Finding Genotype and Phenotype Associations

Oskar Gross

Helsinki October 15, 2010

UNIVERSITY OF HELSINKI
Department of Computer Science
In this paper we present the biological background of genotype and phenotype and give a statement of the problem. We also take a look at different ways for finding associations between genotype and phenotype.

The paper is focused on giving introduction to the genotype-matching problem and gives intuition how current problem can be approached and we show why finding a very good solution is much more complex than it seems in the first glance.

For each method we bring out the most important components and try to give a biological connection why these solutions may improve the results.

ACM Computing Classification System (CCS):
J.3 [Biology and genetics],
I.5 [PATTERN RECOGNITION]
# Contents

1 Introduction  
2 Biological Background  
3 Genotype-Phenotype Matching  
   3.1 Classification Association Rule Mining  
      3.1.1 Conditionally Weighted CMI (CWCMI)  
   3.2 Genome Associations to a Quantitative Trait Network  
   3.3 Markov Blanket Search  
4 Discussion  
References
1 Introduction

In this paper we briefly give the biological motivation for finding the associations between genotype and phenotype and then show what are the different approaches for finding a solution to current problem. For every method we bring out the main ideas and we go more deeply to the state-of-art methods and also see, what are the results and discuss if and how they could be possibly improved.

2 Biological Background

The relation between genotype and phenotype has been and is one of the key questions in genetics. Genotype is the genetic makeup of an organism and phenotype is the set of observable traits which is combined by the effect of genes and environment. While Gregor Mendel found the connection between genotype and phenotype by studying the family patterns (phenotypic traits) for hereditary diseases, it turns out that there are traits which do not follow the rules of Mendel and also traits which controlled by many different genes - these are called multifactorial traits. Finding the associations between genotypes and phenotypes is important for many purposes: to improve diagnostic prediction, introducing individualized therapeutic interventions and understanding biological mechanisms in general [ACCH10]. The motivation from biology is quite obvious and although the associations between phenotype and genotype might look trivial in the first glance, the problem is that frequently the effect of a gene on a specific phenotypic trait is very low and even further there are very rare genes which have a very small effect on the phenotype [ACCH10]. This is also the reason why we need to use more advanced algorithms and statistical methods for finding solutions to current problem. Further we give a short overview of biological terms which are used in current paper.

*Genome-Wide Association Study (GWAS)* - "A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses." [Ins10]

*Single nucleotide polymorphisms (SNPs)* are the most common variation in the hu-
man genome. SNP represents a difference in a single nucleotide in DNA. For example SNP may replace one nucleotide to another in a certain stretch of DNA. Most of the SNPs have no effect on phenotype, but some of them have important role in detecting and predicting illnesses and drug responses. By studying stretches of DNA which harbor a SNP associated with a disease trait, we may find genes that are associated to a disease or other phenotypic traits [HJBJ07].

Phylogenetic trees are for showing the evolutionary relationships between different biological species or other entities.

Quantitative Trait Network (QTN) is a network which is constructed by the quantitative traits of a complex multivariate phenotype. For instance a asthma is a complex multivariate phenotype i.e. it has many symptoms. In the QTN of asthma the symptoms are nodes and they have weighed edges between each other. The weight of the edge is the correlation strength between these two traits [SK09].

3 Genotype-Phenotype Matching

It has been argued that GWAS may not be the best way to find solution to the genotype-phenotype problem. It is brought out that, most of the associations we have found have relatively small effect on the actual phenotypic trait and it has not lead to new genetic tests for predicting or preventing diseases. When designing methods for analyzing the genotype-phenotype connections, we have to take into account that these connections are non-linear. It is emphasised that powerful data mining and machine learning methods need to be developed in order to understand these connections [MAW10]. As follows we give an overview of the state-of-art methods for finding the connections between phenotype and genotype.

