Available online at www.sciencedirect.com
ScienceDirect
Molecular Immunology 44 (2007) 1253–1261
Molecular Immunology
www.elsevier.com/locate/molimm

Virus-epitope vaccine design: Informatic matching the HLA-I polymorphism to the virus genome

Tal Vider-Shalit, Shai Raffaeli, Yoram Louzoun *

Math Department, Bar Ilan University, 52900 Ramat Gan, Israel

Received 5 February 2006; received in revised form 7 June 2006; accepted 8 June 2006
Available online 22 August 2006

Abstract

Attempts to develop peptide vaccines, based on a limited number of peptides face two problems: HLA polymorphism and the high mutation rate of viral epitopes. We have developed a new genomic method that ensures maximal coverage and thus maximal applicability of the peptide vaccine. The same method also promises a large number of epitopes per HLA to prevent escape via mutations. Our design can be applied swiftly in order to face rapidly emerging viral diseases.

We use a genomic scan of all candidate peptides and join them optimally. For a given virus, we use algorithms computing: peptide cleavage probability, transfer through TAP and MHC binding for a large number of HLA alleles. The resulting peptide libraries are pruned for peptides that are not conserved or are too similar to self peptides. We then use a genetic algorithm to produce an optimal protein composed of peptides from this list properly ordered for cleavage. The selected peptides represent an optimal combination to cover all HLA alleles and all viral proteins.

We have applied this method to HCV and found that some HCV proteins (mainly envelope proteins) represent much less peptide than expected. A more detailed analysis of the peptide variability shows a balance between the attempts of the immune system to detect less mutating peptides, and the attempts of viruses to mutate peptides and avoid detection by the immune system. In order to show the applicability of our method, we have further used it on HIV-I, Influenza H3N2 and the Avian Flu Viruses.

Keywords: Polymorphism; Immunomics; DNA vaccine; Informatics; Universal; Virus; HLA

1. Introduction

Rapidly emerging viral diseases, such as the Avian Flu Virus, oblige the development of procedures for the fast development of vaccines. Traditional vaccine development takes many years, while emerging disease can spread worldwide within one year. Furthermore, classical vaccines use attenuated viruses with many dangers requiring many slow safety tests and a very careful development. A less dangerous and far more rapid method is sub-unit vaccines, composed of peptides from viral proteins. The main limit of sub-unit peptide vaccines is the high HLA polymorphism. The HLA has the highest polymorphism among all human loci (Robinson et al., 2001). Thus, any peptide based vaccine containing a limited number of peptides will not be applicable to a large part of the human population, since most T cell epitopes are specific for a single HLA allele or at most to single family. Moreover, the use of a limited number of peptides allow viruses to evade T cells trained on this limited set of peptides through point mutations of their epitopes (Timm et al., 2004). In order to avoid these two limitations, a large number of epitopes for each HLA are required. Note that the probability that a virus would simultaneously mutate multiple epitopes decreases exponentially with the number of epitopes.

Practically all previous attempts (Dietrich et al., 2005; Kast et al., 1991; Wang et al., 2005) to develop vaccines were focused on a single or a limited number of epitopes. For example, using super-type epitopes or promiscuous epitopes (De Groot and Rappuoli, 2004) or focusing on a limited number of alleles (Gulukota, 2003; Gulukota and DeLisi, 1996; Wang et al., 2005). Here, we present a new methodology to design sub-unit vaccines that would cover a large part of the human population and provide a large enough coverage within a given HLA to avoid viral mutation through epitopes mutation. Our method is fully based on genomics and bioinformatics methods, and is thus extremely fast to apply, once the appropriate genomic data is available. We use a set of bioinformatics tools and extensive genomic data to

* Corresponding author. Tel.: +972 3 5317610; fax: +972 3 5317165.
E-mail address: louzouy@math.biu.ac.il (Y. Louzoun).

0161-5890/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.molimm.2006.06.003

detect most candidate epitopes from a viral genome, and select from those an optimal subset of epitopes that yield a maximal coverage of the human population using only conserved epitopes.

Another major advance of our method is the combination of bioinformatics tools for all epitope processing stages, and not only for the MHC-binding stage, as was done in most of the previous attempts (Wang et al., 2005). The combination of algorithms for all processing stages maximize the probability that a computed peptide would indeed is presented. We focus in this method on MHC-I and CD8 cytotoxic T cells (CTL) epitopes. The repertoire of those is more limited and more amenable to informatic estimates. Moreover, those are the main effectors cells of the cellular immune system.

CTL recognize 8–10 amino acid peptides in the context of MHC-I molecules. Approximately 90% of which are nonamers (Robinson et al., 2001). A cleaved nonamer is presented on a MHC-I molecule only if its affinity for the MHC molecule is high enough. This affinity is determined in large part by the anchor and helping residues. The peptides that eventually bind the MHC-I molecules must pass through a succession of stages. They first must be cleaved from a larger protein (Rock et al., 2002), then pass through the TAP machinery (Uebel and Tampe, 1999) into the ER, where they can bind the MHC molecules (Williams et al., 2002). Each one of these steps limits the peptide repertoire. During the last two decades, many models for each of these stages (cleavage, TAP binding and MHC binding) have been developed by us and by other groups (Andersen et al., 2000; Flower, 2003; Kesmir et al., 2002; Nussbaum, 2001). We have combined prediction algorithms, validation methods and the accumulation of genomic data to develop a genomic view of the immune self and foreign repertoire (Louzoun et al., 2006).

An instructive candidate for testing our genomic sub-unit peptide vaccine methodology is the Hepatitis C virus (HCV). HCV is a positive, single-stranded RNA virus of the Flaviviridae family. The size of the HCV genome is 9.6 kb and it carries a single, long open-reading frame (ORF) encoding a polyprotein of about 3010 amino acids. We focus on HCV since the immune response to HCV is CTL based. The spread and frequency of HCV makes it one of the most dangerous viruses. Another appeal of this virus is its high within host mutation rate. During replication, the HCV constantly changes and mutates. It manages to evade the body's immune system (Bowen and Walker, 2005; Erickson et al., 2001; Guglietta et al., 2005; Timm et al., 2004; Weiner et al., 1995) and remains in the body of most infected people. These mutations significantly delay the immune response, since once an immune response was mounted against a virus, a few weeks may be required in order to switch to a new antigen (Kohler et al., 1994; Uebel et al., 1997). However, in contrast with HIV, HCV still has many highly conserved regions allowing us to detect many conserved epitopes. There are six main genotypes groups (1–6), each containing multiple subtypes (a, b, ...). Different subtypes are distributed differently in different parts of the world, and certain genotypes predominate in certain areas.

Within a viral genome, we compute all epitopes predicted to be presented on most MHC-I alleles for the most frequent haplotypes. The resulting list is filtered for all peptides that are not conserved in the virus, or are too similar to self-epitopes (i.e. epitopes presented from the host genome). Although these epitopes could be very immunogenic, their mutability limits their validity as good candidates for a vaccine. From this list, we sort an ordered optimal sub-list of a properly cleaved peptide combination presenting all proteins and all HLA alleles. We have applied our method to HCV, HIV, Influenza and the newly emerging Avian Flu Virus, to show the concept plausibility. When a large amount of viral epitopes are known, we have compared our method with the known epitopes and found a significant overlap.

2. Materials and methods

2.1. Genomic data

Viral and human protein sequences were used for this analysis. The human sequences were obtained from the Ensembl database (Birney et al., 2004; Hubbard et al., 2002). We have only used genes experimentally validated to produce proteins and thus used only known and CCDS human proteins (and not the ones based on ORFs). The hepatitis C protein sequences were obtained from the LANL database (Korber, 2002; Kuiken et al., 2003, 2005) (http://www.hiv.lanl.gov). We used the coding regions from the virus genome which encodes for three structural proteins: Core (C), envelope 1 and 2 (E1, E2) and seven non-structural proteins (p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B).

2.2. Cleavage

Given a peptide and its N and C flanking amino acids, FN–$P_1P_2P_3 \ldots P_n$–FC, a cleavage score is defined as:

$$S(\text{peptide}) = S_1(\text{FN}) + S_2(P_1) + \sum_{i=2}^{n-1} S_3(P_i) + S_4(P_n) + S_5(\text{FC}).$$

FN and FC are the N and C termini flanking regions while P_i are the internal amino acids. This linear score determines the probability that a specific peptide should result from proteasomal cleavage. A peptide with a high score, S, has a high probability of being produced, while a low S score predicts a low creation probability. We have used a simulated annealing algorithm for learning the appropriate values of S_1–S_5. The proteome can modify its structure and its cleavage properties in the presence of interferon gamma. We have incorporated results from both standard and "immuno-proteasome" in our analysis. The algorithm was validated to have much less than 16% false positives and 10% false negatives (Louzoun et al., 2006).

2.3. TAP

The probability that a peptide should bind to the TAP machinery is mainly a function of the residues at the first N and three last C positions. It can be approximately estimated through a linear combination of the binding energies of each residue. Multiple