CHAPTER II

LITERATURE REVIEW

Hepatitis C Virus (HCV)

HCV is a positive-sense single stranded RNA virus that belongs to the family of *Flaviviridae*; the most closely related human viruses are hepatitis G virus, yellow fever and dengue virus. The virion has a 40-60 nm. diameter with spikelike projections and has a lipid coat. The buoyant density has between 1.09 and 1.11 g/mL in sucrose gradient (13,14).

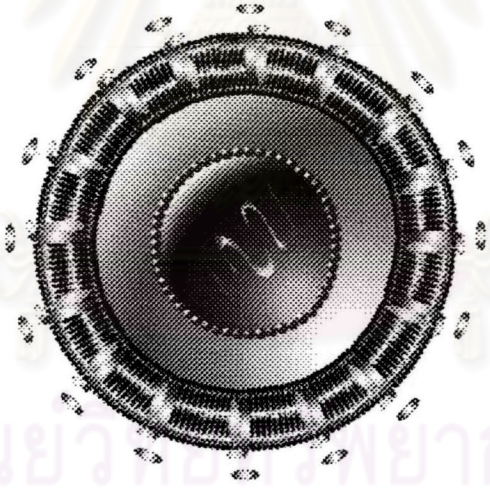

Figure 1.The virion of hepatitis C virus. A protein capsule has peptide nodules that lock into, and allow infection of host cell. The changing of the envelope proteins may help it evade the immune system. Inside the protein capsule is a capsid containing genetic chain that alter the replication cycle of hepatocytes during the natural history of the disease (15).

HCV genome consists of a single open reading frame (ORF) or coding region and two untranslated regions (UTR) or non coding regions (NCR). The large ORF encodes a polyprotein precursor from which individual viral proteins are processed co– and post–translationally through the combined action of host and virus-specific protease (1,16). Most of the information available on the viral proteins of HCV and putative function has been derived from analyses of cDNA expression system.

The HCV RNA is approximately 9,400 ribonucleotides, containing a poly A at the 3' end. The ORF length of each genotype is characteristically different. The ORF in genotype 1 is approximately 9,400 ribonucleotides whereas 9,099 nucleotides in genotype 2 and 9,063 nucleotids in genotype 3 (17). These differences may account for some of the phenotypic differences among genotype. A long open reading frame coding for a polyprotein of 3,010 amino acids that processed into 2 groups of proteins; structural and regulatory proteins. The structural proteins of HCV appear to be processed from the N-terminal region of the polyprotein at least in part though the action of host signal peptidase cleaving after internal signal sequences within the polyprotein. The action results in the production of a basic, the nucleocapsid protein (C) with size approximately 22 kD at the extreme N-terminus of the polyprotein and is followed by two glycosylated proteins (envelope proteins; E). Two regions of the envelope E2 protein, designated hypervariable regions 1 and 2 (HVR-1 and HVR-2), have a high rate of mutation, believed to be the result of selective pressure by virus-specific antibodies. HVR-1 is a 27 amino acid segment in the amino terminus of the second envelop protein which has been identified as the most variable region of the genome (18). E2 also contains the binding site for CD81, a tetraspanin expressed on various cell types including hepatocytes and B lymphocytes that is thought to function as a cellular receptor or co-receptor for the virus (19). The human CD81 is sufficient for binding not only E2 but also HCV particles. HCV can bind to a variety of cells other than hepatocytes because HCV RNA has been found in T lymphocytes, B lymphocytes, and monocytes. Whether virus binding to

CD81 is followed by entry and infection in all cell types is not clear, because it is possible those additional factors are required for HCV fusion or infectivity.

The regulatory proteins or non-structural proteins of HCV appear to be processed from C-terminal region of the polyprotein in large part though the combined action of two encoded proteases. The regulatory proteins contain the NS2 (~23 kD), NS3 (70-72 kD), NS4A (4-10 kD), NS4B (27kD), NS5A (56-58 kD) and NS5B (68-70 kD) protein. The NS2, NS3, NS4A proteins have function as proteases which are involved in the processing of the non-structural region, NS3 protein functions as a NTP binding helicase enzyme activity and NS5B protein functions as RNA-dependent RNA polymerase (20, 21). Nonstructural protein 5A (NS5A) is the amino terminal half of nonstructural protein 5. A region in NS5A have been linked to the response to α-interferon (IFN-α) therapy and called the interferon-sensitivity-determining region (ISDR). The relation between the type of NS5A sequence and the HCV RNA level suggested that NS5A has an important role in HCV replication. Thus, mutations in NS5A may suppress the replication of HCV and increase susceptibility to interferon. NS5A may be a direct target of antiviral proteins induced by interferon (22). However, the function and properties of the other proteins in this coding region such as p7 are less well characterized.

The two untranslated regions (UTR) or non coding regions (NCR) are consist of 5' and 3' NCR. The 5' NCR of HCV RNA is the most highly conserved portion of the genome. The sequence variation of 5' NCR is approximately 10% while as much as 50% or more within E1 region (23, 24). This percentage suggested that 5' NCR may play a very important regulatory role during viral replication, perhaps at the level of translation since a small, hairpin, secondary structure could form at this region. The 5' NCR of HCV ranges in length from 332 to 343 nt and contains up to five AUG codons, depending on the HCV genotype or subtype (25). The translation of HCV ORF is initiated by a cap-independent internal ribosomal entry mechanism mediated by an internal ribosomal entry site (IRES) situated

within the 5' NCR (26). A small, presumed 3' NCR has been identified downstream from the large ORF that contains between 27 and 55 nucleotides depending on the source of virus.

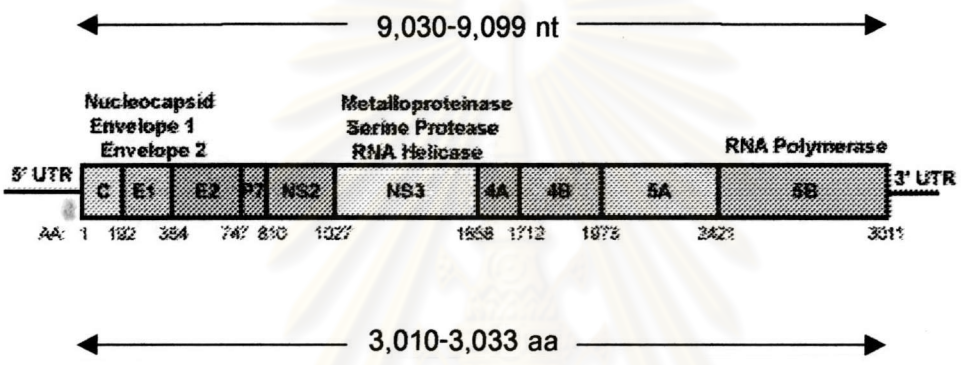

Figure 2. The genetic organization of the HCV genome. The boxed area corresponds to the single open reading frame of the HCV genome. The 5' untranslated regions (UTR) including the internal ribosome entry site (IRES) and play an important role during viral replication.

The natural targets of HCV are hepatocytes and B lymphocytes (27). Like other RNA viruses, viral replication occurs through an RNA-dependent RNA polymerase that lack a proofreading function which results in the rapid evolution of diverse but related quasispecies within an infected person. HCV exhibits high mutation rate at an average of 1.44-1.92 x 10^6 base substitutions per site per year (28). This sequence variation is not evenly distributed over the genome but differs between regions. The 5' NCR is the most highly conserved region of the HCV genome, which is the usual locus of oligonucleotide primers used for detection of HCV RNA by PCR method. The putative core and NS3 region are also conserved and antigens from this region are used in anti-HCV detection assay (29).

HCV Genotype

HCV is classified on the basis of the similarity of nucleotide sequences into major genetic groups designated genotypes. Comparison of the publish sequences HCV has led to the identification of several genotypes that may differ from each other by as much as 33% over the whole viral genome (30). HCV genotypes are numbered in the order of discovery. The more closely related HCV genome within some genotypes is designated subtypes, which are assigned in the order of discovery such as 1a, 1b, 2a, 2b. The complex of genetic variants found within an individual isolate are termed the quasispecies, which particularly in the variable regions of the genome. The genomic sequences of different HCV isolates vary by as much as 35% (7). The degrees of difference in nucleotide sequences among isolates vary from one genomic region to another. At least 6 genotypes and more than 100 subtypes have been identified worldwide. Some genotypes are endemic worldwide, while others maybe restricted to distinct geographical region (6). Although the HCV genotype 1, 2, and 3 appear to have a worldwide distribution, the geographic distribution of HCV genotypes has been report as prevalence varies from one geographic area to another. Approximately 70% of HCV infected patients in the United States are infected with genotype 1a and 1b, whereas genotype 4 is found commonly in the Middle East (31). Genotype 5 is found commonly in South Africa, and genotype 6 is found in Hong Kong, Vietnam and Thailand (32). Genotype 3

is most prevalent in India, Bangladesh, and other parts of Asia (33). The geographic distribution of HCV genotypes from blood donors from four major regions of Thailand was studied. HCV genotype 3a was the most prevalent genotype, followed by genotype 1b and genotype 6 group variants (9,34).

Table1. Terminology commonly used in studies related to HCV genomic heterogeneity (35).

Terminology	Definition	% Nucleotide similarity
Genotype	Genetic heterogeneity among different HCV isolates	65.7-68.9
Subtype	Closely related isolates within each of the major genotypes	76.9-80.1
Quasispecies	Complex of genetic variants within individual isolates	90.8-99

Compared with other variants of HCV, genotype 1 is associated with a poor treatment response and outcome. Genotype 1 is also associated with more severe liver disease and a high incidence of liver cancer (36, 37). Response rates to therapy are lower in patients with genotype 1, as compared to patients infected with other genotype (11). The efficacy of treatment regimens differs on the basis of genotype.

Pathogenesis

Antibody responses are often directed against the viral envelope proteins that serve as neutralizing antibodies. Multiple protein antigens encoded by the viral RNA seem to produce serologic responses in host. The specific type of immune response, humoral or cellular, that associated with protection and clearance of HCV after an acute exposure, the response appears to be type specific. The antibodies are likely directed against epitopes of hypervariable region 1 located in the E2 region and that persistent infection depends on the ability of the virus to continually evade the effects of neutralizing antibody. Due to its

variability, HVR1 has been used extensively as an indicator of viral evolution. The quasispecies of HCV showed the selection of strains to avoid immune pressure (38).

The important components of protective immunity, cytotoxic T lymphocytes (CTLs), against many viral infections. Cytotoxic T lymphocytes (CTLs) are thought to be one of the major hot defense arms against viral infection (39) and are also implicated in the immunopathogenesis of viral infection. The hepatitis C virus specific CTLs have been demonstrated in both the peripheral blood and among liver infiltrating lymphocytes of patients with chronic hepatitis C (40). Viral clearance is associated with the development and persistence of strong virus specific responses by cytotoxic T lymphocytes and helper T lymphocytes (41). The HCV heterogeneity may also be importance in escaping CTL-induced immunity. The single amino acid changes in CTL epitopes result in failure of recognition by HCV specific CTLs (42). These single amino acids changes are found in HCV isolates, hence the need to address the problem of type specificity of immune response. However, the immune responses to the initial strain did not protect against an infection with another strain of HCV (43,44). The HCV infects and replicates in hepatocytes and peripheral blood lymphocytes. The circulating HCV specific CTLs also would expose to hepatocytes that express HCV antigen. CTL response to HCV may not be strong enough to prevent progression of disease in HCV infection. Sustained insufficient CTL responses to HCV may incompletely suppress the outgrowth of HCV and cause chronic liver injury by persistently damaging HCV infected hepatocytes (45). However, a strong CTL response may eventually eliminate HCV from infected individuals and lead to healing of liver injury. The development of vaccine for HCV has been hampered, at least partly, by the great heterogeneity of the HCV genome.

Epidemiology

The prevalence rate among blood donor in Northern Europe is 0.2% to 0.5% and 0.5% to 1.5% in Southern Europe, the United States, and Japan (46, 47). In Southeast Asia, the seropositive rate survey from Indonesia, Malaysia, Vietnam, and the Philippines are between 1.5% to 3.0% (47). The prevalence of antibodies to HCV in Thai population was 1.5% to 5%(48,49). HCV is transmitted mainly by intimate contact with virus contaminated blood. Most hemophilic patients received clotting factors before advent of viral inactivation procedures became infected with HCV (50). Transplantation of organs from infectious donors to the organ recipient also carried a high risk for transmitting infection before donor screening (51). Number of new infections per year has declined from an average of 240,000 in the 1980s to about 25,000 in 2001. Most infections are due to illegal injection drug use. Tattooing and acupuncture have also been reported to be two transmission routes. Needle accidents in health care workers inoculated with blood contaminated needles have led to HCV transmission, but the rate of infection is rather low (52). Transfusion associated cases occurred prior to blood donor screening; now occurs in less than one per million transfused unit of blood. Sexual transmission is possible but rare (53,54). Transmission of HCV infection from mother to infant is uncommon and appears to be more likely if mother is positive for both anti-HCV and HCV RNA than if she is positive for anti-HCV only. The risk of transmission may be much greater if the mother is co-infection with HIV (55). The incubation period of transfusion associated NANBH is 2 to 26 weeks with a mean of 8 weeks (56).

Other clinically important syndromes include coinfections with other viruses, especially HIV-1 and other hepatitis viruses. HIV-1 positive patients were coinfected with HCV in 33% and rose to 75% in the patients with known injection drug used (57).

Signs and symptoms of Hepatitis C Virus (HCV) infection

HCV infection occurs among persons of all ages, but the highest incidence rate of acute hepatitis C is found among persons aged 20 to 39 years, and males predominate. Clinical manifestations can occur within 7 to 8 weeks after exposure to HCV. Hepatitis C is usually clinically mild, with only minimal to moderate elevation of liver enzymes. Persons with acute infection might have jaundice and 10%-20% might have non-specific symptoms such as anorexia, malaise, or abdominal pain (58). However, most patients are asymptomatic but histologic evaluation often reveals evidence of chronic active hepatitis, especially in those whose disease is acquired following transfusion. The various ALT patterns have been observed in patients during follow-up and they might have prolonged periods of normal ALT activity even though they have histologic confirmed chronic hepatitis. These persons include anti-HCV positive with persistently elevated ALT levels, detectable HCV RNA, and a liver biopsy that indicates either portal or bridging fibrosis or at least moderate degrees of inflammation and necrosis (3). In many cases HCV causes persistent infection. Persistence maybe facilitated by the relatively poor immunogenicity of the virus and the low level viraemia. Change in viral antigen, due to mutations in HVR-1, may also contribute to persistence. However, genetic drift in HCV may be independent of the host's immune pressure. An estimated 74 to 86% of persons will have persistent viremia (59). This range may prove to be low as more sensitive tests become available to detect viremia. Most studies have reported that cirrhosis develops in 10% - 20% of persons with chronic hepatitis C over a period of 20-30 years (60). When cirrhosis is established, the rate of development of HCC might be as high as 1% - 4% per year (61). However, drinking alcohol can make liver disease worse.

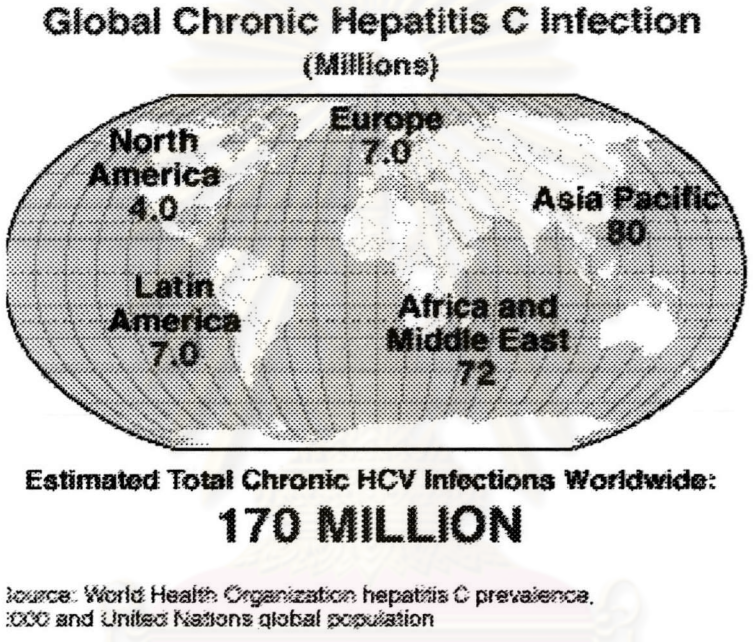

Figure 3. Global Chronic Hepatitis C Infection. HCV infects an estimated 170 million persons worldwide. Most of infections were in the third world or developing countries (61).

Diagnosis

The identification and mapping of the HCV genome was the first step in the development of reliable and accurate diagnostic assays for HCV infection. Diagnostic tests for HCV infection are divided into serologic assays for antibodies and molecular tests for viral particles. To detection of HCV antibodies, immuno assays were used. HCV antigens are coated on the wells of microtiter plates and used to capture and measure the circulating HCV antibodies (62). Some infected patients showed HCV antibodies within 15 weeks of exposure. However, there is a window period during which an individual might be infectious but seronegative.

The detection of HCV RNA aids the diagnosis of HCV infection in many methods. Direct detection of HCV is essential for the correct diagnosis of HCV infected patients. A reliable and reproducible extraction and amplification method may have high diagnostic value. Although there usually not enough viral RNA to permit detection, the application of PCR techniques to amplify reverse transcribed cDNA permits a very sensitive assay for viral RNA circulating in the bloodstream and in the tissue biopsy specimens. HCV RNA can be detected in serum or plasma within 1-2 weeks after exposure to the virus before elevations of circulating ALT and viral antibodies levels and can be used to diagnose HCV infection in chronic NANB hepatitis patients who may be seronegative. PCR makes it possible to detect virus replication in immunocompromised patients during the window period (63). Most RT-PCR assays have a lower limit of detection of 100-1000 viral genome copies per mL. To minimize false-negative results, serum must be separated from cellular components within 2-4 hours after collection, and preferably stored frozen at –20°C or –70°C (64). To perform the PCR assay, it is clear that primer specific for 5' NCR region has been used in most laboratories to develop sensitive RT-PCR based detection assays for HCV RNA since this region show highly conserved among all HCV isolates in many studies (65).

Quantitative assays for measuring the concentration of HCV RNA have been developed, including a quantitative RT-PCR and a branched DNA signal amplification assay. These assays should not be used as a primary test to confirm or exclude the diagnosis of HCV infection or to monitor the endpoint of treatment. Testing for the level of HCV RNA might help predict likelihood of response to antiviral therapy, although measurement of HCV RNA levels has not proven useful in managing patients with hepatitis C (41).

Table 2.Recommendations for testing based on risk for HCV infection (3).

Person	Risk of infection	Testing recommended
Recipients of clotting factors made before 1987	High	Yes
Injecting drug users	High	Yes
Recipients of blood and/or solid organ before 1992	Intermediate	Yes
People with undiagnosed liver problem	Intermediate	Yes
Infants born to infected mothers	Intermediate	After 12-18 months old
Hemodialysis patients	Intermediate	Yes
Healthcare/public safety workers	Low	Only after known exposure
People having sex with multiple partners	Low	No
People having sex with an infected partner	Low	No

Treatment of HCV

Treatment of HCV is apparently influenced by the specific genotype involved, the level of viraemia, interferon dose, duration of treatment and liver histology. Knowledge of the HCV genotype is important because it has predictive value in terms of the response to antiviral therapy. All seem to play a role in predicting response. The individual who infected with genotype 1a or 1b is less likely to have a favorable response to interferon treatment than those who infected with genotype 2 or 3. Because HCV genotype 1b is resistant to interferon, the rate of complete response is only 10 to 40%, which is much lower than those of the other genotypes, such as genotype 2 or 3, with the rates of complete response of 60 to 90%. Interferon and ribavirin are two drugs licensed for the treatment of patients with chronic hepatitis C who are at risk for progression to cirrhosis Interferon can be taken alone or combination with ribavirin. Monotherapy for HCV infection with α-interferon was associated with initial rates of response as high as 40%. (66).

The treatment consists of 3 million Units of α-interferon administered subcutaneously three times a week. The attachment of polyethylene glycol to α-interferon (α- peginterferon) extends the half-life and duration of therapeutic activity of α-interferon. In contrast to α-interferon, α-peginterferon is given only once a week. However, the combination therapy with interferon and ribavirin, a nucleoside analogue, is approved for the naïve treatment of patients with chronic hepatitis C. The virologic response to combination therapy should be assessed at week 24, since elimination of the virus can occur. The individual who infected with genotype 2 or 3 who have a negative PCR assay for HCV RNA can also stop therapy at this time. An additional 24 weeks of treatment is suggested for patients with other genotypes and a negative PCR assay. Combination therapy, using pegalated interferon and ribavirin, can get rid of the virus in up to 40% of those genotype 1 and up to 80% for those with genotype 2 or 3. However, after therapy stopped, some of responder relapsed with a return of serum ALT levels to the pretreatment range or higher (18,66).

Patients treated with interferon may experience some trouble side effect. Early common side effect is a flu-like which usually disappears within 3 weeks. Other side effects include fatigue myalgias, leucopenia and thrombocytopenia, psychiatric syndromes often with depression and autoimmune phenomena resulting in arthritis, thyroidtis, or hemolytic anemia may develop (67,68).

HCV Genotyping assay

HCV has high genomic variability and at least six different genotypes and an increasing number of subtypes have been reported. Since many study indicated that the genotypes of HCV are associated with the development of hepatocellular carcinoma and seem to play a role in predicting response to α-interferon therapy. Many methods have been used to differentiate the HCV genotype in the clinical laboratory setting based on, such as probe hybridization, type specific primers or Restriction Fragment Length Polymorphism (RFLP). However, the reference method for HCV genotyping is sequencing of a specific PCR-amplified portion of the HCV genome and followed by phylogenetic analysis (69). Many regions of HCV genome were used to determine the HCV genotype such as core region, envelope region, NS5B and 5' non coding region. In 1995, Sukiyama, K., et al. determined the HCV genotype by RT-PCR for portion of the HCV 5' non coding region, HCV core region and NS5 region that can be used commonly for HCV genotyping. Some of the HCV RNA failed to detect by RT-nested PCR for a portion in the NS5. These failures may be due to a variation of NS5 (70). Additionally, Seme, K., et al. proposed that the two genotyping methods, genotyping specific primers and nested polymerase chain reaction (nested-PCR), based on the analysis of 5'NCR were found more sensitive than those methods based on the core regions. The specificity of genotyping methods was due to a lack or excess of sequence variation in the target region (71). Because there is a high degree of sequence conservation found within the 5' NCR of the genome that making this

region a target of choice for most nucleic acid amplification based detection assay and was developed for identifying HCV genotypes and subtypes.

On the basis of reverse hybridization, line probe assay (LiPA) was developed. Stuyver, L., et al. proposed that 5' non coding region show 7 variable areas of nucleotide base between each genotype. There are R1 = nucleotides -240 to -233, R2 = nucleotides -167 to -155, R3 = nucleotides -147 to -142, R4 = nucleotides -138 to -132, R5 = nucleotides -128 to -118, R6 = nucleotides -100 to -92 and R7 = nucleotides -81 to -70. These 7 variable 5' non coding region areas were applied to a LiPA. The biotinylated PCR fragments are hybridized to a selection of highly specific immobilized probes that fixed on the strip and then the biotin group in the hybridization complex is revealed by incubation with a streptavidin-alkaline phosphatase complex and the appropriate chromogen compounds. The results indicted that the LiPA is a reliable assay applicable to routine genotyping and subtyping of HCV specimen (72). However, the interpretation of the HCV genotype should be used the present of different pattern lines on strip, some lines not clear. Then, some patterns could not be interpreted.

The alignment technique is used to demonstrate homology between new sequences and existing families of sequences. Related groups of organism are grouped and then determine their genotype. Phylogenetic analysis was used to classify into genotypes and subtypes after computer assisted analysis of the sequence data (73).

Phylogenetic Systematics

Phylogenetic Systematics is the field of biology dealing with identifying and understanding the evolutionary relationships among the many different kinds of organism. A special type of graph that called phylogenetic trees presents the relationship pattern among organism. The aim of phylogenetic classification is to group organisms according to their ancestral lineage that can be studied in two methods, phenetic and cladistic. Phenetic

method, each organism is compared with every other for all characters measured and the number of similarities or differences is calculated. Then, the organisms are grouped in a way that the most similar are grouped close together and the more different ones are linked more distantly. However, the interest in and use of phenetics method has been declining in recent years because of the convergent evolution. The results of this method do not necessarily reflect genetic similarity or evolution reletedness. The lack of evolutionary significance in this method has had little impact on organism classification. In cladistics method, groupings do not depend on whether organisms share physical traits but on their evolution relationships. Related groups of organism are grouped because they share a set of unique features which were not present in distant ancestors. All organisms in a group must share a common ancestor. The amount of nucleotide sequence difference between a pair of genomes from different organisms should indicate how two genomes shared a common ancestor. Two genomes that diverged in the recent past should have fewer differences than two genomes whose common ancestor is more ancient. Then, this method attempts to determine the rates and patterns of change occurring in DNA or protein sequence.

A phylogenetic tree is composed of nodes and branches. Node represents an individual unit or different organisms. This can be either an existing species or and ancestor. The relationships between them are shown as branches, links or edges. The lengths of the branches may or may not be significant. That is, the lengths of branches can be used to indicate the actual evolutionary distances between organisms. A group of two or more organisms or nucleotide sequences that includes both their common ancestor called clade. Then, species are represented by organism that from a clade, genera by those that from a larger clade, families and orders by those that from still larger clade, and so on. As same as this way, phylogenetic trees can also be use to classify the genotypes and subtypes of organisms in species. Phylogenetic trees may also be rooted and unrooted. In rooted trees, there is a node called root that represents a common ancestor. An unrooted

tree only presents the relationship among species, without identifying a common ancestor or evolutionary path.

Phylogenetic analysis consists of four steps.

1. Alignment.
2. Determining the substitution or mutations model of sequence that can consider sequence variation.
3. Phylogenetic tree construction.
4. Phylogenetic evaluation.

Multiple alignment is used to demonstrate homology between new sequences and existing families of sequences. All sequence alignments are carried out to determine the most related sequence. The most similarity clusters or minimum number of mutation sequence would be defined as branches. The positions of the gaps that were introduced during the early alignments of the closely related sequences are not changed as new sequences are added.

Phylogenetic trees can be constructed from different methods.

1. Distance matrix method. The most common method that used to construct a phylogenetics tree based on the distance matrix methods. A distance table is regarded as an nxn matrix and determine the differences between each possible pair of cluster. The number of cluster in the matrix is reduced by one. Then, select the two most similar clusters, that is, the distance function or distance score is minimal. Repeat the analysis until there is only one cluster left.
2. Maximum parsimony method. The phylogenetic trees are constructed on the basis of the minimum number of mutations required to convert one sequence into another. Grouping those sequences that can be interconverted with the smallest number of overall changes generates the trees.

3. Maximum likelihood method. The methods also involve multiple sequence alignment and the analysis of changes at each position of the sequence. The likelihood of different sequence changes at each position is calculated, and these values are multiplied to provide an overall likelihood for each tree. The most reliable tree is that with the maximum likelihood.

The reliability of phylogenetic trees can be test in many ways. Bootstrapping is the one technique that used to test the reliable. If different methods of tree construction give the same result or the data can be resampled to test their statistical significance, this is good evidence that the tree is reliable. The tree constructed by bootstrapping should always match the original tree (74, 75, and 76).

A large number of software packages are available. Some popular programs are versatile and allow distance matrix, parsimony and maximum likelihood method analysis can be carried out. Some programs are free available programs that located on the Internet. In this research, Clustal W was used. Clustal W is a fully automatic program for global multiple alignment of DNA or protein sequences. The program consists of multiple alignment and phylogenetic analysis. In this format, the links are still represented by lines but the ancestral nodes are represented by vertical lines rather than boxes. The scores are calculated as the number of identities in the best alignment divided by the number of residues compared (gap positions are exclude). In Clustal W, before any pair of sequences or prealigned groups of sequences, a table of gap opening penalties for every position in the two sequences. The cDNA sequences have been compared to determine whether the relationship existing between them could have occurred by chance. The smallest unit of comparison is a pair of nucleotide. The maximum match can be defined as the largest number that can be matched while allowing for all possible deletion or insertion (77).