Immunology
The Journal of cells, molecules, systems and technologies

British Society for immunology

IMMUNOLOGY ORIGINAL ARTICLE

Development of a novel clustering tool for linear peptide sequences

Sandeep K. Dhanda,[1] Kerrie Vaughan,[1] Veronique Schulten,[1] Alba Grifoni,[1] Daniela Weiskopf,[1] John Sidney,[1] Bjoern Peters[1,2] and Alessandro Sette[1,2]

[1]*Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA,* and [2]*Department of Medicine, University of California, San Diego, CA, USA*

doi:10.1111/imm.12984
Received 25 May 2018; revised 28 June 2018; accepted 11 July 2018.

Correspondence: Alessandro Sette and Sandeep Kumar Dhanda, Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA, USA. Emails: alex@lji.org (AS) and sdhanda@lji.org (SKD)
Senior author: Alessandro Sette (AS)

Summary

Epitopes identified in large-scale screens of overlapping peptides often share significant levels of sequence identity, complicating the analysis of epitope-related data. Clustering algorithms are often used to facilitate these analyses, but available methods are generally insufficient in their capacity to define biologically meaningful epitope clusters in the context of the immune response. To fulfil this need we developed an algorithm that generates epitope clusters based on representative or consensus sequences. This tool allows the user to cluster peptide sequences on the basis of a specified level of identity by selecting among three different method options. These include the 'clique method', in which all members of the cluster must share the same minimal level of identity with each other, and the 'connected graph method', in which all members of a cluster must share a defined level of identity with at least one other member of the cluster. In cases where it is not possible to define a clear consensus sequence with the connected graph method, a third option provides a novel 'cluster-breaking algorithm' for consensus sequence driven sub-clustering. Herein we demonstrate the tool's clustering performance and applicability using (i) a selection of dengue virus epitopes for the 'clique method', (ii) sets of allergen-derived peptides from related species for the 'connected graph method' and (iii) large data sets of eluted ligand, major histocompatibility complex binding and T-cell recognition data captured within the Immune Epitope Database (IEDB) with the newly developed 'cluster-breaking algorithm'. This novel clustering tool is accessible at http://tools.iedb.org/cluster2/.

Keywords: Allergy; Antigens/Peptides/Epitopes; Bioinformatics>; MHC/HLA; Viral.

Introduction

Immune epitope data pertaining to specific antigenic regions often must be assessed from data generated using several different, although largely overlapping and/or highly similar, sequences. This is typically because the exact sequences used to probe a specific region can differ between different laboratories, studies and methodologies, reflecting, for example, the use of different isolates, or alternative frame shifts.

Similarly, epitope identification strategies often involve the use of overlapping peptide sets, but can also include the testing of candidate peptides identified from major histocompatiblity (MHC) binding predictions,[1,2] as well as the elution of naturally processed MHC ligands, which often occur as nested sequences.[3] Epitope-related sequences might also originate from homologous variants, such as different viral strains, homologous sequences found in related species, repeat sequences or single-nucleotide polymorphisms.[4] Different sequence variants might bind to closely related human leucocyte antigens (e.g. human leucocyte antigen supertypes)[5–7] and, in addition, certain sequences might reflect 'hot spots' of immunodominant regions that are preferentially processed and bound.[8]

To capture this diversity, the Immune Epitope Database (IEDB) was specifically designed to curate each molecular structure (unique epitope) and associated metadata as a distinct entity.[9] However, to perform aggregate analyses on these individual data, there is a clear need for improved clustering strategies that facilitate biologically meaningful response analyses.

The IEDB has developed a tool, called IMMUNOME BROWSER,[10,11] that enables visualization of the immune reactivity of related epitope sequences by displaying reported epitope reactivity along the linear sequence of a specified antigen or reference proteome. The tool is powerful for visualization purposes, but is not well suited to resolve redundancy of sequences in separate and specific clusters, and is also not ideal to analyse sequences that may cluster together, but that are in fact derived from unrelated antigens. Indeed, the identification of regions of homology between unrelated sequences has been associated with a number of important biological phenomena, like molecular mimicry, allergen cross-reactivity, and microbiome modulation of adaptive immune responses.[12–16] Clustering epitope reactivity may also be of interest to generate epitope 'megapools' of minimal composition by elimination of potentially redundant sequences based on a specific degree of protein similarities.[17–20]

Clustering biological data, such as DNA or protein (amino acid) sequences, is a well-studied problem. However, making 'biologically meaningful' clusters based on epitope response data presents unique challenges.[21,22] Here, a clustering approach would be biologically meaningful if it could provide connectivity of the entities in the clusters, provide a cluster representative sequence, and was based on sequence characteristics of each data set (unsupervised strategy). Several attempts have been made in the past to address this issue from different angles. For example, Hammock[23] was developed to identify consensus motifs in large data sets, and PepServe[24] to cluster peptides based on their physiochemical properties. Another application, UCLUST,[25] performs rapid clustering and implements several previously identified clustering algorithms, and Andreatta *et al.*[26,27] developed a tool based on Gibbs sampling. Overall, most of these tools cluster based on the physiochemical properties of peptide sequences or by defining and focusing on shared motifs.

However, none of these tools provide a complete connectivity of the peptides within a cluster and do not generate a clear consensus sequence representing each cluster. The need for elucidation of complete connectivity is key when evaluating cross-reactive responses or epitope homology between different species. Vice versa, generating a clear consensus sequence is important when trying to represent a group of sequences sharing a defined level of identity with a single unique sequence that maintains the same capacity to induce an immune response as the constituent epitopes, such as has been done with the megapool approach.[20]

From a computational standpoint, two main clustering approaches exist: cliques[28] and clustering by connected graph.[29–31] In the clique approach, sequences are clustered into cliques, defined as a group of sequences related to all other members at (or above) the exact desired level of identity. Although this approach allows definition of clear consensus sequences, the main drawback is that the same sequence may be found in multiple cliques, leading to inflation of the number of clusters and ambiguity within clusters. That is, the clique approach provides greater inclusiveness, per se, but it also hinders cluster size reduction, which is desired for immune epitope clustering.

In the connected graph approach, all peptide sequences that are identical to a certain pre-specified percentage level, for example 70%, are clustered together. In this case, any member of the cluster will be at least 70% identical to at least one member of the cluster. Although this approach clusters peptides only once, thereby avoiding redundancy, a drawback is that other members of the cluster might, and often are, related by levels of identity much lower than 70%, which is not suitable when you need to have a clear identity cut-off. In addition, this method cannot always resolve clearly consensus sequences.

To address these issues and apply them in different immunological contexts, we developed a clustering tool that is flexible enough to take into consideration different strategies, and generate consensus sequences from clusters by identifying the most commonly occurring residues at each position in a given length of amino acids based on a multiple sequence alignment. This novel tool allows the user to select different clustering algorithms and strategies depending on the biological question. We illustrate the performance of this method through the resolution of clusters of various epitope data sets available within the IEDB, and with different immunological purposes.

Methods

Dengue virus data set

The dengue virus (DENV) epitope data set was derived from a comprehensive analysis of human leucocyte antigen-matched $CD4^+$ T-cell responses in a Sri Lankan cohort, allowing for the selection of 363 epitopes restricted by 15 DRB1 alleles with phenotypic frequency > 4% in the Sri Lankan population, as previously reported.[18] The set of epitopes covers serotype-specific epitopes from DENV serotypes 1, 2, 3, 4 (all 10 proteins), as well as epitopes conserved between those serotypes. For each epitope a corresponding spot-forming colony (SFC) value is provided based on *in vitro* expansion of DENV-specific T cells and interferon γ enzyme-linked immunospot assay to rank the best epitopes with highest SFC.[32]

Rat and mouse allergen data sets

Allergen epitope data sets were assembled from previous data characterizing rat and mouse antigens. The rat epitopes are derived from the rat allergen *Rat n* I, a major urinary protein, and constitute a set of 19 peptides found to be T-cell-reactive in rat-allergic patients.[33] In addition, our group recently identified 23 peptides from the major mouse allergen, *Mus m* I.[34]

IEDB data sets used for algorithm and tool development

Additional epitope data sets were compiled from the IEDB (http://iedb.org), which curates a vast set of published T-cell response data, as well as MHC class I and II binding and ligand elution (MHCLE) data.

To retrieve relevant sets of epitopes, a query was performed from the IEDB home page targeting 'T cell Assays' (using check boxes in the 'Assay' search panel), and including both positive and negative peptides. As no other selection criteria were included in this initial query, these data represent both human and non-human data. The query was performed on February 2017, at which point the IEDB contained a total of 115 228 peptides tested (positive and negative) in T-cell response assays. The full set of T-cell response data was downloaded to Excel using the Assays tabs (Export T cell Assays Results).

