ELSEVIER

Understanding infectious agents from an *in silico* perspective

Joo Chuan Tong[1,2] and Lisa F.P. Ng[2,3]

[1] Data Mining Department, Institute for Infocomm Research, 1 Fusionopolis Way, 21-01 Connexis South Tower, Singapore 138632, Singapore
[2] Department of Biochemistry, Yong Loo School of Medicine, National University of Singapore, Singapore 117597, Singapore
[3] Singapore Immunology Network, 8A Biomedical Grove, 04-06, Immunos, Singapore 138648, Singapore

Knowledge of infectious diseases now emerging from genomic, proteomic, epidemiological and clinical data can provide insights into the mechanisms of immune function, disease pathogenesis and epidemiology. Here, we describe how considerable advances in computational methods of data mining, mathematical modeling in epidemiology and simulation have been used to enhance our understanding of infectious agents and discuss their impact on the discovery of new therapeutics and controlling their spread.

Introduction

Epidemics, pandemics and outbreaks of infectious diseases have occurred throughout human history. In ancient times, the Athenian plague of 430–427 BC reportedly killed up to one-half of the population of Athens, the Justinian plague of 541–542 AD resulted in more than 100 million deaths, and the Black Death between 1348 and 1350 accounted for more than 100 million deaths. Worldwide, changes in socioeconomic, demographic and environmental factors have led to the resurgence of old and new infectious diseases. Over the past few decades, the world has witnessed not only the increasing problem of drug-resistant pathogens in diseases such as malaria and tuberculosis but also the emergence of new pathogens. These include the rotavirus in 1973, human immunodeficiency virus (HIV) in 1981, hepatitis C virus in 1989, hantavirus in 1993 and the severe acute respiratory syndrome coronavirus (SARS-CoV) in 2002. The re-emergence of epidemic chikungunya virus (CHIKV), previously known to be a benign disease, in Africa, the Indian Ocean, South-East Asia and the Pacific in the past decade has caused severe morbidity with some fatalities. More recently, in April 2009, the triple reassortant influenza A (H1N1) viruses, which contain genes from human, swine and avian influenza A viruses, appeared and have spread to more than 212 countries and overseas territories or communities, causing more than 15,921 deaths over the course of one year.

Corresponding author:. Tong, J.C. (victor@bic.nus.edu.sg)

From the earliest times, human has striven to understand the behaviors of infectious organisms and the mechanisms governing disease transmission. This goal has profoundly shaped modern knowledge of emerging and re-emerging infections. More recently, computational techniques have led the way to a new era by enabling rapid large-scale analyses of pathogens that were not possible using traditional experimental techniques. Here, we survey methods in mathematical modeling in epidemiology, computational biology and bioinformatics that have been used to study infectious diseases and discuss how these works have been translated into benefits for humankind, particularly in molecular epidemiology and in the design of novel therapeutics.

Mathematical models for understanding disease epidemiology

Mathematical models are now routinely used for studying the spread and control of infectious diseases. The history of mathematical epidemiology could be traced to 1760, when Daniel Bernoulli formulated a model to evaluate the effectiveness of variolation of healthy people with the smallpox virus [1]. It was not until the start of the 20th century, however, that mathematical models were applied to the study of the transmission patterns of infectious diseases. They were first used to understand the recurrence of measle epidemics [2] and the incidence and control of malaria [3]. Since then, epidemiology modeling has grown rapidly, fueled by the advent of specialized databases (Table 1) focusing on pathogens and their genes [4]. Some of these methods

1359-6446/06/$ - see front matter © 2010 Published by Elsevier Ltd. doi:10.1016/j.drudis.2010.10.007

TABLE 1

Bioinformatic resource centers for infectious disease research

Resource	Description	Web URL
Immune Epitope Database	Comprehensive repository of MHC-binding peptides, T-cell epitope and B-cell epitope data.	http://www.immuneepitope.org/
The International ImMunoGeneTics information system (IMGT)	Highly integrated resource for sequence, structural and genetic information on immune regulators across multiple species.	http://www.imgt.cines.fr/
The Innate Immune Database (IIDB)	Resource for facilitating gene-specific and systems biology oriented research. Enables integrative analysis of individual immune-active genes or the entire genomic locus.	http://www.db.systemsbiology.net/IIDB
Immunological Database and Analysis Portal (ImmPort)	Portal for accessing references and experiment data for immunologists. Supports production, analysis, archiving and exchange of scientific data.	https://www.immport.org
SYFPEITHI	Database of experimentally verified MHC-binding peptides.	http://www.syfpeithi.de/
MHCBN	Extensive repository of MHC-binding and non-binding peptides.	http://www.imtech.res.in/raghava/mhcbn/
AntiJen	Database containing quantitative binding data for peptides binding to MHC peptides, T-cell epitopes, transporter associated with antigen processing (TAP), B-cell epitopes and protein–protein interactions.	http://www.darrenflower.info/antijen/
Bcipep	Extensive repository of B-cell epitopes.	http://www.imtech.res.in/raghava/bcipep/
AntigenDB	Comprehensive information about a wide range of experimentally-validated antigens cross-linked to epitope data.	http://www.imtech.res.in/raghava/antigendb/
HIV Molecular Immunology Database	HIV-1 cytotoxic and helper T-cell epitopes and antibody-binding sites.	http://www.hiv.lanl.gov/content/immunology/

had been incorporated into successful environmental management programs [5], some in the development of intervention measures and containment strategies [6], some in the design of new therapeutic agents [7], and others in the planning of experiments and hypotheses testing [8].

