Statistical methods for haplotype inference – Part I

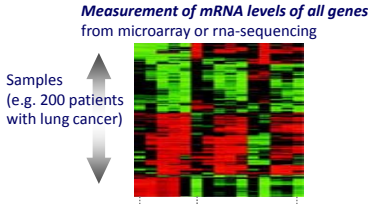
Lecture 3 – May 21th, 2013
GENOME 541, Spring 2013

Su-In Lee
GS & CSE, UW
suinlee@uw.edu

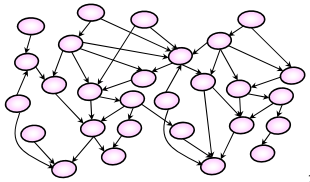
1

Learning gene regulatory networks

- **Input:**

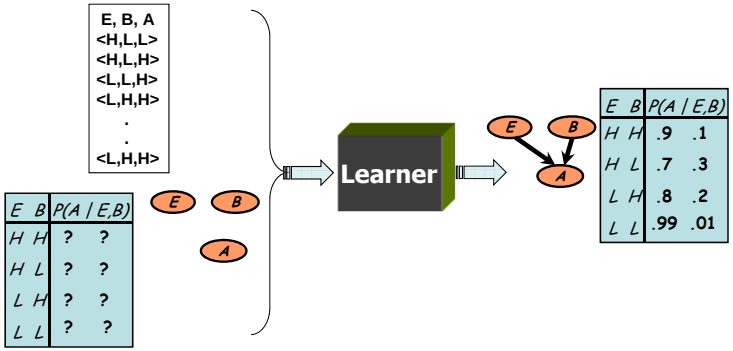


- **Goal:** Reconstruct the *gene regulatory network* underlying genome-wide gene expression
- **Method:** Probabilistic models to represent the regulatory network



2

Unknown structure, complete data

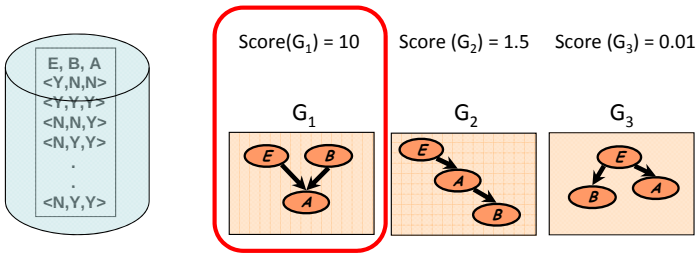


- Network structure is not specified
  - Learner needs to estimate both structure and parameters
- Data does not contain missing values

3

Score-based learning

- Define scoring function that measures how well a certain structure fits the observed data.



- Search for a structure that maximizes the score.

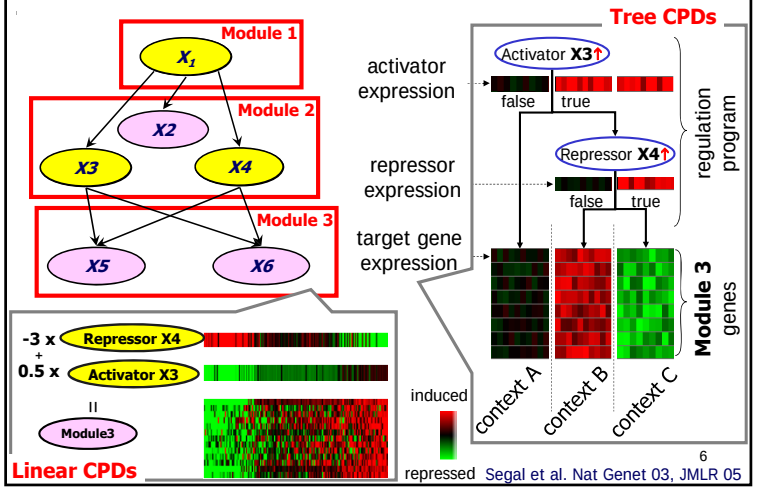
4

Challenges

- Too large search space
  - What is the number of possible structures of *n* genes? $\sim 3^{n^2/2}$
- Computationally costly
- Heuristic approaches may be trapped to local maxima.
- Biologically motivated constraints can alleviate the problems
  - Module-based approach
  - Only the genes in the candidate regulators list can be parents of other variables

