論文内容の要旨

論文題目 Marked Point Process to Analyze Biological Data
（マーク付き点過程によるバイオデータ解析）

氏名　初田 浩志

1 Introduction

1.1 Biological Data Analysis

The primary goal of bioinformatics is developing computational techniques to analyze and understand biological data, especially genomic data. However, it is difficult to define the goal of analyzing and understanding biological data.

This dilemma is common to all fields of natural science. In natural science, we usually do not have "ground truth" data to evaluate the accuracy of methods. This is because we analyze scientific data when we do not know the "ground truth".

Obviously, because we want methods that lead to new biological discoveries, the goal of computational techniques to analyze biological data is to make a biological discovery. This is correct; we need to have deep expertise in biology to decide the goals of computational methods for biological data analysis. However, it is preferable that we have another strategy to determine the goals, because it is efficient to evaluate computational techniques before we conduct biological experiments based on the results of the computational analysis. In this respect, I propose that we should focus on human perception, our way of seeing.

1.2 Human Perception

To understand our way of seeing, it is useful to consider it in terms of psychology of vision. Figure 1(a) shows the well-known image of "Dalmatian dog" proposed by R. Gregory. If we are unfamiliar with this picture, it is likely that it will not make sense. Once the dog silhouette has been seen as shown in Figure 1(b), it becomes clear that there is a dog in the original image. This example illustrates the amazing ability of the human visual system to find meaningful objects in images by using prior knowledge of what we see in them. The important point here is that this is not a particular case; this is our basic visual capability. We can recognize local features (e.g. fragments of silhouette of a dog) only by knowing that they consist in an overall structure (e.g. a dog). Without the prior knowledge of the overall structure, it is difficult to distinguish between a local meaningful object and a noise.

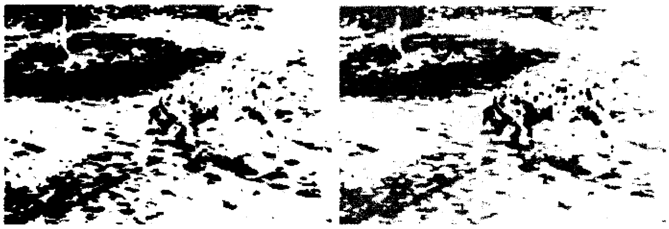


Figure 1 Dalmatian dog proposed by R. Gregory (a) obscure image (b) dog silhouette

1.3 Goal of This Study

Therefore, I define my goal of this study as achieving a computational technique that mimics our way of seeing biological data. More specifically, the purpose of this study is developing a method to analyze biological data by using prior knowledge of what it analyzes. For this reason, I developed a detection method that uses prior knowledge of the structure of the data by using a marked point process.

2 Expression Patterns of the Distributions of Transcription Start Sites Using Marked Point Process

In this study, I propose a novel method to detect expression patterns of the distributions of transcription start sites (TSSs). Figure 2.1 (a) shows a toy example of TSSs distribution. When we see this type of data, we usually put arbitrary "hills" into the data, as shown in Figure 2.1 (b), to interpret them. In other words, we use our prior knowledge of the structure of the data to understand them; we use our prior knowledge that the TSSs distributions can be approximated by a mixture of some simple functions, such as Gauss functions, to see the data. From this viewpoint, I present a method that mimics our way of seeing the data by using prior knowledge of the data structure and applying Gauss functions to the data.

I model the distributions of TSSs by using a marked point process and present a method to detect expression patterns of TSSs. In my methodology, each point corresponds to each nucleotide, and points have attributes to represent the prior knowledge of data structure. The attribute of the point is called mark in the field of point processes; thus, it is a marked point process that I use in this study. Because I employ Gauss functions as the marks, my method puts Gauss functions into TSSs data to detect their patterns, as shown in Figure 2.1 (c). To search good configurations of Gauss functions, I define a Gibbs energy consisting of two terms: data energy and prior energy. The data energy measures consistence of Gauss functions to the TSSs data; the prior energy incorporates some constraints on interactions of Gauss functions. The Gibbs energy can be minimized by using a multiple birth-and-death dynamics coupled with the traditional simulated annealing to find the optimum configuration of the marked points.

