Cookbook Linkage Analysis, Ab Initio

Alejandro A. Schäffer
NCBI

e-mail: schaffer@helix.nih.gov

Customer service through CIT/DCB:

Ed Suh: suh@helix.nih.gov

Jim Tomlin: jtomlin@helix.nih.gov

Carla Bock: bockc@helix.nih.gov

http://bimas.cit.nih.gov/linkage/index.html

Recipe for disease gene hunting

1. Identify 1 or more sets of related patients
2. Connect patients into pedigrees
3. Genotype patients and their families
4. Look for polymorphisms that cosegregate with the disease
5. When linkage is found, narrow the region with more families
6. Once the linkage region is small, either find candidate genes and skip to step 9, or clone DNA from the area and put into artificial hosts (vectors).

7. Sequence the DNA in the region

8. Look for genes (e.g., by computer prediction, exon trapping, cDNA screen, etc.)

9. Screen genes for mutations

10. Prove mutations are causal (proof standards vary)

Ingredients for Cookbook Linkage Analysis

- 1 or more family trees (pedigrees) exhibiting a recurring disease (phenotype)

- Genotypes for as many individuals at as many loci as feasible

- 1 or more UNIX, VMS, or DOS computer(s)

- LINKAGE/FASTLINK and/or VITESSE software packages

- 1/2 cup of statistical theory

I will also briefly discuss GENEHUNTER, SimIBD software.

Desired Results

- Approximate location of disease gene

- Placement of disease gene relative to multiple other loci

- Exclusion of a genome region for containing the disease gene

Caution: All conclusions are obtained with high probability, but not certainty

Goal of this lecture:

Reduce easy cases of linkage analysis to:

1. Family Ascertainment
2. S.M.O.G. (Simple Matter of Genotyping)
3. Looking up instructions
4. File manipulation and keystroking

References:

TO: J. D. Terwilliger and J. Ott, Handbook of Human Genetic Linkage, Johns Hopkins Press, 1994.

JO: Analysis of Human Genetic Linkage, Johns Hopkins Press, 1991

LD: LINKAGE Documentation

FD: FASTLINK Documentation

AD: APM Documentation

CD: GenoCheck Documentation

GD: GENEHUNTER Documentation

SD: SimIDB Documentation

VD: VITESSE Documentation

Linkage analysis software packages

Parametric:

LINKAGE/FASTLINK/GenoCheck , VITESSE

LIPED, MAPMAKER, CRI-MAP

MENDEL, PAP, S.A.G.E.

Non-parametric:

APM,S.A.G.E.,GENEHUNTER,SimIBD

[S.A.G.E. sages: Joan Bailey-Wilson and Alexander Wilson, at NHGRI.]

Parametric means specify model of inheritance

many disease studies still use LINKAGE

Reason: TO and associated mini-courses

Recombinant

Suppose loci A and B are conjectured to be on the same chromosome.

Individual i is a recombinant w.r.t. parent p, if i inherits the A allele from one chromatid of p and the B allele from the other chromatid of p.

Recombination Fraction: Θ

Probability that a child is recombinant. between loci A and B is Θ_{AB}.

Low prob. recombinant $\rightarrow$ linkage

For ordered multiple loci:

$$A \;\; B \;\; C \;\; D$$

define a **vector** $[\Theta_1, \Theta_2, \Theta_3]$

which is an ordered list of recombination fractions between adjacent loci

Recombination fractions vary for male/female parents, but published maps tend to be gender-averaged

Recombination Fraction vs. Genetic Distance

Genetic Distance x_{AB} is the expected number of crossovers between locus A and locus B.

If A and B are on different chromosomes, $\Theta_{AB} = 0.5$, but x_{AB} is not defined.

$\Theta_{AB} \leq 0.5$, while x_{AB} may be larger than 1.

On the same chromosome, $\Theta_{AB} < x_{AB}$ and they are close for small values.

$x_{AB} + x_{BC} = x_{AC}$

$\Theta_{AB} + \Theta_{BC} \neq \Theta_{AC}$

x is measured in Morgans, Θ is unitless

1 Morgan means 1 expected crossover

Relationship between Morgans and Megabases is poorly understood.

Mapping Functions

Mapping functions convert from Θ to x and back.

Canonical application: compute Θ_{AC} knowing Θ_{AB} and Θ_{BC}

1. Apply m.f. to get x_{AB} and x_{BC}

2. Compute $x_{AC} = x_{AB} + x_{BC}$

3. Apply m.f. to get Θ_{AC}

Published maps show Θ between adjacent markers, LINKAGE needs Θ, but you may not want to use loci adjacent on map.

