Review of Literature

2.1 Human Papilloma Virus Classification

Papillomavirus are highly diverse and widespread in nature. They have been recognized primarily in higher vertebrates and are likely to occur in most mammals. Papillomavirus are highly species specific and there is no firm evidence of a papillomavirus from one species causing productive infection in another species. Over 100 human papillomavirus types (HPVs) have been described. A HPV type is defined as a genome whose major capsid protein nucleotide sequence differs from the homologous nucleotide sequence of every other HPV-type by at least 10 percent. Subtypes have 90 to 98 percent sequence similarity to the corresponding type and variants show more than 98 percent sequence homology to the prototype (De Villiers et al., 2004). In an evolutionary perspective HPVs can be separated into a number of groups or genera shown in Figure 2.1. Papillomaviruses are separated according to genotype i.e. differences in the open reading frame of the highly conserved L1 major capsid protein. L1 sequences that share less than 60 percent nucleotide sequence identity belong to different genera. Papillomaviruses within a species share between 71-89 percent nucleotide identity within the L1 open reading frame. The human papillomaviruses are found in the Alpha-, Beta-, Gamma-, Mu- and Nu-genera. The other genera contain papillomaviruses isolated from various mammals and birds (De Villiers et al., 2004; Doorbar, 2006; Forslund, 2007).

HPVs are highly trophic and infect basal epithelial cells of the skin or mucosa and can further be classified into cutaneous or mucosal types. Beta papillomavirus are typically associated with in apparent cutaneous infections in humans; whereas the Alpha papillomavirus group mainly contains the genital/mucosal HPV types. Within each of these groups they can further be designated as high-risk or low-risk according to the propensity for malignant progression of the lesions they cause. Low-risk HPV types 6, 11, 42, 44, 53 and 83 are associated with benign warts in the genital tract (condylomata acuminate). High-risk HPV types 16, 18, 31, 33, 34, 35, 39, 45, 51, 52, 54, 56, 58, 59, 66, 68 and 69 are usually found in malignant lesions (Munoz et al., 2006).
Figure 2.1: Classification of HPV as taxonomic scheme developed by International Committee on Taxonomy of Viruses

2.2 Association with cervical cancer

Harald zur Hausen's laboratory was the first to demonstrate that genital warts contain human papillomavirus (HPV) genomes (Gissmann and Hausen, 1980; De Villiers et al., 1981). Subsequent low-stringency hybridization experiments with HPV sequences isolated from genital warts performed in his laboratory led to the discovery of related HPV sequences in cervical cancer tissues (Durst et al., 1983). Human papillomavirus is a highly prevalent, sexually transmitted infection and has been shown as an etiologic agent for cervical cancer (Munoz et al., 2006). With approximately 500,000 newly diagnosed cases each year and a 50 percent mortality rate, cervical cancer is the second most common cause of cancer-related death in women worldwide (Stanley, 2006).

Oncogenic HPV infection is a necessary but not a sufficient cause of cervical cancer. There are several other factors that may modulate the risk of transition from HPV infection to cervical cancer development in conjunction with HPV risk factors.
tumour development include: persistent infection with high-risk viral types, a large number of lifetime sexual partners, co-infection with human immunodeficiency virus, immune suppression, parity, long term use of oral contraceptives and cigarette smoking (Dillner and Brown, 2004; Steben and Duarte-Franco, 2007). While most cases of cervical cancer are caused by HPV infections, not all women that are infected with high-risk HPV will develop cervical cancer.

In most cases, the immune system is able to fight off the infection. Cervical cancer development is a multistep process. It often takes many years from HPV infection to cancer development and requires both the presence of oncogenic HPV genotypes and the interaction of many host factors (Munoz et al., 2006). The latency period between initial HPV infection and the cancer can be more than 10 years (Frazer, 2007). The premalignant stages can be identified both clinically by speculum examination and in the laboratory by papanicolaou smear analysis.

More than 35 HPV types regularly or sporadically infect the genital tract. Among these types, HPV16 is the most prevalent and accounts for 50 to 60 percent of the cervical cancer cases in most countries, followed by HPV18, who accounts for 10 to 12 percent. The four most common oncogenic HPV types (HPV16, 18, 31, and 45) were found in together about 80 percent of squamous cell carcinomas and HPV types 16, 18, 45, 59 and 33 together account for 94 percent of HPV-types found in adenocarcinomas (Dillner and Brown, 2004). Subsets of HPV types have been established as the causative agents of cervical cancer, since 99 percent of tumours are positive for HPV DNA (Munoz et al., 2006). Even though HPV infections are common among younger women of the age group 19-35 years, cervical cancer is more common in women over the age of 35 years (Wright et al., 2006; Dunne et al., 2007; Schiffman et al., 2007).

