Proteome-wide Epitope Prediction: Leveraging Bioinformatic Technologies in Rational Vaccine Design

Lindsay M.W. Piel, Stephen N. White*

USDA-ARS Animal Disease Research 3003 ADBF, WSU Pullman, WA 99164, USA

*Correspondence should be addressed to Stephen White; stephen_white@wsu.edu

Received date: August 22, 2021, Accepted date: November 23, 2021

Copyright: © 2021 Piel LMW, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Artificial intelligence-based prediction technologies have allowed definition of T-cell epitopes presented by Major Histocompatibility Complex (MHC) molecules with allele-specificity of presentation. While some have utilized these technologies on a smaller scale, recent work has expanded the workable proteome size, leveraged both classes of Major Histocompatibility (MHC) molecules, extended the range of host species assessed during comparative analysis, and incorporated pathogen genetic diversity to highlight broadly useful epitopes. A recent study focused on the zoonotic pathogen Coxiella burnetii exemplifying themes and possibilities for future analyses. These data suggest an expanding role for epitope prediction in rational vaccine design for a very broad range of pathogen and host systems.

Keywords: T-cell epitope, Machine-learning, Artificial intelligence, Major histocompatibility complex, Proteome-wide

Commentary

Vaccine development began in the 1790’s when Edward Jenner used cowpox to confer protection against the smallpox virus [1]. The field of vaccinology has greatly expanded since then, wherein vaccination has been a valuable tool in the decline of many diseases [1,2]. While Jenner’s use of cowpox shares attributes to a live-attenuated vaccine, there are alternate methods of vaccination, which include subunit, conjugate, mRNA, viral vector, and toxoid vaccines [2-4]. Development of these methods was facilitated through greater understanding of the immune response, elucidation of both host and pathogen genetic diversity, and advancement of laboratory techniques [1-3]. The most recent notable advancement in vaccine production was the development of a nucleic acid vaccine to combat the SARS-CoV-2 virus [1]. While advancement in vaccine methodology can be readily seen, many subunit-based vaccines end up generating a predominantly B-cell driven response [1,5].

B-cells are responsible for differentiating into plasma cells and mediating antibody production [1,6,7]. Antibodies are important during the immune response as they mediate opsonization for complement and innate immune cells; however, they can also inactivate circulating viruses [8]. Identification of B-cell immunogens typically relies on antibody responses seen in patients that have survived previous infection with the agent of interest [9-11].

In addition, it is well known that T-cell immunity plays an important role in host defense against infection, and it is becoming increasingly evident that vaccine production needs to incorporate T-cell recognition of pathogens [1,12]. This idea is inherently important to infectious agents that require a T-cell helper 1 (Th1) phenotype, as cellular immunity is the cornerstone to agent clearance [13-15]. It is possible that the ability to use whole cell inactivated or live-attenuated strains has previously limited the requirement to assess T-cell epitopes, or peptides that allow T-cell recognition of a pathogen [1,3]. Still, the need to generate vaccines against agents that are either difficult to culture or those that require a high-level containment facility suggests the necessity of accurately defining T-cell epitopes [1,3,16].

Recognition of agent presence by T-cells relies on the major histocompatibility complexes (MHC) present on the surface of other cell types [17-19]. There are two classes of MHC alleles, namely MHC Class I (MHC I) and MHC Class II (MHC II) [17].
MHCI exists on the surface of most cell types and as such is important in the designation of compromised host cells to cytotoxic T-cells, or CD8+ T-cells [17-20]. Therefore, MHCI is inherently responsible for alerting the immune system to an intracellular pathogen [7,18]. In contrast, MHCI I marks antigen presenting cells (APCs) consisting of dendritic cells, macrophages, and B-cells [7]. Recognition of a loaded MHCI by CD4+ T-cells, or T-helper cells, initiates the production of an organized adaptive immune response, so it is required for response to most pathogens [7,17].

