Functional Enrichment of human protein complexes in Malaria Parasites

Jumoke Soyemi
Department of Computer Science
The Federal Polytechnic
Ilaro, Nigeria
Jumoke.soyemi@federapolyilaro.edu.ng

Itunuoluwa Isewon, Jelili Oyelade & Ezekiel Adebiyi
Department of Computer & Information Sciences
Covenant University
Ota, Nigeria
Ola.oyelade@covenantuniversity.edu.ng

***Abstract*— This study extracted differentially expressed genes (DEG) from a RNA-Seq gene expression experiment of human red blood cells for both case and control. A protein interaction network (PIN) for the DEG at the red blood stage was extracted from protein interaction database. From the protein interaction network built, we identified 64 protein complexes using the molecular complex detection (MCODE) algorithm in Cytoscape. The functional enrichment of the identified protein complexes revealed functions related to rRNA processing, Ribosome biogenesis, RNA metabolic process, cellular process, Nucleic and metabolic process and much more which are active in the RBCs that could be open to invasion by *Plasmodium falciparum*.**

Keywords—Differentially expressed genes; MCODE; Cytoscape; Red blood stage; RNA-Seq gene expressing data.

I. INTRODUCTION

Protein complexes are physical links created among two or more proteins resulting from biochemical events. Vital molecular procedures inside cells are performed using molecular mechanisms from protein-protein interactions [1][2][3]. Protein-protein interactions (PPIs) are crucial for all biological processes. Therefore, identifying PPI networks provide many new insights into protein functions. [4]. An important step in gaining insight into the composite molecular associations in living organisms is to map protein-protein physical interactions since proteins hardly act alone. Most often, they interact to form molecular mechanisms with complex physiochemical dynamic relationships that take part in biological activities at the cellular and systems levels.

Red blood cells also known as erythrocytes happen to be the most popular blood type among blood cells. It is also the vertebrate organism's major methods of transporting oxygen to the tissue of the body through the blood flow in the circulatory system. Cell nucleus and most organelles are absent in human mature red blood and the number of new red blood cells produced per second are two million and four hundred [5]. Erythrocytes grow in the bone marrow, circulating for about 20 seconds within 100 to 120 days in the body after which macrophages reprocess their components. A quadrant of the human body is made up these red blood cells [6][7].

The parasite, *Plasmodium falciparum* is the root initiator of malaria in human. There are virtually more than 100 different *Plasmodium species* of the parasite but only four are responsible for causing this disease in human. *Plasmodium falciparum* happen to be the lethal of the four [8][9]. The infected mosquito after taking in blood meal from an individual it bites; injects its infected saliva into the bloodstream of such an individual in the form of sporozoite. The sporozoite thereafter invades the liver cell where each structure develops into schizont within a week or two. A schizont is a formation with thousands of small rounded merozoites. When the schizont matures, it ruptures to release thousands of merozoites into the bloodstream [10][11].

The invasion of parasite in the erythrocytes takes place when merozoites discharged from infected RBCs attaches to the surface of uninfected RBCs. This invasion often takes place within 30 seconds [12][13]. The first interaction in merozoites and RBCs is random after which positive invasion must actively re-position using actin filaments and myosin-based motors, to make the red cell membrane come in contact with the apical end [14][15].

Therefore, identifying protein complexes responsible for invasion as well as functions of the proteins enriched will give a better insight into invasion process and contribute to knowledge require for future development of therapy against the malaria parasite.

Section I introduces the general concept of the work, section II discusses the material and methods required to implement the study. The results and discussion was done in Section III while section IV concludes the study.

II. MATERIALS AND METHODS

A. RNA-seq Gene Expression Data and Analysis

RNA-seq gene expression data, analyzed for one hundred and sixteen Indonesian patients infected with the malaria parasite, *P. falciparum* was reported in [16]. RNAs from red blood cell containing a mixture of host and parasite transcripts was extracted and mapped the RNA-seq tags to the human and *P. falciparum* reference genomes to separate the respective tags. The study was thus able to simultaneously analyze expression patterns in both humans and parasites. The human case and control data in this case contains 18684 normalized data each. In our study, we analyzed both the case and control data of human red blood cells for differential expressions.