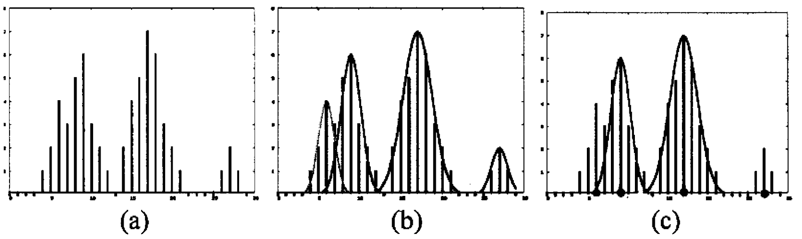
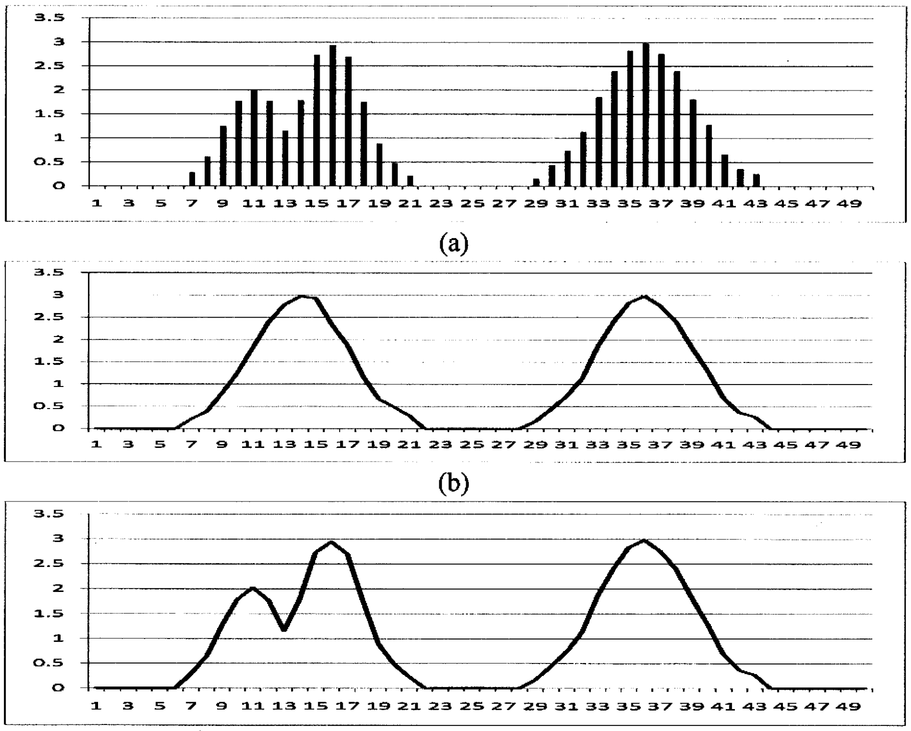


Figure 2 Illustrative diagrams of my method. (a) Toy example of transcription start sites data. (b) An example of our way of seeing: We see four "hills", such as Gauss functions. (c) My method: In my method, each point has Gauss functions; Red point has red function. Yellow function is removed because it overlaps the red function, depending on a parameter. Blue function is removed because it is too small, depending on another parameter.

The proposed method was tested on synthetic and real-world data. The synthetic data is composed of three Gauss functions, and random noise is added to the mixture. In Figure 3, histogram is the synthetic data, and plotted curves are the results of pattern detections of the synthetic data by my method. In Figure 3(a), the leftmost and central functions overlap moderately. It is difficult to definitely decide the number of Gauss functions; there seems to be two functions, and there seems to be three functions at the same time. The proposed method detects two Gauss functions when the data and prior energy terms are equally contributed to the global energy function, as shown in Figure 3(b). It also detects three functions when we increase the weight of data energy, as shown in Figure 3(c). In other words, if we want to undervalue "repulsive power" between adjacent functions, my method can detect functions in such a way. This result demonstrates that the proposed method is useful in that it detects patterns in a way preferred by us.

3 Discussion

Improving our understandings of promoters is one of the most important goals to analyze the distributions of transcription start sites. Because promoters are responsible for transcriptional regulation, we can make a hypothesis that promoters decide expression patterns of the TSSs distributions in the downstream sequence. In other words, expression patterns of the TSSs distributions might be determined by the type of promoter sequences. For this purpose, we need to classify the TSSs data, and one of the most straightforward approaches is using machine learning techniques. In this classification, training data is a set of known promoter sequences and their downstream TSSs distributions, and we can classify TSSs distribution data that is not related with the known promoters on the basis of the training. Although we can perform this classification by using raw TSSs data, we can also utilize the patterns of the TSSs distributions that my method detects. Using the results of my method has an advantage over using raw data, because we can incorporate our assumption of the data into this classification. For example, the distribution in Figure 3(a) is obscure in that deciding the number of Gauss functions in the distribution is difficult. However, we can use Figure 3(b) for the classification, if we think this distribution consists of two Gauss functions based on the knowledge of biology. We can also use Figure 3(c) for the classification, if we think this distribution consists of three Gauss functions based on the knowledge of biology. More specifically, if we think the promoter sequence upstream the distribution in Figure 3(a) has an effect on two points in the downstream region, we should use Figure 3(b) for the classification. If we think the promoter sequence upstream the distribution in Figure 3(a) has an effect on three points in the downstream region, we should use Figure 3(c) for the classification. In this respect, the proposed approach is useful for biologists because they can test their hypotheses by using it.



(a)

(b)

(c)

Figure 3 (a) A synthetic data composed of three Gauss functions.
(b) Pattern detection of (a) using the proposed method 1. The data and prior energy terms are equally considered in the global energy function.
(c) Pattern detection of (a) using the proposed method 2. The data energy is more overvalued than the prior energy in the global energy function.