# AMPLIFICATION, CLONING AND CHARACTERIZATION OF GENOMIC SEQUENCES CODING FOR RNA HELICASE GENE FROM *Aedes aegypti*

**By**

**Ahmad H. Ibrahim**

**Thesis submitted in fulfillment of the requirements**
**for the degree of**
**Master of Science**

**July 2010**

**Acknowledgements**

In the name of Allah, the most beneficent, the merciful, all praise is on to Him, the sustainer of the heavens and earth and all that which is within it. May his blessings be upon the Prophet Mohammad, peace be upon him amen. It is with the wisdom of Allah that I have finished yet another decisive step in the course of my life.

I am heartily thankful to my supervisor, Associate Professor Dr. Mustafa Fadzil Farid Wajidi, whose guidance, encouragement, and support from the initial to the final level enabled me to develop an understanding of the subject. This thesis would not have been possible without his guidance and persistent help. I would like to thank him for being the first person who taught me. I am proud to record that I had the opportunities to work with an exceptionally experienced scientist like him.

It is an honor for me to thank the Dean of Distance Education School, Associate Professor Omar Majid, and Professor Dr. Hanafi Atan for their support and encouragement. It is a pleasure to thank all the Professors, Doctors and all the staffs at the school of Distance education and the School of Biological sciences for their available support in a number of ways, assistance, advice, guidance, equipments and chemicals they had provided. A special thank to the Vector Control and Research Unit, Universiti Sains Malaysia who provide us with mosquito larvae.

I am indebted to many of my colleagues in the School of PJJJ and in lab 409,407 and 411 especially, Hok Chai, Chee Wah, Eugen Ong, Balqis AbulGhani, Aini Hyati Abdulrahim, Sabrina Safaraldin for giving me such a pleasant time when working together with them since I knew them in USM, and for all what they did to support me in my research.

I owe my deepest gratitude to my family, my mother the one who sincerely raised me with her caring and gently love, and to my beloved wife and my kids for being supportive and kind. I would like also to show my gratitude to Universiti Sains Malysia for the fellowship provided to me.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

# TABLES OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBERVIATIONS

| | |
|---|---|
| λ | Lambda |
| µg | Microgram |
| µl | Micro liter |
| µM | Micro molar |
| aa | Amino acid |
| acc. no. | Accession number |
| BLAST | Basic Local Alignment and Search Tool |
| ATP | Adenosine triphosphate |
| bp | Base pair |
| DIG | Digoxigenin (non radioactive DNA labeling ) |
| dNTP | Deoxynucleotide triphosphate |
| EDTA | Ethylendiamineteraacetic |
| IPTG | Isopropyl β-thiogalactopyranoside |
| kb | Kilo base pair |
| M | Molar |
| min | Minute |
| mM | Millimolar |
| ng | Nano gram |
| OD | Optical density |
| ORF | Open reading frame |
| PCR | Polymerase chain reaction |
| RNA | Ribonucleic acid |

| | |
|---|---|
| rpm | Revolution per minute |
| SDS | Sodium dodecyl sulphate |
| TAE | Tris-Acetic Acid EDTA |
| TBE | Tris-Boric Acid EDTA |
| U | Unit |
| w/v | Weigh per volume |
| X-gal | 5-bromo-4-chloro-3-indolyl-β-galactoside |

**AMPLIKASI, PENGKLONAN DAN PENCIRIAN JUJUKAN GENOMIK GEN HELIKASE RNA DARI *Aedes aegypti***

## ABSTRAK

Nyamuk *Aedes aegypti* adalah vektor patogen arbovirus yang penting seperti demam denggi and penyakit lain. Sebanyak 2.5 juta orang yang tinggal di kawasan yang wujudnya wabak virus denggi. Ini memberikan kepentingan untuk mengawal penyakit dan vektornya. Kajian genomik organisma hidup seperti *Aedes aegypti* boleh mempertingkatkan kefahaman tentang vektor penyakit dan menolong kita mengawal penyakit ini. Kajian yang mendalam telah mengenal pasti enzim yang boleh membuka dupleks acid nukleik. Enzim ini dikenali sebagai helikase. Helikase RNA ialah protein yang terlibat dalam beberapa aspek metabolisme RNA, termasuk transkripsi, hiriscantuman pra-mRNA, biogenesis ribosom dan pengangkutan sitoplasma. Suatu fragmen DNA sepanjang 1415 bp yang mengandungi jujukan yang berpadanan hujung 5' mRNA yang mengkodkan RNA helikase putatif telah diamplikasi dan diklon daripada *Aedes aegypti*. Jujukan yang terletak ke hulu daripada primer PCR pada fragmen yang teramplifikasi berpadanan dengan 346 nt yang pertama hujung 5' helikase RNA kotak DEAD bersandaran - ATP *Aedes aegypti* (No akses ensembl: AAEL004456-RA). Jujukan yang berpadanan mengandungi 297 jujukan ekson dan 49 jujukan yang tidak diterjemah. Jujukan yang terletak di hulu mengandungi