JO: Chapter 1 and FD: README.mapfun

Essence of LINKAGE:

Given a fixed order of markers and disease loci, e.g.:

$$M_1\ D\ M_2\ M_3$$

and a candidate Θ vector,

$$[\theta_1, \theta_2, \theta_3]$$

compute a number called the "likelihood" to quantify:

How well does the **data** fit the **theta**?

Higher likelihood is better.

What is FASTLINK?

New versions of LINKAGE main programs

- Faster sequentially
- Run in parallel
- Robust against crashes and user errors
- Substantial new documentation

What is VITESSE?

Program to replace MLINK and LINKMAP

- Faster than FASTLINK
- New running time diagnostics
- "Simple" pedigrees only

"Simple" means each person has $<=$ 2 grand-parents in pedigree, and no loops.

J. Tomlin has a utility program to combine FASTLINK (LINKMAP) and VITESSE output for mixed-simple data sets.

Reference: VD

Maximum Likelihood Estimation

Compare two candidates: Θ_0, Θ_1

H_0: Θ_0 is the correct value

H_1: Θ_1 is the correct value

Compute :

$$\text{Like}(\Theta_1)/\text{Like}(\Theta_0)$$

If this ratio is above a prescribed threshold, then accept H_1, reject H_0; if below a different threshold, accept H_0, reject H_1.

Thresholds depend on the bad consequences to society from false negatives and false positives.

Testing for linkage

Unlinked means that θ between the disease and all other markers is 0.5

Linked means that θ between the disease and some other marker is (substantially) < 0.5. Use the best choice for hypothesis H_1.

Standard threshold for autosomal linkage: 1000

Standard threshold for autosomal nonlinkage: 1/100

Reference: JO, Sec. 4.4

+ +

What's a lod score?

Let L_1 be pedigree likelihood when disease is linked

Let L_0 be pedigree likelihood when disease is unlinked

The **lod score** is $\log_{10}(L_1/L_0)$

Lod score of magnitude $> z$ mathematically implies the statistical probability of error is $< 10^{-z}$, ignoring multiple testing issues.

Caution: What LINKAGE/FASTLINK prints for log(L) may be off by an additive term, but lod scores are correct.

Logarithms

Why do linkage analysis experts talk about lod scores rather than likelihoods?

Logarithms convert multiplication/division into addition/subtraction. This is very useful for multiple data sets.

$x = b^y \leftrightarrow y = \log_b(x)$

$\log(x * y) = \log(x) + \log(y)$

$\log(x/y) = \log(x) - \log(y)$

So, the lod score can be computed as :

$\log_{10}(L_1) - \log_{10}(L_0)$

Stages of a generic linkage study

1. Genome scan to find possible regions

2. Identify the correct region

3. Pin disease gene between two known genes (Fine Mapping/Testing for Order)

Genome Scan

Use 150-400 polymorphic probes at 8–20cM spacing.

Always do two locus analysis between disease and marker probe of known position.

Interested in lodscores above 1 or 2.

Standard marker sets suggested by CHLC.

Two-stage scans can reduce lab costs.

Never, never compare in print the lod score with marker A against the lodscore with marker B

Identify the Correct Region

Use multiple probes at 1–4cM spacing.

Do two or three locus analysis between disease and marker probes of known position.

Interested in lodscores above 3.

Can find suitable markers from a combination of the Marshfield, MIT, and Généthon maps.

Use markers whose relative order is agreed upon.

Testing for marker order

Fixed marker map: $M_1 M_2 M_3$

Find best Θ vector for each of 4 placements of disease D.

Let the best likelihood for each order be L_1, L_2, L_3,L_4.

Compare differences between $\log(L_1)$, $\log(L_2)$, $\log(L_3)$, $\log(L_4)$

If $\log(L_i) - \log(L_j) = t > 0$ then the odds that j is incorrect in favor of i, are: $10^t : 1$

Desired level for t: 3.0

Marker order cautions

May have very high lodscores, yet uncertain marker order

Published maps may be incorrect over small regions

Published Θ values are estimated from small samples

Limited variation among samples

Association between marker alleles and disease phenotypes (linkage disequilibrium) can confound the computations

LINKAGE/FASTLINK main programs:

Differ in: How thetas are chosen, output format

MLINK: Used to vary 1 θ at a time

LINKMAP: Used to vary 1 θ or 2 adjacent θ at a time

ILINK: Used to find a locally optimal Θ vector.

