Spring 2024

PREDICTING FUTURE STATES WITH SPATIAL POINT PROCESSES IN SINGLE MOLECULE RESOLUTION SPATIAL TRANSCRIPTOMICS

Biraaj Rout
University of Texas at Arlington

PREDICTING FUTURE STATES WITH SPATIAL POINT PROCESSES IN SINGLE MOLECULE RESOLUTION SPATIAL TRANSCRIPTOMICS

a thesis presented
by
Biraaj Rout
to
The Department of Computer Science and Engineering

in partial fulfillment of the requirements
for the degree of
Master
in the subject of
Computer Science

University of Texas at Arlington
Arlington, Texas
December 2023

Thesis advisor: Dr. Jacob Luber Biraaj Rout

PREDICTING FUTURE STATES WITH SPATIAL POINT PROCESSES IN SINGLE MOLECULE RESOLUTION SPATIAL TRANSCRIPTOMICS

Abstract

In this thesis, we present an innovative framework centered around the application of Random Forest Regression to forecast the prospective distribution of cells expressing the Sog-D gene (active cells) during the embryogenesis process in Drosophila. Our methodology specifically targets the Anterior-to-posterior (AP) and Dorsal-to-Ventral (DV) axes, unraveling the intricacies of gene expression control in living organisms at super-resolution, single-molecule resolution through whole embryo spatial transcriptomics imaging. The Random Forest Regression model serves as a pivotal tool in predicting the succeeding stage's active cell distribution, capitalizing on the insights obtained from the preceding stage. We integrate temporally resolved, spatial point processes into our analysis, incorporating Ripley's K-function alongside the cell's state at each embryogenesis stage. Our approach yields an average predictive accuracy for active cell distribution, providing a valuable tool akin to RNA Velocity for spatially resolved developmental biology. This framework empowers researchers to extrapolate future spatially resolved gene expression from a singular data point, leveraging features derived from spatial point processes. Through this thesis, we contribute to advancing the understanding of developmental biology, offering a robust methodology for predicting gene expression dynamics at sub-cellular resolutions.

Contents

1

Introduction

In recent years, the landscape of biological research has undergone a transformative evolution, driven by remarkable technological advancements that empower scientists to unravel the intricacies of embryogenesis. High-resolution imaging techniques, such as those enabled by advancements in microscopy and live-cell imaging, have emerged as crucial tools, facilitating the capture of intricate details during embryonic development. These techniques provide unprecedented insights into the regulatory mechanisms governing gene expression patterns [14,17]. This technological progress is particularly crucial in the context of contemporary genomics, where understanding the intricate orchestration of gene regulation is a central challenge. Deciphering how gene expression dynamically shapes spatiotemporal outputs throughout development has become a paramount goal.

The early Drosophila embryo, a well-established model system, stands as a paradigmatic example, offering profound insights into the nuanced control of patterning orchestrated by enhancers. This model has revealed intricate complexities and the dynamic nature inherent in the patterning process. Notably, certain genes are influenced by multiple transiently acting enhancers, orchestrating sequential changes in expression, while others are governed by enhancers with prolonged effects that support spatial alterations over time[7,3,19].

Despite significant strides in genetic and live imaging techniques, the analytical methodologies to fully exploit the wealth of information embedded within real-time imaging of transcriptional dynamics have lagged behind the field. Current methodologies predominantly rely on static parameter cell and transcript tracking techniques, leaving room for innovation and advancement in analytical approaches[6,18]. To address this gap, our study introduces a quantitative approach aimed at systematically assessing mutant enhancer phenotypes. We collected data from wet lab collaborators consisting of live images expressing genes and developed new methods providing crucial information on the timing, levels, and spatial domains of gene expression.

In the utilization of transgenic fly lines, our investigation incorporates live imaging of the GFP signal associated with the MS2 stem-loop reporter sequence. This unique MS2 cassette, comprising 24 repeats of a DNA sequence, generates an RNA stem-loop upon transcription. The stem-loop structure selectively binds to the phage MS2 coat protein (MCP), fused to GFP, resulting in a robust green signal within the nuclei of Drosophila embryos at sites of nascent transcript production. The imaging protocol, optimized for both spatial coverage and temporal resolution, utilizes a Zeiss LSM 900 and captures the entire dorsal-ventral (DV) axis of embryos over a 2-hour period at a notably improved interval of 30 seconds per scan compared to prior studies.

