Quantification of cell phenotype transition manifolds with information geometry

Miao Huang[1,2], Junda Ying[3], Yuxuan Wang[3], Haijun Zhou[1,2], Lei Zhang[3],Weikang Wang[1,2]

[1] CAS Key Laboratory for Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China

[2] School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[3]School of Mathematics, Peking University, Beijing 100091, China

Email: wangwk@itp.ac.cn

Key word:
Cell phenotype transition, Fisher information, manifold, critical point

Abstract

Cell phenotype transition (CPT) is a common occurrence in many biological processes, such as development, and the precise control of CPT is critical in these processes. Recent advancements in single-cell sequencing techniques have revealed that cell transition dynamics in development are confined to low-dimensional manifolds. However, no available method currently exists for quantifying the manifolds from experimental data. Current methods, such as manifold learning, can only preserve topology information and can only be used for visualization.

In this work, we present a novel method to quantify the cell phenotype transition manifolds using information geometry. Specifically, we transform single cells' high-dimensional expression vectors into probability distributions with Gaussian embedding. The Fisher-Rao metric is then naturally defined in this embedding space. With Gaussian embedding, we calculate the Ollivier-Ricci curvature of each single cell. Our analyses revealed that cells with low curvature are related to critical transitions.To further analyze the invariant characteristics of the manifold in CPT, we calculated the information velocity of each single cell based on RNA velocity. The high information velocity regions correspond to the low curvature regions, indicating that the

geometry can guide the dynamics of single cells on the manifolds. This method focuses on analyzing the invariant characteristics of the manifold in CPT, and the results illustrate that it can be a general approach for quantifying the CPT manifold.

Introduction

Biological systems are often complex, yet biological processes can achieve remarkable precision. One example is embryo development, where stem cells transition into different phenotypes to form organized organisms. It is well-known that cell phenotype transition (CPT) is confined to low-dimensional manifolds in phase space (1, 2, 3, 4, 5). Similarly,in the field of neural science, it is also found that neural dynamics is also on low-dimensional manifold(6, 7, 8). The underlying connectivity constrain the population activity of neurons to certain patterns. And these patterns can be decomposed to several independent neural modes(8, 9).

Waddington's landscape analogy is a well-known metaphor for describing embryo development. To study these underlying low-dimensional manifolds, recent advances in single-cell RNA sequencing have made it possible to quantitatively describe the landscape (10, 11). Most current methods define scalar potentials and do not fully capture the dynamics of development, as sequencing is destructive to cells. To address this issue, algorithms have been developed to infer RNA velocity using counts of spliced and unspliced RNA, enabling the learning of a vector field in the gene space of individual cells (4, 12, 13). However, RNA velocity vectors are typically projected onto low-dimensional embedding spaces such as t-SNE and UMAP for visualization, which lack biological meaning and the ability to be studied quantitatively.

In mathematics, Gauss and Riemann raised up the concept of intrinsic geometry of manifold, where the Riemann metric determines the structure of manifold. Defining a Riemann metric of the low-dimensional manifold with real data can be challenging. Current manifold learning methods only preserve topology information in the original space and are often just used as dimension reduction techniques that project high-dimensional data to a low-dimensional embedding space.

Fisher information is a quantitative measure of the information that observables carry about distribution parameters (14). Rao proved that Fisher information can be defined as a Riemann metric in probability space (14, 15). Amari develop information geometry based the Fisher-Rao metric (16, 17). From the perspective of the master equation, the dynamics of cell phenotype transition can be seen as the evolution of a distribution in phase space. By transforming single

cell gene vectors into probability space, the Riemann metric can be defined with Fisher information. One simple strategy for approximating the probability of single cells' gene vectors is Gaussian mixture model (GMM), where each Gaussian distribution component represents the probability density of a stable cell state (18, 19). However, GMM is a coarse-graining approach that is incapable of yielding smooth manifolds. To address this limitation, we utilize the Gaussian embedding method, in which each individual cell is modeled as a multivariate Gaussian distribution. This allows us to calculate the Fisher information of each cell and define the Riemann metric in a smooth and continuous manner (20, 21).

In our study, we introduce a novel approach to quantitatively analyze low-dimensional manifolds from single cell data. Initially, we transform each single cell's sequencing data into a multivariate Gaussian distribution using Gaussian embedding, calculate the Fisher information of each cell and quantify the manifold of CPT (20, 21). By representing each cell as a Gaussian distribution, we can capture its uncertainty and calculate the information distance between individual cells explicitly (14, 22). Moreover, we can define the Fisher metric on pre-defined variables such as eigengenes, using the reproducing kernel Hilbert space method (RKHS) or neural networks with backward propagation (4, 23). By projecting the manifold onto coordinates with biological meaning, we can learn the relationship between critical characteristics and genes or gene communities. Unlike previous analyses, we focus on the invariant properties of the CPT manifold, such as curvature and information velocity (24, 25, 26, 27).

Methods

1 Gaussian embedding of single cell RNA sequencing

Single cell RNA sequencing is a technique that enables the measurement of mRNA abundance in individual cells. Each cell can be represented as a point in a high-dimensional gene expression space, and we calculate pair-wise Euclidean distances between cells to construct a k-nearest neighbor (kNN) graph. In our study, we choose different values of k (10, 15, 20) to validate the robustness of the method (supplement Fig. 2d).

To represent each cell as a continuous probability density function in a latent space, we use the Gaussian embedding method, which has previously been applied in natural language processing (21). Our approach utilizes neural networks to learn the Gaussian embedding of each cell based on its expression and its adjacency with other cells in the gene space R^D (D is the original high-dimensional gene expression space) (20). Each cell is then represented by a multivariate Gaussian distribution in an L-dimensional space R^L, where L is much lower than D (L=3 in application).

In our study, we use a neural network trained with an energy-based loss function to learn the Gaussian embedding of each cell. The energy function represents the pairwise distance between the embedded distributions. In the original work, the energy function is Kullback-Leibler divergence $E_{ij} = D_{KL}(P_i \| P_j)$ (20). In our work, we use the Fisher distance as the energy function,

$$E_{ij} = \sqrt{2\sum_{d=1}^{D}(ln\frac{|(\mu_{id}/\sqrt{2},\sigma_{id}) - (\mu_{jd}/\sqrt{2},-\sigma_{jd})| + |(\mu_{id}/\sqrt{2},\sigma_{id}) - (\mu_{jd}/\sqrt{2},\sigma_{jd})|}{|(\mu_{id}/\sqrt{2},\sigma_{id}) - (\mu_{jd}/\sqrt{2},-\sigma_{jd})| - |(\mu_{id}/\sqrt{2},\sigma_{id}) - (\mu_{jd}/\sqrt{2},\sigma_{jd})|})^2}.$$

And the loss function is defined as $Loss = \sum_{i}\sum_{k<l}\sum_{j_k \in N_{ik}}\sum_{j_l \in N_{il}}(E_{ij_k}^2 + exp(-E_{ij_l}^2))$, here N_{ik} represent the set of nodes j_k that shortest paths between i and j_k equal k.