To retrieve MHC binding and MHCLE data, a similar query was performed targeting (separately) each of these data sets (using the 'Find' feature within the Assay search pane). These queries revealed a total of 64 312 peptides for MHC binding data and 139 614 for MHCLE. From the exported Excel tables, we selected only linear peptides and further categorized each peptide data set as associated with either class I or class II MHC molecules. Because of the unique immunobiology of class I and class II, these data sets were analysed separately in clustering algorithms. The composition and breakdown of the resulting data set is summarized in the Supplementary material (Table S1).

Generation of sequence identity matrices

In-house Python scripts were used to calculate the sequence identity between each peptide pair in each data set. When calculating the identity between any peptide pair, one peptide is aligned to a second peptide in all the possible frames and the number of residues matching is counted for each frame (including the offsets). The alignment with the largest number of matches was used for identity calculations. To scale the level of sequence identity in the range 0–1 (meaning from 0% to 100%), we divided the maximum number of matches in the alignment by the length of the smaller peptide (see eqn i).

$$\text{Sequence identity} = \frac{\text{Maximum number of residue matching in an alignment}}{\text{Length of smaller peptide}}$$

Building a graph network connecting clusters and cliques

The sequence identity matrix calculated as above was then used to create an undirected graph network. Here, an undirected graph network works as a spacial/graphical representation of the data, where nodes within the network represent the clustered peptides and the edges represent the sequence identity defined above for a given threshold. The undirected graph network was generated with the Networkx version 1.11 library of Python language.[35] Using this, all the fully interconnected nodes (called cliques) and the connected components (called a connected graph) were derived. Here, a clique is defined as a set of peptides, wherein all the peptides share sequence identity higher than the threshold with all other members of the clique. By contrast, each connected graph represents a group of peptides, where each peptide shares sequence identity higher than the threshold to at least one of the other peptides in that connected graph. As a result, one peptide can be present in more than one clique, whereas each cluster contains only unique peptides, and no peptide was present in two connected graphs.

Generation of alignments and consensus sequences from peptide data sets

Sequence alignments were generated using Clustal Omega version 1.2.4 for high-quality alignments based on several benchmark studies.[36] From the alignment output, the consensus sequence is derived using the Biopython package.[37] Here, we used a threshold of 51% occurrences to assign a residue at a specific position. An X was assigned for ambiguous residues. These alignments were used to derive consensus sequences from each set of peptides in a cluster.

Cluster breaking algorithm applied on connected graph method and web implementation

A cluster breakdown algorithm was developed scanning each connected graph for the longest possible consensus sequence using an undirected graph network constructed from a set of peptides (as described above). The process starts by extracting the most connected peptide (centre node, N0) in a graph, including all the directly connected peptides (neighbours, N1) of that peptide. Then the next set of peptides (N2) is extracted, and connected to centre node (N0) through N1. Each peptide in N2 is then aligned with a consensus sequence, and peptides that share sequence identity less than the threshold of the consensus sequence are removed. The next alignment is created with the remaining peptides from N2 and all the peptides from N1 (including centre node). A new consensus is then derived from this alignment. The process is iterated until the clear longest consensus sequence that can represent the maximum number of peptides in a connected graph is obtained. These peptides are then removed from the network and the process is repeated in a recursive fashion for each sub-graph. The complete cluster-break pipeline was implemented in the DJANGO framework and can be accessed by users as an online tool. JAVASCRIPTS were used to support the visualization in the server.

Results

Use of a network graph approach to generate cliques and connected graph

Our overall goal was to develop an online tool that would allow the user to apply different clustering strategies to a list of epitope or peptide sequences. Specifically, we used two main methodologies to cluster sequences, one by cliques[28] and the other by connected graph.[29–31]

For this we developed a simple algorithm to cluster sequences using a network graph approach.[35] A network was created using a matrix, where each row represents a peptide pair and their sequence identity. This method creates graphical networks of sequences in which all members of the cluster are homologous to all other members of the connected graph at a specified identity level (see Supplementary material, Fig. S1). Here circles, or nodes, represent the peptides and the solid line is the identity above threshold between any peptide pair. From these graphical networks, peptide sequences that are connected in the network (called the connected graph) can be extracted. The peptides labelled as A, B, C and D are connected graphs shown in Supplementary material (Fig. S1). As an alternative, all the peptides that are fully interconnected (referred to as cliques) can be extracted. The cliques are represented by peptides A, B and C (clique 1) or peptides B, C and D (clique 2) in the example above, as peptides A and D are not connected (see Supplementary material, Fig. S1). Peptides that do not share any homology with remaining peptides (e.g. E in our example), are referred to as singletons. Each peptide will be present in only one connected graph, but can be present in multiple cliques. Singletons are represented the same way in the connected graph and clique approaches.

Application of the clique method to refine a previously identified DENV-derived epitope megapool

Next, we evaluated the performance of our clustering methods using different approaches suited to the different immunological questions behind them. As a first example we used a set of 363 DENV-derived epitopes for which a 'megapool' of peptides was previously generated based on experimental data from human leucocyte antigen-matched CD4$^+$ T-cell responses in a Sri Lanka DENV-endemic population.[18] In this case, we sought to reduce the number of peptides included in the megapool to facilitate experimental approaches without sacrificing biological activity. For this purpose, to obtain clear consensus sequences encompassing peptides achieving the same identity level cut-off despite the possibility of having cluster redundancy, the clique method was applied.

The results are summarized in Table 1, and also shown in the Supplementary material (Tables S2 and S3). Among the original 363 DENV epitopes, 305 were unique in the data set and at 70% sequence identity they were grouped into 198 cliques (see Supplementary material, Table S2). Of those 198 cliques, 118 cliques contained two or more epitopes and, in total, covered 225 epitopes. The remaining 80 epitopes did not share any similarities with the rest of the epitope set and hence were assigned as singletons (clusters containing a single sequence).

Examining the best epitopes (i.e. those with highest magnitude of response, in terms of SFCs) contained in each of the 118 cliques, we observed that this corresponded to 83 unique epitopes (some epitopes were selected as the best epitope in more than one clique). Hence, in the end, the clique clustering generated a set of 80 singletons and 83 unique 'best of the clique' epitopes, or a total of 163 epitopes that could be used to represent the 363 DENV 'megapool' previously identified, equivalent to a 55% reduction in the number of epitopes needed (see Supplementary material, Table S3).

Application of the clustering method in mouse versus rat allergen data

As another demonstration of the peptide-clustering application we analysed data derived from a study of mouse allergens. To date, one major mouse allergen, *Mus m* I, has been identified, to which our laboratory has mapped 23 different T-cell epitopes.[34] A previous study had identified 19 T-cell epitopes[33] from the major rat allergen, *Rat n* I, which exhibits significant homology to *Mus m* I

Table 1. Summary of the dengue virus data clustering using clique approach

Feature	Counts
Total peptides	363
Unique peptides	305
Total cliques	198
Peptides covered in cliques with two or more peptides	225
Cliques with two or more peptides	118
Unique peptides selected in cliques with two or more peptides	83
Singleton cliques	80
Total peptides selected	163

(65% identity, 78% similarity), making these data sets good candidates for cluster analysis.

Here the purpose of the clustering task was to judge whether each epitope derived from the mouse allergen would co-cluster with sequences derived from the rat allergen. For this, the clique approach was not appropriate, because the evaluation of the association requires that each sequence be represented in only one cluster or it will inflate the amount of sequences shared between the two species. Accordingly, the connected graph strategy was used (Table 2).

The analysis revealed that the 23 overlapping *Mus m* I epitopes correspond to 20 unique epitopes, as three epitopes clustered with other epitopes within the same set (70% identity threshold). Of the 23 *Mus m* I epitopes, eight clustered with *Rat n* I peptides. Six of the eight *Mus m* I epitopes that clustered with *Rat n* I peptides were among the top nine in terms of T-cell response magnitude (conversely, only two peptides were in the bottom 14; $P = 0{\cdot}023$ by exact Fisher test). This analysis supported the notion that epitope allergens conserved in multiple species may elicit a more dominant immune response, probably due to increased/repeated exposure by the various homologous allergens.

Analysis of MHCLE data reported in the IEDB

We then tested the two clustering approaches on a larger data set consisting of naturally processed peptides eluted from MHC, using MHCLE data reported in the IEDB. MHCLE peptides are of variable size, and often the sequences eluted are largely nested and/or overlapping.[38,39] For this reason, this type of data was ideal for further validation of our clustering tool and to test the applicability on larger-scale data.

The IEDB has MHCLE data for over 100 000 different peptide sequences. Sets of class I and class II naturally processed peptides were generated by querying the IEDB (as described in the Materials and methods section) and the extracted sequences were clustered using the network graph approach.[35] As an alternative, all the peptides that were fully interconnected (cliques) were also extracted.

Because of their unique immunobiology, class I and II ligands were analysed separately. We expected that the effect of sequence clustering would be more pronounced for class II, because of the well-known 'ragged end' nature of natural class II ligands,[40] allowing ligands of variable length. Indeed, we found that of a total of 105 642 class I and 33 757 class II peptides, 57 455 and 28 523 (respectively) were clustered (Table 3), and 48 187 and 5234 were singletons (not clustering with any other sequence) for class I and class II, respectively. Hence, almost half (46%) of the class I peptides were identified as singletons compared with only 16% for class II; 84% of the class II sequences were found in clusters (Table 3).

