Microarray Data Analysis Toolbox (MDAT): for normalization, adjustment and analysis of gene expression data

Nicholas Knowlton*, Igor M. Dozmorov and Michael Centola

Department of Arthritis and Immunology, Oklahoma Medical Research Foundation, 825 NE 13 Street, RP Rm 461 MS: 58, Oklahoma City, OK 73104, USA

Received on January 21, 2004; revised on June 29, 2004; accepted on July 15, 2004
Advance Access publication July 22, 2004

ABSTRACT
Summary: We introduce a novel Matlab toolbox for microarray data analysis. This toolbox uses normalization based upon a normally distributed background and differential gene expression based on five statistical measures. The objects in this toolbox are open source and can be implemented to suit your application.

Availability: MDAT v1.0 is a Matlab toolbox and requires Matlab to run. MDAT is freely available at http://microarray.omrf.org/publications/2004/knowlton/MDAT.zip
Contact: knowltonn@omrf.ouhsc.edu

1 INTRODUCTION
DNA microarrays are a powerful tool at examining genome-wide changes in gene expression. With statistical scaling and proper normalization, differential genes can be selected on a statistical basis. This approach gives potential pathway targets that allow researchers to reverse engineer the respective biological pathways and reveal possible disturbances. Our software utilizes Matlab R12 (Mathworks, Natick, MA) for analysis.

Matlab is a powerful analysis tool that runs on a series of text scripts that are parsed at execution. This ‘open-architecture’ gives more flexibility and insight than compiled languages (C, Java); furthermore, it allows users to cherry pick pieces suitable for integration into their analysis methods. Since Matlab was designed to deal with massive matrices, it is perfectly suited to microarray data (Venet, 2003; Eichler et al., 2003). This is compounded by the many toolboxes available for statistical applications, including self-organized maps, neural networks and hierarchical clustering.

Our toolbox adds to previous implementations by introducing new algorithms for normalization and differential gene analysis. In brief, our normalization method relies on a number of low-expression genes to provide an estimate of non-specific binding. This information is then used to perform a Z transformation on the data. Once normalized the data are ‘unbiased’ (described in the following sections) to allow comparisons across chips. Fitted data are then used to find a homoscedastic group of gene variances that will be used as an internal standard of measurement noise (Dozmorov et al., 2004). Finally, a Student’s t-test, associative t-test (Dozmorov and Centola, 2003a) and normalized ratio are used to select differentially regulated genes.

2 IMPLEMENTATION
2.1 General comments
The program presented here was based on the methods of Dozmorov and Centola (2003a) and has been utilized in several recent papers (Dozmorov et al., 2003b; Jarvis et al., 2003, 2004) with the use of Clontech membranes and 70mer human oligos. Its universality has been demonstrated on cDNA membranes, cDNA slides, 70mer oligo custom slides and Affymetrix slides.

2.2 Modules
2.2.1 Data trimming
The mean and SD are calculated. Using these as a starting point, data beyond +2 SD above the mean is cut. The mean and SD are recalculated and data beyond −2 SD below the mean is cut. This trimming of outstanding values above and below is continued, further refining the SD estimate, until no additional cuts can be made.

2.2.2 Curve fitting
SD and mean are estimated by given data. Using these estimates, a Gaussian CDF is generated and compared to the actual data’s CDF. This comparison is accomplished by a non-linear least squares minimization procedure. The result is a 10 bar histogram superimposed by a Gaussian distribution corresponding to the optimized estimators (Fig. 1). The user is presented with this picture to determine if the right half of the distribution is distorted by low signal and therefore, is not normally distributed like background signal. If the user agrees that there is a ‘good fit’ (minimal tail distortion) then the parameter estimates are outputted. If there is some visual distortion of the right tail
Fig. 1. Curve fitting. A histogram showing the gene expression distribution is presented. The bottom 30th percentile of data is represented by the diagonal hash bar. Inset, histogram of lower 30th percentile of data processed through the data trimming module. The solid black line is empirically determined by the data while the dotted line is a mirror image of the left half. The dot-dashed line represents the axis of symmetry for this curve fit.

(proposed presence of weak gene expression) the procedure is repeated using a new user-defined mean (histogram bars 1–5) and estimating the new distribution on the bars to the left of the chosen one. For example, if bar 4 is chosen as the new mean. Bars 1–4 would be used to create the left tail and a symmetric right tail is produced that would now extend only to bar 8.

