L'identificazione di persone scomparse e di vittime di disastri di massa tramite profili di DNA

La Razionalità nel giudicare
CISA, Università di Firenze

Fabio Corradi
Dipartimento di Statistica, Informatica, Applicazioni, Università di Firenze

5 Novembre 2021

Finalità della presentazione

- Illustrare due problemi reali:
  - L'identificazione di persone scomparse
  - L'identificazione di vittime di disastri di massa
- Rappresentare *in modo grafico* un problema con molte evidenze, molte ipotesi, molti dati mancanti
- Gestire i risultati tramite la costituzione di una *short list* ottenuta tramite la teoria delle decisioni

Identificazione di persone scomparse: richieste

- Organizzazione preventiva di lungo periodo per la costituzione di una banca dati con la tipizzazione di evidenze genetiche di persone non identificate (unknown)
- Acquisizione, per ogni persona dichiarata scomparsa (missing), del pedigree familiare che lega quest'ultima con i parenti che la stanno cercando, che sono donatori di DNA che viene acquisito
- Un sistema probabilistico per processare le evidenze e così fornire informazione circa le ipotesi di interesse

Identificazione di vittime di disastri di massa: richieste

- Organizzazione tempestiva di rilevazione di campioni di DNA relative alle vittime non identificate unknowns, tipizzazione delle stesse e immissione in un data base.
- Organizzazione di una raccolta di tracce di DNA relative ai parenti delle vittime del disastro, tipizzazione delle stesse e immissione in un data base.
- Va inoltre acquisito per ogni persona dichiarata scomparsa il pedigree familiare che lega quest'ultima ai parenti donatori di DNA

Ulteriori richieste

- I data base degli unknown e dei parenti di persone scomparse debbono essere accessibili al software di identificazione
- Il sistema probabilistico di trattamento delle evidenze per fornire informazione circa ipotesi di interesse deve essere computabile
- Remark: Fornire informazione significa attribuire una probabilità a tutte le ipotesi che sono giudicate di interesse condizionatamente alle evidenze disponibili, seguendo le regole del calcolo delle probabilità

Output dei due sistemi

- Valutazione di ipotesi in casi di estrema difficoltà di identificazione (corpi degradati, assenza di altre evidenze identificative)
- Missing persons: Contemporanea valutazione della probabilità di un *unknown* di essere uno dei *missing* con gestione della probabilità a priori di identificazione basate sulla distanza fra luogo di ritrovamento dell'unknown e i suoi luoghi di residenza.
- Vittime di Disastri: valutazione contemporanea di attribuzione di ogni unknown a una delle famiglie richiedenti. Possibilità di identificazione anche per più vittime in uno stesso nucleo familiare senza richiesta dei parenti

Definizioni

Una famiglia provvede evidenze di DNA, x_f, per rintracciare il suo familiare, M, scomparso. Si valuta se possa essere un *Unknown*, U che viene tipizzato, x_U

- Definiamo $H = \{H_f, H_g\}$
  - H_f: U è la persona scomparsa in quella famiglia ($U \equiv M$)
  - H_g: U è un membro generico di una popolazione di riferimento ($U \neq M$)
- NB: M quasi sempre non è osservato!

