Genetic Linkage Analysis

Lectures 8 - Oct 24, 2011
CSE 527 Computational Biology, Fall 2011
Instructor: Su-In Lee
TA: Christopher Miles
Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

Outline

- Review: disease association studies
  - Association vs linkage analysis

- Genetic linkage analysis
  - Pedigree-based gene mapping
  - Elston-Stewart algorithm

- Systems biology basics
  - Gene regulatory network
Genome-Wide Association Studies

- Any disadvantages?
  - Hypothesis-free: we search the entire genome for associations rather than focusing on small candidate areas.
  - The need for extremely dense searches.
  - The massive number of statistical tests performed presents a potential for false-positive results (multiple hypothesis testing)

Alternative strategy – Linkage analysis
  - It acts as systematic studies of variation, without needing to genotype at each region.
  - Focus on a family or families.
Basic Ideas

- Neighboring genes on the chromosome have a tendency to stick together when passed on to offspring.

- Therefore, if some disease is often passed to offspring along with specific marker-genes, we can conclude that the gene(s) responsible for the disease are located close on the chromosome to these markers.

Outline

- Review: disease association studies
  - Association vs linkage analysis

- Genetic linkage analysis
  - Pedigree-based gene mapping
  - Elston-Stewart algorithm

- Systems biology basics
  - Gene expression data
  - Gene regulatory network
Genetic linkage analysis

- Data
  - Pedigree: set of individuals of known relationship
  - Observed marker genotypes
  - Phenotype data for individuals

- Genetic linkage analysis
  - Goal – Relate sharing of specific chromosomal regions to phenotypic similarity
  - Parametric methods define explicit relationship between phenotypic and genetic similarity
  - Non-parametric methods test for increased sharing among affected individuals

Reading a Pedigree

- Circles are female, squares are males
- Shaded symbols are affected, half-shaded are carriers
- What is the probability to observe a certain pedigree?
Elements of Pedigree Likelihood

- Prior probabilities
  - For founder genotypes

- Transmission probabilities
  - For offspring genotypes, given parents

- Penetrances
  - For individual phenotypes, given genotype

Probabilistic model for a pedigree:
(1) Founder (prior) probabilities

- **Founders** are individuals whose parents are not in the pedigree
  - They may or may not be typed. Either way, we need to assign probabilities to their actual or possible genotypes.
  - This is usually done by assuming Hardy-Weinberg equilibrium (HWE). If the frequency of D is .01, HW says

\[
P(\text{father Dd}) = 2 \times .01 \times .99
\]

- Genotypes of founder couples are (usually) treated as independent.

\[
P(\text{father Dd, mother dd}) = (2 \times .01 \times .99) \times (.99)^2
\]
Probabilistic model for a pedigree: (2) Transmission probabilities I

- According to Mendel's laws, children get their genes from their parents’ genes independently:

\[
\begin{align*}
&\text{Dd} & 1 & \text{Dd} & 2 \\
&3 & \text{dd} \\
\end{align*}
\]

\[P(\text{children 3 dd | father Dd, mother dd}) = \frac{1}{2} \times \frac{1}{2}\]

- The inheritances are independent for different children.

Probabilistic model for a pedigree: (2) Transmission probabilities II

\[
\begin{align*}
&\text{Dd} & 1 & \text{Dd} & 2 \\
&3 & \text{dd} & 4 & \text{Dd} & 5 & \text{DD} \\
\end{align*}
\]

\[P(\text{3 dd, 4 Dd, 5DD | 1 Dd, 2 dd}) = \left(\frac{1}{2} \times \frac{1}{2}\right) \times \left(2 \times \frac{1}{2} \times \frac{1}{2}\right) \times \left(\frac{1}{2} \times \frac{1}{2}\right)\]

- The factor 2 comes from summing over the two mutually exclusive and equiprobable ways 4 get a D and a d.
Probabilistic model for a pedigree: (3) Penetrance probabilities I

- Independent penetrance model
  - Pedigree analyses usually suppose that, given the genotype at all loci, and in some cases age and sex, the chance of having a particular phenotype depends only on genotype at one locus, and is independent of all other factors: genotypes at other loci, environment, genotypes and phenotypes of relative, etc

- Complete penetrance
  - \( DD \)
  - \( P(\text{affected} \mid DD) = 1 \)

- Incomplete penetrance
  - \( DD \)
  - \( P(\text{affected} \mid DD) = .8 \)

