**ABSTRACT**

Markov Cluster Algorithm (MCL) is a fast and scalable unsupervised graph clustering algorithm, based on stochastic flow simulation model. By random walks on the graph, the flow of the graph is estimated for the clusters. It simulates flow using two simple algebraic operations on matrices i.e. expansion and inflation. The expansion operator connects different regions of graph while inflation operator strengthens and weakens the flow. This procedure follows until a steady state is reached and all the values in column are same. This algorithm works efficiently for clusters with small diameters. It is suitable for fast and accurate grouping of protein structures in human body.

Restricted neighborhood search Algorithm (RNSC) is also graph clustering technique that uses stochastic local search. Cost functions are assigned to clusters of a graph to get an optimal cost clustering. It is implemented for un-weighted and undirected graphs and uses heuristic approach. A number of moves are considered in this algorithm. A global move results in an optimal change in cost, diversification move shuffles the clustering randomly and an intensification move is associated with a good cost chosen from restricted portion of the current clustering’s neighborhood. To escape cycling, a tabu list is created to avoid same moves in the algorithm. This algorithm is efficient as it reduces the run time of clustering algorithm. The paper proposes behavior analysis of both the algorithms effectively which is implemented in Python programming language.

**ACKNOWLEDGEMENT**

An understanding such as this one is never the work of a single person and this research work is not an exception. We, at this very outset, thank all those who directly or indirectly helped me in the completion of my research work and encouraged us in moments of our despair.

We wish to express our appreciation to \*\*\*\*\*\*\*\*\*\*\*\*under whom we worked, interacted and whose insights and thoughts helped not only in completion of my research work but also in enhancing my knowledge. We’re highly indebted to his generous help and guidance throughout the development of our research work.

Finally, we would like to heartily thank all my class mates without their help & support our research work would not have been successfully completed.

And last but certainly not the least; we acknowledge our deep gratitude to all the persons concerned.

*- NAME*

*- ROLL NO*

*B.Tech (Final Year)*

*Computer Science and Technology.*

**TABLE OF CONTENTS**

* 1. Abstract ……………………………………………………………………………..i
  2. Acknowledgement…………………………………………………….....................ii
  3. Introduction ………………………………………………………………………..1
  4. Literature Survey ………………………………………………………………….4
  5. Research methodology …………………………………………………………….6
  6. Work plan…………………………………………………………………………..9
  7. Proposed Outcome ………………………………………………………………..9
  8. Bibliography and References………………………………………………….......9
  9. Appendix………………………………………………………………………......11

**INTRODUCTION**

Markov cluster algorithm is an unsupervised clustering algorithm which simulates the flow of the graph. It was first implemented by Stijen van Dogen at the Center for Mathematics and Computer Science in Netherlands. A Markov model is a stochastic model used to model randomly changing systems where it is assumed that future states depend only on the present state and not on the sequence of events that preceded it. Clusters in a graph are characterized by the presence of edges between the nodes of the clusters and the number of longer paths in the clusters should be high.

The MCL algorithm finds the clusters in a graph by a mathematical bootstrapping procedure. This process computes the probabilities of random walks through graph, and uses two operators (i.e. expansion and inflation) that transform one set of probabilities into another. This is done by stochastic matrices which are also known as Markov matrices that uses the concept of random walks on a graph. Expansion process is done by taking the power of a stochastic matrix using normal matrix product (matrix squaring). Inflation step involves the inflating of the resulting matrix, followed by the scaling step such that the resulting matrix is stochastic again i.e. the matrix elements (on each column) correspond to the probability values.

Iterating expansion and inflation results in the separation of graph into different segments. There are no longer any paths between these segments and collection of resulting segments is interpreted as a clustering. Inflation parameter can be altered using parameter ‘r’. By increasing this parameter, the inflation operator can be strengthened which increases the granularity or tightness of the clusters. Expansion causes flow to dissipate within clusters and inflation eliminates flow within clusters. When the matrix reaches a steady state i.e. the matrix elements does not change with further expansion and inflation steps.

Restricted Neighborhood Clustering algorithm (RNSC) is a popular method to detect the complexes in protein-protein interaction networks. It explores the solution space for all possible clusters by minimizing cost function that reflects the number of inter-cluster and intra-cluster edges. This algorithm is a stochastic meta-heuristic technique and can be used in various search space representations. It is a cost based clustering and performs local search iteratively to obtain optimum clustering in an efficient way. The main goal of this algorithm is to find the best cost for the set of clusters from a graph by assigning some cost functions (i.e. naive cost function and scaled cost function). The memory requirement for this algorithm is O (n^2). The complexity of a move in Naive cost function is O (n), which is the size of the restricted neighborhood of a move M.

Naive cost function acts as a pre-processor to produce initial clustering result on a graph and the Scaled function tries to optimize the output from naïve function to reach a global optimal solution. The scaled cost function is applied to evaluate a low cost clustering result. A large number of data structures are used in this algorithm which decreases the runtime of this RNSC algorithm. It first computes all the possible moves and then makes the best possible move available as compared to the other algorithms where initially the move is made and then the cost is computed. A number of moves are being considered in RNSC. A global move is a move that results in a near optimal change in cost. A Diversification move randomly shuffles the clustering and can be achieved through different ways. An Intensification move is associated with a good cost chosen from the restricted part of the current clustering’s neighborhood.

Here the performance of MCL and RNSC is done on the real scale- free graphs. The comparative assessment is measured in terms of cost of clustering, cluster size, modularity index of clustering results and NMI value to find out the better and more efficient algorithm for analyzing the protein-protein interaction network.