Original Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

EXPLORING KEY GENE INTERACTIONS USING PARTICLE SWARM OPTIMIZATION

NIRANJAN.R[1], AMAL PRAKASH.N[1],SREEJA ASHOK[*1], M.V.JUDY[1]

Department of Computer Science & I.T.Amrita School of Arts & Sciences, Kochi, Amrita Vishwa Vidyapeetham, Amrita University, India

ABSTRACT

Clustering is an exploratory method that is widely used for analyzing similarity of data objects. Clustering helps biologist in identifying functional similarity of genes. Most of the techniques employed for clustering genes need prior knowledge of the number of feasible clusters. Here we propose a novel hybrid approach towards gene clustering, which implements Particle Swarm Optimization (PSO) technique to find out closely related clusters by exploring the domain knowledge from gene ontology. The proposed approach is validated using the benchmark dataset and compared the performance with standard community detection algorithms. The results are promising and able to derive meaningful clusters from the dataset.

KEYWORDS: **Clustering, Gene Ontology, PSO, Optimum Path, Semantic Similarity**.

SREEJA ASHOK
Department of Computer Science & I.T.Amrita School of Arts & Sciences, Kochi,Amrita Vishwa Vidyapeetham, Amrita University, India

INTRODUCTION

A lot of research is going on in the field of genetics, which could improve our understanding of genes and their functions. Research labs around the globe are producing large amounts of biological information. This information can be used in various areas of health care; disease diagnosis and treatment, evolutionary biology, anthropology, etc. Efficient methods are needed for processing this large amount of gene data that is being produced. Clustering is the methodology of grouping a set of objects in a specific manner, such that the objects in the same collection, called a cluster, are more identical in one way or other to each other than to those in other cluster groups[24]. The elements in a cluster must have high intra- cluster similarity while keeping a low inter-cluster similarity. Genetic clustering ascertains the degree of similarity of genes and groups them into clusters. These clusters will contain genes which are highly similar to those within the cluster than those which are outside the cluster. This type of grouping of genes can reveal functional similarity of closely related genes.

BACKGROUND

There are many diverse methodologies which are used to discover clusters in large datasets which include Partition-based, Hierarchical, Density based, Grid based and Graph based Clustering[1,17].Partition- based clustering technique divides the data into *k* partitions, where each cluster tries to optimize the clustering criteria. Hierarchical technique involves performing a hierarchical decomposition of the elements, by grouping them into a hierarchy (or tree) of clusters. This can be agglomerative (bottom-up) or divisive (top-down).Density-Based method clusters objects in accordance with the density of objects. In this approach the clusters continues growing till the number of objects in the neighborhood reaches a particular value. DBScan is the most commonly used density based clustering method. Grid based is a space-driven approach which quantizes the object space into a finite number of cells prearranged as a grid structure. Graphs are widely used in modeling variety of entities and their inter-relationships. Objects or entities are represented as nodes or vertices in a complete or connected graph and their relationship is represented as edges. Clustering can be applied into graphical data for the identification of well-connected or well related components within it. Graph-clustering is the process of grouping vertices of a graph into relative clusters, in such a way that there should be large no of edges within each cluster and relatively few between the clusters[3].

Clustering Graphical Data

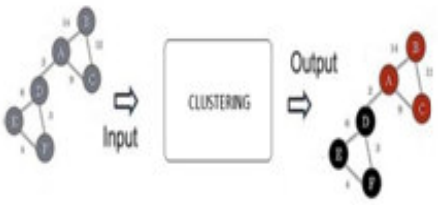

Figure 1

Figure 1 represents a sample clustering using graphical approach. A graph is typically denoted as G (V, E) where, V is the set of Vertices and E is the set of Edges. If each single vertex is connected with an edge to every other vertex in a graph, it forms a Complete Graph. A complete graph will have n (n-1)/2 edges, where n is the number of vertices.

