NCBI NR Protein Database Clustered by Homology Inference

Aryel Marlus Repula de Oliveira, Roberto Tadeu Raittz

Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Programa de Pós Graduação em Bioinformática

Gene clustering is used to infer proteic features such as homology and function, for example, and it improves the performance of metagenomic analysis, computational searches optimization, etc. Methods for inferring homology from sequences have been an intense topic in research in the past years. The most precise techniques are based in statistical approaches, like Hidden Markov Models allied with sequence similarity comparison. This kind of statistical approach, when used in a huge database like NCBI NR, has two main challenges: firstly, the high computational complexity increases the time spent to perform the clustering and secondly, the training of Markov models without a reference, converges to local maximum of probabilities and do not generate a consistent classification. Due to these challenges, public databases are currently limited either in sensibility or size and the available clustering tools also share these same limitations. We propose a new solution to gene clustering, based on pattern recognition of physical and chemical characteristics of proteins with neural networks, allied with a new and fast algorithm for sequence similarity search that creates a matrix representation of the database sequences, consuming less memory and simplifying the computational processing for comparison and machine learning. Preliminary results demonstrate that the new similarity search algorithm is approximately 4 times faster than the CD-HIT tool, which is used to cluster UniRef, the larger clustered protein database available, producing a similar product than our new algorithm. In this new approach, we create a matrix representation of the NCBI NR protein sequences and start the clustering process considering 50% of similarity; later, we train the neural networks considering, initially, 35 consolidated physical and chemical characteristics and validate the results against curated protein clusters and families. We expect to provide a clustered protein database considering homology inference and also provide a new tool for clustering, using the proposed solution.