A Weighted Hypergeometric Statistic for the Enrichment of Gene Sets

Rehman Qureshi, Ahmet Sacan
Center for Integrated Bioinformatics,
School of Biomedical Engineering, Drexel University
Philadelphia, United States
raq22@drexel.edu, as3344@drexel.edu

Abstract—**Microarray data can be analyzed by observing the activity of groups of genes; this approach can provide more relevant information than the analysis of individual gene expression. There are numerous statistical methods available for the enrichment of gene sets; however, these methods often fail to identify relevant gene sets due to the noisy nature of the data. We present weighted hypergeometric and weighted chi-squared methods that rank each gene's contribution to enrichment by the absolute value of the logarithm of its fold change. We demonstrate that these methods can produce more biologically relevant results than the standard hypergeometric test, despite being more conservative and enriching fewer pathways with significant p-values.**

Keywords-microarray; pathway; enrichment; gene set

I. INTRODUCTION

The advent of microarrays for mRNA expression experiments has resulted in several data analysis challenges. Chief among them is the problem of interpreting the lists of significant genes resulting from these experiments. Often the genes identified are too numerous and diverse to form a coherent picture of the behavior observed during the experiment. This challenge has led to the development of statistical methods that compare the gene list to groups of genes with similar classifications, or functions. Typically, a gene list is compared against groups of genes from either the Gene Ontology (GO) [1] or the pathways annotated by the Kyoto Encyclopedia of Genes and Genomes (KEGG) [2]. In this paper we focus on the enrichment of KEGG pathways, however the method we introduce is also applicable to other gene classification datasets.

Huang et al. provided a review and classification of the available enrichment tools [3]. In general, the various methods for enriching gene sets or pathways can be divided into three basic approaches: over-representation analysis (ORA), functional class scoring (FCS), and pathway topology (PT) based methods [3, 4]. In ORA, a list of genes is selected based on certain criteria such as significance, fold change, or both. ORA methods seek to determine if this list is over-represented in any pathways or groups of genes. If a subset of k genes from the list of significant genes is present in a pre-determined gene set, then the probability of finding $X \geq k$ genes is calculated. If the calculated probability is below a pre-determined threshold, then the gene set is considered "enriched" and is deemed relevant to the experiment. Typically, the chi-squared distribution, Fisher's Exact Test, the binomial probability distribution, or the hypergeometric distribution are used to calculate the probability [3].

In FCS methods, typified by Gene Set Enrichment Analysis (GSEA) [5], instead of preselecting significant genes, all the genes in the experiment are utilized in the calculation of the enrichment [4, 5]. Using all the genes for enrichment instead of a pre-selected subset can provide greater statistical power [3]. In the FCS methods, genes are ranked based on their expression. GSEA ranks the genes by their correlation with the classes of samples in the experiment. It then tests if the genes in a particular gene set are randomly distributed throughout the ranked gene list. It generates a probability distribution function by randomly permuting sample labels and generating ranked lists of genes based on their correlation with the permuted sample labels. Thus GSEA is able to estimate the significance of the ranks of the genes from a particular gene set. Estimation using this null distribution corresponds to a weighted Kolmogorov-Smirnov-like statistic [5]. Parametric Analysis of Gene Set Enrichment (PAGE) eliminates the need for permutations by calculating a z-score for each pathway and computing significance using the normal distribution [6].

The PT-based methods take into account the network topology of the genes that interconnect to form a pathway. ORA and FCS methods do not consider the connections between genes in a pathway, and will produce the same results for a pathway even if the edges of a pathway are randomized; however the PT-based methods are unsuitable for enriching GO classifications where network topology is not available [4]. ScorePAGE calculates pairwise similarity between genes in a pathway and divides the pairwise similarity scores by the number of reactions connecting the two genes [4, 7]. Signaling Pathway Impact Analysis (SPIA) assigns perturbation factors to genes in a pathway which are determined by the gene's change in expression as well as a linear function of perturbation factors of other genes in the pathway. The sum of the perturbation factors of the genes in the pathway are used to calculate a statistic called the impact factor of the pathway [4, 8].

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

In this work, we present a novel method for the enrichment of gene sets that can be considered a hybrid of ORA and FCS approaches. Like FCS methods, we consider all genes in the dataset, and we weigh each gene's contribution to its enrichment by its fold change. We propose hypergeometric and chi-squared statistics similar to those used in ORA that are weighted by the fold change of the genes.

