BMC Bioinformatics

RESEARCH

Open Access

SwarmTCR: a computational approach to predict the specificity of T cell receptors

Ryan Ehrlich[1], Larisa Kamga[2], Anna Gil[3], Katherine Luzuriaga[2], Liisa K. Selin[3] and Dario Ghersi[1*]

*Correspondence:
dghersi@unomaha.edu
[1] School of Interdisciplinary Informatics, College of Information Science and Technology, University of Nebraska at Omaha, 1110 S 67TH, Omaha, NE 68182, USA
Full list of author information is available at the end of the article

Abstract

Background: With more T cell receptor sequence data becoming available, the need for bioinformatics approaches to predict T cell receptor specificity is even more pressing. Here we present SwarmTCR, a method that uses labeled sequence data to predict the specificity of T cell receptors using a nearest-neighbor approach. SwarmTCR works by optimizing the weights of the individual CDR regions to maximize classification performance.

Results: We compared the performance of SwarmTCR against another nearest-neighbor method and showed that SwarmTCR performs well both with bulk sequencing data and with single cell data. In addition, we show that the weights returned by SwarmTCR are biologically interpretable.

Conclusions: Computationally predicting the specificity of T cell receptors can be a powerful tool to shed light on the immune response against infectious diseases and cancers, autoimmunity, cancer immunotherapy, and immunopathology. SwarmTCR is distributed freely under the terms of the GPL-3 license. The source code and all sequencing data are available at GitHub (https://github.com/thecodingdoc/SwarmTCR).

Keywords: TCR, Immunoinformatics, Binding specificity

Background

The adaptive immune system plays a critical role in curbing infections and in cancer immunosurveillance, defined as the patrolling of the body by the immune system with the active elimination of precancerous and cancerous cells [1]. CD8+ T lymphocytes are one of the key cell types involved in antiviral responses and cancer immunosurveillance. They perform their function by binding to small peptides presented on the surface of highly polymorphic molecules known as the Major Histocompatibility Complex (MHC) using T Cell Receptors (TCR). TCRs are transmembrane proteins that contain either α and β or γ and δ chains, within which are three loops called Complementarity Determining Regions (CDRs). CDR loops are characterized by both germline loops (CDR1 and CDR2) and the hyper-variable CDR3 loop, which is the product of somatic recombination [2, 3]. These CDR loops are responsible for interacting with the peptide/MHC (pMHC) complex. The diversity of TCR sequences is mostly focused on the CDR regions

and is very large, with numbers in human that are thought to exceed 10^{20} possible distinct receptors [4].

The collection of TCRs possessed by an individual is known as the T cell repertoire, which is shaped over time by the history of infections in combination with stochastic factors, and is in turn responsible for determining the outcome of an immune response. One of the ultimate goals of T cell repertoire analysis is predicting the specificity of the T cells of an individual using sequence information alone [2]. This would entail determining the identity of the peptide(s) that each TCR is capable of recognizing by computationally analyzing the TCR sequences from CD8+ cells circulating in the peripheral blood of an individual. More high-quality TCR sequencing data and peptide binding information specificity will be needed to achieve this objective. On the sequencing front, although the availability of TCR sequencing data is still fairly limited, the field has seen steady progress and technological advancements [5].

Two main technologies are currently available for sequencing TCRs: (1) single cell (SC) sequencing and (2) bulk sequencing (BS). SC TCR sequencing technology allows to reconstruct the complete sequence of a TCR with paired α and β chain sequence information, but its cost is still limiting the amount of available data. In contrast, BS technology is more affordable and has yielded substantially larger amounts of data, but reconstructing the correct α and β chain pairs within a TCR is not possible with this technology.

Being able to map the specificity of human repertoires can equip us with powerful new tools for studying autoimmunity, cancer immunotherapy, and immunopathology [6]. However, for these methods to be broadly applicable it is critical to sample T cell repertoires deeply and in multiple individuals, as well as to account for the diversity of binding topologies to pMHCs with computational approaches. Here we introduce SwarmTCR, a computational method to predict the specificity of TCRs for class I MHC/peptide complexes that compares favorably to the nearest-neighbor based approach TCRdist [2] on both SC and BS data.

TCRdist uses a nearest-neighbor approach with a pairwise sequence alignment score between TCRs as a proximity measure. The two chains are weighted equally, and the CDR3 region is weighted three times more than the other CDR regions. While this is a reasonable choice considering the importance of CDR3 for peptide binding, it does not take into consideration the fact that the two chains and the regions within them might have different levels of involvement in binding to the pMHC, depending upon the peptide being presented and the MHC type. In a recent study, we curated a non-redundant set of TCR/pMHC crystal structures and explored binding topologies of TCR/pMHC complexes and the number of contact residues ($\leq$ 4.5 [7]) made by α and β chains with pMHC [8]. Our results indicated a wide range of TCR binding angles and a variable use of the α (7–25 contacts) and β (6–22 contacts) chains in making contacts with the pMHC. We also computed the number of alpha and beta contacts to the pMHC, determining a ratio of contacts (α/β ratio) for each structure. In some complexes, the α chain had a much larger number of interactions with the pMHC than the β chain (corr. $= 0.77, p < 1.6x10^{-14}$), whereas in other complexes the β made more interactions than the α chain (corr. $= -0.73, p < 2.2x10^{-12}$). In other complexes, we saw an almost equal number of $\alpha\beta$ interactions with the pMHC ($\sim$15 contacts per chain). Taken together,

these results suggest a wide range of binding recognition modes, which should be reflected in a computational method to predict TCR binding specificities.

