MutaGeneSys: Making Diagnostic Predictions Based on Genome-Wide Genotype Data in Association Studies

Julia Stoyanovich [a]*; Itsik Pe'er [a]

[a]Columbia University, Department of Computer Science, 1214 Amsterdam Avenue, New York, NY 10025 USA

ABSTRACT

Summary: We present MutaGeneSys: a system that uses genome-wide genotype data for disease prediction. Our system integrates three data sources: the International HapMap project, whole-genome marker correlation data and the Online Mendelian Inheritance in Man (OMIM) database. It accepts SNP data of individuals as query input and delivers disease susceptibility hypotheses even if the original set of typed SNPs is incomplete. Our system is scalable and flexible: it operates in real time and can be configured on the fly to produce population, technology, and confidence-specific predictions.

Availability: Efforts are underway to deploy our system as part of the NCBI Reference Assembly. Meanwhile, the system may be obtained from the authors.

Contact: jds1@cs.columbia.edu

1 INTRODUCTION

Availability of genetic information continues to revolutionize the way we look at medicine, with an ever stronger trend towards personalized diagnostics and treatment of heritable conditions. One promising research direction considers the evaluation of an individual's susceptibility to disease based on single nucleotide polymorphisms (SNPs) – DNA sequence variations occurring when a single nucleotide in the genome differs between members of the species. Significant attention of the research community is devoted to determining *direct causal association* between the genotype and the phenotype, and many interesting correlations have already been reported. The Online Mendelian Inheritance in Man (OMIM)[6] database is the currently most complete source of associations between high-profile phenotypes and genotypes. A text search of OMIM yields, for example, a correlation between a C-T SNP (rs908832) in exon 14 of the ABCA2 gene and Alzheimer disease, an association of two C-reactive protein (CRP) polymorphisms, rs1800947 and rs1205, with Systemic Lupus Erythematosus and antinuclear antibody production, and a connection between an nsSNP in the IFIH1 gene, rs1990760, and insulin-dependent Diabetes Mellitus.

Fully exploiting genetic information for disease prediction is difficult for two reasons. First and foremost, genetic information remains expensive to collect, and it is currently economically prohibitive to make a complete set of an individual's genotypes of SNPs available for analysis. Nature Genetics' "Question of the Year"[5] announced the sequencing of the entire human genome for $1,000 as a goal for the genetics community. Cost effective methods (e.g. SNP arrays) currently exist for collecting genetic data from 1-5% if all 11 million human SNPs. This calls for the development of techniques that can effectively utilize partial genetic information for disease prediction. The second reason is that OMIM is accessible only in text form, and while there is some amount of cross-reference between OMIM and other NCBI databases, using this data for automated genome-wide diagnostics remains difficult.

Several studies, culminating with the International HapMap project[7], report on a significant amount of correlation among markers in the genome[9]. This genomic redundancy enables one to start with an incomplete set of typed SNPs, and expand this set by including associated proxies. *Indirect association* between proxy genotype and phenotype is precisely the tool that is needed for effective and efficient association analysis.

In the simplest case, SNPs are correlated pairwise, and one of them may be predicted by the other; such correlations are referred to as *single-marker predictors*. Many two-marker and three-marker predictors are also known. Correlations between causal SNPs and their proxies are associated with a coefficient of determination (squared Pearson's correlation coefficient) $r^2 \epsilon [0, 1]$. Marker correlation is population-dependent[8]; the coefficient of determination may also depend on the genotyping technology and the resolution that were used to compile the marker correlation data. For example, according to our marker correlation dataset[2], rs12076827 (allele A) and rs1572970 (allele A) on chromosome 1 together determine that rs1205 will appear with minor allele T, with $r^2 = 0.733$ (in the Japanese and Chinese population). This correlation was established using the Affymetrix GeneChip genotyping technology at 500K resolution. OMIM links rs1205 with Systemic Lupus Erythematosus (SLE) and antinuclear antibody production.

