Graphical Models for Forensic Identification Problems

Steffen Lauritzen[1]
University of Oxford

Graphical Models with Genetic Applications
Warwick, April 2009

[1] Based on joint work with Cowell, Dawid, Mortera, Vicard, and others

Paternity: Is individual A the father of individual B?

Immigration: Is A the mother of B? Are A and B related at all? If so, how?

Crime: Did person A contribute to a given stain, found at the scene of the crime? Who contributed to the stain?

Disaster: Was A among the individuals found in a grave? How many of a named subset of individuals were in the grave? Who were found in a grave?

An area on a chromosome is a *locus*.

The DNA composition, i.e. a particular sequence of the four *bases*, represented by the letters A, C, G and T, on a given locus is an *allele*.

A locus thus corresponds to a (random) variable and an allele to its realised state.

A DNA *marker* is a known locus where the alleles can be identified in the laboratory.

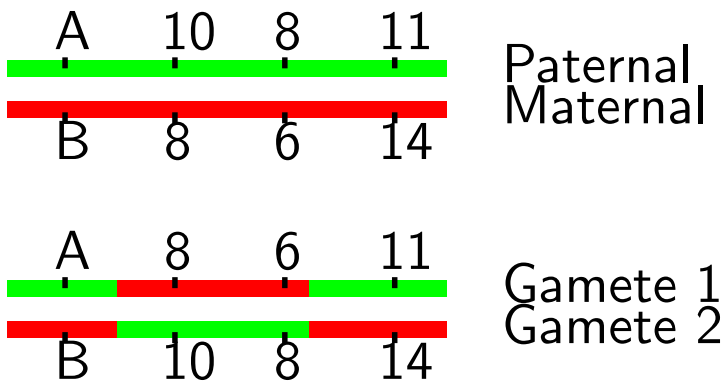
A *genotype* of an individual at a locus is an unordered pair of alleles. One allele comes from the father and one from the mother, but one cannot easily distinguish which is which.

Short Tandem Repeats (STR) are markers with alleles given by integers. If an STR allele is 5, a certain word (e.g. CAGGTG) is repeated exactly 5 times at that locus:

...CAGGTGCAGGTGCAGGTGCAGGTGCAGGTG...

A *DNA profile* is typically a list of genotypes at 10-11 known STR markers.

The homologous chromosome pairs are inherited through the process of forming *gametes*, known as *meiosis:*



A child receives one randomly chosen gamete from each parent to form a new homologous pair.

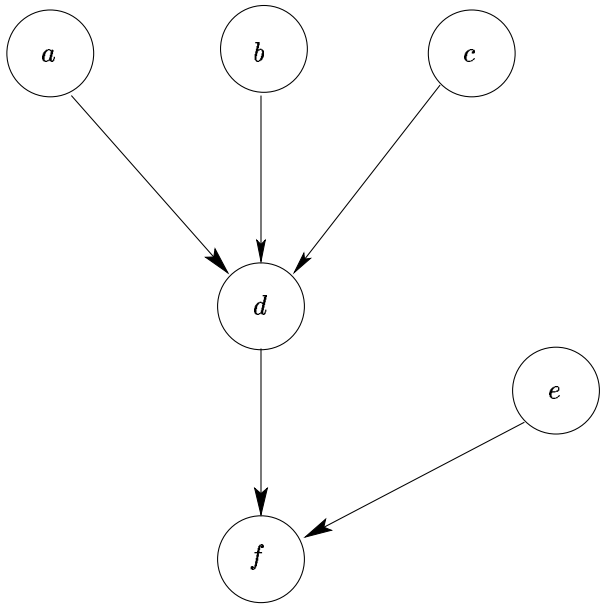
For forensic markers, we can assume independence of alleles within and across markers, as they are located on different chromosomes.

Bayesian network is

- Directed Acyclic Graph (DAG)
- Nodes V represent (random) variables $X_v, v \in V$
- Specify conditional distributions of children given parents: $p(x_v \mid x_{\mathrm{pa}(v)})$
- Joint distribution is then $p(x) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)})$
- Algorithm transforms network into *junction tree* so $p(x_v \mid x_A)$ can be *efficiently computed* for all $v \in V$ and $A \subseteq V$ by *probability propagation.*

Variant calculates revised probabilities $p^*(x_v)$ after *likelihood evidence*

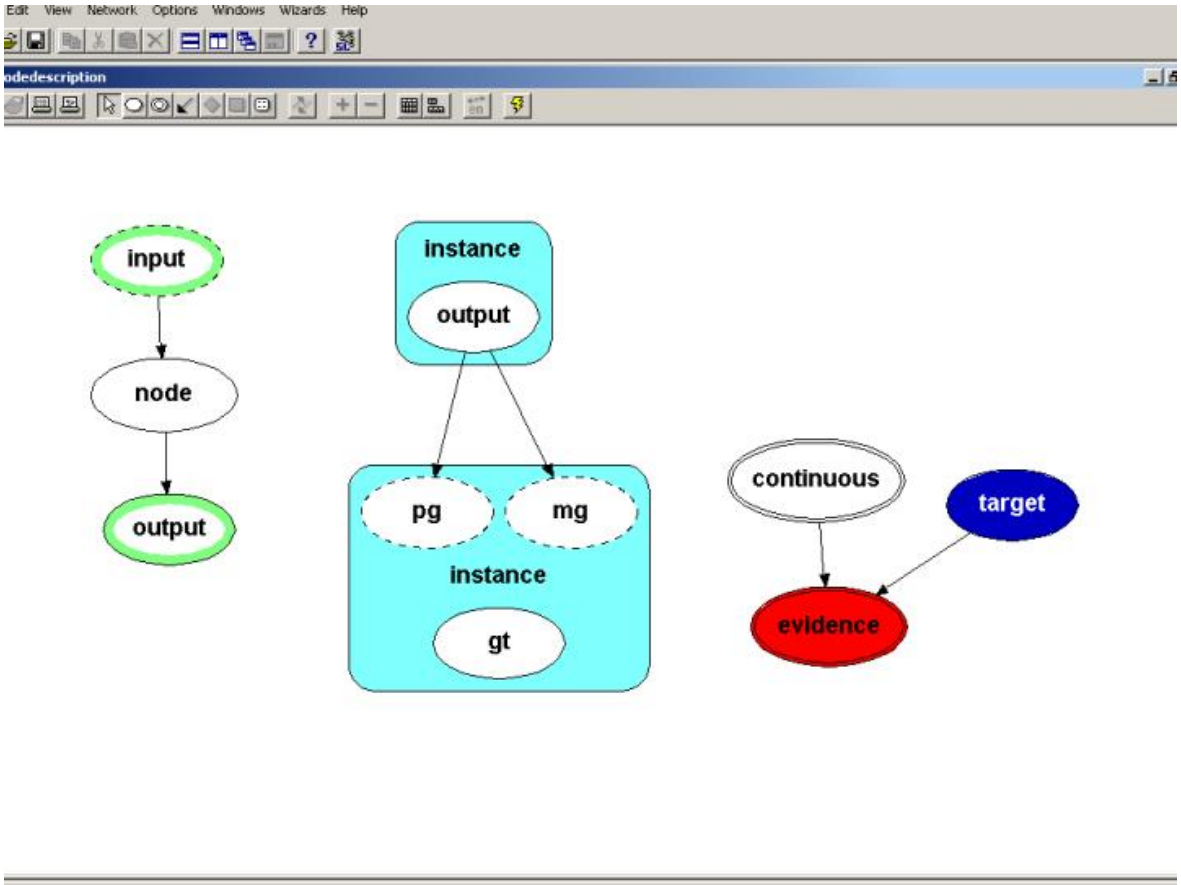
$$p^*(x) \propto \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)}) \prod_{a \in A} L_a(x).$$



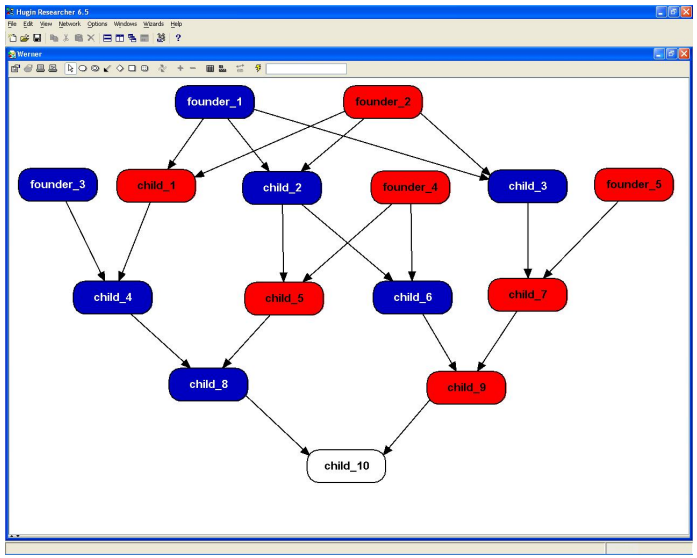
a, b and c (graph) parents of d; f (graph) child of d and e.

$$p(x) = p(x_a)p(x_b)p(x_c)p(x_d \mid x_{\{a,b,c\}})p(x_e)p(x_f \mid x_{\{d,e\}}).$$

- O-O networks have a hierarchical structure where a node can represents a network
- Objects are *instances* of BNs of certain *class*
- Objects have *input* and *output nodes*, and also ordinary nodes
- Instances of a given class have identical conditional probability tables for non-input nodes
- Objects are connected by arrows from output nodes to input nodes. These arrows represent *identity links* whereas arrows between ordinary nodes represent *probabilistic dependence*.

