UNSUPERVISED DISCOVERY OF *TOXOPLASMA GONDII* MOTILITY PHENOTYPES

Mojtaba S. Fazli[1] *, Stephen A. Vella*[2]*, Silvia N.J. Moreno*[2]*, Shannon Quinn*[1]

[1] Department of Computer Science, The University of Georgia, Athens, GA, USA

[2] Department of Cellular Biology, The University of Georgia, Athens, GA, USA

ABSTRACT

Toxoplasma gondii is a parasitic protozoan that causes disseminated toxoplasmosis, a disease that afflicts roughly a third of the worlds population. Its virulence is predicated on its motility and ability to enter and exit nucleated cells; therefore, studies elucidating its mechanism of motility and in particular, its motility patterns in the context of its lytic cycle, are critical to the eventual development of therapeutic strategies. Here, we present an end-to-end computational pipeline for identifying *T. gondii* motility phenotypes in a completely unsupervised, data-driven way. We track the parasites before and after addition of extracellular Ca^{2+} to study its effects on the parasite motility patterns and use this information to parameterize the motion and group it according to similarity of spatiotemporal dynamics.

Index Terms— *Toxoplasma gondii*, computer vision, tracking, spectral clustering, autoregressive models

1. INTRODUCTION

Toxoplasma gondii is the causative agent of toxoplasmosis and is considered one of the most successful parasitic infections. Approximately 1/3 of the worlds population will test positive for *T. gondii*, a parasite capable of infecting any nucleated cell [1]. Though *T. gondii* infection is usually mild or asymptomatic within in healthy host, infections are life-long, and the immunocompromised such as AIDS patients, fetuses, and organ-transplant recipients are at risk of developing severe complications from the reactivation of dormant tissue cysts [1, 2, 3]. Ca^{2+} is a universal signaling molecule that regulates numerous cellular processes in both prokaryotes, unicellular and multicellular eukaryotes [4]. *T. gondii*s pathogenesis is directly linked to its lytic cycle, comprised of invasion, replication, egress, and motility, and progression forward is regulated through Ca^{2+} signaling [5]. Therefore, the motion of *T. gondii* is critical to understanding its lytic cycle and, ultimately, in developing potential therapies. Cytosolic Ca^{2+} oscillations in *T. gondii* precede activation/enhancement of each step of the lytic cycle [5], making these oscillations a critical avenue through which to analyze *T. gondii* pathogenesis. Here we present an unsupervised, data-driven pipeline for identifying *T. gondii* motility patterns in response to cytosolic Ca^{2+}. We used extracellular parasites that express in vivo a derivatized variant of the Green Fluorescent Protein (GFP), GCaMP6f, whose fluorescence is directly proportional to Ca^{2+} concentration [6].

We linked fluctuations of fluorescence (cytosolic Ca^{2+} concentration) to motility patterns. Using our algorithm, we extracted and quantified motility tracks, parameterized the motion trajectories, and use spectral clustering to identify distinct motility phenotypes. In particular, we repeat this procedure for *T. gondii* parasites both before and after addition of extracellular Ca^{2+}, which results in Ca^{2+} influx and stimulation of motility [5]. This not only allows us to identify distinct *T. gondii* motility phenotypes in a completely unsupervised, data-driven way, but also enriches our knowledge of these phenotypes with respect to the conditions under which they manifest. Ideally, this work will help pave the way to understanding key differences in the biology of *T. gondii* and mammalian cells and the potential for future therapeutic drug discovery.

2. METHODS

2.1. Data Acquisition

T. gondii tachyzoites (RH) strain were maintained as described previously using Dulbeccos modified minimal essential media with 1% FBS. GCaMP6f is genetically encoded calcium indicator whose fluorescence is sensitive to cytosolic calcium concentration (Ca^{2+}) and allows us to track Ca^{2+} dynamics in motile cells. Construction of GCaMP6f-expressing parasites was performed by cloning the coding sequence of GCaMP6f into a ptubP30GFP (a selection-less plasmid and a kind gift from Boris Striepen). Plasmids were electroporated into RH parasites and subcloned via cell sorting. Videos were acquired using an LSM 710 confocal microscope on 10% FBS (Fetal Bovine Serum) coated glass bottom cover dishes (MATTEK) [5]. Addition of 2M Thapsigargin and/or 1.8 mM CaCl2 was noted. Images were processed using the FIJI ImageJ software suite [7] and converted to z-stacks for processing in OpenCV. For our experiments, we select 10 videos with average number of 14 cells in each video.

