International Journal on Emerging Technologies 11(2): 516-523(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Markov based Genetic Algorithm (M-GA): To Mine Frequent Sub Components from Molecular Structures

P. Kanmani, O.S. Gnana Prakasi and N. Jayapandian
Assistant Professor, Department of Computer Science and Engineering,
Christ (Deemed to be University), Bengaluru (Karnataka), India.

(Corresponding author: Kanmani P.)
(Received 30 December 2019, Revised 26 February 2020, Accepted 28 February 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Processing the molecular compounds to identify the internal chemical structure is a challenging task in bio-chemical research. Popular approaches, mine the frequent subcomponents from the molecules with chemical and biological properties represented in the form of feature vector histogram. Though this helps to identify the absence or presence of mined feature, calculating the frequency of every frequent substructure involves sub graph isomorphism test which is an NP-Complete process. To overcome the above mentioned bottleneck we proposed Markov based Genetic algorithm (M-GA) in which the chemical descriptors were considered from two-dimensional representations of molecules that classify chemical compounds using mining significant substructure and generates the binary vector that generate pure active classes, singleton reactors, descriptor sets. This method scales down the process of mining substructures that are statistically significant from huge chemical databases. The results shows that the performance of proposed algorithm is improved compared to the existing algorithms.

Keywords: Molecular substructure, Histogram, Markov model, Memetic Operator, Fitness function.

Abbreviations: SVM, Support Vector Machine; RW, Randam Walk; M-GA, Markov based Genetic Algorithm; ROC, Receiver Operating Characteristic; AT, Angiotension II receptor; PI, Protease Inhibitor; RTI, Reverse Transcriptase inhibitor; AUC, Area Under the Curve.

I. INTRODUCTION

Large availability of molecular composites has created high difficulties in the effective design of molecular probing and mining framework systems. Classification of molecular composites is an important task for drug development where collections of molecules are screened and tested for the highest probability of success against a given target. Graphsig scaffolds lead generations are the various mining tools that can significantly mine the represented molecular composites. SVM classifiers play an important role in lead generation and lead optimization and they classify the chemical structures using molecular classification. Neural network technology, Bayesian Model, Kernel-based methodologies statistical or cell based partitioning, recursive partitioning techniques sub graph mining techniques are the various techniques that are used for mining chemical structures. Identification of the chemical or biological properties of molecules using well-known quantization structure activity relationship (QSAR) approaches from IS structure is a major goal of mining the chemical compounds. The critical work in processing the molecular composites is to identify the embedded chemical information in the structure.

A number of classification methods are proposed in [5, 8, 9, 12-14] on the feature vector representation of molecules. Structural key [8, 13] and hashed fingerprints [17] are the existing vector-based approaches. Fixed width bit vector representation is used as the fingerprint structure of a molecule. Every linear paths and cycles are enumerated into finger prints and hashed into a vector resulting in a compact form.

Other methods using structural keys [11,15] have a dictionary of features that can be used for screening every molecules resulting in vector representation. This dictionary information is known in prior and all the information is collected purely based on the domain knowledge. Recent researchers concentrate on dynamic feature selection techniques that employ molecular characterization. Popular approaches mine the frequent subcomponents from the molecules where the mined structures with chemical and biological properties [16] are represented in the form of feature vector histogram. This helps to identify the absence or presence of mined features.

Information ratio and Gain ratio are considered as the attribute selection measures. Based on the information theory, the content of data under [22] concern helps to retrieve the information from the attributes with strong feature values.

This technique uses a frequency threshold value and all the selected subcomponents are used as the descriptors. In the straight forward approach the solution to remove not statistically significant frequent substructures [18]. This approach may not be scaled to large dataset due to two reasons. Our goal is to mine the molecules at a threshold of very low frequency.

Fusion of medical images and multimodal algorithms focuses on classification of similar and dissimilar image signals [23]. Low and high level frequency coefficients help to extract the exact structural properties of critical clinical studies. Hybrid wave fusion algorithm was proposed by the authors to enhance the wavelet domain information extraction methods.

Graphsig paper provides a natural structural representation of molecular in which nodes are formed from atoms, covalent bonds between atoms forms the edges and edge labels are related to bond order. Graph-based representation is used and treated as graph databases and the problem of identifying significant molecular substructures is reduced to significant sub-graph identification.

