A method to improve microarray-based identification of SNPs

Silvia Galfré*^, Francesco Morandin#, Arianna Cozza*, Silvia Pellegrini* and Roberto Marangoni^

*Dipartimento di Patologia Sperimentale, Biotecnologie mediche, Infettivologia ed Epidemiologia, Università di Pisa
#Dipartimento di Matematica, Università di Parma
^Dipartimento di Informatica, Università di Pisa, Via F. Buonarroti 2
56127 Pisa, Italy
marangon@di.unipi.it

Keywords. SNP, microarray, bilinear regression.

Introduction

Single Nucleotide Polymorphism (SNP) represents a variation in sequence (polymorphism) between individuals caused by a change in a single nucleotide. This process is responsible for most of the genetic variation between individuals. Furthermore, the identification of distinct SNPs may play a crucial role in assessing a potential genetic influence for those disorders that do not appear to have a simple genetic transmission. In turn, the identification of genetic risk factors may contribute to determine biological markers of disease that can be used for the preclinical diagnosis of a pathological condition. Early diagnosis is important for enacting successful therapeutic strategies. In order to obtain more informative data, multiple SNPs should be tested simultaneously in the same individuals.
A common protocol used in SNPs investigations is based on Single Base Extension (SBE) followed by microarrays hybridization [1], in which each DNA sample is hybridized on two arrays: one used to explore the existence of “A” and “T” in the SNP locus, the other array for “C” and “T”. To obtain a global evaluation of the frequency with which each SNP is represented in the population, it is necessary to make a quantitative comparison of the signals recorded from the two arrays. Because of many technical reasons, during this step a large quantity of noise is introduced, thus compromising the reliability of the final data.
Here we present a simple approach, based on the usage of three arrays instead of only two, which can address this problem. We also give a statistics method for data processing to be used with the proposed experimental protocol.

Some details

1. The protocol: three-array solution

The proposed protocol is identical to the above described, but one more step is added: a third array is used, in which the hybridization is made by all the marked DNAs used for hybridizing the previous two arrays. The signal recorded from this array is used as a quantitative referring point to compare the intensity of signals of the remaining two arrays.

2. The analysis: bilinear regression

Our data set consists of three arrays, N spots and two channels (red and green color of dyes) for each array. These three arrays are not equivalent: the array 1 contains DNA hybridized in such a way to discriminate between two possible nucleotides (e.g., “A” in the red channel and “T” in the green channel), the array 2 is hybridized to discriminate the remaining couple of nucleotides (e.g., “C” in the red channel, “G” in the green

channel), and the array 3 contains all the labeled DNAs previously used (e.g. “A, C” in the red channel, “T, G” in the green channel).
For each choice of channel, spot and array we can write the following equation:

(0.1) $$I_i = x_i + ex_i + \varepsilon$$

where I_i is the measured intensity, x_i is the true intensity, e and ε are the multiplicative error and the additive error, respectively, and $i = \{1,2,3\}$ refers to the array. The dependencies on the single spot and dye are left implicit, as usual in regression methods. We assume that all the errors are independent of the particular choice of spot, array and dye. It follows:

(0.2) $$I_3 = aI_1 + bI_2 + c$$

where the constants *a*, *b* and *c* are specific for each triple of arrays that we analyze. By estimating these constants, the arrays 1 and 2 become comparable, simply by multiplying the intensities in the array 1 by *a*, and the intensities in the array 2 by *b*. If we substitute the (0.1) in (0.2) and we move all the error terms to the right, we obtain:

(0.3) $$x_3 - ax_1 - bx_2 - c = eax_1 + ebx_2 - ex_3 + a\varepsilon_1 + b\varepsilon_2 - \varepsilon_3$$

Writing this equation for all spots and both dyes, we use a bilinear weighted-least-squares regression to find an estimate of the constants *a*, *b* and *c*. To do this, the variance of each right-hand term should be estimated up to a multiplicative constant. This suggests a recursive approach, starting from a neutral estimate (in our case, *a=1, b=1*), and iterating the regression procedure, using in each iteration the estimations of *a* and *b* obtained in the previous iteration. Several experiments we performed show that this recursive method converges very fast: since the first iteration, the estimation of the searched parameters is already good. So we reduce the method to only one regression calculation, setting at 1 the initial variances. We are currently evaluating if this estimation could be improved applying a quality control filter [2].
Once the values of *a*, *b* and *c* have been calculated, we are able to determine the percentage of a single nucleotide (*dNTP*) present in a particular locus:

(0.4) $$\%\left(dNTP\right) = \frac{I_{ij}}{I_{1r} + I_{1v} + I_{2r} + I_{2v}}$$

where I_{ij} indicates the intensity of the array *i* and of the dye *j* recalculated for both arrays using the proper scale factors (0.2). The calculation of the %(*dNTP*) for all the considered loci will give the total number of SNPs existent in the population.

3. A brief description of the main results obtained

In this study, we used microarrays to examine 23 SNPs related to the neurotrophine genes and their potential role in subjects affect by Alzheimer’s type dementia. We compared the SNPs determination using the standard two-array and the proposed three-array protocols. Our results clearly show a significant improvement in the data analysis, due to uncertainties reduction and noise filtering.

References

[1] J. N. Hirschhorn, P. Sklar, K. Lindblad-Toh, Y. Lim, M. Ruiz-Gutierrez, S. Bolk, B. Langhorst, S. Schaffner, E. Winchester, E.S. Lander, SBE-TAGS: An array-based method for efficient single-nucleotide polymorphism genotyping. *PNAS,* 97, 12164–12169, 2000.

[2] X. Wang, S. Ghosh, S. Guo, Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, 29, 2001.