HAPLOTYPE- AND PATHWAY-BASED AGGREGATIONS FOR THE BAYESIAN ANALYSIS OF RARE VARIANTS

Peter SARKOZY
Advisor: Peter ANTAL

I. Abstract

The statistical analysis of rare variants from new generation sequencing methods has become a central challenge. We discuss the single gene aspect of "equivalent pathway degrading variants" with similar functional effects, which arises from the transcription to post-translation chain, and its multivariate pathway aspect arising from cascades and modules. We propose a stochastic aggregation for incorporating uncertain knowledge, and describe and evaluate a method for the Bayesian analysis of uncertain data using Bayesian networks.

II. Introduction

New generation sequencing methods are rapidly changing the landscape of the research of common diseases, tumorgenetics, and immunogenetics, some already advocate the era of the "common disease-rare variants" [6]. Thus the discovery and use of rare genomic variants with strong effects became a central challenge.

A haplotype is a combination of alleles or SNPs that are within close proximity on a chromosome, and are transmitted together. Since humans contain a set of maternal and paternal chromosomes, each haplotype is present on both, and the indication of which strand a haplotype is from is marked as its phase. Biological pathways are a number of linked biochemical steps, with a start and an end point; e.g. metabolic pathways or signaling pathways. A variation in the DNA (SNP, insertion, deletion, etc.) can be called a rare variant (RV) if it is present in less than 5% of the population.

At first, we present a unified approach to common and rare variants in common disease. Second, we catalog information sources for the univariate, gene-centered aggregation of variants, starting from transcription regulation to post-translational modifications. We also overview information sources for the multivariate, gene-gene/protein-protein associations and pathway based aggregation of variants. Next we demonstrate a method for the Bayesian analysis of uncertain data in the haplotype analysis of asthma.

III. Common and rare variants in common diseases

After decades of research of the rare variants of rare diseases, common variants (CVs) became central in the research of complex, common diseases. The slower-than-expected progress and partial results of the corresponding genome wide association studies (GWAS) resulted in heavy criticism, [6] concludes the failure of "the common disease (CD)-common variant (CV) hypothesis". Furthermore, it argues for a systems biology based evaluation of RVs in CDs.

The single nucleotide polymorphism (SNP) distribution in the human population is as follows: in a human population with 10^9 members, with 10^2 new germline SNPs per person "every point mutation compatible with life is likely present" while for a pair of genomes CVs give most of the variability. A CD is typically related to the combination of various degradations of multiple pathways, whereas the degradation of pathways is caused by both RVs and CVs. In fact, it is probably a tenable hypothesis that the effect strength of CVs and RVs are comparable, meaning that the current CVs are formed mainly by random drift, and proportionally there is no difference in coverage and functional

role in pathway degradation. This implies that the majority of variants with strong effect are rare and we can discuss classes of "equivalent pathway degrading variants".

It is worth considering the multivariate extension of the earlier calculation in order to evaluate the asymptotic version of the CD-CV hypothesis; which states the asymptotic statistical sufficiency of CVs. Despite the presumed equivalence of CVs and RVs, taking into account the above mentioned aggregated minor allele frequency (MAF) for RVs, the chance of observing the causes (or proxies) of effected pathway patterns drops exponentially in a given patient when seeing only the CVs. To put it in other words, the loss of power of CVs for detecting CDs is a major problem escalating with their multifactorial nature, but this leaves the question of the asymptotic sufficiency of CVs for mapping CDs open.

IV. Aggregation of Rare variants

The weak or non-existing linkage of RVs limits the use of haplotypes or chromosomal regions for aggregations. However this property also limits the possibility of discovery of non-functional associations.

In the univariate, gene-centered approach RVs can be aggregated along the transcriptional regulations to post-translational modifications chain as follows: transcription factor binding sites, miRNA binding sites, splice-regulatory element binding sites, and phosphorylation and glycosylation related variations and conserved regions.

