BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 1 2013, pages 8–14
doi:10.1093/bioinformatics/bts621

Sequence analysis

Advance Access publication October 24, 2012

Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach

Massimo Andreatta[1,*], Ole Lund[1] and Morten Nielsen[1,2]

[1]Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby, Denmark and [2]Instituto de Investigaciones Biotecnológicas, Universidad de San Martín, CP 1650 San Martín, Buenos Aires, Argentina

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Proteins recognizing short peptide fragments play a central role in cellular signaling. As a result of high-throughput technologies, peptide-binding protein specificities can be studied using large peptide libraries at dramatically lower cost and time. Interpretation of such large peptide datasets, however, is a complex task, especially when the data contain multiple receptor binding motifs, and/or the motifs are found at different locations within distinct peptides.

Results: The algorithm presented in this article, based on Gibbs sampling, identifies multiple specificities in peptide data by performing two essential tasks simultaneously: alignment and clustering of peptide data. We apply the method to de-convolute binding motifs in a panel of peptide datasets with different degrees of complexity spanning from the simplest case of pre-aligned fixed-length peptides to cases of unaligned peptide datasets of variable length. Example applications described in this article include mixtures of binders to different MHC class I and class II alleles, distinct classes of ligands for SH3 domains and sub-specificities of the HLA-A*02:01 molecule.

Availability: The Gibbs clustering method is available online as a web server at http://www.cbs.dtu.dk/services/GibbsCluster.

Contact: massimo@cbs.dtu.dk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 19, 2012; revised on October 5, 2012; accepted on October 14, 2012

1 INTRODUCTION

Peptides are short amino acid sequences occurring ubiquitously in biological processes, such as metabolism, signal transduction and immune response. They are also extensively used in research to mimic functional or (linear) structural aspects of proteins and protein interactions. The advantage of using peptides lies in the relative ease in generating large libraries of sequences, such as in phage display technologies (Koivunen *et al.*, 1999; Bratkovič, 2010). More recently, developments in high-throughput peptide microarrays have allowed producing large-scale datasets of peptide-ligand interactions and have been applied to various problems including antibody–antigen interactions, peptide-MHC binding, kinase binding motifs and other receptor-ligand interactions (Soen *et al.*, 2003; Schutkowski, 2005; Uttamchandani and Yao, 2008; Halperin *et al.*, 2011).

Identifying receptor-ligand binding motifs within peptide datasets is a highly challenging task for at least two major reasons, which we term alignment and poly-specificity. The alignment problem arises because most receptor motifs are weak and short, making identification of the binding register within the ligands not trivial (Nielsen *et al.*, 2004). The poly-specificity problem arises because receptor-ligand datasets often contain multiple motifs either owing to the experimental setup or to the actual poly-specificity of the receptor (Gfeller, 2012). Several bioinformatics methods have been developed attempting to deal with these challenges and detect subtle sequence signals in peptide datasets, including motif alignment (Bailey *et al.*, 2006), Gibbs sampling (Lawrence *et al.*, 1993), Hidden Markov Models (Noguchi *et al.*, 2002) and artificial neural networks (Nielsen and Lund, 2009). In particular, artificial neural networks have shown a high performance on this kind of data (Wang *et al.*, 2010; Andreatta *et al.*, 2011). Significant correlations between residues have been found in peptide interaction domains (Gfeller *et al.*, 2011). Although positional correlations can be accurately captured by artificial neural networks, the specificities of such domains can, in many cases, more intuitively be represented by multiple position-specific scoring matrices (PSSM) (Bailey and Elkan, 1995; Gfeller *et al.*, 2011; Kim *et al.*, 2012). Multiple PSSMs allow visualizing poly-specificities as sequence logos of the different binding modes.

Although the above methods attempt to deal with the challenges involved in motif identification in peptide datasets, they all suffer from the limitations of only dealing with single specificities or requiring the input data to be pre-aligned to a common motif. In this article, we describe a novel approach for effective alignment and clustering of peptide data going beyond these limitations. In the Gibbs clustering method, alignment and specificity clustering are performed simultaneously by sampling the space of possible solutions using a Gibbs sampling strategy. Each cluster is represented by a PSSM, and the method aims at maximizing the information content of individual matrices while minimizing the overlap between distinct clusters.

2 METHODS

The Gibbs clustering algorithm attempts to group the input peptide data into a number of clusters and for each cluster identify the optimal local sequence alignment based on the optimization of the fitness of the system in terms of Kullback–Leibler distance (KLD) sum of the alignments. The KLD allows measuring the information gain of an observed amino acid distribution compared with a background distribution (the frequency of

*To whom correspondence should be addressed.

each amino acid in random protein sequences). A given alignment can be represented by a log-odds (LO) weight matrix, which summarizes the amino acid preferences for each column of the alignment. Throughout the article, we graphically represent LO matrices using the sequence logo visualization tool Seq2Logo (Thomsen and Nielsen, 2012).

