IACSIT International Journal of Engineering and Technology, Vol.2, No.2, April 2010
ISSN: 1793-8236

Clustering *Plasmodium falciparum* Genes to their Functional Roles Using k-means

Victor Chukwudi Osamor, Ezekiel Femi Adebiyi, and Seydou Doumbia

***Abstract*— We developed recently a new and novel Metric Matrics k-means (MMk-means) clustering algorithm to cluster genes to their functional roles with a view of obtaining further knowledge on many *P. falciparum* genes. To further pursue this aim, in this study, we compare three different k-means algorithms (including MMk-means) results from an *in-vitro* microarray data (Le Roch et al., Science, 2003) with the classification from an *in-vivo* microarray data (Daily et al., Nature, 2007) in other to perform a comparative functional classification of *P. falciparum* genes and further validate the effectiveness of our MMk-means algorithm. Results from this study indicate that the resulting distribution of the comparison of the three algorithms' *in vitro* clusters against the *in vivo* clusters are similar thereby authenticating our MMk-means method and its effectiveness. However, Daily et al. claim that the physiological state (the environmental stress response) of *P. falciparum* in selected malaria-infected patients observed in one of their clusters can not be found in any *in-vitro* clusters is not true as our analysis reveal many *in-vitro* clusters representation in this cluster.**

***Index Terms*— clustering algorithm; effectiveness; functional classification; malaria parasite; genes; in-vivo; in-vitro; microarray.**

I. Introduction

The complete *P. falciparum* lifecycle revolves around three major developmental stages, namely, the mosquito, human liver and human blood stages. The Intraerythrocytic Development Cycle (IDC) represents all of the stages in the development of *P. falciparum* responsible for the symptoms of malaria. It has long been a goal to understand the regulation of gene expression throughout each developmental stage.

The *P. falciparum* Intraerythrocytic Development Cycle (IDC) begins with merozoite invasion of red blood cells (RBCs) and is followed by the formation of the parasitophorous vacuole (PV) during the ring stage. This stage transform to the trophozoite stage characterized by the parasite entering into a highly metabolic maturation phase, prior to parasite replication. During the schizont stage, the cell prepares for reinvasion of new RBCs by replicating and dividing to form up to 32 new merozoites. In preparation for sexual developmental stage development, some of these merozoites differentiates into the gametocytes stages which are taken up by female *Anopheles gambiae* mosquito during blood feed from an infected patient resulting in the formation of sporozoites that migrate into the salivary gland. Using these sporozoites, female *Anopheles gambiae* is able to transmit malaria to an uninfected person through its bite for onward commencement of the human liver and RBC asexual stages.

The genome of *P. falciparum* indicates the presence of approximately 5,400 genes spread across 14 chromosomes, a circular plastid genome and a mitochondrial genome. *P. falciparum* is the causative agent of the deadly form of human malaria, affecting 200–300 million individuals per year worldwide. Insights into the biochemical function and regulation of these genes will provide the foundation for future drug and vaccine development efforts toward eradication of this disease [3]. The need to elucidate *P. falciparum* gene functions has been hampered by the fact that majority of these genes are uncharacterized and have no homology to other species since more than 60% of the predicted open reading frames (ORFs) lack orthologs in other genomes. As this fact underscores the need to elucidate functional roles of genes, many tools that have facilitated the study of model organisms remain elusive or inefficient in *Plasmodium*. Genome-wide expression profiling by microarray technology provides an easy alternative for the functional genomic exploration of *P. falciparum* [3]. Since the IDC is responsible for the symptoms of malaria, it has become the target for the vast majority of antimalarial drugs and vaccine strategies, however, some recent approaches are exploring possibilities for vaccine targeting the parasite at the liver stage [4].

A dependable classification of *P. falciparum* genes into functional and life cycle stages is from the *in-vitro* miocroarray experiment data of Le Roch *et al.*[1]. Daily *et al.* [2] used the non negative matrix factorization (NMF) algorithm [5] to classify the samples expression profiles obtained from the *in vivo* microarray experiments of the parasites from venous blood samples of 43 patients residing in Senegal into three distinct clusters. They tried to use [1] and other existing *in vitro* classifications to explain these three clusters. They found that the profiles of samples in the second cluster were similar to early ring-stage profiles of the D7 strain grown *in vitro* [1] and that the other two clusters were not observed *in vitro*.

They later interpreted these three clusters biological basis by comparing them to an extensive compendium of expression data in the yeast *Sacchromyces cerevisiae.* This

Manuscript received Feb. 23, 2010. This work was supported by Covenant University Staff Development Grant.

V. C. Osamor is with the Bioinformatics Unit, Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria (corresponding author e-mail: vcosamor@ gmail.com).

E. F. Adebiyi is with the Bioinformatics Unit, Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria. (e-mail: eadebiyi@sdsc.edu).

S. Doumbia is with the Malaria Research Training Centre, University of Bamako, Mali, Africa. (e-mail: sdoumbi@MRTCBKO.org).

comparison showed that the three states resemble, first, active growth based on glycolytic metabolism, second, a starvation response accompanied by metabolism of alternative carbon sources, and third, an environmental stress response. And therefore showed that the glycolytic state (depicted by the second cluster) is highly similar to the known profile of the ring state *in vitro* but the other two states have not been observed *in vitro*, and this revealed a previously unknown physiological diversity in the *in vivo* biology of the malaria parasite, in particular evidence for a functional mitochondrion in the asexual-stage parasite.

In this work, our original intension is to further validate the effectiveness of our new and novel MMk-means [6] algorithm, presently under publication consideration review, by comparing three different k-means algorithms (including MMk-means) results on Le Roch et al.[1] *in vitro* microarray data with the *in vivo* microarray data of Daily et al.[2]. We achieved our aim and found that the three algorithms' *in vitro* clusters against the *in vivo* clusters distribution are similar, but we however also found that while the starvation response state (depicted by the first cluster) was not observed in the *in vitro* microarray data, our comparative analysis showed that the environmental stress response state (depicted by the third cluster) can be painted from the *in vitro* data.

II. Methodology And Results

We enumerate next below the data and the algorithms employed.

A. Data used

Daily et al. [2] data were obtained using venous blood samples from *P. falciparum*-infected patients in Senegal. This consisted of patients who presented to the district hospital in Velingara, Senegal, with fever and symptoms suggestive of malaria. Le Roch et al. [1] used lab cultured samples of *P. falciparum* and reported that 2235 genes were significantly expressed. This is shown in row 1 of *Table 1* below. Daily et al. [2] data has 5159 genes in each of the 3 clusters with samples of 8, 17 and 18 respectively. We use SAM (Significant Analysis of Microarray) software [7] to extract the list of significant genes from the three clusters of Daily et al. [2] as listed in row 2 of Table 1.

