PAComplex: a web server to infer peptide antigen families and binding models from TCR–pMHC complexes

I-Hsin Liu¹, Yu-Shu Lo¹ and Jinn-Moon Yang¹,²,*

¹Institute of Bioinformatics and Systems Biology and ²Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan

Received March 8, 2011; Revised April 22, 2011; Accepted May 12, 2011

ABSTRACT

One of the most adaptive immune responses is triggered by specific T-cell receptors (TCR) binding to peptide-major histocompatibility complexes (pMHC). Despite the availability of many prediction servers to identify peptides binding to MHC, these servers are often lacking in peptide–TCR interactions and detailed atomic interacting models. PAComplex is the first web server investigating both pMHC and peptide-TCR interfaces to infer peptide antigens and homologous peptide antigens of a query. This server first identifies significantly similar TCR–pMHC templates (joint Z-value ≥ 4.0) of the query by using antibody–antigen and protein–protein interacting scoring matrices for peptide-TCR and pMHC interfaces, respectively. PAComplex then identifies the homologous peptide antigens of these hit templates from complete pathogen genome databases (>10⁸ peptide candidates from 389 pathogens) and experimental peptide databases (80,057 peptides in 2287 species). Finally, the server outputs peptide antigens and homologous peptide antigens of the query and displays detailed interacting models (e.g. hydrogen bonds and steric interactions in two interfaces) of hitTCR-pMHC templates. Experimental results demonstrate that the proposed server can achieve high prediction accuracy and offer potential peptide antigens across pathogens. We believe that the server is able to provide valuable insights for the peptide vaccine and MHC restriction. The PAComplex server is available at http://PAcomplex.life.nctu.edu.tw.

INTRODUCTION

An immune system protects an organism from diseases by identifying and killing pathogens (1). One of the most adaptive immune responses is triggered by specific T-cell receptors (TCRs) binding to peptide-major histocompatibility complexes (pMHC) molecules. An increasing number of available binding peptide antigens that are reliable (2–4) and high-throughput experiments that provide systematic identification of pMHC interactions explain the growing requirement for fast and accurate computational methods for discovering homologous peptide antigens of a new peptide antigen and developing peptide-based vaccines for pathogens.

Many methods have been proposed for predicting pMHC interactions. These methods can be roughly divided into the sequence-based methods such as motif matching (5,6), matrix methods [e.g. SYFPEITHI (7), MAPPP (8), IEDB (9)] and machine learning approaches [e.g. SVMHC (10)]; and structure-based approaches [e.g. PREDEP (11) and MODPROPEP (12)]. However, these methods are often lack of the TCR and pMHC binding, which is critical to trigger adaptive immune responses. Since the increasing number of TCR–pMHC crystal structures to investigate both pMHC and peptide-TCR interfaces provides further insights for understanding TCR–pMHC interactions and binding mechanisms. Additionally, discovering homologous peptide antigens (called peptide antigen family) to a known peptide antigen often provides a valuable reference for efforts to elucidate the functions of a new peptide antigen.

To address these issues, we propose the PAComplex server for predicting TCR–pMHC interactions and inferring antigen families across organisms of a query protein or a set of peptides. To our best knowledge, PAComplex is the first web server investigating both pMHC and peptide-TCR interfaces to infer peptide antigens and homologous peptide antigens of a query. Additionally, peptide antigen families are derived from a complete pathogen genome database (>10⁸ peptide candidates from 389 pathogens) and experimental peptide databases to demonstrate the feasibility of the PAComplex server and increase the number of potential antigens. Moreover, for a peptide antigen family, the amino acid composition and conservation are

*To whom correspondence should be addressed. Tel: 886 3 5712121 (ext. 56942); Fax: 886 3 5729288; Email: moon@faculty.nctu.edu.tw

© The Author(s) 2011. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
evaluated at each position. Experimental results demonstrate that the server can improve the peptide antigen prediction accuracy and is useful for identifying peptide antigen families by using two interfaces of TCR–pMHC structures. Furthermore, the proposed server provides a valuable reference for efforts to develop peptide vaccines and elucidate MHC restriction and T-cell activation.

