Classification of T cell movement tracks allows for prediction of cell function

Reka K. Kelemen, Gengen F. He,
Hannah L. Woo, Thomas Lane,
Caroline Rempe and Jun Wang

UT-ORNL Graduate School of Genome Science and Technology,
Center for Environmental Biotechnology, and
Departments of Civil and Environmental Engineering,
Geography, and Microbiology,
University of Tennessee,
F337 Walters Life Science,
Knoxville, TN 37996-0840, USA
E-mail: rkelemen@utk.edu
E-mail: ghe@utk.edu
E-mail: hwoo@utk.edu
E-mail: tlane13@utk.edu
E-mail: crempe@utk.edu
E-mail: jwang43@utk.edu

Ian A. Cockburn

Department of Pathogens and Immunity,
John Curtin School of Medical Research,
Australian National University,
GPO Box 334, Canberra City ACT 2600, Australia
E-mail: ian.cockburn@anu.edu.au

Rogerio Amino

Unité de Biologie et Génétique du Paludisme,
Institut Pasteur,
75724, Paris Cedex 15, France
E-mail: rogerio.amino@pasteur.fr

Vitaly V. Ganusov

Department of Microbiology,
University of Tennessee,
M409 Walters Life Science Building,
1414 West Cumberland Avenue,
Knoxville, TN 37996-0845, USA
E-mail: vitaly.ganusov@gmail.com

Michael W. Berry*

EECS Department,
University of Tennessee,
401 Min H. Kao Building,
Knoxville, TN 37996, USA
E-mail: berry@eecs.utk.edu
*Corresponding author

Abstract: Using a unique combination of visual, statistical, and data mining methods, we tested the hypothesis that an immune cell's movement pattern can convey key information about the cell's function, antigen specificity, and environment. We applied clustering, statistical tests, and a support vector machine (SVM) to assess our ability to classify different datasets of imaged flouresently labelled T cells in mouse liver. We additionally saw clusters of different movement patterns of T cells of identical antigenic specificity. We found that the movement patterns of T cells specific and non-specific for malaria parasites are differentiable with 72% accuracy, and that specific cells have a higher tendency to move towards the parasite than non-specific cells. Movements of antigen-specific T cells in uninfected mice vs. infected mice were differentiable with 69.8% accuracy. Wxe additionally saw clusters of different movement patterns of T cells of identical antigenic specificity. We concluded that our combination of methods has the potential to advance the understanding of cell movements in vivo.

Keywords: computational biology; data mining; support vector machine; T cells; cell movement tracks; pattern classification.

Reference to this paper should be made as follows: Kelemen, R.K., He, G.F., Woo, H.L. Lane, T., Rempe, C., Wang, J., Cockburn, I.A., Amino, R., Ganusov, V.V. and Berry, M.W. (2014) 'Classification of T cell movement for prediction of cell function', *Int. J. Computational Biology and Drug Design*, Vol. 7, Nos. 2/3, pp.113–129.

Biographical notes: Reka K. Kelemen is a graduate student pursuing a MS in Life Science through the Graduate School of Genome Science and Technology at the University of Tennessee, Knoxville.

Gengen F. He is a graduate student pursuing a PhD through the Department of Geography at the University of Tennessee, Knoxville.

Hannah L. Woo is a graduate student pursuing a PhD through the Department of Civil and Environmental Engineering at the University of Tennessee, Knoxville.

Thomas Lane is a graduate student pursuing a MS in Life Science through the Graduate School of Genome Science and Technology at the University of Tennessee, Knoxville.

Caroline Rempe is a graduate student pursuing a PhD in Life Science through the Graduate School of Genome Science and Technology at the University of Tennessee, Knoxville.

Jun Wang is a graduate student pursuing a PhD through the Department of Microbiology at the University of Tennessee, Knoxville.

Ian A. Cockburn received his PhD in Biology from the University of Edinburgh in 2004. He is currently an Associate Professor in the John Curtin School of Medical Research at the Australian National University. His research interests include immunity to malaria, T cell migration and function, and B cell differentiation and development.

Rogerio Amino received his PhD in Microbiology and Immunology from the Universidade Federal de Sao Paulo, Brazil. After a postdoctoral fellowship in the first Grand Programme Horizontal at the Institut Pasteur, Paris, France, he moved to the Department of Biochemistry at the Federal University of Sao Paulo, as an Assistant Professor. In 2008, he returned to the Institut Pasteur, where his group is currently using functional imaging to study the determinants implicated in Plasmodium pre-erythrocytic stages survival in the skin and liver of na ive and immunised hosts.

Vitaly V. Ganusov received his PhD in Biology in 2003 from Emory University. He holds a joint position between the Department of Microbiology and Department of Mathematics at the University of Tennessee in Knoxville. His research interests include mathematical modelling of cellular immune responses to viral infections, lymphocyte recirculation kinetics, and within-host pathogen evolution.

Michael W. Berry recieved his PhD in Computer Science in 1990 from the University of Illinois at Urbana-Champaign. He is a Full Professor in the Departments of Electrical Engineering and Computer Science and Mathematics, and is Director of the Center for Intelligent Systems and Machine Learning (CISML) at the University of Tennessee, Knoxville. His research interests include bioinformatics, data/text mining, parallel and scientific computing, and visual analytics.

This paper is a revised and expanded version of a paper entitled 'Classification of T cell movement tracks allows for prediction of cell function' presented at the *International Conference on Intelligent Biology and Medicine*, 13 August, 2013, Nashville, TN, USA.

