Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Bioinformatics in Design of Antiviral Vaccines

Ashesh Nandy, Centre for Interdisciplinary Research and Education, Kolkata, India
Subhash C Basak, University of Minnesota Duluth, Duluth, MN, United States

Viruses and vaccines

Viruses

Viruses are one of the smallest of biological entities that infect all life forms, from bacteria and fungi to animals and plants. The tobacco mosaic virus was the first virus to be so identified and since then over 5000 viruses have been investigated in detail although there are probably millions of viruses in existence and more are being discovered, the latest among these being a cluster of about 1500 viruses in molluscs discovered in late 2016.

Among the viruses that are pathogenic to humans one could count influenza, measles, whooping cough, plague, and polio to name a random few. Their numbers are augmented occasionally by new zoonotic viruses that make the jump from animal hosts to infect humans and create new diseases; notable among these was the human immunodeficiency virus (HIV) that arose from the simian immunodeficiency virus and began infecting humans around the 1980s, and the Zika virus epidemic that is affecting a part of the Americas at this time whose source is believed to be primates in Africa. When new viruses arise against which the human body is not prepared with defense mechanisms, it could lead to widespread epidemics and pandemics; the Spanish flu of 1918 was one such instance which spread rapidly across the world and resulted in over 20 million, also estimated to be 100 million, fatalities; **Table 1** provides a glimpse of the more outstanding viral epidemics in the 20th and 21st centuries to show how pathogenic viral diseases can be even with all our understanding to date.

The human body defends itself against invading foreign pathogens through an elaborate immune response system where the invaders are identified and then systematically eliminated. There is a long history of efforts to aid the process and combat viral

Table 1 Selected viral epidemics and human fatalities in 20th and 21st centuries

Year	*Virus*	*Effect*	*Inferred origins*
1918	H1N1 Spansh FLU	~20–100 million deaths	Birds?
1957–58	Asian flu	1–1.5 million deaths	
1980 to date	HIV-1	~25 million deaths	Common chimpanzee
2002–03	SARS	14%–15% fatalities	
2009	Swine flu	18,000–284,000 deaths	
2013–16	Ebola	90% fatalities	
2015 to date	Zika	>3500 microcephaly cases in Brazil	Rhesus monkey

 https://doi.org/10.1016/B978-0-12-801238-3.10878-5

Table 2 Landmarks in viral epidemics

Year	*Virus*
1545	Smallpox epidemic in India
1699	Yellow fever epidemic in America
1798	Edward Jenner publishes his account of the effects of his smallpox vaccine
1885	Louis Pasteur successfully prevents rabies by postexposure vaccination
1908	Karl Landsteiner and Erwin Popper discover poliovirus
1918	Millions die from the Spanish flu
1935	Max Theiler develops live-attenuated 17D yellow fever vaccine
1947	Zika virus isolated in Uganda
1960–69	Live attenuated vaccines developed for Measles, Mumps, and Rubella
1976	Ebola virus isolated in Zaire
1980	World Health Organization declares global eradication of smallpox
1981	Hepatitis B vaccine is licensed
1986	Yeast-derived recombinant hepatitis B vaccine is licensed
1994	Polio declared eliminated from Americas
2002	Polio declared eliminated from Europe
2006	First HPV vaccine is licensed
2013	Ebola virus outbreak in Africa
2014	Polio declared eliminated from India
2015	Zika epidemic in Brazil
2016	Guinea declared Ebola free after successful vaccination drive

menace; some of the landmark achievements are listed in Table 2, but much more needs to be done to fight against recurrent and new epidemics. Current efforts at rational design of antiviral vaccines to augment this effort are based on our understanding of the microbiology and genomics of viruses and the human immune system, which we recount here very briefly.

Viral Genomics

When not in action against any organism, viruses exist as individual particles which are referred to as virions. The virions consist of a protein shell, called a capsid, with the genome of the virus inside; in the case of some viruses, there is also a lipid envelope covering the capsid. Some viruses, called DNA viruses, carry single-stranded or double-stranded DNA genomes; other viruses, called RNA viruses, may carry double-stranded RNA or single-stranded, positive or negative sense RNA genome. The genome itself could be a continuous strand, as in the case of dengue, or could be in separate pieces, one for each constituent gene as in the case of the influenza genome, and referred to as segmented genome. The genes themselves can be grouped as structural and nonstructural. The structural genes contribute to the make up of the capsid, that is the viral coat, which is formed from many identical protein subunits, which may consist of one or more proteins grouped together, and can take up forms such as icosahedral, helical, prolate, and so on. The nonstructural genes function in replication of the genome, among other functions.

When a virus attacks a cell, a part of the surface proteins will attach to the cell surface through its segments called receptors and then fuse into the interior of the cell in a process known as endocytosis. Inside the cell the outer layers of the virion are removed by viral or host enzymes and the viral genome is released. The viral genes are then involved in synthesis of viral messenger RNA (mRNA), viral protein synthesis, and replication of the viral genome using the machinery of the host cell to produce multiple copies. During replication there may be errors in copying the nucleotides, resulting in a mutated gene. DNA viruses have the ability to use error correcting enzymes to undo the damage; RNA viruses do not have such machinery and so accumulate errors very rapidly. It is estimated that RNA viruses accumulate errors at the rate of $\sim 10^{-4}$–10^{-5} per nucleotide per replication; viruses that are about 10,000 nucleotides long in their genomes such as influenza or dengue viruses will accumulate errors at the rate of about 1 nucleotide per replication. DNA viruses on the other hand will accumulate errors at a rate of $\sim 10^{-8}$–10^{-11} and therefore will be much more stable. Accumulation of errors, called genetic drift, may lead to creation of new strains of the virus; other methods of generation of new strains are through reassortments in segmented genomes where whole genes may be exchanged with another homologous strain, or, more rarely, recombination where parts of a gene may be exchanged with similar segments of another homologous strain.