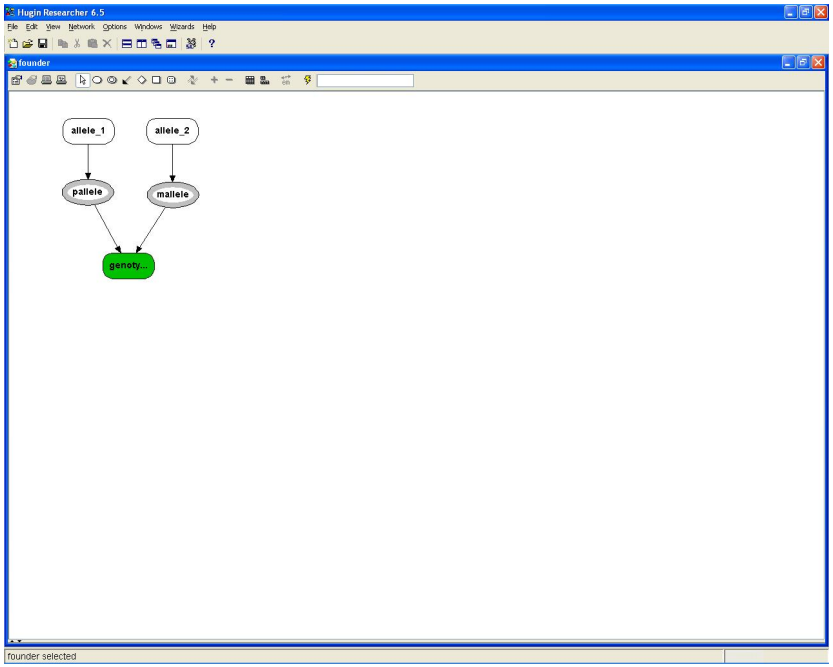


Pedigree as OOBN



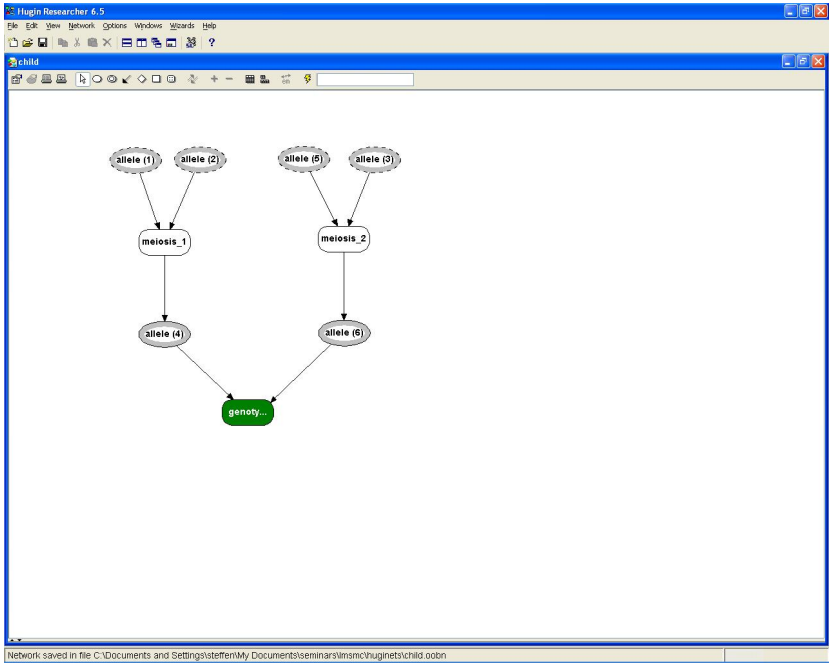
OOBN for a pedigree from study of Werner's syndrome. Each node is itself a BN describing reproductive mechanism.

Founders



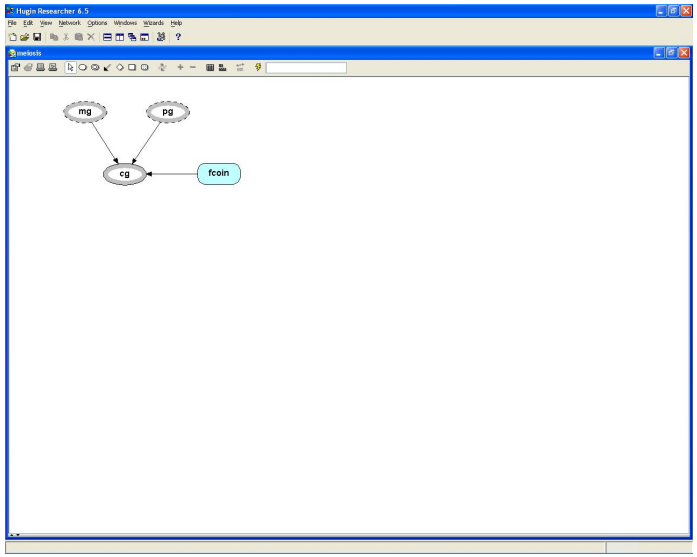
The BN for each founder in a pedigree.

Children



The BN for each child in a pedigree.

Meiosis



The BN for each meiosis in a pedigree. The fair coin is an *inheritance indicator*.

- DNA profiles of *mother*, a *child*, and a *male* individual, known as the *putative father*. Denote this *evidence* by E.
- Query Q to be investigated :

  Is the putative father equal to the true father?

- Weight of evidence reported as a *likelihood ratio:*

$$L = \frac{P(E \mid Q = \texttt{true})}{P(E \mid Q = \texttt{false})}.$$

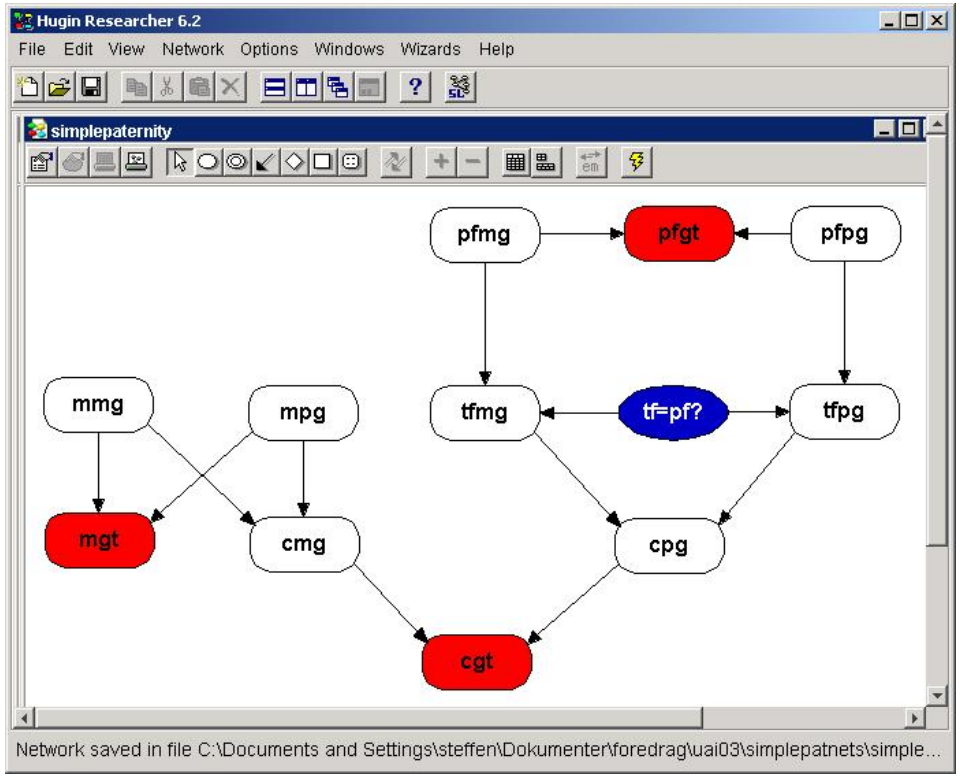
- Make BN with $P(E \mid Q = \texttt{true})$ determined by laws of inheritance and $P(E \mid Q = \texttt{false})$ assuming random genes of putative father.
- Let $P(Q = \texttt{true}) = P(Q = \texttt{false})$ so we have

$$L = \frac{P(E \mid Q = \texttt{true})}{P(E \mid Q = \texttt{false})} = \frac{P(Q = \texttt{true} \mid E)}{P(Q = \texttt{false} \mid E)}$$

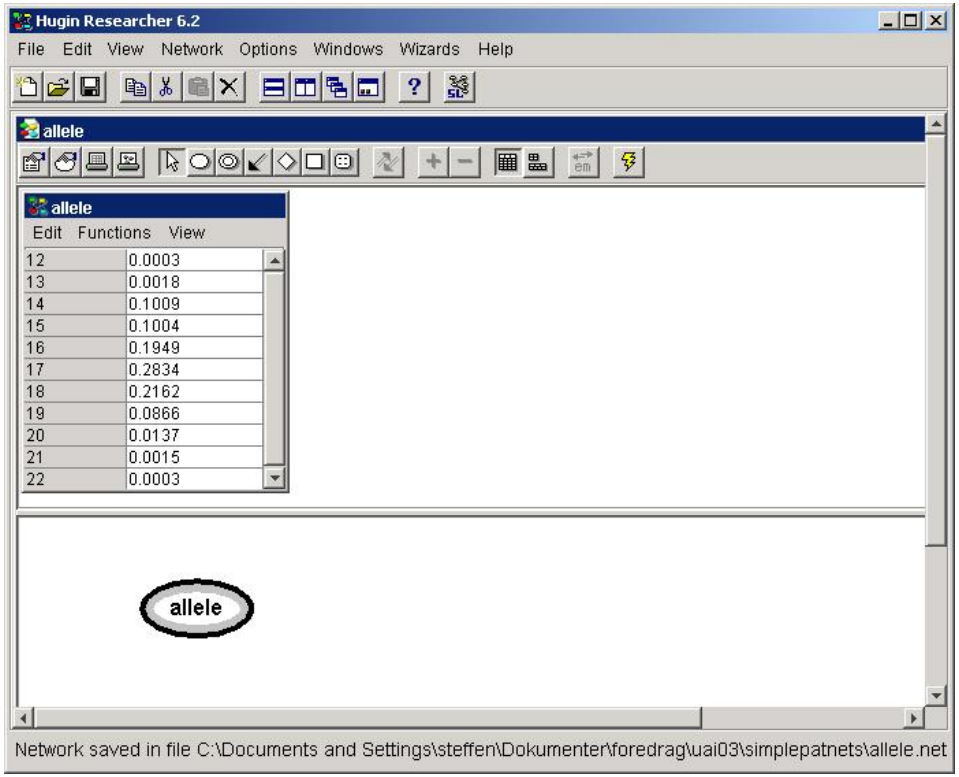
  and compute the latter by probability propagation.

We can make a network for each independent marker and multiply likelihood ratios, or we can make a network incorporating all markers at once.

