

SNPHAP 2007, January 27, 2007

Design and Validation of Methods Searching for Risk Factors in Genotype Case- Control Studies

Dumitru Brinza
Alexander Zelikovsky

Department of Computer Science
Georgia State University



Outline

- SNP, Haplotypes and Genotypes
- Heritable Common Complex Diseases
- Disease Association Search in Case-Control Studies
- Addressing Challenges in DA
- Risk Factor Validation for Reproducibility
- Atomic risk factors/Multi-SNP Combinations
- Maximum Odds Ratio Atomic RF
- Approximate vs Exhaustive Searches
- Datasets/Results
- Conclusions / Related & Future Work

Human Genome – all the genetic material in the chromosomes,
length 3×10^9 base pairs

Difference between any two people occur in 0.1% of genome

SNP – single nucleotide polymorphism site where two or more different nucleotides occur in a large percentage of population.

Diploid – two different copies of each chromosome

Haplotype – description of a single copy (expensive)
example: 00110101 (0 is for major, 1 is for minor allele)

Genotype – description of the mixed two copies
example: 01122110 (0=00, 1=11, 2=01)

n **Complex disease**

- .. Interaction of multiple genes
 - n One mutation does not cause disease
 - n Breakage of all compensatory pathways cause disease
 - n Hard to analyze - 2-gene interaction analysis for a genome-wide scan with 1 million SNPs has 10^{12} pair wise tests
- .. Multiple independent causes
 - n There are different causes and each of these causes can be result of interaction of several genes
 - n Each cause explains certain percentage of cases

n **Common diseases are Complex: > 0.1%.**

- n In NY city, 12% of the population has Type 2 Diabetes

DA Search in Case/Control Study

Given: a population of n genotypes each containing values of m SNPs and disease status

| | SNPs | Disease Status |
|---------------------------|------------------|----------------|
| Case genotypes: | 0101201020102210 | -1 |
| | 0220110210120021 | -1 |
| | 0200120012221110 | -1 |
| | 0020011002212101 | -1 |
| Control genotypes: | 1101202020100110 | 1 |
| | 0120120010100011 | 1 |
| | 0210220002021112 | 1 |
| | 0021011000212120 | 1 |

Find: risk factors (RF) with significantly high odds ratio i.e., pattern/dihaplotype significantly more frequent among cases than among controls

n Computational

- Interaction of multiple genes/SNP's
 - n Too many possibilities – obviously intractable
- Multiple independent causes
 - n Each RF may explain only small portion of case-control study

n Statistical/Reproducing

- Search space / number of possible RF's
 - n Adjust to multiple testing
- Searching engine complexity
 - n Adjust to multiple methods / search complexity

n **Computational**

- Constraint model / reduce search space
 - n Negative effect = **may miss “true” RF’s L**
- **Heuristic search J**
 - n Look for “easy to find” RF’s
 - n May miss only “maliciously hidden” true RF

n **Statistical/Reproducing**

- Validate on different case-control study
 - n That’s obvious but **expensive L**
- **Cross-validate in the same study J**
 - n Usual method for prediction validation

Significance of Risk Factors

n Relative risk (RR) – cohort study

$$RR = \frac{d \cdot (H + D - h - d)}{(D - d)(h + d)}$$

n Odds ratio (OR) – case-control study

$$OR = \frac{d \cdot (H - h)}{h \cdot (D - d)}$$

n P-value

- binomial distribution

$$p = \sum_{k=0}^d \binom{h + d}{k} \left(\frac{D}{H + D} \right)^k \left(\frac{H}{H + D} \right)^{h + d - k}$$

- Searching for risk factors among many SNPs requires multiple testing adjustment of the p-value

n **Multiple-testing adjustment**

- **Bonferroni**

- n *easy to compute*

- n *overly conservative*

- **Randomization**

- n *computationally expensive*

- n *more accurate*

n **Validation rate using Cross-Validation**

- **Leave-One-Out**

- **Leave-Many-Out**

- **Leave-Half-Out**

Atomic Risk Factors, MSCs and Clusters

- Genotype SNP = Boolean function over 2 haplotype SNPs
 - 0 iff $g_0 = (x \text{ NOR } y)$ is TRUE
 - 1 iff $g_1 = (x \text{ AND } y)$ is TRUE
 - 2 iff $g_2 = (x \text{ XOR } y)$ is TRUE
- Single-SNP risk factor = Boolean formula over g_0 , g_1 and g_2
- Complex risk factor (RF) = CNF over single-SNP RF's:
$$g_0^1 (g_0 + g_2)^2 (g_1 + g_2)^3 g_0^5$$
- **Atomic risk factor (ARF)** = unsplittable complex RF's:
$$g_0^1 g_2^2 g_1^3 g_0^5$$
 - n single disease-associated factor
- ARF \leftrightarrow multi-SNP combination (MSC)
 - n **MSC** = subset of SNP with fixed values of SNPs, 0, 1, or 2
- **Cluster**= subset of genotypes with the same MSC

- n **Maximum Odds Ratio Atomic Risk Factor**
 - .. **Given:** genotype case-control study
 - .. **Find:** ARF with the maximum odds ratio