2.3 Genomic organization
HPVs are DNA viruses with a double-stranded closed circular genome of approximately 8 kilo bases (Seedorf et al., 1985). In the Figure 2.2 the genome can be divided into three major regions: i) a ~1.0 kb non-coding long control region (LCR) that contains a variety of cis elements that regulate gene expression and viral replication. LCR has also been referred to as the upstream regulatory region
(URR), ii) a ~4.0 kb early (E) region, that encodes non-structural proteins, iii) a ~3.0 kb late (L) region, that encodes the two capside proteins (Munger et al., 2004). All papillomavirus have a similar genomic organization. Viral genes can be divided into early or late categories dependent on the time of expression. The HPV genome encodes six early proteins (E1, E2, E4, E5, E6 and E7) and two late proteins (L1 and L2). The late genes encode the viral capside proteins whereas the early genes encode proteins involved in viral DNA replication, transcription and cellular transformation (Hebner and Laimins, 2006).

The HPV16 genome contains two major promoters. The early promoter, p97, initiates Transcription upstream of the HPV16 E6 open reading frame, while the differentiation dependent late promoter, p742, located in the HPV 16 E7 open reading frame is activated during the late phase of the HPV life cycle (Ozbun and Meyers, 1997; Klumpp and Laimins, 1999; Doorbar, 2006).

Figure 2.2: Genomic organization of human papillomavirus type 16
2.4 The viral proteins E7
E7 is the main transforming protein. One of its main functions in the HPV life cycle is the binding and degradation of the retinoblastoma (Rb) tumour suppressor family of proteins. The Rb proteins are major regulators of the cell cycle. Hypophosphorylated Rb controls the transition at the G1/S phase of the cell cycle through binding of the E2 family of transcription factors that activate transcription of many components involved in S-phase replication. In normal cells, phosphorylation of Rb by cyclin-kinase complexes leads to a release of E2Fs and transcription of S-phase genes. E7 can override this normal cell cycle control through binding and degradation of hypophosphorylated Rb and thereby constrictively releasing E2F-complexes (Boyer et al., 1996; Jones et al., 1997; Hebner and Laimins, 2006).

Both high-risk and low-risk E7 proteins have been shown to bind to the Rb protein, but low-risk E7 proteins bind with a lower efficiency. As a result the low-risk E7 proteins do not efficiently transform cells (Munger et al., 1989; Gage et al., 1990). E7 can interact with many other cellular proteins, most of which are important regulators of the growth, but the actual biological relevance is uncertain. Another consequence of E7 expression is the induction of genomic instability, as seen in many malignancies. The E7 has been shown to induce abnormal centrosome duplication resulting in multipolar, abnormal mitoses and aneuploidy. The low-risk E7 proteins do not induce similar abnormalities (Duensing et al., 2000; Duensing et al., 2001).

2.5 HPV 16 E7 life cycles
At present there is some controversy as to the precise nature of the receptor for virus entry, but it is thought that heparin sulphate proteoglycan plays a role in initial binding and/or virus uptake (Joyce et al., 1999; Shafti-Keramat et al., 2003). Virus particles are taken into the cell relatively slowly (Culp and Christensen, 2004) and for HPV16 this occurs by clathrin-coated endocytosis. This mode of entry may not be conserved among all HPV types.

Virus particles disassemble in late endosomes, and the viral DNA is transported to the nucleus by the minor capsid protein L2 (Day et al., 2003; Day et al., 2004). Human papillomavirus enters the basal epithelial cells of the skin and mucosa. Infection is thought to occur through micro wounds of the epithelium that expose cells in the
basallayer (the only cell layer in an epithelium that is actively dividing) to viral entry. The HPV life cycle is tightly linked to the differentiation program of the infected host cell, the keratinocytes, with production of mature virions particles restricted to the differentiated suprabasal cells (Munoz et al., 2006). Infection leads to the establishment of the viral genome as a stable episome (without integration into the host cell genome) in cells of the basal layer and the early promoter is activated. In these infected cells, low levels of viral DNA synthesis occur. Viral episome replication occurs synchronously with host cell chromosome replication. Since HPVs do not encode enzymes for DNA replication, production of viral genomes is critically dependent on the host cellular DNA synthesis machinery (Schiffman et al., 2007).