Graph based representation followed by SVM classifier [10] is used to mine pattern and biological activities in the paper [6]. The method for examining supramolecular interactions by analyzing the chemical/biological properties and target scores promotes to ensure and optimize a specific pharmacophore. The proposed approach is an outstanding deviation from existing techniques that justifies compounds on a target-by-target basis. The basic geometry of the pharmacophores is merely the point to bind between compounds and targets as well as properties of compounds [3]. Mining the molecular sub structures having P-value lesser than given threshold helps to examine the structures to generate chemical descriptors.

To identify the similar pharmacophores the analysis over the geometric features helps to categorize assortment space of the sub structures. The feature classification gives both true positive and false negative characteristics of structures based on the chemical activity. Different level of research was performed to authenticate identification of presence of subspaces which leads to biological related activities. The existing model works only with the activeness of molecules and then labels the substance [2, 4]. The proposed model will solve the problem of identifying significant molecular substructures by identifying significant sub-graph. The method also validates the activeness of the molecules present in the composition by measuring the precision, recall and similarity indexing values.

Few mining methods for extracting subcomponents involves sub graph isomorphism test which becomes an NP-Complete process when the size of the database is large. Most of the mining methods focus on the mining feature by setting the threshold value and computation of these values depends on the frequency of the subcomponents. So the complexity of the process is increased.

To overcome, in this paper we proposed a Markov based Genetic algorithm [1] (M-GA) to mine the frequent sub-components from the molecular structure. In this, we considered the two-dimensional representations of the molecules from which the chemical properties are estimated using the proposed binary vector. The main advantage of this method is that the generated scoring classes show the pure active classes, singleton reactors, descriptor set.

II. PROPOSED METHOD

The molecular database NCI/NIH AIDS shown in Table1.is the input set used to mine the frequent substructures by extracting the graph of respective structure under analysis. The molecular descriptor is deployed on the retrieved graphs which are treated in the form of encoded binary strings. The identified strings are treated as chromosomes and the fitness evaluation is computed for the same and the iterations will be continued until optimal convergence is reached. The Markov blanket function based memetic operator is applied on the converged substances to highlight the strong chromosomes which can be categorized for active molecule substance.

The input molecular databases are considered as graph and the problem of identifying a significant substructure is solved by finding most significant molecular substructure. Fig. 1 represents a graph structure of benzene. The labeled graph for the representation of molecules is been structured where the atoms are represented as nodes and edges as covalent bonds in-between the atoms. The edge labels correspond to bond order. To identify the implication of any given substructure we have to quantify the measures using the impact factor called P-value [7]. Mined molecule substructure having P-value lesser that user-specified threshold are used to mine the structures to generate chemical descriptor for classification. We have modified and used Markov based Genetic method to classify the mined structure effectively.

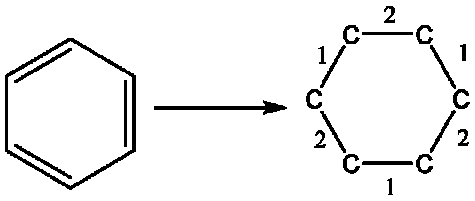


Fig. 1. Conversion of benzene to graph.

A. Domain Knowledge on Molecular Graph

The domain knowledge is converted into graph based on physio-chemical characteristics. Fig. 2. and 3 shows an example of converting domain knowledge into graph. Reliability of atoms allows us to differentiate between similar atoms in different states or aromatic ties and this was performed using Joelib2 atom type [6]. The presences of functional groups are identified for all the molecules and the important functional groups are abstracted out and are represented by a single node. This is a well-known approach and it showed a better performance than before. Performing this node in a molecular graph is represented either as functional groups are identified rather than individual actual atoms. During this performance some structures overlaps which can be solved by abstracting one from these. The order dependency is solved based on frequency and size. Four different ordering is performed based on frequency and size, frequency increase, frequency decrease, size increase and size decrease. All the priorities are done in descending order of the size to avoid overlapping.

