Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 218590, 26 pages
doi:10.1155/2010/218590

Review Article

Emerging Vaccine Informatics

Yongqun He,[1] Rino Rappuoli,[2] Anne S. De Groot,[3,4] and Robert T. Chen[5]

[1] *Department of Microbiology and Immunology, Unit for Laboratory Animal Medicine, Center for Computational Medicine and Bioinformatics, and Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI 48109, USA*
[2] *Novartis Vaccines and Diagnostics, 53100 Siena, Italy*
[3] *EpiVax, Inc., Providence, RI 02903, USA*
[4] *Institute for Immunology and Informatics, University of Rhode Island, Providence, RI 02903, USA*
[5] *HIV Vaccine and Special Studies Team, Centers for Disease Control and Prevention (CDC/DHAP/EB), Atlanta, GA 30333, USA*

Correspondence should be addressed to Yongqun He, yongqunh@med.umich.edu

Received 8 December 2010; Accepted 31 December 2010

Academic Editor: Rodomiro Ortiz

Vaccine informatics is an emerging research area that focuses on development and applications of bioinformatics methods that can be used to facilitate every aspect of the preclinical, clinical, and postlicensure vaccine enterprises. Many immunoinformatics algorithms and resources have been developed to predict T- and B-cell immune epitopes for epitope vaccine development and protective immunity analysis. Vaccine protein candidates are predictable *in silico* from genome sequences using reverse vaccinology. Systematic transcriptomics and proteomics gene expression analyses facilitate rational vaccine design and identification of gene responses that are correlates of protection *in vivo*. Mathematical simulations have been used to model host-pathogen interactions and improve vaccine production and vaccination protocols. Computational methods have also been used for development of immunization registries or immunization information systems, assessment of vaccine safety and efficacy, and immunization modeling. Computational literature mining and databases effectively process, mine, and store large amounts of vaccine literature and data. Vaccine Ontology (VO) has been initiated to integrate various vaccine data and support automated reasoning.

1. Introduction

While the history of vaccines is relatively short, vaccines have contributed to dramatic improvements in public health worldwide. Jenner's description of smallpox prevention in 1796 [1] is the most commonly recognized "start" of vaccine research in European historical documents, although variolation had been practiced in Asia centuries earlier. Critical advances in vaccine science took place in the late 19th and early 20th centuries, by scientists such as Pasteur, Koch, von Behring, Calmette, Guérin, and Ehrlich [2]. Discoveries by these early vaccine researchers contributed to the development of antiserums, antitoxins, and live, attenuated bacterial vaccines. The discovery of tissue culture methods for viral and bacterial propagation *in vitro* during the period from 1930 to 1950 was a technical advance that enabled the development of vaccines against many viruses including measles and polio. Further advances in cell culture techniques, carbohydrate chemistry, molecular biology, and immunology have led to the modern era of "subunit" vaccine development. The recombinant hepatitis B vaccine, one of the first subunit vaccines, was licensed in 1986 [2]. This marked the beginning of the molecular biology phase of vaccine development. At present, human vaccines are used in the prevention of more than thirty infectious diseases. Due to the success of the smallpox eradication campaign in 1960s and 1970s, the powerful impact of vaccines on human health is universally recognized [3]. In addition, there exist a large number of animal vaccines [4].

With the advent of computers and informatics, new approaches have been devised that facilitate vaccine research and development. Immunoinformatics targets the use of mathematical and computational approaches to address immunological questions. Since the 1980s, many immunoinformatics methods have been developed and used to predict T-cell and B-cell immune epitopes [5]. Indeed, many predicted T- and B-cell immune epitopes are possible epitope vaccine targets. Experimentally verified immune epitopes are

now stored in web-based databases which are freely available for further analysis [6]. Immune epitope studies are crucial to uncover basic protective immune mechanisms.

A new era of vaccine research began in 1995, when the complete genome of *Haemophilus influenzae* (a pathogenic bacterium) was published [7]. In parallel with advances in molecular biology and sequencing technology, bioinformatics analysis of microbial genome data has allowed *in silico* selection of vaccine targets. Further advances in the field of immunoinformatics have led to the development of hundreds of new vaccine design algorithms. This novel approach for developing vaccines has been named reverse vaccinology [8] or immunome-derived vaccine design [9]. Reverse vaccinology was first applied to the development of vaccines against serogroup B *Neisseria meningitides* (MenB) [10]. With the availability of multiple genomes sequenced for pathogens, it is now possible to run comparative genomics analyses to find vaccine targets shared by many pathogenic organisms.

In the postgenomics era, high throughput-omics technologies-genomics, transcriptomics, proteomics, and large-scale immunology assays enable the testing and screening of millions of possible vaccine targets in real time. Bioinformatics approaches play a critical role in analyzing large amounts of high throughput data at differing levels, ranging from data normalization, significant gene expression detection, function enrichment, to pathway analysis.

Mathematical simulation methods have also been developed to model various vaccine-associated areas, ranging from analysis of host-pathogen interactions and host-vaccine interactions to cost cost-effectiveness analyses and simulation of vaccination protocols. The mathematical modeling approaches have contributed dramatically to the understanding of fundamental protective immunity and optimization of vaccination procedures and vaccine distribution.

Informatics is also changing postlicensure immunization policies and programs. Computerized immunization registries or immunization information systems (IIS) are effective approaches to track vaccination history. Bioinformatics has widely been used to improve surveillance of (1) vaccine safety using systems such as the Vaccine Adverse Event Reporting System (VAERS, http://vaers.hhs.gov/) [11] and the Vaccine Safety Datalink (VSD) [12] project and (2) vaccine effectiveness for each of the target vaccine preventable diseases via their respective public health surveillance systems. Computational methods have also been applied to model the impact of alternative immunization strategies and to detect outbreaks of vaccine preventable diseases and safety concerns related to vaccinations as well.

With the large amounts of vaccine literature and data becoming available, it is not only challenging but crucial to perform vaccine literature mining, generate well-annotated and comprehensive vaccine databases, and integrate various vaccine data to enhance vaccine research. Computational vaccine literature mining will allow us to efficiently find vaccine information. To effectively organize and analyze the huge amounts of vaccine data produced and published in the postgenomics and information era, many vaccine-related databases, such as the VIOLIN vaccine database and analysis system (http://www.violinet.org/) [13] and AIDS vaccine trials database (http://www.iavireport.org/trials-db/), have been developed and are available on the web. However, relational databases are not ideal for data sharing since different databases may use different schemas and formats. A biomedical ontology is a consensus-based controlled vocabulary of terms and relations, with associated definitions that are logically formulated in such a way as to promote automated reasoning. Ontologies are able to structure complex biomedical domains and relate the myriads of data accumulated in such a fashion as to permit shared understanding of vaccines among different resources. The Vaccine Ontology (VO; http://www.violinet.org/vaccineontology/) is a novel open-access ontology in the domain of vaccine [14]. Recent studies show that VO can be used to support vaccine data integration and improve vaccine literature mining [15, 16].

In summary, vaccine informatics is an emerging field of research that focuses on the development and applications of computational approaches to advance vaccine research and development (R&D) and improve immunization programs. Vaccine informatics plays an important role in every aspect of pre- and postlicensure vaccine enterprises (Figure 1). This paper summarizes the history of vaccine informatics developments in advancing vaccine research and development and immunization programs.

2. Immunoinformatics and Vaccine Design

This section describes immunoinformatics and how it is used for vaccine design and to study protective immune responses to vaccines.

2.1. Brief History of Immunoinformatics Approaches for Vaccine Design. The first immunoinformatics tools for vaccine design were developed in the 1980s by DeLisi and Berzofsky and others [17]. Chief among vaccine design informatics tools are epitope-mapping algorithms. Since the T-cell epitopes are bound in a linear form to the human leukocyte antigen (HLA), the interface between ligands and T-cells can now be modeled with accuracy. A large number of T-cell epitope-mapping algorithms have consequently been developed [18, 19]. These tools now make it possible to start with the entire proteome of a pathogen and rapidly identify putative T-cell epitopes. Such information is immensely valuable for the development of new vaccines, diagnostic purposes, and for studying the pathology of infectious diseases [5, 20–27].

Several different routes for vaccine development have been pursued. One method, which has been used by De Groot and Martin [24, 28], is to synthesize the putative T-cell epitopes and screen peripheral blood mononuclear cells (PBMC) isolated from human subjects infected with the target pathogen (or have a target cancer) for immune response to the epitopes. A T-cell *in vitro* response to a specific peptide epitope (typically measured by ELISA or ELISpot assay) served as an indicator that the protein from which the peptide was derived was expressed, processed, and presented to the immune system in the course of a "natural" immune response. This approach, often considered a means of

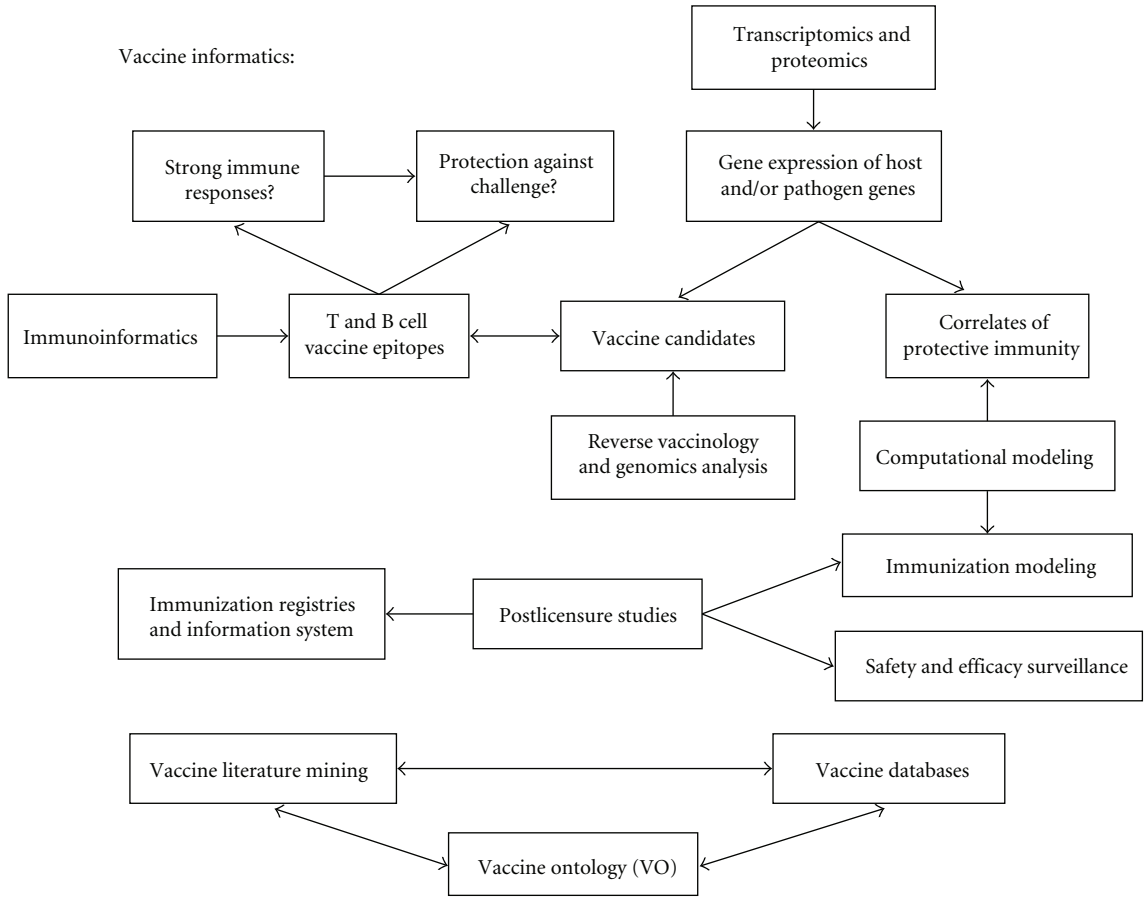

Figure 1: Overview of current vaccine informatics topics.

making epitope-based vaccines (see below), can also be used to identify proteins for use in vaccine development. This approach, described by De Groot and Martin's group as "fishing for antigens using epitopes as bait", has been used to discover new vaccine antigens for *F. tularensis* (a bioterror agent) [28], tuberculosis [24], smallpox [29], and *H. pylori* [30].

The proteome of *M. tuberculosis* (*Mtb*), the etiologic agent of TB, contains almost 4,000 proteins. Evaluating each one using the straightforward but expensive and laborious approach of synthesizing and testing overlapping peptides could take decades. Using epitope mapping tools, it is now possible to screen a whole proteome *in silico*, followed by a finer focus on the resulting sets of peptides [5].

The ability to accurately predict T-cell epitopes from raw genomic data is fundamental to the development of novel vaccines, and serves as the starting point for a number of research projects. Freeing the researchers from the constraints of predetermined sets of "virulence genes" has resulted in some remarkable discoveries. McMurry and De Groot [24] found extraordinary diversity of human immune responses to proteins in the *Mtb* genome that have yet to be ascribed a function, suggesting that human immune response is omnivorous and is not focused on recognition of a single "immunodominant" protein. In addition, these investigators have found a remarkable similarity between *Francisella tularensis* (the etiologic agent for Tularemia) and (human) self, at the epitope level [28]. Thus informatics, starting at the genome, may reveal potential antigenic relationships between human proteins and pathogens, or even commensal organisms, which might predetermine individual immune response, that is, prior exposure to a given pathogen may tune immune response to a second pathogen [31].

An alternative approach, "Reverse Vaccinology," a term coined by Rappuoli, starts with predicting putative vaccine candidates by *in silico* genomics analysis based on yet different criteria. The predicted vaccine candidates (e.g., bacterial surface proteins) are thought to stimulate protective immunity. Candidate proteins can be evaluated experimentally by demonstrating an immune response that correlates with *in vivo* protection [10]. The Reserve Vaccinology approach is well discussed later on (see below).

2.2. MHC Polymorphism, Epitope Variations, and Vaccine Design. The success rate of vaccine development decreases with the increasing variability of the surface antigens of pathogens and the decreasing ability of antibodies to confer protective immunity [32]. Fortunately, vaccine informatics

tools are being developed that increase the accuracy of vaccine target prediction for variable pathogens and help vaccinologists triage antigens.

T-cells are activated by direct interaction with antigen presenting cells (APCs). On the molecular level, the initial interaction occurs between the T-cell receptor and peptides derived from endogenous and exogenous proteins that are bound in the cleft of MHC class I or class II molecules. In general, MHC class I molecules present peptides 8–10 amino acids in length and are predominantly recognized by $CD8^+$ cytotoxic T lymphocytes (CTLs). Class I peptides usually contain an MHC I-allele-specific motif composed of two conserved anchor residues [33–35]. Peptides presented by class II molecules are longer, more variable in size, and have more complex anchor motifs than those presented by class I molecules [36–38]. MHC class II molecules bind peptides consisting of 11–25 amino acids and are recognized by $CD4^+$ T helper (Th) cells.

MHC class I molecules present peptides obtained from proteolytic digestion of endogenously synthesized proteins. Host- or pathogen-derived intracellular proteins are cleaved by a complex of proteases in the proteasome. Small peptide fragments are then typically transported by ATP-dependent transporters associated with antigen processing (TAPs) and also by TAP-independent means into the endoplasmic reticulum (ER), where they form complexes with nascent MHC class I heavy chains and beta-2-microglobulin. The peptide-MHC class I complexes are transported to the cell surface for presentation to the receptors of $CD8^+$ T-cells [39–41].

MHC class II molecules generally bind peptides derived from the cell membrane or from extracellular proteins that have been internalized by APCs. The proteins are initially processed in the MHC class II compartment (MIIC). Inside the MIIC, MHC is initially bound to class II-associated invariant chain peptide (CLIP) which protects the MHC from binding to endogenous peptides. Peptides generated by proteolytic processing within endosomes replace CLIP in a reaction catalyzed by the protein HLA-DM [42, 43]. The class II molecules bound to peptide fragments are transported to the surface of APCs for presentation

To complicate matters further, HLA molecules bind different peptides due to the configuration of their HLA binding pockets. This is the source of genetic diversity of immune responses [34]. Fortunately, there is some conservation between HLA pockets, and both DeLisi and Sette have addressed the issue of HLA coverage for epitope predictions by demonstrating that epitope-based vaccines containing epitopes restricted by selected "supertype" Class I and Class II HLA can provide the broadest possible coverage of the human population [44, 45]. De Groot and Martin have constructed an algorithm, Aggregatrix, which uses the "set cover" method to identify the best set of peptides from a pathogen that would yield the broadest coverage of HLA if included in a vaccine. The Aggregatrix algorithm selects optimized epitope sets which, in terms of immunogenicity and genetic conservation, collectively "cover" a wide variety of known circulating strain variants of a given pathogen and a majority of the common human HLA types [46]. The Conservatrix algorithm is used to identify highly conserved peptide segments contained within multiple isolates of variable pathogens such as retroviruses [47]. The amino acid sequences of protein isolates are parsed into 9 mer frames overlapping by eight amino acids. The resulting peptide set yields a list of unique segments and appearance frequencies. Highly conserved sequences are thought to be important in the evolutionary "fitness" of pathogens and thus are unlikely to change in an attempt to evade the immune system. Conserved sequences can be analyzed using epitope prediction software.

