Finding Associations among SNPs for Prostate Cancer using Collaborative Filtering

Rohit Kugaonkar
University Of Maryland
Baltimore County
Baltimore,USA
rohitku1@umbc.edu

Aryya Gangopadhyay
University Of Maryland
Baltimore County
Baltimore,USA
gangopad@umbc.edu

Yelena Yesha
University Of Maryland
Baltimore County
Baltimore,USA
yeyesha@cs.umbc.edu

Anupam Joshi
University Of Maryland
Baltimore County
Baltimore,USA
joshi@cs.umbc.edu

Yaacov Yesha
University Of Maryland
Baltimore County
Baltimore,USA
yayesha@cs.umbc.edu

Michael A. Grasso
University Of Maryland School of Medicine
Baltimore,USA
mgrasso@umem.org

Mary Brady
NIST
Gaithersburg,USA
mary.brady@nist.gov

Napthali Rishe
Florida International University
Miami,USA
rishen@cis.fiu.edu

ABSTRACT

Prostate cancer is the second leading cause of cancer related deaths among men. Because of the slow growing nature of prostate cancer, sometimes surgical treatment is not required for less aggressive cancers. Recent debates over prostate-specific antigen (PSA) screening have drawn new attention to prostate cancer. Genome-based screening can potentially help in assessing the risk of developing prostate cancer. Due to the complicated nature of prostate cancer, studying the entire genome is essential to find genomic traits. Due to the high cost of studying all Single Nucleotide Polymorphisms (SNPs), it is essential to find tag SNPs which can represent other SNPs. Earlier methods to find tag SNPs using associations between SNPs either use SNP's location information or are based on data of very few SNP markers in each sample. Our study is based on 2300 samples with 550,000 SNPs each. We have not used SNP location information or any predefined standard cut-offs to find tag SNPs. Our approach is based on using collaborative filtering methods to find pairwise associations among SNPs and thus list top-N tag SNPs. We have found 25 tag SNPs which have highest similarities to other SNPs. In addition we found 16 more SNPs which have high correlation with the known high risk SNPs that are associated with prostate cancer. We used some of these newly found SNPs with 5 different classification algorithms and observed some improvement in prostate cancer prediction accuracy over using the original known high risk SNPs.

DTMBIO'12, October 29, 2012, Maui, Hawaii, USA.

Categories and Subject Descriptors

J.3 [**LIFE AND MEDICAL SCIENCES**]: Biology and genetics, Medical information systems; H.2.8 [**DATABASE MANAGEMENT**]: Database applications—*Data mining*

Keywords

Prostate cancer; SNP; SNP association; Collaborative filtering; tag SNPs

1. INTRODUCTION

Prostate cancer is the most common type of cancer found among the American male population. It is the second leading cause of cancer related deaths. Each year about 240,890 new cases of prostate cancer cases are detected [13]. Existing clinical tests for prostate cancer have limited accuracy, with the U.S. Preventive Services Task Force asserting that there are concerns that the harms of testing outweigh the benefits [7]. Also the decision to carry out these tests at an early age is dependent on the family history of a patient. Genome-based screening can potentially help in identifying individuals for which performing further testing is advisable. Single Nucleotide Polymorphisms (SNP) represent single base change in a DNA sequence. SNPs are the most common genetic variations found in the population. SNPs themselves do not cause any disease but help to determine the response to certain drugs and patient's susceptibility to develop a particular disease [6]. In order to study complex diseases such as cancer all of the SNPs need to be studied to find their associations with a particular disease. Genome-wide association studies carry out research on the entire genome and SNPs to find traits for major diseases. Such studies typically compare people with disease (cases) and people without the disease (controls), to see if a particular SNP is more common among cases and thus can be associated with a particular disease [3]. Even with rapid reduction in hardware costs, it is not possible to genotype all

the SNPs. [8] mentions cost and computational complexity as motivation for finding tag SNPs.

2. APPROACH

2.1 Collaborative Filtering (CF)

Collaborative filtering methods are popular on the online shopping websites like Amazon. Collaborative filtering methods are also applied to sensing and monitoring data, financial service institutions, advertising agencies, electronic commerce and web 2.0 applications. Collaborative filtering is also used to suggest research papers based on the users' research interest [4]. To the best of our knowledge, a collaborative filtering approach has not been used to find the associations between SNPs and thus to find the tag SNPs.

We have used a memory-based collaborative filtering algorithm on our entire dataset. In our case, user association is replaced by association among SNPs. Also items in our case refer to corresponding frequency counts of high risk, low risk and controls for each SNP-allele combination.

2.2 Similarity computation

We have used the Pearson correlation and vector cosine similarity computation methods in this paper [5].