Bioinformatic programs delineating T-cell epitopes started to be developed in 2007 [3,20,21]. Labs studying virology-based interaction with the immune system were able to use these programs to identify T-cell epitopes within the entire viral proteome [22]. Due to the size difference between viral and bacterial proteomes, bacteriology-based research continued to narrow the proteins of interest based on other constraints [23-26]. While this methodology can identify proteins that interact with T-cells over the course of infection, there are likely highly qualified immunogens that will be missed by limiting queried proteins. Recently published work has achieved one of the two known proteome-wide T-cell epitope analyses within a bacterium [22,27] while advancing numerous other aspects of larger-scale analysis, such as avoiding induction of autoimmune responses, improved capture of pathogen genetic diversity, leveraging diversity within hosts, and comparing results across different hosts of the same pathogen.

Pathogen protein conservation has been of specific concern during vaccine development, especially when considering profoundly variable genomes or rapidly mutating agents [4,18]. Previous work with bacterial agents has completed proteome-wide alignments to identify the core- and pan-genome. However, while these studies have used a large pool of bacterial isolates, they have not fully considered bacterial groupings within certain species [24,26,28]. Choice isolates will include factors like alternate virulence during inoculation studies, isolation from differing host species, or large genomic rearrangements [27]. Leveraging phylogenetically diverse isolates is recommended [27,29], and this step can be enhanced by choosing isolates which arose from diverse hosts and capture a range of the most important virulence phenotypes [29-32].

While prior vaccine design studies have commonly employed agent conservation to narrow the proteins of interest, there are many investigations that do not consider host homology [22,24,25]. Homology of agent proteins to the proteins found in host species is important to recognize as sensitization of the host immune system to these macromolecules could cause an autoimmune reaction [28]. Previous work examining either allergy responses or autoimmune reactions can suggest acceptable cut-off values for homology between agent and host [28,33]. Genome-wide analysis for each host of interest generates a dataset that is better viewed in matrices rather than individual records, and numerous programs can accomplish such matrix analyses, including BLASTGraber, BlastViewer, BlasterJS, and JAMBLAST [34-36]. At this stage many previous studies have limited the proteins of interest based on antibody responses, surface localization, or secretion of proteins [3,23-26,28,37,38]. However, this technique can fail to identify strong immunogens during initial screening. Therefore, to enhance the outcome of predicted T-cell epitopes, no further protein limitation should be performed.

Bioinformatic tools have been developed to model each processing step for T-cell recognition of an antigen. These levels of processing include proteasomal cleavage, TAP interaction, MHCI/MHCII binding, and T-cell recognition [17,20,39,40]. Of these events, binding of antigens to MHC alleles is the most selective stage for antigen recognition [17]. Tools which define MHC loading of antigen consist of both binding affinity matrices and machine-learning. In silico identification of T-cell epitopes began by using matrices that examined the ability of the MHC binding groove to interact with the R-side groups of amino acids present in agent peptides. This methodology was expanded into machine-learning through generation of support vector machine (SVM) and artificial neural network (ANN) based bioinformatic tools [17,18]. With these programs, data pools that contained either experimentally defined antigens or random peptides were exploited to train tools on alleles of interest [18,20]. Once trained, these programs were able to expand into delineating T-cell epitopes for MHC alleles that had not been directly studied previously [20,21]. Available tools that encompass machine-learning include ANN_PRED, MHC2Pred, ConvMHC, NN_Align 2.3, NetMHC, SVMHC, KISS, SVRMHC, DeepHLA_Pan, and IEDB binding [18,41].

As with the pathogen proteins, maintaining diversity is also imperative when it comes to the host side of vaccine production. Human sequencing of MHC alleles has thus far defined approximately 13,000 different sequences worldwide. This information is organized on the Allele Frequency Net Database (AFND) by geographic region, genomic locus, and data collection standards [42]. This allows investigators to better refine alleles of interest by generation of phylogenetic trees to accomplish preservation of host allele heterogeneity, while encompassing a worldwide distribution. Prior to this work, many researchers would focus on the known supertype alleles, which are suggested to be present in 88% of the population and bind similar antigens compared to one another [18,25,38]. Extending the allelic pool tested from the known supertype alleles not only allows for a decrease in the number of false positives returned from analysis but also permits a larger representation of populations.

