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 Peters1,2 and Alessandro Sette1,2

1Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA, and 2Department 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 histocompatibility (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 IMMUNE BROWSER,\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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\textsuperscript{23} was developed to identify consensus motifs in large data sets, and PepServe\textsuperscript{24} to cluster epitopes based on their physicochemical properties. Another application, UCLUST,\textsuperscript{25} performs rapid clustering and implements several previously identified clustering algorithms, and Andreatta \textit{et al.}\textsuperscript{26,27} developed a tool based on Gibbs sampling. Overall, most of these tools cluster based on the physicochemical 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.\textsuperscript{20}

From a computational standpoint, two main clustering approaches exist: cliques\textsuperscript{28} and clustering by connected graph.\textsuperscript{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 epitopes 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.

\section*{Methods}

\subsection*{Dengue virus data set}

The dengue virus (DENV) epitope data set was derived from a comprehensive analysis of human leucocyte antigen-matched CD4\textsuperscript{+} T-cell responses in a Sri Lankan cohort, allowing for the selection of 363 epitopes restricted by 15 DRB1 alleles with phenotypic frequency \textgreater{} 4\% in the Sri Lankan population, as previously reported.\textsuperscript{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 \textit{in vitro} expansion of DENV-specific T cells and interferon \gamma enzyme-linked immunospot assay to rank the best epitopes with highest SFC.\textsuperscript{32}

\subsection*{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 \textit{Rat n I}, a major urinary protein, and constitute a set of 19 peptides found to be T-cell-reactive in rat-allergic patients.\textsuperscript{33} In addition,
our group recently identified 23 peptides from the major mouse allergen, *Mus m I.*

**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. 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. From the alignment output, the consensus sequence is derived using the Biopython package. 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 breaking 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 and the other by connected graph.

For this we developed a simple algorithm to cluster sequences using a network graph approach. 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, C and D or a total of 163 epitopes that could be used to represent the maximum number of peptides in a connected graph.

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. 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 corresponds 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. A previous study had identified 19 T-cell epitopes from the major rat allergen, Rat n I, which exhibits significant homology to Mus m I

| 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 |
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 cliques 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
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

| Cluster number | Peptide number | Alignment | Position | Description | Peptide |
|----------------|----------------|-----------|----------|-------------|---------|
| 1              | Consensus      | TFQLMXLYGRXKDLSSDIEKE    | –        | –           | TFQLMXLYGRXKDLSSDIEKE |
| 1              | 1              | TFQLMVLYGRXKDLSSDIEKE    | 1 Rat Pep17 | TFLQMVLYGRXKDLSSDIEKE |
| 2              | Consensus      | TFQLVYHMLNKXGETFMQLMYGRXKDLSSDIEKE | – | – | TFQLVYHMLNKXGETFMQLMYGRXKDLSSDIEKE |
| 2              | 1              | TFQLVYHMLNKXGETFMQLMYGRXKDLSSDIEKE | 1 Rat Pep19 | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 2              | 2              | TFQLVYHMLNKXGETFMQLMYGRXKDLSSDIEKE | 7 Mus Pep9 | AHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 3              | Consensus      | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | – | – | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 3              | 1              | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | 1 Rat Pep19 | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 3              | 2              | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | 7 Mus Pep9 | AHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 4              | Consensus      | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | – | – | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 4              | 1              | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | 1 Rat Pep19 | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 4              | 2              | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | 7 Mus Pep9 | AHLINKXGETFMQLMYGRXKDLSSDIEKE |

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

| Cluster number | Peptide number | Alignment | Position | Description | Peptide |
|----------------|----------------|-----------|----------|-------------|---------|
| 1              | Consensus      | TFQLMXLYGRXKDLSSDIEKE    | –        | –           | TFQLMXLYGRXKDLSSDIEKE |
| 1              | 1              | TFQLMVLYGRXKDLSSDIEKE    | 1 Rat Pep17 | TFLQMVLYGRXKDLSSDIEKE |
| 2              | Consensus      | TFQLVYHMLNKXGETFMQLMYGRXKDLSSDIEKE | – | – | TFQLVYHMLNKXGETFMQLMYGRXKDLSSDIEKE |
| 2              | 1              | TFQLVYHMLNKXGETFMQLMYGRXKDLSSDIEKE | 1 Rat Pep19 | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 2              | 2              | TFQLVYHMLNKXGETFMQLMYGRXKDLSSDIEKE | 7 Mus Pep9 | AHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 3              | Consensus      | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | – | – | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 3              | 1              | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | 1 Rat Pep19 | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 3              | 2              | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | 7 Mus Pep9 | AHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 4              | Consensus      | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | – | – | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 4              | 1              | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | 1 Rat Pep19 | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE |
| 4              | 2              | YDRYVMFHLINKXGETFMQLMYGRXKDLSSDIEKE | 7 Mus Pep9 | AHLINKXGETFMQLMYGRXKDLSSDIEKE |

S. K. Dhanda et al.

© 2018 The Authors. Immunology Published by John Wiley & Sons Ltd., Immunology, 155, 331-345
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

---

**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 | 2.78₁        | 57.361        |

MHCLE, major histocompatibility complex ligand elution.