Using the connected graph (where different clusters can be connected through single peptide, see Supplementary material, Fig. S1) clustering method, we found 11 932 and 4683 clusters, respectively, for classes I and II. The average cluster sizes were 4·82 and 6·09 peptides, respectively, indicating that this method can effectively cluster the data. In contrast, we observed 36 732 and 25 941 cliques, and the average clique sizes were 2·78 and 57·36 for the two data sets in the clique method. Hence, there are many more cliques than clusters in class II and the average size of cliques (57·36 peptides, on average) is much greater than the average size of clusters (6·09 peptides, on average). This is due to the fact that many peptides are found in multiple cliques, and indicates that the clique approach was ineffective for data set reduction. As a result, it was not considered further for this particular application.

We then applied the clustering method on the same larger data sets evaluating the distribution of connected graph clusters generated with the class II MHCLE data as a function of the number of peptides contained in each cluster. The distribution is shown in Fig. 1 (red line) for the top 10 clusters (solid red line) and also summarized in Table 4 ('Raw Data' column). Table 4 shows that a single cluster contained more than 1000 peptides and a total of 10 clusters contained between 100 and 999 peptides. Similarly, we observed 21 clusters with 50–99 peptides, 55 clusters with 30–49 peptides, and 439 clusters with 10–29 peptides. Finally, the majority of the clusters (4157 out of 4683: (89%) encompassed fewer than 10 peptides.

However, manual inspection of the 'large membership' clusters (i.e. clusters with > 1000 peptides) revealed that they contained rather heterogeneous sequences with very little actual sequence similarity, often linked by the presence of very short sequence stretches of three or four residues. For several of the large clusters a clear consensus sequence could not be defined (data not shown). We reasoned that the largest and smallest sequences contained in the data pollute the clustering assignments, and so sought to alleviate this problem by instituting appropriate peptide length/size filters, and developing an additional feature on the clustering method.

Resolution is improved by excluding tail-end sequence sizes

To implement an appropriate filter system for peptide length, we reasoned that most biologically relevant data (MHC ligands, MHC binding or T-cell recognition) are associated with epitopes of between 8 and 25 residues.[41–44] To test this assumption, we generated data sets corresponding to class I and class II peptides in the MHLE, MHC binding and T-cell response data.

Across the different data sets, there were few peptides with length less than 8 residues (< 8mer) (Table 5). The

Table 2. The clustering tool output with mouse and rat allergic data

Cluster number	Peptide number	Alignment	Position	Description	Peptide
1	Consensus	TFQLMXLYGRXXDLSSDIKEKFAKLCEA	–	–	–
1	1	TFQLMVLYGRTKDLSSDIKE--------	1	Rat Pep17	TFQLMVLYGRTKDLSSDIKE
1	2	-----GLYGREPDLSSDIKERFA-----	6	Mus Pep3	GLYGREPDLSSDIKERFA
1	3	-------YGREPDLSLDIKEK-------	8	Mus Pep7	YGREPDLSLDIKEK
1	4	--------GRTKDLSSDIKEKFAKLCEA	9	Rat Pep9	GRTKDLSSDIKEKFAKLCEA
2	Consensus	YDRYVMXHLINXKXGETFQLMXLYGRTK	–	–	–
2	1	YDRYVMFHLINFKNGETFQL--------	1	Rat Pep19	YDRYVMFHLINFKNGETFQL
2	2	------AHLINEKDGETFQLM-------	7	Mus Pep9	AHLINEKDGETFQLM
2	3	--------LINFKNGETFQLMVLYGRTK	9	Rat Pep12	LINFKNGETFQLMVLYGRTK
2	4	----------NEKDGETFQLMGLY----	11	Mus Pep6	NEKDGETFQLMGLY
3	Consensus	EENGSMRVFXXHIXVLENSL	–	–	–
3	1	EENGSMRVFMQHIDVLENSL	1	Rat Pep4	EENGSMRVFMQHIDVLENSL
3	2	---GSMRVFVEHIHVLEN--	4	Mus Pep16	GSMRVFVEHIHVLEN
4	Consensus	FXXHIXVLENSLXFKFRIKE	–	–	–
4	1	FMQHIDVLENSLGFKFRIKE	1	Rat Pep6	FMQHIDVLENSLGFKFRIKE
4	2	FVEHIHVLENSLAFK-----	1	Mus Pep2	FVEHIHVLENSLAFK
5	Consensus	RXNIIDLTKTXRCLXARG	–	–	–
5	1	RDNIIDLTKTDRCLQARG	1	Rat Pep14	RDNIIDLTKTDRCLQARG
5	2	-ENIIDLTKTNRCLKA--	2	Mus Pep17	ENIIDLTKTNRCLKA
6	Consensus	GXWFSIXXASXKREKIEENG	–	–	–
6	1	GDWFSIVVASNKREKIEENG	1	Rat Pep8	GDWFSIVVASNKREKIEENG
6	2	-EWFSILLASDKREKI----	2	Mus Pep4	EWFSILLASDKREKI
7	Consensus	EEASSTGRNFNVXKINGEWHTIIL	–	–	–
7	1	EEASSTGRNFNVQKINGEWHTIIL	1	Mus Pep10	EEASSTGRNFNVQKINGEWHTIIL
7	2	----------NVEKINGEWHTIIL	11	Mus Pep13	NVEKINGEWHTIIL
8	Consensus	FVEYDGXNTFTILKTDYDXY	–	–	–
8	1	FVEYDGGNTFTILKTDYDRY	1	Rat Pep7	FVEYDGGNTFTILKTDYDRY
8	2	----DGFNTFTILKTDYDN-	5	Mus Pep5	DGFNTFTILKTDYDN
9	Singleton	TFTILKTDYDRYVMFHLINF	–	Rat Pep18	TFTILKTDYDRYVMFHLINF
10	Singleton	GIYYLNYDGFNTFTI	–	Mus Pep14	GIYYLNYDGFNTFTI
11	Singleton	KTPEDGEYFVEYDGGNTFTI	–	Rat Pep10	KTPEDGEYFVEYDGGNTFTI
12	Singleton	LENSLVLKFHTVRDE	–	Mus Pep8	LENSLVLKFHTVRDE
13	Singleton	LQSGFYSLSSLVTVP	–	Mus Pep21	LQSGFYSLSSLVTVP
14	Singleton	ENSLGFKFRIKENGECRELY	–	Rat Pep5	ENSLGFKFRIKENGECRELY
15	Singleton	EKALVSSVRQRMKCS	–	Mus Pep11	EKALVSSVRQRMKCS
16	Singleton	LEQIHVLENSLVL	–	Mus Pep1	LEQIHVLENSLVL
17	Singleton	DDVVASEALNSVWSGF	–	Mus Pep15	DDVVASEALNSVWSGF
18	Singleton	SRPFIFQEVIDLGGE	–	Mus Pep12	SRPFIFQEVIDLGGE
19	Singleton	DKETLSLEELKALLL	–	Mus Pep20	DKETLSLEELKALLL
20	Singleton	IGGPDDGVITPWQSSF	–	Mus Pep19	IGGPDDGVITPWQSSF
21	Singleton	DIKEKFAKLCEAHGITRDNI	–	Rat Pep2	DIKEKFAKLCEAHGITRDNI
22	Singleton	RELYLVAYKTPEDGEYFVEY	–	Rat Pep15	RELYLVAYKTPEDGEYFVEY
23	Singleton	ILGKLVKDYHLQFHR	–	Mus Pep18	ILGKLVKDYHLQFHR
24	Singleton	TIFISLFLLSVCYSA	–	Mus Pep23	TIFISLFLLSVCYSA
25	Singleton	EELRRLAPITSDPTE	–	Mus Pep22	EELRRLAPITSDPTE
26	Singleton	NLDVAKLNGDWFSIVVASNK	–	Rat Pep13	NLDVAKLNGDWFSIVVASNK
27	Singleton	LCEAHGITRDNIIDLTKTDR	–	Rat Pep11	LCEAHGITRDNIIDLTKTDR
28	Singleton	RIKENGECRELYLVAYKTPE	–	Rat Pep16	RIKENGECRELYLVAYKTPE
29	Singleton	ASNKREKIEENGSMRVFMQH	–	Rat Pep1	ASNKREKIEENGSMRVFMQH
30	Singleton	EEASSTRGNLDVAKLNGDWF	–	Rat Pep3	EEASSTRGNLDVAKLNGDWF

range for < 8mer was from as low as 0·09% (T-cell assay for MHC class I) to 0·63% (MHCLE class I). The majority of the peptides for MHC class I (~98%) were represented by peptides 8–15 amino acids in length for the MHCLE and MHC binding data sets, and 8–20 amino acids for peptides within the T-cell assays data set. To achieve 98% coverage for class II, the length range is higher; we consider peptides 8–25 amino acids in length for MHCLE and 9–21 amino acids for the MHC binding data sets. When we examined the size distribution of all these various ligand categories, we found that filtering the peptide data sets on the 8–25 amino acid size range captured 98% or more of all ligands, except for T-cell assays for MHC class II, where it covers up to 97·24% of the

Table 3. Statistics of peptides from MHCLE data set and their clustering at 70% sequence identity threshold

Features	MHCLE class I	MHCLE class II
No. of peptides	105 642	33 757
No. of peptides clustered	57 455	28 523
No. of singletons	48 187	5234
Singletons (% peptides)	46	16
No. of clusters	11 932	4683
Average size of cluster	4.82	6.09
No. of cliques	36 732	25 941
Average size of cliques	2.78[1]	57.361

MHCLE, major histocompatibility complex ligand elution.