5

The Module networks concept

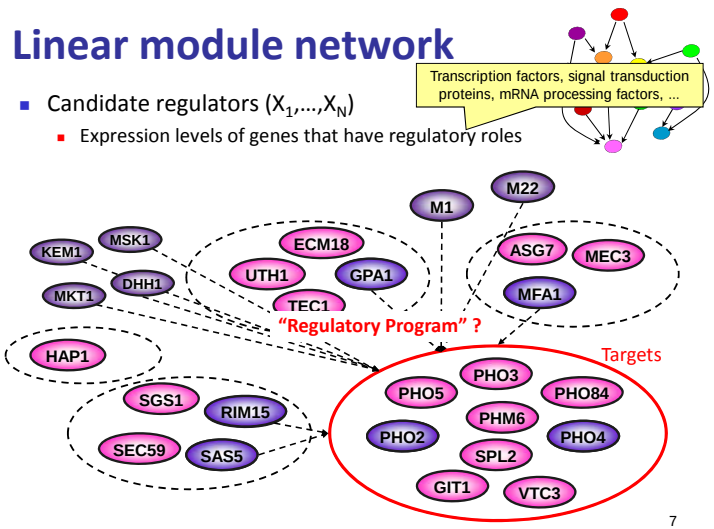


6

Segal et al. Nat Genet 03, JMLR 05

Linear module network

- Candidate regulators ($X_1,\ldots,X_N$)
  - Expression levels of genes that have regulatory roles



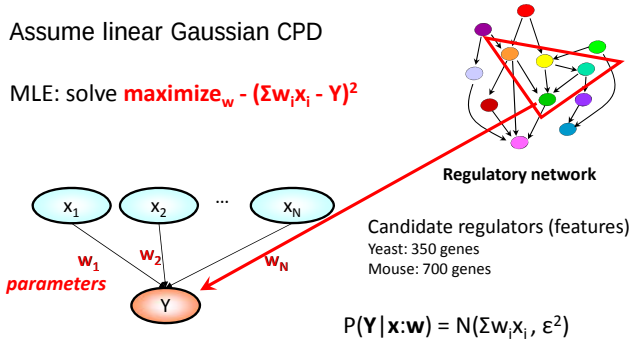
Lee et al., PLoS Genet 2009

7

7

Feature selection via regularization

- Assume linear Gaussian CPD
- MLE: solve **maximize$_w$ - $(\Sigma w_i x_i - Y)^2$**



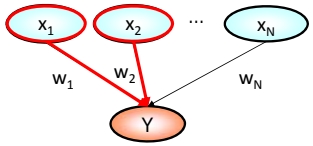
$P(\mathbf{Y}|\mathbf{x}:\mathbf{w}) = N(\Sigma w_i x_i, \varepsilon^2)$

Problem: This objective learns too many regulators

8

L_1 regularization

- **"Select" a subset of regulators**
  - Combinatorial search?
  - Effective feature selection algorithm: L_1 regularization (LASSO)
    [Tibshirani, J. Royal. Statist. Soc B. 1996]
  - minimize$_w$ $(\Sigma w_i x_i - Y)^2 + \Sigma\, C\, |w_i|$: convex optimization!
    $\Rightarrow$ Induces sparsity in the solution **w** (Many w_i's set to zero)



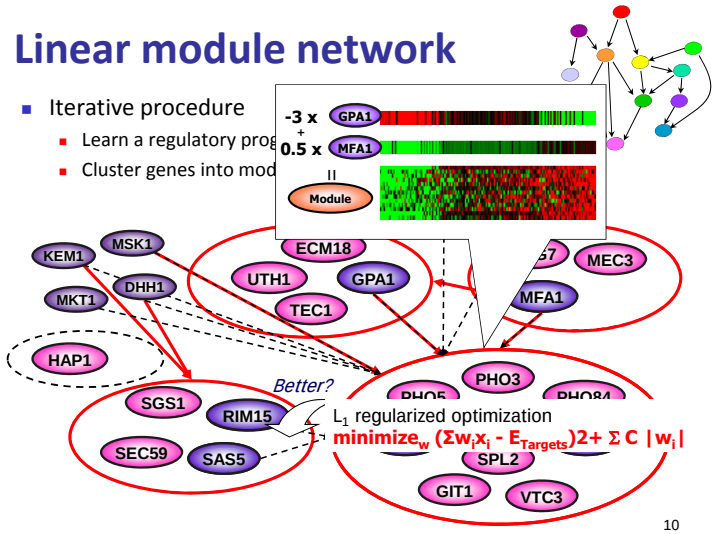
Candidate regulators (features)
Yeast: 350 genes
Mouse: 700 genes

