BIOCHIP JOURNAL, Vol. 3, No. 4, 354-357, December 2009

METASIS: The Meta-analysis Tool for Biochip Data

Mi-Kyung Lee[1] **& Yang-Seok Kim**[1]

[1]College of Oriental Medicine, KyungHee University, Hoegi-dong, Dongdaemun-gu, Seoul, Korea
Correspondence and requests for materials should be addressed to Y.-S. Kim (yskim1158@khu.ac.kr)

Accepted 7 October 2009

Abstract

Recently, many public databases such as CEBS (Chemical Effects in Biological Systems) and GEO (Gene Expression Omnibus) have been developed to provide raw expression data with their experimental conditions. The proper use of publically available raw data can be a very efficient method of making biological discoveries without performing experiments. However, one barrier to this approach is that the experiments from which the data in the databases were generated were performed using many different types of array platforms, which each produce information with different characteristics. Therefore, it is necessary to develop a program to provide a variety of statistical methods to integrate different types of expression data for meta-analysis. We have developed the METASIS meta-analysis software for analysis of expression arrays. METASIS can deal with various experimental data from all kinds of platforms if properly formatted files are provided. METASIS offers state of the art statistical methods, such as t-based modeling, rank product and Fisher's inverse Chi-square method. In addition, Java was used as the programming language for METASIS. The software is available upon request via e-mail (yskim1158@khu.ac.kr).

Keywords: Gene expression, Meta-analysis, T-based, Rank product, Fisher's inverse chi-square

Introduction

Microarray technology assesses the gene expression levels of thousands of genes simultaneously in a high-throughput mode. This technology has been widely used to achieve various goals, such as to understand underlying biological mechanisms, to discover novel subgroups of phenotypes, to examine drug responses, to classify samples into specific phenotype groups and to predict phenotype outcomes. Despite the great contributions that microarray techniques have made, some researchers have argued that microarray-based biological discoveries are not reproducible or robust. These situations are caused by inappropriate analysis or validation and insufficient sample size. Additionally, the fact that there are usually a larger number of probes than samples makes this situation worse[1,2]. To overcome these problems, it has been proposed that information from multiple existing microarray studies such as GEO (Gene Expression Omnibus), CEBS (Chemical Effects in Biological Systems) and ArrayExpress be combined[3-5]. Meta-analysis to combine the results from various sites and apply a specific statistical technique focused on meta data analysis can increase the reliability and generalizability of results. Moreover, meta-analysis allows increased statistical power to obtain a more precise estimate of gene expression differentials, assess the heterogeneity of the overall estimate and produce a single summary measure by combining study results. Recently, meta-analysis has been extensively used and its usefulness has been demonstrated[6]. Therefore, several tools to conduct gene expression meta-analysis have been introduced, including GeneMeta and RankProd. However, these programs are implemented in non familiar interfaces such as R. Of course, some easy-to-use programs have been introduced for meta-analysis such as A-MADMAN, EMAAS and EzArray[7-9]. These programs include data retrieval, management and analysis modules designed to achieve their own goals; however, it is not easy to conduct state of the art statistical methods for meta-analysis with a consistent interface using these programs. To address this problem, we have developed a software system called METASIS. METASIS was written in a Java application environment, which ensures cross-platform compatibility. Additionally, METASIS includes several approaches such as t-based modeling, the rank product method and Fisher's inverse Chi-square method.

System Implementation

Built-in Algorithms

METASIS provides a variety of statistical methods to enable interpretation and obtain biologically meaningful results from various angles. Below is an introduction to the implemented methods. First, a t-based modeling approach was built to integrate the size effect from multiple studies by measuring within- and bet-