Acknowledging the dynamic nature of spatial outputs during embryonic development, our study pioneers an image-processing approach to gather information across both time and space. By concentrating on one lateral half of the embryos, our goal is to predict the distribution of active cells(both horizontally and vertically) at each stage of embryo development. Inspired by innovative concepts such as RNA velocity[11] and spatial proteomics data analysis utilizing Ripley's K-function[4], our proposed pipeline integrates a novel feature extraction method and analysis framework. This innovative approach holds the promise of predicting the future distribution of cells expressing the Sog-D gene, thereby contributing substantially to a deeper understanding of the intricate dynamics of gene expression during embryogenesis.

This work was inspired by previous pioneers in the field[11,4], shaping the trajectory of our research. Furthermore, recent studies such as[9] have emphasized the importance of integrating multi-omic approaches in understanding gene regulation dynamics, providing a broader context for our endeavors. The convergence of advanced imaging techniques and cutting-edge analytical methods represents a frontier in contemporary genomics research, offering unprecedented opportunities to decipher the complexities of gene expression regulation during embryogenesis.

2

Paper

Abstract

In this paper, we introduce a pipeline based on Random Forest Regression to predict the future distribution of cells that are expressed by the Sog-D gene (active cells) in both the Anterior to posterior (AP) and the Dorsal to Ventral (DV) axis of the Drosophila in embryogenesis process. This method provides insights into how cells and living organisms control gene expression in super-resolution whole embryo spatial transcriptomics imaging at sub-cellular, single-molecule resolution. A Random Forest Regression model was used to predict the next stage active distribution based on the previous one. To achieve this goal, we leveraged temporally resolved, spatial point processes by including Ripley's K-function in conjunction with the cell's state in each stage of embryogenesis and found average predictive accuracy of active cell distribution. This tool is analogous to RNA Velocity for spatially resolved developmental biology, from one data point we can predict future spatially resolved gene expression using features from the spatial point processes.

Index terms— Random Forest, Regression, Dorpsophila, Sog-D, Ripley's K-function, transcriptomics, embryogenesis

2.1 Introduction

Recent technological advances have made it possible to capture high-resolution images from the embryogenesis process that help researchers to study gene expression patterns. [10,5]. One of the major challenges of the modern genomics era is to better understand how gene expression is regulated to support spatiotemporal outputs that change over the course of development. The early Drosophila embryo has served as a paradigm for how enhancers control patterning and has demonstrated that the patterning process is complex and dynamic. It is known that multiple, transiently acting enhancers act sequentially to support changing outputs of expression for some genes[5,13,16], whereas other genes are controlled by enhancers that act over a longer period and support changing spatial outputs over time. For example, expression of the gene short gastrulation (sog) is driven by at least two co-acting enhancers that support temporally dynamic expression. Live imaging experiments offer the capacity to analyze gene expression dynamics with increased temporal resolution and linear quantification. However, genetic and live imaging techniques have outpaced analysis techniques to harvest the bountiful information contained within real-time movies of transcriptional dynamics with modern methods confined to static parameter cell and transcript tracking methods [10,12,1]. To assess these mutant enhancer phenotypes systematically, we developed a quantitative approach to measure the spatiotemporal outputs of enhancer-driven MS2-yellow reporter constructs as captured by in vivo imaging to provide information about the timing, levels, and spatial domains of expression. Using transgenic fly lines, we conducted live imaging of the GFP signal associated with the MS2 stem-loop reporter sequence. This MS2 cassette contains 24 repeats of a DNA sequence that produces an RNA stem loop when transcribed. The stem-loop structure is specifically bound by the phage MS2 coat protein (MCP). MCP fused to GFP binds to MS2-containing transcripts (i.e., sog_Distal.MS2) producing a strong green signal within the nuclei of Drosophila embryos at sites of nascent transcript production. In this system, the nuclear GFP signal is only observed as a single dot for every nucleus corresponding to nascent transcription of the one copy of the MS2-containing reporter transgene site integrated into the genome. Furthermore, the nuclear periphery is marked by a fusion of RFP to nuclear pore protein (Nup-RFP) [14]. The imaging protocol was optimized to provide spatial information across the entire dorsal-ventral (DV) axis of embryos with the fastest temporal resolution that also retains embryo viability. In brief, embryos were imaged on Zeiss LSM 900 continuously over the course of 2hr at an interval of 30s per scan (twice as fast compared to previous studies). Importantly, this imaging protocol is not phototoxic to embryos. Because spatial outputs likely change in time across the embryo for many gene expression patterns, we developed an image-processing approach to collect detailed information in both time and space by capturing one lateral half of the embryos. With this qualified imaging dataset, our goal was to predict the distribution of active cells in each stage of the embryo development. Several methods have been proposed for the efficient prediction of temporal variables. Authors in [11] proposed a novel concept called RNA velocity, which is defined as the time derivative of the gene expression. This concept allows for the estimation of the future state of individual cells in standard scRNA-seq protocols. In [4], authors proposed a method to capture spatial proteomics data to map cell states in order to predict cancer patient survival. They utilized Ripley's K-function for capturing spatial features which inspired us in our proposed pipeline. We developed a feature extraction method and analysis pipeline that can be used to predict the future distribution of cells in which the Sog-D gene is expressed.

