Methods for gene mapping and haplotype analysis

Prof. Hannu Toivonen

hannu.toivonen@cs.helsinki.fi

University of Helsinki – Department of Computer Science

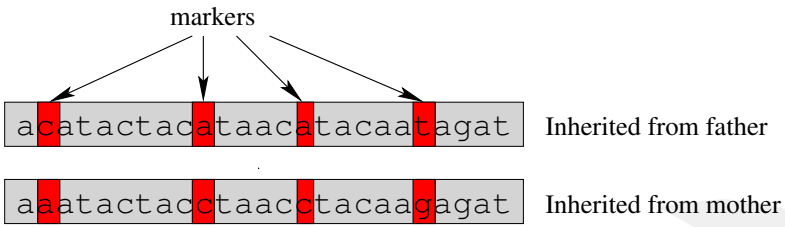
Overview

- Discovery and utilization of patterns in the human genome
  - Shared patterns → family relationships, population history
  - Patterns associated with a disease → gene mapping
- Two kinds of challenges
  1. Obtaining the patterns
  2. Utilizing them, e.g., for gene mapping
- The focus of this talk will be on the first topic

Outline

- Part I
  - Haplotypes: the key data type
  - Gene mapping I: a major application
  - Genotypes: the data available in reality
  - Gene mapping II: using genotype data
- Part II
  - Haplotyping: reconstructing haplotypes from genotypes
  - Novel Markovian methods
  - Experiments

Haplotypes



- Marker: a polymorphic locus in dna
- Allele: a particular variant in a marker (e.g. a/c or 1/2)
- *Haplotype* = string of alleles along a single chromosome: $H_{father} = (c, a, a, t)$, $H_{mother} = (a, c, c, g)$
- Economic but informative representation of dna

Haplotypes

- 1.8 million SNP (single nucleotide) polymorphisms known to date
- Over 10 million SNPs anticipated in the human genome
  - Average distance between SNPs appr. 300 bases
- Figures in a typical gene mapping study
  - some markers in the area of interest
  - 20 – 100.000 markers
  - distance between markers 1.000 – 5.000.000 bases
  - 100 – 1.000 individuals

Haplotypes

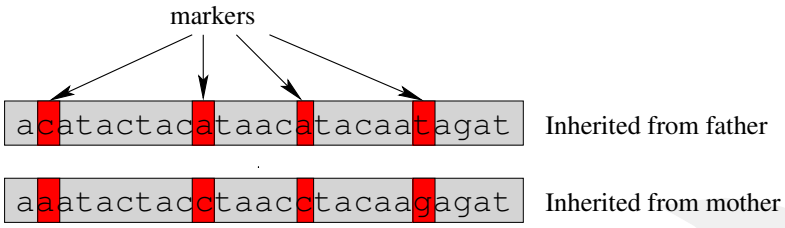
- Haplotypes are fragmented by recombination (in meiosis)
  - A haplotype is a unique mosaic of fragments from the ancestors
- For a geneticist, recombination is an enemy and a friend
  - – Haplotypes are stochastic and fragmented
  - + Fragmentation allows analysis of local patterns
- For us: *pattern* $\approx$ haplotype fragment
- Haplotype fragments are potentially inherited for generations

Gene mapping I

- **The gene mapping problem**: given a set of haplotypes from people who have a hereditary disease (the cases), predict the locus of a disease susceptibility gene
- Usually also a set haplotypes from healthy controls is given
- Outline of a solution:
  - search for haplotype patterns shared by (or over-represented in) the cases
  - predict the gene to be close to the best patterns
  - (this approach is known as association analysis)

- This works in population isolates (such as Finland :-) where individuals carrying the mutated gene have potentially inherited it from a relatively recent common ancestor $\rightarrow$ they share common haplotype patterns around the gene
- Issues:
  - weak effects of genes
  - diagnostic problems
  - gene-gene interactions
  - gene-environment interactions
  - small data sets
  - ...

Haplotypes and genotypes



- Haplotype = string of alleles along a single chromosome: $H_{father} = (c, a, a, t)$, $H_{mother} = (a, c, c, g)$
- *Genotype* = list of unordered allele pairs along the pair of chromosomes: $G = (\{a, c\}, \{a, c\}, \{a, c\}, \{g, t\})$
- Current laboratory techniques produce genotypes, not haplotypes!