2 Riemann manifold with Fisher metric

By representing single cell with Gaussian embedding, we calculate the Fisher information of each cell, which serves as the Riemann metric in probability space. Defining connections like Levi-Civita connection which is essential for curvature in this space with experimental data is still challenging. Hence, we use an alternative method proposed by Yann Ollivier. This method estimates the curvature on a graph by assigning each node a probability measure, and then calculates the coarse Ricci curvature using the following equation (26):

$$\kappa(a,b) = 1 - \frac{W(P_a, P_b)}{d(a,b)}$$

Where $W(P_a, P_b)$ represents the Wasserstein distance between the probability measures of node a and node b and d is the length of shortest path between node a and node b (24). The Wasserstein distance is a mathematical concept that measures the minimum cost of transportation when transforming a probability distribution u to a probability distribution v. The Ricci curvature, on the other hand, controls the overlap of two balls or the overlap between distributions, and is directly related to the Wasserstein distance.

By transforming each single cell into a multivariate Gaussian distribution with diagonal covariance, we can directly calculate the Wasserstein distance between neighbor nodes

$$W(P_a, P_b) = (|\mu_a - \mu_b|^2 + tr(\Sigma_a + \Sigma_b - 2(\Sigma_b^{1/2}\Sigma_a\Sigma_b^{1/2})^{1/2})^{1/2}.$$

In our study, $d(a,b)$ is the Euclidean distance between neighboring cells in gene space. The coarse Ricci curvature of each node is then calculated as the average of the coarse Ricci curvature of all its edges (28).

3 Information velocity and Information length

The information velocity is defined with the following equation:

$$v_p = \sqrt{\frac{d\theta^i}{dt} g_{ij} \frac{d\theta^j}{dt}}$$

where v_p is the information velocity, $\frac{d\theta}{dt}$ represents the velocity of the parameter and θ is (μ, σ) in Gaussian embedding. The information velocity measures the variation rate of the probability density function and reflects the speed of information variation along a path (25, 27, 29). Integration of information velocity along a path in the manifold is defined as information length.

$$L = \int \sqrt{\sum_{ij} g_{ij} d\theta^i d\theta^j} = \int \sqrt{\sum_{ij} g_{ij} \frac{d\theta^i}{dt} \frac{d\theta^j}{dt}} = \int v_p dt.$$

Information length is already used in analyzing thermodynamics and attractor structure (25, 27). To compute the information length between two cells, it is necessary to define the path that links them. Commonly, this path is the geodesic, which represents the shortest route connecting the points. The measure of information along the geodesic in the space of Gaussian embedding is referred to the Fisher distance. We employ the cell graph approach to calculate the information distance: the lengths of the graph edges are determined using the Fisher distance, and the distance between any two cells is then defined as the shortest path length on the cell graph. We assessed the correlation between transcription factors and the variability of information length along the path that connects the initial and terminal states in CPT processes. Through this method, transcription factors that are pivotal in modulating information variation can be identified and selected.

We also calculated the information velocity of each cell using RNA velocity. As RNA velocity reflects the direction of single cell along the path of CPT in the gene expression space, the information velocity of single cell represents the speed of information variation along the transition path of CPT. With RNA velocity $\frac{dx}{dt}$, $\frac{d\theta}{dt}$ can be represented by chain rule $\frac{d\theta}{dt} = \frac{d\theta}{dx}\frac{dx}{dt}$. The $\frac{d\theta}{dx}$ is calculated with *autograd* function of Pytorch or RKHS. For the eigen-gene or PCA coordinates, the velocity is defined as $\frac{d\theta}{dt} = \Sigma_m w_m \frac{dx_m}{dt}$, where w_m is the weight of gene x_m in the corresponding velocity genes.

4 Eigengene analysis

To study the relationship between critical characteristics and genes or gene communities, we also calculate the co-expression profile of different genes based on correlations. Using hierarchical clustering, genes are separated into several modules. The eigengene (EG) is defined as the value of the first principal component of each module $Eg = W_m g_m$ (30, 31). As each gene's velocity is known, the velocity of eigen gene can be calculated.

5 Algorithms for transformation between different coordinate systems

(1) Reproducing kernel Hilbert space

In our study, we use a vector field learning method that is trained with sparse vector data pairs to learn a vector value function in a function Hilbert space. The function Hilbert space is composed of a basis of vector field functions. The optimal function is a linear combination of the base functions in this space (4, 23). Using the Reproducing Kernel Hilbert Space (RKHS) method (code adopted from *dynamo*), we can calculate compute the transformation between (μ, σ) and other coordinates.

(2) Regression with neural networks

We also use a neural network to learn the mapping function between vector pairs. Specifically, we train a simple multi-layer perceptron neural network in PyTorch, where the input is the user-defined coordinates and the output is the (μ, σ) of the Gaussian embedding. Using the *autograd* in PyTorch, we can compute the transformation between (μ, σ) and other coordinates. This allows us to compute the Fisher metric in the coordinates with biological meaning such as eigengenes and principal components, which can further facilitate the quantitative analysis of CPT.

6 Test data generation

In our study, we use several toy models to generate data to verify the feasibility of the Fisher information manifold method. Specifically, we use three different models:

(1) Saddle surface: We randomly sampled data points on a saddle surface, where the original data had a dimensionality of three and lied on a 2D surface $z = k((x - x_0)^2 - (y - y_0)^2)$. And we

add additional seven dimensions of small random data following a Gaussian distribution. These seven dimensions account for minor variances of the data, and the resulting ten-dimensional data points still lie on the 2D saddle surface.

(2) Ellipsoid: We randomly sampled data points on an ellipsoid surface, where the original data had a dimensionality of three and lied on a 2D surface

$x = asin\phi cos\psi$
$y = bsin\phi sin\psi$
$z = acos\phi$
where (ϕ, ψ) are angles and $b > a$.

Similar to the saddle surface model, we added another seven dimensions of noise.

(3) Four-well potential landscape: We sampled data on a 4-well potential surface $U = kx^4 - x^2 + ky^4 - (y - y_0)^2$, which was generated by performing trajectory simulation with Langevin equation. After simulation, the trajectory data were also added with another seven dimensions of noise.

By applying the Fisher information manifold method to these toy models, we are able to verify its feasibility and effectiveness in studying the manifold of cell phenotype transition. These models allow us to evaluate the performance of the method in different scenarios and gain insights into its strengths and limitations.

7 Information velocity of saddle point

Intuitively, Gaussian embeddings capture the local uncertainty of points on the energy landscape. Here we transform the landscape picture into information manifold. However, Gaussian embeddings of points on energy landscape have no direct correspondence, i.e., there is no ground truth for the Gaussian embedding and the subsequent analysis of information velocity. Here we use another method to estimate the Gaussian distributions around saddle points and the information velocity based on dynamics on the landscape.

In a simple double-well potential ($U = ax^2 + by^4 - (y - y_0)^2$) where the depths of the two basins are different, multiple trajectories are initialized from the shallow basin. The evolution of trajectories follows the Langevin equation. The mean and standard deviation of all the simulated trajectories at each time point are calculated and are used as estimation (ground truth) of parameters (μ, σ) in Gaussian embedding. The standard deviations around the saddle point have high values. The velocity of $\frac{d\theta}{dt}$ are derived by using moments equation (18). And we can obtain the information velocity around the saddle point.

8 Density estimation

The local cell density is estimated with k-nearest neighbor and Gaussian kernel (32). The local density is defined with $\rho_i = \Sigma_j exp(-d_{ij}^2/d_c^2)$. Here d_{ij} is the distance between cell i to its neighbor cell j and d_c is a cutoff distance (we set d_c as the maximum value between all neighbor cell pairs). As the experiment data is noisy, we introduce a smoothing step. For a scaler function f like curvature and information velocity defined on single cell, its value is smoothed by the following equation $f_i^s = \Sigma_j f_j exp(-d_{ij}^2/d_c^2)/\rho_i$.