2.2 Methods

We generated super resolution live imaging data expressing *sog* gene (control) and *sog-D* gene (case) in early embryo of *Drosophila* (9 case, 4 control). We conduct pre-processing, feature extraction, training, and testing Fig.2.1. Both the training and testing phases incorporate identical pre-processing and feature extraction steps. The videos shows real time images from embryonic development, which were manually given stage development labels: NC 13 early, NC 13 late, NC 14 A, NC 14 B, NC 14 C, NC 14 D. In the pre-processing step, we used a generalist, deep learning-based segmentation method called Cellpose, which can precisely segment cells in each frame of the embryo development. Active cells were identified based on prevalance of green pixels indicative of gene expression within the cell, and the active mask underwent feature extraction. During this stage, the masked images underwent a gridding procedure with a predetermined size. Subsequently, the entire imaging dataset was transformed into a tabular format, taking into account the spatial information of each cell. We utilized four different metrics to capture both local and global features in a frame including $m1$, $m2$ for both AP and DV axes, Ripley's k-function, and n (total number of cells in each grid). Here, $m1$ and $m2$ denote the first and second moments, respectively, capturing the distribution of active cells at each stage. Furthermore, Ripley's k-function was employed to analyze spatial correlation and quantify deviations from a random spatial distribution. Equation 2.1 illustrates the formula for calculating Ripley's k-function. Where, A is the area under each window with a constant radius, n is the number of data points, d_{ij} is the distance between two points, and e_{ij} is an edge correction weight. Then, the tabular data went through two steps of averaging on each stage

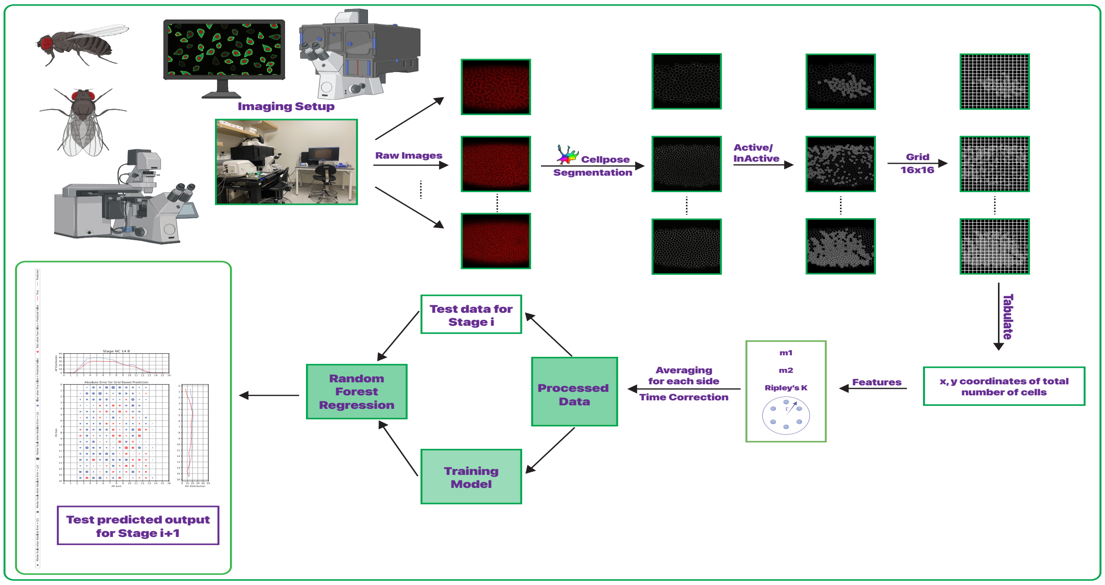

Figure 2.1: Implemented pipeline, starting with the imaging process, followed by subsequent stages involving pre-processing, feature extracting, training ,and testing. These steps collectively aim to predict the distribution of active cells for the next stage.