After replication, the products move into the endoplasmic reticulum and the Golgi apparatus where the different elements are put together. The assembled virus particles, the mature virions, then are released from the host cell by lysis, a process where the viral products break through the cell membrane and are ready for the next infection while the host cell is destroyed. Some viruses acquire a lipid envelope during the lysis process.

Immune Response

All organisms are equipped with some degree of immunity against invading pathogens and microorganisms. This immunity can be considered in two parts—innate immunity and adaptive immunity. Innate immunity is provided by our dry outer skin and mucous

membranes in certain openings like the mouth, nostrils, uterine passages, etc., the body's temperature, the highly acidic environment in the stomach and many others, all of which act to prevent pathogens and microorganisms from harming the host, but are not necessarily specific to any particular pathogen or its varieties. Pathogens and microorganisms that may enter the body through cuts and bruises are ingested by phagocytes, the macrophages and dendritic cells, and finally eliminated.

Adaptive immunity is the body's response to specific pathogens, referred to as antigens, that pass the innate immunity barrier through various means. Adaptive immunity is of two types—humoral immunity that is rooted in blood serum and other body fluids (known as humors in ancient times) and cell-based immunity. In both cases the immune response is mediated primarily by two cell types—B-lymphocytes (B-cell) and T-lymphocytes (T-cell), both produced in the bone marrow. The B-cells mature in the bone marrow and on leaving express antigen-binding receptors, called antibodies which are unique to each B-cell, on their membrane; the T-lymphocytes, or T-cells, move from the bone marrow to the thymus to mature, on leaving which each T-cell expresses a unique membrane-bound molecule called T-cell receptor which also binds to antigens under certain conditions. Each of the antibodies or receptors is encoded by specific genes which undergo random rearrangements and create different antibodies. Every B-cell or T-cell has thousands of antibodies or receptors on their surface, each cell having their own unique and identical antibodies; collectively, there would be about 10^{10} B-cells and 10^{9} T-cells produced in the bone marrow, but these numbers are reduced to some extent by filters to prevent autoimmune threats from the antibodies. These antibodies bind to the antigens with extreme specificity; each antibody can bond with a specific antigen and even a change in one amino acid of the antigen can nullify the bonding. The availability of millions of B- and T-cells with their own antibodies provides diversity and specificity in the antibody–antigen binding against multifarious pathogens.

B-cell receptors can recognize foreign antigens on their own. Naïve B-cells, that is, B-cells that have not yet encountered an antigen, split immediately on encountering an antigen into a memory B-cell and an effector B-cell, the progeny B-cells having identical membrane-bound receptors as the naive B-cell. The memory B-cells survive much longer than the naïve B-cells; effector B-cells live for a short period but release antibodies at a very high rate, about 2000 per second. On first encounter with an antigen, the immune response can take 5–6 days to start and peak at about 14 days; at the second attack the response occurs very fast, in about 1–2 days, and the intensity is very high. Vaccination primes the body for just this kind of response.

T-cells also have a similar response profile. There are two subtypes of T-cells—T-helper cells, T_H, and T-cytotoxic cells, T_C; the T-helper cells each have an associated CD4 membrane glycoprotein, and the T-cytotoxic cells each have a CD8 membrane glycoprotein associated with them. The T-cells recognize antigens when coupled with special protein molecules known as major histocompatibility complex (MHC) molecules which are displayed on the surfaces of cells. The human MHC molecules are expressed by certain HLA (human leukocyte antigen) alleles and contain a cleft at the distal end from the membrane that can contain peptides. There are two main classes of the MHC—MHC Class I which are expressed in almost all vertebrate nucleated cells, and MHC Class II which are expressed in only antigen presenting cells, that is, cells like macrophages, dendrites, etc., that ingest invading antigens by phagocytosis or endocytosis and degrade the complex antigen into small peptides. The peptides get attached to the MHC Class II molecules which then rise to the surface of the cell and are targeted by the T_H cells triggering the mechanisms for their eventual elimination.

Antigens that invade host cells and replicate get their peptides attached in the endoplasmic reticulum to the MHC Class I molecules which then move to the cell surface. The T_C-cells recognize the MHCI-peptide complex and proliferates and differentiates into effector cells known as cytotoxic T-lymphocytes (CTL). The CTLs monitor the affected cells and kill them.

The attachments of the degraded peptides to the MHC molecules are governed by the conformation of the clefts in the distal ends. The HLA that give rise to the MHCs are highly polymorphic and polygenic; a HLA database lists over 1600 HLA class II molecules alone; the alleles differ between individuals and between populations. Understandably, experimental approaches to determine antigenicity and traditional vaccines where effectiveness of the MHC-peptide bonding has to be tested are prohibitively time-consuming and expensive processes. Computational techniques coupled with robust algorithms to determine the binding affinities of peptides to the MHC molecules become a more reasonable approach. Still, there is a dearth of good data on many alleles that make up a population's HLA profile, and dependence on good models of homology, molecular docking, and structural data remain imperative. Bioinformatics therefore lead the current search for effective vaccine targets.

Vaccines

Vaccines relate to adaptive immunity, being very specific to the pathogen addressed. This is of vital importance to the success of the vaccination program. RNA viruses such as influenza mutate very rapidly and vaccines developed against an influenza strain in one year may not work against the strain of the next year; the Centers for Disease Control and Prevention in the USA prescribes different vaccines every year to combat flu strains they expect would be the major trends in the year. Drugs face the same hurdle: drugs such as Relenza developed against influenza have been rendered useless in a few years because of specific mutations in the flu strains, and the current favorite oseltamivir drugs, marketed as Tamiflu, are also reported to be facing resistance in some parts of the United States and Japan due to viral mutations. No vaccines have been devised yet against a common viral disease like dengue; rotavirus, an intestinal viral infection that kills millions of infants each year, have witnessed repeated vaccine failures, albeit with some complications in postvaccination situation. DNA viruses are much more stable and vaccines can be active for far longer periods. This is the reason smallpox could be eradicated, most of the world is almost polio-free, and human papillomavirus can be addressed by commercially available vaccines.