As a first step to leverage these findings, we developed SwarmTCR, a method to predict TCR specificity that automatically learns the optimal set of weights to assign to each CDR region based on classification accuracy in a cross-validation setting (Fig. 1). In addition to CDR1, CDR2, and CDR3, the method also incorporates the CDR2.5 region (a loop between CDR2 and CDR3 that can interact with the pMHC, as discussed in [2]), for a total of four weights per chain. By directly optimizing the weights for CDR regions in an peptide-specific fashion, our method automatically accounts for the diversity in pMHC recognition that is documented in crystal structures (see Methods).

We applied our method to SC and BS data and compared its performance against that of TCRdist. In addition to performing in most cases better than TCRdist, the weights returned by SwarmTCR in SC sequencing data can potentially inform the user about the contribution of the two chains in recognizing the pMHC complex.

Results

Classification performance of SwarmTCR

The rationale for developing SwarmTCR is that receptor α and β chains can be involved in peptide recognition to a variable extent. Figure 2 shows two crystal structures of TCR/pMHC complexes to visually illustrate the fact that the α and β chains can be involved in pMHC binding to a very different extent, depending on the peptide that is being recognized. In the example shown in Fig. 2, out of the total number of residues making contact with the pMHC, one TCR (PDB ID: 4G8G [9]) has 16 (59%) α chain residues and 11 (41%) β chain residues in contact with the pMHC, whereas the other [10] has 9 (39%) α chain residues and 14 (61%) β chain residues in contact with the pMHC. This is consistent with results in the literature [8].

Based on this observation, SwarmTCR optimizes the weights used to compute the CDR alignment scores underpinning the nearest-neighbor classification approach. In contrast, previous attempts at predicting TCR specificity (TCRdist method) used a static weighting scheme with equal α and β chain contributions and fixed CDR loop

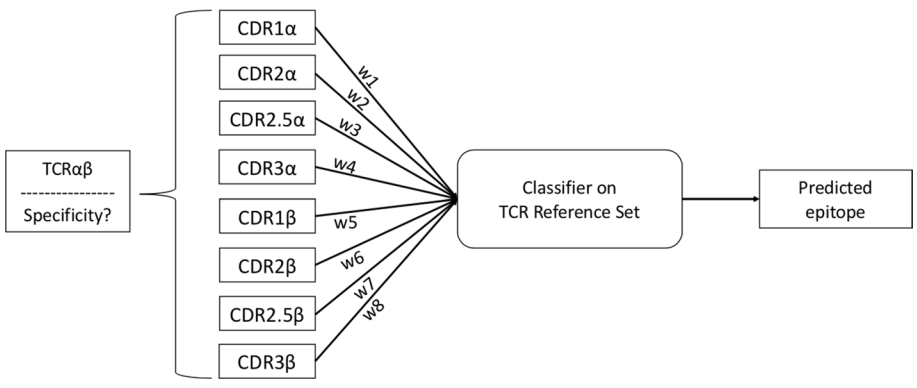
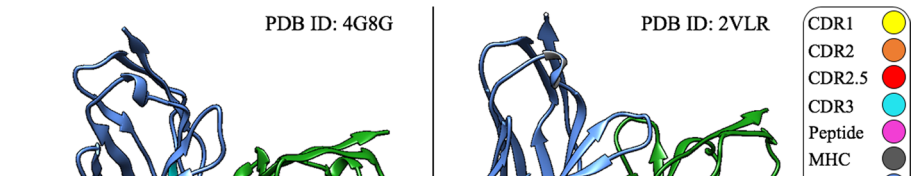


Fig. 1 Overall approach. Here the model for single-cell data is illustrated. Weights for each CDR loop are determined by the optimization step and subsequently tested on the testing set to assess peptide prediction performance



Chain, Target	CDR1	CDR2	CDR2.5	CDR3	Total
α, peptide	4	1	0	3	8
α, MHC	2	3	0	3	8
β, peptide	0	0	0	5	5
β, MHC	2	2	0	2	6

Chain, Target	CDR1	CDR2	CDR2.5	CDR3	Total
α, peptide	0	0	0	3	3
α, MHC	1	1	0	4	6
β, peptide	1	2	0	3	6
β, MHC	0	4	0	4	8

Fig. 2 Contact residues in crystal structures.The complexes shown here (PDB ID: 4G8G, peptide: KRWIILGLNK and PDB ID: 2VLR, peptide: GILGFVFTL) express the need for SwarmTCR. The 4G8G complex illustrates an α driven interaction and 2VLR conversely, a β driven interaction. All protein chains (including the CDR loops) are color-coded to supplement the tables beneath each structure. The tables show the number of contact residues in each CDR loop and a target structure

Fig. 3 CDR weights and performance. These boxplots summarize the results of SwarmTCR and compare them against TCRdist. **A**, **C** SC and BS SwarmTCR results describe weights (y-axis) selected for each CDR loop (x-axis) for each repertoire tested. **B**, **D** SC and BS performance comparison of SwarmTCR and TCRdist compare average precision scores (y-axis) for each repertoire tested (x-axis). *P*-values for performance comparison are defined by two-sample independent t-test

weights [2]. The SwarmTCR method makes no assumptions about chain or CDR loop importance, but learns the weights in an peptide-specific fashion.