METHOD AND IMPLEMENTATION

Homologous peptide antigen

The concept of homologous peptide antigen is the core of this server. We define the homologous peptide antigen (p') of the peptide (p) in template complex as follows: (i) p and p' can be bound by the same MHC forming pMHC and p'MHC, respectively, with the significant interface similarity (Z_{MHC} ≥ 1.645); (ii) pMHC and p'MHC can be recognized by the same TCR with significant TCR-TCR similarity (Z_{TCR} ≥ 1.645); and (iii) TCR-pMHC and TCR-p'MHC share significant complex similarity (Joint Z ≥ 4.0). The joint Z-value (J_z) is defined as

$$J_z = \sqrt{Z_{MHC} \times Z_{TCR}}$$ (1)

The Z_{MHC} and Z_{TCR} of a TCR-p'MHC candidate with interaction score (E) can be calculated by $(E-\mu)/\sigma$, where $\mu$ is the mean and $\sigma$ is the standard deviation from 10,000 random interfaces (Supplementary Figure S1). For a TCR-pMHC template collected from Protein Data Bank (PDB), these 10,000 random interfaces are generated by substituting with another amino acid according to the amino acid composition derived from UniProt (13). Here, J_z ≥ 4.0 is considered a significant similarity according to the statistical analysis of 41 TCR-pMHC structure complexes; 80,057 experimental peptide antigens; and ≥10^6 peptide candidates derived from 864,628 protein sequences in 389 pathogens.

Template-based scoring function

We have recently proposed a template-based scoring function to determine the reliability of protein–protein interactions derived from a 3D-dimer structure (14). For measuring the pMHC interaction score, the scoring function is defined as

$$E_{tot} = E_{VDW} + E_{SP} + E_{sim}$$ (2)

Where $E_{VDW}$ and $E_{SP}$ denote steric force and special energy (i.e. hydrogen bond energy and electrostatic energy), respectively, according to four knowledge-based scoring matrices (14) which have a good achievement between pMHC and protein–protein interactions. $E_{sim}$ refers to the peptide similarity score between p and p'.

To model $E_{VDW}$ and $E_{SP}$ of the peptide-TCR interactions, we developed a new residue-based matrix (Supplementary Figure S2) because the peptide-TCR interface resembles antigen–antibody interactions and differs from protein–protein interfaces (15,16). The matrix is derived from non-redundant set which consists of 62 structural antigen–antibody complexes (including 131 interfaces) constructed by Ponomarenko et al. (17). According to this matrix, the peptide-TCR (antigen–antibody) interface prefers aromatic residues (i.e. Phe, Trp and Tyr), which interact with aliphatic residues (i.e. Ala, Val, Leu, Ile and Met) or long side-chain polar residues (i.e. Gln, His, Arg, Lys and Glu), to form strong van der Waals (VDW) forces (yellow boxes). Additionally, the scores are high if basic residues (i.e. Arg and Lys) interact to acidic residues (i.e. Asp and Glu). Conversely, the scores are low (purple box) when non-polar residues interact with polar residues.

Overview

Figure 1 shows the details of the PAComplex server to predict peptide antigens and search the template-based homologous peptide antigens of a query protein sequence (or a set of peptides) by the following steps (Figure 1A). The server initially divides the query protein sequence into fix length (ranging from 8 to 13) peptides based on selected MHC class I allele and templates. Each peptide (p') is then aligned to the bound peptide (p) of TCR–pMHC template collected from PDB. Next, the peptide antigen is examined by utilizing the template-based scoring function to statistically evaluate the complex similarity (J_z ≥ 4.0) between TCR–pMHC and TCR-p'MHC (Figure 1B and C). For each peptide antigen, the server introduces the potential TCR–pMHC binding models and the detailed residues interactions (e.g. hydrogen bonds and VDW forces) of pMHC and peptide-TCR interfaces (Figure 1D). For the hit templates, the server identifies the homologous peptide antigens with J_z ≥ 4.0 from an experimental peptide database (80,057 peptides in 2287 species) and a complete pathogen genome database (≥10^6 peptide antigen candidates with J_z ≥ 1.645 derived from 864,628 protein sequences of 389 pathogens) (Figure 1B and E). For a peptide antigen family, we measure the amino acid composition and conservation at each position (Figure 1F) by WebLogo program (18). Finally, this server provides peptide antigens, visualization of the TCR–pMHC interaction models, and peptide antigen families with conserved amino acids.

