Development of Cell Trajectory Analysis for Revealing Interactions of Cancer Cells

福井大学審査学位論文［博士(工学)］

Development of Cell Trajectory Analysis for Revealing Interactions of Cancer Cells

がん細胞の相互作用の解明に向けた細胞軌跡の解析技術の開発

2023 年 3 月

Zhuohan Xin

INDEX

ABBREVIATIONS

ECM	Extracellular matrix
EMT	Epithelial-mesenchymal transition
FBS	Fetal bovine serum
FN	Fibronectin
MDS	Multi-dimensional scaling
MSC	Mean silhouette coefficient
PCA	Principal component analysis
PS	Polystyrene
TGF-β	Transforming growth factor β1
TGF (+)	EMT-induced mesenchymal cells by TGF-β
TGF (−)	EMT non-induced epithelial cells
t-SNE	t-distributed stochastic neighbor embedding
UMAP	Uniform manifold approximation and projection
WCSS	Within-cluster sum of squares

CHAPTER 1

GENERAL INTRODUCTION

Abstract

This chapter reviews the progress of research on cell migration, the interactions between cells when collective migration occurs including chemical signals, mechanical signals, and the role of cells in collective migration. These computational methods to investigate migration patterns in collective migration are overviewed.

1.1 Cell Migration

Cell migration is a fundamental act of cellular life activity, which is as important as proliferation, apoptosis, and functional differentiation. This thesis is aimed at clarifying the migration patterns of cell populations in order to understand cell-cell interactions. Cell migration is associated with tissue morphogenesis, immune surveillance, and cancer metastasis during development. Cell migration events *in vivo* involve the migration of single cells and therefore the subject of a large number of *in vitro* studies is dominated by single-cell migration, contributing to the study of various relevant cell behaviors.

Invasion of cancer cells can spread between tissues, which has profound implications for the spread of cancer. During this migration process, cells metastasis as single and multicellular cells through pathways such as the vasculature (Figure 1.1) [1]. The behavior of single-cell migration within tissues is critical to a range of physiological processes such as embryonic development and wound healing in addition to having an impact on tumor development. The dynamic migration process of a cell varies as it progresses through the tissue cycle. This involves cell division as well as the directionality and speed of movement through the tissue [2, 3]. The migration of tumor cells is different from that of cells in

normal tissue, with more intense physical interactions between cells and between cells and the Extracellular matrix (ECM) [4-6]. Tumor cell populations have a greater capacity to metastasize, with significantly reduced adhesion in their years, but at the same time are able to migrate as a population to other tissues.

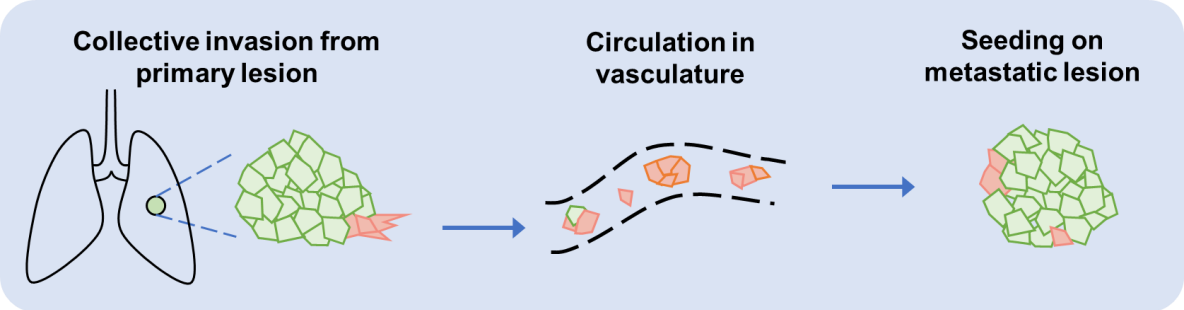


Figure 1.1 Schematic diagram of collective invasion of cancer cells from the primary lesion spread to other lesions through vasculature.

1.2 Collective Migration

The analysis of cell migration has been expanded by researchers with the observation that cancer cells often invade and migrate in collective populations as adherent groups. Collective cell migration is a process whereby a population of cells moves in concert through cell-cell communication and cell-environment interactions. A growing number of studies have shown that clusters of tumor cells acting collectively are more aggressive compared to single cells [7]. Interactions between cells have a long-term effect on cell migration patterns. Inter-cellular coordination keeps them adherent and moving faster and more efficiently overall [8]. A large number of studies have analyzed the interactions from a mechanical and molecular aspect. It requires a complement of different aspects to try to explain the mechanisms of collective cell migration. In this study, the interactions between cells in collective migration are elaborated in terms of cell culture, preparation of mimetic cell scaffolds, and analysis of cell migration trajectories.

Unlike single-cell migration, cells in collective cell migration are coordinated with each other [9-11]. In cell populations, leader cells regulate the interconnectedness of cells in the population through feedback mechanisms mediated by molecular and mechanical signals [12-14].

Typically, leader cells have distinct morphological and migratory behavior and have a role in leading collective invasion in collective migration. Leader cells can remodel the matrix in a variety of ways, such as through force-mediated and polarized assembly of fibronectin (FN) to deform and remodel the matrix [15-17]. The matrix remodeled by the leader cells leaves behind microscopic traces that can be exploited by less invasive follower cells [18-20], which collectively migrate to interact with each other on the basis of the microscopic traces to further expand the area of invasion. The establishment of polarization is due to the presence of two cell types. They differ in morphology and characteristics and are often referred to by researchers as leader and follower cells. Leader cells provide guidance for collective migration through intercellular interactions before other follower cells [21].

1.3 Cell Interactions in Collective Migration

In collective cell migration, adhesion junctions are formed between cells [22]. Collective behavior facilitates the invasion of a cell population into a new region by coordinating mechanosensitive adhesion junctions [23]. However, collective migration in both two-dimensional planes and three-dimensional structures involves interactions that include both force-mediated and molecular signals. The pathways such as PI3K-Rac signaling are involved in actin remodeling and mediating collective migration. Through PI3K-dependent integrin adhesion and modulation of Rho-GTPase signaling, Cadherin-induced regulation of actomyosin contractility at more distant sites of the cell affects global cellular mechanics [24, 25]. Many cells in collective cell migration are able to constantly reorient their movements. In a relatively crowded environment, contact inhibition of

locomotion [26, 27] arises, *i.e.*, cells move in separate directions after contact with each other. In addition, the speed and polarity of the cells change. Collective migration relies on the establishment of polarized cell populations that are asymmetric in their anterior-posterior aspect.

Interactions in collective cell migration include those between cells and those between cells and the matrix (Figure 1.2) [28]. The forces between cells and matrix usually consist of contraction forces driving the cells and adhesive forces between cells and matrix. The active contraction force is driven by the gradient generated by chemotaxis. This drives the actin filaments to produce contractile forces by the actin molecular motor. This corresponds to adhesive forces, which are mediated by the detachment of proteins at focal adhesions [29]. This adhesive force is related to the speed of cell movement relative to the substrate. Adhesive between cells and matrix is usually analyzed in terms of viscous damping forces similar to Stokes drag, and the dynamics of this movement are non-linear and associated with reinforcing feedback [30]. Adhesion between cells is mediated by transmembrane protein complexes and associated with the sliding of connexins [31-33]. Cells are resisted elastically by surrounding cells as they move [34, 35], which to some extent creates a squeeze on the cell causing a change in cell morphology. The pressure causes the cells to

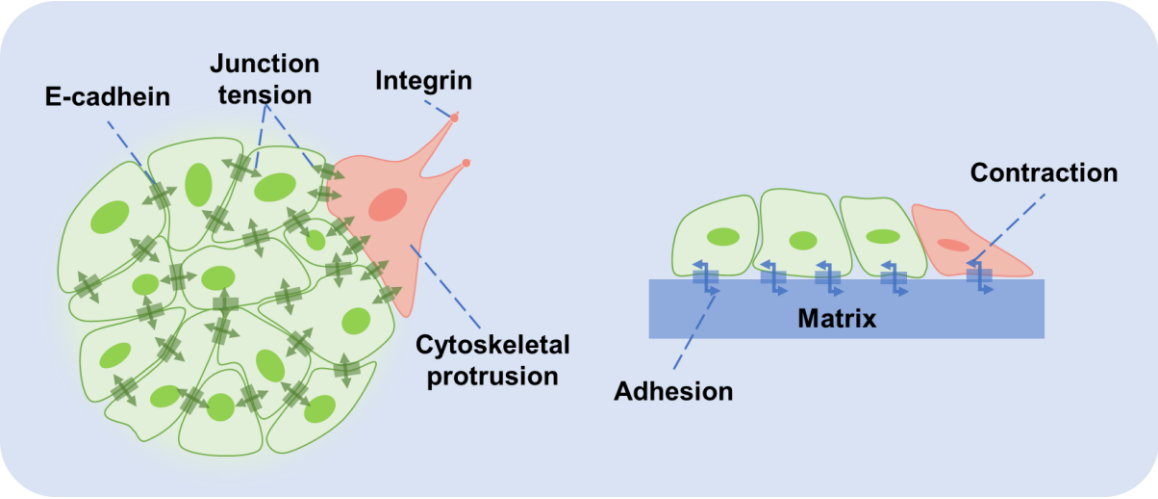


Figure 1.2 Schematic diagram of top and side views of force and cadherin in cell–cell and cell–matrix interactions.

extrude in their original position or insert in a new position [36, 37]. The presence of adhesion and friction allows cells to interact with each other in their movement, providing the impetus for collective migration.

1.4 EMT in Physiological Processes

Epithelial-mesenchymal transformation (EMT) refers to the morphological transformation in which epithelial cells lose their epithelial properties, transform into mesenchymal phenotype, and obtain migration ability (Figure 1.3) [38]. It causes epithelial cells generated at specific sites to detach from epithelial tissues and migrate to other locations. This is the basis of normal development, wound healing, and malignant epithelial tumors. During normal development and tumorigenesis, various stimuli of the microenvironment can induce EMT of epithelial cells through various different signaling pathways. EMT induces changes in cell function and phenotype, such as loss of cell polarity and changes in cell morphology. Invasiveness acquired through the regulation of microRNA-200 family and microRNA-205 genes, as well as the expression changes of transcription inhibitors ZEB1 and SIP1, may be an important step in tumor progression [39, 40].

In the early stages of cancer metastasis, cancer cells leave the primary tumor as individual cells and migrate through the effects of EMT. EMT involves a series of biological changes in which epithelial cells lose their epithelial properties and acquire

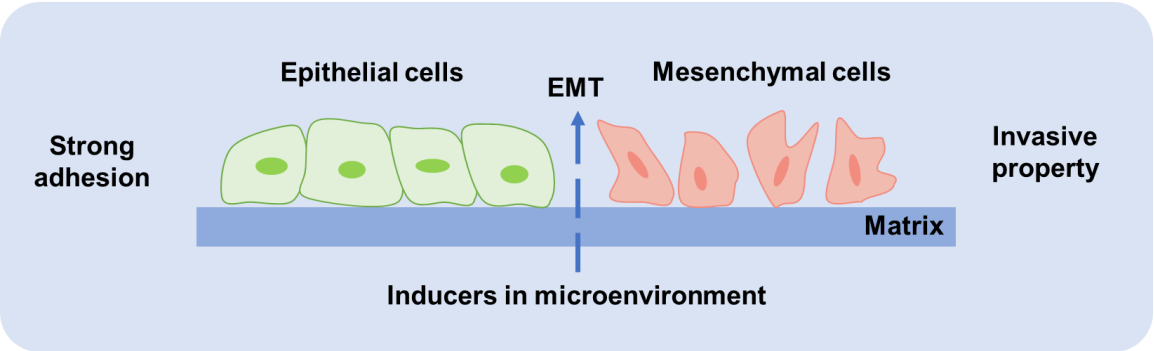


Figure 1.3 Schematic diagram of epithelial -mesenchymal transformation.

mesenchymal properties. When EMT is induced in cells, the cell function and characteristics are greatly altered; such changes include loss of cell polarity, altered cell morphology, and acquisition of invasive capacity, along with downregulation of epithelial cell genes and upregulation of mesenchymal cell genes [41, 42]. Transforming growth factor β1 (TGF-β) is a cell growth factor and representative EMT-inducing factor [43-45]. When TGF-β acts on epithelial cells, various transcriptional regulatory mechanisms are activated through the TGF-β signaling pathway. TGF-β induces EMT by inhibiting the transcription of intercellular adhesion molecules (*e.g.*, E-cadherin), decreasing genes characteristically expressed in epithelial cells, and increasing genes characteristically expressed in mesenchymal cells [46]. In EMT, epithelial and mesenchymal features are considered as binary "on/off." However, *in vitro* experiments showed that epithelial and mesenchymal markers are coexpressed in the same cells [47]. This observation suggests that EMT progressively develops in a state in which epithelial and mesenchymal properties are mixed [48]. Screening for cell surface markers in breast cancer showed that EMT exhibits a distinctive mixed phenotype and develops in a progressive pattern [49]. Furthermore, these mixed tumor cell subpopulations increase the metastatic potential *in vivo* [50].

This is considered to be a key mechanism for epithelial cancer cells to acquire metastatic phenotype through single-cell invasion. These experimental results support EMT induced cell migration in ECM, and fiber structure promotes colonization. Cells migrate through actin aggregation at the front edge of the cell population to produce lamellar cells. For adhesion, they are mainly generated by the molecular mechanism composed of integrins and related adhesion proteins [51-53]. E-cadherin is a major mediator of collective cell interactions [54, 55]. In morphogenesis and cancer models, the loss of E-cadherin is accompanied by the migration mode of EMT, leading to the weakening of cell connectivity, followed by cell separation and the increase of N-cadherin. The collective migration process of cells based on EMT has been studied to varying degrees through molecular

mechanisms and mechanism models [56-58]. More and more studies have shown that the relationship between EMT and cell collective migration is very complex in the process of tumor metastasis. Whether it is cell-cell or cell-matrix interaction, multiple factors will be integrated to affect migration mode [59]. Cell-to-cell interactions are strongly dependent on cadherin. Cadherin is a highly conserved calcium-dependent transmembrane protein that constitutes a major component of adherents' junctions. The expression levels of different types of cadherin are associated with the development of cancer. Epithelial (E-) and neuronal (N-) cadherin form intercellular adhesions. The intracellular cadherin domains connect to β- and α- catenin, that associate with the actin cytoskeleton to mediate mechanotransduction [60]. During EMT, downregulation of E-cadherin expression and an upregulation of N-cadherin expression is usually observed simultaneously. This conversion between cadherin involves a weakening of cell-to-cell junctions [61]. These interactions include molecular mechanisms and have been extensively described in conjunction with mechanistic models [56-58].

1.5 Collective Cell Migration Model *in Vitro*

The phenotype of migrating cells depends on the biochemical composition, stiffness, and overall morphology of the matrix. It is typical to distinguish between their dimensional systems. Migration in a one-dimensional system generally refers to cells migrating along individual collagen fibers, where the cells can only be oriented in line with the fibers. Cells migrating within the vessel or along the surface are migrating in two dimensions, in which case the direction of cell migration is largely not constrained by the orientation of individual fibers but can be more selective within the plane. When cells are enclosed in a matrix, *i.e.*, migrating in a three-dimensional environment, migration is more spatially possible [62]. In most studies, when it comes to interactions in cell migration, the two-dimensional plane is widely used because it possesses operability and ease of observation. Instruments and models for testing adhesion and mechanical distribution based on cell-to-cell interactions

have been widely studied. These models involve cell polarity, morphology and signaling pathways associated with migration [63].

PI3K, PTEN, and PtdIns(3,4,5)P_3 play a role in chemotaxis by controlling the actin skeleton of the cell to regulate cell motility. PtdIns(3,4,5)P_3 generated by PI3K plays a role through different downstream signal components, including GTPase RAC, ARF-GTPASES, and kinase Akt. Some pathways such as PI3K-PTEN and some parallel pathways interact to produce chemotaxis. This chemotaxis involves the production and degradation of PtdIns(3,4,5)P_3 and results in a net accumulation [24]. This net accumulation is at the leading edge of the cell, where it acts to promote the polymerization of actin and allows the cell to produce more pseudopods at the leading edge. Cells are able to undergo directed motility in response to the PtdIns(3,4,5)P_3 gradient in response to chemotaxis [64-66].

The generation of interactions is dependent on the regulation of signals from the microenvironment and therefore three-dimensional models are built to better match realistic physiological structures. The two-dimensional plane has a flat topography that facilitates the cells to be free from directional constraints. ECMs with a two-dimensional structure possess different mechanical properties. Cells exhibit very different migratory behavior towards different ECMs [67-69]. In addition to the stiffness and smoothness of the fiber bundles, which have a strong influence on cell spreading, other physicochemical properties such as composition and fiber density have different inducing effects on cell migration [70]. Higher porosity gives the possibility for cells to diffuse to a greater extent.

1.6 Biomimetic Nanofibrous Structures

Nowadays, electrospinning, phase separation, and molecular self-assembly technologies have been developed to manufacture nanofiber scaffolds. In particular, electrospun technology has been recognized as a more easy and adaptive method to produce ultra-thin fibers. Compared with other technologies, its diameter ranges from submicron to

nanometer fibers. Electrospun technology has been used as the main method to design biomaterial scaffolds. The electrospun fiber has high specific surface area and high porosity, which can imitate the local ECM structure and improve the effective response of cells (Figure 1.4). It is applicable to biomedical applications, including tissue engineering, catalysis, wound management, drug delivery, and filtration [71]. This convenient technology was first discovered by Rayleigh in 1897 [72]. Later, Taylor established the basic principle of electrospinning, which was inspired by the electrically driven ejector in 1969 [73]. Generally, electrospinning equipment uses a high-voltage power supply to generate high potential. Under the action of the electric field, the positively charged polymer solution is attracted to the opposite electrode, causing the solution (Taylor cone) sprayed in a straight line to form a fiber [74]. Recent electrospun fiber manufacturing methods include coaxial [75], side-by-side [76], triaxial [77], multi-fluid [78], and nanostructured fibers. These can not only create many types of structural surfaces but also obtain controllable nanofibers. The most commonly used synthetic polymers for electrospinning are linear aliphatic polyesters, including PLLA, PGA, and PCL. They can easily adjust the mechanical, architectural and degradation characteristics. However, most of them are hydrophobic and lack active binding sites for cell adhesion. Therefore, other modifications are required. Surface modification can be used to improve the physical,

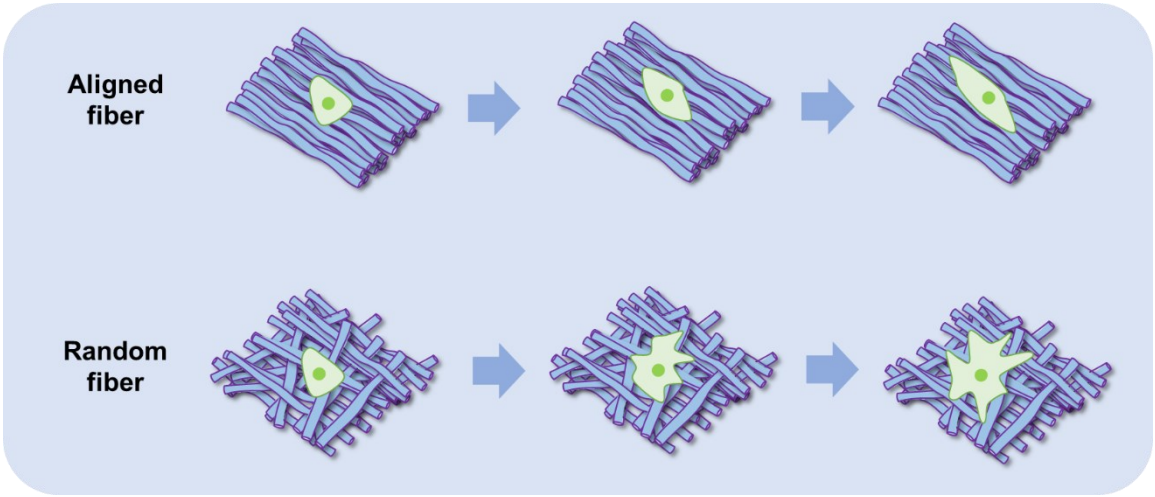


Figure 1.4 Cells migrate on aligned and random fibers that mimic the ECM.

chemical, and biological properties of nanofibers. Methods such as plasma modification, wet chemical method, and surface polymerization create more favorable microenvironment *in vivo* [79]. In addition, water-soluble compounds including polyvinylpyrrolidone [80], polyethylene glycol [81], and polyethylene oxide [82] are often used as drug carriers. They are ideal candidates for biomedical applications. Therefore, people have realized that it is an effective method to produce polymer hybrid nanofibers by mixing two or more polymers, which can maximize their respective advantages.

1.7 Clustering Methods in Biology

In biology, the goal of research often involves a large number of individuals. Each of these individuals has unique characteristics, but in further studies, individuals are closely related to each other. Finding individuals with similar characteristics in a group is helpful for collective events and large-scale analysis. Clustering methods can find similar characteristics in the group structure. In the application of clustering, whether the target groups are similar or not is my point of interest. In flow cytometry, each cell in the population is fluorescently labeled with markers, and when they are passed through a fixed wavelength laser, the cells in the population are excited with fluorescence and the intensity is recorded. The fluorescence intensity is not only related to the excitation wavelength, but most notably to the gene expression level. The observations consist of measurements from different channels, so the cells with different fluorescence intensities in different channels or cells with different gene expressions can be compared. A similar approach can help to identify genes with co-regulatory effects in order to distinguish disease marker genes. In clustering methods, how to define similarity and how many unused groups the samples should be divided into based on similarity are the most important questions. In statistics, the purpose of clustering is to divide a sample into subsets where samples within a subset are more similar than those in different subsets, which is the concept of clustering[83]. The combination of bioinformatics and computer vision offers a new approach to modern

applications of biology. The multi-view learning algorithm uses medical images for clustering analysis, allowing the results of clustering to be combined with views that are sample features more easily captured. The method has shown excellent performance in processing cancer datasets [84]. Object clustering methods on a subset of attributes were used to describe data from metabolomics analyses, and clustering effects based on pathology and intervention effect studies demonstrated better results compared to principal component analysis (PCA) [85]. Clustering is the partitioning of a data set into clusters according to a particular criterion (distance), such that the similarity of data objects within the same cluster is as great as possible, while the differences between data objects no longer within the same cluster are also as great as possible. Unsupervised learning in machine learning is involved here, where the labelling information of the training samples is unknown and the goal is to uncover the intrinsic properties, structure and information of the training samples to provide a basis for further data mining.

1.8 Time-series Clustering

A time-series (or dynamic series) is a series of values of the same statistical indicator arranged in the order of their occurrence in time. Time-series databases contain valuable information that can be obtained through pattern discovery, and clustering a common solution is used to reveal certain regular patterns in time-series datasets. Time-series data is often too large to understand in an intuitive way, and understanding time-series problems is made somewhat easier by time-series clustering, which divides the different time-series data and then analyses them separately. Time-series clustering is one of the most commonly used exploration techniques and is often one of the required processes for more complex data mining algorithms. Representing the cluster structure of a time-series as a visual image can help users quickly understand the data structure, clusters, anomalies, and other rules in a dataset (Figure1.5).

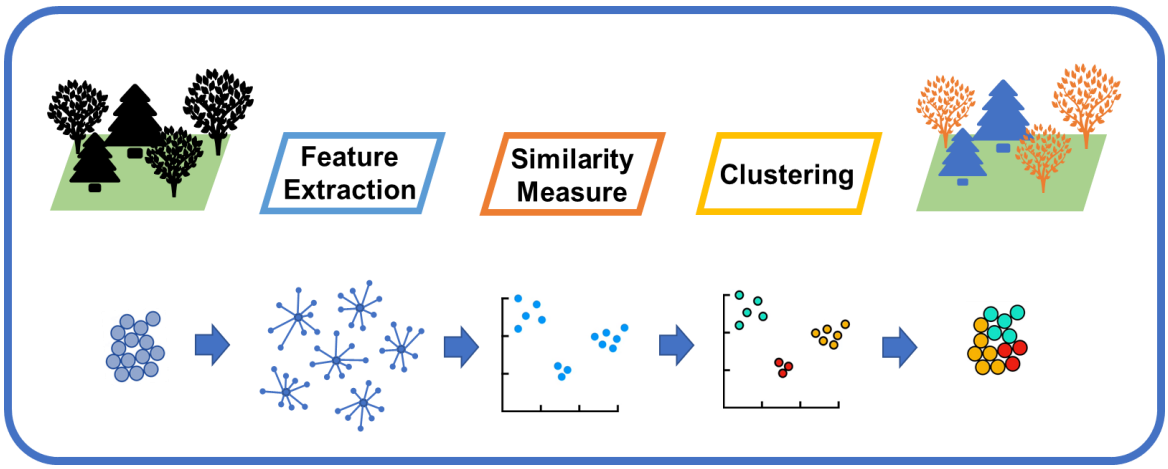


Figure 1.5 Schematic diagram of clustering method including original data collection, feature extraction of individuals, similarity measurement of feature, clustering by similarity.

Data analysis and mining technology is a combination of machine learning algorithms and data access technology, using machine learning to provide statistical analysis, knowledge discovery and other means to analyze large amounts of data, while using data access mechanisms to achieve efficient reading and writing of data [86]. Machine learning has an irreplaceable position in the field of data analysis and mining.

Pattern recognition originated in the field of engineering, while machine learning originated in computer science [87]. The combination of these two different disciplines has brought about the adaptation and development of the field of pattern recognition [88]. As genomic and other sequencing projects continue to evolve, the focus of bioinformatics research is gradually shifting from the accumulation of data to how to interpret that data. In the future, new discoveries in biology will rely heavily on our ability to combine and correlate diverse data in multiple dimensions and at different scales, rather than relying solely on a continued focus on traditional domains.

In research and applications, it is often necessary to observe data containing multiple variables, collect a large amount of data and then analyze it to find patterns. Large multivariate data sets undoubtedly provide a wealth of information for research and applications, but they also increase the workload of data collection to some extent. More

importantly, in many cases, many variables may be correlated with each other, thus increasing the complexity of the problem analysis. If each indicator is analyzed separately, the analysis is often isolated and cannot fully utilize the information in the data, so blindly reducing the indicators can lose a lot of useful information and thus generate wrong conclusions. Therefore, there is a need to find a reasonable method to reduce the number of indicators to be analyzed while minimizing the loss of information contained in the original indicators, in order to achieve a comprehensive analysis of the collected data. Since there are certain correlations among the variables, it is possible to consider turning the closely related variables into as few new variables as possible, so that these new variables are two unrelated, and then fewer integrated indicators can be used to represent the various types of information present in each variable, respectively. Dimensionality reduction is a method of pre-processing high-dimensional feature data, and is a very widely used data pre-processing method. The purpose of dimensionality reduction is to keep the most important features of high-dimensional data and remove the noise and unimportant features, so as to achieve the purpose of improving the speed of data processing. In practical production and application, dimensionality reduction can save me a lot of time and cost within a certain range of information loss.

1.9 Summary

Collective cell migration is a major mechanism of cancer metastasis. Cancer cells can spread between tissues. EMT is one of the bases of tumor metastasis. Many research supported that mesenchymal and epithelial cell act as leader and follower, respectively. To mimic the ECM, the microenvironment and structure are required. Topographical features of the ECM can be recreated by spinning polymers into fibers and depositing them as a thin layer on a surface. Cells can adhere and migration on it. To clarify the complexity of migration phenomena, many methodologies treat all cells together and from a colony aspect to get an overall impression of the migration pattern. It is more accuracy to analyze from a

single-cell aspect. The single-cell property can detail the interaction in collective migration.

In this study, the objective is from the aspect of cell trajectory to investigate the interaction between epithelial and mesenchymal cells. First, the electrospun fiber was used to mimic the ECM structure, epithelial and mesenchymal cells were seeded on fiber with different ratios. The cell positions and migration properties like directions and distance of cells were recorded to see the effect of mesenchymal cells. Second, to further investigate the relationship of epithelial and mesenchymal cells in collective migration, all cells were clustered using machine learning. From the time-lapse observation, the high-dimension time-series data was obtained. Then the dimension reduction and clustering algorithm were conducted. The epithelial and mesenchymal cells were clustered into different groups by similarity. At the last, the time-series clustering method was optimized by comparing parameters.

REFERENCE

[1] Li, L.; He, Y.; Zhao, M.; Jiang, J., Collective cell migration: Implications for wound healing and cancer invasion. *Burns Trauma* **2013,** 1, (1), 21-6. doi: 10.4103/2321-3868.113331.

[2] Bretscher, M. S., On the shape of migrating cells - a 'front-to-back' model. *J Cell Sci* **2008,** 121, (16), 2625-2628. doi: 10.1242/jcs.031120.

[3] Koppen, M.; Fernandez, B. G.; Carvalho, L.; Jacinto, A.; Heisenberg, C. P., Coordinated cell-shape changes control epithelial movement in zebrafish and Drosophila. *Development* **2006,** 133, (14), 2671-81. doi: 10.1242/dev.02439.

[4] Ravasio, A.; Le, A. P.; Saw, T. B.; Tarle, V.; Ong, H. T.; Bertocchi, C.; Mege, R. M.; Lim, C. T.; Gov, N. S.; Ladoux, B., Regulation of epithelial cell organization by tuning cell-substrate adhesion. *Integr Biol (Camb)* **2015,** 7, (10), 1228-41. doi: 10.1039/c5ib00196j.

[5] Ladoux, B.; Nicolas, A., Physically based principles of cell adhesion mechanosensitivity in tissues. *Rep Prog Phys* **2012,** 75, (11), 116601. doi: 10.1088/0034-4885/75/11/116601.

[6] Nier, V.; Jain, S.; Lim, C. T.; Ishihara, S.; Ladoux, B.; Marcq, P., Inference of Internal Stress in a Cell Monolayer. *Biophys J* **2016,** 110, (7), 1625-1635. doi: 10.1016/j.bpj.2016.03.002.

[7] Cheung, K. J.; Ewald, A. J., A collective route to metastasis: Seeding by tumor cell clusters. *Science* **2016,** 352, (6282), 167-169. doi: 10.1126/science.aaf6546.

[8] Collins, T. A.; Yeoman, B. M.; Katira, P., To lead or to herd: optimal strategies for 3D collective migration of cell clusters. *Biomech Model Mechan* **2020,** 19, (5), 1551-1564. doi: 10.1007/s10237-020-01290-y.

[9] Wu, Y.; Ali, M. R. K.; Dong, B.; Han, T.; Chen, K.; Chen, J.; Tang, Y.; Fang, N.; Wang, F.; El-Sayed, M. A., Gold Nanorod Photothermal Therapy Alters Cell Junctions and Actin Network in Inhibiting Cancer Cell Collective Migration. *Acs Nano* **2018,** 12, (9), 9279-9290. doi: 10.1021/acsnano.8b04128.

[10] Clark, A. G.; Vignjevic, D. M., Modes of cancer cell invasion and the role of the microenvironment. *Curr Opin Cell Biol* **2015,** 36, 13-22. doi: 10.1016/j.ceb.2015.06.004.

[11] Friedl, P.; Locker, J.; Sahai, E.; Segall, J. E., Classifying collective cancer cell invasion. *Nat Cell Biol* **2012,** 14, (8), 777-83. doi: 10.1038/ncb2548.

[12] Reffay, M.; Parrini, M. C.; Cochet-Escartin, O.; Ladoux, B.; Buguin, A.; Coscoy, S.; Amblard, F.; Camonis, J.; Silberzan, P., Interplay of RhoA and mechanical forces in collective cell migration driven by leader cells. *Nat Cell Biol* **2014,** 16, (3), 217-23. doi: 10.1038/ncb2917.

[13] Yang, Y.; Jamilpour, N.; Yao, B.; Dean, Z. S.; Riahi, R.; Wong, P. K., Probing Leader Cells in Endothelial Collective Migration by Plasma Lithography Geometric Confinement. *Sci Rep* **2016,** 6, 22707. doi: 10.1038/srep22707.

[14] Omelchenko, T.; Vasiliev, J. M.; Gelfand, I. M.; Feder, H. H.; Bonder, E. M., Rho-dependent formation of epithelial "leader" cells during wound healing. *Proc Natl Acad Sci U S A* **2003,** 100, (19), 10788-93. doi: 10.1073/pnas.1834401100.

[15] Attieh, Y.; Clark, A. G.; Grass, C.; Richon, S.; Pocard, M.; Mariani, P.; Elkhatib, N.; Betz, T.; Gurchenkov, B.; Vignjevic, D. M., Cancer-associated fibroblasts lead tumor invasion through integrin-beta 3-dependent fibronectin assembly. *J Cell Biol* **2017,** 216, (11), 3509-3520. doi: 10.1083/jcb.201702033.

[16] Erdogan, B.; Ao, M.; White, L. M.; Means, A. L.; Brewer, B. M.; Yang, L.; Washington, M. K.; Shi, C.; Franco, O. E.; Weaver, A. M.; Hayward, S. W.; Li, D.; Webb, D. J., Cancer-associated fibroblasts promote directional cancer cell migration by aligning fibronectin. *J Cell Biol* **2017,** 216, (11), 3799-3816. doi: 10.1083/jcb.201704053.