[1]Number of peptides/cliques (one peptide can be present in several cliques).

'#' denotes the count of a particular feature.

Table 4. Distribution of cluster size for MHCLE data set for class II with different approaches

	Number of clusters		
Cluster size	Raw data	Length filtered data1 (8–25 residues)	Length filtered data[1] + Cluster break algorithm
≥ 1000	1	0	0
100–999	10	11	10
50–99	21	17	26
30–49	55	50	49
10–29	439	452	492
< 10	4157	4196	4335
Total	4683	4726	4912

MHCLE, major histocompatibility complex ligand elution.

[1]Length filtered data: to obtain a final list of peptide data sets where short (< 8 amino acids) and long (> 25 residues) peptides have been removed.

peptides. Accordingly, in subsequent analyses, we only considered peptides in the 8–25 residue range, so effectively removing outlier sequences from the top and bottom end of the length distribution for both class I and II (Table 5).

Next, we repeated the cluster analysis considering only peptides in the 8–25 residue range. As expected, several of the large clusters were broken down into smaller ones, and the number of clusters increased from 4683 to 4726. Maximum cluster size dropped from 1006 to 649, with only 11 clusters containing more than 100 peptides (Table 4; 'Length filtered column'). This normalizing effect was also observed in the plot representing the number of peptides in each cluster (Fig. 1, blue line). Also, as expected, the average number of peptides in each of the different clusters was decreased. In most cases a clear consensus sequence could be defined, but for several of the large clusters a clear consensus sequence remained elusive (data not shown).

In conclusion, while the data filtering increased the clustering resolution, the presence of several large and ambiguous clusters suggested that further approaches were required to maximize clustering resolution in the case of large data sets like the ones considered herein.

Development of an algorithm that allows cluster-break for clear separation

Manual inspection of the remaining large clusters revealed that in several cases, these clusters contained numerous loosely connected sub-clusters (Fig. 2a). To

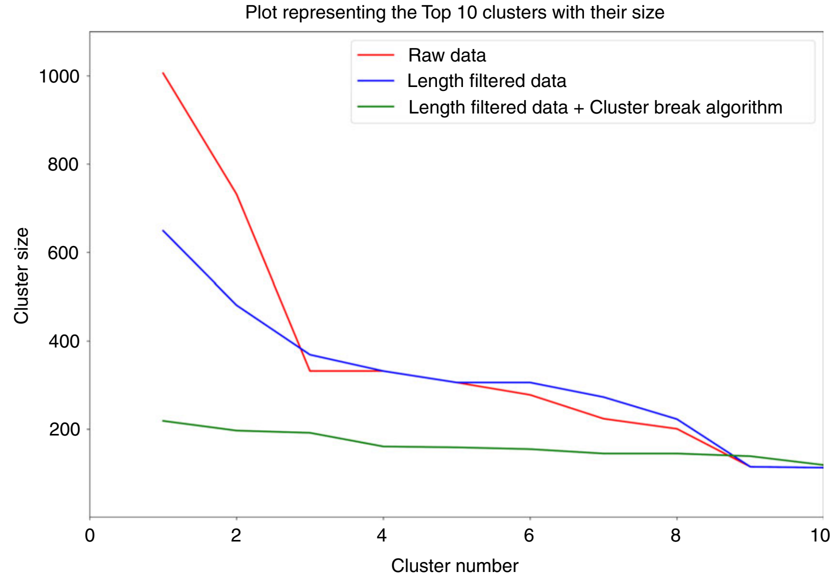


Figure 1. Plot representing the number of peptides in top 10 clusters from major histocompatibility complex class II ligand elution data.

Table 5. Distribution of peptide length in both the classes of MHCLE, MHC binding and T-cell assay data sets

Length	MHCLE I (%)	MHCLE II (%)	MHC binding I (%)	MHC binding II (%)	CD8 T cell (%)	CD4 T cell (%)
< 8	0·63	0·20	0·16	0·60	0·09	0·24
8	7·04	1·10	6·33	0·25	8·24	0·49
9	50·92	1·37	61·51	2·59	44·30	1·13
10	18·38	2·00	24·85	3·54	20·70	2·33
11	11·74	3·35	4·24	2·56	4·73	1·22
12	4·74	5·85	0·44	2·91	0·64	6·57
13	2·72	10·60	0·19	5·50	0·33	2·76
14	1·36	15·08	0·16	3·01	1·28	2·86
15	0·87	16·49	1·01	56·55	14·61	40·69
16	0·38	14·54	0·18	4·07	0·45	6·11
17	0·30	10·07	0·06	3·43	0·34	3·89
18	0·22	6·33	0·06	3·13	1·09	4·52
19	0·17	3·91	0·02	1·56	0·11	2·12
20	0·12	2·68	0·72	7·62	2·54	18·42
21	0·09	1·84	0·05	1·20	0·05	1·64
22	0·07	1·15	0·01	0·30	0·02	0·56
23	0·05	0·94	0·00	0·14	0·08	0·44
24	0·04	0·65	0·01	0·21	0·05	0·46
25	0·02	0·45	0·00	0·30	0·06	1·05
> 25	0·14	1·40	0·00	0·53	0·25	2·52
≥ 8 and ≤ 25	99·22	98·41	99·83	98·87	99·65	97·24

MHC, major histocompatibility complex; MHC I, MHC class I; MHC II, MHC class II; MHCLE, MHC ligand elution; CD8, T-cell data recognized by MHC I; CD4, T-cell data recognized by MHC II.

visualize this feature, we followed the approach of constructing an undirected graph network from peptide sequence identity matrices using a publicly available network package.[35] According to this approach, a circle represents each peptide sequence, and any peptides (circles) that share an indicated level of identity (in this case 70%) are connected by a solid line (see Supplementary material, Fig. S1).

Application of the cluster-break algorithm to various large peptide data sets

When the cluster-break algorithm was applied to the MHC class II ligand elution data set, the number of clusters increased to 4912, and the largest clusters now contained only 219 sequences (Table 4; 'Length filtered' and 'cluster break algorithm' columns). The largest sub-cluster was a 'legitimate' cluster with the consensus sequence 'AASQRMEPRAPWIEQEGPEYWDXETRXVKAHSQTH'. A further smoothing in the distribution of peptides in different clusters is also observed (Fig. 1, green line). Based on these results we surmise that the combination of the peptide size filter and the 'cluster-breaking' algorithm represents a viable clustering strategy for large data sets of this type.

An example of this representation based on a cluster composed of 99 different peptides is shown in Fig. 2(a). Next, a cluster-break algorithm was derived to further dissect clusters into sub-clusters, and achieve definition of the longest consensus sequence (Fig. 2b), as described in the Materials and methods.

To further validate the cluster-break strategy, we performed a similar analysis on the class I MHCLE data set, and MHC binding and T-cell epitopes for both class I and class II data sets. We observed no cluster with more than 1000 members (Table 6). The maximum size for any cluster was 257, found in one case in the MHC class I binding data set (Fig. 3, dashed blue line). The maximum cluster size for class I peptides in the MHCLE and TCR data sets was in the range of 50–99, whereas the range was 100–999 for the remaining data sets. Similar to the MHCLE class II data set, there were few clusters in the size ranges 30–49 or 10–29; the majority of clusters have less than 10 peptides. Figure 3 plots the top 10 clusters

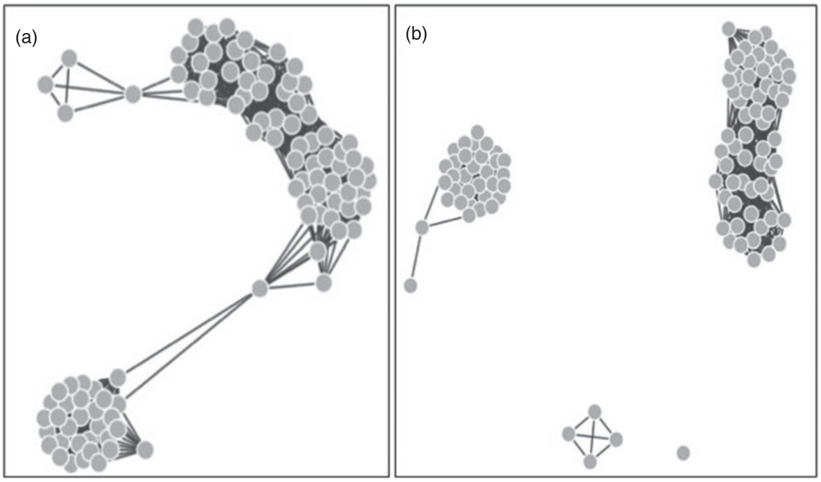


Figure 2. Example cluster visualization before (a) and after (b) cluster-breaking algorithms.