$P(Y|\mathbf{x}:\mathbf{w}) = N(\Sigma w_i x_i, \varepsilon^2)$

9

Linear module network

- Iterative procedure
  - Learn a regulatory prog
  - Cluster genes into mod



Lee et al., PLoS Genet 2009

10

10

Summary

- Basic concepts on Bayesian networks
- Probabilistic models of gene regulatory networks
- Learning algorithms
  - Parameter learning
  - Structure learning
  - Structure discovery
- Evaluation
- Recent probabilistic approaches to reconstructing the regulatory networks

11

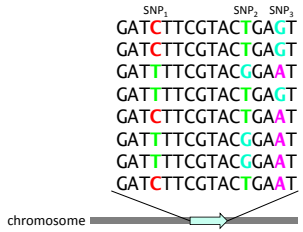
Haplotype inference (5/21, 5/23)

- Background & motivation
- Problem statement
- Statistical methods for haplotype inference
  - Clark's algorithm
  - Expectation Maximization (EM) algorithm *Today*
  - Coalescent-based methods and HMM
  - Haplotype inference on sequence data
- Example applications

12

Genetic variation

- Single nucleotide polymorphism (SNP)
  - Each variant is called an *allele*; each allele has a *frequency*



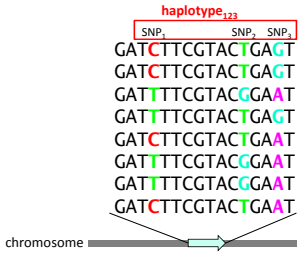
$P(SNP_1 = C) = 4/8$
$P(SNP_1 = T) = 4/8$
:

- How about the relationship between alleles of neighboring SNPs?
  - We need to know about haplotype

13

Haplotype

- A combination of alleles present in a chromosome
- Each haplotype has a *frequency*, which is the proportion of chromosomes of that type in the population
- There are 2^N possible haplotypes
  - But in fact, far fewer are seen in human population



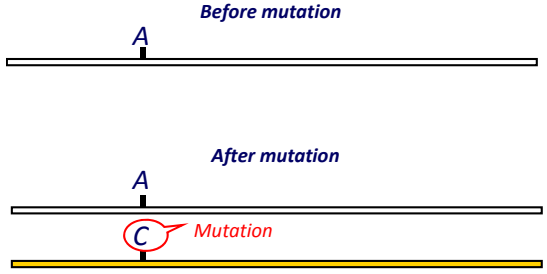
Haplotype frequencies

$P(haplotype_{123} = TGA) = 3/8$
$P(haplotype_{123} = CTG) = 2/8$
$P(haplotype_{123} = CTA) = 2/8$
$P(haplotype_{123} = TTG) = 1/8$

14

History of two neighboring alleles

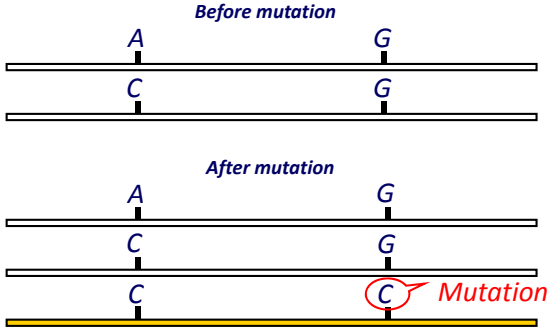
- Alleles that exist today arose through ancient mutation events…



15

History of two neighboring alleles

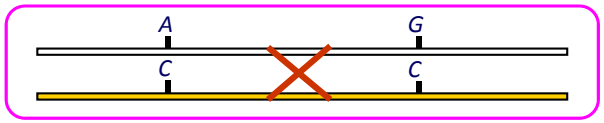
- One allele arose first, and then the other…



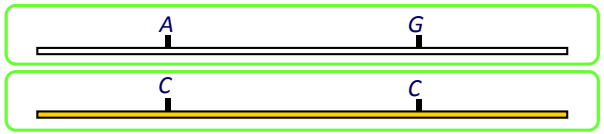
Haplotype: combination of alleles present in a chromosome