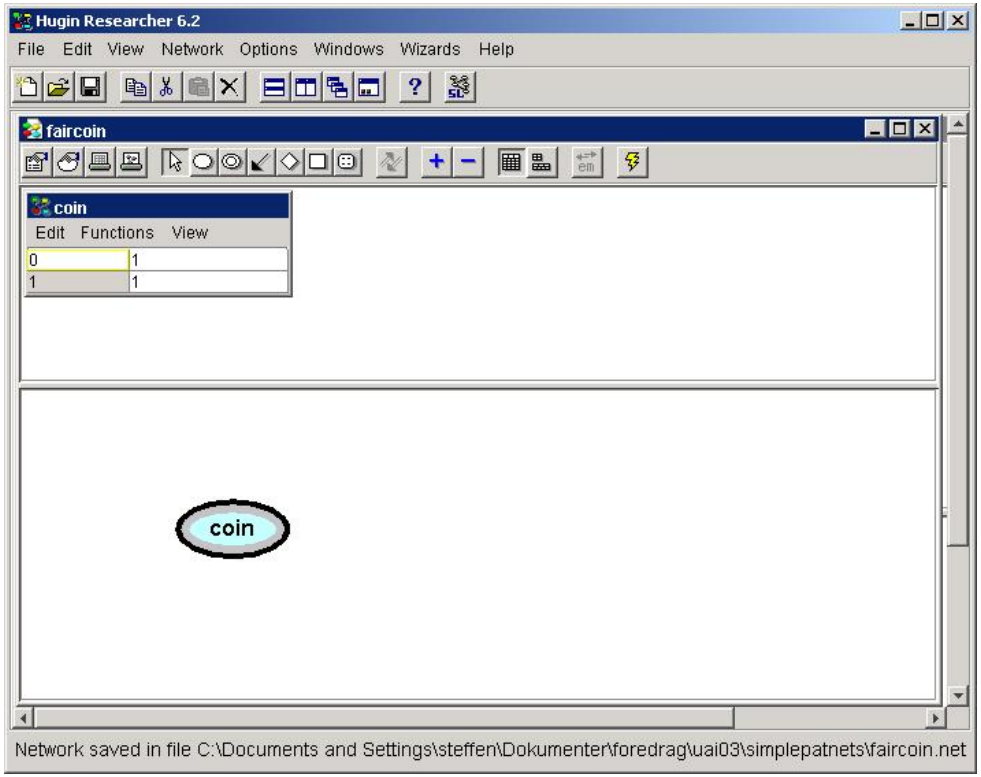
OOBN for paternity case: single marker



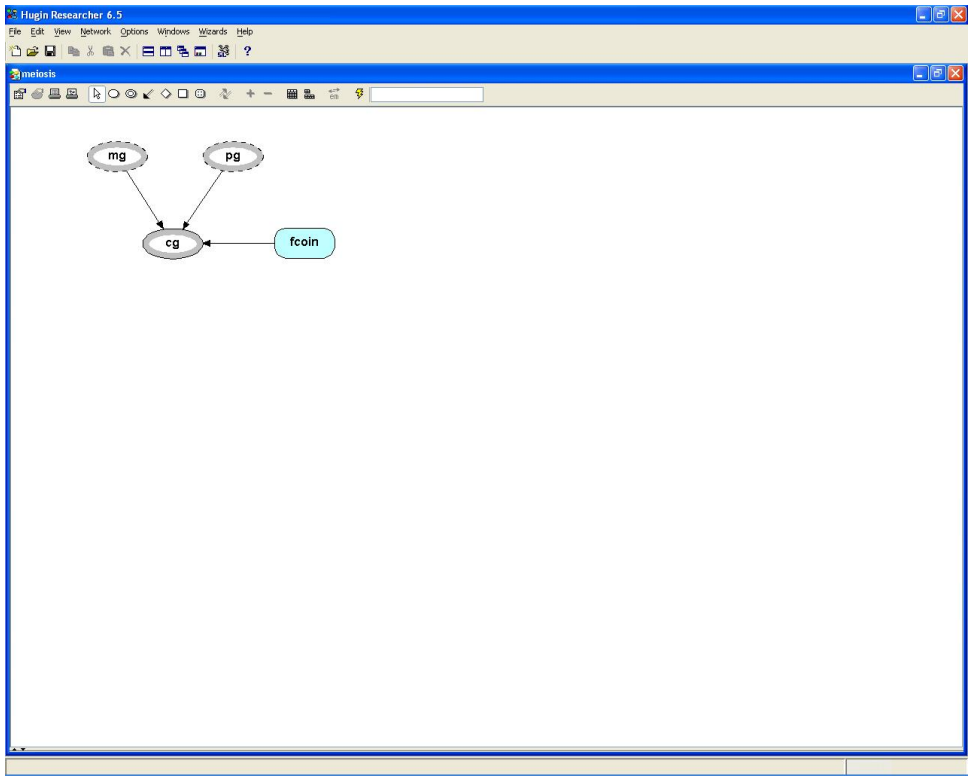
Allele



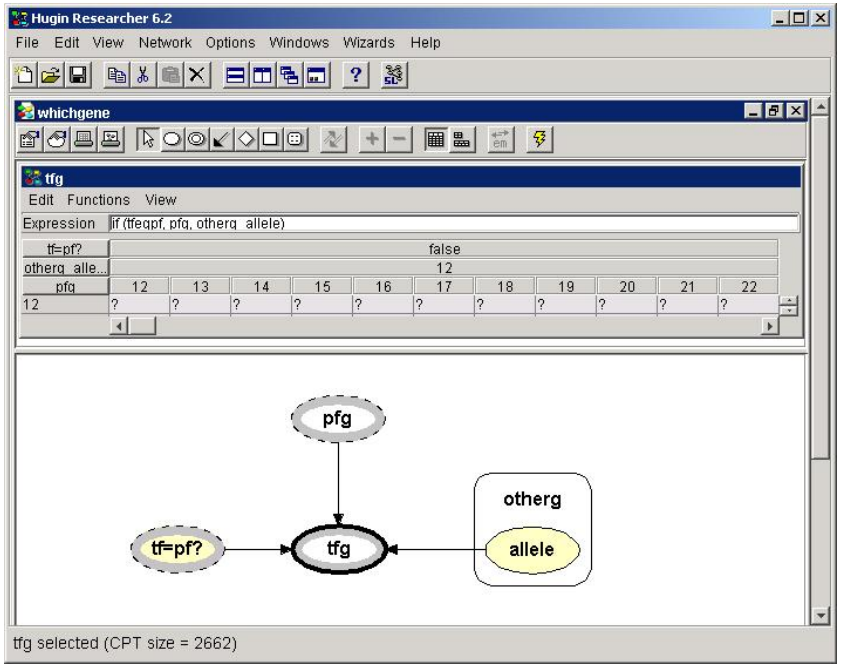
Faircoin



Meiosis

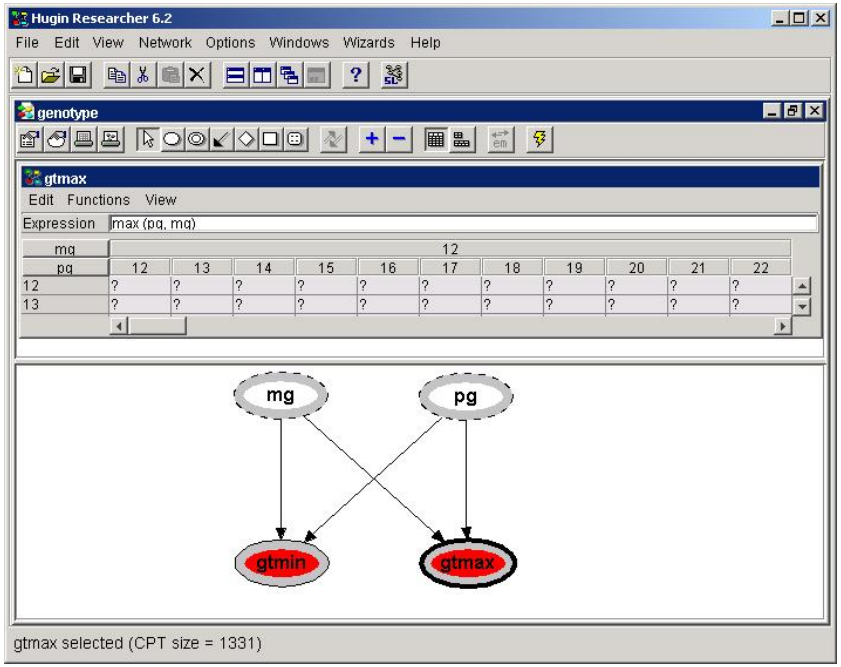


Who is the father?



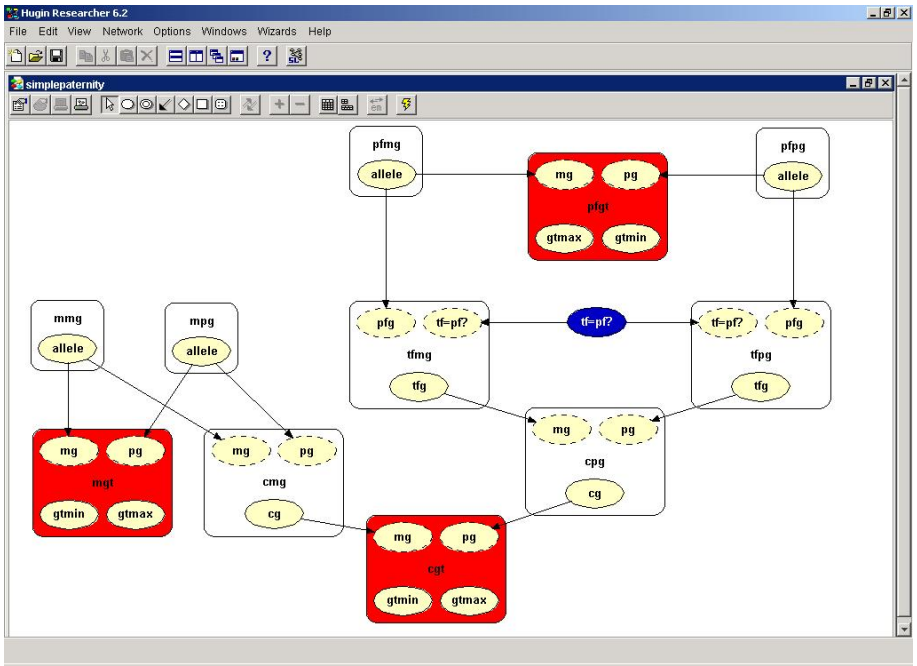
Is the allele from the putative father or random?

