Genome-wide Clustering of Genes[*]

Christian Rengstl

Universitätsklinikum Regensburg
Klinik und Poliklinik für Innere Medizin II
Franz-Josef-Strauß-Allee 11
93053 Regensburg
Germany
christian.rengstl@klinik.uni-regensburg.de

Abstract
With the development of 500K chips, i.e. approximately 500.000 single nucleotide polymorphisms per individual, the quantity and quality of data in genetic researches have risen considerably. As the amount of data makes it hard for any researcher to identify genes and SNPs that are relevant for a specific research problem, it is necessary to organize the data into clusters based on their informativeness. A good approach to cluster the data is to use genetic data, the gene functions and the phenotypes of individuals.

1 Introduction

The recent development of so-called 500K chips in the areas of bio-informatics and genetics has enabled researchers world-wide to perform large high-scale studies on the human genome. The idea behind a 500K chip is that for each individual within a population under consideration around 500.000 bi-allelic genetic markers, single nucleotide polymorphisms (SNPs), are genotyped. As SNPs are the largest source of variation in the human genome and have a very low mutation rate in comparison to other genetic markers they are very suited to conduct genetic studies.

These 500K chips, though, produce enormous amounts of data, which makes it hard for the researcher to focus on interesting genes. A lot of applications in the area of bio-informatics calculate the likelihood stating in how far genes might be

[*] Published in: OSSWALD, Achim; STEMPFHUBER, Maximilian; WOLFF, Christian (eds.) (2007). Open Innovation. Proc. 10th International Symposium for Information Science. Constance: UVK, 361-365.

relevant for diseases under consideration. The results can then be organized in signalling pathway networks, i.e. networks of gene interactions, using pathway browsers like Ingenuity Pathway Analysis [http://www.ingenuity.com]. Using these networks as starting point, genes and SNPs found on those signalling pathways should be clustered in order to improve the researchers' situation concerning focusing on specific genes among the pool of all input genes.

The input for this project is taken from the KORA (Kooperative Gesundheitsforschung in der Region Augsburg) study which comprises 500K chips of 1644 unrelated individuals. The *x* most promising genes will be assembled into signalling pathway networks, which in turn will serve as the input for the clustering procedure.

2 Data

The data is stored on a Windows 2003 server running EnterpriseDB version 8.1, which is an extended version of the more known database PostgreSQL. The current size of the database amounts to 143 gigabytes split over around 300 partitioned tables. The database design ensures that all data can be queried through one master table. Each chromosome has in turn an own master table with up to 22 child tables.

SNP data are available in pure text files with the format as seen in table 1. An example for signalling pathway data can be seen in table 2.

Chromosome	*Position*	*Pid*	*SNP-ID*	*Allele 1*	*Allele 2*
22	014433758	ZZZ000110011	RS915677	A	C
22	014433758	ZZZ000110022	RS915677	C	T
22	014433758	ZZZ000110033	RS915677	G	G
22	014433758	ZZZ000110044	RS915677	-	-

Table 1: Exemplary SNP data

ID	*Genes*	*Score*	*Focus Genes*	*Top Functions*
1	ANLN, APAF1, APC, ...	41	35	Cancer, Cellular Assembly and Organization, ...
2	ACP1, ADRB2, ...	41	35	Cellular Compromise, Immune and Lymphatic System Development
3	BCL2, BID, BIRC3, ...	41	35	Cellular Development, Hematological System Development ...

Table 2: Exemplary signalling pathway data

3 Planned Approach

In order to cluster this amount of data, it is necessary to minimize the data under consideration before the actual clustering process starts. To accomplish this, it has been planned, and already partially implemented, to calculate the informativeness of all SNPs and genes under consideration. However, before the informativeness of a SNP can be calculated, it is necessary to define the degree of interdependence between SNPs on a gene, the so-called "linkage disequilibrium" (LD). This measure is calculated using $r^2 = \frac{D^2}{p_1 q_1 p_2 q_2}$ where $D = x_{11} - p_1 q_1$ with x_{11} being the frequency of the combination of the first alleles of two SNPs, and p_1, p_2, q_1 and q_2 being haplotype frequencies. LD is in so far important as two SNPs that exceed the threshold of usually 0.8 are considered redundant and therefore can be omitted from the further clustering.

The informativeness of SNPs is defined as the relation of the sum of the entropy of the less frequent to the sum of the entropy of the most frequent allele of all individuals. The entropy of a SNP in turn is defined as $H_M = -P(A) * \log P(A)$ where *M* is a biallelic marker and *A* in this case refers to the most frequent allele found on the genetic marker. This way the number of SNPs can be reduced before the clustering process continues. As a signalling pathway can be thought of as a net of links between genes, the PageRank algorithm can be used to deduce the weight of a gene. The informativeness of genes is calculated using an adapted PageRank [Page et al 99] algorithm. In this algorithm all genes are initialized with the sum of the entropy values of all SNPs found on a gene excluding those SNPs that were removed from the initial dataset due to high LD. Like this all genes below a certain threshold, which still has to be defined, can also be excluded from the clustering process.

After the dataset is reduced to only the most relevant items, the actual clustering process can be initiated. As not only the genetic data, i.e. the alleles, of a gene/SNP are interesting for clustering, but also the phenotypical realizations of those genetic data, the phenotypes of the individuals under consideration are also an important part of the clustering process. The aim here is not to cluster genes and SNPs only according to their genetic data but also to include phenotypical data in the clustering. The actual function of a gene within the process of protein synthesis, which can be extracted from protein databases like Swissprot or from the signalling pathway input, will also be considered during clustering. The clustering algorithm that will be implemented for this approach is self-organizing maps. Nevertheless, for evaluation purposes and to combine aspects of several algorithms, more than only one clustering algorithm will be implemented in order to find the one best suited for the clustering of genetic data in combination with both phenotypes, genetic functions and the informativeness of genes/SNPs.

4 References

[Hampe et al 06] Hampe, Jochen; Schreiber, Stefan; Krawczak, Michael (2006). Entropy-based SNP selection for genetic association studies. *Human Genetics*, 114/1 (2006), 36-43.

[Page et al 99] Page, Lawrence; Brin, Sergey; Motwani, Rajeev; Winograd, Terry (1999). The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies Project. http://dbpubs.stanford.edu/pub/showDoc.Fulltext?lang=en&doc=1999-66&format=pdf&compression=&name=1999-66.pdf [January 2007]

[Zhao et al 05] Zhao, Jinying; Boerwinkle, Eric; Xiong, Momiao (2005). An Entropy-Based Statistic for Genomewide Association Studies. *American Journal of Human Genetics*, 77 (2005), 27-40.

[Ziegler & König 06] Ziegler, Andreas; König, Inke Regina (2006). A statistical approach to genetic epidemiology: concepts and applications. Weinheim: Wiley-VCH, 2006.