Efficient Selection of Optimal Time Points Over Biological Time-Series Data

1 Methods

1.1 Problem statement

Our goal is to identify a (small) subset of time points that can be used to accurately reconstruct the expression trajectory for *all* genes or other molecules being profiled. We assume that we can efficiently and cheaply obtain a dense sample for the expression of a very small subset of representative genes (here we use nanostring to profile less than 0.5% of all genes) and attempt to use this subset to determine optimal sampling points for the entire set of genes.

Formally, let G be the set of genes we have profiled in our dense sample, $T = \{t_1, t_2, \ldots, t_T\}$ be the set of all sampled time points. We assume that for each time point we have R repeats for all genes. We denote by e^r_{gt} be the expression value for gene $g \in G$ at time $t \in T$ in the r'th repeat for that time point. We define $D_g = \{e^r_{gt}\,, t \in T, r \in R$ as the complete data for gene g over all replicates and time points T.

To constrain the set of points we select we assume that we have a predefined budget k for the maximum number of time points we can sample in the complete experiment (i.e. for profiling all genes, miRNAs, epigenetic marks etc. using high throughput seq experiments). We are interested in selecting k time points from T which, when using only the data collected at these k points, minimizes the prediction error for the expression values of the unused points. To evaluate such a selection, we use the selected values to obtain a smoothing spline [5, 1, 18] function for each gene and compare the predicted values based on the spline to the measured value for the non-selected points to determine the error. In our problem, t_1 and t_T define the first and end points, so they are always selected. The rest of the points are selected to maximize the following objective 1:

Problem 1. *Given D_g for genes $g \in G$, the number of desired time points k, identify a subset of $k-2$ time points in $T \setminus \{t_1, t_T\}$ which minimizes the prediction error for the expression values of all genes in the remaining time points.*

1.2 Spline assignments

Before discussing the actual procedure we use to select the set of time points, we discuss the method we use to assign splines based on a selected subset of points for each gene. There are two issues that needs to be resolved when assigning such smoothing splines: 1. The number of knots (control points) and 2. their spacing. Past approaches for using splines to model time series gene expression data have usually used the same number of control points for all genes regardless of their trajectories [2, 16], and mostly employed uniform knot placements. However, since our method needs to be able to adapt to any size of k as defined above, we select them indirectly through regularization parameter of the fitted cubic smoothing spline where number of knots will be increased until the smoothing condition is satisfied [18]. Regularization parameter is estimated by leave-one-out cross-validation (LOOCV).

1.3 *TempSelect*: Iterative process to select points

Because of the highly combinatorial nature of the time points, selection problem we rely on a greedy iterative process to select the optimal points as summarized in Figure 1 (See Supplementary Text for pseudocode of the algorithm).

There are three key steps in this algorithm which we discuss in detail below.

- *Selecting the initial set of points:* When using an iterative algorithm to solve non-convex problems with several local minima, a key issue is the appropriate selection of the initial solution set [7, 10]. We have tested a number of methods for performing such initializations. The simplest method we tried is to uniformly select a subset of the points (so if $k = T/4$ we use each 4'th point). Another method we tested is to partition the set of all time points T into $k-1$ intervals of almost equal size. This method determines these boundaries by estimating the cumulative number of points until each time point and selecting time points with cumulative values $\frac{T}{k-1}, 2\frac{T}{k-1}, \ldots, (k-2)\frac{T}{k-1}$ respectively. Then, it uses k interval boundaries including t_1 and t_T as initial solution. Finally, we tested a method that relies on the changes between consecutive time points to select the most important ones for our initial set. Specifically, we sort all points except t_1 and t_T by average absolute difference with respect to its predecessor and successor time points by computing:

$$m_{t_i} = \frac{\sum_{g \in G} |Md(e_{gt_{i-1}}) - Md(e_{gt_i})| + |Md(e_{gt_{i+1}}) - Md(e_{gt_i})|}{2|G|} \quad (1)$$

  where $Md(e_{gt_i})$ is the median expression for gene g at time t_i. We then select the $k-2$ points with maximum m_{t_i} as the initial solution.