Mean and standard deviation of the optimized weights for several peptides are shown in Fig. 3 (numerical values in Additional file 1: Table 1), together with

classification performance for SwarmTCR and TCRdist [2], separately for SC and BS data. To test the robustness of the results, we repeated the same analysis using TCRs from IEDB at different confidence thresholds (0, 2, 3), obtaining similar results (see Methods and Additional file 1: Table 2). In addition, in the absence of true negative data (i.e., data showing which TCRs do not bind to particular epitopes), we randomly shuffled CDR regions within each chain (alpha and beta), for all TCR sequences with our existing single-cell data. As expected, for nearly all repertoires we observed a notable loss in precision, with IAV-M1 showing a less pronounced loss in precision (see Methods and Additional file 1: Table 2).

Additional area under receiver operating characteristic curve (AUROC) analysis can be found in Additional file 1: Table 3, SC true positive rate (TPR)/false positive rate (FPR) boxplots in Additional file 1: Figs. 1 and 2, and BS TPR/FPR boxplots in Additional file 1: Figs. 3 and 4.

Single cell sequencing

SC data provides paired $\alpha\beta$ chain information, i.e., the complete TCR sequence is available. The SwarmTCR optimization procedure for SC data involves the use of eight separate weights, since we have paired α and β chain sequences. The results of our SC analysis show relatively high weight being placed on non-CDR3 loops, although the CDR3 region has high weight for several peptides (Fig. 3A, and Additional file 1: Fig. 5). Interestingly, in the case of the EBV YVL peptide and the Yellow Fever LLW (peptide: LLWNGPMAV) peptide the SwarmTCR optimization procedure assigns more weight to the α chain, suggesting that the α and β chains might have a more or less prominent role in TCR peptide recognition depending on the peptide, which is consistent with the example shown in Fig. 2 and the previous literature [8].

By looking at the results in Fig. 3B and in Fig. 4B (Additional file 1: Table 1), we can see that the largest difference between the classification performance of TCRDist [2] and SwarmTCR is for the EBV YVL and GLC peptides. The optimized weights for these peptides differ substantially from the fixed TCRdist weights. Based on the optimized weights, YVL appears to favor the α chain as noted above, with only CDR2β being weighted more than its α counterpart. This is consistent with results in the literature [11].

PR curves for all SC peptides are shown in Additional file 1: Fig. 6, AUROC results in Additional file 1: Table 3, SC TPR/FPR box plots in Additional file 1: Figs. 1 and 2. The distributions of alignment scores between test and reference TCRs for positive (i.e., binding) and negative (i.e., "non binding") TCRs is shown in Additional file 1: Fig. 7. For most epitopes (with the notable exception of NLV, which has poor performance), we can observe a clear separation of scores between positives and negatives.

SwarmTCR weights correlate with structural contacts

We further explored the potential of SwarmTCR to infer TCR chain usage in binding to the pMHC by extracting contact residue counts from TCR/pMHC crystal structure CDR regions (see Methods). Figure 5A shows that the weights generated by SwarmTCR correlate in a statistically significant manner (PCC $= 0.812$, $p < 0.05$) with actual chain usage for TCR/peptide contacts, compared to TCRdist (PCC $= 0.484$, $p < 0.331$).

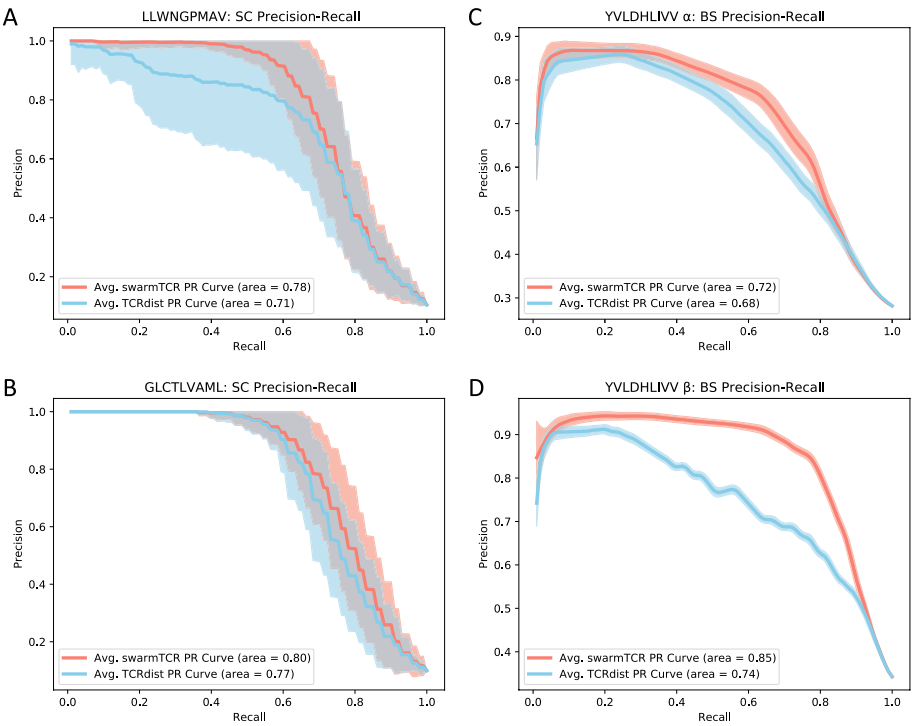


Fig. 4 Precision-recall curves for SwarmTCR and TCRDist.These precision-recall curves show the performance of SwarmTCR and TCRdist on the data used for 50 cross-validation iterations. TCRdist mean curves are in blue and SwarmTCR mean curves are in red, while the shaded regions cover one standard deviation

