BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 15 2013, pages 1904–1907
doi:10.1093/bioinformatics/btt304

*Sequence analysis*

Advance Access publication May 28, 2013

# EpiCombFlu: exploring known influenza epitopes and their combination to design a universal influenza vaccine

Varun Jaiswal, Sree K. Chanumolu, Pankaj Sharma, Rajinder S. Chauhan and Chittaranjan Rout*

Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh 173234, India

Associate Editor: John Hancock

**ABSTRACT**

**Motivation:** Influenza is responsible for half a million deaths annually, and vaccination is the best preventive measure against this pervasive health problem. Influenza vaccines developed from surveillance data of each season are strain-specific, and therefore, are unable to provide protection against pandemic strains arising from antigenic shift and drift. Seasonal epidemics and occasional pandemics of influenza have created a need for a universal influenza vaccine (UIV). Researchers have shown that a combination of conserved epitopes has the potential to be used as a UIV.

**Result:** In the present work, available data on strains, proteins, epitopes and their associated information were used to develop a Web resource, 'EpiCombFlu', which can explore different influenza epitopes and their combinations for conservation among different strains, population coverage and immune response for vaccine design. Forward selection algorithm was implemented in EpiCombFlu to select optimum combination of epitopes that may be expressed and evaluated as potential UIV.

**Availability:** The Web resource is freely available at http://117.211.115.67/influenza/home.html.

**Contact:** chittaranjan.rout@juit.ac.in

**Supplementary information:** Supplementary data are available at Bioinformatics online.

Received on February 26, 2013; revised on May 21, 2013; accepted on May 22, 2013

## 1 INTRODUCTION

The influenza costs half a million lives annually. Despite availability of vaccines, there is always a threat of pandemic from newly emerging virulent strains. Current influenza vaccines, i.e. trivalent inactivated vaccine and live-attenuated vaccine, provide moderate protection, which is either greatly reduced or absent in some seasons (Osterholm *et al.*, 2012). Pandemics of flu in the past have indicated that these vaccines were not efficient against new virulent strains.

The trivalent inactivated vaccine and live-attenuated vaccine influenza vaccines provide protection by producing neutralizing antibodies against surface structural glycoproteins, hemagglutinin (HA) and neuraminidase (NA) (Fiore *et al.*, 2009). But frequent mutations in these surface proteins result in escape of many virulent strains from antibody-mediated immunity (Fiore *et al.*, 2009). The flu pandemic in 2009 and development of resistance strains to ribavirin and oseltamivir drugs (Regoes and Bonhoeffer, 2006) reignited the hunt for a universal influenza vaccine (UIV), which can effectively counter the epidemic and pandemics caused by the influenza virus.

In recent years, epitope-based vaccines have shown their effectiveness against HIV, hepatitis B and influenza viruses in clinical studies (Atsmon *et al.*, 2012; Engler *et al.*, 2001; Gahery *et al.*, 2006). Current approaches have been focusing on B-cell epitope (BCE)- and/or T-cell epitope (TCE)-mediated immune responses to develop a UIV (Goodman *et al.*, 2011; Kaur *et al.*, 2011). Single conserved ectodomain (epitope) of M2 (M2e) protein was developed as a UIV against the influenza virus, but mutations in the middle part of the ectodomain eliminated its potential as a UIV (Wang *et al.*, 2009). A vaccine produced from a single conserved TCE or BCE may not provide broad-specific protection, but a vaccine developed from a cocktail of few conserved epitopes can take it closer to a UIV (Atsmon *et al.*, 2012). Multimeric-001, a vaccine developed by BiondVax and containing a trimeric combination of nine epitopes from the influenza virus, has provided broad protection and is currently in phase-II clinical trials (Atsmon *et al.*, 2012). Recent positive outcomes of epitopes-based vaccines have provided an opportunity that available enormous information (epitope and protein sequences, immunogenic data, strains, etc.) on known epitopes (BCEs and TCEs) in public databases may be explored for the selection of epitopes and/or their combinations to design a potential UIV. Therefore, a Web resource named as 'EpiCombFlu' has been developed, which consists of the 'Epitope information resource' and the 'Epitope combination explorer'. The former is a database containing epitopes' strain coverage and their immunogenic data, whereas the latter explores combinations of epitopes for maximum strains' coverage using forward selection algorithm (FSA).

## 2 METHODS

### 2.1 Data collection of proteins and epitopes of influenza virus

The flow diagram of the proposed method is provided in Figure 1. All available sequences of 12 proteins (complete set of proteins encoded by influenza), HA, NA, NP, M1, M2, PB1, PB2, PB1-F2, NS1, NS2, PA and PA-X (Supplementary Table S1), of the influenza A virus were

*To whom correspondence should be addressed.