1 Introduction

CD8 T cells are part of the immune system and are responsible for seeking and killing damaged, cancerous or infected cells in mammals. In infectious diseases such as malaria, which takes about one million lives annually and affects 300 times more, CD8 T cells have been proven to play crucial roles in the host's ability to clear the parasite during an adaptive immune response (Weiss et al., 1988). It has been long known that CD8 T cells are essential for immunity against malaria, and recently (Schmidt et al., 2008) have shown that reaching a certain large, but definable threshold of memory CD8 T cells specific for malaria antigen is sufficient for clearing the infection in 98% of mice exposed to malaria (Schmidt et al., 2011, 2008).

Despite their importance in the immune response against many infections the mechanisms by which parasite-specific CD8 T cells survey the site of infection and factors which influence T cell behaviour and movement patterns, remain poorly understood. Current hypotheses about CD8 T cell movement are based on in vitro or in vivo immunological experiments, such as the detection of increasing numbers of T cells as the concentration of chemoattractant molecules increases in a certain area. The results of such experiments suggest that T cells are recruited to the site of infection by a concentration gradient of chemokines, which are expressed by the infected cell and other immune cells. Challenges remain in experimentally detecting the radius of such a chemokine concentration gradient and the precise effect that chemokines have on T cell movement in vivo.

With the latest advancements in visualisation techniques, high resolution videos produced by spinning disk confocal and two-photon microscopy give biologists and immunologists a whole new perspective on T cells and their movement in the body (Germain et al., 2012). Numerous studies have employed in vivo or in vitro visualisation techniques using fluorescently labelled immune cells, and analysed their behaviour and movement trajectories to either visually or statistically test for phenomena such as Lévy flights (Harris et al., 2012).

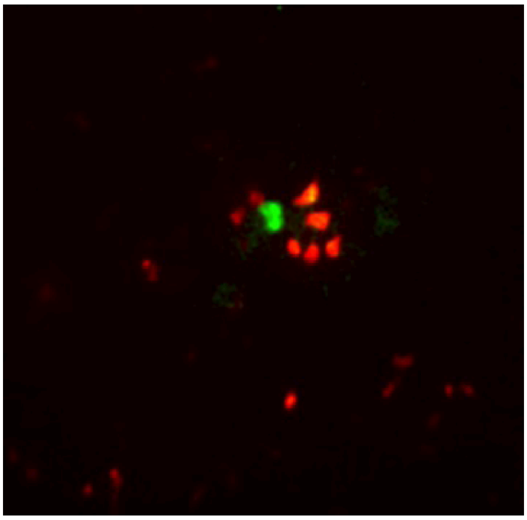
While qualitative analyses are important tools in gaining understanding of the immune response to infections, quantitative approaches to the same questions can give us more unbiased results and in many cases highlight underlying patterns otherwise undetectable. Here, we introduce the use of data mining and spatial and hierarchical clustering in analysing T cell movement coordinate data obtained from in vivo imaging of mouse livers. We test several hypotheses regarding T cell movement such as that T cell movement pattern depends on the T cell's specificity for the infection, the T cell's distance from the infected cell, or the presence of an infection in the body. We also address the question of whether the tissue topology is the key determining factor in T cell movement, and we try to reconstruct the tissue landscape at each imaged site, based on the movement coordinate data.

We find that CD8 T cells do exhibit different movement patterns based on their specificity for the infecting parasite and that our support vector machine (SVM) trained on two types of T cell movement data can distinguish between the two cell types with 72% accuracy. Hierarchical clustering reveals several subtypes of T cell movement even in a homogeneous set of CD8 T cells specific for malaria antigen. We also observe zones of highly travelled areas in videos of the liver, which might be important information for overlaying T cell movement and tissue topology. Overall, we demonstrate several approaches and quantitative, computational tools that are applicable not only to the movement data of cells, but of animals, or any other moving object.

2 Implementation

We have analysed the T cell movement coordinate datasets of several mouse experiments. Experiments were done by Ian Cockburn and Rogerio Amino at Johns Hopkins University (Baltimore, USA) and Pasteur Institute (Paris, France), respectively. These experiments involved in vivo imaging of fluorescently labelled *Plasmodium yoelii* parasites, the causative agents of malaria, and effector CD8 T cells of various epitope specificities (Figure 1). All the imaging was done by confocal spinning disc microscopy in anaesthetised mice over a time period of 30–60 min. Two- or three- dimensional position coordinates of labelled CD8 T lymphocytes and malaria-infected cells were acquired at time steps of 30 s to 4 min, as described by Cockburn et al. (2013). We focused on three datasets that we will refer to as Datasets 1, 2, and 3, as defined below.

Figure 1 Image of malaria-specific PyTCR CD8 T cells and malaria-infected hepatocyte. CD8 T cells are labelled with red fluorescent membrane dye (PKH-26); these cells congregate around a liver cell identified by green fluorescent protein-expressing malaria parasites that are located inside of the cell (see online version for colours)



2.1 Dataset 1: PyTCR and OT-1 T cells in infected mice

The first dataset we analysed consisted of movement coordinates of two different types of effector CD8 T cells in malaria infected mice: one specific for the *Plasmodium yoelii* (Py) epitope SYVPSAEQI, and the other specific for the ovalbumin epitope SIINFEKL, which is unrelated to malaria. We will refer to these two cell types as PyTCR (*Plasmodium yoelii* T cell receptor transgenic) and OT-1 (ovalbumin T cell receptor transgenic). The position coordinates of these two distinctively labelled cell types around a single infected liver cell were acquired every 2 min for a total of 38 min. This experimental setup is suitable for testing whether CD8 T cells specific and CD8 T cells not specific for malaria exhibit different movement patterns when injected together into malaria-infected mice. This dataset contained a total of 28 cells, 11 OT-1 and 17 PyTCR.

2.2 Dataset 2: PyTCR T cells in infected mice

The second dataset contained the movement coordinates of PyTCR cells in livers of malaria-infected mice which were imaged at time intervals of 4 min for a total of 50 min. This dataset is useful for clustering a seemingly homogeneous set of T cells based on their movement attributes. This dataset contained a total of 148 cells.