3.1 Classification Association Rule Mining

As organismal traits are often influenced by combinations of genes, it makes sense to find these multi-feature associations between genotype and phenotype by using the classification association rule mining. In general we would like to find rules in form $G_A, G_B \rightarrow P_A$, where $G_A, G_B$ are genes and $P_A$ is a phenotype trait. Tamura et al. [TD08] proposed a method NETCAR which uses clusters of orthologous groups (COG) of proteins and association rule mining to find relevant associations
between presence of phenotype trait and COG’s. Derived from this idea MacDonald and Beiko [MB10] showed that when applying predictive rule mining approach to genotype-phenotype mapping problem we get slightly more accurate results 100 times faster than using NETCAR.

3.1.1 Conditionally Weighted CMI (CW CMI)

A measure based on mutual information (MI) measure for highlighting genes with relevant biological functions can be further extend to a CW CMI [MB10]. If a phenotypic trait is common within a taxonomic group, then the genes that correlate with the trait may rather be the characteristic of the taxonomic group, not causal of the trait [MB10]. First thing to notice is that some common genes can be explained by common ancestry. The mutual information (MI) measure is given as (1)

\[
I(X;Y) = \sum_{x,y} p_{x,y} \log \frac{p_{x,y}(x,y)}{p_x(x)p_y(y)},
\]

where \( p \) is the probably mass function for the subscripts. MI measures the mutual dependence of two variables and in current case between the phenotypes and genes. The problem with this measure is, that it does not take into account the fact that some genes are explained by common ancestry. For this reason it is reasonable to use the conditional mutual information measure (CMI) (2)

\[
I(X;Y|Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y,z)p_z(z)}{p_{x,z}(x,z)p_{y,z}(y,z)},
\]

where \( p(x,y,z) \) is the joint probability mass function of random variables \( X, Y \) and \( Z \), and \( p \) are the probability mass functions of the subscripts. For applying current measure for the phenotype problem, let \( X \) be the presence or absence of the gene, \( Y \) be the phenotype and \( Z \) be the taxonomic group. The CMI measure gives us then the amount of MI between the genes and phenotype which cannot be explained by common ancestry. An important point is that for any phenotype there is a maximal CMI as it is impossible to share more information than it is present in either factor. Maximum CMI is the entropy which remains after subtracting the MI of phenotype and taxonomic groups, which is defined as

\[
H(Y|Z) = \sum_{y,z} p(y,z) \log p_{y|z}(y|z).
\]

When \( H(Y|Z) \) is positive we can define a normalized weight formula

\[
\alpha(X, Y, Z) = \frac{I(X;Y|Z)}{H(Y|Z)}
\]
The interpretation of current normalization is that if $\alpha = 0$ then there is no evidence that $X$ and $Y$ share any more information than $X$ and $Z$ or $Y$ and $Z$ alone. On the other hand when $\alpha = 1$ then the information shared between $X$ and $Y$ is the maximum amount possible given the taxonomic groups. Classification based on Predictive Association Rules (CPAR) [Jia] builds predictive rules so, that it covers all examples for each class label in a dataset. In general, CPAR is constructed so, that the rules generalize as many samples as possible for a certain class. The samples which were not covered by the generalization are focused in second iteration. Current approach which uses CPAR approach was compared to the NETCAR algorithm. The accuracy of the approach was slightly better, but the main achievement is the running time of the method which is 100 times faster than the NETCAR algorithm.

3.2 Genome Associations to a Quantitative Trait Network

As mentioned before, it is not reasonable to suspect that phenotypic trait depends on single gene and also vice versa we should presume that one gene may influence similar phenotypic traits. One way of finding relations between genotype and phenotype is to map genotype to a group of quantitative traits. The result is that genetic markers which influence subgroups of phenotypic traits can be detected. The following method is proposed by Kim et al. [SK09]. Considering a QTN it may be that highly correlated phenotypic traits in a subnetwork share some common genetic causes. Using this idea it may also be possible to find patterns which we would miss by analyzing single trait at a time. When incorporating molecular and/or clinical phenotype network to the searching for genetic associations increases the power of detecting pleiotropic effects. Proposed algorithm graph-guided fused lasso (GFlasso) searches for genotype markers that are associated to multiple highly correlated traits. Authors proposed several different GFlasso methods. Here we give an overview of the weighted-graph-guided fused lasso ($G_W$Flasso). In general $G_W$Flasso finds an estimate of the regression coefficients as follows:

$$\hat{B}_{GW} = \arg\min_k \sum (y_k - X_k^T \beta_k)^T (y_k - X_k^T \beta_k) + \lambda \sum_k \sum_j |\beta_{jk}| + \gamma \sum_{(m,l) \in E} f(r_{ml}) \sum_j |\beta_{jm} - \text{sign}(r_{ml}) \beta_{jl}|,$$

from which the set of QTLs $S_k, \forall k$ can be uncovered. It is important to mention, that here $X$ is a $N \times J$ matrix of genotypes for $N$ individuals and $J$ SNP’s, where
each element $x_{ij}$ of $X$ is assigned 0, 1, or 2 depending of the number of minor alleles at the $j$-th locus of the $i$th individual, $Y$ is a $N \times K$ matrix of $K$ trait measurements over the same set of individuals where $y_k$ means the $k$th trait of $Y$ and $r_{ml}$ refers to the correlation strength between the traits $m$ and $l$. The last term of the equation is known as the fusion penalty which makes the strength of influence of each marker to become more similar across the strongly correlated phenotypic traits. Kim et al. also propose a optimization algorithm which we do not analyze here further as it is more of technical detail.

Proposed GFlasso method was compared to three other methods: single-SNP/single-trait regression analysis, multivariate regression methods for single output and PCA-based regression method which takes into account trait correlations. The running time of GFlasso for 250 SNPs and 250 phenotypes is approximately 300 minutes. Running the algorithm with 50 SNPs and 1300 number of phenotypes takes around 50 minutes and with 1700 number of phenotypes it takes around 150 minutes. The average test error for GFlasso method was ranging from 10-16% with an average around 12% depending on the trait correlation threshold parameter.

### 3.3 Markov Blanket Search

Most of statistical methods which identify genes consider different expressions, but they do not use the knowledge of protein-protein interactions. It is proposed by Yeh et al. [YLYS10] to identify significant networks and genes based on the phenotype and protein-protein interaction networks and also use the microarray data of the specific disease. The structure is learned with Markov Blanket Search. Markov blanket for a node $A$ is composed of $A$’s parents, its’ children and its’ childrens’ other parents. The conditional entropy is used to fit the model to biological networks.

$$G = H(X_i|M_i) - H(X_i|M_i \cup X_j) + H(X_j|M_j) - H(X_j|M_j \cup X_i),$$

where $G$ denotes the information gain of gene $X_j$ which is one of the neighbourhoods of the Markov blanket of $X_i$ in protein networks. $H(X_i|M_i \cup X_j)$ is the conditional entropy of $X_i$ given its best Markov blanket. Then the Markov blanket search algorithm starts the search from the disease causing gene and performs greedy search to find the best Markov blanket for each gene.

Current method was compared to different statistical tests. The results showed that the precision and recall were significantly higher when the genotype-phenotype information was taken into account. The main idea of T-test is to determine whether
Figure 1: The risk-allele frequency and strength of genetic effect. The figure is based on the figure by Nature Reviews Genetics.

the means of two groups are statistically different from each other and the Wilcoxon test is two measure the distribution differences in two groups. In general the precision rates where slightly better than for T-test or Wilcoxon test. The combination of precision and recall was significantly better than for other compared statistical methods. Mainly current method shows that taking the genotype-phenotype knowledge into account it may increase the opportunity to identifying disease-related genes.

4 Discussion

As we can see there are many approaches for finding the relations between genotype and phenotype. At the moment there does not seem to be a very good method for predicting or detecting relations between the phenotype and genotype. On the other hand, as the search space in the genome is very large, we have methods that reduce the search space a little and by using these methods there is a possibility to get further clues to detect the genes which play role in a certain phenotype. A
very nice illustration of the phenotype-genotype problem is illustrated on Figure 1, where we can see, that we have a lot of common variants which have a very low or intermediate effect, thus it is also very hard to detect these genes.

References