Though number of contacts increase when MHC contacts are included (Fig. 5B), SwarmTCR weights maintain stronger correlation (PCC $= 0.827$, $p < 0.042$) compared to TCRdist (PCC $= 0.645$, $p < 0.166$). We performed the same analysis on CDR regions (Additional file 1: Figs. 8, 9, and 10), obtaining lower correlation values. However, SwarmTCR does appear to capture germline loops with high contact counts. Contact counts for all PDB structures can be seen in Additional file 1: Figs. 11 and 12.

Bulk sequencing

As mentioned in the Introduction, in contrast to SC sequencing bulk sequencing yields the sequence of only the α or the β chain of TCRs but not both. The SwarmTCR optimization procedure was carried out in the same manner as for SC, except that the weights to optimize are four instead of eight, since we have unpaired α or β chain sequences, containing CDR1, CDR2, CDR2.5, and CDR3 regions for a total of four weights per chain. Compared to SC data, the results on BS data show more weight being placed on CDR3 loops, indicating its importance in predicting the specificity of TCR data when using only one chain. While [2] assigns to the CDR3 loop three times the weight of the CDR1, CDR2, and CDR2.5 regions, SwarmTCR assigns to CDR3 between 4 and 64 times the weight of the other regions (using the average weight as a measure), as it can be seen in Fig. 3C and Additional file 1: Fig. 13. These differences between the SwarmTCR weights and the original weights by Dash et al. [2] have a substantial impact on the classification performance for the GLC and YVL peptides using the β chains (Fig. 3D).

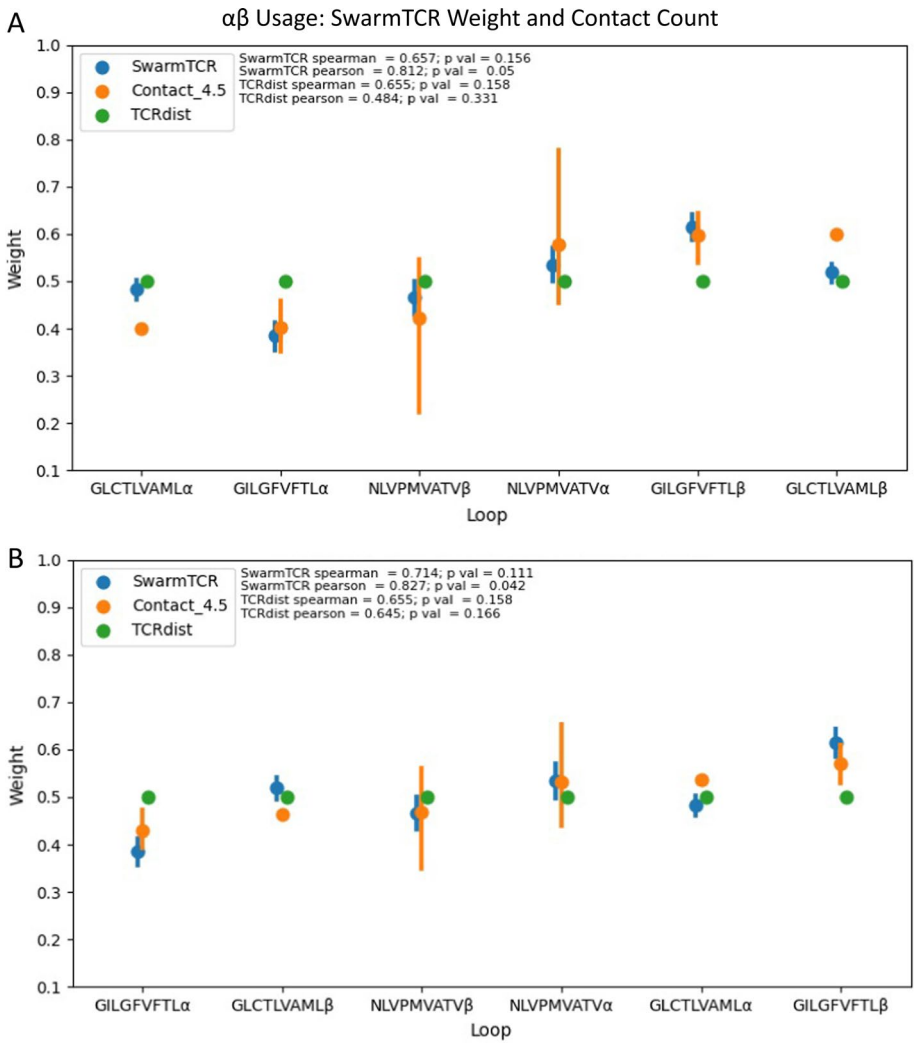


Fig. 5 Comparison of SwarmTCR weights and contact residues in crystal structures. These plots compare $\alpha\,\beta$ chain usage for SwarmTCR and TCRdist, using known crystal structures as a baseline. The x-axis details the normalized weight for each repertoire's chain and the y-axis is ordered in ascending order according to crystal structure average number of contacts. Plot **A** includes the number of CDR loop contacts to the peptide and plot **B** includes number of CDR loop contacts to the pMHC. Pearson and Spearman statistics are located to the right of the legend

Figure 4C and D shows precision-recall (PR) curves for a representative peptide (EBV YVL), obtained by averaging 50 curves, with the shaded region representing one standard deviation above and below the mean. SwarmTCR outperforms the original weights used in [2] for both chains, with a more substantial improvement for the β chain (AUCPR 0.85 with the optimized weights vs. 0.74 with the original TCRdist weights). PR curves for all BS peptides are shown in Additional file 1: Fig. 14, AUROC results in Additional file 1: Table 3, SC TPR/FPR box plots in Additional file 1: Figs. 3 and 4.

