COMPUTATIONAL BIOLOGY

Avoiding common pitfalls when clustering biological data

Tom Ronan, Zhijie Qi, Kristen M. Naegle*

Clustering is an unsupervised learning method, which groups data points based on similarity, and is used to reveal the underlying structure of data. This computational approach is essential to understanding and visualizing the complex data that are acquired in high-throughput multidimensional biological experiments. Clustering enables researchers to make biological inferences for further experiments. Although a powerful technique, inappropriate application can lead biological researchers to waste resources and time in experimental follow-up. We review common pitfalls identified from the published molecular biology literature and present methods to avoid them. Commonly encountered pitfalls relate to the high-dimensional nature of biological data from high-throughput experiments, the failure to consider more than one clustering method for a given problem, and the difficulty in determining whether clustering has produced meaningful results. We present concrete examples of problems and solutions (clustering results) in the form of toy problems and real biological data for these issues. We also discuss ensemble clustering as an easy-to-implement method that enables the exploration of multiple clustering solutions and improves robustness of clustering solutions. Increased awareness of common clustering pitfalls will help researchers avoid overinterpreting or misinterpreting the results and missing valuable insights when clustering biological data.

Introduction

Technological advances in recent decades have resulted in the ability to measure large numbers of molecules, typically across a smaller number of

Differentiating between a meaningful and a random clustering result can be accomplished by applying cluster validation methods, determining statistical and biological significance, accounting for noise, and evaluating