Immunogenicity Prediction by VaxiJen: A Ten Year Overview

Nevena Zaharieva1, Ivan Dimitrov1, Darren R Flower2 and Irini Doytchinova1*

1Faculty of Pharmacy, Medical University of Sofia, Bulgaria
2School of Life & Health Sciences, Aston University, Aston Triangle, Birmingham, UK

Abstract

Vaccination stands alone as an unparalleled medical intervention able to save human lives numbering hundreds of millions. Modern licensed vaccines are either whole organism based or based on single proteins, together with carbohydrate epitope-based vaccines. Single protein or so-called subunit vaccines are prime targets for vaccine design and reverse vaccinology. Immunoinformatics is a branch of bioinformatics focusing on immunology and vaccinology, which includes database compilation, data mining, and epitope prediction. Immunoinformatics can aid in the discovery of subunit vaccines through algorithms for immunogenicity prediction. VaxiJen is the first and, ostensibly, the only widely-used method for distinguishing between immunogens and non-immunogens among proteins of bacterial, viral, parasite, fungal and tumour origin. In this review, we chart the applications of VaxiJen, placing these into context, and explore some of the future directions that research into identifying immunogens might take.

Keywords: VaxiJen; Subunit vaccine; Immunoinformatics; Immunogenicity prediction

Introduction

Vaccination is unquestionably one of the greatest achievements of human civilisation. The WHO estimates that today immunizations prevent between 2 and 3 million deaths annually and protect many more people from illness and disability [1]. Current vaccines are one of several types: live attenuated microorganisms (Hepatitis A virus, Influenza virus, Japanese encephalitis virus, Measles virus, Mumps virus, Poliovirus, Rabies virus, Vaccinia virus, Varicella zoster virus, Yellow fever virus, M. tuberculosis, S. typhi), microorganisms inactivated or killed by heat or chemical treatment (Influenza virus, Japanese encephalitis virus, Mumps virus, Poliovirus, V. cholerae, B. pertussis, Y. pestis), subunit vaccines (proteins from Influenza virus and Hepatitis B virus, toxoids of B. pertussis, C. tetani and C. diphtheriae, conjugates of H. influenzae type b, N. meningitis and S. pneumoniae), recombinant vaccines (Hepatitis B virus surface antigen and Human papillomavirus), or carbohydrate epitope-based vaccines (Pneumovax-23) [2].

The wide use of vaccines during the last 150 years has led to significant reduction (95-97%) in mortality from a wide range of previously deadly diseases, including diphtheria, tetanus, measles, mumps, rubella, pneumonia, hepatitis B and meningitis [3], as well as the total eradication of smallpox and the near eradication of polio. Yet, despite such tremendous success, a long list of deadly diseases still await efficient vaccines.

Vaccine development remains reliant on a gallimaufry of antiquated processes. The knowledge-based design of vaccines is still very much in an initial stage. Immunogenicity is the property of a molecule (protein, lipid or carbohydrate, or a combination thereof) or living organism (virus, bacterium, parasite or fungus) that induce humoral and/or cell-mediated response from the immune system. An immunogen triggering a protective immune response is a protective immunogen. Protective immunogens can make promising vaccine candidates.

Identification of protective immunogens among the proteome of a particular microorganism is the initial step in the process of vaccine design and development. This is the key step in what has come-to-be-known as Reverse Vaccinology [3]. Use of in silico methods in this initial stage can direct and thus greatly shorten subsequent experimental work [4-7]. Moreover, the proper use of in silico methods can replace, reduce and refine the use of time-consuming and often misleading animal experimentation [8].

Immunoinformatics is a branch of bioinformatics focusing on immunology and vaccinology [5]. It includes strategies for databases compilation, data mining and analyses, in silico methods and algorithms for immunogenicity prediction, including tools for the prediction of B-cell and T-cell epitopes. The current state of immunoinformatics has been reviewed recently [9,10].

VaxiJen (http://www.ddg-pharmfac.net/vaxijen/VaxiJen.html; or, alternatively, just type vaxijen into Google) was the first server for prediction of protective antigens, tumour antigens and subunit vaccines [11]. It is also the first alignment-free bioinformatics tool for in silico identification of immunogens. VaxiJen uses Wold’s z-scales [12] to describe the main physicochemical properties of the amino acids building the tested proteins, then converts the derived strings into uniform vectors by auto cross covariance (ACC) [13], selects the relevant variables by genetic algorithm (GA) [14] or stepwise regression and finally, classifies the proteins as protective antigens or non-antigens by partial least squares (PLS)-based discriminant analysis. Initially, the algorithm was trained to identify bacterial protective immunogens [15]. Later, models for viral and tumour immunogens were included [11] and VaxiJen was developed to give free access to the models. The last version of VaxiJen (VaxiJen 2.0) also includes models for identification of parasite and fungal immunogens [16].

Since its launch in 2007, VaxiJen has been widely used to identify candidate subunit vaccines among proteins of bacterial, viral, parasite,
fungal and tumour origin. In the present review, we analyse the diverse applications of VaxiJen. Special attention is given to the experimentally validated predictions. We also take the opportunity to explore a few of the future directions that research into identifying immunogens might take.

**Prediction of bacterial immunogens**

*Staphylococcus aureus* (*S. aureus*) is one of the most important causes of nosocomial and community-acquired infections. Yet no vaccine against *S. aureus* exists. Fragments of the virulence proteins clumping factor A (CfA), iron-regulated surface determinant (IsdB) and gamma hemolysin (Hlg) were selected and predicted as immunogenic by VaxiJen (scores 0.47 and 0.58 at cutoff 0.4) [17]. A recombinant gene containing the three fragments and hydrophobic linkers between them was constructed and expressed in *E. coli* BL21 [18]. Immunisation of BALB/c mice with the recombinant protein evoked antigen-specific antibodies and increased survival following the intraperitoneal challenge with pathogenic *S. aureus*.

