# Characterization of Gene Functional Expression Profiles of Plasmodium Falciparum

Hongbo Xie, Slobodan Vucetic, Hao Sun, Pooja Hedge, Zoran Obradovic

Center for Information Science and Technology, Temple University
1805 N Broad Str., 303 Wachman Hall, Philadelphia, PA 19122, USA

## ABSTRACT

Our objective was to provide functional characterization of gene expression related to Intraerythrocytic Developmental Cycle (IDC) of Plasmodium Falciparum. We explored a hypothesis that genes with same or similar function are likely to have similar expression profiles. Analysis of 1,051 Gene Ontology (GO) terms represented by at least two genes in Plasmodium Falciparum microarray data set revealed that gene expression profiles in 550 of them are significantly ($P<0.05$) correlated. We represented each of the 550 significant GO terms with the functional expression profile defined as average expression profile of all genes annotated with a given GO term. Using K-means clustering, we clustered 199 profiles corresponding to GO molecular functions into 4 groups. This was repeated on 228 profiles corresponding to GO biological process. We quantified the clustering quality by introducing a measure of GO term similarity defined as the minimal distance between two GO terms in GO direct acyclic graph. The results based on this measure showed that the obtained clusterings are biologically relevant which supported our hypothesis that genes with similar functions have similar expressions. This indicated that functional expression profiles can provide a valuable tool for analysis and interpretation of microarray data.

## Keywords

Bioinformatics, Microarrays, Clustering, Gene Ontology, Plasmodium Falciparum

## 1. INTRODUCTION

Microarray technology allows study of expression patterns of thousands of genes simultaneously under different conditions. The sheer amount of data obtained by microarray experiments and complexity of relevant biological knowledge present a number of challenges in biological interpretation of results. The basic underlying assumption in an analysis of microarray data is that expression profiles of functionally related genes are correlated. Following this assumption and given the appropriately preprocessed microarray data, it is a standard practice to cluster genes based on the similarity of their expression profiles and to proceed with the functional analysis of the obtained clusters. The objective of such an analysis is to confirm a specific biological hypothesis, to predict functional properties of less characterized genes, or to uncover new or unexpected biological knowledge. However, it is still an open question of how to perform the functional analysis in an objective scientific manner, as well as how to estimate the biological significance of the obtained clusters.

In this paper we address several important issues surrounding analysis of microarray data through a case study on expression data related to Intraerythrocytic Developmental Cycle (IDC) of Plasmodium Falciparum [1]. In the first part of the study we confirm a hypothesis that genes with same function have correlated expression profiles. Gene Ontology (GO http://www.geneontology.org) annotation of Plasmodium Falciparum genes according to molecular function, biological process and cellular component is used to test the hypothesis. The confirmation of this hypothesis allows us to calculate gene functional expression profile (FEPs) defined as an average expression profile of all genes assigned to a given function. In the second part of the study we use the FEPs to explore a hypothesis that genes with similar function have similar expression profiles. To validate this hypothesis, FEPs are clustered using K-means algorithm and each cluster is examined for functional similarity. We propose an objective measure of clustering quality that is based on the structure of GO hierarchy. We further illustrate how clustering of FEPs can lead to valuable insights into the developmental cycle of Plasmodium Falciparum. Finally, we discuss the consequences of the obtained results and indicate several avenues for the further research.

## 2. MATERIALS and METHODS

### 2.1 Data Set and Data Preprocessing

The data set used in the study is related to microarray hybridizations covering 46 time points during 48-hour intraerythrocytic developmental cycle of *Plasmodium falciparum* [1]. Already normalized dataset ("Complete Set") was downloaded from CAMDA 2004 web site. All arrays in this dataset have been normalized by a linear scalar (global normalization) to set the total sum of Cy3 intensity equal to the total sum of Cy5 intensity across the entire array [1]. Each array consists of 5,080 unique oligonucleotides, 4,525 of which are related to 3,532 unique gene IDs, while the remaining 555 oligonucleotides are without gene assignment. For 990 genes represented by more than one oligonucleotide their expression ratio is calculated as the average expression ratio of the corresponding oligonucleotides. For this study, the missing data were replaced with the average expression ratio over the assay. A log transform was then applied on the whole dataset.