Locally optimal means that it cannot be perturbed by a small amount to get a different vector with a better likelihood.

VITESSE coalesces LINKMAP and MLINK by taking the LINKAGE program name as a parameter.

When to use MLINK

You want a table of 2-point lodscores comparing disease against a variety of markers and for a variety of theta values.

You need to fill in the lodscore at specific theta vectors.

Used for genome scan.

Caution: Avoid publishing positive results based solely on 2-point MLINK analysis. This error is common in the literature. It used to be excusable due to running time limitations.

When to use LINKMAP

You want to draw a multipoint lodscore plot.

You want to find in which gap of a fixed marker order the disease gene lies.

Major Caution: LINKMAP gives relatively little information about whether the disease may be near but outside the set of fixed markers.

Note: LINKMAP will not draw anything, but it will yield a table of values that can be plotted.

When to use ILINK

Whenever you have time to wait

When you do not feel the published thetas are valid for your data set

When you wish to estimate allele frequencies (too advanced for this course)

In the context of GenoCheck to find genotyping errors.

Major Caution: The output of ILINK is hard to understand. See README.ILINK in FD.

ILINK avoids reliance on published distances and may lead to stronger results.

Disease Placement by Recombinants

Suppose the fixed map is

$$M_1 \; M_2 \; M_3$$

Suppose we established linkage between D and this region.

Suppose that there are D/M_1 and D/M_3 recombinants.

Conclude that D is between M_1 and M_3.

This method of inference is incomparable in power to the likelihood method.

Allele Frequencies

Published by CEPH/Généthon for many markers using a reference panel of families. See list of FTP/WWW sites.

Results are quite sensitive to allele frequencies

Allele frequencies can be estimated from input data.

Caution: All families in reference panel are from France and Utah.

Reference: TO, Chapter 22; Jim Tomlin's allele frequency estimation recipe

Genotype and Phenotype

Dominant, Recessive, Co-dominant

Notation:

DD: Homozygous for disease

Dd: Heterozygous for disease

dd: Homozygous for non-disease

Penetrance, f: Probability that an individual of a given genotype is affected

In practice you may know affection status, but not genotype

Penetrance examples

Dominant:

DD: 1.0 Dd: 1.0 dd: 0.0

Recessive:

DD: 1.0 Dd: 0.0 dd: 0.0

Pure Codominant:

DD: 1.0 Dd: $0 < f < 1$ dd: 0.0

Must See: TO, Chapter 9

Penetrance can be used to **model**:

- Age-of-onset effect
- Gender effect
- Subset of symptoms
- Diagnostic uncertainty
- Errors in data
- Environment

LINKAGE auxiliary programs

makeped: converts initial pedigree file into properly formatted file

preplink: interactively defines locus descriptions

lcp: prepares a script for execution of a sequence of main programs

lsp: used within script to extract desired loci from pedigree file

unknown: infers missing genotypes

lrp: converts the results of LINKMAP into a nice table suitable for multipoint graphing

Important: Use the FASTLINK version of unknown

LINKAGE/FASTLINK Troubleshooting

"Nous vous conseillons de surveiller votre mail pendant les quelques premieres minutes apres le lancement. En effet si pour une raison quelconque (et elles sont nombreuses!) votre calcul 'crashe' vous n'attendrez pas inutilement avant d'y remedier."

Lucien Bachner (Infobiogen, Paris)

Reference: README.trouble

PEDCHECK is a useful program to identify "inconsistent" genotypes. E.g., a child has allele 4, but neither parent does.

Mendelian autosomal inheritance

Each person has two alleles at each locus.

A child inherits one allele from the father and one allele from the mother.

The set of sibling alleles S can be written so that

$$S = M \cup F$$

in such a way that M and F are of size 1 or 2 and each sibling has 1 allele in M and 1 allele in F.

Most linkage programs check these rules.

Mendelian X inheritance

A female has two alleles at a locus.

A male has one allele at a locus.

A female inherits her father's allele and one of her mother's alleles.

A male inherits one of his mother's alleles

The set of sibling alleles S can be written so that:

$$S = M \cup F$$

in such a way that M is of size 1, F is of size 1 or 2, each male sibling has 1 allele in F, each female sibling has 1 allele in M and 1 allele in F.

Checking Genotypes Further

Genotypes can be wrong, but still consistent with Mendelian rules.

Wrong genotypes usually increase the estimate of θ.