978-1-5090-4642-3/17/$31.00 ©2017 IEEE

B. *Differentially Expressed Genes (DEG) and GO Functional Enrichment*

To extract DEGs for human RBCs, t-test was the method used. The t-test evaluates the means of two groups whether they are statistically different. This t-test was employed because it is the appropriate analysis method to compare the means of two groups especially in our case that we are looking at the case and control datasets. The Gene ontology functional enrichment of the DEGs was performed using PANTHER Overrepresentation Test (release 20170413) from GO Ontology database[17]. The Bonferroni correction for multiple testing was used for the GO biological process.

C. *Protein Interaction Data*

For human RBCs, Protein interaction data for proteins that are synthesized by genes that are differentially expressed in the RBCs was extracted from BioGRID databases [18]. The PPIs from BioGRID database is a combination of both experimental and computational dataset. BioGRID is an interaction database with data collected through comprehensive curation efforts. The BioGRID is presently expanded to include interactions, chemical associations and post-translational modifications from many publications. g:Profiler is a web tool that performs the extraction of the PPIs from the BioGRID database. g:Profiler also makes it possible for gene lists in the form of protein-protein interaction networks to be interpreted. The tool takes the advantage of a unique statistical approach to discover sub-networks of interaction that are significant in gene lists. g:Profiler performs sub-network enrichment analysis on both single gene and multiple gene lists.

The approach in g:Profiler involves set of major genes and neighbourhood genes with the major genes as the initial input list linked by an association while the neighbourhood genes involve direct interaction partners of the major set that are not related to the initial input. To determine whether the input list has more significant interacting genes than what is expected, hyper genomic test is used. G:Profiler incorporates an extension of this approach for list of ordered genes. In our study, several majors and neighbourhood genes are taken to be similar to our incremental enrichment analysis and the major genes with low p-values are reported as the final result.

D. *Protein Complex Detection with MCODE Algorithm*

Analysis of the network for clusters of the protein interaction network (PIN) using the Molecular Complex Detection (MCODE) algorithm was performed in Cytoscape (version 3.4.0) [19]. Node score cut-off was set at 0.2, false degree cut-off was set at 2, K-core was set at 2 and maximum depth from seed was set at 100. Functional enrichment of our protein complexes was done using Database for Annotation, Visualization and Integrated Discovery (DAVID) [20].

The MCODE algorithm has an advantage over other graph based algorithms because it had a directed mode that could select sub-networks of interest without using the global network thus permitting sub-networks interconnectivity assessment. MCODE [21] is among the pioneer computational approach for complexes prediction from PPI networks. This algorithm implements an agglomerative method that works in three stages; a stage that weighs the protein, another stage that extracts and the last stage which also optional is the post-processing of complexes [22].

III. RESULTS AND DISCUSSION

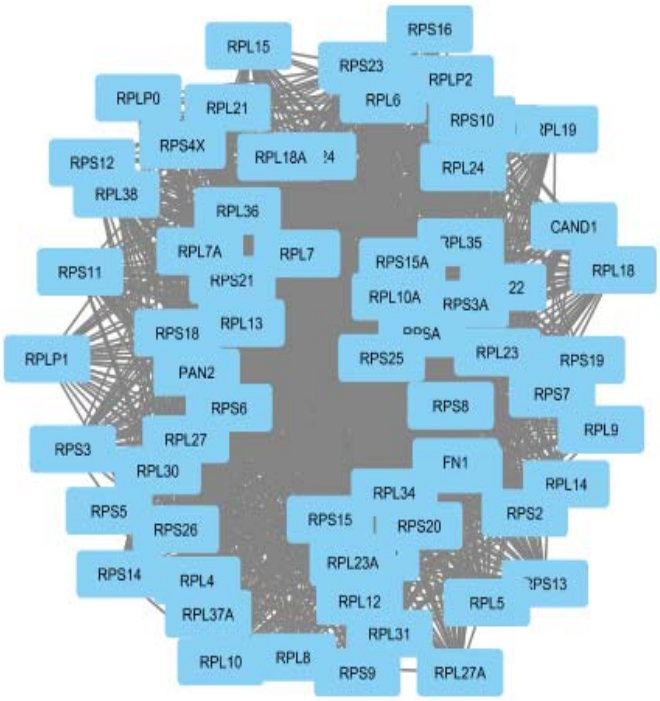
The gene expression data for human RBC returned a total of 23,175 DEGs and a total of 14,111 DEGs (see supplementary file S01) after adjusting the P-value to 0.05 so as to decrease the false discovery rate. The gene ontology of the 14,111 DEGs with P value < 0.05 is shown in table 1. The detailed report of table 1 can be found in supplementary file SO2.