2.3 Algorithm

It should be noted that the normalization and differential gene expression algorithms rely on the use of averages and SD of logarithm data. The mean is defined as $\overline{X} = \frac{1}{n} \log(\prod_{i=1}^{n} x_i)$, where $x_i$ denotes gene $x$ on chip $i$. The SD can be defined in a similar manner.

2.3.1 Normalization Normalization relies on the Gaussian distribution of lowly expressed genes to give unbiased estimators ($\mu, \sigma$) of technical noise. These estimators are determined as follows: the lowest 30th percentile of data is selected to eliminate highly expressed values. Then, the new subset is trimmed and subsequently curve fitted (see above). Once an appropriate fit is determined the data is Z-transformed [$Z = (x - \mu) / \sigma$] yielding mean $= 0$ SD $= 1$. Finally, the data are log-transformed with negative values substituted with the log of the minimum positive value.

2.3.2 Bias fitting Once each chip is normalized to its own technical noise, the chips are adjusted to each other. The basic biological assumption is that the majority of genes are equally expressed; capitalizing on this knowledge we average across groups (Groups 1 to $n$). Individual chips are compared to the groups’ average by a difference (Chip – Average). This log-ratio is trimmed (see above) and a simple linear regression (SLR) ($y_i = \beta_0 + \beta_1x_i$) is performed on the subset of data, where $x_i$ is the log average of all $y_i$s. The data are then adjusted for bias by a transformation of regression parameters [$x = (y - \beta_0) / \beta_1$]. Once all the chips have been processed, the average is re-calculated and the procedure is repeated for a total of three passes, further reducing bias.

2.3.3 Differential gene analysis We start with selection of a ‘reference group’—an internal standard for estimating parameters on equally expressed homoscedastic genes. Creating the reference group is a multi-step iterative process. First, the bias adjustment parameters ($\beta_0, \beta_1$) are used to create residuals in Group 1. Next, the SD of all residuals taken together is calculated along with the SD for every gene individually. An $F$-test is performed on every gene individually. An $F$-test is performed on every gene versus all the other genes. Excluding all genes whose $F$-statistic was significant ($\alpha = 0.05$) or whose SD was higher than the total groups SD reduces the variability in the group. This process is repeated
Table 1. Description of program output and statistical cutoffs

<table>
<thead>
<tr>
<th>A</th>
<th>GenBank accession no.</th>
<th>Group 1</th>
<th>Group 2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>SD</td>
<td>P background</td>
<td>Average</td>
<td>SD</td>
<td>P background</td>
<td>Students t-test</td>
<td>Associative t-test</td>
<td>Ratio G2/G1</td>
<td>Group</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&gt;1.5</td>
<td>A1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;t</td>
<td>&gt;t</td>
<td>&gt;t</td>
<td>&gt;t</td>
<td>&gt;t</td>
<td>&gt;t</td>
<td>&gt;t</td>
<td>&gt;t</td>
<td>&gt;1.5</td>
<td>A2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&lt;t</td>
<td>&gt;1.5</td>
<td>A3</td>
<td></td>
</tr>
<tr>
<td>NM_031943</td>
<td>3.58717</td>
<td>3.137587</td>
<td>4.81E-06</td>
<td>7.462597</td>
<td>3.760824</td>
<td>2.64E-09</td>
<td>0.012552</td>
<td>6.84E-13</td>
<td>2.080358</td>
<td>A1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC015691</td>
<td>13.17225</td>
<td>6.606308</td>
<td>1.55E-15</td>
<td>24.41795</td>
<td>8.382067</td>
<td>&lt;1E-16</td>
<td>0.001672</td>
<td>3.67E-08</td>
<td>1.853741</td>
<td>A1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AH057355</td>
<td>1002.749</td>
<td>499.3588</td>
<td>8.88E-16</td>
<td>323.2207</td>
<td>176.6664</td>
<td>4.05E-08</td>
<td>0.000244</td>
<td>1.76E-12</td>
<td>0.322335</td>
<td>A2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_004098</td>
<td>813.778</td>
<td>482.4975</td>
<td>1.52E-11</td>
<td>228.8166</td>
<td>127.8106</td>
<td>7.84E-08</td>
<td>0.002485</td>
<td>1.19E-11</td>
<td>0.281178</td>
<td>A2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC014161</td>
<td>1.075331</td>
<td>3.106765</td>
<td>0.166233</td>
<td>4.114545</td>
<td>3.020017</td>
<td>4.37E-05</td>
<td>0.009919</td>
<td>0.000912</td>
<td>3.826304</td>
<td>A3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK027180</td>
<td>1.605719</td>
<td>1.849044</td>
<td>0.000514</td>
<td>8.692896</td>
<td>5.676928</td>
<td>4.35E-06</td>
<td>6.65E-05</td>
<td>2.13E-14</td>
<td>5.41371</td>
<td>A3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK023845</td>
<td>246.8043</td>
<td>175.0725</td>
<td>1.71E-08</td>
<td>57.95832</td>
<td>73.76018</td>
<td>0.018408</td>
<td>0.003263</td>
<td>2.95E-06</td>
<td>0.234835</td>
<td>A4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK057626</td>
<td>3.142863</td>
<td>1.8698</td>
<td>1.79E-11</td>
<td>0.608879</td>
<td>1.272454</td>
<td>0.151176</td>
<td>0.00016</td>
<td>1.82E-05</td>
<td>0.193734</td>
<td>A4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data from IBD patients' PBMC. Threshold \( t = 5.2876 \times 10^{-5} \) denoted 't' in the table.