9 Dataset preprocessing

The scRNA-seq datasets of dentate gyrus neurogenesis, pancreatic endocrinogenesis, and EMT of A549 treated by TGF-β were obtained from the GEO website with GEO number GSE95753 (33), GSE132188 (34), GSE121861 (35) respectively.

To focus on the process of CPT in each dataset and reduce the influence of genes that are not related with CPT processes, we use two methods to select the genes that show significant variations in CPT. The first one is selecting genes that show switch-like behaviors by using the method in (36). The other one is DUBStepR based on the correlation between genes which can filter out those genes that are not cell-type-specific (37).

Results

1 Overview of Fisher information manifold

In this study, we employ neural networks to transform high-dimensional single cell data into low-dimensional multivariate Gaussian distributions. To accomplish this, we adopt a two-layer neural network following the Gaussian embedding method (20). First, we represent each single cell as a vector X_i in the multidimensional expression space (Fig. 1a bottom), and compute the pairwise distance $|X_i - X_j|$ between cell i and cell j. Next, we use the k-nearest neighbor cell graph (Fig. 1a top) as input in the neural network (Fig. 1b top). The neural network model is trained with an energy-based learning method (Fig. 1b bottom) (38). The transformed Gaussian distributions represent the uncertainty of cell nodes and reflect the diversity of their neighborhoods separately (Fig. 1c) (20).

The Gaussian embeddings together with Fisher-Rao metric can be used to quantify different properties of the CPT manifold like curvature and information velocity (Fig. 1d). Unlike current manifold learning methods, this method is not designed for visualization. Although we can simply plot the projected high-dimensional cell vectors onto a 2D or 3D (μ, σ) space for visualization purposes, it is important to note that simply projecting high-dimensional manifolds onto a lower-dimensional space can be misleading or even conceptually flawed. This is because lower-dimensional projections cannot accurately keep the true structure of the high-dimensional manifold. Different from previous studies that rely on visualization of projections, our approach enables us to study the intrinsic geometry of the manifold quantitatively and identify invariant properties that are independent of coordinate transformation.

2 Test of Fisher information manifold method

To evaluate the effectiveness of the Fisher information manifold method, we apply it to several test datasets. The first dataset comprises of data points sampled on a saddle surface (Fig. 2a top). The second dataset consists of data points sampled on the surface of a 3D ellipsoid (Fig. 2a middle). We also sample data on a four-well potential surface and simulate multiple trajectories (Fig. 2a bottom). For all the test datasets, we add another 7D random data (Gaussian distribution with small amplitude) as noise to form a 10D space, and construct kNN graphs using the generated samples. The Gaussian embeddings for the toy datasets are calculated following the

procedures in the Methods. Here we plot the embedded data of saddle surface in the space of Gaussian embedding (Fig. 2b and supplement Fig. 2a).

For comparison, we also apply current manifold learning methods, including Isomap, t-SNE, and UMAP to the test datasets (supplement Fig. 2b). To compare the different methods, we calculate the proportion of pairwise Euclidean distances in the projected space to the pairwise Euclidean distances in the original space. As manifold learning methods can only preserve global topology and serve for visualization, the distributions of the relative distance ratios show large variations (supplement Fig. 2c left). But the dispersion of relative proportion of pair-wise distances in the original space to Fisher distances between the transformed distributions are significantly lower (supplement Fig. 2c right), which indicates that Gaussian embedding can better preserve the local metrics and provide a quantification framework for further analysis.

Curvature is a fundamental property in Riemann geometry, but its calculation requires definition of connection, which is difficult to define explicitly. To address this problem, we use coarse Ricci curvature (CRC) for analyzing curvature on complex networks(24, 26) (Methods). In the saddle surface data, data points around saddle points have lower CRC values (Fig. 2c top). In the ellipsoid data, we observe that the data points around both ends of the major axis exhibit higher CRC values compared to data in other regions (Fig. 2c middle). In the four-well potential landscape, the data points on the barriers between basins have lower CRC values, consistent with the findings in the saddle surface data (Fig. 2c bottom). Previous analysis on complex networks has shown that CRC values of edges between nodes in different communities are mostly negative, while edges inside each community usually have positive CRC values (39). The results revealed that in the intermediate region, nodes exhibit lower CRC value.

In addition, we found that in the four-well potential data, the data points in the basins show high CRC values (Fig. 2c bottom). An intuitive explanation is that data points locate inside basins share similar neighbors and their embedded distributions are close, which leads to short Wasserstein distances and high CRC values.

To investigate dynamics on the manifold, we use information velocity defined as

$v_p = \sqrt{\frac{d\theta^i}{dt} g_{ij} \frac{d\theta^j}{dt}}$ to quantify it (Methods). Information velocity measures the variation rate of probability density function. It can be initially calculated in the (μ, σ) coordinates. Based on simulation of trajectories on a double-well potential (Fig. 2d top) (Methods), we found that the information velocity around saddle point show is high (Fig. 2d bottom). That is to say, the variation rate of probability density function near the saddle point is fast, which is just the opposite to the CRC value (Fig. 2d middle).

These conclusions on toy datasets will be used for identification of meta-stable cell states and critical points in CPTs of single-cell RNA sequencing data.

3 Fisher information of single cell RNA sequencing data

We first apply Gaussian embedding method to process single-cell RNA sequencing data of dentate neurogenesis (Fig. 3a). During the process of neurogenesis in dentate gyrus, radial glia-like cells differentiate through nIPCs, neuroblast 1 and 2, immature granule cells, and eventually into mature granule cells (33).

For visualization purposes, we calculate the pseudo-time of each single cell (Fig.3b) and plot the single-cell data in the space of Gaussian embedding (Fig. 3c and supplement Fig. 3a). We also show the Fisher information of μ_1 in Fig. 3d (Other elements in Fisher information matrix are shown in supplement Fig. 3b).

The coordinates (μ, σ) lack direct biological interpretation. To study the relationship between the geometry of CPT manifold and the biological characteristics of the system, it is necessary to calculate the Fisher metric in different coordinate systems, especially those with explicit biological meaning. Here we perform coordinate transformation to obtain the Fisher metric

$$g_{ij} = \int_X \frac{\partial logP}{\partial \theta} \frac{\partial \theta}{\partial u_i} \frac{\partial logP}{\partial \theta} \frac{\partial \theta}{\partial u_j} P dx$$

in the new coordinates $\{u\}$ (calculation of partial derivative $\frac{\partial \theta}{\partial u_i}$ is in Methods) (4, 23).

We first transform the (μ, σ) coordinates to principal component (PC) coordinates. The Fisher information is an useful tool for analyzing informative orthogonal parameters (40). With Gaussian embedding, we can obtain the diagonal Fisher matrix directly. However, the Fisher matrices in PC coordinates may not necessarily be diagonal after coordinate transformation. If a non-diagonal element g_{ij} $(i \neq j)$ of Fisher matrix equals zero, θ_i and θ_j are defined as information orthogonal parameters (40). By analyzing the Fisher matrix in coordinates with biological meaning, we can learn about the orthogonality between these parameters. As shown in Fig. 3e, the Fisher information between PC1 and PC2 in relatively low (around 0) most of time except there exists a peak where pseudo-time equals 0.6. And this is also the time when the value of PC1 and PC2 show drastic changes.