2.3. T-Cell Epitope Mapping. Although textbooks teach that protective immune response is attributed to the development of protective antibodies, the immune response to attenuated intact viruses and subunit vaccines is to a very large degree dependent on T-cell recognition of peptide epitopes bound to MHC. Thus targeting antigens that contain many $CD4^+$ T helper epitopes may lead to the selection of good B-cell antigens as well as immunogens for effective CD8 responses—this is because $CD4^+$ T helper cells are critically important to the development of memory B-cell (antibody) and memory CTL (cytotoxic T-cell) responses, in addition to being active against pathogens on their own. T helper cells have been called the "conductors of the immune system orchestra" [20]. CTLs generally play a role in the containment of viral and bacterial infection [48], and the prevalence of CTLs usually correlates with the rate of pathogen clearance. Regulatory T-cells are also represented among $CD4^+$ T-cells, although some $CD8^+$ Tregs have been described.

T-cell epitope algorithms now achieve a high degree of prediction accuracy (in the range of 90 to 95% Positive Predictive Value). For example, epitope mapping tools can now be compared to other available tools, using the Immune Epitope Database "gold standard" as described by Wang et al. [49]. A list of epitope mapping tools, ancillary algorithms, and their comparative features is provided in Table 1. A number of the epitope mapping tools are available to researchers via the web. These include the tool available at the SYFPEITHI website [50] and an HLA binding prediction tool available on at the National Institutes of Health (BIMAS) [51]. A recently developed set of tools has now been made available through the Immunome Epitope Database. Each of these tools has been described and validated [49]. One such proprietary algorithm, EpiMatrix, is in active use in the pharmaceutical industry [52]. While none of these sites yield exactly the same predictions, all predictions are quite accurate, especially when compared to results obtained with early epitope mapping tools (e.g., SYFPEITHI and BIMAS) [49, 52]. In general, the newer and more actively maintained algorithms tend to outperform the older more static predictive methods.

With many machine learning techniques developed since early 1990s for T-cell epitope predictions, it is possible to comparatively evaluate them through prediction performance assessments [49, 53–55]. Lin et al. compared 30 servers developed by 19 groups that can predict HLA-I binding peptides [53]. Their benchmarking study showed that predictions of six out of seven of HLA-I binding peptides achieved excellent classification accuracy. In general, nonlinear predictors outperform matrix-based predictors,

Table 1: List of online tools for T-cell epitope prediction.

Tool	Website	Type	Class	Refs
ANNPRED	http://www.imtech.res.in/raghava/nhlapred/neural.html	ANN	I	[234]
Bimas	http://www-bimas.cit.nih.gov/molbio/hla_bind/	QM	I	[51]
EpiJen	http://www.ddg-pharmfac.net/epijen/EpiJen/EpiJen.htm	Multi-step algorithm	I	[235]
EPIMHC	http://imed.med.ucm.es/epimhc/	user made Profiles	I, II	[236]
EpiMatrix	http://www.epivax.com/	Matrix-based and pocket profile	I, II	[237]
HLABIND	http://atom.research.microsoft.com/hlabinding/hlabinding.aspx	Adaptive Double Threading	I	[238]
IEDB	http://tools.immuneepitope.org/analyze/html/mhc_binding.html	ARB-QM, SMM-QM, ANN-regression	I	[239]
KISS	http://cbio.ensmp.fr/kiss/	SVM	I	[240]
MHC2PRED	http://www.imtech.res.in/raghava/mhc2pred/	SVM	II	[241]
MHC Pred	http://www.jenner.ac.uk/MHCPred	Partial least-squares-based multivariate statistical method	I, II	[242]
MULTIPRED	http://antigen.i2r.a-star.edu.sg/multipred/	ANN, pHMM	I, II	[243]
MOTIF_SCAN	http://www.hiv.lanl.gov/content/immunology/motif_scan/motif_scan	Sequence Motifs	I, II	—
NetCTL	http://www.cbs.dtu.dk/services/NetCTL/	Multi-step algorithm	I, CTL	[60]
NetCTLSpan	http://www.cbs.dtu.dk/services/NetCTLpan/	Multi-step algorithm	I, CTL	[62]
NetMHC	http://www.cbs.dtu.dk/services/NetMHC/	ANN	I	[244]
netMHCII	http://www.cbs.dtu.dk/services/NetMHCII/	SMM-QM	II	[245]
netMHCpan	http://www.cbs.dtu.dk/services/NetMHCpan/	ANN-regression	I	[246]
netMHCIIpan	http://www.cbs.dtu.dk/services/NetMHCIIpan/	ANN-regression	II	[247]
PEPVAC	http://imed.med.ucm.es/PEPVAC/	Profiles or PSSM	I	[248]
PREDEP	http://margalit.huji.ac.il/Teppred/mhc-bind/index.html	Threading	I	[249]
POPI	http://iclab.life.nctu.edu.tw/POPI/	SVM	I, II	[250]
PROPREDI	http://www.imtech.res.in/raghava/propred1/	QM	I	[251]
PROPRED	http://www.imtech.res.in/raghava/propred/	QM	II	[252]
RANKPEP	http://imed.med.ucm.es/Tools/rankpep.html	Profiles or PSSM	I, II	[253]
SVMHC	http://abi.inf.uni-tuebingen.de/SVMHC	SVM	I, II	[254]
SVRMHC	http://svrmhc.biolead.org/	SVM-regression	I, II	[255]
SYFPEITHI	http://www.syfpeithi.de/	Motif matrices	I, II	[50]
TEPITOPE	http://www.vaccinome.com/	QM	II	[256]
Vaxign	http://www.violinet.org/vaxign/	PSSM	I, II	[110]

Abbreviations: ANN: artificial neural networks; PSSM: position-specific scoring matrix; QM: quantitative matrices; SMM: stabilized matrix method; SVM: support vector machine; multistep algorithm: integrating predictions of proteasomal cleavage, TAP transport efficiency, and MHC class I affinity.

and most predictors can be improved by non-linear transformations of their raw prediction scores [53]. While good performance has been achieved for MHC class I predictions, there is still limited success for prediction of epitopes for HLA class II [54, 55]. The low prediction accuracy of HLA-II binding peptides is due to several factors: (a) insufficient or low-quality training data, (b) difficulty in identifying 9-mer binding cores within longer peptides used for training and lack of consideration of the influence of flanking residues, and (c) relative permissiveness of the binding groove of HLA-II molecules for peptide binding, which limits the stringency of binding [54].

Adequate predictors are lacking for predicting epitopes for HLA-C, HLA-DQ, and HLA-DP. However, Wang et al. have made a significant effort in peptide binding predictions for HLA DR, DP, and DQ molecules [56]. Their research with a large-scale datasets of over 17,000 HLA-peptide binding affinities for 11 HLA DP and DQ alleles found that prediction methodologies developed for HLA DR molecules perform equally well for DP and DQ molecules.

The generation of an MHC class-I epitope starts with the degradation of endogenous proteins into oligomeric fragments by cytosolic proteases, mainly the proteasome. These oligomeric fragments may escape from the attack of amino peptidases by entering the endoplasmic reticulum (ER) by the transporter associated with antigen presentation (TAP) [57]. The prediction algorithms for TAP binding and proteasomal cleavage have been developed [58, 59]. For

example, Peters et al. used a stabilized matrix method to predict TAP affinity of peptides [58]. This scoring method took advantage of the fact that binding of peptides to TAP is mainly determined by the C terminus and three N-terminal residues of a peptide. Predictions of the MHC class I pathway can be improved by predictions of proteasomal cleavage, TAP transport efficiency, and MHC class I binding affinity [58, 60–62].

While many successes have been made in the area of T-cell epitope prediction, the limitations of all these predictors should be noted. Our goal is to identify good vaccine targets that will induce productive immune responses. However, our ability to measure is usually done indirectly: peptide-binding assays, induction and measurement of immune responses *ex vivo*, use of animal models, and so forth. Only a small number of HLA-binding peptides are good targets. In clinical vaccine trials, wrong peptides were often selected using indirect methods and tested [63]. For example, the virulence and tumor maintenance capacity of high-risk Human Papillomavirus 16 (HPV-16) is mediated by two viral oncoproteins, E6 and E7. Of 21 E6 and E7 peptides computed to bind HLA-A*0201, 10 were confirmed through TAP-deficient T2 cell HLA stabilization assay. By testing their physical presence among peptides eluted from HPV-16-transformed epithelial tumor HLA-A*0201 immunoprecipitates, only one epitope (E7(11–19)) highly conserved among HPV-16 strains was detected. This 9-mer serves to direct cytolysis by T-cell lines. However, a related 10-mer (E7(11–20)), previously used as a vaccine candidate, was not detected by immune-precipitation or cytolysis assays. These data underscore the importance of precisely defining CTL epitopes on tumor cells and offer a paradigm for T-cell-based vaccine design [63].

2.4. B-Cell Epitope Mapping. It is important to clarify that limited immunoinformatics tools are currently available to identify B-cell antigens (recognized by antibodies). While humoral, or antibody-based, response represents the first line of defense against most viral and bacterial pathogens, the protein target of this arm of defense is usually too complex to model *in silico*. Antibodies that recognize B-cell epitopes, composed of either linear peptide sequences or conformational determinants, are present only in the three-dimensional form of the antigens. Several B-cell epitope prediction tools, including 3DEX, CEP, and Pepito, are at various stages in development and are in the process of being refined [64–67]. IEDB has collected a list of web prediction tools for B-cell epitope prediction (http://toolsl.immuneepitope.org/main/html/bcell_tools.htm). Unfortunately, the computational resources and modeling complexity required to predict B-cell epitopes are enormous. This complexity is due, in part, to the inherent flexibility in the complementarity-determining regions (CDR) of the antibody and, in part, attributable to posttranslational modifications such as glycosylation, all of which can result in modification of B-cell epitopes.

B-cell epitopes include linear and discontinuous epitopes. Linear epitopes comprise a single continuous stretch of amino acids within a protein sequence. An epitope whose residues are distantly separated in the sequence but have physical proximity through protein folding is named a discontinuous epitope. Although most epitopes are discontinuous [68], experimental epitope detection is primarily for linear epitopes. Tools for prediction of linear B-cell epitopes exist but in general are not predictive [69, 70]. The benchmarking B-cell epitope prediction by Blythe and Flower [69] found that with the best set of scales and parameters, amino acid propensity profiles can predict linear B-cell epitopes only marginally better than random. Such a conclusion has been confirmed by another study where the dismal performance of five predictors was tested against a set of reported linear B-cell epitopes [70].

Although devising accurate B-cell epitope mapping tools remains difficult, the selection of potent B-cell antigens can be accelerated using T-cell epitope mapping tools. When considering B-cell antigens as potential subunit vaccines, it also may be important to also consider their T-cell epitope content since the quality and kinetics of the antibody response is dependent upon the presence of T help. B-cell antigens that contain significant T help may outperform B-cell antigens lacking cognate help. In some cases, an identified T-cell epitope may also contain a B-cell epitope. Different epitopes activate T and B-cells. Despite this observation, it has been widely reported that B-cell epitopes may colocalize near, or overlap, Class II (Th, $CD4^+$) epitopes [71, 72].

2.5. Immunoinformatics-Based Vaccine Design Strategies. Different epitope-based vaccine design strategies exist, for example, mosaic vaccines [73], consensus [74, 75], centralized or ancestor immunogen [76, 77], or COT^+ [78]. Mosaic vaccines are comprised of "mosaic" proteins that are assembled from fragments of natural sequences via a computational optimization method [73]. Many immunogens, such as HIV envelope proteins, have high amino acid sequence divergences. To minimize the genetic differences between vaccine strains and contemporary isolates, immunogenic consensus sequences can be detected and used in vaccine design [74, 75]. Computer programs can also be developed to generate "centralized" vaccine that consists of consensus, ancestor, or center of the tree, modeled from phylogenetic trees. These "centralized" sequences can decrease the genetic distances between the "centralized" and wild-type gene immunogens [76, 77]. In an effort to develop antigens that capture both consensus and mutation sequences among strains, Nickle et al. reconstructed COT^+ antigens by including the ancestral state sequence at the center of phylogenic tree (COT) and extending the COT immunogen through addition of a composite sequence that includes high-frequency variable sites preserved in their native contexts [78]. These epitope-based vaccine designs have proven effective and provided vaccine researchers with different options in rational vaccine design. It is promising to combine various epitope methods to improve target discovery [56, 60, 79].

Integrated systems and workflows for computational vaccinology are likely to be key for automation of vaccine target discovery [80–82]. For example, Sollner et al. introduced the

pBone/pView computational workflow that supports design and execution of immunoinformatics workflow modules, results visualization, and knowledge sharing and reuse [80]. Pappalardo et al. developed ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design, and optimization [81]. Feldhahn et al. developed FRED, an extendable, open source software framework for T-cell eptiope detection that integrates many prediction methods and supports implementation of custom-tailored prediction pipelines [82]. The effectiveness of these systems has been demonstrated with different applications.

The EpiVax vaccine design tools (EpiMatrix, ClustiMer, VaccineCAD, EpiAssembler, BlastiMer) are available to researchers through a portal at the Institute for Immunology and Informatics (the iVAX toolkit) [9]. The team of De Groot, Moise, and Martin have implemented the iVAX toolkit to develop four vaccines, a multiepitope TB vaccine [24], a cross-clade HIV vaccine [74], a prototype *H. pylori* vaccine [83], and a tularemia vaccine [84]. In collaboration with the TRIAD (Translational Immunology Research and Accelerated [vaccine] Development) program at the University of Rhode Island, iVAX is now being used to design additional vaccines including a multipathogen biodefense vaccine against Tularemia and *Burkholderia* spp, an epitope-based vaccine for HCV, and a vaccine derived from the deer tick saliva to prevent the acquisition of Tick-borne-pathogens. In addition, iVAX has recently been used to scan the entire genome of *Salmonella typhi* for vaccine candidates. This program is also accessible to researchers working on Neglected Tropical Diseases through the immunome website http://immunome.org/.

3. Reverse Vaccinology

3.1. Basic Principles of In Silico Antigen Prediction. Initially, when Reverse Vaccinology (RV) was developed, prediction of putative vaccine candidates was based solely on *in silico* analysis of the genome of a single strain. Now that selection criteria have been implemented, however, *in silico* analysis remains the central step in an RV project (see Figure 2).

The first step in the process of genome interpretation, usually referred as gene finding, consists in the prediction and localization of genes onto the chromosome. This is accomplished using prediction programs, which scan the sequence in search of regions that are likely to encode proteins. In prokaryotic systems, the identification of potential coding regions or open reading frames requires implementation of a few basic rules. In the simplest formulation, open reading frames (ORFs) are identified as segments of the same frame comprised between one of the three standard start codons (ATG, TTG, GTG) and one of the three standard stop codons (TAA, TAG, TGA). It is generally accepted that there is approximately one gene for every 1000 DNA base pairs. This suggests that significantly long start-to-stop segments are likely to encode for proteins.

Genome annotation procedures can be automated to different extents. Automated methods for prokaryotic gene finding such as GLIMMER [85], ORPHEUS [86], and GeneMark [87] have been used in genome sequencing projects [88–91]. GLIMMER uses interpolated Markov models, GeneMark uses hidden Markov models, and ORPHEUS is mainly based on codon usage and ribosome binding site statistics derived from annotated genes.

An exhaustive summary of software tools and websites that can be used to obtain bacterial genome annotations was presented by Stothard and Wishart [92].

The annotation procedure allows the translation of the bacterial genome sequence into a list of all the proteins that a bacterium virtually expresses at any time in its life cycle. Each of these amino acid sequences is then compared to the content of public databases of proteins or DNA sequences in an attempt to identify related sequences. When there exist obvious sequence similarities, it is reasonable to transfer this information on the filed sequence to the query. The functional annotation of a protein is sometimes sufficient for the selection of the protein as vaccine candidate, especially when the prediction of protein subcellular localization is uncertain, for example, a protein annotated as fibronectin binding protein may be a good vaccine candidate even when localization algorithms classify it as cytoplasmic. A critical aspect is represented by sequences that lack homologues or contain only remote homologues filed in the databases. ORFs having 20% or less of amino acid identity to any amino acid sequence found in the databases are generally considered to have unreliable homologues. These could represent novel uncharacterized proteins or random open reading frames misidentified as genes. Although homology searches can identify to a limited extent ORFs that are likely to encode functional proteins, experimental authentication by proteomic techniques is usually a more powerful approach for distinguishing genes from random ORFs.

