Valutazioni probabilistiche sui riscontri del DNA a scopo di identificazione criminale*

Marco Bramanti
Dipartimento di Matematica. Politecnico di Milano
Via Bonardi 9. 20133 Milano
marco.bramanti@polimi.it

22 marzo 2010

Sommario

Dopo aver brevemente illustrato in cosa consiste il test del DNA, si discutono alcuni problemi probabilistici legati a questo test e si cerca di stabilire qualche formula per il calcolo delle probabilità di eventi significativi in questo contesto. Ne emergono alcune osservazioni interessanti, dal punto di vista sia qualitativo che quantitativo.

Introduzione

L'uso del test del DNA a scopo di identificazione criminale, utilizzato a partire dal 1985 nel Regno Unito e via via diffuso in molti paesi tra cui l'Italia, accompagnato dalla realizzazione di database del DNA in certi paesi (primi fra tutti ancora il Regno Unito, in seguito gli Stati Uniti e parecchi altri), ha certamente dato un contributo significativo alla lotta contro il crimine, non senza suscitare talvolta accesi dibattiti. Si va dalle posizioni entusiastiche di chi, dati alla mano, sostiene che la realizzazione di ampi database del DNA anche nel nostro paese rivoluzionerebbe la lotta alla criminalità, agli scetticismi di chi pone dei dubbi sul valore probante di questo metodo o vi si oppone per considerazioni legate alla privacy o ad altri aspetti.

In questo articolo, dopo aver illustrato sinteticamente in cosa consiste e su quali basi biologiche poggia il test del DNA, ci si concentrerà su alcuni aspetti probabilistici legati alla valutazione del "valore probante" di questo test, prescindendo dai numerosi altri aspetti del dibattito. Ci occuperemo esclusivamente dell'uso del test del DNA a scopo di identificazione criminale, senza trattare invece quello legato ai "test di paternità", altra possibile applicazione di questa metodologia, che però presenta aspetti e problemi di tipo diverso.

*Pubblicato su: La Matematica nella Società e nella Cultura - Rivista dell'Unione Matematica Italiana, Serie I, Vol.II, n. 3, Dicembre 2009, pp.447-493.

Vedremo che per quanto il test del DNA, se correttamente utilizzato, sia da considerarsi uno strumento di identificazione molto potente, una quantificazione probabilistica del suo valore probatorio appare piuttosto sfuggente. Inoltre, come spesso accade nelle applicazioni del calcolo delle probabilità a questioni relativamente elementari, certi giudizi che possono essere suggeriti dal puro buon senso appaiono ingenui o fuorvianti.

Si può quindi capire almeno in parte l'origine di certe perplessità all'utilizzo in ambito processuale dei risultati di questo test. Al tempo stesso, una maggiore consapevolezza delle problematiche scientifiche e matematiche coinvolte non può che giovare a questo dibattito, anche se non ne esaurisce le sfaccettature.

L'articolo è suddiviso in tre parti. Nella prima, più breve, si descrivono il test e le sue basi biologiche; nella seconda si presentano alcuni problemi probabilistici legati al test; nella terza si cerca di dare qualche risposta alle domande sollevate in precedenza. I dati reali utilizzati nei calcoli e nelle tabelle sono tratti perlopiù da fonti statunitensi (database dell'F.B.I. e relativi documenti), che sono abbondanti e facilmente consultabili dalla rete; i valori numerici ottenuti quindi, per quanto non corrispondenti ai dati statistici della popolazione italiana, sono comunque realistici.

Ringraziamenti. Desidero ringraziare i referee, che hanno letto con grande accuratezza il manoscritto, contribuendo al suo miglioramento sotto molti aspetti.

1 Le basi biologiche del test del DNA

Cominciamo col ricordare alcuni fatti ben noti. Il DNA di un individuo è contenuto in 23 coppie di *cromosomi* (per ogni coppia, uno ereditato dal padre e uno dalla madre); ogni cromosoma consiste di due filamenti di DNA avvolti in una doppia elica. Ogni filamento è una lunga catena di *nucleotidi* contenenti ciascuno una base azotata tra 4 possibili: Adenina, Timina, Citosina, Guanina (abbrevieremo: A, T, C, G). I due filamenti sono legati tra loro nel modo seguente: ogni base azotata del primo si lega ad una base azotata del secondo, e precisamente: A si lega con T, C si lega con G. Per questo motivo l'informazione contenuta nei due filamenti di ogni doppia elica è sostanzialmente la stessa, essendo i due filamenti l'uno il negativo dell'altro. (Questo fatto sarà chiarito ulteriormente quando spiegheremo che cosa esattamente si va ad osservare nel DNA, per eseguire il test). Perciò nel seguito parleremo di filamenti singoli, prescindendo dalla struttura di doppia elica.

Qualche numero: l'intero DNA di un individuo consiste in una sequenza di circa 3 miliardi di nucleotidi, cioè di $3 \cdot 10^9$ "lettere" scelte tra A, T, G, C. Di questa lunga sequenza (il cossiddetto *genoma*), la maggior parte è uguale in ogni individuo (caratterizza cioè la specia umana in quanto tale) e solo circa $3 \cdot 10^6$ nucleotidi cambiano da individuo a individuo. Gemelli identici hanno identico DNA. A parte questo caso, si può pensare che due individui abbiano sempre DNA diverso; tuttavia, è evidente che non è possibile confrontare l'intera

catena del DNA di due persone, ma si deve procedere con confronti parziali, ed è qui che entra la probabilità: se un confronto parziale mostra delle diversità, si può escludere che si tratti della stessa persona, mentre se il confronto parziale mostra coincidenza, si apre una problematica di tipo statistico-probabilistico.

È importante quindi scegliere bene i tratti di DNA da confrontare. Per capire come si effettua questa scelta, dobbiamo spiegare qualcosa di più sulla struttura del nostro genoma. Non tutto il DNA codifica informazioni; le parti codificanti sono i *geni*; tra un gene e l'altro vi sono sequenze non codificanti (che si possono vedere come una sorta di marcatori che servono a separare e individuare le parti codificanti). Tra queste parti non codificanti hanno importanza per il nostro discorso certe sequenze periodiche di basi: tipicamente, una *coppia* o un *quartetto* di basi viene ripetuta per 10-20 volte, come in

A, C, A, C, A, C, A, C, A, C, A, C, A, C, A, C, A, C, A, C, A, C, A, C.

Queste sequenze vengono dette "*short tandem repeat*" (STR), e alcune di esse hanno la proprietà che la lunghezza del periodo è altamente variabile da individuo a individuo (pur non avendo questa variabilità alcuna conseguenza somatica, trattandosi di una porzione non codificante). Questo fatto si chiama *polimorfismo*, ed è la base dei metodi di identificazione mediante il DNA. Più precisamente[1]:

1. Si fissa l'attenzione su alcuni *loci* (specifiche posizioni su specifici cromosomi) in cui si sa esserci un STR, e si conta quante volte è ripetuta la coppia di basi; si tratta di un numero compreso tra 10 o poco meno e 30 o poco più. Si può capire meglio ora perché nel nostro discorso è lecito prescindere dalla struttura a doppia elica del DNA: se un filamento presenta in un certo locus una sequenza A, C ripetuta 10 volte, ad esempio, il filamento gemello presenterà in corrispondenza una sequenza T, G ripetuta 10 volte; ciò che a noi importa in questo discorso è solo il numero di ripetizioni (10), perciò è irrilevante quale dei due filamenti abbiamo osservato.

2. Per ogni *locus* si sa che nella popolazione esiste un certo numero di varianti; ogni variante si dice *allele*; ad esempio, può essere noto che in un certo locus possono esserci, in individui diversi, 12, 13, 14, 15 o 16 ripetizioni di una coppia di basi (ma non un altro numero), mentre in un altro locus possono esserci altre possibilità.

3. Ogni locus viene osservato sempre nei due cromosomi accoppiati (quello materno e quello paterno), per cui in ogni locus noi osserviamo in effetti una coppia di alleli; la coppia si chiama *genotipo*. Non è una coppia ordinata, perché non sappiamo quale allele viene dal padre e quale dalla madre. Se i due alleli della coppia sono uguali, diciamo che quell'individuo in quel locus è *omozigote*; altrimenti che è *eterozigote*. Se in un locus ci sono n possibili alleli, i corrispondenti genotipi sono quindi $n\,(n+1)\,/2$.

4. Si osserva infine quello che accade non in un solo locus, ma in un certo

[1] Qualche riferimento per queste informazioni di base sul test del DNA: [17], [10], [9, Cap. 7]. Per ulteriori approfondimenti: [6], [16].

numero di loci. Il numero di questi loci va crescendo col progredire delle tecniche (si è partiti da 3-5 loci; oggi lo standard dell'F.B.I. è 13 loci).

5. Supponiamo, per fissare le idee, che in ogni locus ci siano 5 possibili alleli; allora in ogni locus ci sono $5 \cdot 6/2 = 15$ genotipi, e su un totale di 13 loci ci saranno $15^{13} = 1.946 \cdot 10^{15}$ possibili *profili*, cioè scelte di un genotipo per ogni locus. Notiamo che si tratta di 6 ordini di grandezza in più rispetto alle persone totali sul pianeta.

6. Se, per fare un primo ragionamento grossolano, supponiamo inoltre che i diversi profili siano equiprobabili, questo significa che *la probabilità che una persona scelta a caso abbia lo stesso profilo di quello del campione trovato sulla scena del crimine* è dell'ordine di 10^{-15}. Questo fatto viene considerato come un indicatore pressoché certo del fatto che una persona che ha lo stesso profilo di quello del campione sia l'effettivo "proprietario" della traccia lasciata. Comunque, come si vede, si tratta di probabilità e non di certezze; una probabilità che come vedremo è molto più difficile di quanto sembri quantificare.

Riguardo all'applicabilità del test, notiamo anzitutto che il DNA si trova nel nucleo cellulare[2], pertanto i resti che interessano sono quelli di cellule dotate di nucleo. Se si tratta di sangue, ad esempio, si possono usare i globuli bianchi, che hanno il nucleo, mentre i globuli rossi no; se si tratta di un capello, sono utili le cellule che circondano la radice, che sono provviste di nucleo, mentre quelle dello stelo non lo sono. Le tecniche odierne consentono di analizzare il DNA a partire da campioni piccolissimi, grazie ad un processo di amplificazione chiamato PCR (Polymerase Chain Reaction), capace di produrre, a partire dal DNA di poche cellule appartenenti ad una persona, miliardi di copie di un frammento definito a priori. I frammenti amplificati tramite PCR vengono poi separati tramite migrazione in un campo elettroforetico e la loro lunghezza viene determinata. Quindi il profilo genetico su (ad esempio) 13 loci può essere determinato da piccolissimi resti[3].

Vale anche la pena di soffermarsi sul fatto che, proprio perché il test del DNA si basa sull'analisi di porzioni non codificanti del genoma, i risultati di questo test non contengono alcuna informazione *somatica* sulla persona. Ciò significa che il profilo ricavato dalla traccia di DNA trovata sulla scena del crimine non è di alcun aiuto nel cercare il colpevole a partire dal suo aspetto (ad esempio, *non può* dirci che "il colpevole è un uomo bianco con gli occhi azzurri"), ma serve solo per eseguire un confronto con una persona già individuata con altri criteri, oppure per fare una ricerca in un database del DNA[4].

Prima di proseguire possiamo ora fissare qualche data che inquadra storica-

[2] Per la verità esiste anche il *DNA mitocondriale*, che si trova anche nelle cellule prive di nucleo, su cui si basa una diversa tecnica di identificazione. In questo articolo tuttavia ci riferiremo esclusivamente al test eseguito sul DNA del nucleo.

[3] Ad esempio la radice di un capello, una goccia di sangue, o anche le minuscole scaglie di pelle che lasciamo su un oggetto quando lo afferriamo.