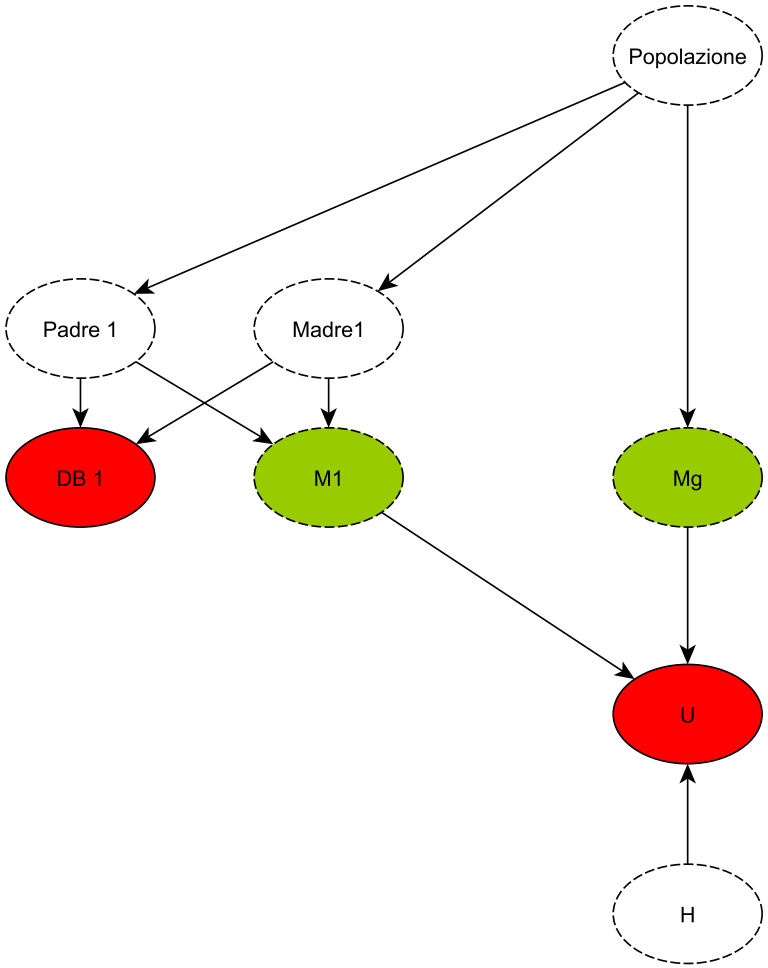
Una famiglia - un missing

- Risultati *analitici* tipicamente offerti in termini di

$$LR = \frac{p(x_U|x_f, H_f)}{p(x_U|H_g)},$$

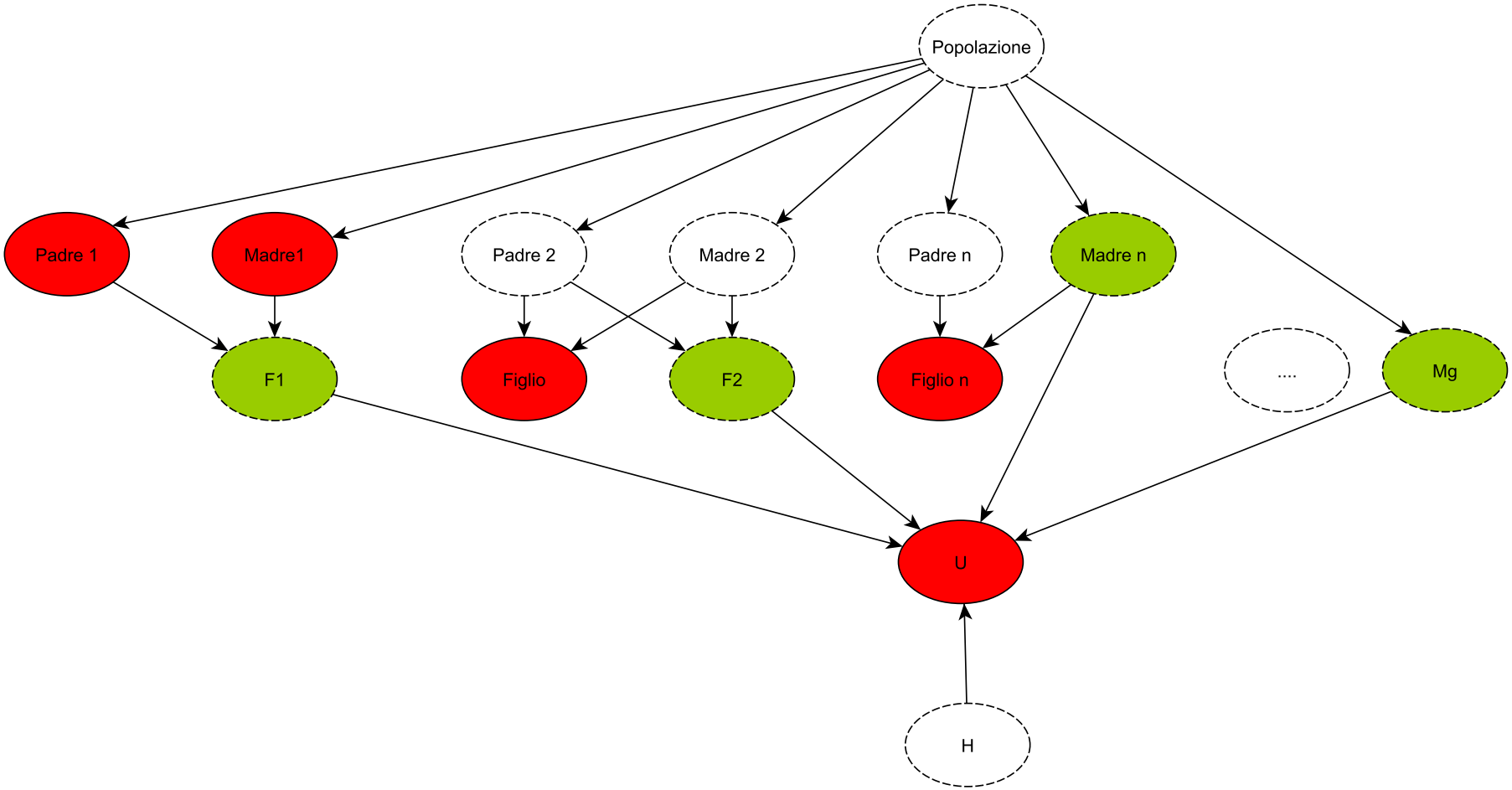
- LR = likelihood ratio = quanto i dati piacciono alle ipotesi
- $LR > 1$ supporta l'ipotesi che $U \equiv M$
- Applicazioni per le valutazioni di parentela [Dawid et al. (2002)], [Egeland et al. (2006)]
- Software dedicati (Familias, DNAview, etc)

Un *unknown* - un missing tramite *Bayesian Networks* (BN)



