SIMULTANEOUS AND SINGLE GENE EXPRESSION: COMPUTATIONAL ANALYSIS FOR MALARIA TREATMENT DISCOVERY

VICTOR CHUKWUDI OSAMOR
CUO3GP0042

Simultaneous and Single Gene Expression: Computational Analysis for Malaria Treatment Discovery

By

Victor Chukwudi Osamor
CUO3GP0042
Department of Computer and Information Sciences
College of Science and Technology
Covenant University

Being

A Thesis Submitted in Partial Fulfillment
of the Requirement for the Award of
Doctor of Philosophy (Ph.D)
in Computer Science of
Covenant University
Ota, Ogun State
Nigeria

CERTIFICATION

We certify that this work was carried out by Victor C. Osamor in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria.

Dr. Ezekiel F. Adebiyi ----------

(Supervisor) Signature & Date

Dr Seydou Doumbia  ----------

(Co-Supervisor) Signature & Date

Dr. Ezekiel F. Adebiyi ----------

(Head of Department) Signature & Date

Prof. E.A. Onibere  ----------

(External Examiner) Signature & Date

DEDICATION

I dedicate this work to the Almighty God without whose miracles, this work would not have been successful. I also dedicate it to my wife, Mrs Ifeoma Patricia Osamor whose support in innumerable ways aided me to conquer all obstacles in course of this work. In addition, I will also like to dedicate it to my biological father Pa Joseph Okafor Osamor, whose push and eagerness propelled me to finish this work. Unfortunately, 24hrs (19^{th}, April 2009) to the submission of this thesis, news filtered in that he has gone to be with Lord. May his gentle soul rest in perfect peace. Adieu Papa!!!

ACKNOWLEDGEMENT

Firstly, I wish to acknowledge Almighty God for it is not of him that willeth nor of him that runneth, but of God that sheweth mercy **(Romans 9:16)**. God has made me to pursue and conquer even when at a time it looked as a deep cloud in outer darkness, God still guided me by showing me a light at the end of the tunnel. God's word ("I returned, and saw under the sun, that the race *is* not to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favour to men of skill; but time and chance happeneth to them all" (**Ecclesiastes 9:11)** kept me alive and burning even in the race to acquire this Doctor of Philosophy (Ph.D) Degree. I am really indebted to God as I mark and pledge to serve him all days of my life. Thank you Bishop (Dr) David Oyedepo for teaching us the winning keys and God bless Covenant University for making this possible. I also want to thank the Vice Chancellor Prof Aize Obayan, new Registrar, Dr Daniel Rotimi and former Registrar Pastor Yemi Nathaniel for their exemplary leadership that gave birth to this success.

My profound gratitude to my Supervisor and current H.O.D. Dr Ezekiel Adebiyi whose meticulous criticisms, corrections, suggestions and guidance to the entire work gave me impetus to work very hard even in the presence of other academic and administrative loads. Sir, I sincerely wish to specially thank you for your effort and rigorous training during this work as I pray that God will also bless your beloved wife (Mrs Adebiyi) for creating time for you to attend to my work. To the former H.O.D. and current Director of Academic Planning Unit (DAPU) Dr C.K. Ayo, I also appreciate you and pray that God will reward you for recognizing my strength and your proffered advice in course of my Ph.D training. This also goes to Prof. Olushola Ojo for his support and guidance as I pray that the Almighthy God will reward you. I also wish to thank Prof N. Okonjo of Chemistry Department, current Dean of College of Science and Technology Prof. James Katende for their prayers and guidance.

To Mr Oyelade J. Olarenwaju, I will like to thank him very much for all the implementation support he gave me in the course of the work. I use this opportunity to thank all staff of Computer and Information Science Department for their cooperation

during this work. Foreign scientists that are worthy of acknowledgement include my co-supervisor Dr Seydou Doumbia of Malaria Research Training Centre (MRTC), University of Bamako in Mali and Dr Doulaye Dembele, my first MATLAB teacher from Plateforme BIOPUCES de Strasbourg, IGBMC, 67404 ILLKIRCH CEDEX, FRANCE for the initial criticism of my first Ph.D proposal. We are grateful to Johanna P. Daily of Department of Immunology and Infectious Disease, Harvard School of Public Health, 665 Huntington Avenue, Boston, Massachusetts 02115, USA and Fahim, A. M. from Faculty of Education, Suez Canal University, Suez city, Egypt, Chris Ding, NERSC Division, Lawrence Berkeley National Lab., UC Berkeley, USA, for useful and constant discussions. We also thank the DeRisi J. from Department of Biochemistry and Biophysics, University of California, San Francisco, USA, and Karine LeRoch from University of California, Riverside, USA., for making their microarray data available through the web. We also will like to thank the three anonymous reviewers of the BMC Bioinformatics Journal for their very useful assessment which improved the quality of Chapter Three beyond what it was when we made our first submission for publication.

The home could not have been more conducive if my lovely queen Mrs Ifeoma Osamor had not made it so. Indeed her support was incalculable arising from her intuitive and instinctive guidance as she aided in some routine mathematical and computer-related tasks. Honey, I thank you so much for sharing this vision and may the Lord strengthen you further and give you more blessings. I thank my father Joseph Osamor whose dream has been to support me up to Doctorate level but unfortunately we lost him, a day to the submission of this thesis. Papa may your soul rest in perfect peace, Amen. I also appreciate my mother, Mrs Regina Osamor for her blessings, prayers and my success, Ifeanyi Peter Osamor and Tina Ebube (nee Osamor) my younger brother and sister respectively for their keen interest in my Ph.D work. It is also a great privilege to use this opportunity to thank my parents-in-law, Mr and Mrs Okpah and the entire Okpahs' family for their support and understanding. Moreso, thanks to the Ebubes' family, Osamors' family, John Onwuegbuzie, Mr Owoloko e.t.c, whose name have not been mentioned but contributed directly or indirectly to the progress of this work. I thank Mr Baderu and Pastor Abolarin of Faith Academy for assisting in the proof-reading of this thesis.

ABSTRACT

The major aim of this work is to develop an efficient and effective k-means algorithm to cluster malaria microarray data to enable the extraction of a functional relationship of genes for malaria treatment discovery. However, traditional k-means and most k-means variants are still computationally expensive for large datasets such as microarray data, which have large datasets with a large dimension size *d*. Huge data is generated and biologists have the challenge of extracting useful information from volumes of microarray data. Firstly, in this work, we develop a novel k-means algorithm, which is simple but more efficient than the traditional k-means and the recent enhanced k-means. Using our method, the new k-means algorithm is able to save significant computation time at each iteration and thus arrive at an $O(nk^2)$ expected run time. Our new algorithm is based on the recently established relationship between principal component analysis and the k-means clustering. We further prove that our algorithm is correct theoretically. Results obtained from testing the algorithm on three biological data and three non-biological data also indicate that our algorithm is empirically faster than other known k-means algorithms. We assessed the quality of our algorithm clusters against the clusters of known structure using the Hubert-Arabie Adjusted Rand index (ARI_{HA}), we found that when k is close to d, the quality is good ($ARI_{HA} > 0.8$) and when k is not close to d, the quality of our new k-means algorithm is excellent ($ARI_{HA} > 0.9$). We compare three different k-means algorithms including our novel Metric Matrics k-means (MMk-means), results from an *in-vitro* microarray data with the classification from an *in-vivo* microarray data in order 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 is similar, thereby authenticating our MMk-means method and its effectiveness. Lastly using clustering, R programming (with Wilcoxon statistical test on this platform) and the new microarray data of *P. yoelli* at the liver stage and the *P. falciparum* microarray data at the blood stages, we extracted twenty nine (29) viable *P. falciparum* and *P. yoelli* genes that can be used for designing a Polymerase Chain Reaction (PCR) primer experiment for the detection of malaria at the liver stage. Due to the intellectual property right, we are unable to list these genes here.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND INFORMATION OF THE STUDY

Fatal human malaria infection is initiated when an infected Anopheline mosquito – *Anopheles gambiae*, injects sporozoites during a human blood meal. After injection, sporozoites enter the bloodstream and go to the liver, where they invade hepatocytes and develop into exoerythrocytic forms (Coppi et al., 2005). These liver stage parasites mature and are released into the Red Blood Cell (RBC) for erythrocytic stage, a form characterized with symptomatic malaria. Four species of the genus *Plasmodium* are responsible for the human malaria out of which *P. falciparum* stands out as the most lethal compared to *P. vivax, P.malariae* and *P.ovale.*

Drug resistance in evolving *Plasmodium falciparum* strains and insecticide resistance of the female *Anopheles* mosquito account for major biomedical catastrophe standing against all efforts to eradicate malaria in Sub-Saharan Africa. Malaria is endemic to more than 100 countries and by far the most costly in terms of human health causing major losses among many African nations including Nigeria. *Plasmodium* species is a protozoan parasite that infects approximately 500 million people annually, killing more than one million, mainly children and pregnant women in Africa (Le Roch et al., 2003; Breman, 2001). Malaria is a global problem as estimates suggest that 40% of the world's population is at risk of malaria (Brown and Reeder, 2002). In a recent PCR (Polymerase Chain Reaction), malaria diagnostics study conducted on 401 children that complained of fever in Lafia, located within the Guinea savanna ecological zone in north-central Nigeria, Oyedeji et al., (2007) reported that 285 patients out of these 401 were infected with malaria. Within this region, malaria transmission was formerly described as stable and uniformly intense through most of the year (Bruce-Chwatt, 1951; Molineaux and Gramiccia, 1980).

There are three main strategies presently attempting to control malaria disease: vaccination, vector control, and drugs. Of these, drug application is currently the main line

of disease control with some level of mosquito control. Despite initially promising results with multicomponent recombinant protein vaccines targeted against the asexual blood stages (Genton et al., 2003) and vaccines directed against the sporozoite stage (Bojang et al., 2001), effective immunization against the disease is not yet available (Yeh et al., 2004). There is, however, a deepening crisis with emerging resistance among malaria parasites to the existing drugs. For these reasons, it is imperative that new lines of drugs be explored before existing drugs lose too much efficacy (Ralph et al., 2001).

Malaria treatment discovery and antimalarial drug development can follow several strategies, ranging from minor modifications of existing agents to the design of novel agents that act against new targets, as available agents are being combined to improve antimalarial regimens (Rosenthal, 2003). Among important efforts that are currently ongoing are the optimization of therapy with available drugs, including the use of combination therapy, the development of analogs of existing agents, the discovery of natural products, the use of compounds that were originally developed against other diseases, the evaluation of drug resistance reversers, and the consideration of new chemotherapeutic targets. The last category benefits from recent advances in malaria research technologies and genomics and is providing new classes of drugs (Rosenthal, 2003).

The concept of gene expression can simply be understood by considering genes as containing the instructions for making messenger RNA (mRNA); but at any moment, each cell engages itself in the production (expression) of mRNA from only a fraction of the genes it carries. If a gene is used to produce mRNA, it is considered "on", otherwise "off". Many factors determine whether a gene is on or off, these include the time of the day, whether or not the cell is actively dividing, its local environment, and chemical signals from other cells. Skin cells, liver cells and nerve cells turn on (express) somewhat different genes and that is in large part, what makes them different. Therefore, an expression profile allows one to deduce a cell's type, state, environment, and other attributes.

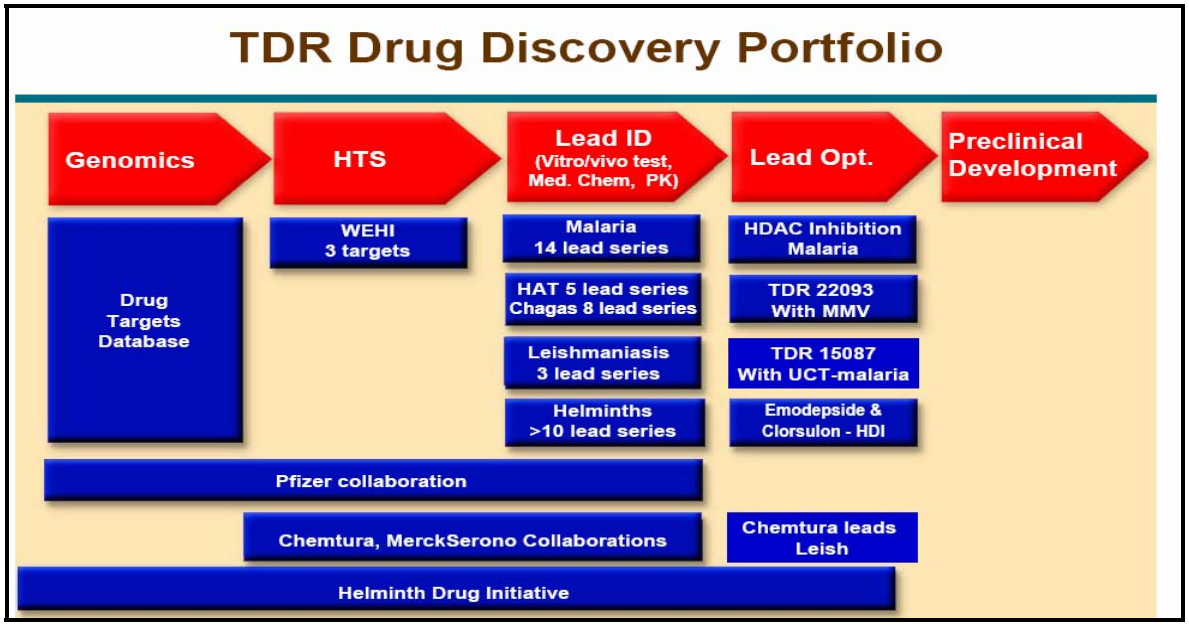
Gene expression profiling experiments often involve measuring the relative amount of mRNA usually called transcript, expressed in two or more experimental conditions. This is because altered levels of a specific sequence of mRNA suggest a changed need for the protein coded for by the mRNA, perhaps indicating a homeostatic response or a pathological condition. Placing expression profiling results in a publicly accessible microarray database, makes it possible for researchers to assess expression patterns beyond the scope of published results, perhaps identifying similarity with their own work.

Gene expression profiling has been commonly used to study the pathogen's or host's responses to each other or to the external stimuli such as drug or vaccine treatments. The fight against malaria is failing and microarray analyses need to keep up the pace to unravel the evolving parasite's gene expression profile, which is a pointer to monitoring the genes involved in malaria's infective metabolic pathway. Gene expression profiles can also be used in studying various state of malaria development in which expression profiles of different disease states at different time points are collected and compared to one another to establish a classifying scheme for purposes like diagnosis and treatments with adequate drugs.

Deoxyribonucleic Acid (DNA) microarray gene expression profiling is a high-throughput measurement of the activity (the expression) of thousands of genes simultaneously (at the same time), to create a global picture of cellular functions. These profiles can, for example, distinguish between cells that are actively dividing, or show how the cells react to a particular treatment. Many experiments of this sort measure an entire genome simultaneously, that is, every gene present in a particular cell. DNA Microarray technology measures the relative activity of previously identified target genes. Tag-based techniques, like serial analysis of gene expression (SAGE, SuperSAGE) are also used for gene expression profiling (Wikipedia, 2008).

To detect or analyse single gene in an environment, we need a PCR-based technology. This is most useful in the quest to detect effectively when a person is infected with the malaria parasite.

Discovery of new malaria treatment benefits from recent advances in malaria research technologies and genomics and is providing new classes of drugs. A number of new antimalarial therapies will likely be needed over the coming years. So, it is important to pursue multiple strategies for drug discovery. The development of resistance in the parasite to effective and inexpensive drugs, the lack of a licensed malaria vaccine, and the fundamental complexity inherent in the malaria parasite means that there is an urgent need to better understand the function of *P. falciparum* genes and their biological role to support the development of new and effective antimalarial strategies (Le Roch et al., 2003). For example, Figure 1 shows a current typical WHO/TDR drug discovery pipeline from genomics (which is the domain of this work) to preclinical development stage.



(Source: Nwaka, 2008)

Figure 1.1: TDR Drug Development Pipeline and Portfolio with Genomics at the Basics.

All contributions of our work are at the genomic level which supports high throughput screening. We therefore engage computational tools such as k-means clustering analysis, principal component analysis, etc to explore the huge data generated from high-throughput experiments involving simultaneous and single gene expression for the purpose of extracting meaningful information that will benefit the malaria treatment and new drug discovery.

1.2 STATEMENT OF THE PROBLEM

Malaria caused by *Plasmodium falciparum* is lethal and responsible for major losses and deaths in Sub-Saharan Africa. On a regular basis, data are generated by researchers. Forming a large part of these huge raw data resources is the Malaria Microarray Data (MMD) arising from the study of gene expression in *Plasmodium*. Researchers are often faced with difficulties or inabilities to:

1) Cluster enormous amount of genomic data at a reasonable time shorter than the runtime of existing algorithms.
2) Analyse and extract useful knowledge from the vast amount of MMD.
3) Find a functional relationship among genes involved in malaria infection to understand the complex biology of the parasite.
4) Enhance a better understanding of malaria disease and provide adequate knowledgebase on how best to apply treatment protocols.

This work is poised at providing viable solutions to the challenges listed above.

1.3 AIM AND OBJECTIVES OF THE STUDY

The aim of this research work is to develop a consortium of analytical tools to cluster genes into their functional roles at improved runtime, with a view to contributing to knowledge on many *P. falciparum* genes.

From the above stated aim, the objectives of the research work are as follows:

1) To improve the run-time of the k-means algorithm for useful clustering of high throughput data at a reasonable time.
2) To improve the quality of experimental results interpretation for biological researchers arising from improved clusters output.
3) To find a functional relationship of genes involved in malaria infection.
4) To enhance a better understanding of the malaria disease and to provide adequate knowledgebase on how best to apply treatment protocols.

1.4 RESEARCH QUESTION

The following are the research questions that will enhance the accomplishment of the proposed research objectives:

1) How can we improve the runtime of our new algorithm to cluster large through-put microarray data for *P.falciparum* genes at a reasonable time?
2) How can the quality of clusters from our new algorithm be assessed to ensure that we have appropriate cluster quality for experimental results interpretation?
3) What is the significance of comparative functional classification of *P. falciparum* genes using k-means clustering?
4) Will our algorithm and clustering result interpretation be able to advance malaria treatment discovery?

1.5 RESEARCH METHODOLOGY

The methodology involves the use of principal component analysis (PCA) to develop a new and novel k-means algorithm for microarray data clustering with Pearson correlation as the distance metric. Our new algorithm is based on the recently established relationship between principal component analysis and the k-means clustering. Using the Ding and He threshold (Ding and He, 2004) and our new theoretical derivation, we are able to determine which of the k clusters are optimally equal to the expected ones; (that is, its members will always remain in the same cluster in subsequent iteration). We shall prove that our algorithm is correct and assessed the quality of our algorithm clusters against the clusters of known structures using the Hubert-Arabie Adjusted Rand index (ARI_{HA}) (Steinley, 2004).

Using C++, we implemented the three variants of k-means algorithms, namely, the Traditional, Overlapped and Enhanced k-means following the design of Fahim et al., (2006). We also implemented a fourth one, our MMk-means algorithm using C++ equipped with a MATLAB gateway code. The C++ program is executed from MATLAB environment that links Borland C++ through the gateway code to exchange data. Borland

C++ accepts the raw microarray data and computes its covariance matrix (r) which is sent to MATLAB for the covariance matrix's eigenvalues computation and returned to C++. Our MMk-means algorithm runs like the traditional k-means algorithm except that it is equipped with a mechanism to determine when a cluster is stable, that is, its membership data points will always remain in the same cluster in each subsequent iteration. The algorithm was developed and tested on a DeLL computer, INTEL® CORE™ DUO CPU T2300 @1.66GHz, 512 RAM, 80GB HDD running on Windows Vista operating system.

In ascertaining the significance of comparative functional classification of *Plasmodium falciparum* genes, we deployed our earlier implemented traditional and MMk-means algorithms to cluster Le Roch et al. (2003) data for k=15. 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.(2003) for k=15 serve as a benchmark to compare the effectiveness of the two algorithms. We employed Relational Database Management System (RDBMS) using Microsoft Access 2003 to map genes (in clusters) of Traditional k-means and MMk-means algorithms to their robust k-means counterpart. This data mining allowed us to compare and contrast traditional k-means and MMk-means from their percentage similarity with Le Roch et al. (2003) clusters.

To further consolidate the validation of our MMk-means algorithm, we carried out comparative analysis of clusters results on Le Roch *et al.* (2003) data as generated by the three (3) algorithms on Daily *et al.* (2007) data. We ran Significant Analysis of Microarray (SAM) (Trusher, et al. 2001) 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. We compared clusters 1-15 from Le Roch *et al.* (2003) data for each of the three k-means algorithms with each cluster of Daily *et al.* (2007) and computed the percentage number of genes common to both. 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.* (2003) on the classification from the *in-vivo* microarray of Daily *et al.* (2007) and compared the distinct physiological states of *P. falciparum* from venous blood of malaria patients for identification of important genes that can advance malaria treatment discovery.

In exploring Polymerase Chain Reaction (PCR)-based dectection of malaria, we analyse the behavior of parasite genes at the liver stage by employing the use of the microarray data of Tarun et al. (2008) and Bozdech et al. (2003a) and their orthologues in PlasmoDB (Kissinger et al., 2002). Our interest is to further analyse the behaviour of these liver stage genes using some knowledge obtained from blood stages of *P. falciparum* 3D7 and HB3 strains from the microarray data of Bozdech et al. (2003a). This idea lends credence to the role of orthologues in functional genomics, as genes in a different species that evolved from a common ancestral gene by speciation retain the same function in the course of evolution (Lewis, 2009).

The Traditional clustering algorithm implemented in Osamor et al., (2009) was deployed and used to cluster Tarun et al., (2008) and Bozdech et al., (2003) microarray data independently. Using guilty by association (GBA) principle, genes in the same cluster are expected to be functionally related and orthologues of Tarun et al. (2008) genes in the same cluster using *P. falciparum* 3D7 and HB3 strains expression are expected to be key genes. The number of cluster input was set at k = 15 to serve as benchmark for effective comparative study with other published result like Le Roch et al. (2003). In addition, the dataset tested seem to have the most stable cluster output at k=15. The resultant output was exported to MS Access relational database management system (RDBMS) for analysis. A significance test was conducted between the two strains (3D7 and HB3) genes using the Wilcoxon's statistics. Based on these statistics and annotation information, the orthologues in *P. yoelli* of the most significant genes for 3D7 and HB3 were recommended as important genes that are likely suitable for PCR-based diagnosis of malaria at the liver stage.

1.6 SIGNIFICANCE OF THE STUDY

Generally, the knowledge of the biology and gene expression pattern of *P. falciparum* will provide an invaluable resource for characterizing the complex roles of individual genes and ultimately the identification of new chemotherapeutic and vaccine candidates (Bozdech et al. 2003b) for antimalaria strategies.

However, the significance of this study includes:

1. Development of a novel algorithm for clustering microarray data for identification of important genes for an enhanced understanding of malaria disease that will help to advance treatment discovery.
2. Obtaining acceptable cluster quality with good effectiveness assessed by standard cluster index.
3. Finding the functional relationship of genes involved in malaria infection.

1.7 CONTRIBUTIONS

Our three main contributions in this work are summarized as follows:

1.7.1 OUR CONTRIBUTIONS: REDUCING THE TIME REQUIREMENT OF K-MEANS ALGORITHM

Since traditional k-means and its variants are still computationally expensive for large datasets such as microarray data, we developed a novel k-means which we shall refer to as MMK-means, which is simple and more efficient than traditional k-means, overlapped and enhanced k-means as designed by Fahim et al (2006). Our new k-means algorithm saves significant computation time at each iteration and thus arrived at an $O(nk^2)$ expected run time.

Mathematically, we also showed that our algorithm is correct. The quality of the clusters generated by our MMk-means were assessed using the Hubert – Arabie Adjusted Rand Index (ARI_{HA}) (Steinley, 2004), against the structure of known clustering result. The results of the exercise show that the quality of MMk-means clusters are desirable.

Note, however, that the new clustering algorithm can be used for other clustering needs as long as an appropriate measure of distance between the centroids and the members is used. This has been demonstrated in the course of this work on three non-biological data.

1.7.2 OUR CONTRIBUTIONS: COMPARATIVE FUNCTIONAL CLASSIFICATION OF *PLASMODIUM FALCIPARUM* GENES USING K-MEANS CLUSTERING

We carried out comparative studies of the clustering analysis of major *P. falciparum* microarray results with the objective of seeing the implication of the different clustering tools applied on the malaria parasite under different microarray experiments. In this work, we demonstrated the biological characteristics of our new algorithm against two other well known k-means clustering algorithms (that has been used on this same biological data) and discovered a new functionality for some set of genes.

By this work, we were able to further validate our new and novel MMk-means algorithm. Results from this study indicate that the resulting distribution of the comparison of the three k-means algorithms' *in-vitro* clusters against the *in-vivo* clusters are similar thereby authenticating our MMk-means method and its effectiveness.

1.7.3 OUR CONTRIBUTIONS: EXPLORING PCR-BASED DETECTION OF MALARIA INFECTION AT LIVER STAGE

In addressing the challenges faced with the identification of useful genes and possible primer information, our *in-silico* prediction in chapter seven points to suggest a new exploratory experimental study for possible PCR-based detection of malaria infection at the liver stage. Using our method, the concept of orthology, R programming, recent microarray data at the liver and blood stages, we predicted twenty nine (29) key genes that will be useful for malaria diagnosis at the liver stage.

1.8 SCOPE AND LIMITATION OF THE STUDY

The work focused on clustering algorithm applied principally to microarray data from *P.falciparum* life cycle. This is to allow for thorough in-depth experimental analysis. Despite the fact that other species of *Plasmodium* and other parasitic organisms of apicomplexan origin can cause disease of importance, this study is limited to human malaria caused by *Plasmodium falciparum.* Different *P. falciparum* data were the only biological data considered among many other *Plasmodium* species because it is the most fatal and considered of much economic importance. Note that we also validated the application of our method to non-biological data.

Microarray data are usually noisy, hence, this study will contend with this limitation by analysing large data sets. Issues relating to Polymerase Chain Reaction (PCR) as regards the development of a diagnostic test on it are explored in chapter 7 of this thesis. The work is limited by the emphasis on *P. falciparum* over other human malaria parasites due to the fatal nature of the *P. falciparum* malaria.

1.9 OUTLINE OF THE THESIS

This write-up is structured in eight chapters and they are as follows: In chapter one, motivation behind the work was elucidated while the background section gave a brief introduction and enumerated the contributions of the work. Also included in this chapter are the statement of the problems, research questions, methodology, aim and objectives, and scope and limitation of work.

The literature review spanned through chapters two, three and four for the purpose of clarity in addressing specific areas of the work. Chapter two highlighted the malaria challenge and existing solutions including global initiatives and local efforts in solving the malaria problem. Chapter three gave an account of the meaning and the technology behind DNA microarray and PCR, taking into consideration the various DNA platforms applications, in respect to malaria treatment discovery and the drawbacks of PCR. In

Chapter four, we discussed the clustering techniques, existing clustering methods, applications and drawbacks of k-means clustering.

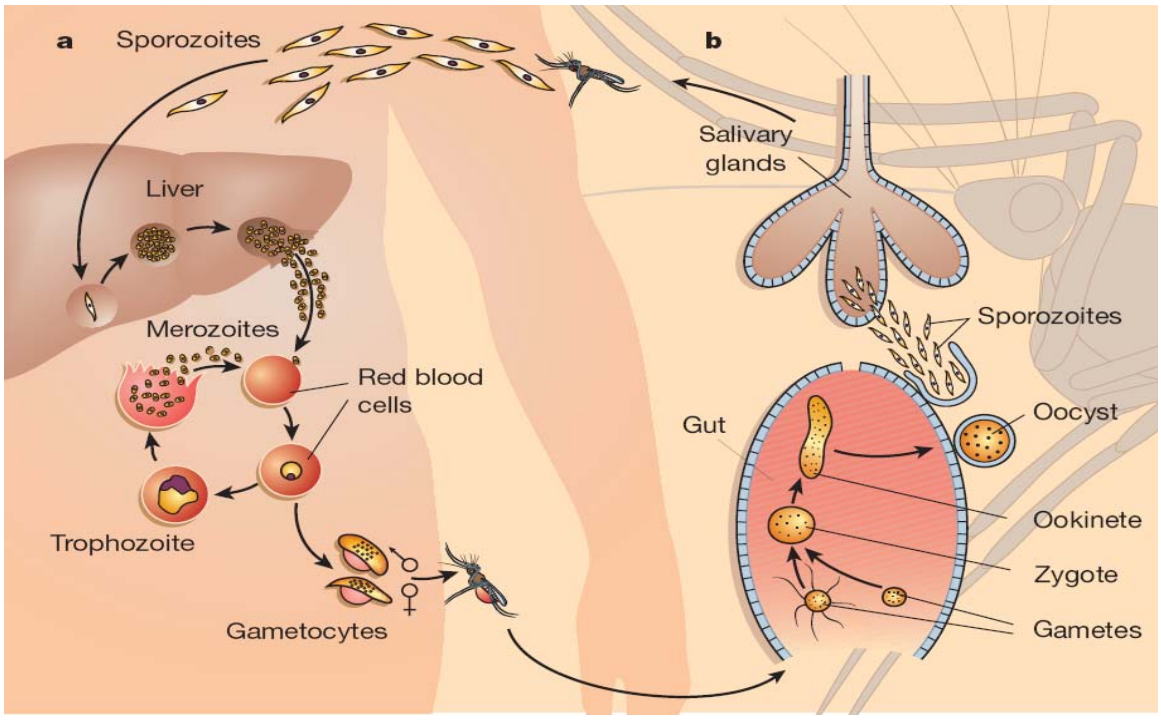
We dedicated each of the next three chapters (five, six and seven) to specific objectives of the work. Chapter five described the ways to solving the problem, the main objective of this work. In its background section, we introduced "Reducing the Time Requirement of k-means Algorithm" and presented previous variants of k-means algorithm. Methods section presents the design of our new algorithm, the Metric Matrices k-means (MMk-means) as well as algorithms correctness and complexity analysis. In result and discussion section, we presented our experimental experience and concluded the chapter. Chapter six reported a set of computational experiments on "Comparative Functional Classification of *Plasmodium falciparum* Genes Using k-means Clustering" to discover the functional role of new sets of genes. Chapter seven discussed the *in-silico* prediction that points to suggest a new exploratory experimental study for possible Polymerase Chain Reaction (PCR)-based detection of malaria infection at the liver stage. In chapter eight, we summarized and concluded the whole work.

CHAPTER TWO

LITERATURE REVIEW 1: THE MALARIA CHALLENGE AND EXISTING SOLUTIONS

2.1 MALARIA TRANSMISSION AND PATHOGENICITY

Plasmodium undergoes developmental life cycle in man and mosquito as depicted in Figure 2.1 (a & b). When a parasite-infected mosquito feeds on a human, it injects the parasites form called sporozoites from its salivary gland into the subcutaneous layer of the skin and into the bloodstream. These migrate to the liver cell forming a quiet liver stage parasite in the parasitophorous vacuole. The co-receptor on sporozoites that mediates invasion involves, in part, the thrombospondin domains on the circumsporozoite protein (CSP) and on thrombospondin-related adhesive protein (TRAP). These domains bind specifically to heparin on hepatocytes. Inside the hepatocyte, each sporozoite develops into tens of thousands of merozoites (Miller, et al. 2002), which can each invade the red blood cells (RBC) on release from the liver.



(Source: Wirth, 2002)

Fig 2.1: Life cycle of the parasite *Plasmodium falciparum* in Man (a) and Mosquito (b).

Furthermore, the blood stage parasites grow and multiply severally, invading many more RBC and releasing the metabolic products arising from RBC degradation and leading to malaria symptoms. To commence sexual stage development, some merozoites undergo several developmental stages namely ring and trophozoite stages and finally differentiate into gametocytes which are picked up by blood-sucking mosquitoes during a bite on an infected person. Eventually, up to 10% of all red blood cells becomes infected and patients' may begin the manifestation of clinical features of malaria, including fever and chills, anaemia and cerebral malaria which can lead to death in case of *Plasmodium falciparum* from female anopheline mosquitoes.

On each mosquito bite of an infected human, it takes up blood containing gametocytes, which develops into male and female reproductive cells (gametes) in the mosquitoes gut, and fusion occurs to form a zygote. The zygote in turn develops into the ookinete, which crosses the wall of the gut and forms a sporozoite-filled oocyst. When the oocyst bursts, the sporozoites move to the mosquito's salivary glands, and the process begins again (Wirth, 2002).

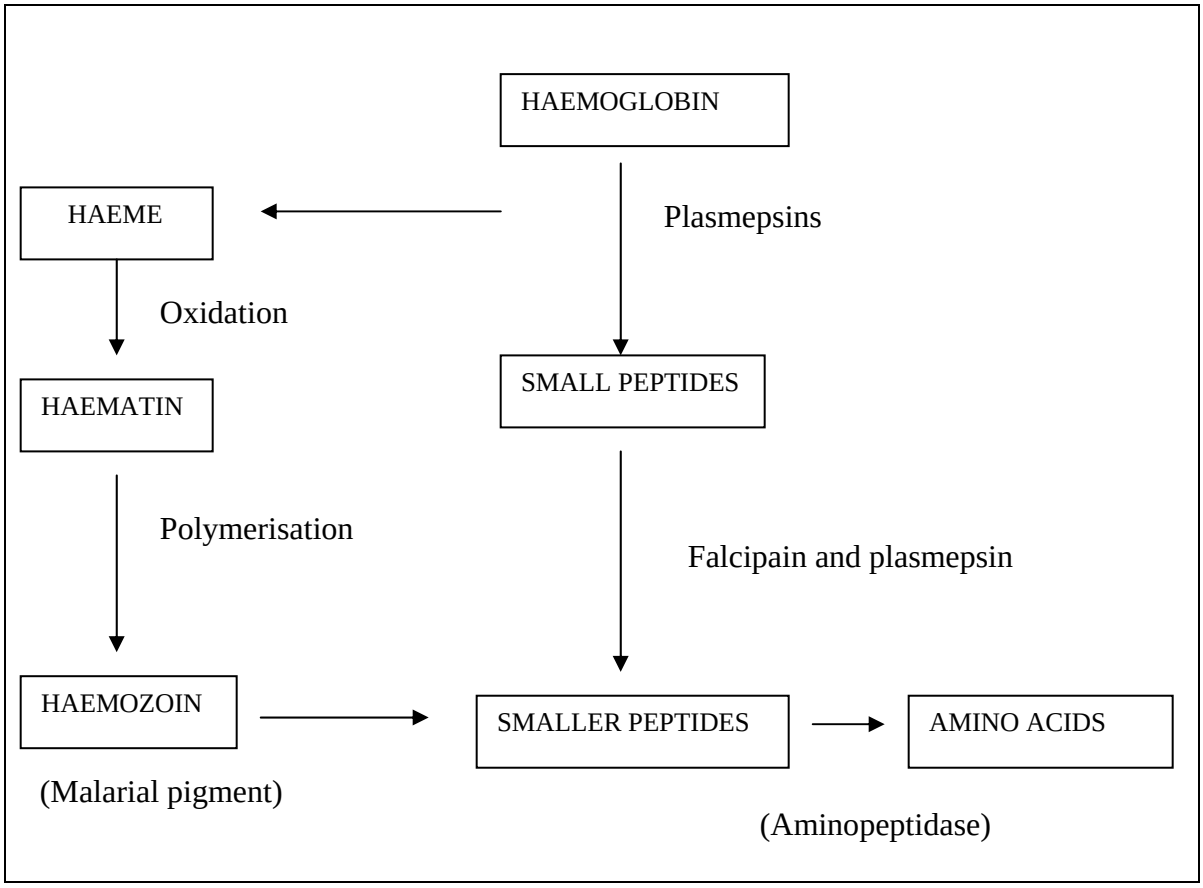
2.1.1 RED BLOOD CELL (RBC) INVASION

In respect to RBC invasion, Miller et al. (2002) noted that what remains completely unknown is which merozoite surface molecules recognize the RBC surface and then signal the start of the invasion process. The parasite induces a vacuole derived from the RBC's plasma membrane and enters the vacuole. Three organelles on the invasive (apical) end of the parasite (rhoptries, micronemes and dense granules) define the phylum Apicomplexa. Receptors that mediate invasion of RBCs by merozoites and invasion of liver by sporozoites are found in micronemes, on the cell surface, and in rhoptries. Identifying the signalling pathways that release organelle contents on contact with a host RBC is a critical issue in parasite biology. Invasion events include releasing essential molecules from apical organelles and initiating the actin–myosin moving junction that brings the parasite inside the vacuole that forms in the RBC.

Although other parasite proteins on the merozoite surface and in apical organelles have been proposed as receptors, there is no direct evidence so far. Because invasion is such a complex series of events from RBC binding, to apical reorientation, to entry, it seems likely that several proteins are required for efficient invasion. For example, evidence has suggested that RBC invasion requires the cleavage of a surface protein on the RBC by an unknown parasite serine protease. Thus, the molecular and cellular events surrounding each step in invasion still remain to be elucidated. Understanding these pathways will give insight into parasite virulence and will facilitate rational vaccine design against merozoite invasion. A single parasite protein, *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), which is expressed at the infected erythrocyte surface connects parasite binding to all the various receptors. PfEMP1 is encoded by the large and diverse *var* gene family that is involved in clonal antigenic variation and has a central role in P. falciparum pathogenesis. Adherence protects the parasite from destruction, as non-adherent mature parasitized RBCs are cleared rapidly in the spleen.

2.1.2 RBC (HAEMOGLOBIN) DEGRADATION

The ability of the blood protein haemoglobin to carry oxygen depends on an iron-containing haeme group, which is made separately in the cell and then binds tightly to a crevice on the globulin protein surface in red blood cells (RBC or erythrocytes). Haemoglobin is degraded by a series of proteases in the digestive food vacuole. The sequential process is represented in Figure 2.2.



(Adapted from http://wisdom.eu-egee.fr/malaria/plasmepsins.pdf)

Figure 2.2: Plasmepsins from *Plasmodium falciparum* Degrades Haemoglobin in RBC.

Plasmepsin I and II attack the haemoglobin breaking it down to haeme and small pepetides. The haeme is converted to haematin by oxidation and further polymerises to haemozoin which causes the high fever. Eventually, smaller peptides result and metabolise into amino acid.

Two homologous plasmepsins I and II are responsible for the initial attack on the hemoglobin Alpha chain between the residues Phe 33 and Leu 34, in the hinge region. This region is highly conserved and responsible for the stability of the haemoglobin tetramer. Upon cleavage, haeme (ferrous +2) is released which is toxic to the parasite and is further oxidized to haematin (ferric +3), also toxic to the parasite. Finally, the haematin is polymerized to haemozoin, the malarial pigment. Both plasmepsin I and II are capable of causing an initial cleavage in the hemoglobin, and the plasmepsins are also capable of several other cleavages after the initial attack.

2.2 SOCIAL AND ECONOMIC IMPACT OF MALARIA

The statistics of Bathurst (2008) elucidated the social and economic impact of malaria and it is highlighted below:

•Afflicts more then 1/3 of the human population;

•Responsible for over 1 million deaths per year of especially children under 5;

•Malaria is curable: 90% of deaths caused by malaria are preventable;

•Annual lost GDP for Africa: $15 billion;

•Costs up to 40% of total public health expenditure;

•Is the cause of up to 50% of in-patient and out-patient care; and

•Costs up to 60% of total household expenditure

This alarming statistics on the threat of malaria have spun global interest to set up initiatives with spelt out responsibilities and goals to combat the malaria pandemic.

2.3 GLOBAL INITIATIVES ON MALARIA PROBLEM

In recent years, several significant objectives and initiatives relevant to the global malaria problem have affirmed the challenge of malaria. The following specific objectives, initiatives, and resolutions (Pan American Health Organisation, 2006) form the basis for the development of and the setting of priorities under the Malaria Plan in most countries:

• The United Nations Millennium Development Goals (MDG) (September 2000)

The malaria plan of MDG is to halt and begin to reverse the incidence of malaria (and other major diseases) by 2015. MDG refers to eight goals that respond to the world's main development challenges to be achieved by 2015. The eight MDGs break down into 21 quantifiable targets that are measured by 60 indicators. These goals were drawn from Millennium Declaration that was adopted by 189 nations- and signed by 147 heads of state and governments during the UN Millennium Summit in September 2000. Listed in definite order, these eight goals (http://www.undp.org/mdg/basics.shtml, 2009) are:

Goal 1: Eradicate extreme poverty and hunger

Goal 2: Achieve universal primary education

Goal 3: Promote gender equality and empower women

Goal 4: Reduce child mortality

Goal 5: Improve maternal health

Goal 6: Combat HIV/AIDS, malaria and other diseases

Goal 7: Ensure environmental sustainability

Goal 8: Develop a Global Partnership for Development

- Medicine for Malaria Venture (MMV) (1999)

  A Swiss non-profit initiative that operates as a public-private partnership for R&D and production of efficacious malaria drugs.

- The Roll Back Malaria (RBM) Initiative (October 1998)

  Halve the malaria burden in participating countries through interventions that are adapted to local needs and reinforcement of the health sector by 2010.

- The Global Malaria Control Strategy (GMCS) (October 1992)
  - Provide early diagnosis and prompt treatment;
  - Plan and implement selective and sustainable preventive measures, including vector control;
  - Detect early, contain or prevent epidemics;
  - Strengthen local capacities in basic and applied research to permit and promote the regular assessment of a country's malaria situation, in particular the ecological, social, and economic determinants of the disease.

2.4 EXISTING DRUGS AND VACCINES FOR MALARIA TREATMENT

2.4.1 SOME AVAILABLE DRUGS

Drugs are chemicals or other substances that alter the function of an organism and are referred as medicines or therapeutic drugs when used for the prevention, treatment and alleviation of diseases as opposed to other hard drugs, such as opiates, which are used illegally. Drugs can be derived from plant, mineral, animal, or synthetic sources. Many early folk medicines, including aspirin, opium, and quinine were derived from plants. Minerals used as medicines include boric acid, Epsom salts, and iodine. Many hormones used to treat a bodily malfunction include insulin for diabetes, or growth hormone to promote proper human development. *Table 1* shows the list of some available malaria drugs as they evolve with time and fail due to resistance, non-compliance, safety and formulation issues (Nwaka, 2008).

Table 2.1: Some Available Malaria Drugs Showing Evolution with Time.

Drug	**Reg.(Yr)**	**Organisations**
Mefloquine	1984	Hoffman La Roche, WRAIR
Halofantrine	1988	GSK, WRAIR
Artemether	1997	Malariaone Poulenc Rorer, Kunmig /TDR
Artemether-lumefantrine	1999	Novartis
Atovaquone+proguanil	2000	GSK
Artemotil (beta-arteether)	2000	Artecef, WRAIR / TDR
Chlorproguanil-dapsone	2003	GSK / TDR
Artesunate-Amodiaquine	2007	Sanofi-Aventis/DNDi

(Source: Nwaka, 2008)

The table also shows respective organizations involved in various antimalarial drug development.

Natural products are the sources of the two most important drugs currently available to treat severe *P. falciparum* malaria, quinine and derivatives of artemisinin. In the case of

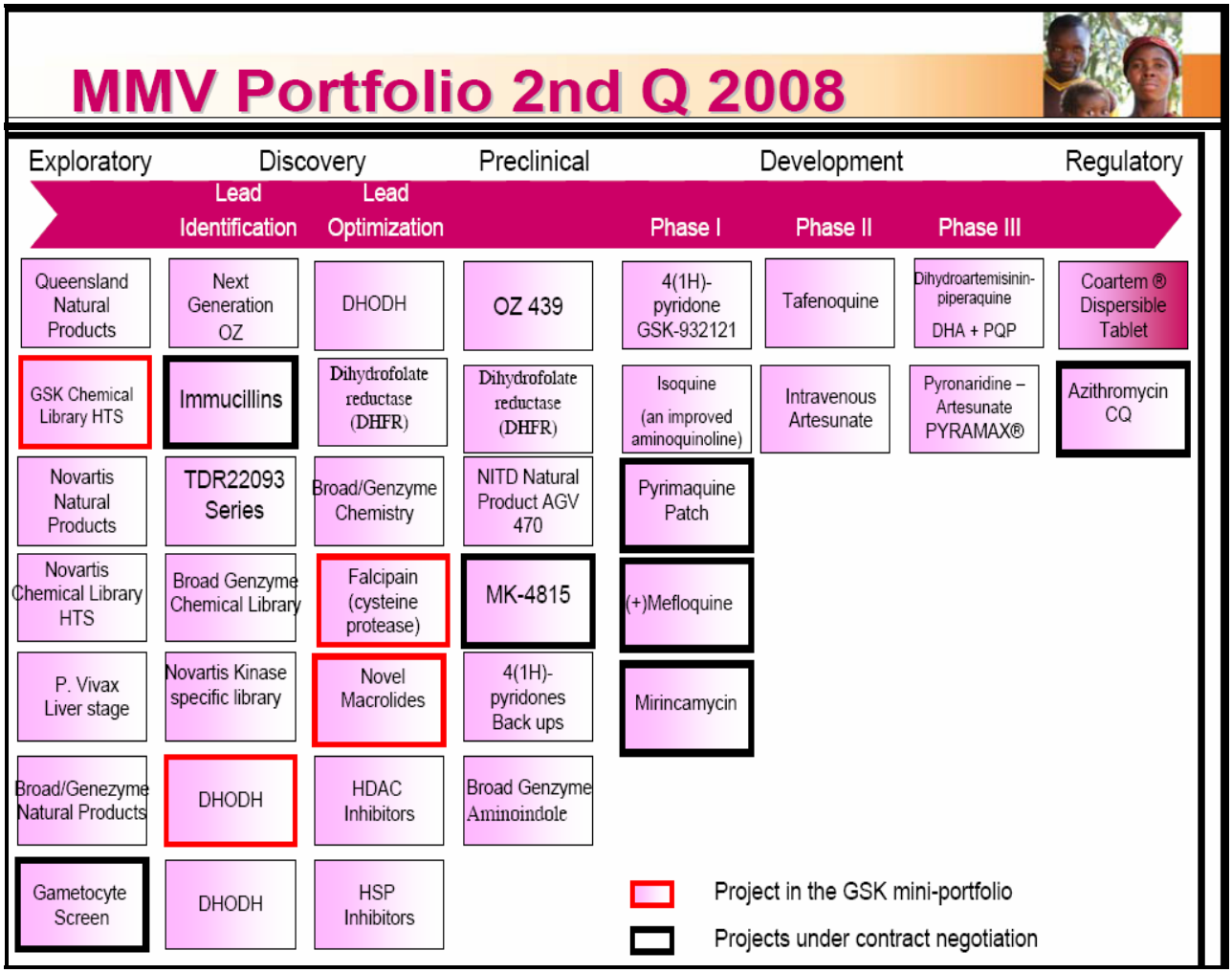
artemisinin, relatively simple chemical modifications of the natural product parent compound have led to a series of highly potent antimalarials that are playing an increasingly important role in the treatment of malaria (Meshnick, 2001). However, the cost of these compounds may be limiting, and so efforts to design fully synthetic endoperoxides that are less expensive to produce are an important priority (Posner et al., 2003; Vennerstrom et al., 2000).