16

Recombination can create more haplotypes

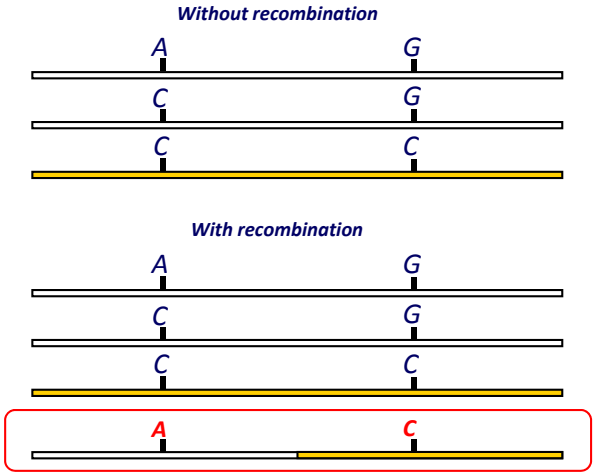


- No recombination (or 2n recombination events)



- Recombination (2n+1 recombination events)

17

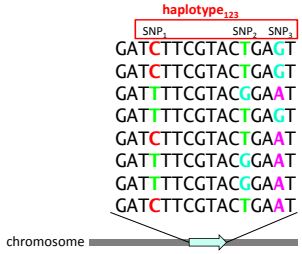


Recombinant haplotype

18

Haplotype

- What determines haplotype frequencies?
  - Recombination rate (*r*) between neighboring alleles in the population
  - *r* is different for different regions in genome
- Linkage disequilibrium (LD)
  - Non-random association of alleles at two or more loci, not necessarily on the same chromosome.



Haplotype frequencies

P(haplotype$_{123}$ = TGA) = 3/8
P(haplotype$_{123}$ = CTG) = 2/8
P(haplotype$_{123}$ = CTA) = 2/8
P(haplotype$_{123}$ = TTG) = 1/8

19

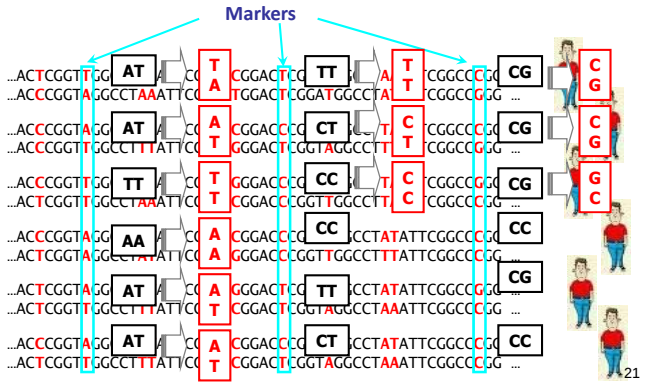
How can we measure haplotypes ?

- Haplotypes can be generated through laboratory-based experimental methods
  - X-chromosome in males
  - Sperm typing
  - Hybrid cell lines
  - Other molecular techniques
- Computational approaches
  - **Input:** Genotype data from individuals in a population
  - **Output:** Haplotypes of each individual in the population

20

Haplotype inference problem

- Sequence and SNP array data generally take the form of unphased genotypes



21

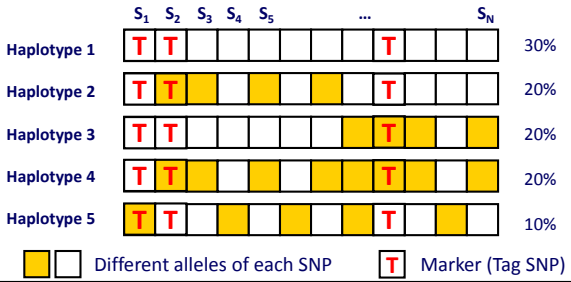
Motivation

- Enormous amounts of genotype data are being generated
  - Inexpensive genome-wide SNP microarrays
  - Whole-genome and whole-exome sequencing tools
- Determination of haplotype phase is increasingly important
  - Characterizing the relationship between genetic variation and disease susceptibility
  - Imputing low frequency variants

22

Why useful in GWAS?

