

Graphical Models for Genetic Analyses

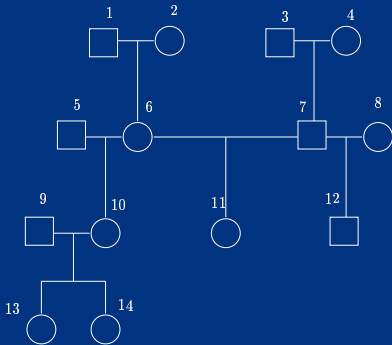
Aarhus University, Fall 2003, Lecture 6

Steffen L. Lauritzen, Aalborg University

Overview

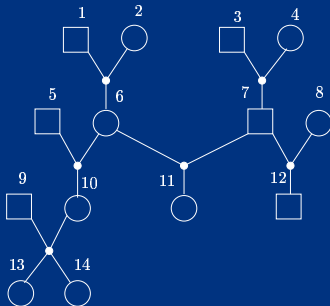
- Pedigrees
- Human DNA
- Bayesian networks and pedigrees
- Problems of linkage
- Pedigree uncertainty
- Forensic genetics

Pedigree



Male individuals represented by squares, female by circles. Individual 6 is child of individuals 1 and 2.

Marriage graph



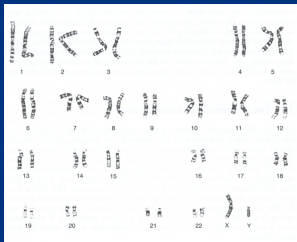
Alternative representation of a pedigree, often known as *marriage graph*.

Pedigree analysis

Analysis of data associated with pedigrees is of interest in several contexts:

- Epidemiology of genetically affected diseases
- Animal breeding
- Genetic counseling
- Forensic identification
- Identifying and understanding genetic relationships

Human chromosomes



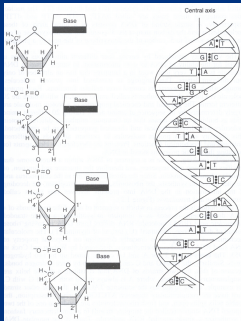
23 pairs of chromosomes in nucleus of human cell.

One pair determines gender: male XY, female XX. Other 22 are *homologous* pairs of DNA molecules.

Only homologous pairs are considered in this lecture.

DNA molecules

A double helix composed by 4 different nucleotides:
C, A, G, and T, binding in pairs C–G and A–T.



Genetic markers

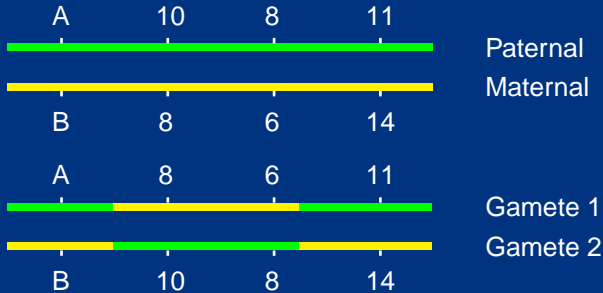
An area on a chromosome is a *locus* and the DNA composition on that area is an *allele*.

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

The *genotype* of an individual is the unordered pair of alleles. Not always observable. The *phenotype* of an individual is an observable characteristic, e.g. eye colour.

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

Meiosis



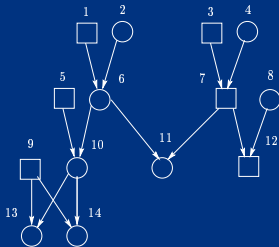
During human reproduction cells form *gametes*, where maternal and paternal DNA is mixed. A child receives one randomly chosen gamete from mother and one from father, to form a new homologous pair.

Bayesian networks for pedigrees

Different alternatives available

- *Genotype network*: Nodes represent genotypes of individuals.
- *Allele network*: Nodes represent alleles, two for each individual.
- *Segregation network*: Nodes represent alleles, two for each individual, or *segregation indicators*, i.e. whether paternal or maternal allele is segregated during meiosis.