*S. aureus* clumping factor B was also identified by VaxiJen as a protective antigen (score 1.09) of *S. aureus* [19]. Shahbaz et al. [20] proposed a hexavalent subunit vaccine containing fragments of agglutinin-like sequence 9 (Als9) (VaxiJen score 0.93), CfA (score 1.53), methicillin resistance determinant protein (PmB) (score 1.15), immunoglobulin G-binding protein A (Spa) (score 1.55), serine-aspartate repeat-containing protein E (SdrE) (score 1.34) and biofilm-associated surface protein (Bap) (score 1.10). Hajighahramani et al. [21] constructed a multi-epitope peptide vaccine containing B- and T-cell epitopes from alpha-enoase (Eno1), CfA and IsdB. VaxiJen scored this candidate-vaccine as a moderately antigenic (score 0.50) [21].

*Streptococcus pneumoniae* (*S. pneumoniae*, pneumococcus) is the major pathogen causing pneumonia, meningitis and sepsis. Currently, two vaccines for prevention from pneumonia caused by *S. pneumoniae* are licensed for adults: 23-valent polysaccharide vaccine (PPV23) and 13-valent conjugate vaccine (PCV13) [22]. Both vaccines contain capsular polysaccharides eliciting weak immunogenicity in children, and are unable to induce effective immune memory [2]. The protein-containing PCV13 is more effective in children than PPV23; the high manufacturing costs limit widespread use in developing countries [23]. *S. pneumoniae* serotype 19F is one of the main pneumococcal serotypes that cause invasive pneumonia in children under 5. Using bioinformatics analysis (including VaxiJen) on surface proteins from *S. pneumoniae* serotype 19F, Tarahomjoo [24] identified four putative candidate-vaccines: cell wall surface anchor family protein, D-alanyl-D-alanine ligase, sulfate transport permease protein C (CysW), competence lipoprotein (ComL), and O-antigen polymerase (uniprot ID: A7ZTH5) – and then analyzed and O-antigen polymerase (uniprot ID: A7ZTH5) – and then analyzed.

**The non-pathogenic, saprophytic Mycobacterium indicus pranii (MIP) provides protection against *M. tuberculosis* infection in mice [35,36]. It is used as an adjunct to chemotherapy in patients with type I and type II category tuberculosis [37]. VaxiJen identified proteins MIP0340, MIP5962 and MIP7697 as the most prominent putative antigens [38]. A comparative analysis of *M. tuberculosis* and MIP using VaxiJen highlighted the importance of the PE/PPE family in host immunomodulation, supporting further the likely potential of MIP as an effective vaccine against tuberculosis [39].

Strains of enterotoxigenic *Escherichia coli* (ETEC) are the most common cause of bacterial diarrhea in travelers to and children in developing countries [40]. ETECs adhere to the host small intestine by colonization factors (CFs) and produce enterotoxins [41]. ETEC secrete heat-stable enterotoxins (STs) and/or heat-labile enterotoxins (LTs). The design of subunit vaccines focus on CFs and toxins. A synthetic chimeric gene, encoding the colonization factors CFA/I, CS2, and CS3 and the B-subunit of LT was designed [42]. VaxiJen showed high antigenicity of the chimeric protein. However, there is no experimental validation of the immunogenicity of the chimeric protein. Similarly, a candidate subunit vaccine, composed of CFB, CSSA, CSBB, and LT, was designed, it’s immunogenicity predicted by VaxiJen, synthesized, and expressed successfully in a prokaryotic host but without subsequent experimental validation [40]. A recombinant protein containing CFB, CFAE and LTB has been tested experimentally [41]. Mice have been immunized with this protein and the antibody titer and specificity of the sera analyzed by ELISA.

Mehla and Ramana applied another approach to identify novel immunogenic proteins from ETEC [43]. They identified proteins shared between virulent ETEC strains E24377A and H10407, but not by commensal *E. coli*. From this initial pool, human homologues were eliminated, the rest were predicted by VaxiJen, with only immunogenic proteins scrutinized for cellular location. Three novel probable immunogenic proteins were identified – putative membrane protein (uniprot ID: ATZG55), uncharacterized protein (uniprot ID: ATZG4) and O-antigen polymerase (uniprot ID: ATZTH5) – and then analyzed for T- and B-cell epitopes.

The enterohemorrhagic *E. coli* (EHEC) strains are major human food-borne pathogens, responsible for bloody diarrhea and hemolytic-uremic syndrome worldwide [44]. The genomes of EHEC strains EDL933 and Sakai have been screened to identify common EHEC antigens absent in nonpathogenic *E. coli* strains. The analysis revealed 897 protein sequences. Applying immunoinformatics (including VaxiJen) their number was reduced to 65. Nine were tested for immunogenicity using a murine gastrointestinal infection model. Two vaccine candidates - Lom-like protein and a putative pilin subunit - significantly induced Th2 cytokines and production of slgA, while the third (a fragment of the type III secretion structural protein Esct) reduced EHEC ecum colonization [44]. Chimeric proteins have been designed containing virulence factors of ETEC and EHEC [45] and ETEC, EHEC and Shigella [46].