Genotype

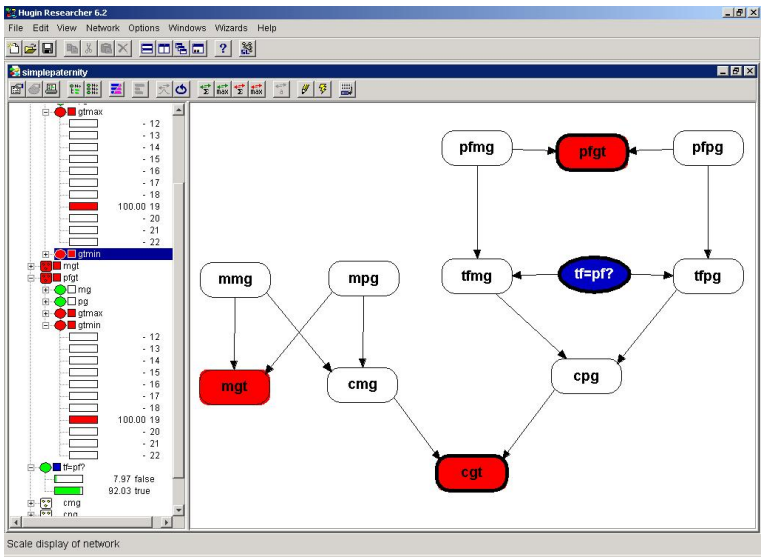


Observation of the smallest and largest allele

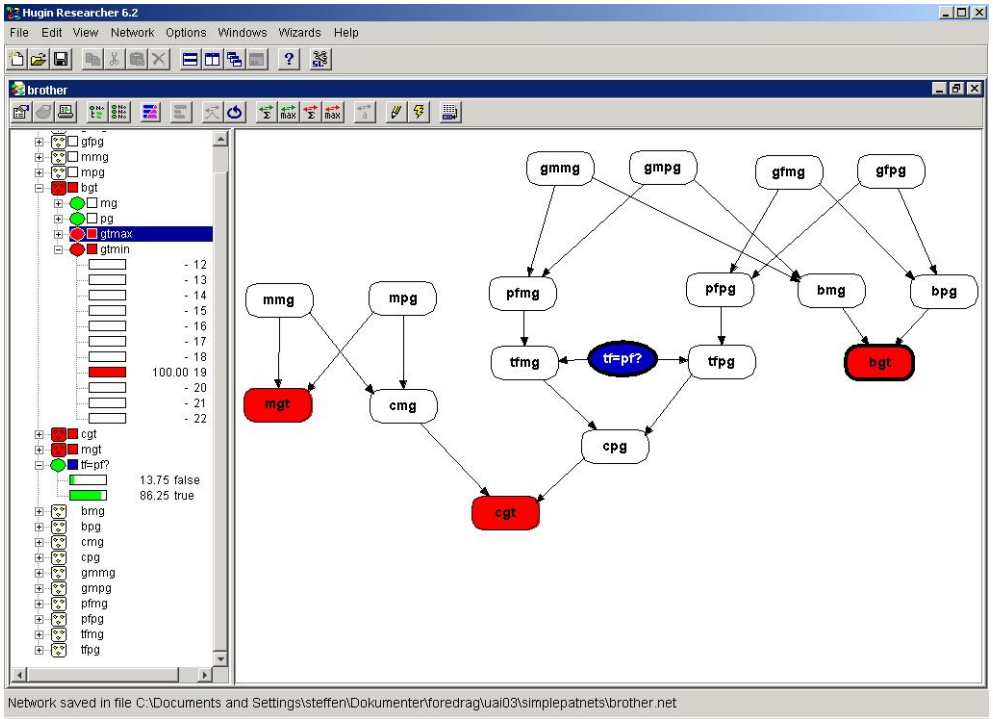
Expanded OOBN



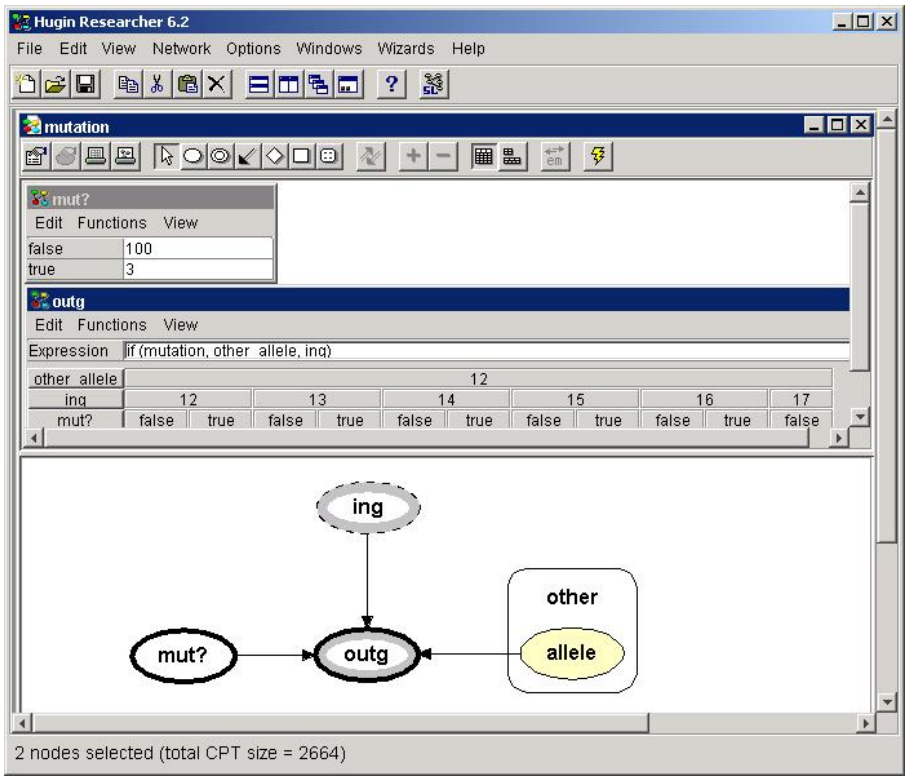
Results



Mother: $(15, 16)$, child: $(15, 19)$, male: $(19, 19)$;
$L = 92.03/7.97 = 11.55$.

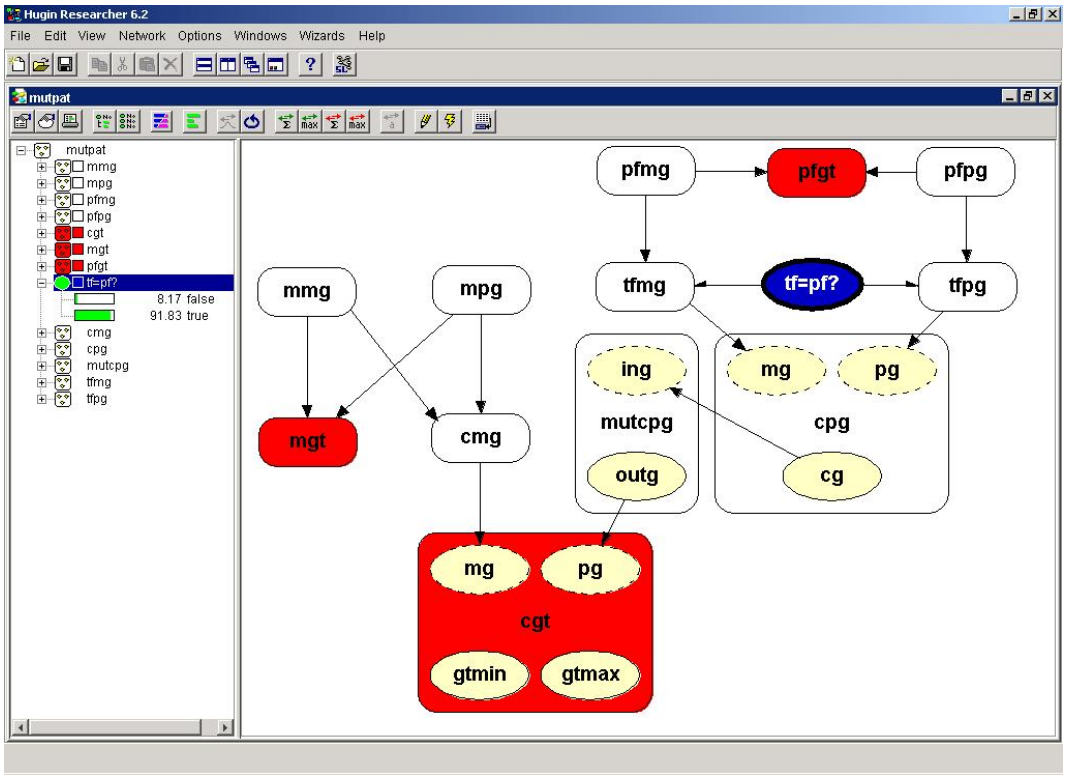


Brother of pf: $(19, 19)$; $L = 86.25/13.75 = 6.27$.



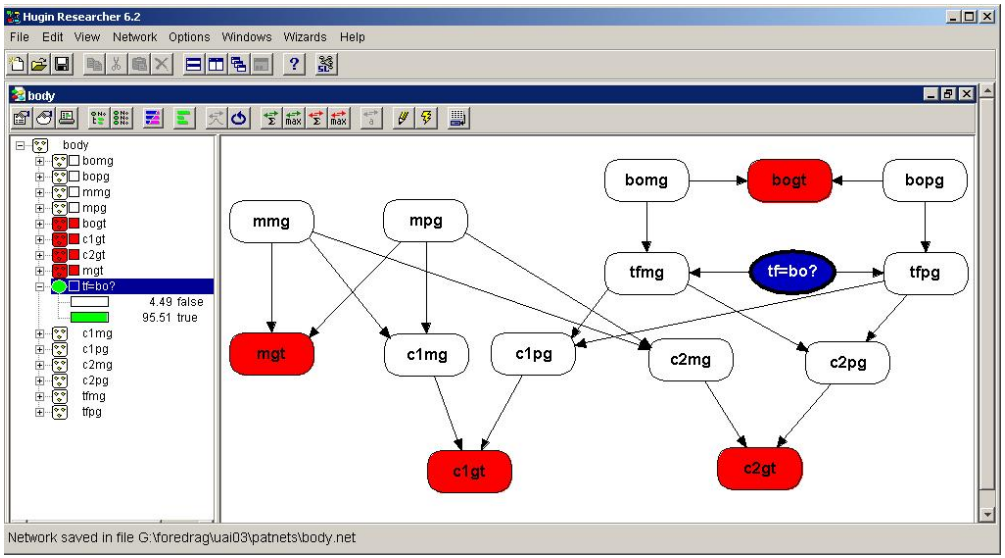
Possible mutation in transmission of alleles

Mutation in male germline



Identification of a *single* dead body is not very different for paternity cases.
For example, if a missing person is known to be a specific member of a family (e.g. the father of two children) and DNA profiles can be found for the body, the mother, and the two children, a minor modification of the paternity network yields the solution.
Problems of identification involving *more than one* body, such as in mass graves and in disasters are more difficult because of their complexity.

