

BMDEExpress: A Software Tool for the Benchmark Dose Analyses of Genomic Data

Longlong Yang¹, Bruce C. Allen² and Russell S. Thomas¹. ¹The Hamner Institutes for Health Sciences, 6 Davis Drive, P.O. Box 1237, Research Triangle Park, NC 27709-2137. ²Bruce Allen Consulting, 101 Corbin Hill Circle, Chapel Hill, NC 27514

TOXICOGENOMICS

ABSTRACT

The analysis of dose-response data is fundamental to the disciplines of toxicology and chemical risk assessment. With the advent of microarray technology in these fields, the number of software tools for analyzing dose-response microarray data has become a limitation. To address this problem, we have developed BMDEExpress, a Java application that combines traditional benchmark dose (BMD) methods with gene ontology (GO) classification in the analysis of dose-response data from microarray experiments. The software application is designed to perform a stepwise analysis beginning with a one-way analysis of variance to identify the subset of genes that demonstrate significant dose-response behavior. The second step of the analysis involves fitting the gene expression data to a selection of statistical models (linear, 2nd polynomial, 3rd polynomial, and power models) and selecting the model that best describes the data with the least amount of complexity. Multiple options are provided for model selection in the software program. Once the best model is selected, the software application queries a client-accessible MySQL database and matches each gene to its corresponding GO categories. Summary values characterizing the central tendencies for the BMDs and BMD lower confidence limits within each GO category are calculated and represent the dose level at which the corresponding cellular process is significantly changed. The software represents a significant advance by allowing users to summarize microarray-based surveys of molecular changes associated with chemical exposure and identify reference doses at which particular cellular processes are altered.

Availability: <http://sourceforge.net/projects/bmdexpress/>

INTRODUCTION

An important part of the risk assessment process is characterizing the dose response behavior of a chemical in question and comparing it to the level of exposure received by the public. To assess the dose response behavior of these chemicals, benchmark dose (BMD) methods are typically employed for estimating reference doses. In the BMD method, dose response data for the toxic endpoint is fit with a statistical model and a BMD is identified that results in a defined level of risk over that observed in control populations. The BMD and the associated lower confidence limit (BMDL) are then used to set standards for human health effects (EPA, 1995).

Microarray technology has been broadly accepted as an efficient and reproducible way to simultaneously measure the expression of thousands of genes. In toxicology, the ability to survey thousands of genes provides a comprehensive assessment of the transcriptional changes resulting from chemical exposure. Bioinformatic methods have been developed to interpret these changes by applying standardized functional annotations to each gene and identifying whether certain biological processes or molecular functions are over- or under-represented (Beissbarth and Speed, 2004; Dennis, et al., 2003). This approach has been referred to as a gene ontology (GO) enrichment analysis and allows large lists of transcriptional alterations to be distilled down into changes in cellular processes such as the immune response, DNA repair, apoptosis, etc.

BMDEExpress integrates BMD methods with GO classification analysis in the examination of microarray dose-response data. The combination of microarray technology with these analysis methods results in a unique risk assessment tool that provides both a comprehensive survey of molecular changes following chemical exposure and dose estimates at which different cellular processes are altered based on a defined increase in risk.

METHODS

BMDEExpress is written in Java with Swing graphical user interface and native functions calling dynamic link library (DLL) written in C and FORTRAN. The software is designed to perform a stepwise analysis on dose-response microarray data that combines BMD methods with GO classification analysis. The program has a series of interfaces that guide the user through the analysis process.

Data Import: Microarray dose-response data are imported as a tab-delimited file with each column as an individual array and the first row listing corresponding doses. Currently, only Affymetrix microarrays are supported (Fig. 1).

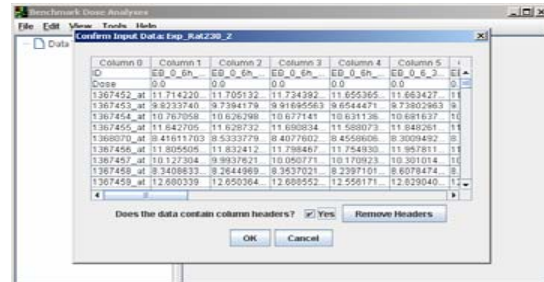


Figure 1: Interface for importing dose-response microarray data. Each microarray sample must be listed as an individual column and the corresponding doses must be provided in the first line of the file. The interface provides an option to remove column headers.

One-Way ANOVA: The first step in the analyses process is to perform a one-way ANOVA to identify probe sets that show significant dose response behavior (Figs. 2 and 3). The software allows the user to adjust the resulting probability values for multiple comparisons using false discovery rate (Benjamini and Hochberg, 1995). The one-way ANOVA is necessary to reduce the computational requirements in the subsequent BMD calculations.

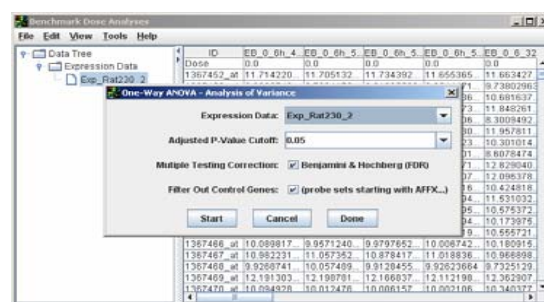


Figure 2: Interface for performing one-way ANOVA, allowing correction for multiple comparisons and removal of control probe sets.

