DECISION SUPPORT SYSTEM FOR DESIGNING SNP ASSOCIATION STUDIES

Gergely HAJÓS
Advisors: Péter ANTAL, Tadeusz DOBROWIECKI

I. Introduction

The aim of the single nucleotide polymorphisms (SNPs) based association studies is to find SNPs relevant to the disease under investigation. Despite of the growing importance of genome-wide association studies (GWAS), the selection of tens out of million SNPs present in the partial genome screening studies (PGSs) is still an important problem and cannot be expected to disappear in the near future.

The presented solution provides help for the researchers in the design of association studies by searching, browsing and selecting biologically and medically relevant SNPs. Another important issue is to consider technological aspects of the laboratory's measurement tools [1] – whether the SNP set is appropriate for measurement. Finally we support the design of the overall measurement, i.e. the scoring of the joint set of potential SNPs, by balancing their individual relevance and internal redundancy. With the integration and automated analysis of biomedical databases and derived datasets, the solution shortens from one or two weeks long preselection period to some days.

Because of the importance of the accumulating potentially related data sets and multistep dependent measurements, we decided to formalize this problem as a sequential decision problem.

II. The method

Earlier SNP properties were collected to estimate the biological and medical relevance of SNP sets with the help of physicians and biologists [2] from the Semmelweis University. The implemented decision support system offers the researchers a graphical interface wherewith it is possible to express their preferences to score these mentioned properties. The program finally selects a SNP set by the biological properties, the given scores, and the correlation between them.

The program performs three consecutive functions controlled by the user (Fig. 1):

1. **Searching and browsing** – The first step is a complex search which is designed to integrate certain SNP searching methods used by the researchers earlier at the DGCI laboratory (i.e.: UCSC [3], HapMap).

2. **Defining priorities** – In the second step it is possible for the users to set coefficients to express preferences about SNP properties. In this step with the help of the decision network, every SNP gets a score expressing its utility and relevance.

3. **Set selection** – Finally the third step performs the set selection, the program attempts to cover the whole given sequence while it chooses relevant and (if necessary) measurable SNPs [4].

The program uses several data sources, databases of the NCBI project such as dbSNP [5], GO [6], HapMap [7]. It performs meta-analysis on the results of text and data mining processes, previous measurements and Bayesian statistical analysis.

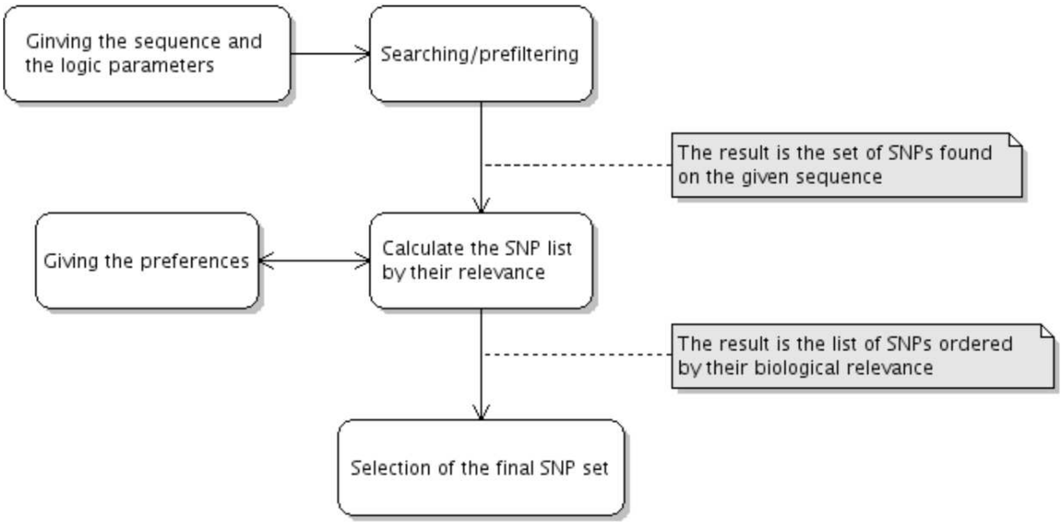
Ginving the sequence and the logic parameters
Searching/prefiltering
The result is the set of SNPs found on the given sequence
Giving the preferences
Calculate the SNP list by their relevance
The result is the list of SNPs ordered by their biological relevance
Selection of the final SNP set

Figure 1: The process of SNP selection

III. Conclusion

The automated and intelligent selection of SNPs is a central issue in modern statistical genetics. We implemented a flexible and principled system, focused on the needs of the SOTE DGCI laboratory with a rich user interface designed to help physicians and biologists.

IV. Future goals

A typical association research project is an iterative sequence of selections, measurements and analysis, while currently available systems — including ours — are only semi sequential, as they can incorporate previous results at best, but they cannot predict the value of further analysis. Such a partially sequential approach is based on three data sources: expert knowledge, publicly available databases, and the results of previous analysis. In the future we want to emphasize the role of potential subsequent measurements. There are two key elements to achieve this in a principled way in the new version of the system:

- modeling the future probabilities of the relevance of SNPs in a hypothetical study with additional measurements. It is solved by constructing typical learning curves for SNPs w.r.t. a given disease (using bootstrap and Bayesian methods [8, 9]),
- approximating the value (expected utility) of their further measurements.

References

[1] Beckman Couler, *Beckman Couler SNPStream genotyping tools*, http://www.mathworks.com/matlabcentral/fileexchange/.

[2] P. Wang, "Snp function portal: a web database for exploring the function implication of snp alleles," *Bioinformatics*, 22(14):523–529, July 2006.

[3] University of California, Santa Cruz, *UCSC Genome Bioinformatics Site*, http://genome.ucsc.edu/.

[4] Z. S. Qin, "An efficient comprehensive search algorithm for tagsnp selection using linkage disequilibrium criteria," *Bioinformatics*, 22(2):220–225, Jan. 2006.

[5] National Center for Biotechnology Information, *dbSNP*, http://www.ncbi.nlm.nih.gov/projects/SNP/.

[6] GO Consortium, *Gene Ontology*, http://www.geneontology.org/.

[7] *International HapMap Project*, http://www.hapmap.org.

[8] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, London, 1993.

[9] P. Antal, "A bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction," *JMLR Workshop and Conference Proceedings 4*, pp. 74–89, 2008.