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.

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

methododlgy :

Feature encoding algorithms

- To capture the key information of CPPs, we used nine different feature encoding algorithms from multiple perspectives, such as compositional information, position-specific information, as well as physicochemical properties, etc. Below, we described how to use the feature encoding algorithms to encode variable- length CPPs into fix-length feature vectors

Amino acid composition

- The amino acid composition □AAC□ is simple but commonly used feature descriptor. For a total of 20 amino acid types, this feature descriptor calculates the frequency of each type occurring in peptide sequences. For example, if the amino acid type i occurs ni times in a given peptide sequence, the frequency of i, denoted as f(i) = ni/L, where L is the length of the sequence. For a given peptide, we yielded a 20-dimensional feature vector by computing the frequencies of 20 different amino acids in the peptide, respectively

Twenty-bit features

- This algorithm is also known as binary profile algorithm, measuring the position- specific information of residues [10]. In this algorithm, each amino acid is encoded into a 20-bit 0/1 vector. (e.g. Ala by 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0□. Therefore, the N□C terminus with 2k residue long is encoded with a 20 □ 2k- dimensional feature vector.

Results :

Comparison of class and probabilistic information for feature representation

- In our feature representation learning scheme, we used two types of information, class and probabilistic information, from the predictions of 45 RF models to encode peptide sequences, respectively. In this section, we discuss which information is more discriminative to classify CPPs from non-CPPs.
- We first compared the two original 45-dimensional feature vectors encoded with class and probabilistic information, respectively. The results are presented in Table 2. As shown in Table 2, we can see that the feature vector using class information performs slightly better than the feature vector using probabilistic information.
- The overall performances in terms of ACC and MCC using class information are 91.1% and 0.823, while that using probabilistic

information are 90.9% and 0.818. To further compare the two types of information

Determination of optimal feature representations

- As described in Feature representation learning scheme, we applied mRMR feature selection technique to the 45-dimensional feature vector generated from a total of 45 RF models. By doing so, we obtained a ranked feature list for the 45 features according to their importance scores calculated by mRMR. The importance scores for all the 45 features are summarized in Supporting Information (Table S2). Afterwards, we used the SFS strategy to find the optimal feature representations from the ranked feature list. We took the features one by one from the list and trained the RF classifier, which was then evaluated with 10 fold cross-validation on the CPP924 data set.