The ensemble of hypothetical proteins can be processed with software programs dedicated to deduce their possible cellular localization. One of the basic assumptions utilized for candidate searches is that a good antigen will be located on the cell surface of a bacterium, where it is readily available for antibody recognition. Several algorithms have been developed that predict the subcellular localization of proteins based solely on the amino acid sequence and composition (see Table 2). The basic assumption made is that the N-terminal sequence of the protein predicts its cellular destination. The presence of a "leader sequence" provides evidence that the proteins will be exported to extra-cytoplasmic compartments. Additional signatures may also be exploited such as the presence of a cleavage site immediately after the leader peptide. Such sites imply that the protein is released into the extra-cellular environment of Gram-positive bacteria or into the periplasmic space of Gram-negatives. Similarly, proteins that contain an LXXC motif, where X is any amino acid, positioned at the end of the leader peptide are often lipoproteins. Anchoring of proteins to the Gram-positive bacterium cell wall often requires a specific carboxy-terminal sorting sequence. This sequence is identified by an LPXTG motif followed by approximately 20 hydrophobic amino acids and a charged tail. In Gram-negative bacteria, additional secretion pathways exist that promote the passage of extracellular proteins across the outer

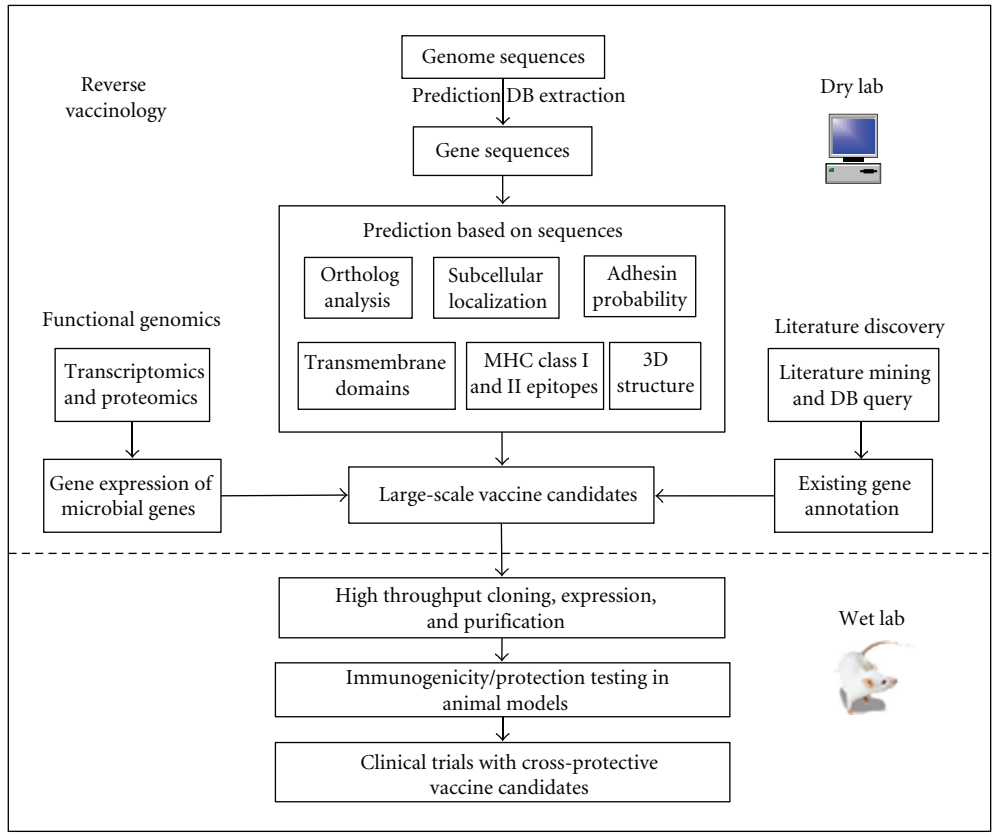

Figure 2: A schematic demonstration of integrative reverse vaccinology strategy towards vaccine development.

membrane. At least six distinct extracellular protein secretion systems have been reported in Gram-negative and Gram-positive bacterium (type I–VI, T1SS–T6SS) that can deliver proteins through the multilayered bacterial cell membrane and in some instances pass directly into the target host cell [93]. The six secretion systems exist in Gram-negative bacteria and the common Gram positive bacteria. Gram positive bacteria contain an additional specific secretion system (type VII) [94]. This increases the variety and complexity of secretion signals, making the identification of outer membrane and secreted proteins yet more challenging [95, 96].

Several computational methods have been generated to predict extracellular proteins in Gram-negative microorganisms [97].

PSORTb is the most widely used tool for predicting subcellular multiple localizations of organelles in Gram-negative bacteria. This program uses biological knowledge to elaborate "if-then" rules, combining information on amino acid composition, similarity to proteins of known subcellular localization, presence of signal peptides, transmembrane helices, and motifs diagnostics of specific subcellular localization. Recently, two predictive methods CELLO [98] and Proteome Analyst [99] have been proposed for Gram-negative bacteria. These programs are providing comparable performances in terms of accuracy and recall with respect to PSORTb [97].

Despite the recent progress, identification of secretion systems components *in silico* and their effectors still mainly relies on the detection of amino acid sequence [94] and the structural [100] similarities of selected proteins. Caution is necessary in applying these predictions, as sequence similarities can be very weak and do not necessarily imply any functional analogy.

In conclusion, by knowing the genome sequence it becomes possible to select using bioinformatics tools to generate a list of potential antigens without cultivating the microorganism. This methodology has a huge advantage over conventional vaccinology approaches for two major reasons. First of all, *in silico* analysis is very fast and cheap, and secondly, proteins not expressed *in vitro* are also identified. However, this approach only provides a prediction of a protein's subcellular localization and it cannot reveal if a protein is expressed and under what conditions. Therefore, use of a bioinformatics approach may need to be complemented with other techniques, for example, a Mass Spectrometry-based approach to aid vaccine candidate prediction. The first RV project employed a single genome. Indeed, at that time there was only one genome available for *N. meningitidis*. Nowadays in most cases, there are more than five genomes available for

Table 2: Tools used for reverse vaccinology.

Tool name	Website URL	Comment	Refs
ORF prediction and genome annotation			
GLIMMER	http://www.cbcb.umd.edu/software/glimmer/	Interpolated Markov models	[85]
ORPHEUS	http://pedant.gsf.de/orpheus/ (not working now)	Codon usage and ribosome binding statistics	[86]
GeneMark	http://exon.gatech.edu/GeneMark/	HMM	[87]
Bacterial protein localization prediction			
PSORTb	http://www.psort.org/psortb/	Multicomponent	[97]
Proteome Analyst	http://webdocs.cs.ualberta.ca/~bioinfo/PA/Sub/	Annotation keywords	[99]
SubLoc	http://www.bioinfo.tsinghua.edu.cn/SubLoc/	SVM	[257]
CELLO	http://cello.life.nctu.edu.tw/	SVM	[98]
PSLpred	http://www.imtech.res.in/raghava/pslpred/	SVM	[258]
LOCtree	http://cubic.bioc.columbia.edu/cgi/var/nair/loctree/query	SVM	[259]
SignalP	http://www.cbs.dtu.dk/services/SignalP/	NN, HMM	[260]
Sequence conservation			
BLAST	http://blast.ncbi.nlm.nih.gov/	Best reciprocal BLAST hit	[261]
OrthoMCL	http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi	SVM	[262]
Transmembrane domain prediction			
HMMTOP	http://www.enzim.hu/hmmtop/	HMM	[263]
PRED-TMBB	http://biophysics.biol.uoa.gr/PRED-TMBB/	HMM (β-barrel)	[264]
TBBpred	http://www.imtech.res.in/raghava/tbbpred/	NN, SVM (β-barrel)	[265]
PROFtmb	http://cubic.bioc.columbia.edu/services/proftmb/	HMM	[266]
Bacterial adhesin prediction			
SPAAN	ftp://203.195.151.45	NN	[112]
Reverse vaccinology software			
NERVE	http://www.bio.unipd.it/molbinfo/ReverseVaccinology-NERVE.html	Multicomponent	[113]
Vaxign	http://www.violinet.org/vaxign/	Web-based, multicomponent	[110]

Abbreviations: HMM: hidden Markov model; NN: neuron network; SVM: support vector machine.

any human pathogen. Therefore, the *in silico* analysis can take advantage of comparative genomics.

3.2. Comparative Genomics and the Pangenome Concept. Today, the number of fully sequenced microbial genomes exceeds 1000 (http://www.ncbi.nlm.nih.gov/bioproject/) (Many are not from pathogens). It is clear that microbial diversity has been vastly underestimated, and a single genome does not exhaust the genomic diversity of any bacterial species [101, 102]. In many cases, an extensive genomic plasticity exists. For example, completion of the genome sequence of *E. coli* O157:H7 revealed that it contains >1,300 strain-specific genes compared to *E. coli* K12, which encode proteins that are involved in virulence and metabolic capabilities [103, 104]. Additional reports have revealed the occurrence of an extensive amount of genomic diversity among the strains of a single species [105–107].

These early findings were formalized with the definition of the bacterial pangenome, as the sum of the genes present in each individual species. This concept was originally introduced during study of the genome variability in eight isolates of *Streptococcus agalactiae* (also known as Group B *streptococcus* or GBS). It was found that each new genome had an average of 30 genes that were not present in any of the previously sequenced genomes. Not every bacterial species has the same level of complexity as GBS. For instance, the pangenome for *Bacillus anthracis* can be adequately described by four genome sequences. Hence, scientists refer to certain species as having an "open" and others a "closed" pangenome. In species with an open pangenome, there

are an unlimited number of new genes found for every genome. In closed pangenomes, there are only a limited number of strain-specific genes. The differences in the nature of the pangenome reflect several factors: differing lifestyles of two organisms, the number of closely related species in the same environment and physiological state, the ability of each species to acquire and stably incorporate foreign DNA (an advantage in niche adaptation from the acquisition of laterally transferred DNA), and the recent evolutionary history of each species. It should be noted that the imperfection of our definition of a bacterial species, for example, *B. anthracis* and *B. cereus,* can be considered the same species by some criteria, may render pangenome analysis more complicated. As the definition of a pangenome improves, the coverage of strains included in a bacterial species will change and thus alter the analysis results.

A pangenome can be divided into three elements: (1) a core genome that is shared by all strains (2) a set of dispensable genes that are shared by some but not all isolates, and (3) a set of strain-specific genes that are unique to each isolate. For *S. agalactiae*, the core genome encodes the basic aspects of *S. agalactiae* biology and was as such predicted to rapidly converge to 80% of the genome in each isolate. Conversely, dispensable and strain-specific genes, which are largely composed of hypothetical, phage-related and transposon-related genes [108], contribute to its genetic diversity. The concept of the pangenome and comparative genomics has practical applications in vaccine research. In fact, while obviously the ideal vaccine candidate is a conserved protein encoded by a gene present in every isolate of the species, in the case of GBS it was shown that the design of a universal protein-based vaccine against GBS was possible using dispensable genes [109]. Of note, capsular-specificity genes and other pathogenicity traits are often identified in an accessory genome. Moving forward, bacterial taxonomy and epidemiology must take into consideration whole genome sequences and not just a few genetic loci, as has been the case so far with methods such as ribosomal RNA sequences, capsular typing, and multilocus sequence typing (MLST). Comparison of the whole genome sequences of GBS strains has shown that the genomic diversity does not necessarily correlate with serotypes or MLST sequence-types. The application of additional whole genome sequence analysis will require that epidemiology studies have a reliable, systematic correlation between strains and disease and permit a standardization of the classification for clinical isolates. These observations are instrumental for developing a protective vaccine that covers a broad range of pathogenic strains.

Comparative genomics is also important for the identification of pathogenic factors since they potentially represent good vaccine candidates. The level of distinction and the function played by carrier versus virulent strains of *streptococci* and *neisseriae*, for example, has been the matter of discussion for a long time and still lacks an answer. There is, as yet, no clear and strict correlation between the presence of apparent virulence factors and the diseases caused by these organisms. The epidemiological evidence is vague and does not provide definitive clues. There may be multiple reasons for this apparent lack of correlation. It is likely that in species that only rarely result in disease, there exist multiple virulence factors and toxins that are uniquely associated with infection. Therefore, comparative genomics can be used to identify the "pathogenicity signature" associated with the most virulent bacterial strains or the strains that are successful in colonization. Comparative genomics can also be used to compare various strains that exhibit different virulence levels, for example, commensal nonpathogenic strains versus virulent ones, to find specific vaccine candidates [110]. The advantages include making a vaccine against commensal strains and narrowing down the pool of vaccine candidates. It is anticipated that most virulence factors will be found in accessory genomes, at least the ones that determine increased pathogenicity. However, presently comparative genomics is not able to identify expression variability that contributes to the different manifestations of pathogenicity of bacterial strains. Hence, functional studies are still critically needed to shed light on the relevance of specific virulence factors.

Another potential application could be studies of certain species of bacterial symbionts such as *Mycoplasma*, *Rickettsiae,* and *Chlamydiae*. These species, instead of acquiring genes during evolution, have actually lost significant levels of their genetic information [111]. Primarily biosynthetic pathway genes have been lost because intracellular bacteria have a relatively constant environment with access to much of what they require for survival. By applying the concept of pangenomics to these species, we would obtain a "microgenome" representative of the set of genes necessary to live in the intracellular niche. Comparing this "microgenome" with the pangenome for free-living species will likely simplify the identification of genes necessary for the microorganism to survive in varying and unfavorable environments. Housekeeping genes specific to the pathogen are considered vaccine candidates.

In comparison to a half decade ago, comparative genomics studies have become incredibly easy to perform. For the most important human pathogens, the average number of genomes for the different available strains is above five. Therefore, for new RV studies, the conservation level of selected antigens can be determined. Antigen conservation level is important since conserved antigens can be used to develop a broad strain protective vaccine [32].

In addition to the basic mechanisms of the RV strategy described above, additional criteria can be added. For example, since outer membrane proteins containing more than one transmembrane helix are difficult to clone and purify [10], the number of transmembrane domains of a candidate protein is often used as an additional filtering criterion. Bacterial adhesins play critical roles in adherence, colonization, and invasion of microbial pathogens to host cells [112]. Therefore, adhesins are essential for bacterial survival and are possible targets for vaccine development. Two RV software programs, NERVE [113] and Vaxign [110], utilize these criteria. Since RV focuses on predicting antigens using protein sequences, immune epitope prediction based on amino acid sequences can also be considered as a criterion for RV vaccine design [110].

Vaxign (http://www.violinet.org/vaxign/) is the first web-based vaccine design program based on genome sequences

utilizing the RV strategy. Predicted features in the Vaxign pipeline include protein subcellular location, transmembrane helices, adhesin 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. Vaxign has been demonstrated to successfully predict vaccine targets for *Brucella* spp. [114, 115] and uropathogenic *Escherichia coli* [110]. Currently, more than 100 genomes have been precomputed using the Vaxign pipeline and available for query in the Vaxign website. Vaxign also performs dynamic vaccine target prediction based on input sequences.

The availability of three-dimensional structure may facilitate epitope prediction and antigen discovery [116, 117]. It would be ideal to also consider inclusion of analysis of high throughput transcriptomics and proteomics data to aide in complementary identification of vaccine candidates.

4. Transcriptomics and Proteomics Data Analysis for Vaccine R&D

Beside genomics methods in vaccine studies (described above), high-throughput transcriptomics and proteomics technologies (i.e., microarray) have been used for vaccine target design and analysis of vaccine-induced host immune responses. These assay systems are able to measure the expression pattern of thousands of genes in parallel, permitting the generation of large amounts of gene expression data. Bioinformatics techniques will play a critical role in analyzing such data and in making novel discoveries. In general, bioinformatics analysis of transcriptomics and proteomics data includes the following: (1) data preprocessing such as data quality controls and normalization, (2) statistical analysis of significantly regulated genes, (3) gene grouping and pattern discovery analyses, and (4) inference of biological pathways and networks [118, 119]. Depending on the specific research goals of any given project, different informatics tools may be applied individually or in combination.

Data processing is important in minimizing the effects of experimental artifacts and random noise. Companies that market microarrays usually provide their own methods for raw data processing and data quality control. For example, the GeneChip Operating Software (GCOS) expression analysis software provided by Affymetrix (Santa Clara, CA) can be used to process image data and the signals from the Affymetrix DNA microarrays [120]. The probe sets of Affymetrix microarray data are labeled present (P), absent (A), or marginal (M) based on the default P values set up in the GCOS system. Such labeling provides a useful approach for gene filtering. Commonly used microarray normalization methods include the Affymetrix MicroArray Suite MAS 5.0 (implemented in GCOS), the Robust Multichip Analysis (RMA) method [121], and the method of Li and Wong [122]. The software programs implementing these methods can be downloaded from the BioConductor (http://www.bioconductor.org/), a repository for open source and open development software programs developed specifically for the analysis and comprehension of omics data [123].

A common task in analyzing microarray data is to identify up- or down-regulated gene lists [124]. Fold changes of gene expression values between treatment group and nontreated controls were first used by biologists. However, this method may miss biologically important genes that exhibit small fold changes but have statistical significance. It also overemphasizes those genes with large fold changes but have little or no statistical significance [119]. Frequently used statistical methods for the determination of significantly changed genes include analysis of variance (ANOVA) [125], significance analysis of microarrays (SAM) [126], and the BioConductor package Linear Models for Microarray Data (LIMMA) [127]. ANOVA is a highly flexible analytical approach and is used in various commercial and open-source software packages [125]. SAM identifies genes with statistically significant expression changes by assimilating a set of gene-specific t-tests [126]. LIMMA uses linear models and empirical Bayesian methods to assess differential expression in microarray experiments [127].

Once the lists of up- or down-regulated genes are determined, they can be grouped into expression classes to identify patterns of gene expression and to provide greater insight into their biological functions and relevance. "Unsupervised and supervised" computational methods can be used for gene clustering analysis [128]. "Unsupervised" methods arrange genes and samples in groups or clusters based solely on the similarities in gene expression. Examples of unsupervised clustering methods include hierarchical clustering [129], self-organizing maps [12], and model-based clustering (e.g., CRCView [130]). "Supervised" methods, for example, EASE [131] and gene set enrichment analysis (GSEA) [132], use sample classifiers and gene expression to identify hypothesis-driven correlations. The Gene Ontology program (GO) is frequently used for gene enrichment analysis by many software programs, such as DAVID [133] and GOStat [134]. Additional GO-based microarray data analysis approaches can be found at http://www.geneontology.org/GO.tools.microarray.shtml.