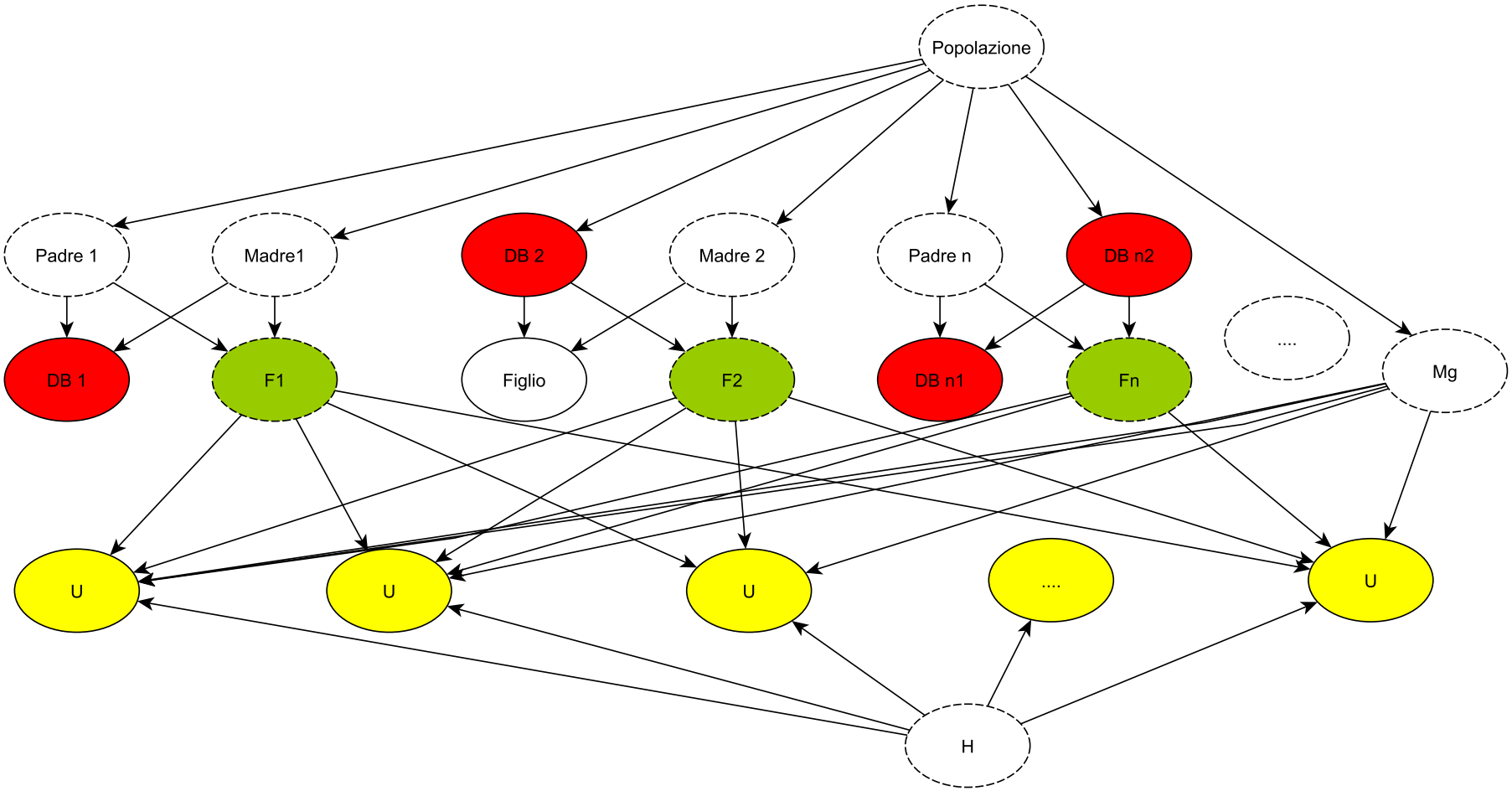
- Nell'esempio un fratello DB1 cerca di stabilre se il fratello M1 è l'Unknown oppure no
- Risultati in termini di probabilità di H_f e H_g (essere / non essere la persona Missing nella famiglia)
- Possibilità di calcolare il tradizionale *LR*

Un *unknown* - molti missing, tramite *BN*



- Ora $H = \{H_{f_1}, H_{f_2}, \ldots, H_{f_n}, H_g\}$
- Valutazione che U sia fra uno dei missing,

Molti *unknowns* - molti missing, tramite *BN*



Molti *unknowns* - molti missing, tramite *BN*

- Assumo $n = 100$ missing e $5 \leq k \leq 95$ unknowns
- Il numero di *possibili attribuzioni* degli unknown ai missing è $|H| = \frac{n!}{(n-k)!}$

k	nn
5	75287520
15	253338471349989280
25	2425192697203419565282 80
35	109506715318799097640868 2802
45	614484712141313085248200 04020
55	614484712141313085248200 04020
65	109506715318799097640868 2802
75	242519269720338500426884
85	253338471349996480
95	75287520

Identificazione di missing analiticamente

- E' stata proposta una via alternativa alle reti e basata sul calcolo dei LR a livello familiare per risolvere il problema delle DB search [Slooten and Meester (2014)]
- Considera una kinship analysis applicata ad una famiglia alla volta (un *unknown* - un missing, ottenendo n+1 risultati in termini di likelihood ratios

$$\mathrm{r} = \{r_0, r_1, \ldots, r_n\}$$

- Si combinano *opportunamente* i risultati ottenuti, per ottenere....

