THE DETECTION AND MOLECULAR CHARACTERISATION OF BACTERIAL SYMBIONTS OF *ANOPHELES GAMBIAE* S.L.

by

CHARLES ADDOQUAYE BROWN

This thesis is submitted to the University of Ghana in partial fulfilment of the requirements for the award of **Ph.D Biochemistry** degree.

SEPTEMBER, 2003.

DECLARATION

This thesis is the result of research work undertaken by Charles A. Brown in the Department of Biochemistry, University of Ghana, under the supervision of Prof. M.D. Wilson and Prof. F. N. Gyang.

(Charles A. Brown)

(Prof. F.N. Gyang)

(Prof. M.D. Wilson)

DEDICATION

This thesis is dedicated to Joana, for all that she went through to enable me complete it, Choch and all the Nii "Atians".

ACKNOWLEDGEMENTS

I am glad to have this opportunity of expressing my deepest gratitude to my supervisors, Prof. M.D. Wilson and Prof. F.N. Gyang, who provided guidance and encouragement throughout the period of this work. Their invaluable comments greatly helped to improve the quality of the work. Special thanks go to Dr D.A. Boakye (Noguchi), Prof. T. Unnasch and Dr Tariq Higazi (both of the University of Alabama at Birmingham, USA) for their diverse contributions.

My sincere thanks goes to Nkem Okoye, Mrs. Anita Ghansah, Mrs Bridgette-Marian Ogoe, Nancy Duah, Janet Midega, Mrs Benedicta Kuivi, Fred Aboagye-Antwi, S. Dadzie, Evans Glah, Helena Baidoo, Naiki Puplampu, Awo Osafo-Addo, F.C. Mills-Robertson, Harry Asmah, Shelly, Tolu, Lydia, Abena, BB and Mrs Susan Adu-Amankwah whose support kept me going. I am equally grateful to Mr Nana M. A. Appawu, Dr K. Bosompem and Dr. K. Koram for their encouragement. A word of appreciation goes to all members of the Parasitology Unit (NMIMR) and lecturers of the Biochemistry Department for their continued interest and encouragement.

Finally, but by no means the least, I thank my very good friends Hassan Hassan, David Mensah, Akosua Boafo, S.B. Ofei, Hetty and Esi Colecraft for their encouragement, moral support and prayers during times of frustration.

God Bless You All.

This work was supported in part with a grant from the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (TDR).

ABSTRACT

The aim of the present study was to identify and characterise bacterial symbionts in *An. gambiae* s.l. mosquitoes, and ultimately select one that occurs in all its life stages. Mosquito larvae and pupae samples were collected from six locations in the Greater Accra region of Ghana and some reared to adults in the laboratory. Wild adult *An. gambiae* mosquitoes from Navrongo and Dodowa (Ghana) and Jaribuni (Kilifi District, Kenya) and laboratory colonies of adult *An. gambiae* mosquitoes from Kilimanjaro (Tanzania), Suakoko (Liberia) and Kisumu (Kenya) and *An. arabiensis* (Wageningen strain) were also studied. Each specimen was surface sterilized before DNA extraction was carried out. The PCR method of Scott *et al.* (1993) was used on each specimen to determine the species of the *An. gambiae* complex. PCRs using universal eubacteria 16S and 23S rDNA oligonucleotide primers were first used to detect the presence of the microorganism's DNA sequences in the mosquito, then PCRs using WOLB16SF1/WOLB16SR1 and *ftsZ1/ftsZ2* primers were carried out on positive reactions to determine whether or not they were *Wolbachia* sp. *In silico* (computational molecular biology) analyses of DNA sequences of *Escherichia coli* and *Pantoae agglomerans* were performed using DIGEST software and the results compared with those obtained by restriction analysis of the amplified 16S and 23S rDNAs with six enzymes. The PCR products were cloned and sequenced using an ABI 377 automated sequencer. Two consensus DNA sequences were generated from the sequence data by pairwise sequence similarity and the phylogram methods. Phylogenetic relationships of the consensus sequences to homologous sequences in the databases were inferred using neighbour-joining (NJ), maximum likelihood (ML), and maximum parsimony (MP) methods. Of the 432 mosquito specimens studied, 373 (94.4%) were identified as *An. gambiae* s.s., 20 (5.1%) as *An. arabiensis*, and one (0.25%) each of *An. merus and An. melas*. 295 *An. gambiae* (74.7%) adults, 29 (7.3%)

pupae and 71 (18%) larvae were studied. DNA fragments of the predicted sizes were successfully amplified using the 16S rDNA and 23S rDNA primers in 85.1% (336/395) and 79.5% (314/395) of the specimens, respectively. Bacterial DNA sequences were amplified from all the sibling species, which consisted of 54 (76.1%) larvae, 26 (89.7%) pupae and 256 (86.8%) adults from both wild and laboratory reared specimens, irrespective of the geographical origin. Out of 281 specimens, PCR positives for 16S and 23S rDNA primers, 94 (33.5%) were positive for the WOL16S rDNA primers but none for the *ftsZ* primers. Failure of PCR with *ftsZ* primers indicated a possible absence of *Wolbachia* infection. The restriction studies revealed that none of the amplified PCR products could be either *E. coli* or *P. agglomerans*. Seven sequences, namely AgA1 (AY247165), AgA2 (AY325810), AgL1 (AY247160), AgL2 (AY247161), AgL3 (AY247162), AgL4 (AY247163) and AgL5 (AY247164) were obtained, four and two with high homology to each other and one, a stand alone. CONSEN4 and CONSEN2 were highly homologous to 16S rDNA sequences of *Paracoccus* and *Rhodobacter* species. The sequence AgL5 was highly homologous to the 16S rDNA sequences of the Cytophaga-Flexibacter-Bacteroides group (CFB) group. The phylogenetic trees constructed indicated that CONSEN4 and CONSEN2 probably belong to the α subdivision of the Proteobacteria whilst AgL5 is a member of the CFB group. Since *Paracoccus* and *Rhodobacter* sp. are Gram-negative aquatic bacteria and members of the CFB group are also aquatic bacteria, it is likely that they occur in the breeding habitats of mosquitoes. Further research is needed; (1) to determine if they are in the breeding habitats and are ingested by larvae, (2) to culture and isolate them, and (3) to determine which part of the adult mosquito they reside in.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

bp	base pairs
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
ddw	de-ionized distilled water
dGTP	deoxyguanosine triphosphate
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide phosphate
dsDNA	double-stranded DNA
dTTP	deoxythymidine triphosphate
EDTA	ethylene diamine tetraacetate. $2H_2O$
EtBr	ethidium bromide
EtOH	ethanol
GPS	global positioning system.
HBI	human blood index
H_2O	water
kb	kilobase
KOH	potassium hydroxide
LB	Luria Bertani broth
M	molar (moles per litre)
µl	microlitre
µM	micromolar
ML	Maximum likelihood
ml	millilitre
MP	Maximum parsimony
mRNA	messenger RNA
Mw	molecular weight
NaOH	sodium hydroxide
NJ	Neighbour-joining
OTU	operational taxonomic unit

PAUP	Phylogenetic Analysis Using Parsimony
PCR	polymerase chain reaction
pH	hydrogen-ion exponent
PHYLIP	PHYLogeny Inference Package
rDNA	ribosomal DNA
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
rRNA	ribosomal RNA
sddH_2O	sterile double distilled water
s.l.	*sensu lato*
s.s	*sensu stricto*
ssDNA	single-stranded DNA
TAE	Tris - acetate EDTA
Tm	melting temperature
Tris	2 -amino-2-(hydroxymethyl)-1,3 propanediol
U	unit

If you would see all of Nature gathered up at one point, in all of her loveliness, and her skill, and her deadliness, and her sex, where would you find a more exquisite symbol than the Mosquito?

- HAVELOCK ELLIS, 1920

CHAPTER ONE

INTRODUCTION

Vector-borne diseases are major causes of death and morbidity worldwide. Not only do they create public health problems, they also present serious obstacles for socio-economic development in tropical countries.

Of the disease-causing vectors, mosquitoes transmit some of the world's most life-threatening and debilitating parasitic and viral diseases, including malaria, Venezuelan equine encephalitis, yellow fever, filariasis and dengue/dengue haemorrhagic fever (Miller, 1992; Priest, 1992; Monath, 1994). Of these, malaria is by far the most prevalent tropical vector-borne disease.

Malaria, the vector of which is the female *Anopheles* mosquito, is estimated to represent 2.3% of the overall global disease burden and 9% of Africa's; it ranks third after pneumococcal acute respiratory tract infections (3.5%) and tuberculosis (2.8%) [WHO, 1999]. An estimated 2.3 billion people, almost one-third of the world's population, are at risk of infection with the malaria parasite (WHO, 1999). The annual incidence of the disease is estimated to be 300 – 500 million clinical cases with mortality accounting for between 1-3 million deaths among children under five years. In sub-Saharan Africa it accounts for more than 90% of the total global incidence and mortality (Breman, 2001; WHO/UNICEF, 2003). It is a major cause of death in pregnant women (Lindsay *et al.*, 2000) and also an important cause of stillbirths and low birth weights (Menendez, 1995; Steketee *et al.*, 2001). These figures are probably underestimates as the statistics used in their collection may vary by a factor of three depending on the method of estimation (WHO, 1999). The 28 million reported cases of malaria in Africa are believed to

represent only 5-10% of the total malaria incidence on the continent (Hamoudi & Sachs, 1999).

Globally, the numbers of malaria cases are increasing and the rate of increase is accelerating. This pattern is illustrated by multifold increases in malaria rates since 1979 in South America (Roberts *et al.*, 1997) accompanied by a rise in the proportion of populations at high risk to the disease. Malaria is appearing in urban areas and in countries that had previously succeeded in eradicating the disease, for example, in urban areas of the Amazon Basin (Roberts & Laughlin, 1999), South and North Korea (Feighner *et al.*, 1998), Armenia, Azerbaijan, and Tajikistan (Curtis, 1999). In Africa, the prevalence of malaria has been escalating at an alarming rate, within the last decade. Between 1994 and 1999, malaria epidemics in 14 sub-Saharan Africa countries caused an unacceptable high number of deaths, many in areas previously free of the disease (OAU, 1997). Additionally, the increase in cases and the altered geographic distribution of malaria is underestimated because accurate information on global incidence is difficult to obtain, since reports are generally fragmentary and irregular (Roberts *et al.*, 2000).

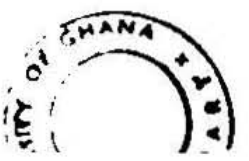
Air travel has brought the threat of the disease to the doorsteps of industrialised countries with an increasing incidence of imported cases and deaths of visitors to endemic areas. Health authorities in many countries are becoming increasingly concerned about the potentially deadly risks of malaria carried into their territory by "jet-setting" mosquitoes that travel on international flights and spread the disease (Gratz *et al.*, 2000). Between 1969 and 1999, 12 countries reported a total of 87 cases of malaria in people living near an airport. France headed the list, with 26 cases, followed by Belgium, 16, and the United Kingdom, 14 cases (WHO, 2000). The occurrence of the relatively large number of cases of airport malaria in Paris and Brussels reflects the large number of flights arriving from

Central and West Africa. These "airport malaria" cases, occurring in or near airports, are distinguished from other cases of imported malaria among persons who contract the infection during a stay in a malarious area and subsequently fall ill (Lusina *et al.*, 2000; Gratz *et al.*, 2000). The majority of the cases were caused by *P. falciparum*. At least five deaths have resulted; all cases occurred among non-immune individuals, accounting for a relatively high mortality of 6% (Gratz et al., 2000).

The prevalence of malaria has an enormous impact on a country's economy incapacitating part of the labour force and thus affecting productivity. The number of DALYs (disability adjusted life years) lost in 2000 to malaria has been estimated to be 42, 280, 000 (WHO/UNICEF, 2003). It is also estimated that a bout of malaria, depending on the severity, typically entails a loss of four or more working days, followed by additional days with reduced work capacity (Shepard *et al.*, 1991; Picard & Mills, 1992; Hempel & Najera, 1996). In Africa, where malaria accounts for up to a third of all hospital admissions, this costs a sum equivalent to over 10 working days, adding to the continent's economic burden. Malaria attacks are also a major cause of school absenteeism and appear to negatively impact on long term learning capacity (McDonald, 1950; Wernsdorfer & McGregor, 1988; Holding & Snow, 2001). Endemic malaria also reduces the growth potential for some industries, notably tourism and transportation, and sharply raises the cost of infrastructure projects and other collective enterprises (Gallup & Sachs, 2001; Sachs & Malaney, 2002). Some researchers have estimated the economic burden of malaria at 0.6-1.0% of GDP in Africa, although recent reports indicate that the economic impact of the disease on national income is likely to be much higher (Breman, 2001; Sachs & Malaney, 2002). The economic loss due to malaria in Africa is in excess of two billion US dollars per year (Okenu, 1999). Malaria thus has social consequences and is a heavy burden on economic development.

Different approaches have been used to control malaria. The control methods include the use of insecticides, chemotherapy and management of the environment. For more than 80 years, insecticides, most notably dichlorodiphenyltrichloroethane (DDT), have been one of the primary means of controlling insect-borne diseases. However, resistance to insecticides has appeared in the major insect vectors from every genus (Brogdon & McAllister, 1998). According to the World Health Organization approximately 125 arthropod species are resistant to at least one, and often two or more, insecticides (WHO, 1992). Resistance has also developed to every chemical class of insecticide, including microbial drugs and insect growth regulators (Brogdon & McAllister, 1998). In many parts of the world where insect-borne diseases cause illness and death, insecticides are available. However, the partial withdrawal of available effective insecticides due to resistance in target species, and the non-renewal of the registration of some insecticides (Lacey & Orr, 1994), coupled with environmental safety issues and the high cost associated with heavy insecticide usage (Georghiou, 1986), make sustaining long term insecticide vector control extremely costly and practically unachievable.

Drugs to treat malaria have been used for thousands of years and WHO and governments of malarial countries have turned toward drug treatment strategies to reduce malarial incidence rates. During the early 1900s, doctors used quinine for malaria therapy. The use of the drug was short-lived and was replaced with chloroquine, a cheap, safe, and effective anti-malarial that gained widespread acceptance among doctors in endemic countries during the 1950s and resulted in an enormous decline in malaria incidence rates (WHO, 1997). But after years of use, because it was either misused or abused, chloroquine-resistant malaria parasites have evolved (NIAID, 2000). From the 1950s to the present, chloroquine resistance has gradually spread to nearly all *falciparum* malaria endemic regions (WHO, 1997; NIAID, 2000).



Newer drug therapies, unfortunately, have not been effective against drug resistant strains of the malaria parasite. The drug mefloquine was introduced in Southeast Asia in the mid-1980s, but complete drug resistance was observed only four years thereafter (Basu, 2000) and is now almost at the same level as chloroquine (Nchinda, 1998). Resistance to a more recent drug atovaquone developed so quickly that resistant strains were detected during clinical trials (Strobel, 1999). Artemisinin and its derivatives (artemether, arteether, and artesunate), which were developed as a result of the more serious drug resistance situation in Asia, are currently the most effective of all anti-malarial drugs. However, it has been shown in laboratory studies that malaria parasites can also become resistant to artemisinin (NIAID, 2000). Moreover, according to scientists studying drug resistance, the malaria parasite frequently mutates and can therefore evolve resistance to nearly any drug (Basu, 2000).

The importance attached to disease control, especially of *P. falciparum* malaria, for example, has motivated research into the development of efficient vaccines. Vaccination against *P. falciparum* and *P. vivax* is seen as the method of intervention with the greatest potential to reduce the morbidity and mortality associated with severe malaria in areas of intense transmission (Miller & Hoffman, 1998) and even contribute to eradicating the disease (Greenwood, 1997). For these reasons, considerable effort and several hundred million dollars have been spent over the last 30 years to develop vaccines (Collins & Paskewitz, 1995). Unfortunately, vaccine development is associated with numerous difficulties: such as antigenic diversity and immune evasion of the parasites: difficulties for the large-scale production of clinical grade material for human trials; lack of good *in vitro* correlates of protection; and lack of adequate or proven delivery systems or adjuvants for use in humans (Alonso & Dgedge, 1999). Although vaccine research

remains a high priority, it is also important to look for alternative approaches to control vector-borne tropical diseases like malaria, while vaccine research continues.

Significant progress has been made in the research and development of new tools for malaria control. Insecticide treated bednets (ITN) and curtains have emerged in recent years as one promising tool. Several studies in 6 African countries have demonstrated the efficacy of ITN in reducing significantly morbidity and mortality in infants (Alonso *et al.*, 1991; Nevill *et al.*, 1996; Binka *et al.*, 1996; Lengeler & Snow, 1996; Abdulla *et al.*, 2001). However, the costs of the nets and treatment still inhibit wide-scale use. In addition, further work remains to be done on different insecticide-fabric combinations and it is not yet known what percentage of the beds in a community need to be equipped with nets to achieve effective malaria control (NIAID, 2000). There is also the threat of physiological and behavioural resistance to pyrethroids (used for impregnating the nets) by mosquitoes, of increase in severe disease in areas of intense transmission due to the reduction in transmission and the possibility that the initial successes of bednets may disappear during long term application, because of the fading of pre-existing immunity levels (NIAID, 2000).

To date, the only effective control programmes for malaria have been those that targeted the mosquito vector (TDR, 2001). Furthermore, since the vector remains the key link in the transmission of malaria, it warrants a new and serious research effort. In this regard, recent biotechnological advances in the area of host-parasite relationships and vector genetics suggest the possibility of revolutionary methods for malaria control (TDR, 2001). The ultimate aim of these biotechnological approaches is to use recombinant DNA technology to replace a highly competent vector population with an otherwise identical population engineered to be an incompatible host for the malarial parasite

(Crampton *et al.*, 1990, 1994; Collins & Besansky, 1994). Thus mosquitoes would exist in their normal environment, but be unable to transmit malaria. This approach is highly desirable because it does not seek to exterminate mosquitoes, as such a step would be technically very difficult and the ecological consequences profound, as mosquitoes are a significant link in the food chain (TDR, 2001).

Several basic problems remain to be solved before such a biotechnology strategy can be clearly implemented. One such problem is finding ways of introducing new or altered genes into the mosquito genome. In principle this can be achieved either by the use of transposable genetic elements (Kidwell & Ribeiro, 1992; Coates *et al.*, 1998), viral transfection (Carlson *et al.*, 1995; Olson *et al.*, 1996) or by genetic manipulation of bacteria symbionts (Beard *et al.*, 1993a; Duravasula *et al.*, 1997). The problem of transfections with exotic viruses is that no efficient entomovirus is currently available and none has been economically produced in sufficient quantity to control disease vectors (Collins and Paskewitz, 1995). Efforts to transform anophelines have been pursued since the first reported case of foreign DNA introduction into *An. gambiae* genome in 1987 (Miller *et al.*, 1987). These were intensified after the successful development of routine transformation techniques using the *Minos* transposable element in the Mediterranean fruitfly *Ceratitis capitata* (Loukeris *et al.*, 1995) and the *Mariner* and *Hermes* elements in *Aedes aegypti* (Coates *et al.*, 1998; Jasinskiene *et al.*, 1998). The ability of the *Minos* transposable element to function as a transformation vector in anopheline mosquitoes has recently been demonstrated (Catterucia *et al.*, 2000).

The potential of using naturally occurring symbiotic bacteria that can be genetically engineered for the control of mosquitoes, on the other hand, has received much less attention. With this approach, the arthropod is not transformed, but the symbiotic bacteria

that it harbours are (Beard *et al.*, 1993a). Such arthropods are called paratransgenic. This approach is guided by the following observations and basic concepts: 1) Throughout the entire developmental cycle many arthropods, especially those that feed on restricted food sources such as blood, cellulose, phloem and stored grains, harbour bacterial symbionts; 2) Some of these symbionts can be cultured and genetically transformed to express a gene whose product kills a pathogen that the arthropod transmits; 3) Normal arthropod symbionts can be replaced with genetically modified symbionts, resulting in a population of arthropod vectors that can no longer transmit disease (Beard *et al.*, 1993a).

A wide variety of bacterial symbionts have been identified in practically every Order of the Insecta and in ticks. In those insects that feed on fluid material such as plant juices or blood, the symbionts appear to provide metabolic products that supplement the host insect's nutrition (Beard *et al.*, 1993a). Examples of such symbionts include *Wolbachia* in drosophilids and parasitic wasps (O'Neil *et al.*, 1992; Stouthamer *et al.*, 1993), Rickettsia-like organisms (RLO) from tsetse flies *Glossina* spp.(Beard *et al.*, 1993b) and ladybeetles (Werren *et al.*, 1994) and *Rhodococcus rhodnii* in the vector of Chagas disease, *Rhodnius prolixus* (Beard *et al.*, 1993a). For mosquitoes, *Wolbachia* has been reported in the Asian tiger mosquito, *Aedes albopictus* (Braig *et al.*, 1994) and in *Culex pipiens* (O'Neil *et al.*, 1992). However, there have been very few reports of bacteria symbionts in Culicids and there is no reason for symbionts not to occur in *An. gambiae* If symbiotic bacteria occur in *An. gambiae*, then it becomes possible to exploit them as vehicles for the introduction of foreign genes that can express anti-*Plasmodium* proteins. The feasibility of using transgenic endosymbionts to alter the disease carrying capacity of a disease vector has been demonstrated in *R. prolixus* (Duravasula *et al.*, 1997). Moreover, using bacteria that have specialised and specific symbiotic association with insect hosts to spread transgenes greatly reduces the chance of unwanted gene spread.

Before the development of assays utilizing molecular techniques, the quantitative documentation of the presence of bacteria in blood-sucking arthropods such as ticks, mites and mosquitoes, depended upon specific staining techniques and serological-based assays. However, biological staining may be unreliable whilst serological-based assays require species-specific antisera (Higgins & Azad, 1995). Moreover, very few symbionts have been cultured outside their host in cell free systems, thereby preventing even a traditional bacterial taxonomic placement (O'Neil, 1997), though in a few cases, symbionts have actually been isolated and grown in artificial culture medium (Welburn *et al.*, 1987; Welburn & Dale, 1997).

Although the fastidious nature of symbionts makes their isolation and *in vitro* cultivation difficult, the advent of polymerase chain reaction (PCR) has greatly aided efforts to characterise these symbionts at the molecular level, and also to monitor changes that occur in vector populations over a period. Ribosomal gene primers specific for *Wol. pipientis* have been used to detect this symbiont in insects belonging to several different Orders (O'Neil *et al.*, 1992). Rousset *et al.* (1992) described the use of 16S and 23S ribosomal RNA (rRNA) primers to amplify segments of these genes from *Wolbachia* in terrestrial isopods, as well as insects. There are several published oligonucleotide primer sets for amplifying sequences of Rickettsiae-like symbionts (O'Neil *et al.*, 1992; Rousset *et al*. 1992a; Higgins and Azad, 1995) and Eubacteria (Weisburg *et al.*, 1991) in several insects. These can be used to screen initially for symbionts in *An. gambiae*. For identification purposes and to investigate the phylogenetic relationships of these bacteria, the sequences of the PCR amplified DNA fragments can be compared to those that have been published and deposited in DNA databanks such as the GENBANK. Such phylogenetic studies can assist in developing successful culture conditions *in vitro* as well as in identifying suitable plasmid transformation vectors.

1.1 Objectives

The ultimate goal of this study is to identify a microbial symbiont that is universal in *An. gambiae* s.l. and which can potentially be genetically manipulated to express anti-*Plasmodium* proteins in adult mosquitoes. The specific objectives are:

1. To use published oligonucleotide primer sequences to amplify by PCR, the DNA sequences of eubacteria and rickettsiae-like organisms (RLO) of *An. gambiae*.
2. To confirm the identity of the symbiont's amplified DNA sequence using restriction analysis.
3. To clone and sequence the amplified DNA fragments.
4. To identify the species of the symbionts by conducting bio-informatic search for homologous sequences in gene database banks.
5. To determine the symbiont's phylogenetic relationships with related organisms.

CHAPTER 2

LITERATURE REVIEW

2.1 Mosquitoes

2.1.1 General introduction

Mosquitoes are perhaps the most familiar of all blood-sucking insects. They are of tremendous significance to man from both the economic and health point of view, because of their role as parasite vectors. They include the only organisms able to transmit human malaria, and, apart from carrying this and other diseases, are almost unrivalled as irritating biting pests.

Aristotle was the first to chronicle mosquitoes, which he referred to as "empis" in his "Historia Animalum" in 300B.C in which he documented their life cycle and metamorphic abilities (Floore, 2000). Although it has been claimed that the name "mosquito" is either a Spanish or a Portuguese word meaning "little fly", Floore (2001) postulates that the word is apparently North American, and dates back to about 1583. In Europe, mosquitoes were called "musketas" by the Spanish, "gnats" by the British, "Les moucherons" or "Les cousins" by the French, and "Stechmucken" or "Schnacke" by the Germans.

2.1.2 Classification and identification

Mosquitoes belong to the order Diptera, sub-order Nematocerca of the class Insecta which is the most dominant group of the Phylum Arthropoda. Mosquitoes are readily distinguished from other similar-looking flies in the sub-order (Nematocerca) by their conspicuous forwardly projecting proboscis, scales on the thorax, legs, abdomen and wing veins, and fringe of scales along the posterior margin of the wings (Service, 1993; Kettle, 1995). All mosquitoes belong to the family Culicidae, which is divided into three

sub-families Culicinae, Anophelinae and Toxorhynchites, and 35 genera (White, 1996). The Culicinae (culicines) and the Anophelinae (anophelines) both contain blood feeding man-biting species that are important disease vectors, and it is important to be able to distinguish between these sub-families. The Figure 2.1 illustrates the criteria which are usually used to distinguish them.

These two mosquito subfamilies also differ significantly in their genomic structures. Anophelinae have heteromorphic sex chromosomes, a small genome size, and repetitive elements that are distributed in a long-period interspersion pattern. In contrast, Culicinae have homomorphic sex chromosomes, and repetitive DNA that is organized in a short-period interspersion pattern (Rai & Black, 1999).

The Anophelinae contain about 450 species in three genera, *Bironella* (which is Australasian in distribution, occurring mainly in the Papuan subregion), *Chagasia* (which is Neotropical) and *Anopheles* (which contains by far the largest number of species, 437, and is nearly worldwide in distribution) [Lehane, 1991; Sallum *et al.*, 2000]. The phylogenetic relationships among *Anopheles*, *Bironella*, and *Chagasia* were evaluated by Harbach and Kitching (1998) using morphological characters. The Anophelinae were found by the authors to be monophyletic, with *Chagasia* occupying a basal position within the subfamily, and *Anopheles* sharing a sister-group relationship with *Bironella*.

The Culicinae, the largest sub-family, is divided into two tribes the Culicini and Sabethini (Service, 1993), with almost 2500 species in over 30 genera (White, 1996). The main medically important genera are *Aedes*, *Culex* and *Mansonia* species.

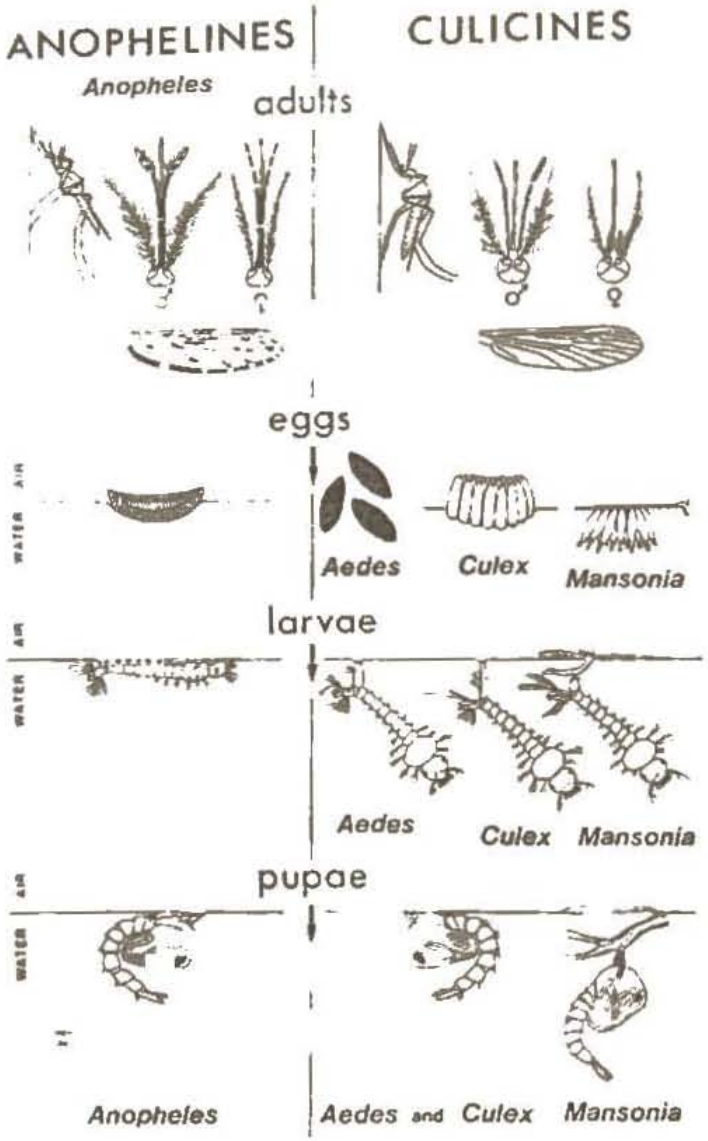


Fig. 2.1 Chart showing the main differences between mosquitoes of the subfamilies Anophelinae (left) and Culicinae (right). (From Service, 1993).

The main characteristics of anophelines which distinguish them from culicines are: adult anophelines rest with their bodies at an angle of 30-45° to the surface with the head, thorax and abdomen in a straight line, have dark and pale scales on the veins of the wings arranged in distinct blocks and palps about as long as the proboscis in both sexes; eggs have floats and are laid singly on the surface of water; larvae lack a siphon and rest with their bodies parallel to the water surface, below the surface film, and pupae have breathing trumpets which are short with a broad opening. In contrast, adult culicines rest with their bodies roughly parallel to the surface, have scales on the wing veins – but not arranged in blocks and palps which are much shorter than the proboscis in females. Eggs are not provided with floats and are either laid separately (*Aedes*), or stacked in a floating raft (*Culex*) or in a mass on floating vegetation (*Mansoni*); larvae have a long or short siphon and rest with their bodies at an angle to the water surface; and pupae have short or long breathing trumpets with a narrow opening.

A high proportion of mosquitoes belong to the genus *Aedes* in which there are more than 1000 species, classified as 40 subgenera. *Aedes* species are distributed throughout the world, especially in temperate countries (White, 1996). Many are vectors of arboviruses, including several of the most important mosquito-borne human viral diseases (e.g. yellow fever, Japanese encephalitis, dengue fever, Rift valley fever and West Nile fever). Representatives of three subgenera of *Aedes* are natural vectors of *Brugia malayi* and *Wuchereria bancrofti* that cause human disease (White, 1996). The subgenus, *Stegomyia*, is probably of the greatest medical interest. *Aedes* (*Stegomyia*) *aegypti* is the most widespread and most dangerous species in this subgenus. It is the principal vector of chikungunya and dengue viruses in almost every outbreak. Urban yellow fever is also mainly transmitted by *Ae. aeygpti* in Central and South America and in West Africa (White, 1996).

About 800 species of *Culex* are known, being classified in 21 subgenera, with many species acting as the vectors of enzootic arboviruses, protozoa and filariae (White, 1989, 1996). The typical subgenus *Culex* contains the majority of species. Japanese encephalitis virus is transmitted mainly by *Culex* spp. in the Oriental region, especially by the following species which breed prolifically in paddy and swamps: *Cx. tritaeniorhynchus*, *Cx. gelidus* and *Cx. vishnui* (Varma, 1989). *Culex theileri* in southern Africa and members of the *Cx. pipiens* complex in Egypt are important vectors of Rift Valley fever virus transmitted from livestock to man. In Australia, *Cx. annulirostris* plays a similar role in the epidemiology of Murray Valley encephalitis (White, 1996). The *Culex pipiens* complex comprises several species, subspecies and forms, with representatives in all parts of the world (Service, 1993). Typical *Cx. pipiens* occurs in temperate countries of the northern hemisphere, spreading through temperate highlands to southern Africa. In Egypt, the *Cx. pipiens* complex is responsible for bancroftian filariasis transmission

(Southgate, 1979) and has been implicated as the main vector in outbreaks of Rift Valley fever. *Culex quinquefasciatus* is the main man-biting tropical member of the *Cx. pipiens* complex, distributed up to 38°N in USA, 30°N in Asia, but only 24°N in Africa. Bancroftian filariasis is largely maintained in tropical villages and towns by *Cx. quinquefasciatus* alone, although West African strains of *W. bancrofti* do not develop in this species of mosquito (White, 1996).