Probabilistic model for a pedigree: (3) Penetrance probabilities II

- Age & sex-dependent penetrance
  - \( DD \) (45)
  - \( P(\text{affected} \mid DD, \text{male}, 45 \text{ y.o.}) = .6 \)
Probabilistic model for a pedigree: Putting all together

- **Assumptions**
  - Penetrance probabilities:
    - \( P(\text{affected} \mid \text{dd}) = 0.1 \), \( P(\text{affected} \mid \text{Dd}) = 0.3 \), \( P(\text{affected} \mid \text{DD}) = 0.8 \)
  - Allele frequency of D is \( 0.01 \)

- The probability of this pedigree is the product:
  - \[(2 \times 0.01 \times 0.99 \times 0.7) \times (2 \times 0.01 \times 0.99 \times 0.3) \times (\frac{1}{2} \times \frac{1}{2} \times 0.9) \times (2 \times \frac{1}{2} \times \frac{1}{2} \times 0.7) \times (\frac{1}{2} \times \frac{1}{2} \times 0.8)\]

Elements of pedigree likelihood

- **Prior probabilities**
  - For founder genotypes e.g. \( P(g_1), P(g_2) \)

- **Transmission probabilities**
  - For offspring genotypes, given parents e.g. \( P(g_4 \mid g_1, g_2) \)

- **Penetrance**
  - For individual phenotypes, given genotype e.g. \( P(x_1 \mid g_1) \)
Elements of pedigree likelihood

- Overall pedigree likelihood
  \[ L = \prod_{f = \text{founders}} P(G_f) \prod_{o,f,m} P(G_o \mid G_f, G_m) \prod_{i = \text{individuals}} P(X_i \mid G_i) \]

  - Probability of founder genotypes
  - Probability of offspring given parents
  - Probability of phenotypes given genotypes

Probabilistic model for a pedigree: Putting all together II

- To write the likelihood of a pedigree given complete data:
  \[ L_c = \prod_{f = \text{founders}} P(G_f) \prod_{o,f,m} P(G_o \mid G_f, G_m) \prod_{i = \text{individuals}} P(X_i \mid G_i) \]

  - We begin by multiplying founder gene frequencies, followed by transmission probabilities of non-founders given their parents, next penetrance probabilities of all the individuals given their genotypes.

- What if there are missing or incomplete data?
  - We must sum over all mutually exclusive possibilities compatible with the observed data.
  \[ L = \sum_{G_1} \cdots \sum_{G_2} \prod_{f = \text{founders}} P(G_f) \prod_{o,f,m} P(G_o \mid G_f, G_m) \prod_{i = \text{individuals}} P(X_i \mid G_i) \]

  - All possible genotypes of individual 1
  - If the individual i’s genotype is known to be \( g_i \), then \( G_i = \{g_i\} \)
Probabilistic model for a pedigree: Putting all together II

What if there are missing or incomplete data?

- We must sum over all mutually exclusive possibilities compatible with the observed data.

$$L = \sum_{g_i=(DD, Dd, dd)} P(G_i = g_1, G_2 = Dd, G_3 = dd, G_4 = Dd, G_5 = DD)$$

Computationally ...

- To write the likelihood of a pedigree:

$$L = \sum_{G_i} \sum_{G_o} \prod_{f=\text{founders}} P(G_f) \prod_{\{o,f,m\}} P(G_o \mid G_f, G_m) \prod_{i=\text{individuals}} P(X_i \mid G_i)$$

- Computation rises exponentially with # people $n$.
- Computation rises exponentially with # markers.
- Challenge is summation over all possible genotypes (or haplotypes) for each individual.
Computationally …

- Two algorithms:
  - The general strategy of beginning with founders, then non-founders, and multiplying and summing as appropriate, has been codified in what is known as the Elston-Stewart algorithm for calculating probabilities over pedigrees.
  - It is one of the two widely used approaches. The other is termed the Lander-Green algorithm and takes a quite different approach.

Elston and Stewart’s insight…

- Focus on “special pedigree” where
  - Every person is either
    - Related to someone in the previous generation
    - Marrying into the pedigree
  - No consanguineous marriages

- Process nuclear families, by fixing the genotype for one parent
  - Conditional on parental genotypes, offsprings are independent
Elston and Stewart’s insight...

