Abstract:

- Sequence clustering attracts attention by developing the metagenomics and microbiomes.
- 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

Related work

1. Liu B, Wu H, Chou KC. Pse-in-One 2.0:

an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nat Sci 2017;09(4):67–91

https://www.scirp.org/html/1-8302870_75771.htm

2. Li YH, Xu JY, Tao L, Li XF, Li S, Zeng X, Chen SY, Zhang P, Qin C, Zhang C, Chen Z. SVM-Prot 2016:

a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. PloS One 2016;11(8):e0155290.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155290

3. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One:

a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res 2015;43(Web Server issue):W65–71.

https://academic.oup.com/nar/article-abstract/43/W1/W65/2467922

4. Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou KC. PseKNC-General:

a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 2014;31(1):119–20

https://academic.oup.com/bioinformatics/article-abstract/31/1/119/2366158