- In a typical short chromosome segment, there are only a few distinct haplotypes
- Carefully selected markers can determine status of others
- We can test for association of untyped SNPs



23

Example

- Holm et al. *Nat Genet* (2011)
  - Used the inferred haplotypes for accurate imputation of a putative rare causal variant in other individuals, to obtain a stronger association signal

24

Example

- Sick sinus syndrome (SSS)
  - Characterized by slow heart rate, sinus arrest and/or failure to increase heart rate with exercise
- Genome-wide association scan of 7.2M SNPs with 792 SSS cases and 37,592 controls

Source	SNP	*P* value	OR	MAF
Directly genotyped	rs1055061	2.2×10^{-5}	1.57	0.055
Imputed from HapMap2	rs10130976	4.4×10^{-7}	1.74	0.048
Imputed from the 1000 Genomes project	14-22399934	5.8×10^{-9}	2.06	0.052
Imputed from the Human1M-Duo chip	rs2231801	1.3×10^{-13}	3.64	0.010
Imputed from the HumanOmni1-Quad chip	rs2231801	1.5×10^{-10}	3.05	0.012
Imputed from the HumanOmni1-Quad chip	**rs28730774**	1.6×10^{-11}	3.49	0.010

- The association analysis yielded association with several correlated SNPs in and near MYH6-MYH7 (never before associated with SSS)

25

Outline

- Background & motivation
- Problem statement
- Statistical methods for haplotype inference
  - Clark's algorithm
  - Expectation Maximization (EM) algorithm *Today*
  - Coalescent-based methods and HMM
  - Haplotype inference on sequence data
- Example applications

26

Typical genotype data

- Two alleles for each individual for each marker
  - Chromosome origin for each allele is unknown

Observation

C G	$marker_1$	{CG} {TC} {GA}
T C	$marker_2$	
G A	$marker_3$	

- Multiple haplotype pairs can fit observed genotype

Possible states

C G T C G A	CTG/GCA	C G C T G A	CCG/GTA
C G C T A G	CCA/GTG	C G T C A G	CTA/GCG

27

Use information on relatives?

- Family information can help determine phase at many markers
- Can you propose examples?
- Genotype: {AT} {AA} {CG}
  - Maternal genotype: {TA} {AA} {CC} → TAC/AAC
  - Paternal genotype: {TT} {AA} {CG} → TAC/TAG
  - Then the haplotype is AAC/TAG

28

Example – inferring haplotypes

- Still, many ambiguities might not be resolved
  - Problem more serious with larger numbers of markers
- Genotype: {AT} {AA} {CG}
  - Maternal genotype: {AT} {AA} {CG}
  - Paternal genotype: {AT} {AA} {CG}
  - Cannot determine unique haplotype
- Problem
  - Determine haplotypes without parental genotypes

29

What if there are no relatives?

- Rely on linkage disequilibrium (LD)
  - LD: non-random association of variants at different sites in the genome
- Assume that population consists of small number of distinct haplotypes

30

Haplotype reconstruction

- Also called, phasing, haplotype inference or haplotyping
- Data
  - Genotype on *N* markers from *M* individuals

Individual *i*

C	G	$marker_1$
T	C	$marker_2$
G	A	$marker_3$

- Goals
  - Frequency estimation of all possible haplotypes
  - Haplotype reconstruction for individuals
  - How many out of all possible haplotypes are plausible in a population?

31

Statistical methods for haplotypes inference

- Let's focus on the methods that are most widely used or historically important
  - Browning and Browning, *Nat Rev Genet*. 2011

Published in final edited form as:
Nat Rev Genet. ; 12(10): 703–714. doi:10.1038/nrg3054.

Haplotype phasing: Existing methods and new developments

Sharon R. Browning[1,*] and **Brian L. Browning**[2,*]
[1]Department of Biostatistics, University of Washington, Seattle WA 98195, USA
[2]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle WA 98195, USA

32

Outline

- Background & motivation
- Problem statement
- Statistical methods for haplotype inference
  - Clark's algorithm
  - Expectation Maximization (EM) algorithm
  - Coalescent-based methods and HMM
  - Haplotype inference on sequence data
- Example applications

33

Clark's haplotyping algorithm

- Clark (1990) *Mol Biol Evol* **7**:111-122
- One of the first published haplotyping algorithms
  - Computationally efficient
  - Very fast and widely used in 1990's
  - More accurate methods are now available