TABLE 1 GENE ONTOLOGY FUNCTIONAL ENRICHMENT

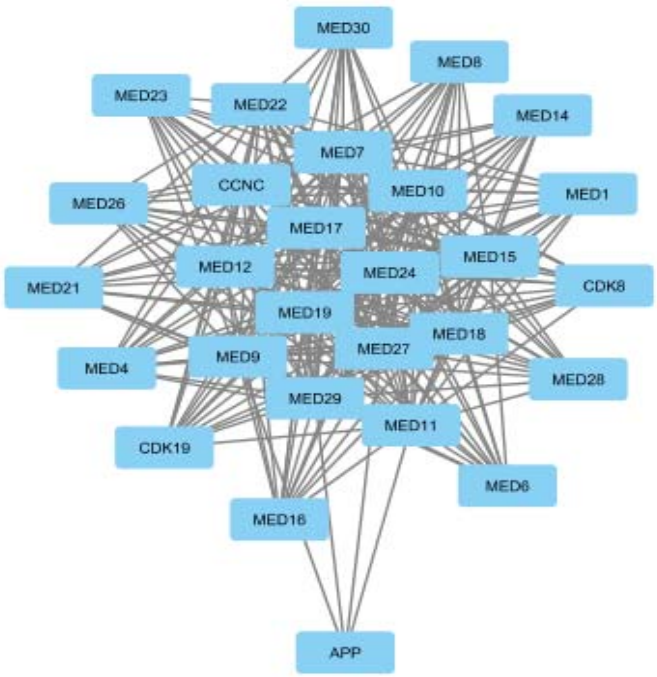
Go biological process	Homo sapiens (Ref)	Result count	Expected	Fold Enrichment	P value
Protein targeting to ER	102	41	17.44	2.35	8.50E-03
Establishment of protein localization to endoplasmic reticulum	106	41	18.12	2.26	21.4E-02
rRNA processing	259	93	44.27	2.10	7.2E-07
Ribosome biogenesis	333	109	56.92	1.91	3.33E-06
Ribonucleoprotein complex biogenesis	466	144	79.66	1.81	2.84E-07
rRNA metabolic process	286	97	48.89	1.98	5.29E-06
ncRNA metabolic process	559	164	95.55	1.72	5.68E-07
RNA metabolic process	3435	706	587.17	1.20	7.98E-04
Cellular macromolecule metabolic process	6977	1384	1192.62	1.16	9.92E-08
Cellular metabolic process	8966	1743	1532.61	1.14	7.65E-09
Cellular process	14947	2708	2554.98	1.06	4.93E-05
Metabolic process	9909	1907	1693.81	1.13	5.20E-09
Macromolecule metabolic process	7726	1501	1320.65	1.14	2.86E-06
Organic substance metabolic process	9462	1815	1617.40	1.12	1.82E-07
Nucleic acid metabolic process	3961	805	677.08	1.19	4.43E-04
Nucleobase-containing compound metabolic process	4523	896	773.14	1.16	4.47E-03
Organic cyclic compound metabolic process	4959	988	847.67	1.17	2.82E-04
Cellular nitrogen compound metabolic process	5142	1025	878.95	1.17	1.18E-04
Nitrogen compound metabolic process	8583	1651	1467.14	1.13	2.54E-06
Heterocycle metabolic process	4698	931	803.06	1.16	2.23E-03
Cellular aromatic compound metabolic process	4735	939	809.38	1.16	1.05E-03
Primary metabolic process	9102	1745	1555.86	1.12	1.05E-06
ncRNA processing	400	134	68.37	1.96	7.08E-09
RNA processing	877	236	149.91	1.57	1.51E-07