Discussion

We introduced SwarmTCR, a computational approach for predicting TCR specificity that maximizes classification performance within a nearest neighbor framework by identifying optimal CDR weights. Compared to the results obtained with fixed TCRdist weights, overall SwarmTCR performs better, with some peptides showing more substantial improvement than others (Fig. 4). We note that in a worst case scenario, with enough data SwarmTCR can always fall back on the weights used by TCRdist if those yield maximum performance during the PSO step.

When comparing CDR weights in SC and BS data, we noticed stark differences between the results obtained with the two data types. In particular, we found that SwarmTCR assigns much more weight to the CDR3 region in BS data, whereas the results on SC data show relatively higher weights for the germline CDR loops. Due to the small size of the SC dataset, the diversity of TCR gene families is likely considerably lower than that found in the BS dataset. Therefore, the lower gene family diversity in the SC dataset compared to BS could partly explain the higher predictive power of gene family (germline loops) in SC data. Another reason for this difference in the weights between the two data types is the presence of paired chains information in SC, where combinations of TCR genes for α and β chains would likely be selected for by the optimization approach. More SC data is needed to further elucidate the issue. Consistent with the substantial differences in size between the two datasets, SC results show higher variance in both performance and weight selection than BS results.

The performance, generarizability, and robustness of computational approaches depend on the quality of the data used for training. An important caveat to consider when using publicly available databases like IEDB and VDJdb is that they might contain data obtained in a specific experimental context and not further validated. For example, confounding factors like bystander activation of CD8+ cells (i.e., activation of T cells that is independent of the TCR [12]) can potentially lead to incorrect assignment of TCR specificity.

An important question to consider is whether the optimized weights can also be interpreted to reflect chain and CDR usage. In other words, if a chain or a CDR region receives a high weight during the optimization step, does that mean that it also makes a large number of contacts with the pMHC? Our results suggest that the optimized weights can point to possible TCR chain and CDR loop usage, as shown in Figs. 3A, 5, and Additional file 1: Table 1 for the GIL TCR/pMHC crystal structure (Fig. 2, PDB ID: 2VLR) with respect to β chain dominance and CDR2β loop usage.

A recent study [11] corroborates these findings, explaining CDR1β and CDR2β's role in pMHC recognition, as well as CDR3β's conserved arginine fitting into a pocket between the peptide and the MHC α2-helix. Additionally, this study explains CDR3β's sequence conservation and notable variability in CDR3α. This likely explains the weighting results of SwarmTCR (GIL, Fig. 3), despite the number of contacts in the 2VLR TCR/pMHC structure. We also note that SwarmTCR's weight results for YVL and LLW repertoires align with findings from this study, indicating the importance of the α chain in pMHC recognition [11].

However, one has to exercise caution when interpreting the weights in a structural sense. As discussed above, the weights are the result of an optimization process designed

to maximize classification performance, and factors other than structural importance can play a role in determining the optimal weights. If we consider the crystal structures and literature mentioned, this is shown by our BS weighting results and differences present in Additional file 1: Figs. 8, 9, and 10. Nonetheless, given large amounts of TCR sequence data, peptide-specific optimal weights can provide helpful information in elucidating TCR/pMHC interactions.

Sequence-based approaches to infer TCR specificity are appealing due to their computational efficiency and the availability of sequence data [2, 6, 13]. However, structural data continue to provide information that expands and sometimes challenges our current understanding of TCR/pMHC interactions. For example, one study found a strong negative correlation between mean CDR3 α, β charge and peptide charge [2]. Another study [3] showed how cross-reactive peptides share similar pMHC features (structural motifs and electrostatic potential) despite having different peptide sequences. These findings point to the importance of factoring in structural information for further improving prediction methods. However, more work needs to be done both at the experimental level (generation of more crystal structures) and the computational level (reliable and scalable modeling of TCRs and pMHC complexes).

Conclusions

Being able to reliably predict TCR specificity will push the boundaries of many disciplines including vaccine design, immunotherapy, cancer research, and disease detection/prevention in new directions. Here we have introduced SwarmTCR, a nearest-neighbor approach that optimizes CDR weights by maximizing classification performance. SwarmTCR was benchmarked on both SC and BS data, and compared against a state-of-the-art methodology, TCRdist. The results showed that SwarmTCR improves the performance of the nearest-neighbor classification approach and that the CDR weights generated in the training phase tend to correlate with the number of contacts made by the CDR regions in crystal structures.

Methods

TCRs sequence data

CD8+ TCR SC and BS data were collected from: (1) the Selin and Luzuriaga Labs at UMASS; (2) VDJdb [14]; and (3) IEDB [15].