- n Clusters with less controls have higher OR
=> MORARF includes finding of max control-free cluster

- n MORARF contains max independent set problem
=> No provably good search for general case-control study

- n Case-control studies do not bother to hide true RF
=> Even simple heuristics may work



Requirements to Approximate search

- **Fast**
 - n longer search needs more adjustment
- **Non-trivial**
 - n exhaustive search is slow
- **Simple**
 - n Occam's razor

Exhaustive Searching Approaches

n Exhaustive search (ES)

- For n genotypes with m SNPs there are $O(n^{km})$ k -SNP MSCs

n Exhaustive Combinatorial Search (CS)

- Drop small (insignificant) clusters
- Search only plausible/maximal MSC's

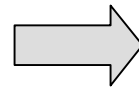
Case-closure of MSC:

- MSC extended with common SNPs values in all cases
- Minimum cluster with the same set of cases

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---------|
| 0 | 1 | 1 | 0 | 1 | 2 | 1 | 0 | 2 | case |
| 2 | 0 | 1 | 1 | 0 | 2 | 0 | 0 | 1 | case |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | case |
| 0 | 1 | 1 | 0 | 1 | 2 | 0 | 0 | 2 | control |
| 0 | 1 | 1 | 0 | 1 | 2 | 0 | 0 | 2 | control |

x x **1** x x **2** x x x
 Present in 2 cases : 2 controls

Case-closure



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---------|
| 0 | 1 | 1 | 0 | 1 | 2 | 1 | 0 | 2 | case |
| 2 | 0 | 1 | 1 | 0 | 2 | 0 | 0 | 1 | case |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | case |
| 0 | 1 | 1 | 0 | 1 | 2 | 0 | 0 | 2 | control |
| 0 | 2 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | control |

x x **1** x x **2** x **0** x
 Present in 2 cases : 1 control

n Combinatorial Search Method (CS):

- Searches only among case-closed MSCs
- Avoids checking of clusters with small number of cases
- Finds significant MSCs faster than ES
- Still too slow for large data
- Further speedup by reducing number of SNPs

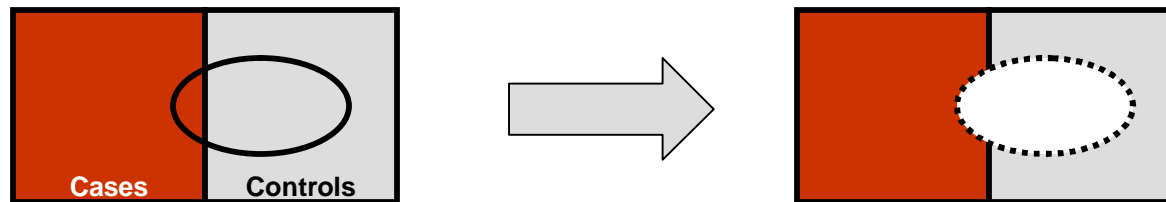
Complimentary Greedy Search (CGS)

n Intuition:

- .. Max OR when no controls – chosen cases do not have simila
- .. Max independent set by removing highest degree vertices

n Fixing an SNP-value

- .. Removes controls **J** -> profit
- .. Removes cases **L** -> expense



n Maximize profit/expense!

n Algorithm:

- .. Starting with empty MSC add SNP-value removing from current cluster max # controls per case

n Extremely fast but inaccurate, trapped in local maximum

AcS – alternating combinatorial search method

$$C \leftarrow \overline{(C - \overline{(C \cap S_0)})}$$

RCGS – Randomized complimentary greedy search method

For $k = 1, \dots, 100$

Randomly permute controls S_0

$C \leftarrow S_1$

For $i = 2, \dots, \ln H$

$C \leftarrow d(C) \cup (H_{\lceil e^i \rceil} \cap h(C^*))$, where C^* is a cluster defined by MSC in the entire S

Repeat until $h(C) > \frac{1}{3}e^i$ or $(h(C) > 0$ and $e^i > H$)

Find 1-SNP combination $X = (s, i)$, where s is a SNP and $i \in \{0, 1, 2\}$

minimizing $(d(C) - d(C \cap X)) / (h(C) - h(C \cap X))$

$C \leftarrow C \cap X$

Update the so far best MSC

5 Data Sets

- n **Crohn's disease (Daly et al)**: inflammatory bowel disease (IBD).

Location: 5q31

Number of SNPs: 103

Population Size: 387

case: 144 control: 243

- n **Autoimmune disorders (Ueda et al)** :

Location: containing gene CD28, CTLA4 and ICONS

Number of SNPs: 108

Population Size: 1024

case: 378 control: 646

- n **Tick-borne encephalitis (Barkash et al)** :

Location: containing gene TLR3, PKR, OAS1, OAS2, and OAS3.