Along with the profound list of human alleles available on most T-cell epitope defining databases, there are certain databases which strive to encompass alternate vertebrate species. The
most common of these being murine alleles of inbred strains of mice [18]. Recent work within the lab has pushed these limits further by incorporating bovine MHCI alleles into the analysis. The importance of expanding host alleles of interest rests in both the zoonotic nature of certain pathogens and in the development of veterinary vaccines. Inclusion of alternate species MHC alleles into machine-learning programs requires two steps. The first of these is expanding the vertebrate allelic sequences available to represent, not only alternate organisms, but breeds and regions specific to these species [43-45]. The second step to increase species representation is to generate elution data based on defined MHC alleles of interest [20,21]. The difficulty in these efforts regards the structural differences between the MHCI and MHCII molecules. MHCI molecules generally bind 9-mer long peptide sequences and have a binding pocket that is closed, making delineation of allele specificity easier to define. In comparison, MHCII molecules have an open binding pocket, increasing the complexity of elucidating the core binding region of studied peptides [17,18]. Reynisson et al. have attempted to solve this problem by using motif deconvolution methods, wherein evaluation of the new program determined a decrease in false positive data [21,46]. At the moment, NetMHCpan 4.1 maintains the largest selection of host species alleles, encompassing human, non-human primates, mouse, swine, bovine, canine, and equine species [18,20]. Furthermore, there is the possibility to use self-defined alleles of interest within certain bioinformatic tools [20]. This may help surpass the initial issue of training data availability for alternate vertebrate species, but one must keep in mind that increasing the evolutionary distance will inevitably affect the predictive value of the program [21].

Beyond the call for increased training on alternate host MHC alleles, there is the paradigm shift to proteome-wide assessment of multiple MHC varieties. Of the two existing proteome-wide T-cell epitope studies, one focused on MHCI based T-cell epitopes and the other determined T-cell epitopes for MHCI and MHCII [22,27]. Assessing MHCII loading of antigens is of major importance as this mediates adaptive immunity organization and response [17]. As mentioned previously, this cellular immunity priming is required for elimination of certain pathogens [1,13,47]. The results obtained from assimilating each of these methods will require alternate evaluation strategies as compared to previous bioinformatic techniques. This is due to previous analysis producing a manageable number of records in relation to the big data produced previously [37,39,48]. Analytical approaches may be comprised of isolating T-cell epitopes which interact with a high number of tested alleles, proteins that have a certain number of T-cell epitopes present, and T-cell epitopes returned during inquiry of both MHC classes [27,39,40]. Notably, this examination should be derived based on the pathogen of interest and the desired vaccine methodology. Following identification of T-cell epitopes, it should be ascertained whether the identified peptides elicit a cellular response when host or model organisms are exposed to the peptides of interest. A method frequently employed during this analysis is the ELISpot assay, which can assess the production of cytokines by isolated T-cells [23,38,49]. This will promote validation or disqualification of the T-cell epitopes previously defined bioinformatically.

A flow-chart encompassing an overview of the presented methodology is depicted in Figure 1. Expansion of this methodology may allow for analysis of pathogens with substantially larger genome sizes, such as apicomplexans [50]. Work on apicomplexan vaccinations has become progressively more important due to the emergence of drug resistance [51]. Generation of an apicomplexan database examining peptide:allele interactions would be of considerable size; however, there are multiple questions that can be considered and answered through use of such a database.

![Figure 1: Overview of methodologic steps for proteome-wide T-cell epitope prediction. Above titles represent the overall step while the contained text annotates certain context that should be considered.](image)

J Cell Immunol. 2021
Volume 3, Issue 6

377
In cases where pathogens can infect multiple hosts, the ability to analyze many hosts simultaneously can harmonize vaccine design efforts to achieve efficiencies in testing and overall cost savings [27]. There are many opportunities for application of proteome-wide epitope prediction analyses in rational vaccine design of pathogens with large proteomes. The benefits can include de novo vaccine situations, as well as T-cell response optimization of older designs. Thus, proteome-wide epitope prediction will be a useful tool in rational vaccine design for a wide variety of pathogens.

References


26. Hisham Y, Ashhab Y. Identification of cross-protective potential antigens against pathogenic Brucella spp. through combining pan-