Haplotypes vs. genotypes

- Haplotypes are a key to most genetic studies
- We will discuss two problems

1. How to reconstruct haplotypes from genotypes?
2. How to do gene mapping using genotypes?

- The first problem will consitute the main body of this talk

Gene mapping II

- Given genotype data, how to do gene mapping?
- Haplotypes are potentially inherited with the disease, not genotypes
- Idea: keep on working with haplotype patterns, just modify how their frequencies are counted
- A slight modification to the previous solution:
  - the frequency of a haplotype pattern is the fraction of genotypes that possibly contain the pattern
- This is an optimistic approach, and more complex weighting schemes are possible
- Experimental result: this really works

Haplotyping

Haplotyping

- Genotype: $(\{1,2\}\{3,3\}\{2,4\}\{2,4\})$
- Possible haplotype configurations:

$\binom{1322}{2344}, \binom{1324}{2342}, \binom{1342}{1324}, \binom{1344}{2322}$

Haplotyping

- Genotype: $(\{1,2\}\{3,3\}\{2,4\}\{2,4\})$
- Possible haplotype configurations:

$\binom{1322}{2344}, \binom{1324}{2342}, \binom{1342}{1324}, \binom{1344}{2322}$

Haplotyping

- Genotype: $(\{1,2\}\{3,3\}\{2,4\}\{2,4\})$
- Possible haplotype configurations:

$\binom{1322}{2344}, \binom{1324}{2342}, \binom{1342}{1324}, \binom{1344}{2322}$

Haplotyping

- Genotype: $(\{1,2\}\{3,3\}\{2,4\}\{2,4\})$
- Possible haplotype configurations:

$\binom{1322}{2344}, \binom{1324}{2342}, \binom{1342}{1324}, \binom{1344}{2322}$

Haplotyping

- Genotype: $(\{1,2\}\{3,3\}\{2,4\}\{2,4\})$
- Possible haplotype configurations:

$\binom{1322}{2344}$, $\binom{1324}{2342}$, $\binom{1342}{1324}$, $\binom{1344}{2322}$

Haplotyping

- Genotype: $(\{1,2\}\{3,3\}\{2,4\}\{2,4\})$
- Possible haplotype configurations:

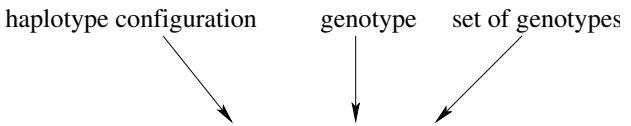
$\binom{1322}{2344}$, $\binom{1324}{2342}$, $\binom{1342}{1324}$, $\binom{1344}{2322}$

- For a genotype G with k heterozygous markers there are 2^{k-1} different haplotype configurations.

Haplotyping

- **The haplotyping problem**:
- Input: a set $\mathcal{G}$ of genotypes
- Output: the most probable haplotype configuration for each genotype $G \in \mathcal{G}$
- Haplotypes of subjects from same population tend to be similar to each other $\Rightarrow$ statistical inference can be used to deduce the underlying haplotypes
- (this is the population-based variant of the problem)

Statistical assumptions



$$P\big(\{H_1, H_2\} \mid G; \mathcal{G}\big) = ?$$

$$\text{argmax}_{\{H_1,H_2\}} P\big(\{H_1, H_2\} \mid G; \mathcal{G}\big) = ?$$

- (We keep on conditioning on $\mathcal{G}$, but do not explicitly mention it anymore)

Statistical assumptions

$$P(\{H_1, H_2\} \mid G) \propto$$

$$\begin{cases} P(\{H_1, H_2\}) & \text{if } \{H_1, H_2\} \text{ compatible with G;} \\ 0 & \text{otherwise.} \end{cases}$$

- Assume Hardy-Weinberg equilibrium and random mating $\Rightarrow$ haplotypes of an invidual are independent of each other:

$$P(\{H_1, H_2\}) = \begin{cases} 2P(H_1)P(H_2) & \text{if } H_1 \neq H_2 \\ P(H_1)^2 & \text{if } H_1 = H_2 \end{cases}$$