[4] L'unica informazione somatica che si ricava facilmente dal DNA è se il suo possessore sia maschio o femmina.

mente il discorso[5]. Come è ben noto, la scoperta della struttura del DNA risale al 1953 e si deve a James Watson e Francis Crick, che ricevettero per questo il premio Nobel per la medicina e fisiologia nel 1962 (insieme a Maurice Wilkins). È solo a partire dagli anni 1980, invece, che si pongono le basi per l'utilizzo del test del DNA:

nel 1985 sulla rivista "Nature" appaiono i lavori [11], [12] di A. Jeffreys, V. Wilson, S. Thein (Leicester University e John Radcliffe Hospital di Oxford, U.K.) in cui viene messo a punto il metodo di identificazione che abbiamo descritto sopra, battezzato "DNA fingerprinting"; il merito della scoperta è comunemente attribuito al primo autore, Alec Jeffreys;

nel 1986 Kary Mullis (v. [14]) scopre la Polymerase Chain Reaction (PCR), sopra descritta come metodo per amplificare piccole porzioni di DNA, permettendo di eseguire il test anche a partire da piccoli resti; per questa scoperta -le cui applicazioni sono molto più vaste di quelle che qui ci interessano- gli viene assegnato il premio Nobel per la chimica nel 1993;

nel 1987 (U.K.) si ha il primo caso di omicidio risolto in base al test del DNA;

nel 1995 in U.K. (più precisamente in Inghilterra, Galles, Scozia) viene stabilito il primo database nazionale del DNA;

nel 1998 in U.S.A. viene stabilito un database nazionale del DNA, permettendo all'F.B.I. di comparare elettronicamente i profili di DNA;

a partire dal 1997 sono stati stabiliti database nazionali del DNA in molti paesi europei, tra cui[6]:

Olanda, Austria (1997), Francia, Germania, Cipro (1998), Belgio, Finlandia (1999), Svezia, Danimarca (2000), Repubblica Ceca (2001), Lituania (2002), Estonia, Slovacchia, Ungheria (2004), Lussenburgo (2006).

2 Problemi probabilistici legati al test del DNA

Un concetto chiave che abbiamo già incontrato nella sezione precedente è quello di *random match probability* (RMP), definita come *la probabilità che una persona scelta a caso in un certo universo abbia lo stesso profilo DNA rispetto ad un profilo prefissato*. L'affermazione fatta in precedenza che la RMP vale circa 10^{-15} si basa su ipotesi grossolane e sarà ridiscussa in seguito. Per ora interessa solo fissare questo concetto, a prescindere dal suo valore quantitativo, e l'idea che si tratti comunque di un numero molto piccolo. Come vedremo, il valore effettivo della RMP varia da profilo a profilo, e può assumere valori diversi anche di molti ordini di grandezza.

Supponiamo ora che il test del DNA dica che il sig. Rossi ha lo stesso profilo

[5] Una cronologia schematica -con qualche data in più rispetto a queste- si trova ad es. in [24]. Per riferimenti storici più ampi, invece, si veda ad es. [8].

[6] Questo elenco è aggiornato fino al 2007. La fonte è [26], che contiene un'approfondita discussione delle normative e prassi europee su questo tema, in particolare dal punto di vista dei problemi etici coinvolti. L'anno indicato talvolta si riferisce all'anno della legge che ha istituito il database; l'effettiva realizzazione può essere successiva.

di quello del campione trovato sulla scena del crimine (diremo brevemente nel seguito: "è positivo al test del DNA"). Le due domande cruciali sono allora:

1. Qual è la probabilità che la traccia trovata sulla scena del crimine appartenga effettivamente al sig. Rossi (e non solo che i due frammenti di DNA coincidano in un certo numero di loci prefissati), se questi è risultato positivo al test?

2. Che conseguenza (processuale ecc.) si può trarre dalla positività al test?

Il seguito del discorso si concentrerà sulla prima domanda, ma facciamo almeno un paio di osservazioni sulla seconda.

Osservazione 1 (Primo disclaimer) *Ci sono resistenze all'uso della probabilità in campo processuale. Si può condannare una persona "solo perché" è molto probabile che abbia commesso il crimine? Solitamente ci si aspetta che una valutazione di colpevolezza sia sostenuta da elementi certi, rispetto ai quali le valutazioni di probabilità sono solo complementari. Ci sono però casi in cui l'esito del test del DNA potrebbe essere l'unico elemento a disposizione, e dobbiamo decidere se fidarci di un'argomentazione probabilistica o no.*

Osservazione 2 (Secondo disclaimer) *Anche ammesso che la traccia trovata sulla scena del crimine sia collegabile con certezza a una persona, ne segue la colpevolezza? Come abbiamo già ricordato, la tecnica di analisi consente di considerare come traccia anche un solo capello o minuscole scaglie di pelle. Il test non dice quando e come queste tracce sono finite sul luogo del crimine. Queste sono situazioni molto diverse, e molto meno probanti, rispetto a quelle in cui, ad esempio, la traccia fosse del sangue lasciato sulla vittima.*

Per non parlare dell'eventuale contaminazione della scena del crimine dovuta a resti lasciati involontariamente da chi scopre il fatto, i tecnici di polizia, ecc. Se non ci si accorge di chi è il vero "proprietario" di quella traccia, si rischia di scagionare il vero colpevole -ammesso che lo si sia trovato con un'indagine indipendente dal DNA-, perché il suo profilo non coincide con quello della traccia. Occorre riflettere quindi anche sull'affermazione, spesso data per scontata, secondo cui il test del DNA "forse non dà la certezza di colpevolezza, ma può dare la certezza di innocenza". Quest'ultima affermazione è vera solo se nel raccogliere le tracce si è seguito un protocollo molto rigoroso di precauzioni.

Chiudiamo queste osservazioni sulle interpretazioni del test, di cui non ci occuperemo oltre. A dispetto di queste doverose precisazioni, nel seguito diremo brevemente che "il sig. Rossi è colpevole" per indicare che è stato lui a lasciare la traccia trovata sulla scena del crimine. Quindi il nostro problema è rispondere alla domanda:

Qual è la probabilità che il sig. Rossi sia colpevole (o innocente), sapendo che è risultato positivo al test del DNA?

Si intende che per eseguire il calcolo dev'essere noto -tra le altre cose- il numero di loci (ad es. 13) su cui è stato eseguito il test. Vediamo ora alcuni problemi specifici che hanno a che fare con questa domanda.

2.1 Problema 1. Le probabilità condizionate inverse

La probabilità che il sig. Rossi sia innocente, sapendo che è risultato positivo al test del DNA, sembra essere parente stretta della RMP introdotta in precedenza; per lo meno, il buon senso dice che se la RMP è un numero piccolissimo, dovrebbe essere piccolissima anche la probabilità di innocenza di una persona risultata positiva al test.

Tuttavia, questo è un tipico esempio di confusione (esplicita o implicita) tra le due probabilità condizionate[7] $P(A|B)$ e $P(B|A)$: indicando con R il sig. Rossi,

$$P(\text{R è innocente}|\text{R è positivo al test}) \neq P(\text{R è pos. al test}|\text{R è innoc.}).$$

La seconda probabilità coincide con la RMP: se Rossi è innocente, lui è solo un individuo scelto a caso che è risultato positivo al test senza essere il "proprietario" di quella traccia di DNA. La prima probabilità è invece quella che ci interessa; o meglio: il complemento a 1 di questa probabilità è la probabilità di colpevolezza che ci interessa.

In vari scritti sull'argomento (ad es. [6], [10]), una RMP molto piccola viene data come forte indicatore di colpevolezza, senza ulteriori discussioni circa le probabilità condizionate. Non necessariamente questi scritti contengono delle falsità: in generale, si guardano bene dal fare affermazioni forti del tipo "questa è la probabilità che il sig. Rossi sia colpevole", tuttavia ciò che non viene detto esplicitamente viene almeno implicitamente suggerito. Questo è un punto importante che si dovrà discutere.

2.2 Problema 2. Il calcolo della RMP

La RMP, anche se non è il nostro obiettivo ultimo, è comunque un numero significativo, come vedremo; un secondo problema è quindi: *come si calcola la RMP*? Occorre chiedersi: per ogni locus quanti alleli ci sono? Con quali frequenze si presentano? Su quale popolazione di riferimento calcoliamo queste frequenze? Siamo certi che il presentarsi di alleli diversi in loci diversi siano eventi indipendenti?

2.3 Problema 3. Banche dati del DNA, "colpo a freddo" e Database Match Probability

Poiché, anche a prescindere da valutazioni quantitative precise, è opinione condivisa che il test del DNA sia uno strumento potente, è naturale suggerire l'istituzione di un database del DNA. Negli USA dal 1998 esiste l'archivio dei profili genetici dell'F.B.I., detto CODIS (combined DNA index system), basato su 13

[7] Ricordiamo che la probabilità condizionata $P(A|B)$ è definita da $P(A \cap B)/P(B)$ (nell'ipotesi $P(B) \neq 0$) e si interpreta come la probabilità che si realizzi l'evento A, sapendo che l'evento B è certamente verificato.

loci. Questa banca dati, oggi la più grande al mondo[8], contiene varie categorie di archivi tra cui: detenuti; archivio medico-legale di campioni prelevati da scene di crimini; resti umani non identificati; parenti volontari di persone scomparse (l'idea è che il DNA di parenti stretti è simile, e si vuole identificare il cadavere di un congiunto scomparso che venisse ritrovato).

Ora, proviamo a confrontare queste due situazioni:

A. È stato commesso un crimine e c'è una traccia di DNA; indagini condotte con metodi tradizionali portano a sospettare pesantemente di una certa persona; questa viene arrestata e viene fatto l'esame del DNA; risulta positivo; viene incriminato.

B. È stato commesso un crimine e c'è una traccia di DNA; non c'è alcun sospettato; viene setacciato un grande database del DNA e si trova una e una sola persona nel database che è positivo al test; la persona in questione viene arrestata e incriminata.

La seconda situazione viene chiamata in gergo "cold hit" (colpo a freddo): qualcuno, che fino al giorno prima non era neppure indagato, viene arrestato e incriminato, in base alla ricerca in un database: costui viene "colpito a freddo", per l'appunto.

Ora, vari ragionamenti di buon senso portano a ritenere le due situazioni molto diverse tra loro. Si sostiene che nel caso B la probabilità di colpevolezza sia molto minore che nel caso A, e tanto minore quanto più ampio è il database considerato; al punto che esistono casi giudiziari reali in cui nel caso B non si è arrivati a una condanna, e la giustizia sembra in stallo. (Si veda il caso Jenkins a cui dedicheremo un paragrafo in seguito).

Un concetto rigoroso legato a questa situazione è quello di *database match probability (DMP)*, definita come la probabiltà che in un database fissato ci sia almeno un individuo che risulta positivo al test (nel raffronto con un campione fissato una volta per tutte).

Un documento del National Research Council degli USA del 1996 (v. [7]), su cui ritorneremo in seguito, contiene le seguenti raccomandazioni, relative all'uso processuale di questi dati probabilistici:

a. Nel caso A, bisogna dire alla giuria qual è la RMP;

b. Nel caso B, bisogna dire alla giuria qual è la DMP, ottenuta moltiplicando la RMP per l'ampiezza m del database.

Una possibile giustificazione della formula di calcolo della DMP contenuta nella raccomandazione b è la seguente: se p è la random match probability e m l'ampiezza del database, in uno schema Bernoulliano di prove indipendenti ripetute si dovrebbe avere:

$$DMP = 1 - (1-p)^m \simeq mp,$$

[8] Alla pagina web del CODIS [21] si trova il dato aggiornato del numero totale di profili (in costante aumento) contenuto in questo archivio: a maggio 2009 erano oltre 7 milioni di profili di pregiudicati (per confronto, solo 6 mesi prima erano mezzo milione in meno). Il paese che ha il più numeroso database del DNA, non in assoluto ma in proporzione alla propria popolazione, è comunque il Regno Unito, dove il 5,2% dei cittadini è nel database (v. [22]).

almeno se mp è piccolo. A sua volta, lo schema Bernoulliano è giustificato dal fatto che stiamo raffrontando ciascun profilo del database, indipendentemente, con un profilo "esterno", fissato una volta per tutte; ogni volta che scegliamo dal database un profilo, la probabilità a priori che esso coincida col profilo fissato è p =RMP.

2.4 Problema 4. Le probabilità condizionate inverse nel caso della ricerca in un database

Proseguiamo la discussione sulla "situazione B" descritta nel punto precedente ("colpo a freddo"). A prescindere da come si calcoli la DMP, siamo sicuri che questa sia la quantità pertinente a valutare la probabilità che ci interessa? La nostra domanda è:

Qual è la probabilità che il sig. Rossi sia innocente, sapendo che è l'unico che è stato trovato positivo al test tra tutti gli individui di un database di m individui?

Posta così la domanda, non sembra molto diversa dalla situazione A: è vero (ed è ovvio!) che più ampio è il database più è facile trovare in esso un riscontro, ma il punto non è il fatto che sia stato trovato *un riscontro qualsiasi*, ma *un riscontro col sig. Rossi*: perché proprio lui, se è innocente? In seguito esamineremo questo problema dal punto di vista delle probabilità condizionate.