The top half of the table consists of statistical rules used to group genes and the bottom half of the table contains two genes for each group to give a better explanation of our statistical cutoffs. In this example, we examined peripheral blood mononuclear cells (PBMCs) from irritable bowel disease (IBD) patients. This example yielded a statistical cutoff of 5.2876e − 005 denoted ‘t’ in the table.
with the new smaller subset until the group is homoscedastic. Once the group is assembled, the data are given to the curve fitting module to estimate the SD.

The selection of differentially expressed genes relies upon five measures; assume Group 1 has \( n \) chips and \( k \) genes and Group 2 has \( m \) chips and \( k \) genes. First, knowledge garnered from the normalization step is used to select only those genes whose signal is above the level of non-specific binding. Then an associative-\( t \)-test (Dozmorov and Centola, 2003a) is performed, with \( m + k - 2 \) degrees of freedom (\( \alpha = 1/k \)), to see if the gene belongs to the group of equally expressed-genes with stable variability. Next, a Student’s \( t \)-test is performed, with \( n + m - 2 \) degrees of freedom (\( \alpha = 0.05 \)), to determine if the genes are equally expressed. Finally, the ratio of gene expressions in Groups 2 and 1 is used to help exclude statistically significant but not biologically significant changes. A summary of these parameters and how they are used to create differential gene groups are explained in Table 1.

This method has several advantages over other implementations. Our method can normalize Cy3 and Cy5 individually on two color cDNA slides, thus taking into account hybridization efficacy. Moreover, similar results are obtained regardless of technology type—Affymetrix, cDNA oligos or cDNA membranes. During differential gene analysis a Student’s \( t \)-test is used for high sensitivity and an associative \( t \)-test is used for high specificity. Selections passing through both criteria have a >95% chance (in-house success rate) of being validated by another method—i.e. quantitative real-time polymerase chain reaction (QRT–PCR). Other methods similar to our method have been proposed (Rocke and Durbin, 2001; Fan et al., 2004). However, there are several critical differences between their methods and ours. When constructing an internal comparison group neither method excludes hypervariable genes (genes with unusually high variability) and without this exclusion specificity is greatly diminished.

### 2.4 Limitations

MDAT is not limited to any specific microarray technology. Nevertheless, quality data are a prerequisite. Owing to the varying quality of several microarray technologies we have made some program options. As described in the first step of analysis, we take expressions at or below the 30th percentile as predominantly expressed below background. This is not always the best choice, so after examining the normalization, the user has the option of defining a new percentile or none at all (i.e. 100%). This gives great flexibility across technologies.

The data are grouped based on five statistical measures and users are presented with nine columns of data (Table 1). Using these data new classification rules can be created and tested, further expanding the versatility of this software.

### 2.5 License and availability

MDAT m files are distributed under the terms of the GNU Public License (GPL) (http://www.gnu.org/licenses/fdl.html). The toolbox is available for download from http://microarray.omrf.org/publications/2004/knowlton/MDAT.zip

### REFERENCES