*Neisseria gonorrhoeae* (*N. gonorrhoeae*), a Gram-negative diplococcus, causes one of the most common sexually transmitted diseases. It first adheres to the epithelium cells using pilli and opa proteins, then penetrates, multiplying on basement membranes, producing lipopolysaccharide endotoxins [47]. Four essential membrane proteins - D-alanine-D-alanine ligase (dil), sulfate transport permease protein C (cysW), competence lipoprotein (comL),
and type IV pilin protein (pIIV) of *N. gonorrhoeae* virulent strain FA 1090 - previously identified as vaccine candidates [48], were studied for antigenicity using VaxiJen and then used to identify epitopes for both B- and T-cell mediated immune responses [49,50]. Bhairamadgi and Katti [46] screened the proteome of *N. gonorrhoeae* strain FA 1090 using immunoinformatic tools (including VaxiJen) and identified a novel immunogen - the hypothetical protein YP_208831.1 – as having the potential to induce both T- and B-cell mediated immunity. The proteomes of four *N. gonorrhoeae* strains FA 1090, TCD_CG08107, NCCP11945 and MS11 were screened to identify probable secretory proteins [51]. From this initial pool, proteins with human homologues were removed, antigens selected by VaxiJen and B- and T-cell epitopes predicted. The final set consisted of 25 antigenic secretory proteins of *N. gonorrhoeae* which could be experimentally validated for vaccine development.

**Helicobacter pylori** (H. pylori) is a human gastric pathogen implicated as the major cause of peptic ulcer and second leading cause of gastric cancer around the world [52]. Based on the genome analysis of 39 H. pylori isolates, Ali et al. [52] selected 28 non-host homolog proteins as therapeutic targets. After selection by VaxiJen and epitope mapping, 3 highly conserved and 2 highly variable putative pathogenicity islands were revealed. Among them 5 potential vaccine candidates – vacA, babA, sabA, fcdA and omp16 – were prioritized [53]. A chimeric gene containing four fragments of FliD, UreB, VacA and CagL with a high density of B- and T-cell epitopes was designed and optimized in terms of solubility, antigenicity (by VaxiJen) and surface accessibility [54] for expression in *E. coli* BL21.

Campylobacter is one of the four key global causes of diarrheal diseases [55] and a major global cause of human gastroenteritis. Campylobacter infections are generally mild, but can be fatal among very young children, the elderly, and the immunosuppressed. Mehta and Ramana [56] screened the genome of *C. jejuni* pathogenic strain NCTC11168 and found 66 proteins non-homologous to the human proteome, 34 of them were deemed to be “drugable”. As *C. jejuni* constitutes the main animal reservoir, poultry vaccination is a promising way to reduce incidence of campylobacteriosis in humans [57]. Meunier et al. [57] screened the genome of the highly virulent Campylobacter jejuni subsp. jejuni 81-176 strain, selecting 24 extracellular and outer membrane proteins. Two of them were the known antigens FlaA and FlaB, the rest were newly identified antigens. After VaxiJen ranking, B- cell epitope mapping, and BLAST searching for conserved regions between C. jejuni and *C. coli*, 14 proteins were identified as potential vaccine candidates. The whole proteome of *C. jejuni* has been investigated for antigenicity, allergenicity, MHC binding, conservancy, and population coverage [58] and several conserved T-cell epitopes have been selected covering more than 80% of the human population.

**Acinetobacter baumannii** (A. baumannii) causes a variety of nosocomial infections of the respiratory tract, bloodstream, skin and soft-tissue infections [59]. Because of its biofilm-forming ability, it is resistant to most antibiotics [60]. The biofilm associated protein (Bap) of *A. baumannii* was searched for a conserved antigenic region using VaxiJen [61], identifying it as an antigen, with the potential to be a diagnostic. A comparative genome analysis was done to identify conserved proteins among five *A. baumannii* strains [62]. Three outer membrane proteins (outer membrane receptor protein, putative penicillin binding protein and glutamate synthase large chain precursor) with high VaxiJen scores were selected and B- and T-cell epitopes predicted. The protein *Abaumannii acinetobactin* (BauA) was investigated using bioinformatics tools (including VaxiJen) and two regions (26-191 of cork domain and 321-635 of barrel domain) selected as candidate vaccines [63]. Thirteen putative antigens were identified by a bioinformatic analysis of 30 *A. baumannii* strains [64]. The set included P pilus assembly protein, pil assembly chaperone, AdeK, PonA, OmpA, general secretion pathway protein D, HuEF receptor, Type VI secretion system OmpA/MotB, TonB dependent siderophore receptor, general secretion pathway protein D, outer membrane protein, peptidoglycan associated lipoprotein and peptidyl- prolyl cis-trans isomerase. The *A. baumannii* phospholipase D (pID) was analyzed for antigenicity by VaxiJen, followed by B- and T-cell epitope mapping [65].

**Vibrio cholerae** (V. cholerae) is a noninvasive gram-negative bacterium causing water borne disease cholera [66]. Despite extant vaccines - inactivated whole organism and B subunit of cholera toxin [2] - development of novel vaccines remains appealing. Barn et al. [66] selected three candidate vaccines (OmpU, UppP and YajC) from the *V. cholerae* strain O195 using bioinformatics tools (including VaxiJen) and searched them for B- and T-cell epitopes. Nezaal et al. [67] have analyzed 6 known V. cholerae protective antigens (OmpW, OmpU, TcpA, TcpF and CTB) to identify promiscuous epitopes, binding to various HLA class II alleles and B-cell epitopes. The identified epitopes were linked together and the fused protein was predicted by VaxiJen to be antigenic. Two B-cell epitopes from Omp containing T-cell epitopes covering HLA class I and class II alleles have been designed as candidate epitope vaccines [68].