To further polish the data, we transformed the resulting data set A to A' according to Holter el at [2]

$$A'_{i,j} = A_{i.j} - A_{*j} - A_{i*} + A_{**}, \qquad (1)$$

where $A_{*j}$ is average of microarray $j$, $A_{i*}$ is average of gene $i$ and $A_{**}$ is the overall average. The resulting dataset A' has a mean of zero among all rows (genes) and columns (microarrays).

## 2.2 Functional Expression Profiles (FEPs)

Genes are grouped based on their Gene Ontology terms provided at http://plasmodb.org. There are 1,569 unique Gene Ontology (GO) terms corresponding to the 3,532 genes. These contain 790 molecular function, 691 biological process, and 88 cellular component GO terms. Within those 1,569 GO terms, 1,051 GO terms have more than 1 associated gene expression profile associated. We further study only GO terms related to processes and functions. For each of the remaining 790+691 GO terms we calculate FEP as the average profile of all genes associated with a given GO term. We observe that the proposed calculation is simpler than FEP calculation in Bozdech et al [1].

## 2.3 Evaluating gene expression correlation

Our goal is to explore each GO term to determine the extent to which applies the hypothesis that genes with identical function have similar expression profiles. We study GO terms associated with at least two genes and use a statistical procedure to identify GO terms with average pairwise gene profile correlation significantly higher than expected by the random model. The random model assumes that genes corresponding to a given GO term are selected at random from the available pool of 3,532 genes. The algorithm used to test the null hypothesis assuming the random model is shown in table 1.

**Table 1. Evaluating the null hypothesis**

Step 1: Average pairwise correlation coefficient is calculated between $n$ gene expression profiles associated to a given GO term.

Step 2: $n$ genes are randomly selected from the whole dataset. Average pairwise correlation coefficient is computed on the random set of genes.

Step 3: Step 2 is repeated 10,000 times, and the proportion of the random sets with average pairwise correlation larger than that of the original gene set is reported as the *p-value*.

FEPs are calculated only for GO terms with p-value less than 0.05.The remaining GO terms are discarded since there is no sufficient evidence that the corresponding genes are correlated.

## 2.4 Clustering of FEPs

The K-means clustering method is applied to a given set of FEPs in order to group GO terms based on the similarity of their expression profiles. The K-means algorithm is an iterative procedure that consists of two steps. In the first step each FEP is assigned to the nearest of K centroids. In the second step, centroids are recalculated as the average of the FEPs assigned to ther clusters. This locally optimal procedure allows minimization of the Euclidean distance between FEPs and the corresponding centroid. Under the hypothesis "genes with similar function have similar expression profiles" two functions from the same cluster should be more similar than two functions from different clusters.

## 2.5 Analysis of FEP clusters using GO term hierarchy

Assigning similarity between GO terms is a highly subjective process. A proxy for measuring their similarity is given by the hierarchical nature of Gene Ontology classification which is organized as a directed acyclic graph. In our study, we define GO term distance as the length of the shortest path between their nodes within the GO hierarchy. As an example, in Figure 1, the shortest path between the terms T2.2.1 and T2.1 is 3 while the shortest path between T2.2.1 and T2.3.2 is 4. The obtained distances confirm the intuition that term T2.2.1 is more similar to T2.1 than to T2.3.2.

Since the consequence of the hypothesis "genes with similar function have similar expression profiles" is that similar GO terms are expected to be grouped into the same cluster, the introduced similarity measure provides an objective means for its testing. Therefore, the average GO term distance within a cluster is expected to be significantly larger than the distance between randomly selected GO terms. Based on such reasoning we developed a procedure that quantifies the quality of clustering. The procedure, presented in Table 2, tests the null hypothesis that assumes the random model defined as the "clusters are obtained by random groupings of GO terms."