and time correcting. Since our goal is to predict the distribution of active cells in each stage and we have different numbers of frames for each stage, we averaged the whole feature values based on each stage. Also, to account for temporal alignment, we implemented a one-stage shift in features, where we utilized the features from the previous stage in the prediction of the current stage. Following the completion of the feature extraction process, the dataset undergoes preparation for training a random forest regression model, a supervised learning algorithm. The outcome of this pipeline is the count of active cells within each grid at a given stage, determined by the features from the preceding stage. Subsequent to training the model, its performance is evaluated using test data. During testing, all pre-processing and feature extraction steps are replicated, and the pre-trained random forest regression model is employed to forecast the count of active cells for each grid across various stages.

$$\hat{K}_r = \frac{A}{n(n-1)} \sum_{i=1}^{n} \sum_{i=1,j\neq i}^{n} 1(d_{ij} \leq r)e_{ij} \qquad (2.1)$$

2.3 Experiment and Results

2.3.1 Main study

As outlined in the methodology section, during the feature extraction phase, square grids were applied to images, and the number of active cells within each grid was predicted. The key challenge was selecting the optimal grid size to enhance performance on test data. Consequently, we replicated the entire process of pre-processing and feature extraction for four distinct grid sizes: 250, 125, 62.5, and 31.25 (where the grid size of 'n' indicates the division of the entire image into n*n squares). We used three different metrics to calculate the model performance on test data for different grid sizes which are rmse (root mean squared error), mae (mean absolute error), and Kullback-Leibler (KL) Divergence. Fig.2.2 shows the experiment for different grid sizes. Our analysis revealed the same increasing trend in both rmse and mae as the grid size increases from 31.25 to 250 which indicated that a smaller grid size corresponds to a lower error. KL Divergence, which we also utilized as a metric, measures how one probability distribution diverges from a second one. Thus, the smaller value for it shows that the two distributions are closer to each other. We used this criterion to see how well the pipeline can capture the trends in the active cell distribution. The KL Divergence for these four different grid sizes showed a different trend. Increasing the grid size from 31.25 to 250 yielded a decrease in KL Divergence. We had two options, the first one was to select 31.25 based on the lower rmse and mae. However, the problem was the average size of the cell was approximately 36 so if we set the grid size to 31.25 we have just one cell in each grid which changes the problem to a classification of active or inactive for each grid which was not our purpose. Another option was to select the optimal grid size based on KL Divergence, which finally, We selected the grid size of 62.5 over 31.25. The decision of selecting 63.5 over 125.0 although the 125 had lower KL Divergence, is attributed to the computational constraints of calculating Ripley's k-function for larger grid sizes in our setup.