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

'₁' denotes the count of a particular feature.

---

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

| Cluster size | Raw data | Length filtered data (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.

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

---

© 2018 The Authors. *Immunology* Published by John Wiley & Sons Ltd., *Immunology*, 155, 331–345

---

**Figure 1. Plot representing the number of peptides in top 10 clusters from major histocompatibility complex class II ligand elution data.**
visualize this feature, we followed the approach of constructing an undirected graph network from peptide sequence identity matrices using a publicly available network package. 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.](image-url)
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 from each data set after applying the cluster-break algorithms.

### Table 6. Distribution of cluster size for different data sets

| Cluster size | MHCL | MHC binding | MHC binding | CD8 | CD4 |
|--------------|------|-------------|-------------|-----|-----|
| ≥ 1000       | 0    | 0           | 0           | 0   | 0   |
| 100–999      | 10   | 0           | 6           | 9   | 0   |
| 50–99        | 26   | 8           | 24          | 14  | 12  |
| 30–49        | 49   | 9           | 60          | 32  | 58  |
| 10–29        | 492  | 15          | 842         | 552 | 1933|
| < 10         | 4335 | 16          | 136         | 6122| 2190|
| Total        | 4912 | 16          | 136         | 6122| 2190|

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.

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 MHCLE + clusters are undetermined when it comes to T-cell assay (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 II) 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 (MHCLE +) (see Supplementary material, Table S4), therefore it cannot be inferred whether the sequence is negative or not tested.

Figure 3. Plot representing the top 10 clusters in different data sets after applying cluster-break algorithm.
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% 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...
通过在给定的文件格式（图5a）中上传。序列可以是纯文本格式，其中每行都有一个单独的序列，或在FASTA格式中。此外，用户可以添加元数据到序列中，例如IEDB肽标识符或其他用户生成的元数据（例如序列来源、序列位置等）在FASTA格式中。接下来，用户可以选择最小阈值以供序列标识，以及是否应用大小排除过滤器（默认，不包括长度过滤器）（图5b）。然后，在第三步，用户会选择三种可用选项中的一个来为聚类方法（图5c）聚类。这三个选项对应于连接图、聚类中断和 clique 方法。在默认设置中，聚类中断算法将被运行。

一旦用户填写了这些选项，提交按钮将自动重定向用户到结果页面。结果页面显示了结果的摘要，包括肽的总数、唯一肽的数目、选择的序列相似度阈值和给定数据集中找到的簇的总数（图5d）。提供导出结果到Excel的功能。

完整的结果将以表格和图形形式提供。表格结果展示共识或肽编号、肽对齐、肽在对齐中的位置、描述和肽序列针对每个簇编号。簇编号表示一个十进制数字，其中整数部分是应用聚类中断算法之前的簇编号，小数部分标识算法识别的子簇。

在当前版本的聚类工具（CLUSTER2）中，我们提供了三种不同的方法，这些方法在肽应用的不同上下文中有用，如表7中总结的那样。Clique方法生成完全互联的肽簇，根据指定的相似度阈值。该方法将为每个簇提供一个清晰的代表序列，但簇的冗余性是预期的。这种方法在产生例如超大池时有用，其中所有成员都具有相同的序列标识符，且能基于用户选择（我们的例子中是SFC值）从每个簇中选择一个肽。可以从中选择唯一的肽以解决冗余性问题。

图5。在线工具的截图。 (a) 在步骤1中指定序列，(b) 在步骤2中选择聚类参数，(c) 在步骤3中选择聚类算法，(d) 表格输出（结果页面），(e) 图形输出（结果页面）。

© 2018 The Authors. Immunology Published by John Wiley & Sons Ltd., Immunology, 155, 331–345

开发一个新型聚类工具
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, PepServe, Hammock and UCLUST. Table 8 provides a summary of the various features assessed in these.

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, 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, PepServe, Gibbs clustering and the Hammock algorithm. 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
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 and biological data mining.

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 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 and biological data mining.

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, tumour homing peptides, haemolytic peptides, therapeutic peptides).