A Complete Graph with 7 Vertices

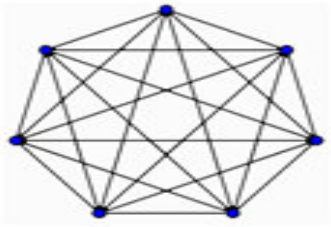

Figure 2

Figure 2 represents complete graph where the interpretation of meaningful relations are very difficult.

The existing clustering algorithms have their own drawbacks. The most commonly used Partition based clustering algorithms like K-Means divides a dataset into a specified no of clusters, that is, user need to explicitly specify the no of clusters. These approaches are extremely sensitive to parameters. The user must provide parameter values that suit the applications requirements[2] . In this paper, we intend to put forward a clustering approach that implements swarm intelligence by using Particle Swarm Optimization technique. A method that does not rely much on user given parameters, and which can analyze data more efficiently

and reveal accurate naturally occurring clusters from genome data.

(i) *Gene Ontology*

Gene ontology[4]consists of functional gene annotations that are hierarchically arranged in such a way that, it shows the relationship between genes and associated biological terms there are three ontologies[11]; There are mainly divided into components, Cellular component which characterizes the parts of a cell and its extracellular environment, Biological process which is the set of molecular events with a defined beginning and termination, relatable to the working and features of cells, tissues, organs and organisms and finally Molecular function that is the molecular level activities of a gene product, like binding or catalysis.

(ii) *Particle Swarm Optimization (PSO)*

Particle Swarm Optimization, categorized under swarm intelligence was developed by James Kennedy and Russell Eberhart in 1995. This method based on the social behavior of creatures resembles the natural behavior of birds flocking or fish schooling[6]. The swarm achieves the best solution interactions between the swarm members. A member with a better solution will inform the other members, and they will move towards it. This continues till an optimal solution is met[7]. PSO is a stochastic population based approach that evaluates the location and velocity of particles in a multi-dimensional space[23]. It updates position and velocity, the global optimum (or G_{OPT}) value which is derived from the replicated social behavior of the particles [8-10,19-22].Here we are implementing the concept of PSO for finding out the optimal path from the graph. In this approach every particle has a current position, personal best position, and global best position. The particles move in the search space to find out the optimal path from the graph. Every particle tries to move towards the particle that has the best solution, that is the global best position. After a specific number of iterations all the particles stop to a particular solution which is optimal. The general formula is given in equation[1]

$$V_{id}^{new} = W \times V_{id}^{old} + C_1 \times rand_{id} \times \left(P_{id} - X_{id}^{old}\right) + C_2 \times rand_{2d} \times \left(P_{gd} - X_{id}^{old}\right) \quad (1)$$

General Pseudo code of the algorithm is detailed below

1. Randomly Generate Initial Population.
2. Repeat, until termination criterion is met.
3. For each particle *i* to max.

3.1 compute the best fitness value of function.
3.2 update best position.
3.3 update global position.

4. End for.

PROPOSED METHOD

Here we propose a clustering approach that uses swarm intelligence for extracting the closely related functional groups of genes. The proposed methodology involves the following steps:

(i) Step1: Compute Semantic Similarity Matrix based on Gene Ontology
(ii) Step2: Generate Semantic Similarity Graph and form Hamilton paths
(iii) Step3: Find the global optimal path from the graph using the PSO based merging approach
(iv) Step4: Using the optimal path extract the clusters from the graph.

Step 1 Semantic Similarity - Measurement based on GO

One of the typical methods to measure semantic similarity is Information Content (IC) based method. IC-based similarity measures depend primarily on two factors; One is the frequency of the two GO terms considered and the other is, their closest common ancestor term within in a specific collection of GO annotations. A rarely used term usually consists of more information. The equation (2) defines the method for calculating the frequency of a term *t*;

$$p(t) = \frac{n_{t'}}{N} \mid t' \in \{t, children\ of\ t\} \quad (2)$$

Where *n* is the number of term *t* and *N*is the total number of terms in GO corpus.
The information content of a GO term is calculated by negative logarithmic probability of the term.
The information content(IC) is defined in equation (3):

$$IC(t) = -\log\left(p(t)\right) \quad (3)$$

Since GO allows multiple parents for each concept, it is possible to have two terms with common parents with multiple paths. IC based methods finds the similarity of GO terms based on their common ancestor term's information content. Such a common ancestor term is also known as Most Informative Information Ancestor (MICA).There are four IC based methods that are mainly used, proposed by Resnik, Jiang, Lin, and Schlicker [12-16].