The success of a vaccine is therefore critically dependent upon the nature of the virus, its genomic content, and understanding of its antigenic properties. How the immune responses are evoked by the viral proteins is fundamental to the understanding of virus action.

Current vaccines

Vaccines in use at present are generally one of several types. The first is the attenuated or live-attenuated virus where the viral genome is altered so that the virus becomes less virulent or harmless, but can still elicit immune response; vaccines against measles, mumps, rubella, and several other viral diseases are of this type. These vaccines have a problem that sometimes mutations in the genome can make it revert back to the more virulent form and infect the human host.

Another conventional vaccine is the inactivated variety where the genome of the virus is removed or killed by heat or chemical treatment and only the empty capsid is used to generate the immune response. Some varieties of influenza and polio vaccines are manufactured in this manner. Depending upon the degree of treatment, such vaccines may require one or more booster shots to retain activity. Also, since the entire virion shell is used for the vaccine, these can cause allergenic reactions in susceptible persons.

A third type of vaccines is the VLP or virus-like particle. In this instance, surface proteins of the virion that can self-assemble into a structure like the original virus are constructed to elicit immune response, often with the aid of an adjuvant. Such VLP vaccines are difficult to design and manufacture. To date VLPs have been licensed against only the hepatitis E virus and human papillomavirus.

An interesting application of these ideas, as yet at an experimental stage, has been reported recently. A group at the Centre for Infectious Disease Research, Seattle, United States, has created a malaria parasite vaccine by genetically deleting three genes that are expressed in the preerythrocytic stage from the *Pseudomonas falciparum* which now becomes incapable of causing infections. The vaccine has been found to evoke good immune response against malarial infection with full safety in mice models and in tests with human volunteers, and further trials are planned.

New Approaches—Bioinformatics

New Paradigm in Vaccine Design

These traditional approaches to vaccines discussed above require long development periods and results are not always satisfactory. Apart from the ever-present threat of attenuated viral vaccines mutating to the pathogenic form, traditional vaccine design generally use the whole protein of the virus coat; the antigens can evoke immune responses but also sometimes cause allergy in humans, a type of autoimmune reaction. Also, a vaccine developed at great cost suitable for one community may not work for another where the people are more susceptible to different strains of the same virus; for instance, human papillomaviruses occur in many types and not all type combinations afflict humans universally.

The rapid advances in recent decades in information technology, high-throughput sequencing of DNA and RNA genomes, the completion of the initial stage of the Human Genome Project, better understanding of immunogenetics and developments in immunoinformatics are providing valuable insights into how the human immune response works to guard against and eliminate foreign pathogens that gain entry into our body. This has led to development of the science of "vaccinomics" that seeks to utilize our understanding of the body's immune reactions to precisely target those parts of the virus, the antigenic determinants, that will enable the immune response to kill the virus. Along with the data on a person's DNA, vaccinomics envisages an ideal scenario, admittedly in the distant future, where these genetic and viral information can be tailored into a vaccine that would be best suited to the individual—personalized immunizations that are safe, efficient, and effective for everyone without worries of side effects. The new discipline of bioinformatics becomes an essential tool to achieve this goal.

Introduction to Bioinformatics

Bioinformatics plays a central role in the new paradigm of vaccine design. Bioinformatics is an interdisciplinary field bringing together concepts of biology, mathematics, statistics, and computer science to provide a quantitative basis for understanding, classifying, modeling, querying, and analyzing biological data. With the development of bioinformatics, it has been possible to construct extensive databases of various biological quantities such as DNA, RNA, and protein sequences, analyze the huge quantities of accumulated data for similarities and dissimilarities to understand newly sequenced sequences and their interrelationships, and gain new insights, for example, where we learnt of the fractal nature of DNA sequences; the newly evolved science of molecular evolution would scarcely have been possible without bioinformatics. Recent progress in and wide dissemination of Internet resources have enabled several tools of bioinformatics analyses to be available and universally accessible across the web. These developments have spurred innovation and progress in biological data analyses and are contributing significantly to new ways of analyzing viral structures and designing vaccines.

The most widely accessed preeminent resources in bioinformatics are provided by the National Center for Biotechnology Information (NCBI) of the National Institute of Health, United States, and the European Molecular Biology Laboratory (EMBL) which have extensive databases of biological information and analytical software. Databases include GenBank, a repository of all sequences at NCBI which are augmented by submissions by researchers of new sequence data, Protein Data Bank (PDB) which contains information on protein sequences and structures, PubChem that contains data on chemical structures, and several other databases; equivalent records and access to the databases are maintained by EMBL and by DDBJ (DNA Databank of Japan), part of

the International Nucleotide Sequence Database Collaboration (INSDC). Nucleotide sequences record evolution of organisms more directly than any other biological information and thus the INSDC constitute an invaluable resource for study of living systems. The initial submissions to these databases are mainly from regional researchers and these are then distributed among the collaborating institutions in approved format for worldwide access. To aid in comparative and homology studies of sequences, popular software tools available with these institutional sources include BLAST (Basic Local Alignment Search Tool), ClustalW for multiple alignment of sequences, Cn3D for viewing the 3D structure of available protein structure data, and many others. Complementing the efforts of these major institutions are numerous other bioinformatics tools developed for specific tasks available across the web. We cover a selection of these in a later section in some detail.

Bioinformatics in Vaccine Design

While bioinformatics provides the tools to analyze diverse sequence data, practical considerations perforce sets limitations to what can be achieved at this time. The personalized vaccine approach of ideal vaccinomics is still a far cry, but a suboptimal approach to vaccinomics that avoids the individualized human genome part and optimizes on the viral information, "reverse vaccinology", is yielding results.