TABLE 1: SHORT STATISTICS ON P. FALCIPARUM MICROARRAY EXPERIMENTAL DATA USED IN OUR COMPARATIVE ANALYSIS.

P, falciparum Microarray Experiment data		**Total No of Genes**	**Timepoints**	**List of Significant genes**
Le Roch *et al.*		5159	16	2235
Daily *et al.*	Cluster 1	5159	8	1471
	Cluster 2	5159	17	3195
	Cluster 3	5159	18	3004

B. Algorithms used

1) SAM (Significant Analysis of Microarrays)

SAM as proposed by Tusher, Tibshirani and Chu [7] is a statistical technique for finding significant genes in a set of microarray experiments. The software implementation allows input to SAM in form of gene expression values from a set of microarray experiments. SAM computes a statistic d_i for each gene i, measuring the strength of the relationship between gene expression and the response variable. It uses repeated permutations of the data to determine if the expression of any gene is significantly related to the response. The cutoff for significance is determined by a tuning parameter **delta (Δ)**, chosen by the user based on the false positive rate. SAM output list of significant genes and considers not only false positive rates, but also false negative rates. For this purpose, a *miss rate* table is also printed. It gives an estimated false negative rate for genes that do not make the list of significant genes. SAM is a licensed software that executes on Windows 2000 or higher, R and Excel 2000 or higher as an Excel add-in.

2) Robust k-means clustering algorithm

The robust k-means clustering algorithm was first used in Le Roch *et al.* [1]. The robust k-means clustering algorithm runs on top of the standard k-means clustering algorithm. Using the Pearson correlation coefficient as the similarity measurement, data were clustered by the standard k-means clustering algorithm independently for 1000 runs. Based on this 1000 results obtained, a probability matrix that any two genes belong to the same cluster is compiled and the run that best approximate the probability matrix is selected. An optimal solution on any given k is obtained as this algorithm eliminates the arbitrariness of any individual k-means run. In Le Roch *et al.* [1], trials were made for k=10, 15, 20, 25, and 30. k=15 was found to produce meaningful classification. Le Roch *et al.* [1] used expression values of 2235 significantly expressed genes across the 16 lifecycle measurements as input and reported that using a k value greater than 20 often yielded clusters with similar expression patterns suggesting that the clusters were over fragmented while on the contrary, the use of k=10 grouped unrelated genes.

3) Traditional k-means clustering algorithm

In k-means clustering, we are given a set of n data points in d-dimensional space R^d and an integer k. The problem is to determine a set of k points in R^d, called centers, so as to minimize the mean squared distance from each data point to its nearest center [8]. To solve this problem, the traditional k-means algorithm was implemented as a gradient descent procedure, which begins at starting cluster centroids (or centers) and iteratively updates these centroids to decrease the mean squared distance from each data point to its nearest center. The asymptotic expected run time for this algorithm is $O(nkl)$, where l is number of iterations.

4) Metric Matrices k-means (MMk-means) clustering algorithm

A new and novel MMk-means algorithm was developed by us in Osamor *et al.* [6] and it is simple but efficient (theoretically and at practical setting via our implementations) than the traditional k-means and the recent enhanced k-means algorithm of Fahim *et al.* [9]. The new algorithm is based on the recently established relationship between principal component analysis and the k-means clustering [10]. In MMk-means, we create a covariance matrix (r) computing the pearson product moment correlation coefficient between the k centroids of the previous and the current iterations and then deduce also k previous and current iterations eigenvalues. Using the Ding and He [10] computed threshold (when it is computationally wise from our new theoretical derivatives), we are able to determine which of the k clusters

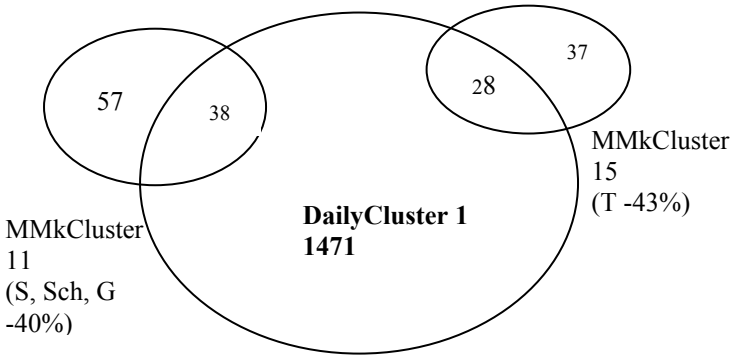
is optimally equal to the expected ones; in other words, stable (that is, its members will always remain in the same cluster in subsequent iterations). Using the above methods, the new k-means algorithm saves significant computation time at each iteration and thus arrived at an $O(nk^2)$ expected run time algorithm. Results obtained from testing the algorithm on five different types of microarray data [6] also indicate that the new MMk-means clustering algorithm is empirically faster than other known k-means algorithms.

From our prior work [6], we have implemented the traditional and MMk-means algorithms respectively. So first, we deployed them to cluster, for k=15, *P. falciparum* microarray data containing 5159 genes and 16 timepoints arising from the work of Le Roch *et al.* [1]. The traditional k-means algorithm is set a gold standard and is used to validate MMk-means algorithm while the Robust k-means clustering results from Le Roch *et al.* [1] for k=15 serve as a benchmark to compare the effectiveness of the two algorithms.