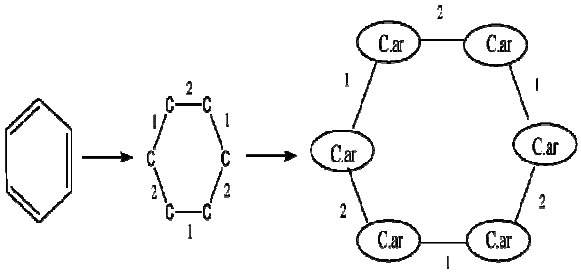


Fig. 2. Domain Knowledge on molecular graph.

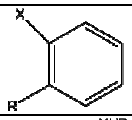
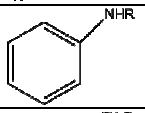
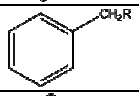
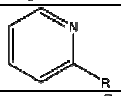
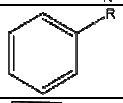
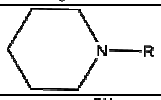
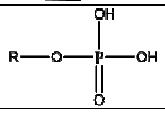
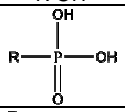
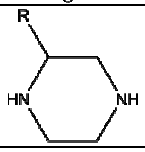
B. Representation of Histogram

A random walk shown in Fig. 4 with restart is applied to convert the functional groups into histogram. This captures the distribution of neighborhood node-node pairs into histogram. This RW starts with a target node and jumps from one node to neighbor, where every neighbor has probability of being a new station for every walker. It is been maintained that the walker will not go very far from one node to neighbor where every neighbor has the probability of being a new station for every walker [19]. It is been maintained that the walker will not go very far from the starting node. It has to capture only the neighborhood and not the complete molecule. The

probability of is that to bring back the walker to the starting node. This RWR is performed at various iterations to calculate node-node pair formed from 2 nodes sharing the same bond like C-1-C.The value of NNP is calculated as in (1)

$$\text{NNP value of T} = \frac{\text{No. of times NNP of type T visited}}{\text{No. of jumps by the walker}} \quad (1)$$

Table 1: NCI/NIH AIDS DATABASE is analyzed to show the performance of the frame work.

S. No.	Functional Group Name	Functional Group Formula	Structure
1.	Halo benzene Derivatives	RC_6H_4X	
2.	Aniline Derivatives	C_6H_5NR	
3.	Toluene Derivatives	$C_6H_5CH_2R$	
4.	Substituted Pyridines	RC_5H_4N	
5.	Substituted Benzene	C_6H_5R	
6.	Piperidine Derivatives	$C_5H_{10}NR$	
7.	Phosphates	RPO_4	
8.	Alcohol	ROH	R-OH
9.	Phospho Derivatives	$RPO(OH)_2$	
10.	Piperazine Derivatives	$C_4H_{10}N_2$	

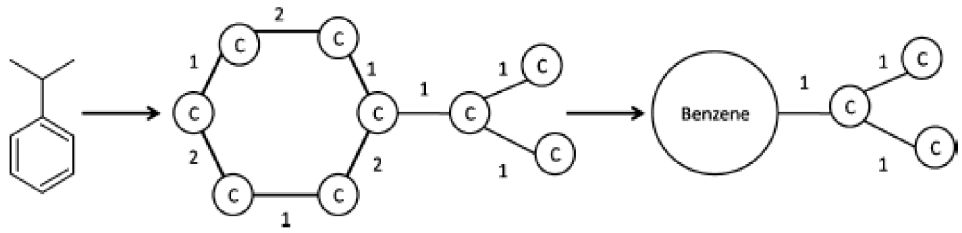
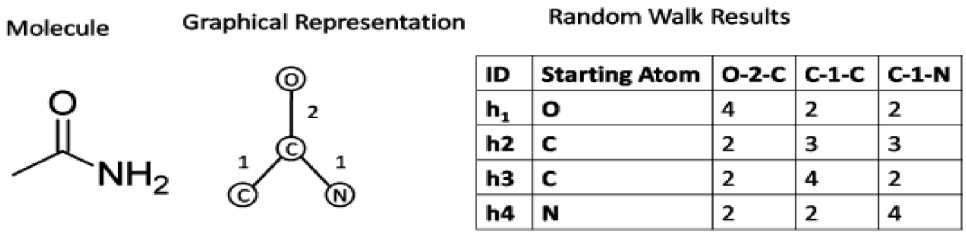


Fig. 3. Extraction of benzene functional group from cumene.