Table 6. Distribution of cluster size for different data sets

	Number of clusters (length filtered data + cluster break algorithm)					
Cluster size	MHCLE II	MHCLE I	MHC binding I	MHC binding II	CD8 T cell	CD4 T cell
≥ 1000	0	0	0	0	0	0
100–999	10	0	6	9	0	6
50–99	26	8	24	14	12	23
30–49	49	9	60	32	58	77
10–29	492	277	507	202	431	431
< 10	4335	15 842	5525	1933	4081	3421
Total	4912	16 136	6122	2190	4582	3958

MHC, major histocompatibility complex; MHC I, MHC class I; MHC II, MHC class II; MHCLE, MHC ligand elution; CD8, T-cell data recognized by MHC I; CD4, T-cell data recognized by MHC II.

from each data set after applying the cluster-break algorithms.

Further analysis of the overlap between MHC binding, MHCLE and T-cell data

We then focused on a different immunological question that can be addressed analysing large data-sets. In this case we compared MHC binding data with MHCLE and T-cell response to identify epitopes that were redundant in the three data sets and could therefore be the best candidates able to be presented and recognized by T cells.

For this purpose, we clustered all the T-cell epitopes, MHC binding and MHCLE data in pairs to analyse the identity sharing between MHC binders or MHC eluted data with T-cell response. As above, class I and class II data were analysed separately. Once we obtained the results of the clustering we further categorized each cluster as positive for T-cell assays if it contained at least one sequence associated with positive T-cell assay results, and as negative if it contained only sequences negative in T-cell assays. A cluster was classified as T-cell assay undetermined if none of its sequences was associated with either positive or negative T-cell recognition. A similar classification was performed to classify clusters according to MHC binding assay status. Unlike MHC binding and T-cell epitope data, which represent both positive and negative assays, for MHCLE data there are only positive data (MHCLEp) (see Supplementary material, Table S4), therefore it cannot be inferred whether the sequence is negative or not tested.

In the case of class I, out of a total of 2719 T-cell assay + clusters, only 515 were undetermined (19%); of the remaining clusters, 2026 were MHC + (74·5%) (Fig. 4a). Similarly, in the case of class II, 1260 (38·6%) of the T-cell assay + clusters were also MHC + (Fig. 4b). Conversely, when we examined MHC + clusters we found that 28·6% and 47·6% were also T-cell assay +, for class I and class II, respectively. Taken together, this analysis supports the notion that MHC binding is necessary, but not sufficient, for T-cell immunogenicity.

When we examined MHCLE data, we found that the vast majority of MHCLEp clusters are undetermined when it comes to T-cell assays (85·1% and 89·9% for class I and class II, respectively) (Fig. 4c,d) or MHC binding (83·5% and 91·5% for class I and class II, respectively) (Fig. 4e,f). In the few cases where the cluster could be categorized, the results were consistent with expectations, in that most (82·4% for class I to 87·6% for class

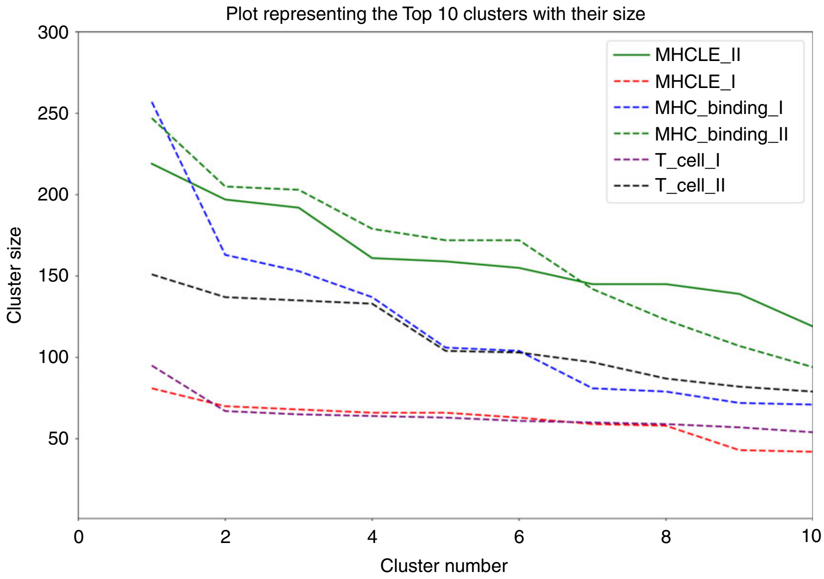


Figure 3. Plot representing the top 10 clusters in different data sets after applying cluster-break algorithm.

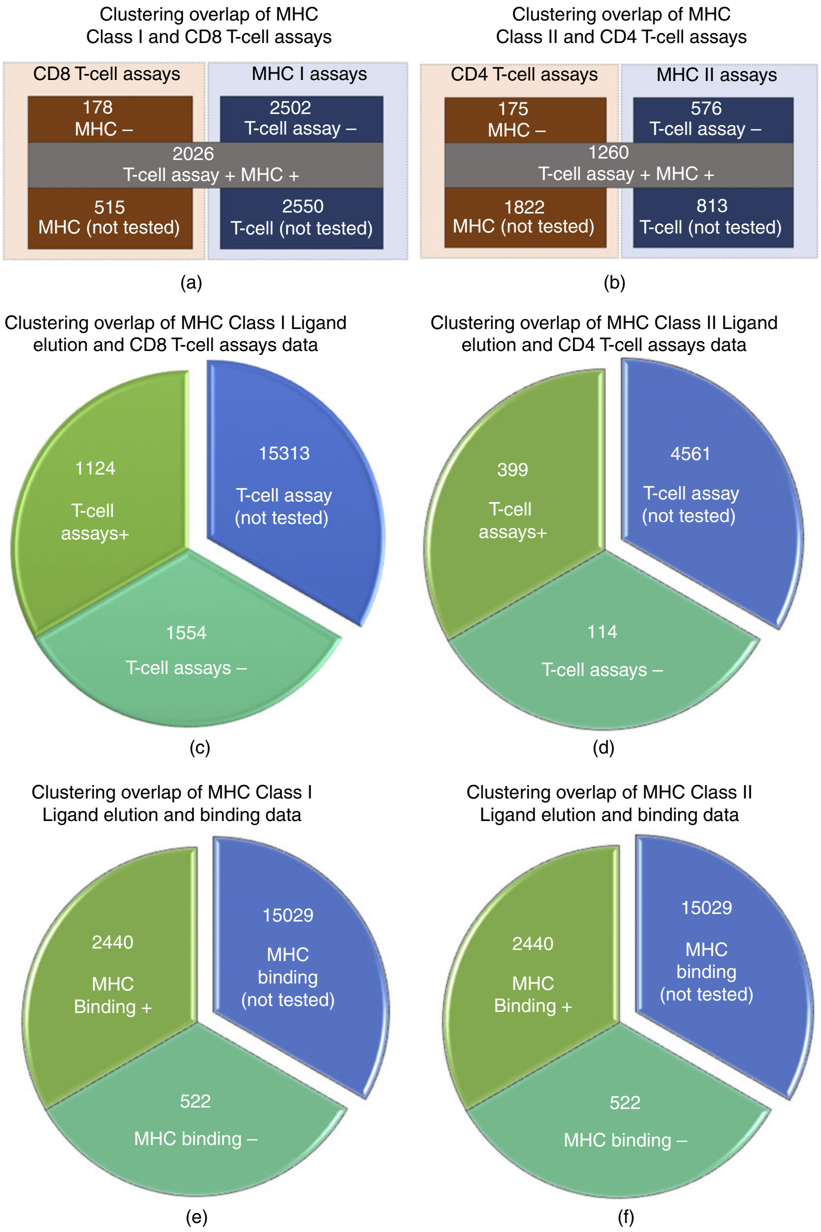


Figure 4. Analysis of overlapping clusters in major histocompatibility complex (MHC) binding, T-cell and MHC ligand elution data. (a) H-chart for overlapping clusters between MHC class I binding and CD8 T-cell assays. (b) H-chart for overlapping clusters between MHC class II binding and CD4 T-cell assays. (c) Pie-chart of overlapping clusters in MHC class I ligand elution data and CD8 T-cell assays. (d) Pie-chart of overlapping clusters in MHC class II ligand elution data and CD4 T-cell assays. (e) Pie-chart of overlapping clusters in MHC class I ligand elution and binding assays data. (f) Pie-chart of overlapping clusters in MHC class II ligand elution and binding assays data.