INPUT, OUTPUT AND OPTIONS

The PAComplex server is easy to use (Figure 2). Users input a protein sequence in FASTA format (or a set of fix-length peptides) and select the parameters (e.g. MHC class I allele and templates) (Figure 2A). The PAComplex server typically infers peptide antigens and homologous peptide antigens of the query within 4 s if the sequence length is ≤ 300. For a query, PAComplex shows the detailed atomic interactions and binding models using Jmol and amino acid profiles (Figure 2C) of homologous peptide antigens from experimental peptide (Figure 2E) and complete pathogen genome databases (Figure 2D). For each peptide antigen, PAComplex also presents the source proteins, organisms and experimental data. In addition, users can download summarized results of query peptides or protein sequences, the modeling TCR–pMHC complex, template structure of TCR–pMHC and peptide family of the template.
Example analysis

Protein P of hepatitis B virus. While affecting over 350 million people worldwide, hepatitis B virus (HBV) infection is a leading cause of liver diseases and hepatocellular carcinoma (19,20). Figure 1 shows the PAComplex derived results using protein P [UniProt (13) accession number: P03155, 750 residues divided into 743 8-mer peptides] of HBV genotype D as the query. Protein P, a multifunctional enzyme, converts the viral RNA genome into dsDNA in viral cytoplasmic capsids. This enzyme displays a DNA polymerase activity that can replicate either DNA or RNA templates, and a ribonuclease H (RNase H) activity that cleaves the RNA strand of RNA–DNA heteroduplexes in a partially processive 3’-to 5’-endonucleasic mode (21,22). For this query, the PAComplex server found three hit peptide antigen candidates (Jz ≥ 4.0; Figure 1C) and 73 homologous peptide antigens in 21 organisms (Supplementary Figure S3A) by using H-2Kb-peptide-TCR template [PDB entry 3CVH (23)] and the experimental peptide database. Among these three hit peptides, the peptide 497–504 (ILILGFRKI) recorded in IEDB (4) is the epitope of protein P and PAComplex presents its binding models and detailed residue interactions of peptide-TCR and pMHC interfaces (Figure 1D). For this query, the PAComplex server found three hit peptide antigen candidates (Jz ≥ 4.0; Figure 1C) and 73 homologous peptide antigens in 21 organisms (Supplementary Figure S3A) by using H-2Kb-peptide-TCR template [PDB entry 3CVH (23)] and the experimental peptide database. Among these three hit peptides, the peptide 497–504 (ILILGFRKI) recorded in IEDB (4) is the epitope of protein P and PAComplex presents its binding models and detailed residue interactions of peptide-TCR and pMHC interfaces (Figure 1D). Position 1 of the homologous peptide antigens prefers the polar residues (e.g. Ser, Thr, Arg and Lys; Figure 1F) and the first position of this peptide (pink) is polar residue Ser forming five hydrogen bonds with residues Tyr7, Glu63 and Tyr171 on MHC molecule (green) (Figure 1D). Additionally, position 7 (Lys) of the homologous peptide antigens prefers the positive residues (Arg and Lys, Figure 1F) and Lys7 of
Figure 2. PAComplex server search results using 50S ribosomal protein L5 (rplE) of *M. pneumoniae* as the query. (A) User interface for inputting the query protein sequence, MHC class I allele, and templates. (B) The peptide antigen candidates ($J_z = 4.0$) of the query. (C) Detailed atomic interactions and binding models with hydrogen bond residues and strong VDW forces. Peptide antigen families from (D) experimental peptide database and (E) complete pathogen genome database, respectively.
The hit candidate is similar to the Rank 4 peptide (SYQRFRRL) and 75–82 (KPPSFPNI) correlate well two homologous peptide antigens (SYQHFRLK and KTPSFPNI), which are epitopes of HBV alpha 1 recorded in IEDB, respectively (orange box in Supplementary Figure S3A). According to the amino acid composition (profile) of this peptide antigen family (Figure 1F), position 7 prefers positive charged residues and positions 5 and 8 prefer non-polar residues. Conversely, the compositions of Positions 2 and 4 are diverse. These two hit antigens match the profile of the antigen family on Positions 1 (polar residues), 5 (conserved residue Phe), 7 (positive or polar residues) and 8 (non-polar residues). For instance, the two hit antigens are conserved on the Position 5 with residue Phe forming strong VDW interactions with MHC [Phe74, Val97, Tyr22, Val9 and Tyr116] and TCR (Phe104 molecules (Supplementary Figure S3B)]. These residue–residue interactions (i.e. Phe-Phe, Phe-Val and Phe-Try) are high scores according to pMHC (14) and peptide-TCR (Supplementary Figure S2) scoring matrices. For peptide SYQHFRLK and SYQHFRLK, they have the positively charged residue type (e.g. Arg and Lys) on Position 7 and different residue types on Position 4. For peptides KPPSFPNI and KTPSFPNI, the only different residue type is located on Position 2. Therefore, these two hit peptides are potential peptide antigens. These results suggest that investigating multiple TCR–pMHC interfaces and the peptide antigen family are useful for predicting peptide antigens and providing valuable insight into MHC restriction and T-cell activation.