The next level of DNA and protein array data analysis is the inference of biological pathways and networks [135, 136]. Several methods have been explored to model gene expression data including simple correlation [137], differential equations [138], neural networks [139], and Bayesian networks [140, 141]. These methods have different advantages and disadvantages [135, 136]. Simple correlation assumes linear and typically pairwise relationships. These limitations render it difficult for the investigator to identify multidimensional relationships between variables [142]. While methods utilizing differential equations are accurate, they are often "hand created" and as such are limited to the use of a small number of variables [142]. In contrast, neural networks make accurate predictions by mapping the data onto a high-dimensional polynomial. This allows the variables to influence each other in complex ways [139]. However, the use of neural networks assumes that everything is affected by the changing variable. This renders it difficult to identify such mechanisms. Bayesian networks (BN) represent a powerful method for identifying causal or apparently causal patterns in gene expression data. A key advantage

of Bayesian networks is that they are relatively agnostic to the complexity of the relationships predicted and can model linear, nonlinear, combinatorial, stochastic, and other types of relationships among variables across multiple levels of biological organizations [143]. However, current Bayesian network approaches are also subject to limitations. For example, the expression levels must be discretized, leading to varying degrees of loss of information [135].

The combined application of transcriptomics and proteomics experiments in conjugation with specialized informatics analyses has many applications in the field of vaccine research and development. First, these "omics" methods can be used to discover vaccine targets for many microorganism-induced diseases as well as cancers [144, 145]. For example, the sexual stages of malarial parasites are essential for transmission of the disease by the mosquito and as such are the targets for malaria vaccine development. To better understand how genes participate in the sexual development process, Young et al. utilized microarrays to profile the transcriptomes of high-purity stage I-V *Plasmodium falciparum* gametocytes [146]. An ontology-based pattern identification algorithm was applied to identify a 246 gene sexual development cluster. Some of the genes have the potential of being used for vaccine development. Sturniolo et al. [147] developed a matrix-based computational algorithm when applied to DNA microarray experiments all data was used successfully to predict human leukocyte antigen (HLA) class II ligands and differentially expressed colon cancer genes. A list of peptides uniquely associated with colon cancer was identified. These are potentially immunogenic. These peptides provide a basis for rational vaccine development against colon cancer.

One practical problem in vaccine investigation is that for most diseases, no immune response correlates well with protection. To solve this issue, systems biology (Omics and bioinformatics) approaches have also been used to detect gene signatures induced in vaccinated hosts (e.g., humans) that correlate and even predict protective immunity. For example, two recently published studies examined early gene signatures induced in humans vaccinated with the attenuated yellow fever vaccine YF17D [148, 149]. Each study analyzed total peripheral-blood mononuclear cells from different cohorts of human volunteers at various time points following vaccination with YF17D. Early effects (3 and 7 days postvaccination) on gene expression were determined using microarrays and were analyzed using bioinformatics approaches. Many genes involved in innate immune response (e.g., Toll-like receptor signaling and inflammasome) were discovered. Gaucher et al. [149] identified a group of transcription factors, including interferon-regulatory factor 7 (IRF7), signal transducer and activator of transcription 1 (STAT2), and ETS2, as key regulators of the early immune response to the YF17D vaccine [149]. YF17D was found to trigger the proliferation of several leukocyte subtypes including macrophages, dendritic cells, natural killer cells, and lymphocytes [149]. Definition of this "baseline" innate immunity response subsequently allowed detection of defective hyperresponse (excessive CCR5 activation) in a YF17D vaccinee who had developed a serious viscerotopic adverse event [150]. In another study, Querec et al. [148] discovered gene signatures that correlate with the magnitude of antigen-specific $CD8^+$ T-cell responses and antibody titers [148]. EIF2AK4, a key gene in the integrated stress response, was found among most of the predictive signatures. The actual predictive capacity of a gene signature was verified using the signatures for $CD8^+$ T-cell responses from the first trial to predict the outcome of the second trial and vice versa. Another distinct early gene signature that included TNFRSF17 (a receptor for B-cell-activating factor) was found to predict the neutralizing antibody titers as late as 90 days following vaccination [148].

Microarray-based methods have also been used to investigate vaccine safety [151]. For example, McKinney et al. used protein microarrays to compare 108 serum cytokines and chemokines in vaccine recipients before and one week after smallpox vaccination [151]. Among 74 individuals studied, 22 experienced systemic adverse events. Machine-learning and statistical analyses identified six cytokines that accurately discriminate between individuals on the basis of their adverse event status. A DNA microarray-based system has also been developed to evaluate the genetic signatures of the toxicity of many vaccines including pertussis vaccine [152] and influenza vaccines [153].

5. Mathematical Simulations for Vaccine R&D

Integrative research, development, and uses of vaccines follow a cyclical fashion where mathematical and computational simulations are connected with experimentation leading to improved accuracy and reduced cost in vaccine R&D [154]. Many mathematic simulations have been developed to support different areas of vaccine research and development (R&D). These studies support various vaccine-associated aspects including vaccine discovery and development, vaccine production and stockpiling, vaccination protocol optimization, vaccine distribution, and vaccine regulation. Here we introduce some striking examples.

Mathematical models have been developed to study the dynamics of host-pathogen and host-vaccine interactions [155]. For example, Kirschner et al. integrate information over relevant biological and temporal scales to generate a model for major histocompatibility complex class II-mediated antigen presentation [156]. This multiscale mathematical model simulates molecular, cellular, tissue, and organ/organism, and the interactions between different levels. This model has been used to answer questions about mechanisms of infection and new strategies for treatment and vaccines. The same group has developed a multifaceted approach to modeling tuberculosis-induced granuloma, a self-organizing structure of immune cells forming in the lung and lymph nodes in response to bacterial invasion [157–159]. Many mathematical models have been developed to understand the mechanisms and limitations of HIV control by humoral and cell-mediated immunity [160]. These studies suggest that $CD8^+$ T-cells do "too little too late" to prevent the establishment of HIV infection. However, passively administered antibody acts very early to reduce the initial viral count and slow HIV growth [160]. Cell culture-based

influenza vaccine manufacturing is of growing importance. Influenza virus is able to replicate and induce apoptosis in host cells. Combined with experiments, Schulze-Horsel et al. have formulated a mathematical model to describe changes in the concentration of uninfected and influenza A virus-infected adherent cells, dynamics of virus particle release, and the time course of the percentage composition of the cell population [161]. This model can be used to characterize and maximize viral titer yield in the bioreactors meant to produce virus for use in influenza vaccines.

Cost-effectiveness analyses (CEA) of vaccination programs can be performed using mathematical modeling [162]. For effective evaluation of cost effectiveness, a model is generally required which considers the relevant biological, clinical, epidemiological, and economic factors of a vaccination program. CEA modeling methods have been categorized based on three main attributes: static/dynamic, stochastic/deterministic, and aggregate/individual based. The modeling methods for CEAs of vaccination programs can be improved in the areas of model choice, construction, assessment, and validation [162]. CFA has been applied to study different vaccination programs such as human papillomavirus (HPV) vaccination [163], influenza vaccination [164], and vacation with pneumococcal conjugate vaccine [165].

Mathematical modeling can be used to simulate and optimize vaccination protocols. The combination of *in silico* and *in vivo* studies has the ability to reduce the time, effort, and cost of vaccine studies by orders of magnitude [166, 167]. For example, Pappalardo et al. designed and implemented SimTriplex, an agent-based model specifically tailored to simulate the effects of tumor-preventive cell vaccines in HER-2/neu transgenic mice prone to mammary carcinoma development [168]. The SimTriplex mathematical model combined with genetic algorithm has been used to search for new vaccination schedules to prevent tumors in HER-2/neu transgenic mice [166, 169, 170]. It has been found that the computational model can be used for simulation of immune responses ("*in silico*" experiments), leading to optimization of vaccine protocols. Pennisi et al. also developed MetastaSim, a hybrid Agent Based-ODE model for the simulation of the Triplex cell vaccine-elicited immune system response against lung metastases in mice [167]. MetastaSim simulates the main features of the immune system. Both innate and adaptive immune responses are covered. This model includes different cell types and molecules, such as dendritic cells, macrophages, cytotoxic lymphocytes, antibodies, antigens, IL-12, and IFN-γ. Their study with MetastaSim demonstrated that it is possible to obtain *in silico* a 45% reduction in the number of vaccinations [167].

Mathematical modeling plays an important role in postlicensure vaccine informatics and in assessing the impact of immunizations against target diseases. For example, Blower et al. developed a mathematic model to predict the tradeoff between efficacy and safety of live attenuated HIV vaccines [171]. More details in this topic are introduced in the following section.

6. Postlicensure Vaccine Informatics

Successful vaccine immunization induces protective immunity in the individual. Equally important for most infectious diseases, when a sufficiently high threshold of a group of individuals is immunized, a "herd effect" is observed at the population level where the incidence of the disease in the remaining unimmunized members of the group is lower than it would be otherwise [172]. Due to the large societal benefits of immunizations, almost all governments (generally at the state/provincial or national levels) organize formal targeted immunization programs to maximize vaccine coverage. The impact of the immunization programs is to reduce the incidence of the targeted disease. For some infectious diseases where the characteristics permit [173] (e.g., smallpox, polio, measles, neonatal tetanus), regional or global initiatives to eliminate or eradicate the targeted disease may be organized. Routine or special immunization programs are incredibly complex to initiate. Ongoing endeavors not uncommonly require careful orchestration and planning for sustained and repeated immunizations of millions of persons annually in most jurisdictions. Vaccine informatics is critical to providing accurate data and facilitates the smooth planning, organization, implementation, and monitoring of almost every aspect of such complex immunization programs. The introduction of each new recommended vaccine into an already crowded pediatric immunization schedule adds to this complexity [174]. We describe next some of the better known postlicensure vaccine informatic systems: tracking immunization history in computerized immunization information systems (IIS) or registries, informatics methods for improving surveillance of vaccine safety and efficacy, and modeling impact of alternative immunization strategies against target diseases.

6.1. Computerized Immunization Information Systems (or Immunization Registries). Accurate tracking of vaccination history is essential to ensure proper completion of the primary immunization schedule and subsequent booster doses. This seemingly straightforward task is nontrivial system-wide when compounded by an increasingly mobile population, immunization schedules of increasing complexity, multiple vaccine manufacturers of the same vaccine, multiple health care providers and/or health insurance for the same individual (a problem in the U.S.), multiple individual with same name, and so forth. Add in small vaccine vials with hard to read small fonts in a busy pediatric clinic serving many crying babies simultaneously, the opportunities for inaccurate or nonrecording of an administered vaccination is substantial in developed and developing countries.

Computerized immunization information systems (IIS) provide an obvious potential solution to these challenges. In the U.S., the first large IISs were organized in Delaware in the early 1970s [175]. This action was followed by several health maintenance organization (HMOs) with the dual purpose of linking the IIS to medical visits for rigorous studies of vaccine safety [176]. The Robert Wood Johnson Foundation funded the All Kids Count I and II programs in the 1990s in multiple communities. This provided an important

impetus to the field [175]. The Centers for Disease Control and Prevention (CDC) now provide some financial and technical assistance for public sector IIS in almost every state (http://www.cdc.gov/vaccines/programs/iis/default.htm). This work is aided by partners such as the American Immunization Registry Association (http://immregistries.org/) and the Public Health Informatics Institute (http://phii.org/). Internationally, Australia [177], Canada [178], and Norway [179] are some of the other countries with active IIS.

IISs also have the potential to provide a foundation of child health registries [175] and electronic health records [180]. While initially focused on routine pediatric immunizations, registries in many locations have been expanded to meet other needs, including adolescent and adult immunizations [181], disasters [182], targeting of at risk populations [182, 183], study vaccine refusal [184], and facilitating accurate and timely reporting of vaccine adverse events [185]. Substantial progress has also been attained in the protection of privacy and confidentiality; in ensuring participation of all immunization providers and recipients, to ensure appropriate functioning of registries, and to ensure sustainable funding for registries [186]. However, challenges remain in exchanging information among different IISs, and across state lines. The National Vaccine Advisory Committee has issued recommendations on how to overcome these challenges (http://www.hhs.gov/nvpo/nvac/IISRecolmmendationsSep08.htm).

6.2. Informatics Methods for Improving Surveillance of Vaccine Safety and Efficacy. Before a vaccine is licensed, it undergoes rigorous testing in preclinical (laboratory and animal) and phased human clinical trials for safety and efficacy [187]. Due mainly to cost (intensive monitoring perprotocol is expensive) and ethical (once a vaccine is determined to be safe and effective, it is no longer ethical to withhold it from others in need) considerations, however, the sample size and duration of followup in prelicensure trials are usually limited. This means surveillance for both vaccine safety and effectiveness [188] in larger immunized population postlicensure and postmarketing is needed. This is challenging because trial conditions (e.g., double-blinding, randomization) that permit straightforward comparison between vaccinated and unvaccinated groups no longer hold. Substantial data collection and adjustments on possible confounders, when possible, are needed to fully analyze and interpret such observational studies.

Post-licensure monitoring for vaccine safety can generally be divided into hypothesis generating and hypothesis testing. Since vaccine coverage for many vaccines can be close to universal, by definition, anyone with a medical adverse event will have previously been vaccinated. Spontaneous reporting or passive surveillance systems like the U.S. Vaccine Adverse Event Reporting System (VAERS, http://vaers.hhs.gov/) in the U.S. [189] and elsewhere [190], where medical problems suspected to be caused by the vaccination can be reported to health authorities, provide the bulk of new vaccine safety hypotheses. Due to the large number of reports (>20,000 annually to VAERS), data mining techniques are beginning to be applied to triage reports worthy of further attention [191].

Once a vaccine safety concern is provisionally identified, based on our understanding of likely pathophysiology and nonrandom clustering of cases in onset time after vaccination, a formal study is usually needed to (1) confirm whether the etiologic link with vaccination is real and not coincidental, and (2) identify the magnitude of the risk (to assist in risk-benefit determination for the immunization). Since these safety concerns are likely to be rare (otherwise they would have been detected pre-licensure), confirmatory pharmacoepidemiologic studies of large vaccinated populations are usually needed to "test the hypothesis". Large national (e.g., Denmark) or population (e.g., Managed Care Organization (MCO)) health care systems, where members have unique personal identifiers and most of the care for both vaccinations (exposure) and medical visits (outcome) are automated, provide an efficient platform for piggy-backing vaccine safety pharmacoepidemiologic studies [192]. The Vaccine Safety Datalink (VSD) project in the US, a consortium of 8 MCO's representing ~3% of the population, has been used as a prototype of how such large linked databases can be used for rigorous vaccine safety studies [176, 193]. Examples include rotavirus vaccine and intussusception [194], thimerosal and neurologic adverse events [195], and vaccinations and central demyelination [196]. Similar large-linked vaccine safety databases have been created in England [197] and Vietnam [198].

Safety issues cannot be assessed directly and can only be inferred from the relative absence of multiple adverse events. Therefore, standardizing the case definitions used to assess adverse events is needed to allow for meaningful comparison of vaccine safety data in various settings. Recognizing this need, the Brighton Collaboration (https://brightoncollaboration.org/public) was formed in 1999 as a voluntary global collaboration to facilitate the development, evaluation, and dissemination of high quality information about the safety of human vaccines in both pre- and post-licensure settings. To date, 28 guidelines and case definitions have been developed and are freely available to users. The case definitions are tiered by the level of evidence available and will differ based on whether the data are gathered in prospective clinical trials or passive postmarketing surveillance and on the level of resource availability (e.g., developed versus developing countries). Since its inception, the Collaboration has helped to form a critical mass of experts interested in vaccine safety that can potentially be convened or accessed as new vaccine safety issues arise. The Brighton Collaboration Viral Vector Vaccine Safety Working Group is exploring using the "wiki" model of mass collaboration for completing and maintaining standard templates on characteristics of various viral vectors [199].

Post-licensure monitoring for vaccine effectiveness is usually done by examining the impact on targeted diseases. For reasonable sensitivity and specificity for monitoring trends of the disease, this usually requires the establishment of some type of public health surveillance system. For example, the recent reintroduction of rotavirus vaccine in the US has resulted in delayed onset and diminished magnitude of

rotavirus activity [200]. A decline in invasive pneumococcal disease was observed after the introduction of conjugate pneumococcal vaccine [201]. Similar data was obtained in developing countries, as was done with introduction of conjugate *Haemophilus influenza* type b vaccine in Mali [202]. When disease remains high or an outbreak occurs despite high vaccine coverage, a special epidemiologic study to assess vaccine effectiveness may be needed. This type of action was undertaken after a posthoneymoon period measles outbreak in Burundi [203], the resurgence of diphtheria in the former Soviet Union [204], and the introduction of a monovalent oral type 1 poliovirus vaccine in India [205].

6.3. Modeling of Impact of Immunizations against Target Diseases. The cyclical nature of epidemics of many infectious diseases such as plague and smallpox in humans (and other animals) was noted by ancient historians prior to the introduction of immunization in modern times [206]. This periodicity was described as a mathematical relationship between susceptible and immune individuals in a population over time that interacted with an external infectious force by Ronald Ross and Anderson Gray McKendrick at the beginning of the 20th Century [207]. It was not until the early 1980's, however, that Anderson and May systematically organized and effectively organized disparate works in population biology, ecology, and epidemiology into mathematical models of infectious diseases that linked the theory with practical translation into public health policy (e.g., vaccinations) [208]. Their 1991 textbook "Infectious Diseases of Humans: Dynamics and Control" [209] has helped to create a cohort of mathematical modelers that have furthered our understanding of transmission of infectious agents within human communities and design programs for their control. As Geographic Information Systems (GISs) that integrate and analyze spatial information become increasingly available for linkage with public health databases [210], this should aid continued refinements in various assumptions used in mathematical models of infectious diseases.

Irrespective of the model or the target disease, a key concept in any mathematical model is the basic reproductive rate or R_0 of a microorganism—the average number of secondary infections produced when one infected individual is introduced into a totally susceptible population. The goal of any control program (e.g., immunizations) is to reduce the R_0 as much as possible. For disease elimination or eradication programs, it must be <1 [211]. Another key concept is "herd immunity", the indirect effect of some vaccines on reduction of disease transmission beyond the protection in actual vaccine recipients [172]. Most mathematical models attempt to describe as accurately as possible the flow of a human population from susceptibility (usually at birth or with the waning of maternally derived immunity), infected (by wild disease or vaccination), and immune states (adjusting for various variables such as mixing) transmission coefficient, vaccine effectiveness, and duration of protection. Each of these variables in turn can be further modeled (e.g., mixing can differ with age-classes or other subpopulations).