- The problem reduces to modeling the distribution $P(H)$

Statistical assumptions

- The standard model assumes haplotypes are inherited as a whole
- The model is just a list of haplotype probabilities, for example:
  $P(ABCDE) = 0.6$
  $P(abCDE) = 0.2$
  $P(AbCdD) = 0.1$
  $P(ABcde) = 0.1$
- This is done practically in all previous work

Our relaxed assumptions

- Markers can be sparsely located
- Many recombinations within the haplotypes $\Rightarrow$ large number of different haplotypes
- Most or all haplotypes can be unique
- Possibly only weak statistical dependencies ("LD") between markers

Haplotyping: solutions

- We will next look at solutions to the haplotyping problem
- Components:
  - Defining models for distribution $P(H)$
  - Finding the pair $\{H_1, H_2\}$ that approximately maximizes $P(H_1)P(H_2)$
- Three increasingly complex Markov models
- An outline of algorithms for using them

Haplotype fragments

- Example haplotypes (all unique)

```
123241
223241
323255
144241
144221
```

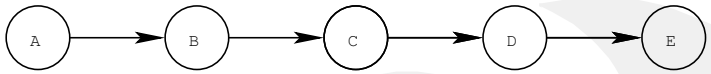
- Frequent fragments

```
-232--      fr=3
---241      fr=3
-2324-      fr=2
1442--      fr=2
```


Markov chain

- Assume independence of non-neighboring markers: consider the haplotype as a (first-order) Markov chain:

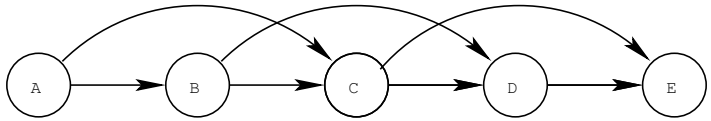
$$P(H) \approx P(H(1)) \prod_{i=2,\ldots,\ell} P(H(i) \mid H(i-1)). \qquad (1)$$



- $P(ABCDE) = P(A) \cdot P(B|A) \cdot P(C|B) \cdot P(D|C) \cdot P(E|D)$

Markov chain of order d (MC-d)

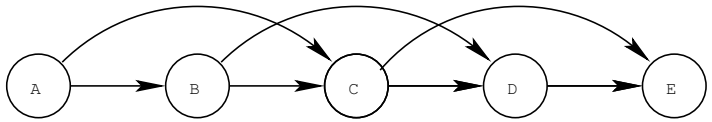
- A neighborhood of several markers (e.g., $d = 2$) is potentially more informative:



$$P(H) \approx P(H(1,d)) \prod_{i=d+1,\ldots,\ell} P(H(i) \mid H(i-d,i-1)). \qquad (2)$$

- $P(ABCDE) = P(AB) \cdot P(C|AB) \cdot P(D|BC) \cdot P(E|CD)$

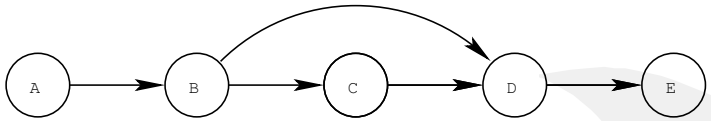
Markov chain of order d (MC-d)



- A problem: selecting a suitable value for d. Too small values will not make use of all the statistical dependencies; too large values will overfit the model to the data.
- A further problem: LD may vary within the marker map; it is possible that no single value of d is suitable for all parts of the map.

Variable order Markov Chain MC-VL

- Goal: adjust the context for each marker of each haplotype individually to obtain flexible balance between generality and informativeness, e.g.:

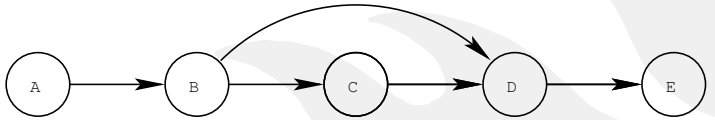


$$P(H) \approx P(H(1)) \prod_{i=2,\ldots,\ell} P(H(i) \mid H(s_i, i-1)), \qquad (3)$$

where $s_i = \min\{s \mid H(s,i) \in \mathcal{F}_{\mathrm{VL}}\}$.