50S ribosomal protein L5 (rplE) of Mycoplasma pneumonia. 50S ribosomal protein L5 (rplE), interacting with 5S rRNA and tRNA, is an essential protein of M. pneumonia, which is the cause of human walking pneumonia (24). Based on the use of the M. pneumonia rplE (Q50306, 180 residues are divided into 172 9-mer peptides) as the query (Figure 2A), the PAComplex server infers one hit candidate (JZ ≥ 4.0; Figure 2B), 95–103 (RMWAFLEKL) and its 66 homologous peptide antigens in 32 organisms, based on the HLA-A0201-peptide-TCR template [PDB entry 2J8U (25); Figure 2D). This server provides the binding model (Figure 2C) and homologous peptide antigens via the experimental peptide database and complete pathogen genome database (Figure 2D and E).

The hit candidate is similar to the Rank 4 peptide (RMWELFDRL, red box) in the peptide family (Figure 2D). But they have three different amino acid types on Positions 4, 7 and 8 whose amino acid compositions of this family are diverse (Figure 2C). Based on binding models and interactions, Position 4 lacks any hydrogen bonds and strong VDW contacts. On the other hand, the hit peptide correlates well with the amino acid profile on the conserved positions (i.e. 1, 2, 3, 5, 6, 9) forming strong VDW forces (the right-side table of Figure 2C) based on the interactions of both peptide-TCR and pMHC interfaces in HLA-A0201-peptide-TCR template (2J8U). Above results imply that the hit peptide of rplE is a potential antigen-activating immune response. Furthermore, PAComplex provides the potential peptide antigens derived from all proteins of the query M. pneumoniae (Figure 2E) and the other 388 pathogens using a complete pathogen genome database. These potential peptide antigens across pathogens can be useful in identifying specific peptides for the target pathogen for vaccine design.