The genus *Mansonia* has only a dozen American species, notably the widespread pest *M. titillans* which breeds in swamps and marshes with floating water lettuce *(Pistia)* and water hyacinth *(Eichornia)* or rooted *Pontederia* (White, 1996). *Mansonia titillans* is an important vector of Venezuelan equine encephalitis and may contribute to transmission of *W. bancrofti*. As with other troublesome *Mansonia* and some *Coquillettidia*, *M titillans* disperses far from the breeding sites in search of hosts to bite. Subgenus *Mansonioides* also has a dozen species, endemic to the Old World tropics, being pests arising from breeding sites in water with plenty of *Eichornia*, *Pistia*, rooted *Isachne* (swamp grass), *Zuzania* and other suitable vegetation (White, 1996). Some arboviruses are transmitted by *Mansonioides* spp., but these species of mosquito are generally refractory to *W. bancrofti*. The bites are more painful than with most other mosquitoes. *Mansonioides* spp. are of particular parasitological interest because, together with *Anopheles (Anopheles)* spp., they are the vectors of brugian filariasis (White, 1989). In South-East Asia, swamp-forest populations of *M. annulata*, *M. uniformis* and especially the sibling species *M. bonneae* and *M dives* share the transmission of zoonotic subperiodic *B. malayi*, which they transmit to man from leaf monkeys and other wild animals (Wharton, 1962).

The third sub-family, the Toxorhynchitinae, has only one genus the *Toxorhynchites*, and contains the world's largest mosquitoes (Service, 1993). There are some 76 mainly

tropical species of *Toxorhynchites* but a few occur in Japan, the maritime area of the far eastern parts of Russia and in the eastern USA northwards towards Canada (Service, 1993). Adults are readily recognised by the large size (up to 19 mm long, and 24 mm wing spread), metallic colouration and prominent curved proboscis. Females are incapable of blood feeding and both sexes suck nectar and other naturally occurring sugary substances. *Toxorhynchites* is consequently not involved in disease transmission (White, 1996).

Approximately 3300 species of Culicidae have been described, more being recognised each year (White, 1996). The most important man-biting species of mosquitoes belong to the genera *Anopheles, Culex, Aedes, Mansonia, Haemogogus, Sabethes* and *Psorophora* (Service, 1980).

2.1.3 Species complexes

Some mosquitoes occur as sibling species (or cryptic species) complexes, pairs or groups of closely related species that are morphological indistinguishable (isomorphic) but reproductively isolated, and which frequently live in the same area (i.e. are sympatric) [Service, 1993]. The first to be recognised as such is the *Anopheles maculipennis* complex (Service, 1993). Similarly, *An. gambiae* was also later shown to be a species complex (Coluzzi, 1966), and many other *Anopheles* species can now be considered as representing complexes, for example, *An. dirus* (Green *et al.*, 1992), *An. culicifacies* (Suguna *et al.*, 1988) and *An. quadrimaculatus* (Narang *et al.*, 1989a).

In many complexes, species-specific differences in ecology and behaviour can substantially affect both disease transmission and the success of methods used for control. For example, a differential response to insecticides has been noted in the *An.*

culicifacies complex in India. In this case, DDT resistance is mainly associated with species B, which is a very poor vector of malaria (Subbarao *et al.*, 1988).

2.1.4 Methods for distinguishing between cryptic species

Species within a complex by definition cannot be distinguished morphologically, or only with great difficulty. In most cases the benchmark definition of cryptic sibling species has been mating incompatibility (Collins & Paskewitz, 1996).

Mating compatibility studies were later replaced by polytene chromosome examinations, which were among the first to provide researchers with tangible genetic markers that could be used to differentiate between species (within cryptic complexes) and chromosomal forms (Coluzzi & Sabatini, 1968; Gatti *et al.*, 1977; Coluzzi *et al.*, 1979; Toure *et al.*, 1998). For routine species identification however, chromosomal markers have major limitations. A number of alternative species-diagnostic procedures have therefore been developed, including identification by electrophoretic analysis of gene-enzyme systems (species-specific allozymes) [Mahon *et al.*, 1976; Miles, 1978; Narang *et al.*, 1981, 1989a,b; Lanzaro *et al.*, 1995], hybridisation with DNA probes (Collins *et al.*, 1987; Gale & Crampton, 1987; Cockburn & Seawright, 1988; Panyim *et al.*, 1988) and quantitative differences in the cuticular hydrocarbon profiles (Carlson & Service, 1979, 1980; Hamilton & Service, 1983; Phillips & Milligan, 1986; Phillips *et al.*, 1988).

From these 'traditional tools' of cytological examination of polytene chromosomes, namely, genetic compatibility, immunological and hybridisation techniques and isozyme analysis (Favia *et al.*, 1994a; Munstermann & Conn, 1997), identification tools have expanded to include a vast array of molecular markers. These contemporary markers range from what are now referred to as 'classical genetic markers' such as mitochondrial

DNAs (Caccone *et al.*, 1996; Besansky *et al.*, 1997; Lehmann *et al.*, 1997; Thelwell *et al.*, 2000), species-specific restriction fragment length polymorphism (RFLP) analysis of ribosomal DNA (Collins *et al.*, 1988; Yasothornsrikul *et al.*, 1988; Torres *et al.*, 2000; Favia *et al.*, 2001), methods used to detect and identify single nucleotide polymorphisms (SNPs) [De Merida *et al.*, 1999] and finally to highly polymorphic markers such as random amplified polymorphic DNAs (RAPDs) [Wilkerson *et al.*, 1993; Favia *et al.*, 1994a], microsatellite DNAs (Zheng *et al.*, 1996, 1997; Lanzaro *et al.*, 1998; Conn *et al.*, 2001; Sharakov *et al.*, 2001a), and amplified fragment length polymorphisms (AFLPs) [Black & Lanzaro, 2001]. More recently, sequence-tagged amplified RAPDs and single-copy markers (Mukabayire *et al.*, 2001) have also been used to distinguish between cryptic species.

One of the greatest advantages of this wide variety of genetic markers is that researchers may choose to utilise any combination of markers or techniques to address multifaceted questions relating to malaria transmission by anopheline mosquitoes (Taylor *et al.*, 2001). These molecular markers have proven useful in a wide variety of applications including molecular taxonomy, evolutionary systematics, population genetics, genetic mapping, and a variety of molecular diagnostics.

2.1.4.1 'Traditional tools'

I) Morphological identification

The analysis of morphological features is still the most widely used technique for taxonomic and systematic studies of anopheline mosquitoes, and the internal classification of the genus *Anopheles* is based largely on morphological characters (Harbach, 1994). The available morphological diagnostic characters, although not reliable, are of definite taxonomic value for the distinction of the saltwater and the fresh

water sibling species of the *An. gambiae* complex. The number of sensilla coeloconica, the value of the palpal index and the shape of the eggs can separate these two groups (Service, 1993). However, none of these characters appears to be completely discriminant as the morphological characters of adults are variable and overlap in many instances (Coluzzi, 1964).

In the field, morphology still remains a very useful tool for identifying unrelated species, and enables the preliminary sorting of material prior to the application, if necessary, of other identification techniques. Such a procedure allows a considerable saving in time and money and the advantages offered by it have ensured its continued use. Unfortunately, however, the proliferation of species groups and complexes within many of the anopheline taxa is rendering the use of morphological characteristics invalid for the identification of important malaria vectors (Beebe & Cooper, 2000). With these cryptic species, it has been proposed that the present morphological studies have not been exhaustive enough and that more detailed observations on all stages of the life-cycle may yet yield diagnostic characters (White, 1977). However, species keys involving obscure characters such as the sculpturing of the egg chorion, larval chaetotaxy, male genitalia or the ratio of various adult female body structures, may not be applicable in the identification of large numbers of field-collected adult female mosquitoes.

ii) Reproductive incompatibility

One of the earliest techniques for species identification is that involving hybridisation experiments where field material of unknown taxa is cross-mated with members of known species. The progeny of these pairings are examined for anomalies in survival rates, sex ratios, gonad structure and fertility. The extent to which these anomalies occur in the offspring has also been used to indicate similarities between parent species (Bryan,

1973). This method requires no sophisticated technology and was widely used in early studies of species groups and complexes (Davidson, 1964; Bryan, 1973; Mahon & Miethke, 1982). Problems with this technique are that it is time consuming and laborious, and requires that a number of reference colonies be maintained. These practical limitations make it inappropriate for large-scale routine identification. Also, the fact that the parents are force mated, artificially overcomes any premating barriers that may have existed naturally in the field (Patterson, 1985; Baimai *et al.*, 1987).

iii) Cytogenetic studies

Cytogenetics involving the karyotyping of polytene chromosomes was one of the earliest tools for the study of anopheline genetics (Norris, 2002). This tool has proven immensely useful for differentiating among sympatric taxa and chromosomal forms, and remains as the only reliable tool to differentiate between all of the *An. gambiae* s.s. chromosomal forms (Coluzzi *et al.*, 1979, 1985); and until recently it was the only tool that could be reliably used to differentiate between all nine members of the *An. funestus* group (Green & Hunt, 1980). Sharakhov *et al.* (2001b) developed a cytogenetic map of *An. funestus* and compared it with the polytene chromosome organisation of *An. gambiae*. In addition, frequencies of chromosome inversion arrangements have been utilised for investigations of population structuring and estimates of effective population size (Taylor *et al.*, 1993; Petrarca *et al.*, 2000; Kamau *et al.*, 2002).

Chromosomal inversions have also been utilised to address phylogenetic issues regarding the origin, maintenance and introgression of inversions between sympatric populations and taxa for many anopheline species groups (Coetzee *et al.*, 1999). Such information has been utilised as evidence for genetic introgression of adaptive chromosomal variants between the two sibling species, *An. gambiae* s.s. and *An. arabiensis* (della Torre *et al.*,

1997). Similar comparisons of shared polytene chromosome banding patterns have been made among putative anopheline species in the New World as well (Pe´rez & Conn, 1992). In both cases, researchers must look to other markers to come to definitive conclusions concerning these homologies.

Among the disadvantages for using this technique in *Anopheles* mosquitoes are that the polytene chromosome preparations must be made from ovarian tissue or fourth instar larvae. This limits the samples to either adult bloodfed female mosquitoes or late instar larvae. In addition to the paucity of experienced personnel trained to read polytene chromosome preparations, markers are not abundant or particularly informative in some species (Lounibos & Conn, 2000). Despite the limitations, this method remains integral for much of the contemporary work.

iv) Cuticular hydrocarbon analysis

Intraspecific differences in the hydrocarbons of the insect cuticle has been capitalised on in the identification of cryptic species. Gas liquid or capillary gas chromatography is used to analyze the hydrocarbons in the cuticular wax extracts, and multivariate statistics is then applied to the chromatographic data to differentiate sibling species by quantitative analysis of the cuticular hydrocarbon peaks (Carlson & Service, 1979, 1980; Hamilton & Service, 1983; Phillips & Milligan, 1986; Phillips *et al.*, 1988). This has been found particularly useful for the differentiation of sibling species of *Anopheles* mosquitoes (Carlson & Service 1979, 1980; Hamilton & Service, 1983; Milligan *et al.*, 1986, Anyanwu *et al.*, 1993). Although the biological implications of variations in cuticular hydrocarbons between very closely related species have not been fully studied, Phillips and Milligan (1986) and Phillips *et al.* (1987) suggested that within such relationships, species-specific differences do exist which tend to highlight the effect of geographical

variation and possible incipient speciation mechanisms. The species-specific function performed by insect cuticular hydrocarbons in various environments has been outlined by Howard and Blomquist (1982) and Phillips *et al.* (1988). They serve mainly to prevent desiccation and to assist in "chemical communication" (Lockey, 1980; Howard & Blomquist, 1982). Milligan *et al.* (1990) argued that the hydrocarbon discrimination is based on the relative concentrations of the component chemicals, rather than their mere presence or absence. Similarly, Hamilton and Service (1983) observed that, although the fourth instars of *An. gambiae* and *An. arabiensis* had similar cuticular hydrocarbons, there were still differences in the relative levels of some of these chemicals. It is thought that hydrocarbons are not just passive protectors against desiccation; they may additionally play a leading role in mate selection/recognition and population divergence. Hydrocarbon differences among sympatric populations of *An. gambiae* s.s. (Phillips *et al.*, 1987) might reflect a semiological function of the compounds, enabling the insect to recognise potential mates in those locations where sibling species coexist. In some insects, part of the mate recognition mechanism has been linked with the detection of specific hydrocarbon compounds and other components of the cuticular lipid layer, such as fatty acids, alcohols, sterols, and aldehydes (Jallon, 1984; Bonavita-Courgourdan *et al.*, 1987; Pescke, 1987).

v) Allozyme/isozyme analysis

Species-specific allozymes analyses have been used for several years in the identification of sibling and cryptic species in mosquitoes (Pasteur *et al.*, 1981; Cianchi *et al.*, 1985). This technique, based on the electrophoretic separation of enzyme variants, allows quantification of the amount of gene differentiation among populations, showing evolutionary diversity up to species level (Bullini & Coluzzi, 1978); biochemical keys are now in use for the differentiation of several species of a multitude of organisms. Early

work, utilising these biochemical markers in *An. albimanus*, were used to document significant amounts of polymorphism among laboratory colonies that were then used in linkage mapping studies (Narang *et al.*, 1981). Careful analysis and use of these techniques led to the discovery of diagnostic allozyme loci for wild sympatric populations of *An. quadrimaculatus* A and B (Lanzaro *et al.*, 1990) and allowed broader use of these tools for the population genetics of natural vector populations (Hii *et al.*, 1988). This tool was also fashioned into a dichotomous electrophoretic taxonomic key for three species within the *An. quadrimaculatus* complex (Narang *et al.*, 1989a, b) and was utilised to delineate two sibling species within the *An. minimus* complex in Thailand (Sucharit *et al.*, 1988). Among the drawbacks of isozyme analysis are that specimens must be fresh or kept frozen until analysis (Richardson *et al.*, 1986) and that the procedure itself requires a relatively large amount of sample material compared to the few nanograms of DNA required for PCR (Norris, 2002). Finally the cost of the technique is relatively high, especially, considering its tediousness and the amount of time needed to perform large-scale studies as compared to other techniques. Nonetheless, isozymes have provided valuable data on which much of the contemporary work on anophelines is based.

vi) Hybridization assays

Most of the first DNA based work for identifying cryptic species were based on hybridization assays that detected species-specific differences in highly repetitive sequences (Collins *et al.*, 2000). Although somewhat time consuming and technical, these assays were more efficient than previous methods, such as polytene chromosome analysis and isoenzyme gel electrophoresis. The approach, however, generally had two major limitations (Collins *et al.*, 2000). The species composition of the cryptic species complex under study had to be known before the assay could be developed, and each species in the complex had to differ from the others in the abundance of one or more

repeat sequences. Despite these limitations, repeat sequence hybridization assays were developed and used for field studies of the *An. gambiae*, *An. dirus* and *An. punctulatus* complexes (Gale & Crampton, 1987, 1988; Panyim *et al.*, 1988; Cooper *et al.*, 1991; Beebe *et al.*, 1994).

2.1.4.2 Classical genetic markers

I) Restriction fragment length polymorphisms (RFLPs)

In DNA strands, restriction fragment length polymorphisms (RFLPs) represent changes in the lengths of DNA between specific enzyme cutting sites brought about by sequence changes. DNA can be cut using one or more restriction endonucleases that recognize sites on the DNA template. The resulting restriction fragments are separated according to size by electrophoresis on agarose gels revealing diagnostic polymorphisms with distinct DNA fragments. However, in most cases, the digestion products appear as a smear with fragments ranging up to approximately 25 kb in size (Loxdale & Lushai, 1998). To gain useful information on these gels, the DNA must first be transferred to a nylon or nitrocellulose membrane by Southern blotting (Sambrook *et al.*, 1989) and selective fragments visualized by hybridization with a suitable labeled DNA probe (Mills, 1984). Favia *et al.* (1997) designed a polymerase chain reaction (PCR-RFLP) assay, which unambiguously separates the *An. gambiae* Mopti form from the Savanna and Bamako forms. This method, based on the presence of a restriction site length polymorphism, has been tested on previously karyotyped sympatric specimens from Mali and Burkina Faso, and it has already been applied to verify the distribution of other molecular markers, as for example the pyrethroid resistance gene (*kdr*) among these chromosomal forms (Chandre *et al.*, 1999).

The main advantage which is applicable to all DNA markers, is that the level of molecular variation detectable is increased. With RFLPs, this is because of numerous restriction enzymes that exist which cut the DNA at different sites and a diverse selection of probes based on hypervariable motifs (Loxdale & Lushai, 1998). By virtue of their ability to detect analogous DNA on both the members of pairs of homologous chromosomes, RFLP based techniques also yield codominant, non-epistatic Mendelian markers which provide great genetic resolution because of the large number of restriction-enzyme-probe combinations available (Zraket *et al.*, 1990). Finding a probe that will be suitable to resolve the population under investigation is the main drawback of this technique, but this is overcome by the availability of suitable probes from parallel studies. Another disadvantage is that such methods utilize probes which need to be screened from genomic libraries if not already available thus leading to increase in costs. These markers are therefore best adopted if probes from parallel studies to those contemplated already exist.

ii) Mitochondrial DNA analysis

Mitochondrial DNA (mtDNA) has been used widely in taxonomic and population studies (Simon *et al.*, 1994) and is a valuable marker to indicate maternal gene flow, as it is predominantly transmitted through maternal lines (Avise, 1991, 1994). Because mtDNA is more numerous than nuclear DNA (the ratio of mitochondria to nuclei being far greater per cell), it can be more easily detected in old samples. Numerous investigations have utilised mtDNA to derive the phylogenies of *An. gambiae* s.l. and to address questions of genetic structuring among members of this species complex (Caccone *et al.*, 1996; Besansky *et al.*, 1997; Lehmann *et al.*, 1997; Thelwell *et al.*, 2000). Mitochondrial DNA has been similarly utilised in investigations of a wide variety of anophelines and anopheline species complexes. Foley *et al.* (1998) used the COII gene to derive the

phylogeny of the Australasian anophelines and De Merida *et al.* (1999) also utilised mtDNA (ND5) and SSCP analysis to examine the population structure of *An. albimanus*. In these examples, mtDNA has proven useful in examinations of phylogenetics and population structure over large geographic scales, but it should be reiterated that this is not true for all anophelines.

Apart from the versatility of its applications, the other advantages of this marker are that sequence analysis can differentiate to the level of race-forms. This is partly because mtDNA mutates approximately 20 times faster than nuclear DNA, so that a 2% divergence in sequences between pairs of lineages separated for less than 10 million years equates to 1 million years since their genetic separation (Brown *et al.*, 1979). However, a confounding factor noted in aphids for example, is the presence of multiple copies of certain mtDNA genes (Sunnucks & Hales, 1996). In such cases, copies of mtDNA genes have become incorporated into the nuclear genome. In one respect, this may prove to be an advantage, as these translocated genes become 'fossilized' in the nuclear genome, and can be a source to evaluate rates of mutation in relation to the equivalent mitochondrial fraction.

Some of the disadvantages are that the level of variability between samples depends on their degree of genetic isolation (an estimate of the lower end of the resolution of genetic isolation would be about 50 years). Mitochondrial DNA has also failed to differentiate cryptic taxa within the *An. maculipennis* complex (Collins *et al.*, 1990).

iii) The polymerase chain reaction applied to nuclear DNA (nDNA)

Ribosomal DNA (rDNA) is a multicopy gene complex found in nuclear DNA (Hillis & Dixon, 1991). The rDNA cistron comprises the 28S, 5.5S and 18S coding regions that are

interspersed with internal transcribed spacer regions (ITS1 and ITS2, either side of the 5.5S gene) [Hillis & Dixon, 1991]. These modules are then repeated in tandem linked by the intergenic spacer region (IGS) [Avise, 1994]. There are several different sites within the rDNA for which universal PCR primers have been constructed (Black, 1991). However, conserved sequences flank regions of hypervariability making this marker versatile at different levels of taxonomy reflecting different rates of evolution (ITS2-28S, Campbell *et al.*, 1993; ITS1–ITS2, Paskewitz *et al.*, 1993; 18S, Pashley *et al.*, 1993).

Many investigators have used these regions to distinguish between members of species populations and complexes of medically important vectors of disease (Paskewitz *et al.*, 1993; Townson & Onapa, 1994; Collins & Paskewitz, 1996). Currently, the most widely used method in the identification of member species of the *An. gambiae* complex is based on species-specific nucleotide sequences in the ribosomal DNA (rDNA) intergenic spacer regions (Scott *et al.*, 1993). This may be used to identify both species and interspecies of hybrids, regardless of life stage, using either extracted DNA or fragments of a specimen. In using this method for the identification of mosquitoes, intact portions of a specimen as small as an egg or the segment of one leg, may be placed directly into the PCR mixture for amplification and analysis (Scott *et al.*, 1993).

This type of diagnostic tool has also been extensively used for differentiation of species within the *An. funestus* group (Hackett *et al.*, 2000), *An. dirus* (Xu *et al.*, 1998), *An. fluviatilis* (Manonmani *et al.*, 2001), and the *An. quadrimaculatus* complex (Cornel *et al.*, 1996). Such studies have readily identified the usefulness of this molecular marker for differentiating at the level of sub-species, races and strains. However, continuous variation of the intergenic spacer (IGS) and multiple copies of the gene within a single individual can confound analysis, PCR amplification and sequencing results. This is well

demonstrated by IGS variability within an individual which has been found to be as great as that across a sample population of the aphid, *Schizaphis graminum* Rondani (Hemiptera: Aphididae) at various spatial scales (Shufran *et al.*, 1991). The costs involved with this marker are similar to any PCR based technique. In general, problems with multiple copies mean that this technique, unlike others, should be used with careful consideration. In some studies to date, contaminants have been amplified along with the study source (Fenton *et al.*, 1994); in this particular case, a novel fungal organism was identified.

Apart from rDNA genes, there are a few other nuclear genes that have been widely applied to phylogenetic investigations. However, Friedlander *et al.* (1992) have reported 14 nuclear genes that may prove useful for higher level phylogenetic analysis over a wide range of taxa, e.g. ENOL, enolase; G6PD, glucose-6-phosphate dehydrogenase and SODPUMP, Na^+/K^+-ATPase. Further sites for short sequences that are proving useful in various taxonomic and population genetic studies have recently been investigated in introns (Lessa & Applebaum, 1993; Palumbi & Baker, 1994; He & Haymer, 1997), non-coding regulatory regions closely associated with genes (Slade *et al.*, 1993).

2.1.4.3 Highly polymorphic markers

i) Microsatellite DNAs

Simple tandem repetitive DNA, more commonly known as microsatellite DNA, has become a popular tool for genetic studies of anophelines. Microsatellites are described as DNA fragments consisting of simple short sequences usually of 2-6 nucleotides (nt), tandemly repeated in more or less uniform tracts up to approximately 10^2 nt long (Tautz, 1993; Chambers & MacAvoy, 2000). Microsatellites are found in the genomes of just about every known organism and organelle. In most eukaryotic organisms,

microsatellites are dispersed throughout the genome (Hamada *et al.*, 1982) and can occur as frequently as every 10 kb (Tautz, 1989). Polymorphism in microsatellites leads to an increased probability of finding heterozygous individuals. Microsatellite loci are also frequently hypervariable, (i.e. they possess several alleles at relatively high frequency) improving their versatility and hence are ideal tools for molecular characterization of individuals and studies of intraspecific variation (Tautz, 1993; Lanzaro *et al.*, 1995).

Exploitation of microsatellite loci has provided biologists with a set of molecular tools with unsurpassed versatility. The key to this versatility lies in the high levels of variability, which are characteristically found at such loci, coupled with the speed and reliability with which this information can be accessed in the laboratory (Chambers & MacAvoy, 2000). Their applications range from estimation of the spatial relationships between chromosome segments to the elucidation of temporal relationships between origins of species and genera. Microsatellite loci have also been described as ideal markers for measuring population level phenomena such as population structure due to their high polymorphism, codominance, abundant presence throughout the genome and relative ease in scoring (Bowcock *et al.*, 1994; Buchanan *et al.*, 1994; Scribner *et al.*, 1994; Lanzaro *et al.*, 1995). Other applications include use in the analysis of laboratory and agricultural organisms and human genome diversity projects to map single gene traits rapidly (Hearne *et al.*, 1992), DNA profiling (Gill *et al.*, 1994) and also in forensic work (Chambers *et al.*, 1997). Microsatellites are almost perfect tools for application in determining the pattern of relationships between individuals because they are capable of highly discriminating biparentally inherited co-dominant markers. In the analysis of *An gambiae* populations, microsatellite DNA analysis has been used to generate a genetic map of the mosquito (Zheng *et al.*, 1993), in the genetic differentiation of populations (Lanzaro *et al.*, 1995) and in the study of population structure (Donnelly *et al.*, 1999).

The advantage of this approach is the ability to detect greater levels of genetic variability as many microsatellite loci, often with numerous alleles, can potentially be screened for ecological use (Evans, 1993). Individual alleles can be scored at particular loci and provide good Mendelian markers. Tri- and tetra-nucleotide repeats provide much better signals, i.e. reduced stutter bands (caused by mispriming inducing templates of smaller size to the original band) and are more easily recorded for size variations (Queller *et al.*, 1993). However, the disadvantages include the requirement to screen several loci for adequate population information (a minimum of four polymorphic loci in clonal organisms and ten and above for sexual populations), thus increasing development time and costs per sample. Although micro-satellite PCR primers are designed to be species specific, cross species allele amplification has been noted (Queller *et al.*, 1993).

ii) Randomly amplified polymorphic DNA (RAPD)

Randomly amplified polymorphic DNA (RAPD) markers have been extensively used to distinguish between members of cryptic species (Williams *et al.*, 1990). The technique is fast, technically easy, and requires little material. Most importantly, no previous nucleotide sequence information is needed for the construction of primers. Many markers can be readily identified for a variety of taxonomic levels and, in comparison with DNA sequencing, the effort and cost are modest so that many individuals can be assayed. This method, which reveals great genetic variability due to the regions in which amplification takes place (Black *et al.*, 1992), is useful in differentiating closely-related individuals, and there are numerous commercially available primer kits which can be used to screen populations. The analysis of RAPDs provides a novel and effective method for distinguishing *Anopheles* species and other organisms according to the banding patterns of their DNA, as well as providing a new means of obtaining genetic markers (Hedrick, 1992).

However, these markers have several drawbacks: (i) they often reveal continuous variation between sample populations; (ii) primer libraries need screening to identify suitable primers and stable genetic polymorphisms, both of which are time consuming and costly; (iii) the dominant nature of the markers makes it impossible to distinguish between homozygous and heterozygous alleles (Carlson *et al.*, 1991); (iv) additional techniques are required for useful Mendelian data to be obtained (Vaughn & Antolin, 1998); and (v) numerous factors such as DNA quality and quantity, Mg^{2+} concentrations, *Taq* source etc. can affect the reproducibility and standardization of reactions (Black, 1993).

2.1.5 Distribution and ecology

Mosquitoes have an almost worldwide distribution, being found throughout the tropics and temperate regions and even well beyond the Arctic Circle; they are absent only from Antarctica and a few islands (Service, 1993). They are found at elevations of 5,500 m and in mines at a depth of 1,250 m below sea level (Service, 1980). Their great diversity of habitats and life-history strategies has allowed them to colonize many contrasting environments. For example, mosquito larvae are found in ponds, swamps, salt-water marshes, pools, tree-holes, polluted water of septic tanks, rice fields, discarded domestic containers, rock pools, plant axils and pitcher plants, and in a variety of other aquatic habitats. Adults are encountered in almost all types of ecological zones, from equatorial rain forests, urbanised conglomerates, cultivated lands to semi-arid areas (Service, 1993). Some genera, however, have a restricted distribution and may be confined to certain areas of the world. The genera *Haemagogus* and *Sabethes*, for example, are found in only Central and South America. Some mosquitoes may occur in only a few countries or localities, for example *An. bwambae*, whereas others such as Cx. *quinquefasciatus* and *Ae. aegypti* are widespread in the tropical regions of the world (Service, 1980).

Most mosquitoes probably disperse only a few hundred metres or so from their emergence sites; for example *Aedes aegypti* usually probably flies only 25-100 m or so. Normally *Anopheles* do not fly more than 2 km, but in certain circumstances they can regularly fly 3-5 km (Service, 1993). The distance mosquitoes fly is determined largely by the environment: if suitable hosts and breeding places are nearby, mosquitoes do not have to disperse far, but if one or both are at a distance, greater dispersal will be necessary.

2.1.6 Medical importance

Mosquitoes cause more human suffering than any other organism, with over one million people dying from mosquito-borne diseases every year. Mosquito-borne diseases continue to cause significant human health problems, largely in the subtropics and tropics, and their incidence has increased significantly within the last 2 decades (World Bank Report, 2001). Not only can mosquitoes carry diseases that afflict humans, but they also transmit several diseases and parasites that dogs and horses are very susceptible to. These include dog heartworm, West Nile virus (WN) and Eastern equine encephalitis (EEE) [White, 1996]. In addition, mosquito bites can cause severe skin irritation through an allergic reaction to the mosquito's saliva causing a red bump and itching. Mosquito vectored diseases include protozoan diseases, i.e., malaria, filarial diseases such as dog heartworm, and arthropod-borne viruses (arboviruses) such as dengue, encephalitis and yellow fever.

Estimates from the World Health Organization indicate that three mosquito-borne diseases are among the leading causes of morbidity and mortality in developing countries around the world (WHO, 2001). Nearly 500 million clinical cases of malaria caused by

infection with *Plasmodium* parasites occur each year, resulting in 2.7 million deaths, mainly in children. Malaria is exclusively transmitted by *Anopheles* mosquitoes.

Lymphatic filariasis is caused by parasitic nematodes and is the second leading cause of permanent and long-term disability worldwide, with 120 million people presenting clinical morbidity (WHO, 2002a). More than half of the world's burden of lymphatic filariasis (LF) is transmitted by *Cx. quinquefasciatus* and other man-biting mosquitoes of the *Cx. pipiens* complex which are responsible for Bancroftian filariasis transmission in the Americas, Egypt, urban East Africa, the Indian subcontinent, Indonesia and southeast Asia (WHO, 2000). In about 40 countries in the African region and Papuan sub-region, *W. bancrofti* is largely transmitted by *Anopheles* mosquitoes that also vector malaria in rural areas. In most Pacific countries, *W. bancrofti* is vectored by aedine mosquitoes (*Aedes* and *Ochlerotatus*) that also transmit arboviruses, notably dengue. Brugian filariasis, transmitted by *Mansonia* and *Anopheles*, is now limited to only 8 oriental countries (White, 1996).

Dengue fever virus, particularly its haemorrhagic form, is a threat to >2.5 billion people, with an annual incidence in the tens of millions and 324,000 deaths per year (WHO, 2002b). Dengue viruses are transmitted to humans through the bites of infective female *Aedes* mosquitoes. Many mosquito species are also vectors of other arboviruses, including several of the most important mosquito-borne human diseases, West Nile virus (WN), eastern equine encephalitis (EEE), western equine encephalitis (WEE), St. Louis encephalitis (SLE), La Crosse (LAC) encephalitis and Japanese encephalitis (JE) [White, 1996]. The arboviruses are the most diverse, numerous and serious diseases transmitted to susceptible vertebrate hosts by mosquitoes and other blood-feeding arthropods.

2.2 The Genus *Anopheles*

Notable distinguishing characteristics of anopheline mosquitoes (from other mosquitoes) include the long palps that are present on both males and females and the characteristic pattern of blocks of dark and pale scales on the wing veins, especially along the costa (top part of the wing) [Fig. 2.2].

Chromosomal data suggest that *Anopheles* is "primitive" within the family Culicidae (Besansky & Collins, 1992). Evidence, including significantly smaller chromosomes and lower nuclear DNA content (Rao & Rai, 1990), and their unique possession of dimorphic sex chromosomes and long-period interspersion of repetitive sequences in the genome (Black & Rai, 1988), support the extensive divergence of *Anopheles* from the other mosquitoes.

The genus *Anopheles* Meigen consists currently of 437 recognised species which divide into six sub-genera; *Anopheles* (185 species), *Cellia* (200 species), *Lophopodomyia* (6 species), *Kertezia* (12 species), *Nyssorhynchus* (29 species), and *Stethomyia* (5 species) [Sallum *et al.*, 2000].