Quantifying disease in an emerging epidemic

Several statistical measures are used frequently to quantify disease in populations and to facilitate disease management (Table 2). One such measure is the incidence of disease, defining the rate at which new cases occur in a population at risk during a specified time period. Because in reality the population at risk is constantly changing because of births, deaths and migrations, this measure might not be a good reflection of the true incidence of disease. One way to solve this is to relate the number of new cases to the person years at risk, which is computed by summing the time during which each individual member of the population is at risk during the measurement period. Another commonly used measure is the prevalence of disease, which is the proportion of the infected in a population at a given point of time. This is often used as an alternative to incidence in cases in which the sample size is small. For diseases with high mortality rates, researchers will be interested in calculating the case fatality rate (CFR) [9]. This is the proportion of infected patients dying from a certain disease during a specified period. For instance, an epidemiological investigation on 1425 patients with SARS-CoV infection in Hong Kong reported up to 28 April 2003 estimated that the CFR was 13.2% for patients younger than 60 years and 43.4% for patients aged 60 years or older [10]. It is used to link mortality to morbidity and can help to measure various aspects or properties of a disease, such as its pathogenicity, severity or virulence. Another important factor that researchers and clinicians are interested in is the clinical attack rate, which is the proportion of infected patients with symptoms of the disease [9]. It was estimated that the median clinical attack rate for the 1889 Russian flu pandemic was 60%, and the CFR ranged from 0.1% to 0.28% [9]. Other statistical measures exist and have been reviewed elsewhere [11].

TABLE 2

Some commonly used measures for quantifying disease in population

Measure	Formulas
Incidence of disease	Occurrence of new cases within a population at risk
	Specified period of time
Prevalence of disease	Number of infected people within a population
	Specified point of time
Case fatality rate	Number of deaths within a population with a particular condition
	Specified period of time
Clinical attack rate	Number of infected people with symptoms of disease
	Total number of infected people
Relative risk	Probability of event occurring in exposed group
	Probability of event occurring in a non-exposed group

Modeling the spatial spread of pathogens

Understanding how a disease is transmitted from one person to the next or spread in a population is important for assessing the risk of infection, for contact tracing and for building a contingency plan to contain the outbreak. Air travel can play an important part in the spread of infectious disease. The transmissions of *Mycobacterium tuberculosis*, SARS-CoV and influenza A virus within the

confined space of airplane cabins have been well documented [12]. Computational analysis on inter-regional influenza spread in the USA from 1996 to 2005 has shown that international air travel might affect the onset of disease outbreak [13]. Crépey and Barthélemy [14] have studied the epidemic patterns of seasonal influenza outbreaks in the USA from 1972 to 2002. The team concluded that air travel flow during this period of time was sufficient to propagate an epidemic throughout the USA and suggested that realistic modeling of the spread of epidemics at the interstate level might only need to take air transportation into account. Yet history has shown that air travel restrictions would have a limited, if any, effect on an outbreak. The 1889 Russian flu pandemic had spread across all of Europe and the USA within four months, despite slower surface travel and much smaller traveler flows. Mathematical modeling has now made it possible for us to compute the speed of disease transmission, as shown recently by Valleron *et al.* [9], which was estimated at the rate of 394 (±255) km/week in Europe and 1015 (±727) km/week in the USA. Several computer simulation experiments on the spread of pandemic influenza have also concluded that air travel restrictions would have limited impact compared with local control measures [15]. An explanation for this might be that the speed of a pandemic is not dependent on the passenger flows between cities but the degree of connectedness of a city network [15]. Attempts are also being made to analyze how a disease spreads within an urban environment [16], quantify the relative importance of different modes of transportation on the regional spread of influenza epidemics [15], and evaluate intervention strategies such as isolation, household quarantine, school or workplace closure, travel restrictions and mass screening at key control nodes including sea ports and airports [17,18].

Analyzing genetic variability

Similarity between related sequences can give clues to the structure, function or homology to the common ancestor [19]. Computational methods that can align sequence features, therefore, are particularly useful. Multiple sequence alignment methods organize the sequences of DNA, RNA or protein to identify regions of similarity that might help explain functional and/or phenotypic variability. The 2009 H1N1 flu was not the first human pandemic caused by influenza A viruses. It shares common ancestry with the 1889 Russian flu that killed approximately 1 million people, the 1918 Spanish flu that reportedly affected approximately 25% of the global population and killed at least 50 million people worldwide, the 1957 Asian flu that resulted in approximately 2 million deaths, and the 1968 Hong Kong flu that caused approximately 1 million deaths. In cases in which ancestry is unclear, multiple sequence alignment methods have been useful for inferring their phylogenetic relationships (Fig. 1). This includes not only identifying globally optimal alignment solutions for studying highly conserved sequences but also identifying maximally homologous subsequences among sets of long sequences for detecting distantly related proteins. By applying phylogenetic analysis to rapidly evolving viruses such as HIV, Bhattacharya *et al.* [20] have shown that viral escape effects instead of immune escape often explain