Gene Expression	3224	773	636.57	1.21	2.95E-05
Translation	386	106	65.98	1.61	2.42E-02
Cellular macromolecule biosynthetic process	3665	752	626.48	1.20	3.37E-04
Macromolecule biosynthetic process	3729	762	636.48	1.20	5.40-03
Organic substance biosynthetic process	4750	935	811.95	1.15	6.40E-03
Biosynthetic process	4817	948	823.40	1.15	5.30E-03
Cellular biosynthetic process	4662	921	796.90	1.16	4.49E-03
Organonitrogen compound biosynthetic process	1434	318	245.12	1.30	1.78E-02
Amide biosynthetic process	475	126	81.19	1.55	1.59E-02
mRNA metabolic process	663	164	113.33	1.45	2.74E-02
Unclassified	3660	479	625.63	.77	0.00E00
Fc-gamma receptor signalling pathway involved in phagocytosis	131	4	22.39	<0.2	1.94E-02
	131	4	22.39	<0.2	1.94E-02
Immune response-regulating cell surface receptor signalling pathway involved in phagocytosis	428	10	73.16	<0.2	1.15E-16
Detection of chemical stimulus involved in sensory perception	474	20	81.02	.25	2.64E-12
Sensory perception of chemical stimulus	534	24	91.28	.26	2.49E-13
Sensory perception	943	106	161.19	.66	1.26E-02
Detection of chemical stimulus	509	24	87.01	.28	2.49E-12
Detection of stimulus	676	59	115.55	.51	2.58E-05
Detection of stimulus involved in sensory perception	525	30	89.29	.33	1.19E-09
Sensory perception of smell	458	10	78.29	<0.2	1.20E-18

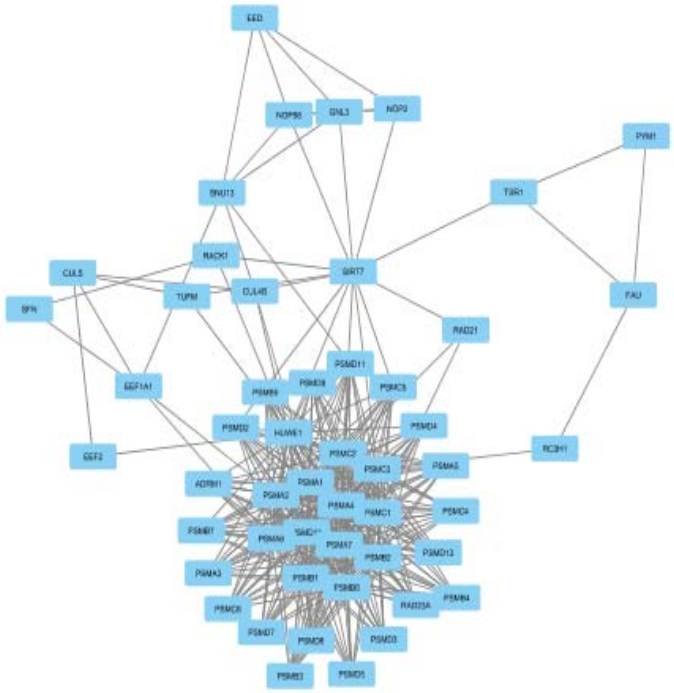
A total of 156,426 PPIs were reported from 14,111 DEGs using g:Profiler web tool and BioGRID PPI database (see supplementary file S03). The protein complex detection of the PIN in Cyroscape returned a total of 64 protein complexes (supplementary file SO4). The visualization of protein complexes were done using MCODE algorithm in Cytoscape. The first six protein complexes visualization are reported in figure 1 below. The density node score, number of nodes and edges for the first six complexes are 51.631-16-1678, 19.704-28-266, 15.04-51-376, 10.435-139-720, 8.696-47-200, 8.027-244-895 respectively. Here, protein complexes that are active in the RBCs that could be opened to invasion have been predicted. The functional enrichment of the identified protein complexes revealed functions related to rRNA processing, Ribosome biogenesis, RNA metabolic process, cellular process, Nucleic and metabolic process and much more which are active in the RBCs that could be open to invasion by *Plasmodium falciparum.*