[17] Glentis, A.; Oertle, P.; Mariani, P.; Chikina, A.; El Marjou, F.; Attieh, Y.; Zaccarini, F.; Lae, M.; Loew, D.; Dingli, F.; Sirven, P.; Schoumacher, M.; Gurchenkov, B. G.; Plodinec, M.; Vignjevic, D. M., Cancer-associated fibroblasts induce metalloprotease-independent cancer cell invasion of the basement membrane. *Nat Commun* **2017,** 8, (1), 924. doi: 10.1038/s41467-017-00985-8.

[18] Wolf, K.; Wu, Y. I.; Liu, Y.; Geiger, J.; Tam, E.; Overall, C.; Stack, M. S.; Friedl, P., Multi-step pericellular proteolysis controls the transition from individual to collective cancer cell invasion. *Nat Cell Biol* **2007,** 9, (8), 893-904. doi: 10.1038/ncb1616.

[19] Gaggioli, C.; Hooper, S.; Hidalgo-Carcedo, C.; Grosse, R.; Marshall, J. F.; Harrington, K.; Sahai, E., Fibroblast-led collective invasion of carcinoma cells with differing roles for RhoGTPases in leading and following cells. *Nat Cell Biol* **2007,** 9, (12), 1392-400. doi: 10.1038/ncb1658.

[20] Carey, S. P.; Starchenko, A.; McGregor, A. L.; Reinhart-King, C. A., Leading malignant cells initiate collective epithelial cell invasion in a three-dimensional heterotypic tumor spheroid model. *Clin Exp Metastasis* **2013,** 30, (5), 615-30. doi: 10.1007/s10585-013-9565-x.

[21] Khalil, A. A.; de Rooij, J., Cadherin mechanotransduction in leader-follower cell specification during collective migration. *Exp Cell Res* **2019,** 376, (1), 86-91. doi: 10.1016/j.yexcr.2019.01.006.

[22] Ladoux, B.; Mege, R. M., Mechanobiology of collective cell behaviours. *Nat Rev Mol Cell Biol* **2017,** 18, (12), 743-757. doi: 10.1038/nrm.2017.98.

[23] Etienne-Manneville, S., Adherens junctions during cell migration. *Subcell Biochem* **2012,** 60, 225-49. doi: 10.1007/978-94-007-4186-7_10.

[24] Kolsch, V.; Charest, P. G.; Firtel, R. A., The regulation of cell motility and chemotaxis by phospholipid signaling. *J Cell Sci* **2008,** 121, (Pt 5), 551-9. doi: 10.1242/jcs.023333.

[25] Khalil, A. A.; de Rooij, J., Cadherin mechanotransduction in leader-follower cell specification during collective migration. *Exp Cell Res* **2019,** 376, (1), 86-91. doi: 10.1016/j.yexcr.2019.01.006.

[26] Carmona-Fontaine, C.; Matthews, H. K.; Kuriyama, S.; Moreno, M.; Dunn, G. A.; Parsons, M.; Stern, C. D.; Mayor, R., Contact inhibition of locomotion in vivo controls neural crest directional migration. *Nature* **2008,** 456, (7224), 957-61. doi: 10.1038/nature07441.

[27] Abercrombie, M.; Heaysman, J. E., Observations on the social behaviour of cells in tissue culture. II. Monolayering of fibroblasts. *Exp Cell Res* **1954,** 6, (2), 293-306. doi: 10.1016/0014-4827(54)90176-7.

[28] Alert, R.; Trepat, X., Physical Models of Collective Cell Migration. *Annu Rev Conden Ma P* **2020,** 11, 77-101. doi: 10.1146/annurev-conmatphys-031218-013516.

[29] Schwarz, U. S.; Safran, S. A., Physics of adherent cells. *Rev Mod Phys* **2013,** 85, (3), 1327-1381. doi: 10.1103/RevModPhys.85.1327.

[30] Elosegui-Artola, A.; Trepat, X.; Roca-Cusachs, P., Control of Mechanotransduction by Molecular Clutch Dynamics. *Trends Cell Biol* **2018,** 28, (5), 356-367. doi: 10.1016/j.tcb.2018.01.008.

[31] Garcia, S.; Hannezo, E.; Elgeti, J.; Joanny, J. F.; Silberzan, P.; Gov, N. S., Physics of active jamming during collective cellular motion in a monolayer. *Proc Natl Acad Sci U S A* **2015,** 112, (50), 15314-9. doi: 10.1073/pnas.1510973112.

[32] Peglion, F.; Llense, F.; Etienne-Manneville, S., Adherens junction treadmilling during collective migration. *Nat Cell Biol* **2014,** 16, (7), 639-51. doi: 10.1038/ncb2985.

[33] Czirok, A.; Varga, K.; Mehes, E.; Szabo, A., Collective cell streams in epithelial monolayers depend on cell adhesion. *New J Phys* **2013,** 15. doi: 10.1088/1367-2630/15/7/075006.

[34] Khalilgharibi, N.; Fouchard, J.; Recho, P.; Charras, G.; Kabla, A., The dynamic mechanical properties of cellularised aggregates. *Curr Opin Cell Biol* **2016,** 42, 113-120. doi: 10.1016/j.ceb.2016.06.003.

[35] Xi, W.; Saw, T. B.; Delacour, D.; Lim, C. T.; Ladoux, B., Material approaches to active tissue mechanics. *Nat Rev Mater* **2019,** 4, (1), 23-44. doi: 10.1038/s41578-018-0066-z.

[36] Beaune, G.; Stirbat, T. V.; Khalifat, N.; Cochet-Escartin, O.; Garcia, S.; Gurchenkov, V. V.; Murrell, M. P.; Dufour, S.; Cuvelier, D.; Brochard-Wyart, F., How cells flow in the spreading of cellular aggregates. *Proc Natl Acad Sci USA* **2014,** 111, (22), 8055-60. doi: 10.1073/pnas.1323788111.

[37] Kocgozlu, L.; Saw, T. B.; Le, A. P.; Yow, I.; Shagirov, M.; Wong, E.; Mege, R. M.; Lim, C. T.; Toyama, Y.; Ladoux, B., Epithelial Cell Packing Induces Distinct Modes of Cell Extrusions. *Curr Biol* **2016,** 26, (21), 2942-2950. doi: 10.1016/j.cub.2016.08.057.

[38] Saenz-de-Santa-Maria, I.; Celada, L.; Chiara, M. D., The Leader Position of Mesenchymal Cells Expressing N-Cadherin in the Collective Migration of Epithelial Cancer. *Cells-Basel* **2020,** 9, (3). doi: 10.3390/cells9030731.

[39] Burk, U.; Schubert, J.; Wellner, U.; Schmalhofer, O.; Vincan, E.; Spaderna, S.; Brabletz, T., A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. *Embo Rep* **2008,** 9, (6), 582-589. doi: 10.1038/embor.2008.74.

[40] Gregory, P. A.; Bert, A. G.; Paterson, E. L.; Barry, S. C.; Tsykin, A.; Farshid, G.; Vadas, M. A.; Khew-Goodall, Y.; Goodall, G. J., The mir-200 family and mir-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol* **2008,** 10, (5), 593-601. doi: 10.1038/ncb1722.

[41] Gregory, P. A.; Bert, A. G.; Paterson, E. L.; Barry, S. C.; Tsykin, A.; Farshid, G.; Vadas, M. A.; Khew-Goodall, Y.; Goodall, G. J., The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol* **2008,** 10, (5), 593-601. doi: 10.1038/ncb1722.

[42] Burk, U.; Schubert, J.; Wellner, U.; Schmalhofer, O.; Vincan, E.; Spaderna, S.; Brabletz, T., A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. *Embo Rep* **2008,** 9, (6), 582-9. doi: 10.1038/embor.2008.74.

[43] Xu, J.; Lamouille, S.; Derynck, R., TGF-beta-induced epithelial to mesenchymal transition. *Cell Res* **2009,** 19, (2), 156-72. doi: 10.1038/cr.2009.5.

[44] Walker, E. J.; Heydet, D.; Veldre, T.; Ghildyal, R., Transcriptomic changes during TGF-beta-mediated differentiation of airway fibroblasts to myofibroblasts. *Sci Rep* **2019,** 9, (1), 20377. doi: 10.1038/s41598-019-56955-1.

[45] Cao, Q.; Deji, Q. Z.; Liu, Y. J.; Ye, W.; Zhaba, W. D.; Jiang, Q.; Yan, F., The role of mechanical stretch and TGF-beta 2 in epithelial-mesenchymal transition of retinal pigment epithelial cells. *Int J Ophthalmol-Chi* **2019,** 12, (12), 1832-1838. doi: 10.18240/ijo.2019.12.03.

[46] Plou, J.; Juste-Lanas, Y.; Olivares, V.; Del Amo, C.; Borau, C.; Garcia-Aznar, J. M., From individual to collective 3D cancer dissemination: roles of collagen concentration and TGF-beta. *Sci Rep* **2018,** 8, (1), 12723. doi: 10.1038/s41598-018-30683-4.

[47] Zavadil, J.; Bottinger, E. P., TGF-beta and epithelial-to-mesenchymal transitions. *Oncogene* **2005,** 24, (37), 5764-74. doi: 10.1038/sj.onc.1208927.

[48] Zhang, J. Y.; Tian, X. J.; Zhang, H.; Teng, Y.; Li, R. Y.; Bai, F.; Elankumaran, S.; Xing, J. H., TGF-beta-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci Signal* **2014,** 7, (345). doi: 10.1126/scisignal.2005304.

[49] Pastushenko, I.; Brisebarre, A.; Sifrim, A.; Fioramonti, M.; Revenco, T.; Boumahdi, S.; Van Keymeulen, A.; Brown, D.; Moers, V.; Lemaire, S.; De Clercq, S.; Minguijon, E.; Balsat, C.; Sokolow, Y.; Dubois, C.; De Cock, F.; Scozzaro, S.; Sopena, F.; Lanas, A.; D'Haene, N.; Salmon, I.; Marine, J. C.; Voet, T.; Sotiropoulou, P. A.; Blanpain, C., Identification of the tumour transition states occurring during EMT. *Nature* **2018,** 556, (7702), 463-468. doi: 10.1038/s41586-018-0040-3.

[50] Huang, R. Y.; Wong, M. K.; Tan, T. Z.; Kuay, K. T.; Ng, A. H.; Chung, V. Y.; Chu, Y. S.; Matsumura, N.; Lai, H. C.; Lee, Y. F.; Sim, W. J.; Chai, C.; Pietschmann, E.; Mori, S.; Low, J. J.; Choolani, M.; Thiery, J. P., An EMT spectrum defines an anoikis-resistant and spheroidogenic intermediate mesenchymal state

that is sensitive to e-cadherin restoration by a src-kinase inhibitor, saracatinib (AZD0530). *Cell Death Dis* **2013,** 4, e915. doi: 10.1038/cddis.2013.442.

[51] Caswell, P. T.; Zech, T., Actin-Basec Cell Protrusion in a 3D Matrix. *Trends Cell Biol* **2018,** 28, (10), 823-834. doi: 10.1016/j.tcb.2018.06.003.

[52] Doyle, A. D.; Wang, F. W.; Matsumoto, K.; Yamada, K. M., One-dimensional topography underlies three-dimensional fibrillar cell migration. *J Cell Biol* **2009,** 184, (4), 481-490. doi: 10.1083/jcb.200810041.

[53] Le Borgne-Rochet, M.; Angevin, L.; Bazellieres, E.; Ordas, L.; Comunale, F.; Denisov, E. V.; Tashireva, L. A.; Perelmuter, V. M.; Bieche, I.; Vacher, S.; Plutoni, C.; Seveno, M.; Bodin, S.; Gauthier-Rouviere, C., P-cadherin-induced decorin secretion is required for collagen fiber alignment and directional collective cell migration. *J Cell Sci* **2019,** 132, (21). doi: 10.1242/jcs.233189.

[54] Thompson, E. W.; Williams, E. D., EMT and MET in carcinoma - clinical observations, regulatory pathways and new models. *Clin Exp Metastas* **2008,** 25, (6), 591-592. doi: 10.1007/s10585-008-9189-8.

[55] Grunert, S.; Jechlinger, M.; Beug, H., Diverse cellular and molecular mechanisms contribute to epithelial plasticity and metastasis. *Nat Rev Mol Cell Bio* **2003,** 4, (8), 657-665. doi: 10.1038/nrm1175.

[56] Kim, J. M.; Lee, M.; Kim, N.; Do Heo, W., Optogenetic toolkit reveals the role of Ca2+ sparklets in coordinated cell migration. *Proc Natl Acad Sci USA* **2016,** 113, (21), 5952-5957. doi: 10.1073/pnas.1518412113.

[57] Becsky, D.; Szabo, K.; Gyulai-Nagy, S.; Gajdos, T.; Bartos, Z.; Balind, A.; Dux, L.; Horvath, P.; Erdelyi, M.; Homolya, L.; Keller-Pinter, A., Syndecan-4 Modulates Cell Polarity and Migration by Influencing Centrosome Positioning and Intracellular Calcium Distribution. *Neuromuscular Disord* **2020,** 30, S170-S171. doi: 10.1016/j.nmd.2020.09.020.

[58] Morrison, J. A.; McLennan, R.; Wolfe, L. A.; Gogol, M. M.; Meier, S.; McKinney, M. C.; Teddy, J. M.; Holmes, L.; Semerad, C. L.; Box, A. C.; Li, H.; Hall, K. E.; Perera, A. G.; Kulesa, P. M., Single-cell transcriptome analysis of avian neural crest migration reveals signatures of invasion and molecular transitions. *Elife* **2017,** 6. doi: 10.7554/eLife.28415.

[59] Capuana, L.; Bostrom, A.; Etienne-Manneville, S., Multicellular scale front-to-rear polarity in collective migration. *Curr Opin Cell Biol* **2020,** 62, 114-122. doi: 10.1016/j.ceb.2019.10.001.

[60] Shapiro, L.; Weis, W. I., Structure and biochemistry of cadherins and catenins. *Cold Spring Harb Perspect Biol* **2009,** 1, (3), a003053. doi: 10.1101/cshperspect.a003053.

[61] Campbell, K.; Casanova, J., A common framework for EMT and collective cell migration. *Development* **2016,** 143, (23), 4291-4300. doi: 10.1242/dev.139071.

[62] Stuelten, C. H.; Parent, C. A.; Montell, D. J., Cell motility in cancer invasion and metastasis: insights from simple model organisms. *Nat Rev Cancer* **2018,** 18, (5), 296-312. doi: 10.1038/nrc.2018.15.

[63] Jain, S.; Cachoux, V. M. L.; Narayana, G. H. N. S.; de Beco, S.; D'Alessandro, J.; Cellerin, V.; Chen, T. C.; Heuze, M. L.; Marcq, P.; Mege, R. M.; Kabla, A. J.; Lim, C. T.; Ladoux, B., The role of single-cell mechanical behaviour and polarity in driving collective cell migration. *Nat Phys* **2020,** 16, (7), 802-+. doi: 10.1038/s41567-020-0875-z.

[64] Iijima, M.; Devreotes, P., Tumor suppressor PTEN mediates sensing of chemoattractant gradients. *Cell* **2002,** 109, (5), 599-610. doi: 10.1016/s0092-8674(02)00745-6.

[65] Huang, Y. E.; Iijima, M.; Parent, C. A.; Funamoto, S.; Firtel, R. A.; Devreotes, P., Receptor-mediated regulation of PI3Ks confines PI(3,4,5)P3 to the leading edge of chemotaxing cells. *Mol Biol Cell* **2003,** 14, (5), 1913-22. doi: 10.1091/mbc.e02-10-0703.

[66] Funamoto, S.; Milan, K.; Meili, R.; Firtel, R. A., Role of phosphatidylinositol 3' kinase and a downstream pleckstrin homology domain-containing protein in controlling chemotaxis in dictyostelium. *J Cell Biol* **2001,** 153, (4), 795-810. doi: 10.1083/jcb.153.4.795.

[67] Kechagia, J. Z.; Ivaska, J.; Roca-Cusachs, P., Integrins as biomechanical sensors of the microenvironment. *Nat Rev Mol Cell Bio* **2019,** 20, (8), 457-473. doi: 10.1038/s41580-019-0134-2.

[68] Chen, B.; Ji, B.; Gao, H., Modeling Active Mechanosensing in Cell-Matrix Interactions. *Annu Rev Biophys* **2015,** 44, 1-32. doi: 10.1146/annurev-biophys-051013-023102.

[69] van Helvert, S.; Storm, C.; Friedl, P., Mechanoreciprocity in cell migration. *Nat Cell Biol* **2018,** 20, (1), 8-20. doi: 10.1038/s41556-017-0012-0.

[70] Artym, V. V.; Swatkoski, S.; Matsumoto, K.; Campbell, C. B.; Petrie, R. J.; Dimitriadis, E. K.; Li, X.; Mueller, S. C.; Bugge, T. H.; Gucek, M.; Yamada, K. M., Dense fibrillar collagen is a potent inducer of invadopodia via a specific signaling network. *J Cell Biol* **2015,** 208, (3), 331-50. doi: 10.1083/jcb.201405099.

[71] Buzgo, M.; Mickova, A.; Rampichova, M.; Doupnik, M., Blend electrospinning, coaxial electrospinning, and emulsion electrospinning techniques. In *Core-Shell Nanostructures for Drug Delivery and Theranostics*, Focarete, M. L.; Tampieri, A., Eds. Woodhead Publishing: **2018**; pp 325-347.

[72] Zheng, Y., Fabrication on bioinspired surfaces. In *Bioinspired Design of Materials Surfaces*, Zheng, Y., Ed. Elsevier: **2019**; pp 99-146.

[73] Taylor, G. I.; Van Dyke, M. D., Electrically driven jets. *Proc R Soc Lond A Math Phys Sci* **1997,** 313, (1515), 453-475. doi: 10.1098/rspa.1969.0205.

[74] Lukáš, D.; Sarkar, A.; Martinová, L.; Vodsed'álková, K.; Lubasová, D.; Chaloupek, J.; Pokorný, P.; Mikeš, P.; Chvojka, J.; Komárek, M., Physical principles of electrospinning (Electrospinning as a nano-scale technology of the twenty-first century). *Textile Progress* **2009,** 41, (2), 59-140. doi: 10.1080/00405160902904641.

[75] Huang, W.-Y.; Hibino, T.; Suye, S.-I.; Fujita, S., Electrospun collagen core/poly-l-lactic acid shell nanofibers for prolonged release of hydrophilic drug. *RSC Adv* **2021,** 11, (10), 5703-5711. doi: 10.1039/d0ra08353d.

[76] Wang, K.; Liu, X. K.; Chen, X. H.; Yu, D. G.; Yang, Y. Y.; Liu, P., Electrospun Hydrophilic Janus Nanocomposites for the Rapid Onset of Therapeutic Action of Helicid. *ACS Appl Mater Interfaces* **2018,** 10, (3), 2859-2867. doi: 10.1021/acsami.7b17663.

[77] Huang, C. K.; Zhang, K.; Gong, Q.; Yu, D. G.; Wang, J.; Tan, X.; Quan, H., Ethylcellulose-based drug nano depots fabricated using a modified triaxial electrospinning. *Int J Biol Macromol* **2020,** 152, 68-76. doi: 10.1016/j.ijbiomac.2020.02.239.

[78] Chang, S. Y.; Wang, M. L.; Zhang, F. Y.; Liu, Y. B.; Liu, X. K.; Yu, D. G.; Shen, H., Sheath-separate-core nanocomposites fabricated using a trifluid electrospinning. *Mater Design* **2020,** 192. doi: 10.1016/j.matdes.2020.108782.

[79] He, C.; Nie, W.; Feng, W., Engineering of biomimetic nanofibrous matrices for drug delivery and tissue engineering. *J Mater Chem B* **2014,** 2, (45), 7828-7848. doi: 10.1039/c4tb01464b.

[80] Jin, W. J.; Lee, H. K.; Jeong, E. H.; Park, W. H.; Youk, J. H., Preparation of polymer nanofibers containing silver nanoparticles by using poly (N-vinylpyrrolidone). *Macromol Rapid Commun* **2005,** 26, (24), 1903-1907. doi: 10.1002/marc.200500569.

[81] Bagheri, H.; Najarzadekan, H.; Roostaie, A., Electrospun polyamide-polyethylene glycol nanofibers for headspace solid-phase microextration. *J Sep Sci* **2014,** 37, (14), 1880-6. doi: 10.1002/jssc.201400037.

[82] Ma, G.; Fang, D.; Liu, Y.; Zhu, X.; Nie, J., Electrospun sodium alginate/poly(ethylene oxide) core-shell nanofibers scaffolds potential for tissue engineering applications. *Carbohydr Polym* **2012,** 87, (1), 737-743. doi: 10.1016/j.carbpol.2011.08.055.

[83] Nugent, R.; Meila, M., An overview of clustering applied to molecular biology. *Methods Mol Biol* **2010,** 620, 369-404. doi: 10.1007/978-1-60761-580-4_12.

[84] Gönen, M.; Margolin, A. A. In *Localized Data Fusion for Kernel k-Means Clustering with Application to Cancer Biology*, Advances in Neural Information Processing Systems, **2014**, Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; Weinberger, K. Q., Eds. Curran Associates, Inc.

[85] Damian, D.; Oresic, M.; Verheij, E.; Meulman, J.; Friedman, J.; Adourian, A.; Morel, N.; Smilde, A.; van der Greef, J., Applications of a new subspace clustering algorithm (COSA) in medical systems biology. *Metabolomics* **2007,** 3, (1), 69-77. doi: 10.1007/s11306-006-0045-z.

[86] Yu, X. Q.; Han, Y. L.; Liu, S. M.; Jiang, W.; Song, Y. C.; Tong, J. Y.; Qiao, T. T.; Lv, Z. W.; Li, D., Analysis of Genetic Alterations Related to DNA Methylation in Testicular Germ Cell Tumors Based on Data Mining. *Cytogenet Genome Res* **2021,** 161, (6-7), 382-393. doi: 10.1159/000516385.

[87] Dimopoulos, A. C.; Nikolaidou, M.; Caballero, F. F.; Engchuan, W.; Sanchez-Niubo, A.; Arndt, H.; Ayuso-Mateos, J. L.; Haro, J. M.; Chatterji, S.; Georgousopoulou, E. N.; Pitsavos, C.; Panagiotakos, D. B., Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *Bmc Med Res Methodol* **2018,** 18, (1), 179. doi: 10.1186/s12874-018-0644-1.

[88] Bulgarevich, D. S.; Tsukamoto, S.; Kasuya, T.; Demura, M.; Watanabe, M., Pattern recognition with machine learning on optical microscopy images of typical metallurgical microstructures. *Sci Rep* **2018,** 8, (1), 2078. doi: 10.1038/s41598-018-20438-6.

CHAPTER 2

QUANTITATIVE ANALYSIS OF COLLECTIVE MIGRATION BY SINGLE-CELL TRACKING

Abstract

This chapter presents the quantitative analysis of cell migration properties. Electrospun fibers were used to mimic the ECM structure. Different proportions of mesenchymal and epithelial cells were seeded on the fibers. The analysis was performed by time-lapse observation combined with manual tracking of cell trajectories.

2.1 Introduction

The basic mechanisms of cell migration have been extensively studied [1]. Cell migration is based on the establishment of a front-to-rear polarity axis involving rearrangement of the cytoskeleton and polarization of the membranes. Underlying this front-to-rear polarity is front-to-rear polarization of the Rho family signaling cascade, whereby Rac1 and CDC42 induce rapid cytoskeletal rearrangements at the front of a cell [2]. This leads to the formation of membrane-like protrusions such as filopodia and lamellipodia. Adhesion between the cell adhesion protein integrin and the ECM is promoted, leading to forward migration of cells [3]. However, during collective migration, cells invade the ECM while maintaining E-cadherin–dependent cell adhesion. During collective migration, the migration mechanism of individual cells occurs for each cell in the population. In addition, there is a leader cell with a highly invasive and ECM remodeling capacity in collective migration and subsequent group

of follower cells [4-6]. The leader and follower cells are defined based only on their relative positions in the cluster and are located at the front and rear of the cluster, respectively. Leader cells are highly invasive and play important roles in ECM remodeling during migration [7-10]. The diverse mixed phenotype is involved in regulating collective migration and forming a “leader–follower” structure at different stages of EMT [11]. These results suggest that EMT-induced cells behave as “leaders,” exhibiting migratory behavior along fibrous structures in the ECM. However, the interactions between “leader” and “follower” cells and role of migration enhancement in metastasis remain unclear.

The microenvironment surrounding the tissues plays an important role in maintaining normal cellular behavior. The microenvironment varies between normal tissues and tumors, suggesting that cancer development and metastasis are influenced by the surrounding microenvironment [12-15]. Classical cell migration assays, such as chemotaxis assays [16, 17] involve the addition of chemokines to the culture environment to induce cell migration according to a concentration gradient. Wound healing assays [18-20] based on scratch assays can be used to evaluate cell migration properties by measuring tissue matrix build-up and associated cell differences in healing. In chemotaxis assays, environments with fixed concentration gradients are uncommon in cancer cell migration *in vivo*; the manner in which cells migrate in scratch assays differs from that in which cancer cells migrate collectively in a 3D environment. Traditional assays cannot adequately track cell population migration, supporting the necessity of constructing cancer cell migration models that simulate the *in vivo* environment. A recently established migration evaluation system mimics the *in vivo* cellular environment. In this system, cells migrate on flat culture dishes coated with FN or ECM gels (*e.g.*, collagen) present in the ECM [21]. Cells present in the ECM in a fibrous structure have an elongated morphology *in vivo*. Therefore, in flat culture dishes or gels without anisotropy, cells have a different morphology and applied tension

compared to their original morphology. The extension and migration directions of pseudopods, which are important for cell migration, may greatly differ from the original morphology [22]. Nanofiber materials fabricated by electrospinning have attracted considerable attention. Because nanofibers mimic fibrous and anisotropic structures *in vivo*, they are expected to be used as scaffolds in regenerative medicine [23-26].

This study focused on leader–follower cell interactions. To quantitatively analyze these interactions and reproducibly observe cell migration on fibers, a method to mimic collective migration were designed by co-culturing TGF-β1-induced EMT mesenchymal cells [TGF (+)] with EMT-negative cells [TGF (−)]. The migration capacity of mimicked cell populations was evaluated. The populations of TGF (+) and TGF (−) cells at different ratios were generated and compared their migratory behaviors. The morphology, trajectory, migration velocity, straightness, and directional angle were evaluated to quantitatively examine the migration of cell populations and effect of TGF (+) on cell migration.

2.2 Materials and Methods

2.2.1 Electrospinning

Tetrahydrofuran was used as the solvent in the electrospun polymer solution to prepare 20 wt % Polystyrene (PS). The PS solution was electrospun into aligned nanofiber structures using a commercial electrospinning setup (NANON, MECC, Fukuoka, Japan). The polymer solution in the syringe was ejected from the needle at a constant flow rate. A high voltage was applied to the needle, and the charged polymer solution was collected using a grounded rotating collector. During electrospinning, the flow rate was 0.1 mL/h and collector speed was 15.7 m/s. The electric field was 2.5 kV/cm. The fibers were treated with O_2 plasma (40 kHz/100 W, 25 Pa, 30 s) and coated with 10 μg/mL FN (37 °C, 2 h).

2.2.2 Cell Culture

NMuMG cells (ATCC CRL-1636) and fluorescently labeled NMuMG cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS). Fluorescent labeling of NMuMG was performed by induction of the pDsRed2-C1 vector. All experiments were approved by the ethics committee of the institution. The medium used to induce EMT in NMuMG–DsRed contained 10 ng/mL TGF-β1 (Sigma) and was cultured for 3 days. Unlabeled NMuMG cells were designated as TGF (−) and EMT-induced NMuMG-DsRed cells were designated as TGF (+). Cell aggregates were prepared by suspending the two cell types at different ratios (1×10^5 cells/mL). The cell suspension (500 µL) was seeded into 24-well grid plates (Elplasia #4445, Kuraray, Tokyo, Japan) for 3D culture and incubated at 37 °C, 5% CO_2, and 95% humidity for 3 h. FluoroBrite DMEM (Gibco) containing 10% FBS was used for fluorescence observation (Figure 2.1). After seeding, cell aggregates formed in the wells. These aggregates were collected and suspended in 500 µL of medium containing 25 µL of CellLight™ Histone 2B-GFP and BacMam 2.0 (Thermo Fisher Scientific K.K., Tokyo, Japan). The aggregates were designated as TGF (*N*), which represented the percentage of TGF (+) cells (N%). After 24 h of pre-incubation on the fiber sheets, time-lapse images were acquired using a Biostation (Nikon, Tokyo, Japan) at 15-min intervals for 24 h.

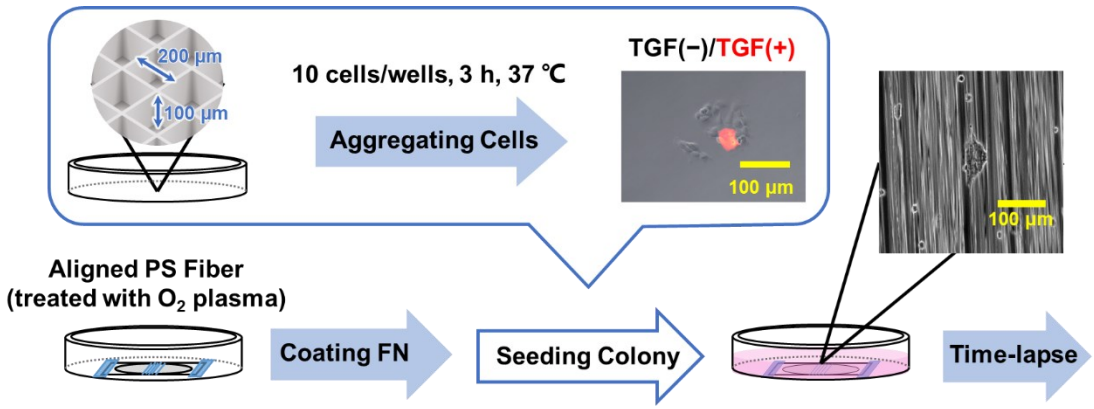


Figure 2.1 Schematic diagram of aligned PS fibers and cell culture.

2.2.3 Morphological Analysis of Colony

Fluorescent images of the colonies were binarized and analyzed using Fiji software. The shape of a cell colony was determined by increased area and circularity. Circularity was determined using Equation (2.1), where *S* and *P* are the area and perimeter of the colony, respectively. A circularity of 1 indicates that the shape is a perfect circle; as the value approaches 0, the shape was considered to have elongated (Figure 2.2).

$$Circularity = 4\pi \times \frac{S}{P^2} \tag{2.1}$$

The definition of the center of mass of a colony was determined using Equation (2.2). This value represents the average of the coordinates of n cells in the collective at a certain time (Figure 2.2). The initial position was set to (0,0).

$$Center\ of\ mass = \frac{1}{n}\sum_{i=1}^{n}(x_i, y_i) \tag{2.2}$$

where x_i, y_i are the coordinates of cell *i*.

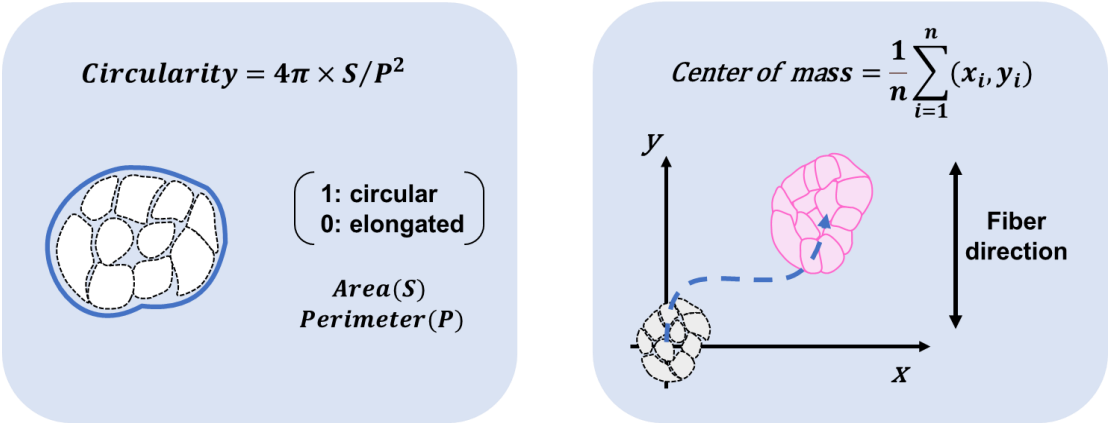


Figure 2.2 The definition of circularity and center of mass.