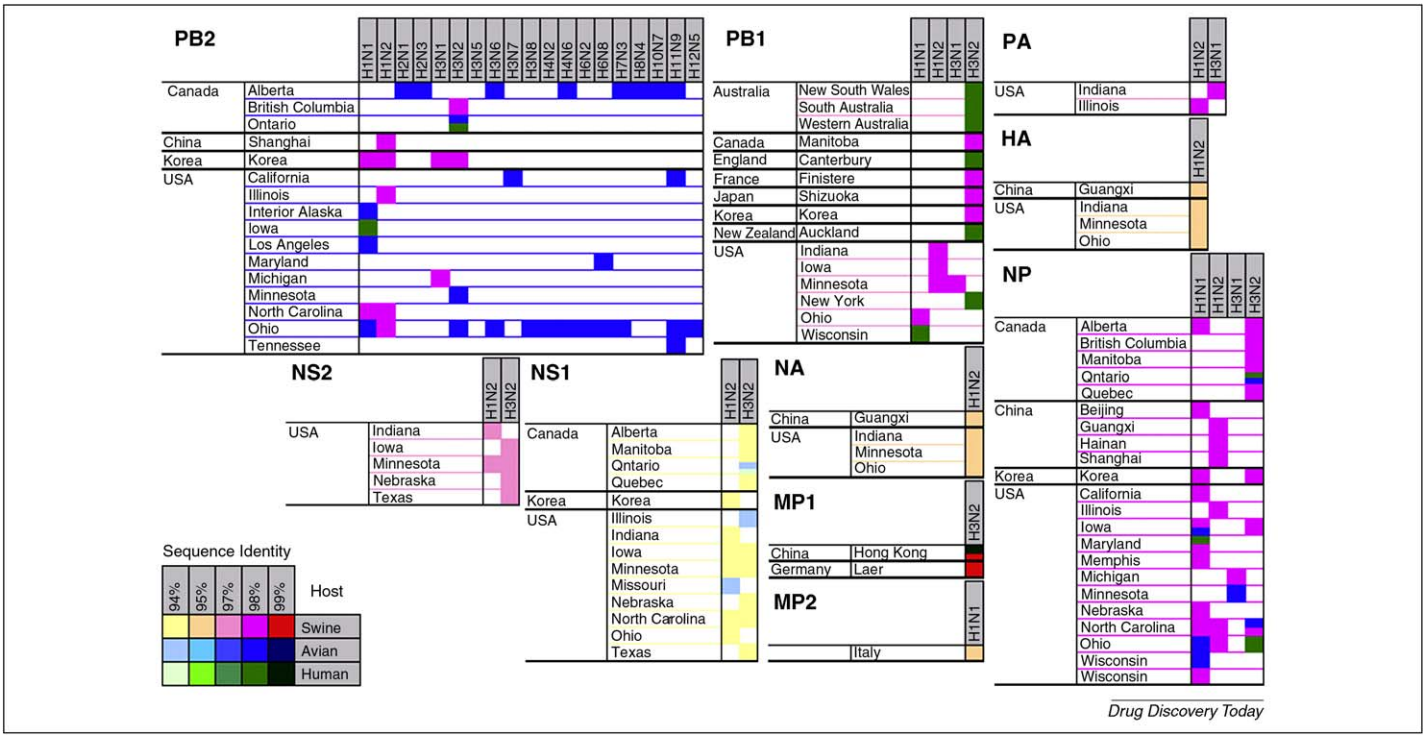

FIGURE 1

Sequence conservation of the 2009 influenza A (H1N1) virus. Influenza A is an enveloped virus that contains eight segments of negative-stranded RNA genome, encoding for 11 proteins: hemagglutinin (HA), nucleocapsid protein (NP), neuraminidase (NA), matrix protein (M), non-structural protein (NS) 1, NS2, polymerase A protein (PA), polymerase basic protein (PB) 1, PB1-F2 and PB2. When two influenza viruses co-infect the same cell, they could swap genes and produce new offspring lineages that contain segments from both parental strains in a process known as reassortment. Here, the sequence homology between proteins of the 2009 triple reassortant influenza A (H1N1) virus, which contain gene segments from human, swine and avian influenza A viruses, and their closest ancestors are shown. Multiple sequence alignment was performed using ClustalX, on 41 012 non-redundant influenza A sequences extracted from GenBank and SwissProt.

apparent human leukocyte antigen (HLA)-mediated immune-escape mutations defined by older analysis methods.

A related concept is the use of information theory to quantify variability in the pathogens' sequences. Theoretical statistics, such as information entropy [21], measure the rate of information transfer in biological sequences. We used this method recently to analyze CHIKV proteomes from its introduction in 1952–2009, and the results indicate that large amounts of 'antigenic switches' (i.e. changes in gene expression at a specific site, which might abrogate binding to HLA molecules or interfere with T-cell response, leading to cellular immune evasion) were clustered over the CHIKV genome [22]. There are also several attempts to identify amino acid residues that are likely to be involved in virus adaptation, for example, by finding interdependencies between mutations in multiple proteins. Miotto *et al.* [23] applied mutual information, an information theoretical statistic that measures the strength of association between a pair of variables, to identify characteristic sites in influenza A proteins where human isolates present conserved mutations. A catalogue of 68 characteristic sites in eight internal proteins was created from 92,343 sequences and used to derive adaptive signatures of influenza A proteomes.

