Chapter 3

Introduction to Algorithms

Senthilkumar Radhakrishnan\textsuperscript{1}, Deepak Kolippakkam\textsuperscript{2},
and Venkatarajan S. Mathura\textsuperscript{3}

\textsuperscript{1}California Institute of Technology, Pasadena, USA
\textsuperscript{2}Harvard University, Boston, USA
\textsuperscript{3}Roskamp Institute, 2040 Whitfield Avenue, Sarasota, FL 34243

Abstract: Computational methods are designed to solve complex problems systematically and efficiently. Classification and selection procedures are often used in biological sequence analysis and other data analysis. This chapter provides an introduction to different methods like clustering, hypothesis testing, and classification methods.

Key words: Bayesian methods, Decision trees, Bayesian, Neural network, Clustering, Support Vector Machines, PCA

1. Introduction

Systematic solutions to complex problems are highly desirable and algorithms are designed to efficiently solve them in a finite time. Rigorous data-analysis and computer science problems use computer programs that implement suitable algorithm to arrive at a solution. These include recursive/iterative methods, graph search algorithms, dynamic programming, greedy algorithms, etc. A thorough discourse on computational algorithms is out of scope for an introductory book but we provide here some of the term definitions of widely used computational methods that are frequently applied in the forthcoming chapters in bioinformatics.

1.1 Classification

The process of dividing a dataset into mutually exclusive groups such that the members of each group are as “close” as possible to one another, and
different groups are as “far” as possible from one another, where distance is measured with respect to specific variables or class labels (for predicting) is known as classification. For instance, determining whether a protein binds to DNA or not based on sequence and structural motifs, cancer type classification based on micro array expression are good examples of classification problems. Classification falls under the category of “supervised learning”. Some of the classification methods widely applied in bioinformatics is Decision trees, Support-vector machine based classifier, Bayesian Classifiers, Neural Network Classifiers, etc.

1.2 Hypothesis Testing

Hypothesis tests are statistical procedures for making rational decisions about the reality of effects. It is an inference technique where either one accepts a null hypothesis (Ho, e.g.: no difference between control and treatment group) or rejects it (that is accepting the alternative). Either one is true, but not both (mutually exclusive and exhaustive). Hypothesis testing is carried out on the observed sample data that represents characteristic population using test statistics. The test statistic quantifies the difference between normally distributed (in case of t-test) hypothetical population and the observed data, which can be used to obtain a p-value. The p-value is the probability to observe sample data assuming null hypothesis is true. A p-value of 0.01 means that the chance of observing sample data is only 1/100 while there is no effect or null is true (no effect on treatment). Since the chance of such observation is very small, we reject null hypothesis and conclude that there is a difference between control and the treatment group. Several statistics exist: t-statistic is for testing mean, F, or Chi-test for testing variance. Hypothesis testing is used in micro array data analysis, sequence analysis, etc.

1.3 Decision Tree

Decision trees are a simple approach to the problem of learning from a set of independent instances. Every node in the decision tree (except the leaf nodes) involves testing of a particular attribute. In most cases, the test at a node compares an attribute value with a constant. Sometimes, the test may be between the values of a set of attributes too. Leaf nodes give the classification labels or the probability distribution over all possible classifications. Decision trees are used in protein secondary structure prediction (Selbig et al., 1999), protein sorting signal prediction, etc.