RESULTS

To evaluate the performance of PAComplex for identifying the peptide antigens and peptide antigen families, we selected two peptide sets, termed BothMT (Figure 3A) and CPD (Figure 3B). BothMT consists of 86 positive and 67 negative octamers with experimental data of both H-2Kb and TCR sides collected from IEDB (4). PAComplex aligned these 153 peptides to six H-2Kb–peptide-TCR complex templates extracted from PDB released on 25 December 2010 to evaluate the accuracies of scoring functions on variant conditions (e.g. single template, multiple templates, single side and both sides). The CPD set, which comprises ≥108 peptide candidates (JZ ≥ 1.645) derived from 864,628 protein sequences of 389 pathogens, was used to evaluate the reliability of homologous peptide antigens and it was collected by the following steps: (i) extract 389 pathogens (e.g. bacteria, archaea and virus) recorded in both IEDB and UniProt (13) databases and their respective complete genomes collected from UniProtdatabase (13); (ii) derive the positive and negative data sets from IEDB for these pathogens; and (iii) extract 41 TCR–pMHC complexes from PDB.

Figure 3A illustrates the receiver operating characteristic (ROC) curves (i.e. true positive and false positive rates) of our scoring functions on single and multiple templates using one interface (i.e. pMHC and peptide-TCR interfaces) and two interfaces (i.e. TCR–pMHC complex). We observed several interesting results: (i) the scoring function using pMHC interface (blue lines) yields a higher accuracy than using peptide-TCR interface (green lines); (ii) using multiple templates (solid lines) is better than using single template (dot lines); and (iii) using two interfaces with multiple templates (red) is the best among these six combinations.

Next, the JZ threshold for reliable homologous peptide antigens is determined by evaluating the PAComplex server on the large-scale CPD data set (Figure 3B). This server was tested on >1010 peptides derived from 864,628 protein sequences of 389 pathogens. Among these peptides, over 108 peptide candidates with JZ ≥ 1.645 were selected for analyzing the relationships between JZ values with both the numbers of positive homologous peptide antigens (blue, recorded in IEDB) and precision (red). When JZ is higher than 4.0, the precision >0.8 and the number of positive antigens exceeds 1600 according to the positive and negative data sets. If the JZ threshold is set to 4.0, the total number of inferring possible peptide antigens surpasses 4000000 (Supplementary Figure S4) statistically derived from 41 TCR–pMHC complexes. The amino acid compositions (profiles) of these 1600 positive peptide
antigens closely correspond to the ones obtained from peptide antigen families (4,000,000 antigens). These experimental results demonstrate that this server achieves high accuracy and is able to provide potential peptide antigens across pathogens.

**CONCLUSIONS**

This work demonstrates the feasibility of using the PAComplex server to identify peptide antigens and homologous peptide antigens. The proposed server provides detailed atomic interactions, binding models, amino acid compositions of peptide families, source proteins and organisms and experimental data. PAComplex server is the first to infer peptide antigens and homologous peptide antigens by considering two TCR–pMHC interfaces from complete pathogen genome and experimental peptide databases. Experimental results demonstrate that the server is highly accurate and capable of providing potential peptide antigens across pathogens. We believe that PAComplex is a fast homologous peptide antigens search server and is able to provide valuable insights into the peptide vaccine, MHC restriction and T-cell activation.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**ACKNOWLEDGEMENTS**

Authors are grateful to both the hardware and software supports of the Structural Bioinformatics Core Facility at National Chiao Tung University.

**FUNDING**

National Science Council, partial support of the Aim for Top University (ATU) plan by Ministry of Education and National Health Research Institutes (NHRI-EX100-10009PI). Funding for open access charge: National Science Council Ministry of Education and National Health Research Institutes.

Conflict of interest statement. None declared.