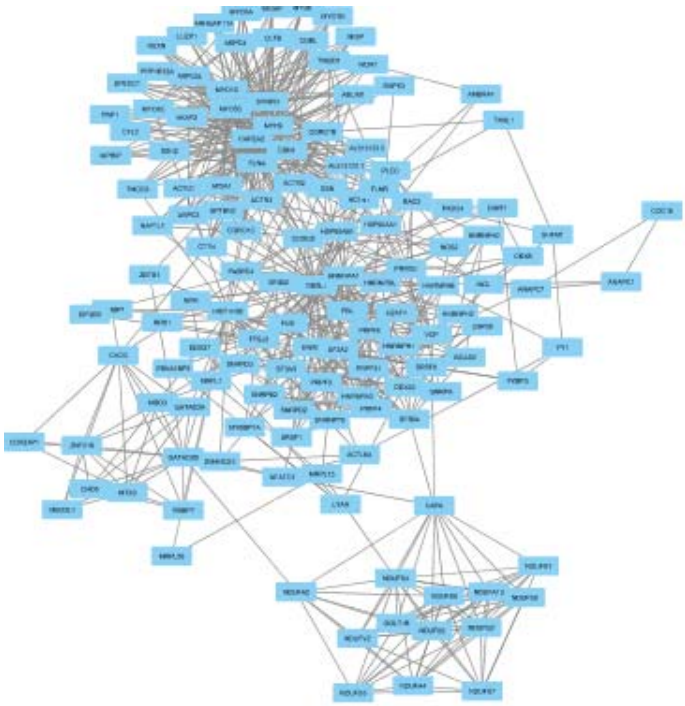
Protein Complex 1



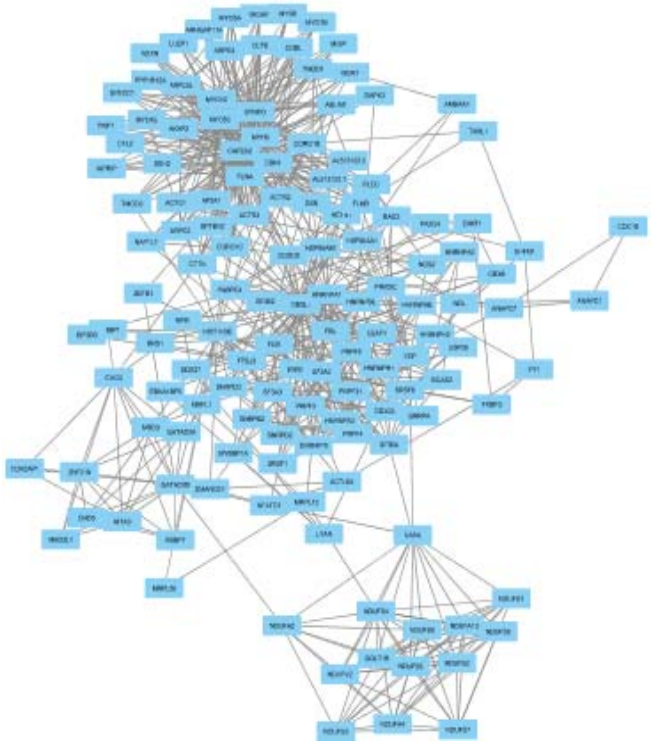
Protein Complex 2



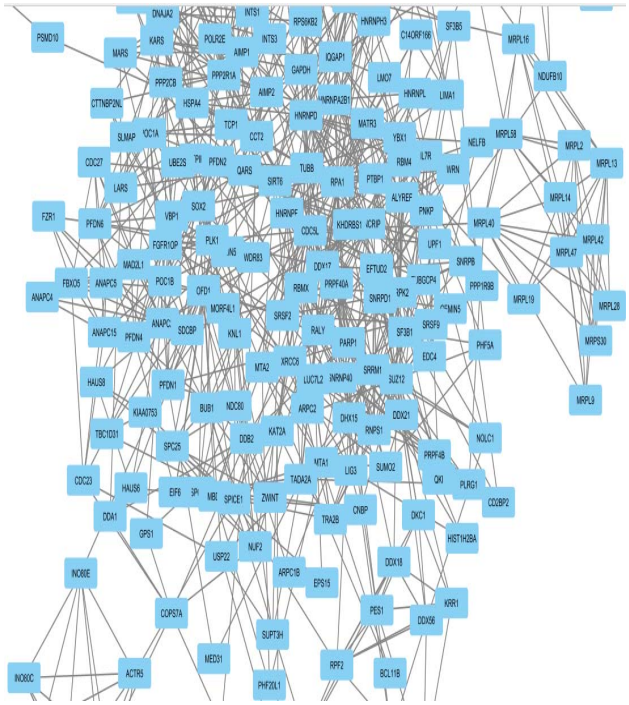
Protein Complex 3



Protein Complex 5



Protein Complex 4



Protein Complex 6

Fig.1. Visualization of the first six protein complexes using MCODE in Cytoscape

TABLE 2 FUNCTIONAL ENRICHMENT OF SOME OF THE PROTEIN COMPLEXES

Clusters	Functions Enriched	Enriched Scores	Count	P_value
Cluster 10 Nodes: 12 Edges: 28 Score: 5.091		5.82		
	Proteasome		8	1.6E-17
	Nucleus		9	2.6E-8
	Threonine protease		4	2.8E-7
	Protease		6	5.2E-7
	Hydrolase		6	5.0E-5
Cluster 11 Nodes: 5 Edges: 5 Score: 4.5		5.11		
	DNA recombination and repair protein Rad51, C-terminal		3	2.2E-5
			3	2.2E-5
	DNA recombination/repair protein RecA/RadB, ATP-binding domain		3	1.7E-4
	DNA-dependent ATPase activity		5	6.8E-4
	ATP binding		4	4.3E-3
	P-loop containing nucleoside triphosphate hydrolase			
Cluster 12 Nodes: 4 Edges: 14 Score: 4		5.11		
	Homologous recombination		4	1.7E-7
	DNA recombination and repair protein Rad51, C-terminal		3	1.2E-6
			3	1.2E-6
	DNA recombination/repair protein RecA/RadB, ATP-binding domain		3	1.6E-5
	DNA-dependent ATPase activity		5	1.2E-4
	ATP binding		4	4.5E-4
	P-loop containing nucleoside triphosphate hydrolase			
Cluster 14 Nodes: 12 Edges: 20 Score: 43.636		0.33		
	extracellular exosome		6	4.9E-2
	Transmembrane helix		3	8.6E-1
	Transmembrane		3	8.6E-1
	integral component of membrane		3	8.6E-1
	Membrane		3	8.6E-1
Cluster 17 Nodes: 4 Edges: 5 Score: 3.33		5.37		
	Basal transcription factors		4	2.0E-7
	Nucleotide excision repair		4	2.3E-7
	Holo TFIIH complex		3	5.1E-7
	RNA polymerase II carboxy-terminal domain kinase activity		3	3.4E-6 4.7E-6