2.1 Log-odds matrices

An LO weight matrix is calculated as $\log(p_{A,j}/q_A)$, where $p_{A,j}$ is the frequency of amino acid A at position j, and q_A is the background frequency of A. These frequencies are calculated as described in Nielsen *et al.*, 2004, including heuristic sequence weighting and pseudo-count correction. To avoid the creation of small highly specialized clusters, we introduce an additional term to the LO matrix calculation to account for the size of the alignment. In our scheme, terms in the PSSM are calculated using:

$$LO_{A,j} = \frac{n}{n+\sigma}\log\frac{p'_{A,j}}{q_A} \quad (1)$$

where n is the number of peptides in the alignment, σ is a weight on small clusters, and $p'_{A,j}$ is the pseudo-count corrected frequency. The function of σ is to flatten the LO matrix when the alignment is composed of few sequences (n small), but its effect is minor when the matrix is constructed on many data points (n large). Practically, it avoids the creation of small and specialized alignments, favouring instead larger and more general ones.

A peptide x can be scored simply by adding the LO values for the amino acid found at each position in x:

$$S = \sum_j LO'_{A,j} \quad (2)$$

where j is the index over the positions in the alignment core, and A is the amino acid found at position j in x. However, when evaluating the fitness of a given sequence x in an alignment (where x is part of the alignment), we must take the precaution of excluding x from the matrix calculation before doing the evaluation. We call $LO'_{A,j}$ the LO matrix made without sequence x.

2.2 Scoring function

In the general case, a Gibbs clustering solution is composed of g clusters, each with a corresponding alignment and LO matrix. When evaluating a clustering solution, we aim to maximize the intra-cluster fitness of the alignment while minimizing the similarity between different clusters. In other words, the distance between points in the same cluster should be as small as possible, whereas the distance between points in different groups should be maximal. In the Gibbs clustering algorithm, we implement this maximization using the relationship:

$$S^*_i = S_i - \lambda \max_{\substack{1\le n\le g \\ n\neq i}}(S_n, 0) \quad (3)$$

where S_i is the score of a given peptide to the LO matrix $LO'_{A,j}$ of cluster i. Note that, as discussed above, the LO matrix of group i is calculated excluding the peptide to be scored. The *max*() part of the equation determines the inter-cluster similarity, i.e. which cluster is the closest to cluster i. If we imagine to have, besides the g clusters given by the data, and additional cluster composed of the universe of natural peptides, the amino acid frequencies $p'_{A,j}$ in this extra group would be equal to the background frequencies q_A for any amino acid A. Thus $\log(q_A/q_A)=0$ in Equation 1, leading to a $LO_{A,j}$ matrix composed of zeros, which gives scores $S_{BG}=0$ for all sequences. This justifies the zero in Equation 3, and provides a generalization for the case where there is only one cluster, with $S^*_i = S_i$.

The parameter λ modulates the weight of inter-cluster similarity on the final sequence score. For $\lambda=0$ overlap between clusters is not penalized, leading to tight but promiscuous clusters. Large λ values put emphasis on inter-cluster similarity, at the expense of consistency within the same group.

Equation 3 defines the energy function of a single sequence in the alignment. The overall score of the alignment/clustering is given by the average score of all sequences in the dataset. The fitness of the system can be thought of as the relative entropy or KLD from the background model made on random peptides.

2.3 Moves of the algorithm

Initially, peptides are distributed randomly in g clusters. Then the algorithm proceeds with a number of 'moves' to align and cluster the sequences and optimize the KLD of the alignment/clustering. The probability of accepting a move is given by:

$$P = \min\left[1, e^{dE/T}\right] \quad (4)$$

where dE is the energy change as a result of the move, and T is a scalar commonly known as the temperature of the system, lowered by discrete steps during the iterations.

The algorithm consists of three different moves: (i) *Single sequence move:* in this move, we attempt to transfer a peptide x from one group G_0 to a destination group G_d chosen at random. The score S^*_o of x in its original cluster is calculated using Equation 3, selecting the core register that gives the highest score. In the same way, S^*_d is obtained for the destination group. The move is then accepted or rejected following Equation 4, where $dE = S^*_d - S^*_o$. (ii) *Simple shift:* this move attempts to move a peptide x within a group, by applying a random shift to the alignment core of x. The score of x is calculated before and after the shift, and the dE between the two configurations determines whether the move is accepted or rejected according to Equation 4. (iii) *Phase shift:* the entire alignment of a group G_o is shifted a random number of positions to the left or to the right. This move may be important if the alignment reaches a local minimum where the sequences are optimally aligned to each other, but the core window is not centered on the most informative motif. As in the other moves, the configurations before and after the move are compared to calculate whether the move is favourable or unfavourable, and accepted/rejected following Equation 4.