- *Iterative improvement step:* After selecting the initial set, we begin the iterative process of refining the subset of selected points. In this step we repeat the following analysis in each iteration. We exhaustively remove

all points from the existing solution (one at a time) and replace it with all points that were not in the selected set (again, one at a time). For each pair of such point, we compute the error resulting from the change (using the splines computed based on the current set of points evaluated on the left out time points), and determine if the new point reduces the error or not. Formally, let $T^- = T \setminus \{t_1, t_T\}$ and C_n be set of points for iteration n. We are interested in finding a point pair ($t_a \in C_n, t_b \in T^- \setminus C_n$) which minimizes the following error ratio for the next iteration $C_{n+} = C_n \setminus \{t_a\} \cup \{t_b\}$:

$$error\ ratio = \frac{error(C_{n+})}{error(C_n)} = \frac{\sum_{g \in G} \sum_{r \in R} \sum_{t \in T \setminus C_{n+}} (\hat{e}_{gt}^{C_{n+}} - e_{gt}^r)^2}{\sum_{g \in G} \sum_{r \in R} \sum_{t \in T \setminus C_n} (\hat{e}_{gt}^{C_n} - e_{gt}^r)^2} \quad (2)$$

where $\hat{e}_{gt}^{C_n}$ is our spline based estimate of the expression of gene g at time t by fitting smoothing spline over points C_n. If there are pairs which leads to an error ratio of less than 1 in the above function, we select the best (lowest error), assign it to C_{n+1} and continue the iterative process. Otherwise we terminate the process and output C_n as the optimal solution. Note that this greedy process is guaranteed to converge to a (local) minima since the number of time points is finite.

- *Fitting smoothing spline:* Third key step of our approach is fitting smoothing spline to every gene independently for selected subset of time points. Smoothing splines are capable of modeling arbitrary nonlinear shapes as well as they do not have the problems seen in other polynomial fitting methods such as Runge's phenomenon. Smoothing splines perform quite well in preventing overfitting [18]. Let $I_g = \{(t, Md(e_{gt})),\ t \in C\}$, and μ be the spline we are interested in fitting, smoothing spline can be found by the following optimization problem which minimizes penalized least-squares error:

$$\min \sum_{(t, y_t) \in I_g} (y_t - \mu(t))^2 + \lambda \int_{t_1}^{t_T} \mu''(x)^2 dx \quad (3)$$

where λ is the regularization parameter which prevents overfitting by affecting the number of knots selected. We estimated regularization parameter by leave-one-out cross-validation (LOOCV) in our experiments.

1.4 Individual vs. Cluster based Evaluation

In section 1.3, we assume that error of each gene has same contribution to the overall error. However, this assumption ignores the fact that expression profiles of genes are correlated with the expression of other genes. To take the correlation between gene profiles into account, we also performed cluster based evaluation of genes where we analyzed the error by weighting each gene in terms of inverse of the numbers of genes in the cluster it belongs. This scheme ensures that each

