Reconstructing Pedigrees with Maximal Likelihood from Genetic Marker Data

Nuala A Sheehan

Departments of Health Sciences and Genetics, University of Leicester

Joint work with James Cussens & Mark Bartlett (University of York)

Medical Research Council Project Grant G1002312

Bristol May 2015

Finding Relatives

Identification of relatives relevant for applications in

- evolution and conservation research
- selective animal breeding
- (genetic) epidemiological research
- genealogical research
- genetic counselling
- forensic identification.

Can view finding relatives as a problem of pedigree reconstruction

Some Basic Genetics

Human DNA packaged into 23 pairs of chromosomes in each normal cell — 22 pairs of autosomes and 1 pair of sex chromosomes.
In any pair, one has DNA of mother and the other of father.

Gene: section of DNA / chromosome coding for a protein and its *position* is a genetic locus

Allele (or DNA variant): a particular sequence at a genetic locus.

Genetic marker: known section of chromosome with known variations of the DNA sequence (alleles) in the population.

Genotype: unordered pair of alleles at a locus (one on each chromosome).

Phenotype — potentially observable characteristic e.g. affected/normal, blood type. Could be the unordered genotype.

Mendel's First Law (1866)

- Each individual has two discrete *factors* (or genes) controlling any given characteristic.
- One is a copy of one of the corresponding pair in his mother and the other is a copy of one of the paternal pair.
- A copy of a randomly selected gene from the two parental genes is passed to each child from a parent, independently for different children and independently of the gene contributed by the other parent.

When genes segregate with probability $\frac{1}{2}$, we have Mendelian segregation — a reasonable assumption for many autosomal traits.

Mendel's Second Law: segregations of genes at different loci are independent. Not true if loci are close to each other or linked.

Hardy-Weinberg Equilibrium

Suppose we have a genetic locus with k alleles $\{A_1, \ldots, A_k\}$ and corresponding population frequencies $\{p_1, \ldots, p_k\}$ with $\sum_{i=1}^{k} p_i = 1$.

Assuming 'random union of gametes' in the founder generation yields founder genotype frequencies of:

$$\begin{array}{llll} A_iA_i & p_i^2 & & \\ A_iA_j & 2p_ip_j & \forall i,j & i \neq j. \end{array}$$

These frequencies can be preserved over generations of random mating in the absence of competing evolutionary forces.

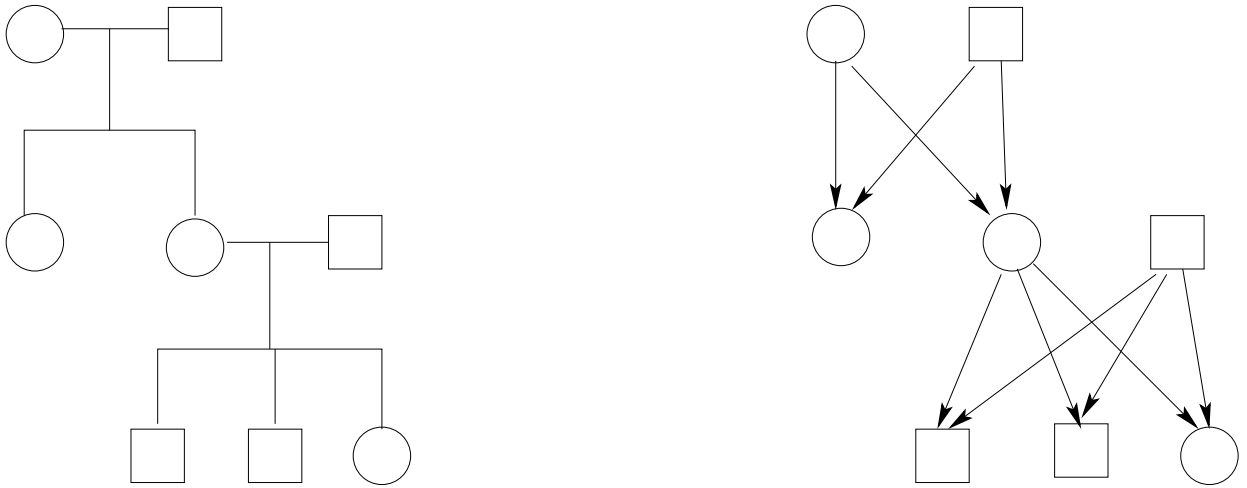
We say that the population is in Hardy-Weinberg equilibrium.

Pedigrees

We formally define a pedigree (or genealogy) to be:

- a set of individuals with a full specification of all the relationships amongst them
- we define a pair of pedigree members to be spouses *only* if they have a common offspring and every spouse pairing is called a marriage
- founders are those with *neither* parent present and either belong to an ancestral baseline generation *or* have married into the pedigree more recently
- the founders are hence unrelated, by definition of the pedigree.

Pedigrees as DAGs



(a) **(b)**

A pedigree is a directed acyclic graph (DAG) with well-defined structural constraints:

– each node has *exactly* two parent nodes (but either can be latent)
– and parents *must* be of opposite sexes.
– a node with no parents is a pedigree founder.

Pedigree Graphs are DAGs

Graphs have directed edges because they are basically depicting patterns of ancestry and descent and so there is a natural ordering whereby parents precede offspring.
Typically omit the arrows as direction is obvious.

Pedigree directed graphs are thus acyclic: an individual cannot be his own ancestor or descendant.

However, computations are typically performed on the *undirected* graph and pedigrees can have plenty of undirected cycles $\Rightarrow$ loops

Consider genetic data on a pedigree with nodes $v \in V$. We can view these nodes as random variables e.g. assigning a genotype to the relevant individual. (There are other representations.)

Likelihood-based Pedigree Reconstruction

In theory: simple to estimate the pedigree connecting a given set of individuals from genetic marker data — consider all possible pedigrees and compute the likelihoods (Thompson 1975)

In practice: enormous number of possibilities $\longrightarrow$ naïve brute force enumeration only feasible for small numbers. (Sheehan & Egeland 2007)

'Sequential' (i.e. *greedy*) algorithms exist which efficiently produce a single high likelihood reconstruction but not necessarily a maximum likelihood (ML) pedigree. (Thompson 1976, Almudevar 2003)

We propose:
Search for maximum likelihood pedigrees using integer linear programming (ILP) and state-of-the-art ILP optimisation solvers.

