Poster H-4

Machine Learning Algorithms for Polymorphism Detection

Authors:
Georg Zeller (*Friedrich Miescher Laboratory of the Max Planck Society*)
Gabriele Schweikert (*MPI Biological Cybernetics*)
Richard Clark (*MPI Developmental Biology*)
Stefan Ossowski (*MPI Developmental Biology*)
Norman Warthman (*MPI Developmental Biology*)
Paul Shinn (*The Salk Institute*)
Kelly Frazer (*Perlegen Sciences Inc.*)
Joe Ecker (*The Salk Institute*)
Daniel Huson (*University of Tübingen*)
Detlef Weigel (*MPI Developmental Biology*)
Bernhard Schölkopf (*MPI Bilogical Cybernetics*)
Gunnar Rätsch (*Friedrich Miescher Laboratory of the Max Planck Society*)

Short Abstract: Based on high-density oligo-nucleotide array measurements and sophisticated machine learning methods, we obtain a genome-wide inventory of polymorphisms (including SNPs, deletions and highly polymorphic regions) in natural populations of Arabidopsis thaliana, representing an unprecedented resource for the study of genetic variation in a multicellular model organism.

Long Abstract:

As extensive studies of natural variation require the identification of sequence differences among complete genomes, there exists a high demand for precise, yet inexpensive high-throughput sequencing techniques. While high-density oligo-nucleotide arrays are capable of rapid and comparatively cheap genomic scans, algorithmic approaches for the accurate identification of sequence polymorphisms from this kind of data remain a challenge [1]. We will present two machine learning based methods tackling the problem of identifying Single Nuceleotide Polymorphisms (SNPs) as well as deletions and highly polymorphic regions. We have collaborated with Perlegen Sciences Inc., to use array hybridization technology for polymorphism discovery in 20 wild strains of the model plant Arabidopsis thaliana, which has a genome of about 125Mb. From this project we obtained nearly 19.2 billion measurements (four 25nt probes for each base on each genomic strand and strain).

For the identification of SNPs, we trained support vector machines (SVMs) on a set of known sequences [2] from each of the 19 non-reference strains. Our approach was two-layered incorporating information from each strain and also across the strains to maximize the SNP prediction accuracy. An important feature of our approach is that each called SNP is given a confidence level that has been lacking in earlier approaches. This allows us to adjust the recovery and accuracy along a ROC curve according to experimentalists' needs. Neighboring polymorphisms (distance $<$25nt) disrupt the signal for SNP detection so that algorithms tend to predict the fewest SNPs in the most highly polymorphic regions. A comparison of our method to a previously used model based method (the model based

method is similar to the one published in [3]) revealed that our approach excels in polymorphic regions that make up much of the genome. We recover many SNPs with a distance to the next polymorphic feature as small as 10bp. Per strain we are able to identify on average about 166,000 SNPs (34% recovery rate) at a false discovery rate of 5% leading to a total of 747,000 predicted SNPs. Some of these cause major functional effects (e.g. premature stop codons, discruption of splice sites, or deletions of coding sequence), of which nearly 600 were confirmed by dideoxy sequencing.

Considerable portions of genomes consist of regions with very high SNP density, where SNP detection algorithms typically fail to reliably identify individual SNPs. We have therefore developed methods to detect highly polymorphic regions containing clusters of SNPs, insertions and deletions. By comparing hybridization signals from target strains to the reference strain, we distinguish between meaningful candidates and other regions with low signal intensity caused by experimental variability. We are currently developing algorithms based on novel label sequence learning techniques (following a discriminative approach related to Hidden Markov SVMs [4]) in order to predict highly polymorphic regions including their composition. In a preliminary attempt we predicted about 700 such regions per strain and a subset of the predictions were validated by dideoxy sequencing, confirming about 100 deletions of length 150bp to 10kb with major effects on genes. First results indicate that more advanced techniques will reduce the length of polymorphic regions that are reliably detectable to 30bp. In summary, we obtain the first whole genome inventory of polymorphisms in an experimentally tractable, multicellular model organism.

References:

[1] D. Gresham, D.M. Ruderfer, S.C. Pratt, J. Schacherer, M.J. Dunham, D. Botstein, L. Kruglyak. Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. Science 311(5769):1932-1936, 2006.

[2] M. Nordborg, T.T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, P. Calabrese, J. Gladstone, R. Goyal, M. Jakobsson, S. Kim, Y. Morozov, B. Padhukasahasram, V. Plagnol, N.A. Rosenberg, C. Shah, J.D. Wall, J.Wang, K. Zhao, T. Kalbfleisch, V. Schulz, M. Kreitman, J. Bergelson. The pattern of polymorphism in A. thaliana. PLoS Biology, 3(7):e196, May 2005.

[3] D.A. Hinds, L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, K.A. Frazer, D.R. Cox. Whole-genome patterns of common DNA variation in three human populations. Science 307(5712):1072-1079, 2005.

[4] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research, 6:1453-1484, 2005.