**Salmonella Typhi** (S. typhi) is a Gram-negative bacterium causing human typhoid fever. Prabhavathy et al. [69] proposed S. typhi proteins OmpLA and LcrC as suitable vaccine candidates based on antigenicity predicted by VaxiJen, followed by prediction of B- and T-cell epitopes, and identification of common epitopes for multiple pathogens. Toobak et al. [70] identified three major outer membrane proteins OmpC, OmpF and OmpA, which were amplified, cloned and expressed. The antigenicity of Omps was predicted by VaxiJen and confirmed by ELISA. Control and all mice immunized with a single Omp gene died within 24 hours of challenge with S. typhi. Mice immunized with two Omps survived for 48-50 hours. Mice immunized with three Omps survived 75 h. Generally, despite the high immunogenicity of Omps, they did not induce long-lasting protection in mice. As mice are not the natural hosts, results in human may well be different [70].

Leptospirosis is caused by the pathogen *Leptospira* is one of the most widespread zoonotic diseases in the world [71]. Victor et al. [72] examined the evolutionary relationships of the outer membrane lipoprotein LipL41 taking 87 sequences from various *Leptospiral serovars* and strains followed by B-cell epitope mapping of conserved regions; identifying 8 LipL41 B-cell epitopes with high VaxiJen score. A similar systemic protein selection, antigenicity prediction and B- and T-cell mapping was applied to Omps of *Treponema pallidum* [73], *Shigella flexneri* [74], *Pseudomonas aeruginosa* [75] and *Clostridium botulinum* [76].

**Prediction of viral immunogens**

Influenza viruses are of four types: A, B, C and D. Type A infects a wide range of avian and mammalian species. Type B almost exclusively infects humans. Type C infects humans, dogs and pigs and causes mild respiratory illness, while type D only infects cattle. Among the four types, type A is the most virulent human pathogen and the most variable. It is classified into subtypes according to the serological reactivity of its surface glycoprotein antigens, hemagglutinin (HA)
and neuramidase (NA) [77]. Gupta et al. [77] collected 86 H1N1 HA protein sequences analyzing them for conservation. The analysis identified 15 conserved regions containing 13 HLA class I and 17 HLA class II epitopes. Their antigenicity was assessed by VaxiJen, with only 4 class I and 9 class II epitopes were proposed as a novel candidate epitope-based vaccine. Moattari et al. [78] compared 10 viral sequences collected from Iranian patients between 2010 and 2013 with 3 vaccine isolates: (A/H1N1)California/2009, A(H1N1) California X-157/2009 and A(H1N1)Brisbane/2007. This study detected several amino acids changes in HA which do not affect the epitope sites, antigenicity (predicted by VaxiJen), or secondary and tertiary structure of HA.

The Human immunodeficiency virus (HIV) causes acquired immunodeficiency syndrome (AIDS). HIV enters the host cell by forming a complex between the viral envelope glycoprotein (Env), the host receptor CD4, and chemokine co-receptors usually CCR5 or CXCR4 [79]. Antiretroviral peptides can inhibit the virus-echceptor interaction by binding either virus envelope proteins or host proteins [80]. Rao et al. [80] compiled a dataset of 110 HIV antiviral peptides and analyzed them for conservancy, antigenicity by VaxiJen, hydrophobicity and antimicrobial activity. Fourteen peptides have been selected for promising AIDS treatment.

Hepatitis C virus (HCV) affects between 130 and 150 million people worldwide and causes chronic liver disease, liver cirrhosis and hepatocellular carcinoma [81]. In silico analyses and B- and T-cell epitope predictions have been made for the structural envelope glycoproteins 1 (E1) [82] and 2 (E2) [83,84] and the non-structural proteins NS3, NS4A, NS5A and NS5B [85]. All predicted epitopes are antigenic according to VaxiJen (VaxiJen score > 0.4).

Zika virus (ZIKV) is a mosquito-borne virus causing mild headache, cutaneous rash, fever, malaise, conjunctivitis and arthralgia [86]. It consists of 10 proteins: capsid, precursor of membrane, envelope and seven non-structural (NS) proteins [87]. The most investigated is the envelope glycoprotein protein. Several B- and T-cell epitopes have been identified by immunoinformatic tools (including VaxiJen) [88-91], but have not been tested experimentally. Dar et al. [92] analyzed 54 full length ZIKV polyprotein sequences identifying 23 HLA class I and 48 HLA class II binders. The most of them are localized in NS5, followed by envelope, NS1 and NS2.

Dengue virus (DENV) is a mosquito-borne virus causing life threatening hemorrhagic fever and shock syndrome [93]. There are four DENV serotypes and several genotypes [94]. The DENV genome encodes 3 structural (capsid C, precursor of membrane prM and envelope E) and 7 non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5). The DENV nucleocapsid is covered by a lipid bilayer containing envelope glycoprotein and membrane proteins [95]. The extramembrane location of these proteins makes them good candidates for vaccine development. The proteins E, prM, NS1 and NS5 have been identified by VaxiJen as probable antigens [96], and several B- and T-cell epitopes have been predicted [96-98]. The non-structural protein NS5 is the most conserved DENV protein making it a good vaccine target [99]. It has low intrinsic antigenicity but contains several potent and promiscuous T- and B-cell epitopes [98,99].

Ebola virus (EBOV) causes a fatal hemorrhagic fever with death rates up to 90% [100]. Fruit bats are the natural host, and disease transmission to human and other primates is mainly via bodily fluids (blood, secretions, semen) [101]. The EBOV proteome has been searched for antigenic proteins using VaxiJen: the L protein (UniprotKB ID: K4G1K7) was identified as the most antigenic (VaxiJen score 0.7024) [102]. Further, a highly promiscuous B- and T-cell epitope was identified as a probable epitope-based vaccine [102]. The glycoprotein 2 (GP2) and viral protein 24 (VP24) have also been analyzed for conservation, antigenicity using VaxiJen, and promiscuous B- and T-cell epitopes [103].