2.2.1 The life cycle of *Anopheles* mosquitoes

The *Anopheles* mosquito goes through four separate and distinct stages of its life cycle; egg, larva, pupa, and adult (Fig. 2.3). Each of these stages can be easily recognized by its special appearance. The immature stages of *Anopheles* are aquatic, whilst the adult is terrestrial. Anopheline larvae exist in a wide variety of different habitats (Service, 1993).

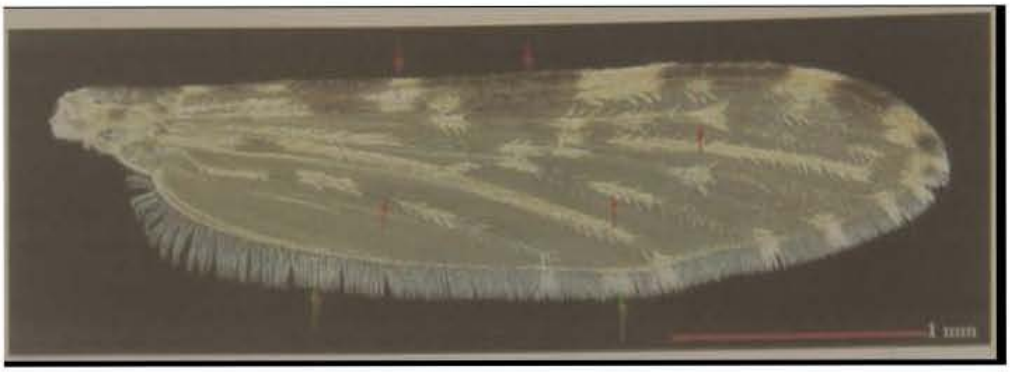


Fig. 2.2 An anopheline wing which is identified by its characteristic pattern. There are blocks of pale and dark scales (red arrows) on the vein, especially along the costa (the top part of the wing). There is also a fringe of narrow, outstanding scales (green arrows) along the bottom of the wing

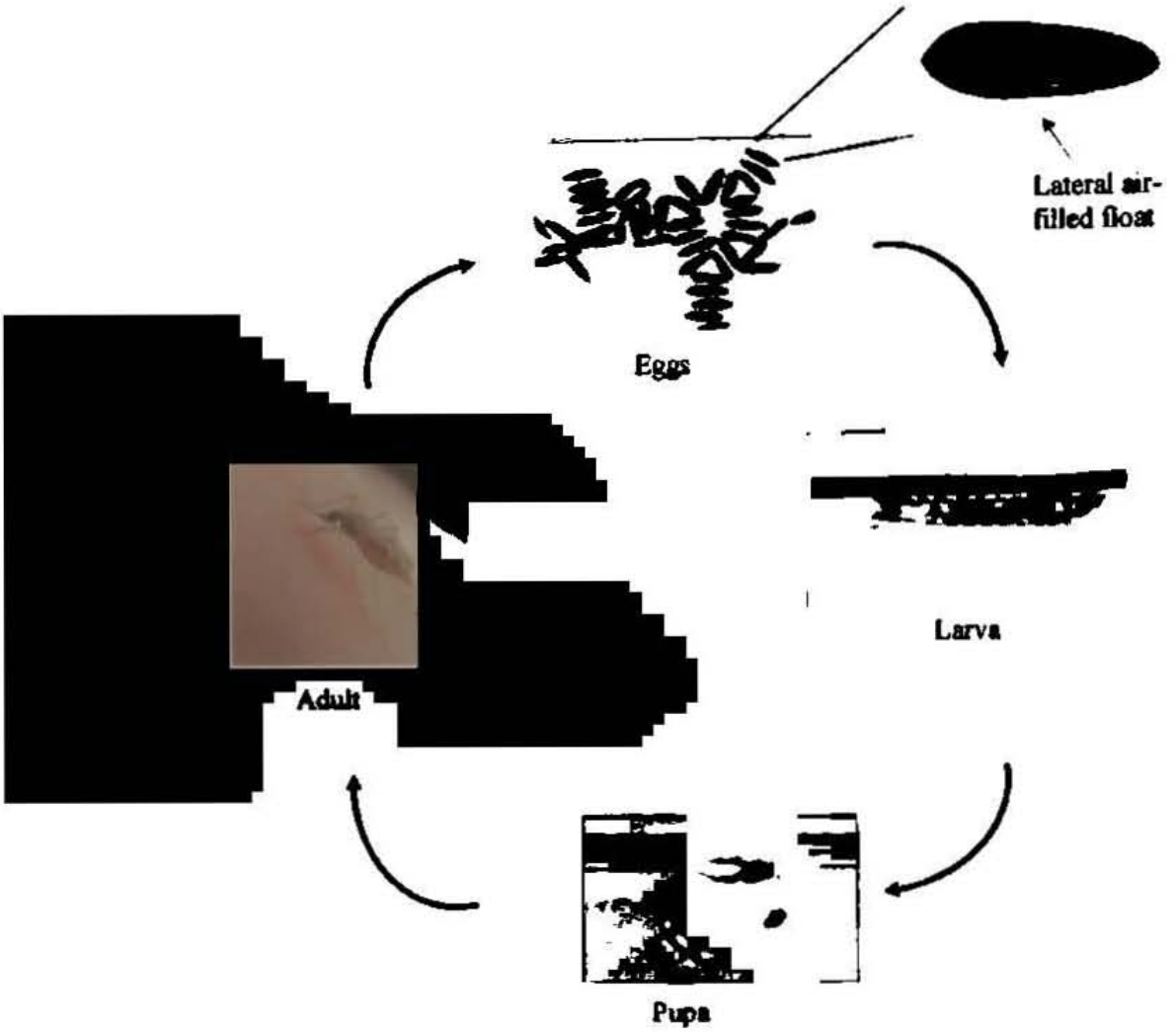


Fig. 2.3 Life cycle of *Anopheles* mosquitoes There are four distinct stages of development (complete metamorphosis): egg, larva, pupa, and adult. The eggs are laid in water and have paired lateral air-filled floats, are boat shaped and may hatch in 2-3 days The larva lies parallel to the water surface and has conspicuous mouth brushes The pupa is comma-shaped, has a fused head and thorax. The pupal stage is a relatively short (two to three days), nonfeeding, transitional stage in which the adult develops within the pupal covering. Both adult females and males drink nectar and other plant juices as energy sources; but only females take a blood meal, utilizing the protein from blood to produce eggs

Some species, such as *An. pseudopunctipennis* Theobald, prefer pools of water in drying streams in relatively dry, sunny areas, while others prefer inundated areas such as rice fields. *Anopheles minimus* in Thailand and the Philippines lives amongst bamboo roots in relatively swift-moving water. *Anopheles gambiae* sensu lato often lives in puddles of water that accumulate in tyre ruts, and *An. stephensi* may live in man-made containers. *Anopheles* larvae as a whole tend to prefer clean, rather than polluted water (Service, 1993).

The *Anopheles* egg is difficult to see with the naked eye, being quite small about 0.5 x 0.1 mm in length (Service, 1993). It is boat-shaped and has two lateral floats which are filled with air. The shape of these floats, as well as the chorion ornamentation may be used to distinguish the subgenera and sometimes the species of anophelines (Hervy *et al.*, 1998). The eggs are laid singly on the surface or on the edge of water and are attached side-to-side or end-to-end, but never end-to-side. The eggs are white when newly laid but turn brown or black within 1-2 hours. In tropical temperatures, eggs hatch in 2-3 days to produce larvae. These eggs cannot withstand desiccation (Service, 1993). Some 100-200 per batch are laid every 2-4 days. Survival and egg development are mainly dependent on temperature and relative humidity; under extreme climatic conditions, mosquitoes may go into hibernation or aestivation, which allows the survival of the species through the winter in temperate climates, or long dry seasons in tropical arid areas (Clements, 1992).

The larvae which hatch from the floating egg have a head, a thorax and nine abdominal segments. The larvae also show numerous morphological peculiarities, the most striking of these being the lack of a siphon on the segment VIII, the presence of palmate setae on the first seven abdominal segments, and the shape of the cephalic setae (Hervy *et al*, 1998). Anopheline larvae are filter feeders of particulates at the surface of water. The

buccal apparatus is surrounded by brushes which create a water flow that brings food particles to the mouth (Service, 1993). A conspicuous distinctive character of *Anopheles* is the way the larval body positions itself parallel to the water surface as against the obliquous style exhibited by *Aedes* and *Culex*. This position might be related to the absence, in the *Anopheles* larva, of the respiratory siphon which in Culicine larvae breaks the surface of water and takes in air into the tracheae.

The duration of the larval stage ranges between 5 and 11 days during which the larvae undergo 3 successive moults, during which they maintain the same morphology, but increase in size from less than 1 mm for the 1st stage larvae to more than 0.5 cm for the 4th instar (Service, 1993). Rates of larval growth are influenced by environmental factors such as water temperature, photoperiodicity, food supply, degree of overcrowding and the species (Service, 1993; White, 1996). The fourth instar larva moults into a pupa, in which the head and thorax are fused to form a common envelope which bears the respiratory trumpets. The pupa is a relatively short stage, usually 2-4 days, during which it does not feed. The actual eclosion takes approximately 12-15 minutes (Service, 1993) and for most species emergence occurs in the early evening or late at night.

After emergence, adult females seek shelter amongst vegetation until they are ready for mating (between a few hours to 2 days) while the mouthparts harden to permit skin penetration and feeding. Males are unable to mate until their genitalia have rotated through 180°, a process that takes 6-24 hours depending on the species (Service, 1993; White, 1996). Mating usually occurs in flight and commonly around dusk. Female monogamy in *An. gambiae* is briefly enforced by the male's deposition of a mating plug into the female's genital chamber after sperm are transferred. The mating plug, formed from proteinaceous secretions of the male's accessory glands, effectively blocks sperm

from subsequent copulations from reaching the spermatheca (Gillies & Chir 1956; Giglioli & Mason 1966; Kettle, 1995).

Female anopheline mosquitoes obtain energy from sugar meals for metabolism and for flight. Flight is needed for mating and for finding a host that will provide a blood meal source (appetitive flight). The blood meal is a protein-rich diet that the mosquito surrounds, after ingestion, with the peritrophic matrix (PM), a thin structure containing chitin and proteins. Digestion requires secreted proteases that penetrate the PM (Clements, 1992). The smaller digestion products are hydrolyzed by microvilli-bound enzymes before absorption by the midgut cells. The blood meal-derived nutrients are processed by the insect fat body (equivalent to the liver and adipose tissue in vertebrates) into egg proteins (vitellogenins) and various lipids associated with lipoproteins. These are then exported through the haemolymph to the insect ovaries. The egg development process takes 2 to 3 days, and no further food intake is needed until after oviposition, when a new cycle of active host finding and blood feeding, digestion, and egg development begins (Clements, 1992).

The adult female may live from a few days to well over a month, going through several cycles of blood feeds and egg-laying. The duration of the whole process, from oviposition to adult eclosion, is temperature-dependent. At the elevated temperatures of the intertropical areas, it takes a minimum of 7 days (Clements, 1992).

2.2.2 *Anopheles* species of medical importance

Only a small number of mosquito species (Table 2.1) are genetically competent to transmit pathogens to a vertebrate host and competence varies within populations of a given species (Gubler *et al.*, 1982; Curtis & Graves, 1983; Christensen & Severson, 1993).

Anopheline mosquitoes are the exclusive vectors of human malaria. A handful of species predominate as the most notorious malaria vectors, but the species and forms involved in the transmission of human malaria worldwide are very diverse (Coetzee *et al.*, 2000; Van Bortel *et al.*, 2001). Malaria vectors are often divided into primary and secondary vectors, but this is rather unsatisfactory because a species can be a so-called primary vector in one area and a secondary vector or even a non-vector in another. White (1982), preferred to classify *Anopheles* as main vectors (widespread, dominant vectors) and subsidiary vectors, i.e. incidental vectors which, by themselves, are incapable of sustaining transmission, or secondary vectors localised in their distribution but able to be principal vectors within their areas. The most important vectors are listed in Table 2.2 according to the epidemiological zones (Fig. 2.4) of Macdonald (1957).

The human-biting index (HBI), which is the proportion of blood meals taken from a human host, has a large impact on a mosquito species's vectorial capacity for malaria. This is illustrated by the fact that the world's major malaria vectors all feed predominantly on humans (Garrett-Jones, 1964).

Table 2.1 *Anopheles* species of medical importance (Hervy *et al.*, 1998).

Arboviruses

Vector	Arbovirus (Endemic Area)
An. brohieri	Shokwe (Ivory Coast)
An. brunnipes	West-Nile (Madagascar)
An. christyi	Rift Valley fever (Kenya)
An. coustani	Rift Valley fever (Zimbabwe, Madagascar)
An. domicola	Wesselsbron (Senegal)
An. flavicosta	Middelburg (Senegal)
An. freetownensis	Wesselsbron (Senegal)
An. funestus	O'Nyong-Nyong (Uganda, Tanzania, Kenya) Wesselsbron (Central African Republic)
An. gambiae	O'Nyong-Nyong (Uganda, Tanzania, Kenya) Wesselsbron (Cameroun)
An. maculipalpis	West-Nile (Madagascar)
An. mascarensis	Mengo (Madagascar)
An. nili	Bagaza, Tataguine (Senegal; Bangui : Burkina Faso)
An. paludis	Gomoka (Central African Republic)
An. pauliani	Rift Valley fever (Madagascar) Andasibe (Madagascar)
An. pharoensis	Rift Valley fever (Kenya); Wesselsbron (Cameroun) Sindbis (Egypt)
An. pretoriensis	Ngari (Senegal)
An. rufipes rufipes	Wesselsbron, Chikungunya and Gomoka (Senegal)

Filariasis

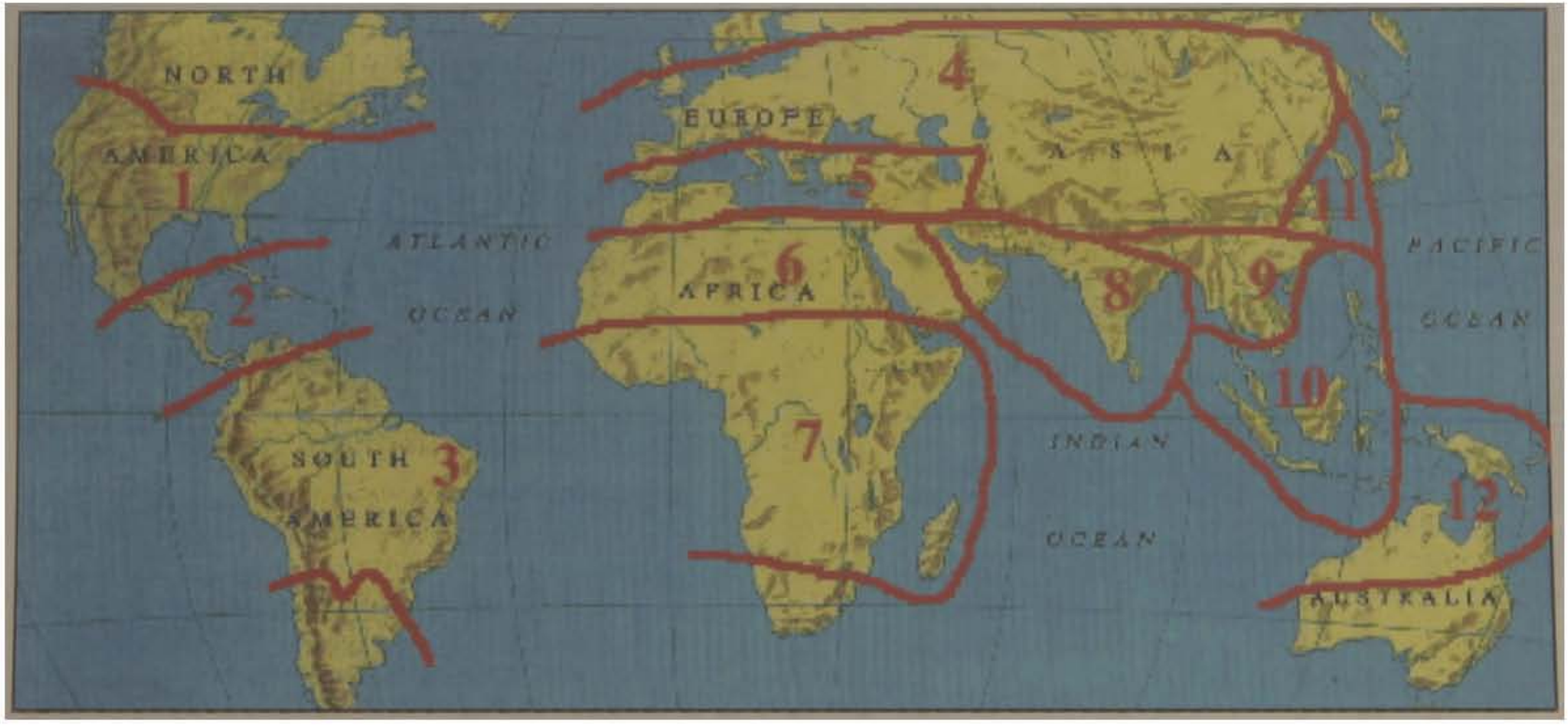
Vector	Places Where Incriminated
An. arabiensis	Main vector of *W. bancrofti* in the Afrotropical region
An. funestus	Main vector of *W. bancrofti* in the Afrotropical and Oriental regions
An. gambiae	Main vector of *W. bancrofti* in the Afrotropical region
An. anthropohagus	Vector of *Brugia malayi* and *W. bancrofti* in the Oriental region
An. pauliani	Secondary vector of *W. bancrofti* in Madagascar
An. albimanus	Vector of *W. bancrofti* in Central and South America.

Table 2.2 Malaria vectors of the world. Zone numbers and geographical naming follow Macdonald (1957). Abbreviations for subgenera given in parentheses are: A, *Anopheles* s.str.; C, *Cellia*; K, *Kerteszia*; N, *Nyssorhynchus*. Main (primary) vectors are distinguished from subsidiary (incidental and local) vectors by bold type. The data have been compiled from Service (1993).

Zones	Vectors	
1. North American	(A.) ***freeborni***	A. (N.) *albimarus*
	A. (A.) *quadrimaculatus*	
2. Central American	A. (A.) ***aztecus***	A. (N.) *albitarsis*
	A. (A.) ***punctimacula***	**A. (N.) *aquasalis***
	A. (A.) *pseudopunctipennis*	**A. (N.) *argyritarsis***
	A. (N.) *albimanus*	**A. (N.) *darlingi***
3 South American	(A.) ***pseudopunctipennis***	**A. (N.) *albitarsis***
	A. (A.) *punctimacula*	**A. (N.) *aquasalis***
	A. (K.) *bellator*	A. (N.) *argyritarsis*
	A. (K.) *cruzii*	A. (N.) *braziliensis*
	A. (K.) *neivai*	**A. (N.) *darlingi***
	A. (N.) *albimanus*	A. (N.) *nuneztovari*
		A. (N.) *triannulatu*
4. North Eurasian	**A. (A.) *atroparvus***	A. (A.) *sacharovi*
	A. (A.) *messeae*	A. (A.) *sinensis*
		A. (C.) *pattoni*
5. Mediterranean	**A. (A.) *atroparvus***	**A. (A.) *sacharovi***
	A. (A.) *claviger*	A. (C.) *hispaniola*
	A. (A.) *labranchiae*	A. (C.) *superpictus*
	A. (A.) *messeae*	
6. Afro-Arabian	A. (C.) *culicifacies*	A. (C.) *multicolor*
	A. (C.) *fluviatilis*	**A. (C.) *pharoensis***
	A. (C.) *hispaniola*	**A. (C.) *sergentii***
7. Afrotropical	**A. (C.) *arabiensis***	A. (C.) *merus*
	A. (C.) *funestus*	A. (C.) *moucheti*
	A. (C.) *gambiae*	A. (C.) *nili*
	A. (C.) *melas*	A. (C.) *pharoensis*

Table 2.2 cont'd

Zones	Vectors	
8. Indo-Iranian	A. (A.) *sacharovi*	A. (C.) *philippinensis*
	A. (C.) *aconitus*	A. (C.) *pulcherrimus*
	A. (C.) *annularis*	A. (C.) *stephensi*
	A. (C.) *culicifacies*	A. (C.) *sundaicus*
	A. (C.) *fluviatilis*	A. (C.) *superpictus*
	A. (C.) *jeyporiensis*	A. (C.) *tessellatus*
	A. (C.) *minimus*	A. (C.) *varuna*
9. Indo-Chinese hills	A. (A.) *nigerrimus*	**A. (C.) *fluviatilis***
	A. (C.) *annularis*	A. (C.) *jeyporiensis*
	A. (C.) *culicifacies*	A. (C.) *maculatus*
	A. (C.) *dirus*	A. (C.) *minimus*
10. Malaysian	A. (A.) *campestris*	A. (C.) *flavirostris*
	A. (A.) *conaldi*	A. (C.) *jeyporiensis*
	A. (A.) *donaldi*	**A. (C.) *leucosphyrus***
	A. (A.) *letifer*	**A. (C.) *ludlowae***
	A. (A.) *nigerrimus*	**A. (C.) *maculatus***
	A. (A.) *whartoni*	A. (C.) *mangyanu*
	A. (C.) *acconitus*	**A. (C.) *minimus***
	A. (C.) *balabacensis*	A. (C.) *philippiensis*
	A. (C.) *dirus*	A. (C.) *subpictus*
		A. (C.) *sundaicus*
11. Chinese	**A. (A.) *anthropophagus***	A. (C.) *balabacensis*
	A. (A.) *sinensis*	A. (C.) *jeyporiensis*
		A. (C.) *pattoni*
12. Australasian	A. (A.) *bancroftii*	A. (C.) *karwari*
	A. (C.) *farauti* type 1	**A. (C.) *koliensis***
	A. (C.) *farauti* type 2	**A. (C.) *punctulatus***
	A. (C.) *hilli*	A. (C.) *subpictus*



1 North American	4 North Eurasian	7 Afrotropical	10 Malaysian
2. Central American	5 Mediterranean	8 Indo-Iranian	11 Chinese
3. South American	6. Afro-Arabian	9 Indo-Chinese hills	12 Australasian

Fig. 2.4 Map of the malaria epidemiological zones (after MacDonald, 1957).

In most parts of Africa, several vectors transmit malaria in each location, in some cases at the same time and in other cases during different seasons, as for example in Dielmo, Senegal (Fontenille *et al.*, 1997). Much variation can be observed between villages a few kilometres apart. In Central Africa, it is not rare to capture four different vectors (e.g. *An. gambiae*, *An., funestus*, *An. nili* and *An. moucheti*) during the same night (Fontenille & Lochouarn, 1999).

Anopheles gambiae sensu stricto and *An. arabiensis*, the two most anthropophilic species, are responsible for more than 3/4 of the world *Plasmodium falciparum* inoculations (Favia & Louis, 1999) and are thus the most efficient vectors in sub-Saharan Africa. However, they differ in the degree of anthropophilly and endophily, with *An. gambiae* s.s. showing an increased propensity towards endophily, endophagy, and anthropophagy compared to *An. arabiensis*. *Anopheles gambiae* s.s, the nominal species, is thus considered to be the most important malaria vector south of the Sahara (Mekuria, 1983).

The importance or competence of *An. gambiae* as a malaria vector is well illustrated by its accidental introduction into Brazil, where in 1938, after series of small but intense local outbreaks, it caused the worst epidemic of malaria recorded there, with over 14,000 deaths in less than 8 months (Soper & Wilson, 1949).

2.3 The *Anopheles Gambiae* Giles Complex

The *Anopheles gambiae* complex is the world's most important malaria vector complex. The complex is a group of morphologically indistinguishable (isomorphic) yet genetically distinct species that vary in their behaviour and vectorial capacity and hence importance as malaria vectors in Africa. Collectively all the members of the complex are referred to as *Anopheles gambiae s.l* (*sensu lato*). The complex comprises six named species (Gillies & Coetzee, 1987), one unnamed species (Hunt *et al.*, 1998) and several incipient species (Coluzzi *et al.*, 1979; Favia *et al.*, 1997). The six named species are: *Anopheles gambiae* Giles, *An. arabiensis* Patton, *An. quadriannulatus* Theobald, *An. merus* Donitz, *An. melas* Theobald and *An. bwambae* White. The latest addition to the complex is provisionally designated *An. quadriannulatus* species B because of its close similarity to this species (Hunt *et al.*, 1998).

The recognition of distinct, cryptic species within the *An. gambiae* complex stems from studies of the inheritance of insecticide resistance. Crosses among some strains of '*gambiae*' produced sterile male offspring (Davidson, 1962; 1964; Davidson & Jackson, 1962; Davidson *et al.*, 1967; Davidson & White, 1972; Davidson & Hunt, 1973). All 30 possible reciprocal crosses of the presently recognized species have been made in the laboratory resulting in fertile F_1 females, but sterile F_1 males, with two exceptions: *An. quadriannulatus* females x *An. gambiae* males and *An. quadriannulatus* females x *An. bwambae* males yield some fertility in F_1 males as well as females. The genetic distinctness of the six formally named species was confirmed by cytogenetic examinations that revealed fixed inversion differences among taxa which are very stable also in areas of sympatry (Coluzzi, 1966; Coluzzi & Sabatini, 1967, 1968, 1969). While not as definitive as the inversion data, allozyme studies also confirmed the genetic distinctness of the six species (Bullini

& Coluzzi, 1973, 1978; Mahon *et al.*, 1976; Miles, 1978). Most recently, diagnostic DNA probes have been developed (Collins *et al.*, 1987, 1988; Scott *et al.*, 1993). Thus these are 'good species' by most definitions, although the possibility of occasional interspecific gene flow (introgression) exists through fertile hybrid females. Hybrids have been observed in nature between *An. arabiensis* and *An. gambiae* in Tanzania, Kenya, Mali, and Nigeria (White, 1970; Coluzzi *et al.*, 1979; Petrarca *et al.*, 1991; Toure *et al.*, 1998), between *An. arabiensis* and *An. quadriannulatus* in Zimbabwe (White, 1974), and between *An. gambiae* and *An. melas* in The Gambia (Bryan, 1979). However, such hybrids are rare and comprise less than 0.1% of the samples analyzed (Coluzzi *et al.*, 1979; Petrarca *et al.*, 1991; Toure *et al.*, 1998). Nevertheless, there is evidence that this degree of hybridization is sufficient to produce introgression of genetic material.

Species within the *Anopheles gambiae* Giles complex differ highly in their vectorial capacity for malaria. Because the sibling species appear equally susceptible to *P. falciparum* Welch infection (Takken *et al.*, 1999), differences in the human blood index (HBI) of the sibling species mostly determine their status as malaria vectors.

2.3.1 Distribution and importance of members of the *Anopheles gambiae* sibling species complex

2.3.1.1 *Anopheles bwambae*

This species is known to occur only in the vicinity of the Buranga hot springs in Bwamba County, Uganda, where it breeds in brackish water from geothermal springs (which present the necessary geothermal conditions for the development of its larvae) together with other halophilic mosquitoes (White, 1985). They were first investigated by Haddow (1945) who drew attention to their biological contrasts with typical *An. gambiae* s.l. but was unaware of the specific taxonomic distinction (White, 1985). *Anopheles*

bwambae species differ from the other species of the group by the presence of a wide pale apical band on the female palps (White, 1985).

Larvae of *An. bwambae* are associated with 'springwater' habitats having much higher conductivity, much greater concentrations of dissolved solids and slightly higher temperature and pH than 'normal' fresh water sites inhabited by larvae of *An. gambiae* s.s. (Harbach *et al.*, 1997). Larval habitats are generally fully sun-exposed, sometimes shaded by low vegetation. Eggs are particularly abundant in ponds directly fed by these springs with muddy water often mixed and polluted by wild animals. Eggs are regularly laid all year round.

Adults of *An. bwambae* occur strictly in a circle of 10 km around the hot springs (White, 1985). During daytime, adult males and females rest mainly in the Semliki Forest on the buttressed bases of large trees, on fallen logs and sticks, and under loose dry leaves (White, 1985). Females also rest among the extensive plantations of bananas fringing the forest, and indoors at villages situated near the forest (White, 1973). The adult populations are abundant all year round.

This opportunist species is always very aggressive. Adult females attack people in roadside villages and forest at the eastern edge of the Semliki Valley (Haddow *et al.*, 1947, 1951; Lumsden, 1951, 1952; White, 1973; Gillett, 1985). In villages, females enter inside houses and feed on humans. However, in the same locality, they can also be attracted by domestic pigs. Their activity is essentially nocturnal, even though in the forest, females bite humans during the day (Lumsden, 1952).

In natural conditions some *An. bwambae* adults have been found with both sporozoites and filariae (White, 1985). However, this species is not suspected to play an important epidemiological role in the transmission of pathogenic agents of humans or domestic animals because of its very restricted distribution.

2.3.1.2 *Anopheles quadriannulatus*

Anopheles quadriannulatus is found restricted to eastern and southern Africa. It is, unlike to the other species, relatively tolerant of the highland cold conditions and has been found in the highlands of Ethiopia and extensively in southern Africa (White, 1974). The Ethiopian population was recently recognised as a distinct species and has been designated as *An. quadriannulatus* species B (Hunt *et al.*, 1998).

The egg-laying sites are similar to those of the other freshwater species of the *gambiae* complex (White, 1974). The sites are always sunny and consist most often of residual puddles or temporary or semi-permanent ponds.

Also, unlike the other member species, *An. quadriannulatus* is aggressive during the first half of the night. Females of the species are considered to be strictly zoophilic (White, 1974; Gibson, 1996; Dekker & Takken, 1998). Although house-resting adult populations of the species have been observed in South Africa and Zimbabwe (Hunt & Mahon, 1986), they were found more frequently in buildings with cattle only, or with human and cattle together (White, 1974). However, anthropophagic behaviour not previously recorded in the species A has been reported (Pates *et al.*, 2001).

Anopheles quadriannulatus has never been implicated in malaria transmission, although recent laboratory experiments have shown that it is susceptible to *P. falciparum* (Takken

et al., 1999). It is also susceptible to the human filarial worm *W. bancrofti* (Hunt & Gunders, 1990).

2.3.1.3 *Anopheles melas*

The distribution as well as some taxonomic characteristics of *An. melas* facilitate its distinction from the other members of the complex. Its distribution area is limited to the coasts of Western Africa (Coluzzi, 1984).

Although *An. melas* larvae can withstand very high salinity they usually develop in waters with salinity which is the same or lower than seawater (Akogbeto, 1995). They are only rarely found in freshwater ponds, and always in small numbers. The larvae of this littoral species essentially colonize estuaries, lagoons and mangrove swamps (Coluzzi, 1984; Diop *et al.*, 2002). However, in Gambia where the hydrographic net permits an upstream tide, this species can occur up to 150 km inland (Snow, 1983).

Adult *An. melas* populations are particularly unstable and their densities may vary throughout the year (Fonseca *et al.*, 1996; Diop *et al.*, 2002). During the rainy season, its density is influenced by both rains and tides. It shows few feeding preferences: can be highly anthropophilic in the presence of man and very zoophilic in his absence (Snow, 1983; Akogbeto, 2000; Diop *et al.*, 2002; Awolola *et al.*, 2002). *Anopheles melas* life expectancy is short and it does not seem to live more than 15 days in natural conditions.

Its low anthropophily and its reduced lifespan make it a bad malaria vector. In endemic malaria zones, where *An. gambiae* s.s. is the main vector, the observed sporozoite indices are at least 10 times smaller in *An. melas* (Bryan, 1983). Even when *An. melas*

is the only species present, the plasmodial indices observed in man are always very low (Akogbeto & Romano, 1999). These data confirm that this species plays a negligible role in malaria transmission, except perhaps when it is very abundant (Diop *et al.*, 2002).

2.3.1.4 *Anopheles merus*

Anopheles merus is one of the species found in East and South Africa as well as on some African islands in the Indian Ocean (Muirhead-Thomson, 1951; Coluzzi, 1984). However, the distribution, as well as some taxonomic characteristics, facilitate its distinction from the members of the complex.

Anopheles merus larvae develop in brackish waters but unlike *An. melas*, they are unable to withstand high salinity (Mosha & Mutero, 1982). This species is particularly abundant in coastal regions where the larvae are often found in brackish water expanses, such as lagoons ponds and swamps and seawater brought by tide diluted with rain (Mosha & Subra, 1992). It can penetrate and survive at great distances where salty water ponds are present.