2.2.4 Migratory Analysis of Single Cell

Fluorescent images of cell nuclei were binarized and analyzed using the Fiji plug-in (TrackMate) to extract the coordinates of motion of each cell at each time point and

trajectory. The velocity, distance, directionality, and angle of cell migration were calculated using these coordinates. The straightness was determined using Equation (2.3) as the ratio of the Euclidean distance of cell migration to the total distance (Figure 2.3).

$$Straightness = \frac{d_{\text{Euclid}}}{d_{\text{Total}}} \tag{2.3}$$

where d_{Euclid} is the Euclidean distance of movement of cells, d_{Total} is the total distance of movement of cells.

The directional angle is the angle between the direction of cell migration and direction perpendicular to the fiber (Figure 2.3). This is used to further investigate the direction difference of cells.

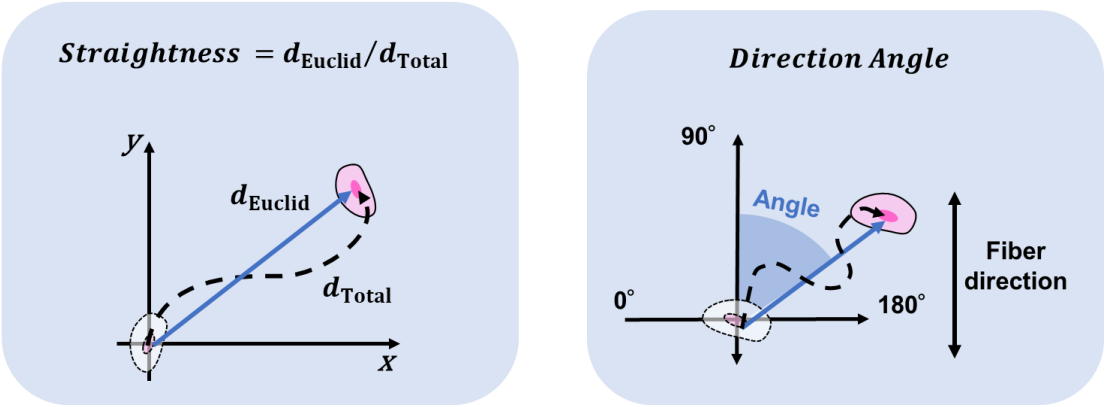


Figure 2.3 The definition of straightness and direction angle.

2.3 Results and Discussion

2.3.1 Cell Migration on Fiber

This thesis characterized the motility of migrating cells in a cell population by analyzing cell behaviors at the single-cell level. The migration of cell populations with different ratios of mesenchymal and epithelial cells were analyzed. First, a system was set up to observe collective migration using different ratios of TGF (+) and TGF (−) cells on directed PS fibers, the surfaces of which were coated with ECM protein (FN). This design mimics the fibrous structure *in vivo* (Figure 2.1). Colony elongation and migration of TGF (+) and TGF (−) cells were evaluated as the migration of cell colonies at different TGF (+) mixing ratios (Figure 2.4). TGF (+) cells were induced with

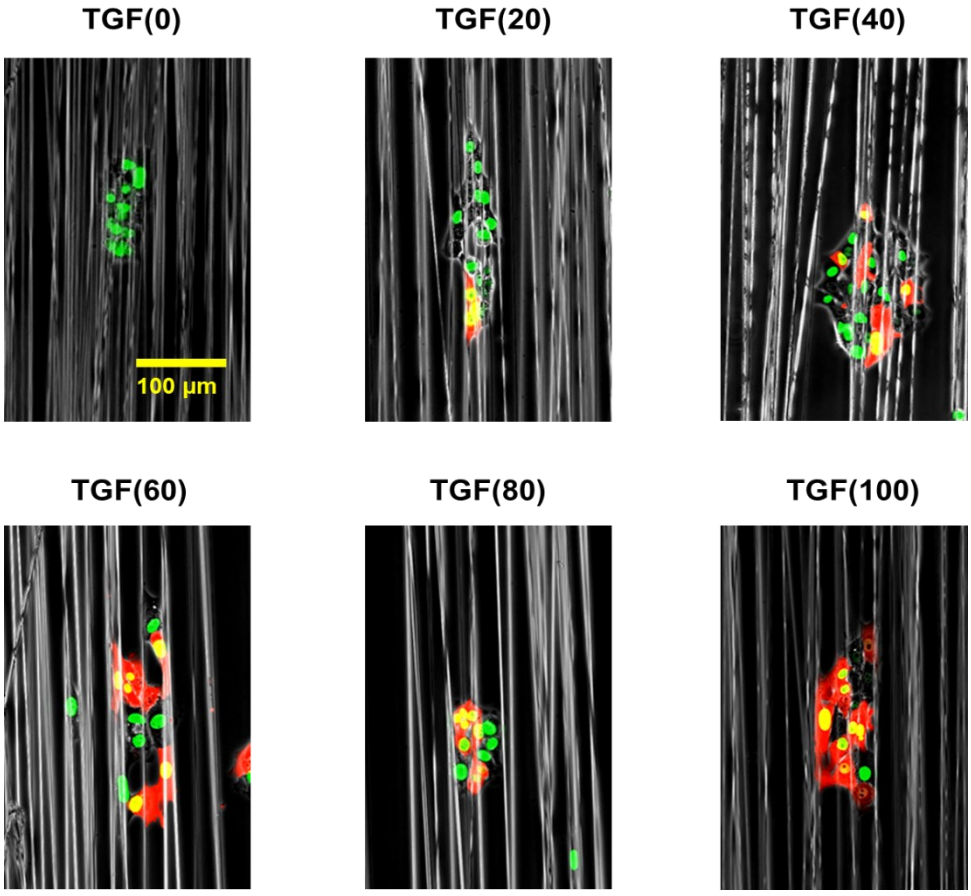


Figure 2.4 TGF(+) cells (red) and TGF(-) cells (green) were inoculated onto the fibers. TGF(*N*) represents the percentage of TGF(+) cells as *N*%.

NMuMG-expressing DsRed to distinguish TGF (+) from TGF (−). Under different percentages of TGF (+) cells, the cells migrated in the colony within 24 h. In the absence of TGF (+) cells [TGF (0)], the cells were strongly attached to each other and the shape of the population did not change significantly. When the percentage of TGF (+) exceeded 50%, the cells were loosely adhered and frequently separated from the colony.

2.3.2 Migration of Colony

Migration from a colony and single-cell perspective were analyzed. First, the colony elongation and migration were observed. Circularity was calculated which is determined by the area and perimeter of the colony. The shapes of cell colonies with different TGF (+)/TGF (−) ratios during migration are shown in Figure 2.5A. TGF (*N*)

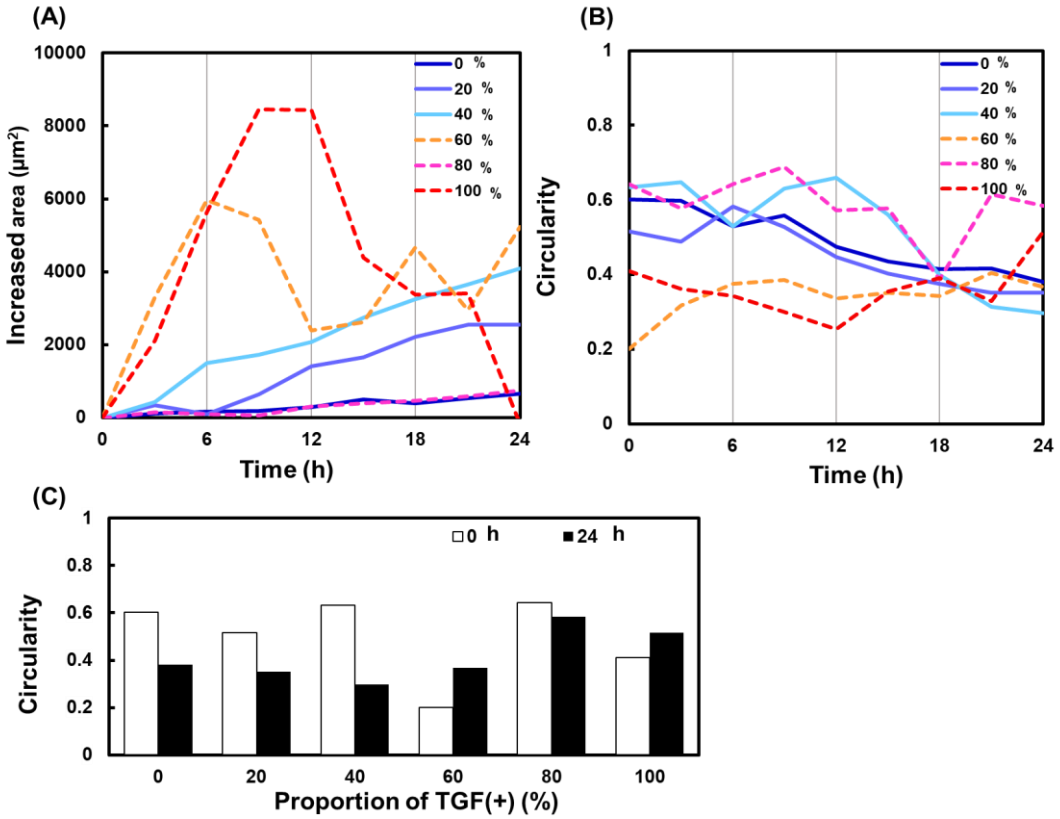


Figure 2.5 Changes in Colony Shape. (A) Area at different TGF (*N*) and time. Increased area is the difference in area calculated every 3 hours. (B) Circularity changes with different time. (C) Circularity at 0 h and 24 h with different TGF (*N*). The *N* is the initial proportion of TGF (+) cells.

represents the percentage of TGF (+) cells as *N*%. In TGF (0)–(40), in which TGF (+) cells accounted for the minority of cells, the area increased as *N* increased and with increasing observation time. Figure 2.5B shows the variation in circularity over time, and Figure 2.5C shows a comparison of circularity at 0 and 24 h. In TGF (0)–(40) cells, the circularity at 24 h decreased significantly compared to that at 0 h, indicating an elongated colony shape. In contrast, in TGF (60)–(100), changes in the area and circularity revealed an unstable state, possibly because of the mesenchymal properties of TGF (+) in most of the colonies, leading to weak cell adhesion and more dispersed migration.

The definition of the center of mass of a colony represents the average of cells coordinates. In Figure 2.6, TGF (0) and TGF (100) did not move as colonies, as their endpoints returned to their original state. TGF (20) moved along the fiber, showing only slight changes in the direction of movement. TGF (80) also moved approximately the same distance but in repeated forward and backward directions. The centers of mass of

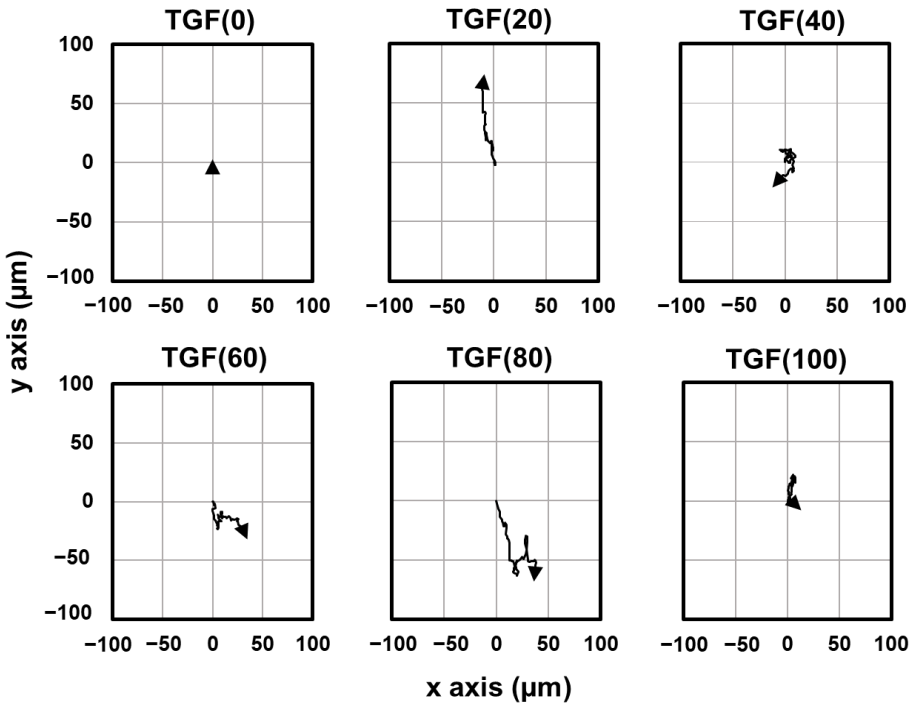


Figure 2.6 Movement of center of mass under different TGF(*N*). The arrow is the direction of the last movement of the center of mass.

TGF (40) and TGF (60) moved similar shorter distances. TGF (+) cell increasing result in the promotion of the entire colony migration most of the time.

2.3.3 Migration of Single Cells

2.3.3.1 Trajectories of Cells

From the aspect a colony, an overall impression of collective migration can be obtained. It's more accuracy to analyze from a single-cell aspect. The single-cell property can detail the interaction in collective migration. Next, the trajectories of each cell were plotted by extracting the time-series data of the coordinates of the nuclei. The

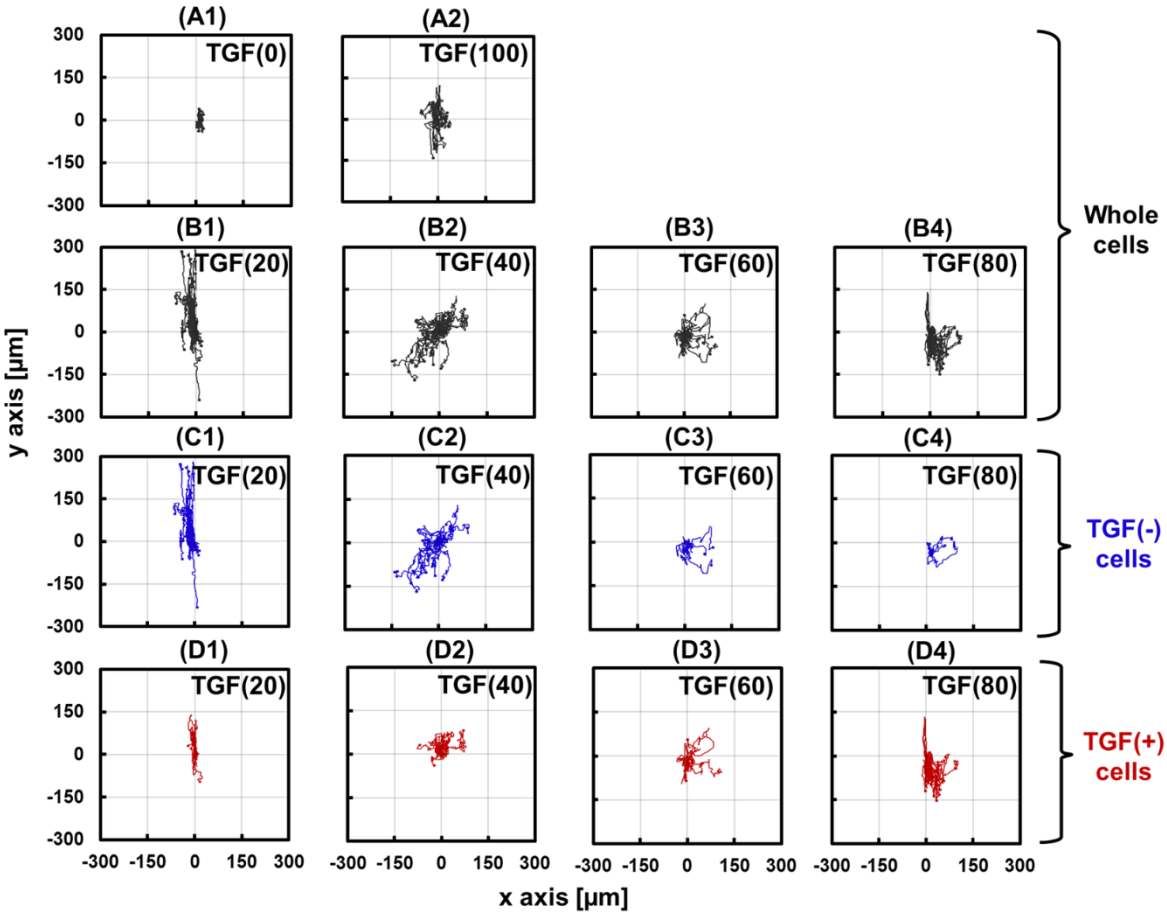


Figure 2.7 Changes in Cell Trajectories. (A1), (A2) and (B1)-(B4) are the trajectories of TGF(+) and TGF(-) under different TGF(*N*). (C1) – (C4) are the trajectories of TGF(-) cells. (D1) – (D4) are the trajectories of TGF(+) cells.

starting point of all cell trajectories was reset to the origin (0,0), as shown in Figure 2.7. The cells migrated mainly along the fibers. TGF (−) cells showed greater migration compared to TGF (+) when $N < 50$ in the direction difference and area expended aspects. TGF (+) and TGF (−) cells belonging to the same colony show similar trajectories in the migration direction when $N > 50$. In TGF (0) cells (Figure 2.7A1), the trajectory did not spread, indicating limited cell migration. In TGF (20) (Figure 2.7B1), the cells migrated mainly along the fibers (y-axis direction). In TGF (40)–(80), the cells migrated along the direction of the fibers and spread in the x-axis direction (Figure 2.7B2–B4). The trajectories of TGF (+) and TGF (−) were separately plotted to investigate migration of these populations independently (Figure 2.7C and D). TGF (−) cells showed greater migration compared to that of TGF (0) cells under all conditions. In addition, TGF (+) and TGF (−) cells belonging to the same colony exhibited similar trajectories in the migration direction. These results suggest that TGF (+) and TGF (−) interact in the same colony and alter the migratory behavior of the entire colony. Interestingly, TGF (−) showed higher migration in TGF (20) and TGF (40) than in TGF (+) (Figure 2.7C1, C2, D1, and D2). This result suggests that the presence of a minority of TGF (+) cells in the colony enhances TGF (−) migration. The presence of TGF (+) cells of majority in a colony showing a limitation to TGF (−) cells migration. When $N >$ 50%, the mesenchymal cells would generate forces to epithelial cells in difference direction which made the epithelial cells lose the consistency of migration and result in a limitation a migration.

2.3.3.2 Straightness of Cells

The straightness was determined here as the ratio of the Euclidean distance of cell migration to the total migration distance. In TGF (0), the straightness was 0.55 ± 0.17 μm but decreased with increasing N, reaching a minimum value of 0.16 ± 0.05 μm in TGF (100) (Figure 2.8A). For TGF (0), intercellular adhesion was strong, all cells move as a unity. They barely change their direction and almost no migration was observed.

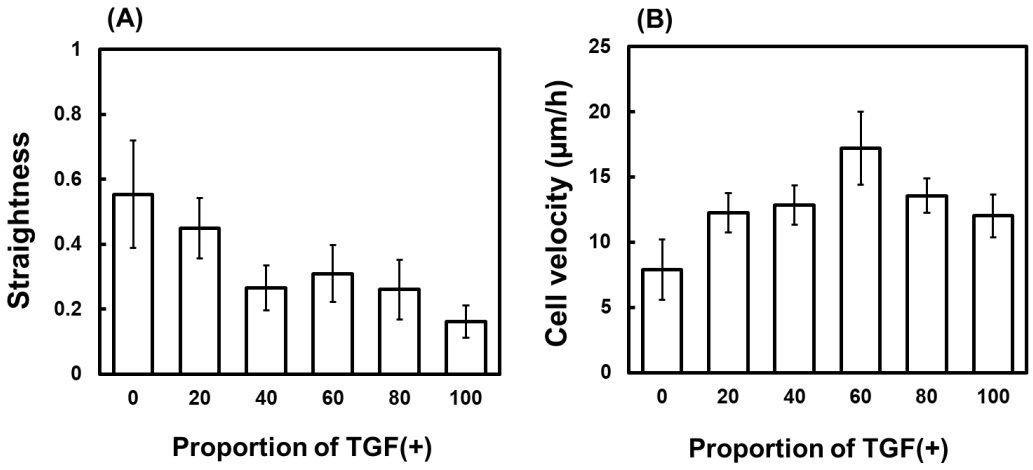


Figure 2.8 Changes in migration direction and velocity. (A) The straightness of all cells with different TGF(*N*). (B) The velocity of all cells with different TGF(*N*). Whiskers are the range of data.

With increasing N, intercellular adhesion loosened, possibly because of a large change in direction related to the increased migratory capacity of each cell. The minimum value of the average migration velocity of the cells was obtained under the TGF (0) condition (7.89 ± 2.33 μm/h). With increasing TGF (+) cells, intercellular adhesion decreased. Because the TGF (+) cells can make TGF (−) cells more active and affect their direction. Cell velocity were increasing before TGF (40) and decrease after TGF (60). When the TGF (+) cells are majority, the active of TGF (−) cells is getting weak. The maximum value was obtained under the TGF (60) condition (17.10 ± 2.31 μm/h) (Figure 2.8B).

2.3.3.3 Directional Angle of Cells

Straightness does not represent the directionality of cell migration. Therefore, the directional angle during cell migration was calculated and displayed as a rose plot (Figure 2.9). The directional angle is the angle between the direction of cell migration and direction perpendicular to the fiber. The mean value of TGF (0) was closest to 90°; as *N* increased, this value differed from 90°, indicating that when the ratio of TGF (+) is higher, fewer cells in the colony migrate along the fibers. Figure 2.10 shows a histogram of the direction angle for each condition; as *N* decreased, a clear peak was

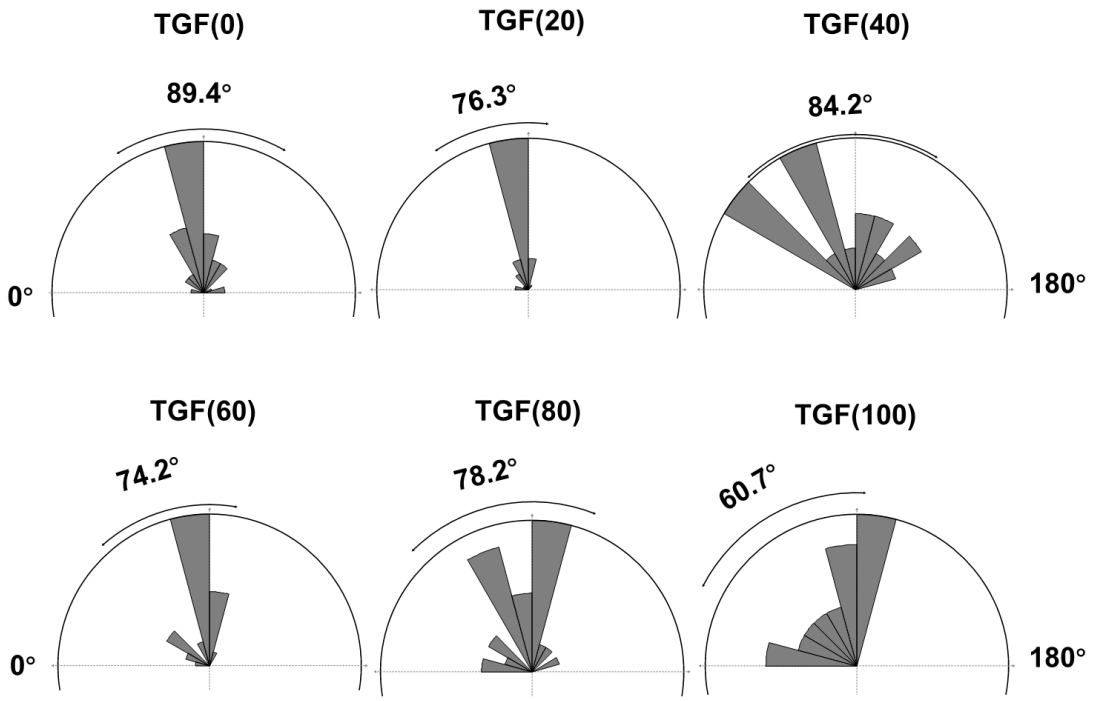


Figure 2.9 The rose plot of distribution of direction angles and median values with different TGF(*N*). The arcs under the median represent 75% of the data.

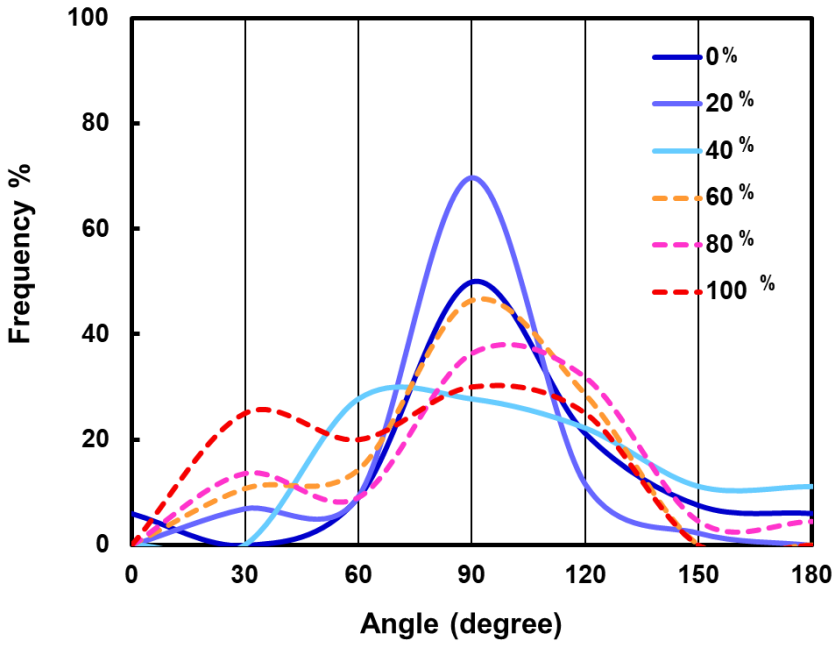


Figure 2.10 The frequency of different TGF(*N*) direction angle values after cubic Bezier curve smoothing. The calculation includes 92 cells.

formed at 90°. This decrease in variability clearly indicates that a change in the ratio of

TGF (+) to TGF (−) alters the migration of the cell population. This indicating that when the ratio of TGF (+) is higher, less and less cells in the colony migrate along the fibers. Their migration become more active and non-directional.

2.3.3.4 The Effect of TGF (+) Cells on the Migration

The thesis examined how the presence of TGF (+) affects each cell in a mixed colony in different situations. In the first situation, $N < 50$. When $N = 20$, TGF (−) exhibited a trajectory along the fiber direction (Figure 2.7C1). When $N = 40$, the migration of TGF (−) was enhanced not only in the fiber direction, but also in other directions (Figure 2.7C2). This pattern was also observed for larger N, which corresponds to the results (Figure 2.7C3, C4). Similar effects were observed not only in the distance of migration, but also in velocity (Figure 2.8C). The velocity of TGF (−) increased with increasing N at $N < 50$. Under this condition, there was more TGF (−), and TGF (+) may have enhanced TGF (−) migration as N increased. This observation is similar to the relationship between leaders and their followers. In collective invasion, leader cells express basal epithelial genes such as cytokeratin-14 and p63 [27]. As the proportion of TGF (+) cells increased, cytokeratin-14 was enriched at the invasion boundary, and the highly migratory cell population became behaviorally and molecularly distinct from other cells, which is factor explaining the formation of leader-follower relationships. Migration mechanisms involving the regulation of cytokeratin-14 alter the physical and chemical properties of cells such as intercellular adhesion and mechanics.

This enhancement does not continue to increase with N; therefore, the second situation was $N > 50$. Notably, the migration range of TGF (−) did not expand but rather shrank (Figure 2.7C3 and C4). Similarly, the migration velocity decreased. Enhanced migration of TGF (+) on TGF (−) is conditional: when the density of TGF (+) is too high, that is, when there are too many leaders, the migration enhancement of followers is inhibited. This causes TGF (−) to lose its centralized leadership, weakening migration

enhancement.

The above two situations used $N = 50$ as the watershed, with the opposite pattern on both sides. A third situation was compared from another perspective to determine the properties that do not have a watershed. Results show that straightness decreased with increasing N (Figure 2.8A), whereas the migration direction moved further away from the fiber with increasing N (Figure 2.10). This result indicates that the presence of TGF (+) significantly enhanced the migration of TGF (−). Enhanced migration increases the total distance and thus decreases straightness. Multiple leaders lead to irregularities in the migration direction of their followers.

The fiber direction can affect the migration of cells. Here, aligned PS fiber was used to mimic the ECM structures. When $N < 50$, most of the TGF (−) cells migrate along the fiber direction. Especially in Figure 2.7 C1, $N = 20$, the cells' trajectories are expended in y axis which is the fiber direction. As TGF (+) cells number growth to $N = 40$, the trajectories can still show an expended in y axis but not the same with $N = 20$. When $N > 50$, the trajectories show a result of random migration which is totally different with the fiber direction. The TGF (+) cells in a majority of the colony can result in the migration direction more randomly.

EMT can be induced by various factors, including transcription factors, growth factors, and microenvironmental miRNAs. TGF-β as an induction factor for EMT in this thesis. TGF-β can induce EMT via signaling pathways such as Smad, RhoA, and MAPK. The results shown here provide insight into the effects of these processes on cell migration behavior. Cells before and after induction of EMT by TGF-β exhibited different migratory properties and showed interactions at different ratios. To further analyze this phenomenon of cellular interactions, The thesis proposed a method based on a combination of time-series clustering and dimensionality reduction to analyze cells with similar migration patterns in Chapter 3. Whether cell trajectories are related to migration patterns remains unclear. In the future, this approach will be applied to analyze cell lines with different proportions of phenotypes and has the potential to

provide an accurately analyzing of interactions based on cell migration trajectories.

2.4 Conclusion

This chapter established a method for observing the migration of cell colonies on fibers. Using this method, the migration of colonies was observed using different ratios of TGF (+) and TGF (−) to quantify the effect of TGF (+) on collective migration. Migration was assessed from two perspectives: that of the entire colony and that of each cell within the colony. When TGF (+) was present at low densities in the colony, the migration distance of individual cells and behavior of the colony were significantly enhanced. This enhancement decreased when the TGF (+) density exceeded a certain level. These collective behaviors were caused by a leader–follower-like structure. The individual cells during collective migration were analyzed. In the future, by precisely analyzing the interactions among individual cells, it will be possible to assess the migratory properties of collective migration more clearly.

REFERENCE

[1] Weiss, F.; Lauffenburger, D.; Friedl, P., Towards targeting of shared mechanisms of cancer metastasis and therapy resistance. *Nat Rev Cancer* **2022,** 22, (3), 157-173. doi: 10.1038/s41568-021-00427-0.

[2] Kolsch, V.; Charest, P. G.; Firtel, R. A., The regulation of cell motility and chemotaxis by phospholipid signaling. *J Cell Sci* **2008,** 121, (Pt 5), 551-9. doi: 10.1242/jcs.023333.

[3] Pollard, T. D.; Cooper, J. A., Actin, a central player in cell shape and movement. *Science* **2009,** 326, (5957), 1208-12. doi: 10.1126/science.1175862.

[4] Mayor, R.; Etienne-Manneville, S., The front and rear of collective cell migration. *Nat Rev Mol Cell Biol* **2016,** 17, (2), 97-109. doi: 10.1038/nrm.2015.14.

[5] Zoeller, E. L.; Pedro, B.; Konen, J.; Dwivedi, B.; Rupji, M.; Sundararaman, N.; Wang, L.; Horton, J. R.; Zhong, C. J.; Barwick, B. G.; Cheng, X. D.; Martinez, E. D.; Torres, M. P.; Kowalski, J.; Marcus, A. I.; Vertino, P. M., Genetic heterogeneity within collective invasion packs drives leader and follower cell phenotypes. *J Cell Sci* **2019,** 132, (19). doi: 10.1242/jcs.231514.

[6] Pedro, B. A.; Konen, J.; Summerbell, E.; Mouw, J. K.; Rupji, M.; Dwivedi, B.; Kowalski, J.; Vertino, P. M.; Marcus, A. I., Dissecting the biology of leader and follower cells in collective cancer invasion. *Cancer Res* **2019,** 79, (13). doi: 10.1158/1538-7445.Sabcs18-4590.

[7] Borghi, N.; Lowndes, M.; Maruthamuthu, V.; Gardel, M. L.; Nelson, W. J., Regulation of cell motile behavior by crosstalk between cadherin- and integrin-mediated adhesions. *Proc Natl Acad Sci U S A* **2010,** 107, (30), 13324-9. doi: 10.1073/pnas.1002662107.

[8] Yang, Y.; Levine, H., Leader-cell-driven epithelial sheet fingering. *Phys Biol* **2020,** 17, (4), 046003. doi: 10.1088/1478-3975/ab907e.

[9] Vishwakarma, M.; Spatz, J. P.; Das, T., Mechanobiology of leader-follower dynamics in epithelial cell migration. *Curr Opin Cell Biol* **2020,** 66, 97-103. doi: 10.1016/j.ceb.2020.05.007.