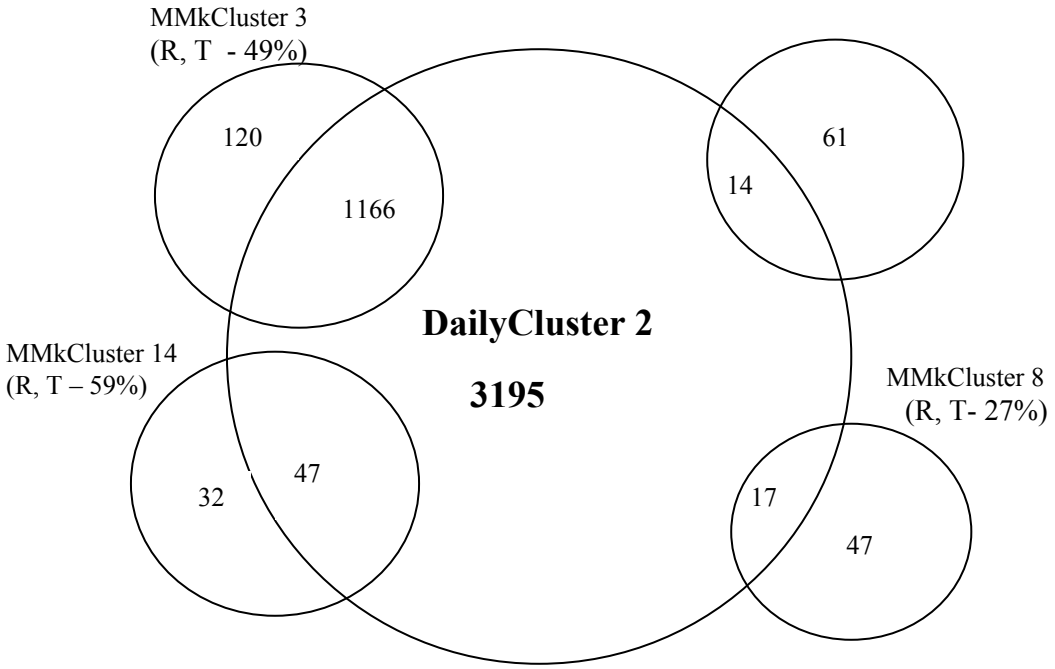
We performed analysis on the clusters output from MMk-means and traditional k-means as depicted in *Table 2*. To map genes (in clusters) of traditional k-means and MMk-means algorithms to their robust k-means counterpart, we employed Relational Database Management System (RDBMS) using Microsoft Access 2003 to design a database involving schema and table relationships for query generation and database interrogation. This data mining allowed us to compare and contrast traditional k-means and MMk-means from their percentage similarity with Le Roch *et al.* [1] clusters (*as recorded in columns 9 and 10 in Table 2*). The correlation coefficient of these data similarity is computed to be 0.7.

To further consolidate the validation of our MMk-means algorithm, we carried out comparative analysis of clusters results on Le Roch *et al.* [1] data as generated by the three (3) algorithms on Daily *et al.* [2] data. Daily *et al.* used Non-negative Matrix Factorisation (NMF) algorithm to cluster their data into three clusters. We ran Significant Analysis of Microarray (SAM) [7] at the settings of delta (Δ) = 0, data type = One Class, to extract list of significant genes that are highly expressed for each of the three clusters (*see Table 1)*. Delta setting of 0 ensures that all the significantly expressed genes are extracted, however we also obtained same number of significantly expressed genes for cluster 1 with $0 \leq \Delta \leq 11.866$, beyond this range, list of significant genes reduces.

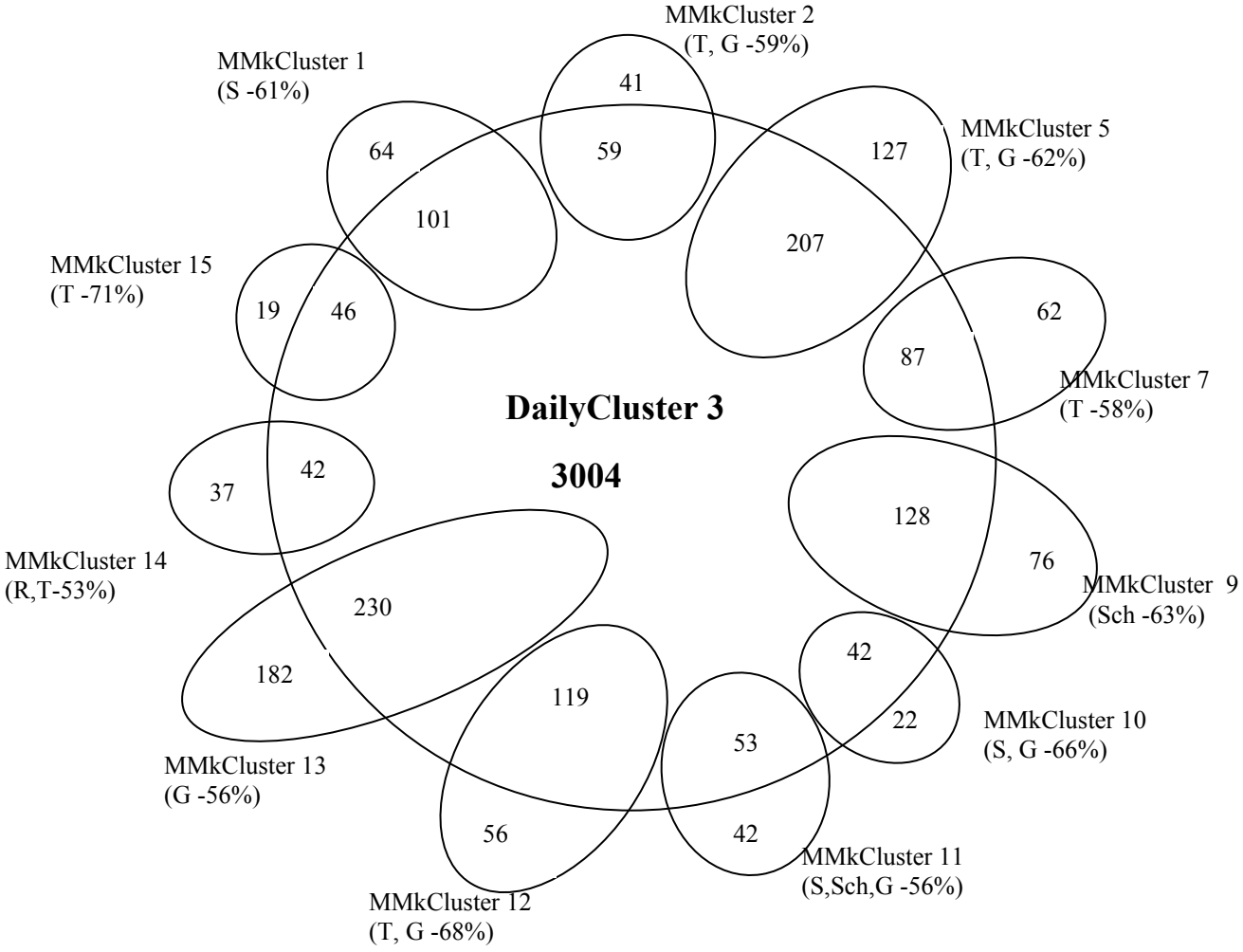
We compared clusters 1-15 from Le Roch *et al.* [1] data for each of the three k-means algorithms with each cluster of Daily *et al.*[2] and computed the percentage number of genes common to both. This resulted into other tables (see Tables 2 and 3). We placed via Venn diagrams the results of the three different k-means algorithms from the *in vitro* microarray data of Le Roch *et al.* [1] on the classification from the *in vivo* microarray of Daily *et al.* [2]. The resulting three venn diagrams are similar. Fig. 1 shows the results of our MMk-means. Fig. 2 depicts that of Robust k-means and Fig. 3 gives the venn diagram describing the results of Traditional k-means algorithm from the *in vitro* microarray data of Le Roch *et al*. [1] on the classification of Daily *et al.* [2]. However, to avoid over cluttering each venn diagram, except for cluster 2 of Daily *et al.* [2], we represented only clusters that pass the following similarity constraint: $n(X \cap Daily\ cluster) \geq 40\%$, where '*X* ' represents any cluster obtained from the runs of Robust, Traditional and MMk-means respectively and '∩' is a set notation that capture the number of elements in the intersection.