ID	Starting Atom	O-2-C	C-1-C	C-1-N
h_1	O	4	2	2
h2	C	2	3	3
h3	C	2	4	2
h4	N	2	2	4

Fig. 4. Random Walk of Carboxmide.

C. Molecular Classification

A significant histogram has to be mined from all the final groups based on the query molecule, a binary vector representation is created. This has been improvised using genetic based M-GA which characteristics the optimum significance of structure [20] to create a molecular descriptors from which the classification is improved to the greater level.

Algorithm 1: Deployment of molecular Descriptor
Input : Histograms collected H, Query Q
Output : Significant histogram S, Binary Representation V
(i) Initialize random population of generated histogram encoded with binary string (ii) While not converged (iii) Evaluation of fitness is performed using the fitness using J(Sc) for all the significant histogram. (iv) Choose the elite chromosome using Markov Blanket based memetic operators. (v) end while

Algorithm 2: Add Operator

Begin
For each node a€Q do
i←o
while i< |H| do
S→H(i)
If S□RWRhist(a) then
V(i)= = i;
i=i+1;
return V;

Algorithm 3: Markov Based Memetic Operator
(i) Select Elite chromosome C_b to perform memetic operator (ii) for z = 1 to Z^2 (iii) Select a unique random value having $0 < a < Z$ (iv) Apply a times, Add operation to generate a new chromosome C'_b (v) Calculate fitness for the modified chromosome based on J(Sc) (vi) Replace the genetype C_b and stop memetic operation (vii) End

Initial Population is randomly analyzed using GA and M-GA optimizes the local search and the fitness is calculated using Eqn. (2)

$$\text{Fitness } (c)=J(Sc) \quad (2)$$

where Sc is the selected feature encoded in the form of chromosome and the objective function evaluates the significance using cross validation or bootstrapping methodology. The one with the best fitness goes for the next generation (21). Thus the fittest is been selected from the mined histogram substructure to retrieve the significant substructure for effective classification.

III. RESULTS AND DISCUSSION

The classification is measured by calculating the area under the ROC curve and the area under curve AUC methods gives the measure for the classification accuracy. All the classifiers are expected to achieve AUC of 0.5 shown in Table 2. The mined structures are further classified into active and inactive by referring the molecular values shown in sample Table 3.

The input molecular structures are properly encoded to generate a set of candidate sub graphs [21]. The generated sub structures and the respective histograms screens the significant sub structures to reduce the area of interest (Fig. 5).

This step helps to eliminate the redundant and irrelevant features based on the predictive capability and Markov operator genetic features of adaptation.

Table 2: AUC Comparison between GraphSig and Proposed system.

Data Set	GraphSig	M-GA
MCF-7	0.90+-0.01	0.90+-0.01
MOLT-4	0.89+-0.02	0.90+-0.02
NCI-H23	0.93+-0.02	0.92+-0.02
OVCAR-8	0.93+-0.02	0.92+-0.02
P388	0.91+-0.02	0.92+-0.02
PC-3	0.93+-0.02	0.93+-0.01
SF-295	0.94+-0.02	0.93+-0.02
SN12C	0.93+-0.02	0.93+-0.03
SW-620	0.93+-0.03	0.93+-0.03
UACC-257	0.93+-0.02	0.92+-0.02
Yeast	0.82+-0.03	0.87+-0.03
Average	**0.913+-0.02**	**0.921+-0.02**

The binary vector is computed by using MBEGA, is a random walk based concept along with dynamic iterations with completeness of classification. The binary classifier thus results in two classes as active molecules and inactive molecules. Fine tuning and cultural local search not only converge to high quality solutions and also improve the computational accuracy and robustness.