The 'simple shift' and 'phase shift' moves have been described before for multiple sequence alignment (Lawrence *et al.*, 1993; Nielsen *et al.*, 2004). The new feature of the Gibbs clustering method is the additional 'single sequence' move, which allows transferring sequences between different clusters. The three moves are generally performed with different frequency. The simple shift move, with the lowest impact among the three moves, is attempted at each iteration. Single sequence moves are performed every F_r iterations. Phase shifts, which affect at the same time all peptides in a given clusters, would generally be the least frequent and occur every F_s iterations, with $F_s > F_r > 1$. Throughout the article, these parameters are fixed to $F_r = 10$ and $F_s = 1000$. The default cooling schedule uses 20 linear temperature steps starting from an initial T of 0.8 down to 10^{-5}.

2.4 Trash cluster to collect spurious sequences

The algorithm allows including an additional cluster, called trash cluster, to collect the peptides that appear not to match any of the motifs being identified. The behaviour of the trash-cluster is identical to any of the other clusters, with the difference that the sequences in the trash cluster do not contribute to the overall score of the system. The trash cluster can be thought of as the universe of all natural peptides (i.e. the background model), and peptides can be moved in and out from the trash cluster with probability defined by the Monte Carlo relationship (Equation 4), where the score to the trash clusters is always equal to the background baseline (zero by default, but can be set to different values to adjust the levels of sensitivity and specificity).

2.5 Measures of clustering quality

As a measure of clustering quality, we used the Adjusted Rand Index (ARI). This measure is based on the well-known Rand index (Rand, 1971), but corrected for chance and class size. We implemented the ARI corrected for chance as in Hubert and Arabie, 1985. As a term of comparison, we also used a modified version of the Matthews correlation coefficient (MCC) extended to more than the conventional two classes (positives and negatives). In the general case where A-mixed specificities are grouped in C clusters, a MCC is initially calculated for each cluster. The true positives for group C_i are given by the class A_i with highest number of sequences in C_i, the false positives by the number of sequences in C_i not belonging to A_i, the false negatives by the number of sequences labelled A_i not found in C_i and the true negatives are all the remaining sequences. The MCC for the entire matrix is then calculated as the average MCC of each cluster. The notation for ARI and MCC calculation is also illustrated in Supplementary Table S1.

2.6 Training from multiple initial seeds

Gibbs sampling is a heuristic rather than a rigorous optimization procedure. Therefore, it cannot guarantee that the most optimal solution is always reached from any starting configuration. A common procedure to boost performance is to repeat the sampling from a number of initial random configurations and select the solution that appears to be optimal in terms of the fitness function that governs the system. Clearly, this is a sound procedure only if optimal fitness (KLD) corresponds to optimal clustering of the data. We investigated the correlation between fitness and quality of the clustering on Major Histocompatibility Complex (MHC) class I datasets containing different number of specificities. Binders to different alleles were combined to obtain mixtures of 5 to 8 alleles, and then the Gibbs clustering algorithm was used to recover the distinct motifs. For each allele combination, we ran the algorithm from 40 random initial configurations, measuring for each the fitness in terms of KLD and the clustering quality in terms of ARI.

In general, we observe that both KLD and ARI tend to decrease as the number of alleles in the mixture increases (Supplementary Fig. S1). Yet, in the case of MHC class I where motifs are very strong and distinct from each other, it is possible to reconstruct with high accuracy even up to 8 different specificities. The same considerations can be made if we measure clustering quality in terms of MCC instead of ARI, which correlates in very similar fashion to KLD (Supplementary Fig. S2). These results show that, only based on the KLD, it is possible to filter out sub-optimal solutions. By running the algorithm from different starting conditions, and selecting solutions with high KLD, the method achieves a higher classification performance. Multiple seeding and automatic selection of the optimal solution are integrated in the Gibbs clustering algorithm.

3 RESULTS

The Gibbs clustering algorithm performs two essential tasks simultaneously: alignment and clustering of peptide data. Here, we use the method to de-convolute binding motifs in a panel of different peptide datasets with different degrees of complexity spanning from the simplest case of pre-aligned fixed-length peptides to cases of unaligned peptide datasets of variable length. More details about the datasets are given as Supplementary Material.

3.1 Pre-aligned data—Mixtures of binders to MHC class I alleles

To benchmark the clustering aspect of the Gibbs algorithm, we used a set of pre-aligned fixed-length peptides with experimentally confirmed binding to representatives of the 12 MHC class I supertypes (see Supplementary Material). These 12 MHC molecules all have highly specific binding motifs with limited mutual overlap (Lund *et al.*, 2004). For each number of alleles $n = \{1, 2, \ldots, 8\}$, 10 different combinations of n alleles were constructed randomly from the pool of the 12 MHC molecules. For each dataset, the algorithm was used to cluster the peptides into $c = \{1, 2, \ldots, 12\}$ groups, and the c with optimal KLD score was recorded. Figure 1 shows the results of this calculation. For $\lambda = 0.5$, the number of predicted motifs correlates well with the actual number of alleles in the dataset. With smaller values of λ, the method tends to over-estimate the number of motifs, whereas for larger λ clusters with shared similarities, they are more heavily penalized and are merged into fewer clusters. The predictions are most consistent (lowest variations in the optimal number of clusters) on mixtures of few alleles. This is a natural consequence of both the increased complexity of the search space, as the number of alleles is increased and the promiscuity of MHC binding peptides. Although the 12 MHC class I molecules share very limited overlap in specificity, a larger collection of alleles increases inevitably the chance of including cross-binding peptides the dataset.