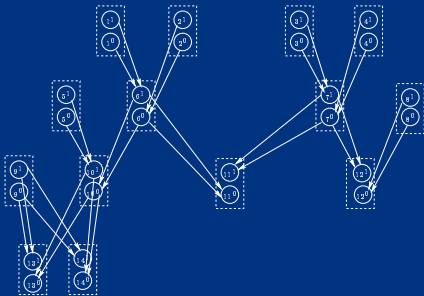
Genotype network



Mendel's first law implies that local Markov property holds. State space at each node is the *genotype*, i.e. unordered pair of alleles.

Size of specification $n(a(a+1)/2)^3 = O(na^6)$.

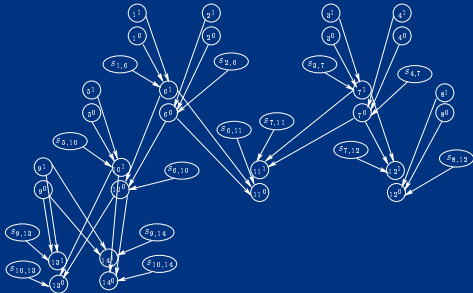
Allele network



Each individual is represented with two alleles. Visually more complex, but computationally simpler.

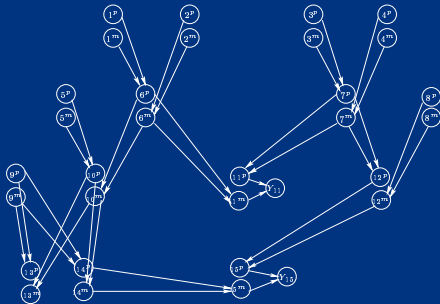
Size of specification $2na^3$.

Segregation network



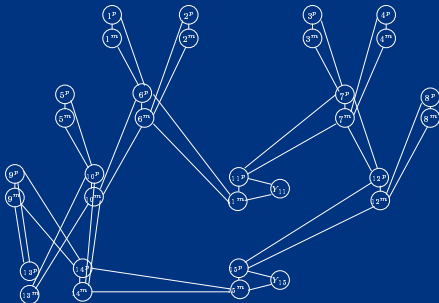
Ovals represent **segregation indicators**: 1 for paternal transmission, 0 for maternal. State at nodes of graph children is deterministic function of states at graph parent nodes. Size $4na^3$, but more expressive.

Adding phenotypic information



Allele network with additional individual and phenotypic information on two individuals.

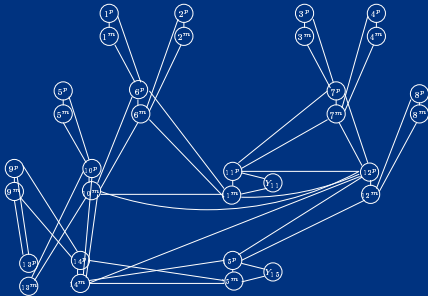
Dependence graph



Dependence graph (moral graph) reflects factorisation of joint density into terms of form

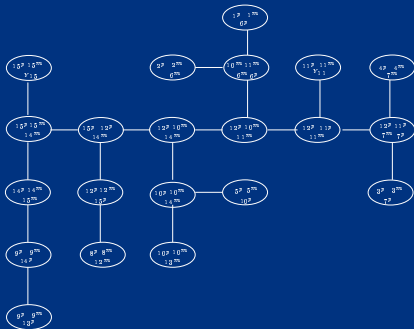
$$\phi(x_{\{v\} \cup \text{pa}(v)}) = p(x_v | x_{\text{pa}(v)}).$$

Triangulation



Links are added to make graph *triangulated*, i.e. so that all cycles of length ≥ 4 have chords.

Junction tree



Cliques are arranged into a tree with $C_1 \cap C_2 \subseteq D$ for all cliques D on path between C_1 and C_2 .

Result of basic computation

Normalisation constant after **COLLINFO** gives *likelihood* for observations $p(x_E)$, useful for estimating unknown parameters of genetic model.

DISTINFO gives $p(x_v | x_E)$ which *predicts* genetic information at other nodes, useful e.g. for identifying inheritance patterns of disease genes.