2.2. Software

We implemented our pipeline using Python 3 and associated scientific computing libraries (NumPy, SciPy, scikit-learn, matplotlib). The core of our tracking algorithm used a combination of tools available in the OpenCV 3.1 computer vision library. The full code for our pipeline is available under the MIT open source license at `https://github.com/quinngroup/toxoplasma-isbi2018`.

2.3. Computational Pipeline

2.3.1. Tracking Algorithm

We used a Kanade-Lucas-Tomasi (KLT) tracker to develop trajectories for T. gondii parasites. KLT is an optical flow approach which uses spatial intensities to conduct a focused search of high-probability object locations in the future [8][9][10]. It is fast, requiring few matches between sequential images. Here, we identify features extracted from objects in each frame and process them through the KLT module to be tracked over the entire video. The resulting tracker has been shown to be robust in our previous work (Fig. 1) [11].
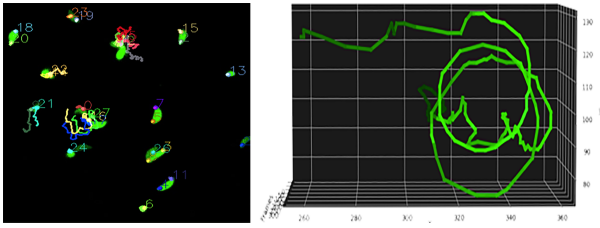

Fig. 1. *(left) Object tracking results using KLT. (right) 3D visualization of a specific cells trajectory. Hue shows the fluorescence of the object, corresponding to Ca^{2+} concentration.*

2.3.2. Preprocessing and normalization of KLT trajectories

KLT trajectories for each parasite object comprised a two-dimensional position vector for each frame. In order to standardize these data to be used in downstream analysis, we employed numerous preprocessing steps. First, we discarded any trajectories below a certain minimum number of frames. This happened if the KLT tracker lost an object (e.g., due to occlusion, movement outside plane of focus). Second, we split the trajectories of each object into two sets, capturing the object trajectory before and after the addition of extracellular Ca^{2+}, to be analyzed distinctly. Third, for each trajectory in the two sets, we truncated the number of frames to be equivalent across all objects.

After applying these preprocessing steps, we ended with a corpus containing 139 trajectories, each capturing the two-dimensional movement of distinct *T. gondii* parasites. Each trajectory consisted of 300 frames, half before and half after the addition of extracellular Ca^{2+}. Our data were represented in the following structures, X and Y (representing x and y coordinates of each object in each frame, respectively):

$$X_{(139,150)} = \begin{bmatrix} x_{(0,0)} & \cdots & x_{(0,149)} \\ \vdots & \ddots & \vdots \\ x_{(138,0)} & \cdots & y_{(138,149)} \end{bmatrix}$$

$$Y_{(139,150)} = \begin{bmatrix} y_{(0,0)} & \cdots & y_{(0,149)} \\ \vdots & \ddots & \vdots \\ y_{(138,0)} & \cdots & y_{(138,149)} \end{bmatrix}$$

2.3.3. Parameterization of normalized KLT trajectories

Performing unsupervised discovery of *T. gondii* motility phenotypes required featurizing the trajectories of each object in such a way as to be invariant to the absolute spatial locations, but sensitive to relative changes over time. Therefore, we used a simple linear model that encodes a Markov-based transition prediction method: an autoregressive (AR) model. AR models are a type of stochastic process, used heavily in modeling linear systems and predicting future observations as a function of some number of previous ones; in this way, it implements the Markov assumption, in which future observations are independent of all others, conditioned on some number of immediate preceding observations. The number of prior observations used to make future predictions is referred to as the order of the system, and follows as:

$$x_t = \sum_j^p A_j x_{t-j} + C \tag{1}$$

where Eq. 1 denotes a p-order system [12]. In our work, we are interested in the transition matrices $\{A1, ..., Ap\}$, which are quantitative encodings of how the system evolves, and therefore represents a parameterization of the system that is not sensitive to absolute spatial positions. In our work, we found a system of order $p = 5$ worked well, whereby each $139x150x2$ object was represented as five 2x2 AR motion parameters:$\{A1, A2, A3, A4, A5\}$. These parameters were flattened and concatenated into a single 20-element feature vector, and this operation was repeated for each of the 139 objects.