2.3 Dataset 3: parasite specific T cells in infected and uninfected mice

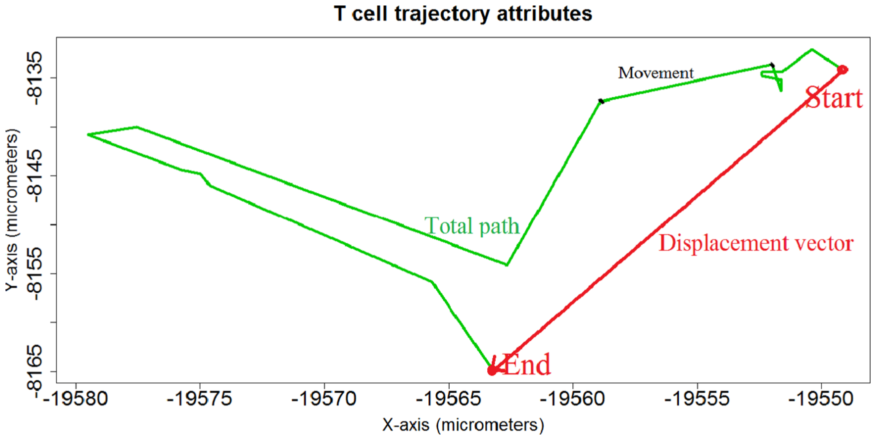
The third dataset was a result of mouse experiments where OT-1 cells were injected into mice, then the mouse was immunised with irradiated malaria parasites that expressed a mutated protein containing the chicken ovalbumin epitope. Six days later the mouse either was challenged with ovalbumin-expressing malaria parasites or remained unchallenged right before the imaging of the liver. Imaging time points were acquired every 30 s for

30 min, therefore this dataset has the largest resolution of T cell movement. This dataset is useful for classifying T cell movement patterns based on whether there is or is not an infection in the body. This dataset contained a total of 406 T cells, 94 in uninfected mice and 312 in infected mice.

2.4 Movement attributes

To characterise T cell movement for classification as discussed in Sections 2.5 and 2.6 (SVM classifier and hierarchical clustering) we calculated several attributes for each cell in Dataset 1 (PyTCR and OT-1 cells) based on the two- or three-dimensional coordinate points. The attributes are total path, maximum, minimum, and average movement, maximum, minimum, average and sum of all angles towards the parasite, total displacement length, total displacement vector, and the *meandering index*, which is the ratio of the total path length to the displacement (Figure 2).

Figure 2 Movement attributes of T cells. The distance between two consecutive positions of a T cell was considered a movement, and the sum of all movements for a T cell was its total path length. The distance and the vector between the start and end points of a T cell are called its displacement and displacement vector (see online version for colours)



To include angles related to T cell movement in our analysis, we used the following calculations for each T cell. The procedures to calculate turning angles of moving T cells are as follows.

Find the average coordinate of the parasite.

$$P =< P_x, P_y, P_z > \tag{1}$$

Calculate parasite-T cell distance vector, PT_n, at each time point n.

$$PT_n =< T_{x,n} - P_x, T_{y,n} - P_y, T_{z,n} - P_z > \tag{2}$$

Calculate the T cell's movement vector, TT_n, between each time point n.

$$TT_n =< T_{x,n} - T_{x,n+1}, T_{y,n} - T_{y,n+1}, T_{z,n} - T_{z,n+1} > \tag{3}$$

Calculate the angle between the parasite-T cell distance vector and T cell's movement vector at time point n.

$$\theta_{PT-TT,n} = \arccos(PT_n \cdot TT_n / ||PT_n||||TT_n||) \tag{4}$$

where $||x||$ denotes Euclidean distance. Calculate the angle between the T cell's movement vector at time point n and its movement vector at the subsequent time point, $n+1$, as

$$\theta_{TT-TT,n} = \arccos(TT_n \cdot TT_{n+1} / ||TT_n||||TT_{n+1}||) \tag{5}$$

so that the scalar distance from parasite can be represented by

$$PT_n = (T_{x,n} - P_x)^2 + (T_{y,n} - P_y)^2 + (T_{z,n} - P_z)^2. \tag{6}$$

2.5 Spatial analysis

To reconstruct a sense of the tissue topology based on the frequency of T cell presence at various sites in the imaged area of the liver, we used a clustering algorithm to create *zones* based on T cell positions with proximity to each other for Datasets 1 (PyTCR and OT-1 cells) and 2 (PyTCR cells). An implementation of the DBSCAN algorithm in Python was utilised for clustering T cell locations across all time points in each video dataset (Ester et al., 1996). No differentiation was made between T cell locations based on whether the cell belonged to a class of parasite specificity, and the T cell locations were not grouped by the cell type that they belonged to. The original testing and visualisation of various clustering algorithms was performed in the ELKI platform (Achtert et al., 2012). The ELKI platform is a powerful tool for assessing clustering methods most likely to yield biologically relevant results, particularly through the visualisation and availability of a multitude of clustering techniques.

The parameters used for all DBSCAN clustering experiments were an epsilon of 10 and a minimum points of 2. The parameter epsilon defines the maximum radial distance from a point within which other points can be considered for the initial point's cluster. The epsilon is a value regarding the maximum distance to look from a point for nearby candidate points to join a cluster. Minimum points is a parameter that specifies how many points are needed to meet the requirements of clustering with one another in order to form a viable cluster. These parameters were selected based on visual assessment of T cell clusters in order to increase their biological relevance in the liver tissue.

2.6 Classifier: support vector machine

We implemented the Python interface of LIBSVM (Chang and Lin, 2011) with Python 2.7.1, making use of the LIBSVM cross-validation function in model training. The above described distance, angle, and zone attributes were used as input into the SVM to train a linear or radial basis function (RBF) model using 3 to 10-fold cross-validation and the no shrinking heuristics option. This was applied to Datasets 1 (PyTCR and OT-1 cells) and 3 (infected and uninfected).