Artesunate has been studied in combination with both sulfadoxine/pyrimethamine (von Seidlein et al., 2000) and amodiaquine (Adjuik et al., 2002) in Africa, with good efficacy. Rosenthal (2003) stated that artemisinin analogs, in particular artesunate and artemether, have recently shown great promise as rapidly acting as potent antimalarials, but the short half-lives of these compounds lead to many late recrudescences after therapy, as seen with artesunate/sulfadoxine/pyrimethamine in Uganda (Dorsey et al., 2002) suggesting that combination therapies are necessary to fully exploit the potency of this class.

Ideally, a combination regimen that prevents resistance development should include at least two agents against which parasite resistance has not yet developed and which have similar pharmacokinetics, so that low blood levels of a single agent will not be present. No such ideal regimen is currently available, although chlorproguanil/dapsone/artesunate may prove to fit this description. Alternatively, the combination of a short-acting, highly potent compound and a longer-acting agent may prove effective, if the initial decrease in parasite burden is so great as to limit subsequent resistance development to the long-acting agent (e.g. artesunate/mefloquine). As another alternative, two drugs with similar pharmacokinetics may prove effective even if resistance to each agent is present in the community (e.g. amodiaquine/sulfadoxine/pyrimethamine). Relatively slow-acting antimalarials (e.g. antibiotics) in combinations like quinine and doxycycline may be effective (Rosenthal, 2003).

Initiatives like Medicine for Malaria Venture (MMV) had projects with drugs at various stages of development as at second quarter of 2008 (shown in Figure 2.3). From figure 2.3, we have the cherry taste and powdery form of Coartem dispersible (from Norvatis), newly formulated for children now at regulatory stage and waiting to be recognized for

usage sometimes in late 2009. Also Azithromycin CQ is formulated to be safer antimalarial at pregnancy. They all pass from exploratory to regulatory stages drug development pipeline.



(Source: Bathurst, 2008)

Figure 2.3: Medicine for Malaria Venture (MMV) Initiative Drug Discovery Portfolio.

2.4.2 MALARIA DRUG RESISTANCE ISSUES

Complicating the process of developing new drugs and treatment strategies for malaria is the problem of drug resistance issues. This is worse particularly regarding the issue of resistance to the most affordable drugs such as chloroquine and Fansidar® (a combination drug of pyrimethamine and sulfadoxine are now widely spread). Some progress has been

made in studying the mechanisms of drug action and drug resistance in malaria parasites, particularly in *Plasmodium falciparum.* These efforts are highlighted by the demonstration of mutations in the parasite's dihydrofolate reductase (DHFR) and dihydropteroate synthase (DHPS) genes conferring resistance on pyrimethamine and sulfadoxine respectively, and by the discovery of mutations in the gene coding for a putative transporter, PfCRT, conferring resistance on chloroquine. Mutations in a homologue of a human multiple-drug-resistant gene, PfMDR1, have also been shown to be associated with responses to multiple drugs (Hayton and Su, 2004). However, except in the case of resistance to antifolate drugs, the mechanisms of action and resistance to most drugs currently in use are essentially unknown or are being debated. But it is believed that there are many more novel ways the parasite uses to engender resistance to drugs.

2.4.3 VACCINES

Zakeri et al., (2007) stated that most experimental pre-erythrocytic stage vaccines are based on or include the circumsporozoite protein (CSP) as an immunogen (any substance or organism that provokes an immune response (produces immunity) when introduced into the body). CSP is the dominant surface protein of the sporozoite and it is used for formulations targeting the pre-erythrocytic stages (the sporozoite and the liver stage parasite). The gene coding for CSP was the first *Plasmodium* gene to be isolated and characterized (Dame et al., 1984; Ellis et al., 1983) and the first *P. falciparum* subunit vaccine tested in human volunteers was based on this protein (Herrington et al, 1987). Today, the most advanced vaccine against malaria, RTS, S, is based on the *P. falciparum* CSP (PfCSP) (Gordon et al., 1995). This vaccine has already undergone two Phase IIb clinical trials in adults and children from The Gambia and Mozambique, respectively (Alonso, 2004; Bojang et al., 2001; Kester et al., 2001) where it provided modest levels of protection. Other examples include, multistage DNA vaccine combination (MuStDO), apical membrane antigen 1 (AMA 1), TRAP/SSP2, synthetic peptide vaccine (SPf66), etc. Generally, protein-based vaccines, DNA-based vaccines, naturally acquired immunity

(NAI) and immunisation with irradiated sporozoite confers partial immunity in rodent and humans (Richie and Saul, 2002).

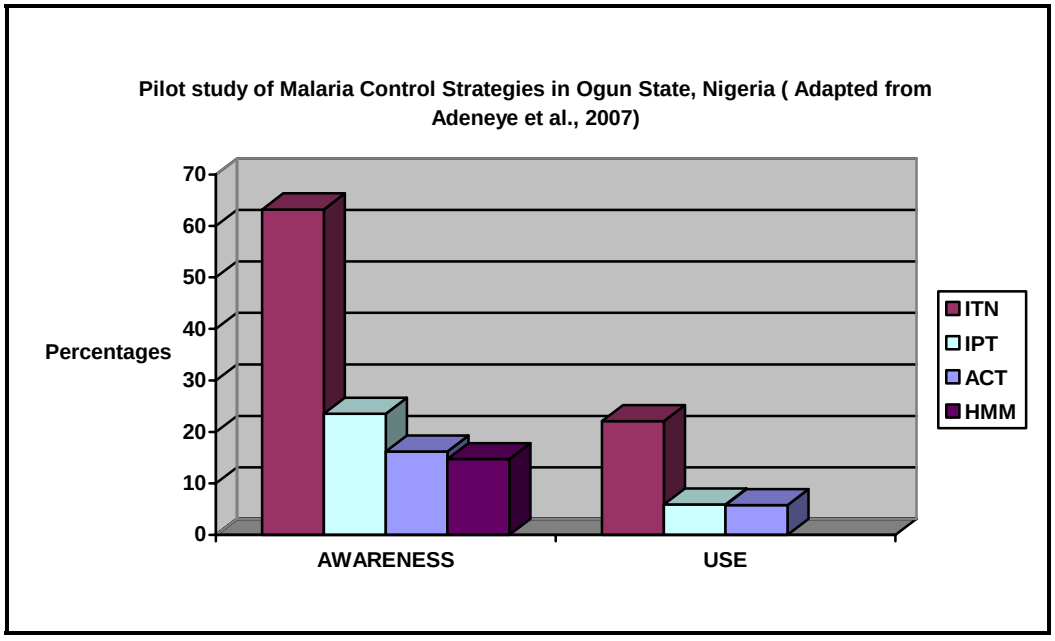
2.5 LOCAL EFFORT TOWARDS MALARIA CONTROL STRATEGIES IN NIGERIA

In Nigeria, malaria control activities are planned and implemented through the Primary Health Care (PHC) system (Federal Ministry of Health, 2005). However, the use of health centres, as the first resort for malaria management has been shown to be low in many African studies including Nigeria. Mothers' malaria treatment-seeking behaviour in rural south-western Nigeria revealed that more than 80% of malaria episodes received treatment outside of the existing government healthcare system (Olaogun et al., 2005; Ajayi and Falade, 2006). The option of malaria treatment at PHC is delayed till the advent of complication and near death. This was attributed to difficulty with access to health centre, scarcity of affordable drugs including antimalarial drugs, perceived deficiencies in the performance of formal health services including poor clinical skills, attitude of health personnel and cultural beliefs (World Health Organisation / United Nations International Children's Emergency Fund, 2003; Feyisetan, et al., 1997). This practice increases morbidity and mortality in addition to contributing to possible emergence of drug resistance (World Health Organisation / United Nations International Children's Emergency Fund, 2003; Okonkwo et al., 2001; Ajayi et al., 2008)

.

World Health Organisation (WHO) initiated the Roll Back Malaria (RBM) Programme in 1998 to halve malaria death world-wide by 2010 (Nabarro and Tayler, 1998) with interventions such as home management of malaria (HMM) (early and appropriate treatment of malaria especially for children less than five years old); intermittent preventive treatment (IPT) of malaria for pregnant women; insecticide treated nets and artemisinin-based combination therapy (ACT) replacing chloroquine and sulfadoxine-pyrimethamine that exhibit parasite resistance (World Health Organisation 2001.; Attaran et al., 2004).

With the approach of 2010 deadline for Roll Back Malaria (RBM), a pilot study was conducted by Adeneye et al. (2007) to assess the awareness, accessibility and use of malaria control strategies among at-risk groups within the context of RBM in Nigeria holo-endemic community, Ijebu-Igbo, in Ogun State, Nigeria. Their results showed that 14.7% and 16.2% of all classes of respondents interviewed were aware of the home management of malaria (HMM) and new antimalarial drug policy programme. Also 63.5% knew about insecticide treated nets (ITNs), while only 22.1% was using the treated material. Only 5.8% of mothers of children less than five years old and none of the pregnant women had taken the new combination drug (ACT). Eight (23.5%) of the 34 pregnant women interviewed knew about intermittent preventive treatment of malaria for pregnant women (IPT). The results of this pilot study showed that efforts need to be intensified to make adequate information and materials relating to the different malaria control strategies more available and accessible at the community level to achieve and sustain the RBM goals, both in Ogun State and in Nigeria in general (Adeneye et al., 2007).



(Adapted from Adeneye et al., 2007)

Figure 2.4: Malaria Control Strategies Awareness Level in parts of Ogun State, Nigeria

The Nigerian Government after 2000 Abuja Declaration, where African Governments agreed to support the RBM strategy of at least 60% at risk-population sleep under ITN, has been promoting RBM interventions through the NetMark initiative (a United States Agency for International Development-funded public-private partnership (United Nations Internation Children's Emergency Fund/ Federal Ministry of Health, 2002). Nigeria's national policy on malaria treatment in 2004 dropped chloroquine and adopted the combination therapy of artemether and lumefantrine (Coartem), artesunate and amodiaquine (Adeneye et al., 2007).

Ajayi et al. (2008) confirms earlier reports that majority of treatment for malaria take place in the home with drugs bought from drug vendors and proposed the evaluation of intervention (health education plus treatment guideline developed using participatory approach) in the use of artemisinin based combination drugs such as artemether-lumefantrine which is now the drug of choice in the treatment of acute uncomplicated malaria in Nigeria.

A part WHO/TDR initiative, African Network for Drugs and Diagnostics Innovation (ANDI) was launched in Abuja in 2008 (http://meeting.tropika.net/andi/, 2009) to promote and sustain African-led R&D innovation through the discovery, development and delivery of affordable new tools including those based on traditional medicines. Malaria treatment discovery is one of the major challenges before this initiative. Osamor and Adebiyi (2007), emphasized the seriousness of institutions and governments in Africa towards eradicating malaria through the use of genomics techniques and microarray technology by citing the interest of The New Partnership for Africa's Development (NEPAD) and WHO/TDR efforts in capacity building in the continent. Examples include the setting-up of ACGT Microarray and African Biosciences facilities in South Africa; series of conferences and workshops organized by African Society of Bioinformatics and Computational Biology (ASBCB); International Workshop on Pattern Discovery in Biology (IWPDB) at Covenant University, Nigeria; and WHO/TDR sponsored functional genomics workshops in Mali.

CHAPTER THREE

LITERATURE REVIEW 2: DNA MICROARRAY AND PCR TECHNOLOGIES

3.1 DNA MICROARRAY TECHNOLOGY

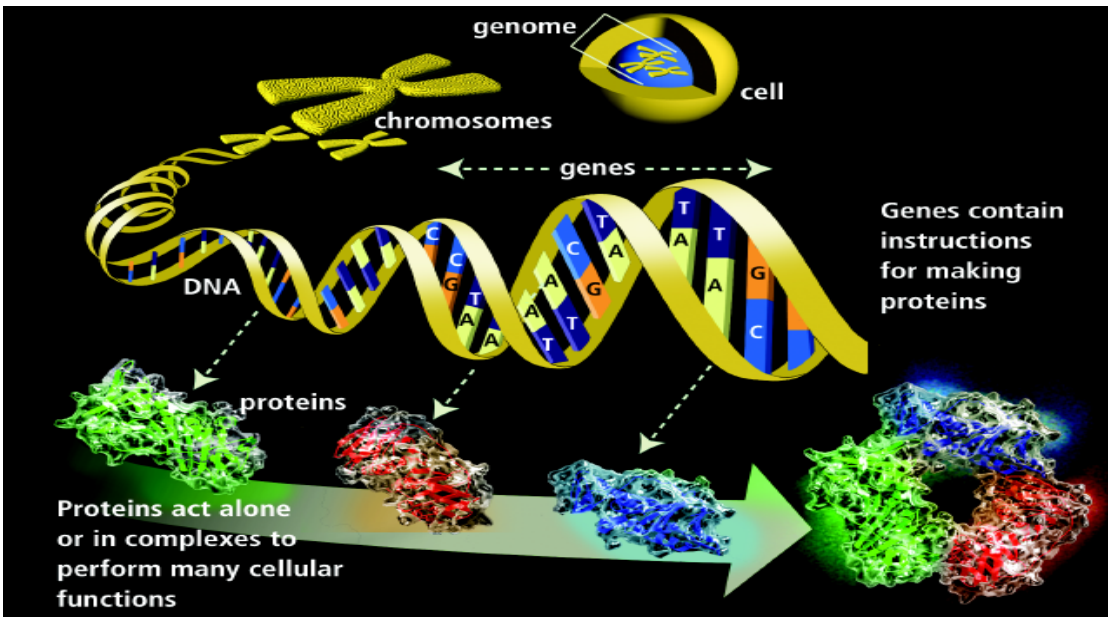
A DNA microarray (also commonly known as *gene or genome chip*, *DNA chip*, or *gene array*) is defined by Wikipedia encyclopedia as a collection of microscopic DNA spots, commonly representing genes, arrayed on a solid surface by covalent attachment to chemically suitable matrices. In the paper of Chen (2006), DNA microarray is defined as an array of tens of thousands of molecular sequences (i.e., probes) mobilized in the form of DNA on a solid and planar platform on a microscopic and high-density scale and can be used in hybridization experiments to parallel measure the quantity of bound homologous sequences in biological samples.

However, we define DNA Microarray as a technology with a grid of nucleic acid molecules of a known composition called probe, placed/immobilized on a solid substrate or slide and used to hybridize messenger RNA (mRNA) from a target cell or tissue of unknown composition to reveal changes in gene expression relative to a control sample. By hybridization, we mean that the four nitrogenous bases of the probe pair up with their complimentary nitrogenous bases in the unknown or tested sample such that Adenine (A) pair Thymine (T)/Uracil (U) and Cytosine (C) pair Guanine (G). Microarray technology, which is also known as "DNA chip" technology, allows the expression behaviour of many thousands of genes to be assessed in a single experiment.

The use of microarrays for gene expression profiling was first published in Schena *et al.* (1995) and the first complete eukaryotic genome (*Saccharomyces cerevisiae*) on a microarray was published in 1997 (Tatusov *et al*., 1997). Microarray experiment involves monitoring gene expression as the cell undergoes some biological processes. These experiments are often used to measure gene expression and therefore are able to detect differences in gene expression between two populations of cells; a test population (disease cell or tissue) versus a control population (normal cell or tissue). However, the experimental and control gene expression values ratio is computed and used. Huge data is

generated and the biologist has the challenge of extracting useful information from volumes of microarray data. Expression levels for tens of thousands of genes can be simultaneously measured in a single hybridization experiment and are collectively called a "gene expression profile".

Each particular cell or tissue in the body has a nucleus bearing a number of chromosomes with genes containing information in its DNA, about the type of needed proteins to be produced by the cell. The characteristic of producing different sets of proteins passes the process called transcription by copying the DNA genetic information of the needed protein to form mRNA (an intermediate product) and finally to protein biomolecules through translation as shown in Figure 3.1.



(Source: Wosik, 2006)

Figure 3.1: Process of Protein Synthesis (DNA=>mRNA=>Protein).

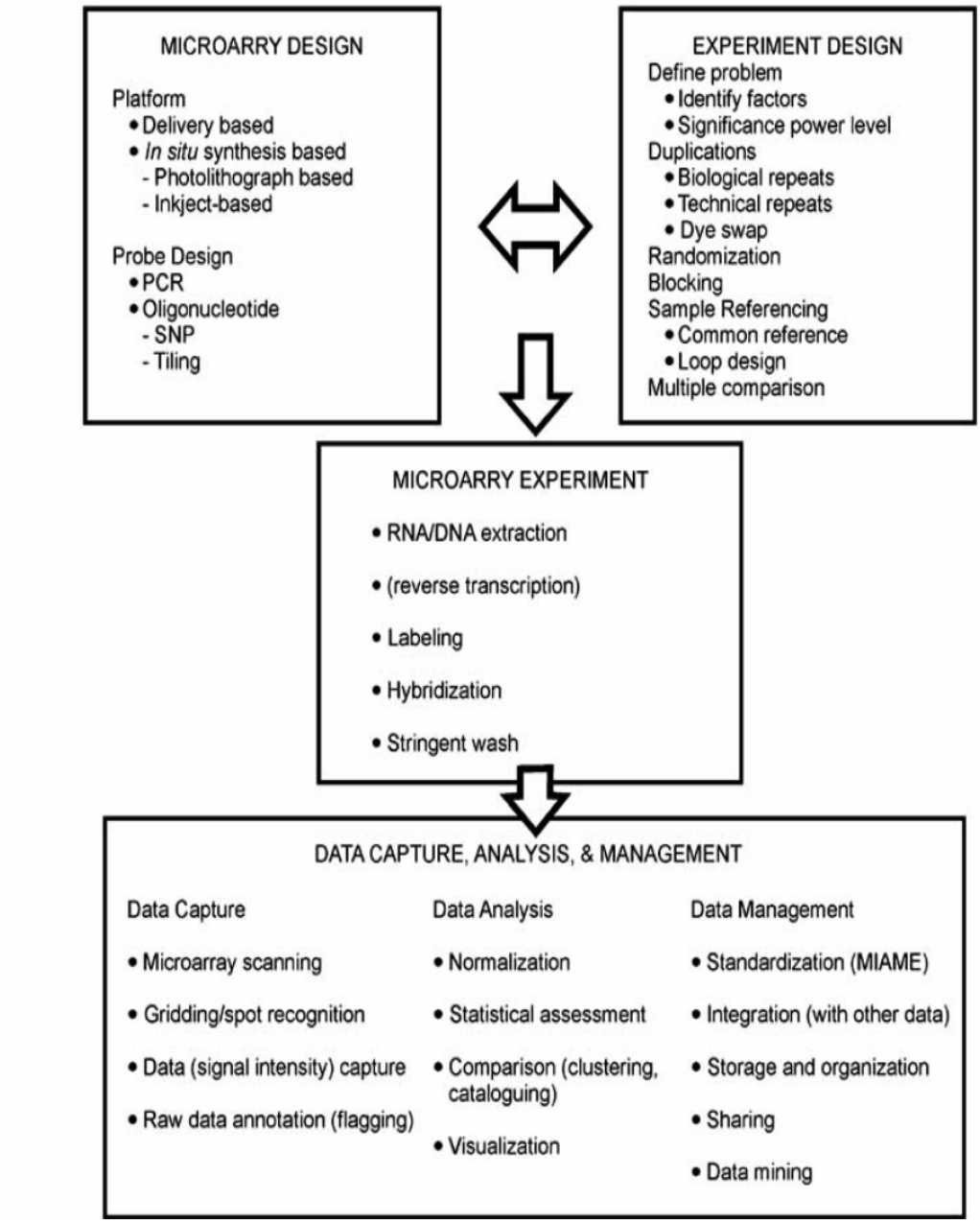
This process involves transcription and translation. The information contained in DNA is used to code for the required protein.

This molecular biology dogma, "DNA => mRNA => Protein" is the basis of microarray technology as DNA microarray measures mRNA transcript as gene expression level. DNA code (genetic code) which is a triplet codon arising from a combination of three(3) of the

four (4) nitrogenous bases: Adenine (A), Thymine (T) ,Guanine (G), and Cytosine (C) form the type of amino acid or protein to be produced while a stop codon terminates the process.

The availability of genomic sequence data for many organisms, has made it possible for scientists to easily design microarrays containing tens of thousands of sequence probes to interrogate the behaviour of all the genes in an organism. This approach has revolutionized our way of performing biological research from the “one-gene-one experiment” approach to a “global” or “genome-wide” systemic study (Chen, 2006).

A typical DNA microarray workflow is depicted in Figure 3.2. Microarray designs in terms of suitable platform and appropriate probes should be made available from the start. The experimental design issues should be well sorted out in advance and required materials provided before the commencement of the experiment. With the experimental design in place, the microarray experiment can be performed, followed by data capture, analysis and management.



(Source: Chen, 2006)

Figure 3.2: A General Workflow of a DNA Microarray Experiment.

The workflow depicts microarray, probe and experiment designs preceding the stage of the actual experiment which is usually followed by data capture and analysis.

3.1.1 DNA MICROARRAY FABRICATION AND PROBE DESIGN

Microarray's fabrication is achieved through two technologies and involves either DNA *deposition* or *in situ* synthesis. While deposition method allows the deposition of PCR-amplified cDNA clones and printing of already synthesized oligonucleotides with fine-pointed pins onto glass slides, *in situ* manufacturing is by photolithography using pre-made masks, ink-jet printing, or electrochemistry on microelectrode arrays. Nucleic acid microarrays primarily use short oligonucleotides (15–25 nucleotides), long oligonucleotides (50–120 nucleotides) and PCR-amplified cDNAs (100–3,000 base pairs) as array elements. Due to varied and evolving technological trend in the function of DNA microarray, it will be convenient to categorize them into two types: One-colour and Two-colour Microarray platform technologies.

3.1.1.1 One Colour Fabrication Technology

In this fabrication method, only one stain is used during hybridization. Examples include Affymetrix, Nimblegen, Agilent, Illumina microarray technologies.

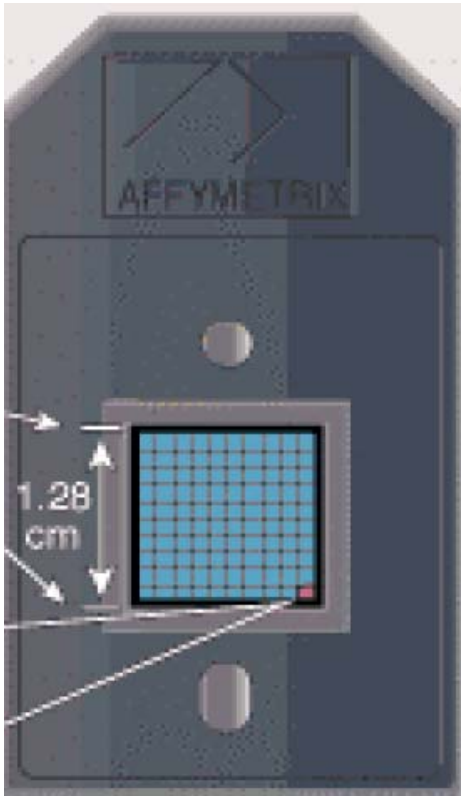
3.1.1.1.1 Affymetrix Oligo Platform

The main features of Affymetrix Oligo Platform as shown in Figure 3.3(a) include:

- High density, up to 40,000 genes on a chip.
- High reliability, 16-20 pairs of probes represent each gene.
- Single-channel hybridization due to high reproducibility between spots and arrays.

In Affymetrix, short probes of 15-25 nucleotides long are possible because they are selected using special algorithms to run on already sequenced genome data to find unique sequences that serve as a representative of each gene in an organism. Figure 3.4 (b) shows a probe with both Perfect match and Mismatch pair aligned to a reference. A probe cell also called feature contains 25 nucleotides and can be a perfect match (PM) or mis-match (MM). Usually, probes are manufactured in pairs such that PM has the same sequence of 25 nucleotides as the MM except for one nucleotide at the middle (13^{th}) position which is complementary. PM hybridizes with the experimental sample to measure the degree of signal intensity while the MM hybridizes to give value for the background subtraction

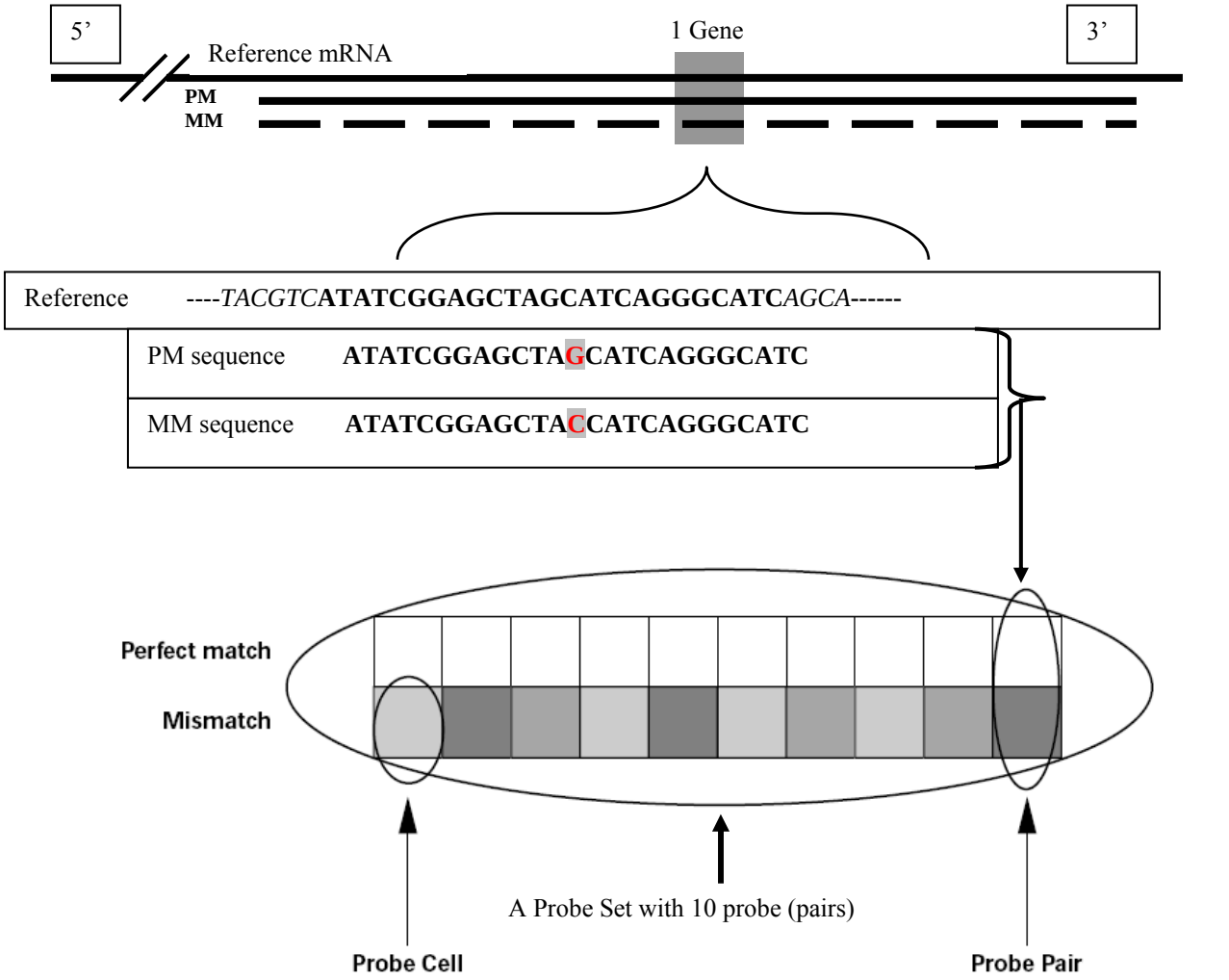
which improves data accuracy. One or more probes can be used to represent a gene and a typical Affymetrix probe contains about 16-20 probes in a probe set.



(Source: Xu and Vernick, 2006)

Figure 3.3 (a) Affymetrix GeneChip.

This is a very handy and flat device with high density capable of containing up to 40,000 genes on a chip. It has high reliability and16-20 pairs of probes represent each gene.

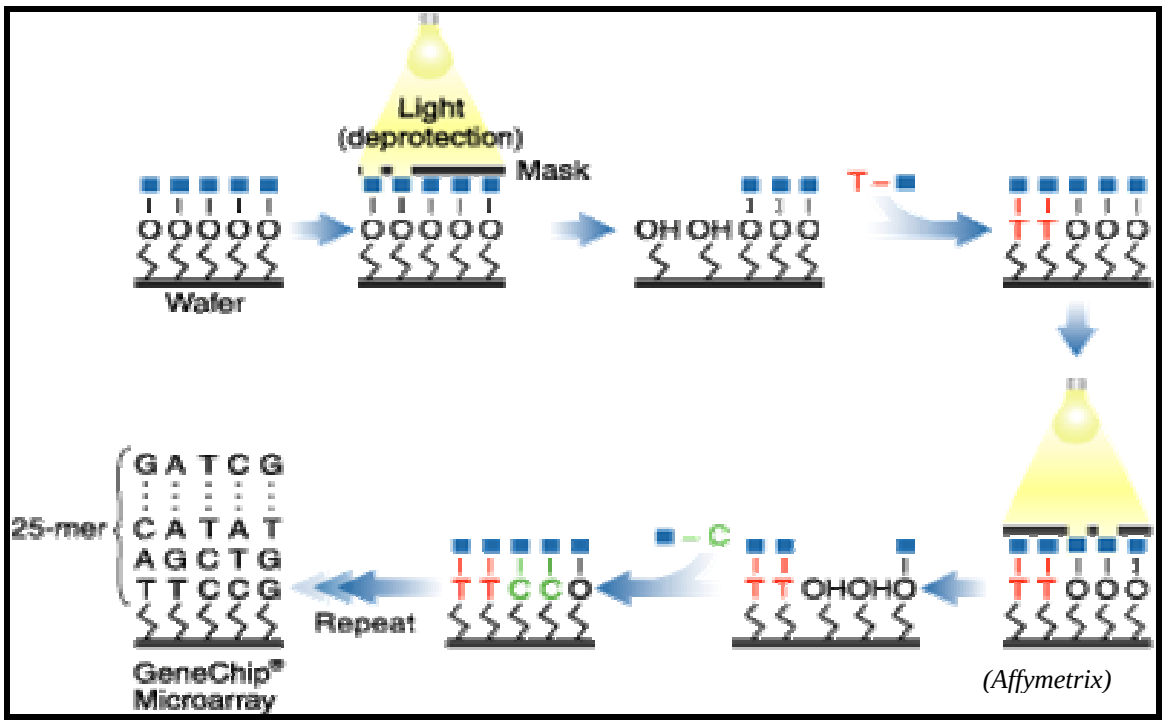


(Adapted from Affymetrix, 2001)

Figure 3.4 (b) Design of Oligonucleotide probe (feature).

It shows its location in the probe set and 1 gene represented by 1 probe. A probe contains 25 nucleotides. The difference between Perfect Match and Mismatch probe is one nucleotide and the Mismatch is to support background correction from noise.

Photolithographic manufacturing process produces GeneChip arrays with millions of probes on a small glass chip or substrate called wafer or Array. The photolithographic process begins by coating a 5″ x 5″ quartz wafer with a light-sensitive chemical compound that prevents coupling between the wafer and the first nucleotide of the DNA probe being created. A physical illustration of the photolithographic manufacturing construction process is shown in figure 3.5.



(Source: Wosik, 2006)

Figure 3.5: Photolithographic Manufacture of Affymetrix Oligonucleotide Array.

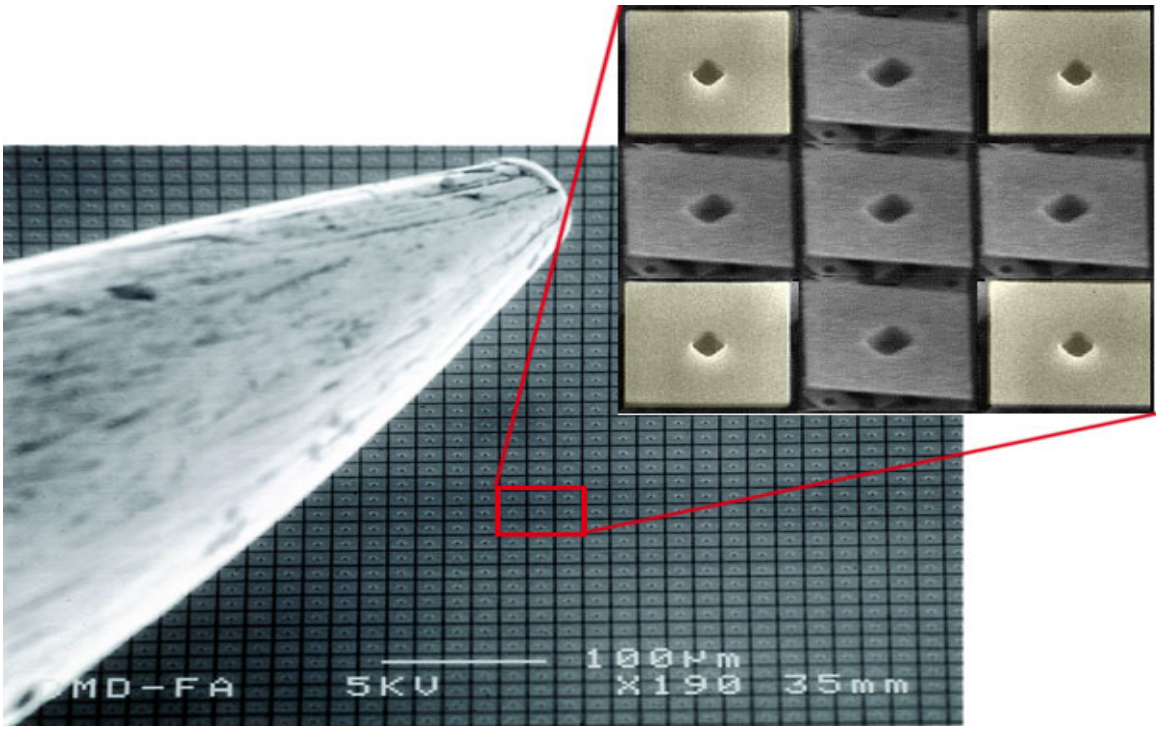
A mask is used to protect the wafer. The shifting of the mask permits light to aid the free nucleotide in solution to attach to the wafer and grow till it reaches 25 nucleotides. The 'O' changes to 'OH' under light and it is in turn displaced by nucleotides A, C, G, T.

3.1.1.1.2 Nimblegen Oligo Platform

The following are the main features of Nimblegen oligo chip:

- Maskless array synthesis controlled by software-driven micro-mirrors;
- Digital Micromirror Device (DMD) is an array of 786,000 aluminum mirrors on a computer chip;
- Each mirror is individually addressable to shine light in specific patterns on the chip;
- Photodeposition chemistry produces arrays of oligonucleotide probes.
- ~400,000 probes/array (~20,000 genes @ 20 probes/gene);
- Array design completely dynamic from one chip to the next (ie, can change 1 nucleotide of one probe, or completely change the sequence of every probe);

An example of Nimblegen oligo chip platform is given in figure 3.6



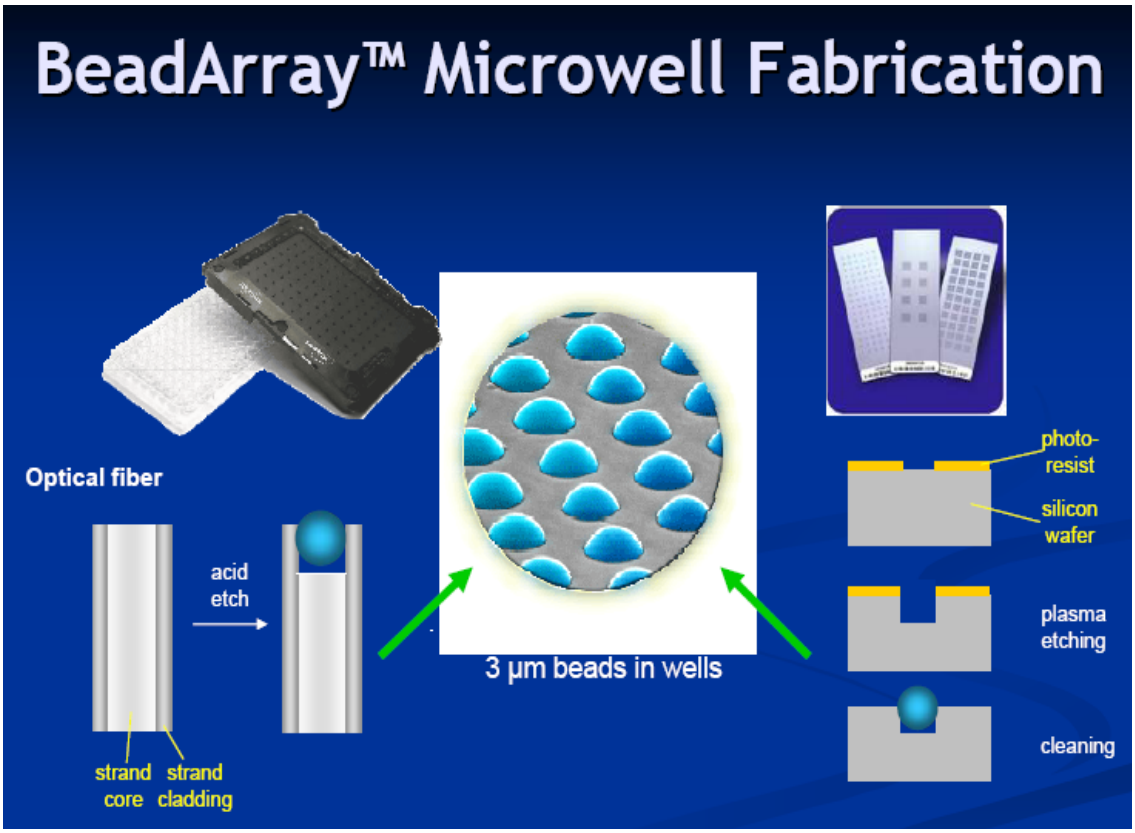
(Source: Xu and Vernick, 2006)

Figure 3.6: Nimblegen Maskless Array Synthesis.

This type of array constitutes a digital micromirror device (DMD) which is an array of 786,000 aluminum mirrors on a computer chip guided by software-driven micro-mirrors. Each mirror is individually addressable to shine light in specific patterns on the chip; Photodeposition chemistry produces arrays of oligonucleotide probes.

3.1.1.1.3 Ambion Illumina Oligo Platform

The technology uses transcript-specific 50mer oligonucleotide DNA probes attached to small (3µM) beads through an "address" linker sequence (see figure 2.9). It is made of optic fibre with a strand core cladding either side. There is a well on the strand core which is covered by the bead on the outside. It is called also called BeadArray.



(Source: WTAC, 2007)

Figure 3.7: Illumina Microarray Fabrication.

This technology uses beads placed in wells created by the use of optic fibres. The chip is similarly made on a silicon wafer. The arrow shows the integration of the various parts that give rise the to the visible beads in wells.

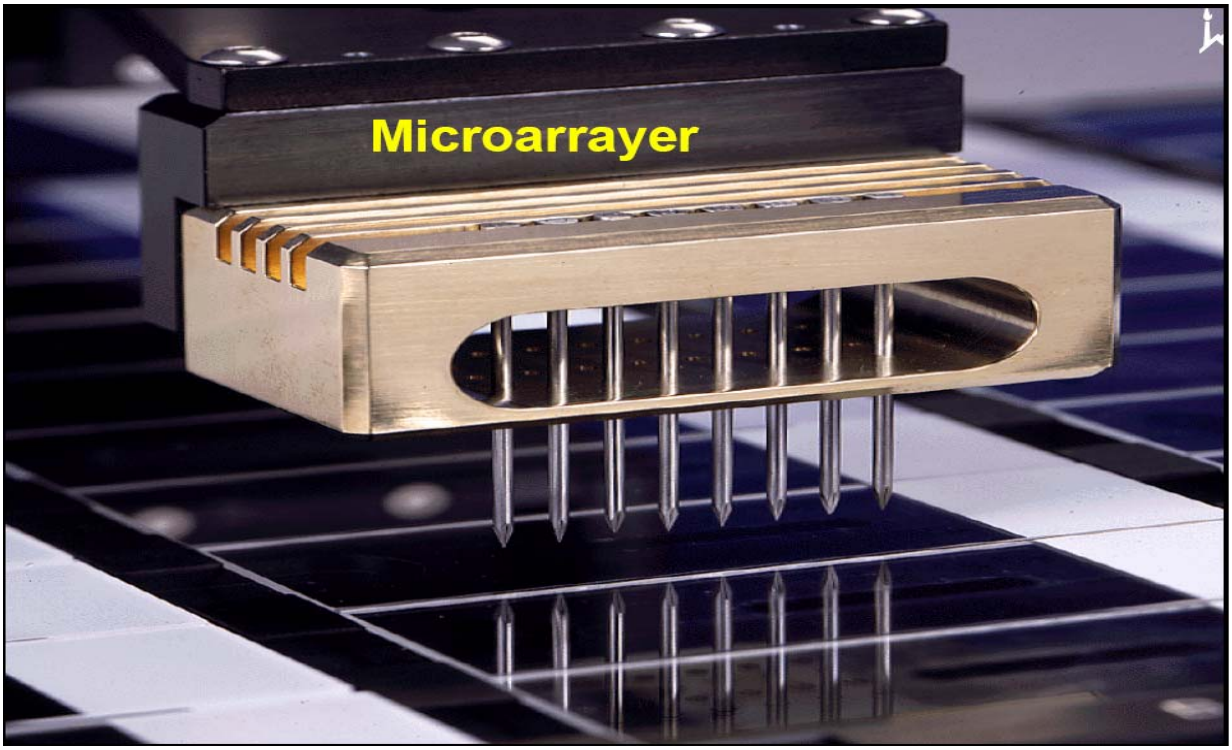
3.1.1.2 Two-Colour Fabrication Technology

A good example of the of two-colour fabrication is the cDNA microarray

3.1.1.2.1 cDNA Microarray Platform

Array fabrication is by DNA clones chosen based on annotation (functional identities, pathways, chromosomal location) or obtained experimentally (EST library from subtractive screen). The probe is designed through polymerase chain reaction (PCR) amplification of the clones and purification under quality control. Spotting on the array is done by robotic arrayer (see figure 3.8). The two types of techniques for printing oligos in the slide by the robotic arrayer are given in figure 3.8. The microarrays containing PCR amplicons are

usually referred to as “cDNA microarrays” as the PCR products are derived from either predicted open reading frames (ORFs) or expressed sequence tags (ESTs).

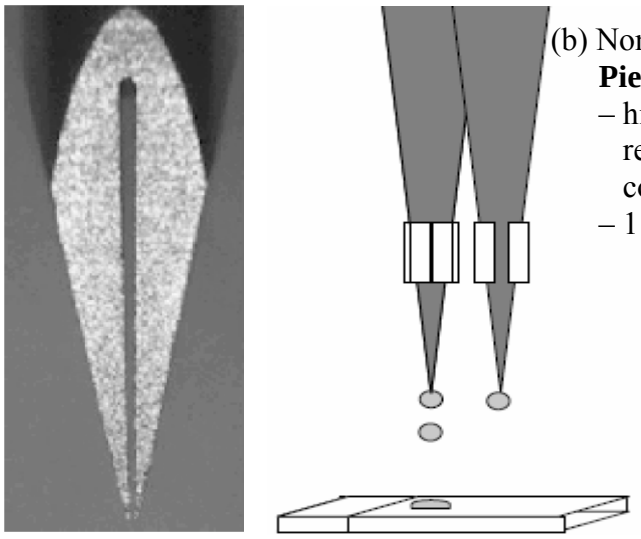


(Source: Xu and Vernick, 2006)

Figure 3.8: Microarrayer- a Robotic Spotting on a Glass Slide for cDNA Microarray.