Crimean-Congo hemorrhagic fever virus (CCHFV) is a tick-borne pathogen causing hemorrhagic fever with fatality rates between 15% and 70% [104]. A total of 80 envelope glycoproteins and 34 RNA-dependent RNA polymerase-L molecules from different CCHFV variants were analyzed for conservation, identifying 4 conserved regions [105]. Two regions were antigenic according to VaxiJen, and one nonamer epitope had the potential to interact with 8 HLA class I and 27 HLA class II alleles [105]. T-cell epitopes were also identified from the proteins GP1 and GP2 [106].

Human coronaviruses (HCoVs) cause mild to severe respiratory tract infections, and are named for the crown-like spike proteins on their surface. 56 outer membrane spike protein sequences from different variants belonging to five types (229E, NL63, HKU1, EMC, and OC43) were retrieved from UniProtKB and assessed for antigenicity using VaxiJen [107]. The spike protein (UniprotKB id: B2KKT9) with the highest VaxiJen score was searched for B- and T-cell epitopes, identifying two peptide sequences as conserved and promiscuous. Shi et al. [108] analyzed nucleocapsid (N) and spike (S) proteins of Middle East respiratory syndrome coronavirus (MERS-CoV) selecting 10 B- and T-cell epitopes as MERS vaccine candidates.

Chikungunya virus (CHIKV) is a mosquito-borne virus causing fever and severe joint pain [109]. The CHIKV proteome was analyzed by VaxiJen, identifying envelope protein 2 as the most immunogenic [110]. Kori et al. [111] have searched the proteomes of three different CHIKV strains for conserved regions and five B- and T-cell epitopes have been predicted from both structural and non-structural proteins.

VaxiJen, followed by B- and T-cell epitope predictions, was used to design epitope-based vaccines against several other viruses: Saint Louis encephalitis virus (SLEV) [112], Nipah virus (NiV) [113], Hantaan virus (HNTV) [114], human cytomegalovirus (HCMV) [115], Lassa virus [116], human enterovirus D-68 (EV-D68) [117], Cardamom mosaic virus (CdMV) [118], Henipavirus [119], Hantavirus [120], human rotavirus A [121], Lymphocytic choriomeningitis virus (LCMV) [122], human adenovirus E (HAdV-E) [34], and human papillomavirus type 16 (HPV) [123].

Prediction of parasite immunogens

Plasmodium falciparum (P. falciparum) causes malaria – the most important parasitic disease, killing one child under 5 year every 120 seconds [124]. P. falciparum has a large genome encoding over 5300 proteins, with stage-specific expression and variation within a single strain [125]. Singh et al. [126] used VaxiJen to identify 22 probable antigens, containing up to 15,000 predicted epitopes binding to HLA class I and II supertypes and covering 95% of the human population. The complete proteome of P. falciparum, after excluding human homologs, was subjected to subcellular localization prediction, with outer membrane proteins analyzed by VaxiJen [127]. Four membrane-associated hypothetical proteins were identified, containing B- and T-cell epitopes. Singh et al. [128] screened 32 extracellular secretory proteins of P. falciparum using VaxiJen, predicting 31 as antigenic and containing many epitopes binding to HLA-A, -B and -DR.

Leishmaniasis are a group of tropical diseases caused by protozoan parasites of genus Leishmania and transmitted to humans by...
hematophagous sandflies [129]. The entire proteome of *Leishmania major* (*L. major*) comprising 8312 proteins was screened for signal peptides and GPI anchors [130]. A set of 151 proteins was selected and subjected to consensus antigenicity prediction using VaxiJen and ANTIGENpro [131], identifying 25 vaccine candidates. *Toxoplasma gondii* (*T. gondii*) is an obligate intracellular parasitic protozoan causing the disease toxoplasmosis [132]. The *T. gondii* major surface antigen, called SAG1 or p30, was screened for conservation and antigenicity using VaxiJen [133], identifying four conserved antigenic regions.

### Prediction of fungal immunogens

*Fusarium circinatum* (*F. circinatum*) causes the economically important disease of pines pitch canker leading to devastating forestry industry losses [134]. The detection of the pathogen in plant growth media and in plant tissues during the early stages of infection is very important [135]. Maphosa et al. [135] conducted comparative genomic studies, identifying 24 unique antigenic proteins. The ORFs of 5 variants were selected and tested by PCR analyses and hybridization assays. The results have showed that three of the selected genes are common and unique to *F. circinatum* and are thus good candidates for rapid, in-the-field diagnostic assays specific to *F. circinatum*.

### Prediction of cancer immunogens

Cancer immunotherapy aims to stimulate immune responses of B- and T-lymphocytes and prevent immune system suppression by IL-10, PGE2 and COX2, secreted by tumor cells [136]. Nezafet et al. [137] designed a multi-epitope polypeptide cancer vaccine containing two tumor-associated antigens E6 and E7 from human papillomavirus as cytotoxic T-cell epitopes, tetanus toxin fragment c (TTToCi) and pan-allelic DR epitope (PADRE) as helper T-cell epitopes, and the TLR4 agonist heparin-binding hemagglutinin (HBHA) as an adjuvant, epitopes separated by proteasome-sensitive linkers. This construct was analyzed for immunogenicity by VaxiJen and ANTIGENpro, and B-cell epitopes identified on the protein surface.

The receptor tyrosine kinase like orphan receptor 1 (*ROR1*) is a transmembrane protein overexpressed in several cancers, including gastric carcinoma and breast cancer [138]. A chimeric receptor consisting of the extracellular domain of ROR1 and the powerful T-cell activator staphylococcal enterotoxin has been constructed as a potent vaccine for breast cancer and analyzed for antigenicity, allergenicity and the presence of B- and T-cell epitopes [139]. A multi-epitope vaccine against breast cancer was designed to include cytotoxic T-cell epitopes from human epidermal growth factor receptor (HER2), mucin 1 protein and heparanase as well as helper T-cell epitopes from survivin and Por B from *Neisseria meningitidis* (TLR2 agonist) as an adjuvant [140].

