Peptide variation governing identity of organisms

Wataru Honda honda@kuicr.kyoto-u.ac.jp
Shuichi Kawashima shuichi@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho Uji, Kyoto 611-0011, Japan

Keywords: MHC, peptide, virus, MHC class I binding motif, KEGG/GENES, VGENES

1 Introduction

Although the immune system in organisms exists to demarcate strictly self from nonself, the boundary have been remaining ambiguous. While an immense number of researches so far have revealed epitopes only on some famous proteins derived from pathogen, comprehensive analysis of self-nonself discrimination have never been carried out. Here we focused on human immune system, and compared peptide variations between human and viruses known to infect human. In a human immune response, major histocompatibility complex (MHC) class I molecules play a pivotal role. Viral proteins translated in a host cell, so called endogenous antigens, are degraded into peptides composed of 8-10 amino acids [1] via the ubiquitin-proteasome pathway. Many of these peptides are loaded and presented by MHC class I molecules on the host cell surface. On the cell surface, MHC class I molecules with loaded viral derived peptides are recognized by Killer T lymphocyte as a token of the viral infected cell [2] . Additionally, these peptides share certain patterns of sequences called MHC class I binding motifs [3] . This well-known fact explains that merely 8~10 residues of amino acid sequences are enough in length to be discriminated from host protein fragments. Here we analyze all 8~10-mer peptide patterns in all virus proteins and survey the frequency to match known motifs. Our results clarify that many of peptides observed only in viruses have a feature to be bound on MHC class I molecules.

2 Datasets and Methods

We selected 659 virus genomes from the KEGG/VGENOME. All human protein sequences and all protein sequences from viruses known to infect human were obtained from the KEGG/GENES and VGENES. We extracted all patterns of 8~10-mer peptides in the virus and human proteins and listed patterns observed only in viruses. Known MHC class I binding motifs were represented in regular expressions based on anchor and auxiliary anchor residues data from SYFPEITHI database[4]. Anchor and auxiliary anchor residues are defined by the score which is calculated by the frequency of amino acids in the respective position in aligned peptides.

Table.1

	Number of amino acids	Number of proteins
H.sapiens	8,359,195	16,468
viruses	3,842,080	24,745

3 Results and Discussions

As shown in the Figure.1, the degree of coincidence in oligopeptide sequence patterns between *H.sapiens* and viruses decline markedly around 5~7-mers. Between 8~10-mers, known as suitable lengths for MHC class I groove, almost all the oligopeptides derived from virus protein sequences are unique for viruses.

Figure.1

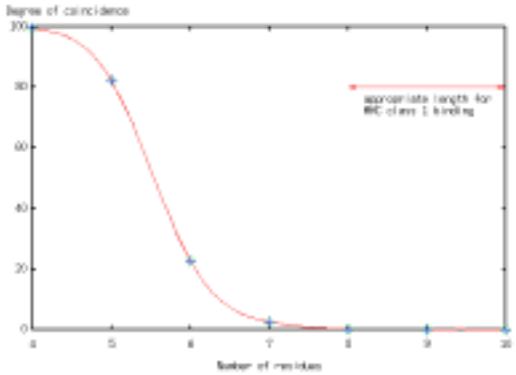

Table.2

	9residues	10residues
All oligopeptides in virus	3,075,199	3,102,211
The number of Oligopeptides unique in virus	3,071,667 99.9%	3,100,029 99.9%
The number of Motif sharing	887,029 28.9%	1,532,018 49.4%

