Abstract

Manal Helal

Defining reference sequences for Nocardia species by similarity and clustering analyses of 16S rRNA gene sequence data

Background: The intra- and inter-species genetic diversity of bacteria and the absence of ‘reference’, the most representative, sequences of individual species present a significant challenge for sequence-based identification. The aims of this study were to determine the utility, and compare the performance of several clustering and classification algorithms to identify the species of 364 sequences of 16S rRNA gene with a defined species in GenBank, and 110 sequences of 16S rRNA gene with no defined species, all within the genus Nocardia. Methods: A total of 364 16S rRNA gene sequences of Nocardia species were studied. In addition, 110 16S rRNA gene sequences assigned only to the Nocardia genus level at the time of submission to GenBank were used for machine learning classification experiments. Different clustering algorithms were compared with a novel algorithm the linear mapping (LM) of the distance matrix. Principal Components Analysis was used for the dimensionality reduction and visualization. Results: The LM algorithm achieved the highest performance and classified the set of 364 16S rRNA sequences into 80 clusters, the majority of which (83.52%) corresponded with the original species. The most representative 16S rRNA sequences for individual Nocardia species have been identified as ‘centroids’ in respective clusters from which the distances to all other sequences were minimized 110 16S rRNA gene sequences with identifications recorded only at the genus level were classified using machine learning methods. Simple kNN machine learning demonstrated the highest performance and classified Nocardia species sequences with an accuracy of 92.7% and a mean frequency of 0.578. Conclusion: The identification of centroids of 16S rRNA gene sequence clusters using novel distance matrix clustering enables the identification of the most representative sequences for each individual species of Nocardia and allows the quantitation of inter- and intra-species variability.