Wilms’ tumor gene WT1 is a zing finger transcription factor overexpressed in leukemias and solid tumors [141]. Khalili et al. [142] identified 44 novel epitopes in WT1 protein and constructed a DNA vaccine containing the predicted epitopes with acceptable population coverage (>65%).

### Prediction of immunogens for diagnostic tools

Leptospirosis is a global zoonotic disease affecting humans and causing severe icteric Weil’s disease, characterized by renal and liver failure [143]. It is caused by spirochetes of the genus *Leptospira*. The pathogenesis-associated leptospiral LigA protein expressed in vivo, has been evaluated as a diagnosis of the acute form of leptospirosis. The C-terminal sequence of LigA (LigA-C) was cloned into pET15b and expressed in *E. coli* [144]. Two B-cell-specific immunogenic epitopes have been predicted and synthesized as peptides for evaluation along with recombinant LigA-C. Selected B-cell epitopes showed increased sensitivity over recombinant LigA-C in single and combination assays for IgM antibody detection, and may be useful in early diagnosis of leptospirosis.

Hepatitis B virus (HBV) and human T lymphotropic virus type I (HTLV-I) are blood- borne viruses. Since HBV and HTLV-I infections are asymptomatic for a long time, people are typically unaware of infection [145,146]. Recombinant proteins from these viruses are used as capture antigens in ELISA blood screening tests [147,148]. A chimeric antigen comprising antigenic fragments of HBV core protein and proteins gp46 and p16 of HTLV-I was constructed, predicted by VaxiJen as antigenic, expressed in *E. coli*, purified and used as coating antigen for specific antibodies. The antibodies were detected successfully by the chimeric protein.

Shawky et al. [150] amplified the region encoding proteins E1 and E2 (HCV-E) from hepatitis C virus (HCV) genotype 4a, cloned it into a plasmid, and used this to immunize mice. The DNA construct was immunogenic, as predicted by VaxiJen. This study also found that combining the HCV-E construct with extracts from *Echinacea purpurea* and *Nigella sativa* prior to immunizing mice significantly increased both humoral and cellular responses.

VaxiJen has also been used to predict the antigenicity of cystatin C developed as a diagnostic tool for accurate estimation of glomerular filtration rate (GFR) [151]. A stable fusion protein was constructed to include immunogenic fragments of several proteins (ApoB-100, hHPSP60 and β-2-GPI) associated with atherosclerosis [152]. It could be used as a vaccine to prevent or modulate atherosclerosis. Bioinformatic tools (including VaxiJen) were used to locate a specific conserved region of ActA, a membrane protein from *Listeria monocytogenes* (*L. monocytogenes*) [153]. The region was used to design an antibody- based diagnostic test for *L. monocytogenes*.

### Prediction of immunogens for veterinary medicine

Brucellosis is a zoonotic illness transmitted from domestic animals to humans. It is caused by Brucella spp. Although a live attenuated vaccine against ovine brucellosis exists, investigations focus on developing a safer subunit vaccine. Several chimeric DNA vaccines have been designed to encode proteins omp19, omp31 and urease [154], BP26, omp31 and TF [155], and GroEL [156]. Based on systematic screening of the exoproteome and secretome of *B. melitensis*, Vishnu et al. [157] identified eight proteins as potential vaccine candidates, including LPS-assembly protein LptD, a polysaccharide export protein, a cell surface protein, heme transporter BhuA, flagellin FltC, 7-alpha-hydroxysteroid dehydrogenase, immunoglobulin-binding protein E1BE, and hemagglutinin. Several B- and T-cell epitopes were predicted, and also from the protein omp25 [158]. All these proteins were assessed to be antigenic by VaxiJen.

*Histophilus somni* (*H. somni*) is an opportunistic bacterial pathogen causing histophilosis in cattle and is associated with thrombotic meningoecephalitis and bovine respiratory disease [159]. The genome of 12 *H. somni* isolates were sequencing, protein coding regions predicted, and several programs (including VaxiJen) used to evaluate the antigenicity, surface exposure scores, and sequence conservation [160]. The first 20 ranked proteins have been analyzed in western blot with bovine serum, with 13 responding to bovine antibodies of *H. somni*.
Presented better vaccine potential than SodC. Aeromonas hydrophila causes aeromoniasis in fish [171]. Rauta et al. [172] analyzed two antigenic outer membrane protein (Aha1) peptides as DNA vaccine candidates including them in a nanoparticle-based delivery system.

Bovine rotavirus and bovine coronavirus are the most important causes of diarrhea in newborn calves as well as pigs and sheep [173]. Rotavirus protein VP8 and coronavirus S2 spike glycoprotein are the major determinants of viral infectivity and neutralization. A chimeric VP8-S2 gene was designed computationally using VaxiJen [173], cloned and sub-cloned into vectors, and transferred into E. coli. The expressed protein was purified and used to immunize hens. The activity and specificity of the isolated and purified anti-VP8-S2 IgY was detected using several experimental assays. The specific anti-VP8-S2 IgY was suggested as a candidate for passive immunization against bovine rotavirus and bovine coronavirus.

Fowl adenovirus serotype 4 (FAV-4) causes hydropericardium syndrome in domestic fowl. The hexon gene was isolated from 3 virus isolates and then cloned. The vectors were analyzed using online bioinformatics tools [174]. All were predicted to be antigens according to VaxiJen and to contain B-cell epitopes.