2.7 Hierarchical clustering

After calculating movement attributes including average movement length, total path length, displacement, and angle between movements for each cell in our collection of videos, we provided this information to a hierarchical clustering algorithm for Datasets 2 (PyTCR cells) and 3 (malaria-infected and uninfected mice). Hierarchical clustering and heatmap generation was done using JMP Genomics 6.0 (SAS Institute Inc. 2012) with WardâŁ™s minimum variance method. All attributes were used for clustering. We additionally used the R-based software packages (*ggplot* and *plyr*) to visualise a series of T cell trajectories on their own, or in relation to the parasite for Dataset 2 (PyTCR).

3 Results and discussion

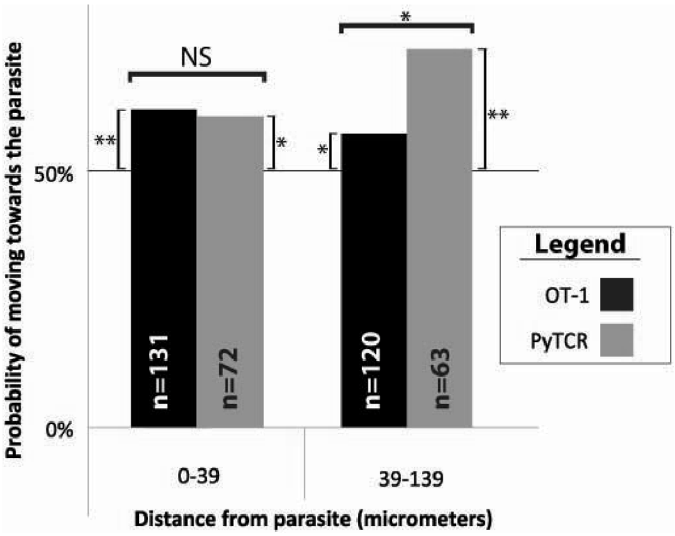
3.1 T cell movement with respect to distance from parasite

Most T cells within Dataset 1 (PyTCR and OT-1 cells) tend to move towards the parasite regardless of cell type. The direction of a trajectory was considered as moving towards the parasite if the angle given by equation 4 was acute. Eighty-two percent of both OT-1 and PyTCR T cells were found to be travelling towards the parasite for more than half the duration of their monitored movement. The distribution of angles between the two cell types were not significantly different based on nonparametric Kruskall-Wallis rank sums test at an alpha level of 0.05. Interestingly, irrespective of the distance, both parasite-specific and non-specific cells had increased preferance to move to the infected cell (more sharp turning angles than expected by chance, $P < 0.025$, binomial test). This finding suggests that OT-1 and PyTCR behaviour are similar in that they will all migrate towards the infection site. However, there may be a bias in our data since the parasite was located in the centre of the frame, so cells further from the parasite are not included in the small ($200 \times 200 \times 50$ μm) field of view. Additionally, this dataset only contained cells with at least six timepoints, which may have excluded cells moving quickly in and out of the frame.

A binomial test is applicable if we consider the orientation data as a Bernoulli trial where ‘moving towards the parasite’ is the success. This statistical method gives the likelihood of observing the probabilities reported in Figure 3 when the expected probability is 50%. We found the likelihood of observing these probabilities is less than 2.5% for both OT-1 and PyTCR cells and both distance ranges greater and less than 40 microns. In other words, it is unlikely we are observing higher probabilities of moving towards the parasite by all cells by random chance. Therefore, the statistics support the possibility of a general bias in T cell behaviour to move towards the parasite.

We found that (parasite-specific) PyTCR cells move towards the parasite 15% more often than non-specific OT-1 cells at distances greater than 40 μm from the parasite (Figure 3, $P = 0.027$, Kruskall-Wallis rank sums). However, no difference in migration towards the infected cell was noted when T cells are within 40 μm of the parasite; the latter length approximates the maximum size of an infected hepatocyte in mice (Figure 3, $P = 0.579$, Student’s t-test). It is unclear why parasite-specific cells are better able to recognise the infection site than non-specific cells. One potential explanation is that malaria parasites (sporozoites) leave an ’antigenic’ trail as the sporozoite penetrates the tissues from the blood. CD8 T cells may be able to recognise such a path using their TCR which skews the moving patterns towards the infection site. In conclusion, our results provide the basis of classifying T cells using a combination of their distance and orientation to the parasite as attributes.

Figure 3 Both antigen-specific (PyTCR) and non-specific (OT-1) effector CD8 T cells are attracted to the infected cell. We calculate the distribution of turning angles for cells of different specificities and calculate the fraction of turns towards the parasite. Direction trends of PyTCR (malaria-specific) and OT-1 (non-specific for malaria) cells (Dataset 1) are based on distance from infection site. The direction of T cell movement at each time point was grouped by distance from parasite: up to 39 μm (near a parasite-infected cell, see Figure 1) or farther than 40 μm (away from a parasite-infected cell). The probability of moving towards the parasite is calculated by dividing the number of angles less than 90° relative to the direction of the infection site by total number of data points at that distance interval. The number of data points analysed at each bin is indicated by n. Distance bins with less than 3 data points are not included