Linkage analysis

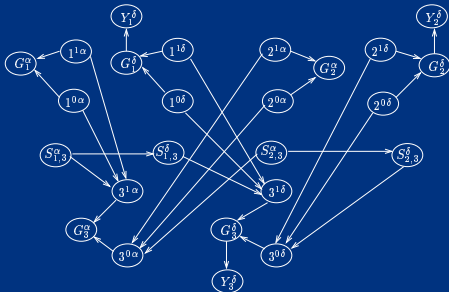
If two loci are close, it is more probable that paternal alleles are inherited together.

recombination fraction between two loci r is the probability that the segregated alleles come from different chromosomes. $0 \leq r \leq 1/2$.

If distances along chromosome are measured in *genetic map distance* $\lambda = -\{\log(1 - 2r)\}/2$ measured in Morgans.

Haldane's model says that cross-overs happen according to a Poisson process of rate 1 per Morgan.

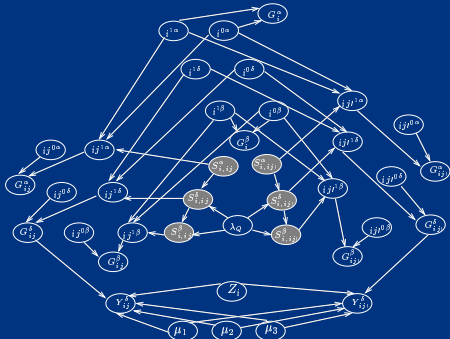
Linkage analysis



Segregation network can express dependence between loci. Here two loci of which one is a known marker.

$$P(S_{i,j}^\delta = 1 | S_{i,j}^\alpha = 1) = 1 - r, \text{ etc.}$$

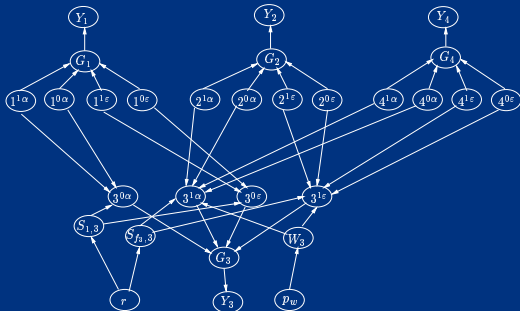
QTL mapping



Three loci: Two known markers, and a possible major gene for a quantitative trait (milk yield) is hunted.

QTL abbreviates **Q**uantitative **T**rait **L**ocus.

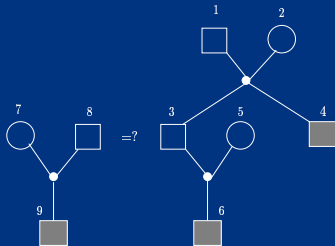
Fur colour of foxes



Female foxes are mated with two males.

Network describes a two-locus linkage model for fur colour of foxes.

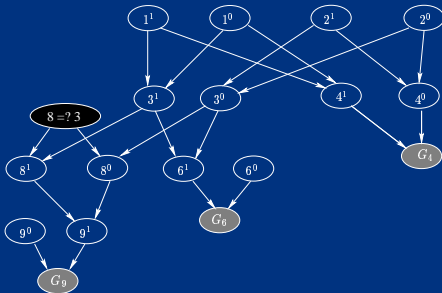
Disputed paternity



Individual 9 claims to be son of individual 3, but with different mother. DNA is available from individuals 9, 6, and 4.

6 is a known son of 3, and 4 is known brother of 3.

BN representation



Bayesian network representation of paternity problem. The black “target node” is binary and indicates whether the alleles from individual 8 are drawn at random or equal to those of 3.

Conclusions

- Many problems associated with pedigree analysis have a natural formulation in terms of Bayesian networks
- Flexibility and modularity of Bayesian networks yields possibilities for incorporating variations of pure pedigree problems
- General local computational algorithm makes case-specific algorithms redundant.
- Variant of algorithm (random propagation) makes efficient basis for developing Markov chain Monte Carlo methods when exact computation becomes unfeasible.