Unidentified body



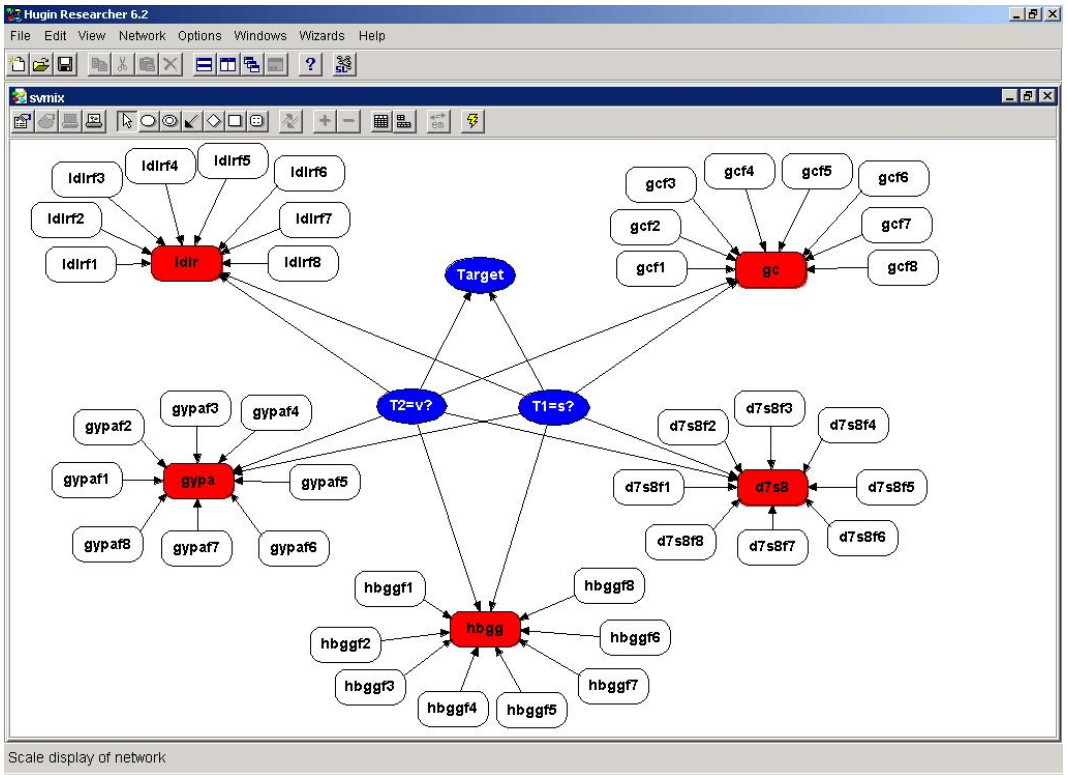
Is the body father of the two children? Same data as for paternity.
Second child $(16, 19)$; $L = 95.51/4.49 = 21.27$.

In *criminal cases* it is not uncommon to find traces where the DNA is a mixture of contributions from several individuals.
This happens for example in *rape* cases, where a vaginal swab typically will contain DNA from the victim as well as the perpetrator, and possibly also from a consensual partner.
But it is also common e.g. in *robberies*, where a balaclava is found on the scene of the crime; these have often been used by several persons.

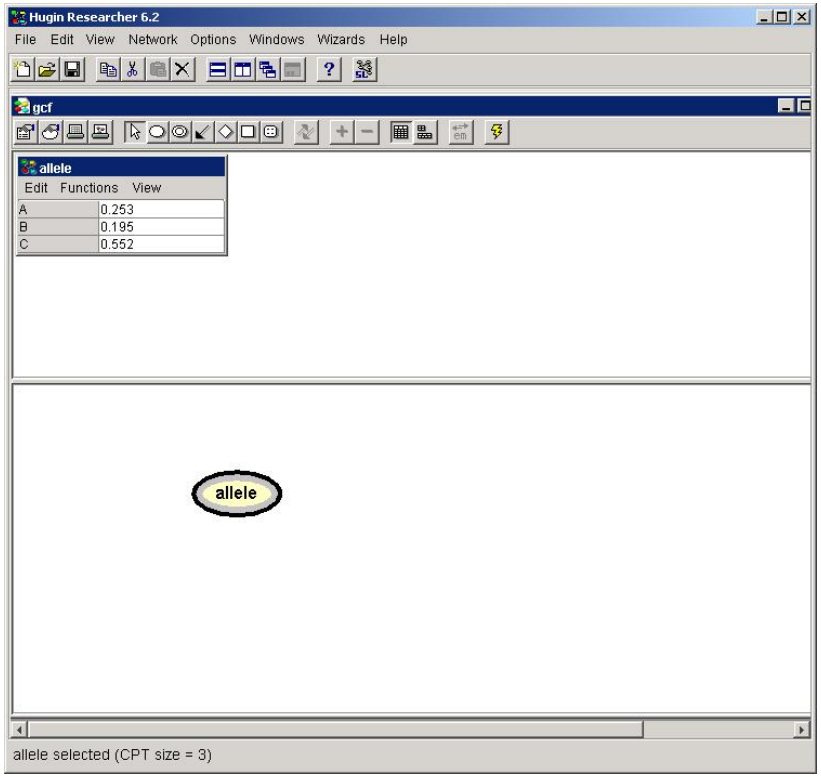
Weir's example

	Marker				
Profile	LDLR	GYPA	HBGG	D7S8	Gc
trace:	B	AB	AB	AB	ABC
victim:	B	AB	AB	AB	AC
suspect:	B	A	A	A	B
p_A	0.433	0.538	0.566	0.543	0.253
p_B	0.567	0.462	0.429	0.457	0.195
p_C	0	0	0.005	0	0.552