A novel development in sequence analysis that has evolved recently is spatio-temporal analysis of infectious disease evolution. Typically, sequence analysis is performed on a collection of biological sequences that are assumed to share an evolutionary relationship. However, the flood of publicly available and newly generated pathogen sequences annotated with host organism, time of isolation and country of origin have opened up the possibility of incorporating these parameters in the study of microbial pathogen evolution. Sheng *et al.* [24] have recently shown how geographical and time information in public biological databases could be integrated with pattern-mining algorithms to study antigenic changes in influenza A viruses. The likelihood of one virus being able to mutate into another form is dependent on whether they exist within a certain time period, the connectivity between the locations where they were collected and their sequence similarity. The method was used to trace the evolution trajectory of H5N1, H1N1 and H3N2 subtypes in Asia, the USA and Europe at different time points. Attempts to unify the epidemiological and evolutionary processes that drive spatio-temporal incidence and phylogenetic patterns at different scales have also been reported. For instance, Grenfell *et al.* [25] introduced a phylodynamic framework to study how pathogen genetic variation – modulated by host immunity, transmission bottlenecks and epidemic dynamics – affects the diversity of epidemiological and phylogenetic patterns in measles, influenza A viruses, HIV, dengue and hepatitis C virus.

Detecting natural selection in molecular evolution

The epidemic behavior of the pathogen could be qualitatively examined by analyzing evolutionary inertia within a focal population [26]. One approach is to estimate natural selection for nucleotide usage at single amino acid sites [27,28]. Because of degeneracy of the genetic code, some point mutations are silent with no amino acid replacements. The neutral theory of molecular evolution, first introduced by Kimura in 1968 [29], states that most nucleotide substitutions are selectively neutral and are fixed by random genetic drift. Because synonymous (silent) substitutions are primarily transparent to natural selection, whereas non-synonymous (replacement) substitutions might be due to strong selective pressure, comparing the fixation rates between non-synonymous (d_N) and synonymous (d_S) substitutions can help to assess the extent of adaptive evolution at highly variable genetic loci [30].

The d_N:d_S ratio (ω), otherwise known as the 'acceptance rate', provides a sensitive measure of selection pressure at the amino acid level [31]. $\omega = 1$ indicates neutral expectation, $\omega < 1$ suggests negative (purifying) selection, and $\omega > 1$ suggests positive (diversifying) selection. A group of genes that often show the $\omega > 1$ relationship are antigenic genes in HIV-1, plasmodia and other parasites [30]. Using a 'relaxed-clock' phylogenetic model to estimate absolute rates of synonymous and non-synonymous substitution through time, Lemey *et al.* [32] showed that disease progression among patients is predicted by synonymous substitution rates, whereas non-synonymous rates evolve within patients as a consequence of changing antibody selective pressure. The hemagglutinin gene from the influenza A virus is probably one of the fastest evolving genes in terms of the rate of nucleotide substitution, which was estimated at 5.7×10^{-3} per site per year [33]. These genes are highly variable so as to enhance the pathogen's ability to evade host defenses.

The simple counting method of Nei and Gojobori [34] is often used for estimating d_N and d_S; however, the reliability of this approach is low when the rate of transitional nucleotide change is higher than that of transversional change. Model-based maximum likelihood methods, such as those proposed by Muse and Gaut [35] and Goldman and Yang [36], are a viable and widely used alternative for this purpose. The original maximum likelihood model of Goldman and Yang [36] assumes a single ω for all lineages and sites and has been extended to account for variation by allowing ω to vary across lineages [37], among substitution sites [35] or both among sites and among lineages [38]. Lineage-specific models assume that ω values do not vary among sites and can detect positive selection for a lineage only if the averaged d_N over all sites is greater than the average d_S. Site-specific models, conversely, allow ω to vary among sites but not among lineages. As such, these models can detect positive selection at individual sites only if the averaged d_N over all lineages is greater than the average d_S. By allowing ω to vary both among sites and among lineages, the extended Goldman and Yang model could be applied to detecting positive selection that occurred at multiple time points and affects multiple sites.

Deciphering host–pathogen interactions for therapeutic designs

Pathogenesis is a multi-step process in which there is continuous cross-talk between invading pathogens and their human host [39]. The ability of invasive parasites to infiltrate the mammalian host requires cell surface contact with host target molecules. Such interaction can take place through specific receptor-mediated mechanisms that could lead to the lysis of host target cells or substrates. The immune system comes into play once a pathogen infiltrates and infects the host using a layered defense mechanism of increasing specificity. This can take the form of innate and/or adaptive immunity through an array of immune receptors including mast cells, phagocytes, basophils, natural killer cells, and T