Likelihood-based Pedigree Reconstruction

As for other current approaches, we will assume:

- a complete sample i.e. all pedigree members are observed with complete marker data. Individuals with unobserved parents are the pedigree founders and their (unobserved) parents are unrelated to other individuals in the sample.

- genotype marker data at unlinked loci (independent markers)

- Hardy-Weinberg proportions for founder genotypes

- Mendelian segregation of genes from parents to offspring.

Pedigree Likelihood

Let $\mathrm{Pa}(v, \mathcal{G})$ denote the *parent set* of v in $\mathcal{G}$: has 0, 1 or 2 elements.

Under the above assumptions:

$$\text{Likelihood: } L(\mathcal{G}) = \prod_{v \in V} \tau(v, \mathrm{Pa}(v, \mathcal{G})).$$

– decomposes into a product of local conditional probabilities, each a probability of an individual's genotype given the parental types

e.g. $\tau(v, \mathrm{Pa}(v, \mathcal{G}))$ = Mendelian probability $\frac{1}{4}, \frac{1}{2}$ or 1 for an individual with two observed parents

$\tau(v, \emptyset) = p(g_v)$ — marginal probability of genotype g_v for a founder.

NOTE: The above decomposition is the definition of a BN.

Optimisation Problem

Aim of a maximum likelihood reconstruction:
– find $\mathcal{G}$ such that $L(\mathcal{G})$ is maximized
$\equiv$ search for an optimal BN with search constrained to valid pedigrees.

Convenient to work with the log-likelihood instead, so we optimize:

$$\text{Log-likelihood: } l(\mathcal{G}) \;=\; \log L(\mathcal{G}) = \sum_{v \in V} \log \tau(v, \mathsf{Pa}(v, \mathcal{G})).$$

NOTE: Because we are working with *unlinked* loci, we only need to consider one locus at a time and take the product (sum) of the likelihoods (log-likelihoods) over all loci.

Integer Linear Programming (ILP)

An integer linear program (ILP) is defined by:

1. a set of variables X, representing unknown quantities, some of which are restricted to be integer values;

2. an objective function of the form $\sum_{x \in X} c_x x$ where the coefficients c_x are fixed constants, and

3. linear equations and inequalities putting joint constraints on the values the variables can take.

ILP optimisation problem: find an assignment of values to X which maximizes the objective function while respecting all constraints.

Integer Linear Programming (ILP)

NP-hard in the general case — no known algorithm to solve in polynomial time.

Still possible to solve exactly e.g. brute force enumeration, albeit ridiculously slow.

There are highly optimized ILP solvers based on (a version of) the simplex algorithm.

Will eventually run out of computer memory or be impractically slow but typically performs well given sufficient time and resources.

Question: Can we encode pedigree reconstruction as an ILP problem?

Encoding Pedigree Learning as an ILP Problem

This can be done for the above assumptions.

Need to define variables (X) that allow for a numeric encoding of any possible pedigree.

Since a pedigree is a full specification of parent-offspring trios, binary variables representing parentage do precisely this:

$$I(W \rightarrow v)(\mathcal{G}) = \begin{cases} 1 & \text{if } v \text{ has parents } W \text{ in } \mathcal{G} \\ 0 & \text{otherwise,} \end{cases}$$

where $W \subseteq V \setminus \{v\}$ and $|W| \leq 2$.

i.e. indicator variables representing all possible parentages of an individual.

Encoding Pedigree Learning as an ILP Problem

Since $I(W \rightarrow v)(\mathcal{G}) = 1$ only when $W = \mathsf{Pa}(v, \mathcal{G})$ the pedigree log-likelihood can be written in the desired form:

$$\begin{aligned} \log L(\mathcal{G}) &= \sum_{v \in V} \log \tau(v, \mathsf{Pa}(v, \mathcal{G})) \\ &\equiv \sum_{v,W} \log \tau(v, W) I(W \rightarrow v)(\mathcal{G}) \end{aligned} \tag{1}$$

Optimisation problem: Find an instantiation of the $I(W \rightarrow v)(\mathcal{G})$ to maximize the above while constraining the search to valid pedigrees

$\longrightarrow$ a maximum likelihood pedigree.

Encoding Pedigree Learning as an ILP Problem

NOTE 1: If it solves, a **guaranteed** maximum likelihood pedigree is returned.

NOTE 2: Can repeatedly solve the above adding an additional constraint each time to prevent a previously returned solution from appearing again
$\longrightarrow$ find k most likely pedigrees.

Dynamic programming methods have also been used to find a single optimal pedigree but are limited to small structures e.g. < 30 individuals (Cowell, 2009).

Constraining the Search to Valid Pedigrees

An ILP formulation is possible if constraints on the $I(W \rightarrow v)(\mathcal{G})$ to rule out invalid assignments can be expressed as linear equations and inequalities.

Since each pedigree member has only one set of parents, any possible pedigree $\mathcal{G}$ determines a joint instantiation of these binary variables

$$I(W \rightarrow v)(\mathcal{G})$$

by setting exactly $|V| = n$ of them to $\mathbf{1}$ and the rest to $\mathbf{0}$.

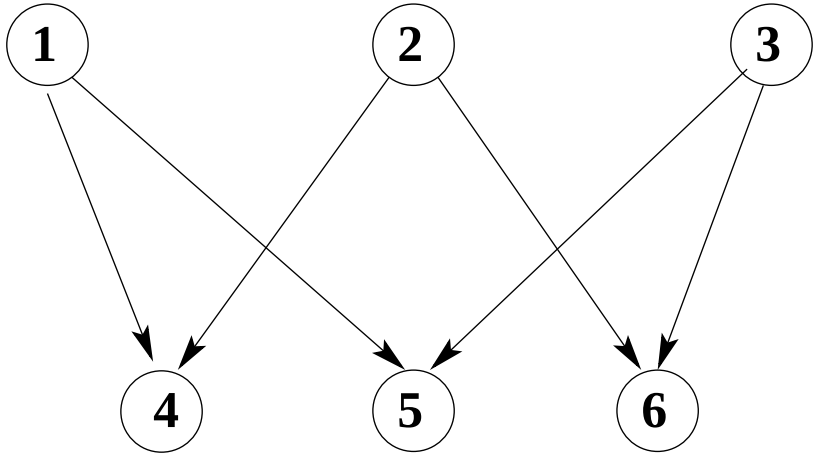
The mapping is not one-to-one as most realisations do not correspond to a valid pedigree.

Constraining the Search to Valid Pedigrees

- Each individual has exactly one parent set:

$$\forall v : \sum_{W} I(W \rightarrow v) = 1, \qquad |W| \leq 2$$

- Sex can be consistently assigned i.e. members of a parent pair are of opposite sex.



(Adapted from Cowell 2009)

Constraining the Search to Valid Pedigrees

Ruling out directed cycles:

Generation numbers: $n(n-1)$ auxiliary variables assigning a value of 0 to each founder and a value greater than that of each parent to a non-founder

— no directed cycles since nobody can be his own ancestor / descendant

(Cussens et al. Genetic Epidemiology 2013).

Cluster constraints: based on observation that any subset $C \in V$ must contain at least one individual with no parent in C (Jaakkola et al. 2010)

$\longrightarrow$ tighter upper bound from linear relaxation & faster solving.

Exponentially many so add as *cutting planes* while solving: 'snip' off interim solutions with cycles.

There are efficient ways to search for good cutting planes

(Cussens 2011, Bartlett & Cussens 2013).

Addressing Model Uncertainty

Probabilistic nature of genetic inheritance $\Rightarrow$ a single reconstruction is not necessarily very useful. Need some measure of variability.

The ILP framework easily permits searching for many high likelihood pedigrees.

Recall: Can repeatedly solve the ILP problem

$\longrightarrow$ find the k most likely pedigrees
— consider relative likelihoods (more peaked distribution means top one more likely to be similar to true pedigree)

$\longrightarrow$ can think about model averaging over 'good' pedigrees
— but takes no account of long tail in distribution.

Maximum Likelihood Versus True Pedigree

There is no guarantee that the most probable i.e. maximum likelihood pedigree is the true pedigree.

Marker data on unrelated individuals may indicate that they are actually related and the observed data could assign a higher probability to an incorrect relationship than the true one (e.g. siblings could look like parent-child).

Whether the maximum likelihood pedigree resembles the true one, or not, has no bearing on the efficiency of a method for finding it.

We would perhaps expect the true pedigree to be among the top most likely pedigrees.

Simulations

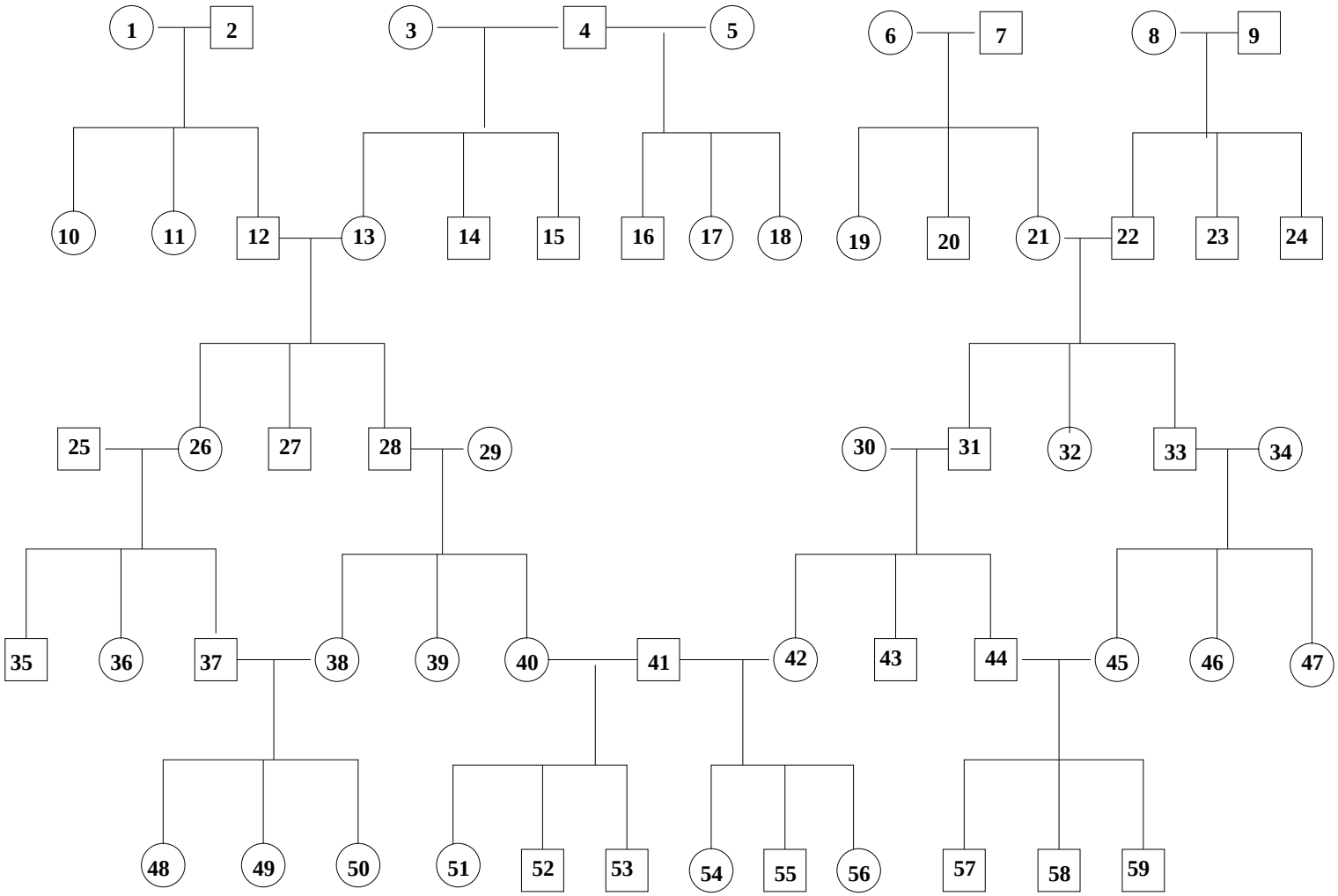
For different pedigree structures

- Simulated 100 sets of complete data on the pedigree from allele frequencies for a typical forensic set of autosomal short tandem repeat (microsatellite) markers. (Highly variable and assumed to be independent).
  - Caucasian allele frequencies for 13 CODIS loci plus two others (Butler 2003)
  - Number of alleles at each marker ranges from 7 to 15.

- Assigned genetic profile to founders at each marker independently assuming Hardy-Weinberg equilibrium and dropped genes from parents to offspring assuming Mendelian transmission.

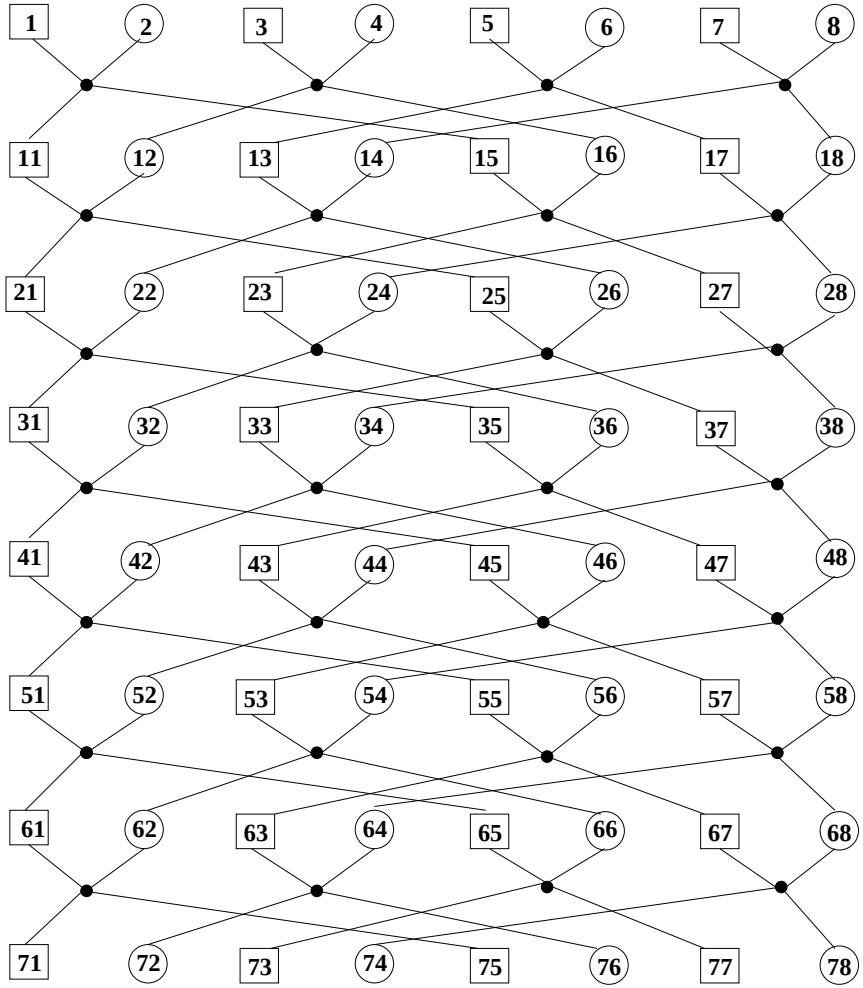
A 'Realistic' 59-member Pedigree

(Almudevar, Theoretical Population Biology, 2003)



Quadruple Second Cousin Regular Mating Structure

(Wright,Genetics, 1921)



No matings between close relatives but everyone related to all 8 founders after 4 generations.

Results: How Good is the Reconstruction?

Evaluation depends on what you want it for!

Demographic/ecological research: focus on general features?
e.g. correct distribution of sibship sizes, numbers of marriages and overall levels of relatedness.

Forensic research: want particular relationships to be correct.

BN learning error rate: number of incorrect parent-offspring links
– not necessarily appropriate measure of success for maximum likelihood approach where goal is to find high likelihood pedigrees quickly

– and may not be best criterion for any given situation.

Summary of Results for 'Small' Pedigrees

(Cusssens et al. 2013 *Genetic Epidemiology*)

For pedigrees of up to about 100 individuals:

- Outperforms competing methods re size (< 30) and solving times.
- Maximum likelihood pedigree found very quickly — seconds.
- Only method to give 'top k' pedigrees in decreasing likelihood order.

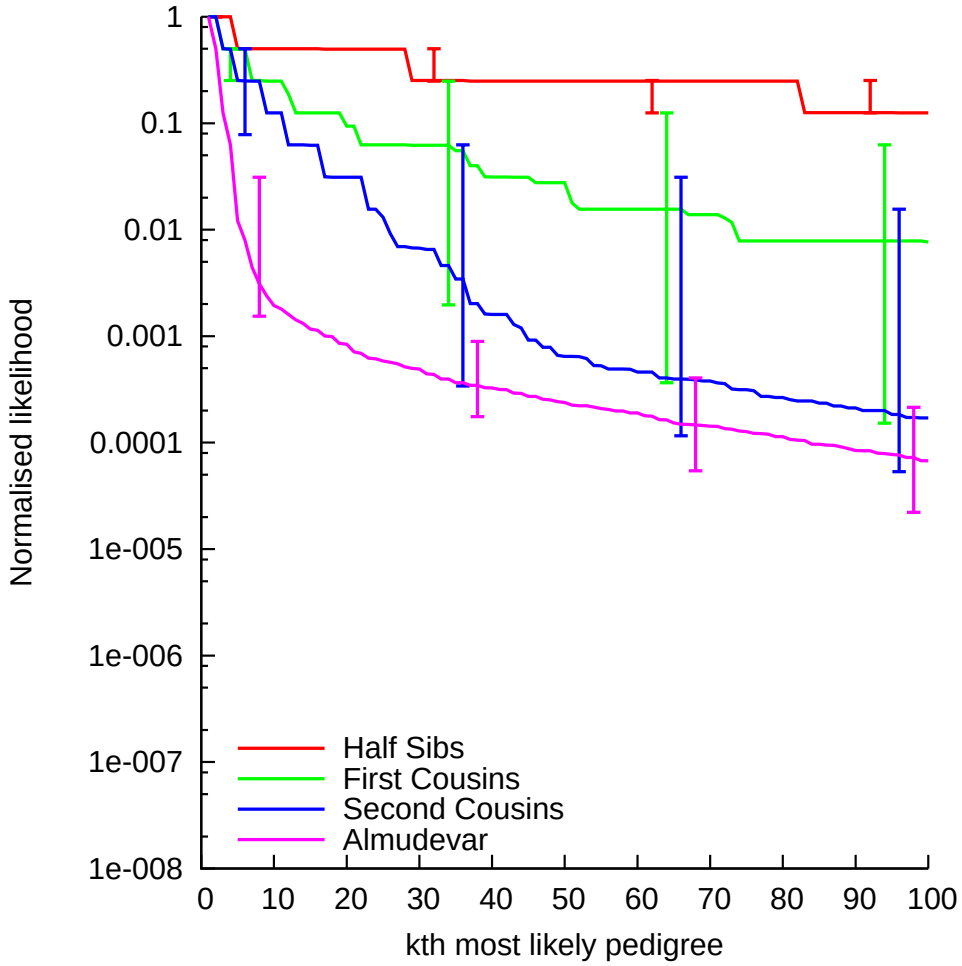
- True pedigree often had maximal likelihood. Generally featured in top 100 unless structure is complex.
- Accuracy deteriorates with increasing inter-relatedness.
  More genetically consistent choices for parent pairs and triplets?
- Need high exclusion probabilities for incorrect parent pairs to get an accurate reconstruction (Thompson 1986).