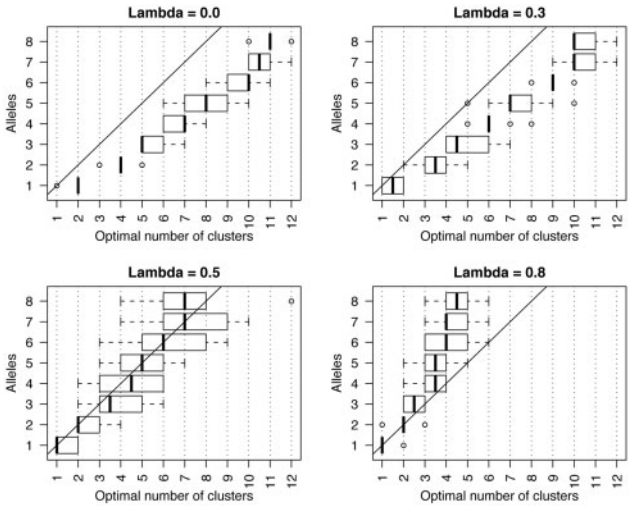

Fig. 1. Box-and-whisker plot showing the optimal number of clusters on mixtures of different MHC class I alleles. The algorithm was run on 10 different random combinations of n alleles, where $n = \{1 \ldots 8\}$, starting with $c = \{1 \ldots 12\}$ clusters for each combination. The optimal number of clusters of each of the 10 combinations is the c with highest KLD of the system. The four panels show the predicted number of clusters for four different values of λ for a fixed value of $\sigma = 10$. With $\lambda = 0.5$, the correlation between number of alleles in the dataset and predicted number of clusters falls approximately on a straight line with slope = 1

3.2 Unaligned data—Mixtures of binders to MHC class II alleles

To demonstrate the performance of the Gibbs clustering method on datasets of unaligned peptides of variable length, we turned to the MHC class II system. As opposed to MHC class I molecules, which in the vast majority of cases interact only with peptides of length between 8 and 10 amino acids, MHC class II molecules

can bind peptides of highly variable length (Rammensee *et al.*, 1999). Binding of a peptide to a MHC class II molecule is primarily determined by a core of 9 amino acids, but the location of the 9-mer core within the peptide is not known a priori. Therefore, MHC class II binding data are by nature unaligned with respect to the binding core.

The Gibbs clustering algorithm was applied to identify motifs in a set of binders to the MHC class II HLA-DRB1*03:01 and HLA-DRB1*04:01 molecules. Compared with MHC class I, class II alleles share a high degree of overlap in their binding specificities. This promiscuity between different MHC class II molecules complicates the performance evaluation of the clustering algorithm, as a peptide may match the motif of multiple alleles, in which case it is not clear in what cluster the sequence should be rightfully placed. To lower this potential degree of cross-binding, the dataset was constructed to include experimentally confirmed binders with weak predicted cross-binding potential (for details refer to Supplementary Material). We maintained the same parameters used for the MHC class I benchmark, except for λ, which was increased to 0.8 to avoid the creation of excessively small and specialized clusters (running the algorithm with $\lambda = 0.5$ resulted, in particular, in the DRB1*03:01 peptides being sub-divided into several small and highly specialized clusters). Additionally, as HLA-DR molecules are known to prefer hydrophobic amino acids at position P1, we imposed a preference for this kind of amino acids in the Gibbs sampling moves as proposed by Nielsen *et al.*, 2004. The algorithm was run multiple times to create 1–4 clusters, each started from five different random configurations. For each cluster size, the solution with the highest KLD score was recorded. The optimal solution indicated the presence of two clusters (Supplementary Fig. S3), and the corresponding motifs are shown in Figure 2. The main distinctive feature in the logos of Figure 2 is the acidic (D) anchor at position P4 and a basic (K/R) anchor at position 6 of the first motif, which are absent in the second logo. These preferences characterize the binding motif of HLA-DRB1*03:01. The classification of the peptides in the two groups (Fig. 2c) demonstrates that most peptides are clustered correctly, with an accuracy of 79% and MCC of 0.59.