Anopheles merus is opportunist when in search of a blood meal, although it has a clear zoophilic preference. Thus, in domestic environment, when cattle are present, horned cattle are much more attractive than man (Mutero *et al.*, 1984). On the other hand, females can be very aggressive to man both indoors and outdoors when cattle are absent. Precipitin tests on engorged females confirm these data in that, those collected inside houses had fed on man and those found outside had fed nearly exclusively on bovids (Mutero *et al.*, 1984). When herds stay overnight near breeding pools, it may be totally absent from the villages though it is both exophilic and endophilic (Mutero *et al.*, 1984; Bushrod, 1979). Females resting outside are usually found along roots and trunks of

mangrove trees as well as in natural holes (termitaries, crab-holes, etc) [Mutero *et al.*, 1984]. The survival rate of this *Anopheles* is low.

Anopheles merus transmits malaria and bancroftian filariasis perfectly in the laboratory (Hunt & Gunders, 1990). However, it has been intimated that it is a bad vector for bancroftian filariasis in natural conditions because of its short lifespan and feeding habits (Southgate & Bryan, 1992). The sporozoic indices of *An. merus* are always smaller than those of *An. gambiae* s.s., although *An. merus* may still play a role as secondary vector in localities where it is particularly abundant.

2.3.1.5 Anopheles arabiensis

Anopheles arabiensis is the member species of the complex with the widest geographical distribution. It is found in the Afro-tropical region in the Sahel, on the plateaux of southern Africa and in Madagascar. It is also present in the countries bordering the Red Sea and Aden Gulf, and most of the islands in the Indian Ocean (Abdullah & Merdan, 1995). Moreover, it penetrates deeply in humid savannas, but not into the forest. Where it occurs in the equatorial rain forest regions, such as in Benin City in Nigeria, it is usually associated with a history of extensive land clearance (Coluzzi *et al.*, 1979), although this is not always the case. *Anopheles arabiensis* shows evidence of being better adapted to less humid environments, and is often the predominant or only member of the complex in arid regions of West Africa, Ethiopia and Sudan (Gillies & Coetzee, 1987). It is also more commonly found resting outside houses and in places such as granaries, than *An. gambiae* s.s.; this also reflects a better adaptation to lower humidity (Donnelly & Townson, 2000).

This species if found mainly in the drier savanna areas with rainfall of less than 1000 mm; a fact which is in keeping with its known habitat preferences (Coluzzi *et al.*, 1979;

Gillies & Coetzee, 1987). Anomalies of *An. arabiensis* occurring in desert areas can be explained by their association with river systems, such as the Niger in Mali, and the Nile in Sudan. In most cases, the ecological context or the distribution, as well as some taxonomic characteristics permit its distinction from the other members of the complex. However, these criteria by themselves make it still impossible to distinguish it from *An. gambiae* s.s which it occurs in sympatry with over most of its distribution area.

The egg-laying sites of *An. arabiensis* are similar to those of the other freshwater species of the complex. These are temporary water collections, usually small, shallow and exposed to the sun, without vegetation (Gillies & DeMeillon, 1968; Gillies & Coetzee 1987). These conditions are found in waters accumulated during the rainy season, in depressions or in cavities such as ditches, ruts, foot or hoof-prints (Beier *et al.*, 1990). *Anopheles arabiensis* appears to exploit permanent, artificial habitats such as rice fields (White *et al.*, 1972, Githeko *et al.*, 1996) and garden wells (Robert *et al.*, 1998). However, consistent differences in habitat use by either *An. gambiae* s.s. or *An. arabiensis* have not been observed, and both species often have been found occupying the same habitat (Service, 1970; White & Rosen 1973, Service *et al.*, 1978, Charlwood & Edoh 1996, Minakawa *et al.*, 1999). Population density of this species varies seasonally in relation to rainfall. It increases quickly with the first rains and the maximum density is reached at the end of the rainy season (White *et al.*, 1972; White & Rosen, 1973) and then decreases as the temporary breeding sites dry up.

The host preference of *An. arabiensis* is more difficult to generalise. Depending on the availability of humans relative to alternative hosts, the HBI of this species is generally high; but in villages where cattle are abundant, this proportion can drop in

favour of bovid feeds (Coluzzi *et al.*, 1979; Gillies & Coetzee. 1987; Githeko *et al.*, 1994). For example, *An. arabiensis* is caught on man and donkeys in Niger while its trophic preferences are the same as those of *An. gambiae* s.s. in some localities in south Senegal (Charlwood *et al.*, 1995). Although one explanation would be that *An. arabiensis* is simply a more opportunistic feeder, its feeding pattern is not always proportional to host abundance. East African populations are generally less anthropophilic than West African ones, while in Madagascar mainly zoophilic populations are known to occur independently of the human-bovid ratio (Ralisoa Randrianasolo & Coluzzi, 1987).

Anopheles arabiensis is one of the main malaria and bancroftian filariasis vectors in the Afro-tropical Region, together with *An. gambiae* s.s. and *An. funestus*. It transmits these diseases in most parts of the continent, on Madagascar, on Reunion and Mauritius islands, as well as on the Arabian Peninsula (White, 1996). This species is often more zoophilic, has a lower lifespan than *An. gambiae* s.s. and thus, it is generally a less efficient vector. However it resists drying better and may be the principal vector of malaria and bancroftian filariasis in dry areas.

2.3.1.6 *Anopheles gambiae sensu stricto*

Anopheles gambiae sensu stricto, the nominal taxon, is the most anthropophilic member of the complex and the main malaria vector in sub-Saharan Africa. Genetic and behavioural variations within this species have been well demonstrated by Coluzzi *et al.* (1979, 1985) who described the different cytological forms (named with the non-Linnean nomenclature Forest, Savanna, Bamako, Mopti, Bissau) which show restricted or no inter-breeding in the field (Bryan *et al.*, 1982; Coluzzi *et al.*, 1985; Touré *et al.*, 1983, 1994, 1998) and

whose distribution depends on environmental factors of climate, breeding sites, etc. (Touré *et al.*, 1994). In sympatric areas, hybrids between Savanna and the other forms have been observed at frequencies lower than expected, but no individuals carrying heterozygous complements of the Mopti and Bamako inversions are seen in nature, even though the two forms produce viable offspring under laboratory conditions (Coluzzi *et al.*, 1985; Persiana *et al.*, 1986). Genotyping X-linked rDNA of *An. gambiae* s.s. has led to the characterization of two molecular forms (M and S) that differ in both the transcribed and nontranscribed spacers in the rDNA repeat unit (Favia *et al.*, 2001; della Torre *et al.*, 2001; Gentile *et al.*, 2002). The relationship between the M and S molecular forms and the chromosomal forms - defined according to nonrandom associations of inversions in chromosome 2 (Coluzzi *et al.*, 2002) - varies according to their ecological and geographic distribution. In some areas of West Africa, such as Mali and Burkina Faso, there is a one-to-one correspondence between the M molecular form and the Mopti chromosomal form. Similarly, the S molecular form always corresponds to the Savanna or Bamako chromosomal form (Favia *et al.*, 1997). In other areas of West Africa, this clear correspondence breaks down (della Torre *et al.*, 2001). Although interbreeding between M and S forms yields fertile progeny, M-S hybrids are rarely observed in nature. Where these forms overlap in time and space, the rate of heterogamous insemination is 31% (Tripet *et al.*, 2001), clearly demonstrating the existence of a pre-mating barrier, albeit an incomplete one.

Anopheles gambiae s.s is extremely versatile regarding tolerance to a wide variety of micro- and macro-environmental conditions, as evidenced by its broad geographic distribution. It is widespread in Africa, though better adapted to wetter regions than to savanna areas (Lindsay *et al.*, 1998). The Forest and Bissau chromosomal forms prevail in humid and coastal areas of West Africa and could both belong to a single potentially

panmictic unit widely distributed also in East Africa. The other three taxa, the Bamako, Mopti and Savanna chromosomal forms, are more adapted to dry environments, are often sympatric with evidence of assortative mating, and overlap the distribution of *An. arabiensis* (Favia *et al.*, 1997). The Bamako chromosomal form appears to have a riverine ecology found in the upper Niger river basin, from southern Mali to northern Guinea; it generally coexists with Mopti and, less frequently, with Savanna. The Mopti chromosomal form is the dominant taxon in the inner delta of the Niger river and in all irrigated areas in Southern Mali and eastwards in Burkina Faso up to the middle Niger river basin (Touré *et al.*, 1994). The Savanna chromosomal form has a wider distribution overlapping both the Bamako and Mopti ranges. It is generally absent from the sahelian zones and the inner part of large irrigation schemes where Mopti becomes dominant. The Savanna form has also various contact zones with Bissau-Forest populations and studies carried out in some of these areas showed incomplete intergradation (Bryan *et al.*, 1982; Petrarca *et al.*, 1983).

The larval stages of the Mopti form of *An. gambiae* s.s. occur mainly in man-made habitats, such as irrigated areas (even in the dry season) while Savanna and Bamako forms have a tendency to breed in more natural sites, exploiting rain-dependent pools for larval development. These affect their spatial and seasonal distribution, since the Mopti form will breed throughout the year and can therefore displace the other forms in irrigated areas, leading to changes in the patterns of malaria entomology (Touré *et al.*, 1994). There is evidence that marginal arid environments of Burkina Faso, where the Savanna and Mopti chromosomal forms correspond to the S and M molecular forms respectively (Favia *et al.*, 1997), these two taxa contrast significantly in the way they exploit limiting resources, such as larval breeding and adult resting habitats. The M molecular form shows the closest association with the domestic environment and larval habitats created

by human activities, whereas the S form is more frequent in rain-dependent temporary breeding sites (Favia *et al.*, 1997).

In many areas in tropical Africa *An. gambiae* s.s. is the main malaria vector; however it is responsible for only part of the transmission in Africa as a whole. *Anopheles gambiae* s.s. is highly anthropophagic and rests predominantly in houses (endophilic) and within the complex, it has the highest vectorial capacity. The vectorial potency of *An. gambiae* s.s. stems from its strong association with humans, that is, its preference for biting humans exacerbated by its capacity to exploit changes in its natural habitat induced by *Homo sapiens* (della Torre *et al.*, 2002). It is an efficient vector of malaria, bancroftian filariasis and some arboviruses.

2.4 Symbiotic Associations

The term symbiosis, which simply means, "living together", was first coined in 1879 by a German botanist, Heinrich Anton De Bary to describe the relationship between certain species of algae and fungi to form lichen (Cheng, 1973). It is however defined as any two organisms living in close association, commonly one living in or on the body of the other, as contrasted with "free living" (Roberts & Janovy, 1996).

The original definition of symbiosis did not include a judgment on whether the partners benefit or harm each other. Currently, most people use the term symbiosis to describe interactions between the symbiont (the smaller organism) and the host (the larger organism) from which both partners benefit; this is also called mutualism (Gray, 2000). If there is a negative effect on one of the partners, it is called a parasitic symbiosis and if there is no beneficial or negative effect it is a commensal symbiosis (Gray, 2000). These clear-cut definitions are not always easy to apply in nature. For example, the bacterium *Pseudomonas aeruginosa* can be found on the skin of humans and does not cause disease, and hence can be called a commensal, but if the person has a severe burn, *P. aeruginosa* can cause infection and become a pathogen (Gray, 2000).

Specifically, under the broad heading of symbiosis, types of associations known as phoresis, commensalism, parasitism and mutualism can be identified. Whether an association is a mutualism, commensalism or parasitism depends on the relative "strengths" of the partners and the balance of power can change over time (Gray, 2000). The associations can also differ in their spatial relations (Roberts & Janovy, 1996). If one of the organisms exists outside the cells of the other, the relation is called ectosymbiosis; if one of the organisms lives inside the cells of the other, it is called endosymbiosis.

Symbiosis may occur between two kinds of plants (e.g. lichen-forming alga and fungus), two kinds of animals (e.g. herbivores and cellulose-digesting gut microorganisms), or a plant and an animal (e.g. fig and fig wasp).

2.4.1 Types of symbiotic associations

2.4.1.1 Phoresis

Phoresis is a fairly loose association and is a specialized form of commensalism in which one organism (phoront), usually smaller than the other, uses the larger organism as a transport host (Cheng, 1973). There is no physiological or biochemical dependence on the part of either participant (Roberts & Janovy, 1996). Examples are bacteria on the legs of a fly or fungus spores on the feet of a beetle.

2.4.1.2 Commensalism

Commensalism, "eating together at the same table", is an association in which one member, usually the smaller, derives benefit from the association, whereas for the other member, the association is neither beneficial nor harmful (The Hutchinson Dictionary of Science, 1998). The relationship can be that of sharing space, substrate, defence, shelter, transport or food. Most often these associations are facultative, that is the commensal will not die if it does not enter into the association as opposed to an obligate association. A classic example of commensalism is the unique relationship between the pilot fish (*Naucrates ductor*) and the shark (Cheng, 1973). The pilot fish accompanies the shark in a free-swimming manner, eating the fragments of food that become available as the shark feeds.

2.4.1.3 Mutualism

Mutualism, as the named suggests, is an association in which both organisms derive mutual benefit (Cheng, 1971). These associations are frequently very intimate and obligatory, since, in most cases, physiological dependence has evolved to such degree that if the two organisms are separated, neither will survive (Roberts & Janovy, 1996). MacInnis (1976) defines mutualism as when each of the interacting species functions as both a host and a parasite. Termites, and the cellulose-digesting protozoans or bacteria that live in their intestines, are an excellent example of mutualism; neither organism can survive without the other.

2.4.1.4 Parasitism

Parasitism is an association in which an organism (parasite) living in or on another living organism (host), obtains from the host part or all of its organic nutriment, usually to the detriment of the host (Roberts & Janovy, 1996). Parasites may cause mechanical injury, such as boring a hole into the host or digging into its skin or other tissues, by stimulating a damaging inflammatory or immune response, or simply by robbing the host of nutrients (Roberts & Janovy, 1996). The relationship may be permanent, as in the case of tapeworms found in the intestines of mammals, or very temporary, as during the feeding of mosquitoes, leeches, and ticks on their hosts' blood. Parasitism is said to be obligatory because the parasite cannot normally survive if it is prevented from making contact with the host (Cheng, 1973).

2.4.2 Modes of how symbionts get together

All symbiotic associations must meet the challenge of maintaining their partnerships over successive generations. In the well-known associations between bacteria and metazoans,

the bacterial symbiont must somehow become intimately associated with its metazoan host. There are three major modes of getting this accomplished.

The first mode involves external transfer. Here, the metazoan host is born free of bacterial symbionts and becomes infected from the environment. This appears to be the case of the squid infections with *Vibrio fischeri*. The squid is born without these bacteria and acquires them from the water (Nuholm *et al.*, 2000). The second method is horizontal (intertaxon) transfer. Here, the hosts are born free of symbionts, but they acquire them from members of the parental generation. The third method, vertical transfer, involves gametes (most commonly, the oocyte) carrying the symbionts to the next generation.

Several types of vertical transmission mechanisms have been observed (Krueger *et al.*, 1996). Females of the oligochaete worms *Inanidrilus leukodermatus* and *I. planus* appear to infect their offspring by smearing symbionts onto their eggs as the eggs pass through symbiont-coated genital pads as they leave the mother (Giere *et al.*, 1991). Certain clams (such as *Solemya reidi, S. velum*, and *Calyptogena soyae*) produce oocytes that contain the microbial symbionts already in their cytoplasm. Polymerase chain reaction techniques and electron microscopy have found bacterial symbionts in the host ovaries, eggs, and larvae (Endow & Ohta, 1990; Cary, 1994; Krueger *et al.*, 1996).

2.4.3 Bacterial symbionts

The cells of eukaryotes constitute the sole habitat of a vast and varied array of specialised prokaryotic lineages.

The bacteria that inhabit animal cells can be divided into several groups (Moran & Baumann, 2000). The most distinctive are 'primary' symbionts that reside within the specialised host cells called bacteriocytes. They have reciprocally beneficial – often reciprocal obligate - relationships with hosts, and occur in many terrestrial arthropods as well as some marine invertebrates. 'Secondary' symbionts and intracellular pathogens are more sporadically associated with host individuals and vary in the tissues occupied. The effects on hosts are usually not known, therefore there is no clear demarcation between symbionts and pathogens; a wide range of interactions certainly exists between non-bacteriocyte associates and their hosts (Bandi *et al.*, 1999). Infection of hosts can be strictly maternal, maternal with occasional horizontal transfer, or entirely horizontal, and there is no absolute correspondence between mode of transmission and effects on host fitness (Moran & Baumann, 2000).

Close associations of bacteria with higher organisms can either have a mutualistic, commensal or parasitic character (Hooper & Gordon, 2001). However, although the final outcome of such associations is entirely different, the analysis of the interactions between these microorganisms and their host species reveals many common themes. For example, in all cases, the bacteria require specific surface structures on the host to adhere to the host (Gross, 2002). They also have to compete for nutrients, and modulate or adapt their genetic programmes to the specific requirements on or inside the host organisms. Moreover, symbionts and parasites can colonize similar niches, either in extracellular locations or inside host cells. Accordingly, many symbionts have close relatives known

to be important pathogens for man or animals. For example, many obligate intracellular symbionts of insects, helminths or trypanosomatides are closely related to pathogens such as *Rickettsia* spp., *Coxiella burnetii*, *Francisella thulariensis*, *Bordetella bronchiseptica* or the *Enterobacteriaceae* (Gross, 2002).

Molecular phylogenetic analyses have revealed the relationships of endosymbionts to several groups of bacteria (Moran & Telang, 1998). Several, including the associates of aphids (Baumann *et al.*, 2000), tsetse flies (Chen *et al.*, 1999), psyllids (Spaulding & von Dohlen, 1998), ants (Schröder *et al.*, 1996) and some bivalves (Peek *et al.*, 1998; Distel, 1998), are related to the enteric bacteria, within the γ-*Proteobacteria*. Other groups of bacteria have also given rise to endosymbionts, such as the Flavobacteria in cockroaches (Bandi *et al.*, 1995). Among the γ-*Proteobacteria*, endosymbionts of different host groups have evolved as independent lineages from nonsymbiotic bacteria (Moran & Telang, 1998).

2.5 Molecular Genetic Analysis

2.5.1 Extraction and purification of genomic DNA

The extraction and purification of DNA is a key step for most protocols in molecular biology and all recombinant DNA techniques. Most DNA *in vivo* is present in association with RNA and proteins. Proteins directly involved in the process of gene expression, such as RNA polymerase regulatory proteins, interact with DNA *in vivo* to form nucleoprotein complexes. DNA polymerase, DNA ligase, various unwinding and supercoiling enzymes, recombination and repair enzymes, and proteins involved in the initiation or maintenance of DNA replication are also associated with DNA *in vivo* (Rodriguez & Tait, 1983). Various tissues are known to contain large quantities of proteins, cell wall materials, phenolic compounds, etc., which tend to contaminate the DNA extract. It is, therefore necessary, first to isolate crude complexes of nucleoproteins from cells and then purify the DNA by separating it from proteins and RNA. Because a plethora of methods exists for extraction and purification of nucleic acids, researchers usually choose the most suitable technique depending on; (i) the target nucleic acid (ssDNA, dsDNA, total RNA, mRNA, etc.), (ii) source organism (mammalian, lower eukaryotes, plants, prokaryotes and viruses), (iii) starting material such as whole organ, tissue, cell culture, blood, (iv) desired results (yield, purity, purification time required) and (v) downstream application (PCR, cloning, labelling, blotting, RT-PCR, cDNA synthesis, RNase protection assays, etc.) [Roche, Applied Science, 2002]. The quality of extracted DNA thus depends on the source material (tissue), presence/ absence of contaminants, method of extraction and purification, and the precautions taken to maintain the integrity of the DNA.

The basic steps involved in DNA extraction and purification are; (i) the release of soluble, high molecular weight DNA from disrupted cells and membranes (ii) dissociation of

DNA-protein complexes by denaturation or proteolysis and (iii) the separation of DNA from other macromolecules (Marmur, 1963; Rodriguez & Tait, 1983). Often, the ideal lysis procedure is a compromise. It must be rigorous enough to fragment the complex starting material (e.g. blood, tissue), yet gentle enough to preserve the target nucleic acid. Common lysis procedures include mechanical disruption, chemical treatment and enzymatic digestion. Mechanical disruption methods include grinding, hypotonic lysis and freeze-thawing. Examples of chemical treatment methods include detergent lysis, use of strong chaotropic agents such as guanidinium thiocyanate and cesium trifluoroacetate and thiol reduction. Enzymatic digestion methods include treatment with either Proteinase K or Pronase. Cell membrane disruption and inactivation of intracellular nucleases may be combined (Roche Applied Science, 2002). For instance, a single solution may contain detergents such as sodium dodecyl sulphate (SDS) to solubilize cell membranes and strong chaotropic salts (for example, ethylene diamine tetra acetate, EDTA) to inactivate intracellular enzymes. After cell lysis and nuclease inactivation, cellular debris may easily be removed by filtration or precipitation.

Methods for purifying nucleic acids from cell extracts are often combinations of extraction/precipitation, chromatography, centrifugation, electrophoresis, and affinity separation. Solvent extraction is often used to eliminate contaminants from nucleic acids. After cell lysis, the cell debris and proteins are removed by denaturation and centrifugation. There are several methods for deproteinizing the lysed cell suspension and these include shaking with a mixture of chloroform/isoamyl alcohol (Marmur, 1963), enzymatic degradation of the proteins with Pronase or Proteinase K (Hotta & Bassel, 1965), and shaking with phenol (Kirby, 1968). The phenol disrupts cellular integrity and denatures proteins, and the final extraction with chloroform removes traces of phenol. Protein in the lysate can also be solubilized by treatment with SDS. Degradation by

DNAse and divalent metal ion contamination is prevented by the presence of chelating agents and by the action of SDS (Marmur, 1963). The DNA is selectively precipitated by the addition of ethanol or isopropanol. Precipitation with alcohol serves to concentrate the high molecular weight DNA whilst removing the small oligonucleotides of DNA and RNA, detergent, and the organic solvents used in the removal of proteins (Rodriguez & Tait, 1983). A subsequent wash with 70% ethanol, followed by brief centrifugation, removes residual salt and moisture. To further purify the DNA, the enzyme ribonuclease (RNase) may be used to digest the contaminating RNA (Marmur, 1963; Rodriguez & Tait, 1983). If the amount of target nucleic acid is low, an inert carrier such as glycogen can be added to the mixture to increase precipitation efficiency.

2.5.2 Restriction endonucleases

The class of enzymes known as restriction endonucleases have played a key role in the development of recombinant DNA technology. These bacterial enzymes possess an endonuclease activity which is directed to a specific sequence of bases in double-stranded DNA. In nature, they serve to protect bacteria from the possible incorporation of foreign DNA into their genomes by digesting such material (Arber & Duissoix, 1962). The bacterium's own DNA is protected by being methylated on A or C residues, which renders it unavailable for digestion by its own enzymes (Smith & Nathans, 1973; Watson *et al*., 1992). It has been suggested that these enzymes may also play a role in promoting "site-specific illegitimate recombination", allowing incoming DNA to be cleaved and incorporated into the chromosome (Rodriguez & Tait, 1983). The term "restriction" arose because it was originally found that certain bacteriophages would not grow on certain bacterial strains; hence they were said to be restricted. Investigation of this phenomenon revealed that it was due to the action of this class of enzymes (Arber, 1979).

Restriction enzymes have been classified into three groups. Types I and III enzymes carry modification (i.e. methylation) and ATP-dependent restriction (cleavage) activities in the same protein (Sambrook *et al.*, 1989). Type III enzymes cut the DNA at the recognition site and then dissociate from the substrate; while Type I enzymes bind to the recognition sequence but cleave at random sites when the DNA loops back to the bound enzyme. Neither Type I nor III restriction enzymes are widely used for molecular biology studies.

Type II restriction/modification systems are binary systems consisting of a restriction endonuclease that cleaves a specific sequence of nucleotides, and a separate methylase that modifies the same recognition sequence (Sambrook *et al.*, 1989). These enzymes recognize specific sequences of four to eight base pairs in length (Watson *et al.*, 1992). The sequences in the two strands of DNA that are recognized by the enzymes possess a two-fold axis of symmetry (Rodriguez & Tait, 1983; Sambrook *et al.*, 1989). The location of cleavage sites within the axis of a dyad symmetry differs from enzyme to enzyme. Some enzymes make cuts which are exactly opposite in the two DNA strands, so that the ends are said to be 'blunt'. Others cleave each strand at similar locations on opposite sides of the axis of symmetry, creating fragments of DNA that carry protruding single-stranded termini (Rodriguez & Tait, 1983; Sambrook *et al*., 1989).

Restriction enzymes that cut specific sequences have been isolated from several hundred bacterial strains, and over 150 different specific cleavage sites have been found (Watson *et al.*, 1992). In order to simplify the naming of these enzymes, a nomenclature has been developed that is based on an abbreviation of the name of the organism from which the enzyme was isolated (Smith & Nathans, 1973; Smith, 1979). The first initial of the genus and the first two initials of the species form the basic name. This may be followed by a strain designation, when the enzyme is present in a specific strain, or a Roman numeral to

differentiate enzymes from the same source. For example, *Hae* II is one of the three enzymes purified from the strain *Haemophilus aegypticus* and *Hinf* I is the enzyme purified from *Haemophilus influenzae* strain f (Rodriguez & Tait, 1983). In general, different restriction enzymes recognize different sequences. However, there are many examples of enzymes isolated from different sources that cleave within the same target sequences. These are known as isoschizomers (Rodriguez & Tait, 1983; Sambrook *et al.*, 1989). Examples of such enzymes are *Hind* III and *Hsu* I (Rodriguez & Tait, 1983).

2.5.3 Restriction fragment length polymorphism (RFLP)

Polymorphism of nucleotide sequences in the DNA of several organisms has proven useful in distinguishing between strains and determining their relatedness to one another (Patterson & Hyypia, 1985).

Until recently, polymorphisms could be detected only if they were expressed by differences in the behaviour of a protein, for example, by differences in enzymatic activity or electrophoretic mobility. This situation changed dramatically with the realisation that sites recognized by restriction endonucleases could be polymorphic (Watson *et al.*, 1992). This is because mutations would have caused the loss of sites at which a particular restriction enzyme can act. Since these sites are only present in the genome of certain species, they are polymorphic.

Polymorphisms detected in this way are known as restriction fragment length polymorphisms. RFLPs can be used to characterize an organism, levels of genotypic diversity, and phylogenetic relationships. They can arise from point mutations leading to either a loss or a gain of a site at which a restriction endonuclease acts. In addition,

deletions or insertions resulting in variations in the number of tandemly repeated DNA sequences can alter the length of fragments between two endonuclease sites (Watson *et al.*, 1992). Since DNA polymorphisms can occur in any type of DNA sequence, it means that an alteration producing a restriction fragment polymorphism length can occur within a coding sequence of a gene, noncoding sequences (introns), sequences between genes, and even DNA with no known function, such as repetitive DNA. DNA polymorphisms can affect any part of the genome and are therefore extremely valuable markers (Rothwell, 1988).

Restriction fragment length polymorphism (RFLP) is most suited for taxonomic studies at the intraspecific level or among closely related taxa (Avise *et al.*, 1989). Presence and/or absence of fragments resulting from changes in recognition sites are used in identifying species and in characterising populations. For example, if the total genomic DNA is digested to completion with a high frequency restriction endonuclease, an extremely large number of fragments are produced that can be separated by electrophoresis and will show up as either weak or strong bands. When DNA from different species are analysed this way the pattern of bands will be different. These restriction spectra will be specific, not only for individual restriction endonucleases but also, for the genomes of the different species (Avise *et al.*, 1989). Thus, RFLPs are used to measure genetic divergence between different populations or related species. The restriction-site difference is effectively a DNA difference, so a measure of the total number of RFLP differences represents a measure of genetic differences, and are therefore important in evolutionary studies (Griffiths *et al.*, 1999)

2.5.4 Agarose gel electrophoresis

Electrophoresis has also played an essential role in the elucidation of the structure, sequence and function of DNA. Electrophoresis through agarose or polyacrylamide gels is the standard method used to separate, identify and to purify DNA fragments (Sambrook *et al.*, 1989). The method is used for DNA fragments generated after endonuclease digestion, before or after enzymatic modification, ligation with other fragments, or after sequencing (Perbal, 1988). The technique is simple, rapid and capable of resolving DNA that cannot be separated adequately by other procedures, such as density gradient centrifugation (Sambrook *et al.*, 1989)

Agarose, which is extracted from seaweed, is a linear polysaccharide (Sambrook *et al.*, 1989; Pharmacia, 1989). It has large pores and allows rapid run times. It is easy to prepare and can be set into a variety of shapes, sizes and porosities, and can be run in a number of different configurations. The choices within these parameters depend primarily on the sizes of the DNA fragments to be separated. Although agarose gels have lower resolving power than polyacrylamide gels, they have a greater range of separation. DNA fragments from 70 bp to approximately 800 kb in length can be separated on agarose gels of various concentrations (Fangman, 1978). The location of DNA within the gel can be determined directly by staining with low concentrations of the fluorescent intercalating dye, ethidium bromide. Bands containing as little as 1-10 ng of DNA can be detected by direct examination of the gel in ultraviolet light (Sharp *et al.*, 1973).

Agarose gels are usually run in an horizontal configuration in an electric field of constant strength and direction. When an electric field is applied across the gel, DNA, which is negatively charged at neutral pH, migrates towards the anode. The rate of migration is

determined by several factors including the concentration of the agarose gel, molecular weight and conformation of the DNA.

Molecules of linear double-stranded DNA in an electric field tend to orientate in an end-on position (Fisher & Dingman, 1971; Aaij & Borst, 1972) and migrate through gel matrices at rates that are inversely proportional to the $\log_{10}$ of the number of base pairs (Helling *et al.*, 1974). Larger molecules migrate more slowly than the smaller molecules through the pores of the gel due to greater frictional drag (Sambrook *et al* . 1989).

The different conformations of DNAs, namely, superhelical circular (form I), nicked circular (form II), and linear (form III) of the same molecular weight, migrate through agarose gels at different rates (Thorne, 1967). The relative mobilities of the three forms depend primarily on the agarose concentration in the gel, but they are also influenced by the strength of the applied current, the ionic strength of the buffer, and the density of the superhelical twists in the form I DNA (Johnson & Grossman, 1977).

A linear DNA fragment of a given size migrates at different rates through gels containing different concentrations of agarose (Sambrook *et al.*, 1989). There is a linear relationship between gel concentration and $\log_{10}$ of the electrophoretic mobility of DNA (Sambrook *et al.*, 1989; Pharmacia, 1989). Thus, by using gels of different concentrations (different porosities), it is possible to resolve a wide range of DNA molecules.

Both the separation and resolution of DNA fragments are affected by the voltage gradient (Southern, 1979; Pharmacia, 1989). At low voltages, the rate of migration of linear DNA fragments is proportional to the voltage applied. However, as the electric field strength is raised, the mobility of high-molecular-weight fragments of DNA increases differentially.

Thus, the effective range of separation in agarose gels decreases as the voltage applied is increased (Sambrook *et al.*, 1989). A balance has to be struck between resolution and separation. Low-molecular-weight fragments diffuse and are thus best separated at fairly high-voltage gradients. Large fragments, however, diffuse very slowly, and the best resolution is achieved by using low-voltage gradients and running for long times (McDonell *et al.*, 1977; Phamacia, 1989).

DNA molecules of larger than 50-100 kb in length migrate through agarose gels at the same rate if the direction of the electric field remains constant (Smith & Cantor, 1987; Sambrook *et al.*, 1989). However, because of the sieving effect of the gel matrix, if the direction of the electric field is altered periodically, the DNA molecules are forced to change course. The time it takes for a molecule to reorient itself in the new electric field depends on its length (Smith & Cantor, 1978). Larger molecules take longer to orient to the new direction of the field; as a result pulse-field gel electrophoresis is used to fractionate extremely large molecules of DNA, up to about 10,000 kb (Smith & Cantor, 1987; Sambrook *et al.*, 1989).

The behaviour of DNA in agarose gels, in contrast to polyacrylamide gels (Allet *et al.*, 1973), is not significantly affected by either the base composition of the DNA (Thomas & Davis, 1975) or by the temperature at which the gel is run (Sambrook *et al.*, 1989). Most agarose gels are run at room temperature, because the relative electrophoretic mobilities of DNA fragments of different sizes do not vary between 4°C and 30°C. However, gels containing less than 0.5% agarose and low-melting temperature agarose gels are rather flimsy, and are best run at 4°C, where they are stronger (Sambrook *et al.*, 1989).

Ethidium bromide, a fluorescent dye that is used to detect DNA in agarose and polyacrylamide gels, reduces the electrophoretic mobility of linear DNA by about 10-15% (Perbal, 1988; Sambrook *et al.*, 1989). The dye intercalates between stacked base pairs, extending the length of linear and nicked circular DNA and making them more rigid (Sambrook *et al.*, 1989).