Reverse vaccinology

The reverse vaccinology model, part of the vaccinomics regime, uses bioinformatics techniques to screen entire genomes of pathogens to determine genes that could lead to good epitopes, the peptides in an antigen to which the antibodies actually bind, and proteins that are surface situated. The selected proteins/peptides are then synthetically produced and tested in animal models. To date this technique has been used against bacterial pathogens and the first human vaccine devised using this method was the Serogroup B meningococcus which causes meningitis. Although there had to be several trials and experimentation, the entire exercise took only a few years whereas traditional vaccine development methods normally take decades. In the case of this particular pathogen there were difficulties in using traditional methods because of autoimmune threats; use of bioinformatics methods were of prime importance in identifying surface antigens that could be used for designing the vaccine.

Peptide vaccines

Extending the reverse vaccinology model to cover viral diseases, many vaccines have been proposed and a large number are undergoing field trials in various phases. Essentially, the method in these cases is to (a) scan the viral genome for surface proteins, (b) analyze the surface proteins to predict the best epitope regions, (c) ensure these epitopes are surface situated, and (d) test them for autoimmune threats (see schematic, **Fig. 1**). Only on successful completion of these steps can the viral peptides matching the epitope requirements be recommended as vaccine targets. These have then to be prepared for in vitro and in vivo tests and field trials launched in animal and human models whose results will finally determine the suitability of such peptide vaccines. Among the first cases reported of success using such a model was against the virulent canine parovirus, which then stimulated similar searches against malaria, swine flu, and other diseases in animals. Proposals have been advanced that ovarian and breast cancers can benefit from the peptides derived from cell surface-associated mucin gene *MUC-1*. Initial field trials have shown that a peptide antigen vaccine against the lymphocytic choriomeningitis virus provided adequate antiviral response; a mixture of four peptides against nonsmall cell lung cancer was found to be safe and capable of generating strong T-cell responses.

The current authors have added a further step to the four mentioned earlier for rational design of peptide vaccines, viz., conserved segments. The issue is that viral genomes, especially RNA viral genomes, undergo rapid mutational changes rendering drugs and vaccines against one strain of a virus rapidly obsolete. To minimize this risk, the first step taken is to analyze as many viral protein sequences as possible for similarities and dissimilarities to determine segments that remain essentially conserved even though the viral genomes undergo numerous mutational changes. The procedure adopted to make this analysis is another application of bioinformatics techniques.

An almost universal approach to determining similarities and dissimilarities in a set of biomolecular sequences of a particular virus depends upon quite sophisticated software that determines segments that are similar across various sequences and introduces gaps where they are not to maximize some predefined measure of similarity. This technique of sequence alignment has enabled discovery of essential similarities between sequences and determine evolutionary relationships. Software such as ClustalW for alignments and MEGA (molecular evolutionary genetics analysis) to determine phylogenetic relationships is very popular for the purpose. However, these are model dependent about how the gaps are to be assessed and how the scoring is to be done to determine the best fit for alignment of multiple sequences; also, alignment for many sequences take a lot of computer time and could rapidly become inefficient in handling over a hundred sequences. Alignment techniques are very reliable when the sequences being aligned are very closely related, whereas results for divergent sequences are not always very reliable and are therefore often not used.

As an alternative approach, alignment-free techniques, barely a couple of decades old, are becoming increasingly prevalent. Several methods were proposed to compute sequence differences by the concept of k-mer/word frequencies where the bases were grouped in batches and analyzed; other techniques involved analyses by substrings within sequences and by information theory. In another technique which has gathered a lot of momentum, the initial impetus was provided by graphical representations of DNA sequences in 2D, 3D, and higher dimensions, which then were assessed quantitatively to provide a measure of the degrees of similarity and dissimilarity between sequences; the quantitative measures are computed either by geometrical means or by matrix

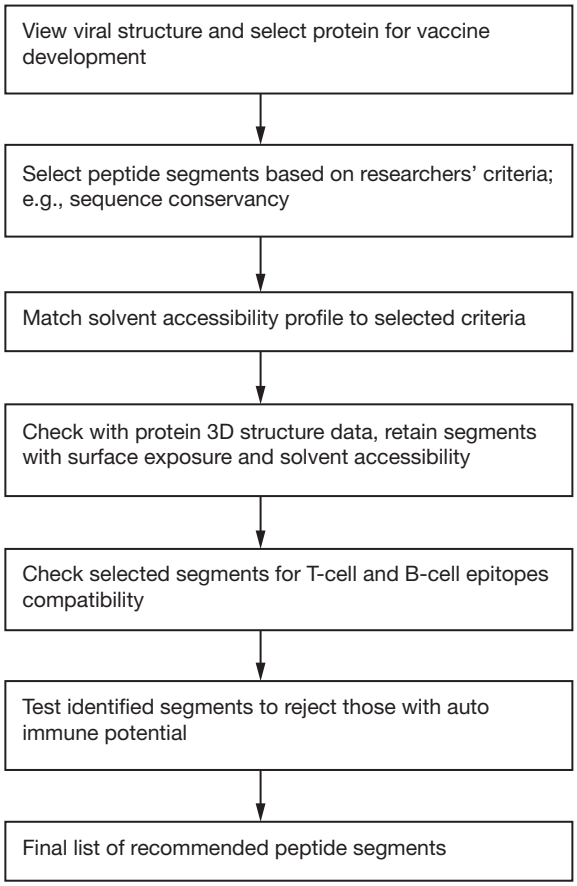

Fig. 1 Work flow chart of peptide selection process. Reproduced from Nandy, A. and Basak, S. C. (2016). A brief review of computer-assisted approaches to rational design of peptide vaccines. *International Journal of Molecular Sciences* **17**, 666.