II) of the MHCLEp clusters were also positive for MHC binding assays. We expected that being positive for MHCLE would have been a better predictor of T-cell assay positivity than simple MHC binding. Surprisingly, this was not the case and we found that 42% and 77·8% of MHCLEp clusters were also T-cell assay +, for class I and class II, respectively. This is in agreement with the notion that MHC binding and ligand processing are necessary, but not sufficient, for T-cell immunogenicity.[45]

Implementation of the three methods in a unique clustering tool online

An online version of the tool described herein was developed, with the aim of making it accessible through the IEDB website. We designed the tool to be as user-friendly as possible by requiring minimum parameter input from the users, but at the same time preserving rigorous functionality.

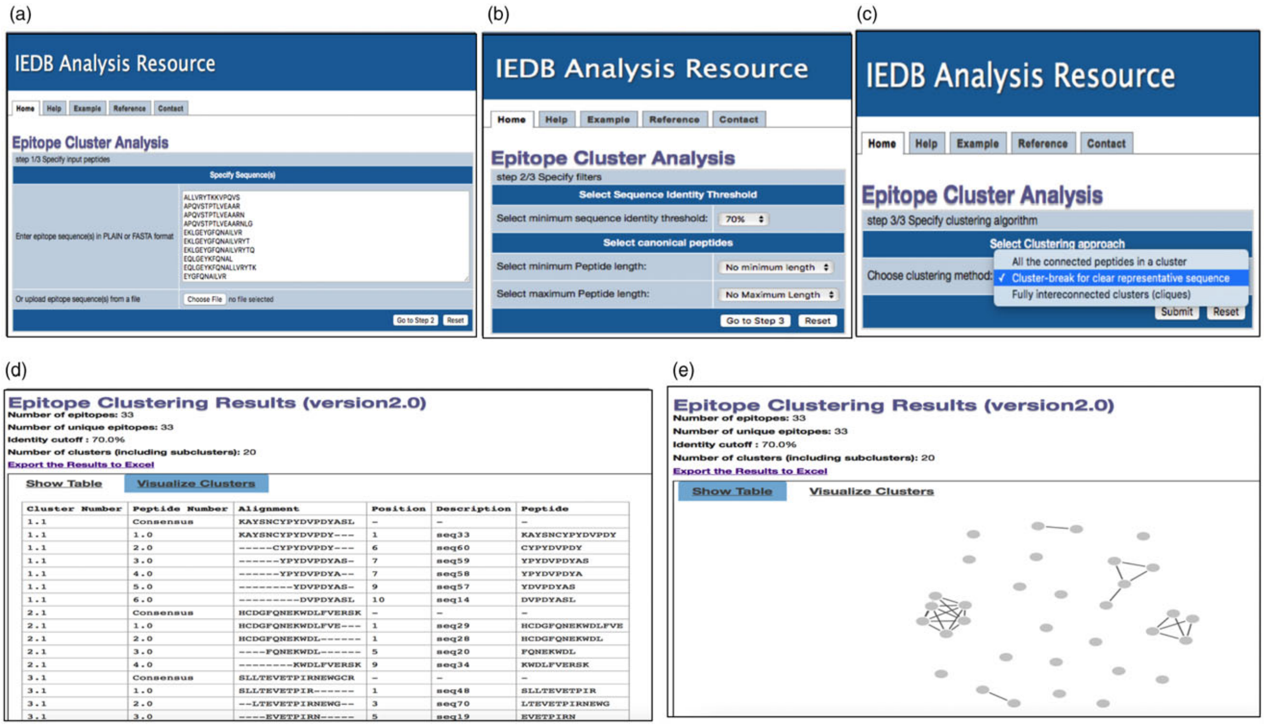
In terms of an input interface, users will provide a set of sequences by pasting directly into a text box or through upload in a given file format (Fig. 5a). The sequences can be in either plain text format, where each line has a separate sequence, or in FASTA format. In addition, the user can attach metadata to the sequences, such as an IEDB peptide identifier or other user-generated metadata (such as the antigen or organism source of the sequence, sequence position etc.) in the FASTA formatted file. Next, the user chooses the minimum threshold for sequence identity, and whether to apply the size-exclusion filter (default, no length filters) (Fig. 5b). And then, in a third step, the user would choose one of three available options for the clustering approach (Fig. 5c). These three options correspond to the connected graphs, cluster-break and cliques approach. In default settings, cluster-break algorithms will be run.

Once the user has filled in these options, a submit button will automatically redirect the user to the results page. The results page displays a summary of results, including the total number of peptides, unique number of peptides, selected identity threshold and total number of clusters found in a given data set (Fig. 5d). The option of exporting the result to Excel will be provided.

The complete results are provided both in tabular and graphic form. The tabular results show the consensus or peptide number, peptide alignment, position of a peptide in the alignment, description and the peptide sequence against each cluster number. The cluster number is represented by a decimal number, where the number before the decimal point is the cluster number before applying the cluster-breaking algorithm and the number following the decimal point identifies each sub-cluster identified by the algorithm.

In the graphic representation (Fig. 5e), each cluster is visualized in the form of networks, where each node is a peptide and the edge is the connection between any two nodes, if they have identity greater than the selected threshold. Single nodes correspond to peptides where no other sequence shares more than the given threshold of identity.

In the current version of clustering tool (Cluster2), we are offering the three different types of approaches described here, which are useful in different contexts of peptide application as summarized in Table 7. The clique approach generates clusters of fully interconnected peptides according to the specified identity threshold. The approach will give, for each cluster, a clear representative sequence, but cluster redundancy would be expected. This approach is useful in generating, for example, megapools where all the peptides belonging to a single cluster have the same sequence identity threshold, and it would be efficacious to select one peptide (based on user discretion, in our case higher magnitude of response in terms of SFCs) from each cluster. Unique peptides can be selected at the end to address redundancy issues.



(a)

Epitope Cluster Analysis

step 1/3 Specify input peptides

Specify Sequence(s)

Enter epitope sequence(s) in PLAIN or FASTA format

```
ALLVRYTKKVPQVS
APQVSTPTLVEAAR
APQVSTPTLVEAARN
APQVSTPTLVEAARNLG
EKLGEYGFQNAILVR
EKLGEYGFQNAILVRYT
EKLGEYGFQNAILVRYTQ
EQLGEYKFQNAL
EQLGEYKFQNALLVRYTK
EYGFQNAILVR
```

Or upload epitope sequence(s) from a file: no file selected

(b)

Epitope Cluster Analysis

step 2/3 Specify filters

Select Sequence Identity Threshold

Select minimum sequence identity threshold: 70%

Select canonical peptides

Select minimum Peptide length: No minimum length

Select maximum Peptide length: No Maximum Length

(c)

Epitope Cluster Analysis

step 3/3 Specify clustering algorithm

Select Clustering approach

Choose clustering method:
- All the connected peptides in a cluster
- ✓ Cluster-break for clear representative sequence
- Fully intereconnected clusters (cliques)

(d)

Epitope Clustering Results (version2.0)

Number of epitopes: 33
Number of unique epitopes: 33
Identity cutoff : 70.0%
Number of clusters (including subclusters): 20
Export the Results to Excel

Cluster Number	Peptide Number	Alignment	Position	Description	Peptide
1.1	Consensus	KAYSNCYPYDVPDYASL	–	–	–
1.1	1.0	KAYSNCYPYDVPDY---	1	seq33	KAYSNCYPYDVPDY
1.1	2.0	-----CYPYDVPDY---	6	seq60	CYPYDVPDY
1.1	3.0	------YPYDVPDYAS-	7	seq59	YPYDVPDYAS
1.1	4.0	------YPYDVPDYA--	7	seq58	YPYDVPDYA
1.1	5.0	--------YDVPDYAS-	9	seq57	YDVPDYAS
1.1	6.0	---------DVPDYASL	10	seq14	DVPDYASL
2.1	Consensus	HCDGFQNEKWDLFVERSK	–	–	–
2.1	1.0	HCDGFQNEKWDLFVE---	1	seq29	HCDGFQNEKWDLFVE
2.1	2.0	HCDGFQNEKWDL------	1	seq28	HCDGFQNEKWDL
2.1	3.0	----FQNEKWDL------	5	seq20	FQNEKWDL
2.1	4.0	--------KWDLFVERSK	9	seq34	KWDLFVERSK
3.1	Consensus	SLLTEVETPIRNEWGCR	–	–	–
3.1	1.0	SLLTEVETPIR------	1	seq48	SLLTEVETPIR
3.1	2.0	--LTEVETPIRNEWG--	3	seq70	LTEVETPIRNEWG
3.1	3.0	----EVETPIRN-----	5	seq19	EVETPIRN

(e)

Epitope Clustering Results (version2.0)

Number of epitopes: 33
Number of unique epitopes: 33
Identity cutoff : 70.0%
Number of clusters (including subclusters): 20
Export the Results to Excel

Figure 5. Screenshots from online tool. (a) Specify Sequence (Step 1), (b) Select clustering parameters (Step 2), (c) Choose clustering algorithm (Step 3), (d) Tabular output (Result page), (e) Graphical Output (Result page).