1048 bp jujukan yang tidak terjemah yang mungkin promoter gen ini. Promoter tidak mengandungi kotak TATA tetapi pada kedudukan -419 terdapat jujukan yang berpadanan dengan kotak TATA. Kaedah berjalan kromosom telah digunakan untuk mendapatkan jujukan 3' kepada jujukan awal yang terklon. Fragmen berjalan PCR prhelgw0.72 DNA telah berjaya diklonkan dan mengesahkan data jujukan sasaran helikase RNA dengan 99% kesamaan (analisis BLAST) kepada jujukan mRNA yang terdapat dalam pengkalan data ensembl. Jujukan  baru ini telah ditambah kepada jujukan yang diperolehi dari amplikasi awal untuk memberi jumlah 2000 bp. Penjajaran jujukan protein dilakukan dengan program ClustalW dan menunjukkan homologi jujukan asid amino dengan helikase RNA mitokondria famili kotak DEAD 28 [*Drosophila melanogaster*]. Protein yang dikodkan  oleh gen baru ini diramalkan terdapat dalam mitokondria.

**AMPLIFICATION, CLONING AND CHARACTERIZATION OF GENOMIC SEQUENCES CODING FOR RNA HELICASE GENE FROM *Aedes aegypti***

## ABSTRACT

The mosquito *Aedes aegypti* is an important vector of arbovirus pathogens, such as dengue fever and other diseases. About 2.5 billion people live in regions where transmission of dengue virus occurs. This makes a vital demand to control these diseases and their vectors. Studies on the genomics of living organisms including *Aedes aegypti* can improve our understanding of the disease vectors and help us to control these diseases. Extensive studies have identified enzymes that are able to unwind complementary strands of a duplex nucleic acid. These enzymes are known as helicases. RNA helicases are proteins involved in several aspects of RNA metabolism, including transcription, pre-mRNA splicing, ribosome biogenesis and cytoplasm transport. A 1415 bp DNA fragment from *Aedes aegypti* that contained sequences matching the 5' end of the mRNA coding for a putative RNA helicase was amplified and cloned from *Aedes aegypti*. Sequences upstream of the PCR primer at the 3' end of our amplified fragment matches exactly the first 346 nt of the 5' end of the *Ae. aegypti* DEAD box ATP- dependent RNA helicase (Ensembl accession no: AAEL004456-RA). The matched sequence consists of 297 exon sequences and 49 untranslated sequences.

Upstream of the matching sequence is 1048 bp untranslated sequence that is presumably the promoter of the gene. The promoter does not appear to contain a TATA box but at position -419 there is a sequence that matches a TATA box. The chromosome walking method was used to amplify sequence 3' of the initially cloned sequences. The PCR walking prhelgw0.72 DNA fragment was cloned and the sequence confirmed the targeted selective amplification of a RNA helicase region with 99% similarity (BLAST analysis) to the mRNA sequence deposited in the ensembl data bank. This new sequence was added to the sequence obtained from the initial amplification to give a total of 2000 bp. Sequence alignment of the protein sequence was performed with ClustalW program and showed that the amino sequences was homologous to the RNA helicase gene of the mitochondrial DEAD BOX 28 family [*Drosophila melanogaster*]. This leads to the prediction that the encoded protein of the newly cloned gene is localized in the mitochondria.

# CHAPTER ONE

# INTRODUCTION

## 1.0 General introduction

Mosquitoes are vectors of many important human diseases. There is a vital demand to develop and control these diseases and their vectors. The mosquito *Aedes aegypti* is an important vector of arbovirus pathogens. The mosquito maintains a close association with human populations and is the principal vector of the etiological agents of yellow fever and dengue fever. According to the World Health Organization (WHO, 2006), and reports from recent studies, about 2.5 billion people live in regions where transmission of dengue virus occurs. This makes dengue an increasingly important public health concern for which no effective therapy currently exists (Suaya *et al*., 2006). Melino & Paci (2007) reported an estimated 100 million cases of dengue fever yearly, together with 250,000 to 500,000 cases of dengue hemorrhagic fever. Dengue fever is caused by four closely related virus serotypes designated; DEN-1, DEN-2, DEN-3 and DEN-4 of the genus *Flavivirus* and family Flaviviridae. Dengue fever (DF) is characterized by fever and bleeding disorders, all of which could progress to high fever, shock and death in extreme cases. DF is a fast growing public health problem in tropical and subtropical countries where the greater part of the world's population reside (Kuno *et al.*, 1998; Twiddy *et al.*, 2002). The advent of genomics can has opened new ways of helping us understand living

organisms better including understanding the biology *Aedes aegypti*. For example, elucidating some of the important features of vector capability can improve our understanding of the mosquito and its association with etiological agents. Characterization of genes in mosquitoes could also help in unravelling the mechanisms involved in resistance that could lead to the development of novel control strategies of the disease vector (Alphey, 2002).