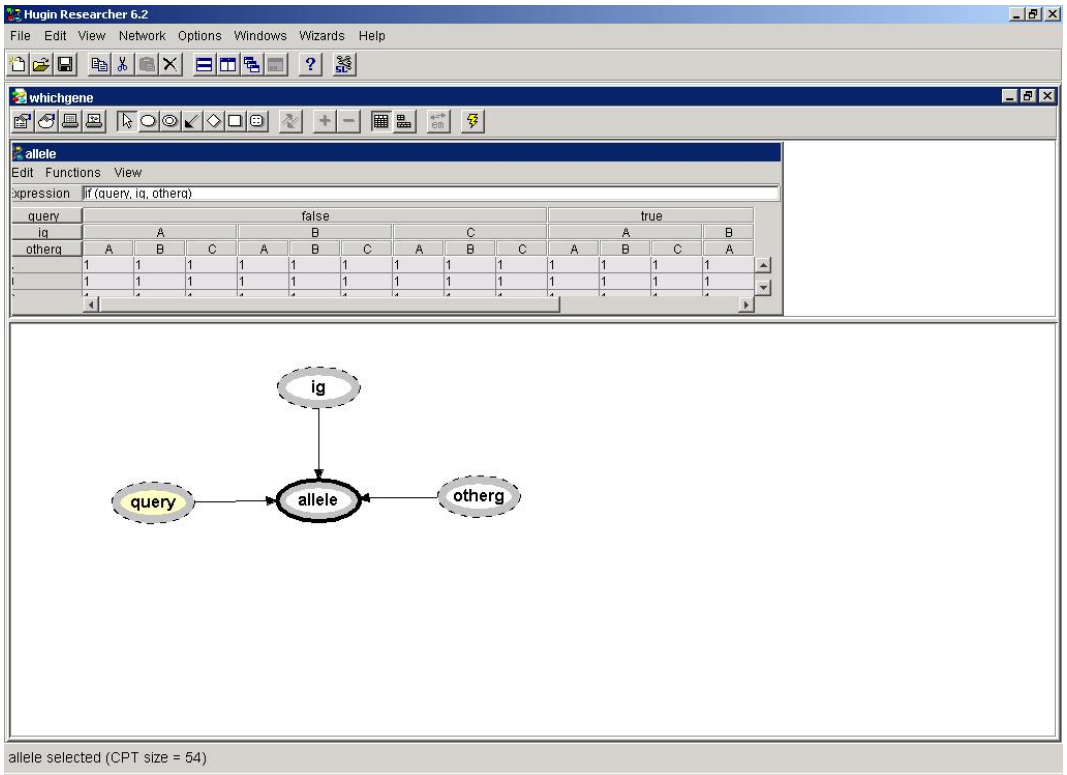
Mixture net for all markers



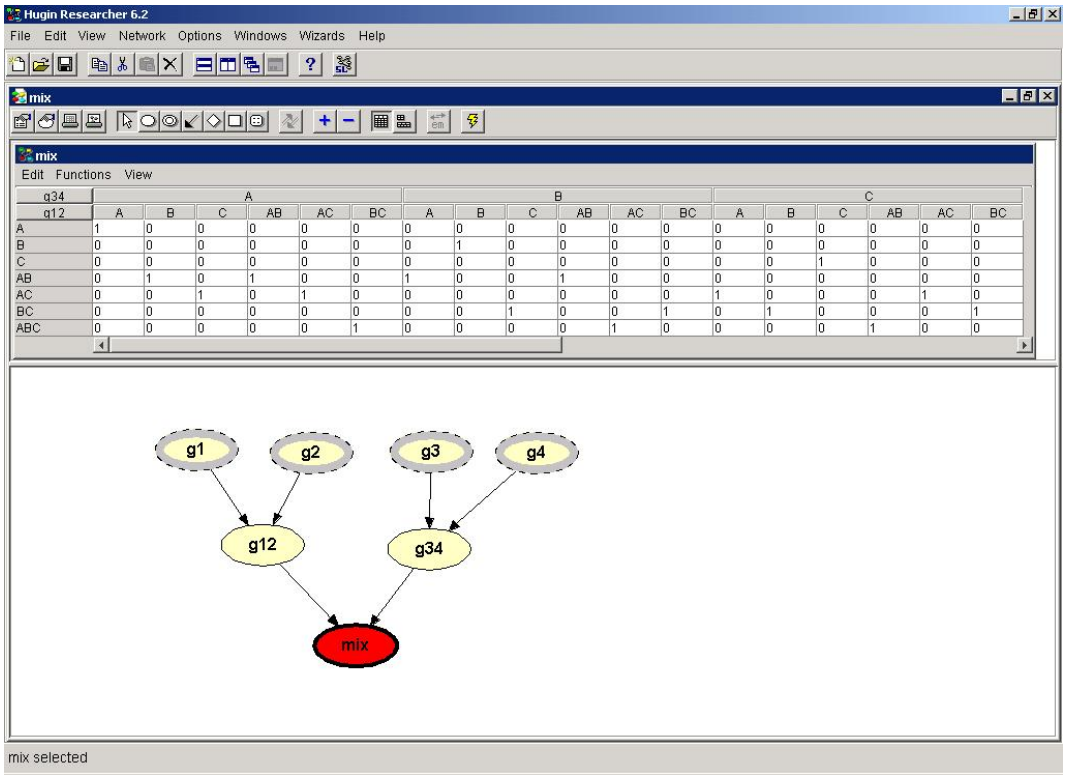
One founder for every marker



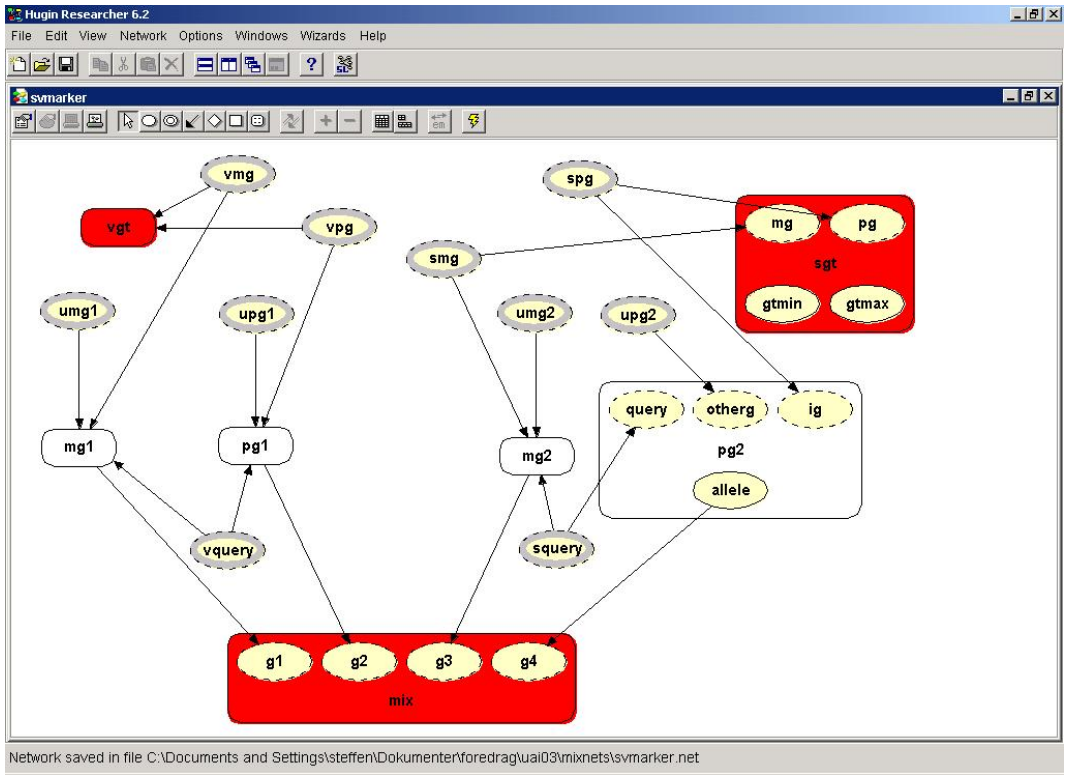
Who contributed to the mixture?



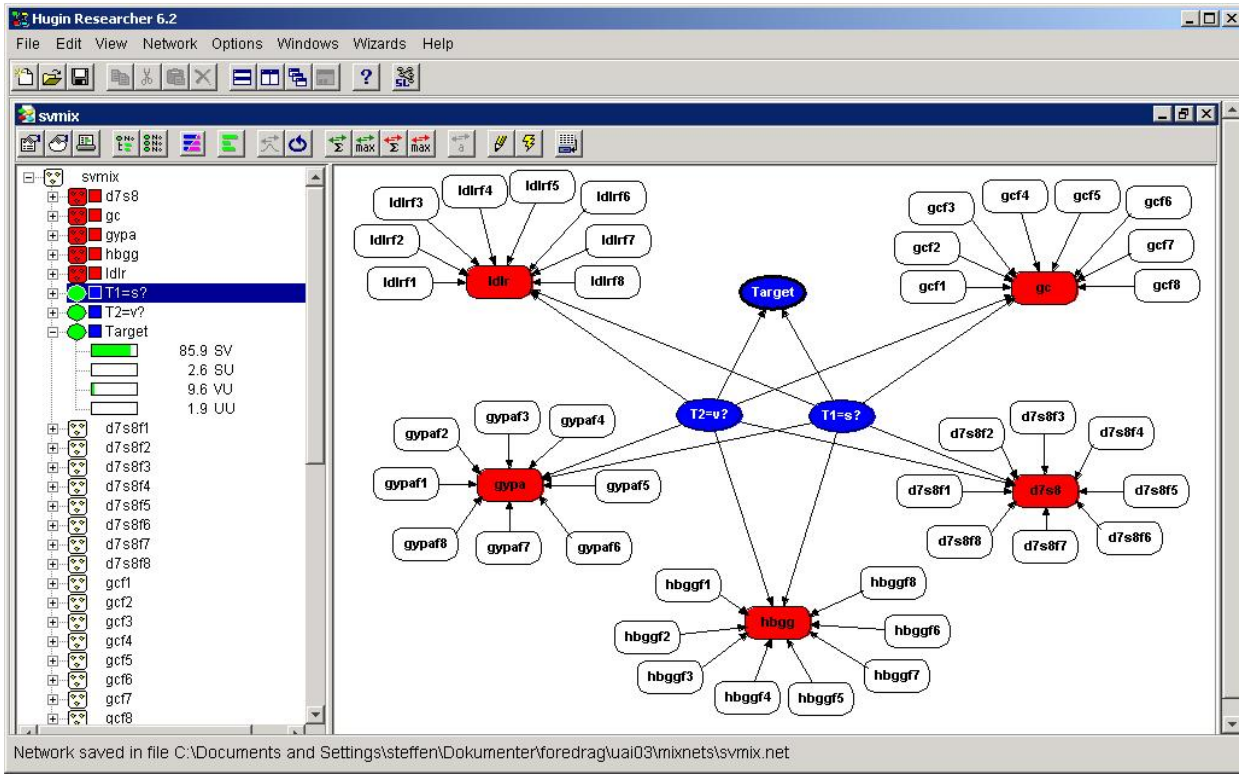
Mixing the DNA



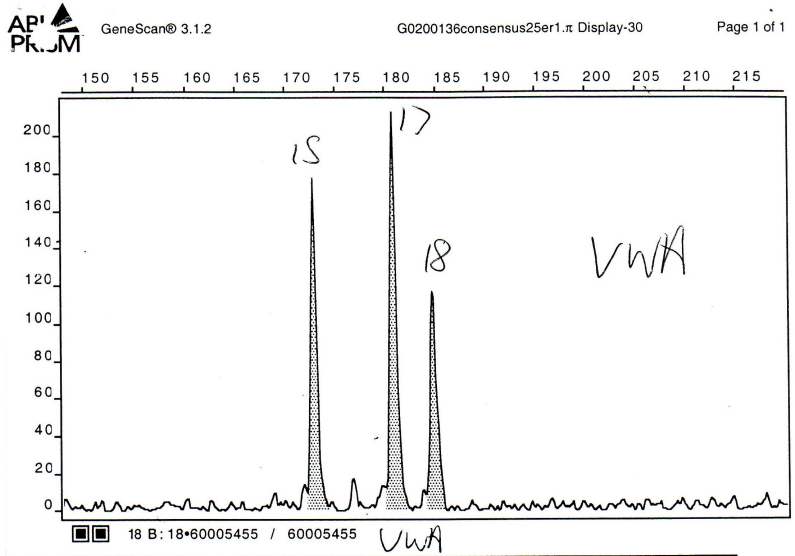
Network for markers



Results from all markers



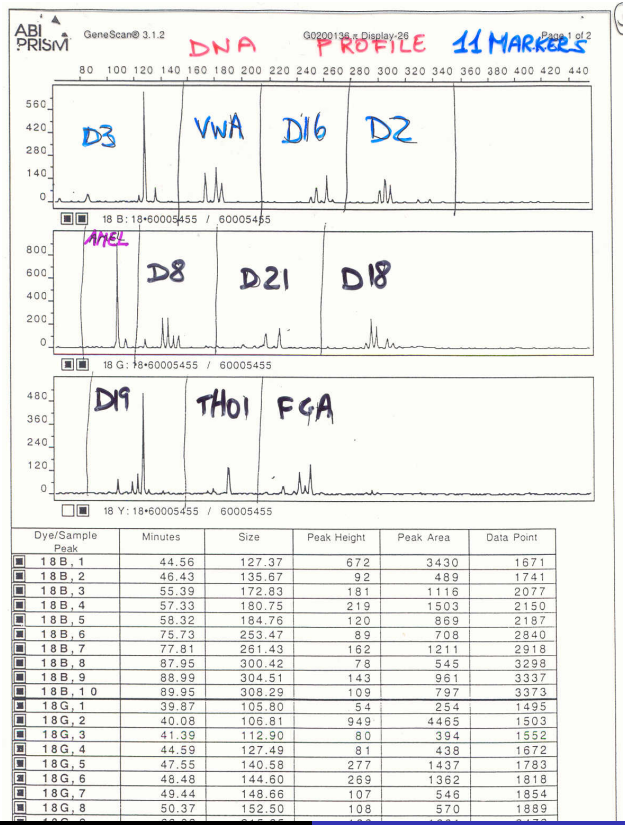
Two-person DNA mixture profile



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
18B,6	55.39	172.83	15 181	1116	2077
18B,7	57.33	180.75	17 219	1503	2150
18B,8	58.32	184.76	18 120	869	2187

Marker vWA with *allele repeat number* $\{15, 17, 18\}$, *peak area* and peak height

DNA profile on 10 markers + Amelogenin



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
18B, 1	44.56	127.37	672	3430	1671
18B, 2	46.43	135.67	92	489	1741
18B, 3	55.39	172.83	181	1116	2077
18B, 4	57.33	180.75	219	1503	2150
18B, 5	58.32	184.76	120	869	2187
18B, 6	75.73	253.47	89	708	2840
18B, 7	77.81	261.43	162	1211	2918
18B, 8	87.95	300.42	78	545	3298
18B, 9	88.99	304.51	143	961	3337
18B, 10	89.95	308.29	109	797	3373
18G, 1	39.87	105.80	54	254	1495
18G, 2	40.08	106.81	949	4465	1503
18G, 3	41.39	112.90	80	394	1552
18G, 4	44.59	127.49	81	438	1672
18G, 5	47.55	140.58	277	1437	1783
18G, 6	48.48	144.60	269	1362	1818
18G, 7	49.44	148.66	107	546	1854
18G, 8	50.37	152.50	108	570	1889

Data from a 1:1 mixture of two individuals p1 and p2.

Marker	Alleles	Peak area	Rel. Weight	p1 gt	p2 gt
D2	17	37624	0.573	17	17
	23	9742	0.148		23
	25	18316	0.279	25	
D3	14	56692	0.344	14	
	15	55256	0.335		15
	16	52793	0.321	16	
D8	8	43569	0.412	8	
	9	17423	0.165		9
	13	16227	0.154		13
	14	28488	0.269	14	

A DNA profile gives information on: *allele repeat number* and corresponding *peak area.*

The *peak weight* W_a is the peak area at allele a multiplied by its allele number, the latter to correct for *preferential amplification.*

Consider a mixture made up from individuals $i \in I$.