Data acquired from the Selin and Luzuriaga labs contained TCRs isolated from HLA A:02:01-restricted, naïve and peptide-specific CD8+ T cells binding to YVL (EBV-BRLF1$_{109}$: HLA-A:02:01 restricted, peptide: YVLDHLIVV), GLC (EBV-BRLF1$_{300}$:HLA-A:02:01 restricted, peptide: GLCTLVAML), and GIL (IAV-M1$_{58}$: HLA-A:02:01 restricted, peptide: GILGFVFTL). SC data were obtained from ex vivo single-cell sequencing of CD8 T cells from peripheral blood mononuclear cells (PBMCs) of four adult donors. Further information on these data can be found in [11]. BS data were obtained from ex vivo bulk sequencing of CD8 T cells from PBMCs of three adult donors. Further information on these data can be found in [16].

Human data from VDJdb was downloaded in January 2018, where paired TCR information is denoted by matching index values and unpaired chains have index values of 0. Complete SC data (confidence value ≥ 1) from the Immune peptide Database (IEDB)

was added to our dataset and used as the default for all analysis. To test the sensitivity of the results to data composition, we also built datasets that had IEDB data with confidence value ≥ 0, ≥ 2, and ≥ 3, respectively. In total, our default SC dataset comprised 1447 TCRs, BS α 21,207 chains, and BS β 25,927 chains (for complete dataset counts see Additional file 1: Table 4). Data is available for download from the Github repository for the project.

CDR information

Our method for predicting the specificity of TCRs requires TCR gene family and complete CDR3 sequence. To obtain this, we retrived all human germline CDR loop information from the International ImMunoGeneTics Information System Gene database (IMGT/GENE-DB) [17]. CDR1 and CDR2 loops can be retrieved directly from the database. However, CDR2.5 needs to be extracted from the IMGT alignment sequence, and is defined by the residues in columns 81-86 of the gapped alignment (F+ORF+in-frame P amino acid sequences with IMGT gaps), as discussed in Dash et al. [2]. After translating the data to protein sequence, we produced non-redundant datasets by removing duplicate TCRs (sequence identity $< 100\%$).

Shuffled CDRs

Since we did not have access to true negative data (i.e., data showing which TCRs do not bind to particular epitopes), we randomly shuffled CDR regions within each α and β chain, to test whether or not a loss of precision would be observed. CDR regions for each TCR were shuffled prior to assigning the receptors to the train and test sets.

Baseline method

We implemented TCRdist [2] as a baseline method for classifying TCRs according to their peptide specificity. TCRdist is based on a nearest-neighbor approach, where the distance between TCRs is obtained from protein sequence alignment scores between TCRs. Using a BLOSUM62 matrix, a protein alignment is performed between any two TCRs using CDR loops 1, 2, 2.5, and 3. Subsequently, CDR 1, 2, and 2.5 are given a weight of 1, where CDR3 is given a weight of 3. Finally, the weighted sum of the CDR loop alignment scores is used as a proximity measure, and TCRs are assigned the peptide specificity of their nearest neighbor [2].

SwarmTCR

The main idea behind SwarmTCR is that the "importance" of the α and β chains as well as the CDR regions within these chains varies depending upon the peptide that is being recognized, as described in the literature [8]. In order to reflect this, SwarmTCR learns optimal weights for each of the eight CDR loops in a peptide-specific fashion. SwarmTCR explores the eight-dimensional (with SC data) or four-dimensional (with BS data) space of CDR weights with Particle Swarm Optimization (PSO), an established optimization technique inspired by the natural flocking behavior of birds that has been shown to achieve good performance in a wide range of optimization contexts [18].

The weights are used in a nearest-neighbor framework as done in TCRdist. We framed this as an optimization problem, where the objective is to identify a set of weights that

maximize classification performance as measured by Average Precision (AP) (eq. 1). AP was selected as the objective function to address the issue of unbalanced datasets (Additional file 1: Table 4), as suggested by [19]. We used Particle Swarm Optimization (PSO) for carrying out the optimization of the weights and maximize AP on a training set.

$$\mathrm{AP} = \sum_{k=1}^{n} P(k)\Delta r(k) \tag{1}$$

AP is determined by the sum over every position of the precision-recall curve where k is the rank of the retrieved TCRs, n is the number of TCRs, $P(k)$ is the precision at cut-off k, and $\Delta r(k)$ is the change in recall from $k-1$ to k [20].

In PSO particles are initially placed in a multidimensional space at random, with each particle representing a possible solution to the optimization problem. At each iteration, particles move with a velocity vector that is a function of both the local best of the particle and the global best. The velocity ($\mathbf{v}$) and position ($\mathbf{p}$) of a particle i are updated at each time step t according to Eq. 2 and 3:

$$\mathbf{v}_i^{t+1} = \omega * \mathbf{v}_i^t + c1 * r1 * (\mathbf{pbest}_i - \mathbf{p}_i^t) + c2 * r2 * (\mathbf{pbest}_g - \mathbf{p}_i^t) \tag{2}$$

$$\mathbf{p}_i^{t+1} = \mathbf{p}_i^t + \mathbf{v}_i^{t+1} \tag{3}$$

where ω is an inertia factor set to 0.5, $c1$ and $c2$ are scaling factors set to 0.5, $r1$ and $r2$ are two random numbers between 0 and 1, $\mathbf{pbest}_i$ is the position of particle i that has resulted in the best value for the objective function so far, while $\mathbf{pbest}_g$ is the global best (i.e., the position corresponding to the best value so far across all particles).

