**MHCcluster**, and method for functional clustering of MHC molecules

Morten Nielsen, Ole Lund and Claus Lundegaard.

Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, Kemitorvet, Lyngby 2800, Denmark

Identification of peptides binding to major histocompatibility complexes (MHC) is a critical step in the understanding of T cell immune responses to pathogenic infections and cancer. The human MHC genomic region (HLA) is extremely polymorphic comprising several thousand alleles, many encoding a distinct MHC molecule. The potentially unique specificity remains uncharacterized for the vast majority of MHC molecules. Likewise, for non-human species including life-stock animals, only a minor fraction of the known MHC class I molecules (MHCI) have been characterized. Here, we describe a tool that will cluster a set of MHCI variants based on their predicted binding specificity. The tool allows for identification of functional similarity between MHCI variants, and provides a graphical and highly intuitive tree-based and heat-map visualization of the functional relationship between MHC molecules.

In short, the **MHCcluster** server allows the user to select a set of MHCI alleles of interest including the option of uploading a set of full-length MHCI protein sequences, and the server return an unrooted tree and a heat-map visualizing the functional similarities between the MHC molecules. The vehicle underlying the **MHCcluster** server is the **NetMHCcons** prediction method [1], which is a consensus of the **NetMHC, NetMHCpan** and **PickPocket** MHC class I peptide binding prediction methods [2-4]. For each user defined MHCI allele, the **MHCcluster** method predicts binding to a set of predefined peptides. Next, the similarity between any two MHC molecules is estimated from the correlation between the union of the top 10% strongest binding peptides for each allele. This similarity is 1 if the two molecules have a perfect binding specificity overlap and -1 if the two molecules share no specificity overlap. Finally is the distance between two molecules defined as 1 – similarity. The distance matrix is converted to a distance tree using UPGMA clustering. To estimate the significance of the MHC distance tree, a large set distance trees is generated using the bootstrap method, and a final tree is summarized in the form of a “greedy” consensus tree with corresponding branch bootstrap values.

![Figure 1](image)

**Figure 1.** Functional clustering for prevalent HLA-A and B molecules using the **MHCcluster** method. The clustering is based on 38 prevalent HLA A and B alleles and is calculated from a set of 50,000 random natural 9mer peptides using 100 bootstrapped distance trees. The left panel gives the consensus tree, and the right panel a heat-map representation of the consensus distance matrix. In both panels is the location of the 12 supertype representatives highlighted. The arrows indicates the HLA-A*68:02 an HLA-A*30:01 alleles discussed in the main text.
Figure 1 shows the tree and heat-map obtained when applying the **MHCluster** method to a set of 38 prevalent HLA-A and B alleles. From the figure, it is clear that the method reproduces the 12 conventional 12 HLA supertypes [5]. However, it is also apparent from the figure that the HLA-A and B molecules contain specificities that are not well characterized by these common 12 supertypes. In particular, it is clear that both the A3 and B7 “supertype clusters” consist HLA molecules with highly divergent specificities. Note, also how the method is capable of correctly identifying the A3-like specificity of the HLA-A*30:01 alleles [6], and the mixed specificity (between A26 and A2) of the HLA-A*68:02 allele (highlighted with arrows in the heat-map of figure 1).

![Figure 1](image1.png)

Figure 2 shows a specificity-tree of the 38 prevalent HLA-A and B alleles included in figure 1 combined with a set of 30 Patr (chimpanzee) A-alleles including the additional chimpanzee A-like MHC class I molecule, Patr-AL (left panel) and 47 Patr-B alleles (left panel). Also, here is the power of the **MHCluster** method apparent. The figure supports the merging notion [7] that chimpanzees have a reduced MHC specificity repertoire compared to humans, in particular that the chimpanzees seem to lack specificities matching the human HLA-A26, B27, B62 and to some extent the HLA-A2 supertypes.

![Figure 2](image2.png)

In conclusion, we have demonstrated that the **MHCluster** method can be used as an effective visual tool to compare functional similarities between large sets of MHC molecules. The method is highly flexible and allows the use to analyze any MHCI variant of interest. **MHCluster** is available at [www.cbs.dtu.dk/services/MHCluster](http://www.cbs.dtu.dk/services/MHCluster).