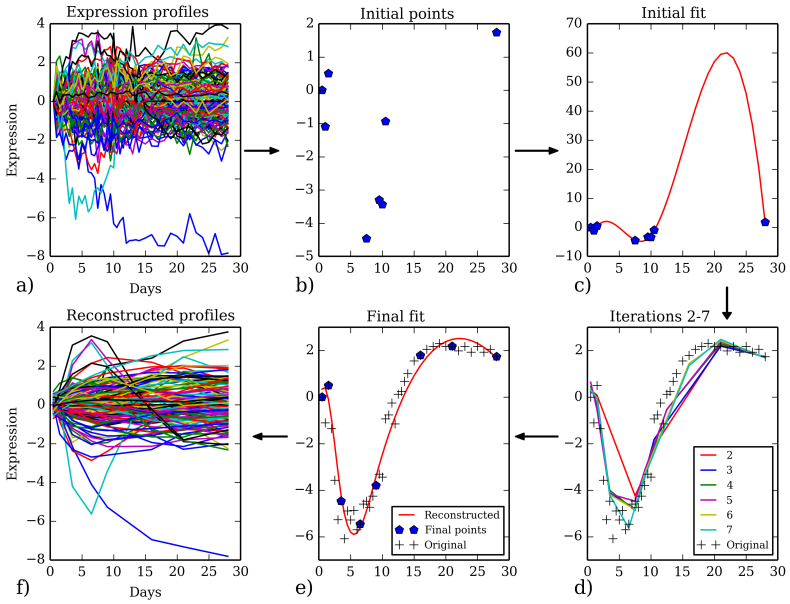


Figure 1: Summary of *TempSelect* execution for selecting 8 points: a) Expression profiles of all considered genes, b) Initially selected points by absolute difference heuristic, c) Reconstructed spline for a single gene over initial points, d) Reconstructed splines over iterations 2-7 for a single gene until convergence, e) Reconstructed and original expression profiles of a single gene, f) Reconstructed profiles of all genes by selected points.

cluster contributes equally to the resulting error rather than each gene. We find clusters by k-means algorithm over time series-data by treating each gene as a point in R^T space as well as over a vector of randomly sampled T time points on fitted spline [3]. We use Bayesian Information Criterion (BIC) to determine the optimal number of clusters [14].

1.5 More Complex Iterative Improvement Procedures

We also propose the following more complex iterative improvement procedures for *TempSelect*:

- We add and remove b time points in each iteration instead of a single point. This increases the complexity of each iteration from $O(kGT^2Q)$ to $O(kGT^{2b}Q)$ where Q is the complexity of fitting a smoothing spline.
- We run simulated annealing to escape from local minima [9]. In this case, we do not always move to a pair of points with the minimum error in each iteration, but instead move to a solution with random pair of points with

probability 1 if its error e^r is lower than error of current solution e^i whereas we move to a solution with probability $e^{-T(e^r-e^i)}$ if $e^r \geq e^i$. Here, T is the temperature that increases by increasing number of iterations and the probability of moving to a solution with larger error decreases over time.

Even though both approaches should escape from local minima theoretically better than the greedy approach we described above, they do not perform significantly better in practical instances.

2 Results

2.1 Datasets and Implementation

We developed a method *TempSelect* to select a subset of k time points from an initial larger set of n points such that the selected subset provides an accurate, yet compact, representation of the temporal trajectory. The method utilizes splines to represent temporal profiles and implements a cross validation strategy to evaluate potential sets of points. Following initialization which is based on the expression values, we employ a greedy search procedure that adds and removes points until a local minima is reached. The resulting set is then used for the larger genomic and epigenetic experiments. To test this method and to demonstrate its ability to reduce time, costs and samples while still providing accurate description of the temporal profiles, we focused on experiments related to lung development in mice. We implemented *TempSelect* in Python. Its implementation, code, detailed results, and datasets are available on `https://github.com/emresefer/geneexpress`.

We have used mRNA, miRNA and methylation data from mouse lung development to test our method. We first profiled the expression of 126 selected genes that are determined to be relevant to lung development using a NanoString array (Methods) and have used these experiments to select a subset of time points for the more global expression and Seq profiling. To test the method, we have also profiled the the expression of a much larger set of randomly selected miRNAs (599). Both datasets contain between 2 and 4 repeats for each time point allowing us to quantify sampling noise as well. We further obtained methylation data for a subset of the time points selected based on the mRNA analysis (Methods).

2.2 *TempSelect* identifies subset of important time points across multiple genes

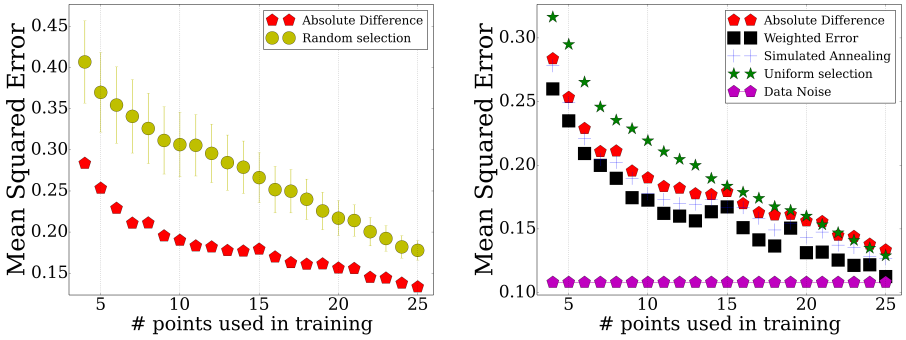
While our method can be used to select any number of time points, to demonstrate its utility we have tested it in the following setting. First we fixed a set of points in advance (first (0.5'th day) and last (28'th day), which are required for any setting and day 7 which was previously determined to be of importance to lung development, see Supplementary Results for other settings). In addition, we have asked *TempSelect* to further select 10 more points (for a total of