34

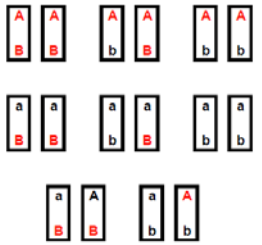
Clark's haplotyping algorithm

- Find unambiguous individuals
  - Initialize a list of known haplotypes
- What kinds of genotypes will these have?
- Unambiguous individuals
  - Homozygous at every locus (e.g. {TT} {AA} {CC})
    Haplotypes: TAC
  - Heterozygous at just one locus (e.g. {TT} {AA} {CG})
    Haplotypes: TAC or TAG

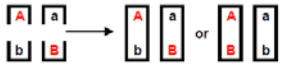
35

Unambiguous vs. ambiguous

- Haplotypes for 2 SNPs (alleles: A/a, B/b)



Unambigous Genotypes
Underlying Haplotype is Known



Ambigous Genotype
Multiple Underlying Genotypes Possible

36

Clark's haplotyping algorithm

- Find unambiguous individuals
  - Initialize a list of known haplotypes
- Resolve ambiguous individuals
  - If possible, use two haplotypes from the list
  - Otherwise, use one known haplotype and augment list
- If unphased individuals remain
  - Assign phase randomly to one individual
  - Augment haplotype list and continue from previous step

37

Parsimonious phasing - example

- Notation (more compact representation)
  - 0/1: homozygous at each locus (00,11)
  - h: heterozygous at each locus (01)

1 0 1 0 0 h	1 0 1 0 0 0 1 0 1 0 0 1
h 0 1 h 0 0	1 0 1 0 0 0 0 0 1 1 0 0
0 h h 1 h 0	0 0 1 1 0 0 0 1 0 1 1 0

38

Parsimony algorithm

- Pros
  - Very fast
  - Can deal with very long sequences
- Cons
  - No homozygotes or single SNP heterozygotes in the data
  - Some haplotypes may remain unresolved
  - Outcome depends on order in which lists are transversed
  - Naïve, not very accurate (no modeling)

39

Outline

- Background & motivation
- Problem statement
- Statistical methods for haplotype inference
  - Clark's algorithm
  - Expectation Maximization (EM) algorithm
  - Coalescent-based methods and HMM
  - Haplotype inference on sequence data
- Example applications

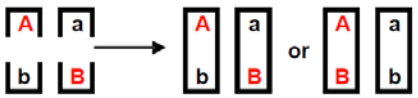
40

The EM haplotyping algorithm

- Excoffier and Slatkin *Mol Biol Evol* (1995); Qin et al. *Am J Hum Genet* (2002); Excoffier and Lischer *Molecular ecology resources* (2010)
- Why EM for haplotyping?
  - EM is a method for MLE with hidden variables.
- What are the hidden variables, parameters?
  - Hidden variables: haplotype state of each individual
  - Parameters: haplotype frequencies

Individual *n*

Haplotype status (hidden variable)

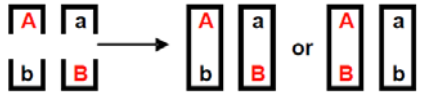


Haplotype frequencies (parameters)

p_{Ab}, p_{aB}, p_{AB}, p_{ab}

41

Assume that we know haplotype frequencies

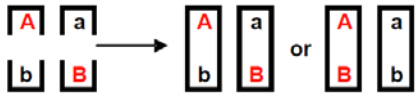


For example, if
$P_{AB} = 0.3$
$P_{ab} = 0.3$
$P_{Ab} = 0.3$
$P_{aB} = 0.1$

- Probability of first outcome:
  - $2P_{Ab}P_{aB} = 0.06$
- Probability of second outcome:
  - $2P_{AB}P_{ab} = 0.18$

42

Conditional probabilities are …



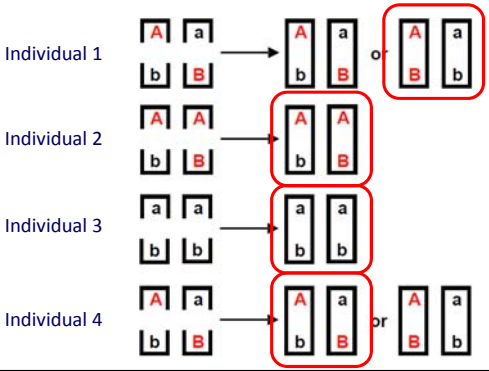
For example, if
$P_{AB} = 0.3$
$P_{ab} = 0.3$
$P_{Ab} = 0.3$
$P_{aB} = 0.1$