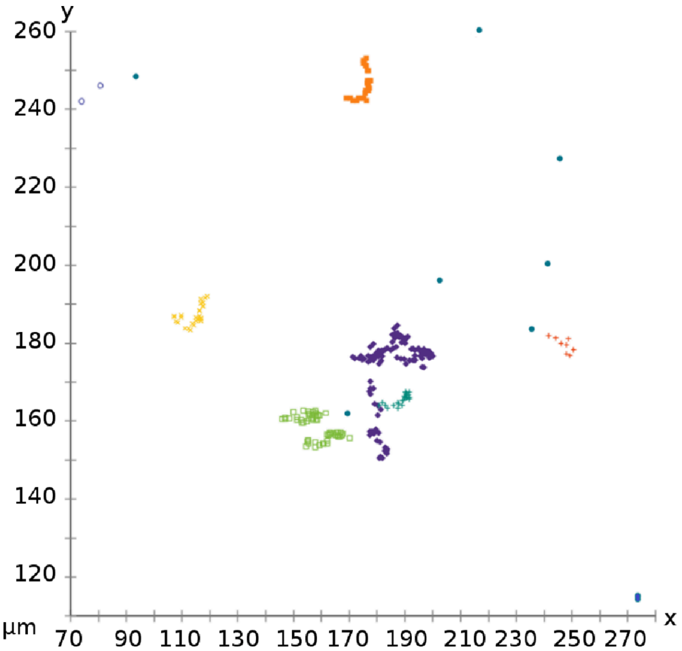


NS = not significant. $*P < 0.05$ and $**P < 0.002$ (unpaired two-tailed t-test, Kruskall-Wallis rank sums, binomial test).

3.2 *Spatial analysis*

Our spatial clustering analysis showed areas of highly travelled sites in the imaged tissue of Datasets 1 (PyTCR and OT-1 cells) and 2 (PyTCR cells), while most of the surveyed area was untouched by T cells. Interestingly, some features of the underlying tissue topology, like *highways* of immune cell traffic can be extrapolated from these figures of points travelled by T cells (Figure 4). Further analysis of high density T cell videos could lead to more highly resolved pictures of the tissue that could potentially predict complete highways of immune cell movement and the chemoattractant concentration gradient.

Figure 4 Clustering reveals highly travelled areas in the imaged area of mouse liver tissue. Colours represent density based clusters of T cell positions at all timepoints from a parasite-infected mouse liver (Dataset 2). Blue circles are cell positions that did not cluster. The colours correspond to categories of the *zone* attribute used in the SVM, which highlight potential areas of intense traffic denoted by many dots and a single colour that may include one or more individual cells (see online version for colours)



3.3 Classifier: support vector machine

We used a SVM to assess whether attributes of T cell movement can be used to differentiate PyTCR from OT-1 T cells in parasite-infected mice (Dataset 1) or parasite-specific T cells in infected or noninfected mice (Dataset 3). A SVM functions as a classifier by mapping the vector data to a higher dimensional space using a kernel function and then by finding a hyperplane that optimally separates the data by maximising the margin between the plane and each class (Solem, 2012). Support vector machines are a common means of supervised (binary) classification that have recently been used in a study by Goodson et al. in 2011 to analyse movement patterns of mouse sperm (Goodson et al., 2011). To our knowledge, we are the first to apply such a method to the classification of T cell movement in live tissues.

We started with a baseline assessment of classification by feeding two basic attributes into the SVM: total path length traversed by a cell and the distance travelled for the first movement. Using a linear kernel and 3-fold cross validation (using the *–v* cross validation option of LIBSVM) on our dataset of 28 cells (17 malaria specific and 11 malaria non-specific) our SVM model gave 64% accuracy. Next, we used the same SVM parameters but added several more distance attributes, then several angle based attributes. All distance attributes (path length, average movement, mininum movement, maximum movement, net x displacement vector, net y displacement vector, net z displacement vector, total displacement) gave an accuracy of 53%, total path length and all angle attributes (mean angle, minimum angle, maximum angle, sum of angles, mean angle to parasite, minimum angle to parasite, maximum angle to parasite, and sum of angles to parasite) gave an accuracy of 64%, and total path length and zoning information gave an accuracy of 57%.

Since the dataset is from a single imaging space, cells move in and out of the frame over the course of the experiment. In an attempt to normalise cells that remain in frame with cells that move in and out of frame, we divided cell data into overlapping *chunks* of six timepoints, the minimum number of time points required to keep a cell in the dataset. We generated the same distance and angle attributes from these chunks of cell time and obtained a 66.9% accuracy with 3-fold cross validation. However, since the total number of data points increased when separated into these chunks, we discovered that this level of cross validation was underfitting the model. At 10-fold cross validation with all distance and angle attributes for time chunks, we observed a 70.6% accuracy. Finally, when we added zoning information into this model, the 10-fold cross validation accuracy increased to 72%. Nevertheless, when we investigated single instances of training on 9/10ths of the data and testing on 1/10th of the data, accuracies varied from 72% to 96% (see Table 1). This could be due to different data distributions in the training and testing folds. Since this dataset was quite small, we did not try subsampling within the data.

Table 1 Summary of support vector machine results: OT-1 and PyTCR data (Dataset 1). The overall accuracy using 10-fold RBF and linear cross-validation was 67.7% and 72%, respectively. The number of correctly or incorrectly classified testing *chunks* of cell timepoints are given in the true and false categories for each cell type; the SVM appears to correctly classify at least 72% of the testing dataset (T = true, F = false)

Model	*Accuracy* (%)	*T-OT1*	*F-OT1*	*T-PyTCR*	*F-PyTCR*
RBF	64	16	9	0	0
RBF	72	18	7	0	0
Linear	80	14	3	6	2
Linear	96	20	1	4	0
Linear	80	16	4	4	1
Linear	84	16	4	5	0
Linear	72	15	6	3	1
Linear	84	20	4	1	0
Linear	92	18	2	5	0
Linear	80	15	5	5	0
Linear	88	17	2	5	1
Linear	76	14	4	5	2