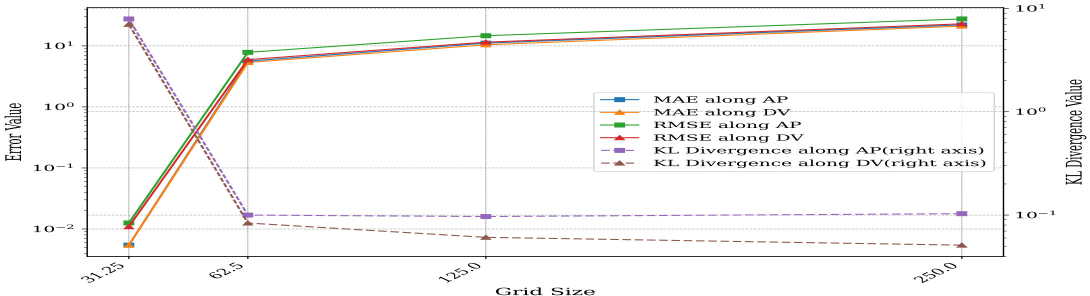

Figure 2.2: The experiment for grid search to find the optimal grid size

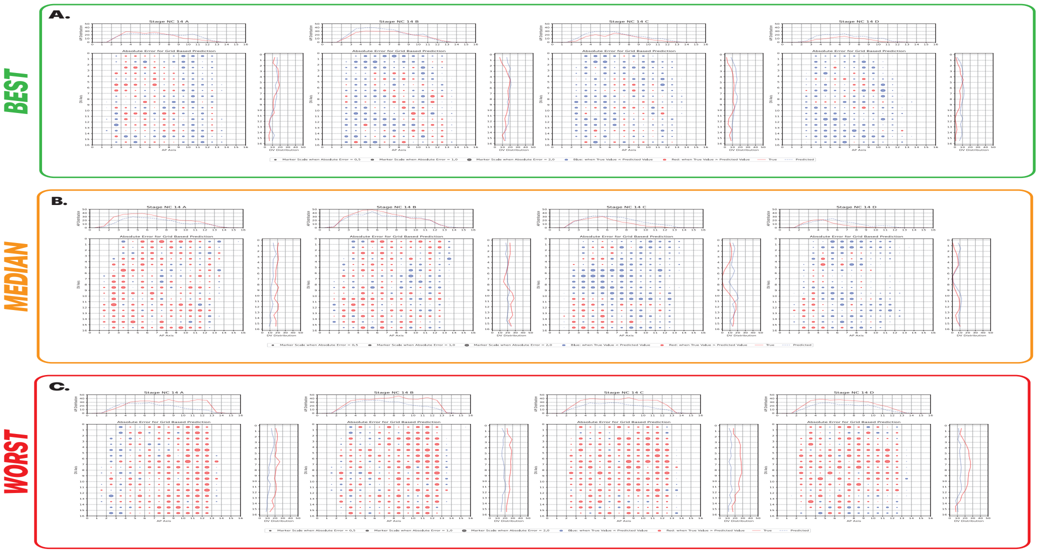

Figure 2.3: The distribution of active cells for the best (A), median (B), and worst (C) accuracy based on mae values. For each A, B, and C from left to right stages are NC 14 A-D. For each stage, the top and right plots show the distribution of active cells along the AP and DV axis respectively. The middle plot shows the absolute error in each grid.

In subsequent experiment, we conducted an ablation study to discern the relative importance of features, identifying those deemed crucial for inclusion in the final release and those that may be omitted. Table 2.1 indicates the performance of different combinations of features. It can be concluded that features of the first row including Ripley's k-function and n are the most important features that we used for training and testing the pipeline. All reported mae values underwent the K-fold cross-validation method to mitigate the influence of random results.

To visualize the performance of the pipeline with selected features and parameters we tested the pre-trained model on the test dataset. Fig 2.3 shows the distribution of active cells for the best, me-

Feature list	mae
n, Ripley's k-function	4.53
m2_DV, n, Ripley's k-function	4.73
m1_DV, n, Ripley's k-function	4.75
m1_DV, m2_AP, n, Ripley's k-function	4.77
m2_AP, n, Ripley's k-function	4.77

Table 2.1: The average mae value on K-fold cross-validation over test dataset for different combinations of features for ablation study.

dian and the worst prediction based on the average mae values.

2.3.2 Case and control study

As we had 4 videos for case (transgenic) and 9 for control, we randomly selected 3 videos from each group for training and 1 for testing. Then, we averaged the *AP_mae*, *DV_mae*, and *mean_mae* for the whole case and control experiments and calculated the difference between case and control for each of these metrics and the results were 1.86, -0.689, and 0.58 respectively. We also utilized cross-validation to avoid overfitting. These results show there is a difference between the performance of our pipeline on case and control in *AP_mean* and *mean_mae*. In other words, our method works better in predicting along AP axis and the mean of AP and DV on control data in comparison with the case one. However, the negative difference between case and control for *DV_mae* indicates that the pipeline works better in predicting the distribution on the DV axis of the case compared to the control. In order to substantiate this assertion, we conducted two additional experiments: First, we leveraged Mixed-Effects modelling, which can account for both fixed effects (like the group: case or control) and random effects (like the variation within videos and stages). The mixed-effects model can help in understanding the influence of these fixed and random effects on our dependent variables like *DV_mae*, *AP_mae*, *mean_mae*. The goal is to understand whether there is a significant difference in any metrics between the case and control groups, accounting for the variability introduced by different stages. The control group has, on average, a lower *AP_mae* compared to the case by about 1.828 units with the *P_value* of 0.003. It shows based on this test, there is a statistically significant difference in *AP_mae* between case and control groups. However, the result for *DV_mae* shows the control group has a higher value by 0.714 units and 0.231 *P_value*. Also, the result for *mean_mae* indicates control has a higher value by -0.557 units and 0.347 *P_value*. Two latter results for *DV_mae* and *mean_mae* cannot indicate any significant difference between case and control because of the high *P_values*. In addition, we implemented another empirical hypothesis testing called Bootstrap method. Bootstrap methods can be used to estimate the distribution of our metrics under the null hypothesis. To implement the bootstrap, we used the same metrics as previous method. we drew samples from the original dataset with replacement, to create a new dataset. Then, for each bootstrap sample, we computed the statistics of interest which are *DV_mae*, *AP_mae*, and *mean_mae*. By analyzing this bootstrap distribution we can find the confidence intervals for each metric. Fig 2.4 shows the Bootstrap distribution of the mean difference in *AP_mae*, *DV_mae*, and *mean_mae*. It indicates that with 95% confidence interval the mean difference of *AP_mae*, (*AP_mae*(case) - *AP_mae*(control)) was between [0.69061964 3.11528348]. It can be concluded that with 95% confidence interval the *AP_mae* for case is at least 0.69061964 units higher than case, which means the performance of the pipeline is better for control outperforms case one. These ranges for *DV_mae* and *mean_mae* are respectively, [-1.65878863 0.27041668] and [-0.33784703 1.5450897]. It can be seen that for *DV_mae* and *mean_mae* the ranges include zero means the performance of control can be better, equal, or worse than the case. The results with the Bootstrap method confirm the results derived from the mixed-effects method, which makes sense given that large amounts of training data are needed to model transgenic effects.