[10] Pedro, B.; Rupji, M.; Dwivedi, B.; Kowalski, J.; Konen, J. M.; Owonikoko, T. K.; Ramalingam, S. S.; Vertino, P. M.; Marcus, A. I., Prognostic significance of an invasive leader cell-derived mutation cluster on chromosome 16q. *Cancer* **2020,** 126, (13), 3140-3150. doi: 10.1002/cncr.32903.

[11] Bocci, F.; Jolly, M. K.; Tripathi, S. C.; Aguilar, M.; Hanash, S. M.; Levine, H.; Onuchic, J. N., Numb prevents a complete epithelial-mesenchymal transition by modulating Notch signalling. *J R Soc Interface* **2017,** 14, (136). doi: 10.1098/rsif.2017.0512.

[12] Gamboa Castro, M.; Leggett, S. E.; Wong, I. Y., Clustering and jamming in epithelial-mesenchymal co-cultures. *Soft Matter* **2016,** 12, (40), 8327-8337. doi: 10.1039/c6sm01287f.

[13] Fontana, F.; Marzagalli, M.; Sommariva, M.; Gagliano, N.; Limonta, P., In Vitro 3D Cultures to Model the Tumor Microenvironment. *Cancers (Basel)* **2021,** 13, (12). doi: 10.3390/cancers13122970.

[14] Tekpli, X.; Lien, T.; Rossevold, A. H.; Nebdal, D.; Borgen, E.; Ohnstad, H. O.; Kyte, J. A.; Vallon-Christersson, J.; Fongaard, M.; Due, E. U.; Svartdal, L. G.; Sveli, M. A. T.; Garred, O.; Osbreac; Frigessi, A.; Sahlberg, K. K.; Sorlie, T.; Russnes, H. G.; Naume, B.; Kristensen, V. N., An independent poor-prognosis subtype of breast cancer defined by a distinct tumor immune microenvironment. *Nat Commun* **2019,** 10, (1), 5499. doi: 10.1038/s41467-019-13329-5.

[15] Feng, C.; Chen, L.; Lu, Y.; Liu, J.; Liang, S.; Lin, Y.; Li, Y.; Dong, C., Programmable Ce6 Delivery via Cyclopamine Based Tumor Microenvironment Modulating Nano-System for Enhanced Photodynamic Therapy in Breast Cancer. *Front Chem* **2019,** 7, 853. doi: 10.3389/fchem.2019.00853.

[16] Yu-Ju Wu, C.; Chen, C. H.; Lin, C. Y.; Feng, L. Y.; Lin, Y. C.; Wei, K. C.; Huang, C. Y.; Fang, J. Y.; Chen, P. Y., CCL5 of glioma-associated microglia/macrophages regulates glioma migration and invasion via calcium-dependent matrix metalloproteinase 2. *Neuro Oncol* **2020,** 22, (2), 253-266. doi: 10.1093/neuonc/noz189.

[17] Sokullu, E.; Cucuk, Z. L.; Sarabi, M. R.; Birtek, M. T.; Bagheri, H. S.; Tasoglu, S., Microfluidic Invasion Chemotaxis Platform for 3D Neurovascular Co-Culture. *Fluids* **2022,** 7, (7). doi: 10.3390/fluids7070238.

[18] Xiao, Y.; Riahi, R.; Torab, P.; Zhang, D. D.; Wong, P. K., Collective Cell Migration in 3D Epithelial Wound Healing. *Acs Nano* **2019,** 13, (2), 1204-1212. doi: 10.1021/acsnano.8b06305.

[19] Vishwakarma, M.; Di Russo, J.; Probst, D.; Schwarz, U. S.; Das, T.; Spatz, J. P., Mechanical interactions among followers determine the emergence of leaders in migrating epithelial cell collectives. *Nature Communications* **2018,** 9. doi: 10.1038/s41467-018-05927-6.

[20] Lin, J. Y.; Lo, K. Y.; Sun, Y. S., A microfluidics-based wound-healing assay for studying the effects of shear stresses, wound widths, and chemicals on the wound-healing process. *Sci Rep* **2019,** 9, (1), 20016. doi: 10.1038/s41598-019-56753-9.

[21] Bandzerewicz, A.; Gadomska-Gajadhur, A., Into the Tissues: Extracellular Matrix and Its Artificial Substitutes: Cell Signalling Mechanisms. *Cells-Basel* **2022,** 11, (5). doi: 10.3390/cells11050914.

[22] Yamada, K. M.; Sixt, M., Mechanisms of 3D cell migration. *Nat Rev Mol Cell Biol* **2019,** 20, (12), 738-752. doi: 10.1038/s41580-019-0172-9.

[23] Jiang, T.; Carbone, E. J.; Lo, K. W. H.; Laurencin, C. T., Electrospinning of polymer nanofibers for tissue regeneration. *Prog Polym Sci* **2015,** 46, 1-24. doi: 10.1016/j.progpolymsci.2014.12.001.

[24] Braghirolli, D. I.; Steffens, D.; Pranke, P., Electrospinning for regenerative medicine: a review of the main topics. *Drug Discov Today* **2014,** 19, (6), 743-53. doi: 10.1016/j.drudis.2014.03.024.

[25] Liu, Y. Y.; Wang, Y.; Zhang, Y.; Wang, B.; Pu, H. Y.; Zhong, S. Y.; Xie, S. R.; Peng, Y.; Luo, J.; Yue, T.; Liu, N., Vascular Scaffold for Guides Endothelial Cell Alignment by Electrohydrodynamic and Electrospinning. *J Biomater Tiss Eng* **2019,** 9, (3), 298-303. doi: 10.1166/jbt.2019.1991.

[26] Hong, J.; Yeo, M.; Yang, G. H.; Kim, G., Cell-Electrospinning and Its Application for Tissue Engineering. *Int J Mol Sci* **2019,** 20, (24). doi: 10.3390/ijms20246208.

[27] Cheung, K. J.; Gabrielson, E.; Werb, Z.; Ewald, A. J., Collective invasion in breast cancer requires a conserved basal epithelial program. *Cell* **2013,** 155, (7), 1639-51. doi: 10.1016/j.cell.2013.11.029.

CHAPTER 3

TIME-SERIES CLUSTERING OF SINGLE-CELL TRAJECTORIES IN COLLECTIVE CELL MIGRATION

Abstract

In this chapter, a trajectory-based time-series clustering method is suggested for further study of cell-cell interactions. After collecting high-dimensional time-series data from time-lapse observations, these data was processed by dimension reduction and clustering algorithms to find similar migration patterns.

3.1 Introduction

The process of collective cell migration can be described to varying degrees using molecular mechanisms and mechanical models [1-3]. However, how these interactions integrate multiple factors that influence migration patterns requires further investigation [4]. This chapter used a simpler approach based on the clustering of single-cell trajectories to explore the migration patterns between cells.

The study of cell migration patterns usually requires recording cell trajectories within observation windows comprising multiple sets of timeseries data. Timeseries data are characterized by high-dimensionality and large data volumes [5]. Clustering such com-plex objects can reveal interesting patterns. As an unsupervised learning technique, k-means clustering is a commonly used clustering method. The main idea behind this clustering method is to minimize the total distance (usually the Euclidean distance) be-tween all the objects in a cluster and their cluster centers. Cluster centers are defined as the average vectors of the objects in the cluster; however, time-series

clustering poses additional challenges. For cell trajectories, this increases the difficulty and computational effort of clustering based on the time dimension and poor clustering performance owing to the different locations of cell trajectories at each time record point, and similar problems have been observed in the phylogenetic analysis [6, 7].

One solution is dimensionality reduction of time-series data and subsequent clustering. Commonly used dimensionality reduction methods to maintain the two-by-two distance structure of the dataset include PCA [8] and multi-dimensional scaling (MDS) techniques [9], and streamwise learning techniques such as t-distributed stochastic neighbor embedding (t-SNE), which can maintain the local structure of the data [10] and uniform manifold approximation and projection (UMAP) [11]. In particular, UMAP is efficient for processing large amounts of data in response to the global structure, providing higher-quality visualization. Clustering is often used in biological disciplines for the identification of functionally relevant genes and functional clustering of gene expression data [12, 13]; however, to my knowledge, this is the first re-port to analyze the range of influence of leader cells on collective migration by classifying the movement pattern with clustering analysis of individual cell tracking.

This chapter attempted to identify similar migration patterns among cells. As an approach to quantitatively analyze cell–cell interactions, co-culture of EMT-induced mesenchymal cells by the addition of TGF-β (referred as TGF (+)) and EMT non-induced cells (TGF (−)) was performed to imitate colony migration on aligned electrospun fibers mimicking the ECM fibrous architecture. Tracking of cell migration by time-lapse observation followed by dimensionality reduction and clustering was performed to find similar migration patterns, where they are not only related to location but also influenced by cell division.

3.2 Materials and Methods

3.2.1 Electrospinning

This chapter used the same electrospinning method as Section 2.2.1. The aligned PS nanofiber was treated with O_2 plasma and coated with FN to mimic the ECM.

3.2.2 Cell Culture and Time-Lapse Observation

The method of preparing the cell culture is the same as Section 2.2.2. To prepare cell aggregates, TGF (−) cells and EMT-induced TGF (+) were seeded on fibers. Time-lapse microscopy was performed for 24 h at intervals of 15 min using Biostation and analyzed using ImageJ (ver. 1.53 e). The captured fluorescence images were binarized and the cells were tracked using a plug-in (TrackMate). The location of the center of the cell nucleus was determined by manual tracking which started from the last time slice to the first time slice backward. The cell was identified on the basis of the locations of the cell nucleus between two adjacent slices and assigned ID numbers to all cells. The slice when cell division occurred was determined by comparing the distance between neighboring cell marker points.

3.2.3 Time-series clustering

Figure 3.1 shows the schematic diagram of time-series clustering of cell trajectories. The Appendix 1 shows the flowchart. After collecting cell migration movies by time-lapse microscopy, cell trajectories were generated manually by period using ImageJ. Second, the trajectories of each period were normalized to convert their starting points to the origin (0,0). Third, the distance matrix is calculated from the Euclidean distance of each pair of trajectories. Fourth, the UMAP dimensionality reduction algorithm to visualize the data in two-dimensions. Fifth, the data are clustered by k-means, where the number of clusters is determined by the mean silhouette

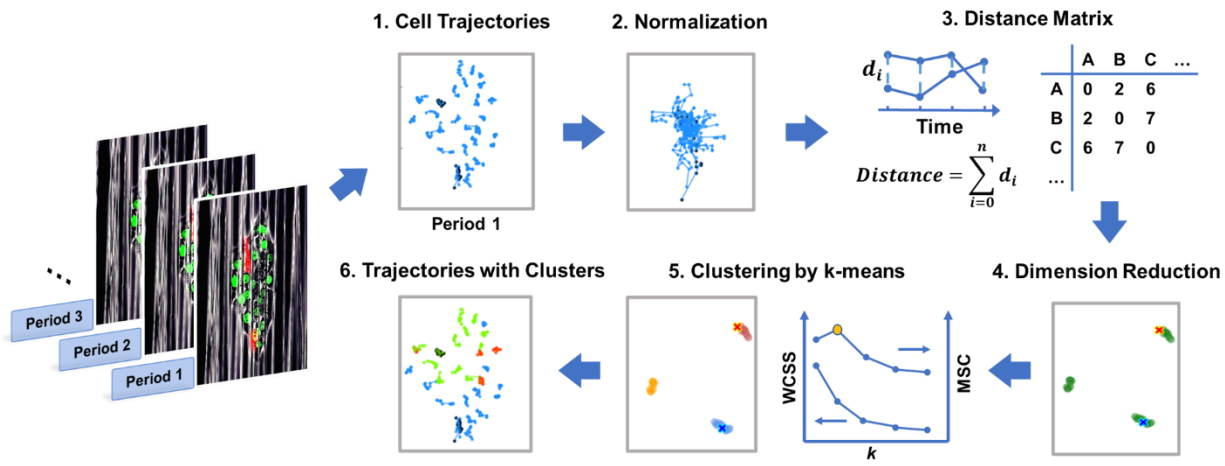


Figure 3.1 Schematic diagram of time-series clustering of cell trajectories in collective cell migration.

coefficient (MSC). Finally, the clusters are combined with the original trajectories to analyze the similar migration patterns in collective cell migration

3.2.3.1 Normalization Cell Trajectory Data

Let

$$V_{\mathrm{i}}^{(\mathrm{p})} \coloneqq \left\langle \boldsymbol{v}_{\mathrm{i,1}}^{(\mathrm{p})}, \dots, \boldsymbol{v}_{\mathrm{i,n}}^{(\mathrm{p})} \right\rangle \tag{3.1}$$

be the ith cell's trajectory of the pth time period, where $\boldsymbol{v}_{\mathrm{i,l}}^{(\mathrm{p})} \in \mathbb{R}^2$ is the cell position at lth time slice of the pth time period and n is the number of time slices in a single time period.

The cells were divided several times during the observation. After manually marking the positions of the cell nuclei and finding mother cells with daughter cells (Figure 3.2), the trajectories of two daughter cells were unified with the mother cell prior to division. To remove the effect of the initial cell positions from the cell trajectory clustering, The method first normalized the ith cell's trajectory of pth time period $V_{\mathrm{i}}^{(\mathrm{p})}$ such that its initial position $\boldsymbol{v}_{\mathrm{i,l}}^{(\mathrm{p})}$ was located at the origin $(0,0)$. The normalized trajectory is given by

$$Z_{\mathrm{i}}^{(\mathrm{p})} \coloneqq \left\langle \boldsymbol{z}_{\mathrm{i,1}}^{(\mathrm{p})}, \dots, \boldsymbol{z}_{\mathrm{i,n}}^{(\mathrm{p})} \right\rangle, \tag{3.2}$$

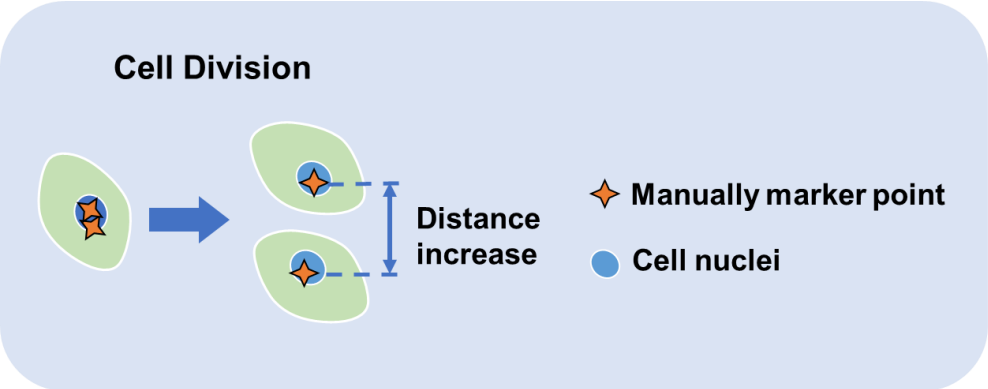


Figure 3.2 Schematic of manually mark the cells' nuclei and the distance increase after cell division occurred.

where $\boldsymbol{z}_{\mathrm{i,l}}^{(\mathrm{p})} := \boldsymbol{v}_{\mathrm{i,l}}^{(\mathrm{p})} - \boldsymbol{v}_{\mathrm{i,1}}^{(\mathrm{p})} = \left(z_{\mathrm{i,l,1}}^{(\mathrm{p})}, z_{\mathrm{i,l,2}}^{(\mathrm{p})}\right)^{\mathsf{T}}$, refer to the Appendix 2, Part 2.

To evaluate the similarity of cell trajectories, the pairwise distance between two trajectories was defined as follows: the distance between two trajectories at the pth time period, $Z_{\mathrm{i}}^{(\mathrm{p})}$ and $Z_{\mathrm{j}}^{(\mathrm{p})}$, is defined by

$$D\left(Z_{\mathrm{i}}^{(\mathrm{p})}, Z_{\mathrm{j}}^{(\mathrm{p})}\right) := \sum_{l=1}^{n} \delta\left(\boldsymbol{z}_{\mathrm{i,l}}^{(\mathrm{p})}, \boldsymbol{z}_{\mathrm{j,l}}^{(\mathrm{p})}\right), \tag{3.3}$$

where $\delta\left(\boldsymbol{z}_{\mathrm{i,l}}^{(\mathrm{p})}, \boldsymbol{z}_{\mathrm{j,l}}^{(\mathrm{p})}\right)$ is the distance between two points, $\boldsymbol{z}_{\mathrm{i,l}}^{(\mathrm{p})}$ and $\boldsymbol{z}_{\mathrm{j,l}}^{(\mathrm{p})}$, which are respectively ith and jth cell positions at lth time slice of pth time period.

This study employed the Euclidean distance to define the distance function $\delta\left(\boldsymbol{z}_{\mathrm{i,l}}^{(\mathrm{p})}, \boldsymbol{z}_{\mathrm{j,l}}^{(\mathrm{p})}\right)$:

$$\delta\left(\boldsymbol{z}_{\mathrm{i,l}}^{(\mathrm{p})}, \boldsymbol{z}_{\mathrm{j,l}}^{(\mathrm{p})}\right) := \sqrt{\left(z_{\mathrm{i,l,1}}^{(\mathrm{p})} - z_{\mathrm{j,l,1}}^{(\mathrm{p})}\right)^2 + \left(z_{\mathrm{i,l,2}}^{(\mathrm{p})} - z_{\mathrm{j,l,2}}^{(\mathrm{p})}\right)^2}. \tag{3.4}$$

Based on the pairwise distance between the two trajectories $D\left(Z_{\mathrm{i}}, Z_{\mathrm{j}}\right)$, the distance matrix is defined as follows:

$$D := \begin{bmatrix} D_{1,1} & \cdots & D_{1,\mathrm{N}} \\ \vdots & \ddots & \vdots \\ D_{\mathrm{N},1} & \cdots & D_{\mathrm{N,N}} \end{bmatrix}, \tag{3.5}$$

where $D_{\mathrm{i,j}} := D\left(Z_{\mathrm{i}}, Z_{\mathrm{j}}\right)$ and N is the total number of cells and their trajectories.

3.2.3.2 Dimension Reduction and Clustering

To visualize the similarity of the cell trajectories, each cell trajectory $Z_{\mathrm{i}}^{(\mathrm{p})}$ was embedded into a two-dimensional space based on the distance matrix D. The t-SNE [10, 14-16] and UMAP [11] were employed as the embedding methods using Python (scikit-learn 1.1.2; umap-learn 0.5.3). The dimensionality reduction algorithm includes distance calculation, but this thesis still performed the distance matrix calculation in advance. The purpose is to further investigate the relationship between the data. By applying these two methods, a set of cell trajectories in pth time period represented in a two-dimensional space was obtained :

$$\mathcal{D}^{(\mathrm{p})} \coloneqq \left\{\boldsymbol{x}_1^{(\mathrm{p})}, \dots, \boldsymbol{x}_{\mathrm{N}}^{(\mathrm{p})}\right\}, \tag{3.6}$$

where $\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})} \in \mathbb{R}^2$ is the ith cell trajectory of the pth time period in two-dimensional space.

Next, the method classified cell trajectories based on their similarities. Here, the k-means clustering method [17] was employed for classification. Using k-means clustering, the N samples of $\mathcal{D}^{(\mathrm{p})}$ were partitioned into $K_{\mathrm{p}}(\leq N)$ sets of the pth time period $M_{\mathrm{p}} = \left\{M_1^{(\mathrm{p})}, M_2^{(\mathrm{p})}, \dots, M_{\mathrm{K_p}}^{(\mathrm{p})}\right\}$ based on the similarities among the samples.

Let

$$\mathcal{M}_{\mathrm{p}} = \left\{\boldsymbol{\mu}_1^{(\mathrm{p})}, \dots, \boldsymbol{\mu}_{\mathrm{K_p}}^{(\mathrm{p})}\right\} \tag{3.7}$$

be a set of the mean of points in each cluster at the pth time period, where $\boldsymbol{\mu}_{\mathrm{k}}^{(\mathrm{p})}$ is the mean of points in the kth cluster at the pth time period. The method also introduced a variable that denotes whether a set $M_{\mathrm{k}}^{(\mathrm{p})}$ contains $\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}$ or not:

$$q_{\mathrm{i,k}}^{(\mathrm{p})} = \begin{cases} 1 & \left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})} \in M_{\mathrm{k}}^{(\mathrm{p})}\right), \\ 0 & \left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})} \notin M_{\mathrm{k}}^{(\mathrm{p})}\right). \end{cases} \tag{3.8}$$

Using the variables, the objective function of k-means clustering, the within-cluster sum of squares (WCSS), is defined as

$$J\left(q_{\mathrm{i,k}}^{(\mathrm{p})}, \boldsymbol{\mu}_{\mathrm{k}}^{(\mathrm{p})}\right) = \sum_{i=1}^{N} \sum_{k=1}^{K_{\mathrm{p}}} q_{\mathrm{i,k}}^{(\mathrm{p})} \left\| \boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})} - \boldsymbol{\mu}_{\mathrm{k}}^{(\mathrm{p})} \right\|^2. \tag{3.9}$$

The optimal solution is a parameter that minimizes the objective function $J\left(q_{\mathrm{i,k}}^{(\mathrm{p})}, \boldsymbol{\mu}_{\mathrm{k}}^{(\mathrm{p})}\right)$. The method numerically solved the optimization problem using the k-means algorithm. Notably, the solution of the k-means algorithm has an initial value dependency. Therefore, several initial values were tested and employed a clustering result that minimized the objective function.

The number of clusters at the pth time period K_{p} is a hyperparameter of k-means clustering. Therefore, the method optimized K_{p} based on the silhouette analysis [18]. Silhouette analysis evaluates clustering results based on the degree of aggregation within clusters and the degree of separation between clusters.

The silhouette coefficients of the samples were calculated at different K_{p} values. This coefficient measures how similar an object is to its cluster compared with other clusters. After the samples were clustered into K_{p} clusters, the average distance of $\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}$ from the other samples in the cluster was calculated for the sample $\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}$ in cluster $M_{\mathrm{I}}^{(\mathrm{p})}$. Within-cluster dissimilarity, which measures how well $\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}$ is assigned to its cluster, is defined as follows:

$$a\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right) = \frac{1}{\left|M_{\mathrm{I}}^{(\mathrm{p})}\right| - 1} \sum_{j \in M_{\mathrm{I}}^{(\mathrm{p})}, j \neq i} \delta\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}, \boldsymbol{x}_{\mathrm{j}}^{(\mathrm{p})}\right), \tag{3.10}$$

where $\left|M_{\mathrm{I}}^{(\mathrm{p})}\right|$ is the number of samples belonging to cluster *I*. The definition of the distance function is the same as that of the distance function in Equation (3.4).

The between-cluster dissimilarity, the average distance between $\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}$ and all samples in the other cluster $M_{\mathrm{J}}^{(\mathrm{p})}$, are defined as follows:

$$b\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right) = \min_{J \neq I} \frac{1}{\left|M_{\mathrm{J}}^{(\mathrm{p})}\right|} \sum_{j \in M_{\mathrm{J}}^{(\mathrm{p})}} \delta\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}, \boldsymbol{x}_{\mathrm{j}}^{(\mathrm{p})}\right). \tag{3.11}$$

The silhouette coefficient of sample $\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}$ is defined as follows:

$$s\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)=\begin{cases}1-\dfrac{a\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)}{b\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)}, & \text{if } a\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)<b\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)\\ 0, & \text{if } a\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)=b\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)\\ \dfrac{b\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)}{a\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)}-1, & \text{if } a\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)>b\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)\end{cases} \tag{3.12}$$

The silhouette coefficient $s\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)$ is a measure of how well the data are clustered over the entire dataset. A larger value indicates a more reasonable clustering of $\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}$.

Based on the silhouette coefficients, the optimal number of clusters K_{p} was obtained:

$$\widehat{K}_{\mathrm{p}}=\underset{K_{\mathrm{p}}}{arg\,max}\,\bar{s}(K_{\mathrm{p}}), \tag{3.13}$$

where $\bar{s}(K_{\mathrm{p}})$ is the MSC, defined as

$$\bar{s}(K_{\mathrm{p}}):=\frac{1}{N}\sum_{i=1}^{N}s\left(\boldsymbol{x}_{i}^{(p)}\right). \tag{3.14}$$

The above equation represents the mean $s\left(\boldsymbol{x}_{\mathrm{i}}^{(\mathrm{p})}\right)$ over all cell trajectories for a specific number of clusters K_{p}. Cell–cell interactions can change during migration, and cell division also perturbs these interactions. Thus, the number of clusters K_{p} may vary over time. Therefore, this process is repeated for each period p and is used to determine the appropriate number of clusters $\widehat{K}_{\mathrm{p}}$, refer to the Appendix 2, Part 3 .

3.2.4 Validation of Robustness

Our method depends on the accuracy of the cell tracking data, particularly the accuracy of the estimated positions of the cell nucleus center. In this study, the positions of the cell nucleus centers were detected manually from bright field images, which might have been affected by the observation noise. Thus, the robustness of our

clustering method was evaluated to the intensity of observation noise in the estimated cell nucleus positions.

As shown in Equation (3.1) $V_{\mathrm{i}}^{(\mathrm{p})} = \left\langle \boldsymbol{v}_{\mathrm{i,1}}^{(\mathrm{p})}, \dots, \boldsymbol{v}_{\mathrm{i,n}}^{(\mathrm{p})} \right\rangle$ denotes the cell trajectories, where $\boldsymbol{v}_{\mathrm{i,l}}^{(\mathrm{p})} \in \mathbb{R}^2$ is the cell position at the *l*th time slice of the pth time period and n is the number of time slices in a single time period.

The method assume that the observation noise of the cell nucleus center position is modeled as two-dimensional Gaussian noise as follows:

$$\{\boldsymbol{\xi}_{\mathrm{l}}\} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_2(\mathbf{0}, \Sigma). \tag{3.15}$$

Here, $\boldsymbol{\xi}_l \in \mathbb{R}^2$ and $\mathcal{N}_2(0, \Sigma)$ denote a two-dimensional Gaussian distribution, where the mean $0.\Sigma$ denotes the variance–covariance matrix, which is defined as

$$\Sigma = \begin{bmatrix} \sigma & 0 \\ 0 & \sigma \end{bmatrix}. \tag{3.16}$$

The true nucleus center $\boldsymbol{v}_{\mathrm{i,l}}^{(\mathrm{p})}$ is unknown, but the method assumed that $\boldsymbol{v}_{\mathrm{i,l}}^{(\mathrm{p})}$ is the true center. The method then generated the test dataset by adding Gaussian noise to the original single-cell trajectory dataset, as follows:

$$V_{\mathrm{i}}{}'^{(\mathrm{p})} = \left\langle \boldsymbol{v}_{\mathrm{i,1}}^{(\mathrm{p})} + \boldsymbol{\xi}_1, \dots, \boldsymbol{v}_{\mathrm{i,n}}^{(\mathrm{p})} + \boldsymbol{\xi}_{\mathrm{n}} \right\rangle. \tag{3.17}$$

Next, the consistency of the clustering results obtained from the datasets was evaluated, original dataset, and generated test datasets. Let $n\left(M_{\mathrm{k}}^{(\mathrm{p})}\right)$ be the number of elements in the kth cluster at the pth time period, and $\boldsymbol{n}^{(\mathrm{p})}$ denote a vector whose element is $n\left(M_{\mathrm{k}}^{(\mathrm{p})}\right)$ for $1 \le k \le K_{\mathrm{p}}$, which is defined as

$$\boldsymbol{n}^{(\mathrm{p})} \coloneqq \left(n\left(M_1^{(\mathrm{p})}\right), n\left(M_2^{(\mathrm{p})}\right), \dots, n\left(M_{\mathrm{K_p}}^{(\mathrm{p})}\right)\right)^{\top}, \tag{3.18}$$

where $n\left(M_{\mathrm{k}}^{(\mathrm{p})}\right)$ for $1 \le k \le K_p$ is sorted in descending order; thus, it satisfies $n\left(M_{\mathrm{k}'}^{(\mathrm{p})}\right) \ge n\left(M_{\mathrm{k}'+1}^{(\mathrm{p})}\right)$.

Let $\boldsymbol{n}_{\mathrm{u}}^{(\mathrm{p})}$ be $\boldsymbol{n}^{(\mathrm{p})}$ obtained from data $\mathcal{D}_{\mathrm{u}}^{(\mathrm{p})}$, where $\mathcal{D}_{\mathrm{u}}^{(\mathrm{p})}$ denotes a set of cell trajectories in the pth time period represented in two-dimensional space (see Equation

(3.6)), $u \geq 1. u = 0$ denotes the index of the original dataset, and $u \geq 1$ denotes that of the test dataset generated by adding Gaussian noise to the original dataset. The consistency of the clustering results obtained from the original and test datasets is defined by the Euclidean distance between the clustering results obtained from the two datasets. Before calculating the consistency, $\boldsymbol{n}_{\mathrm{u}}^{(\mathrm{p})}$ was normalized by the total number of cells in the pth time slice $N^{(\mathrm{p})}$. The normalized clustering result is defined as follows:

$$\boldsymbol{n}_{\mathrm{u}}^{\prime(\mathrm{p})} \coloneqq \left(\frac{n_{\mathrm{u}}\left(M_1^{(\mathrm{p})}\right)}{N^{(\mathrm{p})}}, \ldots, \frac{n_{\mathrm{u}}\left(M_{\mathrm{K_p}}^{(\mathrm{p})}\right)}{N^{(\mathrm{p})}} \right)^{\mathrm{T}}. \tag{3.19}$$

Then, the consistency between the original dataset and the test dataset is defined by the Euclidean distance between $\boldsymbol{n}_0^{\prime(\mathrm{p})}$ and $\boldsymbol{n}_{\mathrm{u}}^{\prime(\mathrm{p})}(u \geq 1)$, which is

$$\delta\left(\boldsymbol{n}_0^{\prime(\mathrm{p})}, \boldsymbol{n}_{\mathrm{u}\geq 1}^{\prime(\mathrm{p})}\right) = \sqrt{\sum_{k=1}^{K_{\mathrm{p}}} \left(\frac{n_0\left(M_{\mathrm{k}}^{(\mathrm{p})}\right)}{N^{(\mathrm{p})}} - \frac{n_{\mathrm{u}\geq 1}\left(M_{\mathrm{k}}^{(\mathrm{p})}\right)}{N^{(\mathrm{p})}} \right)^2}. \tag{3.20}$$

Several test datasets $(u \geq 1)$ were generated using Equation (3.17) and obtained a set of test dataset clustering results.

$$\mathcal{R}^{(\mathrm{p})} \coloneqq \left\{ \boldsymbol{n}_1^{(\mathrm{p})}, \boldsymbol{n}_2^{(\mathrm{p})}, \ldots, \boldsymbol{n}_{\mathrm{h}}^{(\mathrm{p})} \right\}, \tag{3.21}$$

where *h* denotes the number of the test datasets. Finally, the difference score was defined as the mean value of $\delta\left(\boldsymbol{n}_0^{\prime(\mathrm{p})}, \boldsymbol{n}_{\mathrm{u}\geq 1}^{\prime(\mathrm{p})}\right)$ for $u \geq 1$ as follows. A lower difference score indicates a better robustness.

$$\bar{\delta}_{\mathcal{R}^{(\mathrm{p})}} \coloneqq \frac{1}{h} \sum_{u=1}^{h} \delta\left(\boldsymbol{n}_0^{\prime(\mathrm{p})}, \boldsymbol{n}_{\mathrm{u}}^{\prime(\mathrm{p})}\right). \tag{3.22}$$

3.3 Results and Discussion

3.3.1 Cell Tracking

The interaction between mesenchymal and epithelial cells is an important driving force for collective cell migration. Interactions promote the transmission of signals between cells, which guides the migration of follower cells. Thus, multiple cells follow the leader's migration. Here, the influence of leader cells was investigated by observing the cell population in a model in which TGF (+) cells (red labeled) and TGF (−) cells coexist (Figure 3.3). Figure 3.3A,B indicate that the shape of the colony changed gradually from a circle to an ellipse, suggesting that the cell migration was not random but anisotropic due to the aligned electrospun fibers. The migration of each cell was one-directional toward the outside of the colony, which might be due to the local gradient of the soluble factors secreted by cells because the cell culture medium had no intentional mechanism to cause a chemotactic gradient; it is beyond the scope of this study but warrants further investigation.

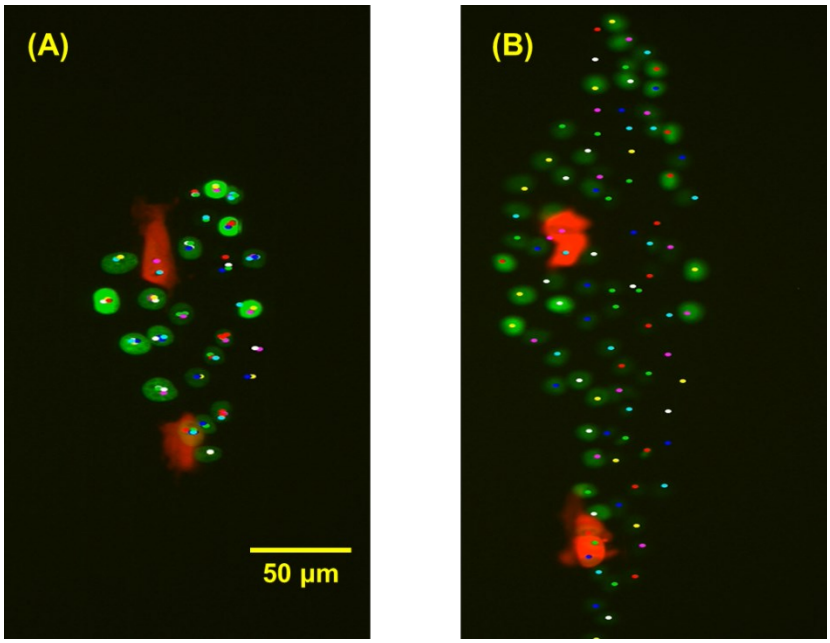


Figure 3.3 Manual markers of cell position at (A) time slice 1 and (B) time slice 96. The TGF (+) and TGF (-) cells are labelled by red and green, respectively.