Historically, one of the more successful integrations of mathematical modeling of vaccine-preventable diseases and immunization program policies has been for measles [203, 211, 212] and rubella [213]. Mathematical models have also been critical for understanding how best to (a) introduce newly licensed vaccines like human papillomavirus vaccine [214], (b) control new emerging public health problems, such as pandemic influenza [215], (c) how best to optimize control of a vaccine-preventable disease (such as impact of pneumococcal conjugate vaccine on emergence of penicillin-resistant strains) [216], or (d) how spatio-temporal variation in birth rates may explain the observed patterns of rotavirus disease after the introduction of new rotavirus vaccine [200].

7. Vaccine Literature Mining, Databases, and Data Integration

Vaccine informatics is dedicated to the acquisition, processing, storage, distribution, analysis, and interpretation of vaccine-associated data by means of computing methods and tools. Advanced DNA sequencing, molecular, cellular, and immunological methods have provided a huge amount of vaccine-related data. These data have been processed and analyzed by exponentially expanded computational power and new algorithms. To facilitate advanced vaccine research and development, the large amounts of vaccine literature data need to be processed and mined. Different types of vaccine databases are also needed to store various vaccine data. Eventually, all these data need to be integrated within the vaccine domain and with other biomedical data for computational reasoning and discovery of new knowledge.

7.1. Vaccine Literature Mining. The papers and authors related to vaccine/vaccination have increased exponentially. Only six vaccine-related papers were published and recorded in PubMed before 1900. In the first half of the 20th century, 1,210 vaccine-related papers were published. This number has increased almost 100-fold in the second half of the last century. In addition, the numbers of vaccine publications have increased exponentially (Figure 3). For example, 6,399 vaccine-related papers were published during the period of 1951–1960, and 96,938 in 2001–2010. Therefore, the number of papers published annually in PubMed has increased more than 15-fold during the past 50 years.

It has become increasingly challenging to retrieve useful vaccine data for research purposes from the huge amount of vaccine literature. Literature mining has been used to facilitate the discovery and analysis of potential vaccine targets. For example, cross-matching and analysis of the literature and *in silico*-derived data allowed the selection of 189 putative vaccine candidates from the entire *Mycobacterium tuberculosis* genome [217]. In this study, the first step towards the selection of vaccine candidates was to accumulate published experimental data from a literature scan of documented studies with a focus on global analyses. The literature sources were then grouped based on different categories (e.g., macrophage). This literature mining approach detected 189 potential vaccine candidates. These were studied further through *in silico* functional analysis and immunoinformatics epitope prediction. A qualitative score

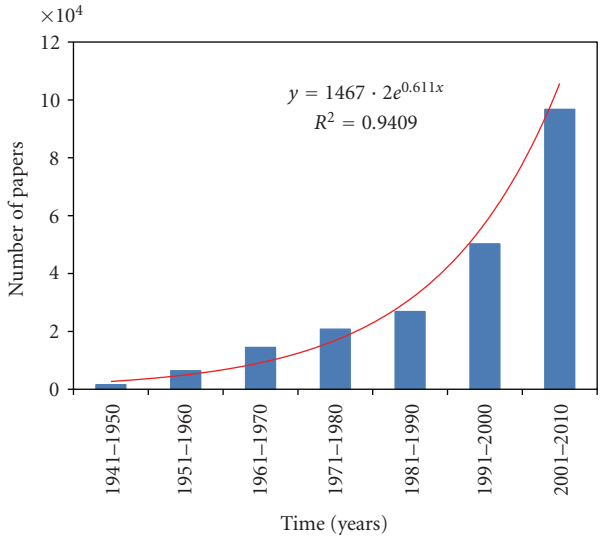

Figure 3: Exponential growth of papers in the area of vaccine and vaccination. The data was obtained by analysis of available papers in PubMed.

was designed based on a total of 14 criteria and used to rank and prioritize the gene list [217].

Literature mining can also be used to analyze vaccine-associated host immune response networks. For example, Ozgur et al. recently applied a literature mining and network centrality analysis [218] to analyze the IFN-γ and vaccine-associated gene networks [219]. Among approximately 1,000 genes found to interact with IFN-γ, 102 genes were predicted to be vaccine-associated and 52 of them were verified by manual curation. The production of IFN-γ is crucial for successful immune response induced by vaccines against various viruses and intracellular bacteria. For example, these include HIV [220], *M. tuberculosis* [221], *Leishmania* spp. [222], and *Brucella* spp. [223]. The discovery of the IFN-γ and vaccine-mediated gene network provides a comprehensive view of the vaccine-induced protective immune network and generates new hypotheses for further experimental testing.

Two literature mining programs presented in the Vaccine Investigation and Online Information Network (VIOLIN; see next section) were developed for general vaccine literature searching and analysis [13]. Vaxpresso (http://www.violinet.org/textpresso/cgi-bin/home) is a vaccine literature mining program using natural language processing (NLP) and ontology-based literature searching [224]. For a list of selected pathogens, Vaxpresso contains all possible vaccine-related papers extracted from PubMed (http://www.ncbi.nlm.nih.gov/pubmed). Vaxpresso is able to retrieve and sort article sentences that match specific keywords and ontology-based categories. Vaxmesh (http://www.violinet.org/litesearch/meshtree/meshtree.php) is a vaccine literature browser based on the Medical Subject Headings (MeSH). MeSH is a controlled vocabulary of medical and scientific terms that is used for indexing PubMed articles in a consistent way supporting PubMed literature mining. Vaxmesh enables users to locate articles using MeSH terms in a hierarchical MeSH tree structure.

7.2. Web-Based Vaccine Databases and Online Resources. Many publicly available vaccine databases and online resources exist (Table 3). For example, the USA CDC Vaccine Information Statements (VISs) system (http://www.cdc.gov/vaccines/pubs/vis/) provides information sheets that explain to vaccine recipients, their parents, or their legal representatives both the benefits and risks of a vaccine. Federal law in the US requires that VISs be handed out for all vaccines before their use. The licensed vaccine information is provided by the U.S. FDA (http://www.fda.gov/BiologicsBloodVaccines/Vaccines/default.htm). The Vaccine Resource Library (VRL, http://www.path.org/vaccineresources/) offers various high quality, scientifically accurate documents and links to specific diseases and topics in immunization.

These databases focus primarily on the clinical uses and regulations of existing vaccines for vaccine users. To store and analyze research data concerning commercial vaccines and vaccines under clinical trials, or in early stages of development, the Vaccine Investigation and Online Information Network (VIOLIN, http://www.violinet.org/) was developed. VIOLIN is a web-based vaccine database and analysis system primarily targeted for vaccine researchers [13]. The VIOLIN vaccine database currently contains more than 2,700 vaccines, or vaccine candidates, for more than 160 pathogens through manual curation from >1500 peer-reviewed papers or other reliable sources. The stored vaccine data includes vaccine preparation, pathogen genes used and gene engineering, vaccine adjuvants and vectors, vaccine-induced host immune responses, and vaccine efficacy in host after virulent challenge. VIOLIN curates more than 500 protective antigens (http://www.violinet.org/protegen/) [225]. Vaccine-related pathogen and host genes are also annotated and available for searching through customized BLAST programs. VIOLIN also stores and processes all possible vaccine literature through different text mining programs [13]. Vaxign, a web-based vaccine design program based on reverse vaccinology strategy [110], is also a program in VIOLIN.

Besides the above databases which focused on vaccine awareness and vaccine research, many other databases are available that are useful for vaccine research and development. For example, more than 65,000 antibody and T-cell epitopes have been deposited in the Immune Epitope Database and Analysis Resource (http://www.immuneepitope.org/) since the database was established in 2004 [6]. These immune epitopes cover a broad range of species including humans, nonhuman primates, rodents, and other animal species as related to all infectious diseases [6]. AntigenDB is an immunoinformatics database of pathogen antigens and store sequences, structures, origins, and epitopes [226].

7.3. Development of a Community-Based Vaccine Ontology (VO). Although public vaccine databases provide help with different aspects of vaccine knowledge and research, it remains a challenge to integrate this disparate body of information on vaccines. Data integration is hampered since the data are often collected using incompatible or poorly described methods for data capture, storage, and

Table 3: Vaccine web resources.

Resource name	Website URL	Comment
Vaccines and Immunization		
WHO Immunization/vaccines	http://www.who.int/immunization/en/	WHO vaccine site
Immunization Action Coalition	http://www.immunize.org/	Vaccination Information for Healthcare Professionals
USA CDC Vaccine Information Statements	http://www.cdc.gov/vaccines/pubs/vis/	Benefits and risks of vaccines
USA CDC listed vaccines	http://www.cdc.gov/vaccines/vpd-vac/vaccines-list.htm	Vaccines used in USA
US FDA licensed vaccine information	http://www.fda.gov/BiologicsBlood-Vaccines/Vaccines/default.htm	Licensed vaccines used in USA
Canada licensed vaccine information	http://www.phac-aspc.gc.ca/dpg-eng.php#vaccines	Canada
DH Immunization at UK	http://www.dh.gov.uk/en/Publichealth/Immunisation/index.htm	Official UK vaccination site
PATH Vaccine Resource Library	http://www.path.org/vaccineresources/	Collection of vaccine resources
NNii: National Network for Immunization Information	http://www.immunizationinfo.org/vaccines	Scientific valid information
GAVI Alliance (Global Alliance for Vaccines and Immunisation)	http://www.gavialliance.org/	Goal: save children's lives
Vaccine Clinical Trials		
Nonhuman primate HIV/SIV vaccine trials database	http://www.hiv.lanl.gov/content/vaccine/home.html	Vaccine studies of HIV/SIV using nonhuman primates
AIDS vaccine trials database	http://www.iavireport.org/trials-db/Pages/default.aspx	AIDS vaccine clinical trials
Clinical trials database (USA NIH)	http://clinicaltrials.gov/	Vaccine or other trials
Vaccine Safety		
WHO Immunization Safety	http://www.who.int/immunization_safety/en/	WHO vaccination safety site
VAERS: Vaccine Adverse Event Reporting System (USA FDA & CDC)	http://vaers.hhs.gov/index	USA vaccine adverse event reporting website
USA CDC-Vaccine Safety	http://www.cdc.gov/vaccinesafety/index.html	USA official vaccine safety site
The Brighton Collaboration	https://brightoncollaboration.org/public	Setting standards in vaccine safety
Vaccine Research Database		
VIOLIN	http://www.violinet.org/	Comprehensive vaccine data
Vaccine Manufacturers		
EVM: European Vaccine Manufacturers	http://www.efpia.org/Content/Default.asp	Collection of European vaccine companies
List of vaccine manufactures	http://www.ontobee.org/browser/rdf.php?o=VO&iri=http://purl.obolibrary.org/obo/VO_0000299	Collected in Vaccine Ontology

dissemination. Integration is also complicated as investigators use independently derived local terminologies and data schemas. These problems can be alleviated through the use of a common ontology, that is, a consensus-based controlled vocabulary of terms and relations, with associated definitions that are logically formulated in such a way as to promote automated reasoning. Ontologies are able to structure complex biomedical domains and relate a myriad of data to allow for a shared understanding of vaccines.

The collaborative, community-based Vaccine Ontology (VO; http://www.violinet.org/vaccineontology/) was recently initiated to promote vaccine data standardization, integration, and computer-assisted reasoning. VO can be used for different applications, including vaccine data integration and

literature mining. Currently, VO contains more than 3,000 terms, including more than 700 vaccines and vaccine candidates that are represented in an appropriately structured ontological hierarchy. These vaccines or vaccine candidates are targeted to 70 pathogens and have been studied in more than 20 animal species (e.g., human, mouse, cattle, and fish). VO also stores terms related to different vaccine components (e.g., protective antigens, vaccine adjuvants and vectors), vaccine-induced immune responses, vaccine adverse events, and protection efficacy. The known relations between these terms are also listed. These representations are readable by computer programs and support computer-assisted reasoning. This knowledge is also exchangeable across multiple scientific domains to facilitate hypothesis generation and validation. This approach will undoubtedly lead to new scientific discoveries.

VO has been used for several different applications. For example, VO, in combination with other ontologies, has been used to model and study vaccine protection investigation [15]. Reported vaccine protection data from different reports can be systematically analyzed [227]. VO can also be used to improve vaccine literature mining. For example, a direct PubMed search for "live attenuated *Brucella* vaccine" returned 69 papers (as of August 2010). VO includes 13 live attenuated *Brucella* vaccines that are defined as "live" and "attenuated". When specific "live, attenuated" *Brucella* vaccine terms are included in a PubMed search, the number of papers found in PubMed increased by more than 10-fold [228, 229]. The application of VO has also enhanced the discovery of IFN-γ and vaccine-associated gene networks [16].

8. Discussion

In summary, vaccine informatics has been widely implemented in the areas of basic vaccine research, translational vaccine development, prolicensure vaccine immunization registry and surveillance, and vaccine data mining and integration.

Vaccine informatics is an emerging interdisciplinary research, with close relationships to several similar research fields. Vaccine informatics overlaps with immunological bioinformatics (or immunoinformatics). The latter field applies informatics technologies to investigate the immune system at a systems biology level [5, 230]. Vaccine informatics emphasizes understanding of vaccine-induced immunity. Vaccine informatics uses information of OMICS (genomics, transcriptomics, proteomics, and metabolomics). This is in contrast to reverse vaccinology that primarily uses genomics, that is, informatics analysis of genome sequences. Other OMICS technologies may also have the potential to aid in rational vaccine design. Recently Poland et al. defined a new area of vaccinomics that will focus on the development of personalized vaccines based on our increasing understanding of genotype information [231]. Vaccine informatics is also closely associated with clinical immunology in the areas of post-licensure vaccine assessment and surveillance. Mathematical modeling also plays an important role in vaccine informatics by modeling various aspects of pre- and post-licensure vaccine research and clinical investigations.

Vaccine informatics still faces many challenges. Many infectious diseases, including HIV/AIDS, tuberculosis, and malaria, still lack effective and safe vaccines. Although extensive progress has been made towards the genetic structure and pathogenesis of HIV and other infectious pathogens, significant gaps in our understanding of host-pathogen interactions still remain [232, 233]. These gaps are attributable to imperfect and nonstandardized animal models, the absence of precise immunological correlates of protection, and the prohibitive cost of confirmatory clinical trials. The development of vaccines against many noninfectious diseases including cancer, autoimmune diseases, and allergy remains a challenge. While many vaccine adverse events are likely genetically determined (and thus predictable), it remains challenging to predict possible vaccine adverse events with available genotype data and possibly design personalized vaccine. These challenges will undoubtedly be met with improved rational vaccine design and a better understanding of fundamental protective immunity mechanisms obtained with improving vaccine informatics technologies.

New bioinformatics technologies are constantly being devised and applied to address various vaccine-related questions using high throughput sequencing, gene expression data, and experimental results from experimental and clinical studies. Efforts during the 21st century vaccinology will witness more successes of application of vaccine informatics in vaccine research.

Acknowledgments

This work has been supported by grant R01AI081062 from the National Institute of Allergy and Infectious Diseases USA. The authors wish to acknowledge the assistance of John Glasser and Gary Urquhart as reviewers for the sections on mathematical modeling and immunization information systems, respectively. Editorial review by Dr. George W. Jourdian is also appreciated. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

References

[1] E. Jenner, *An Inquiry into the Causes and Effects of the Variolae Vaccinae*, Low, London, UK, 1798.

[2] M. R. Hilleman, "Vaccines in historic evolution and perspective: a narrative of vaccine discoveries," *Vaccine*, vol. 18, no. 15, pp. 1436–1447, 2000.

[3] J. Parrino and B. S. Graham, "Smallpox vaccines: past, present, and future," *Journal of Allergy and Clinical Immunology*, vol. 118, no. 6, pp. 1320–1326, 2006.

[4] E. N. T. Meeusen, J. Walker, A. Peters, P. P. Pastoret, and G. Jungersen, "Current status of veterinary vaccines," *Clinical Microbiology Reviews*, vol. 20, no. 3, pp. 489–510, 2007.

[5] A. S. De Groot, H. Sbai, C. S. Aubin, J. McMurry, and W. Martin, "Immuno-informatics: mining genomes for vaccine components," *Immunology and Cell Biology*, vol. 80, no. 3, pp. 255–269, 2002.

[6] B. Peters, J. Sidney, P. Bourne et al., "The immune epitope database and analysis resource: from vision to blueprint," *PLoS Biology*, vol. 3, no. 3, p. e91, 2005.

[7] R. D. Fleischmann, M. D. Adams, O. White et al., "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science*, vol. 269, no. 5223, pp. 496–521, 1995.

[8] R. Rappuoli, "Reverse vaccinology," *Current Opinion in Microbiology*, vol. 3, no. 5, pp. 445–450, 2000.

[9] A.S. De Groot, J. McMurry, L. Moise, and B. Martin, "Epitope-based Immunome-derived vaccines: a strategy for improved design and safety," in *Applications of Immunomics*, A. Falus, Ed., Springer Immunomics Series, Springer, New York, NY, USA, 2009.

[10] M. Pizza, V. Scarlato, V. Masignani et al., "Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing," *Science*, vol. 287, pp. 1816–1820, 2000.