For the dataset of cells in malaria-infected mice vs. uninfected mice, we generated the same set of distance and angle attributes, excluding those attributes using information about parasite location (mean angle to parasite, minimum angle to parasite, maximum angle to parasite, and sum of angles to parasite), on overlapping chunks of six time points for these cells. Cells with fewer than 6 timepoints were excluded. These attributes were used to train a RBF SVM model, with which we observed an accuracy of 87.9%. Looking closer at our data, however, we discovered that the percent of data points labelled ‘infected’ was about 87%. To avoid biasing the classifier to the larger quantity of infected cells, we tried using the built-in weight function of LIBSVM, but this also gave an accuracy of 87%. As a further check, we pseudo-randomly sampled the infected cells for a subset equal to the number of uninfected cells in the dataset. This subset gave a cross-validation accuracy of

69.8% under the same conditions. Since the LIBSVM cross validation option is intended for training, we next made our own split of the chunk data with equal numbers of infected and uninfected cells into 10-folds. We trained with 9/10ths of this data and tested with 1/10th of this data in single instances, and were surprised by the high accuracies this gave (Table 2). The large discrepancy between the 69.8% of LIBSVM's cross validation option and the accuracies around 99% that we observed with single instance train and test sets may be due to an unrepresentative random sampling of infected cell datapoints. Although the number of infected and uninfected cells is the same with this method, the number of timepoints for each cell varies. Our dataset of 6 timepoint chunks has about twice as many datapoints for the infected class as the uninfected class. Nevertheless, our confusion matrix and high accuracy show a good separation of the data with an RBF kernel function (see Table 2).

Table 2 Summary of support vector machine results: infected (*inf*) vs. uninfected (*uni*) (Dataset 3). The overall 10-fold RBF cross-validation accuracy was 69.8%. The number of correctly or incorrectly classified testing *chunks* of cell timepoints are given in the true and false categories for each cell type; the SVM appears to correctly classify at least 97% of the testing dataset in these iterations (T = true, F = false)

Model	*Accuracy* (%)	*T-uni*	*F-uni*	*T-inf*	*F-inf*
RBF	99.2	132	2	238	1
RBF	100	99	0	282	0
RBF	97.9	112	5	225	2
RBF	99.7	114	0	256	1
RBF	99.7	115	0	281	1
RBF	100	96	0	259	0
RBF	99.2	114	0	271	3
RBF	99.5	121	1	61	1
RBF	99.7	134	1	221	0
RBF	99.4	131	2	220	0

Our application of machine learning to the datasets where there are two distinct cell types (malaria-specific vs. nonspecific) or two different experimental conditions (infection or no infection) present yielded a confirmation of those differences in the movement patterns of cells. Despite the small size of the first dataset our support vector machine predicted the right cell type in at least 72% of the cases (Table 3). This result is encouraging for further training on parasite-specific and non-specific cell movement datasets to better outline and understand what differences it makes for a cell to have a receptor for the parasite that is causing an on-going infection. A related classification was based on two experimental conditions, namely having a malaria infection in the liver or not. The SVM trained on the malaria-specific T cell movement coordinate data in these two conditions was able to predict which cell movement takes place in an infected vs. an uninfected mouse with a 70% accuracy. This result indicates that there are differences between the movement attributes of malaria-specific cells in infection and infection-free environments. Since we did run into variation and discrepancies between LIBSVM cross validation accuracy and single instance train and test accuracies, it would be beneficial to include boosting in future analyses of this dataset to reduce bias and hopefully obtain more consistent accuracies.

Table 3 Summary of support vector machine results for Dataset 1. Accuracies vary depending on input attributes, folds used for cross validation, and whether whole cells or overlapping *chunks* of 6 timepoints were used. Cross-validation folds are given, meaning the dataset was divided into 3 or 10 equal parts (folds), trained on *n-1* folds, and tested on 1 fold. Accuracy denotes the percent of the testing dataset that was correctly classified by the support vector machine. All distance attributes include path length, average movement, mininum movement, maximum movement, net x displacement vector, net y displacement vector, net z displacement vector, and total displacement. All angle attributes include mean angle, minimum angle, maximum angle, sum of angles, mean angle to parasite, minimum angle to parasite, maximum angle to parasite, and sum of angles to parasite

Cross val.	*Data type*	*Attributes*	*Accuracy* (%)
3	Whole cell	Path length, 1^{st} distance	64
3	Whole cell	All distance attributes	53
3	Whole cell	Path length, all angle attributes	64
3	Whole cell	Path length, zones	57
3	6 timepoints	All distance, angle attributes	66.9
10	6 timepoints	All distance, angle attributes	70.6
10	6 timepoints	All distance, angle attributes,zones	72

Both of these classifiers could be more accurate, or at least more consistent, with more data for training and testing. Though we did not assess power, we expect that it is too low to be confident that these movement patterns are consistently predictive of the assessed classes.

More experimental data is in the process of being produced. This will increase the power of T cell classification and would enable an independent, blind validation using a dataset from an experiment not included in our training data.