The composition and ionic strength of electrophoresis buffers do also affect the mobility of DNA (Sambrook *et al.*, 1989). In the absence of ions, electrical conductance is minimal and the DNA migrates slowly, if at all. In buffers of high ionic strength, electrical conductance is very efficient and significant amounts of heat are generated. If overheating should occur, the DNA bands will be distorted, the DNA denatured or the gel melts (Sambrook *et al.*, 1989). Several different buffers are available for electrophoresis of double-stranded DNA. These contain EDTA (pH 8.0) with either Tris-acetate (TAE), or Tris-borate (TBE), or Tris-phosphate (TPE) at a final concentration of approximately 50 mM (pH 7.5-7.8) [Sambrook *et al.*, 1989].

2.5.5 Determination of DNA fragment sizes

Estimation of molecular size of nucleic acids in either acrylamide or agarose gels after electrophoresis is important for the identification and the characterisation of restriction fragments and for studies of native and denatured DNA.

To estimate the size of DNA fragments from their mobilities in gel electrophoresis, a relationship is established between the mobilities and the lengths of standard fragments (Southern, 1979; Duggleby *et al.*, 1981; Elder & Southern, 1983). This relationship is then used to calculate the lengths of unknown fragments from their mobilities. The accuracy with which fragment lengths can be estimated depends on the accuracy of the

chosen relationship between mobility and length (Duggleby *et al.*, 1981; Elder & Southern, 1983; Rochelle *et al.*, 1985; Oerter *et al.*, 1990).

Numerous methods have been proposed for graphical and computer analyses of the relationships between mobility and length (Duggleby *et al.*, 1981; Elder & Southern, 1983; Rochell *et al.*, 1985; Oerter *et al.*, 1990). The most commonly used method is the plotting of the logarithm of the molecular weight against the mobility of standard fragments, and estimating the lengths of unknown fragments from the resulting graph (Duggleby *et al*. 1981; Schaeffer & Sederoff, 1981; Elder & Southern, 1983). The standard curves obtained from such plots often show pronounced curvature which may introduce significant subjectivity into the interpolation (Duggleby *et al.*, 1981; Elder & Southern, 1983). As a result, linear models have been developed that more or less fit with experimental data (Aaij & Borst, 1972; Duggleby *et al.*, 1981; Schaeffer & Sederoff 1981; Elder & Southern, 1983).

2.5.6 Polymerase Chain Reaction (PCR)

The polymerase chain reaction (PCR) is an *in vitro* technique for the enzymatic amplification of a specific DNA fragment of interest from small amounts of a longer molecule. The technique was invented by Kary Mullis (Mullis *et al.*, 1986; Mullis & Faloona, 1987), and was originally used to amplify human (β)-globin DNA and for prenatal diagnosis of sickle-cell anaemia (Saiki *et al.*, 1985; Saiki *et al.*, 1986; Embury *et al.*, 1987).

A typical amplification reaction medium includes a thermostable DNA polymerase, oligonucleotide primers, deoxynucleotide triphosphates (dNTPs), reaction buffer, magnesium and optional additives and a template DNA. The reaction is carried out using

an automated thermal cycler (reactor), which takes the reaction through series of different temperatures for varying time durations. Each PCR cycle theoretically doubles the amount of targeted template sequence (amplicon) in the reaction: 10 cycles theoretically multiply the amplicon by a factor of about a thousand; 20 cycles, by a factor of more than a million in a matter of hours (White *et al.*, 1989).

Each cycle of PCR amplification consists of three phases (Fig. 2.5). The first phase is the denaturation or separation of the two strands of the duplex DNA molecule. This is accomplished by briefly heating the reaction mix to temperatures of about 92-95°C. Each strand serves as a template on which a new strand is built. The temperature is reduced during the second phase so that the primers which are oriented with their 3' ends pointing towards each other can anneal to the template (White *et al.*, 1989). The temperature is then raised during the third phase to the optimum for the DNA polymerase to add nucleotides onto the ends of the annealed primers (extension). The time of incubation at this temperature varies according to the length of target being amplified (Saiki, 1989). At the end of the cycle, which lasts for a few minutes, the temperature is raised to 92-95°C for denaturation to commence another cycle. A typical PCR of usually 25 to 30 cycles produces a sufficient amount of DNA for further experimental procedures e.g. cloning.

A major problem encountered in the original PCR procedure was that the DNA polymerase (*Escherichia coli* DNA Polymerase I, Klenow fragment) had to be replenished after every cycle because it is not stable at the high temperatures needed for denaturation (Saiki *et al.*, 1985. Mullis & Faloona, 1987). This problem was solved in

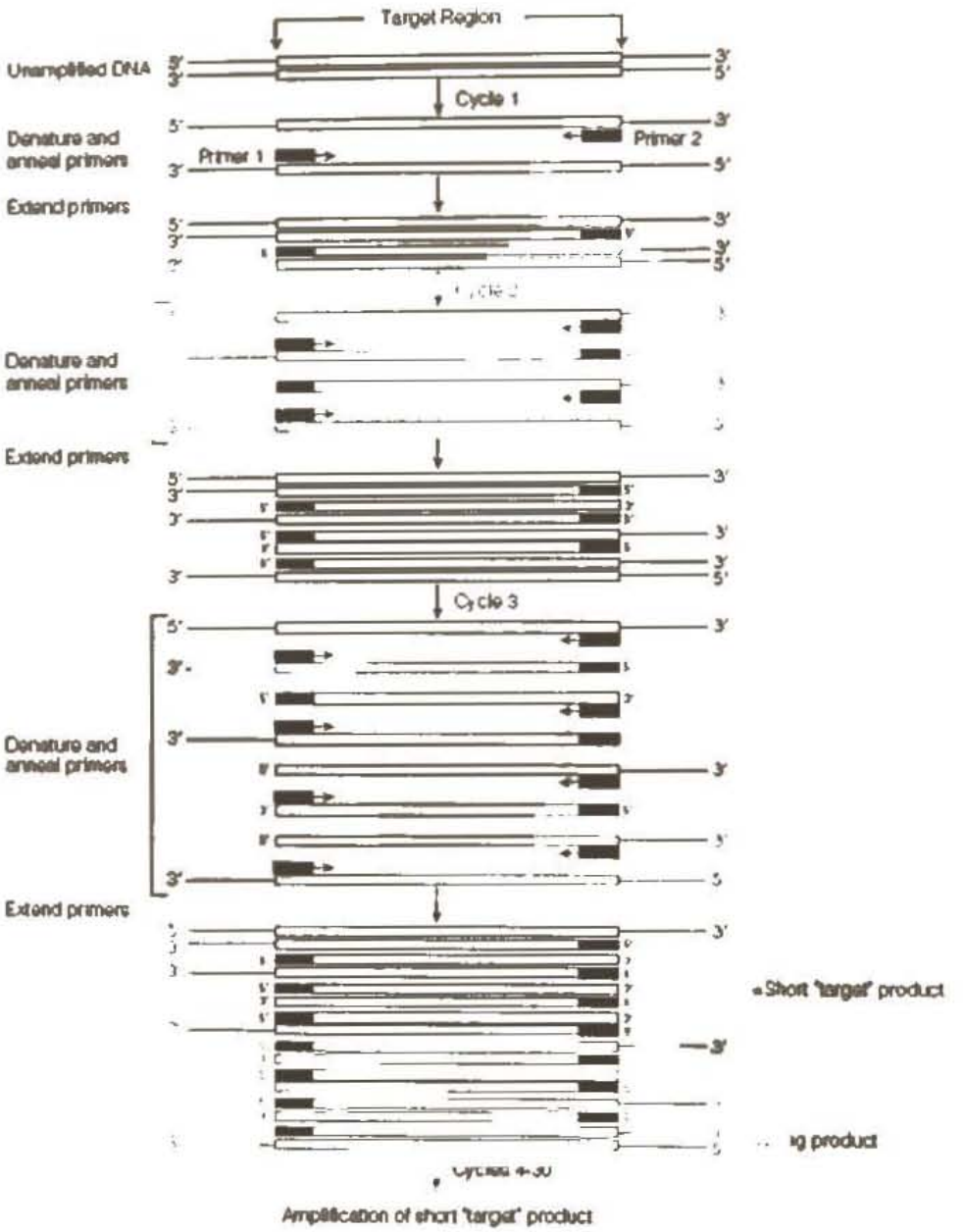


Fig. 2.5 Schematic diagram of the PCR process (Protocols and Application Guide, 3rd Edition, Promega Corporation, Madison, USA).

1987 with the discovery of a heat-stable DNA polymerase called *Taq* (Saiki *et al.*, 1988); an enzyme isolated from the thermophilic bacterium, *Thermus aquaticus*, which inhabits hot springs (Chien *et al.*, 1976). *Taq* DNA polymerase has an optimal extension rate (polymerisation rate) of 35-100 nucleotides per second at between 70-80°C (Newton & Graham, 1997). The higher temperature optimum for the *Taq* polymerase extension allowed the use of higher temperatures for primer annealing and extension, thereby increasing the overall stringency of the PCR reaction and minimising the extension of primers that were mismatched with the template. The increase in the specificity of *Taq* polymerase results in an improved yield of amplified target fragment by reducing the competition by nontarget products for enzyme and primers (White *et al.*, 1989). Apart from *Taq* DNA polymerase, its cloned and modified version, Tth DNA polymerase from *Thermus thermophilus*, *Pfu* DNA polymerase from *Pyrococcus furiosus* and a few others can also be used for PCR (Newton & Graham, 1997).

The annealing step is an important parameter in enhancing the specificity of a PCR reaction. Usually, the temperature chosen for annealing of the primer is a compromise (Saiki *et al.*, 1988). This is because at lower temperatures annealing is more efficient but there is a significantly increased amount of mispriming. At higher temperatures, however, there is increased specificity but the overall efficiency of amplification is decreased. Under standard conditions, the annealing temperature in a reaction should be 5°C lower than the melting temperatures (T_m) of the primers (Innis & Gelfand, 1990). To determine the approximate melting temperatures of the primer-DNA duplex, the equation

$$T_m = 4(G + C) + 2(A + T)$$

(where G = dGTP, C = dCTP, A = dATP and T = dTTP) can be used (Winnacker, 1987). More exact numbers can be calculated by applying algorithms (Rychlik & Rhoads, 1989) that take into account the occurrence of intramolecular interaction, such as, hairpin

structures. Normally, a series of reactions are set up to determine the optimal annealing temperature.

A number of factors influence the specificity of the amplification reactions. The stringency of the annealing step can be controlled to some extent by adjusting the annealing temperature. Minimizing the incubation time during the annealing and extension steps will limit the opportunities for mispriming and extension by molecules of otherwise idle polymerase (Saiki, 1989). Reducing primer and enzyme concentrations also serves to limit mispriming, particularly the type that leads to dimerization. Finally, changing the concentration levels of $MgCl_2$ can further improve specificity, either by increasing the stringency of the reaction or by direct effects on the polymerase itself (Saiki, 1989).

The PCR method can be used with a complex template such as genomic DNA and can amplify a single-copy gene contained therein. It is also capable of amplifying a single molecule of target in a complex mixture of DNAs and RNAs (Saiki *et al.*, 1988) and can, under certain conditions, produce fragments up to 35-42 kbp long (Barnes, 1994; Cheng *et al.*, 1994). The starting material, the template DNA may be single stranded or double stranded DNA. The DNA to be amplified should be free of contaminants or non-target materials to avoid their amplification rather than the target DNA. Normally, subnanogram quantities of the template DNA are used for PCR.

The PCR is used in a wide variety of fields including: molecular biology, environmental science, forensic science, medical science, biotechnology, microbiology, the food industry, diagnostic science, epidemiology, genetics, gene cloning, and many more. It has been used in the diagnosis of genetic disorders (Saiki *et al.*, 1985; 1988), the analysis

of allelic sequence variation (Sambrook *et al.*, 1989), and analysis of mutation or any research that involves the rapid cloning and sequencing of homologous DNA fragments (Gyllensten, 1989). It has also been used for the analysis of individual identity in forensic samples by the amplification of highly polymorphic DNA regions (Higuchi *et al.*, 1988), and the examination of nucleotide sequences from ancient preserved specimens (Paabo *et al.*, 1988; Paabo, 1990). The PCR has been applied to difficult problems in developmental biology. For example, PCR of cDNA has been used to study V-J region combinations in the T-cell receptor δ-chain (Loh *et al.*, 1989) and the examination of the mRNAs for growth factors in small numbers of macrophages isolated from wounds actively undergoing healing (Rappolee *et al.*, 1988).

2.6 Molecular Phylogenetic Analysis

Molecular phylogeny is the study of evolutionary relationships among organisms by using techniques of molecular biology. It began at the turn of the century, even before Mendel's laws were discovered in 1900. Immunochemical studies had shown that serological cross-reactions were stronger for closely related organisms than for distantly related ones. The evolutionary implications of these findings were used by Nuttal (1904) to infer phylogenetic relationships among various groups of animals.

In the 1960s and 1970s the study of molecular phylogeny, using protein sequence data, progressed tremendously. Less expensive and more expedient methods, such as protein electrophoresis, DNA-DNA hybridization, and immunological methods, though less accurate than protein sequencing, were used to study the phylogenetic relationships among populations or closely related species. The application of these methods stimulated the development of measures of genetic distance and tree-making methods (Fitch & Margoliash, 1967; Nei, 1975; Felsenstein, 1988).

The rapid accumulation of DNA sequence data from the late 1970s has had a great impact on molecular phylogeny. DNA sequence data are not only more abundant, but also easier to analyze than protein sequence data. Thus, they have been used in a wide range of applications: reconstructing the ancestral gene sequences from which extant genes are derived; studying the origin and epidemiology of human diseases; inferring the evolution of ecological and behavioral traits through time; estimating historical biogeographic relationships; prioritizing the conservation of endangered species; and reconstructing the historical relationships across all of life (Woese, 1987; Hillis, 1997).

Whatever the application, all phylogenetic analyses require a criterion (known as an optimality criterion) for assessing how well the data fit candidate trees, a method for searching among possible solutions for the tree with the best fit to the data, and a method for assessing confidence in the results (Hillis, 1997).

2.6.1 Molecular phylogeny methods

Molecular phylogeny methods allow, from a given set of aligned sequences, the suggestion of phylogenetic trees which aim at reconstructing the history of successive divergences which took place during the evolution between the considered sequences and their common ancestor.

2.6.1.1 Phylogenetic trees

All life forms, both extant and extinct, have a common origin somewhere in the past, from which they evolved into the plethora of species found today. Thus, all animals, plants, and bacteria are related by descent to each other. Closely related species share a more recent common ancestor than distantly related species. The objectives of phylogenetic studies are to:

i) reconstruct the correct genealogical ties between organisms, and

ii) estimate the time of divergence between organisms since they last shared a common ancestor (Li & Graur, 1991).

The evolutionary relationships between species can be represented by a phylogenetic tree. This is a graph composed of nodes and branches in which only one branch connects any adjacent nodes. The nodes represent taxonomic units, while the branches connecting them reflect their relationships in terms of descent and ancestry. A phylogenetic tree is characterised by its topology (form) and its length (sum of its branch lengths) [Li &

Graur, 1991]. The topology is the pattern of branches found in a tree. The branching pattern (often called branching order) shows the genealogy of the organisms. That is, it shows which species share more common ancestry with which others. The branch length is commonly used to indicate some form of evolutionary distance represented by that branch. The actual, still existing taxonomic units are often called operational taxonomic units (OTUs); a generic term that can represent many types of comparable taxa, for example, a family of organisms, individuals of a single species, a set of related genes or even gene regions, are represented by nodes on the tips of the branches, called external nodes (Weiller *et al.*, 1995). The other nodes are called internal nodes. Internal nodes may be called hypothetical taxonomic units (HTU) to emphasise that they are the hypothetical progenitors of OTUs (Weiller *et al.*, 1995).

Phylogenetic trees can be either rooted or unrooted. A tree where a special node indicating the common ancestor to all OTUs is present is called a rooted tree. An unrooted tree leaves the position of the common ancestor unspecified. The number of bifurcating rooted trees (N_R) for n OTUs is given by

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

for $n \geq 2$ (Li & Graur, 1991).

The number of bifurcating unrooted trees (N_U) for $n \geq 3$ is

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$ (Penny *et al.*, 1982; Li & Graur, 1991).

A *species tree* is a phylogenetic tree representing the evolutionary pathway of a group of species. When a tree is constructed from one gene from each species, the inferred tree is sometimes called a *gene tree* (Nei, 1987). These two are not necessarily the same, as the divergence of the gene can actually predate the divergence of the species. This is not a problem when long-term evolution is studied. Another problem is that the topology of the gene tree does not necessarily coincide with that of the species (Li & Graur, 1991).

2.6.1.2 Search algorithms

Once a criterion has been selected for evaluating the fit of data to trees, it is necessary to search among the universe of possible trees for the optimal solution. A number of algorithms for estimating phylogeny from DNA sequences exist, and it not always clear what the strengths and weaknesses of each method are, or which should be preferred in a given situation (Kuhner & Felsentein, 1994). For a small number of taxa, there are few possible tree topologies (branching arrangements), but the number of distinct trees increases rapidly as a function of the number of taxa. There are only three branching orders for four taxa, but there are more than two million different trees for 10 taxa, and more than 2×10^{182} different trees for 100 taxa (Hillis, 1997).

For smaller data sets, there are two computational procedures, or algorithms, that are guaranteed to find the optimal tree(s) for a particular criterion. One is simply an exhaustive evaluation of all distinct trees (Saitou & Imanishi, 1989), but this can be very time-consuming and is rarely feasible for more than about 12 taxa (Hillis, 1997). The 'branch-and-bound' algorithm (Hendy & Penny, 1982) is a second algorithm that will find the exact solution(s) for 20 or more taxa, depending on the computational resources and the 'noisiness' of the data. It works by ignoring whole classes of trees that can be mathematically excluded as suboptimal compared to a known solution to the problem

(Hillis, 1997). This known solution may not be optimal, but it serves to define an upper 'bound', or limit, for the analysis.

All exhaustive search methods examine all or a large number of the theoretically possible tree topologies and use certain criteria to choose the best one (Saitou & Imanishi, 1989). Their main advantage is that they produce a large number of different trees together with a relative estimate of the likelihood that they represent the phylogeny. This allows the investigator to compare the support for the best tree with the support for the second best, and thus estimate the confidence in the tree obtained. Unfortunately the number of possible trees and thus computing time grows very quickly as the number of taxa increases; the number of bifurcated rooted trees for n OTUs is given by (2n-3)!/(2n-2(n-2))! [Cavalli-Sforza & Edwards, 1967]. This means that for a data set of more than 10 OTUs only a subset of possible trees can be examined, so various strategies are used to search the 'tree space' but there is no algorithm that guarantees that the best possible tree was examined. The algorithms that search for optimal trees can contain an optional final step, global rearrangement, which, involves removing each branch in turn and trying all possible positions for it; this improves performance but approximately triples run time (Kuhner & Felsentein, 1994).

In contrast, the stepwise clustering methods avoid this problem by examining local subtrees first (Saitou & Imanishi, 1989; Hillis, 1997). Typically, the two closest related OTUs are combined to form a cluster. The cluster is then treated like a single OTU representing the ancestor of the OTUs it replaces. Thus the complexity of the data set is reduced by one OTU. This process is repeated, clustering the next closest related OTUs, until all OTU's are combined. The various stepwise clustering algorithms differ in their methods of determining the relationship of OTUs and in combining OTUs to clusters

(Kuhner & Felsentein, 1994; Hillis, 1997). They are usually very fast and can accommodate large numbers of OTUs. As they produce only one tree, the confidence estimators of the exhaustive search methods are not available, although various other statistical methods have been developed for estimating the confidence in the correctness of a tree obtained (Li & Gouy, 1990).

The majority of distance matrix methods use a stepwise clustering to compute trees to represent their relationships while many character state methods adopt the exhaustive search approach (Saitou & Imanishi, 1989; Weiller *et al.*, 1995; Hillis, 1997).

2.6.1.3 Heuristics

For large data sets, it is necessary to rely on approximation techniques (also called heuristics) to find near optimal solutions (Hillis, 1997). Several methods have been devised for getting a quick initial approximation (or point estimate) of the optimal tree, which can be used as a starting point for more thorough analyses. Sometimes these initial approximations are treated as end-points in the estimation procedure, but this practice has some serious drawbacks. Initial point estimates can almost always be improved, except in the case of very small data sets. In addition, point-estimation methods provide no indication of how many other trees (or which trees) may be just as good, or even better, estimates of the phylogeny.

The two most widely used point estimation methods are stepwise addition and neighbor-joining (Hillis, 1997). The former method adds taxa one by one to a growing tree, whereas the latter method starts with all taxa in an unresolved 'star' and adds the internal branches to the tree in a stepwise fashion. Both methods find similar results for small trees, but stepwise addition usually finds better solutions (and is computationally slightly

slower) for large data sets.

Once a point-estimate of a tree has been found, the estimate usually can be improved in a procedure known as branch-swapping (Hillis, 1997). There are several different types of branch swapping, but all of them involve rearranging the branches on an initial tree to look for topologically related trees that are as good or better phylogenetic solutions. If branch swapping finds an equally good or better tree, then the branch swapping is continued on the new tree.

2.6.2 Tree construction methods

Numerous methods for constructing phylogenetic trees have been proposed (Sneath & Sokal, 1973; Felsentein, 1981; Nei, 1987; Swofford & Olsen, 1990). These methods can be classified into two main types: distance matrix methods and discrete-character methods (Li & Grauer, 1991). In distance matrix methods, evolutionary distances (usually the number of nucleotide or amino acid substitutions separating two taxonomic units) are computed for all pairs of data, and a phylogenetic tree is constructed by using an algorithm based on some functional relationships among the distance values. In discrete-character methods, character states (e.g., the nucleotide and amino acid at a site) are used, and the shortest pathway leading to these character states is chosen as the phylogenetic tree. All characters are analysed separately and usually independently from each other (Weiller *et al.*, 1995).

Some types of molecular data (*e.g.*, DNA hybridization data) exist only as distance data; therefore phylogenetic trees for these data can be constructed only by distance methods. By contrast, discrete-character data can usually be converted into distance data and thus they can be analyzed either by distance methods or by discrete-character methods.

There are two major groups of discrete character methods, namely, maximum parsimony methods and maximum likelihood methods. Two types of errors can occur in the construction of phylogenetic trees: topological errors and branch-length errors (Tateno *et al.*, 1982).

2.6.2.1 Discrete character methods

i) Maximum parsimony (MP) and weighted parsimony (WP)

The theoretical basis of the MP method is William of Ockham's philosophical idea that the best hypothesis to explain a process is the one that requires the smallest number of assumptions (Sober, 1988). The method was first proposed by Camin and Sokal (1965) as a method for reconstructing phylogenetic trees from morphological data. Later modifications adjusted the MP method for amino acid (Eck & Dayhoff, 1966) and nucleotide sequences (Fitch, 1971). The latest modification utilises dynamically modified character weighting (Williams & Fitch, 1990) and is called 'weighted parsimony' (WP).

A maximum parsimony tree is the tree that requires the smallest number of evolutionary changes to result in the set OTUs under study (Li & Graur, 1991; Hillis, 1997). It is often possible to find more than one tree with the same minimum number of changes, so that no unique tree can be inferred. Parsimony considers all sites, but not all sites convey information regarding the most parsimonious tree. Only sites that favour some topologies over others are informative and, consequently, only these sites are used in the calculations. For molecular sequence data, sites are informative only when there are at least two different kinds of residues at the site, each of which is represented in at least two of the sequences under study (Li & Graur, 1991; Weiller *et al.*, 1995). This usually restricts drastically the number of positions used in the analysis.

In order to find the most parsimonious tree, first, all the informative sites are identified. Next, the minimum number of substitutions at each informative site is calculated for all possible tree topologies. In the final step, the number of changes over all the informative sites for each tree is summed up and the tree associated with the smallest number of substitutions is chosen (Li & Graur, 1991).

The standard MP method weights all changes equally although for sequence data this is not always desirable (Weiller *et al.*, 1995). Firstly it is sometimes desirable to weight different sequence positions differentially in order, for example, to emphasise the few changes at conserved positions rather than the several at variable positions. Secondly it is often advisable to weight different types of changes differently, for example transversions are usually more significant than transitions, and certain amino acid substitutions more than others. Recently, Williams and Fitch (1990) introduced the dynamically weighted parsimony method (WP) which uses an initial tree (seed tree) to assign different weights to different types of substitution and/or sequence positions. It then uses this weighting scheme to generate a new tree. This process is repeated until the same best tree is obtained in two consecutive runs.

The number of possible topologies increases very sharply with the number of taxa, making computer time a limiting factor for calculations on more than a few sequences. The "branch and bound" method (Hendy and Penny, 1982) can limit the number of topologies that have to be tested, and still find the optimal tree. Even with this optimisation, the method is limited to a small number of sequences. For larger sets, heuristic methods must be used, but these cannot guarantee to find the most parsimonious tree. Parsimony is also very sensitive to unequal rates of evolution in different branches

in the tree (Felsenstein, 1978). Since it is not possible to correct for superimposed evolutionary events (where a given nucleotide position has changed more than once since the two sequences diverged), the presence of distantly related lineages can cause distortion.

The effects of the unequal rates of evolution should be solved in the evolutionary parsimony method proposed by Lake (1987), but this method has been criticized for depending on a number of unrealistic assumptions (Jin & Nei, 1990) and for underestimating branch lengths (Olsen, 1991). Information on evolutionary processes may be incorporated by weighting characters differentially (such as first versus third positions of codons), or by weighting character state changes differentially (such as transitions versus transversions) [Hillis, 1997].

ii) Maximum likelihood (ML)

The maximum-likelihood method for reconstructing phylogenetic relationships has become very popular because of its main advantage of application of a well-defined evolution model to given data set (Felsenstein, 1988).

The ML method was originally proposed by Cavalli-Sforza and Edwards (1967) for gene frequency data. Later, Felsenstein (1981) developed a ML algorithm for nucleotide sequences. Maximum-likelihood is similar to the MP method in that it examines every reasonable tree topology and evaluates the support for each by examining every sequence position.

Phylogenetic tree inference based on the method of maximum likelihood is appealing from both biological and statistical perspectives. The maximum likelihood method tries

to infer the topology that is most consistent with a set of observed data (Felsenstein, 1981; Strimmer & Haeseler, 1996). The possibility that a given topology will produce the observed sequences is calculated for all or many possible topologies (Swofford & Olsen, 1990, 1996). In order to calculate these possibilities, a concrete model of the evolutionary process that converts one sequence into another must be specified. For example, all nucleotides are assumed to be equally frequent and the probability of change of any nucleotide to any other nucleotide is assumed to be the same in the Jukes-Cantor model (Mount, 1999). For each possible tree, the likelihood of finding the actual sequence changes at each column in the aligned sequences is calculated. The probabilities for each aligned position are then multiplied to provide a likelihood for each tree. The tree which provides the maximum likelihood value is the most probable tree.

Just as for the parsimony approach, this method requires much computer time and can only be used on a limited number of sequences. However, it is possible to speed up calculations by parallelizing the algorithm by using approximative techniques (Adachi & Hasegawa, 1994; Olsen *et al.*, 1994).

2.6.2.2 Distance matrix methods

I) Transformation of sequence data to distances

Although sequences are fundamentally character data, phylogenetic analysis based on sequences is often preceded by a prior reduction to distance values relating all pairs of sequences. Many fast methods are available to construct evolutionary trees from distance data (Nei, 1987; Gojobori *et al.*, 1990). Moreover, the use of distances easily allows corrections for multiple mutations at the same site, which are more difficult to take into account when using character data (De Rijk, 1995).

In the distance method, a matrix of distance scores among all of the sequences is first drawn. The distance score is a measure of dissimilarity between the sequences, so that the less similar the sequences, the higher the distance score between them (De Rijk, 1995). Another way of thinking about genetic distance is that it represents the minimum number of changes, including substitutions, insertions and deletions that are required to convert one sequence into the other.

One of the most common ways of summarizing the relationship of two sequences to a number is their fractional similarity. In its simplest form, *sequence similarity* is the number of alignment positions containing identical residues divided by the number of alignment positions compared (De Rijk, 1995). Accounting for gaps is the biggest problem in the calculation of similarity. Sequence comparison analyses can provide either similarity or distance scores, depending on the scoring system used (Smith & Waterman, 1981; von Heijne, 1987). Similarity scores also may be converted to distance scores, based upon assumed models of evolutionary change in sequences (Swofford & Olsen, 1990, 1996).

Additive distances can be fitted to an unrooted tree such that all pairwise distances are equal to the sum of the branch lengths that connect the OTUs (De Rijk, 1995; Weiller *et al.*, 1995). Ultrametric distances are the most constrained. They will precisely fit a tree so that the distance between any two taxa is equal to the sum of the branches joining them, and can be rooted so that all taxa are equidistant from the root (De Rijk, 1995).

ii) Unweighted pair group method with arithmetic mean (UPGMA)

The unweighted pair group method with arithmetic mean (UPGMA) is the simplest method for tree reconstruction. It was originally developed for constructing taxonomic

phenograms, i.e., trees that reflect the phenotypic similarities between OTUs (Sokal & Michener, 1958), but it can also be used to construct phylogenetic trees if the rates of evolution are approximately constant among the different lineages, so that an approximately linear relation exists between evolutionary distance and divergence time (Nei, 1975). For such a relation to hold, linear distance measures such as the number of nucleotide (or amino acid) substitutions should be used.

The tree is constructed by a sequential clustering algorithm (Sneath & Sokal, 1973). Within the matrix, the two OTUs with the smallest distance (i.e., most similar) are clustered into a composite OTU, which will then be treated as a new single OTU. A new matrix is created, where the distances of other OTUs to the composite OTU are calculated. This process is repeated until only two OTUs remain. These are clustered together, and the root is placed at half the calculated distance between the two clusters.

The UPGMA calculates the distances to a composite OTU as the arithmetic mean of the distances between the constituent OTUs in each composite OTU. If a simple average of the distances between the composite OTUs is used, the method is called weighted pair group method with arithmetic mean (WPGMA) [Sneath & Sokal, 1978].

2.6.2.3 Transformed distance methods

If the assumption of rate constancy among lineages does not hold, UPGMA may give an erroneous topology. The topological errors might be remedied, however, by using a correction called the transformed distance method (Farris, 1977; Klotz *et al.*, 1979; Li, 1981). This method uses an outgroup as reference to make corrections for unequal rates and then applies the UPGMA to the new distance matrix to infer the topology of the tree. The outgroup is an OTU which, based on external knowledge such as taxonomic or

paleontologic information, clearly diverges before the common ancestor to all other OTUs (De Rijk, 1995).

It is often difficult to decide which of the taxa is an outgroup. To overcome this problem, Li (1981) proposed a two-stage approach which first infers the root using UPGMA, and then uses the taxa on either side of the root as an outgroup to correct the distances among lineages on the other side.

i) Neighbour relation methods

The neighbour-joining (NJ) method (Saitou & Nei, 1987; Studier & Keppler, 1988) constructs the tree by sequentially finding pairs of neighbours, which are the pairs of OTUs connected by a single interior node. The method is conceptually related to cluster analysis, but removes the assumption that data are ultrametric (De Rijk, 1995). In other words, it does not require that all lineages have diverged by equal amounts. As it does not attempt to cluster the most closely related OTUs, but rather minimises the length of all internal branches and, thus, the length of the entire tree, it can be regarded as parsimony applied to distance data (Weiller *et al.*, 1995).