methods (see Video 1 in the online version at https://doi.org/10.1016/B978-0-12-801238-3.10878-5 for a brief introduction). In the 2D model, specifically, the geometrical approach is intuitive and simple: For a DNA or RNA sequence of N nucleotides, one starts from the origin of a two-dimensional Cartesian coordinate system and proceeds one step in the negative x-direction for an adenine, one step in the positive y-direction for a cytosine, one step in the positive x-direction for a guanine, and one step in the negative y-direction for a thymine (uracil for RNA). Doing this successively for each base in the sequence starting from the beginning traces out a graph in the xy-plain which reflects the base distribution in the sequence. For each base there is a (x, y) coordinate and one can define weighted center of mass

$$\mu_x = \sum_{i=1}^{N} x_i/N \quad , \quad \mu_y = \sum_{i=1}^{N} y_i/N$$

and a graph radius as the distance from the origin to the center of mass,

$$g_R = \sqrt{\mu_x^2 + \mu_y^2}$$

which will be unique to a sequence and form a kind of descriptor of a sequence. The definition also allows distance between two sequences to be measured as

$$\Delta g_R = \sqrt{\left(\mu_x - \mu_x'\right)^2 + \left(\mu_y - \mu_y'\right)^2}$$

which enables a phylogenetic tree to be drawn to show the relationship between the sequences of a closely knit family. The beauty of the graph radius defined earlier is that identical values of the g_R imply identical sequences or sequence segments. This enabled the current authors to trace the movements of avian flu viral strains across countries and determine incidences of recombination events in influenza hemagglutinin. There are other methods of computing sequence descriptors, mainly by matrix methods where the eigenvalues, especially the leading eigenvalue, of the defining matrix serve to identify a sequence, as can be seen in the reviews listed in Further Reading.

The concept of graphical representations was extended by several authors to cover protein sequences, slightly more complicated since there are 20 amino acids to contend with. We adopted a virtual 20-dimensional (20D) rectangular representation where each amino acid was assigned to one axis and the protein sequence plotted in the imaginary system by taking one step at a time along the axis specified by the amino acid in the sequence. This plots out a 20D curve in the hyperspace where, for a protein sequence of n amino acids we could define weighted center of mass by

$$\mu_1 = \sum_{x_1} \Big/ n, \mu_2 = \sum_{x_2} \Big/ n \ldots \mu_{20} = \sum_{x_{20}} \Big/ n$$

and the graph radius

$$p_R = \sqrt{(\mu_1^2 + \mu_2^2 + \ldots + \mu_{20}^2)}$$

where again the p_R, with suitable parameterization described in our papers, were found to identify identical protein sequences when the p_R values were identical. Techniques were developed by the current authors and their group to determine extent of variances between different strains of a virus, movement of identical strains across countries, segments of RNA and protein sequences that remain conserved across many mutations, and many more.

The procedure for rational design of peptide vaccines then encompasses the following steps (see schematic, **Fig. 1**):

(a) Select the surface situated protein of the viral coat; for example, hemagglutinin or neuraminidase of influenza, or the L1 and L2 proteins of the human papillomavirus (HPV);
(b) Scan as many of the viral protein sequences as possible to determine segments that change least, that is, where p_R values for each window of particular segment length change the least between the sequences; segment lengths should be between 9 and 15 amino acids in consonance with the cleft sizes of MHC Class I and Class II proteins;
(c) Analyze the protein sequences for their average solvent accessibility scores to determine which of the above segments qualify as potential surface peptides;
(d) The peptides selected after the above step are then analyzed through appropriate software to predict their potential for linear and conformational epitopes;
(e) The peptides that pass this test should next be checked against a 3D protein structure, if available, to ensure that the selected peptides do indeed lie on the surface and are not covered by neighboring proteins in a multimeric quaternary structure;
(f) The peptides that qualify after this test should be tested next for autoimmune threats by BLAST analysis that compares the identified peptides against human protein sequences.

Peptides that pass these rigorous tests can then be expected to be target vaccines to elicit human immune response. An example of a protein with identified peptide regions in influenza A neuraminidase N1 protein is given in **Fig. 2** and Video 2 in the online version at https://doi.org/10.1016/B978-0-12-801238-3.10878-5.

There are several other considerations, however, before these peptides can be classified as vaccines. There is the question of adequacy of the immune response. Often adjuvants are required to augment the immune recognition process. Then there is a need to enable these peptides on a carrier protein to ensure that they do not fold up into some shape and therefore evade the antibody binding envisaged. Last, but not least, it may be necessary to have more than one peptide vaccine injected at a time for best results, the so-called multivalent vaccine approach. However, the major bioinformatics steps taken as outlined earlier helps to narrow down possible peptides to a few and expedite the vaccine development process far beyond what traditional vaccine development techniques allowed.

Some of the bioinformatics software used most often in the rational design of peptide vaccines as outlined earlier are briefly mentioned later.

Bioinformatics Tools

Epitope Predictions

One of the most important steps in analyzing an invading pathogen for possible vaccine targets is identifying the epitope regions. Accumulated databases of identified epitopes and antigens have established a starting point to determine potential epitopes in new antigenic determinants. The analyses for this exercise have been done using different methodologies including neural networks and QSAR (quantitative structure activity relationships). Among the many epitope prediction software, some of the more widely used web-based servers are the following.

IEDB (Immune Epitope DataBase) Analysis Resource provides a collection of tools for the prediction and analysis of immune epitopes. T-cell and B-cell epitopes prediction tools in the software suite can predict binding affinities for MHC Class I and Class II molecules. It includes tools to predict intrinsic potential for a peptide to be a T-cell epitope based on proteasomal processing, transporter associated with processing (TAP), and MHC Class I binding. The B-cell tools predict linear epitopes, and include Ellipro which analyses a protein antigen's 3D structure in terms of complex geometrical shapes to predict possible linear and conformational epitopes. An important analyses tool in IEDB is population coverage where the fraction of individuals that can respond to a given set of epitopes is predicted based on the HLA genotypic frequencies assumed for the target population.