Next, we examined matching of oligopeptides unique for virus and MHC class I binding motifs. In 120 motifs we used, there are some clear features. For instance, hydrophobic residues are preferred in the carboxy terminus, and aromatic residues are often seen nearby the amino terminus. These allocations of amino acids in the motifs are suitable for binding to MHC hydrophobic groove. In the case of 10-mers, we found almost half of the peptide sequences observed only in virus proteins shared common features with MHC class I binding motifs. Although several motif patterns have a possibility to overestimate the number of peptide matching motifs because of the limitation of our regular expressions, this result shows us that viruses are encoded by totally different amino acid sequences compared with human. This difference is remarkably interesting immunologically.
Additionally, we compared peptide variation of other organisms including three prokaryotes and seven eukaryotes with that of *H.sapiens*. As shown in Figure.2, from virus to *Drosophila melanogaster*, variation of peptides longer than 7 residues are totally different from that of *H.sapiens*. Actually, 99.9% of peptide variations in the bacteria are unique for them, and 99.53% of peptide variations in even *Saccharomyces cerevisiae* that is the most primitive eukaryote in our analysis are their own. While 97.94% of peptide variations in *Ciona intestinalis* known as a lower chordate are unique for them and as for *Mus musculus*, the ratio of unique peptide decreases to 63.7%. Therefore, this figure shows the fact that the number of peptide variation in common with *H.sapiens* have dramatically increased after emergence of protochordate in the process of evolution. Adaptive immune system encompassing MHC molecules also had followed the emergence of the vertebrate. In 2002, draft genome sequence of *C.intestinalis* was published. In this genome, MHC molecules are not present but sequence appearing as origin of MHC was found. The fact above is suggesting that requirement of 7~8-mer peptides for adaptive immune system to demarcate self from nonself is related to advent of adaptive immune system.
In this report, we examined matching of MHC class I binding motifs with peptide derived from human and viruses. Analysis of other antigen presenting molecules, for example MHC class II , are ongoing. To integrate pattern recognition of these molecules will lead us to thorough comprehension of immune system strategy of exclusion, that is, self-nonself discrimination. Moreover, analysis of peptide variation in many other organisms would provide us important evidence to clarify the relationships between evolutionally selection pressure in protein sequence in the context of host-parasite interaction and development of adaptive immune system.

Figure.2

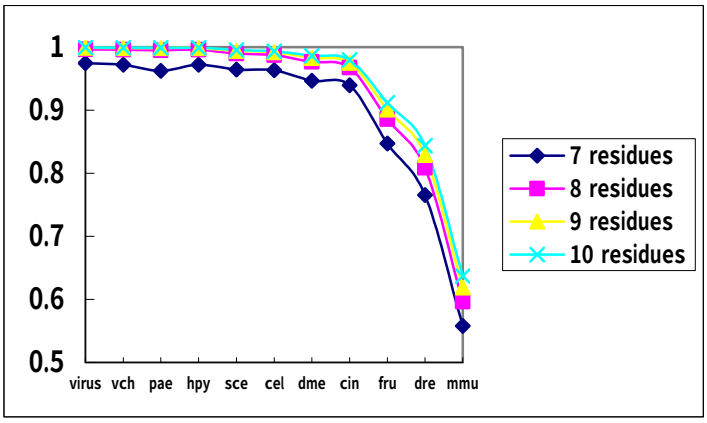

Fig.2: Horizontal axis represents just about evolutionary distance from *H.sapiens*, and vertical axis represents degree of coincidence between various lengths of peptides of each organism and that of *H.sapiens* with the same length.
vch :*Viblio cholerae*
pae :*Pseudomonas aeruginosa*
hpy :*Helicobacter pylori*
sce :*Saccharomyces cerevisiae*
cel :*Caenorhabditis elegans*
dme :*Drosophila melanogaster*
cin :*Ciona intestinalis*
fru :*Fugu rubripes*
dre :*Danio rerio*
mmu :*Mus musculus*

References

[1] Lauvau G, Kakimi K, Niedermann G, Ostankovitch M, Yotnda P, Firat H, Chisari FV, van Endert PM. Human transporters associated with antigen processing (TAPs) select epitope precursor peptides for processing in the endoplasmic reticulum and presentation to T cells.J Exp Med. 1999 Nov 1;190(9):1227-40.
[2] Rock KL, Gramm C, Rothstein L, Clark K, Stein R, Dick L, Hwang D, Goldberg AL. Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules.Cell. 1994 Sep 9;78(5):761-71.
[3] Bouvier M, Wiley DC. Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules. Science. 1994 Jul 15;265(5170):398-402.
[4] Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. 1999 Nov;50(3-4):213-9. Review.