2.2.1 *Pearson correlation based similarity computation*

This method provides the strength of the linear dependence between two users. The output value in the correlation metric are in the range of -1 to +1 where -1 indicates a perfectly negative correlation and +1 indicates a perfectly positive correlation. For this computation, normalization of the dataset using standard deviation and mean is required [11]. The Pearson correlation coefficient r between two data samples (X_i and Y_i) is given by following formula: $r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$. where, $\bar{X}$ and $\bar{Y}$ are the sample means.

2.2.2 *Cosine similarity computation*

The similarity between two vectors is given by measuring the cosine of the angle between them. The similarity values computed using this measure are in the range of -1 to +1 where +1 indicates that two vectors are most similar with each other [12]. Similarity using this measure between two vectors (X_i and Y_i) is given by the following formula:

Similarity= $cos\Theta = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i)^2}}$

3. METHODOLOGY

Algorithm 1 provides the pseudo code for our method. Figure 1 shows the system architecture and the following sections provide the detailed description of each step:

3.1 Data preprocessing and integration

Prostate cancer case (high-low) and control information are stored in a phenotype dataset. That information is mapped with the SNP, allele information in genotype using sample id field.

3.2 Compute frequency count

We found that there are approximately three different allele combinations for each SNP. We computed distinct SNP-allele combinations for each SNP. There are a total of 1,713,739 SNP-allele combinations. For each SNP-allele combination, we computed the frequency counts of high risk, low risk and controls. To fit into the existing analogy of user based collaborative filtering, in our case each SNP-allele combination represents users and the corresponding frequency counts are comparable to the ratings each user has given to different items. Using this step we have considerably reduced the computation by almost 1000 times from samples * SNPs(550,000*2300) to just SNP-allele* controls, low risk case, high risk case frequency count (1713,739*3).

Algorithm 1 Algorithm for clustering SNP-allele pairs based on similarity

Input: A set of n tuples <SNP, allele, case, control>
Output: k clusters: $\{C_1, C_2, \ldots, C_k\}$ (note: k is not an input parameter)

1: Group all SNP-allele pairs (SA) and their corresponding case/control counts
2: **for** $i = 1 \rightarrow n-1$ **do**
3: **for** $j = i+1 \rightarrow n$ **do**
4: Compute similarity (SA_i, SA_j);
5: **end for**
6: Merge i with the most similar node j^*;
 If j^* is already merged with a cluster $C = SA_1, \ldots, SA_l$, add i to cluster C;
7: **end for**

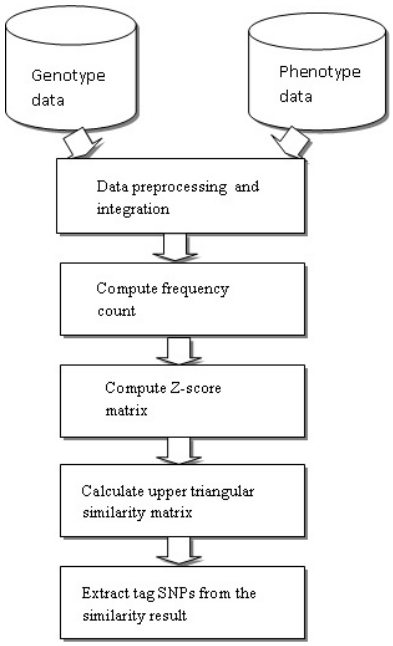

Figure 1: System architecture

3.3 Compute Z-score matrix

This step is not required if we are working with a cosine similarity matrix. For Pearson correlation, before computing similarity between two vectors it is necessary to normalize the sample data, otherwise the correlation will be biased towards variables with higher ranges in sample data. To normalize sample data, we computed the z-score matrix using mean and standard deviation of each vector.

3.4 Calculate similarity matrix

Here we computed similarities over the matrix of 1,713,739 rows (SNP-allele) combination and 3 columns (controls, low risk case and high risk cases) using the Pearson correlation and cosine similarity matrices. For Pearson correlation we used the Z-score matrix computed in the previous step. Ideally each SNP-allele vector needs to be compared with every other SNP-allele vector to find the similarity between two vectors but since the similarity measure is undirected it is same in both directions. The similarity matrix is symmetric which reduces the total computation time by half.

3.5 Extract tag SNPs

This is the final step of our process. In this step, we extract top-N SNP-allele combinations based on their frequency in the similarity matrix. These SNP-alleles represent centroids in the clusters. Since these SNPs have higher similarity with all other SNPs (similarity count), they can be used to represent all these SNPs and thus act as tag SNPs.