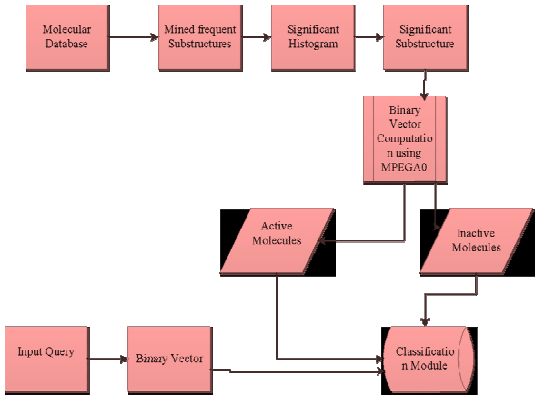
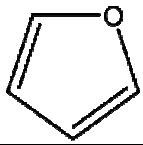
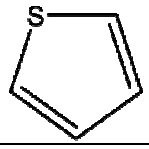
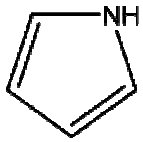
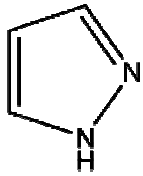
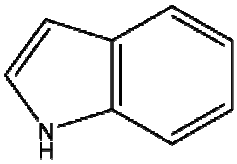
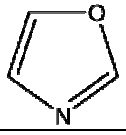
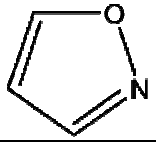
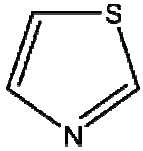


Fig. 5. Block Diagram.

Add and delete operations rank the features in the descending order and the segregation is based on the correlation value. Based on the linear ranking method the features falls under less likely window are eliminated to normalize the Markov blanket module. The selected features are from the most likely window. The experimental result shows that locating local optimum in sufficient precision count under the region of convergence is improved with the support of dimensionality reduction.

All molecules can be determined, as "active" or not, but the higher the value, the higher the dose needed to achieve the inhibitory effect and the higher the chance of "off-target" drug binding events, and hence, potential toxicity. The concentration needed to realize the potential toxicity is molecule-, cell-, tissue-, organ- and organism-specific. Acceptable concentration for depends on the sub structures involved in binding and thus the classification is more supportive for medical analyst and drug prediction process.

Table 3: Structure and Pharmacological activities of sample active components considered for experimentation.

Name & Structure	Pharmacological activities						
	Antihistamine	Anti Microbial	Anti-viral	Anti-Inflammatory	Anti Protozoal	Cns Depressant	Anti Convulsant
Furan 	No	Yes	Yes	Yes	No	No	No
Thiophene 	No	Yes	Yes	Yes	No	No	Yes
Pyrrole 	No	Yes	Yes	Yes	Yes	Yes	Yes
Pyrazole 	No	Yes	Yes	Yes	Yes	Yes	Yes
Indole 	Yes	Yes	Yes	Yes	Yes	Yes	No
Oxazole 	Yes	Yes	Yes	Yes	No	No	No
Isoxazole 	Yes	Yes	Yes	Yes	No	No	No
Thiazole 	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Triazole	Yes	Yes	Yes	Yes	Yes	No	Yes

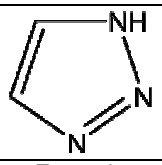
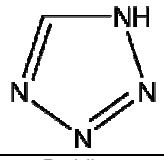
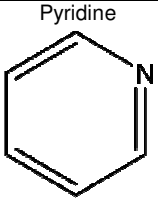
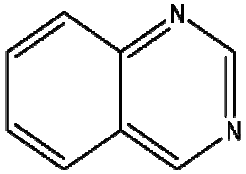
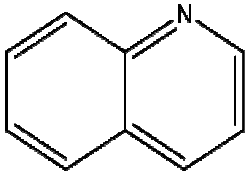
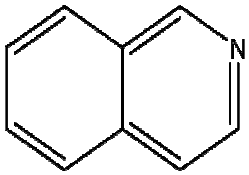
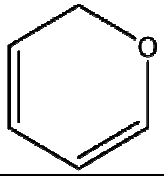
							
Tetrazole 	Yes	Yes	Yes	Yes	Yes	No	No
Pyridine 	Yes	Yes	Yes	No	No	No	No
Quinazoline 	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quinoline 	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Isoquinoline 	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pyran 	No	Yes	Yes	No	No	No	No

Table 4: Adaptability measure by comparing different dataset.