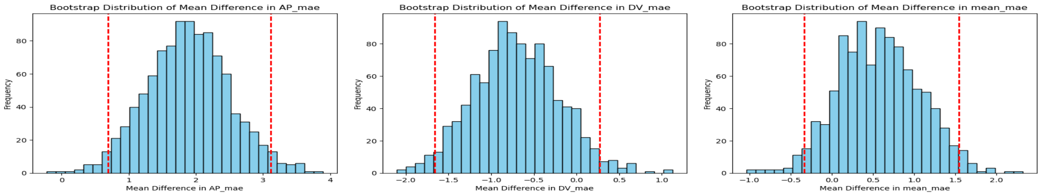

Figure 2.4: The Bootstrap Distribution of Mean Difference in AP_mae, DV_mae, and mean_mae between case and control in 1000 iterations.

2.4 Conclusion

Our work presents several key contributions. Firstly, we have developed a novel and optimized imaging technology that delivers spatial information throughout the entire DV axis of an embryo. Secondly, we introduce an automated pipeline that effectively discriminates cell types with high accuracy. Lastly, our approach enables the accurate prediction of the stage-level distribution of active cells, based on data from the preceding stage.

2.5 Compliance with Ethical Standards

All animal experiments were approved by the UTA IACUC review board. This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of my institution.

2.6 Acknowledgments

This work was supported by the Cancer Prevention and Research Institute of Texas (CPRIT) Recruitment of First-Time, Tenure-Track Faculty Members Grant (RR220015) (JML) and the University of Texas System Rising STARs Award (JML).

3 Conclusion

3.1 Challenges

We were able to use the novel imaging technique with our automated pipeline to predict the stage-level distribution of active cells using the information from previous stages even though we had less data. This is a major advance for the field. Despite these promising results, achieving them yielded substantial difficulties that had to be overcome. One of the major challenges we faced was selecting the optimal grid size to accommodate enough active cells after the segmentation as the grid size cannot be less than the cell size. Also, we were limited by computational time to get Ripley's k for active cell distribution in each grid when it became smaller. Thus, we did the entire process over 5 grid sizes and found 63.5 to be the most optimal.

There were a few challenges during and after segmentation. Before cell segmentation, we had to resize images to a proper size where we don't lose information and after multiple tests, we finalized the size to be 1000*1000. The green pixels in the image signify genes being expressed. Thus, after segmentation of raw images to identify active cells we had to check the intensity of green pixels which was quite uncertain initially but by modifying the intensity value to capture active cells with multiple iterations, we fixed a certain value which was verified by a biologist and obtained a good accuracy.

There are still multiple challenges remaining such as aligning the grouped cell portion in the image to get better gene expression coverage. Also, there are some cells having gene expressions that were missed out by the cellpose segmentation due to the deformed structure of the cell and blurred portions which arose due to cells being 3D and we were segmenting them in 2D. We also need more data to have a wide coverage of later cell stages and not limited by the initial ones.

In the future, we anticipate having more videos to capture different gene expressions and predict the underlying distribution. Also, more ways can be explored to predict active cell distribution of later stages with just a few initial stages. Accuracy for predicting active cell distribution along the DV axis can also be improved by tuning the features after we get more data on various case and control gene expressions.

3.2 Future Directions

I am planning to join the UTA Ph.D. program and extend this initial computational modeling work into larger-scale LLM research with the goal of detecting biosynthetic gene clusters improving previous work at Merck Research Lab[8] building upon previous computational BGC work at Merck, Princeton, and Stanford. Biosynthetic gene clusters (BGCs) are groups of genes in the genome of an organism that are responsible for the production of a specific natural product or bioactive compound. These clusters typically include genes encoding enzymes involved in the biosynthesis of the compound, as well as genes for regulatory elements and other supporting functions. Natural products produced by BGCs include antibiotics, antifungals, anticancer agents, and other bioactive molecules[8,15].

The DeepBGC[8] paper demonstrates reduced false positive rates and an improved ability to identify novel BGC classes compared to existing machine-learning tools. Additionally, random forest classifiers are employed to accurately predict BGC product classes and potential chemical activity. The application of DeepBGC to bacterial genomes reveals previously undetectable putative BGCs, suggesting the potential for the discovery of natural products with novel biological activities. The enhanced accuracy and classification capabilities of DeepBGC make it a valuable tool for in-silico BGC identification.