(a)


MMkCluster 3
(R, T - 49%)
120
1166
61
14
DailyCluster 2
3195
MMkCluster 14
(R, T – 59%)
32
47
MMkCluster 8
(R, T- 27%)
17
47

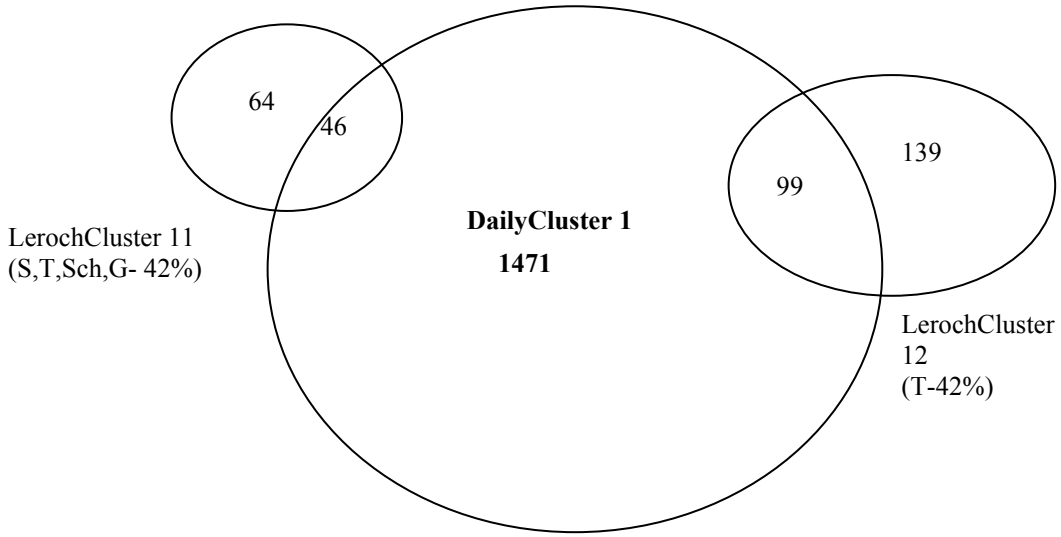
(b)


MMkCluster 2
(T, G -59%)
41
MMkCluster 1
(S -61%)
64
59
127
MMkCluster 5
(T, G -62%)
101
207
MMkCluster 15
(T -71%)
19
46
62
87
MMkCluster 7
(T -58%)
DailyCluster 3
3004
37
42
128
MMkCluster 14
(R,T-53%)
76
MMkCluster 9
(Sch -63%)
230
42
182
119
22
MMkCluster 10
(S, G -66%)
53
MMkCluster 13
(G -56%)
56
42
MMkCluster 11
(S,Sch,G -56%)
MMkCluster 12
(T, G -68%)

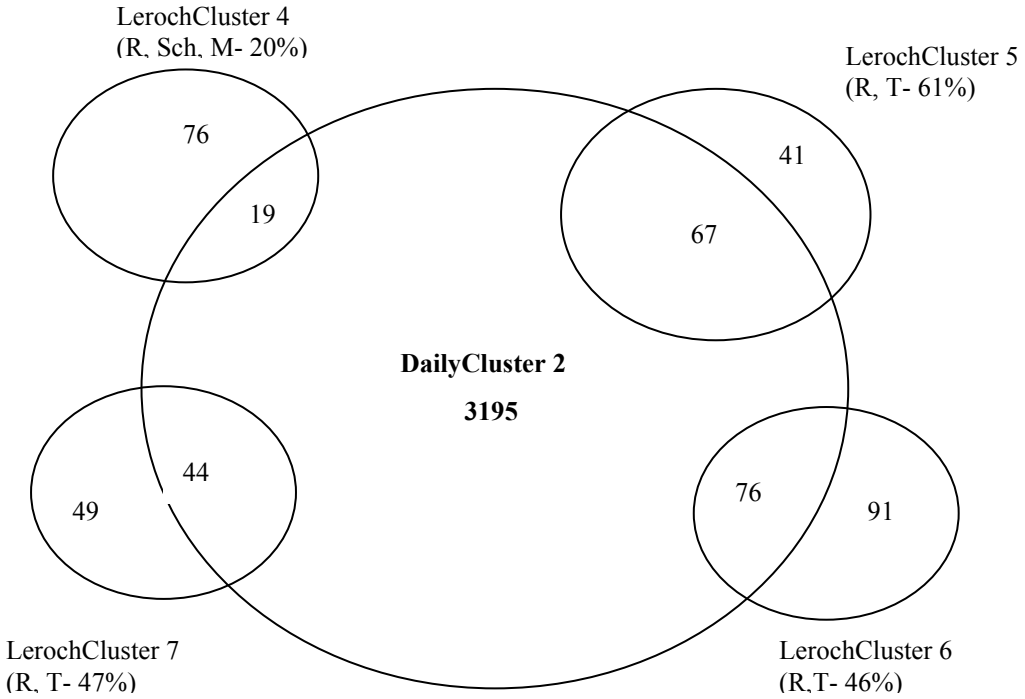
(c)

Figure1a-c: Venn diagram of MMk-means Clustered data of Le Roch *et al.*, [1] and NMF clustered data of Daily *et al.*, [2].