Data Set	GraphSig	M-GA
MCF-7	121	123
MOLT-4	253	254
NCI-H23	158	156
OVCAR-8	154	154
P388	124	130
PC-3	149	150
SF-295	171	169
SN12C	170	171
SW-620	211	210
UACC-257	143	142
Yeast	159	157
Average	**164.8**	**165.09**

Table 5: Precision/ Recall Achieved on MDDR Data set by GraphSig and M-GA.

Class	GraphSig		M-GA	
	Precision	Recall	Precision	Recall
AC	0.97	0.97	0.99	0.97
AT	0.99	0.94	0.98	0.95
PR	0.94	0.92	0.95	0.92
RT	0.92	0.94	0.95	0.94
Average	0.96	0.94	0.97	0.95

The adaptability measure of various data sets is shown in the Table 4. To further identify the adaptability of the proposed methods over the MDDR (MDI Drug Data Report) the precision and recall are identified for the four groups like ACE inhibitors (AC), Angiotension II receptor(AT), Protease inhibitor (PR) and reverse transcriptase inhibitor(RI) are analyzed and shown in Table 5.

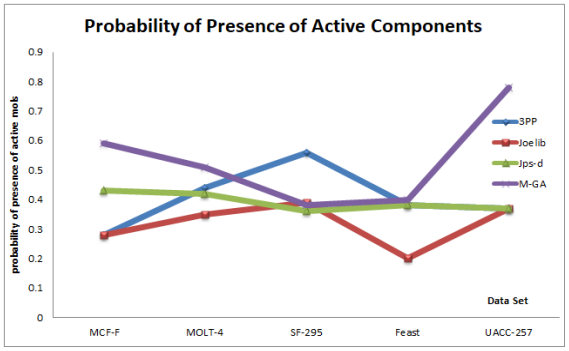


Fig. 6. Probability for the success rate of proposed model with respect to datasets.

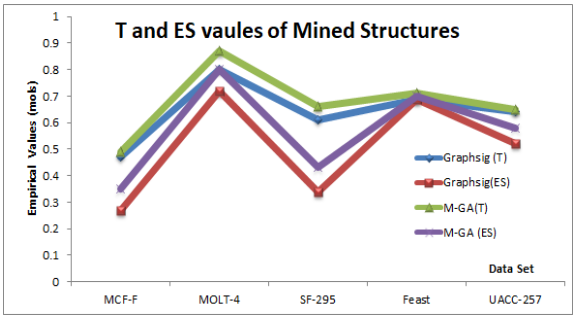


Fig. 7. Comparison of GraphSig and Proposed MGA mined structures in terms of T and ES.

Fig. 6 and 7 shows the chances of occurrence of active and frequent sub components in the mines sub structures. Every mined molecular structure from referred dataset combined with most related substructure from both the sets and the average of these duos' are considered as the similarities. Best results for each are shown in the graph in terms of their empirical values.

IV. CONCLUSION

The proposed work defines a new substructure mining algorithm, MBG, which mines molecular substructures and then classifies the structures into active and inactive molecules. The algorithm does the best classification and a diligently correlated task to bring structural classification based on the molecule compounds.

The structure miner can still be improved by incorporating similarity indexing and clustering to identify the contribution of each chemical substance in high dimension in terms of physio-chemical properties. Given experimental results are applicable in machine learning computational tasks in drug compound prediction and design.

V. FUTURE SCOPE

The proposed system results in better identification of the combined sub molecules and the final structure for the drug. The substructure miner can be designed as network model to train with the combinations of possible compounds and then classifier can be extended to work with the trained knowledge base. The future idea can reduce the time taken to identify the proper combinations and thus lead to perfect classifier in research domain.

Conflict of Interest. No.

REFERENCES

[1]. Xue, L., & Bajorath, J. (2000). Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *Journal of Chemical Information and Computer Sciences*, *40*(3), 801-809.

[2]. Ranu, S., Calhoun, B. T., Singh, A. K., & Swamidass, S. J. (2011). Probabilistic Substructure Mining From Small-Molecule Screens. *Molecular Informatics*, *30*(9), 809-815.

[3]. Nidhi., Glick, M., Davies, J. W., & Jenkins, J. L. (2006). Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *Journal Chemical Information Modeling, 46*(3), 1124–1133.