Genome-wide correlation can be used to augment an individual's genetic information, greatly enhancing its diagnostic value. In the example above, if rs1205 was not typed, but rs12076827 and rs1572970 were, an association with the SLE-causing phenotype could still be determined. Our project is the first step in this direction. The goal of our project is to stream-line the process of correlating SNP information with heritable disorders, and to enable real-time retrieval of disease susceptibility hypotheses on genome-wide scale. The aims of our system are two-fold.

- To create a flexible, extensible, and efficient framework for storage and querying of direct and indirect association data.
- To provide a set of tools for data import and maintenance, and to use these tools to populate the database with currently available direct and indirect association data.

We use and integrate three datasets: the International HapMap project [7], Online Mendelian Inheritance in Man [6] – an on-line repository of articles about human genes and genetic disorders, and a dataset of marker correlations [2].

*to whom correspondence should be addressed

Our effort only partially overlaps systems such as the Genetic Association Database (GAD)[3] and the database of Genotype and Phenotype (dbGaP)[1]. GAD correlates different kinds of genetic data across a variety of data sources, including also the OMIM database. To the best of our knowledge, GAD has no facilities to use individual genotype data as part of a query. dbGaP focuses on archiving and distribution of results of studies that have investigated the interaction of genotype and phenotype. The system provides browsing and navigation facilities that can be customized on a per-study basis, by defining study-specific variables. dbGaP does not provide automatic integration of genome-wide marker correlation data with study results, and could be enhanced by integrating a system such as ours.

Our system is scalable and flexible: it operates in real time and can be configured on the fly to produce population, technology, and confidence-specific predictions. Efforts are underway to deploy our system as part of the NCBI Reference Assembly.

We discuss the design of our system in Section 2, describe some interesting findings and use scenarios Section 3, and conclude in Section 4.

2 METHODS

2.1 Data Representation

To address the first aim of our project, we designed a relational database schema for storage and querying of genotype and association data. We chose to use Oracle 10g, a commercial relational database system, in our implementation. A different relational platform may also be used to host a system such as ours; our decision to use Oracle is based primarily on query performance, and on availability of an XML package that we use for query result formatting.

The database schema is divided into three areas.

Allele frequency and marker correlation data

This is the most important area of the data repository, populated using the techniques described in Section 2.2. HapMap, marker correlation, and OMIM data is parsed and loaded into database tables for permanent storage. We then create *a materialized view* (a materialized logical table) that contains a transitive closure of direct and indirect associations between SNPs on whole genome array and OMIM records. Each row of this view includes information about population, technology, resolution, coefficient of determination, and the assumed source and target alleles. As new correlation data becomes available, it is loaded into permanent storage. The materialized view can then be re-built on demand; and takes less than 5 minutes on currently available data: the entire HapMap genome, marker correlation, and OMIM data.

Genotype and other data pertaining to an individual

If a user wishes to run association study wide or individual-specific diagnostic genotype queries, he must populate these tables. We provide command-line utilities for loading data in HapMap[7], IntraGenDB[4] and standard linkage formats.

Request and result tables

To initiate an association and diagnostic request, the user inserts an entry into the request table, specifying request parameters. Currently supported parameters are: population, technology, resolution, and coefficient of determination. Upon completion of the request, the system populates the result table with discovered SNP-disorder associations. Results may be displayed in XML format, or retrieved directly from the result table.

2.2 Data Processing and Integration

The HapMap dataset is a comprehensive repository of SNP genotypes. We use this dataset to compile population-specific lists of alleles and allele frequencies for the known SNPs. Our prediction dataset consists of single- and two-marker correlations; consequently, these are the types of correlation that our system supports. We provide a command-line utility for loading correlation data. Additional strategies for marker prediction may be accommodated by implementing similar load utilities.

Both HapMap and our marker correlation dataset are clean and non-redundant, available in comma-separated format, and lend themselves well to straight-forward processing. The challenge with these two datasets is the sheer volume of data and, consequently, the size of the full transitive closure of marker correlations. It turns out that while there is a lot of information regarding correlations along the genome (the marker correlation dataset is large), relatively little is still known about correlations between SNPs and heritable disorders. We observe that our system can take advantage of marker correlations only if they ultimately lead to a hypothesis of disease susceptibility, and use available marker-to-disorder data as the limiting factor. In other words, an correlation between SNP_1 and SNP_2 is only useful if at least one of these SNPs is associated with a heritable disorder.