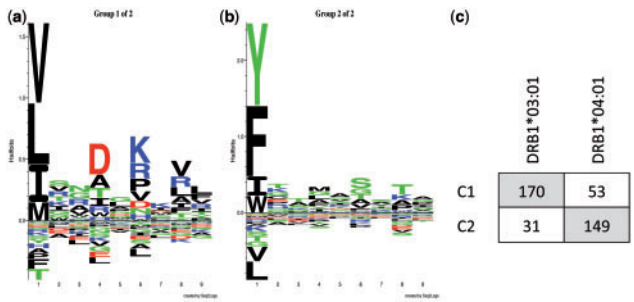

Fig. 2. Reconstructed binding motifs from a mixture of binders to 2 MHC class II alleles. The dataset was composed of 202 and 201 binders to the molecules HLA-DRB1*03:01 and HLA-DRB1*04:01, respectively. In (**a**) and (**b**) are shown the logos of the two motifs identified by the algorithm, with the first cluster predominantly composed of DRB1*03:01 binders and the second of DRB1*04:01 binders. (**c**) confusion matrix for the two classes of binders, the correlation coefficient is MCC = 0.59

3.3 Gibbs clustering as a tool to remove noise from data

In the previous examples, we assumed that all sequences fit into one cluster or another. However, experimental data often contains some level of noise, and hence peptides that may not fit in any of the motifs. The Gibbs clustering algorithm allows, by the inclusion of a trash cluster, a very simple yet highly effective manner to detect such spurious peptides and remove them from the motif identification (see Section 2 for the implementation).

In Supplementary Table S2 is shown the effect of the trash cluster on mixtures of 1, 2, 3 and 4 MHC class I alleles polluted with 50 random peptides. We observed that the majority of the random peptides were placed into the trash cluster, but an average of ~5 peptides were assigned to one of the clusters. This fits the overall expectation as 1–5% of random natural peptides are estimated to bind to a given MHC class I molecules (Rao *et al.*, 2009; Yewdell and Bennink, 1999). Furthermore, most of the random peptides that were inserted into one of the clusters had consistently lower scores than the actual binders (Supplementary Fig. S4). The Gibbs clustering algorithm allows obtaining different levels of sensitivity and specificity by varying the threshold to assign a peptide to the trash cluster. Increasing this threshold would remove more noise (peptides with low cluster score) from the dataset, but at the same time would increase the number of binders placed in the trash. In the experiments with noisy data (Supplementary Table S2), a few sequences measured to be binders to a given allele are assigned to the trash (2 for the 1 clusters case, 2 for 2 clusters, 2 for 3 clusters, 4 for 4 clusters). Interestingly, none of these peptides appear to match the binding motifs of the alleles they were measured to bind to. Using the state-of-the-art MHC class I binding prediction method NetMHCcons (Karosiene *et al.*, 2012), these peptides all show extremely low predicted binding affinity to their respective HLA restriction element (>10 000 nM, see Table 1). Furthermore, an experimental re-examination of three of these peptides confirmed that they are indeed non-binders to their respective HLA molecule (J. Sidney, personal communication). The method was thus able, while grouping distinct specificities into different clusters, to also identify false positives that most likely correspond to erroneous measurements in the experimental assay. Introducing the trash bin for the MHC class II benchmark also led to an improved clustering performance, removing two outlier peptides, maintaining the optimal solution to consist of two clusters and enhancing the performance to MCC = 0.62 (data not shown).

3.4 SH3 domains

The Src Homology 3 domain (SH3 domain) is a small protein interaction module abundantly found in eukaryotes. SH3 domains consist of ~60 amino acids and have been shown to mediate protein–protein interactions by preferentially binding to short proline-rich sequences (Yu *et al.*, 1994). The minimal consensus sequence for SH3 domain binding is composed of two prolines located two amino acids apart (PxxP), but it is commonly recognized that there exist two main classes of binders: class I ligands having a general consensus sequence +xϕPxϕP and class II ligands with consensus sequence ϕPxϕPx+ (where + is a positively charged amino acid, usually R, ϕ is a hydrophobic amino acid, and x any amino acid)

Table 1. Measured, predicted and re-tested binding affinities (in nM) for peptides assigned to the trash cluster

Peptide	HLA	IEDB[a]	Predicted[b]	Validated[c]
DHHFTPQII	A*01:01	62	28 485	24 822
SQTSYQYLI	B*07:02	248	24 349	49 928
NAFGWENAY	B*07:02	350	24 481	—
TVFKGFVNK	B*27:05	235	13 723	—
ELPIVTPAL	B*40:01	314	15 208	—
ADKNLIKCS	B*40:01	316	33 324	76 190

As a rule of thumb, generally affinity < 50 nM identifies a strong binder, 50 nM < affinity < 500 nM a weak binder, affinity > 500 nM non-binders.
[a]Binding affinity deposited in the Immune Epitope Database.
[b]Predicted binding affinities using NetMHCcons.
[c]Re-tested binding affinities after detection as outliers.