**REFERENCES**

1. Marrack, P., Scott-Browne, J.P., Dai, S., Gapin, L. and Kappler, J.W. (2008) Evolutionarily conserved amino acids that control TCR–MHC interaction. *Annu. Rev. Immunol.*, 26, 171–203.
2. Bhasin, M., Singh, H. and Raghava, G.P. (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics*, 19, 665–666.
3. Brusic, V., Rudy, G. and Harrison, L.C. (1998) MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res.*, 26, 368–371.
4. Vita, R., Zarebski, L., Greenbaum, J.A., Enami, H., Hoof, I., Salimi, N., Damle, R., Sette, A. and Peters, B. (2010) The immune epitope database 2.0. *Nucleic Acids Res.*, 38, D554–D564.
5. Rudensky, A., Preston-Hurlbut, P., Hong, S.C., Barlow, A. and Janeway, C.A. Jr (1991) Sequence analysis of peptides bound to MHC class II molecules. *Nature*, 353, 622–627.
6. Hammer, J., Valsasmini, P., Tolba, K., Bolin, D., Higelin, J., Takacs, B. and Sinigaglia, F. (1993) Promiscuous and allele-specific anchors in HLA-DR-binding peptides. *Cell*, 74, 197–203.
7. Schuler, M.M., Nastke, M.D. and Stevanovic, S. (2007) SYFPEITHI: database for searching and T-cell epitope prediction. *Methods Mol. Biol.*, 409, 75–93.
8. Hakenberg, J., Nussbaum, A.K., Schild, H., Rammensee, H.G., Kuttler, C., Holzhuter, H.G., Kloezel, P.M., Kaufmann, S.H. and Mollenkopf, H.J. (2003) MAPPP: MHC class I antigenic peptide processing prediction. *Appl. Bioinformatics*, 2, 155–158.
9. Peters, B. and Sette, A. (2005) Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, 6, 132.
10. Donnes, P. and Kohlbacher, O. (2006) SVMHC: a server for prediction of MHC-binding peptides. *Nucleic Acids Res.*, 34, W194–W197.

11. Altuvia, Y., Sette, A., Sidney, J., Southwood, S. and Margalit, H. (1997) A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Hum. Immunol.*, 58, 1–11.

12. Kumar, N. and Mohanty, D. (2007) MODPROPEP: a program for knowledge-based modeling of protein-peptide complexes. *Nucleic Acids Res.*, 35, W549–W555.

13. Apweiler, R., Martin, M.J., O’Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., Barrett, D., Bely, B., Bingley, M., Binns, D. et al. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, 38, D142–D148.

14. Chen, Y.C., Lo, Y.S., Hsu, W.C. and Yang, J.M. (2007) 3D-partner: a web server to infer interacting partners and binding models. *Nucleic Acids Res.*, 35, W561–W567.

15. Rudolph, M.G., Luz, J.G. and Wilson, I.A. (2002) Structural and thermodynamic correlates of T cell signaling. *Annu. Rev. Biophys. Biomol. Struct.*, 31, 121–149.

16. Rudolph, M.G. and Wilson, I.A. (2002) The specificity of TCR/pMHC interaction. *Curr. Opin. Immunol.*, 14, 52–65.

17. Ponomarenko, J.V. and Bourne, P.E. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct. Biol.*, 7, 64.

18. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, 14, 1188–1190.

19. Bhattacharya, D. and Thio, C.L. (2010) Review of hepatitis B therapeutics. *Clin. Infect. Dis.*, 51, 1201–1208.

20. Dienstag, J.L. (2008) Hepatitis B virus infection. *N. Engl. J. Med.*, 359, 1486–1500.

21. Bartenschlager, R. and Schaller, H. (1988) The amino-terminal domain of the hepadnaviral P-gene encodes the terminal protein (genome-linked protein) believed to prime reverse transcription. *EMBO J.*, 7, 4185–4192.

22. Bartenschlager, R. and Schaller, H. (1992) Hepadnaviral assembly is initiated by polymerase binding to the encapsidation signal in the viral RNA genome. *EMBO J.*, 11, 3413–3420.

23. Mareeva, T., Martinez-Hackert, E. and Sykulev, Y. (2008) How a T cell receptor-like antibody recognizes major histocompatibility complex-bound peptide. *J. Biol. Chem.*, 283, 29053–29059.

24. Miyata, M. (2010) Unique centipede mechanism of Mycoplasma gliding. *Annu. Rev. Microbiol.*, 64, 519–537.

25. Miller, P.J., Paz Y., Conti, B., Riddle, D., Appella, E. and Collins, E.J. (2007) Single MHC mutation eliminates enthalpy associated with T cell receptor binding. *J. Mol. Biol.*, 373, 315–327.