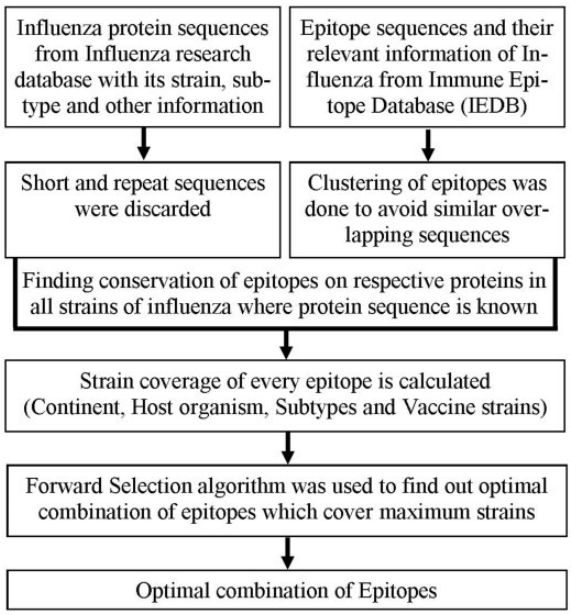


**Fig. 1.** Flow diagram for finding optimal combination of epitopes to design potential UIVs

retrieved from an influenza research database (http://www.fludb.org/). Redundant protein sequences within the same strain were discarded while constricting protein sequences files for the aforementioned 12 proteins. Different sequences of each individual protein (i.e. HA, NA, etc.) and their associated information such as strain name, subtype, country and host were stored. TCE and BCE 'full data' files were downloaded from immune epitope database (http://www.immuneepitope.org/), and all epitopes' source organism (influenza), host organism (human) and 'qualitative measurement' related to immune response (positive, positive-low, positive-intermediate or positive-high) from 11 influenza proteins were extracted separately. Individual epitopes along with their literature reference, immune epitope database Id and information on types of immunogenic response and human leukocyte antigen (HLA) allele (in case of TCE only) were stored in the MySQL database as backend data for development of a Web resource. To avoid similar epitopes from the same region of the protein, the epitopes were clustered, so that only one epitope is selected from each cluster. CD-Hit (http://weizhong-lab.ucsd.edu/cd-hit/) was used for clustering of similar epitopes with threshold identity cutoff value of 80%. Total number of non-overlapping clusters obtained in HA, NA, NP, M1, M2, PB1, PB2, PB1-F2, NS1, NS2 and PA proteins was 238, 72, 74, 33, 9, 48, 16, 4, 24, 10 and 20, respectively. Clustering of epitopes helps in covering different regions of a protein.

### 2.2 Calculation of epitopes' strain coverage among different strains and population coverage

All epitopes were matched against respective proteins to determine their strain coverage. Exact sequence match was taken as conservation cutoff because there is immense inter-residues interaction in T-cell epitope, so even a single amino acid substitution can alter interaction of other residues to T-cell receptor or HLA molecules (Rimmelzwaan *et al.*, 2004). Data of all epitopes in terms of coverage in strains, vaccine strains, continent-wise strains, subtypes and their host organism were calculated by using in-house programs. Individual strain coverage (ISC) of each epitope was computed as the number of strains containing the epitope in their respective protein sequences, and cumulative strain coverage (CSC) for combination of epitopes was determined as the number of all strains containing any epitope in their respective protein sequences. The strain coverage percentage of each epitope was calculated as ISC number multiplied by 100 and divided by total number of strains where that protein sequence was available. Similarly, coverage percentage of combination of epitopes was calculated as CSC number multiplied by 100 and divided by total number of strains where any protein sequence was known. This ISC information was stored in the MySQL database.

Population coverage of TCE-based vaccine is crucial due to polymorphism of major histocompatibility complex molecules, which display distinct peptide-binding specificity. The TCE that binds to several HLAs or common HLA supertype provides maximum population coverage. HLA supertypes of all TCEs [cytotoxic T-cell (CTL) and T-helper (Th) cells] were also stored. Global representatives of the most frequent supertype of HLA-A and HLA-B (A1, A2, A3, B24 and B7) were categorized to provide population coverage information (Sidney *et al.*, 2008).

### 2.3 Database: epitope information resource

Immunogenic information of epitopes and their coverage data were stored in the MySQL database as backend data. The user interface (Web pages) was designed in hypertext markup language. PERL and PHP languages were used in developing a server side program, which retrieves information from the MySQL database for processing user's request. EpiCombFlu's resource can be searched by protein's name, continent's name, host organism's name and/or epitope's type. The results can be refined and tabulated (downloadable in .xls format).