- Conditional probability of first outcome:
  - $2P_{Ab}P_{aB} / (2P_{Ab}P_{aB} + 2P_{AB}P_{ab}) = 0.25$
- Conditional probability of second outcome:
  - $2P_{AB}P_{ab} / (2P_{Ab}P_{aB} + 2P_{AB}P_{ab}) = 0.75$

43

Assume that we know the haplotype state of each individual

- Computing haplotype frequencies is straightforward

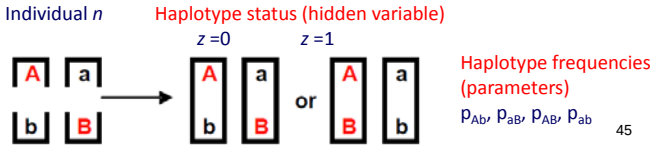


p_{AB} =?
P_{ab} =?
p_{Ab} =?
p_{aB} =?

44

EM as Chicken vs Egg

- If we know haplotype frequencies **p**'s (parameters), we can estimate the haplotype status of individuals **z**'s (hidden variables)
- If we know the haplotype state of each individual **z**'s (hidden variables), we can estimate the haplotype frequencies **p**'s (parameters)



45

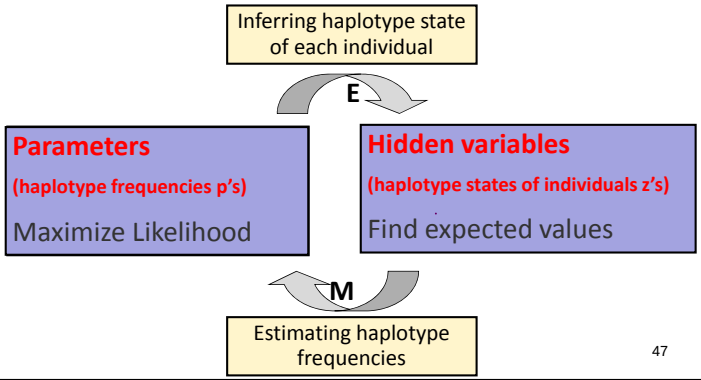
EM as Chicken vs Egg

- If we know haplotype frequencies **p**'s (parameters), we can estimate the haplotype states of individuals **z**'s (hidden variables)
- If we know the haplotype state of individuals **z**'s (hidden variables), we can estimate the haplotype frequencies **p**'s (parameters)
- BUT we know neither; iterate
  - Expectation-step: Estimate **z**'s, given haplotype frequencies **p**'s
  - Maximization-step: Estimate **p**'s, given the haplotype states of individuals **z**'s
- Overall, a clever "hill-climbing" strategy

46

Phasing By EM

- EM: Method for maximum-likelihood parameter inference with hidden variables



47

EM algorithm for haplotyping

- 1. "Guesstimate" haplotype frequencies
- 2. Use current frequency estimates to replace ambiguous genotypes with fractional counts of phased genotypes
- 3. Estimate frequency of each haplotype by counting
- 4. Repeat steps 2 and 3 until frequencies are stable

48

Phasing by EM

Data:

1 0 h h 1	1 0 0 0 1	¼
	1 0 1 1 1	¼
	1 0 0 1 1	¼
	1 0 1 0 1	¼
h 0 0 1 h	0 0 0 1 0	¼
	1 0 0 1 1	¼
	0 0 0 1 1	¼
	1 0 0 1 0	¼
1 h h 1 1	1 0 0 1 1	¼
	1 1 1 1 1	¼
	1 0 1 1 1	¼
	1 1 0 1 1	¼

Phasing by EM

Data:

1 0 h h 1	1 0 0 0 1	¼
	1 0 1 1 1	¼
	1 0 0 1 1	¼
	1 0 1 0 1	¼
h 0 0 1 h	0 0 0 1 0	¼
	1 0 0 1 1	¼
	0 0 0 1 1	¼
	1 0 0 1 0	¼
1 h h 1 1	1 0 0 1 1	¼
	1 1 1 1 1	¼
	1 0 1 1 1	¼
	1 1 0 1 1	¼