**Table 2. Evaluating the null hypothesis**

Given $K$ clusters obtained by clustering of $m$ FEPs:

Step 1: For cluster $C_i$, $i = 1, 2, \ldots K$, with $p_i$ FEPs, FEP(1), ... FEP($p_i$), select FEP($j$) and randomly choose $p_i$-1 FEPs from the pool of the remaining $m$-1 profiles. Sum all the $m$-1 distances between FEP($j$) and randomly selected FEPs and denote it as $d_j$.

Step 2: Repeat Step 1 $p_i$ times to calculate the sum $D_i = d_1+d_2+...+d_{pi}$

Step 3: Repeat Step 2 for each of the $K$ clusters.

Step 4: Repeat Step 3 1,000 times to produce 1,000 cumulative distances $D_i$, $i = 1, 2, \ldots K$, for each of the $K$ random clusters.

Step 5: Calculate cumulative distances $D_i$, $i = 1, 2, \ldots$K, for each of the original $K$ clusters.

Step 6: Report the proportion of random $D_i$'s that are smaller than the original $D_i$ as the *p-value* of the hypothesis.

While the GO term distance definition seems reasonable, we observe its potential drawback. When the shortest path includes the root node on GO hierarchy, it is evident that the corresponding GO terms can be quite unrelated. On the other hand, the resulting distance can be quite small (e.g., the distance between T2 and T3 from Figure 1 is only 2, which is shorter than the distance between T2.2.1 and T2.3.2). Thus, we introduced a penalty on the paths that include the root of GO hierarchy. For this study, all such paths are assigned the value equal to the depth of the GO tree.

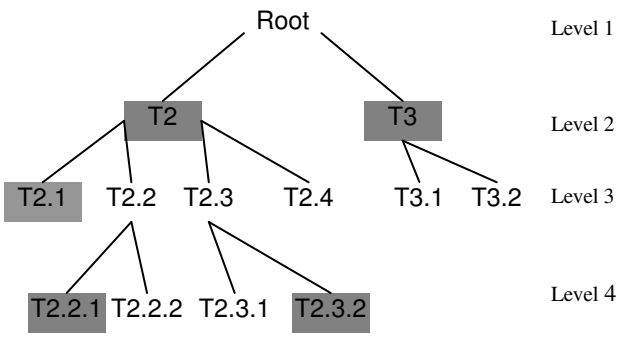


**Figure 1. Structure of GO term tree. Gene Ontology term Tx.y is a sub-function of Tx.**

# 3. RESULTS

## 3.1 Gene expression correlation of different GO terms

Using the procedure described in section 2.2, we calculate p-values for each of the 1,051 molecular function and biological process GO terms associated with at least two genes. In Figure 2 we show the number of GO terms with p-value smaller than the threshold *x*.

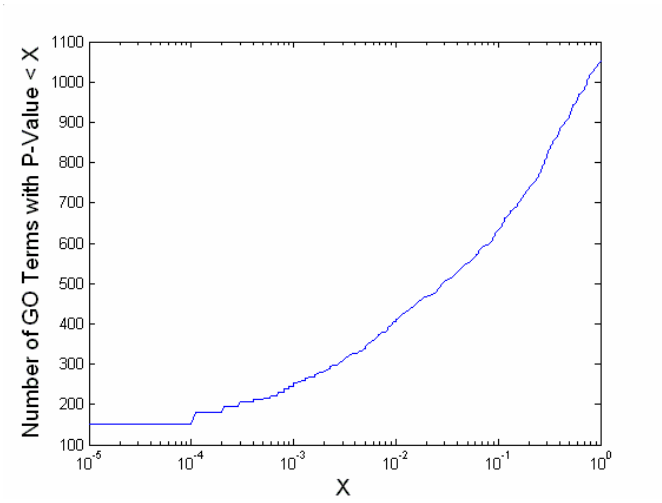


**Figure 2. Cumulative distribution of GO term p-values**

It can be observed that 52.3% (or 550) of the 1,051 GO terms have p-value less than 0.05, 199 of which are molecular function and 229 are biological process GO terms. Additional analysis (data not shown) shows that there is no linear relationship between number of genes associated with a given GO term and average correlation coefficient between these genes.

