Classification of amino acid residue pairs using GMM and EM Algorithm

Higor Coimbra Amorim

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS

Abstract

RID (Residue Interaction Database) is a system built to propose site-directed mutations on 3D protein structures using PDB files. One of the steps of RID process is to indicate which amino acid residue pairs of a protein are able to receive a mutation and then, all of the candidates are classified according to their atomic structure similarities. On RID, these classifications are based on a score produced by the overlap of all candidates atomic structures using LSQKAB. However, for a large dataset of PDB files, for example a dataset with about 16000 elements, as the one used in this research, the overlap made by LSQKAB can be a very large time consuming process. One of the proposals made to replace LSQKAB was the use of an atom to atom distance matrix to replace the residue pairs' PDB files and the use of K-Means clustering algorithm to replace the score overlap classification. Results showed that K-Means was a viable method to cluster residue pairs and could be used inside the RID system, as part of the structural classification process. Following the good results obtained by K-Means, this research proposes the use of a more flexible method of clustering: Gaussian Mixture Models (GMMs) used within the Expectation-Maximization (EM) algorithm. The first results showed that GMMs can also be a viable clustering algorithm and a candidate to replace LSQKAB. The clusters' overall biological similarity for a number of 500 and 750 clusters on the GMM are 4.73%, 0.1% higher, respectively, than the ones obtained by K-Means, based on a dataset of 16383 PDB files of amino acid residue pairs.

Funding:

Link to Video: