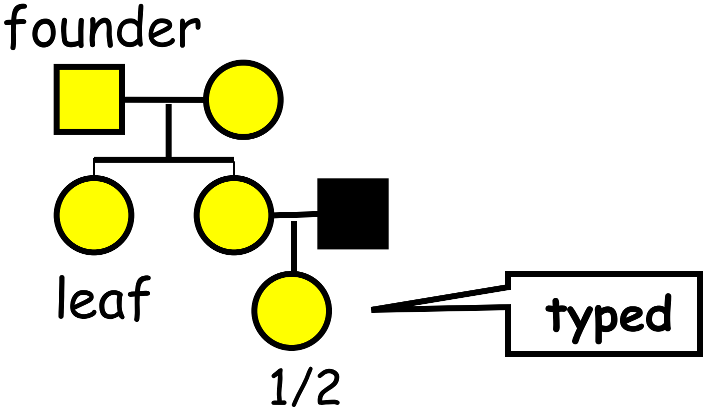
The Elston-Stewart Algorithm

Tutorial

by Ma'ayan Fishelson

The Given Problem

- **Input**: A pedigree + phenotype information about some of the people. These people are called ***typed***.



- **Output**: the probability of the observed data, given some probability model for the transmission of alleles.

Q: What is the probability of the observed data composed of ?

A: There are three types of probability functions: founder probabilities, penetrance probabilities, and transmission probabilities.

Founder Probabilities - One Locus

- **Founders** - individuals whose parents are not in the pedigree. We need to assign probabilities to their genotypes. This is done by assuming **Hardy-Weinberg equilibrium**.

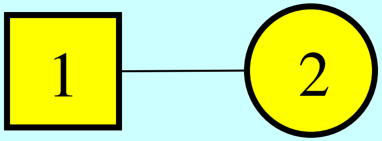
d/h

d-mutant allele
h-normal allele

Suppose the gene frequency of d is 0.05, then:

$$P(d/h) = 2 * 0.05 * 0.95$$

- Genotypes of different founders are treated as **independent**:



d/h h/h

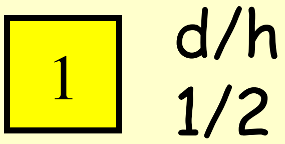
$$Pr(d/h, h/h) = Pr(d/h) * Pr(h/h) = (2 * 0.05 * 0.95)*(0.95)^2$$

Founder Probabilities – Multiple Loci

- According to **linkage equilibrium**, the probability of the multi-locus genotype of founder k is:

$$Pr(x_k) = Pr(x_k^1) * \dots * Pr(x_k^n)$$

Example:



$$Pr(d/h,\ 1/2) \underset{\text{Linkage equilibrium}}{=} Pr(d/h) * Pr(1/2) \underset{\text{Hardy-Weinberg equilibrium}}{=} 4 * Pr(d)*Pr(h) * Pr(1)*Pr(2)$$

Penetrance Probabilities

- **Penetrance**: the probability of the phenotype, given the genotype.

- E.g.,**dominant** disease, **complete penetrance**:

■ d/d Pr(affected|d/d) = 1.0	■ d/h Pr(affected | d/h) = 1.0	□ d/h Pr(affected | h/h) = 0

- E.g., **recessive** disease, **incomplete penetrance**:

■ d/d

Pr(affected | d/d) = 0.7

Can be, for example, sex-dependent, age-dependent, environment-dependent.

Transmission Probabilities

- **Transmission probability**: the probability of a child having a certain genotype given the parents' genotypes.

$$Pr(x_c | x_m, x_f).$$

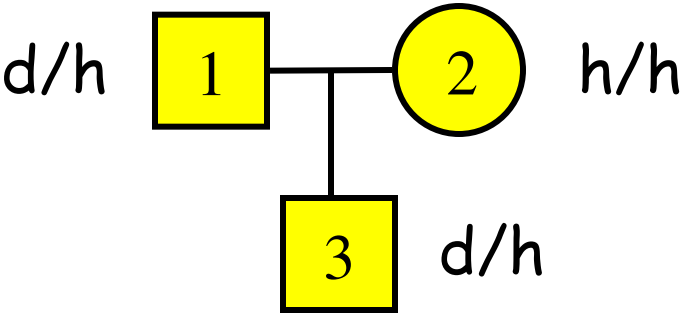
- If we split the ordered genotype x_c into the maternal allele x_{cm} and the paternal allele x_{cf}, we get:

$$Pr(x_c | x_m, x_f) = Pr(x_{cm}|x_m)Pr(x_{cf}|x_f)$$

The inheritance from each parent is independent.

Transmission Probabilities – One locus

- The transmission is according to the 1st **law of Mendel**.



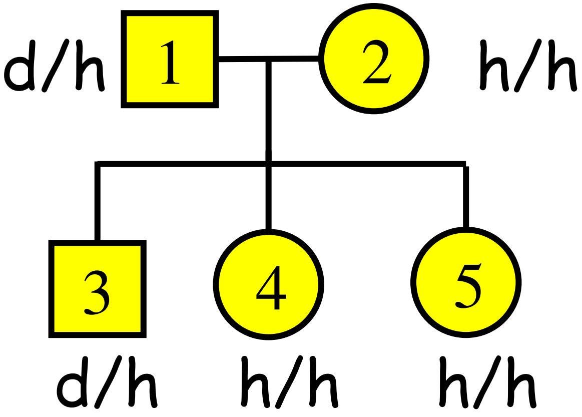
$$\Pr(X_c=d/h \mid X_m=h/h, X_f=d/h) =$$

$$\Pr(X_{cm}=h \mid X_m=h/h) * \Pr(X_{cf}=d \mid X_f=d/h) = 1 * \tfrac{1}{2} = \tfrac{1}{2}$$

We also need to add the inheritance probability of the other phase, but we can see that it's zero !

Transmission Probabilities – One locus

- Different children are independent given the genotypes of their parents.



$\Pr(X_3=d/h, X_4=h/h, x_5=d/h \mid X_1=d/h, X_2=h/h) =$

$= (1 * \frac{1}{2}) * (1 * \frac{1}{2}) * (1 * \frac{1}{2})$

Transmission Probabilities – Multiple Loci

- Let's look at paternal inheritance for example.
- We generate all possible recombination sequences (s_1,s_2,…,s_n), where s_l = 1 or s_l = -1. (2^n sequences for n loci).
- Each sequence determines a selection of paternal alleles p_1,p_2,…,p_n where:

$$p_l = \begin{cases} x_{fM} & \text{if } s_1 \times \cdots \times s_l = 1 \\ x_{fF} & \text{if } s_1 \times \cdots \times s_l = -1, \end{cases}$$

and therefore its probability of inheritance is:

$$\frac{1}{2}[p_1 == x_{kf}^{(1)}]\prod_{l=2}^{n}[p_l == x_{kf}^{(l)}] \times \begin{cases} \theta_l & \text{if } s_l = -1 \\ 1-\theta_l & \text{if } s_l = 1, \end{cases}$$

We need to sum the probabilities of all 2^n recombination sequences.

Calculating the Likelihood of Family Data - Summary

The **likelihood of the data** is the probability of the observed data (the known phenotypes), given certain values for the unknown recombination fractions.

- For a pedigree with m people:

$$L = P(x) = \sum_{g} P(x, g) = \sum_{g} P(x \mid g) P(g),$$

where $x=(x_1,\ldots,x_m)$ and $g=(g_1,\ldots,g_m)$.

Calculating the Likelihood of Family Data - Summary

- G_i : genotype vector for individual i
- Founders: $1..k$
- Non founders: $i \rightarrow m(i), f(i)$

Recombination probabilities

Founder priors by Hardy-Weinberg

$$L(X) = \sum_{G_1}\sum_{G_2}\cdots\sum_{G_m}\left(\left\{ \prod_{founder\, i} \Pr(G_i)\_or\_ \prod_{nonfounder\, i} \Pr(G_i \mid G_{m(i)}, G_{f(i)}) \right\} \prod_{any\, i} \Pr\left(X_i \mid G_i\right) \right)$$

Penetrances

Computational Problem

$$\boxed{L = \sum_{g} P(x \mid g)\, P(g)}$$

Performing a multiple sum over all possible genotype combinations for all members of the pedigree.

<u>Complexity disaster:</u>

- Exponential in #markers
- Exponential in #individuals

Elston-Stewart algorithm

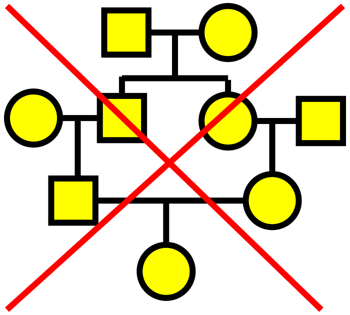
The Elston-Stewart algorithm provides a means for evaluating the multiple sum in a streamlined fashion, for **simple pedigrees**.

More efficient computation

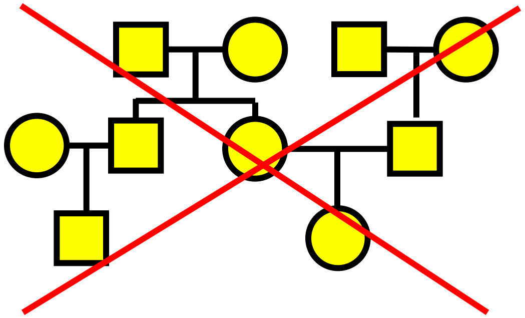
- Exponential in #markers
- Linear in #individuals

Simple Pedigree

- No consanguineous marriages, marriages of blood-related individuals (→ no loops in the pedigree).



- There is **one** pair of founders from which the whole pedigree is generated.



Simple Pedigree

- There is exactly one nuclear family T at the top generation.
- Every other nuclear family has exactly one parent who is a direct descendant of the two parents in family T and one parent who has no ancestors in the pedigree (such a person is called a founder).
- There are no multiple marriages.
- One of the parents in T is treated as the proband.

“Peeling” Order

- Assume that the individuals in the pedigree are ordered such that parents precede their children, then the pedigree likelihood can be represented as:

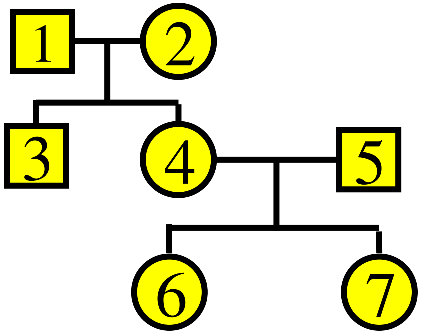
$$L(\theta) = \sum P(x_1 \mid g_1) P(g_1 \mid \cdot) \ldots \left[\sum P(x_m \mid g_m) P(g_m \mid \cdot) \right],$$

where $P(g_i \mid \cdot)$ is:

- $P(g_i)$, if i is a founder, or
- $P(g_i \mid \underbrace{g_{mi}, g_{fi}}_{\text{the genotypes of i's parents}})$, otherwise.

- **In this way, we first sum over all possible genotypes of the children and only then on the possible genotypes for the parents.**

An Example for "Peeling" Order



$h(g_i) = P(x_i|g_i)\ P(gi)$

$h(g_m,g_f,g_c) = P(x_c|g_c)P(g_c|g_m,g_f)$

$$L = \sum_{g_1}\sum_{g_2}\cdots\sum_{g7} h(g_1)h(g_2)h(g_1,g_2,g_3)h(g_1,g_2,g_4)^{*} \\ h(g_5)h(g_4,g_5,g_6)h(g_4,g_5,g_7)$$

According to the Elston-Stewart algorithm:

$$L = \sum_{g_1} h(g_1)\sum_{g_2} h(g_2)\sum_{g_3} h(g_1,g_2,g_3)\sum_{g_4} h(g_1,g_2,g_4)^{*} \\ \sum_{g_5} h(g_5)\sum_{g_6} h(g_4,g_5,g_6)\sum_{g_7} h(g_4,g_5,g_7)$$

Elston-Stewart “Peeling” Order

As can be seen, this “peeling” order, “clips off” branches (sibships) of the pedigree, one after the other, in a ***bottom-up order***.

Elston-Stewart – Computational Complexity

- The computational complexity of the algorithm is linear in the number of people but exponential in the number of loci.

Variation on the Elston-Stewart Algorithm in Fastlink

- **The pedigree traversal order in Fastlink is some modification of the Elston-Stewart algorithm.**
- ***Assume no multiple marriages…***
- **Nuclear family graph**:
  - Vertices: each nuclear family is a vertex.
  - Edges: if some individual is a child in nuclear family x and a parent in nuclear family y, then x and y are connected by and edge x-y which is called a **"down" edge** w.r.t. x and an **"up" edge** w.r.t. y.

Traversal Order

- One individual *A* is chosen to be a “proband”.
- For each genotype *g*, the probability is computed that *A* has genotype *g* conditioned on the known phenotypes for the rest of the pedigree and the assumed recombination fractions.
- The first family that is visited is a family containing the proband, preferably, a family in which he is a child.

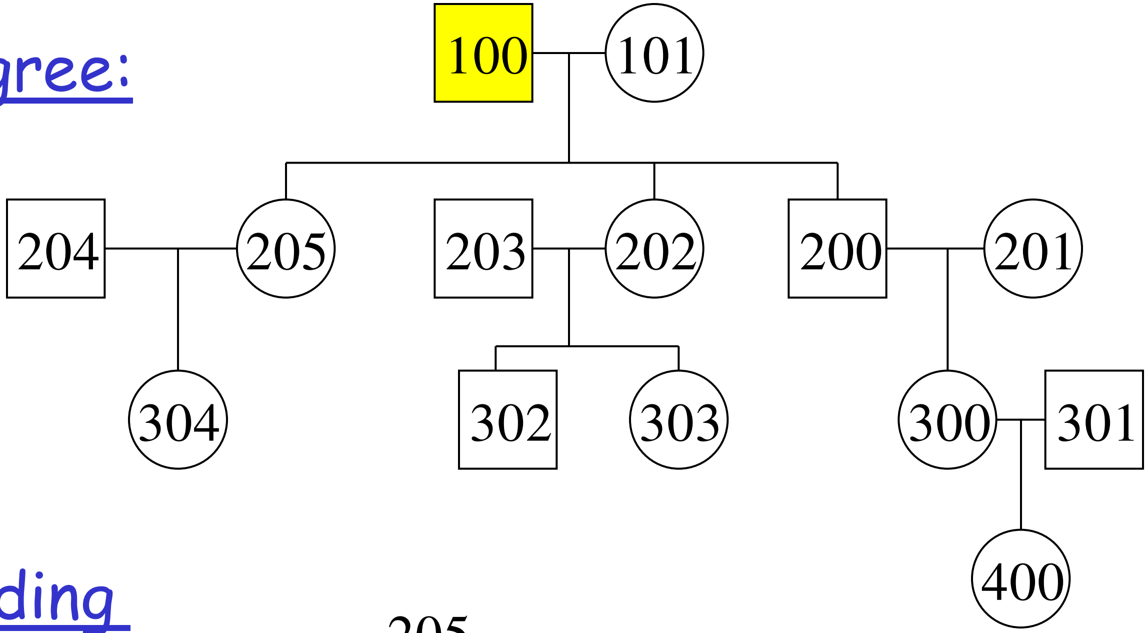
```
Visit(w) {
    While w has an unvisited neighbor x reachable via an up edge:
                    Visit(x);
    While w has an unvisited neighbor y reachable via a down edge:
                    Visit(y);
    Update w;
}
```

Traversal Order - Updates

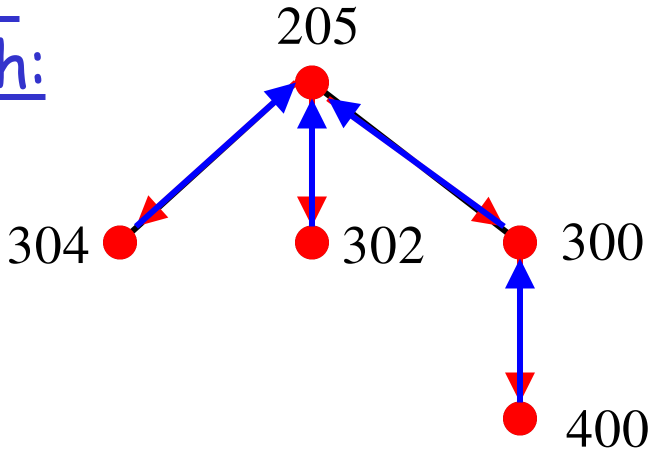
- If nuclear family *w* is reached via a down edge from *z*, the parent in *w* that nuclear families *w* and *z* share, is updated.
- If nuclear family *w* is reached via an up edge from *z*, then the child that *w* and *z* share is updated.

Example 1

An example pedigree:

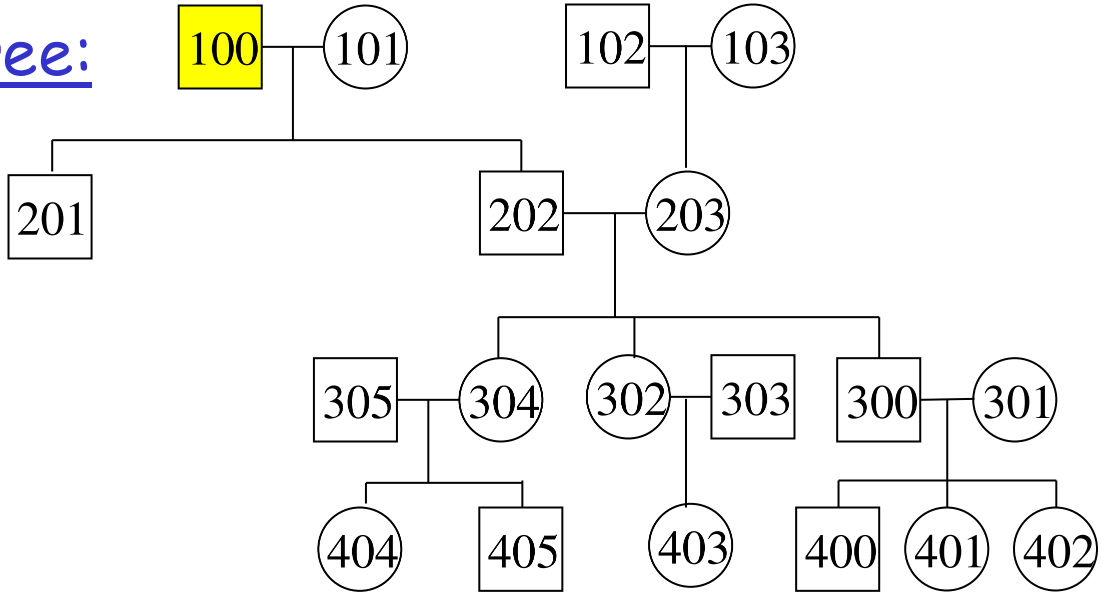


The corresponding nuclear family graph:

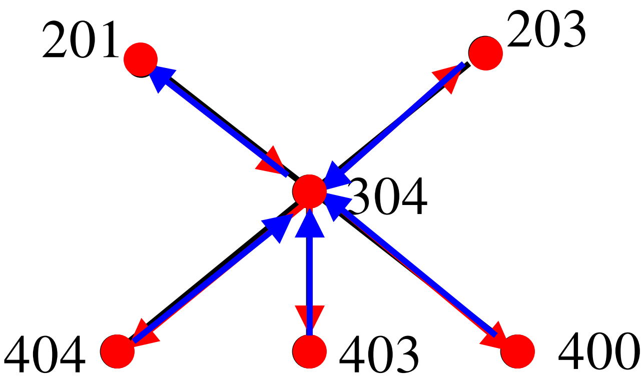


Example 2

An example pedigree:



The corresponding nuclear family graph:


201
203
304
404
403
400