Resnik Method

The method Proposed by Resnik Philip uses the following equation for computing the similarity.

$$sim_{Resnic}(t_1, t_2) = IC(MICA) \quad (4)$$

Lin Method
This method was proposed by Dekang Lin. It is defined as

$$sim_{Lin}(t_1, t_2) = \frac{21C(MICA)}{IC(t_1)+IC(t_2)} \quad (5)$$

RelMethod
TheRel method or Relevance method proposed by Schlicker is a combination of Resnik's method and Lin's method. It uses the following equation to find out the similarity.

$$sim_{Rel}(t_1, t_2) = \frac{21C(MICA)(1-p(MICA))}{IC(t_1)+IC(t_2)} \quad (6)$$

JiangMethod
This method proposed by Jay J. Jiang and David W. Conrath extracts Semantic similarity, based on the equation defined in (7).

$$sim_{Jiang}(t_1, t_2) = 1 - min\big(1, IC(t_1) + IC(t_2) - 2IC(MICA)\big) \quad (7)$$

The similarity matrix is an *(N x N)* square matrix which represents the pair wise similarity between*N* genes. The matrix will have rows and columns equal to gene size. Each cell *(u, v)* of the matrix will contain the similarity between the genes *u* and *v*, where *u* and *v* can be of any value from *1* to *N*.

Step 2 Graph Representation and Creation of Hamiltonian Paths

The similarity matrix can be considered as a fully connected graph with N nodes, with each pair of nodes having an edge connecting them. The edges between the nodes are assigned a weight that signifies the similarity. Hamiltonian paths are derived from the fully connected graph formed from *N*N*similarity matrix. A Hamiltonian path includes all the nodes in the fully connected graphat least once and there should not be any repeated paths. A Hamiltonian*P*will have N nodes or vertices and *(N-1)* edges.

Step 3 Finding the Global Optimum Path
The solution space is defined as a set of Hamiltonian paths derived from the above step. For each iteration, obtain the local best solution and the global best solution which is a function of sum of distance of all edges connected in each path. Next step is to merge the particles with global best solutions in each iteration to derive a set of new solutions. Merging is the process of combining one particle with the global best particle in such a way that the particle converge more closely to the global bestpath. Let a path *P* is represented as$P(p_1,p_2,p_3...p_n)$; G_{BEST}be the most optimum particle andL_{NEW} be the newpath. Set the start node of L_{NEW} is set as start node of G_{BEST}. Optimum edge from G_{BEST} andP_i is added to L_{NEW} until every node in the fully connected graph is added to L_{NEW}. The process repeats till the stopping criterion is met.

Step 4 Clustering based on the derived optimal path
Now we have derived the global optimum path from the similarity matrix of the gene list.The next step is clustering of genes based on the optimum semantic similarity path. The Hamiltonian global optimum path constitutes of a set of genes and their resemblance as edge value. We perform a value based clustering on the path derived to achieve clusters. We conduct a walkthrough over the global optimum path to find contiguous similarity values. The nodes or genes are clustered based on these contiguous similarity values. A set of contiguously occurring genes with similar weight is grouped together to form clusters. Accordingly we discover all the existing clusters from the gene list by individually grouping all those genes which have common weight.