### 2.4 FSA for finding optimal combination of epitopes (Epitope combination explorer)

The main component of EpiCombFlu is the Epitope combination explorer, which implements FSA to explore different combinations of epitopes for UIV. Evaluation of all combinations of epitopes covering approximately all global strains of influenza can result in combinatorial explosion. Therefore, FSA was developed and implemented in the Web server to determine an optimal combination of epitopes covering almost all global strains. FSA takes initial epitope with maximum strain coverage. To select the next one (epitope), it determines CSC for a combination of new and initial epitopes, and the epitope (irrespective of protein) with maximum strain coverage is included into the combination. Similar method is followed to include the third epitope into the combination, and the algorithm incorporates epitopes in an iterative manner till user-mentioned numbers of epitopes are added in the combination. The FSA can also take initial epitope(s) from user and then add new epitopes (according to FSA) to provide optimal combination for maximum strain coverage. The stepwise execution of FSA and its output is provided in the Supplementary Figure (Fig. S1) and 'Tutorial'.

## 3 RESULTS

Web resource 'EpiCombFlu' provides compiled information on epitope's strain coverage, epitope type, host type, literature references, etc. so that user can evaluate the prospective epitope(s) for developing a UIV. The Web resource also provides an analysis facility that adds epitopes automatically on the basis of maximum strain coverage. The inclusion of information on strain and population coverage and both humoral and cellular immune response features in EpiCombFlu is crucial for epitope-based broad-spectrum vaccine (Ben-Yedidia and Arnon, 2005).

### 3.1 Description

The Web resource provides a user-friendly interface to search and retrieve information related to influenza epitopes in 'Epitope information resource'. Output of the search is presented in a tabular format, which contains strain coverage and immunogenic information of epitopes. 'Epitope combination explorer' explores different combinations of influenza epitopes using FSA. There are two options to submit epitope sequences into this module: (i) Selection of an epitope from dropdown menu of corresponding proteins [mouse on epitope sequence in the 'list box' displays a hover showing information about its immune response, and coverage in strains (global, continents-wise and vaccine, strains), subtypes and their host organism] and (ii) insertion of epitope sequence is provided in the text box. The epitope combination explorer provides similar information as 'Epitope information resource' about conservation and immune response, but for a combination of selected epitopes.

### 3.2 Conservation of epitopes according to strain, subtype and host type

Epitope from HA, GLFGAIAGFI, has maximum strain coverage and is conserved in 27 402 strains. Top 10 epitopes from HA protein have conservation range between 27 402 and 13 761 strains. In the Web resource, strain coverage information of each epitope is hovered so that user can select conserved epitopes. Strain coverage data of top 10 epitopes of each protein is given in Supplementary Table S1.

### 3.3 Performance of the FSA for UIV design

When the epitope with maximum strain coverage, GLFGAIAGFI, was used as an initial epitope for FSA and all other epitopes were added automatically based on maximum cumulative strain coverage of combined epitopes (Refer Section 2.4 and Supplementary Fig. S1), nine epitopes were selected from five proteins (4 HA, 2 NA, 1 M1, 1 NP and 1 NS1) covering 51 222 of total 57 414 strains of influenza A virus (Table 1). Epitopes 1, 3 and 7 provide CTL immune responses. Epitopes binding to major histocompatibility complex II allele were taken as Th-cell epitopes. Interestingly, other six epitopes are known to induce Th immune response, which is crucial to encounter viral infection (Tan *et al.*, 2011). First and seventh CTL and second and fifth Th epitopes are known to bind with multiple HLA supertypes, indicating their applicability to global human population.

Epitope-based vaccine, Multimeric-001, is composed of nine epitopes (four BCE and one Th epitope from HA protein, two CTL and one Th from NP protein and one peptide that contains both BCE and CTL epitope from M1 protein), and epitopes forming this vaccine have coverage of only 30 848 strains of influenza A virus (Supplementary Table S2) in comparison with 51 222 strains' coverage for nine-epitope combination discovered in the present work by FSA. Besides coverage, the FSA-discovered combination of epitopes is expected to induce better immune response and population coverage that is important in vaccine design against influenza virus.

## 4 DISCUSSION

The EpiCombFlu resource has been developed to assist vaccinologists in producing epitope-based UIVs. Different combinations of epitopes (potential UIVs with optimum strain coverage) can be identified by using FSA with different initial epitope(s). The combination of epitopes can be expressed as synthetic protein and its UIV potential could be checked. As evolutionary conserved epitopes are from functionally important parts of proteins, these immunogenic parts are, therefore, expected to be retained by new pandemic strains (McMurry *et al.*, 2008). The developed vaccine containing these epitopes is anticipated to prevent or mitigate infection by pandemic strains. The performance of EpiCombFlu server does not depend on initially selected epitopes. Many independent studies were carried out to verify whether selection of different combinations of epitopes can alter the strain coverage (Supplementary Table S2). In first case, instead of taking an epitope with maximum strain coverage as the initial epitope, a 20-length epitope from HA was taken, and FSA was used to find out combination of epitopes with optimal strain coverage. Similarly, in the second case, the same 20-length epitope was taken as the initial epitope, and only $\geq$10-length epitopes were used subsequently for combination by FSA

