Abstract:

- Sequence clustering attracts attention by developing the metagenomics and microbiomics.
- The latest sequencing techniques have decreased costs but produced a massive amounts of DNA/RNA sequences.
- The challenge is to cluster the sequence data by quick and accurate methods.
- There is a gap between algorithm developers and bioinformatics users, because of using 16S ribosomal RNA operational units.
- Software tools produce variant results which are difficult to be analyzed or understanded.
- The different clustering mechanisms enable understanding the massive results, so we selected popular clustering tools.
- By using two independent benchmark datasets these tools explained the key computing principles, analyzed their characters and compared them.
- We aim is to employ clustering tools effectively to analyze big data.

Introduction

- There are two different types of machine learning tasks in clustering, supervised and unsupervised learning, the difference is in the training set (known or unknown labels).
- Machine learning focuses on vectors, features or attributions which all need to be classified or clustered.
- We have some tools that transform DNA/RNA/protein sequences to numeric vectors but clustering sequences directly is the preferred way.
- We review bioinformatics molecular sequence-clustering algorithms and the applications.
- Genes obtained from microarray data are clustered and genes in the same cluster are considered to do the same function.
- Several advanced microarray clustering algorithms was proposed with hierarchical and ensemble clustering.
- Bi-clustering has been employed to avoid the noise in microarray data and the feature, sample selection is performed at the same time.
- Single-cell sequencing data work similar to gene expression data.
- Different gene expressions were clustered by different cell types, so these methods focus on gene expression values Not gene sequences.