4. EXPRIEMENTAL SETUP

4.1 Dataset details

For our experiment we have used the Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer genome-wide association study provided by dbGaP(data base of Genotype and Phenotype) [1]. For this dataset, genotyping (Illumina platform) of 554,291 SNPs is performed and is included in the study. This dataset contains 1,172 prostate cancer patients and 1,157 control patients of European ancestry. Of 1,172 prostate cancer patients 737 are classified as high risk patients and 493 are classified as low risk patients. Aggressiveness of cancer is determined using Gleason score (>= 7) and the stage of prostate cancer (stage 3 or above). The dataset contains only family history and age as phenotype parameters. It does not contain any clinical parameters such PSA test results. Genotype data contains bi-allelic data for each SNP.

4.2 Hardware and software

We used an AMD Opteron machine with 47 processors (12 cores each) and 504 gigabytes of physical memory for our experiments. We used a Mysql server for storing the dataset. The source code is written in MATLAB.

5. RESULTS

5.1 Top-N tag SNPs

We found top-25 tag SNPs from our methodology. We used Pearson correlation based similarity computation for finding these tag SNPs. Table 1 shows top 25 tag SNPs along with their corresponding similarity count. Also a graph indicating 25 tag SNPs along with its subset of similar SNPs is shown in Figure 3. In Figure 3, each node represents a tag SNP. The nodes are colored based on their degrees(similarity count) as follows: the degrees of blue, magenta, green, yellow, maroon, and red nodes are 2, 5, 7, 8, 10, and 11 respectively.

5.2 Other associated SNPs

We found 16 high risk SNP-alleles related to prostate cancer from literature and SNPedia [2]. Through our cosine similarity computations, we found other SNPs which are highly similar (cosine similarity > 0.9) to those 16 SNPs. Table 2 shows these 16 original high risk SNP-alleles along with their associated SNP-alleles and cosine similarity measures.

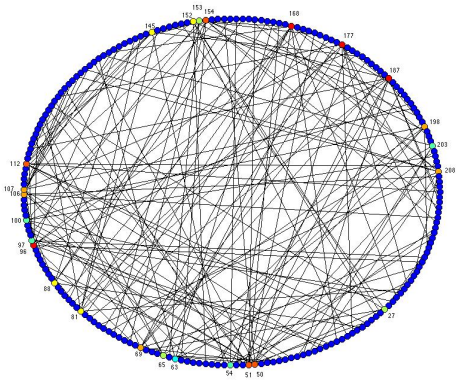

Figure 2: Subgraph of association among SNPs

SNP-allele	Similarity count
rs6632302-TT	12959
rs4077503-AA	8952
rs7700289-CC	8760
rs482877-CC	7460
rs17342020-AA	7138
rs7604484-TT	5636
rs364641-AA	5446
rs9823690-GG	4552
rs11688574-TG	3886
rs7760456-GG	3608
rs2554644-TT	3579
rs1924444-TC	3512
rs359045-AA	3465
rs6942067-AA	3317
rs506776-CC	3164
rs726730-TC	3155
rs9650108-AA	3039
rs7896195-CC	2671
rs8072737-TT	2623
rs305578-CC	2615
rs1427602-AG	2455
rs1552324-AC	2381
rs1401861-GG	2279
rs3734838-GG	2268
rs12509926-TC	2218

Table 1: 25 tag SNP-alleles

5.3 Classification using newly found SNPs

To check the usage of these newly found SNPs using collaborative filtering, we applied an information gain algorithm on all 32 SNPs (16 previously found prostate cancer risk SNPs and 16 newly found SNPs). We used a ranking method in information gain attribute selection. We have selected first 16 SNPs according to the information gain output. We found that among 16 newly found SNPs rs4775919, rs1544872, rs8111157 and rs1998641 are useful and they are associated with rs1859962, rs1447295, rs7652331 and rs10492519 previously found risk SNPs respectively. We applied 5 classification algorithms on these 4 newly found SNPs and 12 previously found risk SNPs. Table 3 shows comparision of prediction accuracy between previously found 16 risk SNPs and 12 previously found and 4 newly found SNPs using the collaborative filtering approach.

Original high risk SNP-allele	Associated SNP-allele	Similarity
rs4430796-AA	rs9525604-TC	0.9923
rs1859962-GG	rs4775919-AG	0.9745
rs6983267-GT	rs9588748-GG	0.9823
rs1447295-AC	rs1544872-TT	1.0000
rs1571801-AA	rs789950-GG	0.9978
rs2107301-TT	rs482355-TC	0.9145
rs1545985-AG	rs4235534-CC	0.9145
rs7652331-TT	rs8111157-AG	0.9951
rs629242-CT	rs7585535-GG	1.0000
rs13149290-CC	rs392306-CC	0.9874
rs251177-CT	rs7878588-TT	0.9733
rs10492519-AG	rs1998641-GG	0.9311
rs5945572-AA	rs7935166-AA	0.9210
rs1456315-AG	rs4242474-CC	0.9921
rs4054823-TT	rs4520319-AG	0.9654
rs4242382-AA	rs4242474-CC	0.9321

