Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Multiple Linear Regression for Index SNP Selection on Unphased Genotypes

Jingwu He and Alex Zelikovsky

Abstract— **The search for the association between complex diseases and single nucleotide polymorphism (SNPs) or haplotypes has recently received great attention. Recent successes in high throughput genotyping technologies drastically increase the length of available SNP sequences. This elevates the importance for the use of a small subset of informative SNPs, called index SNPs [21], accurately representing the rest of the SNPs (i.e., the rest of the SNPs can be highly predicted from the index SNPs). Index SNP selection achieves the compaction of huge unphased genotype data (obtained, e.g., from Affimetrix Map Array) in order to make feasible fine genotype analysis.**

In this paper we propose a novel index SNP selection on unphased genotypes based on multiple linear regression (MLR) SNP prediction. We measure the quality of our index SNP selection algorithm by comparing actual SNPs with the SNPs computationally predicted from chosen index SNPs. We obtain an extremely good prediction rates and compression. For example, for region ENm010 (123 SNPs) [9], we can use 2% of SNPs for representing all SNPs with 93.5% accuracy. An experimental study on 4 ENCODE regions from HapMap [9] shows that our method uses significantly fewer index SNPs (e.g., up to two times less index SNPs to reach 90% prediction accuracy) than the state-of-the-art method of Halperin et al. [8] for genotypes.

Keywords: Single nucleotide polymorphism, Index SNPs, Multiple linear regression.

I. Introduction

The search for the association between complex diseases and single nucleotide polymorphism (SNPs) or haplotypes has recently received great attention. Recent successes in high throughput genotyping technologies drastically increase the length of available SNP sequences. This elevates the importance for the use of a small subset of informative SNPs accurately representing the rest of the SNPs (i.e., the rest of the SNPs can be highly predicted from the index SNPs) thus achieving the compaction of huge unphased genotype data (obtained, e.g., from Affimetrix Map Array) in order to make feasible fine genotype analysis.

Johnson et al.[10] have referred such informative SNPs on haplotypes as *tag SNPs*. Note tag SNPs are commonly referred as informative SNPs on haplotypes. Following [21], We prefer to use the more general term *index SNPs* for referring informative SNPs either on haplotypes or genotypes. An established way of selecting index SNPs [4], [2], [18], [15], [20] is based on linkage disequilibrium (LD) and partition of the entire SNP sequence into blocks, i.e., contiguous SNP segments within which the number of different haplotypes is comparatively small. Due to the low diversity within a block, a small number of index SNPs can predict values of all other SNPs in the same block. An alternative way is to ignore block structure and select index SNPs across the whole region (e.g. [3]). However, predicting other non-index SNPs from its index SNPs on genotypes has received much less attention. Recently, Halperin et al. [8] described a new method for predicting SNPs from the index SNPs on genotypes for maximizing the prediction accuracy.

Following [11], [13], [8], we can formulate the Index SNP Selection Problem (ISSP) as follows:

Index SNP Selection problem (ISSP). Given a sample S of a population P of *individuals* (either haplotypes or genotypes) on m SNPs, select positions of k ($k < m$) SNPs such that for any individual one can predict non-selected SNPs from its k selected SNPs. Our method solves the optimization version of ISSP which asks for k index SNPs *maximizing the prediction accuracy* measured by the percentage of the number of correctly predicted SNPs.

Human SNPs belong two near-identical copies of each chromosome (haplotypes). Most experimental techniques generate for each site an unordered pair of allele readings, one from each haplotype, which is called a genotype. *Phasing*, or splitting a genotype into two haplotypes is usually imputed computationally. The selection of index SNPs can be done either before or after phasing. In this paper we propose a novel index SNP selection on unphased genotypes based on multiple linear regression (MLR) SNP prediction [17]. The MLR method predicts the non-index SNP s so that the predicted s is the closest to its projection on the span of vectors corresponding to index SNPs. The MLR method accumulates information about *all* index SNPs resulting in significantly higher prediction accuracy with the same number of index SNPs than other prediction methods. The previous SNP prediction methods rely either on a single SNP (see, e.g., [4]), a closest pair of index SNPs [8], or small number of index SNPs from the block with limited haplotype diversity [20]. Traditional SNP prediction methods (see, e.g., [3], [20], [8]) use simple majority voting for prediction. Our previous linear reduction tag selection (LR) [11], [13] picks linear independent tag SNPs but cannot easily handle bounds on *prediction error* or number of tags. We measure the quality of index SNP selection based on MLR by measuring the percentage of correctly predicted SNPs. We obtain an extremely good prediction rates and compression.