13). For this setting, the method selected the following points: 0.5, 1.0, 1.5, 2.5, 4, 5, 7, 10, 13.5, 15, 19, 23, 28 out of 40 points. While we do not know the ground truth, the larger focus on the earlier time points determined by the method (with 7 of the 13 points for the first 7 days) makes sense in this context as several aspects of lung differentiation are determined in this early phase [6]. The other 3 weeks were more or less uniformly sampled by our method. This highlights the usefulness of an unbiased approach to sampling time points rather than just uniformly sampling through the time window.

We have also tested the performance of *TempSelect* by using it to select subsets of size 3 to 25 time points and testing how well these can be used to determine the values of nonsampled points. To determine the accuracy of the reconstructed profiles using the selected points, we computed the average mean squared error for points that were not used by the method (Methods). We normalized mRNA dataset by quantile normalization followed by $\log 2$ transformation. The results are presented in Figure 2. The figure includes a comparison of our method with two baseline methods: a random selection of the same number of points and uniform sampling of points within the range being studied, a method that is commonly used for time series expression profiling which ensures that the number of unsampled points between two consecutive time points is approximately same. We have also compared the performance of the different strategies for initializing the set of points as discussed in Method (sorting by absolute differences or by equal partition) and between different methods for searching for the optimal subset (simulated annealing, weighting genes by cluster size, and adding/removing multiple time points per iteration, see Methods). Finally, the figure also presents the repeat noise values which is the theoretical limit for the performance of any profile reconstruction method.

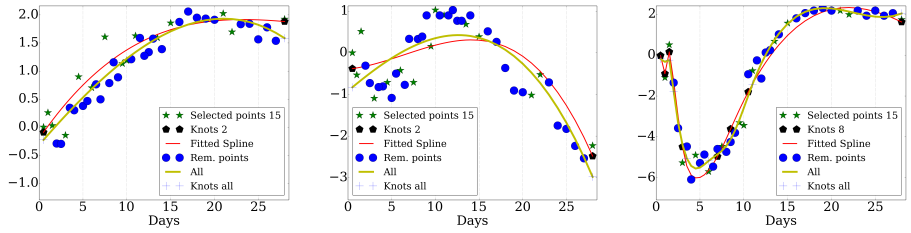
As expected, we find significant performance improvement over randomly selected points in terms of mean squared error. Importantly, we also see a significant and consistent improvement (for all numbers of selected time points) over uniform sampling highlighting the advantage of study specific sampling design. Sorting initial points by absolute values further improves the performance highlighting the importance of initialization when searching large combinatorial spaces. Simulated annealing, weighting, and multiple point selection increases the performance only in a limited way (Multiple point modification results are not presented due to space limitations). As the number of points used by the method increases, it leads to results that are very close to the error represented by noise in the data (0.108) (Supplementary Figure 1 for noise in each time point). Using additional earlier time points (prior to birth) does not change the relative performance of the methods (Supplementary Figure 2).

Figure 3 presents the reconstructed and measured expression values for when using *TempSelect* to select 13 time points (less than a third of the points that were profiled). Note that even though each of these genes had a different trajectory and different inflection points, the selected set of points enable *TempSelect* to fit all of these quite accurately without overfitting (See Supplementary Figs. 3–4 for figures of several other genes and for figures reconstructed by using the best 8 time points as determined by *TempSelect*, respectively).



(a) (b)

Figure 2: Performance of *TempSelect* by increasing number of selected points, a) *TempSelect* with absolute difference heuristic vs Random selection, b) Comparison of *TempSelect* variations with the noise in the data.



