Chapter 3.11
Comparative Genome Annotation Systems

Kwangmin Choi
Indiana University, USA

Sun Kim
Indiana University, USA

ABSTRACT

Understanding the genetic content of a genome is a very important but challenging task. One of the most effective methods to annotate a genome is to compare it to the genomes that are already sequenced and annotated. This chapter is to survey systems that can be used for annotating genomes by comparing multiple genomes and discusses important issues in designing genome comparison systems such as extensibility, scalability, reconfigurability, flexibility, usability, and data mining functionality. We also discuss briefly further issues in developing genome comparison systems where users can perform genome comparison flexibly on the sequence analysis level.

INTRODUCTION

Once a complete genome sequence becomes available, the next and more important goal is to understand the content of the genome. The exponential accumulation of genomic sequence data demands use of computational approaches to systematically analyze huge amount of genomic data. The availability of such genome sequence data and diverse computational techniques has made comparative genomics—research activity to compare sequences of multiple genome sequences—to become useful not only for finding common features in different genomes, but also for understanding evolutionary mechanisms among multiple genomes.

The process of assigning genomic functions to genes is called *genome annotation*, which utilizes diverse domain knowledge sources from sequence data to the contextual information of the whole genome. Currently there exist several methods for genome annotation. Experimental approaches for genome annotation are probably most reliable, but slow and labor-intensive. Genome annotation can be done computationally as well by (1) assigning function(s) to a gene based on its sequence similarity to other genes that are already annotated with well-defined gene functions, (2) assigning function(s) to a gene based on its position in a conserved gene cluster through comparative analysis of multiple genomes, and (3) inferring function via detecting functional coupling.

Genome annotation probably can be done most accurately by comparing a genome with its phylogenetically-related genomes, which we termed as *comparative genome annotation*. Comparing multiple genomes, however, is a very challenging task. First of all, genome comparison requires handling complicated relationships of many entities. For example, comparison of all protein coding genes in three genomes with 2,000 genes in each genome can involve several million pairwise relationships among genes. The way of genome selection raises another problem because the choice of genomes to be compared is entirely subjective and there are numerous combinations of genomes to be compared. Thus it is necessary to develop an information system for comparative genome annotation which can deal with such challenges.

This chapter primarily aims to survey existing systems from the data mining perspective. Well developed data mining tools will be very helpful in handling numerous functional relationships among genes and genomes. For example, an accurate sequence clustering tool can simplify the genome annotation task significantly, since the user can handle a set of sequences as a single unit for the next analysis. Unfortunately, existing data mining tools are not developed to handle a continuous character stream or genomic sequence, thus our discussion from the data mining perspective is to propose what is needed, rather than characterizing existing systems in the terms of traditional data mining perspective. Here we propose the desirable features that comparative genome annotation systems should have:

- **Extensibility:** There are always new resources such as newly sequenced genomes and computational tools. The system should be able to include them as they become available.
- **Reconfigurability:** There are numerous different ways to combine data and tools, so no single system can meet all needs of users. When needed, it is desirable to reconfigure the system for a specific task.
- **Genome selection flexibility:** The choice of genomes to be compared is entirely subjective. Thus users should be able to compare genomes of their choice with different criteria for sequence comparison.
- **Usability:** Genome comparison involves a huge amount of data, so the system should be easy to use. In addition, it should be easy to port to other platforms.
- **Data mining:** Users need to perform a series of analyses to achieve a research goal and the system should also provide high-level data mining tools to simplify genome analysis task. To explain the data mining issue further, assume that a user wants to analyze a sequence s against several genomes, G_1, G_2, G_3. Obviously a user can perform three different comparisons of s to each G_i and then "combine" the results into one, and then proceed for further analysis. A well developed system can easily mitigate the burden of combining three different search results by providing some system functions. Alternatively, it is possible to use a high performance sequence clustering algorithm to generate the output in a single operation.