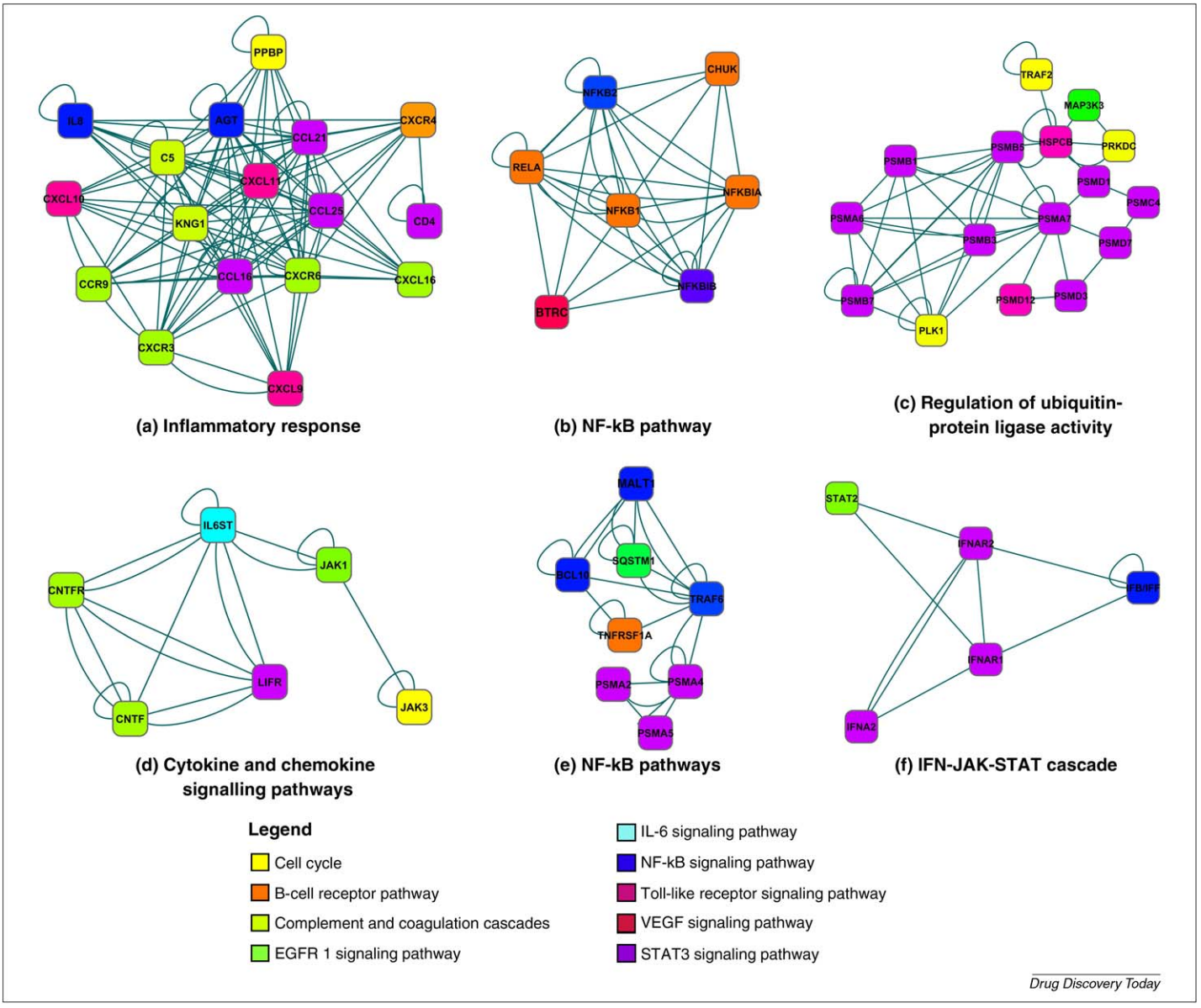

FIGURE 2

Example of PPI sub-networks that might be activated during CHIKV infection, predicted using a support vector machine model. The support vector machine system was trained using 2075 genes and 12,822 PPIs derived from BIOgrid, KEGG, Netpath, MINT, DIP, InAct, Reactome, Ambion and SABiosciences. Cluster A is implicated in the inflammatory response, clusters B and E in the NF-κB pathway, cluster C in the regulation of ubiquitin–protein ligase activity, cluster D in the cytokine and chemokine signaling pathway, and cluster F in the JAK-STAT cascade. Mediator factors linking these clusters include IL8 (in cluster A); NF-κBIA (in cluster B); TRAF2 (in cluster C); IL6ST, JAK1 and JAK3 (in cluster D); TRAF6 (in cluster D); and IFNα and STAT2 (in cluster F).

cells and B cells. Computational methods, such as support vector machines, that predict protein–protein interactions (PPIs) and PPI networks, therefore, are useful not only for identifying effector proteins associated with host infection but also for discovering host immune molecules that are involved in the clearance of pathogen (Fig. 2).

The identification of novel interactions between *Plasmodium falciparum* and human proteins was recently reported by Dyer *et al.* [40]. *P. falciparum* is one of four distinct species of the malaria parasite that infect humans. The parasite is responsible for at least 250 million infections and 1 million deaths each year. Using Bayesian statistics, the team identified 516 *P. falciparum*–human protein interactions and several functionally enriched sub-networks that could serve as the starting point for therapeutic development. The role of NF-κB in regulating gene expression has been analyzed by Shelest et al. [41] using *Pseudomonas aeruginosa* as a model organism. 13,000 genes were screened using weighted matrices, and 135 potential new target genes were identified. Attempts have also been made to simulate the human immune system at the system level. A good example of this is the EU-funded ImmunoGrid project, which aims to develop a virtual human immune system that reflects both the diversity and the relative proportions of its constituent molecules and cells [42].