We also perform coordinate transformation from (μ, σ) to eigengene coordinates. The value of eigengenes is the PC1 of gene modules which are calculated with clustering on gene correlation (Fig. 3f) (Methods). We found that in the neurogenesis process, the non-diagonal element g_{12} in the Fisher matrix of eigengenes are almost 0 all the time. The information orthogonality between both eigengenes reflects the different roles of eigengenes in different stages of this CPT process. Similar with that of PC coordinates, the non-diagonal element g_{12} also exhibits a peak when pseudo-time equals 0.6 (Fig. 3g). The Fisher information and information orthogonality are not static quantities as Fisher matrix varies. Through Fisher information analysis, we can gain different perspectives on the dynamics of the CPT process.

As CPT proceeds, the accumulated variation of information i.e. information length can be calculated (supplement Fig. 3c) (Methods). Specifically, the transcription factors that are correlated with information length probably play important regulatory roles in CPT. This could be a potential method for analyzing gene regulation beyond other methods like correlation or causal relation inference. We found that Zeb1 shows high correlation with information length (Fig. 3h & 3i), which is consistent with the previous studies on Zeb1's import roles in neurogenesis.

We also apply Fisher information analysis on two additional datasets. One is the development of pancreatic endocrine cells (Fig. 4a top and 4b top)(34). During embryonic development Ngn3-low progenitors first transform into Ngn3-high precursors then Fev-high (Fev+) cells. The latter further

develop into endocrine cells including glucagon producing α-cells and β-cells. The variation of Fisher information of μ_1 is shown in Fig. 4c top. And the Fisher information of the two eigengenes (non-diagonal element) exhibits one local minimum value and one local maximum value as CPT proceeds (Fig. 4d top). And both local extrema are deviated from zero, which represents the complex interplay between the eigengenes. The synergetic influences of both eigengenes on the probability distributions varies as CPT proceeds. The correlations of transcriptional factors with information length are shown in Fig. 4e left.

The other is an induced in-vitro transition process, the epithelial-to-mesenchymal transition (EMT) of human A549 cells treated with TGF-β for different durations (35) (Fig. 4a bottom and 4b bottom). In this dataset, A549 cells were treated with TGF-β for different durations (0 day, 8 hours, 1 day, and 3 days). At the initial stage (0 day), most A549 cells are in the epithelial state. After three days of treatment, most cells transition into the mesenchymal state. Comparing with the development datasets, the cancer cells are more heterogeneous, such that the Fisher information of μ_1 exhibits significant different values in the 8h sample (Fig. 4c bottom). The non-diagonal element between the two PCs is non-zero at the beginning of EMT (mostly 0 day). And the two PCs become information-orthogonal after TGF-β treatment (Fig. 4d bottom). The transcriptional factor that are highly correlated with information length are marked in Fig. 4e right.

4 Fisher information manifold can identify critical points

Fisher information analysis on specific coordinates can reveal the regulations in CPT. To learn more about the geometry of the information manifold, we further calculated the coarse Ricci curvature (CRC) on the single-cell RNA sequencing datasets (Fig. 5a) (Methods).

In the dentate gyrus data, we initially calculate the CRC of edges in the cell graph and transform the edge curvature to cell curvature by averaging the curvatures of edges that are linking with the cell. Cells with the lowest and highest pseudo-time values displayed high CRC values, indicating that these cells underwent a transition from one stable state to another stable state. The low-CRC region connects the two stable states (Fig. 5b). Following the test results on toy datasets, we found that this low-curvature region represents the region around the saddle node. In dynamical

systems, saddle nodes connect different attractors or basins. Therefore, this low-CRC region reflects the intermediate or transition states.

For the endocriogenesis data, we identified two regions where the CRC values are low (Fig. 5c left). The two low-CRC regions connect to the high-CRC regions, which reflect the stable and intermediate states in this process. In the EMT dataset, curvature analysis reveals that cells treated with TGF-β for one day have the lowest CRC value, and it connects the two high-CRC regions that are the epithelial and mesenchymal states (Fig. 5c right). The intermediate state is commonly recognized as a partial EMT state, which is related to the collective migration of tumors (41, 42).

As shown in Fig. 2d, the high CRC characterizes the basins on the energy landscape. And the initial and final cell populations in dentate gyrus neurogenesis have high CRC values. We analyzed genes that are highly correlated with CRC values in the marked regions (red box in Fig. 5b), i.e. the critical point and basins. These genes severed as markers of the corresponding cell states in these regions. And the three groups of marker genes belong to the three eigengenes separately (except a few genes in region ii belong to the third eigengene), which indicates the modularized regulation in CPT (Fig. 5d).

It should be noted that CRC is an approximation of Ricci curvature, and it is not based on Fisher metric. But the analyses reveal its connection with the metric. The low-CRC region overlaps with the region where non-diagonal element of Fisher information matrix shows the peak (Fig. 3e & 3g).

5 Geometry guides the dynamics of single cell on CPT manifolds

CRC reveals the structure of the CPT manifold, with low-CRC regions closely related to saddle nodes or barriers between basins. And deriving CRC is independent of dynamics information. We wonder if this geometric feature can reveal the dynamics of single cells on the manifold.

To gain more dynamics information from snapshot sequencing data, Gioele La Manno and his collaborators proposed the RNA velocity method (12). The RNA velocity method assumes simple dynamics between spliced RNA and unspliced RNA and estimates the time derivative of

spliced RNA counts. This method has been further developed by several labs (4, 13). With RNA velocity, we can calculate the information velocity (IV) (Fig. 6a) (Methods) (25, 27, 29).

We found that in all the analyzed CPT processes, cells in the low-CRC region have high IV values, indicating that the high IV value corresponds with the low-CRC value (Fig. 6b). As the low-CRC region is around the saddle node or barrier, probability distributions on the opposite sides of the barrier are distinct. The transition between the two distributions would be fast, leading to a high IV value in this region. In the basins, the IV values are low, indicating that cells are in stable states, and the distribution shows little variation. The results reflect the geometry of underlying manifold guides the dynamics of single cells in CPTs. For better illustration, we plot the CRC and IV of dentate gyrus neurogenesis along pseudo-time in one figure (Fig. 6c).

Additionally, we performed coordinate transformations and calculate IV in different coordinate system. We did dimension reduction with Non-negative matrix factorization (NMF) and calculated the Fisher metric in this coordinate system. The velocities of each dimension can be obtained by the linear combination of different velocity genes' RNA velocity. We found that the distribution of RNA velocity is similar to that in the (μ, σ) coordinates (Fig. 6d). However, the IV value is not necessarily related to the high absolute value of RNA velocity. We calculated the norm of the RNA velocity vectors in the PCA space and plotted the distribution (supplement Fig. 6). There is no necessary connection between its distribution and the distribution of IV. From the definition of IV, it is evident that it is a value that is independent of coordinates. Our analyses proved that it is invariant under coordinate transformation. It should be noted that the choice of coordinates can affect the absolute value of IV.

The information manifold emerges from the complex interactions between genes. Specifically, the saddle region with low-CRC and high IV emerges from the gene regulation mechanism. The eigengenes reflect the control modules in the gene regulation network, and the switching of control modules plays an important role in cell state transition. We plotted the eigengene values across the cells and found that the eigengenes form a relay in the dentate gyrus neurogenesis (Fig. 6e). Furthermore, we found that the low-CRC region or high IV region overlaps with the transitory region between eigengenes (Fig. 6e).