Elenco risultati ottenibili analiticamente o tramite Bayesian Network

1. Probabilità che l'*Unknown* sia un ben specifico missing avendo come alternativa che sia uno degli altri *missing* oppure appartenga al resto della popolazione di missing di cui non si hanno informazioni parentali (cd *Rest*)
2. Probabilità di presenza del missing fra quelli cercati dalle famiglie richiedenti
3. LR a supporto che $U \in \mathcal{M}$ vs $U \in rest$
4. Likelihood ratio a supporto dell'ipotesi che $U \equiv M_j$ versus $U \in Rest \cup \mathcal{M} \setminus M_j$

Approcci a confronto

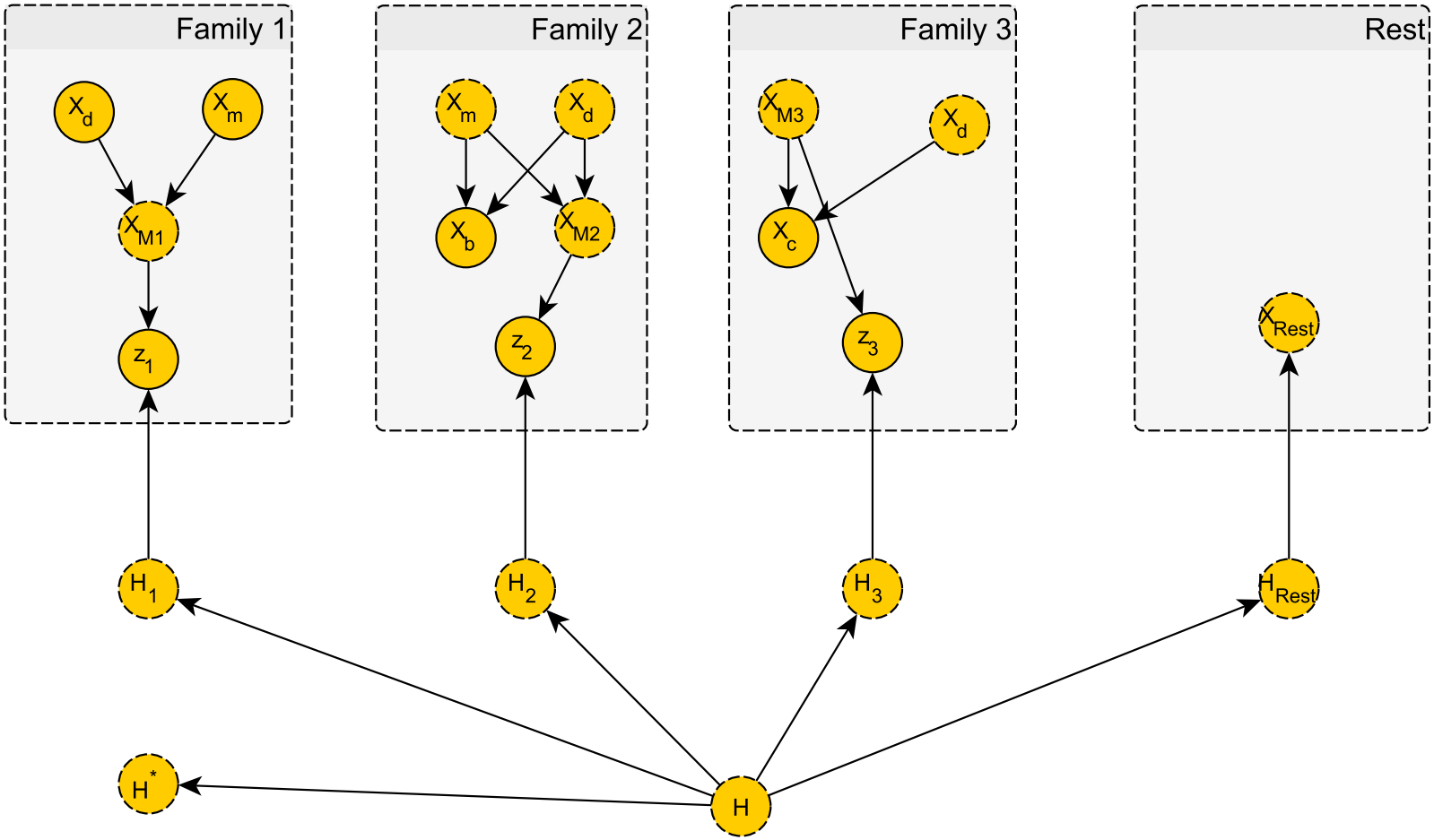
- Risultati coincidenti con i due approcci. Quale dei due è il migliore?
- Identificazione analitica: facilità di combinare dei likekihood ratio r $= \{r_0, r_1, \ldots, r_n$ ottenuti tramite un software specialistico in casi più standard di identificazione (una sola popolazione di riferimento, nessuna incertezza nelle osservazioni etc.)
- Maggiore facilità per le reti di gestire situazioni complesse (popolazioni mixed e admixed, alleli silenti, incertezza sulle stime di probabilità delle frequenze alleliche, mutazioni etc.)
- Sorprendentemente, in questo topic, la soluzione BN ha preceduto nel tempo quella analitica

