Expert Systems with Applications 40 (2013) 6185–6194

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Expert system for clustering prokaryotic species by their metabolic features

Clara Higuera [a,b,*], Gonzalo Pajares [b], Javier Tamames [c], Federico Morán [a]

[a] Dpt. Bioquímica y Biología Molecular I, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, avda. Complutense s/n, 28040 Madrid, Spain
[b] Dpto. Ingeniería del Software e Inteligencia Artificial, Facultad Informática, Universidad Complutense, C/ Prof. José García Santesmases, s/n. 28040 Madrid, Spain
[c] Centro Nacional de Biotecnología, National Research Council (CSIC), c/Darwin, 3. Cantoblanco, 28049 Madrid, Spain

ARTICLE INFO

Keywords:
Expert system
Clustering
Self-organizing Maps
Clustering validity indices
Metabolism
Prokaryotic species

ABSTRACT

Studying the communities of microbial species is highly important since many natural and artificial processes are mediated by groups of microbes rather than by single entities. One way of studying them is the search of common metabolic characteristics among microbial species, which is not only a potential measure for the differentiation and classification of closely-related organisms but also their study allows the finding of common functional properties that may describe the way of life of entire organisms or species. In this work we propose an expert system (ES), making the main contribution, to cluster a complex data set of 365 prokaryotic species by 114 metabolic features, information which may be incomplete for some species. Inspired on the human expert reasoning and based on hierarchical clustering strategies, our proposed ES estimates the optimal number of clusters adequate to divide the dataset and afterwards it starts an iterative process of clustering, based on the Self-organizing Maps (SOM) approach, where it finds relevant clusters at different steps by means of a new validity index inspired on the well-known Davies Bouldin (DB) index. In order to monitor the process and assess the behavior of the ES the partition obtained at each step is validated with the DB validity index. The resulting clusters prove that the use of metabolic features combined with the ES is able to handle a complex dataset that can help in the extraction of underlying information, gaining advantage over other existing approaches, that may relate metabolism with phenotypic, environmental or evolutionary characteristics in prokaryotic species.

1. Introduction

Trying to understand the communities of microbial species is highly important because many natural and artificial processes are mediated by groups of microbes rather than by isolated entities. In order to create artificial communities or manipulate the existing ones it is necessary to comprehend the specific requirements of the individual species and to be able at a long term to predict in which conditions are they able to survive.

One kind of microorganisms which are important in life are the prokaryotes and a way of studying them have been since many years trying to categorize (Hong, Kim, & Lee, 2004) the huge variety of prokaryotic organisms which is itself a challenging task. One of the reasons is the lack of a globally accepted concept of species for prokaryotes and the fact that their taxonomy is continuously being influenced by the advances in microbial population genetics, ecology and genomics (Gevers et al., 2005). When it comes to assign an unknown bacteria to a species the experts usually do it identifying phenotypic or genome similarity.

The traditional method to classify prokaryotes has been since many years the identification of the 16S rRNA (Jain, Wang, Liao, & Boyd, 2009), a sequence highly conserved through evolution. It allows to find differences among microorganisms and build evolutionary trees, also called phylogenetic trees that show the evolutionary relationships among species that are believed to possess a common ancestor.

Although the analysis of 16S rRNA has been widely and successfully applied, experts have started to look for other kinds of information which may shed some light into the differentiation of prokaryotic species. One of them is the search of common metabolic characteristics, which some authors suggest to be not only a potential measure for the classification or differentiation of closely-related organisms (Lee et al., 2012) but also that their study may allow the finding of common functional properties that traditional methods such as the analysis of 16S rRNA is not able to find (Jain et al., 2009).

In biochemistry a metabolic pathway consists of a set of reactions that take place inside the cell, it involves the transformation of substrates into different products necessary for maintaining its

* Corresponding author at: Dpt. Bioquímica y Biología Molecular I, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, avda. Complutense s/n, 28040 Madrid, Spain. Tel.: +34 91 394 4265.
E-mail addresses: pajares@fdi.ucm.es, clarahiguera@ucm.es (C. Higuera).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.05.013

life. These reactions are intermediated by molecules called enzymes, which are responsible for the proper performance of the pathway. Several different pathways can co-exist inside the cell. The collection and spatio-temporal organization of the pathways in the cell, together is called metabolism. However, among species, neither do all present the same pathways nor do all enzymes have equal importance in their metabolism, because different types of metabolism are adapted to the specific conditions of the environment in which a species live including the interactions with other species.

One of the main problems in separating prokaryotic species by their metabolic features is the wide diversity of possible representations and their comparison, such as the number of common enzymes between two organisms, the presence or absence of various metabolic paths and so on (Clemente, Satou, & Valiente, 2005). Another problem lies on the lack of information from some species in relation to others, for instance while *Escherichia coli* is a very studied bacteria and its metabolic pathways have been deeply analyzed others like Rhodococcus Ruber there is almost no information. These are two problems addressed on this paper based on a computational intelligent approach.