The updated version 6 of ”antibiotics and secondary metabolite analysis shell—antiSMASH”[2] is introduced, enhancing microbial genome mining for natural product discovery. This widely used tool now supports 71 cluster types, displays modular structures of Multi-modular Biosynthetic Gene Clusters (BGCs), incorporates a new BGC comparison algorithm, integrates results from other prediction tools, and improves the detection of tailoring enzymes in RiPP clusters. antiSMASH 6 provides researchers with advanced features and expanded capabilities for the characterization of BGCs in bacteria and fungi, thereby facilitating the discovery of novel natural products.

The ongoing research focuses on computationally detecting Biosynthetic Gene Clusters (BGCs), and my role involves advancing the current state of this field. I plan to enhance BGC detection through innovative approaches and methodologies. Subsequently, I aim to collaborate with Professor Joe Buanomo in the field of chemistry to computationally validate the predicted metabolites using Mass Spectrometry. This collaborative effort seeks to bridge the computational predictions with experimental validation, providing a comprehensive and robust exploration of natural product discovery.

BERT-like tokenization and Language Models (LLMs) can significantly enhance the effectiveness of the DeepBGC[8] strategy described in the paper. By employing BERT-like tokenization, which captures contextual information and relationships between words, the model gains a more nuanced understanding of the biosynthetic gene clusters (BGCs) and their associated natural products. This allows DeepBGC to better discern subtle patterns in genomic data, reducing false positive rates in BGC identification. Additionally, LLMs contribute by leveraging pre-trained language representations, enabling the model to generalize across various genomic sequences and identify novel BGC classes more accurately. The comprehensive contextual information provided by BERT-like tokenization, coupled with the generalization capabilities of LLMs, empowers DeepBGC to offer improved precision in BGC identification and classification, ultimately advancing the state of in-silico BGC exploration for natural product discovery.

Understanding the nuances of gene expression during the research for the prediction of future states in single molecule spatial transcriptomic and usage of statistical machine learning methods along with various projects involving application over multiplexed codex images with the usage of deep learning over various tissue and proteomics data I will be pursuing modeling of biological systems at a larger scale.

References

[1] Birnie, A., Plat, A., Korkmaz, C., & Bothma, J. P. (2023). Precisely timed regulation of enhancer activity defines the binary expression pattern of fushi tarazu in the drosophila embryo. *Current Biology*.

[2] Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., van Wezel, G. P., Medema, M., & Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research*, 49(W1), W29–W35.

[3] Cho, C.-Y. & O'Farrell, P. H. (2023). Stepwise modifications of transcriptional hubs link pioneer factor activity to a burst of transcription. *Nature Communications*, 14(1), 4848.

[4] Dayao, M. T., Trevino, A., Kim, H., Ruffalo, M., D'Angio, H. B., Preska, R., Duvvuri, U., Mayer, A. T., & Bar-Joseph, Z. (2023). Deriving spatial features from in situ proteomics imaging to enhance cancer survival analysis. *Bioinformatics*, 39(Supplement_1), i140–i148.

[5] Dunipace, L., Saunders, A., Ashe, H. L., & Stathopoulos, A. (2013). Autoregulatory feedback controls sequential action of cis-regulatory modules at the brinker locus. *Developmental cell*, 26(5), 536–543.

[6] Dutta, S., Djabrayan, N. J.-V., Torquato, S., Shvartsman, S. Y., & Krajnc, M. (2019). Self-similar dynamics of nuclear packing in the early drosophila embryo. *Biophysical journal*, 117(4), 743–750.

[7] Fukaya, T. (2021). Dynamic regulation of anterior-posterior patterning genes in living drosophila embryos. *Current Biology*, 31(10), 2227–2236.

[8] Hannigan, G. D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., Wang, R., Piizzi, G., Temesi, G., Hazuda, D. J., Woelk, C. H., & Bitton, D. A. (2019). A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*, 47(18), e110–e110.

[9] Israel, S., Ernst, M., Psathaki, O. E., Drexler, H. C., Casser, E., Suzuki, Y., Makalowski, W., Boiani, M., Fuellen, G., & Taher, L. (2019). An integrated genome-wide multi-omics analysis of gene expression dynamics in the preimplantation mouse embryo. *Scientific Reports*, 9(1), 13356.

[10] Koromila, T. & Stathopoulos, A. (2019). Distinct roles of broadly expressed repressors support dynamic enhancer action and change in time. *Cell reports*, 28(4), 855–863.