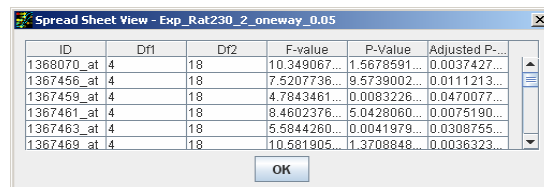


Figure 3: Example output from one-way ANOVA.

Benchmark Dose Analyses: The BMD analyses involves fitting the gene expression data to a selection of statistical models (linear, 2nd polynomial, 3rd polynomial, and power models) and selecting the model that best describes the data (best model) with the least amount of complexity (Fig. 4). The statistical models implemented in the current version of BMDEExpress are modified from the source code of the BMDs developed by the U.S. EPA. The original source code was modified to create a DLL with native C and FORTRAN functions called using a Java Native Interface.

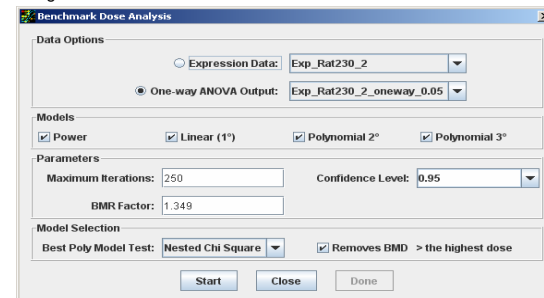


Figure 4: BMD analysis interface with options for input data, models, parameters, and model selection.

The user is allowed to modify several critical parameters associated with the BMD analyses and can also choose the method for model selection. In the first option, a nested likelihood ratio test is used to select among the linear, 2nd polynomial, and 3rd polynomial models followed by an Akaike information criterion (AIC) comparison between the best nested model and the power model (i.e., the model with the lowest AIC is selected). In the second option, a completely AIC-based selection process is performed. In the output from the BMD analyses, the probe set identifier is provided along with selected values for each of the statistical models. These include the BMD, the BMDL, the fit p-value from the likelihood ratio test, the log-likelihood value for the fit of the model, and the AIC (Fig. 5).

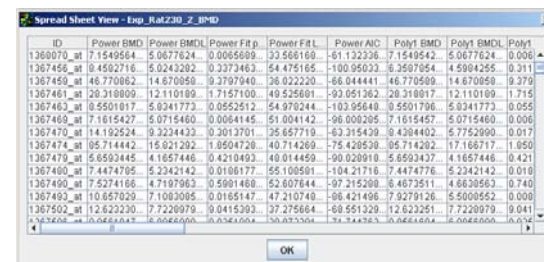


Figure 5: Output of BMD analysis with probe set IDs, and BMD, BMDL, fit p-value, fit log-likelihood value, and AIC value from each model.

Gene Ontology Analyses: Upon completion of the BMD analyses, the user can organize the gene expression changes using GO classification for the individual probe sets. The user can select the biological process, molecular function, or cell component GO classes (Fig. 6). The software application then queries a client-accessible MySQL database and matches each probe set identifier to its corresponding GO categories. Summary values representing the central tendencies and variability for the BMD and BMDL across all genes in the category are calculated and provided as output (Fig. 7). The summary values represent the dose level at which different cellular processes are altered. The GO database accessed by the program is automatically updated to ensure the classifications are current.

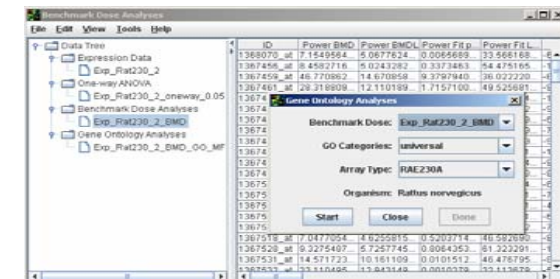


Figure 6: Gene Ontology analysis interface. Probe set identifiers are converted into unique genes and matched to GO terms. The organism is automatically determined based on the probe set identifiers.

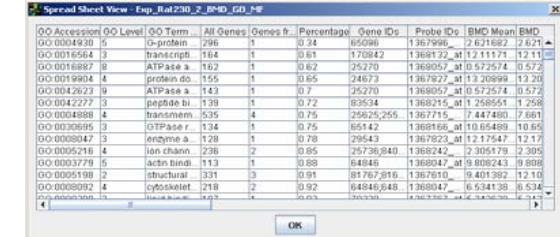


Figure 7: GO analysis output provides a large list of summary values for the BMDs and BMDLs in each GO category including mean, median, percentile rank, and standard deviation.

SUMMARY

The analysis of dose-response data is critical to the fields of toxicology and chemical risk assessment. However, the number of bioinformatic tools available for analyzing dose-response microarray data is limited. BMDEExpress was written to fill this gap by integrating analysis methods that are well established within their respective fields. The software represents a significant advance by allowing users to summarize microarray-based surveys of molecular changes associated with chemical exposure and identify reference doses at which particular cellular processes are altered.

ACKNOWLEDGEMENTS

The authors would like to thank Jeffrey S. Gift and R. Woodrow Setzer for providing helpful suggestions and the source code used in the BMDs software. This work was supported by a grant from American Chemistry Council's Long Range Initiative and a Superfund Program Project Grant (2 P42 ES004911-17).

REFERENCES

Beissbarth, T. and Speed, T.P. (2004) Gostat: find statistically overrepresented Gene Ontologies within a group of genes, *Bioinformatics*, 20, 1464-1465.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B*, 57, 289-300.

Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery, *Genome Biol*, 4, P3.

EPA (1995) The Use of the Benchmark Dose Approach in Health Risk Assessment. Office of Research and Development, U.S. Environmental Protection Agency, Washington, D.C.