Discussion

The quantification of underlying manifold of CPT has emerged as a new field in recent years. In this study, we developed a method based on information geometry to study the intrinsic properties of the CPT manifold. Unlike previous manifold learning methods that are only for visualization in low-dimensional embedding space, we studied the CPT manifold from the perspective of intrinsic geometry and did not rely on visualization. We use the Gaussian embedding method transform the single-cell vector into multivariate Gaussian distributions. With the transformed distribution, Fisher metric and CRC can be calculated. We found that the geometry of manifold, like CRC, is related to dynamical properties like IV in CPT. Moreover, the CRC and IV are invariant under coordinate transformation.

Calculation of the Fisher metric requires building a cell graph from experimental data. The sampling bias can decrease the accuracy of Gaussian embedding. For instance, the local cell density may alter the neighbor diversity, which can affect the estimation of the Fisher metric. Although we found that CRC is not related to local cell density, the sampling bias should be taken into consideration, especially when handling time-course datasets. Furthermore, the sampling is constrained by experimental techniques, and we cannot extrapolate the results into the regions where there are no cells. Here, we focused on using the information of local neighborhood to investigate the properties of single cells.

The intrinsic geometry of the CPT manifold is related to the intrinsic dimension. Estimating the number of intrinsic dimensions is still an open question. An accurate intrinsic dimension can increase the training efficiency of Gaussian embedding , while an incomplete coordinate system probably distort the calculation of the Fisher metric. We tested a range of numbers of latent dimensions to ensure that the transformation is robust. But the method for estimation of intrinsic dimension and its reliability will be further explored.

The Fisher information can be used as order parameter in phase transition. The benefit of using Fisher information is that it does not require defining Hamiltonian or ensemble function as in

equilibrium physics (43). Hence it can be applied on non-equilibrium systems more easily. The peak of non-diagonal elements in the Fisher matrix is similar with the divergence of order parameter in phase transition, which indicates CPT probably share similar characteristics with phase transition. The Fisher information offers a powerful tool for analyzing non-equilibrium phase transition in biological systems.

It is worthy to note that Duluxan Sritharan et. al proposed a method to calculate a Riemann curvature with extrinsic differential geometry (44). Different from the extrinsic geometry method, the method presented here is from the perspective of intrinsic geometry which is invariant under coordinate transformation. It allows us to study the system be transforming into different biological meaningful variables as proposed in the paper like eigengenes. More importantly, we discovered the relation between geometry and dynamics.

Nowadays, RNA velocity has become a powerful method for learning the dynamics of CPT. However, it relies on some strong assumptions and suffers from some flaws when applied to different datasets. Given that the geometry of the CPT manifold is related to the cells' dynamics on the manifold, this method can be further developed to regularize the inferred dynamics like the projection of RNA velocity on the embedding space.

In this work, we mainly analyzed the cell transition from one phenotype to another phenotype. In development it is common that stem cells differentiated into several cell types, which exhibit more complex manifold. Lineage tracing and reconstruction is also a vibrant research field. For short, our preliminary results show that Fisher information can also be used to analyze this kind of manifold (data not shown), which will be further accomplished in future studies.

Figures

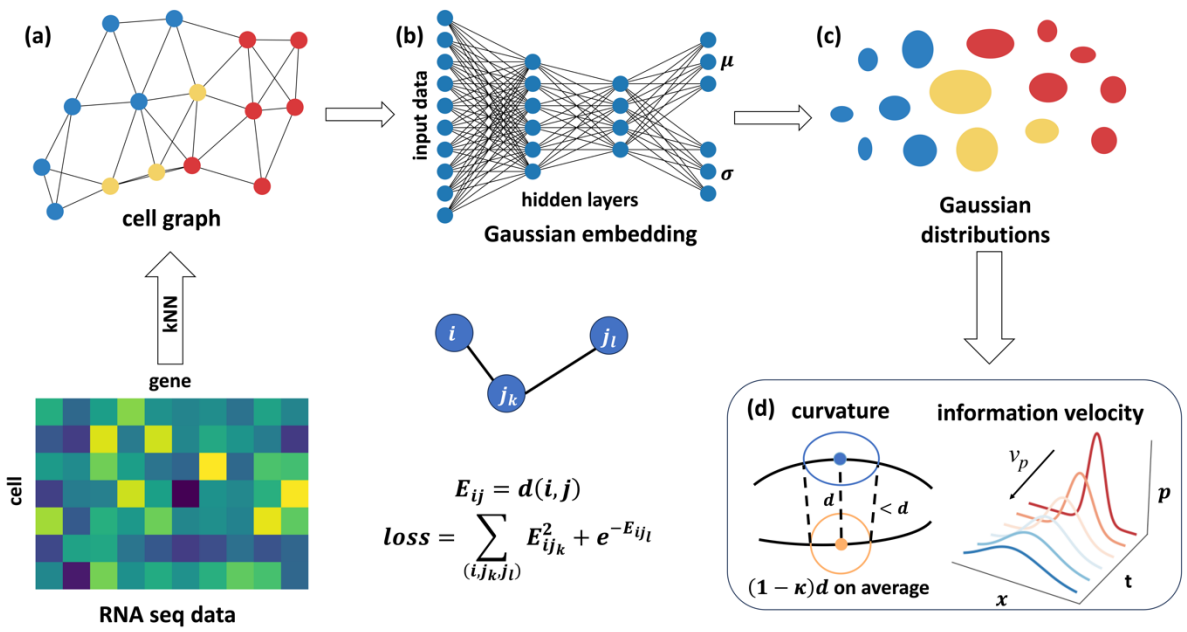

Figure 1 Schematic plot of the Fisher information manifold method, outlining the process of constructing the cell graph, using the Gaussian embedding method, representing single cells as distributions, calculating the metric, finding the transition path of minimum length, and computing the information velocity.

(a) The construction of a KNN cell graph with single cell RNA sequencing data.

(b) The network structure of the Gaussian embedding and training method.

(c) Gaussian embedding of the cell graph by representing single cells as multivariate Gaussian distributions.

(d) Illustration of the concepts of CRC and information velocity.

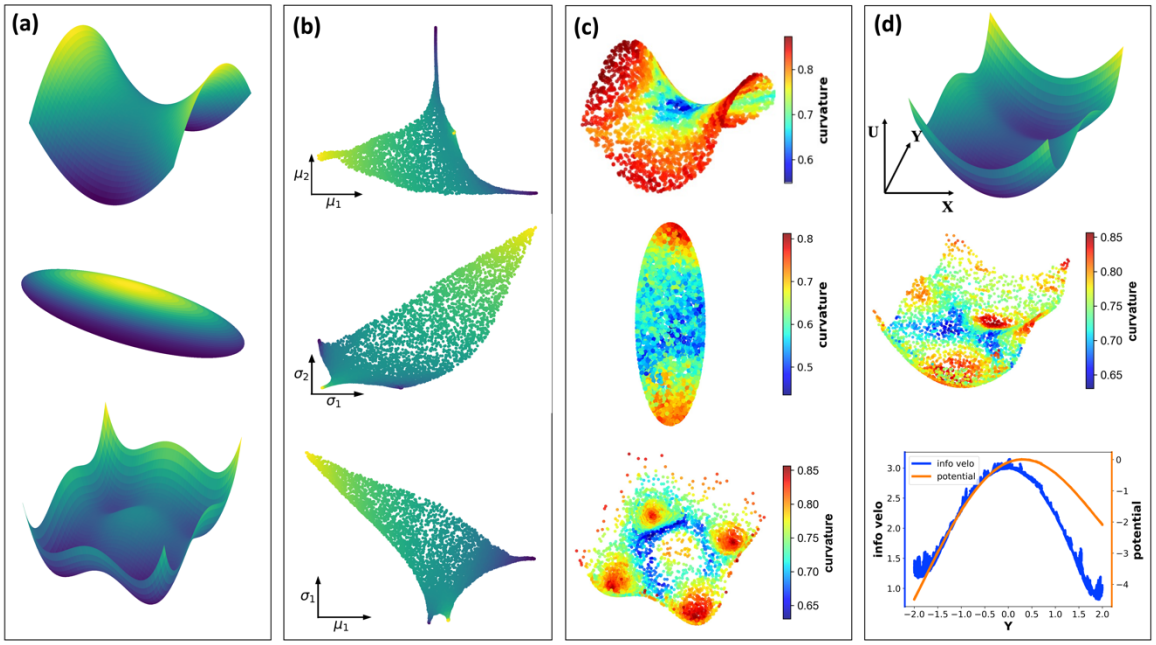

Figure 2 Test of the method on ellipsoid, saddle surface, and 4-well potential.