In contrast to cluster analysis, neighbour-joining keeps track of nodes of the tree rather than taxa or clusters of taxa. The NJ algorithm starts by assuming a bush like tree that has no internal branches. In the first step, it introduces the first internal branch and calculates the length of the resulting tree. The algorithm sequentially connects every possible OTU-pair and finally joins the OTU-pair that yields the shortest tree (Weiller *et al.*, 1995). The length of a branch joining a pair of neighbours, X and Y, to their adjacent node, is the average distance between all OTUs and X as well as Y, less the average distances of all

remaining OTU-pairs. This process is then repeated, always joining two OTUs (neighbours) by introducing the shortest possible internal branch.

ii) Fitch and Margoliash (FM) method

The FM algorithm (Fitch & Margoliash, 1967) initially uses the same clustering method as UPGMA and thus the initial topology is always the same as in UPGMA, but the branch lengths are calculated differently. All trees with closely related topologies are then explored and compared in terms of the so called 'percent standard deviation'. This is essentially a measure that assesses how well the distances in the matrix correlate with the branch length obtained in the tree (patristic distances). The tree with the smallest 'standard deviation' is chosen.

iii) Distance Wagner method (DW)

As the network reconstructed from character state data is often called a Wagner network, Farris (1972) called his method for reconstructing unrooted trees from distance data "Distance Wagner". Like the FM method, the DW method uses a distance matrix to analyse many possible trees. Like many clustering algorithms, the DW method firstly combines the two most closely related OTUs. As the network grows, it tries to fit all remaining OTUs in turn into any of the possible edges, choosing the one that can be connected with the shortest branch. This method minimises the total length of each subtree as it is built (Weiller *et al.*, 1995). As it depends very much on correct estimation of every branch length, it is very susceptible to sampling errors. Improvements to the algorithm have been suggested and these yield the modified Farris method (MF) [Tateno *et al.*, 1982].

iv) Minimum evolution (ME)

The minimum evolution (ME) method (Edwards & Cavalli-Sforza, 1963; Saitou & Imanishi, 1989; Rzhetsky & Nei, 1992; Edwards, 1996) is a distance-based algorithm. The method first computes each branch length of a given tree topology by using Fitch and Margoliash's (1967) procedure for branch-length estimation, and the tree that shows the smallest sum of branch lengths is chosen as the best one. Thus the criterion used in this method is the sum of branch lengths (SBL). Rzhetsky and Nei (1993) showed that the expected value of SBL is smallest for the true topology when unbiased estimates of pairwise distances are used. However, this result does not mean that the topology with the smallest SBL value is the most probable tree (Nei, 1996).

The principle of minimum evolution, which implies search for a tree with minimum mutations to explain differences observed between sequences, was proposed by Cavalli-Sforza and Edwards (1967), who considered a Steiner tree. With the use of Fitch and Margoliash's procedure for branch-length estimation, computation becomes much simpler than the Cavalli-Sforza and Edwards method. The ME method also seems to be related to Dayhoff's (1978) method (Blanken *et al.*, 1982). Although this principle bears some resemblance to that of maximum parsimony, in that it seeks the tree with the lowest overall change in characters, it differs from parsimony in that 'change' is adjusted to account for inferred superimposed events, using a model of evolution (Hillis, 1997). The method is similar to the NJ method because the principle of minimum evolution is also adopted in the NJ method (Saitou & Nei, 1987). and because both the ME and NJ methods are distance methods. Simulation results have shown that the ME method and the NJ method are indeed quite similar. However, the ME method is an exhaustive-search method and examines all possible trees to choose the best one. In contrast, the NJ method

is a stepwise clustering method, and the computational time is much shorter in this method than in the ME method.

2.6.3 Accounting for superimposed events (Nucleotide substitution in a DNA sequence)

A basic process in the evolution of DNA sequences is the change in nucleotides with time. However, as the process of nucleotide substitutions is extremely slow, to detect evolutionary changes in a DNA sequence, comparative methods whereby a given sequence with which it shared a common ancestry in the past are used (Li & Graur, 1991). To study the dynamics of nucleotide substitution, several assumptions regarding the probability of substitution of one nucleotide by another are made. Numerous such mathematical schemes have been proposed (Li & Graur, 1991).

The substitution scheme of Jukes and Cantor's (1969) model is one of the most frequently often used ones to correct for multiple mutations per site. This model starts from the assumptions that all substitutions are independent, that all sequence positions are equally subject to change, that substitutions occur randomly among the four types of nucleotides, and that no insertions or deletions have occurred. For example, if the nucleotide under consideration is a, it will change to T, C, or G with equal probability. Since the model involves only one parameter, it is also called the one-parameter model (Li & Graur, 1991). Based on these pre-assumptions, the authors derived an equation for estimating evolutionary distances from observed dissimilarity.

Several other equations for the estimation of evolutionary distances have been proposed. For example, Kimura (1980) has provided a method for inferring evolutionary distances

based on a model of evolution in which transitions and transversions may occur at different rates. It is therefore often called the 'two parameter model'.

Golding (1983) has shown that applying the Jukes-Cantor correction to sequences composed of sites with unequal evolutionary rates leads to an underestimation of large evolutionary distances. As a result, distant species may seem closer than they actually are and this can cause artificial clustering of long branches (Olsen, 1987). This artefact can be avoided by converting dissimilarities into evolutionary distances according to Jin and Nei (1990). They assume that there is a gamma distribution of substitution rates over the sequence positions. A general equation for applying this idea in the correction of distances was proposed by Rzetsky and Nei (1994). This equation contains a parameter a which depends on the dataset under study.

Other equations are based on substitution models in which the four different nucleotides are not used in equal proportions (e.g. Tajima & Nei, 1984), or where a bias in the direction of change is accounted for (e.g. Tamura & Nei, 1993; Zharkikh, 1994).

An important drawback of most of these models is that they do not consider differences in substitution rate among the sites of a molecule.

2.6.4 Assessing the reliability of a tree (confidence in phylogenetic estimates)

All methods of tree construction make assumptions. When these assumptions are not satisfied, systematic errors can be introduced. Even if all assumptions are satisfied, random errors may occur due to the use of a limited dataset. Methods are therefore needed for testing the statistical significance of any given tree (Hillis, 1997).

There are two different types of methods for testing the reliability of a tree constructed. One is to test the topological difference between the tree and its closely related tree by using certain quantities such as the likelihood value in the maximum likelihood method (Kishino & Hasegawa, 1989) and the sum of all branch lengths in the minimum evolution method (Rzhetsky & Nei, 1992). This type of test is supposed to examine the reliability of every interior branch of the tree, and it is generally a conservative test. The procedure of the test is usually quite complicated and requires considerably more computation.

The other type of test is to examine the reliability of each interior branch whether it is significantly different from zero or not. If a particular interior branch is not significantly different from zero, the possibility of trifurcation of the branches associated or even the other types of bifurcating trees that can be generated by changing the splitting order of the three branches involved cannot be excluded (Kumar *et al.*, 1993). There are two different ways of testing the reliability of an interior branch. One way is to compute the standard error of the interior branch and test the deviation of the branch length from zero, and the other is to use the bootstrap test (Efron, 1982; Felsenstein, 1985).

When the sequences do not provide enough phylogenetic information (e. g. sequences are too short or lacking in variation), no algorithm will produce sensible answers. One way to evaluate whether there is enough phylogenetic signal in the data is to apply tests like jackknifing or bootstrapping (Efron, 1982; Felsenstein, 1985; Li & Gouy, 1990). Both methods indicate whether smaller sample sizes would result in the same tree.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Chemicals, Reagents, Equipment and Software

The sources and/or manufacturers of the chemicals and reagents, software and online analysis servers, and equipment used for the study are listed in Appendix I. The various buffers and solutions used were prepared as described in Appendix II.

3.2 Biological Specimens and Sample Collection

Mosquito larvae and pupae samples were collected from six locations, separated by a few kilometres, in the Greater Accra region of Ghana. The areas were as follows: Madina (5°40'N, 0°9'W), Adenta (5°42'N, 0°9'W), Dzorwulu (5°37'N, 0°12'W), Achimota (5°38'N, 0°14'W), Legon (5°39'N, 0°9'W) and Airport Residential Area (5°36'N, 0°11'W). The global positioning system (GPS) was used to determine the geographical coordinates of each breeding site.

Anopheles mosquito breeding sites were identified by sampling of small pools of stagnant water in the open and in gutters, exposed to sunlight (Figs. 3.1a, b, c and d). The Dzorwulu and Airport Residential sites were characterized by small shallow pools of stagnant rainwater. At the Adenta, Madina and East Legon, the sites were mainly open drains containing stagnant water exposed to sunlight, and these contained a mixture of both *Anopheles* and *Culex* pre-adult stages. The site at Achimota was water flowing from a leaking water supply pipe and exposed to sunlight.

Fig. 3.1a Mosquito collection site at Adenta (open drain with stagnant water exposed to sunlight).

ent from surrounding houses).

Fig. 3.1c Mosquito collection site at Achimota (narrow stretch of water flowing from a broken water pipe).

Fig. 3.1d Mosquito collection site at Dzorwulu (small shallow pool of stagnant water)

The *Anopheles* larvae were identified by their characteristic resting position, with the body parallel to the water surface and just below the surface film. The larvae and pupae were carefully collected into small plastic containers by scooping gently with a 350 ml dipper to avoid injuries. The samples were transported to the laboratory in 2-litre jars that were loosely capped to allow adequate ventilation.

Wild adult *An. gambiae* mosquitoes were also obtained from Navrongo (10°30'N, 1°00'W) and Dodowa (5°51'N, 0°4'W) both in Ghana and Jaribuni (3°00'S, 39°00'E) in the Kilifi District, Kenya. Laboratory colonies of adult *An. gambiae* from Kilimanjaro (Tanzania), Suakoko (Liberia) [which is the strain used as the standard reference for *An. gambiae* s.s.] and Kisumu (Kenya) and *An. arabiensis* (Wageningen strain) were obtained by courtesy of M.D Wilson.

3.3 Laboratory Rearing of Mosquitoes

In the laboratory, the larvae and pupae samples were transferred into 34 x 24 x 5.5 cm plastic basins (Fig.3.2a). In cases where larvae of *Culex* and/or *Aedes* species were present they were identified by their angular position on the water surface and were removed. Using a plastic Pasteur pipette, a few *Anopheles* larvae and pupae were transferred and placed in 1.5 ml eppendorf tubes containing 1 ml of isopropanol and stored at 4°C.

The remaining pupae were similarly transferred into small plastic cups and placed in labelled 25 cm cubic cages (Fig. 3.2b) for adult emergence. The remaining larvae were maintained in the basins filled with water, from the collection site, to a depth of about 2 cm. The larvae were fed once on finely ground Nutrafin goldfish food (Rolf Hagen, USA).

The larvae and pupae were reared to adults under conditions of 27-30°C and 76±2% relative humidity. The rooms had a 12 hr photoperiod. Each morning, the pupae were collected and placed in the appropriately labelled cages for adult emergence. The adults that emerged were transferred from the cages into small paper cups, using an aspirator, and killed by brief refrigeration at 4°C. Adult mosquito samples were preserved on silica gels at room temperature and a few in isopropanol at 4°C.

Fig. 3.2a. Plastic basins used for rearing mosquito larvae

Fig. 3.2b. Cages used for the rearing of pupae and the maintenance of emerged adult mosquitoes

3.4 Identification of *Anopheles* species

3.4.1 Morphological identification

Larvae were identified as *Anopheles* based on the fact that they lie parallel to the water surface and lack a siphon. Adult mosquitoes were identified as *Anopheles* using the markings on the palps, the banding and speckling on the legs and the characteristic pattern of blocks of dark and pale scales on the vein of the wings, especially along the costa (top part of the wing). *Anopheles gambiae* complex species were differentiated from *An. funestus* species using the morphological key of Gillies and de Meillon (1968) and Hervy *et al.* (1998). Briefly, the species of the *An. gambiae* complex have 5 pale spots on the costal margin of the wings, anal vein coloration with 3 white spots, a dark apical fringe and white speckled (or spots in the median part) tibia ornamentation. In contrast, *An. funestus* and other species have 4 pale spots on the costal margin on the wings, entirely dark anal vein coloration, and entirely dark tibia ornamentation.

3.4.2 Molecular identification of sibling species of the *Anopheles gambiae* complex

The PCR method of Scott *et al.* (1993) for the species identification of single specimen of the *An. gambiae* complex was used. The amplification process utilised one universal primer and four species-specific primers each of 20 bases (Table 3.1). The universal primer UN anneals to the same position on the rDNA of each the five species, but the reverse primer GA anneals specifically to *An. gambiae* s.s., ME to both *An. merus* and *An. melas*, AR to *An. arabiensis*, and QD to *An. quadriannulatus*. The sizes of the amplified products are 153 bp for *An. quadriannulatus*, 315 bp for *An. arabiensis*, 390 bp for *An. gambiae* s.s., and 464 or 466 bp for *An. merus* or *An. melas* respectively.

Table 3.1 Sequence details and melting temperatures (Tm) of oligonucleotide primers used for the PCR identification of the *An. gambiae* species complex (Scott *et al.*, 1993).

Primer name	Sequence (5' to 3')	Tm (°C)
UN	GTG TGC CCC TTC CTC GAT GT	58.3
GA	CTG GTT TGG TCG GCA CGT TT	59.3
ME	TGA CCA ACC CAC TCC CTT GA	57.2
AR	AAG TGT CCT TCT CCA TCC TA	47.4
QD	CAG ACC AAG ATG GTT AGT AT	42.7

3.4.2.1 Genomic DNA extraction

Each specimen was initially washed once in a solution of 400 µl EDTA buffer, 10 µl of 25% SDS and 5 µl of 10 mg/ml lysozyme, and incubated at 37°C, with agitation for 1 hr. This washing protocol was designed to wash away external sources of bacterial DNA by lysing bacterial cells on the exterior surface of the specimen. The specimen was then homogenised in a 1.5 ml Eppendorf tube containing 100 µl Bender buffer (preheated at 65°C) using a sterile polypropylene rod followed by incubation at 65°C for 30 min. Then 125 µl of buffer saturated phenol were added to the homogenate and mixed well by vortexing at 1000rpm and spun at 14 krpm for 10 min. The supernatant was transferred to a fresh tube, and 250 µl of chloroform added; it was vortexed briefly and spun at 14 krpm for 10 min. The supernatant was again transferred into a new tube and 250 µl of pre-chilled absolute ethanol and 10 µl of 8 M potassium acetate added, followed by incubation at -40°C for 1 hour. The DNA was pelleted by centrifugation at 10 krpm for 10 min and the supernatant was discarded. Two hundred micro litres of 70% ethanol were added to the pellet, the tube gently swirled and the DNA re-pelleted by centrifugation at 10 krpm for 5 min. The supernatant was discarded and the tube opened and inverted over a paper towel to dry. The dried DNA pellet was redissolved in 25 µl of TE + RNase (5µg/ml) and incubated on ice for about an hour; it was then stored at -20°C until needed.

3.4.2.2 PCR amplification

Each reaction mix of 20 µl contained 1x PCR buffer C (from the Invitrogen PCR Optimizer Kit), 200 µM each of the four deoxyribonucleotide triphosphates (dNTPs), 0.25 µM each of oligonucleotide primers and 0.5 U of DNA *Taq* polymerase enzyme. One microlitre of extracted mosquito DNA was used as template in the amplification reaction. The reaction was thoroughly mixed, centrifuged briefly at 10 krpm and overlaid

with 20 µl of mineral oil (nuclease free) to avoid evaporation and refluxing during thermocycling. The temperature profile for the reactions was 94°C for 3 minutes (initial melt) followed by 35 cycles of 94°C for 30 seconds (denaturation), 50°C for 30 seconds (annealing), 72°C for 1 minute (extension), and a final cycle of 72°C for 10 minutes. For each set of reactions, a negative control, which contained no DNA template was performed. The amplification reactions were carried out using a PCR Express Thermal Cycler (Hybaid Ltd., UK).

3.4.2.3 Analysis of PCR products

Eight microlitres of each PCR product were added to 1 µl of 10x bromophenol blue gel loading dye and electrophoresed in 2.0% agarose gels stained with 0.5 µg/ml ethidium bromide. The gels were prepared and run with 1 X TAE buffer, using either a midi- or maxi-gel system, at 100V for one hour and were visualised and photographed over a UVP dual intensity transilluminator at short wavelength using a Polaroid direct screen instant camera fitted with an orange filter, a hood and a Polaroid Type 667 film. The film was processed as recommended by the manufacturers (Polaroid Inc., USA). The sizes of the PCR products were estimated by comparison with the mobility of a standard 100 bp ladder (Sigma, USA).

3.5 Detection of Bacterial Symbionts Using PCR

To detect the symbiont's DNA sequences in the life stages of *An. gambiae* mosquitoes, PCR with general and specific eubacteria oligonucleotide primer sequences was used in the larvae, pupae and adult mosquito life stages. PCR reactions using general primers designed from sequences of bacterial 16S rDNA (fD1/rD1, fD1/rP1 and fD2/rP2) and 23S rDNA (rRNA F1 and rRNA R) were first carried out, and then another PCR with specific primers (WOLB16SF1/WOLB16SR1 and *ftsZ*1/*ftsZ*2) was carried out to determine if the symbionts detected were those of *Wolbachia* sp. The details of all the primer sequences are given in Table 3.2. The genomic DNA extracted from larvae, pupae and adult *An. gambiae* s.l. specimens (see Section 3.4.2.1) were used as template for the PCR.

The eubacteria 16S rDNA primer sets amplify sequences of a wide variety of bacteria taxa (Weisburg *et al.*, 1991). The 23S rDNA primers were used to confirm the initial results obtained with the 16S rDNA gene primers (Rousset *et al.*, 1992). Both primer sets amplify approximately 1.5 kb DNA fragments.

Each reaction mix of 25 µl contained 1x buffer C (from the Invitrogen PCR Optimizer Kit), 1.5 mM $MgCl_2$, 0.40 mM each of the 4 dNTPs, 0.40 µM of each oligonucleotide primer and 0.5 U of *Taq* polymerase enzyme. One microlitre of a single mosquito DNA extract was used as template in the amplification reaction. The reaction was thoroughly mixed, centrifuged briefly at 10 krpm and overlaid with about 20 µl of nuclease free mineral oil. The amplification reactions were carried out using the PCR Express Thermal Cycler (Hybaid Ltd., UK).

Table 3.2 Details of oligonucleotide primer sequences used for PCR detection of bacterial symbionts in *An. gambiae* s.l. and their melting temperatures.

Gene/Species	Primer name	Sequence (5' to 3')	Size (mers)	Tm (°C)
Eubacterial 16S rRNA	fD1	AGAGTTTGATCCTGGCTCAG	20	53
	rD1	AAGGAGGTGATCCAGCC	17	52
	rP1	ACGGTTACCTTGTTACGACTT	21	53
	fD2	AGAGTTTGATCATGGCTCAG	20	51
	rP2	ACGGCTACCTTGTTACGACTT	21	55
Eubacterial 23S rRNA	23S rRNA F1	CCGAATGGGGAAACCC	16	57
	23S rRNA R	CCACCTGTGTCGGTTT	16	49
Wolbachia sp	WOLB16SF1	AGTCCTGGCTAACTCCGTGCCA	22	64
	WOLB16SR1	TCACCCCAGTCACTGATCCCAC	22	62
	*ftsZ*1	GTATGCCGATTGCAGAGCTTG	21	59
	*ftsZ*2	GCCATGAGTATTCACTTGGCT	21	55

The optimised PCR conditions for the 16S rDNA were: an initial denaturation stage of 94°C for 3 min, followed by 35 cycles of 94°C for 60 s, 50°C for 60 s and 72°C for 2 min and ended with an additional elongation step of 72°C for 10 min. The same reaction conditions were used for the 23S rDNA amplifications, except that the annealing temperature was set at 47°C.

Samples which were positive for the presence of both 16S and 23S bacterial sequences were selected and used for another round of PCR, using previously described *Wolbachia* primers (Holden *et al.*, 1993; Wenseleers & Billen, 2000). The WOLB16SF1 (forward) and WOLB16SR1 (reverse) universal primer pair is highly specific for *Wolbachia*, and amplifies a 1000 bp of the small subunit (SSU) rDNA of all known *Wolbachia* strains, including the recently discovered C and D strains found in nematodes, but no other related bacteria (Wenseleers & Billen, 2000). A further PCR was carried out using the *ftsZ1*/*ftsZ2* primer set which is based on sequences in a gene coding for a protein that initiates cell division in prokaryotes and is broadly reactive with *Wolbachia* species in insects (Holden *et al.*, 1993). The approximate product size for this set of primers is 769 bp.

Each reaction mix of 20 µl contained 1 X PCR buffer C (from the Invitrogen PCR Optimizer Kit), 1.5 mM $MgCl_2$, 0.40 mM each of the four dNTPs, 0.40 µM of each oligonucleotide primer, 0.5 U of *Taq* polymerase enzyme and 0.5 µl of DNA as template. The reaction mix was treated as before and the reactions were carried out using the PCR Express Thermal Cycler (Hybaid Ltd., UK).

For the WOLB16SF1/ WOLB16SR1 primers, the temperature profile was as follows: an initial denaturation at 94°C for 3 min, followed by 35 cycles of 94°C for 60 s, 60°C for 60 s and 72°C for 2 min for 35 cycles, plus one additional cycle with a final 10 min chain elongation step. The same reaction conditions were used for the *ftsZ1*/*ftsZ2* primers, except for the annealing temperature which was set at 55°C.

Template DNA from locally caught brown-banded cockroach *Supella longipalpa* (Blattaria: Blattellidae) nymphs was used as positive control for the presence of *Wolbachia*. Negative controls, containing only double-distilled water, were also included to check for contamination. The PCR products were electrophoresed in a 1.0% agarose gel as described under section 3.4.2.3. The sizes of PCR products were estimated by comparison with the mobility of a standard 1 kb ladder.

3.5.1 Estimation of the concentration of PCR products

Five microlitres of PCR product was diluted to 95 µl with nuclease-free water and the absorbance reading at 260 nm wavelength (A_{260}) was recorded. The concentration of PCR product (X µg/µl) was then estimated by using the formula (Wilfinger *et al.*, 1997) below:

$$A_{260} \times \text{dilution factor} \times 50\ \mu g/\mu l = X\ \mu g/\mu l$$

3.6 Amplified Ribosomal DNA Restriction Analysis (ARDRA)

Restriction analysis of enzymatically amplified 16S rDNA can be used to identify the species of many bacteria genera. This exploits the differences in the DNA sequences that results in different restriction sites. In this variation of ribotyping, the ribosomal DNA is amplified and the product digested with a restriction enzyme, and the diagnostic DNA fragment profiles are visualized following separation by gel electrophoresis, avoiding the need for Southern blotting which involves cumbersome blotting techniques. This method has been used to differentiate bacterial species (Vaneechoutte, 1996; Dijkshoorn *et al.*, 1998).

3.6.1 *In silico* (computational molecular biology) restriction analysis

The most predominant bacteria species detected in *An. gambiae* have been *Escherichia coli* and *Pantoae agglomerans* (Straif *et al.*, 1998). In order to identify whether any of the amplicons might be either of these microorganisms, virtual restriction analyses were performed on their rDNA nucleotide sequences.

The complete 16S rDNA sequences of *E. coli* and *P. agglomerans* and the complete 23S rDNA sequence of *E. coli* were retrieved from the Ribosomal Database Project II (Cole *et al.*, 2003) and copied into the software "Jellyfish" version 1.1 (biowire.com). The FIND function of this programme was used to align the primer pairs used for the PCR amplifications to the homologous rDNA sequences, and the region that they flanked was copied, saved as a new file and then imported into the software programme DIGEST version 1.0 (Nakisa, 1992). *In silico* analysis was performed with the DIGEST software using the enzymes in its database WISCONSI.920. The output included the positions of the DNA sequences that each enzyme cut, the number of fragments and their sizes. It

also listed all enzymes that do not restrict. The results were compared to those obtained by the experimental restriction patterns of the amplified 16S and 23S rDNAs.

3.6.2 Restriction endonuclease digestion and analysis

From the results obtained from the *in silico* analyses, the enzymes that cut at less than six sites, and some which had no sites, were selected for this analyis. *Dra*I, *Eco*RI, *Hind*III, *Pvu*I, *Sal*I and *Xba*I were selected for the analysis of fD1/rD1, fD1/rP1, fd2/rP2 and 23S rRNA F1/R1 amplicons and *Apo*I, *Eco*RI, *Hind*III, *Rsa*I, *Nsp*I, *Hinf*I and *Bam*HI for the WOLB16SF1/ WOLB16SR1 amplicons. The digestions were carried out as recommended by the manufacturers (Sigma-Aldrich, USA). The final reaction volume of 20 µl contained 2-6 µl of the amplified products. The incubation was carried out at 37°C for 2 hours using a heat block. The products were electrophoresed in ethidium bromide-stained 2% agarose gels using 1X TAE buffer and the results were visualised by transillumination. A standard 100 bp ladder was used as molecular weight marker. The sizes of the digested products were calculated using the computer software DNAFRAG version 3.03 (Nash, 1992).

3.7 Cloning of Amplified Bacterial Sequences

The cloning of PCR products allows the generation of relatively large amounts of the amplified region, which can then be conveniently used, whenever needed, for experiments such as sequencing and hybridisation (Newton & Graham, 1997).

The cloning process involves the ligation, using DNA ligase, of the amplified DNA into a suitable plasmid vector to form a recombinant molecule, followed by transformation of the recombinant molecule into suitable competent bacteria cells for replication and multiplication. To select for transformants, "blue-white" colour screening, a well-established means for identifying a ligation product or indicating the presence of a DNA insert, is employed. The method is based upon the ability of β-galactosidase to hydrolyse 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside (X-gal), resulting in the characteristic blue staining of a colony or phage plaque (Horwitz *et al.*, 1964). The multiple cloning region of numerous plasmid and phage vectors is imbedded within the amino-terminal fragment of an isopropyl β-D-thiogalactoside (IPTG)-induced β-galactosidase gene. Successful ligation of a DNA fragment into the multiple cloning site disrupts β-galactosidase expression; without insertion, expression is uninterrupted. Thus, when the *E. coli* host, expressing the carboxyterminal portion of β-galactosidase, is transformed with a plasmid vector or infected with a phage without a DNA insert, a "blue" colony or plaque will result (Ullmann *et al.*, 1967). Transformation or infection of a vector with DNA fragment insertion results in a "white" or clear colony or plaque.

For this study, the TA cloning method was used to clone the amplified bacterial sequences. The TA cloning method takes advantage of the terminal transferase activity of some DNA polymerases such as *Taq* polymerase. This enzyme adds a single, 3'-A overhang to each end of the PCR product (Clark, 1988; Hu, 1993), which makes it

possible to clone it directly into a linearized cloning vector with single, 3'-T overhangs (TA cloning). DNA polymerases with proofreading activity, such as *Pfu* polymerase, cannot be used because they provide blunt-ended PCR products.

3.7.1 dA tailing of PCR products

To enable a high efficiency of cloning of the PCR product in the TA Cloning system (Invitrogen, USA), the addition of 3'A-overhangs post amplification was carried out. Twenty-five microlitre reactions each consisting of 19.75 µl of diluted PCR product (10 in 9.75 µl of water), 2.5 µl 10X PCR buffer C (from the Invitrogen PCR Optimizer Kit), 2.5 µl 2 mM dATP and 0.25 µl *Taq* were set up. Each reaction mix was vortexed, briefly centrifuged and incubated at 72°C for 10 minutes in a thermal cycler. The products were used immediately for the ligation reaction.

3.7.2 Ligation

To estimate the amount of PCR product needed to ligate with 50 ng (20 fmoles) of pCR®2.1 vector, the formula (TA Cloning Manual, Invitrogen, USA) below was used:

$$\text{X ng PCR product} = \frac{(\text{Y bp PCR product})(50 \text{ ng pCR}^{®}2.1 \text{ vector})}{(\text{size in bp of the pCR}^{®}2.1 \text{ vector: } \sim 3900)}$$

where X ng is the amount of PCR product of Y base pairs to be ligated in a 1:1 (vector: insert) molar ratio.

Using the concentration previously determined for the PCR product, (see section 3.5.1), the volume needed to give the required amount, as determined using this formula, was, if necessary, obtained by diluting with sterile water. One vial of pCR®2.1 vectors (Invitrogen, USA) was centrifuged briefly to collect all the liquid in the bottom of the vial. The ligation reaction mix was composed as follows: 1-3 µl dA-tailed PCR product,

1 µl of 10X Ligation Buffer, 2 µl pCR®2.1 vector (25 ng/µl), with sterile water to make up the volume to 9 µl and 1 µl T4 DNA Ligase (4.0 Weiss units). Each tube was vortexed and centrifuged briefly. The ligation mix was incubated at 14°C overnight. The products of the reactions that were not immediately used for transformation, were stored at -20°C until needed.

3.7.3 Transformation experiments

To perform transformation experiments, the temperature of the water bath was set at 42°C and the SOC medium at room temperature. The LB plates (two plates for each ligation/transformation), containing 100 mg/ml ampicillin, were opened and dried in a bacteriological hood under UV light at 37°C for 30 minutes. Each plate was labelled and spread with 40 µl of 40 mg/ml X-Gal and the liquid allowed to soak into the plates for about 15 minutes under sterile conditions. The transformation procedure was carried out following the instruction manual supplied with the TA Cloning® kit, with slight modifications.

The vials containing the ligation reactions were briefly centrifuged and placed on ice. One 50 µl vial of frozen One Shot® chemically competent *E. coli* cells for each ligation/ transformation was thawed on ice. Immediately after the cells had thawed, 2 µl of each ligation reaction was directly pipetted into the vial of competent cells and mixed by gently tapping the base of the vials. They were then incubated on ice for 30 minutes and then heat-shocked at 42°C in a water bath for exactly 30 seconds, avoiding any shaking and mixing. The vials were then removed from the water bath and placed on ice for 2 minutes. Two hundred and fifty microlitres of the SOC medium were added to each vial and the vials shaken horizontally at 37°C for exactly 1 hour at 225 rpm in a shaking incubator. The vials were next removed, placed on ice for a few minutes and centrifuged

for a minute at 10 krpm. The supernatants were discarded and the pellets resuspended in 100 µl of SOC medium. Five microlitres and 95 µl of cells from each transformation vial were spread separately on the LB agar plates. The plates were then incubated at 37°C for at least 18 hours and transferred to a +4°C environment for 2-3 hours for colour development.

To determine the presence of inserts, 10-15 white colonies were randomly picked from selected plates for plasmid isolation. The colonies were grown overnight in 5 ml LB broth containing 50 µg/ml of ampicillin in a rotary shaker at 37°C. Then, 100 µl of the broth were removed and placed in 0.5 ml microfuge tubes and heated at 100°C for 5 minutes in a thermal cycler and placed immediately on ice. The tubes were spun at 10 krpm in a microfuge for 10 minutes and the supernatant used for PCR.

Each PCR reaction mix consisted of 16.05 µl PCR water, 2.5 µl 10x PCR buffer C (from the Invitrogen PCR Optimizer Kit), 2.5 µl 20 mM dNTP mix, 0.625 µl each of the 20 µM Universal M13 Forward (-40) and reverse primers, 0.2 µl *Taq* polymerase (5U/µl) and 0.25 µl of supernatant from the cell preparations. These were thoroughly mixed, centrifuged briefly and overlaid with 50 µl of mineral oil. The PCR reaction cycling profile was 30 cycles of 1 minute at 95°C, 1 minute at 55°C and 1 minute at 72°C, and ended with 7 minutes at 72°C. Ten microlitres of each reaction was electrophoresed in 1% agarose gels and visualised as described before (see section 3.4.2.3). A PCR product band size of approximately a 1000 bp was observed for vectors with insert.

3.7.4 Plasmid DNA isolation

Positive colonies were processed for plasmid DNA isolation. One microlitre of cell culture was pelleted by centrifugation at 10krpm for 2 minutes, and DNA from the cells,

extracted using the QIAprep Spin Miniprep Kit (QIAGEN, USA) protocol as described below. The buffers and reagents used were part of the kit supplied by the manufacturer.

The supernatants were discarded and the pelleted bacterial cells resuspended in 250 µl Buffer P1 containing RNase, making sure that no cell clumps were visible after resuspension. Two hundred and fifty microlitres of Buffer P2 was added and the suspension mixed well by gently inverting the tube 4–6 times. When necessary, the inversion of the tube was continued until the solution became viscous and slightly clear but this was not allowed to proceed for more than 5 min. Then, 350 µl of Buffer N3 was added and the tube inverted immediately but gently 4–6 times to avoid localized precipitation, followed by centrifugation at 10 krpm for 10 min.

The supernatants obtained were transferred onto a QIAprep column and then centrifuged for 1 minute and the flowthrough discarded. The QIAprep spin column was washed with 0.5 ml Buffer PB and centrifuged for 1 minute and the flowthrough again discarded. The QIAprep spin column was again washed with 0.75 ml Buffer PE and centrifuged for another 1 minute and the flowthrough treated as before. An additional 1 minute centrifugation at the same speed was done to remove residual wash buffer. To elute DNA, the QIAprep column was placed in a clean 1.5 ml microcentrifuge tube and 50 µl Buffer EB (10 mM Tris-Cl, pH 8.5) added to the centre of each QIAprep column, left to stand for 1 min, and centrifuged for 1 min. The isolated plasmid DNA was checked after a 0.8% agarose gel electrophoresis and visualised under UV light. The expected band size of the isolated plasmid was approximately 5 kb.

3.8 Identification of Amplified Bacterial Symbiont Sequences and Phylogenetic Analysis

3.8.1 Sequencing of amplified bacterial DNA sequences

The sequencing of the insert DNA in the pCR®2.1 vector was carried out bi-directionally. The M13 reverse primer was used to sequence from the *lac* promoter end of the vector and the M13 forward (-20) Primer from the *lacZá* end (Fig. 3.3).