2.3.4. Establishing pairwise motion similarity

The AR motion parameters are powerful quantitative representations of movement that are invariant to absolute spatial coordinates. Unfortunately, they are geodesics that do not span a Euclidean space [13], and therefore are not amenable to more traditional unsupervised pattern-recognition strategies such as K-means that rely on Euclidean-based pairwise distance metrics. Instead, we constructed several custom kernel matrices to capture pairwise similarity, with the goal of using these matrices as graph embeddings to use in spectral clustering. To compute a pairwise similarity (kernel) matrix, we stacked the AR parameters learned from the before and after sequences of *T. gondii* motility to form two matrices:

$$AR_{Before(139,20)} = \begin{bmatrix} AR_{b(0,0)} & \cdots & AR_{b(0,19)} \\ \vdots & \ddots & \vdots \\ AR_{b(138,0)} & \cdots & AR_{b(138,19)} \end{bmatrix}$$

$$AR_{After(139,20)} = \begin{bmatrix} AR_{a(0,0)} & \cdots & AR_{a(0,19)} \\ \vdots & \ddots & \vdots \\ AR_{a(138,0)} & \cdots & AR_{a(138,19)} \end{bmatrix}$$

We then computed pairwise affinities, or similarities, using the radial-basis function (RBF) or Gaussian kernel. This kernel matrix A has the form:

$$A_{ij} = A_{ji} = e^{-\gamma||x_i - x_j||}$$

where is a free parameter that designates the width of the Gaussian kernel. Unfortunately, heat kernels are very sensitive to the choice of gamma, so the value was chosen by empirical cross-validation, and selected =0.1 (Fig. 2). Other kernels were also tested, including trajectory-based kernels such as spatial covariance [14], angle covariance (using displacement angles instead of position vectors), and Pearson covariance matrices [15]. However, these did not yield low-dimensional embeddings of sufficient rank to be useful in spectral clustering.

2.3.5. Spectral clustering to identify similar motility phenotypes

Clustering methods are a natural choice for unsupervised learning; they typically impose few assumptions on the data, defining only a notion of similarity with which to group data together. K-means is a favorite for its speed and simplicity; unfortunately, in our case, the drawbacks of K-means make it an impractical option. For one, K-means assumes Euclidean data; AR motion parameters are geodesics that do not live in a Euclidean space. For two, K-means assumes isotropic clusters; it is possible that AR motion parameters exhibit isotropy in their space, but without a proper distance metric this cannot be explicitly tested. For these reasons, we chose spectral clustering. Spectral clustering operates on the spectrum of the underlying graph of the data [16], imposed through a similarity computation using a pairwise kernel. Using this affinity kernel, a graph Laplacian is computed, and its principal eigenvectors are used to embed the original data in a low-dimensional space where they are separable by simpler clustering strategies such as K-means. While this requires a full diagonalization of the Laplacian and can therefore pose a computational bottleneck with large data, our 139 trajectories is sufficiently small for to be solved by out-of-the-box spectral clustering implementations, such as scikit-learn [17]
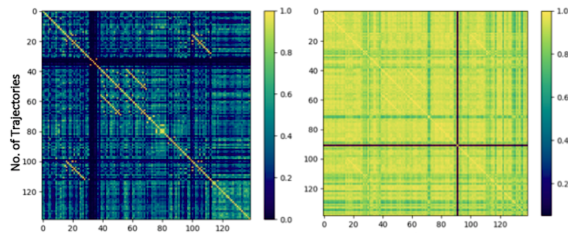

Fig. 2. *RBF kernels from AR motion parameters using gamma = 0.1, from AR motion parameters computed before (left) and after (right) addition of cytosolic Ca^{2+}.*

3. RESULTS AND DISCUSSION

Using established heuristics from previous work on the *T. gondii* parasite and its lytic cycle [5, 11], we chose k = 5 clusters for our analysis. Figure 3 shows a summary of the clustering results. The left column indicates trajectories from before the addition of cytosolic Ca^{2+}, and the right column of after Ca^{2+} addition. Each subfigure is a single object, illustrated by a line showing the extent of its movement over the 150-frame duration of the segment.