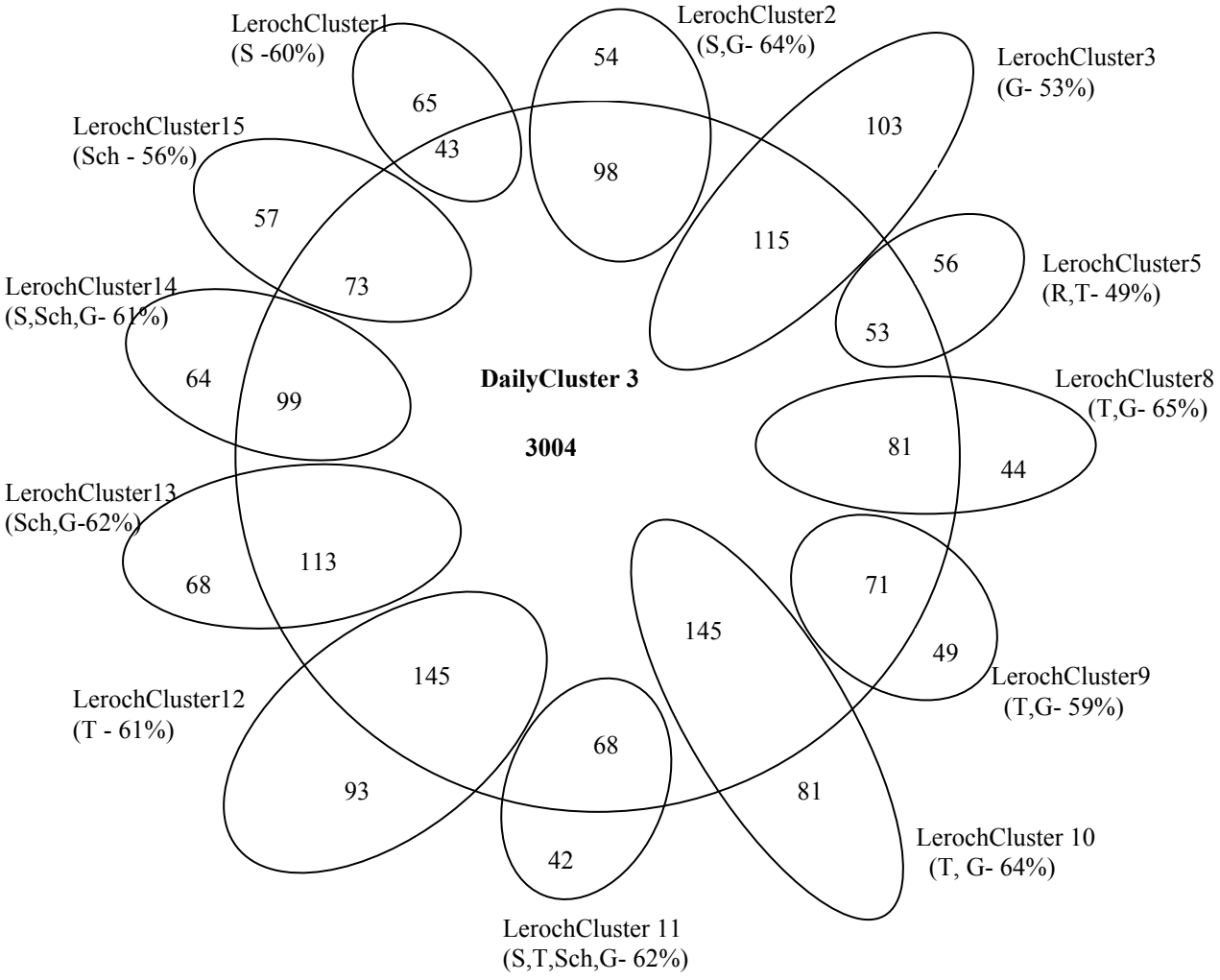
MMkCluster is the Cluster created by MMk-means, DailyCluster is the resulting cluster from Daily *et al.*, 2007. R=Ring stage, T=Trophozoite, S=Sporozoite, Sch=Schizont, G= Gametocyte, and M= Merozoite. Except in DailyCluster 2, an MMkCluster is represented in Venn diagram if its meet the similarity criterion of ≥40% of its gene content present in DailyCluster. This criterion is to avoid over cluttering of the Venn diagram. DailyCluster 2 representation had four Clusters indicted for ring stage parasite without the use of this criterion. (a) Two clusters MMkClusters 11 and 15 had ≥ 40% of entire genes in their cluster present in DailyCluster 1 of 1471 genes. (b) Only MMkClusters 3, 4, 8 and 14 indicted for Ring stage parasites are represented here irrespective of ≥ 40% similarity criterion of genes in DailyCluster 2 with 3195 genes. (c) Eleven MMkClusters have ≥ 40% of their genes content represented in DailyCluster 3 with 3004 genes.



(a)



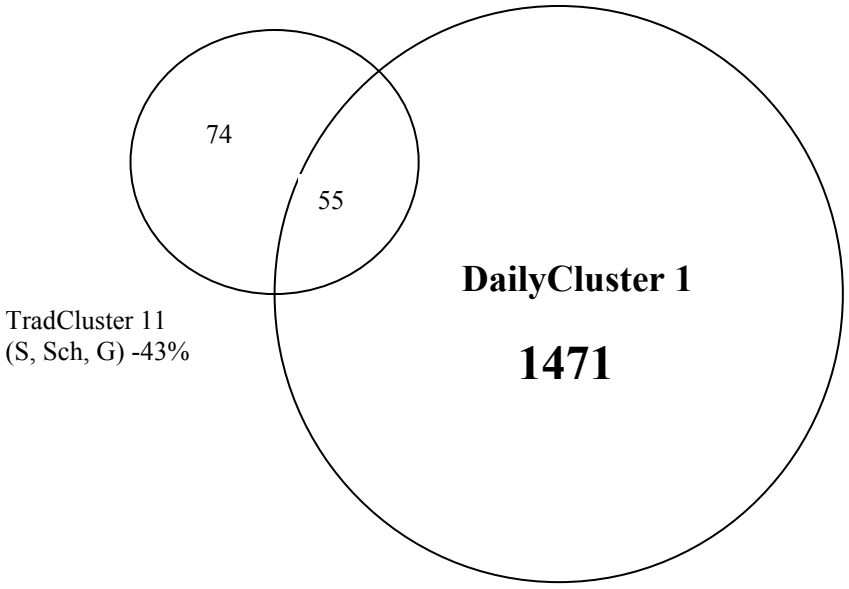
(b)