Table 3.1 Example of manual tracked coordinates of a cell from slice from 1 to slice 96 (24 hours).

Slice	X	Y	Slice	X	Y	Slice	X	Y
1	792	315	33	790	248	65	789	101
2	792	313	34	787	241	66	789	96
3	791	309	35	789	240	67	788	96
4	789	302	36	788	234	68	786	91
5	791	297	37	786	226	69	791	92
6	790	300	38	786	219	70	786	85
7	789	293	39	786	213	71	787	83
8	786	294	40	785	205	72	789	81
9	788	289	41	785	207	73	790	81
10	791	285	42	786	203	74	786	76
11	793	281	43	787	199	75	788	73
12	794	276	44	786	193	76	785	74
13	791	275	45	784	187	77	787	79
14	789	268	46	784	184	78	785	77
15	790	277	47	787	184	79	786	70
16	790	276	48	788	175	80	786	70
17	792	273	49	787	170	81	787	63
18	794	272	50	787	162	82	786	60
19	793	264	51	787	154	83	785	55
20	797	262	52	786	151	84	785	55
21	795	259	53	786	144	85	784	55
22	795	254	54	785	141	86	786	64
23	795	256	55	785	133	87	788	75
24	793	254	56	788	129	88	784	73
25	795	254	57	788	131	89	783	72
26	795	247	58	784	127	90	781	66
27	788	241	59	786	123	91	782	56
28	787	237	60	785	124	92	787	59
29	787	231	61	787	121	93	789	59
30	790	226	62	788	116	94	786	56
31	790	220	63	790	112	95	790	51
32	789	221	64	787	107	96	787	50

At this time, TGF (+) cells were labeled with the red fluorescent protein DsRed to distinguish them from TGF (−). A population of 10–20 cells was observed because its size was the same as that of a colony *in vivo*. In the first time slice, there were 28 cells (Figure 3.3A). In the last time slice, 92 cells were identified (Figure 3.3B). Table 3.1

shows manual labeling coordinates of one cell. These data were used as the input data in the Python codes, refer to the Appendix 2, Part 1. Since the manual tracking started from the last time slice to the first time slice backward, 92 marker points of the cell nucleus were found in each time slice and assigned the ID number. In the cell division event, the cell before the division was defined as the mother and the two cells after division as daughter cells.

Cell division occurs during cell migration. For analysis, it is crucial to identify the correspondence between mother and daughter cells and the time at which cell division occurs. For all the marker points in each time slice, the method calculated the Euclidean distance between each pair of marker points. A sudden change in the distance indicates that cell division had taken place (Figure 3.2). The time-lapse observation recorded the image slices every 15 min. The distance between the two adjacent slices is defined as the step size. For a representative cell, the step size distribution for all pairs of slices is shown in Figure 3.4A. The maximum step size was 6 μm. The distances between the four groups of daughter cells are shown in Figure 3.4B. The graph shows a clear "distance jump" in the curve, where the distance before the "jump" is always very short. The "distance jump" indicates the occurrence of cell division. After the "jump" a completely different distance curve appears. The Euclidean distance between the marker points was greater than 10 μm after cell division. The maximum moving step size of 6 μm is below the "jump distance" of 10 μm, which means that the accuracy is acceptable for identifying when cell division occurs. In combination with fluorescence images, the division relationship between cells can be determined. divisions were identified based on the calculation of the distance between the cell marker points and manually labeled cells. The marker points of the mother and two daughter cells were unified prior to cell division.

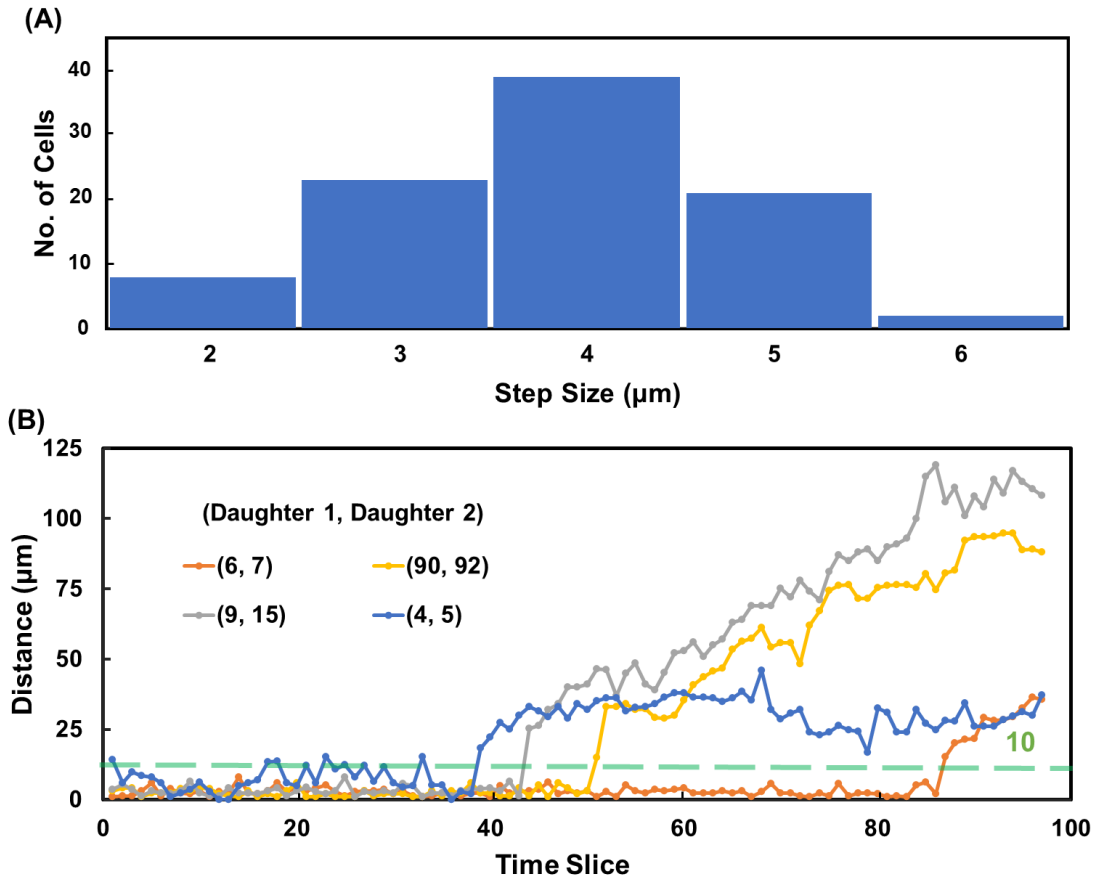


Figure 3.4 (A) The distribution of average moving step size of each cell before division. (B) Euclidean distance between two daughter cells.

3.3.2 Dimension Reduction and Clustering

After the cell tracking, the entire migration process was divided into multiple periods. The trajectories of TGF (+) (black) and TGF (−) (blue) cells during different periods are shown in Figure 3.5. Two groups of TGF (+) cells were observed from the initial position and marked separately. The cell trajectories within each period were normalized. The purpose of normalization is to eliminate the effect of location (Figure 3.6), where “normalization” refers to translating all cell trajectories such that their starting points are located at the origin (0, 0), refer to the Appendix 2, Part 3.

The normalized cell trajectories are used to calculate the distance matrix, which is a series of high-dimensional datasets for which the method need to reduce the

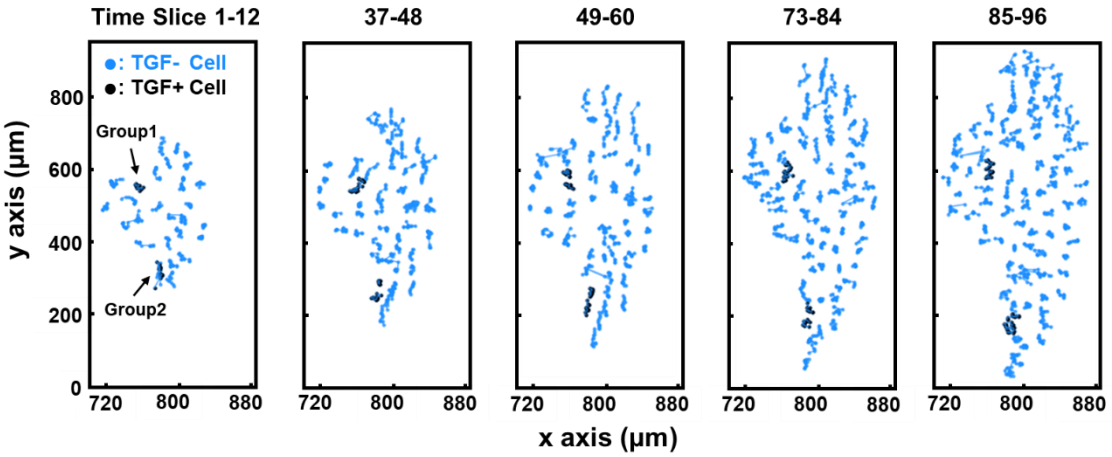


Figure 3.5 The positions of TGF (+) (black) and TGF (-) (blue) cells in different observation windows.

dimensionality for visualization. There are two reasons for reducing the dimension of the data before performing k-means clustering, in addition to visualization. The first reason is data summarization. This method aims to find a low-dimensional structure from high- dimensional cell-trajectory data. Reducing the dimensions before performing k-means clustering is beneficial for capturing the low-dimensional structures. The second reason is robustness. Dimension reduction can help reduce the noise included in high-dimensional data and improve the robustness of this method. Dimension reduction before performing downstream analysis, such as clustering, is also used in single-cell RNA-sequencing data analysis [19, 20]. This method chose two as the number of the dimensions to ensure consistency with the visualization which are

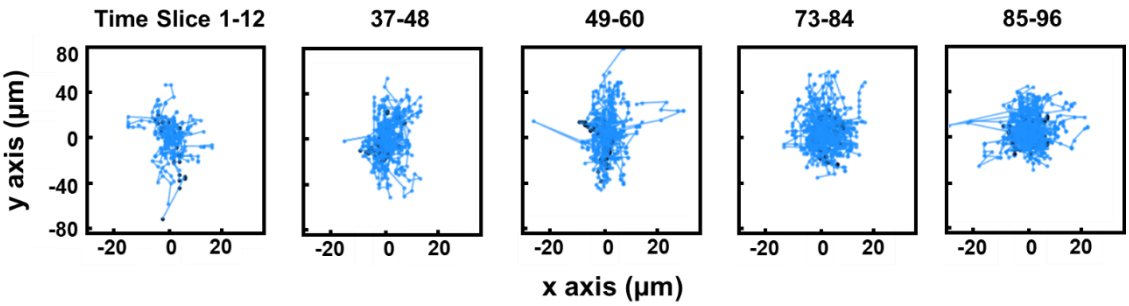


Figure 3.6 The positions of TGF (+) (black) and TGF (-) (blue) cells after normalization, the initial position transferred to (0,0).

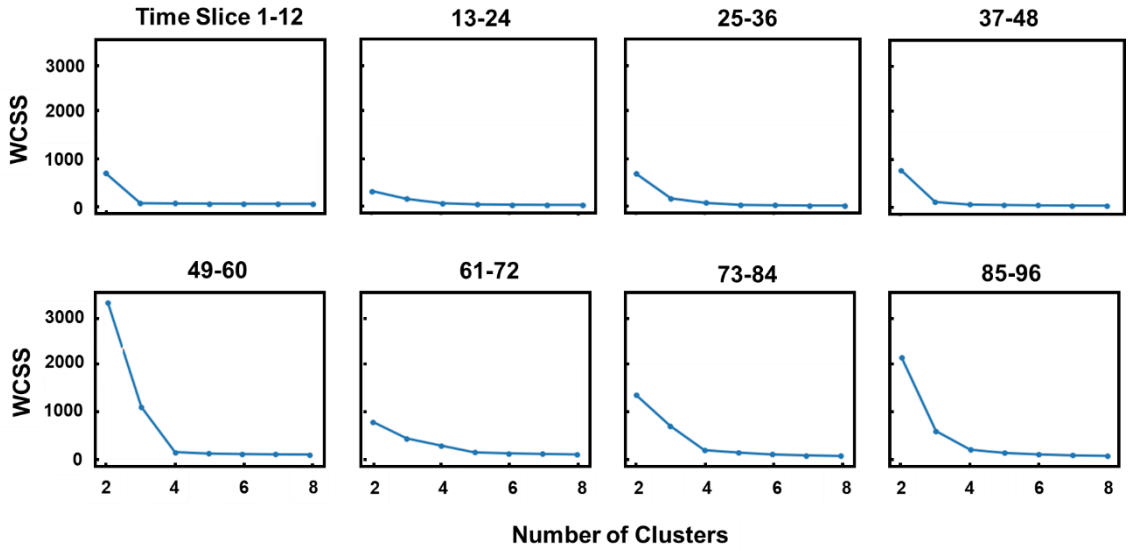


Figure 3.7 WCSS of all samples with the number of k-means clusters of UMAP dimensionality reduction results under different observation periods.

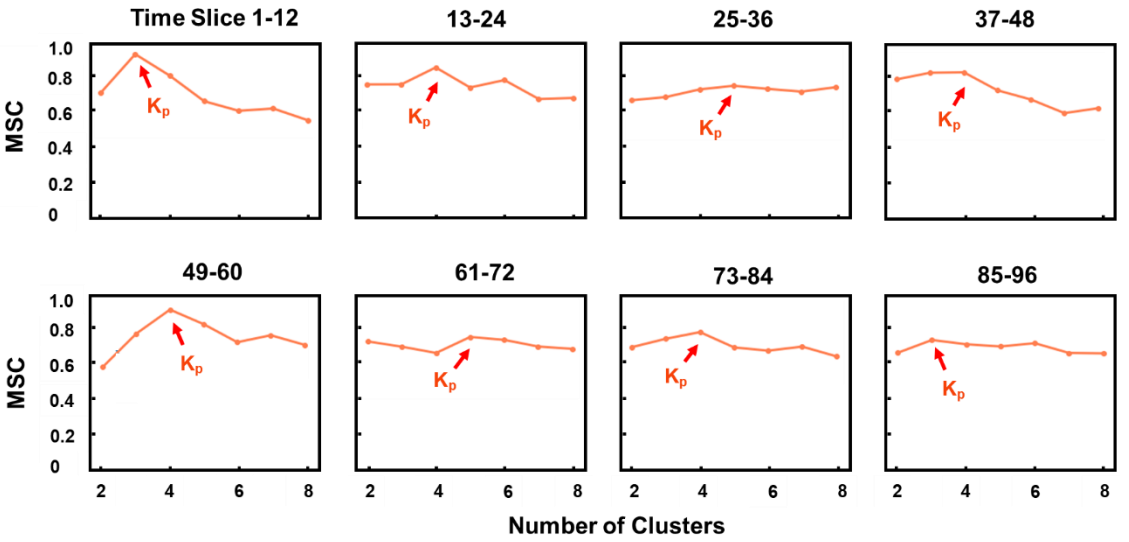


Figure 3.8 MSC of all samples with the number of k-means clusters of UMAP dimensionality reduction results under different observation periods. The K_p value shown by the red arrows are the optimal numbers of clusters.

discussed in Chapter 4. Based on this, the dimension reduction was conducted by the UMAP algorithm. It visualized the similarities in two-dimensions. At the same time, the global structures in high-dimension can be still retained in low-dimension. As the result, each point represents a trajectory. TGF (−) cells and TGF (+) cells were marked

separately. Then all the points will be clusters by k-means algorithm.

The clustering method used the k-means algorithm, k represents the number of clusters. Before conducted clustering process, the cluster number k value should be first defined. Here two methods were used to optimize the k value. The first is WCSS of distance, which result with different k is shown in Figure 3.7. The objective is to minimize the WCSS. The result show that the WCSS is decrease with the increase of k value, which is hard to select an optimal k. Then the Mean silhouette coefficient method, MSC, was combine to determine the appropriate number of clusters (Figure 3.8). MSC can check whether a point is suitable for its cluster. It will calculate the similarity within clusters and between clusters. MSC $\bar{s}(k)$ is in the range of $(-1,1)$; therefore, the closer $\bar{s}(k)$ is to 1, the more reasonable the clustering is, at which point k is the appropriate number of clusters. After a series of calculation, the k value in each observation windows were obtained.

3.3.3 Similarity of Migration Patterns

Figure 3.9 shows the dimensionality reduction results for several periods with TGF (−) cells (green) and TGF (+) cells (red and blue). In each period in Figure 3.9, it is clear that TGF (+) cells of the same group were not always in the same cluster at the same time. TGF (+) and TGF (−) cells assigned to the same cluster showed high similarity and maintained the same migration pattern in terms of collective cell migration. TGF (+) cells with more active migratory behavior exert some influence on TGF (−) cells. TGF (+) cells with a leading role showed interactions with TGF (−) cells through signaling. Pathways such as PI3K-Rac signaling are involved in actin remodeling and mediation of collective migration. Through PI3K-dependent integrin adhesion and modulation of Rho-GTPase signaling, cadherin-induced regulation of actomyosin contractility at more distant sites in the cell affects global cellular mechanics [21, 22]. The present analysis found that cells around TGF (+) cells showed

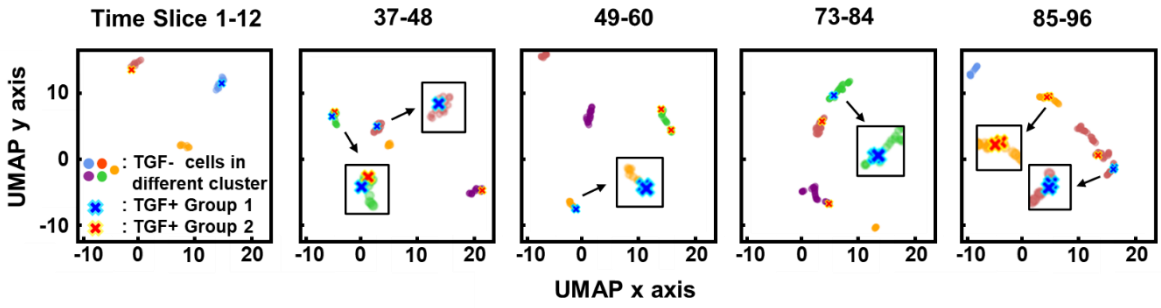


Figure 3.9 Visualization of cell tracks after dimensionality reduction. 2D visualization of TGF (+) ("X" mark) and TGF (-) cell (colored points) trajectories after dimensionality reduction by UMAP and clustering by k-means. Each mark and point represent a trajectory. Colors represent different clusters. The insets show the magnified images of TGF (+) cells when they overlap.

the same migration pattern, and they might have been under the influence of the leader cell. The behaviors observed here may correspond to reports that the leader cell leads the movement of the cell population such that the surrounding follower cells are in tow. Therefore, the methodology suggested in the thesis could be used to detect leader cells within a cell population. This maintained them in the same overall migration direction and distance.

In addition to clusters containing TGF (+) cells, some clusters contained only TGF (−) cells, illustrating that their migration pattern is different from that of TGF (+) cells. TGF (+) cells, as leaders of collective migration, affect nearby TGF (−) cells; however, the effect is relatively narrow. TGF (−) cells with fewer or no interactions migrated based on their migratory patterns.

3.3.4 Positional Similarity

It is expected that the influence of leader cells on follower cells decreases with distance and is stronger only within a certain range. Within the effective range, the interaction strengthened the connections between them. To verify these results, the clustering results were superimposed onto a cell location map. During the previous

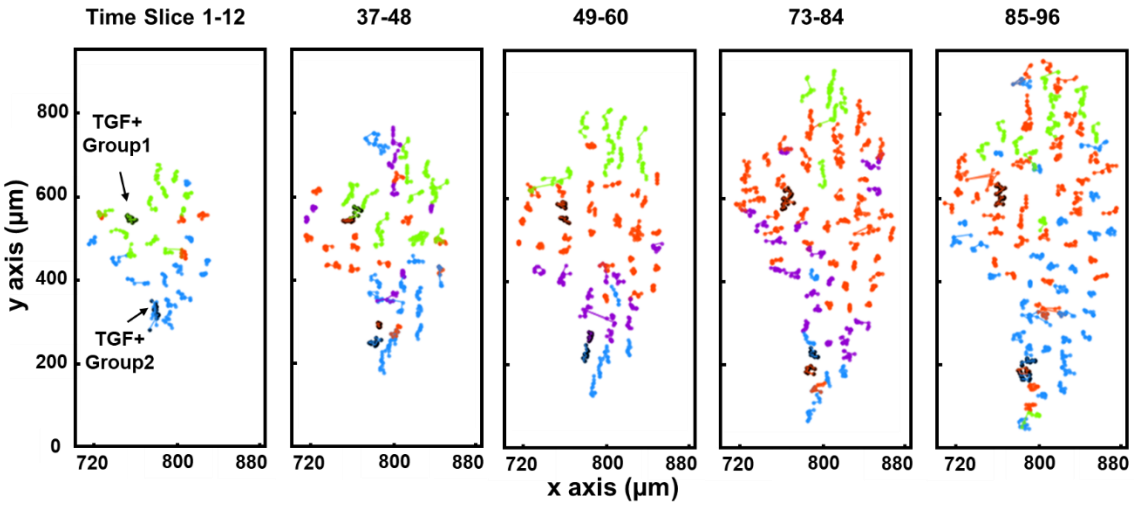


Figure 3.10 Cell trajectories' positions combined with clustering results. Two groups of TGF (+) cells are marked black pointed by arrows. Colors represent different clusters.

normalization process, the location information of the cell trajectories was reset and the similarity of the cell trajectories was determined from the clustering results. However, it is impossible to determine whether cell trajectories with similarities are close to each other in terms of their actual locations. Therefore, the clustering results and actual locations were combined, as shown in Figure 3.10. Notably, the clustering results in Figure 3.10 are the same as those in Figure 3.9; however, the color was adjusted for ease of comparison. The data points for TGF (+) cells in Figure 3.10 are marked in black, whereas the lines between the points remain colored.

Interestingly, after combining the clustering results with actual locations, most cells from the same cluster showed positional similarities. This method reproduces positional similarity, even after normalization eliminates the location information. This worked for most cells, except for individual cells, such as the green cluster in Figure 3.10 slice 37-48, which contained cells belonging to the other clusters.

The aligned fiber was used to mimic the ECM in an *in vitro* model. This environment closely resembles the environment inside an organism. The migration direction of some cells may be influenced by fiber alignment; however, after normalization, the effect of position was eliminated. Normalization makes the method

reliable and the trajectory independent of the ECM. In addition, the calculation of this method treats all cells equally and does not distinguish between leader and follower cells, but the results show cells with similar migration patterns. After superimposing the clustering with the actual locations, the results show that in most cases, there were many TGF (−) cells around TGF (+) cells. From these results, clusters are not related to cells moving in a limited direction; they are related to the similarity of migration. TGF (+) cells were mostly not in the center of each cluster but at the edge of the cluster, which is consistent with the characteristics of the leader cells.

3.3.5 Robustness

Whether this method is effective in identifying cells with similar migration patterns requires further validation. During manual marking, the accuracy of the marked cell nucleus position determines the degree of error. The robustness of the clustering results was verified by adding noise to the cell trajectories. The test data were generated by adding Gaussian noise with a fixed mean and standard deviation, σ (Figure 3.11A). Figure 3.11B shows the original positions of the cell trajectories for time slices 49–60,

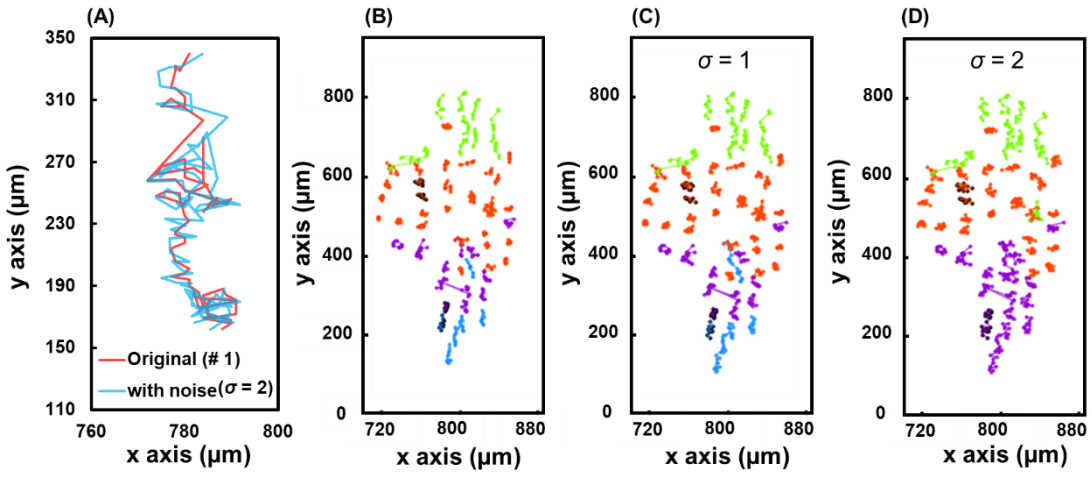


Figure 3.11 The original trajectory of Cell ID 1 (red) and the trajectory with Gaussian noise of standard deviation $\sigma = 2$ (blue). (B) Cell tracks' positions combined with clusters' result under time slice 49-60. (C)-(D) All trajectories are with noise of standard deviations = 1, 2.

and Figure 3.11C, D show the clustering results with $\sigma = 1$ and 2, respectively. A change in σ causes the similar trajectories to become dissimilar. The boundaries between clusters became indistinct and k increased, rendering the clustering meaningless. The maximum value of k for the original clustering results was 5. Therefore, the maximum value of $k = 5$ was applied to noise-containing clustering analysis. The trajectory difference score of the clustering results with noise compared to those without noise is shown in Figure 3.12B. Notably, the cluster color order (cluster index) changed after each adjustment of the σ value. Therefore, the ID of each trajectory in each cluster was recorded and then adjusted the color order as a whole to ensure reasonableness (Figure 3.11B).

As shown in Figure 3.11B, C, D, the distribution of cluster positions did not differ

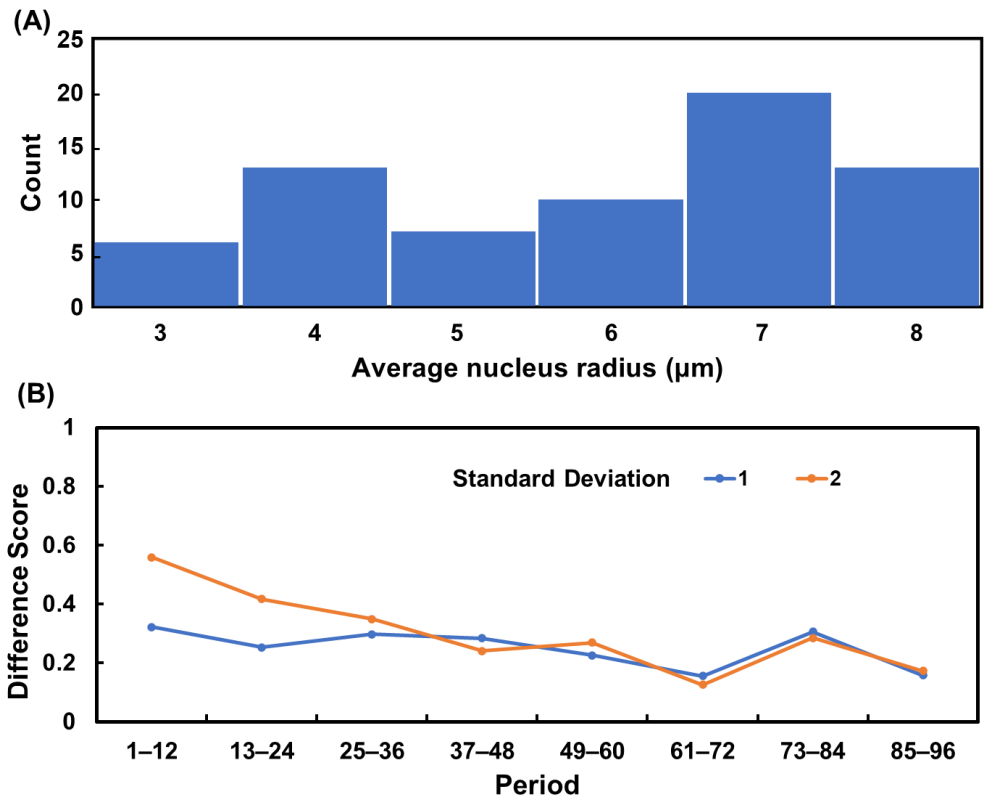


Figure 3.12 Robustness of the method. (A) Histogram of average radius of cell nucleus (μm). (B) The difference scores of all clustering results under differ standard deviation (from the mean of three calculations in each σ). The sample sizes for each time period are 92 cells.

significantly and only the clusters of individual cells changed. The observation error is mainly due to the accuracy of manual labeling. The radius of the nucleus was below 9 μm (Figure 3.12A). Considering the accuracy, the manual labeling of the center of the cell nucleus, $\sigma = 2$ was sufficiently large for the test. The clustering results with noise and without noise were comparable in consistency. This difference may be related to noise altering the original migration trajectory. This is because the change in trajectory affects the migration pattern, including speed and direction. The internal connections between cells that originally belonged to the same cluster are broken. When σ is greater, noise causes severe distortion of the trajectory. Nevertheless, there were no significant difference in the position of each cluster. Therefore, the method is robust to observation errors that may be contained in the cell trajectory data.

3.4 Conclusion

Based on trajectories, a method was proposed to make clustering. The method is combined with UMAP dimensionality reduction and k-means clustering algorithm of machine learning. And succeeded preform the positional similarity. TGF (+) and TGF (−) cells belonging to the same cluster showed similar migration patterns and, within a certain range, migration was consistent. the robustness of the method was validated, it showed stable results even with noise.

Interactions between cells are essential for coordinated and directed collective movements. Although, collective cell migration has been extensively studied, this study focused on exploring the similarity of migration patterns based on migration trajectories. Using a combination of dimensionality reduction and clustering techniques, cells with similar migration patterns were observed to exhibit positional similarities. TGF (+) and TGF (−) cells belonging to the same cluster showed similar migration patterns and, within a certain range, migration was consistent. Validation by adding noise illustrated the robustness of the proposed method. This provides a new perspective for a deeper

understanding of collective cell migration. The method applies to collective cell migration on aligned fibers and models such as wound healing and migration on other substrates. It can also be extended to the cell migration model in a 3D hydrogel matrix if cells can be tracked, indicating its application in tissue engineering and organ development research.

REFERENCE

[1] Morrison, J. A.; McLennan, R.; Wolfe, L. A.; Gogol, M. M.; Meier, S.; McKinney, M. C.; Teddy, J. M.; Holmes, L.; Semerad, C. L.; Box, A. C.; Li, H.; Hall, K. E.; Perera, A. G.; Kulesa, P. M., Single-cell transcriptome analysis of avian neural crest migration reveals signatures of invasion and molecular transitions. *Elife* **2017,** 6. doi: 10.7554/eLife.28415.

[2] Becsky, D.; Szabo, K.; Gyulai-Nagy, S.; Gajdos, T.; Bartos, Z.; Balind, A.; Dux, L.; Horvath, P.; Erdelyi, M.; Homolya, L.; Keller-Pinter, A., Syndecan-4 Modulates Cell Polarity and Migration by Influencing Centrosome Positioning and Intracellular Calcium Distribution. *Front Cell Dev Biol* **2020,** 8, 575227. doi: 10.3389/fcell.2020.575227.

[3] Kim, J. M.; Lee, M.; Kim, N.; Heo, W. D., Optogenetic toolkit reveals the role of Ca2+ sparklets in coordinated cell migration. *Proc Natl Acad Sci U S A* **2016,** 113, (21), 5952-7. doi: 10.1073/pnas.1518412113.

[4] Capuana, L.; Bostrom, A.; Etienne-Manneville, S., Multicellular scale front-to-rear polarity in collective migration. *Curr Opin Cell Biol* **2020,** 62, 114-122. doi: 10.1016/j.ceb.2019.10.001.

[5] Sikka, G.; Rani, S., Recent Techniques of Clustering of Time Series Data: A Survey. *International Journal of Computer Applications* **2012,** 52, (15), 1-9. doi: 10.5120/8282-1278.