Computer-aided drug design

Antiviral drugs could, in essence, target either cellular proteins or viral proteins [43]. Drugs that target cellular proteins could be active against a spectrum of unrelated viruses because many of the

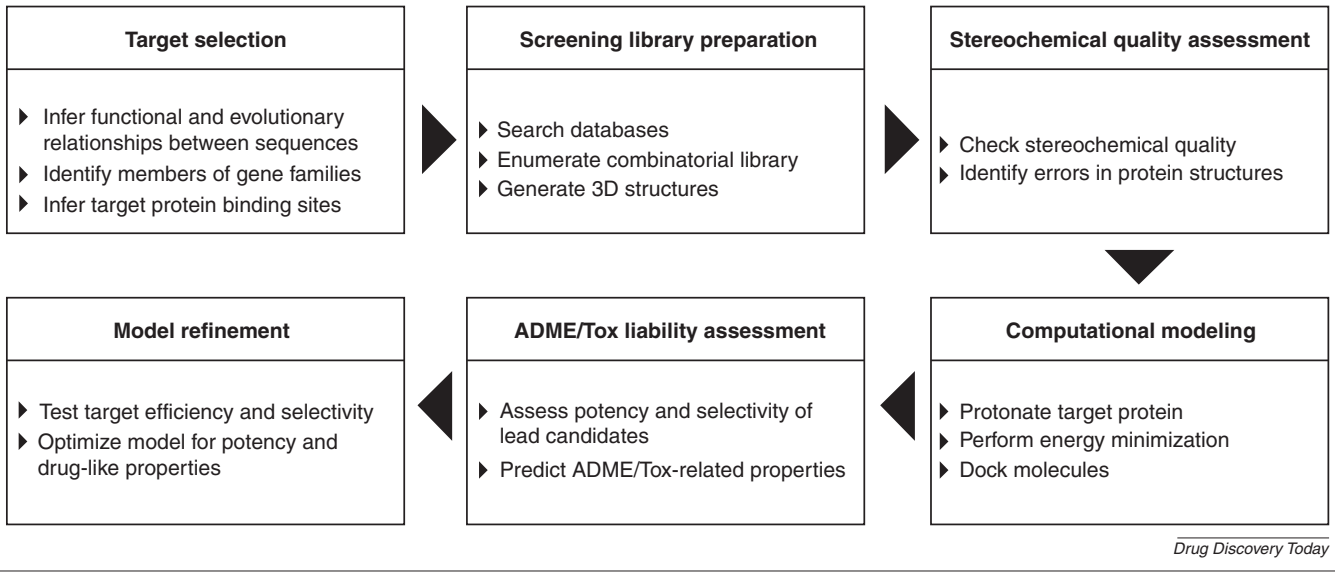

FIGURE 3

Example roadmap for a structure-based virtual screening campaign. A structure-based screening campaign usually comprises the following steps: (i) target selection, (ii) library preparation, (iii) stereochemical quality assessment, (iv) computational modeling, (v) ADME/Tox assessment and (vi) computational optimization.

same cellular proteins are used by different viruses for replication. Such compounds might have less possibility of resistance development but might lead to an increased risk of toxicity. By contrast, drugs that target viral proteins are likely to be more specific and less toxic in nature but with a narrower spectrum of action and higher likelihood of viral drug resistance. The life cycle of a typical virus consists of several stages, from virus attachment to the host cell to the release of the progeny virions from the cell. As such, one could target specific processes in viral infection including virus attachment, uncoating, viral RNA replication, and viral protein synthesis and processing.

The use of structure-guided design methods is important for identifying and selecting protein targets, as well as for identifying hits and screening fragments [44] (Fig. 3). The discovery of novel natural inhibitors for human rhinovirus (HRV) coat protein using a structure-guided search method was described by Rollinger *et al.* [45]. HRVs are small, non-enveloped, single-stranded RNA enteroviruses belonging to the *Picornaviridae* family. The viral capsid protein contains a hydrophobic pocket occupied by a pocket factor. Displacing this pocket factor with small antiviral compounds could trigger conformational changes in the capsid protein, which prevent the virus from uncoating and/or attaching to the cell surfaces. A structure-guided search, based on features characteristic for ligand binding in the hydrophobic pocket, was performed on a database of 9676 plant metabolites endowed with antiviral activity. The strategy eventually led to the discovery of asafetida and its constituent compounds, which have selective inhibitory activity against HRV serotype 2. Other structure-guided lead discovery initiatives targeting viral proteins have also been reported, including those against dengue virus envelope protein [46], HIV type 1 integrase [47] and hepatitis C virus RNA-dependent RNA polymerase [48].

Computer-aided vaccine design

Selecting antigens that could induce effective protective response is difficult because of the combinatorial nature of the human immune system. A large repertoire of immunoglobulins and T-cell receptors is known to exist, generated by mechanisms such as the combinatorial diversity of the variable, diversity and joining genes, the N-diversity, and for immunoglobulins, the somatic hypermutations [49,50]. More than 4600 HLA alleles have been reported to date, and because a fully heterogeneous person can inherit up to six different HLA class I alleles and an equal number of class II alleles, the theoretical number of HLA haplotypes is greater than 10^{12}. In addition, the number of T-cell epitope candidates is more than 10^{11}, and many more B-cell epitope candidates are known to exist.