(a) PDGFRA (b) ELN (c) INMT

Figure 3: Reconstructed expression profiles over genes a) PDGFRA, b) ELN, c) INMT

2.3 Identified time points using mRNA data are appropriate for miRNA profiling

To test the usefulness of our method for predicting the correct sampling rates for other genomic datasets, we next profiled mouse miRNAs for the same developmental process. miRNAs have been known to regulate lung development [15] and several miRNAs are differentially expressed during this developmental process [19]. Several of these are also coordinately activated with various TFs to control specific transitions during development [13]. Thus, any large scale effort to model this process would require the profiling of miRNAs as well. Unlike the mRNA dataset, which utilized prior knowledge to profile less than 1% of all genes, the miRNA dataset profiled almost 600 miRNAs, more than 50% of known mouse miRNAs. Thus, such data represents an unbiased sample and can provide information on whether using one type of genomic data can be helpful

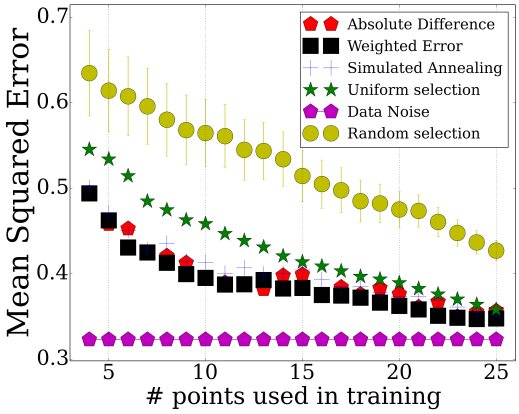


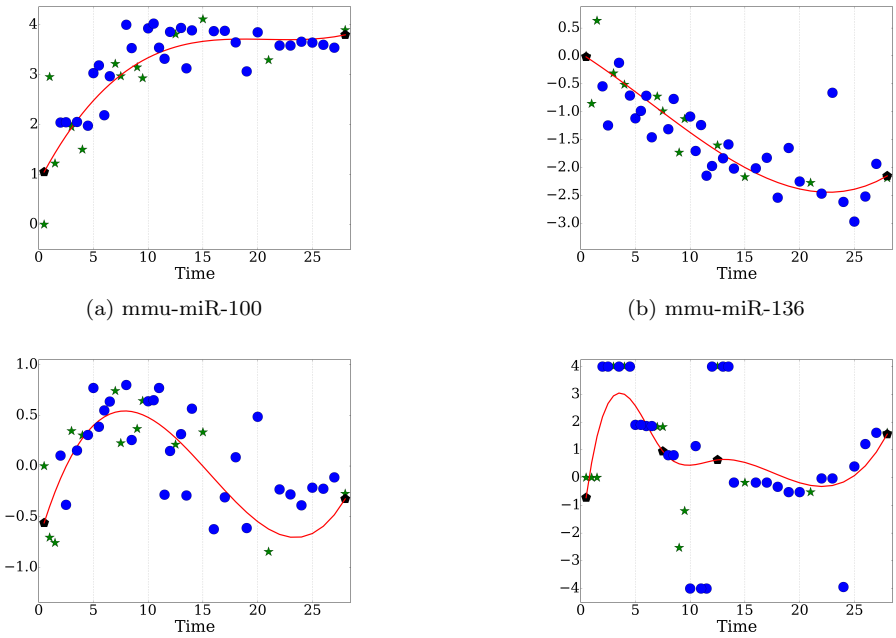
Figure 4: Performance of *TempSelect* by increasing number of selected points over miRNA dataset

for determining rates for other types. In our analysis, we normalized miRNA values by variance mean normalization [4]. We also found miRNA clusters to be enriched for a number of biological processes as well as being noisier than mRNA dataset (See Supplementary Results).

To test *TempSelect* on this dataset, we used the *mRNA* expression data to select the time points and then used the miRNA expression values for the selected time points to reconstruct the complete trajectories for each miRNA. The results are presented in Figure 4. In addition to the comparison included in the mRNA figure, the miRNA figure includes the optimal results for using miRNA data (as opposed to mRNA data) to select the points. As can be seen, the points selected by the mRNA analysis leads to reconstruction that is much better than when using random points ($p < 0.01$ based on randomization analysis) highlighting the relationship between the two datasets and the ability to use one to determine points for the other. Further, performance using the mRNA set is very similar to the performance using the miRNA data itself. For example, when using the 13 selected mRNA points, the average mean squared error is 0.4312 whereas when using the optimal points based on the miRNA data itself the error is 0.4042. More generally, even though the noise in the miRNA data is higher than for the mRNA dataset, relative ordering of the performance of each of the methods is similar to the mRNA results in Figure 2. This serves as a strong indication that mRNAs can serve as a general proxy for selecting time points for other genomic datasets.

