

Computational Approaches to Detecting Gene-Gene Interactions

ISMB'03 Tutorial Proposal

Jason H. Moore

Assistant Professor

Director, Bioinformatics Core

Program in Human Genetics

Vanderbilt University, Nashville, TN, USA

Tutorial Proposal

We have entered an era in genomics and genetic epidemiology in which the generation of genetic information is far outpacing our ability to understand its implications for the prevention, diagnosis, and treatment of common multifactorial human diseases. New technologies such as DNA microarrays (Schena et al. 1995) facilitate measurements of thousands of DNA sequence variations and the gene expression levels of thousands of genes in epidemiological study designs. The availability of vast quantities of genetic information has presented genetic epidemiologists and computational biologists with several important computational and statistical challenges. The first of these challenges is the *variable selection problem*. This problem stems from the growing realization that interactions among multiple genetic and environmental factors are likely to be more important than any one factor for predicting risk of common multifactorial diseases such as essential hypertension (Kardia 2000; Moore and Williams 2002) and sporadic breast cancer (Ritchie et al. 2001). Given interactions play an important role in disease etiology, we need to be considering combinations of genetic variations or gene expression variables rather than one variable at a time in our genetic analysis. The problem is that, when the number of variables is large, there are an effectively infinite number of combinations that could be evaluated. For example, with 10,000 single nucleotide polymorphisms or gene expression variables there are approximately $5 * 10^7$ possible subsets of size two, $1.7 * 10^{11}$ possible subsets of size three, and $4.2 * 10^{14}$ possible subsets of size four. Clearly, the combinatorial magnitude of variable selection precludes an exhaustive search of all possible variable subsets.

The second challenge that genetic epidemiologists and computational biologists face is the *statistical modeling problem*. That is, what is the most appropriate way to model the relationship between combinations of genetic variations or gene expression variables and clinical endpoints? One traditional approach to modeling the relationship between genetic variations or gene expression variables and discrete clinical outcomes is logistic regression (Hosmer and Lemeshow 2000). Logistic regression is a parametric statistical approach for relating one or more independent or explanatory variables to a dependent or outcome variable (e.g. disease status) that follows a binomial distribution. However, as reviewed by Moore and Williams (2002), the number of possible interaction terms grows exponentially as each additional main effect is included in the logistic regression model. Thus, logistic regression, like most parametric statistical methods, is limited in its ability to deal with

interaction data involving many simultaneous factors. Alternative approaches such as multifactor dimensionality reduction (Ritchie et al. 2001, 2003; Hahn and Moore 2003), cellular automata (Moore and Hahn 2002a, 2002b), and symbolic discriminant analysis (Moore et al. 2001; 2002d) are flexible, nonparametric, and genetic model-free but at the cost of ease of computation. In contrast to logistic regression, the functional form of these models must be optimized. As with the variable selection problem, there are an effectively infinite number of possible model forms.

The goals of this tutorial are to

- 1) review the importance of gene-gene and gene-environment interactions for understanding the etiology of common multifactorial diseases such as essential hypertension and sporadic breast cancer,
- 2) review the variable selection and statistical modeling problems for the detection and characterization of gene-gene and gene-environment interactions, and
- 3) review new computational approaches for dealing with these challenges.

The tutorial will focus on new data reduction and pattern recognition approaches such as multifactor dimensionality reduction (Ritchie et al. 2001, 2003; Hahn and Moore 2003), combinatorial partitioning (Nelson et al. 2001; Moore et al. 2002a; 2002b), and cellular automata (Moore and Hahn 2002a, 2002b) for identifying interactions among genetic variations as well as approaches such as symbolic discriminant analysis (Moore et al. 2001; 2002) and dynamics based pattern recognition (Parker and Moore 2001) for identifying interactions among gene expression variables.

Algorithms such as evolutionary computation (Moore and Parker 2001) will be reviewed as approaches for optimizing variable and feature selection. Additionally, we will review new methods for simulating complex gene-gene interactions for the evaluation of new analytical methods (Moore et al. 2002c).

This tutorial will be important for anyone interested in using measures of genetic variation or gene expression for understanding the prevention, diagnosis, and treatment of common, complex multifactorial diseases. Since this is such a new area, the only assumptions made are that participants understand the technologies used to generate the data (e.g. DNA microarrays) and have a fundamental understanding of data analysis. This tutorial would be too introductory only to those actively developing new methods for detecting and characterizing the effects of interactions between multiple genetic and environmental factors on risk of common diseases.

Evidence of a History of Outstanding Teaching Skills

I have a solid record of teaching for graduate level courses at the University of Michigan and Vanderbilt University as well as tutorials for the annual Genetic Analysis of Complex Human Diseases workshop at Duke University

(<http://wwwchg.mc.duke.edu/geneticcourse/>) and Vanderbilt University (<http://phg.mc.vanderbilt.edu/gachd.htm/>). The last three years I have given lectures on biostatistics, population genetics, and quantitative genetics for our graduate level human genetics class (<http://phg.mc.vanderbilt.edu/mpbmain.shtml>). Last year I co-directed a year-long tutorials in physiology course. Further, this year I organized and taught a six-lecture introduction to biostatistics for the biomedical graduate students.