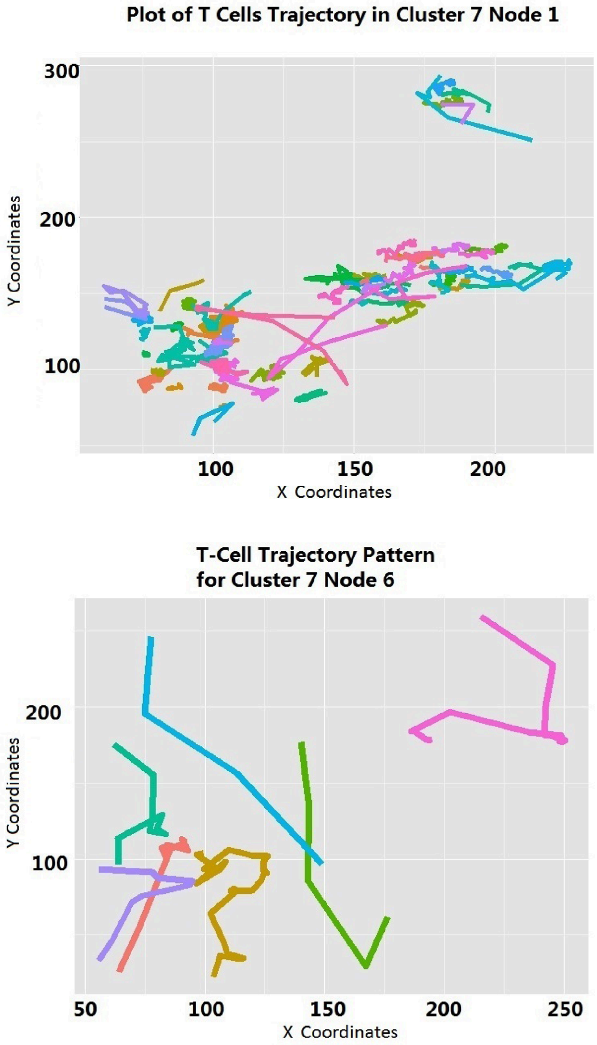
Previous studies of the role of receptor-ligand interactions in T cell migration have shown that the T cells' antigen-specific receptor (TCR) is involved in regulating the motility of the cell based on the presence or absence of the specific parasite antigen on nearby endothelial cells (Ward and Marelli-Berg, 2009). The biological explanation of non-specific cells behaving as differently from infection-specific T cells as if there was no infection in the surveyed tissue is consistent with our findings.

3.4 Hierarchical clustering

Hierarchical clustering of the data containing the movement coordinates of malaria-specific PyTCR T cells (Dataset 2) resulted in seven very distinct clusters and numerous sub-clusters of cells based on their movement. Plotting the tracks of cells in different clusters showed that some cells tend to make small movements while others tend to move in very long movements over the same time interval (Figure 5).

Other clusters have cells with a mixture of long or short movements, but interestingly they have very small overall displacement. This means that in many cases the cells move a lot, but due to sharp angles between movements they change their direction of movement drastically and end up near their starting point. This suggests that there are different movement types and perhaps search strategies even within the same set of cells, which have the same antigenic specificity.

Figure 5 Each of the images above contains malaria-specific CD8 T cell paths that belong to different clusters based on the hierarchical clustering of different movement patterns (Section 3.4). We see that T cells tend to move either in small movements and remain localised (top panel) or in large movements and traverse large distances (bottom panel). Colours represent different cells (see online version for colours)

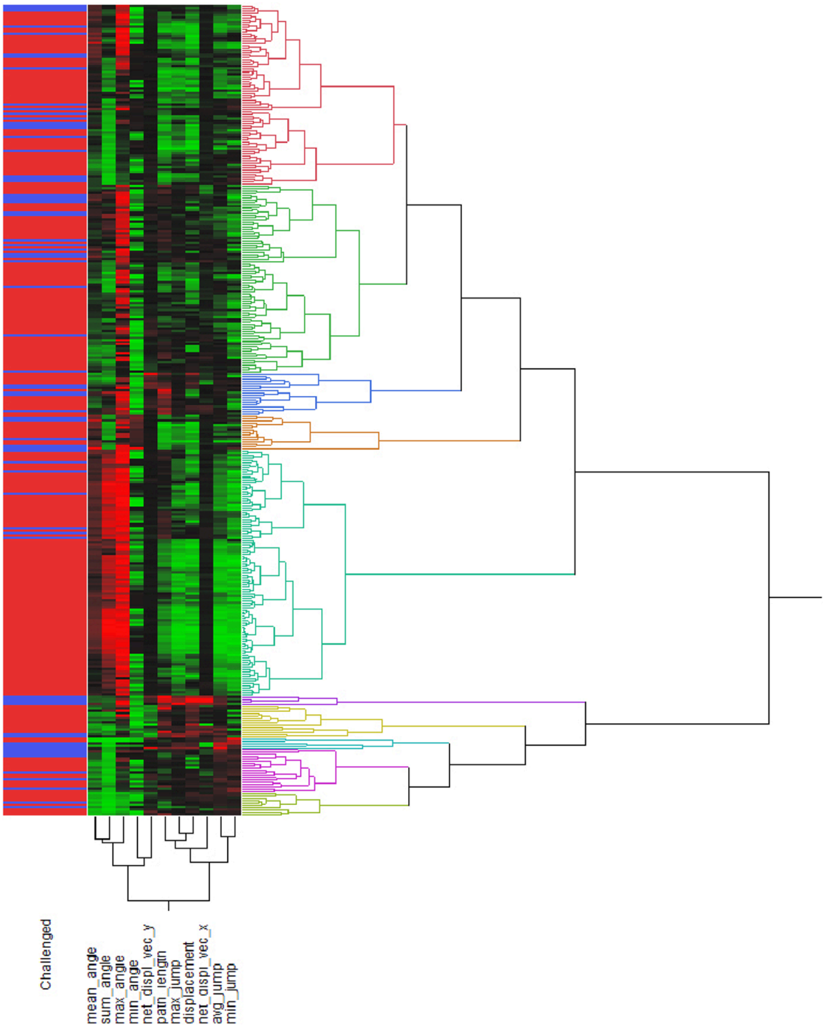


Plot of T Cells Trajectory in Cluster 7 Node 1

T-Cell Trajectory Pattern for Cluster 7 Node 6

We additionally did Ward's hierarchical clustering using all the attributes for malaria-specific T cells in the malaria-infected vs. uninfected mice (Dataset 3; Figure 6). We can see patterns in the attributes that reveal what some clusters were based on, but the clusters do not appear to correspond to infected or uninfected cells (as visualised by the red and blue bars to the left of the heatmap). In Ward's hierarchical clustering, Euclidean distance for each attribute is used in an unsupervised manner to separate the clusters while the SVM uses a supervised approach to find an optimised hyperplane to separate the data. The difference between supervised and unsupervised learning may be the reason that the SVM appears to find differences where the clustering does not.