Figure 5 presents the reconstructed and measured expression values for a few miRNAs using time points identified using the mRNA dataset. Accurate prediction of different miRNA profiles show the importance of identified points for the mRNA dataset. Several of these miRNAs are known to be involved in regulation of lung development. For example, mmu-miR-100 is known to regulate Fgfr3 and Igf1r, mmu-miR-136 targets Tgfb2, mmu-miR-152 targets

Meox2, Robo1, Fbn1, Nfya [11].



(a) mmu-miR-100

(b) mmu-miR-136

(c) mmu-miR-152

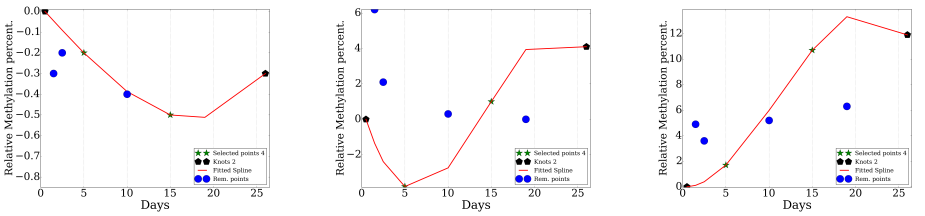
(d) mmu-miR-219

Figure 5: Predicted expression profiles of miRNAs a) mmu-miR-100, b) mmu-miR-136, c) mmu-miR-152, d) mmu-miR-219.

2.4 Selecting time points for Methylation analysis

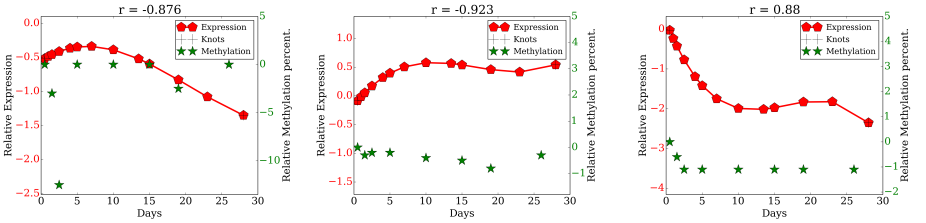
Methylation data has 3 repeats for time points 0.5, 1.5, 2.5, 5, 10, 15, 19, 26 for 266 loci belonging to 13 genes. Among these genes all of them except Zfp536 also exist in mRNA dataset. Supplementary Table 1 summarizes the number of loci for each gene in methylation dataset. We used shifted percentage of methylation at each time point in our analysis which is obtained by subtracting the median percentage of methylation at initial time point (baseline) from all data points for each gene.

In addition to mRNA and miRNA expression data, epigenetic data has been increasingly studied in time series experiments [8, 17, 12]. To test the ability of the mRNA data to determine appropriate points for methylation analysis we profiled the up stream regions of 13 genes at 8 of the 42 time points used for the mRNA and miRNA studies (Methods). We next applied *TempSelect* to the mRNA data of these 8 points to select 4 of them and compared the selected points to those that would have been selected using the methylation data itself. The 4 points identified using the mRNA data (0.5, 5, 15, 26) were



(a) Chromo. 2, 157423995 (SRC) (b) Chromo. 5, 134721315 (ELN) (c) Chromo.12, 112657170 (AKT1)

Figure 6: Reconstructed methylation profiles over several loci (chromosome, position) with corresponding genes.



(a) DNMT3A (b) SRC (c) LOX

Figure 7: Comparison of gene expression and methylation data for genes a) DNMT3A, b) SRC, c) LOX.

exactly the same as the ones selected using the methylation data indicating again that mRNA data is a good proxy for selecting sampling for epigenetic data as well. Figure 6 shows reconstructed splines over the identified points for several genomic methylation loci. Figure 7 presents the methylation and expression curves for 3 genes: DNMT3A, SRC, and LOX genes. As can be seen, in several cases we observed strong negative or postive correlations between the two datasets in the time points we used serving as another indication for the ability to use one dataset to select the sampling points for the other. See Supplementary Table 2 for correlation of all genes and Supplementary Fig. 6 for distribution of correlation for loci of each gene.

