論文の内容の要旨

論文題目　EM algorithms for optimization of Wright Fisher model and reconstruction of spatiotemporal gene expression patterns
（EMアルゴリズムによるWright-Fisherモデルの最適化と遺伝子発現の時空間再構築）

氏　　名　　小嶋　泰弘

Probabilistic models play important roles in extracting knowledge from biological data. In particular, a likelihood of the probabilistic models given the data are maximized over the parameters of the models, and the optimized parameters are regarded as some quantity reflecting the biological processes behind the data. For example, Sandberg *et al.* estimated the kinetics parameters of transcriptional regulation from single cell RNA-seq (scRNA-seq) data utilizing this methodology. However, the optimization of the probabilistic models often faces an obstacle that some variables constituting the models are unobserved. EM algorithms overcame this obstacle in many situations of the biology by marginalizing the log likelihood over the unobserved variables. In this thesis, we describe two studies where EM algorithms are utilized for efficient optimization of a mathematical model of population genetics, Wright-Fisher model (WF), and the reconstruction of time-continuous spatiotemporal gene expression patterns during embryogenesis.

Estimation of population genetic parameters using an EM algorithm and sequence data from experimental evolution populations

Introduction

Lower cost of genome sequencing has enabled us to derive time course sequence data, which provides us the genome wide dynamics of biological processes. In particular, a new experimental method, Evolve and Resequence, provides us time course allele frequency data of numerous single nucleotide polymorphisms (SNPs) during an experimental evolution induced by artificial selective pressure on specific phenotype. Many of researches utilizing this method mainly focused on detecting SNPs which are related to the phenotype selected in the experimental evolution.

On the other hand, time course of allele frequencies has been extensively investigated by mathematical models in the field of population genetics. In particular, the dynamics of allele frequencies is characterized by the strength and dominance of selection in WF model. These parameters optimized for E&R data are expected to help us derive new insight for the evolutionary dynamics behind the experimental evolution, since they provide us quantitative information of the evolutionary selection across numerous SNPs. However, it is difficult to estimate these parameters from E&R data due to the huge number of SNPs.

Recently, several methods are developed for the optimization of WF models to E&R data by improving the efficiency of the estimation, while these methods can not evaluate the confidence of the estimation. Since the simultaneous estimation of s and h is often intrinsically unidentifiable, the estimates which is far away from true values can affect the post analysis of the estimated values. Here, we proposed a new method to efficiently estimate WF parameters from E&R data, utilizing an EM algorithm for continuous time Markov chain Furthermore, our method evaluates estimation confidence by calculating confidence intervals based on empirical Fisher information matrix.

Results and discussion

First, we validated the estimation of our method and compare it with other existing methods using simulated ER data. We found that the values estimated by our method distributed around true values used in simulation. On the other hand, we confirmed that the filtering process based on estimated confidence intervals selectively excludes the estimated values which is far away from true values. In our comparison with other methods, dominance estimation of our method was the most accurate after the filtering process, while the estimation time of our method was about one third of that of the most efficient other method.

Next, we applied our method to a real E&R data where *Drosophila* population was bred in a new thermal condition, which is expected to induce selective pressure on phenotype related to thermal adaptation. As a result, we found that there was a specific peak in the distribution of the selection strength parameters of SNPs within a cosmopolitan inversion, *In(3R)P*, region, which suggested that many SNPs within *In(3R)P* region share the common allele frequency dynamics during this experimental evolution. On the other hand, the estimated dominance parameters are distributed around 1, which indicated that many of deleterious alleles in this experiment are recessive. This observation is consistent with the observation that many of deleterious alleles are recessive in natural population of *Arabidopsis*.

Reconstructing spatiotemporal gene expression patterns during embryogenesis

Introduction

Embryogenesis produces a specific shape of each species from a single fertilized egg by repeated cell division and cell movement. To achieve this complicated process, the behavior of cells in each spatiotemporal context must be accurately specified. One large element determining the cell behavior is the expression of numerous genes that reaches tens of thousands. Hence, the localization of gene expression has been an important clue for revealing the mechanism of each developmental events.

Recently, several methods have enabled us to derive spatial expression patterns of numerous genes at one time. These methods enable us to explore spatial gene expression domains and gene modules sharing spatial expression patterns. On the other hand, these are conducted with one time point or sparse time course presumably due to the high cost of the experiments. Hence, the developmental events captured by this method is limited. Furthermore, the dynamics of these spatial gene expression patterns have not been explored at transcriptome wide manner, since the spatial patterns at each stage are analyzed separately presumably due to extensive cell migration during embryogenesis. Such molecular dynamics is now well captured by single cell RNA-seq (scRNA-seq). In particular, the application of this technology to 12 time points of zebrafish embryogenesis revealed molecular differentiation to divergent cell types. On the other hand, each cell of scRNA-seq has lost their spatial information, which obscure the spatiotemporal dynamics of these differentiation process.

Here, we integrated transcriptome data containing spatial information and single cell RNA-seq data with cell movement data, which is derived by advanced imaging technology, and estimate time continuous spatiotemporal gene expression patterns for numerous genes to explore the spatiotemporal dynamics of cell type differentiation. In our method, we reconstructed original spatial position of scRNA-seq cells and spatial gene expression patterns at the observation time of the transcriptome data utilizing MAP-EM algorithm. From the estimated spatial gene expression patterns at the observation time, we predicted the spatial gene expression patterns at any time points involved in the cell movement data.

Results and discussion

In simulation experiments for validation of our method, we evaluated the estimation performance of scRNA-seq cell position and spatiotemporal gene expression patterns. We found that 75 % of scRNA-seq cells are located at the position closer to true position than 20 % of the embryo diameter, while the Pearson's correlation coefficient between estimated and true spatiotemporal gene expression patterns exceeds 0.8 for 84 % of all genes.

In the application to real data, we confirmed that the predicted spatial gene expression patterns are consistent with spatial patterns observed from real embryo for many genes. The success in the reconstruction was confirmed at two embryonic stages regardless of the existence of spatial gene expression profiles at the corresponding embryonic stages. Furthermore, we explored the spatiotemporal dynamics of cell type differentiation, and found that expression profiles at midbrain and hindbrain boundary (MHB) is closer to that at later embryonic stage than other regions of midbrain and hindbrain. This is presumably because the differentiation process is most progressed at the region close to MHB due to signaling from MHB, which induces the differentiation process into midbrain and hindbrain in real embryo.