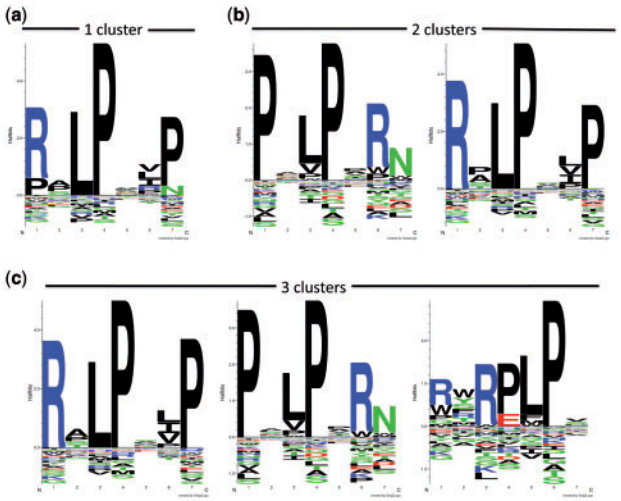

Fig. 3. Sequence motifs on SH3 domain binding data clustered in 1 to 3 clusters. (**a**) Sequence motif of the dataset aligned in one single cluster. The cluster contains 2360 peptides, 97 peptides were discarded to the trash cluster. (**b**) Sequence motifs for SH3 domain data split in two clusters. The two groups are in strong agreement with the canonical class I (right, 1892 peptides) and class II (left, 498 peptides) types of SH3 domain ligands. Sixty-seven peptides were moved to the trash cluster. (**c**) Sequence motifs when the data is split in 3 clusters. The clusters have sizes of 1606, 490 and 305 peptides, respectively, with 56 peptides discarded to the trash cluster

(Mayer, 2001). However, there are a few exceptions to these predominant motifs, and a number of non-consensus ligands have been identified (reviewed in Carducci *et al.*, 2012; Saksela and Permi, 2012).

The Gibbs clustering algorithm was run on a large dataset of 2457 peptides binding to the Src SH3 domain. The peptides are 12 amino acids long and unaligned with respect to the binding motif(s) to the SH3 domain. As the dataset may contain non-consensus ligands and noise, we performed the alignment/clustering with the addition of a trash cluster, which collects peptides that do not match any of the main motifs. To ensure the removal of non-consensus sequences that may only partially match the major motifs, the baseline for the trash cluster was set to a relatively high value of 10. The sequence motifs identified by the Gibbs clustering are shown in Figure 3. Aligning all sequences into a single cluster (Fig. 3a) showed the characteristic PxxP pattern, in this case preceded by a leucine (L) and arginine/proline (R/P) three positions back. Clustering the peptides into two groups revealed the two sequence motifs shown in Figure 3b. They correspond very well to the two known classes of SH3 domain ligand, one with the PxϕPxRN pattern (class II) and the other with pattern RxϕPxϕP (class I). Dividing the dataset further and creating 3 clusters led to the emergence of a new subset of specificity (panel c) besides the two described in the 2-clusters case. Although several exceptions to the two main classes have been discovered (Saksela and Permi, 2012), this RxRPϕP pattern has not, to the best of our knowledge, been described before. Splitting the dataset further to more than 3 clusters does not show new specificities besides those described here.

The two motifs displayed in Figure 3b agree strongly with the results obtained in a previous study (Kim *et al.*, 2012), where the MUSI method was applied to the same phage display dataset. The Gibbs clustering method, however, has the strong advantage compared with MUSI, in that the data do not need to be aligned before clustering. Instead, in the Gibbs clustering method, alignment and clustering are performed simultaneously. In the specific case of SH3 domain binding, where both motifs share a strong common PxxP pattern, a pre-alignment strategy to a common motif, like the one implemented in MUSI, can be successful. However, in the general case, the different motifs will be weak and will not share a common pattern. On such data, it becomes difficult if not impossible to accurately identify the binding core within the peptide dataset using alignment techniques (Nielsen *et al.*, 2004). For instance, by applying the MUSI method on the MHC class II dataset from above, we found the solution with two motifs being sub-optimal compared with a solution with a single motif. Forcing MUSI to generate two clusters, the overall performance was MCC = 0.21, which is significantly lower than what was obtained using the Gibbs clustering method ($P < 0.01$, bootstrap test).

3.5 Sub-specificities of MHC class I molecules

Peptide binding to MHC molecules is one of the most selective steps in determining MHC class I-restricted CTL responses. The strength of this interaction is commonly measured in terms of binding affinity between peptide and MHC complex. However, not all peptides with high affinity are immunogenic, indicating the presence of other factors determining an effective response (Assarsson *et al.*, 2007). Some studies have suggested that the stability of the MHC-peptide complex is a major player in determining immunogenicity (Busch and Pamer, 1998; Geironson *et al.*, 2012; Harndahl *et al.*, 2012).

By means of the Gibbs clustering algorithm, we investigated whether there exist sub-specificities for MHC class I binding and whether these sub-specificities correlated with different levels of affinity and/or stability. For this purpose, we used a dataset recently published by Harndahl *et al.*, 2012 consisting of 650 peptides binding with affinity stronger than 500 nM to HLA-A*02:01 for which also the peptide stability had been measured. We applied the Gibbs clustering algorithm to split the

