Infection, Genetics and Evolution 12 (2012) 1513–1518

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

In silico accelerated identification of structurally conserved CD8+ and CD4+ T-cell epitopes in high-risk HPV types

Shishir K. Gupta [a,1], Mugdha Srivastava [b,1], Bashir A. Akhoon [c], Shailendra K. Gupta [c], Niels Grabe [a,*]

[a] Hamamatsu Tissue Imaging and Analysis (TIGA) Center, National Center for Tumor Diseases (NCT), University of Heidelberg, Germany
[b] Society for Biological Research & Rural Development, Lucknow, UP, India
[c] Department of Bioinformatics, Indian Institute of Toxicology Research (CSIR), P.O. Box 80, MG Marg, Lucknow 226001, India

ARTICLE INFO

Article history:
Received 3 December 2011
Received in revised form 6 February 2012
Accepted 12 February 2012
Available online 24 May 2012

Keywords:
Cervical cancer
HPV
Epitope prediction
T-cell epitopes

ABSTRACT

Primary approach to prevent cervical cancer includes the development of human papillomavirus (HPV) vaccines. Currently available HPV vaccines (Gardasil and Cervarix) predominantly consider HPV16 and HPV18 strains. However, due to ignorance of the other high-risk strains aside from HPV16 and HPV18 during vaccine development, the critical need is to synthesize a vaccine with possible protection against all the high-risk HPV types. One feasible approach is to design a vaccine containing conserved immunogenic peptides that represent the genotypic diversity of all the current and future high-risk HPV types. While the epitopes derived from sequentially conserved regions may undergo mutations, it is worthwhile to explore the structurally conserved regions as a new dimension for epitope prediction. In the present study, 81 structurally conserved peptides were predicted to have immune relevance as T-cell epitopes of all the reported high-risk HPV proteins studied. A small dataset of three epitopes was also recognized as potential vaccine candidates generating both CD8+ and CD4+ responses.

1. Introduction

Cervical cancer is the most important manifestation of genital human papillomavirus (HPV) infection and is one of the leading causes of cancer related mortality in women worldwide. Despite the advent of the Papanicolaou smear test almost 50 years ago, cervical cancer remains the second most common malignant disease (Agorastos et al., 2009). More than 100 different types of HPV strains have been identified (de Villiers et al., 2004) on the basis of epidemiologic and phylogenetic relationship. Among them, a total of 15 HPV types (HPV16, HPV18, HPV31, HPV33, HPV35, HPV39, HPV45, HPV51, HPV52, HPV56, HPV58, HPV59, HPV68, HPV73 and HPV82) are classified as high-risk genotypes (Muñoz et al., 2003; Meijer et al., 2009). Persistent infection with any one of these 15 genotypes of carcinogenic HPV causes almost 94.2% cases of cervical cancer (Smith et al., 2007). In addition, the HPV genome consists of eight HPV genes designated as E or L according to their expression in early or late differentiation stage of the epithelium. E1, E2, E5, E6, and E7 are expressed in the early stages; L1 and L2 in the final stages of differentiation; while E4 expressed throughout the differentiation stages (Schiffman et al., 2007).

The availability of HPV genomic sequences and functional characterization of the genes involved in the virulence have considerably increased our understanding towards the molecular basis of HPV pathogenesis and provide a wealth of information that can be used as new tactics for vaccine designing. Vaccination is recognized as one of the most successful and cost-effective public health strategy to control the burden of many infectious diseases worldwide (Chabot et al., 2004). Two licensed HPV vaccines Gardasil and Cervarix are currently in use but these are not completely effective against all the HPV strains (Gupta et al., 2010a). Gardasil is a tetravalent vaccine composed of virus-like particles (VLPs) of HPV16, HPV18, HPV6 and HPV11 while Cervarix is a bivalent vaccine composed by VLPs of HPV16 and HPV18 (La Torre et al., 2007). Two high risk types are majorly covered by these vaccines therefore essentially there is need of a potential vaccine that may cover all the high risk genotypes of HPV.

The identification of peptides that stimulate T-cell responses, termed T-cell epitopes, is a critical requirement for the development of successful epitopic vaccines. Immunoinformatics has

Abbreviations: ANN, artificial neural network; ARB, average relative binding; BLAST, basic local alignment search tool; HLA, human leukocyte antigen; HPV, human papillomavirus; IEDB, immune epitope database; MHC, major histocompatibility complex; NMR, nuclear magnetic resonance; PDB, protein data bank; PSI-BLAST, position-specific iterative; RMSD, root mean square deviation; SCR, structurally conserved region; SMM, stabilized matrix method; SVM, support vector machine.

* Corresponding author. Address: Hamamatsu Tissue Imaging and Analysis (TIGA) Center, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany. Tel.: +49 6221 54 512 48; fax: +49 6221 54 514 82.

E-mail addresses: shishir.gupta@bioquant.uni-heidelberg.de (S.K. Gupta), mugdha.srivastava@gmail.com (M. Srivastava), bashirakhlaq@gmail.com (B.A. Akhoon), skgupta@iitr.res.in (S.K. Gupta), niels.grabe@bioquant.uni-heidelberg.de (N. Grabe).

[1] These authors contributed equally to this work.

1567-1348/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2012.02.022

attracted considerable attention of experimental biologists as it provides the opportunity for rapid screening and identification of probable vaccine candidates before being subjected to expensive and time consuming in vitro confirmatory studies (Gupta et al., 2010b, 2011). Indeed, immunoinformatics has changed the paradigm of traditional vaccinology and established as a crucial field for accelerating immunology research (Baloria et al., 2011).