[11] W. Zhou, V. Pool, J. K. Iskander et al., "Surveillance for safety after immunization: Vaccine Adverse Event Reporting System (VAERS)—United States, 1991–2001," *MMWR. Surveillance Summaries*, vol. 52, no. 1, pp. 1–24, 2003.

[12] P. Tamayo, D. Slonim, J. Mesirov et al., "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2907–2912, 1999.

[13] Z. Xiang, T. Todd, K. P. Ku et al., "VIOLIN: vaccine investigation and online information network," *Nucleic Acids Research*, vol. 36, no. 1, pp. D923–D928, 2008.

[14] Y. He, L. Cowell, A. D. Diehl et al., "VO: vaccine ontology," in *Proceedings of the 1st International Conference on Biomedical Ontology (ICBO '09)*, Buffalo, NY, USA, August 2009.

[15] R. R. Brinkman, M. Courtot, D. Derom et al., "Modeling biomedical experimental processes with OBI," *Journal of Biomedical Semantics*, vol. 1, supplement 1, p. S7, 2010.

[16] A. Ozgur, Z. Xiang, D. Radev, and Y. He, "Mining of vaccine associated IFN-γ gene interaction networks using the Vaccine Ontology," *Journal of Biomedical Semantics*. 2(Suppl 2):S8, 2011, http://www.jbiomedsem.com/qc/content/2/S2/S8.

[17] C. DeLisi and J. A. Berzofsky, "T-cell antigenic sites tend to be amphipathic structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 20, pp. 7048–7052, 1985.

[18] A. S. De Groot and J. A. Berzofsky, "From genome to vaccine—new immunoinformatics tools for vaccine design," *Methods*, vol. 34, no. 4, pp. 425–428, 2004.

[19] A. S. De Groot and L. Moise, "New tools, new approaches and new ideas for vaccine development," *Expert Review of Vaccines*, vol. 6, no. 2, pp. 125–127, 2007.

[20] J. D. Ahlers, I. M. Belyakov, E. K. Thomas, and J. A. Berzofsky, "High-affinity T helper epitope induces complementary helper and APC polarization, increased CTL, and protection against viral infection," *The Journal of Clinical Investigation*, vol. 108, no. 11, pp. 1677–1685, 2001.

[21] H. Inaba, W. Martin, A. S. De Groot, S. Qin, and L. J. De Groot, "Thyrotropin receptor epitopes and their relation to histocompatibility leukocyte antigen-DR molecules in graves' disease," *Journal of Clinical Endocrinology and Metabolism*, vol. 91, no. 6, pp. 2286–2294, 2006.

[22] A. S. De Groot, B. M. Jesdale, E. Szu, J. R. Schafer, R. M. Chicz, and G. Deocampo, "An interactive web site providing major histocompatibility ligand predictions: application to HIV research," *AIDS Research and Human Retroviruses*, vol. 13, no. 7, pp. 529–531, 1997.

[23] K. B. Bond, B. Sriwanthana, T. W. Hodge et al., "An HLA-directed molecular and bioinformatics approach identifies new HLA-A11 HIV-1 subtype E cytotoxic T lymphocyte epitopes in HIV-1-infected Thais," *AIDS Research and Human Retroviruses*, vol. 17, no. 8, pp. 703–717, 2001.

[24] J. McMurry, H. Sbai, M. L. Gennaro, E. J. Carter, W. Martin, and A. S. De Groot, "Analyzing Mycobacterium tuberculosis proteomes for candidate vaccine epitopes," *Tuberculosis*, vol. 85, no. 1-2, pp. 95–105, 2005.

[25] Y. Dong, S. Demaria, X. Sun et al., "HLA-A2-restricted CD8-cytotoxic-T-Cell responses to novel epitopes in mycobacterium tuberculosis superoxide dismutase, alanine dehydrogenase, and glutamine synthetase," *Infection and Immunity*, vol. 72, no. 4, pp. 2412–2415, 2004.

[26] O. A. Koita, D. Dabitao, I. Mahamadou et al., "Confirmation of immunogenic consensus sequence HIV-1 T-cell epitopes in Bamako, Mali and Providence, Rhode Island," *Human Vaccines*, vol. 2, no. 3, pp. 119–128, 2006.

[27] A. S. De Groot, "Immunomics: discovering new targets for vaccines and therapeutics," *Drug Discovery Today*, vol. 11, no. 5-6, pp. 203–209, 2006.

[28] J. A. McMurry, S. H. Gregory, L. Moise, D. Rivera, S. Buus, and A. S. De Groot, "Diversity of Francisella tularensis Schu4 antigens recognized by T lymphocytes after natural infections in humans: identification of candidate epitopes for inclusion in a rationally designed tularemia vaccine," *Vaccine*, vol. 25, no. 16, pp. 3179–3191, 2007.

[29] L. Moise, J. A. McMurry, S. Buus, S. Frey, W. D. Martin, and A. S. De Groot, "In silico-accelerated identification of conserved and immunogenic variola/vaccinia T-cell epitopes," *Vaccine*, vol. 27, no. 46, pp. 6471–6479, 2009.

[30] L. Moise, J. A. McMurry, J. Pappo et al., "Identification of genome-derived vaccine candidates conserved between human and mouse-adapted strains of H. pylori," *Human Vaccines*, vol. 4, no. 3, pp. 219–223, 2008.

[31] S. K. Kim, M. Cornberg, X. Z. Wang, H. D. Chen, L. K. Selin, and R. M. Welsh, "Private specificities of CD8 T cell responses control patterns of heterologous immunity," *Journal of Experimental Medicine*, vol. 201, no. 4, pp. 523–533, 2005.

[32] R. Rappuoli, "Bridging the knowledge gaps in vaccine design," *Nature Biotechnology*, vol. 25, no. 12, pp. 1361–1366, 2007.

[33] T. Elliott, V. Cerundolo, J. Elvin, and A. Townsend, "Peptide-induced conformational change of the class I heavy chain," *Nature*, vol. 351, no. 6325, pp. 402–406, 1991.

[34] K. Falk, O. Rotzschke, S. Stevanovic, G. Jung, and H. G. Rammensee, "Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules," *Nature*, vol. 351, no. 6324, pp. 290–296, 1991.

[35] O. Rotzschke, K. Falk, S. Stevanovic, G. Jung, P. Walden, and H. G. Rammensee, "Exact prediction of a natural T cell epitope," *European Journal of Immunology*, vol. 21, no. 11, pp. 2891–2894, 1991.

[36] E. R. Unanue, "Cellular studies on antigen presentation by class II MHC molecules," *Current Opinion in Immunology*, vol. 4, no. 1, pp. 63–69, 1992.

[37] J. H. Brown, T. S. Jardetzky, J. C. Gorga et al., "Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1," *Nature*, vol. 364, no. 6432, pp. 33–39, 1993.

[38] R. M. Chicz, R. G. Urban, J. C. Gorga, D. A. A. Vignali, W. S. Lane, and J. L. Strominger, "Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles," *Journal of Experimental Medicine*, vol. 178, no. 1, pp. 27–47, 1993.

[39] R. N. Germain and D. H. Margulies, "The biochemistry and cell biology and antigen processing and presentation," *Annual Review of Immunology*, vol. 11, pp. 403–450, 1993.

[40] P. Cresswell and A. Lanzavecchia, "Antigen processing and recognition," *Current Opinion in Immunology*, vol. 13, no. 1, pp. 11–12, 2001.

[41] E. S. Trombetta and I. Mellman, "Cell biology of antigen processing in vitro and in vivo," *Annual Review of Immunology*, vol. 23, pp. 975–1028, 2005.

[42] E. Appella, E. A. Padlan, and D. F. Hunt, "Analysis of the structure of naturally processed peptides bound by class I and class II major histocompatibility complex molecules," *EXS*, vol. 73, pp. 105–119, 1995.

[43] R. N. Germain, F. Castellino, R. Han et al., "Processing and presentation of endocytically acquired protein antigens by MHC class II and class I molecules," *Immunological Reviews*, no. 151, pp. 5–30, 1996.

[44] A. Sette and J. Sidney, "Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism," *Immunogenetics*, vol. 50, no. 3-4, pp. 201–212, 1999.

[45] A. S. De Groot, J. McMurry, and L. Moise, "Prediction of immunogenicity: in silico paradigms, ex vivo and in vivo correlates," *Current Opinion in Pharmacology*, vol. 8, no. 5, pp. 620–626, 2008.

[46] A. S. De Groot, D. S. Rivera, J. A. McMurry, S. Buus, and W. Martin, "Identification of immunogenic HLA-B7 "Achilles' heel" epitopes within highly conserved regions of HIV," *Vaccine*, vol. 26, no. 24, pp. 3059–3071, 2008.

[47] A. S. De Groot, E. A. Bishop, B. Khan et al., "Engineering immunogenic consensus T helper epitopes for a cross-clade HIV vaccine," *Methods*, vol. 34, no. 4, pp. 476–487, 2004.

[48] H. Plotnicky, D. Cyblat-Chanal, J. P. Aubry et al., "The immunodominant influenza matrix t cell epitope recognized in human induces influenza protection in HLA-A2/K transgenic mice," *Virology*, vol. 309, no. 2, pp. 320–329, 2003.

[49] P. Wang, J. Sidney, C. Dow, B. Mothé, A. Sette, and B. Peters, "A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach," *PLoS Computational Biology*, vol. 4, no. 4, Article ID e1000048, 2008.

[50] H. G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanović, "SYFPEITHI: database for MHC ligands and peptide motifs," *Immunogenetics*, vol. 50, no. 3-4, pp. 213–219, 1999.

[51] K. C. Parker, M. A. Bednarek, and J. E. Coligan, "Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains," *Journal of Immunology*, vol. 152, no. 1, pp. 163–175, 1994.

[52] A. S. De Groot and W. Martin, "Reducing risk, improving outcomes: bioengineering less immunogenic protein therapeutics," *Clinical Immunology*, vol. 131, no. 2, pp. 189–201, 2009.

[53] H. H. Lin, S. Ray, S. Tongchusak, E. L. Reinherz, and V. Brusic, "Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research," *BMC Immunology*, vol. 9, article 8, 2008.

[54] H. H. Lin, G. L. Zhang, S. Tongchusak, E. L. Reinherz, and V. Brusic, "Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research," *BMC Bioinformatics*, vol. 9, no. 12, article S22, 2008.

[55] U. Gowthaman and J. N. Agrewala, "In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion," *Journal of Proteome Research*, vol. 7, no. 1, pp. 154–163, 2008.

[56] P. Wang, J. Sidney, Y. Kim et al., "Peptide binding predictions for HLA DR, DP and DQ molecules," *BMC Bioinformatics*, vol. 11, 2010.

[57] S. Bulik, B. Peters, C. Ebeling, and H. Holzhütter, "Cytosolic processing of proteasomal cleavage products can enhance the presentation efficiency of MHC-1 epitopes," *Genome Informatics*, vol. 15, no. 1, pp. 24–34, 2004.

[58] B. Peters, S. Bulik, R. Tampe, P. M. Van Endert, and H. G. Holzhütter, "Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors," *Journal of Immunology*, vol. 171, no. 4, pp. 1741–1749, 2003.

[59] M. Nielsen, C. Lundegaard, O. Lund, and C. Keşmir, "The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage," *Immunogenetics*, vol. 57, no. 1-2, pp. 33–41, 2005.

[60] M. V. Larsen, C. Lundegaard, K. Lamberth, S. Buus, O. Lund, and M. Nielsen, "Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction," *BMC Bioinformatics*, vol. 8, article 424, 2007.

[61] S. Tenzer, B. Peters, S. Bulik et al., "Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding," *Cellular and Molecular Life Sciences*, vol. 62, no. 9, pp. 1025–1037, 2005.

[62] T. Stranzl, M. V. Larsen, C. Lundegaard, and M. Nielsen, "NetCTLpan: pan-specific MHC class I pathway epitope predictions," *Immunogenetics*, vol. 62, no. 6, pp. 357–368, 2010.

[63] A. B. Riemer, D. B. Keskin, G. Zhang et al., "A conserved E7-derived cytotoxic T lymphocyte epitope expressed on human papillomavirus 16-transformed HLA-A2+ epithelial cancers," *The Journal of Biological Chemistry*, vol. 285, pp. 29608–29622, 2010.

[64] D. Enshell-Seijffers, D. Denisov, B. Groisman et al., "The mapping and reconstitution of a conformational discontinuous B-cell epitope of HIV-1," *Journal of Molecular Biology*, vol. 334, no. 1, pp. 87–101, 2003.

[65] A. Schreiber, M. Humbert, A. Benz, and U. Dietrich, "3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins," *Journal of Computational Chemistry*, vol. 26, no. 9, pp. 879–887, 2005.

[66] U. Kulkarni-Kale, S. Bhosle, and A. S. Kolaskar, "CEP: a conformational epitope prediction server," *Nucleic Acids Research*, vol. 33, no. 2, pp. W168–W171, 2005.

[67] M. J. Sweredoski and P. Baldi, "PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure," *Bioinformatics*, vol. 24, no. 12, pp. 1459–1460, 2008.

[68] D. J. Barlow, M. S. Edwards, and J. M. Thornton, "Continuous and discontinuous protein antigenic determinants," *Nature*, vol. 322, no. 6081, pp. 747–748, 1986.

[69] M. J. Blythe and D. R. Flower, "Benchmarking B cell epitope prediction: underperformance of existing methods," *Protein Science*, vol. 14, no. 1, pp. 246–248, 2005.

[70] U. Reimer, "Prediction of linear B-cell epitopes," *Methods in Molecular Biology*, vol. 524, pp. 335–344, 2009.

[71] E. Rajnavolgyi, N. Nagy, B. Thuresson et al., "A repetitive sequence of Epstein-Barr virus nuclear antigen 6 comprises overlapping T cell epitopes which induce HLA-DR-restricted

CD4(+) T lymphocytes," *International Immunology*, vol. 12, no. 3, pp. 281–293, 2000.

[72] C. M. Graham, B. C. Barnett, I. Hartlmayr et al., "The structural requirements for class II (I-A(d))-restricted T cell recognition of influenza hemagglutinin: b cell epitopes define T cell epitopes," *European Journal of Immunology*, vol. 19, no. 3, pp. 523–528, 1989.

[73] W. Fischer, S. Perkins, J. Theiler et al., "Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants," *Nature Medicine*, vol. 13, no. 1, pp. 100–106, 2007.

[74] A. S. De Groot, L. Marcon, E. A. Bishop et al., "HIV vaccine development by computer assisted design: the GAIA vaccine," *Vaccine*, vol. 23, no. 17-18, pp. 2136–2148, 2005.

[75] B. Gaschen, J. Taylor, K. Yusim et al., "Diversity considerations in HIV-1 vaccine selection," *Science*, vol. 296, no. 5577, pp. 2354–2360, 2002.

[76] F. Gao, B. T. Korber, E. A. Weaver, H. X. Liao, B. H. Hahn, and B. F. Haynes, "Centralized immunogens as a vaccine strategy to overcome HIV-1 diversity," *Expert Review of Vaccines*, vol. 3, no. 4, pp. S161–S168, 2004.

[77] F. Gao, E. A. Weaver, Z. Lu et al., "Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group M consensus envelope glycoprotein," *Journal of Virology*, vol. 79, no. 2, pp. 1154–1163, 2005.

[78] D. C. Nickle, M. Rolland, M. A. Jensen et al., "Coping with viral diversity in HIV vaccine design," *PLoS Computational Biology*, vol. 3, no. 4, article e75, pp. 754–762, 2007.

[79] L. A. McNamara, Y. He, and Z. Yang, "Using epitope predictions to evaluate efficacy and population coverage of the Mtb72f vaccine for tuberculosis," *BMC Immunology*, vol. 11, article 18, 2010.

[80] J. Söllner, A. Heinzel, G. Summer et al., "Concept and application of a computational vaccinology workflow," *Immunome Research*, vol. 6, supplement 2, 2010.

[81] F. Pappalardo, M. D. Halling-Brown, N. Rapin et al., "ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design and optimization," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 330–340, 2009.

[82] M. Feldhahn, P. Dönnes, P. Thiel, and O. Kohlbacher, "FRED—a framework for T-cell epitope detection," *Bioinformatics*, vol. 25, no. 20, pp. 2758–2759, 2009.

[83] L. Moise, J. A. McMurry, J. Pappo et al., "Identification of genome-derived vaccine candidates conserved between human and mouse-adapted strains of H. pylori," *Human Vaccines*, vol. 4, no. 3, pp. 219–223, 2008.

[84] S. H. Gregory, S. Mott, J. Phung et al., "Epitope-based vaccination against pneumonic tularemia," *Vaccine*, vol. 27, no. 39, pp. 5299–5306, 2009.

[85] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg, "Improved microbial gene identification with GLIMMER," *Nucleic Acids Research*, vol. 27, no. 23, pp. 4636–4641, 1999.

[86] D. Frishman, A. Mironov, H. W. Mewes, and M. Gelfand, "Combining diverse evidence for gene recognition in completely sequenced bacterial genomes," *Nucleic Acids Research*, vol. 26, no. 12, pp. 2941–2947, 1998.

[87] J. Besemer, A. Lomsadze, and M. Borodovsky, "GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions," *Nucleic Acids Research*, vol. 29, no. 12, pp. 2607–2618, 2001.

[88] S. T. Fitz-Gibbon, H. Ladner, U. J. Kim, K. O. Stetter, M. I. Simon, and J. H. Miller, "Genome sequence of the hyperthermophilic crenarchaeon Pyrobaculum aerophilum," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 2, pp. 984–989, 2002.

[89] A. M. Cerdeño-Tárraga, A. Efstratiou, L. G. Dover et al., "The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129," *Nucleic Acids Research*, vol. 31, no. 22, pp. 6516–6523, 2003.