Table 2: prostate cancer high risk associated SNPs

Classification algorithm	Accuracy with 16 risk SNPs	Accuracy with 4 newly found SNPs and 12 risk SNPs
Naive bayes	60%	62.44%
Bayesnet	59.778%	62.22%
SMO-polykernel	57.778%	60.8889%
SMO-RBF	56.4444%	61.333%
J48	57.1006%	58.87%

Table 3: Improvement in prediction accuracy with 4 newly found SNPs

We found 4 newly found SNPs, which were not previously known to be associated with prostate cancer, and which we used to increase the accuracy of our prediction algorithm. It is known that rs4775919 is associated with psoriatic arthritis [9] and rs1544872 is associated with type 2 diabetes [10]. However, their association with prostate cancer is not known at this time. More clinical and molecular research is needed for these 4 SNPs to clarify their association with prostate cancer.

6. CONCLUSIONS

We have described a novel method for finding association among a very large number of SNPs without using any gene location information. We have used a collaborative filtering methods to find the association among SNPs. To find the similarity between each pair of SNPs, we used Pearson correlation and Cosine similarity measure. We found 25 tag SNPs which have highest similarity with other SNPs in the dataset. Altogether approximately 400 SNPs can be used to replace other 1713,739 SNPs. Researchers can use these SNPs to facilitate research in prostate cancer and other multi-genetic diseases, and thus reduce the overhead of working on the entire genome. This approach can be easily extended for top-N SNPs. Also we found high similarity (> 0.9) 16 other SNPs which are associated with the previously found SNPs mentioned in the literature. Out of the 16 newly found SNPs, we used 4 SNPs for classification and achieved some improvement in prediction accuracy. We have used our approach on the entire dataset of almost 2300 samples with 550,000 SNPs per sample. Our method can be easily generalized to any large SNP dataset. Our algorithm is scalable for a large dataset with parallelization.

7. ACKNOWLEDGMENTS

The authors would like to thank Database of Genotype and Phenotype (dbGaP) for providing a dataset for prostate cancer research. We would also like to thank the Muticore Computer Center for providing the required resources.

8. REFERENCES

[1] *dbGaP CGEMS Prostate Cancer dataset.* <http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/variable.cgi?study_id=phs000207.v1.p1&phv=82615&phd=&pha=2877&pht=1105&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1 >

[2] *Prostate Cancer SNPs.* SNPedia. August 17 2011. Retrieved May 28, 2012 from < http://www.snpedia.com/index.php/Prostate_cancer>

[3] *Genome–Wide Association Studies.* U.S. Department of Health and Human Services. Retrieved May 28, 2012 from < http://gwas.nih.gov/>

[4] Xiaoyuan Su, Taghi M. Khoshgoftaar, *A Survey of Collaborative Filtering Techniques,* Advances in Artificial Intelligence, 2009.

[5] Anna Huang, *Similarity Measures for Text Document Clustering,* NZCSRSC 2008.

[6] *Introduction to Genetics.* 23andme. Retrieved May 28, 2012 from < https://www.23andme.com/>

[7] Otis W. Brawley, *Prostate Cancer Screening: What We Know, Don't Know, and Believe,* Annals of Internal Medicine, May 2012.

[8] Irina Astrovskaya, Alex Zelikovsky, *Genotype Tagging with Limited Overfitting,* Advances in Bioinformatics and Computational Biology, Lecture Notes in Computer Science, 2009, Volume 5676/2009.

[9] Liu et. al., *A Genome-Wide Association Study of Psoriasis and Psoriatic Arthritis Identifies New Disease Loci,* PLoS Genetics, March 2008, Volume 4 , Issue 3.

[10] Jukka Salonen, Jelena Hypponen, Jari Kaikkonen, Mia Pirskanen,Pekka Uimari, Juha-Matti Aalto, *Novel Genes and Markers in Type 2 Diabetes and Obesity,* Patent Cooperation Treaty, Publication number. WO/2007/128884.

[11] J. L. Rodgers, W. A. Nicewander, *Thirteen ways to look at the correlation coefficient,* The American Statistician, Vol. 42, No. 1. (Feb., 1988), pp. 59-66.

[12] P.N. Tan, M. Steinbach, V. Kumar, newblock *Introduction to Data Mining,* newblock Addison-Wesley (2005), chapter 8.

[13] *Prostate cancer Overview.* American Cancer Society March 9,2012, Retrieved May 28, 2012 from <http://www.cancer.org/acs/groups/cid/documents/webcontent/003072-pdf.pdf >