CGHMltiArray: exact P-values for multi-array comparative genomic hybridization data

Mark A. van de Wiel1,3,*, Serge J. Smeets2, Ruud H. Brakenhoff2 and Bauke Ylstra3
1Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, PO Box 513, 5600 MB, Eindhoven, The Netherlands, 2Department of Otolaryngology—Head and Neck Surgery and 3Microarray Core Facility, VU University Medical Center, Amsterdam, The Netherlands

Received on March 22, 2005; revised on May 3, 2005; accepted on May 4, 2005
Advance Access publication May 6, 2005

ABSTRACT
Summary: We compute P-values, based on the Wilcoxon test with ties, to compare two conditions with array comparative genomic hybridization data, and we provide a simple interface to export and plot these P-values.

Availability: CGHMltiArray is freely available at http://www.win.tue.nl/~markvdw/CGHMltiArray.html
Contact: m.a.v.d.wiel@tue.nl

Supplementary information: Programs, the manual and supplementary information are available on the website.

Array comparative genomic hybridization (array CGH) is applied to the detection of genomic abnormalities in cancer and inheritable DNA copy number aberrations that cause genetic disorders. It is a high-resolution, high-throughput technique that allows for genomewide measurement of chromosomal DNA copy number changes and determination of the associated breakpoints along the chromosomes (Oostlander et al., 2004).

Software such as aCGHsmooth (Jong et al., 2004) and similar programs (Olshen et al., 2004) enables the visualization and identification of aberrated chromosomal regions by individual separately. CGH-Miner (Wang et al., 2005) has additional features to summarize alteration information over groups. We developed CGHMltiArray to integrate array CGH data over individuals by computing P-values per clone and visualizing these to find generic patterns. The program deals with the most common situation: comparison of two conditions.

When considering a suitable statistic to measure generic DNA copy number changes among individuals, we have to consider the nature of array CGH data. Although technical errors may disperse the data somewhat, the data in reality represent discrete levels of genetic aberrations. The normal DNA copy number of mammalian clones is two: one from both the paternal and maternal chromosomes. In particular diseases, such as cancer, changes in the DNA copy number with respect to the ‘normal’ value may occur as a ‘deletion’ (at least one copy is lost) or a ‘gain’ (at least one additional copy is present). These non-normal levels may be further detailed, e.g. by including ‘amplification’, which is a high level of copy number gains.

The granularity of (discretized) CGH data makes the t-statistic, or variations thereof, unsuitable. Moreover, the discrete levels possess a natural ordering, which rules out the Fisher exact test. The Wilcoxon test makes explicit use of both features. However, the data naturally contain many ties, i.e. sets of equal observations. The distribution of the Wilcoxon statistic, and consequently the P-values, depends on the tie structure (Hájek et al., 1999). Hence, it has to be re-computed for each new case.

Define the Wilcoxon statistic W as the sum of the mid-ranks assigned to the smallest sample. The observed value of W is denoted by w. The two-sided P-value is then defined by 2P(W ≤ w) if w ≤ E(W), and 2P(W ≥ w) otherwise, where probabilities and expected values are computed under the null hypothesis of equally likely permutations of the mid-ranks. Since the number of tests is of the order of thousands, one needs a fast calculation method. Moreover, asymptotic theory is often not applicable, because the number of biological replicates per condition is small and the presence of ties worsens the accuracy of asymptotic approximations. For example, when both sample sizes equal eight, cases with asymptotic P-value approximations in the range 0.0001–0.05 correspond to 2–3 times larger exact (true) P-values, which leads to more than a doubling of the number of false calls when using the approximations.

Therefore, a fast algorithm to compute exact P-values is needed. The relevance of such algorithms to solve bioinformatics problems was recently shown by Bejerano et al. (2004). We developed the split-up algorithm (van de Wiel, 2001), which suits the requirements well: it is fast, exact and deals with ties. The algorithm represents the probability distribution of the test statistic under the null hypothesis (‘no change between two conditions’) as a generating function. Baglivo et al. (1996) showed that generating functions are powerful tools to represent null distributions of discrete test statistics. We used the generating function introduced by Streitberg and Röhmel (1986). This generating function is a polynomial in product form, expansion of which would reveal the entire null distribution, but this may be time consuming. The split-up algorithm splits the product into two parts and requires the expansion of these two smaller parts, which is several orders of magnitude faster than full expansion. Then, these two results are efficiently combined to compute the P-value. CGHMltiArray is written in Mathematica (Wolfram, 1999). The basic algorithm is also available as R code and as an executable. To make the algorithm easily accessible, we provide a web implementation too.

The website provides a tool to convert smoothed log2-ratios from other software to input data for CGHMltiArray. First, it transforms observed array CGH log2 values to discretized data: ‘1’ for gains,
Fig. 1. DNA view of \( P \)-values by position for an illustrative dataset.

\`-1\` for losses and \`0\` for normals. It allows for the introduction of extra levels for amplifications or double deletions. Next, it counts occurrences of all levels for both conditions.

The input for CGHMultiArray then consists of a simple text file, the rows of which represent clones in the chromosomal order. When \( \ell \) is the number of levels used (three in this example), the first (second) \( \ell \) columns represent counts of the number of control (treatment) samples attaining level \( j \), \( j = 1, \ldots, \ell \). Optional columns may provide name, chromosomal information and base pair position information, which allows for separate \( P \)-value plots by chromosome. The algorithm stores all count configurations, so \( P \)-value computations for clones with the same configuration are not repeated. More details are provided in the manual. Note that when one would like to use only \( \ell = 2 \) levels, e.g. for comparing gains, the Wilcoxon statistic with tie correction is equivalent to the hypergeometric statistic.

We have used CGHMultiArray to analyze the genomic changes of two groups of 12 head and neck tumors. Illustrative data shown here (also available on the website) are identical to the real data except for permutations of chromosomes. CGHMultiArray generates an exportable list of univariate \( P \)-values, a DNA view of these (Fig. 1) and, optionally, a view by chromosome (data not shown). These views are useful to identify regions with unusually many differential aberrations. One may wish to perform a multiple testing correction to the \( P \)-values afterwards, such as the Benjamini and Yekutieli (2001) FDR rule. In this case, one may want to focus on a limited number of clones or consider chromosomal regions instead of separate clones. An implementation for the latter option is provided on the website.

ACKNOWLEDGEMENTS

Part of this research has been funded by the Dutch BSIK/BRICKS project. We thank Marko Boon, Kyung In Kim and Marcel van Verk for their software support.

REFERENCES