Our approach, in principle, involves the creation of a pseudo-pathway consisting of a set of *Qm* genes, instead of the *m* genes in the original pathway or set. Unlike standard ORA, we create multiple pseudo-genes for each individual gene in the original pathway; we then apply enrichment to the number of "significant" pseudo-genes in the pseudo-pathway. We create "significant" pseudo-genes by duplicating a real gene in proportion to its fold change using the scoring metric described in Section II. We then calculate the probability of finding greater than this number of pseudo significant genes in our pseudo pathway.

II. METHODS

A. Data Management

Pathways and their corresponding sets of genes were extracted from the KEGG database and stored in an SQLite [9] relational database. Storing this information in a relational database allowed faster pathway membership determination, reducing the time required for the experiments and analysis.

Microarray datasets were downloaded from the Gene Expression Omnibus (GEO) [10]. In this study we focused on datasets that were associated with binary phenotypes. We did not utilize any datasets that contained more than two groups of samples. The datasets contained expression values for Affymetrix probes, which were converted to Entrez gene identifiers. The mapping between Entrez gene identifiers and KEGG gene identifiers was stored in the relational database. Since the relationship between Affymetrix probes and Entrez gene identifiers is not one-to-one we would often have to combine information from different Affymetrix probes. We did this by taking the mean of fold changes from different Affymetrix probes and the minimum of their p-values when multiple Affymetrix probes mapped to a single Entrez gene identifier.

B. Enrichment

Quantile normalization was performed on the expression values for each gene in each sample in the dataset. We calculated the fold change of each gene between the two conditions profiled on the array by taking the ratio of the means of the gene for each of the 2 sample groupings. For the standard ORA, which we compared to our method, we calculated the statistical significance of the change in expression for each of the genes using a two-tailed Student's t-test. When performing hypergeometric enrichment we took the set of significant genes with p-value ≤ 0.01. The probability of finding $X>k$ significant genes by random chance in a particular KEGG pathway was calculated using the formula below:

$$P(X > k) = 1 - \sum_{r=0}^{k} \frac{\binom{m}{r} \times \binom{N-m}{n-r}}{\binom{N}{n}} \quad \textbf{(1)}$$

where N is the number of genes on the array, m is the number of significant genes, n is the number of genes in the particular KEGG pathway, and k is the number of genes that are both significant and present in the particular KEGG pathway. This probability was used to determine which KEGG pathways were enriched and to determine their relative ranks.

For our weighted hypergeometric and chi-squared tests, each gene was assigned a score calculated by taking the absolute value of the logarithm of the fold change as shown in the formula below:

$$g_i = \left| log_2\left(fold\ change(gene_i)\right) \right| \quad \textbf{(2)}$$

In each dataset a value Q was calculated by taking the maximum of the gene scores in the dataset. The hypergeometric distribution is a discrete probability distribution function; however our gene scores presented themselves on a continuous scale. In order to utilize the hypergeometric distribution we had to discretize our continuous data. To do this we simply rounded all values to the nearest whole number. For each KEGG pathway we took the sum of the scores of the genes involved in the pathway to calculate k, as shown in the formula below:

$$k = \sum_{i=1}^{n} g_i \quad \textbf{(3)}$$

where n is the number of genes in a particular KEGG pathway. Each individual gene's score g_i, corresponded to the number of copies of that gene that were considered significant in the pseudo pathway. The value k corresponds to the total number of significant genes in the pseudo pathway. We then utilized the hypergeometric distribution to calculate the probability that the pathway score was greater than k according to the formula below

$$P(X > k) = 1 - \sum_{r=0}^{k} \frac{\binom{N}{r} \times \binom{QN-N}{Qn-r}}{\binom{QN}{Qn}} \quad \textbf{(4)}$$

where all of the variables represent the same quantities as they do in Equation 1, except for k which is determined by Equation 3, and all quantities are rounded to the

nearest whole number. We used this p-value to assign relative ranks to the pathways.