Gardasil and Cervarix vaccines are prophylactic, since they immunize the individual against the contraction of the HPV, but unfortunately, they do not eradicate a pre-existing infection (Wain, 2010). However, therapeutic vaccines generates T-cell immune responses to eliminate existing HPV infection and HPV-associated neoplasms (Lin et al., 2010). Epitope based vaccines provide a specific strategy for prophylactic and therapeutic application of pathogen-specific immunity. The identification of epitopes suitable for diagnostic use and for therapeutic or prophylactic intervention is clearly a crucial prerequisite of these strategies. Selection of immunogenic, consensus and conserved epitopes from proteins of major high-risk strains may provide an experimental basis for designing of HPV specific T-cell vaccines that probably target to approx 95% of the causative agents.

Moreover, due to the possibilities of mutations in sequentially conserved consensus T-cell epitopes, derived from multiple sequence alignment of proteins, the relevant vaccines are likely to show subsequent restriction to the T-cell responses. It is well known that the protein structures are more conserved in comparison to protein sequence. Thus, the information of putative structurally conserved epitopes with no or low variant representation is potentially advantageous in avoiding altered peptide efficacy. In this study, we used the *in silico* techniques to identify the structurally conserved T-cell epitopes in HPV proteins among all the high-risk HPV types.

2. Material and methods

2.1. Sequence collection

Amino acid sequences of E1, E2, E6, E7, L1 and L2 proteins of all 15 high-risk HPV types were collected from HPV sequence database (http://hpv-web.lanl.gov/). All the retrieved protein sequences were grouped into six datasets of 15 proteins, one from each high-risk HPV types based on their expression in early or late differentiation stage of epithelium for further analysis. E5 proteins were not included in the study since E5 protein has previously been reported as inducer of down-regulation of MHC class-I (Ashrafi et al., 2006; Kanodia et al., 2007; Miura et al., 2010). Moreover, the E2-E5 region of HPV genome has been shown to be lost when the episomal HPV DNA integrates into host chromosome. Hence, using E2 and E4 epitopes beyond this stage may be futile, consequently E4 proteins were also not considered. However, since E1 and E2 are expressed in higher levels than E6 and E7 early in the progress of an HPV infection, it may be assumed that these proteins may act as good targets for vaccine development against early stages of disease. Therefore, despite the possible loss of E2 proteins during host chromosome integration, E2 was included in this study.

2.2. Structure modeling

PSI-BLAST (Altschul et al., 1997) was used to search the suitable template for modeling of all the proteins in six datasets (E1, E2, E6, E7, L1 and L2). PSI-BLAST scans the profile of the query sequence against each of the template sequences in a database. Based on homology percentage (>30%), less e-value and comparatively high-quality X-ray crystallographic structures (R-values $\leqslant$ 2Å), PDB templates were selected for the 3D modeling of each HPV protein (data not shown). In the absence of X-ray crystallographic template structures, NMR structures were also used. The 3D structure of all the proteins were built by using automated Modeller (Eswar et al., 2006) tool available on $(PS)^2$ server (http://ps2.life.nctu.edu.tw/) (Chen et al., 2006). Further, the stereochemical quality of all the generated 90 models were evaluated using PROCHECK server (Laskowski et al., 1993).

2.3. Energy minimization

Minimum energy arrangements of atoms correspond to stable states of the system and energy minimization procedure can repair distorted geometries of proteins by moving atoms to release internal constraints (Srivastava et al., 2011). Energy minimizations of all the generated models were carried out to eliminate bad contacts between protein atoms and structural water molecules and to correct the stereochemistry of the model followed by model evaluation and validation through Ramachandran plot using PROCHECK v.3.0.1 (Laskowski et al., 1993). Steepest descent approach and conjugate gradient technique with Gromos96 forcefield (Gunsteren et al., 1996), implemented in Swiss-PdbViewer v.4.2 (Kaplan and Littlejohn, 2001) was exploited for energy minimizations of models. Furthermore, energy minimizations and model evaluations were repeatedly performed until getting the best model of each protein with minimal bad contacts and utmost residues in core and allowed regions of Ramachandran plot.

Stereochemical qualities of final energy minimized protein models were statistically compared with that of initial structures in order to identify the significant structural changes during minimization process. From the structural minimization process of all the models, the mean structural refinement of each datasets were expressed as mean ± standard error of the difference (SED) after paired comparisons between initial and minimized dataset. Statistical differences between means were determined by student's t-test and $P < 0.05$ was considered significant.

2.4. Structure clustering

The structural diversity of the HPV proteome was studied by structural superimposition of HPV proteins from all the six protein datasets. Superimposition and calculation of root mean square deviation (RMSD) among structures within each dataset was performed using the Kabsch rotation matrix (Kabsch, 1976, 1978) implemented in MaxCluster v.3.4.2 (http://www.sbg.bio.ic.ac.uk/~maxcluster/). The RMSD was calculated using the superimposition between matched pairs.

$$\mathrm{RMSD} = \sqrt{\sum_{i}^{N} \frac{[di * di]}{N}}$$

where di is the distance between matched pair i, and N is the number of matched pairs.

In order to reduce the biasedness and the size of dataset, cluster analysis was performed using sequence-independent alignment method (Ortiz et al., 2002) and Nearest Neighbour (NN) clustering (Shortle et al., 1998) implemented in MaxCluster within default threshold of RMSD and MaxSub score (Siew et al., 2000). The MaxSub score falls in a range between 0 to 1, where 1 is an identical pair of structures

$$\mathrm{MaxSub} = \frac{1}{N}\sum_{i}^{M}\left[\frac{1}{\left\{1+\left(\frac{di^2}{d^2}\right)\right\}}\right]$$

where di is the distance between identical residues i, d is the distance threshold, M is the number of residues in the MaxSub and