	ATP-dependent DNA helicase activity		3	1.3E-5
	Nucleotide-excision repair		3	5.3E-3
	Positive regulation of transcription from RNA polymerase II promoter		3	
Cluster 20 Nodes: 4 Edges: 5 Score: 3.33		3.58		
	Transmembrane receptor protein		3	8.2E-7
	serine/threonine kinase activity transcription		3	1.4E-6
	from RNA polymerase II promoter		4	5.6E-6
	Receptor signaling protein serine/threonine kinase activity		3	1.4E-3 2.3E-3
	ATP-binding		3	2.5E-3
	Nucleotide-binding		3	6.8E-2
	Receptor			
	Membrane			
Cluster 23 Nodes: 6 Edges: 8 Score: 3.2		2.02		
	Zinc		3	8.1E-3
	Intracellular		3	1.4E-2
	Metal-binding		3	1.5E-2
Cluste r 44 Nodes: 8 Edges: 10 Score: 2.857		0.94		
	Membrane		6	1.4E-2
	Transmembrane helix		4	2.3E-1
	Transmembrane		4	2.3E-1
	Integral component of membrane		4	2.3E-1

IV. CONCLUSION

Computational approach has again proved to be faster and efficient than experimental approach. This study showed the use of computational approach in manipulating and analysing real life biological problem as opposed to time consuming, effort-driven experimental approach. Our study predicted 156,426 protein-protein interactions at the RBCs that might be implicated in the invasion process. Complete knowledge of these protein-protein interactions will provide the needed insight into studying the disease at hand to better understand proteins likely to interact with that of parasite based on functions or structures and such functions have been predicted in the functional enrichment analysis carried out in our study. Further investigation revealed protein complexes that are implication in the invasion process thus creating room of opportunities for better insight into drug development.