[6] Hozumi, Y.; Wang, R.; Yin, C.; Wei, G. W., UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. *Comput Biol Med* **2021,** 131, (December 2020), 104264. doi: 10.1016/j.compbiomed.2021.104264.

[7] Niennattrakul, V.; Ratanamahatana, C. A. In *Inaccuracies of shape averaging method using dynamic time warping for time series data*, Berlin, Heidelberg, **2007**, Shi, Y.; van Albada, G. D.; Dongarra, J.; Sloot, P. M. A., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, pp 513-520.

[8] Wold, S.; Esbensen, K.; Geladi, P., Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **1987,** 2, (1-3), 37-52. doi: 10.1016/0169-7439(87)80084-9.

[9] Mead, A., Review of the Development of Multidimensional Scaling Methods. *The Statistician* **1992,** 41, (1), 27-27. doi: 10.2307/2348634.

[10] van der Maaten, L.; Hinton, G., Visualizing Data using t-SNE. In *J Mach Learn Res*, **2008**; Vol. 9, pp 2579-2605.

[11] McInnes, L.; Healy, J.; Melville, J., Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**, https://arxiv.org/abs/1802.03426.

[12] Fujita, A.; Severino, P.; Kojima, K.; Sato, J. R.; Patriota, A. G.; Miyano, S., Functional clustering of time series gene expression data by Granger causality. *Bmc Syst Biol* **2012,** 6, 137. doi: 10.1186/1752-0509-6-137.

[13] Pyatnitskiy, M.; Mazo, I.; Shkrob, M.; Schwartz, E.; Kotelnikova, E., Clustering gene expression regulators: new approach to disease subtyping. *PLoS One* **2014,** 9, (1), e84955. doi: 10.1371/journal.pone.0084955.

[14] Althuwaynee, O. F.; Aydda, A.; Hwang, I. T.; Lee, Y. K.; Kim, S. W.; Park, H. J.; Lee, M. S.; Park, Y., Uncertainty Reduction of Unlabeled Features in Landslide Inventory Using Machine Learning t-SNE Clustering and Data Mining Apriori Association Rule Algorithms. *Appl Sci (Basel)* **2021,** 11, (2). doi: 10.3390/app11020556.

[15] Hussain, S. M.; Buongiorno, D.; Altini, N.; Berloco, F.; Prencipe, B.; Moschetta, M.; Bevilacqua, V.; Brunetti, A., Shape-Based Breast Lesion Classification Using Digital Tomosynthesis Images: The Role of Explainable Artificial Intelligence. *Appl Sci (Basel)* **2022,** 12, (12). doi: 10.3390/app12126230.

[16] Lonseko, Z. M.; Adjei, P. E.; Du, W. J.; Luo, C. S.; Hu, D. C.; Zhu, L. L.; Gan, T.; Rao, N. N., Gastrointestinal Disease Classification in Endoscopic Images Using Attention-Guided Convolutional Neural Networks. *Appl Sci (Basel)* **2021,** 11, (23). doi: 10.3390/app112311136.

[17] MacQueen, J.; et al., *Some methods for classification and analysis of multivariate observations*. California, **1967**; Vol. 1, p 281-297.

[18] Rousseeuw, P. J., Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **1987,** 20, (C), 53-65. doi: 10.1016/0377-0427(87)90125-7.

[19] Luecken, M. D.; Theis, F. J., Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **2019,** 15, (6), e8746. doi: 10.15252/msb.20188746.

[20] Moon, K. R.; Stanley, J. S.; Burkhardt, D.; van Dijk, D. v.; Wolf, G.; Krishnaswamy, S., Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology* **2018,** 7, 36-46. doi: 10.1016/j.coisb.2017.12.008.

[21] Kolsch, V.; Charest, P. G.; Firtel, R. A., The regulation of cell motility and chemotaxis by phospholipid signaling. *J Cell Sci* **2008,** 121, (Pt 5), 551-9. doi: 10.1242/jcs.023333.

[22] Khalil, A. A.; de Rooij, J., Cadherin mechanotransduction in leader-follower cell specification during collective migration. *Exp Cell Res* **2019,** 376, (1), 86-91. doi: 10.1016/j.yexcr.2019.01.006.

CHAPTER 4

OPTIMIZATION OF TIME-SERIES CLUSTERING PARAMETERS AND CONTROL OF CELL CONDITIONS

Abstract

This chapter shows the optimization process of the dimension reduction algorithm by changing their hyperparameters, the optimization of observation window. Based on the optimal parameters, the time-series clustering was conducted to different ratio of mesenchymal and epithelial cells.

4.1 Introduction

Machine learning has a wide range of applications, and there are opportunities for machine learning algorithms in both the biological and engineering fields [1-3]. The terms "data mining" and "data analysis" are often used in conjunction with each other and are considered to be interchangeable in many contexts. There are various definitions of data mining with different words but close meanings, such as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in huge amounts of data". Both data analysis and data mining help people collect and analyze data, make it into information, and make judgments, so they can be referred to together as data analysis and mining [4-6].

Sequence data will be inter-integrated with structural and functional data gene expression data, biochemical response pathway data phenotypic and clinical data, and a host of other data [7, 8]. Such a large amount of data presents an urgent need for the development of theoretical algorithms and software in the storage, acquisition,

processing, browsing, and visualization of biological information [9-11]. In addition, the complexity of genomic data itself also poses an urgent need for the development of theoretical algorithms and software. Machine learning methods such as neural networks, genetic algorithms, decision trees, and support vector machines are suitable for dealing with such a large amount of data, noise, and lack of unified theory [12-14]. PCA is one of the most widely used linear dimensionality reduction algorithms for data. the main idea of PCA is to map n-dimensional features to k (where $k<n$) dimensions, which are brand new orthogonal features also known as principal components, which are k-dimensional features reconstructed on the basis of the original n-dimensional features, aiming to use the idea of dimensionality reduction to t-SNE is a machine learning algorithm for dimensionality reduction. t-SNE is a nonlinear dimensionality reduction algorithm, which is very suitable for visualizing high-dimensional data down to 2 or 3 dimensions. UMAP [15] is a dimensionality reduction algorithm for high-dimensional data similar to t-SNE. UMAP has two main advantages over t-SNE, it can reflect the global structure, and it runs faster and takes less memory when computing large sample data.

Unsupervised learning is a method of machine learning, as opposed to supervised learning. In practice, supervised learning requires that the true value of the predicted quantity be provided for each sample in training, *i.e.*, labeling the training samples, which is difficult in some applications [16, 17]. For example, in medical diagnosis, to obtain a diagnostic model by supervised learning, a large number of cases and their medical images need to be accurately labeled, which is labor-intensive and inefficient. In this case unsupervised learning methods are usually used, *i.e.*, learning is performed under the condition that no supervised information (the real value of the predicted quantity) is provided.

In unsupervised learning, all data are unlabeled, but these data exhibit a cluster structure, where similar types of data are clustered together [18]. Dividing these unlabeled data into a combination is clustering. In the field of machine learning,

dimensionality reduction refers to reducing the number of random variables under certain constraints [19-22]. Dimensionality reduction can be further subdivided into two major methods: variable selection and feature extraction [21, 22]. Variable selection refers to finding the main variables among the original variables when the data contains a large number of redundant or irrelevant variables, so as to simplify the model and make it easier to be learned by the machine. When the input information of an algorithm is too complex and the number of variables is too large, proper feature extraction is the key to construct an effective model for learning. Feature extraction is the process of constructing informative and non-redundant feature values from the original data, which can help in the subsequent learning process and generalization steps, where the initial data set is reduced to more manageable groups for learning, while maintaining the accuracy and completeness of the description of the original data [23, 24].

In the context of machine learning, hyperparameters are parameters whose values are set before starting the learning process, rather than the parameter data obtained through training [25-28]. Often, optimization of hyperparameters is required to select an optimal set of hyperparameters for the learning machine in order to improve the performance and effectiveness of learning [29]. Machine learning model tuning is an optimization problem. There are a set of hyperparameters and my goal is to find the right combination of their values, which can help the method find the minimum (*e.g.*, loss) or maximum (*e.g.*, accuracy) of the function.

This chapter involving optimization aims to find the hyperparameters of algorithm and length of observation window that make the time-series clustering perform best. In the clustering algorithm for cell migration trajectories mentioned in Chapter 3, the ratio of TGF (+) cells to TGF (−) cells is fixed. This chapter optimized the algorithm by comparing the clustering results generated by the hyperparameters of the dimensionality reduction algorithm and apply them to different proportions of both cells to validate the algorithm in finding similar migration pattern.

4.2 Method

4.2.1 Optimization of time-series clustering

In the algorithm proposed in Chapter 3, dimensionality reduction is one of the most important steps. Through the high-dimensionalities of single image data, a single image is transformed into a collection of data in a high-dimensional space, and its non-linear dimensionality is reduced to seek the low-dimensional representation vector of its high-dimensional data manifold structure, which is used as the feature expression vector of the image data. Thus, the problem of high-dimensional image recognition is transformed into the recognition problem of feature expression vector, which greatly reduces the complexity of calculation, reduces the recognition error caused by redundant information, and improves the recognition accuracy. The application of the nonlinear dimensionality reduction method to the image data recognition problem is feasible in practice, computationally simple, and can greatly improve the effectiveness of commonly used methods and obtain better recognition results.

Visualization of high-dimensional data is very important, people can only understand two-dimensional three-dimensional data, so from the high-dimensional data by means of reducing the dimensionality to make people easy to understand. Commonly used dimensionality reduction methods such as UMAP and t-SNE both have the same idea that the relative information of high-dimensional data samples is reflected in the lower dimensionality [30-34]. The advantage of UMAP over t-SNE is its ability to reflect both local and global structures while maintaining relative global distances in a low-dimensional space. Each of the two-dimensionality reduction algorithms has its own characteristics, and their hyperparameters determine the effect of clustering. Therefore, the optimization uses different ("perplexity" in t-SNE; "n_neighbors" in UMAP " and "min_dist") parameter combinations to run the dimensionality reduction algorithm and compare the results.

4.2.2 Time-series clustering at different cell ratios

In the clustering analysis in Chapter 3, the ratio of two types of cells was fixed and finally found that the clustering of cell trajectories could reflect the role of cells. To further validate the generalizability of the method in order to further reveal the interactions between leader and follower cells, the analysis using different data sets were performed. The cell colonies with different TGF (+)/TGF (−) ratios are described in Chapter 2. Cell migration images obtained by time-lapse microscopy at different scales were used in a time-series clustering algorithm.

The migration process was divided into different time slices and each time slice recorded the position of the cells on the fiber. The observation window consists of several time slices, so the whole migration process can be divided into several observation windows. The setting of the length of the observation window is crucial, which determines the final clustering result. An appropriate observation window correctly reflects the cell-cell interactions and the current cell state. An inappropriate time window distribution may lead to difficulties in finding the correct migration pattern. Therefore, the results of different observation windows were compared and optimized them in combination with the algorithm parameters.

4.3 Result and discussion

4.3.1 Effect of parameters on clustering

The clustering similarity measure is geometric distance, commonly used distances are Euclidean distance, Manhattan distance. The principle of partition-based clustering is that given a data set containing N points, the partitioning method will construct k groupings; each grouping represents a cluster, where each grouping contains at least one data point, and each data point belongs to one and only one grouping; for a given value of k, the algorithm first gives an initialized grouping method, and then changes

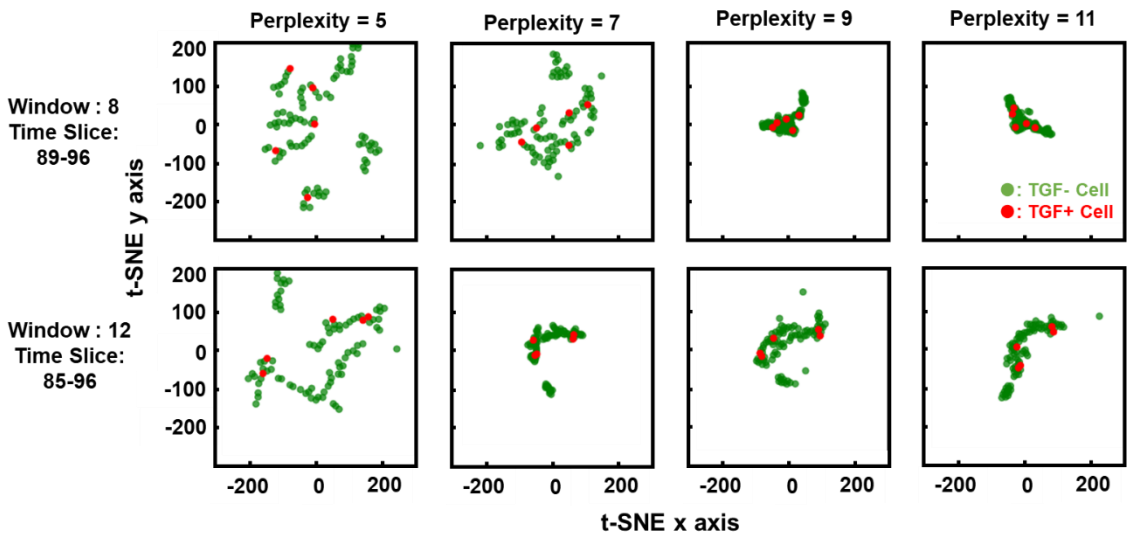


Figure 4.1 The dimensionality reduction of cell trajectories by t-SNE under different parameters and observation windows.

the groupings by iterative methods until the criterion function converges. The commonly used k-means clustering method is simple and intuitive, easy to implement and takes relatively little time to compute. k-means produces clusters that are relatively tight, reflecting the closeness of the samples within the cluster around the center of mass. However, it is difficult to predict the exact number of clusters and is sensitive to the initial value setting. k-means mainly finds circular or spherical clusters and does not work well for clusters of different shapes and densities. The method to optimize the *k* value of each observation window is important which was described in Chapter 3.

The capabilities of dimension reduction allow me to obtain appropriate clustering results. The method calculated using the t-SNE for comparison (Figure 4.1). The results of t-SNE show no obvious regularity of the data points. When the hyperparameter "perplexity" is 5 or 7, the two-dimensional visualization results show more scattered data points and no obvious aggregation characteristics. When the "perplexity" is too large, the data points are too concentrated in one cluster and do not show the aggregation results of different clusters. Obvious clustering and more concentrated data within clusters are suitable results for dimensionality reduction. As a comparison,

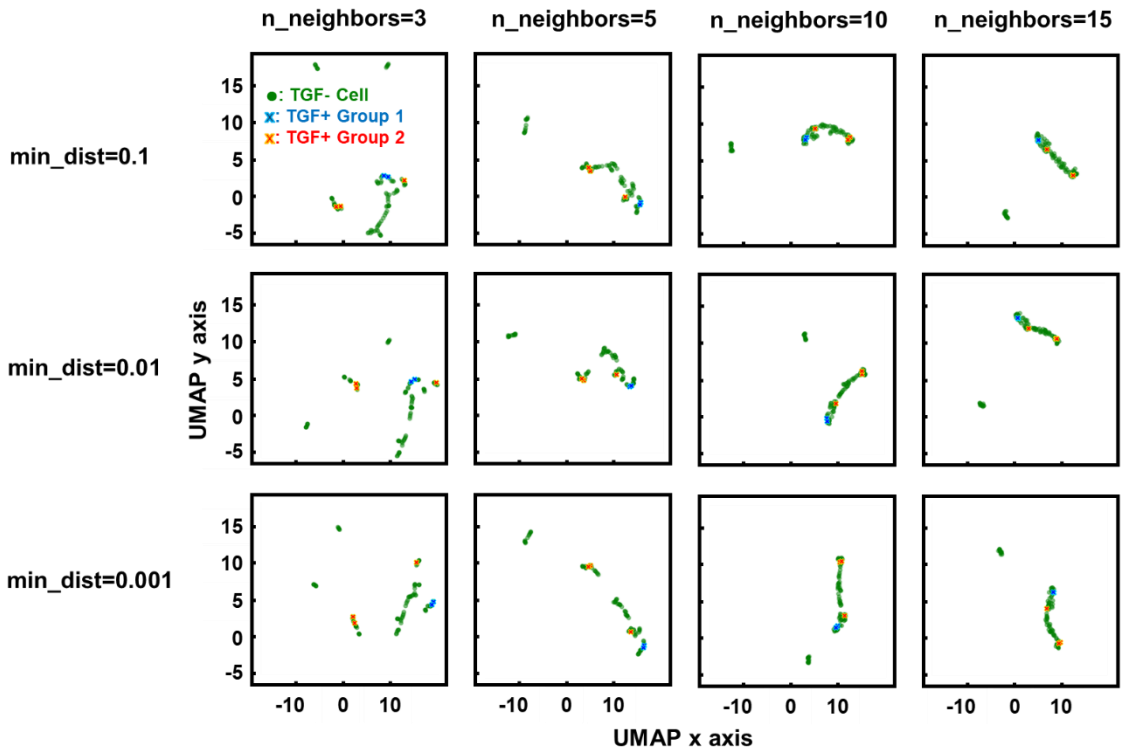


Figure 4.2 The dimensionality reduction of cell trajectories under different UMAP parameters (observation window = 12, time slice 85-96).

UMAP showed a more apparent aggregation of data points in the same cluster(Figure 4.2). In addition, there were more apparent boundaries between the clusters.

The optimal observation window was selected such that the distance between the clusters was larger in the first and last time periods and the data points within the clusters were closer together (Figure 4.3). An observation window of 12 time slices was selected. As shown in Figure 4.2, the parameters for UMAP were chosen, such that the distance between clusters was larger and the data points within the clusters were closer.

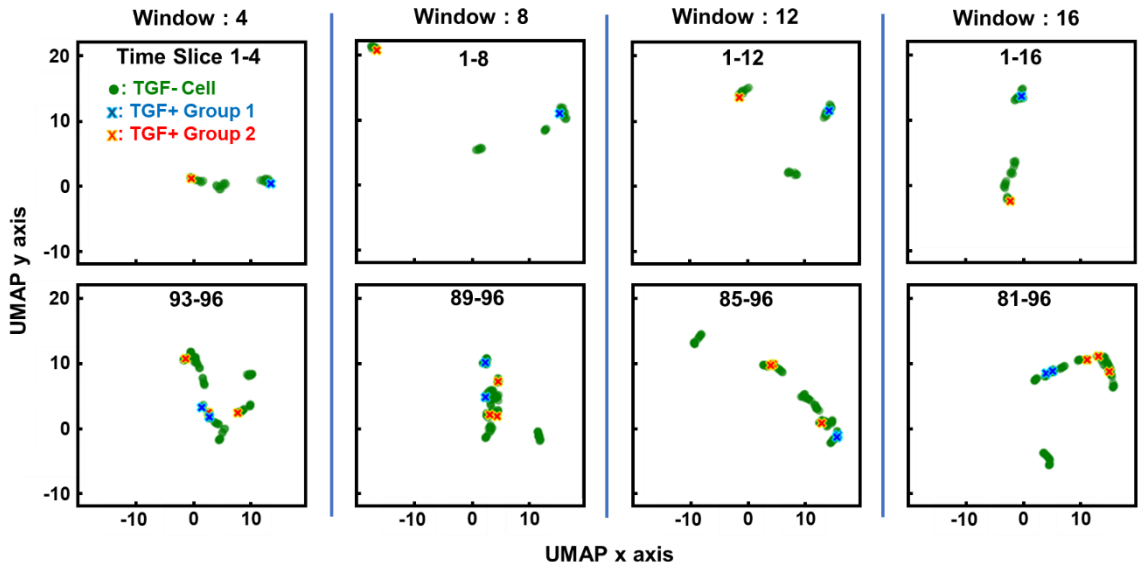


Figure 4.3 Cell tracks after dimensionality reduction when select different observation windows. The first and last periods showed that window 12 has clear boundaries of clusters and TGF (+) cells in the same group tend to in the same cluster.

4.3.2 Clustering of different cell ratios

The cell trajectories in Chapter 3 were analyzed using a time-series clustering approach with a ratio of TGF (+) and TGF (−) cells of 1:4. Here, the cell migration experiment was repeated with different cell ratios of TGF (+) and TGF (−) cells of 1:4, 2:3, 3:2, 4:1, and they are represented by TGF(20), TGF(40), TGF(60), and TGF(80) respectively. The previous conclusion showed that the clustering results of cells are influenced by the cell phenotype. TGF (+) cells as leader cells have a guiding effect on the migration of TGF (−) cells as follower cells. Cell migration trajectories were first normalized to eliminate the effect of position, and the starting point of all trajectories was set at the origin (0,0). The normalized data were then dimensionally reduced and clustered to obtain a two-dimensional visualization of the clustering results. Each data point represents a trajectory, and from the clustering results it can be found that data points with similar migration patterns are clustered together, which represents the existence of interactions between these cells. When the clustering results were

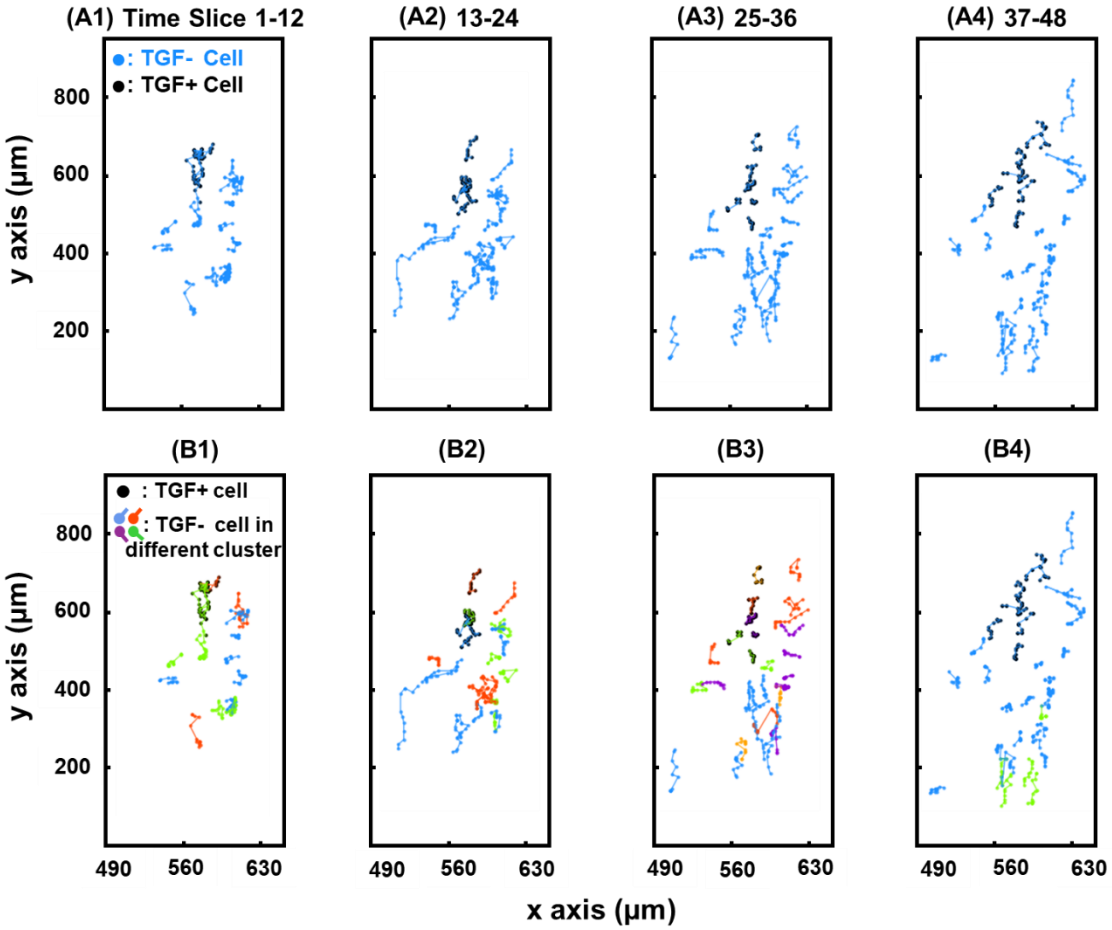


Figure 4.4 Before and after cell clustering with TGF(20). (A1-A5) The positions of TGF (+) (black) and TGF (-) (blue) cells before normalization in different observation windows. (B1-B5) Cell trajectories' positions combined with clustering results.

superimposed on the trajectory map before normalization, the trajectories of cells with similar migration patterns also had positional similarity.

Here, a clustering analysis was performed using different ratios of cell migration data from Chapter 2. Figure 4.4A shows the cell migration trajectory of TGF(20), and B shows the results after clustering. TGF (+) cells belong to the green and blue clusters within the first and last observation window, respectively. Within the last window, the blue clusters are significantly larger in extent than the green clusters in the first window, indicating a gradual expansion of the influence of the leader cells. In the middle two windows, the extent of the following cells belonging to the same cluster as the TGF (+)

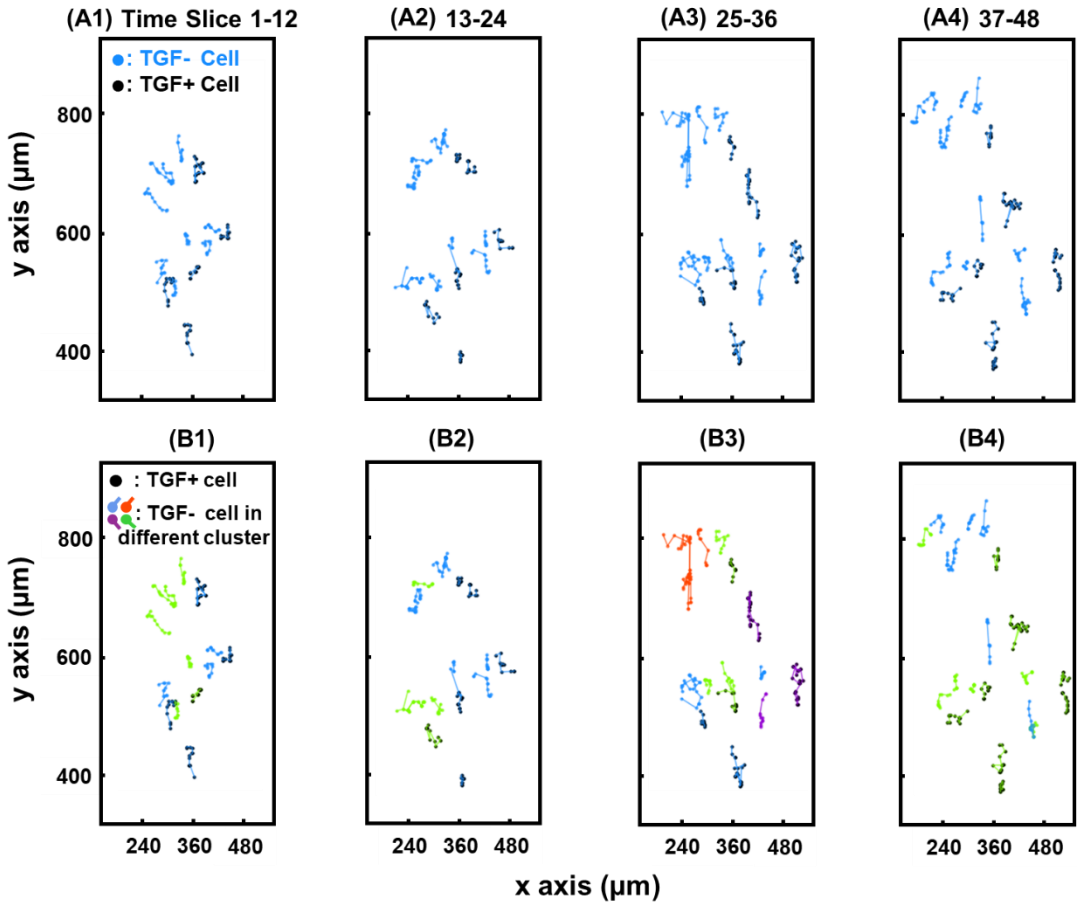


Figure 4.5 Before and after cell clustering with TGF(40). (A1-A5) The positions of TGF (+) (black) and TGF (-) (blue) cells before normalization in different observation windows. (B1-B5) Cell trajectories' positions combined with clustering results.

cells is not obvious, and only the cluster boundaries located in the lower right corner are very clear. The clearly bounded clusters of following cells indicate that their migration is not influenced by the leader cells, which is related to the distance between them and the cell state they are in. Figure 4.5 The TGF (+) cells increased to 40%. Compared to TGF (20), the cluster boundaries are more pronounced in the clustering results for TGF (40). The wide influence of leader cells in all observation windows, especially in the first two windows where leader cells are present in the green and blue clusters, indicates that all cells are more or less influenced by TGF (+) cells. In the third window the orange clusters maintained exclusive migration patterns and did not align with the leader cells, but in the last window these exclusive patterns disappeared. This

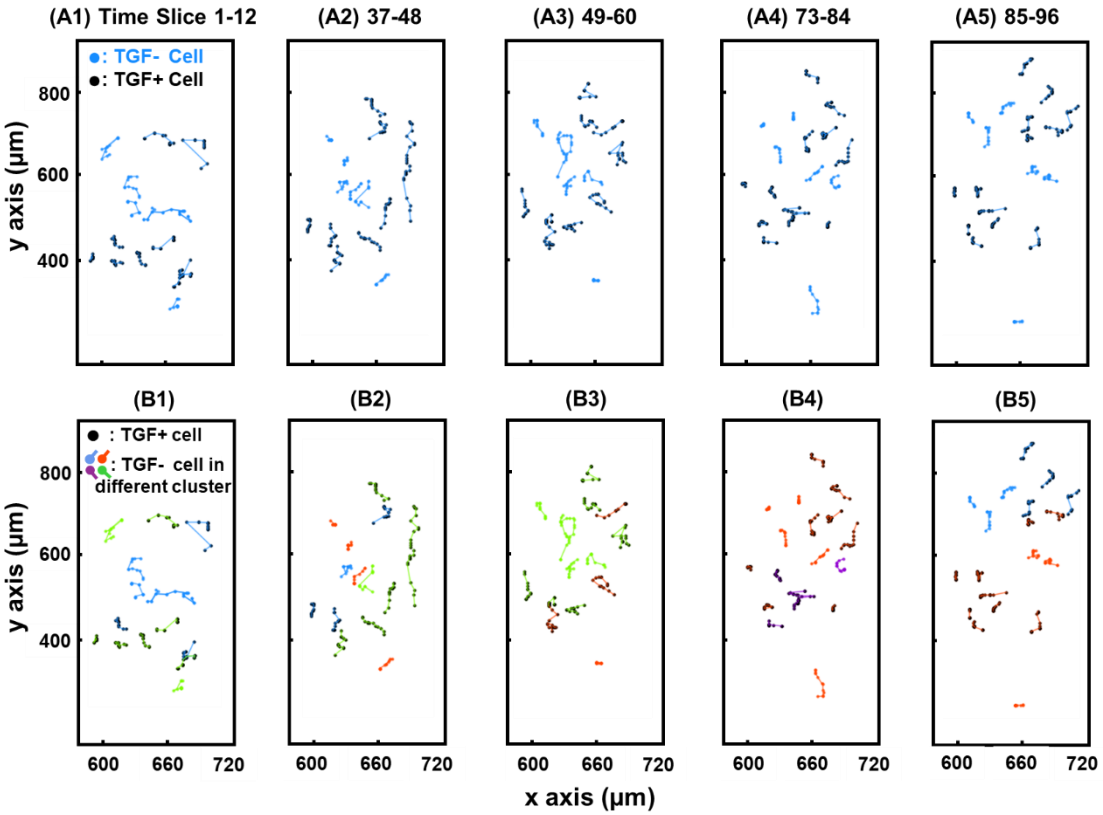


Figure 4.6 Before and after cell clustering with TGF(60). (A1-A5) The positions of TGF (+) (black) and TGF (-) (blue) cells before normalization in different observation windows. (B1-B5) Cell trajectories' positions combined with clustering results.

implies that the influence of TGF (+) cells may not be continuous and that the effect changes with the cellular environment and the cellular state, also in relation to the increase in the proportion of leader cells. Similar clustering results appear in TGF(60), as shown in Figure 4.6, where TGF (+) cells appear in almost all clusters, and since the number of leader cells already occupies the majority, the migration pattern of following cells will mostly be consistent with the leader cells. The more active migration pattern of mesenchymal cells drives the migration of epithelial cells. Such a result is more obvious in TGF(80), where the TGF (−) cells in Figure 4.7 are almost surrounded by TGF (+) cells, and the following cells almost lose their migratory characteristics, their migratory pattern being dominated by the leader cells, thus exhibiting a migratory pattern consistent with that of TGF (+) cells.