Variable order Markov Chain MC-VL

- How to select the number of markers in each context?
- Solution: use the largest frequent context, where "frequent" means frequency of at least some constant c
- Motivation: use the longest context for which there is sufficient evidence



- $P(ABCDE) = P(A) \cdot P(B|A) \cdot P(C|B) \cdot P(D|BC) \cdot P(E|D)$

Estimating conditional probabilities

- (Conditional) haplotype fragment probabilities are needed in the above models
- Probabilities are estimated by the corresponding frequencies
- Example: $P(D|ABC) \approx \frac{fr(ABCD-)}{fr(ABC--)}$
- ...but how to estimate haplotype frequencies from genotype data?

Fragment frequency estimation

- genotype #het.markers weight

genotype	#het.markers	weight
...{3,4}{2,3}{3,3}...	2	0.5
...{3,4}{2,3}{3,4}...	3	0.25
...{1,3}{1,2}{1,4}...	3	0.25
...{3,3}{3,3}{4,4}...	0	2.0

- haplotype fragments
  - `- 3 2 3 -`
  - `- 3 2 4 -`
  - `- 3 3 4 -`

Fragment frequency estimation

- genotype #het.markers weight
  ...{3,4}{2,3}{3,3}... 2 0.5
  ...{3,4}{2,3}{3,4}... 3 0.25
  ...{1,3}{1,2}{1,4}... 3 0.25
  ...{3,3}{3,3}{4,4}... 0 2.0
- haplotype fragments
  - 3 2 3 - (freq $= 0.5 + 0.25 = 0.75$)
  - 3 2 4 -
  - 3 3 4 -

Fragment frequency estimation

- genotype #het.markers weight
  ...{3,4}{2,3}{3,3}... 2 0.5
  ...{3,4}{2,3}{3,4}... 3 0.25
  ...{1,3}{1,2}{1,4}... 3 0.25
  ...{3,3}{3,3}{4,4}... 0 2.0
- haplotype fragments
  - 3 2 3 - (freq $= 0.75$)
  - 3 2 4 - (freq $= 0.25 + 0.25 = 0.5$)
  - 3 3 4 -

Fragment frequency estimation

- genotype #het.markers weight
  ...{3,4}{2,3}{3,3}... 2 0.5
  ...{3,4}{2,3}{3,4}... 3 0.25
  ...{1,3}{1,2}{1,4}... 3 0.25
  ...{3,3}{3,3}{4,4}... 0 2.0
- haplotype fragments
  - 3 2 3 - (freq $= 0.75$)
  - 3 2 4 - (freq $= 0.5$)
  - 3 3 4 - (freq $= 0.25 + 2.0 = 2.25$)

Fragment frequency estimation

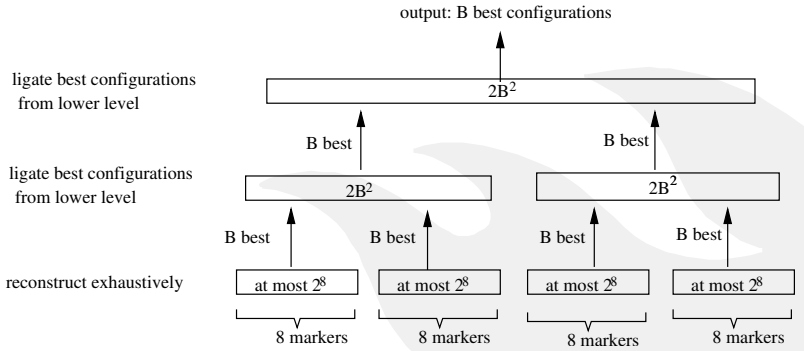
- For the fixed order Markov chain of order d, just enumerate all fragments of lenght d and $d + 1$ and compute their frequencies as described above
- For the variable order Markov chain, depth-first search is used to find the set of frequent fragments: start from frequent fragments of length 1, and expand fragments to the right until the frequency drops below a given threshold c

Haplotype reconstruction

- Task: Find $\text{argmax}_{\{H_1,H_2\}\text{compatible with G}} P(H_1)P(H_2)$ for each genotype $G \in \mathcal{G}$
- Problem: the number of haplotype configurations for a genotype G is exponential in the number of heterozygous markers in G
  $\Rightarrow$ exhaustive search through of all possible haplotype configurations is not practical.

Haplotype reconstruction

- We use a divide-and conquer approach (motivated by the related "PL" approach by Niu et al. (2002)) in the haplotype reconstruction step to restrict the search space.