In the multivariate, pathway degrading approach RVs can be aggregated w.r.t. pathway knowledge bases and gene-gene/protein-protein associations.

In each case both nominal and quantitative variables V'_i can be induced based on the original sets of variables $\underline{V}$ using deterministic transformations $V'_i = f_i(\underline{V})$. Typically, there is considerable uncertainty over these transformations, and it is more practical to expect a Bayesian transformation, in which each deterministic transformation has a prior distribution $p(F_i = f_i)$. This implicitly defines a conditional distribution for stochastic mapping (assuming discrete F_i for simplicity)

$$p_i(V`_i|\underline{V}) = \sum p(f_i)f_i)(\underline{V}), \qquad (1)$$

V. Haplotype based aggregation

Haplotype block selection can be performed automatically with several freely available tools, as well as based on expert opinions. The PHASE [8] software can utilize certainty data generated in the previous steps (e.g. image processing, quality control) of the analysis, and output posteriors for the haplotype blocks. Aggregating multiple SNP's into haplotype blocks will result in increased cardinality per variable (since an SNP can only have 3 different states), with many low sample count rare haplotypes. We can handle this effect by implementing the following disease models for each block:

- Dominant disease model: Presence of a specific haplotype on either chromosome can increase risk. This is the general disease model used in most studies
- Recessive disease model: Both chromosomes must have a specific haplotype in order for it have an effect
- Merging of rare haplotypes

The method of maternal and paternal haplotype analysis using sample doubling can be described as follows. We separate the maternal and paternal chromosomes from each sample, because direct joining of the two would cause an undesirable cardinality increase. We also can sample the posterior haplotype distributions produced by PHASE. Next we create two samples from each data point, by splitting the data along the two chromosomes. Because the phase of the aggregations with relation to each other is unknown, we will generate multiple data files, each with a different, randomly

generated global phase, and use the advantages of Bayesian model averaging to average out the effect of unknown global phases

VI. Pre-processing vs. post-processing aggregation

There are various numerically and statistically motivated transformation techniques in the data preprocessing phase, such as normalization, standardization, and dimensionality reduction. These methods can be very valuable in RV aggregation, although the discrete nature of the data excludes many standard solutions. However the real challenge is to incorporate prior knowledge in RV transformation.

The incorporation of such priors in data analysis is already common place, although not in the data preprocessing phase and not for detecting interactions, but in the post processing phase (see Table 1.), such as in the Gene Ontology (GO) annotation analysis or in the Gene Set Enrichment Analysis (GSEA) (for Bayesian aggregation in the post processing phase, see [2].

Aggregation types		Pre-processing	Post-processing
Chromosomal	Based on haplotypes or gene regions	With PHASE or direct joining	Sub Markov Blanket sets
Functional	Aggregating downstream of the functional pathway, or even across the entire pathway.	Along the transcriptional regulations to post-translational modifications chain	GO annotation, GSEA

Table 1: Methods of aggregation along different dimensions. Columns represent pre- and post-processing aggregation, while rows represent variable subset selection methods.

VII.Analyzing uncertain data

Because of uncertainty in RV transformation and aggregation, the analysis of uncertain data is a central theme in the analysis of rare variants. We will concentrate on the Bayesian statistical framework for the analysis, particularly because of its ability to incorporate priors and aggregate posteriors [2].Assuming a distribution over possible datasets, various approaches are as follows

1. using only the most probable data set,
2. using multiple data sets with high probability,
3. Monte Carlo data-averaging in Bayesian model-averaging (MC-DA-BMA).

The Bayesian averaging over model properties is done using Metropolis-Hastings algorithms (M-H) [2]. To avoid multiple burn-in in case of multiple data sets, we can mix data-averaging and model-averaging in a joint Metropolis-Hastings scheme, in the M-H-within-Gibbs, which is a hybrid of M-H and Gibbs sampling, the Gibbs sampling steps and the M-H steps can follow each other successively [3]. The Gibbs sampling steps can be used to generate uncertain and missing values, then using the completed data set a structure can be sampled in the M-H step.