One can estimate θ between markers using ILINK.

Estimated $\theta >>$ published θ is a good clue that there are errors.

Use GenoCheck (derived from ILINK) to identify the most suspect person/locus pairs.

Convince laboratory personnel to follow-up based on GenoCheck output.

Reference: CD

Checking the status of a run

On UNIX: Use ps (with flags) and look for a process running one of: ilink, linkmap, mlink.

Make sure that process status is R

For FASTLINK

Each likelihood evaluation takes roughly the same amount of time.

Programs take a checkpoint after every likelihood evaluation.

Use: ls -lt check*

to see when last checkpoint was taken.

For VITESSE

Use parental pairs number

Reference: FD: README.time

Non-parametric analysis

Affected relatives have disease for the same reason

Basic method is affected sibpairs

Newer software considers affected relative pairs

Useful when mode of inheritance is unclear

APM, GENEHUNTER, SimIBD

Identical By State vs. Identical By Descent

Reference: TO, Chapter 26

APM

A_{ij} = probability that i,j share an allele IBS

A = Sum of A_{ij} over all affected i,j pairs

Can also weight A by allele frequency.

Compare observed A to expected.

Very sensitive to allele frequencies.

Empirical p-values.

GENEHUNTER

General approach using inheritance vectors.

Inher. vector: each allele from grandpa/grandma?

Given a probability distribution of inheritance vectors, conditional on phenotypes, compare this to uniform distribution.

Two different statistics: NPL_{pairs}, NPL_{all}

NPL_{pairs} somewhat analogous to A, but using IBD.

Obtain a normally distributed Z-statistic for p-values.

Reference: GD

SimIBD

A different statistic: SimIBD

Improves APM by conditioning simulation on affection status.

Improves APM by moving away from IBS to IBD.

You get empirical p-values.

Currently does only 2-point analysis

Reference: SD

Filing a bug report

A run of a program is an experiment.

An effective bug report enables the software developer to exactly reproduce the experiment without bias as to the outcome.

Use e-mail, not phone or FAX, but beware that mailers can mangle files.

Send files as ASCII, not BinHex

State what problems occurred, but describe only effects (phenotypes) that you observe, not causes (genotypes) that you conjecture.

Reference: FD: README.bugreport

Refereeing a linkage analysis paper

Look for:

- List of marker loci used
- Description and justification of penetrance function
- Allele frequency for disease
- Method for obtaining marker allele frequencies

Look for:

- 2-point (disease with each locus) LOD score table

- Multipoint analyses strengthening 2-point results

- Explanation of what software was used

These 7 items are **not sufficient**, but ought to be **necessary** for a "parametric" analysis. If any of the 7 items is missing, do not recommend the manuscript be accepted.

Installation Options:

1. Invite me to make a “field trip” to your lab.

2. Use galaxy.nih.gov and contact Jim Tomlin (jtomlin@helix.nih.gov)

3. Retrieve via ftp and install yourself

ftp: login as anonymous, leave full e-mail as password

Useful FTP/WWW/Email addresses:

Local help and pointers:

http://bimas.cit.nih.gov/linkage/index.html

Software locally available on galaxy.nih.gov. Accessible with an account on helix.nih.gov and a request.

http://bimas.cit.nih.gov/linkage/galaxy1.html

All Software Catalog:

http://linkage.rockefeller.edu

Mailing List: fastlink-list@fastlink.nih.gov

APM:

ftp watson.hgen.pitt.edu

under pub/apm

FASTLINK:

ftp fastlink.nih.gov

under pub/fastlink

GENEHUNTER: (1 URL split into 2 lines)

http://waldo.wi.mit.edu/ftp/

distribution/software/genehunter/gh2

GENOCHECK:

ftp softlib.cs.rice.edu

under pub/GenoCheck

LINKAGE for DOS and utilities:

ftp linkage.rockefeller.edu

under software/linkage/DOS

under software/utilities

under software/linkage/UNIX

and other subdirectories also

PEDCHECK:

ftp watson.hgen.pitt.edu

under pub/pedcheck

SimIBD:

ftp watson.hgen.pitt.edu

under pub/simibd

VITESSE:

ftp watson.hgen.pitt.edu

under pub/vitesse

Généthon allele frequencies:

ftp://ftp.genethon.fr/pub/Gmap/Nature-1995

Marshfield maps

http://research.marshfieldclinic.org/genetics/

follow various links to find markers or build maps

MIT map

http://www.genome.wi.mit.edu/

click on Human Physical Mapping Project