Scaled Likelihoods for Top 100 Pedigrees

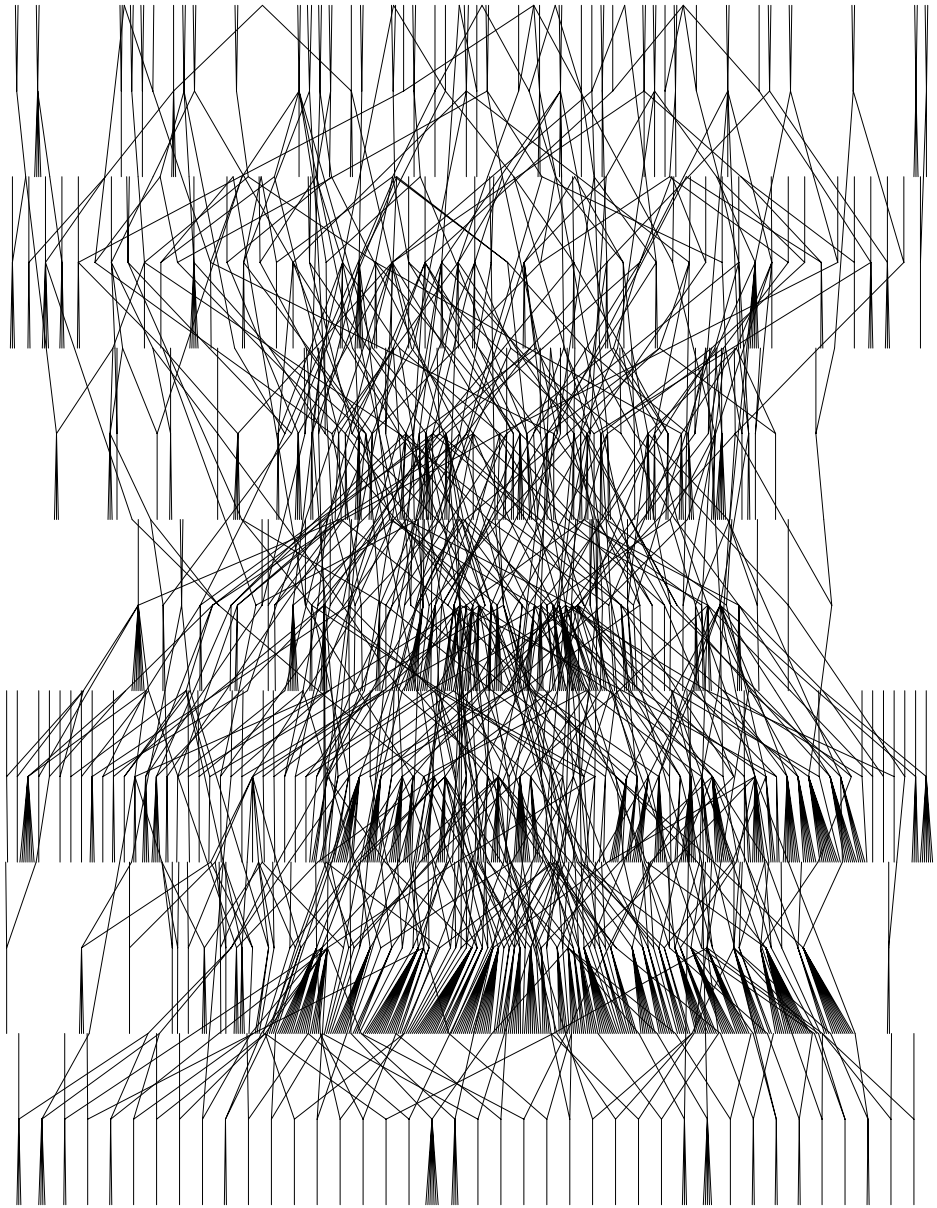
Median values for 100 datasets

Realistic frequencies

4 pedigrees of differing complexity



Reconstructing Big Pedigrees



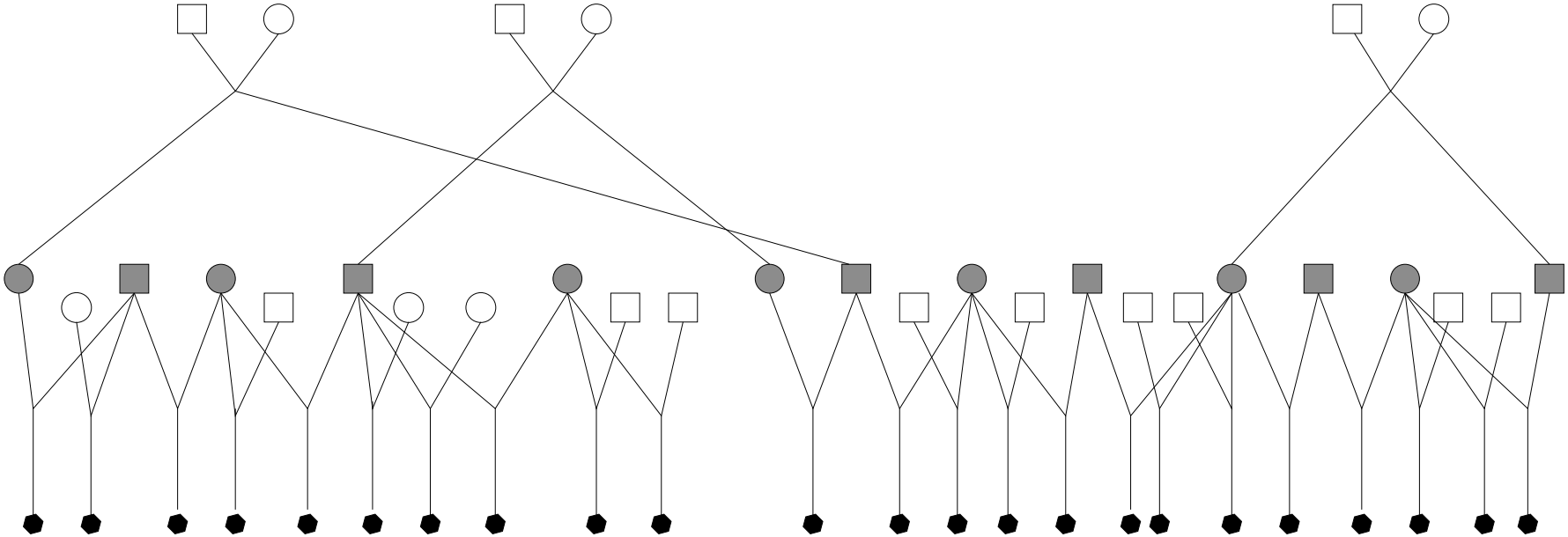
1614 Polar Eskimos

The 'Polar Eskimo Genealogy'

(Gilberg et al. 1978, Edwards 1992).

- Thule, North West Greenland — 'northernmost native human population on Earth'
- Population prior to 1818 can be regarded as a breeding isolate. 1614 descendants from this breeding isolate ranging over 7 generations from a birth in about 1805 to one in 1975.
- Remarkably low levels of inbreeding
  - highest inbreeding coefficient is $\frac{1}{8}$ — single uncle-niece marriage
  - only 5 first cousin marriages — one via a half-first cousin pairing.
- Lots of inter-relatedness by marriage $\longrightarrow$ multiple inter-connecting undirected cycles
- Likelihood calculations intractable even for a single locus when there is missing data. (Sheehan 1992).

Marriage Loops in the Polar Eskimo Pedigree



Evaluating an Eskimo Reconstruction

Unreasonable to expect all 2778 parent-child graph edges to be correct!

Local Features:

- family (sibship) sizes
- number of marriages for a given individual
- particular relationships between certain individuals

Global Features:

- number of pedigree components
- number of founders (pedigree *width*)
- number of generations (pedigree *depth*)
- number of pedigree descendants for certain individuals

Evaluating the Reconstruction

Solving times:

Cluster constraints versus **generation numbers** to rule out cycles
$\longrightarrow$ 3-100 *hours* versus 3-42 minutes (average 10) on 100 datasets.

True pedigree has:

- 225 founders
- 11 distinct components with 1581 in the largest (then 11 and ≤ 6)
- 563 individuals with 1 marriage,129 with 2, 21 with 3 & 8 with 4
- 3 sibling pairs and 23 marriages in the marriage chain structure
- One individual with 10 offspring from two marriages
- One individual with 4 marriages and 452 descendants in the pedigree

- One uncle-niece marriage

How Good is the Reconstruction?

BN learning error rate:

There are 2778 edges in the true pedigree.

Recall: proportion of all real edges found
—high recall $\Rightarrow$ most of the correct edges have been found in the reconstruction (could have many incorrect ones also).

Precision: proportion of constructed edges that are real
— high precision $\Rightarrow$ more correct than incorrect edges in the reconstruction (may not have them all).

Recall and Precision are analogous to *Sensitivity* and *Positive Predictive Value* in diagnostic testing.

How 'Good' is the Reconstruction?

- Seems 'good' at getting local features right e.g. marriage chain, number of children of a single individual and uncle-niece marriage.
- Not so good at global features — will try to create as many relatives as possible to optimize the likelihood
  $\longrightarrow$ too few founders (but usually correct), too few components and large one too big, too many generations, too many descendants.
- Distribution of marriages (0, 1, 2, 3, or 4) reasonably good but overestimates number of multiple marriages.
- Edges?: proportion of all real edges (directed) found and proportion of reconstructed edges that are true — both $\sim 95\%$
  – not sure what this is telling us though!

Using Prior Information

The genealogy file is in standard triplet format (individual, father, mother) where everyone has 0 or 2 parents. (Founder labels were created for missing parents.)

'Telling' the algorithm that everyone has 0 to 2 parents leads to a vast improvement in solving time: average ~ 38 *seconds*, range 6.5-242 seconds to find a single maximum likelihood reconstruction over the 100 simulated datasets.

Average $\#$ parent sets to be considered reduced from 7.3 to 2.1.

It is now easy to knock out the top k pedigrees: top 100 pedigrees *always* had the same likelihood $\longrightarrow$

Maximum likelihood pedigree not necessarily unique: number of solutions depends on size and complexity.