Newcastle disease virus (NDV) causes Newcastle disease, an extremely infectious viral disease affecting most bird species. Motamedi et al. [175] used an in silico approach, assembling potential and conserved epitopic regions of hemagglutinin–neuraminidase (HN) and fusion (F) glycoproteins of NDV to induce multipotent responses against the virus. Epitope predictions have shown that the hypothetical synthetic construct could induce immature B and T cell epitopes. Most regions of the construct had a high antigenic propensity and surface accessibility.

The cattle tick, Rhipicephalus microplus, is an obligate hematophagous ectoparasite of cattle occurring in the tropical and subtropical regions of the world and is a vector of disease causing pathogens, such as Babesia bovis, Babesia bigemina and Anaplasma marginale [176]. A systematic approach using a combination of functional genomics (DNA microarrays) techniques and a pipeline of in silico predictions of subcellular localization and protective antigenicity using VaxiJen was used to identify novel anti-tick candidate vaccines [177]. 791 candidates were identified, of which 176 were membrane-associated and 86 secreted soluble proteins. Five predicted membrane-associated antigenic peptides were selected and synthetic peptides designed and then tested using polyclonal antisera from BALB/c mice immunized with a crude extract of tick midgut membrane proteins. Three showed antigenicity higher than that of Bm86 peptides. Moreover, 19 novel transmembrane proteins were identified using LS-MS/MS and suggested as putative tick targets [178]. Aguirre et al. [179] identified using immunoinformatics an antigenic peptide from ATAQ protein, a putative Bm86 homolog. The pure peptide and its conjugate with Keyhole Limpet Hemocyanin (KLH) were tested for immunogenicity in mice, rabbits and cattle. Between 35% and 47% of the animals developed a consistent immune response.

Dogs are the domestic reservoirs of the parasite Leishmania infantum (L. infantum) causing visceral leishmaniasis (VL) in humans and dogs [180]. The control of canine VL could reduce human infection rates. Agalliou et al. [180] analyzed the total protein extract from late-log phase L. infantum promastigotes using two-dimensional western blots and probing with sera from asymptomatic and symptomatic dogs. Forty-two protein spots were found to react differentially with IgG from asymptomatic dogs. Of these, 21 were identified by mass
spectrometry and predicted with ANTIGENPro and VaxiJen for antigenicity. Six proteins were identified as novel candidate antigens able to be developed as vaccines or diagnostic tests.

Other methods for immunogenicity prediction

Following in the wake of VaxiJen, several groups have sort to emulate its success or to develop alternative approaches for the identification of whole protein antigens from microbial life. NERVE (New Enhanced Reverse Vaccinology Environment) was the first in silico pipeline for identification of the best vaccine candidates from whole proteomes of bacterial pathogens [181]. Reverse vaccinology is a method build on genome-based antigen discovery [182] and it has been successfully applied to the development of an innovative vaccine against N. meningitidis serogroup B [183]. NERVE uses four filters for selection of immunogens: localization, topology, probability of being adsenin and similarity to human proteins. The threshold values have been tuned on 10 proteomes containing known immunogens. As a result, NERVE ranks the most probable vaccine candidates.

VaxiJen is the first web-based vaccine design system that predicts vaccine targets based on genome sequences using the strategy of reverse vaccinology [184]. VaxiJen predicts protein subcellular location, transmembrane helices, adhesion probability, conservation to human and/or mouse proteins, sequence exclusion from genome(s) of nonpathogenic strain(s), and epitope binding to MHC class I and class II. The precomputed VaxiJen database contains prediction of vaccine targets for >70 genomes. VaxiJen also performs dynamic vaccine target prediction based on input sequences.

ANTIGENPro is a two-stage architecture based on structural features as sequence length, molecular weight, absolute charge per residue, turn-forming residues fraction, hydropathy, etc. and support vector machine (SVM) classifier [131]. The algorithm has been trained on two datasets: (i) antigens that elicit a strong antibody response in protected individuals but not in unprotected individuals, using human immunoglobulin reactivity data obtained from protein microarray analyses; and (ii) known protective antigens from the literature. ANTIGENPro has correctly classified 82% of the known protective antigens when trained using only the protein microarray datasets. The accuracy on the combined dataset was 76% by cross-validation experiments. ANTIGENPro has performed well also on an external pathogen proteomes.

Vaccine has been built around the concept of linked resources [185]. It consists of two parts: part A builds the proteome using gene predictors and similarity searches; and part B predicts protein characteristics such as localization, topology, number of transmembrane helices, binding to MHC class I and class II. The output is a list of vaccine candidates ranked according to the average probability of the classifiers used.

Jenner-Predict server has been developed for prediction of protein vaccine candidates (PVCs) from proteomes of bacterial pathogens [186]. The server targets host-pathogen interactions and pathogenesis by considering known functional domains from protein classes such as adhesin, virulence, invasin, porin, flagellin, colonization, toxin, choline-binding, penicillin-binding, transferring-binding, fibronectin-binding and solute-binding. It predicts non-cytosolic proteins containing these domains as PVCs. It also provides vaccine potential of PVCs in terms of their possible immunogenicity by comparing with experimentally known IEDB epitopes, absence of autoimmunity and conservation in different strains. Predicted PVCs are prioritized so that only few prospective PVCs could be validated experimentally. The performance of web server has been evaluated against known protective antigens from diverse classes of bacteria.

iVAX is an integrated set of tools for triaging candidate antigens, selecting immunogenic and conserved T cell epitopes, eliminating regulatory T cell epitopes, and optimizing antigens for immunogenicity and protection against disease [187]. iVAX has been applied to vaccine development programs for emerging infectious diseases, cancer antigens and biodefense targets. Several iVAX vaccine design projects have had success in pre-clinical studies in animal models and are progressing toward clinical studies.