Un *Toy example*

L'esempio è presentato per evidenziare la interrelazione che si crea fra le richieste di identificazione da parte delle famiglie e la persona da identificare

- $N = 100$ persone dichiarate missing. $n = 3$ famiglie hanno fornito materiale biologico e pedigree
- Consideriamo un solo Locus diallelico (A,B), sia $p(A) = 0.1$, $p(B) = 0.9$
- Osserviamo $U = (A, A)$ (very lucky)
- Abbiamo solo tre famiglie
  - F1: Madre e Padre (AA,AA) cercano il figlio
  - F2: Fratello (AB) cerca fratello
  - F3: Madre (AB) cerca figlio

Un *Toy example, rappresentazione*



Un *Toy example: risultati*

- In ottica non informativa iniziale la probabilità che ciascuno dei tre missing sia l'Unknown è $\frac{1}{100} = 0.01$.
- Dopo che tutta l'evidenza è stata considerata abbiamo

$\mathcal{F}$	Missing	x_f	$p(H = i \mid x_U, x_f)$	$LR(U = M_j)$
f_1	figlio	$\{x_d, x_m\} = AA$	0.499	96.022
f_2	fratello	$x_b = AB$	0.017	1.663
f_3	madre	$x_c = AB$	0.014	1.390

- Risultati contraddittori: LR favorisce l'identificazione di M1 ma la sua probabilità a posteriori non è così elevata in assoluto ma cmq si è passati da 0.01 a 0.499
- Risultato accettabile solo scartando la possibilità di acquisire più evidenze

Formare una short list

- Nel caso i confronti fra l'unknown e i missing siano moltissimi (decine di migliaia), si ottengono altrettante probabilità a posteriori e LR, uno per famiglia
- Il passo successivo è quello di decidere quali verificare presso le famiglie
- Le verifiche spesso sono difficoltose (e penose) e si cerca di restringerle alle situazioni più promettenti

I° proposta (puramente probabilistica)

[Slooten and Meester (2014)] individuano una short list di k missing individuals con il valore più alto del prodotto $r_j p(H = j)$, e tale che

$$\sum_{j=1}^{k} r_j p(H = j) \geq \alpha \sum_{i=1}^{n} r_i p(H = i).$$

- Nel caso di uguali probabilità a priori di identificazione equivale a individuare i k individui che, assieme, hanno α volte la somma totale dei LR
- *Se l'Unknown è compreso fra i missing cercati dalle famiglie* il metodo garantisce l'identificazione con probabilità α

I° proposta (probabilistica)

