論 文 の 内 容 の 要 旨

論文題目

Identification of Interactions in expression Quantitative Trait Data via the Minimum Description Length

（最小記載長を用いた定量的発現形質データからの相互作用の同定）

氏　　名　　ジョージ　チャルキディス

With increasing amounts of genomic data it becomes possible to glimpse deeper into the underlying biological mechanism of disease. The hope of genomic analyses are not only to increase our knowledge and understanding of disease, but also find new cures and treatment methods.

Even though genome-wide association studies (GWAS) helped reveal a magnitude of genetic variations to be associated with phenotypes, may it be a disease like Type 2 Diabetes or a more general phenotypic trait like the body mass index (BMI), the actual mechanisms of how those genetic variants exert their influence on the phenotype have for the great majority of discovered associations remained unclear.

With respect to single nucleotide polymorphism (SNP) genetic variants, one reason for this lack of understanding is that several SNPs, which have been associated with a certain disease via GWAS, appear to originate from non-coding regions of the genome, which means that they are not contained in any known gene that codes for a protein. Disease associated SNPs in protein coding genes are easier to discern, because they can alter or even stop the encoding of a protein and initiate a chain reaction leading to disease pathogenesis.

Recent advances in genome technology have enabled scientists to also sequence the RNA of a cell and thus capture the transcriptomic landscape. The transcriptome acquired via RNA-sequencing reveals the gene expressions of a person.

Initial studies on transcriptome variability have shown that although genetic variation is responsible for producing differing phenotypes, there is mounting evidence that changes in the transcriptomic landscape are more directly linked to phenotypes. Furthermore, transcriptomic variability seems to be able to code for more diversified phenotypes.

Although a phenotype can be any observable trait, in this thesis we focus on disease phenotypes.

To improve our understanding of the involvement of genetic variants in disease susceptibility, it is important to discover the links between SNPs and variations in gene expressions. By doing so, explanations of the underlying molecular mechanisms of disease can be found. Interactions of SNPs with gene expressions that are associated with disease reveal parts of molecular pathways that make up the underlying machinery of disease.

In order to study the complex relation between SNPs, transcriptome, and disease, a new type of genetic study is emerging which is called expression quantitative trait locus study (eQTL). One aim of eQTL studies is to enhance our knowledge regarding the functional relationships of the many SNPs discovered by GWAS. Because of the fact that many GWAS SNPs fall into non-coding regions, it is very well possible that those disease associated SNPs exert their influence on disease via regulatory mechanisms by changing the transcription activity of genes which in turn have an effect on disease.

As a consequence of this hypothesis, lots of next-generation DNA Sequence, health care, and medical data are pooled into eQTL datasets with the hope of uncovering interactions, i.e. relationships between SNPs and gene expressions, that could help us clarify the molecular mechanisms involved in disease pathogenesis.

A further goal is to utilize these novel findings for improving diagnostic capabilities through the identification of new biomarkers or even enable the tailoring and guidance of treatment approaches, which would constitute a form of personalized medicine.

It is very important to detect the associations between SNPs and gene expressions that constitute the interaction networks involved in disease and recover them as accurately and completely as possible.

In this thesis we propose an information theoretic tool named qMAP for detecting SNP-transcript associations in eQTL data and extracting the disease associated interactions based on the minimum description length principle (MDL).

The tool qMAP implements a detection strategy adhering to the MDL idea that SNPs which influence gene expressions can be used as a statistical model to describe those gene expressions for constructing a more compact codification of them.

From data compression it is well known that good data descriptions yield short codelengths, which means that the better the SNP genotype can predict the gene expression pattern the shorter the resulting codelengths of the encoded gene expression pattern will be.

The encoding approach used for this task is the normalized maximum likelihood (NML) code. To this end, the dynamic programming method of Kontkanen and Myllymaki (KM) is extended and modified to work in a mixed environment that consists of continuous gene expressions and discrete genotypes, i.e. SNPs.

The strength of an association between a SNP and gene transcript is determined by an MDL-score that we propose in this thesis. The MDL-score is derived from the various NML codes needed to describe the interaction phenomena in eQTL data, with higher MDL-scores indicating stronger associations and functional relationships between SNPs and gene expressions.

For calculating the MDL-score a two-dimensional grid is laid on the data which is spanned by the axis for gene expressions and genotype respectively. Via the extended KM-algorithm, the grid is compartmentalized by optimizing for the stochastic complexity of the eQTL interaction pair, which is the minimization of the normalized maximum likelihood code.

Since the compartmentalization of the genotype axis is determined by the encoding procedure for the genetic variants, the KM-algorithm is extended to incorporate this information and an algorithm is created that optimizes the NML-code for a mixed pair of continuous gene expressions and discrete SNPs random variables. The optimization of the grid structure with respect to the NML-code yields the resulting description lengths of the SNP-transcript pairs from which the MDL-score for functional relationship and association strength is derived.

The entire analysis program is called qMAP, which stands for quantitative MDL association program, and is implemented in the Python language.

Using simulated eQTL data which contain a known interaction network of SNP-gene expression associations, the detection rate performance for varying sample sizes, reaching from 10 to 380, has been benchmarked and analyzed for qMAP, the genome analysis toolkit PLINK, the maximal information coefficient MIC, and MI-KDE (a custom implementation that uses kernel density estimation (KDE) to calculate the mutual information (MI) in a mixed environment).

For this simulated eQTL study, qMAP can improve the detection rate of correct SNP-gene expression associations to 78% from 57.3% for PLINK and 52% for MIC. With appropriate parameter tuning MI-KDE also achieves a detection rate of 78% and shows equivalent performance to qMAP.

Compared to the other state-of-the-art approaches, qMAP displayed the highest reconstruction capability by delivering the most accurate and complete image of the interaction network that was contained in the synthetic eQTL dataset.

Since qMAP showed outstanding performance on simulated datasets, it was also applied to a real human cortical gene expression eQTL dataset. The obtained results were compared against the reported associations between SNPs and gene expressions in the original study and the concordance of results was confirmed. This shows that qMAP also performs well in real biological settings.

With an improved detection rate, the MDL-based eQTL analysis program qMAP can be a helpful tool for biologists and physicians who want to extract more information out of eQTL data by identifying the interaction networks that are made up of the various functional relationships between SNPs and gene expression which are associated to disease.

Consequently, qMAP can enable better insights into the molecular underpinnings of the workings of disease and elucidate information regarding disease susceptibility.