2.5 Problema 5. Il paradosso del database dell'Arizona

La difficoltà a fare calcoli precisi di probabilità, nelle situazioni che stiamo descrivendo, unita alla nostra difficoltà nel figurarci intuitivamente che cosa può succedere quando sono in gioco numeri molto grandi o molto piccoli, fa sì che certi fatti statistici empiricamente osservati ci sembrino paradossali. Per alcune persone, già inclini per mentalità a gettare un'ombra di dubbio sul potere predittivo del calcolo delle probabilità e del test del DNA, questi paradossi sono la dimostrazione evidente della fallibilità del metodo.

Un esempio vistoso di questo fatto è il seguente: nel 2005 la banca dati DNA dei prigionieri dell'Arizona conteneva 65000 profili realizzati sulla base di 13 loci. Un'analisi del database rivelò che 144 individui avevano profili corrispondenti in 9 loci (si intende: 9 loci qualsiasi su 13, a priori non prefissati); un altro piccolo gruppo aveva una corrispondenza in 10 loci, due profili coincidevano in 11 loci e altri due avevano 12 loci identici.[9] Ora, queste frequenze ci sembrano esageratamente grandi rispetto alla nostra aspettativa di buon senso. Se, per fare una stima grossolana, supponiamo che la probabilità che due individui diversi abbiano uno stesso genotipo in uno specifico locus sia pari a 1/10 ([10]) la RMP per profili di 9 loci sarebbe 10^{-9}. Supponiamo, per semplificare le cose, che

[9] Citato in [9], p.114.

[10] Si tratta di un ordine di grandezza ragionevole. In seguito calcoleremo il valore più accurato 0.0744.

il database dell'Arizona contenesse profili di 9 loci, o che comunque siano stati ispezionati solo i primi 9 loci, ad esempio (il che in realtà cambia il problema rispetto a come l'abbiamo formulato, ma aiuta a fissare le idee); la probabilità di avere almeno due profili uguali su 65000 individui sarebbe (nell'approssimazione bernoulliana descritta nel par. 2.3, che è quella suggerita in [7]) dell'ordine di

$$65000 \cdot 10^{-9} = 6.5 \cdot 10^{-5},$$

e l'eventualità di trovarne addirittura 144 sembra irreale[11].

In realtà, questa analisi si basa su un insidioso quanto grossolano errore: la confusione tra la *probabilità che almeno due profili nel database siano uguali tra loro* con la *probabilità che almeno un profilo nel database sia uguale ad un profilo esterno, fissato una volta per tutte*. Che si tratti di due numeri diversi, e che il primo possa essere molto maggiore del secondo, si capisce subito pensando che se, come caso limite, l'ampiezza del database fosse maggiore del numero totale dei possibili profili[12], potrebbe ancora succedere che nessuno di essi sia uguale ad un profilo esterno fissato, ma certamente almeno due profili nel database dovrebbero essere uguali tra loro. Può essere utile anche il prossimo semplice esempio numerico:

Esempio. Una variabile può assumere 10 valori diversi $x_1, x_2, ..., x_{10}$, equiprobabili. La probabilità che, su 5 osservazioni indipendenti, ce ne sia almeno una in cui è assunto il valore x_1 è:

$$1 - \left(1 - \frac{1}{10}\right)^5 \simeq 0.4.$$

Invece, la probabilità che, su 5 osservazioni indipendenti, ce ne siano almeno 2 uguali tra loro è:

$$1 - \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{10^5} \simeq 0.7.$$

Dunque la coincidenza di almeno due (o almeno k) profili nel database "tra loro" è un evento di probabilità ben maggiore della coincidenza con un profilo esterno. Se poi si passa al problema di trovare corrispondenze in 9 loci *qualsiasi* su 13, la probabilità di trovare almeno due (o almeno k) profili nel database con questa corrispondenza parziale dovrebbe aumentare parecchio; tuttavia il calcolo della probabilità esatta diventa molto difficile (ne riparleremo), per cui non si arriva ad un numero da confrontare con la constatazione empirica. Rimane dunque spazio per l'incredulità, o per la domanda: c'è qualcosa nelle nostre ipotesi che non funziona? E cosa?

[11] Il calcolo fatto con la formula più precisa $1 - (1-p)^m$ anziché mp non dà differenze apprezzabili.

[12] Questo è praticamente impossibile dato il numero elevatissimo di profili possibili, ma il ragionamento è comunque sufficiente a dimostrare che è stata fatta una confusione tra due concetti diversi.

2.6 Problema 6. Il test in due passi su loci indipendenti

I dubbi sollevati sulla probabilità di colpevolezza nel caso di "cold hit" hanno suggerito un astuto correttivo al metodo da seguire in questo caso, correttivo che diventa sempre più praticabile, con la crescente facilità di esaminare numerosi loci.

Supponiamo, dunque, che la ricerca su un database abbia fornito una e una sola corrispondenza tra il campione di riferimento e un'individuo del database e che questa corrispondenza sia stata ottenuta con un raffronto su, poniamo, 8 loci prefissati. Ora abbiamo un indiziato ben preciso; lo trattiamo come tratteremmo una persona su cui sono caduti i sospetti per motivi legati all'indagine, indipendenti dal DNA: lo arrestiamo e gli facciamo l'esame del DNA, ma su *altri* 5 loci, diversi da quelli che sono stati usati fin qui. Questo presuppone che il campione prelevato sulla scena del crimine sia sufficientemente buono da offrirci $8 + 5$ loci da utilizzare. (I numeri 8 e 5 sono solo un esempio, naturalmente). Ora il confronto su questi 5 loci è fatto su un singolo indiziato, non setacciando l'intero database; se otteniamo ancora coincidenza in 5 loci su 5, la nostra fiducia dovrebbe essere paragonabile, qualitativamente, a quella ottenuta nel caso del raffronto su un unico indiziato. Quantitativamente, avremo in generale meno loci di quelli che avremmo solitamente, ma questo è un problema superabile con la tecnica. L'idea sembra buona. È stata raccomandata nel primo rapporto steso dal National Research Council americano su questi temi, nel 1992 (v. [6]). Il secondo rapporto, del 1996 (v. [7], già citato, che chiameremo NRC2), tuttavia, si espresse negativamente su questo modo di procedere, commentandolo così:

"Tale procedura è sensata, ma spreca informazioni, e se vengono usati troppi loci per l'identificazione del sospetto, potrebbero non rimanerne abbastanza per un'adeguata analisi successiva. Una seconda procedura consiste nell'applicare una semplice correzione: moltiplicare le probabilità di corrispondenza (RMP) per la dimensione della banca dati esaminata. Questo è il metodo che raccomandiamo"[13]. A mio parere, un passo indietro.

Eseguiremo in seguito qualche calcolo di probabilità legate a questo procedimento.

2.7 Intermezzo. Il caso Jenkins

Esponiamo in estrema sintesi un caso giudiziario esemplare dei problemi processuali legati all'uso del calcolo delle probabilità in un caso di "cold hit". Di questo caso e del relativo dibattito si trova ampia documentazione sia in letteratura che in rete[14].

Il 4/6/1999 il sig. Dolinger viene assassinato nella sua casa a Washington a colpi di pugnale. Abiti insanguinati vengono trovati in casa. Le indagini portano a sospettare del sig. Watson, su cui pesano vari indizi. Ma il test del DNA lo scagiona, e viene rilasciato. L'F.B.I. inizia a ricercare corrispondenze tra il DNA del sangue trovato sulla scena del crimine e il database CODIS, ma

[13] Citato in [9], p. 111.

[14] Per maggiori dettagli su questa vicenda si veda ad esempio [13, § I.A.], [9, pp.99 sgg.].

la ricerca dà esito negativo. Sei mesi dopo, novembre 1999, il profilo genetico del campione di sangue viene mandato al dipartimento di medicina legale dello stato della Virginia, che lo confronta col suo database di 101905 profili di criminali. Questa volta si trova una corrispondenza, che riguarda solo 8 loci dei 13 della banca dati CODIS, perché la banca dati della Virginia, più vecchia, archiviava solo i dati di quegli 8 loci. Si risale alla persona corrispondente al campione, sig. Jenkins. Costui viene trovato, e a fine dicembre 1999 si esegue un nuovo test del DNA, questa volta su tutti i 13 loci del database CODIS: gli 8 del database della Virginia più altri 5. Siamo quindi in un caso di test in due passi su loci indipendenti, come descritto sopra. La corrispondenza è su tutti i 13 loci. L'F.B.I. stima che la RMP relativa al profilo di Jenkins[15] è dell'ordine di 10^{-18}. Sulla base di questa informazione il 13 gennaio 2000 Jenkins è arrestato.

Tuttavia, in seguito cominciano le obiezioni al metodo. Jenkins è stato individuato inizialmente sulla base di una corrispondenza su 8 loci in un database di circa 100000 profili. Il calcolo fatto in base alle raccomandazioni del NRC2 dà per la RMP su 8 loci[16] $p = 10^{-8}$, e quindi per la DMP su un database di $n = 10^5$ profili,

$$DMP = 10^5 \cdot 10^{-8} = 0.001.$$

Questo numero, implicitamente considerato la probabilità di coincidenza fortuita, sembra essere troppo grande per dare la certezza di colpevolezza. L'esito è che, a tutto il 2008, il processo contro Jenkins era ancora pendente. Torneremo in seguito a discutere queste conclusioni.

3 Qualche risposta

Affronteremo ora quantitativamente i vari problemi posti in precedenza. A volte arriveremo a qualche risposta, a volte ci renderemo semplicemente conto più da vicino della complessità del problema.

3.1 Le probabilità condizionate inverse

Trattiamo per prima cosa la "situazione A" descritta nel par. 2.3: una e una sola persona, già sospettata per motivi indipendenti dall'esame del DNA, viene sottoposta a test e si ottiene la corrispondenza col DNA del campione trovato sulla scena del crimine, in 13 loci su 13 (o un altro numero che possiamo ipotizzare). L'altra situazione (ricerca nel database) sarà discussa nei par. 3.4 e 3.5.

[15] Questo numero sembra esageratamente piccolo ma, come vedremo nella seconda parte, è verosimile.

[16] Questo valore è in grave disaccordo con quello sopra riportato di 10^{-18}. In parte il disaccordo si deve al diverso significato dei due numeri: il primo è la RMP di uno *specifico* profilo (che può essere particolarmente raro), il secondo è un valore standardizzato che vuole essere rappresentativo di un valore "tipico". In secondo luogo, le raccomandazione del NRC sono prudenti in senso garantista, quindi tendono a "gonfiare" i valori della RMP (che in un certo senso è una probabilità di innocenza).

Sia J il sospettato. Diciamo "J è colpevole" per intendere "Il DNA trovato sulla scena del crimine appartiene a J ". (Come già detto nell'introduzione, le due cose non coincidono, ma il ragionamento probabilistico può applicarsi solo a questa affermazione).

Sia p la RMP della popolazione, cioè la probabilità che una persona scelta a caso nella popolazione considerata abbia il DNA coincidente con quello trovato sulla scena del crimine (che ora *non* supponiamo noto, ma anch'esso generico), in tutti i loci considerati. Questo numero si può vedere anche come la probabilità che due persone scelte a caso nella popolazione abbiano il DNA coincidente in tutti i loci considerati. Per ora non vogliamo calcolare p ma, supponendola nota, trarne delle conseguenze.

Si hanno le seguenti probabilità condizionate:

$$P\left(J \text{ positivo al test}|J \text{ innocente}\right) = p;$$
$$P\left(J \text{ positivo al test}|J \text{ colpevole}\right) = 1.$$

Stiamo supponendo che J sia già sospettato per motivi indipendenti dal test; questo significa che la probabilità che J sia colpevole, prima di eseguire il test, è $c > 0$. Rispetto all'esperimento aleatorio "eseguire il test e confrontare il profilo ottenuto con quello della traccia", c è la *probabilità a priori* che J sia colpevole. Se ora si esegue il test e si trova che J è positivo, il teorema di Bayes dà, indicando con "J pos." l'evento "J è positivo al test":

$$\begin{aligned} &P\left(J \text{ innoc.}|J \text{ pos.}\right) = \\ &= \frac{P\left(J \text{ pos.}|J \text{ innoc.}\right) P\left(J \text{ innoc.}\right)}{P\left(J \text{ pos.}|J \text{ innoc.}\right) P\left(J \text{ innoc.}\right) + P\left(J \text{ pos.}|J \text{ colpev.}\right) P\left(J \text{ colpev.}\right)} = \\ &= \frac{p\left(1-c\right)}{p\left(1-c\right)+c}. \end{aligned}$$

Perciò la probabilità di colpevolezza sapendo che J è risultato positivo al test (che d'ora in poi chiameremo "probabilità di colpevolezza a posteriori" e indicheremo con π), è

$$\pi = 1 - \frac{p\left(1-c\right)}{p\left(1-c\right)+c} = \frac{1}{1+p\left(\frac{1}{c}-1\right)}. \tag{1}$$

Facciamo qualche esempio numerico:

Se $p = 10^{-13}$	e $c = 0.01$	$\pi \simeq 1 - 10^{-11}$
Se $p = 10^{-11}$	e $c = 0.01$	$\pi \simeq 1 - 10^{-9}$
Se $p = 10^{-8}$	e $c = 0.01$	$\pi \simeq 1 - 10^{-6}$
Se $p = 10^{-7}$	e $c = 0.1$	$\pi \simeq 1 - 10^{-6}$
Se $p = 10^{-7}$	e $c = 0.00001$	$\pi \simeq 1 - 10^{-2}$

Ricordiamo che siamo nell'ipotesi che J sia stato sottoposto al test del DNA in quanto già sospettato per altri motivi, il che significa che la probabilità a priori di colpevolezza, c, è ritenuta significativamente discosta da zero (ad es.