PSO based gene clustering process flow

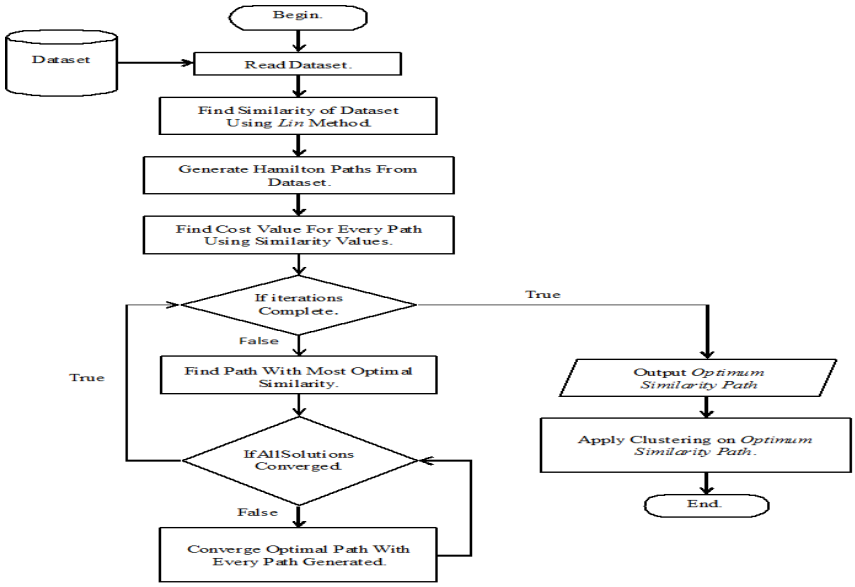

Figure 3

Figure 3 represents the flowchart of the proposed approach

Table 1

Description of variables used in the algorithm

Label.	Description.
geneList	List of Illumina's TruSight Target Cancer Genes.
N	No. of genes in *geneList*.
rMAT	*N* * *N* similarity matrix of *geneList* computed using *Lin* Method.
solCount	No. of initial solutions generated.
solutionMATRIX	*N* * (*N*+1) matrix of initial solution paths along with path similarity score
G_{OPT}	Solution with most optimum similarity score in *solutionMATRIX* at any iteration.
L_{OPT}	Solution with their local optimum similarity score in *solutionMATRIX* at any iteration.
L_{NEW}	New solution obtained by converging G_{OPT} and L_{OPT}.
itr	No. of iterations to be performed.
edgeSim	Similarity value of two Genes G_i and G_{i+1}.
pathSim	Total Similarity value of a Hamilton path $solution_i$.

Table 2

Algorithm For Finding Optimum Gene Similarity Path.

optimumGeneSimilarityPath (*geneList, N, itr, solCount, rMAT*)
Output: *Optimum Gene Similarity Path.*
Begin.
solCount← Input Number of initial solutions to be generated.
itr← Input Number of iterations performed on initial solutions.
rMAT←$sim_{Lin}(t_1, t_2) = \frac{2IC(MICA)}{IC(t_1)+IC(t_2)}$
solutionMATRIX←**generateInitialSolutions** (*geneList, solCount*).
Repeat until *itr*
G_{OPT}←$solution_i$ from *solutionMATRIX* with most optimal similarity score.
For every $solution_i$ in *solutionMATRIX*
L_{OPT}←$solution_i$.
Optimum Gene Similarity Path←**convergeSolutions** (G_{OPT}, L_{OPT}).
End.

Table 3

Algorithm For Generating Initial Solutions.

GenerateInitialSolutions (*geneList, solCount, rMAT*)
Output: *solutionMATRIX*(*solCount* * *N*+1)
Begin.
Repeat until *solCount*
$solutionMATRIX_i$←**createHamiltonPath** (*geneList*).
$solutionMATRIX_i$ ["SimilarityScore"] ←*pathSim*.
End.

Table 4
Algorithm For Creating Hamilton Paths.

```
createHamiltonPath (geneList, rMAT)
Output: solution_i, pathSim.
Begin.
    Select a random node/gene G_i from geneList.
    Set pathSim←0.
    Repeat until path has all nodes from geneList exactly one time.
        Get G_i+1 from geneListsuch that edge (G_i, G_i+1) is optimum among edges
        starting from G_i.
        Set edgeSim←rMAT [G_i, G_i+1].
        Set pathSim←pathSim + edgeSim.
End.
```

