg D. A.; Keyes K. M. (1999). Population data on thirteen codis core short tandem repeat loci in african americans, u.s. caucasian, hispanics, bahamians, jamaicans and trinidadians. *Journal of Forensic Sciences*.

[Butler(2005)] Butler J. M. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Elvisier Academic Press, seconda edizione.

[Clayton *e altri*(1998)] Clayton T. M.; Whitaker J. P.; Sparkes R.; Gill P. (1998). Analysis and interpretation of mixed forensic stains using dna str profiling. *Forensic Science International*.

[Corp.(2002)] Corp. P. (2002). Genetic identity reference information.

[Cowell *e altri*(1999)] Cowell R. G.; Dawid A. P.; Lauritzen S. L.; Spiegelhalter D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer.

[Cowell *e altri*(2013)] Cowell R. G.; Graversen T.; Lauritzen S. L.; Mortera J. (2013). Analysis of dna mixtures with artefacts.

[Højsgaard(2013)] Højsgaard S. (2013). grain: Graphical independence networks.

[Jensen e Nielsen(2001)] Jensen F. V.; Nielsen T. D. (2001). *Bayesian networks and decision graphs*. Springer.

[Kupferschmid *e altri*(1999)] Kupferschmid T. D.; Calicchio T.; Budowle B. (1999). Maine caucasian population dna database using twelve short tandem repeat loci. *Journal of Forensic Sciences*.

[Mortera *e altri*(2003)] Mortera J.; Dawid A. P.; Lauritzen S. L. (2003). Probabilistic expert systems for dna mixture profiling. *Theoretical Population Biology*.

[Taroni *e altri*(2006)] Taroni F.; Aitken C.; Biefdermann A. (2006). *Bayesian Networks and Probabilistic Inference in Forensic Science*. Wiley.