PROBABILISTIC MODELING OF UNCERTAINTY IN GENOTYPING STUDIES

Peter Sárközy
Advisors: Péter Antal, Tadeusz Dobrowiecki

I. Introduction

In this article we present a probabilistic approach to model uncertainties in genotyping with an explicit representation of rejection and a probabilistic framework to cope with such uncertain data.

While the use of noise models is common in gene expression data analysis, where they cope with similar normalization and feature extraction problems [1], such models are missing in genetic association studies (GAS). Beside measurement problems, uncertainty also arises in haplotype reconstruction as well as later stages of the GAS analysis chain. Our planned framework will allow the explicit propagation of uncertainty from measurement through haplotype reconstruction to data, and allow us to fully utilize all of the information contained in a genotyping measurement.

II. The measurement

The DNA segments containing the single nucleotide polymorphisms (SNPs) in question from multiple samples are amplified by a polymerase chain reaction (PCR), and then pipetted into spots on a plate that contains their complementary strands bound to the plate substrate. The strands are marked during PCR by different color dyes to denote whether the SNP is wild type or mutant type. The dyes fluoresce under specific monochromatic frequencies, thus the images can be recorded. Each dye emits light of a different wavelength. Some systems use a third control dye mixed in to act as a control value and can be used for normalization. Brightness is nonlinear with the amount of hybridized marked strands.

Each plate is made up of a large number of wells (48-384), which contain the DNA from a single sample, and contain multiple spots (12-64) for each SNP being measured. Different batches of chemical agents are used on a plate, to insure that a single bad chemical doesn't cause complete failure, and also for optimizing the pipetting system's workflow.

III. Image processing and clustering

In our studies we analyzed two different genotyping systems, Beckman Coulter's SNPstream and the Applied Biosystems' (ABI) TaqMan probe based assay. These systems advertise high call rate and high data concordance, but in practice these results vary. The noise characteristics in the image processing phase of both systems is quite similar, as well as the sample layout and physical characteristics of each measurement. We derive quality metrics such as circularity, noise levels, evenness, signal-to-noise ratio, etc. from each sample based on the image characteristics that we can later use in the quality assessment of each sample.

A 3 level noise model can be applied to the systems, with noise parameters applicable at the plate level (total background noise, intensity offset, chemical batch errors), well level (local noise, large artifacts, dust) and spot level (total intensity, circularity). We assume a multivariate quantized errors-in-variables model for this noise, as seen in Eq. (1).

$$y_t = \log(\underline{x}_t^*) + \underline{\varepsilon}_t\,,\ \underline{x}_t = \underline{x}_t^* + \underline{\eta}_t\,. \tag{1}$$

In the clustering part of the measurement, where we take all of the samples for one SNP and plot the intensities of each dye to produce genotype calls, the two systems use different coordinate systems, but these are easily transformed. The samples are grouped into clusters to determine the genotype associated with each one. The clusters distribution characteristics and the samples position relative to its corresponding and neighboring clusters lets us derive the final certainty score for the sample. The problem is a classification problem with rejection [2], where we are trying to get the most reliable results out of our dataset, while preserving its integrity. We must not only classify and provide the quantized classification or rejection result, as do most commercial solutions, but also must provide our level of confidence in our decision. Changing the rejection threshold will result in higher area under the curve of the receiver operating characteristic.

IV. Results of the analysis

We were able to provide more accurate genotype calls than commercial solutions, from the image data alone. We were also able to assign uncertainty data to each sample which correlated very closely with the errors made by other commercial genotyping software. We had marked success on calling genotypes on hard to analyze SNPs.

We compared 768 samples of a single SNP. We used a validated set from a TaqMan based probe (very accurate but slow and expensive) of this SNP. The comparison was between the calls made by the SNPstream application suite and our system The TaqMan probe based system was used as a reference, because its primer is highly optimized for a single SNP and generates very accurate calls, unlike the SNPstream system (48 primers per plate). The SNP chosen for this was one that was difficult to assay with SNPstream, because of its low average spot intensities and high noise levels.

SNPstream called 72 SNPS erroneously out of 768 compared to the TaqMan assay, while our application only called 56 errors.

Altogether there were 36 instances where our application had produced calls different from the SNPstream application, and all of these instances had very high associated uncertainty metrics. The distance from cluster center and the signal-to-noise ratio were the most significant. All the spots that were called differently had very high uncertainty metrics, thus likely to be false.

We found several errors in their measurement protocol. The plates were not washed well enough from residual non-hybridized fluorescent primers, as well as there being wipe marks on the plates. These resulted in many erroneous calls from large noise artifacts. We also managed to pinpoint that the pipetting machine applied gradually increasing concentrations of chemicals on the plates, because the contents of the vials settled fast. Adding a mixing step before each sample load will eliminate this problem.

V. Further plans

Quality measures can be successfully used to revise previous measurements with a variable certainty threshold, and to create probabilistic models for genotype calls that can aid in haplotype reconstruction. Applying these methods to large sets of data can allow us to refine probabilistic models that can be used for other uncertain data sets as well [3]. Conventional methods of utilizing the output of genotyping calls can be augmented by running these methods multiple times with a variable certainty threshold.

References

[1] E. Wit, J, McClure, "Statistics for Microarrays", *Department of Statistics,University of Glasgow, UK*, pp. 57-62, 2004

[2] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, B. De Moor, "Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection" *Artificial Intelligence in Medicine 29*, pp. 39–60, 2003

[3] E. Halperin et al., "Tag SNP selection in genotype data for maximizing SNP prediction accuracy", *Bioinformatics,* vol. 21, no. 1, 2005