(a) Data points sampled on an ellipsoid surface (top), saddle surface (middle) and simulated distribution of data on a 4-well potential landscape embeddings (bottom).

(b) Embedding of saddle surface onto the Gaussian embedding space with different coordinates.

(c) CRC distributions of different test datasets, including the saddle surface (top), ellipsoid surface (middle) and 4-well potential (bottom).

(d) CRC distribution (middle) and Information velocity of saddle point (bottom) in double well potential (top).

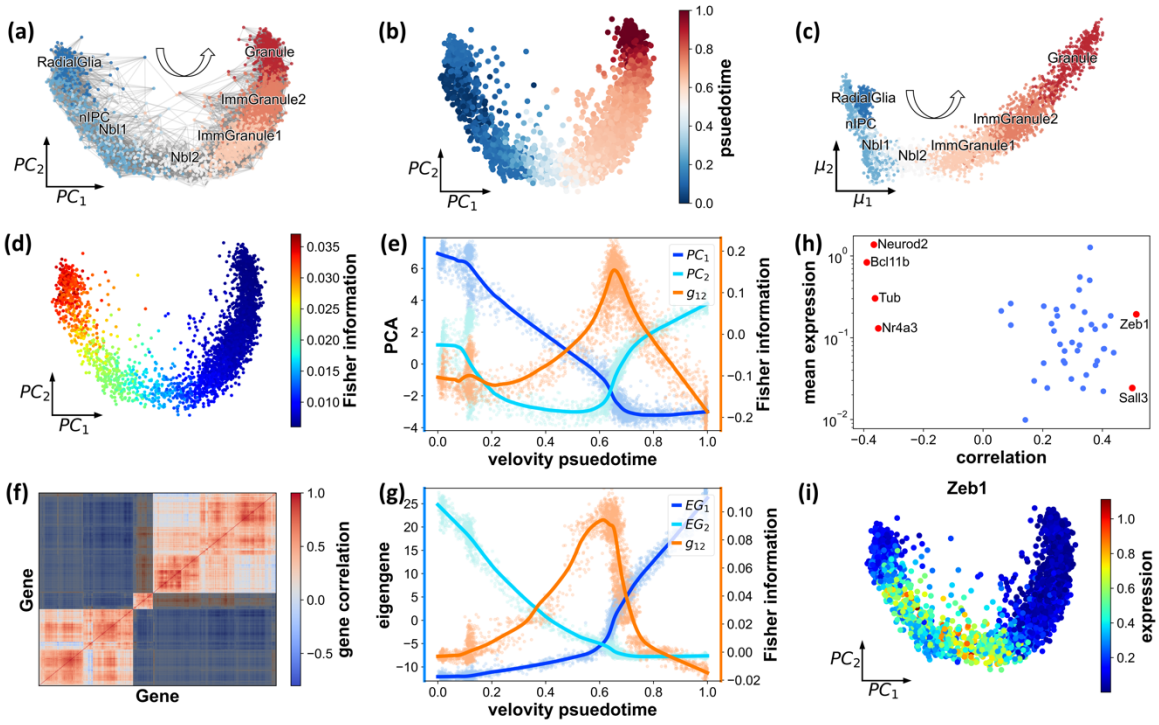

Figure 3 Fisher information analyses of Dentate gyrus neurogenesis.

(a) KNN graph of Dentate gyrus neurogenesis. Color represents cell type.

(b) Pseudo-time analysis of Dentate gyrus neurogenesis. Color represents pseudo-time.

(c) Gaussian embedding of Dentate gyrus neurogenesis onto the space of means.

(d) Variation of Gaussian embedding coordinate μ_1 and its corresponding Fisher information (First diagonal element of Fisher matrix) in dentate gyrus neurogenesis. Color represents value of Fisher information.

(e) One non-diagonal element in Fisher information matrix on PC coordinates (PC1-PC2) and the corresponding values of PCs (with respect to pseudo-time).

(f) Illustration of eigengenes derived from clustering on gene correlations. Color represents correlation and bright regions are the clustering for eigengenes.

(g) One non-diagonal element in Fisher information matrix on eigengene coordinates (EG1-EG2) and the corresponding values of eigengenes (with respect to pseudo-time).

(h) Correlation analysis of different transcriptional factors with information length. The genes that have high absolute correlation values with information length are marked with red.

(i) Variation of Zeb1 in dentate gyrus neurogenesis.

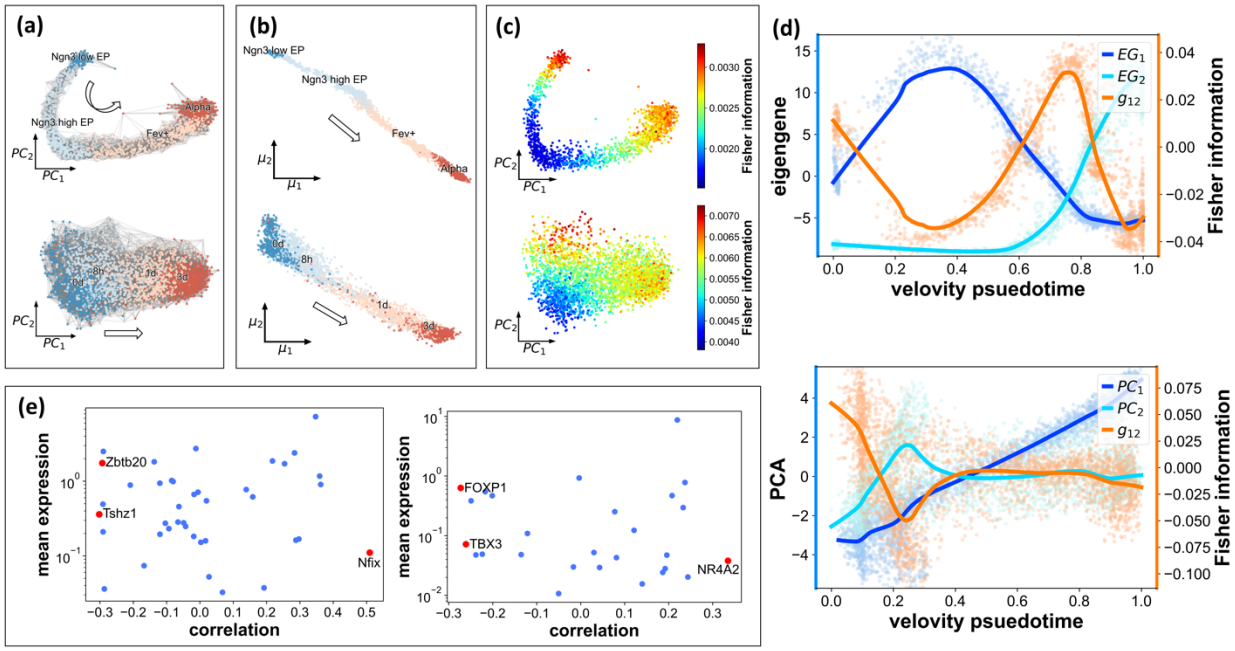

Figure 4 Fisher information analyses of endocrinogenesis and EMT of A549 cells induced by TGF-β.