The pointed pin is used to print oligos on a glass slide. This technique of printing can be either contact or non-contact printing as in figure 3.8.

(a) Contact Printing
Pins
-uptake ~0.25 ul
-dispense ~0.6 nl
-approx 1-10ng DNA per spots



(b) Non-Contact Printing
Piezoelectric or ink jet
– higher spot-to-spot reproducibility than contact printing
– 1 drop ~100 pl

(Source: Xu and Vernick, 2006)

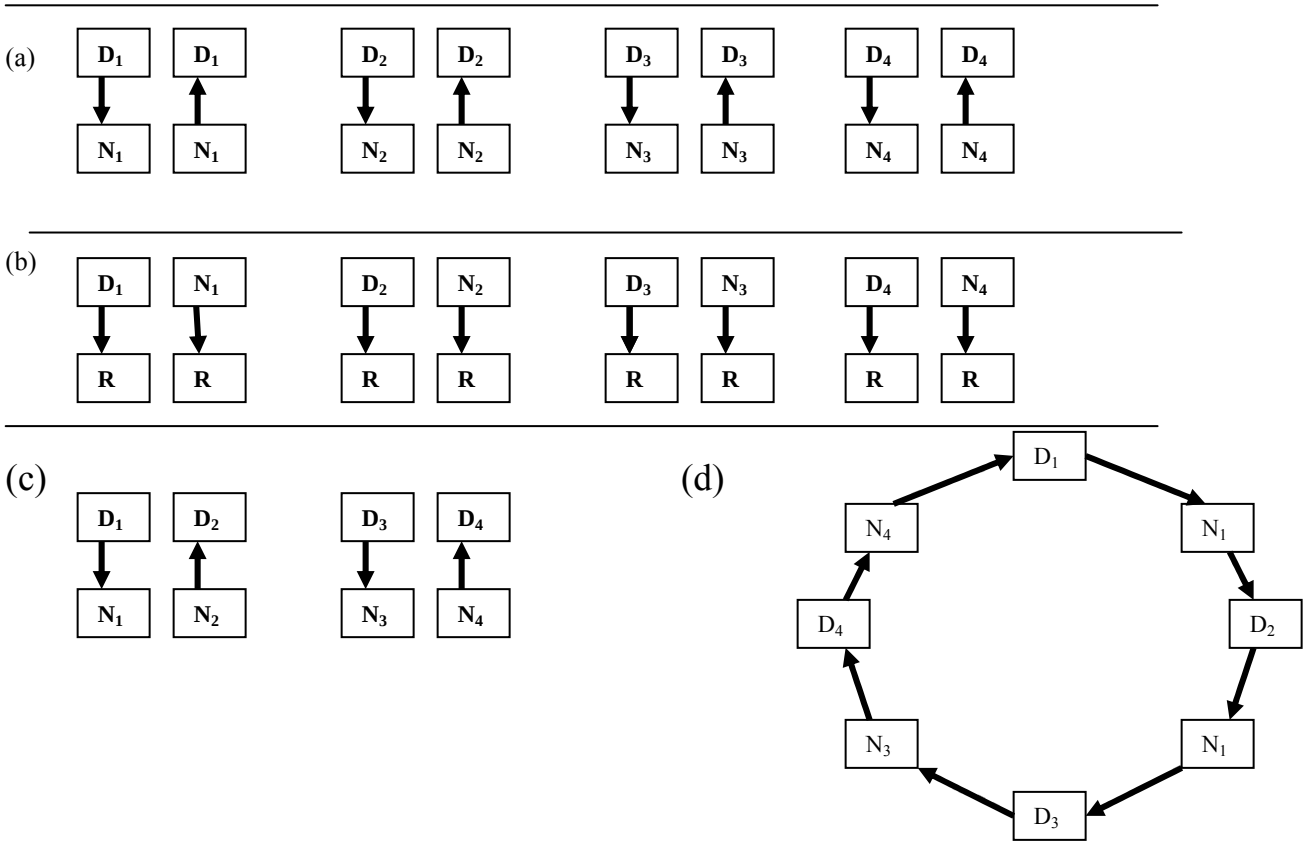
Figure 3.9: (a) Contact and (b) Non-Contact Printing of Oligos on Slide.

Contact printing as in (a) uses pin that dispenses ~ 0.6nl upon uptake of ~0.25ul oligos by direct contact or touch of the pin on the slide while non-contact printing is done using ink jet that drops the oligo on a slide without touching it. Microlitre (u*l*) = $10^{-6}l$, Nanolitre (n*l*) =$10^{-9}l$, Picolitre (p*l*)= $10^{-12}l$.

3.1.2 EXPERIMENTAL DESIGN

Often neglected is a well thought-out relevant experimental design to address the appropriate comparison to be made in microarray studies for answering suitable biological questions. Microarray experimental design allows researchers to test and vary the input variables that impact on the microarray experiment to get correct output. These involve three major principles: Replication, Randomisation and Blocking (Draghici, 2003). In replication, it duplicates, or repeats the same experiment more than once but varies one factor like changing the location of the same probe on the array to monitor its behaviour. This gives an estimate of the experimental error to conclude as to whether or not the differences observed in the data are significant. Randomisation means that experimental data can be monitored by allowing probes to be placed on the array in no particular order (random). Blocking allows the experimenter to keep all nuisance factors (not of interest but can affect the experiment outcome) while interesting factor is varied (Quackenbush, 2001).

Some microarray specific types of experimental design include Direct-dye swap design, Balanced block design, Reference design and Loop design (see figure 2.12). In reference design, many sample conditions or time points t_1, t_2,…,t_n are pair wise compared to only one reference sample *ref.* Samples 1 to n are measured once while *ref* is measured *n* times. Vinciotti *et al.* (2005) studied the relative efficiency of both a loop and a reference design using the same RNA preparations. Their results of these experiments show that (1) the loop design attains a much higher precision than the reference design, (2) multiplicative spot effects are a large source of variability, and if they are not accounted for in the mathematical model, for example, by taking log-ratios or including spot effects, then the model will perform poorly.

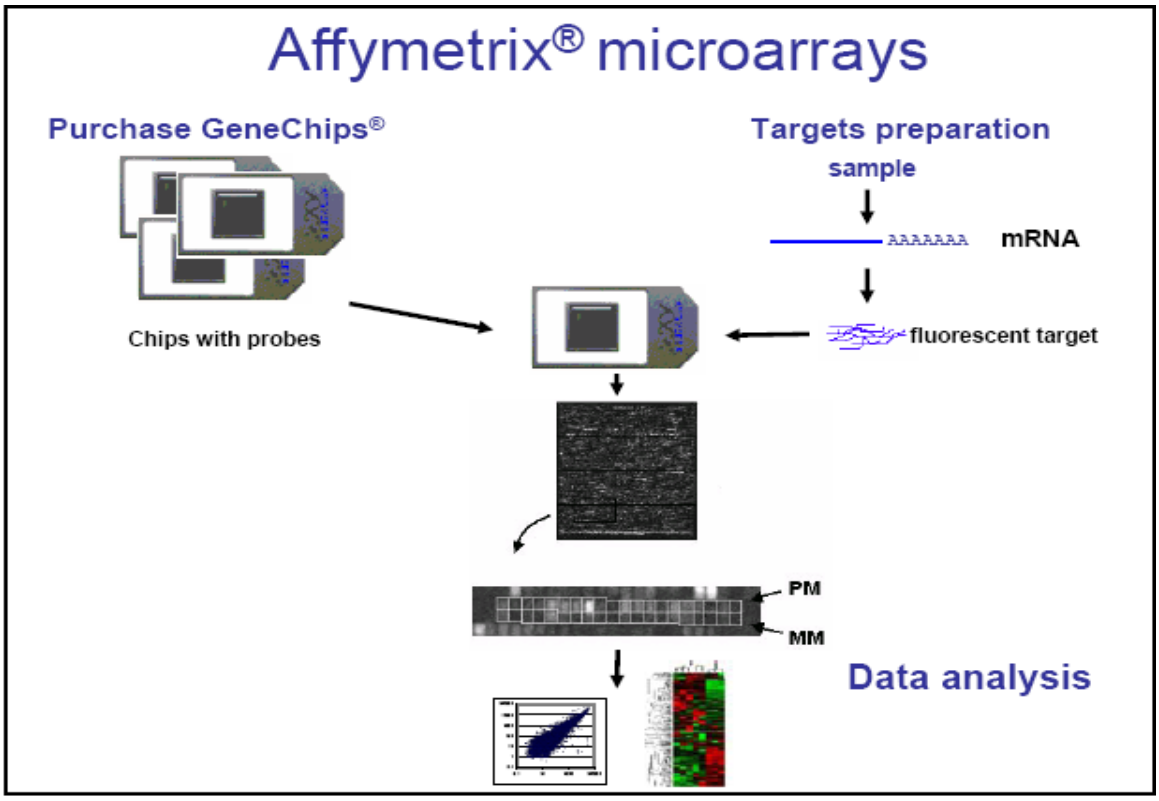


(Quackenbush, 2005)

Figure 3.10: (a) Direct comparison with Dye Swap , (b) Reference Design (c) Balanced Block Design (e) Loop Design
Boxes= Individual samples to be compared, D "Diseased" and N "Normal"
Arrow= Hybridisation assay with tail end for 1st dye and head end for 2nd dye.

3.1.3 OLIGONUCLEOTIDE MICROARRAY EXPERIMENT

Once the microarray is constructed by Affymetrix, Oligonucleotide chip experiment requires the preparation of a sample for GeneChip arrays. Messenger RNA (mRNA) is extracted from the cell and converted to cDNA as shown in figure 3.11. It then undergoes amplification and labeling where the target mRNA population is labeled; typically with a fluorescent dye, so that hybridization to the probe spot can be detected when scanned with a laser (Gibson, 2003). Fragmentation and hybridization of the sample to the 25-mer oligos on the surface to the chip takes place under an appropriate temperature. The next step is the washing of unhybridized material, the chip, scanned in a confocal laser scanner and the image analyzed by computer. This approach provides a way to use directly the growing body of sequence information for experimental investigations (Wosik, 2006). However, one sample is hybridized on one array, unlike cDNA which is capable of hybridizing two distinctly labeled (R and G) samples on one array.

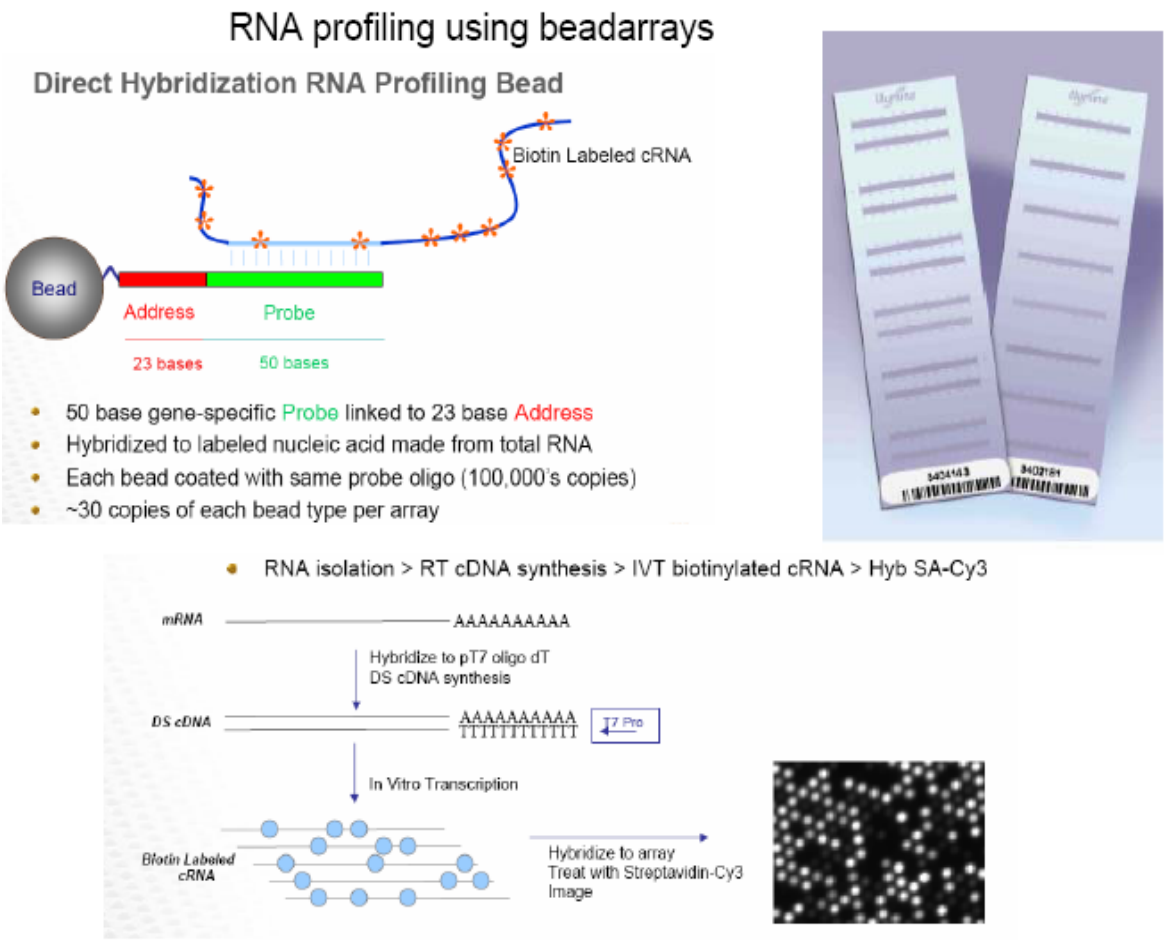


(Source: Wosik, 2006)

Figure 3.11: Oligonucleotide Chip Experiment.

The chip was bought from Affymetrix and target sample is capable of fluorescence using PM and MM on hybridisation.

Illumina follows the RNA isolation step from sample to obtain cDNA by RT, which is used for IVT biotinylated cRNA and hybridisation with Cy3. Biotin-labeled cRNA is hybridized to the array and the array is stained using a post-hybridization cocktail containing strepatavidin-Cy3 (see figure 3.12). This allows six RNA samples to be analysed per slide. However, if our RNA sample is small, we can apply a two round amplification kit which allows us to start with smaller quantities of RNA.



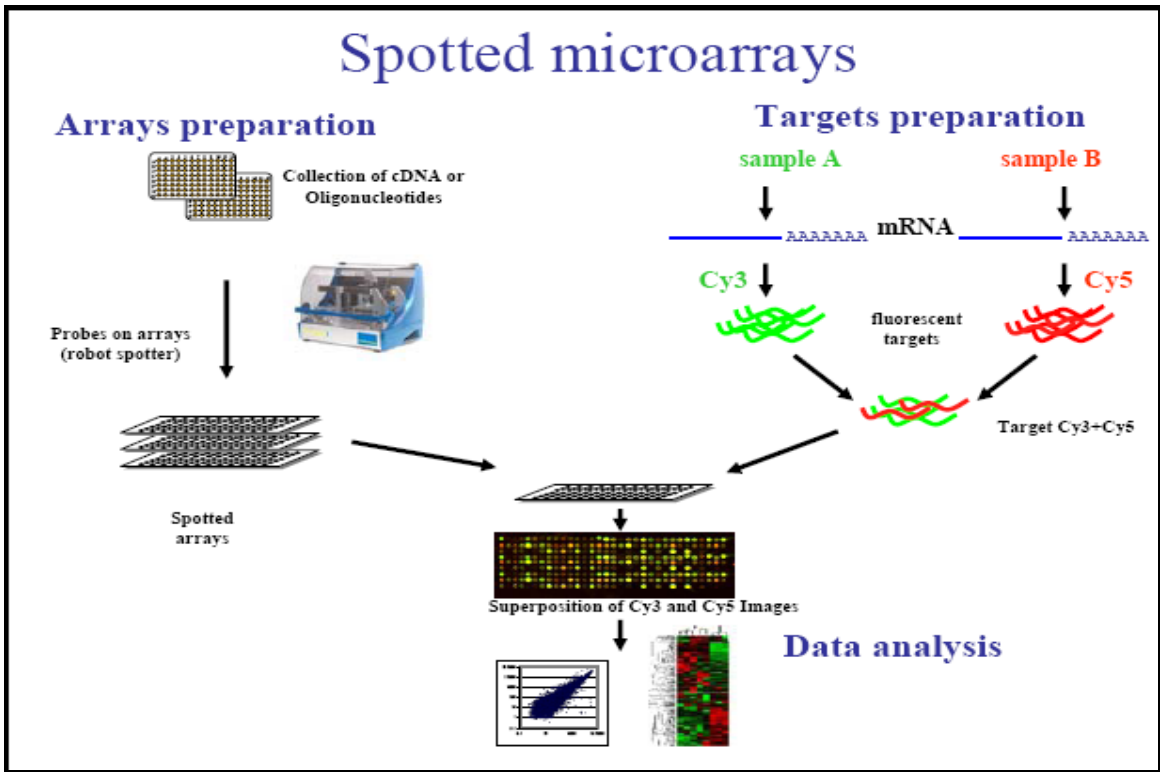
(Source: WTAC, 2007)

Figure 3.12: Illumina Microarray Experiment.

This requires RNA isolation and converted into cDNA using Reverse Transcriptase to obtain IVT biotinylated cRNA which hybridizes with Cy3**.** The bead has 23 addressable bases and 50bases used for probe.

3.1.4 cDNA MICROARRAY EXPERIMENT

Spotted or cDNA or Two-channel microarrays consist of thousands of individual DNA sequences called probe printed in a high density array on a glass microscopic slide using a robotic arrayer. The process of probe printing is shown in Figure 3.13 during array preparation. During target preparation, mRNA is extracted from two samples called A and B to be studied. These target samples mRNA are reverse-transcribed into cDNA and labeled using different florescent dyes where Cy3 represents green and Cy5 represents red. Labeled samples are mixed together and competitively hybridized with the probe on the array, to give rise to image for analysis. Relative abundance of spotted DNA sequences or probe in two samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array.



(Source: Wosik, 2006)

Figure 3.13: cDNA Microarray experiment.

The Target preparation is for two chanels (Red and Green) which hybridize competitively on the chip. Required for hybridization solution preparation are the following: RNA extraction (experimental & reference samples), RNA labeling , Enzymatically synthesize first-strand cDNA, Reverse transcriptase, Incorporate fluorescently labeled deoxyribonucleotides (dNTPs), Cyanine5 (Cy5) labels Experimental sample (red), Cyanine3 (Cy3) labels Reference sample (green), cDNA purification, mix the two labeled RNAs. Data acquisition requires GenePix software and need is a chip description file (CDF).

3.1.5 Creation of Microarray Images and Data analysis

After competitive hybridization, images of slides are taken by Microarray Scanner which makes fluorescent measurement of each dye. The ratio of the fluorescence intensities for each spot is indicative of the relative abundance of the corresponding DNA sequence in the nucleic acid samples. cDNA microarray image processing steps generate two main quantities R (red) and G (green) for each spot on the array, thereby measuring transcript abundance for red and green mRNA labeled samples. The R and G values are usually combined or normalized into a single log-intensity ratio, $log_2\ R/G$, measuring relative transcript abundance in the two samples. A positive log-ratio denotes gene over-expression, while a negative log-ratio denotes gene under-expression. Normalisation needs to be done before clustering for further data analysis so as to identify and remove systematic sources of variation such as different labeling efficiencies, scanning properties of dyes used, and print tip of robotic arrayer during probe spotting.

Microarray gene expression profiles are often subjected to cluster analysis and pathway mapping to unveil groups of co-regulated genes – a practice that is referred to as regulatory network and metabolic pathways discovery or reconstruction. A common early step in microarray data analysis is log transformation. Typically, log base 10 is used (Affymetrix, 2001); however, log base 2 or natural log will work equally well. Log transformation has several important effects on the data. The most critical reason for log transform microarray data is that some of the error in the signal intensity measurement increases as the magnitude of signal intensity increases. That is, small numbers have less error in an absolute sense than higher numbers. Fortunately, higher numbers have roughly the same percentage error as small numbers. This roughly constant factor can be simply calculated and subtracted to normalize the data, once the signals have been log transformed. There are additional effects of logging that make log-transformed microarray data more closely fit statistical assumptions, when applying statistical test methods. Log transformation makes data more symmetrical, one of the standard assumptions of normality. Log transformation also reduces the influence of a single measurement. Means on a log scale are more like geometric means, which are resistant to the effects of outliers, and it follows that outliers result in better estimates of variance. So, by log transforming data, common statistical

methods are made more reasonable and provide more accurate insights for the biologist. To uncover hidden knowledge buried in the huge data, further analysis, using tools like k-means clustering is desirable.

3.1.6 MICROARRAY SOFTWARE SUPPORT

There are various commercial and open source software products existing currently to support DNA microarray analysis. For this work, we consider a free open source software TM4 (Saeed *et al.,* 2003) from The Institute of Genomic Research (TIGR) used for cDNA microarray analysis, a commercial GeneChip Operating Software (GCOS) from Affymetrix and dChip software for Oligonucleotide microarray.

3.1.6.1 TIGR TM4

The Institute of Genome Research (TIGR) TM4 suite of tools consists of four major applications, Microarray Data Manager (MADAM), TIGR_Spotfinder, Microarray Data Analysis System (MIDAS), and Multiexperiment Viewer (MeV), as well as a Minimal Information About a Microarray Experiment (MIAME)-compliant MySQL database, all of which are freely available to the scientific research community at http://www.tigr.org/software. The MADAM data entry interface provides access to data associated with a microarray study. It has a navigation panel on the left-hand side which leads users through the process of data entry during a microarray experiment. TIGR Spotfinder provides image processing with direct connections to the microarray database. MIDAS allows users to define data normalization and filtering protocol using a simple graphical scripting interface. MeV allows users to apply a number of sophisticated data mining tools to their array data and provides integrated graphical depictions of the results from the analyses conducted. Three of the TM4 applications, MADAM, MIDAS, and MeV, were developed in Java and have been tested on Microsoft ® Windows, Linux, Unix, and MacOS X platforms; TIGR Spotfinder was written in C/C++ and runs only on Windows systems (Saeed *et al.,* 2003).

3.1.6.2 GeneChip Operating Software (GCOS)

GCOS is proprietary software currently used to analyse Affymetrix Oligonucleotide microarray after hybridization of probe with target samples, since Microarray Suite (MAS) has been discontinued. Image of the slide is captured via a scanner and expression values are generated into a *.dat file (Data File). The software derives the *.cel file (Cell Intensity File) from a *.dat file and automatically creates it upon opening a *.dat file. It contains a single intensity value for each probe cell delineated by the grid (calculated by the Cell Analysis algorithm). Chip File *.chp, the output file generated from the analysis of a probe array contains a qualitative and quantitative analysis for every probe set. Report File *.rpt is a text file summarizing data quality information for a single experiment and is generated from the analysis of output file (*.chp). There are also other output files involved in the use of GCOS such as *.cab (Cab File) which is a compressed file that is a backup copy of a process or publish database, project, sample, and/or experiment. *.txt and *.xls are standard formats for text files and spreadsheet files and GCOS exports text in these file formats. The Library Files (probe information) *.cif, *.cdf, and *.psi contain information about the probe array design characteristics, probe utilization and content, scanning and analysis parameters. These files are unique for each probe array type. Fluidics Files include *.bin and *.mac. The fluidics files contain information about the washing, staining, and/or hybridization steps for a particular array format.

3.1.6.3 DNA-Chip Analyser (dChip)

DNA-Chip Analyzer (dChip) is a software package for probe-level and high-level analysis of Affymetrix gene expression microarrays and SNP microarrays (Li and Wong, 2001a; Li and Wong, 2001b; Lin et al., 2004). However, gene expression or SNP data from other microarray platforms can also be analyzed by importing as external dataset. High-level analysis in dChip includes comparing samples, hierarchical clustering, Loss Of Heterozygosity (LOH) and copy number analysis of SNP arrays. To use dChip, the user needs to provide Affymetrix array data files (in CEL or DAT format, or see public CEL files), and the CDF file (Chip Description File). Obtain the dChip and gene information files and CEL file if required into a local directory and double click to start the dChip program. Affy conversion tool will convert all CEL files in a directory from version 4 to

version 3, while leaving the file name the same. You can check the CEL file's change in size to confirm if conversion is done and also ensure that CEL files are not read-only. If conversion fails, DAT files can also be read by dChip instead of CEL files.

To read in cDNA array data, an external data file with every two columns as the green and red channel intensities from one array (e.g. obtained from GenePix GPR file), is read into dChip by "Analysis/Get external file" before continuing with data analysis. dChip is a freeware, single executable program developed on Windows 2000 but preferring windows NT/XP computers with 512 Megabytes memory for maximal operation. dChip is written in Visual C++ 6.0 and uses Windows-specific functions for graphic tasks, and the source code is freely available for academic purposes.

3.1.7 APPLICATIONS OF DNA MICROARRAY

3.1.7.1 Gene Expression Profiling Applications

Gene expression profiling applications include the following:

- Pathogenesis studies
- Pathogen's responses to drugs
- Pathogen's responses to host
- Host's responses to infection
- Host's responses to Treatments
- Host response to Vaccines

Expression levels for tens of thousands of genes can be simultaneously measured in a single hybridization experiment and are collectively called a "gene expression profile". In the gene expression profiling experiments, the biological samples that the probes are designed to interrogate are RNA extracted from cells or tissues. This type of experiments answer the question of "what genes and how much of them are expressed in the biological sample?". The RNA molecules are first converted to cDNA by reverse transcription and labeled with a fluorescent dye. The expression level of a gene are measured as the light

intensities emitted, after excitation with laser light, by the fluorescent dye attached to the cDNA that bound the homologous probes on the array (Chen, 2006).

Gene expression profiles of the host to pathogen may also be used in diagnosis for identifying possible pathogens. DNA microarray technology enables scientists to perform global survey of novel virulence factors, antimicrobial drug resistance genes, and potential vaccine targets by monitoring the transcription profiles of the pathogens in response to host environments.

For example, two recent studies used the microarray approach to monitor the gene expression of the malaria pathogen *Plasmodium falciparum* in the host cells and have identified potential vaccine candidates or drug targets. Daily et al. (2007) studied the gene expression profiles of *P. falciparum* that was isolated from blood samples of infected patients and compared them with the *in-vitro* profiles of a reference *P. falciparum* strain at the ring stage. A new family of hypothetical protein that may encode surface antigens were found to be over-expressed in the *in-vivo* samples, making these potential candidates for vaccine development.

Gaur et al. (2006) identified new virulence genes by comparing gene expression profiles between two *P. falciparum* clones. The *P. falciparum* Dd2, a parasitic clone that requires sialic acid residues on the erythrocyte surface for successful invasion, is capable of undergoing a genotypic switch to become a subclone Dd2 (NM), which can invade erythrocytes without the sialic acid residues. By comparing the expression profiles of these two parasitic clones, four novel genes were initially identified to be up-regulated in the sialic independent clone Dd2 (NM). Two of these genes, PfRH4 and PEBL, were confirmed by RT-PCR and the expression of PfRH4 at protein level was further confirmed to be only in Dd2 (NM) (Chen, 2006).

3.1.7.2 Gynotyping Applications

Genotyping applications include the following:

- Pathogen Identification
- Drug Resistance Survey
- Host Susceptibility
- Pathogen Cataloging
- Vaccine Re-Evaluation

In the genotyping experiments, the targets are DNA extracted from the biological samples and the probes are designed to survey the sequence variations in or among the samples. “Single nucleotide polymorphism (SNP) microarray” is an example of genotyping microarray. A variation of the SNP microarray is called “sequencing microarray” or “re-sequencing microarray” and can be used to re-sequence a specific region of a closely related genome, of which the sequences have to be decoded.

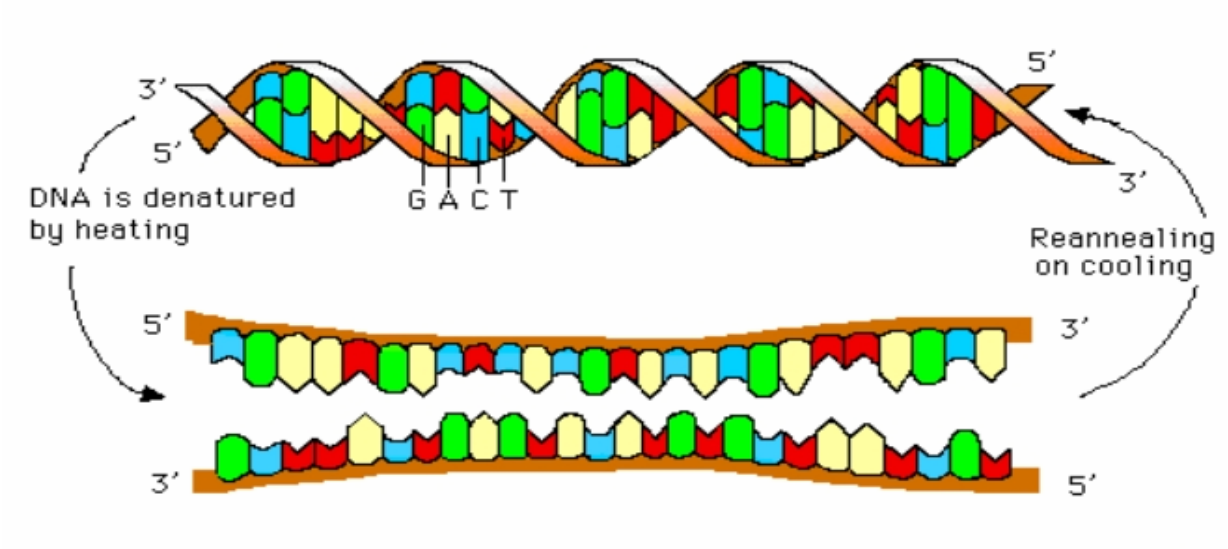
The most direct and perhaps also the most widely used applications of DNA microarray technology in infectious diseases fall in this category. DNA microarrays allow quick identification of the pathogens based on the unique sequence signature detectable by the large number of the probes on the array.

A good example is the identification of a new corona virus that caused the severe acute respiratory syndrome (SARS) epidemic outbreak in 2003. Before the outbreak, Wang et al. (2003) had devised a microarray intended for detecting the widest possible range of both known and unknown viruses. This viral microarray platform contained probes representing all the approximately 1,000 known virus sequences at the time from GenBank. In March 2003, Wang et al. (2003) used this microarray to quickly identify the viral agent in SARS samples as a new type of corona virus.

3.2 PCR TECHNOLOGY

3.2.1 WHAT IS PCR TECHNOLOGY?

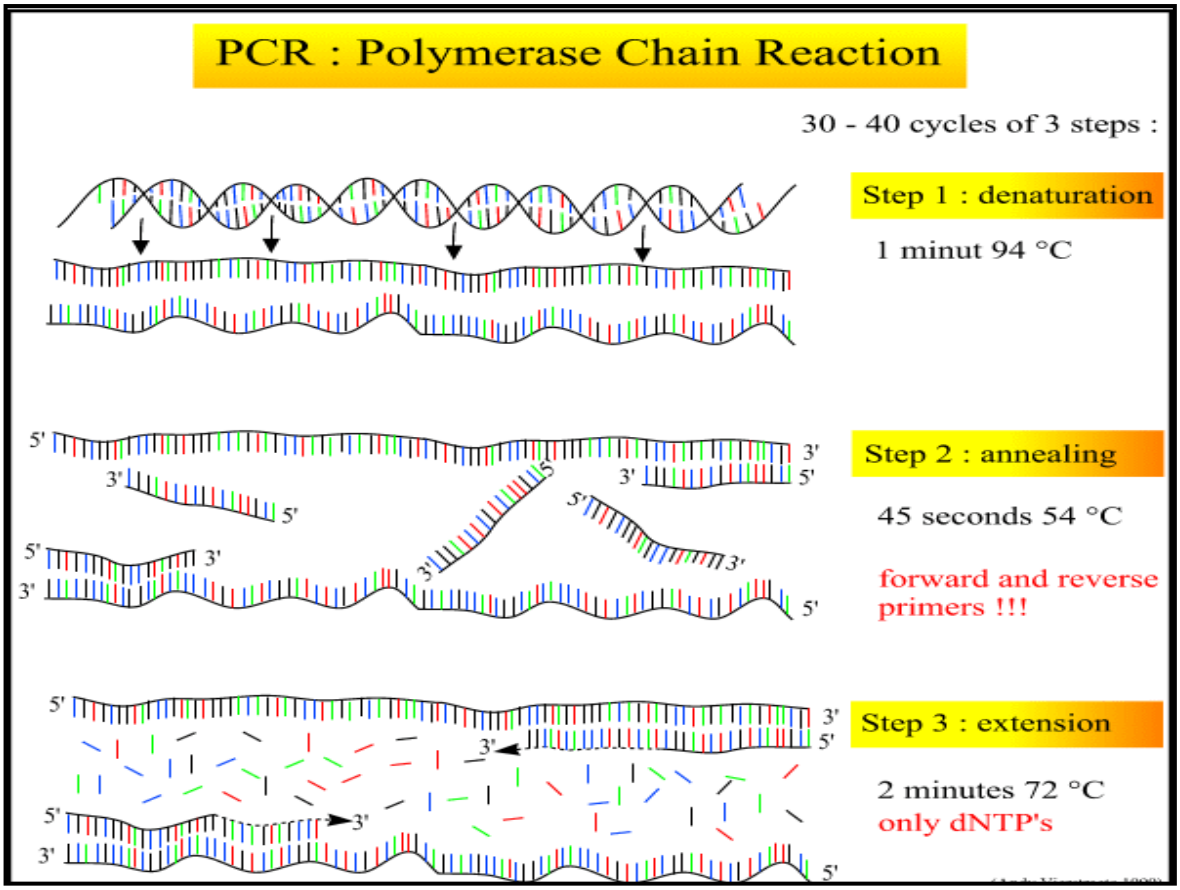
The polymerase chain reaction (PCR) is a common method for amplifying DNA or making unlimited copies of genes of interest. PCR is a cyclic process with the doubling of DNA targets at each cycle. The strands in each targeted DNA duplex are separated by heating and then cooled to allow primers to bind them. The DNA polymerases extend the primers by adding nucleotides to them (Anderson, 2006). In this way, duplicates of the original DNA-strand targets are produced (see Figure 3.14 - 3.15).



(Source : Xu and Vernick, 2006)

Figure 3.14: The Concept of Denaturation and Reannealing Process of Double Stranded DNA Molecule.

When heat is applied to a dsDNA, it separates but it anneals again on cooling.

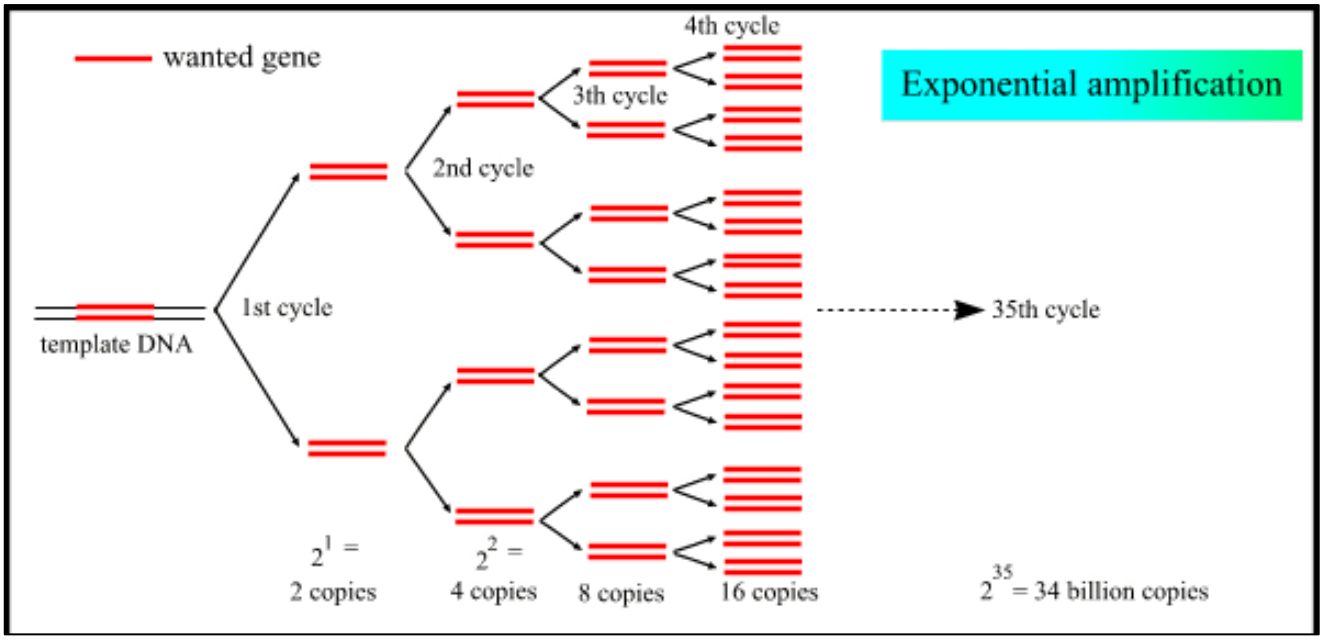


(Adapted from Mullis, 1990)

Figure 3.15: Polymerase Chain Reaction (PCR) Stages (Denaturation, Annealing and Extension).

Target double stranded DNA is heated up to separate and allow the primers to anneal at a lowered temperature, before polymerase comes to extend each strand.

PCR was invented by Kary Mullis in 1982 when at Cetus Corporation. Beginning with a single molecule of genetic material DNA, the PCR can generate several billion similar molecules in one afternoon (Mullis, 1990). This is exemplified in Figure 3.16.



(Adapted from Mullis, 1990)

Figure 3.15: Polymerase Chain Reaction (PCR) Exponential Synthesis.

PCR is a cyclic process with the doubling of DNA targets at each cycle. This depicts exponential synthesis of a DNA fragment up to the 35th cycle yielding about 34 billion copies.

In order to robustly detect and quantify gene expression from small amounts of RNA, amplification of the gene transcript is necessary. For mRNA-based PCR the RNA sample is first reverse transcribed to cDNA with reverse transcriptase. Development of PCR technologies based on reverse transcription and fluorophores permits measurement of DNA amplification during PCR in real time, i.e., the amplified product is measured at each PCR cycle. The data thus generated can be analysed by computer software to calculate relative gene expression in several samples, or mRNA copy number. Real-time PCR can also be applied to the detection and quantification of DNA in samples to determine the presence and abundance of a particular DNA sequence in these samples (Wikipedia, 2007).

Real-time PCR as a highly sensitive technique for the continuous on-line monitoring of PCR-amplified products has been developed for the quantification of nucleic acids in literature (Higuchi et al., 1993; Bassler et al., 1995; Morrison et al., 1998, 1999; Abe et al., 1999; Donovan et al., 2000). It is based on the detection and measurement in 'real-time' of fluorescence emitted proportionally to the synthesis of the PCR product. The fluorescent

signals required for detection can be obtained by labelling the PCR product with either a fluorescent dye (Morrison et al., 1998, 1999) or a fluorogenic probe (Bassler et al., 1995; Abe et al., 1999; Donovan et al., 2000). Fluorescent dyes (i.e. SYBR Green I) bind nonspecifically to any PCR product generated, while the fluorogenic probes are designed to release fluorescence after they are annealed to the specific target sequence. Thus, fluorogenic probes can be considered sequence-specific detection reagents.

The liver stages of malaria parasites have been traditionally studied using histopathological methods (Yoeli et al., 1965; Khan et al., 1992), which involve labour-intensive and time consuming procedures. The development of methods such as reverse transcription-PCR (Briones et al., 1996; Vernick et al., 1996), which can detect malaria-specific nucleotide sequences, partially overcame these limitations.

Orlandi-Pradines et al. (2006) noted that the CSP antigen is actively expressed only in the sporozoite stage and is generally used as a reference for estimation of immunologic exposure to malaria transmission (Druilhe et al., 1986). The pre-erythrocytic antigens tested also included liver stage antigen 1 (LSA1), which is expressed only in the hepatic stage, sporozoite threonine- and asparagines rich protein (STARP), and sporozoite and liver-stage antigen (SALSA), which are expressed both at the sporozoite and hepatic stages (Druilhe and Fidock, 1998).

3.3.2 PCR TECHNOLOGY IN MALARIA TREATMENT DISCOVERY

In reviewing what the PCR technology has delivered towards the treatment discovery of malaria, we discussed some PCR-based works on malaria parasite.

Human *Plasmodium falciparum* malaria parasite SALSA is a small protein localized on the sporozoite membrane surface and that it continues to be expressed during the liver stage. Bottius et al.(1996), using the PCR DNA amplification technique for the gene encoding SALSA in seven culture-adapted strains (five Asiatic and two African strains) and 16 isolates from Senegal, suggested that the SALSA protein be completely conserved among

P. falciparum isolates and pointed out the scarce homology (<30%) between SALSA and other *P. falciparum* antigens.

Nardin *et al*. (1999) investigations underscore the importance of the liver stage as a target for vaccine development, as the inhibition of parasite growth in hepatocytes can result in the reduction or complete ablation of the erythrocytic stages, thus attenuating or eliminating the symptoms and the pathology of the disease.

In an attempt to stimulate studies aimed at evaluating very precisely the efficacy of anti-malarial drug treatments and vaccination regimens, Brun˜a-Romero et al., (2001) applied a real time PCR-based assay that detects and measures parasite loads in the livers of mice exposed to the bite of a single malaria-infected Anopheles mosquito.

Fallon *et al.*, (2003) described a polymerase chain reaction (PCR) assay that detects avian malarial infection across divergent host species and parasite lineages representing both *Plasmodium spp.* and *Haemoproteus spp.* The assay is based on nucleotide primers designed to amplify a 286-bp fragment of ribosomal RNA (rRNA) coding sequence within the 6-kb mitochondrial DNA malaria genome. They claimed that the rRNA malarial assay outperformed other published PCR diagnostic methods for detecting avian infections. Fallon et al., (2003) noted that the development of molecular technology has made screening for these parasites faster and more reliable. Interest in developing a single screening assay that accurately detects diverse strains of nonhuman malaria, including infections missed by microscopic examination, has resulted in a number of polymerase chain reaction (PCR) assays (Feldman et al., 1995; Li et al., 1995; Perkins et al., 1998; Bensch et al., 2000; Richard et al., 2002) and a serological technique (Atkinson et al., 2001).

Considering the study of the malaria parasite biology and treatment discovery, the issue of how malaria parasites exit their host cells after completion of reproduction remains largely unsolved. Aly and Matuschewski (2005) attempted to validate a vital function of malaria cysteine proteases in active parasite egress, using a target gene that can be analyzed

functionally by real time PCR. They described a complete arrest of *Plasmodium* sporozoite egress from *Anopheles* midgut oocysts by targeted disruption of a stage-specific cysteine protease. Their findings show that sporozoites exit oocysts by parasite-dependent proteolysis rather than by passive oocyst rupture arising from parasite growth. They stated that malaria cysteine proteases are necessary for egress of invasive stages from their intracellular compartment and propose that similar cysteine protease–dependent mechanisms occur during egress from liver-stage and blood-stage schizonts.

Oyedeji *et al* (2007) used patients blood samples in a recent PCR (Polymerase Chain Reaction) malaria diagnostics study conducted on 401 children that complained of fever in Lafia, north-central Nigeria. They reported that 285 patients out of these 401 were infected with malaria. They conducted PCR on *stevor*, *SSUrRNA* and *MSA2* genes for comparison and assessment of PCR-based detection of *P. falciparum* in human blood sample. It was reported that *stevor* gene amplification has the highest sensitivity. Hence the most suitable for the parasite detection

PCR-based methods have been consistently shown to be a powerful tool for malaria diagnosis (Coleman et al., 2006; Berry et al., 2005; Cox-Singh et al., 1997; Di Santi et al., 2004).

3.3.3 DRAWBACKS OF PCR TECHNOLOGY

Despite having higher sensitivity and specificity in detecting *Plasmodium* infections, the use of PCR-based techniques in routine diagnosis is limited because of its logistical and technical difficulties (Hanscheid and Grobusch, 2002). The following are the drawbacks of evolving PCR technology for malaria treatment discovery and diagnosis:

(1) In malaria parasites, most genes like cytochrome b gene has a high AT content (approximately 73%), making it difficult to design effective primers. In addition, designing 'universal' primer assays is complicated by sequence variation in *Plasmodium spp.* (Bensch et al., 2000; Richard et al., 2002; Ricklefs and Fallon, 2002).

(2) PCR-based techniques for routine diagnosis are labour intensive and costly to maintain and this is in agreement with Oyedeji et al., (2007).

(3) Despite the fact that PCR-based assays have better sensitivity than conventional microscopy and antigen-based diagnostic tests (Tham et al., 1999), observations from the study of Oyedeji et al. (2007) showed that the level of sensitivity for PCR could vary depending on the approach employed (e.g. in terms of protocol) and the characteristic of the target sequence of the chosen assay. There is no current standard set for PCR-based malaria diagnosis. Hence for all amplification techniques, it is not known if the sensitivity of PCR is high enough to justify their use as a reference or standard in the diagnosis of *P. falciparum* infection.

As the scale of PCR technological studies grows, PCR diagnosis of malaria will play an increasing role in epidemiology with the development of high-throughput techniques that could facilitate a large-scale analysis of samples within a short period.