A similar approach was applied to the chi-squared statistic. The chi-squared statistic has been used to calculate an approximation of the exact probability, which is determined by the hypergeometric distribution. The chi-squared statistic is often employed due to the difficulty of calculating hypergeometric probabilities for large populations. The chi-squared statistic [11] is determined using the 2x2 table shown in Table I. The values from Table I are used in the equation shown below

$$\chi^2 = \frac{N(|n_{11}n_{22} - n_{12}n_{21}|)^2}{N_{1r}N_{2r}N_{1c}N_{2c}} \quad \textbf{(5)}$$

We utilize a chi-squared distribution with 1 degree of freedom, which is calculated from Table I as follows:

$$df = (r-1)(c-1) \quad \textbf{(6)}$$

where r is the number of rows in the table and c is the number of columns. We compute a weighted chi-squared statistic by constructing a table similar to Table I, but instead of the significant genes column we use the pathway score calculated by Equation 3, and the sum of the scores of all the genes on the array is used in place of the total number of genes in the pathway. Unlike the hypergeometric probability distribution, the chi-squared probability distribution is continuous, so we need not discretize our data.

III. Results

We demonstrate our methods using a microarray dataset from Njau et al. that profiles the differences between *C. Pneumoniae* infected dendritic cells and mock-infected control cells [12]. Njau *et al.* did not provide enrichment analysis for their data. We compared the results of standard hypergeometric enrichment of KEGG pathways to the results of weighted hypergeometric and weighted chi-squared enrichment. Table II shows the results of the hypergeometric enrichment. The results of the weighted hypergeometric enrichment are presented in Table III, and the results of the weighted chi-squared enrichment are shown in Table IV. In each table the top-10 most significant pathways are shown.

TABLE I. The 2x2 table used to calculate the chi-squared statistic.

	Genes on Array	**Significant Genes**	
In Pathway	n_{11}	n_{12}	$N_{1r}=n_{11}+n_{12}$
Not in Pathway	n_{21}	n_{22}	$N_{2r}=n_{21}+n_{22}$
	$N_{1c}=n_{11}+n_{21}$	$N_{2c}=n_{12}+n_{22}$	$N=n_{11}+n_{12}+n_{21}+n_{22}$

TABLE II. The results of hypergeometric enrichment of the genes that are significant at the 0.01 level

Pathway	**p-value**	**FDR**	**significant genes**
Nitrogen metabolism	0.0003	0.0805	4
Biotin metabolism	0.0008	0.0997	1
Prion diseases	0.0023	0.1867	4
Natural killer cell mediated cytotoxicity	0.0025	0.1521	9
Gap junction	0.0025	0.1246	7
Cytokine-cytokine receptor interaction	0.0026	0.1073	15
ErbB signaling pathway	0.0027	0.0960	7
Osteoclast differentiation	0.0030	0.0910	9
Non-small cell lung cancer	0.0032	0.0871	5
Vibrio cholerae infection	0.0043	0.1047	5

TABLE III. The results of weighted hypergeometric enrichment of the dataset.

Pathway	**P-Value**	**FDR**	**Score**
Glycosphingolipid biosynthesis - globo series	0.0110	1	22
Glycosphingolipid biosynthesis - ganglio series	0.0126	1	23
Glycosaminoglycan degradation	0.0193	1	26
Pantothenate and CoA biosynthesis	0.0305	1	24
D-Arginine and D-ornithine metabolism	0.0567	1	2
Protein export	0.0635	1	26
Vitamin digestion and absorption	0.0755	1	29
Thiamine metabolism	0.1011	1	6
Primary bile acid biosynthesis	0.1088	1	19
Ether lipid metabolism	0.1143	1	40

TABLE IV. The results of weighted chi-squared enrichment of the dataset

Pathway	**P-Value**	**FDR**	**Score**
Drug metabolism - cytochrome P450	0.0345	1	35.98
Amyotrophic lateral sclerosis (ALS)	0.0762	1	34.94
Pathways in cancer	0.0814	1	268.33
Cell adhesion molecules (CAMs)	0.1021	1	97.04
Calcium signaling pathway	0.1144	1	140.01
NOD-like receptor signaling pathway	0.1315	1	39.69
Glycosphingolipid biosynthesis - ganglio series	0.1418	1	23.47
Glyoxylate and dicarboxylate metabolism	0.1431	1	9.82
Glycosphingolipid biosynthesis - globo series	0.1456	1	22.15
Axon guidance	0.1734	1	102.64