Research on dengue viruses (DENV) has led to the characterization of large number of virus-encoded proteins and enzymes as well as envelope and capsid proteins, polymerases, helicases and proteases. The mechanism of action involved in the access of DENV into host cells is becoming better understood at the molecular level. The development of vaccines against dengue is an active area of research, which is complicated by the presence of four DENV serotypes and the lack of suitable animal model for dengue disease. The access to molecules of potential therapeutic interest has long been a matter of great concern. Up to the present moment, no anti-DENV drug has been reported (Smith & Helenius 2004; Selisko *et al* 2007). The alternative strategy for combating dengue fever is to identify low molecular weight molecules that could selectively block the function of the proteins encoded by these viruses. These molecules control the intracellular traffic of DENV proteins in the infected cell. Currently these strategies are under investigations (Lum *et al*., 2007; Chene, 2009). Given the difficulties of finding a vaccine for dengue fever, the major method of controlling vector-

borne diseases is still by elimination of their vectors. Because of this, epidemiological and entomological studies are needed order to develop and deliver solutions, which can respond to the main risk concerns of dengue.

## 1.1 The mosquito *Aedes aegypti*

The mosquito *Aedes aegypti* (Figure 1.1) is a morphologically, biologically and physiologically complex species.

**Figure 1.1**: Adult female *Aedes aegypti* feeding on a human arm, taken from CDC website URL: http://phil.cdc.gov/phil/details.asp?pid=9260

*Aedes aegypti* is classified as;

| | |
|---|---|
| Kingdom: | Animalia |
| Phylum: | Arthropoda |
| Class: | Insecta |
| Order: | Diptera |
| Family: | Culicidae |
| Genus: | *Aedes* |
| Subgenus: | Stegomyia |
| Species: | *Aedes aegypti* |

Two notable forms of *Aedes aegypti* are found in nature; first a light-colored form of *Aedes aegypti aegypti* which is common in the tropics and subtropics and breeds specifically in domestic environments; second a dark form, *Aedes aegypti formosus,* is found mainly in African jungles and breeds in tree-holes and sometimes in rock-holes ( Failloux *et al*., 2002). The jungle species of *Aedes aegypti* from tropical Africa probably first became domesticated as a result of human water-storage spaces, which became ideal breeding environments. These populations have differentiated into domestic populations known as *Aedes aegypti* (Bosio *et al.*, 2000; Failloux *et al*., 2002). About 2.5 billion people live in the tropical and subtropical areas of the world, with approximately 120 million people every year making trips to these regions (Elder & Lloyd 2007; Suaya *et al*., 2006). It has been estimated that hundreds of thousands of hospitalizations and about 20 000 deaths occur each year due to incidence of dengue fever. However, most of these cases

occur in those countries that are economically disadvantaged and are facing many other public health problems (Suaya *et al*., 2006). The distribution of the risk of dengue virus infection is shown in Figure 1.2.

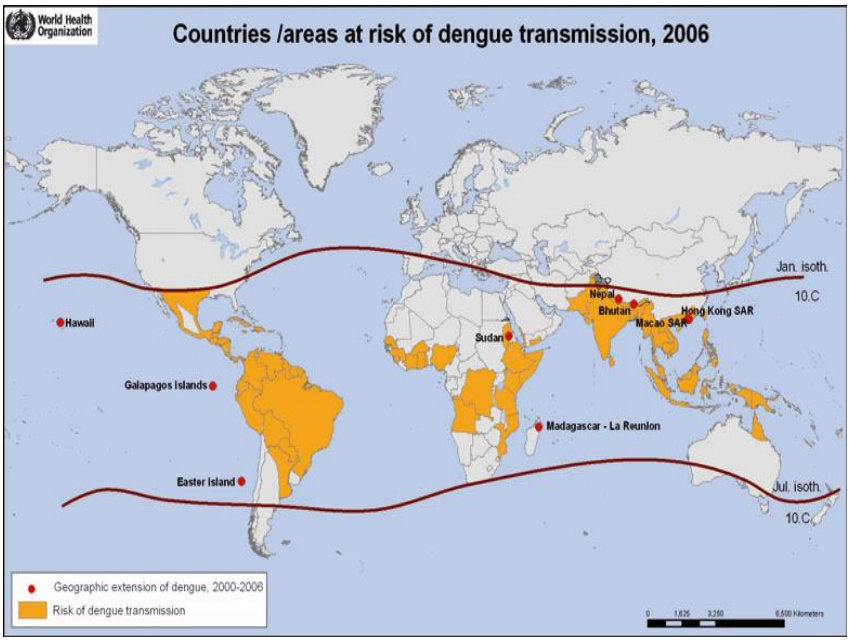


**Figure1.2**: The world map of countries showing risk of dengue virus infections in 2006, according to the WHO (Taken from Melino & Paci, 2007).