We currently use OMIM, a repository of publications about human genes and genetic disorders, as our data source for marker to disorder associations. Associations between SNPs and diseases are not readily available and we resort to parsing this information from the text. We use a very simple method for this extraction: we process OMIM record by record, looking for occurrences of *rs* numbers (cross-references from OMIM to dbSNP), in the text of the article. We then assume that the mentioned SNP is associated with the heritable trait to which the current OMIM record pertains.

We made an assumption when associating SNPs with OMIM records: for lack of information, when OMIM mentions that an SNP is associated with a disorder, we assume that it refers to *the minor allele*.

There is nothing to prevent one from using more sophisticated Natural Language Processing techniques, or work with a different dataset that has genotype to phenotype associations readily available in machine-readable form. We provide parsing and data load utilities for OMIM. Similar utilities may be implemented for alternative datasets.

3 RESULTS

As mentioned earlier, our database contains a significant amount of SNP and marker correlation data, but only a limited number of SNP to OMIM associations. Across all populations and platforms we store over 10 million SNP records, close to 50 million single correlation records, and over 20 million double correlation records. However, out of 18000 articles in the OMIM repository, only 187 mention associations between heritable disorders and SNPs, with 133 unique participating SNPs. Combining OMIM with marker correlation data, we are able to make diagnostic predictions for additional 328 unique single-marker pairs, and 396 double-marker sets. The dataset is enriched with a total of 1312 population-specific correlations. The number of diagnostic predictions will grow as more information about direct associations between SNPs and heritable

disorders becomes available, i.e. when OMIM is extended, or a new data source becomes available.

For an example of the effectiveness of MutaGeneSys, consider age related macular degeneration (ARMD). According to OMIM, two SNPs are implicated in this disorder: rs3793784 in the ERCC6 gene and rs380390 in the CFH gene. MutaGeneSys contains 72 additional SNPs that are associated with the ARMD. As another example, Systemic Lupus Erythematosus (SLE) is associated with two CRP polymorphisms in OMIM; MutaGeneSys uses 15 additional unique SNPs to diagnose SLE.

MutaGeneSys is installed on a Pentium 4 with a 3GHz CPU and only 512MB of RAM, running Red Hat Enterprise Linux 4 and Oracle 10g Enterprise Edition. We tested the performance of our system on IntraGenDB[4] data, and found that diagnostic association requests run in interactive time: in under 5 seconds for full genotype scans of approximately 317,000 SNPs per DNA sample. This time does not include loading the genetic information into the database, which took an additional 60-90 seconds. Running our system on a faster machine with more RAM will likely further improve its performance.

4 CONCLUSION

We presented MutaGeneSys: a scalable system that uses genome-wide genotype data for disease prediction. MutaGeneSys allows detection of individuals susceptible to OMIM disorders among participants of whole genome association studies, a yet unexplored perspective of such data. In the longer run, we believe this system and its successors will pave the way for using whole genome SNP arrays as practical diagnostic tools, advancing them from bench to bedside.

ACKNOWLEDGMENTS

This material is based in part upon work supported by the National Institute of Health grant 1U54CA121852-01A1. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Institute of Health.

This material is based in part upon work supported by the National Science Foundation under Grant s IIS-0121239. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1]dbGaP: Genotype and phenotype. www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gap.

[2]Evaluating and improving power of whole genome products. www.cs.columbia.edu/ itsik/StandardGenotyping.htm.

[3]Genetic association database. geneticassociationdb.nih.gov.

[4]IntraGenDB population genetics database. intragen.cu-genome.org.

[5]Nature genetics: Question of the year. www.nature.com/ng/qoty.

[6]Online mendelian inheritance in men (OMIM). www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.

[7]Internaltional HapMap consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, 2005.

[8]de Bakker et al. Transferability of tag SNPs in genetic association studies in multiple populations. *Nature Genetics*, 38:1298–1303, 2006.

[9]Pe'er, de Bakker, Maller, Yelensky, Altshuler, and Daly. Evaluating and improving power in whole genome association studies using fixed marker sets. *Nature Genetics*, 38(6):663–7, 2006.