[90] J. Wei, M. B. Goldberg, V. Burland et al., "Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T," *Infection and Immunity*, vol. 71, no. 5, pp. 2775–2786, 2003.

[91] M. P. McLeod, X. Qin, S. E. Karpathy et al., "Complete genome sequence of Rickettsia typhi and comparison with sequences of other rickettsiae," *Journal of Bacteriology*, vol. 186, no. 17, pp. 5842–5855, 2004.

[92] P. Stothard and D. S. Wishart, "Automated bacterial genome analysis and annotation," *Current Opinion in Microbiology*, vol. 9, no. 5, pp. 505–510, 2006.

[93] A. Economou, P. J. Christie, R. C. Fernandez, T. Palmer, G. V. Plano, and A. P. Pugsley, "Secretion by numbers: protein traffic in prokaryotes," *Molecular Microbiology*, vol. 62, no. 2, pp. 308–319, 2006.

[94] T. T. Tseng, B. M. Tyler, and J. C. Setubal, "Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology," *BMC Microbiology*, vol. 9, no. 1, article S2, 2009.

[95] H. Tjalsma, A. Bolhuis, J. D. H. Jongbloed, S. Bron, and J. M. Van Dijl, "Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome," *Microbiology and Molecular Biology Reviews*, vol. 64, no. 3, pp. 515–547, 2000.

[96] H. Tjalsma, H. Antelmann, J. D. H. Jongbloed et al., "Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 2, pp. 207–233, 2004.

[97] J. L. Gardy and F. S. L. Brinkman, "Methods for predicting bacterial protein subcellular localization," *Nature Reviews Microbiology*, vol. 4, no. 10, pp. 741–751, 2006.

[98] C. S. Yu, C. J. Lin, and J. K. Hwang, "Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions," *Protein Science*, vol. 13, no. 5, pp. 1402–1406, 2004.

[99] Z. Lu, D. Szafron, R. Greiner et al., "Predicting subcellular localization of proteins using machine-learned classifiers," *Bioinformatics*, vol. 20, no. 4, pp. 547–556, 2004.

[100] A. P. Tampakaki, V. E. Fadouloglou, A. D. Gazi, N. J. Panopoulos, and M. Kokkinidis, "Conserved features of type III secretion," *Cellular Microbiology*, vol. 6, no. 9, pp. 805–816, 2004.

[101] P. D. Schloss and JO. Handelsman, "Status of the microbial census," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 4, pp. 686–691, 2004.

[102] F. M. Cohan and E. B. Perry, "A Systematics for Discovering the Fundamental Units of Bacterial Diversity," *Current Biology*, vol. 17, no. 10, pp. R373–R386, 2007.

[103] N. T. Perna, G. Plunkett, V. Burland et al., "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7," *Nature*, vol. 409, no. 6819, pp. 529–533, 2001.

[104] Y. Zhang, C. Laing, M. Steele et al., "Genome evolution in major Escherichia coli O157:H7 lineages," *BMC Genomics*, vol. 8, 2007.

[105] M. Brochet, E. Couvé, M. Zouine et al., "Genomic diversity and evolution within the species Streptococcus agalactiae," *Microbes and Infection*, vol. 8, no. 5, pp. 1227–1243, 2006.

[106] H. Tettelin, V. Masignani, M. J. Cieslewicz et al., "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 13950–13955, 2005.

[107] E. Brzuszkiewicz, H. Brüggemann, H. Liesegang et al., "How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic Escherichia coli strains," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 34, pp. 12879–12884, 2006.

[108] H. Tettelin, D. Medini, C. Donati, and V. Masignani, "Towards a universal group B Streptococcus vaccine using multistrain genome analysis," *Expert Review of Vaccines*, vol. 5, no. 5, pp. 687–694, 2006.

[109] D. Maione, I. Margarit, C. D. Rinaudo et al., "Immunology: identification of a universal group B Streptococcus vaccine by multiple genome screen," *Science*, vol. 309, no. 5731, pp. 148–150, 2005.

[110] Y. He, Z. Xiang, and H. L.T. Mobley, "Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development," *Journal of Biomedicine and Biotechnology*, vol. 2010, Article ID 297505, 2010 pages, 2010.

[111] H. Ochman and N. A. Moran, "Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis," *Science*, vol. 292, no. 5519, pp. 1096–1098, 2001.

[112] G. Sachdeva, K. Kumar, P. Jain, and S. Ramachandran, "SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks," *Bioinformatics*, vol. 21, no. 4, pp. 483–491, 2005.

[113] S. Vivona, F. Bernante, and F. Filippini, "NERVE: new enhanced reverse vaccinology environment," *BMC Biotechnology*, vol. 6, article 35, 2006.

[114] Y. He and Z. Xiang, "Bioinformatics analysis of Brucella vaccines and vaccine targets using VIOLIN," *Immunome Research*, vol. 6, supplement 1, p. S5, 2010.

[115] Z. Xiang and Y. He, "Procedia in vaccinology," in *Proceedings of the 2nd Global Congress on Vaccines*, vol. 1, pp. 23–29, Boston, Mass, USA, 2009.

[116] D. Serruto and R. Rappuoli, "Post-genomic vaccine development," *The FEBS Letters*, vol. 580, no. 12, pp. 2985–2992, 2006.

[117] M. Mora, C. Donati, D. Medini, A. Covacci, and R. Rappuoli, "Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach," *Current Opinion in Microbiology*, vol. 9, no. 5, pp. 532–536, 2006.

[118] E. L. Hendrickson, R. J. Lamont, and M. Hackett, "Tools for interpreting large-scale protein profiling in microbiology," *Journal of Dental Research*, vol. 87, no. 11, pp. 1004–1015, 2008.

[119] Y. Liang and A. Kelemen, "Associating phenotypes with molecular events: recent statistical advances and challenges underpinning microarray experiments," *Functional and Integrative Genomics*, vol. 6, no. 1, pp. 1–13, 2006.

[120] Affymetrix, *GeneChip Expression Data Analysis Fundamentals*, Affymetrix, Inc., Santa Clara, Calif, USA, 2004.

[121] R. A. Irizarry, B. Hobbs, F. Collin et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[122] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 1, pp. 31–36, 2001.

[123] R. C. Gentleman, V. J. Carey, D. M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, p. R80, 2004.

[124] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, no. 4, pp. 546–554, 2002.

[125] J. F. Ayroles and G. Gibson, "Analysis of variance of microarray data," *Methods in Enzymology*, vol. 411, pp. 214–233, 2006.

[126] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[127] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 3, 2004.

[128] S. Raychaudhuri, P. D. Sutphin, J. T. Chang, and R. B. Altman, "Basic microarray analysis: grouping and feature reduction," *Trends in Biotechnology*, vol. 19, no. 5, pp. 189–193, 2001.

[129] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[130] Z. Xiang, Z. S. Qin, and Y. He, "CRCView: a web server for analyzing and visualizing microarray gene expression data using model-based clustering," *Bioinformatics*, vol. 23, no. 14, pp. 1843–1845, 2007.

[131] D. A. Hosack, G. Dennis Jr., B. T. Sherman, H. C. Lane, and R. A. Lempicki, "Identifying biological themes within lists of genes with EASE," *Genome Biology*, vol. 4, no. 10, p. R70, 2003.

[132] A. Bild and P. G. Febbo, "Application of a priori established gene sets to discover biologically important differential expression in microarray data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15278–15279, 2005.

[133] D. A. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[134] T. Beißbarth and T. P. Speed, "GOstat: find statistically overrepresented Gene Ontologies with a group of genes," *Bioinformatics*, vol. 20, no. 9, pp. 1464–1465, 2004.

[135] Y. U. Xia, H. Yu, R. Jansen et al., "Analyzing cellular biochemistry in terms of molecular networks," *Annual Review of Biochemistry*, vol. 73, pp. 1051–1087, 2004.

[136] J. Goutsias and N. H. Lee, "Computational and experimental approaches for modeling gene regulatory networks," *Current Pharmaceutical Design*, vol. 13, no. 14, pp. 1415–1436, 2007.

[137] J. Hardin, A. Mitani, L. Hicks, and B. VanKoten, "A robust measure of correlation between two genes on a microarray," *BMC Bioinformatics*, vol. 8, Article 220, 2007.

[138] P. A. Morel, S. Ta'asan, B. F. Morel, D. E. Kirschner, and J. L. Flynn, "New insights into mathematical modeling of the

immune system," *Immunologic Research*, vol. 36, no. 1–3, pp. 157–165, 2006.

[139] G. Dreyfus, *Neural Networks : Methodology and Applications*, Springer, New York, NY, USA, 2005.

[140] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[141] Z. Xiang, R. M. Minter, X. Bi, P. J. Woolf, and Y. He, "miniTUBA: medical inference by network integration of temporal data using Bayesian analysis," *Bioinformatics*, vol. 23, no. 18, pp. 2423–2432, 2007.

[142] A. Stuart, J. K. Ord, and S. F. Arnold, *Kendall's Advanced Theory of Statistics*, Oxford University Press, New York, NY, USA, 2004.

[143] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to Bayesian network inference for generating causal networks from observational biological data," *Bioinformatics*, vol. 20, no. 18, pp. 3594–3603, 2004.

[144] N. Dhiman, R. Bonilla, D. O'Kane J, and G. A. Poland, "Gene expression microarrays: a 21st century tool for directed vaccine design," *Vaccine*, vol. 20, no. 1-2, pp. 22–30, 2001.

[145] T. Chen, "DNA microarrays—an armory for combating infectious diseases in the new century," *Infectious Disorders—Drug Targets*, vol. 6, no. 3, pp. 263–279, 2006.

[146] J. A. Young, Q. L. Fivelman, P. L. Blair et al., "The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification," *Molecular and Biochemical Parasitology*, vol. 143, no. 1, pp. 67–79, 2005.

[147] T. Sturniolo, E. Bono, J. Ding et al., "Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices," *Nature Biotechnology*, vol. 17, no. 6, pp. 555–561, 1999.

[148] T. D. Querec, R. S. Akondy, E. K. Lee et al., "Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans," *Nature Immunology*, vol. 10, no. 1, pp. 116–125, 2009.

[149] D. Gaucher, R. Therrien, N. Kettaf et al., "Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses," *Journal of Experimental Medicine*, vol. 205, no. 13, pp. 3119–3131, 2008.

[150] B. Pulendran, J. Miller, T. D. Querec et al., "Case of yellow fever vaccine-associated viscerotropic disease with prolonged viremia, robust adaptive immune responses, and polymorphisms in CCR5 and RANTES genes," *Journal of Infectious Diseases*, vol. 198, no. 4, pp. 500–507, 2008.

[151] B. A. McKinney, D. M. Reif, M. T. Rock et al., "Cytokine expression patterns associated with systemic adverse events following smallpox immunization," *Journal of Infectious Diseases*, vol. 194, no. 4, pp. 444–453, 2006.

[152] I. Hamaguchi, J.-i. Imai, H. Momose et al., "Two vaccine toxicity-related genes Agp and Hpx could prove useful for pertussis vaccine safety control," *Vaccine*, vol. 25, no. 17, pp. 3355–3364, 2007.

[153] T. Mizukami, J. I. Imai, I. Hamaguchi et al., "Application of DNA microarray technology to influenza A/Vietnam/1194/2004 (H5N1) vaccine safety evaluation," *Vaccine*, vol. 26, no. 18, pp. 2270–2283, 2008.

[154] V. Brusic and N. Petrovsky, "Immunoinformatics and its relevance to understanding human immune disease," *Expert Review of Clinical Immunology*, vol. 1, pp. 145–157, 2005.

[155] C. A.A. Beauchemin and A. Handel, "A review of mathematical models of influenza A infections within a host or cell culture: lessons learned and challenges ahead," *BMC Public Health*, vol. 11, supplement 1, p. S7, 2011.

[156] D. E. Kirschner, S. T. Chang, T. W. Riggs, N. Perry, and J. J. Linderman, "Toward a multiscale model of antigen presentation in immunity," *Immunological Reviews*, vol. 216, no. 1, pp. 93–118, 2007.

[157] M. A. Fishman and A. S. Perelson, "Th1/Th2 cross regulation," *Journal of Theoretical Biology*, vol. 170, no. 1, pp. 25–56, 1994.

[158] B. F. Keele, E. E. Giorgi, J. F. Salazar-Gonzalez et al., "Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 21, pp. 7552–7557, 2008.

[159] S. Marino, J. J. Linderman, and D. E. Kirschner, "A multifaceted approach to modeling the immune response in tuberculosis," *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, In Press.

[160] M. P. Davenport, R. M. Ribeiro, L. Zhang, D. P. Wilson, and A. S. Perelson, "Understanding the mechanisms and limitations of immune control of HIV," *Immunological Reviews*, vol. 216, no. 1, pp. 164–175, 2007.

[161] J. Schulze-Horsel, M. Schulze, G. Agalaridis, Y. Genzel, and U. Reichl, "Infection dynamics and virus-induced apoptosis in cell culture-based influenza vaccine production-Flow cytometry and mathematical modeling," *Vaccine*, vol. 27, no. 20, pp. 2712–2722, 2009.

[162] S. Y. Kim and S. J. Goldie, "Cost-effectiveness analyses of vaccination programmes: a focused review of modelling approaches," *PharmacoEconomics*, vol. 26, no. 3, pp. 191–215, 2008.

[163] S. J. Goldie, J. J. Kim, K. Kobus et al., "Cost-effectiveness of HPV 16, 18 vaccination in Brazil," *Vaccine*, vol. 25, no. 33, pp. 6257–6270, 2007.

[164] S. Aballéa, J. Chancellor, M. Martin et al., "The cost-effectiveness of influenza vaccination for people aged 50 to 64 years: an international model," *Value in Health*, vol. 10, no. 2, pp. 98–116, 2007.

[165] D. J. Isaacman, D. R. Strutton, E. A. Kalpas et al., "The impact of indirect (herd) protection on the cost-effectiveness of pneumococcal conjugate vaccine," *Clinical Therapeutics*, vol. 30, no. 2, pp. 341–357, 2008.

[166] A. Palladini, G. Nicoletti, F. Pappalardo et al., "In silico modeling and in vivo efficacy of cancer-preventive vaccinations," *Cancer Research*, vol. 70, pp. 7775–7763, 2010.

[167] M. Pennisi, F. Pappalardo, A. Palladini et al., "Modeling the competition between lung metastases and the immune system using agents," *BMC Bioinformatics*, vol. 11, supplement 7, article S13, 2010.

[168] F. Pappalardo, P. L. Lollini, F. Castiglione, and S. Motta, "Modeling and simulation of cancer immunoprevention vaccine," *Bioinformatics*, vol. 21, no. 12, pp. 2891–2897, 2005.

[169] P.-L. Lollini, S. Motta, and F. Pappalardo, "Discovery of cancer vaccination protocols with a genetic algorithm driving an agent based simulator," *BMC Bioinformatics*, vol. 7, article 352, 2006.

[170] F. Pappalardo, M. Pennisi, F. Castiglione, and S. Motta, "Vaccine protocols optimization: in silico experiences," *Biotechnology Advances*, vol. 28, no. 1, pp. 82–93, 2010.

[171] S. M. Blower, K. Koelle, D. E. Kirschner, and J. Mills, "Live attenuated HIV vaccines: predicting the tradeoff between efficacy and safety," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 6, pp. 3618–3623, 2001.

[172] T. J. John and R. Samuel, "Herd immunity and herd effect: new insights and definitions," *European Journal of Epidemiology*, vol. 16, no. 7, pp. 601–606, 2000.

[173] "Recommendations of the International Task Force for Disease Eradication," *MMWR—Recommendations and Reports*, vol. 42, no. 16, pp. 1–38, 1993.

[174] National Center for Immunization and Respiratory Diseases, "General recommendations on immunization—recommendations of the Advisory Committee on Immunization Practices (ACIP)," *MMWR—Recommendations and Reports*, vol. 60, no. 2, pp. 1–64, 2011.

[175] D. Wood, K. N. Saarlas, M. Inkelas, and B. T. Matyas, "Immunization registries in the United States: implications for the practice of public health in a changing health care system," *Annual Review of Public Health*, vol. 20, pp. 231–255, 1999.

[176] R. T. Chen, J. W. Glasser, P. H. Rhodes et al., "Vaccine safety datalink project: a new tool for improving vaccine safety monitoring in the United States," *Pediatrics*, vol. 99, no. 6, pp. 765–773, 1997.

[177] B. P. Hull, S. L. Deeks, and P. B. McIntyre, "The Australian Childhood Immunisation Register-A model for universal immunisation registers?" *Vaccine*, vol. 27, no. 37, pp. 5054–5060, 2009.

[178] "National standards for immunization coverage assessment: recommendations from the Canadian Immunization Registry Network," *Canada Communicable Disease Report*, vol. 31, pp. 93–97, 2005.

[179] C. Trewin, H. B. Strand, and E. K. Grøholt, "Norhealth: norwegian health information system," *Scandinavian Journal of Public Health*, vol. 36, no. 7, pp. 685–689, 2008.

[180] M. L. Popovich, J. J. Aramini, and M. Garcia, "Immunizations: the first step in a personal health record to empower patients," *Stud Health Technol Inform*, vol. 137, pp. 286–295, 2008.

[181] D. B. Fishbein, B. C. Willis, W. M. Cassidy et al., "Determining indications for adult vaccination: patient self-assessment, medical record, or both?" *Vaccine*, vol. 24, no. 6, pp. 803–818, 2006.

[182] J. A. Boom, A. C. Dragsbaek, and C. S. Nelson, "The success of an immunization information system in the wake of Hurricane Katrina," *Pediatrics*, vol. 119, no. 6, pp. 1213–1217, 2007.