Perhaps the most relevant to the ISMB tutorial is the Genetic Analysis of Complex Human Diseases workshop at Duke University. I have participated in this workshop the last three years giving tutorials on biostatistics, quantitative genetics, computational approaches to genetic analysis, and microarray data analysis. I have consistently received excellent student feedback and have been among the highest rated faculty at the course. The following are actual quotes from students that have taken the course:

“Dr. Moore provided one of the best explained approaches to data reduction approaches that I have ever heard.” –2001 Genetic Analysis of Complex Human Diseases workshop, Duke University

“All instructors were effective, however, Scott, Moore and Martin did a really good job.” –2001 Genetic Analysis of Complex Human Diseases workshop, Duke University

As further evidence of my teaching ability, I presented a tutorial on “Non-Traditional Approaches for the Analysis of High-Dimensional Genetic Data” at the 2002 Joint Statistical Meeting and have been invited to present a three-hour tutorial on bioinformatics and genetics at the 2003 Pacific Symposium on Biocomputing.

References

Hahn L.W., Moore J.H. Multifactor dimensionality reduction ideally discriminates between discrete clinical endpoints using multilocus genotypes, submitted (2003).

Hahn, L.W., Ritchie, M.D., and Moore, J.H. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics, in press (2003).

Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: John Wiley & Sons, Inc. (2000).

Kardia SLR. Context-dependent genetic effects in hypertension. Current Hypertension

Reports 2: 32-38 (2000).

Moore, J.H. and Hahn, L.W. A cellular automata approach to detecting interactions among single-nucleotide polymorphisms in complex multifactorial diseases. Pacific Symposium on Biocomputing, 2002: 53-64 (2002a).

Moore, J.H., Hahn, L.W. Cellular automata and genetic algorithms for parallel problem solving in human genetics. In: Merelo, J.J., Panagiotis, A., Beyer, H.-G. (eds) Lecture Notes in Computer Science 2439, pp 821-830, Springer-Verlag, Berlin (2002b).

Moore, J.H., Hahn, L.W., Ritchie, M.D., Thornton, T.A., White, B.C. Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics. In: W.B. Langdon, E. Cantu-Paz, K. Mathias, R. Roy, D. Davis, R. Poli, K. Balakrishnan, V. Honavar, G. Rudolph, J. Wegener, L. Bull, M.A. Potter, A.C. Schultz, J.F. Miller, E. Burke, and N. Jonoska (eds). Proceedings of the Genetic and Evolutionary Computation Conference, Morgan Kaufmann Publishers, San Francisco, pp 1150-55 (2002c).

Moore, J.H., Lamb, J.M., Brown, N.J., Vaughan, D.E. A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the *ACE I/D* and *PAI-1 4G/5G* polymorphisms on plasma PAI-1 levels. Clinical Genetics, in press (2002a).

Moore, J.H., Smolkin, M.E., Lamb, J.M., Brown, N.J., Vaughan, D.E. The relationship between plasma t-PA and PAI-1 levels is dependent on epistatic effects of the *ACE I/D* and *PAI-1 4G/5G* polymorphisms. Clinical Genetics, in press (2002b).

Moore, J.H. and Parker, J.S. Evolutionary computation in microarray data analysis. In Lin, S. and Johnson, K. (eds), Methods of Microarray Data Analysis, Kluwer Academic Publishers, Boston (2001).

Moore, J.H., Parker, J.S. and Hahn, L.W. Symbolic discriminant analysis for mining gene expression patterns. In De Raedt, L., Flach, P. (eds) Lecture Notes in Artificial Intelligence 2167, 372-381, Springer-Verlag, Berlin (2001).

Moore, J.H., Parker, J.S., Olsen, N.J., Aune, T. Symbolic discriminant analysis of microarray data in autoimmune disease. Genetic Epidemiology 23, 57-69 (2002d).

Moore, J.H. and Williams, S.M. New strategies for identifying gene-gene interactions in hypertension. Annals of Medicine 34, 1-8 (2002).

Nelson M., Kardina S.L.R., Ferrell R.E., Sing C.F. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Research 11:458-470 (2001).

Parker, J.S. and Moore, J.H. Dynamics based pattern recognition and parallel genetic algorithms for the analysis of multivariate gene expression data. Proceedings of the 2001 Genetic and Evolutionary Computation Conference Workshop Program, San Francisco, pp 433-436 (2001).

Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. and Moore, J.H. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. American Journal of Human Genetics 69, 138-147 (2001).

Ritchie, M.D., Hahn, L.W. and Moore, J.H. Power of multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions. Genetic Epidemiology, in press (2003).

Schena M., Shalon D., Davis R.W., Brown P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467-470 (1995).