ACKNOWLEDGMENT

This research was supported by H3ABioNet via a NHGRI grant number U41HG006941.

REFERENCES

[1] T. Pawson, and P. Nash, "Protein–protein interactions define specificity in signal transduction", Genes and development, vol. 14, pp. 1027-1047, 2006.

[2] J. Espadalar, O. Romero-Isart, R.M. Jackson and B. Oliva, "Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships", Bioinformatics, vol. 21, pp. 3360-3368, 2005.

[3] M. Jiang, Y. Chen and Y. Zhang, "Identification of hepatocellular carcinoma related genes with k-th shortest paths in a protein protein interaction network", Mol BioSyst, vol. 9, pp. 2720-2728, 2013.

[4] H. Ge, A.J. Walhout and M. Vidal, "Integrating 'omic' information: a bridge between genomics and systems biology", Trends Genet, vol. 19, pp. 551–560, 2003.

[5] E. Sackmann, Biological Membranes Architecture and Function. Germany: Elsevier Science, 1995.

[6] F. Pierige, S. Serafini, L. Rossi and M. Magnani, "Cell-based drug delivery", Advanced Drug Delivery Reviews, vol. 60, pp. 286–295, 2008.

[7] A.S. Paul, E.S. Egan and M.T. Duraisingh, " Host–parasite interactions that guide red blood cell invasion by malaria parasites", Curr Opin Hematol, vol. 22 pp. 220–226, 2015.

[8] M.C. James and L.H. Stephen, Malaria. Medical Microbiology 4th Eds. NCBI Resources, 1996.

[9] A.F. Cowman, D. Berry and J. Baum, "The cellular and molecular basis for malaria parasite invasion of the human red blood cell", The Journal of Cell Biology (JCB), vol. 198, pp. 961 – 971, 2012.

[10] L.E. Ziady and N. Small, Prevent and Control Infection: Application Made Easy, Verlag: Juta Academy, 2005.

[11] K.G. Le Roch, D.W. Chung and N. Ponts, "Genomics and integrated systems biology in Plasmodium falciparum: a path to malaria control and eradication", Parasite Immunol, vol. 34 pp. 50–60, 2012.

[12] P.R. Gilson, and B.S. Crabb, " Morphology and kinetics of the three distinct phases of Red blood cell invasion by Plasmodium falciparum merozoites", Int J Parasitol, vol. 39, pp. 91-96, 2009

[13] N.J. White, S. Pukrittayakamee, T.T. Hien, M.A. Faiz, O.A. Mokuolu and A.M. Dondorf, “Malaria”, Lancet, vol. 383, pp. 723-735, 2014.

[14] L Bannister and G. Mitchell, “The ins, outs and roundabouts of malaria”, Trends Parasitol, vol. 19, pp. 209 – 213, 2003

[15] M. Koch and J. Baum, “The mechanics of malaria parasite invasion of the human erythrocyte – towards a reassessment of the host cell contribution”, Cellular Microbiology, vol. 18, pp. 319-329, 2016

[16] J. Yamagishi, A. Natori, M.E. Tolba, A.E. Mongan, C. Sugimoto, T. Katayama and Y. Eshita, “Interactive transcriptome analysis of malaria patients and infecting Plasmodium falciparum”, Genome research, vol. 24, pp. 1433-1444, 2014

[17] PANTHER: http://pantherdb.org/

[18] BioGRID : https://thebiogrid.org/

[19] CYTOSCAPE: http://www.cytoscape.org/

[20] G. Dennis, B.T. Sherman, D.A. Hosack, J. Yang, W. Gao. H.C. Lane and R.A. Lempicki, “DAVID”, Database for Annotation Visualization and Integrated Discovery”, Genome Biol., vol. 4, pp. 60-70, 2003.

[21] G.D. Bader and C.W.V. Hogue, “An automated method for finding molecular complexes in large protein interaction networks ”, BMC Bioinformatics, vol. 4, pp. , 2003.

[22] S. Srihari and H.W. Leong, “A survey of computational methods for protein complex prediction from protein interaction networks”, J. Comp. Biol. Bioinf., vol. 11, pp. 1-27, 2013.