The cluster approach groups all directly or indirectly connected peptides, based on specified identity threshold. In this case, peptides belonging to the same cluster might have a sequence identity below the threshold, but no peptide redundancy is observed. This approach would be useful, for example, to create non-redundant clusters of peptides that can be used to analyse cross-reactivity with a minimal peptide set. For larger data sets, representative sequence/s might not be clear.

The cluster-break approach is an extension of the above-mentioned cluster approach, where clusters are broken down to sub-clusters to obtain a clear representative sequence from each sub-cluster and to address issues of large membership. Hence, this approach is very useful in applications where it would be helpful to reduce the number of peptides derived from larger data sets.

Comparison with the existing algorithms

Clustering is a well-studied problem in biological data, and has been applied to address different biological questions like creating gene networks, evolutionary relationships and to find sequence homologues. Due to the different purposes for which various algorithms have been developed, it is somewhat challenging to compare an algorithm with others based on a benchmarking data set. Instead, here, we compared the features of different algorithms for peptide clustering (Table 8). In addition to the approaches described herein, we considered, for feature comparison, GibbsCluster,[26,27] PepServe,[24] Hammock[23] and UCLUST.[25] Table 8 provides a summary of the various features assessed in these.

Table 7. Features and applications of different clustering approaches implemented in Cluster2

Features	Clique	Connected clusters	Cluster-break
Clear consensus sequence	Mostly clear, as peptides are fully interconnected	May have many 'X' at different ambiguous positions	Clusters are broken down to get maximum possible clear consensus
Large membership issue	No	Yes	Resolved
Redundancy in clusters	Yes	No	No
Same sequence identity threshold between peptides in clusters	Yes	No	No
Application	Generating mega pools	Cross-reactivity of small data sets	Cross-reactivity of large data sets

Some critically important features are taken into consideration for comparison, which include user input sequences, free online tool, graphical visualization, connectivity of peptides within cluster, cluster representative sequence, dealing with large data sets, basis of calculating identities among peptides and unsupervised nature of clustering (Table 8). As evident from the comparison, the new version of the clustering tool (Cluster2) offers many missing features over existing tools. Overhang identity calculations, visualizing connectivity and providing clear representative sequences are some unique features implemented in the tool. Additionally, our novel cluster-breaking algorithm provides unique way of clustering in large data sets.

Discussion

Peptide clustering has several potential applications in an immunological context. Although there are various algorithms available for data clustering,[46–48] the identification of regions with specific homology, as well as representation of those with clear consensus sequences, has been challenging for the development of appropriate clustering algorithms.

In the past, several methods have been published investigating different aspects of peptide clustering, such as UCLUST,[25] PepServe,[24] Gibbs clustering[26,27] and the Hammock algorithm.[23] However, none of these tools provides the complete connectivity of the peptides within a cluster and a biologically meaningful consensus sequence representing each cluster, features that are of great utility, particularly in the context of immunological studies. In addition, as many available clustering tools use a specific approach, it is challenging for immunologists to select the appropriate algorithms able to address the biological question of interest.

Here, we extend graph theory applications to peptide clustering, where connectivity is defined by shared sequence identity cut-off. Accordingly, peptides that are fully homologous to each other are grouped in cliques, whereas peptides that are homologous to a point are grouped together in connected components. Both of these approaches have their own pros and cons. In the clique approach, each group can give a clear representative sequence, but a peptide can be present in several groups, hindering data set reduction. In the peptide-connected approach, no peptide would be present in two groups, but the group may not provide a clear representative sequence. In addition, we report an alternative approach, where no peptide would be present in more than one

Table 8. Feature-based comparison of different clustering algorithms

Feature	GibbsCluster	PepServe	Hammock	UCLUST	Cluster1.0	Cluster2.0
Input sequences	Amino acid sequence	Amino acids subjected to retrieval	Amino acid sequence	Amino acid sequence	Amino acid sequence	Amino acid sequence
Freely available online tool	Yes	Yes	Yes (Galaxy)	No	Yes	Yes
Graphical visualization	Yes	Yes	No	No	No	Yes
Provides connectivity in a cluster	No	No	No	No	No	Yes
Cluster representative sequence	Motif	No	Main sequence	No	No	Consensus sequence
Large membership issue	NA	NA	NA	Yes	Yes	Resolved
Overhang sequences identity calculation	NA	NA	NA	Yes, but consider only aligned region	No	Yes
Clustering basis	Supervised	Unsupervised	Unsupervised	Unsupervised	Unsupervised	Unsupervised

NA: feature cannot be compared.

group, and also each group can provide a clear representative sequence. In this approach, we develop a novel cluster-break technique allowing derivation of a clear consensus sequence.

The three different approaches are designed to address different biological questions. The clique approach selects a clear consensus sequence out of each clique and has all the peptides grouped on a specific identity level cut-off. This was applied on DENV data, and can be extended in all the cases where redundancy is not important, while a precise identity level is required in each cluster. The reason for applying the clique approach in general is that the peptides in each clique are fully interconnected and selecting the peptide with highest response can represent the remaining peptides in that clique. Hence, the clique approach is suitable for highly redundant data with sequence overlap and or high sequence homology. A similar clique-based approach has previously been applied to analyse the glycan structures[49] and biological data mining.[50]

Alternatively, the cluster approach based on connected components, contains sequences with a lower level of identity with respect to the one used as cut-off, but where no peptide redundancy is observed in each cluster. The cluster approach produced significant and meaningful clusters for T-cell epitopes derived from rat and mouse allergens.[34] In this instance, the resulting clusters shared high interconnectivity and were smaller in size, so a clear representative/consensus sequence was observed and similarities among the two species were not inflated by the presence of redundancies.

Finally, in the context of larger data sets, we observed that neither the clique approach nor the connected components provide legitimate clusters. For this reason, we generated an algorithm to break down clusters in a systematic manner to extract a clear representative/consensus sequence from each sub-cluster. This cluster-break approach has shown to be the most suitable in the context of variable and larger data sets, such as in the case of known epitopes derived from the IEDB.[9] This approach is appropriate for applications where the user, for example, would want to derive minimal, non-redundant, and clear representation from large data sets or identify common sequences between different data sets, as in the case of epitopes able to both bind to MHC and induce a T-cell response. It is worth noting that comparable efficacy for both CD4$^+$ and CD8$^+$ T cells clustering analysis is observed in the context of an IEDB large data set example for the cluster-break method, suggesting that the tool shows equal efficacy in both types of data set.

Herein we describe the creation of a freely available online tool for peptide clustering. The tool provides a user-friendly format to support clustering of disparate epitope data sets. The unique tool incorporates multiple approaches to analyse different types of data set and to address different immunological questions. We applied the tool to in-house experimentally derived CD4$^+$ restricted epitopes, however the tool is equally applicable to CD8$^+$ T-cell epitopes, as shown for the IEDB-derived data set. In addition to the epitopes, the tool can also be applied to cluster and explore interesting patterns in any linear peptide data set (such as cell-penetrating peptides,[51,52] tumour homing peptides,[53] haemolytic peptides,[54] therapeutic peptides[55]).

Acknowledgements

The authors acknowledge the funding supported by U.S. National Institutes of Health grants HHSN27220140045C, HHSN272201200010C and 1P01AI106695-01A1 to AS.

Author's contribution

SKD wrote the code, and analysed the data with AS and BP. VS contributed the rat and mouse allergy section. KV provided the larger data set. AG and DW analysed the DENV data. AS and BP conceived and supervised the project. JS edited the manuscript. All authors approved the manuscript.

Disclosure

The authors declare no competing financial interests.

References

1 Vivona S, Gardy JL, Ramachandran S, Brinkman FS, Raghava GP, Flower DR *et al.* Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. *Trends Biotechnol* 2008; **26**:190–200.

2 Dhanda SK, Usmani SS, Agrawal P, Nagpal G, Gautam A, Raghava GPS. Novel *in silico* tools for designing peptide-based subunit vaccines and immunotherapeutics. *Brief Bioinform* 2017; **18**:467–78.

3 Paradela A, Alvarez I, Garcia-Peydro M, Sesma L, Ramos M, Vazquez J *et al.* Limited diversity of peptides related to an alloreactive T cell epitope in the HLA-B27-bound peptide repertoire results from restrictions at multiple steps along the processing-loading pathway. *J Immunol* 2000; **164**:329–37.

4 Altenburg AF, Rimmelzwaan GF, de Vries RD. Virus-specific T cells as correlate of (cross-)protective immunity against influenza. *Vaccine* 2015; **33**:500–6.

5 Sette A, Sidney J. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* 1999; **50**:201–12.

6 Fikes JD, Sette A. Design of multi-epitope, analogue-based cancer vaccines. *Expert Opin Biol Ther* 2003; **3**:985–93.