$c = 0.01$, $c = 0.1$); in questo caso, come si vede, la probabilità a posteriori di colpevolezza è elevatissima, per valori tipici di p. (Anzi, come vedremo in seguito p può essere anche molto più piccolo, e quindi π ancora più grande).

Valori di π sempre elevati, ma forse non tanto da eliminare "ogni ragionevole dubbio", si riscontrano solo se c è molto piccolo (cioè se non abbiamo indizi indipendenti dal test).

La formula (1) ci dice che per calcolare la probabilità di colpevolezza a posteriori del test occorre conoscere due quantità:

1) la RMP del profilo, dato ovviamente cruciale nel metodo del DNA, sul cui calcolo torneremo;

2) la probabilità a priori di colpevolezza, intesa come forza delle prove che avevamo prima di eseguire il test del DNA.

Il secondo punto è quello che in certi approcci a queste discussioni probabilistiche viene ignorato, e di cui il ragionamento mediante il teorema di Bayes mostra che occorre tener conto[17]. D'altro canto, questo è ragionevole: sappiamo che uno dei problemi legati al test del DNA sta nella differenza tra il caso in cui la persona sottoposta a test è sospettato a priori e il caso in cui non lo è; questa differenza cruciale, tuttavia, non è di tipo zero / uno, ma dipende quantitativamente dalla *forza delle prove a priori* contro il sospettato.

Proprio per la difficoltà di quantificare la probabilità a priori di colpevolezza, alcuni autori preferiscono rinunciare a calcolare la probabilità a posteriori (mediante il teorema di Bayes) e utilizzano indici diversi, come il *rapporto di verosimiglianza* tra l'ipotesi di colpevolezza e quella di innocenza. Una spiegazione e discussione di questo concetto, in cui qui non entreremo, si trova ad esempio in [2, sec. 4].

3.2 Il calcolo della RMP

3.2.1 La distribuzione degli alleli su una popolazione e il calcolo della RMP di un profilo specifico

Ricordiamo le domande che ci siamo poste, preliminari al calcolo della RMP: per ogni locus quanti alleli ci sono? Con quali frequenze si presentano? Su quale popolazione di riferimento calcoliamo queste frequenze? Siamo certi che il presentarsi di alleli diversi in loci diversi siano eventi indipendenti?

Quella che segue è una tabella della distribuzione degli alleli nel locus D16S539, presa dal database del CODIS (U.S.A.) (v. [21], [20]); gli alleli rappresentati sono 8, e le frequenze relative sono calcolate su 4 distinti campioni, corrispondenti a 4 gruppi etnici diversi (come si vede, le frequenze variano con l'etnia). L'ultima riga (es. "$2N = 586$") significa che il campione è di $N = 293$ persone, cioè $2N = 586$ alleli, perché ogni persona contribuisce alla statistica con entrambi gli alleli del genotipo che si trova nel locus; si noti che i due alleli della stessa persona provengono dai due genitori diversi, e quindi è corretto considerarli dati

[17] Uno scritto su questi argomenti in cui invece si dà ampio spazio al ragionamento bayesiano è [13].

indipendenti; naturalmente è essenziale che nel campione non ci siano parenti.

Locus D16S539. Cromosoma: coppia n° 16				
Quartetto di basi ripetuto: G, A, T, A				
ALLELE	**FREQUENZA**			
	Caucasian	**Black**	**Asian**	**East Indian**
8	0,02	0,02	0,01	0,07
9	0,12	0,22	0,27	0,16
10	0,05	0,13	0,11	0,09
11	0,29	0,29	0,27	0,33
12	0,35	0,17	0,23	0,20
13	0,15	0,15	0,10	0,14
14	0,03	0,02	0,01	0,03
15	0,00	0,00	0,003	0,00
2N	**586**	**358**	**390**	**334**

Tabella 1

Per la popolazione caucasica, ad es., il numero di alleli è 7 (si noti che l'allele 15 ha frequenza nulla), quindi il numero dei genotipi è $7 \cdot 8/2 = 28$.

A partire dalla tavola delle frequenze degli alleli, dobbiamo per prima cosa calcolare una tavola delle probabilità dei genotipi. Si procede così: se, ad esempio, l'allele A_1 ha probabilità p_1 e l'allele A_2 ha probabilità p_2, i genotipi che possono formare sono:

		Allele paterno	
		A_1	A_2
Allele materno	A_1	(A_1, A_1): p_1^2	(A_1, A_2): $p_1 p_2$
	A_2	(A_2, A_1): $p_2 p_1$	(A_2, A_2): p_2^2

Tabella 2

La regola del prodotto delle probabilità discende dall'indipendenza tra i contributi dei due genitori. In pratica (A_2, A_1) e (A_1, A_2) sono lo stesso genotipo (nelle coppie di cromosomi non è distinguibile quale sia quello donato dal padre e quale dalla madre), che però dev'essere pesato con probabilità $2p_1p_2$ (due caselle della tabella). Quindi la regola è:

in un locus eterozigote di alleli A_i, A_j : $2p_ip_j$;

in un locus omozigote di allele A_i : p_i^2.

In questo caso ad esempio otterremmo la seguente tabella di probabilità per

i 28 genotipi possibili:

Probabilità genotipi nel locus D16S539, polazione caucasica							
Alleli	8	9	10	11	12	13	14
8	0,0003						
9	0,0039	0,0135					
10	0,0017	0,0114	0,0024				
11	0,0099	0,0673	0,0284	0,0841			
12	0,0120	0,0817	0,0345	0,2042	0,1239		
13	0,0050	0,0343	0,0145	0,0858	0,1042	0,0219	
14	0,0009	0,0063	0,0026	0,0157	0,0190	0,0080	0,0007

Tabella 3

La variabilità va ora da un minimo di 0.0003 a un massimo di 0.2042.

Dal punto di vista di tutti i calcoli successivi, non ha nessuna importanza come si chiamano questi genotipi (es. $(8, 9)$ piuttosto che $(13, 10)$); conta solo che sono 28, e hanno quelle 28 probabilità.

Presentiamo ora la tabella delle frequenze alleliche per la sola popolazione caucasica, relativa a 9 loci[18] (per contenere i calcoli, consideriamo 9 loci anziché 13):

D3S1358		vWA		FGA		D8S1179		D21S11		D18S51		D5S818		D13S317		D7S820	
12	0,005	13	0,002	16	0,002	8	0,012	26	0,005	10	0,005	7	0,002	8	0,111	7	0,020
13	0,005	14	0,082	18	0,009	9	0,010	27	0,036	11	0,012	8	0,002	9	0,087	8	0,142
14	0,114	15	0,109	19	0,080	10	0,085	28	0,160	12	0,119	9	0,027	10	0,070	9	0,140
15	0,258	16	0,227	20	0,154	11	0,080	29	0,251	13	0,119	10	0,068	11	0,275	10	0,287
16	0,241	17	0,271	21	0,184	12	0,137	29,2	0,002	14	0,179	11	0,336	12	0,319	11	0,227
17	0,217	18	0,212	21,2	0,002	13	0,316	30	0,229	15	0,167	12	0,389	13	0,082	12	0,140
18	0,148	19	0,084	22	0,174	14	0,225	30,2	0,017	16	0,140	13	0,160	14	0,056	13	0,038
19	0,010	20	0,014	22,2	0,007	15	0,111	31	0,085	17	0,121	14	0,015			14	0,007
20	0,002			23	0,143	16	0,024	31,2	0,089	18	0,061						
				24	0,142			32	0,012	19	0,031						
				25	0,067			32,2	0,078	20	0,019						
				26	0,031			33,2	0,032	21	0,012						
				27	0,007			34	0,002	22	0,005						
								34,2	0,002	23	0,007						
										24	0,002						

Tabella 4

Notiamo che gli alleli relativi a loci diversi non vengono mai confrontati: l'allele 13 nel primo locus e nel secondo non hanno niente in comune, potremmo eliminare i nomi delle classi.

[18] Per informazioni più dettagliate su questi loci (collocazione sui cromosomi, gruppo di basi ripetute, ecc.) si rimanda a [5].

Osserviamo che il numero di alleli in ogni locus varia da un minimo di 7 a un massimo di 15, con frequenze relative comprese tra un minimo di 0.002 a un massimo di 0.389. C'è una grande variabilità, quindi. Il numero totale dei genotipi possibili per ciascun locus è $\frac{n(n+1)}{2}$ cioè:

$$45 \qquad 36 \qquad 91 \qquad 45 \qquad 105 \qquad 120 \qquad 36 \qquad 28 \qquad 36$$

e quindi il numero di profili possibili (su 9 loci) è

$$N = 45 \cdot 36 \cdot 91 \cdot 45 \cdot 105 \cdot 120 \cdot 36 \cdot 28 \cdot 36 = 3.03 \cdot 10^{15}. \tag{2}$$

Dovremmo ora, per ogni locus, costruire la tabella delle probabilità dei vari genotipi. A questo punto *il calcolo della probabilità di un particolare profilo di 9 loci si ottiene semplicemente moltiplicando tra loro le 9 probabilità dei genotipi corrispondenti, nell'ipotesi che le variabili aleatorie* "genotipo che si presenta nel k-esimo locus" (per $k = 1, 2, ..., 9$) *siano indipendenti.* (Questa è la cosiddetta "regola del prodotto").

Un simpatico calcolatore automatico della frequenza di uno specifico profilo è utilizzabile al sito [15]: si sceglie il database di riferimento (corrispondente a un gruppo etnico), si immette per ogni locus la coppia di valori degli alleli, e il calcolatore fornisce la RMP di quel profilo.

L'ipotesi di *indipendenza* è naturalmente cruciale. A questo proposito, cominciamo col dire che i 9 loci presi in considerazione nella Tabella 4 si trovano su 9 differenti coppie di cromosomi; dal punto di vista del meccanismo biologico (meiosi[19]) che sta all'origine della variabilità casuale del patrimonio genetico trasmesso da genitore a figlio, non ci dovrebbe essere alcuna relazione, a priori, tra questi loci. (Si dice che il *linkage genetico* è nullo). I 13 loci presi in considerazione dal CODIS si trovano su *dodici* diverse coppie di cromosomi: i loci D5S818 e CSF1P0 si trovano entrambi sulla coppia n° 5; tuttavia sono separati tra loro da circa 26.3 milioni di basi, laddove una distanza di un milione di basi è considerata sufficiente a rendere trascurabile il linkage genetico (v. [5, p. 256]). Dal punto di vista dei meccanismi biologici di base è quindi ragionevole aspettarsi l'indipendenza tra il genotipo di questi loci.

Una verifica statistica, a posteriori, dell'indipendenza tra i loci non è così agevole, tuttavia. Per eseguire test statistici di indipendenza occorrerebbe un database di dati disaggregati, cioè contenente per ogni individuo esaminato il profilo dei 9 loci (e non solo le frequenze relative degli alleli locus per locus) estremamente numeroso: anche pensando di testare l'indipendenza dei loci *a due a due* (e non l'indipendenza simultanea dei 9 loci), si tratta di costruire una tabella di contingenza che, nel caso ad esempio dei primi due loci della tabella, ha $45 \cdot 36 = 1620$ classi; affinché il test sia significativo ogni classe deve contenere almeno 5 individui, quindi occorre un database di almeno 8000

[19] Nella meiosi una cellula con corredo cromosomico diploide (cioè con 23 coppie di cromosomi) dà origine a quattro cellule con corredo cromosomico aploide (cioè ciascuna con 23 cromosomi singoli). Ogni cromosoma singolo nelle 4 cellule figlie è ottenuto ricombinando fra loro porzioni dei due cromosomi corrispondenti della cellula madre.

individui. Inoltre molti genotipi hanno frequenze relative inferiori all'1%, che porta a classi di contingenza con frequenza relativa inferiore allo 0.01%, e quindi un database di almeno 50000 individui. Invece, le statistiche di questo tipo sono fatte solitamente su database di qualche centinaio o poche migliaia di individui. Ma c'è anche un'altra ragione, che spiegheremo fra poco, per cui è difficile pensare di testare l'indipendenza con metodi statistici.