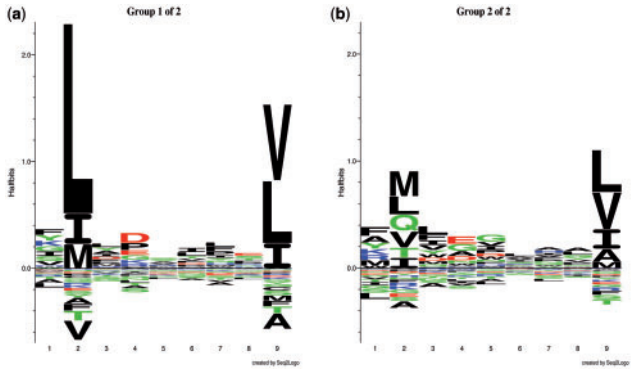

Fig. 4. Sub-motifs of HLA-A*02:01 binding specificity. The peptides in the two clusters have similar affinity but differ significantly in stability. The sequence logo in the left panel is composed mainly of stable peptides (Th ≈ 5.7 h), whereas peptides in the second group have lower stability (Th ≈ 2.1 h)

dataset in two clusters using default parameters and investigated the properties of the sequences in the two groups.

The sequence motifs for the resulting clusters are shown in Figure 4. The first cluster (G1), composed of 441 sequences, was highly specific in terms of amino acid preference, with [LIM] at P2 and [VLI] at P9. The contribution from other positions is secondary. The second cluster (G2) is more promiscuous at both anchor positions P2 and P9, especially at P2 where several amino acids other than L, I and M are allowed. The peptides in the two groups had a median binding affinity of 6 nM and 9 nM, for G1 and G2, respectively. This difference is not significant ($P = 0.095$, Wilcoxon rank-sum test). In contrast, we observed that peptides in G1 have a significantly higher stability compared with G2 ($P < 10^{-6}$, Wilcoxon rank-sum test): the median half-life of the MHC-peptide complex in G1 is Th ≈ 5.7 h, whereas in G2, it is only Th ≈ 2.1 h.

From these results, we can conclude that the method identified subtle differences between the binders to HLA-A*02:01 that appear to differentiate stable binders from unstable binders. In particular, as previously noted peptide-HLA-A*02:01 complexes appear to be destabilized by a sub-optimal amino acid in just one of the two anchor positions and in particular position P2 (Harndahl *et al.*, 2012).

4 DISCUSSION

We proposed an efficient algorithm to identify multiple specificities in peptide datasets. The applications of the method are numerous, ranging from the deconvolution of poly-specificities contained in a dataset, to the analysis of sub-specificities within a known binding motif. The algorithm aims at identifying the solution (the set of clusters and corresponding alignments) that optimally fits the peptide dataset. The optimal solution is automatically selected and the identified binding motifs are visualized as individual sequence logos. Using a panel of benchmark datasets, we have demonstrated the power of the Gibbs clustering method in deconvoluting poly-specificities contained both in pre-aligned and unaligned peptide datasets covering the MHC class I, MHC class II and human SH3 domain systems.

Gibbs sampling is a powerful approach to explore large spaces of possible solutions. In the case of amino acid sequences, there are immense possible ways of aligning and clustering them as soon as the number of sequences becomes bigger than a handful. The probabilistic nature of Gibbs sampling allows efficient sampling of the search space and convergence towards a state of high fitness of the system. Compared with other motif identification methods, Gibbs clustering is unique in that it incorporates alignment and clustering in a set of alternative sampling moves, allowing for simultaneous identification of clusters and optimal sequence alignment. This property makes the method capable of identifying subtle and relatively weak binding motifs (as demonstrated for the case of MHC class II binding motifs), but it comes at the price of computational speed. Analysing the 400 peptides in the MHC class II binding dataset takes a little >5 min using Gibbs clustering. This running time is reduced to 15 s using the MUSI algorithm (Kim *et al.*, 2012) yet at the cost of a dramatic and significant drop in accuracy.

In a general situation, it is not known a priori how many motifs are contained in a dataset. When presented with a set of experimental data, the investigator ideally wants a definitive answer to the question: 'How many motifs are contained in my data?' Unfortunately, the answer is not unambiguous, not so much for a fault of mathematical and computational methods, rather for the ambiguity of the question. The answer depends on the level of resolution that is expected for the particular problem at hand. If the goal is a rough classification of sequences based on global differences, then the resulting number of clusters will be small. Conversely, more partitions would be produced if we were searching for subtler distinguishing sequence characteristics. The 'true' number of clusters is therefore not an objective answer but depends on the kind of biological question that is being asked. In the Gibbs clustering algorithm, we introduce a parameter λ that aims to modulate the degree of resolution required by the user. High λ penalizes overlap between clusters and tends to create coarser clusters, whereas low λ results in smaller and specialized clusters. For example, we showed that for a certain value of λ, we could accurately identify the number of MHC class I molecules contained in a dataset of mixed specificities. In another example, we split one of these very same specificities into sub-motifs and looked for subtle differences in a rather homogenous population of peptides. These are not the extremes: one could conceive partitioning the data further into more specialized sub-populations, as well as obtaining a coarser picture of similarities between alleles. The same data may have different levels of resolution depending on the aim of the analysis, and the investigator should keep this in mind when using a classification method like the one presented here. The Gibbs clustering method in its current form is limited to handle situations where motifs are of uniform length. Likewise, the method can only handle amino acid input data. The reason for this limitation is that most of its unique features like pseudo-count estimates from Blosum substitution matrices and sequence weighting of are specific for amino acid data.