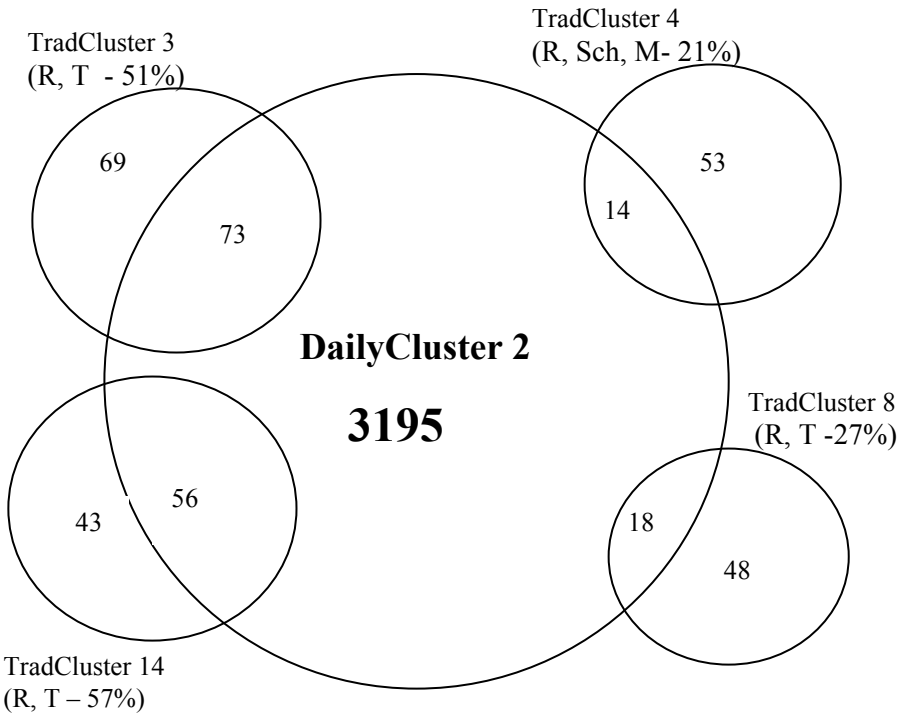
(c)

Figure 2a-c: Venn diagram of Robust k-means clustereddata of Leroch *et al.*, [1] and NMF clustered data [2].

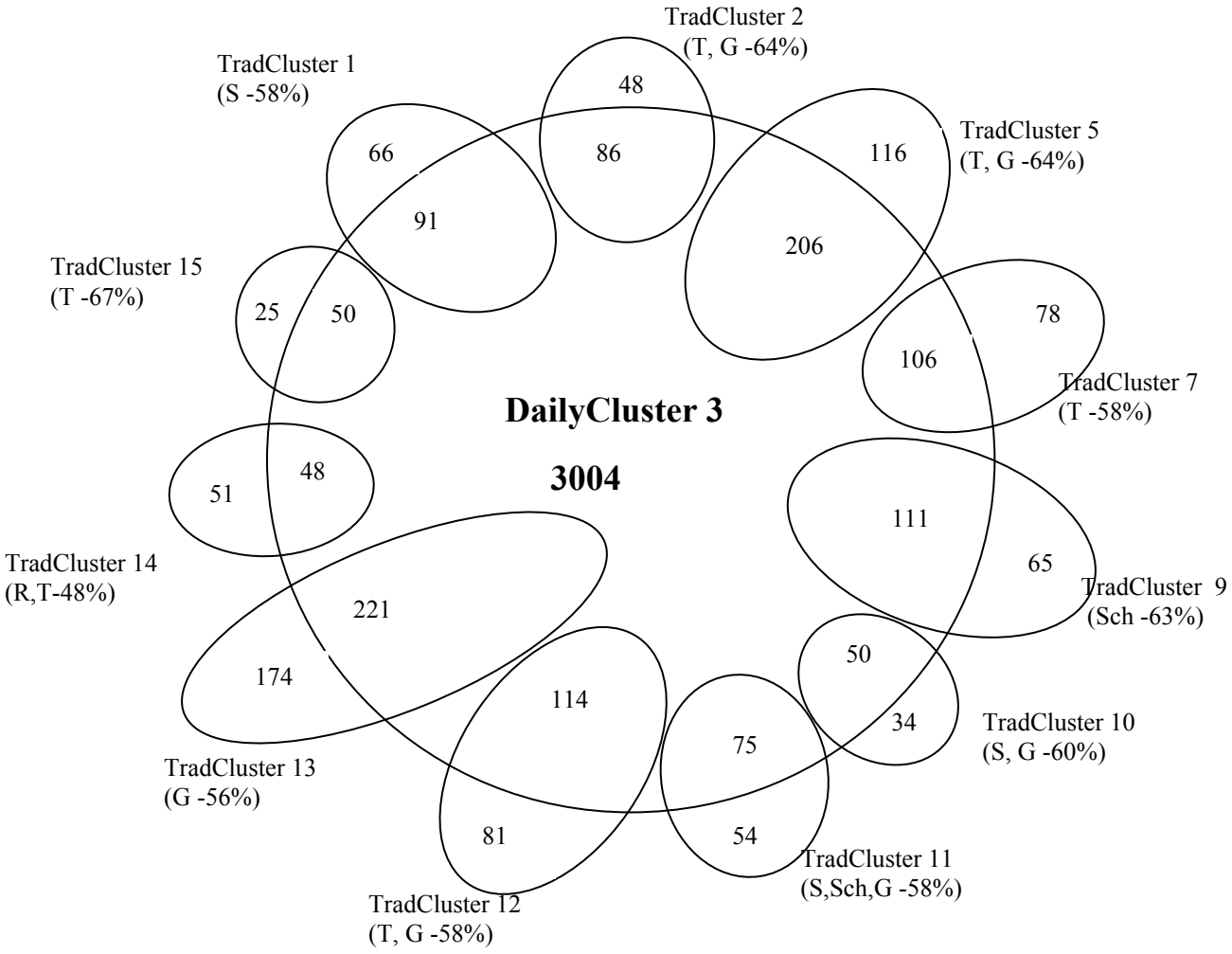
Except for DailyCluster 2, a LerochCluster is represented in venn diagram if number of genes found at intersection of each Leroch et al. cluster with any DailyCluster is ≥ 40% . Only DailyCluster 2 representation had four clusters indicted for ring stage parasite and represented without considering the criterion of ≥40%. DailyCluster 1 and 3 representation follows the ≥40% number of genes at each intersection of LerochCluster and DailyCluster. LerochCluster = Cluster created by Robust k-means, R=Ring stage, T=Trophozoite, S=Sporozoite, Sch=Schizont, G= Gametocyte, M=Merozoite. % = Proportion of the total number of genes in each cluster found at the intersection of that LerochCluster and DailyCluster multiply by 100. (a) Two clusters LerochClusters 11 and 12 had ≥40% of entire genes in their cluster present in DailyCluster 1 of 1471 genes. This criterion is to allow for a clear comparison and avoid clustering of diagrams. (b) Only LerochClusters 4, 5, 6, 7 are the four Robust k-means clusters indicted for Ring stage parasites and represented here irrespective of criterion ≥ 40% of genes in their specific cluster being present in DailyCluster 2 with 3195 genes. (c) Eleven LerochClusters have ≥ 40% of their gene content were represented in DailyCluster 3 with 3004 genes.