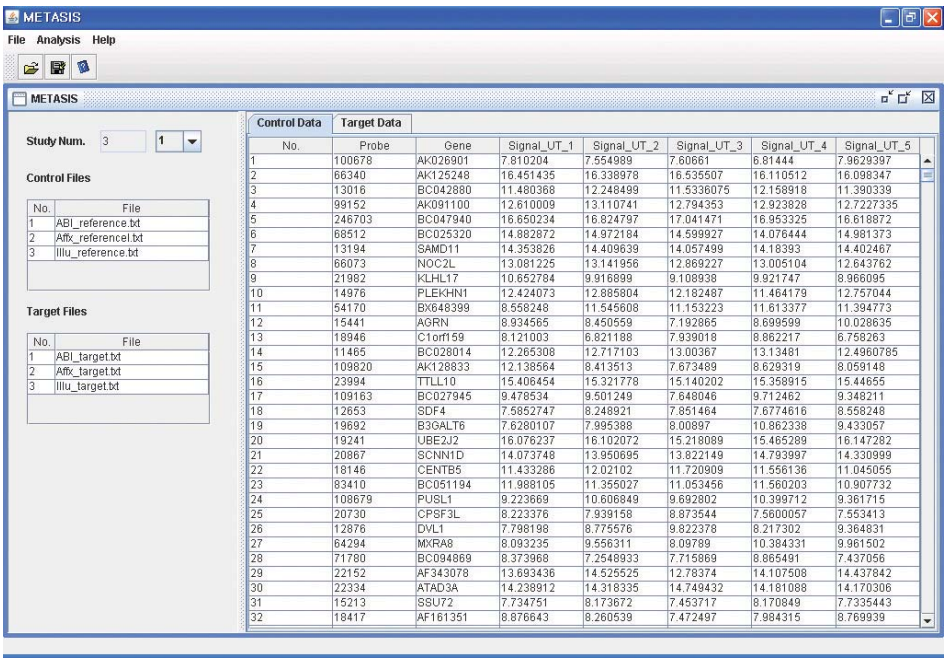


Figure 1. Input data display. The left panel provides general information regarding the submitted input files. The right panel shows the loaded data in a tabular format.

ween-study variation[10]. Following is an overview of the T-based method:

$$y_i = \theta_i + \varepsilon_i,\ \varepsilon_i \sim N(0, s_i^2)$$

$$\theta_i = \mu + \delta_i,\ \delta_i \sim N(0, \tau^2)$$

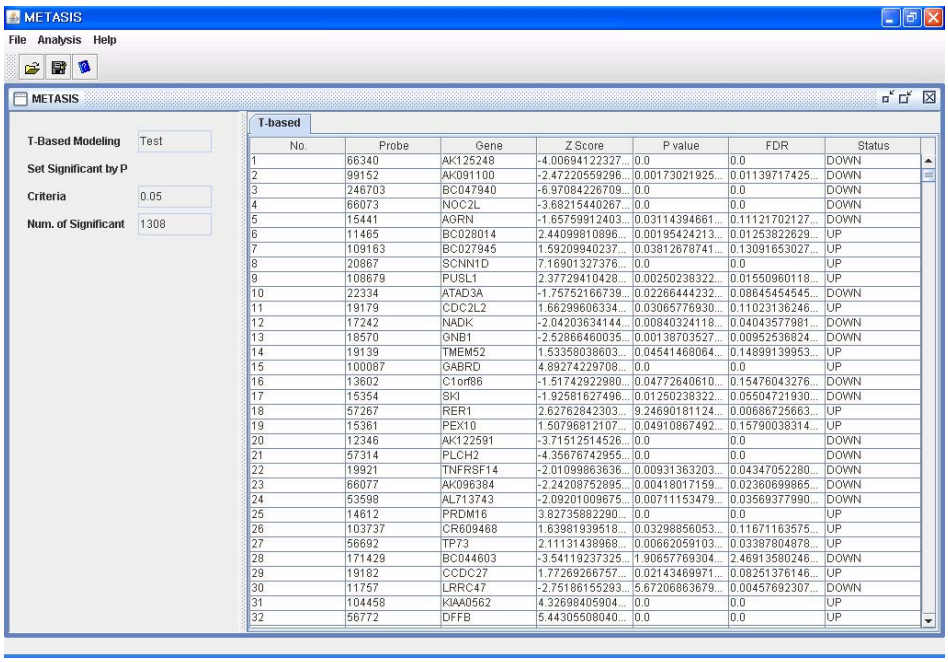
In the above formula, μ is the overall mean of interest and y_i is the observed effect size for independent studies, while s_i^2 and τ^2 represent the within study variance of sampling error conditioned on the *i*th study and the between-study variance between studies, respectively. The T-based modeling is based on one of two statistical models, the fixed-effect model or the random-effects model. The fixed-effects model (FEM) assumes that the differences in the observed effect sizes are from sampling error alone ($\tau^2 = 0$). Conversely, the random-effects model (REM) assumes that the observed effect in any given study arises for two distinct reasons: true variation in effect sizes and sample error. Therefore, the FEM can be considered as a special case of the REM. The question of which model is appropriate for a given study can be tested by determining the homogeneity of study effects by calculating the Cochran's homogeneity test statistic. To assess the statistical significance from the combined results, T-based modeling calculates p-values from an assumed standard normal distribution using obtained Z-scores for each gene. However, another statistical figure reporting the false discovery rate is also calculated based on permutation to compensate for small sample size and violation of the normality. Second, RankProd, a non-parametric statistic, assumes fewer statistical distributions than t-based modeling[11]. Specifically, it assumes that the probability of finding a specific item among the top r of n items in a list is p=r/n under the null hypothesis that the order of all items is random. Multiplying these probabilities equals the rank product, $RP = \pi_i (r_i/n_i)$, where r_i is the rank of the item in the *i*-th list and n_i is the total number of items in the *i*-th list. This statistic can be expressed by the following formula:

$$RP_g^{up} = \Pi_{i=1}^{k} r_{i,g}^{up} / n_i$$

Genes with the smallest RP values are the most interesting candidates. If the same gene appears at the top of the list in a replicate experiment, one's confidence will increase and further replicates may produce reproducible results. To assess the statistical significance for RP values of each gene, RankProd employs a permutation-based estimation procedure. Specifically, it counts how many permutated RP values smaller than or equal to a gene calculated RP value occur in a given random experiment. Third, METASIS employs Fisher's Inverse $\chi 2$ test to compute a combined statistic from the P-value obtained from the individual dataset as follows:

$$S = -2\log(\Pi_i p_i)$$

This statistic follows a $\chi 2$ distribution with 2I degrees of freedom under the joint null hypothesis and thus P-values from the combined statistic can be cal-



No.	Probe	Gene	Z Score	P value	FDR	Status
1	66340	AK125248	-4.00694122327...	0.0	0.0	DOWN
2	99152	AK091100	-2.47220559296...	0.00173021925...	0.01139717425...	DOWN
3	246703	BC047940	-6.97084226709...	0.0	0.0	DOWN
4	66073	NOC2L	-3.68215440267...	0.0	0.0	DOWN
5	15441	AGRN	-1.65759912403...	0.03114394661...	0.11121702127...	DOWN
6	11465	BC028014	2.44099810896...	0.00195424213...	0.01253822629...	UP
7	109163	BC027945	1.59209940237...	0.03812678741...	0.13091653027...	UP
8	20867	SCNN1D	7.16901327376...	0.0	0.0	UP
9	108679	PUSL1	2.37729410428...	0.00250238322...	0.01550960118...	UP
10	22334	ATAD3A	-1.75752166739...	0.02266444232...	0.08645454545...	DOWN
11	19179	CDC2L2	1.66299606334...	0.03065776930...	0.11023136246...	UP
12	17242	NADK	-2.04203634144...	0.00840324118...	0.04043577981...	DOWN
13	18570	GNB1	-2.52866460035...	0.00138703527...	0.00952536824...	DOWN
14	19139	TMEM52	1.53358038603...	0.04541468064...	0.14899139953...	UP
15	100087	GABRD	4.89274229708...	0.0	0.0	UP
16	13602	C1orf86	-1.51742922980...	0.04772640610...	0.15476043276...	DOWN
17	15354	SKI	-1.92581627496...	0.01250238322...	0.05504721930...	DOWN
18	57267	RER1	2.62762842303...	9.24690181124...	0.00686725663...	UP
19	15361	PEX10	1.50796812107...	0.04910867492...	0.15790038314...	UP
20	12346	AK122591	-3.71512514526...	0.0	0.0	DOWN
21	57314	PLCH2	-4.35676742955...	0.0	0.0	DOWN
22	19921	TNFRSF14	-2.01099863636...	0.00931363203...	0.04347052280...	DOWN
23	66077	AK096384	-2.24208752895...	0.00418017159...	0.02360699865...	DOWN
24	53598	AL713743	-2.09201009675...	0.00711153479...	0.03569377990...	DOWN
25	14612	PRDM16	3.82735882290...	0.0	0.0	UP
26	103737	CR609468	1.63981939518...	0.03298856053...	0.11671163575...	UP
27	56692	TP73	2.11131438968...	0.00662059103...	0.03387804878...	UP
28	171429	BC044603	-3.54119237325...	1.90657769304...	2.46913580246...	DOWN
29	19182	CCDC27	1.77269266757...	0.02143469971...	0.08251376146...	UP
30	11757	LRRC47	-2.75186155293...	5.67206863679...	0.00457692307...	DOWN
31	104458	KIAA0562	4.32698405904...	0.0	0.0	UP
32	56772	DFFB	5.44305508040...	0.0	0.0	UP