[183] K. A. Feemster, C. V. Spain, M. Eberhart, S. Pati, and B. Watson, "Identifying infants at increased risk for late initiation of immunizations: maternal and provider characteristics," *Public Health Reports*, vol. 124, no. 1, pp. 42–53, 2009.

[184] F. Wei, J. P. Mullooly, M. Goodman et al., "Identification and characteristics of vaccine refusers," *BMC Pediatrics*, vol. 9, no. 1, article 18, 2009.

[185] V. L. Hinrichsen, B. Kruskal, M. A. O'Brien, T. A. Lieu, and R. Platt, "Using electronic medical records to enhance detection and reporting of vaccine adverse events," *Journal of the American Medical Informatics Association*, vol. 14, no. 6, pp. 731–735, 2007.

[186] A. R. Hinman, G. A. Urquhart, R. A. Strikas et al., "Immunization information systems: national vaccine advisory committee progress report, 2007," *Journal of Public Health Management and Practice*, vol. 13, no. 6, pp. 553–558, 2007.

[187] C. P. Farrington and E. Miller, "Vaccine trials," *Molecular Biotechnology*, vol. 17, no. 1, pp. 43–58, 2001.

[188] D. S. Fedson, "Measuring protection: efficacy versus effectiveness," *Developments in biological standardization*, vol. 95, pp. 195–201, 1998.

[189] R. T. Chen, S. C. Rastogi, J. R. Mullen et al., "The vaccine adverse event reporting system (VAERS)," *Vaccine*, vol. 12, no. 6, pp. 542–550, 1994.

[190] K. S. Lankinen, S. Pastila, T. Kilpi, H. Nohynek, P. H. Mäkelä, and P. Olin, "Vaccinovigilance in Europe—need for timeliness, standardization and resource," *Bulletin of the World Health Organization*, vol. 82, no. 11, pp. 828–835, 2004.

[191] D. Banks, E. J. Woo, D. R. Burwen, P. Perucci, M. M. Braun, and R. Ball, "Comparing data mining methods on the VAERS database," *Pharmacoepidemiology and Drug Safety*, vol. 14, no. 9, pp. 601–609, 2005.

[192] T. Verstraeten, F. DeStefano, R. T. Chen, and E. Miller, "Vaccine safety surveillance using large linked databases: opportunities, hazards and proposed guidelines," *Expert Review of Vaccines*, vol. 2, no. 1, pp. 21–29, 2003.

[193] J. Baggs, J. Gee, E. Lewis et al., "The Vaccine Safety Datalink: a model for monitoring immunization safety," *Pediatrics*, vol. 127, supplement 1, pp. S45–S53, 2011.

[194] P. Kramarz, E. K. France, F. Destefano et al., "Population-based study of rotavirus vaccination and intussusception," *Pediatric Infectious Disease Journal*, vol. 20, no. 4, pp. 410–416, 2001.

[195] W. W. Thompson, C. Price, B. Goodson et al., "Early thimerosal exposure and neuropsychological outcomes at 7 to 10 years," *The New England Journal of Medicine*, vol. 357, no. 13, pp. 1281–1292, 2007.

[196] F. DeStefano, T. Verstraeten, L. A. Jackson et al., "Vaccinations and risk of central nervous system demyelinating diseases in adults," *Archives of Neurology*, vol. 60, pp. 504–509, 2003.

[197] P. Farrington, S. Pugh, A. Colville et al., "A new method for active surveillance of adverse events from diphtheria/tetanus/pertussis and measles/mumps/rubella vaccines," *The Lancet*, vol. 345, no. 8949, pp. 567–569, 1995.

[198] M. Ali, C. G. Do, J. D. Clemens et al., "The use of a computerized database to monitor vaccine safety in Viet Nam," *Bulletin of the World Health Organization*, vol. 83, no. 8, pp. 604–610, 2005.

[199] D. Tapscott, *Wikinomics*, Penguin, New York, NY, USA, 2008.

[200] V. E. Pitzer, C. Viboud, L. Simonsen et al., "Demographic variability, vaccination, and the spatiotemporal dynamics of rotavirus epidemics," *Science*, vol. 325, no. 5938, pp. 290–294, 2009.

[201] C. G. Whitney, M. M. Farley, J. Hadler et al., "Decline in invasive pneumococcal disease after the introduction of protein-polysaccharide conjugate vaccine," *The New England Journal of Medicine*, vol. 348, no. 18, pp. 1737–1746, 2003.

[202] S. O. Sow, M. D. Tapia, S. Diallo et al., "Haemophilus influenzae type b conjugate vaccine introduction in Mali: impact on disease burden and serologic correlate of protection," *American Journal of Tropical Medicine and Hygiene*, vol. 80, no. 6, pp. 1033–1038, 2009.

[203] R. T. Chen, R. Weierbach, Z. Bisoffi et al., "A 'post-honeymoon period' measles outbreak in Muyinga sector, Burundi," *International Journal of Epidemiology*, vol. 23, no. 1, pp. 185–193, 1994.

[204] R. T. Chen, I. R. Hardy, P. H. Rhodes, D. K. Tyshchenko, A.V. Moiseeva, and V. F. Marievsky, "Ukraine, 1992: first assessment of diphtheria vaccine effectiveness during the recent resurgence of diphtheria in the Former Soviet Union," *Journal of Infectious Diseases*, vol. 181, supplement 1, pp. S178–S183, 2000.

[205] N. C. Grassly, J. Wenger, S. Durrani et al., "Protective efficacy of a monovalent oral type 1 poliovirus vaccine: a case-control study," *The Lancet*, vol. 369, no. 9570, pp. 1356–1362, 2007.

[206] W. H. McNeill, *Plagues and Peoples*, Blackwell Scientific Publications, Oxford, UK, 1976.

[207] M. A. Kermack, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A*, vol. 115, pp. 700–721, 1927.

[208] R. M. Anderson and R. M. May, "Directly transmitted infectious diseases: control by vaccination," *Science*, vol. 215, no. 4536, pp. 1053–1060, 1982.

[209] R. M. Anderson, R. M. May, and B. Anderson, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford, UK, 1991.

[210] K. Ellen and S. M. Cromley, *GIS and Public Health*, Guilford Press, New York, NY, USA, 2002.

[211] N. J. Gay, "The theory of measles elimination: implications for the design of elimination strategies," *Journal of Infectious Diseases*, vol. 189, no. 1, pp. S27–S35, 2004.

[212] H. R. Babad, D. J. Nokes, N. J. Gay, E. Miller, P. Morgan-Capner, and R. M. Anderson, "Predicting the impact of measles vaccination in England and Wales: model validation and analysis of policy options," *Epidemiology and Infection*, vol. 114, no. 2, pp. 319–344, 1995.

[213] J. M. Best, C. Castillo-Solorzano, J. S. Spika et al., "Reducing the global burden of congenital rubella syndrome: report of the World Health Organization Steering Committee on Research Related to Measles and Rubella Vaccines and Vaccination, June 2004," *Journal of Infectious Diseases*, vol. 192, no. 11, pp. 1890–1897, 2005.

[214] J. J. Kim, "Mathematical model of HPV provides insight into impacts of risk factors and vaccine," *PLoS Medicine*, vol. 3, no. 5, article e164, pp. 587–588, 2006.

[215] N. M. Ferguson, D. A. T. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke, "Strategies for mitigating an influenza pandemic," *Nature*, vol. 442, no. 7101, pp. 448–452, 2006.

[216] L. Temime, D. Guillemot, and P. Y. Boëlle, "Short- and long-term effects of pneumococcal conjugate vaccination of children on penicillin resistance," *Antimicrobial Agents and Chemotherapy*, vol. 48, no. 6, pp. 2206–2213, 2004.

[217] A. Zvi, N. Ariel, J. Fulkerson, J. C. Sadoff, and A. Shafferman, "Whole genome identification of Mycobacterium tuberculosis vaccine candidates by comprehensive data mining and bioinformatic analyses," *BMC Medical Genomics*, vol. 1, p. 18, 2008.

[218] A. Özgür, T. Vu, G. Erkan, and D. R. Radev, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network," *Bioinformatics*, vol. 24, no. 13, pp. i277–i285, 2008.

[219] A. Ozgur, D. R. Radev, Z. Xiang, and Y. He, "Literature-based discovery of IFN-γ and vaccine-mediated gene interaction networks," *Journal of Biomedicine and Biotechnology*, vol. 2010, Article ID 426479, p. 13, 2010.

[220] H. Streeck, N. Frahm, and B. D. Walker, "The role of IFN-γ Elispot assay in HIV vaccine research," *Nature Protocols*, vol. 4, no. 4, pp. 461–469, 2009.

[221] H. A. Fletcher, "Correlates of immune protection from tuberculosis," *Current Molecular Medicine*, vol. 7, no. 3, pp. 319–325, 2007.

[222] P. Mansueto, G. Vitale, G. Di Lorenzo, G. B. Rini, S. Mansueto, and E. Cillari, "Immunopathology of leishmaniasis: an update," *International Journal of Immunopathology and Pharmacology*, vol. 20, no. 3, pp. 435–445, 2007.

[223] Y. He, R. Vemulapalli, A. Zeytun, and G. G. Schurig, "Induction of specific cytotoxic lymphocytes in mice vaccinated with Brucella abortus RB51," *Infection and Immunity*, vol. 69, no. 9, pp. 5502–5508, 2001.

[224] H. M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS Biology*, vol. 2, no. 11, article e309, 2004.

[225] B. Yang, S. Sayers, Z. Xiang, and Y. He, "Protegen: a web-based protective antigen database and analysis system," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D1073–D1078, 2011.

[226] H. R. Ansari, D. R. Flower, and G. P. S. Raghava, "AntigenDB: an immunoinformatics database of pathogen antigens," *Nucleic Acids Research*, vol. 38, no. 1, Article ID gkp830, pp. D847–D853, 2009.

[227] Y. He, Z. Xiang, T. Todd et al., "Bio-Ontologies 2010: semantic applications in life sciences," in *Proceedings of the 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB '10)*, p. 4, Boston, Mass, USA, July 2010.

[228] Z. Xiang and Y. He, "Improvement of pubmed literature searching using biomedical ontology," in *Proceedings of the 1st International Conference on Biomedical Ontology (ICBO '09)*, Buffalo, NY, USA, July 2009.

[229] J. Hur, Z. Xiang, E. Feldman, and Y. He, "Bio-Ontologies 2010: semantic applications in life sciences," in *Proceedings of the 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB '10)*, Boston, Mass, USA, July 2010.

[230] O. Lund, M. Nielsen, C. Lundergaard, C. Kesmir, and S. Brunak, *Immunological Bioinformatics*, MIT Press, Cambridge, Mass, USA, 2005.

[231] G. A. Poland, I. G. Ovsyannikova, and R. M. Jacobson, "Personalized vaccines: the emerging field of vaccinomics," *Expert Opinion on Biological Therapy*, vol. 8, no. 11, pp. 1659–1667, 2008.

[232] M. Sullivan, "Moving candidate vaccines into development from research: lessons from HIV," *Immunology and Cell Biology*, vol. 87, no. 5, pp. 366–370, 2009.

[233] S. K. Pierce and L. H. Miller, "World Malaria Day 2009: what malaria knows about the immune system that immunologists still do not," *Journal of Immunology*, vol. 182, no. 9, pp. 5171–5177, 2009.

[234] M. Bhasin and G. P. S. Raghava, "A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes," *Journal of Biosciences*, vol. 32, no. 1, pp. 31–42, 2007.

[235] I. A. Doytchinova, P. Guan, and D. R. Flower, "EpiJen: a server for multistep T cell epitope prediction," *BMC Bioinformatics*, vol. 7, article 131, 2006.

[236] P. A. Reche, H. Zhang, J. P. Glutting, and E. L. Reinherz, "EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology," *Bioinformatics*, vol. 21, no. 9, pp. 2140–2141, 2005.

[237] A. S. De Groot, M. Ardito, E. M. McClaine, L. Moise, and W. D. Martin, "Immunoinformatic comparison of T-cell epitopes contained in novel swine-origin influenza A (H1N1) virus with epitopes in 2008-2009 conventional influenza vaccine," *Vaccine*, vol. 27, no. 42, pp. 5740–5747, 2009.

[238] N. Jojic, M. Reyes-Gomez, D. Heckerman, C. Kadie, and O. Schueler-Furman, "Learning MHC I—peptide binding," *Bioinformatics*, vol. 22, no. 14, pp. e227–e235, 2006.

[239] R. Vita, L. Zarebski, J. A. Greenbaum et al., "The immune epitope database 2.0," *Nucleic Acids Research*, vol. 38, no. 1, Article ID gkp1004, pp. D854–D862, 2009.

[240] L. Jacob and J. P. Vert, "Efficient peptide-MHC-I binding prediction for alleles with few known binders," *Bioinformatics*, vol. 24, no. 3, pp. 358–366, 2008.

[241] M. Bhasin and G. P. S. Raghava, "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence," *Bioinformatics*, vol. 20, no. 3, pp. 421–423, 2004.

[242] P. Guan, I. A. Doytchinova, C. Zygouri, and D. R. Flower, "MHCPred: bringing a quantitative dimension to the online prediction of MHC binding," *Applied Bioinformatics*, vol. 2, no. 1, pp. 63–66, 2003.

[243] G. L. Zhang, A. M. Khan, K. N. Srinivasan, J. T. August, and V. Brusic, "MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides," *Nucleic Acids Research*, vol. 33, no. 2, pp. W172–W179, 2005.

[244] S. Buus, S. L. Lauemøller, P. Worning et al., "Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach," *Tissue Antigens*, vol. 62, no. 5, pp. 378–384, 2003.

[245] M. Nielsen and O. Lund, "NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction," *BMC Bioinformatics*, vol. 10, article 1471, p. 296, 2009.

[246] I. Hoof, B. Peters, J. Sidney et al., "NetMHCpan, a method for MHC class i binding prediction beyond humans," *Immunogenetics*, vol. 61, no. 1, pp. 1–13, 2009.

[247] M. Nielsen, C. Lundegaard, T. Blicher et al., "Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan," *PLoS Computational Biology*, vol. 4, no. 7, Article ID e1000107, 2008.

[248] P. A. Reche and E. L. Reinherz, "PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands," *Nucleic Acids Research*, vol. 33, no. 2, pp. W138–W142, 2005.

[249] O. Schueler-Furman, Y. Altuvia, A. Sette, and H. Margalit, "Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles," *Protein Science*, vol. 9, no. 9, pp. 1838–1846, 2000.

[250] C. W. Tung and S. Y. Ho, "POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties," *Bioinformatics*, vol. 23, no. 8, pp. 942–949, 2007.

[251] H. Singh and G. P. S. Raghava, "ProPred1: prediction of promiscuous MHC class-I binding sites," *Bioinformatics*, vol. 19, no. 8, pp. 1009–1014, 2003.

[252] H. Singh and G. P. S. Raghava, "ProPred: prediction of HLA-DR binding sites," *Bioinformatics*, vol. 17, no. 12, pp. 1236–1237, 2002.

[253] P. A. Reche, J. P. Glutting, H. Zhang, and E. L. Reinherz, "Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles," *Immunogenetics*, vol. 56, no. 6, pp. 405–419, 2004.

[254] P. Dönnes and A. Elofsson, "Prediction of MHC class I binding peptides, using SVMHC," *BMC Bioinformatics*, vol. 3, article 25, 2002.

[255] W. Liu, J. Wan, X. Meng, D. R. Flower, and T. Li, "In silico prediction of peptide-MHC binding affinity using SVRMHC," *Methods in Molecular Biology*, vol. 409, pp. 283–291, 2007.

[256] H. Bian and J. Hammer, "Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE," *Methods*, vol. 34, no. 4, pp. 468–475, 2004.

[257] H. U. Chen, NI. Huang, and Z. Sun, "SubLoc: a server/client suite for protein subcellular location based on SOAP," *Bioinformatics*, vol. 22, no. 3, pp. 376–377, 2006.

[258] M. Bhasin, A. Garg, and G. P. S. Raghava, "PSLpred: prediction of subcellular localization of bacterial proteins," *Bioinformatics*, vol. 21, no. 10, pp. 2522–2524, 2005.

[259] R. Nair and B. Rost, "Mimicking cellular sorting improves prediction of subcellular localization," *Journal of Molecular Biology*, vol. 348, no. 1, pp. 85–100, 2005.

[260] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, "Locating proteins in the cell using TargetP, SignalP and related tools," *Nature Protocols*, vol. 2, no. 4, pp. 953–971, 2007.

[261] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[262] F. Chen, A. J. Mackey, C. J. Stoeckert, and D. S. Roos, "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups," *Nucleic Acids Research*, vol. 34, pp. D363–368, 2006.

[263] G. E. Tusnády and I. Simon, "The HMMTOP transmembrane topology prediction server," *Bioinformatics*, vol. 17, no. 9, pp. 849–850, 2001.

[264] P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas, "PRED-TMBB: a web server for predicting the topology of β-barrel outer membrane proteins," *Nucleic Acids Research*, vol. 32, pp. W400–W404, 2004.

[265] N. K. Natt, H. Kaur, and G. P. S. Raghava, "Prediction of transmembrane regions of β-barrel proteins using ANN- and SVM-based methods," *Proteins*, vol. 56, no. 1, pp. 11–18, 2004.

[266] H. R. Bigelow, D. S. Petrey, J. Liu, D. Przybylski, and B. Rost, "Predicting transmembrane beta-barrels in proteomes," *Nucleic Acids Research*, vol. 32, no. 8, pp. 2566–2577, 2004.