Experiments

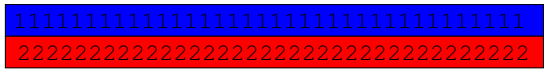
- Performance of the methods
- Sensitivity to parameters
- Comparison to three state-of-the-art approaches:
  - plem: EM algorithm with partition ligation
  - snphap: EM algorithm with sequential pruning
  - Phase: MCMC with coalescence prior
- All treat haplotypes as non-divisible units ($\approx$ no recombinations)
- Controlled experiments with simulated data

Simulated data

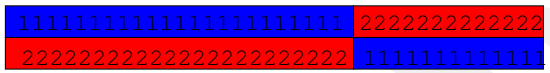
- The simulated setting corresponds to a association study in a population isolate
- 20 independent founders, random mating for 20 generations, no immigration, uniform recombination rate, final population size 100000
- SNP or microsatellite (with 6 alleles/marker) markers
- 32 markers, sample of 500 genotypes
- Main parameter: marker spacing, which ranges between 0.01-1cM between each adjacent pair of markers; giving total map length of 0.31-31 cM.
- 10 independents simulations per setting, over which results are averaged

Performance measure

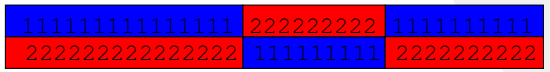
- "switch distance" = number of neighboring phase relations reconstructed incorrectly.



Original haplotype configuratior



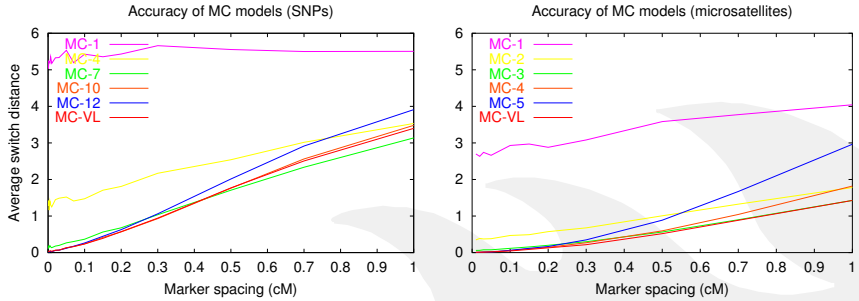
Switch distance=1



Switch distance=2

Results for simulated data

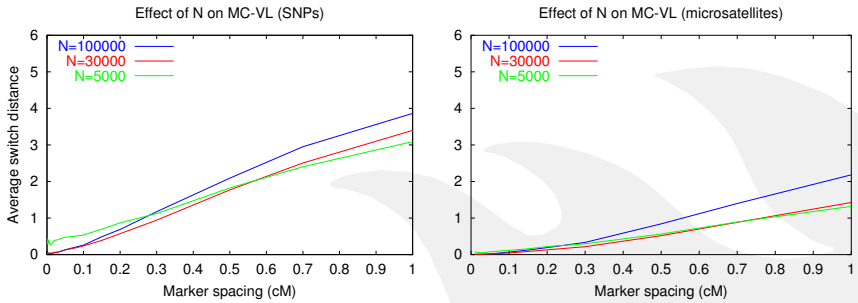
Effect of parameter d (order of Markov chain) of model MC-d