First, we observe that object trajectories are significantly richer and more dynamic after the addition of cytosolic Ca^{2+} than before; we saw this in Fig. 1 (right), and is well-understood that since Ca^{2+} regulates the *T. gondii* lytic cycle, we can expect more activity when those concentrations are high [4, 5]. Second, and most relevant to the goal of this work, is that we can observe qualitatively unmistakable similarity of trajectory dynamics, especially in the after-Ca^{2+} clusters. Most of the before-Ca^{2+} trajectories are largely immotile (Fig. 3, bottom left), but there are clusters that show similar motion motifs, including full or partial circles (Fig. 3, upper left). The after-Ca^{2+} clusters show similar groupings of motifs: a clear partial-twirl, sometimes combined with a straight trajectory (Fig. 3, lower right), or seemingly random motion with a long outward arc (Fig. 3, upper right). Particularly when viewed in context (Fig. 4), we can clearly
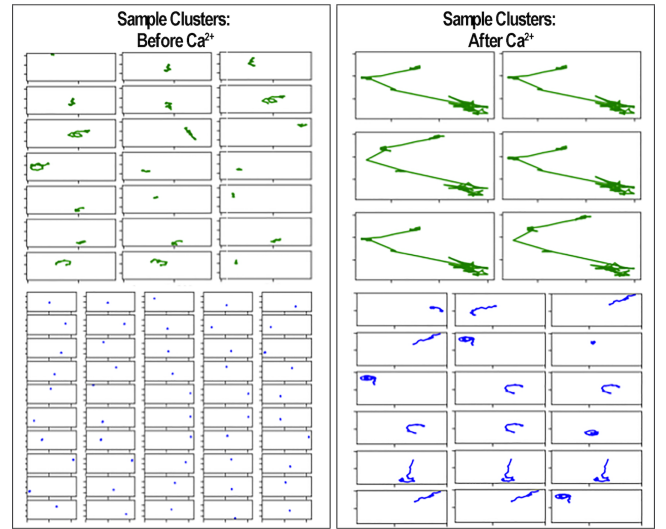

Fig. 3. *Clustering results of trajectories before (**left column**) and after (**right column**) addition of cytosolic Ca^{2+}. Two sample clusters are shown in each column. Subplots are aligned to center for easy viewing.*

identify several distinct *T. gondii* motility phenotypes such as helical, twirling, and circularthat are identified through manual heuristics [1-5].

There are limitations to this approach. First, while the RBF kernel encodes nonlinear relationships, it still relies on a Euclidean-style pairwise comparison. Second, despite the empirical performance of AR parameters in encoding two-dimensional spatial movement, this is not an ideal application of this technique; it is likely that the relative simplicity of *T. gondii* motility patterns, and the small number of motion motifs, are responsible for the relatively good empirical per-

formance. Third, the combination of small quantity of data and low-resolution temporal models precludes a deeper exploration of the manifold of *T. gondii* motility phenotypes. Nevertheless, the results strongly suggest such a manifold exists, and could be elucidated in future work.
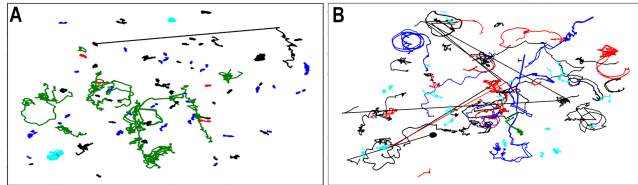

Fig. 4. *Aggregated plotting of trajectories before (A) and after (B) addition of cytosolic Ca^{2+}. Each object is colored by its cluster identifier, and placed according to its original absolute spatial location. Here, some trajectories(straight lines mostly in part B) are artifacts and were excluded from the final clustering results.*