In conclusion, we believe the Gibbs clustering method to be both a highly accurate and very user-friendly tool that will allow researchers to interpret peptide datasets in terms of receptor specificities in a highly intuitive manner. Therefore, we expect it to become an important tool as large-scale peptide chip

technologies grow to be a cost-effective and accessible platform for investigation of protein-ligand interactions. The method is highly customizable and publicly available as an online web-server at http://www.cbs.dtu.dk/services/GibbsCluster.

ACKNOWLEDGEMENTS

The authors thank John Sidney (La Jolla Institute for Allergy and Immunology, California, USA) for the binding affinity validation of the predicted MHC class I outliers.

Funding: The research leading to these results received funding from the European Union Seventh Framework Program FP7/2007-2013 under grant agreement n° 222773).

Conflict of Interest: none declared.

REFERENCES

Andreatta,M. *et al.* (2011) NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. *PLoS One*, **6**, e26781.

Assarsson,E. *et al.* (2007) A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection. *J. Immunol.*, **178**, 7890–7901.

Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.*, **21**, 51–80.

Bailey,T.L. *et al.* (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.

Bratkovič,T. (2010) Progress in phage display: evolution of the technique and its applications. *Cell. Mol. Life Sci.*, **67**, 749–767.

Busch,D.H. and Pamer,E.G. (1998) MHC class I/peptide stability: implications for immunodominance, in vitro proliferation, and diversity of responding CTL. *J. Immunol.*, **160**, 4441–4448.

Carducci,M. *et al.* (2012) The protein interaction network mediated by human SH3 domains. *Biotechnol. Adv.*, **30**, 4–15.

Geironson,L. *et al.* (2012) Stability of peptide-HLA-I complexes and tapasin folding facilitation—tools to define immunogenic peptides. *FEBS Lett.*, **586**, 1336–1343.

Gfeller,D. (2012) Uncovering new aspects of protein interactions through analysis of specificity landscapes in peptide recognition domains. *FEBS Lett.*, **586**, 2764–2772.

Gfeller,D. *et al.* (2011) The multiple-specificity landscape of molecular peptide recognition domains. *Mol. Syst. Biol.*, **7**, 484.

Halperin,R.F. *et al.* (2011) Exploring antibody recognition of sequence space through random-sequence peptide microarrays. *Mol. Cell. Proteomics*, **10**, M110.000786.

Harndahl,M. *et al.* (2012) Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *Eur. J. Immunol.*, **42**, 1405–1416.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Karosiene,E. *et al.* (2012) NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics*, **64**, 177–186.

Kim,T. *et al.* (2012) MUSI: an integrated system for identifying multiple specificity from large peptide or nucleic acid data sets. *Nucleic Acids Res.*, **40**, e47.

Koivunen,E. *et al.* (1999) Identification of receptor ligands with phage display peptide libraries. *J. Nucl. Med.*, **40**, 883–888.

Lawrence,C.E. *et al.* (1993) Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lund,O. *et al.* (2004) Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics*, **55**, 797–810.

Mayer,B.J. (2001) SH3 domains: complexity in moderation. *J. Cell Sci.*, **114**, 1253–1263.

Nielsen,M. and Lund,O. (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics*, **10**, 296.

Nielsen,M. *et al.* (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, **20**, 1388–1397.

Noguchi,H. *et al.* (2002) Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J. Biosci. Bioeng.*, **94**, 264–270.

Rammensee,H. *et al.* (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.

Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 846–850.

Rao,X. *et al.* (2009) A comparative study of HLA binding affinity and ligand diversity: implications for generating immunodominant CD8+ T cell responses. *J. Immunol.*, **182**, 1526–1532.

Saksela,K. and Permi,P. (2012) SH3 domain ligand specificity: what's the consensus and where's the specificity. *FEBS Lett.*, **586**, 2609–2614.

Schutkowski,M *et al.* (2005) Peptide arrays for kinase profiling. *ChemBioChem*, **6**, 513–521.

Soen,Y. *et al.* (2003) Detection and characterization of cellular immune responses using peptide-MHC microarrays. *PLoS Biol.*, **1**, e65.

Thomsen,M.C. and Nielsen,M. (2012) Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.*, **40**, 1–7.

Uttamchandani,M. and Yao,S.Q. (2008) Peptide microarrays: next generation biochips for detection, diagnostics and high-throughput screening. *Curr. Pharm. Des.*, **14**, 2428–2438.

Wang,P. *et al.* (2010) Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinformatics*, **11**, 568.

Yewdell,J.W. and Bennink,J.R. (1999) Mechanisms of viral interference with MHC class I antigen processing and presentation. *Annu. Rev. Cell Dev. Biol.*, **15**, 579–606.

Yu,H. *et al.* (1994) Structural basis for the binding of proline-rich peptides to SH3 domains. *Cell*, **76**, 933–945.