(a) KNN graph of endocrinogenesis (top) and EMT (bottom). Color represents cell type.

(b) Gaussian embedding of endocrinogenesis (top) and EMT (bottom) onto the space of means of Gaussian embeddings.

(c) Variation of μ_1 and its Fisher information in endocrinogenesis (top) and EMT (bottom). Color represents value of Fisher information.

(d) One non-diagonal element in Fisher information matrix on eigengene coordinates and the corresponding values of eigengenes (with respect to pseudo-time) in endocrinogenesis (top). One non-diagonal element in Fisher information matrix on PC coordinates and the corresponding values of PCs (with respect to pseudo-time) in EMT (bottom).

(e) Correlation of transcriptional factors with information length in endocrinogenesis (left) and EMT (right). The genes that have high absolute correlation values with information length are marked with red.

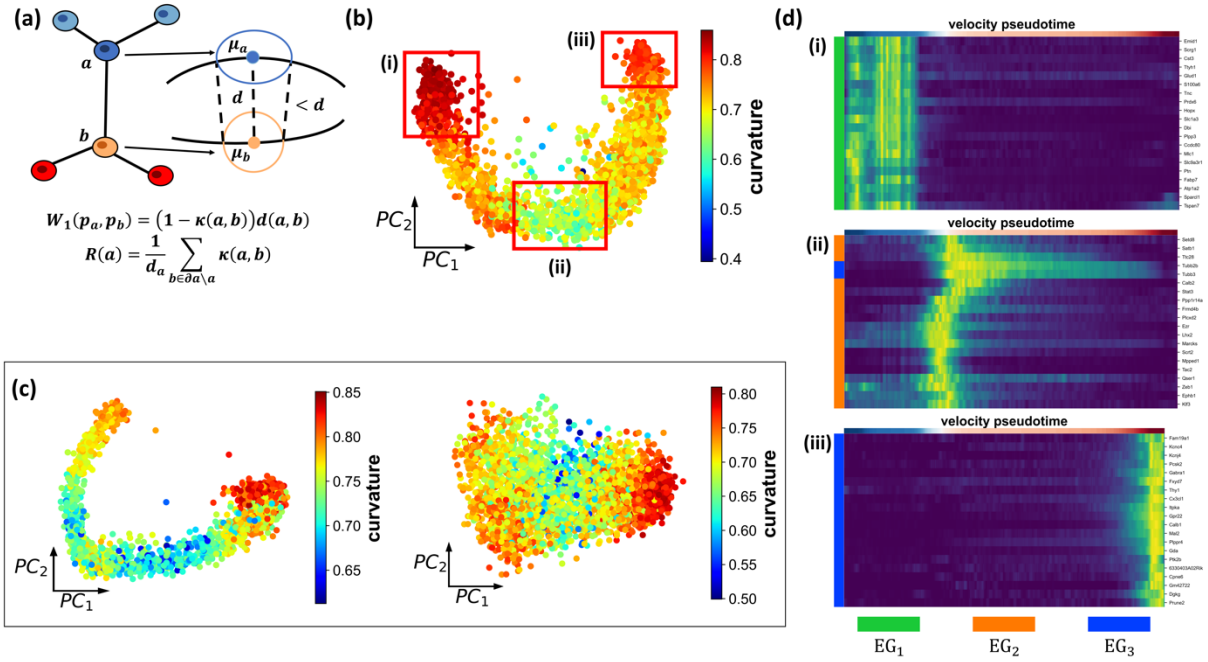

Figure 5 Identification of the critical point and attractors in different CPT processes, including dentate gyrus neurogenesis, endocrinogenesis, and EMT of A549 cells induced by TGF-β.

(a) Schematic plot of coarse Ricci curvature estimation in single cell RNA sequencing data.

(b) Identification of the critical point and attractors with coarse Ricci curvature in dentate gyrus neurogenesis. Red boxes are regions that cells have extreme values of CRC.

(c) Identification of the critical point and attractors with coarse Ricci curvature in endocrinogenesis (left) and EMT of A549 cells induced by TGF-β (right).

(d) Heatmap displaying genes with high correlation to CRC values, highlighted within the red boxes labeled i, ii, and iii. The color bars to the left of the heatmap indicate the eigengenes or communities associated with each gene in the corresponding row.

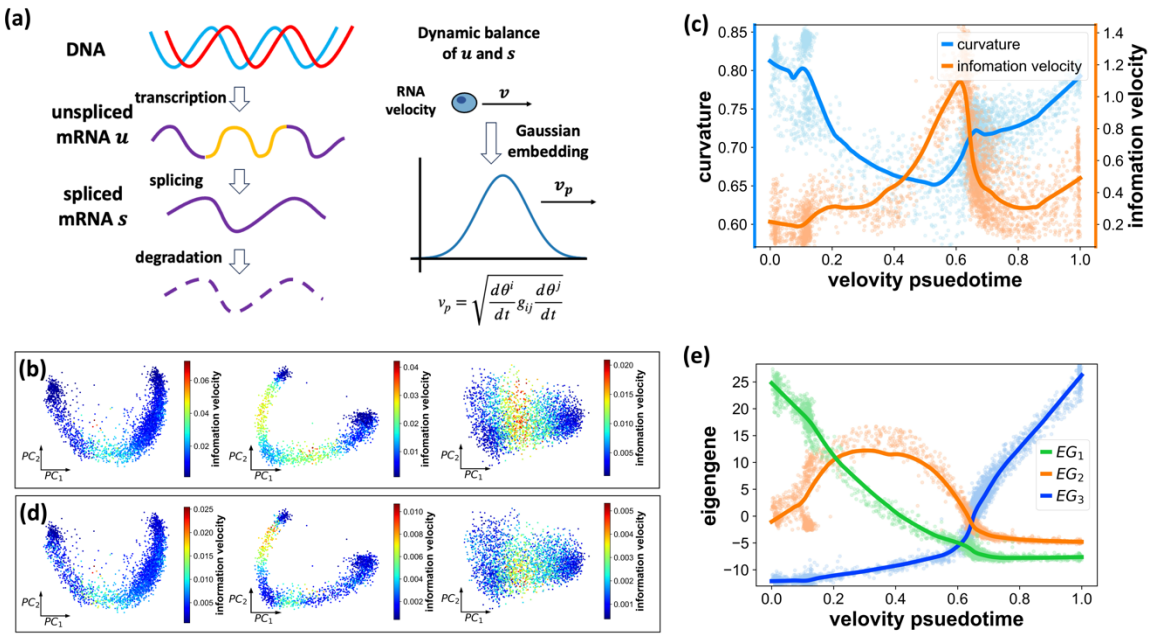

Figure 6 Information velocity in different latent spaces and its invariance under coordinate transformation.