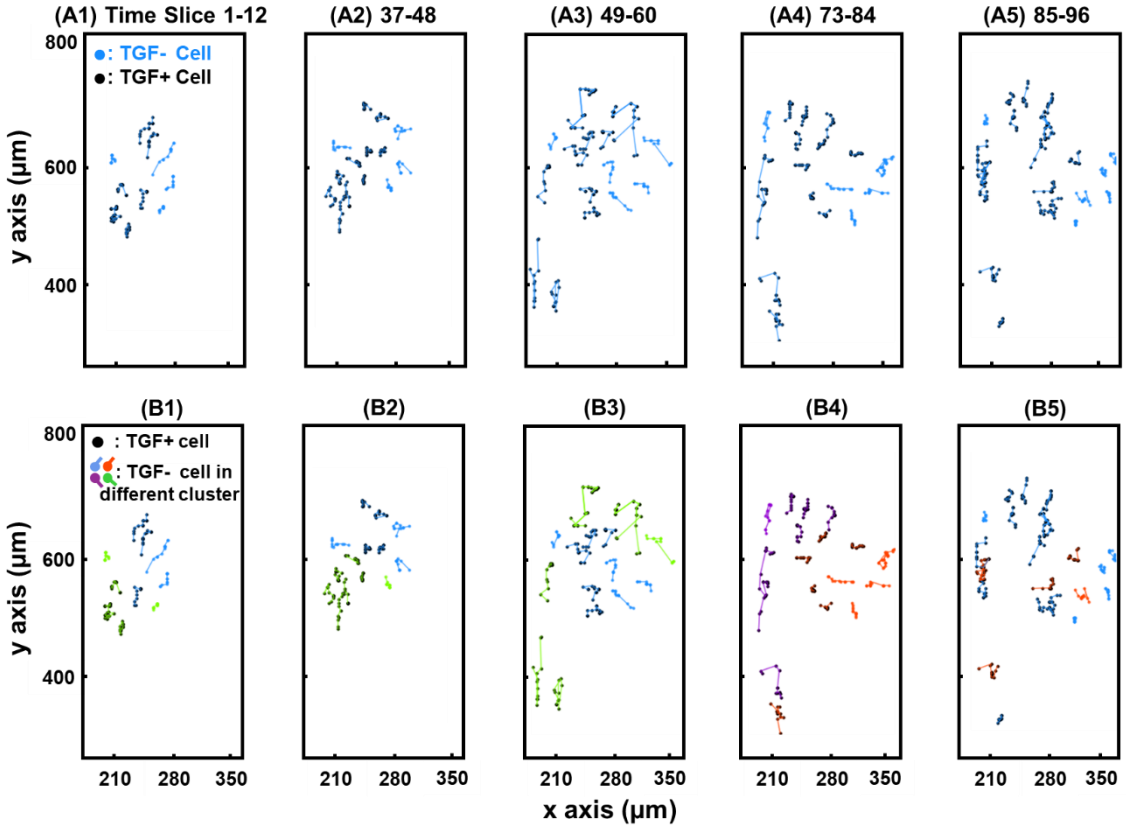


Figure 4.7 Before and after cell clustering with TGF(80). (A1-A5) The positions of TGF (+) (black) and TGF (-) (blue) cells before normalization in different observation windows. (B1-B5) Cell trajectories' positions combined with clustering results.

4.3.3 Correlation between cell division and clustering

In the process of clustering, the choice of observation window, the environment around the cells, and the state of the cells can affect the results of clustering. Cell division is an important factor that affects the migration patterns. To further analyze the relationship between cell division and migration patterns, here, based on the results in Chapter 3, cell lineages was drawn. In Figure 4.8, there were two TGF (+) cells in the first generation that were distant from each other. Multiple daughter cells were observed after cell division. Until the last time slice, five TGF (+) cells were present. According to each division, TGF (+) was referred to as two groups, where Group 1 included TGF (+) Cell ID 4 and 5 and Group 2 included TGF (+) Cell ID 1 to 3 (refer to Figure 3.5).

Different clusters are represented by different colors, and it is easy to see that the color of the cluster's changes for the most part before and after cell division. Also, the clusters of daughter cells and mother cells are mostly different. This indicates that the migration pattern of the cells starts to change before the division, and this change makes them different from the migration characteristics of the surrounding cells. This phenomenon continues until the end of cell division, which leads to changes in the cell's surroundings, which is also one of the factors affecting the migration pattern of the cells. Thus, the clustering of cell migration patterns is influenced by multiple factors, which are linked to each other, and their combined effect leads to a constant change in the migration characteristics of the cells, which makes the clustering more difficult.

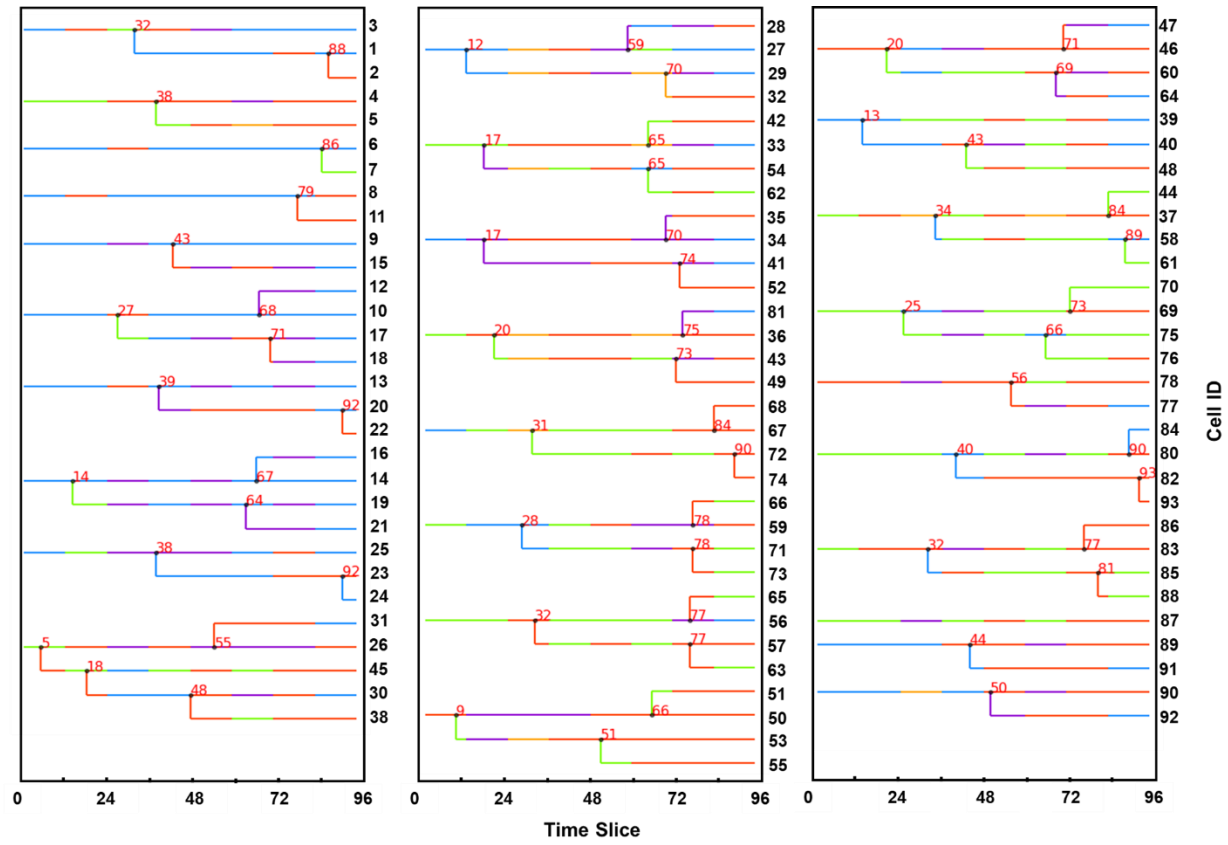


Figure 4.8 Cell lineage tree. Daughter cells are divided from mother cells at different time slice (red font). Colors represent the clusters. Cell ID 1-5 are TGF (+) cells (Group1: ID 4-5, Group2: ID 1-2), Cell ID 6-93 are TGF (-) cells.

4.4 Conclusion

The UMAP and t-SNE dimension reduction algorithms in time-series clustering method were compared and optimized their hyperparameters. Different observation windows were also compared to find the reasonable clustering. In combination of the clustering results, the cell lineage tree was made. Most of the time, the cell division accompany with the change of clustering results.

Cell migration is a process that involves multiple factors, and clustering of cell migration trajectories can identify cells with similar migration patterns. The interaction of multiple factors makes the optimization of clustering methods more difficult, but also improves the accuracy of clustering results during the optimization process. Cell migration is influenced not only by external factors such as the surrounding environment, but also by internal factors such as cell division. The ratio of different cell types can expand the scope of this influence and make it easier to observe. In terms of clustering methods, the choice of observation window is crucial for clustering, as a suitable window can accurately distinguish cells in different states; the parameters of the algorithm also affect the clustering results, and the optimal combination of parameters is a guarantee of reasonable clustering. These influencing factors together lead to the changing patterns of cell migration, but the optimized algorithm can further clarify the interactions between cells.

REFERENCE

[1] Jung, D. H.; Kim, Y.; Cho, H. H.; Lee, B.; Suh, S. J.; Heo, J. H.; Lee, J. H., Automatic quantification of living cells via a non-invasive achromatic colorimetric sensor through machine learning-assisted image analysis using a smartphone. *Chem Eng J* **2022,** 450. doi: 10.1016/j.cej.2022.138281.

[2] Liu, Y.; Wang, Z.; Zhou, Z.; Xiong, T., Analysis and comparison of machine learning methods for blood identification using single-cell laser tweezer Raman spectroscopy. *Spectrochim Acta A Mol Biomol Spectrosc* **2022,** 277, 121274. doi: 10.1016/j.saa.2022.121274.

[3] Zhang, J.; Cui, S.; Shen, L.; Gao, Y.; Liu, W.; Zhang, C.; Zhuang, S., Promotion of Bladder Cancer Cell Metastasis by 2-Mercaptobenzothiazole via Its Activation of Aryl Hydrocarbon Receptor Transcription: Molecular Dynamics Simulations, Cell-Based Assays, and Machine Learning-Driven Prediction. *Environ Sci Technol* **2022,** 56, (18), 13254-13263. doi: 10.1021/acs.est.2c05178.

[4] Cendon-Florez, Y.; Pippa, R.; Boffo, S.; Odero, M. D.; Giordano, A., Data mining analysis of the PP2A-cell cycle axis in breast and prostate cancer patients. *Cancer Res* **2020,** 80, (16). doi: 10.1158/1538-7445.Am2020-107.

[5] Lin, J.; Chen, L. Z.; Wu, D. J.; Lin, J. X.; Liu, B.; Guo, C. R., Potential Diagnostic and Prognostic Values of CBX8 Expression in Liver Hepatocellular Carcinoma, Kidney Renal Clear Cell Carcinoma, and Ovarian Cancer: A Study Based on TCGA Data Mining. *Comput Math Method M* **2022,** 2022. doi: 10.1155/2022/1372879.

[6] Qi, X. J.; Guo, Z. H.; Chen, Q. Y.; Lan, W. N.; Chen, Z. Z.; Chen, W. M.; Lin, L. Z., A Data Mining-Based Analysis of Core Herbs on Different Patterns (Zheng) of Non-Small Cell Lung Cancer. *Evid-Based Compl Alt* **2021,** 2021. doi: 10.1155/2021/3621677.

[7] Fujita, A.; Severino, P.; Kojima, K.; Sato, J. R.; Patriota, A. G.; Miyano, S., Functional clustering of time series gene expression data by Granger causality. *Bmc Syst Biol* **2012,** 6, 137. doi: 10.1186/1752-0509-6-137.

[8] Shi, C. M.; Wei, B. T.; Wei, S. L.; Wang, W.; Liu, H.; Liu, J. L., A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip J Wirel Comm* **2021,** 2021, (1). doi: 10.1186/s13638-021-01910-w.

[9] Barnard, A. S.; Opletal, G., Predicting structure/property relationships in multi-dimensional nanoparticle data using t-distributed stochastic neighbour embedding and machine learning. *Nanoscale* **2019,** 11, (48), 23165-23172. doi: 10.1039/c9nr03940f.

[10] Stark, G. F.; Hart, G. R.; Nartowt, B. J.; Deng, J., Predicting breast cancer risk using personal health data and machine learning models. *PLoS One* **2019,** 14, (12). doi: 10.1371/journal.pone.0226765.

[11] Aghabozorgi, S.; Shirkhorshidi, A. S.; Wah, T. Y., Time-series clustering - A decade review. *Inform Syst* **2015,** 53, 16-38. doi: 10.1016/j.is.2015.04.007.

[12] Ploszaj-Mazurek, M.; Rynska, E.; Grochulska-Salak, M., Methods to Optimize Carbon Footprint of Buildings in Regenerative Architectural Design with the Use of Machine Learning, Convolutional Neural Network, and Parametric Design. *Energies* **2020,** 13, (20). doi: 10.3390/en13205289.

[13] Oneto, L.; Bunte, K.; Sperduti, A., Advances in artificial neural networks, machine learning and computational intelligence. *Neurocomputing* **2020,** 416, 172-176. doi: 10.1016/j.neucom.2020.03.059.

[14] Kasabov, N. K.; Doborjeh, M. G.; Doborjeh, Z. G., Mapping, Learning, Visualization, Classification, and Understanding of fMRI Data in the NeuCube Evolving Spatiotemporal Data Machine of Spiking Neural Networks. *IEEE Trans Neural Netw Learn Syst* **2017,** 28, (4), 887-899. doi: 10.1109/TNNLS.2016.2612890.

[15] McInnes, L.; Healy, J.; Melville, J., Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**, https://arxiv.org/abs/1802.03426.

[16] Taylor, J. N.; Mochizuki, K.; Hashimoto, K.; Kumamoto, Y.; Harada, Y.; Fujita, K.; Komatsuzaki, T., High-Resolution Raman Microscopic Detection of Follicular Thyroid Cancer Cells with Unsupervised Machine Learning. *J Phys Chem B* **2019,** 123, (20), 4358-4372. doi: 10.1021/acs.jpcb.9b01159.

[17] Puvanesarajah, S.; Hodge, J. M.; Evans, J. L.; Seo, W.; Yi, M.; Fritz, M. M.; Macheski-Preston, M.; Gansler, T.; Gapstur, S. M.; Gaudet, M. M., Unsupervised deep-learning to identify histopathological features among breast cancers in the Cancer Prevention Study-II Nutrition Cohort. *Cancer Res* **2019,** 79, (13). doi: 10.1158/1538-7445.Sabcs18-2417.

[18] Lin, E.; Mukherjee, S.; Kannan, S., A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis. *Bmc Bioinformatics* **2020,** 21, (1), 64. doi: 10.1186/s12859-020-3401-5.

[19] Heiser, C. N.; Lau, K. S., A Quantitative Framework for Evaluating Single-Cell Data Structure Preservation by Dimensionality Reduction Techniques. *Cell Rep* **2020,** 31, (5), 107576. doi: 10.1016/j.celrep.2020.107576.

[20] Sun, X.; Liu, Y.; An, L., Ensemble dimensionality reduction and feature gene extraction for single-cell RNA-seq data. *Nat Commun* **2020,** 11, (1), 5853. doi: 10.1038/s41467-020-19465-7.

[21] Becht, E.; McInnes, L.; Healy, J.; Dutertre, C. A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W., Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* **2019,** 37, (1), 38-+. doi: 10.1038/nbt.4314.

[22] Joswiak, M.; Peng, Y.; Castillo, I.; Chiang, L. H., Dimensionality reduction for visualizing industrial chemical process data. *Control Eng Pract* **2019,** 93. doi: 10.1016/j.conengprac.2019.104189.

[23] Zisis, T.; Bruckner, D. B.; Brandstatter, T.; Siow, W. X.; d'Alessandro, J.; Vollmar, A. M.; Broedersz, C. P.; Zahler, S., Disentangling cadherin-mediated cell-cell interactions in collective cancer cell migration. *Biophys J* **2022,** 121, (1), 44-60. doi: 10.1016/j.bpj.2021.12.006.

[24] Han, P.; Wang, W. Q.; Shi, Q. Y.; Yue, J. C., A combined online-learning model with K-means clustering and GRU neural networks for trajectory prediction. *Ad Hoc Netw* **2021,** 117. doi: 10.1016/j.adhoc.2021.102476.

[25] Sarkar, D.; Khan, T.; Talukdar, F. A., Hyperparameters optimization of neural network using improved particle swarm optimization for modeling of electromagnetic inverse problems. *Int J Microw Wirel T* **2021**. doi: 10.1017/S1759078721001690.

[26] Jasmine, E. M.; Milton, A., The role of hyperparameters in predicting rainfall using n-hidden-layered networks. *Nat Hazards* **2022,** 111, (1), 489-505. doi: 10.1007/s11069-021-05063-3.

[27] Wen, L.; Ye, X. C.; Gao, L., A new automatic machine learning based hyperparameter optimization for workpiece quality prediction. *Meas Control-Uk* **2020,** 53, (7-8), 1088-1098. doi: 10.1177/0020294020932347.

[28] Palaniswamy, S. K.; Venkatesan, R., Hyperparameters tuning of ensemble model for software effort estimation. *J Amb Intel Hum Comp* **2021,** 12, (6), 6579-6589. doi: 10.1007/s12652-020-02277-4.

[29] Blume, S.; Benedens, T.; Schramm, D., Hyperparameter Optimization Techniques for Designing Software Sensors Based on Artificial Neural Networks. *Sensors (Basel)* **2021,** 21, (24). doi: 10.3390/s21248435.

[30] Hozumi, Y.; Wang, R.; Yin, C.; Wei, G. W., UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. *Comput Biol Med* **2021,** 131, 104264. doi: 10.1016/j.compbiomed.2021.104264.

[31] Yang, Y.; Sun, H.; Zhang, Y.; Zhang, T.; Gong, J.; Wei, Y.; Duan, Y. G.; Shu, M.; Yang, Y.; Wu, D.; Yu, D., Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell Rep* **2021,** 36, (4), 109442. doi: 10.1016/j.celrep.2021.109442.

[32] Linderman, G. C.; Rachh, M.; Hoskins, J. G.; Steinerberger, S.; Kluger, Y., Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods* **2019,** 16, (3), 243-245. doi: 10.1038/s41592-018-0308-4.

[33] Kobak, D.; Berens, P., The art of using t-SNE for single-cell transcriptomics. *Nat Commun* **2019,** 10, (1), 5416. doi: 10.1038/s41467-019-13056-x.

[34] Zhou, B.; Jin, W., Visualization of Single Cell RNA-Seq Data Using t-SNE in R. *Methods Mol Biol* **2020,** 2117, 159-167. doi: 10.1007/978-1-0716-0301-7_8.

CHAPTER 5

PROSPECTIVE OF CLUSTERING METHOD

IN CELL MIGRATION

Abstract

This chapter elucidates the applications of time-series clustering algorithms, especially the combination of machine learning and bioinformatics. Clustering and prediction are not limited to cell migration, and with the customization of algorithms, which can be applied to a wide range of biological and medical fields.

5.1 Bioinformatics and Machine Learning

Machine learning provides important technological support for many interdisciplinary disciplines. Bioinformatics attempts to use information technology to study the phenomena and laws of life, such as the implementation of the genome project and the exciting prospect of genetic drugs. Bioinformatics research involves the entire process from 'life phenomena' to pattern discovery. This includes data acquisition, data management, data analysis, simulation, and experimentation. The "data analysis" is precisely the stage for machine learning techniques. Various machine learning techniques have already shone in this field.

Increasingly, machine learning is seen in bioinformatics applications, such as finding usable patterns in data and then making predictions [1, 2]. Typically, these predictive models are used to operationalize processes to optimize the decision-making process, but at the same time, they can also provide key insights and information to report on strategic decisions [3, 4].

The study of cell migration is an important part of biology. In recent years, researchers have conducted in-depth analysis and exploration of the mechanisms of cell migration. Cells produce movement after receiving migration signals or sensing some molecular concentration gradients. When migration occurs, the anterior part produces a pseudopod capable of extending forward, the rear part produces a contraction, and when the two behaviors alternate, the cells migrate forward. At the same time, new adhesion relationships are formed between the cells, and molecular signals are released that can coordinate multicellular migration. Collective migration based on physical and chemical signals has been extensively studied. Different migration models are able to match an increasing number of instances. However, model calculations in biology often involve large volumes of raw data, which require techniques and methods with relevant processing capabilities for deeper analysis [5]. The combination of multiple disciplines can be a good solution to this problem, and computational science can provide more arithmetic power and higher efficiency.

5.2 Tracking and Clustering Methods in Cancer Analyze

Genomic information is now widely used for the accurate treatment of cancer. Since individual types of histological data represent only a single viewpoint and are subject to data noise and bias, multiple types of histological data are required for accurate prediction. However, effective integration of multi-omics data is challenging due to the large number of redundant variables in multi-omics data but relatively small sample sizes.

5.2.1 Tracking Targeted Markers in Cancer Cells

For cancer patients, receiving treatment when cancer has not yet metastasized often leads to good treatment outcomes. However, once cancer has metastasized, the patient's treatment prospects are significantly reduced. According to statistics, 90% of

cancer patients die because of cancer metastasis. However, the process of cancer metastasis is not fully understood. Which of the thousands of in situ cancer cells are important causes of cancer metastasis? What changes have they produced during the metastasis process and to which tissues in the body have they metastasized? If each cancer cell could be given a unique "tag", it would be possible to track their evolution and metastasis, as well as that of their progeny.

For example, genes expressing Cas9 enzyme and guidance RNA directing them to cut specific regions of cancer cell genome were introduced into cancer cells. With the continuous division and proliferation of cancer cells, Cas9 enzyme will continue to cut in these designated areas, and the DNA repair mechanism of cancer cells will continue to repair these gaps. During the repair process, various insertion sequences will be introduced. These inserted sequences will be passed down from generation to generation along with cell division and become unique "tags" carried by each cancer cell [6]. This provides a possibility for tracking the migration of cancer cells. Other anti-cancer technologies include encapsulating small molecules of anti-cancer drugs in nanoparticles and labeling them with ligands that target only cancer cell markers in order to focus high concentrations of the drug on tumor cell sites with no effect on healthy tissue. Targeting only one cancer marker is not sufficient, and mounting two ligands on liposomes, each targeting a surface protein of metastatic cancer cells, can detect those cancer cells that are "missed" by the single-ligand nanoparticles [7]. In addition, the recognition efficiency of the dual-ligand nanoparticles was also high. A similar approach has been taken with chaperonin to act as a marker in the blood to indicate cancer cells, thus revealing more clearly the spreading cancer. This new biomarker can detect more cancer cells in the blood [8]. Cancer cells require a large number of proteins to survive and circulate in the body, and the chaperonin complex allows proteins to fold into a functional three-dimensional shape without which important proteins needed by cancer cells cannot be formed, and all cells contain the

chaperonin complex, but cancer cells have higher levels.

5.2.2 Computer Algorithms in Cancer Tracking

In addition to the use of invasive biopsies or the use of contrast agents to track cancer, the approach through optical and computer algorithms is widely used. Multiphoton microscopy works by transmitting a laser into the tissue [9]. The short pulses maintain a small average power and do not damage the tissue. As the different tissue components interact with the laser, they emit signals that are then retrieved by the microscope to form an image. Automated image processing algorithms to reveal unique textural features and statistical model analysis are used to distinguish healthy from diseased tissue. Metastatic and clonal history integrative analysis allows researchers to infer past metastatic processes from the DNA sequence data obtained so far [10]. A clearer understanding of the history of cancer migration has been obtained.

The addition of computational science provides new solutions for solving large-volume and high-dimensional data. The clustering method used in this study is based on the dimensionality reduction algorithm and clustering algorithm in machine learning for collective cell migration. Time-lapse observation microscopy is a frequently used instrument in biology. For cell movement, wound healing, time-lapse observation gives me a time-series of data where cells exhibit different migratory properties at each moment. These time-dependent features make up high-dimensional raw data that are difficult to compare visually, so the advantages of dimensionality reduction algorithms are revealed. Representing high-dimensional data in two or three-dimensions enables visualization of unimaginable data, and then clustering can make the results more reliable. The downscaling and clustering algorithms are not only applicable to multidimensional time-series of migration trajectories, but also to other targets with multidimensional characteristics. For example, by sequencing mRNA for each cell to distinguish which genes are activated, a large amount of gene transcript data can be

more efficiently classified using dimensionality reduction clustering. Such applications can be extended to the identification and discovery of cancer cells to facilitate more precise treatment delivery.

5.3 Customization of Machine Learning Algorithms

A large number of biological experiments have accumulated tens of thousands of pieces of bioinformatic data. How to effectively collect, organize, retrieve, and analyze the data to extract the rules from them, and to translate them into theories, so as to read the bioinformatics to guide the research work, has posed a high demand on bioinformatics, and also posed a challenge to information theory and technology [11, 12]. Data mining, an emerging technology based on digital databases, statistics, and artificial intelligence, provides biologists with unprecedented data analysis tools for the analysis and extraction of gene and protein information.

The time-series clustering method introduced in Chapter 3 involves dimension reduction and clustering algorithms. The application of this method is not limited by UMAP and k-means algorithms. Besides these algorithms, the method can combine with other algorithms to process more datasets (Table 5.1). In the era of big data, researchers are often faced with hundreds of samples and tens of thousands of gene expression matrices. How to extract valuable information from this huge amount of data has become a priority.

Methods of data dimensionality reduction can be divided into linear and non-linear dimensionality reduction. The PCA method is one of the most widely used algorithms for data dimensionality reduction [13-15]. It retains the main features of the data, at the observation point level, simplifying the complexity of the observed objects, and these smaller feature matrices are easier to visualize and analyze. At the same time, it can help to determine correlations between data points. LDA (Linear discriminant analysis) is a supervised learning technique for dimensionality reduction, which means that each

sample of the dataset has a class output, unlike PCA (unsupervised learning). The idea behind LDA is to maximize the between-class mean and minimize the within-class variance [16]. The data is projected in low-dimensions and the projection points are as close as possible to each other for the same class of data and as far away as possible from the centroids of the projection points for different classes of data. The MDS method assumes that the dataset lies on a smooth, low-dimensional non-linear manifold, and the distance-preserving method assumes that the manifold can be defined by the pairwise distances of its points [17]. A low-dimensional mapping can be found by keeping one or more features of the high-dimensional space and attempting to keep the two-by-two distances between points constant. MDS uses the dissimilarity matrix as the original input, rather than from the original data matrix, thus enabling a better study of differences within sets of preserved relational data. The algorithms require that the distances between samples in the original space are maintained in the low-dimensional space. For most clustering algorithms, distance is an important property for classifying samples into categories, so when the distance is kept constant after dimensionality reduction, then it is equivalent to keeping the relative spatial relationships of the sample's constant.

Table 5.1 The combination of clustering and dimension reduction algorithm.

Clustering algorithm	Dimension reduction algorithm
k-means GMM DBSCAN Spectral clustering Hierarchical clustering	PCA MDS t-SNE UMAP

Clustering is an unsupervised learning algorithm where the data does not need to be labelled. It is different from classification, which is a supervised learning of labelled data. Clustering is the partitioning of a data set into classes or clusters according to some metric (*e.g.*, distance between samples) so that the similarity of elements within classes is as great as possible and the similarity of elements between classes is as small as possible, by which similar data are clustered to achieve the effect of clustering. This thesis conducted k-means algorithm, besides many other algorithms, can be selected. k-means is unable to cluster two classes with the same mean (same cluster centroid) and GMM (Gaussian mixture model) was proposed to address this shortcoming. GMM does this by maximizing the posterior probabilities of the selected components [18]. The posterior probability of each data point indicates the likelihood of belonging to each class, rather than determining that it belongs to a class exactly, hence the term soft clustering. It may be more appropriate than k-means clustering when the class sizes are different and there are correlations between the clusters. In data mining and statistics, hierarchical clustering is a method of cluster analysis that seeks to build a hierarchy of clusters. In multivariate statistics, spectral clustering techniques make use of the spectrum of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a density-based clustering algorithm that does not require a pre-specified number of clusters to be clustered and has an indeterminate number of clusters [19]. It defines a "cluster" as the largest set of densely connected points. It is able to find arbitrarily shaped classes, whereas k-means can only find convex shapes, and DBSCAN is also very noise-resistant, finding arbitrarily shaped clusters in noisy data.

5.4 Prospects of Time-series Clustering Method

Genomic information is now widely used for the accurate treatment of cancer. As

individual types of histological data represent only a single phenomenon and are subject to data noise and variation, multiple types of histological data are required to make accurate predictions. However, effective integration of multi-omics data is challenging due to the large number of redundant variables in multi-omics data but relatively small sample sizes.

Machine learning is premised on algorithm training that provides specific input data that can predict output values within a certain probability interval [20]. Machine learning is inductive rather than inferential, which is related to probabilities not final conclusions [21]. These algorithms are constructed and predictive models are obtained. The model can be analyzed directly on the original data and applied to the new data to predict some desired information. The output of the model can be a classification, a hidden relationship, an attribute or an estimate, etc. Building machine learning models is an iterative process that requires labelling data and hands-on experimentation. With the development of deep learning techniques, integrating multi-omics data extracts representative features. However, the generated models are very poorly applied due to the influence of data noise. In addition, previous studies have typically focused on individual cancer types, without comprehensive testing of pan-cancers.

The combination of bioinformatics and machine learning is an interdisciplinary component that quantifies information about organisms and studies their effects on their interactions [22]. Machine learning is now widely used in image recognition research, using known training sets to predict outcomes for the type of data, while deep learning models can predict and downscale analysis with greater power and flexibility. With the right training data, deep learning can automatically learn features and patterns with little human intervention. In the future, machine learning can also be used to improve the interpretability of biological data and to transform biological image information data into actionable clinical information through recognition and clustering (Figure 5.1), improving disease diagnosis protocols to minimize drug side effects and maximize

Figure 5.1 Future applications of time-series clustering method in combine with image recognition.

efficacy. Manual statistical analysis is slow due to the number of variables involved, and machine learning can help shorten the process.

The clustering approach in this study was able to uncover the intrinsic effects of the intersection of multiple factors on migration. This differs from the previous molecular and physical perspectives that required measurement and analysis of proteins and forces. It provides a new perspective for studying collective cell migration, especially the interactions between multiple cells when they coexist. In the future, methods can combine more influencing factors and improve the precision of cellular localization to improve the efficiency of clustering. This provides a new selectivity for the study of cancer cell invasion.

REFERENCE

[1] Serra, A.; Galdi, P.; Tagliaferri, R., Machine learning for bioinformatics and neuroimaging. *Wires Data Min Knowl* **2018,** 8, (5). doi: 10.1002/widm.1248.

[2] Zhang, L. C.; Liu, M. J.; Zhang, Z. J.; Chen, D.; Chen, G.; Liu, M. Y., Machine learning based identification of hub genes in renal clear cell carcinoma using multi-omics data. *Methods* **2022,** 207, 110-117. doi: 10.1016/j.ymeth.2022.09.008.

[3] Zou, Q.; Liu, Q., Advanced Machine Learning Techniques for Bioinformatics. *Ieee Acm T Comput Bi* **2019,** 16, (4), 1182-1183. doi: 10.1109/Tcbb.2019.2919039.

[4] Hossain, M. A.; Saiful Islam, S. M.; Quinn, J. M. W.; Huq, F.; Moni, M. A., Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. *J Biomed Inform* **2019,** 100, 103313. doi: 10.1016/j.jbi.2019.103313.

[5] Zou, Q., Latest Machine Learning Techniques for Biomedicine and Bioinformatics. *Curr Bioinform* **2019,** 14, (3), 176-177. doi: 10.2174/1574893614031902201128 55.

[6] Quinn, J. J.; Jones, M. G.; Okimoto, R. A.; Nanjo, S.; Chan, M. M.; Yosef, N.; Bivona, T. G.; Weissman, J. S., Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* **2021,** 371, (6532), 909-+. doi: 10.1126/science.abc1944.

[7] Doolittle, E.; Peiris, P. M.; Doron, G.; Goldberg, A.; Tucci, S.; Rao, S.; Shah, S.; Sylvestre, M.; Govender, P.; Turan, O.; Lee, Z.; Schiemann, W. P.; Karathanasis, E., Spatiotemporal Targeting of a Dual-Ligand Nanoparticle to Cancer Metastasis. *Acs Nano* **2015,** 9, (8), 8012-21. doi: 10.1021/acsnano.5b01552.

[8] Cox, A.; Martini, A.; Ghozlan, H.; Moroose, R.; Zhu, X.; Lee, E.; Khaled, A. S.; Barr, L.; Alemany, C.; Fanaian, N.; Griffith, E.; Sause, R.; Litherland, S. A.; Khaled, A. R., Chaperonin containing TCP1 as a marker for identification of circulating tumor cells in blood. *PLoS One* **2022,** 17, (6), e0264651. doi: 10.1371/journal.pone.0264651.

[9] Pouli, D.; Genega, E. M.; Sullivan, T. B.; Rieger-Christ, K. M.; Wright, V.; Georgakoudi, I.; Schnelldorfer, T., Two-photon images reveal unique texture features for label-free identification of

ovarian cancer peritoneal metastases. *Biomed Opt Express* **2019,** 10, (9), 4479-4488. doi: 10.1364/BOE.10.004479.