## 1.2 Strategies to control *Aedes aegypti* population

Many strategies are also currently employed to control *Aedes* mosquitoes, such as destruction of breeding sites or by larviciding with insecticides. In addition, predatory *Mesocyclops* spp have been demonstrated to control mosquito larvae. The adult mosquitoes are also controlled by the spraying of insecticide treated materials on indoor household items due to the

failure of outdoor sprayed insecticides reaching indoor mosquito populations (McCall & Kittayapong 2007).

Mosquito pests have been controlled almost exclusively with insecticides since the introduction of DDT in 1940s. However, the development of resistance to several insecticides possibly together with the increased awareness of the environmental, human and animal health hazards when in contact with these chemicals has encouraged the search for new insecticidal compounds, novel molecular targets and alternative control methods (Sadasivaiah, *et al*., 2007; Selisko *et al* 2007).

In the search for novel tools for vector control, Alphey (2002) has proposed the development of genetic engineering tools that could be used on mosquito vectors as an alternative strategy for mosquito control. Therefore, efforts are focused on genes that enhance insect immunity to pathogens or genes that will reduce the reproductive capacity of mosquito populations (Alphey *et al*., 2002). There is also hope that naturally occurring arboviruses can be customized to express and silence genes in mosquitoes (Macer, 2005).

Current efforts on the genomics of *Aedes aegypti* will further contribute to the arsenal of strategies for mosquito control. The efforts in sequencing of *Aedes aegypti* DNA were planned to provide new opportunities for research into development of novel insecticides and possible genetic modifications to

prevent the spread of arboviruses. The genome of the *Aedes aegypti* mosquito has been totally sequenced. The published data included about 1376 million base pairs containing an estimated 15,419 protein-encoding genes (Nene *et al*., 2007). The knowledge gathered on these genes and proteins will provide researchers with a new array of targets that can be used for studies leading to more effective control measures in attempts to reduce or eliminate the disease vector.

It is proposed here that the RNA helicase gene can be used as one of the molecular targets for controlling *Aedes* mosquitoes. RNA helicase is a ubiquitous enzyme that is involved in many essential cellular functions. Therefore, disruption of this gene or protein may lead to total or reduced disfunction of the mosquito's metabolism thus causing death.

## 1.3 Research Objectives

Extensive research has been conducted on helicase genes in many species. However, information about the ATP-dependent RNA helicase genes from *Aedes aegypti* is still scant. The research objectives of this study are as follows:

(i) To amplify and clone the genomic sequences of the RNA helicase gene from *Aedes aegypti*.

(ii) To determine of the nucleotide sequences of the RNA helicase gene from *Aedes aegypti* and comparing with the sequences of other known RNA helicase.

(iii) Characterization of the RNA helicase gene from *Aedes aegypti* by looking for specific motifs or domains.

# CHAPTER TWO

# LITERATURE REVIEW

The use of genomics in recent years has improved our understanding of a variety of complex phenotypes and helps elucidate the fundamental processes that characterize living systems. By understanding each gene structure, function and location, researchers are offered an ever greater opportunity to rapidly interpret, assess and create new theories on how biological systems work. To take advantage of this however, we first need to identify all the hereditary components of a particular cell and unravel the mysteries of the genome by conducting sequencing research projects (Hughes *et al.*, 2004; Severson, *et al.*, 2004).

## 2.1 The cellular role of helicase

The nucleic acids in the cells often have to be unwound so that cellular process may proceed using the information found in these modules. Helicases are a diverse class of enzymes that have the ability to unwind nucleic acid duplexes with a separate directional polarity (Bennett *et al.*, 1999; Singleton *et al.*, 2007). They can be divided into DNA or RNA helicases, depending upon the nucleic acid substrate on which they act. Most helicases can unwind either DNA or RNA, but some can unwind both. Similarly, helicases can be classified into families and superfamilies based on their primary sequences (Cordin *et al*., 2006; Jankowsky & Fairman, 2007).

The enzymes moves directionally along a nucleic acid phosphodiester backbone, separating two annealed nucleic acid strands (i.e. DNA, RNA, or RNA-DNA hybrid) using energy derived from ATP hydrolysis (Cordin *et al.*, 2006; Linder, 2006).

Helicases are involved in many aspects of cellular metabolism including DNA replication transcription, splicing, translation, RNA export, ribosome biogenesis, pre-mRNA processing, DNA repair, recombination and the regulation of mRNA stability (Lohman & Bjornson, 1996; Eisen *et al*., 1998; Matsui *et al.*, 2006). In general, the RNA helicases are involved in RNA metabolism whereas DNA helicases act on DNA as their template (Linder, 2006; Cordin *et al.*, 2006; Singleton & Wigley, 2002).

DNA and RNA sequence analysis as well as biochemical and genetic experiments, reveal that helicases are composed of two 'core' domains with highly conserved motifs (Tanner *et al*., 2003). The two core domain are called the motif I (or Walker A) and motif II (or Walker B). The Walker A motif was defined classically with amino acid sequence GxxxxGKT, where the three final residues are GK (T/S). On the other hand the Walker B motif, originally defined as a single aspartic acid residue, has the general form DExx (Gorbalenya & Koonin, 1993).

DEAD-box helicase take its name after the four-amino-acid sequence Asp-Glu-Ala-Asp (DEAD), which is part of the specialized version of the walker B motif (de la Cruz *et al.*, 1999; Linder, 2006). Although the DEAD-box proteins are the largest and probably best studied family with regards to biological function, only a few RNA helicases from the DEAD-box family have been characterized biochemically (Caruthers & McKay, 2002; Rocak & Linder, 2004, Owttrim, 2006). DEAD/H-box protein families have a conserved fragment of approximately 400 flanking amino acid residues which maintains significant sequence conservation, including the classical helicase motifs. Linder *et al*. (2006) reported that DEAD-box proteins are characterized by a set of conserved motifs and several hundreds DEAD-box proteins can be identified in genetic databases based on these criteria. At the sequence level, RNA helicases are identified by the presence of seven to nine conserved motifs. These motifs are involved in binding an NTP, specifically ATP and using the energy of hydrolysis to unwind dsRNA.

## 2.2 Basic structure of helicases / DEAD-box proteins

Helicases are classified by structure and sequence into several superfamilies. All currently known helicases belong to four helicase superfamilies viz, SF1, SF2, SF3 and SF4 (Jankowsky & Fairman, 2007). Each superfamily can have specific variations in the conserved sequence motifs, as well as other family-specific properties. Despite the fact that the overall helicase sequence homology is low, conserved regions or motifs

within each family have been identified. A survey of these DEAD box proteins shows that they vary in size from approximately 400 residues to greater than 1200 residues (Korolev *et al*., 1998; Caruthers & McKay, 2002; Cordin *et al*., 2006).

Interestingly, the DEAD-box proteins are quite unique in many of its sequence motifs. The motifs are named as F, Q, I, Ia, Ib, II, III, IV, V and VI (Figure 2.1).

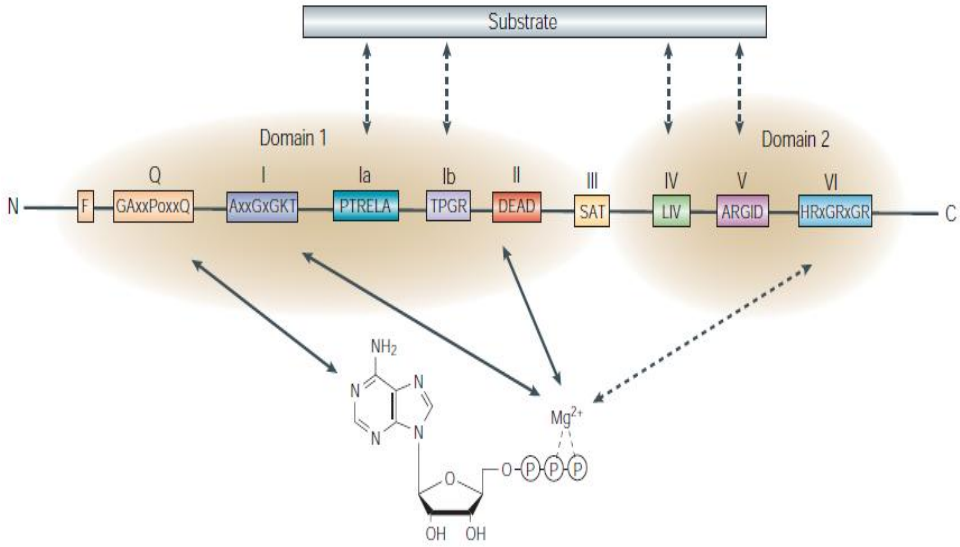


**Figure 2.1**: The conserved motifs of DEAD-box proteins and their interaction with ATP. A sequence alignment of the DEAD-box proteins has revealed nine conserved motifs, (taken from Rocak & Linder, 2004).

Motif I is identified as the Walker A motif, whereas motif II is identified as the Walker B motif. All the helicases contain the Walker A and B motifs. Walker A and Walker B are the best known motifs; both of them are present in all the DEAD-box protein families. The two Walker motifs are also found in many other NTPases. Mutational analysis has revealed that Walker A is responsible for ATP binding and Walker B is involved in ATP hydrolysis. Motif III is associated with the unwinding function and motif VI is required for RNA binding (Schmidt *et al*., 2002; Tanner *et al*., 2003; Cordin *et al*., 2004; Linder, 2006).

## 2.3 Biological activities of helicase

Recently, it has become clearer that helicase motifs are actually characteristic of proteins that are able to move towards nucleic acid strands. Therefore they were named translocases (Koonin & Rudd, 1996; Bork & Koonin1993; Cordin *et al.,* 2006; Fuller-Pace, 2006; Singleton *et al.,* 2007). DEAD-box proteins possess both RNA-dependent ATPase activity and ATP-dependent RNA helicase activity, and are responsible for duplex RNA unwinding as well as facilitating the rearrangement of the RNA structure (Matsui *et al.*, 2006).

### 2.3.1 Function during cell division

Denaturing of double-stranded DNA in the laboratory into two single-strands normally occurs at temperatures over 90°C (Devlin, 2006). To

unwind the parental DNA strands at physiological melting temperature, helicase enzymes separate the strands by generating a replication fork. The single-stranded DNA which results from this helicase activity is layered with single-stranded DNA-binding protein (SSB) that blocks the re-annealing parental strands and avoids hairpins and other structures from appearing in the single-stranded DNA.

Helicases function as molecular motors and display a wide array of biochemical activities. Helicases are capable of unwinding and completing translocation along double or single-stranded nucleic acids using the chemical energy of nucleotide tri-phosphate (NTP) hydrolysis (Linder, 2000; Silverman *et al.*, 2004; Eoff & Raney, 2005; Jankowsky and Fairman, 2007). Furthermore, helicases utilize the energy of ATP hydrolysis to open double-stranded DNA into two single strands (Abdel-Monem & Hoffmann-Berling, 1976; Egelman, 1998; Jankowsky & Fairman, 2007).

### 2.3.2 Function as "coupling" factors

The RNA helicases also have roles within large complexes, such as in the ribosome or the spliceosome by acting together with protein factors in the complexes to modulate their activity. Sequence analysis of several genomes revealed that DEXD/H-box proteins are the largest family of helicase enzymes largely involved in catalyzing ATP-dependent remodeling of RNA-

protein complexes and the unwinding of RNA duplexes (Jankowsky & Bowers, 2006).

Silverman *et al.* (2004) suggested that pre-mRNA splicing is a dynamic process through which the trans-acting proteins interact in the spliceosome to remove introns in an ordered manner. A number of DExD/H-box proteins contain the partner domains or outer areas of the helicase domain which confer specificity in spliceosome or pre-ribosome complexes. This is achieved perhaps by serving as the sites for interaction with regulatory proteins. Additionally, protein cofactors are able to interact with RNA helicases to promote target recognition and helicase activity. The processing activity of a helicase is commonly regulated by further protein components or "coupling" factors, which may interact with the helicase both directly and indirectly using nucleic acid components of the system (von Hippel & Delagoutte, 2001).

Furthermore, to allow the RNA helicase to join in multiple cellular processes, it is likely that those protein cofactors engage the helicase to form an appropriate complex. The DExD/H-box proteins frequently have related cofactors to facilitate the control of activity and coordinate functions such as the helicase eukaryotic translation initiation factor 4A (eIF4A) and cofactor eIF4B as well as the hepatitis C virus helicase NS3. Protein cofactors can act physically to modulate helicase activity through direct protein-protein

interactions or indirectly by utilizing a complex of proteins. Cofactors act cooperatively by bringing the helicase in contact with the RNA. The helicase-cofactor complex acts by physically changing the shape of the protein. This modulation, in turn, increases RNA helicase activity. (Silverman *et al.*, 2003; Silverman *et al*., 2004) [Figure 2.2].

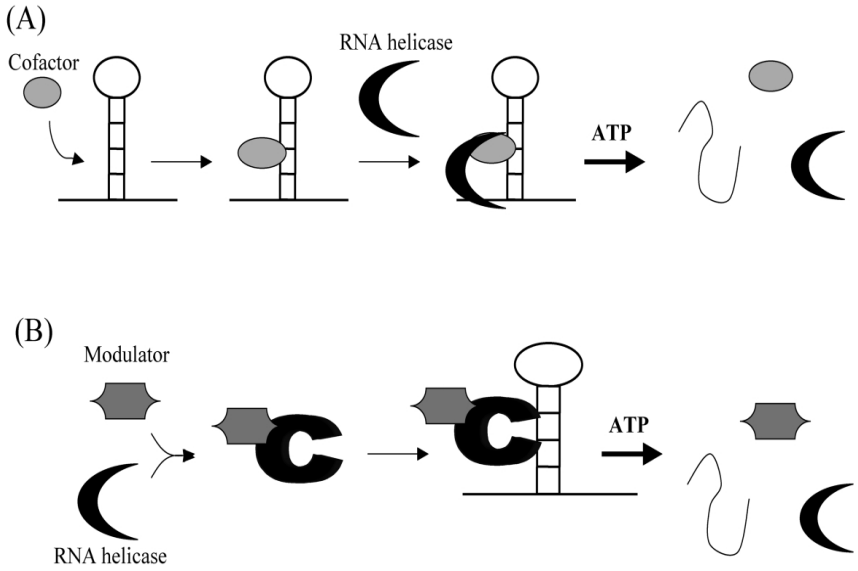


**Figure 2.2 A)** Bridging Model of DExD/H-box cooperative cofactor. The cofactor binds to the RNA target, and increases the ability of the RNA helicase to associate with the RNA target.

**B)** Modulate model of cooperative protein functions in DExD/H-box proteins (Taken from Silverman *et al*., 2003).

### 2.3.3 Translocation activities of helicase

Many different mechanisms have been implicated for translocation in cellular processes, which separate double stranded nucleic acid into single strands. Patel and Donmez (2006) suggested that base pair separation happens at the junction of single-stranded and duplex regions. Mackintosh and Raney (2006) showed that translocation is coupled to ATP hydrolysis. It is believed that translocations involving helicases occur along the single-stranded nucleic acid in a fixed direction and disjoints the complimentary strand requiring hydrolysis of ATP at each step.

Kim *et al.* (1998) studied the structure of HCV NS3 bound to a DNA substrate. They observed that the DNA strand bound to the N terminal and C terminal domains and to the inter-domain cleft. From this structure, a model was proposed suggesting that ATP binding and hydrolysis could cause the inter-domain cleft to open and close, thereby allowing the two helicase domains to move relative to each other via the flexible linker region during translocation. This opening and closing of the DNA cleft could alter DNA binding and thus allows the protein to move along the DNA substrate, thereby enabling duplex unwinding via subsequent displacement of the complementary strand (Tanner & Linder, 2001; Linder, 2006; Jankowsky & Fairman, 2007).

## 2.4 Relationships among helicases

The relationships among helicase proteins were identified by different genetic approaches. Despite their sequence variations, the amino acid sequences expose a high degree of conservation of different motifs. The conserved motifs are found in the core region which is flanked by 5' and 3' external sequences of various lengths (Linder, 2000). The RNA helicases were classified into different superfamilies (SF) based on the sequence similarities of a few hundred proteins that were available at that time. RNA helicases have been divided into three large superfamilies and two smaller families (Gorbalenya & Koonin, 1993; Singleton *et al.*, 2007).

van Brabant *et al.* (2000) acknowledged that many biochemically uncharacterized proteins were designated as putative helicases based on homology to known helicases. Therefore, the analysis of amino acid sequence alignments has allowed the phylogenetic grouping of these "helicase pretenders" into certain families. At the same time, individual point mutations can accumulate, ultimately giving rise to enormous sequence diversity.

## 2.5 Common characteristics of helicases

DNA and RNA helicases share some common characteristics and most contain short, conserved amino-acid fingerprints called helicase motifs. In general, eukaryotes have more DEAD-box proteins then prokaryotes. Many

helicases also have requirements for specific structures in their nucleic acid substrates. For example, many helicases require either a 3' or 5' overhang adjacent to a duplex nucleic acid molecule in order to unwind the duplex (Tuteja & Tuteja, 2004). Helicases need to hold multiple DNA binding sites in order to function (Patel & Donmez, 2006). The active forms of most helicases are shown to be oligomeric, generally dimeric or hexameric (Lohman & Bjornson, 1996). Characteristic properties of helicases correspond to their oligomeric condition and the manner of binding to the nucleic acid at the unwinding junction. However, there is still some debate regarding the role of specific sub-domains in the overall mechanism for DNA unwinding (Mackintosh & Raney, 2006).

The helicase processivity and rate by which helicases unwind different substrates can vary by order of performance. The DNA helicases can unwind substrates several kilobases long, while most RNA helicases unwind substrates of less than 100 bases. Therefore, most RNA helicases do not appear to act on long RNA duplexes in the cell. The RNA helicases have, in addition, also been shown to participate in each stage of RNA metabolism (Cordin *et al*., 2006).

## 2.6 Classification and discovery of helicases

Since the discovery of the earliest DNA helicase in *Escherichia coli* in 1976 and the earliest detection of eukaryotic ones in *Liliaceae* (Lily) in 1978,

multiple DNA helicases have been isolated from different organisms. A large number of these enzymes have been isolated from both prokaryotic and eukaryotic cells and the number is still increasing (Tuteja & Tuteja, 2004). Gorbalenya & Koonin (1993) described three superfamilies and two families of putative helicases on the basis of primary structure analyses. The classification of helicases is based on the occurrence and characteristics of conserved motifs in the primary sequence. Helicases are classified into five main groups (named SF1 to SF5). This was done after it was recognized that a smaller subset of related helicase proteins obviously share short conserved amino acid fingerprints or motifs (Kikuma *et al*., 2004 and Cordin *et al*., 2006).

### 2.6.1 Classification of helicases

Sequence analyses of helicases from various groups of organisms have revealed a series of short and conserved amino acid motifs. Helicases are classified by their structure and sequence into five superfamilies, the largest of which are SF1 and SF2 (Gorbalenya & Koonin, 1993; Koonin & Rudd, 1996; Hall & Matson 1999 and Singleton *et al.*, 2007).

It has been observed that in SF1 & SF2, between 7 to 9 conserved motifs are generally clustered in the central 200-700 amino acids known as the 'core-region' of the protein. On the other hand, SF3 is a small family that has only three motifs, including the two classical ATP-binding motifs.

Members of SF3 exist in RNA and DNA viruses and include small putative helicase domains of approximately 100 amino acid residues (Caruthers & McKay, 2002; Shankar & Tuteja, 2008).

Caruthers and McKay (2002) opined that SF4, containing the *E. coli* DnaB-like hexameric helicases, forms another smaller, distinct group. The DnaB protein has five motifs, unwind DNA in the 5' to 3' direction and generally forming hexameric ring structures. SF4 also contains enzymes such as the bacterial transcription termination factor. A study by Patel and Picha (2000) showed that the known hexameric helicases were discovered in a ring-shaped structure. They also observed that *E. coli* DnaB and Rho were among the first proteins shown to form hexamers and functioned as helicases. The SF3 and SF4 are generally hexameric helicases that are basically of bacterial or viral origin (Soultanas & Wigley, 2000; Cordin *et al*., 2006).

### 2.6.2 Superfamily 1 (SF1)

Gorbalenya and Koonin (1993) and Koonin and Rudd (1996) proved that DNA and RNA helicases of SF1 are characterized by seven conserved motifs (I, 1a, II, III, IV, V and VI). Five proximal motifs are separated from the two distal motifs by a spacer, which varies widely in length from about 50 to 500 amino acid residues. It has been suggested that the distal motifs belong to a separate domain (Gorbalenya & Koonin, 1993). All helicases contain two common motifs known as the Walker motifs shared by these SF1

members. Members of the SF1 helicase include PcrA, Rep and *E. coli* Uvrd (Lee & Yang, 2006; Shankar & Tuteja 2008).

### 2.6.3 Superfamily 2 (SF2)

Helicase superfamily 2 (SF2) has over 100 known members in humans and is implicated in diverse cellular processes. Although SF2 proteins are fairly similar in sequence, enzymes of this helicase superfamily are diverse in their functions. The SF2 was originally defined by the presence of seven conserved sequence motifs (Gorbalenya & Koonin, 1993). According to the average number of base pairs unwound per helicase binding event (processivity), the SF2 proteins may unwind either DNA or RNA in a processive or non-processive fashion and in either a 3' to 5' or 5' to 3' direction. Evidence suggests that the SF2 helicases include many diverse enzymes that are best known for catalyzing ATP-driven separation of nucleic acid duplexes. Some proteins are specific to ATP while others can utilize any nucleotide for hydrolysis.

### 2.6.4 Similarities between SF1 and SF2

SF1 and SF2 helicases are the largest and most closely related and are widely found in viruses, prokaryotes and eukaryotes (Hall & Matson, 1999). SF1 and SF2 helicases have a 3' to 5' directionality. In both superfamilies conserved motifs have been identified and the specific characteristics of the motifs are used to sub-classify the proteins into different families.

Anantharaman *et al*. (2002) recognized that SF1 and SF2 were defined by similar sets of seven motifs that typically range ~400 amino acids. The Walker A and B sequences match up to motifs I and II (GxGKT/S and DExx, respectively, where x can be any amino acid). SF1 is characterized by variations of the VALTR sequence in motif VI, while SF2 helicases have H/QrxgRxgR in this region.

### 2.6.5 Helicase DEAD-box families

More than 500 cellular RNA helicases have been considered and classified into different families. An analysis of genomics sequence data identified considerable numbers of open reading frames containing some or all of the characteristic helicase motifs and has allowed classification of the respective gene product into one of the helicase classes. Shanker &Tuteja (2008) provided evidence that the family of DEAD-box RNA helicases contained characteristic sets of conserved sequence motifs, called a helicase core, usually flanked by specific amino- and carboxyl-terminal domains which vary widely in length and sequence.

The degree of amino acid similarity and the organization of the conserved regions of helicase motifs, has led to the grouping of all the helicases including many putative helicases into distinct families. Furthermore, the respective motifs defined by the alignment of the members

of all families, possibly share their functional characters (Rocak & Linder, 2004). Based on the analysis of crystal structures, genetic data and biochemical experiments, the individual roles of each motif has been proposed (Tanner and Linder, 2001; Schmidt *et al.*, 2002; Tanner *et al.*, 2003; Wagner *et al.*, 2005; Cordin *et al*., 2004 and Linder, 2006).

### 2.6.6 RNA helicases

With the growing number of identified putative RNA helicases and other related proteins it became clear that the members of SF2 could further be categorized into different subgroups. The classification of these proteins is based on the amino acid sequence of conserved motifs such as the DExD/H-box proteins. The DExD/H helicases are further divided into two subgroups named for a conserved Asp-Glu-Ala-Asp (DEAD) amino acid motif or the DEAH-box, which contain Asp-Glu-Ala-His (DEAH) motif (Pause *et al.* 1993; Svitkin *et al*. 2001; Tanner, 2003; Rocak *et al.*, 2005).

Based on the observation that most typical members of SF2 show showing RNA helicase activity *in vitro* helicases, it has been suggested that DExD/H proteins act mainly as ATP-dependent RNA helicases. Although DExD/H box proteins have shown a significant sequence and structural homology within their conserved 'helicase' core, their flanking N- and C-terminal domains are extremely at variance and provide specificity of function during interaction with particular RNA substrates or other