Jingwu He is supported by Georgia State University Molecular Basis of Disease Fellowship.

Alex Zelikovsky is partially supported by NIH Award 1 P20 GM065762-01A1.

Jingwu he is a PH.D student, Computer Science, Georgia State University, Atlanta, GA, 30318 jingwu@cs.gsu.com

Alex Zelikovsky is with Faculty Computer Science, Georgia State University, Atlanta, GA, 30318 alexz@cs.gsu.edu

For example, for region ENm010 (123 SNPs) [9], we can use 2% of SNPs for representing all SNPs with 93.5% accuracy. An experimental study on 4 ENCODE regions from HapMap [9] shows that our method uses significantly fewer index SNPs (e.g., up to two times less index SNPs to reach 90% prediction accuracy) than the state-of-art method of Halperin et al. [8] for genotypes. Moreover, Our program gives users a freedom to select number of index SNPs. The number of index SNPs selected by the users depends on the complexity of fine genotype analysis.

The rest of the paper is organized as follows. In the next section we describes our MLR SNP prediction algorithm and stepwise index SNP selection. Section III presents an experimental results and discussions.

II. Multiple linear regression SNP prediction and stepwise index SNP selection

In this section we first give an intuition of multiple linear regression SNP prediction and describe the prediction and selection algorithms in details.

A. SNP prediction based on multiple linear regression

The index SNPs are selected based on the sample population with an intention to derive conclusions about the entire population. Statistical analysis may ensure that high prediction quality of non-index SNPs is not a coincidence. If certain SNPs are highly correlated (i.e. in linkage disequilibrium) in the sample, then we would expect that this correlation will be observed in the entire population. Therefore, it would be highly desirable that the index SNPs taking part in predicting non-index SNP should correlate with the non-index SNP. Therefore, haplotype tags have been selected based on the squared correlation R^2 between true and predicted SNPs [6] and true and predicted halotype dosage in [19]. The general purpose of multiple linear regression is to learn the relationship between several independent variables and a dependent variable. The regression coefficients represent the independent contributions of each independent variable to the prediction of the dependent variable. Multiple linear regression method implicitly relies the correlation between SNPs. If the index SNP (or SNPs) and a predicted SNP are not correlated, the regression coefficients will not contribute to the prediction of SNP.

Below, we describe how our MLR algorithm predicts the value of a single unknown non-index SNP s in the genotype g (each unknown SNP in g is separately predicted) based on the values of all other SNPs in g (i.e., k index SNPs) and the values of *all* SNPs in the sample genotypes. Usually, a genotype is represented by a vector with coordinates 0,1, or 2, where 0 represents the homozygous site with major allele, 1 represents the homozygous site with minor allele, and 2 represents the heterozygous site. The sample population S together with the index-restricted individual x are represented as a matrix M. The matrix M has $n+1$ rows corresponding to n sample individuals and the individual x and $k+1$ columns corresponding to k index SNPs and a single non-index SNP s. All values in M are known except the value of s in x. For genotypes, there are 3 possible resolutions s_0, s_1, and s_2 corresponding to SNP values 0, 1, or 2, respectively. The SNP prediction method should chose correct resolution of s.

The proposed MLR SNP prediction method considers all possible resolutions of s together with the set of index SNPs T as the vectors in $(n+1)$-dimensional Euclidean space. It assumes that the most probable resolution of s should be the "closest" to the spanning space of T. The distance between resolution of s and T is measured between s and its projection on the vector space $span(T)$, the span of the set of index SNPs T (see Figure 1).