After division of basal cells one daughter cell remains in the basal layer while the other one move up and starts to terminally differentiate. HPV episomes are also replicated and distributed evenly among daughter cells. Differentiated cells contain little or no replicative machinery; therefore the virus will be unable to propagate if cells are permitted to terminally differentiate. Therefore the virus expresses proteins that stimulate G1/S progression. However, the virus still requires a certain level of differentiation because the shift from early to late promoter, which transcribes capsids protein mRNA, is mediated in differentiated cells. As HPV positive cells differentiate, the late promoter is activated leading to expression of late genes and new virions are produced. The factors determining promoter shift have not yet been determined. In the uppermost layers of the epithelium DNA is packaged into newly formed virus capsids and shed into the environment as a cargo within epithelial squamae (Munger et al., 2004; Munoz et al., 2006).

2.6 Principles of HPV vaccine for cervical cancer
Two types of HPV vaccine are under development: (A) prophylactic vaccines to prevent HPV infection and associated lesions, and (B) therapeutic vaccines to induce regression or remission of established pre-cancer lesions or cervical cancer. Current prophylactic HPV vaccines are based on purified virus-like particles (VLPs), i.e. the viral capsid, without the viral DNA, composed of the main envelope protein L1 of the oncogenic HPV types. Risk of infection with HPV is directly proportional to age. Prophylactic HPV vaccines are found to be effective in treating HPV infection and this is shown by a number of randomized clinical trials (Munoz et al., 1996; Stanley,
Gardasil, a prophylactic vaccine shows maximum efficacy if used by girls’ pre-exposure to HPV. Prophylactic HPV vaccines are associated with VLPs and are made up of HPV L1 proteins at present. As VLPs, has no viral DNA therefore these are not associated with the development of infection even though their structure is similar to actual virus morphologically. By intra-muscular (I/M) injection, high levels of systemic anti HPV L1, immunoglobulin G antibodies are formed due to VLPs (Dunne et al., 2006). Recently, Peptide vaccines have shown the advantages of safety and ease of production; however, their weak immunogenic properties and the need for HLA matching must be overcome.

2.7 Gene expression profile of HPV 16 E7 in squamous cervical carcinoma

Genomic destabilization is a key process for cancer development. Profile of HPV-induced carcinogenesis is the integration of the HPV genome into a host chromosome. HPV genome integration often occurs near common fragile sites of the human genome (Thorland et al., 2003), but there are no apparent hot spots for integration and no evidence for insertional mutagenesis (Ziegert et al., 2003). Integration follows a more specific pattern with respect to the HPV genome.

The gene expression of HPV16 E7 oncoprotein in primary human epithelial cells causes genomic instability. High-risk HPV E7 oncoprotein independently induces genomic instability in normal human cells (White et al., 1994). They cooperate to generate mitotic defects and aneuploidy through the induction of centrosome abnormalities in normal human epithelial cells, and the characteristic multipolar mitoses in cervical lesions are caused by centrosome abnormalities (Duensing et al., 2000). In contrast, low-risk HPV E7 proteins are not capable of inducing centrosome abnormalities. Centrosome abnormalities and associated mitotic defects are apparent in cells that, similar to low-grade HPV-associated lesions, express episomal HPV-16 at a low copy number (Duensing et al., 2001), and their incidence increases in cells with integrated HPV (Pett et al., 2004). Centrosome abnormalities have also been detected in cervical cancer.

The gene expression profile in HPV 16 positive cervical carcinoma was done by genome wide analysis. Therefore the human whole genome codlink microarrays were
used, which have the more sensitivity to detect transcript unit per cell to also distinguish amongst highly homologous sequences. This array contains a broad range of genes derived from publicly available m-RNA sequences (over 55,000 genes and ESTs probe) (Ramakrishnan et al., 2002). After cluster filtering 6,007 probes corresponding to 3,248 characterized transcripts were obtained. These data are reanalyzed using a supervised approach with significance analysis of micro-array (SAM) (Tusher et al., 2001), which predict 1,007 and 1,060 genes that were significantly up and down regulated (Carlous et al., 2007).

2.8 Effects of HPV E7 on cellular metabolism

A series of intriguing experiments has suggested that expression of high risk HPV E7 proteins can directly induce fundamental cellular metabolic alterations that also commonly occur in a variety of cancer cell lines. The glycolytic enzyme M2 pyruvate kinase (PK) can adopt a highly active tetrameric form with a high affinity for phosphoenolpyruvate, and a dimeric form with a low affinity for this crucial metabolite. Like many cancer cell lines, HPV E7 transformed cell lines contain increased amounts of the dimeric form of M2 PK and show a decreased glycolytic flux. Dimeric M2 PK persists in E7 expressing cells despite high levels of fructose 1, 6 bisphosphate that normally stimulate the formation of the high affinity, tetrameric form. Yeast two hybrid analyses revealed an interaction between E7 and M2 PK, and it has been hypothesized that E7 may directly disrupt the normal metabolic regulation of M2 PK (Zwerschke et al., 1999; Mazurek et al., 2001).

In addition, HPV-16 E7 has also been reported to interact with and allosterically activate α-glycosidase (Zwerschke et al., 2000). Activation of this enzyme results in the depletion of intracellular glycogen stores, similar to what has been observed in human cancers (Bannasch et al., 1997) including cervical lesions (Pedersen et al., 1975). The activation of these enzymes in E7 expressing cells may drive cellular hyper proliferation are alternatively represent a consequence of the altered energy requirements of rapidly growing transformed cells.

2.9 Human TMEM 50 A protein

Transmembrane protein 50A is a protein that in humans is encoded by the TMEM50A gene (Flegel and Wagner, 2000). The TMEM50A gene is conserved in chimpanzee,
Rhesus monkey, dog, cow, mouse, rat, chicken, zebra fish, fruit fly, and mosquito. This gene is located on chromosome 1 p36.11 in the human (Homo sapiens) genome. Its mRNA sequence is 2284 nucleotides in length and includes seven exons. The coding sequence is from base pairs 151 to 624. It encodes TMEM 50 A Human protein expressed in almost all human tissues, but evidence from EST profiles through NCBI, suggests that its expression may be slightly higher in parathyroid tissues and brain tissues. It also seems to be expressed higher during the neonate and juvenile development stages. The EST is one of 1043 sequences matched to TMEM 50 A. Its ORF name is represented as “UNQ386/PR0718”. The cellular location of this protein was predicted by using PSORT II that predicts TMEM50A is most likely found in the cells plasma membrane or the endoplasmic reticulum. TMHMM shows that TMEM50A has four transmembrane regions. This was further confirmed by similar results found in TMEM50A orthologs and the neutral charge found in these regions using SAPS program in biology workbench. Investigation of several GEO profiles showed that TMEM50A is highly up regulated in late stage cervical cancer. This may suggest that TMEM50A has some function that may be causing or is caused directly by cervical cancer. Although few studies are available to confirm this idea, more studies may offer suggestions that use TMEM50A for treatment of late stage cervical cancer.

2.10 Bioinformatics approaches

The various aspects of human papilloma virus 16 genome that are discussed in this thesis can be analyzed with the help of bioinformatics and computational methods. Numerous databases are now available which contain both sequence and functionality information. Most of these are accessible over the internet through convenient web browser interfaces. Many also permit retrieval of sequence information using online servers. Sequence databases now contain the nucleotide and predicted amino acid sequences of virtually every gene in model microbes. The genome sequence data of HPV 16 can be downloaded from NCBI. Major biological databases, which can be used, are DDBJ, SwissProt, and PDB. Also there are various online as well as offline softwares like COGs, BLASTP, T-COFFEE, GenTHREADER, ExPASy tools, Phylogenetic analysis tools.
2.11 In-Silico proteome analysis of HPV 16 E7 and human TMEM 50 A

The proteome analysis of human papilloma virus 16 E7 and human TMEM 50 A was carried out by using bioinformatics approaches for understanding the various biological mechanisms involved in genome annotation, metabolic pathways activities etc. The putative uncharacterized proteins can be annotated with the sequence, motif search, Domain prediction, protein-protein interaction, Sub cellular information, Signal peptide, physiochemical properties and structural information with the help of different bioinformatics approaches. This also includes BLAST searches to determine identities/similarities or homologous to known proteins, the protein sequence is examined for the presence of functional domains using some databases PROSITE, PRINTS, InterPro Scan, ProDom, Pfam and SMART, subjected to searches for motifs and protein secondary structure prediction (Gert et al., 2005). Analysis of such proteins will enhance our understanding in designing and development for diagnosis of cervical cancer caused by human papilloma virus 16 E7 antigen proteins.

2.12 Immunoinformatics and vaccine designing

The accelerating growth of bioinformatics techniques and applications along with the substantial amount of experimental data has made a significant impact on the immunology research. This has led to a rapid growth in the field of computation immunology, and a number of immunology-focused resources and software, which help in understanding the properties of the whole immune system, have become available (Brusic and Petrovsky, 2005). This has given rise to a new field, called immunoinformatics. Immunoinformatics can be described as a branch of bioinformatics concerned with in silico analysis and modeling of immunological data and problems. Immunoinformatics research stresses mostly on the design and study of algorithms for mapping potential B- and T-cell epitopes of HPV type 16 E7, which speeds up the time and lowers the cost needed for laboratory analysis of pathogen gene products. Using such tools and information, an immunologist can analyse the sequence areas with potential binding sites, which in turn leads to the development of new vaccines. The methodology of analysing the Hpv genome to identify potential antigenic proteins is known as ‘reverse vaccinology’ (Tomar and De, 2010).

This is mainly beneficial because conventional methods need to dedicate time to HPV pathogen cultivation and subsequent protein extraction. Although pathogens grow
quickly, extraction of their proteins and then testing of those proteins on a large scale is expensive and time-consuming. Immunoinformatics is capable of reducing time and saving resources for the development of relevant vaccines by revealing virulence genes and surface-associated proteins. Vaccines are agents—either molecular or supermolecular which can stimulate protective immunity against HPV 16 E7 and cervical cancer which they cause. There are two main forms of support for vaccine discovery provided by bioinformatics. The first is technically indistinguishable from support for more general target discovery, and include genomic annotation for both host proteins humanTMEM 50 A and humanpapilloma virus 16 E7 protein sequences and immunotranscriptomics, the application of microarray analysis of the immune system. The other kind of support is focused on immunoinformatics, addresses problem such as the accurate prediction of epitopes. Currently, prediction of T-cell epitopes is essentially confined to predictions of varying accuracy of peptide binding of major histocompatibility complex. Binding of peptide to classI MHC are reasonably accurate, at least for well characterised alleles (Lafuente and Reche, 2009).

However, several comparative studies have shown recently that the prediction of class II MHC binding prediction T-cell epitopes is typically poor (EI-Manzalawy et al., 2008; Linet et al., 2008), and same as for structure-derived prediction of MHC-mediated epitopes (Knapp et al., 2009; Zhanget et al., 2010). The prediction of B cell epitopes remains primitive or depends on an exclusive knowledge of protein structure, methods for T cell epitope prediction displaying not inconsiderable algorithmic sophistication have been and continue to be developed. IEDB is the dominant force in current immunoinformatics and stands as one of the principle achievements of the field. The number and quality of epitopes within it are an enormous improvement over all that has gone before (Vita et al., 2010).

2.13 In-Silico Epitope prediction
The in silico prediction of peptide binding affinities to MHC proteins is a very important first step in the process of vaccine design and development. Peptides which act as T-cell epitopes bind to MHC molecules; thus all T-cell epitopes are MHC binders but not all MHC binders are T-cell epitopes. Binding to a MHC protein is a necessary but not a sufficient condition for a peptide to be an epitope. Peptides presented by an MHC on the cell surface have either an intracellular or an extracellular
origin. MHC class-I molecules, present on most cell types, present peptides primarily from protein synthesized within the cell (endogenous processing pathway). MHC class-II molecules, expressed on a restricted number of cell types, such as dendrite cells, B cells and macrophages, can present peptides derived from endocyted extracellular proteins (exogenous processing pathway (Rotzschke and Falk, 1994)). A principal feature of MHC molecules is their allelic polymorphism: 3,411 human leukocyte antigens (HLA) class-I and 1,222 HLA class II molecules were listed by the ImMunoGeneT-ics/HLA (IMGT) database in July 2010 (Robinson et al., 2003). Most of the known predictors are ligand-based, starting with motif-searching algorithms, progressing though different quantitative matrices, to more sophisticated machine-learning methods, such as ANN, HMM and SVM. Among these servers used in the present study, NetMHCpan and Consensus (ANN, SMM) were used for MHC class-I binding prediction and MHC class-II predicted by using CombLib (SMM) and Sturniolo method (Surniolo et al., 1999). The aim of present study was to identify and map the B-cell and T-cell epitopes in the complete genome encoding putative protein sequence of human papillomavirus type 16 E7. These predicted epitopes were also modelled using HHPred modelling sever and after this, selected modelled epitopes were docked with modelled TMEM 50 A protein using PatchDock server which gives the generation of candidate vaccines structures visualised with Discovery Studio 3.5 client tool. These obtained lead molecules were simulated for energy-minimization using molecular dynamics program GROMACS 4.0.6 package in Linux environment.