This result from Figure 2 to a large extent validates the hypothesis "genes with identical function have similar expression profiles." It also reveals that, for a given microarray experiment, one can expect a large fraction of functions that do not follow the underlying hypothesis. While this result is well known to biologists, we believe that our procedure is very effective tool for gaining a further insight into this phenomenon. Thus, ranking of GO terms based on their p-value could be useful in rapid identification of functions that are closely related with the specific developmental cycle of *Plasmodium falciparum*.

It is interesting to note that of 12 FEPs referenced by Bozdech et al (Figure 2 in [1]) two of them have p-value higher than 0.05. For example, the average correlation coefficient among genes associated with 'Ribonucleotide Synthesis' function is only 0.258 (p-value = 0.11), which weakens the claim that is related to the Ring stage of IDC.

## 3.2 Clustering FEPs

We constructed FEPs for the 199 molecular function and 229 are biological process GO terms shown to be significant in section 3.1. We proceeded by clustering the 199 molecular function FEPs into 4 groups by using the K-means algorithm. This was repeated for the 229 biological process FEPs. The clustering results are shown in figures 3 and 4.

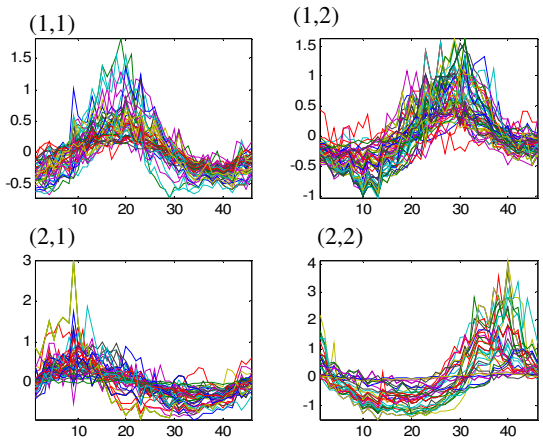


**Figure 3. K-mean clustering profiles of 199 function FEPs**

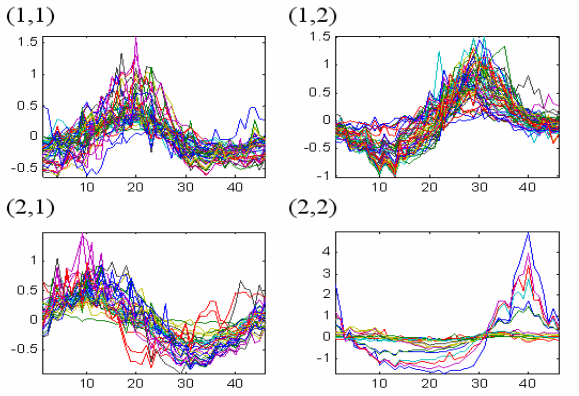


**Figure 4. K-mean clustering profiles of 238 processes FEPs**

All obtained clusters clearly show the periodic nature in agreement with findings of Bozdech et al [1]. The only exception is probably the last cluster in both figures that contains FEPs with quite flat profiles. It is interesting to note that the 4 clusters from Figures 3 and 4 are quite similar, although they correspond to quite different gene functional classification.

The 4 clusters of molecular function GO terms can be assigned into the Ring, Trophozoite and Schizont stages according to

