Chapter II
Data Mining for Biologists

Koji Tsuda
Max Planck Institute for Biological Cybernetics, Germany

ABSTRACT

In this tutorial chapter, the author reviews basics about frequent pattern mining algorithms, including itemset mining, association rule mining, and graph mining. These algorithms can find frequently appearing substructures in discrete data. They can discover structural motifs, for example, from mutation data, protein structures, and chemical compounds. As they have been primarily used for business data, biological applications are not so common yet, but their potential impact would be large. Recent advances in computers including multicore machines and ever increasing memory capacity support the application of such methods to larger datasets. The author explains technical aspects of the algorithms, but do not go into details. Current biological applications are summarized and possible future directions are given.

INTRODUCTION

As new techniques for obtaining biological data continue to be introduced and a huge amount of data is accumulated in a large number of databases, the importance of data mining is ever increasing (Han & Kamber 2000). Modern high throughput technologies create comprehensive information about genome sequences, gene expression, polymorphism, protein interactions etc. However, extracting important information is not a trivial task. It is not only because of the huge amount of data. It is hard to define what is the important information in a current context mathematically. In collaboration among computer scientists and biologists, clear definition of the task is the key to success. It is often the case that computer scientists require biologists to define the mathematical definition of what they want from data. For example, it might be a statistical classifier predicting some phenotype from SNP data. But when biologists do not have sufficient knowledge about data mining methods, or computer scientists do not understand the biologists' motivation well, such collaboration might end up in failure. One typical failure is to define an impossible task which cannot be solved by any reasonable data mining method. This tutorial chapter is supposed to give computational biologists some background knowledge about data mining methods.

There are many different classes of methods in data mining. Most popular ones include clustering, supervised classification, regression, sequence alignment etc. There are good introductory books about data mining methods in biology, for example, (Schoelkopf et al., 2004). In this chapter, we cannot focus on all techniques, but concentrate on frequent pattern mining algorithms and their possible applications to biological data.

The most popular method of pattern mining algorithms is itemset mining (Agrawal & Srikant, 1994), which has been extensively applied to market basket analysis. Suppose a supermarket welcoming many customers every day. Each customer buys more than one item. When each item is indexed by an integer, each transaction (i.e., the items bought by a customer) is summarized as a set of integers (Figure 1). Then, the goal of itemset mining is to enumerate the subsets of items which occurred in more than m transactions. The frequency threshold m is called the minimum support parameter. Found subsets are called patterns. In a popular example of market basket analysis, it is discovered that beer and diaper are likely to be sold simultaneously. The frequent co-occurrence of two items is considered as very strong signal of association between them. Biological objects are often represented as an itemset. For example, a promoter sequence can be represented as a set of sequence motifs. A patient is a set of genotypes at several quantitative loci. Applications to gene expression data can be seen, e.g., in (Creighton & Hanash, 2003; Tamura and D'haeseleer, 2008).

The idea of frequently appearing substructure is commonly exploited in sequence motif analysis (Bailey et al., 2006). Their objective is to find a frequently appearing substring from a set of strings. One point which discriminates motif analysis from frequent pattern mining is that the motif analysis finds only one motif at a time, whereas frequent pattern mining enumerates all frequently appearing patterns. For example, MEME (Bailey et al., 2006) can derive a probabilistic motif from a set of k-mers. Even though there are many possible motifs, it can extract one motif because it is based on a mixture model of two components, one of which is the background model. It could be extended to deal with multiple motifs by increasing the number of components, but then the optimization of parameters gets more difficult (Blekas et al., 2003).

Pattern mining algorithms are combinatorial methods that enumerate all solutions satisfying a set of pre-specified criteria (e.g., frequency). We call the other methods which create only one optimal solution "optimization methods". Often, the enumerative nature of itemset mining is misunderstood as confusing. Itemset mining creates a lot of frequent patterns which have more or less the same quality. On the other hand, a conventional optimization method show only one solution, which is conceived as "clear". However, in many situations in biological data analysis, the amount of data is far too short for determining unique solution with absolute confidence. In my opinion, multiple solutions by pattern mining methods are a natural consequence of the reality. An advantage is that biologists can select good ones from the solutions by their own biological knowledge. Those who anticipate a magic box that provides one reliable solution which is both statistically significant and biologically meaningful might find enumerative methods useless. However, the only one solution by optimization methods is easily perturbed by noise and premature finish of parameter optimization (i.e., local minima).

Itemset mining has been extended to more structured data, such as transaction sequences (Pei et al., 2004), trees (Asai et al., 2002) and labeled graphs (Yan & Han, 2002). In this tutorial, we focus on graph mining, because other techniques are considered as specialization of graph mining. Much of real data is described as labeled graphs. For example, chemical compounds are represented as graphs where nodes represent atoms, and edges represent bonds among them (Kazius et al., 2006). Graph mining can be applied to the 3D structure of proteins as well

Figure 1. Examples of frequent itemsets