L'indipendenza è solitamente[20] considerata verificata per una popolazione abbastanza omogenea dal punto di vista etnico, mentre certamente non è verificata, a rigore, per una popolazione multietnica.[21]

Facciamo un esempio numerico scolastico, ma sufficiente a spiegare l'idea:

Popolazione A

		Locus 1		
		Genotipo 1a	Genotipo 1b	
Locus 2	Genotipo 2a	1/4	1/4	1/2
	Genotipo 2b	1/4	1/4	1/2
		1/2	1/2	

Popolazione B

		Locus 1		
		Genotipo 1a	Genotipo 1b	
Locus 2	Genotipo 2a	4/9	2/9	2/3
	Genotipo 2b	2/9	1/9	1/3
		2/3	1/3	

Popolazione A unione B (ugualmente numerose)

		Locus 1		
		Genotipo 1a	Genotipo 1b	
Locus 2	Genotipo 2a	0,3056	0,2361	0,54
	Genotipo 2b	0,2361	0,1806	0,42
		0,54	0,42	

Frequenze di A unione B nell'ip. di indipendenza

		Locus 1	
		Genotipo 1a	Genotipo 1b
Locus 2	Genotipo 2a	0,2916	0,2268
	Genotipo 2b	0,2268	0,1764

Tabella 5

In ciascuna delle popolazioni A e B le variabili "genotipo nel locus 1" e "genotipo nel locus 2" sono indipendenti; tuttavia, le frequenze relative sono molto diverse nelle due popolazioni; nella popolazione $A \cup B$ (A, B ugualmente numerose) le due variabili non sono più indipendenti. D'altro canto, la discrepanza tra la frequenza reale e quella che si avrebbe (con le stesse marginali) nell'ipotesi di indipendenza non è numericamente molto elevata. Questo significa che una simile discrepanza sarebbe difficilmente messa in evidenza da un test statistico condotto *sulla popolazione complessiva*. Il modo corretto di evidenziare la diversa frequenza relativa nelle diverse sottopopolazioni è quello di campionare separatamente le sottopopolazioni (che sospettiamo a priori siano diverse).

Nel seguito darò per scontata l'indipendenza dei loci, e quindi la regola del prodotto. Questo equivale a ragionare all'interno di una popolazione omogenea

[20] si veda ad es. [21].

[21] L'interessante articolo [16, v. sec. 3 e 8] contiene un'ampia discussione del problema dell'indipendenza. La tesi di quel lavoro è che gli studi accurati fatti per verificare l'indipendenza o non hanno dato motivo di dubitarne, o hanno mostrato delle deviazioni dall'indipendenza che hanno scarsa rilevanza dal punto di vista dei risultati numerici.

(ad esempio, un certo gruppo etnico degli U.S.A.). Naturalmente questo significa che la RMP così calcolata ha il significato di probabilità che una persona scelta a caso in *quella* popolazione abbia quel prefissato profilo (e questo, indipendentemente dal fatto che la persona che ha lasciato la traccia appartenga o meno a quella popolazione).

Se non vogliamo o non possiamo dare per scontata l'omogeneità della popolazione, possiamo applicare dei correttivi prudenziali (cioè garantisti), e approssimare le valutazioni di probabilità *per eccesso* (infatti la RMP è una probabilità di *innocenza*). In [6] si suggerisce la prassi seguente. Ogni volta che si vuole calcolare la RMP di uno specifico profilo, tutte le frequenze vanno arrotondate verso l'alto in base a certi criteri (dette "ceiling principle"):

1) aumentare la frequenza allelica q in base alla stima dell'intervallo di confidenza[22] al 95% per q;

2) prendere il massimo tra il valore trovato e quello convenzionale di 0.05 (assunto come minimo ragionevole per la frequenza allelica);

3) se si vuole confrontare il sospettato con la popolazione complessiva (di tutte le razze) si assume come frequenza di riferimento per quell'allele quella massima tra i vari gruppi razziali, prima di applicare i ragionamenti 1) e 2).

Queste precauzioni generano valori molto più grandi per la probabilità di un profilo specifico, rispetto a quelli calcolati applicando semplicemente la "regola del prodotto".

3.2.2 Calcolo della RMP per una popolazione

Se non stiamo ragionando su un caso specifico (cioè non abbiamo un profilo fissato con cui confrontare gli altri), ma vogliamo fare ragionamenti di validità generale per una certa popolazione, la RMP che ci interessa è la probabilità che due persone scelte a caso nella popolazione abbiano lo stesso profilo, o analogamente, la probabilità che una persona scelta a caso abbia lo stesso profilo di un profilo fissato (ma a noi ignoto, e scelto anch'esso a caso).

Consideriamo, come nella sezione precedente, profili di 9 loci. Ragioniamo con il teorema delle probabilità totali. La probabilità p che due profili scelti a caso siano uguali è, indicando con "$I = II$" l'evento "il primo profilo scelto a caso è uguale al secondo profilo scelto a caso" e con "$II = h$" l'evento "il secondo profilo scelto a caso è uguale all' h-esimo profilo", nell'elenco degli $N = 3.03 \cdot 10^{15}$ possibili (v. (2)):

$$p = \sum_{h=1}^{N} P\left(I = II | II = h\right) \cdot P\left(II = h\right).$$

[22] Per chi non conosce il concetto di intervallo di confidenza, rinunciando all'idea di spiegarlo in poche righe, mi limiterò a dire quanto segue. Il valore q della frequenza relativa di un certo allele (da cui dipende il calcolo della probabilità dei genotipi e quindi dei profili) in realtà non è noto sulla popolazione complessiva, ma è stimato da un piccolo campione. Ciò che la statistica permette di dire, in base a un certo dato campionario, è, ad es. che "il valore vero di q, con una confidenza del 95%, è minore di 0.15" In tal caso, qui si suggerisce di assumere $q = 0.15$.

Indichiamo ora con:

$$p_j^i, \text{ per } j = 1, 2, ..., 9; i = 1, 2, ..., k_j$$

(con k_j numero di genotipi del j-esimo locus) la probabilità dell'i-esimo genotipo nel j-esimo locus. Un profilo, individuato da una 9-upla

$$(i_1, i_2, ..., i_9) \text{ con } i_j \in \{1, 2, ..., k_j\}$$

ha probabilità

$$q(i_1, i_2, ..., i_9) = \prod_{j=1}^{9} p_j^{i_j}.$$

Notiamo che se l' h-esimo profilo è individuato da $(i_1, i_2, ..., i_9)$, si ha

$$P(I = II | II = h) = P(I = h) = q(i_1, i_2, ..., i_9),$$

perciò

$$p = \sum_{\substack{i_1=1,...,k_1 \\ i_2=1,...,k_2 \\ \cdots \\ i_9=1,...,k_9}} q(i_1, i_2, ..., i_9)^2 = \sum_{\substack{i_1=1,...,k_1 \\ i_2=1,...,k_2 \\ \cdots \\ i_9=1,...,k_9}} \left(\prod_{j=1}^{9} p_j^{i_j} \right)^2. \tag{3}$$

Scritta così, la formula precedente non è agevole per il calcolo effettivo (si noti il numero elevato di addendi che contiene). Il seguente ragionamento probabilistico permette di riscriverla in forma più semplice[23].

Definiamo la variabile aleatoria (v.a.)

$$X_j = \text{probabilità del genotipo che si presenta nel } j\text{-esimo locus di un profilo scelto a caso}$$

(per $j = 1, 2, ..., 9$); sia

$$X = \prod_{j=1}^{9} X_j = \text{probabilità di un profilo scelto a caso.}$$

La v.a. X assume i valori possibili $q(i_1, i_2, ..., i_9)$, ciascuno con probabilità $q(i_1, i_2, ..., i_9)$; dunque confrontando con (3) vediamo che

$$p = EX$$

(valore atteso di X). D'altro canto le v.a. $X_1, X_2, ..., X_9$ sono indipendenti. Perciò

$$EX = \prod_{j=1}^{9} EX_j.$$

[23] Si potrebbe arrivare alla stessa conclusione per via puramente algebrica, ma è più istruttivo l'argomento qui proposto.

Ora la v.a. X_j assume i possibili valori p_j^i $(i = 1, 2, ..., k_j)$, ciascuno con probabilità p_j^i; perciò

$$EX_j = \sum_{i=1}^{k_j} \left(p_j^i\right)^2 .$$

Questo numero si calcola sommando k_j termini, cioè (al massimo) poco più di un centinaio; poi si tratta di moltiplicare tra loro i 9 numeri EX_j, e il gioco è fatto. Troviamo cioè la formula molto più semplice:

$$p = \prod_{j=1}^{9} \left(\sum_{i=1}^{k_j} \left(p_j^i\right)^2 \right) . \tag{4}$$

Nella (4) le probabilità p_j^i sono quelle dei *genotipi*; per il calcolo effettivo conviene riscrivere la formula precedente in termini di probabilità degli *alleli*. Ragioniamo sul j-esimo locus; se q_j^i è la probabilità dell'i-esimo allele nel j-esimo locus, si vede facilmente, ricordando la diversa regola per il calcolo delle probabilità di genotipi omozigoti ed eterozigoti (v. tabella 2), che

$$\sum_{i=1}^{k_j} \left(p_j^i\right)^2 = \sum_{i=1}^{n_j} \left(q_j^i\right)^4 + \sum_{i<k} \left(2q_j^i q_j^k\right)^2 = 2 \left(\sum_{i=1}^{n_j} \left(q_j^i\right)^2 \right)^2 - \sum_{i=1}^{n_j} \left(q_j^i\right)^4$$

(dove n_j è il numero di alleli nel j-esimo locus). In definitiva, troviamo

$$p = \prod_{j=1}^{9} \left[2 \left(\sum_{i=1}^{n_j} \left(q_j^i\right)^2 \right)^2 - \sum_{i=1}^{n_j} \left(q_j^i\right)^4 \right] . \tag{5}$$

L'ultima formula trovata si può implementare direttamente a partire dalla tabella delle frequenze alleliche, che per 9 loci contiene poco più di un centinaio di numeri.

Calcoliamo ora la *RMP per la popolazione caucasica (su 9 loci)*, sulla base

della tabella di distribuzione allelica del database CODIS.

n° alleli	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5	Locus 6	Locus 7	Locus 8	Locus 9	
1	0,258	0,271	0,184	0,316	0,251	0,179	0,389	0,319	0,287	
2	0,241	0,227	0,174	0,225	0,229	0,167	0,336	0,275	0,227	
3	0,217	0,212	0,154	0,137	0,160	0,140	0,160	0,111	0,142	
4	0,148	0,109	0,143	0,111	0,089	0,121	0,068	0,087	0,140	
5	0,114	0,084	0,142	0,085	0,085	0,119	0,027	0,082	0,140	
6	0,010	0,082	0,080	0,080	0,078	0,119	0,015	0,070	0,038	
7	0,005	0,014	0,067	0,024	0,036	0,061	0,002	0,056	0,020	
8	0,005	0,002	0,031	0,012	0,032	0,031	0,002		0,007	
9	0,002		0,009	0,010	0,017	0,019				
10			0,007		0,012	0,012				
11			0,007		0,005	0,012				
12			0,002		0,002	0,007				
13			0,002		0,002	0,005				
14					0,002	0,005				
15						0,002				
Somma dei quadrati	0,20659536	0,195775	0,140498	0,196016	0,165062	0,127927	0,295403	0,212036	0,195155	
Doppio quadrato della somma dei quadrati	0,085363286	0,076655701	0,039479376	0,076844545	0,054490928	0,032730635	0,174525865	0,089918531	0,076170948	
Somma quarta potenza	0,010647895	0,010304977	0,003512108	0,013131717	0,007529336	0,002818954	0,036320875	0,016362595	0,010617044	
Differenza	0,074715391	0,066350724	0,035967268	0,063712828	0,046961592	0,029911681	0,138204989	0,073555935	0,065553904	1,06344E-11

Tabella 6

Utilizzando la formula (5) troviamo (v. tabella 6):

$$p_9 = 1.06 \cdot 10^{-11}. \tag{6}$$

Nel par. 3.1 abbiamo calcolato che per questo valore di p e una probabilità a priori di colpevolezza pari a 0.01, la probabilità di colpevolezza a posteriori nel caso di positività al test è dell'ordine di $1 - 10^{-9}$.