[10] El-Kebir, M.; Satas, G.; Raphael, B. J., Inferring parsimonious migration histories for metastatic cancers. *Nat Genet* **2018,** 50, (5), 718-726. doi: 10.1038/s41588-018-0106-z.

[11] Wong, K. K. L.; Deng, X. F.; Ng, E. Y. K., A Special Section on Machine Intelligence Applied to Bioinformatics and Statistical Analysis. *J Med Imag Health In* **2020,** 10, (5), 1216-1218. doi: 10.1166/jmihi.2020.3002.

[12] Cho, Y. R.; Kang, M., Interpretable machine learning in bioinformatics. *Methods* **2020,** 179, 1-2. doi: 10.1016/j.ymeth.2020.05.024.

[13] Kong, X. Z.; Song, Y.; Liu, J. X.; Zheng, C. H.; Yuan, S. S.; Wang, J.; Dai, L. Y., Joint Lp-Norm and L(2,1)-Norm Constrained Graph Laplacian PCA for Robust Tumor Sample Clustering and Gene Network Module Discovery. *Front Genet* **2021,** 12, 621317. doi: 10.3389/fgene.2021.621317.

[14] Fujisawa, K.; Shimo, M.; Taguchi, Y. H.; Ikematsu, S.; Miyata, R., PCA-based unsupervised feature extraction for gene expression analysis of COVID-19 patients. *Sci Rep* **2021,** 11, (1), 17351. doi: 10.1038/s41598-021-95698-w.

[15] Chen, Z.; Gong, F.; Wan, L.; Ma, L., RobustClone: a robust PCA method for tumor clone and evolution inference from single-cell sequencing data. *Bioinformatics* **2020,** 36, (11), 3299-3306. doi: 10.1093/bioinformatics/btaa172.

[16] Jiang, T.; Liu, X. P.; Zhang, C.; Yin, C. A. H.; Liu, H. Z., Overview of Trends in Global Single Cell Research Based on Bibliometric Analysis and LDA Model (2009-2019). *J Data Info Sci* **2021,** 6, (2), 163-178. doi: 10.2478/jdis-2021-0008.

[17] Jin, B.; Fu, C.; Jin, Y.; Yang, W.; Li, S.; Zhang, G.; Wang, Z., An Adaptive Unsupervised Feature Selection Algorithm Based on MDS for Tumor Gene Data Classification. *Sensors (Basel)* **2021,** 21, (11). doi: 10.3390/s21113627.

[18] Wang, M.; Chen, J. Y., A GMM-IG framework for selecting genes as expression panel biomarkers. *Artif Intell Med* **2010,** 48, (2-3), 75-82. doi: 10.1016/j.artmed.2009.07.006.

[19] Zhao, Y.; Liu, X.; Li, X., An improved DBSCAN algorithm based on cell-like P systems with promoters and inhibitors. *PLoS One* **2018,** 13, (12), e0200751. doi: 10.1371/journal.pone.0200751.

[20] Jung, D. H.; Kim, Y.; Cho, H. H.; Lee, B.; Suh, S. J.; Heo, J. H.; Lee, J. H., Automatic quantification of living cells via a non-invasive achromatic colorimetric sensor through machine learning-assisted image analysis using a smartphone. *Chem Eng J* **2022,** 450. doi: 10.1016/j.cej.2022.138281.

[21] Rana, H. K.; Akhtar, M. R.; Islam, M. B.; Ahmed, M. B.; Lio, P.; Huq, F.; Quinn, J. M. W.; Moni, M. A., Machine Learning and Bioinformatics Models to Identify Pathways that Mediate Influences of Welding Fumes on Cancer Progression. *Sci Rep* **2020,** 10, (1), 2795. doi: 10.1038/s41598-020-57916-9.

[22] Koido, M.; Hon, C. C.; Koyama, S.; Kawaji, H.; Murakawa, Y.; Ishigaki, K.; Ito, K.; Sese, J.; Parrish, N. F.; Kamatani, Y.; Carninci, P.; Terao, C., Prediction of the cell-type-specific transcription of non-coding RNAs from genome sequences via machine learning. *Nat Biomed Eng* **2022**. doi: 10.1038/s41551-022-00961-8.

SUMMARY

This thesis focused on the bioinformatics with machine learning to investigate the interaction in collective cell migration. Studies of the migratory behavior of cells are needed to investigate and control metastasis. Cell migration generates a huge amount of dataset which can be processed by machine learning more efferently. The time-series clustering method can reproduce the migration similarity in the presence of different ratios of cells.

In Chapter 1, the recent progress of collective cell invasion and migration analysis was reviewed. EMT is one of the main mechanisms of cancer metastasis, in which epithelial cells acquire mesenchymal properties and the ability to leave the population to invade other regions of the body. In collective migration, highly migratory cells are found at the front of the cell population. Cells can interact with each other which can be detected by many equipment and methods. Most of the methods treat all cells together and get an overall impression, the more accurate aspect of single-cell is needed and can be achieved by clustering method.

In Chapter 2, the migration properties of epithelial and mesenchymal cells were examined during collective migration at the single-cell level. Different mixed ratios of cell populations were compared. Collective migration was quantitatively analyzed from two perspectives: cell migration within the colony and migration of the entire colony. Analysis of the effect of the cell mixing ratio on migration behavior showed that a small number of highly migratory cells enhanced some of the migratory properties of other cells. The results provide useful insights into the cellular interactions in collective cell migration of cancer cell invasion.

In Chapter 3, to fully comprehend metastasis, the methodology of analysis of individual cell migration in tissue should be well developed. Extracting and classifying cells with similar migratory characteristics in a colony would facilitate an

understanding of complex cell migration patterns. Here, electrospun fiber was used as the ECM for the *in vitro* modeling of collective cell migration, clustering of mesenchymal and epithelial cells based on trajectories, and analysis of collective migration patterns based on trajectory similarity. the trajectories were normalized to eliminate the effect of cell location on clustering and used UMAP to perform dimensionality reduction on the time-series data before clustering. When the clustering results were superimposed on the trajectories before normalization, the results still exhibited positional similarity, thereby demonstrating that this method can identify cells with similar migration patterns. The same cluster contained both mesenchymal and epithelial cells, and this result was related to cell location and cell division.

In Chapter 4, the time-series clustering method was further optimized. The parameters of the algorithm can affect the results in a different level. UMAP dimension reduction was conducted, besides UMAP, t-SNE is another algorithm with good performance. The hyperparameters of UMAP and t-SNE were changed and prepared to find better visualization results. The length observation window was also based on the same principle to generate a more reasonable result. The clustering results with different ratios of cells showed the location similarity under optimal conditions of algorithm.

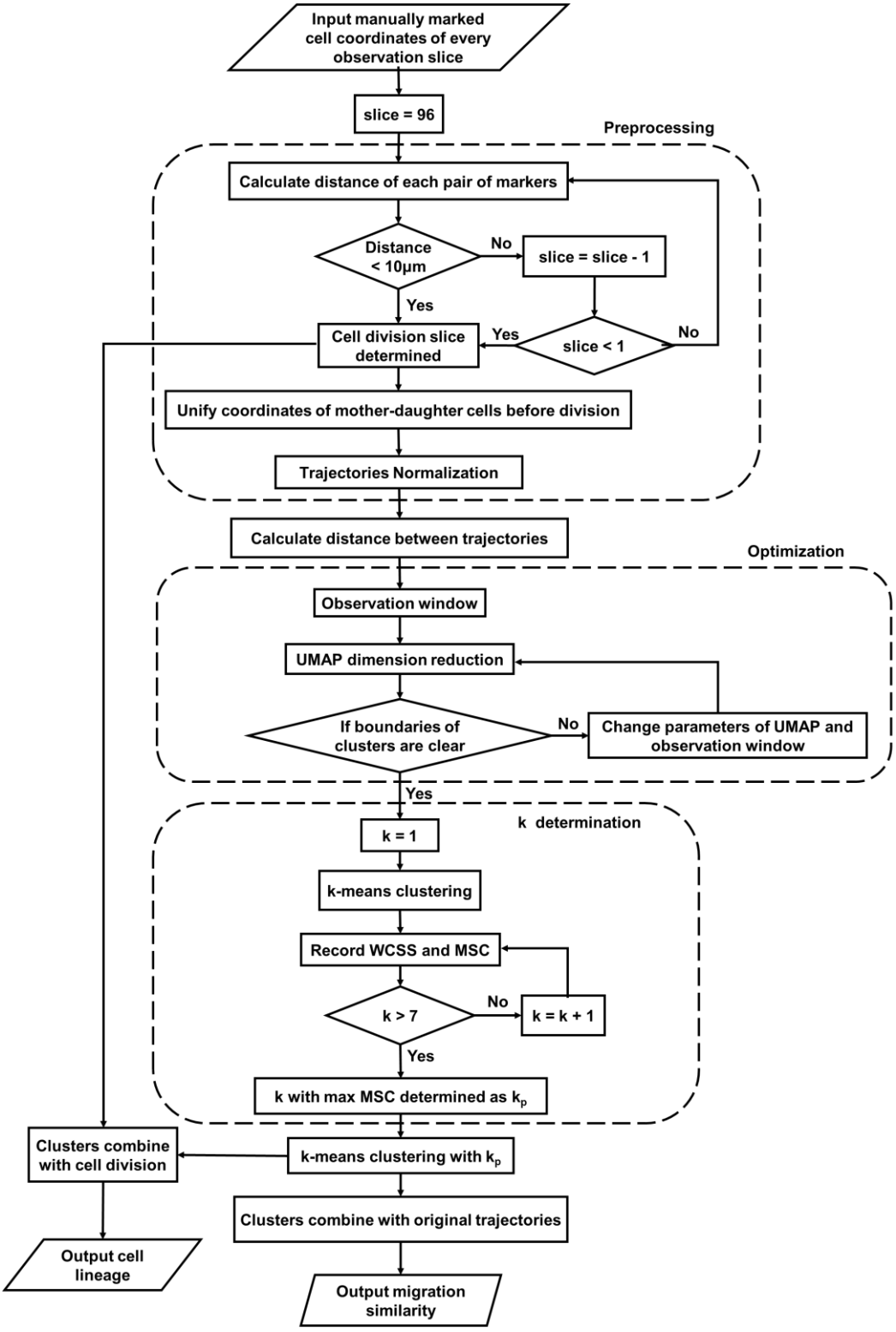
In Chapter 5, the applications of bioinformatics were clarified with machine learning. The biology information usually includes a huge amount of dataset like proteins, cells and genes which made the processing low efficient. Machine learning can deal with huge volumes of data efficiency and find the relationship between individuals, especially the hidden relationship. In the future, the tendency to be interdisciplinary and the combination of different subjects will solve many questions easily. Time-series clustering is an example to process biological questions with machine learning. It can be developed in combination with other algorithms and used in fields like prediction and diagnosis.

The interactions between cells are further investigated in the single-cell trajectory

aspect in this thesis. The time-series clustering method can reproduce the locational similarity in the presence of different of ratios of epithelial and mesenchymal cells. In a specific range, the ratio of mesenchymal cells can affect the migration properties of epithelial cells. These data and methods highlight the reliability of time-series clustering in identifying consistent migration patterns during collective cell migration. It provides new insights into the epithelial–mesenchymal interactions that affect migration patterns. The method can combine with different algorithm to deeply generate more relationship in many fields.

APPENDIX 1

The simplified flowchart of the time-series clustering algorithm.



Step 1, import the coordinates of all cells that were manually labeled.

Step 2, the distances of the marked points are calculated to unify the cell positions before cell division. The observation process is cut into multiple observation windows. Each window is normalized to the trajectory.

Step 3, UMAP dimensionality reduction is performed on the high-dimensional time-series data. After the optimal k-value is determined using MSC, k-means clustering is performed.

Step 4, for each observation window the clustering results are combined with the original trajectories and used to analyze similar migration patterns.

APPENDIX 2

The simplified Python code of the time-series clustering algorithm.

Part 1: Import the manually marked cells' coordinates.

```
import pandas as pd
df_TGF_both = pd.read_csv('/Original.csv')
DM_manual = pd.read_csv(' /DM_manual.csv')
pip install tslearn
pip install umap-learn
```

Part 2: Preprocess of cells' coordinates.

Part 2.1: The distance between marker points is calculated to identify the occurrence of cell division.

```
classinformation = df_TGF_both['TGF'].unique()
for temp in classinformation:
   temp_data = df_TGF_both[df_TGF_both['TGF'].isin([temp])]
   exec('df_TGF%s = temp_data'%temp)
df_TGFp = df_TGFP
df_TGFm = df_TGFN.reset_index(drop=True)
import numpy as np
import math
df_TGFm1 = df_TGFm.set_index(['Track n','Slice n'])
Track_h1 = []
Track_h2 = []
Slice_h1 = []
dis_p = []
for i in range(6,94):
  for n in range(1,98):
    x_dt1 = df_TGFm1.loc[(i,n),'X']
    y_dt1 = df_TGFm1.loc[(i,n),'Y']
    for a in range(6,94):
      if a > i:
        x_dt2 = df_TGFm1.loc[(a,n),'X']
        y_dt2 = df_TGFm1.loc[(a,n),'Y']
        dis = math.sqrt((x_dt1-x_dt2)**2 + (y_dt1-y_dt2)**2)
        if dis < 10:
```

```
        dis_p1 = [dis]
        disx1 = pd.DataFrame([dis_p1])
        disx2 = disx1.apply(int, axis=1)
        dis_p.append(disx2)
        Track_h1.append([i])
        Slice_h1.append([n])
        Track_h2.append([a])
list_p = pd.DataFrame({'Track_h1': np.ravel(Track_h1),'Track_h2': np.ravel(Track_h2),
          'Slice_h1': np.ravel(Slice_h1),'Distance': np.ravel(dis_p)})
list_p1 = list_p.sort_values(by = ['Track_h1','Track_h2'],ascending=True).reset_index(d
rop=True)
df_TGFp1 = df_TGFp.set_index(['Track n','Slice n'])
Track_k1 = []
Track_k2 = []
Slice_k1 = []
dis_pr = []
for i in range(1,6):
  for n in range(1,98):
    x_dt1r = df_TGFp1.loc[(i,n),'X']
    y_dt1r = df_TGFp1.loc[(i,n),'Y']
    for a in range(1,6):
      if a > i:
        x_dt2r = df_TGFp1.loc[(a,n),'X']
        y_dt2r = df_TGFp1.loc[(a,n),'Y']
        disr = math.sqrt((x_dt1r-x_dt2r)**2 + (y_dt1r-y_dt2r)**2)
        if disr < 10:
          dis_p1r = [disr]
          disx1r = pd.DataFrame([dis_p1r])
          disx2r = disx1r.apply(int, axis=1)
          dis_pr.append(disx2r)
          Track_k1.append([i])
          Slice_k1.append([n])
          Track_k2.append([a])
list_pr = pd.DataFrame({'Track_k1': np.ravel(Track_k1),'Track_k2': np.ravel(Track_k2),
          'Slice_k1': np.ravel(Slice_k1),'Distance': np.ravel(dis_pr)})
list_p1r = list_pr.sort_values(by = ['Track_k1','Track_k2'],ascending=True).reset_index
(drop=True)
```

Part 2.2: Uniform cell marker point trajectories prior to division.

```
import copy
df_TGFb1 = df_TGF_both.set_index(['Track n','Slice n'])
```

```
df_TGFb2 = copy.deepcopy(df_TGFb1)
m = DM_manual['Track_mo'].values
n = DM_manual['Track_dt'].values
s = DM_manual['Slice'].values
for m,n,s in zip(m,n,s):
    for i in range(1,s+1):
      X_mo = df_TGFb2.loc[(m,i),'X']
      Y_mo = df_TGFb2.loc[(m,i),'Y']
      df_TGFb2['X'].loc[(n,i)] = X_mo
      df_TGFb2['Y'].loc[(n,i)] = Y_mo
df_TGFb3 = df_TGFb2.reset_index()
df_Bo_Uni = df_TGFb3.reset_index(drop=True)
```

Part 2.3: Normalization. At each observation period, all cell trajectories are moved to let their starting points to the origin (0, 0)

```
ob_win = 12
cell_index_b1 = df_Bo_Uni['Track n'].unique()
x_list_b1=[]
y_list_b1=[]
x_list_b2=[]
y_list_b2=[]
for name in df_Bo_Uni['Track n'].unique():
    X_list1 = df_Bo_Uni[df_Bo_Uni['Track n'] == name]["X"].values
    Y_list1 = df_Bo_Uni[df_Bo_Uni['Track n'] == name]["Y"].values
    for a in df_Bo_Uni['Slice n'].unique():
      if a % ob_win == 0:
        initial_X1 = X_list1[a-ob_win]
        initial_Y1 = Y_list1[a-ob_win]
        x_list_b1[a-ob_win:a] = X_list1[a-ob_win:a] - initial_X1
        y_list_b1[a-ob_win:a] = Y_list1[a-ob_win:a] - initial_Y1
        if a + ob_win > 97:
          initial_X1 = X_list1[96]
          initial_Y1 = Y_list1[96]
          x_list_b1[a:97] = X_list1[a:97] - initial_X1
          y_list_b1[a:97] = Y_list1[a:97] - initial_Y1
          break
    x_list_b2.extend(x_list_b1)
    y_list_b2.extend(y_list_b1)
df_relative1 = pd.DataFrame({'X relative': np.ravel(x_list_b2),'Y relative': np.ravel(y
_list_b2)})
df_Bo_Uni_Re = pd.concat([df_Bo_Uni, df_relative1], axis=1)
```

```
df_Bo_Uni_Re.to_csv('Both_Uni_Relative.csv')
from matplotlib.pyplot import MultipleLocator
import matplotlib.pyplot as plt
def plotting(df_TGFm, df_TGFp, X_column, Y_column, cluster_pred):
    cell_index_m = df_TGFm['Track n'].unique()
    x_list_m = [[] for i in range(len(cell_index_m))]
    y_list_m = [[] for i in range(len(cell_index_m))]
    for i, name in enumerate(cell_index_m):
        x_list_m[i] = df_TGFm[df_TGFm['Track n'] == name][X_column].values
        y_list_m[i] = df_TGFm[df_TGFm['Track n'] == name][Y_column].values
    if df_TGFp is not False:
      cell_index_p = df_TGFp['Track n'].unique()
      x_list_p = [[] for i in range(len(cell_index_p))]
      y_list_p = [[] for i in range(len(cell_index_p))]
      for i, name in enumerate(cell_index_p):
        x_list_p[i] = df_TGFp[df_TGFp['Track n'] == name][X_column].values
        y_list_p[i] = df_TGFp[df_TGFp['Track n'] == name][Y_column].values
    if X_column=='X' and Y_column=='Y':
      fig,ax = plt.subplots(figsize=(3.5,6))
    colors = ["dodgerblue","lawngreen","orangered","darkviolet","orange"]
    if  cluster_pred is not False:
      for i, color in enumerate(cluster_pred):
        ax.scatter(x_list_m[i], y_list_m[i], color=colors[color], linestyle='solid')
        ax.plot(x_list_m[i], y_list_m[i], color=colors[color], linestyle='solid')
    if df_TGFp is not False:
      for i in range(len(cell_index_p)):
        ax.scatter(x_list_p[i], y_list_p[i], color='black', linestyle='solid',)
    fig.tight_layout()
    return
Start_slice=[]
End_slice=[]
for start_slice, end_slice in zip(range(1,97,12),range(12,98,12)):
    Bo_Uni_Re = copy.deepcopy(df_Bo_Uni_Re)
    for i in range(1,start_slice):
      Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Slice n'].isin([i]))]
    for a in range(end_slice+1, 98):
      Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Slice n'].isin([a]))]
    for b in DM_manual['Slice'].values:
      track_del1 = DM_manual.loc[DM_manual['Slice']==b]['Track_dt'].unique()
      if end_slice <= b:
        for n in track_del1:
```

```
        Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Track n'].isin([n]))]
    data_2=[]
    data_3=[]
    for num in Bo_Uni_Re['Track n'].unique():
      data_2 = np.array(Bo_Uni_Re[Bo_Uni_Re['Track n']==num][["X relative","Y relative
"]])
      data_3.append([token for st in data_2 for token in st])
      Start_slice.extend([start_slice])
      End_slice.extend([end_slice])
    data_ori = np.array(data_3)
    data_name = Bo_Uni_Re['Track n'].unique()
    Uni_Re_2_P = Bo_Uni_Re[Bo_Uni_Re['TGF'].isin(['P'])]
    Uni_Re_2_N = Bo_Uni_Re[Bo_Uni_Re['TGF'].isin(['N'])]
    plotting(Bo_Uni_Re, Uni_Re_2_P, "X", "Y", cluster_pred)
    plt.show()
```

Part 3: Time-series clustering

Part 3.1: The optimal number of clusters of each observation window.

```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.colors
import umap
def WCSS(start_slice, end_slice):
    Bo_Uni_Re = copy.deepcopy(df_Bo_Uni_Re)
    for i in range(1,start_slice):
      Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Slice n'].isin([i]))]
    for a in range(end_slice+1, 98):
      Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Slice n'].isin([a]))]
    for b in DM_manual['Slice'].values:
      track_del1 = DM_manual.loc[DM_manual['Slice']==b]['Track_dt'].unique()
      if end_slice <= b:
        for n in track_del1:
          Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Track n'].isin([n]))]
    data_2=[]
    data_3=[]
    for num in Bo_Uni_Re['Track n'].unique():
        data_2 = np.array(Bo_Uni_Re[Bo_Uni_Re['Track n']==num][["X relative","Y relative
"]])
        data_3.append([token for st in data_2 for token in st])
    data_ori = np.array(data_3)
    data_name = Bo_Uni_Re['Track n'].unique()
```

```
Bo_Uni_Re1 = copy.deepcopy(Bo_Uni_Re)
orig = Bo_Uni_Re1.set_index(['Track n','Slice n'])
eu_dis = []
for i,j in enumerate(Bo_Uni_Re1['Track n'].unique()):
  if j == Bo_Uni_Re1['Track n'].unique()[-1]:
    break
  for d in range (1,j+1):
    Bo_Uni_Re1 = Bo_Uni_Re1[(-Bo_Uni_Re1['Track n'].isin([d]))]
  for m in Bo_Uni_Re1['Track n'].unique():
    euclid1 = 0
    for n in range(start_slice,end_slice+1):
      x_e1 = orig.loc[(j,n),'X relative']
      y_e1 = orig.loc[(j,n),'Y relative']
      x_e2 = orig.loc[(m,n),'X relative']
      y_e2 = orig.loc[(m,n),'Y relative']
      euclid = sqrt((x_e1-x_e2)**2+(y_e1-y_e2)**2)
      euclid1 = euclid1 + euclid
    eu_dis.append(euclid1)
k = len(eu_dis)
n_matrix = int((1+int((1+8*k)**0.5))/2)
half_matrix = np.zeros((n_matrix,n_matrix))
start_index = 0
for row in range(n_matrix-1):
  end_index = start_index+(n_matrix-1-row)
  half_matrix[row,row+1:] = eu_dis[start_index:end_index]
  start_index = end_index
symme_matrix = half_matrix + half_matrix.T
reducer = umap.UMAP(random_state=1, n_neighbors=5, min_dist=0.001)
embedding = reducer.fit_transform(symme_matrix)
WCSS = []
Scores = []
for k in range(2,9):
  estimator = KMeans(n_clusters = k, random_state = 1).fit(embedding)
  WCSS.append(estimator.inertia_)
  Scores.append(silhouette_score(embedding,estimator.labels_,metric='euclidean'))
X = range(2,9)
WCSS_2f = [float('{:.1f}'.format(i)) for i in WCSS]
fig, ax1 = plt.subplots(1, 1, figsize=(5,4))
ax2 = ax1.twinx()
ax1.plot(X, WCSS, 'o-', linewidth=3)
ax2.plot(X, Scores, 'o-', color='coral', linewidth=3)
```

```
    ax1.set_xlabel('Number of clusters (k)',fontsize=15)
    ax1.set_ylabel('WCSS',fontsize=15)
    ax2.set_ylabel('MSC', fontsize=15)
    plt.show()
for a, b in zip(range(1,97,12),range(12,98,12)):
  WCSS(a,b)
```

Part 3.2: UMAP dimension reduction and k-means clustering.

```
def period(start_slice, end_slice, leader_number, cluster):
    Bo_Uni_Re = copy.deepcopy(df_Bo_Uni_Re)
    for i in range(1,start_slice):
      Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Slice n'].isin([i]))]
    for a in range(end_slice+1, 98):
      Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Slice n'].isin([a]))]
    for b in DM_manual['Slice'].values:
      track_del1 = DM_manual.loc[DM_manual['Slice']==b]['Track_dt'].unique()
      if end_slice <= b:
        for n in track_del1:
          Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Track n'].isin([n]))]
    data_2=[]
    data_3=[]
    for num in Bo_Uni_Re['Track n'].unique():
       data_2 = np.array(Bo_Uni_Re[Bo_Uni_Re['Track n']==num][["X relative","Y relative
"]])
       data_3.append([token for st in data_2 for token in st])
    data_ori = np.array(data_3)
    data_name = Bo_Uni_Re['Track n'].unique()
    Bo_Uni_Re1 = copy.deepcopy(Bo_Uni_Re)
    orig = Bo_Uni_Re1.set_index(['Track n','Slice n'])
    eu_dis = []
    for i,j in enumerate(Bo_Uni_Re1['Track n'].unique()):
      if j == Bo_Uni_Re1['Track n'].unique()[-1]:
        break
      for d in range (1,j+1):
        Bo_Uni_Re1 = Bo_Uni_Re1[(-Bo_Uni_Re1['Track n'].isin([d]))]
      for m in Bo_Uni_Re1['Track n'].unique():
        euclid1 = 0
        for n in range(start_slice,end_slice+1):
          x_e1 = orig.loc[(j,n),'X relative']
          y_e1 = orig.loc[(j,n),'Y relative']
          x_e2 = orig.loc[(m,n),'X relative']
```

```
        y_e2 = orig.loc[(m,n),'Y relative']
        euclid = sqrt((x_e1-x_e2)**2+(y_e1-y_e2)**2)
        euclid1 = euclid1 + euclid
      eu_dis.append(euclid1)
  k = len(eu_dis)
  n_matrix = int((1+int((1+8*k)**0.5))/2)
  half_matrix = np.zeros((n_matrix,n_matrix))
  start_index = 0
  for row in range(n_matrix-1):
    end_index = start_index+(n_matrix-1-row)
    half_matrix[row,row+1:] = eu_dis[start_index:end_index]
    start_index = end_index
  symme_matrix = half_matrix + half_matrix.T
  reducer = umap.UMAP(random_state=1, n_neighbors=5, min_dist=0.001)
  embedding = reducer.fit_transform(symme_matrix)
  cluster_pred = KMeans(n_clusters=cluster, random_state=1).fit_predict(embedding)
  n=leader_number
  plt.figure(figsize=(6,6))
  plt.scatter(embedding[n:94, 0], embedding[n:94, 1], c=cluster_pred[n:], cmap = matpl
otlib.colors.ListedColormap(["indianred","purple","cornflowerblue","limegreen","orange
"]))
  plt.show()
  print('Cell ID:',data_name)
  print('Cluster ID:',cluster_pred)
  return data_name
for a,b,c,d in zip(range(1,97,12),range(12,98,12),(2,2,3,4,4,4,4,5),(3,4,5,4,4,5,4,3)):
 period(a,b,c,d)
```

Part 3.3: Clustering results combine with original trajectories to generate similar migration pattern.

```
Cell_ID=[]
Cluster_ID=[]
Start_slice=[]
End_slice=[]
for start_slice, end_slice, cluster in zip(range(1,97,12),range(12,98,12),(3,4,5,4,4,5,
4,3)):
  Bo_Uni_Re = copy.deepcopy(df_Bo_Uni_Re)
  for i in range(1,start_slice):
    Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Slice n'].isin([i]))]
  for a in range(end_slice+1, 98):
    Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Slice n'].isin([a]))]
```

```
    for b in DM_manual['Slice'].values:
      track_del1 = DM_manual.loc[DM_manual['Slice']==b]['Track_dt'].unique()
      if end_slice <= b:
        for n in track_del1:
          Bo_Uni_Re = Bo_Uni_Re[(-Bo_Uni_Re['Track n'].isin([n]))]
    data_2=[]
    data_3=[]
    for num in Bo_Uni_Re['Track n'].unique():
      data_2 = np.array(Bo_Uni_Re[Bo_Uni_Re['Track n']==num][["X relative","Y relative
"]])
      data_3.append([token for st in data_2 for token in st])
      Start_slice.extend([start_slice])
      End_slice.extend([end_slice])
    data_ori = np.array(data_3)
    data_name = Bo_Uni_Re['Track n'].unique()
    Bo_Uni_Re1 = copy.deepcopy(Bo_Uni_Re)
    orig = Bo_Uni_Re1.set_index(['Track n','Slice n'])
    eu_dis = []
    for i,j in enumerate(Bo_Uni_Re1['Track n'].unique()):
      if j == Bo_Uni_Re1['Track n'].unique()[-1]:
        break
      for d in range (1,j+1):
        Bo_Uni_Re1 = Bo_Uni_Re1[(-Bo_Uni_Re1['Track n'].isin([d]))]
      for m in Bo_Uni_Re1['Track n'].unique():
        euclid1 = 0
        for n in range(start_slice,end_slice+1):
          x_e1 = orig.loc[(j,n),'X relative']
          y_e1 = orig.loc[(j,n),'Y relative']
          x_e2 = orig.loc[(m,n),'X relative']
          y_e2 = orig.loc[(m,n),'Y relative']
          euclid = sqrt((x_e1-x_e2)**2+(y_e1-y_e2)**2)
          euclid1 = euclid1 + euclid
        eu_dis.append(euclid1)
    k = len(eu_dis)
    n_matrix = int((1+int((1+8*k)**0.5))/2)
    half_matrix = np.zeros((n_matrix,n_matrix))
    start_index = 0
    for row in range(n_matrix-1):
      end_index = start_index+(n_matrix-1-row)
      half_matrix[row,row+1:] = eu_dis[start_index:end_index]
      start_index = end_index
```

```
symme_matrix = half_matrix + half_matrix.T
reducer = umap.UMAP(random_state=10, n_neighbors=4, min_dist=0.001)
embedding = reducer.fit_transform(symme_matrix)
cluster_pred = KMeans(n_clusters=cluster, random_state=1).fit_predict(embedding)
Uni_Re_2_N = Bo_Uni_Re[Bo_Uni_Re['TGF'].isin(['N'])]
plotting(Bo_Uni_Re, Uni_Re_2_N, "X", "Y", cluster_pred)
plt.show()
```

LIST OF ACHIEVEMENTS

Publications

1. Zhuohan Xin, Masashi K. Kajita, Keiko Deguchi, Shin-ichiro Suye and Satoshi Fujita. Time-Series Clustering of Single-Cell Trajectories in Collective Cell Migration. *Cancers*, **2022**, 14, (19), 4587. DOI: 10.3390/cancers14194587.
2. Zhuohan Xin, Keiko Deguchi, Shin-ichiro Suye, and Satoshi Fujita. Quantitative Analysis of Collective Migration by Single-Cell Tracking Aimed at Understanding Cancer Metastasis. *International Journal of Molecular Sciences*, **2022**, 23, (20), 12372. DOI: 10.3390/ijms232012372.
3. Geometrically Controlled in Nanofiber Construct for Controlling Glioblastoma Invasion. In preparation.

Presentations

1. Zhuohan Xin, Masashi K. Kajita, Keiko Deguchi, Shin-ichiro Suye and Satoshi Fujita. Time-series Clustering of Single Cell Trajectories in Collective Cell Migration. *The 43rd Annual Meeting of the Japanese Society for Biomaterials and the 8th Asian Biomaterials Congress*. Nagoya, Japan, **2021**.
2. Zhuohan Xin, Masashi K. Kajita, Keiko Deguchi, Shin-ichiro Suye and Satoshi Fujita. Time-series Clustering of Single Cell Trajectories in Collective Cell Migration. *The 11th Hokuriku-Shinetsu Block Young Researchers' Conference of the Japanese Society for Biomaterials*. Ueda. Japan, **2022**.

ACKNOWLEDGMENT

First of all, I would like to express my heartfelt gratitude to all those who have helped me in pursuing my Ph.D. My sincere and hearty thanks and appreciations go firstly to my supervisor, Prof. Satoshi Fujita, whose suggestions and encouragement have given me much insight into my research.

In 2020, my doctoral career started right after the global outbreak of the epidemic, and since then the long struggle has begun. Everything started to become difficult and all the activities turned online. It was a challenge for me to complete the scientific task. I appreciate the help of Keiko Deguchi for the experimental support. Thanks to the detailed and professional help of Assistant Prof. Masashi Kajita, which enabled me to complete the important research content. I am also extremely grateful to Prof. Shin-ichiro Suye and all my lab members who have kindly provided me with assistance and companionship in the course of preparing my thesis. It has been a great privilege and joy to benefit from your personality and diligence, which I will treasure my whole life.

In addition, I would like to express my gratefulness to my family for their unfailing love and unwavering support for me while the entire family endured great hardship. This is my motivation to be ahead in life and a permanent reminder of my appreciation.

My gratitude to you knows no bounds.

2022.10

Fukui