CHAPTER FOUR

LITERATURE REVIEW 3: THE CLUSTERING TECHNIQUES: EXISTING METHODS, APPLICATIONS AND DRAWBACKS

4.1 DEFINITION, HISTORY AND APPLICATIONS OF CLUSTERING

Cluster analysis is to discover the natural grouping(s) of a set of patterns, points, or objects. “Cluster analysis” first appeared as a phrase in 1954 and was suggested as a tool used to understand anthropological data (Clements, 1954). Biologists called it “numerical taxonomy” owing to the early research done on hierarchical clustering, a technique that aided them to create hierarchy of different species for analyzing their relationship systematically and understanding their phylogeny.

Cluster analysis is described in Webster dictionary as a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics. Single-link clustering (Sneath, 1957), Complete-link clustering and Average-link clustering (Sokal and Michener, 1958) first appeared in 1957, 1948, and 1958 respectively. The most popular partitional clustering algorithm, k-means, has been proposed several times in the literature: Lloyd in 1957 (Lloyd, 1957), and MacQueen in 1967 (MacQueen, 1967).

Cluster analysis finds its need in any discipline that involves analysis of multivariate data. Although not exhaustive, some important applications of clustering can be enumerated as:

4.1.1 IMAGE SEGMENTATION

This is an important problem in computer vision and can be formulated as a clustering problem (Jain and Flynn, 1996). Image segmentation for computer vision is one of the most important issues involved in building intelligent, autonomous systems whose major contribution is in the area of image understanding. In order to understand an image, the first thing a computer must do is to segment the image into several parts. In a satellite image, we may want to divide the image automatically into buildings, water, forest, and

agriculture. Data clustering plays an integral role in image segmentation algorithms (Hamerly, 2003).

4.1.2 DATA COMPRESSION

Clustering of data ensures that each cluster has some set of objects that belong to that cluster. In that sense, we may wish to represent the set of those objects with a form of description that easily replaces the set of objects in the cluster. If we do not need to store all of the objects, this offers an opportunity not to store all the objects but instead replace it with sets of description such as number of objects and boundary of the set of objects, thereby ensuring data compression.

4.1.3 REDUCTION OF SEARCH SPACE FOR FAST DATA ACCESS

A common type of query in databases is searching for the database object nearest to some query object. In databases, we search for desired objects and results using suitable queries to obtain them in form of output. Fast search and access is implemented when we cluster the data in the database in such a way that related data belong to same cluster. If we cluster the data in the database before any query execution, then we can do a two-level search which can be faster; searching first, the nearest cluster, and then doing a local search for the nearest object in that cluster. Documents can be clustered to generate topical hierarchies for information access or retrieval (Bhatia and Deogun, 1998).

4.1.4 FUNCTIONAL GENOMICS ANALYSIS

Clustering algorithm is also applicable in the study of genome data (Baldi and Hatfield, 2002). It can be used to find complex relationships within populations of gene expression values from DNA microarray data. We apply, for example, k-means algorithm numerically to cluster the gene expression values, because gene in the same cluster provides clues that they are performing a similar function. The concept of co-regulation and co-expression of genes in functional genomics is an important feature which k-means clustering tool can provide information to assist researchers in doing data analysis and data interpretation.

4.2 OVERVIEW OF CLUSTERING ALGORITHMS

A clustering algorithm is either hierarchical or partitional as shown in Figure 4.1. Hierarchical algorithms create successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once. For the hierarchical variants, we have the agglomerative and divisive clustering. However, in partitional clustering, we have QT-Clustering (Heyer et al., 1999), Self Organising Map (SOM) (Tamayo et al, 1999) and Traditional k-means which have evolved in recent years for high level analysis. A number of k-means variants algorithms exist and some of them include the Fuzzy C-means (Dembele and Kastner, 2003), X-means (Pelleg and Moore, 2000), G-means (Hamerly and Elkan, 2003), and PG means (Feng and Hamerly, 2006)

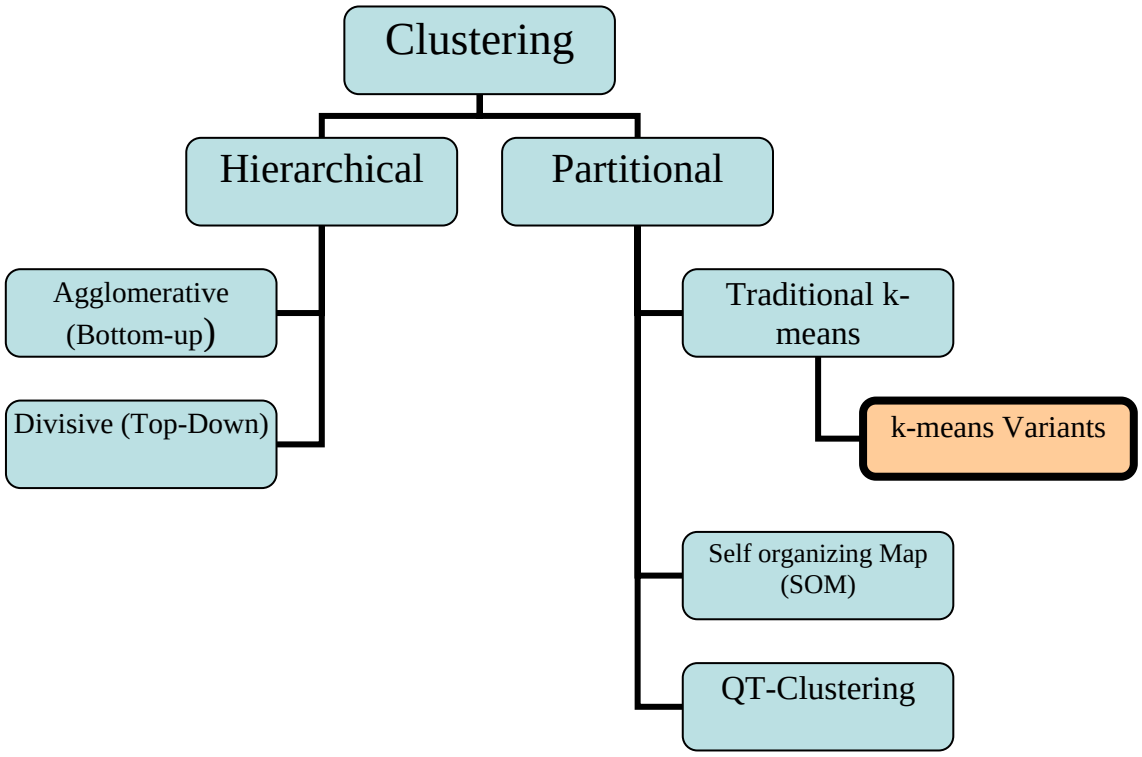


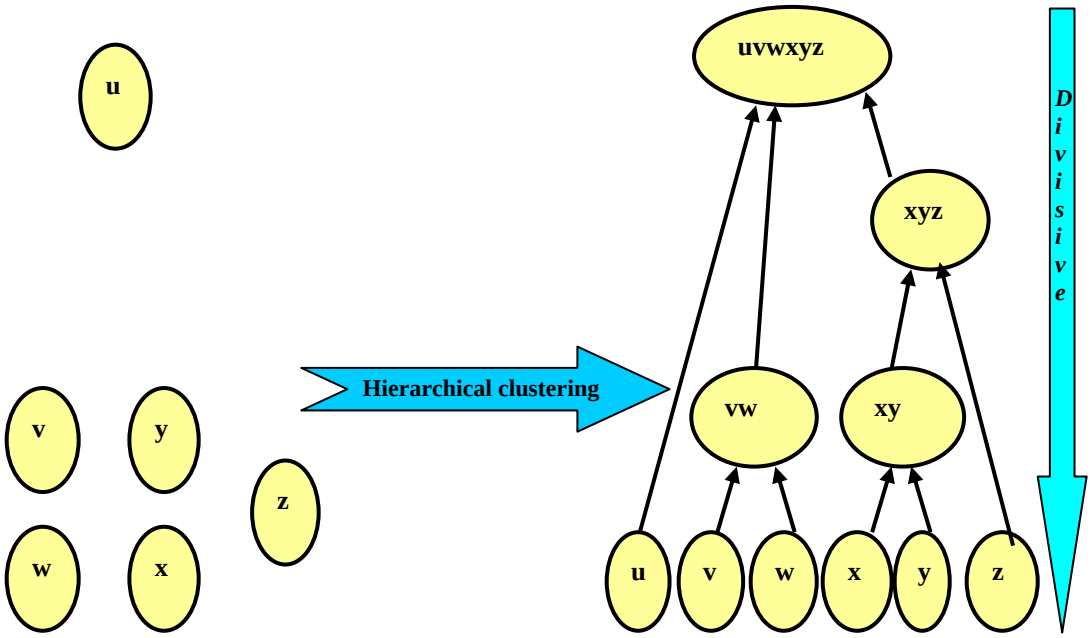
Figure 4.1: Evolving Clustering Algorithms and k-means variants.

These are the major classes of clustering with partitional clustering giving rise to traditional k-means. Several k-means variants have emerged for different applications using various models.

4.2.1 HIERARCHICAL CLUSTERING

Hierarchical algorithms can be agglomerative (bottom-up) or divisive (top-down). A hierarchical clustering algorithm becomes agglomerative (Härdle and Simar, 1998), if it starts from isolated data patterns and coalesces the nearest pattern or groups as specified by the threshold, in bottom-up fashion, forming hierarchies whereas divisive starts from whole set of patterns that divides along the top-down direction into successive smaller clusters (see Figure 4.2).

Using six raw data elements {u}, {v}, {w}, {x}, {y} and {z} as in figure 4.2 below with the position or location of each individual data reflecting the distances between them, which ultimately determine the data to merge into a cluster. Usually, we want to take the two closest elements, therefore we must define a distance d (element1, element2) between a pair of elements.



(Source: Wikipedia, 2006)

Figure 4.2: Agglomerative and Divisive Clustering.

Agglomerative is usually a bottom up approach while Divisive clustering is a top-down approach.

One can also construct a distance matrix showing the closeness of individual data points to each other. Merging two closest raw data points v and w result in the following clusters {v

w}, {u}, {x}, {y} and {z}. In order to merge further, we need to compute the distance between {u} and {v w} hence the need to define the distance between two clusters. {x} merges with {y} to form {x y} which further merges to {z}. Finally, clusters {u}, {v w}, {x y z} merge into a single whole cluster {uvwxyz}. Computing distances between clusters can be done using single linkage, complete linkage or average linkage. Distance *d(x,y)* can be computed by using a distance metric measure like the Euclidean distance. However, other distance metric measures include the Pearson Correlation, Mahalanobis (City block) distance and Chebychev, which is like City Block, which instead of summing the differences, takes the maximum. Correlation coefficient of 1 means perfectly correlated (giving zero distance), a correlation coefficient of 0 means uncorrelated (giving unit distance), and a correlation coefficient of –1 means oppositely correlated (Biodiscovery, 2001).

Several problems are shared by these hierarchical methods. Decisions to join two elements are based only on the distance between those elements, and once elements are joined they cannot be separated. This is a local decision-making scheme that does not consider the data as a whole, and it may lead to mistakes in the overall clustering. In addition, for large data sets, the hierarchical tree is extremely complex, and the choice of location for cutting the tree is unclear (Heyer et al., 1999).

4.2.2 PARTITIONAL CLUSTERING AND TRADITIONAL K-MEANS

In partitional clustering, there are distinct data groups such that each data belongs to a group called partition. One good example is the traditional k-means algorithm and some of its variants. The word "*k-means*" indicates that the algorithm takes as an input a user predefined number of clusters, which is the k from its name, while *means* stands for an average representing the average location of all the members of a particular cluster.

k-means algorithm is a simple, iterative procedure, in which each cluster has only one *centroid* which moves based on the computed means of data belonging to that cluster. *Centroid* is an artificial point in the space of data which represents an average location of the particular cluster. The coordinates of this centroid point are averages of attribute

values of all examples that belong to the cluster. We define k-means clustering as an algorithm used to group objects or data points into user-defined number of classes called clusters based on certain attributes, whereby each data point presented is allocated to a cluster whose centroid maintains the shortest distance to that data point.

k-means clustering (MacQueen, 1967) is the most common partitioning algorithm. The goal and objective function of k-means algorithm is to minimize dissimilarity in the elements within each cluster, while maximizing this value between elements in different clusters. A simplified representation of k-means algorithm as adapted from Teknomo (undated) and Sammy (undated) is given in Figure 4.3.

Step 1: Select k initial cluster centroids, c_1, c_2, c_3, ..., c_k. One way to do this is
either:
Take first k instances or
Random sampling of k elements or
Take any random partition in k clusters and computing their centroids

Step 2: Assign each instance x in the input data to the cluster whose centroid is the nearest to x.

Step 3: For each cluster, re-compute its centroid based on which elements are contained within it.

Step 4: Go to (2) until convergence is achieved, i.e. until a pass through the input data causes no new assignments.

(Source: Adapted from Teknomo (2006); Sammy (2006))

Figure 4.3: A Simplified Representation of Traditional k-means Algorithm.

There are three possible ways to select your initial centroids: first k instances, random sampling or random partition. Data are assigned to their nearest centroid and new k centroids are computed for the next iteration iteratively till convergence is reached.

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 m_j, j=1,2,3,….k, in R^d, called *centers*, to minimize the mean squared distance from each data point to its nearest center (Kanungo et al, 2004). The objective function is:

$$1/n \sum_{i=1}^{n} \left[\min_{j} d^2(x_i,m_j) \right] \qquad (2.1)$$

Where $d^2(x_i,m_j)$ denotes the metric used (Euclidean distance) for example, distance between x_i and m_j for j=1,2, 3,…,k.

The problem in Eq.(2.1) is to find k cluster centroids, such that the average squared Euclidean distance (MSE) between a data point and its nearest cluster centroid is minimized. The approximate solution to Eq. (2.1) is easily implemented by k-means.

The k-means algorithm is popular and easy to implement, scalable with speed of convergence always to a local minimum as the global minimum is NP-complete. The particular local minimum found depends on the starting cluster centroids. The algorithm updates cluster centroid iteratively to decrease the objective function in Eq. (2.1) till local minimum is found. Its computational complexity is O(nk*l*), where n= total number of dataset objects, k=cluster number and *l*= number of iterations.

4.3 EXISTING K-MEANS CLUSTERING ALGORITHM AND RELATED WORKS

The traditional k-means has evolved over time as many algorithm designers employ clever ways of redesigning k-means to improve its efficiency.

4.3.1 FUZZY C-MEANS

This is also referred to fuzzy k-means. k-means clustering algorithm operates with a concept referred to as *hard partitioning* which implies that each data point is assigned to one and only one cluster. This is synonymous with what we have in a classical set where an element is either a member or a non-member of a set and as a result implements the

principle of *no sitting on the fence*. However, in truth, points on the edge of a cluster or near another cluster may not be as much inside that cluster as points in the cluster center, hence the need for fuzzy clustering. Fuzzy clustering applies the technique of fuzzy logic to ensure that each data point does not belong to only one cluster but has a degree to which it also belongs to other clusters. This is called *degree of membership.* The Fuzzy c-means is implemented using MATLAB and visual C++. Dembele and Kastner (2003) in their work focused on the method of choosing appropriate fuzzy parameter *m* for microarray data clustering, since the fixed value of *m=2* failed to give a good result. Dembele and Kastner (2003), proposed an empirical method, based on the distribution of distances between genes in a given data set to determine an adequate value for *m*.

4.3.2 K-MEDOIDS ALGORITHM

This iterative algorithm is similar in approach to k-means, but it imposes an additional constraint: that the centers that are used to represent the data are taken from the dataset itself. Thus a "medoid" is a datapoint that best represents a set of data. Because of this constraint, k-medoids can operate on data which do not live in a metric space, as long as the data can be described in terms of pairwise distances for the datapoints being clustered. However, the initial construction of the pairwise distance matrix D requires time $O(dn^2)$ and the search for a new medoid (each iteration) takes in expected time $O(n^2)$. Therefore, it is not a linear-time algorithm for clustering, and is of little practical interest for moderately-sized and large datasets (Hamerly, 2003).

4.3.3 DENSITY OF POINTS CLUSTERING (DPC)

DPC by Wicker et al. (2002) divide cluster and test whether it should be divided or not. If it is not divided, then there will be only one cluster in the data set, otherwise there are at least two clusters that will be further iteratively divided if necessary. The division is attempted through the k-means method with k = 2. The test is based on point density measures. When the density measure between two possible clusters is too small compared to the density measure inside both clusters, the two clusters are kept because they are not well connected to each other.

4.3.4 X-MEANS

Pelleg and Moore (2000) proposed a scheme for learning k, which they call X-means. The algorithm searches over many values of k and scores each clustering model using the so-called Bayesian Information Criterion. The user only specifies by guessing a range in which the true k will lie and X-means chooses the model with the best BIC score on the data. X-means improves the speed of a naïve k-means by embedding the dataset in a multiresolution kd-tree, storing sufficient statistics at the nodes. It also uses a fast algorithm that allows additional geometric computation, blacklisting, that maintains a list of just those centroids that need to be considered for a given region.

4.3.5 OVERLAPPED AND ENHANCED K-MEANS

Fahim et al. (2006) proposed what they called Overlapped and Enhanced k-means algorithm which uses a simple data structure to keep some information on each iteration for usage in the next iteration. For each data point, its distance to the centroid of its nearest member cluster is stored for that iteration so that there will be no need for its distance computation if it is near its centroid at the next iteration. The scheme can improve the computational speed of k-means algorithm by the magnitude in the total number of distance calculations and the overall computational time. Their k-means implementation is based on two functions called *distance()* and *distance_new()* used for each algorithm as follows:

(1) Overlapped k-means: Obtained when these two functions are executed a number of times, one after the other, starting with *distance()* as first and followed by *distance_new()* alternatively so that there is overlap between these two functions.

(2) Enhanced k-means: This is obtained by executing the function *distance()* twice while function *distance_new()* is executed for the reminder of the iterations.

The reader is referred to chapter 5 for more details, but briefly, the function *distance ()* is similar to basic function of k-means algorithm but it has an additional feature of having a simple data structure to keep the distance between each point and its nearest cluster. The

distance new () encapsulates the decision for a point to either stay in its old cluster assigned to it in the previous iteration or be reassigned to a new cluster if the computed distance is larger than the distance to the old centre.

4.4 DRAWBACKS OF EXISTING K-MEANS METHODS

Generally, k-means is known to converge at local optima as noted in Steinley (2003). However, the drawbacks of k-means include the following:

1) Most k-means clustering tool performs poorly on large datasets, hence there is the need to cluster enormous amount of genomic data at a reasonable time shorter than the runtime of existing algorithms.

2) To the best of our knowledge, existing methods are not well equipped for analysing and extraction of useful knowledge from the vast amount of Malaria Microarray Data (MMD) for the purpose of finding a functional relationship of genes involved in malaria infection or understanding the complex biology of the parasite.

3) Many clustering tools are developed based on some specific clustering goals in the mind of the developer at the time of their development, hence, different models and statistical assumptions were made. It means that we may not be very sure of the output obtained from an engineering-based clustering tool versus a social science based clustering tool or transcriptome-based clustering tool as they may give completely different results on the same data.

4) Many of the clustering tools such as enhanced k-means lack effective evaluation of their cluster quality (Fahim et al., 2006).

CHAPTER FIVE

REDUCING THE TIME REQUIREMENT OF K-MEANS ALGORITHM

5.1 INTRODUCTION

Clustering is the unsupervised grouping of objects into classes without any *a priori* knowledge of the datasets to be analyzed. In this case, there is no supervisor to teach the systems first on how to classify the known sets of data points. Given *X* n dataset points, $x_1, x_2, x_3, \ldots, x_n$, contained in d-dimensional space R^d, the process of clustering can be formally stated as: to seek partitions $X_1, X_2, X_3, \ldots, X_k$ such that every x_i , $i = 1,2,3,\ldots,n$, falls into one of these regions and no x_i falls into two regions. Partitions $X_1, X_2, X_3, \ldots, X_k$ satisfy the following: $X_1 \cup X_2 \cup X_3 \ldots \cup X_k = X$ and $X_i \cap X_j = 0 \quad \forall i \neq j$, where $\cup$ and $\cap$ stand for union and intersection respectively. The definition further stated that we are to cluster, or form into each class, data points x_i that are as similar as possible, hence we need what is called a similarity measure (or dissimilarity measure) usually given in a numerical form to serve as an indicator of the degree of resemblance or natural association between a data and groups of data (Bow, 1984). The dissimilarity measure ρ (as used in k-means) is expected to satisfy the following: $\rho(x_i,x_i) = 0$ and $\rho(x_i,x_j) \neq 0 \quad \forall i \neq j$.

We define clustering from genomics point of view as a data analysis tool that puts genes into groups called clusters so that the degree of association is strong between gene members of the same cluster and weak between gene members of different clusters. Hierarchical algorithms create successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once. For the hierarchical variants, we have the agglomerative and divisive clustering. However, in partitional clustering, we have QT (Quality Threshold) clustering (Heyer et al, 1999), Self Organising Map (SOM) (Tamayo et al., 1999) and Standard k-means **(**MacQueen, 1967) which has been evolving in recent years for high dimensional data analysis.

Formally, 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 (Kanungo et al., 2004; Fahim et al., 2006).

The k-means method has been shown to be effective in producing good clustering results for many practical applications such as data compression and vector quantization (Duda and Hart, 1973) as well as pattern recognition and pattern classification (Gersho and Gray, 1992). It has also found application in data mining and knowledge discovery (Fayyad et al, 1996) image segmentation (Jain and Flynn, 1996; Hamerly and Elkan, 2003) and gene expression (Baldi and Hatfield, 2002). The traditional k-means algorithm requires in expectation, *O(nkl)* run time where *l* is the number of k-means iterations. This time was said to be reduced in (Fahim et al., 2006) to *O(nk)* but we found their algorithm to still run in *O(nkl)* expectation time. Note that this can still be computationally expensive for large datasets, such as the microarray data, where we have large datasets with large dimension size *d*.

For efficient and effective analysis of microarray data, we developed a novel Pearson correlation-based Metric Matrices k-means (MMk-means) with a better run-time $O(nk^2)$ than the Traditional k-means and other variants of k-means algorithm like Overlapped and Enhanced k-means algorithms developed in Fahim et al., (2006).

5.2 PREVIOUS VARIANTS OF THE ALGORITHM

The word “k-means” indicates that the algorithm takes as an input a user predefined number of clusters, which is the “k” from its name, while “means” stands for an average representing the average location of all the members of a particular cluster. There are two existing basic versions of k-means clustering, a non-adaptive version introduced by Lloyd (1957) and an adaptive version introduced by MacQueen (1967). A number of k-means variants algorithms exist and they include fuzzy c-means (Bezdek, 1981), X-means (Pelleg and Moore, 2000), G-means (Hamerly and Elkan, 2003), and PG means (Feng and Hamerly, 2006) and Fuzzy J-Means (Belacel et al., 2002; Belacel et al., 2004).

A generalized pseudocode of Traditional k-means algorithm is given in Figure 4.3. An overview of the Traditional k-means algorithm and its recent variants was presented by (Fahim et al., 2006). Figure 4.3 in a more explicit format is given in Figure 5.1, where MSE denotes the mean squared error.

```
//TRADITIONAL K-MEANS
1 MSE=largenumber;
2 Select initial cluster centroids mj // Randomly or first k genes;
3 Do
4       OldMSE=MSE;
5       MSE1=0;
6       For j=1 to k
7               mj=0; nj=0;
8       endfor
9       For i=1 to n
10              For j=1 to k
11                      Compute squared Euclidean  distance d^2(xi, mj);
12              endfor
13              Find the closest centroid mj to xi;
14              mj=mj+xi; nj=nj+1;
15              MSE1=MSE1+d^2 (xi, mj);
16      endfor
17      For j=1 to k
18              nj=max(nj, 1); mj=mj/nj;
19      endfor
20      MSE=MSE1;
21  while (MSE<OldMSE)
```

Figure 5.1: Pseudocode of Traditional k-means *(Fahim et al., 2006)*

Fahim *et al*. (2006) designed two new variants of k-means algorithms noting that if the distance between a data point and the current centroid (new center) of the cluster that it was assigned to in the previous iteration is less than or equal to the distance of the data point to its previous centroid (old centre), then the point remains in that cluster and there is no need to compute its distance to the other *k*-1 centers. To do this, they introduced two arrays, namely *Clusterid* and *Pointdis* to keep track of the centroid to which each point is assigned to and the distance between this point and its centroid. They used two sub-procedures in *Figure 5.2a* and *Figure 5.2b* to design two variants of k-means algorithm in *Fig5.3a* and *Fig5.3b* respectively. These algorithms are shown as

```
//Function distance()
//assign each point to its nearest cluster
1 For i=1 to n
2       For j=1 to k
3               Compute squared Euclidean distance d^2(x_i, m_j);
4       endfor
5       Find the closest centroid mj to x_i;
6       m_j=m_j+x_i; n_j=n_j+1;
7       MSE=MSE+d^2(x_i, m_j);
8       Clusterid[i]=number of the closest centroid;
9       Pointdis[i]=Euclidean distance to the closest centroid;
10 endfor
11 For j=1 to k
12      m_j=m_j/n_j;
13 endfor
```

Figure 5.2a: Pseudocode of Function distance( ) **(***Fahim et al., 2006)*

```
//Function distance_new()
//assign each point to its nearest cluster
1 For i=1 to n
        Compute squared Euclidean distance d^2(x_i, Clusterid[i]);
        If (d^2(x_i, Clusterid[i])<=Pointdis[i])
                Point stay in its cluster;
2       Else
3               For j=1 to k
4                       Compute squared Euclidean distance d^2(x_i, m_j);
5               endfor
6       Find the closest centroid m_j to x_i;
7       m_j=m_j+x_i; n_j=n_j+1;
8       MSE=MSE+d^2(x_i, m_j);
9       Clustered[i]=number of the closest centroid;
10      Pointdis[i]=Euclidean distance to the closest centroid;
11 endfor
12 For j=1 to k
13      m_j=m_j/n_j;
14 endfor
```

Figure 5.2b: Pseudocode of Function distance_new( ) *(Fahim et al., 2006)*

```
1 MSE=largenumber;
2 Select initial cluster centroids; // Randomly or //first k elements
3   Do
4           OldMSE=MSE;
5           MSE1=0;
6           For j=1 to k
7                   n_j=0;
8           endfor
9           distance()
10          distance_new()
11          MSE=MSE1;
12  while (MSE<OldMSE
```

Figure 5.3a: Pseudocode of Overlapped k-means

```
1MSE=largenumber;
2 iteration=0
3 Select initial cluster centroids; // Randomly or  //first k elements
4       Do
5               iteration+=1
6               OldMSE=MSE;
7               MSE1=0;
8               For j=1 to k
9                       n_j=0;
10              endfor
11              if(iteration≤2)
12              distance();
13              else
14                      distance_new();
15              MSE=MSE1;
16    while (MSE<OldMSE)
```

Figure 5.3b: Pseudocode of Enhanced k-means

Fahim *et al*., used $n\sum_{i=1}^{l} 1/i$ to estimate the total number of data points for each iteration that moved from their clusters during the number of k-means iterations, *l*. They showed that the cost of using an enhanced k-means algorithm is approximately *O(nk)*. We observed that the total number of data points for each iteration that moved from its clusters during the k-means iterations is not strictly monotonically decreasing and thus the Overlapped and Enhanced k-means algorithm eventually still costs *O(nkl)* run time in expectation.

5.3 METHODOLOGY

5.3.1 BASIC DEFINITIONS

For simplicity of our algorithm, the following basic definitions were adduced:

(a) Notation

||.|| denotes the Euclidean norm of a vector. The trace of a matrix *X, i.e.*, the sum of its diagonal elements, is denoted as trace (A). The Frobenius norm of a matrix $||X||_F = \sqrt{trace(X^T X)}$. I_n denotes identity matrix of order n.

(b) Centroid: Centriod, sometimes called centre, is an artificial point in the space of data and it represents an average location of a particular cluster. The coordinate of this centroid point is an average of attribute values of all data that belong to the cluster.

(c) Correlation (r): A measure of relationship between two or more sets of data. Correlation is of two types namely Spearman rank correlation coefficient that uses ordinal values (ranked) while Pearson product moment correlation coefficient uses interval or ration value. Types of correlation relationship can be positive ($0< r \leq 1$) , negative ($0 > r \geq -1$) or no correlation $r = 0$.

(d) Metric Matrix: A kxk matrix corresponding to values of the correlation coefficient (r) between the centroids of the previous (pm_j) and current (m_j) iterations respectively, where $0 < j \leq k$.

(e) Ding-He Threshold: It is an interval determined by Ding and He (2004), used in our new k-means algorithm to determine whether a cluster must remain without further clustering or be subjected to further clustering.

(f) Minimum MMk-means Iteration (MMI): The minimum number of k-means iteration times required before the Ding-He interval is applied in our new k-means algorithm.

(g) $diff_j$: An absolute value obtained from the subtraction of the current iteration eigenvalues (e_j) from the previous iteration eigenvalues (pe_j) and it serves as an indicator to terminate clustering for each cluster. Each eigenvalues set is obtained from the corresponding Metric Matrix.

(h) Some Set Notations:

set[j]: $1 \leq j \leq k$ is the set referring to cluster *j*.

add[*i*]: Is a function to add data point into a cluster, where *i* is the index of a data point.

set[j].n_j : Is the size of cluster *j*, that is number of data points in a cluster *j*.

5.3.2 ALGORITHM DESIGN FOR OUR NEW MMK-MEANS

Our MMk-means algorithm (see *Figure 3.4)* runs like the traditional k-means algorithm except that it is equipped with a mechanism to determine when a cluster is stable, that is, its membership data points will always remain in the same cluster in each subsequent iteration. This is an improvement on the Overlapped and Enhanced variants of k-means algorithms introduced by Fahim et al., (2006). They equipped their algorithms with the ability to detect the stability of a data point but MMk-means is equipped with the mechanism to detect the stability of a cluster representing a whole bunch of data points.

To do this, we use a simple data structure to indicate when a data point belongs to a stable cluster. We use the recently established relationship between principal component analysis and k-means clustering to design a mechanism for determining when the whole data points in a cluster are stable. We create a covariance matrix (r), computing the Pearson product moment correlation coefficient between the k centroids of the previous and current iterations and then deduce k previous and current iterations eigenvalues. The difference of these eigenvalues for each cluster is computed and checked to see if it satisfies (that is, lies within) the Ding-He interval. If it does, the corresponding cluster is considered stable and there is no need to compute its data point distances with the current centroid of the cluster or the rest k-1 centroids.

The mechanism so explained is prescribed in the subprocedure Compute_MM of *Figure 5.5* and this function is being executed when the current total iterations number is greater than MMI-1.

```
1   m=Compute_multiplier(k, d, X)// or Compute_factor(k, d, X)
2   compute_eigenvalues=false; check_stability=false
3   iteration = 0
4   MSE=large number;
5    Select initial cluster centroids  //  Randomly or first k genes
6   Initialise
7   adj_x[i].bool=0
8   adj_x[i].dist=0
9   adj_x[i].j=0
10  Do
11              iteration += 1
12              OldMSE=MSE;
13              MSE=0;
14              For j=1 to k
15                              pm_j=m_j; n_j=0;
16              endfor
17              For i=1 to n
18                              if(adj_x[i].bool == 'F' {       // 'F' = False
19                                      For j=1 to k
20                                              Compute distance d^2(x_i, m_j);
21                                      endfor
22                                      Find the closest centroid m_j to x_i; (Store in dist=
                                        d^2(x_i, m_j))
23                                      adj_x[i].dist= dist;
24                                      adj_x[i].j=j
25                                      set[j]= add[i]
26                                      m_j=m_j+x_i; n_j= n_j+1;
27                              } else {    // point stays in its cluster
28                                      dist = adj_x[i].dist
29                              }
30                               MSE1=MSE1+dist
31              endfor
32              For j=1 to k
33                      n_j=max(n_j, 1); m_j=m_j/n_j; // n_j can only be max between 0 and 1
34              endfor
35              if (iteration>1) {
36                      if (compute_eigenvalues==true)
37                              check_stability=true
38                      if((1-mMSE_l/MSE)*100≤0.7)
39                              compute_eigenvalues=true
40              }
41              Compute_MM(pm_j, m_j, iteration)
42      while (MSE<OldMSE)
```

Figure 5.4: Pseudocode of Our Main Program for MMk-means

```
1  Compute_MM(pm_j, m_j, iteration){
2        if (compute_eigenvalues==true){ //if (iteration > MMI-1) is implemented here
3              Compute r using pm_j  and m_j
4              Compute its eigenvalue into e_j
5            if (check_stability==true){ //if(iteration > MMI) is also implemented here
6                  For j=1 to k {
7                        diff_j= | pe_j-e_j|
8                        if (diff_j <Ding-He H_1 && diff_j > Ding-He L_0) {
9                              For(i=1 to set[ j ].n_j)
10                             adj_x[set[ j ][i]].bool= 1
11                       }
12                 }
13           }
14           pe_j = e_j
15       }
16 }
```

Figure 5.5: Pseudocode of our Compute MM Sub-program for MMk-means

For any *n* dataset points, given the total number of k-means iteration *l* required, we can actually set MMI = *l*/2, but note that *l* is unknown until a traditional k-means algorithm is executed. We know that for a given clustering procedure, k-means algorithm aims at minimizing the first Mean Squared Error (MSE_1), through a number of iterations, *l*, distributing all data points into clusters, to arrive at an optimal (minimized) Mean Squared Error (MSE_l). Therefore, we estimate the required Minimum MMk-means Iteration (*MMI*) to be bounded by $0 < MMI \leq MSE_1.k/MSE_l$. For a given set of *n* data points, the first iteration of a traditional k-means algorithm can be used to determine MSE_1 in *O(nk)* time. For the given n dataset points, we can form the *d*-by-*n* matrix $X = [x_1,\ldots, x_n]$. Centring each data point around the origin, such that $y_i = x_i - \bar{x}$ and $\bar{x} = \sum_i x_i/n$,

Ding and He (2004) showed that:

Theorem 1.

The optimal Mean Squared Error (MSE_l) is tightly bounded from below and above by

$$n\overline{y^2} - \sum_{i=1}^{k-1} \lambda_i < MSE_l < n\overline{y^2}, \qquad (1)$$

where $\overline{y^2} = \sum_i \mathbf{y}_i^{\mathrm{T}} \mathbf{y}_i / n$ *and* λ_i *are the eigenvalues of the covariance matrix*

$YY^T = \sum_i (x_i - \bar{x})(x_i - \bar{x})^T$.

Therefore, using equation (1) above, we can compute MSE_l in *O(n)* time. Empirical testing followed by personal communication (Ding, 2008) shows that equation (1) above does not hold for large k and data with high dimensional (*d*). So equation (1) will not estimate MSE_l for all k and *d* as we desire in our new k-means algorithm.

It is also obtained in Zha et al. (2002) that:

Theorem 2.

The optimal Mean Squared Error (MSE_l*) is bounded from below by*

$MSE_l \geq trace(X^T X) - \max_{A^T A = I_k} trace(A^T X^T X A) = \sum_{i=k+1}^{d} \sigma_i^2(X)$,

where $\sigma_i(X)$ *is the i largest singular value of X and A is an arbitrary orthonormal*

matrix.

Ding and He (2004) indicated that the lower bound in theorem 2 is not asymptotically tight as in theorem 1. From theorem 2 above, we observed that although the equation does not correspondingly estimate MSE_l for large *k* and high *d*, it possesses a

distribution that mimics the series needed to estimate MSE_l for large *k* and high *d.*

Using this observation, we are able to estimate approximately a multiplier we called *m*, that is useful in the prediction of MSE_l from MSE_1 and consequently determine *MMI*.

Observation 1.

From equation (2), we can estimate $m = 1 - \{\sum_{i=k+1}^{d}(\sigma_i(X) - \sigma_{i+1}(X))\}/(d-k)\}$

and consequently find $MSE_l \cong m MSE_i$. Note that σ_i (X) is the *i* largest singular value of

X. We encapsulate the computation of the multiplier *m* in an implicit subprocedure Compute_multiplier.

Observation 2.

For each iteration, given $MSE_i \cong mMSE_1$*, we can determine MMI, by estimating the*

distance of the current iteration MSE away from the final and optimal MSE, MSE_i *and*

when

*(1-mMSE$_1$/current iteration MSE)*100% ≤0.7%, MMI is equal to the current total iterations number.*

5.3.3 ALGORITHM CORRECTNESS AND COMPLEXITY ANALYSIS

To prove the correctness of our new and novel k-means algorithm, we will need the following definitions and theorem from Kumar *et al.* (2004) and Fan (1949) respectively.

Definition 1. *Given a set of k points K, which we also denote as centers, define the k-means cost of X, set of n points in d-dimensional space R^d, with respect to K,* $\Delta(X, K)$*,*

as

$$\Delta(X, K) = \sum_{x \in X} (x, K)^2,$$

where d(x, K) denotes the distance between x and the closest point to x in K.

Definition 2. *For a set of points X, define the centroid, C(X), of X as the point*

$\sum_{x \in X} x/|X|$. *For any point* $a \in R^d$, *it follows that*

$$\Delta(X, a) = \Delta\big(X, c(x)\big) + |X|.\Delta(c(X), a).$$

An important result, that we shall see soon, how it relates centroid of each partition X_j to an eigenvalue, is given by Fan (1949) and stated as:

Theorem 3. *Let H be a symmetric matrix with eigenvalues* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, *and the corresponding eigenvector* $U = [u_1, \ldots, u_n]$. *Then*

$$\lambda_1 + \lambda_2 + \ldots + \lambda_k = \max_{A^T A = I_k} trace(A^T HA).$$

Moreover, the optimal A is given by* $A^* = [u_1 \ldots u_k]Q$ *with Q an arbitrary orthogonal matrix.*

Let the centroids at each k-means iteration be $m_1^i, m_2^i, \ldots, m_k^i$, $1 \leq i \leq l$, where l is the total number of k-means iterations. Now, we will also need the following lemma.

Lemma 1. *For a partition* X_j *at a t-iteration* $\leq l$, *let* $diff_j = |pe_j - e_j|$, *if Ding-He* $H_1 < diff_j <$ *Ding-He* L_0 *then* $\Delta\big(X_j^t, m_j^t\big) = \Delta\big(X_j^i, m_j^i\big)$ *for* $t < i \leq l$.

Proof. Note that r in sub-procedure Compute_MM of figure 7, which is the key mechanism we used to identify stable partitions, is the k x k correlation coefficient matrix generated between the centroids of the previous and current iterations of the k-means algorithm. Note further that r is a covariance matrix (Rodgers et al., 1988).

Zha *et al.* (2002) in their attempt to prove theorem 2 (stated above) showed that theoretically

$$MSE = \sum_{j=1}^{k} MSE_j = \sum_{j=1}^{k} \left(trace(X_j^T X_j) - \left(\frac{e^T}{\sqrt{n_j}}\right)(X_j^T X_j)\left(\frac{e}{\sqrt{n_j}}\right) \right)$$

$$= trace(X^T X) - trace(A^T X^T XA), \qquad (3)$$

where A is an n x k orthonormal matrix given by

$$A = \begin{matrix} n_1 \\ \vdots \\ n_k \end{matrix} \begin{pmatrix} \frac{e}{\sqrt{n_1}} & & \\ & \frac{e}{\sqrt{n_2}} & \\ & \vdots & \\ & & \frac{e}{\sqrt{n_k}} \end{pmatrix}, \qquad (4)$$

e is a vector of appropriate dimension with all elements equal to one and $n_j, 1 \leq j \leq k$

are numbers of data points in each cluster. Note that the minimization of equation (3) is equivalent to

$$max = \{ trace(A^T X^T XA) \mid A \; of \; the \; form \; of \; (4) \}.$$

It is also shown in Zha et al. (2002) that

$$trace(A^T X^T XA) = \sum_{j=1}^{k} n_j ||m_j||^2. \qquad (5)$$

Equations (3) and (5) relate minimized MSE to maximizing the sum of the centriods. Note also that theorem 3 above relates centroids of each partition to an eigenvalue. Iteratively, r in Compute_MM relates centroids of previous and current iterations respectively and therefore from equation (5) and theorem 3, its eigenvalues characterize the iterative minimized MSE of each partition and $diff_j$ is an estimate of how close is

the minimized MSE for a partition (in terms of its centroid) to the optimal one. Since Ding and He [2] had shown an upper and lower bound to expect this, then if Ding-He $\mathbf{H_1} < \mathbf{diff_j} <$ Ding-He $\mathbf{L_\infty}$, the centroid of the corresponding partition X_j virtually

does not change in subsequent iterations.

This translates to $\Delta(X_j^t, m_j^t) = \Delta(X_j^i, m_j^i)$ for $t < i \leq l$ from definition 2.

□

We now prove the correctness of MMk-means algorithm in the theorem that follows.

Theorem 4. *Given a point set X, MMk-means returns a k-means solution on input X.*

Proof. We should note that our algorithm maintains the following loop invariant:

Invariant: Let $M = \{m_1^i, \ldots, m_k^i\}, for\ i = 0, \ldots, l-1\ and\ \forall\ 1 \leq j \leq k$

(1) $\Delta(X_j^{i+1}, m_j^{i+1}) \leq \Delta(X_j^i, m_j^i)$.

(2) The set X is a subset of $X_1 \cup X_2 \cup \ldots \cup X_k$.

It is straight forward to note that for *i=0*, the invariant holds. Now, let's assume that the invariant holds for some fixed *i=p*, it remains to show that the invariant holds for *i=p+1* as well, then we are done.

Based on our assumption, for *i=p*,

$\Delta(X_j^{p+1}, m_j^{p+1}) \leq \Delta(X_j^p, m_j^p)\ \forall\ j$ (6)

For *i=p+1*, we have to show for every *j* that it is either

$\Delta(X_j^{p+2}, m_j^{p+2}) = \Delta(X_j^{p+1}, m_j^{p+1})$ (7) or

$\Delta(X_j^{p+2}, m_j^{p+2}) < \Delta(X_j^{p+1}, m_j^{p+1})$ (8)

Note that for a partition X_j, if $\Delta(X_j^{p+1}, m_j^{p+1}) = \Delta(X_j^p, m_j^p)$ from (1) above then using lemma 1, $\Delta(X_j^{p+2}, m_j^{p+2}) = \Delta(X_j^{p+1}, m_j^{p+1})$ for all iterations later on. This proves (7) above.

Now if $\Delta(X_j^{p+1}, m_j^{p+1}) < \Delta(X_j^p, m_j^p)$ from (6), it remains to prove that

i) $\Delta(X_j^{p+2}, m_j^{p+2}) < \Delta(X_j^{p+1}, m_j^{p+1})$ or

ii) $\Delta(X_j^{p+2}, m_j^{p+2}) = \Delta(X_j^{p+1}, m_j^{p+1})$.

Lemma 1 indicates the condition to expect ii), so we are done as regards this. If this condition is not valid for a particular partition X_j then $m_j^{p+2} \neq m_j^{p+1}$. From definition 1,

$$\Delta(X_j^{p+2}, m_j^{p+2}) = \Delta(X_j^{p+2}, a) - |X_j^{p+2}|.\Delta(m_j^{p+2}, a) \quad (9) \text{ and}$$

$$\Delta(X_j^{p+1}, m_j^{p+1}) = \Delta(X_j^{p+1}, a) - |X_j^{p+1}|.\Delta(m_j^{p+1}, a) \quad (10)$$

for any point $a \in R^d$. Note that for our MMk-means algorithm and infact any other k-means algorithm, if $m_j^{p+2} \neq m_j^{p+1}$, then $\Delta(m_j^{p+2}, a) < \Delta(m_j^{p+1}, a)$ and therefore from equations (9) and (10), $\Delta(X_j^{p+2}, m_j^{p+2}) < \Delta(X_j^{p+1}, m_j^{p+1})$. This completes the proof for (8) above.

It now remains to show that for each iteration *i* in our MMk-means algorithm, the input set *X* is a subset of $X_1^i \cup X_2^i \cup \ldots \cup X_k^i$. This is actually straight forward. Note that for an iterations *i*, $x \in X$, there exists only one closest m_j^i centroid to *x* from *M* for a particular partition X_j^i, such that $x \in X_j^i$. This shows that all $x \in X$ belongs to a partition X_j^i for $1 \leq j \leq k$and therefore *X* is a subset of $X_1 \cup X_2 \cup \ldots \cup X_k$. □

Theorem 5. *Our new and novel MMk-means algorithm runs in O(nk^2) expected time.*

Proof.

Using the devices enumerated under the algorithm design section above, our new k-means algorithm is presented in *Figures 6 and 7*. Based on the value of *MMI*, the number of our MMk-means total iterations is *O(k),* so that our new k-means algorithm runs in *O(nk^2)* expectation time.

□

5.3.4 EXPERIMENTAL DATA USED

To compare the efficiency and effectiveness of our algorithm, in comparison to the Traditional k-means, the Overlapped and Enhanced k-means algorithms, we tested them using both biological and non-biological data. Details of these files are given next.

5.3.4.1 Biological Data (Malaria Microarray Data)

We tested the algorithms using normalized microarray expression data at varying timepoints for *P. falciparum* microarray experiment data from Bozdech et al. (2003a) and Le Roch et al (2003) as depicted in *Table 5.1*.