[4]. Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal, 17*(6), 441–458.

[5]. Ranu, S., & Singh, A. K. (2009). Mining Statistically Significant Molecular Substructures for E□cient Molecular Classification. *Journal Chemical Information Modeling, 49*(11), 2537–2550.

[6]. Swamidass, S. J., Chen, J., Bruand, J., Phung, P., Ralaivola, L., & Baldi, P. (2005). Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics, 21*(1), 359–368.

[7]. Swamidass, S. J., Azencott, C. A., Daily, K., & Baldi, P. (2010). A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics, 26*(10), 1348–1356.

[8]. Ning, X., Rangwala, H., & Karypis, G. (2009). Multi-Assay-Based Structure-Activity Relationship Models: Improving Structure-Activity Relationship Models by Incorporating Activity Information from Related Targets.

Journal Chemical Information Modeling, 49(11), 2444–2456.
[9]. Kumar, R., Tyagi, S.,& Sharma, M. (2013). Memetic Algorithm: Hybridization of Hill Climbing with Selection Operator. *International Journal of Soft Computing and Engineering*, *3*(2), 140-145.
[10]. Alba, E., Garcia-Nieto, J., Jourdan, L., & Talbi, E. G. (2007). Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. *In Proceedings of the International Conference on IEEE Congress on Evolutionary Computation (CEC'07), Singapore,* 25-28, 284-290.
[11]. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences, 96*(12), 6745-6750.
[12]. Lu, H., & Tonge, P. J. (2010). Drug-target residence time: critical information for lead optimization. *Current Opinion in Chemical Biology, 14*(4), 467–474.
[13]. Abrams, C., & Bussi, G. (2013). Enhanced sampling in molecular dynamics using meta dynamics, replica-exchange, and temperature-acceleration. *Entropy, 16*(1), 163–199.
[14]. De Vivo, M., Masetti, M., Bottegoni, G., & Cavalli, A. (2016). Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry*, *59*(9), 4035-4061.
[15]. Safi, M., & Lilien, R. H. (2012). Efficient a priori identification of drug resistant mutations using Dead-End Elimination and MM-PBSA. *Journal of chemical information and modeling*, *52*(6), 1529-1541.
[16]. Wallach, I., Jaitly, N., & Lilien, R. (2010). A structure-based approach for mapping adverse drug reactions to the perturbation of underlying biological pathways. *PloS one*, *5*(8), 1-11.
[17]. Taylor, R., Miller, R. H., Miller, R. D., Porter, M., Dalgleish, J., & Prince, J. T. (2015). Automated structural classification of lipids by machine learning. *Bioinformatics*, *31*(5), 621-625.
[18]. Huan, J., Wang, W., Prins, J., & Yang, J. (2004). SPIN: mining maximal frequent subgraphs from graph databases, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA,* 581 – 586.
[19]. Yan, X., & Han, J. (2002). gSpan: graph-based substructure pattern mining. *Proceedings of the IEEE International Conference on Data Mining, Maebashi City, Japan,* 721-724.
[20]. Huang, C. F. (2004). A Systematic Study of Genetic Algorithms with Genotype Editing, *Proceedings of the IEEE International Conference on Genetic and Evolutionary Computation, Seattle, WA, USA,* 1233-1245.
[21]. Lameijer, E. W., Bäck, T., Kok, J. N., & Ijzerman, A. P. (2005). Evolutionary algorithms in drug design. *Natural Computing*, *4*(3), 177-243.
[22]. Dewangan, A. K., Shrivas. A. K., & Kumar, P. (2016). Classification of Thyroid Disease with Feature Selection Technique. *International Journal of Engineering and Techniques*, Vol. 2 Issue 3.
[23]. Kumar, A., & Khandelwal, A. (2019). A Review on Multimodal Medical Image Fusion Based on DWT. *International Journal of Electrical, Electronics and Computer Engineering, 8*(1), 55-58.

How to cite this article: Kanmani, P., Prakasi, K. O. S. G. and Jayapandian, N. (2020). Markov based Genetic Algorithm (M-GA): To Mine Frequent Sub Components from Molecular Structures. *International Journal on Emerging Technologies*, *11*(2): 516–523.