Frequencies	
0 0 0 1 0	1/12
0 0 0 1 1	1/12
1 0 0 0 1	1/12
1 0 0 1 0	1/12
1 0 0 1 1	3/12
1 0 1 0 1	1/12
1 0 1 1 1	2/12
1 1 0 1 1	1/12
1 1 1 1 1	1/12

Phasing by EM

Data:

1 0 h h 1	1 0 0 0 1	¼
	1 0 1 1 1	¼
	1 0 0 1 1	¼
	1 0 1 0 1	¼
h 0 0 1 h	0 0 0 1 0	¼
	1 0 0 1 1	¼
	0 0 0 1 1	¼
	1 0 0 1 0	¼
1 h h 1 1	1 0 0 1 1	¼
	1 1 1 1 1	¼
	1 0 1 1 1	¼
	1 1 0 1 1	¼

Haplotypes

0.4
0.6
0.75
0.25
0.6
0.4

Frequencies	
0 0 0 1 0	1/12
0 0 0 1 1	1/12
1 0 0 0 1	1/12
1 0 0 1 0	1/12
1 0 0 1 1	3/12
1 0 1 0 1	1/12
1 0 1 1 1	2/12
1 1 0 1 1	1/12
1 1 1 1 1	1/12

Expectation

Phasing by EM

Maximization

Data:

1 0 h h 1	1 0 0 0 1	¼
	1 0 1 1 1	¼
	1 0 0 1 1	¼
	1 0 1 0 1	¼
h 0 0 1 h	0 0 0 1 0	¼
	1 0 0 1 1	¼
	0 0 0 1 1	¼
	1 0 0 1 0	¼
1 h h 1 1	1 0 0 1 1	¼
	1 1 1 1 1	¼
	1 0 1 1 1	¼
	1 1 0 1 1	¼

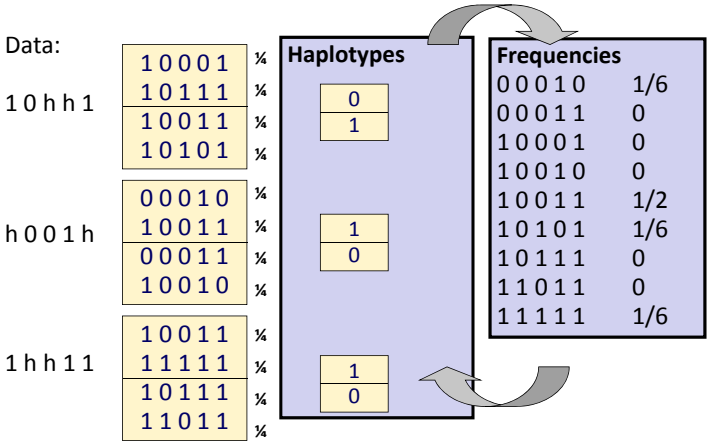
Haplotypes

0.4
0.6
0.75
0.25
0.6
0.4

Frequencies	
0 0 0 1 0	.125
0 0 0 1 1	.042
1 0 0 0 1	.067
1 0 0 1 0	.042
1 0 0 1 1	.325
1 0 1 0 1	.1
1 0 1 1 1	.067
1 1 0 1 1	.067
1 1 1 1 1	.1

Expectation

Phasing by EM



53

Computational cost (for SNPs)

- Consider sets of *m* unphased genotypes
  - Markers 1..*m*

For example, if m=10

- If markers are bi-allelic
  - 2^m possible haplotypes = 1024
  - 2^{m-1} (2^m + 1) possible haplotype pairs = 524,800
  - 3^m distinct observed genotypes = 59,049
  - 2^{n-1} reconstructions for n heterozygous loci = 512

54

EM algorithm

- Pros
  - More accurate than Clark's method
  - Fully or partially phased individuals contribute most of the information
- Cons
  - Estimate depends on starting point: need to run multiple times on different starting points
  - Implementation may become computationally expensive: cost grows rapidly with number of markers
    - For each individual, the number of possible haplotypes is 2^m, where m is the number of makers
    - Typically run for short sequences with < 25 SNPs
  - No modeling on haplotypes

55

Outline

- Background & motivation
- Problem statement
- Statistical methods for haplotype inference
  - Clark's algorithm
  - Expectation Maximization (EM) algorithm
  - Coalescent-based methods and HMM
  - Haplotype inference on sequence data
- Example applications

56