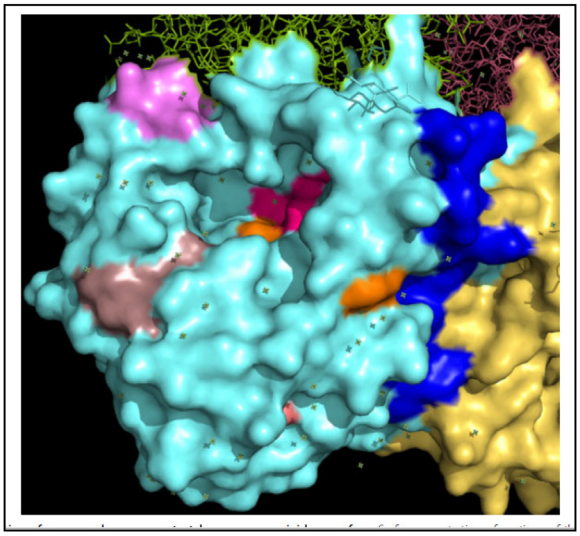

Fig. 2 The 3D space filling model of the neuraminidase protein showing surface exposed, conserved peptide segments, here shown in different colors, on one monomer (in cyan) of the quaternary structure. All peptides can be seen in Video 2. The 3D space filling model of the neuraminidase protein showing surface exposed, conserved peptide segments, here shown in different colors, on one monomer (in cyan) of the quaternary structure. All peptides can be seen in Video 2 in the online version at https://doi.org/10.1016/B978-0-12-801238-3.10878-5. Reproduced from Ghosh, A.; Nandy, A.; Nandy, P. (2010). Computational analysis and determination of a highly conserved surface exposed segment in H5N1 avian flu and H1N1 swine flu neuraminidase. *BMC Structural Biology*. **10**, 6, Fig. 3.

The ABCPred web server uses artificial neural network to predict B-cell epitopes in an antigen sequence. The training set uses 700 B-cell epitopes and 700 non-B-cell epitopes (random peptides) of maximum length of 20 residues and has achieved an accuracy of 65.93% in epitope prediction.

VaxiJen is the first on-line server for alignment-free prediction of protective antigens based on physicochemical properties of proteins. It derives from the principal amino acid properties of the protein sequences and uses bacterial, viral, and tumor protein datasets to design models for protein antigenicity prediction. Test sets showed prediction accuracy of between 70% and 89%.

MHCPred uses an additive method to predict the binding affinity of the peptides to the MHC Class I and Class II molecules. It also predicts the TAP binding affinity.

EpiJen predicts binding affinity for MHC Class I molecules for T-cell epitope prediction by taking three types of interactions into account: the interaction between individual amino acids and the binding site, the interaction between adjacent and every second amino acids, and their effects on binding. The modeling considers stages of the antigen presentation by the MHC Class I, including proteasome cleavage and TAP binding, to predict the most effective epitopes. The results are often concise, predicting only a small percent of possible peptides as epitopes.

NetCTL1.2 server offers a tool to predict cytotoxic T-lymphocyte (CTL) epitopes in protein sequences. It uses artificial neural networks to predict the CTL epitopes for MHC Class I binding and proteasomal cleavage. The method also integrates TAP transport efficiency. The prediction covers 12 MHC super types and has been trained on a set of 886 known MHC class I ligands.

Average Solvent Accessibility (ASA) Prediction Tools

One of the main considerations in effectiveness of antibody binding to epitopes is the accessibility to the antigen. It is important therefore to determine solvent accessibility through the hydrophobicity indexes. There are several software available to determine such indexes. SABLE, I-TASSER, HHPred, etc. are commonly used for the purpose. They provide indications amino acid-wise of surface accessibility of the residues with accuracies in excess of 70%. Coupled with the epitope prediction data, these provide an indication of which epitopes could be accessible for the antibody–antigen binding.

3D Structure

3D structure information of the viral surface protein under consideration for vaccine design is useful to ensure that the epitopes considered surface situated by the ASA profile computation are not covered by any neighboring protein in a multimeric structure, for example, in influenza hemagglutinin or neuraminidase. Such information is also required for the Ellipro analysis, especially for conformational epitopes which are the dominant MHC Class II epitopes in viruses such as human papillomavirus. Commonly used software to view 3D structure data are Cn3D available from the NCBI website, and PyMOL and RasMOL which are powerful molecular graphics visualization tools with capabilities to highlight segments of user's interest. Fig. 2 shows one such monomeric structure in the quadromeric influenza A neuraminidase protein viewed through the Cn3D software.

Thus, numerous tools are available in the new bioinformatics era, and more tools and servers are constantly being added. It is up to the user to make the best use of the available data, information and analytics to derive the end results desired.

The New Age Vaccines

The availability of bioinformatics tools and advances made in genetics, immunogenomics, recombinant DNA technology, and others have provided the impetus to design the new age peptide vaccines which have considerable advantages over traditional vaccines:

- The products can be manufactured as synthetic peptides like any chemicals with stringent purity and quality controls;
- Production of peptides is simple, fast, and cost-effective;
- Problems associated with biological contamination of the antigens are removed;
- The products are typically stable under normal storage conditions;
- The peptides can be primed to avoid allergic or autoimmune responses, ideally, and in principle, can be designed for individual needs.

However, there are problems as well. Peptides are very poor immunogens on their own and require adjuvants (immune stimulants) to be effective. They are also very susceptible to enzymatic degradation. The choice of an epitope is a crucial step in the design of peptide vaccines. These epitopes should be able to induce strong, long-lasting humoral and cellular immunity against the pathogen, but sometimes are not the immunodominant types desired. Ultimately, the chosen epitope needs to be highly conserved to cover a variety of viral strains or be a mixture of several epitopes for the required vaccine.

Overall, however, production of peptide vaccines leads to improvements in time and cost. The initial screening for possible vaccine targets is now more focused and lab procedures and field trials can start within a few weeks of the start of a project. From the in silico processes of identification of suitable epitopes and design of synthetic peptide vaccines to in vitro tests and in vivo field trials, the entire process is much more rapid and efficient. In fact, in the case of the SARS (Severe Acute Respiratory Syndrome) outbreak in Asia in 2003, the initial indications were available in late 2002–early 2003, the genomic sequences were available by May 2003, and the first results of an experimental vaccine by December 2003, an extremely fast response indeed.