their transcriptionally active time as shown in Tables 3 and 4. They clearly indicate that many biochemical functions and processes are transcriptional active at different stages of ICD. More specifically, our results confirm that functions such as transcription machinery listed in [1] belong to the ring stage. They also reveal additional functions related to the ring stage such as oxidoreductase, peptide binding and proteasome ATPase, as well as processes such as regulation of cell proliferation, stress response and regulation of cell cycle. Similarly, functions such as RNA binding, deoxyhypusine synthase and iron transportation, and processes such as TCA intermediate metabolism, DNA replication and deoxyribonucleotide biosynthesis have peak expression at Trophozoite stage and confirm results from [1]. Functions such as metallopeptidase, protein translocase and translation machinery, and processes such as mRNA splicing, regulation of translation and mRNA metabolism have peak expression level at schizont stage. It is interesting to note that the clusters from Figures 3 and 4 do not reveal clear boundaries between stages as some functions are active across the boundaries. For example cluster (2,2) of Figure 3 includes 35 FEPs with peak expression towards the end of Schizont stage. The summary of the clusters' association with different ICD stages is shown in Tables 3 and 4.

**Table 3. Correlation of clusters with P. falciparum ICD transcriptome stages of molecular function GO terms**

| Cluster index | Stages | Number of FEPs |
|---|---|---|
| (1,1) | Trophozoite | 48 |
| (1,2) | Schizont | 63 |
| (2,1) | Ring | 53 |
| (2,2) | Schizont-Early Ring | 35 |

**Table 4. Correlation of clusters with P. falciparum ICD transcriptome stages of biological process GO terms**

| Cluster index | Stages | Number of FEPs |
|---|---|---|
| (1,1) | Trophozoite | 78 |
| (1,2) | Schizont | 80 |
| (2,1) | Ring | 50 |
| (2,2) | Schizont-Early Ring | 20 |

## 3.3 Statistical evaluation of the FEP clusters

Using the procedure described in section 2.5, we evaluated the quality of the obtained clustering. The penalty for root crossing equal to 23 was used for biological processes while the penalty equal to 21 was used for molecular functions. A detailed statistical analysis for each cluster is shown in Tables 5 and 6.

For molecular function clustering, it can be seen that clusters (1,1) and (1,2) have much smaller overall distance than any of the 1,000 random clusters, cluster (2,1) has smaller overall distance than 89% of random clusters, while cluster (2,2) has larger distance than any of the 1,000 random clusters. When summarizing the results for 4 separate clusters it can be seen that the resulting within-cluster distance is always significantly lower than that of 1,000 random clusterings. Figure 5 illustrates a histogram of the total distances obtained by random clusterings and compares them with the distances of the obtained molecular function FEP clustering. This result validates the hypothesis "genes with similar function have similar expression profiles." It also shows that the proposed definition for GO term distance is effective, despite the fact that it depends on the GO hierarchy that is not optimized for such purpose. Thus, this distance measure could prove very useful for evaluation of the quality of different gene expression clustering algorithms.

Similarly to Table 5 and Figure 5, Table 6 and Figure 6 provide another confirmation that the underlying hypothesis is valid. However, we observe that the difference between distances of the molecular process clusters and 1,000 random clusterings is rather small. Further research is needed to reveal the reason for such a result.

**Table 5. Sum of GO term distances within each cluster (times 10,000) for the 1,000 random clusters and original molecular function clusters. The resulting p-values for each cluster.**

| | Cluster (1,1) | Cluster (1,2) | Cluster (2,1) | Cluster (2,2) | Total |
|---|---|---|---|---|---|
| Random clusters Mean±Std | 2.94 ± 0.048 | 5.19 ± 0.058 | 4.00 ± 0.046 | 2.20 ± 0.033 | 14.3 ± 0.090 |
| Original clusters | 2.20 | 4.56 | 3.94 | 2.46 | 13.2 |
| p-value | 0.0 | 0.0 | 0.11 | 1.0 | 0.0 |

**Table 6. Sum of GO term distances within each cluster (times 10,000) for the 1,000 random clusters and original biological processes clusters. The resulting p-values for each cluster.**

| | Cluster (1,1) | Cluster (1,2) | Cluster (2,1) | Cluster (2,2) | Total |
|---|---|---|---|---|---|
| Random clusters Mean±Std | 5.61 ± 0.043 | 5.57 ± 0.040 | 2.64 ± 0.030 | 0.37 ± 0.011 | 14.2 ± 0.064 |
| Original clusters | 5.54 | 5.15 | 2.96 | 0.37 | 14.0 |
| p-value | 0.047 | 0.0 | 1.0 | 0.52 | 0.0 |