The optimization is set to terminate if the swarm moves $\leq 10^{-8}$ from its best position or if the change in the swarm's best objective value is $\leq 10^{-8}$. The swarm size is set to 25, with a maximum number of 20 iterations.

The SwarmTCR model

We define as "training set" the TCRs used to obtain an optimal set of weights maximizing average precision, and test set as the TCRs where the performance of the optimal weights is evaluated. Within both sets, we have a reference subset containing labeled TCRs and a sample subset that the nearest neighbor approach compares against the reference subset to infer peptide labels for the TCRs.

Training and test sets for SC and BS data were constructed differently due to data availability, with BS data being much more abundant than SC data. For BS, the training and test sets were filled using a 50/50 split, and for both training and test sets half of the TCRs specific for a particular peptide were placed into the reference subset and the remainder into the sample subset (Fig. 6A).

For SC, 30% of all TCRs specific for a peptide were placed into the reference subset for both training and testing sets. The same reference subset was used in training and testing due to limited amounts of SC data (see Additional file 1: Table 4). In order to create the sample subsets for training and testing, the remaining 70% (TCR specific) undergoes another 70/30 split (Fig. 6B). We note that the sample reference sets are distinct for

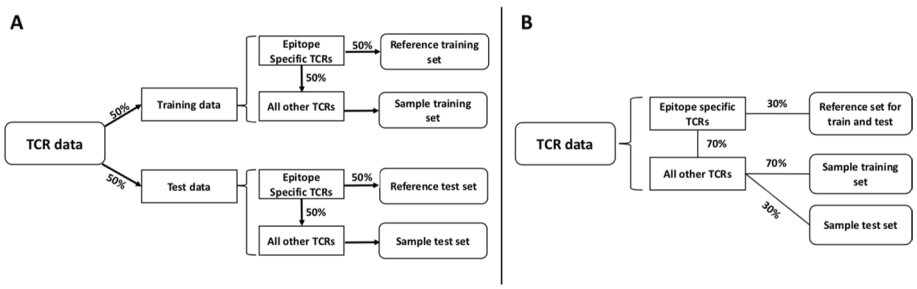


Fig. 6 Data allocation. The allocation of data for both BS (**A**) and SC (**B**) model is illustrated above. All TCRs are split proportionally into the sample and reference sets. As described in the literature, the same SC reference set is used in training and testing (due to abundance of data, this was not necessary in the BS analysis). The subset labeled peptide specific TCRs are the group of TCRs specific for the peptide tested by SwarmTCR

training and test. We also ensured that the different proportions of TCR peptide specificities were equally represented in training and test sets.

Once data was randomly allocated into the training and test sets as described above, we performed the PSO procedure on the training set. Each solution (optimal set of weights maximizing average precision) was then applied to the test set. Cross-validation was performed using repeated random sub-sampling for 50 iterations on both SC and BS datasets.

Crystal structure contacts and SwarmTCR output

We searched the Protein Data Bank (PDB) for TCR/pMHC crystal structure complexes containing one of the peptides in our TCR repertoires, to compare α and β chain usage and CDR loop usage with SwarmTCR weights.

We found nine complexes with the GILGFVFTL peptide (PDB IDs: 1OGA, 2VLJ, 2VLK, 2VLR, 5EUO, 5E6I, 5ISZ, 5JHD, 5TEZ), three complexes with the NVLPMVATV peptide (PDB IDs: 3GSN, 5D2L, 5D2N), and one complex with the GLCTLVAML peptide (PDB ID: 3O4L). Using the distance threshold discussed in a previous publication [8], we extracted CDR region residues within 4.5Å to the target (peptide, pMHC).

We then compared contact residue counts from the crystal structures to the SwarmTCR weights for each repertoire and the default TCRdist weight set.

Abbreviations

TCR: T cell receptor; pMHC: Peptide/major histocompatibility complex; PR: Precision-recall; CDR: Complementarity determining region; SC: Single cell; BS: Bulk sequencing; PBMC: Peripheral blood mononuclear cell.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04335-w.

Additional file 1. Supplementary Figures and Tables.

Acknowledgements

We thank Robin Brody for technical assistance and Dr. Matteo Ligorio and Sean West for reading the manuscript.

Authors' contributions
RE and DG conceived the project; RE and DG developed the software; LK and AG collected the experimental data. KL and LKS provided ideas and suggestions for the method. RE and DG wrote the manuscript. All authors read and approved the final version of the manuscript.

Funding
This work has been supported in part by a Nebraska Systems Science grant (DG) and by NIH grant AI49320 (KL and LK); UL1TR001453 (KL).

Availability of data and materials
Project name: SwarmTCR. Project home page: https://github.com/thecodingdoc/SwarmTCR, v1.0. Operating system(s): Platform independent (command-line software). Programming language: C++. License: GPL-3

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Interdisciplinary Informatics, College of Information Science and Technology, University of Nebraska at Omaha, 1110 S 67TH, Omaha, NE 68182, USA. [2]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA, USA. [3]Department of Pathology, University of Massachusetts Medical School, Worcester, MA, USA.