**Table 1.** FSA identified nine epitopes, and their immune response and strain coverage information[a]

| Sr. No | Epitope sequence | Protein name | B-cell | T-cell | Th-cell | No. CSC | No. ISC |
|---|---|---|---|---|---|---|---|
| 1 | GLFGAIAGFI | HA | N | Y | N | 27 402 | 27 402 |
| 2 | IYWTIVKPGDILLINS | HA | N | N | Y | 40 535 | 13 178 |
| 3 | KTRPILSPLTK | M1 | N | Y | N | 44 668 | 24 518 |
| 4 | STDTVDTVLEKNVTVTHS | HA | N | N | Y | 47 747 | 16 562 |
| 5 | RTFFLTQGALLNDKHSN | NA | N | N | Y | 48 960 | 14 624 |
| 6 | DRLRRDQKS | NS1 | N | N | Y | 49 629 | 14 918 |
| 7 | ILRGSVAHK | NP | N | Y | N | 50 182 | 18 800 |
| 8 | EQLSSVSSFERFE | HA | N | N | Y | 50 729 | 17 048 |
| 9 | CVCINGTCTVVMTDGSA | NA | N | N | Y | 51 222 | 8646 |

[a]Refer section 2.4.
*Note*: 'Y' and 'N' denote 'Yes' and 'No', respectively.
CSC, cumulative strain coverage of combined epitope; ISC, individual strain coverage.

(Supplementary Table S2). The FSA method was able to find combination of nine epitopes having ~90% strain coverage. Even when four BCEs with low strain coverage were used as starting epitopes, FSA was effective in getting epitope combinations having >85% strain coverage. This outcome of FSA justifies that the strain coverage is not dependent on selection of initial epitopes. Even if the user likes to include one or more epitopes (i.e. known to be highly immunogenic) for developing a UIV, the Web resource (FSA) can easily combine other database epitopes to get maximum strain coverage (Supplementary Table S2).

*Conflict of Interest*: none declared.

## REFERENCES


Atsmon,J. *et al.* (2012) Safety and immunogenicity of multimeric-001–a novel universal influenza vaccine. *J. Clin. Immunol.*, **32**, 595–603.

Ben-Yedidia,T. and Arnon,R. (2005) Towards an epitope-based human vaccine for influenza. *Hum. Vaccin.*, **1**, 95–101.

Engler,O.B. *et al.* (2001) Peptide vaccines against hepatitis B virus: from animal model to human studies. *Mol. Immunol.*, **38**, 457–465.

Fiore,A.E. *et al.* (2009) Seasonal influenza vaccines. *Curr. Top. Microbiol. Immunol.*, **333**, 43–82.

Gahery,H. *et al.* (2006) New CD4+ and CD8+ T cell responses induced in chronically HIV type-1-infected patients after immunizations with an HIV type 1 lipopeptide vaccine. *AIDS Res. Hum. Retroviruses*, **22**, 684–694.

Goodman,A.G. *et al.* (2011) A human multi-epitope recombinant vaccinia virus as a universal T cell vaccine candidate against influenza virus. *PLoS One*, **6**, e25938.

Kaur,K. *et al.* (2011) Targeting B cell responses in universal influenza vaccine design. *Trends Immunol.*, **32**, 524–531.

McMurry,J.A. *et al.* (2008) A call to cellular & humoral arms: enlisting cognate T cell help to develop broad-spectrum vaccines against influenza A. *Hum. Vaccin.*, **4**, 148–157.

Osterholm,M.T. *et al.* (2012) Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. *Lancet Infect. Dis.*, **12**, 36–44.

Regoes,R.R. and Bonhoeffer,S. (2006) Emergence of drug-resistant influenza virus: population dynamical considerations. *Science*, **312**, 389–391.

Rimmelzwaan,G.F. *et al.* (2004) Sequence variation in the influenza A virus nucleoprotein associated with escape from cytotoxic T lymphocytes. *Virus Res.*, **103**, 97–100.

Sidney,J. *et al.* (2008) HLA class I supertypes: a revised and updated classification. *BMC Immunol.*, **9**, 1.

Tan,P.T. *et al.* (2011) Highly conserved influenza A sequences as T cell epitopes-based vaccine targets to address the viral variability. *Hum. Vaccin*, **7**, 402–409.

Wang,Y. *et al.* (2009) Monoclonal antibody recognizing SLLTEVET epitope of M2 protein potently inhibited the replication of influenza A viruses in MDCK cells. *Biochem. Biophys. Res. Commun.*, **385**, 118–122.