[11] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. (2018). Rna velocity of single cells. *Nature*, 560(7719), 494–498.

[12] Lim, B., Heist, T., Levine, M., & Fukaya, T. (2018). Visualization of transvection in living drosophila embryos. *Molecular cell*, 70(2), 287–296.

[13] Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell*, 167(5), 1170–1187.

[14] Lucas, T., Ferraro, T., Roelens, B., Chanes, J. D. L. H., Walczak, A. M., Coppey, M., & Dostatni, N. (2013). Live imaging of bicoid-dependent transcription in drosophila embryos. *Current biology*, 23(21), 2135–2139.

[15] Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., de Bruijn, I., Chooi, Y. H., Claesen, J., Coates, R. C., Cruz-Morales, P., Duddela, S., Düsterhus, S., Edwards, D. J., Fewer, D. P., Garg, N., Geiger, C., Gomez-Escribano, J. P., Greule, A., Hadjithomas, M., Haines, A. S., Helfrich, E. J. N., Hillwig, M. L., Ishida, K., Jones, A. C., Jones, C. S., Jungmann, K., Kegler, C., Kim, H. U., Kötter, P., Krug, D., Masschelein, J., Melnik, A. V., Mantovani, S. M., Monroe, E. A., Moore, M., Moss, N., Nützmann, H.-W., Pan, G., Pati, A., Petras, D., Reen, F. J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N. J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A. K., Balibar, C. J., Balskus, E. P., Barona-Gómez, F., Bechthold, A., Bode, H. B., Borriss, R., Brady, S. F., Brakhage, A. A., Caffrey, P., Cheng, Y.-Q., Clardy, J., Cox, R. J., De Mot, R., Donadio, S., Donia, M. S., van der Donk, W. A., Dorrestein, P. C., Doyle, S., Driessen, A. J. M., Ehling-Schulz, M., Entian, K.-D., Fischbach, M. A., Gerwick, L., Gerwick, W. H., Gross, H., Gust, B., Hertweck, C., Höfte, M., Jensen, S. E., Ju, J., Katz, L., Kaysser, L., Klassen, J. L., Keller, N. P., Kormanec, J., Kuipers, O. P., Kuzuyama, T., Kyrpides, N. C., Kwon, H.-J., Lautru, S., Lavigne, R., Lee, C. Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Méndez, C., Metsä-Ketelä, M., Micklefield, J., Mitchell, D. A., Moore, B. S., Moreira, L. M., Müller, R., Neilan, B. A., Nett, M., Nielsen, J., O'Gara, F., Oikawa, H., Osbourn, A., Osburne, M. S., Ostash, B., Payne, S. M., Pernodet, J.-L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J. M., Salas, J. A., Schmitt, E. K., Scott, B., Seipke, R. F., Shen, B., Sherman, D. H., Sivonen, K., Smanski, M. J., Sosio, M., Stegmann, E., Süssmuth, R. D., Tahlan, K., Thomas, C. M., Tang, Y., Truman, A. W., Viaud, M., Walton, J. D., Walsh, C. T., Weber, T., van Wezel, G. P., Wilkinson, B., Willey, J. M., Wohlleben, W., Wright, G. D., Ziemert, N., Zhang, C., Zotchev, S. B., Breitling, R., Takano, E., & Glöckner, F. O. (2015). Minimum information about a biosynthetic gene cluster. *Nature Chemical Biology*, 11(9), 625–631.

[16] Perry, M. W., Bothma, J. P., Luu, R. D., & Levine, M. (2012). Precision of hunchback expression in the drosophila embryo. *Current biology*, 22(23), 2247–2252.

[17] Pimmett, V. L., Dejean, M., Fernandez, C., Trullo, A., Bertrand, E., Radulescu, O., & Lagha, M. (2021). Quantitative imaging of transcription in living drosophila embryos reveals the impact of core promoter motifs on promoter state dynamics. *Nature communications*, 12(1), 4504.

[18] Valentino, M., Ortega, B. M., Ulrich, B., Doyle, D. A., Farnum, E. D., Joiner, D. A., Gavis, E. R., & Niepielko, M. G. (2022). Computational modeling offers new insight into drosophila germ granule development. *Biophysical journal*, 121(8), 1465–1482.

[19] Wang, M., Hu, Q., Lv, T., Wang, Y., Lan, Q., Xiang, R., Tu, Z., Wei, Y., Han, K., Shi, C., et al. (2022). High-resolution 3d spatiotemporal transcriptomic maps of developing drosophila embryos and larvae. *Developmental Cell*, 57(10), 1271–1283.