Using the standard hypergeometric distribution results in 56 pathways enriched at the 0.05 level or below. As shown in Table III the weighted hypergeometric enrichment resulted in only 4 significant pathways, and as shown in Table IV the weighted chi-squared enrichment resulted in only one significant pathway. The significant pathway from the weighted chi-squared enrichment was ranked 132nd by hypergeometric enrichment and had a p-value of 0.23; that pathway contained two significant genes. The two glycosphingolipid biosynthesis pathways, whose role in the immune system is well-known, were the

most significant pathways under weighted hypergeometric enrichment; but were not detected by the standard hypergeometric enrichment because none of the genes in the pathway were significant despite their large fold changes. Although it contained no genes that were significant at the 0.01 level, the mean change in expression of the genes in the glycosphingolipid biosynthesis—globo series pathway was over three-fold. Furthermore, this pathway contained genes that were both highly upregulated and genes that were highly downregulated. Table V shows the p-values and fold changes of the genes in this pathway. The KEGG glycosphingolipid biosynthesis pathway is shown in Fig. 1.

IV. Discussion

Our weighted hypergeometric and weighted chi-squared enrichment methods are based on the idea of extending ORA to account for the magnitude of the fold change in genes and to avoid the pre-selection of significant genes. Our method allows all genes on the array to contribute to the enrichment in proportion to the change in their expression in the experiment. Our approach is a blend of the ORA methods and the FCS methods. Unlike GSEA [5] we are able to identify pathways containing both up and downregulated genes through the use of the score we assign to each gene. Our score does not assign opposite signs to the magnitude of up and downregulation. A similar idea to this has already been applied to GSEA to enable the detection of pathways with bidirectional gene expression [13].

It is evident from the results presented herein, that our weighted enrichment methods are much more conservative than the standard analysis. Because the weighted hypergeometric and weighted chi-squared enrichment are so conservative, when they are corrected for multiple comparisons they produce no significant results. We applied the Benjamini-Hochberg false discovery rate (FDR) correction [14] to the results of the enrichment analysis, as shown in Tables II-IV. When applying the FDR to the hypergeometric enrichment, as displayed in Table II, we observe that there are no pathways that can be considered significant at the 0.05 level, and the weighted enrichment methods produce Benjamini-Hochberg corrected p-values of 1 for all pathways. Gold et al. demonstrated that the Benjamini-Hochberg FDR, while useful for correcting the p-values of the individual genes on the microarray, is not suitable for the analysis of pathway enrichment because FDR is highly variable and requires many more comparisons than the number of comparisons performed during pathway enrichment [15]. It has been demonstrated that the most commonly used corrections for multiple comparisons are overly conservative and result in decreased power [3, 16]. Furthermore, the p-values resulting from enrichment analyses can be fragile and sensitive to non-statistical aspects of their determination such as the sources of data or the handling of the mapping of gene identifiers between different conventions; these issues cannot be fixed by correction for multiple comparisons [3]. Huang et al. suggest that the results of enrichment analyses are merely guidelines for the investigator and their relevance should be considered with respect to *a priori* biological knowledge [3]. As a result of these issues, we only included the FDR values in order to be thorough in the presentation of our method. As an alternative to the correction for multiple comparisons we considered the top-k pathways produced by the different methods, typically setting k to 10.

We evaluated the biological relevance of the relative rankings of the pathways produced by the various methods. The dataset we used involves the effect of *C. Pneumoniae* infection on dendritic cells. Thus, we would expect immune function related pathways to be enriched. The standard hypergeometric method produced at least two pathways related to immune function: natural killer cell mediated cytotoxicity and *V. Cholerae* infection. The other pathways in the top-10 may or may not be involved in immune processes. The top-2 pathways are nitrogen metabolism and biotin metabolism. In the results from the weighted hypergeometric enrichment the top-2 pathways are both types of glycosphingolipid biosynthesis, and the third ranked pathway involves glycosaminoglycan degradation. Glycans and glycosylation are critical to the antigen-presenting function of dendritic cells [17]. The role of glycosphingolipids in the immune system is well-known. Glycosphingolipids are surface proteins that are embedded in the plasma membrane that can act as cell-surface antigens [18, 19]. The standard hypergeometric method does not enrich these pathways, however our weighted hypergeometric method is able to capture this activity as enriched pathways and help illuminate the effect of the infection on the cells. In addition, despite finding fewer pathways that were significant at the 0.05 level, the weighted chi-squared method also identified the glycosphingolipid synthesis pathways in its top-10 pathways. However, the weighted hypergeometric method gave these pathways a better rank than the weighted chi-squared method. These pathways are likely to have greater biological relevance than the top-ranked pathways generated by the standard hypergeometric method.

