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

1 Methods

In this paper, we propose an efficient framework to identify subset of time points jointly over gene expression time-series data for multiple genes. Our proposed solution is different than the existing sampling based approaches since the existing methods focus mostly on selecting subset of important time points only over single gene. In addition, several other methods are based on active learning which is not always valid for biological gene expression data due to the nature of the experiments. For instance, multiple genes can be sampled simultaneously once the time point is fixed which may not be modeled perfectly by active learning based approaches. In general, our proposed solution is important since:

- There is still a space for improvement to identify subset of important time points since previous solutions are all based on heuristics. For instance, [1] samples gene expression values at uniformly distributed time points which is not guaranteed to be the optimal.

- Efficient methods can greatly decrease experiment cost without trading off accuracy of the expression values.

More formally, let G be the set of genes which expression we are interested in measuring/predicting, and $T = \{t_1, t_2, \ldots, t_T\}$ be the set of all sampled time points. We assume that experiments is repeated D times; expression of each gene over each time points is mesured D times. In these experiments, let e^d_{gt} be the expression value for gene $g \in G$, $t \in T$ across d'th repeat of the experiment. We define $D_g = \{e^d_{gt}\,, t \in T, d \in 1, \ldots, D\}$ as the data for gene g over all replicates and time points T.

We assume that we have a predefined budget k which is the maximum number of time points we can sample (we simply assume that sampling each time point has same cost). We are interested in selecting k number of time points which minimizes the prediction error of the rest of the unselected time points where we predict expression values of unsampled time points by smoothing splines. In our problem, t_1 and t_T define the first and end points, so they will always be a part of solution. It can be defined formally as in Problem 1:

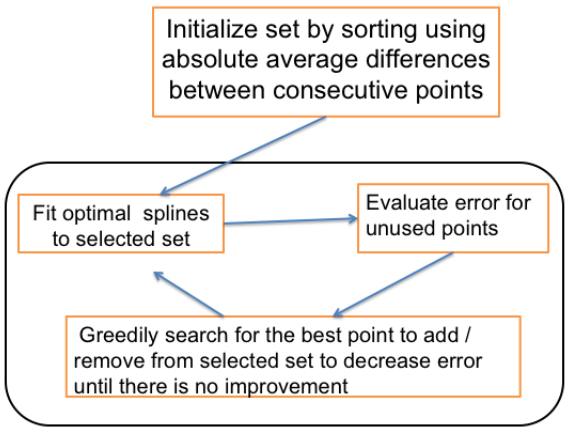

Figure 1: Summary of our Method

Problem 1 *Given D_g for genes $g \in G$, and number of points k to select, we are interested in identifying the subset of $k-2$ time points among $T \setminus \{t_1, t_T\}$ which minimizes the prediction error of the expression values of remaining time points by smoothing splines.*

1.1 The Algorithm

Here, we propose the following iterative solution to tackle Problem 1. Initially, we select $k-2$ number of time points (t_1 and t_T are already in the solution) by one of the heuristics that are defined in detail in Section 1.2. Then, we keep adding one point among the remaining points and remove one point until there is no improvement in the error of the remaining time points. In each iteration, we select the point pair with the minimum error among all $(k-2)(T-k)$ point pairs as our solution, and keep iterating until there is no improvement. Figure 1 summarizes the iterative approach.

1.2 Initial Point Selection

Initial selection of k points has a significant effect on the results. We found uniform partition and absolute value difference heuristics to perform the best.

Uniform partition heuristic partitions the set of all time points T into $k-1$ intervals of almost equal size by dynamic programming. Then, it uses k interval boundaries including t_1 and t_T as initial solution. On the other hand, we absolute value difference heuristic to perform better than uniform partition

heuristic. In this case, we sort all points except t_1 and t_T by average absolute difference with respect to the neighbouring time points as in:

$$m_{t_i} = \frac{\sum_{g \in G} \sum_{d \in D} |e^d_{gt_{i-1}} - e^d_{gt_{i+1}}|}{2 \sum_{g \in G} |D_g|} \tag{1}$$

where then it selects the top $k-2$ points with maximum m_{t_i} as initial solution.

1.3 Iterative Step

Iterative step exhaustively takes out each single point from the existing solution and adds a point from the remaining points into the solution. Let C be the current solution of time points, and C^* be the best solution identified by iterative step. We are interested in finding a point pair (t_a, t_d) among $(k-2)(T-k)$ point pairs which minimizes the following error for new solution $C^* = C \setminus \{t_d\} \cup \{t_a\}$:

$$error_{C^*} = \frac{\sum_{g \in G} \sum_{d \in D} \sum_{t \in T \setminus C^*} |\hat{e}_{gt} - e^d_{gt}|}{\sum_{g \in G} \sum_{d \in D} \sum_{t \in T \setminus C^*} 1} \tag{2}$$

where $\hat{e}_{gt}$ is our estimate of the expression of gene g at time t by fitting a smoothing spline. This iterative step continues until there is no improvement in the error of the remaining time points.

2 Results

References

[1] 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.