7 Ou D, Mitchell LA, Tingle AJ. HLA-DR restrictive supertypes dominate promiscuous T cell recognition: association of multiple HLA-DR molecules with susceptibility to autoimmune diseases. *J Rheumatol* 1997; **24**:253–61.

8 Ayyoub M, Merlo A, Hesdorffer CS, Speiser D, Rimoldi D, Cerottini JC *et al.* Distinct but overlapping T helper epitopes in the 37-58 region of SSX-2. *Clin Immunol* 2005; **114**:70–8.

9 Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* 2015; **43**: D405–12.

10 Fleri W, Paul S, Dhanda SK, Mahajan S, Xu X, Peters B *et al.* The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Front Immunol* 2017; **8**:278.

11 Dhanda SK, Vita R, Ha B, Grifoni A, Peters B, Sette A. ImmunomeBrowser: a tool to aggregate and visualize complex and heterogeneous epitopes in reference protein. *Bioinformatics* 2018; https://doi.org/10.1093/bioinformatics/bty463.

12 Mukherjee S, Warwicker J, Chandra N. Deciphering complex patterns of class-I HLA-peptide cross-reactivity via hierarchical grouping. *Immunol Cell Biol* 2015; **93**:522–32.

13 Stufano A, Capone G, Pesetti B, Polimeno L, Kanduc D. Clustering of rare peptide segments in the HCV immunome. *Self Nonself* 2010; **1**:154–62.

14 Lucchese G, Capone G, Kanduc D. Peptide sharing between influenza A H1N1 hemagglutinin and human axon guidance proteins. *Schizophr Bull* 2014; **40**:362–75.

15 Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 2017; **547**:89–93.

16 Backert L, Johannes Kowalewski D, Walz S, Schuster H, Berlin C, Christoph Neidert M *et al.* A meta-analysis of HLA peptidome composition in different hematological entities: entity-specific dividing lines and "pan-leukemia" antigens. *Oncotarget* 2017; **8**:43915–24.

17 Lindestam Arlehamn CS, McKinney DM, Carpenter C, Paul S, Rozot V, Makgotlho E *et al.* A quantitative analysis of complexity of human pathogen-specific CD4 T cell responses in healthy *M. tuberculosis* infected South Africans. *PLoS Pathog* 2016; **12**: e1005760.

18 Weiskopf D, Angelo M, Zapardiel J, Seumois G, de Silva A, de Silva AD *et al.* DENV-specific CD4 T-cells dominantly recognize capsid-derived epitopes and display a cytotoxic phenotype. *J Immunol* 2016; **196**:147.13.

19 Hinz D, Oseroff C, Pham J, Sidney J, Peters B, Sette A. Definition of a pool of epitopes that recapitulates the T cell reactivity against major house dust mite allergens. *Clin Exp Allergy* 2015; **45**:1601–12.

20 Carrasco Pro S, Sidney J, Paul S, Lindestam Arlehamn C, Weiskopf D, Peters B *et al.* Automatic generation of validated specific epitope sets. *J Immunol Res* 2015; **2015**:763461.

21 Zhao Y, Karypis G. Clustering in life sciences. *Methods Mol Biol* 2003; **224**:183–218.

22 Rottger R. Clustering of biological datasets in the era of big data. *J Integr Bioinform* 2016; **13**:300.

23 Krejci A, Hupp TR, Lexa M, Vojtesek B, Muller P. Hammock: a hidden Markov model-based peptide clustering algorithm to identify protein-interaction consensus motifs in large datasets. *Bioinformatics* 2016; **32**:9–16.

24 Alexandridou A, Dovrolis N, Tsangaris GT, Nikita K, Spyrou G. PepServe: a web server for peptide analysis, clustering and visualization. *Nucleic Acids Res* 2011; **39**: W381–4.

25 Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010; **26**:2460–1.

26 Andreatta M, Alvarez B, Nielsen M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res* 2017; **45**:W458–63.

27 Andreatta M, Lund O, Nielsen M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* 2013; **29**:8–14.

28 Peay ER. Hierarchical Clique Structures. *Sociometry* 1974; **37**:54–65.

29 Gibbons A. Algorithmic Graph Theory. Cambridge, Cambridgeshire; New York: Cambridge University Press, 1985.

30 Bondy JA, Murty USR. Graph Theory with Applications. North Holland: Elsevier Science Ltd, 1976.

31 West DB. Introduction to Graph Theory. Upper Saddle River, NJ: Prentice Hall, 2008.

32 Weiskopf D, Angelo MA, Grifoni A, O'Rourke PH, Sidney J, Paul S *et al.* HLA-DRB1 alleles are associated with different magnitudes of dengue virus-specific CD4$^+$ T-cell responses. *J Infect Dis* 2016; **214**:1117–24.

33 Jeal H, Draper A, Harris J, Taylor AN, Cullinan P, Jones M. Determination of the T cell epitopes of the lipocalin allergen, *Rat n* 1. *Clin Exp Allergy* 2004; **34**:1919–25.

34 Schulten V, Westernberg L, Birrueta G, Sidney J, Paul S, Busse P *et al.* Allergen and epitope targets of mouse-specific T cell responses in allergy and asthma. *Front Immunol* 2018; **9**:235.

35 Aric AH, Daniel AS, Pieter JS. Exploring network structure, dynamics, and function using NetworkX. 2008:11–5.

36 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li WZ *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011; **7**:539.

37 Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009; **25**:1422–3.

38 Sette A. Naturally Processed Peptides. Basel: Karger, 1993.

39 Schellens IMM, Hoof I, Meiring HD, Spijkers SNM, Poelen MCM, van Gaans-van den Brink JAM *et al.* Comprehensive analysis of the naturally processed peptide repertoire: differences between HLA-A and B in the immunopeptidome. *PLoS One* 2015; **10**: e0136417.

40 Hunt DF, Michel H, Dickinson TA, Shabanowitz J, Cox AL, Sakaguchi K *et al.* Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science* 1992; **256**:1817–20.

41 Chicz RM, Urban RG, Lane WS, Gorga JC, Stern LJ, Vignali DA *et al.* Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* 1992; **358**:764–8.

42 Sercarz EE, Maverakis E. MHC-guided processing: binding of large antigen fragments. *Nat Rev Immunol* 2003; **3**:621–9.

43 Chang ST, Ghosh D, Kirschner DE, Linderman JJ. Peptide length-based prediction of peptide-MHC class II binding. *Bioinformatics* 2006; **22**:2761–7.

44 Rist MJ, Theodossis A, Croft NP, Neller MA, Welland A, Chen Z *et al.* HLA peptide length preferences control CD8$^+$ T cell responses. *J Immunol* 2013; **191**:561–71.

45 Dhanda SK, Karosiene E, Edwards L, Grifoni A, Paul S, Andreatta M *et al.* Predicting HLA CD4 immunogenicity in human populations. *Front Immunol* 2018; **9**:1369.

46 Aggarwal CC, Reddy CK. Data Clustering: Algorithms and Applications. Boca Raton, FL: Taylor & Francis, 2013.

47 Xu R, Wunsch DC 2nd. Clustering algorithms in biomedical research: a review. *IEEE Rev Biomed Eng* 2010; **3**:120–54.

48 Andreopoulos B, An A, Wang X, Schroeder M. A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief Bioinform* 2009; **10**:297–314.

49 Fukagawa D, Tamura T, Takasu A, Tomita E, Akutsu T. A clique-based method for the edit distance between unordered trees and its application to analysis of glycan structures. *BMC Bioinformatics* 2011; **12**(Suppl 1):S13.

50 Matsunaga T, Yonemori C, Tomita E, Muramatsu M. Clique-based data mining for related genes in a biomedical database. *BMC Bioinformatics* 2009; **10**:205.

51 Agrawal P, Bhalla S, Usmani SS, Singh S, Chaudhary K, Raghava GP *et al.* CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic Acids Res* 2016; **44**:D1098–103.

52 Gautam A, Singh H, Tyagi A, Chaudhary K, Kumar R, Kapoor P *et al.* CPPsite: a curated database of cell penetrating peptides. *Database (Oxford)* 2012; **2012**: bas015.

53 Kapoor P, Singh H, Gautam A, Chaudhary K, Kumar R, Raghava GP. TumorHoPe: a database of tumor homing peptides. *PLoS One* 2012; **7**:e35187.

54 Gautam A, Chaudhary K, Singh S, Joshi A, Anand P, Tuknait A *et al.* Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res* 2014; **42**:D444–9.

55 Singh S, Chaudhary K, Dhanda SK, Bhalla S, Usmani SS, Gautam A *et al.* SATPdb: a database of structurally annotated therapeutic peptides. *Nucleic Acids Res* 2015; **44**: D1119–26.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article:

Fig. S1. Example representation of clustering using graph theory.

Table S1. Number of epitopes for major histocompatibility complex (MHC) Class I and Class II from each data set.

Table S2. Output of clustering tool for dengue virus data.

Table S3. Selected peptides from dengue virus data in all the cliques and singletons.

Table S4. Analysis of overlapping clusters among different data sets.