Se vogliamo conoscere l'RMP della popolazione caucasica su 13 loci anziché 9, possiamo eseguire un calcolo analogo a partire da una tabella di frequenze relativa a tutti e 13 i loci interessati (che qui non è riportata); si trova:

$$p_{13} = 2.15 \cdot 10^{-15}.$$

È significativo anche il numero

$$p_1 = \left(2.15 \cdot 10^{-15}\right)^{1/13} = 0.0744, \tag{7}$$

che dà, come media pesata, la probabilità di uguaglianza tra due genotipi a caso. In altre parole: se ogni locus avesse lo stesso numero di genotipi, e questi fossero tutti equiprobabili di probabilità 0.0744, la RMP della popolazione sarebbe uguale a quella che abbiamo calcolato a partire dalla distribuzione reale, non uniforme. Il ricorso a questa distribuzione fittizia dei genotipi, che è uniforme ma in un certo senso equivalente a quella reale, ci tornerà utile nel prossimo paragrafo.

A titolo di confronto, il valore di p_1 fornito dall'F.B.I. nello studio della popolazione caucasica[24] è

$$p_1 = 1/13.66 = 0.0732064,$$

[24] Questo dato è tratto da [4], che a sua volta cita il Journal of Forensic Science, Vol 44 number 6.

un valore abbastanza vicino a quello trovato in (7).

Lo spirito di questa sezione era quello di calcolare dei *valori medi* realistici per la RMP. Quanto questo sia importante si capisce meglio osservando quanto ampio è, per contro, il *range* dei valori che p può assumere. Per avere un'idea di quanto può variare la probabilità di un profilo, possiamo moltiplicare tra loro le frequenze di 9 genotipi scegliendo una volta per ogni locus il valore massimo, e una volta il valore minimo[25]. Usando i dati della tabella 6 della frequenza allelica, otteniamo:

$$\begin{aligned} p_{9,\max} &\simeq 5.71 \cdot 10^{-9} \\ p_{9,\min} &\simeq 6.29 \cdot 10^{-44}. \end{aligned} \tag{8}$$

Come si vede, la variabilità è enorme. Di fronte a questi dati non sarebbe stato facile azzardare quale possa essere un valore "tipico" per la probabilità di un profilo di 9 loci, senza un calcolo più accurato. Ad esempio, ricordiamo che la RMP su 13 loci calcolata dall'F.B.I. nel caso Jenkins era 10^{-18}, un valore che, tenuto conto dell'ampiezza del range, non ci appare ora troppo lontano dal valore tipico sopra calcolato, $p_{13} = 2.15 \cdot 10^{-15}$.

3.3 Il calcolo della DMP

Ricordiamo che la DMP è la probabilità che, in un database di m profili, ci sia almeno un match con un profilo prefissato.

Questo concetto è stato introdotto in relazione al cosidetto problema del "colpo a freddo" (l'incriminazione di una persona il cui profilo DNA si trova in un database, e viene trovato corrispondente a quello della traccia) perché, nell'interpretazione corrente, la probabilità di innocenza di una persona accusata mediante colpo a freddo sarebbe parente stretta della DMP.

In realtà a mio modo di vedere tra queste due quantità non c'è relazione. Ci occupiamo comunque del problema del calcolo della DMP per due motivi:

1) è propedeutico rispetto al problema ben più difficile di discutere il "paradosso del database dell'Arizona" (o meglio, di calcolare in quel contesto la probabilità a priori dell'*evento raro* realizzatosi);

2) ci dà un'indicazione quantitativa sull'utilità di avere a disposizione questi database del DNA.

Infatti, e qui sta secondo me il punto chiave, poiché la DMP è la probabilità di trovare almeno un riscontro nel database, questo numero è un indice dell'utilità di eseguire la ricerca nel database, mentre *non è* un indice del grado di colpevolezza o innocenza dell'eventuale persona individuata con questo metodo. Ma questo sarà approfondito nel prossimo paragrafo.

Il calcolo della DMP è presto fatto (come sostanzialmente già indicato nella prima parte).

[25] Più precisamente, il genotipo di probabilità minima ha per probabilità il quadrato della minima probabilità degli alleli (genotipo omozigote), mentra il genotipo di probabilità massima ha per probabilità il doppio prodotto delle due probabilità maggiori tra gli alleli (genotipo eterozigote).

Se p è la RMP con il profilo fissato, allora la variabile che conta il numero di profili nel database che coincidono con quello fissato è una binomiale $B(m,p)$, dove m è l'ampiezza del database. Il numero atteso di profili uguali a quello di riferimento è mp, e la probabilità di almeno un match è

$$1-(1-p)^m,$$

che a sua volta è ancora uguale circa a mp, almeno se p è molto piccolo rispetto ad $1/m$. Ad esempio, per un database di 65000 profili di 9 loci, la probabilità di almeno un match con un profilo che ha p uguale al valore medio (6) calcolato nella sezione precedente avremmo

$$65000 \cdot 1.06 \cdot 10^{-11} = 6.89 \cdot 10^{-7}.$$

Si badi, però, che *questo numero non è affatto uguale alla probabilità che almeno due profili nel database coicidano tra loro*, e *non rappresenta la probabilità che uno specifico profilo nel database corrisponda a quello di riferimento.*

3.4 Il calcolo della DCP e il paradosso del database dell'Arizona

Vogliamo ora occuparci della probabilità dell'evento "almeno due profili nel database sono *uguali tra loro*". Chiameremo questo numero Database Coincidence Probability, DCP. Il termine non è standard, perché questo numero è a volte confuso[26] con la DMP.

3.4.1 Calcolo della DCP per una distribuzione uniforme

Cominciamo dal caso, irrealistico ma utile per avvicinare gradualmente il problema nella sua generalità, di una popolazione in cui in ogni locus ci sia lo stesso numero di genotipi, tutti equiprobabili. Quest'eventualità è in realtà impossibile: se i genotipi omozigoti sono tutti equiprobabili, allora necessariamente gli alleli sono tutti equiprobabili, ma allora i genotipi eterozigoti hanno probabilità doppia rispetto ai genotipi omozigoti[27]. Tuttavia proseguiamo nella finzione, che ci servirà per proporre successivamente un argomento di media.

Sia n il numero di genotipi in ogni locus, 9 il numero dei loci (tanto per fissare le idee) e m il numero di individui del database.

Il numero di profili possibili è dunque n^9; se $m > n^9$ *certamente* almeno due profili nel database sono uguali; supponiamo quindi $m \leq n^9$. Se i genotipi sono tutti equiprobabili, il calcolo della probabilità di almeno un riscontro nel

[26] Nella discussione del caso del database dell'Arizona, la confusione tra queste due quantità si trova ad esempio in [10], [4].

[27] Se potessimo distinguere l'allele ereditato dal padre da quello ereditato dalla madre potremmo distinguere, sia pur fittiziamente, il genotipo (A,B) dal genotipo (B,A); in tal caso l'ipotesi di alleli equiprobabili implicherebbe quella di genotipi equiprobabili.

database è perfettamente analogo al ben noto "problema dei compleanni"[28], e si affronta con la combinatoria. Infatti sotto le nostre ipotesi anche i *profili* sono tutti equiprobabili. Calcoliamo la probabilità che i profili nel database siano tutti diversi; la DCP è il complemento a 1 di questa probabilità.

Il primo individuo del database può scegliere il suo profilo tra gli n^9 possibili; il secondo può sceglierlo tra gli $n^9 - 1$ diversi da quello del primo individuo, il secondo tra gli $n^9 - 2$ diversi da quelli dei primi due individui, e così via fino all'm-esimo individuo del database, che può sceglierlo in $n^9 - m - 1$. D'altro canto il numero totale di modi in cui si possono scegliere i profili per m individui è $\left(n^9\right)^m$, perciò si ha:

$$DCP = 1 - \frac{n^9\left(n^9-1\right)\left(n^9-2\right)...\left(n^9-m+1\right)}{\left(n^9\right)^m}. \tag{9}$$

La (9) si può riscrivere anche nella forma seguente (dove si è posto per comodità $M = n^9$):

$$DCP = 1 - \left(1-\frac{1}{M}\right)\left(1-\frac{2}{M}\right)...\left(1-\frac{m-1}{M}\right).$$

Questa scrittura è utile in vista della seguente formula di approssimazione:

$$\prod_{k=1}^{m-1}\left(1-\frac{k}{M}\right) = \exp\left(\log\prod_{k=1}^{m-1}\left(1-\frac{k}{M}\right)\right) = \exp\left(\sum_{k=1}^{m-1}\log\left(1-\frac{k}{M}\right)\right)$$
$$\simeq \exp\left(-\sum_{k=1}^{m-1}\frac{k}{M}\right) \simeq \exp\left(-\frac{m^2}{2M}\right),$$

da cui

$$DCP \simeq 1 - \exp\left(-\frac{m^2}{2M}\right), \tag{10}$$

che è molto più comodo da calcolare, e dà un'approssimazione piuttosto accurata.

Esempio 3 *Supponiamo di avere un database di* 65000 *profili (le dimensioni di quello dell'Arizona considerato nell'esempio). Supponiamo che la popolazione abbia una distribuzione con* 10 *genotipi possibili ed equiprobabili in ogni locus. Allora, per* 9 *loci si ha* $n^9 = 10^9 > 65000$, *e*

$$DCP = 1-\frac{10^9\left(10^9-1\right)...\left(10^9-65000+1\right)}{\left(10^9\right)^{65000}} \simeq 1-\exp\left(-\frac{65000^2}{2\cdot 10^9}\right) \simeq 0.879065.$$

[28] Si tratta di questo: se in una stanza ci sono k persone, qual è la probabilità che almeno due di esse compiano gli anni lo stesso giorno? Naturalmente la risposta dipende da k. Ciò che è sorprendente per l'intuizione è che il minimo numero di persone per cui questa probabilità è $> 1/2$ è 23, un numero che ci sembra piccolo rispetto a 365, che è il numero di possibili compleanni. Come si vede, la probabilità di *almeno una coincidenza* è più alta di ciò che il buon senso suggerirebbe.

(A titolo di confronto, il valore che si ottiene con la formula esatta anziché con l'approssimazione dell'esponenziale è 0.879066*). Stesso calcolo per* 21 *genotipi possibili:*

$$DCP \simeq 1 - \exp\left(-\frac{65000^2}{2 \cdot 21^9}\right) \simeq 0.002656,$$

un probabilità molto più piccola, ma non così piccola. Come si vede la DCP è più alta di quanto ci si aspetta intuitivamente (così come accade nel problema dei compleanni). Confrontiamo, per inciso, con la DMP:

$$DMP \simeq mp \text{ con } p = RMP = 1/n^9.$$

Si avrebbe:

$$\text{per } n = 10, DMP \simeq \frac{65000}{10^9} = 0.000065,$$
$$\text{per } n = 21, DMP \simeq \frac{65000}{21^9} = 8.18351 \cdot 10^{-8}$$

valori molto lontani da quelli della corrispondente DCP.

3.4.2 Calcolo della DCP per una distribuzione non uniforme

Facciamo ora l'ipotesi realistica che la distribuzione allelica nella popolazione non sia uniforme. Come potremmo calcolare la DCP? Il problema è notevolmente più complicato. La probabilità dell'evento "nel database i profili sono tutti diversi" non si lascia esprimere come rapporto "casi favorevoli/casi possibili", perché i profili non sono equiprobabili.

Per scrivere una formula esatta, bisogna ragionare nello spazio di probabilità di *tutti i profili possibili*. Sia N il numero totale di profili[29], di probabilità $q_1, q_2, ..., q_N$; sia m il numero di individui del database (ovviamente è $m \ll N$). Calcoliamo, al solito, la probabilità complementare di quella che ci interessa, ossia la probabilità che gli m individui abbiano tutti profili diversi. Prima scegliamo, tra tutti gli N profili, m profili (diversi tra loro) che apparterranno agli m individui del database: questi avranno ciascuno probabilità

$$q_{k_1}, q_{k_2}, ..., q_{k_m}$$

con

$$k_1 < k_2 < ... < k_m;\ k_j \in \{1, 2, ..., N\}, \tag{11}$$

e la probabilità di aver scelto congiuntamente questi m profili in un ordine prefissato è data dal prodotto di questi q_j; poi li permutiamo in tutti i modi possibili, e quindi questo prodotto va moltiplicato per $m!$; infine sommiamo al variare degli indici in tutti i modi ammissibili, cioè compatibili con (11). Si ottiene:

$$DCP = 1 - \sum_{\substack{k_1<k_2<...<k_m \\ k_j \in \{1,2,...,N\}}} \left(\prod_{j=1}^{m} q_{k_j}\right) m! \tag{12}$$

[29] Abbiamo calcolato in precedenza, per la popolazione caucasica, $N = 3.03 \cdot 10^{15}$ per profili di 9 loci.