(a) Schematic plot of calculation of information velocity with RNA velocity method.

(b) Calculation of the information velocity in the(μ, σ) space in dentate gyrus neurogenesis (left), endocrinogenesis (middle), and EMT of A549 cells induced by TGF-β (right).

(c) Correspondence of low-CRC value and high information velocity in dentate gyrus neurogenesis (with respect to pseudotime).

(d) Information velocity in NMF embedding space in dentate gyrus neurogenesis (left), endocrinogenesis (middle), and EMT of A549 cells induced by TGF-β (right).

(e) The relation between manifold geometry and gene regulation networks (switch of eigengenes)

Code availability

The scripts are available on GitHub at https://github.com/HuangMiao999/Information-Geometry-on-Cell-Manifolds.

The authors declare no conflict of interest.

Acknowledgement

The work is supported by National Natural Science Foundation of China Grants No.12247104 and National Innovation Institute of Defense Technology Grant No. 22TQ0904ZT01025. We also thank Dr. Li Li (Institute of Theoretical Physics, Chinese Academy of Sciences) for insightful discussion on Riemann geometry.

References

1. François P, Jutras-Dubé L. Landscape, bifurcations, geometry for development. Current Opinion in Systems Biology. 2018;11:129-36.
2. Corson F, Siggia ED. Geometry, epistasis, and developmental patterning. Proceedings of the National Academy of Sciences. 2012;109(15):5568-75.
3. Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, et al. CellRank for directed single-cell fate mapping. Nature Methods. 2022;19(2):159-70.
4. Qiu X, Zhang Y, Martin-Rufino JD, Weng C, Hosseinzadeh S, Yang D, et al. Mapping transcriptomic vector fields of single cells. Cell. 2022;185(4):690-711.e45.
5. Puy A, Beneventano P, Levin SA, Lo Piano S, Portaluri T, Saltelli A. Models with higher effective dimensions tend to produce more uncertain estimates. Science Advances. 2022;8(42):eabn9450.
6. Gallego JA, Perich MG, Miller LE, Solla SA. Neural Manifolds for the Control of Movement. Neuron. 2017;94(5):978-84.
7. Pillai AS, Jirsa VK. Symmetry Breaking in Space-Time Hierarchies Shapes Brain Dynamics and Behavior. Neuron. 2017;94(5):1010-26.
8. Elsayed GF, Lara AH, Kaufman MT, Churchland MM, Cunningham JP. Reorganization between preparatory and movement population responses in motor cortex. Nature Communications. 2016;7(1):13239.
9. Sadtler PT, Quick KM, Golub MD, Chase SM, Ryu SI, Tyler-Kabara EC, et al. Neural constraints on learning. Nature. 2014;512(7515):423-6.
10. Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nature Communications. 2017;8(1):15599.
11. Shi J, Teschendorff AE, Chen W, Chen L, Li T. Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. Briefings in Bioinformatics. 2020;21(1):248-61.
12. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560(7719):494-8.
13. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. Nature Biotechnology. 2020;38(12):1408-14.

14. Pinele J, Strapasson JE, Costa SIR. The Fisher–Rao Distance between Multivariate Normal Distributions: Special Cases, Bounds and Applications. Entropy. 2020;22(4):404.
15. Amari S-i. Information Geometry and Its Applications: Springer; 2016.
16. Amari S-i. Information Geometry and Its Applications: Springer Publishing Company, Incorporated; 2016.
17. Nielsen F. An Elementary Introduction to Information Geometry. Entropy. 2020;22(10):1100.
18. Kang X, Li C. A Dimension Reduction Approach for Energy Landscape: Identifying Intermediate States in Metabolism-EMT Network. Advanced Science. 2021;8(10):2003133.
19. Pearce P, Woodhouse FG, Forrow A, Kelly A, Kusumaatmaja H, Dunkel J. Learning dynamical information from static protein and sequencing data. Nature Communications. 2019;10(1):1-8.
20. Bojchevski A, Günnemann S. Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking. arXiv: Machine Learning. 2017.
21. Vilnis L, McCallum A. Word Representations via Gaussian Embedding. CoRR. 2014;abs/1412.6623.
22. Costa SIR, Santos SA, Strapasson JE. Fisher information distance: A geometrical reading. Discrete Applied Mathematics. 2015;197:59-69.
23. Ma J, Zhao J, Tian J, Bai X, Tu Z. Regularized vector field learning with sparse approximation for mismatch removal. Pattern Recognition. 2013;46(12):3519-32.
24. Sia J, Jonckheere E, Bogdan P. Ollivier-Ricci Curvature-Based Method to Community Detection in Complex Networks. Scientific Reports. 2019;9(1):9800.
25. Kim EJ, Lee U, Heseltine J, Hollerbach R. Geometric structure and geodesic in a solvable model of nonequilibrium process. Phys Rev E. 2016;93(6):062127.
26. Ollivier Y. Ricci curvature of Markov chains on metric spaces. Journal of Functional Analysis. 2009;256(3):810-64.
27. Kim E-j. Investigating Information Geometry in Classical and Quantum Systems through Information Length. Entropy. 2018;20(8):574.
28. Wu W, Hu G, Yu F. Ricci Curvature-Based Semi-Supervised Learning on an Attributed Network. Entropy [Internet]. 2021; 23(3).
29. Crooks GE. Measuring Thermodynamic Length. Phys Rev Lett. 2007;99(10):100602.

30. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology. 2007;1(1):54.
31. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1):559.
32. Du M, Ding S, Jia H. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. Knowledge-Based Systems. 2016;99:135-45.
33. Hochgerner H, Zeisel A, Lönnerberg P, Linnarsson S. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. Nature Neuroscience. 2018;21(2):290-9.
34. Bastidas-Ponce A, Tritschler S, Dony L, Scheibner K, Tarquis-Medina M, Salinno C, et al. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. Development. 2019;146(12).
35. Cook DP, Vanderhyden BC. Context specificity of the EMT transcriptional response. Nature Communications. 2020;11(1):2142.
36. Wang W, Poe D, Ni K, Xing J. Transiently increased intercommunity regulation characterizes concerted cell phenotypic transition. bioRxiv. 2021:2021.09.21.461257.
37. Ranjan B, Sun W, Park J, Mishra K, Schmidt F, Xie R, et al. DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. Nature Communications. 2021;12(1):5849.
38. Lecun Y, Chopra S, Hadsell R. A tutorial on energy-based learning. 2006.
39. Ni C-C, Lin Y-Y, Luo F, Gao J. Community Detection on Networks with Ricci Flow. Scientific Reports. 2019;9(1):9984.
40. Cox DR, Reid N. Parameter Orthogonality and Approximate Conditional Inference. Journal of the Royal Statistical Society: Series B (Methodological). 1987;49(1):1-18.
41. Nieto MA, Huang RY, Jackson RA, Thiery JP. EMT: 2016. Cell. 2016;166(1):21-45.
42. Pastushenko I, Blanpain C. EMT Transition States during Tumor Progression and Metastasis. Trends Cell Biol. 2019;29(3):212-26.
43. Prokopenko M, Lizier JT, Obst O, Wang XR. Relating Fisher information to order parameters. Physical Review E. 2011;84(4):041116.

44. Sritharan D, Wang S, Hormoz S. Computing the Riemannian curvature of image patch and single-cell RNA sequencing data manifolds using extrinsic differential geometry. Proc Natl Acad Sci U S A. 2021;118(29).