Table 5
Algorithm For Converging The Initial Solutions Generated.

```
convergeSolutions (G_OPT, L_OPT)
Output: L_NEW.
Begin.
    Set Start node of L_NEW← Start node of G_OPT.
    Repeat until all nodes in geneList are added to L_NEW
        Set G_SUC← Successor node of L_NEW [i] in G_OPT.
        Set P_SUC← Successor node of L_NEW [i] in L_OPT.
        Set L_NEW[i+1]← Optimum edge in (L_NEW [i], G_SUC) and (L_NEW [i], P_SUC).
    Set L_NEW← Optimum path in L_NEW, G_OPT.
End.
```

RESULT AND EXPERIMENT ANALYSIS

To implement the methodology proposed above, we acquired a cancer gene list from Illumina, Inc., a prominent Biotechnology company developing integrated systems for the analysis of genetic variation and biological function. The target cancer gene list has been published online particularly for research purposes. The *92* genes targeted have been identified with an association of predilection towards cancer. Genes included are related with two of the most common cancers; colon and breast along with some rare cancers. We employ information content-based Lin method to find out the semantic similarity between the genes from the extracted cancer gene list.

Semantic Similarity path derived using PSO based clustering

Figure 4
Figure 4 represents the global optimum solution derived from the set of 92 genes. We necessitate this global optimum for our clustering process

The outcome that we derived here is in the form of a similarity matrix, where rows and columns simultaneously depict the genes. And values show their similarity measure in accordance with the Semantic Similarity Measurement context based on Gene Ontology. R Programming Language and associated packages have been brought in for our implementation scenario. A set of initial solutions is generated for our convergence procedure. To begin with, a random gene is chosen from the *92* target cancer gene list. The next gene selected will have less similarity value and so on. That is, minimum similarity edges are added to the solution based on the last gene added. It goes on until all genes are added to a single solution. Like this we have a set of solutions each of which begin on the random gene we selected.The set of initial solutions we derived is made to converge using our methodology. The convergence will give the global optimum solution of the similarity matrix from the initial solutions. We necessitate this global optimum for our clustering process. This global optimum is the most optimal Hamilton Path from our similarity measures. From this path, we form the clusters by grouping together the genes based on weight.

Genes clustered using PSO

Figure 5
Figure 5represents the clustered graph which shows the gene interactions in each clusters.

We have analyzed the same data set using the prominent network community detection method *walktrap*[18]. It performs short random walks of steps specified as its parameter. Then the result of this random walk is used to determine separate communities in a bottom-up manner. The outcome shows only two clusters formed from the cancer gene list contrasting the result of our methodology. It implies the convergence mechanism formulated effectively reveal clusters from the dataset.

Table 5
Clusters formed by using community detection method walktrap

Clusters	Gene Symbols
Cluster 1	AIP, ALK, APC, ATM,BMPR1A,BMPR1A, BUB1B, CDC73, CDH1, CDK4, CDKN1C, CDKN2A, CEBPA, CEP57, CHEK2, CYLD, EGFR, EPCAM, EXT1, EXT2, FANCA, FANCB, FANCC, FANCD2,FANCF, FANCI,FANCL,FANCM,FH,FLCN,GPC3,KIT,MAX,MET,MUTYH,NF1,NF2,NSD1, PRF1, PRKARIA, PTCH1, PTEN, RBI, RET, SDHAF2, SLXL, STK11, SUFU, TMEM127, TP53, TSC1, TSC2, VHL.
Cluster 2	BAP1, BLM, BRCA2, BRIP1, DDB2, DICER1, DIS3L2, ERCC2, ERCC3, ERCC4, ERCC5, EZH2, FANCG, GATA2,HNF1A,HRAS,MEN1,MLH1,MSH2,MSH6,NBN,PALB2,PHOX2B, PMS1, PMS2, RAD51C, RAD51D, RECQL4, RUNX1, SBDS, SDHB, SDHC, SDHD, SMAD4, SMARCBI, WRN, WT1, XPA, XPC.

CONCLUSION

The proposed method could able to detect the optimum clusters in a gene network, which helps in proper identification of the functional relationships of similar genes. These gene relationships that we exposed will be instrumental in pharmacogenomics, disease prognosis, etc. The performance is evaluated by comparing clustering sizes using existing community detection algorithms. The future plan is to apply the same methodology in large dataset and exploring the common functionalities that differentiates each cluster using feature extractions.