Si può osservare che nel caso particolare della distribuzione uniforme si ritrova la formula (9): infatti in quel caso i profili sono tutti equiprobabili, quindi

$$q_{k_j} = \frac{1}{N} \text{ per ogni } j,$$

e si ha

$$DCP = 1 - \sum_{\substack{k_1<k_2<\ldots<k_m \\ k_j\in\{1,2,\ldots,N\}}} \left(\frac{1}{N^m}\right) m! = 1 - \left(\frac{1}{N^m}\right) m! \sum_{\substack{k_1<k_2<\ldots<k_m \\ k_j\in\{1,2,\ldots,N\}}} 1;$$

ora il numero di scelte degli indici k_j è il numero di scelte di m profili tra N, cioè $\left(\begin{array}{c} N \\ m \end{array} \right)$, e

$$\left(\frac{1}{N^m}\right) m! \left(\begin{array}{c} N \\ m \end{array} \right) = \frac{N\,(N-1)\,(N-2)\,\ldots\,(N-m+1)}{N^m},$$

che è la formula già trovata ($N = n^9$).

La (12) è, come la prima formula (3) trovata per la RMP di una popolazione, una formula "astronomica", praticamente inutilizzabile: per un database di 65000 individui, e un numero di profili possibili $N = 3.03 \cdot 10^{15}$ come calcolato in precedenza, la sommatoria ha $\left(\begin{array}{c} N \\ m \end{array} \right)$, cioè qualcosa come $5.47 \cdot 10^{721680}$ addendi!

Cerchiamo allora un approccio diverso, di tipo approssimato anziché esatto, che porti a calcoli praticabili. Sfruttiamo per questo l'idea di "distribuzione uniforme equivalente alla distribuzione reale", accennata alla fine del par. 3.2.2. Potremmo procedere così:

1. Si calcola, per la distribuzione allelica reale (quindi non uniforme) il numero $p_9 = RMP$, secondo la formula (5) relativa a 9 loci.

2. Si calcola la DCP di una popolazione "a genotipi equiprobabili" la cui RMP valga il p_9 calcolato al punto precedente. In pratica, si calcola DCP con la formula (9), dove si è posto $n^9 = 1/p_9$ (anche se questo numero non è un intero!), oppure con la formula approssimata (10), dove si è posto $M = 1/p_9$:

$$DCP' = 1 - \exp\left(-\frac{m^2 p_9}{2}\right).$$

Esempio 4 *Per la popolazione caucasica abbiamo calcolato in precedenza (v. (6)) $p_9 = 1.06 \cdot 10^{-11}$. Applichiamo quindi la (10) con $M = 1/p_9$. Otteniamo*[30]

$$DCP' \simeq 1 - \exp\left(-\frac{1.06 \cdot 10^{-11} \cdot 65000^2}{2}\right) = 0.0221436$$

[30] Se invece della formula mediante l'approssimazione gaussiana utilizzassimo la (9) troveremmo 0.0221433..., ossia un risultato coincidente fino alla quinta cifra significativa.

(Usiamo il simbolo DCP' anziché DCP per ricordarci che è solo un'approssimazione del valore vero di DCP). Un valore molto più grande, giustamente, del corrispondente valore di RMP. Confrontiamo con la DMP:

$$DMP = mp = 65000 \cdot 1.06 \cdot 10^{-11} = 6.89 \cdot 10^{-7},$$

che è un valore molto più basso.

L'idea del procedimento precedente è sostituire ad una distribuzione non uniforme un'altra uniforme che però abbia meno genotipi in ogni locus e quindi molti meno profili possibili. La minor variabilità dovuta al minor numero di genotipi compensa la maggior variabilità dovuta alla distribuzione uniforme, in modo da dare la stessa RMP (per definizione) e perciò, speriamo, una DCP non troppo diversa. Nell'esempio qui sopra, la distribuzione non uniforme originale ha un numero di profili possibili $N = 3.03 \cdot 10^{15}$; la distribuzione uniforme "equivalente" ne ha "soltanto" $M = 1/p_9 = 9.43 \cdot 10^{10}$. In termini di alleli, invece di averne un numero variabile da 7 a 15, la distribuzione uniforme approssimante ne ha un numero (costante) pari ("circa") a 5, che dà $n = 15$ genotipi e $p_1 \simeq 1/15 = 0.066$, coerentemente al fatto che il corrispondente valore p_9 è $0.066^9 = 2.6 \cdot 10^{-11}$, valore confrontabile al valore vero di $p_9 = 1.06 \cdot 10^{-11}$.

Naturalmente l'affermazione che DCP' sia un'approssimazione accettabile della DCP va giustificata. In effetti si può dimostrare che nel caso dell'esempio 4 vale la seguente stima a priori dell'errore commesso:

$$|DCP - DCP'| \leq 0.0125, \tag{13}$$

il che porta al seguente intervallo per il valore vero della DCP:

$$0.0096 < DCP < 0.0346.$$

La dimostrazione di questo risultato è piuttosto laboriosa e, dato il taglio espositivo di questo articolo, ritengo opportuno ometterla. Il lettore interessato può trovare tutti i dettagli in [3]. Aggiungo soltanto che la stima numerica è ottenuta come applicazione di una stima teorica che coinvolge i parametri p, m, il numero $q_{\max} = \max_i q_i$ e le quantità

$$E\left(X^n\right) = \sum_{i=1}^{N} \left(q_i\right)^{n+1},$$

per $n = 2, 3, ...$ interpretabili come momenti della variabile X che assume il valore q_i con probabilità q_i. Questi ultimi possono essere calcolati a partire dalla tabella delle frequenze della distribuzione degli alleli, con un procedimento analogo a quello con cui abbiamo calcolato la RMP della popolazione (che non è altro che $E\left(X\right)$). Per ottenere la stima (13) sono stati utilizzati i momenti fino a $n = 20$.

3.4.3 Il paradosso del database dell'Arizona

Il precedente calcolo della DCP, che dà un valore compreso tra circa l'1% e il 3.5%, suggerisce che non sia un evento così raro trovare *due* profili uguali nel database, per quanto sia ben più raro trovarne *molti* uguali. Nel caso reale del database dell'Arizona, tuttavia, l' "evento raro" in discussione è che ci sia coincidenza in 9 loci *qualsiasi scelti* tra i 13, il che dovrebbe aumentare di molto la probabilità. Tuttavia, finora non sono stato capace di calcolare questa probabilità. Senz'altro gli approcci "bernoulliani" proposti in [10], [4] sono grossolanamente sbagliati, per i motivi già discussi. Il problema potrebbe anche essere aperto e se qualche lettore trova la soluzione, sarò ben lieto se me la comunicherà. Volendo precisare il problema, per verificare se la coincidenza osservata nel caso dell'Arizona sia un evento a priori molto raro oppure no, occorrerebbe calcolare la probabilità del seguente evento, o almeno stimarne l'ordine di grandezza:

"Almeno 144 profili, tra i 65000 del database, sono uguali in 9 loci (qualsiasi) su 13, ma non sono uguali in 10 loci o più".

Il tutto tenendo conto del fatto che i profili non sono affatto equiprobabili, ma le loro probabilità si possono calcolare in base a tabelle di frequenze alleliche sui 13 loci.

3.5 Ricerca in un database e "cold hit"

3.5.1 Le probabilità condizionate inverse nel caso della ricerca in un database

Consideriamo ora il caso in cui a priori non abbiamo alcun sospettato, ma confrontando sistematicamente il DNA del campione trovato sulla scena del crimine con un database di m campioni, troviamo uno e un sol individuo per cui si ha corrispondenza in 9 loci su 9 (o un altro numero prefissato). Per semplicità non considereremo l'eventualità che nel database si trovi più di un riscontro. Chiameremo J l'unico individuo del database che è risultato positivo al raffronto.

Sia p l'RMP di riferimento (se parliamo di un caso concreto, sarà la probabilità del profilo della traccia trovata sulla scena del crimine e non il generico RMP della popolazione) e m l'ampiezza del database.

Sia DMS l'evento "database match specifico" ossia "J è risultato positivo al raffronto e nessun altro nel database lo è";

sia C l'evento "J è colpevole", sempre da intendersi nel senso limitativo di "il DNA trovato sulla scena del crimine appartiene a J";

sia I l'evento "J è innocente e il colpevole è interno al database";

sia E l'evento "il colpevole è esterno al database (e quindi in particolare J è innocente)".

Ci interessa calcolare la probabilità a posteriori di colpevolezza,

$$\pi = P(C|DMS) = \frac{P(DMS|C)\,P(C)}{P(DMS|C)\,P(C) + P(DMS|E)\,P(E) + P(DMS|I)\,P(I)}.$$

Si ha:

$$P\left(DMS|C\right) = \left(1-p\right)^{m-1}$$

(se J è colpevole è certo che J corrisponderà, ma vogliamo escludere che ci siano falsi positivi tra tutti gli altri $m-1$);

$$P\left(DMS|E\right) = p\left(1-p\right)^{m-1}$$

(come prima per i falsi positivi, e inoltre se J è innocente la probabilità che risulti positivo è p; si badi che non ci stiamo chiedendo qual è la probabilità che almeno un individuo nel database sia positivo, ma la probabilità che lo sia proprio J);

$$P\left(DMS|I\right) = 0$$

(se il colpevole è interno al database e non è J, è impossibile che *nessuno oltre a* J sia risultato positivo al test).

Dobbiamo ancora, però, introdurre la probabilità a priori di colpevolezza di J, ossia il numero

$$P\left(C\right) = c_0.$$

Questa è il vero problema. Per il presupposto innocentista[31] dovremmo porre $c_0 = 0$, e così il discorso si arena. D'altro canto, nella logica investigativa, è chiaro che $c_0 > 0$. Dopo tutto il delitto è avvenuto, qualcuno l'ha commesso, il numero delle persone sul pianeta è finito, quindi non può essere $c_0 = 0$ per ciascun individuo. La domanda da porsi è:

"Qual è la probabilità (prima di eseguire il test) che quel particolare individuo del database sia colpevole?"

Il problema sarà assegnare a c_0 un ordine di grandezza sensato. Ad esempio, potremmo porre $c_0 = 1/N$ dove N è l'ampiezza della popolazione a cui "deve" appartenere il colpevole, ammesso che si sappia valutare questo N. Ad ogni modo, si ha:

$$\begin{aligned}\pi\left(c_0, p, m\right) &= P\left(C|DMS\right) = \\ &= \frac{\left(1-p\right)^{m-1} c_0}{\left(1-p\right)^{m-1} c_0 + p\left(1-p\right)^{m-1} P\left(E\right)} = \frac{c_0}{c_0 + pP\left(E\right)}.\end{aligned}$$

Poiché $P\left(E\right) \leq 1 - c_0$ si ha

$$\pi\left(c_0, p, m\right) \geq \frac{c_0}{c_0 + p\left(1-c_0\right)},$$

formula che dà una limitazione inferiore della probabilità di colpevolezza, *indipendente dall'ampiezza del campione.*

Notiamo che invece $P\left(E\right)$ dipende implicitamente da m. Se, come caso limite, fossimo certi a priori che il colpevole appartiene al database (che è come dire che abbiamo esteso il database all'universo dei sospettati), otterremmo $\pi = 1$, il che è ovvio (in questo caso, l'unico positivo al test è il colpevole).

[31] Ricordiamo che in questo caso, prima di eseguire il test, su J non pesa alcun indizio!

Solitamente si presenta però la situazione opposta: il database è *piccolo* in confronto alla popolazione dei potenziali colpevoli, quindi $P(E)$ è poco inferiore a 1. Tenendo anche conto del fatto che ragionevolmente è $p \ll c_0$ (dato che i valori tipici di $1/p$ hanno ordini di grandezza maggiori del numero di persone sul pianeta) deduciamo allora

$$P(C|DMS) \simeq \frac{c_0}{c_0 + p} \tag{14}$$

La cosa sorprendente è che questa probabilità (diversamente dalla DMP) *non dipende esplicitamente dall'ampiezza m del database*[32]. Anzi, questa formula coincide con la (1), ottenuta nel caso del test eseguito su un unico sospettato, con l'importante differenza di come valutiamo la probabilità a priori di colpevolezza, che nel caso del test su un unico sospettato sarà un numero significativamente discosto da zero, mentre nel caso della ricerca in un database potrebbe avere un valore piccolissimo.