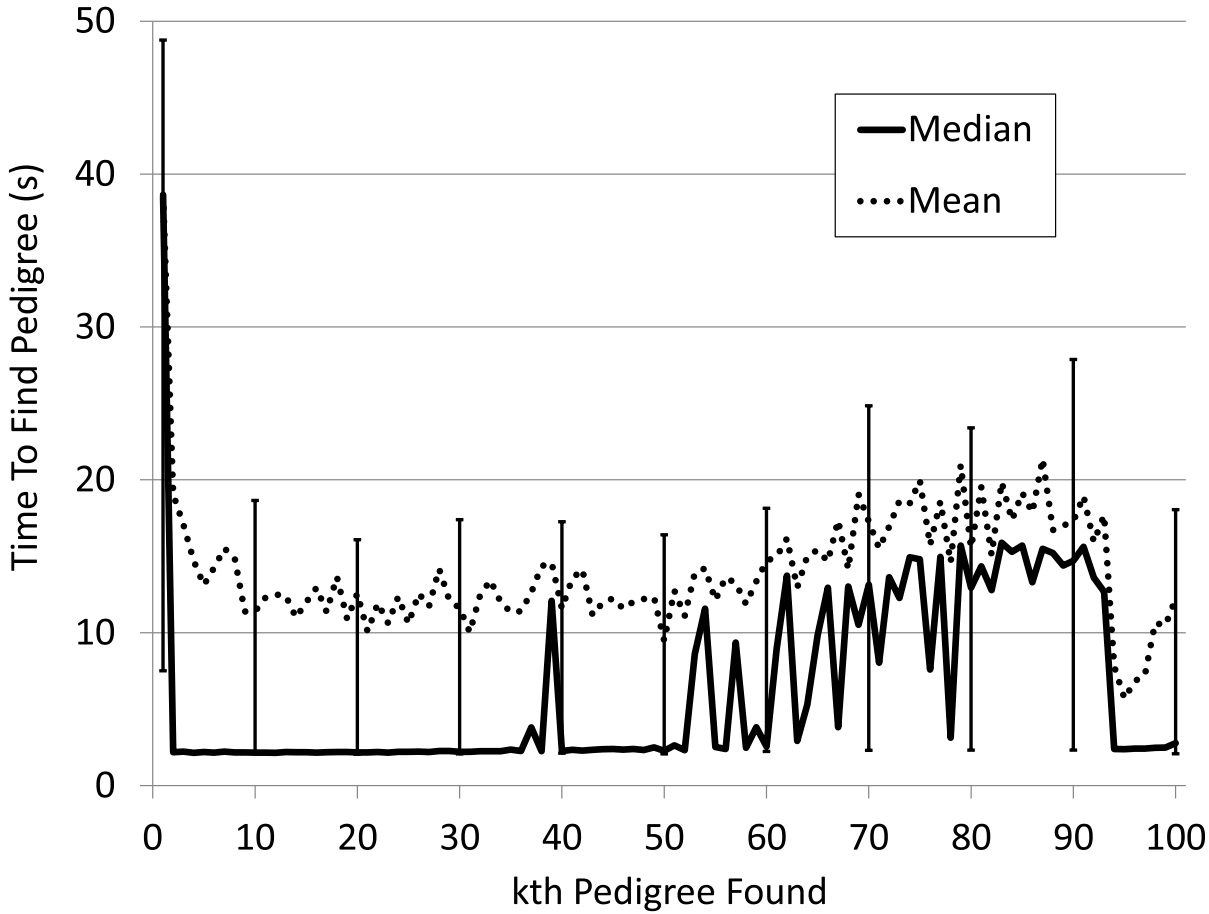
Results for 0 or 2 Parents

- 99% identified correct number of founders: 92% were true founders.
- Almost always got the number of components right and the right number in each.
- No better on marriage distribution: reasonable but never right.
- Much better on getting the number of pedigree descendants right and the number of generations.

Global features seem to be sorted. Local features still good.

Likelihoods tend to be higher for $0, 1, 2$ parent case.
Reconstruction much closer to true pedigree in $0, 2$ parent case.

Solving Times for 0 or 2 parents



Additional Marker Data

Without prior information on size of parent sets.

- Doubled number of markers to 30 (additional 15 has same frequencies as before), 100 datasets, a single reconstruction
- Average number of parent sets reduced from 7.3 to 5.4
- Solving times reduced: < 1 minute always
- Similarity to true pedigree significantly improved: 77% exact and remainder had 1 or 2 incorrect parentage assignments

$\longrightarrow$ 30 markers sufficient to rule out most possible parent sets except those corresponding to (possible reversed) true edges — task reduced to orienting edges.

Comparison with an Approximate Approach

FRANz (Riester et al. 2009) uses simulated annealing (Almudevar 2003) for the complete data case. Very quick.

Compared on same 100 simulated dataets for 15 markers and no additional (prior) information.

- Solving times: median lower for ILP, average lower for FRANz.
- FRANz never found any of the maximum likelihood pedigrees.
  3 reconstructions were *less* likely than true pedigree.
- FRANz reconstructions also *less* similar to true pedigree in terms of recovering local and global features
  $\longrightarrow$ finding a pedigree with maximal likelihood not just theoretically interesting in this case.

We've cracked big — now we need to get useful.

Complications & Future Work

Core of ILP formulation is the simple decomposition of the pedigree likelihood — directed local Markov property for BNs. Crucially depends on Mendelian segregation $\longrightarrow$ different representation?

Also breaks down for
– incomplete data — sum over all possibilities for missing people
– linked markers e.g. dense SNPs
— siblings no longer independent given parental genotypes.

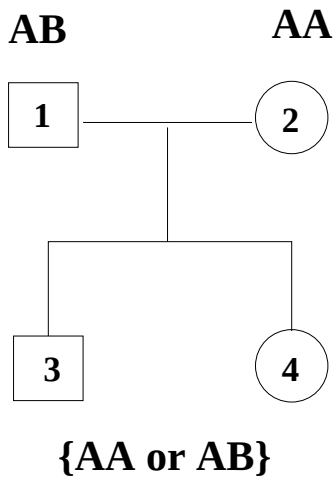
Need to add extra variables for missing data and unknown phase
— will return most probable combination of pedigree structure *and* haplotypes in latter case. (ILP has been used for haplotype inference.)