Number of SNPs: 41

Population Size: 75

case: 21 control: 54

- n **Lung cancer (Dragani et al)** :

Number of SNPs: 141

Population Size: 500

case: 260 control: 240

- n **Rheumatoid Arthritis (GAW15)** :

Number of SNPs: 2300

Population Size: 920

case: 460 control: 460

Search Results

Table 1. Comparison of 5 methods searching atomic risk factors represented by multi-SNP combinations on five real datasets.

| Search method | Risk factor with maximum odds ratio (OR) | | | | | | # with MT-adj. p<0.05 | runtime sec. |
|--------------------------------|--|----------------|------------|---------------|------------------------|---------------|-----------------------|--------------|
| | OR | (OR) 95%CI | case freq. | control freq. | unadjusted p-value | # SNPs in MSC | | |
| Lung cancer | | | | | | | | |
| ES(1) | 13.89 | 1.37 - 140.13 | 0.03 | 0.00 | 7.36×10^{-3} | 1 | 0 | 0.5 |
| ES(2) | 26.63 | 2.69 - 262.69 | 0.05 | 0.00 | 1.00×10^{-4} | 2 | 0 | 21.7 |
| CS(1) | 24.77 | 2.50 - 244.84 | 0.04 | 0.00 | 1.85×10^{-4} | 7 | 2 | 0.6 |
| CS(2) | 38.02 | 3.87 - 372.29 | 0.07 | 0.00 | 2.51×10^{-6} | 3 | 2 | 18.2 |
| ACS(1) | 24.77 | 2.50 - 244.84 | 0.04 | 0.00 | 1.85×10^{-4} | 7 | 2 | 1.0 |
| ACS(2) | 41.92 | 4.28 - 409.80 | 0.07 | 0.00 | 7.36×10^{-7} | 3 | 6 | 25.5 |
| CGS | 72.92 | 7.49 - 708.00 | 0.12 | 0.00 | 7.36×10^{-11} | 12 | 1 | 0.3 |
| RCGS | 97.82 | 10.08 - 947.55 | 0.15 | 0.00 | 8.58×10^{-14} | 12 | 16 | 14.0 |
| Tick-borne encephalitis | | | | | | | | |
| ES(1) | 7.50 | 2.32 - 24.10 | 0.38 | 0.08 | 2.44×10^{-3} | 1 | 0 | 0.1 |
| ES(2) | 30.71 | 2.76 - 326.30 | 0.19 | 0.00 | 1.90×10^{-3} | 2 | 0 | 0.3 |
| CS(1) | 57.33 | 5.27 - 593.95 | 0.31 | 0.00 | 4.44×10^{-5} | 12 | 1 | 0.1 |
| CS(2) | 57.33 | 5.27 - 593.95 | 0.31 | 0.00 | 4.44×10^{-5} | 4 | 0 | 0.5 |
| ACS(1) | 57.33 | 5.27 - 593.95 | 0.31 | 0.00 | 4.44×10^{-5} | 12 | 1 | 0.1 |
| ACS(2) | 57.33 | 5.27 - 593.95 | 0.31 | 0.00 | 4.44×10^{-5} | 4 | 0 | 1.0 |
| CGS | 30.71 | 2.76 - 326.30 | 0.19 | 0.00 | 1.90×10^{-3} | 9 | 0 | 0.1 |
| RCGS | 57.33 | 5.27 - 593.95 | 0.31 | 0.00 | 4.44×10^{-5} | 5 | 6 | 11.5 |

Validation Results

Table 2. Leave-half-out cross-validation of 4 disease-association search methods on 4 real datasets. The validation rate for a method is the portion of MCS found on the training half that have been validated (i.e., have p-value $< 5\%$) on the testing half. The significance rate is the portion of MSC on the training half that stay significant after multiple testing adjustment.

| Data | Search method | validation rate % | significance rate % |
|--------------------------|---------------|-------------------|---------------------|
| Lung cancer (Germans) | ES(1) | 7.0 | 31.0 |
| | CS(1) | 1.0 | 27.0 |
| | CGS | 64.0 | 59.0 |
| | RCGS | 71.6 | 87.4 |
| Chron's disease | ES(1) | 0.0 | 0.0 |
| | CS(1) | 0.0 | 0.0 |
| | CGS | 4.0 | 14.0 |
| | RCGS | 25.0 | 0.0 |
| Tick-borne encephalitis | ES(1) | 0.0 | 93.0 |
| | CS(1) | 0.0 | 95.0 |
| | CGS | 2.0 | 10.0 |
| | RCGS | 1.0 | 31.3 |
| Autoimmune disorder | ES(1) | 1.0 | 17.0 |
| | CS(1) | 0.0 | 0.0 |
| | CGS | 3.0 | 10.0 |
| | RCGS | 2.0 | 16.7 |



Conclusions

- n Approximate search methods find more significant RF's**
- n RF found by approximate searches have higher cross-validation rate**
 - .. Significant MSC's are better cross-validated**
- n Significant MSC's with many SNPs (>10) can be efficiently found and confirmed**
- n RCGS (randomized methods) is better than CGS (deterministic methods)**

n More randomized methods

- Simulated Annealing/Gibbs Sampler/HMM
- But they are slower L

n Indexing (have our MLR tagging)

- Find MSCs in samples reduced to index/tag SNPs
- May have more power (?)

n Disease Susceptibility Prediction

- Use found RF for prediction rather prediction for RF search