VIII. Results

We performed transformations in a partial genetic association study in asthma, both in the gene-centered and in the pathway-centered approach. Here we illustrate our Bayesian network based methods for the Bayesian analysis of uncertain data for haplotypes in the gene-centered approach.

Because of computational reasons such separation of blocking, phasing (haplotype inference [1]) and data analysis is a practical choice followed in many systems (see e.g. HapScope [5], HAPLOT [4], GEVALT [7]). We similarly follow this decomposition, in which the biomedical expert specifies

the blocks a priori (corresponding to unrelated chromosomal regions), then the PHASE method is applied for each block [8] to generate a maximum a posteriori distribution over phased genotyped data, and finally we apply our Bayesian model-based approach [2].

We applied the Monte Carlo data averaging in Bayesian model averaging (MC-DA-BMA) method in a genetic association study in asthma research investigating 56 SNPs, 15 genes in chromosome 11 and 14 (settings: burn--in: 1000000, step: 5000000, 10 random datasets from phasing distribution, see Table 2.).

Region Name	Average MBM posterior	Average MBM Posterior (ML)	Variance of MBM posterior
FRMD6 START HT1	0.638754	0.643055	0.0946
FRMD6 START HT2	0.723333	0.657325	0.0250
AHNAK HT 1	0.160561	0.152456	0.0187
AHNAK HT 2	0.611516	0.589275	0.0058

Table 2: BNF maximum likelihood and the averaged results on chromosome 11 and 14. Where HT is the haplotype.

IX. Conclusions

We focused on the Bayesian statistical framework for the analysis, particularly because of its ability to incorporate priors and to aggregate posteriors [2]. We summarized various types of aggregations of rare variants, proposed and implemented a stochastic aggregation scheme, which can be used both in the pre-processing and post-processing phases. We implemented various sampling schemes to cope with uncertain data sets using parallel computing information resources (for availability, see http://mitpc40.mit.bme.hu:8080). We demonstrated these methods for the Bayesian analysis of uncertain data in the haplotype analysis of asthma.

To cope with challenges of the analysis of RVs it is worth noting that beside the aggregation of the effects of RVs along pathways, their effects can be aggregated along multiple quantitative traits (QT) as well, for which new statistical methodologies are needed. QT analysis truly avoids some serious problems of case-control studies, such as binary oversimplification and population confounding. These requirements are neatly covered by a causal network analysis. Furthermore the Bayesian statistical framework offers many advantages for this analysis, such as the use of background knowledge as prior and in the posterior interpretation, meta-analysis, and automated correction for multiple testing. The application of RVs from NGS using a (Bayesian, causal) network analysis may define a third phase in genetic research after linkage analysis and association analysis.

References

[1] Clark A.G. Inference of haplotypes from PCR-amplified samples of diploid populations. Molecular Biology and Evolution, 7(2):111-122, 1990.

[2] P. Antal, A. Millinghoffer, G. Hullam, Cs. Szalai, and A. Falus. A Bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction. JMLR Proceeding, 4:74-89, 2008.

[3] S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. The American Statistician, 49(4):327-335, 1995.

[4] Kidd K. K. Gu S, Pakstis AJ. Haplot: a graphical comparison of haplotype blocks, tag snp sets and snp variation for multiple populations. Bioinformatics, 21(20):3938-3939, 2005.

[5] William L. Rowe, Jinghui Zhang. Hapscope: a software system for automated and visual analysis of functionally annotated haplotypes. Nucleic Acids Research, 30(23):5213-5221, 2002.

[6] J. McClellan and MC. King. Genetic heterogeneity in human disease. Cell, 141:210-217,2010.

[7] Ron Shamir Ofir Davidovich, Gad Kimmel. Gevalt: An integrated software tool for genotype analysis. BMC Bioinformatics, 8(36):2105-2112, 2007.

[8] M. Stephens and P. Donnelly. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. The American Society of Human Genetics, 73(5):1162-1169, 2003.