GReMLIN: A graph mining strategy to infer protein ligand interaction patterns

Charles A. Santana[1], Fabio R. Cerqueira[1], Carlos H. da Silveira[2], Alexandre V. Fassio[3], Raquel C. de Melo-Minardi[3], Sabrina de A. Silveira[1]

Universidade Federal de Viçosa[1], Universidade Federal de Itajubá[2] and Universidade Federal de Minas Gerais[3]

Interaction between proteins and ligands are relevant in many biological process. Such interactions have gained more attention as the comprehension of protein-ligand molecular recognition is an important step to ligand prediction, target identification and drug design. This work proposes GreMLIN, a strategy to search patterns in protein-ligand interactions based on frequent subgraph mining. Here, we investigated if it is possible to find patterns that characterize protein-ligand interactions in a set of selected proteins. This patterns can be key factors to understand and support the recognition molecular process. Moreover, if such patterns exist, we believe that they can represent an important step in the prediction of the protein-ligand interaction.

Our strategy models protein-ligand interfaces as bipartite graphs where nodes represent protein or ligand atoms and edges represent interactions among them. The nodes and edges are labeled with physicochemical proprieties of atoms and a distance criteria. A clustering analysis is performed on graphs to characterize them according their similarities and differences, and a subgraph mining algorithm is applied to search for relevant patterns on protein-ligand interfaces in each cluster.

We collected structural data of protein-ligand complexes in Protein Data Bank (PDB) to validate our strategy and show their applicability. There are two datasets: (i) the CDK (Ciclin dependent kinases) dataset that have 73 PDB entries with identical sequences coupled with different ligands; and (ii) the Ricin dataset with 29 PDB entries, which share sequence identity greater than or equal to 50% with ricin template 2AAI chain A. Both datasets have biological relevance, but with different characteristics. Our strategy was able to find frequent substructures with considerable cardinality in the protein-ligand interfaces in the CDK and Ricin datasets. We provide the results of our strategy for the test datasets in a prototype interactive tool to visualize and explore the patterns found in protein-ligand interactions. Also, we provide a schematic 2D graph representation of such interactions and a 3D representation of these interactions in a molecule viewer. Availability: *http://homepages.dcc.ufmg.br/~alexandrefassio/gremlin/*. Financial support: CAPES, CNPq, FAPEMIG.