VacSol is a high throughput in silico pipeline for vaccine candidate prediction from bacterial pathogens [188]. It consists of known tools like BLAST [189], PSORT [190], HMMTOP [191], ABCPred [192], ProPred [193] and databases DEG 10 [194], VFDB [195] and UniProt [196], integrated to work consequently. It is freely available to download from https://sourceforge.net/projects/vacsol/.

Proteome analysis is a bioinformatics tool for discovery of bacterial vaccine candidates based on “protective signatures” [197]. Authors have collected a database of all known protective antigens of 38 bacterial pathogens, have analysed them using BLAST [189], ClustalW [198], Smart [199] and Pfam [200] and have identified common structural features like function/biological role (toxins, iron-uptake systems, adhesins, etc.) and/or structural organization (multiple internal structural motifs). Support vector machine (SVM) classification has been applied to derive a model for discrimination between bacterial protective antigens and non-antigens [201]. Initially, the training set consisted of 136 antigens and 136 non-antigens (most of them taken from VaxiJen datasets), lately the training set was increased to 200 antigens and 200 non-antigens [202]. The validation data showed a better performance of the SVM model comparing to VaxiJen. The model is not accessible online.

Discussion

The role that immunoinformatics plays in vaccine design develops in several ways. Helping to design transgenic whole-organism pathogens, which cannot grow or cause harm, through designing out one or more virulence factors and so rendering the organism effectively harmless or by inducing severely compromised reproductive capacity of a virus or other microorganism [203]; Helping to design epitope ensemble vaccines. Such efforts fall into two camps: un-validated prediction-only methods that predict supposedly high-binding and more modern approaches that use immunoinformatics to select rather than predict the best epitopes suitable for forming a vaccine [204,205]; Helping to identify immunogenic and potentially protective single proteins from the genome of a given pathogenic microorganism. Such methodologies come in two main guises: “pipelines” or networks of methods and algorithms that together are able to select appropriate proteins [206,207] and single methods that seek to predict immunogens. VaxiJen typifies such an approach. VaxiJen is now a decade old. While so-called pipelines for vaccine design abound, methods that address the more fundamental questions of immunogenicity and antigenicity do not. VaxiJen stands alone; almost. Why is this? The main answer lies with the difficulty of the task – were it easy to improve on extant result then solutions would abound, as they have in other areas of immunoinformatics and bioinformatics - and the complex perception of the accuracy and veracity of the outcome. While the research objective of VaxiJen is both obvious and profound, current implementation is certainly sub-optimal. We need to improve and to test VaxiJen rigorously.

More specifically, we at least need to address the following: First, we need much more “positive” data; that is, carefully curated and validated,

Citation: Zaharieva N, Dimitrov I, Flower DR, Doytchinova I (2017) Immunogenicity Prediction by VaxiJen: A Ten Year Overview. J Proteomics Bioinform, an open access journal ISSN: 0974-276X
examples of protective antigens from pathogenic microorganisms and cancer. Databases of such do exist - AntigenDB [206] is, for example, a dedicated resource directly addressing this, while IEDB [206] also contains similar data, but as something of an afterthought, rather than being its focus. IEDB concentrates on epitopes and related information not antigens and protein immunogens. What we need is more and better data resources of this type upon which to draw upon. Thus there is a need for a concerted effort to enlarge, deepen, and broaden available data collations; much as has been done for epitopes [207].

At the same time, we also need much better and much more carefully constructed negative training sets and learning protocols. We need to balance the selection of negative test sets so that any signal present in the analysis reflects antigenicity and no other quality. We need to select similar protein lengths, similar origin species, similar subcellular locations, and similar functions: the list goes on. Imagine we wanted to separate a particular evolutionarily-conserved family of membrane protein, a fair test would be other membrane proteins of similar length from the same species; not soluble or fibrous proteins selected from a distant branch of the tree-of-life.

Better representations of the sequence data: currently VaxiJen employs Wold’s z-scales to characterise proteins using ACC transform. This works, but it is not clear that this is optimal. Other descriptors are available. Single descriptors characterising the whole sequence [208] and other multivariate descriptors of sequences. One could envisage a phase space of disjoint descriptor variables from which we could use variable selection protocols to select a compact and near-optimal choice of indicative variables.

Better algorithms: Artificial Intelligence or AI now is focused on the development of deep-learning algorithms. Powerful machine learning toolkits, such as Weka [209], are already available, and these are more than capable of delivering robust and extensible methods provided the data and the data representation are adequate. Nonetheless, as new algorithms do appear we must be open and embrace them. Avoiding complacency must be our mantra.

More validation: better protocols for establishing the immunogenicity and recall of identified vaccines. This is the world of experimentalist, and here what is needed is a fast, straightforward methodology which can be used to give much more consistent and much more accurate estimates of individual proteins.

More generally, as with most computational studies of real world problems, there exists a pressing need for experimental validation. The publication of an ever increasing stockpile of papers relating to the in silico analyses of pathogen genomes and virtual proteomes, have generated many potential vaccine candidates. Such papers typically use methodology largely embodied in web-servers; operating such systems is facile, and the resulting analysis straightforward. We need to highlight the severe limitations of all such studies in the absence of proper experimental validation. Although many papers are technically sound, their utility is hard to quantify and their significance questionable. Likewise, more examples of experimentally validated VaxiJen predictions are needed. A majority of papers citing the use of VaxiJen do not contain any such verification. Publishing unverified papers ultimately becomes counterproductive. Other studies give credence to their computational results [18,41,44,149,150,160,164,166,167,173,179] by combining vaccine design with experimental validation in animal models. Even in the days of AI hysteria, prediction without validation exerts little influence and convinces few.

Reference