X-axis: task difficulty; Y-axis: error

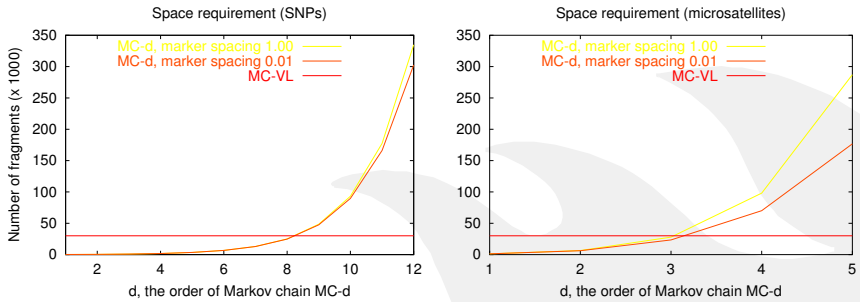
Results for simulated data

Effect of parameter N (number of fragments) of model MC-VL



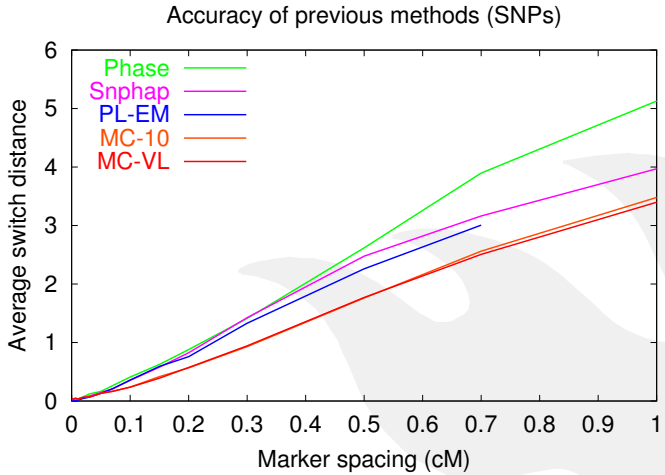
Results for simulated data

Empirical space requirement



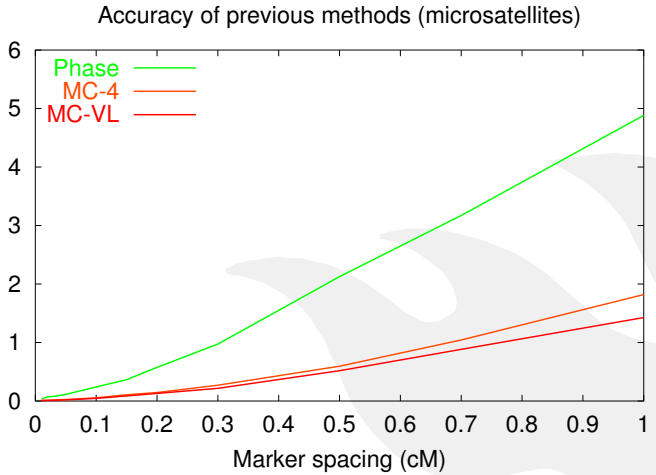
Results for simulated data

Comparison with existing methods, SNP data



Results for simulated data

Comparison with Phase, microsatellite data



Results for Daly data

- Daly data: a real data set with densely spaced markers
- 103 SNP markers, 147 genotypes, approx. 10 % missing data.
  Results:

Method	Error rate	Avg. Switch distance
SNPHAP	0.410	1.292
PHASE	0.973	N/A
MC-9	0.483	0.932
MC-VL	0.449	0.905

Running times

- Proportional to space complexity
  - MC-d: exponential in d, linear in number of markers
  - MC-VL: linear in number of patterns used
  - in practice: 1 – 3 minutes
- Previous solutions
  - SNPHAP: 2 – 20 seconds
  - PL-EM: 3 – 100 seconds
  - PHASE: 5 – 30 hours

Summary

- Novel statistical haplotyping methods suitable for long marker maps, typically used in gene mapping studies
- Exploits local dependencies (LD) with a Markov chain model; variable order Markov chain is used for improved adaptivity
- With simulated data, outperforms competing methods when the distance between neighboring markers is at least 0.05 cM
- The method was competitive also with the real and dense Daly data
- Implementation and data sets are available at: http://www.cs.helsinki.fi/group/genetics/haplotyping.html

On-going and future work

- An EM-like, iterative algorithm to estimate fragment frequencies
- $\rightarrow$ better fit of the Markovian models
- New fragment-based models for haplotype frequencies
- A sequential reconstruction algorithm
- Other genetic applications for haplotype modeling (discovery of haplotype blocks, reconstruction of founders, ...?)
- Other applications for variable order Markov chains
- Better ways to choose the variable order of a Markov chain

Acknowledgements

- Haplotyping: Lauri Eronen and Floris Geerts
- http://www.cs.helsinki.fi/group/genetics/haplotyping.html
- Lauri Eronen, Floris Geerts, and Hannu Toivonen: A Markov chain approach to reconstruction of long haplotypes. *Pacific Symposium on Biocomputing (PSB 2004)*, 104-115, Hawaii, USA, January 2004. World Scientific.