Received: 8 November 2020 Accepted: 16 August 2021
Published online: 07 September 2021

References

1. Swann JB, Smyth MJ. Immune surveillance of tumors. J Clin Investig. 2007;117(5):1137–46. https://doi.org/10.1172/JCI31405.
2. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, Thomas PG. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature. 2017;547(7661):89–93. https://doi.org/10.1038/nature22383.NIHMS150003.
3. Antunes DA, Rigo MM, Freitas MV, Mendes MFA, Sinigaglia M, Lizée G, Kavraki LE, Selin LK, Cornberg M, Vieira GF. Interpreting T-cell cross-reactivity through structure: implications for TCR-based cancer immunotherapy. Front Immunol. 2017;8:1–16. https://doi.org/10.3389/fimmu.2017.01210.
4. Zarnitsyna VI, Evavold BD, Schoettle LN, Blattman JN, Antia R. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. Front Immunol. 2013;4(485):1–11. https://doi.org/10.3389/fimmu.2013.00485.
5. De Simone M, Rossetti G, Pagani M. Single cell T cell receptor sequencing: techniques and future challenges. Front Immunol. 2018;9:1638. https://doi.org/10.3389/fimmu.2018.01638.
6. Schönbach C, Ranganathan S, Brusic V. Immunoinformatics. Berlin: Springer; 2008.
7. Heringa J, Argos P. Side-chain clusters in protein structures and their role in protein folding. J Mol Biol. 1991;220(1):151–71. https://doi.org/10.1016/0022-2836(91)90388-M.
8. Ehrlich R, Ghersi D. Analyzing T cell receptor alpha/beta usage in binding to the pMHC. In: Proceedings - 2017 IEEE international conference on bioinformatics and biomedicine, BIBM 2017 (2017). https://doi.org/10.1109/BIBM.2017.8217629.
9. Ishizuka J, Stewart-Jones GBE, van der Merwe A, Bell JI, McMichael AJ, Jones EY. The structural dynamics and energetics of an immunodominant T cell receptor are programmed by its Vβ domain. Immunity. 2008;28(2):171–82. https://doi.org/10.1016/j.immuni.2007.12.018.
10. Ladell K, Hashimoto M, Iglesias MC, Wilmann PG, McLaren JE, Gras S, Chikata T, Kuse N, Fastenackels S, Gostick E, Bridgeman JS, Venturi V, Arkoub ZA, Agut H, van Bockel DJ, Almeida JR, Douek DC, Meyer L, Venet A, Takiguchi M, Rossjohn J, Price DA, Appay V. A molecular basis for the control of preimmune escape variants by HIV-specific CD8+ T cells. Immunity. 2013;38(3):425–36. https://doi.org/10.1016/j.immuni.2012.11.021.
11. Kamga L, Gil A, Song I, Brody R, Ghersi D, Aslan N, Stern LJ, Selin LK, Luzuriaga K. CDR3α drives selection of the immunodominant Epstein Barr virus (EBV) BRLF1-specific CD8 T cell receptor repertoire in primary infection. PLoS Pathog. 2019;15(11):1–24. https://doi.org/10.1371/journal.ppat.1008122.
12. Kim T-S, Shin E-C. The activation of bystander CD8(+) T cells and their roles in viral infection. Exp Mol Med. 2019;51(12):1–9. https://doi.org/10.1038/s12276-019-0316-1.
13. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, Arlehamn CSL, Sette A, Boyd SD, Thomas J. Identifying specificity groups in the T cell receptor repertoire. Nature. 2018;547(7661):94–8. https://doi.org/10.1038/nature22976.Identifying.
14. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, Komech EA, Sycheva AL, Koneva AE, Egorov ES, Eliseev AV, Van Dyk E, Dash P, Attaf M, Rius C, Ladell K, McLaren JE, Matthews KK, Clemens EB, Douek DC, Luciani F, Van Baarle D, Kedzierska K, Kesmir C, Thomas PG, Price DA, Sewell AK, Chudakov DM. VDJdb: a curated database of

T-cell receptor sequences with known antigen specificity. Nucleic Acids Res. 2018;46(D1):419–27. https://doi.org/10.1093/nar/gkx760.
15. Ponomarenko J, Papangelopoulos N, Zajonc DM, Peters B, Sette A, Bourne PE. IEDB-3D: structural data within the immune epitope database. Nucleic Acids Res. 2011;39:1164–70. https://doi.org/10.1093/nar/gkq888.
16. Gil A, Kamga L, Chirravuri-Venkata R, Aslan N, Clark F, Ghersi D, Luzuriaga K, Selin LK. Epstein-Barr virus epitope-major histocompatibility complex interaction combined with convergent recombination drives selection of diverse T cell receptor α and β repertoires. mBio. 2020;11(2):00250–20. https://doi.org/10.1128/mBio.00250-20.
17. Chaume D, Lefranc M-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 2005;33:256–61. https://doi.org/10.1093/nar/gki010.
18. Kennedy J. Particle swarm optimization: tutorial. encyclopedia of machine learning. 2010. https://doi.org/10.1109/ICNN.1995.488968.
19. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):1–21. https://doi.org/10.1371/journal.pone.0118432.
20. Su W, Yuan Y, Zhu M. A relationship between the average precision and the area under the ROC curve. ICTIR 2015—proceedings of the 2015 ACM SIGIR international conference on the theory of information retrieval (2015). https://doi.org/10.1145/2808194.2809481.

Publisher's Note