4. CONCLUSION

In this work, we have demonstrated a computational pipeline for tracking, extracting, parameterizing, and clustering motion trajectories of *T. gondii* parasites in an initial effort to elucidate discrete patterns in the parasites lytic cycle. This is a first step toward a mechanistic understanding of the parasites virulence as a function of its motility patterns. We have also demonstrated the application of a linear time series (AR) model to the parameterization of two-dimensional spatial motility, and its effectiveness in clustering for recognizing gross motion patterns. Given the relatively short tracks of the *T. gondii* objects, we felt AR models (i.e., a linear model) would be sufficient to parameterize motion. This technique could be applied pretty much as-is to any trajectory analysis of discrete objects.

In the future, we anticipate refining the elements of this pipeline to extract finer-grained motility patterns. Specifically, we anticipate deep learning to be the next evolution of this pipeline: using segmentation and tracking tools such as Faster RCNN to track the objects, and parameterizing the motility sequences using recurrent neural networks instead of linear systems.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] Geita Saadatnia and Majid Golkar, "A review on human toxoplasmosis," *Scandinavian journal of infectious diseases*, vol. 44, no. 11, pp. 805–814, 2012.

[2] Francis Derouin and H Pelloux, "Prevention of toxoplasmosis in transplant patients," *Clinical Microbiology and Infection*, vol. 14, no. 12, pp. 1089–1101, 2008.

[3] Jeffrey L Jones, Monica E Parise, and Anthony E Fiore, "Neglected parasitic infections in the united states: toxoplasmosis," *The American journal of tropical medicine and hygiene*, vol. 90, no. 5, pp. 794–799, 2014.

[4] David E Clapham, "Calcium signaling," *Cell*, vol. 80, no. 2, pp. 259–268, 1995.

[5] Lucas Borges-Pereira, Alexandre Budu, Ciara A McKnight, Christina A Moore, Stephen A Vella, Miryam A Hortua Triana, Jing Liu, Celia RS Garcia, Douglas A Pace, and Silvia NJ Moreno, "Calcium signaling throughout the toxoplasma gondii lytic cycle a study using genetically encoded calcium indicators," *Journal of Biological Chemistry*, vol. 290, no. 45, pp. 26914–26926, 2015.

[6] Tsai-Wen Chen, Trevor J Wardill, Yi Sun, Stefan R Pulver, Sabine L Renninger, Amy Baohan, Eric R Schreiter, Rex A Kerr, Michael B Orger, Vivek Jayaraman, et al., "Ultrasensitive fluorescent proteins for imaging neuronal activity," *Nature*, vol. 499, no. 7458, pp. 295–300, 2013.

[7] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al., "Fiji: an open-source platform for biological-image analysis," *Nature methods*, vol. 9, no. 7, pp. 676–682, 2012.

[8] Bruce D Lucas, Takeo Kanade, et al., "An iterative image registration technique with an application to stereo vision," 1981.

[9] Carlo Tomasi and Takeo Kanade, "Detection and tracking of point features," 1991.

[10] Jianbo Shi et al., "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

[11] Mojtaba Sedigh Fazli, Stephen Andrew Vella, Silvia NJ Moreno, and Shannon Quinn, "Computational motility tracking of calcium dynamics in toxoplasma gondii," *arXiv preprint arXiv:1708.01871*, 2017.

[12] P. BOURKE, "Auto-regression analysis (ar)," *http://local.wasp.uwa.edu.au/pbourkke/other/ar/*, 1998.

[13] Franco Woolfe and Andrew Fitzgibbon, "Shift-invariant dynamic texture recognition," *Computer Vision–ECCV 2006*, pp. 549–562, 2006.

[14] Andrew Gordon Wilson, "Covariance kernels for fast automatic pattern discovery and extrapolation with gaussian processes," *University of Cambridge*, 2014.

[15] Bruce Weaver and Karl L Wuensch, "Spss and sas programs for comparing pearson correlations and ols regression coefficients," *Behavior Research Methods*, vol. 45, no. 3, pp. 880–895, 2013.

[16] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[17] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.