Figure 2. Meta-analysis result display. The left panel provides the applied methods and parameters. The right panel displays the resulting data in a tabular format.

culated. This method is the most straightforward one employed by METASIS because it computes a combined statistic from the P-values obtained from the analysis of the individual datasets. However, it is impossible to estimate the average magnitude of differential expression by working with the P-values.

Data Requirements and User Interface

METASIS can deal with multiple experimental data for all types of platforms if properly formatted text files are provided. To perform T-based modeling or rank product, METASIS accepts a tab-delimited text file as an input file to include fold change values for each sample. The first two columns record the probe identifier and gene symbol, while the rest of the columns calculate the actual fold change value. The loaded data sets are represented in a tabular format such as the one shown in Figure 1. After METASIS applies the meta-analysis process, it expresses the resulting data sets in a tabular form that includes the gene name, P-value and expression status as shown in Figure 2. The input file for Fisher's Inverse test includes a p-value for each study. Similar to a T-based modeling and rank product, it includes a probe identifier and gene symbol in the first two columns. METASIS also expresses the resulting data in a tabular form. For each process, the resulting data can be saved in a text file format.

Conclusions

METASIS is a java-based gene expression meta-analysis program that enables users to combine multiple gene expression data sets and to conduct various meta-analysis statistical methods in a consistent interface. Recently, several meta-analysis tools for gene expression datasets have been introduced such as A-MADMAN, EMAAS and EzArray. These programs work well for their specific purposes; however, they cannot be used to apply state of the art statistical methods focused on meta-analysis techniques such as T-based modeling, rank product and Fisher's Inverse $\chi 2$ test. As these methods are well known meta-analysis tools that are widely used in various research fields, METASIS will provide desirable and meaningful results for users who wish to perform gene expression meta-analysis to evaluate biological or clinical data. Finally, METASIS has minimal software and hardware requirements and all levels of users can operate it without difficulty.

Acknowledgements

Mi-Kyung Lee was supported by a Graduate Research Scholarship from the Graduate School of Kyung-Hee University in 2009.

References

1. Ramasamy, A., Mondry, A., Holmes, C.C. & Altman, D.G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* **5**, e184

(2008).
2. Bammler, T. *et al.* Standardizing global gene expression analysis between laboratories and across platforms. *Nat. Methods* **2**, 351-356 (2005).
3. GEO (Gene Expression Omnibus) http://www.ncbi.nlm.nih.gov/geo/
4. Waters, M. *et al.* CEBS-Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics data. *Nucleic Acids Research* **36** (Database issue), D892-900 (2008).
5. ArrayExpress-http://www.ebi.ac.uk/microarray-as/ae/
6. Hong, F. & Breitling, R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* **24**, 374-382 (2008).
7. Bisognin, A. *et al.* A-MADMAN: annotation-based microarray data meta-analysis tool. *BMC Bioinformatics* **10**, 201 (2009).
8. Barton, G. *et al.* EMAAS: an extensible grid-based rich internet application for microarray data analysis and management. *BMC Bioinformatics* **9**, 493 (2008).
9. Zhu, Y. & Xu, W. EzArray: a web-based highly automated Affymetrix expression array data management and analysis system. *BMC Bioinformatics* **9**, 46 (2008).
10. Choi, J.K., Yu, U., Kim, S. & Yoo, O.J. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19**(Suppl 1), i84-i90 (2003).
11. Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* **573**, 83-92 (2004).