In the early days, vaccines were primarily developed using dead or weakened forms of pathogens. More recently, over the past two decades, advances in information technologies have provided the basis for systematic discovery of immunogenic epitopes for subunit or peptide-based vaccine design [51,52]. Much emphasis has been placed on computationally identifying evolutionarily conserved amino acid sequences on pathogen proteomes, which are immunologically relevant as potential T-cell epitopes [53]. Such epitopes could offer broader protection across diverse subtypes and are particularly useful against pathogens with pandemic potential, such as the influenza A virus [54]. Another approach is to identify 'promiscuous' peptides that could bind to a wide repertoire of HLA molecules [55]. By making sure that the most frequent HLA molecules will bind to at least one of the peptides in the vaccine cocktail, this method enables the design of broad-based peptide vaccines with improved population coverage. Halling-Brown *et al.* [56] have shown that vaccine antigens contain fewer predicted HLA-binding peptides that control bacterial proteins in most subcellular locations. A third strategy would thus be to identify a protein that contains a single immuno-dominant epitope.

Much effort has also been devoted to developing tools that can help identify B-cell epitopes on antigen sequences. B-cell epitopes can be either linear or conformational in nature [49]. Although only 10% of B-cell epitopes are linear, they have been the subject of intense interest in recent years because they are considered easier

to design than their conformational counterparts [57]. *In silico* screens of small-molecule chemokine antagonists have also been reported [58] and used for the discovery of vaccine adjuvants augmenting human T-cell proliferation.

Concluding remarks

Here, we have discussed ways in which mathematical modeling in epidemiology, computational biology and bioinformatics has been used to better our understanding of infectious agents. Although the first mathematical model in epidemiology was reported in 1760, it took one and a half centuries for deterministic epidemiology to take off and a further century before it was widely embraced [59]. With the rapid growth in the variety of analytic tools and the increasing availability of large volumes of genomic, functional, clinical and epidemiological data in scientific literature, public databases and clinical records, we are now in the midst of a golden era of infectious disease research. One important challenge will be how to integrate the methods of various technology advances and make sense of the data generated using these techniques. In the next few years, it is expected that more sophisticated methods will emerge to enable integrated data analysis and higher level experimental design. This will not only enhance our understanding of the molecular biology and pathogenesis of infectious diseases but also enable the design of next-generation diagnostics and therapeutics and new control strategies to contain their spread.