- Riprendendo l'esempio considero di verificare i due missing che hanno ottenuto il LR più alto
- *Se l'Unknown è compreso fra i missing cercati dalle famiglie* il metodo stabilisce la relazione

$$97.685 = \alpha \cdot 0.99075$$

ovvero assicura di identificare l'unknown nel

$$\alpha = \frac{97.685}{99.075} = 98.59\%$$

dei casi

- Poichè non so se l'unknown è presente fra i missing cercati, i missing $\{1, 2\}$ assicurano (realmente) una probabilità di successo pari a 0.509

II° proposta (probabilità + teoria delle decisioni)

[Boreale and Corradi (2016)], e [Corradi (2016)] propongono una short list fondata sulla teoria delle decisioni

Assunzioni

- Assumo di poter stabilire il costo (medio) di ogni singola verifica e lo pongo = 1.
- Assumo di poter stabilire che il risultato positivo di identificazione dell'unknown, vale *a* volte il costo della verifica
- Ambedue le quantità variano da caso a caso

II° proposta (probabilità + teoria delle decisioni)

Regola decisionale ottima

- Includo un missing nella lista e procedo alle verifiche se $a \cdot p(H = i|x_u, x_f) > 1$

- Ovvero includo un missing nella verifica se il valore atteso di identificare l'unknown con quel missing è superiore al costo unitario dello sforzo richiesto per la verifica .

II° proposta (probabilità + teoria delle decisioni)

Risultati con il Toy example

a	1	2	3	...	12	...	15
$\mathcal{D}^*$	$\emptyset$	$\emptyset$	1	...	1, 2	...	1, 2, 3

- Il *Missing 1* viene scrutinato solo se l'identificazione vale 3 unità di costo di una investigazione.
- Questo è motivato da $p(H = 1|x_u, x_f) = 0.492$

- Per arrivare a scrutinare *Missing 2* debbo avere $a = 12$.

Bibliografia

- Boreale, M. and Corradi, F. (2016) Searching secrets rationally *International Journal of Approximate Reasoning*, **69**, 133–146.
- Cavallini, D. and Corradi, F. (2005) Forensic identification of relatives of individuals included in a database of DNA profiles. *Biometrika*, **93, 3**, 525-536.
- Corradi, F. (2016). The identification of missing persons making use of DNA profiles. *Applied Statistics*, **27**, 269-281.
- Chung, Y.K., Hu, Y.Q. and Fung, W.K. (2010) Evaluation of DNA mixtures from database search. *Biometrics*, **66**, 233-238.
- Commissario straordinario del Governo per le persone scomparse. *XIII Relazione semestrale*, Direzione Centrale della Polizia Criminale, Roma, 1–45

- Corradi, F. (2010) Mass fatality incident identification based on nuclear DNA evidence *Journal of Machine Learning Research*, **9**, 105–112.
- Dawid, A. P.(2001) Comment on Stockmar's Likelihood ratios for evalu•ating DNA evidence when the suspect is found through a database search *Biometrics*, **57**, 976-978.
- Dawid, A. P., Mortera, J., Pascali, V.L., Van Boxel, D. (2002) Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, 29, 577–595.
- Dràbek, J. (2009) Validation of software for calculating the likelihood ratio for parentage and kinship *Forensic Science International: Genetics*, **3**(2),112–118.
- Egeland, T., Kulle, B., Andreassen, R. (2006) Essen-Möller and identification based on DNA *Chance*, **19**(2), 27–31.

- Gittelson, S., Biederman, A. , Bozza, S., Taroni, F. (2012) The database search problem: A question of rational decision making *Forensic Science International*, **222**, 186–199.
- Lauritzen, S.L. and Sheehan, N. A. (2003) Graphical models for genetic analyses *Statistical Science*, **18**, 489-514
- Slooten, K. and Meester, R (2014) Probabilistic strategies for familial DNA searching. *Journal of the Royal Statistical Society series C*, **63,3**, 361-384.
- Stockmarr, A. (1999) Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search *Biometrics*, **55**, 671-677.
- Taroni,F., Biederman, A., Bozza, S., Garbolino, P., Aitken, C. *Bayesian Network for probabilistic inference and decision analysis in Forensic Science*. John Wiley & Sons, 2014.