Table 5.1 - Short statistics on the three microarray experimental data used in the testing of our algorithm and the other three variants of k-means algorithm

P.f Microarray Experimental data	**Total No Of Genes**	**Time points**
Bozdech et al, (2003a)- 3D7 strain data	4596	53
Bozdech et al., (2003a) – HB3 strain data	4313	48
Le Roch et al, (2003) 3D7 strain data	5159	16

The source of *P. falciparum* for the microarray experiment (Bozdect et al., 2003a) is from a large-scale lab culture grown using parasitisized Red Blood Cells (RBC). From Table 5.1, Bozdech et al. (2003a) microarray experiment described a complete asexual intraerythrocytic developmental cycle (IDC) of 3D7 and HB3 strains of P. *falciparum* such as early ring stage, late ring stage, early trophozoite stage, late trophozoite stage, early schizont stage, late schizont stage and gametocyte stage. By analyzing the IDC transcriptome of the 3D7 strain and HB3 strain of *P. falciparum*, they were able to demonstrate that at least 60% of the genome is transcriptionally active. Le Roch et al., (2003) investigated nine (9) developmental stages by extracting total RNA from sporozoites; from six periodic intracellular asexual blood stages (grown in culture and synchronized by means of two independent methods, a 5% D-sorbitol treatment and a temperature cycling incubator); and from merozoites and mature stage IV and V gametocytes. This design allowed one of the 367,226 probes to be placed, on average, every 150 bases on both DNA strands and they used robust k-means to cluster their 3D7 strain *P. falciparum* microarray data.

5.3.4.2 Non-Biological Data

In order to determine clearly the behaviour of our algorithm on non-biological data, we decided to deploy them using Fahim et al. (2006), datasets as in Table 5.2.

Table 5.2 - Non-Biological data used in the testing of our algorithm and the other three variants of k-means algorithm

Dataset	No of Records	No of Attributes
Abalone	4177	7
Wind	6574	12
Letter	20000	16

In Table 5.2, Abalone dataset described with 8 attributes represents physical measurements of abalone (sea organism). Wind dataset described by 12 attributes represents measurements on wind from 1/1/1961 to 31/12/1978. Letter dataset represents the image of English capital letters. The image consists of a large number of black-and-white rectangular pixel displayed as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15.

5.4 EXPERIMENTATION EXPERIENCE AND RESULTS

The number of genes in Table 5.2 ranges from 4313 – 5159 while the number of time-points is from 16-53. We executed our algorithms on both biological and non-biological data with the following values of k to include 15, 17, 19, 20, 21, 23, 25. The system used is a DeLL computer on MS Windows Vista OS, INTEL® CORE™ DUO CPU T2300 @1.66GHz, 512 RAM, 80GB HDD. The results obtained were plotted. The plots of minimized Mean Standard Error (MSE) versus k values help to measure clusters quality (that is effectiveness) and run time (in sec) versus k helps to measure each algorithm efficiency empirically. For our three malaria microarray data, these plots are shown in *Figure 5.6a – 5.6b*.

5.4.1 MEASURE OF QUALITY USING MSE AND SPEED VIA RUNTIME

The Figures 5.6 show the results obtained when we measured cluster quality, using obtained MSE vs k plots. We also recorded the speed of the algorithms comparatively using the empirical run time vs k plots. The result description for each plot is given under each figure.

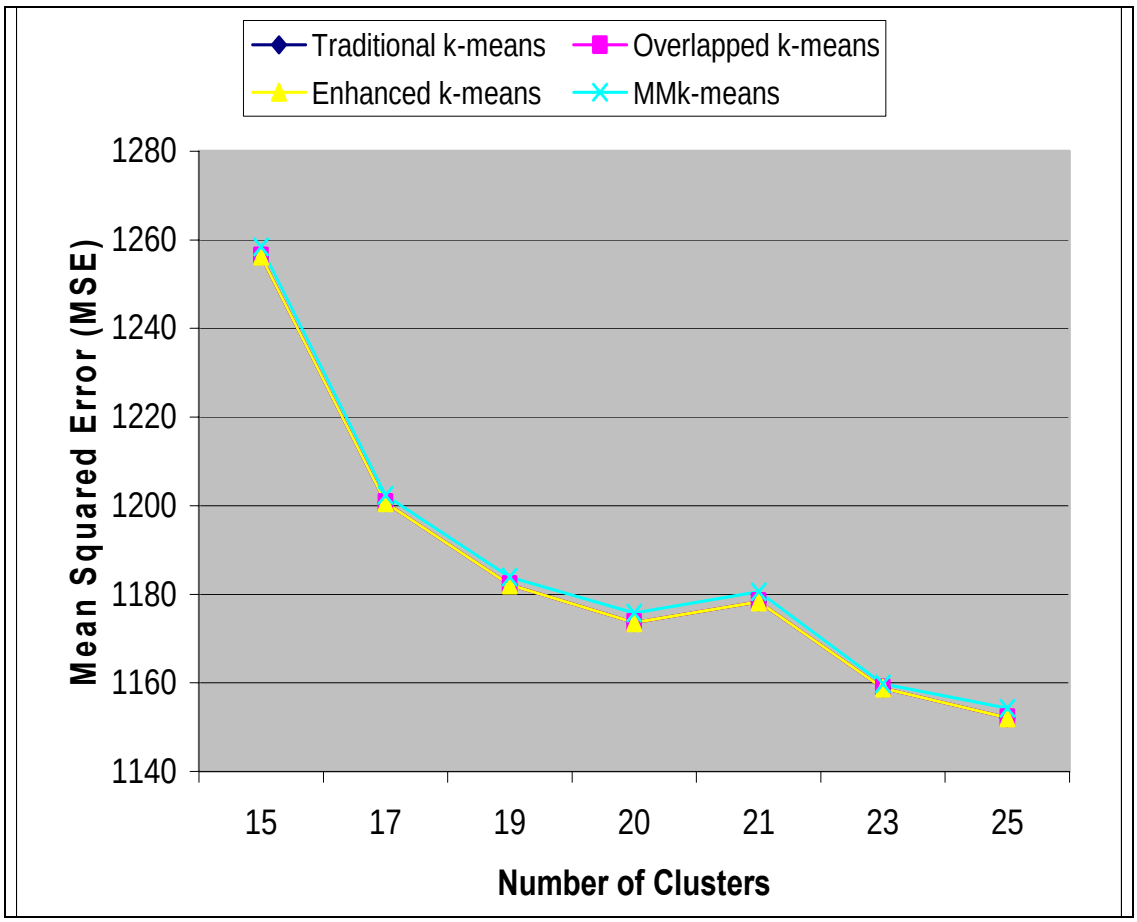


Figure 5.6a: Quality of Clusters (Bozdech et al. (2003a) 3D7 Microarray Dataset)

The qualities of clusters for the four algorithms are similar. The MSE decreases gradually as the number of clusters increases except for k=21 that has a higher MSE than its preceding cluster, k = 20.

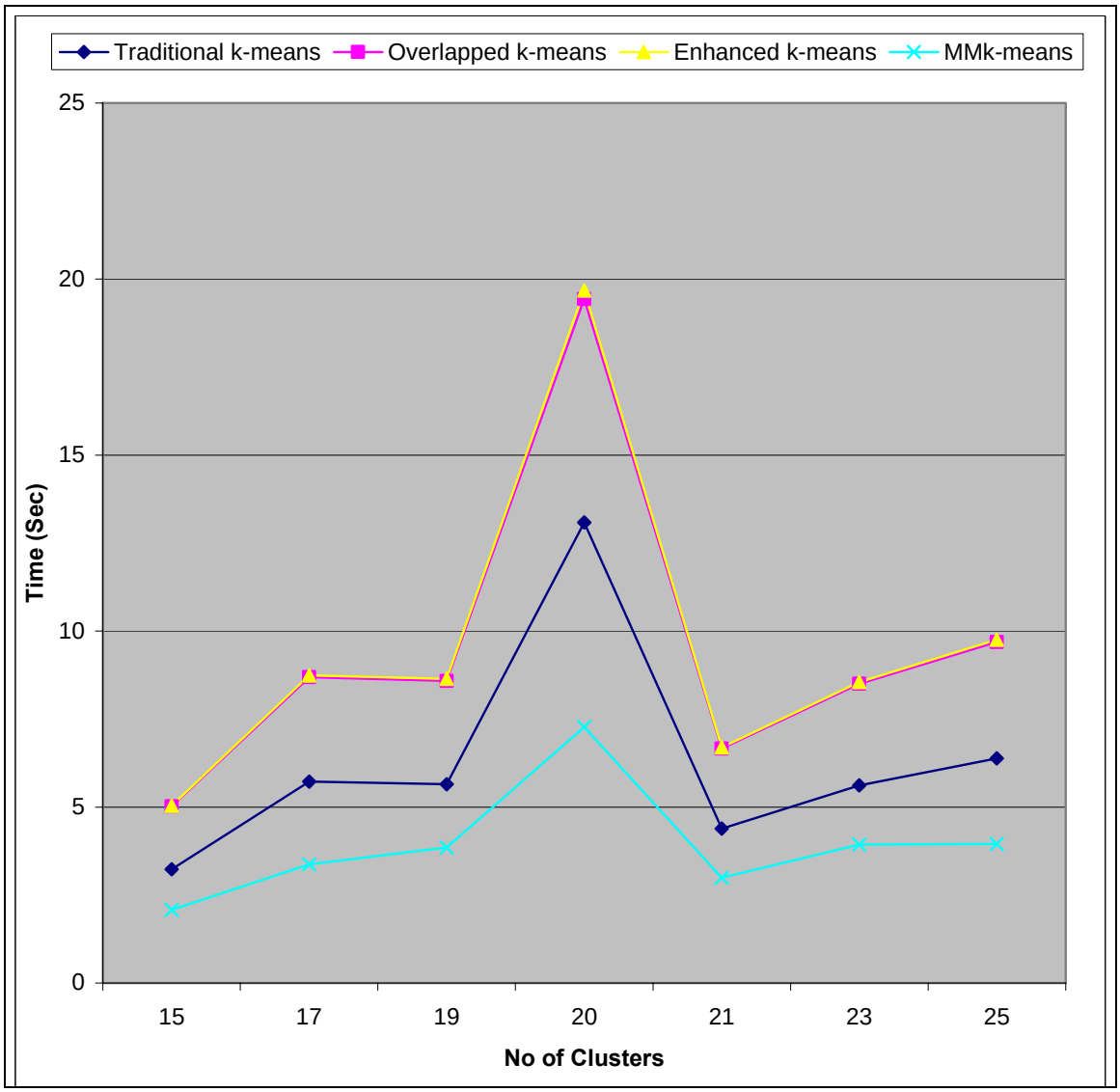


Figure 5.6b: Execution Time (Bozdech et al. (2003a) 3D7 Microarray Dataset)

The plot shows that our MMk-means has the fastest run-time for tested number of clusters, $15 \leq k \leq 25$. Comparatively, k=20 took the longest run-time for all the four algorithms, implying that this is a function of the nature of the data used.

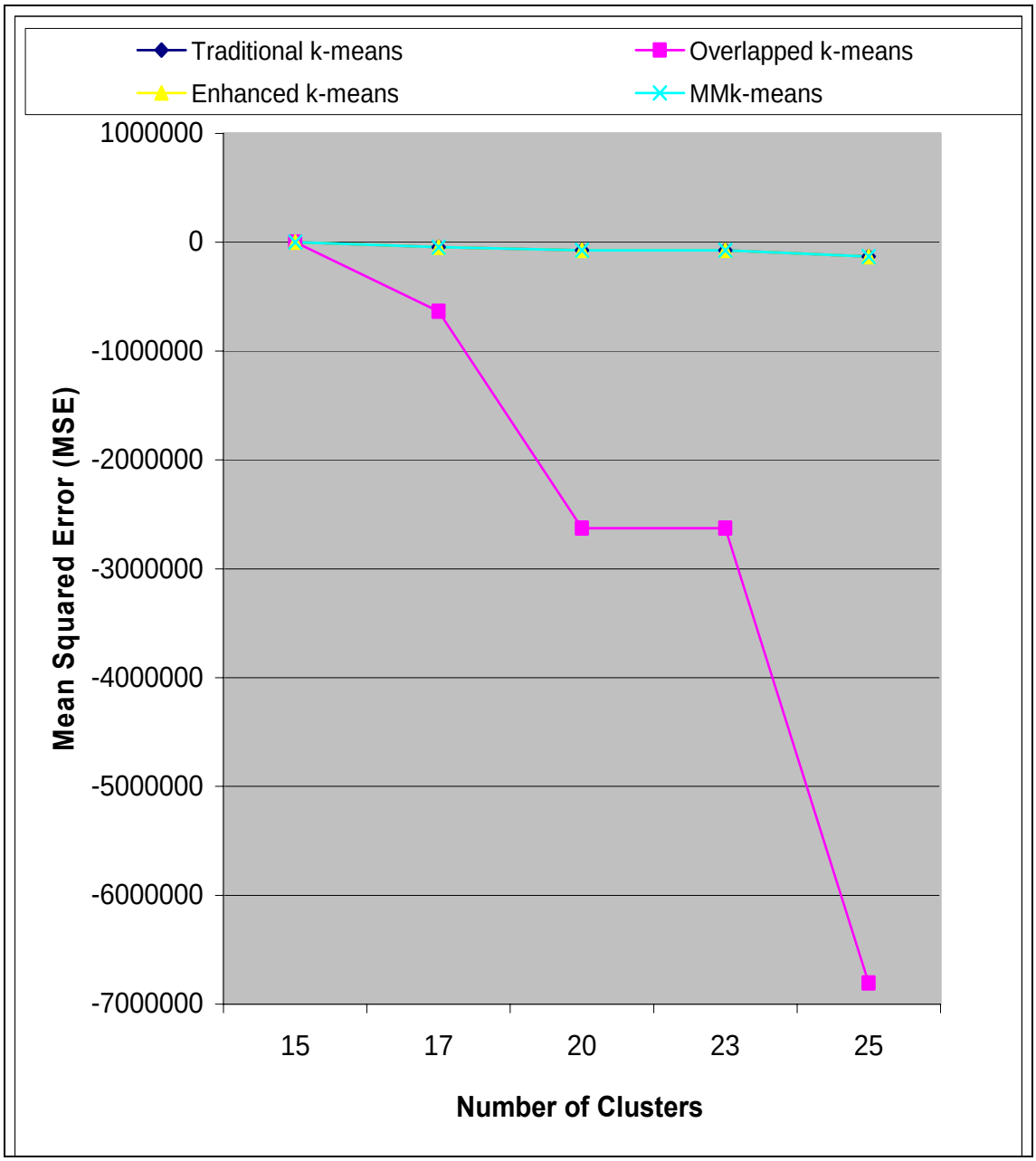


Figure 5.7a: Quality of Clusters (Le Roch et al. (2003) 3D7 Microarray Dataset)

Effective and efficient clustering was achieved only for k=15 while higher values of k create inefficient clustering as many empty clusters are created in the process because $k \geq d$ where d is the dimension of the test microarray data.

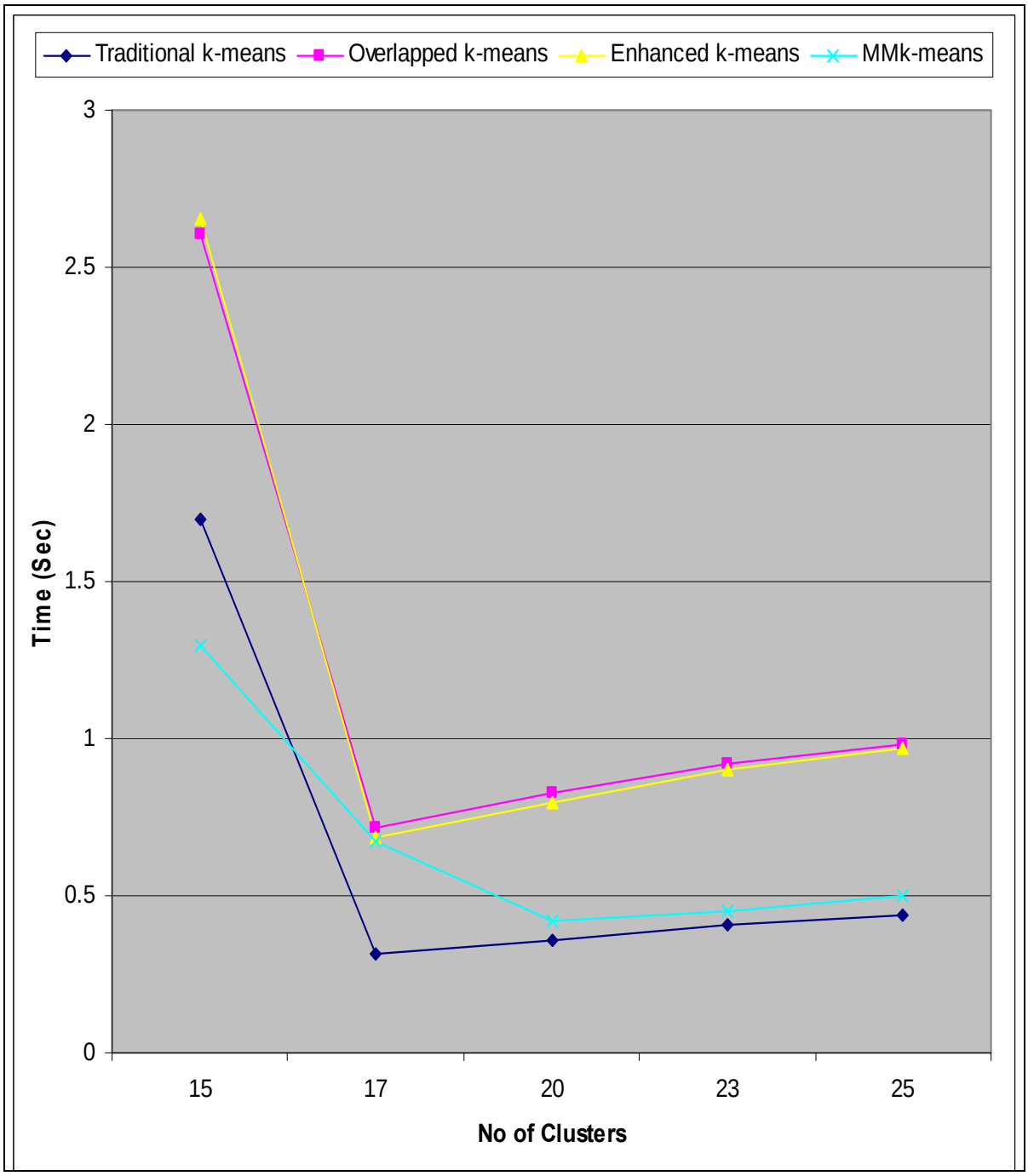


Figure 5.7b: Execution Time (Le Roch et al. (2003) 3D7 Microarray dataset)

Our MMk-means performed best only at k=15 as traditional k-means performed slightly better at other values of k because $k \geq d$. Overlapped and Enhanced were the slowest in all cases.

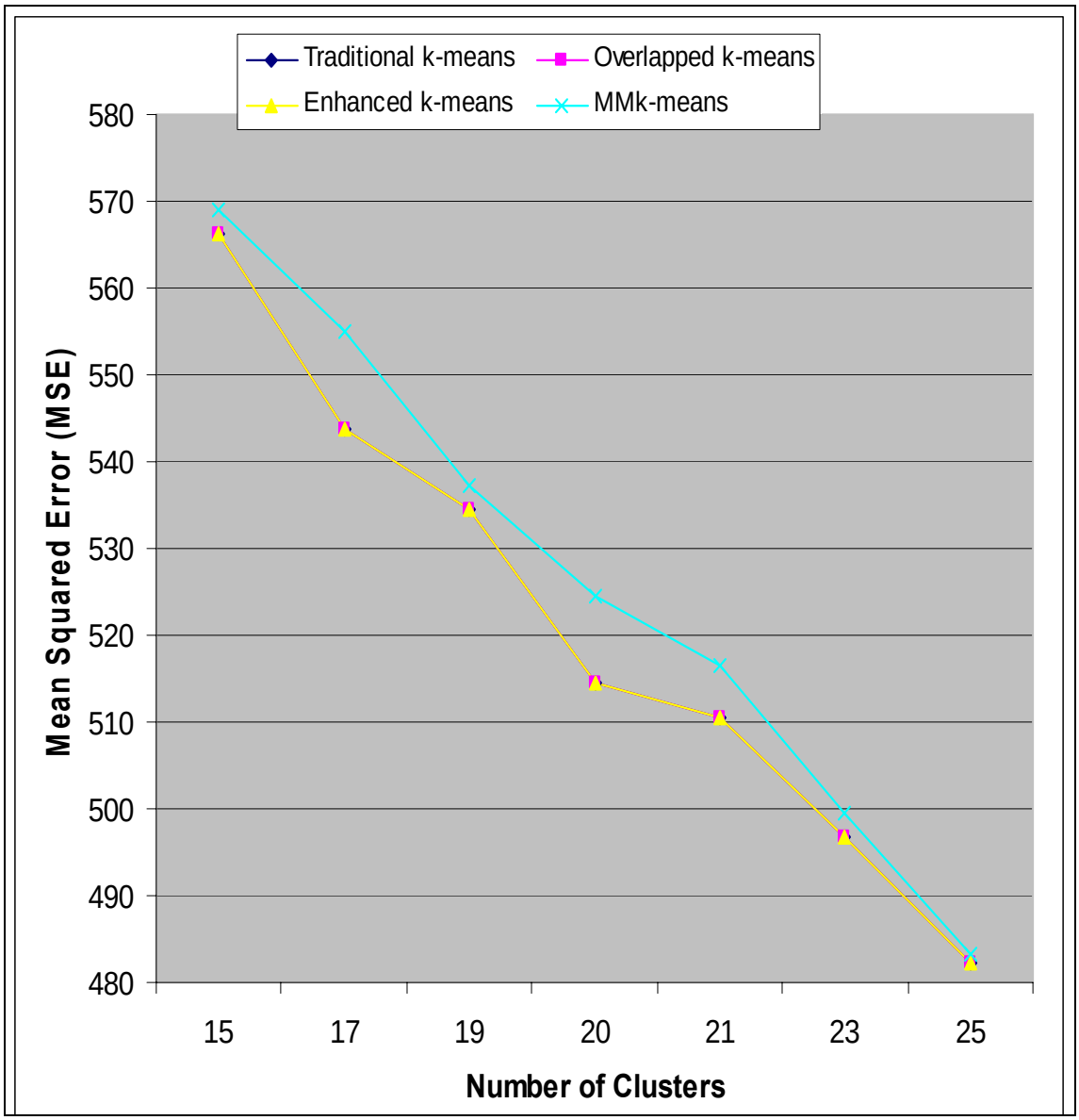


Figure 5.8a: Quality of Clusters (Bozdech et al. (2003a) HB3 Microarray Dataset)

The qualities of clusters for the four algorithms are comparatively similar. The MSE decreases gradually as the number of clusters increases throughout all values of k used in the testing for the four algorithms.

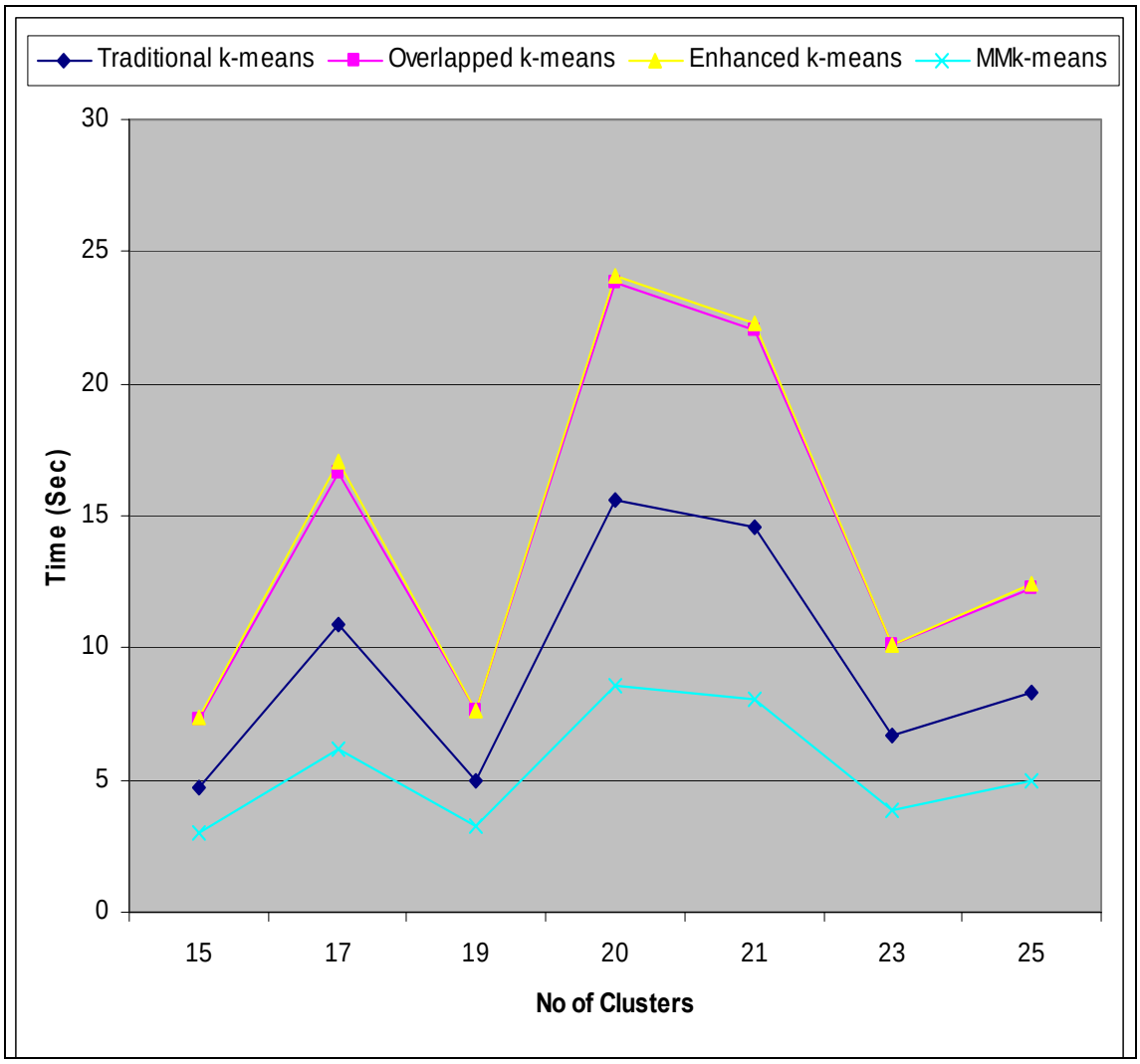


Figure 5.8b: Execution Time (Bozdech et al. (2003a) HB3 Microarray Dataset)

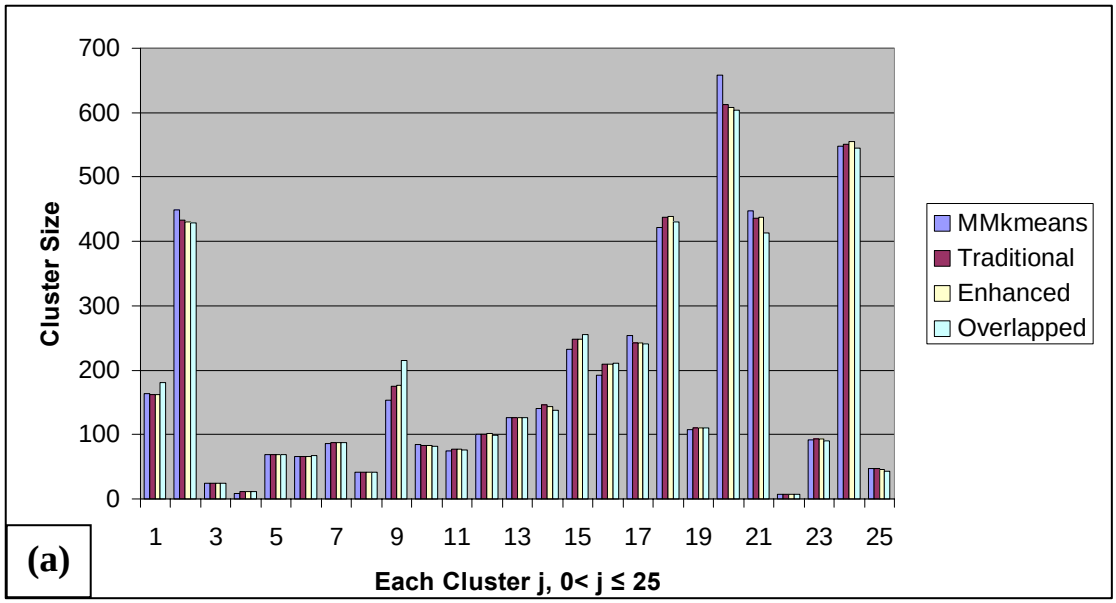
The plot shows that our MMk-means have the fastest run-time for all tested k values, $15 \leq k \leq 25$ while Overlapped and Enhanced k-means are the slowest.

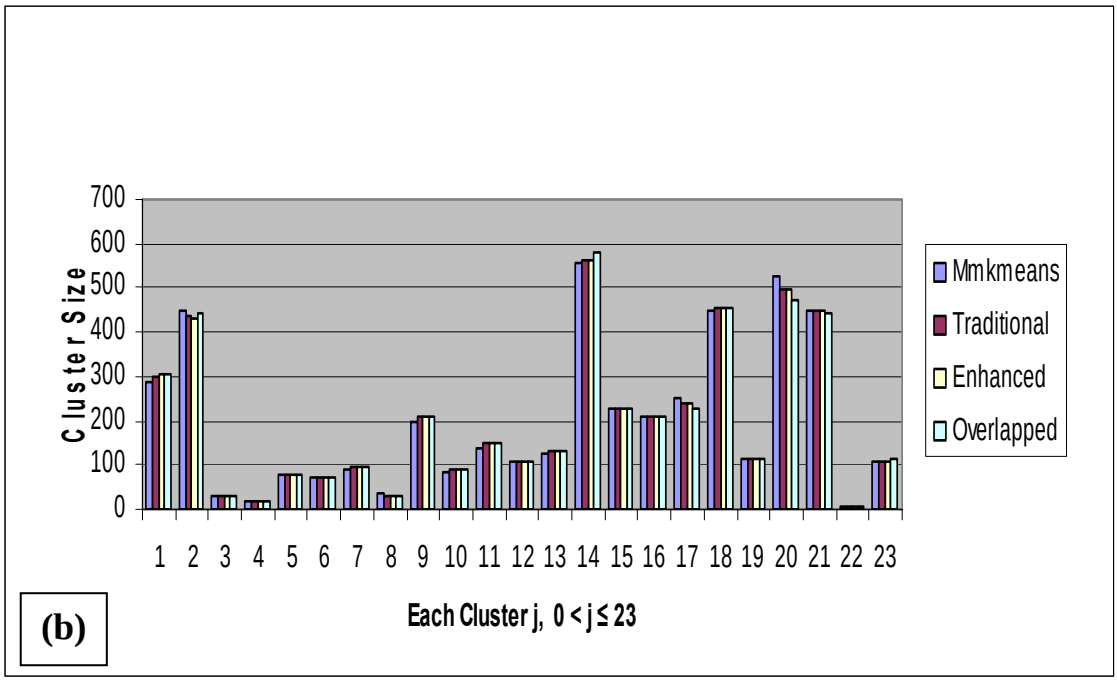
It was observed that eigenvalues of the correlation coefficient matrix r decreases along the diagonal matrix from top to bottom for all iterations before the last iteration and changes interestingly at the last iteration by increasing from top to bottom. It should be noted that the stability condition for clustering as measured by $diff_j$ of line 7 in *Figure 5.5* does not apply appropriately to negative gene expression values we have in Le Roch *et al*. (2003) data. The theoretical reason is given in Ding and He (2004). We observed that nevertheless our new algorithm quality of clustering compared excellently with the Traditional k-means.

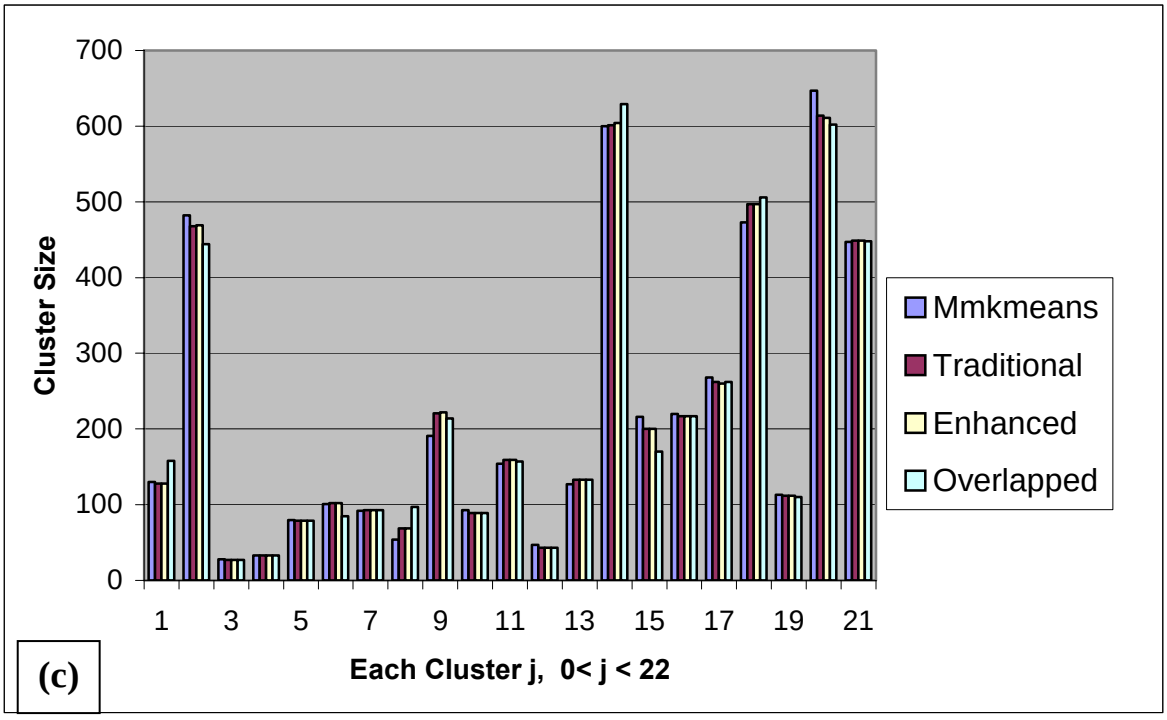
We found out that the algorithms of Fahim *et al*. (2006) were slower than the Traditional k-means contrary to the claim of the authors. Whenever k (number of clusters) < *d* (dimension or timepoints), effective clustering is achieved for the four algorithms and our MMk-means has the best empirical runtime. Overlapped and Enhanced k-means are the slowest in all cases. Results are displayed in the Figures below. Empty clusters are created by all the algorithms if k > *d* as the clustering becomes irregular, similar to results for 15 > k > 25 using Le Roch et al. (2003) data exemplified in *Figure 5.7a-b* above.

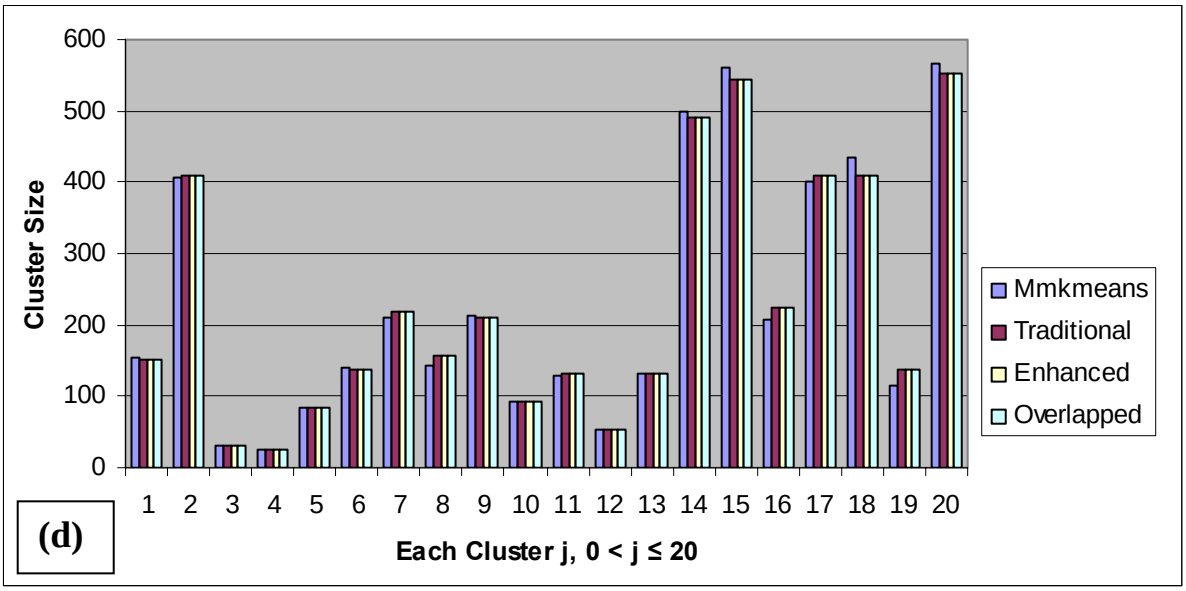
5.4.2 MEASURE OF QUALITY VIA CLUSTER COUNT DISTRIBUTION

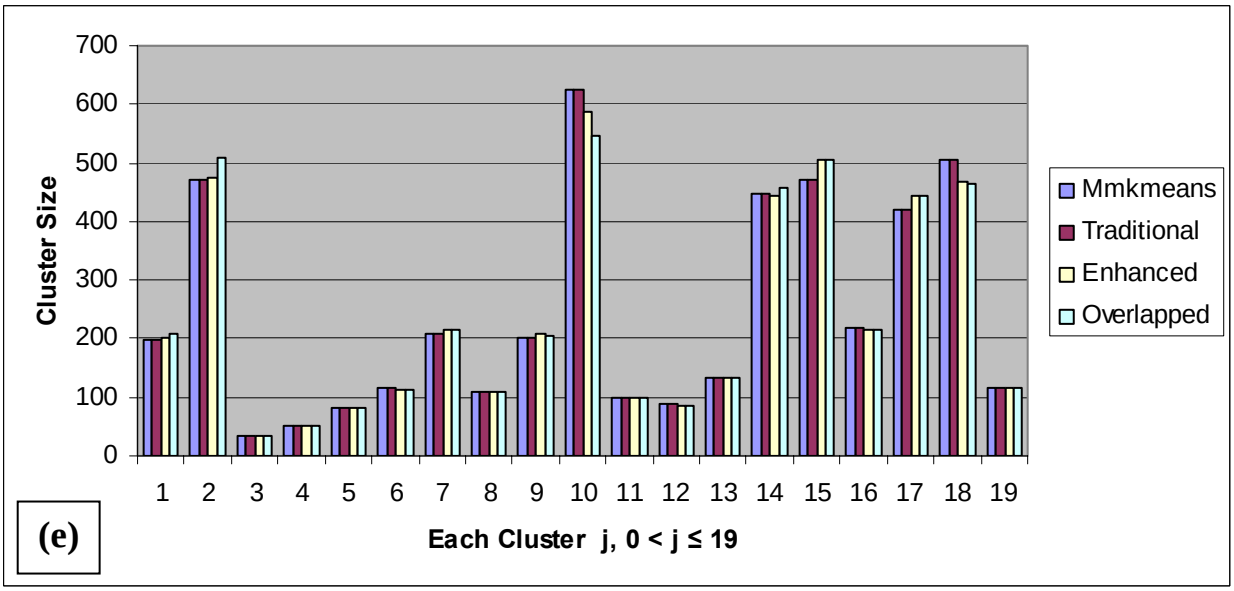
However, the histograms in *Figure 5.8a-g* are to help us study if each cluster count maintains a similar distribution for different k values for the four algorithms. It shows that in all cases, their distributions are closely similar, supporting the argument of comparable cluster quality created by our MMk-means and the other algorithms.

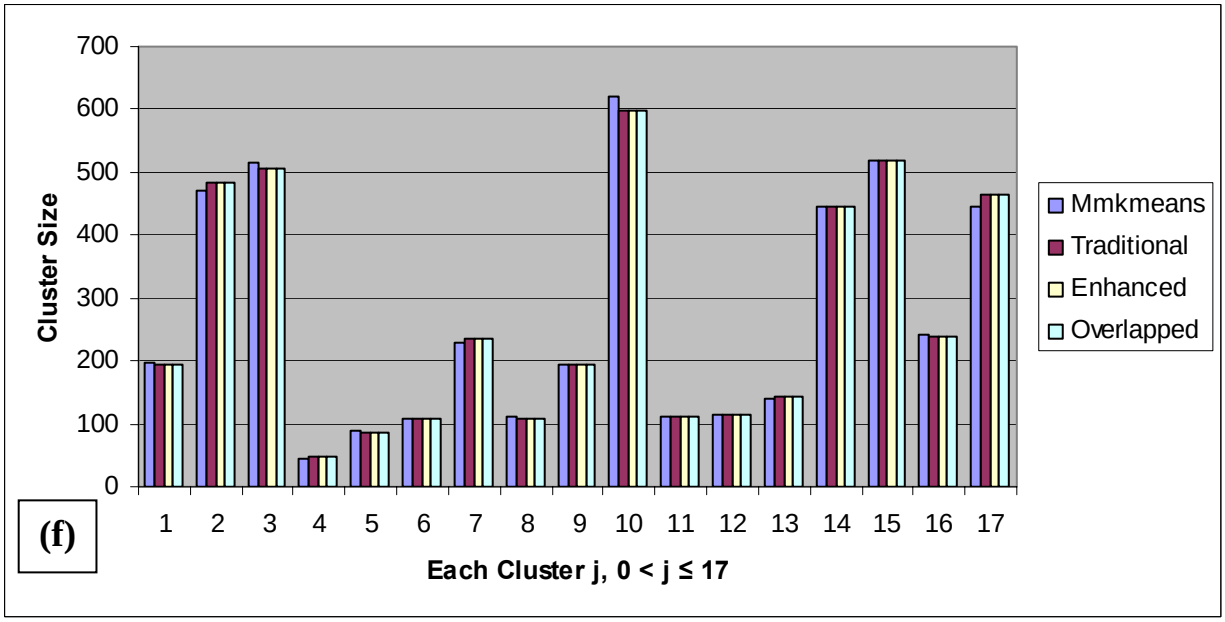



700
600
500
400
300
200
100
0
Cluster Size
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Each Cluster j, 0 < j ≤ 23
Mmkmeans
Traditional
Enhanced
Overlapped
(b)


700
600
500
400
300
200
100
0
Cluster Size
1 3 5 7 9 11 13 15 17 19 21
Each Cluster j, 0< j < 22
Mmkmeans
Traditional
Enhanced
Overlapped
(c)


600
500
400
300
200
100
0
Cluster Size
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Each Cluster j, 0 < j ≤ 20
Mmkmeans
Traditional
Enhanced
Overlapped
(d)


700
600
500
400
300
200
100
0
Cluster Size
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Each Cluster j, 0 < j ≤ 19
Mmkmeans
Traditional
Enhanced
Overlapped
(e)


700
600
500
400
300
200
100
0
Cluster Size
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Each Cluster j, 0 < j ≤ 17
Mmkmeans
Traditional
Enhanced
Overlapped
(f)

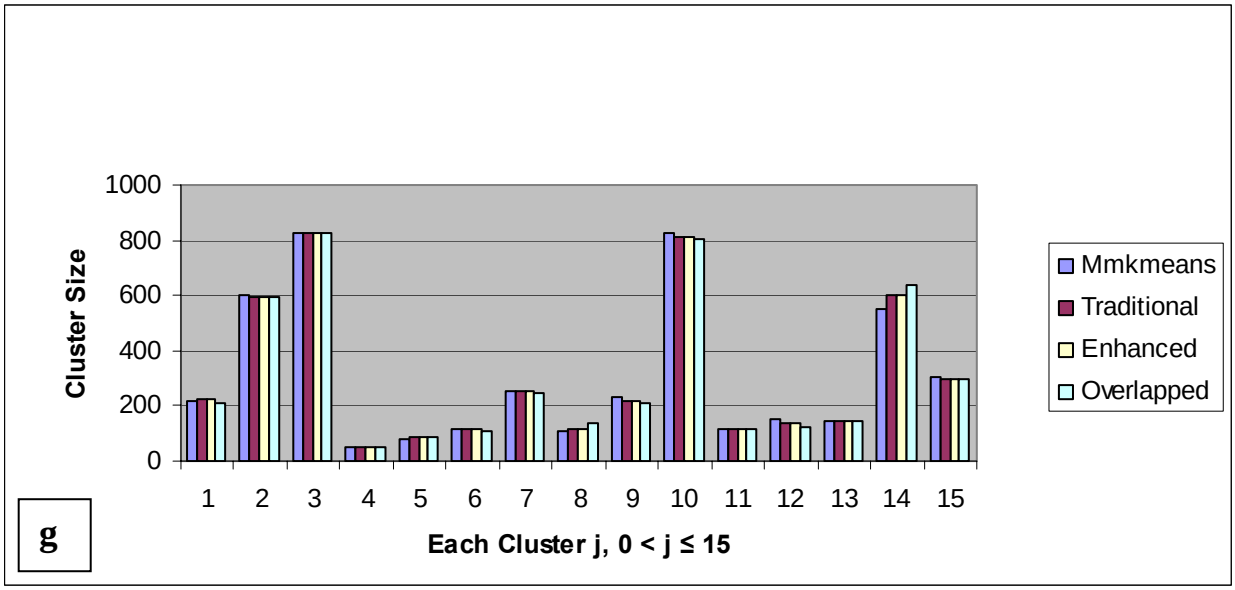


Fig 5.9a-g: Distribution of Cluster size for four k-means algorithms on Bozdech et al. (2003a) 3D7 Microarray Dataset

These plots help us to study if each cluster membership count maintains a similar distribution for different k values for the four algorithms. It shows that for all cases, their distributions are closely similar, supporting the argument of comparable cluster quality created by our MMk-means and the other algorithms.