- The (pre-amplification) proportions of DNA $\theta = \{\theta_i, i \in I\}$ are assumed *constant across markers*,

Consider a mixture made up from individuals $i \in I$.

- The (pre-amplification) proportions of DNA $\theta = \{\theta_i, i \in I\}$ are assumed *constant across markers*,
- the weight W_{ia} roughly proportional to the *amount of DNA* of type a possessed by individual i;

Consider a mixture made up from individuals $i \in I$.

- The (pre-amplification) proportions of DNA $\theta = \{\theta_i, i \in I\}$ are assumed *constant across markers*,
- the weight W_{ia} roughly proportional to the *amount of DNA* of type a possessed by individual i;
- W_a is the *sum* of the allele a weights of all contributors.

Consider a mixture made up from individuals $i \in I$.

- The (pre-amplification) proportions of DNA $\theta = \{\theta_i, i \in I\}$ are assumed *constant across markers*,
- the weight W_{ia} roughly proportional to the *amount of DNA* of type a possessed by individual i;
- W_a is the *sum* of the allele a weights of all contributors.
- W_{ia}, are independent for fixed θ and *Gamma distributed*: $W_{ia} \sim \Gamma(\rho\gamma_i n_{ia}, \eta)$, where

Consider a mixture made up from individuals $i \in I$.

- The (pre-amplification) proportions of DNA $\theta = \{\theta_i, i \in I\}$ are assumed *constant across markers*,
- the weight W_{ia} roughly proportional to the *amount of DNA* of type a possessed by individual i;
- W_a is the *sum* of the allele a weights of all contributors.
- W_{ia}, are independent for fixed θ and *Gamma distributed*: $W_{ia} \sim \Gamma(\rho\gamma_i n_{ia}, \eta)$, where
  - $\gamma_i = \gamma\theta_i$ is the *amount of DNA* from individual i in mixture;

Consider a mixture made up from individuals $i \in I$.

- The (pre-amplification) proportions of DNA $\theta = \{\theta_i, i \in I\}$ are assumed *constant across markers*,
- the weight W_{ia} roughly proportional to the *amount of DNA* of type a possessed by individual i;
- W_a is the *sum* of the allele a weights of all contributors.
- W_{ia}, are independent for fixed θ and *Gamma distributed*: $W_{ia} \sim \Gamma(\rho\gamma_i n_{ia}, \eta)$, where
  - $\gamma_i = \gamma\theta_i$ is the *amount of DNA* from individual i in mixture;
  - θ_i is the *proportion of DNA* (fraction) from individual i;

Consider a mixture made up from individuals $i \in I$.

- The (pre-amplification) proportions of DNA $\theta = \{\theta_i, i \in I\}$ are assumed *constant across markers*,
- the weight W_{ia} roughly proportional to the *amount of DNA* of type a possessed by individual i;
- W_a is the *sum* of the allele a weights of all contributors.
- W_{ia}, are independent for fixed θ and *Gamma distributed*: $W_{ia} \sim \Gamma(\rho\gamma_i n_{ia}, \eta)$, where
  - $\gamma_i = \gamma\theta_i$ is the *amount of DNA* from individual i in mixture;
  - θ_i is the *proportion of DNA* (fraction) from individual i;
  - n_{ia} is the *number of alleles* of type a carried by individual i;

Consider a mixture made up from individuals $i \in I$.

- The (pre-amplification) proportions of DNA $\theta = \{\theta_i, i \in I\}$ are assumed *constant across markers*,
- the weight W_{ia} roughly proportional to the *amount of DNA* of type a possessed by individual i;
- W_a is the *sum* of the allele a weights of all contributors.
- W_{ia}, are independent for fixed θ and *Gamma distributed*: $W_{ia} \sim \Gamma(\rho\gamma_i n_{ia}, \eta)$, where
  - $\gamma_i = \gamma\theta_i$ is the *amount of DNA* from individual i in mixture;
  - θ_i is the *proportion of DNA* (fraction) from individual i;
  - n_{ia} is the *number of alleles* of type a carried by individual i;
  - η determines *scale* and ρ is the *amplification factor*.

R_a denotes *relative weights* $R_a = W_{+a}/W_{++}$ so

$$\{R_a, a \in A\} \sim Dir(\rho B_a, a \in A),$$

where $B_a = \gamma \sum_i \theta_i n_{ia}$ is the weighted allele number and $B_+ = \sum_a B_a = 2\gamma$ is twice the total amount of DNA γ.

Note η disappears and

$$\mathbb{E}(R_a) = \mu_a = B_a/B_+ = \sum_i \theta_i n_{ia}/2$$

and

$$\mathbb{V}(R_a) = \mu_a(1-\mu_a)/(\rho B_+ + 1) = \sigma^2 \mu_a(1-\mu_a).$$

We used $\sigma^2 = 0.01$ which conforms with values of a minor/major peak area ratio reported in the literature.

Evidential calculation

Population gene frequencies are assumed to be *known*. The *evidence* is for example:

$$\mathcal{E} = \{\text{sgt, vgt, mixture profile}\},$$

where sgt,vgt are genotypes of a *suspect* and a *victim*.

The *hypotheses* are for example

$$H_0 : s\&v, \quad H_1 : U\&v.$$

The *weight of the evidence* is the likelihood ratio:

$$LR = \frac{\Pr(\mathcal{E} \mid H_0)}{\Pr(\mathcal{E} \mid H_1)} = \frac{\Pr(H_0 \mid \mathcal{E})}{\Pr(H_1 \mid \mathcal{E})} \frac{\Pr(H_1)}{\Pr(H_0)}.$$

Separation of DNA profiles

Identifying the genotype of each of the possibly unknown contributors to the mixture.

Calculate either

$$P\{\text{sgt} \mid \text{vgt}, \text{mixture}\}$$

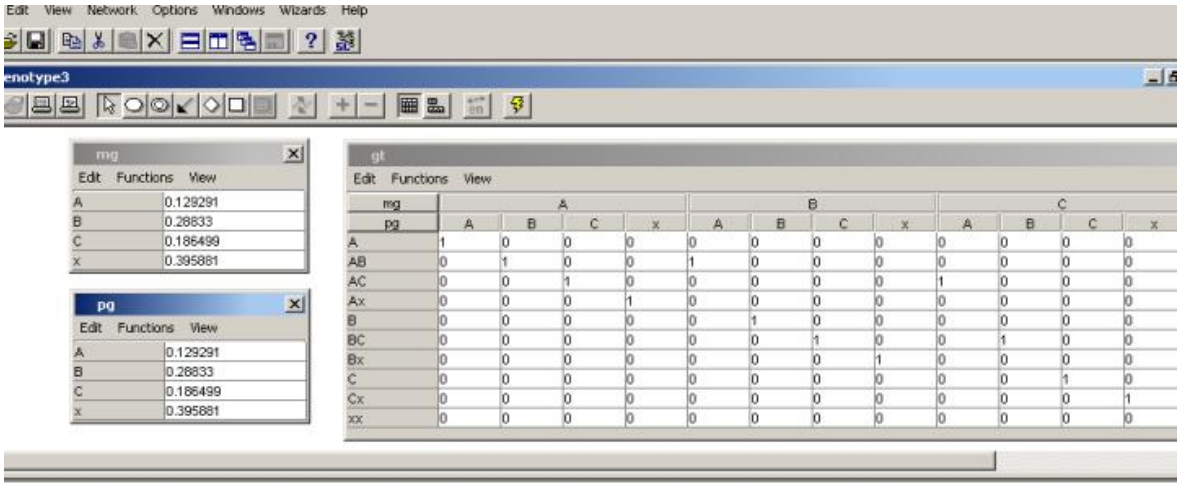
or

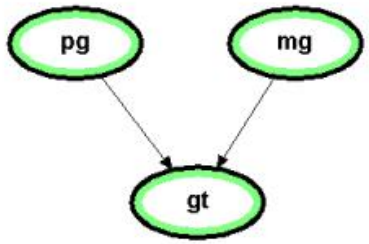
$$P\{\text{p1gt}, \text{p2gt} \mid \text{mixture}\}$$

and find most probable combination.

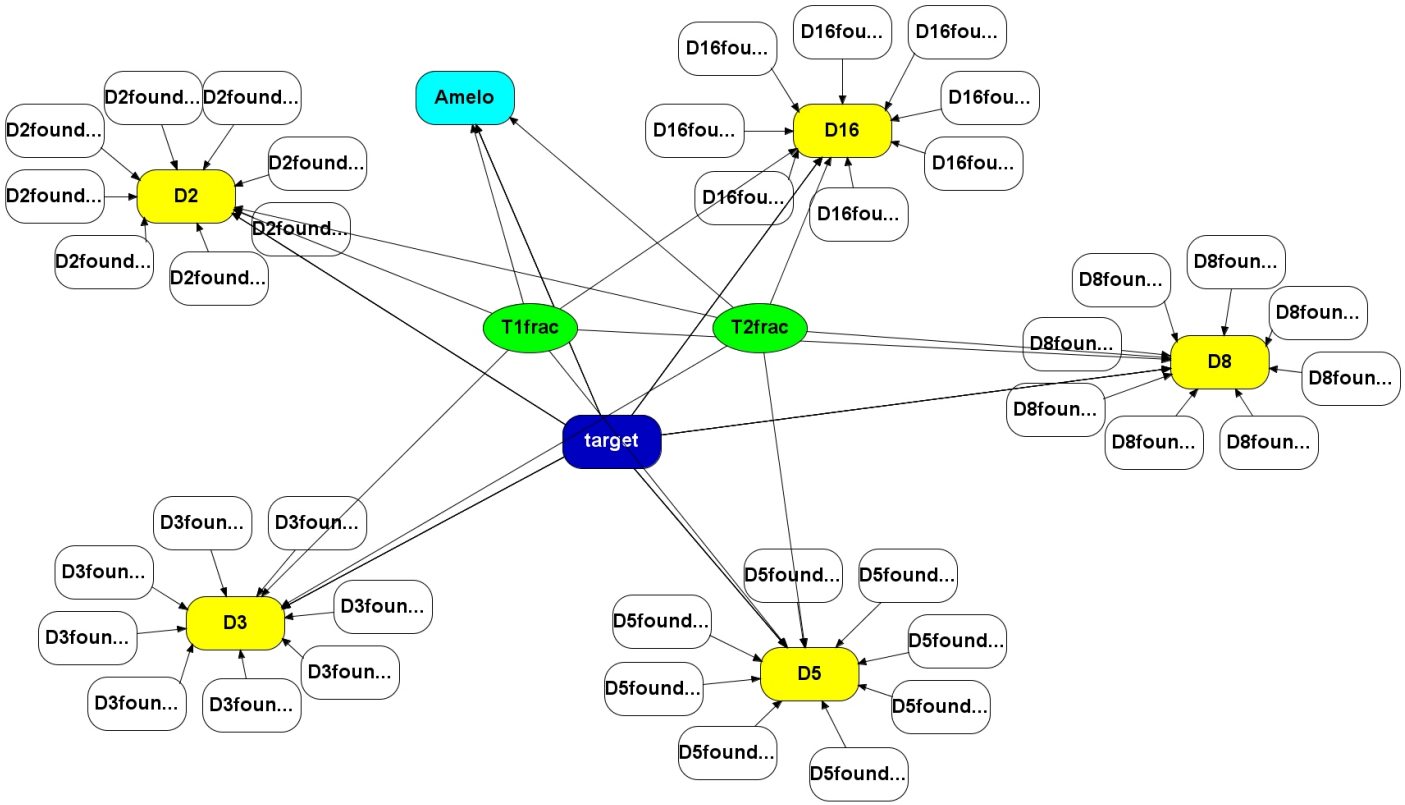
Important in investigative phase.