Figure 6 Hierarchical clustering of parasite-specific T cell movement patterns in an infected mouse (red bars) and an uninfected mouse (blue bars) based on mean angle, sum of angles, max. angle, min. angle, net y displacement, path length, max movement, total displacement, net x displacement, avg. movement, and min. movement. Clustering patterns are visible in the attribute similarities, but do not appear to correlate with whether a CD8 T cell is from an infected or uninfected mouse (see online version for colours)



4 Conclusions and future work

This study has tackled the novel interdisciplinary problem of classifying T cell movement track data using computational tools. The experimental design for Dataset 1 used two distinct T cell types with a true functional difference. Since our aim was to glean functional information based on movement, we were addressing the functional difference indirectly with our attempts to predict it. Unlike Dataset 1, Datasets 2 and 3 contain a single cell type, so we were merely observing variation in movement patterns in Dataset 2 and testing whether T cell movement is predictably different between infected and uninfected mice in Dataset 3. As the levels of accuracy of our SVMs suggest in the case of both Dataset 1 and Dataset 3, functional differences as well as environmental conditions are predictable based on the chosen movement attributes of CD8 T cells. Future work in T cell movement classification will benefit from recent technological advances in visualisation techniques that enable high quality images of cell movement in vivo. With the availability of the underlying tissue topology information in videos of the liver, it is possible to map the movement patterns onto the landscape and test crucial hypotheses about the main determining factors of T cell movement in the tissue. Our widely applicable methods are useful in all organs and organisms in which similar cell movement data can be collected. However, since the necessary labelling and in vivo microscopy for this experiment is not possible to do in humans, the closest human experiments would have to be conducted with human T cells in humanised mice or on an ICAM-1 protein surface (Kuras et al., 2012). Comparing T cell movements between different tissues, and also movements of different cell types using our methods contributes to our understanding of the immune system and opens up a new perspective on the classification of cell types.

References

Achtert, E., Goldhofer, S., Kriegel, H., Schubert, E. and Zimek, A. (2012), 'Evaluation of clusterings metrics and visual support', *Proceedings of the 28th International Conference on Data Engineering (ICDE)*, pp.1285–1288.

Chang, C-C. and Lin, C-J. (2011) 'LIBSVM: a library for support vector machines', *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp.27:1–27:27, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

Cockburn, I.A., Amino, R.c-a., Kelemen, R.K., Kuo, S.C., Tse, S., Radtke, A., Mac-Daniel, L., Ganusov, V.V., Zavala, F. and Menard, R. (2013) 'In vivo imaging of CD8+ T cell-mediated elimination of malaria liver stages', *PNAS*, Vol. 110, No. 22, pp.9090–9095.

Ester, M., Kriegel, H., Sander, J. and Xu, X. (1996) 'A density-based algorithm for discovering clusters in large spatial databases with noise', *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* pp.226–231.

Germain, R.N., Robey, E.A. and Cahalan, M.D. (2012) 'A decade of imaging cellular motility and interaction dynamics in the immune system', *Science*, Vol. 336, No. 6089, pp.1676–1681.

Goodson, S.G., Zhang, Z., Tsuruta, J.K., Wang, W. and O'Brien, D.A. (2011) 'Classification of mouse sperm motility patterns using an automated multiclass support vector machines model', *Biol. Reprod.*, Vol. 84, No. 6, pp.1207–1215.

Harris, T.H., Banigan, E.J., Christian, D.A., Konradt, C., Tait Wojno, E.D., Norose, K., Wilson, E.H., John, B., Weninger, W., Luster, A.D., Liu, A.J. and Hunter, C.A. (2012) 'Generalized Lvy walks and the role of chemokines in migration of effector CD8+ T cells', *Nature*, Vol. 486, No. 7404, pp.545–548.

Kuras, Z., Yun, Y.H., Chimote, A.A., Neumeier, L. and Conforti, L. (2012) 'KCa3.1 and TRPM7 channels at the uropod regulate migration of activated human T cells', *PLoS ONE*, Vol. 7, No. 8, pp.e43859.

Schmidt, N.W., Butler, N.S. and Harty, J.T. (2011) 'Plasmodium-host interactions directly influence the threshold of memory CD8 T cells required for protective immunity', *J. Immunol.*, Vol. 186, No. 10, pp.5873–5884.

Schmidt, N.W., Podyminogin, R.L., Butler, N.S., Badovinac, V.P., Tucker, B.J., Bahjat, K.S., Lauer, P., Reyes-Sandoval, A., Hutchings, C.L., Moore, A.C., Gilbert, S.C., Hill, A.V., Bartholomay, L.C. and Harty, J.T. (2008) 'Memory CD8 T cell responses exceeding a large but definable threshold provide long-term immunity to malaria', *Proc. Natl. Acad. Sci. USA*, Vol. 105, No. 37, pp.14017–14022.

Solem, J. (2012) *Programming Computer Vision with Python: Tools and Algorithms for Analyzing Images*, Oreilly and Associate Series, O'Reilly Media, Incorporated, See http://books.google.com/books?id=gf2LxrmMzXAC

Ward, S.G. and Marelli-Berg, F.M. (2009) 'Mechanisms of chemokine and antigen-dependent T-lymphocyte navigation', *Biochem. J.*, Vol. 418, No. 1, pp.13–27.

Weiss, W.R., Sedegah, M., Beaudoin, R.L., Miller, L.H. and Good, M.F. (1988) 'CD8+ T cells (cytotoxic/suppressors) are required for protection in mice immunized with malaria sporozoites', *Proc. Natl. Acad. Sci. USA*, Vol. 85, No. 2, pp.573–576.