TABLE V. The p-values and fold changes of the genes in the glycosphingolipid biosynthesis-globo pathway

Entrez Gene ID	Symbol	P-Value	Fold Change
2523	Fut1	0.1546	1.0791
2524	FUT2	0.0851	1.5068
2717	Gla	0.0488	7.7938
3073	HexA	0.2710	5.0730
3074	Hexb	0.1669	7.7906
4668	nagA	0.0637	2.1720
6482	ST3GAL1	0.2759	2.5360
6483	ST3GAL2	0.1305	2.0677
6489	ST8SIA1	0.3579	7.7559
8706	B3galnt1	0.0292	0.8635
10317	B3galt5	0.4758	3.8532
10690	fut9	0.2756	1.4185
26301	Gbgt1	0.3385	0.0973
53947	A4GALT	0.6219	0.6198

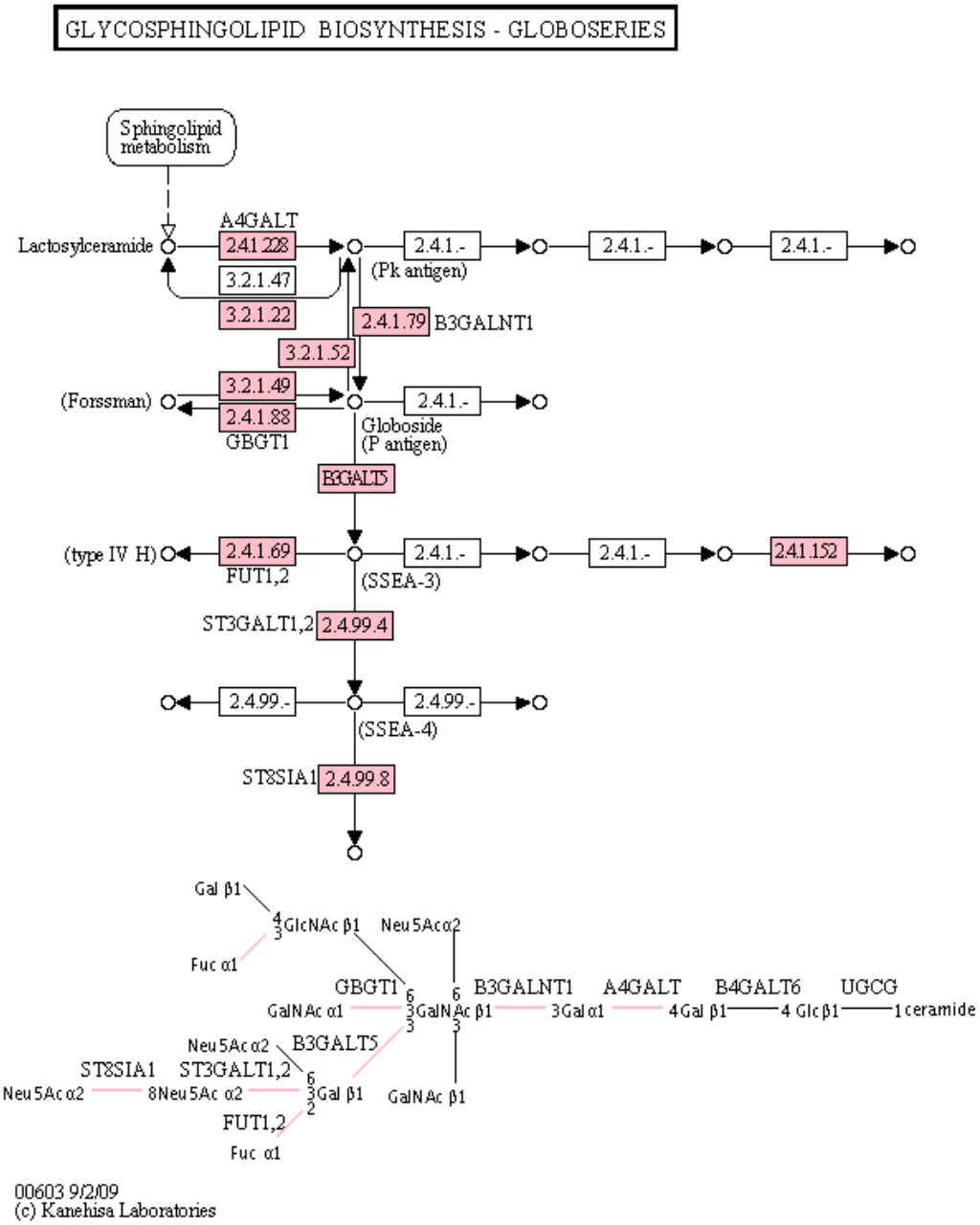

Figure 1The glycosphingolipid biosynthesis-globo series KEGG pathway, with Entrez Genes detected on the array colored pink; this was the top-ranked pathway by weighted hypergeometric enrichment.

V. Conclusion

We have presented weighted hypergeometric and weighted chi-squared methods for the enrichment of gene sets. We demonstrated the ability of these methods to produce more relevant results when enriching KEGG pathways than traditional hypergeometric enrichment, despite the problem of correction for multiple comparisons. Although we demonstrated our methods on the KEGG pathways, they are equally applicable to the Gene Ontology classifications, the evaluation of which we leave as a future work. Ultimately, we advocate the usage of multiple enrichment methods, including the one presented herein, to obtain diverse results and the usage of *a priori* biological knowledge to interpret the context and relevance of those results.

REFERENCES

1. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al*: **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Research* 2004, **32**(Database issue):D258-261.
2. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Research* 2000, **28**(1):27-30.
3. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Research* 2009, **37**(1):1-13.
4. Khatri P, Sirota M, Butte AJ: **Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges**. *PLoS Comput Biol* 2012, **8**(2):e1002375.
5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545-15550.
6. Kim S-Y, Volsky D: **PAGE: Parametric Analysis of Gene Set Enrichment**. *BMC Bioinformatics* 2005, **6**(1):144.