Clustering techniques such as Self-organizing Maps (SOM) or Neural Networks have been often used with success in the fields of medicine, biology, biochemistry, ecology and microbiology when it comes to group or separate elements by common characteristics (Park, Céréghino, Compin, & Lek, 2003; Rabow, Shoemaker, Sausville, & Covell, 2002; Stegmayer, Gerard, & Milone, 2012; Szaleniec, 2012). Both of them have proven to be useful in the difficult task of inferring inherent information from groups of prokaryotic species and also to predict certain behavior or life styles (Bohlin, Skjerve, & Ussery, 2009; Larsen, Field, & Gilbert, 2012; Suen, Goldman, & Welch, 2007). When the number of samples and features involved is large, SOM can be very appropriate for cluster analysis when looking for underlying hidden patterns in data, what could lead to the formulation of new hypothesis. However few have been done in the field of clustering prokaryotic species by their metabolic features. To date, the clustering methods which use these kind of features have been, (to our knowledge), above all hierarchical clustering methods applied to the reconstruction of phylogenetic or evolutionary relationships among species.

The output of a hierarchical clustering is a tree called dendrogram which reflects the possible hierarchical clustering structure of the data. One of the advantages of the method is that the dendrogram has shown to result significantly similar to an evolutionary hierarchy when applied to prokaryotic and eukaryotic bacteria.

Hong et al. (2004) use a complete-linkage hierarchical clustering to construct a phylogenetic tree that represents the similarity of metabolic profiles of a set of 43 microorganisms. The method helped them to study the changes in metabolism as a result of the evolutionary process. Clemente et al. (2005) and Jaume Casasnovas (n.d.) utilize an average-link hierarchical clustering to reconstruct phylogenetic relationships from other metabolic features. (Jaume Casasnovas, n.d.) enhanced a year later the method by using fuzzy clustering. Other authors like Jain et al. (2009) found functional similarities among clustered species such as adaptation to cold environments or ability to suppress agriculture pathogens by using a novel method based on hierarchical clustering to generate comparison trees based on characteristics collected from metabolic networks of bacteria. Nevertheless, the authors of this work use a very reduced sample of only twenty species.

Even though hierarchical clustering has shown good results in terms of reconstructing phylogenetic trees, in order to be able to find inherent common functionalities different from metabolism among elements of a group, the structure of a dendrogram does not seem very helpful. The reason is that the expert must decide the appropriate level or scale of clustering to start considering groups. Hierarchical clustering does not actually create clusters, but compute only a hierarchical representation of the data set (Sander, Qin, Lu, Niu, & Kovarsky, 2003). Another problem of hierarchical clustering is that for big data sets, of hundreds of elements, it is extremely difficult to identify and visualize relationships between elements.

An alternative are the classical unsupervised clustering algorithms, also called partitioning algorithms, which divide unlabelled data into defined groups (clusters) of similar elements. The fact that the output of these methods is a set of clusters makes easy the biological interpretation of the results and the possible extraction of common underlying information in data.

However, clustering complex datasets is a very hard and arduous task. When applying clustering methods many problems arise: algorithms are usually very sensitive to data which may be noisy or incomplete, also similar algorithms can result in much worse performance than others when applied to the same dataset, so there exist a need of selecting the best method for the data being used. Moreover there are cases in which the optimal number of clusters that best describes the topology of the data is unknown. This is most of the times the case of biological data, that in contrast to other fields there is not any a priori information about it. A later stage of the clustering process is the assessment of the final partition, which is also particularly difficult when dealing with biomedical data, especially in cases of lack of experimental scientific corroboration or expert supervision.

In this work we are facing the problem of clustering a set of 365 prokaryotic species, which represent a considerable number of species, where hierarchical clustering approaches becomes ineffective because of the problems described above. Concerning classical clustering approaches the most promising method was SOM, its performance is normally tested based on indices that measure the quality of the clustering, one of them is the well-known Davies Bouldin index (DB) (Davies & Bouldin, 1979). We have tested different clustering strategies for our dataset, including Fuzzy Clustering (Duda, Hart, & Stork, 2001) and Learning Vector Quantization (Pandya & Macy, n.d.). We have verified that such methods obtained not only a worse distribution of the species in clusters than SOM, confirmed by the experts, but also worse values of DB. Thus, because of the performance of SOM together with the idea of hierarchical strategies based on the reasons expressed below we design an expert system (ES) exploiting the performance of SOM and applying a hierarchical philosophy valid for the large dataset analyzed with the aim of proving its validity for large datasets of microbial species.

The design of the ES is based on what would be common in human reasoning with such complex data. When the human expert is faced with this type of data, a common practice is to begin separating the data with less difficulty making progresses towards higher levels of difficulty. At each step, a minimum degree of confidence is required about the appropriate progress. This proposal with its tested effectiveness makes the main contribution of this paper.

2. Materials and methods

2.1. Material

The metabolic features selected for the clustering were the percentages of annotated enzymes that a set of species possess of certain metabolic pathways. This way, each species will have a vector of pathways assigned which contains that percentage. This value is an indicator of how complete is the pathway in a species. A high value would mean that the species contains all or most of the enzymes of a pathway while a zero value would mean that the