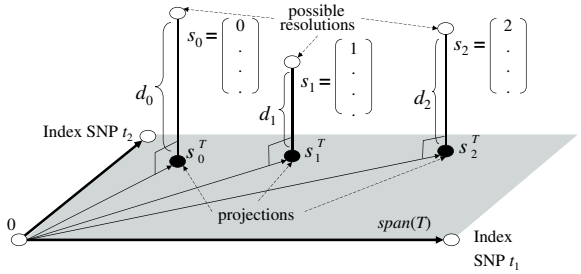

Fig. 1. MLR SNP Prediction Algorithm. Three possible resolutions s_0,s_1, and s_2 of s are projected on the span of index SNPs (a dark plane). The unknown SNP value is predicted 1 since the distance between s_1 and its projection s_1^T is the shorter than for s_0 and s_2.

Computationally, the distance between a resolution s_i and T is measured as $dist(T, s_i) = |T \cdot (T^t \cdot T)^{-1} \cdot T^t \cdot s_i - s_i|$. The runtime of the MLR SNP prediction algorithm is $O(kn^2)$. In general, there are $m-k$ non-index SNPs in each individual but the matrix $T \cdot (T^t \cdot T)^{-1} \cdot T^t$ is the same for all these non-index SNPs and should be computed only once. Thus, the total runtime for predicting a complete individual is $O(kn(n+m))$.

As noted in [2], one cannot straightforwardly apply linear combinations of column-SNPs since *equivalent* columns can be linearly independent. This problem has been overcome by a change in notation [12]: the former $\{0, 1, 2\}$ notation becomes $\{-1, 1, 0\}$. Such change of notations allows a genotype g to become a linear combination of two haplotypes h and h', i.e., $g = \frac{h+h'}{2}$.

B. Stepwise index SNP selection

Given number k of index SNPs to be selected, it is better to choose the k index SNPs minimizing number of errors when predicting the non-index SNPs in the sample. Unfortunately, the exhaustive search is prohibitively slow. We propose to apply the following *Stepwise Index SNP Algorithm* (SISA). SISA starts with the best index SNP t_0, i.e., the SNP minimizing the error when predicting alone all other SNPs. Then SISA finds such index SNP t_1, which would be the best extension of $\{t_0\}$ for predicting other SNPs with minimal error, and continues adding best index SNPs until reaching the set of index SNPs of the given size k. It is a greedy algorithm, however statisticians use term *stepwise* instead [19]. The runtime of SISA is $O(m^3(n+m)^2)$. Note

that SISA produces a *hereditary* set of index SNPs, i.e., the chosen k index SNPs contain the chosen $k-1$ index SNPs. The hereditary property of chosen index SNPs allows to extend without retyping the set of index SNPs in case of obtaining additional funding.

III. Experiment Results

The following genotype datasets of 4 ENCODE regions from HapMap [9] are used to measure the quality of our SNP prediction and index SNP selection algorithms as well as comparison with the results of [8]. We assume the data are completed. Before using our application, the missing data should be imputed. We use GERBIL algorithms [7] for resolving missing data.

ENr123 from 45 Han Chinese from Beijing (HCB) and 44 Japanese from Tokyo (JPT). The total number of SNPs are 63 after cleaning the SNPs who have only one allele.

ENm010 from 45 Han Chinese from Beijing (HCB) and 44 Japanese from Tokyo (JPT). The total number of SNPs are 123 after cleaning the SNPs who have only one allele.

We report the prediction accuracy measured by the percentage of correctly predicted non-index SNPs. Table I compares MLR, STAMPA, and LR on prediction accuracy and running time by using the number of index SNPs (2, 5, 10, 15, 20, 25) on 4 ENCODE regions. Figure 2 shows the prediction accuracy as a function of number of index SNPs on ENr123-HCB and ENm010-HCB regions. We obtain an extremely good prediction rates and compression. For example, for region ENm010-HCB (123 SNPs), we can use 2% of SNPs for representing all SNPs with 93.5% prediction accuracy, while STAMPA needs 5% of all SNPs for reaching same prediction accuracy. MLR is faster than STAMPA when few number of index SNPs are required. All experiments are performed on a computer with Intel Pentium 4, 3.06Ghz processor and 2 GB of RAM.