(a)



(b)



(c)

Figure 3a-c: Venn diagram of Traditional k-means clustered data of Leroch et al. [1] and NMF clustered data of Daily et al. [2].

Except in DailyCluster 2, for a TradCluster to be represented in Venn diagram, it will require the criterion of ≥ 40% gene content of each TradCluster to be present in DailyCluster. Only DailyCluster 2 representation had four Clusters indicted for ring stage parasite without considering the criterion of >=40%. DailyCluster 1 and 3 representation follows the >=40% gene content of each TradCluster to be present in DailyCluster. TradCluster= Cluster created by Traditional k-means, R=Ring stage, T=Trophozoite, S=Sporozoite, Sch=Schizont, G= Gametocyte, M=Merozoite. % = Proportion of the total number of genes in each cluster found at the intersection of that TradCluster and DailyCluster multiplied by 100. (a) Only TradCluster 11 had ≥40% of 129 entire genes in its cluster present in DailyCluster 1 of 1471genes. This criterion is to allow for clear comparison and avoid cluttering of diagram. (b) Only TradClusters 3, 4, 8, 14 are the four traditional kmeans clusters indicted for Ring stage parasites and represented here irrespective of criterion ≥40% of genes in their specific cluster being present in DailyCluster 2 with 3195 genes. (c) Eleven TradClusters has ≥40% of their gene content represented in DailyCluster 3 with 3004 genes.

Table 2 portrays the similarity of mm and traditional k-means respectively to robust k-means. the MMk-means and traditional k-means have same cluster id. for example, cluster 2 above has 100 and 134 highly expressed genes from MMk-means and traditional k-means respectively. out of this number, 40 and 61 of these genes from mm and traditional k-means respectively are the same with those found in Le Roch cluster 9. column 1 contains cluster id (clusters 1-15 created by MMk-means and traditional k-means for k=15), column 2 contains MMkmeans clusters membership count, column 3 contains number of only differentially expressed genes found in each cluster for MMk-means, column 4 contains tradk-means clusters membership count, column 5 contains number of only differentially expressed genes found in each cluster for traditional k-means, column 6 contains number of genes in each MMk-means cluster mapped to same gene id in same or different cluster of robust k-means, column 7 indicates number of genes in each traditional k-means cluster mapped to same gene id in same or different cluster of robust k-means, column 8 indicates approximate corresponding Le Roch cluster id to each cluster id of the traditional and mm k-means respectively (based on each robust k-means cluster having the maximum number of genes mapped to a particular traditional k-means and MMk-means cluster, we assigned an approximate corresponding cluster number as Le Roch cluster). column 9 indicates MMk-means cluster % similarity with le roch cluster (c/a) (percentage of genes common to both MMk-means cluster and robust k-means cluster for only the highly expressed genes), column 10 indicates tradk-means cluster % similarity with le roch cluster (d/a) (percentage of genes common to both traditional k-means cluster and robust k-means cluster for only the highly expressed genes). The correlation coefficient between traditional and mm k-means percentage similarity with Le Roch *et al* [1]. Clusters respectively (columns 9 and 10) shows positive correlation with a value of 0.7.

III. Discussion

The correlation coefficient of 0.7 computed from Table 2 results indicates that the MMk-means and the traditional k-means algorithms comparison to Robust k-means shows similar effectiveness. In the same vein, the results of the Venn diagrams are similar, furthering the authentication of the accuracy of MMk-means algorithm. Based on the average of 0.54 spearman rank correlation, Daily *et al.* [2] reported that the *in vivo* profiles of Cluster 2 samples were similar to early ring-stage profiles of the 3D7 strain grown *in vitro* by Le roch *et al.*[1]. We obtained this as shown in Figure 1b, where we obtained 20%, 61%, 46%, and 47% similarity respectively for each of the 4 clusters indicted to contain genes that coded for the ring-stage of the parasite.

We also verified Daily *et al.* [2] claim that the *in-vivo* expression profiles of samples in clusters 1 & 3 were not similar to those of rings (0.12 & 0.26) or late stages (0.06 & 0.01) of the asexual parasite life cycle *in-vitro,* but were only weakly similar to the profiles of other developmental states such as gametocytes (0.31 & 0.23) or sporozoite (0.35 & 0.33). For cluster 1, this is evident from Figure 1a as only 1 out of 15 *in-vitro* clusters formed a reasonable intersection with it. However, cluster 3 comparison with the *in-vitro* clusters is not in accordance with their claim (see Figure 1c), because 11 clusters out of 15 *in-vitro* clusters formed reasonable intersection with cluster 3, showing that the physiological state (the environmental stress response) of *P. falciparum* in the selected malaria-infected patients observed in cluster 3 actually exist in the *in-vitro* profiling data of Le Roch *et al.*[1].

IV. Conclusion

This work authenticated our new and novel MMk-means algorithm [11] and also delivered a biological viable result that is missing in Daily *et al.*[2] results. We achieved our aim and found that the three algorithms *in-vitro* clusters against the *in-vivo* clusters distribution are similar. We however, also found that while the starvation response state (depicted by the first cluster) was not observed in the *in-vitro* microarray data, our comparative analysis showed that the environmental stress response state (depicted by the third cluster) can be painted from the *in-vitro* data. Part of our results had been published in Osamor et al. [12].

Acknowledgment

This work was funded by the Covenant University Staff Development grant. We are grateful to J. P. Daily and J. Oyelade, for useful discussions. We also appreciate K. Le Roch and J. P. Daily for making their microarray data available. This work was concluded, completely written and revised with the other authors, while EA was a Guest scientist Dec./Jan., 2008 at the German Cancer Research Center, Heidelberg, Germany.