3 Conclusion

We develop a method *TempSelect* to efficiently identify subset of important time points over densely sampled gene expression profiles. We show that these points can be used as candidates for high-throughput profiling experiments as well as other temporal experiments such as methylation. Additionally, identified points can serve as a proven benchmark to reduce the experimental cost.

References

[1] Ziv Bar-Joseph, Georg K Gerber, David K Gifford, Tommi S Jaakkola, and Itamar Simon. Continuous representations of time-series gene expression data. *Journal of Computational Biology*, 10(3-4):341–356, 2003.

[2] Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13(8):552–564, 2012.

[3] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[4] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[5] Carl De Boor. A practical guide to splines. *Mathematics of Computation*, 1978.

[6] Martin Guilliams, Ismé De Kleer, Sandrine Henri, Sijranke Post, Leen Vanhoutte, Sofie De Prijck, Kim Deswarte, Bernard Malissen, Hamida Hammad, and Bart N Lambrecht. Alveolar macrophages develop from fetal monocytes that differentiate into long-lived cells in the first week of life via gm-csf. *The Journal of experimental medicine*, 210(10):1977–1992, 2013.

[7] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[8] Tal Kafri, Mira Ariel, Michael Brandeis, Ruth Shemer, Lance Urven, John McCarrey, Howard Cedar, and Aharon Razin. Developmental pattern of gene-specific dna methylation in the mouse embryo and germ line. *Genes & development*, 6(5):705–714, 1992.

[9] Scott Kirkpatrick, C Daniel Gelatt, Mario P Vecchi, et al. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[10] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[11] Antonia P Popova, J Kelley Bentley, Tracy X Cui, Michelle N Richardson, Marisa J Linn, Jing Lei, Qiang Chen, Adam M Goldsmith, Gloria S Pryhuber, and Marc B Hershenson. Reduced platelet-derived growth factor receptor expression is a primary feature of human bronchopulmonary dysplasia. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 307(3):L231–L239, 2014.

[12] Eberhard Schneider, Galyna Pliushch, Nady El Hajj, Danuta Galetzka, Alexander Puhl, Martin Schorsch, Katrin Frauenknecht, Thomas Riepert,

Achim Tresch, Annette M Müller, et al. Spatial, temporal and interindividual epigenetic variation of functionally important dna methylation patterns. *Nucleic acids research*, 38(12):3880–3890, 2010.

[13] Marcel H Schulz, Kusum V Pandit, Christian L Lino Cardenas, Namasivayam Ambalavanan, Naftali Kaminski, and Ziv Bar-Joseph. Reconstructing dynamic microrna-regulated interaction networks. *Proceedings of the National Academy of Sciences*, 110(39):15686–15691, 2013.

[14] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[15] Roberto Sessa and Akiko Hata. Role of micrornas in lung development and pulmonary diseases. *Pulmonary circulation*, 3(2):315, 2013.

[16] Rohit Singh, Nathan Palmer, David Gifford, Bonnie Berger, and Ziv Bar-Joseph. Active learning for sampling in time-series experiments with application to gene expression analysis. In *Proceedings of the 22nd international conference on Machine learning*, pages 832–839. ACM, 2005.

[17] Rudolf P Talens, Dorret I Boomsma, Elmar W Tobi, Dennis Kremer, J Wouter Jukema, Gonneke Willemsen, Hein Putter, P Eline Slagboom, and Bastiaan T Heijmans. Variation, patterns, and temporal stability of dna methylation: considerations for epigenetic epidemiology. *The FASEB Journal*, 24(9):3135–3144, 2010.

[18] Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

[19] Andrew E Williams, Sterghios A Moschos, Mark M Perry, Peter J Barnes, and Mark A Lindsay. Maternally imprinted micrornas are differentially expressed during mouse and human lung development. *Developmental Dynamics*, 236(2):572–580, 2007.