We further explored biological process FEP cluster (2,2) which includes 20 GO terms. It is evident from Figure 4 that two types of expression profiles can be distinguished; one has very low variation of normalized expression levels over time, while other has a large variation with periodic behavior that characterizes other clusters. Its peak expression level corresponds to the early Ring stage. The low variation group of functional group includes 13 GO terms. Those 13 terms belong to widely different biological process families such as transportation and biosynthesis. Most of the 7 GO terms with large expression variation over time belong to cell growth and/or maintenance family. It indicates that it would be beneficial to further divide (2,2) cluster into two clusters. This would certainly result in significant reduction in within-cluster distance and provide better insight into ICD.

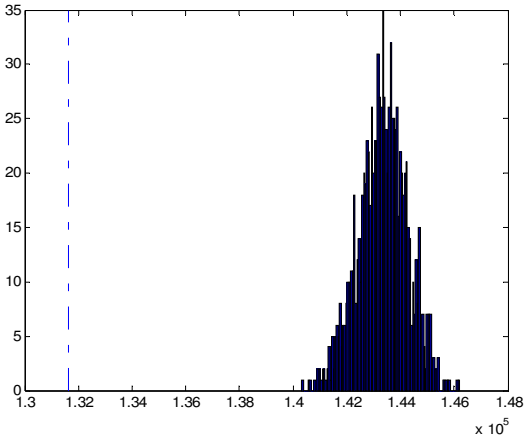


**Figure 5. Distribution of distances for 1,000 random clusterings of 199 molecular function FEPs. The dashed line is the distance for the K-means clustering.**

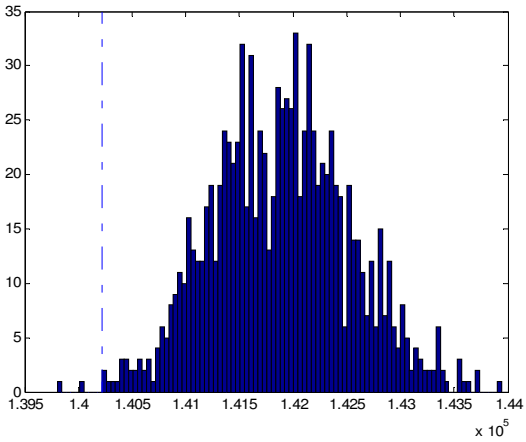


**Figure 6. Distribution of distances for 1,000 random clusterings of 199 biological processes FEPs. The dashed line is the distance for the K-means clustering.**

# 4. CONCLUSION

We studied transcriptome of Intraerythrocytic Developmental Cycle of Plasmodium falciparum based on the provided microarray dataset. Our objective was to validate a hypothesis that genes with same or similar function are likely to have similar expression profiles. We selected a set of functions based on Gene Ontology classification containing genes with highly correlated expression profiles. This set of functions covered more than 50% of the studied GO terms thus confirming the hypothesis that genes with identical function have similar expression profiles. Clustering of FEPs of the significant GO terms, followed by the objective analysis of their quality confirmed the weaker hypothesis that genes with similar function have similar expression profiles. The introduced measure showed the potential to be a useful tool in comparing different gene expression clustering algorithms. In depth study of obtained FEP clusters helped in gaining valuable biological insights into the nature of IDC of Plasmodium falciparum. Further study following these promising results is likely to lead to a useful novel methodology for microarray data analysis.

**ACKNOWLEDGEMENTS**

This study was supported in part by *The Pennsylvania Department of Health* grant to Z. Obradovic and S. Vucetic.

# 5. REFERENCES

[1] Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL, The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum, PLoS Biol. 2003 Oct; 1(1):E5.

[2] Neal S. Holter, Madhusmita Mitra, Amos Maritan, Fundamental patterns underlying gene expression profiles: Simplicity from complexity, Proc Natl Acad Sci 97 (2000) 8409–8414.