7. Rahnenfuhrer J, Domingues FS, Maydt J, Lengauer T: **Calculating the statistical significance of changes in pathway activity from gene expression data**. *Stat Appl Genet Mol Biol* 2004, **3**:Article16.
8. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-s, Kim CJ, Kusanovic JP, Romero R: **A novel signaling pathway impact analysis**. *Bioinformatics* 2009, **25**(1):75-82.
9. Hipp DR: **SQLite**. In.; 2007.
10. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002, **30**(1):207-210.
11. Drăghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA: **Global functional profiling of gene expression**. *Genomics* 2003, **81**(2):98-104.
12. Njau F, Geffers R, Thalmann J, Haller H, Wagner AD: **Restriction of Chlamydia pneumoniae replication in human dendritic cell by activation of indoleamine 2,3-dioxygenase**. *Microbes and Infection* 2009, **11**(13):1002-1010.
13. Saxena V, Orgill D, Kohane I: **Absolute enrichment: gene set enrichment analysis for homeostatic systems**. *Nucleic Acids Research* 2006, **34**(22):e151.
14. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **57**(1):289-300.
15. Gold DL, Miecznikowski JC, Liu S: **Error control variability in pathway-based microarray analysis**. *Bioinformatics* 2009, **25**(17):2216-2221.
16. Bluthgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D: **Biological profiling of gene groups utilizing Gene Ontology**. *Genome informatics International Conference on Genome Informatics* 2005, **16**(1):106-115.
17. Erbacher A, Gieseke F, Handgretinger R, Müller I: **Dendritic cells: Functional aspects of glycosylation and lectins**. *Human Immunology* 2009, **70**(5):308-312.
18. Ichikawa S, Hirabayashi Y: **Glucosylceramide synthase and glycosphingolipid synthesis**. *Trends in cell biology* 1998, **8**(5):198-202.
19. Uemura A, Watarai S, Iwasaki T, Kodama H: **Induction of Immune Responses against Glycosphingolipid Antigens: Comparison of Antibody Responses in Mice Immunized with Antigen Associated with Liposomes Prepared from Various Phospholipids**. *Journal of Veterinary Medical Science* 2005, **67**(12):1197-1201.