To demonstrate the biological characteristics of our new algorithm against other well known k-means clustering algorithms, in Osamor *et al*. (2009), we compared three different k-means algorithms (Robust, Traditional and MM k-means respectively) results from an *in-vitro* microarray data of Le Roch et al. (2003) with the classification from an *in-vivo* microarray data of Daily *et al.* (2007). Our aim was to perform a comparative functional classification of *P. falciparum* genes with a view to obtaining further knowledge on many *P. falciparum* genes. Interestingly, we discovered a new functional group for some set of genes.

5.4.3 MEASURE OF QUALITY VIA THE HUBERT – ARABIE ADJUSTED RAND INDEX (ARI_{HA})

It has been noted that the plots of MSE do not provide a compelling arguments about the similarity of the solutions of two or more clustering algorithms (Steinley, 2004; Steinley, 2006). To further ascertain the quality of our new algorithm on the three microarray data of *Table 5.1*, we assessed the quality of its clusters against the clusters of the known structure using the Hubert-Arabie Adjusted Rand index (ARI_{HA}) (Steinley, 2004). The result of this assessment is given in *Table 5.3(a)*.

Table 5.3(a): ARI_{HA} Computation for Biological data

	Traditional k-means						**Enhanced k-means**					
	Bozdech et al.(2003)-3D7 strain		Bozdech et al.(2003)-HB3 strain		Le Roch et al. (2003)		Bozdech et al.(2003)-3D7 strain		Bozdech et al.(2003)-HB3 strain		Le Roch et al. (2003)	
	k=15	k=20	k=15	k=20	k=10	k=15	k=15	k=20	k=15	k=20	k=10	k=15
MMk-means	0.9480	0.9170	0.9068	0.6488	0.9352	0.6643						
Enhanced k-means	0.9935	1.0000	0.9901	0.9967	0.9717	0.9728						
Overlapped k-means	0.9635	1.0000	0.9707	0.9920	0.8837	0.8682	0.9636	1.0000	0.9720	0.9917	0.8891	0.8916

For each data, Bozdech et al. (2003a) 3D7 and HB3 strains (Bozdech et al., 2003a) and Le Roch et al. (2003), we used two values of k to demonstrate the effect of changing k values on the clusters quality of the clustering algorithms. We considered the structure of the Traditional k-means as the known structure and compared the clusters of MM, Enhanced and Overlapped k-means respectively with it. In a separate column, we also compared the structure of Enhanced k-means with that of Overlapped k-means. This is to assess the two k-means algorithms presented in Fahim et al. (2006). We found out that Enhanced and Overlapped k-means respectively produced similar clusters and their structures were similar to that of the Traditional k-means. For MMk-means, this is also the case and we found categorically that when k is close to d, the quality of its clusters

is good (ARI_{HA} > 0.8) and when k is not close to d, the quality is excellent (ARI_{HA} > 0.9).

To test the behavior of our new algorithm on non-biological data, we used the data, Fahim et al. (2006), deployed their implementations in their experimental experience. These include a Letter image, an Abalone and a Wind datasets. Details on these datasets are given in *Table 5.2*. We assessed also here the quality of our new k-means algorithm, that of Enhanced and Overlapped k-means respectively using the Hubert-Arabie Adjusted Rand index (ARI_{HA}) (Steinley, 2004). We set the Traditional k-means algorithm clusters from these datasets as the known ones. The result of this exercise is given in *Table 5.3(b)*.

Table 5.3(b) ARI_{HA} Computation for Non-Biological data

	Traditional k-means						**Enhanced k-means**					
	Abalone		Wind		Letter		Abalone		Wind		Letter	
	k=5	k=7	k=5	k=12	k=5	k=10	k=5	k=7	k=5	k=12	k=5	k=10
MMk-means	0.8472	0.6045	1.0000	0.9205	0.8623	0.8015						
Enhanced k-means	0.9454	0.9837	0.9992	0.9997	0.9930	1.0000						
Overlapped k-means	0.9540	0.9004	0.9895	0.9821	0.9875	1.0000	0.9544	0.9064	0.9904	0.9818	0.9879	1.0000

Table 5(b) shows Hubert-Arabie Adjusted Rand Index (ARI_{HA}) Cluster Quality Computation Result for Non-biological data. For each data, Abalone and Wind, Letter Image of Fahim et al. (2006), we used two values of k to demonstrate the effect of changing k values on the clusters quality of the clustering algorithms. We considered the structure of the Traditional k-means as the known structure and compared the clusters of MM, Enhanced and Overlapped k-means respectively with it. In a separate (last) column, we also compared the structure of the Enhanced k-means with that of Overlapped k-means.

The quality of MMk-means clusters is similar to what we observed from that of the biological data.

5.5 DISCUSSION ON IMPLEMENTATION ISSUES

Using C++, we implemented the three variants of k-means algorithms, namely, the Traditional, Overlapped and Enhanced k-means following Fahim et al. (2006) design. We also implemented a fourth one, our MMk-means algorithm using C++ and MATLAB. They are available under the GNU open-source license. Please see <http://sourceforge.net> for more details. From (Ding and He, 2004), we used their experimentally determined interval: 0.5 – 1.5%, which indicates when a cluster is optimally equal to the expected ones.

In computing distance as stated in *Figure 3.4,* line 19, we use Pearson correlation which fits best for microarray data since we are interested in coexpressed and coregulated genes. When we say that genes are co-expressed (referring to co-expressed genes), we mean that such sets of genes have similar expression patterns for some biological processes or functions. There are two reasons for interest in coexpressed genes. First, there is evidence that many functionally related genes are coexpressed. For example, genes coding for elements of a protein complex are likely to have similar expression patterns which ultimately can be used to identify previously uncharacterized genes. The second reason for interest in coexpressed genes is that coexpression may reveal much about the genes' regulatory systems. For example, if a single regulatory system controls two genes, then we might expect the genes to be coexpressed, which gives a good clue of the organism's regulatory network. In general, there is likely to be a relationship between coexpression and coregulation (Heyer et al., 1999).

5.6 CONCLUSION

A major contribution of this work so far, has been the development of a novel Metric Matrices k-means. The efficiency of our algorithm maintains a better result than Traditional, Overlapped and Enhanced k-means algorithms. Its effectiveness is quite comparable to results obtained by these other variants of k-means algorithm for most values of k as demonstrated in Osamor et al. (2009) and also in this work.

In this work, emphases are on the reduction of the time requirement of the k-means algorithm and its application to microarray data due to the desire to create a tool for malaria research. However, the new clustering algorithm can be used for other clustering needs as long as an appropriate measure of distance between the centroids and the members is used. This was demonstrated above on three non-biological data of *Table 5.2*.

CHAPTER SIX

COMPARATIVE FUNCTIONAL CLASSIFICATION OF *PLASMODIUM FALCIPARUM* GENES USING K-MEANS CLUSTERING

6.1 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 transforms 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 differentiate 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 towards eradication of this disease (Bozdech et al., 2003b). 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) lacks 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* (Bozdech et al, 2003a). 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.

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. (2003). Daily et al. (2007) used the non negative matrix factorization (NMF) algorithm (Brunet et al., 2004) 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 (Le Roch et al., 2003) 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 3D7 strain grown *in-vitro* (Le Roch et al., 2003) and that the other two clusters were not observed *in-vitro*.

They later interpreted these three clusters biological bases by comparing them with an extensive compendium of expression data in the yeast *Sacchromyces cerevisiae.* This 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. It, 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 of a functional mitochondrion in the asexual-stage parasite.

In this work, our original intention is to further validate the effectiveness of our new and novel MMk-means (Osamor et al., under review) algorithm, presently under publication consideration review, by comparing three different k-means algorithms (including MMk-

means) results on Le Roch et al. (2003) *in-vitro* microarray data with the *in-vivo* microarray data of Daily et al. (2007). 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. Our results had been published in Osamor et al. (2009).

6.2 METHODOLOGY AND RESULTS

We enumerate next the data and the algorithms employed.

6.2.1 DATA USED

Daily *et al.* (2007) data were obtained using venous blood samples from *P. falciparum*-infected patients in Senegal. This cohort consisted of patients who presented to the district hospital in Velingara, Senegal, with fever and symptoms suggestive of malaria. Le Roch *et al.* (2003) used lab cultured samples of *P. falciparum* and reported that 2235 genes were significantly expressed. This is shown in row 1 of Table 6.1 below. Daily *et al.* (2007) 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 (Tusher *et al.*, 2001) to extract the list of significant genes from the three clusters of Daily *et al*. (2007) as listed in row 2 of Table 6.1.

Table 6.1: 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

6.2.2 ALGORITHMS USED

6.2.2.1 SAM (Significant Analysis of Microarrays)

SAM, as proposed by Tusher et al., (2001) 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 genes is significantly related to the response. The cut-off for significance is determined by a tuning parameter delta (Δ), chosen by the user based on the false positive rate. SAM outputs a 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 programming and Excel 2000 or higher as an Excel add-in.

6.2.2.2 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. 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.

6.2.2.3 Robust k-means Clustering Algorithm

The robust k-means clustering algorithm was first used in Le Roch *et al.* (2003). 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.* (2003), trials were made for k=10, 15, 20, 25, and 30. k=15 was found to produce meaningful classification. Le Roch *et al.* (2003) 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.

6.2.2.4 Metric Matrices k-means (MMk-means) Clustering Algorithm

A new and novel MMk-means algorithm was developed by us in Osamor *et al.* (under review) and it is simple but more 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.* (2006). The new algorithm is based on the recently established relationship between principal component analysis and the k-means clustering (Ding and He, 2004). 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 (2004) 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 arrives at an $O(nk^2)$ expected run time algorithm. Results obtained from testing the algorithm on five different types of microarray data (Osamor et al., under review) also indicate that the new MMk-means clustering algorithm is empirically faster than other known k-means algorithms.

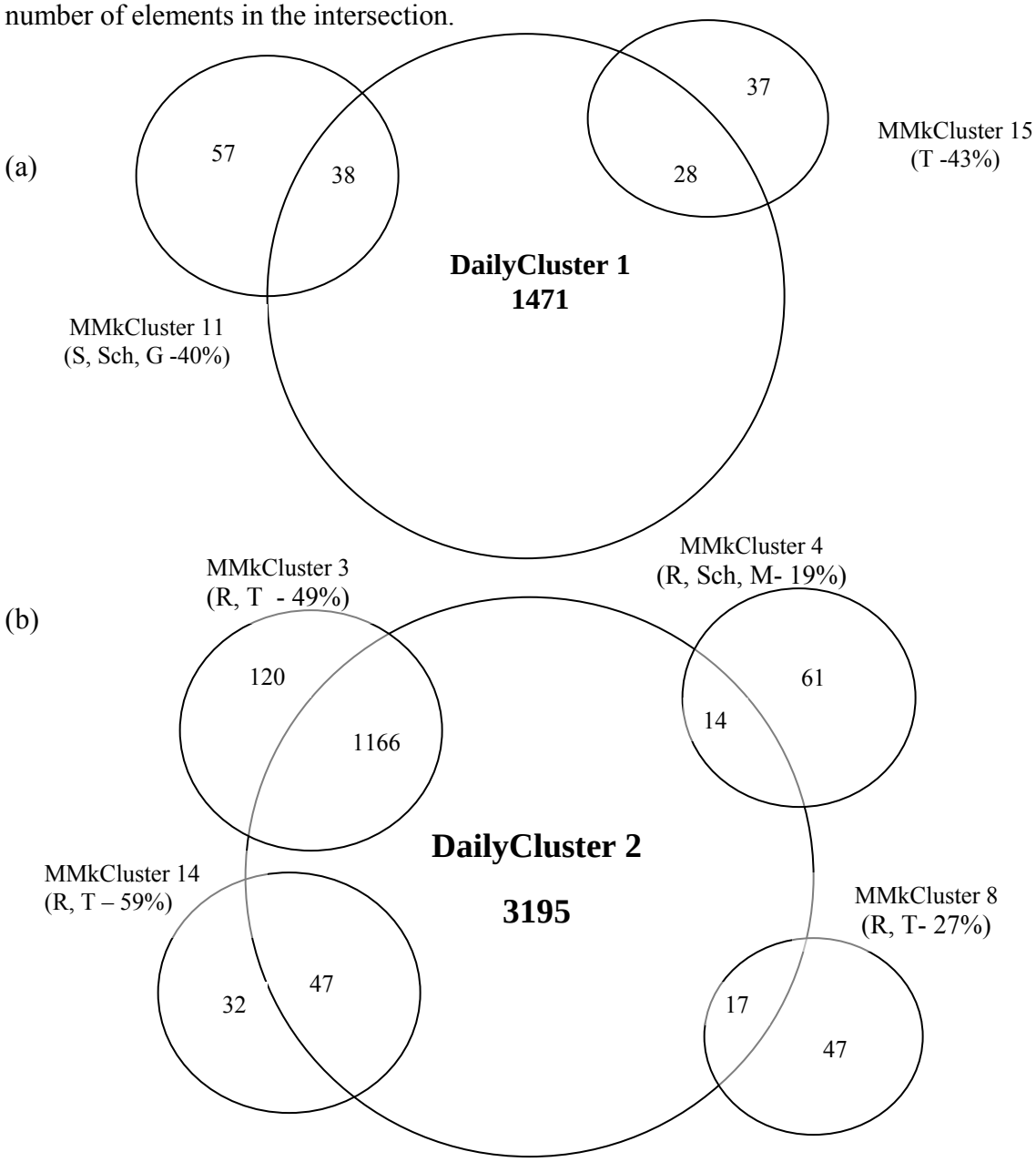
From our previous work (Osamor et al., under review), we implemented the traditional and MMk-means algorithms respectively. First, we deployed them to clusters, for k=15, *P. falciparum* microarray data containing 5159 genes and 16 timepoints arising from the work of Le Roch *et al.* (2003). The traditional k-means algorithm is set a gold standard and is used to validate MMk-means algorithm while the Robust k-means clustering results in Le Roch *et al.* (2003) for k=15 serves 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 6.2*. 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.* (2003) clusters (*as recorded in columns 9 and 10 in Table 6.2*). 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.* (2003) data as generated by the three (3) algorithms on Daily *et al.* 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) (Tusher et al., 2001) at the settings of delta (Δ) = 0, data type = One Class, to extract a list of significant genes that are highly expressed for each of the three clusters (*see Table 6.1)*. Delta setting of 0 ensures that all the significantly expressed genes are extracted. However, we also obtained the same number of significantly expressed genes for cluster 1 with $0 \leq \Delta \leq 11.866$, beyond this range, the list of significant genes reduces.

We compared clusters 1-15 from Le Roch *et al.* data for each of the three k-means algorithms with each cluster of Daily *et al.* and computed the percentage number of genes common to both. This resulted in three tables. These are given in Tables 6.3 – 6.5.

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.* (2003) on the classification from the *in-vivo* microarray of Daily *et al.* (2007). The resulting three venn diagrams are similar. Fig. 6.1 shows the results of our MMk-means. Fig. 6.2 depicts that of Robust k-means and Fig. 6.3 gives the venn diagram describing the results of Traditional k-means algorithm from the *in vitro* microarray data of Le Roch et al (2003) on the classification from the microarray of Daily et al (2007). Note that, to avoid over clustering each venn diagram, except for cluster 2 of Daily *et al.* We represented only clusters that pass the following similarity constraint: n*(X ∩Daily cluster) ≥ 40%,* where '*X* ' represents any cluster obtained from the runs of Robust, Traditional and MMk-means respectively and '∩' is a set notation that captures the number of elements in the intersection.



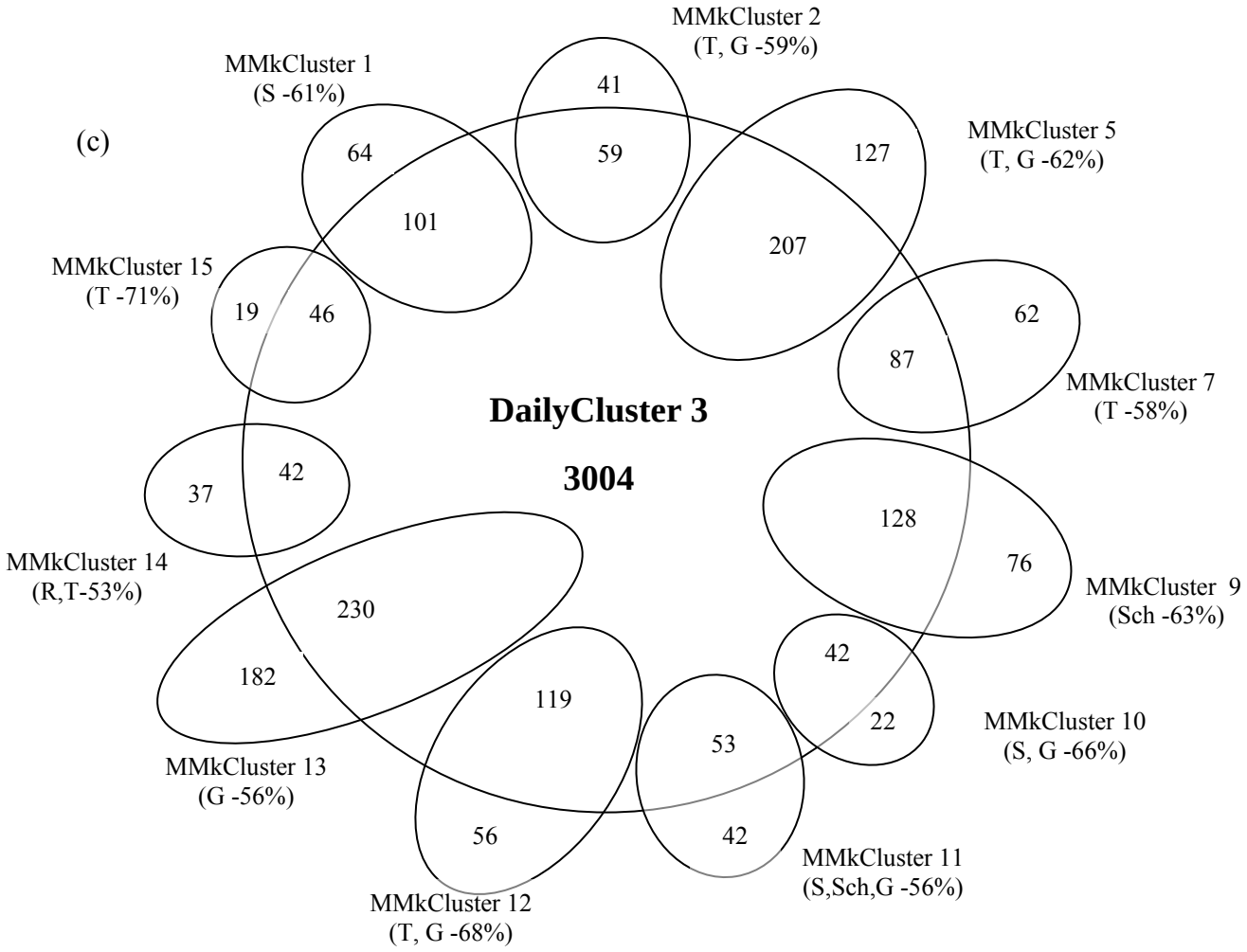
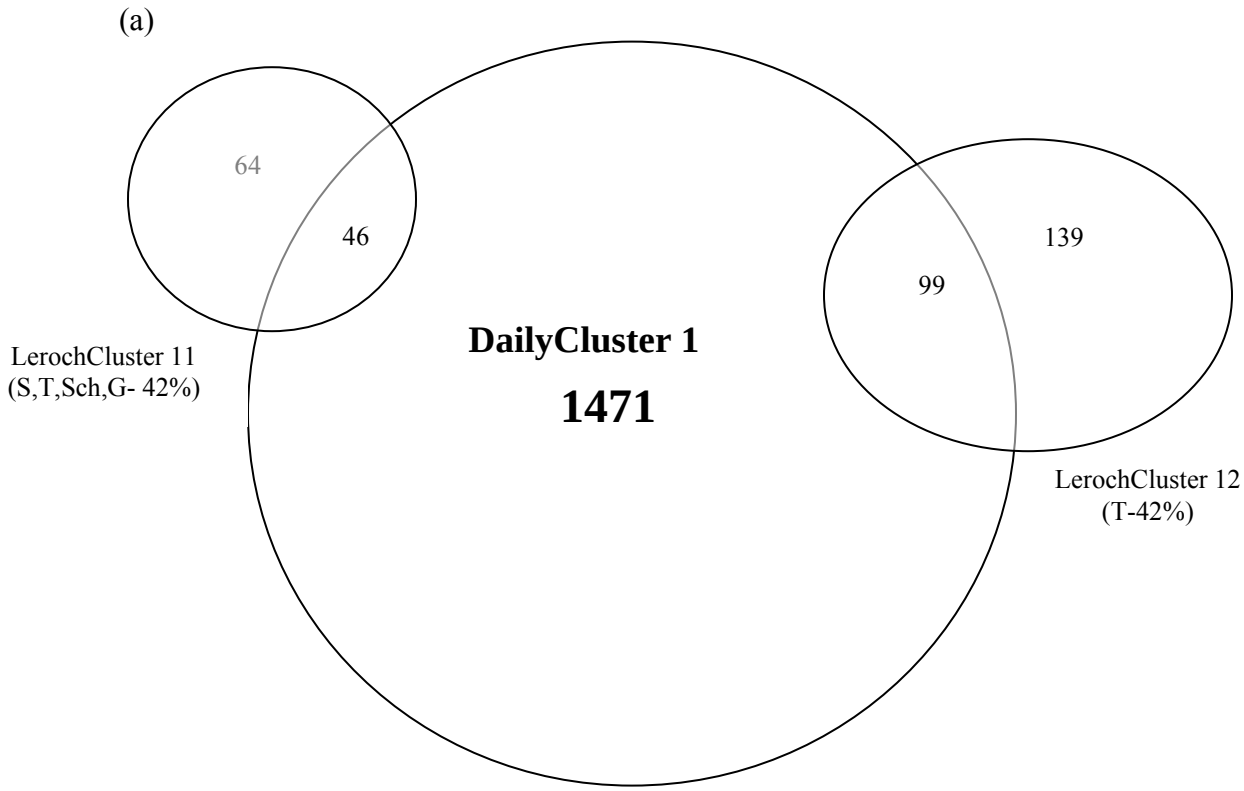
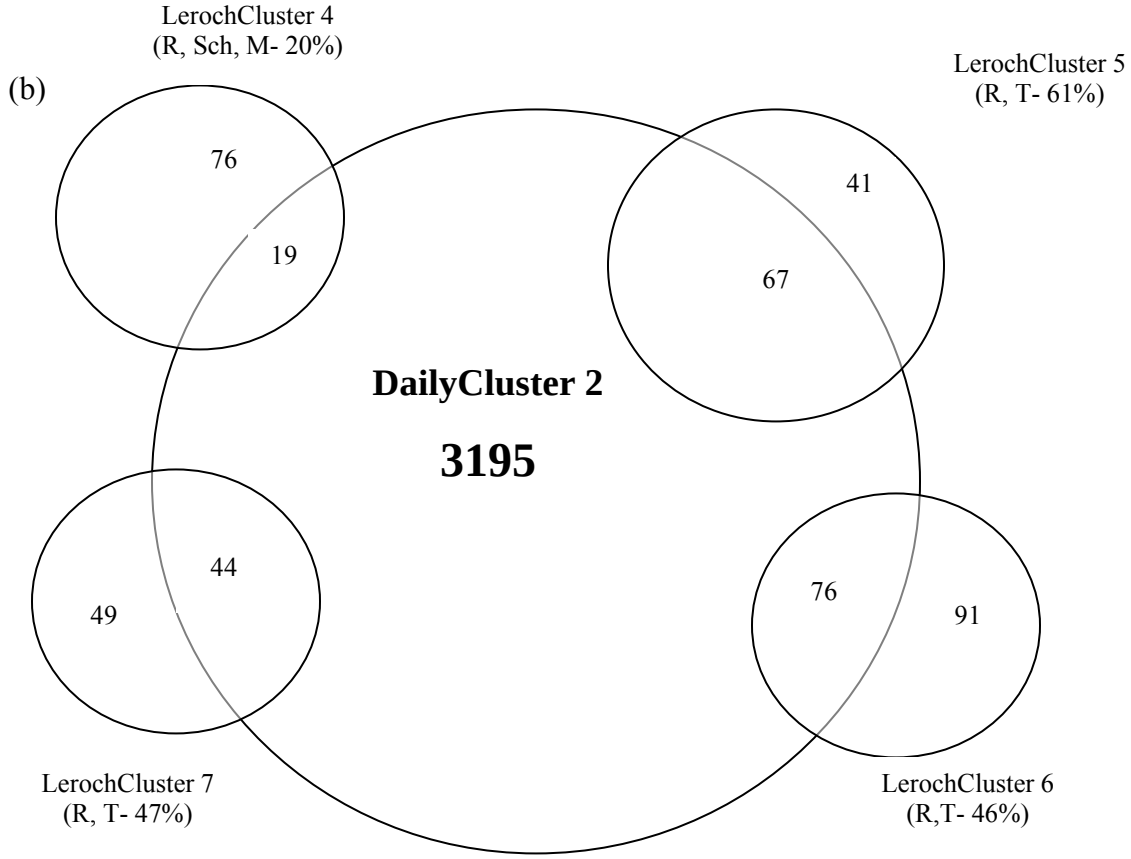


Figure 6.1a-c: Venn diagram of MMk-means Clustered data of Le Roch *et al.*, 2003 and NMF clustered data of Daily *et al.*, 2007.

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)
64
46
139
99
DailyCluster 1
1471
LerochCluster 11
(S,T,Sch,G- 42%)
LerochCluster 12
(T-42%)


LerochCluster 4
(R, Sch, M- 20%)
LerochCluster 5
(R, T- 61%)
(b)
76
19
41
67
DailyCluster 2
3195
44
49
76
91
LerochCluster 7
(R, T- 47%)
LerochCluster 6
(R,T- 46%)

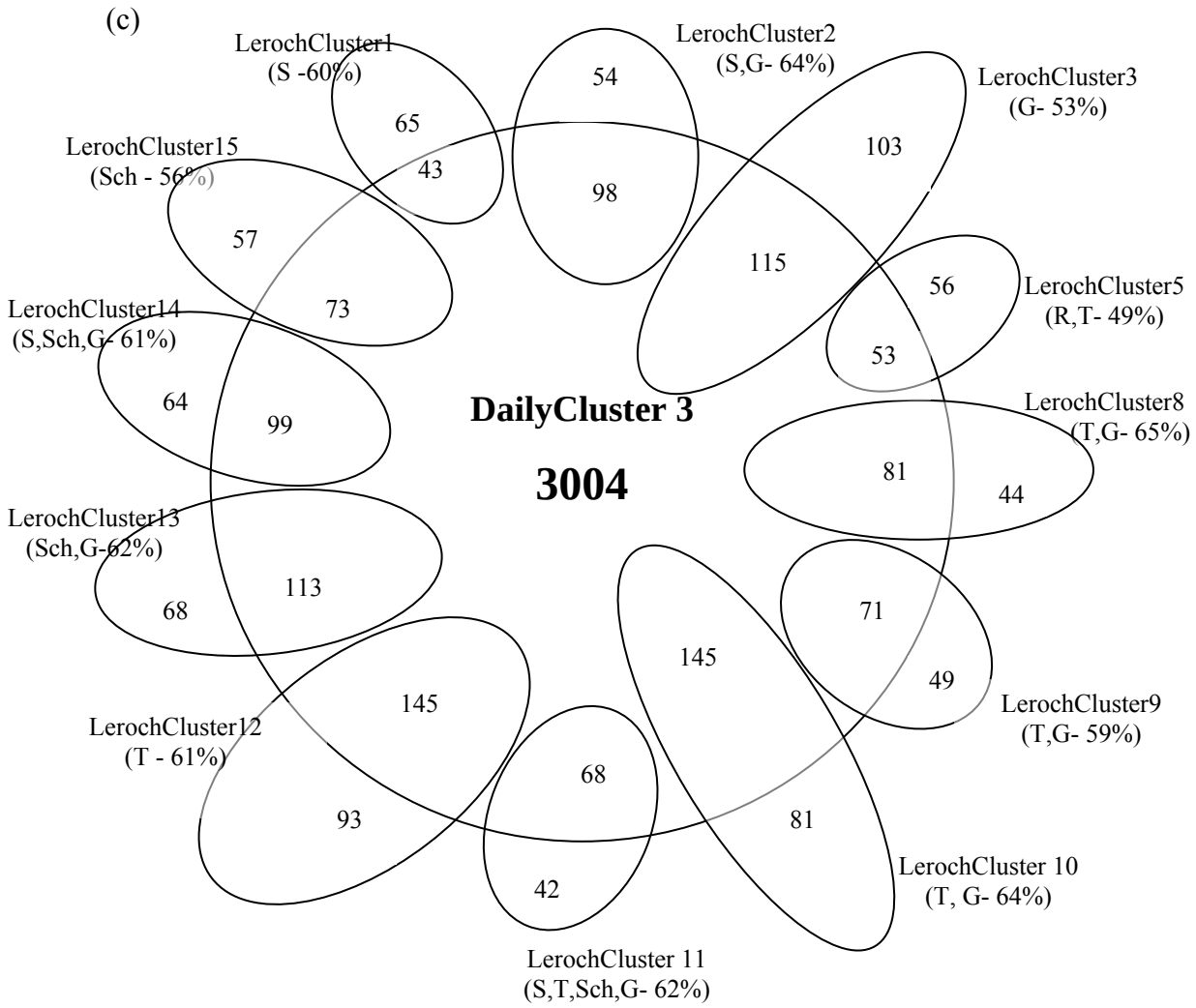
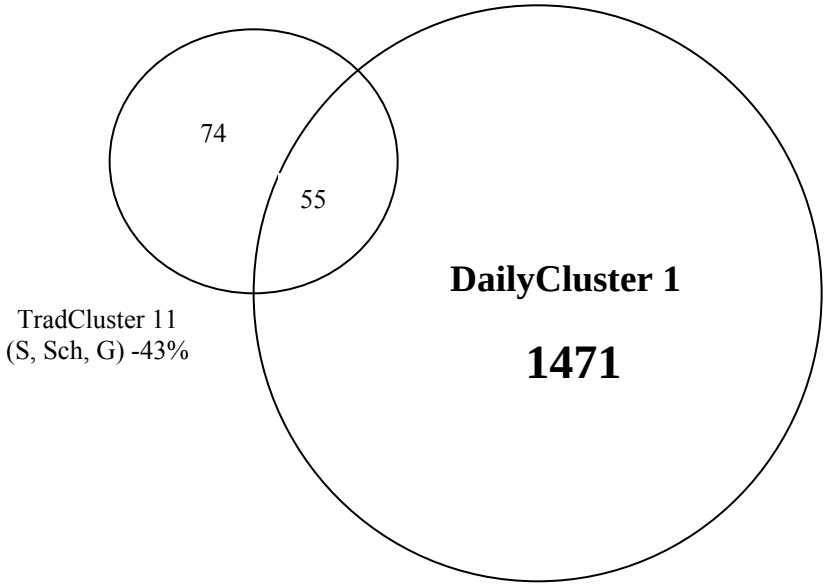


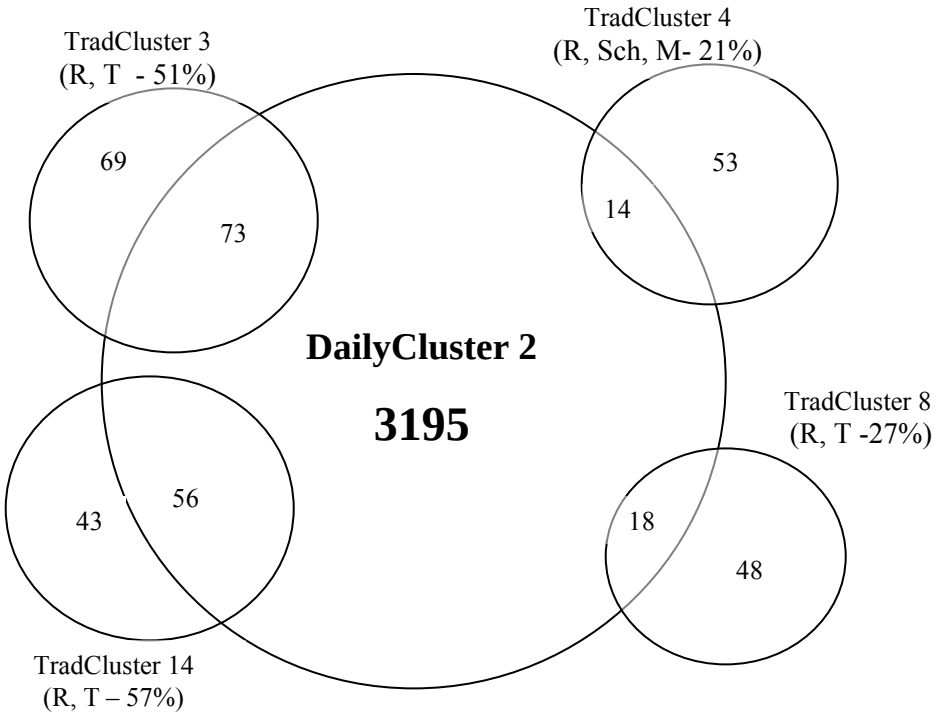
Figure 6.2a-c: Venn diagram of Robust k-means clustered data of Leroch *et al.*, (2003) and NMF clustered data of Daily *et al.* (2007).

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 MmkCluster 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)


74
55
DailyCluster 1
1471
TradCluster 11
(S, Sch, G) -43%

(b)


TradCluster 3
(R, T - 51%)
TradCluster 4
(R, Sch, M- 21%)
69
73
53
14
DailyCluster 2
3195
TradCluster 8
(R, T -27%)
43
56
18
48
TradCluster 14
(R, T – 57%)

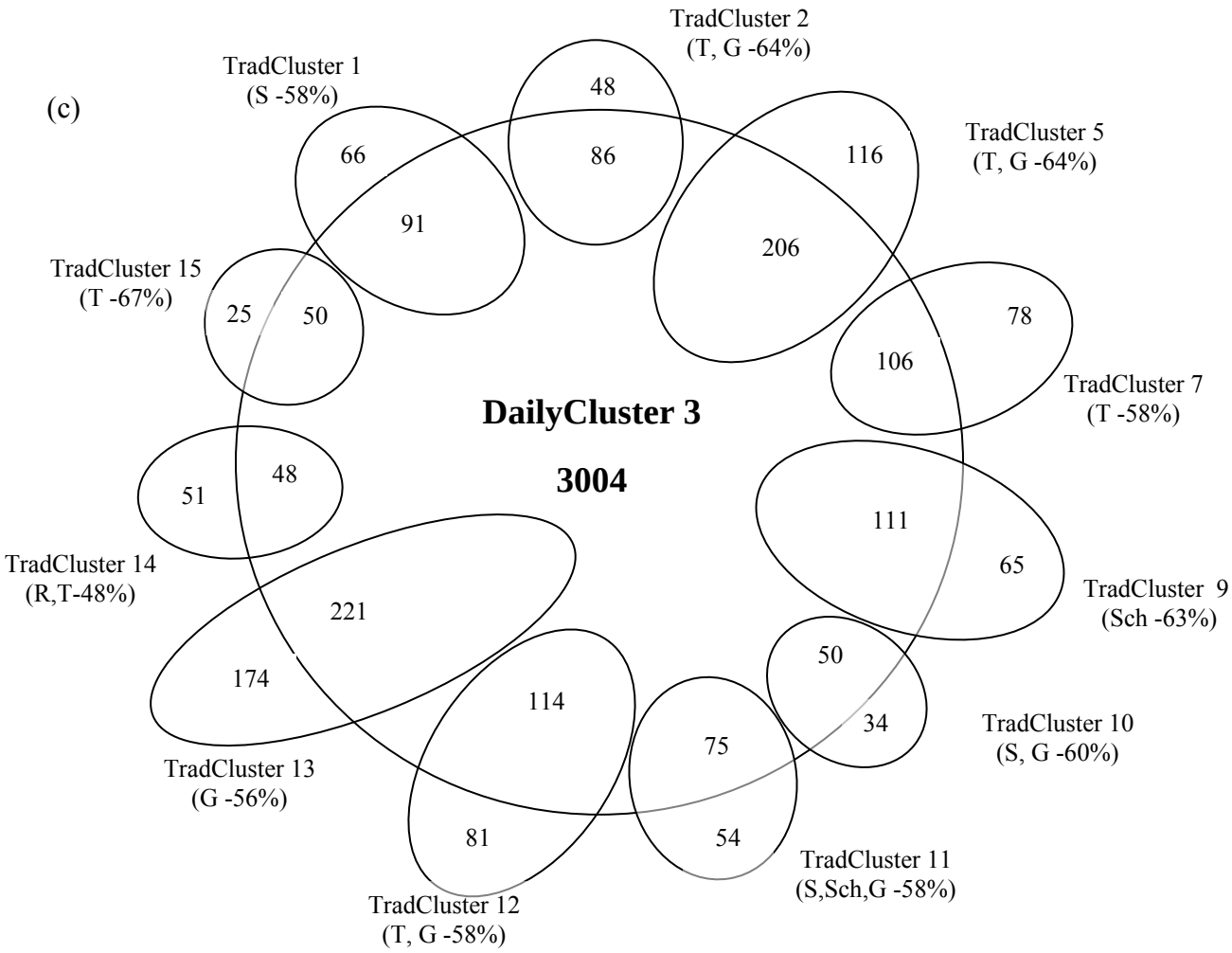


Figure 6.3a-c: Venn diagram of Traditional k-means clustered data of Leroch et al. (2003) and NMF clustered data of Daily et al. 2007.

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 6.2: MMk-means and Traditional k-means clusters with their equivalent corresponding clusters in Le Roch et al., (2003).
(See page 112 for table description)

Cluster ID (k=15)	MMk-means Member Count	MMk-means Diff. Exp count (a)	Tradk-means Member Count	Tradk-means Diff. Exp count (b)	No of MMk-means Genes in Equiv. Le Roch cluster (c)	No of Tradk-means Genes 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 6.3: Analysis of traditional k-means Clustered data of Le Roch *et al.* 2003 and NMF clustered data of Daily *et al.* 2007

(See page 112 for table description)

TRADk-means Cluster ID	Traditional k15 Diff Exp Gene count (a)	Stages	DAILY07_CLST1 vs tradkmeansLEROCH03			DAILY07_CLST2 vs tradkmeansLEROCH03			DAILY07_CLST3 vs tradkmeansLEROCH03		
			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	157	Sporozoite	55	102	35%	102	55	65%	91	66	58%
2	134	Trophozoite, Gametocyte	44	90	33%	86	48	64%	86	48	64%
3	142	Ring, Trophozoite	33	109	23%	73	69	51%	48	94	34%
4	67	Ring,Schizont, Merozoite	16	51	24%	14	53	21%	9	58	13%
5	322	Trophozoite, Gametocyte	101	221	31%	196	126	61%	206	116	64%
6	10	Sporozoite	3	7	30%	2	8	20%	1	9	10%
7	184	Trophozoite	67	117	36%	113	71	61%	106	78	58%
8	66	Ring, Trophozoite	12	54	18%	18	48	27%	15	51	23%
9	176	Schizont	60	116	34%	109	67	62%	111	65	63%
10	84	Sporozoite, Gametocyte	22	62	26%	53	31	63%	50	34	60%
11	129	Sporozoite, Schizont, Gametocyte	55	74	43%	82	47	64%	75	54	58%
12	195	Trophozoite, Gametocyte	59	136	30%	111	84	57%	114	81	58%
13	395	Gametocyte	106	289	27%	255	140	65%	221	174	56%
14	99	Ring,Early Trophozoite	33	66	33%	56	43	57%	48	51	48%
15	75	Trophozoite	28	47	37%	43	32	57%	50	25	67%
	2235		694	1541		1313			1231		

Table 6.4 : Analysis of Mmkmeans Clustered data of Le Roch *et al.* 2003 and NMF clustered data of Daily *et al.* (2007)

(See page 113 for table description)

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		

Table 6.5: Analysis of Robust k-means Clustered data of Le Roch *et al*. 2003 and NMF clustered data of Daily *et al.* 2007

(See page 113 for table description)

Le Roch Cluster ID	Gene Count (a)	Stages	DAILY07_CLST1 vs robustLEROCH03			DAILY07_CLST2 vs robustLEROCH03			DAILY07_CLST3 vs robustLEROCH03		
			No of Genes Present in DailyCLST 1 (b)	No of Genes Absent in DailyCLST1 (a-b)	% age of Genes Present in DailyCLST 1 (b/a)	No of Genes Present in DailyCLST 2 (c)	No of Genes Absent in DailyCLST 2 (a-c)	% age of Genes Present in DailyCLST 2 (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	108	Sporozoite	33	75	31%	73	35	68%	65	43	60%
2	152	Sporozoite, Gametocyte	60	92	39%	104	48	68%	98	54	64%
3	218	Gametocyte	42	176	19%	126	92	58%	115	103	53%
4	95	Ring,Schizont, Merozoite	23	72	24%	19	76	20%	12	83	13%
5	109	Ring,Early Trophozoite	34	75	31%	67	42	61%	53	56	49%
6	167	Ring, Trophozoite	47	120	28%	76	91	46%	65	102	39%
7	93	Ring, Trophozoite	17	76	18%	44	49	47%	28	65	30%
8	125	Trophozoite, Gametocyte	37	88	30%	68	57	54%	81	44	65%
9	120	Trophozoite, Gametocyte	32	88	27%	82	38	68%	71	49	59%
10	226	Trophozoite, Gametocyte	67	159	30%	136	90	60%	145	81	64%
11	110	Sporozoite,Trophozoite,Schizont, Gametocyte	46	64	42%	53	57	48%	68	42	62%
12	238	Trophozoite	99	139	42%	160	78	67%	145	93	61%
13	181	Schizont, Gametocyte	54	127	30%	116	65	64%	113	68	62%
14	163	Sporozoite, Schizont, Gametocyte	59	104	36%	111	52	68%	99	64	61%
15	130	Schizont	44	86	34%	78	52	60%	73	57	56%
	2235		694			1313			1231		

Description of Tables 6.2-6.5

Table 6.2: MMk-means and Traditional k-means clusters with their equivalent corresponding clusters in Le Roch et al., (2003).

This table portrays the similarity of mm and traditional k-means respectively to Robust k-means. The MMk-means and Traditional k-means have the 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 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 MMk-means 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 MMk-means percentage similarity with Le Roch et al. clusters respectively (columns 9 and 10) shows positive correlation with a value of 0.7.

Table 6.3: Analysis of traditional k-means Clustered data of Le Roch *et al.* 2003 and NMF clustered data of Daily *et al.* 2007.

This table portrays the percentage of genes from traditional k-means clustering analysis for *in-vitro* Le Roch *et al.* data compared to that present in NMF clustered Daily et al. *in-vivo* data. Col= Column, CLST = Cluster, tradkmeansLEROCH03= Clusters obtained from Traditional k-means on Le Roch *et al.* data, DAILY07_CLST= Daily *et al.* clusters. Col 1= TRADk-means Cluster ID (15 clusters from Traditional k-means), Col 2 = Traditional k15 Diff Exp Gene count (a) (No of genes in each cluster for only 2235 highly expressed genes), Col 3=Stages (Prevailing parasite stages description for each cluster), Col 4= DAILY07_CLST1 vs tradk-meansLEROCH03 (Containing No of genes present in both Traditional k-means clustered Le Roch *et al*. data and Daily *et al.* Cluster 1, also percentage No of genes from Traditional k-means Clusters present in Daily Cluster 1) , Col 5= DAILY07_CLST2 vs tradk-meansLeroch03 (containing No of genes present in both Traditional k-means clustered

Le Roch *et al*. data and Daily *et al.* cluster 2, also percentage No of gene from Traditional k-means clusters from Le Roch *et al.* data present in Daily Cluster 2), Col 6= DAILY07_CLST3 vs tradkmeansLEROCH03 (Containing No of genes present in both Traditional k-means clustered Le Roch *et al.* data and Daily et al. cluster 3, similarly, percentage No of genes from Traditional k-means on Le Roch data Clusters and present in Daily Cluster 3).

Table 6.4 : Analysis of Mmkmeans Clustered data of Le Roch *et al.* 2003 and NMF clustered data of Daily *et al.* (2007).