IV. Conclusions and Future Work

We have suggested a multiple linear regression method for index SNP selection. Our method obtains an extremely good prediction rates and compression. For example, for region ENm010 (123 SNPs) [9], we can use 2% of SNPs for representing all SNPs with 93.5% accuracy. In the future, we will apply multiple linear regression method for disease association.

References

[1] Affymetrix (2005) *http://www.affymetrix.com/products/arrays/*.

[2] Avi-Itzhak, H.I., Su, X. and de la Vega, F.M. (2003) 'Selection of minimum subsets of single nucleotide polymorphism to capture haplotype block diversity', *Proceedings of Pacific Symposium on Biocomputing*, Vol. 8, pp. 466–477.

[3] Bafna, V., Halldorsson, B.V., Schwartz, R.S., Clark, A.G. and Istrail, S. (2003) 'Haplotypes and informative SNP selection algorithms: don't block out information', *Proceedings of the Seventh International Conference on Research in Computational Molecular Biology*, pp. 19–27.

[4] Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004) 'Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium', *American Journal of Human Genetics*, Vol. 74, No. 1, pp. 106–120.

[5] Clark, A. (2003) 'Finding genes underlying risk of complex disease by linkage disequilibrium mapping', *Current Opinion in Genetics & Development*, Vol. 13, No. 3, pp. 296–302.

[6] Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003) 'Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power', Human Heredity, Vol. 56, pp. 18–31.

[7] Kimmel, G., and Shamir R.(2004). 'GERBIL: Genotype resolution and block identification using likelihood', *PNAS*, Vol. 102, pp 158–162.

[8] Halperin, E., Kimmel, G. and Shamir, R. (2005) 'Tag SNP Selection in Genotype Data for Maximizing SNP Prediciton Accuracy', *Bioinformatics*, Vol. 21, pp. 195-203.

[9] International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796, *http://www.hapmap.org*.

[10] Johnson, L., Esposito, L., Barratt, BJ., Smith, AN., Heward J, Genova GD, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RCJ, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SCL, Clayton DG, Todd JA (2001) 'Haplotype tagging for the identification of common disease genes'. *Nature Genet* **29**, 233–237.

[11] He, J. and Zelikovsky, A. (2004) 'Linear Reduction Methods for Tag SNP Selection', *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology (EMBC'04)*, pp. 2840–2843.

[12] He, J. and Zelikovsky, A. (2004) 'Linear Reduction for Haplotype Inference', *Proceedings of the Workshop on Algorithms in Bioinformatics (WABI'04)*, Vol. 3240, pp. 242–253.

[13] He, J., Westbrooks, K. and Zelikovsky, A. (2005) 'Linear Reduction Method for Predictive and Informative Tag SNP Selection', *International Journal Bioinformtics Research and Applications*, Vol. 3, pp. 249-260.

[14] Huang, Y.H., Zhang, K., Chen, T. and Chao, K.M. (2004) 'Approximation Algorithms for the Selection of Robust Tag SNPs', *Proceedings of the Workshop on Algorithms in Bioinformatics (WABI'04)*, Vol. 3240, pp. 278–289.

[15] Patil, N., Berno, A., Hinds, D., Barrett, W., Doshi, J., Hacker, C., Kautzer, C., Lee, D., Marjoribanks, C., McDonough, D., Nguyen, B., Norris, M., Sheehan, J., Shen, N., Stern, D., Stokowski, R., Thomas, D., Trulson, M., Vyas, K., Frazer, K., Fodor, S. and Cox, D. (2001) 'Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome', *Science*, Vol. 294, pp. 1719–1723.

[16] Sebastiani, P., Lazarus, R., Weiss, S., Kunkel, L., Kohane, I., and Ramoni, M. (2003) 'Minimal haplotype tagging', *Proceedings of the National Academy of Sciences*, Vol. 100, pp. 9900–9905.

[17] StatSoft, Inc. (1999) Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/stathome.html.