Data l'incertezza sulla valutazione quantitativa di c_0, questo algoritmo sembra essere scarsamente utilizzabile[33]. Tuttavia, si consideri il seguente

Esempio 5 *Supponiamo che si stia cercando nel database del CODIS (13 loci) il colpevole di un crimine commesso in USA. Se valutiamo $c_0 = 10^{-8}$ (il che corrisponde all'idea grossolana che un pregiudicato che si trova nel database sia sospettato non meno della media dei cittadini americani, che sono dell'ordine di 10^8) e utilizziamo il valore medio calcolato in precedenza (v. par. 3.2.2) $p_{13} = 2.15 \cdot 10^{-15}$, otteniamo*

$$\pi \simeq \frac{c_0}{c_0 + p} = \frac{10^{-8}}{10^{-8} + 2.15 \cdot 10^{-15}} = \frac{1}{1 + 2.15 \cdot 10^{-7}} \simeq 1 - 2.15 \cdot 10^{-7},$$

che come indice di colpevolezza non è male.

3.5.2 Il test in due passi su loci indipendenti

Si può anche pensare di utilizzare il "colpo a freddo" ottenuto nella ricerca in un database come punto di partenza per un confronto più approfondito (nello spirito dei raffronti fatti quando si ha un unico forte indiziato):

1. Inizialmente non ho alcun sospettato; lo scan del database offre una e una sola corrispondenza, poniamo in 8 loci su 8, per l'individuo J.

2. L'individuo J diventa allora il nostro sospettato, che sottoponiamo a un nuovo raffronto del DNA, su (poniamo) ulteriori 5 loci.

Supponiamo che risulti positivo anche a questo raffronto. L'intero procedimento si può allora rappresentare così.

[32] Si confronti quest'affermazione con la raccomandazione fatta nel documento ufficiale [7] e ricordata nel par. 2.3: "nel caso di un riscontro nel database, bisogna comunicare alla giuria il valore della DMP, ossia mp".

[33] Difatti vari autori utilizzano approcci diversi, ad esempio ricorrendo ai già citati rapporti di verosimiglianza. Si vedano ad es. [1], [25] per discussioni del problema della ricerca in un database, alternative a quella qui presentata.

Applichiamo la (1), e calcoliamo la probabilità di colpevolezza a posteriori del (secondo) raffronto sui 5 loci,

$$\pi = \frac{1}{1 + p_5 \left(\frac{1}{c} - 1\right)},$$

dove c è la probabilità di colpevolezza in base al (primo) raffronto su 8 loci:

$$c \simeq \frac{c_0}{c_0 + p_8}$$

e c_0 è la probabilità di colpevolezza a priori. Sostituendo la seconda nella prima si trova:

$$\pi = \frac{1}{1 + \frac{p_5 p_8}{c_0}}.$$

La cosa interessante è che i numeri p_8 e p_5 compaiono nella formula solo mediante il loro prodotto $p_8 p_5 = p_{13}$ che ha il significato di RMP per il raffronto su tutti e 13 i loci. Ossia:

le nostre conclusioni non dipendono dal modo in cui abbiamo "spezzato in due" le informazioni che avevamo: un primo test su 6 *loci seguito da un secondo su* 7 *avrebbe dato gli stessi risultati.*

(Quello che cambia è la probabilità di identificare uno e un solo sospettato nello scan del database: se usiamo poca informazione nel primo passo, rischiamo di trovare più sospettati). In definitiva, possiamo riscrivere la nostra conclusione come:

$$\pi \simeq \frac{1}{1 + \frac{p_{13}}{c_0}}.$$

Ad esempio, nel caso Jenkins, se $p_{13} = 10^{-18}$ come sosteneva l'F.B.I., e poniamo $c_0 = 10^{-8}$ (si veda l'argomentazione nell'esempio della sezione precedente), otteniamo

$$\pi = \frac{1}{1 + 10^{-10}} \simeq 1 - 10^{-10}.$$

Dalla discussione di quest'ultima sezione raccogliamo quindi la seguente conclusione:

fissata l'informazione in nostro possesso (numero di loci a disposizione per il confronto) il procedimento in due tempi non è di per sé meglio del procedimento in un passo solo ("cold hit").

Solo se l'accertamento successivo è l'occasione per esaminare loci *ulteriori* (che non si trovavano nel campione presente nel database, ma si possono ricavare dal sospettato, e che supponiamo presenti nel campione proveniente dalla scena del crimine), questa procedura aggiunge qualcosa alle nostre informazioni. Si noti che proprio questa seconda evenienza è quella che si è realizzata nel caso Jenkins, in cui il database della Virginia su cui è stato fatto lo scan conteneva 8 loci, mentre il campione lasciato sulla scena del crimine consentiva di analizzarne 13.

Conclusioni

Sintetizziamo le principali affermazioni a cui siamo giunti nel corso della parte 3.

1. L'ipotesi di *indipendenza dei genotipi* presenti in loci diversi, almeno all'interno di una popolazione etnicamente omogenea, è un'ipotesi cruciale su cui si basa tutto il calcolo delle probabilità dei profili. Quest'ipotesi, che nella prassi è generalmente accettata, è ragionevole dal punto di vista dei meccanismi biologici ma difficilmente può essere verificata a posteriori mediante test statistici di indipendenza, a causa della mole di dati statistici che sarebbero richiesti (v. par. 3.2.1). Su questo punto probabilmente sarebbe necessaria una riflessione ulteriore (ma si ricordi anche quanto osservato nella nota 18, par. 3.2.1). Nel seguito di queste conclusioni partiremo dal presupposto (standard) che tale ipotesi sia verificata.

2. Calcolare la *probabilità di un profilo specifico* (v. par. 3.2.1) è allora banale. Meno banale, e certamente interessante per le applicazioni ai calcoli teorici che abbiamo fatto, è la determinazione della probabilità che un profilo scelto a caso da una popolazione coincida con un profilo fissato ma ignoto. Si tratta della *RMP di una popolazione*, che abbiamo mostrato come calcolare nel par. 3.2.2.

3. Qual è la probabilità che un individuo sia l'effettivo proprietario della traccia di DNA trovata sul luogo del crimine, se è risultato positivo al test del DNA? Sebbene il calcolo fatto mediante il teorema di Bayes lasci un certo alone di indeterminatezza dovuta alla difficoltà di quantificare la *probabilità a priori di colpevolezza*, abbiamo visto con esempi numerici che questa probabilità è comunque *molto alta*, sia nel caso del test su un unico indiziato (v. par. 3.1) sia nel caso della ricerca in un database (v. par. 3.5.1): il test del DNA è uno strumento di identificazione potente, anche se non è possibile distillare univocamente un numero che si possa chiamare "probabilità di colpevolezza".

4. Il caso di corrispondenza trovata mediante *ricerca in un database* è quello che offre spesso motivi di perplessità. Oltre a quanto appena ricordato (v. punto 3), abbiamo mostrato a questo riguardo che: la DMP (probabilità di trovare almeno un match con un certo profilo esterno in un database, v. par. 3.3), facilmente calcolabile, non va confusa con la probabilità di innocenza a posteriori (che tra l'altro *non dipende sensibilmente dall'ampiezza del database*, v. par. 3.5.1, almeno quando questo sia piccolo rispetto alla popolazione dei potenziali colpevoli), né con la probabilità di trovare almeno due profili uguali in un database (DCP, v. par. 3.4), che può essere molto più alta della DMP. Inoltre: fissato il numero totale di loci a disposizione, la ricerca in un database "in due tempi" (v. par. 3.5.2), cioè usando una parte dei loci per individuare un sospettato e un'altra parte dei loci per eseguire un raffronto ulteriore, non dà risultati migliori della ricerca in un solo tempo (mentre è ovviamente meglio se è l'occasione per esaminare loci *ulteriori*).

5. Abbiamo dedicato un certo spazio (v. par. 3.4) a mostrare come calcolare la DCP in modo approssimato. Questo calcolo suggerisce che non sia così improbabile trovare due profili parzialmente coincidenti in un grande database,

e che quest'eventualità non dovrebbe portare quindi, di per sé, a dubitare nel potere identificativo del test del DNA.

Riferimenti bibliografici

[1] D. J. Balding: *The DNA database search controversy*, Biometrics 58 (2002), no. 1, 241–244.

[2] D. A. Berry: *Inferences Using DNA Profiling in Forensic Identification and Paternity Cases*, Statistical Science, Vol. 6, No. 2 (1991), 175-189.

[3] M. Bramanti: *Un risultato di approssimazione per spazi di probabilità finiti non uniformi,* Novembre 2009. Documento scaricabile all'indirizzo: http://www1.mate.polimi.it/~bramanti/pubblica/prob_approx.pdf

[4] C. Brenner: *Arizona DNA Database Matches*, January 8, 2007: http://dna-view.com/ArizonaMatch.htm

[5] J.M. Butler: *Genetic and genomics of core short tandem repeat loci used in human identity testing*, J. Forensic Sci. 51 (2006), pp. 253–265.

[6] Committee on DNA Technology in Forensic Science, National Research Council: *DNA Technology in Forensic Science*, National Academy Press, 1992.

[7] Committee on DNA Technology in Forensic Science, National Research Council: *An Update: The Evaluation of Forensic DNA Evidence*, National Academy Press, 1996. Riassunto scaricabile da: http://www.nap.edu/catalog/5141.html

[8] L. A. Derksen: *Agency and Structure in the History of DNA Profiling: The Stabilization and Standardization of a New Technology.* PhD Thesis, Department of Sociology and Science Studies Program, University of California, San Diego. (2003). Scaricabile da: http://web.viu.ca/derksenl/Publications/DNA%20Profiling%20History.htm

[9] K. Devlin, G. Lorden: *Il matematico e il detective.* Longanesi, Milano, 2008.

[10] K. Devlin: *Scientific Heat about Cold Hits,* Unfinished draft, 2007. Scaricabile da: http://stanford.academia.edu/KeithDevlin/Papers

[11] A. Jeffreys, V. Wilson, S. Thein: *Hypervariable "minisatellite" regions in human DNA*, Nature 314 (1985), 67-73.

[12] A. Jeffreys, V. Wilson, S. Thein: *Individuals specific fingerprints of human DNA*, Nature 316 (1985), 76-79.

[13] D. H. Kaye: *Rounding Up the Usual Suspects: A Logical and Legal Analysis of DNA Trawling Cases,* North Carolina Law Review, Vol. 87, No. 2, (2009), 425-503. Scaricabile al sito:
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1134205

[14] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich: *Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction*, Cold Spring Harbour Symposium on Quantitative Biology 51 (1986), 263-273.

[15] Random Match Probability Calculator:
http://www.csfs.ca/pplus/profiler.htm

[16] U. Ricci: *DNA e crimine: dalla traccia biologica all'identificazione genetica*, Laurus Robuffo, Roma II ed. 2004.

[17] U. Ricci: *Genetica forense e identificazione personale.* Il Giornale del Linguaggio Universale n° 2 (2006). Articolo scaricabile da:
http://saccone.dba.unict.it/didattica/geneticamutagenesi/test_1.pdf

[18] U. Ricci, C. Previdere, P. Fattorini, F. Corradi: *La prova del DNA per la ricerca della verità. Aspetti giuridici, biologici e probabilistici*, Giuffrè, 2006. Indice ed introduzione sono visibili alla pagina web:
http://www.aifo-italia.it/all/la_prova_del_dna.pdf

[19] K. Roeder: *DNA fingerprinting: a review of the controversy.* Statist. Sci. 9 (1994), no. 2, 222–278.

[20] Sito della Canadian Society of Forensic Science (contiene dati statistici sulla distribuzione allelica nella popolazione nordamericana)
http://www.csfs.ca/
In particolare, le tavole di frequenza allelica si possono scaricare come file Excel al link:
http://www.csfs.ca/strdnadata/CFSalldata.zip

[21] Sito del CODIS (database DNA dell'F.B.I.):
http://www.fbi.gov/hq/lab/html/codis1.htm

[22] Sito del database DNA in U.K.:
http://www.homeoffice.gov.uk/science-research/using-science/dna-database/

[23] Sito del Gruppo Genetisti Forensi Italiani (con statistiche sulle frequenze alleliche nella popolazione italiana):
http://www.gefi-forensicdna.it/

[24] Sito dell'agenzia governativa australiana Crimtrac:
http://www.crimtrac.gov.au
/systems_projects/KeyDatesintheHistoryofDNAProfiling.html

[25] A. Stockmarr: *Likelihood Ratios for Evaluating DNA Evidence When the Suspect is Found Through a Database Search*, Biometrics 55 (1999), 671-677.

[26] N. Van Camp, K. Dierickx: *National Forensic DNA Databases in the EU*, European Ethical-Legal Papers N°9, Leuven, 2007. Scaricabile dalla pagina web del Centre for Biomedical Ethics and Law of the Catholic University of Leuven: https://www.kuleuven.be/cbmer /page.php?LAN=E&ID=383&TID=0&FILE=subject&PAGE=1