This table portrays the percentage of genes from MMk -means clustering analysis of *in-vitro* Le Roch data that are present in Daily et al. *in-vivo* data. Col= Column, CLST = Cluster, MMk-meansLeRoch03= Clusters from MMk-means on Le Roch data, DAILY07_CLST= Daily *et al.* clusters. Col 1= MMk-means Cluster ID (15 clusters from MMk-means), Col 2= MMk-means k15 Diff Exp Gene count (a) (No of genes in each cluster for only 2235 highly expressed genes), Col 3=Stages (Prevailing parasite stages description for each cluster), Col 4= DAILY07_CLST1 vs MMk-meansLeroch03 (Containing No of genes present in both MMk-means clustered Leroch data and Daily Cluster 1 and percentage No of gene from MMk-means Clusters present in Daily Cluster 1) , Col 5= DAILY07_CLST2 vs Mmk-meansLeroch03 (containing No of genes present in both MMk-means clustered Le Roch et al. data and Daily et al. cluster 2, also percentage No of gene from MMk-means Clusters from Le Roch data present in Daily Cluster 2), Col 6= DAILY07_CLST3 vs MMk-meansLeroch03 (Containing No of genes present in both MMk-means clustered Le Roch et al. data and Daily cluster 3, similarly, percentage No of genes from MMk-means Le Roch data Clusters present in Daily Cluster 3).

Table 6.5: Analysis of Robust k-means Clustered data of Le Roch *et al*. 2003 and NMF clustered data of Daily *et al.* 2007.

This table portrays the percentage of genes from Robust k-means clustering analysis of *in-vitro* Le Roch *et al.* data that are present in Daily *et al. in-vivo* data. Col= Column, CLST = Cluster, robustLEROCH03= Clusters from Robust k-means, DAILY07_CLST= Daily *et al.* clusters. Col 1= Le Roch Cluster ID (15 clusters from Robust k-means), Col 2= Gene count(a) (No of genes in each cluster for only 2235 highly expressed genes), Col 3=Stages (Prevailing parasite stages description for each cluster), Col 4= DAILY07_CLST1 vs robustLEROCH03 (Containing No of genes present in both Leroch and Daily cluster 1 and percentage No of gene form Le Roch Clusters present in Daily Cluster 1) , Col 5= DAILY07_CLST 2 vs robustLEROCH03 (containing No of genes present in both Leroch and Daily cluster 2 and percentage No of gene from Le Roch Clusters present in Daily Cluster 2), Col 6= DAILY07_CLST3 vs robustLEROCH03 (Containing No of genes present in both Leroch and Daily cluster 3 and percentage No of genes from Le Roch Clusters present in Daily Cluster 3).

6.3 DISCUSSION

The correlation coefficient of 0.7 computed from Table 6.2 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 (see Figure 6.1a-c to 6.3a-c), furthering the authentication of the accuracy of MMk-means algorithm.

Based on the average of 0.54 spearman rank correlation, Daily *et al.* (2007) 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.* (2003). We obtained this as shown in Figure 6.1b, where we obtained 20%, 61%, 46%, and 47% similarity respectively for each of the 4 clusters indicated to contain genes that coded for the ring-stage of the parasite.

We also verified Daily *et al.* (2007) 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 in Figure 6.1a 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 6.1c), because 11 clusters out of 15 *in-vitro* clusters formed a 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 exists in the *in-vitro* profiling data of Le Roch *et al.* (2003)

6.4 CONCLUSION

This work authenticates our new and novel MMk-means algorithm and also delivers a biological viable result that is missing in Daily *et al.* (2007) results.

CHAPTER SEVEN

EXPLORING PCR-BASED DETECTION OF MALARIA INFECTION AT THE LIVER STAGE

7.1 INTRODUCTION

Waiting for the detection of malaria at the blood stage can lead to delayed treatment that may engender serious complication and death. The attachment of erythrocytes infected with *Plasmodium falciparum* to the microvessels of the brain leads to a pathological condition known as cerebral malaria that can result in death. There are no effective therapeutic means for alleviating this pathology (Land et al., 1995) as adhesin proteins on the surface of the parasite-infected red blood cell aid malaria disease complication. It is therefore imperative to note that many lives will be saved if these parasites can be detected and treated at the asymptomatic liver stage instead of waiting till the disease manifestation at blood stage. Having introduced the concept of Polymerase Chain Reaction (PCR) in chapter 3, in this chapter, we explored the basis of using PCR to detect malaria at the liver stage.

Malaria transmission involves three different developmental stages namely: the human liver stage, the human blood stages and the mosquito stage. Symptoms of malaria are expressed at the human blood stage. Generally, there is no doubt that there are some available drugs that can cure the diseases, but the problem in most cases is poor or late diagnoses resulting in complications and even death. The rationale for this study is to explore a diagnostic technique for detecting malaria at the liver stage, so that timely intervention can be made to alleviate the problem of the disease. Diagnostics on biochip has been making in-road into modern healthcare at a faster pedestal especially Point-of-Care than the lab-based diagnostics. The ultimate breakthrough may be to translate the result to a liver-based malaria diagnostics chip comparable to diagnostic chip used for detecting other diseases like HIV.

The use of microscopic examination of Giemsa-stained blood smears remains the cheapest and most commonly used method for the malaria diagnosis at the blood stage. Microscopy

has its own bottle-necks in that its sensitivity is limited particularly when parasitaemia is low or when parasite morphology is altered. It is also time-consuming (Coleman *et al*., 2002) and requires a highly technical expert. There is no way to determine early invasion and infection since malaria diagnosis is carried out at the human blood stage of the parasite which accompanies the manifestation of the symptom for reported cases. Since the human liver stage is asymptomatic and precedes the blood stage, early diagnosis of the parasite at the liver stage will help intercept the havoc timely, through the use of some vaccines / drugs to abort the progression from liver stage to blood stage and thereby eradicate the malaria parasite at liver stage.

In-vivo experimental access to liver stages of human malaria parasites is practically prohibited and therefore rodent model malaria parasites have been used for *in-vivo* studies. However, genome-wide liver stage (LS) gene expression was profiled by using Green fluorescent protein-tagged *Plasmodium yoelii* (PyGFP) to efficiently isolate Liver Stage infected hepatocytes from the rodent host (Tarun et al., 2008). Our interest is to use this only known microarray information on liver stage of *P. yoelli* to find orthologue genes in *P. falciparum,* which can be used to develop a PCR experiment useful for the detection of malaria at the liver stage.

7.2 COMPILATION OF PROTEINS RELEVANT FOR THE PARASITE SURVIVAL AT LIVER STAGE

The compilation of the proteins that has been described to date as being relevant for the parasite's survival and development in the hepatic stage is represented either in the micronemes/rhopteries or surface of the sporozoite as shown in Figure 7.1 (Garcia et al., 2006). These include:

(1) Circumsporozoite (CS),
(2) Thrombospondin-Related Anonymous Protein (TRAP)
(3) Sporozoite Threonine-And Asparagine-Rich Protein (STARP)
(4) Liver-Stage Antigen 1 (LSA-1)
(5) Liver-Stage Antigen (LSA-3)

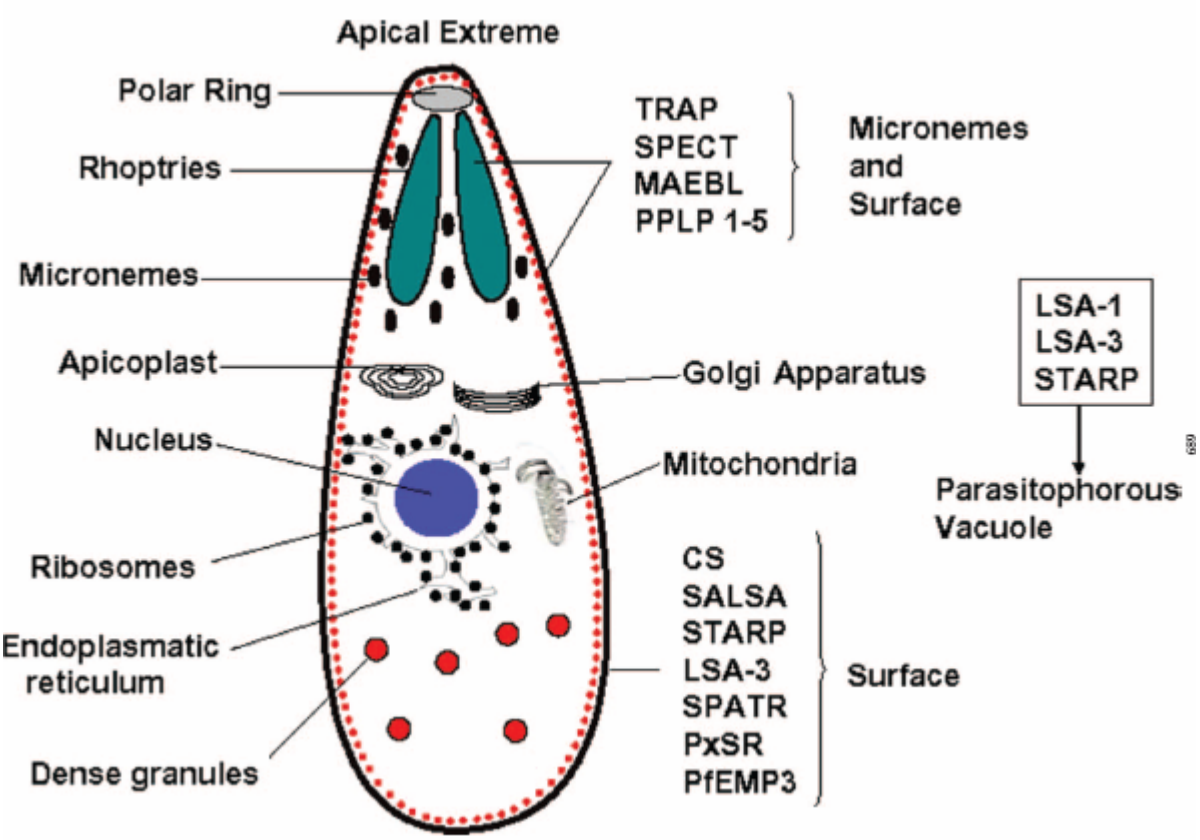
(6) Sporozoite And Liver-Stage Antigen (SALSA)

(7) Sporozoite Microneme Protein Essential For Cell Transversal (SPECT),

(8) Spect2/*Plasmodium* Perforin-Like Protein 1 (PPLP)

(9) Apical Membrane 3 Antigen/ Membrane Apical Erythrocyte Binding-like Erythrocyte Binding-Like Protein (MAEBL)

(10) *P. falciparum* Secreted Protein with Altered Thrombospondin Repeat (PfSPATR)



Source: (Garcia et al., 2006)

Figure 7.1: Diagram of Internal Structure of a Sporozoite.

The micronemes and surface at the anterior region contains TRAP, SPECT MAEBL and PPLPs while the surface abdominal region contains CSP, SALSA, LSA-3, SPATR, PxSR, PfEMP. LSA-1, LSA-3 and STARP are found in the parasitophorous vacuole

Liver stage parasite protein regions that are able to induce an immune cellular response have been reported; some high-binding-activity peptide (HABP) sequences identified in these proteins contain T epitopes. Proteins such as CS, LSA-1, LSA-3, SALSA, STARP,

and TRAP possess T epitopes and could be relevant in inducing an effective immune response against the parasite (Garcia *et al*., 2003; Garcia *et al*., 2004; Garcia *et al*., 2006, Lo´pez *et al*.,2001; Lopez, *et al*. 2003; Puentes, *et al*. 2004; Suarez, *et al*. 2001).

We give details as regards each as follows:

Circumsporozoite (CS): The CS protein is found in all the different *Plasmodium* parasites and is localized on the sporozoite surface. The genes encoding the CS protein are localized in *P. falciparum* genome chromosome 3 and have been clearly characterized (Ancsin and Kisilevsky, 2004). CS protein is an important multifunctional molecule for the parasite, fulfilling different roles (depending on the developmental stage) that are vital for the parasite's development. It has been reported that the CS protein is involved in the invasion of a mosquito's salivary glands sporozoite binding to liver cells and inactivating host cell protein synthesis machinery.

Thrombospondin-Related Anonymous Protein (TRAP): TRAP belongs to a family of functionally homologous proteins involved in parasite mobility due to gliding and cell penetration, suggesting that it could be a mediator for parasite ligand interactions with substrate receptors involved in parasite mobility and with host receptors involved in invasion. TRAP has a cytoplasmatic tail whose primary sequence is not conserved in the different *P. falciparum* isolates but shows special characteristics: it is acid residue rich (18 to 30%) and contains tryptophan in the penultimate or antepenultimate residue. Besides, TRAP is liberated onto the substrate during locomotion due to gliding, and during this movement.

Sporozoite Threonine-And Asparagine-Rich Protein (STARP): STARP is a molecule that is consistently expressed on sporozoite surface and was identified and cloned for the first time in *P. falciparum* laboratory strains and field isolates (Fidock et al., 1994) obtained from a broad range of regions where malaria is endemic. It is also present in other *Plasmodium* species and has a highly conserved structure (Garcia et al., 2006). Immunofluorescence and immunoelectron microscopy assays using immune sera directed

against the protein's central and C-terminal regions have revealed that STARP is expressed on the sporozoite surface during the intrahepatic stage and has also been detected in early ring stages.

Liver-Stage Antigen 1 and 3: The LSA-1 protein has a molecular mass of 240 kDa and contains 1,909 amino acids. The LSA-1 protein is expressed during early trophozoite and mature parasite stages (Garcia et al 2006). It is localized around developed schizont and is found in the parasitophorous vacuole surrounding exoerythrocytic merozoites. The *P. falciparum* LSA-3 protein LSA-3 was the first preerythrocytic-stage protein used to test immunological responses among volunteers immunized with irradiated sporozoites. While the immunised volunteers became protected, unimmunised people were not protected from the disease (Daubersies et al., 2000). LSA-3 has a 200-kDa molecular mass and has 1,786 amino acids. The LSA-3 protein is expressed in sporozoites and is present both in the parasitophorous vacuole and on the periphery of mature hepatic merozoites.

Sporozoite And Liver-Stage Antigen (SALSA): Garcia et al., (2006) reported that SALSA protein synthesis begins during the sporozoite stage, possibly during the maturation process in a mosquito's salivary glands; its production increases during hepatic schizogony. Bottius et al. (1996), using the PCR DNA amplification technique for the gene encoding SALSA in seven culture-adapted strains (five Asiatic and two African strains) and 16 isolates from Senegal, suggested that the SALSA protein is completely conserved among *P. falciparum* isolates and pointed out the scarce homology between SALSA and other *P. falciparum* antigens.

Sporozoite Microneme Protein Essential For Cell Transversal (SPECT): However, SPECT is a protein that has been involved in sporozoite infection of liver cells and localized on sporozoite micronemes, suggesting that it is possibly involved in sporozoite mobility during invasion. *In vitro* cell invasion assays have shown that sporozoites containing a deletion of the gene encoding SPECT (spect-disrupted parasites) completely lose their ability to pass through Kupffer cells; however, these sporozoites preserve their normal ability to infect hepatocytes, suggesting that SPECT is involved in the ability of sporozoites to pass through cells. Disruption of the gene encoding SPECT does not affect

parasite proliferation during the intraerythrocyte stage or a mosquito's development (Ishino et al., 2004).

***Plasmodium* Perforin-Like Protein 1:** SPECT2/*Plasmodium* Perforin-Like Protein 1 is localized on sporozoite micronemes, suggesting that this protein could be involved in invasion (Ishino et al., 2004). Immunofluorescence assays have suggested that this protein's expression is restricted to the salivary gland sporozoite stage. Interrupting the gene encoding the SPECT2 protein in sporozoites does not affect parasite development in the mosquito or the number of sporozoites residing in its stomach and salivary glands (Ishino et al., 2004).

Membrane Apical Erythrocyte Binding-like Erythrocyte Binding-Like Protein (MAEBL) and *P. falciparum* Secreted Protein with Altered Thrombospondin Repeat (PfSPATR) : Blair et al., (2002) noted that MAEBL protein is localized on the salivary gland sporozoite surface, in free merozoites, and in late-stage schizonts. Parasites that have the disrupted gene encoding *P. berghei* MAEBL lose their ability to infect salivary glands. However, they maintain their mobility due to gliding, indicating that MAEBL is not essential for *in-vitro* mobility (Kariu et al., 2002). PfSPATR immunoelectron microscopy studies have shown that PfSPATR protein is localized on the sporozoite surface and around the parasite's rhoptries during the erythrocyte asexual stage and on the infected erythrocyte membrane (Chattopadhyay, 2003).

Generally, it should be observed that some of the proteins found in the hepatic stage of the parasite life cycle may also occur at the invertebrate host stage as well as the red blood cell stage. An investigator on one stage-specific protein that could be useful for PCR must first identify proteins that occur or are highly expressed at only one stage of the parasite development. Oyedeji et al., (2007) identified three genes which express themselves at blood stage of malaria infection for their PCR-based comparative malaria diagnosis.

7.3 EXISTING MALARIA DIAGNOSTIC TOOLS

The limitations of diagnosing malaria by light microscopy of Giemsa-stained smears have led to the development of several new techniques (Hänscheid, 1999) that aim to simplify and speed up diagnosis and increase sensitivity. Results have been obtained using fluorescent dyes (eg. with the quantitative buffy coat, QBC®) (Levine *et al.,* 1989) and simple dipstick tests to detect various antigens (World Health Organisation, 1996; Makler *et al.,*1998) as well as with PCR, regarded as the new reference method because of its superior sensitivity and specificity (Oyedeji et al. 2007; Snounou *et al.*, 1993).

Tham et al., (1999) described a sensitive and reliable two-step PCR-based amplification assay for the diagnosis of malaria at blood stage. *Plasmodium* infections were diagnosed by use of a genus-specific primer set, with two distinct primer sets designed to specifically detect either *P. falciparum* or *Plasmodium vivax*. Quantitative Buffy Coat (QBC; Becton Dickinson) analysis for malaria, which is a fluorescent microscopic examination of capillary-centrifuged blood, was performed in tandem with the thick-film Giemsa stain analysis. Their blood samples were also assayed with two commercially available test kits ParaSight-F (Becton Dickinson) and ICT Malaria Pf (ICT Diagnostics). Both test kits are based on immunological detection of the *P. falciparum* histidine-rich protein 2 and PCR performed beter as it recorded no false positive or false negative.

Tham et al., (1999) strategy for the PCR amplification was the detection of a malaria infection with genus-specific primers made from the conserved large-subunit rRNA gene, and detection of *P. falciparum* and *P. vivax*, done with primers made from the *coxI* gene. These primers were then used in a multiplex PCR system. The 18S rRNA gene has been used as a DNA target for the differentiation of plasmodial species by nested PCR (Snounou, 1996; Tahar et al., 1997) and reverse transcription-PCR. Other DNA targets such as the circumsporozoite protein gene (Sethabutr et al., 1992; Tahar et al., 1997) have also been investigated for species-specific regions. Tan et al. (1997) demonstrated that the large-subunit rRNA gene is extensively conserved within *Plasmodium* species and is suitable as a genus-specific DNA target region.

Hänscheid et al., (2000) noted that, all the above stated tests have the inherent disadvantage that they have to be specifically requested by an alert clinician who suspects the presence of disease. Absence of this suspicion is a main reason for high numbers of misdiagnoses (Kain *et al.*, 1998). Standard automated laboratory tests such as automated full blood counts, (FBCs), routinely performed in the analysis of febrile patients in many countries, have not been found to help significantly in a specific diagnosis of malaria (Marshall *et al.*, 1990; Giacomini *et al.*, 1991). Of concern is the number of false positives with the FBC analyser, due to persisting haemozoin-containing white blood cells (WBCs) in the circulation (Hänscheid et al., 2000).

A number of *P. falciparum* polymorphic genetic markers namely merozoite surface protein 1 (MSP-1), merozoite surface protein 2 (MSP-2) and glutamine rich protein (GLURP) are the most widely used in PCR, but several different protocols have been independently developed. It is known that differences in protocols (sample collection, storage, DNA extraction, amplification conditions and detection of product) can influence the specificity and sensitivity of the PCR amplification. Thus comparison of results from different studies is at present limited (Björkman et al., 1998).

Results from the PCR-based assay indicate that the *stevor* gene amplification is the most sensitive technique for detection of *P. falciparum* (Oyedeji et al., 2007). Oyedeji *et al.* (2007) conducted PCR of *stevor*, *SSUrRNA (Small Subunit rRNA)* and *MSA2 (Merozoite Surface Antigene)* genes for comparison and assessment of PCR-based detection of *P.falciparum.* It was reported that *stevor* gene amplification has the highest sensitivity hence the most sensitive technique for the parasite detection (Oyedeji *et al.,* 2007).

Silvie et al., (2008) did not work directly on malaria diagnosis but investigated the conserved *Plasmodium* asparagine-rich protein specifically expressed in sporozoites and early liver stage, and was therefore termed SLARP (Sorozoite and Liver stage Asparagine-Rich Protein). Using PCR, Silvie et al. (2008), showed that SLARP controls the initiation of *Plasmodium* liver stage development.

Bruna-Romerio et. al., (2001) adapted a real-time PCR to develop an assay for detection and quantification of the liver stage of *P. yoelli* parasite in mice infected through the bite of a single *Anopheles* mosquito. Bruna-Romerio et al. (2001) was unable to compare his result with a similar assay done by McKenna et al. (2000) due to the unfortunate absence of basic information regarding the sequence of primers and probes, as well as reaction conditions.

In view of the limitation of Bruna-Romerio et al. (2001), determination of important genes, is key before primers and probes can be produced for successful PCR that can be used for malaria detection at liver stage. Notwithstanding the importance of early and accurate diagnosis in malaria treatment discovery, most of the existing diagnostics gave little attention to malaria detection at the liver stage, hence the need to explore detection at this level of the parasite life cycle.

7.4 METHODOLOGY AND RESULTS

Not much microarray work has been done on the liver stage of *Plasmodium* parasite. However, one promising microarray work was done by Tarun et al. (2008) on the liver stage of *P. yoelli* parasite in mice. To deeply analyse the behavior of parasite genes at liver stage, we employed the use of the microarray data of Tarun et al. (2008) and searched for their orthologues in *P. falciparum* using the orthoMCL algorithm in PlasmoDB (Kissinger et al., 2002). Our interest is to further analyse the behaviour of these liver stage genes using some knowledge obtained from blood stage of *P. falciparum* 3D7 and HB3 strains from the microarray data of Bozdech et al. (2003a). This idea lends credence to the role of orthologues in functional genomics, as genes in a different species that evolved from a common ancestral gene by speciation, retain the same function in the course of evolution (Lewis, 2009).

7.4.1 DATA USED

We used the data listed in Table 7.1 below. Tarun et al. (2008) isolated liver stage-infected hepatocytes from *P. yoelli* Green florescent Protein (PyGFP)-infected mice and Sporozoites from P. yoelii-infected *A. stephensi* mosquitoes were isolated from midguts at day 10 and from salivary glands at day 15 after infectious blood meal. Tarun et al. (2008) represented 1985 genes on a microarray slide and performed a high-throughput experiment on *P. yoelli* under 18 different experimental conditions. Since our interest is to study the behaviour of *P. falciparum* genes at the liver stage whose orthologues are represented in *P. yoelli* data (Tarun et al., 2008), we employed the expression data of Bozdech et al. (2003) for 3D7 and HB3 strains. This is to identify *P. falciparum* genes that have an extensively different behaviour at the liver stage from what is obtained at the blood stages.

Bozdech et al. (2003a) microarray experiment used lab cultered *P. falciparum* as in section 5.3.4.1 to describe a complete asexual intraerythrocytic developmental cycle (IDC) of 3D7 and HB3 strains such as early ring stage, late ring stage, early trophozoite stage, late trophozoite stage, early schizont stage, late schizont stage and gametocyte stage. By analyzing the IDC transcriptome of the 3D7 strain and HB3 strain (Llinás et al., 2005) of *P. falciparum* for 4596 and 4313 genes respectively, they were able to demonstrate that at least 60% of the genome is transcriptionally active.

Table 7.1: Microarray data of *P. yoelli* and *P. falciparum* with *P.yoelli* orthologues

Experimental data	**Total No Of Genes**	**Time points**
Bozdech et al, (2003)- *P. falciparum* 3D7 strain data	4596	53
Bozdech et al., (2003) – *P. falciparum* HB3 strain data	4313	46
Tarun et al, (2008)- *P. yoelli* data	1985	18
P.yoelli orthologue in *P. falciparum* 3D7	1180	53
P.yoelli orthologue in *P. falciparum* HB3	1163	46

Bozdech et al. (2003a) microarray experiment described a complete asexual intraerythrocytic developmental cycle (IDC) of 3D7 and HB3 strains of P. *falciparum* such as early ring stage, late ring stage, early trophozoite stage, late trophozoite stage, early schizont stage, late schizont stage and gametocyte stage. By analyzing the IDC transcriptome of the 3D7 strain and HB3 strain of *P. falciparum* for 4596 and 4313 genes respectively, they were able to demonstrate that at least 60% of the genome is transcriptionally active. Tarun et al., (2008) represented 1985 genes on a microarray slide and performed a high-throughput experiment on *P. yoelli.* Searching PlasmoDB for the orthologues of these 1985 genes, we obtained 1180 for *P. yoelli* orthologue in *P. falciparum* 3D7 and 1163 *P. yoelli* orthologue in *P. falciparum* HB3 from Bozdech et al., (2003a).

7.4.2 SEARCHING AND MAPPING GENES FOR ORTHOLOGUES

7.4.2.1 Preliminary Search for Genes that Code for Specific Liver Stage Proteins

A preliminary search on the PlasmoDB for genes that code for each protein responsible for parasite survival and development in the liver stage as discussed in section 7.2 was conducted. The result produced the following genes and/or orthologues depicted in Table 7.2. Each protein was searched individually and no gene retrieved for SALSA and SPEC2/PPLP. This may be that no corresponding gene was submitted to the gene bank repository or the orthoMCL algorithm failed to retrieve the concerned genes.

The associated genes that code for *Plasmodium* proteins were verified to assess the level of their representation on Tarun et al. (2008) microarray slide. Surprisingly, many of the retrieved genes were not represented on the array.

Table 7.2: Liver Stage *Plasmodium* Proteins and their Coding Genes

	***PLASMODIUM* PROTEINS**	**ASSOCIATED GENES FROM PLASMODB**
1	Circumsporozoite (CS)	PY01663, PY03168, PY07368
2	Thrombospondin-Related Anonymous Protein (TRAP)	PY01828, PY02417, PY02475, PY03378, PY04302, PY05174.
3	Sporozoite Threonine-And Asparagine-Rich Protein (STARP)	PY00217, PY05105
4	Liver-Stage Antigen 1 (LSA-1)	PF10_0356 (PY02217, PY02913, PY03361, PY04214, PY04691, PY05279)
5	Liver-Stage Antigen 3 (LSA-3)	PFB0915w
6	Sporozoite And Liver-Stage Antigen (SALSA)	None
7	Sporozoite Microneme Protein Essential For Cell Transversal (SPECT)	PF13_0197 (PY02149), PFC0755c (PY00154), PFF1325c (PY04641), PFF1370w, (PY05545)
8	Spect2/*Plasmodium* Perforin-Like Protein 1 (PPLP)	None
9	Apical Membrane 3 Antigen/ Membrane Apical Erythrocyte Binding-like Erythrocyte Binding-Like Protein (MAEBL)	PY01844, PY02049, PY03020, PY03552, PY04797
10	*P. falciparum* Secreted Protein with Altered Thrombospondin Repeat (PfSPATR)	PY02498, PY02991, PY04732

Each protein name is used as a search keyword on the PlasmoDB website. Many of the proteins retrieved more than one genes that code them, signifying that more than one gene is responsible for coding a protein.

7.4.2.2 Multiple Search for Orthologues Using Gene List

We searched PlasmoDB for the orthologues of 1985 *P. yoelli* genes represented on Tarun et al., (2008) microarray and obtained 1459 *P. yoelli* orthologues in *P. falciparum* 3D7, HB3 and Dd2 strains. We mapped and extracted the expression values of these *P. yoelli* orthologues from Bozdech et al., (2003a) data and obtained 1180 genes for *P. yoelli* orthologue in *P. falciparum* 3D7 and 1163 genes for *P. yoelli* orthologue in *P. falciparum* HB3. We did not use Dd2 strain because the extracted orthologues were very few.

In addition, the *P. yoelli* orthologues found were represented in a venn diagram (Osamor et al., 2009) as depicted in Figure 7.2. It shows that 1139 genes were found to be represented in both *P. falciparum* 3D7 and *P. falciparum* HB3 strains.

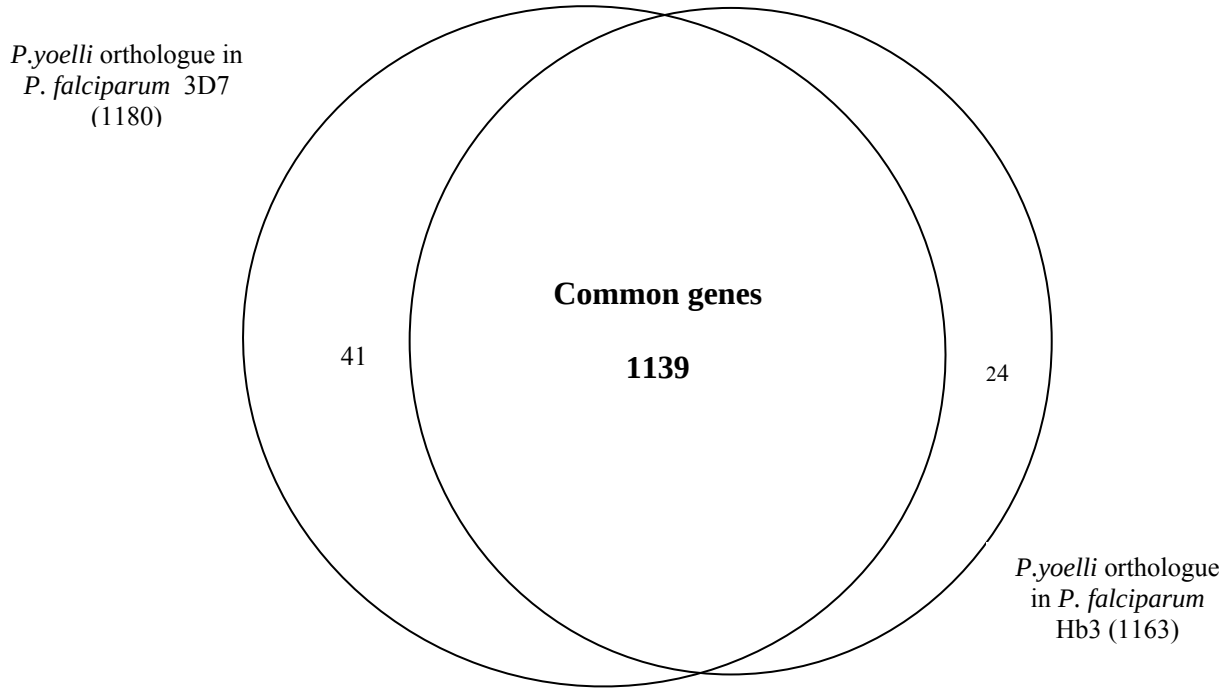


Figure 7.2: Venn Diagram Showing The Number of Common Genes for *P.yoelli* orthologues in Two Strains of *P. falciparum.*

Using the *P. yoelli* genes represented on microarray data of Tarun et al. (2008), *P. yoelli* orthologues in *P. falciparum* 3D7 and HB3 strains were retrieved from the microarray data of Bozdech et al. (2003a).

7.4.3 TRADITIONAL K-MEANS CLUSTERING AND GENES EXPRESSION SIGNIFICANCE TEST

The Traditional clustering algorithm implemented in Osamor et al., (under review) and in chapter five was deployed and used to cluster the 1985 genes of Tarun et al., (2008) microarray data and 1139 *P. falciparum* 3D7 genes and 1139 *P. falciparum* HB3 genes from Bozdech et al., (200a3) microarray data independently. Using guilty by association (GBA) principle (Le Roch et al. (2003), genes in the same cluster are expected to be functionally related and orthologues of Tarun et al. (2008) genes in the same cluster using *P. falciparum* 3D7 and HB3 strains expression are expected to be key genes that maybe

useful for designing a PCR experiment for the detection of malaria at the liver stage. The number of cluster input was set at k = 15 and the resultant output was exported to MS Access relational database management system (RDBMS) for further analysis and query. Based on a simple select query and crosstab query generated from the cluster output, we obtained the information contained in Table 7.3, which is a 15 x 15 matrix comparing and counting the number of genes common to each pair of clusters for *P. falciparum* 3D7 and HB3 strains expression data as extracted above. The diagonal elements represent the number of genes in two clusters (Osamor *et al.*, 2009; Osamor *et al.*, under review) when the cluster ID's are the same.

Table 7.3 Common Genes Matrix for *P. yoelli* orthologues in *P. falciparum* 3D7 and HB3 Strains Clusters

		P. falciparum HB3 Orthologue Clusters															
	Cluster No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total Of Gene ID
P.falciparum 3D7 Orthologues Clusters	1	**63**	24	1	2	3			3	1	1		33	1	1	7	140
	2	20	**27**	1		1	1	4	32		1		7	5	1	12	112
	3	3	2	**1**	3	5	1	3	5		1	1	2		1	1	29
	4	4	1			4			1		3		4	2			19
	5	2		30	4	**3**		2		4	2		2	1	1	1	52
	6	3		3	25	14			1		3		7		2		58
	7	7	2	2	1	3	8	**3**	32			4	6	60	11	10	149
	8	26	2	2	1	5			**1**	2		2	67	2	2	2	114
	9			1	1	2	1	1		**3**	1	3			1		14
	10	10	10			3			6			1	6	4	2		42
	11	3	2	5	1	4	1	2	3	38	6	**31**	4	1	10		111
	12	3	2	1		3		3	1		2	5	**2**	3	4	5	34
	13			1		5					2		2		1		11
	14	9		2	14	12			3	3	1	1	55	2	**1**	1	104
	15	3	5	2		3	1	3	5	3	1	26		29	65	**4**	150

Using the genes represented in Tarun et al. (2008), *P. yoelli* orthologues in *P. falciparum* Pf3D7 and HB3 strains were retrieved from microarray data of Bozdech et al. (2003) and classified into 15 clusters using traditional k-means implemented in Osamor et al. (under review). This resulted in a 15x15 matrix showing the number of common genes in every two clusters of *P. falciparum* 3D7 and *P. falciparum* HB3 orthologues. The diagonal elements represent the number of common genes in two clusters with equal ID number.

Table 7.4: Comparative Table for *P. yoelli* orthologues in *P.falciparum*

TRAD k-means Cluster ID	Clusters of P.yoelli from Tarun et al.		P.y orthologues in P.f 3D7 of Bozdech et al.		P.y orthologues in P.f HB3 of Bozdech et al.			
	No of Genes Present/ Cluster	Percentage of Genes Present (T1)	No of Genes Present/ Cluster (T2)	Percentage of Genes Present using the Total Gene clustered	No of Genes Present/ Cluster (T3)	Percentage of Genes Present using the Total Genes clustered	No of Genes Common to 3D7 & HB3	Percentage of Genes Present using the Totality of the Genes in this Category
1	167	8.41	140	12.29	156	13.70	63	45
2	64	3.22	112	9.83	77	6.76	27	19
3	81	4.08	29	2.55	52	4.57	1	1
4	213	10.73	19	1.67	52	4.57	0	0
5	192	9.67	52	4.57	70	6.15	3	2
6	57	2.87	58	5.09	13	1.14	0	0
7	137	6.90	149	13.08	21	1.84	3	2
8	226	11.39	114	10.01	93	8.17	1	1
9	142	7.15	14	1.23	54	4.74	3	2
10	107	5.39	42	3.69	24	2.11	0	0
11	65	3.27	111	9.75	74	6.50	31	22
12	65	3.27	34	2.99	197	17.30	2	1
13	148	7.46	11	0.97	110	9.66	0	0
14	166	8.36	104	9.13	103	9.04	1	1
15	155	7.81	150	13.17	43	3.78	4	3
Total	**1985**		**1139**		**1139**		**139**	

The table illustrates the *P. yoelli* orthologues in both 3D7 and Hb3 strains with their various percentage score. Column 9 indicates the number of common genes in two clusters with the same ID number from 3D7 and HB3 strains. Relatively high percentage score was obtained for cluster 1, 2 and 11 giving an indication that each of these 3 clusters is closely similar in both strains (3D7 and HB3) and may make a good choice for a gene to be used for PCR detection of malaria parasite at liver stage. More explanation is given for these genes in column 10 as to the choice of PF13_0227, PFL1700C, PF13_0227 and PF13_0358. 139 genes were found to be in corresponding clusters for both 3D7 and HB3.

The result of clustering expression values of *P. yoelli* (Tarun et al., 2008), *P. falciparum* 3D7 and HB3 (Bozdech et al., 2003a) is presented in Table 7.4 with the number of genes assigned to each corresponding clusters stated. The table further illustrates the *P. yoelli* orthologues in both 3D7 and HB3 strains with their various percentage score based on the number of genes per cluster. In Table 7.4 (row 9), the number of common genes ie the

intersection of 3D7 and HB3 is recorded per cluster. This makes a total of 139 genes. We extracted the expression values for only the 139 genes found to be in corresponding clusters for both 3D7 and HB3 strains as shown in Table 7.4 (column 8).

Next, using R programming, we performed a Wilcoxon statistical test significance test to extract from these 139 *P. falciparum* genes, those that are highly similar (p-values ≥ 0.5). From these 139 genes, we found that only 54 are highly similar. We find the *P. yoelli* orthologues for these 54 genes using PlasmoDB and seek from Tarun *et al.* (2008) microarray expression data, the expression values of these orthologues at the liver stage. We could not locate one of these 54 genes on Tarun *et al.* (2008) microarray experimental data and we are left with potential 53 viable candidates for PCR test. Finally, we compare the expression values of the 53 genes at the liver and blood stages. The idea behind this is that the genes that have highly dissimilar gene expression at the liver and the blood stages will be the theoretical/statistical viable candidate. Doing this we arrived at 29 genes set. Due to intellectual property right, we will not be able to list these genes.

7.5 DISCUSSION

Plasmodium proteins responsible for the parasite survival and development at the liver stage and their associated genes were verified to assess the level of their representation on Tarun *et al*. (2008) microarray slide. Surprisingly, many of the genes retrieved for each protein were not represented on the array. It is arguably right to say that the key genes used already for PCR testing at the liver stage seem missing in Tarun *et al.* (2008) microarray data. We therefore propose a call for a better robust microarray that will incorporate the missing genes.

On searching the PlasmoDB using Tarun *et al*. (2008) gene list containing 1985 genes, our search for *P. yoelli* orthologues retrieved only 1459 genes as *P. falciparum* orthologues. This indicates that not all *P. yoelli* genes have orthologues in *P. falciparum* and possibly some genes may have more than one orthologue. It is also important to note that the absence of orthologues in some species does not mean that the ortholog does not exist, it may not have been detected by the OrthoMCL algorithm used in EuPathDB (Omar and Kissinger, 2009).

In the output from traditional k-means clustering, we obtained a relatively high percentage score for common genes in cluster 1, 2 and 11 (see table 7.4, column 10) giving an indication that genes of 3D7 and HB3 strains for these 3 clusters are highly functionally related and confirmed to be important based on their presence in the two strains. Using R programming and Wilcoxon significance statistical test available on this platform, we arrived at 29 *P. falciparum / P.yoelli* gene that are highly statistical viable candidates to setup a PCR for the detection of malaria at the liver stage.

7.6 CONCLUSION

PCR diagnostic assay can easily be developed for mass screenings through automation and could thus be an effective diagnostic tool that is sensitive, specific, and less labour intensive than the current methods being used. We would like to take advantage of this system for the development of a simple, predictive and reliable test of the diagnosis of malaria at the liver stage. Here we have reported the application of a PCR-based test for the diagnosis of malaria infections in a clinical environment with emphasis on the choice of suitable genes for diagnosis. However, owing to little liver stage microarray expression data for the parasite transcriptome and the absence of key genes on the array of Tarun *et al.*, (2008) liver stage transcriptome analysis, we propose a call for better and robust microarray that will incorporate the missing genes.

CHAPTER EIGHT

CONCLUSION AND FUTURE WORK

The high rate of morbidity and mortality occasioned by malaria devastation in many African countries, parts of America and Asia, has attracted great attention to public health. Africa suffers infrastructural decay and poverty occasioned by malaria burden which seems to have persisted over time. A keen observation around most hospitals in Nigeria and Africa in general show that many patients are plagued by malaria compared to other diseases. The recent alarming statistics given by Bathurst (2008) on the global social and economic impact of malaria has generated attention on malaria problem. With affliction of more then 1/3 of the human population and recording 1 million deaths per year, malaria challenge costs up to 40% of total public health expenditure and an annual lost Gross Domestic Product (GDP) of $15 billion in Africa. Many global initiatives currently addressing the malaria issues include Millennium Development Goals (MDG), Medicine for Malaria Venture (MMV), Roll Back Malaria (RBM), Global Malaria Control Strategy (GMCS) with support from Gates foundation.

Malaria control activities in Nigeria are planned and implemented through the Primary Health Care (PHC) system (Federal Ministry of Health, 2005). However, the use of health centres, as the first resort for malaria management has been shown to be low in many African studies including Nigeria. The option of malaria treatment at PHC is delayed till the advent of complication and near death. This was attributed to difficulty with access to health centre, scarcity of affordable drugs including antimalarial drugs, perceived deficiencies in the performance of formal health services including poor clinical skills, attitude of health personnel and cultural beliefs (World Health Organisation/ United Nations International Children's Emergency Fund, 2003; Feyisetan, et al., 1997). Some current malaria intervention strategies are HMM (home management of malaria), ITNs (insecticide treated nets), IPT (intermittent preventive treatment) and ACT (Artemesinin-based combination drug). Due to development of resistance in the parasite, lack of licensed vaccine, fundamental complexity (2 hosts and intricate life cycle with complex gene regulatory mechanisms) inherent in the parasite, multiple drug strategies - novel and combination or optimization therapies using analogues has been demonstrated to be useful.

Following local intensified effort against malaria in 2004, Nigeria evolved a national policy on malaria treatment by dropping chloroquine as first line drug treatment and adopted the combination therapy of artemether and lumefantrine (Coartem), artesunate and amodiaquine (Adeneye et al., 2007). Again, there is increasing fear that the parasite will soon show prevalence of resistance on the new artemisinin-based drugs despite its expensive cost relative to chloroquine. A new approach is desirable hence the challenging need to eradicate malaria in endemic Sub-Saharan African nations and other endemic regions through the application of functional genomics and computational tools capable of evolving new malaria treatment strategies.

We developed recently a new and novel Metric Matrics k-means (MMk-means) clustering algorithm to cluster genes according to their functional roles with a view to obtaining further knowledge on many *P. falciparum* genes. Using Pearson correlation as the distance metric, Ding and He threshold (Ding and He, 2004) and our new theoretical derivation, we are able to determine which of the k clusters as a bunch are stable (with its data members in the same cluster retained in subsequent iteration). Using the above methods, in our new k-means algorithm, we were able to save significant computation time at each iteration and thus arrived at an $O(nk^2)$ expected run time. Applying the technique of PCA and its relationship to k-means in MMk-means, we present a k-means algorithm with improved runtime while maintaining good cluster quality as evaluated using Hubert-Arabie Adjusted Rand Index (ARI_{HA}). The result from ARI_{HA} was demonstrated for both biological and non-biological data. We also proved theoretically that our new and novel k-means algorithm is correct.

By using the *in-vitro* microarray data of Le Roch et al. (2003) and with the classification from an *in-vivo* microarray data of Daily et al. (2007), we performed a comparative functional classification of *P. falciparum* genes and further validated the effectiveness of our MMk-means algorithm. We were able to deliver improved cluster quality using our MMk-means.

However, Daily et al.'s (2007) claim, that the physiological state (the environmental stress response) of *P. falciparum* in selected malaria-infected patients observed in one of their clusters cannot be found in any *in-vitro* clusters, was not observed in this work, as our analysis reveals many *in-vitro* clusters representation in that their *in-vivo* cluster. The implication of this result in malaria treatment is that the wide difference between *in-vitro* and *in-vivo* reports on *P. falciparum* infection and virulence may not be largely different in reality. This will be a good lead for drug design and diagnostics especially at the clinical trial stage comparative to whatever result investigators must have obtained *in-vitro*. The gene clusters identified should be closely studied further for drug and diagnostic purposes.

The scarcity of technology required to detect malaria infection at the liver stage could lead to early intervention at the asymptomatic stage and quick recovery. However, waiting for the detection of malaria at the blood stage can lead to delayed treatment that may engender serious complication and death. In addressing the challenges faced with the identification of useful genes and possible primer information, our *in-silico* prediction points to suggest a new exploratory experimental study for possible PCR-based detection of malaria infection at the liver stage. Using our method and the concept of orthology, we predicted some key genes that will be useful for malaria diagnosis at the liver stage. Due to intellectual property right, we were unable to list these twenty nine genes here.

In future work, we will incorporate an added capability of fitting adequate number of gene clusters required for a given microarray input data set and deploy our MMk-means to the analysis of an *in-vivo* malaria microarray data. Due to little work as regards the liver stage microarray expression data for the parasite transcriptome and the absence of key genes on the array of *Tarun et al.* (2008) liver stage transcriptome analysis, we propose a call for better and robust microarray that will incorporate the missing genes.

Data used for part of this work, were obtained from simultaneous gene expression studies carried out in 2003 with different deoxyribonucleic acid (DNA) microarray technologies (Le Roch et al., 2003; Bozdech et al., 2003a), during the early days of genome sequencing of *P. falciparum*. Probe and primer designs were challenging due to scarcity of sequence information for most high throughput technologies like DNA microarray and Polymerase

Chain Reaction (PCR). With the current technological advancement, we recommend a fresh microarray experiment to obtain a more reliable data for further analysis.

REFERENCES

- Abe A., Inoue K., Tanaka T., Kato J., Kajiyama N., Kawaguchi R., Tanaka S., Yoshiba M. and Kohara M. (1999): **Quantitation of hepatitis B virus genomic DNA by real-time detection PCR**. *Journal of Clinical Microbiology, 37, 2899–903*

- Adeneye A.K., Jegede A.S., Mafe M.A. **and** Nwokocha E.E. (2007): **A Pilot Study to Evaluate Malaria Control Strategies in Ogun State, Nigeria**. *World Health and Population, 9(2):83-94*

- Adjuik M., Agnamey P., Babiker A., Borrmann S., Brasseur P., Cisse M., Cobelens F., Diallo S., Faucher J. F. and Garner P. (2002): **Amodiaquine-artesunate versus amodiaquine for uncomplicated *Plasmodium falciparum* malaria in African children: a randomised, multicentre trial.** *Lancet* **359**, 1365 -1372.

- Affymetrix Inc. (2001): **GeneChip Expression Analysis Algorithm Tutorial.**

- Ajayi I.O. and Falade C.O. (2006): **Pre-hospital treatment of febrile illness in children attending the General outpatients' clinic, University College Hospital, Ibadan Nigeria.** *African Journal Medicine and Medical Sciences*, 35**:**85-91.

- Ajayi I. O., Falade C. O., Bamgboye E. A., Oduola A.M.J. and Kale O. O. (2008): **Assessment of a Treatment Guideline to Improve Home Management of Malaria in Children in the Rural South-West Nigeria.** *Malaria Journal*, **7:**24doi:10.1186/1475-2875-7-24.

- Atkinson C. T, Dusek R. J. and Lease J. K. (2001): **Serological responses and immunity to superinfection with avian malaria in experimentally-infected Hawaii Amakihi.** *Journal of Wildlife Diseases 37: 20-27.*

- Alonso P.L., Sacarlal J., Aponte J.J., Leach A., Macete E., Milman J., Mandomando I., Spiessens B., Guinovart C., Espasa M., Bassat Q., Aide P., Ofori-Anyinam O., Navia M.M., Corachan S., Ceuppens M., Dubois M.C., Demoitie M.A., Dubovsky F., Menendez C., Tornieporth N., Ballou W.R., Thompson R. and Cohen J. (2004): **Efficacy of the RTS,S/AS02A vaccine against *Plasmodium falciparum* infection and disease in young African children: randomized controlled trial.** *Lancet 364:* 1411–1420.

- Aly A.S.I. and Matuschewski K. (2005): **A malarial cysteine protease is necessary for *Plasmodium* sporozoite egress from oocysts.** *The Journal of Experimental Medicine, 202(2) 225–230.* [Available online @ www.jem.org/cgi/doi/10.1084/jem.20050545]

- Ancsin, J. B. and Kisilevsky R. (2004): **A binding site for highly sulfated heparan sulfate is identified in the N terminus of the circumsporozoite protein: significance for malarial sporozoite attachment to hepatocytes.** *Journal of Biological Chemistry, 279:21824–21832.*

- Anderson J. (2006): **Design and Applications in Molecular Biology Research: PCR and Primer Design.** *World Health Organisation / Tropical Disease Research Workshop Training manual, Malaria Research Training Centre, Mali.*

- Attaran A., Barnes K. I., Curtis C., d'Alessandro U., Fanello C. I., Galinski M. R., Kokwaro G., Looareesuwan S., Makanga M., Mutabingwa T. K., Talisuna A., Trape J. F. and Watkins W. M. (2004): **WHO, the Global Fund and Medical Malpractice in Malaria Treatment.** *The Lancet 363:237-240.*

- Baldi P. and Hatfield G.W. (2002)**: DNA Microarrays and Gene Expression.** *Cambridge University Press, UK.*

- Bassler H.A., Flood S.J., Livak K.J., Marmaro J., Knorr R. and Batt C.A., (1995): **Use of a fluorogenic probe in a PCR-based assay for the detection of Listeria monocytogenes.** *Applied Environmental Microbiology, 61, 3724–8.*

- Bathurst I. (2008): **Initiatives of MMV in Antimalaria Drug Design**. *International Conference on "Drug Design and Discovery for Developing Countries"*, International Centre for Science – United Nations Industrial Development Organisation, Trieste, Italy. [http://www.ics.trieste.it/portal/ Activity Document.aspx?id=5713

- Belacel N., Cuperlovic-Culf M., Laflamme M. and Ouellette R. (2004): **Fuzzy J-Means and VNS Methods for Clustering Genes from Microarray Data**. *International Journal of Bioinformatics, Oxford University Press, NRC 46546.*

- Belacel N., Hansen P. and Mladenovic N.(2002): **Fuzzy J-Means: a new heuristic for fuzzy clustering**. *Pattern Recognition,* **35,** 2193-2200.

- Bensch S., Stjernman M., Hasselquist D., Ostman O., Hansson B., Westerdahl H. and Pinheiro T. R. (2000): **Host specificity in avian blood parasites: A study of Plasmodium and Haemopro- teus mitochondrial DNA amplified from birds.** *Proceedings of the Royal Society of London Series B 267: 1583-1589*

- Berry A., Fabre R., Benoit-Vical F., Cassaing S. and Magnaval J. F. (2005): **Contribution of PCR-based methods to diagnosis and management of imported malaria.** *Medecine Tropical: rev du corps de santé colonial*, **65:**176-183.

- Bezdek J. C. (1981): **Pattern Recognition with Fuzzy Objective Function Algorithms.** *Plenum Press, New York.*

- Bhatia S. K. and Deogun J. S. (1998)**: Conceptual clustering in information retrieval.** *IEEE Transactions on Systems, Man and Cybernetics, Part B, 28(3):427-436.*

- Biodiscovery (2001): **GeneSight User Manual version 3.1.3.**

- Björkman A., do Rosário V. E., Snounou G. and Walliker D. (1998): **Standardizing PCR for Molecular Epidemiology Studies of Malaria.** *Parasitology Today, 11(1*), *Pp 19-23 doi:10.1016/S0169-4758(97)01201-5*

- Blair, P. L., Kappe S. H., Maciel J. E., Balu B. and Adams J. H. **(**2002): ***Plasmodium falciparum* MAEBL is a unique member of the *ebl* family**. *Molecular and Biochemical Parasitology. **122:**35–44.*

- Bojang K.A., Milligan P.J.M., Pinder M., Vigneron L., Alloueche A., Kester K.E., Ballou W.R., Conway D.J., Reece W.H.H., Gothard P., Yamuah L., Delchambre M., Voss G., Greenwood B.M., Hill A.V.S., McAdam K.P., Tornieporth N., Cohen J.D. and Doherty T. (2001): **Efficacy of RTS,S/AS02 malaria vaccine against *Plasmodium falciparum* infection in semi-immune adult men in The Gambia: a randomised trial.** *Lancet 358:* 1927–1934.

- Bottius E., BenMohamed L., Brahimi K., Gras H., Lepers J. P., Raharimalala L., Aikawa M., Meis J., Slierendregt B., Tartar A., Thomas A. and Druilhe P. (1996): **A novel *Plasmodium falciparum* sporozoite and liver stage antigen (SALSA) defines major B, T helper, and CTL epitopes.** *Journal of Immunology, **156:**2874–2884.*

- Bow S-T (1984): **Pattern Recognition: Applications to Large Data-Set Problems**. *Marcel Denker Incorporation, New York.*

- Bozdech Z., Llinas, M., Pulliam B. L., Wong E. D., Zhu J. and DeRisi J. L. (2003a): **The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*** *Public Library of Science Biology*, **1**, E5.

- Bozdech Z., Zhu J., Joachimiak M.P., Cohen F.E., Pulliam B. and DeRisi,J.L. (2003b)**: Expression pr**o**filing** of the **schizont and trophozoite stages** of *Plasmodium falciparum* **with a long-oligonucleotide microarray**. *Genome Biology*, **4**, R9

- Breman J. G. (2001): **The ears of the hippopotamus: manifestations, determinations and estimates of th. malaria burden.** *Am J Trop Med Hyg*, Supplement, 1–2**:**1-11

- Briones M. R., Tsuji M. and Nussenzweig V. (1996): **The large difference in infectivity for mice of Plasmodium berghei and Plasmodium yoelii sporozoites cannot be correlated with their ability to enter into hepatocytes.** *Molecular and Biochemical Parasitology,* 77, 7–17.

- Brown G. V. and Reeder J. C. (2002): **Malaria Vaccines**. *The Medical Journal of Australia,* 177: 230–231.

- Bruce-Chwatt L. J (1951): **Malaria in Nigeria.** *Bulletin of the World Health* Organization 4**:**301-327

- Brun˜a-Romero O., Hafalla J. C.R., Gonza´lez-Aseguinolaza G., Sano G., Tsuji M. and Zavala F. (2001): **Detection of malaria liver-stages in mice infected through the bite of a single Anopheles mosquito using a highly sensitive real-time PCR.** *International Journal for Parasitology 31, 1499–1502.*

- Brunet J. P., Tamayo P., Golub T. R. and Mesirov J. P. (2004): **Metagenes and molecular pattern discovery using matrix factorization.** *Proceeding of the National Academy of Science, USA,* 101, 4164–4169.

- Cerami C., Frevert U., Sinnis P., Takacs B., Clavijo P., Santos M. J. and Nussenzweig V. (1992): **The basolateral domain of the hepatocyte plasma membrane bears**

receptors for the circumsporozoite protein of *Plasmodium falciparum* sporozoites. *Cell 70:1021–1033*

- Chattopadhyay, R., D. Rathore, H. Fujioka, S. Kumar, P. de la Vega, D. Haynes, K. Moch, D. Fryauff, R. Wang, D. J. Carucci and S. L. Hoffman (2003): **PfSPATR, a *Plasmodium falciparum* protein containing an altered thrombospondin type I repeat domain is expressed at several stages of the parasite life cycle and is the target of inhibitory antibodies.** *The Journal of Biological Chemistry, 278:25977–25981.*

- Chen T. (2006): **DNA Microarrays - An Armory for Combating Infectious Diseases in the New Century.** *Infectious Disorders - Drug Targets 6(3):263-279.*

- Clements F. E. (1954): **Use of cluster analysis with anthropological data.** *American Anthropologist,New Series, Part 1, 56(2):180-199.*

- Coleman R.E., Maneechai N., Rachaphaew N., Kumpitak C, Miller R.S., Soyseng V., Thimasarn K. and Sattabongkot J. (2002): **Comparison of field and expert laboratory microscopy for active surveillance for asymptomatic Plasmodium falciparum and Plasmodium vivax in western Thailand.** *American Journal Tropical Medicine Hygiene,* 67:141-144.

- Coleman R.E., Sattabongkot J., Promstaporm S., Maneechai N., Tippayachai B., Kengluecha A., Rachapaew N., Zollner G., Miller R.S., Vaughan J.A., Thimasarn K. and Khuntirat B. (2006): **Comparison of PCR and microscopy for the detection of asymptomatic malaria in a *Plasmodium falciparum/viva*x endemic area in Thailand.** *Malaria Journal,* **5:**121.

- Coppi A., Pinzon-Ortiz C., Hutter C., and Sinnis P. (2005): **The *Plasmodium* circumsporozoite protein is proteolytically processed during cell invasion.** *Journal of Experimental Medicine, Vol. 201, No. 1, 27–33.*

- Cox-Singh J., Mahayet S., Abdullah M.S. and Singh B. (1997): **Increased sensitivity of malaria detection by nested polymerase chain reaction using simple sampling and DNA extraction.** *International Journal of Parasitology*, **27:**1575-1577.

- Daily J. P., Scanfeld D., Pochet N., Le Roch K., Plouffe D., Kamal M., Sarr O., Mboup S., Ndir O., Wypij D., Levasseur K., Thomas E., Tamayo P., Dong C., Zhou Y., Lander E. S., Ndiaye D., Wirth D., Winzeler E. A., Mesirov J. P. and Regev A. (2007): **Distinct physiological states of *Plasmodium falciparum* in malaria-infected patients.** *Nature* 450, 1091-1095.

- Dame J. B., Williams J. L., McCutchan T. F., Weber J. L., Wirtz R. A., Hockmeyer W.T., Maloy W.L., Haynes J.D., Schneider I., Roberts D.R., Sanders G.S., Reddy E.P., Diggs C.L. and Miller L.H. (1984): **Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite *Plasmodium falciparum.*** *Science, 225:* 593–599.

- Darken C. and Moody J. (1990): **Fast adaptive k-means clustering: Some empirical results**. *International Joint Conference on Neural Networks, 2, 233-238.*

- Daubersies P., Thomas A. W., Millet P., Brahimi K., Langermans J. A. M., Ollomo B., BenMohamed L., Slierendregt B., Eling W., Van Belkum A., Dubreuil G., Meis J. F., Gue´rin-Marchand C., Cayphas S., Cohen J., Gras-Masse H., and Druilhe P. (2000): **Protection against Plasmodium falciparum malaria in chimpanzees by mmunization with the conserved preerythrocytic liver stage antigen 3**. *Nature Medicine, 6:1258–1263.*

- Dembélé D. and Kastner P. (2003): **Fuzzy C-means Method for Clustering Microarray Data**, *Bioinformatics 19(8): 973-980*.

- Ding C. and He X. (2004): **K-means Clustering via Principal Components Analysis**. *Association for Computing Machinery International Conference Proceeding Series, 69*.

- Ding C. (2008): **Personal communication via the emails**, 2008.

- Di Santi S.M., Kirchgatter K., Brunialti K.C., Oliveira A.M., Ferreira S.R. and Boulos M. (2004): **PCR -- based diagnosis to evaluate the performance of malaria reference centers.** *Revista do Instituto de Medicina Tropical de Sao Paulo*, 46:183-187.

- Donovan J.W., Ladetto M., Zou G., Neuberg D., Poor C., Bowers D. and Gribben J.G. (2000): **Immunoglobulin heavy-chain consensus probes for real-time PCR quantification of residual disease in acute lymphoblastic leukemia.** *Blood 95, 2651–8.*

- Dorsey G., Njama D., Kamya M. R., Cattamanchi A., Kyabayinze D., Staedke S. G., Gasasira A. and Rosenthal P. J. (2002): **Sulfadoxine/pyrimethamine alone or with amodiaquine or artesunate for treatment of uncomplicated malaria: a longitudinal randomised trial.** *Lancet **360**, 2031 -2038.*

- Draghici, S. (2003): **Data Analysis Tools for DNA Microarrays.** *Chapman & Hall/CRC: New York.*

- Druilhe P.R.L. and Fidock D. (1998): **Immunity to liver stages.** Sherman I.W., ed. *Malaria: Parasite Biology, Pathogenesis and Protection*. Washington, DC: American Society for Microbiology Press, 513–540.

- Druilhe P., Pradier O., Marc J.P., Miltgen F., Mazier D. and Parent G., (1986): **Levels of antibodies to *Plasmodium falciparum* sporozoite surface antigens reflect malaria transmission rates and are persistent in the absence of reinfection**. *Infection and Immunity, 53:* 393–397.

- Duda R.O. and Hart P.E (1973): **Pattern Classification and Scene Analysis.** John Wiley & Sons, New York.

- Ecclesiastes 9:11, **The Holy Bible**, The King James Version.

- Ellis J, Ozaki L.S., Gwadz R.W., Cochrane A.H., Nussenzweig V., Nussenzweig R.S. and Godson G.N. (1983): **Cloning and expression in *E. coli* of the malarial sporozoite surface antigen gene from *Plasmodium knowlesi***. *Nature 302:* 536–538.

- Fahim A.M., Salem A.M., Torkey F.A. and Ramadan M.A (2006): **An efficient enhanced k-means clustering algorithm.** *Journal of Zhejiang University SCIENCE A 7(10):1626-1633.* Available online at www.zju.edu.cn/jzus.

- Fallon S. M, Ricklefs R. E., Swanson B. L. and Bermingham E (2003): **Detecting Avian Malaria: An Improved Polymerase Chain Reaction Diagnostic.** *Journal of Parasitology, 89(5), pp. 1044-1047.*

- Fan, K. (1949): **On a theorem of Weyl concerning eigenvalues of linear transformations**. *Proceeding of the National Academy of Science, 35, 652-655.*

- Fayyad U.M., Piatetsky-Shapiro G., Smyth P. and Uthurusamy R. (1996): **Advances in Knowledge Discovery and Data Mining**. *Association for the Advancement of Artificial Intelligence/Massachusetts Institute of Technology Press.*

- Federal Ministry of Health [Nigeria] (2005): **National Antimalarial Treatment Policy.** *Federal Ministry of Health, National Malaria and Vector Control Division, Abuja, Nigeria.*

- Feldman R. A., Freed L. A. and Cann R. L. (1995): **A PCR test for avian malaria in Hawaiian birds.** *Molecular Ecology 4: 663-673*

- Feng Y. and Hamerly G. (2006): **PG-means: learning the number of clusters in data**. *Proceedings of the twentieth annual conference on neural information processing*

systems (NIPS). Available at http://cs.baylor.edu/~hamerly/papers /nips_06_ pgmeans .pdf

- Feyisetan B. J., Asa S. and Ebigbola J. A. (1997): **Mother's management of childhood diseases in Yoruba land: the influence of cultural beliefs.** *Health Transition Review,* **7:**221-234.

- Fidock D. A., Bottius E., Brahimi K., Moelans I. I., Aikawa M., Konings R. N., Certa U., Olafsson P., Kaidoh T., Asavanich A., Guerin-Marchand C. and Druilhe P. (1994): **Cloning and characterization of a novel Plasmodium falciparum sporozoite surface antigen, STARP.** *Molecular and Biochemical Parasitology, 64:219–232.*

- Garcia J. E., Puentes A., Lopez R., Vera R., Suarez J., Rodriguez L., Curtidor H., Ocampo M., Tovar D., Forero M., Bermudez A., Cortes J., Urquiza M. and Patarroyo M. E. **(**2003): **Peptides of the liver stage antigen-1 (LSA-1) of *Plasmodium falciparum* bind to human hepatocytes.** *Peptides **24:**647–657.*

- Garcia J. E., Curtidor H., Lopez R., Rodriguez L., Vera R., Valbuena J., Rosas J., Ocampo M., Puentes A., Forero M., Patarroyo M. A. and Patarroyo M. E. (2004): **Liver stage antigen 3 *Plasmodium falciparum* peptides specifically interacting with HepG2 cells.** *Journal of Molecular Medicine, 82:600–611.*

- Garcia J. E., Alvaro P. and Patarroyo M. E. (2006)**: Developmental Biology of Sporozoite-Host Interactions in *Plasmodium falciparum* Malaria: Implications for Vaccine Design.** *Clinical Microbiology Reviews 19, 4:686-707.*

- Gaur D., Furuya T., Mu J., Jiang L. B., Su X. Z. and Miller L. H. (2006): **Upregulation of expression of the reticulocyte homology gene 4 in the *Plasmodium falciparum* clone Dd2 is associated with a switch in the erythrocyte invasion pathway.** *Molecular and Biochemical Parasitology, 145, 205.*

- Genton B., Anders R.F., Alpers, M.P. and Reeder J.C. (2003): **The malaria vaccine development program in Papua New Guinea.** *Trends in Parasitology,* 19: 264–270.

- Gersho A., Gray R.M. (1992): **Vector Quantization and Signal Compression**. *Boston, Kluwer Academic.*

- Giacomini T., Lusina D., Foubard S., Baledent F., Guibert F. and Lepennec M. P. (1991): **Dangers of hematological automated analysis for malaria diagnosis.** *Bulletin de la Societe de Pathologie Exotique,* 84, pp. 330–333.

- Gibson G. (2003): **Microarray Analysis**. *Public Library of Science Biology*, 1(1): e15 doi:10.1371/journal.pbio.0000015.

- Gordon D.M., McGovern T.W., Krzych U, Cohen J.C., Schneider I, LaChance R., Heppner D.G., Yuan G., Hollingdale M., Slaoui M., Hauser P., Voet P., Sadoff J. C. and Ballou W. R. (1995): **Safety, immunogenicity, and efficacy of a recombinantly produced *Plasmodium falciparum* circumsporozoite protein-hepatitis B surface antigen subunit vaccine.** *Journal of Infectious Diseases, 171:* 1576–1585.

- Hamerly G. J. (2003): **Learning structure and concepts in data through data clustering.** *Ph.D Thesis, University of California*, San Diego.

- Hamerly G. and Elkan C. (2003): **Learning the k in kmeans.** In *proceedings of the seventeenth annual conference on neural information processing systems (NIPS).* Available at http://www.citeseer.ist.psu.edu/hamerly03learning.html

- Hänscheid T. (1999): **Diagnosis of malaria: review of alternatives to conventional** microscopy. *Clinical and Laboratory Haematology*, 21, pp. 235–245.

- Hänscheid T., Valadas E. and Grobusch M. P. (2000): **Automated Malaria Diagnosis Using Pigment Detection.** *Parasitology Today,16 (12) Pp 549-551*

- Hanscheid T. and Grobusch M.P. (2002): **How useful is PCR in the diagnosis of malaria?** *Trends in Parasitology,* 18:395-398

- Hayton K and Su X.-z. (2004): **Genetic and Biochemical Aspects of Drug Resistance in Malaria Parasites.** *Infectious Disorder-Drug Targets* **4(1)**: 1-10. [http://www.bentham.org/cdtid/contabs/cdtid4-1.htm]

- Herrington D.A., Clyde D.F., Losonsky G.A., Cortesia M.J., Murphy J.R., Davis J.R., Baqar S., Felix A.M., Heimer E.P., Gillessen D., Nardin E.H., Nussenzweig R.S., Nussenzweig V., Hollingdale M.R. and Levine M.M. (1987): **Safety and immunogenicity in man of a synthetic peptide malaria vaccine against *Plasmodium falciparum* sporozoites**. *Nature 328:* 257–259.

- Heyer L.J., Kruglyak S. and Yooseph, S. (1999): **Exploring Expression Data: Identification and Analysis of Coexpressed Genes.** *Genome Research* 9:1106-1115.

- Higuchi R., Fockler C., Dollinger G. and Watson R. (1993): **Kinetic PCR analysis: real-time monitoring of DNA amplification reactions.** *Biotechnology (NY)* 11, 1026–30.

- http://meeting.tropika.net/andi/. Retrieved on 20 March, 2009.

- http://wisdom.eu-egee.fr/malaria/plasmepsins.pdf. Retrieved 5 April, 2009.

- http://www.undp.org/mdg/basics.shtml. Retrieved 15 June, 2009.

- Ishino, T., K. Yano, Y. Chinzei, and M. Yuda (2004): **Cell-passage activity is required for the malaria parasite to cross the liver sinusoidal cell layer.** *Public Library of Science Biology, 2:77–84.*

- Jain A.K and Flynn P. (1996): **Image segmentation using clustering.** In *Advances in Image Understanding, pages 65-83. Institute of Electrical and Electronics Engineers Computer Society Press.*

- Kain K.C., Harrington M. A., Tennyson S. and Keystone J. S. (1998): **Imported malaria: prospective analysis of problems in diagnosis and management.** *Clinical Infectious Diseases,* 27, pp. 142–147.

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

- Kariu, T., M. Yuda, K. Yano and Y. Chinzei. **(**2002): **MAEBL is essential for malarial sporozoite infection of the mosquito salivary gland**. *Journal of Experimental Medicine, 195:1317–1323.*

- Khan Z.M., Ng C. and Vanderberg J.P. (1992): **Early hepatic stages of Plasmodium berghei: release of circumsporozoite protein and host cellular inflammatory response.** *Infection and Immunity, 60, 264–70.*

- Kissinger J. C., Brunk B. P., Crabtree J., Fraunholz M. J., Garjria B., Milgram A. J., Pearson D.S., Schug J., Bahl A., Diskin, S.J., Ginsburg H., Grant G. R., Gupta D., Labo P., Li L., Mailman M.D., McWeeney S. K. Whetzel P., Stoeckert C. J. and Roos D. S. (*2002*). **The Plasmodium genome database.** *Nature 419: 490-492.*

- Kumar A., Sabharwal Y. and Sen S. **A simple linear time (1+ε) approximation algorithm for k-means clustering in any dimensions**. *Proceedings of the 45th Annual Institute of Electrical and Electronics Engineers Symposium on Foundation of Computer Science, 2004, 454-462.*

- Land K. M., Sherman I. W., Gysin J. and Crandall I. E. (1995): **Anti-adhesive antibodies and peptides as potential therapeutics for Plasmodium falciparum malaria.** *Parasitology Today, 11(1), Pp 19-23, doi:10.1016/0169-4758(95)80100-6.*

- Le Roch K. G., Zhou Y., Blair P. L., Grainger M., Moch J. K., Haynes J. D., Dela Voga P., Holder A. A., Batalov S., Carucci D. J. and Winzeler E. A. (2003**): Discovery of** g**ene function by expression pr**ofi**ling** of the **malaria parasite life cycle**. *Science***,** *301, 1503–1508.*

- Levine R.A., Wardlaw S. C. and Patton C. L. (1989): **Detection of haematoparasites using quantitative buffy coat analysis tubes.** *Parasitology Today*, 5, pp. 132–134.

- Lewis C. (2009): **Definition of Homologue, Orthologue and Paralogue.** Available on http://homepage.usask.ca/~ct1271/857/def_homolog.shtml. Retrieved on April 4, 2009.

- Li C. and Wong W. H. (2001a): **Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection.** *Proceedings of the National Academy of Science* Vol. 98, 31-36.

- Li C. and Wong W. H. (2001b): **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application.** *Genome Biology* 2(8): research0032.1-0032.11.

- Li J., Wirtz R. A., Mcconkey G. A., Sattabongkot J., Waters A. P., Rogers M. J. and Mccutchan T. F. (1995). ***Plasmodium*: Genus- conserved primers for species identification and quantitation.** *Experimental Parasitology 81: 182-19*

- Lin M., Wei L-J., Sellers W. R., Lieberfarb M., Wong W. H. and Li C. (2004): dChipSNP: Significance Curve and Clustering of SNP-Array-Based Loss-of-Heterozygosity Data. *Bioinformatics*. 20: 1233-1240.

- Llinás M., Bozdech Z., Wong E.D., Adai A.T., DeRisi J.L. (2006): **Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains.** *Nucleic Acids Research 34(4):1166-73.*

- Lloyd S.P. (1957): **Least squares quantization in PCM.** *Bell Laboratories Internal Technical Report, Institute of Electrical and Electronics Engineers Transaction on Information Theory.*

- Lo´pez, R., H. Curtidor, M. Urquiza, J. Garcia, A. Puentes, J. Suarez, M. Ocampo, R. Vera, L. E. Rodriguez, F. Castillo, G. Cifuentes, and M. E. Patarroyo. (2001). ***Plasmodium falciparum:*** **binding studies of peptide derived from the sporozoite surface protein 2 to Hep G2 cells**. *The Journal of Peptide Research, 58:285–292.*

- Lopez, R., Garcia J., Puentes A., Curtidor H., Ocampo M., Vera R.,Rodriguez L. E, Suarez J., Urquiza M.,. Rodriguez A. L, Reyes C. A., Granados C. G., and Patarroyo M. E. (2003): **Identification of specific Hep G2 cell binding regions in Plasmodium falciparum sporozoite-threonine-asparagine- rich protein (STARP).** *Vaccine 21:2404–2411*.

- MacQueen J. (1967): **Some methods for classification and analysis of multi-variate observations.** In *Proc. of the Fifth Berkeley Symp. on Math., Statistics and Probability, LeCam, L.M., and Neyman, J., (eds.), Berkeley: University of California Press*.

- Makler M.T., Piper R.C. and Milhous W.K.(1998): **Lactate dehydrogenase and the diagnosis of malaria.** *Parasitology Today, 14, pp. 376–377.*

- Marshall B. A., Theil K. S., and Brandt J. T. (1990): **Abnormalities of leukocyte histograms resulting from micro-organisms.** *American Journal of Clinical Pathology,* 93, pp. 526–532.

- McKenna K.C., Tsuji M., Sarzotti M., Sacci Jr. J.B., Witney A.A. and Azad A.F., (2000): Gamma delta T cells are a component of early immunity against preerythrocytic malaria parasites. *Infection and Immunity, 68, 2224–30.*
- Meshnick S. R. (2001): **Artemisinin and its derivatives.** In *Antimalarial Chemotherapy: Mechanisms of Action, Resistance, and New Directions in Drug Discovery* (*ed. P. J. Rosenthal), pp. 191-201. Totowa, NJ: Humana Press.*

- Miller L. H., Baruch D. I., Marsh K. and Doumbo O. K. (2002)**: The Pathogenic Basis of Malaria.** *Nature 415, 673-679.*

- Molineaux L. and Gramiccia G. (1980): **The Garki project: Research on the epidemiology and control of malaria in the Sudan savanna of West Africa.** *World Health Organization, Geneva.*

- Morrison T.B., Weis J.J. and Wittwer C.T. (1998): **Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification.** *Biotechniques 24, 954–62.*

- Morrison T.B., Ma Y., Weis J.H. and Weis J.J. (1999): **Rapid and sensitive quantification of Borrelia burgdorferi-infected mouse tissues by continuous fluorescent monitoring of PCR**. *Journal of Clinical Microbiology, 37, 987–92.*

- Mullis K.B. (1990): **The Unusual Origin of the Polymerase Chain Reaction.** *Scientific American, April 1990, pp 56-64.*

- Nabarro D.N. and Tayler E.M. (1998): **The Roll Back malaria Campaign**. *Science 280 (5372): 2067-8.*

- Nardin, E., Zavala, F., Nussenzweig, V. and Nussenzweig, R.S., (1999): **Preerythrocytic malaria vaccine: mechanisms of protective immunity and human vaccine trials.** *Parassitologia 41, 397–402.*

- Nwaka S. (2008): **Drug Discovery Networks for Infectious Tropical Disease.** *International Conference on "Drug Design and Discovery for Developing Countries"*, International Centre of Science – United Nations Industrial Development Organisation, Trieste, Italy. [http://www.ics.trieste.it/portal/ Activity Document.aspx?id=5713]

- Okonkwo P. O., Akpala C. O., Okafor H. U., Mbah A. U. and Nwaiwu O. (2001): **Compliance to correct dose of chloroquine in uncomplicated malaria correlates with improvement in the condition of rural Nigerian children.** *Transaction of the Royal Society of Tropical Medicine and Hygiene*, 95:320-324.

- Olaogun A .A., Ayandiran O., Olasode O .A., Adebayo A. and Omokhodion F. (2005): **Home management of childhood febrile illness in a rural community in Nigeria.** *Australia Journal of rural Health*, **13:**97-101.

- Omar H. and Kissinger J. (2009): **PlasmoDB Support: Ortholog Genes of P.yoelli.** *Communication via Email.*

- Orlandi-Pradines E., Penhoat K., Durand C., Pons C., Bay C., Pradines B., Fusai T., Josse R., Dubrous P., Meynard J., Durand J., Migliani R., Boutin J., Druilhe P. and Rogier C. (2006): **Antibody Responses To Several Malaria Pre-Erythrocytic Antigens As A Marker Of Malaria Exposure Among Travelers.** *American Journal of Tropical Medicine and Hygiene,* 74(6), pp. 979–985.

- Osamor V.C and Adebiyi E.F. (2007): **Microarray Technology: An experimental and computational way to eradicate malaria in Africa**. *Proceedings of 9th*

International Conference of Nigeria Computer Society on Millennium Development Goals in Nigeria, Owerri, Nigeria.

- Osamor, V., Adebiyi, E. and Doumbia, S. (2009): **Comparative functional classification of *Plasmodium falciparum* genes using k-means clustering**. *Int. Conf. on Bioinformatics and Biomedical Technology, Singapore. International Electrical and Electronics Engineers (IEEE) Computer Society Press, pp. 491 – 496.*

- Osamor V., Adebiyi E., Oyelade J. and Doumbia S. (under review): **Reducing the Time Requirement of k-means Algorithm.** *BioMedCentral Bioinformatics.*

- Oyedeji S. I., Awobode H. O., Monday G. C., Kendjo E., Kremsner P. G. and Kun Jurgen F. (2007): **Comparison of PCR-based detection of Plasmodium falciparum infections based on single and multicopy genes.** *Malaria Journal 6:112.*

- Pan American Health Organisation (2006): **Regional Strategic Plan for Malaria in the Americas 2006-2010**. Washington, D.C: *Pan American Health Organization, ISBN 92 75 12641 0.*

- Pelleg D. and Moore A. (2000): **X-means: Extending K-means with efficient estimation of the number of clusters**. In *Proceedings of the 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann, San Francisco, California.

- Perkins S. L., Osgood S. M. and Schall J. J. (1998): **Use of PCR for detection of subpatent infections of lizard malaria: Implications for epizootiology.** *Molecular Ecology 7: 1587-1590.*

- Posner G. H., Paik I. H., Sur S., McRiner A. J., Borstnik K., Xie, S. and Shapiro T. A. (2003): **Orally active, antimalarial, anticancer, artemisinin-derived trioxane dimers with high stability and efficacy.** *Journal of Medical Chemistry, 46, 1060 - 1065.*

- Puentes, A., J. Garcia, R. Vera, R. Lopez, J. Suarez, L. Rodriguez, Curtidor H., Ocamp M.o, Tovar D., Forero M., Bermudez A., Cortes J., Urquiza M., and Patarroyo M. E. (2004): **Sporozoite and liver stage antigen Plasmodium falciparum peptides bind specifically to human hepatocytes.** *Vaccine 22:1150–1156.*

- Quackenbush J. (2001): **Computational analysis of microarray data**. *Nature Review Genetics*, 2, 418-427.

- Quackenbush J. (2005): **Using DNA Microarrays to Assay Gene Expression.** In Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Third Edition, *Eds Baxevanis A.D. and Ouellette B.F. F., John Wiley & Sons*.

- Ralph S. A., D'Ombrain M. C. and McFadden G. I. (2001): **The apicoplast as an antimalarial drug target.** *Drug Resistance Updates, 4, 145–151.*

- Richard E. A., Sehgal R. N. M., Jones H. I. and Smith T. B. (2002): **A comparative analysis of PCR-based detection methods for avian malaria.** *Journal of Parasitology 88: 819-822*

- Richie T. L. and Saul A. (2002): **Progress and challenges for malaria vaccines.** *Nature, 415, 694-701.*

- Ricklefs R. E. and Fallon S. M. (2002): **Diversification and host switching in avian malaria parasites.** *Proceedings of the Royal Society of London Series B, 269: 885-892.*

- Rodgers, J. L. and Nicewander, W. A. (1988): **Thirteen ways to look at the correlation coefficient**. *The American Statistician, 42(1), 59-66.*

- Romans 9:16, **The Holy Bible**, The King James Version.

- Rosenthal P. J. (2003): **Antimalarial drug discovery: old and new approaches.** *The Journal of Experimental Biology* 206, 3735-3744, doi: 10.1242/jeb.00589

- Saeed A. I, Sharov V., White J., Li J., Liang W., Bhagabati N., Braisted J., Klapa M., Currier T., Thiagarajan M., Sturn A., Snuffin M., Rezantsev A., Popov D., Ryltsov A., Kostukovich E., Borisovsky I., Liu Z., Vinsavich A., Trush V. and Quackenbush J. (2003): **TM4: A Free, Open-Source System for Microarray Data Management and Analysis**. *BioTechniques 34:374-378.*

- Sammy L. (2006): **k-Means Clustering & Finding k**. [http://www.codeodor.com]. Retrieved: December 5, 2006.

- Sethabutr O., Brown A. E., Panyim S., Kain K. C., Webster H. K. and Echeverria P. (1992): **Detection of *Plasmodium falciparum* by polymerase chain reaction in a field study.** *Journal of Infectious Diseases, 166:145–148.*

- Schena M., Shalon D., Davis R.W. and Brown P.O. (1995): **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science, 270 (5235): 467-70.*

- Silvie O., Goetz K. and Matuschewski K. (2008): **A Sporozoite Asparagine- Rich Protein Controls Initiation of *Plasmodium* Liver Stage Development.** *Public Library of Science Pathogen,* 4(6): e1000086, doi: 10.1371/journal.ppat.1000086.

- Sneath P.H. (1957): **The application of computers to taxonomy.** *Journal of. General Microbiology, 17:201-226.*

- Snounou G. (1996): **Detection and identification of the four malaria parasite species infecting humans by PCR amplification.** *Methods in Molecular Biology, 50:263–291.*

- Snounou G., Viriyakosol S., Jarra W., Sodsri T. and Brown K. N. (1993): **Identification of the 4 human malaria parasite species in field samples by the polymerase-chain-reaction and detection of a high prevalence of mixed infections.** *Molecular and. Biochemical Parasitology,* 58, pp. 283–292.

- Sokal R.R and. Michener C.D. (1958): **A statistical method for evaluating systematic relationships.** *University of Kansas Science Bulletin, 38:1409-1438.*

- Steinley, D. (2003): **Local optima in K-means clustering: What you don't know may hurt you.** *Psychological Methods,* 8, 294–304.

- Steinley D. (2004): **Properties of the Hubert-Arabie Adjusted Rand index**. *Psychological Methods, 9, 386-396.*

- Steinley, D. (2006): **K-means clustering: A half-century synthesis.** *British Journal of Mathematical and Statistical Psychology, 59, 1–34.*

- Suarez, J. E., Urquiza M., Puentes A., Garcia J. E., Curtidor H., Ocampo M., Lopez R., Rodriguez L. E., Vera R., Cubillos M., Torres M. H. and Patarroyo M. E. (2001): ***Plasmodium falciparum*** **circumsporozoite (CS) protein peptides specifically bind to HepG2 cells.** *Vaccine, 19:4487–4495.*

- Tahar R, Ringwald P. and Basco L K. (1997): **Diagnosis of *Plasmodium malariae* infection by the polymerase chain reaction**. *Transaction of the Royal Society of Tropical Medicine and Hygiene, 91:410–411*.

- Tamayo P., Slonim D., Mesirov J., Zhu Q, Kitareewan S., Dmitrovsky E., Lander E. and Golub T. (1999): **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proceedings of the National Academy of Sciences, USA, 96:2907-2912.*

- Tan T. M. C., Nelson J. S., Ng H. C., Ting R. C. Y. and Kara U. A. K. (1997): **Direct PCR amplification and sequence analysis of extrachromosomal *Plasmodium* DNA from dried blood spots.** *Acta Tropica, 68:105–114.*

- Tarun A. S., Peng X., Dumpit R. F, Ogata Y., Silva-Rivera H., Camargo N., Daly T. M., Bergman L. W. and Kappe S. H. I. (2008): **A combined transcriptome and proteome survey of malaria parasite liver stages**. *Proceedings of the National Academy of Sciences, 105 (1) 305-310.*

- Tatusov R. L., Koonin E. V. and Lipman D. J. (1997): **A genomic perspective on protein families.** *Science 278, 631*

- Teknomo K. (2006): **K-Means Clustering Tutorials**. [http:\\people.revoledu.com\kardi\ tutorial\kMean\]. Retrieved: December 2006

- Tham J. M., Lee S. H., Tan T. M. C., Ting R. C. Y., and Kara U. A. K.(1999): **Detection and Species Determination of Malaria Parasites by PCR: Comparison with Microscopy and with ParaSight-F and ICT Malaria Pf Tests in a Clinical Environment.** *Journal of Clinical Microbiology, 37(5): 1269–1273.*

- Tusher V. G., Tibshirani R. and Chu G., (2001): **Significance analysis of microarrays applied to the ionizing radiation response.** *Proceedings of the National Academy of Sciences, USA*, 98:5116–5121.

- United Nations International Children's Emergency Fund and Federal Ministry of Health [Nigeria] (2002): **Treated Bednets in Nigeria: Analysis of the Market for Bednets Insecticides and ITNs in Nigeria.** *Abuja: United Nation Information and Children Emergency Fund / Federal Ministry of Health [Nigeria]*.

- Vennerstrom J. L., Dong Y., Andersen S. L., Ager A. L. J., Fu H., Miller R. E., Wesche D. L., Kyle D. E., Gerena L., Walters S. M., Wood J. K., Edwards G., Holme A. D.,

Mclean W. G. and Milhous W. K. (2000): **Synthesis and antimalarial activity of sixteen dispiro-1,2,4,5-tetraoxanes: alkyl-substituted 7,8,15,16-tetraoxadispiro [5.2.5. 2]hexadecanes.** *Journal of Medicinal Chemistry, 43, 2753 -2758.*

- Vernick K.D., Keister D.B., Toure A. and Toure Y.T. (1996) : **Quantification of *Plasmodium falciparum* sporozoites by ribosomal RNA detection.** *American Journal of Tropical Medicine and Hygiene, 54, 430–8.*

- Vinciotti V., Khanin R., D'alimonte D., Liu X., Cattini N., Hotchkiss G., Bucca G., De Jesus O., Rasaiyaah J., Smith C. P., Kellam P. and Wit E. (2005): **An experimental evaluation of a loop versus a reference design for two-channel microarrays.** *Bioinformatics 21(4):492-501*. [Available at http://bioinformatics.oxfordjournals.org/cgi/content/ full/21/4/492].

- von Seidlein L., Milligan P., Pinder M., Bojang K., Anyalebechi C., Gosling R., Coleman R., Ude J. I., Sadiq A., Duraisingh M., Warhurst D., Alloueche A. Targett G., McAdam K., Greenwood B., Walraven G., Olliaro P. and Doherty T. (2000): **Efficacy of artesunate plus pyrimethamine-sulphadoxine for uncomplicated malaria in Gambian children: a double-blind, randomised, controlled trial.** *Lancet*, 355, 352 - 357.

- Wang D. , Urisman A., Liu Y. T., Springer M., Ksiazek T. G., Erdman D. D., Mardis E. R., Hickenbotham M., Magrini V., Eldred J., Latreille J. P., Wilson R. K., Ganem D. and DeRisi J. L. (2003): **Microarray-based detection and genotyping of viral pathogens**, *Proceedings of the National Academy of Sciences, U S A,* 99, pp. 15687–15692.

- Wellcome Trust Advanced Course (2007): **Functional Genomics and System Biology**. *The Sixty Second Wellcome Trust Advanced Course, Cambridge, UK.*

- World Health Organisation (1996): **A rapid dipstick antigen capture assay for the diagnosis of falciparum malaria.** *Bulletin of World Health Organisation, 74, 47–54.*

- World Health Organisation (2001): **Antimalarial Drug Combination Therapy Report of a WHO Technical Consultation**. *Geneva: World Health Organisation.* (WHO/CDS/RBM/ 2001.35).

- World Health Organisation (*2006*): **Guidelines for the treatment of malaria.** *World Health Organisation, WHO/HTM/MAC/2006.1108.*

- World Health Organisation / United Nations International Children's Emergency Fund (2003): **Africa malaria report.** *WHO/CDS/MAL/2003. 1093, Geneva: World Health Organisation*. [http://mosquito.who.int/amd2003/amr2003/pdf/amr2003.pdf]

- Wicker N., Dembele D., Raffelsberger W., and Poch O. (2002): **Density of points clustering, application to transcriptomic data analysis.** *Nucleic Acids Research, 15; 30(18): 3992–4000.*

- Wikipedia (2005): **Euclidian distance**. [http://en.wikipedia.org/wiki/Euclidean_ distance]. Retrieved on 20-2-2005.

- Wikipedia (2008): **Gene expression profiling.** http://en.wikipedia.org/wiki/ Expression_profiling. Retrieved on 24-12-2008.

- Wikipedia (2007): **Real-Time Polymerase Chain Reaction.** [http://en.wikipedia.org/wiki/Real-time_polymerase_chain_reaction]. Retrieved on 20-2-2007.

- Wikipedia (2006): **Hierarchical Clustering.** [http://en.wikipedia.org/wiki/ Hierarchical_clustering]. Retrieved on 12-11-2006.

- Wirth D. F. (2002): **The Parasite Genome, Biological Revelations**. Nature 419, 495-496.

- Wosik E. (2006): **Oligonucleotide Array.** http://cnx.org/content/m12385/latest/ Retrieved on 31-3-2007.

- Xu J. and Vernick K. (2006): **Introduction to Microarray Technology.** *World Health Organisation / Tropical Disease Research Workshop Training manual, Malaria Research Training Centre, Mali.*

- Yeh I., Hanekamp T., Tsoka S., Karp P. D. and Altman R. B. (2004): **Computational Analysis of Plasmodium falciparum Metabolism: Organizing Genomic Information to Facilitate Drug Discovery.** *Genome Research* 2004, 14: 917-924

- Yoeli M., Vanderberg J., Upmanis R.S. and Most H. (1965): **Primary tissue phase of *Plasmodium berghei* in different experimental hosts.** *Nature 208, 903.*

- Zakeri S., Avazalipoor M., Mehrizi A. A., Djadid N.D. and Snounou G. (2007): **Restricted T-Cell Epitope Diversity In The Circumsporozoite Protein From *Plasmodium falciparum* Populations Prevalent in Iran.** *American Journal of Tropical Medicine and Hygiene, 76(6), pp. 1046–1051.*

- Zha H., Ding C., Gu M., He X. and Simon H. D. (2002): **Spectral relaxation for K-means clustering**. *Advances in Neural Information Processing Systems*, 14: 1057-1064.