[18] R. Judson, B. Salisbury, J. Schneider, A. Windemuth, and J. C. Stephens. How many SNPs does a genome-wide haplotype map require? Pharmacogenomics, 3:379–391, 2002.

[19] Stram, D., Haiman, C., Hirschhorn, J., Altshuler, D., Kolonel, L., Henderson, B. and Pike, M. (2003) 'Choosing haplotype-tagging SNPs based on unphased genotype data using as preliminary sample of unrelated subjects with an example from the multiethnic cohort study', *Human Heredity*, Vol. 55, pp. 27–36.

[20] Zhang, K., Qin, Z., Liu, J., Chen, T., Waterman, M., and Sun, F. (2004) 'Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies', *Genome Research*, Vol. 14, pp. 908–916.

[21] Zhang P., Sheng H. and Uehara R. (2004) A double classification tree search algorithm for index SNP selection, *BMC Bioinformatics*, Vol. 5, pp. 89–95

TABLE I

THE COMPARISON OF MLR AND STAMPA ON PREDICTION ACCURACY AND RUNNING TIME BY USING THE NUMBER OF INDEX SNPS (2, 5, 10, 15, 20, 25) ON 4 ENCODE REGIONS. TOTAL NUMBER OF SNPS IN EACH DATASET IS IN THE PARENTHESIS.

Data	Measure	Methods	2	5	10	15	20	25
ENr123-HCB (63)	prediction accuracy	MLR	0.803	0.928	0.981	0.992	0.998	0.999
		STAMPA	0.744	0.903	0.937	0.952	0.960	0.969
		LR	0.654	0.742	0.812	0.874	0.907	0.923
	running time (s)	MLR	0.247	0.633	1.893	3.798	6.345	9.357
		STAMPA	4.109	4.136	4.180	4.267	4.385	4.425
ENr123-JPT (63)	prediction accuracy	MLR	0.814	0.938	0.980	0.994	0.998	1
		STAMPA	0.792	0.909	0.953	0.968	0.981	0.986
ENm010-HCB (123)	prediction accuracy	MLR	0.935	0.955	0.968	0.979	0.989	0.995
		STAMPA	0.895	0.938	0.956	0.960	0.966	0.969
		LR	0.753	0.791	0.832	0.894	0.912	0.938
	running time (s)	MLR	0.763	1.388	3.978	10.308	13.345	15.357
		STAMPA	3.451	3.462	3.652	3.655	3.852	4.425
ENm010-JPT (123)	prediction accuracy	MLR	0.911	0.959	0.972	0.985	0.990	0.997
		STAMPA	0.903	0.942	0.959	0.968	0.977	0.989

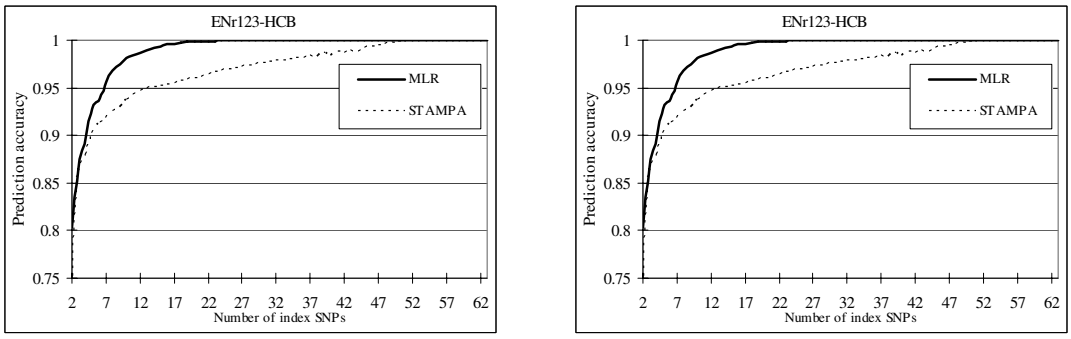

Fig. 2. Prediction accuracy as a function of number of index SNPs on two datatsets: (Left) ENr123-HCB and (Right) ENm010-HCB.