Acknowledgements

The authors acknowledge the funding supported by U.S. National Institutes of Health grants HHSN27220140045C, HHSN27220120010C 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 JB, Ramachandran S, Brinkman FS, Raghava GP, Flower DR et al. Computer-aided biotechnology from immune-informatics to reverse vaccinology. Trends Biotechnol 2008; 26:190–200.
2 Dhanda SK, Usman SI, Agraval P, Nagal G, Gautham 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-B7-rebound peptide repertoire results from restrictions at multiple steps along the processing-loading pathway. J Immunol 2000; 164:328–37.
4 Altenburg AF, Rimmelzwaan GF, de Vries RD. Virus-specific T cells as correlate of (cross-)protective immunity against influenza. Vaccine 2013; 31:500–6.
5 Sette A, Sidney J. Nine major HLA class I supertypes account for the vast majority of HLA-A and -B polymorphisms. 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 Oh 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 Biomol Struct Dyn 1997; 15:243–51.
8 Ayyoub M, Merko A, Heedorffer CS, Spenser D, Rimidoli D, Cerrottini JC et al. Distinct but overlapping T helper epitopes in the 37-58 region of S25X. 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 2010; 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, Hu B, Grifoni A, Peters B, Sette A. ImmuneBrowser: 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 Stafano A, Capone G, Pesetti B, Polimeno L, Kanduc D. Characterizing of rare peptide segments in the HCV immunome. Self Nontol 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; 40:561–75.
15 Dush P, Flore-Girardot AJ, Hertz T, Wang GC, Sharma S, Souquette A et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature 2017; 547:89–95.
16 Backert L, Johannes Kowalewski D, Walt S, Schuster H, Berlin C, Christoph Neidert M et al. A meta-analysis of HLA peptide 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, Rozen V, Mak golitho 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 Wisniewski D, Pangulo M, Zapidjian I, Soumios 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 Hinu D, Overoff 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, Wisniewski D, Peters B et al. Automatic generation of validated specific epitope sets. J Immunol 2015; 195:763461.
21 Zhao Y, Karypis G. Clustering in life sciences. Methods Mol Biol 2003; 224:183–218.
22 Rötter R. Clustering of biological datasets in the era of big data. J Integ Bioinform 2016; 13:30.
23 Krejci A, Hupph TR, Leux M, Vojtoek 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 Alexandrou A, Dovrolis N, Tsangarides GT, Nikita K, Syrrou 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; 26:2660–1.
26 Andreatta M, Alvarez B, Nielsen M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. Nucleic Acids Res 2017; 45:W548–W53.
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 Pyro ER. Hierarchical Clue Structures. Syst Biol 1984; 37:54–65.
29 Gibbons A. Algorithmic Graph Theory. Cambridge, Cambridgebridge, New York: Cambridge University Press, 1985.
30 Bondy JA, Murty USR. Graph Theory with Applications. New York: Elsevier Science Ltd, 1976.
31 West DB. Introduction to Graph Theory. Upper Saddle River, NJ: Prentice Hall, 2000.
32 Wisniewski 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; 214:1117–21.
33 Jael H, Draper A, Harris J, Taylor AN, Cullinan P, Jones M. Determination of the T-cell epitopes of the lipocalin allergen, Bet v 1. Clin Exp Allergy 2004; 34:1919–25.
34 Schulten V, Westernberg L, Bursuta G, Sidney J, Paul S, Buser P et al. Allergen and epitope targets of mouse-specific T-cell responses in allergy and asthma. Front Immunol 2018; 9:235.
35 Ari C, Daniel AS, Pieter JS. Exploring network structure, dynamics, and function using NetworkX. 2008;11–5.
36 Sivers F, Wilm A, Dinet D, Gibson TJ, Karpuski K, Li WZ et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Synt Biol 2011; 7:539.
37 Cock PJ, Antao T, Chang JT, 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 Schellen BMM, Hoof I, Meiring HD, Spijkers SNM, Poelen MCM, van Gaans-van den Brink IAM 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, Michiel H, Dickinson TA, Sabhanowitz JS, 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 Chacz RM, Urban RG, Lane WS, Gorga SC, Stern LJ, Vignali DA et al. Predominant naturally processed peptides bound to HLA-DR1 are derived from HMC-related molecules and are heterogeneous in size. Nature 1992; 358:764–8.
42 Sercarz EE, Mavrides K. MHC-guided processing: binding of large antigen fragments. Nat Rev Immunol 2003; 3:821–9.
43 Chang ST, Ghosh D, Kirchner DE, Linderman JJ. Peptide length-based prediction of peptide-MHC class II binding. Bioinformatics 2006; 22:2761–7.
44 Rist MJ, Theodossi 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, Karoosie E, Edwards L, Grifoni A, Paul S, Andreatta MA et al. Predicting HLA-CDM immunogenicity in human populations. Front Immunol 2018; 9:1369.
46 Aggarwal GC, 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; 3:320–54.
48 Andropoulos 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 Polakowa D, Taniura 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.
Development of a novel clustering tool

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.