References

1 Bernoulli, D. (1760) Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. In *Mémoires de Mathématiques et de Physique*. Académie Royale des Sciences pp. 1–45
2 Hamer, W.H. (1906) Epidemic disease in England. *Lancet* 1, 733–739
3 Ross, R. (1905) The logical basis of the sanitary policy of mosquito reduction. *Science* 22, 689–699
4 Greene, J.M. *et al.* (2007) National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. *Infect. Immun.* 75, 3212–3219
5 Bradley, D.J. (1994) Watson, Swellengrebel and species sanitation: environmental and ecological aspects. *Parassitologia* 36, 137–147
6 Ferguson, N.M. *et al.* (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 437, 209–214
7 Botten, J. *et al.* (2010) Coverage of related pathogenic species by multivalent and cross-protective vaccine design: arenaviruses as a model system. *Microbiol. Mol. Biol. Rev.* 74, 157–170
8 Brusic, V. and Ranganathan, S. (2008) Critical technologies for bioinformatics. *Brief. Bioinform.* 9, 261–262
9 Valleron, A.J. *et al.* (2010) Transmissibility and geographic spread of the 1889 influenza pandemic. *Proc. Natl. Acad. Sci. U. S. A.* 107, 8778–8781
10 Donnelly, C.A. *et al.* (2003) Epidemiological determinants of spread of casual agent of severe acute respiratory syndrome in Hong Kong. *Lancet* 361, 1761–1766
11 Merril, R.M. and Timmreck, T.C., eds) (2006) *Introduction to Epidemiology*, Jones & Bartlett Pub
12 Mangili, A. and Gendreau, M.A. (2005) Transmission of infectious diseases during commercial air travel. *Lancet* 365, 989–996
13 Brownstein, J.S. *et al.* (2006) Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLoS Med.* 3, e401
14 Crépey, P. and Barthélemy, M. (2007) Detecting robust patterns in the spread of epidemics: a case study of influenza in the United States and France. *Am. J. Epidemiol.* 166, 1244–1251
15 Viboud, C. *et al.* (2006) Air travel and the spread of influenza: important caveats. *PLoS Med.* 3, e503
16 Ferguson, N.M. *et al.* (2006) Strategies for mitigating an influenza pandemic. *Nature* 442, 448–452
17 Borkowski, M. *et al.* (2009) Epidemic modeling with discrete-space scheduled walkers: extensions and research opportunities. *BMC Public Health* 9 (Suppl. 1), S14
18 Fu, X. *et al.* (2009) Key node selection for containing infectious disease spread using particle swarm optimization. *Proc. IEEE SIS', 2009* 109–113
19 Kemena, C. and Notredame, C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* 25, 2455–2465
20 Bhattacharya, T. *et al.* (2007) Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* 315, 1583–1586
21 Shannon, C.E. (1950) Prediction and entropy of printed English. *Bell Syst. Tech. J.* 30, 50–64
22 Tong, J.C. *et al.* (2010) HLA class I restriction as a possible driving force for Chikungunya evolution. *PLoS One* 5, e9291
23 Miotto, O. *et al.* (2010) Complete-proteome mapping of human influenza A adaptive mutations: implications for human transmissibility of zoonotic strains. *PLoS One* 5, e9025
24 Sheng, C. *et al.* (2010) Mining mutation patterns in biological sequences. In *Proceedings of the 20th International Conference on Data Engineering*, Los Angeles, USA
25 Grenfell, B.T. *et al.* (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303, 327–332
26 Bennett, S.N. *et al.* (2003) Selection-driven evolution of emergent dengue virus. *Mol. Biol. Evol.* 20, 1650–1658
27 Suzuki, Y. and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16, 1315–1328
28 Suzuki, Y. *et al.* (2001) ADAPTSITE: detecting natural selection at single amino acid sites. *Bioinformatics* 17, 660–661
29 Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* 217, 624–626
30 Nei, M. (2005) Selectionism and neutralism in molecular evolution. *Mol. Biol. Evol.* 22, 2318–2342
31 Miyata, T. and Yasunaga, T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *J. Mol. Evol.* 16, 23–36
32 Lemey, P. *et al.* (2007) Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLOS Comput. Biol.* 3, e29
33 Air, G.M. (1981) Sequence relationships among the hemagglutinin genes of 12 subtypes of influenza A virus. *Proc. Natl. Acad. Sci. U. S. A.* 78, 7639–7643
34 Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426
35 Muse, S.V. and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11, 715–724
36 Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–736
37 Kosakovsky Pond, S.L. *et al.* (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol. Biol. Evol.* 25, 1809–1824
38 Yang, Z. and Nielsen, R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19, 908–917
39 Enquist, L.W. *et al.* eds (1999) *Principles of Virology: Molecular Biology, Pathogenesis, and Control*, Am. Soc. Microbiol.
40 Dyer, M.D. *et al.* (2007) Computational prediction of host–pathogen protein-protein interactions. *Bioinformatics* 23, i159–i166
41 Shelest, E. *et al.* (2003) Prediction of potential C/ECP/NF-κB composite elements using matrix-based search methods. *In Silico Biol.* 3, 71–79
42 Halling-Brown, M. *et al.* (2010) ImmunoGrid: towards agent-based simulations of the human immune system at a natural scale. *Philos. Transact. A. Math. Phys. Eng. Sci.* 368, 2799–2815
43 De Clercq, E. (2002) Strategies in the design of antiviral drugs. *Nat. Rev. Drug Discov.* 1, 13–25
44 Congreve, M. *et al.* (2005) Structural biology and drug discovery. *Drug Discov. Today* 10, 895–907
45 Rollinger, J.M. *et al.* (2008) Structure-based virtual screening for the discovery of natural inhibitors for human rhinovirus coat protein. *J. Med. Chem.* 51, 842–851
46 Zhou, Z. *et al.* (2008) Antiviral compounds discovered by virtual screening of small-molecule libraries against dengue virus E protein. *ACS Chem. Biol.* 3, 765–775
47 Liao, C. *et al.* (2007) Virtual screening application of a model of full-length HIV-1 integrase complexed with viral DNA. *Bioorg. Med. Chem. Lett.* 17, 5361–5365
48 Ryu, K. *et al.* (2009) Identification of novel inhibitors of HCV RNA-dependent RNA polymerase by pharmacophore-based virtual screening and *in vitro* evaluation. *Bioorg. Med. Chem.* 17, 2975–2982

49 Lefranc, M.-P. and Lefranc, G. (2001) *The Immunoglobulin FactsBook*. Academic Press 458 pp.
50 Lefranc, M.-P. and Lefranc, G. (2001) *The T cell receptor FactsBook*. Academic Press 398 p.
51 Flower, D.R. (2007) Immunoinformatics and the *in silico* prediction of immunogenicity. An introduction. *Methods Mol. Biol.* 409, 1–15
52 Yang, X. and Yu, X. (2009) An introduction to epitope prediction methods and software. *Rev. Med. Virol.* 19, 77–96
53 Koo, Q.Y. *et al.* (2009) Conservation and variability of West Nile virus proteins. *PLoS One* 4, e5352
54 Ekiert, D.C. *et al.* (2009) Antibody recognition of a highly conserved influenza virus epitope. *Science* 324, 246–251
55 Zhang, H. *et al.* (2009) Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. *Bioinformatics* 25, 83–89
56 Halling-Brown, M. *et al.* (2009) Proteins accessible to immune surveillance show significant T-cell epitope depletion: implications for vaccine design. *Mol. Immunol.* 46, 2699–2705
57 Tong, J.C. and Ren, E.C. (2009) Immunoinformatics: current trends and future directions. *Drug Discov. Today* 14, 684–689
58 Davies, M.N. *et al.* (2009) Toward the discovery of vaccine adjuvants: coupling *in silico* screening and *in vitro* analysis of antagonist binding to human and mouse CCR4 receptors. *PLoS One* 4, e8084
59 McKenzie, F.E. and Samba, E.M. (2004) The role of mathematical modeling in evidence-based malaria control. *Am. J. Trop. Med. Hyg.* 71 (Suppl. 2), 94–96