Interest in peptide vaccines has grown over the years, especially with indications of good results against viruses that cause cancers. The number of peptide vaccines listed in ClinicalTrials.gov website of the NIH, United States lists at present 591 projects under various phases of field trials. Of these, 466 relate to anticancer peptide vaccine trials, 92 are peptide vaccines against HIV, 16 are for influenza viruses, the trials being conducted by institutions, academia, and commercial firms. The majority are at the Phase I and Phase II trials, but 16 projects are listed in Phase III, including a few on enterovirus and Alzheimer's disease but with the majority related to various cancers.

Some of the projects to design peptide vaccines are promising, and indeed the first successful peptide vaccine was against the virulent canine parovirus as mentioned earlier and there have been such vaccines developed against other viral diseases in animals, but to date no such vaccine has been licensed for humans. However, significant contributions have been made to rationally design peptide vaccines against influenza, rotavirus, coronavirus, and certain other viruses, of which the hypervariable flu virus has presented the greatest challenges. These efforts have been taken a step further to design a universal flu vaccine, a long cherished goal, that ideally could provide permanent protection against the flu virus with a single shot or perhaps with booster shots every 5–10 years. The targets here are generally the viral hemagglutinin protein's conserved segments which are common to most flu strains so as to provoke effective antibody response against the flu virus. The estimates of almost a quarter to half a million lives being lost annually to flu according to World Health Organization are indicative of the importance of such a vaccine.

The hepatitis C virus is another virus with a high genomic variability where traditional vaccine methods have not proved very successful. Now efforts are under way to develop synthetic peptide vaccines based on CTL T-epitopes determined on the viral core protein which have been seen to evoke good immune response in mice models. A peptide vaccine against multiple sclerosis has entered Phase 3 trial stage, epitopes of human papillomavirus protein E5 as peptide vaccine candidates provided strong cell-mediated immunity (See video abstract, available under the Relevant Websites section for epitopes identified in human papillomavirus L1 protein; **Fig. 3** provides a snapshot of one peptide identified in the HPV35 L1 protein pentameric structure identified by the type of bioinformatics analysis outlined here.). A combination chemotherapy together with peptide vaccine therapy is also in Phase 3 trials to determine how effective that may be in treating patients with locally advanced or metastatic pancreatic cancer, and the ClinicalTrials website lists a number of other peptide vaccine candidates in various stages of clinical trials.

Bioinformatics in Vaccine Design: Problems and Prospects

Bioinformatics are giving us a valuable service in identifying potential epitope regions in a pathogen with safeguards built in such as no autoimmune threats, most effective immune response capability and best population coverage. Traditional methods of vaccine design will take years to achieve parallel results and even then are unlikely to test for as wide a population coverage as bioinformatics studies can do.

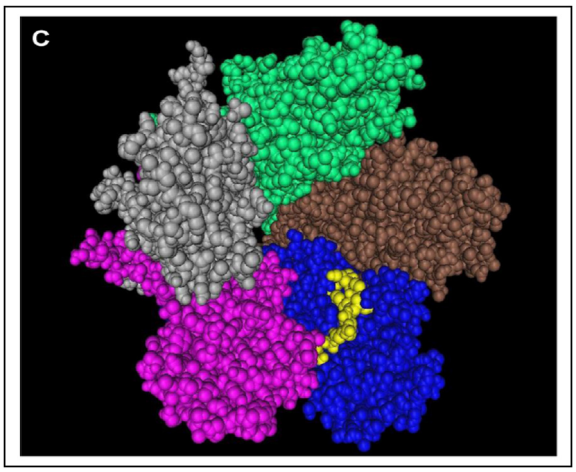

Fig. 3 The 3D space filling model of HPV35 L1 surface protein pentameric structure showing one of the peptides (in *yellow* on the monomer protein in *blue*) identified as surface accessible, conserved segment with high epitope potential.

So much, however, is yet only a part of the total effort required to take a vaccine from the computer to the marketplace. The theoretically derived peptides have to be next synthesized and tested in the laboratory against mouse models to ensure the predicted immune responses do actually occur. To administer the synthetic peptides, one has to determine the best carrier proteins and take care that the peptides do not fold up in vivo and destroy the antibody–antigen binding possibility. To enhance the effectiveness of the administered vaccine, suitable adjuvants can be used. Use of adjuvants is quite common and well recognized; an adjuvant like AS04 is a part of the Cerverix HPV vaccine.

An added advantage of using peptide vaccines over traditional whole protein vaccines is combining different peptides into one vaccine, called multivalent vaccines, to enable as wide a coverage as possible so that one or the other epitope will elicit adequate immune response. The principle is well-known: Gardasil, a 9-valent VLP vaccine is designed against nine types of human papillomavirus; a childhood vaccine known as triple antigen is designed against diphtheria, tetanus, and whooping cough where the toxoids of diphtheria and tetanus are used along with whole killed cells of whooping cough. These are examples from currently used traditional vaccines; the potential for multivalent vaccines will increase many fold in the peptide vaccine era.

In spite of the apparently enormous developments to date, the bioinformatics era in vaccine design can be considered to be at the beginning stages and much more needs to be done. The entire exercise of determining T-cell and B-cell epitopes depends acutely on accuracy of genomic and proteomic data and that is by no means guaranteed as yet. For instance, there are still significantly large number of gene and genomic sequences where one or more of the constituents remain ambiguous: for example, out of 22 genomes of the Zika virus, 9 have 1 or more nucleotides, and therefore the associated amino acids, as yet to be identified with certainty. Since antibodies are extremely specific, and one amino acid difference can cancel the antibody–antigen binding, errors in the database can make the entire bioinformatics exercise in vaccine design in vain. Highly reliable data is an imperative in such bioinformatics search for suitable vaccine targets.

Mathematical and statistical techniques constitute a major pillar of the bioinformatics exercise. While a number of different techniques are being used to determine efficient epitopes, getting around the effects of mutational changes requires identification of conserved sequences and segments. There are a number of alignment-free techniques to facilitate this search, but the robustness of these techniques requires detailed analysis. Combinations of these techniques can be expected to provide more precise and dependable results.