- Conditional on parental genotypes, offsprings are independent
- Thus, avoid nested sums, and produce likelihood whose cost increases linearly with the number of offspring

\[
L = \sum_{G_n} \sum_{G_j} \cdots \sum_{G_{o1}} P(X_m | G_n)P(G_n | G_{o1})P(X_j | G_j)P(G_j) \prod_{o \in \text{one}} P(X_o | G_o)P(G_o | G_n, G_j)
\]

\[
= \sum_{G_n} P(X_m | G_n)P(G_n | G_{o1})\sum_{G_j} P(X_j | G_j)P(G_j) \prod_{o \in \text{one}} P(X_o | G_o)P(G_o | G_n, G_j)
\]

Successive Conditional Probabilities

- Starting at the bottom of the pedigree...
- Calculate conditional probabilities by fixing genotypes for one parent
- Specifically, calculate \( H_k (G_k) \)
  - Probability of descendants and spouse for person \( k \)
  - Conditional on a particular genotype \( G_k \)
Formulae …

- So for each parent, calculate
  \[ H_{parent}(G_{parent}) = \sum_{G_{spouse}} P(X_{spouse} | G_{spouse}) P(G_{spouse}) \]
  \[ \prod_o \sum_{G_o} P(X_o | G_o) P(G_o | G_{parent} G_{spouse}) H_o(G_o) \]

- By convention, for individuals with no descendants
  \[ H_{leaf}(G_{leaf}) = 1 \]

Final likelihood

- After processing all nuclear family units

Simple sum gives the overall pedigree likelihood

\[ L = \sum_{G_{founder}} P(X_{founder} | G_{founder}) P(G_{founder}) H_{founder}(G_{founder}) \]

\[ L = \sum_{G_1} \ldots \sum_{G_i} \prod_{\ell \text{-founders}} P(G_{\ell}) \prod_{(o,m)} P(G_o | G_f, G_m) \prod_{i \text{-individuals}} P(X_i | G_i) \]

\[ = \sum_{G_{founder}} P(G_{founder}) P(X_{founder} | G_{founder}) \sum_{G_{nonfounder}} \prod_{(o,m)} P(G_o | G_f, G_m) \prod_{i \text{-nonfounders}} P(X_i | G_i) \]

\[ P(X, \text{given genotypes} | G_{founder}) = H_{founder}(G_{founder}) \]
What next?

- Computation of the pedigree likelihood

  For every marker, we want to

  - Compute the pedigree likelihood for each marker and choose the marker that is closely linked to the disease gene.

Further Reading

- Part I
Outline

- Review: disease association studies
  - Association vs linkage analysis

- Genetic linkage analysis
  - Pedigree-based gene mapping
  - Elston-Stewart algorithm

- Systems biology basics
  - Review: gene regulation
  - Gene expression data
  - Gene regulatory network

Review: Gene Regulation

```
AGATATGGATTGTTAGGATTTATGCGCGTCAGTGACTACGCATGTTACGCACCTACGACTAGGTAATGATTGATC
```

```
DNA
AGATATGGATTGTTAGGATTTATGCGCGTCAGTGACTACGCATGTTACGCACCTACGACTAGGTAATGATTGATC
```

```
RNA
AUGUGGAUUGUU
```

```
Protein
```

```
AUGCGCGUC
```

```
AUGUUACGCACCUAC
```

```
AUGAUUGAU
```

```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
Gene expression data

- **Down-regulated**
- **诱导**
- **Repressed**

Experiments (samples) $^i$

$E_{ij}$ - RNA level of gene $j$ in experiment $i$

Co-expression genes? ⇒ functionally related?

Goal: Inferring regulatory networks

- Infer the regulatory network that controls gene expression
- Causality relationships among $e_{1:Q}$
  - $A$ and $B$ regulate the expression of $C$
    (A and B are regulators of C)
- Bayesian networks

$q \approx 2 \times 10^4$ (for human)
Clustering expression profiles

Hierarchical agglomerative

- Compute all pairwise distances
- Merge closest pair
Clustering expression profiles

Limitations:

- No explanation on what caused expression of each gene
- (No regulatory mechanism)

Co-regulated genes cluster together

Infer gene function

Goal: Inferring regulatory networks

Infer the regulatory network that controls gene expression
- Causality relationships among $e_1, Q$

$A$ and $B$ regulate the expression of $C$
($A$ and $B$ are regulators of $C$)

Bayesian networks