Five microlitres of plasmid DNA was diluted to 80 µl with nuclease-free water and the concentration of the plasmid DNA was determined by absorbance reading at 260 nm wavelength (A_{260}). The concentration of plasmid DNA (Y µg/µl) was estimated by using the formula (Wilfinger *et al*., 1997) below:

$$A_{260} \text{ x dilution factor x } 50 \text{ µg/µl} = Y \text{ µg/µl}$$

The sequencing reaction mix contained plasmid DNA (final concentration of 1 µg/µl), 0.65 µl of 5 µM each of either the forward or reverse M13 primer and nuclease-free water to a final volume of 9 µl and the reaction carried out using the ABI PRISM Big Dye Terminator Cycle Sequencing Ready Reaction Kit (PE-Applied Biosystems, USA) and analysed with the Perkin Elmer ABI 377 automated sequence analyser (Applied Biosystems, USA). Raw chromatogram reads were base-called using PHRED (Ewing *et al*., 1998).

3.8.1.1 Sequence editing

The sequence data obtained were copied into the MacDNASIS (version 3.2) software package (Hitachi Software Engineering Co., Ltd., Japan) and manually edited to remove the *Eco*RI sites and the "GGTCC" sequences that flank the inserted PCR amplified fragment on each side of the sequenced primer ends. The complete sequence of the PCR product was generated from the "shotgun" fragments using the AUTO CONNECTION

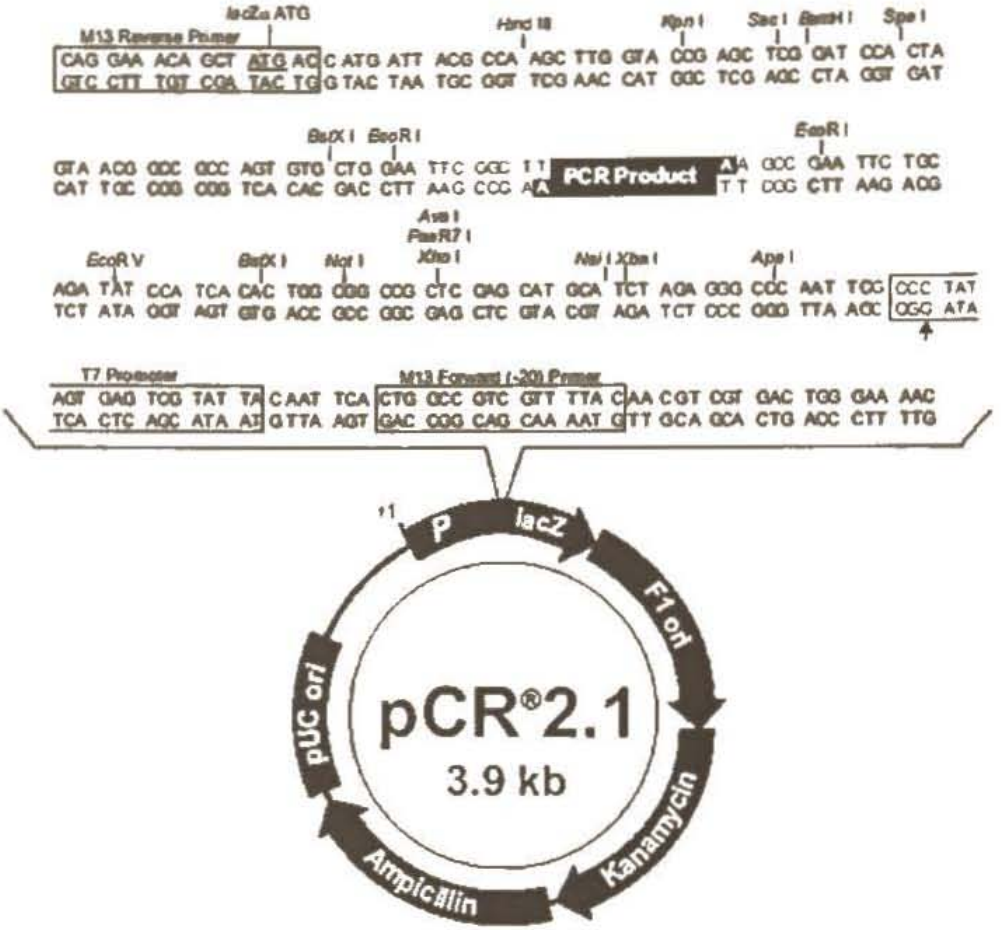


Fig. 3.3 Map of the linearized vector pCR®2.1. The sequence of the multiple cloning site is shown with a PCR product inserted by TA Cloning®. *Eco*R I sites flank the inserted PCR product on each side. The arrow indicates the start of transcription for the T7 RNA polymerase.

option under the CONTIG manager function of the software package. An overlap minimum size of 20 bp and a minimum matching percentage of 60% were selected.

3.8.1.2 VecScreen and chimera detection

Prior to sequence analysis, sequences were first screened to reveal and remove any of the vectors, using the VecScreen programme from the BLAST webpage of the National Center for Biotechnology Information (NCBI) website (Altschul *et al.*, 1990, 1997). VecScreen searches a query for segments that match any sequence in a specialized non-redundant vector database. The search uses BLAST parameters preset for optimal detection of vector contamination. Those segments of the query that match vector sequences are categorised according to the strength of the match, and their locations are displayed.

Several studies have indicated that a significant fraction of the 16S rDNA sequences obtained by PCR amplification of genomic DNAs extracted from environmental samples could be recombinants or chimeric (as a consequence of PCR coamplification from mixed genomes (Liesack *et al.*, 1991; Choi *et al.*, 1994; Kopczynski *et al* , 1994; Wang & Wang, 1996). Chimera formation is thought to occur when a prematurely terminated amplicon re-anneals to a foreign DNA strand and is copied to completion in the following PCR cycles. This results in a sequence composed of two or more phylogenetically distinct parent sequences and, when comparatively analysed with other 16S rDNA sequences, suggests the presence of a non-existent organism (Hugenholtz & Huber, 2003). To check for possible chimeric sequences created during amplification, the CHECK_CHIMERA software version 2.7 of the Ribosomal Database Project II (Maidak *et al.*, 2001) was used. This program can help to determine if a sequence is composed of two fragments that are

similar to clearly different sequences from the database and uploaded user sequences, i.e. if the sequence is of chimeric origin. The default settings for the program were used.

3.8.1.3 Sequence similarity and DNA database searches

After editing, the sequences were submitted to the standard nucleotide-nucleotide BLAST database search programme (blastn 2.2.1) of the NCBI website (Altschul *et* al., 1990; Altschul *et* al., 1997) to search for closely related sequences in the non-redundant GenBank database. The default parameter settings were used. Related sequences were acquired by using the Batch Entrez programme (NCBI).

Percent similarity between the sequences obtained was determined by using the maximum matching option under the compare function in the DNASIS software package. This function compares two DNA/amino acid sequences and aligns them for maximum homology. Homology plots were also constructed using the Homology option under the compare function of the software.

3.8.1.4 Construction of consensus sequence

Based on the results of the homology plots, sequences showing over 80% similarity to one another were grouped and aligned to generate a consensus sequence using the multiple alignment function in DNASIS. The automatic Higgins option was used with the following multiple alignment parameters: gap penalty of 300, fixed gap penalty of 20, floating gap penalty of 20, k-tuple value of 2, window size value of 5 and number of top diagonal value of 5. The alignments were completed manually and the consensus sequences generated subjected to BLAST searches as described previously.

3.8.2 Phylogeny construction

Sequences homologous to the consensus sequences were retrieved from the DNA databanks for comparisons. For sequences of groups of nearly identical groups only single representatives were used for phylogenetic analysis. Reference 16S rDNAs used in the analyses were obtained from the GenBank and are shown in Tables 3.3a, 3.3b and 3.3c.

The sequence data were aligned using the CLUSTAL W package (Thompson *et al.*, 1994) integrated into the MacVector 7.1 software suite (Accelerys, Pharmacopeia Inc., USA). For both pairwise and multiple alignments, the open gap penalty of 30 and extended gap penalty of 10 were applied. A slow alignment mode was selected for pairwise alignments. For multiple alignments, a delay divergence of 60% was selected and the transitions were weighted.

The regions of the DNA sequences that aligned ambiguously (base positions which were of indeterminate identity, insertions and deletions, and alignment gaps), most of which corresponded to variable regions of SSU rRNA, were not used for the subsequent determination of phylogenetic relationships. A number of algorithms, including neighbour-joining, maximum likelihood and maximum parsimony, for inferring phylogeny were used to evaluate the sequence relationships.

Table 3.3a Bacteria species and strains used for the phylogenetic analysis of CONSEN4.

GenBank accession number	Organism
X78717	Purple bacterium (unnamed)
AB041770	*Paracoccus kawasakiensis*
Y16930	*P. denitrificans*
D16429	*Rhodobacter blastica*
Y12703	*P. marcusii*
AB006899	*P. carotinifaciens*
AF365994	Marine alpha proteobacterium BBAT3
X53855	*R. sphaeroides*
AB017797	*Rhodobacter sp.* TCRI 5
D16427	*R. capsulatus*
AY014179	*P. yeeii strain G3060*
D70847	*R. azotoformans*
D32238	*P. alkaliphilus*
AY005463	*Ruegeria* sp
D32241	*P. kocurii*
AF245634	Uncultured *Roseobacter* sp.
X78315	*R. algocolus*
D88523	*Agrobacterium gelatinovorum*
D32240	*P. aminovorans*

Table 3.3b Bacteria species and strains used for the phylogenetic analysis of CONSEN2.

GenBank accession number	**Organism**
AJ440751.1	Gram-negative bacterium MM 1
AF502217.1	Uncultured bacterium clone HP1B64
X78717.1	Purple bacterium (unnamed) strain SW2
Y16930.1	*Paracoccus denitrificans*
AF229874.1	*Paracoccus* sp. 4FB8
D32238.1	*P. alkaliphilus*
AY014176.1	*P. aminophilus*
AF527586.1	Uncultured bacterium clone LPB54
AF136850.1	*Ketogulonogenium robustum*
AY014173.1	*P. yeeii* strain G1212
AF445668.1	Uncultured alpha proteobacterium clone SM1C11
AF368183.1	Uncultured *Rhodobacter* group bacterium clone SBRT155
AB017799.1	*Rhodobacter* sp. TCRI

Table 3.3c Bacteria species and strains used for the phylogenetic analysis of AgL5.

GenBank accession number	Organism
Y18833	*Hymenobacter roseosalivarius*
AY279405	Endosymbiont of *Aphytis sp*
AY279402	Endosymbiont of *Aspidiotus nerii*
AY274139.1	Uncultured bacterium clone D133
AF382107.1	Uncultured bacterium clone ZA2626c
AF336356.1	Bacterium Wuba47
D12658.1	*Cytophaga aurantiaca*
AY038780.1	Uncultured CFB group bacterium clone TAF-B66
AJ244689.1	*Cyclobacterium* sp. V4.MS.32
AF361200.1	Uncultured *Cytophagales* bacterium clone 30
AJ011917	*Flectobacillus sp.*
AF337883.2	Uncultured gold mine bacterium D28
AY279404	Endosymbiont of *Aphytis lingnanensis*
AF408275.1	Uncultured *Hymenobacter* sp. clone KL-2-4-9
AJ400340.1	Uncultured marine bacterium ZD0203
Y18835	*Taxeobacter ocellatus*

3.8.2.1 Neighbour-joining (NJ) analysis

The neighbour joining (NJ) programme in the PHYLIP package version 3.5c (Felsenstein, 1993) for inferring phylogenetic relationships, also contains programs carrying out other related tasks using different algorithms on different kinds of data.

SEQBOOT was first used to bootstrap the sequences with 1000 replicates (Felsenstein, 1985). DNADIST was then used to calculate evolutionary distances from the bootstrapped output data with the Kimura two-parameter model for nucleotide change (Kimura, 1980) using a transition-transversion ratio, estimated from the data with the software PUZZLE 4.0.2 (Strimmer & von Haeseler, 1996) and one category of substitution rates. Phylogenetic trees were constructed from the evolutionary sequence data using NEIGHBOR (Saitou and Nei, 1987). A consensus tree was constructed using CONSENSE with the treefile from NEIGHBOR.

3.8.2.2 Maximum likelihood analysis (ML)

Phylogenetic trees were also derived from the molecular sequence data by the maximum likelihood (ML) method carried out using the quartet puzzling (QP) analysis with the programme PUZZLE 4.0.2. PUZZLE is ANSI C compliant and a PHYLIP compatible program that implements the quartet puzzling method for reconstructing tree topologies from character state data. Quartet puzzling is a method that applies maximum likelihood tree reconstruction to all possible quartets of taxa and subsequently tries to combine most of the four-taxa maximum likelihood trees to construct an overall tree.

A 1000 puzzling steps using the Hasegawa, Kishino, and Yano (HKY) substitution model (Hasegawa *et al*., 1985) with an estimated transition-transversion ratio, and site-to-site substitution rate variation modelled on a gamma distribution with four categories and the

shape parameter were estimated from the data. For parameter estimation, quartet sampling (for substitution process) plus NJ tree (for rate variation) were used.

3.8.2.3 Maximum parsimony analysis (MP)

A MP tree requires the smallest number of evolutionary changes to result in the set of OTUs under the study (Hillis, 1997). For each tree to be evaluated, the minimum possible number of changes for each 'character' (nucleotide position or morphological trait) is calculated, and the minimum number of changes across all characters are totalled, to obtain the parsimony score. The best tree is the one that requires the fewest changes across all characters (Hillis, 1997). Information on evolutionary processes may be incorporated by weighting characters differentially (such as first versus third positions of codons), or by weighting character state changes differentially (such as transitions versus transversions).

Maximum parsimony (MP) analysis was accomplished with the Phylogenetic analysis using parsimony (PAUP) version 3.0r program (Swofford, 1991). A stepmatrix was used to give transversions twice the weight of transitions, and gaps were treated as missing data. Two heuristic search strategies were used for each analysis, with addseq = simple, and the second with addseq = random and 10 replicates. Branch swapping was by tree bisection-reconnection.

Consensus (50% majority rule) trees were constructed by using uncorrected NJ distances with 1,000 bootstrap replicates. These trees excluded groupings that occurred in less than 50% of the replicates and negative branch lengths were prohibited.

3.8.3 Nucleotide Sequence Accession Numbers

The amplified 16S rDNA sequences of the bacteria from the mosquitoes, after identification, were submitted to the DNA Databank of Japan (DDJB), European Molecular Biology Laboratory (EMBL) and GenBank (at NCBI) nucleotide sequence databases.

CHAPTER FOUR

RESULTS

4.1 Mosquito Species Identification

Genomic DNAs for PCR were extracted from 432 mosquito specimens comprising of 91 larvae, 37 pupae and 304 adult females (Fig 4.1) and these were used for the studies.

Out of the 432 mosquito specimens processed, PCR amplification for species identification was successful for 395 (91.4%) while amplification failed for 37 (8.6%). Of the 395 PCR positive specimens, 373 (94.4%) were identified as *An gambiae* s.s., 20 (5.1%) as *An. arabiensis*, 1 (0.25%) as *An. merus* and 1 (0.25%) *An. melas* (Fig. 4.2 and Table 4.1). The 395 mosquito specimens comprised 295 (74.7%) adults, 29 (7.3%) pupae and 71 (18%) larvae. The details of the species identified and of the life-stage forms are also given in Table 4.1.

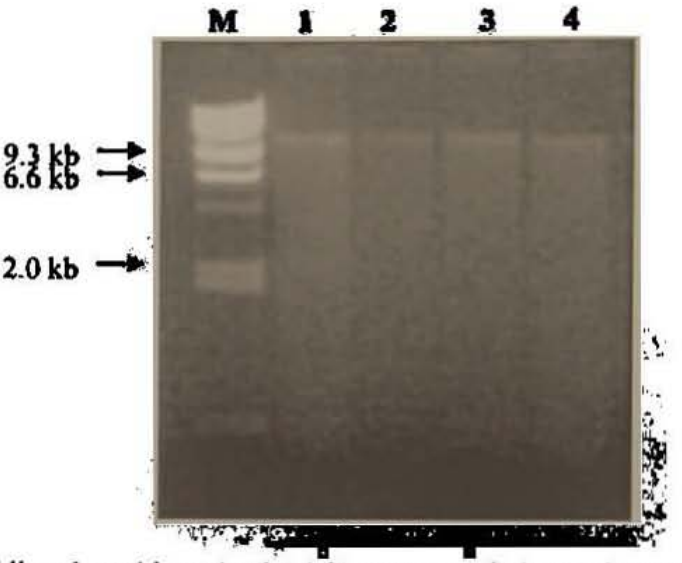


Fig 4.1 Ethidium bromide-stained 0.8 % agarose gel electrophoregram of genomic DNA extracted from *An. gambiae* s.l. mosquitoes. Lane M = λ DNA - *Hind* III digest; lane 1 = adult; lanes 2 and 4= larvae; lane 3 = pupa.

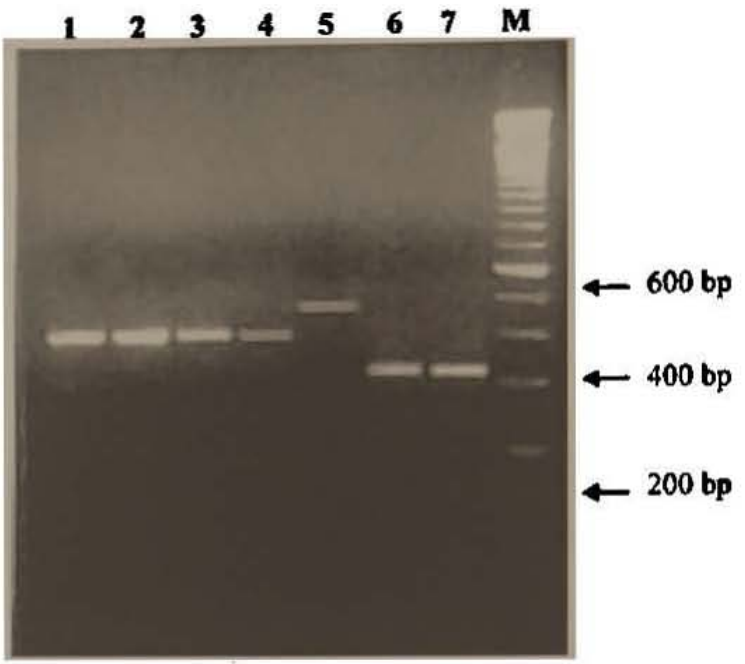


Fig 4.2 Ethidium bromide-stained 2.0 % agarose gel electrophoregram of DNA bands produced by the rDNA-PCR identification method for members of the *An. gambiae* complex. Lanes 1 and 2 = *An. gambiae* s.s adults; lane 3 = *An. gambiae* s.s pupa; lane 4 = *An. gambiae* s.s larva; lane 5 = *An. merus* adult; lanes 6 and 7 = *An. arabiensis* adults; lane M = 100 bp ladder (Sigma-Aldrich, USA)

Table 4.1 Distribution of species identified among the different life stage forms of *An. gambiae* complex species using the PCR method of Scott *et al.* (1993).

Life-stage	Number of *An. gambiae* complex species			
	gambiae s.s	*arabiensis*	*merus*	*melas*
Larvae	71	0	0	0
Pupae	28	0	0	1
Adults	274	20	1	0
TOTAL (%)	**373 (94.4)**	**20 (5.1)**	**1 (0.25)**	**1 (0.25)**

4.2 PCR Detection of Bacteria DNA sequences in *An. gambiae* s.l.

The 395 *An. gambiae* s.l specimens were examined for the presence of symbiotic bacteria using the eubacteria 16S and 23S primers. Out of the total of 395 specimens studied, DNA fragments of the predicted sizes were successfully amplified in 336 (85.1%) and 314 (79.5%) of them for the 16S rDNA and 23S rDNA primers, respectively (Table 4.2). Bacterial DNA sequences were amplified from all the sibling species used and also from larvae, pupae and adult mosquito specimens (Figs. 4.3 and 4.4). Bacterial DNA sequences were also amplified from both wild and laboratory reared adult mosquitoes. The PCR amplifications that were positive for both 16S and 23S rRNA primers produced single DNA bands in all cases and negative controls with no DNA template consistently gave no amplification product. Bacteria DNA sequences were amplified from specimens from all the mosquito collection sites.

A total of 281 (71.1%) PCR positive specimens for both 16S and 23S rDNA primers, were selected and screened for *Wolbachia* using the WOL16S rDNA and *ftsZ* primers. Using the WOL16S rDNA primers, 94 (33.5%) specimens yielded amplified DNA product of the expected 1.0 kb size (Fig. 4.5). The DNA fragments were amplified from larvae, pupae and adult mosquito specimens (Table 4.3). The PCR amplifications that were positive for the WOL16S rDNA primers produced single DNA bands in all cases and negative controls with no DNA template consistently gave no amplification product. None of the mosquito specimens, however, amplified DNA with the *ftsZ* primer pair although all amplifications of the positive control were successful (Fig. 4.6).

Table 4.2. PCR amplification of bacteria 16S and 23S rDNA gene sequences in *An. gambiae* s.l. mosquitoes.

Life stage	Number	Number (%) positive for	
		16S rDNA	23S rDNA
Larvae	71	54 (76.1)	50 (70.4)
Pupae	29	26 (89.7)	26 (89.7)
Adults	295	256 (86.8)	238 (80.7)
Total	**395**	**336 (85.1)**	**314 (79.5)**

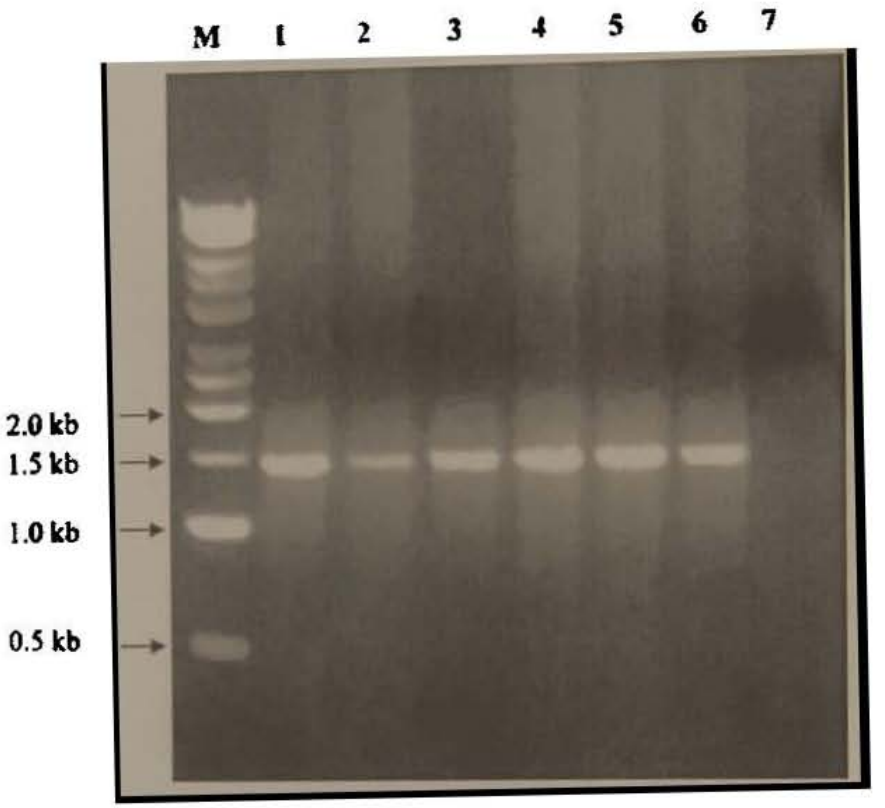


Fig. 4.3 Ethidium bromide-stained 1.0% agarose gel electrophoregram of amplified bacteria 16S rDNA sequences from mosquito specimens using rD2/rP2 primers. Lane M = 1 kb ladder (MBI Fermentas, Germany); lane 1 = *An. gambiae* s.s. larva; lane 2 = *An. gambiae* s.s. larva; lane 3 = *An gambiae* s.s. pupa; lane 4 = *An gambiae* s.s. adult; lane 5 = *An. arabiensis* adult; lane 6 = positive control (cockroach nymph: *Supella longipalpus*); lane 7 = negative control.

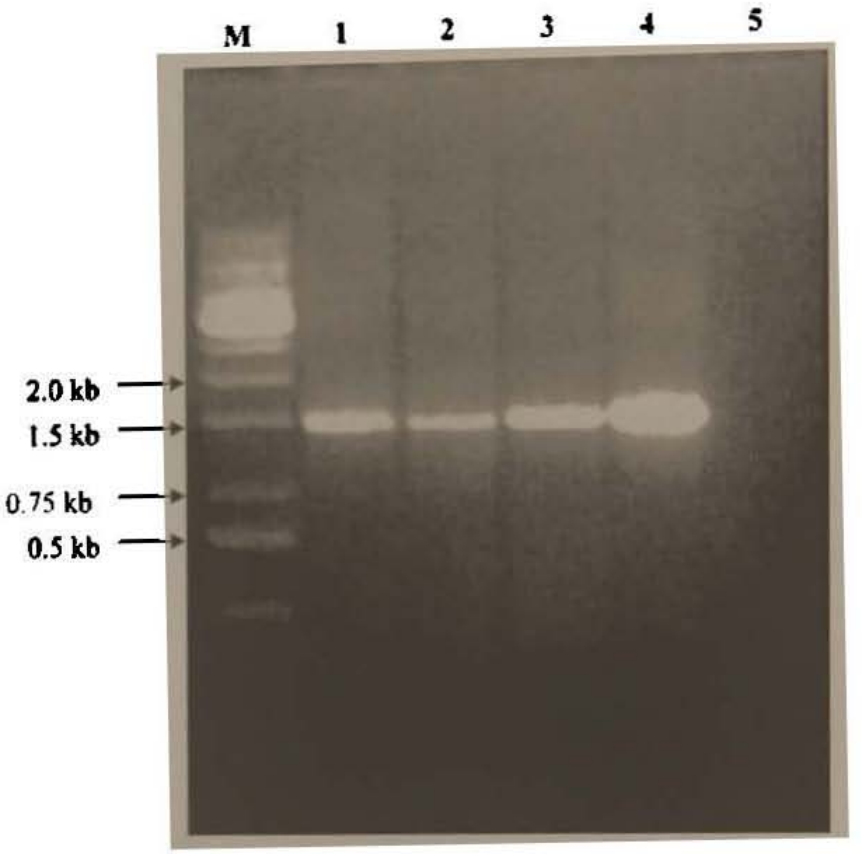


Fig. 4.4 Ethidium bromide-stained 1.0% agarose gel electrophoregram of amplified bacteria 23S rDNA sequences from mosquito specimens using WA1/WA3 primers. Lane M, 1 kb ladder (MBI Fermentas, Germany); lane 1 = *An. gambiae* s.s. larva; lane 2 = *An. gambiae* s.s. pupae; lane 3 = *An. gambiae* s.s. adult; lane 4 = positive control (cockroach nymph: ***Supella longipalpus***); lane 5 = negative control

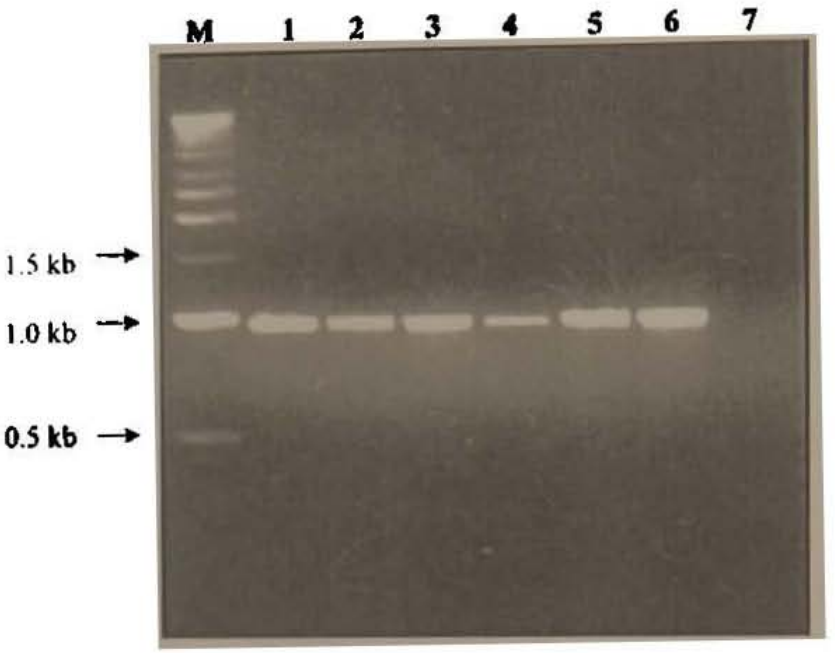


Fig. 4.5 Ethidium bromide-stained 1.0% agarose gel electrophoregram of amplified bacteria 16S rDNA sequences from mosquito specimens using WOL16S primers. Lane M = 1 kb ladder (Sigma-Aldrich, USA); lane 1 = positive control (cockroach nymph: *Supella longipalpus*); lane 2 = *An. gambiae* s.s. larva; lane 3 = *An. gambiae* s.s. pupa; lane 4 = *An. gambiae* s.s. adult; lane 5 = *An. gambiae* s.s. adult; lane 6 = *An. arabiensis* adult; lane 7 = negative control

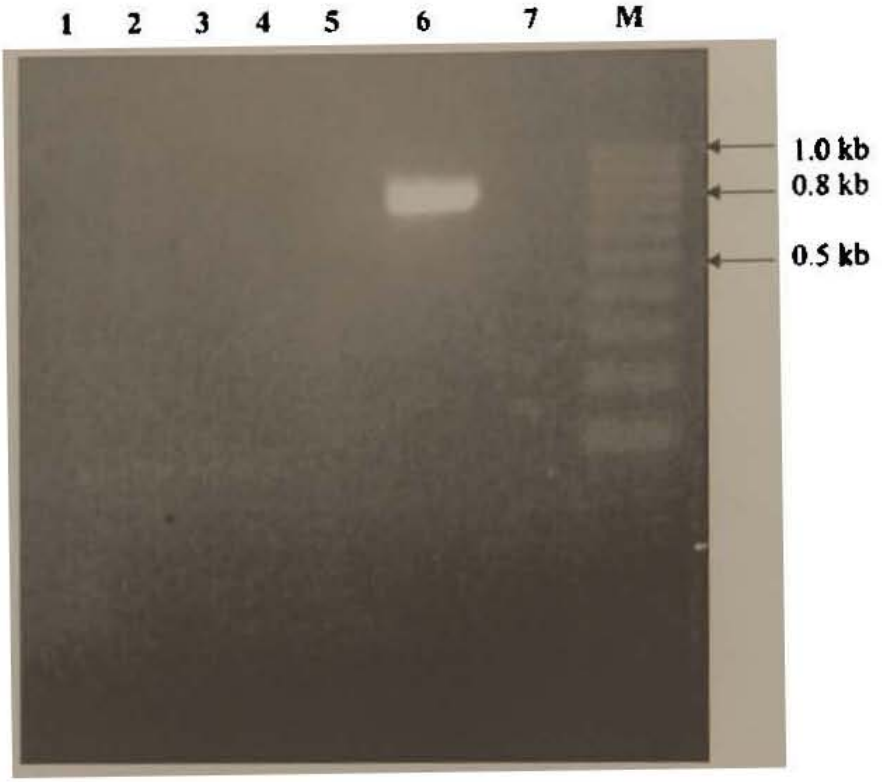


Fig. 4.6 Ethidium bromide-stained 1.0% agarose gel electrophoregram of amplified bacteria 16S rDNA sequences using *ftsZ* primers. Lane 1 = *An. gambiae* s.s. larva; lane 2 = *An. gambiae* s.s. pupa; lane 3 = *An. gambiae* s.s. adult; lane 4 = *An. gambiae* s.s. adult; lane 5 = *An. arabiensis* adult; lane 6 = positive control (cockroach nymph: *Supella longipalpus*); lane 7 = negative control; lane M = 100 bp molecular weight ladder (Gibco BRL, UK)

Table 4.3 PCR amplification of *Wolbachia ftsZ* and 16S rDNA gene sequences in specimens of *An. gambiae* s.l. mosquitoes.