The necessity of more robust techniques that perhaps can provide the best results for vaccine targets with analyses of fewer sequences than needed with current techniques is best judged in trying to cope with viral epidemics. Such epidemics strike fairly rapidly, last over a few months to a year, and then die out either due to mutational changes or enough people acquiring immunity to act as a deterrent to further spread of the virus. Given that a fresh epidemic will arise only from a new virus or a new strain of an old virus, gathering enough genomic data and going through normal dry and wet labs for a fool-proof vaccine within this time schedule is outside the realms of possibility with current traditional or bioinformatics techniques. The approach to the SARS epidemic of 2003 was an exception; the Ebola virus epidemic of 2015 could be controlled by a new vaccine, but that was because the vaccine had already been designed and was ready for field trials, but was brought into live use in the face of exigency. In the case of a future epidemic we will need much better, rapid-result bioinformatics and wet lab techniques to contain such viral attacks. The world is not ready for effective handling of epidemics and pandemics as yet.

Supplementary data to this article can be found online at https://doi.org/10.1016/B978-0-12-801238-3.10878-5.

Further Reading

Basak, S. C., & Nandy, A. (2016). Computer-assisted approaches as decision support systems in the overall strategy of combating emerging diseases: Some comments regarding drug design, vaccinomics, and genomic surveillance of the Zika virus. *Current Computer-Aided Drug Design, 12*, 1–3.

Dey, S., De, A., & Nandy, A. (2016). Rational design of peptide vaccines against multiple types of human papillomavirus. *Cancer Informatics, 15*(S1), 1–16. https://doi.org/10.4137/CIN.S39071.

Ghosh, A., Nandy, A., & Nandy, P. (2010). Computational analysis and determination of a highly conserved surface exposed segment in H5N1 avian flu and H1N1 swine flu neuraminidase. *BMC Structural Biology, 10*, 6.

Goldsby, R. A., Kindt, T. J., Osborne, B. A., & Kuby, J. (2003). *Immunology* (4th edn.). New York: W H Freeman & Co.

Lewis R (2004) Vaccines: Victims of their own success? The Scientist, http://www.the scientist.com/?articles.view/articleNo/15802/title/Vaccines-Victims-of-Their-Own-Success (accessed 04.25.17).

Li, W., Joshi, M. D., Singhania, S., Ramsey, K. H., & Murthy, A. K. (2014). Peptide vaccines: Progress and challenges. *Vaccine, 2*, 515–536. https://doi.org/10.3390/vaccines2030515.

Maron, D.F. 2015. Fact or fiction?: Vaccines are dangerous. Sc American March. https://www.scientificamerican.com/article/fact-or-fiction-vaccines-are-dangerous/ (Accessed 24th April 2017).

Moisa, A. A., & Kolesanova, E. F. (2012). Chapter 11. Synthetic peptide vaccines. In R. Priti (Ed.), *Insight and control of infectious disease in global scenario* (pp. 201–228). Rijeka, Croatia: INTECH. Available at: http://www.intechopen.com/books/insight-and-control-of-infectious-disease-in-global-scenario/syntheticpeptide-Vaccines.

Moyer, M.W. 2010. Vaccinomics: Scientists are devising your personal vaccine. Sc American https://www.scientificamerican.com/article/vaccinomics-personal-vaccine/ (Accessed 24th April 2017).

Nandy, A., & Basak, S. C. (2016). A brief review of computer-assisted approaches to rational design of peptide vaccines. *International Journal of Molecular Sciences, 17*, 666. https://doi.org/10.3390/ijms17050666.

Nandy, A., Harle, M., & Basak, S. C. (2006). Mathematical descriptors of DNA sequences: Development and applications. *Arkivoc*, (9), 211–238.

Poland, G. A., Kennedy, R. B., & Ovsyannikova, I. G. (2011). Vaccinomics and personalized vaccinology: Is science leading us toward a new path of directed vaccine development and discovery? *PLoS Pathogens, 7*, e1002344. https://doi.org/10.1371/journal.ppat.1002344.

Poland, G. A., Whitaker, J. A., Poland, C. M., Ovsyannikova, I. G., & Kennedy, R. B. (2016). Vaccinology in the third millennium: Scientific and social challenges. *Current Opinion in Virology, 17*, 116–125. https://doi.org/10.1016/j.coviro.2016.03.003.

Purcell, A. W., McCluskey, J., & Rossjohn, J. (2007). More than one reason to rethink the use of peptides in vaccinedesign. *Nature Reviews. Drug Discovery, 6*, 404–414.

Rappuoli, R. (2001). Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine, 19*, 2688–2691.

Sarkar, T., Das, S., De, A., Nandy, P., Chattopadhyay, S., Chawla-Sarkar, M., & Nandy, A. (2015). H7N9 influenza outbreak in China 2013: In silico analyses of conserved segments of the hemagglutinin as a basis for the selection of peptide vaccine targets. *Computational Biology and Chemistry, 59*, 8–15.

Singluff, C. L. (2011). The present and future of peptide vaccines for cancer single or multiple, long or short, aloneor in combination? *Cancer Journal, 17*, 343–350.

Skwarczynski, M., & Toth, I. (2016). Peptide-based synthetic vaccines. *Chemical Science, 7*, 842.

Sobolev, B. N., Olenina, L. V., Kolesanova, E. F., Poroikov, V. V., & Archakov, A. I. (2005). Computer design of vaccines: Approaches, software tools and informational resources. *Current Computer-Aided Drug Design, 1*, 207–222.

Wong, S.-S., & Webby, R. J. (2013). Traditional and new influenza vaccines. *Clinical Microbiology Reviews, 26*, 476–492.

Relevant Websites

Video Abstract (4:36) Dey, S., De, A., & Nandy, A. (2016). Rational design of peptide vaccines against multiple types of human papillomavirus. *Cancer Informatics, 15*(S1), 1–16. - https://www.youtube.com/watch?v=kpgKyoFbUWg.