More intensive likelihood computations $\longrightarrow$ approximations.

Will also need to account for genotyping error and/or mutations.

Suppose Inheritance is not Mendelian

For the following nuclear family with typed parents, 1 and 2, note that their offspring can only have two possible genotypes, AA or AB.



Consider the case where an individual transmits a copy of his *paternal* gene to his offspring with probability τ_1 and his *maternal* gene with probability $1 - \tau_1$.

Consider the conditional probability of a second child being AA given the genotype of the first child.

Suppose Inheritance is not Mendelian

We can ignore inheritance from the mother. If the first child is AA,

$$P(G_4 = \{A, A\}|g_1, g_2, g_3 = AA) = \tau_1^2 + (1 - \tau_1)^2$$

— both inherit a copy of the same paternal *or* maternal gene from their father.

If the first child is AB

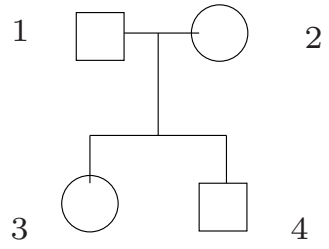
$$P(G_4 = \{A, A\}|g_1, g_2, g_3 = AB) = 2\tau_1(1 - \tau_1)$$

— they inherit copies of different paternal genes.

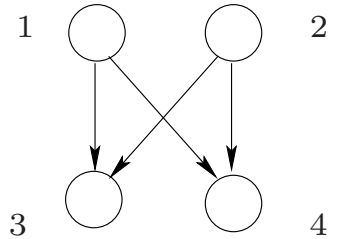
Note that $\tau_1^2 + (1 - \tau_1)^2 = 2\tau_1(1 - \tau_1)$ only if $\tau_1 = \frac{1}{2}$ i.e. genotypes of siblings are not conditionally independent given the genotypes of their parents *unless* inheritance in Mendelian.

Other BN Representations

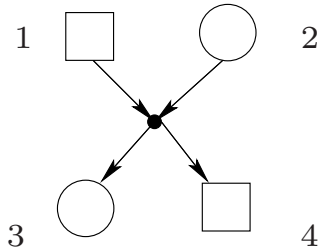
Different properties & computational challenges.Lauritzen & Sheehan (2003)



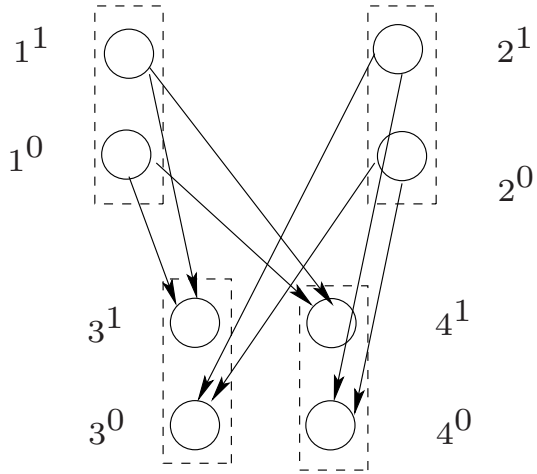
Standard



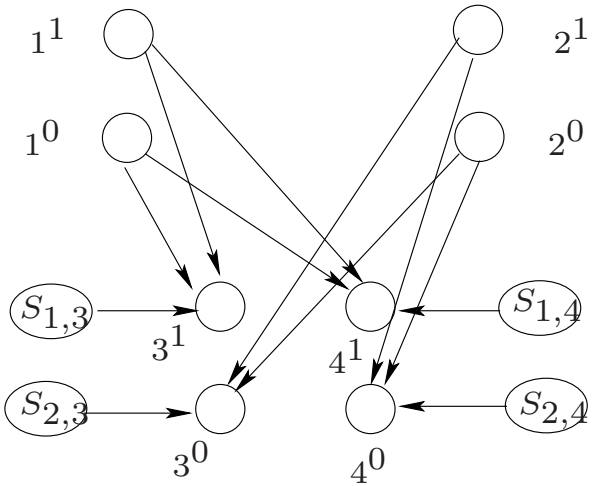
Genotype Network



Marriage Node Graph



Allele Network



Segregation Network

Complications & Future Work

In reality, prior information typically available and should be used
– demographic information on mating patterns
– numbers of marriages and family sizes
– breeding ages and generation gap
– knowledge of specific relationships
– sex and ages for some individuals.

Prior information can be incorporated
– as additional constraints ('hard' prior information)
– as additional log prior terms in the objective function

$\longrightarrow$ helps eliminate 'unreasonable' structures e.g. a pedigree that fits the data best might represent unrealistically high levels of inbreeding etc.

Conclusions

- Can now begin to investigate the properties of maximum likelihood pedigree reconstructions
- A maximum likelihood reconstruction is not necessarily unique
- Performance of approximate approaches that deliver a high likelihood, but not necessarily optimal solution, can now be evaluated
- Only ML approach that can get the first n most likely pedigrees
- ML appears to be performing well with only marker data but will need prior information to guide search for accurate reconstruction on big problems
- 15 microsatellites reasonable for *complete* data (30 better!): will need many more markers for missing / latent data
- ILP approach compares well with other approaches and seems to be more amenable to generalization to more complex problems.

Some References

Sheehan, N A et al. (2014) *Theoretical Population Biology* **97**:11–19.

Cussens, J. et al (2013) *Genetic Epidemiology* **37**: 69–83.

Jaakkola, T. et al (2010) *Journal of Machine Learning Research Workshops and Conference Proceedings* **9**: 258–365.

Riester. M et al. (2009) *Bioinformatics* **25:** 2134–2139.

Sheehan, N A & Egeland, T (2007) *Annals of Human Genetics* **71**: 501–518.

Almudevar, A. (2003) *Theoretical Population Biology* **63**: 63–75.

Thompson, E A 1976) *Social Sciences Information* **15**: 477–526.

Software: GOBNILP — freely available from
http://www.cs.york.ac.uk/aig/projects/prilp/