References

[1] K. G. Le Roch, et al., "Discovery of gene function by expression profiling of the malaria parasite life cycle," Science 2003, 301, 1503–1508.

[2] J. P. Daily, D. Scanfeld, N. Pochet, K. Le Roch, et al., "Distinct physiological states of *Plasmodium falciparum* in malaria-infected patients," *Nature* 2007, 450, 1091-1095.

[3] Z. Bozdech, M. Llinas, B. L. Pulliam, E. D.Wong, J. Zhu, J.L. DeRisi, "The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium* ***falciparum,"*** *PLoS Biol* 2003, 1, E5.

[4] A.S. Tarun, X. Peng, R.F.Dumpit, Y. Ogata , H. Silva-Rivera, et al.., "A combined transcriptome and proteome survey of malaria parasite liver stages," PNAS 2008, 105 (1) 305-310.

[5] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," Proc. Natl Acad. Sci. USA 2004, 101,4164–4169.

[6] V. Osamor, E. Adebiyi, J. Oyelade and S. Doumbia, "Reducing the Time Requirement of k-means Algorithm," unpublished.

[7] V. Tusher, R. Tibshirani and G. Chu, "Significance analysis of microarrays applied to transcriptional responses to ionizing radiation," *Proc. Natl. Acad. Sci. USA* 2001, 98:5116–5121.

[8] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman and A.Y. Wu, "A local search approximation algorithm for k-means clustering", *Computational Geometry* 2004, *28(2-3):89-112.*

[9] A.M. Fahim, A.M. Salem, F.A. Torkey, M.A. Ramadan, "An efficient enhanced kmeans clustering algorithm", Journal of Zhejiang University SCIENCE A, 2006, 7(10):1626- 633. Available online at www.zju.edu.cn/jzus

[10] C. Ding and X. He, "K-means Clustering via Principal Components Analysis", *.ACM Int. Conf. Proc. Series, 2004, 69.*

[11] V. C. Osamor: Simultenous and Single Gene Expression Computational Analyisis for Malaria Treatment Discovery, PhD Thesis, Covenant University, Ota, Nigeria, 2009.

[12] V. Osamor. , E. Adebiyi., S. Doumbia: Comparative Functional Classification of Plasmodium falciparum Genes using k-means Clustering. IACSIT-SC, IEEE Computer Society Press, (2009), Pp 491-495. Available online at http://doi.ieeecomputersociety.org/ 10.1109/IACSIT-SC.2009.107

TABLE 2: MMK-MEANS AND TRADITIONAL K-MEANS CLUSTERS WITH THEIR EQUIVALENT CORRESPONDING CLUSTERS IN LE ROCH ET AL., 2003

Cluster ID (k=15)	**MMk-means Member Count**	**MMk-means Diff. Exp count (a)**	**Tradk-means Member Count**	**Traditional Diff Exp count (b)**	**No of MMk-means Genes in Equiv. Le Roch Cluster (c)**	**No of Tradk-means Genes in Equiv. Le Roch Cluster (d)**	**Approx. Corresp. Le Roch Cluster ID**	**Mmk-means % Similarity with Le Roch Clusters (c/a)**	**Tradk-means % Similarity with Le Roch Clusters (d/a)**
1	478	165	443	157	101	100	1	61%	64%
2	233	100	347	134	40	61	9	40%	46%
3	574	236	383	142	116	93	6	49%	65%
4	178	75	166	67	66	60	4	88%	90%
5	743	334	678	322	146	152	10	44%	47%
6	147	18	137	10	4	4	1	22%	40%
7	290	149	366	184	68	92	12	46%	50%
8	163	64	167	66	24	24	7	38%	36%
9	350	204	342	176	116	62	15	57%	35%
10	142	64	176	84	51	62	2	80%	74%
11	172	95	216	129	69	56	14	73%	43%
12	442	175	456	195	67	100	8	38%	51%
13	655	412	627	395	212	209	3	51%	53%
14	426	79	440	99	44	52	5	56%	53%
15	166	65	215	75	45	35	12	69%	47%
	5159	**2235**	**5159**	**2235**					

TABLE 3 : ANALYSIS OF MMKMEANS CLUSTERED DATA OF LE ROCH *ET AL.* 2003 AND NMF CLUSTERED DATA OF DAILY *ET AL.* (2007)

MMk-means Cluster ID	**MMk-means k15 Diff Exp Gene count (a)**	**Stages**	**DAILY07_CLST1 vs MMk-meansLeroch03**			**DAILY07_CLST2 vs Mmk-meansLeroch03**			**DAILY07_CLST3 vs Mmk-meansLeroch03**		
			No of Genes Present in DailyCLST1 (b)	No of Genes Absent in DailyCLST1 (a-b)	% age of Genes Present in DailyCLST1 (b/a)	No of Genes Present in DailyCLST2 (c)	No of Genes Absent in DailyCLST2 (a-c)	% age of Genes Present in DailyCLST2 (c/a)	No of Genes Present in DailyCLST3 (d)	No of Genes Absent in DailyCLST3 (a-d)	% age of Genes Present in DailyCLST3 (d/a)
1	165	Sporozoite	58	107	35%	107	58	65%	101	64	61%
2	100	Trophozoite, Gametocyte	35	65	35%	66	34	66%	59	41	59%
3	236	Ring, Trophozoite	67	169	28%	116	120	49%	87	149	37%
4	75	Ring,Schizont, Merozoite	20	55	27%	14	61	19%	11	64	15%
5	334	Trophozoite, Gametocyte	102	232	31%	197	137	59%	207	127	62%
6	18	Sporozoite	5	13	28%	8	10	44%	5	13	28%
7	149	Trophozoite	51	98	34%	90	59	60%	87	62	58%
8	64	Ring, Trophozoite	12	52	19%	17	47	27%	14	50	22%
9	204	Schizont	76	128	37%	134	70	66%	128	76	63%
10	64	Sporozoite, Gametocyte	18	46	28%	39	25	61%	42	22	66%
11	95	Sporozoite, Schizont, Gametocyte	38	57	40%	57	38	60%	53	42	56%
12	175	Trophozoite, Gametocyte	53	122	30%	108	67	62%	119	56	68%
13	412	Gametocyte	109	303	26%	268	144	65%	230	182	56%
14	79	Ring,Early Trophozoite	22	57	28%	47	32	59%	42	37	53%
15	65	Trophozoite	28	37	43%	45	20	69%	46	19	71%
	2235		694			1313			1231		