So is evidential calculation which can be used to decide whether it is worthwhile to search for supporting evidence against a suspect.

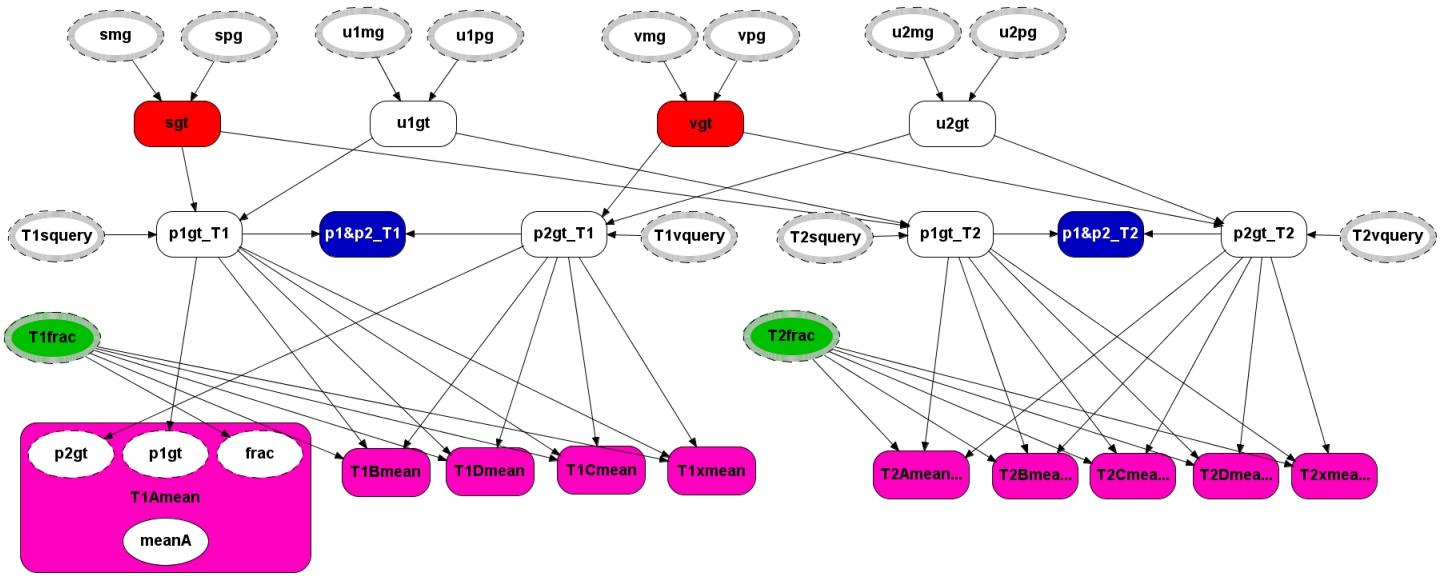




Master network for two DNA traces



Marker network for two DNA traces



Representation of evidence in peak areas

Data on peak areas are thus for each marker m of the form

$$R_a = r_a, a \in A.$$

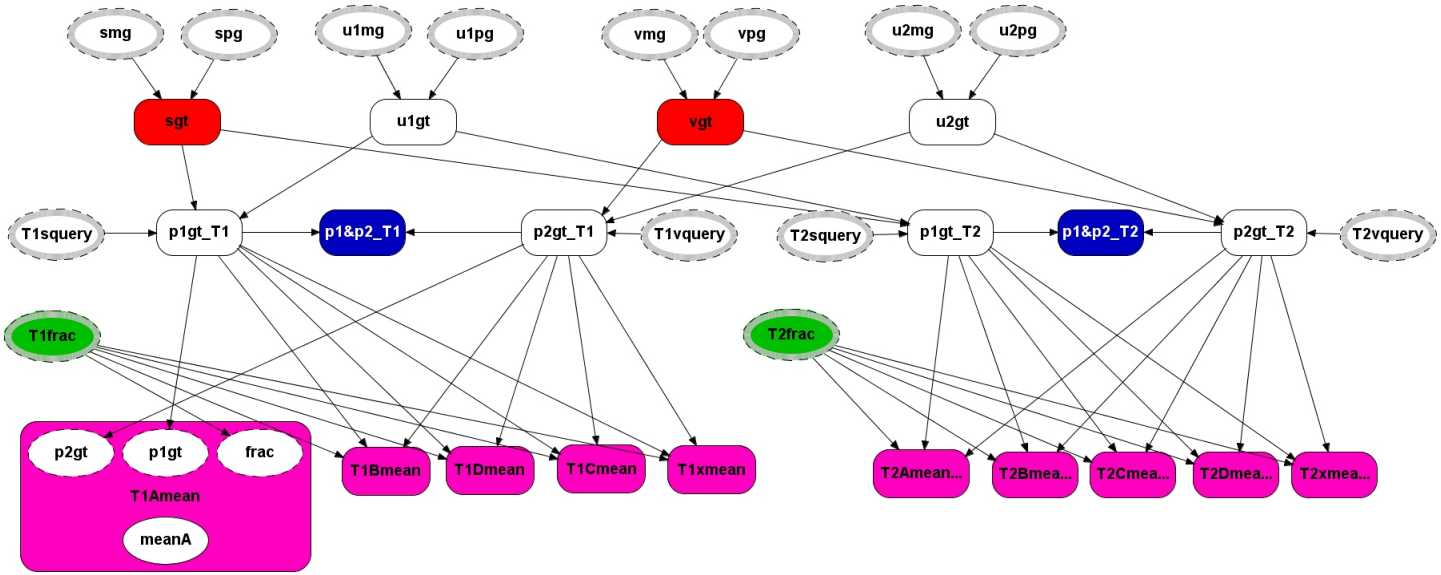
Associated *evidence* is represented in the form of a *likelihood function* on the unknown mean vector $\mu_a, a \in A$ as

$$L \propto \prod_{a \in A} \frac{r_a^{2\rho\gamma\mu_a - 1}}{\Gamma(2\rho\gamma\mu_a)} \propto \prod_{a \in A} \frac{r_a^{\mu_a(\sigma^{-2}-1)}}{\Gamma\{\mu_a(\sigma^{-2}-1)\}}.$$

where we have used that $B_a = 2\gamma\mu_a$ and $\sigma^2 = (\rho B_+ + 1)^{-1}$.

Thus the *joint likelihood evidence factorizes into evidence for each allele a separately.*

Representing evidence from peak areas



The following *likelihood evidence* is inserted in the mean nodes and propagated throughout the network

Prepared mixture in 1:1 ratio which is hard to separate. (Effective fraction $\theta \neq 0.5$)? *Predicted genotypes of p1 and p2 correct on all 11 markers* (excerpt).

Marker	p1 gt	p2 gt	Prob.
D2	17 25	17 23	0.458
D3	14 16	15 15	0.815
D8	8 14	9 13	0.647
D16	9 11	11 11	0.608

Incorrect identifications in red.

Correct on	T1 only 1:1? all	T2 only 1:1 9 out of 11 markers	T1 & T2 all
D2	0.4582	0.3838	0.6956
D3	0.8152	*0.4854*	**0.8531**
D8	0.6471	0.4831	0.7357
D16	0.6078	0.7534	0.7877

Note the *increase in probabilities for D3*, which was *incorrectly* identified when analysing T2 by itself.

Assuming common contributors, using the profile of one contributor in all three separations.

Correct on	T1 only 1:1:1 3 out of 14	T2 only 1:2:5 11 out of 14	T1 & T2 all
D2	*0.178*	1.000	1.000
D3	*0.285*	0.768	0.987
D5	*0.432*	*0.190*	**0.883**
D16	*0.171*	*0.299*	**0.967**

Note the *increase in probabilities* for the profiles *on markers D5 and D16*, none of which were correctly identified with a single mixture analysis.

- *Identification* and *separation* problems can be solved in the same network.

- *Identification* and *separation* problems can be solved in the same network.
- *All uncertainties* associated with the analysis are quantified.

- *Identification* and *separation* problems can be solved in the same network.
- *All uncertainties* associated with the analysis are quantified.
- *Modularity and flexibility* of the OOBN allows easy extension to similar but different situations.

- *Identification* and *separation* problems can be solved in the same network.
- *All uncertainties* associated with the analysis are quantified.
- *Modularity and flexibility* of the OOBN allows easy extension to similar but different situations.
- Sensitivity to the scaling factors γ, σ^2 used to model variation in amplification and measurement processes. Calibration needed.

- *Identification* and *separation* problems can be solved in the same network.
- *All uncertainties* associated with the analysis are quantified.
- *Modularity and flexibility* of the OOBN allows easy extension to similar but different situations.
- Sensitivity to the scaling factors γ, σ^2 used to model variation in amplification and measurement processes. Calibration needed.
- Need to incorporate artifacts such as stutter peaks and drop-outs. Recent work shows success in this.

Cowell, R. G., Lauritzen, S. L. and Mortera, J (2007). Identification and separation of DNA mixtures using peak area information. *Forensic Science International* **166**, 28–34.

Cowell, R. G., Lauritzen, S.L. and Mortera, J. (2007). A Gamma model for DNA Mixture Analyses. *Bayesian Analysis* **2**, 333–348.

Cowell, R. G., Lauritzen, S.L. and Mortera, J. (2008). Probabilistic modelling for DNA mixture analysis. *Forensic Science International Genetics: Supplement Series*, **1**, 640-642.

Dawid, A. P., Mortera, J., Pascali, V. L., and van Boxel, D. W. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, **29**, 577–95.

Mortera, J., Dawid, A. P., and Lauritzen, S. L. (2003). Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, **63**, 191–205.