ACKNOWLEDGEMENT

This work is supported by the DST Funded Project, (SR/CSI/81/2011) under Cognitive Science Research Initiative in the Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham University, Kochi.

CONFLICT OF INTEREST

The authors state that the present manuscript presents no conflict of interest.

REFERENCES

1. Han J, Kamber M, Pei J. Data mining: concepts and techniques. Elsevier; 2011 Jun 9.
2. Madhulatha TS. An overview on clustering methods. arXiv preprint arXiv:1205.1117. 2012 May 5.
3. S. E. Schaeffer, “Graph clustering,” Computer Science Review, vol. 1, no. 1, pp. 27–64, 2007
4. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004 Jan 1;32(suppl 1):D258-61.
5. Nagar A. Functional Analysis of Genes Using the Gene Ontology: Gene Similarity, Clustering, and Classification. ProQuest; 2008.
6. Rini DP, Shamsuddin SM, Yuhaniz SS. Particle swarm optimization: technique, system and

challenges. International Journal of Computer Applications. 2011 Jan;14(1):19-26.
7. Bhagat A, Jain S. A FUZZY BASED PSO APPROACH TO FIND OPTIMAL ROUTE IN MOBILE ADHOC NETWORK.
8. Toofani A. Solving routing problem using particle swarm optimization. International Journal of Computer Applications. 2012 Jan 1;52(18).
9. Kennedy, James. "Particle swarm optimization." *Encyclopedia of machine learning*. Springer US, 2011. 760-766.
10. Kuo RJ, Syu YJ, Chen ZY, Tien FC. Integration of particle swarm optimization and genetic algorithm for dynamic clustering. Information Sciences. 2012 Jul 15;195:124-40.
11. Kustra R, Zagdanski A. Incorporating gene ontology in clustering gene expression data. InComputer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on 2006 Jun 22 (pp. 555-563). IEEE.
12. Yu G. GO-terms Semantic Similarity Measures. Bioinformatics. 2010 Sep 1;26(7):976-8.
13. Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res.(JAIR). 1999 Jul 11;11:95-130.
14. Lin D. An information-theoretic definition of similarity. InICML 1998 Jul 24 (Vol. 98, pp. 296-304).
15. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC bioinformatics. 2006 Jun 15;7(1):302.
16. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008. 1997 Sep 20.
17. Sreeja Ashok, M.V.Judy. A novel iterative partitioning approach for building prime clusters. International Journal of Advanced Intelligence Paradigms. 2015;7(3-4):313-25.
18. Sonumol NS, Uma VR, Ashok S, Judy MV. Community Detection in Multidimensional Genomic Dataset. International Journal of Artificial Intelligence™. 2015 Sep 22;13(2):109-17.
19. Kiranyaz S, Ince T, Gabbouj M. Multidimensional particle swarm optimization for machine learning and pattern recognition. Springer; 2014.
20. Grosan C, Abraham A, Chis M. Swarm intelligence in data mining. InSwarm Intelligence in Data Mining 2006 (pp. 1-20). Springer Berlin Heidelberg.
21. Talukder S. *Mathematical modelling and applications of particle swarm optimization* (Doctoral dissertation, Blekinge Institute of Technology).
22. Kumbharana NS, Pandey GM. A comparative study of ACO, GA and SA for solving travelling salesman problem. International Journal of Societal Applications of Computer Science. 2013 Feb;2(2):224-8.
23. Anusuya Venkatesan, Dr Latha Parthiban And K.Arul. ROI Detection And Segmentation Of Medical Images Using Optimized Thresholding And Clustering. International Journal of Pharma and Bio Sciences. 2013 July; 4(3): (B) 1235 – 1245.
24. NALINI D, Revathi S. An Enhanced Clustering Algorithm Implemented On Biological Data In Data Mining. International Journal of Pharma and BioSciences. 2013 ; 4(2): 1281.