Life stage	Number	Number positive for	
		ftsZ(%)	WOL16S (%)
Larvae	40	0 (0)	23 (57.5)
Pupae	20	0 (0)	2 (10.0)
Adults	221	0 (0)	69 (31.2)
Total	**281**	**0 (0)**	**94 (33.5)**

4.3 Amplified Ribosomal DNA Restriction Analysis (ARDRA)

All the amplified 16S and 23S rDNA amplicons were analysed using six restriction enzymes *Dra*I, *Eco*RI, *Hind*III, *Pvu*I, *Sal*I and *Xba*I. For all the amplicons analysed as such, no difference in restriction patterns was observed. With the exception of the *Eco*RI restriction enzyme, none of the other enzymes cleaved the 16S rDNA amplicons (Fig 4.7). The *Eco*RI enzyme cleaved the 16S rDNA amplicons into 2 fragments which were approximately 880 and 680 bp in sizes.

The 23S rDNA amplicons were cleaved by only *Dra*I and *Sal*I. Both enzymes cleaved to give two fragments which were 1020 bp and 540 bp for *Dra*I, and 1260 bp and 340 bp for *Sal*I (Fig. 4.8). The sizes of the restricted fragments obtained with each of the 6 enzymes are given in Table 4.4.

Comparison of the results of the experimental restriction patterns of 16S and 23S rDNA amplicons and those obtained from the *in silico* analyses for *E. coli* and *P. agglomerans* revealed that none of the products could be that of either *E. coli* or *P. agglomerans* (Table 4.4). For example, both the 16S rDNA amplicons and the homologous *E. coli* and *P. agglomerans* 16S rDNA sequences were cleaved by *Eco*RI, but only *E. coli* possessed restriction sites for *Hind*III. Furthermore, none of the amplicons was cleaved with *Sal*I in contrast to both *E. coli* and *P. agglomerans*. For the 23S rDNA, the amplicons and *E. coli* sequences possessed sites for both *Dra*I and *Sal*I but only *E. coli* rDNA was cleaved by *Eco*RI. Moreover, the sizes of the restriction fragments differed in the case of the *Sal*I digest which for the amplicons resulted in fragments sizes 1260 bp and 340 bp, whilst for *E. coli* they were 1340 and 160 bp.

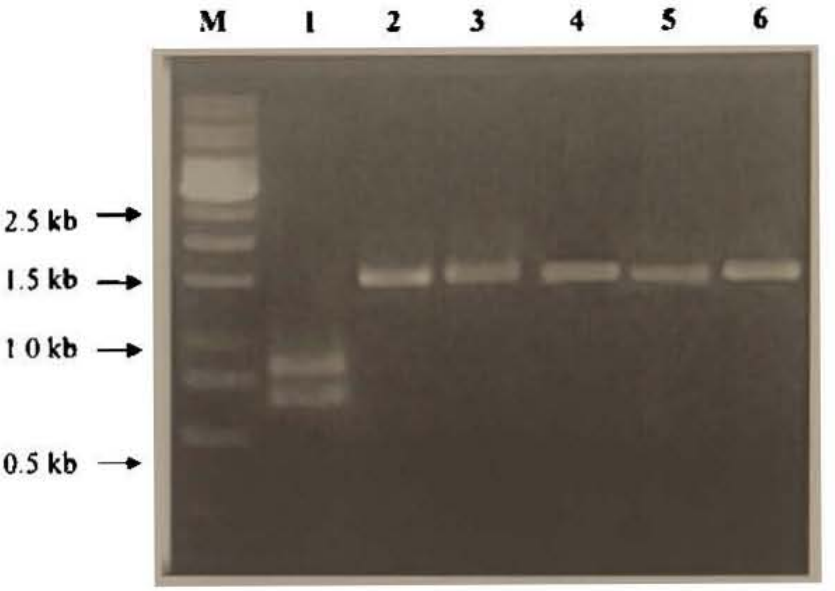


Fig 4.7 Ethidium bromide-stained 2.0% agarose gel electrophoregram of restriction enzyme digests of 16S rDNA PCR products. Lane M = 1 kb molecular weight marker (MBI Fermentas, Germany); lane 1 = *Eco*RI; lane 2 = *Hind*III; lane 3 = *Sal*I; lane 4 = *Dra*I; lane 5 = *Pvu*I; lane 6 = *Xba*I

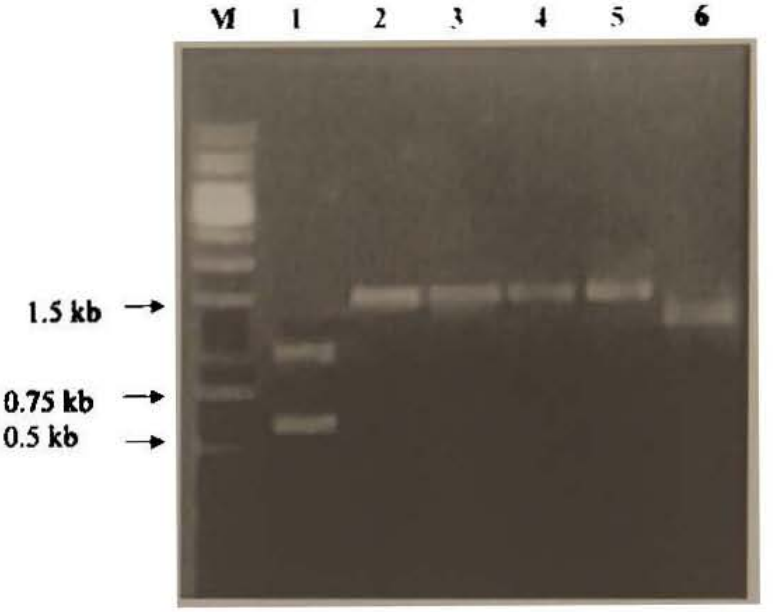


Fig 4.8 Ethidium bromide-stained 2.0% agarose gel electrophoregram of restriction enzyme digests of 23S rDNA PCR products. Lane M = 1 kb molecular weight marker (MBI Fermentas, Germany); lane 1 = *Dra*I; lane 2 = *Pvu*I; lane 3 = *Xba*I; lane 4 = *Eco*RI; lane 5 = *Hind*III; lane 6 = *Sal*I

Table 4.4 Observed and expected fragment sizes for amplicons, *E. coli* and *P. agglomerans* rDNAs after digestion with the various restriction enzymes (- denotes no restriction)

Restriction enzymes	Restriction fragment sizes (bp) for 16S rDNA primers			Restriction fragment sizes (bp) for 23S rDNA primers	
	Amplicons	*E. coli**	P. agglomerans*	Amplicons	*E. coli**
Dra I	-	-	-	1020 540	1080 420
Eco RI	880 680	870 670	836 659	-	840 670
Hind III	-	890 570 80	-	-	-
Pvu I	-	-	-	-	-
Sal I	-	820 670 50	806 636 53	1260 340	1340 160
Xba I	-	-	-	-	-

*Restriction fragments sizes shown are those obtained from the *in silico* analyses using the software DIGEST.

All the amplicons obtained using the WOL16S primers were digested with *Apo*I, *Eco*RI, *Hind*III, *Rsa*I, *Nsp*I, *Hinf*I and *Bam*HI. None of the amplicons was cleaved by the *Hind*III restriction enzyme (Fig. 4.9). *Nsp*I *Rsa*I and *Apo*I, each cut once to produce the same two fragments in all the amplicons analysed. The restriction fragments sizes were estimated to be 800 and 180 bp for *Apo*I (Fig. 4.10), 524 and 418 bp for *Rsa*I (Fig. 4.11) and 585 and 333 bp for *Nsp*I (Fig. 4. 11).

Restriction with *Eco*RI, *Bam*HI and *Hinf*I gave multiple banding patterns. Complex patterns, with the sum of the restriction bands being greater than the size of undigested PCR-amplified WOL16S rDNAs were obtained with the *Eco*RI and *Hinf*I. The *Eco*RI digestion (Fig. 4.12) yielded 3 fragments of sizes 1039, 590 and 436 bp and 4 fragments of sizes 1039, 736, 590 and 436 bp. The *Hinf*I digestion (Fig. 4.13) yielded fragments of 871 bp, 871 and 693 bp and 800 bp in sizes. For the *Bam*HI restriction, 3 distinct restriction patterns with estimated fragment sizes of 746, 668, and 693 bp respectively were observed (Fig. 4.13).

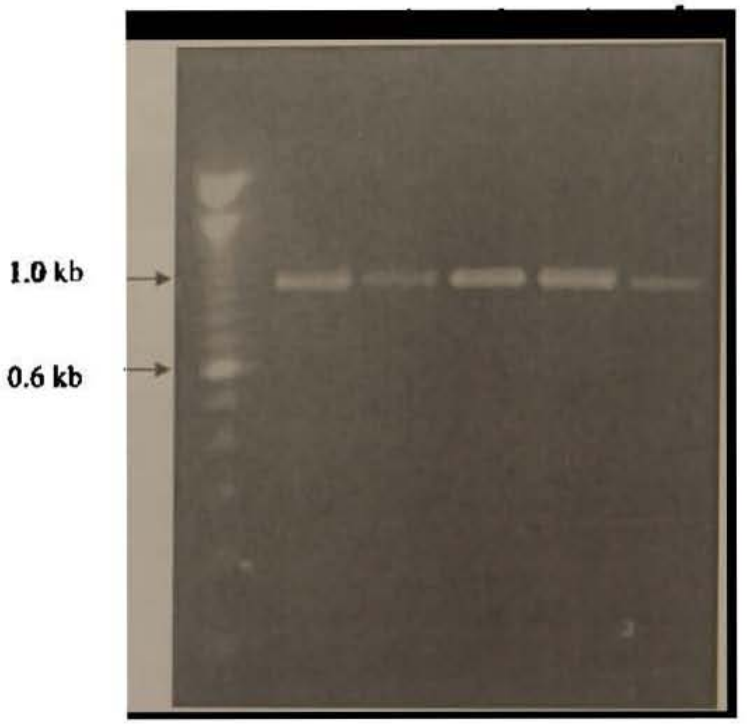


Fig 4.9 Ethidium bromide-stained 2.0% agarose gel electrophoregram of *Hind*III restriction enzyme digests of WOL16S rDNA PCR products. Lane M = 100 bp molecular weight marker (Sigma-Aldrich, USA); lanes 1 – 4 = digested bacterial PCR products from different mosquitoes; lane 5 = undigested PCR product.

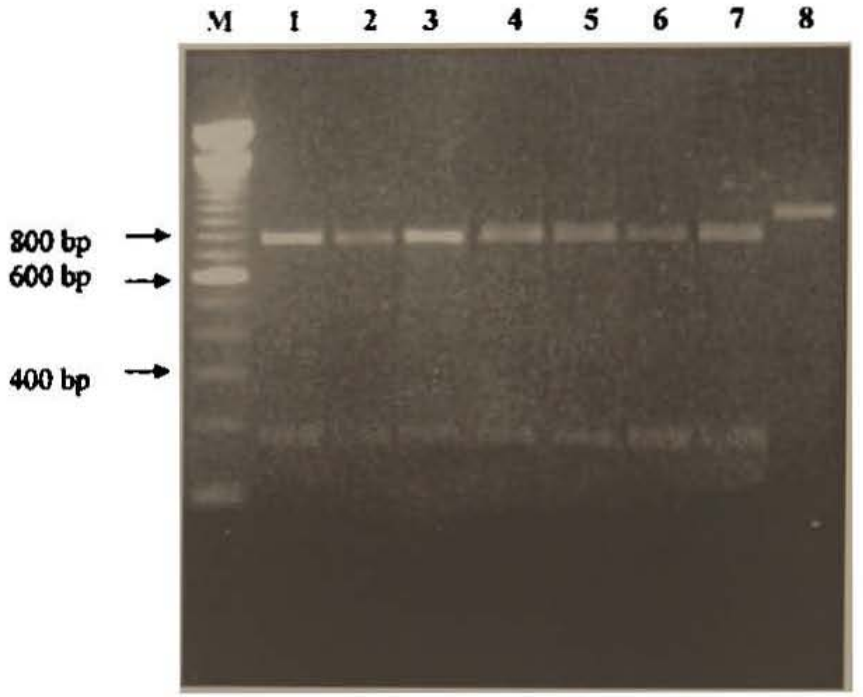


Fig 4.10 Ethidium bromide-stained 2.0% agarose gel electrophoregram of *Apo*I restriction enzyme digests of WOL16S rDNA PCR products. Lane: M, 100 bp molecular weight marker (Sigma-Aldrich, USA); lanes 1-7 = bacterial PCR products from different mosquitoes digested with *Apo*I; lane 8 = undigested PCR product.

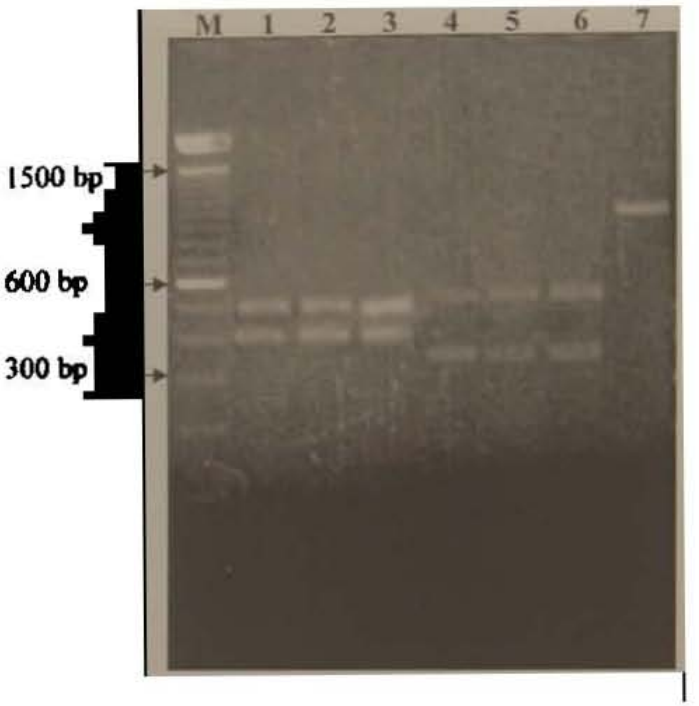


Fig 4.11 Ethidium bromide-stained 2.0% agarose gel electrophoregram of *RsaI* and *NspI* restriction enzyme digests of WOL16S rDNA PCR products. Lane M = 100 bp molecular weight marker (Sigma-Aldrich, USA); lanes 1-3 = bacterial PCR products from different mosquitoes digested with *RsaI*; lanes 4-6 = bacterial PCR products from different mosquitoes digested with *NspI*; 7, lane 7 = undigested PCR product.

4.4 Bacterial Sequences in 16S rDNA

Plasmids isolated from 12 clones with insert (Figs 4.14 & 4.15) were partially sequenced. Using the VecScreen search, five were found to contain sequences identified to be either of vector origin or gave poor quality reads and were therefore rejected. Seven sequences with sizes ranging from 810-992 bp (excluding the eubacterial primer sequences used for amplification) were obtained (Figs. 4.16 – 4.22). Six of these sequences, designated AgA1, AgA2, AgL1, AgL2, AgL3 and AgL4 were obtained from the WOL16S primer PCR products whilst the sequence designated AgL5 was obtained from the eubacteria 16S rDNA primer product. None of the 23S rDNA PCR products was sequenced.

4.4.1. AgA1 (GenBank accession number AY247165)

The details of the AgA1 DNA sequence are shown in Fig. 4. 16. Briefly, it is 992 bp with molecular weights of 306.78 and 611.05 kD for the single and double stranded DNA, respectively. It is composed of 24.4% A, 24.5% C, 30.0% G and 21.1% T and the percentage GC is 54.5%.

4.4.2 AgA2 (GenBank accession number AY325810)

The details of the AgA2 DNA sequence are shown in Fig. 4. 17. It is 991 bp with molecular weights of 306.57 and 610.45 kD for the single and double stranded DNA, respectively. It is composed of 24.8% A, 24.9% C, 30.6% G and 19.7% T and the percentage GC is 55.5%.

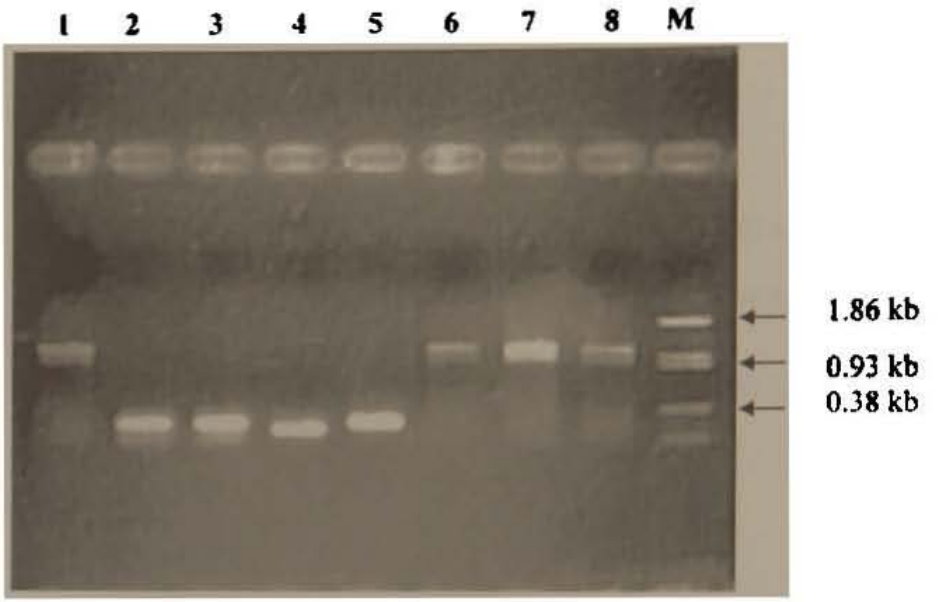


Fig 4.14 Ethidium bromide-stained 1.0% agarose gel electrophoregram showing the presence/ absence of inserts (~ 1000 bp).

Lane M = pBR322 DNA *Bst*N1 digest; lanes 1, 6, 7, and 8 = clones with insert; lanes 2, 3, 4 and 5 = clones without insert

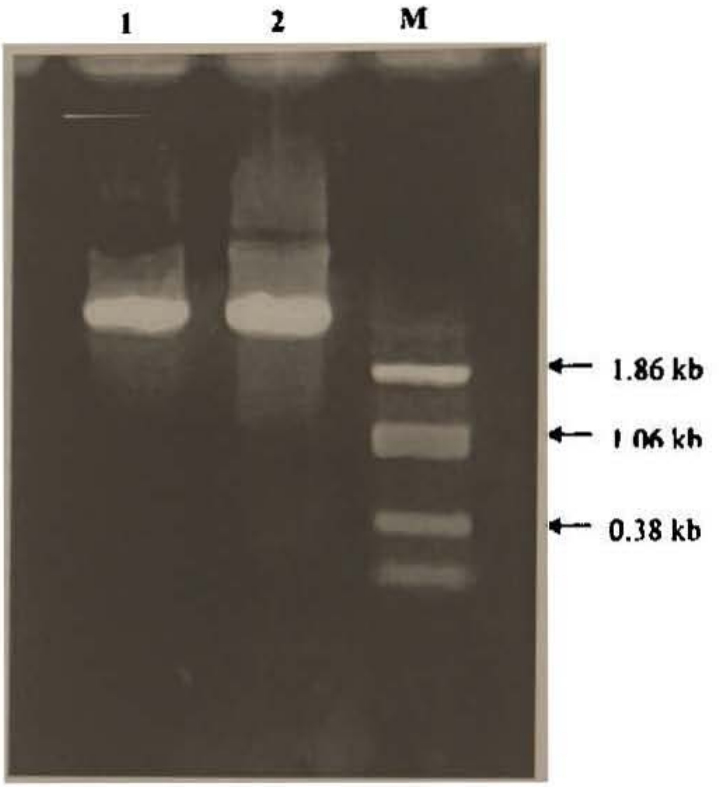


Fig 4.15 Ethidium bromide-stained 0.8% agarose gel electrophoregram of isolated plasmid DNAs.

Lane M = pBR322 DNA *Bst*N1 digest; Lanes 1 and 2 = plasmid DNA isolates.

```
   1 agtcctggct aactccgtgc cagcagccgc ggtaatacgg agggggctag
  51 cgttgttcgg aattactggg cgtaaagcgc acgtaggcgg accagaaagt
 101 tgggggtgaa atcccggggc tcaaccccgg aactgcctcc aaaactcctg
 151 gtcttgagtt cgagagaggt gagtggaatt ccgagtgtag aggtgaaatt
 201 cgtagatatt cggaggaaca ccagtggcga aggcggctca ctggctcgat
 251 actgacgctg aggtgcgaaa gcgtggggag caagcaggat tagataccct
 301 ggtagtccac gccgtagacg atgaatgcca gtcgtcggaa agtatacttg
 351 tcggttgaca cacctaacgg attaagcatt ccgcctgggg agtacggtcg
 401 caagattaaa actcaaaagg aatttgacgg ggggcccgca caaagcggtg
 451 gagcatgtgg tttaattcaa agcaacgcgc agaaccttac caacccttga
 501 catcctgatc gcggttagtg aagacacttt ccttcagttc ggctggatca
 551 gtgacaagtg ctgcattggc tgtcgtcaac ttcgttgtcg tgaaaatgtt
 601 cgggttaagt ccggcaacga agcgcaaccc cacgtcctta gttgccatca
 651 ttcagttggg cactctaggg aaactgccga tgataagtcg gaggaaggtg
 701 tggatgacgt caagtcctca tggcccttac gggttgggct acacacgtgc
 751 tacaatggca gtgacaatgg gttaatccct aaaagctgtc tcagttcgga
 801 ttggggtctg caactcgacc ccatgaagtc ggaatcgcta gtaatcgcgt
 851 aacagcatga cgcggtgaat acgttcccgg gccttgtaca caccgcccgt
 901 cacaccatgg gaattggttc tacccgacga cgctgcgcta accctacggc
 951 aggcaggcgg ccacggtggg atcagtcact ggggtgaaag cc
```

Fig. 4.16 Details of bacterial DNA sequence AgA1 (GenBank accession number AY247165).

```
   1 agtcctggct aactccgtgc cagcagccgc ggtaatacgg agggggctag
  51 cgttgttcgg aattactggg cgtaaagcgc acgtaggcgg aacagaaagt
 101 cagaggtgag atcccagggc tcaacctcgg aactgccttt gaaactcctg
 151 ttcttgaggt cgagagaggt gagtggaatt ccgagtgtag aggtgaaatt
 201 cgtagatatt cggaggaaca ccagtggcga aggcggctca ctggctcgat
 251 actgacgctg aggtgcgaaa gcgtggggag caaacaggat tagataccct
 301 ggtagtccac gccgtaaacg atgaatgcca gtcgtcgggg gaagcatgct
 351 attcggtgac acacctaacg gattaagcat tcccgcctgg ggggtacgcc
 401 cgcaaaggtt aaaactcaaa ggaattgacg ggggcccgca caagcggtgg
 451 agcatgtggt tcaattagaa gcaacgcgca gaaccttacc aacccttaac
 501 atggcagtga ccgttccaga gatggtcctt tctcgcaaag agacacttgc
 551 acacaagtgc tgcatggctt gtcgtcaact tcgtgtcgtg agatgtttcg
 601 ggttaagtcc ggcaacgagc gccaacccca cgccttcagt tgccagcatt
 651 cagttgggca ctctgaagga actgccggtg ataagccgga ggaaggtgtg
 701 gatgacgtca agtcctcatg gcccttacgg gttgggctac acacgtgcta
 751 caatggtggt gacaatgggg taatcccaaa aagccatctc agttcggatt
 801 gtcgtctgca actcggcggc atgaagtcgg aatcgctagt aatcgcgtaa
 851 cagcatgacg cggtgaatac gttcccgggc cttgtacaca ccgcccgtca
 901 caccatggga attggatcca cccgaaggcg gtgcgccaac cagcaatgga
 951 ggcagccgac cacggtggga tcagtcactg gggtgaaagc c
```

Fig. 4.17 Details of bacterial DNA sequence AgA2 (GenBank accession number AY325810)

```
   1 agtcctagct aactccgtgc cagcagccgc ggtaatacgg aggggctag
  51 cgttgttcgg aattactggg cgtaaagcgc acgtaggcgg atcagaaagt
 101 cagaggtgaa atcccagggc tcaaccttgg aactgccttt gaaactcctg
 151 atcttgaggt cgagagaggt gagtggaatt ccgagtgtag aggtgaaatt
 201 cgtagatatt cggaggaaca ccagtggcga aggcggctca ctggctcgat
 251 actgacgctg aggtgcgaaa gcgtggggag caaacaggat tagataccct
 301 ggtagtccac gccgtaaacg atgaatgcca gtcgtcgggt agcatgctat
 351 tcggtgacac acctaacgga ttaggcatcc cgcctgggga gtacgcccgc
 401 aaggttaaaa ctcaaaggaa ttgacggggg gcccgcacaa gcggtggagc
 451 atgtggttta attagaagca acgcgcagaa ccttaccaac ccttgacatg
 501 gcagtgaccg ttccagagat ggtcctttct cgcaagaaac acttgcacac
 551 aagtgcttgc atggctgtcg tcagctcgtg tcgtgagatg ttcggttaaa
 601 tccggcaacg aacgcaaccc acgccttcag ttgccagcat tcagttgggc
 651 actctgaagg aactgccggt gataagccgg aggaaggtgt ggatgacgtc
 701 aagtcctcat ggcccttacg ggttgggcta cacacgtgct acaatggtgg
 751 tgacaatggg gtaatcccaa aaagccatct cagttcggat tgtcgtctgc
 801 aactcggcgg catgaagtcg gaatcgctag taatcgcgta acagcatgac
 851 gcggtgaata cgttcccggg ccttgtacac accgcccgtc acaccatggg
 901 aattggatcc acccgaaggc ggtgcgccag ccagcaatgg aggcagccga
 951 ccacggtggg atcagtcact ggggtgaaag cc
```

Fig. 4.18 Details of bacterial DNA sequence AgL1 (GenBank accession number AY247160).

```
   1 agtcctggct aactccgtgc cagcagccgc ggtaatacgg agggggctag
  51 cgttgttcgg aattactggg cgtaaagcgc acgtaggcgg atcagaaagt
 101 cagaggtgaa atcccagggc tcaaccttgg aactgccttt gaaactcctg
 151 atcttgaggt cgagagaggt gagtggaatt ccgagtgtag aggtgaaatt
 201 cgtagatatt cggaggaaca ccagtggcga aggcggctca ctggctcgat
 251 actgacgctg aggtgcgaaa gcgtggggag caaacaggat tagataccct
 301 ggtagtccac gccgtaaacg atgaatgcca gtcgtcgggt agcatgctat
 351 tcggtgacac acctaacgga ttaagcatcc cgcctgggga gtacggccgc
 401 aaggttaaaa ctcaaaggaa ttgacggggg cccgcacaaa gcggtggagc
 451 atgtggttta attagaagca acgcgcagaa ccttaccaac ccttgacatg
 501 gcagtgaccc gttccagaga tggtcctttc tcgcaagaga cacttgcaca
 551 caagtgctgc atggctgtcg tcaactcgtg tcgtgagatg tttcggttaa
 601 gttccggcaa cgaaccgcaa cccacgcctt caattgccag cattcagttg
 651 ggcactctga aggaactgcc ggtgataagc cggaggaagg tgtggatgac
 701 gtcaagtcct catggccctt acgggttggg ctacacacgt gctacaatgg
 751 tggtgacaat ggggtaatcc caaaaagcca tctcagttcg gattgtcgtc
 801 tgcaactcgg cggcatgaag tcggaatcgc tagtaatcgc gtaacagcat
 851 gacgcggtga atacgttccc gggccttgta cacaccgccc gtcacaccat
 901 gggaattgga tccacccgaa ggcggtgcgc caaccagcaa tggaggcagc
 951 cgaccacggt gggatcagtc actggggtga aagcc
```

Fig. 4.19 Details of bacterial DNA sequence AgL2 (GenBank accession number AY247161).

```
  1 ggctttcacc ccagtgactg atcccaccgt ggtcggctgc ctccattgct
 51 ggttggcgca ccgccttcgg gtggatccaa ttcccatggt gtgacgggcg
101 gtgtgtacaa ggcccgggaa cgtattcacc gcgtcatgct gttacgcgat
151 tactagcgat tccgacttca tgccgccgag ttgcagacga caatccgaac
201 tgagatggct ttttgagatt accccattgt caccaccatt gtagcacgtg
251 tgtagcccaa cccgtaaggg ccatgaggac ttgacgtcat ccacaccttc
301 ctacggctta tcaccggcag ttccttcaga gtgcccaact gaatgctggc
351 aactgaaggc gtgggttgcg ctcgttgccg gacttaaccg aaaatctcac
401 aacacgagct gacgacagcc atgcagcacc tgtgtgcaag tgtctcttgc
451 gaagaaagga ccatctctgg aacggtcact gccatgtcga gggttggtaa
501 aggttctgcg ccgttgcttc gaattaaacc acatgctcca cccgcttgtg
551 cggggccccc cgtcaattcc tttgagtttt aaccttgcgc gccgtactcc
601 cccaagcggg aatgctttaa tttccgttag ggtggtgtca acccgaaata
651 acatgctacc cgacgactgg cattcatcgt ttacggcgtg ggctaccagg
701 gtatctaatc ctgtttgctc cccacgcttt cgcacctcag cgtcagtatc
751 gagccagtga gccgccttcg ccactggtgt tcctccgaat atctangaat
801 ttcacctcta cactcg
```

Fig. 4.20 Details of bacterial DNA sequence AgL3 (GenBank accession number AY247162).

```
  1 ggctttcacc ccagtgactg atcccaccgg ggtcggctgc ctccattgct
 51 ggttggcgca ccgccttcgg gtggatccaa ttcccatggt gtgacgggcg
101 gtgtgtacaa ggcccgggaa cgtattcacc gcgtcatgct gttacgcgat
151 tactagcgat tccgacttca tgccgccgag ttgcagacga caattcgaac
201 tgagatggct ttttgggatt accccattgt taccaccatt gtatcacgtg
251 tgtggcccaa cccgtaaggg ccatgaggac ttgacgtcat ccacaccttc
301 ctccggctta tcaccggcag ttccttcana gtgcccaact gaatgctggc
351 aactgaaagc gtgggttgcg ctcgttgccg gaattaaccg aacatctcaa
401 gacaaaagct gacgacagcc atgcagcacc tgtgtgcagt gtctcttgcg
451 agaaaggacc atctctggaa cggtcactgc catgtcaagg gttggtaaag
501 gttctgcgcc tttgctttcg aattaaacca catgctccaa ccgnttgtgc
551 cggcccccgt caaattcctt tgagttttaa cctttgcggc cggtacttcc
601 ccaggcggga atgcttaatt ccgttaaggt gtgtcaaccg aatagcatgc
651 tacccgacga ctggcattca tcgtttacgg cgtggactac cagggtatct
701 aatcctgttt gctccccacg ctttcgcacc tcagcgtcag tatcgagcca
751 gtgggccgcc ttcgccactg gtgttcctcc gaatatctan gaatttcacc
801 tctacactcg
```

Fig. 4.21 Details of bacterial DNA sequence AgL4 (GenBank accession number AY247163).

```
  1 ggctntacgg ctacctntgt tacgngactt agccncncan attacctgtt
 51 ctaccctaat ccncgttctt gacgcggaac tggcttaggg tctaccagac
101 tttcatggct caggatgaac gcgtagcggc taggcctaat acaagtgtcc
151 cggaacggta tgaaccgggt gcataaggca atgacacact gacgcactag
201 gtgcgtaatt cacagctatg catacgctac ctgagttgct gatccccgaa
251 tcagaacttt ggagnnaacg gagatttcga gataccgcat ccagtaagac
301 agtgnatcat ggctaccctc atgtaccgcc catgtgtaac agcgtgtgtc
351 ggcccatagg cgatagagtg gcgcatgagt ccaacttatg ttacgtctgg
401 tccacgtaac ggcttcactc tcaagtacga tgattgcggt aggcaggctt
451 gatctgagag gataagtccc ccacactggc acctgtagat acgggccaga
501 ctaacaacag ggagttgcag cacgtagcgg aacattaacc caacacctcg
551 acggacgagt ctgaccacag ccatgcagca cgcaggatga aagtcgttat
601 gcagactaat ctaaactgct aaattattcc ttcacattct agcctaagaa
651 aaagttcctt gcgagggaaa gtcaccgaac tataaccaca tactccaccc
701 gcttactcct gccaccaccc gtcggtaatt cctttgaagg tgagaccgtt
751 ggccgacggt acccccctca aatggtgcgt agacggttaa ccgattagtg
801 gtgaaaacgn cgtcacgatt nagtgcattg anctgtng
```

Fig. 4.22 Details of bacterial DNA sequence AgL5 (GenBank accession number AY247164).